Universidad Nacional de Córdoba

Maestría en Estadística Aplicada

Introducción al Análisis Estadístico

Matías Alfonso *
Guillermina Senn †
Andrés Peña M. ‡

Cohorte 2018-2019
Córdoba-Argentina

Junio 2018

Guía de Actividades 2

Es importante que la utilización de cada una de las herramientas estadísticas esté debidamente justificada. Se evaluará no sólo la aplicación de las mismas sino también la capacidad para identificar adecuadamente las estrategias y los modelos apropiados.

La debida fundamentación del trabajo realizado es esencial para que el docente pueda evaluar los logros obtenidos en el proceso de enseñanza-aprendizaje, y pueda realizar los ajustes necesarios antes de llegar a las instancias finales de evaluación.

Si bien los problemas están planteados bajo la forma de ejercicios, no se limite a la mera aplicación de fórmulas, sino que haga una defensa conceptual de la solución planteada, y elabore conclusiones concretas en función del problema analizado.

Docentes: Dra. Patricia Caro y Mgter. Mariana Gonzalez.

*matias.alfonso@gmail.com
†senn.guillermina@gmail.com
‡andres.pena.montalvo@gmail.com

I. Actividad 1

Seguiremos en esta actividad trabajando con los salarios de los profesores en una universidad de Ohio, U.S.A. entre 1993-1994 para investigar si existían pruebas de la desigualdad de género en los salarios de la universidad como así también estudiar la correlación del salario con otras variables.

Se sabe que la variable Salario durante el año académico de los profesores titulares sigue una distribución con media de 61133 dólares y desviación estándar de 9500 dólares.

a) Se elige una muestra de 16 profesores (con reemplazo) y se desea determinar la probabilidad de que la media muestral difiera del promedio poblacional en menos de 1000 dólares.

1) Realicen el cálculo considerando que la variable Salario se distribuye Normal. Dejen indicada la distribución de probabilidad de la variable Media Muestral. Justifiquen.

Sabemos que en la población la variable salario X sigue una distribución normal:

$$X \sim N(\mu,\ \sigma^2)$$

$$X \sim N(61333,\ 9500^2)$$

entonces la distribución de probabilidad de la variable aleatoria $\bar{X}$ (Media Muestral) es la siguiente:

$$\bar{X} \sim N(\mu,\ \frac{\sigma}{\sqrt{n}})$$

$$\bar{X} \sim N(61333,\ 9500/\sqrt{16})$$

Queremos calcular entonces:

$$\begin{aligned} P(|\bar{X} - \mu| < 1000) &= P(-1000 < \bar{X} - \mu < 1000) \\ &= P(-1000 + \mu < \bar{X} < 1000 + \mu) \\ &= P(-1000 + 61133 < \bar{X} < 1000 + 61133) \\ &= P(60133 < \bar{X} < 62133) = F(62133) - F(60133) \\ &= 0{,}3263 \end{aligned}$$

```
mu <- 61133
sd <- 9500
n <- 16

pnorm(62133, mu, sd/sqrt(n)) - pnorm(60133, mu, sd/sqrt(n))

## [1] 0.33
```

2) Realicen el cálculo considerando que la variable Salario tiene distribución desconocida. Justifiquen

Como la población tiene distribución desconocida y el tamaño de la muestra es chico, se podría aplicar la desigualdad de Chevyshev.

$$P(|\bar{X} - \mu| \leq d) \geq 1 - \frac{\sigma_{\bar{x}}^2}{d^2} \;=\; P(|\bar{X} - \mu| \leq 1000) \geq 1 - \frac{\frac{9500^2}{16}}{1000^2}$$
$$= -4{,}6106$$

```
1 - 9500^2/(1000^2 * 16)
```

```
## [1] -4.6
```

Sin embargo la regla de Chevyshev sólo permite establecer cotas de probabilidad para intervalos simétricos, en los que el número d que se suma y resta a la media es mayor que la desviación estándar. En este caso como $d = 1000$ es menor que $\sigma = 9500$ la probabilidad es inferior a cero, por lo cual la desigualdad no es aplicable. Si se tuviera la información muestral detallada se podría optar por realizar una transformación (logarítmica por ejemplo) a la variable Salario para que cumpla propiedades de normalidad y poder darle el tratamiento como tal.

b) Se elige una muestra de 64 profesores (con reemplazo) y se desea determinar la probabilidad de que la media muestral difiera del promedio poblacional en menos de 1000 dólares. Realicen el cálculo y justifiquen.

