Pontificia Universidad JAVERIANA Bogotá

MAESTRÍA EN EPIDEMIOLOGÍA CLÍNICA

BIOESTADÍSTICA AVANZADA

MÓDULO I

Semana 4

Diagnóstico del modelo

Material de contenido y aplicación

Carlos Javier Rincón R.

Introducción

En esta semana revisaremos como realizar el diagnóstico del modelo de regresión lineal. Consiste en identificar datos atípicos, evaluar los supuestos del modelo y finalmente se revisar el concepto de colinealidad. Se presentan 3 ejercicios de aplicación que deben ser replicados por los estudiantes localmente, para afianzar los conceptos presentados.

Los supuestos del modelo

Al realizar el ajuste de un modelo de regresión lineal, a continuación se debe realizar una exploración o diagnóstico sobre tres aspecto que pueden afectar los resultados y conclusiones obtenidas: Datos atípicos, los supuestos del modelo y colinealidad. A continuación se discuten cada uno de estos aspectos.

Datos atípicos

Los datos atípicos son valores observados en las variables de interés, que se consideran muy distintos (muy pequeños o muy grandes) a los valores en general observados en toda la población de estudio y que pueden ocasionar una desviación de la relación entre las variables independientes y el desenlace. Para evaluar que “tan distintos” son, se debe contar con el conocimiento por parte de expertos sobre cada una de las variables en el contexto en el cual se está investigando; esto permite diferenciar cuales datos atípicos no son plausibles y cuales aunque sean extraños si se pueden presentar. La anterior diferenciación permite tomar una decisión, si el dato no es plausible se revisa si es un error de medición y se corrige, de no ser posible esta corrección se elimina este valor (no se eliminan todas las observaciones de ese sujeto). Si el valor atípico es plausible, se revisa igualmente si pudo ser un error de medición, si no, se debe dejar este valor.

Posterior a la corrección de los errores de medición, una primera aproximación para identificar datos atípicos es a través de estadísticas descriptiva, identificando para cada variable los cinco sujetos con valores más altos y los cinco sujetos con valores más bajos. Adicionalmente se pueden obtener gráficos de cajas y bigotes para cada variable, y gráficos de dispersión entre cada variable independiente y el desenlace.

Ahora, los datos atípicos se pueden identificar de una forma más objetiva a partir de distintas medidas. La primera se denomina `leverage` y se denota por h_i que se obtiene para cada sujeto, basado en la comparación entre sus valores observados $x_{i1}, x_{i2}, \ldots, x_{ip}$ y el promedio de los valores observados en cada variable $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$ a través de la distancia geométrica. Tenemos que hi toma valores entre 0 y 1 y se identifican como datos atípicos las observaciones con $h_i > 2(p + 1)/n$, donde p y n son el número de variables independientes y el número de sujetos respectivamente.

La identificación de datos atípicos también se puede realizar a partir de los residuales del modelo, basado en que las observaciones atípicas presentarán residuales más altos en valor absoluto. La estimación de los residuales se obtiene de la diferencia $\hat{e}_i = y_i \ \hat{y}_i$, pero estos se encuentran expresados en las unidades de medida del desenlace, lo cual impide tener un único valor de referencia para distintos desenlaces que determine cuando se puede considerar o no una dato atípico. Por eso, se plantean los residuales estandarizados:

$$z_i = \frac{\hat{e}_i}{\hat{\sigma}_{y|x}}$$

y los residuales studentizados:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}_{y|x}\sqrt{1 - h_i}}$$

donde $\hat{\sigma}_{y|x}$ es la raíz cuadrado de la estimación del cuadrado medio del error. Bajo el cumplimiento del supuesto de normalidad y homocedasticidad descritos en las dos primera semanas, z_i sigue una distribución normal estándar y r_i sigue una distribución **t_student** con $n - p - 1$ grados de libertad. Bajo estas distribuciones se pueden identificar como datos atípicos los sujetos con residuales en valor absoluto superiores al percentil 95 correspondiente a cada distribución.

Otra medida utilizada frecuentemente, son los residuales de **Jackknife** que se obtiene bajo la siguiente expresión:

$$r_{(-i)} = \frac{\hat{e}_i}{\hat{\sigma}_{y|x,-i}\sqrt{1-h_i}}$$

donde $\hat{\sigma}_{y|x,-i}$ es la estimación de la desviación estándar de y dado x después de retirar la observación i. $r_{(-i)}$ tiene un distribución t_student con $n - p - 2$ grados de libertad por lo cual se identificará como dato atípico las observaciones con residuales de Jackknife superiores en valor absoluto al percentil 95 de esta distribución.

Como última medida tenemos la distancia de **Cook**, que se basa en la influencia que puede tener una observación en la estimación de los coeficientes $\beta t's$ del modelo. Es decir, identificar las observaciones que al retirarlas generan un mayor cambio en los coeficientes estimados del modelo. La distancia de Cook se obtiene con la siguiente expresión:

$$d_i = (\frac{1}{p+1})(\frac{h_i}{1-h_i})r_i^2$$

y si $d_i > 1$ se identifica la observación como atípica.

Ejemplo de aplicación 1

Utilizaremos la base de datos de 40 fetos en quienes se evaluaron las siguientes variables:

- DBP: diámetro biparietal en milímetros (mm).
- LCC: longitud cráneo-caudal en milímetros (mm).
- DOF: diámetro occipitofrontal en milímetros (mm).
- edad: edad de la madre (menor o igual a 30 años, 31 a 35 años y mayor a 35 años).
- sexo: sexo del feto (H,M).

Los datos observados se registran a continuación:

```
id<-c(1:40) # identificador del sujeto

DBP<-c(18,19,19.2,21,26,24.4,23,16.1,18.8,17,22,20.2,22.5,17.2,
20.3,19,22,20.3,19.5,19.5,20.8,23.8,22,22,21.8,19.6,18.6,
22.5,26.4,19.5,26.4,24.1,21.3,17.1,24.7,22.8,17.7,35.1,
19.3,23.3)

LCC<-c(50,58.9,55.5,64,82,79.1,72.3,49,56.9,61,63.4,54.5,74.8,
51,61.3,74,67,60.1,62.1,56.8,72.2,73.2,64,65.1,64.9,60.4,
58,73.7,82.7,59,82.5,68.1,55,49,76.9,72.6,62.2,90.8,64.9,
71.4)

DOF<-c(23,26.1,24.5,30,34,29.3,24.9,21.3,23,24,26.3,23.4,27.3,22.1,
26.1,29,29,23.6,23.1,25.5,27.4,29,24,28.5,26.3,27,25.3,22.8,
31.9,25.7,33.3,28.3,25.5,22.2,30.07,31,23.9,37.8,24.5,30.2)

edad<-c("mayor a 35 años","31 a 35 años","31 a 35 años","mayor a 35 años",
"31 a 35 años","mayor a 35 años","31 a 35 años","31 a 35 años",
"menor o igual a 30 años","mayor a 35 años","31 a 35 años",
"31 a 35 años","menor o igual a 30 años","mayor a 35 años",
"31 a 35 años","menor o igual a 30 años","mayor a 35 años",
"31 a 35 años","31 a 35 años","31 a 35 años","menor o igual a 30 años",
"31 a 35 años","31 a 35 años","31 a 35 años","31 a 35 años",
"menor o igual a 30 años","menor o igual a 30 años",
"menor o igual a 30 años","31 a 35 años","31 a 35 años","31 a 35 años",
"mayor a 35 años","menor o igual a 30 años","31 a 35 años",
"mayor a 35 años","menor o igual a 30 años","mayor a 35 años",
"mayor a 35 años","31 a 35 años","mayor a 35 años")

sexo<-c("M","H","H","H","H","H","M","H","M","M","M","H","H","M","M","M",
"M","H","M","M","H","H","H","M","H","M","H","M","M","H","M","M",
"M","M","M","M","H","H","H","M")

bd<-data.frame(id,DBP,LCC,DOF,edad,sexo) # Base de datos

bd$edad<-factor(bd$edad,levels = c("menor o igual a 30 años",
                                   "31 a 35 años","mayor a 35 años"))

bd$sexo<-factor(bd$sexo,levels = c("H","M"))
```

Inicialmente, se identifican los sujetos con valores extremos listando los cinco valores menores y los cinco valores mayores en cada una de las variables cuantitativas.

```
library(tidyverse)
# organizar la base de datos por la variable DBP:
bd<-arrange(bd,DBP)
# Reportar los cinco valores menores con sus ID respectivos
rbind(bd$id[1:5],bd$DBP[1:5])
```

```
## [,1] [,2] [,3] [,4] [,5]
## [1,] 8.0 10 34.0 14.0 37.0
## [2,] 16.1 17 17.1 17.2 17.7
```

```
# Reportar los cinco valores mayores con sus ID respectivos
rbind(bd$id[36:40],bd$DBP[36:40])
```

```
## [,1] [,2] [,3] [,4] [,5]
## [1,] 35.0 5 29.0 31.0 38.0
## [2,] 24.7 26 26.4 26.4 35.1
```

```
# Se repite el proceso para LCC y DOF:
bd<-arrange(bd,LCC)
rbind(bd$id[1:5],bd$LCC[1:5])
```

```
## [,1] [,2] [,3] [,4] [,5]
## [1,] 8 34 1 14 12.0
## [2,] 49 49 50 51 54.5
```

```
rbind(bd$id[36:40],bd$LCC[36:40])
```

```
## [,1] [,2] [,3] [,4] [,5]
## [1,] 6.0 5 31.0 29.0 38.0
## [2,] 79.1 82 82.5 82.7 90.8
bd<-arrange(bd,DOF)
```

```
rbind(bd$id[1:5],bd$DOF[1:5])
```

