Taller 1 de AR

Joan Nicolas Castro Cortes

2022-04-05

Ejercicio 2.1

En la tabla B.1 del apéndice aparecen datos sobre el desempeño de los 26 equipos de la Liga Nacional de Fútbol en 1976. Se cree que la cantidad de yardas ganadas por tierra por los contrarios (x_8) tiene un efecto sobre la cantidad de juegos que gana un equipo (y).

```
tableB1 <- read.csv("C:/Users/nico9/Documents/Notebooks/Analisis de regresion/r/LinearModels/tableB1.cs
head(tableB1)
```

```
##          Team  y x_.1. x_.2. x_.3. x_.4. x_.5. x_.6. x_.7. x_.8. x_.9.
## 1  Washington 10  2113  1985  38,9   647     4   868  59,7  2205  1917
## 2   Minnesota 11  2003  2855  38,8   613     3   615    55  2096  1575
## 3 New England 11  2957  1737  40,1   600    14   914  65,6  1847  2175
## 4     Oakland 13  2285  2905  41,6   453    -4   957  61,4  1903  2476
## 5  Pittsburgh 10  2971  1666  39,2   538    15   836  66,1  1457  1866
## 6   Baltimore 11  2309  2927  39,7   741     8   786    61  1848  2339
```

y: Games won (per 14 - game season)
x1: Rushing yards (season)
x2: Passing yards (season)
x3: Punting average (yards/punt)
x4: Field goal percentage (FGs made/FGs attempted 2season)
x5: Turnover differential (turnovers acquired – turnovers lost)
x6: Penalty yards (season)
x7: Percent rushing (rushing plays/total plays)
x8: Opponents ' rushing yards (season)
x9: Opponents ' passing yards (season)

a. Ajustar un modelo de regresión lineal simple que relacione los juegos ganados, con las yardas ganadas por tierra por los contrarios, x_8.

Solución:

Vamos a hacer uso del método de los mínimos cuadrados para ajustar un modelo de regresión lineal simple. Teniendo en cuenta los estimadores insesgados de los parámetros β_0 y β_1 del modelo $y = \beta_0 + \beta_1 x$ que son

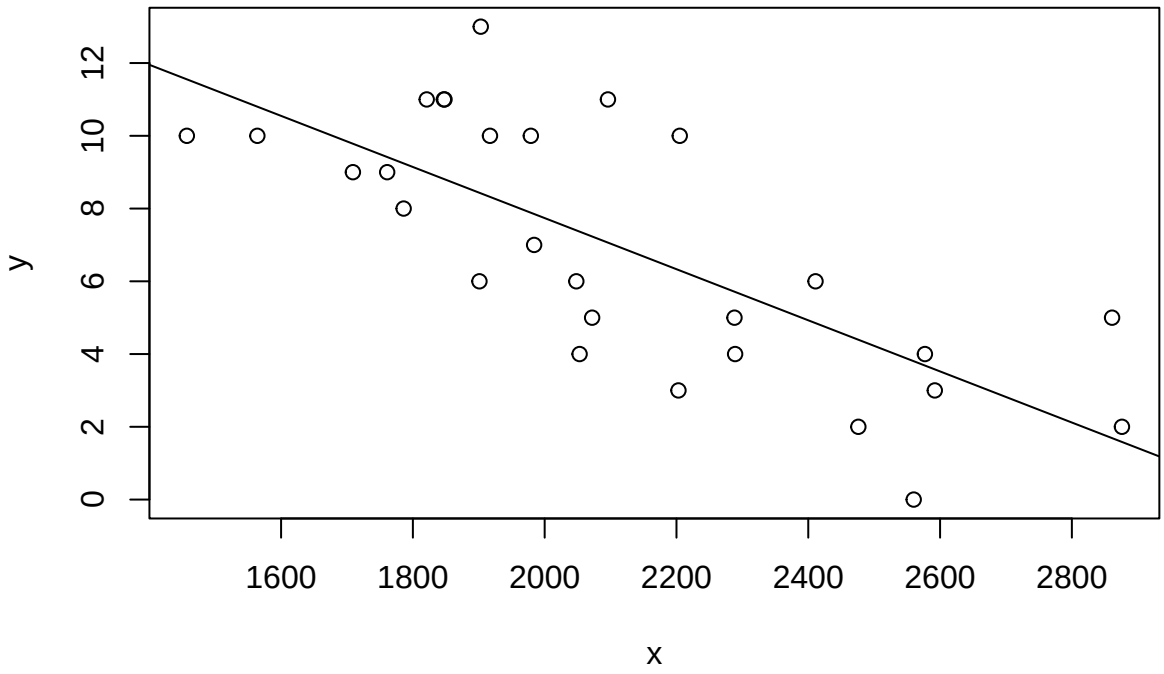
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

y

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

tenemos el modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Por lo tanto el modelo ajustado se calcula de la siguiente manera:

```r
x = tableB1[,10]
y = tableB1[,2]
meanx = mean(x)
meany = mean(y)
Sxy = sum(((x-meanx))*y)
Sxx = sum(((x-meanx)^2))
beta1 = Sxy/Sxx
beta0 = meany - beta1*meanx

plot(x,y)
abline(a=beta0,b=beta1) # Los parametros son el intercepto y la pendiente que calculamos
```



b. Formar la tabla de análisis de varianza y probar el significado de la regresión.

Solución:

- *Anova*

Para desarrollar el análisis de varianza tengamos en cuenta la identidad fundamental del análisis de varianza para el modelo de regresión que nos dice

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$
$$SS_T=SS_R+SS_{\text{Res}}$$

donde al lado izquierdo tenemos la suma de los cuadrados de las observaciones corregidas (SS_T) y al derecho

la suma de cuadrados del modelo (SS_R) mas la suma del cuadrado de los residuales (SS_{Res}).

También tenemos que los grados de libertad correspondientes vienen dados de la siguiente manera

$$df_{\text{T}} = df_{\text{R}} + df_{\text{Res}}$$
$$n - 1 = 1 + (n - 2)$$

dado que para SS_T se pierde un grado de libertad al ajustar de la forma $\sum_{i=1}^{n}(y_i - \bar{y})$ en las desviaciones $(y_i - \bar{y})$, SS_T está determinado por un único parámetro ($\hat{\beta}_1$ en $SS_R = \hat{\beta}_1 S_{xy}$), y SS_{Res} tiene dos grados de libertad menos dado que al ajustar $\sum_{i=1}^{n}(y_i - \hat{y}_i)$ se pierden dos grados de libertad dado que las desviaciones $(y_i - \hat{y}_i)$ son resultado de estimar $\hat{\beta}_0$ y $\hat{\beta}_1$

De este modo al hacer el análisis de varianza con la hipótesis nula $H_0 = \beta_1 = 0$ tenemos que el estadístico de prueba es

$$F_0 = \frac{SS_R/df_R}{SS_{\text{Res}}/df_{\text{Res}}} = \frac{SS_R/1}{SS_{\text{Res}}/(n-2)} = \frac{MS_R}{MS_{\text{Res}}} = \frac{MS_R}{\hat{\sigma}^2} \sim F_{1,n-2}$$

dado que $SS_R = MS_R/\sigma^2 \sim \chi^2_1$ y $SS_{\text{Res}} = MS_{\text{Res}}/\sigma^2 \sim \chi^2_{n-2}$. Donde rechazaremos la hipótesis $H_0 = \beta_1 = 0$ si $F_0 > F_{1-\alpha,1,n-2}$

Tenemos que para un $\alpha = 0.001$ el análisis de la varianza viene dado por el siguiente cálculo

```
alpha = 0.001

SS_T = sum((y^2))-(((sum(y))^2)/length(x))
SS_Res = SS_T - (beta1*Sxy)
SS_R = beta1*Sxy

df_T = length(x) -1
df_Res = length(x) - 2
df_R = 1

MS_Res = SS_Res/df_Res
MS_R = SS_R/df_R

F0 = MS_R/MS_Res
F0test = qf(1-alpha,df1 = 1,df2 = (length(x)-2))
pvalue_F0 = 1-pf(F0,df1 = 1,df2 = (length(x)-2))

SS_R
```

```
## [1] 178.0923
```

```
SS_Res
```

```
## [1] 148.872
```

```
SS_T
```

```
## [1] 326.9643
```

```
df_R
```

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

```
df_Res
```

```
## [1] 26
```

```
df_T
```

```
## [1] 27
```

```
MS_R
```

```
## [1] 178.0923
```

```
MS_Res
```

```
## [1] 5.725845
```

```
F0
```

```
## [1] 31.10324
```

```
F0test
```

```
## [1] 13.73897
```

```
pvalue_F0
```

```
## [1] 7.380709e-06
```

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_(0)
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_{Res}
Residual	$SS_{\text{Res}} = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_{Res}	
Total	SS_T	$n-1$		

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P value
Regression	178.0923	1	178.092	31.10324	7.380709e-06
Residual	148.872	26	5.725845		
Total	326.9643	27			

Dado que tenemos que $F_0 = 31.10324 > F_{1-\alpha,1,n-2} = 13.73897$ ***rechazamos la hipótesis que nos dice que*** $\beta_1 = 0$ ***con una significancia del*** 0.1% ***y un p valor de*** $7.380709e^{-06}$

- *Significancia de la regresión*

Para la significancia de la regresión tengamos en cuenta que, en este caso, no conocemos la varianza poblacional; entonces para el test de hipótesis $H_0 = \beta_1 = \beta_{10} = 0$ tenemos el estadístico de prueba

$$t_0 = \frac{\hat{\beta}_1 - \beta_{01}}{\sqrt{MS_{\text{Res}}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{01}}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

dado que $MS_{\text{Res}} = \hat{\sigma}^2$ es un estimador insesgado de σ^2, $(n-2)MS_{\text{Res}}/\sigma^2 \sim \chi^2_{n-2}$, y MS_{Res} con $\hat{\beta}_1$ son independientes. Donde rechazamos la hipótesis, que nos dice que tenemos una pendiente nula en este caso, si $|t_0| > t_{1-\alpha/2,n-2}$.

De este modo tenemos que para un $\alpha = 0.05$ el test de significancia de la regresión viene dado por el siguiente cálculo

```r
alpha = 0.05
se_beta1 = sqrt(MS_Res/Sxx)
t0 = (beta1-0)/se_beta1 # el estadístico de prueba preguntando si \hat{\beta}_1 = \beta_{10} = 0
t0test = qt((1 - alpha/2),df=(length(x)-2))
pvalue_t0 = 2*pt(-abs(t0),df=(length(x)-2))

t0
```

```
## [1] -5.577028
```

```r
t0test
```

```
## [1] 2.055529
```

```r
pvalue_t0
```

```
## [1] 7.380709e-06
```

Dado que tenemos que $|t_0| = 5.577028 > t_{1-\alpha/2,n-2} = 2.055529$ ***rechazamos la hipótesis que nos dice que*** $\beta_1 = 0$ ***con una significancia del*** 5% ***y un p valor de*** $7.380709e^{-06}$

Podemos verificar los resultados en R haciendo uso de las funciones `lm()` para crear el objeto correspondiente al modelo lineal, `summary()` para ver la significancia de la regresión y el intercepto y `anova()` para ver el análisis de varianza.

```r
xylm <- lm(y ~ x)
```

```r
summary(xylm)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.804 -1.591 -0.647  2.032  4.580
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.788251   2.696233   8.081 1.46e-08 ***
## x           -0.007025   0.001260  -5.577 7.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 26 degrees of freedom
## Multiple R-squared:  0.5447, Adjusted R-squared:  0.5272
## F-statistic:  31.1 on 1 and 26 DF,  p-value: 7.381e-06
```

```r
anova(xylm)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 178.09 178.092  31.103 7.381e-06 ***
## Residuals 26 148.87   5.726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. Determinar un intervalo de confianza de 95% para la pendiente.

Solución:

Teniendo en cuenta que

$$t_0 = \frac{\hat{\beta}_1 - \beta_{01}}{\sqrt{MS_{\text{Res}}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{01}}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

por lo tanto un intervalo de confianza de $100(1-\alpha)\%$ de la pendiente β_1 está dado por

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\,\text{se}\left(\hat{\beta}_1\right) \leq \hat{\beta}_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\,\text{se}\left(\hat{\beta}_1\right)$$

de este modo tenemos que el intervalo de confianza de $100(1-\alpha)\% = 95\%$ de la pendiente dadas las observaciones es

```
alpha = 0.05
se_beta1 = sqrt(MS_Res/Sxx)

IC_beta1_inf = beta1-(qt((1-alpha/2),df=(length(x)-2))*(se_beta1))
IC_beta1_sup = beta1+(qt((1-alpha/2),df=(length(x)-2))*(se_beta1))
IC_beta1 = c(IC_beta1_inf,IC_beta1_sup)
IC_beta1
```

```
## [1] -0.009614347 -0.004435854
```

Con una confianza del 95% el valor real del parámetro β_1 ***se encuentra entre*** $(-0.009614347, -0.004435854)$

d. ¿Qué porcentaje de variabilidad total da y, y explica este modelo?

Solución:

Dado que SS_T es la medida de la variabilidad en y sin considerar el efecto de la variable regresora x y SS_{Res} es la medida de la variabilidad sobrante despues de considerar la variable regresora x tenemos que el coeficiente de determinación $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{\text{Res}}}{SS_T}$ es considerado también la proporción de la variación explicada por el regresor x. Por lo que tenemos

```
R2 = SS_R/SS_T
R2
```

```
## [1] 0.5446843
```

El 54.44% ***de la fuerza de la variabilidad se explica en el modelo de regresión***

Para explicar el modelo tengamos en cuenta primero que deberíamos suponer que el intercepto es $\beta_0 = 14$ dado que es plausible que si un equipo tiene 0 yardas gandas durante una temporada casi seguramente gane los 14 partidos de la temporada. Esto se refuerza con el hecho de tener una pendiente negativa puesto que a más yardas ganadas por los contrarios menos juegos ganan los equipos.

e. Determinar un intervalo de confianza de 95% para la cantidad promedio de juegos ganados, si la distancia ganada por tierra por los contrarios se limita a 2000 yardas.

Solución:

Nos piden calcular la respuesta media $E(y)$ de juegos ganados para un número para un número de yardas ganadas del equipo contrario $x_0 = 2000$. Dado que x_0 se encuentra dentro de los valores de yardas con los que se planteo el modelo, podemos decir que un estimador insesgado de $E(y|x_0)$ es

$$\widehat{E(y \mid x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

y su varianza es

$$\begin{aligned}\operatorname{Var}\left(\hat{\mu}_{y|x_0}\right) &= \operatorname{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) = \operatorname{Var}\left[\bar{y} + \hat{\beta}_1\left(x_0 - \bar{x}\right)\right] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2\left(x_0 - \bar{x}\right)^2}{S_{xx}} = \sigma^2\left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]\end{aligned}$$

Tambien vemos que como $E(y|x) = \hat{\mu}_{y|x_0}$ es es una combinación lineal de la y_i tenemos que

$$\hat{\mu}_{y|x_0} \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]\right)$$

Ahora para construir un estimador insesgado como $\operatorname{Cov}(\bar{y}, \hat{\beta}_1) = 0$

$$\frac{\hat{\mu}_{y|x_0} - E\left(y \mid x_0\right)}{\sqrt{MS_{\text{Res}}\left(1/n + \left(x_0 - \bar{x}\right)^2 / S_{xx}\right)}} \sim t_{(n-2)}$$

lo que nos deja que un intervalos de confianza de $100(1-\alpha)\%$ de la respuesta media en un punto $x = x_0$ es

$$\begin{aligned}\hat{\mu}_{y|x_0} - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right)} \leq \\ E\left(y \mid x_0\right) \leq \\ \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right)}\end{aligned}$$

entonces ***el intervalo de confianza del*** 95% ***de la respuesta media de juegos ganados para una avance de limitado a*** 2000 ***yardas es***

```
x0 = 2000

muyx0 = beta0 + beta1*x0
se_muyx0 = sqrt( MS_Res*( (1/length(x)) + ((x0-meanx)^2)/Sxx ) )
IC_muyx0_inf = muyx0 - (qt((1-alpha/2),df=(length(x)-2)))*se_muyx0
IC_muyx0_sup = muyx0 + (qt((1-alpha/2),df=(length(x)-2)))*se_muyx0
IC_muyx0 = c(IC_muyx0_inf,IC_muyx0_sup )
IC_muyx0
```

```
## [1] 6.765753 8.710348
```

Ejercicio 2.2

Supóngase que se quiere usar el modelo desarrollado en el problema 2.1 para pronosticar la cantidad de juegos que ganará un equipo si puede limitar los avances por tierra de sus contrarios a 1800 yardas. Determinar un estimado de punto de la cantidad de juegos ganados cuando.$x_8 = 1800$. Determinar un intervalo de predicción de 90% para la cantidad de juegos ganados.