En este caso, como $n > 30$, podemos utilizar el Teorema Central del Límite, y realizar una aproximación por la distribución normal.

$$\begin{aligned} P(\mu_0 < \bar{X} < \mu_1) &= P(60133 < \bar{X} < 62133) \\ &= P(\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} < Z < \frac{\overline{X}_n - \mu_1}{\sigma/\sqrt{n}}) \\ &= \Phi_{0,1}(\frac{62133 - 61133}{9500/\sqrt{64}}) - \Phi_{0,1}(\frac{60133 - 61133}{9500/\sqrt{64}}) \\ &= 0{,}6003 \end{aligned}$$

```
mu <- 61133
sd <- 9500
n <- 64

pnorm((62133-mu)/(sd/sqrt(n)))-pnorm((60133-mu)/(sd/sqrt(n))) #Estandarizando
```

```
## [1] 0.6
```

```
pnorm(62133, mu, sd/sqrt(n)) - pnorm(60133, mu, sd/sqrt(n)) #Directo
```

```
## [1] 0.6
```

II. Actividad 2

Los ejercicios que siguen deben resolverse a partir de la información de la base de datos "Universidad".

a) Construyan un intervalo del 98 % de confianza para la proporción de docentes mujeres, analizado por cargo en la Universidad. Interpreten.

Primero calculamos la proporción de docentes mujeres en la Universidad:

```
tabla <- table(factor(universidad$genero, labels = c("Femenino",
                                                     "Masculino")))
prop <- as.data.frame(prop.table(tabla)); names(prop)<-c("Género", "p")
xtable(prop, "Proporción estimada")
```

	Género	p
1	Femenino	0.25
2	Masculino	0.75

Cuadro 1: Proporción estimada

Por otro lado, $\hat{p} * n = 128$

```
p <- prop[prop$Género=="Femenino", "p"]
n <- nrow(universidad)

n*p

## [1] 128
```

Como n*p es mayor a 5, podemos utilizar la aproximación normal de $\hat{p}$. Entonces:

$$P(-z < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z) = 1 - \alpha$$

$$LIC = \hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$LSC = \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Establecemos un nivel de significación de 0.02 ($\alpha = 0{,}02$) y buscamos los valores de $z_{\frac{\alpha}{2}}$ y $z_{\frac{(1-\alpha)}{2}}$ y construimos el intervalo de confianza para p.

```
confint <- p + c(1, -1) * qnorm(.01) * sqrt(p * (1-p) / n)
confint

## [1] 0.20 0.29
```

Tenemos entonces un 98 % de probabilidad de que el intervalo 0.2 y 0.29 contenga la proporción de mujeres en la población.

b) Determinen el tamaño de muestra (con reemplazo) necesario para reducir a 0,05 la amplitud del intervalo de confianza anterior, para los profesores Titulares. Consideren un valor de p=0,05.

Sabemos que la amplitud del intervalo de confianza es $LS - LI = 2e$, entonces:

$$\begin{aligned} 2e &= 0{,}05 \\ e &= 0{,}025 \end{aligned}$$

Ahora bien, sabemos que:

$$\begin{aligned} n &= \frac{z^2_{0{,}99}\hat{p}(1-\hat{p})}{e^2} \\ &\approx 810 \end{aligned}$$

```
z <- qnorm(.99)
p <- 0.05
e <- 0.025
z^2 * p * (1-p) / e^2

## [1] 411
```

III. Actividad 3

¿Existen evidencias para apoyar la hipótesis de que la varianza de la antigüedad es significativamente distinta de 50 horas? Considere un nivel de significación del 0,01. En caso de corresponder construya una estimación por intervalo del 99 % e interprete. ¿Qué supuesto/s debe hacer para aplicar la prueba? Evalúe si se cumplen. Plantee las hipótesis de trabajo y el estadístico de prueba[1].