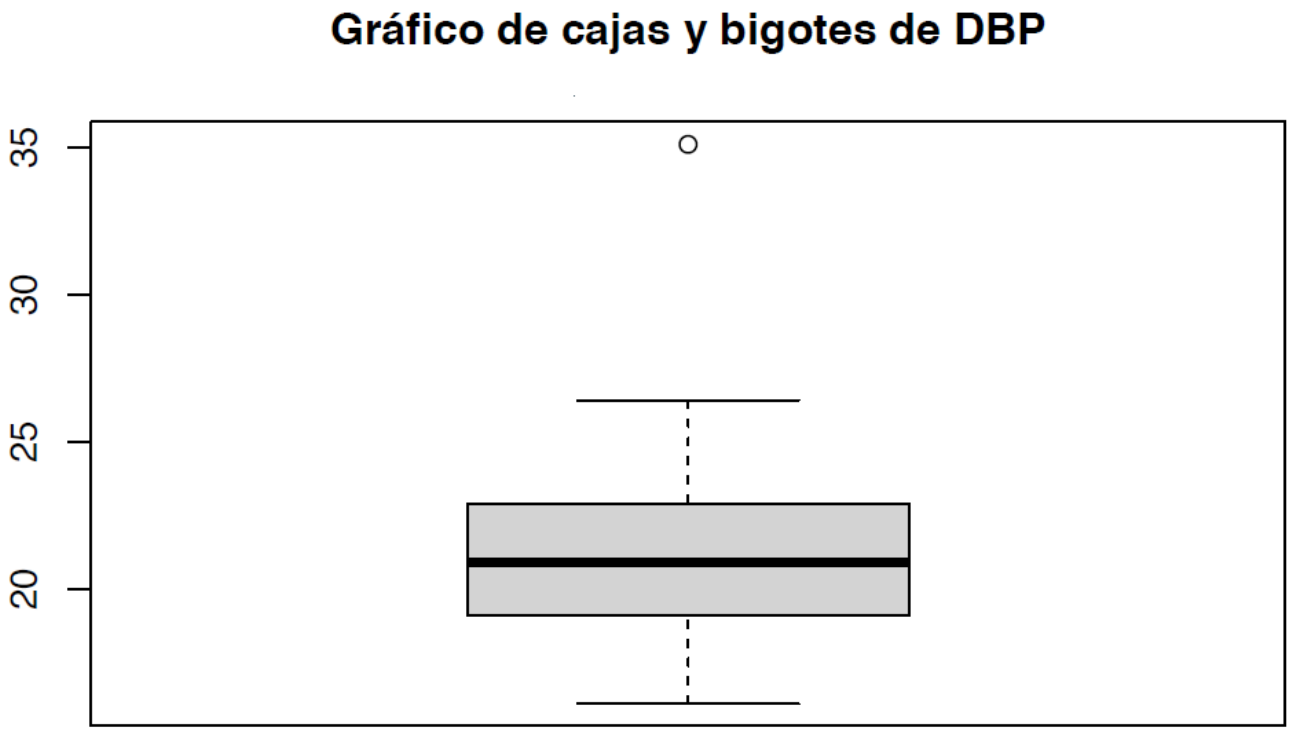
```
## [,1] [,2] [,3] [,4] [,5]
## [1,] 8.0 14.0 34.0 28.0 1
## [2,] 21.3 22.1 22.2 22.8 23
```

```
rbind(bd$id[36:40],bd$DOF[36:40])
```

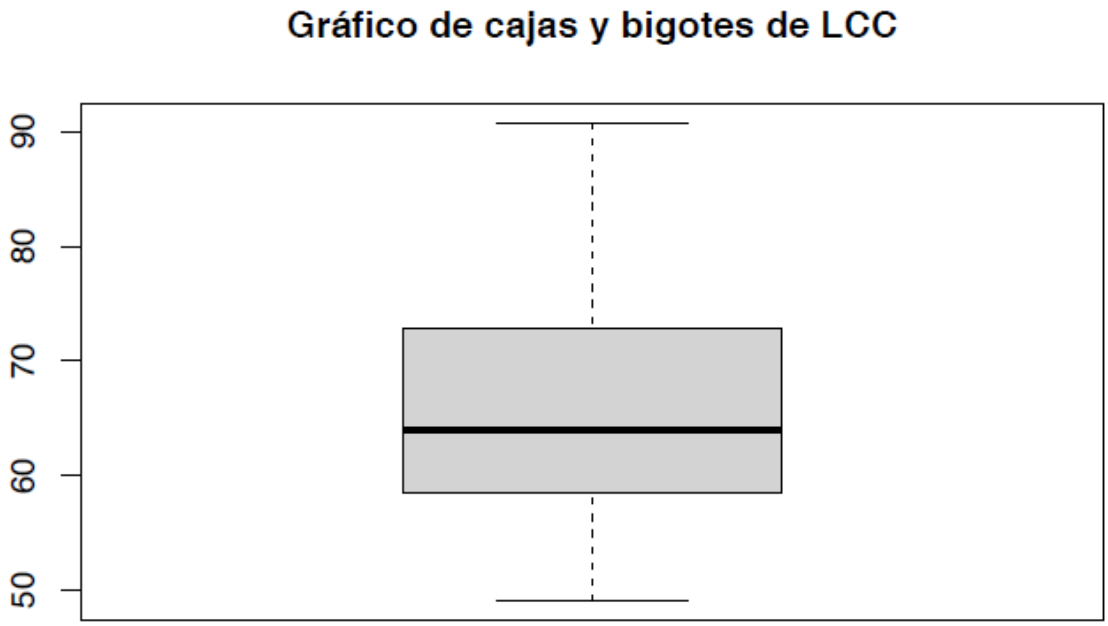
```
## [,1] [,2] [,3] [,4] [,5]
## [1,] 36 29.0 31.0 5 38.0
## [2,] 31 31.9 33.3 34 37.8
```

A continuación obtenemos gráficos de cajas y bigotes para cada variable cuantitativa:

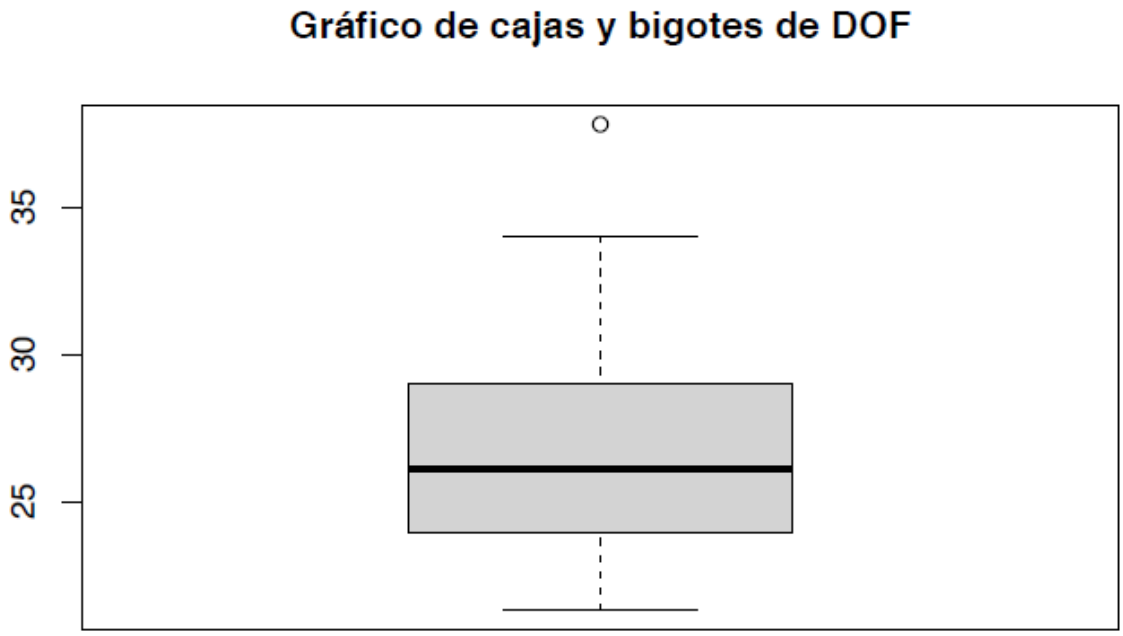
```
boxplot(bd$DBP,main="Gráfico de cajas y bigotes de DBP")
```




```
boxplot(bd$LCC,main="Gráfico de cajas y bigotes de LCC")
```

