Examen final

Juan Cantero Jimenez

2/21/2022

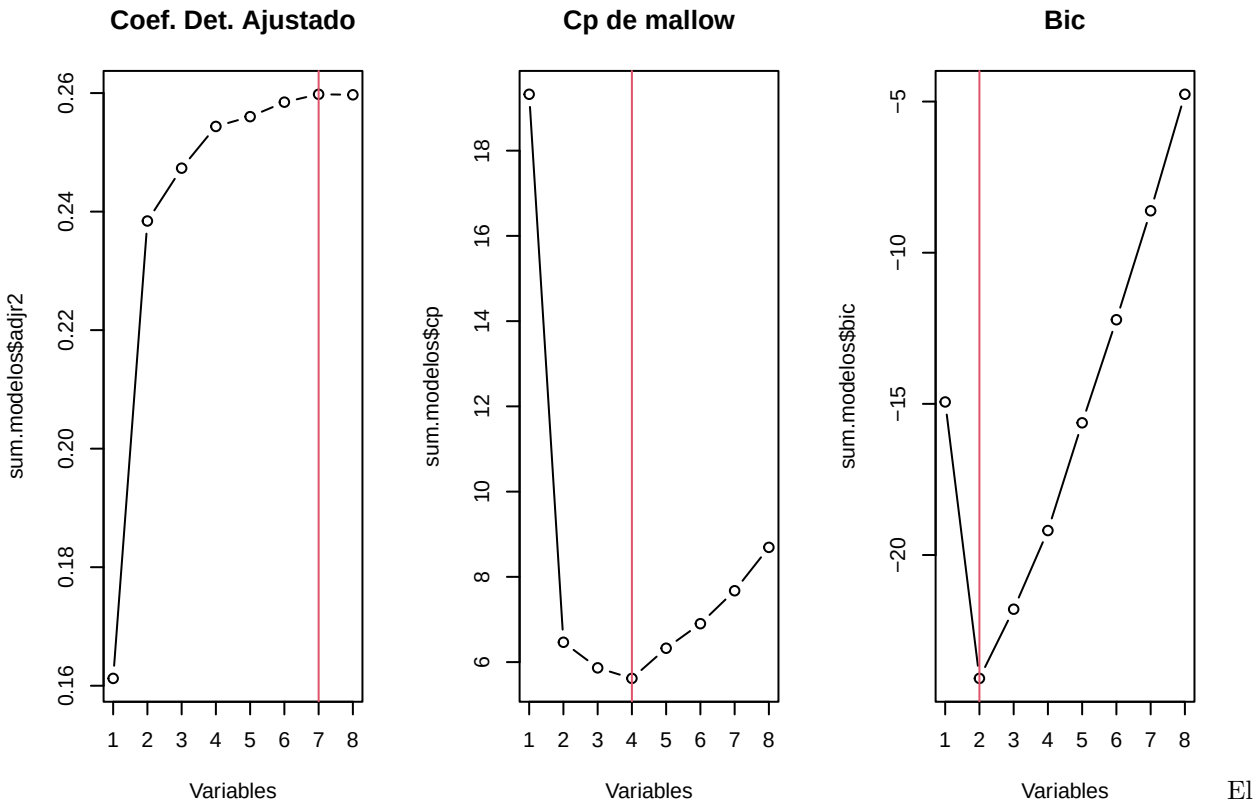
Ejercicio 1

```
load("pesca.Rdata")
str(pesca)
```

```
## 'data.frame':    135 obs. of  8 variables:
##  $ Grupo        : Factor w/ 2 levels "Control","Pescador": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Edad         : int  45 38 24 41 43 58 45 46 46 46 ...
##  $ ResTime      : int  6 13 2 2 11 2 6 0 14 5 ...
##  $ Altura       : int  175 173 168 183 175 176 184 170 175 175 ...
##  $ Peso         : int  70 73 66 80 78 75 85 68 80 75 ...
##  $ PescadoSem   : int  14 7 7 7 21 21 21 7 21 7 ...
##  $ PartesPescado: Factor w/ 4 levels "Ninguna","Carne",..: 3 2 3 2 2 2 2 3 2 2 ...
##  $ Mercurio     : num  4.48 4.79 3.86 11.44 10.85 ...
```

```
modelos <- leaps::regsubsets(Mercurio ~ ., data=pesca)
sum.modelos <-summary(modelos)
```

```
par(mfrow=c(1,3))
plot(1:8, sum.modelos$adjr2, xlab="Variables", main="Coef. Det. Ajustado", type="b")
abline(v = which.max(sum.modelos$adjr2), col=2)
plot(1:8, sum.modelos$cp, xlab="Variables", main="Cp de mallow", type="b")
abline(v = which.min(sum.modelos$cp), col=2)
plot(1:8, sum.modelos$bic, xlab="Variables", main="Bic", type="b")
abline(v = which.min(sum.modelos$bic), col=2)
```



El modelo con mejor ajuste según BIC es el modelo que contiene dos covariables.

