

# Tema 4. La inferencia en el modelo de regresión

Gustavo A. García

[ggarci24@eafit.edu.co](mailto:ggarci24@eafit.edu.co)

Econometría para la Toma de Decisiones

Maestría en Economía Aplicada

Escuela de Finanzas, Economía y Gobierno

Universidad EAFIT

Link slides en formato **html**

Link slides en formato **PDF**

## En este tema

- Introducción
- Pruebas de hipótesis
- Intervalos de confianza
- La estrecha relación entre los intervalos de confianza y las pruebas de hipótesis
- Ejercicio aplicado en R

# Lecturas

- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. [Cap. 4](#)
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. [Cap. 5](#)

# Introducción

El modelo de RLS presenta la siguiente estructura:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

En la estimación de los parámetros del modelo tenemos

Parámetro	Estimador	Varianza estimada
$\beta_0$	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$	$\hat{\sigma}_{\hat{\beta}_0}^2 = \frac{\hat{\sigma}_u^2 \sum X_i^2}{n \sum x_i^2}$
$\beta_1$	$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$	$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}_u^2}{\sum x_i^2}$
$\sigma_u^2$	$\hat{\sigma}_u^2 = \frac{\sum \hat{u}^2}{n-2}$	-

# Introducción

- El método estadístico intenta decir cosas sobre los parámetros poblacionales con base en los estadísticos muestrales
- En el caso del modelo de RLS, consiste en decir algo acerca de  $\beta_0$  y  $\beta_1$  con base en  $\hat{\beta}_0$  y  $\hat{\beta}_1$
- Lo anterior implica construir intervalos de confianza y pruebas de hipótesis para  $\beta_0$  y  $\beta_1$

# Pruebas de hipótesis

- Ahora se quiere verificar estadísticamente una afirmación como la siguiente:  $\beta_1 = \beta_{10}$ , esto es, verificar la hipótesis nula ( $H_0$ ):  $H_0 : \beta_1 = \beta_{10}$
- En estadística las hipótesis se rechazan o no se rechazan
- Lo importante en la inferencia estadística es:
  - suponer que  $H_0$  es cierta
  - encontrar la distribución muestral bajo  $H_0$
  - observar la realidad bajo el supuesto de  $H_0$  cierta
  - si lo observado es poco probable  $\implies$  rechazar  $H_0$   
si lo observado es probable  $\implies$  no rechazar  $H_0$
- En consecuencia, las hipótesis nulas ( $H_0$ ) que se verifican son del tipo igualdad a, ya que es bajo este supuesto que se dan las distribuciones muestrales conocidas
- Cuando se esta bajo hipótesis nulas del tipo  $>$ ,  $<$  o  $\neq$  se tienen otras distribuciones

# Pruebas de hipótesis

Tenemos que el método estadístico de toma de decisiones implica:

- Formular una hipótesis nula (en términos de igualdad) y una hipótesis alternativa

$$H_0 : \beta_1 = \beta_{10}$$

$$H_A : \begin{aligned} &\beta_1 < \beta_{10} \text{ ó} \\ &\beta_1 \neq \beta_{10} \text{ ó} \\ &\beta_1 > \beta_{10} \end{aligned}$$

- Hay que encontrar la distribución muestral del estadígrafo apropiado, bajo  $H_0$

$$\text{Bajo } H_0 \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{N-2} \text{ gdl}$$

- Dado esto se define el nivel de significancia aceptable en la prueba ( $\epsilon$ )



# Pruebas de hipótesis

- No se debe olvidar que cualquier decisión que se tome se hace en condiciones de incertidumbre:

$H_0$ Decisión \ Realidad	Cierta	Falsa
Rechaza	Error tipo I	Decisión correcta
No rechaza	Decisión correcta	Error tipo II

- $Prob(\text{Cometer error tipo I}) = \epsilon \implies$  Nivel de significancia
- $1 - Prob(\text{Cometer error tipo II}) \implies$  Potencia de la prueba

# Pruebas de hipótesis

La mecánica es

- Se formula el contraste

$$H_0 : \beta_1 = \beta_{10}$$

$$H_A : \begin{aligned} &\beta_1 < \beta_{10} \text{ ó} \\ &\beta_1 \neq \beta_{10} \text{ ó} \\ &\beta_1 > \beta_{10} \end{aligned}$$

- Bajo  $H_0$  cierto el estadístico de prueba ( $t_0$ ) será:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{N-2} \text{gdl}$$

- Se establece una regla de decisión en función de  $H_0$ :

- si  $H_A : \beta_1 < \beta_{10} \implies t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}_{\hat{\beta}_1}} < -t_{N-2}(\epsilon) \implies \text{Rechazo } H_0$
- si  $H_A : \beta_1 \neq \beta_{10} \implies |t_0| = \frac{|\hat{\beta}_1 - \beta_{10}|}{\hat{\sigma}_{\hat{\beta}_1}} > t_{N-2}(\epsilon/2) \implies \text{Rechazo } H_0$
- si  $H_A : \beta_1 > \beta_{10} \implies t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}_{\hat{\beta}_1}} > t_{N-2}(\epsilon) \implies \text{Rechazo } H_0$