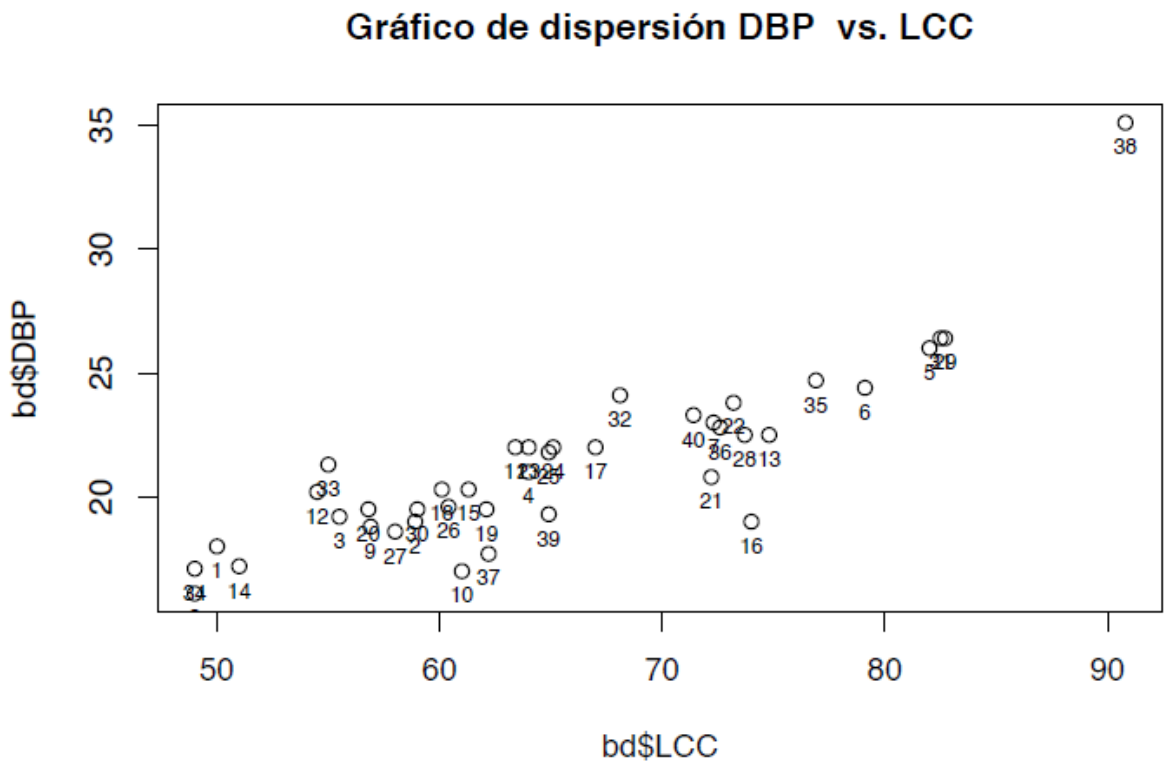



```
boxplot(bd$DOF,main="Gráfico de cajas y bigotes de DOF")
```

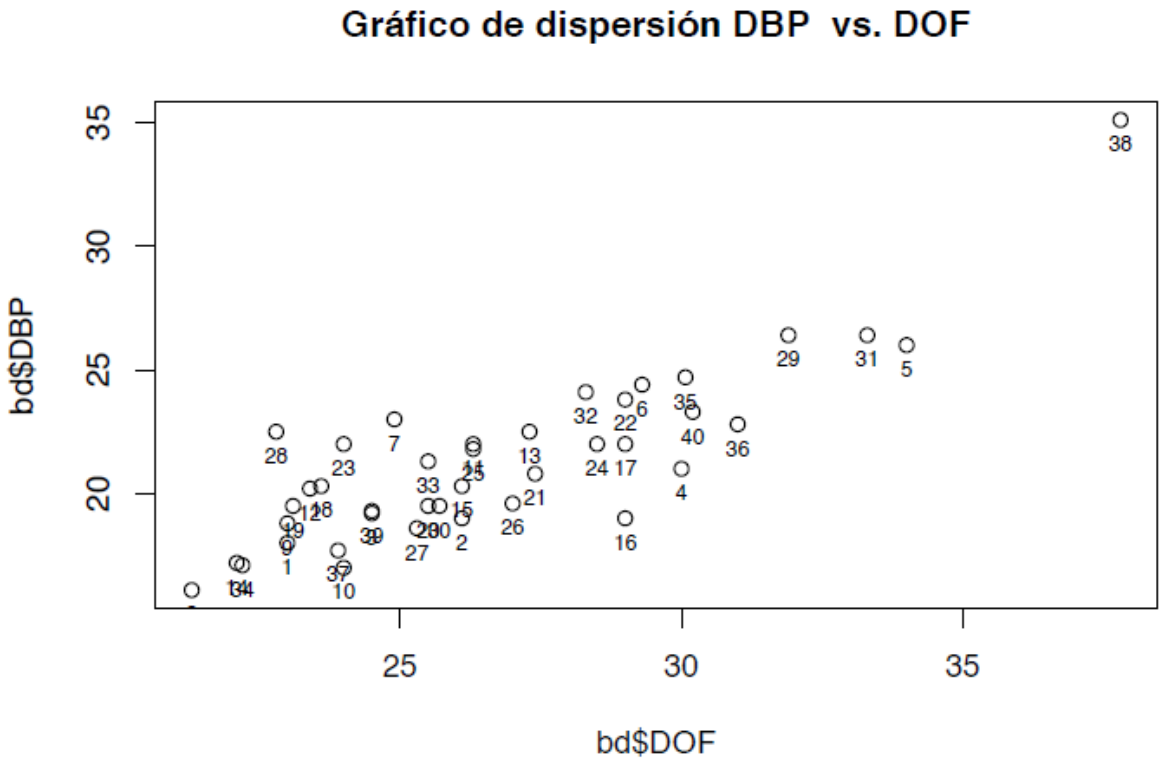


Finalmente, gráficos de dispersión entre cada variable cuantitativa versus el desenlace DBP.

```
plot(bd$LCC,bd$DBP,main="Gráfico de dispersión DBP vs. LCC")
text(bd$LCC,bd$DBP, labels=bd$id, cex= 0.7,pos=1)
```




```
plot(bd$DOF,bd$DBP,main="Gráfico de dispersión DBP vs. DOF")
text(bd$DOF,bd$DBP, labels=bd$id, cex= 0.7,pos=1)
```



De la descripción anterior, para efectos prácticos del ejercicio asumimos que todos los valores observados son plausibles y que no hay errores de medición. Además, podemos inicialmente identificar al sujeto *id* = 38 quién presenta un valor muy alto en las tres variables cuantitativas.

Ahora, obtendremos las siguientes medidas: h_i (leverage), z_i (residuales estandarizados), r_i (residuales estudentizados), r_{-i} (residuales de Jackknife) y d_i (distancia de Cook). Primero, reordenamos la base de datos por *id* (no es obligatorio, pero facilita identificar a los sujetos dentro de la base de datos), ajustamos nuestro modelo de

regresión lineal y de este extraemos los residuales y la estimación del CME $\left(\hat{\sigma}^2_{y|x}\right)$.

```
bd<-arrange(bd,id)

m<-lm(DBP~LCC+DOF+edad+sexo,data=bd) # ajuste del modelo
summary(m)
```

```
##
## Call:
## lm(formula = DBP ~ LCC + DOF + edad + sexo, data = bd)
##
## Residuals:
## Min 1Q Median 3Q Max
## -4.0099 -0.6101 -0.1220 0.8183 4.8845
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.55734 2.05576 -0.758 0.453938
## LCC 0.17913 0.04752 3.770 0.000623 ***
## DOF 0.38877 0.13385 2.905 0.006422 **
## edad31 a 35 años 1.04573 0.66413 1.575 0.124616
## edadmayor a 35 años 0.81194 0.74371 1.092 0.282624
## sexoM 0.03695 0.52534 0.070 0.944336
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.615 on 34 degrees of freedom
## Multiple R-squared: 0.8099, Adjusted R-squared: 0.782
## F-statistic: 28.98 on 5 and 34 DF, p-value: 2.4e-11
```

```
r.i<-m$residuals # los residuales del modelo
cme<-sum(r.i^2)/(40-5-1);cme # estimación del CME
```