Si, el resultado habría cambiado si se hubiera escogido como criterio de selección R2 ajustado o Cp de Mallow. Si se hubiera escogido el primero como criterio de selección, se obtendría como óptimo un modelo con 7 covariables, así si se hubiera escogido el critero de Cp de Mallow se obtendría como óptimo un modelo con 4 covariables. Estos modelos son los siguientes:

```
print("R2 ajustado")
```

```
## [1] "R2 ajustado"
```

```
sum.modelos$which[which.max(sum.modelos$adjr2),]
```

```
##           (Intercept)         GrupoPescador                  Edad
##                  TRUE                  TRUE                  TRUE
##               ResTime               Altura                  Peso
##                  TRUE                 FALSE                  TRUE
##            PescadoSem    PartesPescadoCarne     PartesPescadoTodo
##                  TRUE                  TRUE                 FALSE
## PartesPescadoVísceras
##                  TRUE
```

```
print("Cp de Mallow")
```

```
## [1] "Cp de Mallow"
```

```
sum.modelos$which[which.min(sum.modelos$cp),]
```

```
##           (Intercept)         GrupoPescador                  Edad
##                  TRUE                  TRUE                 FALSE
##               ResTime               Altura                  Peso
##                 FALSE                 FALSE                  TRUE
```

```
##            PescadoSem    PartesPescadoCarne     PartesPescadoTodo
##                  TRUE                  TRUE                 FALSE
## PartesPescadoVísceras
##                 FALSE
```

```
print("BIC")
```

```
## [1] "BIC"
```

```
sum.modelos$which[which.min(sum.modelos$bic),]
```

```
##           (Intercept)         GrupoPescador                  Edad
##                  TRUE                 FALSE                 FALSE
##               ResTime               Altura                  Peso
##                 FALSE                 FALSE                  TRUE
##            PescadoSem    PartesPescadoCarne     PartesPescadoTodo
##                  TRUE                 FALSE                 FALSE
## PartesPescadoVísceras
##                 FALSE
```

Como se puede observar el número total de variables escogidas no es lo único que cambia entre los modelos óptimos para los distintos parámetros, las variables escogidas tambien cambian.

Por último ajustamos el modelo optimo para el parámtro BIC:

```
lm1 <- lm(Mercurio ~ PescadoSem + Peso, data = pesca)
summary(lm1)
```

```
##
## Call:
## lm(formula = Mercurio ~ PescadoSem + Peso, data = pesca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8344 -1.3096 -0.2953  0.6279 11.8572
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.07682    2.44481  -4.122 6.60e-05 ***
## PescadoSem    0.15884    0.04175   3.805 0.000216 ***
## Peso          0.17518    0.03322   5.273 5.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.564 on 132 degrees of freedom
## Multiple R-squared:  0.2498, Adjusted R-squared:  0.2384
## F-statistic: 21.97 on 2 and 132 DF,  p-value: 5.787e-09
```

Ejercicio 2

```
lm2 <- lm(Mercurio ~ PescadoSem + Peso + Grupo:PescadoSem + Grupo:Peso, data=pesca)
summary(lm2)
```

```
##
## Call:
## lm(formula = Mercurio ~ PescadoSem + Peso + Grupo:PescadoSem +
##     Grupo:Peso, data = pesca)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0705 -1.1391 -0.2026  0.6525 11.4266
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -9.26610    2.48392  -3.730 0.000284 ***
## PescadoSem                1.09393    0.60324   1.813 0.072076 .
## Peso                      0.14464    0.03671   3.939 0.000133 ***
## PescadoSem:GrupoPescador -0.99702    0.60622  -1.645 0.102461
## Peso:GrupoPescador        0.02897    0.01248   2.322 0.021797 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.532 on 130 degrees of freedom
## Multiple R-squared:  0.2797, Adjusted R-squared:  0.2575
## F-statistic: 12.62 on 4 and 130 DF,  p-value: 1.053e-08
```

```
anova1 <-anova(lm1, lm2)
anova1
```

```
## Analysis of Variance Table
##
## Model 1: Mercurio ~ PescadoSem + Peso
## Model 2: Mercurio ~ PescadoSem + Peso + Grupo:PescadoSem + Grupo:Peso
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    132 868.07
## 2    130 833.49  2    34.588 2.6974 0.07115 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En base al test ANOVA realizado entre el modelo optimo encontrado en el ejercicio 1 y este con la interacción de grupos, se puede concluir que no existen diferencias significativas entre ambos. Esto indica que el efecto de la interacción es de poca mágnitudad.

