Aplicación 2.8: Ecuación de salarios en un banco americano

J. Ramajo

2020

Supongamos que se está interesado en determinar las causas de la existencia de diferencias significativas entre los salarios de los empleados de una entidad bancaria americana. Disponemos de datos sobre los salarios (*SALARIO*) de una muestra de 200 empleados del banco y de algunas variables referentes a las características de estos: el nivel de educación en años de escolarización (*EDUC*), el número de años de experiencia en el banco (*EXPER*), el sexo (*SEXO*) y la raza (*RAZA*) de cada trabajador.

La teoría del capital humano sugiere que las rentas básicas percibidas se determinan básicamente según el grado de preparación y la experiencia de los individuos. Por esta razón, se propone usar la siguiente ecuación Minceriana (en honor del economista Jacob Mincer, precursor de la función de ingresos), que relaciona el nivel de salarios con cada una de estas variables:

library(readr)  
library(car)

## Loading required package: carData

library(MASS)  
library(effects)

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

library(tidyverse)

## ── Attaching packages ──────────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.1 ✓ dplyr 1.0.0  
## ✓ tibble 3.0.1 ✓ stringr 1.4.0  
## ✓ tidyr 1.1.0 ✓ forcats 0.5.0  
## ✓ purrr 0.3.4

## ── Conflicts ─────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::recode() masks car::recode()  
## x dplyr::select() masks MASS::select()  
## x purrr::some() masks car::some()

library(RcmdrMisc)

## Loading required package: sandwich

library(sfsmisc)

##   
## Attaching package: 'sfsmisc'

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

#  
SAL\_BANCO <- read\_csv("SAL\_BANCO.csv")

## Parsed with column specification:  
## cols(  
## EDAD = col\_double(),  
## EDUC = col\_double(),  
## EXPER = col\_double(),  
## RAZA = col\_double(),  
## SALARIO = col\_double(),  
## SEXO = col\_double()  
## )

str(SAL\_BANCO)

## tibble [200 × 6] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ EDAD : num [1:200] 28.5 41.9 41.2 46.2 35.2 ...  
## $ EDUC : num [1:200] 16 19 15 12 16 19 12 12 15 16 ...  
## $ EXPER : num [1:200] 0.25 13 12 20 5.75 ...  
## $ RAZA : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...  
## $ SALARIO: num [1:200] 16079 28351 22800 12301 20500 ...  
## $ SEXO : num [1:200] 0 0 0 0 0 0 0 0 0 0 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. EDAD = col\_double(),  
## .. EDUC = col\_double(),  
## .. EXPER = col\_double(),  
## .. RAZA = col\_double(),  
## .. SALARIO = col\_double(),  
## .. SEXO = col\_double()  
## .. )

head(SAL\_BANCO, n=10)

## # A tibble: 10 x 6  
## EDAD EDUC EXPER RAZA SALARIO SEXO  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 28.5 16 0.25 0 16079. 0  
## 2 41.9 19 13 0 28351. 0  
## 3 41.2 15 12 0 22800. 0  
## 4 46.2 12 20 0 12301. 0  
## 5 35.2 16 5.75 0 20500. 0  
## 6 30.1 19 2.92 0 27250. 0  
## 7 44.5 12 18 0 9000. 0  
## 8 27.8 12 3.42 0 10980. 0  
## 9 35.4 15 11.1 0 16020. 0  
## 10 34.3 16 5.67 0 21950. 0

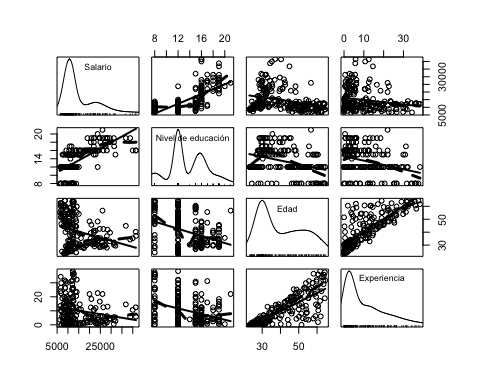
dim(SAL\_BANCO)

## [1] 200 6

summary(SAL\_BANCO)

## EDAD EDUC EXPER RAZA   
## Min. :23.00 Min. : 8.00 Min. : 0.000 Min. :0.00   
## 1st Qu.:29.75 1st Qu.:12.00 1st Qu.: 2.170 1st Qu.:0.00   
## Median :37.34 Median :12.00 Median : 6.085 Median :0.00   
## Mean :40.54 Mean :13.52 Mean :10.200 Mean :0.29   
## 3rd Qu.:51.50 3rd Qu.:16.00 3rd Qu.:15.940 3rd Qu.:1.00   
## Max. :64.50 Max. :21.00 Max. :38.330 Max. :1.00   
## SALARIO SEXO   
## Min. : 6360 Min. :0.000   
## 1st Qu.: 9645 1st Qu.:0.000   
## Median :12000 Median :0.000   
## Mean :14722 Mean :0.455   
## 3rd Qu.:18817 3rd Qu.:1.000   
## Max. :41498 Max. :1.000

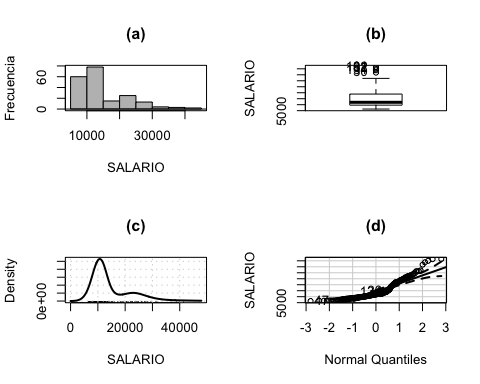
#  
# Matriz 'scatterplot' de los datos  
scatterplotMatrix(~SALARIO + EDUC + EDAD + EXPER, data=SAL\_BANCO,   
 var.labels=c("Salario", "Nivel de educación",   
 "Edad", "Experiencia"),  
 smooth=list(smoother=loessLine, var=FALSE, lwd.smooth=3),   
 col="black")



# Análisis univariante de la variable SALARIO  
#  
par(mfrow=c(2, 2))  
Hist(SAL\_BANCO$SALARIO, xlab="SALARIO", ylab="Frecuencia",   
 col="gray", main="(a)")  
Boxplot(~SALARIO, data=SAL\_BANCO, main="(b)", ylab="SALARIO")

