Punto 2° Parcial: Ajuste de un modelo de R.L.M

Universidad Nacional de Colombia
Analisís de Regresión 2022-1S
Medellín, Colombia
2022

Daniel Villa 1005087556 Juan Pablo Vanegas 1000640165

UNIVERSIDAD NACIONAL DE COLOMBIA

Contents

1 Objetivos:

Crear un modelo ajustado de R.L.M. por el cual se pueda predecir la estatura de un individuo (discrimiando por genero) sabiendo las estaturas de los padres (madre y padre) utilizando el software estadístico *R*.

1.1 Objetivos específicos

- Plantear el modelo de R.L.M.
- Interpretar los parámetros del modelo.
- Determinar si el efecto de las estaturas de los padres sobre la estatura del sujeto es significativo.
- Interpretar nuestro R^2.
- Validar los supuestos del modelo.
- Aplicar la prueba de falta de ajuste.

2 Antecedentes Relevantes

La población encuestada, pertenece a estudiantes de la Universidad Nacional de Colombia sede Medellín de diferentes carreras, es decir, la mayoria de los sujetos de la muestra son jovenes entre los 18 y los 25 años, además decidimos que solamente aquellos que tenian la posibilidad de saber las estaturas de sus padres entraban a nuestra base de datos, ya que el porceso sería más arduo si tomamos datos donde nos faltan llenar valores en las celdas correspondientes.

3 Variables de respuesta:

En nuestro caso será la estatura del sujeto (Hombre o Mujer) para ajustar un modelo para predecir por medio de nuestras variables predictoras la esturura del sujeto.

4 Variable de Control:

En este caso tendremos 3 varibles haciendo de este un modelo de R.L.M.

1. Estatura del Padre.
2. Estatura del Madre.
3. Genero del sujeto.

5 Creación del Modelo

Nuestro modelo tendra la forma de:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i,\ i = 1, ..., 30 \ \ donde\ \varepsilon_i \sim N(0, \sigma^2)$$

Para llegar a lo anterior primero miraremos nuestra base de datos y como se comportan estas variables *(algunos datos...)*.

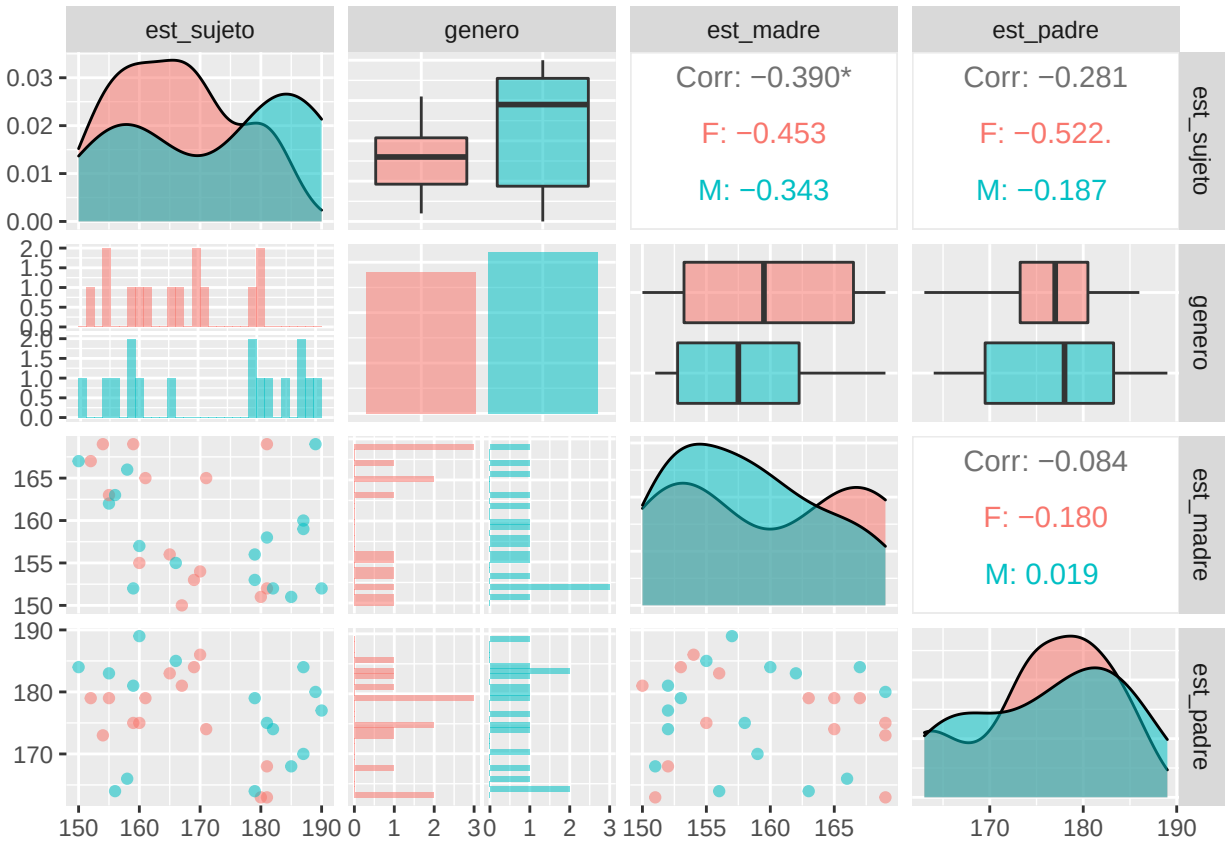
est_sujeto	genero	exp_sujeto	est_madre	est_padre	ced_sujeto	ced_madre	ced_padre	exp_padre	exp
180	F	1/11/2017	151	163	2567	1262	8323	18/10/1965	1/1
181	M	22/6/2016	158	175	6044	5350	0316	21/6/1984	15/
187	M	26/9/2020	159	170	6946	4240	0659	31/12/1977	6/
155	M	8/10/2020	162	183	7023	2121	4144	21/1/1970	4/
159	F	21/6/2016	169	175	1791	7832	3727	11/6/1987	19/
.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	
150	M	17/11/2016	167	184	8126	9626	9357	30/6/1981	13/
154	F	5/2/2020	169	173	9538	2408	4497	12/9/1982	30/
167	F	7/9/2021	150	181	2894	7612	4657	22/3/1980	31/
156	M	18/6/2020	163	164	8849	4763	8827	9/6/1989	23/
170	F	4/8/2017	154	186	9227	4016	1264	29/7/1968	10/