```
glm1 <- glm(Mercurio ~ PescadoSem + Peso, data=pesca)
glm2 <- glm(Mercurio ~ PescadoSem + Peso + Grupo:PescadoSem + Grupo:Peso, data=pesca)
loocv1 <- boot::cv.glm(pesca, glm1)
loocv2 <- boot::cv.glm(pesca, glm2)
loocv1$delta[2]
```

```
## [1] 6.916
```

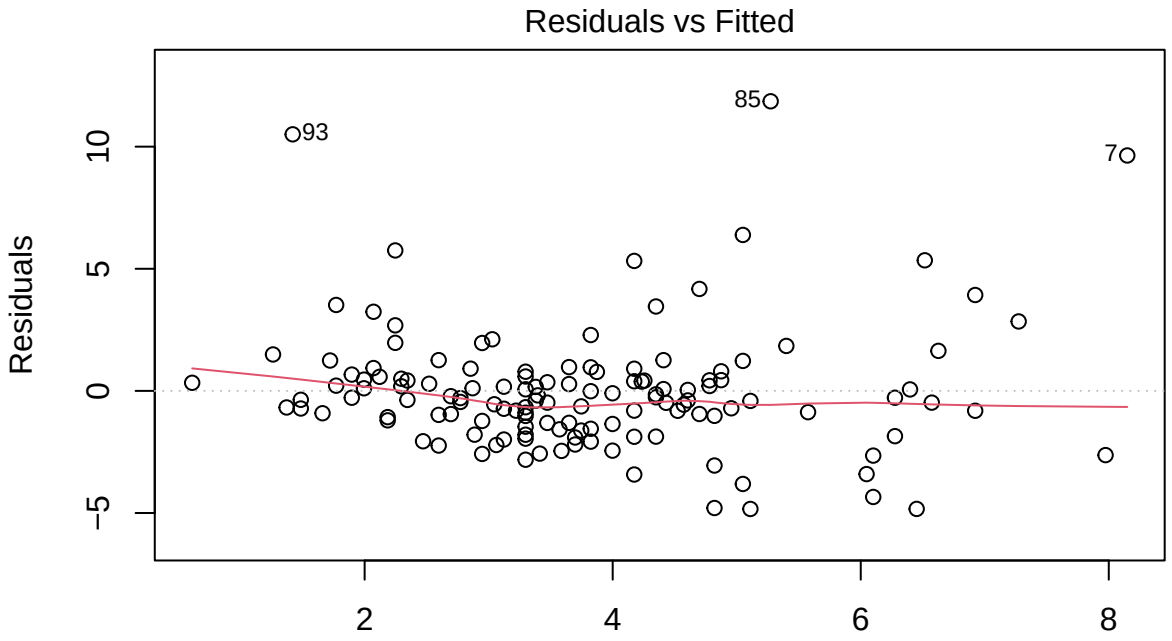
```
loocv2$delta[2]
