Ayudantía 8

Respuestas

1. Considere el siguiente modelo:

$$\log(\text{salary})_i = \beta_0 + \beta_1 log\ (sales)_i + \beta_2 finance_i + \beta_3 consprod_i + \beta_4 utility_i + \mu_i$$

donde *"salary"* es el salario anual del director general de la empresa en miles de dólares, *"sales"* son las ventas anuales de la empresa en millones de dólares, y "*finance*", "*consprod*" y "*utility*" son variables binarias que indican el sector en el que opera la empresa (sector financiero, sector de bienes consumo y sector servicios). El sector omitido es el sector industrial.

a) Estime el modelo utilizando los datos del fichero CEOSAL1 y presente los resultados en forma de ecuación.

b) Manteniendo fijas las ventas, calcule la diferencia porcentual promedio en el salario estimado entre los sectores financiero e industrial ¿Es esta diferencia estadísticamente significativa al 1 por ciento?

c) Manteniendo fijas las ventas, ¿Cuál es en promedio la diferencia porcentual en el salario estimado entre el sector de bienes de consumo y el sector financiero? Contraste si la diferencia es estadísticamente significativa al 5%.

Paso 0.1: Remuevo todo el entorno y configuro mi environment.

```
rm(list=ls())
getwd() ## Directorio inicial
```

```
## [1] "C:/Users/IgnacioSepulveda/Documents/Ayudantias/econometrics"
```

```
setwd("C:/Users/IgnacioSepulveda/Documents/Ayudantias/econometrics") ## Directorio nuevo
getwd() ## Check
```

```
## [1] "C:/Users/IgnacioSepulveda/Documents/Ayudantias/econometrics"
```

Paso 0.2: Carga las librerías necesarias, **si no las tengo las instalo**

```
## install.package("tidyverse") si no lo tengo instalado, saco #, y lo corro
## install.packages("stargazer")
## install.packages("haven")
library("tidyverse")
library("haven")
library("stargazer")
```

1

a)

Paso 1: Cargo la base,

```
df=read_dta('CEOSAL1.dta')
```

Paso 2: Veo que tal los datos,

```
#str(df)
```

Paso 2: Corro la regresión y obtengo los coeficientes,

```
reg_1a=lm(lsalary~lsales+finance+consprod+utility,data=df)
```

Paso 3: Finalmente armo la ecuación,

$$log(salary)_i = 4.9 + 0.244log(sales)_i + 0.124finance_i + 0.239consprod_i - 0.352utility_i$$

b)

Aquí debemos notar del enunciado que tanto finance ,consprod y utility son tipos de industrias, por lo tanto son variables dummy donde su coeficiente es una difencia de medias con respecto al industrial. Entonces basta con observar el p-value de la tabla.

Queremos hacer el siguiente test,

$H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$

Rechazamos H_0 cuando p-value $< \alpha$

```
reg_1a %>% summary()
```

```
##
## Call:
## lm(formula = lsalary ~ lsales + finance + consprod + utility,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20301 -0.24100  0.00333  0.16418  2.60009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.88998    0.27487  17.790  < 2e-16 ***
## lsales       0.24426    0.03209   7.613 9.78e-13 ***
## finance      0.12406    0.08926   1.390 0.166091
## consprod     0.23850    0.08295   2.875 0.004464 **
## utility     -0.35315    0.09681  -3.648 0.000336 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4662 on 204 degrees of freedom
## Multiple R-squared:  0.3356, Adjusted R-squared:  0.3225
## F-statistic: 25.76 on 4 and 204 DF,  p-value: < 2.2e-16
```

Como p-value = 0.166 no podemos rechazar la nula de que la medias son estadisticamente iguales.

c)

El β que nos da la diferencia es simplemente la resta del β asociado a consprod menos el asociado a finance, dado que se cancela la media del industrial dado que ambos lo usan como base, por lo tanto sabemos que el β,o diferecias de medias, sera igual a 0.114 . El problema es que el error estandar no podemos calcularlo tan facil. La formula seria la siguiente $se(\beta_1 - \beta_2) = \sqrt{Var(\beta_1 - \beta_2)} = \sqrt{Var(\beta_1) - Var(\beta_2) + 2Cov(\beta_1, \beta_2)}$

Una forma más fácil es simplemente cambiar la base de la regresión y dado que queremos comparar consumo vs finance podríamos dejar a finance como base.

