EJERCICIOS de ECONOMETRÍA BÁSICA

Juan Equiza Goñi

Universidad de Navarra

EXERCISES of BASIC ECONOMETRICS

ISBN: 978-84-8081-625-0

Servicio de Publicaciones de la Universidad de Navarra, Pamplona (España)

Pag

INTRODUCCIÓN

Este documento es una recopilación de los materiales que he preparado en los últimos años para la asignatura "Econometría básica" de la Universidad de Navarra. Los alumnos de esta asignatura han estudiado antes Cálculo, Álgebra, Probabilidad y Estadística. El manual en el que nos apoyamos es "Introducción a la Econometría" de James Stock y Mark Watson, un libro excelente y que se encuentra disponible tanto en inglés como en castellano. Este curso básico abarca los 8 o 9 primeros temas. Los autores facilitan esquemas en inglés para su presentación: yo los resumí, adapté y traduje al castellano.

Además, consciente de la importancia de la evaluación para canalizar e incentivar el esfuerzo de los alumnos, he diseñado unos ejemplos de exámenes parciales y finales para su estudio. La primera pregunta siempre es teórica: pido hacer alguna demostración o derivación matemática de entre un listado que saben que tendrán disponible durante el examen. Junto al listado, los alumnos también disponen de un formulario pues lo importante no es conocer fórmulas sino entender su correcta aplicación. La segunda y tercera suelen ser ejercicios numéricos elaborados en base a estudios o datos de otros libros de texto, sobretodo "Introducción a la Econometría. Un enfoque moderno" de Jeffrey M. Wooldridge. Algunos de estos enunciados presentan resultados obtenidos con el software libre GRETL que usamos en clase.

En muchos casos hay dos versiones de cada ejemplo de examen (a y b). La razón es que, si se cree entender las respuestas a la versión (a), se tendría que poder resolver la (b) sin problemas. Los alumnos que se ejerciten con los ejemplos de exámenes parciales deberían trabajar con la Lista de demostraciones y la Formulario de las páginas 5 y 6 al lado (pues podrán disponer de esa hoja el día del examen). No ocurre así con el Apéndice de demostraciones y su Tabla resumen, que es un material de apoyo para mi docencia. Si los estudiantes practican con ejemplos de exámenes finales, deberían tener la Lista de demostraciones y Formulario de las páginas 28 y 29 al lado. Estos listados de resultados ya demostrados son a los que se refiere la pregunta 1 de exámenes parciales y finales. Al final, aporto una guía con un ejemplo de cómo realizar un proyecto de investigación (opcional) que ayudó a muchos alumnos a comprender mejor la materia.

INTRODUCTION

This document collects the materials that I prepared for the course Econometrics I at the University of Navarra in the last years. Students of this course have already followed courses in Calculus, Algebra, Probability and Statistics. The reference textbook for the course is Introduction to Econometrics by James Stock and Mark Watson, an excellent book that is available both in English and Spanish. The course covers the first 8 or 9 chapters of the book. The authors provide slide presentations in English: I summarized and adjusted them, as well as translated them to Spanish.

Aware of the importance of the evaluation system in channeling and incentivizing students' efforts, I designed mock midterm and final exams. The first question always concerns theory: I ask for writing proofs or mathematical derivations selected from a list that students know that they have at the time of the exam. With the list, I provide students with a table of formulas during the test emphasizing that they do not need to know them by heart, but applying them correctly. The second and third questions consist on exercises based on studies and data from other textbooks, mainly "Introduction to Econometrics. A modern approach" by Jeffrey M. Wooldridge. Some of the exercises show results obtained with the free software GRETL that we use in class.

In most cases, I provide two versions of each example of exam (a & b). The reason is that, if the student understood well the solution to version (a), she must solve version (b) without trouble. Students practicing with examples of midterm exams should work with the List of Proofs and Formulae in pages 53 & 54 next to them (as they have it available the day of the test). This does not apply to the Appendix of Proofs and its Summary Table which is simply a supportive teaching material for the course. If students practice for the final exam, they should work with the List of Proofs and Formulae in pages 76 & 77. Both list of proofs are the ones referred to in question 1 of the tests. Finally, I provide a guide and an example for students that chose to do an (optional) research project which, in most cases, they found very useful for understanding the material of the course.

EN CASTELLANO

Lista de RESULTADOS BÁSICOS demostrados en el APÉNDICE para EXAMEN PARCIAL

(1) $E(aX + c) = a\mu_X + c$

(2) $var(X) = E[X^2] - \mu_X^2$

(3) $var(aX + c) = a^2\sigma_X^2$

(4) $E(aX + bY + c) = a\mu_X + b\mu_Y + c$

(5) $cov(X, Y) = E(XY) - \mu_X\mu_Y$

(6) $cov(X, X) = \sigma_X^2$

(7) $cov(aX, c + bY) = ab\ \sigma_{XY}$

(8) $cov(aX + bY + c, Z) = a\ \sigma_{XZ} + b\ \sigma_{YZ}$

(9) $var(aX + bY + c) = a^2\ var(X) + b^2\ var(Y) + 2ab \cdot cov(X, Y)$

$\equiv a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\ \sigma_{XY}$

(10) $E(aX + bY + c|Z) = aE(X|Z) + bE(Y|Z) + c$

(11) $Var(aX + bY + c|Z) = a^2\ var(X|Z) + b^2\ var(Y|Z) + 2ab \cdot cov(X, Y|Z)$

(12) $E[E(X|Y)] = E(X) \equiv \mu_X$ Ley de Esperanzas Iteradas (LIE)

(13) $P(X = x_i \cap Y = y_j) = P(X = x_i) \cdot P(Y = y_j)\ \forall\ x_i, y_j \leftrightarrow X$ e Y independientes

(14) $X\ e\ Y\ \ independientes\ \ \rightarrow\ \ corr(X, Y) = 0$

Lista de DEMOSTRACIONES hechas en CLASE y EJERCICIOS antes del PARCIAL

(I) $E(\bar{Y}) = \mu_Y$ para una muestra $\{Y_1, Y_2, \dots, Y_n\}$ i.i.d.

(II) $S_X^2 = \left(\frac{n}{n-1}\right)\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2\right)$ es la varianza de la muestra $\{X_1, X_2, \dots, X_n\}$

(III) $S_{XY} = \left(\frac{n}{n-1}\right)\left(\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\,\bar{y}\right)$ para la muestra $\{(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)\}$

(IV) resuelve el problema de MCO para estimar $Y_i = \beta_0 + \beta_1 X_i + u_i$ si $X_i = 0\ \forall\ i$ y demuestra que $\hat{\beta}_0 = \bar{y}$

(V) resuelve el problema de MCO para estimar $Y_i = \beta_0 + \beta_1 X_i + u_i$ y demuestra que $\sum_{i=1}^n \hat{u}_i = 0$, y $\sum_{i=1}^n \hat{u}_i x_i = 0$, así como que $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$, y $\widehat{\beta_1} = S_{XY}/S_X^2$

(VI) $S_{\hat{u}X} = 0$ cuando estimas $Y_i = \beta_0 + \beta_1 X_i + u_i$ usando MCO

(VII) si $E(u|X) = constante$, entonces $corr(u, X) = 0$

(VIII) si $E(u|X) = 0$, entonces $\beta_1 = \frac{\Delta E(Y|X)}{\Delta X}$ en el modelo $Y_i = \beta_0 + \beta_1 X_i + u_i$ …

(IX) …pero $\beta_1 = E(Y|D = 1) - E(Y|D = 0)$ en $Y_i = \beta_0 + \beta_1 D_i + u_i$ si D es binaria

(X) $SCT = SCE + SCR$ cuando estimas $Y_i = \beta_0 + \beta_1 X_i + u_i$ usando MCO

Momentos / Parámetros Poblacionales	**Estimación muestral** de Momentos / Parámetros
$E(X) = \sum_{j=1}^{m} x_j \, P(X = x_j) \equiv \mu_X$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \hat{\mu}_X$
$var(X) = E[(X - \mu_X)^2] \equiv \sigma_X^2$	$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2 \equiv \hat{\sigma}_X^2$
$sd(X) = \sqrt{\sigma_X^2} \equiv \sigma_X$	$s_X = \sqrt{s_X^2} \equiv \hat{\sigma}_X$
$cov(X,Y) =$ $= E[(X - \mu_X)(Y - \mu_Y)] \equiv \sigma_{XY}$	$s_{XY} =$ $= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \equiv \hat{\sigma}_{XY}$
$corr(X,Y) = {}^{\sigma_{XY}}/_{\sigma_X \sigma_Y} \equiv \rho_{XY}$	$r_{XY} = {}^{s_{XY}}/_{s_X s_Y} \equiv \hat{\rho}_{XY}$
$Y = \beta_0 + \beta_1 X_1 + u \quad \rightarrow \hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 \; ; \; \hat{u} = Y - \hat{Y}$ $\{ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u \quad \rightarrow \hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_k} X_k \; ; \; \hat{u} = Y - \hat{Y} \}$	
$\beta_1 = \frac{\Delta E(Y \mid X_1)}{\Delta X_1} \left\{ = \frac{\Delta E(Y \mid X_1, X_2)}{\Delta X_1} \right\}$	$\widehat{\beta_1} = \frac{s_{YX}}{s_X^2} \left\{ = \frac{s_{Y1} s_{22} - s_{12} s_{Y2}}{s_{11} s_{22} - s_{12}^2} \right\}$
$\beta_0 = E(Y \mid X = 0)$ $\{ = E(Y \mid X_1 = 0, X_2 = 0) \}$	$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$ $\{ = \overline{Y} - \widehat{\beta_1}\overline{X}_1 - \widehat{\beta_2}\overline{X}_2 \}$
||||||||||||||||||||||||||||||||||	||||||||||||||||||||||||||||||||||
$E(\overline{X}) = \mu_X \equiv \mu_{\overline{X}}$	
$var(\overline{X}) = \frac{1}{n}\sigma_X^2 \equiv \sigma_{\overline{X}}^2$	$\hat{\sigma}_{\overline{X}}^2 = \frac{1}{n}\hat{\sigma}_X^2 \equiv s_{\overline{X}}^2$
$sd(\overline{X}) = \frac{1}{\sqrt{n}}\sigma_X \equiv \sigma_{\overline{X}}$	$\hat{\sigma}_{\overline{X}} = \frac{1}{\sqrt{n}}\hat{\sigma}_X \equiv SE(\overline{X})$
$E(\widehat{\beta_1}) \cong \beta_1 + \rho_{Xu} \, {}^{\sigma_u}/_{\sigma_X}$	
$var(\widehat{\beta_1}) = \frac{1}{n}\frac{var([X - \mu_X]u)}{[var(X)]^2} \equiv \sigma_{\widehat{\beta_1}}^2$	$\hat{\sigma}_{\widehat{\beta_1}}^2 = \frac{1}{n}\frac{\sum_{1=1}^{n} \hat{u}_i^{\,2}(x_i - \overline{x})^2/(n-2)}{[\sum_{1=1}^{n}(x_i - \overline{x})^2/n]^2}$
$sd(\widehat{\beta_1}) = \sqrt{\sigma_{\widehat{\beta_1}}^2} \equiv \sigma_{\widehat{\beta_1}}$	$\hat{\sigma}_{\widehat{\beta_1}} = \sqrt{\hat{\sigma}_{\widehat{\beta_1}}^2} \equiv SE(\widehat{\beta_1})$
|||||||| $R^2 = \frac{SCE}{SCT} = \frac{\sum_{1=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{\sum_{1=1}^{n}(y_i - \bar{y})^2}$;;	$SER = \sqrt{\sum_{i=1}^{n} \hat{u}_i^2/(n-2)}$ ||||||||

-2.575, -1.96 y -1.64 dejan 0.5, 2.5 y 5% de probabilidad en la cola izquierda de la N(0,1)

APÉNDICE: demostraciones fáciles+tabla resumen

Este documento demuestra varias propiedades de las variables aleatorias que usamos con frecuencia en la asignatura. Entender esta materia es importante. La mayor parte de las demostraciones se sacaron del libro de texto de Stock y Watson. Para un resumen de los conceptos y de la anotación, ver la tabla en la última página.

Sea X una variable aleatoria (discreta) que toma valores{ $X=x_1$, $X=x_2$, …, $X=x_k$} con probabilidades { $P(X=x_1)$, $P(X=x_2)$, … , $P(X=x_k)$ } tales que, por definición, { $P(X=x_1)+P(X=x_2)+\ldots+P(X=x_k)=1$ }, entonces

(A) $E(X) = \sum_{i=1}^{k} x_i P(X = x_i) \equiv \mu_X$ Media o Esperanza

(B) $var(X) = E[(X-\mu_X)^2] \underset{A}{=} \sum_{i=1}^{k}(x_i-\mu_X)^2 P(X=x_i) \equiv \sigma_X^2$ Varianza

(C) $sd(X) = \sqrt{\sigma_X^2} \equiv \sigma_X$ Desviación Típica

Sean a, b y c constantes (b se usará luego), entonces

(1) $E(aX+c)$

$$\underset{A}{=} \sum_{i=1}^{k}(ax_i+c)P(X=x_i)$$
$$= \sum_{i=1}^{k} ax_i P(X=x_i) + \sum_{i=1}^{k} cP(X=x_i)$$
$$= a\sum_{i=1}^{k} x_i P(X=x_i) + c\sum_{i=1}^{k} P(X=x_i)$$
$$\underset{A}{=} aE(X)+c \equiv a\mu_X + c$$

(2) $var(X)$

$$\underset{B}{=} E[(X-\mu_X)^2]$$
$$= E[X^2+(\mu_X)^2 - 2\mu_X X]$$
$$\underset{1}{=} E[X^2] + \mu_X^2 - 2\mu_X E[X]$$
$$= E[X^2] - \mu_X^2$$

(3) $var(aX + c)$ $\overset{=}{B1}$ $E\{[(aX + c) - (a\mu_X + c)]^2\}$

$= E\{[a(X - \mu_X)]^2\}$

$\overset{=}{1}\ a^2\ E[(X - \mu_X)^2]$

$\overset{=}{B}\ a^2\ var(X) \equiv a^2\sigma_X^2$

Sea Y otra variable aleatoria (discreta) que toma valores { $Y=y_1$, $Y=y_2$, …, $Y=y_m$}. Existe una distribución conjunta de X e Y con probabilidades

{ $P(x_1, y_1)$, $P(x_1, y_2)$, …, $P(x_1, y_j)$,…, $P(x_1, y_m)$;

$P(x_2, y_1)$, $P(x_2, y_2)$, …, $P(x_2, y_j)$,…, $P(x_2, y_m)$;

… ,

$P(x_i, y_1)$, $P(x_i, y_2)$, …, $P(x_i, y_j)$,…, $P(x_i, y_m)$; … ,

$P(x_k, y_1)$, $P(x_k, y_2)$, …, $P(x_k, y_j)$,…, $P(x_k, y_m)$ }

donde, en general, $P(x_i, y_j)$ es anotación corta para $P(X = x_i \cap Y = y_j)$, y la suma de probabilidades para todas las combinaciones de X e Y es 1.

Además, las distribuciones de probabilidad individuales para X e Y son,

(D) $P(X = x_i) = \sum_{j=1}^{m} P(X = x_i \cap Y = y_j)$

y $P(Y = y_j) = \sum_{i=1}^{k} P(X = x_i \cap Y = y_j)$

es decir, son las probabilidades marginales, de modo que definimos:

(A') $E(Y) = \sum_{j=1}^{m} y_j P(Y = y_j) \equiv \mu_Y$ ¡como para X !

(B') $var(Y) = E[(Y - \mu_Y)^2] = E[Y^2] - \mu_Y^2 \equiv \sigma_Y^2$ ¡como para X !

(C') $st(Y) = \sqrt{\sigma_Y^2} \equiv \sigma_Y$ ¡como para X !

(4) $E(aX+bY+c) \underset{A}{=} \sum_{i=1}^{k}\sum_{j=1}^{m}(ax_i+by_j+c)P(X=x_i \cap Y=y_j)$

$= \sum_i \sum_j ax_i P(x_i,y_j) + \sum_i \sum_j by_j P(x_i,y_j) + \sum_i \sum_j cP(x_i,y_j)$

$= a\sum_{i=1}^{k} x_i \sum_{j=1}^{m} P(x_i,y_j) + b\sum_{j=1}^{m} y_j \sum_{i=1}^{k} P(x_i,y_j) + c\cdot 1$

$\underset{D}{=} a\sum_{i=1}^{k} x_i P(X=x_i) + b\sum_{j=1}^{m} y_j P(Y=y_j) + c$

$\underset{A}{=} aE[X] + bE[Y] + c \equiv a\mu_X + b\mu_Y + c$

y, en general, si en lugar de dos variables X e Y tuviéramos variables $\{X_1, X_2, \ldots, X_s, \ldots, X_S\}$ y constantes $\{a_1, a_2, \ldots, a_s, \ldots, a_S, c\}$

(4) $E(c+\sum_{s=1}^{S} a_s X_s) = c + \sum_{s=1}^{S} a_s E(X_s) \equiv c + \sum_{s=1}^{S} a_s \mu_{Xs}$

Definamos

(E) $cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] \equiv \sigma_{XY}$

entonces

(5) $cov(X,Y) \quad \underset{E}{=} E[(X-\mu_X)(Y-\mu_Y)] =$

$= E(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y)$

$\underset{4}{=} E(XY) - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y$

$= E(XY) - \mu_X\mu_Y$

(6) $cov(X,X) \quad \underset{E}{=} E[(X-\mu_X)(X-\mu_X)]$

$\underset{B}{=} var(X) \equiv \sigma_X^2$

(7) $cov(aX, c+bY) \underset{E}{=}\underset{4}{=} E[(aX-a\mu_X)(c+bY-c-b\mu_Y)]$ ¡o usa (5)!

$= E[ab(X-\mu_X)(Y-\mu_Y)]$

$\underset{4}{=} ab\,E[(X-\mu_X)(Y-\mu_Y)]$

$\underset{E}{=} ab\,cov(X,Y) \equiv ab\,\sigma_{XY}$

(8) $cov(aX+bY+c,Z) \overset{=}{7} cov(aX+bY,Z)$

$\overset{=}{E}\ \overset{=}{4}\ E[\{(aX+bY)-(a\mu_X+b\mu_Y)\}\{Z-\mu_Z\}]$ ¡o usa (5)!

$= E[\{a(X-\mu_X)+b(Y-\mu_Y)\}\{Z-\mu_Z\}]$

$\overset{=}{4}\ aE[(X-\mu_X)(Z-\mu_Z)]+bE[(Y-\mu_Y)(Z-\mu_Z)]$

$\overset{=}{E}\ a\ cov(X,Z)+b\ cov(Y,Z) \equiv a\ \sigma_{XZ}+b\ \sigma_{YZ}$

y, en general, si tenemos variables aleatorias $\{X_1, X_2, \ldots, X_s, \ldots, X_S\}$ y $\{Y_1, Y_2, \ldots, Y_t, \ldots, Y_T\}$ y constantes $\{a_1, a_2, \ldots, a_s, \ldots, a_S ; b_1, b_2, \ldots, b_t, \ldots, b_T ; c\}$

(8) $cov(\sum_s a_s X_s , c+\sum_t b_t Y_t) = \sum_s \sum_t a_s b_t cov(X_s, Y_t) \equiv \sum_s \sum_t a_s b_t \sigma_{XsYt}$

(9) $var(aX+bY+c) \overset{=}{3} var(aX+bY)$

$\overset{=}{B}\overset{=}{4}\ \mathrm{E}[\{(aX+bY)-(a\mu_X+b\mu_Y)\}^2]$ ¡o usa (2)!

$= \mathrm{E}[\{a(X-\mu_X)+b(Y-\mu_Y)\}^2]$

$= \mathrm{E}[a^2(X-\mu_X)^2+b^2(Y-\mu_Y)^2+2ab(X-\mu_X)(Y-\mu_Y)]$

$\overset{=}{4}\ a^2\mathrm{E}[(X-\mu_X)^2]+b^2E[(Y-\mu_Y)^2]+2abE[(X-\mu_X)(Y-\mu_Y)]$

$\overset{=}{B}\overset{=}{E}\ a^2var(X)+b^2var(Y)+2ab\cdot cov(X,Y)$

$\equiv a^2\sigma_X^2+b^2\sigma_Y^2+2ab\ \sigma_{XY}$

y, en general, si tenemos variables aleatorias $\{X_1, X_2, \ldots, X_s, \ldots, X_S\}$ ay constantes $\{a_1, a_2, \ldots, a_s, \ldots, a_S ; c\}$

(9) $var(\sum_s a_s X_s + c) = \sum_s a_s^2\ var(X_s) + 2\sum_s \sum_{t\neq s} a_s\ a_t\ cov(X_s, X_t)$

$\equiv \sum_s a_s^2\ \sigma_{Xs}^2 + 2\sum_s \sum_{t\neq s} a_s\ a_t\ \sigma_{XsXt}$

El <u>Teorema de Bayes</u> dice que la probabilidad condicional es

(F) $P(X=x_i|Y=y_j) = P(X=x_i \cap Y=y_j) \Big/ P(Y=y_j)$

luego

(G) $E(X|Y=y_j) = \sum_{i=1}^{k} x_i P(X=x_i|Y=y_j)$ ¡ o simplemente $E(X|Y)$!

(H) $Var(X|Y=y_j) = E\{(X-\mu_X)^2|Y=y_j\}$ ¡ o simplemente $Var(X|Y)$!

$\overset{=}{G} \sum_{i=1}^{k}(x_i-\mu_X)^2\, P(X=x_i|Y=y_j)$

Por tanto, si consideramos una variable aleatoria extra Z,

(10) $E(aX+bY+c|Z) \overset{=}{4}\overset{=}{G}\ aE(X|Z)+bE(Y|Z)+c$

(11) $Var(aX+bY+c|Z) \overset{=}{9}\overset{=}{H}$

$a^2\ var(X|Z)+b^2\ var(Y|Z)+2ab\cdot cov(X,Y|Z)$

y, en general,

(10) $E(c+\sum_{s=1}^{S} a_s X_s\,|Z) \overset{=}{4}\overset{=}{G}\ c+\sum_{s=1}^{S} a_s E(X_s|Z)$

(11) $Var(\sum_s a_s X_s + c|Z) \overset{=}{9}\overset{=}{H}$

$\sum_s a_s^2\ var(X_s|Z)+2\sum_s\sum_{t\neq s} a_s\ a_t\ \ cov(X_s,X_t|Z)$

donde los detalles son omitidos porque las demostraciones son casi (4) y (9)

(12) $E[E(X|Y)] = E\big[E(X|Y=y_j)\big] \overset{=}{A} \sum_{j=1}^{m} E(X|Y=y_j)P(Y=y_j)$

$\overset{=}{G} \sum_{j=1}^{m}\big[\sum_{i=1}^{k} x_i P(X=x_i|Y=y_j)\big]P(Y=y_j)$

$\overset{=}{F} \sum_{j=1}^{m}\sum_{i=1}^{k} x_i\, P(X=x_i\cap Y=y_j)$

$= \sum_{i=1}^{k}\sum_{j=1}^{m} x_i P(X=x_i\cap Y=y_j)$

$= \sum_{i=1}^{k}\big[x_i \sum_{j=1}^{m} P(X=x_i\cap Y=y_j)\big]$

$\overset{=}{D} \sum_{i=1}^{k} x_i\, P(X=x_i)$

$\overset{=}{A} E(X) \equiv \mu_X$ Ley de Esperanzas Iteradas (LIE)

Las variables aleatorias X e Y son independientes $\leftrightarrow$ ("si y solo si")

(I) $P(X=x_i|Y=y_j) = P(X=x_i)\ \ \forall\, x_i, y_j$

Luego

(13) $P(X=x_i\cap Y=y_j) \overset{=}{F}\overset{=}{I} P(X=x_i)\cdot P(Y=y_j)\ \forall\, x_i, y_j$

Por último, recuerde la definición

(J) $corr(X,Y) = {}^{\sigma_{XY}}/_{\sigma_X \sigma_Y} \equiv \rho_{XY}$

luego

(14) *Si X e Y son independientes , entonces* $corr(X,Y) = 0$

X e Y son variables aleatorias (no constantes) así que $\sigma_X > 0$ *y* $\sigma_Y > 0$.

Luego $corr(X,Y) = 0 \leftrightarrow \sigma_{XY} = 0$ o, dado (5), $E(XY) = E(X)E(Y)$.

$$E(XY) \underset{A}{=} \sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j P\left(X = x_i \cap Y = y_j\right)$$

$$\underset{I}{=} \sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j P(X = x_i)\, P\left(Y = y_j\right)$$

$$= \sum_{i=1}^{k} x_i P(X = x_i) \sum_{j=1}^{m} y_j P\left(Y = y_j\right)$$

$$\underset{A}{=} E(X)E(Y)$$

Usted debe entender la distinción entre los momentos de una variable aleatoria y sus estimaciones (con sombrero^)	
Momentos de una variable aleatoria X	Estimaciones muestrales de los momentos
$E(X) = \sum_{i=1}^{k} x_i\, P(X = x_i) \equiv \mu_X$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \hat{\mu}_X$
$var(X) = E[(X - \mu_X)^2] \equiv \sigma_X^2$	$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \equiv \hat{\sigma}_X^2$
$sd(X) = \sqrt{\sigma_X^2} \equiv \sigma_X$	$s_X = \sqrt{s_X^2} \equiv \hat{\sigma}_X$
$cov(X,Y) =$ $= E[(X - \mu_X)(Y - \mu_Y)] \equiv \sigma_{XY}$	$s_{XY} =$ $= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \equiv \hat{\sigma}_{XY}$
$corr(X,Y) = {}^{\sigma_{XY}}/_{\sigma_X \sigma_Y} \equiv \rho_{XY}$	$r_{XY} = {}^{s_{XY}}/_{s_X s_Y} \equiv \hat{\rho}_{XY}$
Además, los estadísticos de la derecha son funciones de los datos, o sea variables aleatorias cuyas concreciones dependen de la muestra. Por ejemplo, X es la altura de los europeos y queremos estimar μ_X, la *media* o *valor esperado* de la altura de los europeos, E(X). Juan elige aleatoriamente una muestra representativa de 500 españoles y obtiene $\overline{x} = 1.67$: esta es su estimación para μ_X ($\hat{\mu}_X$). Giampiero saca otra muestra de 500 italianos y obtiene $\overline{x} = 1.65$. Catiana elige otra muestra representativa de 500 alemanes que da $\overline{x} = 1.70$. Luego {1.65, 1.67, 1.70} son concreciones de la variable aleatoria $\overline{X}$, la altura media de una muestra representativa de 500 europeos. Esta variable aleatoria $\overline{X}$ tiene también momentos: media, varianza, etc, que podemos estimar con datos.	
$E(\overline{X}) = \mu_X \equiv \mu_{\overline{X}}$	$\hat{\mu}_{\overline{X}} = \overline{x}$
$var(\overline{X}) = \frac{1}{n}\sigma_X^2 \equiv \sigma_{\overline{X}}^2$	$\hat{\sigma}_{\overline{X}}^2 = \frac{1}{n}\hat{\sigma}_X^2 \equiv s_{\overline{X}}^2$
$sd(\overline{X}) = \frac{1}{\sqrt{n}}\sigma_X \equiv \sigma_{\overline{X}}$	$\hat{\sigma}_{\overline{X}} = \frac{1}{\sqrt{n}}\hat{\sigma}_X \equiv SE(\overline{X})$

Ejemplo 1 de EXAMEN PARCIAL

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) Se pedirá ***demostrar*** *UNO de los enunciados de la última página:*

o bien el resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos,*

o bien, la derivación *YYY* hecha en clase, ejercicios…

2. (3.5 puntos) *El gobierno quiere averiguar si impartir asignaturas en inglés como Artes plásticas, Música o Ciencias en la enseñanza primaria tiene un impacto en la formación de los alumnos sobre estas materias.*

a. Diseña un experimento (RCT) que dé respuesta a la pregunta del gobierno. Explícalo brevemente y señala explícitamente qué rasgo fundamental nos daría certeza de estar midiendo el efecto de impartir las asignaturas en inglés y no el de otros factores.

En lugar de hacer un experimento (RCT), seleccionamos aleatoriamente notas observadas para 500 estudiantes de 3º de primaria en un test sobre estas materias en junio de 2017 (que incluye tanto estudiantes en inglés como en castellano). Hallamos una puntuación media de 81 (sobre 100) y una desviación típica de 15.2 puntos.

b. Construye un intervalo de confianza del 95% para la puntuación media poblacional. ¿Podemos decir con al menos esta misma confianza que los alumnos, en promedio, aprueban el test ($\mu \geq 50 puntos$)?

En la muestra, 208 estudiantes hicieron las asignaturas en inglés, mientras que 292, en castellano. La media para los primeros es 79.8 puntos (con una desviación típica de 14.5 puntos) y 81.9 puntos para los segundos (con 16.7 puntos de desviación típica).

c. ¿Podemos afirmar al nivel de significación del 5% que los alumnos en inglés aprenden tanto como los de castellano ($\Delta=0$)? Ilustra con estos dos gráficos tu respuesta.

$\hat{\Delta} \sim N(?,?)$ $\quad$ $t \sim N(0,1)$

d. Dibuja en estas mismas gráficas cuál es el mínimo nivel de significación necesario para rechazar $\Delta=0$ (o repite lo necesario del dibujo para hacerlo clara pero cómodamente).

3. (4.5 puntos) *Queremos estudiar el efecto de la formación académica (ED, medida como años o cursos académicos superados) en los ingresos por hora (AHE, average hourly earnings, en dolares). Para ello, obtenemos datos observacionales de una muestra aleatoria de 2829 trabajadores a tiempo completo de 29-30 años y estimamos:*

$$(\mathbf{1}) \quad \widehat{AHE} = -7.29 + 1.93\,ED, \qquad R^2 = 0.16, \qquad SER = 10.29$$

(1.1) (0.08) *errores típicos robustos a la heterocedasticidad*

a. Interpreta -7.29 y 1.93 en términos de años de formación y dólares ganados

b. Da el intervalo de confianza del 95% para el cambio esperado en *AHE* si ΔED=2 años

c. ¿Cuáles serían los ingresos estimados para un trabajador con 30 años de educación? ¿Crees que esta predicción es fiable? Justifica usando argumentos estadísticos

En general, hacer una carrera universitaria implica alcanzar los 16 años de formación. Construimos una variable ficticia igual a 1 si ED≥16 años (y 0 en otro caso) y estimamos:

(2) MCO, usando las observaciones 1-2829
Variable dependiente: **ahe**
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
const	16.5009	0.209406	78.7987	<0.0001	***
dummy (binaria)	8.57931	0.435689	19.6913	<0.0001	***

Media de la vble. dep.	19.83983	D.T. de la vble. dep.	11.23823
Suma de cuad. residuos	307670.5	D.T. de la regresión	10.43230
R-cuadrado	0.138589	R-cuadrado corregido	0.138284
F(1, 2827)	387.7491	Valor p (de F)	5.38e-81

d. ¿Cuál es la diferencia media en salario de aquellos que cruzaron el umbral de los 16 años (con los que no)? ¿Es significativamente distinta de cero a un nivel del 10, 5 y 1%?

e. ¿Cambiaría el SER si no marcamos *errores standard robustos a la heterocedasticidad*?

