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

TAREA 1

- π El informe con la solución de la tarea debe ser entregado antes del miércoles 7 de septiembre (5pm).
- π La solución puede ser entregada en grupos de máximo 3 personas.
- π Utilice procedimientos explícitos y justifique sus respuestas.

Parte A. Problemas Teóricos y Conceptuales.

Problema 1. Fundamentos de Inferencia Paramétrica.
Se tiene una muestra aleatoria $Y_1, Y_2, \ldots, Y_n$ de una población $Y \sim Normal(0,1)$. Demuestre:

1. $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim Normal(0, 1/n)$
2. $\sum_{i=1}^{n} Y_i^2 \sim \chi^2(n)$
3. $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2(n-1)$
4. $\frac{n\bar{Y}^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} \sim c \cdot F(1, n-1)$; con $c = ?$

Ayuda: Puede usar la siguiente información para las demostraciones:

- Si $Z \sim Normal(0,1)$, entonces $Z^2 \sim \chi^2(1)$
- Si $Z_1 \sim \chi^2(v_1)$ y $Z_2 \sim \chi^2(v_2)$, y Z_1 y Z_2 son independientes, entonces $(Z_1 + Z_2) \sim \chi^2(v_1 + v_2)$
- $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ y $\bar{Y}$ son independientes en el caso que $Y_i \sim_{i.i.d.} Normal$

Problema 2. Fundamentos distribuciones multivariadas: Propiedades.
Suponga que Z_1 y Z_2 son dos vectores Normales multivariados en p y k dimensiones respectivamente. Esto es,

$$Z_1 \sim Normal_p\left(\mu_1\ ,\ \Sigma_1\right)$$
$$Z_2 \sim Normal_k\left(\mu_2\ ,\ \Sigma_2\right)$$

La *Covarianza* entre estos dos vectores se define como una matriz de $p \times k$, la cual tiene como elementos las covarianzas correspondientes. Esto es:

$$Cov(Z_1, Z_2) = \left[cov\left(Z_{1i}, Z_{2j}\right)\right]_{ij}$$

Si esta matriz contiene solo ceros, entonces los vectores Z_1 y Z_2 son independientes. Con esta definición, se tiene que si se definen los nuevos vectores aleatorios:

$$W_1 = A_1 Z_1$$
$$W_2 = A_2 Z_2$$

Para matrices constantes A_1 y A_2, entonces:

$$Cov(W_1\ ,\ W_2) = A_1 Cov(Z_1\ ,\ Z_2) A_2^T$$

Recordemos que una requerimiento fundamental para construir intervalos de confianza o pruebas de hiótesis para los parámetros β, es que el estimador de la varianza $\hat{\sigma}^2$ y los estimadores de los parámetros de la recta $\hat{\beta}$ sean independientes. Para esto se requiere que $\hat{Y}$ y $Y - \hat{Y}$ sean vectores indeendientes. Usando las definiciones anteriores, muestre que este requerimiento se cumple.

Problema 3. Fundamentos de Inferencia Paramétrica.
Se tiene una muestra aleatoria $Y_1, Y_2, \ldots, Y_n$ de una población $Y \sim Normal(\mu, \sigma^2)$. El estimador convencional de la varianza es

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}.$$

Sin embargo, el estimador por máxima verosimilitud es $\hat{\sigma}^2 = \frac{(n-1)S^2}{n}$. Cuál de los dos es mejor estimador? Ayuda: Use el ECM como criterio.

Problema 4. Fundamentos de Inferencia Paramétrica.
Se tiene una muestra aleatoria $Y_1, Y_2, \ldots, Y_n$ de una población $Y \sim f_Y(y)$. Se sabe además que $\mu = \mathbb{E}(Y)$ y $\sigma^2 = \text{Var}(Y)$ son dos parámetros desconocidos (media y varianza de la población). Se define un estimador lineal insesgado de μ como:

$$\hat{\mu} = \sum_{i=1}^{n} a_i Y_i$$

donde $a_1, \ldots, a_n$ son constantes arbitrarias, y $\mathbb{E}(\hat{\mu}) = \mu$.

1. Muestre que $\sum_{i=1}^{n} a_i = 1$
2. Demuestre que de todos los posibles estimadores insesgados lineales para μ, $\bar{Y}$ es el que tiene menor varianza (BLUE).

