TP1 Regresión Lineal Simple

Marcos Buccellato

2023-06-01

Table of Contents

# 1.1.Correlación

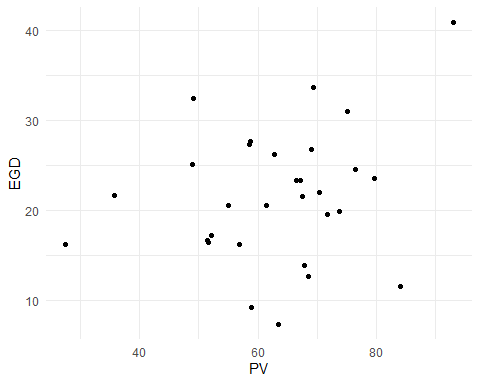
## Ejercicio1.1.

En el archivo grasacerdos.xlsx se encuentran los datos del peso vivo (PV,en Kg) y al epesor de grasa dorsal(EGD,en mm) de 30 lechones elegidos al azar de una población de porcinos Duroc Jersey del Oeste de la provincia de Buenos Aires. Se pide

grasa <- read\_excel('grasacerdos.xlsx')  
grasa$PV <- as.numeric(gsub(",", ".", grasa$PV))  
grasa$EGD <- as.numeric(gsub(",", ".", grasa$EGD))

### (a)Dibujar el diagrama de dispersión e interpretarlo.

ggplot(grasa, aes(PV, EGD)) +   
 geom\_point() + theme\_minimal()

 Como cambia el EGD en función del crecimiento del PV.

### (b)Calcular el coeficiente de correlación muestral y explíquelo.

biNormTest <- mvn(data = grasa[2:3], mvnTest = "hz")  
corCoeff <- cor(grasa$PV,grasa$EGD, method = "pearson")  
print(biNormTest$multivariateNormality)

## Test HZ p value MVN  
## 1 Henze-Zirkler 0.2539437 0.9049686 YES

Como el test de binormalidad de las variables dio positivo, puedo usar el coeficiente de perason como indicador de la correlación entre las variables.

### (c)¿Hay suficiente evidencia para admitir asociación entre el peso y el espesor de grasa? (α=0,05). Verifique los supuestos para decidir el indicador que va a utilizar.

corTest <- cor.test(grasa$PV,grasa$EGD, method = "pearson")

El resultado de un test de correlación (usando Pearson según se vio en (b)) dio negativo con un p valor de 0.1749942 que es mayor al = 0,05

## Ejercicio1.2.

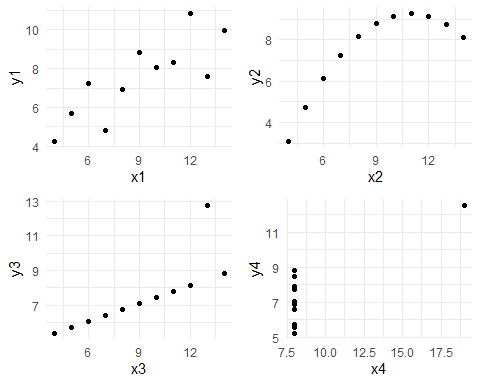
Los datos del cuarteto de Anscombe se encuentran en el archivo anscombe.xlsx

anscombre <- read\_excel('anscombe.xlsx')

Se pide explorar los datos de la siguiente manera:

### (a) Graficar los cuatro pares de datos en un diagrama de dispersión cada uno.

c1 <- ggplot(anscombe, aes(x1, y1)) +   
 geom\_point() + theme\_minimal()  
c2 <- ggplot(anscombe, aes(x2, y2))+   
 geom\_point() + theme\_minimal()  
c3 <- ggplot(anscombe, aes(x3, y3))+   
 geom\_point() + theme\_minimal()  
c4 <- ggplot(anscombe, aes(x4, y4))+   
 geom\_point() + theme\_minimal()  
grid.arrange(c1,c2,c3,c4, ncol = 2, nrow = 2)



### (b) Hallar los valores medios de las variables para cada par de datos.

colMeans(anscombe)

## x1 x2 x3 x4 y1 y2 y3 y4   
## 9.000000 9.000000 9.000000 9.000000 7.500909 7.500909 7.500000 7.500909

### (c) Hallar los valores de la dispersión para cada conjunto de datos.

sapply(anscombe, sd)

## x1 x2 x3 x4 y1 y2 y3 y4   
## 3.316625 3.316625 3.316625 3.316625 2.031568 2.031657 2.030424 2.030579

### (d) Hallar el coeficiente muestral de correlación lineal en cada caso.

mvn(data = anscombe[c(1,5)], mvnTest = "hz")$multivariateNormality$MVN

## [1] "YES"

mvn(data = anscombe[c(2,6)], mvnTest = "hz")$multivariateNormality$MVN

## [1] "NO"

mvn(data = anscombe[c(3,7)], mvnTest = "hz")$multivariateNormality$MVN

## [1] "NO"

mvn(data = anscombe[c(4,8)], mvnTest = "hz")$multivariateNormality$MVN

## [1] "NO"

cor.test(anscombe$x1,anscombe$y1,method="pearson")$p.value

## [1] 0.002169629

cor.test(anscombe$x2,anscombe$y2,method="spearman")$p.value

## [1] 0.02305887

cor.test(anscombe$x3,anscombe$y3,method="spearman")$p.value

## [1] 0

cor.test(anscombe$x4,anscombe$y4,method="spearman")$p.value

## [1] 0.1173068

### (e) Observar, comentar y concluir.

Los primeros pares de variables son binormales por lo tanto puedo usar pearson. Hay correlación porque ep p-valor es bajo. EN los siguientes no podría usar pearson porque no son binormales, por tal motivo uso spearman. Hay correlación salvo en el caso del cuarto par de variables.

# 1.2.Modelo Lineal Simple

## Ejercicio1.3.

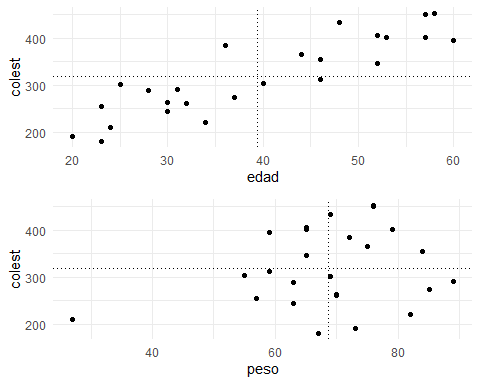
El archivo peso\_edad\_colest.xlsx disponible en contiene registros correspondiente sa 25 individuos respecto de supeso, su edad y el nivel de colesterol total en sangre.

colest <- read\_excel('peso\_edad\_colest.xlsx')

Se pide:

### (a) Realizar el diagrama de dispersión de colesterol en función de la edad y de colesterol en función de peso. Le parece adecuado ajustar un modelo lineal para alguno de estos dos pares de variables?

