

Soluciones al boletín de problemas 4

Capítulo 18

1. 20 y 25

5. (i) histograma de la suma. Tiende a convertirse en una curva normal.
(ii) histograma del producto.
(iii) histograma de los números que se van a retirar.

Capítulo 20.

5. peso medio por invitado : 150 lbs.

$$4 \text{ tons} = 8000 \text{ lbs.}$$

$$50 \times 150 = 7500 \text{ lbs.}$$

$$SE = 35 \times \sqrt{50} = 247,5$$

$$\therefore 7500 \pm 2 SE = 7500 \pm 2(247,5) = 7995 \text{ y } 7005. \text{ *SE = Error estándar.}$$

Y el rango entre 7005 lbs. y 7995 lbs. abarca más del 95,45 % de la suma de los pesos de las 50 personas seleccionadas. Por tanto, al ser el porcentaje del grupo 8000 lbs., se halla muy a la derecha de la curva, aproximadamente **2,275**. ($100 - 95,45 = 4,55$, $4,55/2 = 2,275$)

6. (ii) El tamaño de la muestra es un porcentaje del 0,1 de la población total de cada estado. Para California, el tamaño de la muestra es 30.000, y para Nevada es 1.000. Con un tamaño mayor, es de esperar una mayor exactitud en California que en Nevada.

8. Población total : 30.000 Total demócratas : 12.000

$$\Pr(\text{Demócratas}) = 12.000/30.000 = 0,4$$

Una posibilidad de 50-50 implica que la distribución teórica del muestreo es simétrica, ya que ésta es simétrica alrededor de la media estimada.

$$\therefore E(\text{Demócratas en la muestra}) = 0,4 \times 1.000 [\Pr(\text{Dem}) \times \text{tamaño muestra}] = \mathbf{400}.$$

Capítulo 21.

1. Se supone que un 15,8 % del total de los hogares estadounidenses dispone de computadora. Por tanto, $E(\text{hogares con computadora en una ciudad de 25.000 habitantes}) = 25.000 \times 0,158 = 3.950$.

a. Para calcular la media y el error estándar (SE) de la media:

$$79/500 = 0,158 \text{ (15,8 \%)}. SE = [\sqrt{(0,158) \times (1-0,158)}] / \sqrt{(500)} = 0,365 / \sqrt{(500)} = 0,0163 \text{ (1,63 \%)}.$$

\therefore El porcentaje estimado de hogares con computadora en la ciudad es un **15,8 %** : el margen de error estaría en torno a un **1,63 %**.

b. CI (intervalo de confianza) = $0,158 \pm 2 \times 0,0163 = 0,1906$ y $0,1254$.

Por tanto, los intervalos de confianza son 12,54 % y 19,06%.

2. $\Pr(\text{hogares de la muestra que disponen de nevera}) = 498/500 = 0,996$ (99,6).

$$SE = [\sqrt{(0,996) \times (1-0,996)}] / \sqrt{(500)} = 0,00282 \text{ (0,282\%)}$$

a. El porcentaje estimado de hogares que disponen de nevera es un **99,6 %**; el margen de error estaría en torno a un **0,282 %**.

b. $CI = 0,996 \pm 2 \times 0,00282 = 0,1,00164$ y $0,99036$. El límite superior del intervalo de confianza es mayor que el 100 %. En este caso no es posible crear el intervalo superior, pero el límite inferior del intervalo es el 99,036 %.

12. (i) irrelevante

(ii) un histograma de los números retirados.

(iii) un histograma de probabilidad para la suma.

14. tamaño de la muestra = 1.500.

Pr(arrendadores en la ciudad que refleja la muestra) = $1035/1500 = 0,69$ (69 %).

E(arrendadores en la muestra) = 0,69.

SE(de arrendadores en la muestra) = $[\sqrt{(0,69)(1-0,69)}] / \sqrt{1500} = 0,012$ (1,2%).

a. El valor previsto del porcentaje de personas de la muestra que alquilan inmuebles es **exactamente** 69%.

*nota: se nos pide hallar el valor y el SE previstos de la muestra, no la población que podemos estimar a partir de ésta. Por consiguiente, todos los valores son idénticos a los de las cifras calculadas a partir de la muestra.

b. El SE del porcentaje de personas de la muestra que alquilan **calculado a partir de los datos** es un 1,2 %.