```

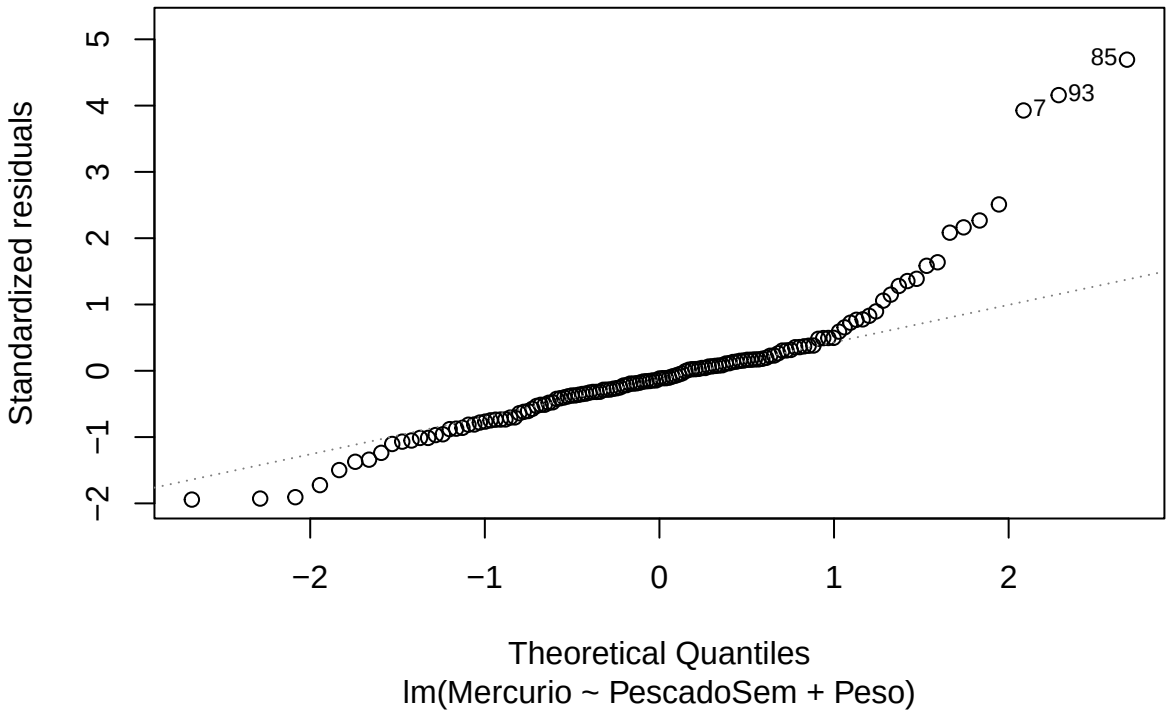
```
## [1] 6.793036
```

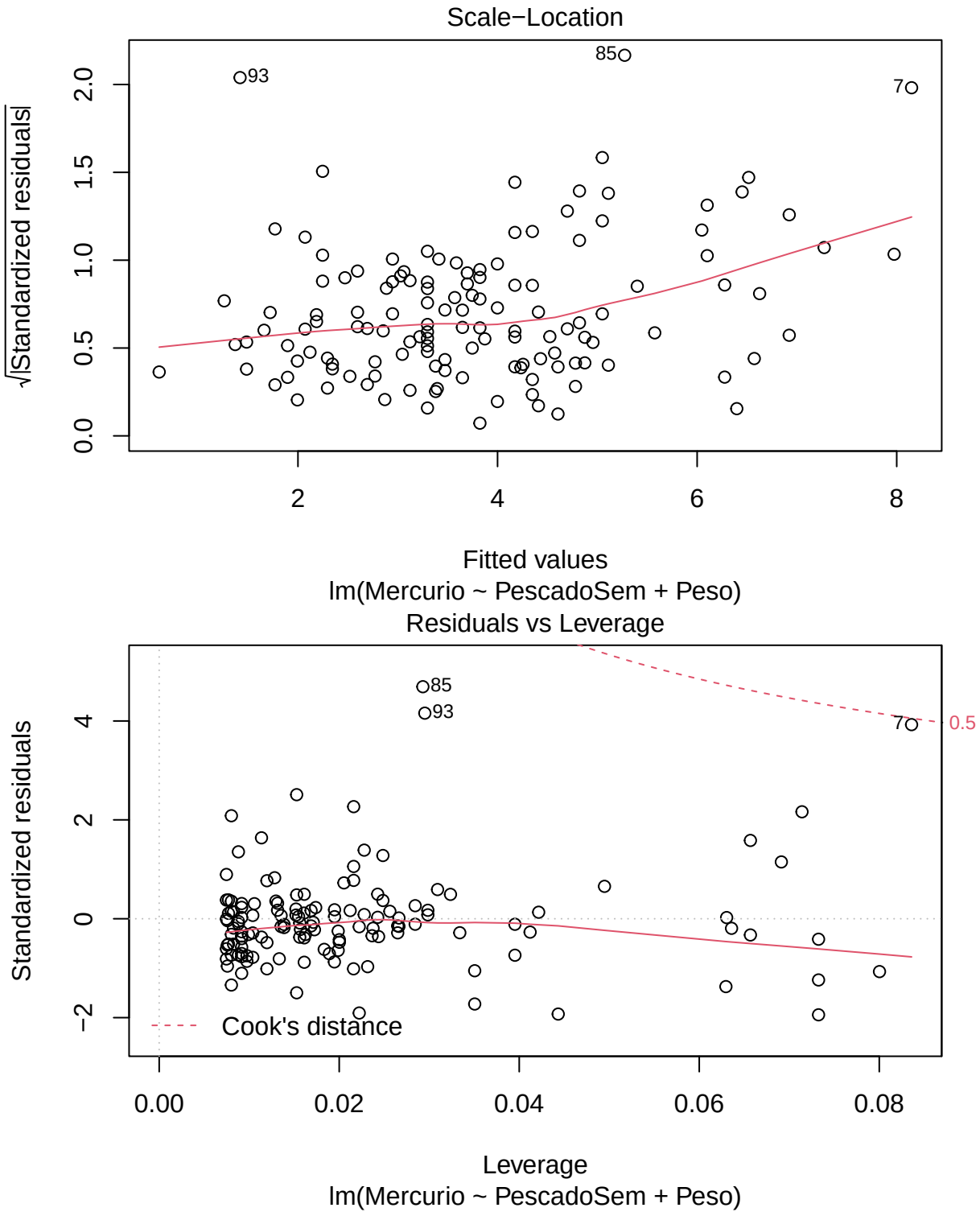
Si, la respuesta habría cambiado si se hubiera usado el LOOCV, en este caso el modelo más optimo es el que posee la interacción pues la segunda componente de delta es menor para este.

Ejercicio 3

