**Manual de prácticas segundo parcial:**

load("acciones.Rdata")

1. Estime un modelo en que el DEG esté en función de la balanza de pagos e interprete la pendiente.
2. Analice si se cumple el supuesto de normalidad de los errores comparando una prueba de Shapiro y un qq-Plot, ambos con residuos estandarizados. Coinciden o no estos dos diagnósticos.
3. Explique en menos de tres renglones cuál es el efecto de la violación del supuesto de normalidad en un modelo lineal gaussiano.
4. Estime los residuos estandarizados, los leverage y los dfBetas para el beta1 del modelo analizado, y haga un gráfico de burbujas de residuos estandarizados en el eje de las abscisas, los leverage en el eje de las ordenadas, y el tamaño de las burbujas que sea proporcional al DFbeta de la pendiente. Diga si existe un valor extremo que sea a la vez influyente, identifíquelo en el gráfico con el nombre del país, y explique cómo llegó a esa conclusión.
5. Vuelva a estimar un modelo excluyendo el caso que es valor extremo, y diga cómo cambiaron las estimaciones del modelo.

load("supermerc2.Rdata")

1. Estime un modelo (con el nombre mod.1) en el que se prediga las ventas según las siguientes variables: q.empacado, q.carnes, amb.interno, num.pasillos, area.tienda y distancia, haga un summary del mismo, e interprete los coeficientes de num.pasillos y q.carnes.
2. Estime los coeficientes estandarizados de esta ecuación. Después, usando la función ya programada en R, grafique los gráficos de regresión parcial para todas las variables y argumente si la variable con el menor coeficiente estandarizado es también la variable que menor aporta a la explicación del modelo según estos gráficos. Explique cómo llegó a esa conclusión.
3. Analice el VIF de este modelo, quite una variable del modelo original, estime un nuevo modelo (con el nombre mod.2) sin esa variable, haga un summary del modelo y diga si la multicolinealidad estaba afectando al modelo del inciso a y cómo llegó a esa conclusión. (Se le da libertad de escoger cuál variable quitar, pero tiene que ser una que indique problemas de multicolinealidad).
4. Utilice el procedimiento del modelo con el menor BIC usando la función de subsets para seleccionar un modelo más parsimonioso, y diga si este modelo coincide con el modelo que seleccionó anteriormente.
5. Explique si los 2 modelos de los incisos anteriores presentan evidencia de que la multicolinealidad les esté generando problemas de interpretabilidad.
6. Haga el siguiente procedimiento:
7. En un proceso de 1000 iteraciones, siga los siguientes pasos:

* Tome una muestra de 30 observaciones y defina t como la posición en el vector.
* Genere una variable X1 con distribución normal con media igual a 4 y desviación estándar igual a 1.
* Genere una variable de error con distribución moral estándar (media igual a 0 y desviación estándar igual a 1).
* Genere una variable Y=2+1∗X1+errorY=2+1∗X1+error
* Genere otra variable Ystar[t]=Y[t−1]∗0.4Ystar[t]=Y[t−1]∗0.4. (En otras palabras, Y con rezago de orden 1. Dada esta operacionalización, pierde la primera observación, porque va a tener un NA en la primera posición de Ystar).
* Estime un modelo de regresión en el que Ystar está en función de X1, con los casos de X que van desde la posición 1 a la 29.
* Guarde el cuadrado medio de error en un vector de almacenamiento.
* Estime un modelo de regresión con los casos que van desde la posición 1 a la 29 en el que Y está en función de X1.
* Guarde el cuadrado medio de error en un vector de almacenamiento.

