XS-2130 Modelos de Regresión Aplicados II Sem 2022 Grupo 2

Examen Parcial 2.

Se entrega el lunes 25 de octubre aula, al correo electrónico [Gilbert.brenes.camacho@gmail.com](mailto:Gilbert.brenes.camacho@gmail.com). (En su defecto, se enviará por otro medio previa autorización del profesor). Envían un solo documento de Word (o pdf de Acrobat) en el que aparezca cada pregunta que se efectúa, las respuestas respectivas pegando los comandos o resultados de R y una copia del script de R pegado al final del documento de Word. Si hacen un markdown, el documento final tiene que estar compilado en Word o pdf, y tiene que contener el texto de las preguntas (con su respectivo puntaje). Si el texto de las preguntas no aparece, el profesor puede bajar puntos por no seguir instrucciones.El nombre del archivo debe tener su nombre. Por ejemplo, gilbertbrenesexamparcial2.docx.

Total de puntos: 50 ptos.

**Preguntas teóricas**

Pregunta 1. Explique por qué la violación del supuesto de independencia de los errores genera problemas en la interpretación de los intervalos de confianza de una predicción. (3 ptos.)

Cuando no hay independencia existe correlación entre las variables, y al generar intervalos de confianza para las predicciones, estos no se van a ajustar adecuadamente,

Pregunta 2. Explique cómo el R2 ajustado penaliza a los modelos poco parsimoniosos. (3 ptos.)

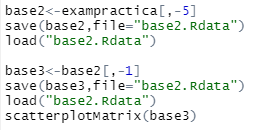
Porque incluye variables que no aportan tanto el modelo entonces esto hace creer el CME y esto hace disminuir el R cuadrado, es decir, hay modelos que no requieren de tantas variables y logra predecir mejor que otros modelos con muchas variables.

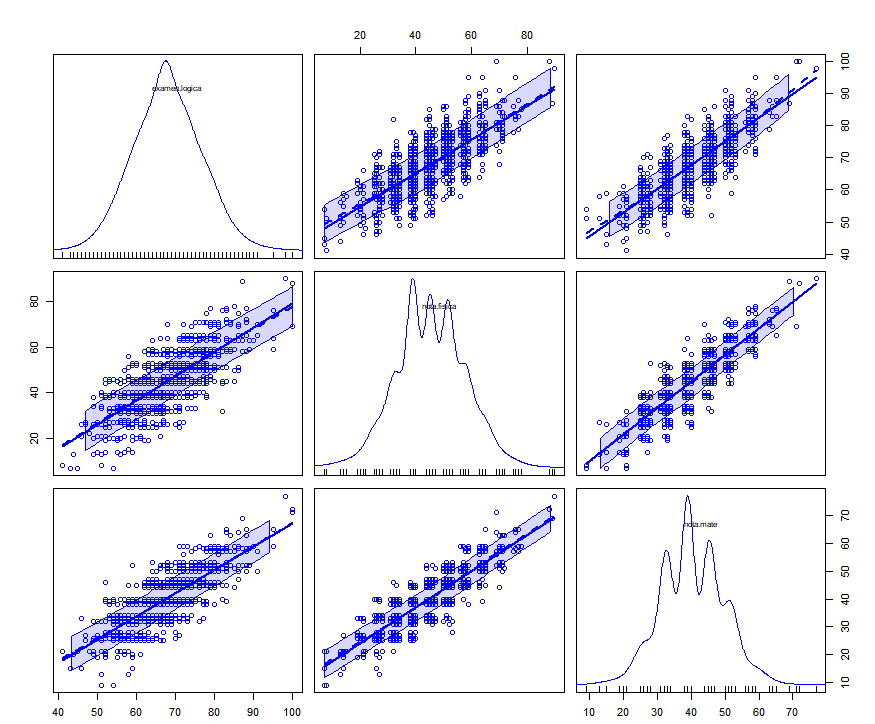
**Preguntas prácticas y semi/teóricas**

Pregunta 1.

