Manual de prácticas

Solución

Pregunta 1.

Explique en menos de 5 renglones. ¿Cuál es la base teórica por la que se argumenta que el Método de Mínimos Cuadrados Ponderados corrige el problema de la heteroscedasticidad?.

El modelo de Mínimos Cuadrados Ponderados corrige el problema de heteroscedasticidad, debido a que al ponderar por el inverso de la variancia se controlan los valores extremos que provocan la violación del supuesto.

Pregunta 2.

Explique en menos de 5 renglones. ¿Cuál es la base teórica por la que se argumenta que el Método de Componentes Principales corrige el problema de multicolinealidad?

Al utilizar el método de componentes principales, la multicolinealidad se corrige porque al tomar en cuenta la correlación de las antiguas variables, se crean nuevas, las la cuales tienen una correlación cero entre ellas.

Pregunta 3.

Se le da un archivo bridges.Rdata en formato R. Corresponde a una muestra de 45 proyectos de construcción de puentes en EEUU. Se quiere investigar qué características del diseño predicen el tiempo para realizar la obra.

La base de datos contiene entonces las siguientes variables:

- **case:** Identificador del caso
- **time:** Tiempo en días-persona
- **darea:** Área superficial, en miles de pies cuadrados
- **ccost:** Costo de construcción, en miles de dólares
- **dwgs:** Número de dibujos estructurales
- **length:** Largo del puente (en pies)
- **spans:** Número de soportes

```
load("bridge.Rdata")
```

a) *Estime un modelo de regresión gaussiano (el tradicional de mínimos cuadrados ordinarios) en el que se prediga el tiempo para realizar las obras en función del resto de variables (excepto case por ser el identificador). Llámelo mod1. Presente el summary del modelo*

```
mod1 <- lm(time~darea+ccost+dwgs+length+spans, data = bridge)
summary(mod1)
```

```
##
## Call:
## lm(formula = time ~ darea + ccost + dwgs + length + spans, data = brid
ge)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.816 -26.797  -9.674  24.882 180.443
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34.83256   25.03837  -1.391    0.172
## darea         0.24675    1.63170   0.151    0.881
## ccost        -0.02107    0.07143  -0.295    0.770
## dwgs         19.68195    4.08583   4.817 2.23e-05 ***
## length        0.05186    0.10378   0.500    0.620
## spans        15.50455   10.14243   1.529    0.134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.31 on 39 degrees of freedom
## Multiple R-squared:  0.7101, Adjusted R-squared:  0.6729
## F-statistic:  19.1 on 5 and 39 DF,  p-value: 1.435e-09
```

b) *Analice la normalidad de los residuos con una prueba de Jarque Bera. Plantee la hipótesis nula y alternativa, y responda adecuadamente. Conteste: ¿Se puede suponer normalidad de los errores? ¿Por qué sí o por qué no?*

H_0: Los residuos se distribuyen como una normal con media $\mu = 0$ y varianza σ^2.

H_1: Los residuos no se distribuyen como una normal con media $\mu = 0$ y varianza σ^2.

```
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
jarque.bera.test(residuals(mod1))
```

```
##
##  Jarque Bera Test
##
## data:  residuals(mod1)
## X-squared = 17.992, df = 2, p-value = 0.0001239
```

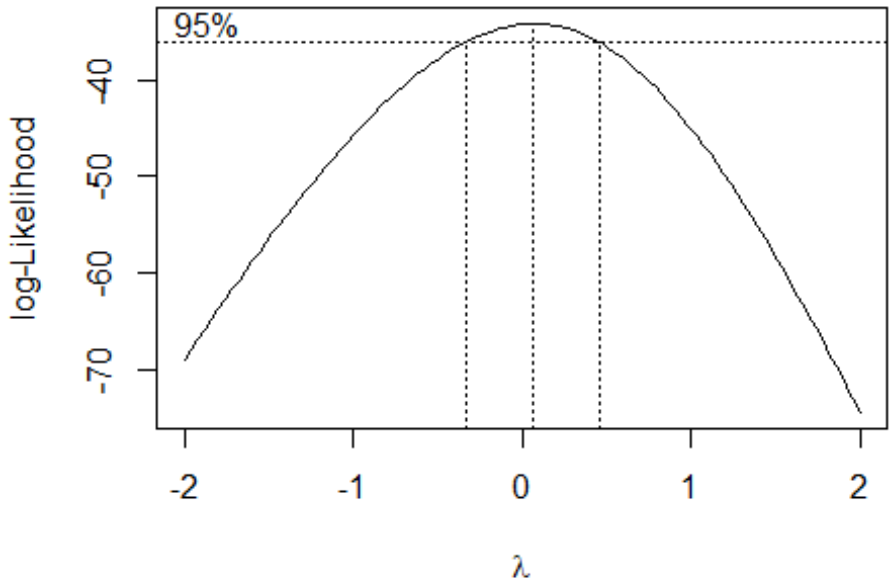
Hay suficiente evidencia estadística para rechazar la hipótesis nula de que los residuos se distribuyen como una normal con media $\mu = 0$ y varianza σ^2. Esto con una significancia del 5%.

c) *Utilice el método de BoxCox para encontrar una transformación para la variable dependiente. Pegue el gráfico en el examen. Escoja un exponente redondeado que sea razonable de acuerdo al gráfico de BoxCox*

```
library(car)

## Loading required package: carData

library(MASS)
boxcox(time~darea+ccost+dwgs+length+spans, data = bridge)
```



Se escoge la transformación logarítmica

```
round(vif(mod1), 2)

##  darea  ccost   dwgs length  spans
##   4.48   6.84   2.09   6.24   4.17
```

d) *Estime un nuevo modelo gaussiano con la variable dependiente transformada, llame al modelo mod2, y verifique con la prueba de Jarque Bera si ya se puede suponer normalidad. No es necesario que plantee las hipótesis nula y alternativa. Solo señale la conclusión.*

```
mod2 <- lm(log(time)~darea+ccost+dwgs+length+spans, data = bridge)
summary(mod2)
```

```
##
## Call:
## lm(formula = log(time) ~ darea + ccost + dwgs + length + spans,
##     data = bridge)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66908 -0.18936 -0.05747  0.30321  0.64653
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.584e+00  1.553e-01  23.083  < 2e-16 ***
## darea        8.997e-04  1.012e-02   0.089    0.930
## ccost       -8.874e-06  4.429e-04  -0.020    0.984
## dwgs         1.354e-01  2.534e-02   5.344 4.22e-06 ***
## length       7.861e-05  6.435e-04   0.122    0.903
## spans        1.019e-01  6.289e-02   1.620    0.113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.343 on 39 degrees of freedom
## Multiple R-squared:  0.7328, Adjusted R-squared:  0.6986
## F-statistic: 21.39 on 5 and 39 DF,  p-value: 3.044e-10
```

```
jarque.bera.test(residuals(mod2))
```

```
##
##  Jarque Bera Test
##
## data:  residuals(mod2)
## X-squared = 0.90136, df = 2, p-value = 0.6372
```