```
tableB1 <- read.csv("C:/Users/nico9/Documents/Notebooks/Analisis de regresion/r/LinearModels/tableB1.csv
head(tableB1)
```

```
##          Team  y x_.1. x_.2. x_.3. x_.4. x_.5. x_.6. x_.7. x_.8. x_.9.
## 1  Washington 10  2113  1985  38,9   647     4   868  59,7  2205  1917
## 2   Minnesota 11  2003  2855  38,8   613     3   615    55  2096  1575
## 3 New England 11  2957  1737  40,1   600    14   914  65,6  1847  2175
## 4     Oakland 13  2285  2905  41,6   453    -4   957  61,4  1903  2476
## 5  Pittsburgh 10  2971  1666  39,2   538    15   836  66,1  1457  1866
## 6   Baltimore 11  2309  2927  39,7   741     8   786    61  1848  2339
```

Solución:

Se hizo uso del método de los mínimos cuadrados para ajustar un modelo de regresión lineal simple del ejercicio 2.1. Teniendo en cuenta los estimadores insesgados de los parámetros β_0 y β_1 del modelo $y = \beta_0 + \beta_1 x$ que son

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

y

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

de este modo tenemos el modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Por lo tanto el modelo ajustado se calculó en el ejercicio 2.1 de la siguiente manera:

```
x = tableB1[,10] # x_8
y = tableB1[,2] # y
meanx = mean(x)
meany = mean(y)
Sxy = sum(((x-meanx))*y)
Sxx = sum(((x-meanx)^2))
beta1 = Sxy/Sxx
beta0 = meany - beta1*meanx
```

donde la estimación puntual de los juegos ganados viene dada por $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

```
x_0 = 1800

y_0 = beta0 + beta1*x_0
y_0
```

```
## [1] 9.14307
```

Por lo tanto se espera, segun el modelo, que un equipo que restringe el avance de yardas del equipo contrario a $x_0 = 1800$ ***según el modelo de regresión gane*** 9.14 ***partidos.***

Ahora para construir un intervalo de predicción tengamos en cuenta la variable aleatoria

$$\psi = y_0 - \hat{y}_0$$

donde $y_0 = \beta_0 + \beta_1 x_0$ y $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Como son combinaciones lineales de distribuciones normales vemos que ψ tiene distribución normal de media cero y varianza

$$\text{Var}(\psi) = \text{Var}\left(y_0 - \hat{y}_0\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

dado que y_0 y $\hat{y}_0$ son independientes. De este modo tenemos que la desviación estándar estimada es la estadística apropiada y así el intervalo de predicción es

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

entonces tenemos que

```
alpha = 0.1
x0=1800

SS_T = sum((y^2))-(((sum(y))^2)/length(x)) # Suma de cuadrados de las observaciones corregidas
SS_Res = SS_T - (beta1*Sxy) # Suma de cuadrados de los residuales
SS_R = beta1*Sxy # Suma de cuadrados de la regresion o suma de cuadrados del modelo

df_T = length(x) -1 # Grados de libertad de las observaciones corregidas
df_Res = length(x) - 2 # Grados de libertad de la suma de cuadrados de los residuales
df_R = 1 # Grados de libertad de la suma de cuadrados de la regresión


MS_Res = SS_Res/df_Res
MS_R = SS_R/df_R
MS_Res = SS_Res/df_Res

y0 = beta0 + beta1*x0
se_y0 = sqrt( MS_Res*(1 + (1/length(x)) + ((x0-meanx)^2)/Sxx ) ) # desviación de \hat{y}_0

IC_y0_inf = y0 - qt((1-alpha/2),df=(length(x)-2))*se_y0
IC_y0_sup = y0 + qt((1-alpha/2),df=(length(x)-2))*se_y0
IC_y0 = c(IC_y0_inf,IC_y0_sup)
IC_y0
```

```
## [1]  4.936392 13.349749
```

Por lo tanto un intervalo de confianza del 90% ***de predicción de numero de juegos ganados si se restringe el avance de yardas del equipo contrario a*** $x_0 = 1800$ ***es de*** $(4.93639213.349749)$

Ejercicio 2.4

La tabla B.3 del apéndice contiene datos sobre el rendimiento de la gasolina, en millas, de 32 automóviles diferentes.

```
tableB3 <- read.csv("C:/Users/nico9/Documents/Notebooks/Analisis de regresion/r/LinearModels/tableB3.cs
head(tableB3)
```

```
##      Automobile     y x_.1. x_.2. x_.3.    x_.4.    x_.5. x_.6. x_.7. x_.8.
## 1        Apollo 18.90   350   165   260  8.0 : 1 2.56 : 1     4     3  2003
## 2         Omega 17.00   350   170   275  8.5 : 1 2.56 : 1     4     3  1996
## 3          Nova 20.00   250   105   185 8.25 : 1 2.73 : 1     1     3  1967
## 4       Monarch 18.25   351   143   255  8.0 : 1 3.00 : 1     2     3  1999
## 5        Duster 20.07   225    95   170  8.4 : 1 2.76 : 1     1     3  1941
## 6 Jenson\nConv. 11.20   440   215   330  8.2 : 1 2.88 : 1     4     3  1845
##   x_.9. x_.10. x_.11.
## 1   699   3910      A
## 2   729   2860      A
## 3   722   3510      A
## 4   740   3890      A
## 5   718   3365      M
## 6    69   4215      A
```

y : Miles/gallon	x_6 : Carburetor (barrels)
x_1 : Displacement (cubic in.)	x_7 : No. of transmission speeds
x_2 : Horsepower (ft-lb)	x_8 : Overall length (in.)
x_3 : Torqne (ft-lb)	x_9 : Width (in.)
x_4 : Compression ratio	x_{10} : Weight (lb)
x_5 : Rear axle ratio	x_{11} : Type of transmission (A automatic; M manual)

Source: Motor Trend, 1975.

a. Ajustar un modelo de regresión lineal simple que relacione el rendimiento de la gasolina y (millas por galón) y la cilindrada del motor x_1 (pulgadas cúbicas).

Solución:

Vamos a hacer uso del método de los mínimos cuadrados para ajustar un modelo de regresión lineal simple. Teniendo en cuenta los estimadores insesgados de los parámetros β_0 y β_1 del modelo $y = \beta_0 + \beta_1 x$ que son

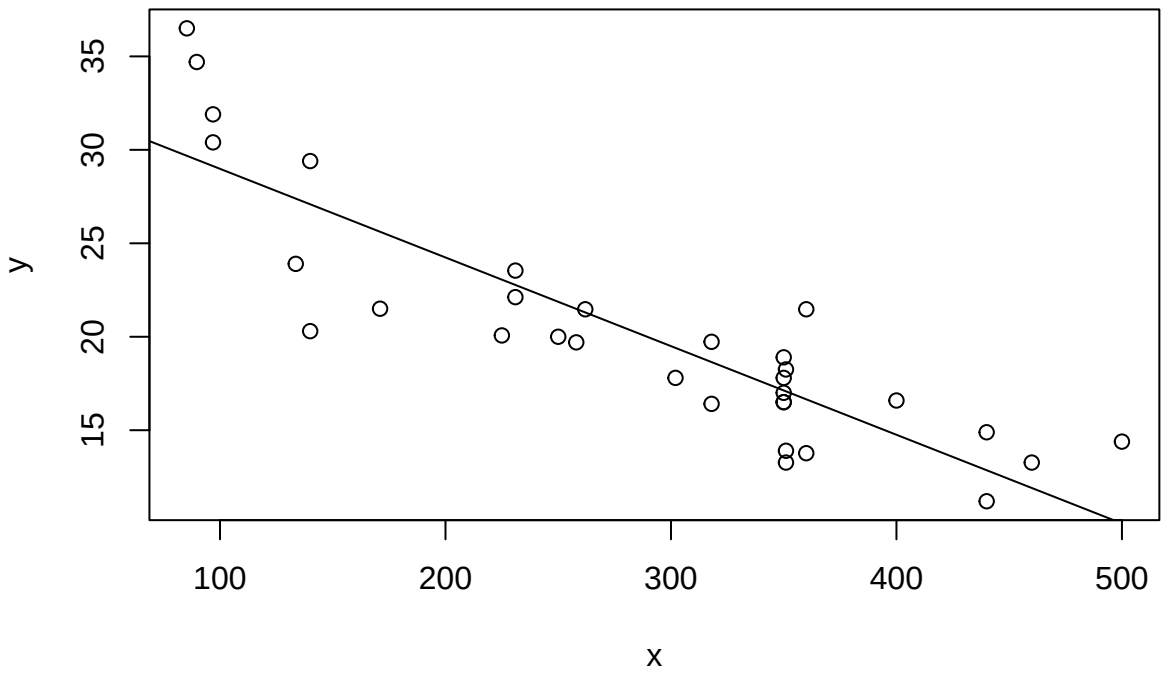
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

y

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

tenemos el modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Por lo tanto el modelo ajustado se calcula de la siguiente manera:

```
x = tableB3[,3]
y = tableB3[,2]
meanx = mean(x)
meany = mean(y)
Sxy = sum(((x-meanx))*y)
Sxx = sum(((x-meanx)^2))
beta1 = Sxy/Sxx
beta0 = meany - beta1*meanx

plot(x,y)
abline(a=beta0,b=beta1) # Los parametros son el intercepto y la pendiente que calculamos
```



b. Formar la tabla de análisis de varianza y prueba de significancia de la regresión.

Solución:

- *Anova*

Para desarrollar el análisis de varianza tengamos en cuenta la identidad fundamental del análisis de varianza para el modelo de regresión que nos dice

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$
$$SS_T=SS_R+SS_{\text{Res}}$$

donde al lado izquierdo tenemos la suma de los cuadrados de las observaciones corregidas (SS_T) y al derecho la suma de cuadrados del modelo (SS_R) mas la suma del cuadrado de los residuales (SS_{Res}).

También tenemos que los grados de libertad correspondientes vienen dados de la siguiente manera

$$df_{\text{T}}=df_{\text{R}}+df_{\text{Res}}$$
$$n-1=1+(n-2)$$

dado que para SS_T se pierde un grado de libertad al ajustar de la forma $\sum_{i=1}^{n}(y_i-\bar{y})$ en las desviaciones $(y_i-\bar{y})$; SS_T está determinado por un único parámetro ($\hat{\beta}_1$ en $SS_R=\hat{\beta}_1 S_{xy}$), y SS_{Res} tiene dos grados de libertad menos dado que al ajustar $\sum_{i=1}^{n}(y_i-\hat{y}_i)$ se pierden dos grados de libertad dado que las desviaciones $(y_i-\hat{y}_i)$ son resultado de estimar $\hat{\beta}_0$ y $\hat{\beta}_1$

De este modo al hacer el análisis de variancia con la hipótesis nula $H_0 = \beta_1 = 0$ tenemos que el estadístico de prueba es

$$F_0 = \frac{SS_R/df_R}{SS_{\text{Res}/df_{\text{Res}}}} = \frac{SS_R/1}{SS_{\text{Res}}/(n-2)} = \frac{MS_R}{MS_{\text{Res}}} = \frac{MS_R}{\hat{\sigma}^2} \sim F_{1,n-2}$$

dado que $SS_R = MS_R/\sigma^2 \sim \chi^2_1$ y $SS_{\text{Res}} = MS_{\text{Res}}/\sigma^2 \sim \chi^2_{n-2}$. Donde rechazaremos la hipótesis $H_0 = \beta_1 = 0$ si $F_0 > F_{1-\alpha,1,n-2}$

De este modo tenemos que para un $\alpha = 0.001$ el análisis de la varianza viene dado por el siguiente cálculo

```
alpha = 0.001

SS_T = sum((y^2))-(((sum(y))^2)/length(x))
SS_Res = SS_T - (beta1*Sxy)
MS_Res = SS_Res/(length(x) - 2)
SS_R = beta1*Sxy
df_R = 1
df_Res = length(x) - 2
df_T = length(x) -1
MS_R = SS_R/df_R
MS_Res = SS_Res/df_Res
F0 = MS_R/MS_Res
F0test = qf(1-alpha,df1 = 1,df2 = (length(x)-2))
pvalue_F0 = 1-pf(F0,df1 = 1,df2 = (length(x)-2))

SS_R
```

```
## [1] 955.3404
```

```
SS_Res
```

```
## [1] 282.2037
```

```
SS_T
```

```
## [1] 1237.544
```

```
df_R
```

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

```
df_Res
```

```
## [1] 30
```

```
df_T
```

```
## [1] 31
```

```
MS_R
```

```
## [1] 955.3404
```

```
MS_Res
```

```
## [1] 9.406791
```

```
F0
```

```
## [1] 101.5586
```

```
F0test
```

```
## [1] 13.29301
```

```
pvalue_F0
```

```
## [1] 3.820033e-11
```

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_(0)
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_{Res}
Residual	$SS_{\text{Res}} = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_{Res}	
Total	SS_T	$n-1$		

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P value
Regression	955.3404	1	955.3404	101.5586	3.820033e-11
Residual	282.2037	30	9.406791		
Total	1237.544	31			

Dado que tenemos que $F_0 = 101.5586 > F_{1-\alpha,1,n-2} = 13.29301$ ***rechazamos la hipótesis que nos dice que*** $\beta_1 = 0$ ***con una significancia del*** 0.1% ***y un p valor de*** $3.820033e^{-11}$

- *Significancia de la regresión*

Para la significancia de la regresión tengamos en cuenta que como no conocemos la varianza poblacional para el test de hipótesis $H_0 = \beta_1 = \beta_{10} = 0$ tenemos el estadístico de prueba

$$t_0 = \frac{\hat{\beta}_1 - \beta_{01}}{\sqrt{MS_{\text{Res}}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{01}}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

dado que $MS_{\text{Res}} = \hat{\sigma}^2$ es un estimador insesgado de σ^2, $(n-2)MS_{\text{Res}}/\sigma^2 \sim \chi^2_{n-2}$ y MS_{Res} con $\hat{\beta}_1$ son independientes. Donde rechazamos la pendiente nula en este caso si $|t_0| > t_{1-\alpha/2,n-2}$.

De este modo tenemos que para un $\alpha = 0.05$ el test de significancia de la regresión viene dado por el siguiente cálculo

```
alpha = 0.05
se_beta1 = sqrt(MS_Res/Sxx)
t0 = (beta1-0)/se_beta1 # el estadístico de prueba preguntando si \hat{\beta}_1 = \beta_{10} = 0
t0test = qt((1 - alpha/2),df=(length(x)-2))
pvalue_t0 = 2*pt(-abs(t0),df=(length(x)-2))

t0
```

```
## [1] -10.07763
```

```
t0test
```

```
## [1] 2.042272
```

```
pvalue_t0
```

```
## [1] 3.820034e-11
```

Dado que tenemos que $|t_0| = 10.07763 > t_{1-\alpha/2,n-2} = 2.042272$ ***rechazamos la hipótesis que nos dice que*** $\beta_1 = 0$ ***con una significancia del*** 5% ***y un p valor de*** $3.820034e^{-11}$

Podemos verificar los resultados en R haciendo uso de las funciones `lm()` para crear el objeto correspondiente al modelo lineal, `summary()` para ver la significancia de la regresión y el intercepto y `anova()` para ver el análisis de varianza.

```
xylm <- lm(y ~ x)
```

```
summary(xylm)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7875 -1.9616  0.0206  1.7878  6.8182
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.727439   1.445559   23.33  < 2e-16 ***
## x           -0.047428   0.004706  -10.08 3.82e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.067 on 30 degrees of freedom
## Multiple R-squared:  0.772,  Adjusted R-squared:  0.7644
## F-statistic: 101.6 on 1 and 30 DF,  p-value: 3.82e-11
```

```
anova(xylm)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x          1 955.34  955.34  101.56 3.82e-11 ***
## Residuals 30 282.20    9.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. ¿Qué porcentaje de la variabilidad total del rendimiento de la gasolina explica la relación lineal con la cilindrada del motor?

Solución:

Dado que SS_T es la medida de la variabilidad en y sin considerar el efecto de la variable regresora x y SS_{Res} es la medida de la variabilidad sobrante después de considerar la variable regresora x tenemos que el coeficiente de determinación $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{\text{Res}}}{SS_T}$ es considerado también la proporción de la variación explicada por el regresor x. Por lo que tenemos

```
R2 = SS_R/SS_T
R2
```

```
## [1] 0.7719647
```

El 77.19% ***de la fuerza de la variabilidad se explica en el modelo de regresión***

d. Determinar un intervalo de confianza de 95% para el rendimiento promedio de gasolina, si el desplaza-

miento del motor es 275 pulg^3.