Capítulo 23.

10. tamaño de la población = 80.000

SD (desviación estándar) = 1,75.

tamaño de la muestra = 625

media de personas por hogar = 2,30.

a. Verdadero.

SE = $1,75 / \sqrt{625} = 0,07$

b. Falso.

No tiene sentido calcular el intervalo de confianza de la muestra. Calculamos éste para comprobar que los cálculos caen dentro del rango de la población.

c. Verdadero.

$2,30 \pm 2 \times 0,07 = 2,44$ y $2,16$.

d. Falso.

No es más que la interpretación errónea de un intervalo de confianza.

e. Falso.

El teorema del límite central sostiene que, si se repite el diseño de la muestra de la población, las medias de la muestra toman la forma de una curva normal.

f. Verdadero.

Ya explicado más arriba.

12. 400 no es una muestra, sino el tamaño de una población. El intervalo de confianza se usa para confirmar la exactitud de las estimaciones obtenidas a partir de una muestra. Por lo tanto, en este caso, el intervalo de confianza carece de sentido.

Capítulo 26.

2. Pr(números rojos) = $18/38 = 0,474$

tamaño de la muestra = 3800 números rojos en la muestra = 1890.

Pr(números rojos en la muestra) = $1890/3800 = 0,497$

a. H_0 : Pr (números rojos) = 0,474

* interpretación : la diferencia entre 0,474(población) y 0,497(muestra) se debe a un error de probabilidad. O BIEN, 0,479 se obtiene por un error casual.

H_1 : Pr (números rojos) > 0,474

* interpretación : la diferencia entre 0,474(población) y 0,497(muestra) no se debe a un error casual sino a un efecto sistemático.

- b. $Z = (0,497 - 0,474) / SE$
 $SE = SD / \sqrt{3800} = [\sqrt{(0,474) \times (1-0,474)}] / \sqrt{(3800)} = 0,0081$
 $\therefore Z = (0,497368 - 0,473684) / 0,0081 = 2,924$
p-valor = $1 - 0,99825 = 0,00175$. (menos de 0,05; 5 % de nivel de significancia).
- c. Tanto el resultado de Z como el p-valor indican demasiados rojos, y no es un error casual.

4. población = 900 estudiantes ; media final = 63 & SD (desviación estándar) = 20
sección a = 30 estudiantes; media final = 55

H_0 = la media del final es igual a 63

H_1 = la media del final no es igual a 63

$SE = 20 / \sqrt{30} = 3,651$

$Z = (55 - 63) / 3,651 = - 2,19$

p-valor = 0,0139

\therefore Tanto el resultado de Z como el p-valor muestran que la diferencia entre la media de población y la media de la muestra no se debe a un error casual. La sección de este profesor auxiliar presenta un rendimiento por debajo de la media.

6. convocatoria = 350 ; mujeres = 102. $\Pr(\text{mujeres en la convocatoria}) = 102/350 = 0,2914$.
grupo del jurado = 100 ; mujeres en el grupo = 9. $\Pr(\text{mujeres en el jurado}) = 9/100 = 0,09$.
Sin embargo, una mayoría (es decir, más de la mitad) de las personas con capacidad para formar parte del jurado en el distrito eran mujeres. ¿Es una selección adecuada?

- a. media = 0,2914 ; y suponemos que (al menos) el 50% de la población son mujeres. $SE = [\sqrt{(0,5) \times (1-0,5)}] / \sqrt{(350)} = 0,0267$.

$Z = (0,2914 - 0,5) / 0,0267 = -7,6142$

p-valor = 0,0000...1

Por tanto, la escasa representación de mujeres en la selección de la convocatoria para el jurado no se debe a un error casual; se debe a que se ha hecho algo mal.

- b. $E(\text{mujeres en el jurado}) = 0,2914 \times 100 = 29,14$. Como 102 de las 350 personas de la convocatoria son mujeres, debería haber 29 mujeres en el jurado. Número real = 9 (0,09)

$SE = [\sqrt{(0,2914) \times (1-0,2914)}] / \sqrt{(100)} = 0,0454$

$Z = (0,09 - 0,2914) / 0,0454 = - 4,4361$

p-valor = 0,001

Nuevamente, la escasa representación de mujeres en el jurado es estadísticamente significativa.

