Examen Final de Modelos Lineales 1

Pontificia Universidad Católica del Perú

Justo Andrés Manrique Urbina*

11 de Julio de 2018

El presente informe contiene las respuestas del alumno respecto al examen final de Modelos Lineales 1.

```
# Carga de librerías
library(formatR)
library(robustbase)
library(car)
library(stargazer)
library(MASS)
library(agricolae)
library(multcomp)
par(mfrow=c(2,2))
```

1. Pregunta 1

Ver al final del informe.

2. Pregunta 2

Ver al final del informe.

3. Pregunta 3

Se procede con la carga de los datos en R y renombre de las variables para facilitar el análisis.

```
rm(list = ls())
# Paso 1: Carga de datos
swiss_1 <- data.frame(swiss)

# Paso 2: Renombre de columnas para un análisis más fácil
names(swiss_1)[1] <- "Y"  #Variable Fertility a Y.
names(swiss_1)[2] <- "A1"  #Variable Agriculture a A1.
names(swiss_1)[3] <- "A2"  #Variable Examination a A2.
names(swiss_1)[4] <- "A3"  #Variable Education a A3.
```

*e-mail: ja.manrique@pm.me

```
names(swiss_1)[5] <- "A4"  #Variable Catholic a A4.
names(swiss_1)[6] <- "A5"  #Variable Infant Mortality a A5.
```

Posteriormente, se efectúa un análisis descriptivo de los datos. En este se puede apreciar lo siguiente:

- La base de datos Swiss se compone de 47 observaciones.
- Las variables con mayor variabilidad son las variables A3 y A4, puesto que su desviación estándar es similar a su media (esto daría un coeficiente de variación cercano a 1). Posteriormente, le siguen las variables A2, A1, A5 e Y en ese orden.

Ver cuadro a continuación:

```
stargazer(title ="Análisis descriptivo de datos: Swiss",swiss_1,
iqr=FALSE,flip = TRUE)
```

Cuadro 1: Análisis descriptivo de datos: Swiss

Statistic	Y	A1	A2	A3	A4	A5
N	47	47	47	47	47	47
Mean	70.143	50.660	16.489	10.979	41.144	19.943
St. Dev.	12.492	22.711	7.978	9.615	41.705	2.913
Min	35.000	1	3	1	2.150	10.800
Pctl(25)	64.700	35.9	12	6	5.195	18.150
Pctl(75)	78.450	67.7	22	12	93.125	21.700
Max	92.500	90	37	53	100.000	26.600

Con el propósito de evaluar, en primera instancia, posibles indicios de multicolinealidad entre variables, así como identificar las relaciones de las covariables con la variable respuesta, se generó una matriz de correlación con todas las variables contenidas en la base de datos. En esta matriz, se puede apreciar lo siguiente:

- La variable A1 tiene una correlación negativa medianamente fuerte con las variables A2 y A3. Por otro lado, dicha variable tiene una correlación positiva mediana con la variable A4.
- La variable A2 tiene una correlación positiva medianamente fuerte con la variable A3. Por otro lado, dicha variable tiene una correlación negativa medianamente fuerte con la variable A4.

La matriz de correlaciones se puede observar en el Cuadro 2, conforme el siguiente código:

```
correlation.matrix <- cor(swiss_1[,c("Y","A1","A2",
"A3","A4","A5")])
stargazer(correlation.matrix, title="Matriz de correlaciones")
```

Cuadro 2: Matriz de correlaciones

	Y	A1	A2	A3	A4	A5
Y	1	0,353	-0,646	-0,664	0,464	0,417
A1	0,353	1	-0,687	-0,640	0,401	-0,061
A2	-0,646	-0,687	1	0,698	-0,573	-0,114
A3	-0,664	-0,640	0,698	1	-0,154	-0,099
A4	0,464	0,401	-0,573	-0,154	1	0,175
A5	0,417	-0,061	-0,114	-0,099	0,175	1

Posteriormente, se efectuó una regresión lineal mediante mínimos cuadrados ordinarios, tomando en consideración lo indicado anteriormente. En dicha regresión se observa que:

- La variable A2 no es estadísticamente significativa (su p-valor es mayor a 0.05). Por ello, procederemos a descartarla y generar un segundo modelo de regresión lineal.