Solución:

Nos piden calcular la respuesta media $E(y)$ de rendimiento promedio de gasolina para un desplazamiento del motor de $x_0 = 275$ pulg^3. Dado que x_0 se encuentra dentro de los valores de desplazamiento de motor que se plantearon el modelo podemos decir que un estimador insesgado de $E(y|x_0)$ es

$$\widehat{E\left(y \mid x_0\right)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

y su varianza es

$$\begin{aligned}\operatorname{Var}\left(\hat{\mu}_{y|x_0}\right) &= \operatorname{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) = \operatorname{Var}\left[\bar{y} + \hat{\beta}_1\left(x_0 - \bar{x}\right)\right] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2\left(x_0 - \bar{x}\right)^2}{S_{xx}} = \sigma^2\left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]\end{aligned}$$

Tambien vemos que como $E(y|x) = \hat{\mu}_{y|x_0}$ es es una combinación lineal de la y_i, luego tenemos que

$$\hat{\mu}_{y|x_0} \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]\right)$$

Ahora para construir un estimador insesgado, como $\operatorname{Cov}(\bar{y}, \hat{\beta}_1) = 0$,

$$\frac{\hat{\mu}_{y|x_0} - E\left(y \mid x_0\right)}{\sqrt{MS_{\text{Res}}\left(1/n + \left(x_0 - \bar{x}\right)^2 / S_{xx}\right)}} \sim t_{(n-2)}$$

lo que nos deja que un intervalos de confianza de $100(1-\alpha)\%$ de la respuesta media en un punto $x = x_0$ es

$$\begin{aligned}\hat{\mu}_{y|x_0} - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right)} \leq \\ E\left(y \mid x_0\right) \leq \\ \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right)}\end{aligned}$$

entonces ***el intervalo de confianza del*** 95% ***de la respuesta media de rendimiento promedio del motor para un desplazamiento del motor de*** 275 ***pulg***^{3} ***es***

```
x0 = 275

muyx0 = beta0 + beta1*x0
se_muyx0 = sqrt( MS_Res*( (1/length(x)) + ((x0-meanx)^2)/Sxx ) )
IC_muyx0_inf = muyx0 - (qt((1-alpha/2),df=(length(x)-2)))*se_muyx0
IC_muyx0_sup = muyx0 + (qt((1-alpha/2),df=(length(x)-2)))*se_muyx0
IC_muyx0 = c(IC_muyx0_inf,IC_muyx0_sup )
IC_muyx0
```

```
## [1] 19.57343 21.79589
```

e. Suponer que se desea pronosticar el rendimiento de gasolina que tiene un coche con motor de 275 pulg^3. Determine un estimado puntual para el rendimiento. Determinar un intervalo de predicción de 95% para el rendimiento.

Solución:

La estimación puntual de los juegos ganados viene dada por $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

```
x_0 = 275

y_0 = beta0 + beta1*x_0
y_0
```

```
## [1] 20.68466
```

Por lo tanto se espera que un automóvil con un desplazamiento de cilindraje de $x_0 = 275$ ***según el modelo de regresión tenga un rendimiento de gasolina de*** 20.68466

Ahora para construir un intervalo de predicción tengamos en cuenta la variable aleatoria

$$\psi = y_0 - \hat{y}_0$$

donde $y_0 = \beta_0 + \beta_1 x_0$ y $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Como son combinaciones lineales de distribuciones normales vemos que ψ tiene distribución normal de media cero y varianza

$$\text{Var}(\psi) = \text{Var}\left(y_0 - \hat{y}_0\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

dado que y_0 y $\hat{y}_0$ son independientes. De este modo tenemos que la desviación estándar estimada es la estadística apropiada y así el intervalo de predicción es

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

entonces tenemos que

```
alpha = 0.05
x0=275

SS_T = sum((y^2))-(((sum(y))^2)/length(x)) # Suma de cuadrados de las observaciones corregidas
SS_Res = SS_T - (beta1*Sxy) # Suma de cuadrados de los residuales
SS_R = beta1*Sxy # Suma de cuadrados de la regresion o suma de cuadrados del modelo

df_T = length(x) -1 # Grados de libertad de las observaciones corregidas
df_Res = length(x) - 2 # Grados de libertad de la suma de cuadrados de los residuales
df_R = 1 # Grados de libertad de la suma de cuadrados de la regresión
```

```
MS_Res = SS_Res/df_Res
MS_R = SS_R/df_R
MS_Res = SS_Res/df_Res

y0 = beta0 + beta1*x0
se_y0 = sqrt( MS_Res*(1 + (1/length(x)) + ((x0-meanx)^2)/Sxx ) ) # desviación de \hat{y}_0

IC_y0_inf = y0 - qt((1-alpha/2),df=(length(x)-2))*se_y0
IC_y0_sup = y0 + qt((1-alpha/2),df=(length(x)-2))*se_y0
IC_y0 = c(IC_y0_inf,IC_y0_sup)
IC_y0
```

```
## [1] 14.32311 27.04622
```

Por lo tanto un intervalo de confianza del 95% ***de predicción del rendimiento de la gasolina dado una cilindrada de*** $x_0 = 275$ ***es de*** $(14.32311, 27.04622)$

f. Comparar los dos intervalos obtenidos en las partes d y e. Explicar la diferencia entre ellos. ¿Cuál es más amplio y por qué?

Solución:

En cuestiones de amplitud de los intervalos tenemos que

```
IC_muyx0_sup- IC_muyx0_inf
```

```
## [1] 2.222457
```

```
IC_y0_sup-IC_y0_inf
```

```
## [1] 12.72311
```

esto se debe a que la eficiencia en términos de la amplitud del intervalo es mejor en valores x_0 cercanos a $\bar{x}$ que es

```
meanx
```

```
## [1] 284.7312
```

esta eficiencia puede cambiar a medida que intentemos estimar un intervalo para un valor x_0 muy alejado del promedio.

Ejercicio 2.6

La tabla B.4 del apéndice presenta datos de 27 casas vendidas en Erie, Pensilvana.

```
tableB4 <- read.csv("C:/Users/nico9/Documents/Notebooks/Analisis de regresion/r/LinearModels/tableB4.csv
head(tableB4)
```

```
##      y  x_.1. x_.2. x_.3. x_.4. x_.5. x_.6. x_.7. x_.8. x_.9.
## 1 25.9 4.9176    10 34720  9980    10     7     4    42     0
## 2 29.5 5.0208    10 35310 15000    20     7     4    62     0
## 3 27.9 4.5429    10 22750 11750    10     6     3    40     0
## 4 25.9 4.5573    10 40500 12320    10     6     3    54     0
## 5 29.9 5.0597    10 44550 11210    10     6     3    42     0
## 6 29.9 3.8910    10 44550  9880    10     6     3    56     0
```

y: Sale price of the house/1000

x1: Taxes (local, school, county)/1000

x2: Number of baths

x3: Lot size (sq ft $\times$ 1000)

x4: Living space (sq ft $\times$ 1000)

x5: Number of garage stalls

x6: Number of rooms

x7: Number of bedrooms

x8: Age of the home (years)

x9: Number of fireplaces

Source: " Prediction, Linear Regression and Minimum Sum of Relative Errors, " by S. C. Narula and J. F. Wellington, Technometrics, 19, 1977. Also see " Letter to the Editor, " Technometrics, 22, 1980.

a. Ajustar un modelo de regresión lineal simple que relacione el precio de venta de la casa con los impuestos actuales (x_1).

Solución:

Vamos a hacer uso del método de los mínimos cuadrados para ajustar un modelo de regresión lineal simple al precio de las casas explicado por el monto de impuestos actuales. Teniendo en cuenta los estimadores insesgados de los parámetros β_0 y β_1 del modelo $y = \beta_0 + \beta_1 x$ que son

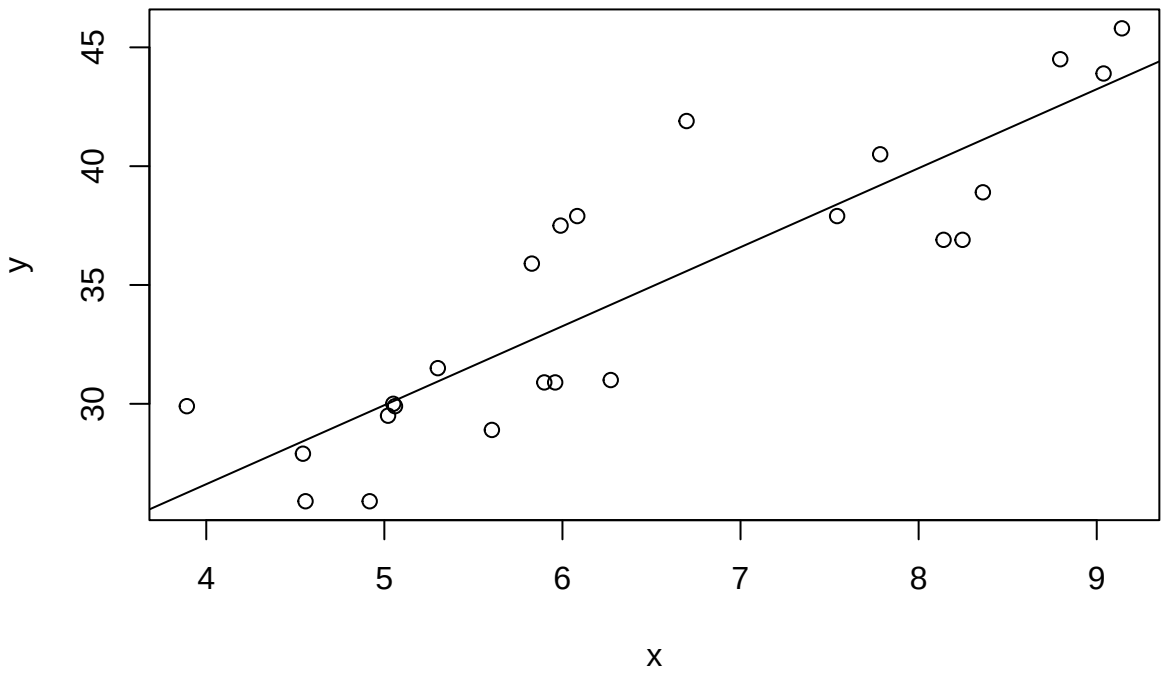
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

y

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

tenemos el modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Por lo tanto el modelo ajustado se calcula de la siguiente manera:

```
x = tableB4[,2]
y = tableB4[,1]
meanx = mean(x)
meany = mean(y)
Sxy = sum(((x-meanx))*y)
Sxx = sum(((x-meanx)^2))
beta1 = Sxy/Sxx
beta0 = meany - beta1*meanx

plot(x,y)
abline(a=beta0,b=beta1) # Los parametros son el intercepto y la pendiente que calculamos
```



b. Probar la significancia de la regresión.

Solución:

Para la significancia de la regresión tengamos en cuenta que, en este caso, no conocemos la varianza poblacional; entonces para el test de hipótesis $H_0 = \beta_1 = \beta_{10} = 0$ tenemos el estadístico de prueba

$$t_0 = \frac{\hat{\beta}_1 - \beta_{01}}{\sqrt{MS_{\text{Res}}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{01}}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

dado que $MS_{\text{Res}} = \hat{\sigma}^2$ es un estimador insesgado de σ^2, $(n-2)MS_{\text{Res}}/\sigma^2 \sim \chi^2_{n-2}$, y MS_{Res} con $\hat{\beta}_1$ son independientes. Donde rechazamos la hipótesis, que nos dice que tenemos una pendiente nula en este caso, si $|t_0| > t_{1-\alpha/2,n-2}$.

De este modo tenemos que para un $\alpha = 0.05$ el test de significancia de la regresión viene dado por el siguiente cálculo

```
alpha = 0.05

SS_T = sum((y^2))-(((sum(y))^2)/length(x)) # Suma de cuadrados de las observaciones corregidas
SS_Res = SS_T - (beta1*Sxy) # Suma de cuadrados de los residuales
SS_R = beta1*Sxy # Suma de cuadrados de la regresion o suma de cuadrados del modelo

df_T = length(x) -1 # Grados de libertad de las observaciones corregidas
df_Res = length(x) - 2 # Grados de libertad de la suma de cuadrados de los residuales
df_R = 1 # Grados de libertad de la suma de cuadrados de la regresión
```

```r
MS_Res = SS_Res/df_Res
MS_R = SS_R/df_R

se_beta1 = sqrt(MS_Res/Sxx)

t0 = (beta1-0)/se_beta1 # el estadístico de prueba preguntando si \hat{\beta}_1 = \beta_{10} = 0
t0test = qt((1 - alpha/2),df=(length(x)-2))
pvalue_t0 = 2*pt(-abs(t0),df=(length(x)-2))

t0
```

```
## [1] 8.517998
```

```r
t0test
```

```
## [1] 2.073873
```

```r
pvalue_t0
```

```
## [1] 2.051257e-08
```

Dado que tenemos que $|t_0| = 8.517998 > t_{1-\alpha/2,n-2} = 2.073873$ ***rechazamos la hipótesis que nos dice que*** $\beta_1 = 0$ ***con una significancia del*** 5% ***y un p valor de*** $2.051257e^{-08}$

Podemos verificar los resultados con la función `lm()` para crear un objeto de modelo lineal y `summary` para ver las diferentes caracteríticas de este

```r
xylm <- lm(y ~ x)
summary(xylm)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8343 -2.3157 -0.3669  1.9787  6.3168
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.3202     2.5717   5.179 3.42e-05 ***
## x             3.3244     0.3903   8.518 2.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 22 degrees of freedom
## Multiple R-squared:  0.7673, Adjusted R-squared:  0.7568
## F-statistic: 72.56 on 1 and 22 DF,  p-value: 2.051e-08
```

c. ¿Qué porcentaje de la variabilidad total del precio de venta queda explicado con este modelo?

Solución:

Dado que SS_T es la medida de la variabilidad en y sin considerar el efecto de la variable regresora x y SS_{Res} es la medida de la variabilidad sobrante después de considerar la variable regresora x tenemos que el coeficiente de determinación $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{\text{Res}}}{SS_T}$ es considerado también la proporción de la variación explicada por el regresor x. Por lo que tenemos

```
R2 = SS_R/SS_T
R2
```

```
## [1] 0.7673344
```

El 76.73% ***de la fuerza de la variabilidad se explica en el modelo de regresión***

d. Determinar un intervalo de confianza de 95% para β_1

Solución:

Teniendo en cuenta que

$$t_0 = \frac{\hat{\beta}_1 - \beta_{01}}{\sqrt{MS_{\text{Res}}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{01}}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

por lo tanto un intervalo de confianza de $100(1-\alpha)\%$ de la pendiente β_1 está dado por

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\,\text{se}\left(\hat{\beta}_1\right) \leq \hat{\beta}_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\,\text{se}\left(\hat{\beta}_1\right)$$

de este modo tenemos que el intervalo de confianza de $100(1-\alpha)\% = 95\%$ de la pendiente dadas las observaciones es

```
alpha = 0.05
se_beta1 = sqrt(MS_Res/Sxx)

IC_beta1_inf = beta1-(qt((1-alpha/2),df=(length(x)-2))*(se_beta1))
IC_beta1_sup = beta1+(qt((1-alpha/2),df=(length(x)-2))*(se_beta1))
IC_beta1 = c(IC_beta1_inf,IC_beta1_sup)
IC_beta1
```

```
## [1] 2.514988 4.133754
```

Con una confianza del 95% el valor real del parámetro β_1 ***se encuentra entre*** $(-0.009614347, -0.004435854)$

e. Determinar un intervalo de confianza de 95% para el precio promedio de venta de una casa, para la cual los impuestos actuales son \$750.