## [1] "30" "32" "82" "101" "131"

densityPlot(SAL\_BANCO$SALARIO, from=0, normalize=TRUE,   
 xlab="SALARIO", main="(c)")  
qqPlot(~SALARIO, data=SAL\_BANCO, ylab="SALARIO",   
 xlab="Normal Quantiles", main="(d)",  
 id=list(method=c(TRUE, rep(FALSE, 132), TRUE)), col.lines="black")



## [1] 47 134 26

#  
with(SAL\_BANCO, hist(SALARIO))  
with(SAL\_BANCO, hist(log(SALARIO)))  
# Simetría de los boxplots  
symbox(~SALARIO, data=SAL\_BANCO, xlab=expression("Potencias,"~lambda), ylab="",   
 powers = c(-1, -0.5, 0, 0.33, 0.5, 1))  
mtext(2, 1, text=expression(t[BC]("SALARIO",~lambda)))  
#  
# Estimación del lambda de la transformación  
#  
S(pt <- powerTransform(SALARIO ~ 1, data=SAL\_BANCO))

## bcPower Transformation to Normality   
## Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd  
## Y1 -0.7274 -1 -1.0467 -0.408  
##   
## Likelihood ratio test that transformation parameter is equal to 0  
## (log transformation)  
## LRT df pval  
## LR test, lambda = (0) 20.54316 1 5.8302e-06  
##   
## Likelihood ratio test that no transformation is needed  
## LRT df pval  
## LR test, lambda = (1) 118.5095 1 < 2.22e-16

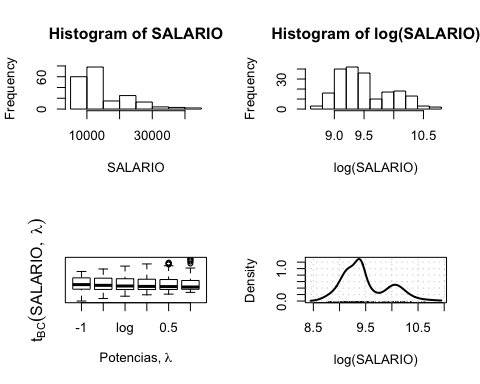
pt$lambda # estimated lambda

## Y1   
## -0.7273628

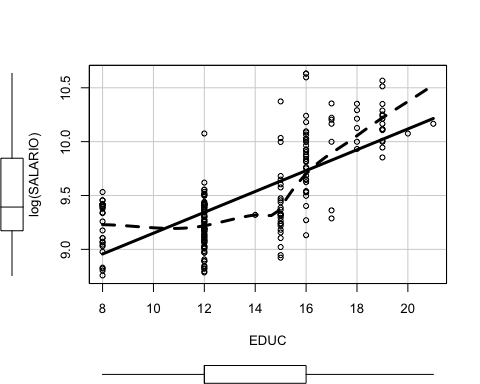
sqrt(pt$invHess) # SE

## [,1]  
## [1,] 0.1629263

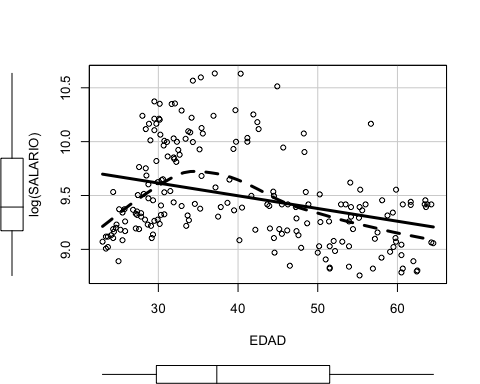
# Densidad de la variable SALARIO log-transformada  
#  
densityPlot(~log(SALARIO), data=SAL\_BANCO, adjust=0.75, xlab="log(SALARIO)")  
basicPowerAxis(0, side="above", at=c(1, 5, 10, 20, 50, 100),   
 axis.title="")



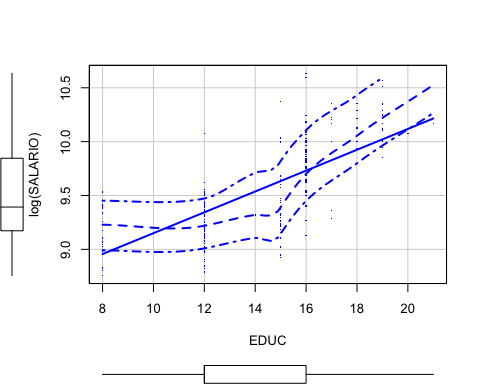
# Análisis bivariante  
#  
# Scatterplots  
#  
scatterplot(log(SALARIO) ~ EDUC, data=SAL\_BANCO, smooth=list(smoother=loessLine, var=FALSE,   
 lwd.smooth=3), col="black",  
 regLine=list(lwd=3),  
 xlab="EDUC",   
 ylab="log(SALARIO)")



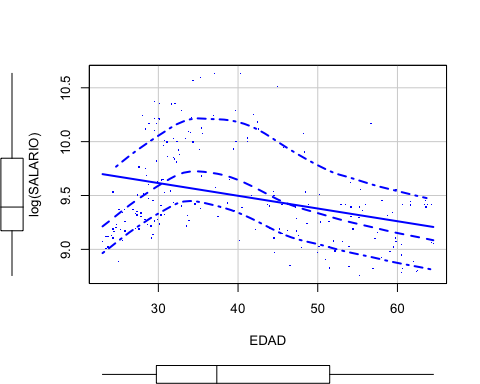
#  
scatterplot(log(SALARIO) ~ EDAD, data=SAL\_BANCO, smooth=list(smoother=loessLine, var=FALSE,   
 lwd.smooth=3), col="black",  
 regLine=list(lwd=3),  
 xlab="EDAD",   
 ylab="log(SALARIO)")



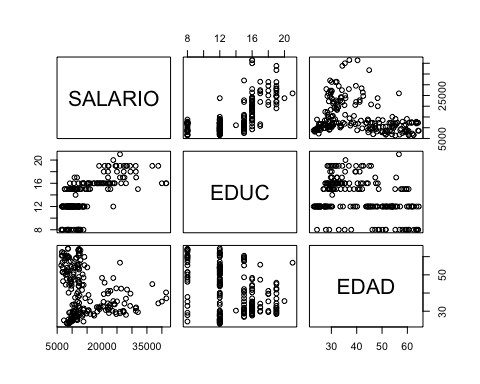
#  
scatterplot(log(SALARIO) ~ EDUC, data=SAL\_BANCO, pch=".")