Con frecuencia los que reciben una formación larga conocen lenguas extranjeras *(aunque sea más fruto de viajes, motivación… que de las clases del instituto, universidad, etc)*

f. Explica qué implica omitir la variable *conocimiento de lenguas extranjeras* en nuestra regresión (1). Usa una fórmula matemática para justificar tu respuesta nítidamente.

g. Haz dos gráficos: (i) uno que muestre qué supuesto de nuestro modelo falla y (ii) otro con sus consecuencias al usar una muestra para estimar el efecto causal de *ED* en *AHE*.

Ejemplo 2a de EXAMEN PARCIAL

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) Se pedirá ***demostrar*** *UNO de los enunciados de la última página:*

o bien el resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos,*

o bien, la derivación *YYY* hecha en clase, ejercicios…

2. (3.5 puntos) *El gobierno de Bélgica quiere investigar el efecto en la formación de alumnos de Primaria de un plan de uso intensivo de nuevas tecnologías en el aula. El aprendizaje se mide con un examen estandarizado.*

a. ¿Qué rasgo clave en un experimento (RCT) aseguraría que se mide ese efecto y no otro?

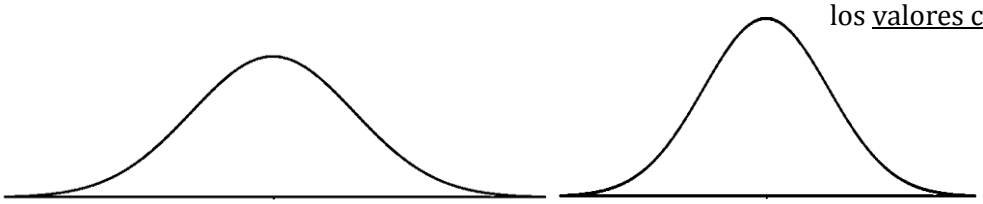
En lugar de datos experimentales, inspeccionamos las notas en un test de una muestra de 500 estudiantes que ya habían implementado (o no) un plan similar. Hallamos una puntuación media de 715 (sobre 800) y una desviación típica de 225 puntos.

b. Construye un intervalo de confianza del 99% para la puntuación media poblacional

c. ¿Rechazamos al nivel de significación del 1% que la puntuación media poblacional es 685 puntos? Si fuera el caso, ¿por qué sería obvio que rechazaríamos también al 5%?

En la muestra, 200 alumnos pertenecen a centros que siguieron este tipo de plan, mientras que 300 no. La media para los primeros es 721 puntos (con una desviación típica de 230 puntos) y 711 puntos para los segundos (con 222 puntos de desviación típica).

d. ¿Rechazamos al nivel de significación del 5% que los alumnos que siguen el plan aprenden lo mismo que los que no? Ilustra abajo tu respuesta, rellena los 3 "?" y localiza los valores críticos.



$? \sim N(?, ?)$ $t \sim N(0,1)$

e. Da un modelo lineal de un regresor y una hipótesis que, contrastada, responda a (**d**)

3. (4.5 puntos) *Queremos estudiar el efecto de la* altura *(medida en pulgadas) en la variable "flujo"* renta *(en dólares por año). Para ello, obtenemos datos observacionales de una muestra aleatoria de 7896 hombres trabajadores y estimamos:*

$$(\mathbf{1}) \quad \widehat{renta} = -43130 + 1306.9\ altura, \qquad R^2 = 0.02, \qquad SER = 26671$$

(6925) (98.86) *errores típicos robustos a la heterocedasticidad*

a. Interpreta -43130 (*el valor, no sólo el signo*) en términos de pulgadas y dólares ganados

b. Reescribe el modelo estimado cambiando la variable *altura* a cm (1cm=2.54pulgadas)

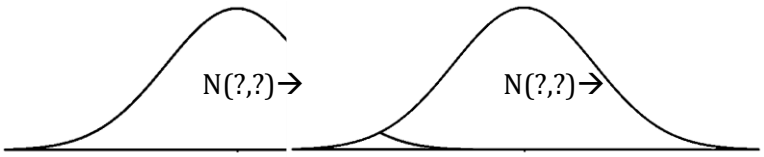
Reestimamos el modelo usando una muestra de 9974 mujeres con altura *en pulgadas:*

(2) MCO, usando las observaciones 1-9974
Variable dependiente: **renta**
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
const	12650.9	6299.15	2.0083	0.0446	**
altura	511.222	97.5846	5.2388	<0.0001	***

Media de la vble. dep.	45621.00	D.T. de la vble. dep.	26835.43
Suma de cuad. residuos	7.16e+12	D.T. de la regresión	26800.90
R-cuadrado	0.002672	R-cuadrado corregido	0.002572
F(1, 9972)	27.44459	Valor p (de F)	1.65e-07

c. Construye un intervalo de confianza del 95% para la diferencia en el efecto de la *altura* en la *renta* entre hombres y mujeres



Ilustra tu respuesta en esta gráfica como hicimos en clase. Rellena los 6 "?"

d. ¿Cambiaría este intervalo si no hubiésemos pedido *errores standard robustos a la heterocedasticidad* al estimar ambas regresiones y hubiera heterocedasticidad? Razona

Muchas rentas altas provienen de rendimientos financieros de herencias, es decir, de la riqueza heredada *que, por su relación con la nutrición en la infancia, se asocia con* altura

e. Explica qué implica omitir la variable *riqueza heredada* en la regresión (1). Usa una fórmula matemática para justificar tu respuesta nítidamente

f. Ilustra en dos gráficos: (i) qué supuesto de nuestro modelo falla y (ii) las consecuencias al usar una muestra para estimar el efecto causal de *altura* en *renta*.

Ejemplo 2b de EXAMEN PARCIAL

(si entendiste la solución al Ejemplo 2a, sabrías responder este ejemplo similar)

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) Se pedirá ***demostrar*** *UNO de los enunciados de la última página:*

o bien el resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos,*

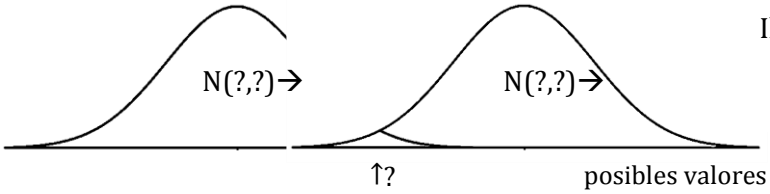
o bien, la derivación *YYY* hecha en clase, ejercicios…

2. (3.5 puntos) *El gobierno de Reino Unido quiere investigar el efecto en la formación de alumnos de Primaria de recibir una educación que separe chicos y chicas o una educación con aula mixta. Su aprendizaje se mide con un examen estandarizado*

a. ¿Qué rasgo clave en un experimento (RCT) aseguraría que se mide ese efecto y no otro?

En lugar de datos experimentales contamos con una muestra de notas observadas en un test realizado por 400 estudiantes tanto de la enseñanza "mixta" como de la "diferenciada". Hallamos una puntuación media de 720 (sobre 800) y una desviación típica de 225 puntos.

b. Construye un intervalo de confianza del 95% para la puntuación media poblacional



Ilustra tu respuesta en este gráfico como hicimos en clase. Rellena los 6 "?"

c. ¿Rechazamos al nivel de significación del 5% que la puntuación media poblacional es 695 puntos? Si fuera el caso, ¿por qué sería obvio que rechazásemos también al 10%?

En la muestra, 150 alumnos de enseñanza diferenciada, mientras que 250 de la mixta. La media para los primeros es 723 puntos (con una desviación típica de 220 puntos) y 718.2 puntos para los segundos (con 231 puntos de desviación típica).

d. ¿Rechazamos al nivel de significación del 5% que los alumnos de enseñanza mixta tienen el mismo nivel de aprendizaje que los de "diferenciada"?

e. Da un modelo lineal de un regresor y una hipótesis que, contrastada, responda a (d)

3. (4.5 puntos) *Queremos estudiar el efecto de la* altura *(medida en pulgadas) en la* renta *(en dólares por año). Para ello, obtenemos datos observacionales de una muestra aleatoria de 7896 hombres con trabajo y estimamos el siguiente modelo (bastante malo, véase R^2):*

$$(\mathbf{1}) \quad \widehat{renta} = -43130 + 1306.9\ altura, \qquad R^2 = 0.02, \qquad SER = 26671$$

(6925) (98.86) *errores típicos robustos a la heterocedasticidad*

a. Interpreta 1306.9 (*el valor, no sólo el signo*) en términos de pulgadas y dólares ganados

b. ¿Contempla nuestro modelo casos como Amancio Ortega o Bill Gates? Razona

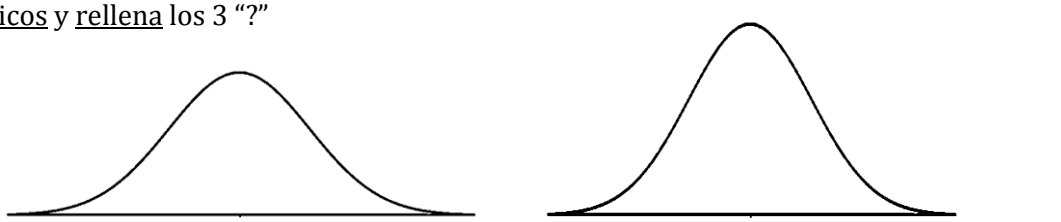
Reestimamos el modelo usando una muestra de 9974 mujeres con altura *en pulgadas:*

Modelo **(2)** MCO, usando las observaciones 1-9974
Variable dependiente: **renta**
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
const	12650.9	6299.15	2.0083	0.0446	**
altura	511.222	97.5846	5.2388	<0.0001	***

Media de la vble. dep.	45621.00	D.T. de la vble. dep.	26835.43
Suma de cuad. residuos	7.16e+12	D.T. de la regresión	26800.90
R-cuadrado	0.002672	R-cuadrado corregido	0.002572
F(1, 9972)	27.44459	Valor p (de F)	1.65e-07

c. Contrasta al nivel de significación del 5% si la diferencia del efecto de la *altura* en la *renta* entre los hombres y en las mujeres es nula. Ilústralo abajo, localiza todos los valores críticos y rellena los 3 "?"



? ~ N(?, ?) t~N(0,1)

d. ¿Podría cambiar la conclusión del contraste si no hubiésemos pedido *errores standard robustos a la heterocedasticidad* al estimar ambas regresiones y sí la hubiera? Razona

Muchas rentas bajas provienen de profesiones donde la fuerza física es importante, de modo que su contrario (la debilidad) *se asocia con rentas altas y negativamente con la* altura

e. Explica qué implica omitir la variable *debilidad física* en la regresión (1). Usa una fórmula matemática para justificar tu respuesta nítidamente

f. Ilustra en dos gráficos: (i) qué supuesto de nuestro modelo falla y (ii) las consecuencias al usar una muestra para estimar el efecto causal de *altura* en *renta*.

Ejemplo 3a de EXAMEN PARCIAL

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

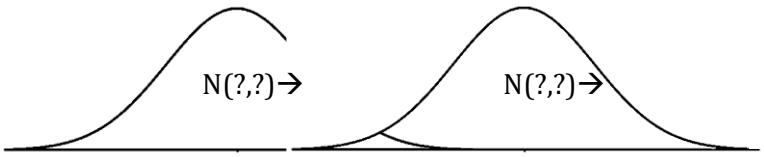
1. (2 puntos) Se pedirá ***demostrar*** *UNO de los enunciados de la última página:*

o bien el resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos,*

o bien, la derivación *YYY* hecha en clase, ejercicios…

2. (4 puntos) *En base a estimaciones anteriores, el gobierno belga supone que la renta media en Bélgica es 1764€ al mes. Para comprobar este valor, su Agencia Estadística encuesta a 500 trabajadores en la capital, Bruselas, el centro político y financiero belga, y halla una renta media muestral de 1785€ (y una desviación típica igual a 335€).*

a. Construye el intervalo de confianza del 95% para la renta media en Bélgica. ¿Rechazamos al nivel de significación del 5% la renta media belga que asume el gobierno?



Ilustra tu respuesta en este gráfico como hicimos en clase . Rellena los 6 "?"

El director de la Agencia Estadística pide ampliar la encuesta a otros 500 trabajadores de Amberes, la otra gran ciudad belga conocida por su gran puerto y su industria química. La media muestral para estos 500 trabajadores es 1545€ (y la desviación típica 295€). En total, para los 1000 trabajadores, la media es 1665€ (y la desviación típica total 321€).

b. En base a la muestra de 1000 trabajadores. ¿Rechazamos ahora que la renta media belga sea 1764€ al nivel de significación del 5%?

c. ¿Rechazamos con 5% de signif. que la renta media es igual en Bruselas y Amberes?

d. Da un modelo lineal de un regresor y una hipótesis que, contrastada, responda a (c)

e. ¿Qué factor omitido debe ser tenido en cuenta en (d) o (c) para poder pronosticar si la renta de un emigrante español cualquiera en Bélgica sería la misma en las dos ciudades?

f. ¿Qué rasgo de un experimento (RCT) ayudaría hacer el pronóstico en (e)? Explica

3. (4 puntos) *Queremos estudiar el efecto de la atención médica prenatal (*nvisit*, medida en visitas médicas de la madre antes del parto) en el peso del bebé (*pesobb*, en gramos). Recogemos una muestra de 2418 recién nacidos de madres no fumadoras y estimamos:*

(**1**) $\widehat{pesobb} = 3066.4 + 32.7\ nvisit, \qquad R^2 = 0.04, \qquad SER = 573.5$

(49.8) (4.2) (*errores standard robustos a la heterocedasticidad)*

a. Interpreta 3066.4 en términos de visitas médicas y/o peso al nacer

b. Contrasta al nivel de signif. del 1% si la variación esperada en el peso del bebé al acudir una vez más al médico la madre es nulo o no. Rellena los 3 "?" e ilustra tu respuesta en ambas figuras indicando sus valores críticos

? ~ N(?, ?) t~N(0,1)

c. ¿Cambiaría el R^2 en (1) si NO pedimos o "marcamos" *errores standard robustos a la heterocedasticidad* y hay heterocedasticidad en los datos? Explica muy brevemente

Estimamos el mismo modelo usando datos de recién nacidos de madres que fumaron durante el embarazo y obtenemos:

(2) MCO, usando las observaciones 1-582
Variable dependiente: **pesobb**
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
const	2790.63	82.9246	33.6526	<0.0001	***
nvisit	38.1389	7.36556	5.1780	<0.0001	***

Media de la vble. dep.	3178.832	D.T. de la vble. dep.	580.0068
Suma de cuad. residuos	1.80e+08	D.T. de la regresión	557.5502
R-cuadrado	0.077527	R-cuadrado corregido	0.075937
F(1, 580)	26.81170	Valor p (de F)	3.10e-07

d. Dibuja dos diagramas de dispersión para los datos correspondientes a madres que sí fumaron durante el embarazo y las que no, así como los dos modelos estimados (ten en cuenta las medidas de bondad de ajuste)

e. ¿Rechazamos al nivel de signif. del 5% que el efecto de una visita extra al médico en el peso del bebé es el mismo para los nacidos de mujeres fumadoras y NO fumadoras?

Ejemplo 3b de EXAMEN PARCIAL

(si entendiste la solución al Ejemplo 3a, sabrías responder este ejemplo similar)

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) Se pedirá ***demostrar*** *UNO de los enunciados de la última página:*

o bien el resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos,*

o bien, la derivación *YYY* hecha en clase, ejercicios…

2. (4 puntos) *En base a estimaciones anteriores, el gobierno escocés supone que la renta media en Escocia es £1115 al mes. Para comprobar este valor, su Agencia Estadística encuesta a 400 trabajadores en la capital, Edimburgo, la capital financiera y política, y halla una renta media muestral de £1180 (y una desviación típica igual a £240).*

a. Basándonos en la encuesta, ¿rechazamos al nivel de signif. del 5% que la renta media en Escocia sea igual a la asumida según las antiguas estimaciones? Ilustra situando los va--lores críticos en ambas figuras y rellena los3 '?'

El director de la Agencia Estadística pide ampliar la encuesta a otros 400 trabajadores de Glasgow, la otra gran ciudad escocesa, que es muy industrial: conocida por sus astilleros y sector manufacturero. La media muestral para estos 400 trabajadores es £1020 (y la desviación típica £198). En total, para los 800 trabajadores, la media es £1100 (y la desviación típica total £220).

b. Construye un intervalo de confianza del 95% para la renta media en Escocia en base a la muestra de 800 trabajadores. ¿Rechazamos ahora que la media sea £1115 al nivel 5%?

c. ¿Rechazamos con 5% de signif. que la renta media es igual en Edimburgo y Glasgow?

d. Da un modelo lineal de un regresor y una hipótesis que, contrastada, responda a (**c**)

e. ¿Qué factor omitido debe ser tenido en cuenta en (**c**) o (**d**) para poder pronosticar si la renta de un emigrante español cualquiera en Escocia sería la misma en las dos ciudades?

f. ¿Qué rasgo de un experimento (RCT) ayudaría a responder a (**e**)? Explica brevemente

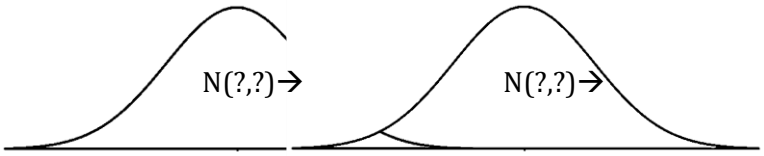
3. (4 puntos) *Queremos estudiar el efecto de la atención médica prenatal (*nvisit*, medida en visitas médicas de la madre antes del parto) en el peso del bebé (*pesobb*, en gramos). Para ello, recogemos una muestra de 3000 recién nacidos y estimamos:*

(**1**) $\widehat{pesobb} = 2979.9 + 36.7\ nvisit, \qquad R^2 = 0.05, \qquad SER = 576.7$

(43.5) (3.7) (*errores standard robustos a la heterocedasticidad)*

a. Interpreta 2979.93 y 36.66 en términos de visitas médicas y/o peso al nacer

b. ¿Cuál es el intervalo del 95% de confianza para el efecto esperado en *peso* de Δ*nvisit*=2?



↑? posibles valores de $\Delta\widehat{pesobb}$

Ilustra tu respuesta en la figura como hicimos en clase y rellena los 5 "?"

c. ¿Es este efecto significativo si usamos un nivel de signif. del 5%? Razona brevemente

d. ¿Cambiaría el intervalo de confianza si no pedimos o "marcamos" *errores standard robustos a la heterocedasticidad* y hubiera heterocedasticidad? Explica muy brevemente

e. Escribe la interpretación del R^2

En realidad, la mayoría de las madres van al médico al menos 12 veces antes de dar a luz. Construimos una variable binaria *que toma el valor 1 si* nvisit≥12 *(o si no, 0) y estimamos:*

(2) MCO, usando las observaciones 1-3000

Variable dependiente: **pesobb**

Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
const	3275.23	16.241	201.6636	<0.0001	***
dummy (binaria)	215.127	21.2711	10.1135	<0.0001	***

Media de la vble. dep.	3382.934	D.T. de la vble. dep.	592.1629
Suma de cuad. residuos	1.02e+09	D.T. de la regresión	582.4056
R-cuadrado	0.033006	R-cuadrado corregido	0.032683
F(1, 2998)	102.2838	Valor p (de F)	1.15e-23

f. ¿Cuál es el peso que predecimos para un bebé cualquiera cuya madre visitó al médico más de 12 veces según nuestra estimación **(2)**?

Ejemplo 4a de EXAMEN PARCIAL

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) Se pedirá ***demostrar*** *UNO de los enunciados de la última página:*

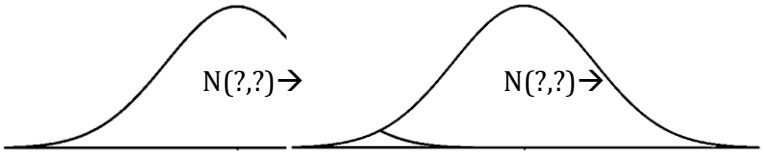
o bien el resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos,*

o bien, la derivación *YYY* hecha en clase, ejercicios…

2. (4.5 puntos) *Studentville, una ciudad de EEUU, quiere concienciar a sus universitarios sobre los problemas de colesterol causados por la comida rápida. Se sabe que el colesterol ideal para jóvenes de unos 20 años es 190mg/dL, pero una muestra de 400 estudiantes de la John Nerd University da una media de 205 mg/dL (y desviación típica de 80 mg/dL). Los alumnos de esta prestigiosa universidad, situada en Studentville, consumen comida rápida tres o más veces por semana de media.*

a. Basándonos en esta muestra, ¿en qué rango podemos decir que se encuentra el colesterol medio de un universitario de Studentville con un 95% de confianza?

b. ¿podemos afirmar que el colesterol medio de los universitarios de Studentville es significativamente (nivel de signif. del 5%) distinto al nivel ideal? Añade una curva a esta figura para ilustrar cómo llegas a tu conclusión. Rellena los 6 "?"



En Studentville, también se encuentra Health&Muscles College, una pequeña universidad apenas conocida salvo por su equipo de atletismo y su título de salud deportiva. No obstante, sus estudiantes son un 20% de todos los universitarios de Studentville, por lo que una muestra representativa no puede ignorarlos. Se recogen datos de 100 alumnos de Health&Muscles College y obtenemos un colesterol medio de 192mg/dL (y una desviación típica de 10mg/dL). Y estos 100 alumnos consumen de media comida rápida menos de 3 veces por semana. La desviación típica del colesterol en los 500 universitarios es 71.7 mg/dL

c. Basándonos ahora en la muestra completa de 500 alumnos, ¿descartamos al nivel de signif. del 5% que los universitarios de Studentville tengan el nivel de colesterol ideal?

d. ¿Rechazamos al nivel de signif. del 5% que los alumnos de *John Nerd University* tienen un nivel de colesterol medio igual al de los estudiantes de *Health&Muscles College*?

e. Da un modelo lineal de un regresor y una hipótesis que, contrastada, responda a (**d**)

f. ¿Por qué no podemos concluir en (d) que si todos universitarios de Studentville pasan a consumir menos de 3 veces/semana comida rápida alcanzarían un nivel de colesterol más cercano al ideal? Además, justifícalo para el modelo de (e) usando una fórmula

g. Señala muy claramente cuál es el rasgo esencial de un experimento (RCT) que pueda responder a "¿variar el consumo de comida rápida causa cambios en el colesterol?"

3. (3.5 puntos) *Una inmobiliaria estudia el efecto de la antigüedad o* edad *(en años) de las viviendas en su* precio *(en $) de venta. Para ello, obtiene datos observacionales de una muestra aleatoria de 1080 viviendas y estima:*

(1) MCO, usando las observaciones 1-1080
Variable dependiente: **precio**
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	184.05	6.05	30.44	<0.0001	***
edad	−1.49	0.197	−7.583	<0.0001	***

Media de la vble. dep.	154.86	D.T. de la vble. dep.	122.91
Suma de cuad. residuos	1.56e+10	D.T. de la regresión	120.26
R-cuadrado	0.243518	R-cuadrado corregido	0.242630
F(1, 1078)	57.50533	Valor p (de F)	7.25e-14

a. Interpreta el valor numérico (*¡y no solamente su signo!*) −1.49 e indica sus unidades

b. ¿Cuál sería el precio estimado de una vivienda tan nueva que acaba de construirse?

c. Calcula el intervalo del 95% de confianza para la pérdida de valor de una casa "nueva" que tarda 3 años en venderse.

Se repite el análisis distinguiendo entre viviendas de estilo tradicional y contemporáneo:

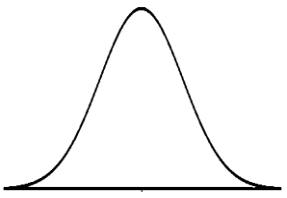
Contemporáneo: (**2**) $\widehat{Precio} = 205.19 - 1.894\, Edad,\ R^2 = 0.05$

n_C=498 (6.046) (0.197) *errores típicos robustos a la heterocedast.*

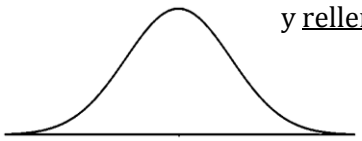
Tradicional: (**3**) $\widehat{Precio} = 164.16 - 1.056\, Edad,\ R^2 = 0.34$

n_T=582 (5.879) (0.252) *errores típicos robustos a la heterocedasticidad*

d. ¿Descartaríamos al nivel de signif. del 5% que la edad afecta de igual modo al valor de las viviendas de estilo tradicional y contemporáneo (Δ=0)? Ilústralo abajo indicando los valores críticos en ambas figuras y rellena los "?"



$\hat{\Delta} \sim N(?, ?)$



$t \sim N(0,1)$

e. Dibuja dos diagramas de dispersión con las regresiones (2) y (3); *¡ojo también a sus R^2!*

Ejemplo 4b de EXAMEN PARCIAL

(si entendiste la solución al Ejemplo 4a, sabrías responder este ejemplo similar)

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) Se pedirá ***demostrar*** *UNO de los enunciados de la última página:*

o bien el resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos,*

o bien, la derivación *YYY* hecha en clase, ejercicios…

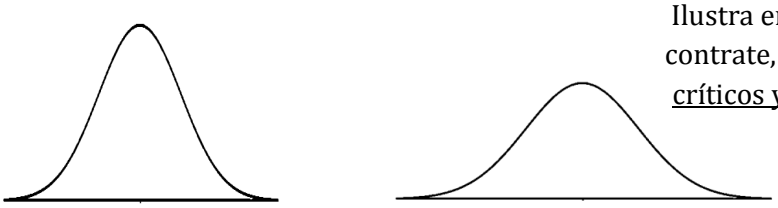
2. (3.5 puntos) *Una importante agencia de publicidad quiere promocionar sus servicios documentando el impacto que esta tiene en las ventas. Para empezar recoge datos de las ventas de sus 105 empresas-cliente actualmente y calcula unas ventas medias semanales iguales a 37.84 millones de $ (con una desviación típica de $1.95 millones).*

a. Basándonos en esta muestra actual, ¿en qué rango se encuentran con un 95% de confianza las ventas medias para cualquier empresa si recibe los servicios de la agencia?

La agencia observa que la cifra de ventas media semanal para las 54 empresas-cliente que más invierten en publicidad (por encima de la media) es $38.57 millones (con desviación típica de $1.88 millones), mientras que las 51 que gastan por debajo de la media venden $37.06 millones en promedio (con desviación típica igual a $1.73 millones).

b. ¿Rechazamos al nivel de signif. del 5% que las ventas medias de las empresas que gastan 'poco' en publicidad (menos de la media) sean iguales a las que gastan 'mucho'?

Ilustra en las dos figuras el contrate, indica sus valores críticos y rellena los dos '?'



$\hat{\Delta} \sim N(?, ?)$ $\quad$ $t \sim N(0,1)$

c. Da un modelo lineal de un regresor y una hipótesis que, contrastada, responda a (b)

d. ¿Por qué no podemos concluir en (b) que variar el gasto en publicidad causa cambios en ventas? Explícalo también en el contexto del modelo en (c) usando una fórmula

e. Señala muy claramente cuál es el rasgo esencial de un experimento (RCT) que pueda responder a la pregunta de si "¿invertir más en publicidad aumenta las ventas?"

3. (4.5 puntos) *Una inmobiliaria quiere estudiar el efecto de la* superficie *(en "pies²" o "ft²" o "squared feet") de las viviendas en California en su* precio *(en $) de venta. Para ello, obtenemos datos observacionales de una muestra aleatoria de 1080 viviendas y estimamos:*

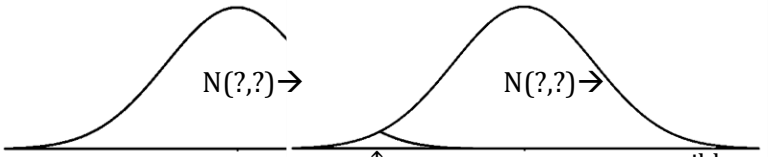
$$(\mathbf{1}) \quad \widehat{precio} = -60861.5 + 92.75\ superficie,\ R^2 = 0.58,\ SER = 79822$$

……. (19491) (9.2) *errores típicos robustos a la heterocedast.*

a. Interpreta el número (*¡y no solamente su signo!*) 92.75 y di si es significativo al 5%

b. Interpreta el valor 79822 (*¡escribe el número en tu frase!*) e indica sus unidades

c. Calcula el intervalo del 95% de confianza para la diferencia de precio entre dos casas con las mismas características pero una mide 500 ft² más que la otra. Ilustra tu intervalo en las figuras y rellena los 6 "?"



posibles valores de ?

d. Pero, siendo críticos, ¿podría ser que la estimación del efecto de la superficie en el precio de las viviendas en (1) esté sesgada porque capte también el número de aseos y habitaciones? Explica qué signo tendría el sesgo y justifica tu respuesta con una fórmula

La superficie media de las viviendas de California es 2325 ft² (y desviación típica de 1008)

e. *La casa donde murió Michael Jackson (que mide 17171 ft²) fue vendida por 18.1 millones de dólares*: calcula el residuo de estimar su valor y explica su anormal tamaño

Para atraer a inversores españoles la agencia inmobiliaria altera su estudio de dos maneras: por un lado, cambia las unidades de "ft²" a "m²" (1 ft² = 0.093 m²) (…)

f. Aplica el cambio de unidades a *m²* y reescribe lo necesario de toda la regresión (1)

(…) *y, por otro, crea la variable binaria "*grande*" igual a 1 si una vivienda mide más de 3333 ft² (o 0 si no) y estima:* **(2)** MCO, usando las observaciones 1-1080

Variable dependiente: **precio**
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	129635	1663.91	77.91	<0.0001	***
grande	214538	23216.9	9.241	<0.0001	***

Media de la vble. dep.	154863.2	D.T. de la vble. dep.	122912.8
Suma de cuad. residuos	1.11e+13	D.T. de la regresión	101670.0
R-cuadrado	0.316421	R-cuadrado corregido	0.315787
F(1, 1078)	85.38850	Valor p (de F)	1.27e-19

g. Interpreta brevemente la $\widehat{\beta_1}$ de la regresión (**2**) y di si es significativa al 5%

h. Explica por qué el R² sólo puede bajar al estimar el modelo (**2**) en lugar de (**1**)

Lista de RESULTADOS BÁSICOS demostrados en el APÉNDICE para el FINAL

(1) $E(aX+c) = aE(X) + c$

(2) $var(X) = E[X^2] - E(X)^2$

(3) $var(aX+c) = a^2 var(X)$

(4) $E(aX+bY+c) = aE(X) + bE(Y) + c$

(5) $cov(X,Y) = E(XY) - E(X)E(Y)$

(6) $cov(X,X) = var(X)$

(7) $cov(aX, c+bY) = ab\ cov(X,Y)$

(8) $cov(aX+bY+c, Z) = a\ cov(X,Z) + b\ cov(Y,Z)$

(9) $var(aX+bY+c) = a^2\ var(X) + b^2\ var(Y) + 2ab\ cov(X,Y)$

(12) $E[E(X|Y)] = E(X)$ Ley de Esperanzas Iteradas (LIE)

(13) $P(X = x_i \cap Y = y_j) = P(X = x_i) \cdot P(Y = y_j)\ \forall\ x_i, y_j \leftrightarrow X$ e Y independientes

(14) X *e* Y *independientes* $\rightarrow\ corr(X,Y) = 0$

Lista de DEMOSTRACIONES hechas en CLASE, EJERCICIOS…para el FINAL

(I) $E(\bar{Y}) = \mu_Y$ para una muestra $\{Y_1, Y_2, \dots, Y_n\}$ i.i.d.

