Estadística 2 - trabajo final

Andrea Gomez Vargas

2024-10-21

library(rstatix)  
library(haven)  
library(olsrr)  
library(ggpubr)  
library(apa)  
library(apaTables)  
library(plyr)  
library(haven)  
library(nortest)  
library(goftest)  
library(olsrr)

## Ejercicio ANOVA

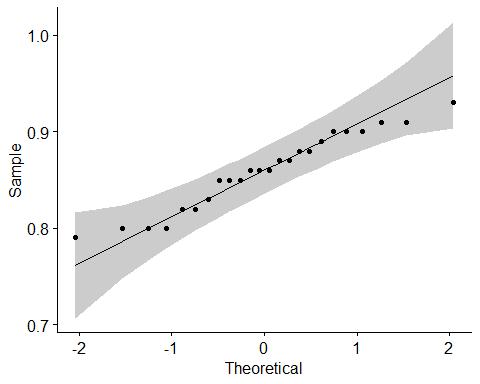
datos <- data.frame(densidad = c(0.9,0.91,0.87,0.9,0.9,0.8,  
 0.86,0.88,0.89,0.82,0.91,0.83,  
 0.93,0.86,0.88,0.85,0.87,0.8,  
 0.79,0.86,0.82,0.85,0.8,0.85),  
 harina = factor(rep\_len(c("General", "Tortas"), length.out = 24)),  
 azucar = c(rep("0",3),rep("50",3),rep("75",3),rep("100",3),rep("0",3),rep("50",3),rep("75",3),rep("100",3))  
 )

* Chequeo de normalidad y homocedasticidad
* Chequeo que se cumplan los supuestos de normalidad y homocedasticidad de variancias para poder hacer el analisis de variancias
* testeo de normalidad de las variables

shapiro.test(datos$densidad)

##   
## Shapiro-Wilk normality test  
##   
## data: datos$densidad  
## W = 0.95557, p-value = 0.356

ggpubr::ggqqplot(datos, "densidad")



como el pv del test de normalidad de Shapiro-Wilk es *0.356 > 0.05* no rechazo la hipotesis nula de normalidad de la variable densidad de las tortas

* testeo de homocedasticidad de varianzas

levene\_test(datos, densidad ~ harina \* azucar)

## # A tibble: 1 × 4  
## df1 df2 statistic p  
## <int> <int> <dbl> <dbl>  
## 1 7 16 2.76 0.0440

* Como el pv del test de homocedasticidad de variancias de Levene es 0.88 > 0.05 no rechazo la hipotesis nula de homocedasticidad
* Anova de dos factores con interaccion

anova1 <- aov(densidad ~ harina \* azucar, data=datos)  
summary(anova1)

## Df Sum Sq Mean Sq F value Pr(>F)  
## harina 1 0.001838 0.001837 1.256 0.279  
## azucar 3 0.007383 0.002461 1.683 0.211  
## harina:azucar 3 0.004075 0.001358 0.929 0.450  
## Residuals 16 0.023400 0.001463

1. Existen diferencias estadísticamente significativas debidas al tipo de harina utilizada?

En base a la evidencia observada se puede concluir que no existen diferencias significativas para la densidad media de las tortas segun el tipo de harina utilizado con un nivel de significacion del 5% cuando el porcentaje de concentracion de zucar y la interaccion entre la concentraccion de azucar y el tipo de harina se consideran en el modelo.

1. Existen diferencias estadísticamente significativas debidas al nivel de endulzamiento utilizado?

En base a la evidencia observada se puede concluir que no existen diferencias significativas para la densidad media de las tortas segun la concentracion de azucar utilizada con un nivel de significacion del 5% cuando el tipo de harina y la interaccion entre la concentracion de azucar y el tipo de harina se consideran en el modelo.

