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

TAREA 2

- π La solución de la tarea debe ser entregado antes del lunes 24 de octubre (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. Escriba su propia función para hacer pruebas sobre los betas.
Como ya lo vimos en clase, las pruebas F para restricciones en los parámetros *beta* de la recta incluyen la mayor parte de las pruebas que se pueden hacer sobre los modelos de regresión (significancia global, significancia de cada variable, combinaciones lineales, etc).
Como habrá visto, la función `lm` de R no tiene la opción de inferir sobre restricciones generales, sino que hace la prueba de significancia global y las individuales.
Lea el capítulo 10 del documento de introducción a R (`intro-R.pdf`) sobre como crear sus propias funciones, y con esto:

1. Cree su propia función para estimar un modelo de regresión lineal en donde los inputs son: matrix $\mathbf{X}$, vector $\mathbf{Y}$, el vector $\mathbf{c}$ y la matriz $\mathbf{A}$ (si se quiere hacer una prueba general con restricciones), y el vector de subíndices que se quieren incluir en la hipótesis nula de una prueba de significancia parcial. Puede usar como base las funciones vistas en clase, si lo prefiere.

2. La función debe tener una input binario en el que el usuario decide si quiere hacer una prueba con restricciones generales o una significancia parcial.

3. La función debe dar como salida: la prueba de significancia global, las significancias individuales de cada uno de los betas (con pruebas de hipótesis o intervalos de confianza) y el resultado de la prueba asociada a $\mathbf{A}$ o a los índices de los betas en la prueba parcial, incluyendo el *p-value*.

4. Póngale el nombre de su preferencia a la función, añádale los demás atributos que quiera y úsela a su conveniencia.

Problema 2. Visualizando las regiones de confianza.
Hasta ahora sabemos hacer intervalos de confianza asociados a cada uno de los parámetros β_j para $j = 0, 1, \ldots, k$. Sin embargo, estos están relacionados a pruebas individuales, con conjuntas.
Para solucionar esto, vamos a hacer regiones conjuntas de confianza para pares de parámetros. Para esto tenga en cuenta que cada par de estimadores $\hat{\beta}_i$ y $\hat{\beta}_j$ para $i, j \in \{0, \ldots, k\}$ se distribuyen como un vector normal bivariado (ervisar las propiedades de la distribución normal multivariada), Esto es:

$$(\hat{\beta}_i, \hat{\beta}_j)^T \sim Normal_2\left((\beta_i, \beta_j)^T, \sigma^2 \mathbf{W}_{ij}\right)$$

Donde la matriz $\mathbf{W}_{ij}$ corresponde a los elementos correspondientes a las varianza y covarianza, esto es:

$$\mathbf{W}_{ij} = \begin{bmatrix} h_{ii} & h_{ij} \\ h_{ij} & h_{jj} \end{bmatrix}$$

Por facilidad en la notación, llamemos $\hat{\beta}_{ij} = (\hat{\beta}_i, \hat{\beta}_j)^T$ y $\beta_{ij} = (\beta_i, \beta_j)^T$.
Para construir el la región de confianza siga los siguientes pasos:

1. Determine la distribución de
$$\sigma^{-1} \left(\mathbf{W}_{ij}\right)^{-\frac{1}{2}} \left[\hat{\beta}_{ij} - \beta_{ij}\right]$$

2. Determine la distribución de
$$\sigma^{-2} \left[\hat{\beta}_{ij} - \beta_{ij}\right]^T \mathbf{W}_{ij}^{-1} \left[\hat{\beta}_{ij} - \beta_{ij}\right]$$

3. Como la varianza de error (σ^2) no se conoce, determine la distribución de:
$$\left(\frac{1}{2 \cdot MSE}\right) \left[\hat{\beta}_{ij} - \beta_{ij}\right]^T \mathbf{W}_{ij}^{-1} \left[\hat{\beta}_{ij} - \beta_{ij}\right]$$

4. Use la expresión del numeral anterior para definir una región (en este caso una elipse) que forma una región bidimensional con una confianza de $1 - \alpha$.

