

List 1

Jesus Gamboa

2026-01-16

Pregunta 1

Se utilizará el dataset cars, disponible en R, que contiene observaciones de:

- speed: velocidad del automóvil (millas por hora)
- dist: distancia de frenado (pies)

Considerar alfa = 0.05 cuando sea necesario

- a) Estimar e interpretar el coeficiente de correlación de Pearson entre la velocidad y la distancia de frenado

`cars`

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
## 7    10   18
## 8    10   26
## 9    10   34
## 10   11   17
## 11   11   28
## 12   12   14
## 13   12   20
## 14   12   24
## 15   12   28
## 16   13   26
## 17   13   34
## 18   13   34
## 19   13   46
## 20   14   26
## 21   14   36
## 22   14   60
## 23   14   80
## 24   15   20
## 25   15   26
## 26   15   54
## 27   16   32
## 28   16   40
## 29   17   32
## 30   17   40
```

```

## 31   17   50
## 32   18   42
## 33   18   56
## 34   18   76
## 35   18   84
## 36   19   36
## 37   19   46
## 38   19   68
## 39   20   32
## 40   20   48
## 41   20   52
## 42   20   56
## 43   20   64
## 44   22   66
## 45   23   54
## 46   24   70
## 47   24   92
## 48   24   93
## 49   24  120
## 50   25   85

cor(cars$speed, cars$dist)

```

`## [1] 0.8068949`

El coeficiente de correlación de Pearson entre la velocidad y la distancia de frenado es 0.807, lo que significa que existe una asociación lineal directa / positiva fuerte / alta entre las variables.

b) ¿Se puede concluir que la correlación es estadísticamente distinta de cero?

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$\alpha = 0.05$

```
cor.test(cars$speed, cars$dist, method = "pearson")
```

```

##
##  Pearson's product-moment correlation
##
## data:  cars$speed and cars$dist
## t = 9.464, df = 48, p-value = 1.49e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6816422 0.8862036
## sample estimates:
##        cor
## 0.8068949

```

Dado que el pvalor (1.49×10^{-12}) es menor al nivel de significancia, se rechaza la hipótesis nula. Por lo tanto, existe evidencia para concluir que la correlación es estadísticamente distinta de cero.

c) Formular el modelo de regresión lineal simple que relaciona la distancia de frenado con la velocidad

$$Y = \beta_0 + \beta_1 X + \epsilon$$

donde:

Y : distancia de frenado

X : velocidad del auto

β_0 : intercepto del modelo

β_1 : pendiente del modelo

ϵ : error del modelo

d) Estimar de manera puntual los coeficientes del modelo e interpretarlos.

```
modelo1 = lm(dist ~ speed, cars)
modelo1 |> coef()
```

```
## (Intercept)      speed
## -17.579095     3.932409
```

$$\hat{Y} = -17.58 + 3.93X$$

- -17.58: no tiene interpretación porque la distancia no puede ser negativa, además el problema no tiene sentido si $x = 0$.
- 3.93: por cada milla por hora adicional de velocidad del auto, se espera que la distancia necesaria para frenar aumente en promedio 3.93 metros.

e) Construir e interpretar un intervalo de confianza para la pendiente del modelo.

```
modelo1 |> confint(level = 0.95)
```

```
##                   2.5 %    97.5 %
## (Intercept) -31.167850 -3.990340
## speed        3.096964  4.767853
```

Con un 95% de confianza, **por cada milla por hora** adicional de velocidad, se espera que la distancia promedio de frenado se incremente entre 3.1 y 4.77 **pies**.

f) ¿Existe una relación lineal significativa entre la velocidad y la distancia?

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$\alpha = 0.05$

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.4.3
anova = aov(dist ~ speed, cars) |> tidy()
anova
```

```
## # A tibble: 2 x 6
##   term        df  sumsq meansq statistic  p.value
##   <chr>     <dbl> <dbl>  <dbl>     <dbl>
## 1 speed       1 21185. 21185.     89.6  1.49e-12
## 2 Residuals   48 11354.   237.      NA     NA
```