Definimos nuestras variables:

- `est_sujeto`: Estatura en cm del sujeto encuestado.
- `genero`: Genero del sujeto encuestado.
- `exp_sujeto`: Fecha de expedición de la cedula del sujeto.
- `est_madre`: Estatura de la madre del sujeto en cm.
- `est_padre`: Estatura del padre del sujeto en cm.
- `ced_sujeto`: Ultimos cuatro dijitos de la cedula del sujeto encuestado.
- `ced_madre`: Ultimos cuatro dijitos de la cedula de la madre.
- `ced_padre`: Ultimos cuatro dijitos de la cedula del padre.
- `exp_madre`: Fecha de expedición de la cedula de la madre.
- `exp_padre`: Fecha de expedición de la cedula del padre.

6 Análisis de correlación

Comenzamos representando los datos en una nube de puntos múltiple, donde vemos la relación entre cada par de variables.



Según la sigueinte tabla veremos que tan debil o fuerte son los datos respecto a otros:

Figure 1: corr_guide

como podemos apreciar nuestras 3 correlaciones entre cada variable predictora y nuestra Y es una *Correlación Negativa débil* ya que ninguno supera ± 0.5.

7 Ajuste del modelo

```
##
## Call:
## lm(formula = est ~ est_pa + est_ma + genero, data = stature)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.854  -6.954   0.101   4.956  26.325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 381.5306    73.4427   5.195 2.01e-05 ***
## est_pa       -0.5417     0.2766  -1.959   0.0609 .
## est_ma       -0.7519     0.3204  -2.346   0.0269 *
## generoM       5.7221     4.1315   1.385   0.1778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.2 on 26 degrees of freedom
## Multiple R-squared:  0.3025, Adjusted R-squared:  0.222
## F-statistic: 3.758 on 3 and 26 DF,  p-value: 0.02294
```

El error típico residual es `11.2`, la $R^2 = 0.3025$, aunque para el modelo múltiple es mejor fijarnos en su valor ajustado $R^2_a = 0.222$. Esto que significa que la recta de regresión explica el 22.22% de la variabilidad del modelo. Además, `F = 3.758` con una significación $p < 0.05$, lo que nos dice que nuestro modelo de regresión resulta un poco mejor que el modelo básico.

8 Comparación de modelos

Pretendemos seleccionar el "mejor" subconjunto de predictores por varias razones:

1. Explicar los datos de la manera más simple. Debemos eliminar predictores redundantes.
2. Predictores innecesarios añade ruido a las estimaciones.
3. La causa de la multicolinealidad es tener demasiadas variables tratando de hacer el mismo trabajo. Eliminar el exceso de predictores ayuda a la interpretación del modelo.
4. Si vamos a utilizar el modelo para la predicción, podemos ahorrar tiempo y/o dinero al no medir predictores redundantes.

Puesto que tenemos dos variables explicativas disponemos de 6 modelos posibles:

$$modelo\ 1: \ est \sim est_{ma} + est_{pa} + genero$$

$$modelo\ 2: \ est \sim est_{ma} + genero$$

$$modelo\ 3: \ est \sim est_{pa} + genero$$

$$modelo\ 4: \ est \sim est_{ma}$$

$$modelo\ 5: \ est \sim est_{pa}$$

$$modelo\ 6: \ est \sim genero$$

Vamos a ajustar cada uno de los modelos

```
##
## Call:
## lm(formula = est ~ est_ma + genero, data = stature)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.075  -9.035  -1.057   8.030  23.859
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 278.2976    53.7713   5.176  1.9e-05 ***
## est_ma       -0.7020     0.3358  -2.091   0.0461 *
## generoM       5.4878     4.3413   1.264   0.2170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.77 on 27 degrees of freedom
## Multiple R-squared:  0.1995, Adjusted R-squared:  0.1402
## F-statistic: 3.365 on 2 and 27 DF,  p-value: 0.04957
```

```
##
## Call:
## lm(formula = est ~ est_pa + genero, data = stature)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.784  -9.021   2.940   8.162  18.059
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 252.2777    52.4691   4.808  5.1e-05 ***
## est_pa       -0.4902     0.2978  -1.646    0.111
## generoM       6.9006     4.4298   1.558    0.131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.1 on 27 degrees of freedom
## Multiple R-squared:  0.1548, Adjusted R-squared:  0.09216
## F-statistic: 2.472 on 2 and 27 DF,  p-value: 0.1033
```

```
##
## Call:
## lm(formula = est ~ est_ma, data = stature)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.884 -10.284  -3.723   7.131  26.948
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 289.6144    53.5839   5.405 9.18e-06 ***
## est_ma       -0.7548     0.3367  -2.242   0.0331 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.9 on 28 degrees of freedom
## Multiple R-squared:  0.1522, Adjusted R-squared:  0.1219
## F-statistic: 5.025 on 1 and 28 DF,  p-value: 0.0331
```

```
##
## Call:
## lm(formula = est ~ est_pa, data = stature)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.3440 -10.9008   0.4737   9.9393  21.2097
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 252.7696    53.7882   4.699 6.31e-05 ***
## est_pa       -0.4721     0.3051  -1.548    0.133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.4 on 28 degrees of freedom
## Multiple R-squared:  0.0788, Adjusted R-squared:  0.0459
## F-statistic: 2.395 on 1 and 28 DF,  p-value: 0.1329
```

```
##
## Call:
## lm(formula = est ~ genero, data = stature)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.688 -11.821   1.929  11.562  17.312
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  166.071      3.330  49.874   <2e-16 ***
## generoM        6.616      4.560   1.451    0.158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.46 on 28 degrees of freedom
## Multiple R-squared:  0.06994,    Adjusted R-squared:  0.03672
## F-statistic: 2.106 on 1 and 28 DF,  p-value: 0.1579
```