Solución:

Nos piden calcular la respuesta media $E(y)$ de precio promedio de venta para una casa donde los impuestos actuales son $x_0 = 750$. Dado que x_0 se encuentra dentro de los valores de yardas con los que se planteo el modelo, podemos decir que un estimador insesgado de $E(y|x_0)$ es

$$\widehat{E(y \mid x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

y su varianza es

$$\begin{aligned}\text{Var}\left(\hat{\mu}_{y|x_0}\right) &= \text{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) = \text{Var}\left[\bar{y} + \hat{\beta}_1\left(x_0 - \bar{x}\right)\right] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2\left(x_0 - \bar{x}\right)^2}{S_{xx}} = \sigma^2\left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]\end{aligned}$$

Tambien vemos que como $E(y|x) = \hat{\mu}_{y|x_0}$ es es una combinación lineal de la y_i tenemos que

$$\hat{\mu}_{y|x_0} \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

Ahora para construir un estimador insesgado como $\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$

$$\frac{\hat{\mu}_{y|x_0} - E\left(y \mid x_0\right)}{\sqrt{MS_{\text{Res}}\left(1/n + \left(x_0 - \bar{x}\right)^2 / S_{xx}\right)}} \sim t_{(n-2)}$$

lo que nos deja que un intervalos de confianza de $100(1-\alpha)\%$ de la respuesta media en un punto $x = x_0$ es

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq E\left(y \mid x_0\right) \leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

entonces ***el intervalo de confianza del*** 95% ***de la respuesta media del valor de una casa la cual paga actualmente un total de*** $\$750$ ***en impuestos es***

```
x0 = 0.750 # Los precios reales estan divididos por 1000

muyx0 = beta0 + beta1*x0
se_muyx0 = sqrt( MS_Res*( (1/length(x)) + ((x0-meanx)^2)/Sxx ) )
IC_muyx0_inf = muyx0 - (qt((1-alpha/2),df=(length(x)-2)))*se_muyx0
IC_muyx0_sup = muyx0 + (qt((1-alpha/2),df=(length(x)-2)))*se_muyx0
IC_muyx0 = c(IC_muyx0_inf,IC_muyx0_sup )
IC_muyx0
```

```
## [1] 11.06792 20.55899
```

> *Nota: Tengamos en cuenta que la mínima observación con la cual se hizo el modelo es* $x_{(1)=3.891}$ *es decir que este modelo es apropiado para estimar valores promedio de venta de las casas siempre y cuando los valores* x_0 *se encuentren entre el mínimo y el máximo de la muestra.*

Ejercicio 2.8

Para los datos de la planta de oxígeno en el problema 2.7, suponer que la pureza y el porcentaje de hidrocarburos son variables aleatorias con distribución normal conjunta.

```
x<-c(1.02,  1.11,  1.43,  1.11,  1.01,  0.95,  1.11,  0.87,  1.43,  1.02,  1.46,  1.55,  1.55

y<-c(86.91, 89.85,  90.28,  86.34,  92.58,  87.33,  86.29,  91.86,  95.61,  89.86,  96.73,  99.42,  98.

meanx = mean(x)
meany = mean(y)

Sxy = sum(((x-meanx))*y)
Sxx = sum(((x-meanx)^2))
```

```
beta1 = Sxy/Sxx # Pendiente
beta0 = meany - beta1*meanx #intersepto

SS_T = sum((y^2))-(((sum(y))^2)/length(x)) # Suma de cuadrados de las observaciones corregidas
SS_Res = SS_T - (beta1*Sxy) # Suma de cuadrados de los residuales
SS_R = beta1*Sxy # Suma de cuadrados de la regresion o suma de cuadrados del modelo

df_T = length(x) -1 # Grados de libertad de las observaciones corregidas
df_Res = length(x) - 2 # Grados de libertad de la suma de cuadrados de los residuales
df_R = 1 # Grados de libertad de la suma de cuadrados de la regresión


MS_Res = SS_Res/df_Res
MS_R = SS_R/df_R
MS_Res = SS_Res/df_Res
```

a. ¿Cuál es la correlación entre la pureza del oxígeno y el porcentaje de hidrocarburos?

Solución:

Teniendo en cuenta que el estimador de la correlación dada una muestra que es

$$r = \frac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)}{\left[\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2\right]^{1/2}} = \frac{S_{xy}}{\left[S_{xx} S S_{\mathrm{T}}\right]^{1/2}}$$

tenemos que para el ejemplo anterior

```
r = Sxy/((Sxx*SS_T)^(1/2))
r
```

```
## [1] 0.6237968
```

b. Probar la hipótesis que $\rho = 0$.

Solución:

Teniendo en cuenta el estadístico de prueba

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)}$$

donde rechazamos la hipótesis de correlación nula si $|t_0| > t_{\alpha/2,n-2}$. De este modo

```
alpha = 0.05
t0 = r*(df_Res^(1/2))/((1-(r^2))^(1/2))
t0test = qt(1-alpha/2,df=df_Res)
pvalue = 2*pt(-abs(t0),df=df_Res)

t0
```

```
## [1] 3.386119
```

```
t0test
```

```
## [1] 2.100922
```

```
pvalue
```

```
## [1] 0.003291122
```

Dado que tenemos que $|t_0| = 3.386119 > t_{1-\alpha/2,n-2} = 2.100922$ ***rechazamos la hipótesis que nos dice que*** $\rho = 0$ ***con una significancia del*** 5% ***y un p valor de*** 0.003291122

c. Establecer un intervalo de confianza de 95% para $\rho = 0$.

Solución:

Para construir un intervalo de confianza veamos que

$$\tanh\left(\operatorname{arctanh} r - \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\operatorname{arctanh} r + \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right)$$

de este modo un intervalo de confianza del 95% para ρ

```
IC_rho_inf = tanh( atanh(r) - (qnorm( 1-alpha/2 )/((length( x )-3)^(1/2))) )
IC_rho_sup = tanh( atanh(r) + (qnorm( 1-alpha/2 )/((length( x )-3)^(1/2))) )
IC_rho = c(IC_rho_inf,IC_rho_sup)
IC_rho
```

```
## [1] 0.2503961 0.8356439
```

Ejercicio 2.10

A continuación se muestran el peso y la presión sistólica sanguínea de 26 hombres seleccionados al azar, en el grupo de edades de 25 a 30. Suponer que el peso y la presión sanguínea (BP) tienen distribución normal conjunta.

```
x<-c(130,   133,   150,   128,   151,   146,   150,   140,   148,   125,   133,   135,   150
y<-c(165,   167,   180,   155,   212,   175,   190,   210,   200,   149,   158,   169,   170
```

a. Determine una recta de regresión que relacione la presión sistólica sanguínea con el peso.

Solución:

Vamos a hacer uso del método de los mínimos cuadrados para ajustar un modelo de regresión lineal simple. Teniendo en cuenta los estimadores insesgados de los parámetros β_0 y β_1 del modelo poblacional $y = \beta_0 + \beta_1 x$ que son

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

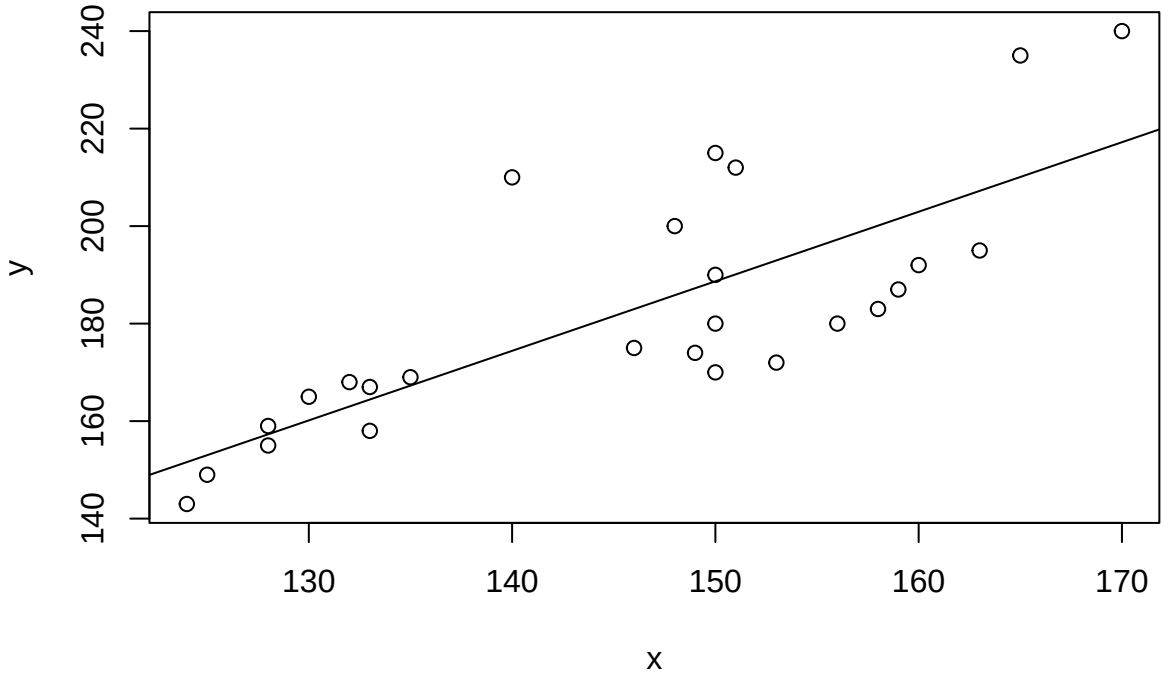
y

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

tenemos el modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Por lo tanto el modelo ajustado se calcula de la siguiente manera:

```
meanx = mean(x)
meany = mean(y)
Sxy = sum(((x-meanx))*y)
Sxx = sum(((x-meanx)^2))
beta1 = Sxy/Sxx
beta0 = meany - beta1*meanx

SS_T = sum((y^2))-(((sum(y))^2)/length(x)) # Suma de cuadrados de las observaciones corregidas
SS_Res = SS_T - (beta1*Sxy) # Suma de cuadrados de los residuales
SS_R = beta1*Sxy # Suma de cuadrados de la regresion o suma de cuadrados del modelo
```

```
df_T = length(x) -1 # Grados de libertad de las observaciones corregidas
df_Res = length(x) - 2 # Grados de libertad de la suma de cuadrados de los residuales
df_R = 1 # Grados de libertad de la suma de cuadrados de la regresión


MS_Res = SS_Res/df_Res
MS_R = SS_R/df_R
MS_Res = SS_Res/df_Res

plot(x,y)
abline(a=beta0,b=beta1) # Los parametros son el intercepto y la pendiente que calculamos
```



b. Estimar el coeficiente de correlación.

Solución:

Teniendo en cuenta que el estimador de la correlación dada una muestra que es

$$r = \frac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)}{\left[\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2\right]^{1/2}} = \frac{S_{xy}}{\left[S_{xx} SS_{\mathrm{T}}\right]^{1/2}}$$

tenemos que para ver la correlación entre el peso de una persona y la presión sistólica según la muestra suministrada es

```
r = Sxy/((Sxx*SS_T)^(1/2))
r
```

```
## [1] 0.7734903
```

c. Probar la hipótesis que $\rho = 0$.

Solución:

Teniendo en cuenta el estadístico de prueba

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)}$$

donde rechazamos la hipótesis de correlación nula si $|t_0| > t_{\alpha/2,n-2}$. De este modo

```
alpha = 0.05
t0 = r*(df_Res^(1/2))/((1-(r^2))^(1/2))
t0test = qt(1-alpha/2,df=df_Res)
pvalue = 2*pt(-abs(t0),df=df_Res)

t0
```

```
## [1] 5.978644
```

```
t0test
```

```
## [1] 2.063899
```

```
pvalue
```

```
## [1] 3.591105e-06
```

Dado que tenemos que $|t_0| = 5.978644 > t_{1-\alpha/2,n-2} = 2.063899$ ***rechazamos la hipótesis que nos dice que*** $\rho = 0$ ***con una significancia del*** 5% ***y un p valor de*** $3.591105e^{-06}$

d. Probar la hipótesis que $\rho = 0.6$.

Solución:

Para un test $H_0 : \rho = \rho_0$ contra $H_1 : \rho \neq \rho_0$ donde para $n \geq 25$ tenemos que la estadística

$$Z = \operatorname{arctanh} r = \frac{1}{2} \ln \frac{1+r}{1-r}$$

tiene aproximadamente distribución normal con media

$$\mu_Z = \operatorname{arctanh} \rho = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

y varianza

$$\sigma_Z^2 = (n-3)^{-1}$$

por lo tanto para verificar la hipótesis $H_0 : \rho = \rho_0$ tenemos que calcular la estadítica

$$Z_0 = (\operatorname{arctanh} r - \operatorname{arctanh} \rho_0)\,(n-3)^{1/2}$$

donde debemos rechazar $H_0 : \rho = \rho_0$ si $|Z_0| > Z_{1-\alpha/2}$. De este modo para ver si la correlación entre el peso de una persona y su presión sistólica es $H_0 : \rho = 0.6$ tenemos

```
p0 = 0.6
alpha = 0.05

Z0 = (atanh(r) - atanh(p0))*((length(x)-3)^(1/2))
Z0test = qnorm(1-alpha/2)
pvalue_Z0 = 2*pnorm(-abs(Z0))

Z0
```

```
## [1] 1.610495
```

```
Z0test
```

```
## [1] 1.959964
```

```
pvalue_Z0
```

```
## [1] 0.1072899
```

Dado que NO tenemos que $|Z_0| = 1.610495 > Z_{1-\alpha/2} = 1.959964$ ***fallamos al rechazar la hipótesis que nos dice que*** $\rho = 0.6$ ***con una significancia del*** 5% ***y un p valor de*** 0.1072899

e. Determinar un coeficiente de confianza de 95% para ρ.

Solución:

Para construir un intervalo de confianza veamos que

$$\tanh\left(\operatorname{arctanh} r - \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\operatorname{arctanh} r + \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right)$$

de este modo un intervalo de confianza del 95% para ρ

```
IC_rho_inf = tanh( atanh(r) - (qnorm( 1-alpha/2 )/((length( x )-3)^(1/2))) )
IC_rho_sup = tanh( atanh(r) + (qnorm( 1-alpha/2 )/((length( x )-3)^(1/2))) )
IC_rho = c(IC_rho_inf,IC_rho_sup)
IC_rho
```

```
## [1] 0.5513214 0.8932215
```

Ejercicio 2.12

Se cree que la cantidad de libras de vapor usadas en una planta por mes está relacionada con la temperatura ambiente promedio. A continuación se presentan los consumos y las temperaturas del último año.

```
y<-c(185.79,214.47,288.03,424.84,454.68,539.03,621.55,675.06,562.03,452.93,369.95,273.98)
x<-c(21,24,32,47,50,59,68,74,62,50,41,30)
x
```

```
##  [1] 21 24 32 47 50 59 68 74 62 50 41 30
```

```
y
```

```
##  [1] 185.79 214.47 288.03 424.84 454.68 539.03 621.55 675.06 562.03 452.93
## [11] 369.95 273.98
```

a. Ajustar un modelo de regresión lineal simple a los datos.

Solución:

Vamos a hacer uso del método de los mínimos cuadrados para ajustar un modelo de regresión lineal simple. Teniendo en cuenta los estimadores insesgados de los parámetros β_0 y β_1 del modelo $y = \beta_0 + \beta_1 x$ que son

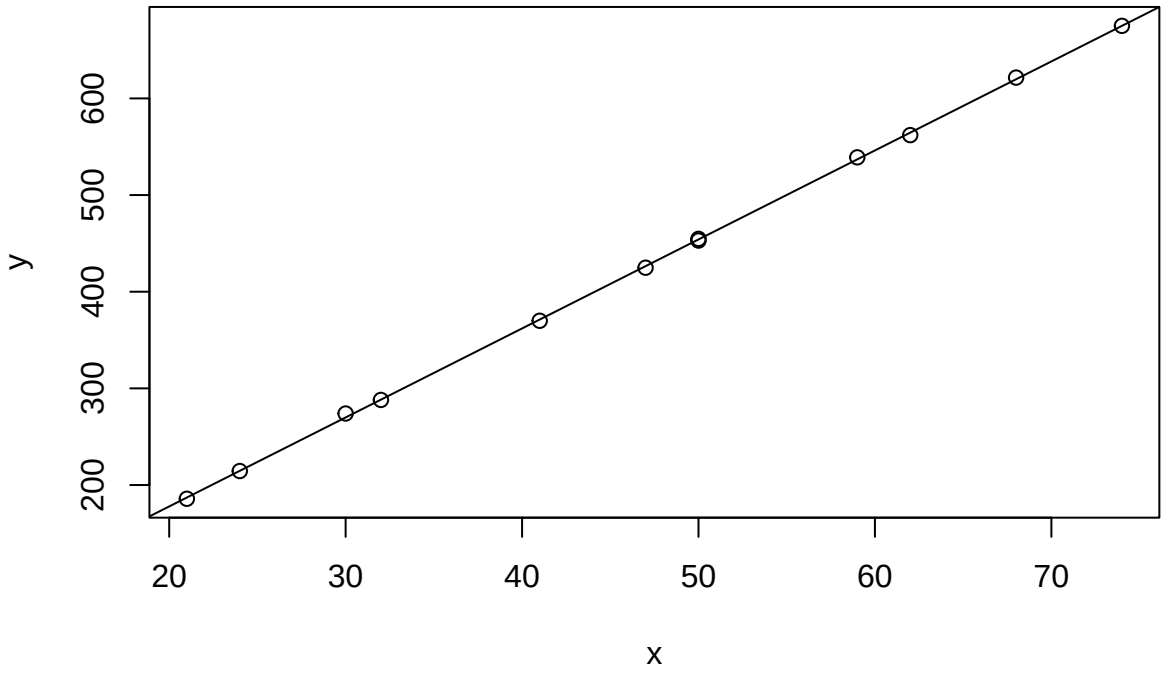
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

y

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

tenemos el modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Por lo tanto el modelo ajustado se calcula de la siguiente manera:

```
meanx = mean(x)
meany = mean(y)
Sxy = sum(((x-meanx))*y)
Sxx = sum(((x-meanx)^2))
beta1 = Sxy/Sxx
beta0 = meany - beta1*meanx

plot(x,y)
abline(a=beta0,b=beta1) # Los parametros son el intercepto y la pendiente que calculamos
```



b. Probar la significancia de la regresión.