p-valor = 1.49×10^{-12}

alfa = 0.05

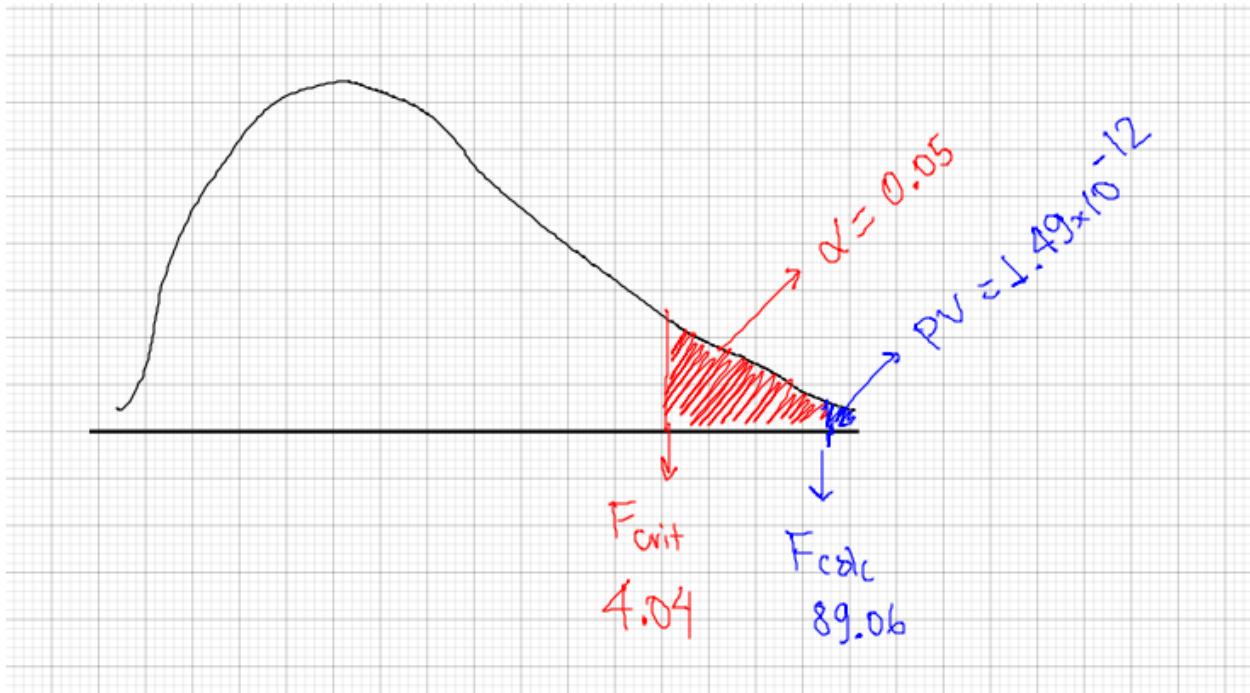
Como el pvalor es menor que el nivel de significancia, se rechaza la hipótesis nula

$F_{calc} = 89.6$

$F_{crit} = F_{0.95,1,48} = 4.04$

```
qf(0.95, 1, 48)
```

```
## [1] 4.042652
```



- g) ¿Puede afirmarse que por cada milla por hora adicional, al automóvil le toma, en promedio, más de 4 pies frenar?