Se tomó una muestra de 1000 estudiantes de un circuito escolar al que se les hizo un examen especial de lógica para prepararlos para el examen de admisión. De esos 1000 estudiantes, 50 viven en un distrito rural y 950 en el distrito cabecera. Se quiere analizar con un modelo de regresion cuál es la relación entre la nota del examen de lógica y las notas obtenidas en matemática y físca en el trimestre previo, pues se sabe que los cursos de física y matemática promueven el pensamiento lógico-matemático. El archivo exampractica.Rdata contiene la variable id (identificador), la variable examen.logica (medida en puntos), nota.fisica (medida en puntos con base 100), nota.mate (medida en puntos con base 100) y una variable dicotómica distrito que es igual a 1 si los estudiantes son del distrito rural, y 0 si son del distrito cabecera. Con dichos datos conteste las siguientes preguntas.

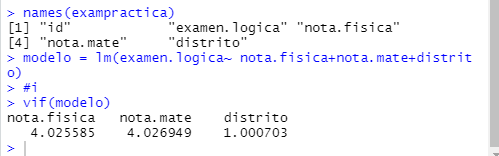
1. Con la función scatterplot.matrix, haga una matriz de gráficos entre la variable respuesta y las dos predictoras cuantitativas. Explique qué observa en términos de la relación entre las variables (3 ptos).





Tanto entre las variables independientes como en la dependiente con las independientes se puede observar que hay cierta correlación entre las variables.

1. Estime un modelo de regresión gaussiana en el que la nota del examen de lógica esté en función de la nota de física, la nota de matemática y el distrito y con este modelo:
   1. Calcule los VIFS y argumente si hay evidencia de multicolinealidad. (3 ptos.)

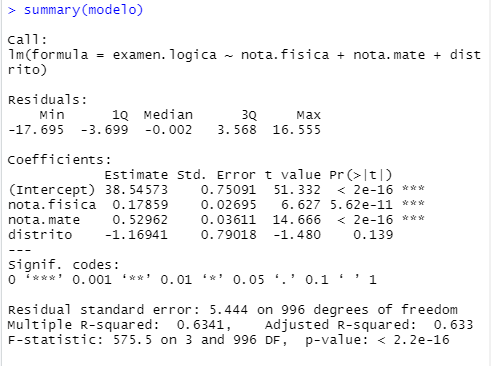


No sabemos si realmente hay multicolinealidad, porque las variables nota.fisica y nota.mate se encuentran en zona de indecisión

* 1. Presentando el summary del modelo, planteé las hipótesis nulas asociadas con las pendientes y contrástelas con un 5% de significancia. (3 ptos.)

H0: las variables que predicen el modelo son significativas

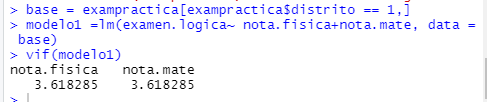
H1: las variables que predicen el modelo no son significativas



1. Con base en la pregunta anterior, argumente si la multicolinealidad está afectando las conclusiones del modelo (3 ptos.)

La multicolinealidad está afectando a la variable distrito porque por el p value estamos observando que esta variable no es significativa

1. Estime un modelo de regresión gaussiana en el la nota del examen de lógica esté en función de la nota de física y la nota de matemática, pero solo para los estudiantes del distrito rural y con dicho modelo:
   1. Calcule los VIFS y argumente si hay evidencia de multicolinealidad. (3 ptos.)

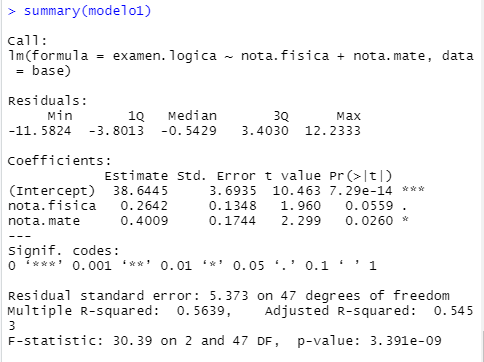


No hay evidencia de multicolinealidad entre las variables del modelo

* 1. Presentando el summary del modelo, planteé las hipótesis nulas asociadas con las pendientes y contrástelas con un 5% de significancia. (3 ptos.)