1. Existe interacción estadísticamente significativa entre los factores considerados?

Como el pv del test para el efecto de la interaccion es 0.04 < 0.05 con un nivel de significacion del 5% se puede concluir que existe interaccion significativa entre el tipo de harina y la concentracion de azucar. Es decir, que la diferencia en la densidad media de las tortas cuando se usa harina común o harina para tortas es distinta segun el nivel de concentracion de azucar.

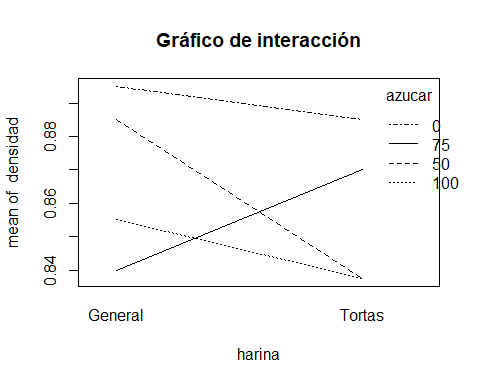
anova\_test(datos, densidad ~ harina \* azucar)

## ANOVA Table (type II tests)  
##   
## Effect DFn DFd F p p<.05 ges  
## 1 harina 1 16 0.462 0.507 0.028  
## 2 azucar 3 16 1.683 0.211 0.240  
## 3 harina:azucar 3 16 0.929 0.450 0.148

La interaccion entre el tipo de harina el porcentaje de concentracion de azucar explica el 38,8% de la variancia de la densisdad de las tortas.

1. Realizar los gráficos de medias e interpretar.

with(datos, interaction.plot(harina, azucar, densidad,   
 fun = mean, main = "Gráfico de interacción"))



aggregate(densidad ~ harina + azucar, datos, mean)

## harina azucar densidad  
## 1 General 0 0.8950  
## 2 Tortas 0 0.8850  
## 3 General 100 0.8550  
## 4 Tortas 100 0.8375  
## 5 General 50 0.8850  
## 6 Tortas 50 0.8375  
## 7 General 75 0.8400  
## 8 Tortas 75 0.8700

Como se observa en el grafico para un porcentaje de concentracion de azucar fijo la densidad de las tortas es mayor cuando se usa harina general en comparacion con cuando se usa harina para tortas, excepto cuando la concentracion de azucar es de 100 donde la densidad de las tortas es menor cuando se usa harina general

## Ejercicio Regresión

datos <- read\_sav("EVALUACION/p081.sav")

1. Cuál es el nivel de asociación lineal de las variables predictoras con la variable Sales? Comentar.

m0 <- lm(SALES ~ AGE + HS + INCOME + BLACK + FEMALE + PRICE, data = datos)  
  
# Correlacion de cada predictor con la variable dependiente  
ols\_correlations(m0)

## Correlations   
## -------------------------------------------  
## Variable Zero Order Partial Part   
## -------------------------------------------  
## AGE 0.227 0.207 0.174   
## HS 0.067 -0.011 -0.009   
## INCOME 0.326 0.269 0.230   
## BLACK 0.190 0.110 0.091   
## FEMALE 0.146 -0.029 -0.024   
## PRICE -0.301 -0.430 -0.392   
## -------------------------------------------

Todas las variables presentan correlaciones linealres positivas excepto por el precio promedio del paquete de cigarrillos (PRICE) que presenta una correlacion lineal negativa. Las variables que presentan las mayores correlaciones lineales con la variable SALES (cantidad de cigarrillos vendidos per capita) son AGE (edad mediana) y PRICE que superan el 0.3 de valor de correlacion en valor absoluto. El resto de las variables presentan valores de correlacion lineal leves, menores al 0.3

1. Realizar una regresión lineal múltiple, seleccionando los mejores predictores entre las variables independientes disponibles, utilizando un método de selección automática. Describir el proceso de selección automática utilizado. (Sug. Considerar como probabilidad de entrada 0.10 y de salida del modelo 0.15.)