$$H_0 : \beta_1 \leq 4$$

$$H_1 : \beta_1 > 4$$

$$\alpha = 0.05$$

```
modelo1 |> tidy()
```

```
## # A tibble: 2 x 5
##   term      estimate std.error statistic p.value
##   <chr>      <dbl>     <dbl>      <dbl>    <dbl>
## 1 (Intercept) -17.6      6.76     -2.60 1.23e- 2
## 2 speed        3.93      0.416      9.46 1.49e-12
```

$$t_{calc} = \frac{3.93 - 4}{0.416} = -0.1683$$

$$(3.93 - 4) / 0.416$$

```
## [1] -0.1682692
```

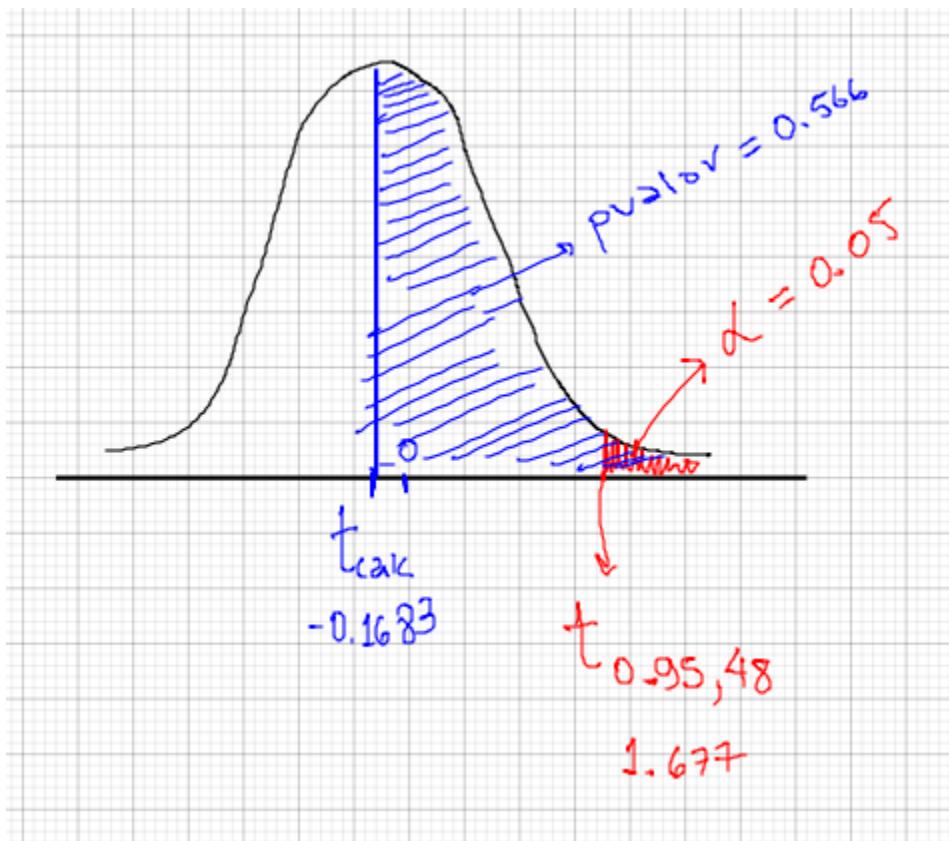
```
qt(0.95, 48)
```

```
## [1] 1.677224
```

$$pvalor = P(t > -0.1683) = 0.5664728$$

$$1 - pt(-0.1683, 48)$$

```
## [1] 0.5664728
```



No se rechaza la hipótesis nula.

En conclusión, no existe evidencia para afirmar que por cada milla por hora adicional, al automóvil le toma, en promedio, más de 4 pies frenar.