Ambos modelos de regresión se pueden observar en el Cuadro 3 al final de esta sección, conforme el siguiente código:

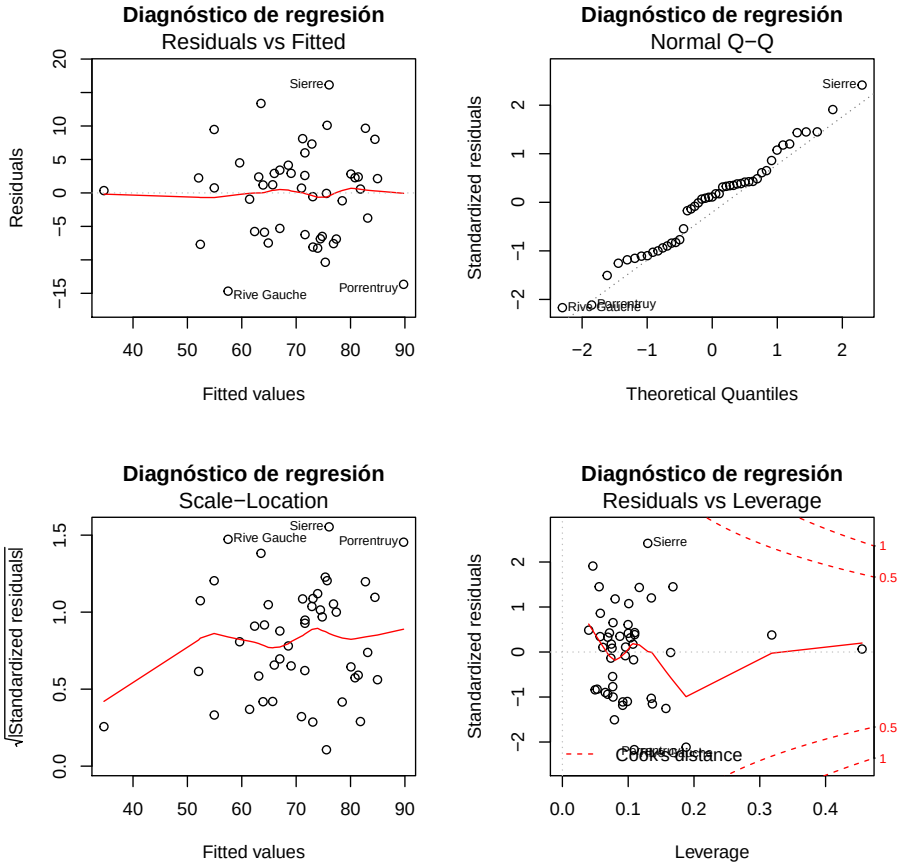
```
lm1_swiss <- lm(formula = Y ~ A1+A2+A3+A4+A5,
data=swiss_1)
lm2_swiss <- lm(formula = Y ~ A1+A3+A4+A5,
data=swiss_1)
stargazer(lm1_swiss,lm2_swiss, title="Regresión Lineal")
```

Cuadro 3: Regresión Lineal

	Dependent variable:	
	Y	
	(1)	(2)
A1	-0.172^{**}	-0.155^{**}
	(0.070)	(0.068)
A2	-0.258	
	(0.254)	
A3	-0.871^{***}	-0.980^{***}
	(0.183)	(0.148)
A4	0.104^{***}	0.125^{***}
	(0.035)	(0.029)
A5	1.077^{***}	1.078^{***}
	(0.382)	(0.382)
Constant	66.915^{***}	62.101^{***}
	(10.706)	(9.605)
Observations	47	47
R^2	0.707	0.699
Adjusted R^2	0.671	0.671
Residual Std. Error	7.165 (df = 41)	7.168 (df = 42)
F Statistic	19.761^{***} (df = 5; 41)	24.424^{***} (df = 4; 42)
Note:		*p<0.1; **p<0.05; ***p<0.01