ols\_step\_both\_p(m0, p\_enter = 0.10, p\_remove = 0.15)

##   
##   
## Stepwise Summary   
## -------------------------------------------------------------------------  
## Step Variable AIC SBC SBIC R2 Adj. R2   
## -------------------------------------------------------------------------  
## 0 Base Model 501.451 505.315 356.203 0.00000 0.00000   
## 1 INCOME (+) 497.718 503.514 352.481 0.10632 0.08808   
## 2 PRICE (+) 490.760 498.487 346.323 0.25028 0.21904   
## 3 AGE (+) 489.024 498.683 345.297 0.30324 0.25877   
## -------------------------------------------------------------------------  
##   
## Final Model Output   
## ------------------  
##   
## Model Summary   
## ----------------------------------------------------------------  
## R 0.551 RMSE 26.506   
## R-Squared 0.303 MSE 762.362   
## Adj. R-Squared 0.259 Coef. Var 22.717   
## Pred R-Squared 0.189 AIC 489.024   
## MAE 16.607 SBC 498.683   
## ----------------------------------------------------------------  
## RMSE: Root Mean Square Error   
## MSE: Mean Square Error   
## MAE: Mean Absolute Error   
## AIC: Akaike Information Criteria   
## SBC: Schwarz Bayesian Criteria   
##   
## ANOVA   
## --------------------------------------------------------------------  
## Sum of   
## Squares DF Mean Square F Sig.   
## --------------------------------------------------------------------  
## Regression 15594.426 3 5198.142 6.818 7e-04   
## Residual 35831.020 47 762.362   
## Total 51425.445 50   
## --------------------------------------------------------------------  
##   
## Parameter Estimates   
## ------------------------------------------------------------------------------------------  
## model Beta Std. Error Std. Beta t Sig lower upper   
## ------------------------------------------------------------------------------------------  
## (Intercept) 64.248 61.933 1.037 0.305 -60.345 188.841   
## INCOME 0.019 0.007 0.358 2.801 0.007 0.005 0.033   
## PRICE -3.399 0.989 -0.438 -3.436 0.001 -5.389 -1.409   
## AGE 4.156 2.199 0.243 1.890 0.065 -0.267 8.579   
## ------------------------------------------------------------------------------------------

Se utiliza un metodo de seleccion secuencial para la definicion del modelo final. En particular se utilizo el Metodo Stepwise. Este metodo comienza realizando todos los modelos de regresion simple entre la variable dependiente (SALES) y cada una de las variables explicativas y se elige como modelo inicial aquel con mayor valor del coeficiente de determinacion R^2, es decir, aquel da el mayor incremento de la suma de cuadrados de la regresion (SSR). A partir del modelo inicial se incorporan variables secuencialmente tomando en cada paso aquella que al ser sumada al modelo genera el mayor incremento de SSR y a su vez tiene una significacion individual mayor a la cota elgida, en este caso que el test de que el parametro asociado a esa variable sea igual a 0 tenga pv menor de 0,1. A su vez en cada paso se elimina la variable con menor nivel de significacion individual siempre que esta sea menor que una cota determinada, en este caso que el test de que el parametro asociado a esa variable sea igual a 0 tenga un pvalue mayor a 0,15.

Este proceso de seleccion termina en un modelo para la variable SALES (cantidad de cigarrillos vendidos) con los regresores INCOME (ingreso personal per capita), PRICE (precio promedio del paquete de cigarrilos) y AGE (edad mediana)

m1 <- lm(SALES ~ INCOME + PRICE + AGE, datos)  
summary(m1)