# Pruebas de hipótesis

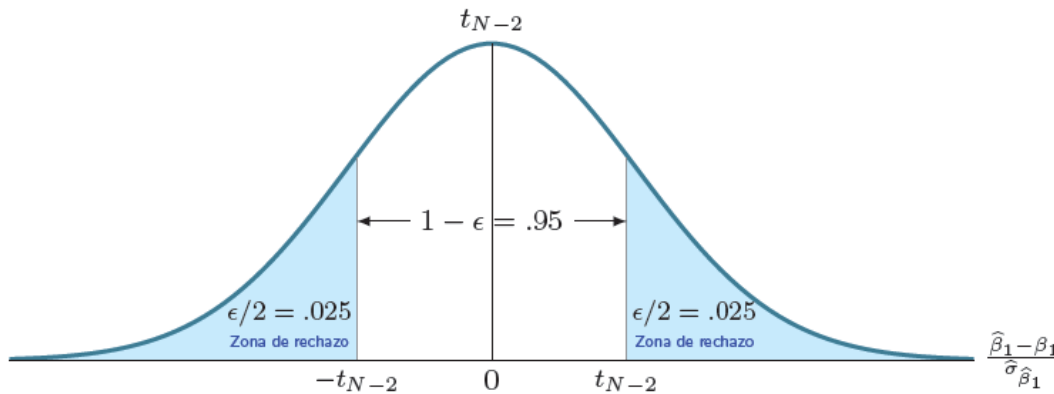
- *p-value*: la probabilidad del límite derecho de  $H_0$  bajo el supuesto de que es cierta
- La regla es rechazar  $H_0$  si

$$p\text{-value} < \epsilon$$

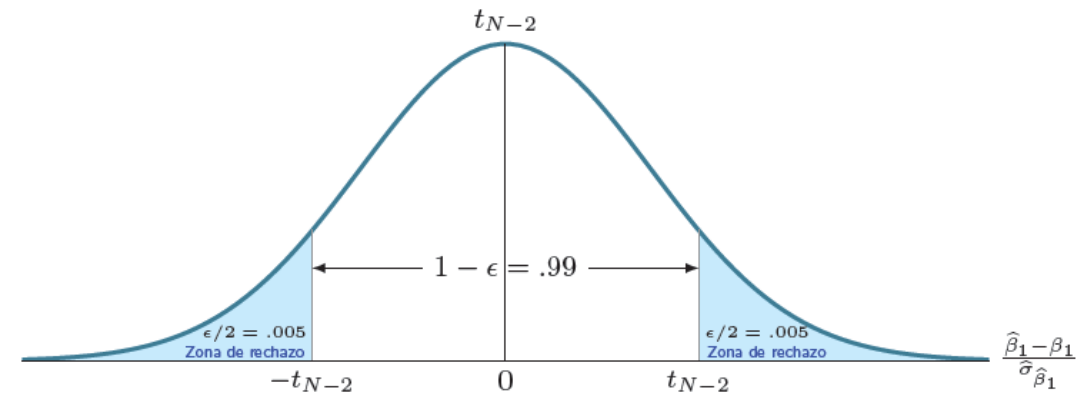
Donde el *p-value*  $(t_0) = 2(1 - F(|t_0|, m))$ ,  $m$  son los grados de libertad

- Gráficamente sería:

Nivel de significancia  $= \epsilon = 0.05$

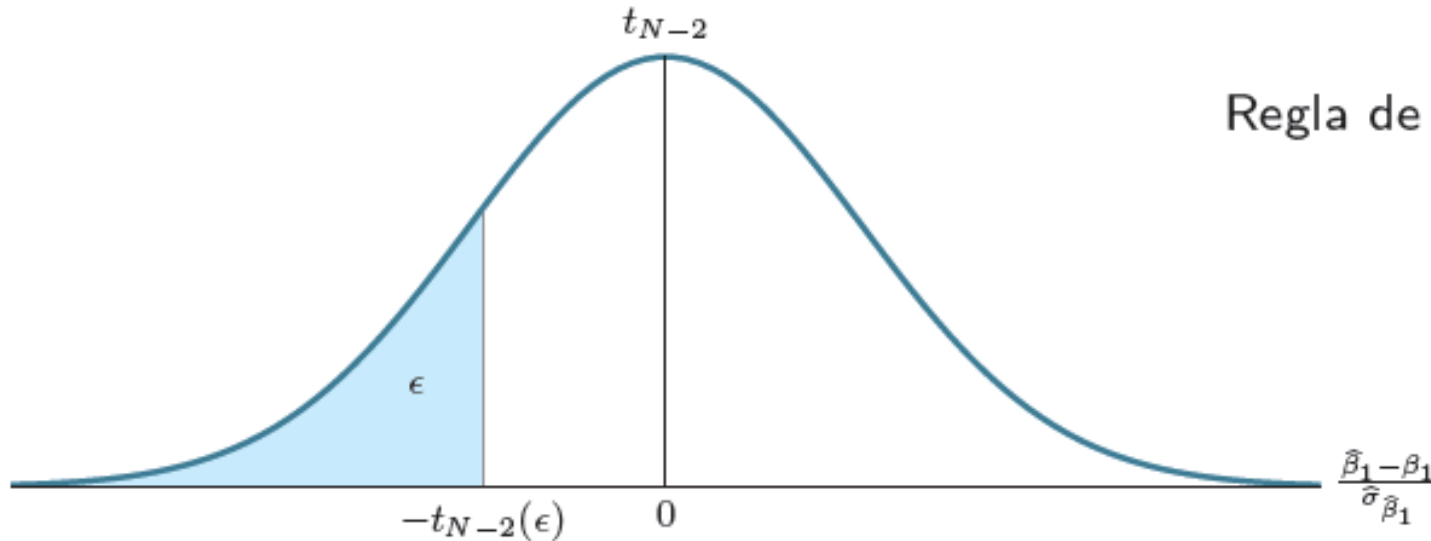


Nivel de significancia  $= \epsilon = 0.01$



# Pruebas de hipótesis

- En el nivel de significancia ( $\epsilon$ ) se tiene el número mágico del 5%
- Si es del tipo  $H_A : \beta_1 < \beta_{10}$ , se trata de una prueba con cola situada a la izquierda