```
plot(lm1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Mercurio ~ PescadoSem + Peso)
93
85
7


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Mercurio ~ PescadoSem + Peso)
85
93
7



Como se puede ver en el qqplot de los residuos, existe una gran desviación sobre la normalidad. Para contrastar esto se decide realizar un test de Kolmogorov Smirnov para comprobar la hipotesis de normalidad:

```
ks.test(x = rstudent(lm1), y = "pt", df=nrow(pesca)-2-2)
```

```
##
##  One-sample Kolmogorov-Smirnov test
```

```
##
## data:  rstudent(lm1)
## D = 0.1671, p-value = 0.001063
## alternative hypothesis: two-sided
```

Como se puede observar la hipotesis nula de normalidad se descarta para estos residuos. Así concluimos que la hipótesis normalidad no se cumple para este modelo.

Respecto de la hipotesis de linealidad, en el plot anterior de Residual vs Fitted así como en el de Scale-location se puede apreciar cierta tendencia en los residuos. Así se concluye que la hipotesis de linealidad no se cumple para el modelo.

Por último si atendemos a la hipótesis de homocedasticidad del modelo, si se observa el plot de Residual vs Fitted se podrá observar como la distribución de los puntos no es homogenea. Para contrastar de forma adecuada esta hipótesis se decide realizar un test de Levene:

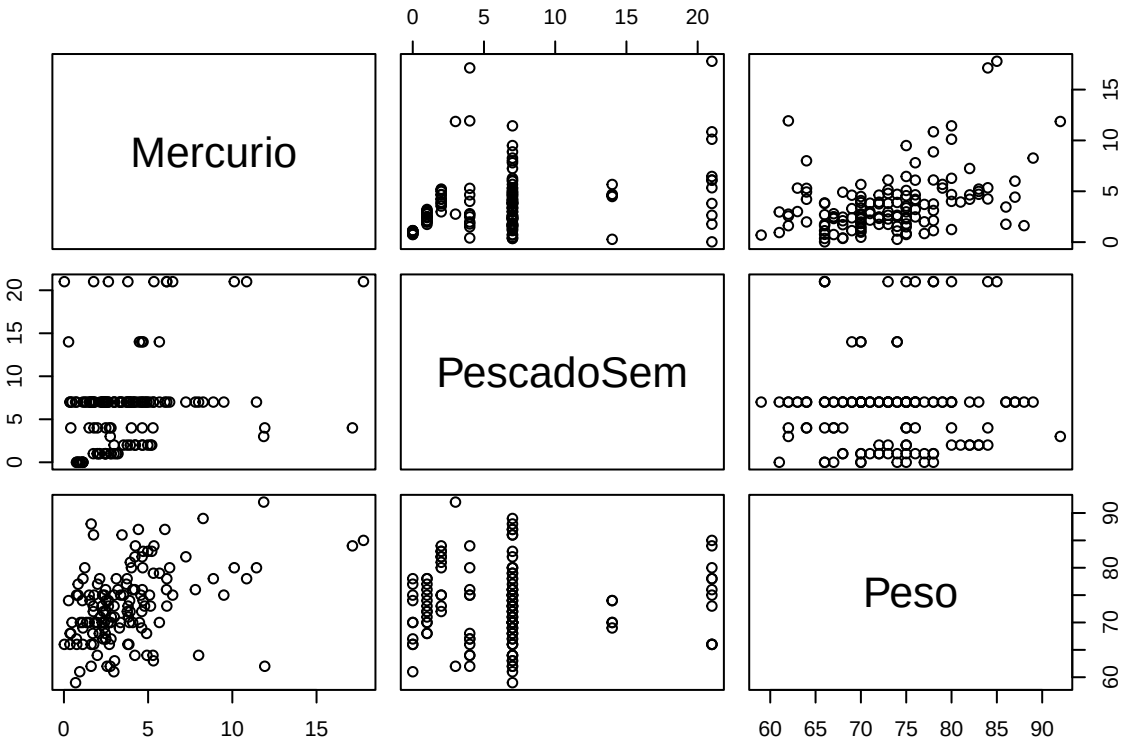
```
grupos <- cut(lm1$fitted.values, quantile(lm1$fitted.values, (0:4)/4),
include.lowest = TRUE)
lawstat::levene.test(rstandard(lm1), grupos)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  rstandard(lm1)
## Test Statistic = 6.3531, p-value = 0.0004698
```

En base a los resultados obtenidos podemos concluir que el modelo posee un problema de heterocidasticidad.