##   
## Call:  
## lm(formula = SALES ~ INCOME + PRICE + AGE, data = datos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -50.430 -13.853 -4.962 6.691 128.947   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.248260 61.933008 1.037 0.30487   
## INCOME 0.019281 0.006883 2.801 0.00737 \*\*  
## PRICE -3.399234 0.989172 -3.436 0.00124 \*\*  
## AGE 4.155908 2.198699 1.890 0.06491 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 27.61 on 47 degrees of freedom  
## Multiple R-squared: 0.3032, Adjusted R-squared: 0.2588   
## F-statistic: 6.818 on 3 and 47 DF, p-value: 0.0006565

1. Qué información da el coeficiente de determinación?

El coeficiente de determinacion (R^2) es una medida que ayuda a evaluar la bondad del ajuste del modelo de regresion. Es la proporcion de variancia de la variable dependiente, SALES en nuestro caso, que es explicada por el modelo. Es una medida que varia entre 0 y 1, es mejor mientras mas se acerque a 1.

El coeficiente de correlacion aumenta mientras mas variables se incluyan en el modelo de regresion. Es por eso que para el caso de regresion multiple es preferible usar el coeficiente de correlacion ajustado que penaliza al coeficiente de correlacion del modelo por cada variable que se suma al modelo

1. Cuáles son los supuestos necesarios para definir la prueba inferencial de los estimadores de los parámeros?

Para que la inferencia sobre los parametros del modelo de regresion sea valida se debe cumplir que los errores del modelo sean aleatorios y tengan una distribucion normal con media 0 y variancia constante

1. Analizar la bondad del ajuste del modelo obtenido, comentando los indicadores y/o test que considera.

summary(m1)

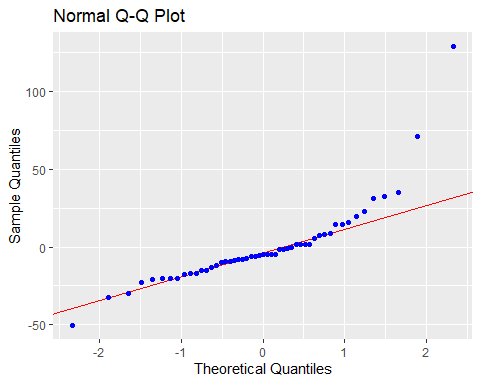
##   
## Call:  
## lm(formula = SALES ~ INCOME + PRICE + AGE, data = datos)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -50.430 -13.853 -4.962 6.691 128.947   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.248260 61.933008 1.037 0.30487   
## INCOME 0.019281 0.006883 2.801 0.00737 \*\*  
## PRICE -3.399234 0.989172 -3.436 0.00124 \*\*  
## AGE 4.155908 2.198699 1.890 0.06491 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 27.61 on 47 degrees of freedom  
## Multiple R-squared: 0.3032, Adjusted R-squared: 0.2588   
## F-statistic: 6.818 on 3 and 47 DF, p-value: 0.0006565

Para evaluar la bondad del ajuste lo primero que se evalua es que los predictores sean significativos para explicar la variable dependiente, es decir que se pueda determinar que el valor del parametro asociado a cada predictor sea distinto 0 en presencia del resto de las variables. Para esto se evalua el pv asociado al test de significacion para cada variable del modelo. En el caso del modelo evaluado las variables INCOME y PRICE son significativas comparandolas con un nivel de significacion del 5%, y la variable AGE es significativa comparandola con un nivel de significacion del 10%.

Otra medida para evaluar la bondad del ajuste es el R^2 ajustado que en el caso del modelo es de 0,259 el cual se puede considerar un valor bajo por lo que el modelo no presentaria un buen ajuste.

1. Realizar un análisis de los residuos del modelo para evaluar el cumplimiento de los supuestos. Para esto, realizar gráficos de los residuos con el valor predicho.