No hay suficiente evidencia estadística para rechazar la hipótesis nula de que los residuos se distribuyen como una normal con media $\mu = 0$ y varianza σ^2. Por lo tanto, se puede concluir que si se cumple el supuesto de normalidad.

e) *Utilice el método de Bootstrap para estimar intervalos no paramétricos para ambos modelos (mod1 y mod2), y compárelos con los intervalos paramétricos teóricos de ambos modelos (mod1 y mod2). A partir de la comparación con los intervalos de bootstrap, argumente si hay evidencia de que la transformación mejora las estimaciones del modelo.*

```
round(mod1$coef, 2)
```

```
## (Intercept)       darea       ccost        dwgs      length       span
s
##      -34.83        0.25       -0.02       19.68        0.05       15.5
0
```

```
Lim.inf <- mod1$coef-qt(0.975, length(bridge$time)-length(mod1$coef))*sum
mary(mod1)$sigma*(diag(summary(mod1)$cov.unscaled))^0.5
Lim.sup <- mod1$coef+qt(0.975, length(bridge$time)-length(mod1$coef))*sum
```

```
mary(mod1)$sigma*(diag(summary(mod1)$cov.unscaled))^0.5


Lim.inf2 <- mod2$coef-qt(0.975, length(bridge$time)-length(mod2$coef))*su
mmary(mod2)$sigma*(diag(summary(mod2)$cov.unscaled))^0.5
Lim.sup2 <- mod2$coef+qt(0.975, length(bridge$time)-length(mod2$coef))*su
mmary(mod2)$sigma*(diag(summary(mod2)$cov.unscaled))^0.5

library(boot)

coefic = function(y, x, d) {
    lm(y[d]~x[d, ])$coef
}

coef <- boot(bridge$time, coefic, R = 1000, x = cbind(bridge$darea, bridg
e$ccost, bridge$dwgs, bridge$length, bridge$spans))

rbind(boot.ci(coef, index = 1, type = "perc")$percent, boot.ci(coef,index
=2, type = "perc")$percent,
      boot.ci(coef, index = 3, type = "perc")$percent, boot.ci(coef, inde
x = 4, type = "perc")$percent,
      boot.ci(coef, index = 5, type = "perc")$percent, boot.ci(coef, inde
x = 6, type = "perc")$percent)

##      conf
## [1,] 0.95 25.03 975.98 -111.9155808 12.9830066
## [2,] 0.95 25.03 975.98   -6.9355997  3.8079146
## [3,] 0.95 25.03 975.98   -0.2232259  0.2318421
## [4,] 0.95 25.03 975.98    8.8734070 34.5731768
## [5,] 0.95 25.03 975.98   -0.2230059  0.4454990
## [6,] 0.95 25.03 975.98   -6.4876634 50.8129638

cbind(Lim.inf, mod1$coef, Lim.sup)

##                 Lim.inf                 Lim.sup
## (Intercept) -85.4774442 -34.83255956 15.8123251
## darea        -3.0536675   0.24675152  3.5471705
## ccost        -0.1655524  -0.02106905  0.1234143
## dwgs         11.4175737  19.68194556 27.9463174
## length       -0.1580568   0.05185508  0.2617670
## spans        -5.0104512  15.50454556 36.0195423

coef2 <- boot(log(bridge$time), coefic, R = 1000, x = cbind(bridge$darea,
bridge$ccost, bridge$dwgs, bridge$length, bridge$spans))

rbind(boot.ci(coef2, index = 1, type = "perc")$percent, boot.ci(coef2, in
dex = 2, type = "perc")$percent,
      boot.ci(coef2, index = 3, type = "perc")$percent, boot.ci(coef2, in
dex = 4, type = "perc")$percent,
      boot.ci(coef2, index = 5, type = "perc")$percent, boot.ci(coef2, in
dex = 6, type = "perc")$percent)
```

```
##      conf
## [1,] 0.95 25.03 975.98  3.134154103 3.825663514
## [2,] 0.95 25.03 975.98 -0.033130832 0.021566477
## [3,] 0.95 25.03 975.98 -0.001215309 0.001620613
## [4,] 0.95 25.03 975.98  0.087379589 0.211617291
## [5,] 0.95 25.03 975.98 -0.001755376 0.002436574
## [6,] 0.95 25.03 975.98 -0.033033845 0.294941613
```

```
cbind(Lim.inf2, mod2$coef, Lim.sup2)
```

```
##                  Lim.inf2                   Lim.sup2
## (Intercept)  3.2698631996  3.583913e+00 3.8979636750
## darea       -0.0195662806  8.997030e-04 0.0213656865
## ccost       -0.0009048191 -8.874007e-06 0.0008870711
## dwgs         0.0841372875  1.353849e-01 0.1866324535
## length      -0.0012230555  7.861368e-05 0.0013802829
## spans       -0.0253460003  1.018680e-01 0.2290820487
```

En realidad no hay una mejora sustancial en el modelo debido a que con el método boostrap se observa que los intervalos en su longitud se mantienen muy parecidos y en la cantidad de coeficientes que son significativos.