5. Investigue la función `contour` de `R` (`https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/contour.html`), y úsela para graficar la región (elipse) de confianza dado los valores $\hat{\beta}_{ij}$, $\mathbf{W}_{ij}$ y el MSE. Puede usar cualquier otra función que considere pertinente para graficar

6. Cree una función donde los inputs son los datos $\mathbf{X}$, $\mathbf{Y}$, el nivel de significancia (α) y el par de parámetros betas (por ejemplo β_1 y β_3). Como resultado de la función, se debe producir un gráfico con la elipse de confianza para los dos parámetros seleccionados.

Parte B. Problemas Teóricos y Conceptuales.

Problema 3. Formas de variables *dummies*.
Para modelar una variable categórica con m clases, debido a que se crea un problema de multicolinealidad perfecta (modelo no estimable), decidimos utilizar una clase base e incluir en el modelo de regresión $m - 1$ variables *dummies* de la forma

$$Z_j = \begin{cases} 1 & \text{clase j} \\ 0 & \text{otra clase} \end{cases} \qquad j = 1, 2, \ldots, m-1$$

que en un modelo con una variable continua y sin interacciones queda como:

$$y = \beta_0 + \beta_1 X + \gamma_1 Z_1 + \gamma_2 Z_2 + \cdots + \gamma_{m-1} Z_{m-1} + \varepsilon.$$

En este caso, los parámetros γ_j se interpretan como los cambios en el intercepto de la clase j con respecto a la clase base. Esto implica que todas la interpretaciones son con respecto a la clase base.
Otra forma de resolver este problema (que no vimos en clase) es poner otro tipo de restricciones, por ejemplo, la restricción de suma cero. Esto es, supongamos que definimos los párámetros τ_j como los efectos de estar en la clase j, donde el intercepto para cada clase se define como:

$$\beta_{0j} = \beta_0 + \tau_j$$

donde $\beta_0 = \frac{\sum_{j=1}^m n_j \beta_{0j}}{\sum_{j=1}^m n_j}$, siendo n_j el número de datos en cada clase j. Esto es, β_0 puede ser interpretado con como el intercepto general del modelo (promedio de interceptos). Para usar esta parametrización, se usa la restricción de que:

$$\sum_{j=1}^{m} n_j \tau_j = 0$$

Para lo cual, se usa una clase base (por ejemplo al m) en la cual se puede modelar como $\tau_m = -\sum_{j=1}^{m-1} \frac{n_j}{n_m}\tau_j$ y con esto se cambian las variables dummies. Para más información, busque la sección "*Factor Effects Model with Weighted Mean*" en el capítulo 16 del libro *Applied Linear Models* de Kutner, et al.

Usando los datos Carseats, se usan como variables explicativas continuas Advertising e Income, además de la variable categórica ShelveLoc. Usando la parametrización con restricción de suma cero, y considerando un modelo con todas las dummies y todas las interacciones:

1. Es cierto que la posición en el estante no influye sobre las ventas?
2. Es cierto que los modelos en todas las clases son paralelos (los efectos no cambian)?
3. Son los modelos para las ventas posición media en el estante, y para la posición buena iguales?
4. Es cierto que el efecto de la publicidad es igual en la posición mala que en la media?

Parte C. Problemas aplicados con datos reales.

Problema 4. Modelos No-Lineales
La base de datos `productivity.txt` contiene la variable de respuesta *productividad* que mide la productividad de trabajadores de construcción (número de tareas manuales por hora) en términos de los la temperatura ambiental ($x1$) y el tiempo (en días) que llevan haciendo la terea repetitiva en las mismas condiciones (lugar y tipo de tarea) ($x2$).

1. Estime un modelo de regresión lineal y concluya sobre la influencia de los dos factores en mención.
2. Se cree que el modelo es no-lineal, por lo cual es mejor usar un modelo polinomial de orden 2. Cree que el modelo resultante debe ser aditivo?
3. Si se sabe que la temperatura ambiental es de 2, estime el efecto que tiene el tiempo que se lleva haciendo la tarea (cuando este es 2.5) sobre la productividad.
4. Si en un día particuale la temperatura ambiental es de 2.5 y y un trabajador llava 3 días haciendo la misma labor en el mismo sitio, construya un intervalo de predicción para la productividad del trabajador.

