Manual de prácticas

Preguntas

#### **Pregunta 1.**

*Explique en menos de 5 renglones. ¿Cuál es la base teórica por la que se argumenta que el Método de Mínimos Cuadrados Ponderados corrige el problema de la heteroscedasticidad?.*

#### **Pregunta 2.**

*Explique en menos de 5 renglones. ¿Cuál es la base teórica por la que se argumenta que el Método de Componentes Principales corrige el problema de multicolinealidad?*

#### **Pregunta 3.**

*Se le da un archivo bridges.Rdata en formato R. Corresponde a una muestra de 45 proyectos de construcción de puentes en EEUU. Se quiere investiga*

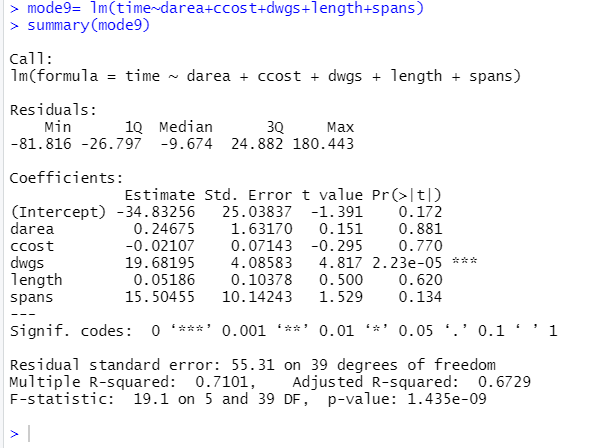
*#ar qué características del diseño predicen el tiempo para realizar la obra.*

La base de datos contiene entonces las siguientes variables:

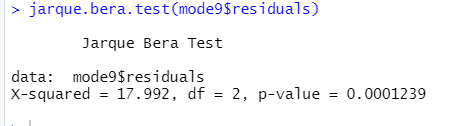
* **case:** Identificador del caso
* **time:** Tiempo en días-persona
* **darea:** Área superficial, en miles de pies cuadrados
* **ccost:** Costo de construcción, en miles de dólares
* **dwgs:** Número de dibujos estructurales
* **length:** Largo del puente (en pies)
* **spans:** Número de soportes

load("bridge.Rdata")

1. *Estime un modelo de regresión gaussiano (el tradicional de mínimos cuadrados ordinarios) en el que se prediga el tiempo para realizar las obras en función del resto de variables (excepto case por ser el identificador). Llámelo mod1. Presente el summary del modelo*



1. *Analice la normalidad de los residuos con una prueba de Jarque Bera. Plantee la hipótesis nula y alternativa, y responda adecuadamente. Conteste: ¿Se puede suponer normalidad de los errores? ¿Por qué sí o por qué no?*

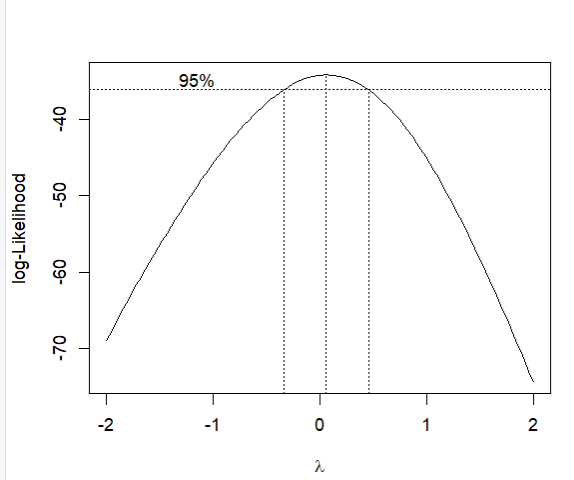


H0= miu=0

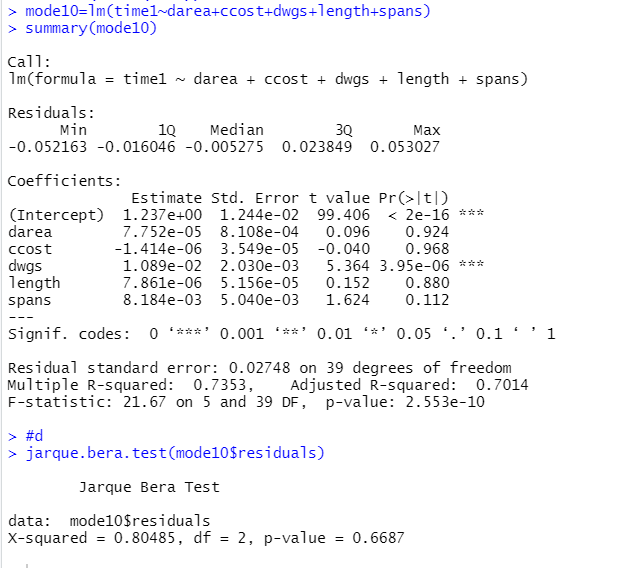
H1=miu<>0

Hay suficiente evidencia estadística para rechazar la hipótesis nula de que los errores se distribuyen normalmente

