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

Examen Parcial 3.

Se entrega el jueves 1 de diciembre a más tardar 11am, al correo electrónico [gilbert.brenes.camacho@gmail.com](mailto:gilbert.brenes.camacho@gmail.com). Pueden enviar un solo documento de Word 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. También pueden generar un documento con Markdown o Quarto, pero necesariamente tienen que venir transcritas las preguntas con sus respectivos puntajes (de lo contrario se les bajará puntos). El nombre del archivo debe tener su nombre. Por ejemplo, gilbertbrenesexamen3.docx.

Total de puntos: 40 ptos.

Pregunta 1.

Explique en menos de 5 renglones. Desde el principio del semestre, se argumentó que la verosimilitud es equivalente a una probabilidad. Sin embargo, no son exactamente el mismo concepto. Explique cuál es la diferencia entre una probabilidad y una verosimilitud (2 ptos.).

En el caso de la verosimilitud esta busca (o toma en cuenta) los valores máximos o mínimos de un parámetro a partir de los datos con los que se están trabajando a partir de esto se infieren las probabilidades, sin embargo, en las probabilidades no ocurre lo mismo, con esta se puede realizar predicciones hacia una variable o dato especifico.

Pregunta 2.

Explique en menos de 5 renglones. Explique por qué, a partir de una regresió logística, en un gráfico de predichos en el eje de las abscisas contra residuos de deviancia en el eje de las ordenadas, se forman dos nubes u olas de puntos claramente diferenciadas (2 ptos.)

Porque en este tipo de regresión la variable dependiente es binaria, entonces al esta tomar el valor de cero y al hacer el grafico, no va a presentar ningún otro tipo de comportamiento que no sea lineal en cero.

Pregunta 3.

Los intervalos paramétricos tradicionales para un coeficiente de un modelo de regresión lineal son simétricos con respecto de la estimación puntual. O sea, la diferencia entre la estimación del coeficiente y su límite inferior es equivalente la diferencia entre el límite superior y la estimación del coeficiente. Explique por qué los intervalos de confianza no paramétricos de bootstrap en general NO necesariamente son simétricos (2 ptos).

Porque estos intervalos no paramétricos se calculan a partir de las medidas de posición, tanto del percentil 25 y el percentil 75, es decir lo que equivaldría a el RIC, y esto se calcula a partir del promedio de cada coeficiente remuestreado una determinada cantidad de veces. Por lo que no necesariamente estos intervalos tendrían un comportamiento simétrico.

Pregunta 4.

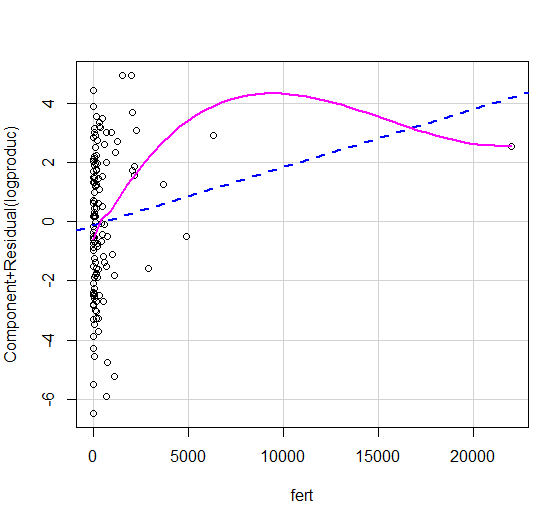
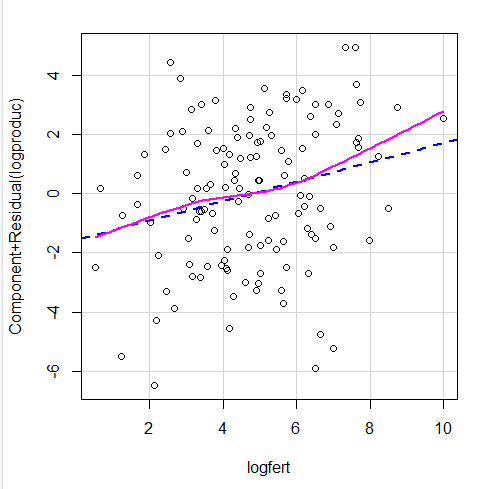
(Basado en Shively, 1998). Se le da un archivo maices2.Rdata en formato R. Corresponde a una muestra de 129 sembradíos de maíz en Filipinas. Se quiere investigar las variables que producen proyectos de construcción de puentes en EEUU. Se quiere investigar qué características del diseño predicen el tiempo para realizar la obra. La base de datos contiene entonces las siguientes variables:

|  |  |
| --- | --- |
| id | Identificador del sembradío |
| logproduc | logaritmo de producción de maíz en Kg por hectárea |
| cerca | Variable binaria: 1=Sembrado en cerca, 0=Otro tipo de sembrado |
| logtrabajo | logaritmo de la cantidad de días hombre por hectárea |
| fert | Cantidad de fertillizante, en Kg por hectárea |
| epocaseca | Época de medición: 1=Época seca, 0=Época lluviosa |
| ocupacion | Porcentaje de parcela destinada a la producción (en puntos porcentuales) |

Conteste con R las siguientes preguntas.

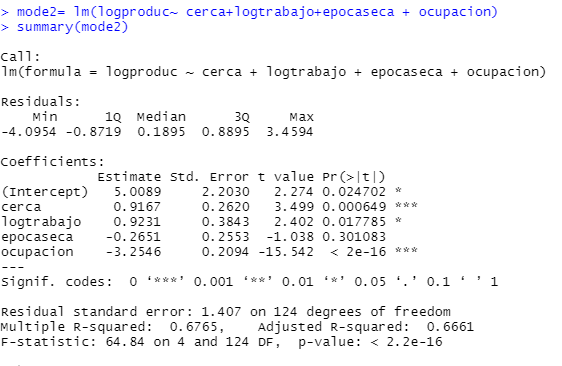
1. Genere un gráfico tipo crPlots a partir de un modelo en el que logproduc esté en función de fertilizante. Diga cómo es la forma de las curvas que aparecen en el gráfico y argumente cuál es la mejor transformación que se le puede hacer a la variable fert para poder suponer linealidad (Para ganarse todos los puntos, la transformación tiene que estar bien justificada a partir de las formas del gráfico) (4 ptos.)

La mejor forma de transformar la variable fert es con el logaritmo, al tener valores un poco extremos en x y en y, y también por tener la mayor cantidad de datos acumulados, es más conveniente usar este tipo de transformaciones.

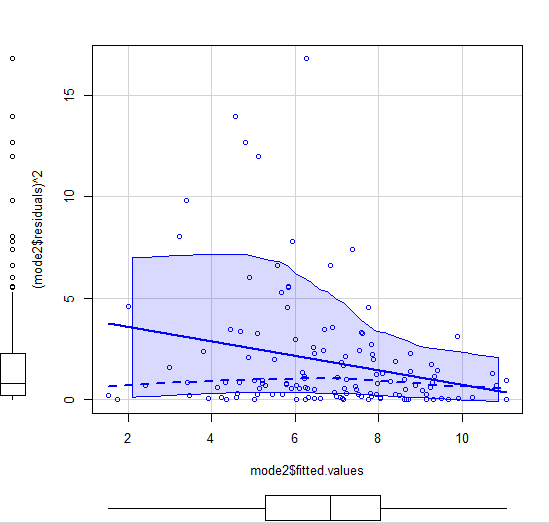
1. Estime un modelo en el que logproduc esté en función de cerca, logtrabajo, epocaseca y ocupacion (o sea, el modelo NO incluye fert). Con este modelo, conteste las siguientes preguntas:
   1. Independientemente de la probabilidad asociada, interprete el coeficiente de la variable cerca. (Idealmente con una interpretación fácil de entender). (4 ptos.)