Para evitar la elección subjetiva del mejor modelo, podemos comparar todos los modelos mediante una tabla ANOVA conjunta para cada par de modelos.

Nota: Se escoje el de menor error estandar (RSS)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
27	3740.781	NA	NA	NA	NA
27	3949.951	0	-209.1702	NA	NA

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
27	3740.781	NA	NA	NA	NA
28	3962.171	-1	-221.3902	1.597938	0.2170003

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
27	3740.781	NA	NA	NA	NA
28	4304.956	-1	-564.1753	4.072073	0.0536355

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
27	3740.781	NA	NA	NA	NA
28	4346.366	-1	-605.5851	4.370958	0.0461022

```
## [1] "Mejor modelo del 2 al 6 con menor RSS"
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
27	3740.781	NA	NA	NA	NA
26	3259.715	1	481.066	3.837058	0.0609385

Comparando ambas tablas anova deducimos que el modelo que mejor se ajusta a los datos es el `modelo2` pues reduce el error estándar.

9 Selección del "mejor" modelo

Existen distintos métodos a la hora de construir un modelo complejo de regresión con varios predictores

- El **método jerárquico** en el que se seleccionan los predictores basándose en un trabajo anterior y el investigador decide en qué orden introducir las variables predictoras al modelo.
- El **método de entrada forzada** en el que todas las variables entran a la fuerza en el modelo simultáneamente.
- Los **métodos paso a paso** que se basan en un criterio matemático para decidir el orden en que los predictores entran en el modelo.

Nosotros vamos a utilizar en R los métodos paso a paso

9.1 Resumen de las variables de interes

```
##       est        genero     est_ma          est_pa
##  Min.   :150.0   F:14   Min.   :150.0   Min.   :163.0
##  1st Qu.:159.0   M:16   1st Qu.:153.0   1st Qu.:170.8
##  Median :168.0          Median :157.5   Median :178.0
##  Mean   :169.6          Mean   :159.0   Mean   :176.2
##  3rd Qu.:181.0          3rd Qu.:165.0   3rd Qu.:182.5
##  Max.   :190.0          Max.   :169.0   Max.   :189.0
```

En todas las variables explicativas los valores de la media y la mediana son muy cercanos, lo cual es muy bueno.

9.2 Correlación