f) *Estime un modelo heteroscedástico (modelo lineal generalizado doble dglm) con la especificación del modelo1 y llámelo mod1h. A partir de las estimaciones de este modelo, justifique si hay presencia de heteroscedasticidad o no, y cómo lo sabe.*

```
library(dglm)
```

```
## Loading required package: statmod
```

```
mod1h <- dglm(time~darea+ccost+dwgs+length+spans, dformula = ~darea+ccost
+dwgs+length+spans, data = bridge)
summary(mod1h)
```

```
##
## Call: dglm(formula = time ~ darea + ccost + dwgs + length + spans,
##     dformula = ~darea + ccost + dwgs + length + spans, data = bridge)
##
## Mean Coefficients:
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -33.7035378 21.04550532 -1.601460 0.1173451930
## darea         2.0795362  1.27626895  1.629387 0.1112830568
## ccost        -0.1434774  0.04380360 -3.275470 0.0022182764
## dwgs         17.9138017  4.31629223  4.150275 0.0001742771
## length        0.1863209  0.07998924  2.329325 0.0251133652
## spans        17.0985236  8.08232903  2.115544 0.0408229182
## (Dispersion Parameters for gaussian family estimated as below )
##
##     Scaled Null Deviance: 223.4888 on 44 degrees of freedom
## Scaled Residual Deviance: 44.99996 on 39 degrees of freedom
##
```

```
## Dispersion Coefficients:
##                  Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)  3.8103580459 0.640146495  5.9523220 2.643646e-09
## darea       -0.0459882774 0.041716980 -1.1023875 2.702932e-01
## ccost       -0.0009313161 0.001826256 -0.5099592 6.100800e-01
## dwgs         0.4238291282 0.104460869  4.0573004 4.964320e-05
## length      -0.0037357031 0.002653267 -1.4079637 1.591418e-01
## spans        0.7834290326 0.259307595  3.0212344 2.517464e-03
## (Dispersion parameter for Gamma family taken to be 2 )
##
##     Scaled Null Deviance: 77.00734 on 44 degrees of freedom
## Scaled Residual Deviance: 42.4805 on 39 degrees of freedom
##
## Minus Twice the Log-Likelihood: 454.253
## Number of Alternating Iterations: 22

cbind(summary(mod1)$coef, summary(mod1h)$coef)

##                 Estimate  Std. Error    t value     Pr(>|t|)    Estima
te
## (Intercept) -34.83255956 25.03837050 -1.3911672 1.720607e-01 -33.70353
78
## darea         0.24675152  1.63169715  0.1512238 8.805787e-01   2.07953
62
## ccost        -0.02106905  0.07143127 -0.2949556 7.695923e-01  -0.14347
74
## dwgs         19.68194556  4.08583028  4.8171226 2.226518e-05  17.91380
17
## length        0.05185508  0.10377854  0.4996705 6.201146e-01   0.18632
09
## spans        15.50454556 10.14242785  1.5286819 1.344144e-01  17.09852
36
##              Std. Error   t value     Pr(>|t|)
## (Intercept) 21.04550532 -1.601460 0.1173451930
## darea        1.27626895  1.629387 0.1112830568
## ccost        0.04380360 -3.275470 0.0022182764
## dwgs         4.31629223  4.150275 0.0001742771
## length       0.07998924  2.329325 0.0251133652
## spans        8.08232903  2.115544 0.0408229182
```

Con el modelo heterocedástico, se puede decir que viola el supuesto de homocedasticidad porque desde su concepción no corrige esta violación si no que la toma en cuenta y recalcula la variabilidad de la variable dependiente, restringida por las independientes escogidas. Sin embargo, al ser una manera de un método de corrección para la violación del supuesto de homocedasticidad. Se puede tener homoscedasticidad algo que se observa con la significancia de los coeficientes de dispersión.

g) *Estime la prueba de variancia no constante (ncvTest) para el modelo 1 y el modelo 2. Plantee una sola vez la hipótesis nula y alternativa. Diga si hay alguna diferencia en el*

supuesto de homoscedasticidad entre el modelo 1 y el modelo 2, y cómo una transformación de Box Cox (que es para corregir violaciones a la normalidad) puede generar similitudes o diferencias en los resultados de estas pruebas

H_0: Las variancias de las variables independientes son constantes.

H_1: Las variancias de las variables independientes no son constantes.

```
ncvTest(mod1)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 16.08882, Df = 1, p = 6.044e-05

ncvTest(mod2)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3.425152, Df = 1, p = 0.06421
```

La transformación de box Cox puede generar diferencias esto debido a que contiene una transformación, es un modelo diferente el cual como en el anterior se le deben verificar los supuestos.

En el modelo 1 (regresión gaussiano) se rechaza la hipótesis nula de que las variancias de las variables independientes son constantes, mientras que en el modelo 2 (transformación en la variable dependiente) no se rechaza.

Por lo que si hay diferencias entre los modelos respecto al supuesto de homocedasticidad y se preferiría usar el segundo modelo dado que corrige el problema de la heterocedasticidad (variancias no constantes).

Pregunta 4.

La Asociación Nacional de Productores de Cebolla encarga un estudio para analizar qué condiciones son las que les permite tener ganancias o pérdidas a los vendedores de cebolla en las Ferias del Agricultor. A la Asociación no le interesa el monto de la ganancia, sino si tuvo pérdidas o ganancias cada día de la feria del agricultor. Toma una muestra de 2500 cebolleros y los entrevista por teléfono el día después de la Feria. Se construye una base de datos cebolleros.dta con las siguientes variables:

- **id:** identificador del vendedor
- **morada:** Si vende cebolla seca amarilla o morada (1=morada, 0=amarilla)
- **personasenpuesto:** Cuántas personas lo ayudan a vender.
- **ganancia8:** Variable binaria: 1= ganancia, 0=pérdida

Se estimó un modelo de regresión logística que generó las siguientes estimaciones:

```
> summary(glm(ganancia8~morada+personasenpuesto,
family=binomial(link="logit"),data=cebolleros))
```

```
Call:
glm(formula = ganancia8 ~ morada + personasenpuesto, family =
binomial(link = "logit"),
    data = cebolleros)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9397  -0.8472  -0.7500   1.5161   1.7090

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.19624    0.09162 -13.056  < 2e-16 ***
morada            0.32027    0.08998   3.559 0.000372 ***
personasenpuesto  0.03590    0.03045   1.179 0.238281
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2959.1  on 2499  degrees of freedom
Residual deviance: 2944.7  on 2497  degrees of freedom
AIC: 2950.7

Number of Fisher Scoring iterations: 4
```

a) *Interprete los coeficientes de las variables “morada” y “personasenpuesto”, haciendo la transformación apropiada*

```
(exp(0.32027)-1)*100
```

```
## [1] 37.74996
```

Entre las cebollas moradas los odds de ganancia son 38% más altos que los odds de ganancia entre las cebollas amarillas. Esto manteniendo las demás variables constantes.

```
(exp(0.03590)-1)*100
```

```
## [1] 3.655219
```

Por cada persona que tiene extra trabajando en el puesto los odds de ganancia aumentan en un 3,66%. Manteniendo las demás variables constantes.

b) *Explique cuál es la lógica de usar la prueba de Hosmer y Lemeshow como una prueba de bondad de ajuste para una regresión logística.*

Utilizar esta prueba lo que hace es calificar los valores predichos en 10 deciles, los cuales van a ser contrastados con los valores esperados. Con esto se obtiene una prueba chi cuadrado de bondad de ajuste. Teniendo así el estadístico de prueba de hipótesis la cual ayuda a llegar a la conclusión de si se tiene un buen ajuste o no.