El modelo ajustado no es satisfactorio pues, las hipotesis de Normalidad, linealidad y homocedasticidad no se cumplen.

```
pairs(pesca[, c("Mercurio", "PescadoSem", "Peso")])
```



Una de las explicaciones posibles para la no adecuación del modelo puede ser que las relaciones asumidas como lineales

entre las covariables y la variable respuesta no se cumplan realmente. Si se observa el plot anterior se podra observar como la variable Peso y Mercurio parecen tener relación de tipo cuadrático.

Ejercicio 4

```
pesca$Mercuriolog <- log(pesca$Mercurio)
```

```
lm3 <- lm(Mercuriolog ~ PescadoSem + Peso + PescadoSem:Grupo + Peso:Grupo, data = pesca)
summary(lm3)
```

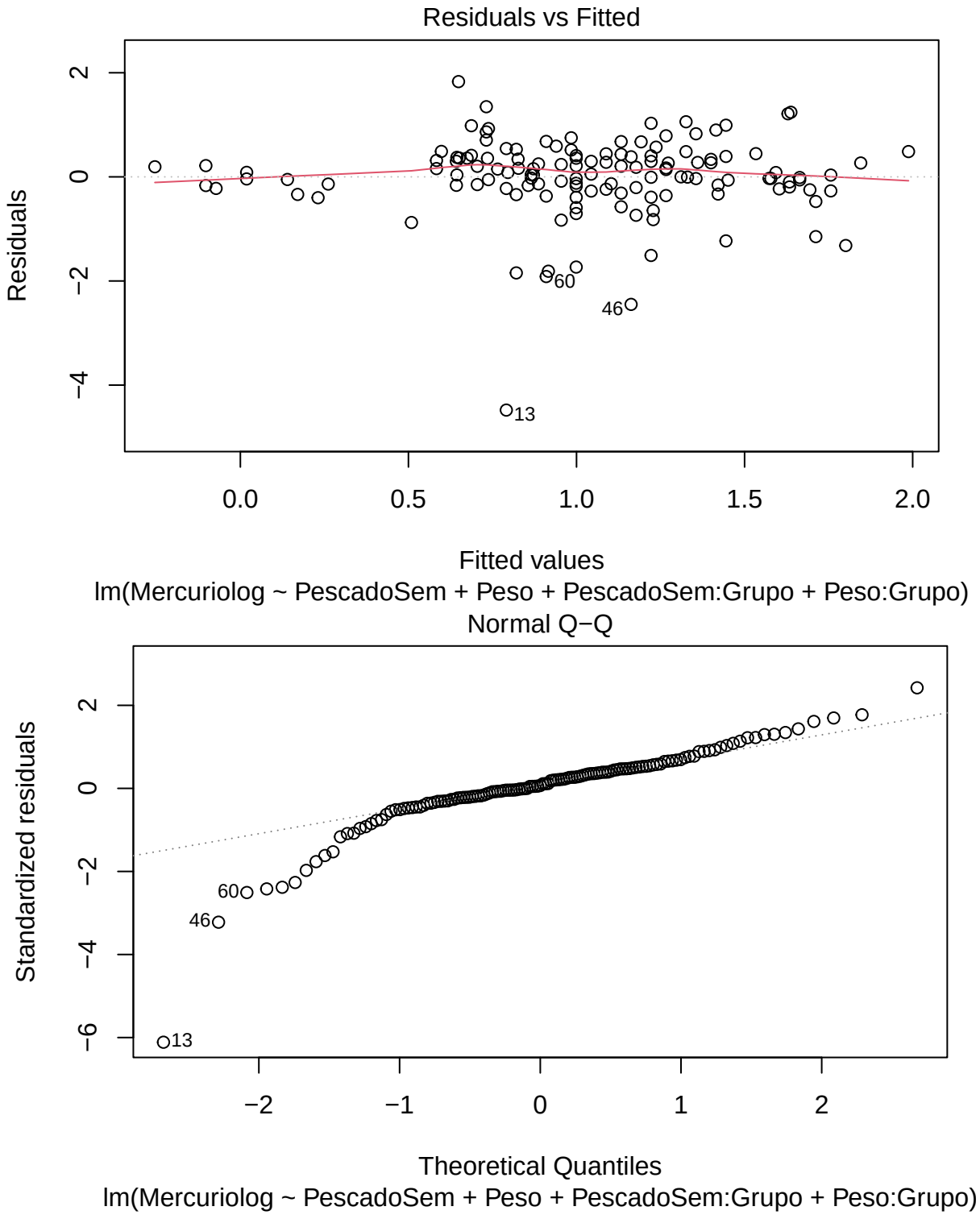
```
##
## Call:
## lm(formula = Mercuriolog ~ PescadoSem + Peso + PescadoSem:Grupo +
##     Peso:Grupo, data = pesca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4801 -0.2226  0.0523  0.3797  1.8296
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2.104809   0.754721  -2.789 0.006085 **
## PescadoSem                 0.625407   0.183291   3.412 0.000860 ***
## Peso                       0.030337   0.011155   2.719 0.007433 **
## PescadoSem:GrupoPescador -0.627527   0.184197  -3.407 0.000875 ***
## Peso:GrupoPescador         0.014216   0.003791   3.750 0.000265 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7694 on 130 degrees of freedom
## Multiple R-squared:  0.2461, Adjusted R-squared:  0.2229
## F-statistic: 10.61 on 4 and 130 DF,  p-value: 1.803e-07
```