Tenemos una función que nos relaciona la varianza muestral con la varianza poblacional, a saber:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

[1]En esta pregunta hay algún error en la redacción. Pregunta si la varianza es diferente a 50 horas. Sin embargo la antigüedad está medida en años, y al ser la varianza, son años[2]

Buscamos entonces

$$P\left(\chi^2_{\frac{\alpha}{2}} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\frac{(1-\alpha)}{2}}\right)$$
$$= P\left(\frac{(n-1)s^2}{\chi^2_{\frac{(1-\alpha)}{2}}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}\right)$$

Identificamos las hipótesis:

$$H_0 : \sigma^2 = 50$$
$$H_1 : \sigma^2 \neq 50$$

Buscamos los valores para $\chi^2_{\frac{\sigma}{2}}$ y $\chi^2_{\frac{(1-\sigma)}{2}}$ considerando $\alpha = 0{,}01$ y evaluamos si el χ_{obs} se encuentra dentro del intervalo.

```
s <- sd(universidad$antiguedad)
n <- length(universidad$antiguedad)
alpha <- .01

## Chi observado
(n-1)*s^2/50

## [1] 889

## Chi teóricos
qchisq(c(alpha/2, (1-alpha)/2), n-1)

## [1] 434 512
```

Como el χ_{obs} se encuentra en la zona de rechazo, entonces rechazo la H_0. Es decir que la varianza es significativamente distinta ($\alpha = 0{,}01$) a 50 aos^2. Construimos un intervalo de confianza para la varianza,

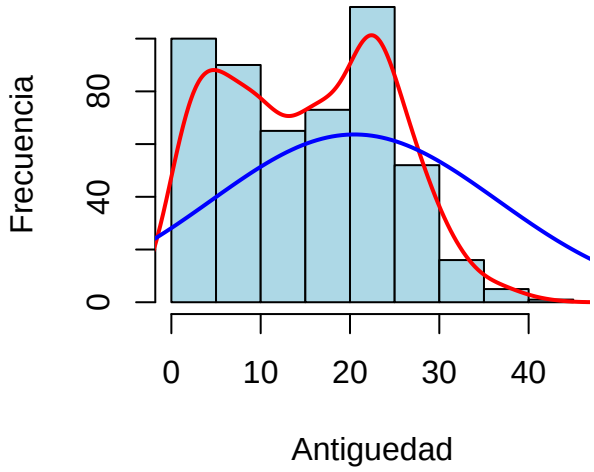
$$P\left(\frac{(n-1)s^2}{\chi^2_{\frac{(1-\alpha)}{2}}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}\right)$$

Construimos el intervalo:

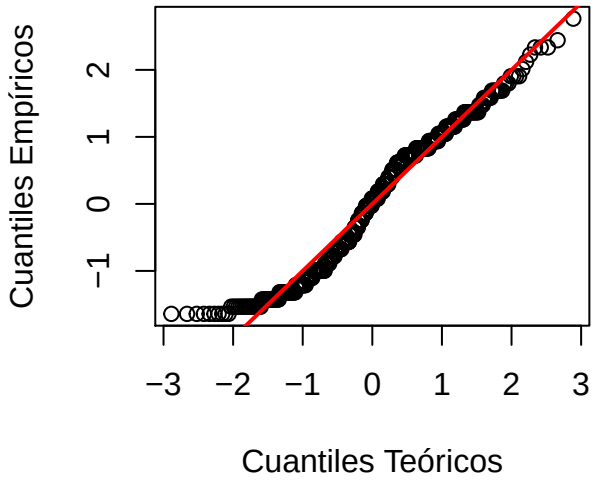
```
## Intervalo
(n-1)*s^2/qchisq(c(0.995, .005), n-1)

## [1]  74 102
```

La prueba de hipótesis supone que la variable se distribuye de manera normal. Evaluemos si estos supuestos se cumplen:

```r
## Creo el histograma
h <- hist(universidad$antiguedad, col = "lightblue",
          xlab = "Antiguedad",
          ylab = "Frecuencia",
          main = "")

## Calculamos la densidad empírica
d <- density(universidad$antiguedad)
x <- d$x
y <- d$y * length(universidad$antiguedad) * diff(h$mids[1:2])

## Agregamos la densidad empírica al gráfico
lines(x, y, col = "red", lwd = 2)

## Cuantiles teóricos
xfit <- seq(min(x),max(x),length=100)
yfit <- dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```