Pregunta 5.

Se le da el archivo dependiente.Rdata. Este contiene una sola variable: dependiente. Con un ciclo, encuentre la estimación de lambda y el valor de la logverosimilitud del modelo,

usando el método de máxima verosimilitud tal que los residuos del modelo en el que la variable dependiente es igual a (dependiente^lambda) se distribuya aproximadamente normal (En otras palabras, las estimaciones de la transformación de BoxCox, pero no use la transformación de BoxCox sino una programación con ciclos.). Tiene que evaluar 100 valores en el ciclo, entre 0.01 a 1.

Se sabe que el error estándar residual es igual a: 0.0836, el Beta0=1.85 (bajo la transformación) y que la función de verosimilitud del método de BoxCox es:

$$L(\beta_0\beta_1,\dots,\beta_{p-1},\sigma^2,\lambda) = (2\pi\sigma^2)^{-n/2} * exp\left[-\frac{1}{2} * \sum_{i=1}^{n}\left(Y_i^{\lambda} - \beta_0 - \beta_1 X_{1,i} - \cdots - \beta_{p-1}X_{p-1,i}\right)^2\right]$$

```
load("dependiente.Rdata")

almacen <- matrix(rep(NA, 200), ncol = 2)

for (i in 1:100) {

    lambda <- i/100
    beta0 <- 1.85
    logvero <- sum(log(dnorm(I(dependiente^lambda), mean = beta0, sd = 0.
0836)))
    almacen[i, 1] <- lambda
        almacen[i, 2] <- logvero


}

max(almacen[, 2])

## [1] 212.9851

almacen[almacen[, 2] == max(almacen[, 2])]

## [1]   0.2200 212.9851
```

Pregunta 6.

Explique en menos de 5 renglones. El procedimiento de Bootstrap utiliza remuestreo (una cantidad grande de muestras con reemplazo de la muestra original). Si hay presencia de heteroscedasticidad, explique qué característica(s) tiene el procedimiento de bootstrap para que los intervalos no paramétricos de bootstrap para una pendiente sean preferibles a los intervalos de confianza estimados con la fórmula teórica $\hat{\beta} \pm t_{(1-\alpha/2,n-p)} * error.estándar(\hat{\beta})$.

En el procedimiento bootstrap se generan estimaciones con cada remuestra por lo que la distribución empírica que se genera para el estimador toma en cuenta la distribución propia de los datos (o sea, la distribución heteroscedástica).

Pregunta 7.

La pandemia de COVID-19 ha popularizado una serie de términos que se restringían a la jerga científica de las disciplinas de la salud: reproducibilidad, carga viral, período de latencia, etc. Suponga que surge un nuevo patógeno infeccioso, el SARS-COV3, que no es mortal pero sí enferma a las personas durante un período prolongado. Virólogos recogen especímenes de saliva de una muestra de 50 pacientes infectados con el nuevo patógeno, y miden la carga viral 3 días después del primer síntoma de fiebre (temperatura corporal mayor o igual a 38°C). Ellos tienen la hipótesis que la carga viral varía por edad y sexo, así que le piden a usted que estime un modelo de regresión que sirva para predecir la carga viral (variable carga.viral10, en copias por mL) en función de la edad (medida en años) y el sexo (variable mujer, que es igual a 1 si la persona es mujer, y 0 si es hombre). La información está en la base de datos sarscovtri.Rdata. Conteste las siguientes preguntas:

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

```
## 'data.frame':    50 obs. of  4 variables:
##  $ X1.50        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ carga.viral10: num  52.2 42.1 27.6 26.3 53.3 ...
##  $ edad         : num  29 43 37 40 30 50 30 47 46 28 ...
##  $ mujer        : int  0 1 0 0 1 1 0 1 1 1 ...
```

a) *Estime un modelo lineal gaussiano (el modelo de regresión tradicional) y analice el supuesto de normalidad con un qqPlot de los residuos y una prueba de Shapiro (use un alfa de 5%). Diga qué concluye sobre el supuesto de normalidad*

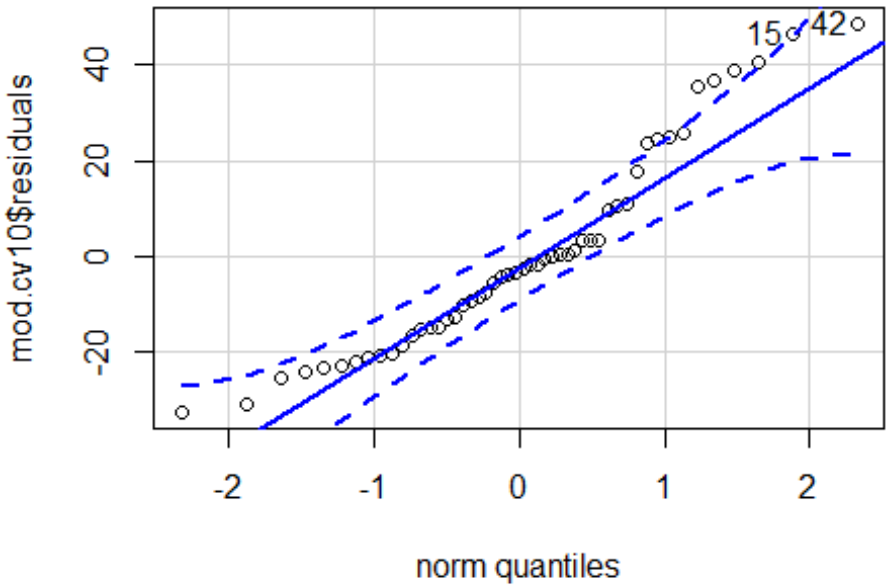
```
mod.cv10 <- lm(carga.viral10~edad+mujer, data = sarscovtri)
summary(mod.cv10)
```

```
##
## Call:
## lm(formula = carga.viral10 ~ edad + mujer, data = sarscovtri)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.483 -15.237  -2.882  10.269  48.670
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.6056    15.6190   3.688 0.000585 ***
## edad         -0.1811     0.3699  -0.489 0.626810
## mujer         0.7408     6.3871   0.116 0.908162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.39 on 47 degrees of freedom
## Multiple R-squared:  0.005422,   Adjusted R-squared:  -0.0369
## F-statistic: 0.1281 on 2 and 47 DF,  p-value: 0.8801
```

```
shapiro.test(mod.cv10$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  mod.cv10$residuals
## W = 0.93301, p-value = 0.007197

library(car)
qqPlot(mod.cv10$residuals)
```




```
## [1] 42 15
```