Una vez definido el modelo a utilizar, se procede a evaluar los supuestos del modelo mediante los siguientes gráficos:

```
par(mfrow=c(2,2))
plot(lm2_swiss,main = "Diagnóstico de regresión")
```



En estos se aprecia lo siguiente:

- La gráfica Residuals vs Fitted"tiene como objetivo identificar si los residuales tienen un comportamiento no lineal. En esta, no se observa una relación no lineal entre los residuales; por lo tanto, se puede inferir no existe una relación no lineal que habría que modelar.
- La gráfica "Normal Q-Q"permite identificar si los residuos están normalmente distribuidos. En el caso los residuos no sigan, en general, una línea recta, sería un indicador que los errores no están distribuidos normalmente. En base al cuadro presentado, pueden existir indicios de no-normalidad. Esto se pondrá a prueba a través del test Shapiro-Wilks.
- La gráfica "Scale-Location"permite identificar si los errores son homocedásticos o no. En ese sentido, si hubiere algún grado de linealidad en esta gráfica, ello implicaría que los errores no tienen varianza constante

por lo que no serían homocedásticos. En base al cuadro presentado, un grado de linealidad en la medida que el valor predicho aumenta; por lo tanto, se podría concluir que los errores son heterocedásticos.

- La gráfica Residuals vs Leverage"permite identificar puntos aberrantes dentro de la base de datos. En base al cuadro presentado, la observación "Sierra"tiene una distancia de Cooks mayor que las otras observaciones. Sin embargo, no supera los umbrales como para tener un impacto significativo.

Finalmente, se efectuaron tests estadísticos de homocedasticidad y normalidad de los errores a fin de validar los supuestos de la regresión lineal múltiple bajo mínimos cuadrados ordinarios. Ver resultados a continuación:

- Prueba de Normalidad de Errores (Shapiro-Wilks)
  - El test de Shapiro-Wilks está orientado a identificar si los errores siguen una distribución normal, a fin de validar los supuestos de la regresión. La hipótesis nula de dicho test es que una determinada muestra proviene de una población normalmente distribuida (Camiz, 2018). Conforme se aprecia en el cuadro posterior, el p-valor del test es mayor a 0.05 por lo que se no serechaza dicha hipótesis nula. Por lo tanto, los residuos seguirían una distribución normal cumpliendo con el supuesto de la regresión.

```
##
##  Shapiro-Wilk normality test
##
## data:  lm2_swiss$residuals
## W = 0.97657, p-value = 0.459
```

- Prueba de Homocedasticidad de Errores (Prueba de Breusch-Pagan)
  - El test de Breusch-Pagan está orientado a identificar si los errores tienen una varianza homocedástica, a fin de validar los supuestos de la regresión. La hipótesis nula de dicho test es que los errores son homocedásticos (Camiz,2018). Conforme se aprecia en el cuadro posterior, el p-valor del test es menor a 0.05, por lo que no se rechaza la hipótesis nula. Por lo tanto, los residuos tendrían varianza constante cumpliendo con el supuesto de la regresión.

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.4687214    Df = 1     p = 0.493576
```

3.1. Corrección de la multicolinealidad

Con el propósito de corregir la multicolinealidad entre A1, A2 y A3, se propuso crear una variable de interacción entre estas tres, conforme se puede ser en el proceso posterior:

```
lm3_swiss <- lm(formula = Y ~ A1:A2:A3+A4+A5,
data=swiss_1)
summary(lm3_swiss)

