MODELOS ESTADÍSTICOS LINEALES - IIND 4100
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
Periodo 2022-20. Profesor: Carlos Valencia

TAREA 3

- π El informe con la solución de la tarea debe ser entregado antes del miércoles 23 de noviembre (5pm).
- π La solución puede ser entregada en grupos de máximo 3 personas.
- π Utilice procedimientos explícitos y justifique sus respuestas.

Parte A. Problemas Computacionales.

Problema 1. Bootstrap para pruebas con restricciones lineales.

El método de *bootstrap* no sólamente sirve para construir intervalos de confianza, sino para contrastar pruebas de hipótesis. En este problema, realizaremos una prueba de hipótesis gráfica por medio de *bootstrap*. En este caso, se hace una prueba de hipótesis con restricciones lineales del tipo $H_0 : \mathbf{A}\beta = c$, en donde $q = 2$.
Tenga en cuenta que la hipótesis nula se puede escribir como:

$$H_0 : \begin{cases} \mathbf{a}_1^T\beta = c_1 \\ \mathbf{a}_2^T\beta = c_2 \end{cases}$$

Con las repeticiones del *bootstrap*, calcule en cada iteración los valores de $\mathbf{a}_1^T\hat{\beta}$ y $\mathbf{a}_2^T\hat{\beta}$. Luego con estos B puntos, se puede construir la elipse de confianza empírica que cubre el 95% de los datos. Después, en el mismo gráfico puede dibujar el punto (c_1, c_2), y si este cae dentr de la elipse, no se rechaza la hipótesis nula.

1. Construya una el código que haga el procedimietno que se describe para constrastar H_0 de manera gráfica. Puede usar como guía el siguiente código:

```
library(ellipse)
x1=rnorm(500,20,3)
x2=.5*x1+rnorm(500,1,3)
plot(x1,x2)
x=cbind(x1,x2)
el=ellipse(x,centre=colMeans(x),level=0.95)
plot(x1,x2)
points(el,type="l")
points(25,5,col=2,lwd=2)
```

2. En la base de datos `masa-corporal` que vimos en la clase 13, se determinó que había un problema de multicolinealidad entre las variables. Por esto, puede ser adecuado usar regresión con penalización tipo *Ridge*. Usando *Ridge*, y la prueba que acaba de diseñar, determine si es cierta la afirmación que el coeficiente de grasa corporal (cmi) esperado para dospersonas es 22 y 12 respectivamente, si se sabe que para la persona 1 clic=24, leg=56 y arm=26. Para la persona 2, clic=20, leg=43 y arm=24.

Problema 2. Método de Estimación Robusta.

La estimación robusta resulta ser bastante útil para evitar que algunos puntos observados creen sesgo enla recta estimada. Además de la regresión con minimización de las desviaciones en valor absoluto,

otra solución al problema, es cambiar la distribución de los erroes para que no sea normal, sino que permita la existencia de datos atípicos (sobre todo en Y), esto es *fat tailed distribution* (`https://en.wikipedia.org/wiki/Fat-tailed_distribution`). En ese caso, supondremos que los errores del modelo son proporcionales a una variable aleatoria con distribución t de *Student*. Es decir, se define el modelo

$$Y_i = X_i^T \beta + \varepsilon_i$$

Donde $\frac{\varepsilon_i}{\sigma} \sim t(v)$, donde v son los grados de libertad correspondientes, y σ es un parámetro que permita adapatar la dispersión de los datos. En este modelo, se deben estimar los parámetros de la recta, así como el parámetro de dispersión σ. Sin embargo, no es posible usar mínimos cuadrados dado que la distribución no es normal y se tendrían valores influyentes. Para estimar, se debe hacer por el método de máxima verosimilitud.
Para simplificar el problema, suponga que $p = 2$, es decir, se deben estimar el intercepto, la pendiente y σ. Esto es,

$$\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma} \right) \sim t(v)$$

para cada $i = 1, 2, \ldots, n$, y son independientes entre si.

1. Costruya una función en `R` que tenga como inputs el vector $\mathbf{Y}$, la matrix $\mathbf{X}$, los parámetros β_0 y β_1 y el parámetro de dispersión σ. El output de la función debe ser la verosimilitud calculada. Pueden usar las funciones relacionadas con la distribución t en `R`.
   Tenga en cuenta que esta es la fucnión de verosimilitud $L(\beta_0, \beta_1, \sigma | \mathbf{Y}, \mathbf{X})$ en térrminos de los tres parámetros, la cual se debe optimizar.