Regla de decisión: rechazar  $H_0$  al nivel de significancia  $\epsilon$  si  $t_0 < -t_{(N-2)}(\epsilon)$  con  $t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}_{\hat{\beta}_1}}$

- Con el uso del  $p$ -value:

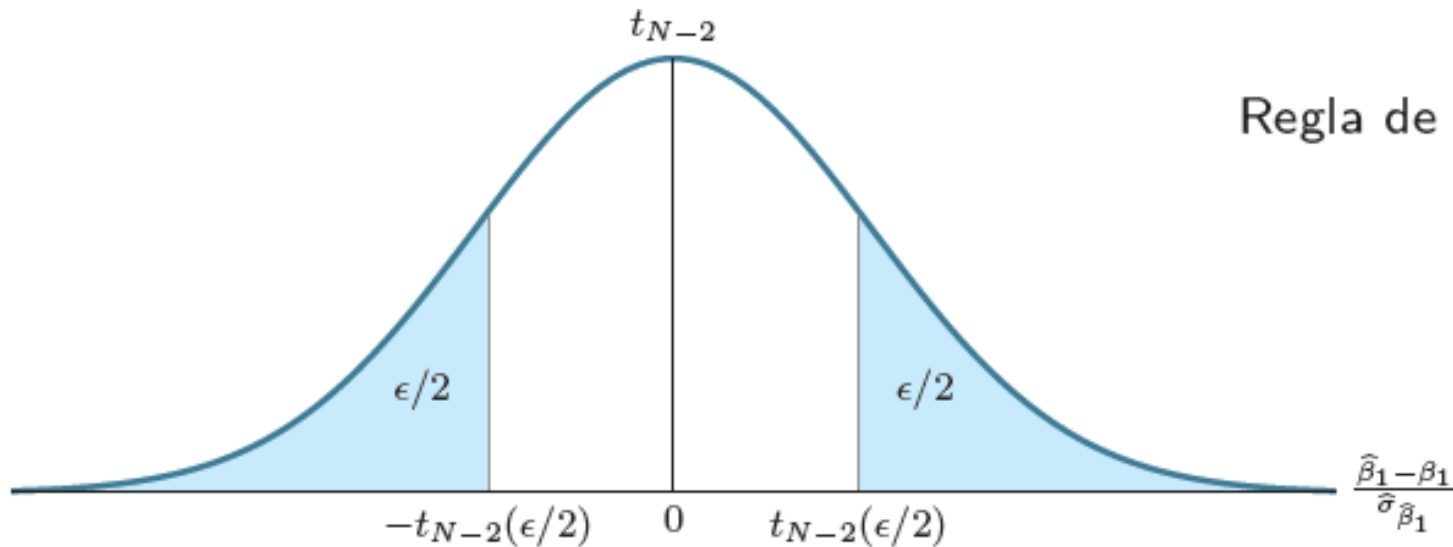
$$p\text{-value} = \int_{-\infty}^{t_0} t_{N-2} dt$$

Esto es exactamente el nivel marginal de significancia en el cual se puede rechazar  $H_0$ . La regla de decisión sería:

Rechazar  $H_0$  si  $p\text{-value} < \epsilon$

# Pruebas de hipótesis

- Si la hipótesis alternativa es del tipo  $H_A : \beta_1 \neq \beta_{10}$ , se tiene una prueba de dos colas