1. Cuando termine el ciclo, compare las medias de los 2 cuadrados medios de error, y diga cómo ilustra esta comparación el efecto que tiene la autocorrelación sobre las estimaciones de modelos de regresión.
2. El criterio “delete-one” (“quitar uno”) se utiliza para el cálculo del PRESS y de los residuos estudentizados. Explique cómo este criterio del “delete-one” para un valor extremo hace que |Yi−Y^i(i)||Yi−Y^i(i)| sea mayor que el |Yi−Y^i||Yi−Y^i|.
3. Explique por qué el BIC penaliza en un mayor grado la cantidad de predictores de un modelo que el AIC.
4. Explique por qué se dice que el Cp de Mallows es un indicador de sesgo en la variancia del error.

load("baseex1.Rdata")

1. Estime un modelo de regresión gaussiana (con el nombre mod1) en el que se prediga la cuota de préstamos en función del resto de variables (excepto id), haga un summary del mismo, e interprete los coeficientes de escoljefe y pobmenor5.
2. Los siguientes puntos se hacen con mod1:

* i. Calcule los vifs del modelo mod1 y explique a qué conclusión se llega sobre la multicolinealidad en mod1 a partir de los valores del vif.
* **ii.** Estime dos modelos más. El modelo mod1a tiene la misma especificación del mod1, pero excluyendo la variable escoljefe. El modelo mod1b tiene la misma especificación del mod1, pero excluye la variable numvehic. Haga una tabla con el R cuadrado ajustado y el P2 de los modelos mod1, mod1a y mod1b y escoja uno de los 3 modelos, justificando su selección.
* **iii.** Los modelos mod1a y mod1b son modelos reducidos. El mod1 es el modelo original. Dada la estructura multivariada de las variables del modelo (en otras palabras, las relaciones entre las variables), diga por qué el R2 ajustado de uno de los modelos reducidos es mayor que el R2 ajustado del modelo original, pero el R2 ajustado del otro modelo reducido es menor que del original.
* **iv.** Haga un histograma de la variable dependiente y un histograma de los residuos de mod1. Explique por qué el histograma de la variable dependiente es más asimétrico que el de los residuos.
* **v.** Dado que estos son datos reales, explique por qué es esperable que la variable dependiente del modelo NO tenga una distribución normal en Costa Rica.

1. Estime un modelo de regresión con menos variables predictoras, usando un algoritmo stepwise basado en el AIC, y llámelo mod2. Diseñe gráficos de dispersión de la variable dependiente vs cada una de las variables predictoras seleccionadas en mod2. Explique claramente en cuáles gráficos se observa evidencia de linealidad en la relación entre la dependiente y las predictoras.
2. Los siguientes puntos se hacen con mod2:

i. Con el mod2, diseñe un gráfico de residuos estandarizados vs leverage, y diga el número de identificación del valor más extremo en X y el valor más extremo en Y.

* **ii.** Explique qué características hacen que el valor más extremo en Y sea extremo, y qué características hacen que el valor más extremo en X sea extremo.
* iii. Haga un gráfico de residuos estandarizados contra leverage, con burbujas ponderadas por los defeffit en valor absoluto. Escriba en el gráfico los números de identificador de los dos casos con los Dffit más altos (en valor absoluto). Y sabiendo la fórmula de los Dffit que se le suministra más abajo, explique si la magnitud de los Dffits más altos se debe al componente de residuo o al componente de leverage.



1. En clase se dijo que la prueba de White es más robusta que la prueba de Breusch-Pagan para detectar homoscedasticidad cuando se viola el supuesto de normalidad. Para ello se propone la siguiente simulación. Diseñe un proceso para calcular la potencia de la prueba de Breusch Pagan y la prueba de White con las siguientes características:

* Tome un tamaño de muestra de 30.
* Simule dos variables X1 y X2 con distribución normal con media 0 y variancia 1.
* Simule una variable Z con distribución exponencial con rate=0.1 (Por definición, la variancia de una distribución exponencial cambia con su valor esperado, por lo que un error con distribución exponencial supone un modelo lineal con heteroscedasticidad).
* Cree una variable EE=Z−mean(z)sd(z)EE=Z−mean(z)sd(z).
* Cree una variable YY en función de X1 y X2 y con la variable EE definiendo el error (usted decide los valores de beta0, beta1 y beta2).
* Calcule la potencia de la prueba de Breusch Pagan y de White para un modelo de regresión lineal en que YY esté en función de X1 y X2, con una significancia del 5%.