##
## Call:
## lm(formula = Y ~ A1:A2:A3 + A4 + A5, data = swiss_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.650  -5.673   1.296   4.538  16.498
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.8722164  9.9379831   4.415  6.7e-05 ***
## A4           0.0893817  0.0351441   2.543  0.01466 *
## A5           1.4487495  0.4848625   2.988  0.00463 **
## A1:A2:A3    -0.0008858  0.0002666  -3.322  0.00183 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.428 on 43 degrees of freedom
## Multiple R-squared:  0.4676,Adjusted R-squared:  0.4304
## F-statistic: 12.59 on 3 and 43 DF,  p-value: 4.858e-06
```

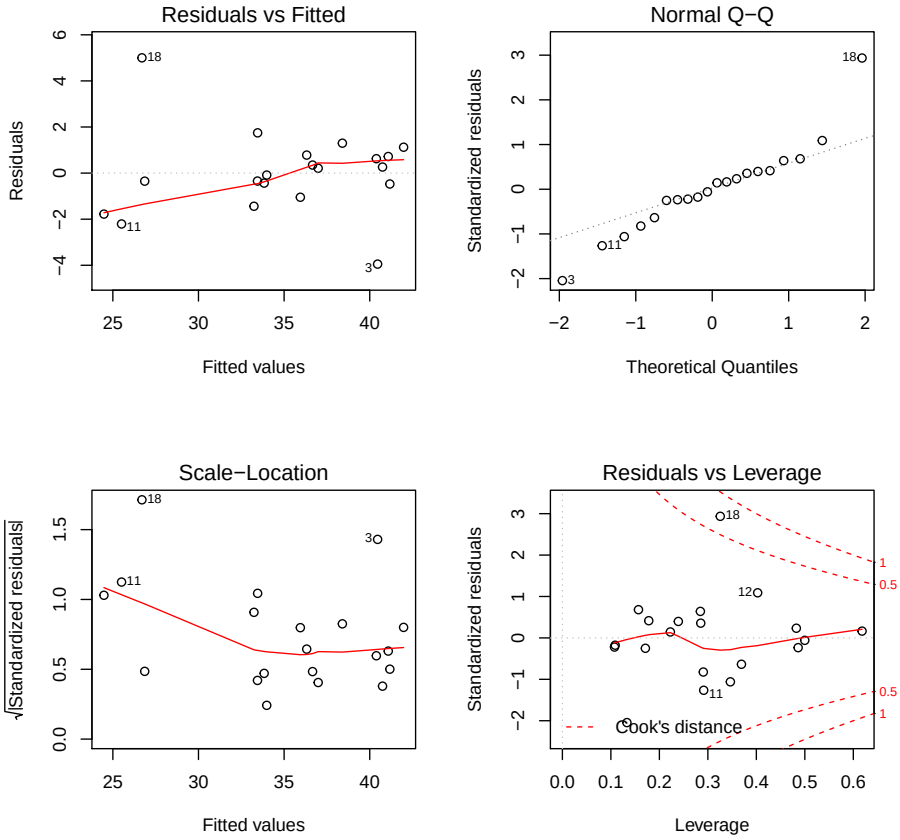
Se observa que la interacción es estadísticamente significativa e impacta negativamente en la variable respuesta «Fertilidad». Asimismo, se observa que las variables A4 y A5 tienen un impacto positivo conforme se aprecia en sus coeficientes.

4. Pregunta 4

4.1. Pregunta 4.a)

Se ajustó un modelo de regresión por mínimos cuadrados ordinarios y se ejecutó la prueba de diagnóstico a través del comando plot() en R. En dicho diagnóstico se prestó atención en la prueba «Residuals vs. Leverage» con el propósito de identificar las observaciones atípicas. Se observó que la observación 18 es una observación atípica (supera el umbral de 0.5). Ver gráfico a continuación:

```
#### Inicio de la pregunta 4 ####
rm(list = ls())
data(coleman)
par(mfrow = c(2, 2))
fit <- lm(formula = Y ~ salaryP + fatherWc + sstatus + teacherSc + motherLev,
    data = coleman)
plot(fit)
```



Posteriormente, se generó un nuevo modelo de regresión eliminando dicha observación, y se observa que:

- El nuevo modelo de regresión tiene como significativas a todas las variables (exceptuando la variable «salaryP»), mientras que el primer modelo de regresión solo mantiene como significativos a las variables «sstatus» y «teacherSC».
- El nuevo modelo de regresión tiene un R ajustado de 0.949, mientras que el anterior de 0.873.
- La variable «motherLev» aumentó su coeficiente, de -1.811 a -4.571 con el nuevo modelo de regresión.
- La constante del nuevo modelo aumentó 34.287, de 19.949 del modelo anterior.

En conclusión, la eliminación de la observación atípica cambió drásticamente la significancia de las variables así como sus coeficientes. Ver código en R dónde se observa la comparación de ambos modelos.