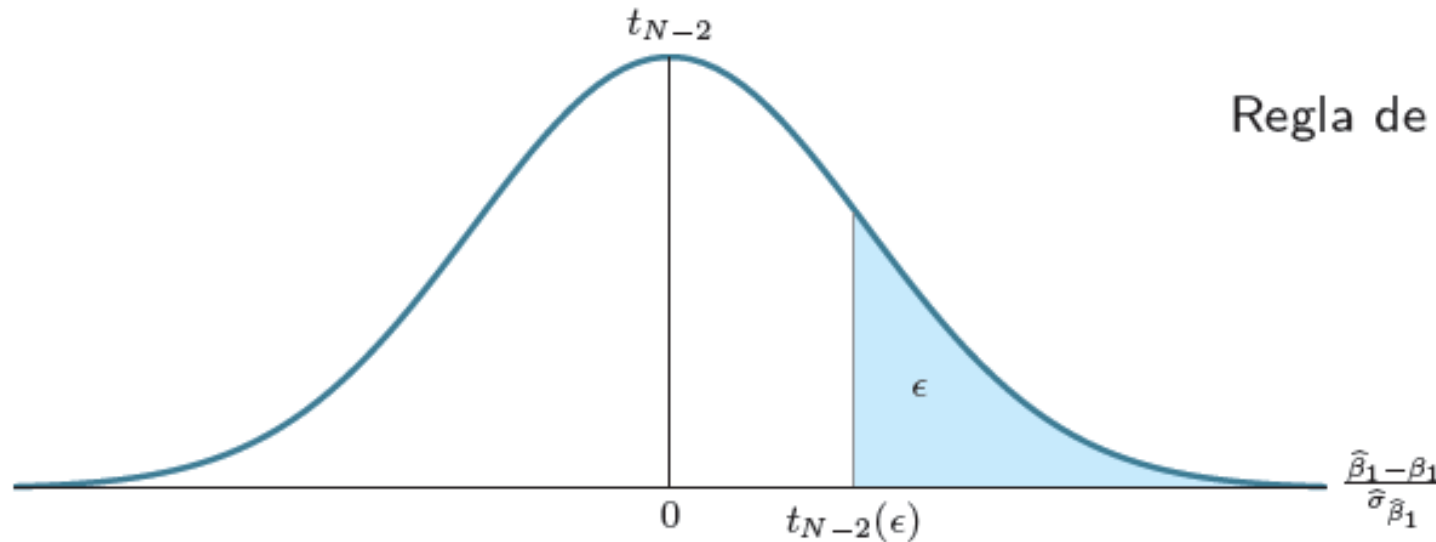
Regla de decisión: rechazar  $H_0$  al nivel de significancia  $\epsilon$  si  $|t_0| > t_{(N-2)}(\epsilon/2)$  con  $t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma} \hat{\beta}_1}$

- En términos del *p-value* la regla de decisión sería:

Rechazar  $H_0$  si *p-value*  $< \epsilon$

# Pruebas de hipótesis

- Si la hipótesis alternativa es del tipo  $H_A : \beta_1 > \beta_{10}$ , se tiene una prueba con cola a la derecha



Regla de decisión: rechazar  $H_0$  al nivel de significancia  $\epsilon$  si  $t_0 > t_{(N-2)}(\epsilon)$   
con  $t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}_{\hat{\beta}_1}}$

- En términos del *p-value* la regla de decisión sería:

Rechazar  $H_0$  si *p-value*  $< \epsilon$

# Intervalos de confianza

## Definición

- Es la probabilidad de que dos valores extremos contengan el parámetro desconocido
- Son unos límites probabilísticos que contienen al verdadero parámetro (en este caso  $\beta_1$ ) con una probabilidad de  $1 - \epsilon$  (nivel de confianza)

## Definición matemática

$$Prob \left[ \hat{\beta}_1 - \hat{\sigma}_{\hat{\beta}_1} t_{N-2}(\epsilon/2) \leq \beta_1 \leq \hat{\beta}_1 + \hat{\sigma}_{\hat{\beta}_1} t_{N-2}(\epsilon/2) \right] = 1 - \epsilon$$

## Interpretación

- La probabilidad de que el intervalo que va desde  $\hat{\beta}_1 - \hat{\sigma}_{\hat{\beta}_1} t_{N-2}(\epsilon/2)$  hasta  $\hat{\beta}_1 + \hat{\sigma}_{\hat{\beta}_1} t_{N-2}(\epsilon/2)$  contenga el verdadero valor de  $\beta_1$  es  $1 - \epsilon$
- El intervalo de confianza  $\hat{\beta}_1 \pm \hat{\sigma}_{\hat{\beta}_1} t_{N-2}(\epsilon/2)$  contiene a  $\beta$  con una probabilidad de  $1 - \epsilon$
- En el  $(1 - \epsilon)\%$  de los casos el intervalo contendrá el parámetro  $\beta_1$
- $IC_{(1-\epsilon)}(\beta_1) = \hat{\beta}_1 \pm \hat{\sigma}_{\hat{\beta}_1} t_{N-2}(\epsilon/2)$  = Estimador  $\pm$  Error de estimación (Valor t-student)

# La estrecha relación entre los intervalos de confianza y las pruebas de hipótesis

- Se rechaza la hipótesis nula  $H_0 : \beta_1 = \beta_{10}$  a un nivel de significancia  $\epsilon$ , cuando  $\beta_{10}$  cae por fuera del correspondiente  $100(1 - \epsilon)\%$  intervalo de confianza
- En nuestro ejemplo wage-educ,  $\beta_{10} = 0$  no cae dentro del intervalo de confianza del 95% (0.436, 0.645), y por tanto, usando este enfoque, nosotros de nuevo rechazamos  $H_0 : \beta_1 = 0$  a un nivel de significancia del 5%
- De hecho, se rechaza cualquier hipótesis nula donde  $\beta_{10}$  no esta contenido en el intervalo de confianza (0.436, 0.645)
- Por otro lado, Si el IC(95%) contiene el cero,  $\beta_1$  no es significativo al 5%



# Ejemplo en R

Se tiene una base de datos de corte transversal de 526 trabajadores correspondientes a 1976 para los Estados unidos. *wage* son los salarios en dólares por hora y *educ* los años de educación. Se desea estimar el siguiente modelo:

$$wage = \beta_0 + \beta_1 educ + u$$

```
library(haven); library(summarytools); library(Hmisc); library(tidyverse)
data <- read_stata("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")
# Descripción de la base de datos: http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.des
names(data)
```

