Ejercicios Regresión

Gloria

2025-04-15

Contents

Capítulo 1. Regresión Lineal Simple **1**
Correlación 1
Modelo Lineal Simple 8

Capítulo 1. Regresión Lineal Simple

Correlación

Ejercicio 1.1.

En el archivo grasacerdos.xlsx se encuentran los datos del peso vivo (PV, en Kg) y al espesor de grasa dorsal (EGD, en mm) de 30 lechones elegidos al azar de una población de porcinos Duroc Jersey del Oeste de la provincia de Buenos Aires. Se pide:
(a) Dibujar el diagrama de dispersión e interpretarlo.
(b) Calcular el coeficiente de correlación muestral y explíquelo.
(c) ¿Hay suficiente evidencia para admitir asociación entre el peso y el espesor de grasa? ($\alpha = 0,05$). Verifique los supuestos para decidir el indicador que va a utilizar.

```
## # A tibble: 5 x 3
##     Obs PV    EGD
##   <dbl> <chr> <chr>
## 1     1 56,81 16,19
## 2     2 70,40 22,00
## 3     3 71,73 19,52
## 4     4 75,10 31,00
## 5     5 79,65 23,58
```

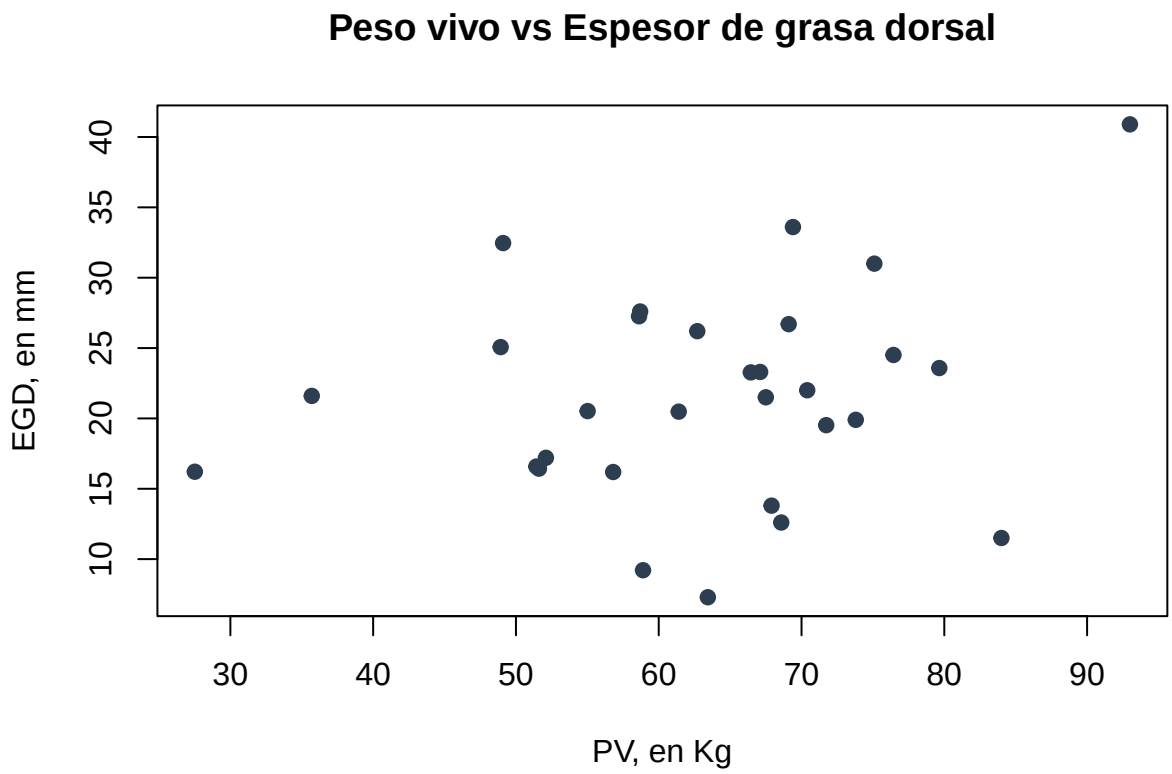
Cambiamos las "," por "." y visualizamos cuantos nulos hay por columna. Convertimos las columnas de PV y EGD a numéricas. Pedimos el summary del dataframe.

```
## tibble [30 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Obs: num [1:30] 1 2 3 4 5 6 7 8 9 10 ...
##  $ PV : chr [1:30] "56.81" "70.40" "71.73" "75.10" ...
##  $ EGD: chr [1:30] "16.19" "22.00" "19.52" "31.00" ...
## NULL
```