- c. Conclusión: algo se está haciendo mal. Es muy poco probable que una elección de jurado de este tipo se produzca por casualidad.

7. total de pacientes en un mes = 1022

días impares : 580 días pares : 442

debería estar dividido por igual y mostrar una tasa de entradas de 50-50 si no hubiera errores.

$\Pr(\text{días pares en la muestra}) = 580/1022 = 0,5675$

$\Pr(\text{previsto(días impares)}) = 0,05$

$SE = [\sqrt{(0,5) \times (1-0,5)}] / \sqrt{(1022)} = 0,0156$

$Z = (0,5675 - 0,5) / 0,0156 = 4,32$

p-valor = 0,0008

Por los resultados de Z y del p -valor se deduce que hay más gente que acude al hospital en días impares. No podemos estar, pues, de acuerdo con el observador que lo considera como una cuestión de cara o cruz.

Capítulo 29.

1. (a) Verdadero. Aunque la diferencia sea muy significativa (p.ej., $p = 0,01$), sigue existiendo la posibilidad (si bien muy remota) de que dicha diferencia se deba a un error casual. Esto es precisamente lo que indica el p -valor.
(b) Falso. El que un número sea estadísticamente significativo no sólo depende del número real, sino también del tamaño de la muestra.
(c) Puede ser tanto verdadero como falso. Un p -valor de 0,047 y otro de 0,052 son magnitudes muy parecidas, pero su tratamiento puede variar. Por ejemplo, cuando un investigador fija el valor crítico en 0,05 (que es casi siempre), la estimación con p -valor 0,052 no es significativa, y la hipótesis nula no se debería rechazar, mientras que la estimación con p -valor 0,047 se considera estadísticamente significativa y la hipótesis nula sí se debería rechazar.
2. (i) ¿Se debe la diferencia a la casualidad?
El contraste de hipótesis tiene por objeto comprobar si la diferencia entre los valores previstos y los observados se debe a la casualidad. Por ello, los resultados de Z son diferencias (intuitivamente) normalizadas y los p -valores representan la probabilidad de que los resultados de Z normalizados puedan deberse al azar. En apariencia, cuanto más pequeño sea un p -valor menor será la probabilidad de que la diferencia se deba a un error casual.
3. promedio por caja = 50
 X_1 : tamaño de la muestra = 100, $SE = SD / \sqrt{100} = 10/10 = 1$
 X_2 : tamaño de la muestra = 300, $SE = SD / \sqrt{900} = 10 / 30 = 0,3333$
La afirmación es FALSA. Los resultados de Z y los p -valores no sólo dependen de las diferencias de promedios, sino también de errores estándar. La muestra del investigador 2 es de mayor tamaño, y da como resultado diferentes errores estándar en ambos investigadores. Por tanto, el investigador cuyos resultados de Z (no el promedio) se alejen más de 0 obtendrá el valor más pequeño, lo que podría ser el caso del investigador 2.
6. $\beta = 0,07$; $SE = 0,05$
 $Z = 0,07 / 0,05 = 1,4$
Aunque no hemos fijado el valor crítico, la opinión más extendida nos dice que $Z = 1,96$ y p -valor $\leq 0,05$ son los valores límite para la significancia estadística. Aquí, el resultado de Z no es significativo con respecto al p -valor = 0,05, lo que confirma que "no hay impacto". Sin embargo, si fijamos un valor límite mayor que 0,05, como $p = 0,01$, llegamos a una conclusión totalmente distinta: el impacto sí es estadísticamente significativo. Por tanto, para ser precisos, podemos concluir que es más posible que exista una relación positiva entre inflación y comportamiento electoral, aunque la proporción real de esa influencia no está calculada con precisión.
8. empleo femenino en los Estados Unidos = 50,4 % en 1985.
empleo femenino en los Estados Unidos = 54,1 % en 1993.
 - a. Se nos pregunta si el índice de empleo femenino en los Estados Unidos entre 1985 y 1993 es estadísticamente significativo. Aunque se basen en estudios de población, si los índices de

- a. $t = (X - 0) / 1 = 2,09 = 0,025$.
- b. 0,05
- c. $t = 2,09$
- d. $t = 2,85$.