El modelo cambiaría al siguiente,

$$log(salary)_i = \beta_0 + \beta_1 log(sales)_i + \beta_2 indus_i + \beta_3 consprod_i + \beta_4 utility_i$$

```
reg_1c=lm(lsalary~lsales+indus+consprod+utility,data=df)
reg_1c %>% summary()
```

```
##
## Call:
## lm(formula = lsalary ~ lsales + indus + consprod + utility, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20301 -0.24100  0.00333  0.16418  2.60009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.01405    0.27710  18.095  < 2e-16 ***
## lsales       0.24426    0.03209   7.613 9.78e-13 ***
## indus       -0.12406    0.08926  -1.390    0.166
## consprod     0.11444    0.09142   1.252    0.212
## utility     -0.47722    0.10413  -4.583 7.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4662 on 204 degrees of freedom
## Multiple R-squared:  0.3356, Adjusted R-squared:  0.3225
## F-statistic: 25.76 on 4 and 204 DF,  p-value: < 2.2e-16
```

Vemos que efectivamente el valor de 0.1144, pero su se es de 0.0914. Con un p-value = 2.12 no podemos rechazar al 5% entonces no es significativa.

2

2. Para analizar el salario de los profesores universitarios se utiliza el modelo:

$$salario_i = \beta_0 + \beta_1 hombre_i + \beta_2 blanco_i + \beta_3 exper_i + \mu_i$$

donde *"salario"* es el salario anual del profesor en miles de dólares, *"exper"* son los años de experiencia docente, *"Hombre"* es una variable binaria que vale 1 si el profesor es hombre (0 si es mujer) y "*Blanco*" es otra variable binaria que vale 1 si el profesor es de raza blanca (0 raza no blanca).

a) Determine el salario medio para:
a.1) Hombres de raza blanca.
a.2) Mujeres de raza blanca.
a.3) Hombres de raza no blanca.

a.4) Mujeres de raza no blanca.

b) ¿Cuál es la diferencia en el salario medio entre:
b.1) Hombres blancos y mujeres blancas con la misma experiencia laboral?
b.2) Hombres blancos y hombres no blancos con la misma experiencia laboral?
b.3) Mujeres blancas y mujeres no blancas con la misma experiencia laboral?
b.4) Hombres no blancos y mujeres no blancas con la misma experiencia laboral?

c) ¿Cómo contrastaría la hipótesis de que no existe diferencia en el salario medio entre blancos y no blancos con la misma experiencia laboral y el mismo género?

Desarrollo en la pizarra.

3

3. Considere el siguiente modelo:

$$colgpa_i = \beta_0 + \beta_1 hsize_i + \beta_2 hsperc_i + \beta_3 sat_i + \beta_4 female_i + \beta_5 athlete_i + \mu_i$$

donde *"colgpa"* es la calificación media acumulada en la universidad, *"hsize"* es el número de alumnos en la promoción de bachillerato (en centenares), *"hsperc" es* el percentil que ocupa en la distribución de calificaciones de los alumnos del instituto que se graduaron el mismo año (definido de forma que, por ejemplo, hsperc = 5 se refiere al cinco por ciento de los mejores alumnos que se gradúan), *"sat"* es la puntuación en el test SAT de aptitud escolar, *"female"* es una variable ficticia que vale 1 si el estudiante es mujer (0 si es hombre) y *"athlete"* es otra variable ficticia que vale 1 si el estudiante es atleta (0 si no lo es).

a) Estime el modelo utilizando los datos del fichero GPA2 y presente los resultados en forma de ecuación. ¿Cuál es la diferencia estimada en la nota media de la universidad entre los atletas y los que no lo son? ¿Cómo interpreta ese coeficiente? ¿Cómo contrastaría que no existe diferencia promedio en la calificación media acumulada de la universidad entre los atletas y los que no lo son?

b) ¿Cuál es el promedio de la calificación media acumulada en la universidad de las mujeres? ¿Y la de los hombres?

c) ¿Cuánto predice el modelo que será la calificación media acumulada en la universidad cuando el estudiante viene de una promoción de 100 alumnos de bachillerato, está dentro del 10% de los mejores estudiantes que se gradúan, obtuvo 800 puntos en el test de aptitud escolar, es mujer y no es atleta?

d) Suprima "sat" del modelo y vuelva a estimar la ecuación. ¿Cuál es ahora la diferencia estimada por ser atleta? Explique por qué la estimación es diferente de la obtenida en el apartado a.