- h) Construir un intervalo de confianza para la desviación estándar del error del modelo.
(ver diapositivas 37 y 38)
- i) ¿Qué porcentaje de la variabilidad de la distancia es explicada por la velocidad?
(ver diapositivas 39 y 40)
- j) Estimar la distancia media de frenado cuando la velocidad es de 18.45 millas por hora.
(ver diapositivas 41, 42 y 43)

Pregunta 2

Se utilizará el dataset diamonds, disponible en el paquete ggplot2 R, que contiene observaciones de precios y otros atributos de 54 mil diamantes

- a) Estimar e interpretar el coeficiente de correlación de Pearson entre el precio del diamante y al menos dos de las variables independientes. Sugerencia: Usar una matriz y/o gráfica de correlaciones

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
head(diamonds)
```

```
## # A tibble: 6 x 10
```

```

##   carat cut      color clarity depth table price     x     y     z
##   <dbl> <ord>    <ord> <ord>    <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55  326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61  326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65  327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58  334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58  335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57  336  3.94  3.96  2.48

datos = diamonds[,-c(2,3,4)]
head(datos)

## # A tibble: 6 x 7
##   carat depth table price     x     y     z
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  0.23  61.5    55  326  3.95  3.98  2.43
## 2  0.21  59.8    61  326  3.89  3.84  2.31
## 3  0.23  56.9    65  327  4.05  4.07  2.31
## 4  0.29  62.4    58  334  4.2   4.23  2.63
## 5  0.31  63.3    58  335  4.34  4.35  2.75
## 6  0.24  62.8    57  336  3.94  3.96  2.48

datos |> cor() |> round(2)

##       carat depth table price     x     y     z
## carat  1.00  0.03  0.18  0.92  0.98  0.95  0.95
## depth   0.03  1.00 -0.30 -0.01 -0.03 -0.03  0.09
## table   0.18 -0.30  1.00  0.13  0.20  0.18  0.15
## price   0.92 -0.01  0.13  1.00  0.88  0.87  0.86
## x       0.98 -0.03  0.20  0.88  1.00  0.97  0.97
## y       0.95 -0.03  0.18  0.87  0.97  1.00  0.95
## z       0.95  0.09  0.15  0.86  0.97  0.95  1.00

```

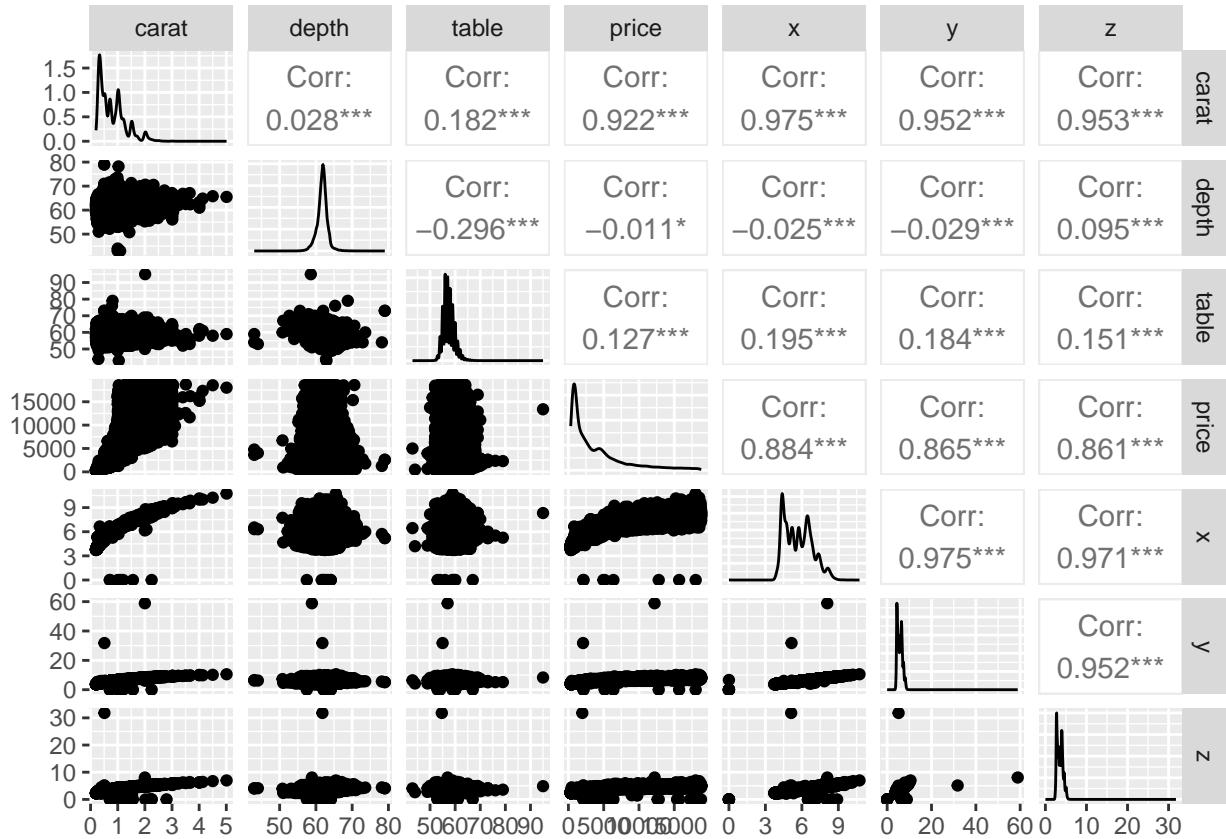
El precio del diamante tiene una muy fuerte asociación lineal directa con su peso, así también con su longitud, ancho y profundidad. En contraste a ello, su asociación es muy baja con el porcentaje total de profundidad y baja con el ancho de la parte superior del diamante en relación con el punto más ancho.