H0: las variables que predicen el modelo son significativas

H1: las variables que predicen el modelo no son significativas

* 1. 

1. Argumente si hay diferencias en el efecto de la multicolinealidad sobre las conclusiones del modelo para toda la población y el modelo solo para el distrito rural, y si hay o no hay diferencias, por qué ocurre (3 ptos.)

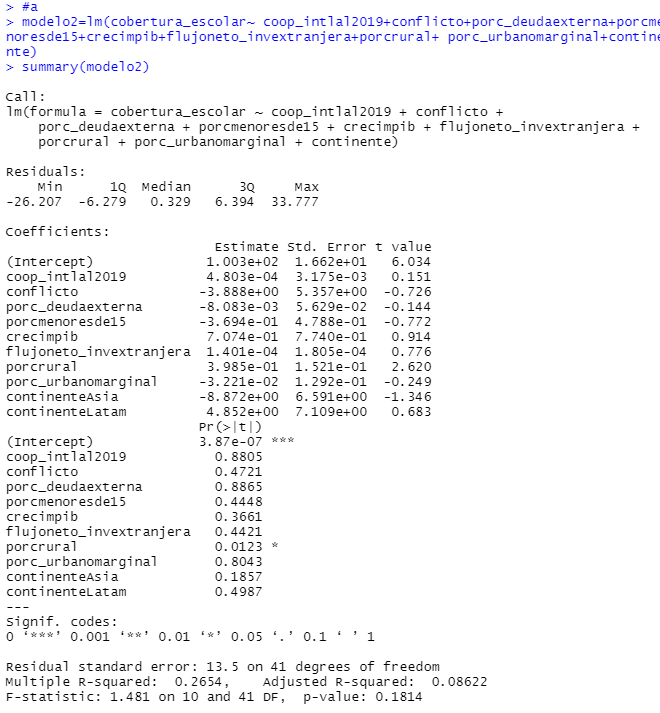
Si hay diferencias porque a nivel de toda la población vemos que por el método del p value en la nota de física se presenta multicolinealidad mientras que a nivel de toda la población esto no ocurría y por medio de los vif vemos que a nivel del distrito no presenta multicolinealidad sin embargo a nivel poblacional teníamos una zona de indecisión

Pregunta 2. Con una muestra de países en desarrollo se busca estimar una ecuación de regresión para analizar cuáles variables están asociadas con la cobertura escolar. Para ello, se construye el archivo de datos conocido como baspais.Rdata. Este tiene las siguientes variables:

|  |  |
| --- | --- |
| Pais | Nombre del país |
| paiscodigo | Abreviatura del nombredel país |
| coop\_intlal2019 | Monto de cooperación internacional (en millones de dólares) |
| conflicto | Variable binaria: 1=Conflicto bélico, 0=Sin conflicto |
| cobertura\_escolar | Porcentaje de niños matriculados en educación primaria (en puntos porcentuales) |
| porc\_deudaexterna | Deuda externa como porcentaje del PIB (en puntos porcentuales) |
| porcmenoresde15 | Porcentaje de la población con edad menor de 15 años (en puntos porcentuales) |
| crecimpib | Crecimiento relativo del PIB (en puntos porcentuales) |
| flujoneto\_invextranjera | Flujo neto de inversión extranjera directa (en millones de dólares) |
| porcrural | Porcentaje de la población que vive en zona rual (en puntos porcentuales) |
| porc\_urbanomarginal | Porcentaje de la población viviendo en barrios urbano marginales (en puntos porcentuales) |
| continente | Continente en el que está localizado cada país |
|  |  |
|  |  |

Con dichos datos, realice los siguientes análisis.

1. Estime un modelo de regresión gaussiana en el que prediga la cobertura escolar con base en el resto de variables predictoras (excepto obviamente pais y paiscodigo), presente el summary, e interprete el coeficiente de conflicto (independientemente del pvalue). (3 ptos.)