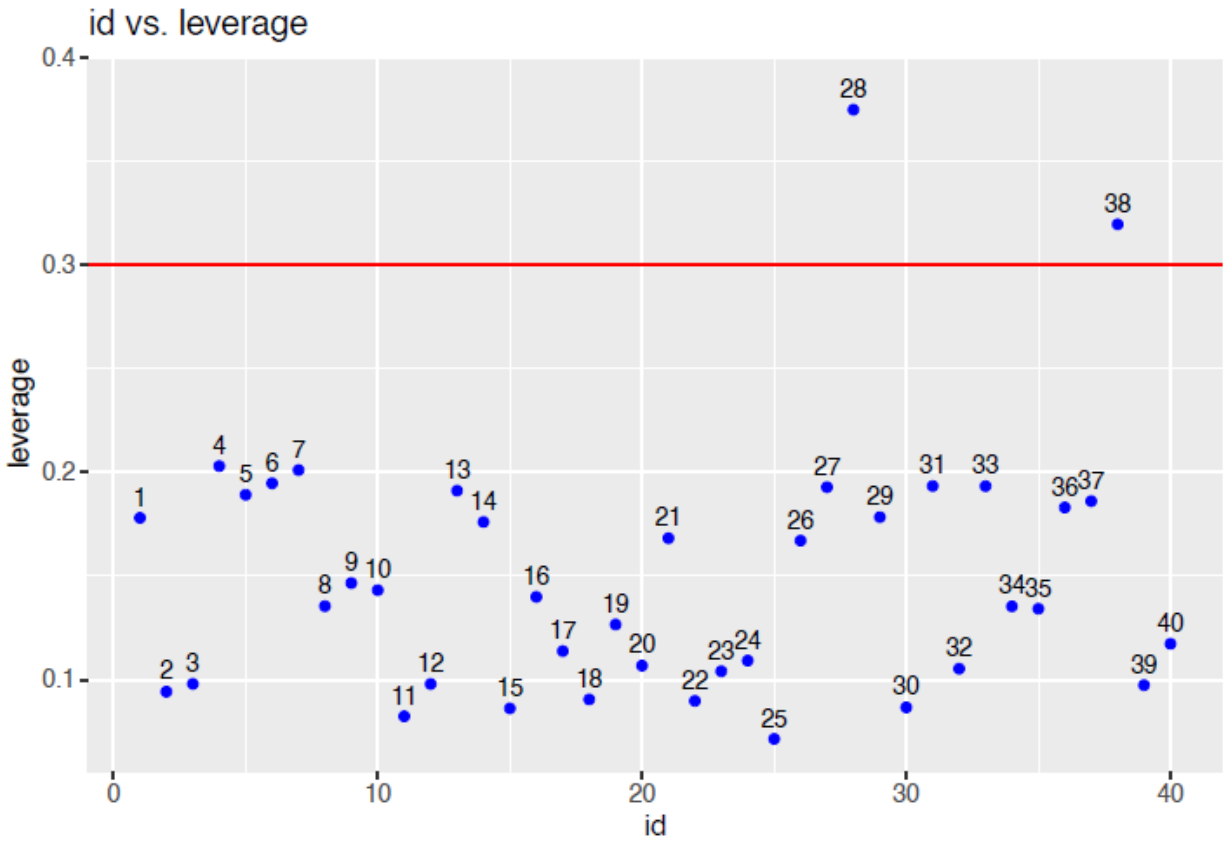
```
## [1] 2.607503
```

Ahora, vamos a generar las cinco medidas (los residuales en valor absoluto) y las guardamos en una nueva base de datos.

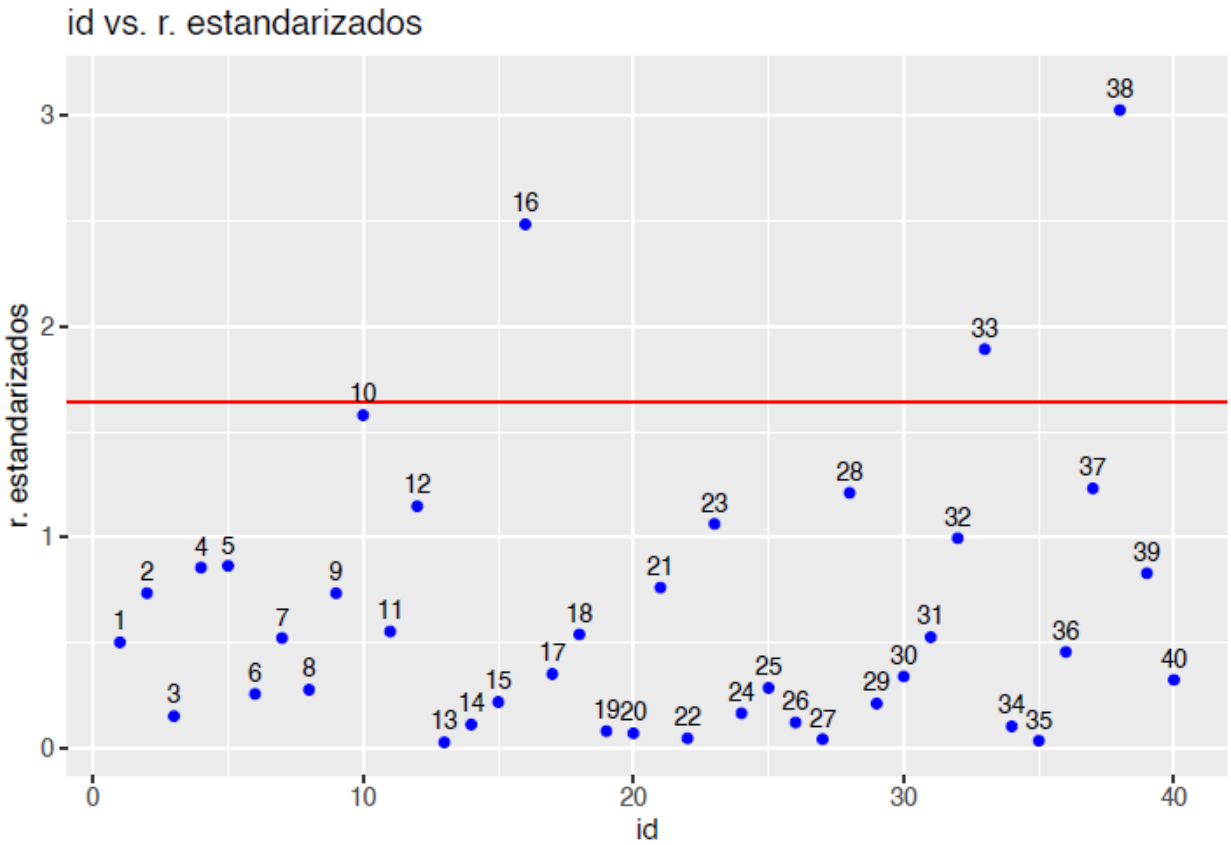
```
h.i<-hatvalues(m) # leverage
r.z<-abs(r.i/sqrt(cme)) # residuales estandarizados
r.st<-abs(r.i/(sqrt(cme)*sqrt(1-h.i))) # residuales estudentizados
r.jack<-abs(rstudent(m)) # residuales de jackkife
cook<-cooks.distance(m) # distancia de cook

# nueva base de datos:
m.datos.atipicos<-data.frame(bd$id,h.i,r.z,r.st,r.jack,cook)
```

A partir de gráficos de dispersión de cada medida versus el id del sujeto, podemos identificar las observaciones influyentes utilizando su punto de referencia respectivo:

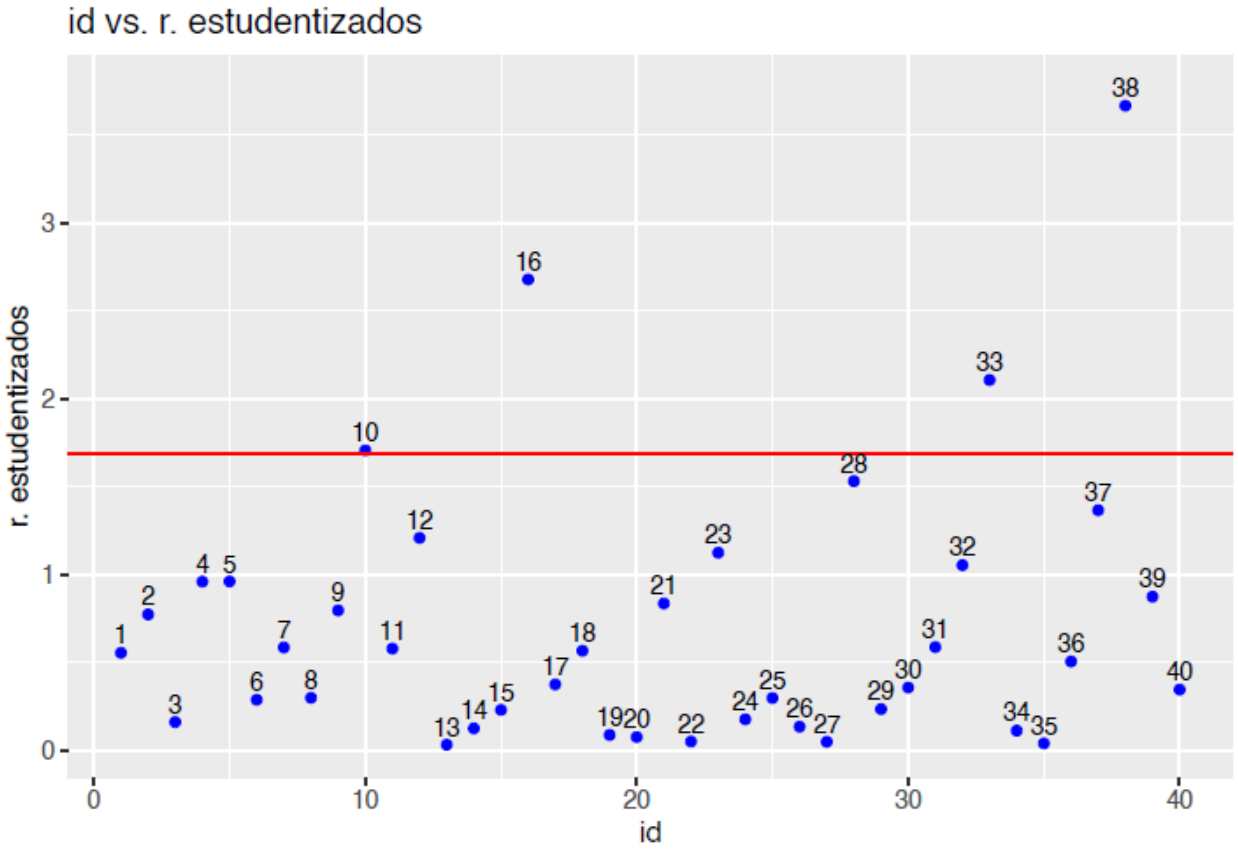
```
ggplot(aes(x=bd.id,y=h.i),data=m.datos.atipicos)+
        geom_point(col="blue")+
        geom_hline(yintercept =2*6/40,col="red")+
        labs(title = "id vs. leverage",y="leverage",x="id")+
        geom_text(aes(label=bd.id), size=3,nudge_x=0,nudge_y=0.01)
```