(II) $S_X^2 = \left(\frac{n}{n-1}\right)\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2\right)$ es la varianza de la muestra $\{X_1, X_2, \dots, X_n\}$

(III) $S_{XY} = \left(\frac{n}{n-1}\right)\left(\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\,\bar{y}\right)$ para la muestra $\{Y_1, Y_2, \dots, Y_n, X_1, X_2, \dots, X_n\}$

(IV) *resuelve el problema de MCO* para estimar $Y_i = \beta_0 + \beta_1 X_i + u_i$ y demuestra que $\sum_{i=1}^n \hat{u}_i = 0$, y $\sum_{i=1}^n \hat{u}_i x_i = 0$, así como $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$, y $\widehat{\beta_1} = S_{XY}/S_X^2$

(V) resuelve de nuevo para $Y_i = \beta_0 + \beta_1 X_i + u_i$ con $x_i = 0\ \forall\ i$ y muestra que $\hat{\beta}_0 = \bar{y}$

(VI) $S_{\hat{u}X} = 0$ cuando estimas $Y_i = \beta_0 + \beta_1 X_i + u_i$ usando MCO

(VII) si $E(u|X) = constante$, entonces $corr(u,X) = 0$

(VIII) en $Y = \beta_0 + \beta_1 X + u$: si X sube en 1 unidad, Y aumenta en β_1 unidades *de media*

(IX) … pero $\beta_1 = E(Y|D=1) - E(Y|D=0)$ en $Y_i = \beta_0 + \beta_1 D_i + u_i$ si D es binaria

(X) en $Y = \beta_0 + \beta_1 \ln(X) + u$: si X sube un 1%, Y aumenta en $0.01 \cdot \beta_1$ unidades *de media*

(XI) en $\ln(Y) = \beta_0 + \beta_1 X + u$: si X aumenta en 1 unidad, Y sube $(100 \cdot \beta_1)$% *de media*

(XII) en $\ln(Y) = \beta_0 + \beta_1 \ln(X) + u$: si X sube un 1%, Y aumenta un β_1% *de media*

(XIII) en $Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \cdot D) + u$: el efecto *medio* de X en Y sube β_3 si $D=1$ y el efecto *esperado* de D en Y depende, asimismo, del nivel de X

(XIV) $SCT = SCE + SCR$ cuando estimas $Y = \beta_0 + \beta_1 X + u$ usando MCO

Momentos / Parámetros	Estimaciones muestrales / Estadísticos
$E(X) = \sum_{j=1}^{m} x_j\, P(X = x_j) \equiv \mu_X$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \hat{\mu}_X$
$var(X) = E[(X-\mu_X)^2] \equiv \sigma_X^2$	$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \equiv \hat{\sigma}_X^2$
$sd(X) = \sqrt{\sigma_X^2} \equiv \sigma_X$	$s_X = \sqrt{s_X^2} \equiv \hat{\sigma}_X$
$cov(X,Y) =$ $= E[(X-\mu_X)(Y-\mu_Y)] \equiv \sigma_{XY}$	$s_{XY} =$ $= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \equiv \hat{\sigma}_{XY}$
$corr(X,Y) = {}^{\sigma_{XY}}/_{\sigma_X\sigma_Y} \equiv \rho_{XY}$	$r_{XY} = {}^{s_{XY}}/_{s_X s_Y} \equiv \hat{\rho}_{XY}$
$Y = \beta_0 + \beta_1 X_1 + u \quad \rightarrow \hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 \;;\; \hat{u} = Y - \hat{Y}$ $\{ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u \quad \rightarrow \hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_k} X_k \quad ;\; \hat{u} = Y - \hat{Y} \}$	
$\beta_1 = \frac{\Delta E(Y\,\vert X_1)}{\Delta X_1} \left\{ = \frac{\Delta E(Y\,\vert X_1,\ldots,X_k)}{\Delta X_1} \right\}$	$\widehat{\beta_1} = \frac{s_{YX}}{s_X^2} \left\{ = \frac{s_{Y1}s_{22} - s_{12}s_{Y2}}{s_{11}s_{22} - s_{12}^2} \;\; si \;\; k = 2 \right\}$
$\beta_0 = E(Y\,\vert X_1 = 0)$ $\{ = E(Y\,\vert X_1 = X_2 = \cdots = X_k = 0) \}$	$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}_1$ $\{ = \overline{Y} - \widehat{\beta_1}\overline{X}_1 - \widehat{\beta_2}\overline{X}_2 - \cdots - \widehat{\beta_k}\overline{X}_k \}$
||||||||||||||||||||||||||||||||||||	||||||||||||||||||||||||||||||||||||
$E(\overline{X}) = \mu_{\overline{X}} \equiv \mu_X$	
$var(\overline{X}) = \frac{1}{n}\sigma_X^2 \equiv \sigma_{\overline{X}}^2$	$\hat{\sigma}_{\overline{X}}^2 = \frac{1}{n}\hat{\sigma}_X^2 \equiv s_{\overline{X}}^2$
$sd(\overline{X}) = \frac{1}{\sqrt{n}}\sigma_X \equiv \sigma_{\overline{X}}$	$\hat{\sigma}_{\overline{X}} = \frac{1}{\sqrt{n}}\hat{\sigma}_X \equiv SE(\overline{X})$
$E(\widehat{\beta_1}) \cong \beta_1 + \rho_{Xu}\,{}^{\sigma_u}/_{\sigma_X}$	
$var(\widehat{\beta_1}) = \frac{1}{n}\frac{var((X-\mu_X)u)}{(var(X))^2} \equiv \sigma_{\widehat{\beta_1}}^2$ $\{o\ complicada\ si\ \ k > 1\}$	$\hat{\sigma}_{\widehat{\beta_1}}^2 = \frac{1}{n}\frac{\sum_{1=1}^{n}\hat{u}_i{}^2(x_i-\overline{x})^2/(n-2)}{[\sum_{1=1}^{n}(x_i-\overline{x})^2/n]^2}$ $\{o\ complicada\ si\ \ k > 1\}$
$sd(\widehat{\beta_1}) = \sqrt{\sigma_{\widehat{\beta_1}}^2} \equiv \sigma_{\widehat{\beta_1}}$	$\hat{\sigma}_{\widehat{\beta_1}} = \sqrt{\hat{\sigma}_{\widehat{\beta_1}}^2} \equiv SE(\widehat{\beta_1})$
||||||| $R^2 = \frac{SCE}{SCT} = \frac{\sum_{1=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{1=1}^{n}(y_i-\bar{y})^2}$; $\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2)$; $SER = \sqrt{\sum_{i=1}^{n}\hat{u}_i^2/(n-k-1)}$ |||||||	
Para contrastar $H_0: \beta_1 = 0\ \&\ \beta_2 = 0$ $F = \frac{t_1^2 + t_2^2 - 2\hat{\rho}_{12}t_1t_2}{2(1-\hat{\rho}_{12}^2)}$	(con homocedasticidad) $F = \frac{(R_{unres}^2 - R_{res}^2)/q}{(1-R_{unres}^2)/(n-k-1)} \;\; donde\ q = 2$

-2.575, -1.96 y -1.64 dejan 0.5, 2.5 y 5% de probabilidad en la cola izquierda de la N(0,1)
Los valores críticos al 5% de una χ_q^2/q con q=1, 2, 3, 4 y 5 son 3.84, 3, 2.6, 2.37 y 2.21

EJEMPLO de EXAMEN FINAL 1ª

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) *Demuestra* de los enunciados de la hoja-formulario**:**

a. resultado XXX del Apéndice, *usando álgebra y los resultados anteriores a estos*

b. la derivación YYY hecha en clase, ejercicios…

2. (3.5 puntos) *Un grupo defensor del medio ambiente mantiene que bajar el precio del transporte público (principalmente en autobús) en las ciudades de EEUU podría aumentar su uso (y, quizá, mejorar la calidad del aire). Queremos comprobar ese efecto en la demanda:*

a. ¿Qué rasgo de un experimento nos garantiza que podemos medir el efecto del precio del billete en el número de usuarios, y no el de cualquier otro factor?

En la práctica, no podemos hacer experimentos, pero tenemos datos observacionales para 120 ciudades sobre: la demanda de autobuses (BUS), en miles de pasajeros; el precio del billete (PR), precio de la gasolina (PGAS), y renta media de la ciudad per capita (INC), en dólares; población (POP), en millones de habitantes, y densidad (DENS) en habitantes por milla²; y si se encuentran en el Northeast (NE), Midwest (MW), South (S) o West (W), todas binarias. Estimamos:

Modelo 1: MCO, usando las observaciones 1-120
Variable dependiente: BUS
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
const	2377.56	848.674	2.8015	0.0060	***
PR	−324.026	228.508	−1.4180	0.1590	
PGAS	775.732	952.583	0.8143	0.4172	
INC	−0.197698	0.0307104	−6.4375	<0.0001	***
POP	1.58668	0.0933463	16.9978	<0.0001	***
DENS	0.143072	0.0228438	6.2631	<0.0001	***
NE	−43.8876	187.265	−0.2344	0.8151	
MW	127.599	**???**	**???**	0.5006	
S	−19.9891	178.139	−0.1122	0.9109	

Media de la vble. dep.	1933.175	D.T. de la vble. dep.	2411.236
Suma de cuad. residuos	54800219	D.T. de la regresión	702.6348
R-cuadrado	0.920794	R-cuadrado corregido	0.915086
F(8, 111)	116.4465	Valor p (de F)	2.33e-50

con la siguiente tabla de varianzas-covarianzas entre los coeficientes estimados:

const	***PR***	***PGAS***	***INC***	***POP***	***DENS***	***NE***	***MW***	***S***	
720247	14432.5	-617688	-12.263	-8.4279	7.32843	-15109	-16230	-2725.0	***const***
	52215.9	-102145	1.38831	-9.7479	1.90096	2091.82	-3976.4	-1978.9	***PR***
		907415	-3.6533	12.819	-6.6559	-6165.5	-5533.9	-18755	***PGAS***
			9.431e-4	5.31e-4	-3.34e-4	0.05319	0.49802	0.15146	***INC***
				0.00871	-0.0014	-1.1700	0.81179	1.06821	***POP***
					5.218-4	0.18554	-0.3269	-3.0e-4	***DENS***
						35068.2	17984.1	18086.2	***NE***
							35659.8	18250.4	***MW***
								31733.4	***S***

b. Contrasta H_0: $\beta_{PR} \geq 0$ usando como alternativa la hipótesis del grupo defensor del medio ambiente. Usa los resultados para argumentar en contra del *lobby*

c. Escribe la hipótesis del contraste de la regresión, ¿rechazamos al nivel de signif. del 5%?

d. Rellena los **dos** "???" en la salida de Gretl

e. Reescribe esta misma regresión si hubiéramos omitido *NE* en lugar de *W*

f. ¿Es el transporte en autobús urbano un bien inferior? Ilustra tu contraste (también al 5%) para H_0: $\beta_{INC} \leq 0$: valores críticos, valores estimados o estadísticos y rellena los '?'

?~N(? , ?) t~N(0,1)

3. (4.5 puntos) *Usando datos homocedásticos sobre salarios, etc, de 269 jugadores de la NBA, queremos determinar el efecto de los tantos marcados y rebotes recogidos por partido en su sueldo. En base a las columnas* ***A*** *-* ***E****, responde a cada apartado* ***a-e****:*

a. *La mayor parte de los jugadores recogen cuatro rebotes por partido.* Según el modelo **A** dibuja dos posibles diagramas de dispersión con datos de esos jugadores: uno para ***points*** y **ln(*wages*),** y otro para ***points*** y ***wages*****.**

b. Interpreta en **B** (mencionando el número en tu frase) $\widehat{\beta_0}$, equal to 5.56

c. Sombrea en gráficos como en **2.e,** el p-valor del contraste de significatividad de β_4 en **C**

d. Explica brevemente por qué $\widehat{\beta_4}$ pasa a ser negativo en ***D*** mientras era positivo en ***C***

e. Contrasta conjuntamente en **E** si estar casado y tener hijos tiene un efecto en el salario (al nivel de significación del 5%)

Variable dependiente: natural log de salarios (en miles de $) , i.e. **ln(*wages*)** ; **n=269**					
Regresores	***(A)***	***(B)***	***(C)***	***(D)***	***(E)***
1. *points*	0.075 (.0076)	0.109 (.0119)	0.101 (.0123)	0.090 (.0122)	0.090 (.0122)
2. *rebounds*	0.062 (.0152)	0.144 (.0259)	0.125 (.0269)	0.117 (.0261)	0.118 (.0266)
3. *points × rebounds*		-0.007 (.0015)	-0.006 (.0016)	-0.005 (.0017)	-0.005 (.0017)
4. *age*			0.060 (.0120)	-0.055 (.0306)	-0.057 (.0315)
5. *exper*				0.196 (.0541)	0.196 (.0541)
6. *exper²*				-0.005 (.0028)	*-0.005 (.0038)*
7. *married*					*0.022 (.0839)*
8. *children*					*0.018 (.0761)*
0. constante	5.90 (.0951)	5.56 (.1392)	4.01 (.3466)	6.49 (.7200)	6.54 (.7358)
R^2	0.4121	0.4353	0.4873	0.5207	0.5210

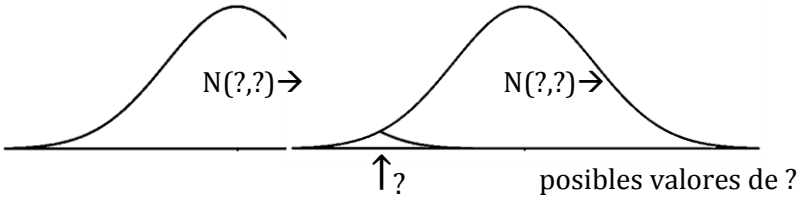
Las estimaciones en cursiva NO rechazan que el coeficiente sea igual a cero al nivel de significación del 10%. Married y children son variables binarias que toman el valor 1 o 0

*En base a las correspondientes columnas (de **A** a **E**) responde a las preguntas **f** a **h**:*

f. ¿Qué modelo elegirías tomando como criterio el R^2 corregido (o $\overline{R^2}$) entre ***C***, ***D*** y ***E***?

g. Según el modelo de regresión más explicativo, ¿cuál es la diferencia esperada en **ln(*wages*)** para dos jugadores de la NBA con básicamente las mismas características pero donde uno es 2 años más joven que el otro? Enuncia, también, qué diferencia aproximada habrá entre sus ***wages*** o salarios (si

h. Ilustra abajo usando un intervalo de 95% de confianza para la diferencia en **ln(*wages*)** o ***wages*** entre estos dos jugadores que esta no es significativa (-mente distinta de 0). Hazlo añadiendo una curva al gráfico y rellena 6 '?'



EJEMPLO de EXAMEN FINAL 1b

(si entendiste la solución al Ejemplo 1a, sabrías responder este ejemplo similar)

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) *Demuestra* de los enunciados de la hoja-formulario**:**

a. resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos*

b. la derivación *YYY* hecha en clase, ejercicios…

2. (4.5 puntos) *Una Agencia Federal quiere averiguar si la demanda de servicios de autobús en las ciudades de EEUU podría aumentar subvencionando el coste de los billetes:*

a. ¿Qué rasgo de un experimento nos garantiza que podemos medir el efecto del precio del billete en el número de usuarios, y no el de cualquier otro factor?

En la práctica, no podemos hacer experimentos, pero tenemos datos observacionales para 120 ciudades sobre: la demanda de autobuses (BUS), en miles de pasajeros; el precio del billete (PR), precio de la gasolina (PGAS), y renta media de la ciudad per capita (INC), en dólares; población (POP) en millones de habitantes, y densidad (DENS), en habitantes por milla^2; y si se encuentran en el Northeast (NE), Midwest (MW), South (S) o West (W), todas binarias. Estimamos:

Modelo 1: MCO, usando las observaciones 1-120
Variable dependiente: BUS
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
const	2377.56	848.674	2.8015	0.0060	***
PR	−324.026	228.508	−1.4180	0.1590	
PGAS	775.732	952.583	0.8143	0.4172	
INC	−0.197698	0.0307104	−6.4375	<0.0001	***
POP	1.58668	0.0933463	16.9978	<0.0001	***
DENS	0.143072	0.0228438	6.2631	<0.0001	***
NE	−43.8876	187.265	−0.2344	0.8151	
MW	127.599	188.838	0.6757	0.5006	
S	−19.9891	178.139	−0.1122	0.9109	

Media de la vble. dep.	1933.175	D.T. de la vble. dep.	2411.236
Suma de cuad. residuos	54800219	D.T. de la regresión	702.6348
R-cuadrado	0.920794	R-cuadrado corregido	0.915086
F(8, 111)	116.4465	Valor p (de F)	2.33e-50

con la siguiente tabla de varianzas-covarianzas entre los coeficientes estimados:

const	***PR***	***PGAS***	***INC***	***POP***	***DENS***	***NE***	***MW***	***S***	
720247	14432.5	-617688	-12.263	-8.4279	7.32843	-15109	-16230	-2725.0	***const***
	52215.9	-102145	1.38831	-9.7479	1.90096	2091.82	-3976.4	-1978.9	***PR***
		907415	-3.6533	12.819	-6.6559	-6165.5	-5533.9	-18755	***PGAS***
			???	5.31e-4	-3.34e-4	0.05319	0.49802	0.15146	***INC***
				0.00871	-0.0014	-1.1700	0.81179	1.06821	***POP***
					5.218-4	0.18554	-0.3269	-3.0e-4	***DENS***
						35068.2	17984.1	18086.2	***NE***
							35659.8	18250.4	***MW***
								31733.4	***S***

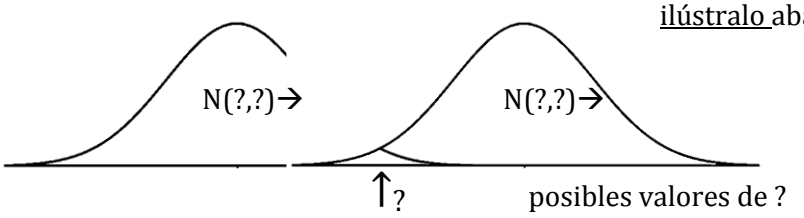
b. Explica si el efecto del precio de los billetes y la gasolina tiene el signo esperado

c. Los precios de los billetes y de la gasolina ¿son conjuntamente significativos para la demanda de servicios de autobús al nivel del 5%? Ilustra tu contraste dibujando la distribución del estadístico y su correspondiente p-valor

d. Rellena los dos "???" en la matriz de varianzas-covarianzas

e. ¿Cuál es el efecto estimado en los usuarios de autobús de una afluencia de inmigrantes que aumenta la población en 2000 residentes y la densidad en 1 habitante por *milla*^{2}?

f. Calcula el intervalo de 95% de confianza para el efecto en la demanda de autobús e ilústralo abajo rellenando los 6 '?'



g. Interpreta el coeficiente -19.99 (-20 aprox) para la variable binaria *S* (South)

3. (3.5 puntos) *Usando datos homocedásticos sobre salarios, etc, de 269 jugadores de la NBA, queremos determinar el efecto de los tantos marcados y rebotes recogidos por partido en su sueldo. En base a las columnas* ***A*** *-* ***E****, responde a cada apartado* ***a-e****:*

a. Si se omite la variable ***rebounds***, se estima $\widehat{\ln(wages)} = 6 + 0.093points$, explica por qué la beta estimada para ***point****s* es mayor que en la regresión ***A*** usando una fórmula

b. Da el cambio en **ln(*wages*)** estimado en **B** para un jugador que marca 1 punto más en cada partido pero nunca recoge rebotes. Haz lo mismo para un jugador que coge 4 rebotes por partido. Explica con tus palabras el efecto no-lineal de 1 punto extra en **ln(*wages*)**

c. Interpreta 0.06 como efecto en *wages*. Explica cómo es posible estimar $\widehat{\beta_4} > 0$

d. Dibuja un gráfico que muestre el efecto no-lineal de la experiencia en **ln(*wages*)**

e. Da la hipótesis, estadístico, valor crítico y conclusión del contraste de la regresión en **E**

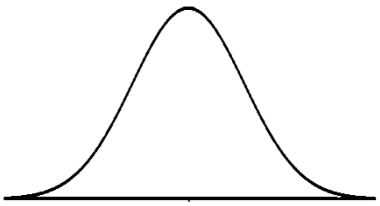
Variable dependiente: natural log de salarios (en miles de $) , i.e. **ln(*wages*)** ; **n=269**					
Regresores	***(A)***	***(B)***	***(C)***	***(D)***	***(E)***
1. *points*	0.075 (.0076)	0.109 (.0119)	0.101 (.0123)	0.090 (.0122)	0.090 (.0122)
2. *rebounds*	0.062 (.0152)	0.144 (.0259)	0.125 (.0269)	0.117 (.0261)	0.118 (.0266)
3. *points × rebounds*		-0.007 (.0015)	-0.006 (.0016)	-0.005 (.0017)	-0.005 (.0017)
4. *age*			0.060 (.0120)	-0.055 (.0306)	-0.057 (.0315)
5. *exper*				0.196 (.0541)	0.196 (.0541)
6. *exper²*				-0.005 (.0028)	*-0.005 (.0038)*
7. *married*					*0.022 (.0839)*
8. *children*					*0.018 (.0761)*
0. constante	5.90 (.0951)	5.56 (.1392)	4.01 (.3466)	6.49 (.7200)	6.54 (.7358)
R^2	0.4121	0.4353	0.4873	0.5207	0.5210

Las estimaciones en cursiva NO rechazan que el coeficiente sea igual a cero al nivel de significación del 10%. Married y children son variables binarias que toman el valor 1 o 0

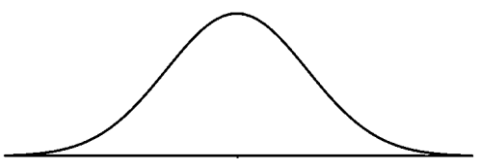
*En base a la columna correspondiente (de **A** a **E**) responde a las preguntas **f** y **g**:*

f. ¿Qué modelo tendrá el SER más alto? Justifica tu respuesta brevemente

g. Según el modelo de regresión más explicativo, ilustra el contraste de significatividad de la variable ***age*** que usa un nivel de significación del 10%. Muestra valores críticos y las estimaciones o estadísticos en los dos gráficos de abajo, y rellena los tres '?'



?~N(? , ?)



t~N(0,1)

EJEMPLO de EXAMEN FINAL 2a

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) *Demuestra* de los enunciados de la hoja-formulario**:**

a. resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos*

b. la derivación *YYY* hecha en clase, ejercicios…

2. (4 puntos) *Queremos medir el efecto de las ventas del producto de una empresa en el sueldo de su CEO (i.e. Chief Executive Officer, su más alto jefe ejecutivo).*

a. *¿cuál de estos experimentos NO asegura diferencias en salario de CEOs se deban a que su empresa venda más; o sea, NO mide un efecto causal de las ventas en el salario? Explica*
i. Selecciona una muestra de varias empresas diferentes con CEOs y aumenta aleatoriamente la demanda de los productos de la mitad de ellas. (…)
ii. Selecciona mediante muestreo aleatorio varias empresas con CEOs y aumenta la demanda de los productos de aquella mitad de empresas más pequeñas. (…)
iii. Selecciona mediante muestreo aleatorio varias empresas con CEOs y aumenta la demanda de los productos de la primera mitad seleccionada. (…)
iv. Selecciona una muestra de varias empresas pequeñas con CEOs y aumenta aleatoriamente la demanda de los productos de la mitad de ellas. (…)
(…) Compara el salario medio de los CEOs de las empresas cuyas ventas se aumentaron artificialmente con el de CEOs de empresas cuyas ventas no cambiaron

En lugar de hacer experimentos, usamos datos observacionales de 177 CEOs para estimar un modelo de regresión multivariante. Algunas variables se incluyen tomando logaritmos; por ejemplo, la variable dependiente **l_salary**=ln(**salary**), donde **salary** se mide en miles de $. *Esta regresión incluye factores como ventas o* ***sales*** *(en millones de $ y logs)***,** *el valor de mercado de la empresa (****mktval,*** *también en millones de $ y logs), el número de años trabajando en la empresa o "tenure in the company" (****comten****), y de años como CEO de la empresa o "tenure as CEO" (****ceoten****). Estimamos el Modelo 1.*

b. Interpreta $\widehat{\beta_0}$ como la estimación (¿de qué?) para un CEO con un perfil (¿cuál?)

c. ¿Cuál es el efecto esperado estimado en el salario o ***salary*** de un CEO si las ventas o ***sales*** de su empresa aumentan un 10%? ¿Es este efecto significativo a un nivel del 5%?

d. *Los expertos hablan del efecto "superstar": las empresas que contratan como CEOs a buenos gestores que no trabajaron antes en la empresa les ofrecen salarios relativamente más altos.* ¿Podemos refutar su afirmación con un nivel de significación del 5%?

Modelo 1: MCO, usando las observaciones 1-177
Variable dependiente: **l_salary**
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

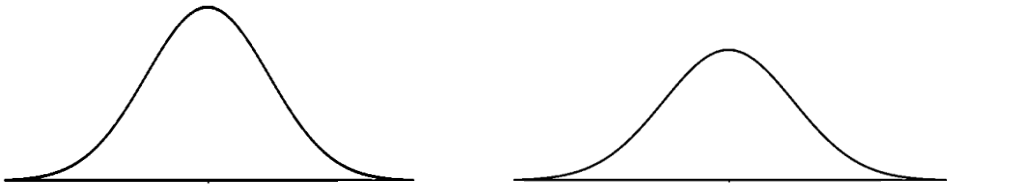
		Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
(0)	const	4.576	0.278	16.443	<0.0001	***
(1)	**l_sales**	0.192	0.036	5.371	<0.0001	***
(2)	**l_mktval**	0.094	0.049	1.928	0.0555	*
(3)	**ceoten**	0.017	0.007	2.281	0.0238	**
(4)	**comten**	−0.009	0.003	−3.035	0.0028	***

Media de la vble. dep.	6.5828	D.T. de la vble. dep.	0.6061
Suma de cuad. residuos	42.1259	D.T. de la regresión	0.4949
R-cuadrado	0.34836	R-cuadrado corregido	0.3332
F(4, 172)	26.6225	Valor p (de F)	3.40e-17

Matriz de covarianzas de los coeficientes *NB: 2e-003 = 0.002*

	(0) **const**	(1) **l_sales**	(2) **l_mktval**	(3) **ceoten**	(4) **comten**	
(0)	0.07746	-0.00017	-0.00896	-0.0012	-5.005e-005	**const**
(1)		0.00128	-0.00125	-7.61e-005	1.405e-005	**l_sales**
(2)			0.00238	0.00021	-2.91e-005	**l_mktval**
(3)				**5.502e-005**	-6.33e-006	**ceoten**
(4)					9.623e-006	**comten**

e. *Un periodista dice que la elasticidad del sueldo de un CEO con las ventas es 0.5.* ¿Podemos rechazar su afirmación con un nivel de significación del 5%? Ilustra tu contraste localizando en estas dos gráficas los valores críticos, tu estimación o estadístico, y rellenando los "?"



f. Interpreta **5.502e-005** en la matriz de covarianzas: di qué es, explica qué significa, y menciona las unidades correspondientes a este número

3. (4 puntos) *La Domestic Affairs Federal Agency de EEUU investiga el nivel de riqueza financiera de los trabajadores. Basándose en una muestra de 3637 trabajadores mayores de 25 años, esta agencia estima varias regresiones que explican sus activos financieros netos (**nettfa**, en miles de $) con la renta anual (**inc**, también en miles de $), su edad o **age** (en años por encima de 25) y **age²**, así como si participan en un plan de pensiones llamado 401k (**p401k** toma valor 1 si participa; si no, 0). Responde con las siguientes estimaciones **A – D** y errores standard robustos a la heterocedasticidad a cada pregunta **a-d:***

Variable dependiente: activos financieros netos o ***nettfa*****,** en miles de $; **n=3637**				
Regresores	**(A)**	**(B)**	**(C)**	**(D)**
(1) *inc*	1.17 (.107)	1.13 (.106)	1.08 (.110)	0.879 (.357)
(2) *age*		-0.107 (.435)	-0.009 (.445)	-0.012 (.445)
(3) age^2		0.041 (.014)	0.039 (.014)	0.039 (.014)
(4) *p401k*			17.37 (2.52)	6.14 (14.99)
(5) *p401k × inc*				0.26 (.369)
(0) constante	-25.1 (4.34)	-36.4 (4.09)	-47.1 (4.07)	-38.6 (12.04)
R^2	0.16	0.20	0.20	0.21
F (test de la regresión)	121.2	87.9	102.7	85.5

*NB: si, por ejemplo, el trabajador tiene 27 años, entonces ***age***=2 y ***age²*** = 4

a. ¿Cómo se interpreta la pendiente? Comenta su valor (mencionando sus unidades) en tu frase y NO solamente el signo

b. contrasta individualmente si β_2 y β_3 son significativos, y después dibuja un diagrama de dispersión probable para las edades y cantidad de activos financieros netos

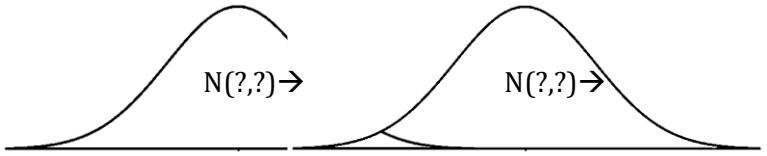
c. interpreta (usando su valor en tu frase) la beta estimada para ***p401k***

d. escribe y dibuja las "rectas" que explican la relación entre la renta y la riqueza financiera para trabajadores de 25 años (i.e. sus valores de ***age*** y ***age²*** son 0)

*En base al **modelo más explicativo** de todas las columnas responde a las preguntas **e-h**:*

e. escribe las hipótesis y recalcula el estadístico-*F* del "contraste de la regresión" *ignorando cualquier heterocedasticidad en los datos* y aclara si llegas a la misma conclusión que usando el estadístico *F* de la tabla del enunciado

f. calcula el intervalo de confianza del 95% para la diferencia en ***nettfa*** entre dos trabajadores que NO participan en el plan de pensiones y sólo difieren en que uno gana *$2000* anuales menos que el otro. Ilústralo abajo y rellena los 6 '?'



g. ¿es esta diferencia significativa (-mente distinta de 0)? Justifica tu respuesta e ilústrala añadiendo una curva a la figura de arriba

EJEMPLO de EXAMEN FINAL 2b

(si entendiste la solución al Ejemplo 2a, sabrías responder este ejemplo similar)

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) *Demuestra* de los enunciados de la hoja-formulario**:**

a. resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos*

b. la derivación *YYY* hecha en clase, ejercicios…

2. (4 puntos) *Queremos medir el efecto de las ventas del producto de una empresa en el sueldo de su CEO (i.e. Chief Executive Officer, su más alto jefe ejecutivo).*

a. *¿cuál de estos experimentos ASEGURA que diferencias en salario de CEOs se deben a que su empresa venda más; es decir, MIDE un efecto causal de las ventas en el salario? Explica*
i. Selecciona mediante muestreo aleatorio varias empresas y sus CEOs y divídela en dos grupos: aquellas cuyas ventas son superiores a la media y aquellas que son inferiores (…)
i. Selecciona por muestreo aleatorio varias empresas y sus CEOs y divídela en dos grupos: aquellas cuyas ventas son superiores a la mediana y aquellas que son inferiores (…)
iii. Selecciona mediante muestreo aleatorio varias empresas y sus CEOs y aumenta la demanda de los productos de aquella mitad de empresas que más venden. (…)
iv. Selecciona una muestra de varias empresas y sus CEOs y aumenta aleatoriamente la demanda de los productos de la mitad de ellas. (…)
(…) Compara el salario medio de los CEOs de ambos grupos de empresas.

