

Soluciones al boletín de problemas 3

Capítulo 9.

4. **Respuesta: prácticamente lo mismo.** Considere dos diagramas de dispersión superpuestos y ligeramente cambiados de posición. Desplace el diagrama de las mujeres al suroeste (baje la altura y el peso) para reflejar las diferencias en altura y peso medio. Resulta que, aproximadamente, el cambio pasa por la "línea de desviación estándar". Por lo tanto, el nuevo diagrama de dispersión debería tener casi la misma correlación que los dos anteriores.

7. De nuevo, se trata de una correlación ecológica basada en el porcentaje, lo que suele dar lugar a un coeficiente de correlación exagerado. Además, el coeficiente de correlación no significa necesariamente causalidad. Sí que existe una considerable asociación entre el porcentaje de población de nativos y los votos recibidos por Johnson. Sin embargo, tal vez existan otros factores que vinculen la población nativa y los votos de Johnson, como el estatus socioeconómico. Además, existe una posible falacia de agregación. Si la fracción de votantes en un condado es muy pequeña y corresponde a un grupo específico de la población total, entonces los votantes son diferentes comparados con la población.

Capítulo 10.

- A – (i)
B – (iii)
C – (ii)

3. altura media de los maridos = 68 Desviación estándar (D.E.) de los maridos = 2,7
altura media de las mujeres = 63 Desviación estándar (D.E.) de las mujeres = 2,5 $r = 0,25$

$$r = \frac{\text{Cov}(\text{weight, height})}{SD_w \times SD_h} = 0,25$$

$$\exists_w = r \times \frac{SD_w}{SD_h} = \frac{0,25 \times 2,5}{2,7}$$

(a) $\frac{0,25 \times 2,5 \times 4}{2,7} \approx 1$, altura prevista de una mujer = 64

(b) $\frac{0,25 \times 2,5 \times -4}{2,7} \approx -1$, altura prevista de una mujer = 62

(c) $\frac{0,25 \times 2,5 \times 0}{2,7} = 0$, altura prevista de una mujer = 63

(d) no tenemos información de la altura de los maridos, por lo tanto, tenemos que predecir que la altura de una mujer es una media, en concreto, 63.

7. Ambos doctores están equivocados. Esta pregunta es sobre la falacia y sobre el efecto de la regresión. Cuando la primera medida es demasiado alta o demasiado baja, la segunda tiende a regresar a la media.

Capítulo 11.

1. (v) $\sqrt{(1-r^2)}$ x D.E. de y

2. Sí, hay algo mal. $\sqrt{(1-r^2)} = 3,12$, y $2 \times \text{r.m.s.} = 6,24$ lo que cubre el 95% de los datos. Incluso si la media es 0, $\pm 6,24$ está demasiado arriba y abajo. Como la escala GPA es normalmente 4,0, no tiene sentido que el valor más alto de los datos pueda ser 6,24.

6. NO. Correlación no es causalidad. Podemos inferir que un estudiante que hace su tarea tiende a tener un GPA mejor, porque probablemente sea estudioso. Sin embargo, no podemos aseverar que hacer las tareas hace aumentar el GPA del estudiante.

9. NO. Se trata otra vez de un efecto de la regresión. El novato del año es el jugador más notable de la temporada, lo que implica que se trata de un valor extremo. En el segundo año, lo normal es que regrese al nivel medio.

Capítulo 12.

4. (a) **alrededor de 1.** La línea representa la media de los datos. Todos los puntos de datos se encuentran entre cero y cuatro, así podemos adivinar que la D.E. de y es 1. (recuerde, $2x$ D.E. suele abarcar el 95 % de los datos). Y la D.E. de x es el r.m.s. para prever y por su media. Por lo tanto, debería estar alrededor de 1.
(b) NO. La línea de regresión parece bajar hacia la izquierda.