```r
## Realizamos un qq-plot
z <- with(universidad, (antiguedad - mean(antiguedad))/sd(antiguedad))
cuantiles <- cumsum(rep(1/(length(z) + 1), length(z)))
cuantiles <- qnorm(cuantiles)

plot(cuantiles, z[order(z)],
     xlab = "Cuantiles Teóricos",
     ylab = "Cuantiles Empíricos")
abline(0, 1, col = "red", lwd = 2)
```




```
## Prueba de Normalidad - Shapiro Wilks
shapiro.test(universidad$antiguedad)

##
## Results of Hypothesis Test
## --------------------------
##
## Alternative Hypothesis:
##
## Test Name:                          Shapiro-Wilk normality test
##
## Data:                               universidad$antiguedad
##
## Test Statistic:                     W = 0.96
##
## P-value:                            0.00000000033

## Prueba de Normalidad - Kolmorogov Smirnof
media<-mean(universidad$antiguedad)
desv<-sd(universidad$antiguedad)
ks.test(universidad$antiguedad, "pnorm",
        mean=media, sd=desv)

##
## Results of Hypothesis Test
## --------------------------
##
## Alternative Hypothesis:             two-sided
##
```

```
## Test Name:                          One-sample Kolmogorov-Smirnov test
##
## Data:                               universidad$antiguedad
##
## Test Statistic:                     D = 0.093
##
## P-value:                            0.00029
```

```
## Asimetría y curtosis
skewness(universidad$antiguedad)
```

```
## [1] 0.11
```

```
kurtosis(universidad$antiguedad)
```

```
## [1] -0.99
```

Si bien la prueba de Shapiro-Wilks da significativa, este procedimiento es sensible a pequeñas desviaciones cuando el tamaño de la muestra es grande, como en este caso. La distribución es levemente asimétrica a la derecha y platicúrtica. El qqplot muestra un ajuste regularmente bueno a la distribución normal.

IV. Actividad 4

a) Prueben, mediante una prueba estadística adecuada, si existen evidencias de diferencias estadísticamente significativas en el salario de los profesores según su género. No olviden el planteo de las correspondientes hipótesis, así como el análisis de los resultados obtenidos. Trabajen con un nivel de significación del 5 %.

Para evaluar si existen diferencias significativas en el salario de los profesores según su género, realizamos una prueba t para diferencia de medias, con muestras independientes y varianzas supuestas desiguales. Las hipótesis estadísticas son:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Veamos los resultados:

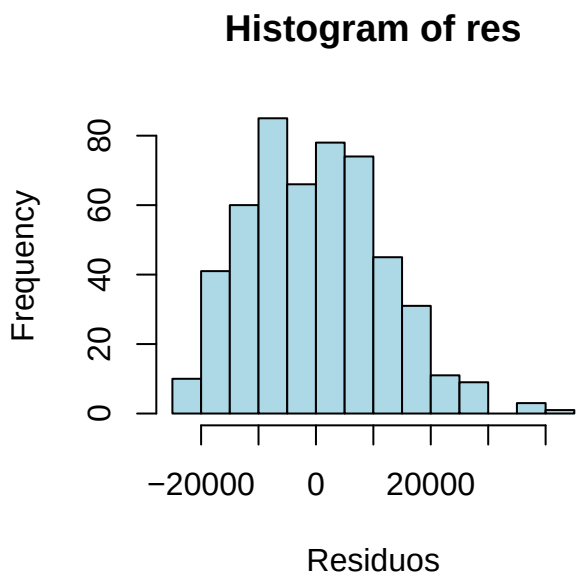
```
with(universidad, t.test(salario[genero == 1], salario[genero == 0],
                         var.equal = FALSE))
```