5. $X \sim T(3, 2,25)$

$$t = (X - 3) / \sqrt{2,25} \quad \text{d.f.} = 20$$

a. $\Pr(X > 1,155) = (1,155 - 3) / \sqrt{2,25} = -1,845 / 1,5 = -1,23$.

Conforme a la t-tabla, el área cubierta por encima de $-1,23$ con un d.f. de 20 ronda el 85 %.

b. $(X - 3) / 1,5$ con un d.f. de 20, para cubrir el 99 %, $\pm t$ deberá ser 2,85.

Nota: la ecuación para calcular los resultados de z o de t es:

$$\frac{X - \text{media}}{\text{SE}}$$

Generalmente, $(X - \text{media})$ se calcula en términos absolutos. El orden no importa en los test de dos colas, pero si el test es de una cola, hay que asegurarse de que el orden no sea $(\text{media} - X)$. No es un problema grave si tiene la intuición adecuada (ya que se puede convertir al contexto de una distribución normal), pero puede dar lugar a confusiones.

Parte III.

Pregunta A

1. La primera parte del problema consiste en efectuar la regresión múltiple que prevea la elección de dormitorio.

```
. reg firstchoice yearbuilt roomsize
```

Source	SS	df	MS			
Model	3963.17801	2	1981.58901	Number of obs =	10	
Residual	6530.42199	7	932.917427	F(2, 7) =	2.12	
				Prob > F =	0.1901	
				R-squared =	0.3777	
				Adj R-squared =	0.1999	
Total	10493.60	9	1165.95556	Root MSE =	30.544	

firstchoice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearbuilt	.9285734	.8766258	1.06	0.325	-1.144317	3.001464
roomsize	-.1171777	.688022	-0.17	0.870	-1.744091	1.509736
_cons	-1717.581	1616.999	-1.06	0.323	-5541.176	2106.013

2. En la segunda parte se pide ejecutar los dos componentes bivariados de la parte (1).

```
. reg firstchoice yearbuilt
```

Source	SS	df	MS			
Model	3936.11796	1	3936.11796	Number of obs =	10	
Residual	6557.48204	8	819.685255	F(1, 8) =	4.80	
				Prob > F =	0.0598	
				R-squared =	0.3751	
				Adj R-squared =	0.2970	
Total	10493.60	9	1165.95556	Root MSE =	28.63	

firstchoice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearbuilt	.7945975	.3626074	2.19	0.060	-.0415767	1.630772
_cons	-1473.848	705.5833	-2.09	0.070	-3100.926	153.2298

```
. reg firstchoice roomsize
```

Source	SS	df	MS			
Model	2916.41791	1	2916.41791	Number of obs =	10	
Residual	7577.18209	8	947.147761	F(1, 8) =	3.08	
				Prob > F =	0.1174	
				R-squared =	0.2779	
				Adj R-squared =	0.1877	
Total	10493.60	9	1165.95556	Root MSE =	30.776	

firstchoice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
roomsize	.5368167	.3059215	1.75	0.117	-.1686396	1.242273
_cons	-5.423699	45.29416	-0.12	0.908	-109.8722	99.02482

La comparación entre el supuesto bivariante y el multivariante muestra indicios de que existe algún tipo de error en éste último. Los errores estándar son demasiado amplios en la regresión multivariante en comparación con la bivariante y los coeficientes muestran cambios muy sustanciales, hasta el punto de que incluso la variable del tamaño de la habitación (roomsized) tiene distinto signo. Ello parece indicar la existencia de multicolinealidad (dos variables independientes que miden el mismo factor subyacente). En este caso, cabe la posibilidad de que, cuanto más moderno sea el edificio, más grandes sean las habitaciones, como respuesta a las preferencias expresadas por los estudiantes durante años. Para comprobarlo, veamos la siguiente regresión:

```
. reg roomsized yearbuilt
```

Source	SS	df	MS	Number of obs = 10		
Model	8149.61788	1	8149.61788	F(1, 8)	=	33.08
Residual	1970.78212	8	246.347765	Prob > F	=	0.0004
				R-squared	=	0.8053
				Adj R-squared	=	0.7809
Total	10120.40	9	1124.48889	Root MSE	=	15.695

roomsized	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearbuilt	1.143357	.1987867	5.75	0.000	.6849536	1.60176
_cons	-2080.029	386.8112	-5.38	0.001	-2972.017	-1188.041

Resulta evidente que el tamaño de las habitaciones es mayor en los edificios más antiguos. Más del 80% de la varianza en el tamaño se explica por el año de construcción, (yearbuilt) lo que demuestra que, efectivamente, hay un problema de multicolinealidad.

