Tarea 2, versión B. Regresión lineal múltiple

Gonzalo Pérez, Sheyla Barradas y Luis Genaro Coria

Semestre 2024-2

La tarea se deberá subir al classroom antes de las 11:59 PM del 17 de mayo de 2024. El examen presencial es el día 24 de mayo.

Favor de argumentar con detalle las respuestas.

NOTA. En caso de que se identifiquen respuestas iguales en otras tareas, se procederá a la anulación de las tareas involucrados.

NOTA. Incluir el(los) nombre(s) completo(s) de la(s) persona(s) que está(n) resolviendo los ejercicios. Equipos de máximo tres integrantes.

Usar una confianza de 95 % o una significancia de .05 en los casos en donde no se requiera otro nivel de forma explícita. En el caso de realizar alguna transformación de las variables, se tiene que hacer explícita la variable que se usa y la interpretación en las pruebas de hipótesis o intervalos de confianza.

1. (1.5 puntos)

Considere el modelo de regresión

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon,$$

y los estimadores obtenidos por mínimos cuadrados escritos en forma matricial

$$\widehat{\boldsymbol{\beta}} = (X^t X)^{-1} X^t \boldsymbol{y}.$$

Usando la matriz proyección $\boldsymbol{H}$ y sus propiedades, indique a qué es igual

- $\boldsymbol{e}^t \boldsymbol{X}$, donde $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ y $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.
- $Cov(\boldsymbol{e}, \widehat{\boldsymbol{y}})$.

2. (1.5 puntos)

Considere el modelo de regresión siguiente:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (3x_i^2 - 2) + \epsilon_i, \quad i = 1, 2, 3,$$

donde $x_1 = -1, x_2 = 0, x_3 = 1$.

I. Defina la matrix diseño $\boldsymbol{X}$ asociada a este modelo. Calcule $\boldsymbol{X}^t\boldsymbol{X}$ y su inversa.
II. Dé las expresiones de los estimadores por mínimos cuadrados ordinarios de β_0, β_1 y β_2: $\widehat{\beta}_0$, $\widehat{\beta}_1$ y $\widehat{\beta}_2$. Deberán ser expresiones en términos de y_1, y_2, y_3.
III. Muestre que los estimadores por mínimos cuadrados ordinarios del modelo reducido cuando se supone $\beta_2 = 0$ no se alteran, es decir, que $\widehat{\beta}_0^* = \widehat{\beta}_0$ y $\widehat{\beta}_1^* = \widehat{\beta}_1$, donde $\widehat{\beta}_0^*$ y $\widehat{\beta}_1^*$ son los estimadores por mínimos cuadrados del modelo

$$y_i = \beta_0^* + \beta_1^* x_i + \epsilon_i^*, \quad i = 1, 2, 3.$$

3. (2 puntos)

La Compañía Kenton Food desea comparar 4 diferentes diseños de empaque de un nuevo cereal. Veinte tiendas, con aproximadamente igual volumen de ventas y perfil de clientes, fueron seleccionadas como unidades experimentales. A cada una de las tiendas se le asignó uno de los empaques de forma aleatoria, de manera que cada empaque fuera asignado a 5 tiendas distintas. Las ventas, en número de casos, fueron observadas durante un período de estudio de 2 semanas:

ventas	12	10	15	17	11	11	17	16	14	15	27	34	22	26	28	23	20	18	17
empaque	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4

Un incendio ocurrió en una de las tiendas durante el período de estudio y dado que esto cambia las condiciones de venta con respecto a las otras tiendas se decidió eliminar la medición de esa tienda. El número de ventas de esa tienda se excluye de la tabla anterior.

Asuma que se cumplen todos los supuestos de un problema tipo ANOVA.

I. Presente un gráfico para describir los datos, por ejemplo, un boxplot por tipo de empaque.
II. Ajuste un modelo de regresión lineal múltiple adecuado. Indique de acuerdo a los parámetros del modelo la expresión del número de ventas promedio por cada tipo de empaque y dé estimaciones puntuales.
III. Escriba las hipótesis que se contrastan con la prueba F asociada a la tabla ANOVA, ejecute ésta e interprete. Use $\alpha = .05$.
IV. ¿Se puede considerar que el diseño del empaque afecta las ventas promedio? Use $\alpha = .05$. Argumente indicando con claridad qué hipótesis se están contrastando en términos de los parámetros del modelo ajustado.
V. Realicé la prueba de hipótesis simultánea asociada a la igualdad de las ventas promedio entre todos los posibles pares de diferentes empaques. Use $\alpha = .05$. Interprete los resultados.
VI. Suponga que los ejecutivos de la empresa tienen la sospecha de que el diseño de empaque 3 es el que aumenta las ventas en comparación con el resto de empaques. Realice una prueba de hipótesis para argumentar en favor o en contra de esta hipótesis de acuerdo con los datos observados. Use $\alpha = .05$

4. (2.5 puntos)

Una institución de investigación realiza un estudio para analizar los efectos de un nuevo tratamiento para controlar los niveles altos de ansiedad. Para eso consideran un puntaje (a mayor valor mayores niveles de ansiedad) y definen un conjunto experimental con 120 individuos que en ese puntaje presentan valores similares al inicio del estudio, 60 son hombres y 60 mujeres. En el mercado se sabe que hay otro tratamiento que se usa comúnmente para este fin, de manera que de forma aleatoria han divido a los 120 individuos en tres grupos: 40 a los que no se aplicó ningún tratamiento (control), 40 a los que se aplicó el tratamiento actual (Trat1) y 40 a los que se aplicó el nuevo tratamiento (Trat2); 20 hombres y 20 mujeres en cada grupo. Los datos se presentan en el archivo *Ex4B.csv*.