En lugar de hacer experimentos, usamos datos observacionales de 177 CEOs para estimar un modelo de regresión multivariante. Algunas variables se incluyen tomando logaritmos; por ejemplo, la variable dependiente **l_salary**=ln(**salary**), donde **salary** se mide en miles de $. *Esta regresión incluye factores como ventas o* ***sales*** *(en millones de $ y logs)***,** *el valor de mercado de la empresa (****mktval,*** *también en millones de $ y logs), el número de años trabajando en la empresa o "tenure in the company" (****comten****), y de años como CEO de la empresa o "tenure as CEO" (****ceoten****). Estimamos el Modelo 1.*

b.¿cuál es el efecto estimado en el sueldo o ***salary*** de un CEO de un aumento del 1% en ventas o ***sales***?

c. ¿cuál es el efecto estimado en el sueldo o ***salary*** de un CEO de un aumento del 5% en el valor de la empresa o ***mktval***? ¿Es este efecto significativo a un nivel del 5%?

d. Comenta brevemente si $\widehat{\beta_3}$ tiene el signo esperado

Modelo 1: MCO, usando las observaciones 1-177
Variable dependiente: **l_salary**
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

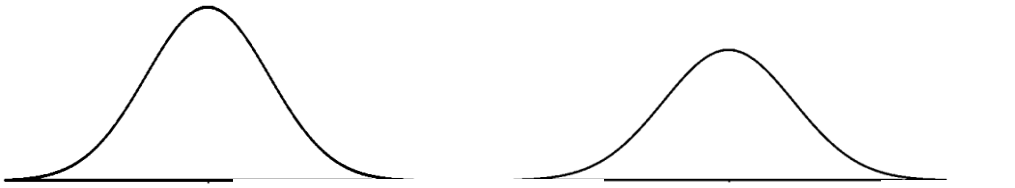
		Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
(0)	const	4.576	0.278	16.443	<0.0001	***
(1)	**l_sales**	0.192	0.036	5.371	<0.0001	***
(2)	**l_mktval**	0.094	0.049	1.928	0.0555	*
(3)	**ceoten**	0.017	0.007	2.281	0.0238	**
(4)	**comten**	−0.009	0.003	−3.035	0.0028	***

Media de la vble. dep.	6.5828	D.T. de la vble. dep.	0.6061
Suma de cuad. residuos	42.1259	D.T. de la regresión	0.4949
R-cuadrado	0.34836	R-cuadrado corregido	0.3332
F(4, 172)	26.6225	Valor p (de F)	3.40e-17

Matriz de covarianzas de los coeficientes *NB: 2e-003 = 0.002*

	(0) **const**	(1) **l_sales**	(2) **l_mktval**	(3) **ceoten**	(4) **comten**	
(0)	0.07746	-0.00017	-0.00896	-0.0012	-5.005e-005	**const**
(1)		0.00128	-0.00125	-7.61e-005	1.405e-005	**l_sales**
(2)			0.00238	0.00021	-2.91e-005	**l_mktval**
(3)				**5.502e-005**	-6.33e-006	**ceoten**
(4)					9.623e-006	**comten**

e. Contrasta al nivel de significación del 5% si dos CEOs que trabajan para empresas muy similares pero que tienen distinta "tenure" ganarían salarios diferentes. En concreto, la única diferencia es que un CEO ha trabajado 2 años más en la empresa que el otro, y 1 de esos 2 años como CEO. Da hipótesis, estimación o estadístico, valor crítico y rellena los "?"



$\Delta\ln(\widehat{salary}) \sim N(\,?\,,\,?\,)$ $\quad$ $t \sim N(0,1)$

f. Escribe las hipótesis del "contraste de la regresión", estadístico y justifica tu conclusión

3. (4 puntos) *La Domestic Affairs Federal Agency de EEUU investiga el nivel de riqueza financiera de los trabajadores. Basándose en una muestra de 3637 trabajadores mayores de 25 años, esta agencia estima varias regresiones que explican sus activos financieros netos (**nettfa**, en miles de $) con la renta anual (**inc**, también en miles de $), su edad o **age** (en años por encima de 25) y **age²**, así como si participan en un plan de pensiones llamado 401k (**p401k** toma valor 1 si participa; si no, 0). Responde con las siguientes estimaciones **A – D** y errores standard robustos a la heterocedasticidad a cada pregunta **a-d:***

Variable dependiente: activos financieros netos o ***nettfa*****,** en miles de $; **n=3637**				
Regresores	**(A)**	**(B)**	**(C)**	**(D)**
(1) *inc*	1.17 (.107)	1.13 (.106)	1.08 (.110)	0.879 (.357)
(2) *age*		-0.107 (.435)	-0.009 (.445)	-0.012 (.445)
(3) age^2		0.041 (.014)	0.039 (.014)	0.039 (.014)
(4) *p401k*			17.37 (2.52)	6.14 (14.99)
(5) *p401k × inc*				0.26 (.369)
(0) constante	-25.1 (4.34)	-36.4 (4.09)	-47.1 (4.07)	-38.6 (12.04)
R^2	0.16	0.20	0.20	0.21
F (test de la regresión)	121.2	87.9	102.7	85.5

*NB: si, por ejemplo, el trabajador tiene 27 años, entonces ***age***=2 y ***age²*** = 4

a. ¿Cómo se interpreta la ordenada en el origen (intercepto)? Comenta su valor y menciona sus unidades en tu frase

b. contrasta en **B**, *asumiendo que los datos son homocedásticos*, si debemos controlar por el hecho de que nuestra muestra contiene trabajadores de distintas edades. Escribe las hipótesis, calcula el estadístico, describe su distribución, y los valores críticos para un nivel de significación del 5%, y tu conclusión

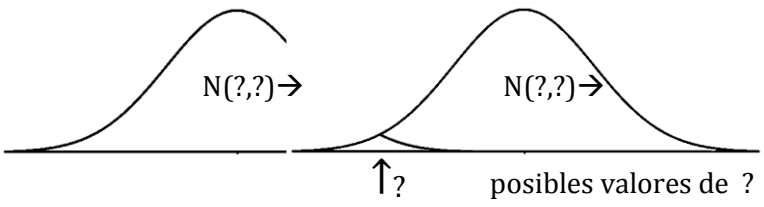
c. reescribe las betas estimadas (y NO sus SE), el R^2 y estadístico-F del "contraste de la regresión" para **C** si se hubiera incluido la variable ***nonp401k*** (que toma el valor 1 si el trabajador NO participa en el plan de pensiones y 0 si sí participa) en lugar de ***p401k***

d. *un trabajador de 25 años gana $24000 anuales de media.* ¿Qué efecto tendría participar en plan de pensiones 401k en su ***nettfa*** según **D**? ¿Y si gana *$20000*? ¿Y si gana *$28000*?

e. ¿cuál es el modelo con mayor R^2 corregido (o $\overline{R^2}$) de todas las columnas? Calcúlalo

En base al ***modelo con mayor R^2 corregido*** *(o $\overline{R^2}$), responde a las preguntas* ***f*** *y* ***g****:*

f. computa e ilustra abajo el intervalo del 95% de confianza para la diferencia en ***nettfa*** entre dos trabajadores de la misma edad que NO participan en el plan de pensiones 401k pero uno gana *$3000* menos que el otro. Rellena los 6 '?'



g. ¿podríamos incluir ambas variables ***p401k*** y ***nonp401k*** en esa regresión? Da nombre a las posibles consecuencias de hacerlo y explica lo que este significa

EJEMPLO de EXAMEN FINAL 3a

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) *Demuestra* de los enunciados de la hoja-formulario**:**

a. resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos*

b. la derivación *YYY* hecha en clase, ejercicios…

2. (4 puntos) *El Servicio Regional de Salud (SRS) necesita medir el efecto en la salud de un bebé de una* ***visita extra*** *al médico de la madre durante el embarazo. Su salud se mide en su peso al nacer (****weight****, en gramos).*

a. *¿Cuál de estos experimentos ASEGURA que diferencias en* ***weight*** *se deban a una* ***visita médica extra****; o sea, MIDE el efecto causal de una visita extra en el peso del bebé? Explica*
i. Selecciona una muestra de bebés y divídela en dos grupos: aquellos cuya madre visitó menos al médico que la media y aquellos cuya madre recibió más atención médica (…)
ii. Selecciona una muestra de diferentes madres y haz que una mitad aleatoria de entre ellas reciba una visita extra de atención médica prenatal. (…)
iii. Selecciona mediante muestreo aleatorio varias madres y haz que reciban una visita extra de atención médica prenatal aquellas que son madres primerizas. (…)
iv. Selecciona por muestreo aleatorio varios bebés y divídelos en dos grupos: aquellos que son el primer hijo de la madre o primogénitos y aquellos que no lo son. (…)
(…) Compara el peso medio de los bebés en los dos grupos.

En lugar de experimentos, el SRS estima con datos observacionales una regresión. El modelo incluye una relación cuadrática entre ***weight*** *y el número de visitas al médico (o sea incluye* ***visit*** *y* ***visit²****). La regresión también controla por el efecto de* ***drink****: el promedio de bebidas alcohólicas consumidas semanalmente por la madre durante el embarazo (entre 0 y 8).*

Modelo 1: MCO, usando las observaciones 1-1651
Variable dependiente: **weight**
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
(0) const	3102.36	98.4668	31.5067	<0.0001	***
(1) **visit**	36.2292	12.925	2.8030	0.0051	***
(2) **visit2**	−0.75307	0.4105	−1.8345	0.0668	*
(3) **drink**	−48.9098	25.4192	−1.9241	0.0545	*

Media de la vble. dep.	3409.520	D.T. de la vble. dep.	575.1008
Suma de cuad. residuos	5.39e+08	D.T. de la regresión	572.0800
R-cuadrado	0.012277	R-cuadrado corregido	0.010478
F(3, 1647)	5.577219	Valor p (de F)	0.000833

Matriz de covarianzas de los coeficientes *NB: 2e-003 = 0.002*

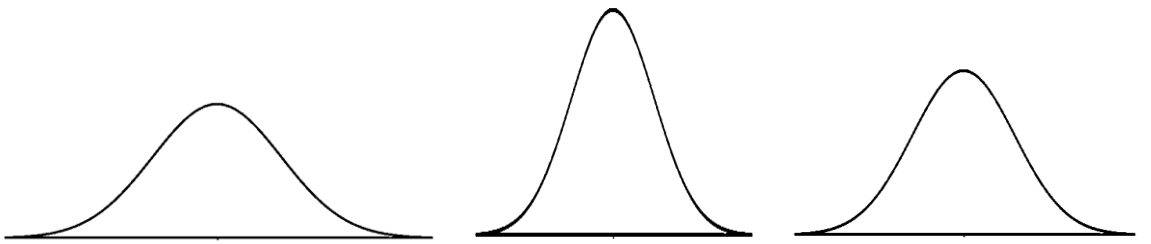
	(0)	(1)	(2)	(3)	
	const	**visit**	**visit2**	**drink**	
(0)	9695.70	-1229.76	33.8000	25.4714	**const**
(1)		167.055	-4.99249	-5.24652	**visit**
(2)			0.168510	0.0394247	**visit2**
(3)				646.138	**drink**

b. Dibuja un diagrama de dispersión probable para el número de visitas médicas y el peso del bebé al nacer para madres con un consumo de bebidas alcohólicas semanal similar

c. Contrasta al nivel de signif. del 5% si el efecto de las visitas en **weight** es no-lineal

d. Interpreta $\widehat{\beta_0}$ como la estimación (¿de qué?) para un bebé con un tipo de perfil (¿cuál?)

e. Estima la brecha media en el peso de los bebés esperados cuyas madres sólo difieren en que una consume dos copas más de alcohol que la otra. ¿Es significativa (-mente distinta de cero)? Ilustra tu contraste situando en los gráficos de abajo los valores críticos relevantes y las estimaciones (o estadísticos) y rellena los "?"



$\Delta\widehat{weight} \sim \mathrm{N}(?,?)$ $\quad\quad$ $\hat{\beta} \sim \mathrm{N}(?,?)$ $\quad\quad$ $\mathrm{t}\sim\mathrm{N}(0,1)$

f. Contrasta con un nivel de signif. del 5% si nuestro modelo multivariante es preferible a un modelo de un único regresor más simple que explique **weight** sólo con **visits** (y const)

3. (4 puntos) *Queremos medir la relación entre el tiempo dedicado a dormir y trabajar (**sleep** & **totwrk**, en minutos por semana) controlando el efecto de la formación (**educ,** en años y tomando logaritmos), su sexo, y una variable binaria **kids** que capta si los individuos tienen hijos menores de 3 años. Usando cada una de las regresiones **A - D** y sus errores standard robustos a la heterocedasticidad, responde la pregunta correspondiente de **a-d***

a. Interpreta el valor numérico (*¡y no solamente su signo!*) -0.15 e indica sus unidades

b. ¿cuál correlation entre ***totwrk*** & ***ln(educ)*** ? Razona aplicando una formula a **A** y **B**

c. reescribe la regresion **C** (las betas y NO sus errores standard) para el caso en que hubiéramos incluido ***female*** (en lugar de ***male***) para tener en cuenta el sexo en el modelo

d. contrasta en **D** *asumiendo homocedasticidad* y al nivel de significación del 5% si tener hijos menores de 3 años afecta de algún modo a tus predicciones de minutos de sueño

Variable dependiente: minutos semanales de sueño o ***sleep*** por la noche ; **n=706**					
Regresores		**(A)**	**(B)**	**(C)**	**(D)**
(1)	*totwrk*	-0.150 (.018)	-0.150 (.019)	-0.166 (.020)	-0.179 (.021)
(2)	*ln(educ)*		-170.6 (57.2)	-174.2 (57.1)	-166.9 (57.8)
(3)	*male*			90.96 (35.4)	89.17 (35.6)
(4)	*kids*				-239.7 (124.7)
(5)	*kids × totwrk*				0.107 (.055)
(0)	constant	3586 (42.0)	4011 (145.6)	4006 (145.4)	4017.6 (147)
R^2		0.1032	0.1130	0.1218	0.1273
F (contraste de la regresión)		65.69	40.04	28.68	18.25

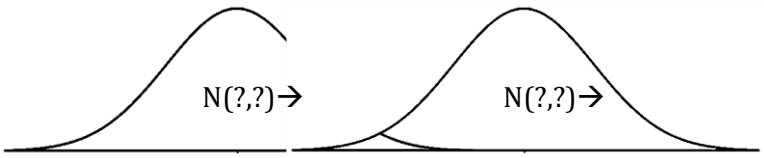
*NB: todos los trabajadores tienen al menos 1 año de formación, luego ***educ*** ≥ 1 y ***ln(educ)*** ≥ 0

*En base al **modelo con mayor R^2 corregido entre B, C y D,** responde a las preguntas **e - g**:*

e. *considera el caso de hombres con hijos menores de 3 años y 1 año de formación.* ¿Cuál es el efecto estimado en su sueño de un aumento en 1 hora de su tiempo dedicado a trabajar?

f. contrasta con un nivel de signif. del 5% si el efecto de ***totwrk*** en ***sleep*** para una trabajadora sin hijos y 1 año de formación es igual al predicho por la regresión **A** para otra trabajadora con iguales características

g. calcula el rango del 95% de confianza para la brecha media de sueño entre dos trabajadores (sin hijos) que sólo se diferencian en que uno trabaja 30 minutos por semana más que el otro. ¿Es esta brecha significativa (-mente distinta de cero)? Ilustra tu contraste abajo: añadiendo una curva al gráfico y rellenando los 6 "?"



↑? posibles estimaciones de ?

EJEMPLO de EXAMEN FINAL 3b

(si entendiste la solución al Ejemplo 3a, sabrías responder este ejemplo similar)

Por favor, lee las preguntas con atención y da una respuesta completa y clara. ¡Suerte!

1. (2 puntos) *Demuestra* de los enunciados de la hoja-formulario**:**

a. resultado *XXX* del Apéndice, *usando álgebra y los resultados anteriores a estos*

b. la derivación *YYY* hecha en clase, ejercicios…

2. (4 puntos) *El Servicio Regional de Salud (SRS) necesita medir el efecto en la salud de un bebé de una* ***visita extra*** *al médico de la madre durante el embarazo. Su salud se mide en su peso al nacer (****weight****, en gramos).*

a. *¿Cuál de estos experimentos NO asegura que diferencias en* ***weight*** *se deban a una* ***visita médica extra****; o sea, NO mide el efecto causal de una visita extra en el peso del bebé? Explica*
i. Selecciona mediante muestreo aleatorio varias madres primerizas y haz que reciban una visita extra de atención médica prenatal la mitad de ellas. (…)
ii. Selecciona mediante muestreo aleatorio varias madres y haz que reciban una visita extra de atención médica prenatal la primera mitad seleccionada. (…)
iii. Selecciona una muestra de diferentes madres y haz que una mitad aleatoria de entre ellas reciba una visita extra de atención médica prenatal. (…)
iv. Selecciona mediante muestreo aleatorio varias madres y haz que reciban una visita extra de atención médica prenatal aquellas que son madres primerizas. (…)
(…) Compara el peso medio de los bebés cuyas madres recibieron una visita de atención médica extra con el de aquellos cuyas madres no recibieron más atención médica prenatal

En lugar de experimentos, el SRS estima con datos observacionales una regresión. El modelo incluye una relación cuadrática entre ***weight*** *y el número de visitas al médico (o sea incluye* ***visit*** *y* ***visit²****). La regresión también controla por el efecto de* ***drink****: el promedio de bebidas alcohólicas consumidas semanalmente por la madre durante el embarazo (entre 0 y 8).*

Modelo 1: MCO, usando las observaciones 1-1651
Variable dependiente: **weight**
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*Valor p*	
(0) const	3102.36	98.4668	31.5067	<0.0001	***
(1) **visit**	36.2292	12.925	2.8030	0.0051	***
(2) **visit2**	−0.75307	0.4105	−1.8345	0.0668	*
(3) **drink**	−48.9098	25.4192	−1.9241	0.0545	*

Media de la vble. dep.	3409.520	D.T. de la vble. dep.	575.1008
Suma de cuad. residuos	5.39e+08	D.T. de la regresión	572.0800
R-cuadrado	0.012277	R-cuadrado corregido	0.010478
F(3, 1647)	5.577219	Valor p (de F)	0.000833

Matriz de covarianzas de los coeficientes *NB: 2e-003 = 0.002*

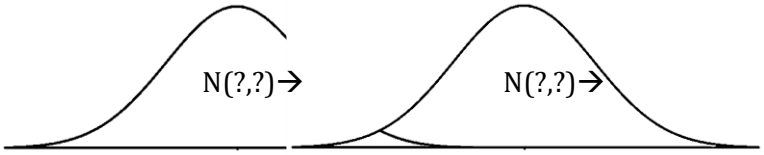
	(0) const	(1) visit	(2) visit2	(3) drink	
(0)	9695.70	-1229.76	33.8000	25.4714	**const**
(1)		167.055	-4.99249	-5.24652	**visit**
(2)			0.168510	0.0394247	**visit2**
(3)				646.138	**drink**

b. Si NO controlamos por el efecto del alcohol consumido, el nuevo modelo estimado es $\widehat{weight} = 3103.5 + 36.2288\, visit - 0.7529\, visit^2$. ¿Cuál es más o menos la correlación entre el número de visitas médicas pre-parto y ***drink**s* en los datos? Justifica tu respuesta

c. Contrasta al nivel de significación 5% si la atención médica previa al parto es un factor relevante para explicar el peso del bebé al nacer. Da las hipótesis, estadístico, su distribución, valores críticos y conclusión

d. ¿Cuál es el efecto estimado en el peso de los bebés de una política social que aumente de 9 a 10 el número medio de visitas médicas a la madre durante el embarazo?

e. ¿Y cuál sería el efecto estimado en el peso de los bebés (***weight***) de una política que suba de 19 a 20 el número medio de visitas? Construye el intervalo del 95% de confianza e ilústralo abajo rellenando los 5 '?'



posibles estimaciones del efecto

f. ¿Es esta estimación (del efecto de pasar de 19 a 20 visitas) significativa (-mente distinta de cero) al 5% de nivel de significación? Responde y sitúa esta H_0 en el gráfico de arriba

g. ¿Es tu anterior respuesta a **f** coherente con la que diste en **c** ? Explica brevemente

3. (4 puntos) *Queremos medir la relación entre el tiempo dedicado a dormir y trabajar (**sleep** & **totwrk**, en minutos por semana) controlando el efecto de la formación (**educ,** en años y tomando logaritmos), su sexo, y una variable binaria **kids** que capta si los individuos tienen hijos menores de 3 años. Usando cada una de las regresiones **A - D** y sus errores standard robustos a la heterocedasticidad, responde la pregunta correspondiente de **a-d***

a. reescribe la regresión **A** para el caso en que ***sleep*** y ***totwrk*** se midiesen en *horas* por semana en lugar de *minutos* (cambia SÓLO las betas estimadas, NO sus errores standard)

b. dibuja un diagrama de dispersión probable para los valores observados de ***sleep*** y ***education*** (sin tomar logaritmos) para trabajadores con valores similares de ***totwrk***

c. interpreta (incluyendo su valor y unidades en tu frase) la beta estimada para ***male***

d. *considera el caso de **mujeres** con un año de formación* (***educ***=1). Escribe y dibuja las dos "rectas" estimadas implícitamente que describen la relación-oposición entre el tiempo dedicado a dormir y a trabajar por aquellas con hijos menores de 3 años y sin ellos

Variable dependiente: minutos semanales de sueño o ***sleep*** por la noche ; **n=706**					
Regresores		**(A)**	**(B)**	**(C)**	**(D)**
(1)	*totwrk*	-0.150 (.018)	-0.150 (.019)	-0.166 (.020)	-0.179 (.021)
(2)	*ln(educ)*		-170.6 (57.2)	-174.2 (57.1)	-166.9 (57.8)
(3)	*male*			90.96 (35.4)	89.17 (35.6)
(4)	*kids*				-239.7 (124.7)
(5)	*kids × totwrk*				0.107 (.055)
(0)	constant	3586 (42.0)	4011 (145.6)	4006 (145.4)	4017.6 (147)
R^2		0.1032	0.1130	0.1218	0.1273
F (contraste de la regresión)		65.69	40.04	28.68	18.25

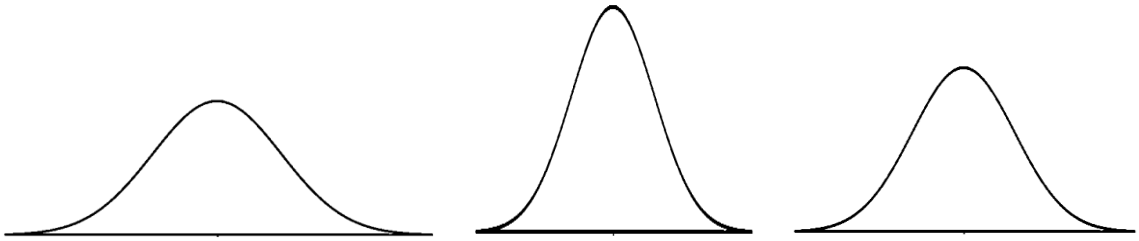
*NB: todos los trabajadores tienen al menos 1 año de formación, luego ***educ*** ≥ 1 y ***ln(educ)*** ≥ 0

*Basándote en **el modelo más explicativo** responde a las preguntas **e** - **g**:*

e. ¿llegas a la misma conclusión que en el "contraste de la regresión" de la tabla si lo rehaces al nivel de significación del 5% *ignorando la heteroscedasticidad* de los datos?

f. contrasta *asumiendo homoscedasticidad* si tener hijos menores de 3 años tiene cualquier impacto en la horas dedicadas al sueño usando un nivel de significación del 5%

g. estima la brecha media de sueño entre dos mujeres (sin hijos) si una trabaja 1 hora semanal más que la otra. ¿Es esta brecha significativa (-mente distinta de cero)? Ilustra tu contraste situando en cada figura de abajo los valores críticos y estimados (o estadísticos) y rellena los "?"



$\widehat{\Delta sleep} \sim N(?,?)$ $\hat{\beta} \sim N(?,?)$ $t \sim N(0,1)$

PROYECTO DE INVESTIGACIÓN

(hasta un 10% extra en la nota total)

Resumen de la asignatura: con frecuencia los economistas queremos medir un "efecto causal", o sea, el impacto de una "variable de interés" en otra "dependiente", manteniendo todo lo demás constante (e.g. el tamaño de la clase en el aprendizaje, el género en los salarios, el estudio en las notas…). Muchas veces no podemos hacer experimentos (RCTs) donde el "asignar el tratamiento aleatoriamente" nos garantiza que el impacto de los otros factores no varía, i.e. se cumple que $E(u|X) = 0$. En ese caso, la diferencia de medias, que puede estimarse como $\widehat{\beta_1}$ en $Y = \beta_0 + \beta_1 Binaria + u$, se interpreta como un efecto causal. Lo mismo se podría decir de la $\widehat{\beta_1}$ estimada en el modelo $Y = \beta_0 + \beta_1 X + u$.

Si usamos datos observacionales, donde $E(u|X) = 0$ raramente se cumple, $\widehat{\beta_1}$ es una mezcla del efecto de nuestra variable de interés y de factores omitidos (i.e. $\widehat{\beta_1}$ está sesgada por el conocimiento del inglés, la profesión, inteligencia…). La solución es incluir estos factores $X_2 \dots X_k$ en el modelo y estimar $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$ para que $E(u|X_1, X_2, \dots, X_k) = 0$ sea realista y, $\widehat{\beta_1}$ pueda leerse como el efecto causal de X_1 en *Y.*

Informe 1: RESPONDE antes del XXX estas PREGUNTAS en 1 página

1. ¿Cuál es el efecto causal que quieres estimar? ¿Por qué lo consideras interesante?

2. Describe el experimento (RCT) que mediría el efecto causal que te interesa?

3. Tú NO harás experimentos sino que usarás datos observacionales, ¿cuál es tu fuente y el tamaño de tu muestra "n"? Copia y pega un enlace o describe tu recopilación de datos

4. ¿Qué otros factores (además de tu variable de interés) afectan a la variable Y ?

5. ¿Cuáles de ellos estarán correlacionados con nuestra variable de interés sesgando $\widehat{\beta_1}$? ¿Cuáles de estos planeas incluir en tu modelo como "variables de control": i.e., variables que NO incluyes para medir su efecto causal SINO para evitar sesgo en tu la estimación del efecto causal de tu interés?

6. ¿Hay alguna hipótesis interesante que puedas contrastar en el modelo y con tus datos?

Informe 2: para el YYY, ENTREGA 1 o 2 páginas con RESULTADOS

1. Copia el diagrama de dispersión de tus variables dependiente y de interés. Comenta brevemente si hay heterocedasticidad en los datos o visos de efectos no-lineales

2. Recoge en una tabla el modelo univariante estimado que busca captar el efecto causal de interés. Aporta también 3 o 4 regresiones más con variables de *control* y defínelas

3. Haz comentarios sobre los posibles sesgos que anticipaste en tu informe de noviembre

4. Documenta la bondad de ajuste de tus 4 o 5 regresiones, así como cualquier otro asunto que no anticipaste en noviembre pero que has trabajado en tu proyecto: especificaciones no-lineales, factores no significativos, validez interna y externa, sesgo por selección…

5. Y responde: ¿te ha gustado el proyecto? ¿por qué? ¿qué dificultades has encontrado?

EJEMPLO de INFORME 1 para el XXX **Nombre:**

1. ¿Cuál es el efecto causal que quieres estimar? ¿Por qué lo consideras interesante?

El efecto de que haya alumno más en la clase en el apredizaje del grupo. Me interesa porque, si el efecto es grande, los gobiernos pueden invertir en profesores y mejorar la formación.

2. Describe el experimento (RCT) que mediría el efecto causal que te interesa?

Cogería una muestra representative de estudiantes y los dividiría aleatoriamente en grupo "tratado" (aquellos que obligo a ir a clases grandes) y grupo de "control" (aquellos que fuerzo a estudiar en clases pequeñas). Después, compruebo si la nota media en un test estandarizado son significativamente distintas.

3. Tú NO harás experimentos sino que usarás datos observacionales, ¿cuál es tu fuente y el tamaño de tu muestra "n"? Copia y pega un enlace o describe tu recopilación de datos

http://vincentarelbundock.github.io/Rdatasets/datasets.html Item: Caschool n=420

4. ¿Qué otros factores (además de tu variable de interés) afectan a la variable *Y* ?

*El porcentaje de alumnos que no saben inglés (y lo están aprendiendo, **EL_pct**) afectará a las notas, así como el número de ordenadores por alumno (**comp_stu**) y otras medidas de gasto en educación (**expn_stu**). Por supuesto, la renta media del distrito (**avginc**) puede afectar a las notas. Hay otros factores que mi base de datos no incluye: la inteligencia de los estudiantes, la formación de los padres, la calidad de la formación preescolar, etc*

5. ¿Cuáles de ellos estarán correlacionados con nuestra variable de interés sesgando $\widehat{\beta_1}$? ¿Cuáles de estos planeas incluir en tu modelo como "variables de control": i.e., variables que NO incluyes para medir su efecto causal SINO para evitar sesgo en tu la estimación del efecto causal de tu interés?