Solución:

Vamos a desarrollar un análisis de varianza. Tengamos en cuenta la identidad fundamental del análisis de varianza para el modelo de regresión que nos dice

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$
$$SS_T=SS_R+SS_{\text{Res}}$$

donde al lado izquierdo tenemos la suma de los cuadrados de las observaciones corregidas (SS_T) y al derecho la suma de cuadrados del modelo (SS_R) mas la suma del cuadrado de los residuales (SS_{Res}).

También tenemos que los grados de libertad correspondientes vienen dados de la siguiente manera

$$df_{\text{T}}=df_{\text{R}}+df_{\text{Res}}$$
$$n-1=1+(n-2)$$

dado que para SS_T se pierde un grado de libertad al ajustar de la forma $\sum_{i=1}^{n}(y_i-\bar{y})$ en las desviaciones $(y_i-\bar{y})$, SS_T está determinado por un único parámetro ($\hat{\beta}_1$ en $SS_R=\hat{\beta}_1 S_{xy}$), y SS_{Res} tiene dos grados de libertad menos dado que al ajustar $\sum_{i=1}^{n}(y_i-\hat{y}_i)$ se pierden dos grados de libertad dado que las desviaciones $(y_i-\hat{y}_i)$ son resultado de estimar $\hat{\beta}_0$ y $\hat{\beta}_1$

De este modo al hacer el análisis de varianza con la hipótesis nula $H_0=\beta_1=0$ tenemos que el estadístico de prueba es

$$F_0=\frac{SS_R/df_R}{SS_{\text{Res}/df_{\text{Res}}}}=\frac{SS_R/1}{SS_{\text{Res}}/(n-2)}=\frac{MS_R}{MS_{\text{Res}}}=\frac{MS_R}{\hat{\sigma}^2}\sim F_{1,n-2}$$

dado que $SS_R=MS_R/\sigma^2\sim\chi^2_1$ y $SS_{\text{Res}}=MS_{\text{Res}}/\sigma^2\sim\chi^2_{n-2}$. Donde rechazaremos la hipótesis $H_0=\beta_1=0$ si $F_0>F_{1-\alpha,1,n-2}$

Tenemos que para un $\alpha=0.001$ el análisis de la varianza viene dado por el siguiente cálculo

```
alpha = 0.001

SS_T = sum((y^2))-(((sum(y))^2)/length(x)) # Suma de cuadrados de las observaciones corregidas
SS_Res = SS_T - (beta1*Sxy) # Suma de cuadrados de los residuales
SS_R = beta1*Sxy # Suma de cuadrados de la regresion o suma de cuadrados del modelo

df_T = length(x) -1 # Grados de libertad de las observaciones corregidas
df_Res = length(x) - 2 # Grados de libertad de la suma de cuadrados de los residuales
df_R = 1 # Grados de libertad de la suma de cuadrados de la regresión

MS_Res = SS_Res/df_Res
MS_R = SS_R/df_R

F0 = MS_R/MS_Res
F0test = qf(1-alpha,df1 = 1,df2 = (length(x)-2))
pvalue_F0 = 1-pf(F0,df1 = 1,df2 = (length(x)-2))

SS_R
```

```
## [1] 280589.6
```

```
SS_Res
```

```
## [1] 37.8547
```

```
SS_T
```

```
## [1] 280627.4
```

```
df_R
```

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

```
df_Res
```

```
## [1] 10
```

```
df_T
```

```
## [1] 11
```

```
MS_R
```

```
## [1] 280589.6
```

```
MS_Res
```

```
## [1] 3.78547
```

```
F0
```

```
## [1] 74122.78
```

```
F0test
```

```
## [1] 21.0396
```

```
pvalue_F0
```

```
## [1] 0
```

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_(0)
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_{Res}
Residual	$SS_{\text{Res}} = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_{Res}	
Total	SS_T	$n-1$		

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P value
Regression	280589.6	1	280589.6	74122.78	0
Residual	37.8547	10	3.78547		
Total	280627.4	11			

```
xylm = lm(y ~ x)
anova(xylm)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 280590  280590   74123 < 2.2e-16 ***
## Residuals 10     38       4
```

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

Dado que tenemos que $F_0 = 74122.78 > F_{1-\alpha,1,n-2} = 21.0396$ ***rechazamos la hipótesis que nos dice que*** $\beta_1 = 0$ ***con una significancia del*** 0.1% ***y un p valor de*** 0

c. En la administración de la planta se cree que un aumento de 1 grado en la temperatura ambiente promedio hace aumentar 10 000 libras el consumo mensual de vapor. ¿Estos datos respaldan la afirmación?

Solución:

Vamos a hacer una verificación de hipótesis donde $H_0 : \beta_1 = 10$ y $H_1 : \beta_1 \neq 10$ dado que la variable respuesta está sobre 1000. Ahora, teniendo en cuenta que, en este caso, no conocemos la varianza poblacional; entonces para el test de hipótesis $H_0 : \beta_1 = \beta_{10} = 10$ tenemos el estadístico de prueba

$$t_0 = \frac{\hat{\beta}_1 - \beta_{01}}{\sqrt{MS_{\text{Res}}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{01}}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 10}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

dado que $MS_{\text{Res}} = \hat{\sigma}^2$ es un estimador insesgado de σ^2, $(n-2)MS_{\text{Res}}/\sigma^2 \sim \chi^2_{n-2}$, y MS_{Res} con $\hat{\beta}_1$ son independientes. Donde rechazamos la hipótesis, que nos dice que tenemos una pendiente de 10, si $|t_0| > t_{1-\alpha/2,n-2}$.

De este modo tenemos que para un $\alpha = 0.05$ el test sobre la pendiente viene dado por el siguiente cálculo

```
alpha = 0.05
se_beta1 = sqrt(MS_Res/Sxx)
t0 = (beta1-10)/se_beta1 # el estadístico de prueba preguntando si \hat{\beta}_1 = \beta_{10} = 0
t0test = qt((1 - alpha/2),df=(length(x)-2))
pvalue_t0 = 2*pt(-abs(t0),df=(length(x)-2))

t0
```

```
## [1] -23.40222
```

```
t0test
```

```
## [1] 2.228139
```

```
pvalue_t0
```

```
## [1] 4.597358e-10
```

Dado que tenemos que $|t_0| = 23.40222 > t_{1-\alpha/2,n-2} = 2.228139$ ***rechazamos la hipótesis que nos dice que*** $\beta_1 = 10$ ***con una significancia del*** 5% ***y un p valor de*** $4.597358e^{-10}$

d. Determinar un intervalo de predicción de 99% para el uso de vapor en un mes con temperatura ambiente promedio de 58°.

Solución:

Para construir un intervalo de predicción tengamos en cuenta la variable aleatoria

$$\psi = y_0 - \hat{y}_0$$

donde $y_0 = \beta_0 + \beta_1 x_0$ y $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Como son combinaciones lineales de distribuciones normales vemos que ψ tiene distribución normal de media cero y varianza

$$\text{Var}(\psi) = \text{Var}\left(y_0 - \hat{y}_0\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

dado que y_0 y $\hat{y}_0$ son independientes. De este modo tenemos que la desviación estándar estimada es la estadística apropiada y así el intervalo de predicción es

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

entonces tenemos que

```
alpha = 0.01
x0=58

y0 = beta0 + beta1*x0
se_y0 = sqrt( MS_Res*(1 + (1/length(x)) + ((x0-meanx)^2)/Sxx ) ) # desviación de \hat{y}_0

IC_y0_inf = y0 - qt((1-alpha/2),df=(length(x)-2))*se_y0
IC_y0_sup = y0 + qt((1-alpha/2),df=(length(x)-2))*se_y0
IC_y0 = c(IC_y0_inf,IC_y0_sup)
IC_y0
```

```
## [1] 521.2237 534.2944
```

Ejercicio 2.14

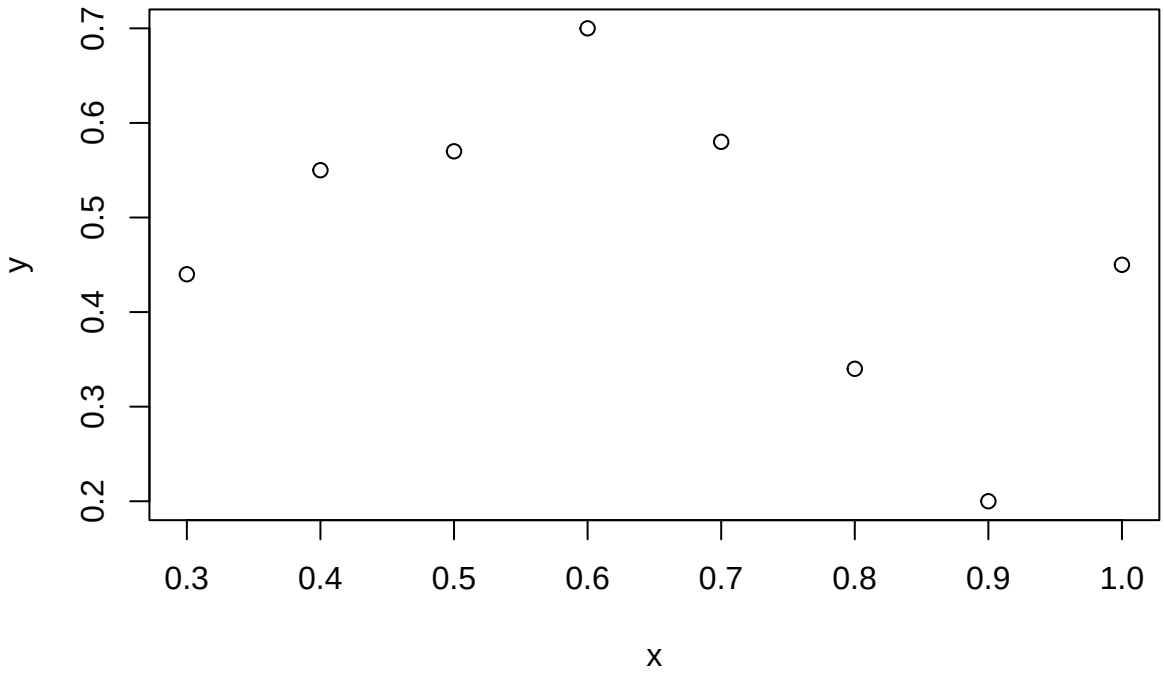
Hsuie, Ma y Tsai ("Separación y caracterización de copoliésteres termotrópicos del ácido p-hidroxibenzoico, ácido sebácico e hidroquinona", Journal of Applied Polymer Science, 56, 471-476, 1995) estudian el efecto de la relación molar del ácido sebácico (el regresor) sobre la viscosidad intrínseca de los copoliésteres (la respuesta). La siguiente tabla muestra los datos.

```
x = c(1.0,0.9,0.8,0.7,0.6,0.5,0.4,0.3)
y = c(0.45, 0.20, 0.34, 0.58, 0.70, 0.57, 0.55, 0.44)
```

a. Trazar un diagrama de dispersión de los datos.

Solución: Para trazar el diagrama de disperción de los datos vamos a usar la función `plot()` de R

```
plot(x = x,y = y)
```



b. Estimar la ecuación de predicción.

Solución:

Vamos a hacer uso del método de los mínimos cuadrados para ajustar un modelo de regresión lineal simple. Teniendo en cuenta los estimadores insesgados de los parámetros β_0 y β_1 del modelo $y = \beta_0 + \beta_1 x$ que son

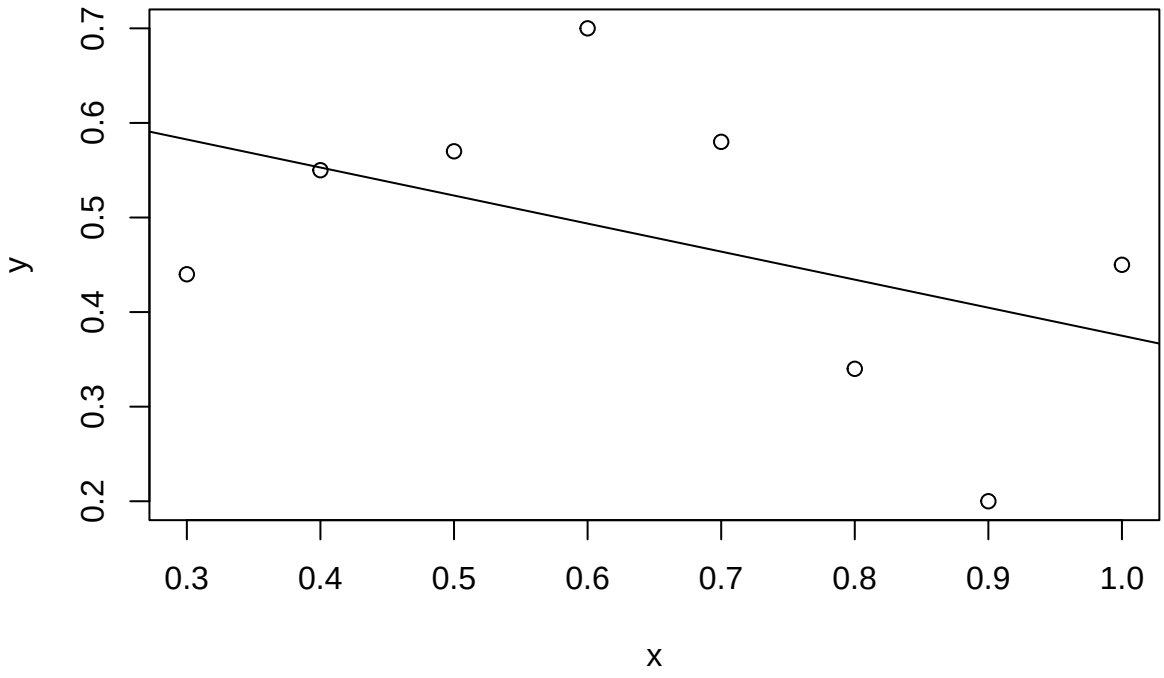
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

y

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

tenemos el modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Por lo tanto el modelo ajustado se calcula de la siguiente manera:

```
meanx = mean(x)
meany = mean(y)
Sxy = sum(((x-meanx))*y)
Sxx = sum(((x-meanx)^2))
beta1 = Sxy/Sxx
beta0 = meany - beta1*meanx

plot(x,y)
abline(a=beta0,b=beta1) # Los parametros son el intercepto y la pendiente que calculamos
```



c. Hacer un análisis completo y adecuado (pruebas estadísticas, cálculo de R2,etcétera).

Solución:

- *Significancia*

En un principio vamos a hacer una prueba de significancia de la regresión. Para la significancia de la regresión tengamos en cuenta que, en este caso, no conocemos la varianza poblacional; entonces para el test de hipótesis $H_0 = \beta_1 = \beta_{10} = 0$ tenemos el estadístico de prueba

$$t_0 = \frac{\hat{\beta}_1 - \beta_{01}}{\sqrt{MS_{\text{Res}}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{01}}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

dado que $MS_{\text{Res}} = \hat{\sigma}^2$ es un estimador insesgado de σ^2, $(n-2)MS_{\text{Res}}/\sigma^2 \sim \chi^2_{n-2}$, y MS_{Res} con $\hat{\beta}_1$ son independientes. Donde rechazamos la hipótesis, que nos dice que tenemos una pendiente nula en este caso, si $|t_0| > t_{1-\alpha/2,n-2}$.