Prueba de hipótesis.

$H0$: Errores se distribuyen normalente

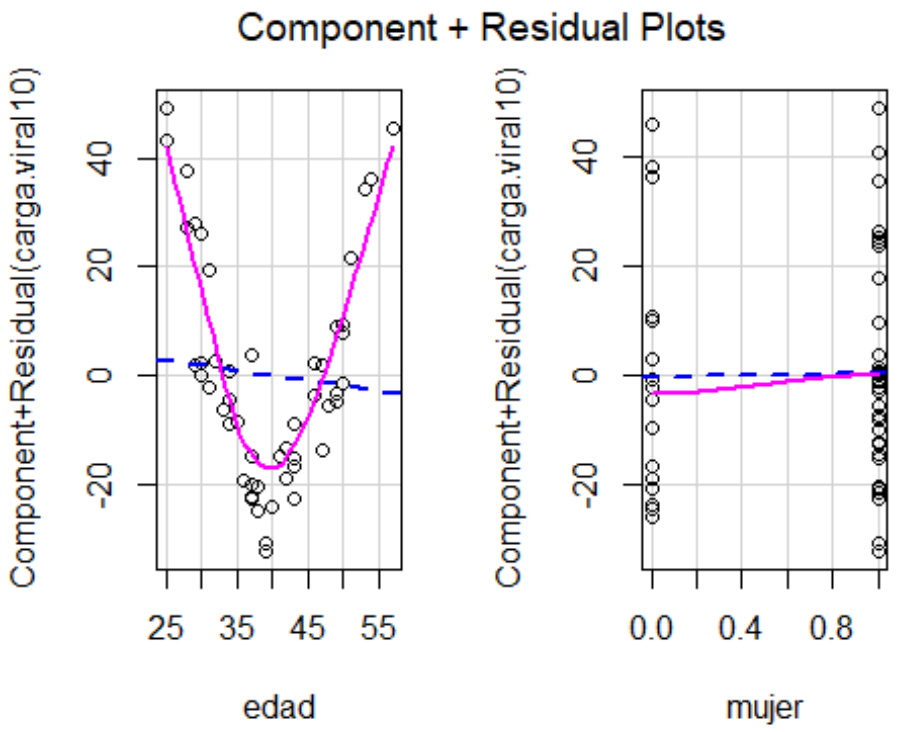
H_1: Errores no se distribuyen normalmente.

Interpretación: Se rechaza la hipótesis nula con un alfa del 5 % que los errores se distribuyen normalmente, manteniendo las demás variables constantes.

Además según el análisis gráfico no tenemos normalidad por los valores extremos que se presentan en la parte superior y en general hay puntos que estan afuera de las bandas de confianza.

b) *Haga gráficos crPlots con la ecuación y diga si hay alguna transformación que usted sugeriría para las variables predictoras. Explique su elección*

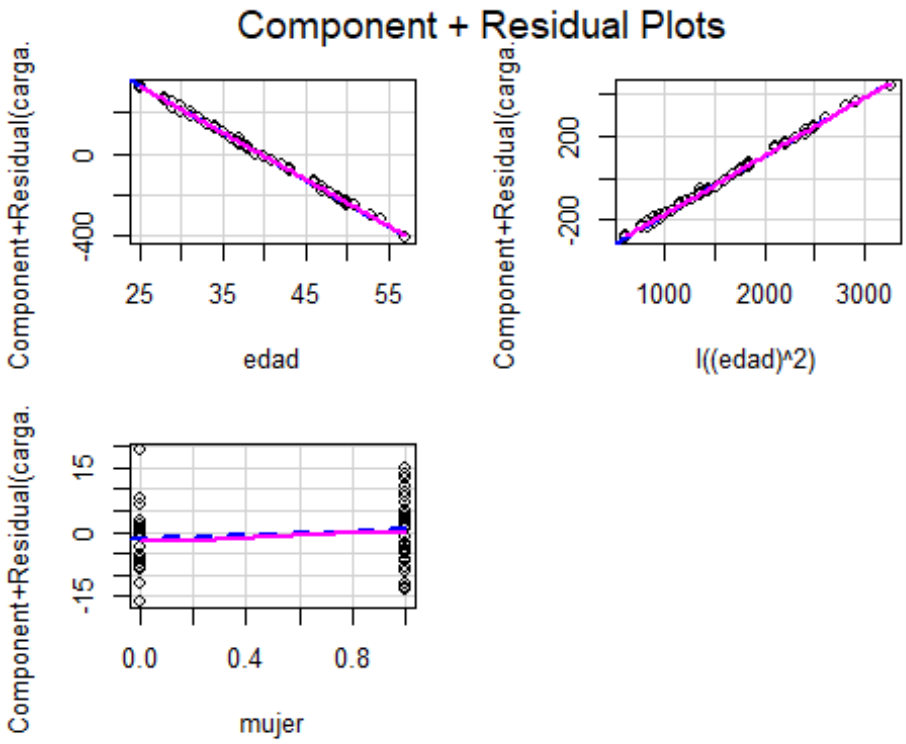
```
crPlots(mod.cv10)
```



Según el análisis gráfico una transformación cuadrática es la más apropiada por la forma (concava hacia arriba), para la variable edad.