```
##
## Results of Hypothesis Test
## --------------------------
##
```

```
## Null Hypothesis:                  difference in means = 0
##
## Alternative Hypothesis:           True difference in means is not equal to 0
##
## Test Name:                        Welch Two Sample t-test
##
## Estimated Parameter(s):           mean of x = 53499
##                                   mean of y = 42917
##
## Data:                             salario[genero == 1] and salario[genero == 0]
##
## Test Statistic:                   t = 10
##
## Test Statistic Parameter:         df = 297
##
## P-value:                          0.000000000000000000026
##
## 95% Confidence Interval:          LCL =  8551
##                                   UCL = 12614
```

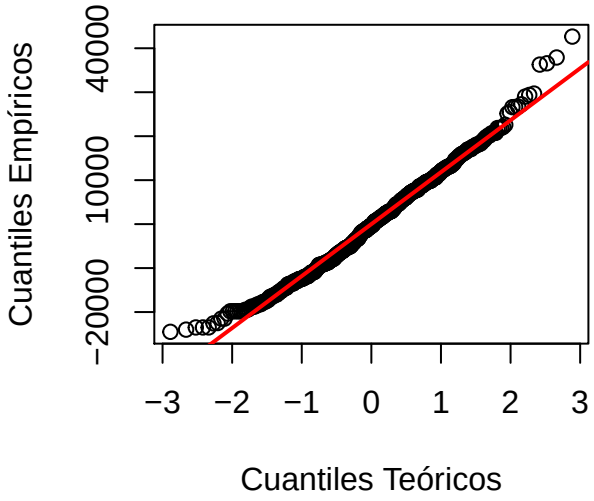
```
## Supuesto de normalidad
res <- with(universidad, c(salario[genero == 0] - mean(salario[genero == 0]),
                           salario[genero == 1] - mean(salario[genero == 1])))
```

```
hist(res, col = "lightblue", xlab = "Residuos")
```



```r
## Realizamos un qq-plot
cuantiles <- cumsum(rep(1/(length(res) + 1), length(res)))
cuantiles <- qnorm(cuantiles)
```

```r
plot(cuantiles, res[order(res)],
     xlab = "Cuantiles Teóricos",
     ylab = "Cuantiles Empíricos")
abline(0, sd(res), col = "red", lwd = 2)
```




```r
## Shapiro-Wilk
shapiro.test(res)
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Alternative Hypothesis:
##
## Test Name:                          Shapiro-Wilk normality test
##
## Data:                               res
##
## Test Statistic:                     W = 0.99
##
## P-value:                            0.0002
```

```r
## Prueba de Normalidad - Kolmorogov Smirnof
```

```r
#media<-0
media<-mean(res)
desv<-sd(res)
ks.test(res, "pnorm",
        mean=media, sd=desv)

##
## Results of Hypothesis Test
## --------------------------
##
## Alternative Hypothesis:          two-sided
##
## Test Name:                       One-sample Kolmogorov-Smirnov test
##
## Data:                            res
##
## Test Statistic:                  D = 0.046
##
## P-value:                         0.23

## Prueba de igualdad de varianzas
with(universidad, var.test(salario[genero == 0], salario[genero == 1]))

##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 ratio of variances = 1
##
## Alternative Hypothesis:          True ratio of variances is not equal to 1
##
## Test Name:                       F test to compare two variances
##
## Estimated Parameter(s):          ratio of variances = 0.53
##
## Data:                            salario[genero == 0] and salario[genero == 1]
##
## Test Statistic:                  F = 0.53
##
## Test Statistic Parameters:       num df   = 127
##                                  denom df = 385
##
## P-value:                         0.000039
##
## 95% Confidence Interval:         LCL = 0.40
##                                  UCL = 0.71

##Prueba de Levene para la igualdad de varianzas
levene.test(universidad$salario, universidad$genero, location="mean")
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Alternative Hypothesis:
##
## Test Name:                        classical Levene's test based on the absolute deviat
##
## Data:                             universidad$salario
##
## Test Statistic:                   Test Statistic = 15
##
## P-value:                          0.00011
```

La prueba de normalidad Shapiro-Wilk arroja una diferencia significativa para el supuesto de normalidad. Sin embargo, esta prueba es sensible a pequeñas desviaciones cuando el tamaño de la muestra es grande. Mediante análisis gráficos de los residuos podemos observar que son aproximadamente normales. La prueba para la igualdad de varianzas nos dio una diferencia significativa ($F_{(127;385)} = 0.53$, p <0.01) para el supuesto de igualdad, por lo que realizamos una prueba t para diferencia de medias con varianzas supestas desiguales. Observamos una diferencia significativa para la diferencia de medias ($t_{(297,23)} = 10.25$, p <0.01). El salario de las mujeres (M = 42916.60) es significativamente menor al de los hombres (M = 53499.24). Hay un 95 % de confianza de que las mujeres cobren entre 8213 U$D y 12953 U$D menos que los hombres.

b) En caso de corresponder construyan una estimación por intervalo del 95 % e interpreten.

V. Actividad 5

Propongan y realicen otra prueba de hipótesis de las estudiadas en la materia, que sea coherente con los objetivos del estudio[2]. Justifiquen la elección del instrumento estadístico e interpreten los resultados obtenidos.

Considerando que nos interesa saber si existe desigualdades de género en los salarios de la universidad, podemos preguntarnos si la rentabilidad de los salarios es diferente para hombres y mujeres. Para ello, realizamos una prueba t para la diferencia de medias. Las hipótesis estadísticas son:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Veamos los resultados:

[2]Recuerden que el propósito del estudio fue investigar si existían pruebas de la desigualdad de género en los salarios de la universidad como así también estudiar la correlación del salario con otras variables.