```r
fit_2 <- lm(formula = Y ~ salaryP + fatherWc + sstatus + teacherSc + motherLev,
    data = coleman[-c(18), ])
par(mfrow = c(2, 2))
stargazer(title = "Regresión Lineal", fit, fit_2)
```

Cuadro 4: Regresión Lineal

	Dependent variable:	
	Y	
	(1)	(2)
salaryP	−1.793	−1.617*
	(1.233)	(0.794)
fatherWc	0.044	0.085**
	(0.053)	(0.035)
sstatus	0.556***	0.674***
	(0.093)	(0.065)
teacherSc	1.110**	1.110***
	(0.434)	(0.279)
motherLev	−1.811	−4.571***
	(2.027)	(1.437)
Constant	19.949	34.287***
	(13.628)	(9.312)
Observations	20	19
R^2	0.906	0.963
Adjusted R^2	0.873	0.949
Residual Std. Error	2.074 (df = 14)	1.334 (df = 13)
F Statistic	27.085*** (df = 5; 14)	68.274*** (df = 5; 13)
Note:		*p<0.1; **p<0.05; ***p<0.01

4.2. Pregunta 4.b)

Se procedió con la generación de la regresión robusta, ver código generado en R:

```
rob.fit <- rlm(formula = Y ~ salaryP + fatherWc + sstatus + teacherSc + motherLev,
    data = coleman)
stargazer(title = "Regresión Lineal", fit, fit_2, rob.fit)
```

Cuadro 5: Regresión Lineal

	Dependent variable:		
	Y		
	OLS		*robust linear*
	(1)	(2)	(3)
salaryP	−1.793	−1.617*	−1.621**
	(1.233)	(0.794)	(0.695)
fatherWc	0.044	0.085**	0.075**
	(0.053)	(0.035)	(0.030)
sstatus	0.556***	0.674***	0.640***
	(0.093)	(0.065)	(0.052)
teacherSc	1.110**	1.110***	1.156***
	(0.434)	(0.279)	(0.244)
motherLev	−1.811	−4.571***	−3.520***
	(2.027)	(1.437)	(1.143)
Constant	19.949	34.287***	27.350***
	(13.628)	(9.312)	(7.681)
Observations	20	19	20
R^2	0.906	0.963	
Adjusted R^2	0.873	0.949	
Residual Std. Error	2.074 (df = 14)	1.334 (df = 13)	0.746 (df = 14)
F Statistic	27.085*** (df = 5; 14)	68.274*** (df = 5; 13)	
Note:			*p<0.1; **p<0.05; ***p<0.01

Se puede observar que:

- Las variables contenidas en la regresión robusta, la cual tiene 20 observaciones, son estadísticamente significativas a pesar de la existencia de una variable atípica. Dicho modelo es similar, respecto a la significancia de sus

variables, con el modelo lineal bajo mínimos cuadrados ordinarios generado con solo las 19 observaciones (es decir, dónde se elimina la variable atípica).

- Respecto a los coeficientes, la regresión robusta tiene mayor similitud con el modelo lineal bajo mínimos cuadrados ordinarios que solo contiene 19 observaciones. Sin embargo, la regresión robusta presenta diferencias por los coeficientes de las variables «motherLev» y la constante (-3.520 a -4.571 y 27.350 a 34.287 respectivamente).

5. Pregunta 5

5.1. Pregunta 5.a)

- El factor de bloque consiste en la variable Día, puesto que en esta se prueban cada uno de los silos. Asimismo, es imposible aleatorizar la variable Día.
- El factor de tratamiento es compuesto por la variable Silos.

5.2. Pregunta 5.b)

Las hipótesis estadísticas se encuentran al final del informe.
Ver a continuación el modelo estadístico:

```
rm(list=ls())
silos <- read.table("D:/Justo Andrés/Dropbox/Maestría en Estadística/2018 - 1/Modelos Lineales 1/Tra
mod_silos<-lm(temperatura~Silo+as.factor(Dia),silos)
anova_silos<-anova(mod_silos)
anova_silos

## Analysis of Variance Table
##
## Response: temperatura
##                Df Sum Sq Mean Sq F value Pr(>F)
## Silo            4   4.46   1.115  0.6904 0.6092
## as.factor(Dia)  4   9.76   2.440  1.5108 0.2460
## Residuals      16  25.84   1.615
```

Se observa que ambas covariables no son significativas (su p-valor es mayor al umbral de 0.05).

5.3. Pregunta 5.c)

Para efectuar el análisis de medias, se utilizó el método de mínimas diferencias. Ver a continuación la ejecución del código:

```
mediat<-tapply(silos$temperatura,silos$Silo,mean)

difAB <- abs(mediat[1]-mediat[2])
difAC <- abs(mediat[1]-mediat[3])
difAD <- abs(mediat[1]-mediat[4])
```

```r
difAE <- abs(mediat[1]-mediat[5])
difBC <- abs(mediat[2]-mediat[3])
difBD <- abs(mediat[2]-mediat[4])
difBE <- abs(mediat[2]-mediat[5])
difCD <- abs(mediat[3]-mediat[4])
difCE <- abs(mediat[3]-mediat[5])
difDE <- abs(mediat[4]-mediat[5])

CME <- anova_silos$`Mean Sq`[3]
t<- qt(0.975,anova_silos$Df[3])

LSD <- t*sqrt((2*CME)/4)
vecdif <- c(difAB,difAC,difAD,difAE,difBC,difBD,difBE,difCD,difCE,difDE)
nombres <- c("difAB","difAC","difAD","difAE","difBC",
"difBD","difBE","difCD","difCE","difDE")

for(i in 1:10)
{
  if(vecdif[i]>LSD)
    print(paste(nombres[i],"Significativa"))
  else
    print(paste(nombres[i],"No significativa"))
}
```

```
## [1] "difAB No significativa"
## [1] "difAC No significativa"
## [1] "difAD No significativa"
## [1] "difAE No significativa"
## [1] "difBC No significativa"
## [1] "difBD No significativa"
## [1] "difBE No significativa"
## [1] "difCD No significativa"
## [1] "difCE No significativa"
## [1] "difDE No significativa"
```

Se observa que no existen diferencias significativas. Asimismo, se realizó el análisis de diferencias mediante el método Tukey. Ver a continuación el código:

```r
amod_1 <- aov(temperatura ~ Silo + as.factor(Dia), data = silos)
compmet_1 <- glht(amod_1, linfct = mcp(Silo = "Tukey"))
summary(compmet_1)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = temperatura ~ Silo + as.factor(Dia), data = silos)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## B - A == 0   0.9000     0.8037   1.120    0.794
## C - A == 0   0.1000     0.8037   0.124    1.000
```