c) *Estime un nuevo modelo lineal gaussiano con la transformación sugerida. Compare la bondad de ajuste de este modelo con el modelo del inciso (a), y diga si la transformación mejoró en algo la bondad de ajuste del modelo. [Nota: Para obtener todos los puntos de la pregunta, tiene que enseñarme un nuevo modelo con un crecimiento sustancial en el indicador de bondad de ajuste]*

```
mod.cv11 <- lm(carga.viral10~edad+I((edad)^2)+mujer, data = sarscovtri)
crPlots(mod.cv11)
```




```
summary(mod.cv10)$adj.r.squared
```

```
## [1] -0.03690087
```

```
summary(mod.cv11)$adj.r.squared
```

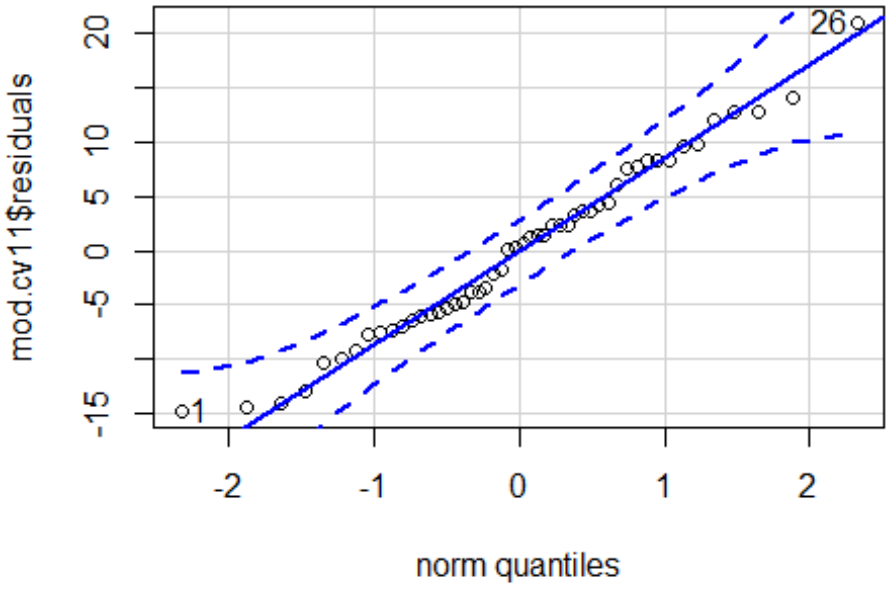
```
## [1] 0.8381702
```

d) *Haga un gráfico qqPlot de los residuos y una prueba de Shapiro (al 5% de significancia) con el nuevo modelo, y diga qué concluye sobre la violación al supuesto de normalidad.*

```
shapiro.test(mod.cv11$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod.cv11$residuals
## W = 0.98185, p-value = 0.6323
```

```
qqPlot(mod.cv11$residuals)
```




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

Según el análisis gráfico se puede suponer normalidad por que los valores no son tan extremos y en general no hay puntos que estén por fuera de las bandas de confianza.

Prueba de hipótesis.

$H0$: Errores se distribuyen normalente

H_1: Errores no se distribuyen normalmente.

Interpretación: No se rechaza la hipótesis nula con un alfa del 5 % que los errores se distribuyen normalmente, manteniendo las demás variables constantes.

e) *A partir del análisis anterior, explique cómo se relacionan los supuestos de linealidad y normalidad.*

Si la relación entre las predictoras y la dependiente no es lineal, al estimar un modelo con especificación errónea, se van a generar residuos extremos, produciendo violación a la normalidad.

Pregunta 8.

Se le da el archivo promociones.Rdata de 100 clientes de una tienda que contestaron un cuestionario para determinar quiénes estarían dispuestos a recibir promociones por Internet. La tienda quiere escoger un modelo qué determine cuáles variables predicen correctamente el deseo de recibir información, por lo que decide estimar un modelo

logístico para determinar la probabilidad de que un cliente quiera recibir este tipo de emails. Las variables son las siguientes:

```
load("promociones.Rdata")
names(promociones)

## [1] "edad"      "mujer"     "gastos"    "promocion" "apellido"
```

a) *Estime un modelo de regresión en el que se prediga la probabilidad de querer recibir emails de promoción, en función de la edad, el ser mujer y los gastos mensuales, e interprete el coeficiente de la variable edad.*

```
mod.promocion <- glm(promocion~edad+mujer+gastos, family = binomial(link
= "logit"), data = promociones)
exp(coefficients(mod.promocion))

## (Intercept)        edad       mujer      gastos
##   1.9805450   1.0113216   2.8193390   0.9944644

(1.0113216-1)*100

## [1] 1.13216
```

Por cada año adicional de edad, los odds de recibir un email aumentan en un 1.1%, manteniendo constante las demás variables.

b) *Estime un modelo de regresión en el que se prediga la probabilidad de recibir emails de promoción en función de mujer únicamente. Conociendo los resultados del beta0 y del beta1, programe un procedimiento en el que pruebe varios valores de beta0 y beta1 en la función de verosimilitud del modelo logístico para variable binaria (en clase lo vimos con ciclos), para encontrar las estimaciones de máxima verosimilitud para Beta0 y Beta1. [Recomendación: Acepto que pruebe solo 100 valores para Beta0 y 100 valores para Beta1, o sea un ciclo con 100mil corridas]*

La función de log verosimilitud en una regresión logística sería:

```
mod.promo2 <- glm(promocion~mujer, family = binomial(link = "logit"), dat
a = promociones)
summary(mod.promo2)

##
## Call:
## glm(formula = promocion ~ mujer, family = binomial(link = "logit"),
##     data = promociones)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8576  -0.4942  -0.4942  -0.4942   2.0801
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.0412     0.3361  -6.073 1.26e-09 ***
```

```
## mujer         1.2303     0.6885   1.787    0.074 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80.993  on 99  degrees of freedom
## Residual deviance: 78.119  on 98  degrees of freedom
## AIC: 82.119
##
## Number of Fisher Scoring iterations: 4
```

```
almacen.beta0  <- rep(NA, 10000)
almacen.beta1  <- rep(NA, 10000)
almacen.loglik <- rep(NA, 10000)

beta0 <- seq(-10,-0.1, 0.1)
beta1 <- seq(0.1, 10, 0.1)

for (i in 1:100) {

  for (j in 1:100) {
    almacen.beta0[(i-1)*100+j] <- beta0[j]
    almacen.beta1[(i-1)*100+j] <- beta1[i]

    betax = beta0[j]+beta1[i]*promociones$mujer
    veros = ((exp(betax)/(1+exp(betax)))^(promociones$promocion))*((1-exp
(betax)/(1+exp(betax)))^(1-promociones$promocion))
    almacen.loglik[(i-1)*100+j] <- sum(log(veros))

  }

}

almacen.beta0[almacen.loglik == max(almacen.loglik)]
```

```
## [1] -2
```

```
almacen.beta1[almacen.loglik == max(almacen.loglik)]
```

```
## [1] 1.2
```

Pregunta 9.