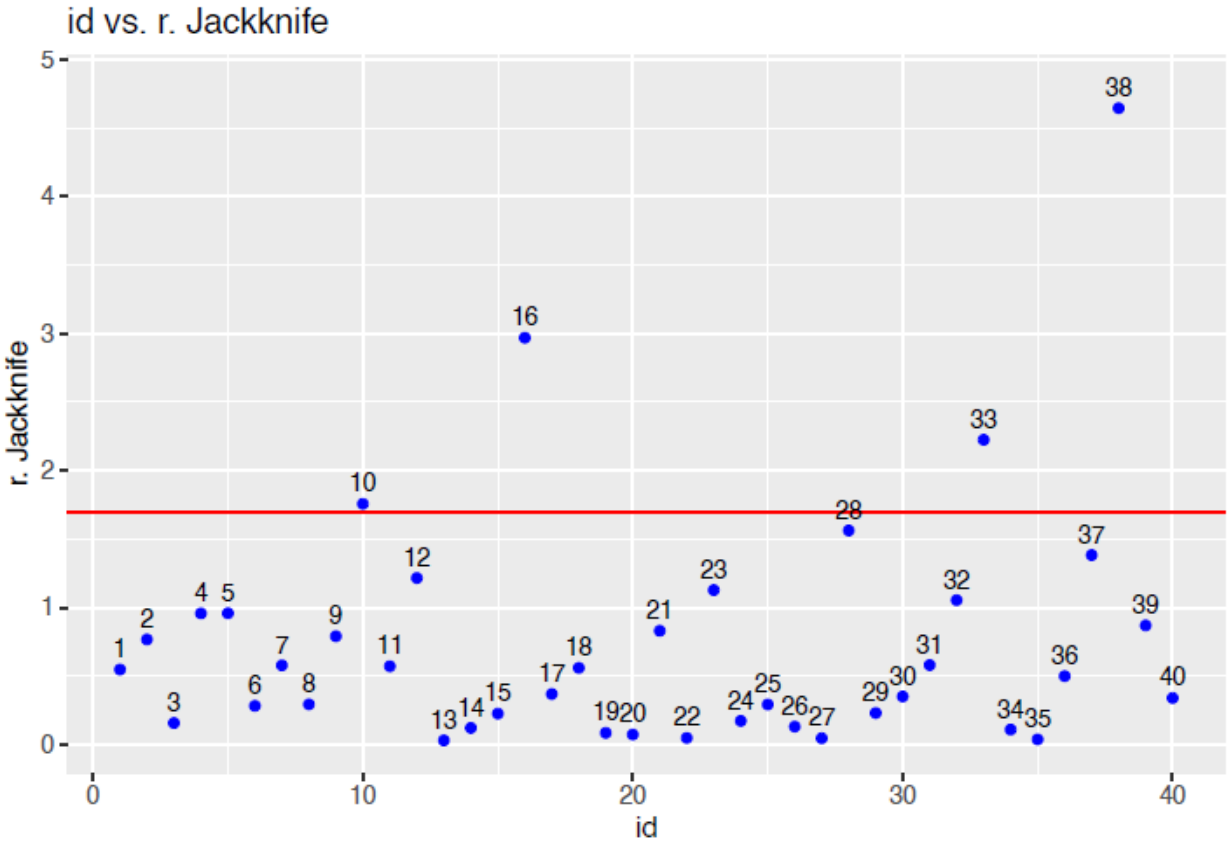

```
ggplot(aes(x=bd.id,y=r.z),data=m.datos.atipicos)+
geom_point(col="blue")+
geom_hline(yintercept =qnorm(0.95),col="red")+
labs(title = "id vs. r. estandarizados",y="r. estandarizados",x="id")+
        geom_text(aes(label=bd.id), size=3,nudge_x=0,nudge_y=0.1)
```



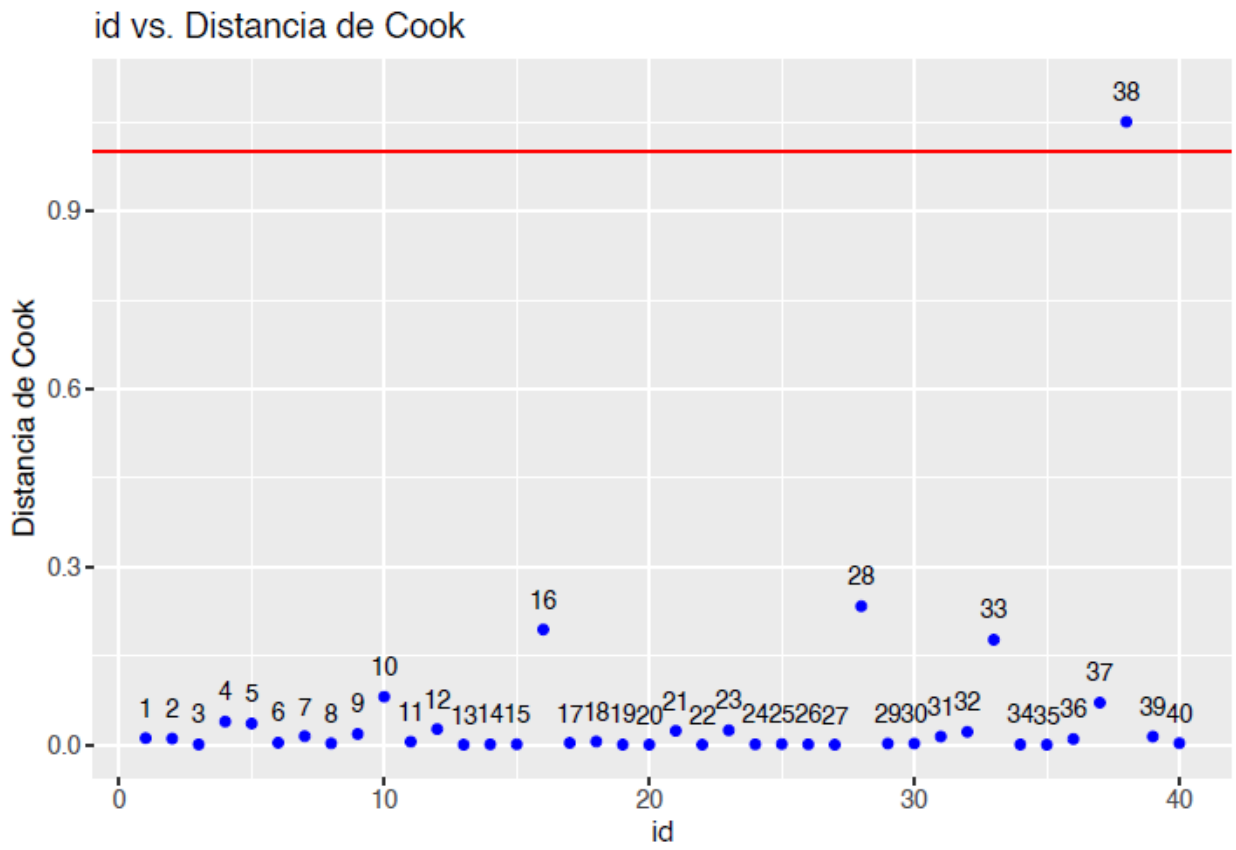

```
ggplot(aes(x=bd.id,y=r.st),data=m.datos.atipicos)+
geom_point(col="blue")+
geom_hline(yintercept =qt(0.95,40-5-1),col="red")+
labs(title = "id vs. r. estudentizados",y="r. estudentizados",x="id")+
        geom_text(aes(label=bd.id), size=3,nudge_x=0,nudge_y=0.1)
```




```
ggplot(aes(x=bd.id,y=r.jack),data=m.datos.atipicos)+
  geom_point(col="blue")+
  geom_hline(yintercept =qt(0.95,40-5-2),col="red")+
  labs(title = "id vs. r. Jackknife",y="r. Jackknife",x="id")+
  geom_text(aes(label=bd.id), size=3,nudge_x=0,nudge_y=0.15)
```




```
ggplot(aes(x=bd.id,y=cook),data=m.datos.atipicos)+
  geom_point(col="blue")+
  geom_hline(yintercept =1,col="red")+
  labs(title = "id vs. Distancia de Cook",y="Distancia de Cook",x="id")+
  geom_text(aes(label=bd.id), size=3,nudge_x=0,nudge_y=0.05)
```



De los 5 gráficos anteriores, h_i identifica a los sujetos $id = 28\ y\ 38$ como posible datos atípicos, los tres tipos de residuales identifican a los sujetos con $id = 10, 16, 33\ y\ 38$ y la distancia de Cook identifica únicamente al sujeto $id = 38$. Como resultado de este análisis y apoyado en la discusión con expertos se decide retirar las observaciones de los sujetos con $id = 16\ y\ 38$ de la base de datos. El sujeto $id = 38$, presenta valores muy altos en todas las variables cuantitativas, y el sujeto $id = 16$, presente un valor muy bajo en DBP, pero por el contrario presenta valores muy altos en LCC y DOF.