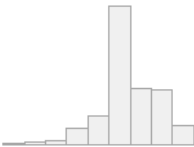
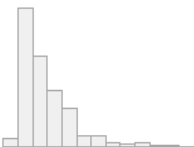
```
[1] "wage"      "educ"      "exper"     "tenure"    "nonwhite"  "female"
[7] "married"   "numdep"    "smsa"      "northcen"  "south"     "west"
[13] "construc"  "ndurman"   "trcommpu"  "trade"     "services"  "profserv"
[19] "profocc"   "clerocc"   "servocc"   "lwage"     "expersq"   "tenursq"
```

```
descr(data[,c("educ", "wage")], stats = "common", transpose = TRUE, headings = FALSE)
```

	Media	Dev.std.	Min	Mediana	Max	Num.Válido	Pct.Válido
educ	12.56	2.77	0.00	12.00	18.00	526.00	100.00
wage	5.90	3.69	0.53	4.65	24.98	526.00	100.00

# Ejemplo en R

```
st_options(lang = "es", footnote=NA, headings = FALSE)
print(dfSummary(data[,c("educ","wage")], valid.col = FALSE, silent=FALSE), method = "render", varnumbers=F)
```

Variable	Estadísticas / Valores	Frec. (% sobre válidos)	Gráfico	Perdidos
educ [numeric]	Media (d-s) : 12.6 (2.8) min < mediana < max: 0 < 12 < 18 RI (CV) : 2 (0.2)	18 valores distintos		0 (0.0%)
wage [numeric]	Media (d-s) : 5.9 (3.7) min < mediana < max: 0.5 < 4.7 < 25 RI (CV) : 3.6 (0.6)	241 valores distintos		0 (0.0%)

# Ejemplo en R

```
# Correlaciones
cor(data[,c("wage", "educ")])
```

```
      wage      educ
wage 1.0000000 0.4059033
educ 0.4059033 1.0000000
```

```
# Correlaciones con significancia estadística
rcorr(as.matrix(data[,c("wage", "educ")]))
```

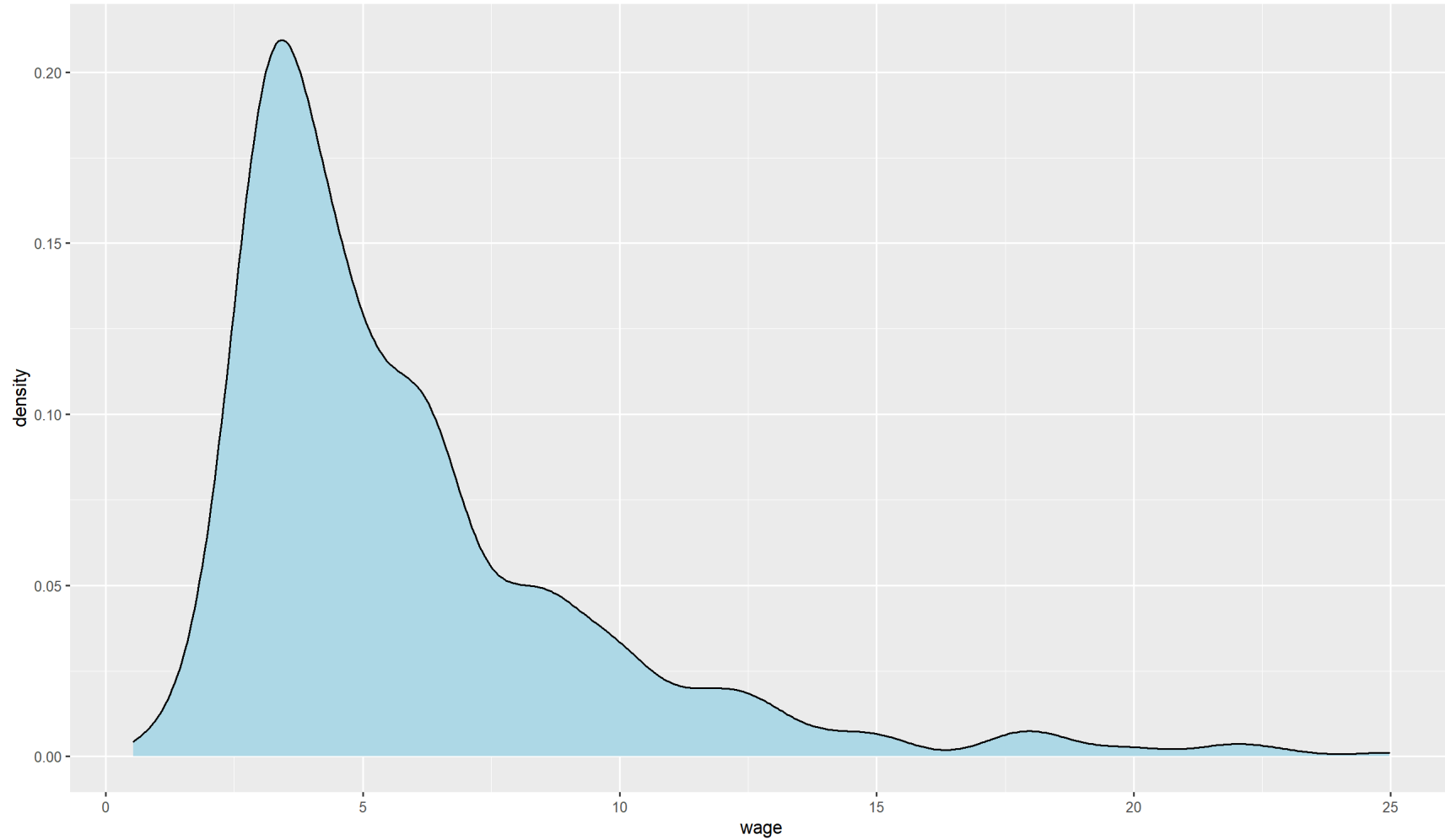
```
      wage educ
wage 1.00 0.41
educ 0.41 1.00
```