Si en una regresión con "p" predictores, se utiliza un Análisis de Componentes Principales con todos los predictores, explique en qué condiciones y por qué la suma de los valores característicos ("lambdas") estimados con el modelo es equivalente a "p".

Cuando se usan variables estandarizadas, el ACP se hace sobre la matriz de correlaciones, en la que la diagonal está compuesta de unos. Entonces, se sabe que la

suma de los valores característicos lambdas es igual a la traza de la matriz, Y dado que la matriz tiene solo unos en la diagonal, la traza es igual a p.

Pregunta 10.

Se le da el archivo enignorte que contiene datos de la Encuesta Nacional de Ingresos y Gastos de los Hogares 2018, referente a las zonas urbanas de la Región Norte. Se quiere predecir el gasto en alimentos en función del ingreso total del hogar, la cantidad de miembros del hogar y la edad del jefe. Los nombres de las variables se tienen a continuación

```
load("enignorte.Rdata")
names(enignorte)
```

```
## [1] "ID_REGION"     "ID_ZONA"       "id"            "gastoalim"
## [5] "ingresototal"  "miembroshogar" "edadjefe"      "resid"
```

Con base en dicha base de datos conteste las siguientes preguntas:

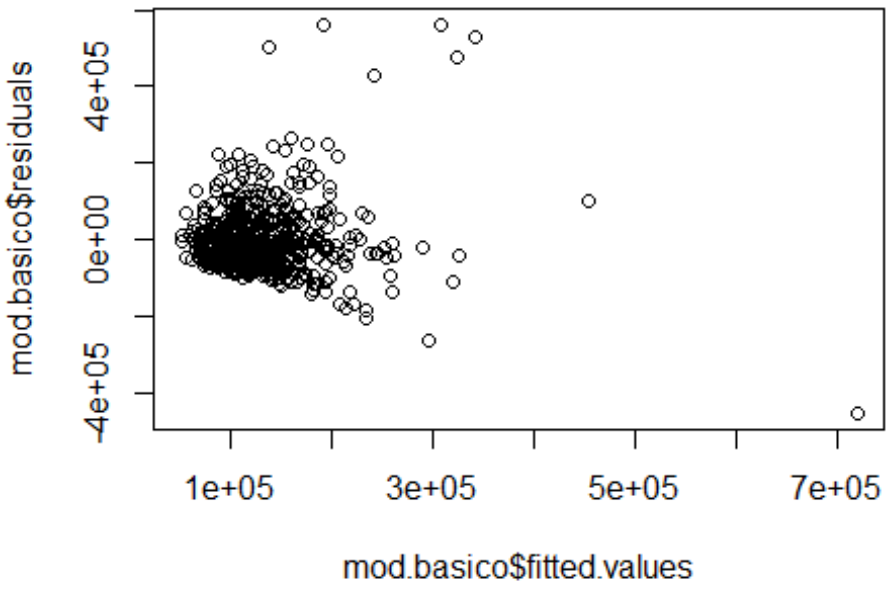
a) *Estime un modelo de regresión gaussiano en el que se prediga el gasto en alimentos en función del ingreso total del hogar, la cantidad de miembros del hogar y la edad del jefe. Muestre el summary del modelo.*

```
mod.basico <- lm(gastoalim~ingresototal+miembroshogar+edadjefe, data = en
ignorte)
summary(mod.basico)
```

```
##
## Call:
## lm(formula = gastoalim ~ ingresototal + miembroshogar + edadjefe,
##     data = enignorte)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -452876  -52673  -15869   35875  561053
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   15950.78   15817.75   1.008   0.3137
## ingresototal    280.38      27.26  10.286  < 2e-16 ***
## miembroshogar 18783.19    2418.55   7.766 3.94e-14 ***
## edadjefe        560.66     246.45   2.275   0.0233 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93020 on 553 degrees of freedom
## Multiple R-squared:  0.252,  Adjusted R-squared:  0.2479
## F-statistic:  62.1 on 3 and 553 DF,  p-value: < 2.2e-16
```

b) *Haga un gráfico de residuos contra predichos y diga cómo evalúa el supuesto de homoscedasticidad.*

```
plot(mod.basico$residuals~mod.basico$fitted.values)
```



Hay heteroscedasticidad.

c) *Estime un modelo de mínimos cuadrados ponderados con dos iteraciones y en que los residuos absolutos sean predichos por los predictores para calcular los ponderadores. Muestre el summary del modelo final.*

Minimos Cuadrados Ponderados:

```
abs.res1   <- abs(residuals(mod.basico))
mod.ponde1 <- lm(abs.res1~ingresototal+miembroshogar+edadjefe, data = eni
gnorte)
ponde1     <- 1/abs(fitted(mod.ponde1))
mod.iter1  <- lm(gastoalim~ingresototal+miembroshogar+edadjefe, data = en
ignorte, weights = I(ponde1))

abs.res2   <- abs(residuals(mod.iter1))
mod.ponde2 <- lm(abs.res2~ingresototal+miembroshogar+edadjefe, data = eni
gnorte)
ponde2     <- 1/abs(fitted(mod.ponde2))
mod.iter2  <- lm(gastoalim~ingresototal+miembroshogar+edadjefe, data = en
ignorte, weights = I(ponde2))

cbind(mod.basico$coefficients, mod.iter1$coefficients, mod.iter2$coeffici
ents)
```

```
##                     [,1]       [,2]       [,3]
## (Intercept)   15950.7798 20940.0528 21166.9426
## ingresototal    280.3805   333.3651   335.4608
## miembroshogar 18783.1906 18376.1491 18370.8753
## edadjefe        560.6566   375.4249   366.7524
```

```
summary(mod.iter2)
```

```
##
## Call:
## lm(formula = gastoalim ~ ingresototal + miembroshogar + edadjefe,
##     data = enignorte, weights = I(ponde2))
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -760.03 -227.96  -56.97  154.36 2149.36
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21166.94   13010.98   1.627   0.1043
## ingresototal    335.46      40.97   8.188 1.85e-15 ***
## miembroshogar 18370.88    2070.25   8.874  < 2e-16 ***
## edadjefe        366.75     199.95   1.834   0.0672 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 332.8 on 553 degrees of freedom
## Multiple R-squared:  0.2349, Adjusted R-squared:  0.2307
## F-statistic: 56.59 on 3 and 553 DF,  p-value: < 2.2e-16
```

d) *Estime un modelo hesteroscedástico (un modelo lineal generalizado gaussiano doble) con las variables predictoras en las ecuaciones de la media y la variancia, y a un 5% de significancia diga si hay variables asociadas con heteroscedasticidad.*