Si el sembrado se realiza en cerca, se espera que, la producción de maíz aumente, en promedio 91.67% kg por hectárea, esto con respecto a otros tipos de sembrado.



Genere un gráfico con la función scatterplot de los valores predichos en el eje de las abscisas y los residuos al cuadrado en el eje de las ordenadas. Interprete dicho gráfico en términos del cumplimiento del supuesto de homoscedasiticidad. Justifique su respuesta a partir de la forma del gráfico (4 ptos.)

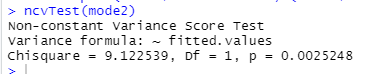




Según el área que se puede observar no se presenta una forma semejante a la de un cuadrado, por tanto, no podemos suponer homocedasticidad

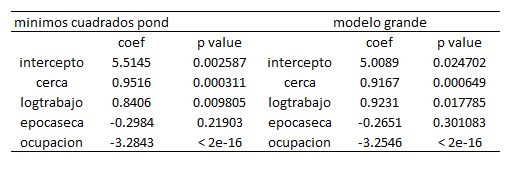
* 1. Con una prueba ncvTest (variante de la prueba de Breusch-Pagan) al 5% de signficancia, analice el cumplimiento del supuesto de homoscedasticidad. Tiene que plantear las hipótesis nula y alternativa, y concluir apropiadamente (4 ptos.)

#Ho: sigmai=sigma

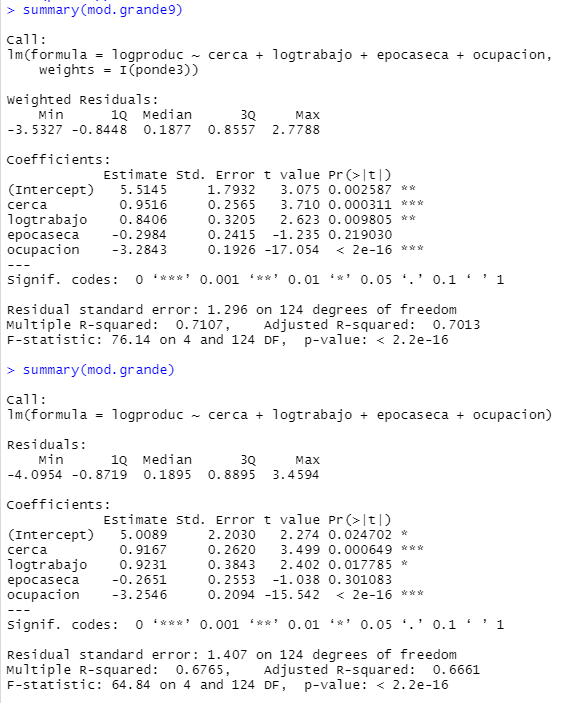
#Hi: sigmai<>sigma

Hay suficiente evidencia estadística para rechazar la hipótesis nula de varianzas iguales

1. Con los mismos predictores del punto anterior, estime un modelo lineal gaussiano con mínimos cuadrados ponderados en el que la ecuación auxiliar para estimar los ponderadores sea a partir de los residuos en valor absoluto. Use tres iteraciones. Haga un cuadro en el que compare los coeficientes estimados del modelo original sin ponderar (el de la pregunta anteriro) con los coeficientes estimados del modelo con mínimos cuadrados ponderados, así como los respectivos pvalues. A partir de esta comparación, explique si el cumplimiento o no del supuesto de homoscedasticidad estaba afectando las conclusiones sobre las variables que predicen la producción (5 ptos.).

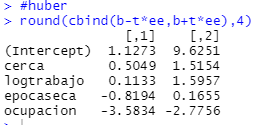
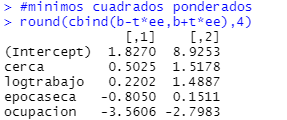


No afectó porque las variables que eran significativas se mantuvieron.