2. Como es difícil encontrar la solución del máximo de $L(\beta_0, \beta_1, \sigma | \mathbf{Y}, \mathbf{X})$ analíticamente, se deben usar métodos numéricos (por ejemplo Newton-Raphson - `https://en.wikipedia.org/wiki/Newton\%27s_method`). Por ejemplo, para optimizar la función puede usar un método general construido en `R` como la función `optim` (`https://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html`). Usando la función `optim`, maximice la verosimilitud $L(\beta_0, \beta_1, \sigma | \mathbf{Y}, \mathbf{X})$ mediante un algoritmo iterativo en el cual se maximiza para cada parámetro a la vez:

   I. Empiece con valores iniciales para $\beta_0^{(0)}$, $\beta_1^{(0)}$ y $\sigma^{(0)}$.

   II. Repita hasta convergencia para la iteración i:

      i. Maximice la función de verosimilitud con respecto a β_0 con valores fijos $\beta_1^{(i-1)}$ y $\sigma^{(i-1)}$. Guarde el máximo como $\beta_0^{(i)}$

      ii. Maximice la función de verosimilitud con respecto a β_1 con valores fijos $\beta_0^{(i)}$ y $\sigma^{(i-1)}$. Guarde el máximo como $\beta_1^{(i)}$

      i. Maximice la función de verosimilitud con respecto a σ con valores fijos $\beta_0^{(i)}$ y $\beta_1^{(i)}$. Guarde el máximo como $\sigma^{(i)}$

3. Construya una mecanismo de inferencia por medio de *bootstrap* para construir intervalos de confianza para los tres parámetros de interés. Use los datos anexos `datospunto2.txt` y resporte los intervalos de confianza para β_0, β_1 y σ con 95% y con 99% de confianza. Puede usar 8 grados de libertad.

Parte B. Problemas Teóricos, Conceptuales y Experimentales.

Problema BONO (10 puntos). Uso de *leverage* para medidas de desempeño.
Si $\mathbf{H}$ es la matriz de proyección en Ω definida por el estimador de mínimocs cuadrados, entonces los

elementos de la diagonal ℓ_{ii} son los valores de *leverage* que se miden para cada observación. Como se ha visto, estos valores tienen muchos usos, como por ejemplo determinar la varianza de los residuales o determinar datos atípicos en $\mathbf{X}$. Uno de los usos más útiles, es que permiten calcular la distancia de un punto Y_i a la recta estimada sin ese punto, sin necesidad de correr el modelo con los $n-1$ datos restantes. Es decir:

$$\left(Y_i - \hat{Y}_{i,(i)}\right) = \frac{Yi - \hat{Y}_i}{1 - \ell_{ii}}$$

donde $\hat{Y}_{i,(i)}$ es la estimación del valor esperado Y_i estimada con el modelo que usa todos los datos, menos la observación i.
Esto quiere decir, que si se define la métrica (`https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Leave-one-out_cross-validation`):

$$LOO = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Yi - \hat{Y}_i}{1 - \ell_{ii}}\right)^2$$

entonces, se puede medir el desempeño del modelo, algo parecido a como lo hace el R^2_{adj}, el C_p de Mallows o el AIC.

1. Demuestre que
$$\ell_{ii} = \frac{d\hat{Y}_i}{dY_i}$$
2. Con la información del numeral anterior, demuestre que:
$$\frac{dY_i}{de_i} = \frac{1}{1 - \ell_{ii}}$$
3. Con la información de los dos numerales anteriores, demuestre (explique) porqué
$$d_i = \frac{e_i}{1 - \ell_{ii}}$$
4. Con los datos de `masa-corporal` que vimos en las clases de multicolinealidad, seleccione el modelo más apropiado para explicar el índice de grasa corporal. Compare el resultado con el obtenido si usa: R^2_{adj}, C_p y AIC.

Problema 3. Problemas de especificación y selección de modelos.
Los datos `CDI` contienen infrmación de 440 municipios (counties) de USA con cirtas variables demográficas. Los datos los puede obtener de la página con los recursos del libro *Applied Linear Models* que encuentra en la bibliografía del curso. En particualr, los datos los encuentra en: `http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData` en el apéndice C.2. Una descripción de los datos es la siguiente:

Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 years old or older
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI labor force that is unemployed
15	Per capita income	Per capita income of 1990 CDI population (dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE, 2 = NC, 3 = S, 4 = W

Source: Geospatial and Statistical Data Center, University of Virginia.