11. $\text{Pend.} = r \times \frac{SD_y}{SD_x} = 0,0000617$

$$\therefore 0,37 \times \frac{SD_y}{SD_x} = 0,0000617$$

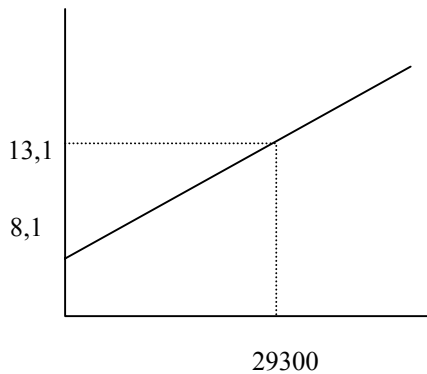
Cuando $x = 0$, $y = 8,1$ años.

$$8,1 = 13,1 - y, y = 5.$$

$$5 = 29300 \times \text{año}/\$$$

$$\text{año}/\$ = 5/29300 = 0,00017065, \text{ lo que difiere del coeficiente de pendiente en la ecuación. } (0,0000617)$$

Sin embargo, no lo podemos saber hasta que no conozcamos la desviación estándar.



Parte II

A.1.

```
. reg abortion attend
```

Source	SS	df	MS	Number of obs =	1748
Model	309.274994	1	309.274994	F(1, 1746) =	301.17
Residual	1793.01334	1746	1.02692631	Prob > F =	0.0000
				R-squared =	0.1471
				Adj R-squared =	0.1466
Total	2102.28833	1747	1.20337054	Root MSE =	1.0134

abortion	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
attend	.2362405	.0136129	17.35	0.000	.2095411 .2629398
_cons	2.166483	.0476642	45.45	0.000	2.072998 2.259968

```
. predict yhat1
```

Esta regresión muestra como la asistencia a misa tiene una incidencia absoluta y elocuente desde el punto de vista estadístico, sobre el modo de entender la ley del aborto. Observe la codificación inversa de ambas variables. La regresión sólo explica el 14% de la varianza, así que hay otros factores o se necesita una transformación.

A.2.

```
. gen attend2=attend^2  
(generados 18 valores faltantes)
```

```
. reg abortion attend attend2
```

Source	SS	df	MS	Number of obs =	1748
Model	323.516155	2	161.758077	F(2, 1745) =	158.69
Residual	1778.77217	1745	1.01935368	Prob > F =	0.0000
				R-squared =	0.1539
				Adj R-squared =	0.1529
Total	2102.28833	1747	1.20337054	Root MSE =	1.0096

abortion	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
attend	.4361105	.0551665	7.91	0.000	.3279111 .54431
attend2	-.0369555	.0098871	-3.74	0.000	-.0563473 -.0175637
_cons	2.016965	.0620911	32.48	0.000	1.895184 2.138745

```
. predict yhat2
```

En mi opinión los datos estaban un poco curvados, así que introduje una transformación polinómica, que son difíciles de interpretar, por eso es importante el gráfico en A.3. De nuevo la regresión es bastante elocuente y absoluta en el universo de valores posibles de asistencia (vea A.3.), pero el modelo transformado sólo explica un poco más de la varianza.

A.3.

```
. graph abortion yhat1 yhat2 attend, connect (.ss) symbol (Oii) jitter(3)  
see attached graph PSet3GraphA3
```

Como se desprende del gráfico, el modelo lineal se queda corto en la predicción de la incidencia para la mayoría del rango de asistencia y se excede en la predicción de los extremos.

B.1.

```
. gen lnbooks=ln(books)
```

```
. reg reading lnbooks
```

Source	SS	df	MS	Number of obs =	40
Model	639.043888	1	639.043888	F(1, 38) =	8.41
Residual	2886.05611	38	75.948845	Prob > F =	0.0062
				R-squared =	0.1813
				Adj R-squared =	0.1597
Total	3525.10	39	90.3871795	Root MSE =	8.7149

reading	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnbooks	9.226154	3.180651	2.90	0.006	2.787264 15.66504
_cons	139.6288	24.59164	5.68	0.000	89.84563 189.412

Elegí registrar los libros para silenciar el efecto de algunos valores extremos como Iowa y Kansas. Puede observar la diferencia consultando los gráficos adjuntos PSet3GraphB1 y PSet3GraphB1ln.