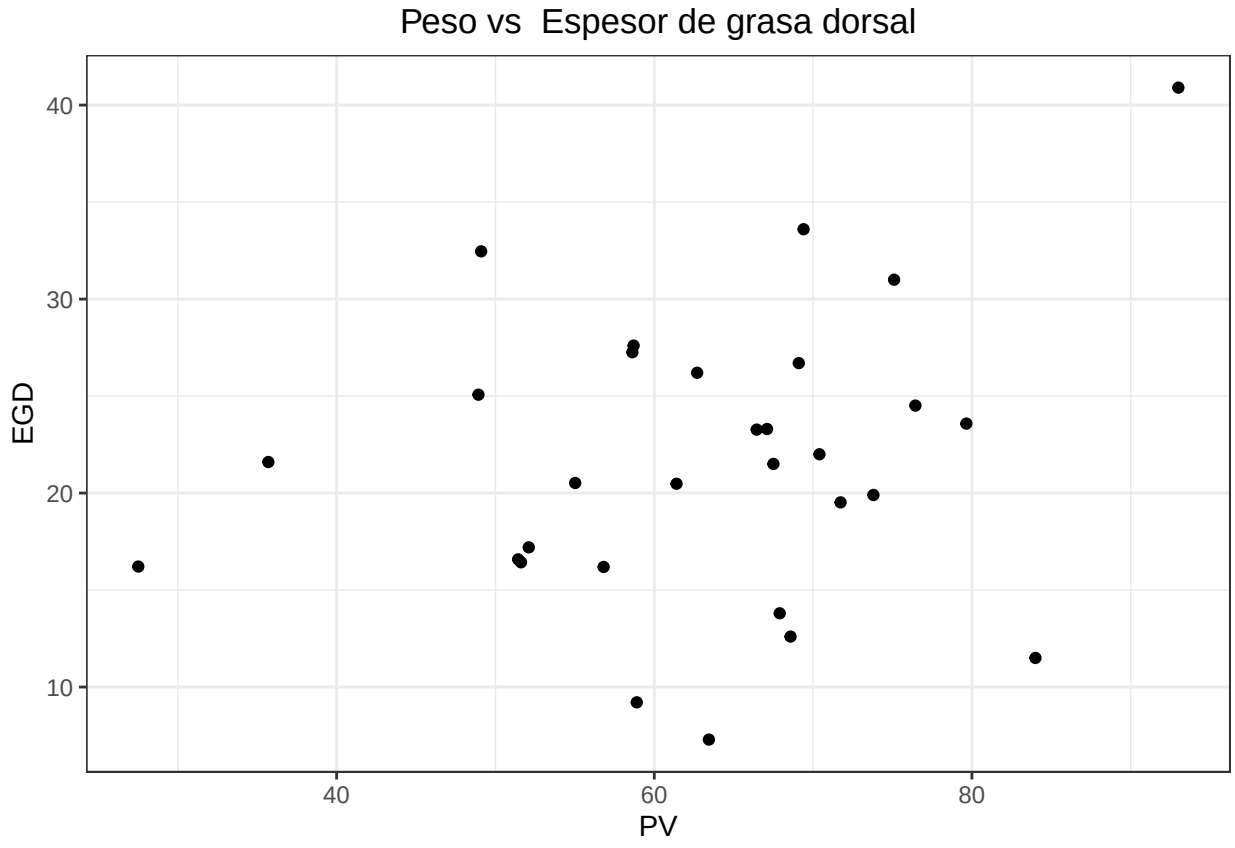
```
## Obs  PV EGD
##   0   0   0
```

```
##        PV             EGD
##  Min.   :27.51   Min.   : 7.29
```

```
##  1st Qu.:55.47   1st Qu.:16.47
##  Median :64.94   Median :21.55
##  Mean   :63.07   Mean   :21.60
##  3rd Qu.:70.15   3rd Qu.:25.92
##  Max.   :93.00   Max.   :40.90
```