***EL_pct** se asocial positivamente con **STR** pero baja las notas, por lo que ignorar este factor sesga $\widehat{\beta_1}$ a la baja. Los factores **comp_stu, expn_stu** o **avginc** mejorarán las notas y están negativamente correlacionadas con **STR**, por tanto su omisión sesga $\widehat{\beta_1}$ a la baja. Planeo incluir, al menos, **EL_pct** y **comp_stu** pero no controlaré por factores sobre los que no tengo datos. En cualquier caso, la inteligencia no creo que esté correlacionadas con **STR** (por tanto no creará sesgo) y el resto sólo puede que tengan una ligera correlación negativa con **STR**.*

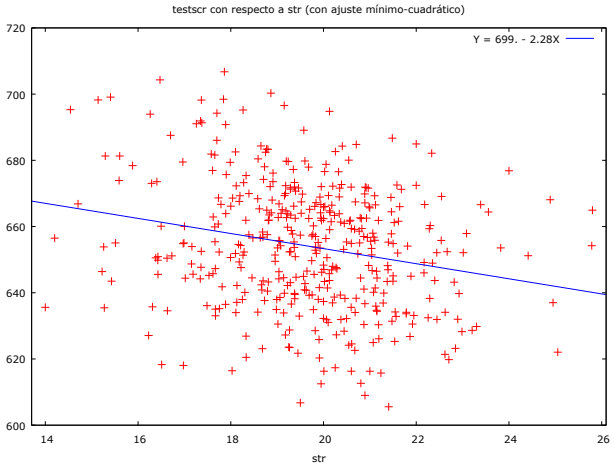
6. ¿Hay alguna hipótesis interesante que puedas contrastar en el modelo y con tus datos?

*Podría contrastar si los recursos en educación (**comp_stu** y **STR**) son relevantes; también si el efecto de **STR** en las notas depende de otras variables: del conocimiento de inglés de los estudiantes, su acceso a ordenadores, el número de matriculados en Primaria, etc.*

EJEMPLO de <u>INFORME 2</u> para el YYY **Nombre:**

1. El objetivo de mi proyecto es medir el impacto del tamaño de la clase (medido por *STR*: el número de alumnos por profesor) en el aprendizaje escolar (medido como la nota media en un examen estandarizado al final de Primaria). Tengo datos de 420 distritos escolares en California donde *STR* está entre 14 y 25 y las notas entre 600 y 710 puntos.

Primero, el **diagrama de dispersión** de notas y *STR* muestra un impacto negativo de *STR* en las notas. En concreto, el modelo de regresión univariante estima que cada alumno extra en la clase baja las notas en 2.28 puntos, de media (ver tabla). Aunque, el efecto es estadísticamente significativo al 1%, *STR* sólo explica el 5% de los cambios en las notas. No hay signos de **heteroscedasticidad** o **no-linealidad** en el efecto de *STR* en las notas.



2. La tabla muestra 5 regresiones que producen un análisis más riguroso del efecto causal

Variable dependiente: nota media en un test estandarizado, i.e. ***test scores*** ;					**n=420**
Regresores	***(A)***	***(B)***	***(C)***	***(D)***	***(E)***
1. ***STR***	-2.28***(.52)	-1.10***(.43)	-0.85**(.43)	-1.36***(.46)	-1.02***(.53)
2. ***EL_pct***		-0.65***(.03)	-0.63***(.03)	-0.69***(.03)	-0.69***(.03)
3. ***comp_stu***			27.27**(12.6)	30.83**(12.5)	31.07**(12.5)
4. ***Henroll***				8.15***(1.6)	41.42**(19.5)
5. ***Henroll×STR***					-1.65***(.95)
0. ***constante***	698.9***(10.4)	686.0***(8.73)	677.1***(9.20)	684.9***(9.47)	678.2***(10.8)
R^2	0.05	0.43	0.43	0.47	0.47
$\bar{R}^2$	0.05	0.43	0.43	0.46	0.46

STR = nº de alumnos por profesor (el promedio para cada distrito)

EL_pct = porcentaje de alumnos que están aprendiendo inglés (promedio en cada distrito)

comp_stu = nº de ordenadores por alumno (el promedio para cada distrito)

Henroll es binaria: =1 si el distrito tiene un número de alumnos de Primaria por encima de la media, y =0, en otro caso

Henroll×STR es un factor de interacción que capta la diferencia entre el efecto de *STR* en las notas cuando *Henroll=1* y su efecto cuando *Henroll=0*

3. La **regresión A** es el modelo univariante que ya he comentado.

La **regresión B** tiene en cuenta ***EL_pct***. Veo que el efecto estimado en la regresión **A** estaba sesgado a la baja pues captaba dos efectos negativos: el efecto "puro" de las clases grandes (de media, 1.10 puntos menos por cada alumno extra en el aula) y el impacto de que haya más alumnos aprendiendo inglés (de media, 0.65 puntos menos por punto porcentual). Ambos factores están positivamente correlacionados. El R^2 crece notablemente indicando que la bondad de ajuste mejora.

La **regresión C** añade el número (medio) de ordenadores por alumno como variable de control. *comp_stu* tiene un efecto positivo en *test scores* (27.27 puntos más por cada ordenador extra por alumno, de media) y está negativamente correlacionado con *STR*, por tanto omitir este factor en el modelo **B** genera sesgo a la baja en mi estimación del efecto causal. El R^2 sube muy poco pues la mayor parte de la información de *comp_stu* ya estaba incluida en *STR* y *EL_pct*, pero aun así, el R^2 corregido sube de 0.42 en **B** a 0.43 en **C**.

La **regresión D** añade la variable binaria *Henroll* que distingue los distritos con muchos alumnos en Primaria (ciudades grandes, áreas densamente pobladas) de los distritos con menos alumnos que la media. Al incorporar al modelo este factor omitido el efecto de *STR* en *test scores* estimado baja, por tanto la estimación anterior estaba sesgada al alza. Esto significa que los distritos con muchos estudiantes (*Henroll=1*) son habitualmente aquellos con mayor *STR* (lógico). El R^2 y R^2 corregido siguen subiendo por lo que **D** explica mejor los datos sobre *test scores* que **C**.

Por último, dado el pequeño cambio en el efecto estimado de *STR* en *test scores* al pasar de **C** a **D**, **l**a **regresión E** investiga posibles efectos no-lineales de *STR* en *test scores.* En concreto, estudia si este efecto es distinto en distritos con muchos estudiantes (*Henroll=1*) y en distritos con pocos alumnos. Para ello, añado el factor de interacción *STR*Henroll* cuyo coeficiente es -1.65. Esto significa que, cuando el distrito tiene muchos estudiantes, *STR* baja *test scores* 1.65 puntos por alumno extra más que si el distrito tuviera pocos alumnos. Este coeficiente es significativo al 10%. Por tanto, el efecto estimado para *STR* en *test scores* cuando un distrito tiene pocos estudiantes es de -1 punto por alumno extra por profesor. Pero, si el distrito tiene muchos estudiantes, entonces el efecto estimado es de -2.65 puntos por alumno extra (quizá sean distritos colapsados, disfuncionales). El R^2 y R^2 corregido suben tan poco que no se aprecian diferencias al redondear a decimales.

4. Dado el pequeño cambio en R^2 corregido entre los modelos **C** y **E**, vale la pena contrastar en **E** si el nivel de estudiantes importa; H_0: $\boldsymbol{\beta_4 = \beta_5 = 0}$ vs H_1: $\beta_4 \neq 0, o\ \beta_5 \neq 0, o\ ambos$. Estimo F=13.85 (>3) por tanto rechazo y concluyo que el número de alumnos es relevante. Puedo contrastar también en **E** si los factores que dependen de los recursos dedicados a la educación (sobretodo, profesores y ordenadores, *STR y comp_stu*) importan; H_0: $\boldsymbol{\beta_1 = \beta_3 = \beta_5 = 0}$ vs H_1: $alguno \neq 0$. Estimo F=8.28 (>2.6) y rechazo: los recursos importan. El modelo **E** llega a replicar un 47% de la variabilidad en las notas: ¡no está mal!

5. Pienso que trabajar con datos, interpretar las regresiones y encontrarles el sentido es **interesante**. ¡Espero que a ti también te haya gustado!

IN ENGLISH

List of BASIC RESULTS shown in the APPENDIX for the MIDTERM EXAM

(1) $E(aX + c) = a\mu_X + c$

(2) $var(X) = E[X^2] - \mu_X^2$

(3) $var(aX + c) = a^2\sigma_X^2$

(4) $E(aX + bY + c) = a\mu_X + b\mu_Y + c$

(5) $cov(X, Y) = E(XY) - \mu_X\mu_Y$

(6) $cov(X, X) = \sigma_X^2$

(7) $cov(aX, c + bY) = ab\ \sigma_{XY}$

(8) $cov(aX + bY + c, Z) = a\ \sigma_{XZ} + b\ \sigma_{YZ}$

(9) $var(aX + bY + c) = a^2\ var(X) + b^2\ var(Y) + 2ab \cdot cov(X, Y)$

$\equiv a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\ \sigma_{XY}$

(10) $E(aX + bY + c|Z) = aE(X|Z) + bE(Y|Z) + c$

(11) $Var(aX + bY + c|Z) = a^2\ var(X|Z) + b^2\ var(Y|Z) + 2ab \cdot cov(X, Y|Z)$

(12) $E[E(X|Y)] = E(X) \equiv \mu_X$ Law of Iterated Expectations(LIE)

(13) $P(X = x_i \cap Y = y_j) = P(X = x_i) \cdot P(Y = y_j)\ \forall\ x_i, y_j \leftrightarrow X$ e Y independent

(14) $X\ \&\ Y\ \ independent\ \ \rightarrow\ \ corr(X, Y) = 0$

List of PROOFS from LECTURES, PROBLEM SETS.... before the MIDTERM EXAM

(I) $E(\bar{Y}) = \mu_Y$ for a sample $\{Y_1, Y_2, \dots, Y_n\}$ i.i.d.

(II) $S_X^2 = \left(\frac{n}{n-1}\right)\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2\right)$ is the variance of the sample $\{X_1, X_2, \dots, X_n\}$

(III) $S_{XY} = \left(\frac{n}{n-1}\right)\left(\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\,\bar{y}\right)$ for the sample $\{(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)\}$

(IV) solve the OLS problem to estimate $Y_i = \beta_0 + \beta_1 X_i + u_i$ when $x_i = 0\ \forall\ i$ and show that then $\hat{\beta}_0 = \bar{y}$

(V) solve the OLS problem to estimate $Y_i = \beta_0 + \beta_1 X_i + u_i$ and show that then $\sum_{i=1}^n \hat{u}_i = 0$, & $\sum_{i=1}^n \hat{u}_i x_i = 0$, as well as $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$, & $\widehat{\beta_1} = S_{XY}/S_X^2$

(VI) $S_{\hat{u}X} = 0$ when you estimate $Y_i = \beta_0 + \beta_1 X_i + u_i$ using OLS

(VII) if $E(u|X) = constant$, then $corr(u, X) = 0$

(VIII) if $E(u|X) = 0$, then $\beta_1 = \frac{\Delta E(Y|X)}{\Delta X}$ in the model $Y_i = \beta_0 + \beta_1 X_i + u_i$ ...

(IX) ... but $\beta_1 = E(Y|D = 1) - E(Y|D = 0)$ in $Y_i = \beta_0 + \beta_1 D_i + u_i$ if D is dummy

(X) $TSS = ESS + SSR$ when you estimate $Y_i = \beta_0 + \beta_1 X_i + u_i$ using OLS

Moments / Parameters in the population	**Sample-estimated** moments / parameters
$E(X) = \sum_{j=1}^{m} x_j\, P(X = x_j) \equiv \mu_X$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \hat{\mu}_X$
$var(X) = E[(X-\mu_X)^2] \equiv \sigma_X^2$	$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \equiv \hat{\sigma}_X^2$
$sd(X) = \sqrt{\sigma_X^2} \equiv \sigma_X$	$s_X = \sqrt{s_X^2} \equiv \hat{\sigma}_X$
$cov(X,Y) =$ $= E[(X-\mu_X)(Y-\mu_Y)] \equiv \sigma_{XY}$	$s_{XY} =$ $= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \equiv \hat{\sigma}_{XY}$
$corr(X,Y) = {}^{\sigma_{XY}}/_{\sigma_X\sigma_Y} \equiv \rho_{XY}$	$r_{XY} = {}^{s_{XY}}/_{s_X s_Y} \equiv \hat{\rho}_{XY}$
$Y = \beta_0 + \beta_1 X_1 + u \quad \rightarrow \hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 \;;\; \hat{u} = Y - \hat{Y}$ $\{\, Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u \quad \rightarrow \hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_k} X_k \;;\; \hat{u} = Y - \hat{Y} \,\}$	
$\beta_1 = \frac{\Delta E(Y\,\vert X_1)}{\Delta X_1} \left\{= \frac{\Delta E(Y\,\vert X_1, X_2)}{\Delta X_1}\right\}$	$\widehat{\beta_1} = \frac{s_{YX}}{s_X^2} \left\{= \frac{s_{Y1}s_{22} - s_{12}s_{Y2}}{s_{11}s_{22} - s_{12}^2}\right\}$
$\beta_0 = E(Y\,\vert X = 0)$ $\{= E(Y\,\vert X_1 = 0, X_2 = 0)\}$	$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$ $\{= \overline{Y} - \widehat{\beta_1}\overline{X}_1 - \widehat{\beta_2}\overline{X}_2\}$
|||||||||||||||||||||||||||||||||||	|||||||||||||||||||||||||||||||||||
$E(\overline{X}) = \mu_X \equiv \mu_{\overline{X}}$	
$var(\overline{X}) = \frac{1}{n}\sigma_X^2 \equiv \sigma_{\overline{X}}^2$	$\hat{\sigma}_{\overline{X}}^2 = \frac{1}{n}\hat{\sigma}_X^2 \equiv s_{\overline{X}}^2$
$std(\overline{X}) = \frac{1}{\sqrt{n}}\sigma_X \equiv \sigma_{\overline{X}}$	$\hat{\sigma}_{\overline{X}} = \frac{1}{\sqrt{n}}\hat{\sigma}_X \equiv SE(\overline{X})$
$E(\widehat{\beta_1}) \cong \beta_1 + \rho_{Xu}\, {}^{\sigma_u}/_{\sigma_X}$	
$var(\widehat{\beta_1}) = \frac{1}{n}\frac{var([X-\mu_X]u)}{[var(X)]^2} \equiv \sigma_{\widehat{\beta_1}}^2$	$\hat{\sigma}_{\widehat{\beta_1}}^2 = \frac{1}{n}\frac{\sum_{1=1}^{n}\hat{u}_i{}^2(x_i-\overline{x})^2/(n-2)}{[\sum_{1=1}^{n}(x_i-\overline{x})^2/n]^2}$
$sd(\widehat{\beta_1}) = \sqrt{\sigma_{\widehat{\beta_1}}^2} \equiv \sigma_{\widehat{\beta_1}}$	$\hat{\sigma}_{\widehat{\beta_1}} = \sqrt{\hat{\sigma}_{\widehat{\beta_1}}^2} \equiv SE(\widehat{\beta_1})$
|||||||| $R^2 = \frac{ESS}{TSS} = \frac{\sum_{1=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}{\sum_{1=1}^{n}(y_i-\bar{y})^2}$;	$SER = \sqrt{\sum_{i=1}^{n}\hat{u}_i^2/(n-2)}$ ||||||||

The values -2.575, -1.96 & -1.64 leave 0.5, 2.5 & 5% probability in the left-tail of the N(0,1)

APPENDIX: easy proofs+summary table

Here you can find proofs for several properties about random variables frequently used in the course. This is important material of the course. Most of them have been extracted from the reference book by Stock & Watson. For a summary of concepts and notation, see table in last page

Let X be a (discrete) random variable that takes values { $X=x_1$, $X=x_2$, …, $X=x_k$ } with probabilities { $P(X=x_1)$, $P(X=x_2)$, … , $P(X=x_k)$ } such that, by definition, { $P(X=x_1)+P(X=x_2)+\ldots+P(X=x_k) = \sum_{i=1}^{k} P(X=x_i) = 1$ }, then

(A) $E(X) = \sum_{i=1}^{k} x_i P(X=x_i) \equiv \mu_X$ Mean or Expected value

(B) $var(X) = E[(X-\mu_X)^2] \underset{A}{=} \sum_{i=1}^{k} (x_i-\mu_X)^2 P(X=x_i) \equiv \sigma_X^2$ Variance

(C) $sd(X) = \sqrt{\sigma_X^2} \equiv \sigma_X$ Standard deviation

Let a, b & c be constants (b will be used later), then

(1) $E(aX+c) \underset{A}{=} \sum_{i=1}^{k} (ax_i+c) P(X=x_i)$

$= \sum_{i=1}^{k} ax_i P(X=x_i) + \sum_{i=1}^{k} cP(X=x_i)$

$= a\sum_{i=1}^{k} x_i P(X=x_i) + c\sum_{i=1}^{k} P(X=x_i)$

$\underset{A}{=} aE(X) + c \equiv a\mu_X + c$

(2) $var(X) \underset{B}{=} E[(X-\mu_X)^2]$

$= E[X^2 + (\mu_X)^2 - 2\mu_X X]$

$\underset{1}{=} E[X^2] + \mu_X^2 - 2\mu_X E[X]$

$= E[X^2] - \mu_X^2$

(3) $var(aX + c)$ $\overset{\overset{=}{=}}{B1}$ $E\{[(aX + c) - (a\mu_X + c)]^2\}$

$= E\{[a(X - \mu_X)]^2\}$

$\overset{=}{1}$ $a^2\, E[(X - \mu_X)^2]$

$\overset{=}{B}$ $a^2\, var(X) \equiv a^2\sigma_X^2$

Let Y be another (discrete) random variable that takes values { $Y=y_1$, $Y=y_2$, …, $Y=y_m$}. There is a joint distribution of X & Y with probabilities

{ $P(x_1, y_1)$, $P(x_1, y_2)$, …, $P(x_1, y_j)$,…, $P(x_1, y_m)$;

$P(x_2, y_1)$, $P(x_2, y_2)$, …, $P(x_2, y_j)$,…, $P(x_2, y_m)$;

… ,

$P(x_i, y_1)$, $P(x_i, y_2)$, …, $P(x_i, y_j)$,…, $P(x_i, y_m)$; … ,

$P(x_k, y_1)$, $P(x_k, y_2)$, …, $P(x_k, y_j)$,…, $P(x_k, y_m)$ }

where, in general, $P(x_i, y_j)$ is short notation for $P(X = x_i \cap Y = y_j)$, and the sum of probabilities for all combinations of X & Y equals 1.

Also, we define individual probability distributions of X &Y as

(D) $P(X = x_i) = \sum_{j=1}^{m} P(X = x_i \cap Y = y_j)$

and $P(Y = y_j) = \sum_{i=1}^{k} P(X = x_i \cap Y = y_j)$

, in other words, as marginal probabilities , then we already know that

(A') $E(Y) = \sum_{j=1}^{m} y_j P(Y = y_j) \equiv \mu_Y$ as for X !

(B') $var(Y) = E[(Y - \mu_Y)^2] = E[Y^2] - \mu_Y^2 \equiv \sigma_Y^2$ as for X !

(C') $sd(Y) = \sqrt{\sigma_Y^2} \equiv \sigma_Y$ as for X !

(4) $E(aX+bY+c) \underset{A}{=} \sum_{i=1}^{k}\sum_{j=1}^{m}(ax_i+by_j+c)P(X=x_i \cap Y=y_j)$

$= \sum_i \sum_j ax_i P(x_i,y_j) + \sum_i \sum_j by_j P(x_i,y_j) + \sum_i \sum_j cP(x_i,y_j)$

$= a\sum_{i=1}^{k} x_i \sum_{j=1}^{m} P(x_i,y_j) + b\sum_{j=1}^{m} y_j \sum_{i=1}^{k} P(x_i,y_j) + c$

$\underset{D}{=} a\sum_{i=1}^{k} x_i P(X=x_i) + b\sum_{j=1}^{m} y_j P(Y=y_j) + c$

$\underset{A}{=} aE[X] + bE[Y] + c \equiv a\mu_X + b\mu_Y + c$

and, in general, if instead of two random variables X & Y we have random variables $\{X_1, X_2, \ldots, X_s, \ldots, X_S\}$ and constants $\{a_1, a_2, \ldots, a_s, \ldots, a_S, c\}$

(4) $E(c+\sum_{s=1}^{S} a_s X_s) = c + \sum_{s=1}^{S} a_s E(X_s) \equiv c + \sum_{s=1}^{S} a_s \mu_{Xs}$

Let´s define

(E) $cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] \equiv \sigma_{XY}$

then

(5) $cov(X,Y) \quad \underset{E}{=} E[(X-\mu_X)(Y-\mu_Y)] =$

$= E(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y)$

$\underset{4}{=} E(XY) - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y$

$= E(XY) - \mu_X\mu_Y$

(6) $cov(X,X) \quad \underset{E}{=} E[(X-\mu_X)(X-\mu_X)]$

$\underset{B}{=} var(X) \equiv \sigma_X^2$

(7) $cov(aX, c+bY) \underset{E}{=}\underset{4}{=} E[(aX-a\mu_X)(c+bY-c-b\mu_Y)]$ or use (5)!

$= E[ab(X-\mu_X)(Y-\mu_Y)]$

$\underset{4}{=} ab\, E[(X-\mu_X)(Y-\mu_Y)]$

$\underset{E}{=} ab\, cov(X,Y) \equiv ab\, \sigma_{XY}$

(8) $cov(aX+bY+c,Z) \overset{=}{7} cov(aX+bY,Z)$

$\overset{=}{E}\ \overset{=}{4}\ E[\{(aX+bY)-(a\mu_X+b\mu_Y)\}\{Z-\mu_Z\}]$ or use (5)!

$= E[\{a(X-\mu_X)+b(Y-\mu_Y)\}\{Z-\mu_Z\}]$

$\overset{=}{4}\ aE[(X-\mu_X)(Z-\mu_Z)]+bE[(Y-\mu_Y)(Z-\mu_Z)]$

$\overset{=}{E}\ a\ cov(X,Z)+b\ cov(Y,Z) \equiv a\ \sigma_{XZ}+b\ \sigma_{YZ}$

and, in general, if we have random variables $\{X_1, X_2, ..., X_s, ..., X_S\}$ & $\{Y_1, Y_2, ..., Y_t, ..., Y_T\}$ and constants $\{a_1, a_2, ..., a_s, ..., a_S; b_1, b_2, ..., b_t, ..., b_T; c\}$

(8) $cov(\sum_s a_s X_s, c+\sum_t b_t Y_t) = \sum_s \sum_t a_s b_t cov(X_s, Y_t) \equiv \sum_s \sum_t a_s b_t \sigma_{XsYt}$

(9) $var(aX+bY+c) \overset{=}{3} var(aX+bY)$

$\overset{=}{B}\overset{=}{4}\ \mathrm{E}[\{(aX+bY)-(a\mu_X+b\mu_Y)\}^2]$ or use (2)!

$= \mathrm{E}[\{a(X-\mu_X)+b(Y-\mu_Y)\}^2]$

$= \mathrm{E}[a^2(X-\mu_X)^2+b^2(Y-\mu_Y)^2+2ab(X-\mu_X)(Y-\mu_Y)]$

$\overset{=}{4}\ a^2\mathrm{E}[(X-\mu_X)^2]+b^2E[(Y-\mu_Y)^2]+2abE[(X-\mu_X)(Y-\mu_Y)]$

$\overset{=}{B}\overset{=}{E}\ a^2var(X)+b^2var(Y)+2ab\cdot cov(X,Y)$

$\equiv a^2\sigma_X^2+b^2\sigma_Y^2+2ab\ \sigma_{XY}$

and, in general, if we have random variables $\{X_1, X_2, ..., X_s, ..., X_S\}$ and constants $\{a_1, a_2, ..., a_s, ..., a_S; c\}$

(9) $var(\sum_s a_s X_s + c) = \sum_s a_s^2\ var(X_s) + 2\sum_s \sum_{t\neq s} a_s\ a_t\ cov(X_s, X_t)$

$\equiv \sum_s a_s^2\ \sigma_{Xs}^2 + 2\sum_s \sum_{t\neq s} a_s\ a_t\ \sigma_{XsXt}$

Bayes' Theorem says that the conditional probability is

(F) $P(X=x_i|Y=y_j) = P(X=x_i \cap Y=y_j) \Big/ P(Y=y_j)$

so

(G) $E(X|Y=y_j) = \sum_{i=1}^{k} x_i P(X=x_i|Y=y_j)$ or simply $E(X|Y)$!

(H) $Var(X|Y=y_j) = E\{(X-\mu_X)^2|Y=y_j\}$ or simply $Var(X|Y)$!

$\overset{=}{G} \sum_{i=1}^{k}(x_i-\mu_X)^2\, P(X=x_i|Y=y_j)$

Thus, if we consider an extra random variable Z,

(10) $E(aX+bY+c|Z) \overset{=}{4} \overset{=}{G}\ aE(X|Z)+bE(Y|Z)+c$

(11) $Var(aX+bY+c|Z) \overset{=}{9} \overset{=}{H}$

$a^2\ var(X|Z)+b^2\ var(Y|Z)+2ab\cdot cov(X,Y|Z)$

and, in general,

(10) $E(c+\sum_{s=1}^{S} a_s X_s\,|Z) \overset{=}{4} \overset{=}{G}\ c+\sum_{s=1}^{S} a_s E(X_s|Z)$

(11) $Var(\sum_s a_s X_s + c|Z) \overset{=}{9} \overset{=}{H}$

$\sum_s a_s^2\ var(X_s|Z)+2\sum_s\sum_{t\neq s} a_s\ a_t\ \ cov(X_s,X_t|Z)$

where details were omitted as these proofs are very similar to (4) & (9)

(12) $E[E(X|Y)] = E\big[E(X|Y=y_j)\big] \overset{=}{A} \sum_{j=1}^{m} E(X|Y=y_j)P(Y=y_j)$

$\overset{=}{G} \sum_{j=1}^{m}\big[\sum_{i=1}^{k} x_i P(X=x_i|Y=y_j)\big]P(Y=y_j)$

$\overset{=}{F} \sum_{j=1}^{m}\sum_{i=1}^{k} x_i\, P(X=x_i\cap Y=y_j)$

$= \sum_{i=1}^{k}\sum_{j=1}^{m} x_i P(X=x_i\cap Y=y_j)$

$= \sum_{i=1}^{k} x_i\big[\sum_{j=1}^{m} P(X=x_i\cap Y=y_j)\big]$

$\overset{=}{D} \sum_{i=1}^{k} x_i\, P(X=x_i)$

$\overset{=}{A} E(X)\equiv\mu_X$ Law of Iterated Expectations (LIE)

The random variables X & Y are independent iff

(I) $P(X=x_i|Y=y_j) = P(X=x_i)\ \ \forall\, x_i, y_j$

Then

(13) $P(X=x_i\cap Y=y_j) \overset{=}{F}\overset{=}{I} P(X=x_i)\cdot P(Y=y_j)\ \forall\, x_i, y_j$

Finally, remember

(J) $corr(X,Y) = {}^{\sigma_{XY}}/_{\sigma_X \sigma_Y} \equiv \rho_{XY}$

Thus,

(14) $If\ X\ \&\ Y\ independent\ ,\ then\ corr(X,Y) = 0$

X & Y are random variables (not constants), therefore $\sigma_X > 0\ \&\ \sigma_Y > 0$.

Thus, $corr(X,Y) = 0$ iff $cov(X,Y) = 0$ or, by (5), $E(XY) = E(X)E(Y)$.

$$E(XY) \underset{A}{=} \sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j P\left(X = x_i \cap Y = y_j\right)$$

$$\underset{I}{=} \sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j P(X = x_i)\, P\left(Y = y_j\right)$$

$$= \sum_{i=1}^{k} x_i P(X = x_i) \sum_{j=1}^{m} y_j P\left(Y = y_j\right)$$

$$\underset{A}{=} E(X)E(Y)$$

You should understand the distinction between the moments of a random variable and their estimates (with a hat ^)	
Moments of random variable X	Sample-estimated moments
$E(X) = \sum_{i=1}^{k} x_i \, P(X = x_i) \equiv \mu_X$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \hat{\mu}_X$
$var(X) = E[(X - \mu_X)^2] \equiv \sigma_X^2$	$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \equiv \hat{\sigma}_X^2$
$sd(X) = \sqrt{\sigma_X^2} \equiv \sigma_X$	$s_X = \sqrt{s_X^2} \equiv \hat{\sigma}_X$
$cov(X,Y) =$ $= E[(X - \mu_X)(Y - \mu_Y)] \equiv \sigma_{XY}$	$s_{XY} =$ $= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \equiv \hat{\sigma}_{XY}$
$corr(X,Y) = {\sigma_{XY}}/{\sigma_X \sigma_Y} \equiv \rho_{XY}$	$r_{XY} = {s_{XY}}/{s_X s_Y} \equiv \hat{\rho}_{XY}$
Moreover, the statistics on the right are also functions of the data, thus random variables whose realization depends on the sample. For example, X is the height of Europeans and we want to estimate μ_X, the unknown *mean* or *expected value* of the height of Europeans, E(X). Juan selects randomly a representative sample of 500 Spaniards and obtains $\overline{x} = 1.67$: this is his estimate of μ_X ($\hat{\mu}_X$). But Giampiero takes another sample of 500 Italians and gets $\overline{x} = 1.65$. And Catiana draws a representative sample of 500 Germans that gives $\overline{x} = 1.70$. Thus, {1.65, 1.67, 1.70} are realizations of the random variable $\overline{X}$, the average height of a representative sample of 500 Europeans. This random variable $\overline{X}$ also has moments: mean, variance, etc; that we can estimate with data.	
$E(\overline{X}) = \mu_X \equiv \mu_{\overline{X}}$	$\hat{\mu}_{\overline{X}} = \overline{x}$
$var(\overline{X}) = \frac{1}{n}\sigma_X^2 \equiv \sigma_{\overline{X}}^2$	$\hat{\sigma}_{\overline{X}}^2 = \frac{1}{n}\hat{\sigma}_X^2 \equiv s_{\overline{X}}^2$
$sd(\overline{X}) = \sqrt{\frac{1}{n}\sigma_X^2} \equiv \sigma_{\overline{X}}$	$\hat{\sigma}_{\overline{X}} = \sqrt{\frac{1}{n}\hat{\sigma}_X^2} \equiv SE(\overline{X})$

MOCK 1 of MIDTERM EXAM

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. (2 points) You will be asked to ***prove*** *ONE from the last page*

either result *XXX* from the Appendix *using algebra & any of the other results listed before*

or the proof *YYY* that we did in class, exercises…

2. (3.5 points) *The regional government would like to know whether teaching in English courses like Arts, Music or Natural Sciences during primary school has a negative effect in how much the students learn about the subject.*

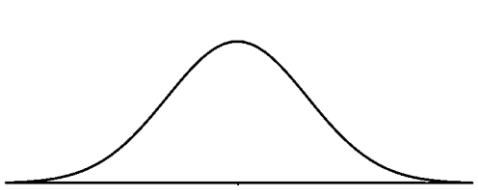
a. Describe how to run an experiment that gives an answer to the government. Explain it briefly and state explicitly which key feature will assure that it will measure the effect of teaching in English and not that of any other factor.