promedios <- colMeans(colest)  
  
c1 <- ggplot(colest, aes(edad, colest)) +   
 geom\_point() +  
 geom\_vline(xintercept=promedios[2],linetype="dotted") +   
 geom\_hline(yintercept=promedios[3],linetype="dotted") +   
 theme\_minimal()  
c2 <- ggplot(colest, aes(peso, colest)) +   
 geom\_point() +   
 geom\_vline(xintercept=promedios[1],linetype="dotted") +  
 geom\_hline(yintercept=promedios[3],linetype="dotted") +   
 theme\_minimal()  
grid.arrange(c1,c2, ncol = 1, nrow = 2)



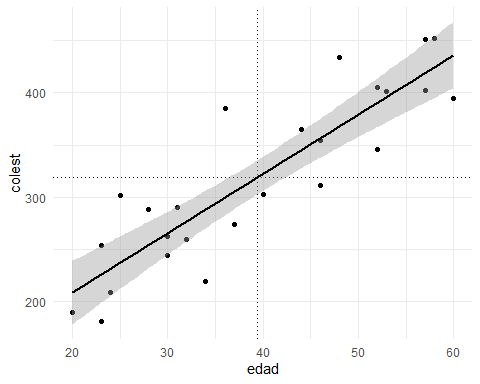
En aparencia colest ~ peso parecería poder ajustarse linealmente.

### (b) Estime los coeficientes del modelo lineal para el colesterol en función de la edad.

model <- lm(colest ~ edad, data = colest)  
model$coefficients

## (Intercept) edad   
## 95.502004 5.670842

(c1+ geom\_smooth(method = "lm", se = TRUE, color = "black") )



### (c) Estime intervalos de confianza del 95% para los coeficientes del modelo y compare estos resultados con el test de Wald para los coeficientes.Le parece que hay asociación entre estos test y el test de la regresión?

cálculo de coeficientes de confianza:

ci <- confint(model, level = 0.95)  
ci

## 2.5 % 97.5 %  
## (Intercept) 41.190390 149.813618  
## edad 4.358216 6.983467

test de Wald:

#install.packages("aod")  
library(aod)  
coef(model)

## (Intercept) edad   
## 95.502004 5.670842

wald.test(Sigma = vcov(model), b = coef(model), Terms = 1)

## Wald test:  
## ----------  
##   
## Chi-squared test:  
## X2 = 13.2, df = 1, P(> X2) = 0.00028

No se muy bien que concluir. Pero entiendo que el test de wald en este caso probaría que existe un coeficiente que es diferente de cero de forma significativa. En este caso el intervalo del coeficiente no incluye al cero, por lo cual sería consistente.

### (d) A partir de esta recta estime los valores de E(Y) para x=25 años y x=48 años. Podría estimarse el valor de E(Y) para x=80 años?

predict(model, newdata = data.frame(edad = c(25,80)))

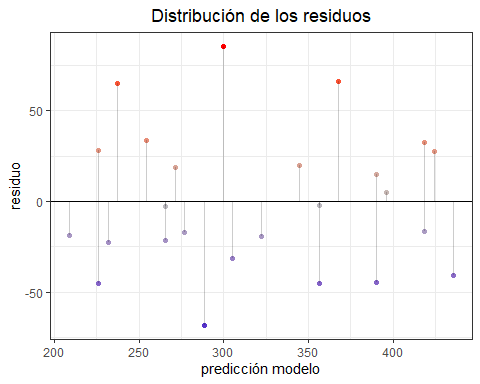
## 1 2   
## 237.2730 549.1693

### (e) Testee la normalidad de los residuos y haga un gráfico para ver si son homocedásticos.

shapiro.test(model$residuals)

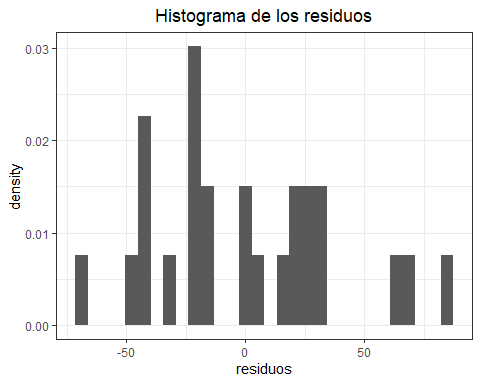
##   
## Shapiro-Wilk normality test  
##   
## data: model$residuals  
## W = 0.96478, p-value = 0.5175

colest2<-colest  
colest2$prediccion <- model$fitted.values   
colest2$residuos <- model$residuals  
  
  
ggplot(data = colest2, aes(x = prediccion, y = residuos)) +   
 geom\_point(aes(color = residuos)) +   
 scale\_color\_gradient2(low = "blue3", mid = "grey", high = "red") +   
 geom\_hline(yintercept = 0) + geom\_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +   
 labs(title = "Distribución de los residuos", x = "predicción modelo", y = "residuo") +   
 theme\_bw() +   
 theme(plot.title = element\_text(hjust = 0.5), legend.position = "none")

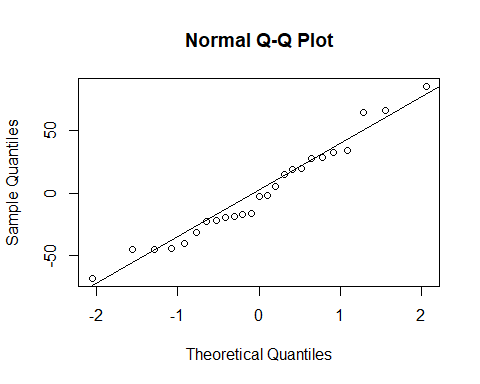


ggplot(data = colest2, aes(x = residuos)) + geom\_histogram(aes(y = ..density..)) +   
 labs(title = "Histograma de los residuos") + theme\_bw() +   
 theme(plot.title = element\_text(hjust = 0.5))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



qqnorm(model$residuals)   
qqline(model$residuals)



Están normalmente distribuidos porque el test falló en rechazar la hipótesis nula de que no están normalmente distribuidos. En el qqplot se pueden ver los valores alineados a la diagonal, lo cual indica coincidencia cercana con los cuantiles teóricos.

#install.packages("lmtest")  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

bptest(model)

##   
## studentized Breusch-Pagan test  
##   
## data: model  
## BP = 0.15821, df = 1, p-value = 0.6908

Concluimos que se rechaza heteroscedasticidad porque el p-valor para rechazar homocedasticidad no es menor a 0.05 siendo los residuos normales y la relación lineal.

Habría que ver la independencia de las variables para ver si el resultado se mantiene, aplicamos el test de Durbin-Watson:

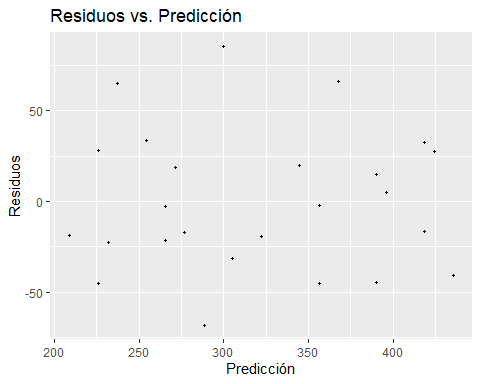
library(car)  
dwt(model)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.1019283 1.783079 0.576  
## Alternative hypothesis: rho != 0

La hipótesis nula de este test es la no autocorrelación, en este caso no se puede descartar la misma. Esto implica que no se sostiene el supuesto de independencia. Por tal motivo la conclusión del test anterior no se puede aplicar.