#gráfico de los residuos para evaluar normalidad   
ols\_plot\_resid\_qq(m1)

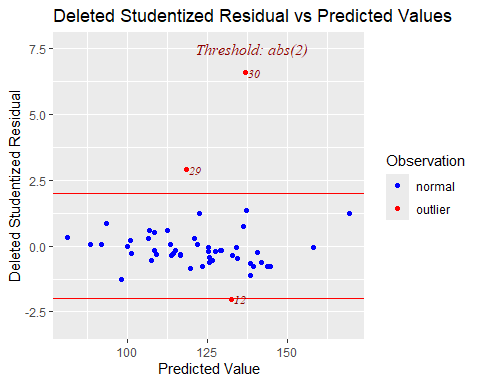


#Test de normalidad de los residuos   
shapiro.test(residuals(m1))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(m1)  
## W = 0.77937, p-value = 2.472e-07

Por el QQ-norm, en el cual se observa que la cola de la distribucion de los residuos se aleja de la distribucion normal, y por el pv del test de normalidad de Shapiro-Wilk que es menor que 0.05 se rechaza la hipotesis de normalidad de los residuos.

#Grafico de los residuos vs los valores predichos  
ols\_plot\_resid\_stud\_fit(m1, print\_plot = TRUE)



En el grafico de los residuos vs los valores predichos no parece haber evidencia de que la variancia de los residuos no sea constante ni que exista una relacion no lineal. Si se observa evidencia de observaciones que pueden ser consideradas outliers y que pueden estar afectando al ajuste de regresion.

1. Analizar la colinealidad de las variables predictoras presentes en la ecuación.

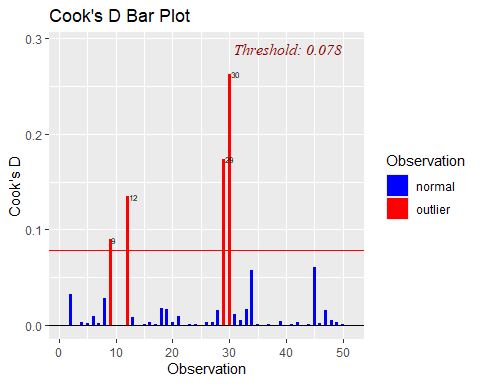
ols\_vif\_tol(m1)

## Variables Tolerance VIF  
## 1 INCOME 0.9098781 1.099048  
## 2 PRICE 0.9142128 1.093837  
## 3 AGE 0.8952388 1.117020

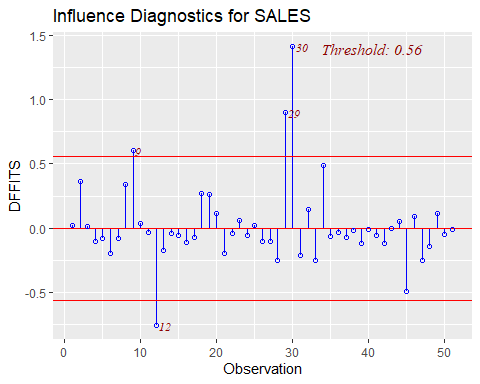
Para determinar la existencia de multicolinealidad se evalua la correlacion lineal entre los predictores del modelo. Como para ninguna de las variables independientes del modelo el variance inflation factor (VIF) es mayor que 10 se concluye que no existe multicolinealidad entre los regresores INCOME, PRICE y AGE.

1. Analizar la presencia de observaciones atípicas y/o influyentes. Comentar y resolver según el caso.

# Distancia de Cook  
ols\_plot\_cooksd\_bar(m1)



# dffits  
ols\_plot\_dffits(m1)



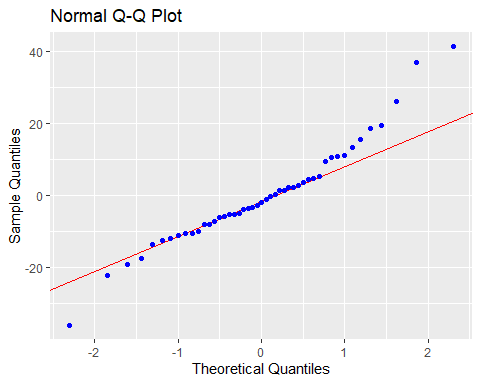
Segun el criterio de la distancia de Cook las observaciones 9, 12, 29 y 30 son outliers, ya sea por parte de la variable dependiente o de las regresoras. Segun el criterio DFFIT se comprueba que las cuatro observaciones identificadas son influyentes ya que el cambio en el valor ajustado supera el threshold cuando se omite cada una de esos casos.