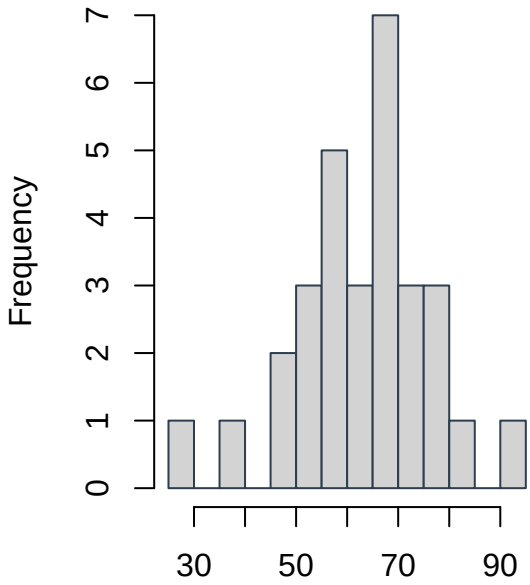

Peso vivo vs Espesor de grasa dorsal
EGD, en mm
10
15
20
25
30
35
40
30
40
50
60
70
80
90
PV, en Kg

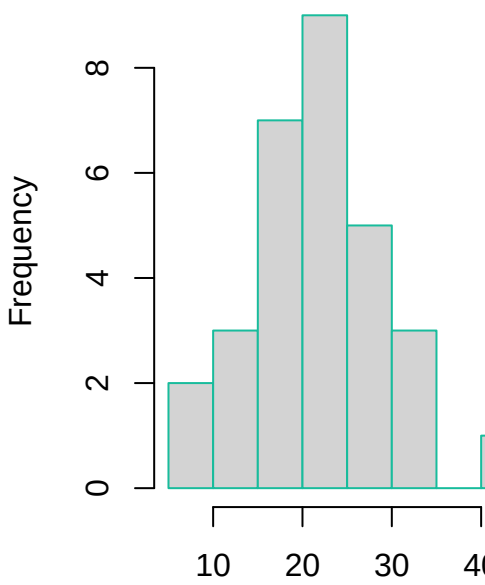
(1.1.a)


Peso vs Espesor de grasa dorsal
EGD
10
20
30
40
40
60
80
PV

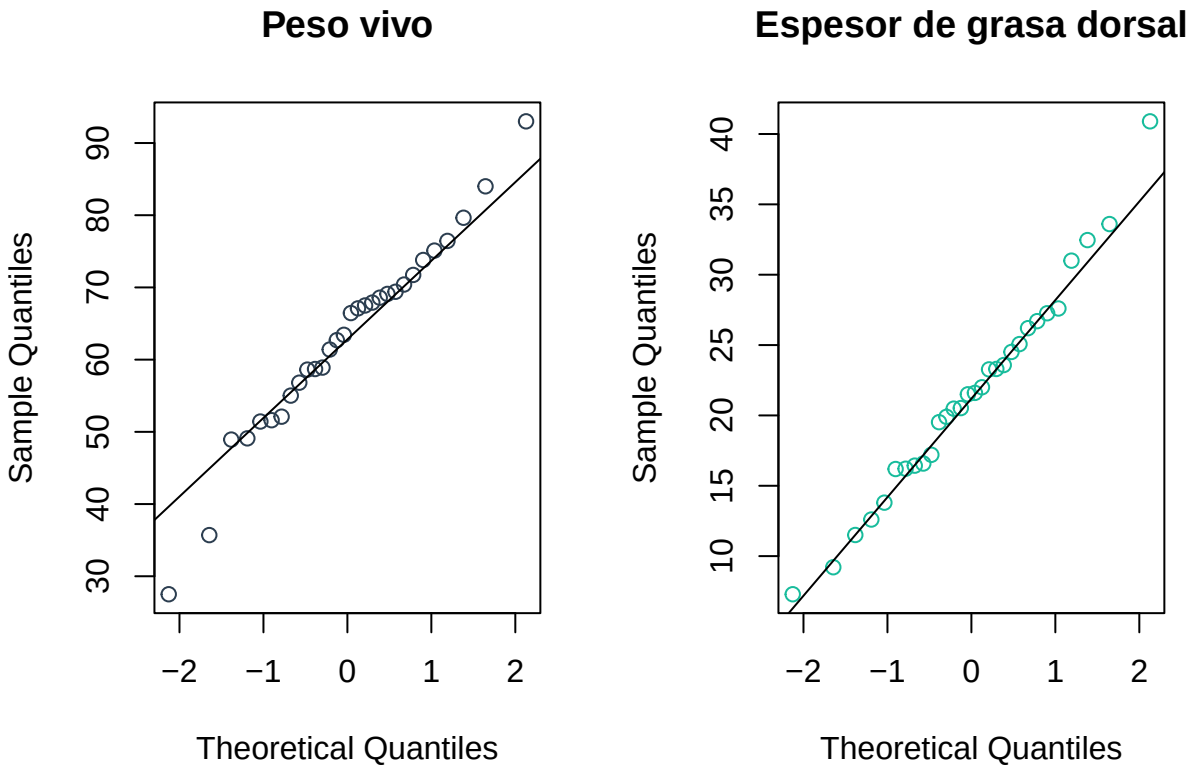
En los datos visualizados no pareciera haber asociación entre las variables.