Graficamos:

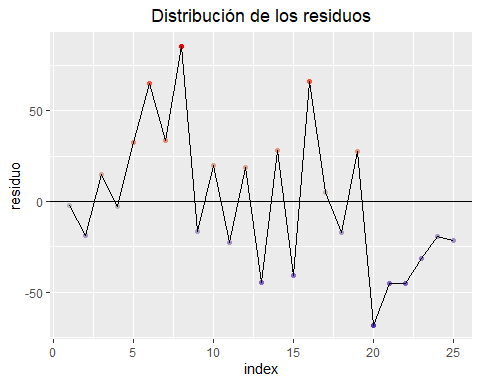
ggplot(colest2, aes(x = prediccion, y = residuos)) +  
 geom\_point(shape = 20) +  
 labs(x = "Predicción", y = "Residuos",  
 title = "Residuos vs. Predicción")



No parece haber un patrón entre los residuos, por lo cual no se puede ver heterocedasticidad. En clase el código para ver esto era este:

ggplot(data = colest2, aes(x = seq\_along(residuos), y = residuos)) +   
 geom\_point(aes(color = residuos)) +   
 scale\_color\_gradient2(low = "blue3", mid = "grey", high = "red") +   
 geom\_line(size = 0.3) + labs(title = "Distribución de los residuos", x = "index", y = "residuo")+   
 geom\_hline(yintercept = 0) +   
 theme(plot.title = element\_text(hjust = 0.5), legend.position = "none")

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



No me queda claro como interpretarlo en este caso

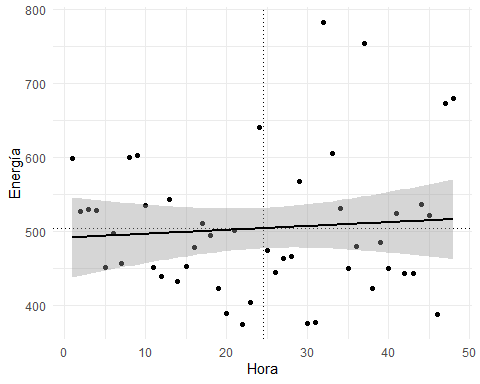
## Ejercicio1.4. Una empresa desarrolló un sistema de energía solar para calentar el agua para una caldera que es parte del sistema de energía del proceso productivo.Existe el interés de controlar la estabilidad del sistema, para ello se monitorea el mismo y se registran los datos cada hora. Los datos se encuentran disponibles en el archivo energia.xlsx

energia <- read\_excel('energia.xlsx')

### (a) Realizar el diagrama de dispersión y evaluar si un modelo de regresión lineal es adecuado.

promediosE <- colMeans(energia)  
ggplot(energia, aes(Hora, Energía)) +   
 geom\_point() +  
 geom\_vline(xintercept=promediosE[1],linetype="dotted") +   
 geom\_hline(yintercept=promediosE[2],linetype="dotted") +   
 geom\_smooth(method = "lm", se = TRUE, color = "black") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



No parecería ser adecuado un modelo lineal

### (b) Estimar un modelo lineal y verificar la normalidad de los residuos del mismo.

modelE <- lm(Energía ~ Hora, data = energia)  
  
shapiro.test(modelE$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: modelE$residuals  
## W = 0.92988, p-value = 0.006752

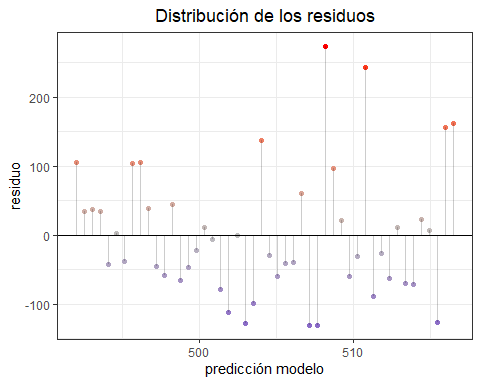
No están normalmente distribuidos porque el test rechaza la hipótesis nula de que no están normalmente distribuidos.

summary(modelE)

##   
## Call:  
## lm(formula = Energía ~ Hora, data = energia)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -131.12 -60.60 -24.31 37.29 273.84   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 491.4894 27.5044 17.869 <2e-16 \*\*\*  
## Hora 0.5208 0.9772 0.533 0.597   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 93.79 on 46 degrees of freedom  
## Multiple R-squared: 0.006138, Adjusted R-squared: -0.01547   
## F-statistic: 0.2841 on 1 and 46 DF, p-value: 0.5966

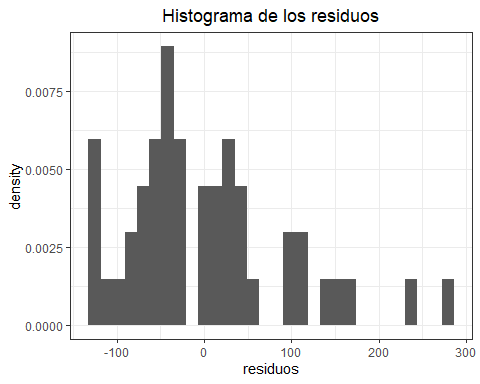
Más aun porque el R ajustado es bajisimo, el p-value de los coefficientes tambien y el p-value del F-test también…..

energia2<-energia  
energia2$prediccion <- modelE$fitted.values   
energia2$residuos <- modelE$residuals  
  
  
ggplot(data = energia2, aes(x = prediccion, y = residuos)) +   
 geom\_point(aes(color = residuos)) +   
 scale\_color\_gradient2(low = "blue3", mid = "grey", high = "red") +   
 geom\_hline(yintercept = 0) + geom\_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +   
 labs(title = "Distribución de los residuos", x = "predicción modelo", y = "residuo") +   
 theme\_bw() +   
 theme(plot.title = element\_text(hjust = 0.5), legend.position = "none")

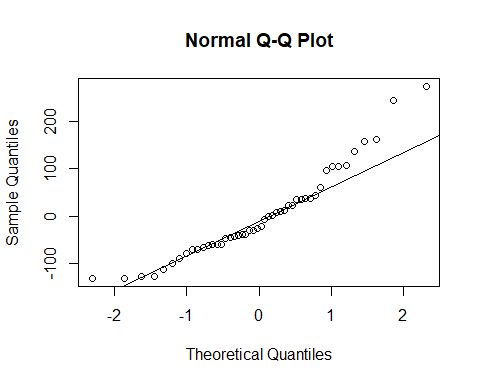


ggplot(data = energia2, aes(x = residuos)) + geom\_histogram(aes(y = ..density..)) +   
 labs(title = "Histograma de los residuos") + theme\_bw() +   
 theme(plot.title = element\_text(hjust = 0.5))