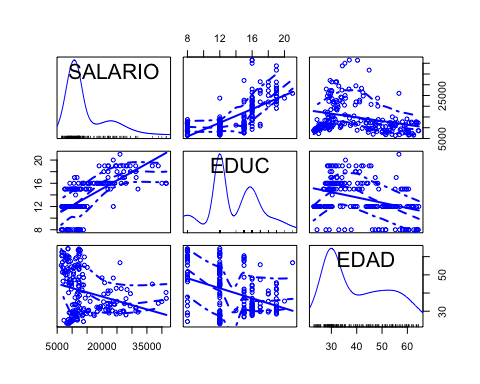
scatterplot(log(SALARIO) ~ EDAD, data=SAL\_BANCO, pch=".")



#  
pairs( ~ SALARIO + EDUC + EDAD, data=SAL\_BANCO)



scatterplotMatrix( ~ SALARIO + EDUC + EDAD, data=SAL\_BANCO)



library(PerformanceAnalytics)

## Loading required package: xts

## Loading required package: zoo

##   
## Attaching package: 'zoo'

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

##   
## Attaching package: 'xts'

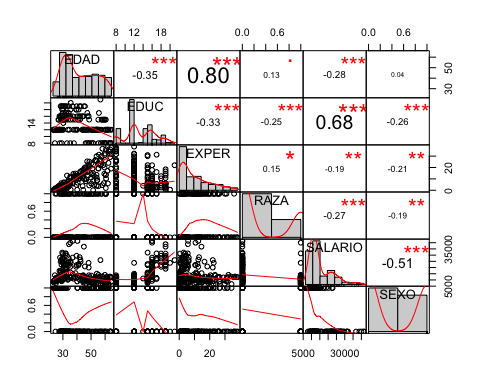
## The following object is masked from 'package:sfsmisc':  
##   
## last

## The following objects are masked from 'package:dplyr':  
##   
## first, last

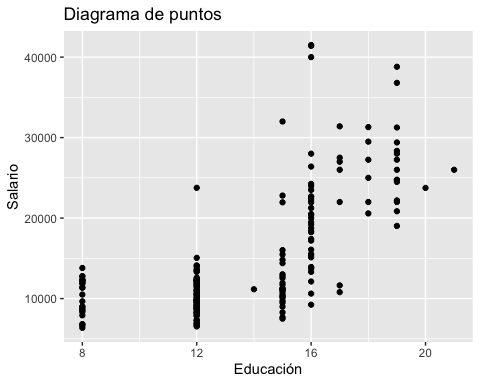
##   
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':  
##   
## legend

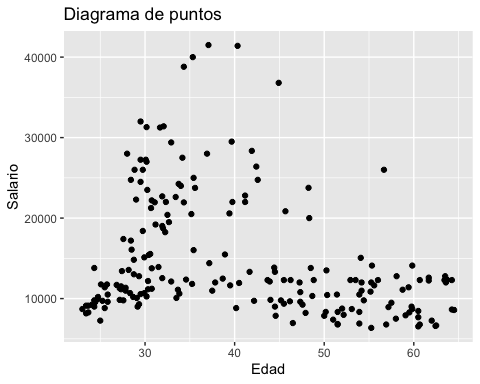
chart.Correlation(SAL\_BANCO, method="pearson", histogram=TRUE, pch=16)



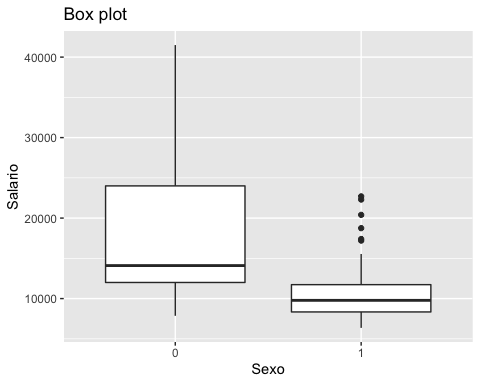
#  
# Gráficos bidimensionales  
#  
ggplot(SAL\_BANCO, aes(x=EDUC, y=SALARIO)) + geom\_point() + labs(title="Diagrama de puntos", x="Educación", y="Salario")



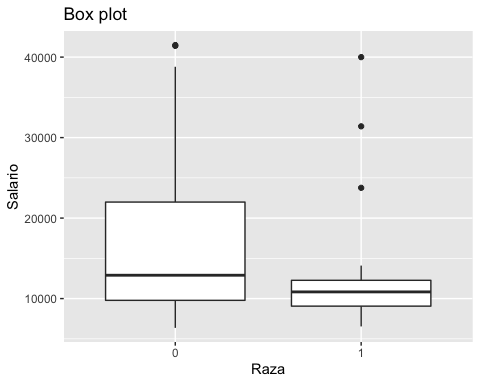
ggplot(SAL\_BANCO, aes(x=EDAD, y=SALARIO)) + geom\_point() + labs(title="Diagrama de puntos", x="Edad", y="Salario")



ggplot(SAL\_BANCO, aes(x=as.factor(SEXO), y=SALARIO)) + geom\_boxplot() + labs(title="Box plot", x="Sexo", y="Salario")

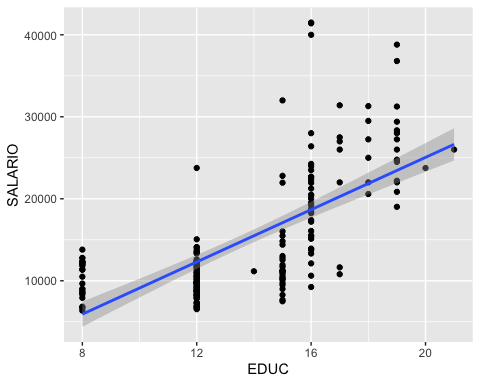


ggplot(SAL\_BANCO, aes(x=as.factor(RAZA), y=SALARIO)) + geom\_boxplot() + labs(title="Box plot", x="Raza", y="Salario")



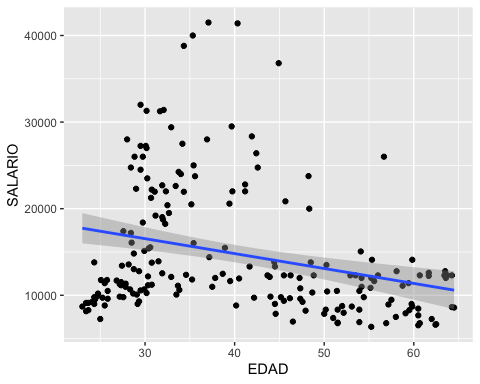
#  
ggplot(SAL\_BANCO, aes(x=EDUC, y=SALARIO)) + geom\_point() + geom\_smooth(method = 'lm')