Como la regresion puede verse fuertemente afectada por la presencia de outliers en este caso se toma la determinacion de eliminar estas observaciones y volver a ajustar el modelo. Otra opcion podria ser realizar un ajuste de regresion robusto que no se viera tan influenciado por la presencia de estos outliers

datos2 <- datos[-c(9,12,29,30), ]  
  
m2 <- lm(SALES ~ AGE + INCOME + PRICE, data = datos2)  
summary(m2)

##   
## Call:  
## lm(formula = SALES ~ AGE + INCOME + PRICE, data = datos2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -35.966 -8.048 -1.907 4.986 41.255   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.495426 33.385323 2.741 0.008896 \*\*   
## AGE 2.724622 1.205695 2.260 0.028956 \*   
## INCOME 0.015148 0.004149 3.651 0.000703 \*\*\*  
## PRICE -2.782824 0.561699 -4.954 1.18e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.61 on 43 degrees of freedom  
## Multiple R-squared: 0.4567, Adjusted R-squared: 0.4188   
## F-statistic: 12.05 on 3 and 43 DF, p-value: 7.411e-06

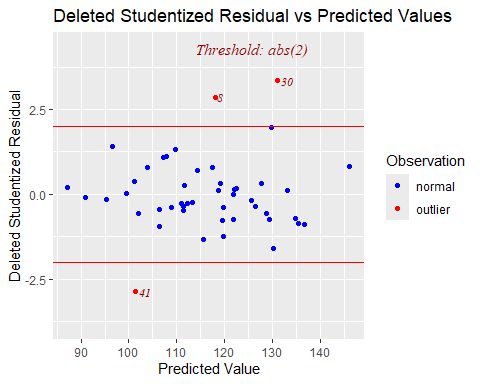
#gráfico de los residuos para evaluar normalidad   
ols\_plot\_resid\_qq(m2)



#Test de normalidad de los residuos   
shapiro.test(residuals(m2))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(m2)  
## W = 0.95481, p-value = 0.06699

#Grafico de los residuos vs los valores predichos  
ols\_plot\_resid\_stud\_fit(m2, print\_plot = TRUE)



La eliminacion de las observaciones detectadas como outliers modifica significativamente el ajuste de regresion. La significacion de los parametros que acompañan a los regresores aumento siendo el pv asociado a cada parametro menor a 0,05 en todos los casos. El R^2 ajustado aumento a 0,412. Por ultimo el test de normalidad de los residuos que previamente se rechazaba ahora arroja un pv de 0,067 > 0,05 por lo que ya no se rechazaria el supuesto de normalidad.

## Ejercicio Regresión logística

1. Calcular el riesgo relativo y los odds ratio de la variable dependiente con cada una las variables dicotómicas. Analizar los resultados.
2. Cuál es la definición de odds ratio? Qué información suministra y de qué manera puede calcularse utilizando la regresión logística?
3. Calcule los odds ratio de cada una de las variables predictoras disponibles con la variable dependiente? Comentar.
4. Realizar una regresión logística múltiple, seleccionando los mejores predictores entre las variables independientes disponibles, utilizando un método de selección automática.
5. Según el modelo obtenido, cuáles son los principales factores de riesgo del bajo peso y cuál es la magnitud de su efecto?
6. Cuáles son los supuestos necesarios para definir la prueba inferencial de los estimadores de los parámetros?
7. Indicar porcentaje de casos bien predichos por el modelo.