1. *Utilice el método de BoxCox para encontrar una transformación para la variable dependiente. Pegue el gráfico en el examen. Escoja un exponente redondeado que sea razonable de acuerdo al gráfico de BoxCox*

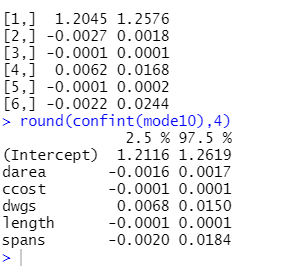
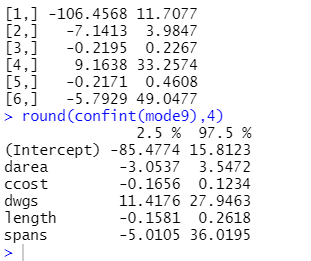


**d)** *Estime un nuevo modelo gaussiano con la variable dependiente transformada, llame al modelo mod2, y verifique con la prueba de Jarque Bera si ya se puede suponer normalidad. No es necesario que plantee las hipótesis nula y alternativa. Solo señale la conclusión.*



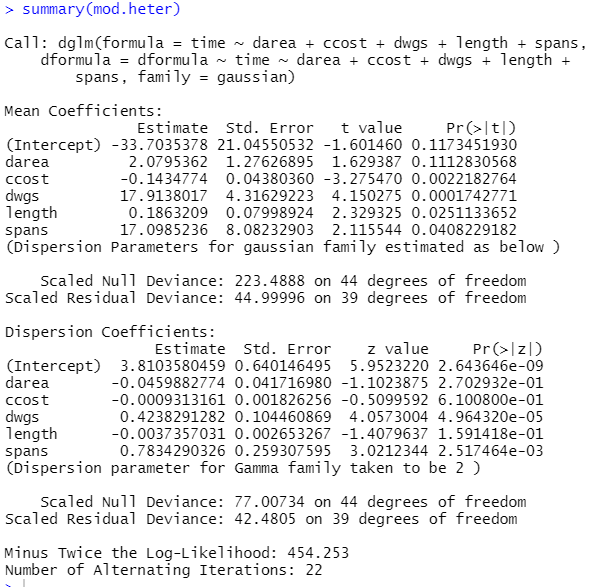
Ya se puede suponer normalidad, es decir, no hay suficiente evidencia estadística para rechazar la hipótesis nula de que los errores se distribuyen normalmente.

1. *Utilice el método de Bootstrap para estimar intervalos no paramétricos para ambos modelos (mod1 y mod2), y compárelos con los intervalos paramétricos teóricos de ambos modelos (mod1 y mod2). A partir de la comparación con los intervalos de bootstrap, argumente si hay evidencia de que la transformación mejora las estimaciones del modelo.*



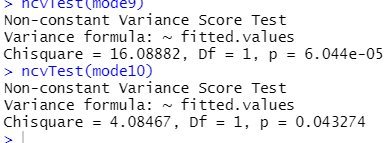
Vemos que en el modelo transformado los limites de los modelos con muy similares, sin embargo en el modelo sin trasformaciones el intervalo con boostrap se vuelve mas amplio por lo que es menos precisa la inferencia.

1. *Estime un modelo heteroscedástico (modelo lineal generalizado doble dglm) con la especificación del modelo1 y llámelo mod1h. A partir de las estimaciones de este modelo, justifique si hay presencia de heteroscedasticidad o no, y cómo lo sabe.*



No todos los coeficientes son significativos, la gran mayoría si lo son, asi que podemos asumir que si hay homocedasticidad

1. *Estime la prueba de variancia no constante (ncvTest) para el modelo 1 y el modelo 2. Plantee una sola vez la hipótesis nula y alternativa. Diga si hay alguna diferencia en el supuesto de homoscedasticidad entre el modelo 1 y el modelo 2, y cómo una transformación de Box Cox (que es para corregir violaciones a la normalidad) puede generar similitudes o diferencias en los resultados de estas pruebas*