```

library(GGally)

## Warning: package 'GGally' was built under R version 4.4.3
datos |> ggpairs()

```



b) ¿Se puede concluir que la correlación entre el precio y la profundidad del diamante es estadísticamente menor a cero?

$$H_0 : \rho \geq 0$$

$$H_1 : \rho < 0$$

$$\alpha = 0.05$$

```
cor.test(datos$price, datos$z, method = "pearson", alternative = "less")
```

```
##
## Pearson's product-moment correlation
##
## data: datos$price and datos$z
## t = 393.6, df = 53938, p-value = 1
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
## -1.0000000 0.8630674
## sample estimates:
##      cor
## 0.8612494
```

$$pvalor = 1$$

No se rechaza la hipótesis nula

No existe evidencia estadística de que la correlación entre el precio y la profundidad del diamante es estadísticamente menor a cero.

- c) Formular el modelo de regresión lineal múltiple que relaciona el precio del diamante con su peso, longitud, ancho y profundidad.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

donde:

- Y : precio del diamante
- X_1 : peso del diamante
- X_2 : longitud del diamante
- X_3 : ancho del diamante
- X_4 : profundidad del diamante
- β_0 : intercepto del modelo
- $\beta_1, \beta_2, \beta_3, \beta_4$: coeficientes del modelo
- ϵ : error del modelo.

- d) Estimar de manera puntual los coeficientes del modelo e interpretarlos

```
modelo2 = lm(price ~ carat + x + y + z, datos)
modelo2 |> coef()
```

```
## (Intercept)      carat          x          y          z
##   1921.1740  10233.9134   -884.2091    166.0384   -576.2035
```

$$\hat{Y} = 1921.17 + 10233.91X_1 - 884.21X_2 + 166.0384X_3 - 576.20X_4$$

- 1921.17: no se interpreta porque no tiene sentido que el peso, longitud, ancho ni profundidad del diamante sean cero.
- 10233.91: por cada unidad adicional de peso, el precio promedio se incrementará en 10233.91 dólares, manteniendo constante la longitud, el ancho y la profundidad del diamante.
- -884.21: por cada milímetro adicional de longitud del diamante, el precio promedio disminuye en 884.21 dólares, manteniendo constante el peso, el ancho y la profundidad del diamante.
- 166.0384: por cada milímetro adicional de ancho del diamante, el precio promedio se incrementa en 166.0384 dólares, manteniendo constante el peso, la longitud y la profundidad del diamante.
- -576.20: por cada milímetro adicional de profundidad del diamante, el precio promedio disminuye en 576.20 dólares, manteniendo constante el peso, la longitud y el ancho del diamante.

- e) Construir e interpretar un intervalo de confianza para el coeficiente del ancho

