Práctica 3 - Modelos Lineales 2

Justo Manrique Urbina - 20091107; Juan Pablo Moreano - 20184093

12/03/2019

Pregunta 1

Pregunta 1.a)

Tenemos que la función de verosimilitud para $Y_j \sim Bernoulli(\mu_j)$ es la siguiente:

$$L(\cdot) = \prod_{j=1}^{n} \mu_j^{y_j}(1-\mu_j)^{(1-y_j)}.$$

Posteriormente, hallamos la log-verosimilitud de los datos, la cual es:

$$l(\cdot) = \sum_{j=1}^{n} y_j \log(u_j) + (1-y_j)\log(1-\mu_j).$$

$$= \sum_{j=1}^{n} y_j \log(\mu_j) + \log(1-\mu_j) - y_j \log(1-\mu_j).$$

$$= \sum_{j=1}^{n} \log(1-\mu_j) + y_j \log(\frac{\mu_j}{1-\mu_j}).$$

$$= \sum_{j=1}^{n} \log(1-\mu_j) + y_j \beta x_j.$$

Recordemos que:

$$\mu_j = \frac{1}{1+e^{-\beta x_j}}.$$

Por lo tanto, tenemos que:

$$l(\cdot) = \sum_{j=1}^{n} \log(e^{-\beta x_j}) - \log(1+e^{-\beta x_j}) + y_j \beta x_j.$$

Y finalmente:

$$l(\cdot) = \sum_{j=1}^{n} -\log(1+e^{-\beta x_j}) + \beta x_j (y_j - 1).$$

La función de Score es:

$$\sum_{j=1}^{n} -\frac{e^{-\beta x_j}}{1+e^{-\beta x_j}} + x_j(y_j-1).$$

$$\sum_{j=1}^{n} -\frac{1}{e^{\beta x_j}+1} + x_j(y_j-1).$$

La información de Fisher es:

$$\sum_{j=1}^{n} \frac{x_j e^{\beta x_j}}{(e^{\beta x_j}+1)^2}$$

La ecuación que debe resolverse para hallar β es la siguiente:

$$\sum_{j=1}^{n} -\frac{1}{e^{\beta x_j}+1} + x_j(y_j - 1) = 0.$$

Pregunta 2

Los datos en el archivo peso.csv consisten de 189 observaciones de 10 variables, y las siguientes variables:

- low: variable indicadora del bajo peso al nacimiento del infante ('1'=peso por debajo 2.5 kg, '0'peso desde 2.5 kg ó más)
- smoke: estatus de fumadora de la madre ('0'=no fumadora, '1'=fumadora)
- ht: historia de hipertensión de la madre ('1'=historia de hipertensión, '0'=no hipertensión)
- lwt: Peso de la madre en el último periodo, valor numérico en libras.

El objetivo del estudio es modelar el peso de nacimiento bajo para los infantes.

Pregunta 2.a)

Analice los datos y describa en detalle sus modelos ajustados en R.

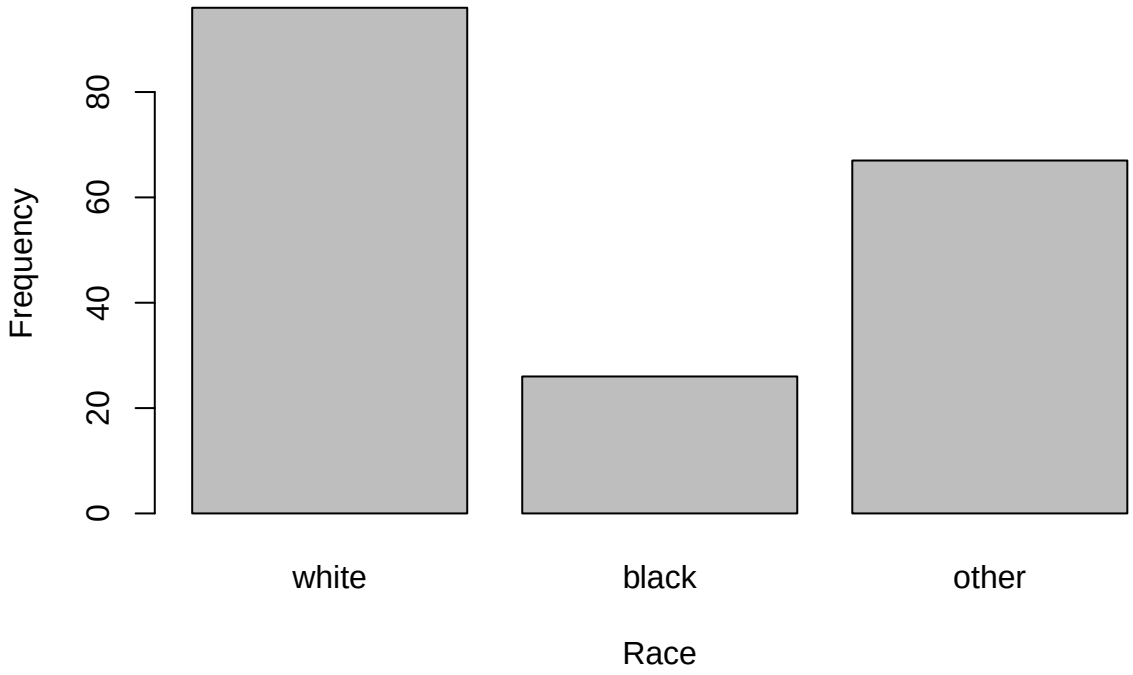
```
setwd("~/Documents/maestria-pucp/2019-2/modelos-lineales-2/practica-3")
datos <- read.csv("peso.csv", sep=",")
```

```
#Convertir a factor las variables
datos$low <- factor(datos$low, levels = c(0,1),labels = c("No", "Yes"))
datos$race <- factor(datos$race, levels = c(1:3), labels=c("white","black","other"))
datos$smoke <- factor(datos$smoke, levels = c(0,1), labels = c("No", "Yes"))
datos$ht <- factor(datos$ht, levels = c(0,1), labels = c("No", "Yes"))
datos$ui <- factor(datos$ui, levels = c(0,1),labels = c("No", "Yes"))
datos$ptl <- factor(datos$ptl)
datos$ftv <- factor(datos$ftv)
```

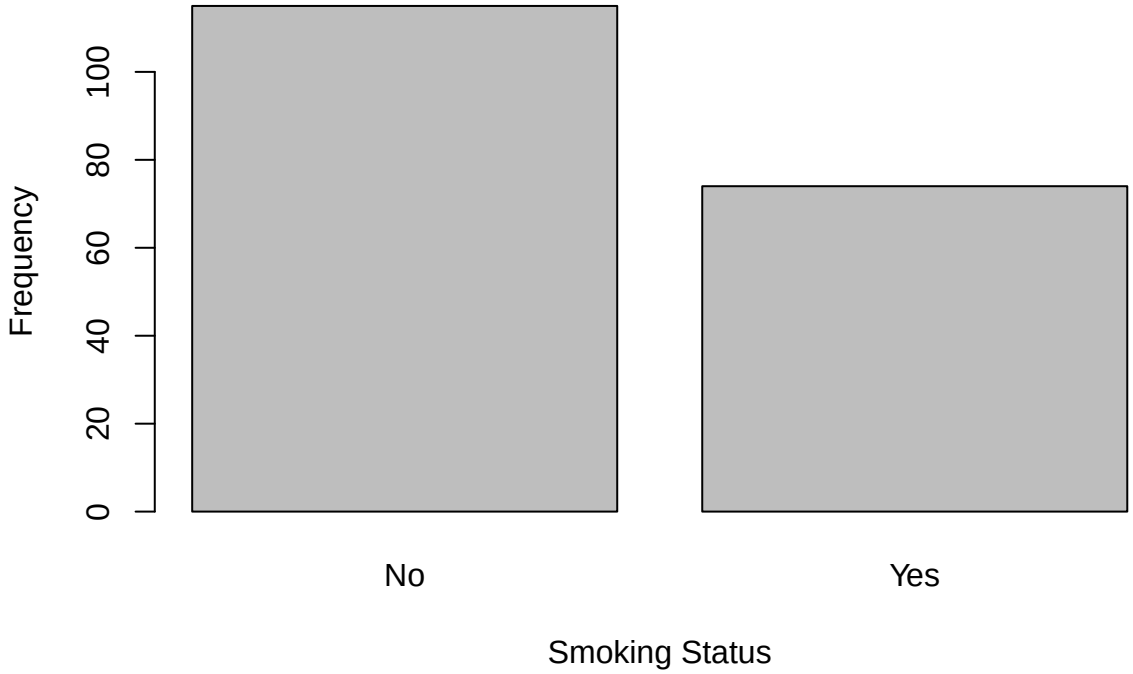
Análisis Exploratorio

Gráfico de barras de cada variable categórica (low, race, smoke)

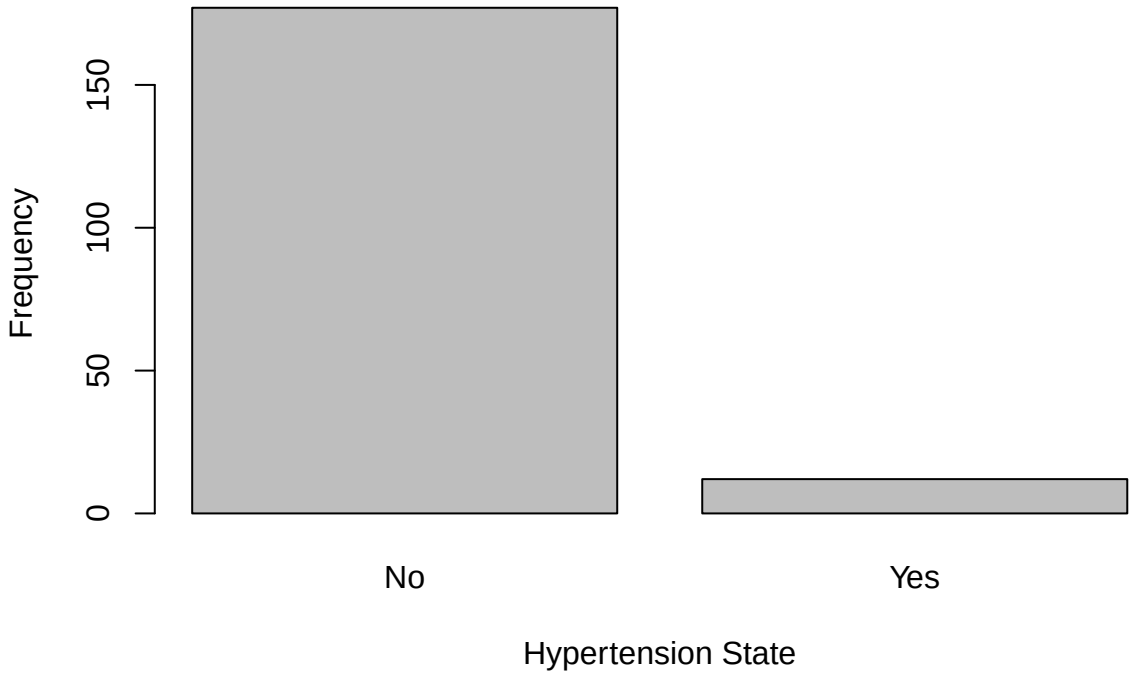
```
barplot(table(datos$race), xlab= "Race", ylab="Frequency")
```