```
with(universidad, t.test(rentabilidad[genero == 1], rentabilidad[genero == 0],
                         var.equal = TRUE))
```

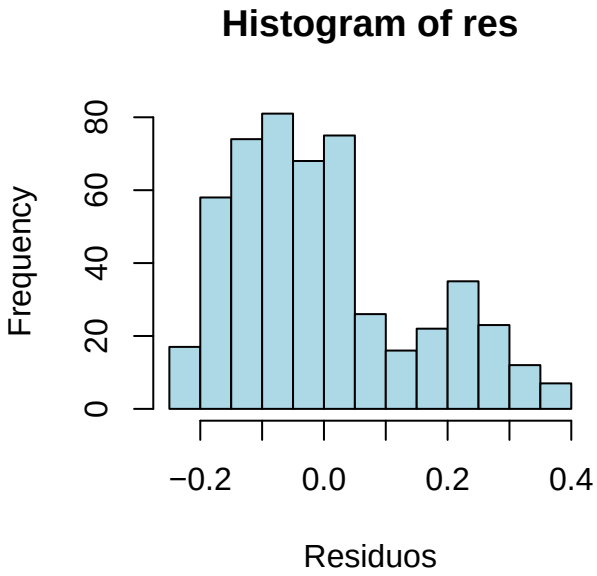
```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 difference in means = 0
##
## Alternative Hypothesis:          True difference in means is not equal to 0
##
## Test Name:                        Two Sample t-test
##
## Estimated Parameter(s):          mean of x = 0.96
##                                  mean of y = 0.90
##
## Data:                            rentabilidad[genero == 1] and rentabilidad[genero == 0]
##
## Test Statistic:                  t = 4.2
##
## Test Statistic Parameter:        df = 512
##
## P-value:                         0.000036
##
## 95% Confidence Interval:         LCL = 0.033
##                                  UCL = 0.092
```

Los salarios de las mujeres son significativamente menos rentables que los de los hombres, p <0.01, 95 % CI[0.033, 0.092]. Es decir que hay un 95 % de confianza en que el trabajo de las mujeres es entre un 3.31 % y un 9.2 % menos rentable que el de los hombres. Recordemos que la rentabilidad está definida como el ratio entre el salario promedio nacional pagado en la disciplina y el promedio nacional de todas las disciplinas.

Evaluemos si se cumplen los supuestos de normalidad y homogeneidad de varianzas.

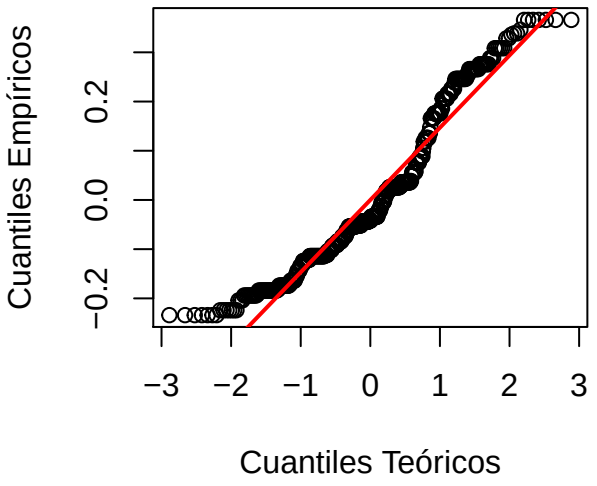
```
res <- with(universidad, c(rentabilidad[genero == 0] - mean(rentabilidad[genero == 0]),
                           rentabilidad[genero == 1] - mean(rentabilidad[genero == 1])))
```