1. Ahora estime un modelo de regresión robusta de Huber, y compare los intervalos de confianza de este modelo con el modelo con mínimos cuadrados ponderados. ¿Se parecen los resultados o no? Explique cuáles son las razones teórico-metodológicas de las similitudes o diferencias entre estas estimaciones (5 ptos.)

La regresión robusta tratar de controlar los valores extremos y los mínimos cuadrados ponderados al tratar de corregir por lo tanto el cálculo de los intervalos presentarían resultados similares, por el comportamiento de las correcciones

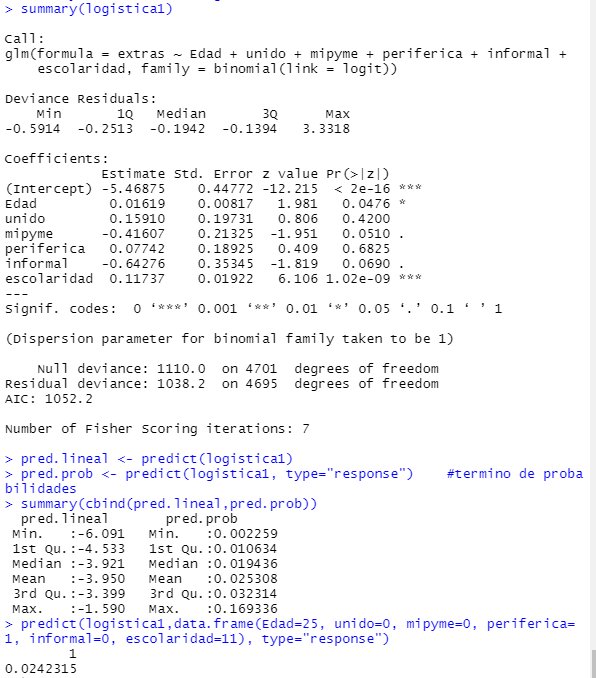
Pregunta 5.

Se le da el archivo ece2020horas.Rdata que contiene una submuestra de asalariados de la Encuesta Continua de Empleo del 3er trimestre de 2020. Se desea estimar una ecuación que prediga la probabilidad de que un asalariado trabaje horas extras. Las variables se describen en un cuadro más abajo. La variable dependiente es extras. Con base en esta definición conteste las siguientes preguntas:

|  |  |
| --- | --- |
| Variable | Explicación |
| id | Identificador de entrevistado |
| Sexo | 1=Hombre, 2=Mujer |
| Edad | Edad en años cumplidos |
| Calificacion\_ocupacion\_COCR11 | Nivel de calificación de la ocupación: 1=Alta, 2=Media, 3=Baja |
| Sector\_actividad | 1=Primario, 2=Secundario, 3=Servicios |
| unido | Estado conyugal: 1=Unido, 0=No unido |
| mipyme | Labora en una Micro o Mediana Empresa: 1=Sí, 0=No |
| periferica | Provincia de residencia: 1=Guanacaste, Puntarenas y Limón, 0= Resto |
| informal | Formalidad del empleo: 1=Informal, 0=Formal |
| escolaridad | Años de escolaridad |
| extras | Variable binaria: 1=Hace horas extras, 0=No hace horas extras |

1. Estime un modelo logístico en el que la probabilidad de trabajar horas extras esté en función de la edad, estado conyugal, tipo de empresa en la que labora, provincia de residencia, grado de formalidad del empleo, y escolaridad. Note que las variables predictoras cualitativa ya son binarias entonces no le recomiendo pasarlas a factor. Déjelas tipo numérico. Con base en esta ecuación, prediga la probabilidad de que labore horas extras un trabajador con las siguientes características: persona soltera de 25 años con 11 años de escolaridad y que vive en Guanacaste, y que trabaja en una empresa que NO es MIPYME del sector formal (4 ptos.)