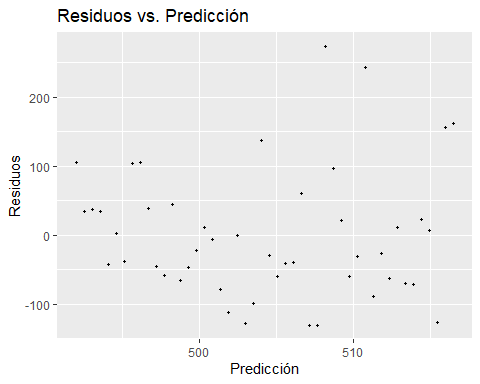
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



qqnorm(modelE$residuals)   
qqline(modelE$residuals)

 Graficamos:

ggplot(energia2, aes(x = prediccion, y = residuos)) +  
 geom\_point(shape = 20) +  
 labs(x = "Predicción", y = "Residuos",  
 title = "Residuos vs. Predicción")

 Aparentemente en este caso no hay normalidad porque hay agrupamiento de los residuos habiendo algunos puntos que se alejan mucho del resto

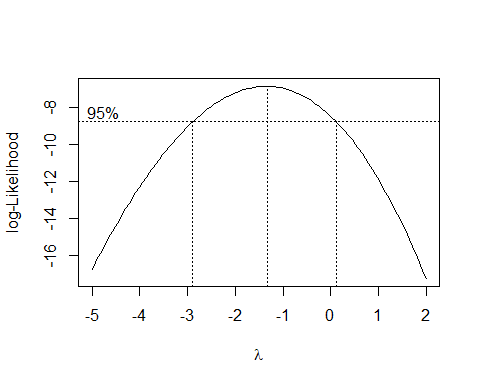
### (c) En caso de rechazar este supuesto buscar una transformación lineal para este modelo y aplicarla.

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

box\_cox\_result <-boxcox(Energía ~ Hora, lambda = -5:2, data = energia)



best\_box\_cox <- box\_cox\_result$x[which.max(box\_cox\_result$y)]  
modelE2 <- lm((Energía)^(best\_box\_cox) ~ Hora, data = energia)  
summary(modelE2)

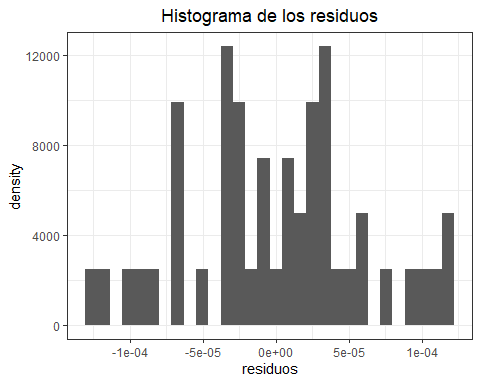
##   
## Call:  
## lm(formula = (Energía)^(best\_box\_cox) ~ Hora, data = energia)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.290e-04 -3.263e-05 3.849e-06 3.599e-05 1.150e-04   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.779e-04 1.787e-05 15.55 <2e-16 \*\*\*  
## Hora -1.251e-08 6.350e-07 -0.02 0.984   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.094e-05 on 46 degrees of freedom  
## Multiple R-squared: 8.444e-06, Adjusted R-squared: -0.02173   
## F-statistic: 0.0003884 on 1 and 46 DF, p-value: 0.9844

shapiro.test(modelE2$residuals)

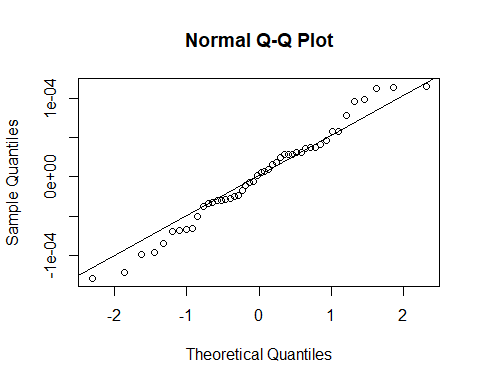
##   
## Shapiro-Wilk normality test  
##   
## data: modelE2$residuals  
## W = 0.98002, p-value = 0.5796

energia3<-energia  
energia3$Energía <- log(energia$Energía)  
energia3$prediccion <- modelE2$fitted.values   
energia3$residuos <- modelE2$residuals  
  
  
ggplot(data = energia3, aes(x = residuos)) + geom\_histogram(aes(y = ..density..)) +   
 labs(title = "Histograma de los residuos") + theme\_bw() +   
 theme(plot.title = element\_text(hjust = 0.5))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



qqnorm(modelE2$residuals)   
qqline(modelE2$residuals)



### (d) Realizar el análisis diagnóstico del nuevo modelo y estimar un intervalo de confianza y un intervalo de predicción para 27.5 hs con ambos modelos. Comparar los intervalos.

shapiro.test(modelE2$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: modelE2$residuals  
## W = 0.98002, p-value = 0.5796

Con el test de shapiro-wilk ahora no se descarta normalidad.

Los intervalos de confianza son:

ci <- confint(modelE2, level = 0.95)  
ci

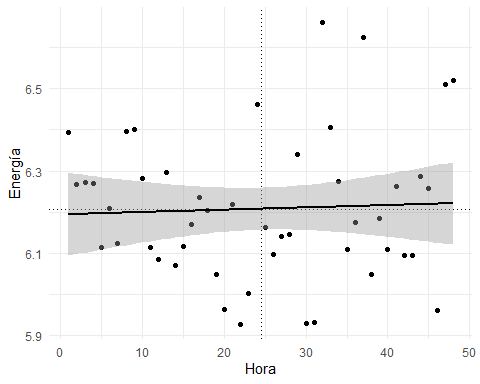
## 2.5 % 97.5 %  
## (Intercept) 2.419201e-04 3.138658e-04  
## Hora -1.290618e-06 1.265591e-06

Predicción:

predict(modelE2, newdata = data.frame(Hora = c(27.5)))

## 1   
## 0.0002775488

promediosE2 <- colMeans(energia3)  
ggplot(energia3, aes(Hora, Energía)) +   
 geom\_point() +  
 geom\_vline(xintercept=promediosE2[1],linetype="dotted") +   
 geom\_hline(yintercept=promediosE2[2],linetype="dotted") +   
 geom\_smooth(method = "lm", se = TRUE, color = "black") +  
 theme\_minimal()



No termino de entender el sentido de este ejercicio, a las claras se ve que no hay un fit muy adecuado con una recta ¿para que tanta manipulación? Es como querer encajarlo a martillazos

# 1.4.Tratamiento de la heterocedasticidad

## Ejercicio1.5. Se obtuvieron datos históricos del mercado inmobiliario de una ciudad de Nueva Taipei,en Taiwan.La base es inmobiliaria.xlsx.