Instead of running an experiment, 500 randomly selected students of 3rd year of primary school did a standardized test about these subjects in June 2017 (including both students in English and in Spanish). The average for all students was 81 (out of 100) points and the standard deviation 15.2 points.

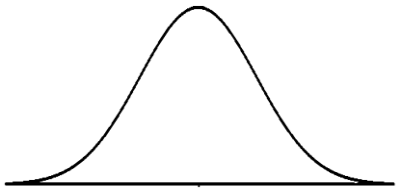
b. Construct a 95% confidence interval for the mean test score in the population. Can we say with at least that confidence that students, on average, pass the test ($\mu \geq 50 points$)?

In the sample, 208 students had taken these courses in English while the other 292 studied in Spanish. The average score for the former is 79.8 points (with a standard deviation of 14.5 points) and 81.9 points for the latter (with 16.7 points as standard deviation).

c. Is there statistically significant evidence (at 5%) supporting that students in Spanish and in English learn equally ($\Delta=0$)? Illustrate with the figures below your answer.



$\hat{\Delta} \sim N(?,?)$



$t \sim N(0,1)$

d. Draw also in both figures minimum significance level that we need to reject $\Delta=0$

3. (4.5 points) *We would like to study the effect of academic education (ED, measured in years successfully passed) on the salaries of workers (AHE, average hourly earnings in dollars). For this purpose, we collect a random sample of 2829 full-time workers aged 29-30 and estimate (using heteroskedasticity-robust standard errors):*

$$(1) \qquad \widehat{AHE} = -7.29 + 1.93\,ED, \qquad R^2 = 0.16, \qquad SER = 10.29$$

$$(1.1) \quad (0.08)$$

a. Interpret -7.29 & 1.93 in terms of years of education and dollars earned

b. Give the 95% confidence interval for the expected effect on *AHE* of ΔED=2 years

c. Which would be the predicted earnings for someone with 30 years of education? Do you consider this estimate reliable? Why or why not? Provide statistical arguments

Actually, obtaining a bachelor degree requires a minimum of 16 years of education. We construct a dummy *variable that takes the value 1 if ED≥16 years (and 0 otherwise) and we estimate:* **(2)** MCO, OLS, using observations 1-2829

Dependent variable: **ahe**
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	16.5009	0.209406	78.7987	<0.0001	***
dummy	8.57931	0.435689	19.6913	<0.0001	***

Mean dependent var	19.83983	S.D. dependent var	11.23823
Sum squared resid	307670.5	S.E. of regression	10.43230
R-squared	0.138589	Adjusted R-squared	0.138284
F(1, 2827)	387.7491	P-value (F)	5.38e-81

d. What is the difference in earnings predicted by crossing the 16-years threshold? Is the latter difference statistically different from 0 at the 10, 5 and 1% significance levels?

e. Would the SER change if we do not ask for *heteroskedasticity-robust standard errors?*

Often students with many years of education have a high knowledge of foreign languages *(but this is not the result of time devoted to study; instead it is due to trips, motivation…)*

f. Explain the implications of omitting the variable *foreign language knowledge* in our regression (1). Use a mathematical formula to justify your answer very clearly.

g. Draw two graphs: (i) one showing which model assumption fails and (ii) another with the consequences in using our sample to do inference about the causal effect of *ED* in *AHE*.

MOCK 2 of MIDTERM EXAM

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. (2 points) You will be asked to ***prove*** *ONE from the last page*

either result *XXX* from the Appendix *using algebra & any of the other results listed before*

or the proof *YYY* that we did in class, exercises…

2. (3.5 points) *The government of Belgium would like to know the effect of a plan fostering the use of new technologies in Primary school on students learning. They could measure their education or "learning" based on a standardized test.*

a. Write the key feature of an experiment (RCT) that measures this effect (not another)

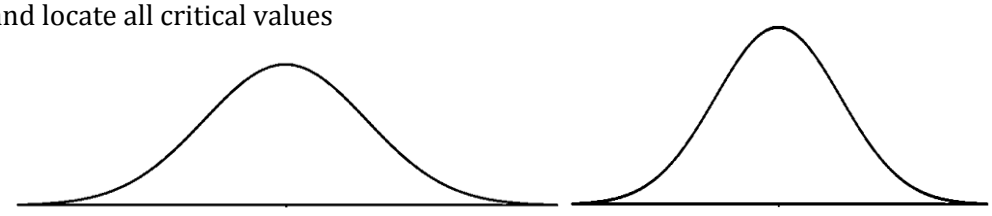
Instead of running an experiment, we use observational data on a standardized test for 500 randomly selected students from schools that had already implemented (or not) a similar plan. The average score for all students was 720 (out of 800) points and the standard deviation 227 points.

b. Compute a 99% confidence interval for the mean test score in Belgium.

c. Do we reject using a 1% significance level that the mean test score is 685 points? Explain briefly why, if that were the case, we would obviously reject using a 5% level

In the sample, 200 students belong to schools that had already implemented a similar plan while 300 study in schools without plan. The average for the former students is 721 points (with standard deviation of 230 points) and, for the latter, it is 711points (with 222 points for standard deviation)

d. Do we reject that mean test scores are equal for those studying in schools with a plan and without it using a 5% significance level? Illustrate your answer below, fill in the 3 "?" and locate all critical values



$? \sim N(?, ?)$ $t \sim N(0,1)$

e. Write a single-regressor model and the hypothesis that, run and tested with our data, would yield the exact same answer as the test in (**d**)

3. (4.5 points) *We would like to study the effect of* height *(measured in inches) on* income *(in $ per year). For this purpose, we collect data from a sample of 7896 male workers and estimate the following (pretty bad, see the R^2) model:*

$$(\mathbf{1}) \quad \widehat{income} = -43130 + 1306.9\, height, \qquad R^2 = 0.02, \qquad SER = 26671$$

(6925) (98.86) *heteroskedasticity-robust standard errors*

a. Interpret -43130 (*the number, not only the sign!*) in terms of inches and dollars

b. Rewrite regression (1) so that the variable *height* is measured in cm (1cm=2.54inches)

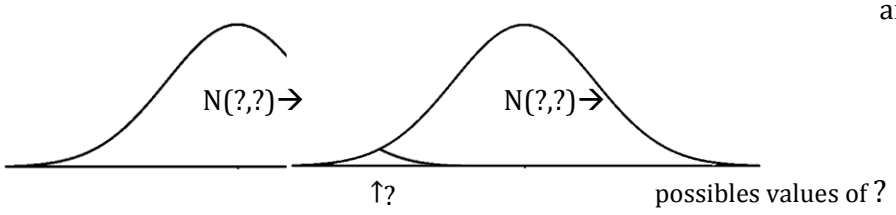
We estimate again our model with data of height *in inches for 9974 female workers:*

(2) OLS, using observations 1-9974
Dependent variable: **income**
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	12650.9	6299.15	2.0083	0.0446	**
height	511.222	97.5846	5.2388	<0.0001	***

Mean dependent var.	45621.00	S.D. dependent var	26835.43
Sum squared resid	7.16e+12	S.E. of regression	26800.90
R-squared	0.002672	Adjusted R-squared	0.002572
F(1, 9972)	27.44459	P-value (F)	1.65e-07

c. Construct a 95% confidence interval for the difference in the effect of *height* on *income* between male and female workers. Illustrate your answer in this graph as we did in class and fill in the 6 "?"



d. Would the confidence interval in (c) change if we had NOT asked for *heteroscedasticity-robust standard errors* and there IS heteroscedasticity in the data? Explain very briefly

High income is very often the result of returns on heritages, that is, on inherited wealth *that, due to its impact on nutrition during childhood, it is positively associated with* height

e. Explain the effect of omitting the factor inherited wealth in regression (1). Use a mathematical formula to justify and discipline your reasoning

f. Illustrate in two graphs: (i) which assumption of our model fails and (ii) which are the consequences when we use our sample to estimate the effect of *height* on *income*

MOCK 2b of MIDTERM EXAM

(if you understood the solution to Mock 2a, you must be able to answer this mock)

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. (2 points) You will be asked to ***prove*** *ONE from the last page*

either result *XXX* from the Appendix *using algebra & any of the other results listed before*

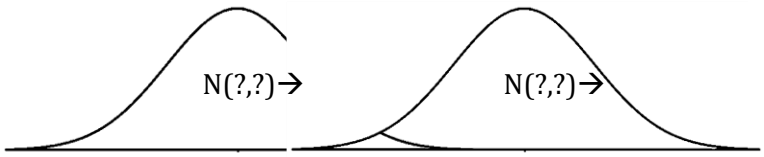
or the proof *YYY* that we did in class, exercises…

2. (3.5 points) *The UK government would like to know the effect of separating boys and girls in Primary school on their education ("single-sex" in contrast to "co-education"). They could measure their education or "learning" based on a standardized test.*

a. Write the key feature of an experiment (RCT) that measures this effect (not another)

Instead of running an experiment, we use observational data for 400 randomly selected students (including both types of education) that did a standardized test. The average score for all students was 720 (out of 800) points and the standard deviation 227 points.

b. Construct a 95% confidence interval for the mean of UK test scores. Illustrate your answer in this picture as we did in class and fill-in the 6 "?"



c. Do we reject using a 5% significance level that the mean UK test score is 695 points? Explain briefly why, if that were the case, we would obviously reject using a 10% level

In the sample, 150 students belong to "single-sex" schools while 250 receive "co-education". The average for the former students is 723 points (with standard deviation of 220 points) and, for the latter, it is 718.2 points (with 231 points for standard deviation)

d. Do we reject that mean test scores are equal for "single-sex" or "co-education" students using a 5% significance level?

e. Write a single-regressor model and the hypothesis that, run and tested with our data, would yield the exact same answer as the test in (d)

3. (4.5 points) *We would like to study the effect of* height *(measured in inches) on* income *(in $ per year). For this purpose, we collect data from a sample of 7896 male workers and estimate the following (pretty bad, see the R^2) model:*

$$(\mathbf{1}) \quad \widehat{income} = -43130 + 1306.9\, height, \qquad R^2 = 0.02, \qquad SER = 26671$$

(6925) (98.86) *heteroskedasticity-robust standard errors*

a. Interpret 1306.9 (*the number, not only the sign!*) in terms of inches and dollars

b. Can we use our model to estimate income for Amancio Ortega or Bill Gates? Explain

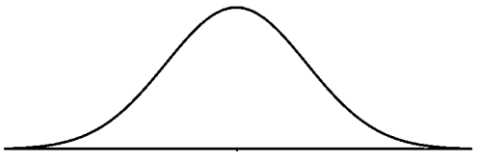
We estimate again our model with data for 9974 female workers:

Model 2: OLS, using observations 1-9974
Dependent variable: **income**
Heteroskedasticity-robust standard errors, variant HC1

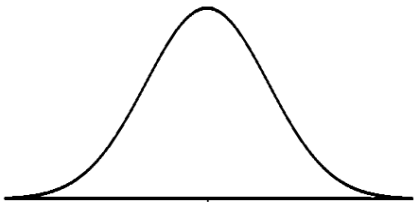
	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	12650.9	6299.15	2.0083	0.0446	**
height	511.222	97.5846	5.2388	<0.0001	***

Mean dependent var.	45621.00	S.D. dependent var	26835.43
Sum squared resid	7.16e+12	S.E. of regression	26800.90
R-squared	0.002672	Adjusted R-squared	0.002572
F(1, 9972)	27.44459	P-value (F)	1.65e-07

c. Test with a 5% signif. level whether the effect of height on income is equal for male and female workers. Illustrate your test using the two figures below, mark their critical values, and fill in the 3 '?'



? ~ N(?, ?)



t~N(0,1)

d. Could our answer to the test in (c) change if we had not asked for *heteroscedasticity-robust standard errors* and there is heteroscedasticity in the data? Explain very briefly

Low income is often linked to jobs where physical strength *is important, therefore the opposite (* weakness *) is associated with high income and negatively correlated with* height

e. Explain the effect of omitting the factor physical weakness in regression (1). Use a mathematical formula to justify and discipline your reasoning

f. Illustrate in two graphs: (i) which assumption of our model fails and (ii) which are the consequences when we use our sample to estimate the effect of *height* on *income*

MOCK 3 of MIDTERM EXAM

Please, read the questions carefully and provide a complete & clear answer. Good luck!

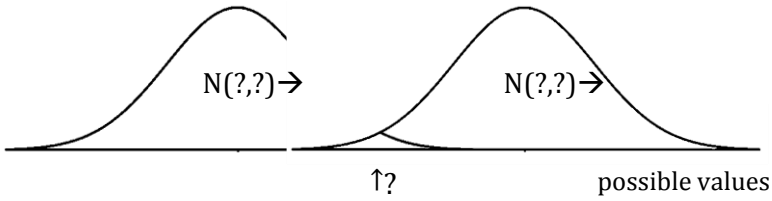
1. (2 points) You will be asked to ***prove*** *ONE from the last page*

either result *XXX* from the Appendix *using algebra & any of the other results listed before*

or the proof *YYY* that we did in class, exercises…

2. (4 points) *Based on previous estimates, the Belgian government assumed that mean earnings in Belgium were 1764€ per month. To verify this conjecture, the Statistics Agency run a survey of 500 workers in the capital, Brussels, a financial and political city, and found average earnings to be 1785€ (and a standard deviation of 335€).*

a. Construct a 95% confidence interval for the mean of Belgian earnings. Do we reject the value assumed by the government for mean earnings using a 5% significance level?



Illustrate your answer in this picture as we did in class. Fill in the 6 "?"

The Head of the Statistics Agency ordered to extend the survey to another 500 workers from the city of Antwerp, well-known by their chemical industry and big port. Average earnings of these workers are 1545€ (and the standard deviation is 295€). The total average earnings in the complete sample of 1000 workers are 1665€ (and the total standard deviation is 321).

b. Based on the full sample of 1000 workers, do we reject that mean earnings in Belgium are 1764€ using a 5% significance level?

c. Can we reject that earnings are the same in Brussels and Antwerp at 5% signif. level?

d. Write a single-regressor model and the hypothesis that, run and tested using our survey data, would yield the exact same answer as the test in (c)

e. What obvious omitted factor must be taken into account in (c) or (d) to predict whether earnings of any Spaniard looking for a job in Belgium would be equal in both cities?

f. What would be the key feature of an experiment (RCT) that gives the prediction in (e)? Explain briefly

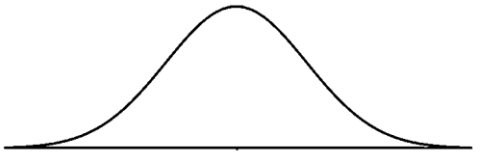
3. (4 points) *We would like to study the effect of the mother's prenatal medical care (*nprevist*, measured in number of prenatal visits) on weight at birth (*birthweight*, measured in grams). We collect a sample of 2418 births from non-smoker mothers and estimate:*

$$(\mathbf{1}) \qquad \widehat{birthweight} = 3066.4 + 32.7\, nprevist, \qquad R^2 = 0.04, \qquad SER = 573.5$$

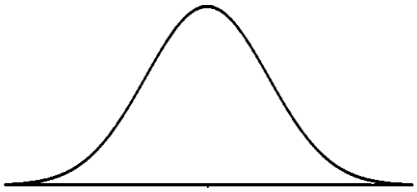
(49.8) (4.2) (*heteroskedasticity-robust standard errors*)

a. Interpret 3066.4 in terms of doctor visits and/or weight of infants

b. Test at a 1% significance level whether the expected change in birth weight resulting from one extra prenatal visit is equal to 0 or not. Fill in the 3 "?" and illustrate adding in both figures the corresponding critical values



$? \sim N(?, ?)$



$t \sim N(0,1)$

c. Would the R^2 in (1) change if we had NOT asked for *heteroscedasticity-robust standard errors* and there is presence of heteroscedasticity? Explain very briefly

The model was also estimated using data on births whose mother smoked during pregnancy and we obtained:

(2) OLS, using observations 1-582
Dependent variable: **birthweight**
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	2790.63	82.9246	33.6526	<0.0001	***
nprevist	38.1389	7.36556	5.1780	<0.0001	***

Mean dependent var.	3178.832	S.D. dependent var	580.0068
Sum squared resid	1.80e+08	S.E. of regression	557.5502
R-squared	0.077527	Adjusted R-squared	0.075937
F(1, 580)	26.81170	P-value (F)	3.10e-07

d. Draw two scatterplots representing the data for births from non-smokers and those that did smoke and both estimated models (according to the goodness of fit measured)

e. Do we reject that the effect of an extra visit to the doctor on birth weight is the same for mothers that smoked during pregnancy and those that did not (at 5% signif. level)?

MOCK 3b of MIDTERM EXAM

(if you understood the solution to Mock 3a, you must be able to answer this mock)

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. (2 points) You will be asked to ***prove*** *ONE from the last page*

either result *XXX* from the Appendix *using algebra & any of the other results listed before*

or the proof *YYY* that we did in class, exercises…

2. (4 points) *Based on previous estimates, the Scottish government assumed that mean earnings in Scotland were £1115 per month. To verify this conjecture, the Statistics Agency run a survey of 400 workers in the capital, Edinburgh, a financial and political city, and found average earnings to be £1180 (and a standard deviation of £240).*

a. Based on this evidence, do we reject the value assumed by the government for mean earnings at 5% significance level? Show below, place all critical values and fill in the 3 '?'

The Head of the Statistics Agency ordered to extend the survey to another 400 workers from the city of Glasgow, well-known by their ship-building industry and other manufacturing sector. Average earnings of these workers are £1020 (and the standard deviation is £198). The total average earnings in the complete sample of 800 workers are £1100 (and the total standard deviation is £220).

b. Using the sample of 800 workers, construct a 95% confidence interval for the mean of Scottish earnings. Do we reject now that the mean is £1115 at the significance 5% level?

c. Do we reject that earnings are the same in Edinburgh and Glasgow at 5% signif. level?

d. Write a single-regressor model and the hypothesis that, run and tested with our survey data, would yield the exact same answer as the test in (c)

e. What omitted factor must be taken into account in (**c**) & (**d**) to predict whether earnings of any Spaniard looking for a job in Scotland would be equal in both cities?

f. What would be the key feature of an experiment (RCT) that gives the prediction in (**e**)? Explain briefly

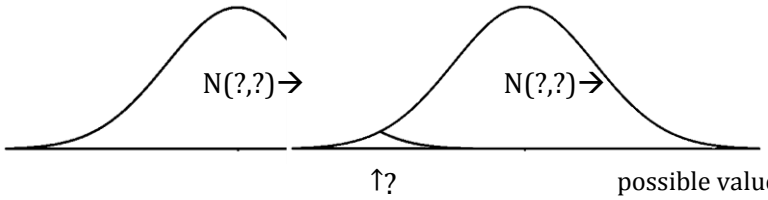
3. (4 points) *We would like to study the effect of the mother's prenatal medical care (*nprevist*, measured in number of prenatal visits) on weight at birth (*birthweight*, measured in grams). For this purpose, we collect a random sample of 3000 births and estimate:*

$$(\mathbf{1}) \qquad \widehat{birthweight} = 2979.9 + 36.7\ nprevist, \qquad R^2 = 0.05, \qquad SER = 576.7$$

(43.5) (3.7) (*heteroskedasticity-robust standard errors*)

a. Interpret 2979.93 & 36.66 in terms of doctor visits and weight of infants

b. Give the 95% confidence interval for the expected effect on *birthweight* of Δ*nprevist*=2



Illustrate your answer in this picture as we did in class. Fill in the 5 "?"

c. Is the latter effect significant if we use a 5% significance level? Explain very briefly

d. Would this interval change if we had not asked for *heteroscedasticity-robust standard errors* and there is presence of heteroscedasticity? Explain very briefly

e. Write the interpretation of the R^2

Actually, most of the mothers had visited the doctor at least 12 times before giving birth. We construct a dummy *variable that takes the value 1 if nprevist≥12 (and 0 otherwise) and we estimate:*

(2) OLS, using observations 1-3000
Dependent variable: **birthweight**
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	3275.23	16.241	201.6636	<0.0001	***
dummy	215.127	21.2711	10.1135	<0.0001	***

Mean dependent var.	3382.934	S.D. dependent var	592.1629
Sum squared resid	1.02e+09	S.E. of regression	582.4056
R-squared	0.033006	Adjusted R-squared	0.032683
F(1, 2998)	102.2838	P-value (F)	1.15e-23

f. What is the infant weight predicted by our estimated model (2) for a baby whose mother visited the doctor more than 12 times during her pregnancy?

MOCK 4 of MIDTERM EXAM

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. (2 points) You will be asked to ***prove*** *ONE from the last page*

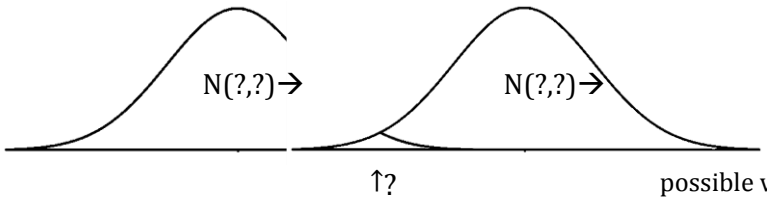
either result *XXX* from the Appendix *using algebra & any of the other results listed before*

or the proof *YYY* that we did in class, exercises…

2. (4.5 points) *Studentville, a US city, searches for raising awareness about high cholesterol levels due to fast-food consumption among university students. The ideal cholesterol level for 20-years-old people is 190mg/dL, but the mean in a sample of 400 students from John Nerd University is 205 mg/dL (and 80 mg/dL, the standard deviation). The students of this prestigious university, located in Studentville, eat fast-food for 3 or more meals per week.*

a. Based on this sample, what is the 95% confidence range for the mean cholesterol level among university students in Studentville?

b. Can we say that mean cholesterol among students in Studentville is significantly (use 5% signif. level) different from the ideal one? Add a curve in the next figure and use it to illustrate your conclusion. Also, fill in the 6 "?"



It is also located in Studentville Health&Muscles College, which is a very small university, only known for the Athletics team and a program on Sports Medicine. However, students from that college represent 20% of all university students in Studentville, thus they cannot be ignored. Therefore, they collect data on 100 students from Health&Muscles College and get an average cholesterol of 192mg/dL (and standard deviation of 10mg/dL). These students eat fast-food less than 3 meals per week. The standard deviation in cholesterol levels for all 500 university students is 71.7 mg/dL

c. Based now on the complete sample of 500 students, do we reject using a 5% signif. level that university students in Studentville have the ideal mean cholesterol level?

d. Do we reject using a 5% signif. level that mean cholesterol among students in *John Nerd University* is equal to that of the students in *Health&Muscles College*?

e. Write a single-regressor model and the hypothesis about it that, run and tested using our data, would yield the exact same answer as the test in (d)

f. Explain why we cannot conclude from (d) that if students in Studentville changed to eating fast-food less than 3 meals/week they will reach a mean level of cholesterol closer to the ideal one? Justify it also in the context of the model in (e) using a formula

g. Write very clearly the main characteristic of an experiment (RCT) which can answer the question of whether "fast-food consumption causes changes in cholesterol levels"

3. (3.5 points) *A real state agency would like to study the effect of* age *(in years) in sales* price *(in $) for homes. For this purpose, it gets data from a random sample of 1080 homes and runs:* **(1)** OLS, using observations 1-1080

Dependent variable: **price**
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	184.05	6.05	30.44	<0.0001	***
age	−1.49	0.197	−7.583	<0.0001	***

Mean dependent var	154.86	S.D. dependent var	122.91
Sum squared resid	1.56e+10	S.E. of regression	120.26
R-squared	0.243518	Adjusted R-squared	0.242630
F(1, 1078)	57.50533	P-value (F)	7.25e-14

a. Interpret the value −1.49(*the number, not only the sign!*) & <u>say if it is significant at 5%</u>

b. Which would be the estimated price for a new home that has just been built?

c. Compute the 95% confidence interval for the loss of value in a "new" house that is sold 3 years after its construction

The analysis is repeated distinguishing between homes of traditional y contemporary style:

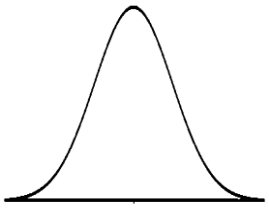
Contemporary: (**2**) $\widehat{Price} = 205.19 - 1.894\, Age,\ R^2 = 0.05$

n_C=498 (6.046) (0.197) *heteroskedasticity-robust standard errors*

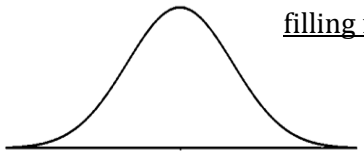
Traditional: (**3**) $\widehat{Price} = 164.16 - 1.056\, Age,\ R^2 = 0.34$

n_T=582 (5.879) (0.252) *heteroskedasticity-robust standard errors*

d. Would you reject at a 5% signif. level that *age* has the same effect on the price of homes of contemporary and traditional style (Δ=0)? Illustrate below , adding to the two figures <u>their critical values</u> and <u>filling in the '?'</u>



$\hat{\Delta} \sim N(?, ?)$



$t \sim N(0,1)$

e. Draw <u>two</u> scatterplots with regressions (2) y (3) *and <u>make sure they reflect the R^2</u>*

MOCK 4b of MIDTERM EXAM

(if you understood the solution to Mock 4a, you must be able to answer this mock)

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. (2 points) You will be asked to ***prove*** *ONE from the last page*

either result *XX* from the Appendix *using algebra & any of the other results listed before*

or the proof *YY* that we did in class, exercises…

2. (3.5 points) *An important marketing agency searches to promote publicity spending by documenting its impact on companies' sales. First, it collects sales data from 105 current client-firms and calculates that their weekly average sales are $37.84 million (with standard deviation equal to $1.95 million).*

a. Based in the current sample, what is the 95% confidence range for mean sales in any firm receiving marketing services from this agency?

The agency also finds that average weekly sales in the 54 client-firms with high (above-average) spending in publicity are $38.57 million (with standard deviation of $1.88 million), while the 51 firms with low (below-average) spending reach $37.06 million in average weekly sales (with standard deviation of $1.73 million).

b. Do we reject using a 5% signif. level that mean sales for firms with high publicity spending are equal to that in firms whose publicity spending is low? Illustrate your test using the two figures below, mark their critical values, and fill in the two '?'

$\hat{\Delta} \sim N(?, ?)$ $\quad$ $t \sim N(0,1)$

c. Write a single-regressor model and the hypothesis about it that, run and tested using our current client-firms data, would yield the exact same answer as the test in (b)

d. Why cannot we deduce from (b) that varying publicity spending causes changes in sales? Explain it also in the context of the model in (c) using a formula in your reasoning

e. Write very clearly the main characteristic of an experiment (RCT) which can answer the question of whether "higher publicity spending increases company sales"

3. (4.5 points) *A real state agency would like to study the effect of home* size *(measured in "ft²" or "squared feet") in California on their* price *(in $). For this purpose, we collect a random sample of 1080 homes and estimate:*

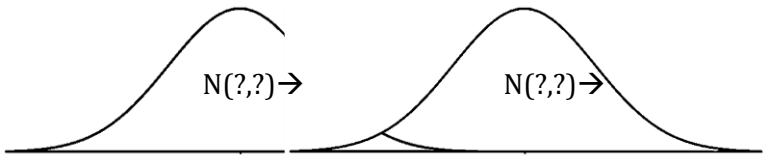
$$(\mathbf{1}) \quad \widehat{price} = -60861.5 + 92.75\ size,\ R^2 = 0.58,\ SER = 79822$$

……. (19491) (9.2) *heteroskedasticity-robust standard errors*

a. Interpret the value 92.75 (*the number, not only the sign!*) & <u>say if it is significant at 5%</u>

b. Interpret the value 79822 (*write the number in your sentence!*) <u>stating clearly the units</u>

c. Compute the 95% confidence interval for the difference in prices for two homes with the same characteristics but one measures 500 ft² more than the other. Illustrate the interval below , and <u>fill in the 6 '?'</u>



d. But, reconsidering (c), could it be possible that the effect of the size on price estimated in (1) is biased because it is capturing also the effect of a higher number of baths and bedrooms? Explain the sign of that bias & y justify your answer <u>using a formula</u>

The average size of homes in California is 2325 ft² (with standard deviation of 1008 ft²)

e. Michael Jackson's home, where he died, measures 17171 ft² and it was sold for $18.1 million: <u>compute the residual</u> for its estimated price & <u>explain why it is abnormaly large</u>

The real state agency presents the study differently to attract Spanish investors: on the one hand, it changes units from "ft²" to "m²" (1 ft² = 0.093 m²) and (…)

f. Apply the change to *m²* & rewrite <u>all that should be rewritten</u> in regression (1)

(…) *on the other, creates the variable "*large*" equal to 1 if the home is larger than 3333 ft² (0 otherwise) & runs:* **(2)** OLS, using observations 1-1080

Dependent variable: **price**
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	129635	1663.91	77.91	<0.0001	***
large	214538	23216.9	9.241	<0.0001	***

Mean dependent var	154863.2	S.D. dependent var	122912.8
Sum squared resid	1.11e+13	S.E. of regression	101670.0
R-squared	0.316421	Adjusted R-squared	0.315787
F(1, 1078)	85.38850	P-value (F)	1.27e-19

g. Interpret briefly the value of $\widehat{\beta_1}$ in regression (**2**) & <u>say if it is significant at the 1% level</u>

h. Explain why the R² could <u>only drop</u> when we estimate model (**2**) instead of (**1**)

List of BASIC RESULTS shown in the APPENDIX for FINAL EXAM

(1) $E(aX + c) = aE(X) + c$

(2) $var(X) = E[X^2] - E(X)^2$

(3) $var(aX + c) = a^2 var(X)$

(4) $E(aX + bY + c) = aE(X) + bE(Y) + c$

(5) $cov(X, Y) = E(XY) - E(X)E(Y)$

(6) $cov(X, X) = var(X)$

(7) $cov(aX, c + bY) = ab\ cov(X, Y)$

(8) $cov(aX + bY + c, Z) = a\ cov(X, Z) + b\ cov(Y, Z)$

(9) $var(aX + bY + c) = a^2\ var(X) + b^2\ var(Y) + 2ab\ cov(X, Y)$

(12) $E[E(X|Y)] = E(X)$ Law of Iterated Expectations(LIE)

(13) $P(X = x_i \cap Y = y_j) = P(X = x_i) \cdot P(Y = y_j)\ \forall\ x_i, y_j \leftrightarrow X$ e Y independent

(14) X & Y $independent \rightarrow corr(X, Y) = 0$

List of PROOFS from LECTURES, PROBLEM SETS.... for FINAL EXAM

(I) $E(\bar{Y}) = \mu_Y$ for a sample $\{Y_1, Y_2, \dots, Y_n\}$ i.i.d.