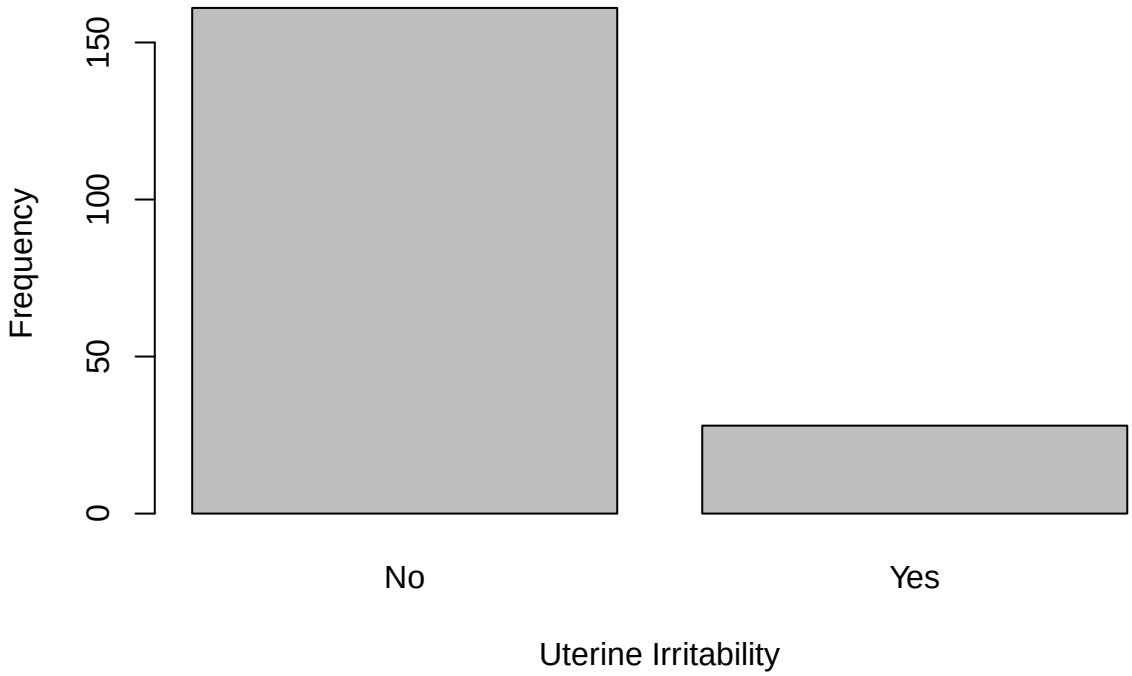

```
barplot(table(datos$smoke), xlab= "Smoking Status", ylab="Frequency")
```




```
barplot(table(datos$ht), xlab= "Hypertension State", ylab="Frequency")
```




```
barplot(table(datos$ui), xlab= "Uterine Irritability", ylab="Frequency")
```




```
barplot(table(datos$ptl), xlab= "Previous Premature Labors", ylab="Frequency")
```

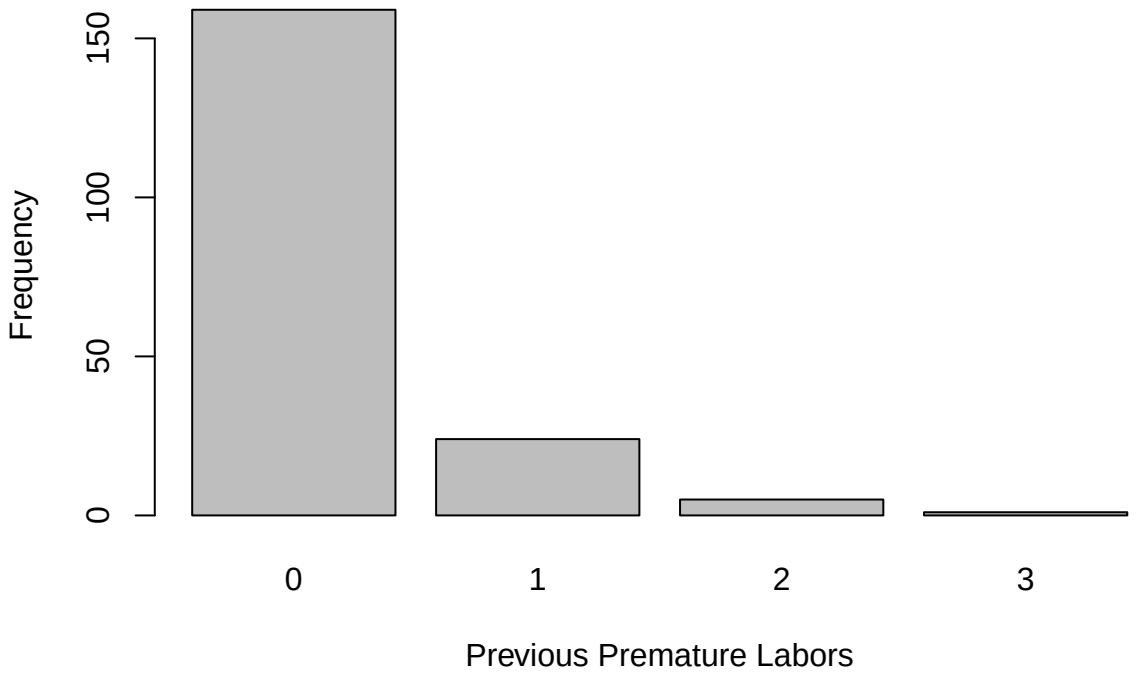
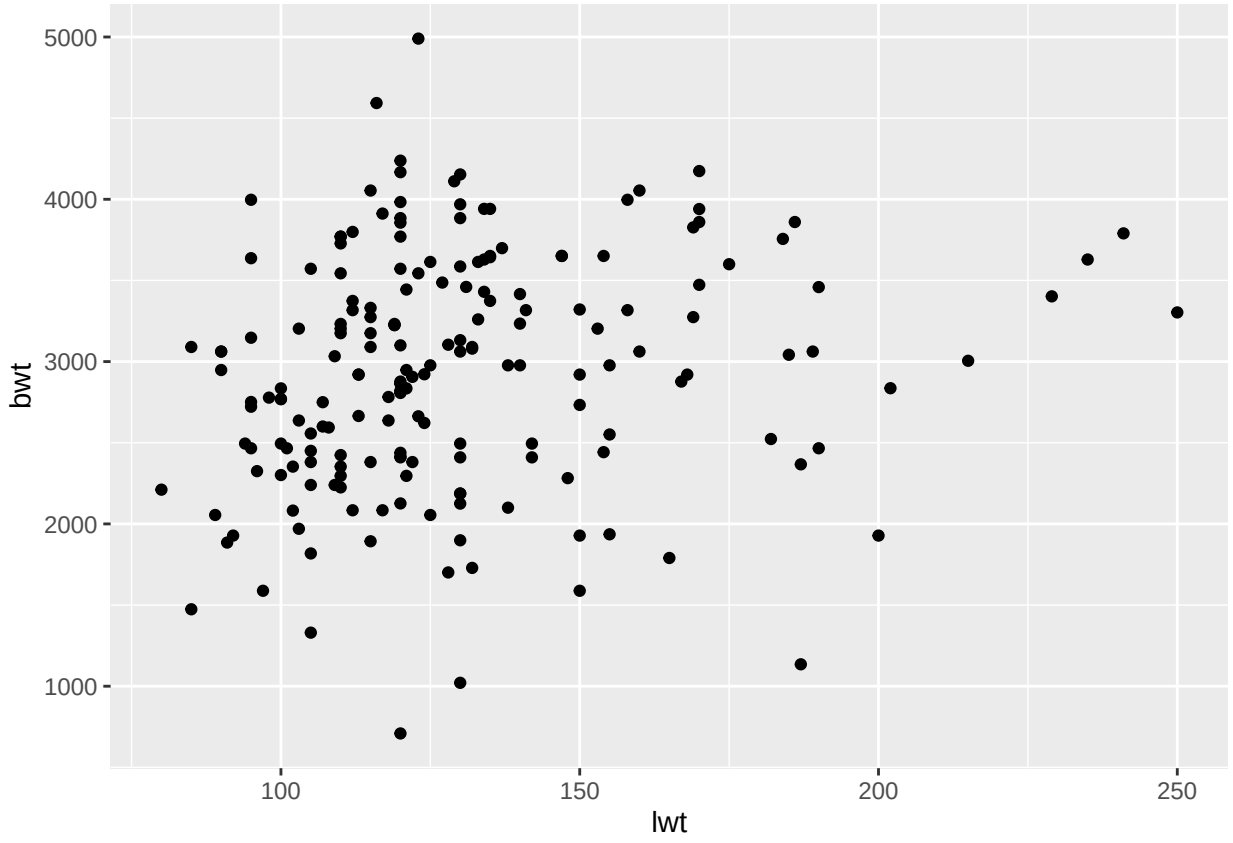


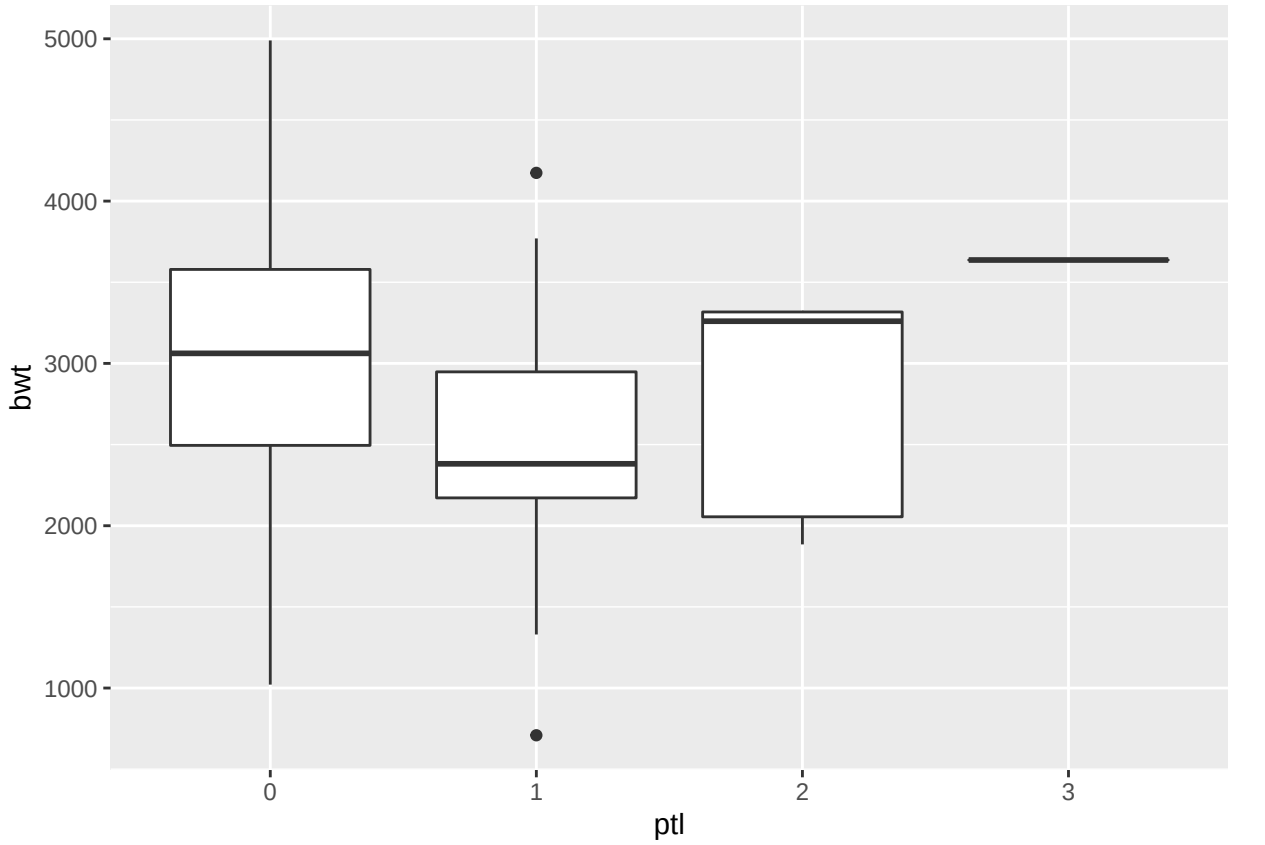
Diagrama de dispersión del peso de la madre (lwt) frente al peso al nacer (bwt)

```
library(ggplot2)
ggplot(data=datos, aes(x=lwt, y=bwt)) +
  geom_point()
```