Las características son:

edad: Edad de la propiedad (en años). distancia: La distancia a la estación de transporte más cercana (en metros). negocios: Cantidad de negocios de conveniencia en las cercanías a una distancia realizable a pie. latitud: Latitud de la ubicación de la propiedad (en grados). longitud: Longitud de la ubicación de la propiedad (en grados). precio: Precio por metro cuadrado (en miles de dólares)

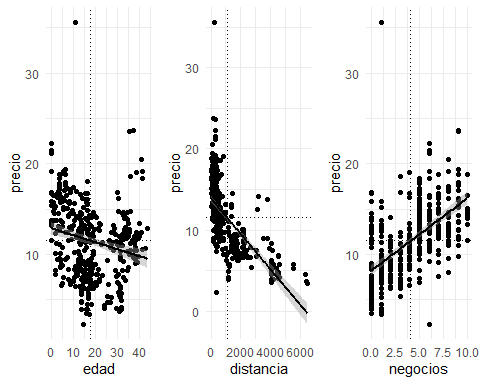
propiedades <- read\_delim('inmobiliaria.csv',col\_names = TRUE,delim=";")

Se quiere investigar si el precio de las propiedades puede ser estimado en función de alguna de las variables disponibles.

### (a) Analizar si el precio depende de alguna de las variables.

No me queda claro el sentido de trabajar con latitud y longitud, sobre todo por separado. Quizas estaría bueno hacer una regresion multivariada en ese caso. En tal caso haría algo así: <http://www.geo.hunter.cuny.edu/~ssun/R-Spatial/spregression.html>

promediosP <- colMeans(propiedades)  
c1 <- ggplot(propiedades, aes(edad, precio)) +   
 geom\_point() +  
 geom\_vline(xintercept=promediosP[1],linetype="dotted") +   
 geom\_hline(yintercept=promediosP[6],linetype="dotted") +  
 geom\_smooth(method = "lm", se = TRUE, color = "black") +  
 theme\_minimal()  
c2 <- ggplot(propiedades, aes(distancia, precio)) +   
 geom\_point() +   
 geom\_vline(xintercept=promediosP[2],linetype="dotted") +  
 geom\_hline(yintercept=promediosP[6],linetype="dotted") +   
 geom\_smooth(method = "lm", se = TRUE, color = "black") +  
 theme\_minimal()  
c3 <- ggplot(propiedades, aes(negocios, precio)) +   
 geom\_point() +  
 geom\_vline(xintercept=promediosP[3],linetype="dotted") +   
 geom\_hline(yintercept=promediosP[6],linetype="dotted") +   
 geom\_smooth(method = "lm", se = TRUE, color = "black") +  
 theme\_minimal()  
  
grid.arrange(c1,c2,c3, ncol = 3, nrow = 1)



biNormTest <- mvn(data = propiedades[c(6,1)], mvnTest = "hz")  
print(biNormTest$multivariateNormality$MVN)

## [1] "NO"

biNormTest <- mvn(data = propiedades[c(6,2)], mvnTest = "hz")  
print(biNormTest$multivariateNormality$MVN)

## [1] "NO"

biNormTest <- mvn(data = propiedades[c(6,3)], mvnTest = "hz")  
print(biNormTest$multivariateNormality$MVN)

## [1] "NO"

En ninguno de los tres casos da positivo el test de binormalidad, por lo tanto para ver la correlación voy a usar spearman

cor.test(propiedades$precio,propiedades$edad,method="spearman")$p.value

## [1] 5.210699e-09

cor.test(propiedades$precio,propiedades$distancia,method="spearman")$p.value

## [1] 2.824113e-83

cor.test(propiedades$precio,propiedades$negocios,method="spearman")$p.value

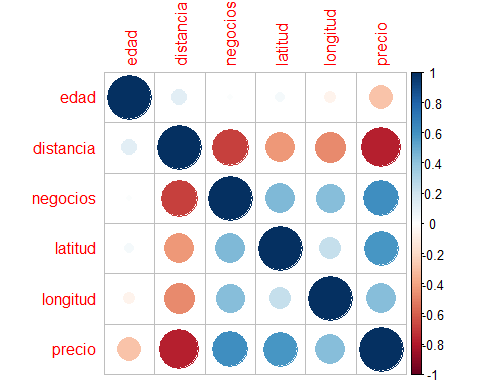
## [1] 9.711186e-45

En este caso estarían todas correlacionadas

library(corrplot)

## corrplot 0.92 loaded

corrplot(cor(propiedades,method="s"))



### (b) Estudiar la linealidad de la relación precio-distancia.

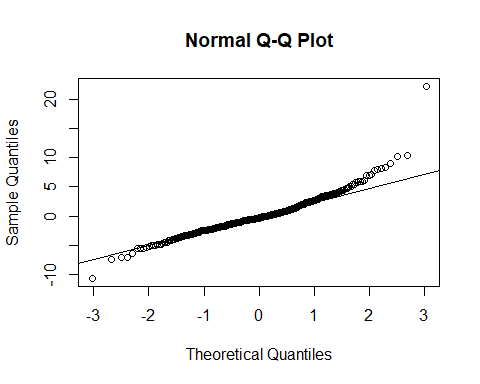
modelProp <- lm(precio ~ distancia, data = propiedades)  
  
shapiro.test(modelProp$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: modelProp$residuals  
## W = 0.93207, p-value = 1.085e-12

No estan normalmente distribuidos los residuos porque el test valida la hipotesis alternativa de que no lo estan.

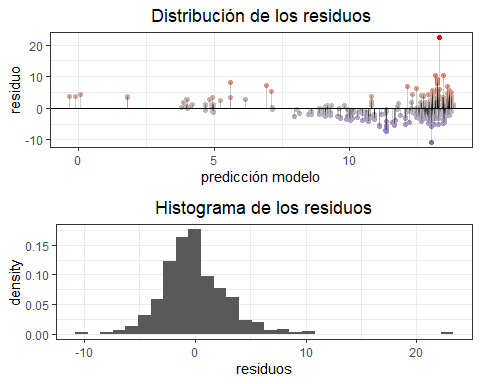
### (c) Estimar los coeficientes del modelo y realizar el análisis diagnóstico de los residuos del mismo. Utilizar para este análisis los gráficos de residuos versus valores ajustados, el qq-plot de los residuos, la grafica de residuos versus leverage.

prop2<-propiedades  
prop2$prediccion <- modelProp$fitted.values   
prop2$residuos <- modelProp$residuals  
  
d1 <- ggplot(data = prop2, aes(x = prediccion, y = residuos)) +   
 geom\_point(aes(color = residuos)) +   
 scale\_color\_gradient2(low = "blue3", mid = "grey", high = "red") +   
 geom\_hline(yintercept = 0) + geom\_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +   
 labs(title = "Distribución de los residuos", x = "predicción modelo", y = "residuo") +   
 theme\_bw() +   
 theme(plot.title = element\_text(hjust = 0.5), legend.position = "none")  
  
  
d2<- ggplot(data = prop2, aes(x = residuos)) + geom\_histogram(aes(y = ..density..)) +   
 labs(title = "Histograma de los residuos") + theme\_bw() +   
 theme(plot.title = element\_text(hjust = 0.5))  
qqnorm(modelProp$residuals)   
qqline(modelProp$residuals)



grid.arrange(d1,d2,nrow = 2)

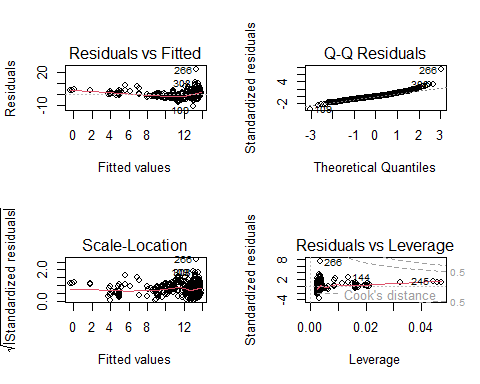
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



De todas maneras, esto parece ser más producto de outliers que otra cosa.