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

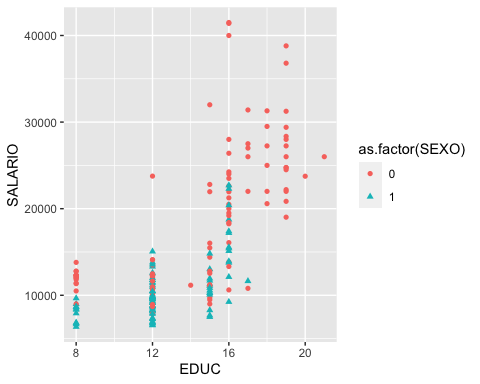


ggplot(SAL\_BANCO, aes(x=EDAD, y=SALARIO)) + geom\_point() + geom\_smooth(method = 'lm')

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

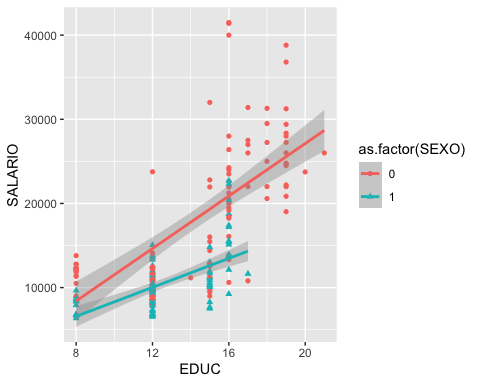


#  
# Variables de control (cambiar SEXO por RAZA para analizar los cambios)  
#  
ggplot(SAL\_BANCO, aes(x=EDUC, y=SALARIO, color=as.factor(SEXO), shape=as.factor(SEXO))) + geom\_point()

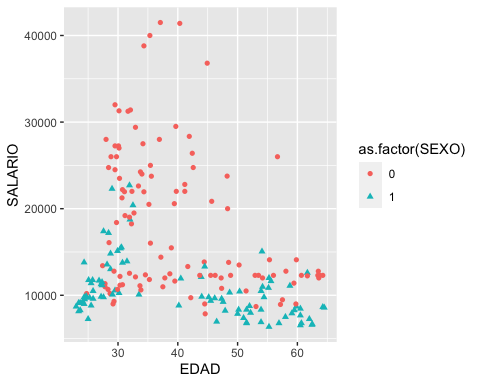


ggplot(SAL\_BANCO, aes(x=EDUC, y=SALARIO, color=as.factor(SEXO), shape=as.factor(SEXO))) + geom\_point() + geom\_smooth(method = 'lm')

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

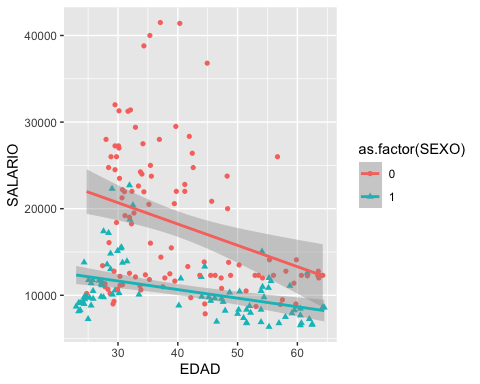


ggplot(SAL\_BANCO, aes(x=EDAD, y=SALARIO, color=as.factor(SEXO), shape=as.factor(SEXO))) + geom\_point()

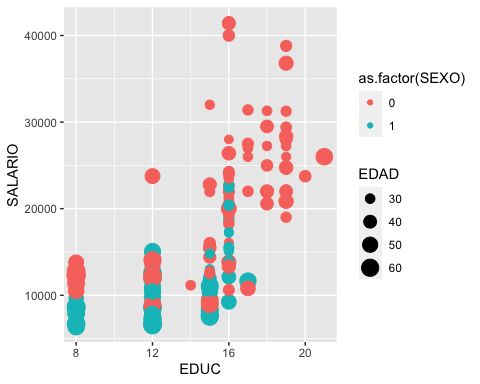


ggplot(SAL\_BANCO, aes(x=EDAD, y=SALARIO, color=as.factor(SEXO), shape=as.factor(SEXO))) + geom\_point() + geom\_smooth(method = 'lm')

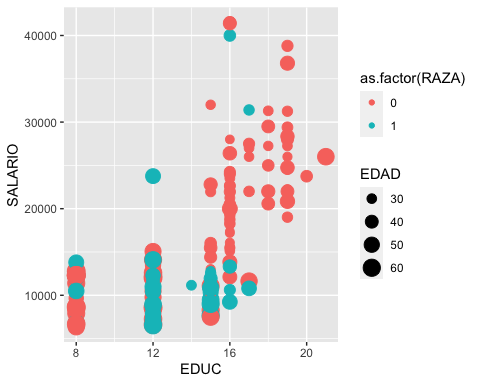
## `geom\_smooth()` using formula 'y ~ x'



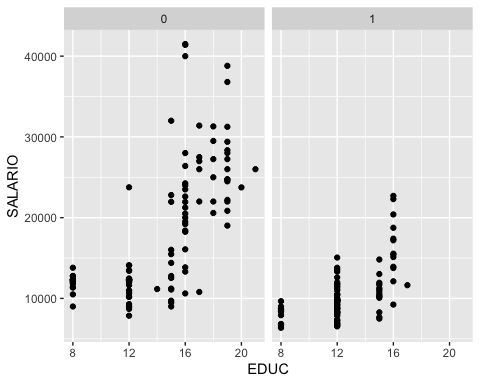
#  
ggplot(SAL\_BANCO, aes(x=EDUC, y=SALARIO, color=as.factor(SEXO), size=EDAD)) + geom\_point()



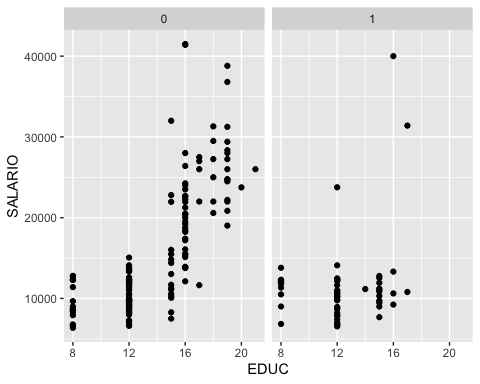
ggplot(SAL\_BANCO, aes(x=EDUC, y=SALARIO, color=as.factor(RAZA), size=EDAD)) + geom\_point()