El siguiente código se puede utilizar para eliminar filas de la base de datos, pero debemos verificar que el id corresponda a la fila correspondiente (como ocurre en nuestro caso):

```
bd<-arrange(bd,id)
bd<-bd[-c(16,38),]
```

Evaluación de los supuestos del modelo

En las primeras sesiones de la asignatura, se presentaron los supuestos que soportan los resultados obtenidos al ajustar un modelo de regresión lineal. La evaluación de estos supuestos se realiza en ocasiones a través de gráficos y en otras a través de pruebas estadísticas. Un primer supuesto a evaluar es el supuesto de linealidad, donde se plantea que la media de y para un valor fijo de x $(\mu_{y|x})$ se expresa a partir de una combinación lineal de las p variables independientes. Una primera evaluación es construir p gráficos de dispersión tomando cada una de las variables independientes versus el desenlace y. Esta no es una exploración definitiva ya que no considera en conjunto la relación lineal de las variables independientes y el desenlace, pero si permite identificar variables con ciertos patrones de comportamiento particulares o también variables sin ninguna relación con el desenlace. Ahora, una mejor aproximación son los gráficos de regresiones parciales, donde se contrastan los residuales obtenidos del ajuste de los dos siguientes modelos:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{k-1} x_{ik-1}$$

y

$$x_{ik} = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_{k-1} x_{ik-1}$$

El primero ajusta y en función de todas las variables x menos una (x_{ik}), y el segundo modelo ajusta la variable que se retiro en el modelo anterior en función del resto de variables independientes. Los residuales de cada uno de estos dos modelos representan los valores de y_i y de x_{ik} que no ha sido explicado por el resto de las variables independientes y se esperaría que estos tengan una relación lineal entre sí.

Uno de los gráficos más utilizados que permite evaluar los supuestos de linealidad y homocedasticidad, es el gráfico de dispersión entre los valores estimados del desenlace $(\hat{y})$ y los residuales del modelo (estandarizados, estudentizados o jackknife). En la siguiente figura se presentan tres posibles comportamientos de este gráfico:

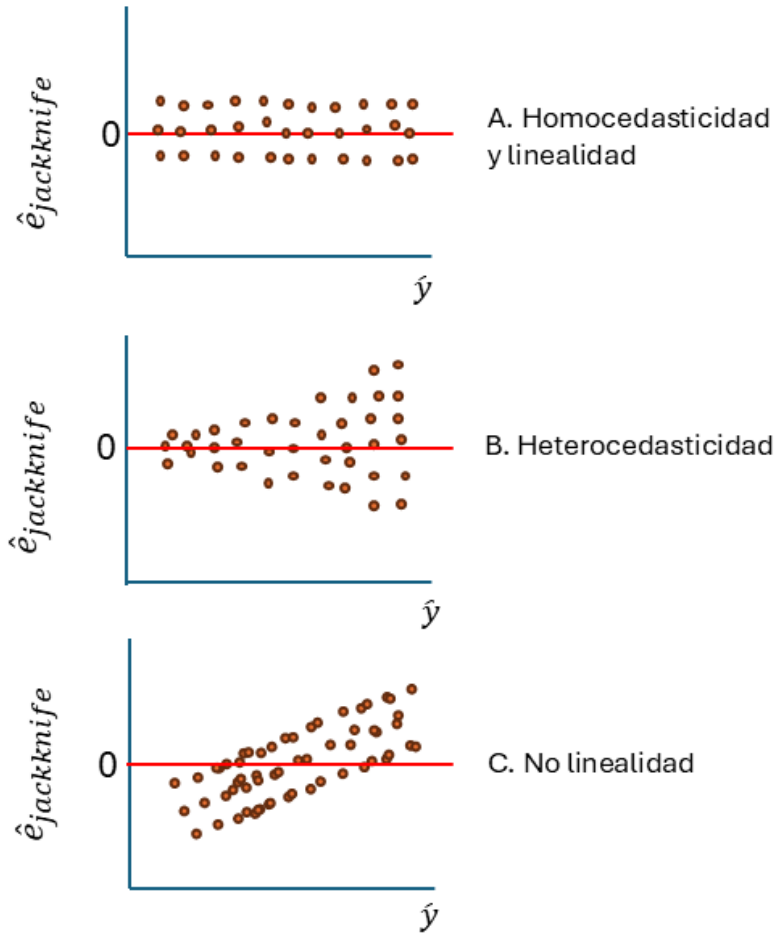


Figure 1: y estimado vs. los residuales de Jackknife

De estos encontramos que:

- **Figura A.** El modelo cumple con el supuesto de linealidad y los residuales tienen un comportamiento aleatorio alrededor de cero sin un patrón particular; se observa una dispersión similar en distinto valores de $\hat{y}_i$ indicando el cumplimiento del supuesto de homocedasticidad.
- **Figura B.** A distintos valores de $\hat{y}$ hay mayor dispersión de los residuales violándose el supuestos de homocedasticidad.

- **Figura C.** hay un comportamiento particular que no se está incluyendo en el modelo y que se puede presentar por violación del supuesto de linealidad o la no incorporación de una variable independiente importante.

Finalmente, frente al supuesto de normalidad sobre los errores: $e \sim N(0, \sigma 2)$, este se puede evaluar con pruebas clásicas de bondad y ajuste como la prueba de lilliefors, la prueba de Shapiro-Wilk o Shapiro-Francia cuya hipótesis nula es que los errores tienen una distribución normal. El resultado de estas pruebas se puede acompañar del gráfico $Q - Q$ plot que contrasta los percentiles de los residuales estimados versus los percentiles teóricos de una distribución normal estándar, esperando observar puntos sobre la linea identidad (recta diagonal a 45 grados: ver ejemplo de práctica).