Otra forma de verlo:

par(mfrow=c(2,2))  
plot(modelProp)



par(mfrow=c(1,1))

El dato 266 al menos parece ser outlier

### (d) Aplicar los test de Durbin-Watson Breush-Pagan.

#Durbin-Watson  
library(lmtest)  
dwtest(modelProp,alternative ="two.sided",iterations=1000)

##   
## Durbin-Watson test  
##   
## data: modelProp  
## DW = 2.1607, p-value = 0.1037  
## alternative hypothesis: true autocorrelation is not 0

No rechaza la hipótesis de NO autocorrelación y no tienen distribución normal…. no serían independientes pero No me queda claro como sería el tema cuando no es normal la distribución como interpretar el test de independencia.

### (e) Analice la presencia de outlier y verifique si coinciden con los puntos influyentes.

summary(influence.measures(model = modelProp))

## Potentially influential observations of  
## lm(formula = precio ~ distancia, data = propiedades) :  
##   
## dfb.1\_ dfb.dstn dffit cov.r cook.d hat   
## 9 -0.10 0.23 0.24\_\* 1.03\_\* 0.03 0.03\_\*  
## 12 0.15 -0.08 0.15 0.97\_\* 0.01 0.00   
## 26 -0.05 0.12 0.13 1.02\_\* 0.01 0.02\_\*  
## 31 -0.04 0.13 0.14 1.02\_\* 0.01 0.02\_\*  
## 36 0.00 -0.01 -0.01 1.02\_\* 0.00 0.02\_\*  
## 37 -0.01 0.02 0.02 1.02\_\* 0.00 0.02\_\*  
## 44 -0.01 0.01 0.02 1.03\_\* 0.00 0.02\_\*  
## 45 0.00 0.00 0.00 1.03\_\* 0.00 0.02\_\*  
## 51 -0.08 -0.01 -0.12 0.98\_\* 0.01 0.00   
## 54 -0.05 0.13 0.13 1.02\_\* 0.01 0.02\_\*  
## 69 -0.01 0.05 0.05 1.02\_\* 0.00 0.02\_\*  
## 83 -0.01 0.02 0.02 1.02\_\* 0.00 0.02\_\*  
## 85 -0.03 0.09 0.10 1.02\_\* 0.01 0.01\_\*  
## 101 0.16 -0.08 0.16 0.97\_\* 0.01 0.00   
## 109 -0.20 0.10 -0.20 0.95\_\* 0.02 0.00   
## 112 -0.13 0.27 0.28\_\* 1.04\_\* 0.04 0.05\_\*  
## 113 0.01 -0.03 -0.03 1.02\_\* 0.00 0.02\_\*  
## 122 0.10 -0.03 0.12 0.98\_\* 0.01 0.00   
## 144 -0.08 0.29 0.31\_\* 0.98\_\* 0.05 0.01   
## 150 -0.02 0.05 0.06 1.02\_\* 0.00 0.02\_\*  
## 151 0.00 -0.01 -0.01 1.02\_\* 0.00 0.02\_\*  
## 158 0.02 -0.06 -0.06 1.02\_\* 0.00 0.02\_\*  
## 162 0.17 -0.09 0.18 0.97\_\* 0.02 0.00   
## 166 -0.01 0.02 0.02 1.03\_\* 0.00 0.02\_\*  
## 172 -0.03 0.09 0.09 1.02\_\* 0.00 0.02\_\*  
## 176 -0.01 0.02 0.03 1.02\_\* 0.00 0.02\_\*  
## 179 -0.01 0.03 0.03 1.02\_\* 0.00 0.02\_\*  
## 185 -0.02 0.05 0.05 1.02\_\* 0.00 0.02\_\*  
## 216 0.21 -0.12 0.21 0.95\_\* 0.02 0.00   
## 222 0.01 -0.04 -0.04 1.02\_\* 0.00 0.02\_\*  
## 224 -0.04 0.19 0.23\_\* 0.99 0.03 0.01   
## 227 0.01 -0.02 -0.02 1.02\_\* 0.00 0.02\_\*  
## 228 -0.02 0.05 0.05 1.02\_\* 0.00 0.02\_\*  
## 245 -0.14 0.31 0.32\_\* 1.04\_\* 0.05 0.04\_\*  
## 247 -0.08 -0.01 -0.12 0.98\_\* 0.01 0.00   
## 251 -0.09 0.20 0.21\_\* 1.03\_\* 0.02 0.03\_\*  
## 266 0.46 -0.25 0.46\_\* 0.76\_\* 0.09 0.00   
## 294 0.00 0.01 0.01 1.02\_\* 0.00 0.02\_\*  
## 308 0.20 -0.10 0.20 0.95\_\* 0.02 0.00   
## 316 -0.01 0.04 0.04 1.02\_\* 0.00 0.02\_\*  
## 325 -0.02 0.05 0.05 1.02\_\* 0.00 0.02\_\*  
## 326 -0.09 -0.01 -0.12 0.98\_\* 0.01 0.00   
## 327 -0.01 0.03 0.03 1.03\_\* 0.00 0.02\_\*  
## 343 -0.13 0.27 0.28\_\* 1.05\_\* 0.04 0.05\_\*  
## 375 0.15 -0.08 0.15 0.98\_\* 0.01 0.00   
## 380 0.01 -0.04 -0.04 1.02\_\* 0.00 0.02\_\*  
## 385 0.14 -0.09 0.14 0.98\_\* 0.01 0.00   
## 390 -0.04 0.11 0.12 1.02\_\* 0.01 0.02\_\*  
## 405 0.00 -0.01 -0.01 1.02\_\* 0.00 0.02\_\*

# 1.5. Cuadrados Mínimos Ponderados

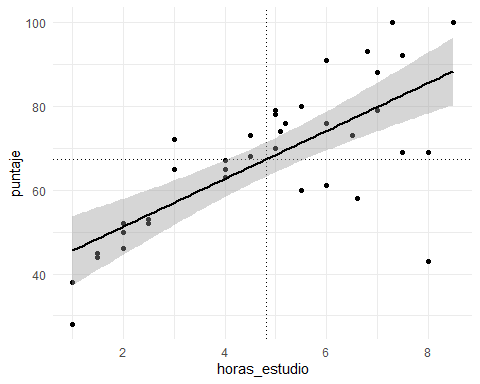
## Ejercicio1.6. En la base estudio.xlsx se encuentran registradas las horas de estudios referidas por un conjunto de estudiantes y su calificación en la evaluación final.