```
##
##  Pearson's product-moment correlation
##
## data:  df$est and df$est_ma
## t = -2.2416, df = 28, p-value = 0.0331
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.65788264 -0.03466696
## sample estimates:
##       cor
## -0.3900645
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$est and df$est_pa
## t = -1.5476, df = 28, p-value = 0.1329
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.58211099  0.08850857
```

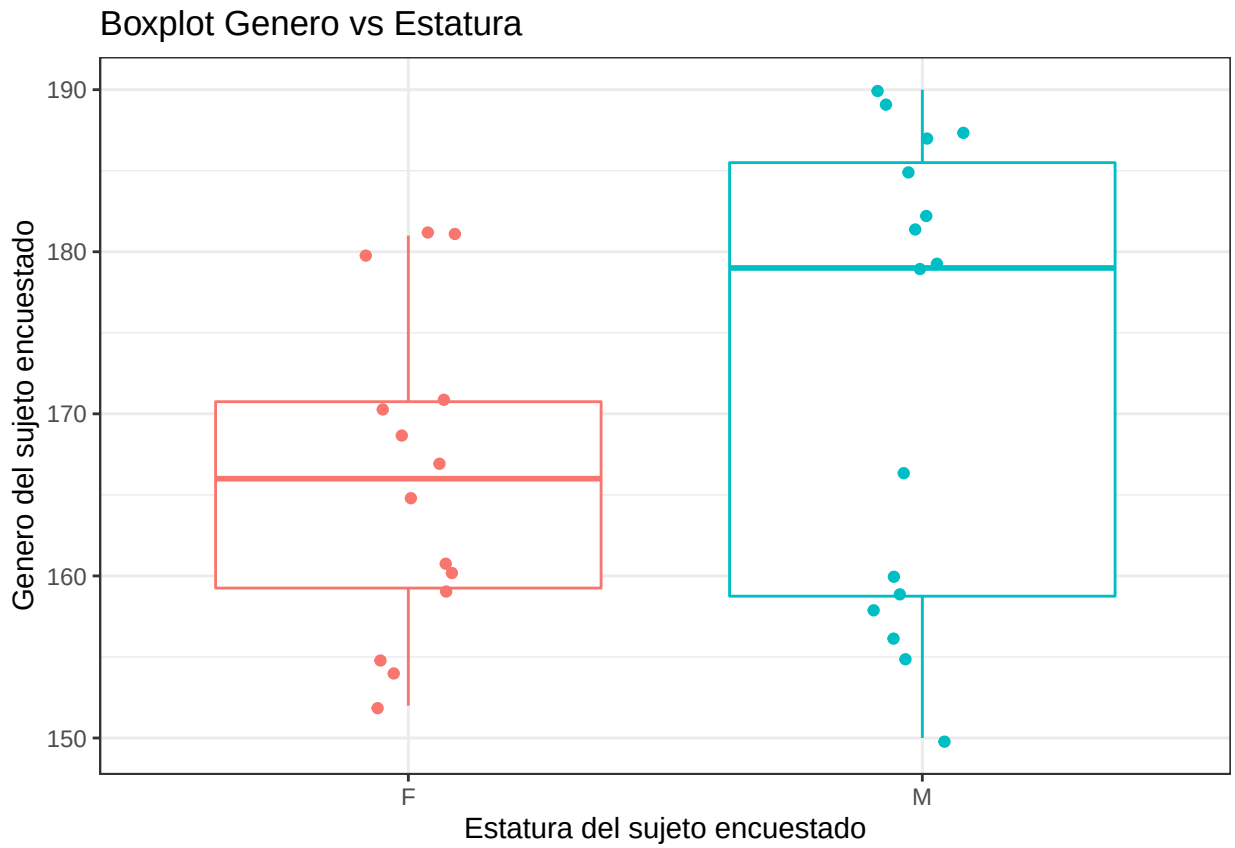
```
## sample estimates:
##        cor
## -0.2807117
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$est_pa and df$est_ma
## t = -0.44376, df = 28, p-value = 0.6606
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4308663  0.2852892
## sample estimates:
##         cor
## -0.08356935
```

Como podemos ver nuestras variables estan correlacionada negativamente, además tenemos que no superan el umbral de ± 0.5 lo que conlleva a que tenga una correlación debil.

Si miramos sus $p - value$ podemos encontrar que solamente para el caso de las madres se rechaza ña hipotesís H_0 donde se asume que la correlación entre este tipo de variables pude ser igual a cero, es decir, que de alguna forma los datos estan correlacionados negativamente entre la estatura del sujeto y de la madre.

Vamos a aplicar los tres métodos a nuestros modelos para cómo funciona cada uno de ellos



El análisis gráfico y de correlación muestran una relación lineal significativa entre la variable `est` y `est_ma`. La variable genero parece influir de forma significativa en la estaura. Ambas variables pueden ser buenos predictores en un modelo lineal múltiple para la variable dependiente `est`.

Esto confirma que nos quedaremos con el `modelo2` por ende a este trabajaremos los datos.

Volvemos a refrescar un poco la memoria con el **modelo2**:

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 278.2976107 53.7712777  5.175581 1.900584e-05
## est_ma       -0.7020405  0.3357947 -2.090684 4.610222e-02
## generoM       5.4877921  4.3412792  1.264096 2.170003e-01
```

se nota puede notar viendo el valor $Pr(|t|)$ que la variable genero es la unica que se rechaza, dentro de la hipotesís que el $\beta_i = 0$, es decir, nuestr modelo solamente se complementaria de los datos de la estatura de la madre.

pero observando el $R^2 = 0.1402$ podemos ver que solamente el kodelo explica el 14.02% de la variabilidad observada en la estatura de los sujetos; nuestro modelo junto con el valor $p - value = 0.04957 \approx 0.05$ nos demuestra que nuestros datos no son significativos.