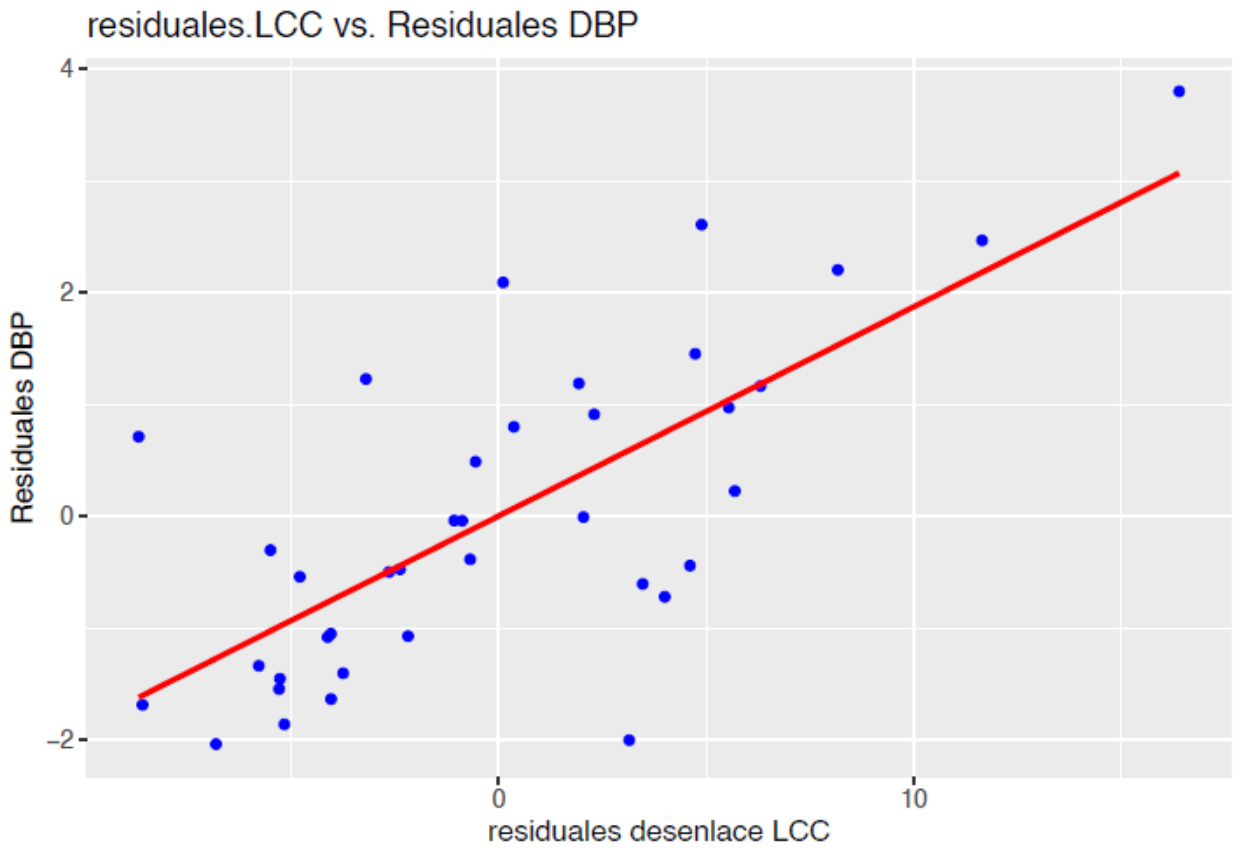
Ejemplo de aplicación 2

Seguiremos trabajando con los datos observados de los 38 fetos (2 observaciones eliminadas por datos atípicos en la práctica anterior) y ajustamos el modelo de regresión con esta base de datos.

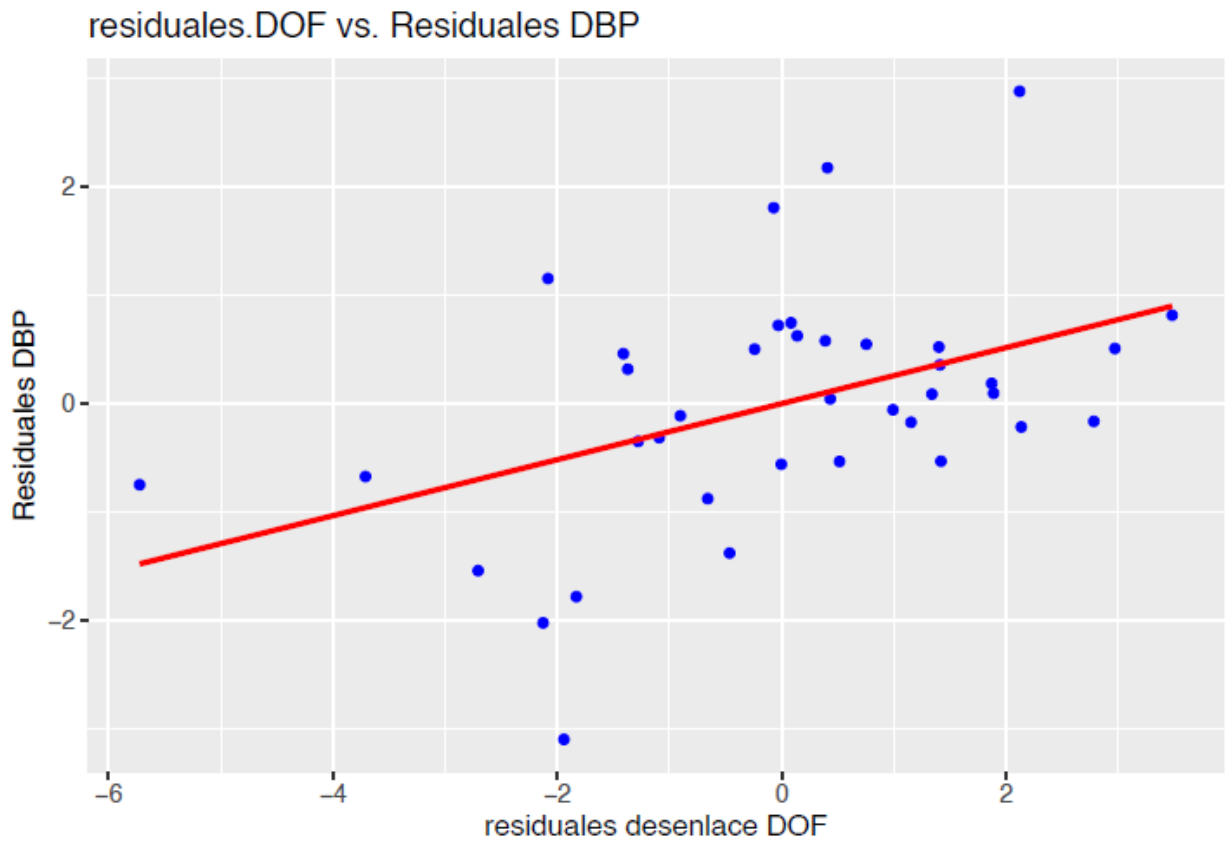
```
m<-lm(DBP~LCC+DOF+edad+sexo,data = bd)
```

A continuación, vamos a obtener los gráficos de regresiones parciales para cada una de las dos variables cuantitativas independientes (LCC y DOF). Para cada una de estas, ajustamos los dos modelos y contrastamos sus residuales en un gráficos de dispersión como se ejemplifica en el siguiente código:

```
# Para la variable LCC:

m.sin_LCC<-lm(DBP~DOF+edad+sexo,data=bd) # modelo sin LCC
m.LCC<-lm(LCC~DOF+edad+sexo,data=bd) # modelo desenlace LCC

rp.LCC<-data.frame(Res.Y=m.sin_LCC$residuals,Res.LCC=m.LCC$residuals) # bd residuales

ggplot(aes(x=Res.LCC,y=Res.Y),data=rp.LCC)+
        geom_point(col="blue")+
        geom_smooth(method = lm,se=FALSE,col="red")+
        labs(title = "residuales.LCC vs. Residuales DBP",
                x="residuales desenlace LCC",
                y="Residuales DBP")
```



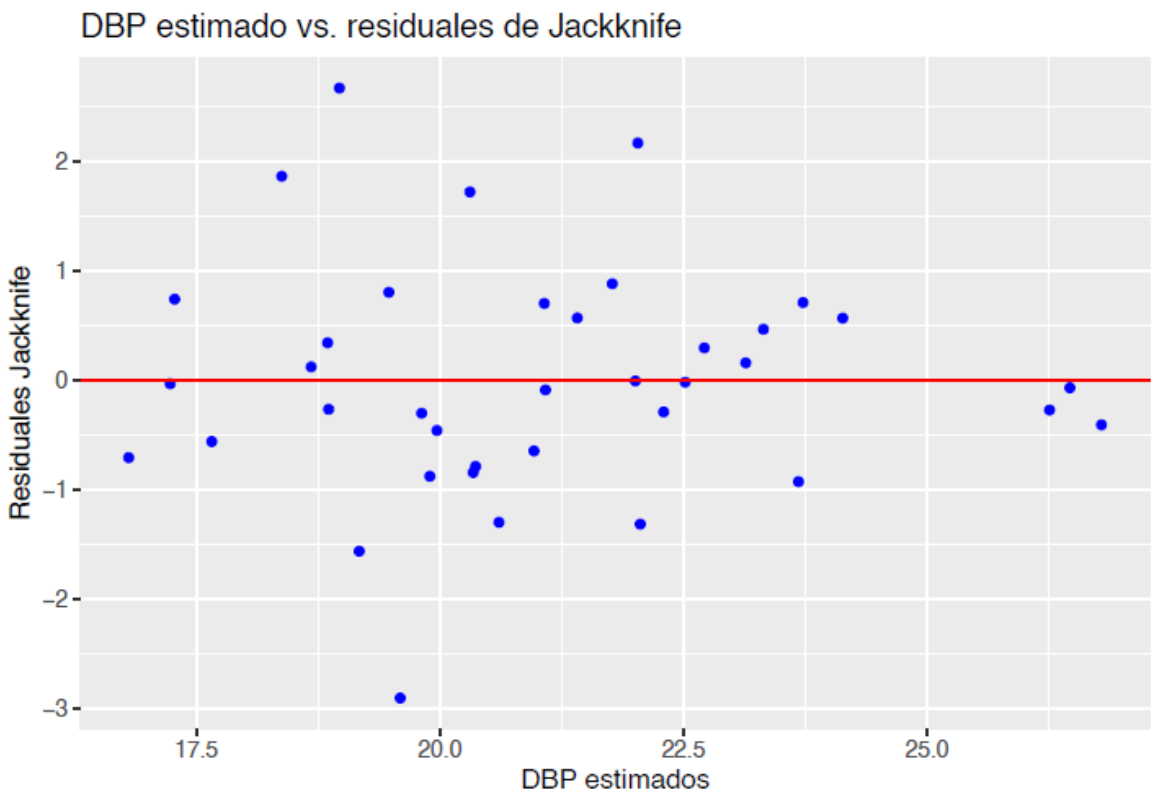

```
# Para la variable DOF:

m.sin_DOF<-lm(DBP~LCC+edad+sexo,data=bd)
m.DOF<-lm(DOF~LCC+edad+sexo,data=bd)

rp.DOF<-data.frame(Res.Y=m.sin_DOF$residuals,Res.DOF=m.DOF$residuals)

ggplot(aes(x=Res.DOF,y=Res.Y),data=rp.DOF)+
        geom_point(col="blue")+
        geom_smooth(method = lm,se=FALSE,col="red")+
        labs(title = "residuales.DOF vs. Residuales DBP",
                x="residuales desenlace DOF",
                y="Residuales DBP")
```



En los dos gráficos anteriores, se observa una tendencia lineal entre los dos residuales indicando que no hay una violación importante del supuesto de linealidad.