Problema 5. Fundamentos del Modelo de Regresión Lineal.
En un modelo de regresión lineal con n observaciones y $k = p - 1$ variables, el estimador de mínimos cuadrados es $\hat{\beta} \in \mathbb{R}^p$, y $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} \in \mathbb{R}^n$ es la predicción en los puntos de la muestra. Sea $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_k) \in \mathbb{R}^{n \times p}$ la matriz de las variables independientes, donde $\mathbf{X}_0 = (1, 1, \ldots, 1)$.
Se define el subespacio de $\mathbb{R}^n$ de todas las posibles combinaciones lineales de $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_k$ como:

$$\Omega = \{\mathbf{X}\theta | \theta \in \mathbb{R}^p\}.$$

Por definición, Ω tiene dimesnión $p \leq n$.

1. Si $p = n$, que relación hay entre $\mathbf{Y}$ y $\hat{\mathbf{Y}}$?
2. Argumente porque $(\mathbf{Y} - \hat{\mathbf{Y}})$ debe ser ortogonal a cualquier $\omega \in \Omega$.

Ahora suponga que se hace una transformación lineal para crean nuevas varibales $\mathbf{Z}$ tal que la nueva matriz de datos es:

$$\mathbf{Z} = \mathbf{X}A$$

donde $A \in \mathbb{R}^{p \times p}$ es una matriz de rango completo.

3. Si se define

$$\Omega_z = \{\mathbf{Z}\theta | \theta \in \mathbb{R}^p\}$$

entonces, ¿qué relación existe entre $\hat{\mathbf{Y}}$ y la proyección de $\mathbf{Y}$ en Ω_Z?

4. Con el resultado del numeral anterior, y usando los datos `measure.txt`, que contienen mediciones de los contornos de pecho, cintura y cadera de 20 personas, use el contorno del pecho como la variable de respuesta (Y) y las otras dos como variables independientes (X_1 y X_2). Estime el modelo de regresión correspondiente.
   Ahora se quieren cambiar las variables (cintura y cadera) a las variables: tamaño (W_1: cintura más cadera) y forma (W_2: cintura menos cadera).

   - Estime el modelo en el que se explica Y a partir de W_1 y W_2 como $\mathbf{Y} = \mathbf{W}\theta + \varepsilon$. Encuentre los vectores $\hat{\mathbf{Y}}$ para cada modelo, con las variables X y otro con las variables W. Compare los dos vectores y concluya.
   - Construya un intervalo de confianza para el valor esperado del contorno del pecho cuando el tamaño es 60 y la forma es cero usando los dos modelos, es decir cuando la cintura y la cadera son respectivamente 30. Use cada modelo por separado y comparelos dos intervalos. Concluya sobre los resultados.

Problema 6. Fundamentos de Álgebra para Modelos Lineales.
(Proposición vista en clase). Sean dos vectores a, z en $\mathbb{R}^p$ y $\mathbf{M}$ una matriz en $\mathbb{R}^{p\times p}$. Si se definen las funciones:

$$\alpha_1 = a^T z \qquad \alpha_2 = z^T \mathbf{M} z$$

Las derivadas respectias de α_1 y α_2 con respecto a vector z (*Jacobiano*) se definen como:

$$\frac{d\alpha}{dz} = \begin{bmatrix} \frac{d\alpha}{dz_1} \\ \frac{d\alpha}{dz_2} \\ \vdots \\ \frac{d\alpha}{dz_p} \end{bmatrix}$$

Muestre que:

1. $\frac{d\alpha_1}{dz} = a$
2. $\frac{d\alpha_2}{dz} = (\mathbf{M} + \mathbf{M}^T)z$

Problema 7. Fundamentos de Inferencia Paramétrica.
Se tiene una muestra aleatoria $Y_1, Y_2, \ldots, Y_n$ de una población $Y \sim Exponential(\theta)$ con función de densidad $f(x) = \frac{1}{\theta}\exp^{\frac{-x}{\theta}}$, donde $\mathbb{E}(Y) = \theta$ y $\mathrm{Var}(Y) = \theta^2$.

1. Halle el estimador de máxima verosimilitud $\hat{\theta}$.
2. Comente sobre la doistribución de $\frac{\hat{\theta}-\theta}{\sqrt{\frac{\theta^2}{n}}}$ cuando n representa un tamaño de muestra grande.
3. Consulte sobre el teorema se Slutsky (`https://en.wikipedia.org/wiki/Slutsky\%27s_theorem`) y úselo junto con la infrmación del numeral anterior para construir un intervalo de confianza (1-α) para el parámetro θ.
4. Suponga que el número de entradas que ocurren al baño del 5 piso del edificio ML, se comportan como un proceso Poisson con tasa $\frac{1}{\theta}$. Suponga que en una semana, el promedio de entradas por hora es de 21.4 y que el día laboral tiene 12 horas. Construya un intervalo de confianza para la media del tiempo entre arribos con una confianza del 95%.