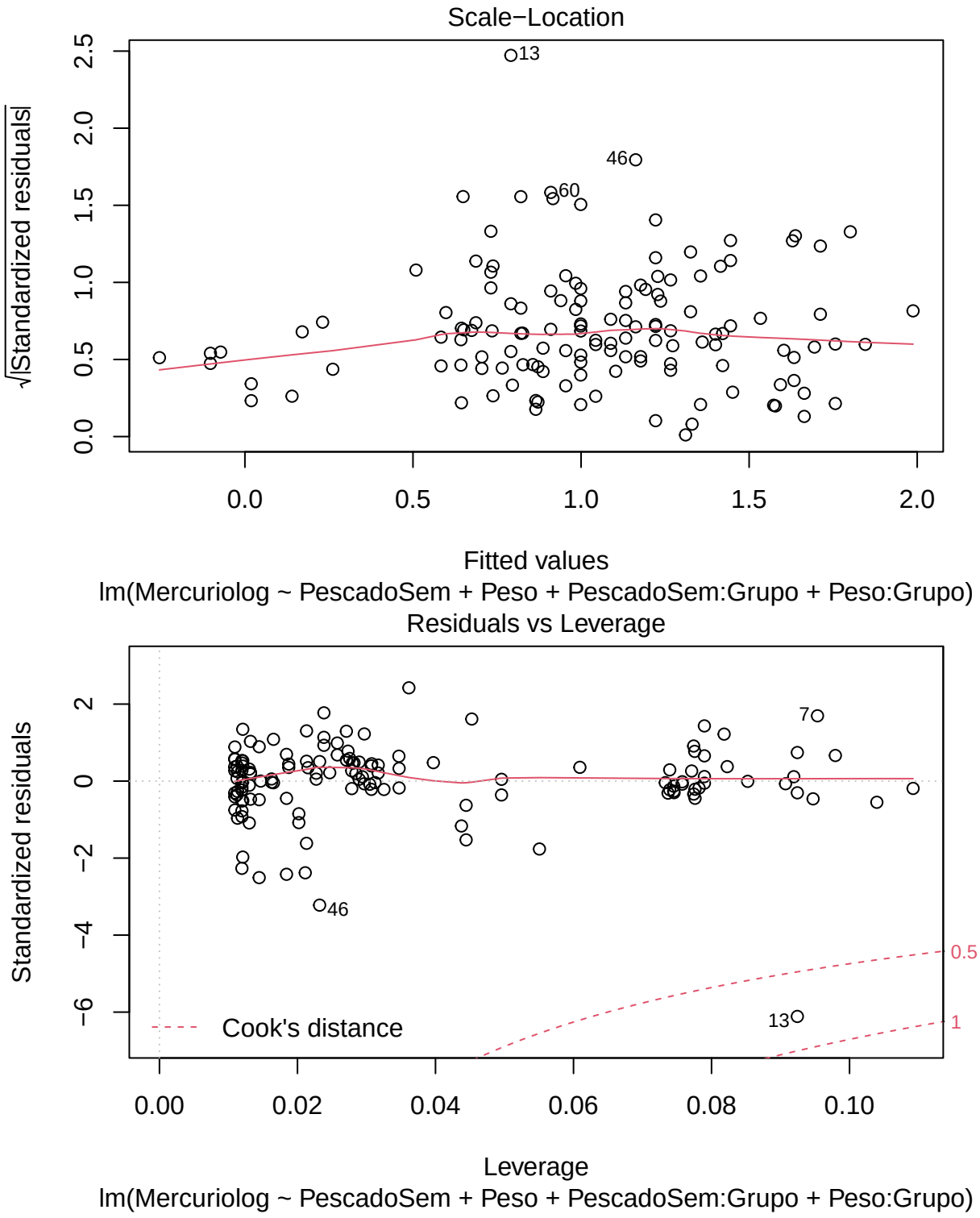
```
coef(lm3)
```

```
##              (Intercept)               PescadoSem                     Peso
##              -2.10480871               0.62540747              0.03033668
## PescadoSem:GrupoPescador       Peso:GrupoPescador
##              -0.62752656               0.01421636
```