3. ¿Qué modelo es el mejor? Evidentemente, el supuesto multivariante es el que no se debe emplear, según lo descrito en el apartado(2). Dado que parece claro que el tamaño de la habitación es una función de la modernidad del edificio, y que los edificios más nuevos tienen también otras ventajas, el mejor modelo será posiblemente el del caso bivariante, usando el año de construcción (yearbuilt) como variable independiente.

Aparte de las razones teóricas para emplear el caso bivariante con la variable yearbuilt, este modelo es también el único que presenta un coeficiente estadísticamente significativo y el valor R² más alto (0,37), lo que refuerza la confianza en esta solución.

Es una lástima, no obstante, desperdiciar todos los datos aportados por la variable del tamaño de las habitaciones. Si la regresión formaba parte de un estudio más amplio en el que era importante controlar las características físicas del edificio, pero sólo como medio para eliminar el sesgo de las variables omitidas, una posible solución sería crear una escala que combinara las variables "yearbuilt" y "roomsized". Para ello, restaríamos cada variable de su media, dividiéndola por su desviación estándar y sumando a continuación todos los resultados de Z. Tendríamos entonces una medida sin unidad de la "calidad del edificio" capaz de predecir la primera preferencia (firstchoice) mejor que cualquier variable por sí sola. En este caso resulta que no se produce una mejora efectiva al crear una variable combinada de estas características (véase más abajo), sino que más bien parece tratarse de un conjunto de variables colineales.

```
summ yearbuilt roomsized
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yearbuilt	10	1945.7	26.31877	1910	1981
roomsize	10	144.6	33.5334	97	200

```
. gen zyearbuilt=(1945.7-yearbuilt)/26.31877
```

```
. gen zroomsize=(144.6- roomsize)/33.5334
```

```
. gen quality= zyearbuilt+ zroomsize
```

```
. reg firstchoice quality
```

Source	SS	df	MS	Number of obs =	10
Model	3591.4989	1	3591.4989	F(1, 8) =	4.16
Residual	6902.1011	8	862.762637	Prob > F =	0.0756
				R-squared =	0.3423
				Adj R-squared =	0.2600
Total	10493.60	9	1165.95556	Root MSE =	29.373

firstchoice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
quality	-10.25477	5.026131	-2.04	0.076	-21.84505 1.335507
_cons	72.2	9.288502	7.77	0.000	50.78068 93.61932

Una vez más, soy partidario de aplicar la regresión bivalente con la variable "yearbuilt". A partir de esta regresión variable combinada, vemos que la t-estadística y R² se han hecho más pequeñas con respecto a "yearbuilt".

Pregunta B

```
. reg cvote82 pvote80 newtown;
```

Source	SS	df	MS	Number of obs =	64
Model	.429321715	2	.214660857	F(2, 61) =	111.19
Residual	.117767969	61	.001930622	Prob > F =	0.0000
Total	.547089684	63	.008683963	R-squared =	0.7847
				Adj R-squared =	0.7777
				Root MSE =	.04394

cvote82	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pvote80	1.088667	.08096	13.45	0.000	.9267773 1.250556
newtown	-.0674145	.0110094	-6.12	0.000	-.0894291 -.0454
_cons	-.0025771	.057215	-0.05	0.964	-.1169856 .1118314

Courter obtuvo peores resultados en las ciudades "nuevas" por un porcentaje de voto del 0,067. Es decir, perdió un porcentaje de votos del 0,067 en estas ciudades.