#### **Pregunta 4.**

*La Asociación Nacional de Productores de Cebolla encarga un estudio para analizar qué condiciones son las que les permite tener ganancias o pérdidas a los vendedores de cebolla en las Ferias del Agricultor. A la Asociación no le interesa el monto de la ganancia, sino si tuvo pérdidas o ganancias cada día de la feria del agricultor. Toma una muestra de 2500 cebolleros y los entrevista por teléfono el día después de la Feria. Se construye una base de datos cebolleros.dta con las siguientes variables:*

* **id:** identificador del vendedor
* **morada:** Si vende cebolla seca amarilla o morada (1=morada, 0=amarilla)
* **personasenpuesto:** Cuántas personas lo ayudan a vender.
* **ganancia8:** Variable binaria: 1= ganancia, 0=pérdida

Se estimó un modelo de regresión logística que generó las siguientes estimaciones:

> summary(glm(ganancia8~morada+personasenpuesto, family=binomial(link="logit"),data=cebolleros))

Call:

glm(formula = ganancia8 ~ morada + personasenpuesto, family = binomial(link = "logit"),

data = cebolleros)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.9397 -0.8472 -0.7500 1.5161 1.7090

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.19624 0.09162 -13.056 < 2e-16 \*\*\*

morada 0.32027 0.08998 3.559 0.000372 \*\*\*

personasenpuesto 0.03590 0.03045 1.179 0.238281

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2959.1 on 2499 degrees of freedom

Residual deviance: 2944.7 on 2497 degrees of freedom

AIC: 2950.7

Number of Fisher Scoring iterations: 4

**a)** *Interprete los coeficientes de las variables “morada” y “personasenpuesto”, haciendo la transformación apropiada*

**b)** *Explique cuál es la lógica de usar la prueba de Hosmer y Lemeshow como una prueba de bondad de ajuste para una regresión logística.*

#### **Pregunta 5.**

*Se le da el archivo dependiente.Rdata. Este contiene una sola variable: dependiente. Con un ciclo, encuentre la estimación de lambda y el valor de la logverosimilitud del modelo, usando el método de máxima verosimilitud tal que los residuos del modelo en el que la variable dependiente es igual a (dependiente^lambda) se distribuya aproximadamente normal (En otras palabras, las estimaciones de la transformación de BoxCox, pero no use la transformación de BoxCox sino una programación con ciclos.). Tiene que evaluar 100 valores en el ciclo, entre 0.01 a 1.*

*Se sabe que el error estándar residual es igual a: 0.0836, el Beta0=1.85 (bajo la transformación) y que la función de verosimilitud del método de BoxCox es:*



#### **Pregunta 6.**

*Explique en menos de 5 renglones. El procedimiento de Bootstrap utiliza remuestreo (una cantidad grande de muestras con reemplazo de la muestra original). Si hay presencia de heteroscedasticidad, explique qué característica(s) tiene el procedimiento de bootstrap para que los intervalos no paramétricos de bootstrap para una pendiente sean preferibles a los intervalos de confianza estimados con la fórmula teórica* .

#### **Pregunta 7.**

*La pandemia de COVID-19 ha popularizado una serie de términos que se restringían a la jerga científica de las disciplinas de la salud: reproducibilidad, carga viral, período de latencia, etc. Suponga que surge un nuevo patógeno infeccioso, el SARS-COV3, que no es mortal pero sí enferma a las personas durante un período prolongado. Virólogos recogen especímenes de saliva de una muestra de 50 pacientes infectados con el nuevo patógeno, y miden la carga viral 3 días después del primer síntoma de fiebre (temperatura corporal mayor o igual a 38°C). Ellos tienen la hipótesis que la carga viral varía por edad y sexo, así que le piden a usted que estime un modelo de regresión que sirva para predecir la carga viral (variable carga.viral10, en copias por mL) en función de la edad (medida en años) y el sexo (variable mujer, que es igual a 1 si la persona es mujer, y 0 si es hombre). La información está en la base de datos sarscovtri.Rdata. Conteste las siguientes preguntas:*

load("sarscovtri.Rdata")  
str(sarscovtri)

## 'data.frame': 50 obs. of 4 variables:  
## $ X1.50 : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ carga.viral10: num 52.2 42.1 27.6 26.3 53.3 ...  
## $ edad : num 29 43 37 40 30 50 30 47 46 28 ...  
## $ mujer : int 0 1 0 0 1 1 0 1 1 1 ...