(1.1.b) En primer lugar analizo la normalidad de las variables mediante los gráficos de los histogramas, los qq-





plots y la prueba de Shapiro.




```
##
##  Shapiro-Wilk normality test
##
## data:  PV
## W = 0.97533, p-value = 0.6925
```

```
##
##  Shapiro-Wilk normality test
##
## data:  EGD
## W = 0.98514, p-value = 0.9395
```

Por el test de Shapiro Wilk no se puede rechazar la normalidad de los datos en ninguno de los dos casos.
Análisis de la normalidad multivariada - Test de Henze Zirkler

```
#Análisis de normalidad bivariada
library(MVN)
attach(grasacerdos)
peso_egd=data.frame(PV,EGD)
#Usamos Test Henze-Zirkler para evaluar normalidad multivariada (bivariada en este caso)
respuesta_testHZ<-mvn(peso_egd , mvnTest = "hz")
print(respuesta_testHZ$multivariateNormality)
```

```
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.2539437 0.9049686 YES
```

El test da por resultado que las variables son normales bivariadas.

La correlación entre ambas variables es:

```
## [1] 0.2543434
```

La correlación entre las variables es baja: Si el valor de r es cercano a 0, indica que no existe una tendencia creciente o decreciente entre las variables estudiadas.

```
##
##  Pearson's product-moment correlation
##
## data:  PV and EGD
## t = 1.3916, df = 28, p-value = 0.175
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1166112  0.5630217
## sample estimates:
##       cor
## 0.2543434
```

El test arroja que el valor p de la prueba de Pearson es 0.175.

```
##            PV       EGD
## PV  1.0000000 0.2543434
## EGD 0.2543434 1.0000000
```