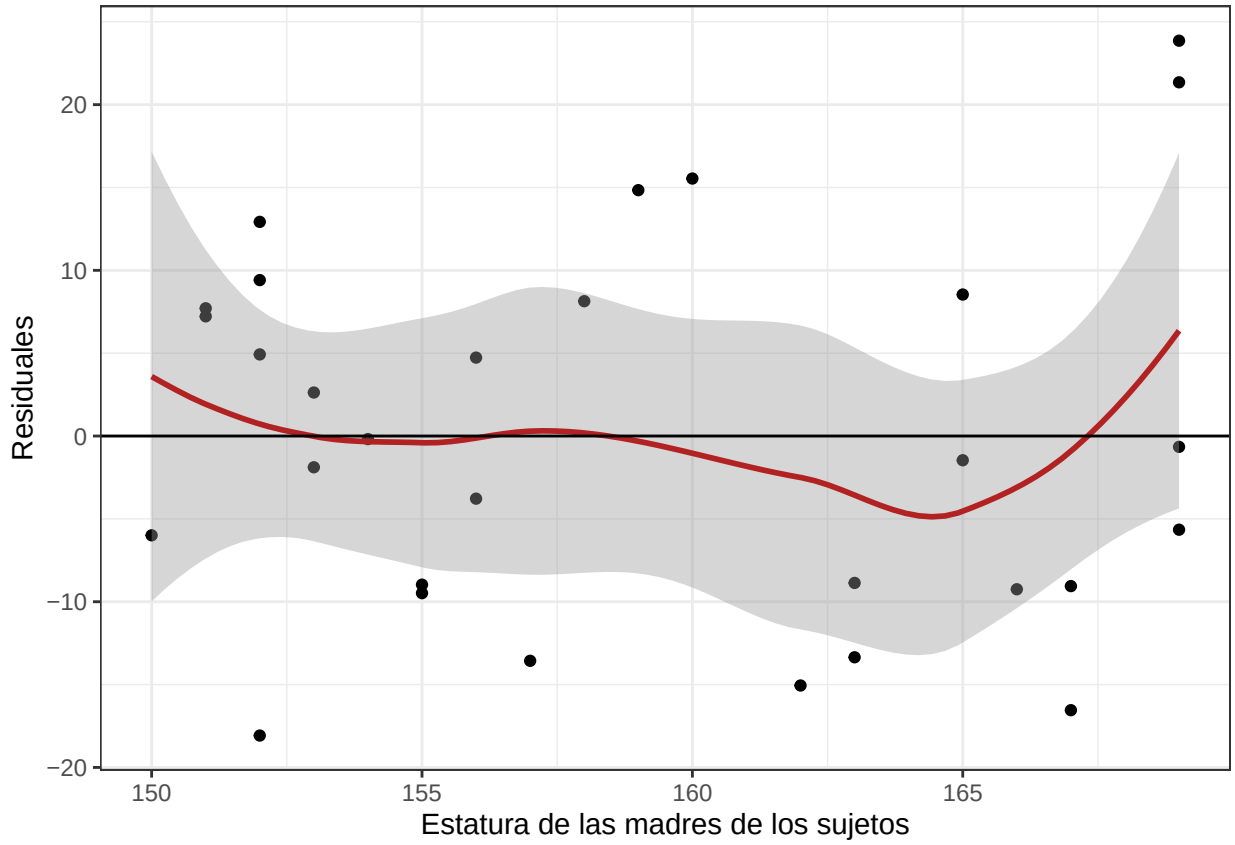
10 Elección de los predictores

En este caso, al solo haber dos predictores, a partir del summary del modelo se identifica que solo la variable `est_ma` es importante.

```
##
## Call:
## lm(formula = est ~ est_ma + genero, data = stature)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.075  -9.035  -1.057   8.030  23.859
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 278.2976    53.7713   5.176  1.9e-05 ***
## est_ma       -0.7020     0.3358  -2.091   0.0461 *
## generoM       5.4878     4.3413   1.264   0.2170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.77 on 27 degrees of freedom
## Multiple R-squared:  0.1995, Adjusted R-squared:  0.1402
## F-statistic: 3.365 on 2 and 27 DF,  p-value: 0.04957
```

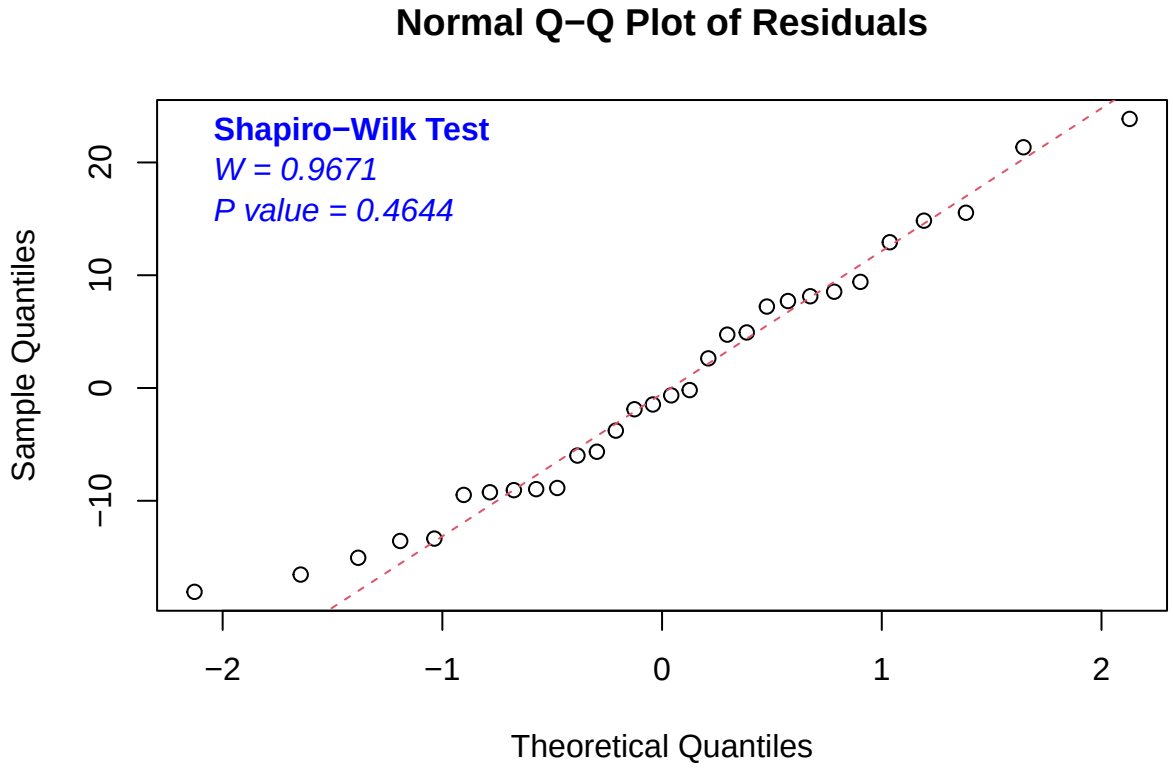
11 Condiciones para la regresión múltiple lineal

11.1 Relación lineal entre los predictores numéricos y la variable dependiente:



Se satisface la condición de linealidad. Se aprecian posibles datos atípicos.