Paso 1: Veo que tal los datos y remuevo anteriores,

```
rm(list=ls()) ## Removemos todo lo anterior
df=read_dta('gpa2.dta')
```

Paso 2: Veo que tal los datos,

```
#str(df)
```

a)

```
reg_3a=lm(colgpa~hsize+hsperc+sat+female+athlete,data=df)
reg_3a
```

```
##
## Call:
## lm(formula = colgpa ~ hsize + hsperc + sat + female + athlete,
##     data = df)
##
## Coefficients:
## (Intercept)        hsize       hsperc          sat       female      athlete
##    1.194217    -0.024486    -0.012957     0.001648     0.156322     0.170143
```

$colga_i = 1.194 + 0.024hsize_i - 0.013hsperc_i + 0.002sat_i + 0.156female_i + 0.17athlete_i$

Con el β asociado a atleta podemos saber la diferencia, el cual muestra que los atletas en promedio ganan 0.17 mas que los no atleta, cuando todo lo demás se mantiene constante.

Para saber si es significativa vemos su p-value.

```
reg_3a %>% summary()
```

```
##
## Call:
## lm(formula = colgpa ~ hsize + hsperc + sat + female + athlete,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.71136 -0.35031  0.03175  0.38878  1.88487
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.194e+00  7.622e-02  15.668  < 2e-16 ***
## hsize       -2.449e-02  4.988e-03  -4.909 9.51e-07 ***
## hsperc      -1.296e-02  5.597e-04 -23.152  < 2e-16 ***
## sat          1.648e-03  6.684e-05  24.656  < 2e-16 ***
## female       1.563e-01  1.800e-02   8.685  < 2e-16 ***
## athlete      1.701e-01  4.236e-02   4.016 6.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5546 on 4131 degrees of freedom
## Multiple R-squared:  0.2918, Adjusted R-squared:  0.2909
## F-statistic: 340.4 on 5 and 4131 DF,  p-value: < 2.2e-16
```

El p-value es bien chico por lo tanto podemos rechazar incluso al 1%, lo que implica de que atleta es significativa y existe una diferencia estadística entre las medias de sus salarios.

b)

Para las mujeres es $\beta_0 + \beta_{female} = 1.194 + 0.156 = 1.35$

Para los hombre es $\beta_0 = 1.194$

c)

```
### Creamos el nuevo df
nueva_data=data.frame(hsize=1,hsperc=10,sat=800,female=1,athlete=0)
paste("predecimos....")
```

```
## [1] "predecimos...."
```

```
reg_3a %>% predict(newdata=nueva_data)
```

```
##        1
## 2.514908
```

d)

Recordamos que la formula es la siguiente sesgo $= \beta_{sat} \frac{Cov(athlete,sat)}{Var(athlete)}$.

Entonces como esperamos que sea β_{sat}? Mejores notas debiesen implicar un salario...

Y la $Cov(athlete, sat)$? Qué pasa si calculamos su correlación?

```
cor(df$sat,df$athlete)
```

```
## [1] -0.1850938
```

Vemos que su relación lineal es negativa, por lo tanto nuestro sesgo debiese ser negativo como el β_{sat} es positivo, lo que implicaría una subestimación. Osea el coeficiente que nuevo sera menor que el anterior.

Comprobemos....

```
reg_3d=lm(colgpa~hsize+hsperc+female+athlete,data=df)
stargazer(reg_3a,reg_3d,type='latex',header=FALSE)
```

Table 1:

	Dependent variable:	
	colgpa	
	(1)	(2)
hsize	-0.024^{***}	-0.017^{***}
	(0.005)	(0.005)
hsperc	-0.013^{***}	-0.017^{***}
	(0.001)	(0.001)
sat	0.002^{***}	
	(0.0001)	
female	0.156^{***}	0.060^{***}
	(0.018)	(0.019)
athlete	0.170^{***}	0.006
	(0.042)	(0.045)
Constant	1.194^{***}	2.996^{***}
	(0.076)	(0.023)
Observations	4,137	4,137
R^2	0.292	0.188
Adjusted R^2	0.291	0.187
Residual Std. Error	0.555 (df = 4131)	0.594 (df = 4132)
F Statistic	340.369^{***} (df = 5; 4131)	238.458^{***} (df = 4; 4132)
Note:		*p<0.1; **p<0.05; ***p<0.01