Problema 8. Proyecciones y distribuciones.
Se tiene un modelo de regresión lineal de forma tal que :

$$\mathbf{Y} \sim Normal_{40}\left(\mathbf{X}\beta, \sigma^2 \mathbb{I}_{40\times 40}\right).$$

Donde el número de variables explicativas (k) es 9. Se define la matrix de proyección $\mathbf{P}_\Omega = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$.

1. Explique porqué $(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbb{I} - \mathbf{P}_\Omega)(\mathbf{Y} - \mathbf{X}\beta) = \|\mathbf{Y} - \mathbf{P}_\Omega\mathbf{Y}\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$
2. Use el resultado anterior para encontrar la distribución de $\frac{1}{\sigma^2}\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$
3. Con un argumento similar al del literal (a), explique porqué si el modelo es significativo $\frac{1}{\sigma^2}\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2$ no se distribuye χ^2 con 39 grados de libertad.

Parte B. Problemas aplicados con datos reales.

Problema 9. *Carseats sales Data.*
Usando los datos `Carseats` contenidos en la libreria de `R ISLR`, se busca un modelo que explique las ventas por almacén `Sales` de sillas infantiles para carro dependiendo de diferentes variables (precio, precio de la competencia, edad promedio del vecindario, etc.). Para entender los datos use:

```
library(ISLR)
?Carseats
data=Carseats
head(data)
data=data[,-c(7,10,11)]  #remove non-continuous variables
head(data)
```

Usando solamente las variables continuas (las últimas dos lineas del código remueven las categóricas):

1. Estime un modelo de regresión lineal e interprete cada uno de los parámetros de la recta.
2. Es cierta la hipótesis que afirma que la edad promedio del vecindario (`Age`) no debe influir sobre la ventas de sillas?
3. Un analista cree que el efecto del precio, es el mismo que el de el precio de la competencia pero con signo negativo. Se puede contradecir al analista con los datos?
4. Es cierta la afirmación de que la edad, la población y la educación del vecindario no son variables que afecten el nivel de las ventas de sillas?.
5. Aunque en general que una variable sea significativa no implica causalidad en el cambio de Y (sólamente correlación), es adecuado asumir que las variables precio, precio de la competencia y gasto en publicidad si tienen un efecto directo en el cambio de las ventas de sillas. Suponga que la competencia baja el precio en \$10. La tienda reacciona bajando el precio en \$4 e invirtiendo \$5 en publicidad. Cree que estas acciones son suficientes para no disminuir las ventas de sillas en el largo plazo?
6. Si Usted quiere iniciar una nueva tienda que vende estas sillas, donde el precio de la competencia es 120, el ingreso es 95, el gasto en publicidad es 7, la población es 300, el precio es 100, la edad es 45 y la educación es 11; calcule in intervalo de confianza para el valor medio de ventas se sillas. Calcule tamibién el intervalo de predicción y compare los dos intervalos.

7. Otra opción para iniciar una tienda es en un vecindario donde donde el precio de la competencia es 100, el ingreso es 95, el gasto en publicidad es 8, la población es 400, el precio es 95, la edad es 40 y la educación es 10. Se venderán (en valor esperado) más sillas en esta nueva tienda que en la del numeral anterios? (Realice pruebas estadísticas).

8. Un reconocido economista afirma que el efecto del precio es el mismo que el de la competencia pero con signo contrario, y además que el efecto de la edad es 1.5 veces el de la educación. Cree Usted que los datos contradicen la afirmación del economista?

Problema 10. Problema Libre.
Escoja unos datos que estén relacionados a un problema inferencial de su preferencia (por ejemplo. afecta la cuarentena el contagio?, de qué depende el precio de la vivienda?, etc). Debe haber una variable de interés (Y) y al menos 3 variables explicativas ($k \geq 3$). Recuerde que para un problema de regresión, Y debe ser continua.

1. Haga una descripción de los datos y del problema general sobre el cual quiere inferir. Muestre algunas gráficas que resuman el comportamiento de los datos.

2. Realice las siguientes pruebas:

   - Significancia individual de cada variable
   - Significancia global del modelo
   - Significancia parcial del algunas variables
   - Una combinación lineal de betas (que tenga sentido dentro del contexto del problema)