De este modo tenemos que para un $\alpha = 0.05$ el test de significancia de la regresión viene dado por el siguiente cálculo

```
alpha = 0.05

SS_T = sum((y^2))-(((sum(y))^2)/length(x)) # Suma de cuadrados de las observaciones corregidas
SS_Res = SS_T - (beta1*Sxy) # Suma de cuadrados de los residuales
SS_R = beta1*Sxy # Suma de cuadrados de la regresion o suma de cuadrados del modelo

df_T = length(x) -1 # Grados de libertad de las observaciones corregidas
df_Res = length(x) - 2 # Grados de libertad de la suma de cuadrados de los residuales
```

```
df_R = 1 # Grados de libertad de la suma de cuadrados de la regresión

MS_Res = SS_Res/df_Res
MS_R = SS_R/df_R

se_beta1 = sqrt(MS_Res/Sxx)
t0 = (beta1-0)/se_beta1 # el estadístico de prueba preguntando si \hat{\beta}_1 = \beta_{10} = 0
t0test = qt((1 - alpha/2),df=(length(x)-2))
pvalue_t0 = 2*pt(-abs(t0),df=(length(x)-2))

t0
```

```
## [1] -1.280803
```

```
t0test
```

```
## [1] 2.446912
```

```
pvalue_t0
```

```
## [1] 0.2475409
```

Dado que NO tenemos que $|t_0| = 1.280803 > t_{1-\alpha/2,n-2} = 2.446912$ ***fallaríamos al rechazar la hipótesis que nos dice que*** $\beta_1 = 0$ ***con una significancia del*** 5% ***y un p valor de*** 0.2475409

- *Varianza explicada por el modelo*

Dado que SS_T es la medida de la variabilidad en y sin considerar el efecto de la variable regresora x y SS_{Res} es la medida de la variabilidad sobrante después de considerar la variable regresora x tenemos que el coeficiente de determinación $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{\text{Res}}}{SS_T}$ es considerado también la proporción de la variación explicada por el regresor x. Por lo que tenemos

```
R2 = SS_R/SS_T
R2
```

```
## [1] 0.2147065
```

Es decir que el modelo explica un 21.47% ***de la variabilidad total en la viscosidad.***

- *Asociación lineal*

Teniendo en cuenta que el estimador de la correlación dada una muestra que es

$$r = \frac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)}{\left[\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2\right]^{1/2}} = \frac{S_{xy}}{\left[S_{xx} SS_{\text{T}}\right]^{1/2}}$$

tenemos que

```
r = Sxy/((Sxx*SS_T)^(1/2))
r
```

```
## [1] -0.4633643
```

Ahora para verificar que la hipótesis que $\rho = 0$, teniendo en cuenta el estadístico de prueba

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)}$$

donde rechazamos la hipótesis de correlación nula si $|t_0| > t_{\alpha/2,n-2}$. De este modo

```r
alpha = 0.05
t0 = r*(df_Res^(1/2))/((1-(r^2))^(1/2))
t0test = qt(1-alpha/2,df=df_Res)
pvalue = 2*pt(-abs(t0),df=df_Res)

t0
```

```
## [1] -1.280803
```

```r
t0test
```

```
## [1] 2.446912
```

```r
pvalue
```

```
## [1] 0.2475409
```

Dado que NO tenemos que $|t_0| = 1.280803 > t_{1-\alpha/2,n-2} = 2.446912$ ***fallariamos al rechazar la hipótesis que nos dice que*** $\rho = 0$ ***con una significancia del*** 5% ***y un p valor de*** 0.2475409

Ejercicio 2.16

Carroll y Spiegelman ("The Effects of Ignoring Small Measurement Errors in Precision Instrument Calibration", Journal of Quality Technology, 18, 170-173, 1986) examinan la relación entre la presión en un tanque y el volumen de líquido. La siguiente tabla muestra los datos. Use un paquete adecuado de programas estadísticos para efectuar un análisis de ellos. Comente los resultados obtenidos con la rutina del programa.

Solución: Vamos a dejar el volumen como la respuesta a la presión. Primero cargamos los datos en R

```r
y = c(2084, 2084,  2273,  2273,  2273,  2463,  2463,  2651,  2652,  2652,  2842,  2842,  303
x = c(4599, 4600,  5044,  5043,  5044,  5488,  5487,  5931,  5932,  5932,  6380,  6380,  681
```

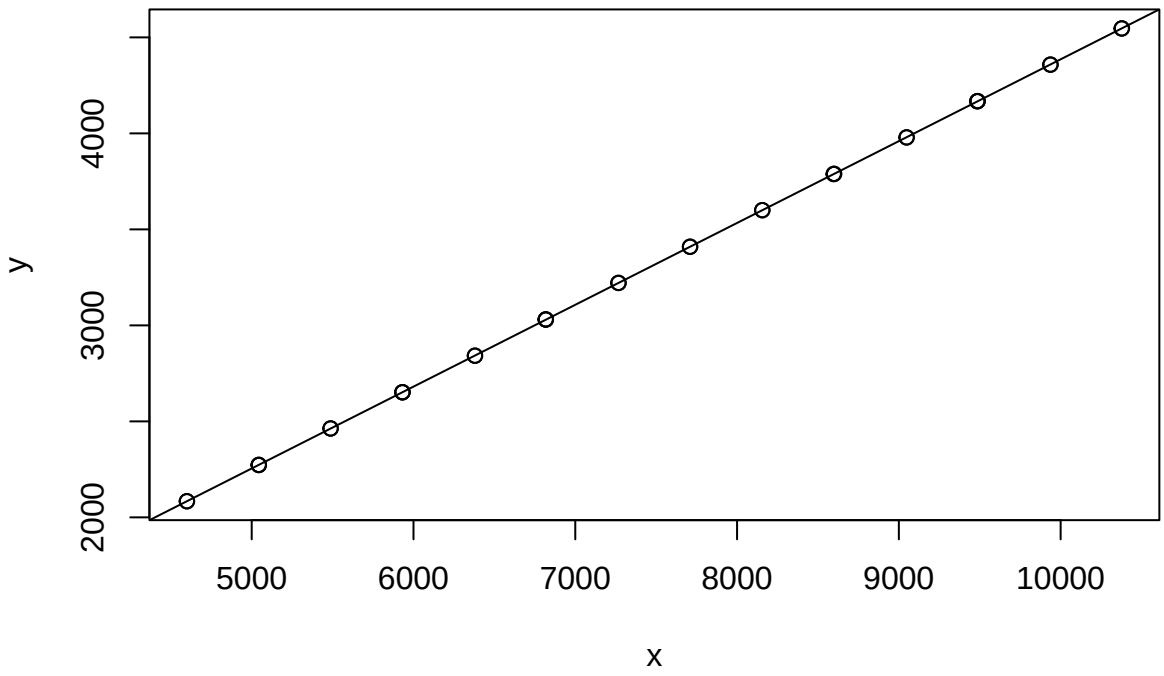
Ahora vamos a usar la función en R de `lm()` la cual nos crea un objeto de modelo lineal

```r
xylm = lm(y ~x)
xylm
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##    123.8780       0.4262
```

donde podemos ver que la estimación via mínimos cuadrados nos estima una $\hat{\beta}_0 = 123.8780$ y $\hat{\beta}_1 = 0.4262$. De este modo nuestro modelo queda $\hat{y} = 123.8780 + 0.4262x$ y lo podemos ver gráficamente con el comando `plot()`.

```r
plot(x,y)
abline(xylm)
```



- *Significancia de la regresión*

Podemos ver la significancia de la regresión haciendo uso de la función `summary()` en R la cual describe con más detalle el objeto generado por la función `lm()` o tambien podemos hacer un análisis de varianza con la función `anova()`

```
summary(xylm)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30006 -0.44769 -0.00476  0.36105  1.82191
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.239e+02  5.498e-01   225.3   <2e-16 ***
## x           4.262e-01  7.187e-05  5930.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7327 on 31 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.517e+07 on 1 and 31 DF,  p-value: < 2.2e-16
```

```r
anova(xylm)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df   Sum Sq  Mean Sq  F value    Pr(>F)
## x          1 18879734 18879734 35165232 < 2.2e-16 ***
## Residuals 31       17        1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
F0test = qf(1-0.001,df1 = 1,df2 = (length(x)-2))
F0test
```

```
## [1] 13.20201
```

Es decir que dado el análisis de varianza dado por la función tenemos que $F_0 = 35165232 > F_{1-\alpha,1,n-2} = 13.20201$ ***rechazamos la hipótesis que nos dice que*** $\beta_1 = 0$ ***con una significancia del*** 0.1% ***y un p valor de*** $2.2e^{-16}$

- *Intervalo de confianza del intercepto y la pendiente*

Con el la función `confint()` de r podemos generar un intervalo de confianza para los parámetros del modelo lineal. De este modo para una confianza fija del 99% para ambos parámetros tenemos los intervalos siguientes

```r
confint(xylm,level = 0.99)
```

```
##                   0.5 %      99.5 %
## (Intercept) 122.3693869 125.3866691
## x             0.4259873   0.4263818
```

- *Predicción de futuras observaciones*

```r
predict(xylm,interval = "prediction",level = 0.99)
```

```
## Warning in predict.lm(xylm, interval = "prediction", level = 0.99): predictions on current data refe
```

```
##         fit      lwr      upr
## 1  2083.901 2081.784 2086.017
## 2  2084.327 2082.211 2086.443
## 3  2273.553 2271.458 2275.648
## 4  2273.127 2271.032 2275.222
## 5  2273.553 2271.458 2275.648
## 6  2462.779 2460.702 2464.856
## 7  2462.353 2460.276 2464.430
## 8  2651.579 2649.516 2653.641
## 9  2652.005 2649.942 2654.067
## 10 2652.005 2649.942 2654.067
## 11 2842.935 2840.884 2844.987
## 12 2842.935 2840.884 2844.987
## 13 3029.604 3027.560 3031.649
## 14 3029.178 3027.134 3031.223
## 15 3029.604 3027.560 3031.649
## 16 3220.535 3218.494 3222.576
## 17 3221.387 3219.346 3223.428
## 18 3409.335 3407.293 3411.376
## 19 3409.761 3407.719 3411.802
## 20 3599.839 3597.793 3601.885
```

```
## 21 3599.839 3597.793 3601.885
## 22 3787.787 3785.733 3789.840
## 23 3788.639 3786.585 3790.693
## 24 3789.065 3787.012 3791.119
## 25 3979.996 3977.930 3982.061
## 26 3979.996 3977.930 3982.061
## 27 4165.812 4163.732 4167.893
## 28 4167.091 4165.010 4169.171
## 29 4167.091 4165.010 4169.171
## 30 4358.448 4356.348 4360.547
## 31 4359.300 4357.201 4361.399
## 32 4546.395 4544.274 4548.516
## 33 4547.247 4545.126 4549.369
```

- *Correlación muestral*

```
cor(x,y)
```

```
## [1] 0.9999996
```

Ejercicio 2.17

Para el modelo de regresión lineal simple $y = 50 + 10x + \varepsilon$, donde ε tiene $NID(0, 16)$, suponer que se usan $n = 20$ pares de observaciones para ajustar este modelo. Generar 500 muestras de 20 observaciones, tomando una observación para cada valor de $x = 1, 1.5, 2, 2.5..., 10$ para cada muestra.

Solución: Vamos a crear una tabla de 20 filas por 500 columnas donde vamos a guardar las observaciones generadas por el modelo

```
x = seq(0.5,10,by = 0.5)
Y = data.frame(x)

length(x) # Se deja desde 0.5 para poder tener 20 muestras pareadas diferentes
```

```
## [1] 20
```

Ahora vamos a iterar 500 veces para conseguir las observaciones asociadas al modelo. Para generar el valor del error usamos la función en R `rnorm()`

```
k = 500

for (i in 1:k) {
  e = rnorm(20,mean = 0,sd = sqrt(16))
  y = 50 + 10*x + e
  Y[,i+1] = y
  names(Y)[i+1] = i
}

dim(Y)
```

```
## [1]  20 501
```

a. Para cada muestra, calcular los estimados de la pendiente y la ordenada al origen por mínimos cuadrado. Trazar histogramas de los valores muéstrales de β_0 y β_1. Comentar la forma de esos histogramas.

Solución: Vamos primero a calcular los valores de la pendiente y la ordenada via mínimos cuadrados

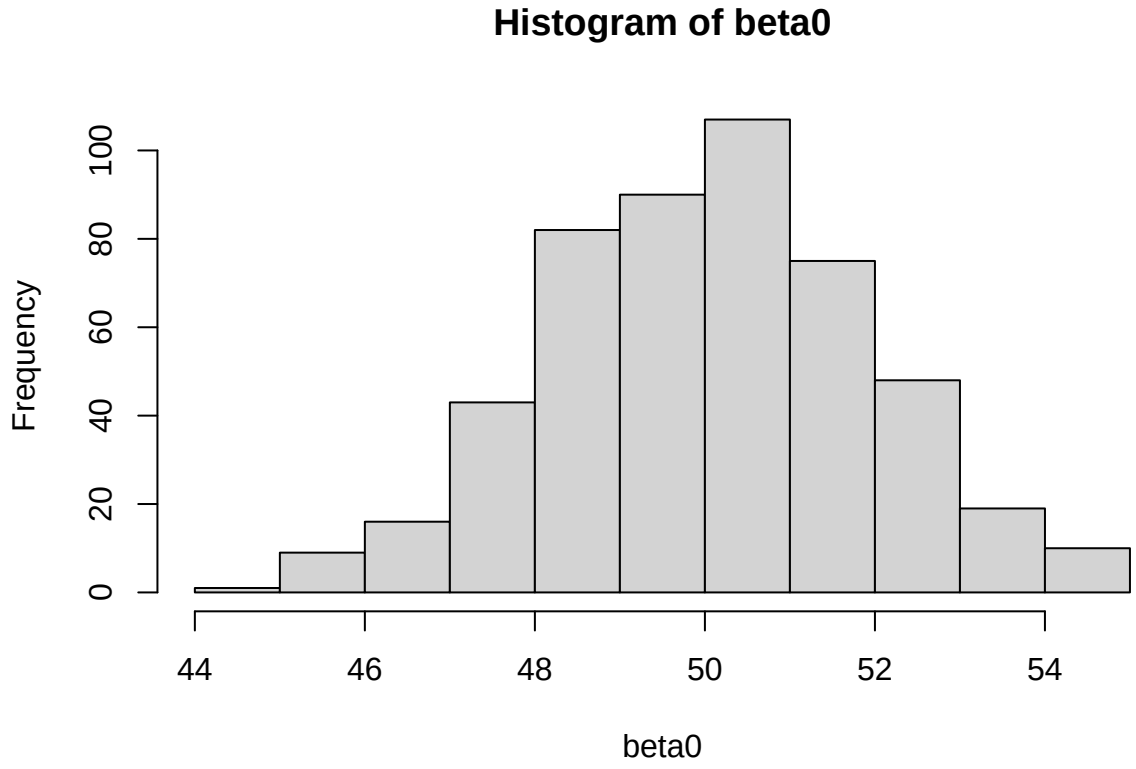
```
B = c("Beta_0","Beta_1")
xYlm = data.frame(B)
```

```r
for (i in 1:k) {
  lmss = lm(Y[,i+1] ~ Y[,1])
  xYlm[,i+1] = lmss$coefficients
}

dim(xYlm)
```

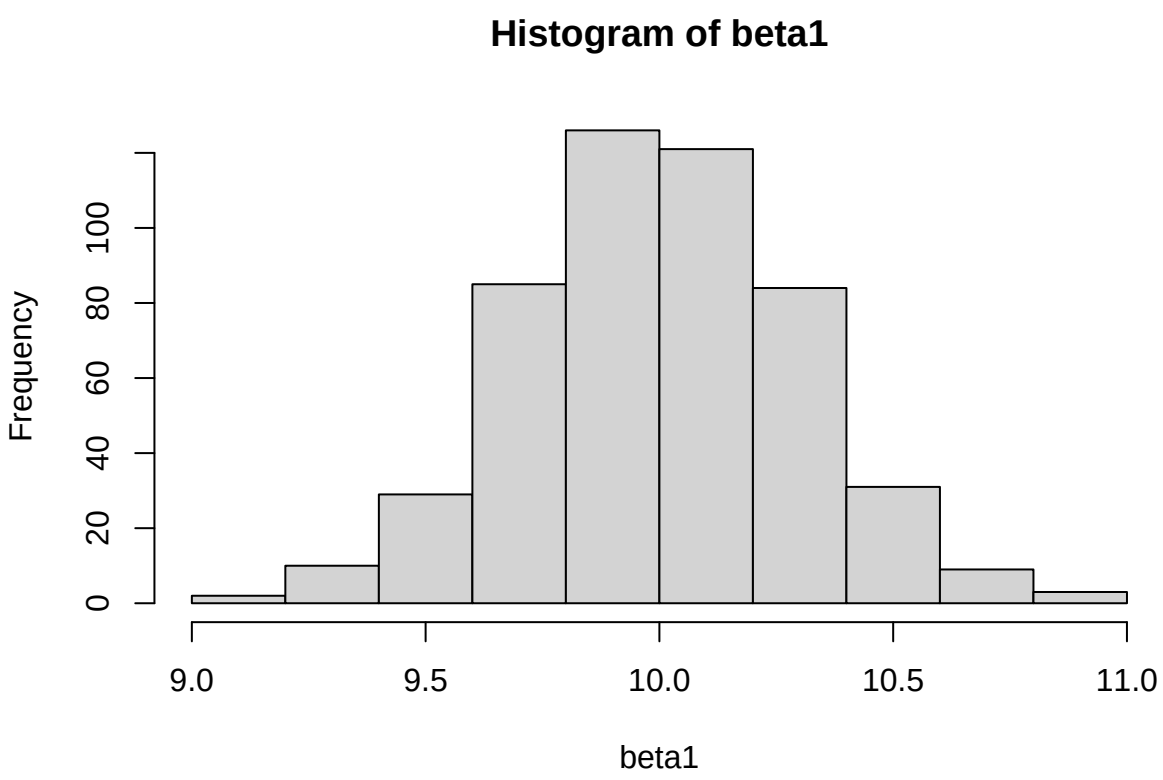
```
## [1]   2 501
```

ahora generamos con la función `hist()` de R los histogramas correspondientes a los parámetros

```r
beta0 = as.numeric(xYlm[1,2:501])
beta1 = as.numeric(xYlm[2,2:501])
hist(x=beta0)
```




```r
hist(x=beta1)
```