(II) $S_X^2 = \left(\frac{n}{n-1}\right)\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2\right)$ is the variance of the sample $\{X_1, X_2, \dots, X_n\}$

(III) $S_{XY} = \left(\frac{n}{n-1}\right)\left(\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\,\bar{y}\right)$ for the sample $\{Y_1, Y_2, \dots, Y_n, X_1, X_2, \dots, X_n\}$

(IV) solve the OLS problem to estimate $Y_i = \beta_0 + \beta_1 X_i + u_i$ and show that then $\sum_{i=1}^n \hat{u}_i = 0$, & $\sum_{i=1}^n \hat{u}_i x_i = 0$, as well as $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$, & $\widehat{\beta_1} = S_{XY}/S_X^2$

(V) solve the OLS problem for $Y_i = \beta_0 + \beta_1 X_i + u_i$ if $x_i = 0\ \forall\ i$ and show that $\hat{\beta}_0 = \bar{y}$

(VI) $S_{\hat{u}X} = 0$ when you estimate $Y = \beta_0 + \beta_1 X + u$ using OLS

(VII) if $E(u|X) = constant$, then $corr(u, X) = 0$

(VIII) in $Y = \beta_0 + \beta_1 X + u$: if X increases in 1 unit, Y rises in β_1 units *on average...*

(IX) ... but $\beta_1 = E(Y|D = 1) - E(Y|D = 0)$ in $Y = \beta_0 + \beta_1 X + u$ if D is dummy

(X) in $Y = \beta_0 + \beta_1 \ln(X) + u$: if X increases in 1%, Y rises in $0.01 \cdot \beta_1$ units *on average*

(XI) in $\ln(Y) = \beta_0 + \beta_1 X + u$: if X increases in 1 unit, Y rises in $(100 \cdot \beta_1)\%$ *on average*

(XII) in $\ln(Y) = \beta_0 + \beta_1 \ln(X) + u$: if X increases in 1%, Y rises in $\beta_1\%$ *on average*

(XIII) $Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \cdot D) + u$: the *average* effect of X on Y rises in β_3 if $D=1$ and the *expected* effect of D on Y depends, at the same time, on the level of X

(XIV) $TSS = SSR + ESS$ when you estimate $Y = \beta_0 + \beta_1 X + u$ using OLS

Moments / Parameters in the population	**Sample-estimated** moments/parameters
$E(X) = \sum_{j=1}^{m} x_j\, P(X = x_j) \equiv \mu_X$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \hat{\mu}_X$
$var(X) = E[(X-\mu_X)^2] \equiv \sigma_X^2$	$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \equiv \hat{\sigma}_X^2$
$sd(X) = \sqrt{\sigma_X^2} \equiv \sigma_X$	$s_X = \sqrt{s_X^2} \equiv \hat{\sigma}_X$
$cov(X,Y) =$ $= E[(X-\mu_X)(Y-\mu_Y)] \equiv \sigma_{XY}$	$s_{XY} =$ $= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \equiv \hat{\sigma}_{XY}$
$corr(X,Y) = {}^{\sigma_{XY}}/_{\sigma_X\sigma_Y} \equiv \rho_{XY}$	$r_{XY} = {}^{s_{XY}}/_{s_X s_Y} \equiv \hat{\rho}_{XY}$
$Y = \beta_0 + \beta_1 X_1 + u \quad \rightarrow \hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 \;;\; \hat{u} = Y - \hat{Y}$ $\{ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u \quad \rightarrow \hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_k} X_k \quad ; \hat{u} = Y - \hat{Y} \}$	
$\beta_1 = \frac{\Delta E(Y\,\vert X_1)}{\Delta X_1} \left\{ = \frac{\Delta E(Y\,\vert X_1, \ldots, X_k)}{\Delta X_1} \right\}$	$\widehat{\beta_1} = \frac{s_{YX}}{s_X^2} \left\{ = \frac{s_{Y1}s_{22} - s_{12}s_{Y2}}{s_{11}s_{22} - s_{12}^2} \;\; si \;\; k = 2 \right\}$
$\beta_0 = E(Y\,\vert X_1 = 0)$ $\{ = E(Y\,\vert X_1 = X_2 = \cdots = X_k = 0) \}$	$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}_1$ $\{ = \bar{Y} - \widehat{\beta_1}\bar{X}_1 - \widehat{\beta_2}\bar{X}_2 - \cdots - \widehat{\beta_k}\bar{X}_k \}$
||||||||||||||||||||||||||||||||||||	||||||||||||||||||||||||||||||||||||
$E(\overline{X}) = \mu_{\bar{X}} \equiv \mu_X$	
$var(\overline{X}) = \frac{1}{n}\sigma_X^2 \equiv \sigma_{\bar{X}}^2$	$\hat{\sigma}_{\bar{X}}^2 = \frac{1}{n}\hat{\sigma}_X^2 \equiv s_{\bar{X}}^2$
$sd(\overline{X}) = \frac{1}{\sqrt{n}}\sigma_X \equiv \sigma_{\bar{X}}$	$\hat{\sigma}_{\bar{X}} = \frac{1}{\sqrt{n}}\hat{\sigma}_X \equiv SE(\overline{X})$
$E(\widehat{\beta_1}) \cong \beta_1 + \rho_{Xu}\, {}^{\sigma_u}/_{\sigma_X}$	
$var(\widehat{\beta_1}) = \frac{1}{n}\frac{var((X-\mu_X)u)}{(var(X))^2} \equiv \sigma_{\widehat{\beta_1}}^2$ $\{but\ complicated\ if \quad k > 1\}$	$\hat{\sigma}_{\widehat{\beta_1}}^2 = \frac{1}{n}\frac{\sum_{1=1}^{n}\hat{u}_i{}^2(x_i-\overline{x})^2/(n-2)}{[\sum_{1=1}^{n}(x_i-\overline{x})^2/n]^2}$ $\{but\ complicated\ if \quad k > 1\}$
$sd(\widehat{\beta_1}) = \sqrt{\sigma_{\widehat{\beta_1}}^2} \equiv \sigma_{\widehat{\beta_1}}$	$\hat{\sigma}_{\widehat{\beta_1}} = \sqrt{\hat{\sigma}_{\widehat{\beta_1}}^2} \equiv SE(\widehat{\beta_1})$
||||||| $R^2 = \frac{SCE}{SCT} = \frac{\sum_{1=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}{\sum_{1=1}^{n}(y_i-\bar{y})^2}$; $\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2)$; $SER = \sqrt{\sum_{i=1}^{n}\hat{u}_i^2/(n-k-1)}$ |||||||	
To test $H_0: \beta_1 = 0\ \&\ \beta_2 = 0$ $F = \frac{t_1^2 + t_2^2 - 2\hat{\rho}_{12}t_1t_2}{2(1-\hat{\rho}_{12}^2)}$	(with homoscedasticity) $F = \frac{(R_{un}^2 - R_{res}^2)/q}{(1-R_{un}^2)/(n-k-1)} \;\; where\ q = 2$

The values -2.575, -1.96 & -1.64 leave 0.5, 2.5 & 5% probability in the left-tail of the N(0,1)
The 5% critical values of the χ_q^2/q with q=1, 2, 3, 4 & 5 are 3.84, 3, 2.6, 2.37 & 2.21

MOCK FINAL EXAM 1a

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. Two proofs (2 points) from the last page

a. result *XXX* from the Appendix *using algebra & any of the other results listed before*

b. the proof *YYY* that we did in class, exercises…

2. (3.5 points) *An environmentalist lobby says that lowering public transport prices (mainly bus transport) in US cities will increase its use (and, thus, decrease air pollution). We would like to test that demand of bus services will react in that way to ticket prices.*

a. Describe very briefly the key feature of an experiment that would guarantee that we measure the effect of ticket prices on the number of users, and not that of any other factor

In practice, we cannot run experiments, but we obtain observational data from 120 cities on their demand of bus services (BUS), in thousands of users; ticket prices (PR), gasoline prices (PGAS), and average income per capita (INC), in dollars; population (POP) in millions of inhabitants, and density, in number of inhabitants per mile² (DENS); and whether they are located in the Northeast (NE), Midwest (MW), South (S) or West (W), all dummy variables. We estimate:

Model 1: OLS, using observations 1-120
Dependent variable: BUS
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	2377.56	848.674	2.8015	0.0060	***
PR	−324.026	228.508	−1.4180	0.1590	
PGAS	775.732	952.583	0.8143	0.4172	
INC	−0.197698	0.0307104	−6.4375	<0.0001	***
POP	1.58668	0.0933463	16.9978	<0.0001	***
DENS	0.143072	0.0228438	6.2631	<0.0001	***
NE	−43.8876	187.265	−0.2344	0.8151	
MW	127.599	**???**	**???**	0.5006	
S	−19.9891	178.139	−0.1122	0.9109	

Mean dependent var	1933.175	S.D. dependent var	2411.236
Sum squared resid	54800219	S.E. of regression	702.6348
R-squared	0.920794	Adjusted R-squared	0.915086
F(8, 111)	116.4465	P-value (F)	2.33e-50

b. Test H_0: $\beta_{PR} \geq 0$ against the claim of the lobby and use it to argue briefly against them

c. What is the hypothesis of the test of the regression? Is it rejected at 5% signif. level?

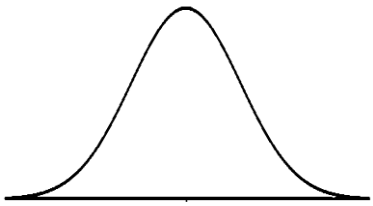
The variance-covariance matrix for the estimated coefficients:

const	***PR***	***PGAS***	***INC***	***POP***	***DENS***	***NE***	***MW***	***S***	
720247	14432.5	-617688	-12.263	-8.4279	7.32843	-15109	-16230	-2725.0	***const***
	52215.9	-102145	1.38831	-9.7479	1.90096	2091.82	-3976.4	-1978.9	***PR***
		907415	-3.6533	12.819	-6.6559	-6165.5	-5533.9	-18755	***PGAS***
			9.431e-4	5.31e-4	-3.34e-4	0.05319	0.49802	0.15146	***INC***
				0.00871	-0.0014	-1.1700	0.81179	1.06821	***POP***
					5.218-4	0.18554	-0.3269	-3.0e-4	***DENS***
						35068.2	17984.1	18086.2	***NE***
							35659.8	18250.4	***MW***
								31733.4	***S***

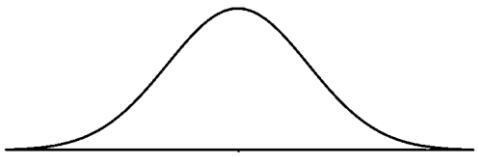
d. Fill in the **two** "???" on the Gretl output

e. Rewrite the regression if we omit NE and include instead W

f. Check if public transport services constitute an inferior good by testing (also at 5% significance level) H_0: $\beta_{INC} \leq 0$. Locate below critical values, estimated value or statistic and fill in the "?"



?~N(? , ?)



t~N(0,1)

3. (4.5 points) *Using homoscedastic data on salaries, etc, of 269 NBA players, we would like to determine the effect of their points scored and rebounds reached per game on their wages. Based on the corresponding column from **A** to **E**, answer each question in **a-e**:*

a. *Most players get four rebounds per game.* Based on **A**, draw two plausible scatterplots for those players: one of ***points*** against **ln(*wages*)** and another of ***points*** against ***wages***

b. Interpret in **B** (including the number & units in your sentence) $\widehat{\beta_0}$ equal to 5.56

c. Show, in graphs like those in **2.e,** the p-value of the significance test for β_4 in **C**

d. Explain briefly why $\widehat{\beta_4}$ has become negative in ***D*** while it was positive in regression ***C***

e. Test jointly whether being married and having children is relevant for wages (5% level)

Dependent variable: natural log of wages (in thousands of $) , i.e. **ln(*wages*)** ; **n=269**					
Regressors	***(A)***	***(B)***	***(C)***	***(D)***	***(E)***
1. *points*	0.075 (.0076)	0.109 (.0119)	0.101 (.0123)	0.090 (.0122)	0.090 (.0122)
2. *rebounds*	0.062 (.0152)	0.144 (.0259)	0.125 (.0269)	0.117 (.0261)	0.118 (.0266)
3. *points × rebounds*		-0.007 (.0015)	-0.006 (.0016)	-0.005 (.0017)	-0.005 (.0017)
4. *age*			0.060 (.0120)	-0.055 (.0306)	-0.057 (.0315)
5. *exper*				0.196 (.0541)	0.196 (.0541)
6. $exper^2$				-0.005 (.0028)	*-0.005 (.0032)*
7. *married*					*0.022 (.0839)*
8. *children*					*0.018 (.0761)*
0. constant	5.90 (.0951)	5.56 (.1392)	4.01 (.3466)	6.49 (.7200)	6.54 (.7358)
R^2	0.4121	0.4353	0.4873	0.5207	0.5210

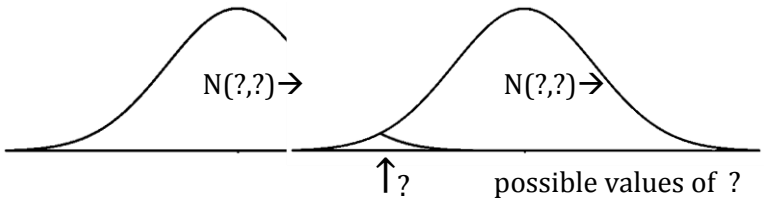
Estimates in Italics are those that are NOT different from zero at 10% significance level.
Married and children are dummy variables that take the value 0 or 1

*Based on the corresponding columns from **A** to **E** answer questions **f** to **h**:*

f. Which model would you choose based on the adjusted R^2 (or $\overline{R^2}$) among ***C***, ***D*** and ***E***?

g. According to the most explanatory regression model, what is the expected difference in **ln(*wages*)** for two NBA players that are basically equal in all characteristics but one is 2 years younger than the other? State which is the approximated difference in ***wages***, then

h. Illustrate below, using a 95% confidence interval, that the difference in **ln(*wages*)** or ***wages*** between these two players is not significant (-ly different from zero). Add a curve and fill in the 6 '?' to do so



MOCK FINAL EXAM 1b

(if you understood the solution to Mock 1a, you must be able to answer this mock)

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. Two proofs (2 points) from the last page

a. result *XXX* from the Appendix *using algebra & any of the other results listed before*

b. the proof *YYY* that we did in class, exercises…

2. (4.5 points) *A Federal Agency would like to know whether demand of bus services in US cities will increase if it subsidizes bus tickets.*

a. Describe very briefly the key feature of an experiment that would guarantee that we measure the effect of ticket prices on the number of users, and not that of any other factor

In practice, we cannot run experiments, but we obtain observational data from 120 cities on their demand of bus services (BUS), in thousands of users; ticket prices (PR), gasoline prices (PGAS), and average income per capita (INC), in dollars; population (POP) in millions of inhabitants, and density, in number of inhabitants per mile² (DENS); and whether they are located in the Northeast (NE), Midwest (MW), South (S) or West (W), all dummy variables. We estimate:

Model 1: OLS, using observations 1-120
Dependent variable: BUS
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	2377.56	848.674	2.8015	0.0060	***
PR	−324.026	228.508	−1.4180	0.1590	
PGAS	775.732	952.583	0.8143	0.4172	
INC	−0.197698	0.0307104	−6.4375	<0.0001	***
POP	1.58668	0.0933463	16.9978	<0.0001	***
DENS	0.143072	0.0228438	6.2631	<0.0001	***
NE	−43.8876	187.265	−0.2344	0.8151	
MW	127.599	188.838	0.6757	0.5006	
S	−19.9891	178.139	−0.1122	0.9109	

Mean dependent var	1933.175	S.D. dependent var	2411.236
Sum squared resid	54800219	S.E. of regression	702.6348
R-squared	0.920794	Adjusted R-squared	0.915086
F(8, 111)	116.4465	P-value (F)	2.33e-50

b. Explain whether the effect of tickets & gasoline prices have the expected sign

c. Are tickets & gasoline prices jointly significant factors for bus services demand at 5%?

Illustrate the test drawing the distribution of your statistic and the corresponding p-value

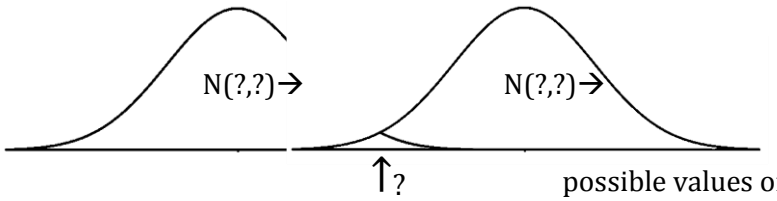
The variance-covariance matrix for the estimated coefficients:

const	***PR***	***PGAS***	***INC***	***POP***	***DENS***	***NE***	***MW***	***S***	
720247	14432.5	-617688	-12.263	-8.4279	7.32843	-15109	-16230	-2725.0	***const***
	52215.9	-102145	1.38831	-9.7479	1.90096	2091.82	-3976.4	-1978.9	***PR***
		907415	-3.6533	12.819	-6.6559	-6165.5	-5533.9	-18755	***PGAS***
			???	5.31e-4	-3.34e-4	0.05319	0.49802	0.15146	***INC***
				0.00871	-0.0014	-1.1700	0.81179	1.06821	***POP***
					5.218-4	0.18554	-0.3269	-3.0e-4	***DENS***
						???	17984.1	18086.2	***NE***
							35659.8	18250.4	***MW***
								31733.4	***S***

d. Fill in the two "???"on the variance-covariance matrix

e. What is the estimated effect on the demand of bus services of a wave of immigrants that increases population in 2000 residents and density in 1 inhabitant per $mile^2$?

f. Calculate the 95% confidence interval for the effect in the number of users and illustrate below, filling in the 6 '?'



possible values of ?

g. Interpret the coefficient -19.99 (-20 approx) for the binary variable S (South)

3. (3.5 points) *Using homoscedastic data on salaries, etc, of 269 NBA players, we would like to determine the effect of their points scored and rebounds reached per game on their wages. Based on the corresponding column from **A** to **E**, answer each question in **a-e**:*

a. If we omit the variable ***rebounds***, we estimate $\ln(\widehat{wages}) = 6.1 + 0.093points$, explain why the estimated coefficient of *points* is larger than in regression ***A*** using a formula

b. Give the estimated change in **ln(*wages*)** for a player that scores 1 extra point in all games but never gets a rebound according to regression **B**. Do the same for a player that reaches 4 rebounds per game. Describe in your own words the non-linear effect that an extra point has on **ln(*wages*)**

c. Interpret 0.06 as an effect on ***wages***. Explain how it is possible to estimate $\widehat{\beta_4} > 0$

d. Draw a figure illustrating the estimated non-linear effect of experience on **ln(*wages*)**

e. Give hypothesis, statistic, critical value and conclusion of the *test of the regression* for ***E***

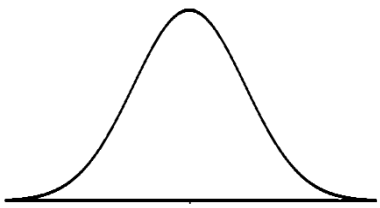
Dependent variable: natural log of wages (in thousands of $) , i.e. *ln(WAGE)* ; **n=269**					
Regressors	***(A)***	***(B)***	***(C)***	***(D)***	***(E)***
1. *points*	0.075 (.0076)	0.109 (.0119)	0.101 (.0123)	0.090 (.0122)	0.090 (.0122)
2. *rebounds*	0.062 (.0152)	0.144 (.0259)	0.125 (.0269)	0.117 (.0261)	0.118 (.0266)
3. *points × rebounds*		-0.007 (.0015)	-0.006 (.0016)	-0.005 (.0017)	-0.005 (.0017)
4. *age*			0.060 (.0120)	-0.055 (.0306)	-0.057 (.0315)
5. *exper*				0.196 (.0541)	0.196 (.0541)
6. $exper^2$				-0.005 (.0028)	*-0.005 (.0032)*
7. *married*					*0.022 (.0839)*
8. *children*					*0.018 (.0761)*
0. constant	5.90 (.0951)	5.56 (.1392)	4.01 (.3466)	6.49 (.7200)	6.54 (.7358)
R^2	0.4121	0.4353	0.4873	0.5207	0.5210

Estimates in Italics are those that are NOT different from zero at 10% significance level.
Married and children are dummy variables that take the value 0 or 1

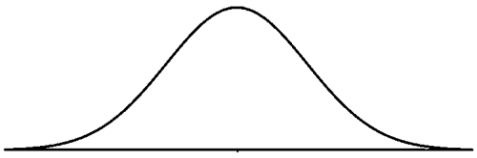
*Based on the corresponding columns from **A** to **E** answer questions **f** and **g**:*

f. Which model would have the highest SER? Justify your answer briefly

g. Illustrate, according to the most explanatory regression model, the test on whether ***age*** is a relevant determinant of wages at a 10% significance level. Show critical values and estimated value or statistic in the two figures below, and fill in the '?'



?~N(? , ?)



t~N(0,1)

MOCK FINAL EXAM 2a

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. Two proofs (2 points) from the last page

a. result XXX from the Appendix *using algebra & any of the other results listed before*

b. the proof YYY that we did in class, exercises…

2. (4 points) *We want to measure the effect of a firm's sales on the salary of its CEO.*

a. *Which of these experiments does NOT ensure that differences in salaries are due to higher sales; that is, it does NOT measure a causal effect of sales on CEOs' salaries? Explain briefly*
i. Select a sample of very different CEOs and firms, and randomly increase the demand of the products sold by half of them. (…)
ii. Select by random sampling some CEOs and firms, and artificially increase the demand of the products sold by the half of firms with smaller size. (…)
iii. Select by random sampling some CEOs and firms, and increase the demand of the products sold by the half of companies whose CEO was selected first. (…)
iv. Select a sample of small business CEOs and firms, and randomly increase the demand of the products sold by half of them. (…)
(…) Then, compare the average salary of the CEOs of companies whose sales were artificially increased with that of CEOs whose companies' sales were not changed

Instead of running an experiment, we estimated using observational data on 177 CEOs a regression model. Some variables were included taking natural logarithms; for example, the dependent variable **l_salary**=ln(**salary**), where **salary** is measured in thousands of $. *The regression includes factors like* ***sales*** *(in millions of $ and logs)***,** *the market value of the firm (****mktval,*** *also in millions of $ and logs), the number of years in the current company or "tenure in the company" (****comten****), or years as CEO on the current firm or "tenure as CEO" (****ceoten****).*

b. Interpret $\widehat{\beta_0}$ as an estimated value (of what?) for a CEO with a particular profile (which?)

c. What is the expected effect on the CEO´s ***salary*** (*without log!*) of a 10% increase in the ***sales*** (*also no log!*) of the company? Is this effect statistically significant at 5% level?

d. *Experts claim that there is a "superstar" effect: firms that hire CEOs from outside the company search for highly regarded candidates, and their salaries are bid up.* Do you find evidence against this claim at 5% significance level?

Model 1: OLS, using observations 1-177
Dependent variable: **l_salary**
Heteroskedasticity-robust standard errors, variant HC1

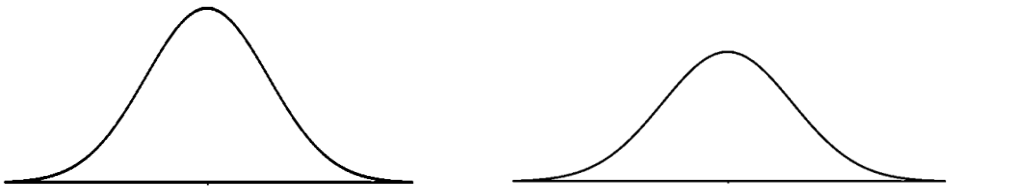
		Coefficient	*Std. Error*	*t-ratio*	*p-value*	
(0)	const	4.576	0.278	16.443	<0.0001	***
(1)	**l_sales**	0.192	0.036	5.371	<0.0001	***
(2)	**l_mktval**	0.094	0.049	1.928	0.0555	*
(3)	**ceoten**	0.017	0.007	2.281	0.0238	**
(4)	**comten**	−0.009	0.003	−3.035	0.0028	***

Mean dependent var	6.5828	S.D. dependent var	0.6061
Sum squared resid	42.1259	S.E. of regression	0.4949
R-squared	0.34836	Adjusted R-squared	0.3332
F(4, 172)	26.6225	P-value (F)	3.40e-17

Covariance matrix <u>for the coefficients</u> *NB: 2e-003 = 0.002*

	(0) **const**	(1) **l_sales**	(2) **l_mktval**	(3) **ceoten**	(4) **comten**	
(0)	0.07746	-0.00017	-0.00896	-0.0012	-5.005e-005	**const**
(1)		0.00128	-0.00125	-7.61e-005	1.405e-005	**l_sales**
(2)			0.00238	0.00021	-2.91e-005	**l_mktval**
(3)				**5.502e-005**	-6.33e-006	**ceoten**
(4)					9.623e-006	**comten**

e. *A journalist claims that the <u>elasticity</u> of a CEO's **salary** to **sales** is 0.5.* <u>Can you reject</u> this claim at 5% significance level? Illustrate the test locating in both figures below the critical values, the estimated value or statistic, and filling in the "?"



?~N(? , ?) t~N(0,1)

f. <u>Interpret</u> the number **5.502e-005** in the covariance matrix: say what <u>it is</u>, explain what <u>it means</u>, and state the <u>units</u> corresponding to this number.

3. (4 points) *The US revenue agency is concerned about the financial wealth of US workers. Based on a sample of 3637 over 25-years-old workers, this agency runs several regressions that explain their net financial wealth (**nettfa**, in thousands of $) with annual income (**inc**, also in thousands of $), **age** (in years <u>over 25</u>) and **age²**, and whether the worker takes part on a pension plan called 401k (**p401k** is equal to 1 if she does, 0 otherwise). Using the following estimated models featuring <u>heteroscedasticity-robust standard errors</u>, answer questions **a-d** corresponding to each column from **A** to **D:***

Dependent variable: net total financial assets or ***nettfa***, in thousands of $; **n=3637**				
Regressors	**(A)**	**(B)**	**(C)**	**(D)**
(1) *inc*	1.17 (.107)	1.13 (.106)	1.08 (.110)	0.879 (.357)
(2) *age*		-0.107 (.435)	-0.009 (.445)	-0.012 (.445)
(3) age^2		0.041 (.014)	0.039 (.014)	0.039 (.014)
(4) *p401k*			17.37 (2.52)	6.14 (14.99)
(5) *p401k × inc*				0.26 (.369)
(0) constant	-25.1 (4.34)	-36.4 (4.09)	-47.1 (4.07)	-38.6 (12.04)
R^2	0.16	0.20	0.20	0.21
F (test of the regression)	121.2	87.9	102.7	85.5

*NB: if ***age***=2, that worker is 27-years-old, and ***age²*** takes the value 4 in that case

a. what is the interpretation of the slope coefficient in regression **A**? Comment its value (mentioning units) in your sentence and NOT only the sign

b. test individually whether β_2 & β_3 are significant in model **B**, and then draw a plausible scatterplot for the ages and holdings of net financial assets observed in the data

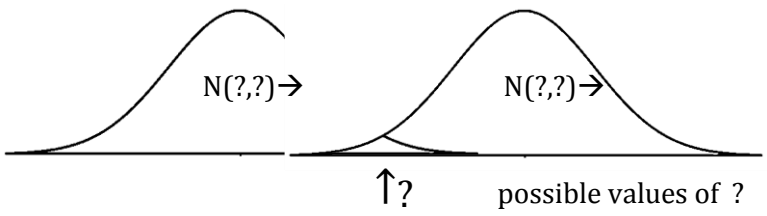
c. interpret (using its value in your sentence) the estimated coefficient for ***p401k*** in **C**

d. write and draw the "straight lines" in model **D** explaining the relation between income and financial wealth for 25-years-old workers (i.e. their values for ***age*** and ***age²*** are 0)

*Based on the **most explanatory model** from all columns, answer questions **e** to **g**:*

e. give the hypothesis, recalculate the *F*-statistic of the "test of the regression" *ignoring any heteroscedasticity* of the data and say if you decide the same thing than using the *F* reported

f. compute the 95% confidence interval for the difference in ***nettfa*** between two workers that do NOT take part in the pension plan and they only differ in that one earns $2000 less than the other. Illustrate below and fill in the 6 '?'



g. Is the previous difference significant (-ly different from 0)? Justify your answer and illustrate it by adding a curve to the figure above

MOCK FINAL EXAM 2b

(if you understood the solution to Mock 2a, you must be able to answer this mock)

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. Two proofs (2 points) from the last page

a. result *XXX* from the Appendix *using algebra & any of the other results listed before*

b. the proof *YYY* that we did in class, exercises…

2. (4 points) *We want to measure the effect of a firm's sales on the salary of its CEO.*

a. *Which of these experiments does NOT ensure that differences in salaries are due to higher sales; that is, it does NOT measure a causal effect of sales on CEOs' salaries? Explain briefly*
i. Select by random sampling some CEOs and their firms, and divide them in two groups: those whose sales are above average and those whose sales stay below average. (…)
ii. Select by random sampling some CEOs and their firms, and divide them in two groups: those whose sales are above the median and those whose sales stay below that value. (…)
iii. Select by random sampling some CEOs and firms, and artificially increase the demand of the products sold by the half of firms with highest sales. (…)
iv. Select a sample of small business CEOs and firms, and randomly increase the demand of the products sold by half of them. (…)
(…) Then, compare the average salary of the CEOs of the two groups of companies.

Instead of running an experiment, we estimated using observational data on 177 CEOs a regression model. Some variables were included taking natural logarithms; for example, the dependent variable l_salary=ln(***salary***), where ***salary*** is measured in thousands of $. *The regression includes factors like* ***sales*** *(in millions of $ and logs)****,*** *the market value of the firm (****mktval,*** *also in millions of $ and logs), the number of years in the current company or "tenure in the company" (****comten****), or years as CEO on the current firm or "tenure as CEO" (****ceoten****).*

b. What is the estimated effect on a CEO's ***salary*** (*no log!*) of a 1% rise in ***sales*** (*no log!*)

c. What is the estimated effect on a CEO's ***salary*** (*notice: the variable in levels*) of a 5% increase in ***mktval*** (*note, also, no log!*)? Is this effect statistically significant at 5% level?

d. Discuss briefly whether $\widehat{\beta_3}$ has the expected sign

Model 1: OLS, using observations 1-177
Dependent variable: **l_salary**
Heteroskedasticity-robust standard errors, variant HC1

		Coefficient	*Std. Error*	*t-ratio*	*p-value*	
(0)	const	4.576	0.278	16.443	<0.0001	***
(1)	**l_sales**	0.192	0.036	5.371	<0.0001	***
(2)	**l_mktval**	0.094	0.049	1.928	0.0555	*
(3)	**ceoten**	0.017	0.007	2.281	0.0238	**
(4)	**comten**	−0.009	0.003	−3.035	0.0028	***

Mean dependent var	6.5828	S.D. dependent var	0.6061
Sum squared resid	42.1259	S.E. of regression	0.4949
R-squared	0.34836	Adjusted R-squared	0.3332
F(4, 172)	26.6225	P-value (F)	3.40e-17

Covariance matrix <u>for the coefficients</u> *NB: 2e-003 = 0.002*

	(0) **const**	(1) **l_sales**	(2) **l_mktval**	(3) **ceoten**	(4) **comten**	
(0)	0.07746	-0.00017	-0.00896	-0.0012	-5.005e-005	**const**
(1)		0.00128	-0.00125	-7.61e-005	1.405e-005	**l_sales**
(2)			0.00238	0.00021	-2.91e-005	**l_mktval**
(3)				5.502e-005	-6.33e-006	**ceoten**
(4)					9.623e-006	**comten**

e. <u>Test</u> with a 5% significance level if two CEOs who work for very similar firms and have different tenures would earn different salaries. <u>In particular,</u> their only difference is that one CEO has been working for the company 2 more years than the other, and 1 of those 2 years as CEO. <u>Illustrate</u> the test locating in both figures below the critical values, the estimated value or statistic, and filling in the "?"