n= 526

```
P
      wage educ
wage      0
educ      0
```

# Ejemplo en R

```
# Densidades de los salarios  
ggplot(data, aes(x=wage)) +  
  geom_density(fill="lightblue")
```

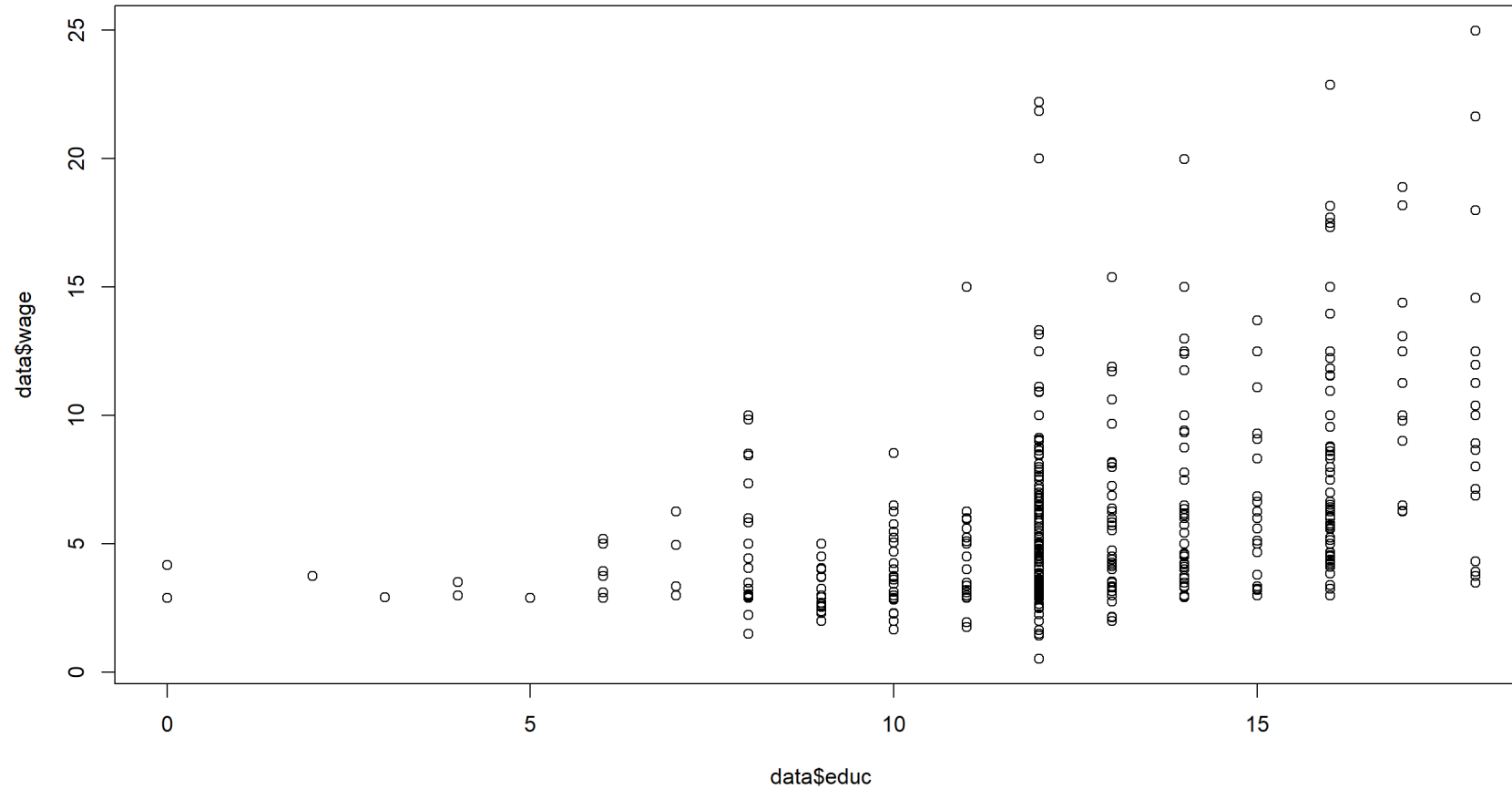


# Ejemplo en R

```
# Densidades de los salarios por género
data <- data |> mutate(sex = case_when(female==1~"F", female==0~"M"))
ggplot(data, aes(x=wage, fill=sex)) + geom_density(alpha=0.3) +
  scale_fill_manual(name="Género", labels = c("Mujer", "Hombre"), values=c("red", "blue")) + labs(x= "Salario", y="Densidad")
```