```
. reg cvote82 pvote80;
```

Source	SS	df	MS	Number of obs =	64
Model	.35693151	1	.35693151	F(1, 62) =	116.38
Residual	.190158174	62	.003067067	Prob > F =	0.0000
Total	.547089684	63	.008683963	R-squared =	0.6524
				Adj R-squared =	0.6468
				Root MSE =	.05538

cvote82	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pvote80	1.100501	.102014	10.79	0.000	.896578 1.304424
_cons	-.0466514	.0715417	-0.65	0.517	-.1896611 .0963583

```
. bys newtown : reg cvote82 pvote80;
```

```
-> newtown = 0
```

Source	SS	df	MS	Number of obs =	30
Model	.028806365	1	.028806365	F(1, 28) =	24.45
Residual	.032991833	28	.00117828	Prob > F =	0.0000
Total	.061798198	29	.002130972	R-squared =	0.4661
				Adj R-squared =	0.4471
				Root MSE =	.03433

cvote82	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pvote80	.7077853	.1431468	4.94	0.000	.4145625 1.001008
_cons	.2639357	.1003594	2.63	0.014	.0583587 .4695127

```
-> newtown = 1
```

Source	SS	df	MS	Number of obs =	34
Model	.33065637	1	.33065637	F(1, 32) =	142.20
Residual	.074410661	32	.002325333	Prob > F =	0.0000
Total	.405067031	33	.012274759	R-squared =	0.8163
				Adj R-squared =	0.8106
				Root MSE =	.04822

cvote82	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pvote80	1.181061	.0990436	11.92	0.000	.9793155 1.382806
_cons	-.1343421	.0694759	-1.93	0.062	-.2758598 .0071755

Si nos fijamos en la primera regresión, vemos que el partidismo fue un factor global decisivo. La regresión muestra que Courter recibió un 1,1 % más de votos cuando el porcentaje de votos a favor de Reagan en una ciudad aumentaba en un 1 %. Sin embargo, la desviación disminuye en los distritos nuevos y aumenta en los ya existentes. En un distrito nuevo, Courter recibió un 1,18 % más de votos cuando el voto a Reagan de los electores de ese distrito se incrementaba en un 1 %. En los distritos ya existentes, el efecto partidista se atenuaba en 0,708 %. De donde se deduce que el partidismo tiene mayor efecto en ciudades "nuevas". Los votos obtenidos por Courter en las otras ciudades se deben a otra razón; concretamente a las ventajas de ostentar el poder y ser más conocido.

```
. gen newt_p = newtown*pvote80;
```

```
. reg cvote82 pvote80 newtown newt_p;
```

Source	SS	df	MS	Number of obs =	64
Model	.43968719	3	.146562397	F(3, 60) =	81.88
Residual	.107402494	60	.001790042	Prob > F =	0.0000
Total	.547089684	63	.008683963	R-squared =	0.8037
				Adj R-squared =	0.7939
				Root MSE =	.04231

cvote82	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pvote80	.7077853	.1764367	4.01	0.000	.3548594 1.060711
newtown	-.3982779	.1379026	-2.89	0.005	-.6741242 -.1224315
newt_p	.4732753	.1966758	2.41	0.019	.0798653 .8666854
_cons	.2639357	.1236988	2.13	0.037	.0165013 .5113702

$$cvote\ 82 = \beta$$

$$cvote\ 82 = \beta_0 + \beta_1\ pvote80 + \beta_2\ newtown + \beta_3\ pvote80 * newtown + \varepsilon$$
 para una ciudad nueva: pendiente = $\beta_1 + \beta_3$ & intercepción = $\beta_0 + \beta_2$
 para una ciudad ya existente: pendiente = β_1 & intercepción = β_0

Al incluir los términos de interacción se tienen en cuenta los distintos efectos del partidismo en ambos tipos de ciudades, lo que da el mismo resultado que si se hubieran efectuado dos regresiones por separado.

El resultado proporciona los mismos coeficientes que en las dos regresiones anteriores, lo que confirma el mayor peso del partidismo en las ciudades nuevas y del mandato en las ya existentes.

Pregunta C

```
. reg rate93q totfac totstu;
```

Source	SS	df	MS	Number of obs =	109
Model	34.7971203	2	17.3985601	F(2, 106) =	45.02
Residual	40.9693812	106	.386503596	Prob > F =	0.0000
				R-squared =	0.4593
				Adj R-squared =	0.4491
Total	75.7665015	108	.70154168	Root MSE =	.62169

rate93q	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
totfac	.0198392	.0054385	3.65	0.000	.0090569	.0306215
totstu	.0054181	.001323	4.10	0.000	.0027951	.0080412
_cons	1.964446	.1103758	17.80	0.000	1.745615	2.183276