estudio16 <- read\_delim('estudio.csv',col\_names = TRUE,delim=";",show\_col\_types = FALSE)

### (a) Ajuste un modelo de regresión simple para estimar la nota final en función de las horas dedicadas al estudio.

model16 <- lm(puntaje ~ horas\_estudio, data = estudio16)  
promedios16 <- colMeans(estudio16)  
ggplot(estudio16, aes(horas\_estudio, puntaje)) +   
 geom\_point() +  
 geom\_vline(xintercept=promedios16[1],linetype="dotted") +   
 geom\_hline(yintercept=promedios16[2],linetype="dotted") +  
 geom\_smooth(method = "lm", se = TRUE, color = "black") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



### (b) Estudie el cumplimiento de los supuestos del modelo, gráfica y analíticamente.

**Análisis gráfico**

De la gráfica que se puede ver en el punto **a)** encontramos que hay algunos puntos que parecen ser outliers e influyentes. Vemos que en apariencia la estimación lineal puede ser apropiada pero es probable que estos puntos influyentes afecten la estimación.

**Indicadores**

Testeamos la binormalidad primero para ver que método usamos para estimar correlación.

biNormTest16 <- mvn(data = estudio16, mvnTest = "hz")  
print(biNormTest16$multivariateNormality)

## Test HZ p value MVN  
## 1 Henze-Zirkler 1.239238 0.00368416 NO

Como este test da negativo, entonces debo usar el método de Spearman para ver la correlación.

corTest16 <- cor.test(estudio16$horas\_estudio, estudio16$puntaje, method = "spearman")

## Warning in cor.test.default(estudio16$horas\_estudio, estudio16$puntaje, :  
## Cannot compute exact p-value with ties

corTest16

##   
## Spearman's rank correlation rho  
##   
## data: estudio16$horas\_estudio and estudio16$puntaje  
## S = 2827.4, p-value = 7.18e-06  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.664846

Da una correlación positiva con un p-valor muy bajo, verificando lo sospechado en el análisis gráfico.

summary(model16)

##   
## Call:  
## lm(formula = puntaje ~ horas\_estudio, data = estudio16)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -42.506 -4.395 1.623 8.204 18.491   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.8305 4.9007 8.127 1.43e-09 \*\*\*  
## horas\_estudio 5.7094 0.9288 6.147 4.94e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.24 on 35 degrees of freedom  
## Multiple R-squared: 0.5192, Adjusted R-squared: 0.5054   
## F-statistic: 37.79 on 1 and 35 DF, p-value: 4.943e-07

Si vemos los resultados del modelo lineal podemos verificar lo mismo. Tenemos bajos p-valores en la prueba de wald de los coeficientes y un p-valor bajo el test F. Vemos tambien que tanto R como R-ajustado nos indican que esta relación explica el 50% de los datos y esto coincide con el análisis previo.

**Normalidad de los residuos**

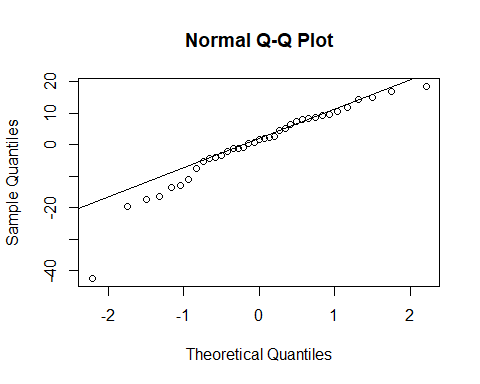
Estudiemos si hacer una regresión lineal es la mejor opción:

shapiro.test(model16$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: model16$residuals  
## W = 0.92105, p-value = 0.01196

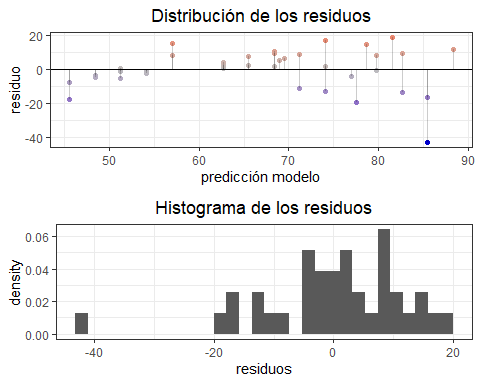
La hipotesis nula de este test es que los residuos están normalmente distribuidos y la alternativa es que no lo están, como el test rechaza la hipótesis nula, concluimos que no están normalmente distribuidos. Veamoslo en un qqplot

qqnorm(model16$residuals)   
qqline(model16$residuals)

 Vemos algunos puntos en los extremos que se alejan. Pero parte de las discrepancias podrían deberse a outliers.

model16b <- data.frame(prediccion = model16$fitted.values,  
 residuos = model16$residuals)  
  
  
d1\_16 <- ggplot(data = model16b, aes(x = prediccion, y = residuos)) +   
 geom\_point(aes(color = residuos)) +   
 scale\_color\_gradient2(low = "blue3", mid = "grey", high = "red") +   
 geom\_hline(yintercept = 0) + geom\_segment(aes(xend = prediccion, yend = 0), alpha = 0.2) +   
 labs(title = "Distribución de los residuos", x = "predicción modelo", y = "residuo") +   
 theme\_bw() +   
 theme(plot.title = element\_text(hjust = 0.5), legend.position = "none")  
  
  
d2\_16<- ggplot(data = model16b, aes(x = residuos)) + geom\_histogram(aes(y = ..density..)) +   
 labs(title = "Histograma de los residuos") + theme\_bw() +   
 theme(plot.title = element\_text(hjust = 0.5))  
  
grid.arrange(d1\_16,d2\_16,nrow = 2)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 Vemos una cierta fórma cónica en los residuos y que la distribución no parece muy normal.

Veamos ahora si los residuos estan aoutocorrelacionados, es decir si hay algún tipo de relación entre lo que paso en una medida y las siguientes. A priori conceptualmente no debería haberlas, pero por ahí detectamos autocorrelación y eso nos podría cuestionar la forma en la que se tomaron las mediciones por ejemplo.

#Durbin-Watson  
dwtest(model16,alternative ="two.sided",iterations=1000)

##   
## Durbin-Watson test  
##   
## data: model16  
## DW = 1.8264, p-value = 0.477  
## alternative hypothesis: true autocorrelation is not 0

La hipótesis nula es que NO hay autocorrelación.

Analizemos la heterocedasticidad de los residuos.

bptest(model16)

##   
## studentized Breusch-Pagan test  
##   
## data: model16  
## BP = 4.781, df = 1, p-value = 0.02877

Los residuos son heterocedasticos porque se rechaza la hipotesis nula de homocedasticidad. Es decir que la varianza no es constante para los diversos valores estimados.