```
hist(res, col = "lightblue", xlab = "Residuos")
```




```
## Realizamos un qq-plot
cuantiles <- cumsum(rep(1/(length(res) + 1), length(res)))
cuantiles <- qnorm(cuantiles)
```

```
plot(cuantiles, res[order(res)],
     xlab = "Cuantiles Teóricos",
     ylab = "Cuantiles Empíricos")
abline(0, sd(res), col = "red", lwd = 2)
```



```
## Shapiro-Wilk
shapiro.test(res)

##
## Results of Hypothesis Test
## --------------------------
##
## Alternative Hypothesis:
##
## Test Name:                          Shapiro-Wilk normality test
##
## Data:                               res
##
## Test Statistic:                     W = 0.94
##
## P-value:                            0.000000000000077

## Prueba de igualdad de varianzas
with(universidad, var.test(rentabilidad[genero == 0], rentabilidad[genero == 1]))

##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                    ratio of variances = 1
##
## Alternative Hypothesis:             True ratio of variances is not equal to 1
##
## Test Name:                          F test to compare two variances
##
## Estimated Parameter(s):             ratio of variances = 0.75
##
## Data:                               rentabilidad[genero == 0] and rentabilidad[genero == 1]
##
## Test Statistic:                     F = 0.75
##
## Test Statistic Parameters:          num df   = 127
##                                     denom df = 385
##
## P-value:                            0.055
##
## 95% Confidence Interval:            LCL = 0.57
##                                     UCL = 1.01
```

Los análisis gráficos y la prueba Shapiro-Wilk muestran que es poco probable que la población tenga una distribución normal. Habrá que considerar otras opciones para evaluar la diferencia de medias.

VI. Actividad 6

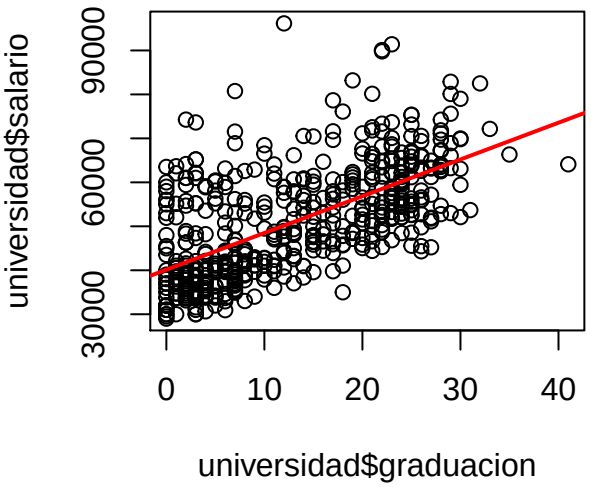
Se sabe que los salarios pueden variar considerablemente según diferentes factores. Estudien la relación entre las variables Salario y Graduación (tiempo desde la graduación en años) mediante un modelo de regresión lineal. Incluyan el análisis de los residuos del modelo. Interpreten los resultados obtenidos.

Se realizó una regresión lineal simple para predecir el salario en función del tiempo de graduación.

```
lm <- lm(universidad$salario ~ universidad$graduacion)
xtable(summary(lm))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	40117.2423	739.4218	54.25	0.0000
universidad$graduacion	836.1744	46.3766	18.03	0.0000