```
. corr rate93q totfac totstu, cov;
(obs=109)
```

	rate93q	totfac	totstu
rate93q	.701542		
totfac	7.55772	217.865	
totstu	31.7926	597.155	3681.26

$b_1 = Cov(X_1 Y) / Var(X_1) - b_2 Cov(X_1 X_2) / Var(X_1)$
 convertir esta fórmula a:

$$Cov(X_1 Y) / Var(X_1) = b_1 + b_2 Cov(X_1 X_2) / Var(X_1)$$

Aquí, b_1 es un efecto directo y $b_2 Cov(X_1 X_2) / Var(X_1)$ es un efecto indirecto (la misma lógica para b_2).

Introduciendo las cifras de la tabla varianza-covarianza, obtenemos las siguientes respuestas:

$$0,034689 = 0,0198392 + 0,0054181 \times (597,155/217,865)$$

$$0,008636 = 0,0054181 + 0,0198392 \times (597,155/3681,26)$$

	Efecto bruto	Efecto directo	Efecto indirecto
totfac	0,034689	0,0198392	0,01485
totstu	0,008636	0,0054184	0,0032182

Parte C.2.

He utilizado las siguientes variables:

pub_fac : ratio entre el número total de publicaciones del programa en el periodo 1988-1992 y el número de facultades adscritas a él. He supuesto que, si el programa resulta efectivo, la ratio entre ambas magnitudes será alta.

myd : lapso de tiempo medio (en años) desde que se inicia el posgrado hasta que se obtiene el título. Se trata de una mediana de los varios títulos que se conceden en el año, distribuidos proporcionalmente durante ese año (un dato que considero importante). Para que resulte eficaz (económico y productivo), el programa debería reducir el tiempo de obtención del posgrado.

suppfac : porcentaje de facultades adscritas al programa que recibieron ayuda a la investigación entre 1988 y 1992. La calidad y eficacia del programa depende de este tipo de apoyo, tanto institucional como externo.

fac_stu : he creado la variable de ratio facultad - estudiante (**fac_stu**) dividiendo el número total de facultades entre el número total de estudiantes: $gen\ fac_stu = totstu/totfac$

```
. reg rate93e pub_fac myd fac_stu suppfac;
```

Source	SS	df	MS	Number of obs = 109		
Model	36.9698165	4	9.24245412	F(4, 104)	=	28.09
Residual	34.2244696	104	.329081438	Prob > F	=	0.0000
-----				R-squared	=	0.5193
Total	71.194286	108	.659206352	Adj R-squared	=	0.5008
-----				Root MSE	=	.57366

rate93e	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pub_fac	.0546024	.0173104	3.15	0.002	.0202753	.0889295
myd	-.1197654	.0336794	-3.56	0.001	-.1865528	-.0529779
fac_stu	.0671608	.0320218	2.10	0.038	.0036604	.1306611
suppfac	.0189985	.003511	5.41	0.000	.0120362	.0259609
_cons	2.413536	.3219996	7.50	0.000	1.774999	3.052073

Naturalmente, siempre hay que fijarse en las gráficas bivariantes. Se adjuntan del siguiente modo:

Archivo	Gráfico de:
PSet4-C3-pub_fac	rate93e and pub_fac
PSet4-C3-myd	rate93e and myd
PSet4-C3-fac_stu	rate93e and fac_stu
PSet4-C3-suppfac	rate93e and suppfac

pub_fac and myd tienen un número altísimo de valores atípicos, lo que se debe posiblemente a un error de entrada de datos. Si lo omitimos, tenemos:

Archivo	Gráfico de:
PSet4-C3-NEWpub_fac	rate93e and pub_fac w/ outlier omitted
PSet4-C3-NEWmyd	rate93e and myd w/ outlier omitted