La variable de interés en el modelo es el número total de crímenes serios (Y).

1. Revise si hay datos influyentes en el modelo. ¿Puede explicar porqué son influyentes y si corresponden a errores de medición?

2. Se quiere estimar el efecto que tiene el desempleo sobre el número total de crímenes serios, dado que se cree que existe una relación de causalidad. En un modelo que considere bien especificado, encuentre el efecto que tiene el desempleo sobre Y. Justifique su respuesta, y explique por qué considera que su modelo está bien especificado. En particular, revise que no tenga variables omitidas importantes.

Parte C. Problemas aplicados con datos reales.

Problema 4. *Ventas de secadoras eléctricas en Bogotá.*
El uso de secadoras de ropa eléctricas en Colombia tiene poca incidencia.Sin embargo, en los últimos años las ventas han venido creciendo considerablemente. Para entender el modelo de ventas, un gran almacén de cadena se hizo un estudio en el cual quiere explicar de qué dependen las ventas por almacén. Como variables independientes uso: precio promedio, consumo de energía, gasto en publicidad y precio promedio de la compentencia en los almacenes más cercanos (de las cadenas de competencia directa). Los datos se encuentran en el archivo `secadoras.txt`

1. Revise todos los supuesto del modelo que sean pertinentes. Realice las pruebas que considere necesarias.

2. Encuentre una manera de solucionar el problema de las varianzas.

3. Un reconocido economista afirma que el efecto del precio es el mismo que el de la competencia pero son signo contrario, y además que el efecto del consumo de energía es cuatro veces que el

del gasto en publicidad con signo contrario. Cree Usted que los datos contradicen la afirmación del economista? Use al menos dos formas diferentes adecuadas para probar esta hipótesis.

Problema 5. *Efectos de medidas preventivas en la severidad del coronavirus.*
Una de las preguntas más comunes en política de salud pública en la actualidad, es determinar el efecto de las medidas de restricción de movilidad o de actividades que impliquen reducir el contacto entre personas. Esto es por ejemplo, las cuarentenas, el pico y cédula, la prohibición de clases presenciales, uso de tapabocas, etc. Por conveniencia, estas medidas han sido agrupadas en una sola métrica conocide com el *Stringency Index*. Para entender cómo se calcula, puede revisar los documentos relacionados en: `https://covidtracker.bsg.ox.ac.uk/`.
En un modelo simplificado, se quiere probar el efecto de esta medidas en Colombia (usando el *stringency index*) sobre la tasa de mortalidad usando el modelo:

$$\log(MR_t) = \beta_0 + \beta_1 SI_t + \beta_2 \log(NC_t) + \varepsilon_t$$

Donde MR_t corresponde a la tasa de mortalidad (número de muertes por Covid por cada millón de habitantes) en el día t, NC_t corresponde al número de casos confirmados (por cada millón de habitantes) durante los últimos 14 días antes de t, y SI_t corresponde al promedio del stringency index para los 14 días anteriores a t, es decir:

$$SI_t = \frac{\sum_{i=1}^{14} strind_{t-i}}{14}$$

$$NC_t = \sum_{i=1}^{14} nc_{t-i}$$

donde $strind_t$ es la medida de stringency index en el día t y nc_{t-i} es el número de casos confirmados en el día $t-1$. Para evitar problemas con los logaritmos, cuando las variables MR_t o NC_t tomen el valor de cero, entonces aproximelos como 1. Si considera que en la base da datos existen variables importantes para tener un modelo mejor específicado, no dude en usarlas.
Note que este modelo es definido sobre series de tiempo. Como periodo de estudie, use datos desde el noviembre 1 del 2020 hasta noviembre 1 de 2021. Para obtener los datos de las variables, puede revisar: `https://ourworldindata.org/grapher/covid-stringency-index`.

1. Estime los parámetros del modelo e interpréte los intervalos de confianza correspondientes. Tenga en cuenta los logaritmos. Encuentra coherencia en los resultados.

2. Considera que el modelo tiene problemas de heteroscedasticidad?

3. Dado que el modelo se stima con series de tiempo, es posibel que se presenten probelmas de autocorrelación. Revise si existe autocorrelación positiva. En caso que sea así, estime el modelo correcto y realice el intervalo de confianza para cada parámetro (incluyendo el intercepto).