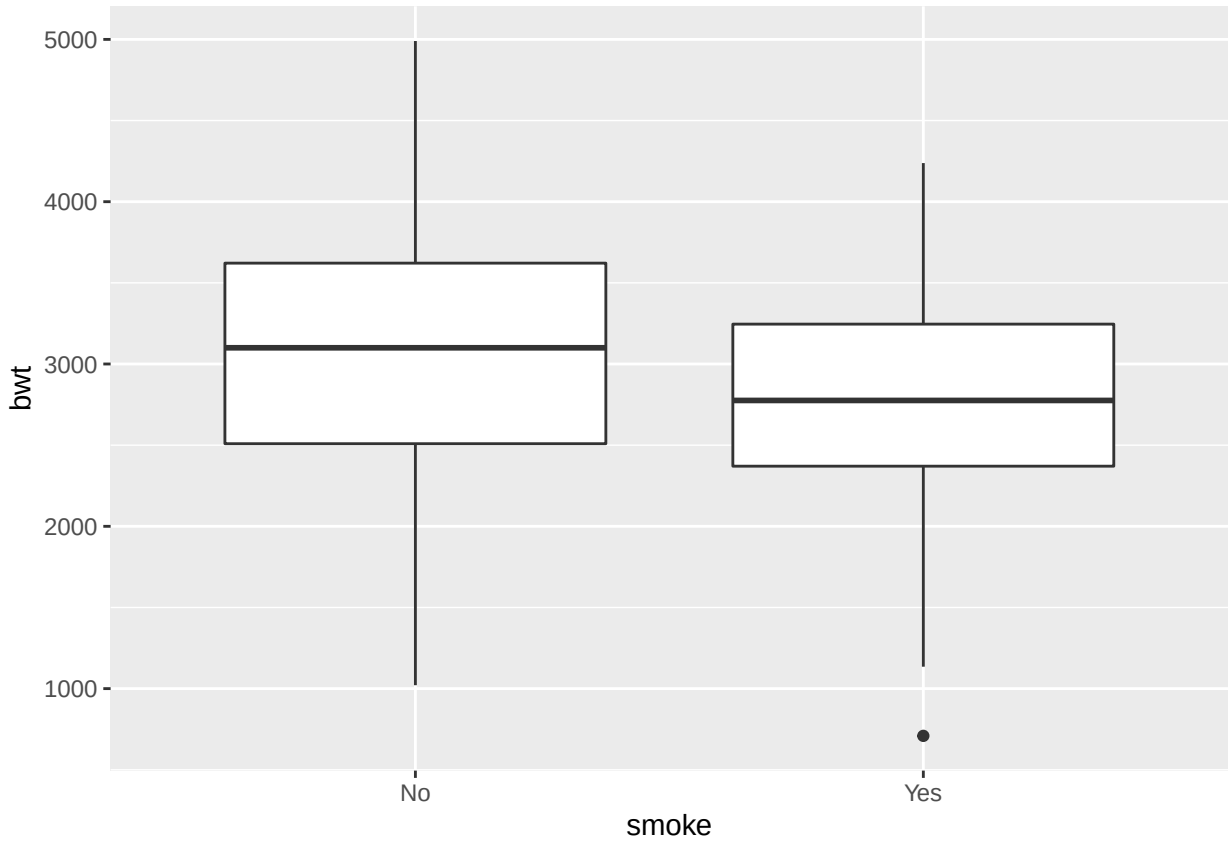


Diagramas de caja de variables categóricas en relación con el peso al nacer

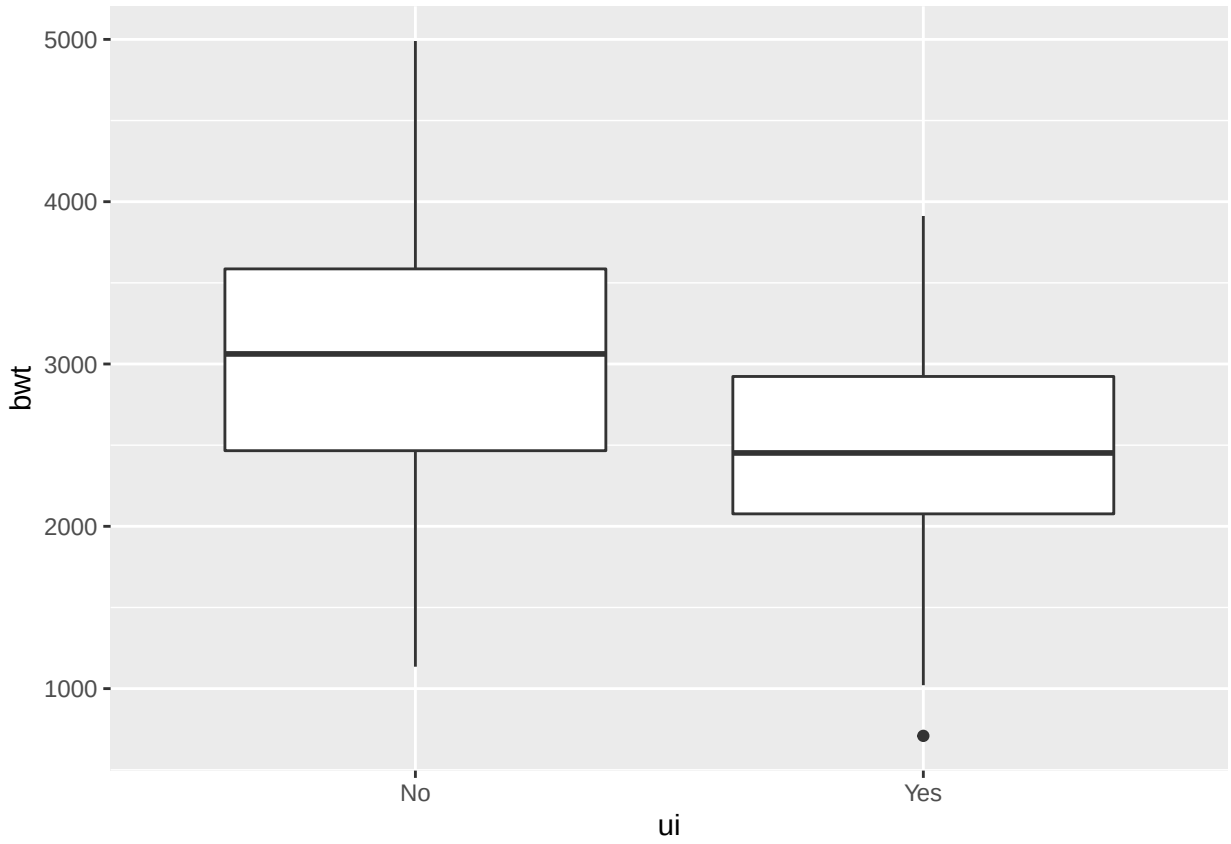
```
ggplot(data=datos, aes(x=ptl, y=bwt)) + geom_boxplot()
```




```
ggplot(data=datos, aes(x=smoke, y=bwt)) + geom_boxplot()
```




```
ggplot(data=datos, aes(x=ui, y=bwt)) + geom_boxplot()
```




```
ggplot(data=datos, aes(x=race, y=bwt)) + geom_boxplot()
```

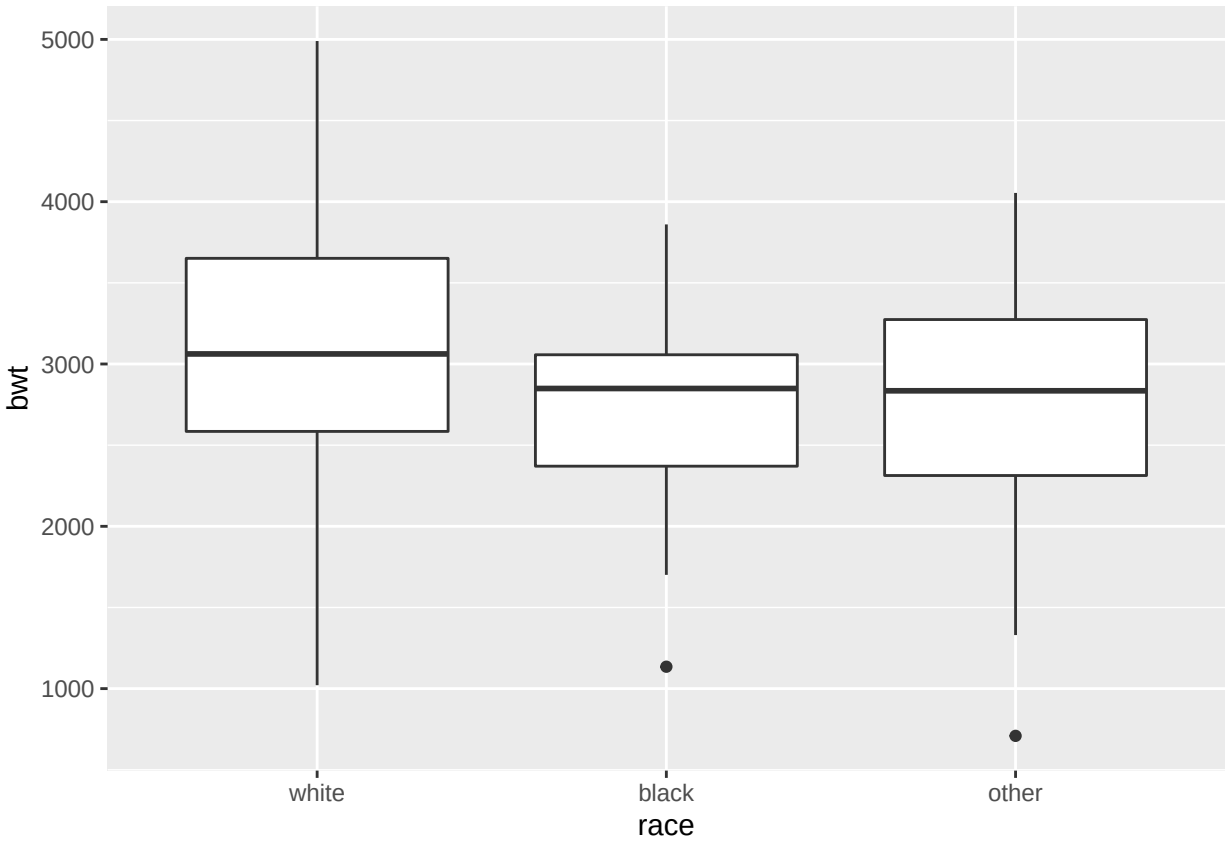


Gráfico de barras de cada variable categórica (low, race, smoke)

```
library(grid)
library(gridExtra)
grid.arrange(
  ggplot(data=datos, mapping=aes(x=low, fill=low)) +
    geom_bar(),
  ggplot(data=datos, mapping=aes(x=low, y = ..prop.., group=1)) +
    geom_bar(),
  ggplot(data=datos, mapping=aes(x=race, fill=race)) +
    geom_bar(),
  ggplot(data=datos, mapping=aes(x=race, y = ..prop.., group=1)) +
    geom_bar(),
  ggplot(data=datos, mapping=aes(x=smoke, fill=smoke)) +
    geom_bar(),
  ggplot(data=datos, mapping=aes(x=smoke, y = ..prop.., group=1)) +
    geom_bar(),
  nrow=3,
  top="Bar plots para Variables categoricas")
```