```
. reg rate93e pub_fac myd fac_stu suppfac
```

Source	SS	df	MS	Number of obs = 108		
Model	37.0885633	4	9.27214082	F(4, 103)	=	28.02
Residual	34.0812883	103	.330886295	Prob > F	=	0.0000
Total	71.1698516	107	.6651388	R-squared	=	0.5211
				Adj R-squared	=	0.5025
				Root MSE	=	.57523

rate93e	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pub_fac	.0706044	.0298839	2.36	0.020	.0113367	.1298721
myd	-.0974805	.047835	-2.04	0.044	-.1923499	-.0026111
fac_stu	.059939	.0339344	1.77	0.080	-.0073618	.1272399
suppfac	.0180133	.0038259	4.71	0.000	.0104256	.0256011
_cons	2.2304	.4263319	5.23	0.000	1.384872	3.075929

Al aumentar la ratio entre la publicación y el porcentaje de facultades adscritas y el apoyo a la investigación, la eficacia del programa aumenta hasta niveles estadísticamente significativos. Asimismo, al aumentar el tiempo medio de permanencia de cada estudiante de posgrado en el programa, la eficacia de éste disminuye, lo que implica que un programa más efectivo haría que los estudiantes obtuvieran su título en menos tiempo. Vemos que la parte de cada estudiante en la facultad deja de ser significativa al nivel de 0,05 una vez omitidos los valores atípicos. Estos valores, además, provocan un cálculo demasiado alto de myd y demasiado bajo de pub_fac, lo que es importante desde el punto de vista de la política a seguir.

Si nos fijamos en el gráfico de rate93e y fac_stu, vemos un valor atípico que posiblemente no es un error de entrada de datos (al menos no de modo evidente). Aplicando el logaritmo natural a la creación de lnfac_stu obtenemos una relación de apariencia mucho más lineal. Existe ausencia de linealidad en la gráfica de of rate93e y pub_fac, que también se corrige sin problemas con el logaritmo natural. (Ver gráficas). Aplicando una regresión con lnfac_stu y lnpub_fac obtenemos:

Archivo	Gráfico de:
PSet4-C3-lnpub_fac	rate93e and log of pub_fac
PSet4-C3-lnfac_stu	rate93e and log of fac_stu

```
. reg rate93e lnpub_fac myd lnfac_stu suppfac
```

Source	SS	df	MS	Number of obs = 106		
Model	33.8241391	4	8.45603476	F(4, 101)	=	31.01
Residual	27.5428644	101	.272701628	Prob > F	=	0.0000
				R-squared	=	0.5512
				Adj R-squared	=	0.5334
Total	61.3670034	105	.584447652	Root MSE	=	.52221

rate93e	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnpub_fac	.3852297	.1155941	3.33	0.001	.1559222	.6145372
myd	-.1151531	.0455439	-2.53	0.013	-.2054999	-.0248063
lnfac_stu	.2108884	.0871956	2.42	0.017	.0379158	.3838609
suppfac	.0116026	.0037141	3.12	0.002	.0042348	.0189704
_cons	2.499508	.4151781	6.02	0.000	1.675907	3.32311

Esto nos permite obtener un 3% más de poder explicativo en nuestro R^2 , así como que todas nuestras variables tengan t-resultados muy por encima de 2. La ratio facultad-estudiantes demuestra ser tan importante como habíamos previsto.

Parte C.3.

```
. reg rate93e lnpub_fac myd lnfac_stu suppfac, beta
```

Source	SS	df	MS	Number of obs = 106		
Model	33.8241391	4	8.45603476	F(4, 101)	=	31.01
Residual	27.5428644	101	.272701628	Prob > F	=	0.0000
				R-squared	=	0.5512
				Adj R-squared	=	0.5334
Total	61.3670034	105	.584447652	Root MSE	=	.52221

rate93e	Coef.	Std. Err.	t	P> t	Beta
lnpub_fac	.3852297	.1155941	3.33	0.001	.3409101
myd	-.1151531	.0455439	-2.53	0.013	-.1817654
lnfac_stu	.2108884	.0871956	2.42	0.017	.1829051
suppfac	.0116026	.0037141	3.12	0.002	.2903878
_cons	2.499508	.4151781	6.02	0.000	.

He utilizado el comando beta para crear coeficientes estandarizados. Estos coeficientes se limitan a sustituir la escala de las variables por desviaciones estándar de la media, lo que da como resultado coeficientes sin unidad, permitiéndonos comparar las variables por sus efectos relativos. En este caso, el registro de publicaciones de la facultad es el más significativo en lo que se refiere a la eficacia del programa, seguido de cerca por el apoyo a la investigación.