```
modelo2 |> confint(level = 0.95)
```

```
##               2.5 %      97.5 %
## (Intercept) 1716.6013  2125.7467
## carat       10110.5571 10357.2697
## x           -963.5315 -804.8867
## y            115.3557  216.7211
## z           -653.1970 -499.2100
```

Por cada milímetro adicional de ancho del diamante, el precio promedio se incrementa entre 115.36 y 216.72 dólares, manteniendo constante el peso, la longitud y la profundidad del diamante.

f) ¿Existe una relación lineal significativa entre el precio de un diamante y su peso, longitud, ancho y profundidad?

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

\$H_1\$: Al menos un β es distinto de cero

$$\alpha = 0.05$$

```
X = model.matrix(price ~ carat + x + y + z, data=datos)
anova = aov(price ~ X, datos) |> tidy()
anova
```

```
## # A tibble: 2 x 6
##   term      df    sumsq     meansq statistic p.value
##   <chr>    <dbl>    <dbl>     <dbl>     <dbl>    <dbl>
## 1 X         4 733203492539. 183300873135. 78920.     0
## 2 Residuals 53935 125269642979.       2322604.     NA      NA
```

pvalue = 0, por lo tanto se rechaza la hipótesis nula.

En conclusión, al menos una de las variables contribuye en la construcción del modelo.

g) ¿Puede afirmarse que por cada milímetro adicional de longitud, el precio promedio disminuye en 800 dólares?

$$H_0 : \beta_2 = -800$$

$$H_1 : \beta_2 \neq -800$$

$$\alpha = 0.05$$

```
modelo2 |> tidy()
```

```
## # A tibble: 5 x 5
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) 1921.     104.     18.4  1.98e- 75
## 2 carat       10234.    62.9     163.     0
## 3 x          -884.      40.5    -21.8  2.32e-105
## 4 y           166.      25.9     6.42  1.36e- 10
## 5 z          -576.      39.3    -14.7  1.28e- 48
```

$$t_{calc} = \frac{-884.2091 - (-800)}{40.5} = -2.08$$