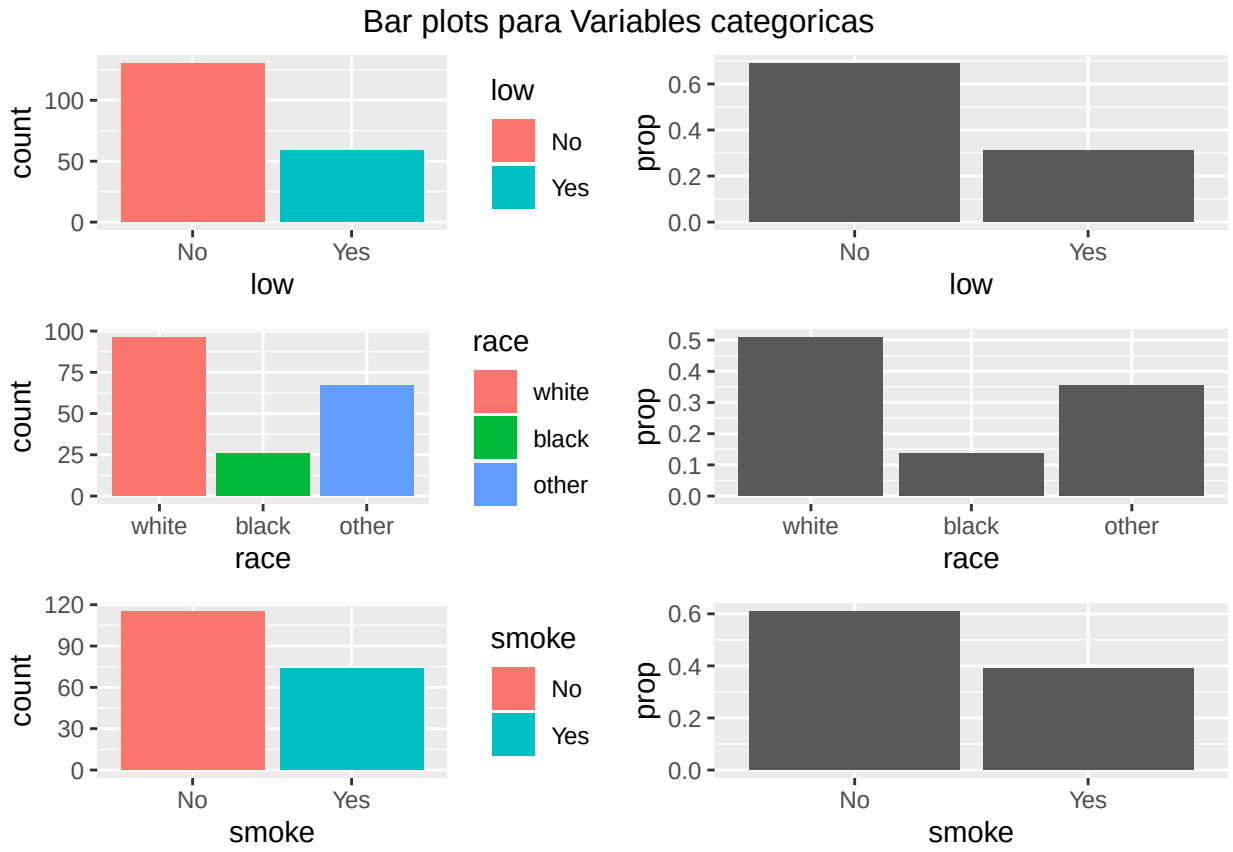
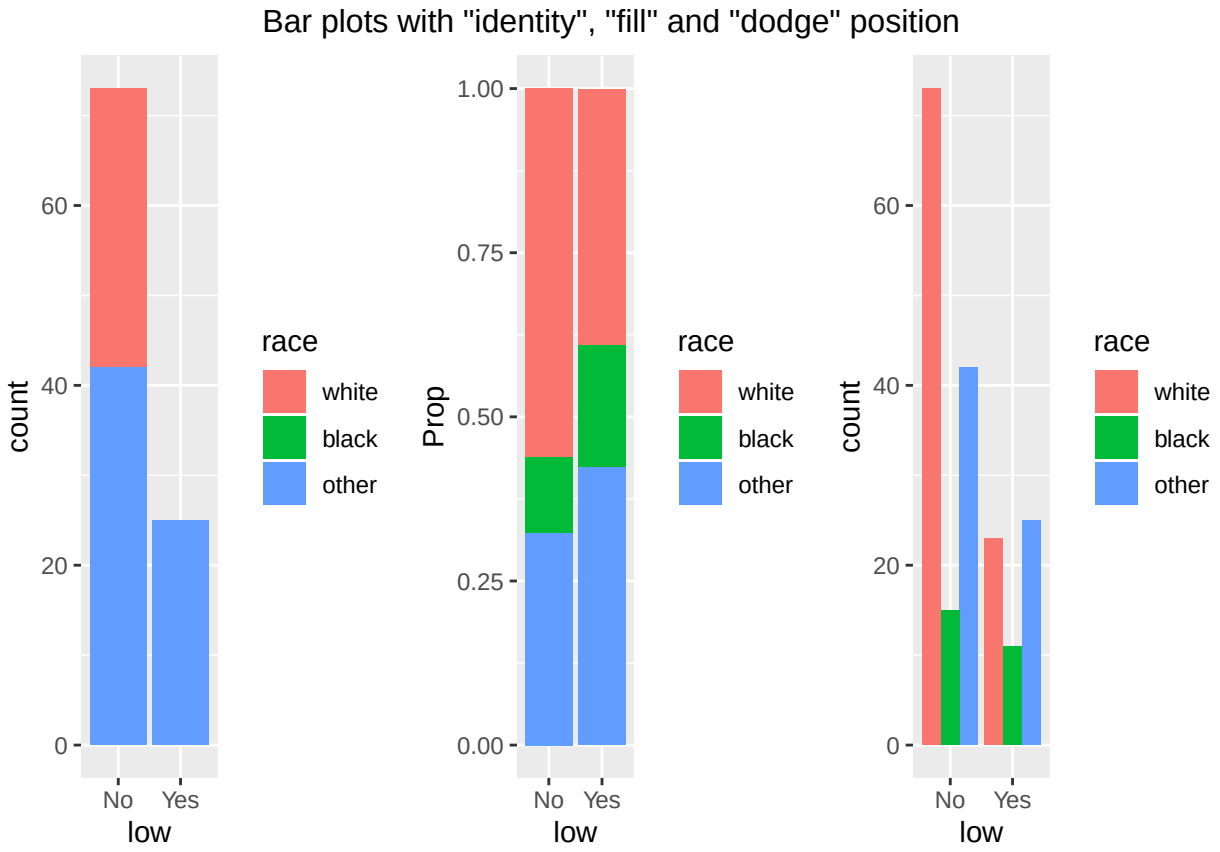


Gráfico de barras de la variable low en relación con otras variables categóricas

```
grid.arrange(
  ggplot(data=datos, mapping=aes(x=low, fill=race)) +
    geom_bar(position = "identity"),
  ggplot(data=datos, mapping=aes(x=low, fill=race)) +
    geom_bar(position = "fill") +
    scale_y_continuous(name="Prop"),
  ggplot(data=datos, mapping=aes(x=low, fill=race)) +
    geom_bar(position = "dodge" ),
  nrow=1,
  top='Bar plots with "identity", "fill" and "dodge" position ')
```



Se procederá a seleccionar cuatro variables del conjunto de datos para realizar este análisis.

La variable dependiente "low" es un indicador de peso al nacer menor a 2.5 kg en el cual "1" un peso al nacer poco saludable menor a 2.5 kg y "0" representa un pesosaludable al nacer superior a 2.5 kg.

Las variables independientes utilizadas en este estudio son "smoke", "race" y "ht". La variable "smoke" midió si la madre fumaba durante el embarazo, la variable raza se recodificó para "1" para representar el blanco y "0" para representar todas las demás variables y la variable "ht" mide si la madre tenía antecedentes de hipertensión o no.

```
library(tidyverse)
library(magrittr)
library(dplyr)
library(Zelig)
library(pander)
library(texreg)
library(visreg)
library(lmtest)
library(visreg)
library(MASS)
```

Se procedió a realizar una modificación en el csv original con el siguiente comando: $datos2 < -datos$

Y con éste cambio procemos a exponer los modelos desarrollados

```
datos_final <- read.csv("peso_final.csv", sep=",")
```

Primer Modelo Propuesto

En el primer modelo, se buscó determinar si un niño que nace con un peso inferior al normal es afectado por

la variable 'raza'. Mirando los resultados podemos concluir que ser blanco tuvo un efecto negativo sobre si un niño nace por debajo del peso normal (−0.6954) , viendo los pvalue observamos que los resultados son estadísticamente significativos.

```
modelo1 <- glm(low ~ race, family = binomial, data = datos_final)
summary(modelo1)
```

```
##
## Call:
## glm(formula = low ~ race, family = binomial, data = datos_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9895  -0.9895  -0.7401   1.3777   1.6905
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4595     0.2129  -2.159   0.0309 *
## racewhite    -0.6954     0.3202  -2.172   0.0298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 229.86  on 187  degrees of freedom
## AIC: 233.86
##
## Number of Fisher Scoring iterations: 4
```

Segundo Modelo Propuesto

En el segundo modelo se procedió a agregar la variable 'humo', en éste modelo podemos ver que las madres que fumaron durante el embarazo tenían una mayor probabilidad (1.1130) de dar a luz a un bebé que estaría por debajo del peso normal al nacer de 2.5kg , ésta relación es significativa a un nivel de confianza de 0.o1

```
modelo2 <- glm(low ~ race + smoke, family = binomial, data = datos_final)
summary(modelo2)
```

```
##
## Call:
## glm(formula = low ~ race + smoke, family = binomial, data = datos_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3402  -0.8840  -0.5433   1.4968   1.9930
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.7382     0.2379  -3.103  0.00191 **
## racewhite    -1.1003     0.3645  -3.019  0.00254 **
## smoke         1.1130     0.3643   3.056  0.00225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 219.98  on 186  degrees of freedom
## AIC: 225.98
##
## Number of Fisher Scoring iterations: 4
```

Tercer Modelo Propuesto

Ahora se procede a agregar a la variable que hace referencia a 'hipertensión', tener antecedentes de tabaquismo y presión arterial alta puede aumentar las posibilidades de que el peso al nacer de un niño esté por debajo de lo normal.

En este modelo, vemos que la hipertensión aumenta la probabilidad de que el peso al nacer de un niño sea inferior a 2.5 kg es (1.1725) , sin embargo este indicador no es estadísticamente significativo con pvalue > 0.05

```
modelo3 <- glm(low ~ race + smoke + hypertension, family = binomial, data = datos_final)
summary(modelo3)
```

```
##
## Call:
## glm(formula = low ~ race + smoke + hypertension, family = binomial,
##     data = datos_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3371  -0.8496  -0.5222   1.0587   2.0297
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.8331     0.2460  -3.386 0.000708 ***
## racewhite     -1.0904     0.3686  -2.958 0.003095 **
## smoke          1.1190     0.3681   3.040 0.002367 **
## hypertension   1.1725     0.6225   1.883 0.059633 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 216.38  on 185  degrees of freedom
## AIC: 224.38
##
## Number of Fisher Scoring iterations: 4
```

Cuarto Modelo Propuesto

Ahora examinamos la variable 'raza' y la interacción entre el tabaquismo e hiperesión para analizar el impacto que tiene sobre la probabilidad de que un niño nazca por debajo del peso noral de 2.5kg.

El modelo muestra que las personas de raza blanca son aún más propensas a tener un hijo que tiene un peso normal al nacer y que el efecto de fumar es negativo , sin embargo no hay relación entre la hipertensión y tener un hijo de menor peso al nacer.

```
modelo4 <- glm(low ~ race + smoke * hypertension, family = binomial, data = datos_final)
summary(modelo4)
```

```
##
```

```
## Call:
## glm(formula = low ~ race + smoke * hypertension, family = binomial,
##     data = datos_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3706  -0.8456  -0.5213   1.0561   2.0314
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.8446     0.2518  -3.354 0.000797 ***
## racewhite           -1.0828     0.3700  -2.926 0.003431 **
## smoke                1.1367     0.3766   3.018 0.002544 **
## hypertension         1.2881     0.8128   1.585 0.113016
## smoke:hypertension  -0.2821     1.2652  -0.223 0.823550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 216.33  on 184  degrees of freedom
## AIC: 226.33
##
## Number of Fisher Scoring iterations: 4
```

Pregunta 2.b)

Defina y analice la o las pruebas de hipótesis pertinentes para evaluar cuál de los dos modelos presenta un mejor ajuste para los datos. Evalue la bondad de ajuste del modelo escogido.

Likelihood Ratio Test

La prueba ilustra que el $Modelo4$ es el mejor ajuste porque tiene la desviación más pequeña. Una desviación menor en el Modelo 4 significa que se ajusta mejor a los datos.

```
anova(modelo1, modelo2, modelo3, modelo4, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: low ~ race
## Model 2: low ~ race + smoke
## Model 3: low ~ race + smoke + hypertension
## Model 4: low ~ race + smoke * hypertension
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       187     229.86
## 2       186     219.98  1   9.8802 0.001671 **
## 3       185     216.38  1   3.5949 0.057956 .
## 4       184     216.33  1   0.0495 0.823892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De la siguiente prueba, vemos que el segundo modelo es el más adecuado porque tiene el AIC bajo (225.98) y el BIC más bajo (235.70), esto puede deberse al hecho de que se encuentra que la hipertensión no era un indicador relevante.