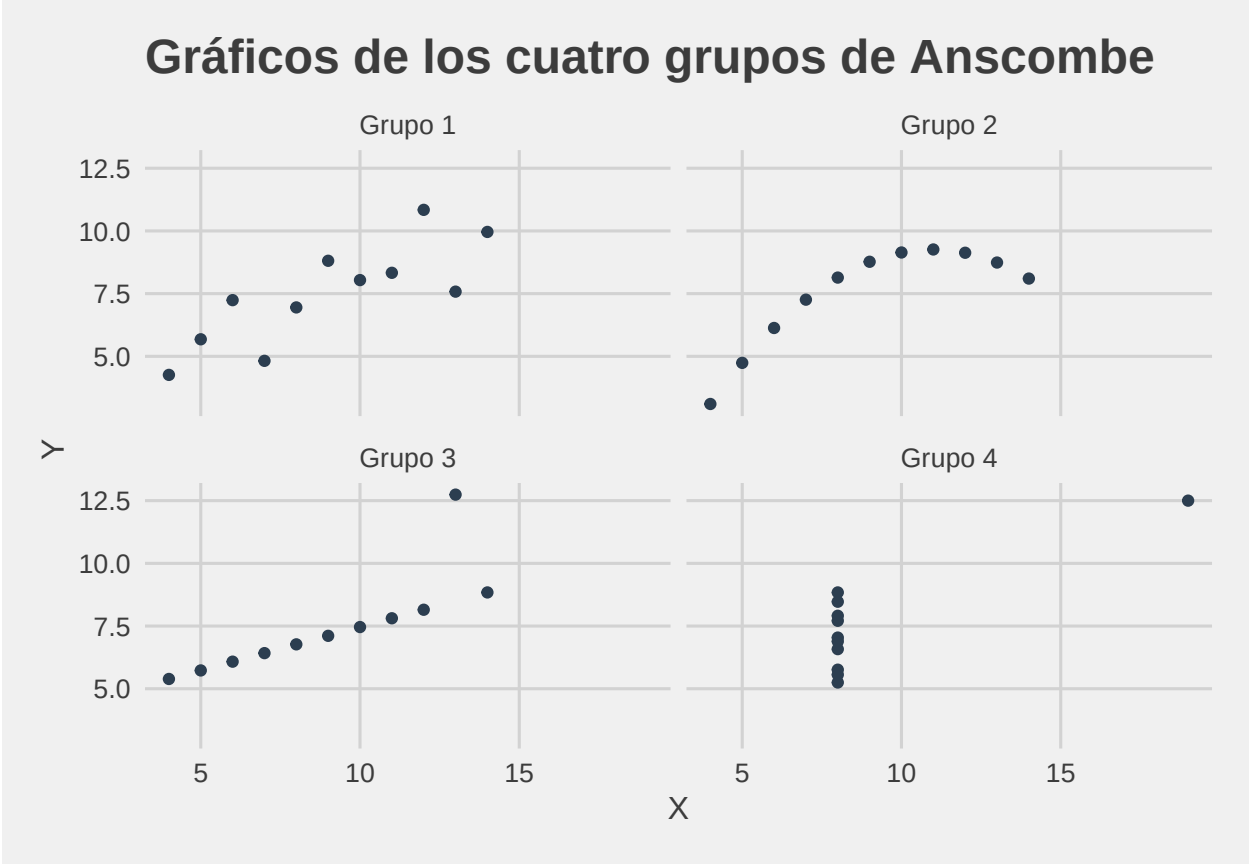
Si bien no es necesario aplicar el coeficiente de Spearman pues se satisfacen los supuestos, igualmente lo hago.

```
##
##  Spearman's rank correlation rho
##
## data:  PV and EGD
## S = 3748, p-value = 0.3785
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1661846
```

Ejercicio 1.2.

Los datos del cuarteto de Anscombe se encuentran en el archivo anscombe.xlsx Se pide explorar los datos de la siguiente manera:
(a) Graficar los cuatro pares de datos en un diagrama de dispersión cada uno.
(b) Hallar los valores medios de las variables para cada para de datos.
(c) Hallar los valores de la dispersión para cada conjunto de datos.
(d) Hallar el coeficiente muestral de correlación lineal en cada caso.
(e) Observar, comentar y concluir.

Grupo	media_x	media_y	var_x	var_y	correl	rcuad
Grupo 1	9	7.500909	11	4.127269	0.8164205	0.6665425
Grupo 2	9	7.500909	11	4.127629	0.8162365	0.6662420
Grupo 3	9	7.500000	11	4.122620	0.8162867	0.6663240
Grupo 4	9	7.500909	11	4.123249	0.8165214	0.6667073