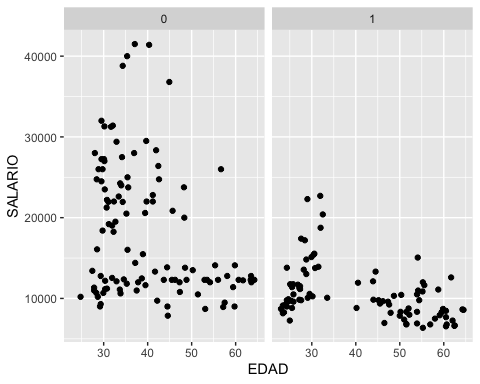
#  
ggplot(SAL\_BANCO, aes(x=EDUC, y=SALARIO)) + geom\_point() + facet\_wrap(~ as.factor(SEXO))



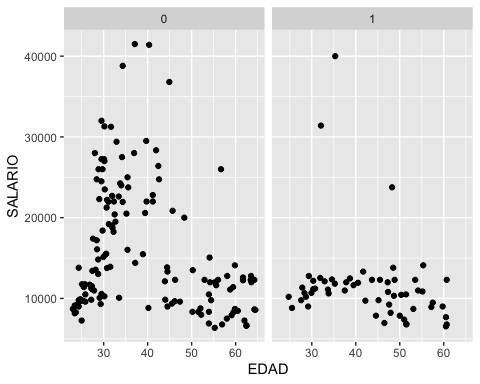
ggplot(SAL\_BANCO, aes(x=EDUC, y=SALARIO)) + geom\_point() + facet\_wrap(~ as.factor(RAZA))



#  
ggplot(SAL\_BANCO, aes(x=EDAD, y=SALARIO)) + geom\_point() + facet\_wrap(~ as.factor(SEXO))



ggplot(SAL\_BANCO, aes(x=EDAD, y=SALARIO)) + geom\_point() + facet\_wrap(~ as.factor(RAZA))

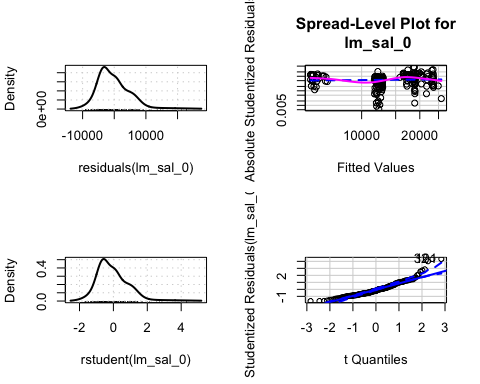


#  
# Modelo de regresión lineal  
# R usa la notación de Wilkinson-Rogers para especificar modelos: variable respuesta ~ variables explicativas  
# El símbolo (~) se lee como “es modelizada como función de”.   
# Otros símbolos utilizados son:  
# + inclusión de variable  
# - exclusión de variable (no substracción)  
# ∗ incluir variables y sus interacciones  
# : interaccionar dos variables  
# ∧ interacción de variables hasta un grado especificado (no un exponente)  
# Para obviar un símbolo de modelo, usar la función I().  
# Ejemplos:  
# y ~ x1 + x2 + x3 (regresión múltiple)  
# y ~ . (incluir como regresores todas las variables de la base de datos)  
# y ~ x1 + x2 - 1 (excluir la constante del modelo)  
# y ~ x1 + x2 + x1:x2 (incluir interacción entre x1 y x2)  
# y ~ x1 \* x2 (incluir x1, x2 y su interacción - mismo resultado que el modelo anterior)  
# y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3 (modelo con interacciones de doble y triple vía)  
# y ~ x1 \* x2 \* x3 (igual que el modelo anterior)  
# y ~ (x1 + x2 + x3)ˆ2 (interacciones de dbole-vía)  
# y ~ x1 + I(x1ˆ2) + x2 (regresión cuadrática en x1 más variables x2)  
# y ~ poly(x1, 2, raw = TRUE) + x2 (igual que el modelo anterior)  
# Modelo lineal  
S(lm\_sal\_0 <- lm(SALARIO ~ EDUC + EDAD, data = SAL\_BANCO))

## Call: lm(formula = SALARIO ~ EDUC + EDAD, data = SAL\_BANCO)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5039.79 2626.87 -1.919 0.0565 .   
## EDUC 1551.98 131.09 11.839 <2e-16 \*\*\*  
## EDAD -30.12 33.83 -0.890 0.3744   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard deviation: 5524 on 197 degrees of freedom  
## Multiple R-squared: 0.463  
## F-statistic: 84.92 on 2 and 197 DF, p-value: < 2.2e-16   
## AIC BIC   
## 4019.31 4032.51

# Diagnósticos  
# 'residuals' -errores estimados y `rstudent()` - residuos estudentizados  
# `densityPlot()` chequeo de la distribución de los errores (densidades estimadas)  
densityPlot(residuals(lm\_sal\_0))  
densityPlot(rstudent(lm\_sal\_0))  
spreadLevelPlot(lm\_sal\_0)  
## Suggested power transformation: 0.8973957

# `qqPlot()` chequeo de errores no-normales (comparación de los residuos estudentizados con una distribution t)  
qqPlot(lm\_sal\_0)



## [1] 32 101

# Chequeo de 'outliers' en la regresión  
#  
max(hatvalues(lm\_sal\_0))

## [1] 0.05872325

which.max(hatvalues(lm\_sal\_0))

## 21   
## 21

#  
outlierTest(lm\_sal\_0)

## rstudent unadjusted p-value Bonferroni p  
## 101 4.331607 2.3616e-05 0.0047231  
## 32 4.331018 2.3673e-05 0.0047347  
## 82 4.011607 8.5743e-05 0.0171490

#  
max(cooks.distance(lm\_sal\_0))

## [1] 0.05421915

which.max(cooks.distance(lm\_sal\_0))

## 30   
## 30

#  
max(abs(dffits(lm\_sal\_0)))

## [1] 0.4088903

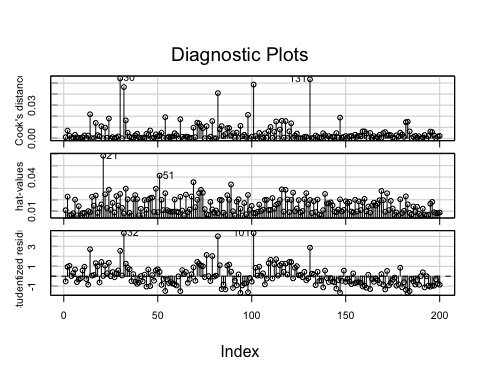
which.max(abs(dffits(lm\_sal\_0)))