Problema 5. *Efectos de la personalidad en la producción intelectual bajo estrés.*
En la base de datos adjunta `personality_scores.txt` ecnuentra datos relacionados a un experimento en el cual se le pide a un grupo de personas que resuelva problemas lógicos, metemáticos y de comprensión general, bajo condiciones de estrés (por ejemplo, con distracciones o presión por resolverlas de manera rápida). La variable `score` es el puntaje de cada persona.
La personalidad de cada participante es medida usando el método de las cinco grandes (o método OCEAN), en la cual, se consideran cinco factores representatitocs de la personalidad: apertura a la experiencia (O), responsabilidad (C), extraversión (E), agradable (A) y neuroticismo (N). Para más información sobre este método de evaliación de la personalidad, puede consultar `en.wikipedia.org/wiki/Big_Five_personality_traits`. Los cinco factores deben ser calculados a través de un cuestionario de 44 preguntas, cada una de ellas relacionada con un factor de personalidad. Algunas de estas preguntas son positivas (suman) y otras son negativas o en reverso (restan). En el documento adjunto `big5_questions.pdf` encuentra la explicación de cómo compilar estos cuestionarios.
Además de la personalidad, se encuentra la variable `satisfaction` en la cual se le pide al participante una evaluación general sobre la satisfaccción con el trabajo en el cual se desempeña.

1. Dado que la base de datos posee algunos datos faltantes (NA), impute estos datos con los promedios de la respectiva variable.

2. Compile el cuestionario para encontrar los cinco factores de personalidad de cada individuo.

Ahora, con los cinco factores calculados, más la variable de satisfacción se intentarán demostrar ciertas hipótesis psicológicas. Dado que se intuye que el efceto de algunos variables no es lineal, es mejor considerar un modelo polinomial (cuadrático). Pruebe las siguientes hipótesis:

2. Es cierto que el efecto de C sobre el desempeño depende (interactúa) con la satisfacción?

3. Es cierto que A y O no deben influir sobre el desempeño?

4. Se cree que para niveles bajos y altos de C el desempeño es alto, pero para niveles promedio de C, el desempeño es menor. Ayuda: Use los términos cuadráticos.

5. En un modelo que considera apropiado, el efecto de E es lineal.

6. Con un modelo apropiado, resuma en palabras cuál es el efecto que tiene la personalidad sobre la producción intelectual bajo estrés.

Para que tenga alguna guía de cómo se hacen estos estudios (aunque con otros modelos), se adjunta un artículo de referencia (`team_personality_productivity.pdf`).

Problema 6. *Interacciones y multicolinealidad.*
En la base de datos "mtcars", se usan las variables continuas: disp (desplazamiento), hp (potencia), y wt (peso del carro). Además se usan las siguientes variables categóricas: dispocición de los cilindrios (variable vs) y tipo de transmisión (variable am). Use:

```
library(ISLR)
data=mtcars
head(data)
data$vs=as.factor(data$vs)
data$am=as.factor(data$am)
```

En un modelo que contenga las tres variables continuas y las variable categóricas (con todas las interacciones entre continuas y dummies):

1. Influye el tipo de caja en los modelos?

2. Vale la pena considerar las interacciones en los modelos, o son todos paralelos?

3. Es igual el efecto de la variable wt en el consumo de combustible para la caja manual con motor en línea que para caja automática con motor en v?

4. Es igual el modelo para la caja manual con motor en línea que para caja automática con motor en v?

5. Un vendedor ofrece dos modelos. El carro 1 tiene disp=20, hp=150, wt=1500, caja manual y motor en v. El carro 2 tiene disp=17, hp=100, wt=2600, caja automática y motor en línea. El vendedor afirma que los consumos medios de combustible son respectivamente 18 y 21. Es cierta la afirmación del vendedor?

6. Al parecer las pruebas tienden a no dar significativas en los puntos anteriores. Pruebe que el modelo tiene problemas de multicolinealidad, y explique qué relación tiene esto con las interacciones que se inluyen en el modelo.

7. Corra el modelo sin intereacciones y determine si siguen existiendo problemas de multicolinealidad.