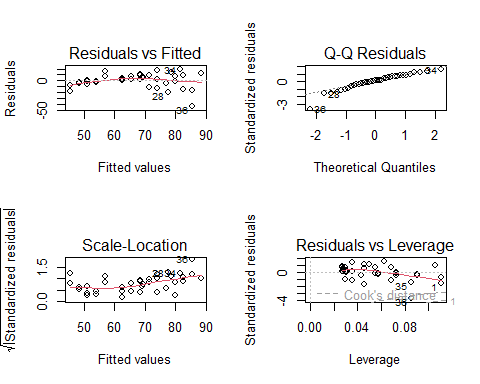
**Resumen:**

* Normalidad: NO
* Independencia: Sí (que no sean normales no afecta?)
* Homcedasticidad: NO

**Outliers**

Hagamos un análisis gráfico de otra forma menos vistosos pero más rápido:

par(mfrow=c(2,2))  
plot(model16)



par(mfrow=c(1,1))

Ya habíamos adelantado que podía haber outliers que sean puntos influyentes. De este análisis gráfico rápido, surgen algunos posibles candidatos.

summary(influence.measures(model = model16))

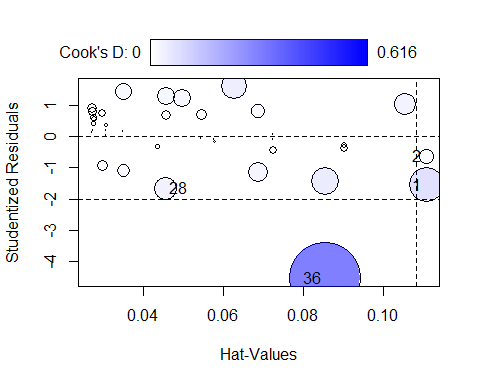
## Potentially influential observations of  
## lm(formula = puntaje ~ horas\_estudio, data = estudio16) :  
##   
## dfb.1\_ dfb.hrs\_ dffit cov.r cook.d hat   
## 36 0.73 -1.15\_\* -1.39\_\* 0.45\_\* 0.62 0.09

El punto 36 parece ser el más obvio candidato

outlierTest(model16)

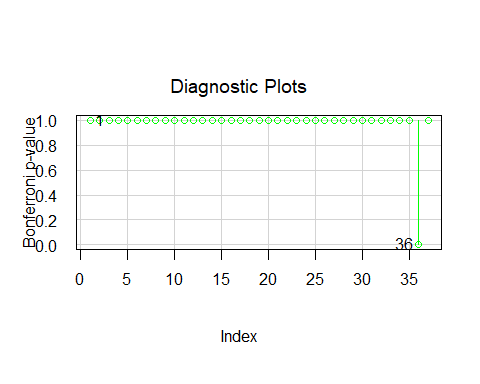
## rstudent unadjusted p-value Bonferroni p  
## 36 -4.530962 6.908e-05 0.002556

influencePlot(model = model16)



## StudRes Hat CookD  
## 1 -1.5489171 0.11057691 0.14340334  
## 2 -0.6474729 0.11057691 0.02649941  
## 28 -1.6723453 0.04544423 0.06332250  
## 36 -4.5309623 0.08554262 0.61631903

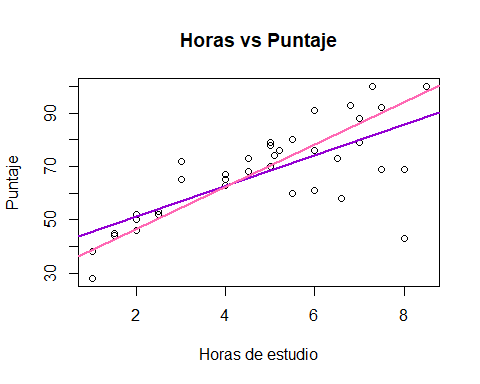
influenceIndexPlot(model16, vars='Bonf', las=1,col='green')

 Podemos concluir que el punto 36 es un outlier

### (c) Ajuste un modelo de mínimos cuadrados ponderados definiendo los pesos de tal manera que las observaciones con menor varianza tengan más peso.

Creamos un vector de ponderaciones que devuelve los valores que salen de la regresión lineal entre los residuos en valor absoluto del modelo original y sus valores estimados y al reciproco lo elevamos al cuadrado. Es decir que construimos un estimador de el residuo dado su valor estimado, lo elevamos al cuadrado para potenciar la diferencia y luego aplicamos la reciproca para que esa diferencia penalice más.

ww<-1 / lm(abs(model16$residuals) ~ model16$fitted.values)$fitted.values^2  
  
plot(estudio16$horas\_estudio,estudio16$puntaje,xlab="Horas de estudio",ylab="Puntaje",  
 main="Horas vs Puntaje")  
  
abline(model16,col="darkviolet",lwd=2)  
  
model16\_ww<- lm(puntaje ~ horas\_estudio, data = estudio16,weights =ww)  
abline(model16\_ww,col="hotpink",lwd=2)

 Comparando graficamente vemos que ambas regresiones son diferentes y la ajustada no se ve tan influenciada por el valor 36, por tal motivo tiene una pendiente mayor. Pero este ajuste también suaviza los efectos de cualquier otro residuo alejado ### (d) Realice el análisis diagnóstico del segundo modelo ajustado. Hagamos el diagnóstico anterior:

**Modelo lineal**

summary(model16\_ww)

##   
## Call:  
## lm(formula = puntaje ~ horas\_estudio, data = estudio16, weights = ww)  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -4.1319 -0.1878 0.3319 0.7847 3.0358   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.4831 2.3204 13.14 4.32e-15 \*\*\*  
## horas\_estudio 7.9253 0.7407 10.70 1.40e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.397 on 35 degrees of freedom  
## Multiple R-squared: 0.7658, Adjusted R-squared: 0.7592   
## F-statistic: 114.5 on 1 and 35 DF, p-value: 1.404e-12

Los p-valores del test de wald, los coeficientes dan ok. Pero ahora el R y R ajustado explican el 76% de los datos, lo cual mejora mucho el modelo.

**Normalidad de los residuos**

shapiro.test(model16\_ww$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: model16\_ww$residuals  
## W = 0.82071, p-value = 3.498e-05

La hipotesis nula de este test es que los residuos están normalmente distribuidos y la alternativa es que no lo están, como el test rechaza la hipótesis nula, concluimos que no están normalmente distribuidos al igual que el modelo anterior.

**Independencia de los residuos**

Como se hace esto para modelos ponderados

**Homocedasticidad**

bptest(model16\_ww)

##   
## studentized Breusch-Pagan test  
##   
## data: model16\_ww  
## BP = 0.042377, df = 1, p-value = 0.8369

Los residuos son homocedasticos porque no se rechaza en este caso la hipotesis nula de homocedasticidad.

**Resumen**

* Normalidad de residuos: NO
* Independencia: SÍ
* Homocedasticidad: SÍ

### (e) Compare ambos ajustes realizados y concluya.

Se compararon los modelos y en los puntos anteriores.