## 30   
## 30

# Medidas de influencia  
S(influence.measures(lm\_sal\_0))

## Potentially influential observations of  
## lm(formula = SALARIO ~ EDUC + EDAD, data = SAL\_BANCO) :  
##   
## dfb.1\_ dfb.EDUC dfb.EDAD dffit cov.r cook.d hat   
## 14 0.09 0.03 -0.15 0.26 0.92\_\* 0.02 0.01   
## 21 -0.01 0.01 0.00 0.01 1.08\_\* 0.00 0.06\_\*  
## 30 -0.31 0.36 0.19 0.41\_\* 0.94\_\* 0.05 0.03   
## 32 -0.10 0.22 0.00 0.39\_\* 0.78\_\* 0.05 0.01   
## 51 0.09 -0.07 -0.06 0.09 1.06\_\* 0.00 0.04   
## 82 -0.06 0.19 -0.04 0.36 0.81\_\* 0.04 0.01   
## 89 0.08 -0.07 -0.05 0.09 1.05\_\* 0.00 0.03   
## 101 -0.17 0.25 0.09 0.40\_\* 0.78\_\* 0.05 0.01   
## 131 -0.21 0.34 0.02 0.41\_\* 0.92\_\* 0.05 0.02

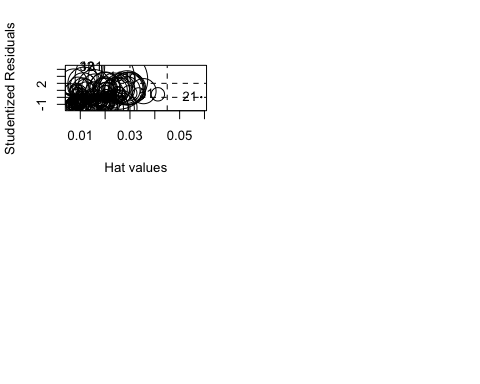
influenceIndexPlot(lm\_sal\_0, vars=c("Cook", "hat", "Studentized"))

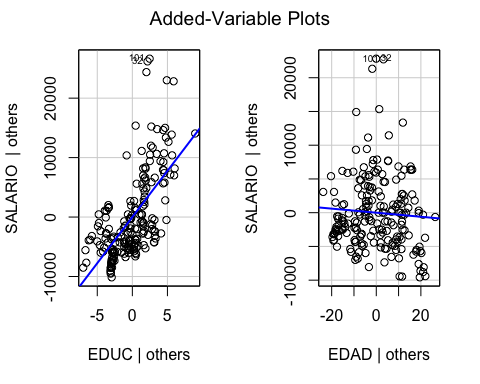


influencePlot(lm\_sal\_0, xlab="Hat values")

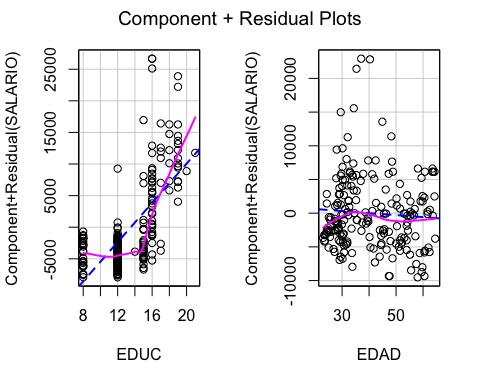
## StudRes Hat CookD  
## 21 0.0291055 0.058723246 1.770638e-05  
## 30 2.5477581 0.025110343 5.421915e-02  
## 32 4.3310179 0.008026966 4.641173e-02  
## 51 0.4348382 0.041240086 2.722295e-03  
## 101 4.3316069 0.008412244 4.867037e-02  
## 131 2.8644299 0.019845816 5.342317e-02

# Gráficos de variable añadida, buscando casos influyentes  
avPlots(lm\_sal\_0, id=list(cex=0.60, method="mahal"))





# Chequeo de no linealidad: gráficos de componente+residuo  
crPlots(lm\_sal\_0, smooth=list(span=0.7))



# Chequeo de varianza no constante:  
ncvTest(lm\_sal\_0)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 14.92596, Df = 1, p = 0.00011181

ncvTest(lm\_sal\_0, var.formula= ~ EDUC + EDAD)

## Non-constant Variance Score Test   
## Variance formula: ~ EDUC + EDAD   
## Chisquare = 18.27925, Df = 2, p = 0.00010733

#  
# Modelo log-lineal  
#  
S(lm\_sal <- lm(log(SALARIO) ~ EDUC + EDAD, data = SAL\_BANCO))

## Call: lm(formula = log(SALARIO) ~ EDUC + EDAD, data = SAL\_BANCO)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.381805 0.150665 55.632 <2e-16 \*\*\*  
## EDUC 0.092165 0.007519 12.258 <2e-16 \*\*\*  
## EDAD -0.003374 0.001941 -1.739 0.0837 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard deviation: 0.3168 on 197 degrees of freedom  
## Multiple R-squared: 0.4945  
## F-statistic: 96.35 on 2 and 197 DF, p-value: < 2.2e-16   
## AIC BIC   
## 112.81 126.01

brief(lm\_sal)

## (Intercept) EDUC EDAD  
## Estimate 8.382 0.09216 -0.00337  
## Std. Error 0.151 0.00752 0.00194  
##   
## Residual SD = 0.317 on 197 df, R-squared = 0.494

coef(lm\_sal)

## (Intercept) EDUC EDAD   
## 8.38180464 0.09216472 -0.00337382

confint(lm\_sal)

## 2.5 % 97.5 %  
## (Intercept) 8.084681719 8.6789275586  
## EDUC 0.077337314 0.1069921306  
## EDAD -0.007200724 0.0004530834

#  
# Introducción de factores de control  
# Boxplots  
#  
Boxplot(log(SALARIO) ~ SEXO, data=SAL\_BANCO, id=list(location="lr"),   
 ylab="SALARIO", xlab="SEXO")

## [1] "50" "63" "154"

#  
Boxplot(log(SALARIO) ~ SEXO, data=SAL\_BANCO, id=list(location="lr"),   
 ylab="SALARIO", xlab="RAZA")

## [1] "50" "63" "154"

#  
lm\_sal\_factors <- lm(log(SALARIO) ~ SEXO + RAZA + EDUC + EDAD, data = SAL\_BANCO)  
S(lm\_sal\_factors)

## Call: lm(formula = log(SALARIO) ~ SEXO + RAZA + EDUC + EDAD, data = SAL\_BANCO)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.027626 0.129264 69.839 < 2e-16 \*\*\*  
## SEXO -0.421681 0.037356 -11.288 < 2e-16 \*\*\*  
## RAZA -0.239739 0.040782 -5.879 1.77e-08 \*\*\*  
## EDUC 0.065614 0.006270 10.465 < 2e-16 \*\*\*  
## EDAD -0.004002 0.001500 -2.669 0.00825 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard deviation: 0.2443 on 195 degrees of freedom  
## Multiple R-squared: 0.7025  
## F-statistic: 115.1 on 4 and 195 DF, p-value: < 2.2e-16   
## AIC BIC   
## 10.78 30.57

compareCoefs(lm\_sal, lm\_sal\_factors)