```
## D - A == 0   1.0000     0.8037   1.244    0.727
## E - A == 0   0.2000     0.8037   0.249    0.999
## C - B == 0  -0.8000     0.8037  -0.995    0.854
## D - B == 0   0.1000     0.8037   0.124    1.000
## E - B == 0  -0.7000     0.8037  -0.871    0.903
## D - C == 0   0.9000     0.8037   1.120    0.794
## E - C == 0   0.1000     0.8037   0.124    1.000
## E - D == 0  -0.8000     0.8037  -0.995    0.854
## (Adjusted p values reported -- single-step method)
```

Se observa que no existen diferencias significativas.

6. Pregunta 6

6.1. Pregunta 6.a)

Respuesta al final del Informe.

6.2. Pregunta 6.b)

Ver a continuación la ejecución del código R:

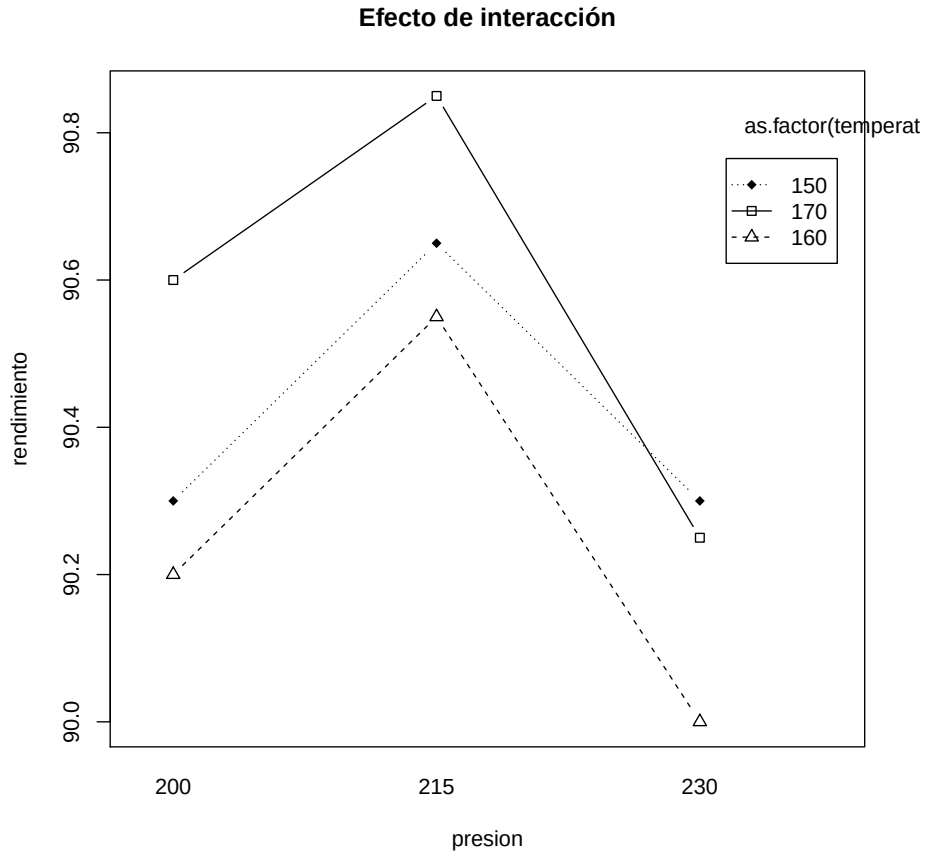
```
rm(list = ls())
rdmto <- read.table("D:/Justo Andrés/Dropbox/Maestría en Estadística/2018 - 1/Modelos Lineales 1/Tra
    header = TRUE)

mod_rdmto <- lm(rendimiento ~ as.factor(presion) + as.factor(temperatura) +
    presion * temperatura, rdmto)

anova_rdmto <- anova(mod_rdmto)
anova_rdmto

## Analysis of Variance Table
##
## Response: rendimiento
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
## as.factor(presion)       2 0.76778 0.38389 27.4797 3.313e-05 ***
## as.factor(temperatura)   2 0.30111 0.15056 10.7771  0.002092 **
## presion:temperatura      1 0.06125 0.06125  4.3844  0.058171 .
## Residuals               12 0.16764 0.01397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se observa, a través del p-valor, que las variables «presion» y «temperatura» son significativas. La interacción de ambas variables está muy próxima a la significancia estadística.