11.2 Distribución normal de los residuos:



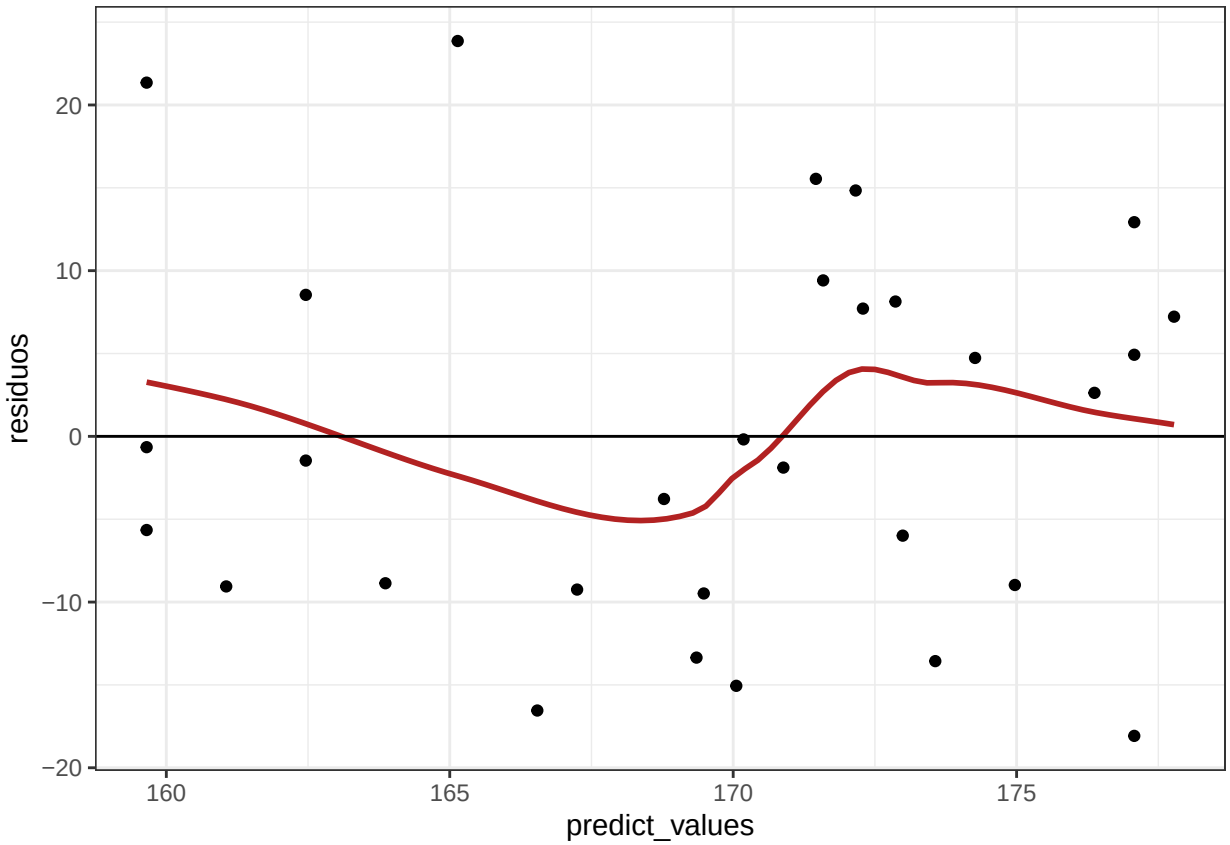
apesar de los datos atipicos se observa que cumple con la normalidad de los residuos.

Nos surge una duda, *¿estos datos atípicos estarán influenciando nuestros datos de alguna manera?*

11.3 Variabilidad constante de los residuos:

```
ggplot(data = data.frame(predict_values = predict(modelo2),
                                residuos = residuals(modelo2)),
             aes(x = predict_values, y = residuos))+
         geom_point()+
         geom_smooth(color = "firebrick", se = FALSE)+
         geom_hline(yintercept = 0)+
         theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
bptest(modelo2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo2
## BP = 8.9341, df = 2, p-value = 0.01148
```

Como podemos ver nuestros ε_i no tienen un varianza constante, apra ello miraremos estos datos atípicos y si hay necesidad de transformar los datos o aplicar un modelo donde las varianzas no sean constantes.

Nota: No multicolinialidad: Dado que solo hay un predictor cuantitativo no se puede dar colinialidad.

11.4 Autocorrelación:

```
dwt(modelo2,alternative = "two.sided")
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.03978963      2.063677   0.844
##  Alternative hypothesis: rho != 0
```

No hay evidencia de autocorrelación

11.5 Identificación de posibles valores atípicos o influyentes

```
outlierTest(modelo2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
```