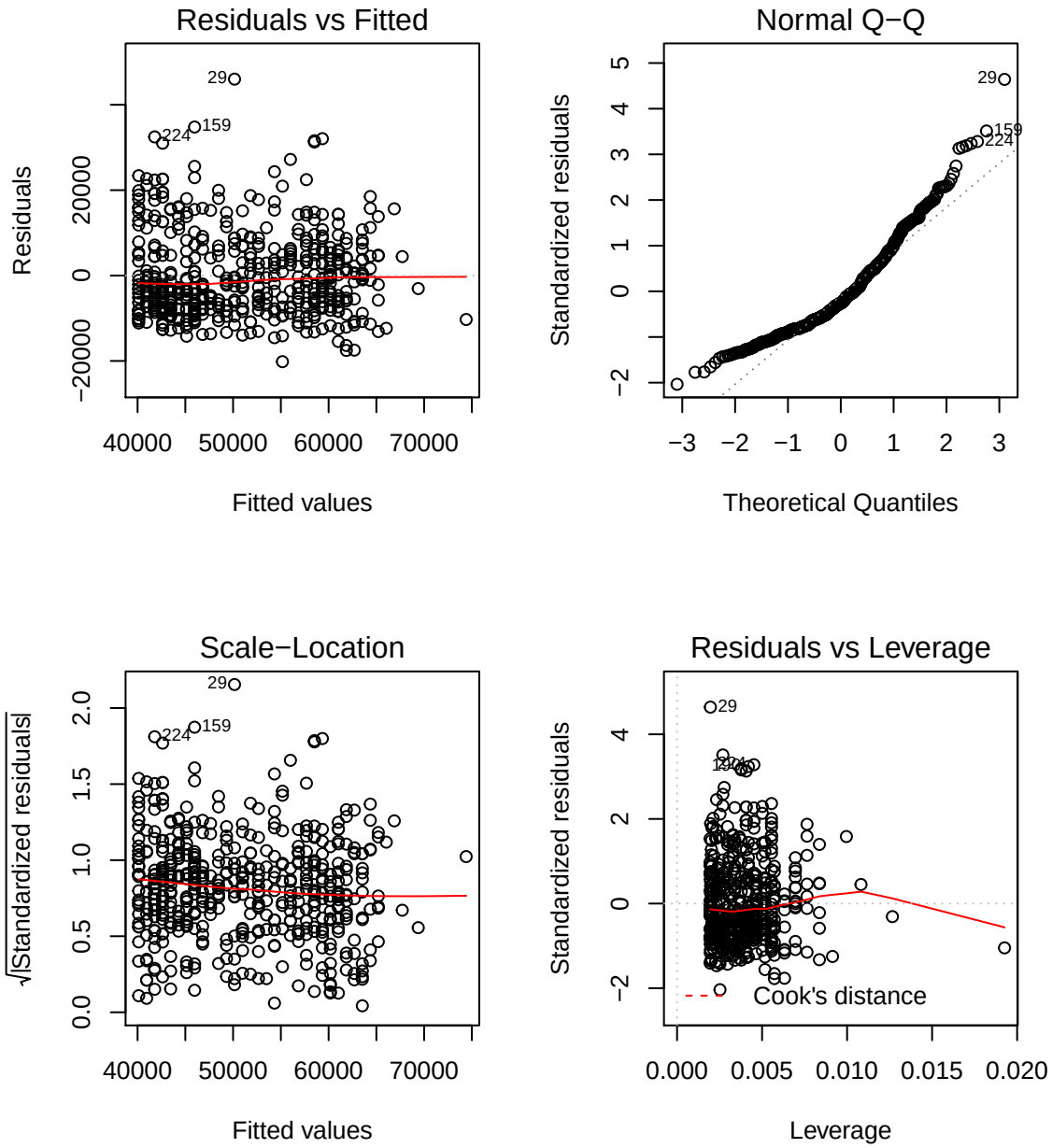
```
plot(universidad$salario ~ universidad$graduacion)
abline(lm, col = "red", lwd = 2)
```



El análisis arrojó un efecto significativo ($F_{(1,512)} = 325,1, p < 0,01$)) para la regresión. El salario predicho por el modelo es igual a $40114{,}24 + 836{,}17$ (tiempo de graduación) para el tiempo medido en años. El salario se incrementó 836.17 USD por año transcurrido desde la graduación (p <0.01).

Los modelos lineales suponen que los residuos se distribuyen de manera normal y homogénea. Observemos los residuos del modelo.

```
par(mfrow=c(2,2))
plot(lm)
```



El análisis gráfico de los residuos nos muestra que el ajuste normal es regular.

```
ks.test(lm$residuals, "pnorm", mean(lm$residuals), sd(lm$residuals))


##
## Results of Hypothesis Test
## --------------------------
```

```
##
## Alternative Hypothesis:          two-sided
##
## Test Name:                       One-sample Kolmogorov-Smirnov test
##
## Data:                            lm$residuals
##
## Test Statistic:                  D = 0.11
##
## P-value:                         0.000014
```