El coeficiente de PescadoSem indica que un aumento en una unidad en esta covariables, aumenta el valor de la variable respueta (Mercuriolog) en 0.625 unidades. De igual forma el coeficiente de la covariable Peso indica que un aumento en una unidad en esta, aumentara la variable respuesta en 0.0303 unidades. El valor del coeficiente de de la interacción entre Grupo e PescadoSem indica que el aumento de una unidad en el PescadoSem para los individuos del grupo pescador produce un descenso en la variable respuesta de 0.62540747 - 0.62752656 =-0.00211909, mientras que para los individuos del grupo control esta seria de 0.6254. De forma análoga un auemento de una unidad en el peso para los individuos del grupo pescador produce ahora un aumento de 0.03033668 + 0.01421636 = 0.04455304 mientras que para los individuos del grupo control esta seria de 0.03033668.

```
plot(lm3)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Mercuriolog ~ PescadoSem + Peso + PescadoSem:Grupo + Peso:Grupo)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Mercuriolog ~ PescadoSem + Peso + PescadoSem:Grupo + Peso:Grupo)




```
ks.test(x = rstudent(lm3), y = "pt", df=nrow(pesca)-3-2)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstudent(lm3)
```

```
## D = 0.15641, p-value = 0.002707
## alternative hypothesis: two-sided
```

```
grupos <- cut(lm3$fitted.values, quantile(lm3$fitted.values, (0:4)/4),
include.lowest = TRUE)
lawstat::levene.test(rstandard(lm3), grupos)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  rstandard(lm3)
## Test Statistic = 0.054381, p-value = 0.9832
```

La hipotesis de normalidad de los residuos no se cumple (vease qqplot así como resultados del test de Kolmogorov Smirnov realizado, p-valor=0.0027), la hipótesis de homocedasticidad se cumple (vease el test de Levene realizado, p-valor=0.9832) y por último la hipótesis de linealidad tampoco se cumple pues se puede apreciar cierta tendencia en el plot de Scale-location, sin embargo es necesario destacar que esta tendencia se ha reducido con respecto al modelo sin la transformación en la variable respuesta. A la luz de estos resultados sigo sin dar por bueno este modelo, aunque los residuos parecen tener menor tendencia, y la hipótesis de homocedasticidad si se cumple para los residuos del modelo.

Ejercicio 5

```
individuo1 <- data.frame(Grupo = as.factor("Control"),
                         PescadoSem = 5,
                         Peso = 80)
individuo2 <- data.frame(Grupo = as.factor("Pescador"),
                         PescadoSem = 5,
                         Peso = 80)
predict(lm3, newdata = individuo1, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 3.449163 2.043259 4.855067
```

```
predict(lm3, newdata = individuo2, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 1.448839 1.200811 1.696867
```

Puesto que los intervalos de confianza al 95% de ambos valores predichos por el modelo no se solapan entre ellos, se puede considerar estadísticamente significativa la diferencia entre el valor log(Mercurio) para el grupo control y el de pescador.

```
predict(lm3, newdata = individuo1, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 3.449163 1.377139 5.521187
```

```
predict(lm3, newdata = individuo2, interval = "prediction")
```

```
##        fit        lwr      upr
## 1 1.448839 -0.0933143 2.990992
```

Puesto que los intervalos de predicción para los valores predichos se solapan, en base a estos intervalos de predicción, las diferencias encontradas entre el individuo del grupo pescador y el de control no son estadisticamente significativas.

En base a estos resultados se puede concluir que no existe diferencias estadisticamente significativas entre los valores predichos por este modelo, pues aunque los intervalos de confianza no se solapan, estos hacen referencia a la recta de regresión, no a los valores predichos. Por el contrario los intervalos de predicción si se solapan.