La probabilidad de que una persona con las características antes mencionadas la probabilidad que labore horas extras es de 0,0242

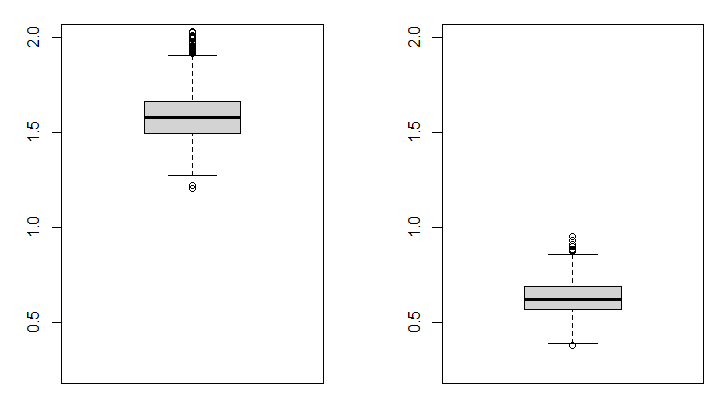


Pregunta 5. Si la regresión robusta controla el efecto de los valores extremos tanto en la estimación de los coeficientes como en los errores estándar, entonces uno podría suponer que, ante la presencia de valores extremos, los coeficientes tienen menores errores estándar y están más cercanos al valor poblacional. Con 1000 corridas y un tamaño de muestra de 50, cree una variable dependiente aleatoria con la siguiente especificación:

Donde:

|  |  |  |
| --- | --- | --- |
|  | = | 0.5 |
|  | = | 1 |
|  | = | Variable aleatoria generada de una distribución uniforme con mínimo=0 y máximo=1 |
|  | = | 10 |
|  | = | Variable binaria en la que los primeros 45 valores son iguales a 0 y los últimos 5 valores iguales a 1 |
|  | = | Varible de error con distribución normal con media 0 y desviación estándar igual a uno |

El componente es para generar valores extremos. Dentro de cada de las mil corridas, estime dos modelos de regresión lineal simple en el que yi esté en función únicamente de x1i; el primer modelo es una regresión gaussiana básica y el segundo modelo es una regresión robusta con la ponderación de huber. En cada una de las 1000 corridas, guarde el valor estimado y su respectivo erorr estándar. Con gráficos de cajas o de violines compare las distribuciones de los y de los errores estándar entre la regresión básica y la regresión robusta. En 5 renglones, qué concluye sobre la afirmación inicial del problema (4 ptos.).



El primer grafico corresponde al modelo de regresión lineal simple y el segundo grafico corresponde al modelo de regresión robusta.

Vemos que en el modelo de regresión robusta se presentan menos valores extremos y en efecto se logra apreciar que la variabilidad del segundo grafico es menor, y con respecto a que estén más cercanos al valor poblacional en efecto se cumple, vemos que el RIC que equivale al 50% de los datos es más pequeño que en el caso de la regresión lineal simple.

attach(maices2)

names(maices2)

logproduct1= exp(logproduc)

logfert=exp(fert)

mode1= lm(logproduct1~ logfert)

crPlots(mode1)

mod.grande= lm(logproduc~ cerca+logtrabajo+epocaseca + ocupacion)

summary(mod.grande)

scatterplot(mode2$fitted.values, (mode2$residuals)^2)

ncvTest(mod.grande)

ncvTest(mod.grande9)

abs.res1=abs(residuals(mod.grande))

mod.ponde1=lm(abs.res1~cerca+logtrabajo+epocaseca + ocupacion)

ponde1=1/abs(fitted(mod.ponde1))

mod.grande7=lm(logproduc~cerca+logtrabajo+epocaseca + ocupacion, weights=I(ponde1))

summary(mod.grande7)

summary(mod.grande)

ncvTest(mod.grande)

ncvTest(mod.grande7)

round(cbind(mod.grande$coef,mod.grande7$coef),4)

###Se comparan los coeficientes para analizar si cambian mucho o no.