```
screenreg(list(modelo1, modelo2, modelo3, modelo4))
```

```
##
## ========================================================================
##                      Model 1    Model 2     Model 3      Model 4
## ------------------------------------------------------------------------
## (Intercept)           -0.46 *    -0.74 **    -0.83 ***    -0.84 ***
##                       (0.21)     (0.24)      (0.25)       (0.25)
## racewhite             -0.70 *    -1.10 **    -1.09 **     -1.08 **
##                       (0.32)     (0.36)      (0.37)       (0.37)
## smoke                             1.11 **     1.12 **      1.14 **
##                                   (0.36)      (0.37)       (0.38)
## hypertension                                   1.17         1.29
##                                                (0.62)       (0.81)
## smoke:hypertension                                          -0.28
##                                                             (1.27)
## ------------------------------------------------------------------------
## AIC                   233.86     225.98      224.38       226.33
## BIC                   240.34     235.70      237.35       242.54
## Log Likelihood      -114.93    -109.99     -108.19      -108.17
## Deviance              229.86     219.98      216.38       216.33
## Num. obs.             189        189         189          189
## ========================================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

Pregunta 2.c)

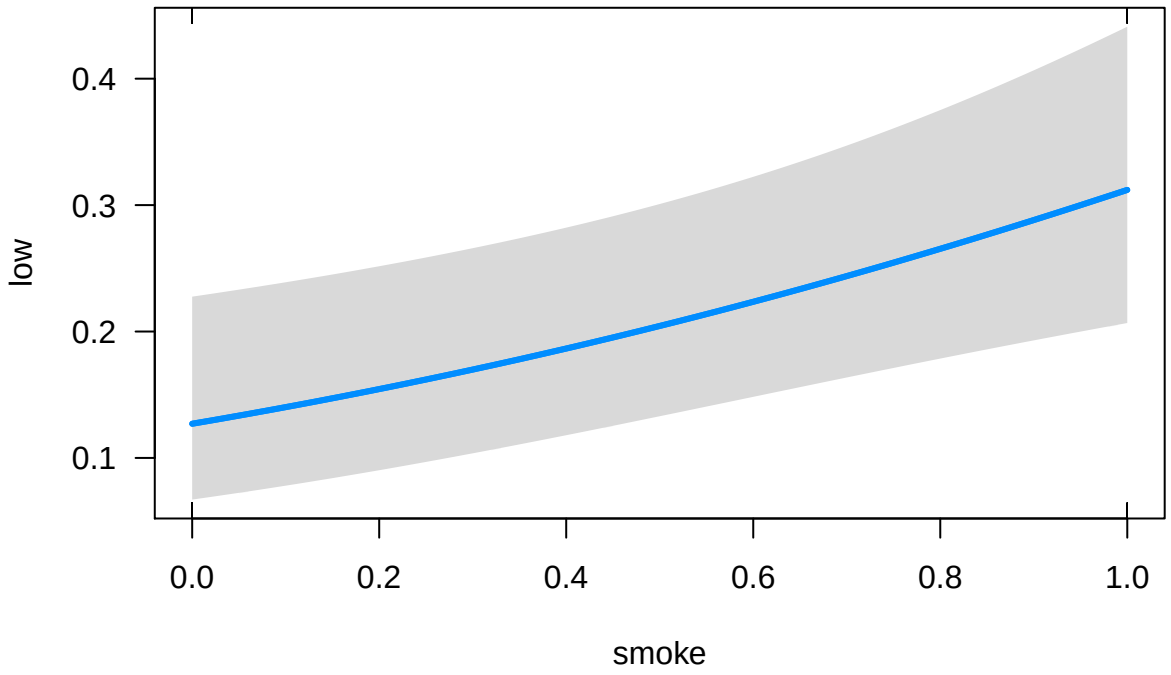
Interprete el coeficiente de regresión de la variable explicativa smoke para el modelo escogido.

Los gráficos muestran la relación por la cual si las madres que respondieron afirmativamente al hábito de fumar durante el embarazo tenían una mayor probabilidad de nacer con un hijo que estaba por debajo del peso normal.

En el modelo elegido podemos ver que las madres que fumaron durante el embarazo tenían una mayor probabilidad (1.1367) de dar a luz a un bebé que estaría por debajo del peso normal al nacer de 2.5kg.

```
library(visreg)
visreg(modelo4, "smoke", scale = "response")
```

```
## Conditions used in construction of plot
## race: white
## hypertension: 0
```



Pregunta 2.d)

Considere una madre que fumaba durante el embarazo, tiene una historia de hipertensión y tiene un peso de 130 (libras) en el último periodo. Describa cómo se realiza una predicción sobre la probabilidad de que el infante tenga peso bajo. ¿Cuál es dicha probabilidad para el modelo escogido?

Nota: si su modelo final, tiene más variables explicativas, puede asumir valores viables para ellas y realizar la predicción solicitada

Se asumió, aparte de lo indicado en la pregunta que la raza de la madre era no-blanca. Ver a continuación el código R:

```
pred <- data.frame(smoke=1,hypertension=1,race ="non-white")
predict.glm(modelo4,newdata = pred,type = "response")
```

```
##         1
## 0.7855136
```

La probabilidad de que el infante tenga peso bajo es de 78.55%.

Pregunta 2.e)

Evalue la capacidad predictiva de los modelos ajustados.

Para evaluar la capacidad predictiva de los modelos, evaluamos el AIC de cada modelo. Ver código R a continuación.

```
cbind(
  rbind(AIC(modelo1),AIC(modelo2),AIC(modelo3),AIC(modelo4))
  )
```

```
##          [,1]
## [1,] 233.8573
## [2,] 225.9772
## [3,] 224.3822
## [4,] 226.3327
```

Observamos que el modelo 3 tiene mejor capacidad predictiva, pues tiene el AIC más bajo. Luego de ello, el modelo 2 y 4 y finalmente el 1.

Pregunta 3

Pregunta 3.a)

¿Cuál de los cuatro modelos Ud. escogería para analizar el conjunto de datos? Justifique su elección del modo más amplio posible, en base a los resultados presentados.

Para ello, primero cargamos los datos y luego generamos las regresiones en base a las características solicitadas:

```
p3 = read.csv("SLID-Ontario.csv")

p3 = cbind(p3,x = p3$yearsEducation-mean(p3$yearsEducation))
p3 = cbind(p3,x2 = p3$x^2)

mod1 <- glm(compositeHourlyWages ~ x,family = gaussian(link="identity"),data=p3)
mod2 <- glm(compositeHourlyWages ~ x + x2,family = gaussian(link="identity"),data=p3)
mod3 <- glm(compositeHourlyWages ~ x,family = Gamma(link="log"),data = p3)
mod4 <- glm(compositeHourlyWages ~ x + x2,family = Gamma(link="log"),data = p3)
```

Posteriormente, evaluaremos cada modelo en base al AIC, BIC y devianza. Para ello, creamos una tabla en dónde la primera columna es el AIC, la segunda el BIC y la tercera la devianza de cada modelo (enumerados de arriba hacia abajo como Modelo 1, Modelo 2, Modelo 3 y Modelo 4). Ver cuadro a continuación:

```
cbind(
  rbind(AIC(mod1),AIC(mod2),AIC(mod3),AIC(mod4)),
  rbind(BIC(mod1),BIC(mod2),BIC(mod3),BIC(mod4)),
  rbind(mod1$deviance,mod2$deviance,mod3$deviance,mod4$deviance)
  )
```

```
##          [,1]     [,2]        [,3]
## [1,] 27410.74 27429.62 222292.3874
## [2,] 27318.24 27343.42 217098.8231
## [3,] 26466.78 26485.66    896.3073
## [4,] 26398.14 26423.31    881.1591
```

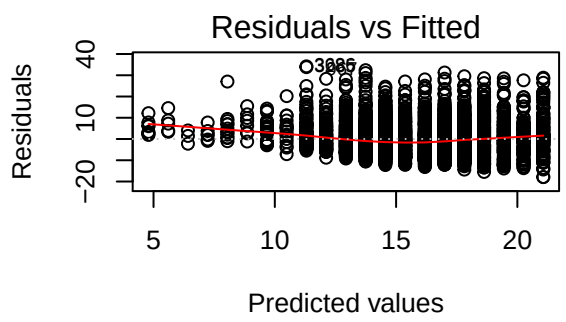
Del presente cuadro, se puede observar lo siguiente:

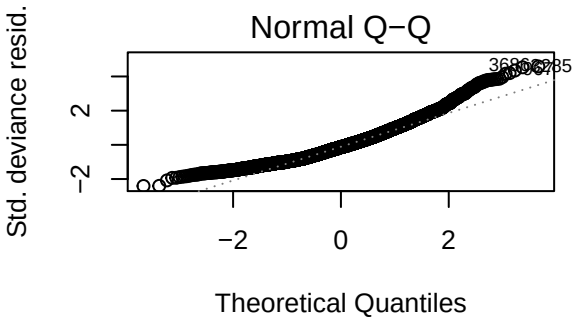
- Que el modelo 4 tiene el AIC y BIC más bajo de los 4 modelos presentados, siendo estos 26,398.14 y 26,423.31 respectivamente.
- Que los modelos en dónde la variable objetivo tiene distribución normal (modelos 1 y 2) tienen devianzas muy altas a diferencia de los modelos cuya variable objetivo tiene distribución gamma (modelos 3 y 4).

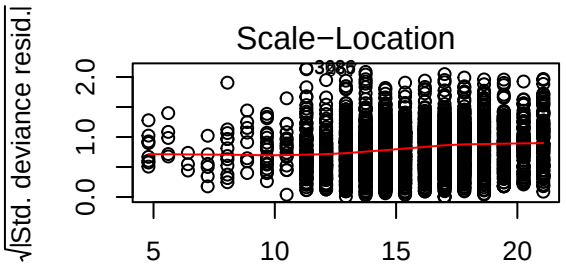
Asimismo, revisamos los gráficos de residuales de cada modelo. Ver a continuación:

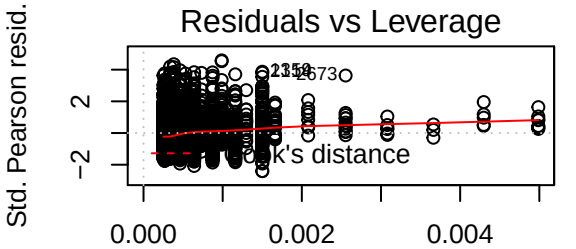
Modelo 1

```
par(mfrow =c(2,2))
plot(mod1)
```



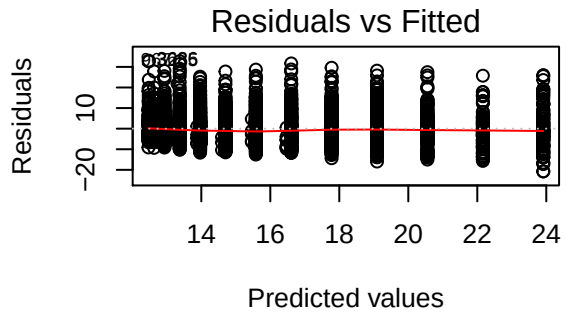


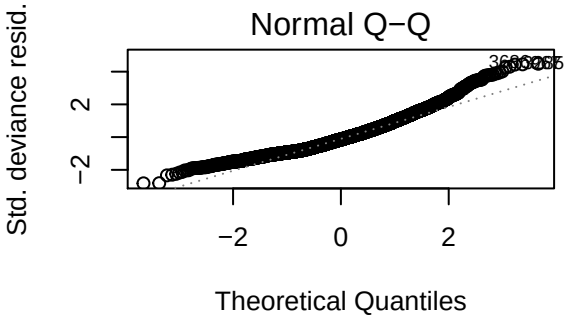


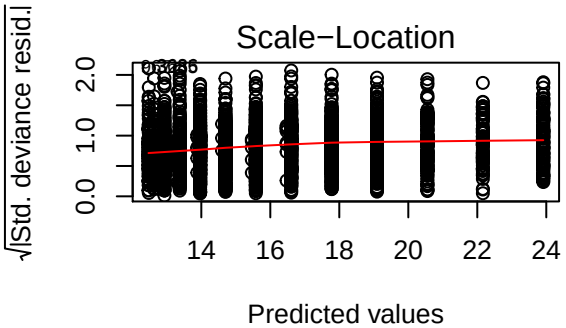


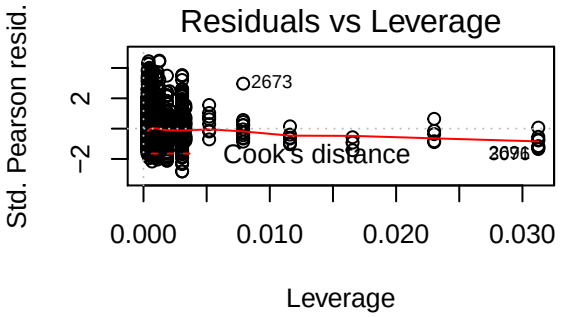
Modelo 2

```
par(mfrow =c(2,2))
plot(mod2)
```