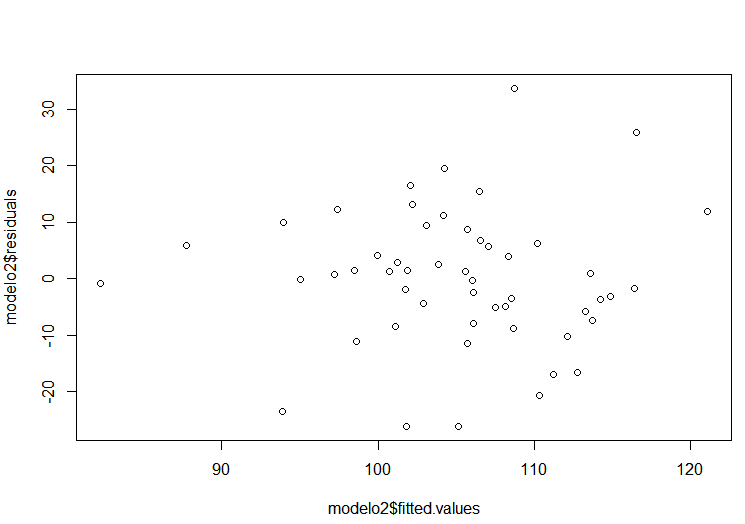


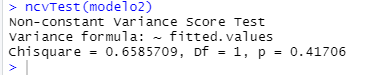
B2= por cada aumento de un punto porcentual de niños matriculados en educación primaria se espera que en promedio la matricula cuando hay conflictos bélicos 3,888 menos que cuando no se presentan conflictos

1. Con base en el modelo anterior, analice el supuesto de homoscedasticidad con un gráfico de predichos contra residuos, y con la variación de la prueba de Breusch-Pagan basada en los valores predichos (al 5% de significancia). ¿Qué concluye? (5 ptos.)

H0: Sigma\_i = Sigma

H1: Sigma\_i <> Sigma

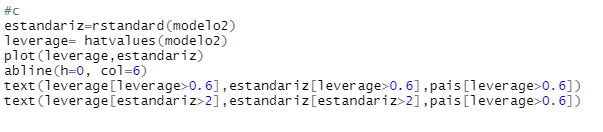


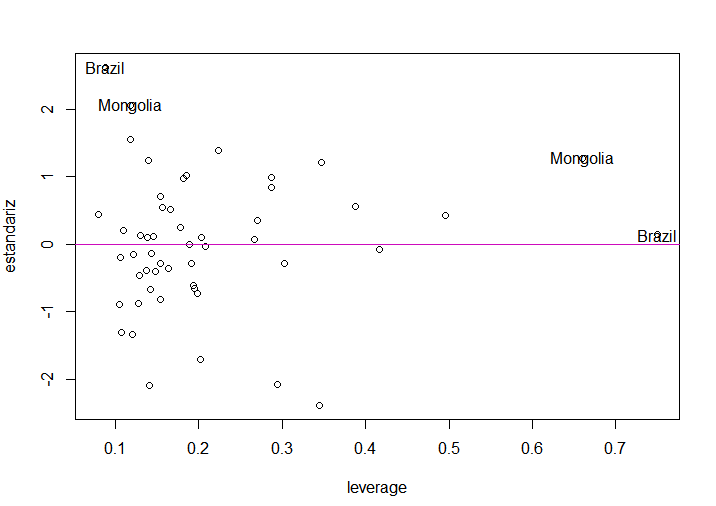


Al 5% de significancia hay suficiente evidencia estadística para no rechazar la hipótesis nula de que la homocedasticidad es diferente

1. Analice si hay valores extremos en las variables predictoras, y SOLO SI LOS HAY, escoja los dos casos que son más extremos según el estadístico apropiado, identífiquelos y explique por qué son extremos (4 ptos.)