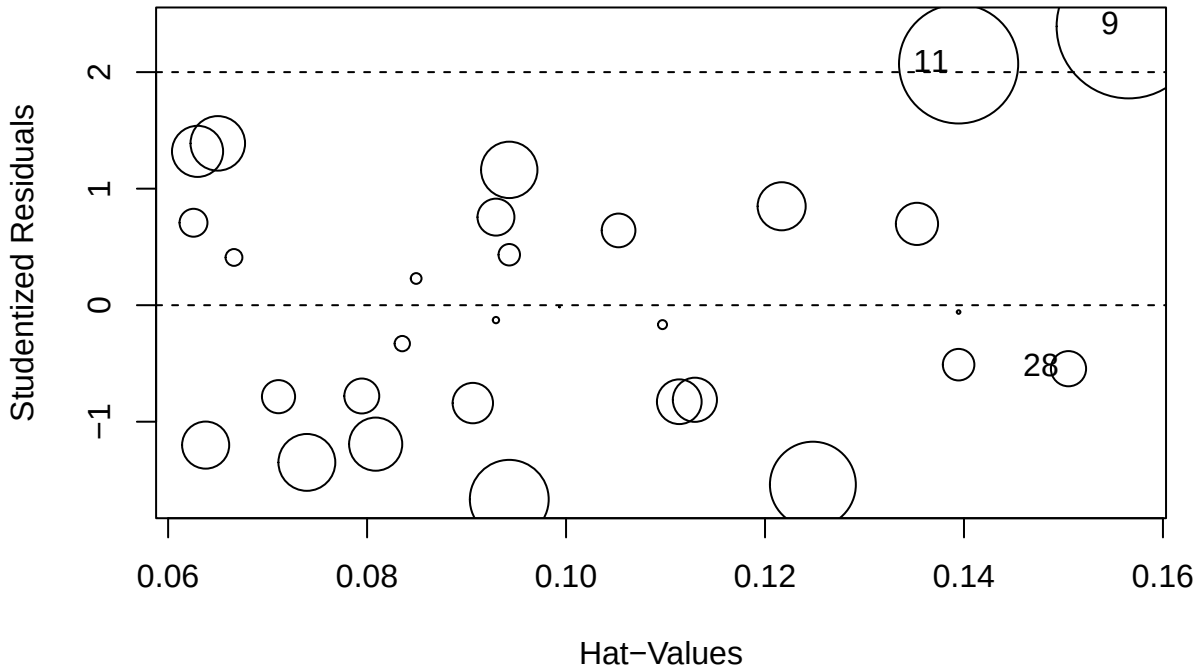
```
## Largest |rstudent|:
##   rstudent unadjusted p-value Bonferroni p
## 9 2.392447           0.024254      0.72762
```

Tal como se apreció en el estudio de normalidad de los residuos, la observación 13 tiene un residuo estandarizado >2 (más de 2 veces la desviación estándar de los residuos) por lo que se considera un dato atípico. El siguiente paso es determinar si es influyente.

```
summary(influence.measures(modelo2))
```

```
## Potentially influential observations of
##   lm(formula = est ~ est_ma + genero, data = stature) :
##
##   dfb.1_ dfb.est_ dfb.gnrM dffit   cov.r cook.d hat
## 9 -0.80   0.80     0.54     1.03_*  0.73  0.30   0.16
```

```
influencePlot(modelo2)
```



	StudRes	Hat	CookD
9	2.3924473	0.1565516	0.3013984
11	2.0706534	0.1394605	0.2064776
28	-0.5450432	0.1505058	0.0180131

El análisis muestran varias observaciones influyentes aunque ninguna excede la distancia de Cook > 1, pero como lo vimos antes, el dato en nuestra modelo tiene influencia si $Cook > \frac{4}{n}$, donde $n = 30$, es decir, $Cook > 0.1333$; nuestros datos influenciadores y que superan, pero al no superar las distnacias del HAT, en fin. no hay datos atípicos en los datos, porque que se llega a la conclusión que este modelo no sirve para predecir o hacer inferencia sobre los datos de la madre y su hijo.

12 Apendice:

12.1 listas de Figuras:

1. Correlation plot.

2. Residuales vs Est_Madre.
3. QQnorm norm Residuales.
4. Predict vs Residuos
5. Infuence plot.

12.2 Codigo:

```
## ----message=FALSE, warning=FALSE, include=FALSE-------------------------
library(readr)
library(tidyverse)
library(kableExtra)
library(magrittr)
library(ggExtra)
library(GGally)
library(janitor)
library(tidystats)
library(car)
library(faraway)
library(lmtest)
library(graphics)
# lectura de la base de datos:
stature <- read_csv("estaturas.csv")


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
kable(rbind(head(stature, n = 5),rep(".", ncol(stature)),
            rep(".", ncol(stature)),rep(".", ncol(stature)),
            tail(stature, n = 5)),digits = 30, align = "c")


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
stature$genero %<>% as.factor()
stature$exp_sujeto %<>% as.Date(format="%d/%m/%Y")
stature$exp_padre %<>% as.Date(format="%d/%m/%Y")
stature$exp_madre %<>% as.Date(format="%d/%m/%Y")
stature$ced_madre %<>%  as.character()
stature %>% ggpairs(.,columns=c(1,2,4,5), aes(color=genero,alpha=0.5))


## ----message=FALSE, warning=FALSE, include=FALSE-------------------------
colnames(stature) <- c("est","genero","exp","est_ma","est_pa","ced",
                       "ced_ma","ced_pa","exp_pa","exp_ma")


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
modelo.m <- lm(est ~ est_pa + est_ma + genero, data = stature)
summary(modelo.m)


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
modelo2 <- lm(est~est_ma+genero, data=stature)
modelo3 <- lm(est~est_pa+genero, data=stature)
modelo4 <- lm(est~est_ma, data=stature)
```

```
modelo5 <- lm(est~est_pa, data=stature)
modelo6 <- lm(est~genero, data=stature)

summary(modelo2)
summary(modelo3)
summary(modelo4)
summary(modelo5)
summary(modelo6)


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
anova(modelo2,modelo3)
anova(modelo2,modelo4)
anova(modelo2,modelo5)
anova(modelo2,modelo6)
print("Mejor modelo del 2 al 6 con menor RSS")
anova(modelo2,modelo.m)


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
# Eliminamos la variables que no nos interesan en el modelo:
df <- stature[, c(1,2,4,5)]
summary(df)


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
cor.test(df$est, df$est_ma)
cor.test(df$est, df$est_pa)
cor.test(df$est_pa, df$est_ma)


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
df %>% ggplot(., aes(x=genero, y=est, color=genero))+
  geom_boxplot()+
  geom_jitter(width = 0.1)+
  theme_bw()+theme(legend.position = "none")+
  labs(x="Estatura del sujeto encuestado",
       y="Genero del sujeto encuestado",
       title = "Boxplot Genero vs Estatura")


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
summary(modelo2)$coefficients


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
summary(modelo2)


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
df %>% ggplot(.,aes(x=est_ma, y=modelo2$residuals))+
  geom_point()+
  geom_smooth(color="firebrick")+
  geom_hline(yintercept = 0)+
```

```
  theme_bw()+
  labs(x="Estatura de las madres de los sujetos",
       y="Residuales")


## ----echo=FALSE, message=FALSE, warning=FALSE-----------------------------
myQQnorm <- function(modelo, student = F, ...){
  if(student){
    res <- rstandard(modelo)
    lab.plot <- "Normal Q-Q Plot of Studentized Residuals"
  } else {
    res <- residuals(modelo)
    lab.plot <- "Normal Q-Q Plot of Residuals"
  }
  shapiro <- shapiro.test(res)
  shapvalue <- ifelse(shapiro$p.value < 0.001, "P value < 0.001", paste("P value = ", round(shapiro$p.va
  shapstat <- paste("W = ", round(shapiro$statistic, 4), sep = "")
  q <- qqnorm(res, plot.it = FALSE)
  qqnorm(res, main = lab.plot, ...)
  qqline(res, lty = 2, col = 2)
  text(min(q$x, na.rm = TRUE), max(q$y, na.rm = TRUE)*0.95, pos = 4, 'Shapiro-Wilk Test', col = "blue",
  text(min(q$x, na.rm = TRUE), max(q$y, na.rm = TRUE)*0.80, pos = 4, shapstat, col = "blue", font = 3)
  text(min(q$x, na.rm = TRUE), max(q$y, na.rm = TRUE)*0.65, pos = 4, shapvalue, col = "blue", font = 3)
}

modelo2 %>% myQQnorm()


## -------------------------------------------------------------------------
ggplot(data = data.frame(predict_values = predict(modelo2),
                                residuos = residuals(modelo2)),
              aes(x = predict_values, y = residuos))+
         geom_point()+
         geom_smooth(color = "firebrick", se = FALSE)+
         geom_hline(yintercept = 0)+
         theme_bw()

bptest(modelo2)


## -------------------------------------------------------------------------
dwt(modelo2,alternative = "two.sided")


## -------------------------------------------------------------------------
outlierTest(modelo2)


## -------------------------------------------------------------------------
summary(influence.measures(modelo2))
influencePlot(modelo2)
```

13 Referencias:

[**1**] Coleman, D. E., & Montgomery, D. C. (1993). A Systematic Approach to Planning for a Designed Industrial Experiment. Technometrics, 35(1), 1–12. https://doi.org/10.2307/1269280

[**2**] The jamovi project (2021). jamovi. (Version 2.2) [Computer Software]. Retrieved from https://www.jamovi.org.

[**3**] R Core Team (2021). R: A Language and environment for statistical computing. (Version 4.0) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2021-04-01).

[**4**] Fox, J., & Weisberg, S. (2020). car: Companion to Applied Regression. [R package]. Retrieved from https://cran.r-project.org/package=car.

[**5**] Ali S. Hadi, S. C. &. (2006). Linear models with r (4th edition.). John Wiley & Sons. Retrieved from http://samples.sainsburysebooks.co.uk/9780470055458_sample_381725.pdf

[**6**] Ferrari, D., & Head, T. (2010). Regression in r. part i: Simple linear regression. UCLA Department of Statistics Statistical Consulting Center. Retrieved October 13, 2014, from http://scc.stat.ucla.edu/page_attachments/0000/0139/reg_1.pdf

[**7**] Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using r (1st edition.). Sage Publications Ltd.

[**8**] J.Faraway, J. (2009). Linear models with r (1st edition.). Taylor & Francis e-Library. Retrieved from http://home.ufam.edu.br/jcardoso/PPGMAT537/Linear%20Models%20with%20R.pdf

[**9**] Kabacoff, R. (2014). Creating a figure arrangement with fine control. Retrieved October 13, 2014, from http://www.statmethods.net/advgraphs/layout.html

[**10**] Pérez, J. L. (2014). La estadística: Una orqueta hecha instrumento. Retrieved October 13, 2014, from http://estadisticaorquestainstrumento.wordpress.com/

[**11**] Sánchez, J. G. P. (2011). Regresión lineal simple. Universidad Politécnica de Madrid. Retrieved October 13, 2014, from http://ocw.upm.es/estadistica-e-investigacion-operativa/introduccion-a-la-estadistica-basica-el-diseno-de-experimentos-y-la-regresion-lineal/contenidos/Material-de-clase/Regresion.pdf

[**12**] (SCG), S. S. C. G. (2013). Multiple linear regression (r). San Diego State University. Retrieved October 13, 2014, from http://scg.sdsu.edu/mlr-r/

[**13**] SPSS. (2007). Análisis de regresión lineal: El procedimiento regresión lineal. IBM SPSS Statistics. Retrieved October 13, 2014, from http://pendientedemigracion.ucm.es/info/socivmyt/paginas/D_departamento/materiales/analisis_datosyMultivariable/18reglin_SPSS.pdf