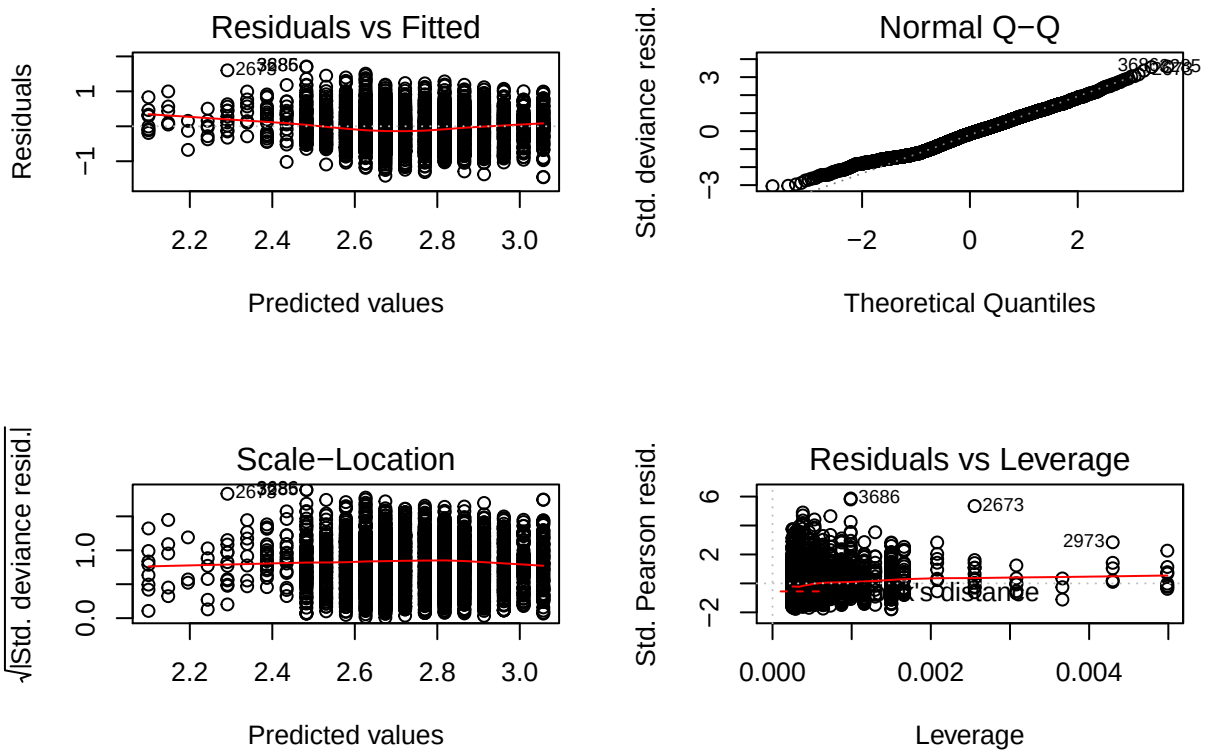






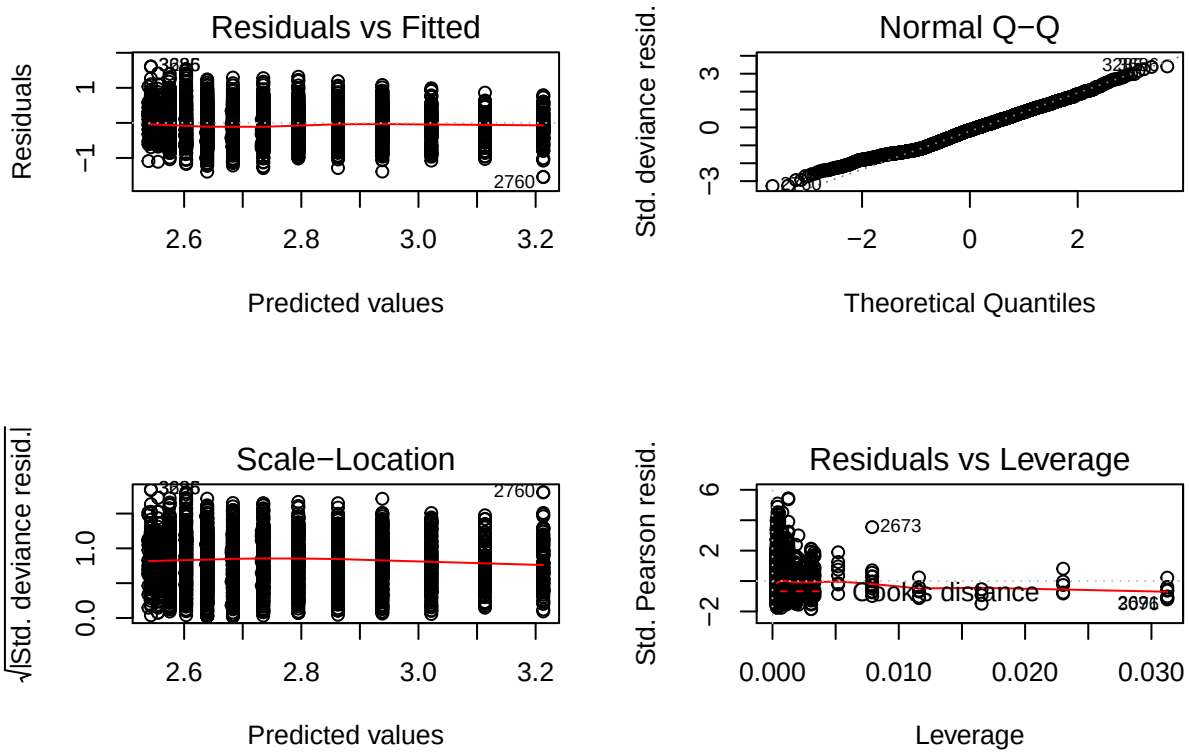
Modelo 3

```
par(mfrow =c(2,2))
plot(mod3)
```



Modelo 4

```
par(mfrow =c(2,2))
plot(mod4)
```



En base a los gráficos presentados, se observa lo siguiente:

- Se observa que los residuales de los modelos 3 y 4 tienen mejor ajuste a la recta en el qqplot que los modelos 1 y 2.
- Entre dichos modelos, se observa mayor uniformidad de los residuales y los valores predichos en el modelo 4.

En conclusión, de los 4 modelos generados escogeríamos el modelo 4, pues es el que tiene menor AIC, BIC, devianza y presenta un mejor ajuste a los datos.

Pregunta 3.b)

Para el modelo 3, interprete los parámetros en términos del problema. Repita el procedimiento para el modelo 4, considerando los parámetros e^{β_0} y $e^{\beta_1/2\beta_2}$.

Modelo 3

```
exp(mod3$coefficients[1:2])-1
```

```
## (Intercept)           x
## 14.36536466  0.04902732
```

Se observa lo siguiente:

- La media esperada de la renta media horaria, asumiendo que no existiese efecto de los años de estudio (es decir, que β es igual a 0) es de 15.36 reales
- Por cada año de estudio adicional, la media esperada de la renta media horaria (en reales) del jefe o jefes de domicilio aumente en 0.05%.

Pregunta 4

Analice el conjunto de datos en el archivo *hojas.csv* con mediciones en tilos de hoja pequeña cultivados en cierto pais, el objetivo es modelar la biomasa de hojas. El conjunto de datos consta de 385 observaciones, y

veremos las siguientes variables:

- Follaje: la biomasa de hojas, en kg (materia seca al horno).
- DAP: el diámetro del árbol, en cm.
- Origen: el origen del árbol; uno de Coppice, natural, plantado. Se usa codificación, Coppice es la categoría de referencia.

Pregunta 4.a)

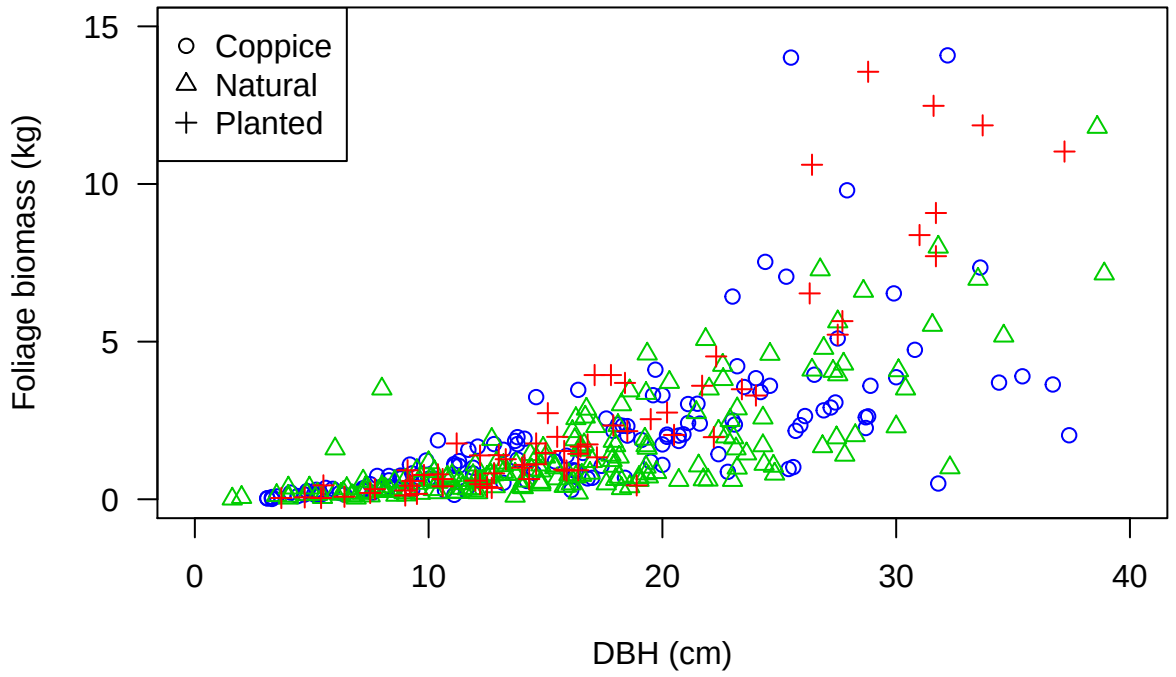
Realice un análisis exploratorio y defina detalladamente sus MLG ajustados en R para estos datos.

```
base <- read.csv("hojas.csv", sep=",")
```

Análisis Exploratorio

Foliage vs DBH

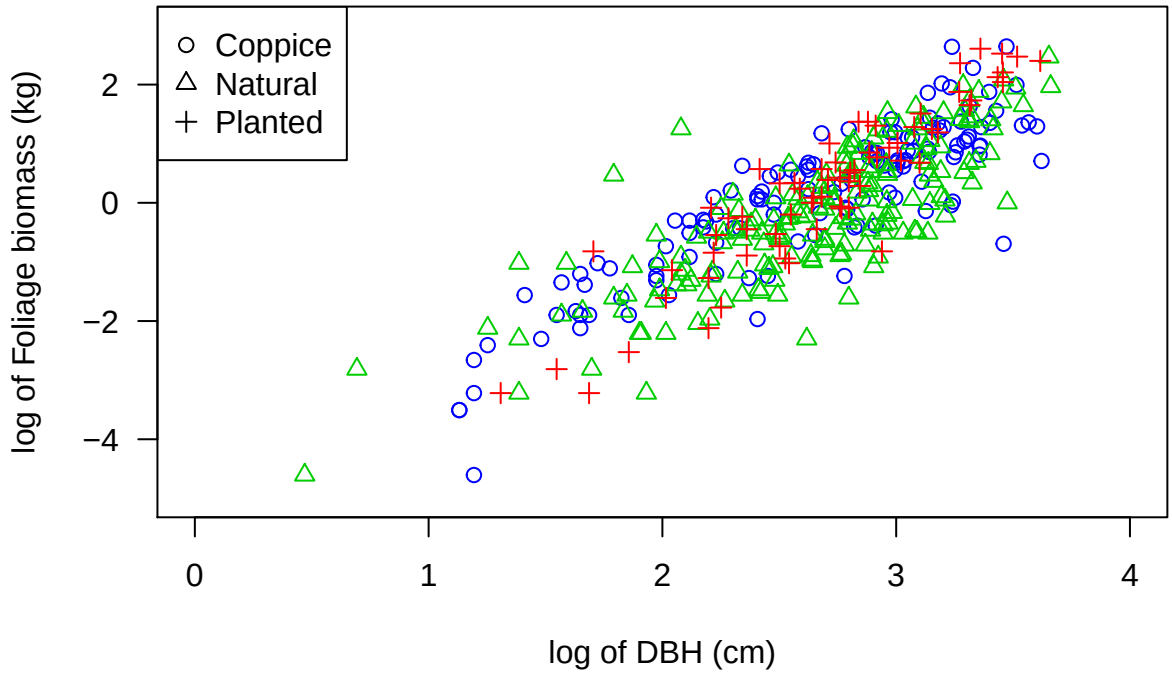
```
plot(Foliage ~ DBH, type="n", las=1,
     xlab="DBH (cm)", ylab="Foliage biomass (kg)",
     ylim = c(0, 15), xlim=c(0, 40), data=base)
points(Foliage ~ DBH, data=subset(base, Origin=="Coppice"),
       pch=1 , col=4 )
points(Foliage ~ DBH, data=subset(base, Origin=="Natural"),
       pch=2 , col=3 )
points(Foliage ~ DBH, data=subset(base, Origin=="Planted"),
       pch=3 , col=2)
legend("topleft", pch=c(1, 2, 3),
       legend=c("Coppice", "Natural","Planted"))
```