**a)** *Estime un modelo lineal gaussiano (el modelo de regresión tradicional) y analice el supuesto de normalidad con un qqPlot de los residuos y una prueba de Shapiro (use un alfa de 5%). Diga qué concluye sobre el supuesto de normalidad*

**b)** *Haga gráficos crPlots con la ecuación y diga si hay alguna transformación que usted sugeriría para las variables predictoras. Explique su elección*

**c)** *Estime un nuevo modelo lineal gaussiano con la transformación sugerida. Compare la bondad de ajuste de este modelo con el modelo del inciso (a), y diga si la transformación mejoró en algo la bondad de ajuste del modelo. [Nota: Para obtener todos los puntos de la pregunta, tiene que enseñarme un nuevo modelo con un crecimiento sustancial en el indicador de bondad de ajuste]*

**d)** *Haga un gráfico qqPlot de los residuos y una prueba de Shapiro (al 5% de significancia) con el nuevo modelo, y diga qué concluye sobre la violación al supuesto de normalidad.*

**e)** *A partir del análisis anterior, explique cómo se relacionan los supuestos de linealidad y normalidad.*

#### **Pregunta 8.**

*Se le da el archivo promociones.Rdata de 100 clientes de una tienda que contestaron un cuestionario para determinar quiénes estarían dispuestos a recibir promociones por Internet. La tienda quiere escoger un modelo qué determine cuáles variables predicen correctamente el deseo de recibir información, por lo que decide estimar un modelo logístico para determinar la probabilidad de que un cliente quiera recibir este tipo de emails. Las variables son las siguientes:*

load("promociones.Rdata")  
names(promociones)

## [1] "edad" "mujer" "gastos" "promocion" "apellido"

**a)** *Estime un modelo de regresión en el que se prediga la probabilidad de querer recibir emails de promoción, en función de la edad, el ser mujer y los gastos mensuales, e interprete el coeficiente de la variable edad.*

**b)** *Estime un modelo de regresión en el que se prediga la probabilidad de recibir emails de promoción en función de mujer únicamente. Conociendo los resultados del beta0 y del beta1, programe un procedimiento en el que pruebe varios valores de beta0 y beta1 en la función de verosimilitud del modelo logístico para variable binaria (en clase lo vimos con ciclos), para encontrar las estimaciones de máxima verosimilitud para Beta0 y Beta1. [Recomendación: Acepto que pruebe solo 100 valores para Beta0 y 100 valores para Beta1, o sea un ciclo con 100mil corridas]*

#### **Pregunta 9.**

*Si en una regresión con “p” predictores, se utiliza un Análisis de Componentes Principales con todos los predictores, explique en qué condiciones y por qué la suma de los valores característicos (“lambdas”) estimados con el modelo es equivalente a “p”.*

#### **Pregunta 10.**

*Se le da el archivo enignorte que contiene datos de la Encuesta Nacional de Ingresos y Gastos de los Hogares 2018, referente a las zonas urbanas de la Región Norte. Se quiere predecir el gasto en alimentos en función del ingreso total del hogar, la cantidad de miembros del hogar y la edad del jefe. Los nombres de las variables se tienen a continuación*

load("enignorte.Rdata")  
names(enignorte)

## [1] "ID\_REGION" "ID\_ZONA" "id" "gastoalim"   
## [5] "ingresototal" "miembroshogar" "edadjefe" "resid"

Con base en dicha base de datos conteste las siguientes preguntas:

**a)** *Estime un modelo de regresión gaussiano en el que se prediga el gasto en alimentos en función del ingreso total del hogar, la cantidad de miembros del hogar y la edad del jefe. Muestre el summary del modelo.*

**b)** *Haga un gráfico de residuos contra predichos y diga cómo evalúa el supuesto de homoscedasticidad.*

**c)** *Estime un modelo de mínimos cuadrados ponderados con dos iteraciones y en que los residuos absolutos sean predichos por los predictores para calcular los ponderadores. Muestre el summary del modelo final.*

**d)** *Estime un modelo hesteroscedástico (un modelo lineal generalizado gaussiano doble) con las variables predictoras en las ecuaciones de la media y la variancia, y a un 5% de significancia diga si hay variables asociadas con heteroscedasticidad.*

**e)** *Haga un cuadro en el que compare los coeficientes y los p-values de los 3 modelos (guassiano en inciso (a), mínimos cuadrados ponderados en inciso (c) y heteroscedástico en inciso (d)), y diga en qué se parecen y se diferencian las conclusiones de los tres modelos (use un alfa de 5%).*

**f)** *En términos del procedimiento de estimación, diga en qué se parecen en general los modelos estimados con mínimos cuadrados ponderados y los modelos heteroscedásticos.*