```
(-884.2091+800)/40.5
```

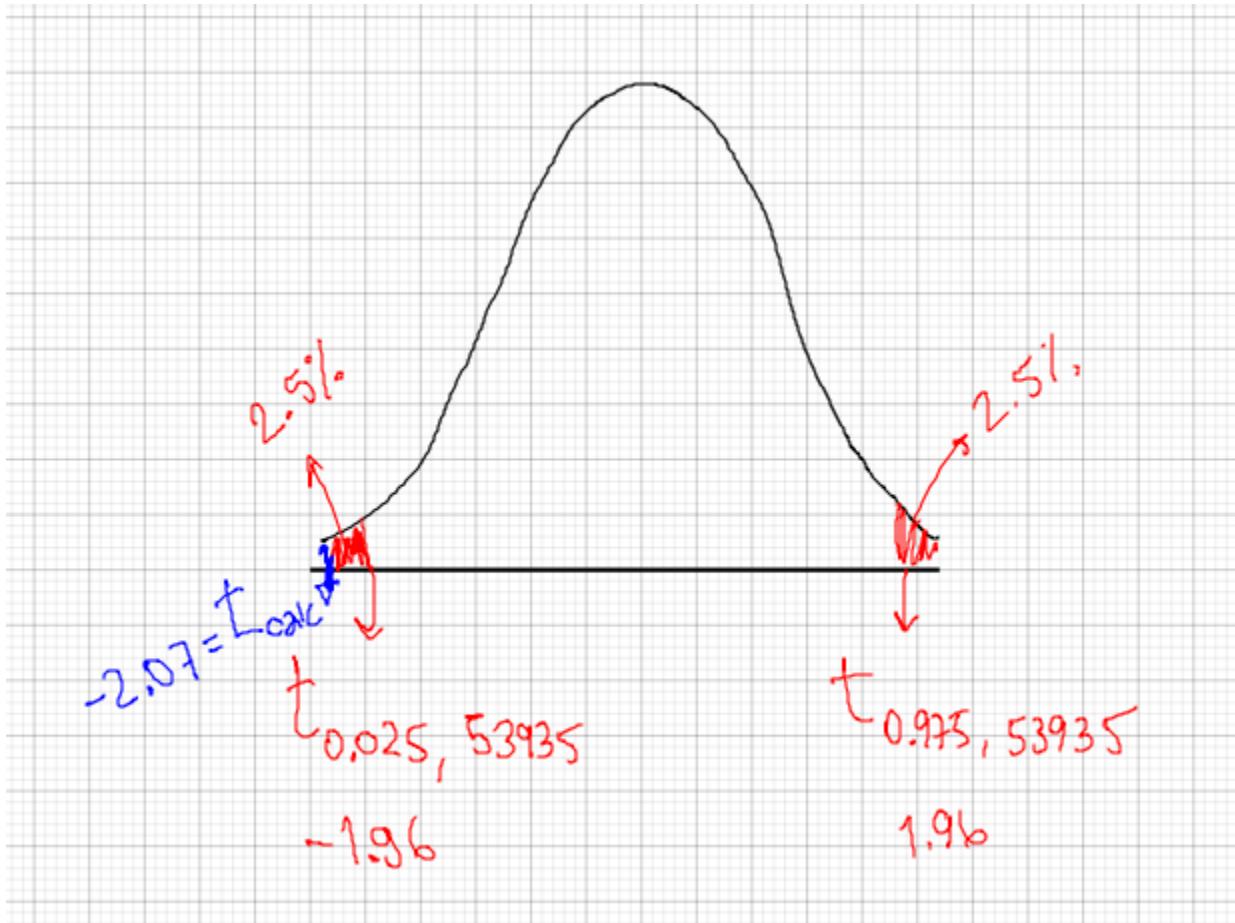
```
## [1] -2.079237
```

```
qt(0.025, 53935)
```

```
## [1] -1.960008
```

```
qt(0.975, 53935)
```

```
## [1] 1.960008
```



Se rechaza H_0 .

En conclusión, no puede afirmarse que por cada milímetro adicional de longitud, el precio promedio disminuye en 800 dólares.

- h) ¿Puede afirmarse que por cada milímetro adicional de ancho, el precio promedio aumenta como máximo 200 dólares?

(Guiarse de la pregunta g)

- i) ¿Qué porcentaje de la variabilidad del precio es explicada por el modelo?

```
summary(modelo2)$adj.r.squared
```

```
## [1] 0.8540677
```

El 85.4% de la variabilidad del precio es explicada por el modelo de regresión lineal.

- j) Estimar de manera intervalar el precio medio de un diamante con peso 0.25, longitud 4 mm, ancho 4.1 mm y profundidad 2.7 mm.

```
predict(modelo2,
        data.frame(carat = 0.25,
                   x = 4,
                   y = 4.1,
                   z = 2.7),
        interval = "confidence")
```

```
##      fit      lwr      upr
```

```
## 1 67.824 33.57465 102.0733
k ) Predecir de manera intervalar el precio de un diamante con peso 0.25, longitud 4 mm, ancho 4.1 mm y profundidad 2.7 mm.

predict(modelo2,
        data.frame(carat = 0.25,
                   x = 4,
                   y = 4.1,
                   z = 2.7),
        interval = "prediction")

##      fit      lwr      upr
## 1 67.824 -2919.442 3055.09
```