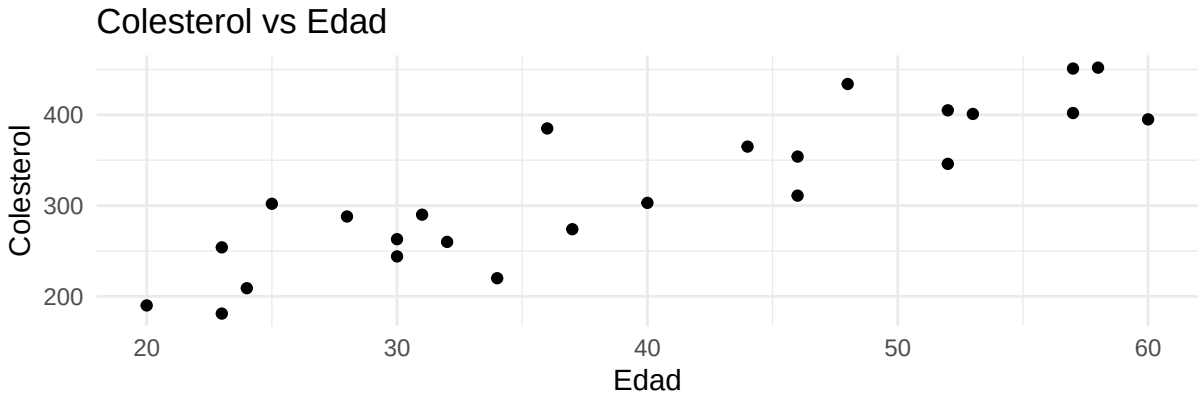
(a)

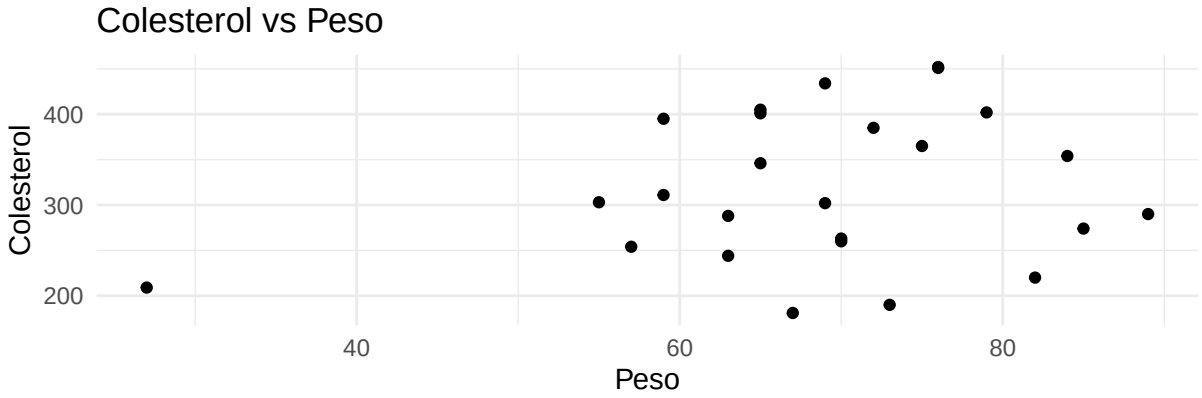
(b)

Modelo Lineal Simple

Ejercicio 1.3.

El archivo peso_edad_colest.xlsx disponible en contiene regis tros correspondientes a 25 individuos respecto de su peso, su edad y el nivel de colesterol total en sangre. Se pide:
(a) Realizar el diagrama de dispersión de colesterol en función de la edad y de colesterol en función de peso. Le parece adecuado ajustar un modelo lineal para alguno de estos dos pares de variables?
(b) Estime los coeficientes del modelo lineal para el colesterol en función de la edad.
(c) Estime intervalos de confianza del 95% para los coeficientes del modelo y compare estos resultados con el test de Wald para los coeficientes. Le parece que hay asociación entre estos test y el test de la regresión?
(d) A partir de esta recta estime los valores de E(Y) para x = 25 años y x =48años. Podría estimarse el valor de E(Y) para x = 80 años?
(e) Testee la normalidad de los residuos y haga un gráfico para ver si son homocedásticos.





(a)

Viendo los gráficos de dispersión, pareciera que el colesterol en función de la edad tiene una relación lineal más clara que el colesterol en función del peso. Por lo tanto, es más adecuado ajustar un modelo lineal para el colesterol en función de la edad.

(b)

```
##
## Call:
## lm(formula = colest ~ edad, data = peso_edad_colest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.311 -22.602  -2.627  27.589  85.348
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  95.5020    26.2545   3.638  0.00138 **
## edad          5.6708     0.6345   8.937 6.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.69 on 23 degrees of freedom
## Multiple R-squared:  0.7764, Adjusted R-squared:  0.7667
## F-statistic: 79.87 on 1 and 23 DF,  p-value: 6.094e-09
```

####(c)