A continuamos, se presenta el gráfico de dispersión entre los desenlaces estimados $\hat{y}_i i$ y los residuales de Jackknife.

```
# base de datos con los residuales y los valores estimados de y:
res.fit<-data.frame(yhat=m$fitted.values,jack=rstudent(m))

ggplot(aes(x=yhat,y=jack),data=res.fit)+
        geom_point(col="blue")+
        geom_hline(yintercept =0,col="red")+
        labs(title = "DBP estimado vs. residuales de Jackknife",
                x="DBP estimados",y="Residuales Jackknife")
```



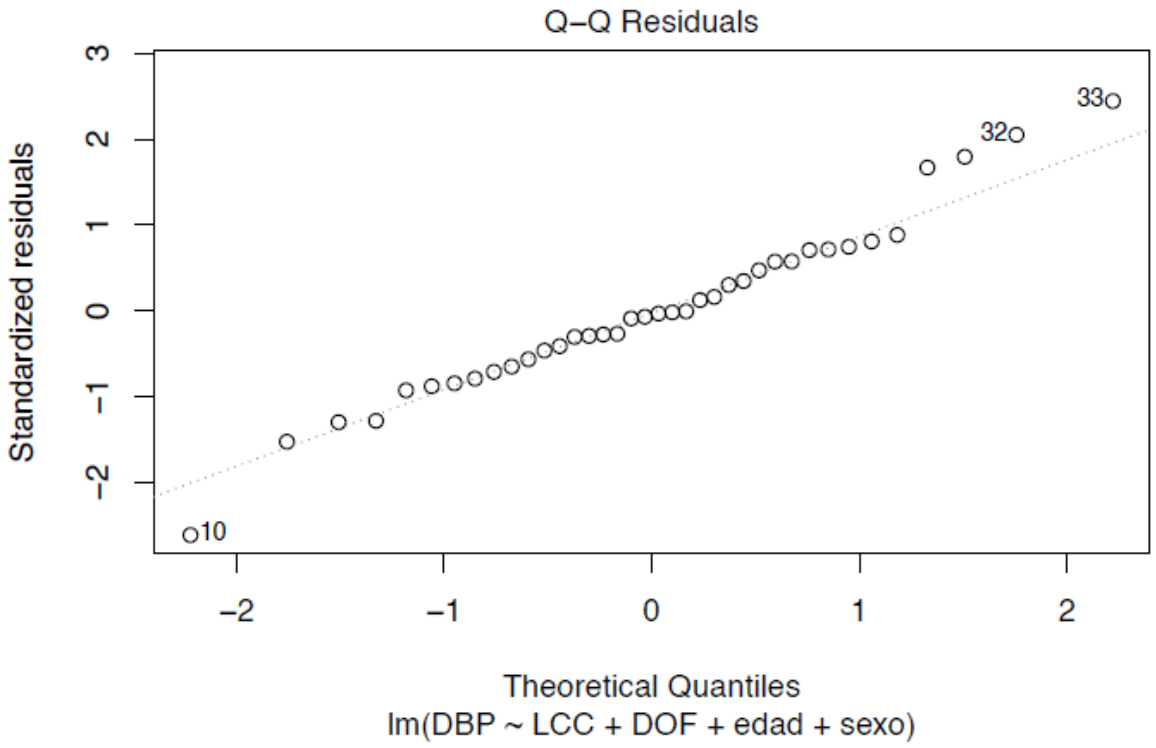
En general no se observa un patrón específico con los residuales alrededor de cero ni un cambio drástico en la dispersión indicando que no hay una violación importante del supuesto de linealidad y homocedasticidad.

Finalmente evaluamos el supuesto de normalidad sobre los residuales con la prueba de lilliefors y el gráfico qqplot:

```
library(nortest)
lillie.test(m$residuals)
```

```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: m$residuals
## D = 0.10045, p-value = 0.432
```

```
plot(m,2)
```



Los residuales del modelos cumplen el supuesto de normalidad.

Colinealidad

Los problemas por **colinealidad** se presentan cuando hay una fuerte relación lineal entre las variables independientes incluidas en el modelo; esto puede generar dos problemas: 1. se presentan estimaciones inestable de los coeficientes, es decir grandes cambios en estos ante leves cambio en los valores de las variables independientes originales y 2. se pueden presentar aumentos (inflaciones) de los errores estándar de los coeficientes estimados ocultando su significancia.

Una exploración preliminar para establecer si hay alguna pareja de variables independientes con una fuerte relación lineal es calculando la matriz de correlación entre todas las posibles parejas de variables independientes, valores con correlaciones muy fuertes ($\rho > 0.75$) sugieren problemas de colinealidad.

Ahora, el diagnóstico de colinealidad se basa en los coeficientes de determinación R^2 producto de ajustar regresiones lineales tomando cada una de las variables independiente como desenlace en función del resto de las variables independientes. Es decir, por ejemplo, para la variable x_2, ajustamos el siguiente modelo:

$$x_2 = b_0 + b_1x_1 + b_3x_3 + \ldots b_px_p$$

de esta regresión, obtenemos su coeficiente de determinación que podemos denotar como R_2^2 e indica la relación lineal entre la variable x_2 y el resto de variables independientes sugiriendo un problema de colinealidad. Repitiendo el proceso para cada variable independiente, tenemos los coeficiente de determinación que denotaremos por R_j^2 donde el subíndice j indica que es el R_2 asociado al modelo donde la variable desenlace es x_j. Los R_j^2 cercanos a 1 (> 0.9) indican que la variable x_j puede presentar un problema de colinealidad con el resto de las variables independientes.

Dos medida adicionales, basadas en los R_j^2, son el **factores de inflación de la varianza** (VIF_j por sus siglas en inglés) y la **Tolerancia**. Los VIF_j son valores proporcionales a los errores estándar de los coeficientes y se obtienen con la siguiente expresión:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Un $VIFj > 10$ sugiere un problemas de colinealidad.

La medida de tolerancia se obtiene con la siguiente expresión:

$$Tolerancia_j = \frac{1}{VIF_j} = 1 - R_j^2$$

Valores cercanos a cero ($Tolerancia j < 0.1$) indicando posibles problemas de colinealidad.

Ejemplo de aplicación 3

Continuamos trabajando con la base de datos de fetos. Primero obtenemos la matriz de correlación entre cada pareja de variables independientes incluidas en el modelo. Dado que las variables edad y sexo son variables categóricas que se ingresan al modelo como variables indicadoras, para el cálculo de la matriz de correlación creamos estas variables y construimos una matriz únicamente con los valores de las variables independientes:

```
library(fastDummies)
bd<-dummy_cols(bd,select_columns = c("edad","sexo"))
X<-matrix(c(bd$LCC,bd$DOF,
            bd$`edad_31 a 35 años`,
            bd$`edad_mayor a 35 años`,
            bd$sexo_M)
          ,nrow = 38,ncol = 5)
cor(X)
```

```
## [,1] [,2] [,3] [,4] [,5]
## [1,] 1.00000000 0.80079322 -0.06331739 0.02902934 0.01259972
## [2,] 0.80079322 1.00000000 -0.06427897 0.10916059 0.05436834
## [3,] -0.06331739 -0.06427897 1.00000000 -0.62994079 -0.21757513
## [4,] 0.02902934 0.10916059 -0.62994079 1.00000000 0.17712298
## [5,] 0.01259972 0.05436834 -0.21757513 0.17712298 1.00000000
```

La matriz de correlación indica, que puede presentarse un problema de colinealidad dada la alta correlación $(\rho > 0.75)$ entre las variables LCC y DOF. A continuación, utilizando la función `multicollinearity` del paquete `performance` para obtener los valor de la inflación de la varianza y la tolerancia para cada variable:

```
library(performance)
multicollinearity(m)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
## Term VIF VIF 95% CI Increased SE Tolerance Tolerance 95% CI
## LCC 2.85 [1.98, 4.50] 1.69 0.35 [0.22, 0.50]
## DOF 2.88 [2.00, 4.54] 1.70 0.35 [0.22, 0.50]
## edad 1.09 [1.00, 3.59] 1.04 0.92 [0.28, 1.00]
## sexo 1.06 [1.00, 8.60] 1.03 0.95 [0.12, 1.00]
```

Ninguno de los valores del VIF (> 10) o la tolerancia (> 0.1) indica un posible problema de colinealidad.

Lecturas complementarias

1. David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, Eli S. Rosenberg. 14. Regression Diagnostics.
   En: Applied Regression Analysis and Other Multivariable Methods. Fifth Edition. Boston, MA: Cengage Learning; 2014. p. 339-400.