## Calls:  
## 1: lm(formula = log(SALARIO) ~ EDUC + EDAD, data = SAL\_BANCO)  
## 2: lm(formula = log(SALARIO) ~ SEXO + RAZA + EDUC + EDAD, data = SAL\_BANCO)  
##   
## Model 1 Model 2  
## (Intercept) 8.382 9.028  
## SE 0.151 0.129  
##   
## EDUC 0.09216 0.06561  
## SE 0.00752 0.00627  
##   
## EDAD -0.00337 -0.00400  
## SE 0.00194 0.00150  
##   
## SEXO -0.4217  
## SE 0.0374  
##   
## RAZA -0.2397  
## SE 0.0408  
##

#  
anova(lm\_sal\_factors)

## Analysis of Variance Table  
##   
## Response: log(SALARIO)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## SEXO 1 12.1696 12.1696 203.8913 < 2.2e-16 \*\*\*  
## RAZA 1 6.2225 6.2225 104.2525 < 2.2e-16 \*\*\*  
## EDUC 1 8.6657 8.6657 145.1865 < 2.2e-16 \*\*\*  
## EDAD 1 0.4251 0.4251 7.1227 0.008252 \*\*   
## Residuals 195 11.6389 0.0597   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

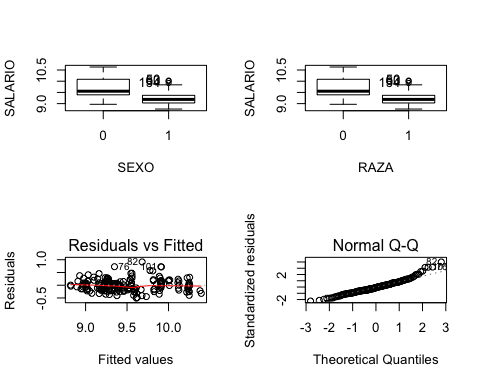
#  
# Modelo cuadrático con factores de control  
#  
S(lm\_sal\_poly <- lm(log(SALARIO) ~ SEXO + RAZA + poly(EDUC,2, raw=TRUE) + poly(EDAD,2, raw=TRUE), data = SAL\_BANCO))

## Call: lm(formula = log(SALARIO) ~ SEXO + RAZA + poly(EDUC, 2, raw = TRUE) +  
## poly(EDAD, 2, raw = TRUE), data = SAL\_BANCO)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.5376718 0.3576590 26.667 < 2e-16 \*\*\*  
## SEXO -0.3496025 0.0382804 -9.133 < 2e-16 \*\*\*  
## RAZA -0.2328608 0.0414828 -5.613 6.83e-08 \*\*\*  
## poly(EDUC, 2, raw = TRUE)1 -0.1112075 0.0406899 -2.733 0.00686 \*\*   
## poly(EDUC, 2, raw = TRUE)2 0.0064279 0.0015116 4.252 3.29e-05 \*\*\*  
## poly(EDAD, 2, raw = TRUE)1 0.0283828 0.0131379 2.160 0.03198 \*   
## poly(EDAD, 2, raw = TRUE)2 -0.0003939 0.0001537 -2.563 0.01115 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard deviation: 0.2313 on 193 degrees of freedom  
## Multiple R-squared: 0.7362  
## F-statistic: 89.76 on 6 and 193 DF, p-value: < 2.2e-16   
## AIC BIC   
## -9.24 17.14

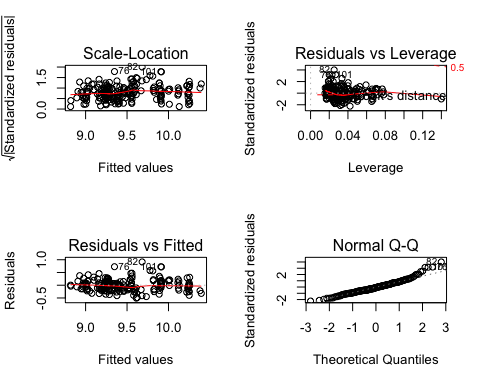
# Comparación con el modelo simple  
anova(lm\_sal, lm\_sal\_poly)

## Analysis of Variance Table  
##   
## Model 1: log(SALARIO) ~ EDUC + EDAD  
## Model 2: log(SALARIO) ~ SEXO + RAZA + poly(EDUC, 2, raw = TRUE) + poly(EDAD,   
## 2, raw = TRUE)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 197 19.777   
## 2 193 10.322 4 9.4554 44.202 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Chequeo por defecto  
plot(lm\_sal\_poly)



# Chequeo de las hipótesis de linealidad (y varianza constante)  
plot(lm\_sal\_poly, which = 1)  
# Chequeo de la hipótesis de normalidad  
plot(lm\_sal\_poly, which = 2)



# Chequeo de las hipótesis de varianza constante  
plot(lm\_sal\_poly, which = 3)  
# Chequeo de observaciones atípicas e influyentes  
plot(lm\_sal\_poly, which = 5)  
plot(lm\_sal\_poly, which = 4)  
#  
# Análisis de efectos  
summary(lm\_sal\_poly)

##   
## Call:  
## lm(formula = log(SALARIO) ~ SEXO + RAZA + poly(EDUC, 2, raw = TRUE) +   
## poly(EDAD, 2, raw = TRUE), data = SAL\_BANCO)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.50692 -0.14152 -0.02066 0.12946 0.91440   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.5376718 0.3576590 26.667 < 2e-16 \*\*\*  
## SEXO -0.3496025 0.0382804 -9.133 < 2e-16 \*\*\*  
## RAZA -0.2328608 0.0414828 -5.613 6.83e-08 \*\*\*  
## poly(EDUC, 2, raw = TRUE)1 -0.1112075 0.0406899 -2.733 0.00686 \*\*   
## poly(EDUC, 2, raw = TRUE)2 0.0064279 0.0015116 4.252 3.29e-05 \*\*\*  
## poly(EDAD, 2, raw = TRUE)1 0.0283828 0.0131379 2.160 0.03198 \*   
## poly(EDAD, 2, raw = TRUE)2 -0.0003939 0.0001537 -2.563 0.01115 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2313 on 193 degrees of freedom  
## Multiple R-squared: 0.7362, Adjusted R-squared: 0.728   
## F-statistic: 89.76 on 6 and 193 DF, p-value: < 2.2e-16