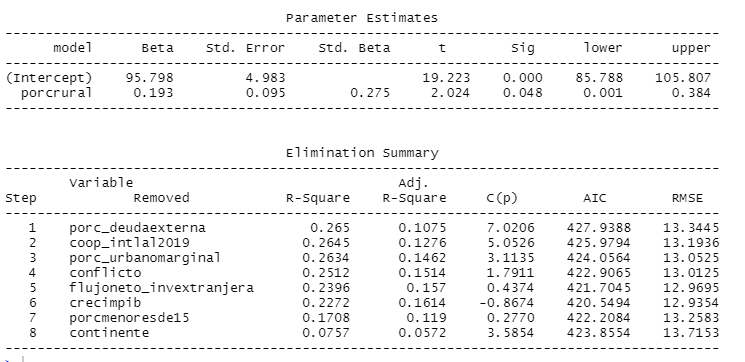
Los valores extremos son Mongolia y Brazil



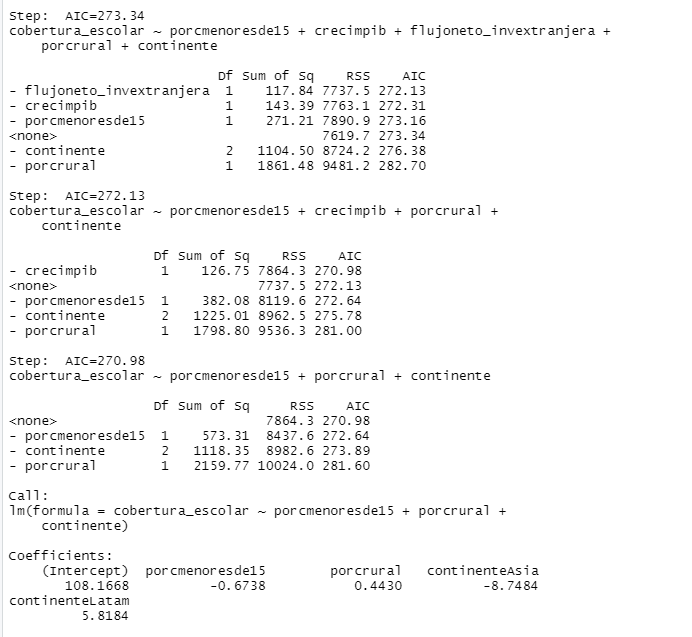


1. A partir del modelo estimado al principio y con las funciones predeterminadas de R, use el algoritmo basado en los pvalues y el algoritmo basado en el AIC para seleccionar predictores con un procedimiento “hacia atrás” (“backwards”) y obtener un modelo más parsimonioso. Diga si se obtiene la misma ecuación (4 ptos.)





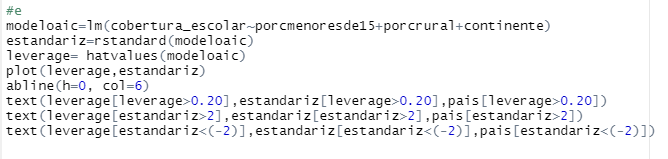


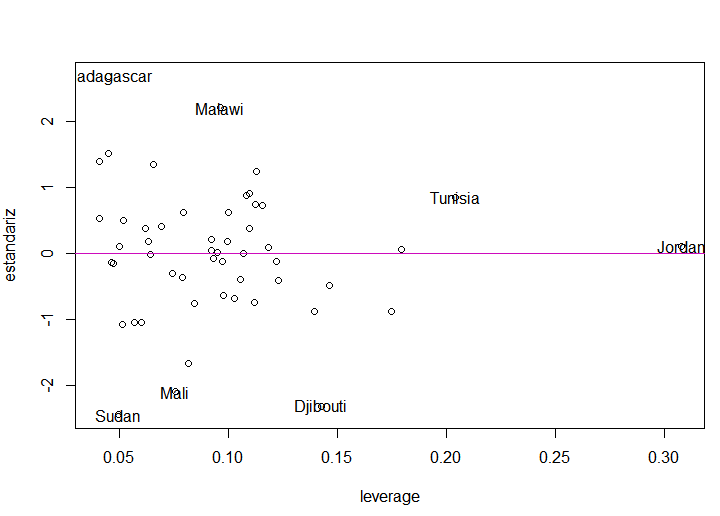


No se obtienen la misma ecuación, con el método del pvale la variable predictora que deja es porcrural, y con el método del AIC las variables predictoras que incluyen son cobertura\_escolar, porcmenoresde15, porcrural y continente

1. Con el modelo reducido con base en el AIC, analice si hay valores extremos en las variables predictoras, y SOLO SI LOS HAY, escoja los dos casos que son más extremos según el estadístico apropiado. Diga si se llega a la misma conclusión que en el inciso ( c ), y por qué sí o por qué no se llega a la misma conclusión (4 ptos.)