Como deberíamos esperar el los valores mas recurrentes de las estimaciones de los parámetros están entre el valor real de los parámetros dado que los estimadores son insesgados. La varianza de los estimadores son $\text{Var}\left(\hat{\beta}_0\right) = \text{Var}(\bar{y}) + \bar{x}^2 \text{Var}\left(\hat{\beta}_1\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$ ***y*** $\text{Var}\left(\hat{\beta}_1\right) = \sigma^2 \sum_{i=1}^n c_i^2 = \frac{\sigma^2 \sum_{i=1}^n (x_i - \bar{x})^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$***. Además sabemos que tienen distribución normal*** $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$ ***y*** $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$

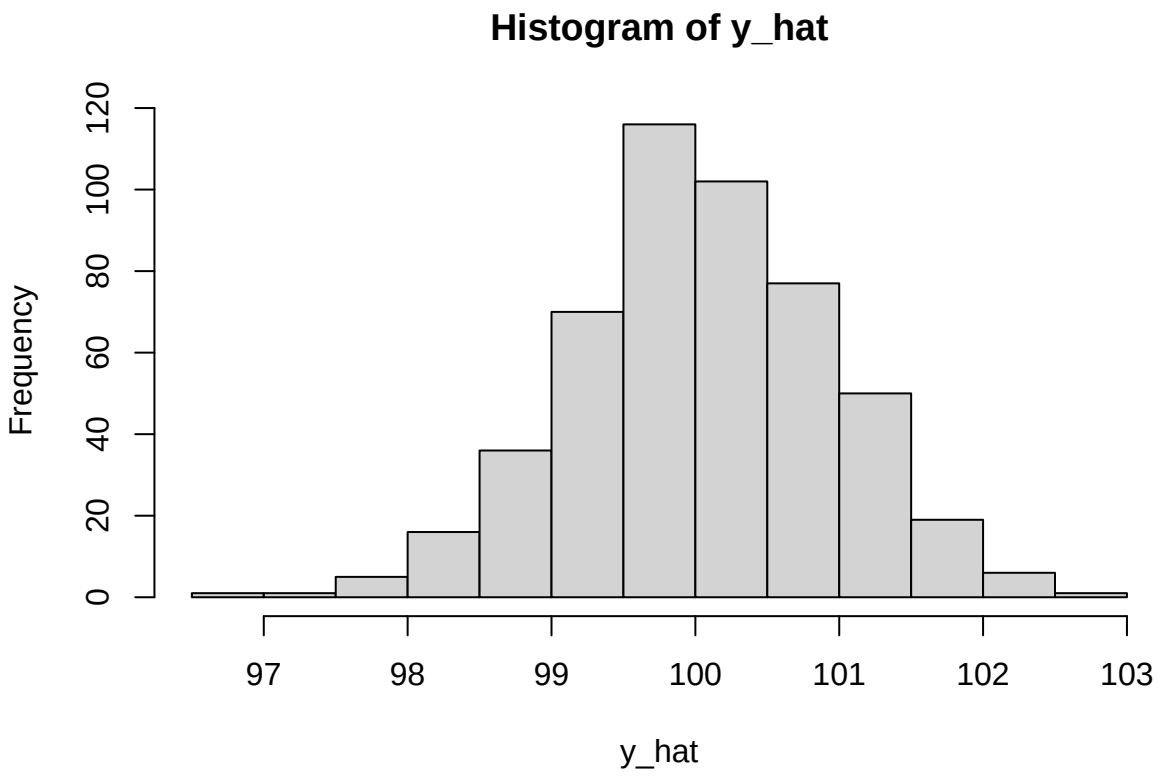
b. Para cada muestra, calcular un estimado de $E(y|x = 5)$. Trazar un histograma de los estimados obtenidos. Comentar la forma del histograma.

Solución:

```
x0 = 5
xY_hat = data.frame(x0)

for (i in 1:k) {
  xY_hat[,i+1] = xYlm[1,i+1] + xYlm[2,i+1]*x0
}
dim(xY_hat)
```

```
## [1]   1 501
```

```
y_hat = as.numeric(xY_hat[1,2:501])
hist(y_hat)
```



En este caso tambien tenemos que las estimaciones puntuales se concentran más alrrededor del valor real. Tambien la distribución satisface la normalidad esperada de las demsotraciones.

c. Determinar un intervalo de confianza de 95% para la pendiente en cada muestra. ¿Cuántos de los intervalos contienen el valor verdadero $\beta_1 = 10$? ¿Es lo que se esperaba?

Solución:

Teniendo en cuenta que

$$t_0 = \frac{\hat{\beta}_1 - \beta_{01}}{\sqrt{MS_{\text{Res}}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{01}}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

por lo tanto un intervalo de confianza de $100(1-\alpha)\%$ de la pendiente β_1 está dado por

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\,\text{se}\left(\hat{\beta}_1\right) \leq \hat{\beta}_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\,\text{se}\left(\hat{\beta}_1\right)$$

de este modo tenemos que los intervalos de confianza de $100(1-\alpha)\% = 95\%$ de las pendientes dadas las observaciones generadas son

```
alpha = 0.05

IC_beta = c("IC_beta1_inf","IC_beta1_sup")
IC_B = data.frame(IC_beta)

for (i in 1:k) {
  meanx = mean(x)
```

```
  meany = mean(Y[,i+1])
  Sxy = sum(((x-meanx))*Y[,i+1])
  Sxx = sum(((x-meanx)^2))

  SS_T = sum((Y[,i+1]^2))-(((sum(Y[,i+1]))^2)/length(x))
  SS_Res = SS_T - (xYlm[2,i+1]*Sxy)
  SS_R = xYlm[2,i+1]*Sxy

  df_T = length(x) -1
  df_Res = length(x) - 2
  df_R = 1

  MS_Res = SS_Res/df_Res
  MS_R = SS_R/df_R
  se_beta1 = sqrt(MS_Res/Sxx)

  IC_B[1,i+1] = xYlm[2,i+1]-(qt((1-alpha/2),df=(length(x)-2))*(se_beta1))
  IC_B[2,i+1] = xYlm[2,i+1]+(qt((1-alpha/2),df=(length(x)-2))*(se_beta1))
}

dim(IC_B)
```

```
## [1]   2 501
```

Ahora vamos a sumar el número intervalos que contienen el valor del parámetro y lo dividimos por el número de simulaciones que hicimos.

```
sum(IC_B[1,2:501] < 10 & IC_B[2,2:501] > 10)/k
```

```
## [1] 0.942
```

que es lo que esperabamos.

d. Para cada estimado de $E(y|x = 5)$ en la parte b, calcular el intervalo de confianza de 95%. ¿Cuántos de esos intervalos contienen el valor verdadero de $E(y|x = 5) = 100$? ¿Es lo que se esperaba?

Solución:

Nos piden calcular la respuesta media $E(y)$ de las simulaciones para un número $x_0 = 100$. Dado que x_0 se encuentra dentro de los valores con los que se planteo el modelo, podemos decir que un estimador insesgado de $E(y|x_0)$ es

$$\widehat{E\left(y \mid x_0\right)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

y su varianza es

$$\begin{aligned}\operatorname{Var}\left(\hat{\mu}_{y|x_0}\right) &= \operatorname{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) = \operatorname{Var}\left[\bar{y} + \hat{\beta}_1\left(x_0 - \bar{x}\right)\right] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2\left(x_0 - \bar{x}\right)^2}{S_{xx}} = \sigma^2\left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]\end{aligned}$$

Tambien vemos que como $E(y|x) = \hat{\mu}_{y|x_0}$ es es una combinación lineal de la y_i tenemos que

$$\hat{\mu}_{y|x_0} \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]\right)$$

Ahora para construir un estimador insesgado como $\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$

$$\frac{\hat{\mu}_{y|x_0} - E\left(y \mid x_0\right)}{\sqrt{MS_{\text{Res}}\left(1/n + \left(x_0 - \bar{x}\right)^2 / S_{xx}\right)}} \sim t_{(n-2)}$$

lo que nos deja que un intervalos de confianza de $100(1-\alpha)\%$ de la respuesta media en un punto $x = x_0$ es

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq E\left(y \mid x_0\right) \leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

entonces ***los intervalos de confianza del*** 95% ***de la respuesta media con*** $x_0 = 100$ ***son***

```
alpha = 0.05
x0 = 5

IC_muyx0 = c("IC_muyx0_inf","IC_muyx0_sup")
IC_muyx0 = data.frame(IC_muyx0)

for (i in 1:k) {
  meanx = mean(x)
  meany = mean(Y[,i+1])
  Sxy = sum(((x-meanx))*Y[,i+1])
  Sxx = sum(((x-meanx)^2))

  SS_T = sum((Y[,i+1]^2))-(((sum(Y[,i+1]))^2)/length(x))
  SS_Res = SS_T - (xYlm[2,i+1]*Sxy)
  SS_R = xYlm[2,i+1]*Sxy

  df_T = length(x) -1
  df_Res = length(x) - 2
  df_R = 1

  MS_Res = SS_Res/df_Res
  MS_R = SS_R/df_R
  muyx0 =xYlm[1,i+1] + xYlm[2,i+1]*x0
  se_muyx0 = sqrt( MS_Res*( (1/length(x)) + ((x0-meanx)^2)/Sxx ) )

  IC_muyx0[1,i+1] = muyx0 - (qt((1-alpha/2),df=(length(x)-2)))*se_muyx0
  IC_muyx0[2,i+1] = muyx0 + (qt((1-alpha/2),df=(length(x)-2)))*se_muyx0
}

dim(IC_muyx0)
```

```
## [1]   2 501
```

Ahora vamos a sumar el número intervalos que contienen el valor esperado y lo dividimos por el número de simulaciones que hicimos.

```
sum(IC_muyx0[1,2:501] < 100 & IC_muyx0[2,2:501] > 100)/k
```

```
## [1] 0.944
```

que es lo que esperábamos, dado que tenemos que tener en cuenta la aleatoriedad de la generación de los datos. Si hacemos más iteraciones entonces los valores que están contenidos son cada vez mas a la confianza fijada.

Ejercicio 2.18

Repetir el problema 2.17 usando sólo 10 observaciones para cada muestra y tomando una observación de cada nivel $x = 1, 2, 3, ..., 10$. ¿Qué impacto tiene usar $n = 10$ sobre las respuestas en el problema 2.17? Comparar las longitudes de los intervalos de confianza y el aspecto de los histogramas.

Solución: Vamos a ejecutar el mismo código del punto anterior.

Primero vamos crear una tabla de 10 filas por 500 columnas donde vamos a guardar las observaciones generadas por el modelo

```
x = seq(1,10,by = 1)
Y = data.frame(x)

length(x)
```

```
## [1] 10
```

Ahora vamos a iterar 500 veces para conseguir las observaciones asociadas al modelo. Para generar el valor del error usamos la función en R `rnorm()`

```
k = 500

for (i in 1:k) {
  e = rnorm(10,mean = 0,sd = sqrt(16))
  y = 50 + 10*x + e
  Y[,i+1] = y
  names(Y)[i+1] = i
}

dim(Y)
```

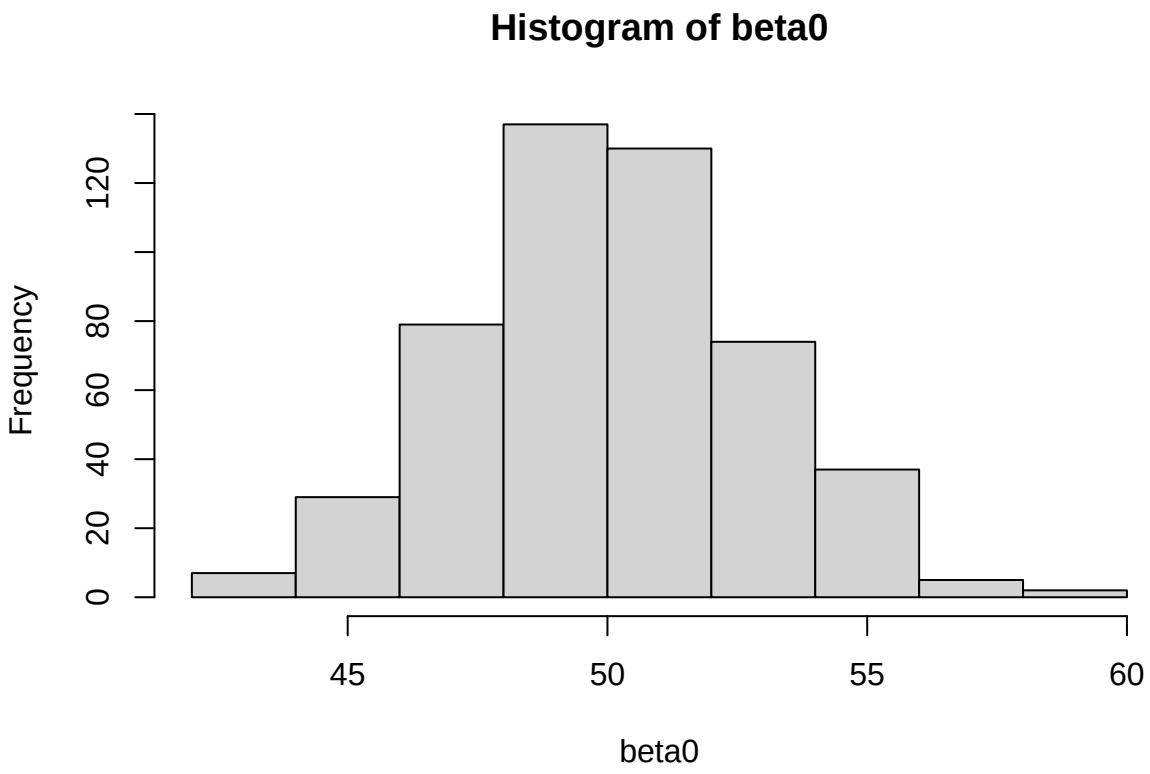
```
## [1]  10 501
```

- *Pendiente e intercepto*

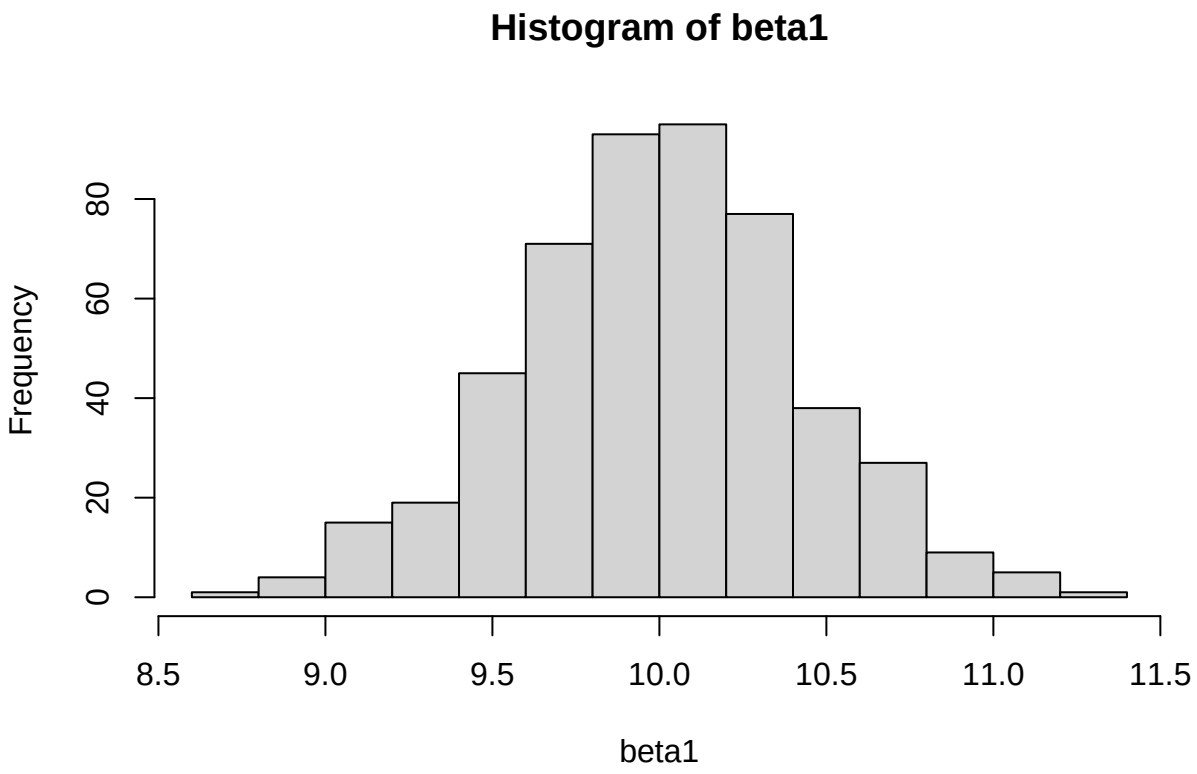
```
B = c("Beta_0","Beta_1")
xYlm = data.frame(B)

for (i in 1:k) {
  lmss = lm(Y[,i+1] ~ Y[,1])
  xYlm[,i+1] = lmss$coefficients
}

dim(xYlm)
```

```
## [1]   2 501
```

```
beta0 = as.numeric(xYlm[1,2:501])
beta1 = as.numeric(xYlm[2,2:501])
hist(x=beta0)
```