Modelos heteroscedastico:

```
library(dglm)
mod.het <- dglm(gastoalim~ingresototal+miembroshogar+edadjefe, dformula =
~ingresototal+miembroshogar+edadjefe,
                family ="gaussian", data = enignorte)
summary(mod.het)
```

```
##
## Call: dglm(formula = gastoalim ~ ingresototal + miembroshogar + edadje
fe,
##     dformula = ~ingresototal + miembroshogar + edadjefe, family = "gau
ssian",
##     data = enignorte)
##
## Mean Coefficients:
##                 Estimate  Std. Error  t value     Pr(>|t|)
```

```
## (Intercept)   19639.4340 12154.90545 1.615762 1.067161e-01
## ingresototal    328.5931    54.39434 6.040943 2.815304e-09
## miembroshogar 19798.8610  2140.18264 9.251015 4.820676e-19
## edadjefe        318.5321   179.32297 1.776304 7.623261e-02
## (Dispersion Parameters for gaussian family estimated as below )
##
##     Scaled Null Deviance: 705.698 on 556 degrees of freedom
## Scaled Residual Deviance: 556.9994 on 553 degrees of freedom
##
## Dispersion Coefficients:
##                   Estimate   Std. Error    z value     Pr(>|z|)
## (Intercept)   21.419725231 0.2404800065 89.0707113 0.000000e+00
## ingresototal   0.004907503 0.0004144072 11.8422237 2.361086e-32
## miembroshogar  0.155535282 0.0367696719  4.2299883 2.337035e-05
## edadjefe       0.001938826 0.0037468298  0.5174576 6.048367e-01
## (Dispersion parameter for Gamma family taken to be 2 )
##
##     Scaled Null Deviance: 1076.987 on 556 degrees of freedom
## Scaled Residual Deviance: 865.1128 on 553 degrees of freedom
##
## Minus Twice the Log-Likelihood: 14114.23
## Number of Alternating Iterations: 6

round(summary(mod.het)$coefficients, 4)

##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19639.4340 12154.9054  1.6158   0.1067
## ingresototal    328.5931    54.3943  6.0409   0.0000
## miembroshogar 19798.8610  2140.1826  9.2510   0.0000
## edadjefe        318.5321   179.3230  1.7763   0.0762
```

Con un 5% de signicancia hay suficiente evidencia estadística para rechazar la hipótesis nula de homocedasticidad para las variables ingresototal y miembroshogar, manteniendo las demás variables constantess. Es decir, estas dos variables están asociadas con la heterocedaticidad.

e) *Haga un cuadro en el que compare los coeficientes y los p-values de los 3 modelos (guassiano en inciso (a), mínimos cuadrados ponderados en inciso (c) y heteroscedástico en inciso (d)), y diga en qué se parecen y se diferencian las conclusiones de los tres modelos (use un alfa de 5%).*

```
cuadro.het <- as.matrix(cbind(summary(mod.basico)$coefficients[, c(1, 4)]
, summary(mod.iter2)$coefficients[, c(1, 4)],
                              summary(mod.het)$coefficients[, c(1, 4)]))

colnames(cuadro.het) <- c("Basico.coef", "Basico.pvalue", "MC Ponderados.
coef",
                          "MCPonderados.pvalue", "DGLM.coef", "DGLM.pvalu
e")
```

```
round(cuadro.het, 2)
```

```
##                Basico.coef Basico.pvalue MC Ponderados.coef MCPonderado
s.pvalue
## (Intercept)      15950.78          0.31           21166.94
0.10
## ingresototal       280.38          0.00             335.46
0.00
## miembroshogar    18783.19          0.00           18370.88
0.00
## edadjefe           560.66          0.02             366.75
0.07
##                DGLM.coef DGLM.pvalue
## (Intercept)    19639.43        0.11
## ingresototal     328.59        0.00
## miembroshogar  19798.86        0.00
## edadjefe         318.53        0.08
```

En el primer modelo gaussiano del inciso a) con un alfa del 5% las variables ingresototal, miembroshogar y edadjefe son significativamente distintas de 0, en el modelo de mínimos cuadrados ponderados del inciso c) con un alfa del 5% las variables ingresototal y miembroshogar son significativamente distintas de 0 y por último en el modelo heterocedástico del inciso d) se llega a la misma conclusión, con un alfa del 5% las variables ingresototal y miembros hogar son significativamente distintas de 0.

Por lo tanto en términos de p-value se puede ver que se parecen mucho los tres modelos, la única diferencia notable es que en el modelo original había un coeficiente extra significativamente distinto de 0.

En términos de la magnitud de los coeficientes en el caso de los interceptos, en el modelo del inciso a) si las demás variables independientes son 0 el gasto en alimentación va a ser de 15.950 colones mientras que en el modelo del inciso c) va a ser de 21.166 colones y en el modelo del inciso d) es de 19.639 colones. Por lo tanto se puede apreciar que los coeficientes de los interceptos pese a ser distintos su diferencia no es relevante grande.

f) *En términos del procedimiento de estimación, diga en qué se parecen en general los modelos estimados con mínimos cuadrados ponderados y los modelos heteroscedásticos.*

En Mínimos Cuadrados Ponderados, para obtener los ponderadores, se estima una regresión en el que el estimador de la variancia (los residuos) está en función de las variables predictoras. En el modelo heteroscedástico, existe una de las ecuaciones en las que la variancia está en función de las variables predictoras.