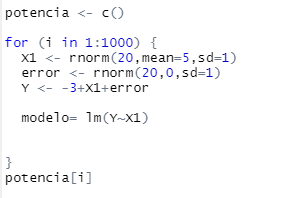
No se llega a las mismas conclusiones es decir aún hay valores extremos, sin embargos estos no son los mismos que se encontraron en el primer modelo.

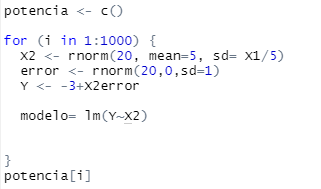




Pregunta 3.

Haga el siguiente procedimiento:

1. Escenario homoscedástico
   1. Genere una variable independiente normal “equis” con tamaño de muestra 20, media 5 y desviación estándar igual a 1. Además, genere un vector “error” de errores normales con media 0 y desviación estándar igual a1.
   2. Cree una variable aleatoria “ye” surgida del siguiente modelo de regresión lineal: .
   3. Estime un modelo de regresión lineal simple con la variable “ye” en función de X.
   4. Repite este procedimiento 1000 veces, y calcule en qué proporción de las 1000 veces se rechaza la hipótesis nula de que beta1=0 (potencia de la prueba).
   5. 
2. Escenario heteroscedástico.
   1. Repita todo el procedimiento del punto 1
   2. Tiene que crear la variable “equis”, pero ahora la variable independiente es “equis2”, que es una variable aleatoria normal con tamaño de muestra 20, media igual a 5, y desviación estándar igual a “equis”/5.



Ya no sé cómo seguir

1. Compare las potencias del escenario homoscedástico con el heteroscedástico y diga cuál es el posible mecanismo de las diferencias (3 ptos.)

Pregunta 4. Optativa. (2 ptos.)

Diga cuál fue el grupo de edad en el que más bajó la fecundidad desde 2001 hasta 2020 en Uruguay y cuál es el método anticonceptivo que se cree que contribuyó a esa reducción.

El método que se cree que contribuyó fue el método de planificación familiar y el grupo de edad que ha venido disminuyendo la tasa de fecundidad son los jóvenes menores de 25 años (no recuerdo exactamente la edad) , es decir los embarazos adolescentes

install.packages("olsrr")

install.packages("qpcR")

install.packages("leaps")

install.packages("faraway")

install.packages("lmtest")

install.packages("e1071")

library(car)

library(olsrr)

library(qpcR)

library(leaps)

library(faraway)

library(lmtest)

library(tseries)

library(e1071)

#a

#excluyendo distrito

base2<-exampractica[,-5]

save(base2,file="base2.Rdata")

load("base2.Rdata")

#excluyendo id

base3<-base2[,-1]

save(base3,file="base2.Rdata")

load("base2.Rdata")

scatterplotMatrix(base3)

#preg1 b

attach(exampractica)

names(exampractica)

modelo = lm(examen.logica~ nota.fisica+nota.mate+distrito)

#i

#No sabemos si realmente hay multicolinealidad, porque las variables nota.fisica y nota.mate se encuentran en zona de indecisión

vif(modelo)

#ii

#H0: las variables que predicen el modelo son significativas

#H1: las variables que predicen el modelo no son significativas

summary(modelo)

#c

#La multicolinealidad está afectando a la variable distrito porque por el p value estamos observando que esta variable no es significativa

#d

View(exampractica)

base = exampractica[exampractica$distrito == 1,]

modelo1 =lm(examen.logica~ nota.fisica+nota.mate, data = base)

#i

vif(modelo1)

#ii

summary(modelo1)

#e

#Si hay diferencias porque a nivel de toda la población vemos que por el método del p value en la nota de física se presenta multicolinealidad mientras que a nivel de toda la población esto no ocurría y por medio de los vif vemos que a nivel del distrito no presenta multicolinealidad sin embargo a nivel poblacional teníamos una zona de indecisión