Foliage vs DBH, en escala logarítmica:

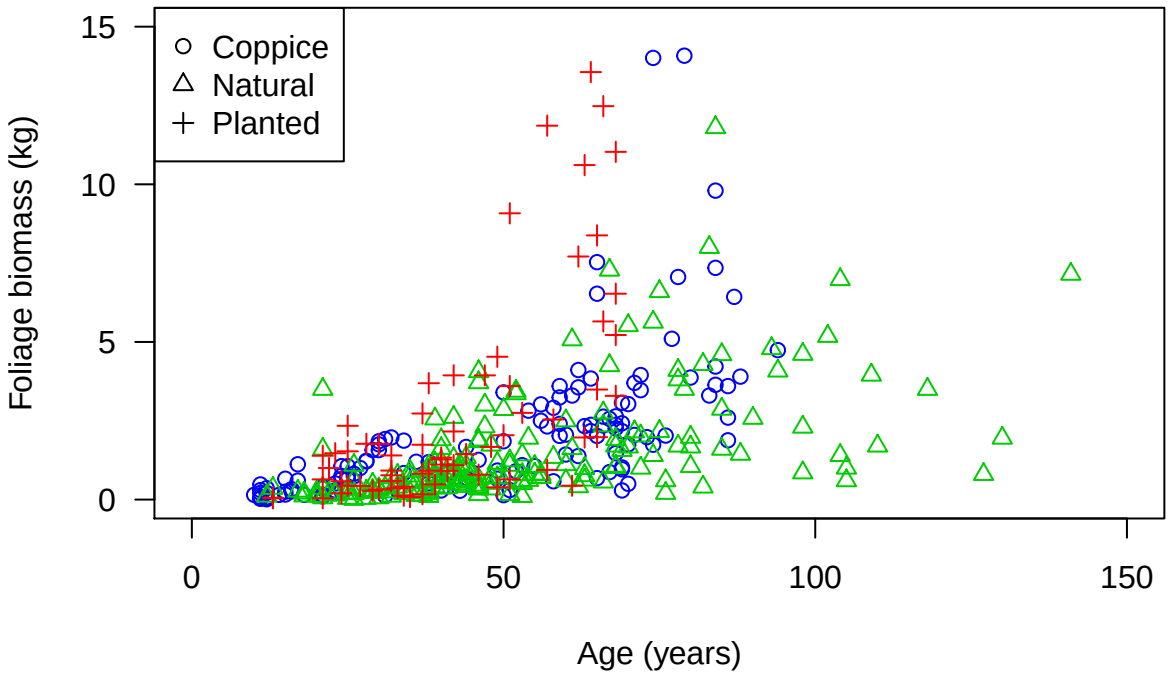
```
plot( log(Foliage) ~ log(DBH), type="n", las=1,
      xlab="log of DBH (cm)", ylab="log of Foliage biomass (kg)",
      ylim = c(-5, 3), xlim=c(0, 4), data=base)
points( log(Foliage) ~ log(DBH), data=subset(base, Origin=="Coppice"),
```

```
        pch=1 , col=4 )
points( log(Foliage) ~ log(DBH), data=subset(base, Origin=="Natural"),
        pch=2 , col=3 )
points( log(Foliage) ~ log(DBH), data=subset(base, Origin=="Planted"),
        pch=3 , col=2 )
legend("topleft", pch=c(1, 2, 3),
       legend=c("Coppice", "Natural","Planted"))
```



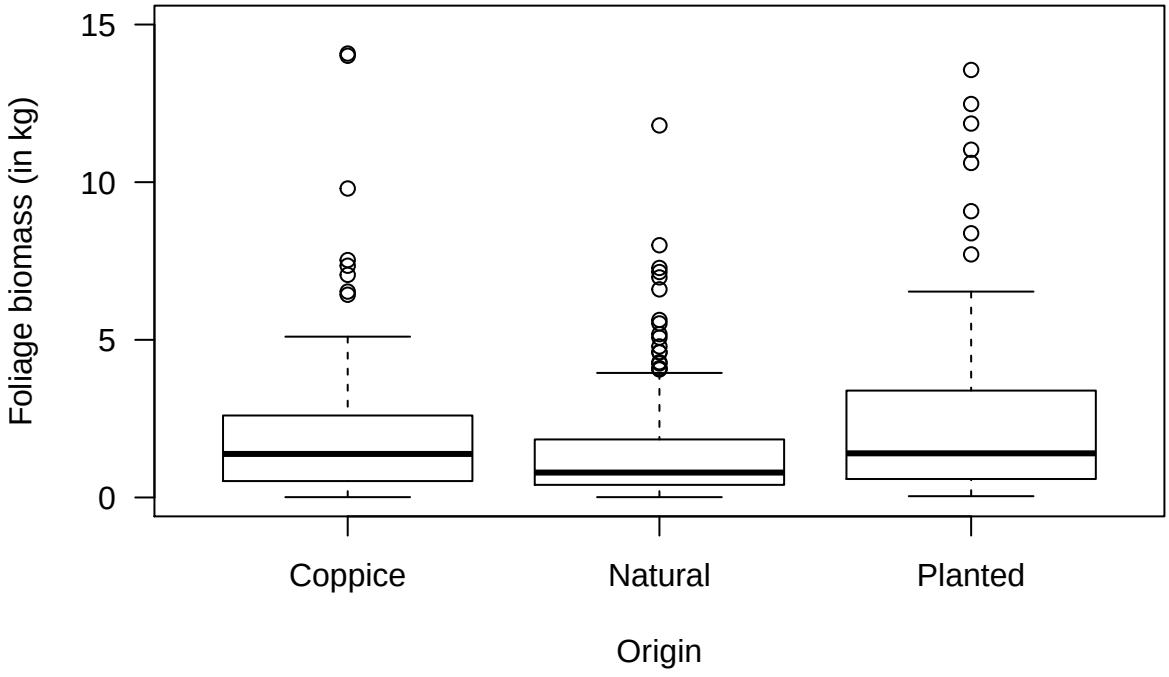
Foliage vs Age

```
plot(Foliage ~ Age, type="n", las=1,
     xlab="Age (years)", ylab="Foliage biomass (kg)",
     ylim = c(0, 15), xlim=c(0, 150), data=base)
points(Foliage ~ Age, data=subset(base, Origin=="Coppice"), pch=1,col=4)
points(Foliage ~ Age, data=subset(base, Origin=="Natural"), pch=2,col=3)
points(Foliage ~ Age, data=subset(base, Origin=="Planted"), pch=3,col=2)
legend("topleft", pch=c(1, 2, 3),
       legend=c("Coppice", "Natural","Planted"))
```



Foliage vs Origin

```
plot( Foliage ~ Origin, data=base, ylim=c(0, 15),
      las=1, ylab="Foliage biomass (in kg)")
```



La respuesta siempre es positiva, la varianza en la biomasa del follaje aumenta a medida que aumenta la media y existe una relación entre follaje biomasa y la variable DBH; al igual que el follaje biomasa y edad. El efecto con origen es más difícil de apreciar.

Modelos Propuestos

```
base.log <- glm( Foliage ~ Origin * log(DBH), family=Gamma(link="log"),
                 data=base)
phi.log.mle <- deviance(base.log)/length(base$Foliage)
phi.log.md <- deviance(base.log)/df.residual(base.log)
phi.log.pearson <- summary( base.log )$dispersion
```

```
base.iG <- glm( Foliage ~ Origin * log(DBH),
                family=inverse.gaussian(link="log"), data=base)
phi.iG.mle <- deviance(base.iG)/length(base$Foliage)
phi.iG.md <- deviance(base.iG)/df.residual(base.iG)
phi.iG.pearson <- summary( base.iG )$dispersion
```

```
c( "MLE"=phi.log.mle, "Mean dev."=phi.log.md, "Pearson"=phi.log.pearson)
```

```
##       MLE Mean dev.   Pearson
## 0.3965961 0.4028747 0.5443774
```

```
c( "MLE"=phi.iG.mle, "Mean dev."=phi.iG.md, "Pearson"=phi.iG.pearson)
```

```
##       MLE Mean dev.   Pearson
##  1.056659  1.073387  1.255992
```

AIC sugiere que la gamma glm es preferible a la inversa gaussiana glm

```
c( "Gamma:"=AIC(base.log), "inv. Gauss.:"=AIC(base.iG) )
```

```
##       Gamma: inv. Gauss.:
##     750.3267    1089.5297
```

Pregunta 4.b)

En base a los resultados presentados para los ¿Ud. cree que los datos siguen una distribución exponencial? ¿Qué parámetro debe analizar para ello?

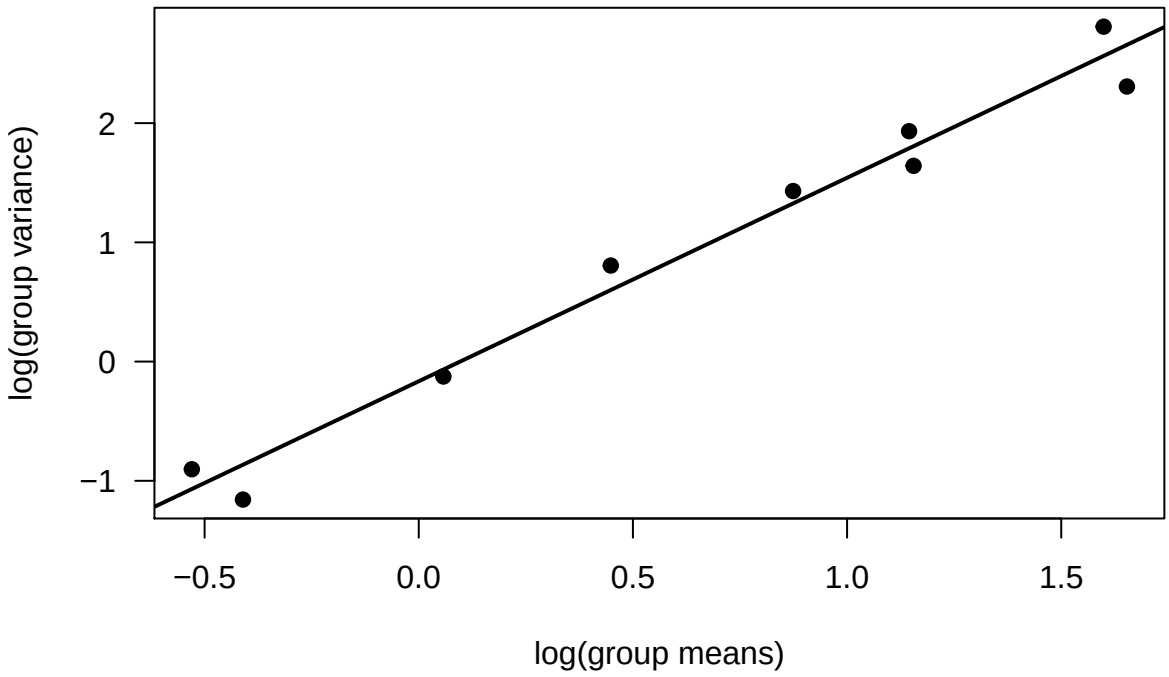
Procedemos a dividir en grupos más pequeños los datos para calcular la media y la varianza de cada grupo , los siguientes gráficos muestran que la varianza aumenta a medida que aumenta la media.

```
base$AgeGrp <- cut(base$Age, breaks=4 )
```

```
var <- with( base, tapply(Foliage, list(AgeGrp, Origin), "var" ) )
mean <- with( base, tapply(Foliage, list(AgeGrp, Origin), "mean" ) )
plot( log(var) ~ log(mean), las=1, pch=19,
      xlab="log(group means)", ylab="log(group variance)")
mf.lm <- lm( c(log(var)) ~ c(log(mean)) )
coef( mf.lm )
```

```
##  (Intercept) c(log(mean))
##    -0.165002     1.706453
```

```
abline( coef( mf.lm ), lwd=2)
```



La pendiente de la línea es un poco menor que 2 , entonces aproximadamente $log(groupvariance) \propto 2\ddot{O}log(groupmean)$ En otras palabas $V(\mu) \approx \mu^2$ corresponde a una distribución gamma.

Existen dos situaciones comunes donde se conoce ϕ, en situaciones donde sigue una distribución normal, las variaciones de muestra se pueden modelar usando una distribución de chi-cuadrado, que es una distribución gamma con $\phi = 2$.