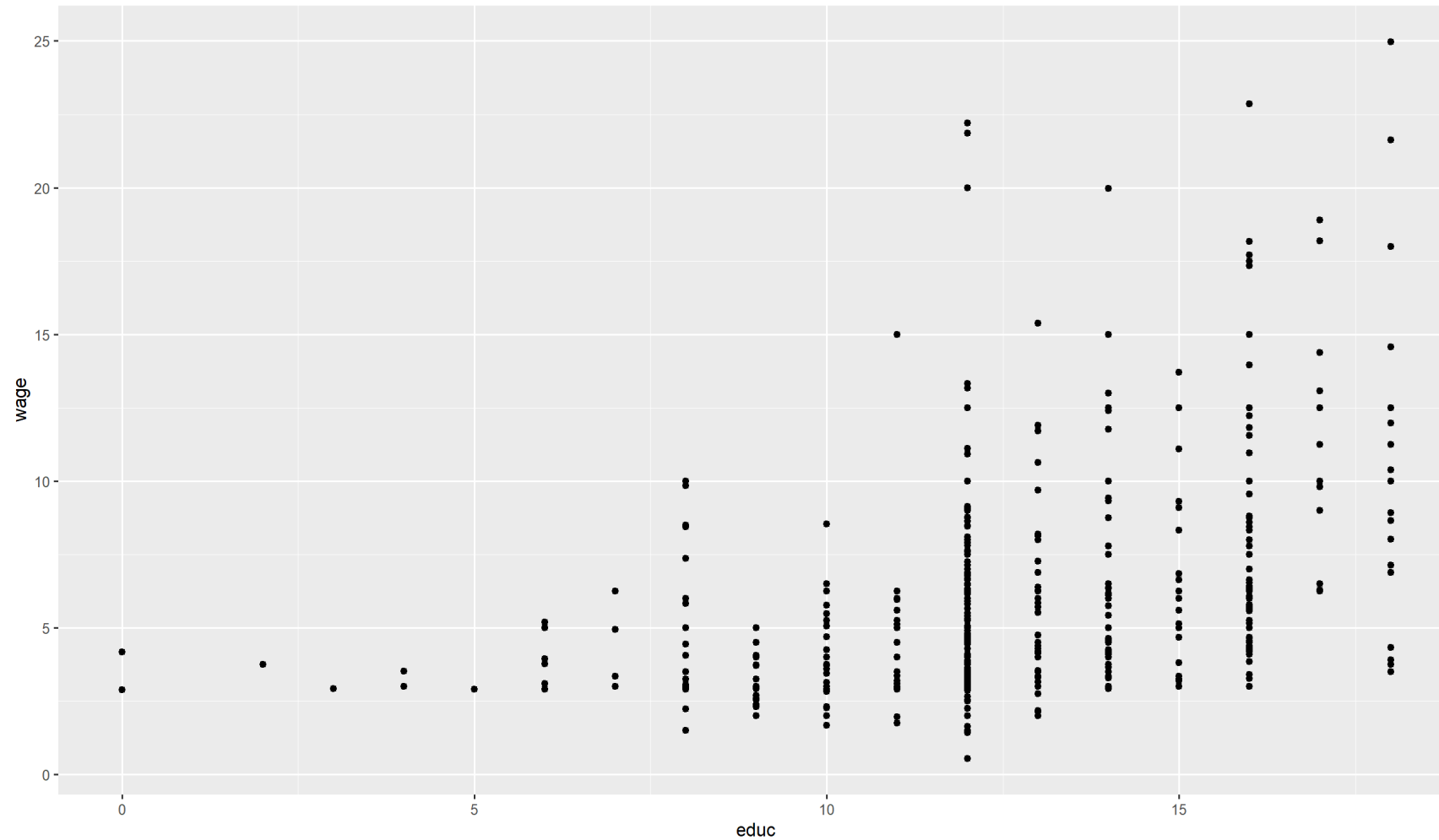
# Ejemplo en R

```
plot(data$wage~data$educ)
```