###Noten los cambios en tamhogar, escoljefe, alquilada, prestamcasa y pobmenores

###2 iteración

abs.res2=abs(residuals(mod.grande7))

mod.ponde2=lm(abs.res2~cerca+logtrabajo+epocaseca + ocupacion)

ponde2=1/abs(fitted(mod.ponde2))

mod.grande8=lm(logproduc~cerca+logtrabajo+epocaseca + ocupacion, weights=I(ponde2))

summary(mod.grande8)

summary(mod.grande)

round(cbind(mod.grande$coef,mod.grande7$coef,mod.grande8$coef),4)

###3 iteración

abs.res3=abs(residuals(mod.grande8))

mod.ponde3=lm(abs.res3~cerca+logtrabajo+epocaseca + ocupacion)

ponde3=1/abs(fitted(mod.ponde3))

mod.grande9=lm(logproduc~cerca+logtrabajo+epocaseca + ocupacion, weights=I(ponde3))

summary(mod.grande9)

summary(mod.grande)

mod.huber=rlm(logproduc~ cerca+logtrabajo+epocaseca + ocupacion)

summary(mod.huber, cor=FALSE)

ee=summary(mod.huber)$coef[,2]

glib=length(maices2$logproduc)-length(mod.huber$coef)

t=qt(.975,glib)

b=mod.huber$coef

#huber

round(cbind(b-t\*ee,b+t\*ee),4)

mod.grande9=lm(logproduc~cerca+logtrabajo+epocaseca + ocupacion, weights=I(ponde3))

summary(mod.grande9, cor=FALSE)

ee=summary(mod.grande9)$coef[,2]

glib=length(maices2$logproduc)-length(mod.huber$coef)

t=qt(.975,glib)

b=mod.huber$coef

#minimos cuadrados ponderados

round(cbind(b-t\*ee,b+t\*ee),4)

Lim.inf=mode1$coef-qt(0.975,length(logproduc)-length(mode1$coef))summary(mode1)$sigma(diag(summary(mode1)$cov.unscaled))^0.5

Lim.sup=mode1$coef+qt(0.975,length(logproduc)-length(mode1$coef))summary(mode1)$sigma(diag(summary(mode1)$cov.unscaled))^0.5

round(cbind(Lim.inf,mode1$coef,Lim.sup),4)

#xbarra+-1.96\*s/sqrt(n)

attach(ece20horas)

names(ece20horas)

logistica1=glm(extras~Edad+unido+mipyme+periferica+ informal+escolaridad,family=binomial(link=logit))

summary(logistica1)

plot(logistica1$fitted,logistica1$residuals)

pred.lineal <- predict(logistica1)

pred.prob <- predict(logistica1, type="response") #termino de probabilidades

summary(cbind(pred.lineal,pred.prob))

predict(logistica1,data.frame(Edad=25, unido=0, mipyme=0, periferica=1, informal=0, escolaridad=11), type="response")

ee=summary(mod.huber)$coef[,2]

glib=length(maices2$logproduc)-length(mod.huber$coef)

t=qt(.975,glib)

b=mod.huber$coef

round(cbind(b-t\*ee,b+t\*ee),4)

almacenee=rep(NA,1000)

almacenb1g=rep(NA,1000)

almacenee1=rep(NA,1000)

almacenb1r=rep(NA,1000)

for (i in 1:1000) {

x1=runif(50,0,1)

x2=c(rep(0,45),rep(1,5))

e=rnorm(50,0,1)

yi=0.5+1\*x1+10\*x2+e

mod=lm(yi~x1)

mod.huber1=rlm(yi~x1)

almacenee[i]=summary(mod)$coefficients[2,1]

almacenb1g[i]=summary(mod)$coefficients[2,2]

almacenee1[i]=summary(mod.huber1)$coefficients[2,1]

almacenb1r[i]=summary(mod.huber1)$coefficients[2,2]

}

par(mfrow=c(1,2))

boxplot(almacenb1g,ylim=c(0.25,2))

boxplot(almacenb1r,ylim=c(0.25,2))