$\Delta\ln(\widehat{salary}) \sim N(?, ?)$ $\quad t \sim N(0,1)$

f. <u>Write</u> the hypotheses & statistic of the <u>"test of the regression"</u> and <u>justify</u> your conclusion

3. (4 points) *The US revenue agency is concerned about the financial wealth of US workers. Based on a sample of 3637 over 25-years-old workers, this agency runs several regressions that explain their net financial wealth (**nettfa**, in thousands of $) with annual income (**inc**, also in thousands of $), **age** (in years <u>over 25</u>) and **age²**, and whether the worker takes part on a pension plan called 401k (**p401k** is equal to 1 if she/he does, 0 otherwise). Using the following estimated models featuring <u>heteroscedasticity-robust standard errors</u>, answer questions **a-d** corresponding to each column from **A** to **D:***

Dependent variable: net total financial assets or *nettfa*, in thousands of $; **n=3637**				
Regressors	**(A)**	**(B)**	**(C)**	**(D)**
(1) *inc*	1.17 (.107)	1.13 (.106)	1.08 (.110)	0.879 (.357)
(2) *age*		-0.107 (.435)	-0.009 (.445)	-0.012 (.445)
(3) age^2		0.041 (.014)	0.039 (.014)	0.039 (.014)
(4) *p401k*			17.37 (2.52)	6.14 (14.99)
(5) *p401k × inc*				0.26 (.369)
(0) constant	-25.1 (4.34)	-36.4 (4.09)	-47.1 (4.07)	-38.6 (12.04)
R^2	0.16	0.20	0.20	0.21
F (test of the regression)	121.2	87.9	102.7	85.5

*NB: if age=2, that worker is 27-years-old, and age^2 takes the value 4 in that case

a. interpret the intercept of regression **A** specifying its value and units in your sentence

b. test in **B**, *assuming that the data was homoscedastic*, whether we should control for the fact that in our sample there are workers of different ages. Write the hypotheses, compute the statistic, state its distribution, and give the 5% critical values with your conclusion

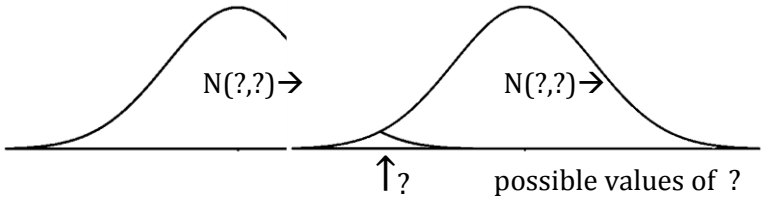
c. rewrite the estimated betas (not their SE), the R^2 and F-statistic of the "test of the regression" of model **C** if we had included the variable ***nonp401k*** (that takes the value 1 if the worker is NOT participating in the pension plan and 0 otherwise) instead of ***p401k***

d. *a 25-years-old worker earns $24000 annually on average.* What is the effect on her ***nettfa*** of being in the 401k pension plan? And if she earns *$20000*? And if she earns *$28000*?

e. which is the model with the highest adjusted R^2 (or $\overline{R^2}$) from all columns? Calculate it

Based on the **model with the highest adjusted R^2** (or $\overline{R^2}$), answer questions ***f*** & ***g***:

f. compute & illustrate below the 95% confidence interval for the difference in ***nettfa*** between two workers of the same age who do NOT take part in the 401k pension plan and one earns *$3000* less than the other. Fill in the 6 '?'



g. could we include both variables ***p401k*** & ***nonp401k*** in that regression? Give a name to the possible consequences of doing so and explain what it means

MOCK FINAL EXAM 3a

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. Two proofs (2 points) from the last page

a. result *XXX* from the Appendix *using algebra & any of the other results listed before*

b. the proof *YYY* that we did in class, exercises…

2. (4 points) *The Regional Health Service (RHS) needs to quantify the effect on the health of a baby of an **extra visit** to the doctor of the mother during pregnancy. The baby's health is measured as his or her weight at birth (**weight**, in grams).*

a. *Which of these experiments ENSURES that changes in **weight** are due to **more doctor visits**; in other words, it does MEASURES a* causal effect *of **visits** on **weight**? Explain briefly*
i. Select a sample of babies and divided them in two groups: those whose mother received below average prenatal care & those that received above average prenatal care (…)
ii. Select a sample of mothers and randomly provide extra prenatal care to half of them. (…)
iii. Select by random sampling some mothers and provide extra prenatal care to those that bear their first child. (…)
iv. Select a random sample of mothers and divided them in two groups: those that bear their first child and those that have already delivered at least one child (…)
(…) Then, compare the average weight at birth of the babies delivered by the two groups.

*Instead of an experiment, the RHS estimated with observational data a regression. The model allows for a quadratic relation between **weight** and the number of visits to the doctor during pregnancy (including, thus, **visit** and **visit²**). The regression also controls for **drink**: the average number of weekly alcoholic drinks that the mother takes during pregnancy (ranking from 0 to 8).*

Model 1: OLS, using observations 1-1651
Dependent variable: **weight**
Heteroskedasticity-robust standard errors, variant HC1

		Coefficient	*Std. Error*	*t-ratio*	*p-value*	
(0)	const	3102.36	98.4668	31.5067	<0.0001	***
(1)	**visit**	36.2292	12.925	2.8030	0.0051	***
(2)	**visit2**	−0.75307	0.4105	−1.8345	0.0668	*
(3)	**drink**	−48.9098	25.4192	−1.9241	0.0545	*

Mean dependent var.	3409.520	S.D. dependent var.	575.1008
Sum squared resid	5.39e+08	S.E. of regression	572.0800
R-squared	0.012277	Adjusted R-squared	0.010478
F(3, 1647)	5.577219	P-value (F)	0.000833

Covariance matrix <u>for the coefficients</u> *NB: 2e-003 = 0.002*

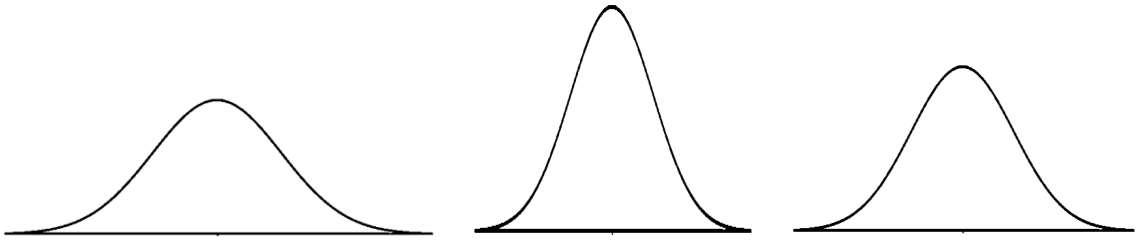
	(0)	(1)	(2)	(3)	
	const	**visit**	**visit2**	**drink**	
(0)	9695.70	-1229.76	33.8000	25.4714	**const**
(1)		167.055	-4.99249	-5.24652	**visit**
(2)			0.168510	0.0394247	**visit2**
(3)				646.138	**drink**

b. <u>Draw</u> a plausible <u>scatterplot</u> of the number of visits to the doctor and the weight of the baby at birth for mothers having the same number of weekly alcoholic drinks

c. <u>Test</u> with a 5% signif. level whether the effect of prenatal care on **weight** is non-linear

d. <u>Interpret</u> $\widehat{\beta_0}$ as the estimate (<u>of what?</u>) for a baby with a particular profile (<u>which?</u>)

e. estimate the <u>average weight gap</u> between expected babies whose mothers only differ in that one has two drinks more than another. Is it <u>significant</u> (-ly different from zero) based on a 5% signif. level? <u>Illustrate your test</u> placing in the each of the figures below the relevant <u>critical</u> and <u>estimated values</u> (or statistic) and <u>fill in</u> the "?"



$\Delta\widehat{weight} \sim N(?,?)$ $\hat{\beta} \sim N(?,?)$ $t \sim N(0,1)$

f. <u>Test</u> with a 5% signif. level whether our multivariate model is preferable to a simpler <u>single-regressor</u> model that predicts **weight** with **visits**

3. (4 points) *We want to quantify the trade-off between time spent sleeping and working (**sleep** & **totwrk**, in minutes per week) controlling for **education** (which is measured in years and is included in logs), gender, and a dummy variable **kids** capturing if the worker has children less than 3-years-old. Using each of the following estimated models **A** to **D** and, <u>based on the heteroscedasticity-robust standard errors</u>, answer each question **a-d***

a. <u>interpret</u> 0.15 in regression **A** (*the number, not only the sign*) & state clearly the <u>units</u>

b. what is the <u>correlation</u> between ***totwrk*** & ***ln(educ)***? <u>Justify</u> using a <u>formula</u> in **A** & **B**

c. <u>rewrite</u> regression **C** (the <u>coefficients</u> and NOT their standard errors) for the case where we had included ***female*** (instead of ***male***) to account for gender in our model

d. <u>test</u> in **D** *assuming homoscedasticity* with a 5% significance level whether having under-3-years-old kids affects in any way our predictions of the amount of time spent sleeping

Dependent variable: minutes per week of ***sleep*** during night time ; **n=706**					
	Regressors	**(A)**	**(B)**	**(C)**	**(D)**
(1)	*totwrk*	-0.150 (.018)	-0.150 (.019)	-0.166 (.020)	-0.179 (.021)
(2)	*ln(educ)*		-170.6 (57.2)	-174.2 (57.1)	-166.9 (57.8)
(3)	*male*			90.96 (35.4)	89.17 (35.6)
(4)	*kids*				-239.7 (124.7)
(5)	*kids × totwrk*				0.107 (.055)
(0)	constant	3586 (42.0)	4011 (145.6)	4006 (145.4)	4017.6 (147)
R^2		0.1032	0.1130	0.1218	0.1273
F (test of the regression)		65.69	40.04	28.68	18.25

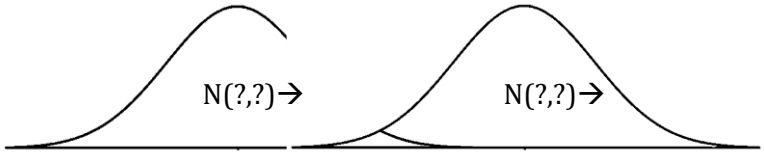
*NB: all workers have at least 1 year of education, thus ***educ*** ≥ 1 and ***ln(educ)*** ≥ 0

*Based on **the model with highest <u>adjusted R^2</u> among B, C & D,** answer questions **e** to **g**:*

e. *consider the case of **men** having **kids** less-than-3-year-old and 1 year of **education**.* What is the <u>estimated effect</u> on their ***sleep*** of a rise in their time devoted to work of 1 hour?

f. <u>test</u> using a 5% significance level whether the <u>effect</u> of ***totwrk*** on ***sleep*** in that model for a female worker without children and 1 year of education is <u>equal to</u> the one that regression **A** would capture for that same worker

g. compute a 95% confidence range for the <u>average sleep gap</u> between two workers (without children) when they only differ in that one works 30 minutes per week more than the other. Is it <u>significant</u> (-ly different from zero)? <u>Illustrate your test</u> below adding a figure and <u>fill in</u> the 6 "?"



MOCK FINAL EXAM 3b

Please, read the questions carefully and provide a complete & clear answer. Good luck!

1. Two proofs (2 points) from the last page

a. result XXX from the Appendix *using algebra & any of the other results listed before*

b. the proof YYY that we did in class, exercises…

2. (4 points) *The Regional Health Service (RHS) needs to quantify the effect on the health of a baby of an **extra visit** to the doctor of the mother during pregnancy. The baby's health is measured as his or her weight at birth (**weight**, in grams).*

a. *Which of these experiments does NOT ensure that changes in **weight** are due to an **extra visit to the doctor**; in other words, it does NOT measure a* causal effect*? Explain briefly*
i. Select a sample of first-time mothers and randomly provide extra prenatal care to half of them. (…)
ii. Select by random sampling some mothers and provide extra prenatal care to the half that was selected first. (…)
iii. Select a sample of very different mothers and randomly provide extra prenatal care to half of them. (…)
iv. Select by random sampling some mothers and provide extra prenatal care to those that bear their first child. (…)
(…) Then, compare the average weight at birth of the babies delivered by the two groups of mothers.

*Instead of an experiment, the RHS estimated with observational data a regression. The model allows for a quadratic relation between **weight** and the number of visits to the doctor during pregnancy (including, thus, **visit** and **visit²**). The regression also controls for **drink**: the average number of weekly alcoholic drinks that the mother takes during pregnancy (ranking from 0 to 8).*

Model 1: OLS, using observations 1-1651
Dependent variable: **weight**
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
(0) const	3102.36	98.4668	31.5067	<0.0001	***
(1) **visit**	36.2292	12.925	2.8030	0.0051	***
(2) **visit2**	−0.75307	0.4105	−1.8345	0.0668	*
(3) **drink**	−48.9098	25.4192	−1.9241	0.0545	*

Mean dependent var.	3409.520	S.D. dependent var.	575.1008
Sum squared resid	5.39e+08	S.E. of regression	572.0800
R-squared	0.012277	Adjusted R-squared	0.010478
F(3, 1647)	5.577219	P-value (F)	0.000833

Covariance matrix for the coefficients *NB: 2e-003 = 0.002*

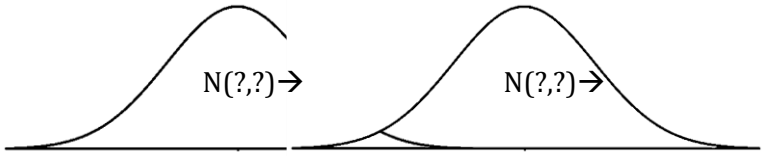
	(0) const	(1) visit	(2) visit2	(3) drink	
(0)	9695.70	-1229.76	33.8000	25.4714	**const**
(1)		167.055	-4.99249	-5.24652	**visit**
(2)			0.168510	0.0394247	**visit2**
(3)				646.138	**drink**

b. If you do not control for drinking habits of the mother, the new estimated model becomes $\widehat{weight} = 3103.5 + 36.2288\, visit - 0.7529\, visit^2$. What is the correlation between the number of prenatal visits and drinks in the data? Justify your guess

c. Test with a 5% significance level whether prenatal healthcare is a relevant factor in explaining the weight at birth of the baby. Give the hypotheses, statistic, its distribution, critical value and conclusion

d. What is the estimated effect on the babies' weight of a policy increasing the average prenatal care from 9 to 10 visits to the doctor?

e. What is now the estimated effect of a policy increasing the average prenatal care from 19 to 20 visits to the doctor? Construct a 95% confidence interval and illustrate it below filling in the 5 "?"



possible estimates of the effect

f. Is this effect (that of passing from 19 to 20 visits to the doctor) significant (-ly different from zero) at 5% level? Answer and locate this H_0 in the graph above

g. Is your previous answer to **f** coherent or consistent with that of (**c**) ? Explain briefly

3. (4 points) *We want to quantify the trade-off between time spent sleeping and working (**sleep** & **totwrk**, in minutes per week) controlling for **education** (which is measured in years and is included in logs), gender, and a dummy variable **kids** capturing if the worker has children less than 3-years-old. Using each of the following estimated models **A** to **D** and, based on the heteroscedasticity-robust standard errors, answer each question **a-d***

a. rewrite regression **A** for the case where ***sleep*** & ***totwrk*** were measured in *hours* per week instead of *minutes* (change ONLY the estimated betas, NOT the standard errors)

b. provide a plausible scatterplot for the observed values of ***sleep*** & ***educ*** (*without taking logs*!) for workers with similar values of ***totwrk*** according to regression **B**

c. interpret (using its value & units in your sentence) the estimated beta of ***male*** in **C**

d. *consider the case of **women** with 1 year of **education**.* Write and draw the two "straight lines" implicitly estimated that explain the trade-off between time spent sleeping and working for those women with young kids and without them

Dependent variable: minutes per week of ***sleep*** during night time ; **n=706**					
	Regressors	**(A)**	**(B)**	**(C)**	**(D)**
(1)	*totwrk*	-0.150 (.018)	-0.150 (.019)	-0.166 (.020)	-0.179 (.021)
(2)	*ln(educ)*		-170.6 (57.2)	-174.2 (57.1)	-166.9 (57.8)
(3)	*male*			90.96 (35.4)	89.17 (35.6)
(4)	*kids*				-239.7 (124.7)
(5)	*kids × totwrk*				0.107 (.055)
(0)	constant	3586 (42.0)	4011 (145.6)	4006 (145.4)	4017.6 (147)
R^2		0.1032	0.1130	0.1218	0.1273
F (test of the regression)		65.69	40.04	28.68	18.25

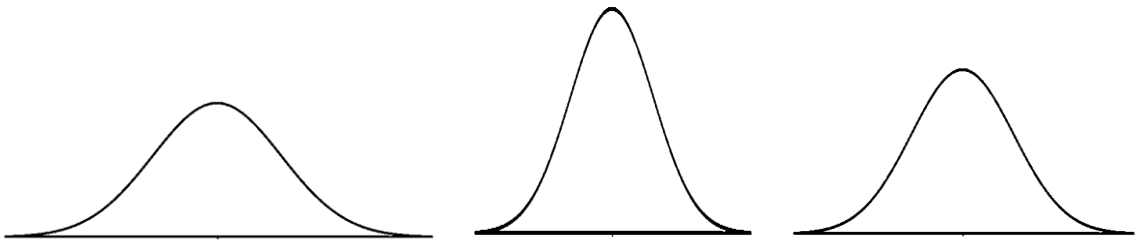
*NB: all workers have at least 1 year of education, thus ***educ*** ≥ 1 and ***ln(educ)*** ≥ 0

*Based now on **the most explanatory model** answer questions **e** to **g**:*

e. would you arrive to the same conclusion from the "test of the regression" in the table if you redo the test at 5% significance level *ignoring any heteroscedasticity* in the data?

f. test *assuming homoscedasticity* and using a 5% significance level whether having under-3-years-old kids has any impact on the amount of time spent sleeping

g. estimate the average sleep gap between two women (without children) when one works 1 hour per week less than another. Is it significant (-ly different from zero)? Illustrate your test placing in the each of the figures below the relevant critical and estimated values (or statistic) and fill in the "?"



$\Delta\widehat{sleep} \sim N(?,?)$ $\hat{\beta} \sim N(?,?)$ $t \sim N(0,1)$

RESEARCH PROJECT

(up to 10% extra on the total grade)

Our course in a nutshell: economists often want to measure a "causal effect", that is, the impact of one "variable of interest" on a "dependent" variable, leaving all other factors constant (e.g. the size of the class on test scores, gender on wages, study on grades…). Often they cannot do experiments (RCTs) where "assigning treatment randomly" would guarantee that other factors are kept fixed, i.e. $E(u|X) = 0$ holds. In that case, the difference in expected values, which can be estimated as $\widehat{\beta_1}$ in $Y = \beta_0 + \beta_1 Binary + u$, is interpreted as a causal effect. The same applies to the estimated $\widehat{\beta_1}$ in $Y = \beta_0 + \beta_1 X + u$.

Using instead observational data, where $E(u|X) = 0$ rarely holds, $\widehat{\beta_1}$ is a mix of the effect of our variable of interest and that of other omitted factors (i.e. $\widehat{\beta_1}$ is biased due to English knowledge, occupation, intelligence…). The solution is to include these factors $X_2 \dots X_k$ in the model and run $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$ so that $E(u|X_1, X_2, \dots, X_k) = 0$ becomes more plausible and, thus, $\widehat{\beta_1}$ can be interpreted as the causal effect of X_1 on *Y.*

Report 1: ANSWER by XXX these QUESTIONS about your PROJECT in 1 page

1. What is the causal effect that you want to estimate? What motivates your interest?

2. Describe the experiment (RCT) that would measure the causal effect of your interest?

3. You will NOT do an experiment. Instead you will use a sample of observational data, what is your source and "n"? Copy & paste a link, or describe briefly your data collection

4. What other factors (apart from your variable of interest) have an impact on your Y ?

5. Which of them could be correlated with your variable of interest causing bias in $\widehat{\beta_1}$? Which of them to you plan to include in your model as "variables of control": i.e., variables NOT included to estimate their causal effect BUT to avoid bias in the estimated effect of your variable of interest?

6. Is there any interesting hypotheses that could be tested with your model and data?

Report 2: by YYY, SUBMIT 1 or 2 pages with your RESULTS & COMMENTS

1. Copy the scatterplot of your dependent variable against your variable of interest. Discuss briefly whether there is heteroscedasticity in the data or non-linear effects

2. Report on a table your estimated single-regressor model targeting the causal effect of your interest. Report also 3 or 4 regressions more with *control* variables and define them

3. Discuss your findings regarding the issues that you raised on you report of November

4. Comment the goodness of fit of your 4 or 5 regressions, as well as any other issues that you did not anticipate in November and you came across while working with data: non-linear specifications, non-significant factors, internal & external validity, selection bias…

5. Answer briefly: did you like the project? why? what difficulties did you encounter?

EXAMPLE of REPORT 1 due on XXXX **Name:**

1. What is the causal effect that you want to estimate? What motivates your interest?

The effect of an extra student in the class on academic learning. It is interesting because, if this effect is large, governments can invest in teachers and improve educational outcomes

2. Describe the experiment (RCT) that would measure the causal effect of your interest?

Take a representative sample of students and divide them randomly into "treatment" group (those that I would force to go to large classes) and "control" group (those that I will force to study in small classes). Then, I would check if average scores on a standardized test are significantly different.

3. You will NOT do an experiment. Instead you will use a sample of observational data, what is your source and "n"? Copy & paste a link, or describe briefly your data collection

http://vincentarelbundock.github.io/Rdatasets/datasets.html Item: Caschool n=420

4. What other factors (apart from your variable of interest) have an impact on your *Y* ?

*Test scores are affected by the percentage of English learners in the district (**EL_pct**), the number of computers that students have (**comp_stu**) and other measures of expenditures (**expn_stu**). Of course, average income in the district (**avginc**) would also have an effect. There are other factors not included in the database: student's intelligence, education of parents, quality of kinder garden education, etc*

5. Which of them could be correlated with your variable of interest causing bias in $\widehat{\beta_1}$? Which of them to you plan to include in your model as "variables of control": i.e., variables NOT included to estimate their causal effect BUT to avoid bias in the estimated effect of your variable of interest?

***EL_pct** is positively correlated with **STR** but it lowers **test scores**, thus ignoring it should bias $\widehat{\beta_1}$ downwards. The factors **comp_stu, expn_stu** or **avginc** should push up **test scores** and be negatively correlated with **STR**, thus omitting them should bias $\widehat{\beta_1}$ downwards. I plan to include, at least, **EL_pct & comp_stu.** I will not control for factors not included in my database. Nonetheless, intelligence is probably not correlated with **STR** (thus, no bias) and the rest only slightly negatively correlated (generating a very minor downward bias).*

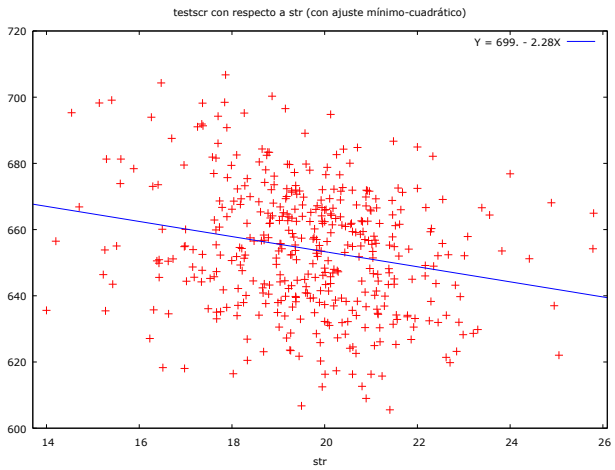
6. Is there any interesting hypotheses that could be tested with your model and data?

*I can test whether resources (**comp_stu** & **STR**) matter, as well as any non-linear effect of **STR** on **test scores**: whether my causal effect of interest depends on English knowledge of students, their access to computers, enrollment in the district, etc.*

EXAMPLE of <u>REPORT 2</u> due on YYY **Name:**

1. The goal of my study is to measure the impact of the size of the class (measured as the *STR*: the average number of students per teacher) in academic learning (measured by average scores on a standardized test at the end of Primary school). I obtain data for 420 school districts in California whose *STR* ranges in 14 to 25 and *test scores* from 600 to 710.

First, I look at the **scatterplot** of *test scores* against *STR*: I observe a negative impact of *STR* on *test scores*. In particular the single-regressor model estimates that every extra student in class will lower test scores in 2.28 points, on average (see table). Although the effect is statistically significant at 1%, *STR* only explains 5% of changes in *test scores*. There are no signs of **heteroscedasticity** or **non-linear** effects of *STR* on *test scores*.



2. The table below shows 5 regression estimates resulting in a refinement of my analysis

Dependent variable: average district score on a standardized test, i.e. ***test scores*** ;					**n=420**
Regressors	***(A)***	***(B)***	***(C)***	***(D)***	***(E)***
1. ***STR***	-2.28***(.52)	-1.10***(.43)	-0.85**(.43)	-1.36***(.46)	-1.02***(.53)
2. ***EL_pct***		-0.65***(.03)	-0.63***(.03)	-0.69***(.03)	-0.69***(.03)
3. ***comp_stu***			27.27**(12.6)	30.83**(12.5)	31.07**(12.5)
4. ***Henroll***				8.15***(1.6)	41.42**(19.5)
5. ***Henroll×STR***					-1.65***(.95)
0. ***constant***	698.9***(10.4)	686.0***(8.73)	677.1***(9.20)	684.9***(9.47)	678.2***(10.8)
R^2	0.05	0.43	0.43	0.47	0.47
$\bar{R}^2$	0.05	0.43	0.43	0.46	0.46

STR = # of students per teacher (average for each district)

EL_pct = percentage of students that are English learners (average for each district)

comp_stu = # of computers per student (average for each district)

Henroll is binary: =1 if the district has a number of students enrolled above-average, and =0, otherwise

Henroll×STR is an interaction term that captures the difference between the effect of *STR* on *test scores* when *Henroll=1* and this effect when *Henroll=0*

3. **Regression A** is the single-regressor model and has already been discussed.

Regression B controls for the percentage of English learners. My estimated effect in regression **A** was downward biased capturing two negative effects: the "pure" effect of larger classes (on average, 1.10 points less per extra student) and the impact of more students that are learning English (on average, 0.65 less points per percentage point). Both factors are positively correlated. The R^2 grows notably showing better fit of the data.

Regression C adds the (average) number of computers per student in the district as a control variable. *comp_stu* has a positive effect on *test scores* (27.27 points more per extra computer per student, on average) and it is negatively correlated with *STR*, thus omitting this factor in model **B** caused downward bias on my estimated causal effect. The R^2 grows very little meaning that most of the information of *comp_stu* was already captured by *STR* & *EL_pct*. Nonetheless the adjusted R^2 rises from 0.42 in **B** to 0.43 in **C**.

Regression D adds a binary variable *Henroll* that identifies districts with many students enrolled (large cities, very populated areas) from districts with below-average enrollment. Controlling for this omitted factor lowers my estimated effect of *STR* on *test scores*, thus previous estimates were upward biased. This means that districts with many students enrolled (*Henroll=1*) are typically those with higher *STR* (plausible). The R^2 & adjusted R^2 keep rising implying that regression **D** fits better than **C** the data on *test scores*.

Finally, given the remarkable change of the estimated effect of *STR* on *test scores* from model **C** to **D**, **regression E** considers possible non-linear effects of *STR* on *test scores.* In particular, it captures whether this effect is different in districts with many enrolled students (*Henroll=1*) and in those with a low enrollment. To do so I add the interaction term *STR*Henroll* whose coefficient equals -1.65 meaning that, when the district has many students enrolled, *STR* lowers *test scores* 1.65 points per extra student more than when the district has below-average enrollment. This coefficient is significant at 10% level. Thus, the estimated negative effect of *STR* on *test scores* when the district does not have many students enrolled is 1 point for each extra student per teacher. However, if the district has many students enrolled, this effect is -2.65 points for each extra student. The R^2 & adjusted R^2 increase so little that there are no differences after rounding to decimals.

4. Given the small difference between the adjusted R^2 of models **C** & **E**, it is worth testing in **E** if the level of enrollment in districts matters at all; H_0: $\boldsymbol{\beta_4 = \beta_5 = 0}$ vs H_1: $either\ \beta_4 \neq 0, or\ \beta_5 \neq 0, or\ both$. I estimate F=13.85 (>3) thus I reject and conclude that enrollment matters. I can easily test in model **E** if those factors that depend on educational resources (mainly, teachers & computers, *STR & comp_stu*) matter; H_0: $\boldsymbol{\beta_1 = \beta_3 = \beta_5 = 0}$ vs H_1: $any\ \neq 0$. I estimate F=8.28 (>2.6) thus I reject and resources matter. My refined model **E** is able to replicate 47% of the variability in test scores: not bad!

5. I consider that playing around with data, interpreting the output of regressions, and thinking whether it make sense to me is **fun**. I hope you liked it too! Best wishes!

BIBLIOGRAFÍA

Stock, J. y Watson, M (2012). Introducción a la Econometría (3ª edición). Editorial Pearson

Wooldridge, J.M. (2013). Introducción a la econometría. Un enfoque moderno. Editorial CENGAGE Learning

AGRADECIMIENTOS

Al equipo que mantiene el programa libre GRETL
www.gretl.sourceforge.net/es

A mis alumnos

REFERENCES

Stock, J. y Watson, M (2015). Introduction to Econometrics (updated 3rd Global edition). Pearson Education Limited

Wooldridge, J.M. (2009). Introduction to econometrics. A modern approach. (4th edition). South-Western, a part of CENGAGE Learning

THANKS

To the team maintaining the free software GRETL
www.gretl.sourceforge.net

To my students