6.3. Pregunta 6.c)

Ver a continuación la ejecución del código R:

```
with(rdmto, (interaction.plot(as.factor(presion), as.factor(temperatura), rendimiento,
    type = "b", pch = c(18, 24, 22), leg.bty = "o", main = "Efecto de interacción",
    xlab = "presion", ylab = "rendimiento")))
```




```
## NULL
```

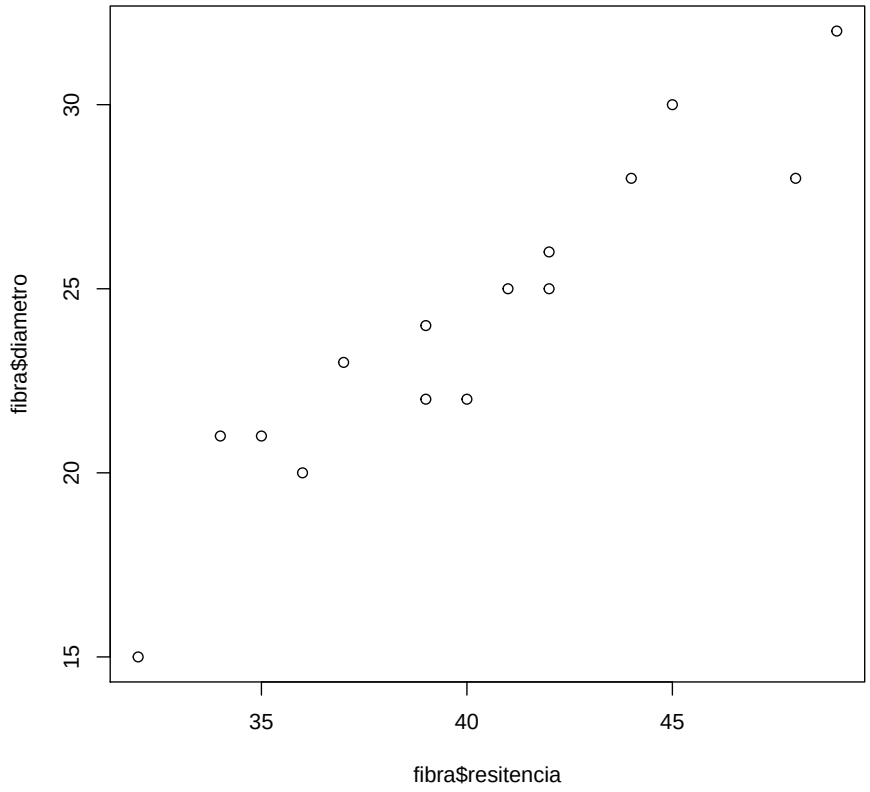
Se observa que, con el propósito de obtener un mayor rendimiento, la presión debe mantenerse en el valor «215» y la temperatura en «170».

7. Pregunta 7

7.1. Pregunta 7.a)

Ver la generación del gráfico mediante R:

```
rm(list = ls())
fibra <- read.table("D:/Justo Andrés/Dropbox/Maestría en Estadística/2018 - 1/Modelos Lineales 1/Tra
    header = TRUE)
plot(fibra$resitencia, fibra$diametro)
```




```
cor(fibra$resitencia, fibra$diametro)

## [1] 0.938542
```

Se observa que existe una relación lineal fuerte entre ambas variables (con un índice de correlación de 0.94). Esto podría deberse a que un mayor diámetro contiene mayor peso, lo cual podría hacerle más resistente.

7.2. Pregunta 7.b)

Las hipótesis estadísticas se encuentran al final del informe.
Ver a continuación el código en R respecto al modelo estadístico:

```
mod <- lm(resitencia ~ diametro + maquina, fibra)
anova_fibra <- Anova(mod, type = "III")
anova_fibra

## Anova Table (Type III tests)
##
## Response: resitencia
##              Sum Sq Df F value    Pr(>F)
```

```
## (Intercept)  87.434  1 34.3664 0.0001089 ***
## diametro    178.014  1 69.9694 4.264e-06 ***
## maquina      13.284  2  2.6106 0.1180839
## Residuals    27.986 11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.3. Pregunta 7.c)

Las máquinas, de acuerdo al p-valor obtenido en el ANCOVA anterior, no influye en la resistencia del monofilamento (su p-valor es mayor a 0.05).