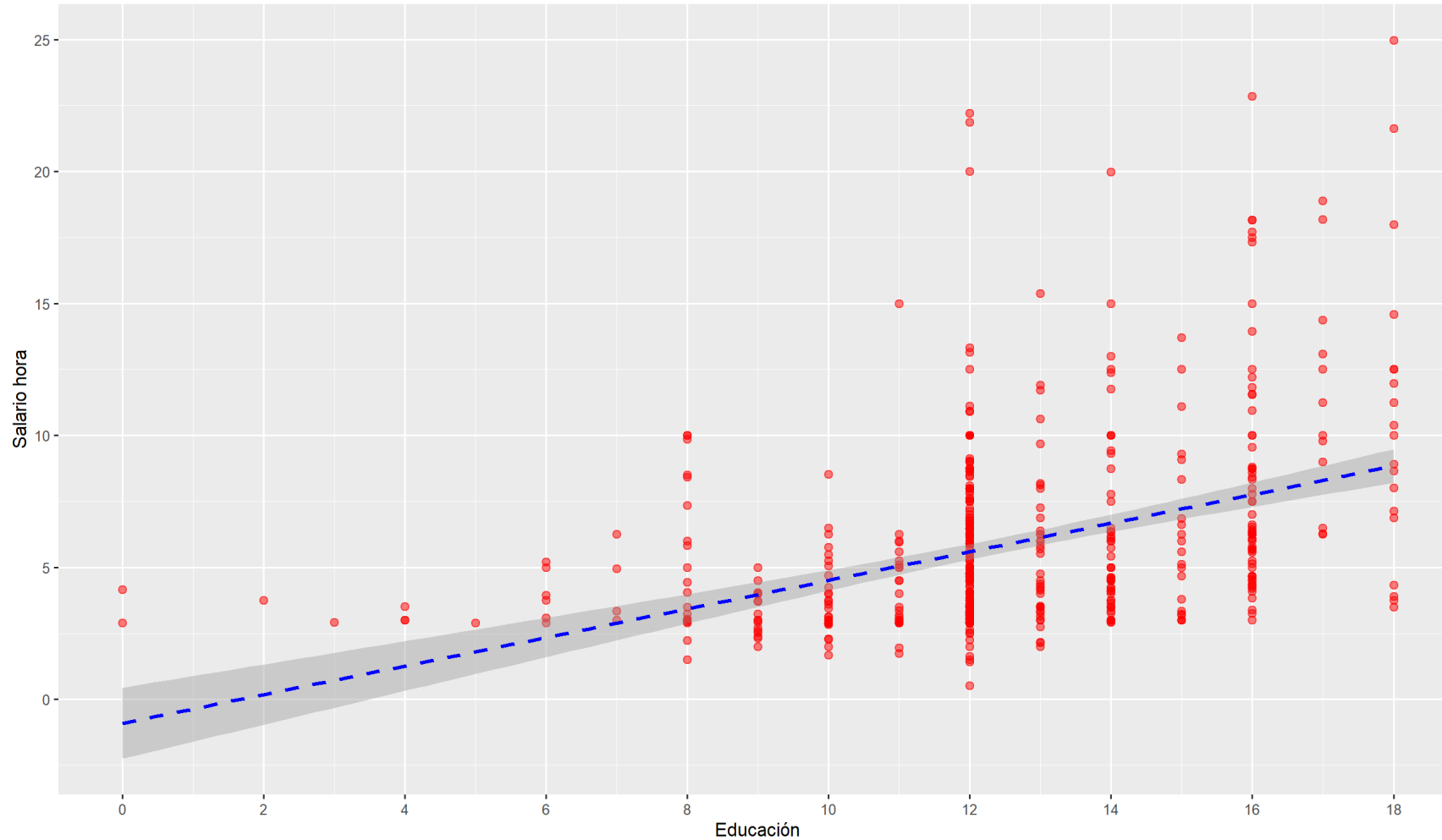
# Ejemplo en R

```
ggplot(data, aes(x=educ, y=wage)) +  
  geom_point()
```



# Ejemplo en R

```
ggplot(data, aes(x=educ, y=wage)) +  
  geom_point(alpha=0.5, color="red", size=2) + geom_smooth(formula=y~x, method=lm, linetype="dashed", color="blue") +  
  labs(x= "Educación", y="Salario hora") + scale_x_continuous(breaks = seq(0, 18, by = 2))
```





# Ejemplo en R

```
# El modelo de regresión
modelo <- lm(wage~educ, data=data)
summary(modelo)
```

Call:

```
lm(formula = wage ~ educ, data = data)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-5.3396	-2.1501	-0.9674	1.1921	16.6085

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.90485	0.68497	-1.321	0.187
educ	0.54136	0.05325	10.167	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.378 on 524 degrees of freedom

Multiple R-squared: 0.1648, Adjusted R-squared: 0.1632

F-statistic: 103.4 on 1 and 524 DF, p-value: < 2.2e-16

```
# Intervalos de confianza
confint(modelo, level = 0.95)
```

	2.5 %	97.5 %
(Intercept)	-2.2504719	0.4407687
educ	0.4367534	0.6459651

# Ejemplo en R

```
freq(data$female, headings = F)
```

	Frec.	% Válido	% Válido acu.	% Total	% Total acu.
0	274	52.09	52.09	52.09	52.09
1	252	47.91	100.00	47.91	100.00
<NA>	0			0.00	100.00
Total	526	100.00	100.00	100.00	100.00

```
# El modelo de regresión para mujeres
modelo_f <- lm(wage~educ, data=subset(data,female==1))
summary(modelo_f)
```

Call:  
lm(formula = wage ~ educ, data = subset(data, female == 1))

Residuals:

Min	1Q	Median	3Q	Max
-3.9137	-1.3212	-0.6352	0.6474	14.4654

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.99803	0.72851	-1.37	0.172
educ	0.45348	0.05799	7.82	1.48e-13 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.272 on 250 degrees of freedom  
Multiple R-squared: 0.1965, Adjusted R-squared: 0.1933  
F-statistic: 61.15 on 1 and 250 DF, p-value: 1.482e-13

```
# El modelo de regresión para hombres
modelo_m <- lm(wage~educ, data=subset(data,female==0))
summary(modelo_m)
```

Call:  
lm(formula = wage ~ educ, data = subset(data, female == 0))

Residuals:

Min	1Q	Median	3Q	Max
-6.1611	-2.7532	-0.7192	1.7725	15.5258

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.20050	1.01646	0.197	0.844
educ	0.53948	0.07739	6.971	2.38e-11 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.84 on 272 degrees of freedom  
Multiple R-squared: 0.1516, Adjusted R-squared: 0.1485  
F-statistic: 48.6 on 1 and 272 DF, p-value: 2.378e-11