Histogram of beta0
Frequency
0
20
40
60
80
120
45
50
55
60
beta0

```
hist(x=beta1)
```



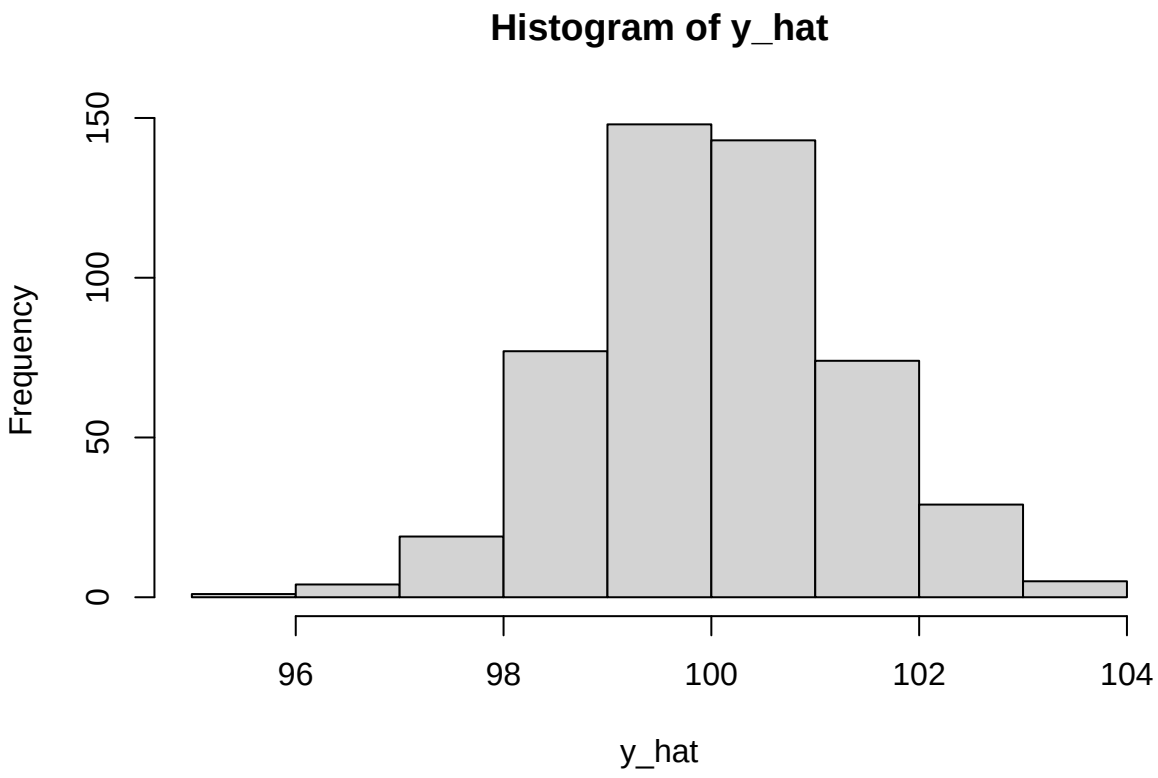
Podemos ver que la varianza en los histogramas es más alta cuando se tienen menos observaciones

- *Estimando una respuesta puntual*

```
x0 = 5
xY_hat = data.frame(x0)

for (i in 1:k) {
  xY_hat[,i+1] = xYlm[1,i+1] + xYlm[2,i+1]*x0
}
dim(xY_hat)
```

```
## [1]   1 501
```

```
y_hat = as.numeric(xY_hat[1,2:501])
hist(y_hat)
```



Nuevamente vemos que la varianza es más alta cuando se tiene una muestra más pequeña

- *Intervalo de confianza para la pendiente*

```
alpha = 0.05

IC_beta = c("IC_beta1_inf","IC_beta1_sup")
IC_B = data.frame(IC_beta)

for (i in 1:k) {
  meanx = mean(x)
  meany = mean(Y[,i+1])
  Sxy = sum(((x-meanx))*Y[,i+1])
  Sxx = sum(((x-meanx)^2))

  SS_T = sum((Y[,i+1]^2))-(((sum(Y[,i+1]))^2)/length(x))
  SS_Res = SS_T - (xYlm[2,i+1]*Sxy)
  SS_R = xYlm[2,i+1]*Sxy

  df_T = length(x) -1
  df_Res = length(x) - 2
  df_R = 1

  MS_Res = SS_Res/df_Res
  MS_R = SS_R/df_R
  se_beta1 = sqrt(MS_Res/Sxx)
```

```
  IC_B[1,i+1] = xYlm[2,i+1]-(qt((1-alpha/2),df=(length(x)-2))*(se_beta1))
  IC_B[2,i+1] = xYlm[2,i+1]+(qt((1-alpha/2),df=(length(x)-2))*(se_beta1))
}

dim(IC_B)
```

```
## [1]   2 501
```

Las longitudes de los intervalos son mayores cuando se tiene un tamaño de muestra más pequeño

- *Estimando una respuesta media*

```
alpha = 0.05
x0 = 5

IC_muyx0 = c("IC_muyx0_inf","IC_muyx0_sup")
IC_muyx0 = data.frame(IC_muyx0)

for (i in 1:k) {
  meanx = mean(x)
  meany = mean(Y[,i+1])
  Sxy = sum(((x-meanx))*Y[,i+1])
  Sxx = sum(((x-meanx)^2))

  SS_T = sum((Y[,i+1]^2))-(((sum(Y[,i+1]))^2)/length(x))
  SS_Res = SS_T - (xYlm[2,i+1]*Sxy)
  SS_R = xYlm[2,i+1]*Sxy

  df_T = length(x) -1
  df_Res = length(x) - 2
  df_R = 1

  MS_Res = SS_Res/df_Res
  MS_R = SS_R/df_R
  muyx0 =xYlm[1,i+1] + xYlm[2,i+1]*x0
  se_muyx0 = sqrt( MS_Res*( (1/length(x)) + ((x0-meanx)^2)/Sxx ) )

  IC_muyx0[1,i+1] = muyx0 - (qt((1-alpha/2),df=(length(x)-2)))*se_muyx0
  IC_muyx0[2,i+1] = muyx0 + (qt((1-alpha/2),df=(length(x)-2)))*se_muyx0
}

dim(IC_muyx0)
```

```
## [1]   2 501
```

Las longitudes de los intervalos son mayores cuando se tiene un tamaño de muestra más pequeño

Ejercicio 2.20

Se tiene el modelo de regresión lineal simple $y = \beta_0 * \beta_1 x + \varepsilon$, con $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ y ε no correlacionada.

a. Demostrar que $E(MS_R) = \sigma^2 + \beta_1^2 S_{xx}$

Solución:

Tenemos que

$$
\begin{aligned}
E(MS_R) &= E(\frac{SS_R}{df_R}) \\
&= \frac{1}{df_R} E(SS_R) \\
&= \frac{1}{df_R} E(\hat{\beta}_1 S_{xy}) \\
&= \frac{1}{df_R} E(\hat{\beta}_1{}^2 S_{xx}) \\
&= \frac{S_{xx}}{df_R} E(\hat{\beta}_1{}^2) \\
&= \frac{S_{xx}}{df_R} E(V(\hat{\beta}_1) + (E(\hat{\beta}_1))^2) \\
&= \frac{S_{xx}}{df_R} E(\frac{\sigma^2}{S_{xx}} + \beta_1^2) \\
&= \frac{\sigma^2}{df_R} + \frac{S_{xx}\beta_1^2}{df_R}
\end{aligned}
$$

donde al tener que $df_R = 1$ entonces

$$
E(MS_R) = \sigma^2 + S_{xx}\beta_1^2
$$

b. Demostrar que $E(MS_{Res}) = \sigma^2$

Solución:

Tenemos que

$$
\begin{aligned}
E(MS_{Res}) &= E(\frac{SS_{Res}}{df_{Res}}) \\
&= \frac{1}{df_{Res}} E(SS_{Res}) \\
&= \frac{1}{df_{Res}} (\sigma^2 df_{Res}) \\
&= \sigma^2
\end{aligned}
$$

Ejercicio 2.22

Considérese el estimador $7\sigma^2$ de máxima verosimilitud de σ^2 en el modelo de regresión lineal simple. Se sabe que σ^2 es un estimador sesgado de σ^2.

a. Demostrar la cantidad de sesgo en $\tilde{\sigma}^2$

Solución:

$\tilde{\sigma}^2 = \text{SSE}/n$. Luego, $E\left(\tilde{\sigma}^2\right) = \frac{n-2}{n}\sigma^2$ entonces el sesgo es $\left(1 - \frac{n-2}{n}\right)\sigma^2$

b. ¿Qué sucede con el sesgo a medida que se hace grande el tamaño n de la muestra?

Solución:

A medida que n crece el sesgo del estimador tiende a cero

Ejercicio 2.24

Se tienen los datos del problema 2.12. Supóngase que el consumo de Vapor y la temperatura ambiente tienen distribución normal conjunta.

```
y<-c(185.79,214.47,288.03,424.84,454.68,539.03,621.55,675.06,562.03,452.93,369.95,273.98)
x<-c(21,24,32,47,50,59,68,74,62,50,41,30)
```

a. Determinar la correlación entre el consumo de vapor y la temperatura ambiente promedio mensual.

Solución:

Tenemos el modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Por lo tanto el modelo ajustado se calcula de la siguiente manera:

```
meanx = mean(x)
meany = mean(y)
Sxy = sum(((x-meanx))*y)
Sxx = sum(((x-meanx)^2))
beta1 = Sxy/Sxx
beta0 = meany - beta1*meanx

SS_T = sum((y^2))-(((sum(y))^2)/length(x)) # Suma de cuadrados de las observaciones corregidas
SS_Res = SS_T - (beta1*Sxy) # Suma de cuadrados de los residuales
SS_R = beta1*Sxy # Suma de cuadrados de la regresion o suma de cuadrados del modelo

df_T = length(x) -1 # Grados de libertad de las observaciones corregidas
df_Res = length(x) - 2 # Grados de libertad de la suma de cuadrados de los residuales
df_R = 1 # Grados de libertad de la suma de cuadrados de la regresión


MS_Res = SS_Res/df_Res
MS_R = SS_R/df_R
MS_Res = SS_Res/df_Res
```

Ahora teniendo en cuenta que el estimador de la correlación dada una muestra que es

$$r = \frac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)}{\left[\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2\right]^{1/2}} = \frac{S_{xy}}{\left[S_{xx} S S_{\mathrm{T}}\right]^{1/2}}$$

tenemos que para ver la correlación entre el peso de una persona y la presión sistólica según la muestra suministrada es

```
r = Sxy/((Sxx*SS_T)^(1/2))
r
```

```
## [1] 0.9999326
```

b. Probar la hipótesis que $\rho = 0$.

Solución:

Teniendo en cuenta el estadístico de prueba

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)}$$

donde rechazamos la hipótesis de correlación nula si $|t_0| > t_{\alpha/2,n-2}$. De este modo

```
alpha = 0.05
t0 = r*(df_Res^(1/2))/((1-(r^2))^(1/2))
t0test = qt(1-alpha/2,df=df_Res)
pvalue = 2*pt(-abs(t0),df=df_Res)

t0
```

```
## [1] 272.255
```

```
t0test
```

```
## [1] 2.228139
```

```
pvalue
```

```
## [1] 1.099192e-20
```

Dado que tenemos que $|t_0| = 272.255 > t_{1-\alpha/2,n-2} = 2.228139$ ***rechazamos la hipótesis que nos dice que*** $\rho = 0$ ***con una significancia del*** 5% ***y un p valor de*** $1.099192e^{-20}$

c. Probar la hipótesis que $\rho = 0.5$.

Solución:

Para un test $H_0 : \rho = \rho_0$ contra $H_1 : \rho \neq \rho_0$ donde para $n \geq 25$ tenemos que la estadística

$$Z = \operatorname{arctanh} r = \frac{1}{2} \ln \frac{1+r}{1-r}$$

tiene aproximadamente distribución normal con media

$$\mu_Z = \operatorname{arctanh} \rho = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

y varianza

$$\sigma_Z^2 = (n-3)^{-1}$$

por lo tanto para verificar la hipótesis $H_0 : \rho = \rho_0$ tenemos que calcular la estadítica

$$Z_0 = (\operatorname{arctanh} r - \operatorname{arctanh} \rho_0)\,(n-3)^{1/2}$$

donde debemos rechazar $H_0 : \rho = \rho_0$ si $|Z_0| > Z_{1-\alpha/2}$. De este modo para ver si la correlación entre el peso de una persona y su presión sistólica es $H_0 : \rho = 0.5$ tenemos

```
p0 = 0.5
alpha = 0.05

Z0 = (atanh(r) - atanh(p0))*((length(x)-3)^(1/2))
Z0test = qnorm(1-alpha/2)
pvalue_Z0 = 2*pnorm(-abs(Z0))

Z0
```

```
## [1] 13.79796
```

```
Z0test
```

```
## [1] 1.959964
```

```
pvalue_Z0
```

```
## [1] 2.621524e-43
```

Dado que tenemos que $|Z_0| = 13.79796 > Z_{1-\alpha/2} = 1.959964$ ***rechazamos la hipótesis que nos dice que*** $\rho = 0.5$ ***con una significancia del*** 5% ***y un p valor de*** $2.621524e^{-43}$

d. Determinar un intervalo de confianza de 99% para ρ.

Solución:

Para construir un intervalo de confianza veamos que

$$\tanh\left(\operatorname{arctanh} r - \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\operatorname{arctanh} r + \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right)$$

de este modo un intervalo de confianza del 95% para ρ

```
IC_rho_inf = tanh( atanh(r) - (qnorm( 1-alpha/2 )/((length( x )-3)^(1/2))) )
IC_rho_sup = tanh( atanh(r) + (qnorm( 1-alpha/2 )/((length( x )-3)^(1/2))) )
IC_rho = c(IC_rho_inf,IC_rho_sup)
IC_rho
```

```
## [1] 0.9997509 0.9999817
```

Ejercicio 2.26

Se tiene el modelo de regresión lineal simple $y = \beta_0 + \beta_1 x + \varepsilon$ donde tenemos que β_0 es conocida

a. Determinar el estimador de β_1

Solución:

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n}\left(y_i\ -\beta_0 - \hat{\beta}_1 xi\right) x_i \\ \hat{\beta}1\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n}\left(y_i - \beta_0\right) x_i \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}\left(y_i - \beta_0\right) x_i}{\sum_{i=1}^{n} x_i^2} \end{aligned}$$

b. Cual es la varianza del estimador de la pendiente encontrado en a

Solución:

$$\begin{aligned} \operatorname{Var}(\hat{\beta}1) &= \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \operatorname{Var}\left(\sum_{i=1}^{n} y_i x_i\right) \\ &= \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2}\left(\sum_{i=1}^{n} x_i^2\right)\sigma^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \end{aligned}$$

c. Determinar un intervalo de confianza para la pendiente.

Solución:

Tenemos que

$$\frac{\hat{\beta}1 - \beta 1}{\sqrt{MS_E/\sum x_i^2}} \sim t_{n-2}$$

entonces podemos definir el intervalo de confianza como

$$\hat{\beta}1 \pm t_{\alpha/2,n-2}\sqrt{MS_E/\sum x_i^2}$$

Donde vemos que es más estrecho que cuando ambos son desconocidos.