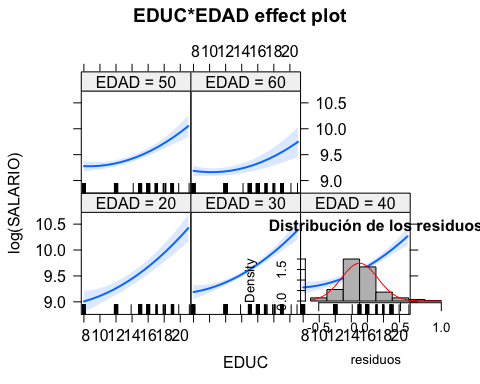
plot(Effect("EDUC", lm\_sal\_poly))  
plot(Effect("EDAD", lm\_sal\_poly))  
#  
# Modelo con efectos de interacción  
#  
lm\_sal\_poly\_int <- lm(log(SALARIO) ~ SEXO + RAZA + poly(EDUC,2, raw=TRUE) + poly(EDAD,2, raw=TRUE) + EDUC:EDAD, data = SAL\_BANCO)  
S(lm\_sal\_poly\_int)

## Call: lm(formula = log(SALARIO) ~ SEXO + RAZA + poly(EDUC, 2, raw = TRUE) +  
## poly(EDAD, 2, raw = TRUE) + EDUC:EDAD, data = SAL\_BANCO)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.3047154 0.5204150 15.958 < 2e-16 \*\*\*  
## SEXO -0.3326206 0.0377729 -8.806 7.61e-16 \*\*\*  
## RAZA -0.2029661 0.0415909 -4.880 2.22e-06 \*\*\*  
## poly(EDUC, 2, raw = TRUE)1 -0.0053869 0.0517288 -0.104 0.917169   
## poly(EDUC, 2, raw = TRUE)2 0.0050769 0.0015360 3.305 0.001132 \*\*   
## poly(EDAD, 2, raw = TRUE)1 0.0537647 0.0150921 3.562 0.000463 \*\*\*  
## poly(EDAD, 2, raw = TRUE)2 -0.0004511 0.0001512 -2.984 0.003219 \*\*   
## EDUC:EDAD -0.0016467 0.0005151 -3.197 0.001625 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard deviation: 0.2259 on 192 degrees of freedom  
## Multiple R-squared: 0.7495  
## F-statistic: 82.07 on 7 and 192 DF, p-value: < 2.2e-16   
## AIC BIC   
## -17.62 12.07

plot(Effect("EDUC", lm\_sal\_poly\_int))  
plot(Effect("EDAD", lm\_sal\_poly\_int))  
plot(Effect(c("EDUC","EDAD"), lm\_sal\_poly\_int))  
#  
# Diagnóstico del modelo  
#  
# Normalidad de los residuos  
# Distribución de los residuos  
library(tseries)

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

r <- resid(lm\_sal\_poly\_int)  
rbar <- mean(r)  
sdr <- sd(r)  
hist(r, col="grey", freq=FALSE, main="Distribución de los residuos",  
 ylab="Density", xlab="residuos")  
curve(dnorm(x, rbar, sdr), col=2, add=TRUE,  
 ylab="Density", xlab="r")



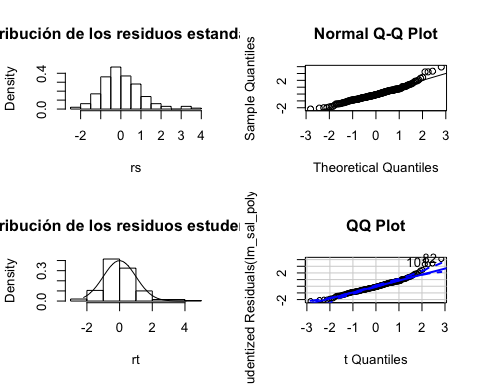
#   
# Residuos estandarizados  
rs<-(r-rbar)/sdr  
hist(rs, freq=FALSE,  
 main="Distribución de los residuos estandarizados")  
#  
qqnorm(rs)   
abline(0,1)   
#  
jarque.bera.test(r) #(package 'tseries')

##   
## Jarque Bera Test  
##   
## data: r  
## X-squared = 46.864, df = 2, p-value = 6.662e-11

shapiro.test(r)

##   
## Shapiro-Wilk normality test  
##   
## data: r  
## W = 0.96165, p-value = 2.993e-05

#  
# Distribución de los residuos estudentizados  
library(MASS)  
rt <- studres(lm\_sal\_poly\_int)  
hist(rt, freq=FALSE,  
 main="Distribución de los residuos estudentizados")  
xfit<-seq(min(rt),max(rt),length=40)  
yfit<-dnorm(xfit)  
lines(xfit, yfit)   
#  
# QQ plot de los residuos estudentizados  
qqPlot(lm\_sal\_poly\_int, main="QQ Plot")



## [1] 82 101

#  
# Multicolinealidad  
# Matriz de correlaciones  
cor(SAL\_BANCO)

## EDAD EDUC EXPER RAZA SALARIO SEXO  
## EDAD 1.00000000 -0.3549893 0.8013086 0.1270033 -0.2844342 0.04033283  
## EDUC -0.35498933 1.0000000 -0.3251096 -0.2460025 0.6788401 -0.26246839  
## EXPER 0.80130862 -0.3251096 1.0000000 0.1459087 -0.1915098 -0.21361832  
## RAZA 0.12700328 -0.2460025 0.1459087 1.0000000 -0.2720387 -0.18565209  
## SALARIO -0.28443420 0.6788401 -0.1915098 -0.2720387 1.0000000 -0.50955103  
## SEXO 0.04033283 -0.2624684 -0.2136183 -0.1856521 -0.5095510 1.00000000

# Factores de inflación de la varianza  
vif(lm\_sal\_poly\_int)

## GVIF Df GVIF^(1/(2\*Df))  
## SEXO 1.386378 1 1.177445  
## RAZA 1.395619 1 1.181363  
## poly(EDUC, 2, raw = TRUE) 24.863252 2 2.233004  
## poly(EDAD, 2, raw = TRUE) 34.127079 2 2.416990  
## EDUC:EDAD 28.606751 1 5.348528

sqrt(vif(lm\_sal\_poly\_int)) > 2 # problema de multicolinealidad

## GVIF Df GVIF^(1/(2\*Df))  
## SEXO FALSE FALSE FALSE  
## RAZA FALSE FALSE FALSE  
## poly(EDUC, 2, raw = TRUE) TRUE FALSE FALSE  
## poly(EDAD, 2, raw = TRUE) TRUE FALSE FALSE  
## EDUC:EDAD TRUE FALSE TRUE