#pregunta 2

attach(baspais)

names(baspais)

#a

modelo2=lm(cobertura\_escolar~ coop\_intlal2019+conflicto+porc\_deudaexterna+porcmenoresde15+crecimpib+flujoneto\_invextranjera+porcrural+ porc\_urbanomarginal+continente)

summary(modelo2)

#B2= por cada aumento de un punto porcentual de niños matriculados en educación primaria se espera que en promedio la matricula cuando hay conflictos bélicos 3,888 menos que cuando no se presentan conflictos

#b

###H0: Sigma\_i = Sigma

###H1: Sigma\_i <> Sigma

plot(modelo2$fitted.values,modelo2$residuals)

ncvTest(modelo2)

#Al 5% de significancia hay suficiente evidencia estadística para no rechazar la hipótesis nula de que la homocedasticidad es diferente

#c

estandariz=rstandard(modelo2)

leverage= hatvalues(modelo2)

plot(leverage,estandariz)

abline(h=0, col=6)

text(leverage[leverage>0.6],estandariz[leverage>0.6],pais[leverage>0.6])

text(leverage[estandariz>2],estandariz[estandariz>2],pais[estandariz>2])

text(leverage[estandariz<(-2)],estandariz[estandariz<(-2)],pais[estandariz<(-2)])

#d

#pvalue

ols\_step\_backward\_p(modelo2, prem=0.05, progress=TRUE)

#aic

step(modelo2, direction="backward")

#No se obtienen la misma ecuación, con el método del pvale la variable predictora que deja es porcrural, y con el método del AIC las variables predictoras que incluyen son cobertura\_escolar, porcmenoresde15, porcrural y continente

#e

modeloaic=lm(cobertura\_escolar~porcmenoresde15+porcrural+continente)

estandariz=rstandard(modeloaic)

leverage= hatvalues(modeloaic)

plot(leverage,estandariz)

abline(h=0, col=6)

text(leverage[leverage>0.20],estandariz[leverage>0.20],pais[leverage>0.20])

text(leverage[estandariz>2],estandariz[estandariz>2],pais[estandariz>2])

text(leverage[estandariz<(-2)],estandariz[estandariz<(-2)],pais[estandariz<(-2)])

summary(cbind(as.numeric(cobertura\_escolar),as.numeric(porcmenoresde15),as.numeric(porcrural),as.numeric(continente))

gastoocio3[identif==192 | identif==2 | identif==158,c(1,5,8,11,12)]

#pregunta 3

potencia <- c()

for (i in 1:1000) {

X2 <- rnorm(20, mean=5, sd= X1/5)

error <- rnorm(20,0,sd=1)

Y <- -3+X2error

modelo= lm(Y~X2)

}

potencia[i]

#pregunta 4

#El método que se cree que contribuyó fue el método de planificación familiar y el grupo de edad que ha venido disminuyendo la tasa de fecundidad son los jóvenes menores de 25 años (no recuerdo exactamente la edad) , es decir los embarazos adolescentes

Fórmulas:

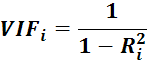


**qi=^**ε**2i /(SCE/n)**

θ**=0.5\*(SCReg\*)**

θ**~**χ**2(p-1)g.l.**



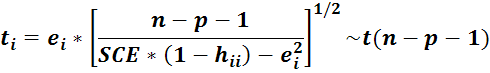






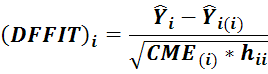


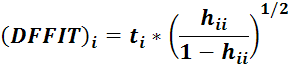


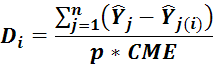


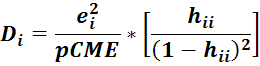


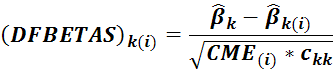
hii>2p/n















AIC = -2 *log-likelihood* + 2p

BIC = -2 *log-likelihood* + *p* log(*n*)

-2 *log-likelihood* = *n* log(*SCRes*/*n*)