B.2.

```
. predict yhat
```

```
. gen resid= reading-yhat
```

Ver gráfico adjunto PSet3GraphB2

Probar también:

```
. rvfplot, s([state]) yline(0)
```

El gráfico residual debería tener el aspecto de ruido alrededor de una línea recta, pero entre 7.8 y 8 hay cierta curva. Podría darse aquí el concepto heteroscedástico (correlación entre la variable independiente y los residuales), pero más importante es la presencia de valores atípicos significativos, sobre todo Washington DC, con una predicción muy pobre. También preocupan Hawaii y Connecticut.

B.3.

La ecuación de regresión anterior es:

$$\text{reading} = \ln(\text{books}) * 9.226 + 139.629$$

Como una línea de regresión siempre atraviesa el punto (X_{avg} , Y_{avg}) el valor calculado por STATA de X_{avg} (la media de lnbooks) se puede utilizar en lugar de resolver lnbooks. Es 7,780 lo que se traduce en 2392,275 libros.

Podemos resolver el cálculo del número de libros que generan un incremento de 5 puntos en la lectura restando del número de libros necesario para crear una puntuación de lectura de 215,85 a partir de la media de libros.

Resolución para libros:

$$215,85 = \ln(\text{books}) * 9,226 + 139,629$$
$$76,221 = \ln(\text{books}) * 9,226$$
$$76,221 / 9,226 = \ln(\text{books})$$
$$8,262 = \ln(\text{books})$$
$$e^{8,262} = \text{books}$$
$$3873,834 = \text{books}$$

Por lo tanto, para incrementar en 5 puntos los niveles de lectura, necesitamos
(3873,834 - 2392,275) 1481,559 más libros por cada cien estudiantes

C.1.

```
. gen lnchaspnd = ln(chaspnd)
(7 missing values generated)
```

```
. reg incvote lnchaspnd
```

Source	SS	df	MS	Number of obs =	26
Model	4.0910e+12	1	4.0910e+12	F(1, 24) =	6.13
Residual	1.6010e+13	24	6.6708e+11	Prob > F =	0.0207
				R-squared =	0.2035
				Adj R-squared =	0.1703
Total	2.0101e+13	25	8.0403e+11	Root MSE =	8.2e+05

incvote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnchaspnd	183283.2	74010.97	2.48	0.021	30532.07 336034.3
_cons	-1501158	1006837	-1.49	0.149	-3579167 576850.1

De nuevo, utilizo un log natural de la variable independiente para linealizarla. Al utilizar ingresos o gastos, el log natural es una transformación habitual. En este caso no se ganaba mucho haciéndolo (véanse gráficos PSet3GraphC1 y PSet3GraphC1ln).

La regresión resultante parece indicar el resultado contraintuitivo de que el gasto del contendiente tiene un efecto absoluto y significativo sobre los votos de los titulares del cargo. Aquí, lo que obra en contra nuestra es probablemente los estados grandes. Éstos requieren más gasto de los candidatos para que el aumento en los gastos generales se relacione con el número de votos posible más que el efecto que intentamos medir. Tenga también en cuenta que el gasto del titular en el cargo no se controla.

C.2.

```
. gen twoptyvote=incvote/(incvote+chavote)
(generados 4 valores faltantes)
```

```
. reg twoptyvote lnchaspnd
```

Source	SS	df	MS	Number of obs =	26
Model	.13145517	1	.13145517	F(1, 24) =	45.84
Residual	.06882913	24	.00286788	Prob > F =	0.0000
				R-squared =	0.6563
				Adj R-squared =	0.6420
Total	.2002843	25	.008011372	Root MSE =	.05355

twoptyvote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnchaspnd	-.0328547	.0048528	-6.77	0.000	-.0428703 -.0228391
_cons	1.05211	.0660163	15.94	0.000	.915859 1.188361

Esto tiene mejor pinta. No sólo el coeficiente es significativo, sino que el signo se encuentra en la dirección esperada. Si utilizamos un porcentaje de votos para silenciar el efecto del tamaño del estado como una variable desconcertante, moldeamos mejor los datos. Este 2° modelo explica alrededor del 66% de la varianza.

C.3.

Claramente, el modelo correcto es el segundo, que implica que cuanto más dinero gasta un candidato, menor es la cuota de voto del titular y, por lo tanto, la probabilidad de victoria de éste.