En segundo lugar, la **distribución exponencial** que en resumen es un distribución gamma con $\phi = 1$

Pregunta 4.c)

Nos fijamos en un árbol de Coppice. Si el registro (DAP) del árbo aumenta en 1 cm , ¿Cuál es el cambio en la media de la biomasa foliar estimada?

Pregunta 4.d)

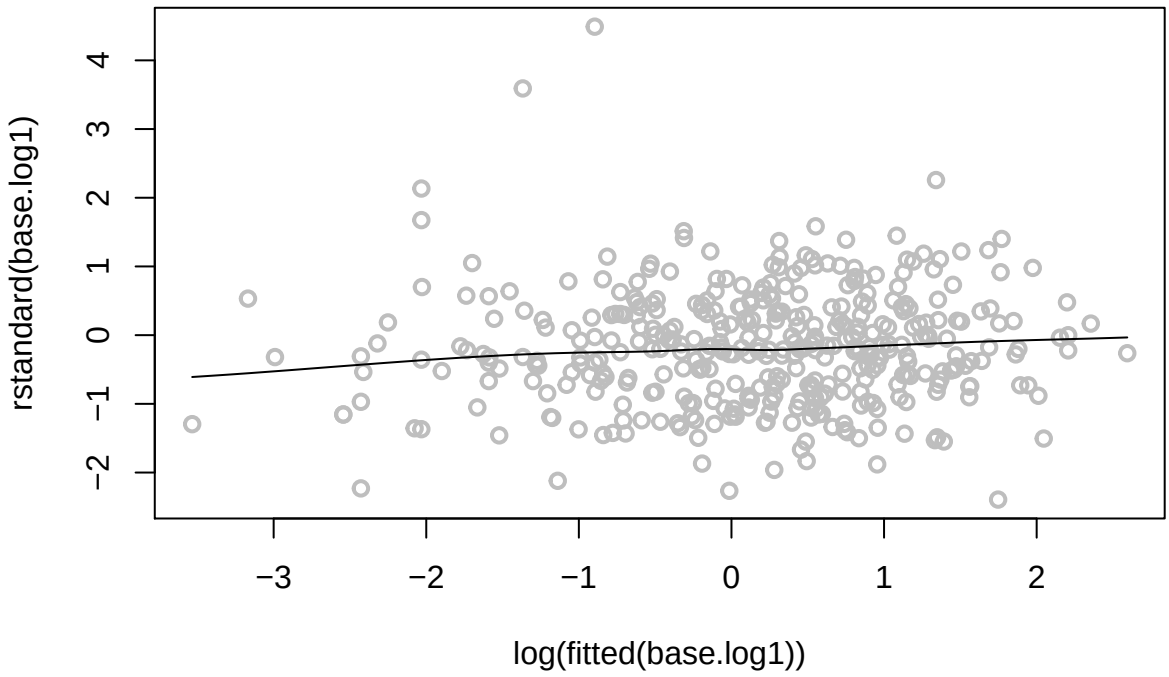
Se realiza una prueba de Anova para comparar el modelo con interacción entre log (DAP) y Origin con el modelo donde la interacción entre log (DAP) y Origin no se incluyen en el predictor lineal. ¿Cuál es su conclusión? ¿Cambia su conclusión según el AIC?

```
base.log1 <- glm( Foliage ~ Origin * log(DBH),
family=Gamma(link="log"), data=base)
```

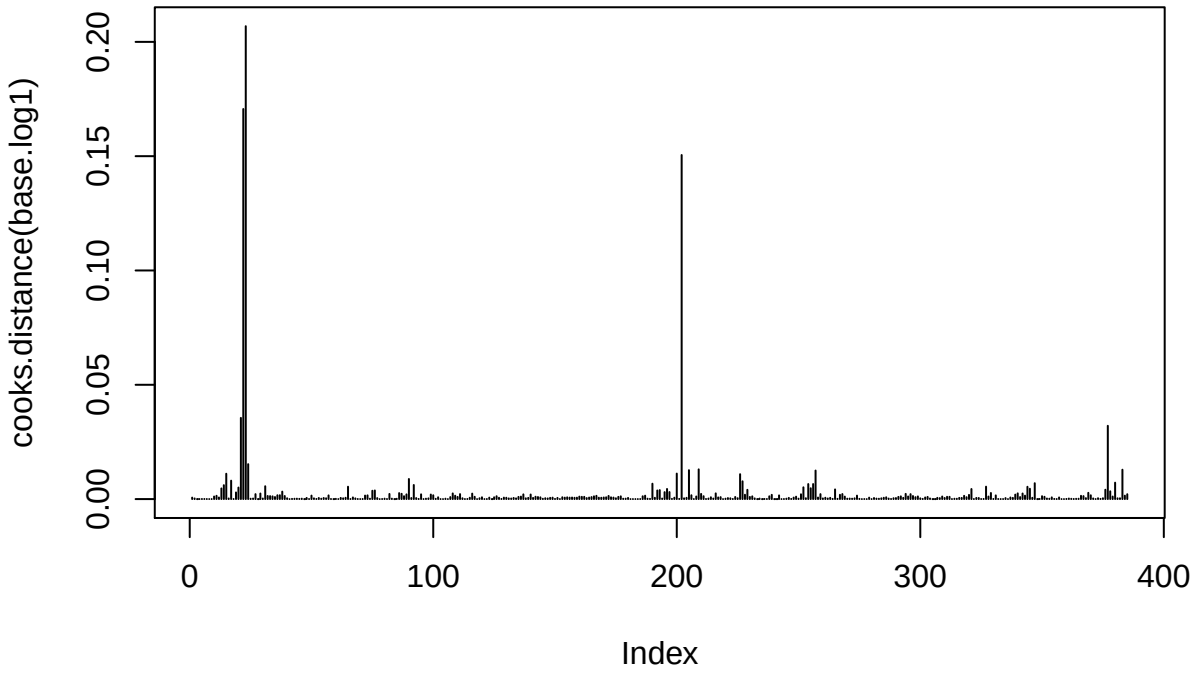
```
base.log2 <- glm( Foliage ~ Origin * DBH,
family=Gamma(link="log"), data=base)
```

```
par(mfrow=c(2, 3))
```

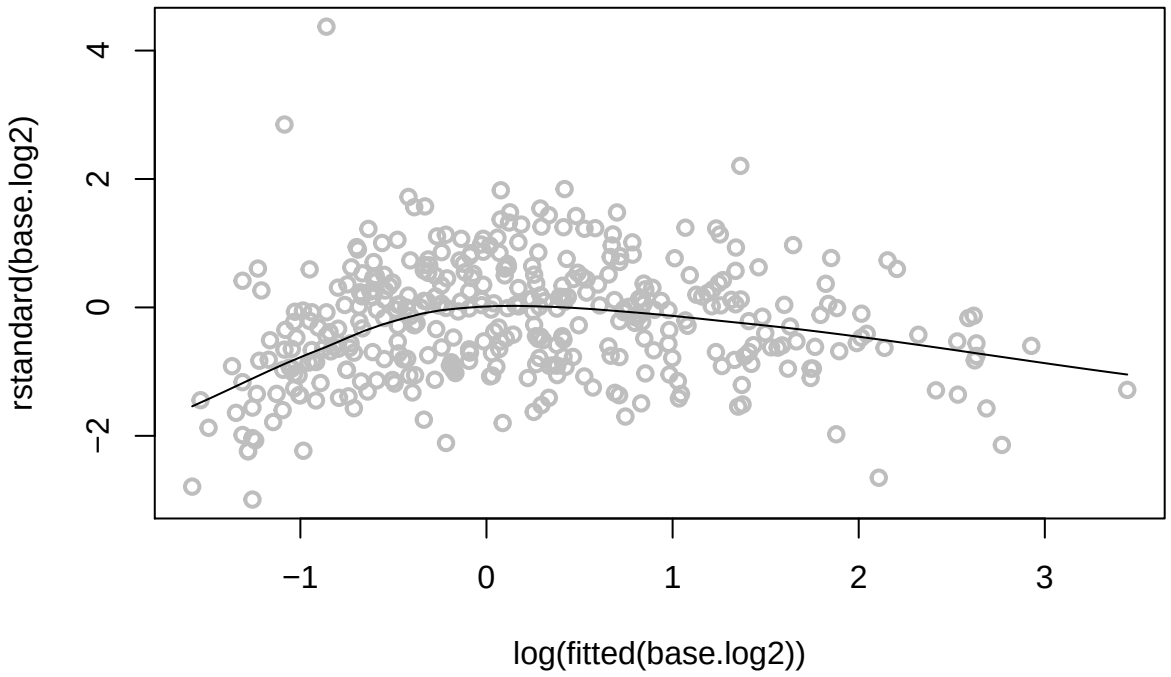
```
scatter.smooth( log(fitted(base.log1)), rstandard(base.log1),
col="gray", lwd=2 )
```



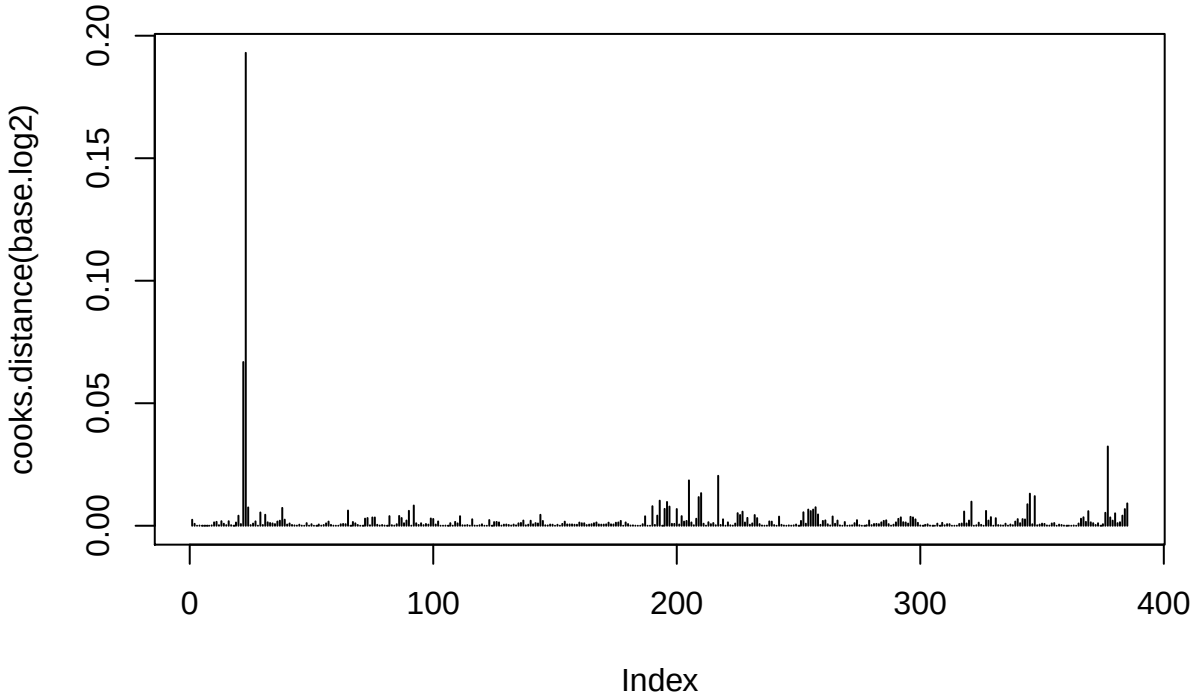

```
plot(cooks.distance(base.log1), type="h")
```




```
scatter.smooth( log(fitted(base.log2)), rstandard(base.log2),
col="gray", lwd=2 )
```




```
plot(cooks.distance(base.log2), type="h")
```




```
colSums(influence.measures(base.log1)$is.inf)
```

```
##   dfb.1_ dfb.OrgN dfb.OrgP dfb.1(DB dfb.ON:( dfb.OP:(    dffit    cov.r
##        0        0        0        0        0        0        7       29
##   cook.d      hat
##        0       18
```

```r
colSums(influence.measures(base.log2)$is.inf)
```

```
##   dfb.1_ dfb.OrgN dfb.OrgP  dfb.DBH dfb.ON:D dfb.OP:D    dffit    cov.r
##        0        0        0        0        0        0       14       27
##   cook.d      hat
##        0       14
```

AIC sugiere que el primer modelo es preferible frente al segundo modelo.

```r
c( "Gamma:"=AIC(base.log1), "inv. Gauss.:"=AIC(base.log2) )
```

```
##       Gamma: inv. Gauss.:
##     750.3267     820.3632
```