Los investigadores sospechan que para el nuevo tratamiento podría existir un efecto diferenciado de acuerdo con el sexo, por lo que consideran conveniente incluir esta variable en el análisis.

(Para este ejercicio no se requiere verificar supuestos del modelo, asuma que se cumplen)

I. Realice un análisis descriptivo de los datos. Dado que las dos covariables son categóricas, incluya un boxplot para cada posible combinación de niveles que se pueden observar en esas dos variables categóricas (*boxplot(Puntaje~Trat+Sexo, ...)*). Comente lo que observe.
II. Considerando un modelo de regresión que incluye las dos variables categóricas de forma individual y también su interacción, dé la expresión del puntaje promedio para cada valor de las variables categóricas, es decir: $E(puntaje; Trat = k, Sexo = l)$, con $k \in \{Control, Trat1, Trat2\}$ y $l \in \{Hombre, Mujer\}$; así como la estimación puntual correspondiente.

III. Escriba las hipótesis que se contrastan con la tabla ANOVA, calcule ésta e interprete. Use $\alpha = .05$.
IV. ¿Se puede considerar que el sexo tiene un efecto en el puntaje, es decir, al menos para un tratamiento existe un efecto diferenciado en el puntaje derivado del sexo de los individuos? Use una prueba F con $\alpha = .025$. Interprete.

- **Hint: aquí** $H_0 : E(puntaje; Trat = k, Sexo = Hombre) = E(puntaje; Trat = k, Sexo = Mujer) \; \forall \; k \in \{Control, Trat1, Trat2\}$**.**
- En caso de no rechazar H_0, considere el modelo reducido eliminando la variable Sexo; pero si se rechaza H_0, considere una prueba simultánea que ayude a identificar para qué tratamiento se puede considerar que el sexo tiene un efecto, con los resultados de esa prueba reduza el modelo si es posible.

V. En caso de que en el inciso anterior se haya reducido el modelo, ajuste de nuevo la regresión y dé la expresión del puntaje promedio para cada valor en las variables categóricas: $E(puntaje; Trat = k, Sexo = l)$, con $k \in \{Control, Trat1, Trat2\}$ y $l \in \{Hombre, Mujer\}$; así como estimaciones puntuales.
VI. Realice una prueba de hipótesis para argumentar en favor o en contra de la hipótesis: *el nuevo tratamiento tiene el mejor desempeño.* Use $\alpha = .05$
VII. Realice una prueba de hipótesis para argumentar en favor o en contra de la hipótesis: *el nuevo tratamiento tiene el mejor desempeño en mujeres, aunque el tratamiento actual lo tiene en hombres.* Use $\alpha = .05$

Nota. Suponga que tiene dos variables categóricas, una con K niveles y la otra con J niveles. Para usar el modelo de regresión con interacciones se requiere incluir $K - 1$ y $J - 1$ variables binarias asociadas a los efectos principales de los niveles, además de $(K-1) \times (J-1)$ variables binarias asociadas a las interacciones. Las variables de las interacciones se construyen como el producto de las $K - 1$ y $J - 1$ variables binarias que se introducen en el modelo.

5. (2.5 puntos)

Suponga que una empresa farmacéutica está ofreciendo al gobierno un nuevo medicamento para tratar a pacientes con la enfermedad Covid-19. El costo del medicamento es considerable y para tomar una buena decisión se han acercado a usted para analizar los datos que ha compartido la empresa farmacéutica. El archivo Ex5.csv contiene la información: *Ant* es el número total de anticuerpos, $Trat$ es una variable con dos niveles dependiendo si se aplicó o no el nuevo medicamento. Se sabe que tener mayores anticuerpos evita que se desarrolle una versión grave de la enfermedad y la empresa afirma que eso se logra al aplicar el medicamento, pues los pacientes que recibieron el medicamento tienen más anticuerpos que los que sólo recibieron placebo. También se sabe que la generación de anticuerpos es diferente dependiendo de la edad de los individuos y se sospecha que eso también podría afectar la efectividad del medicamento, así que al diseñar el experimento se seleccionaron al azar 100 personas de 300 que presentaban síntomas leves al iniciar el cuadro de la enfermedad a los que se les administró el medicamento, al resto se les dió sólo seguimiento. En todos los pacientes se capturó la edad y se procuró tener pacientes en el rango entre 16 y 60 años en ambos grupos. No se sospecha de otro aspecto que pudiera modificar la evaluación del medicamento.

(para este ejercicio no se requiere verificar supuestos del modelo, asuma que se cumplen)

I. Realice un análisis descriptivo de los datos considerando tanto la información de la edad como de la administración o no del medicamento.
II. Ajuste un modelo adecuado para evaluar la efectividad del medicamento ajustando por la edad de los pacientes. Es decir, un modelo que incluya como explicativas las variables edad, la binaria asociada a la administración del medicamento y la interacción obtenida como el producto de estas dos.
III. De acuerdo con el modelo ajustado, indique las expresiones asociadas a la relación de la generación promedio de anticuerpos con la edad en a) el grupo control y b) en el grupo que recibe el medicamento.
IV. ¿Se puede decir que la edad afecta de la misma forma la generación de anticuerpos en el grupo control que en el grupo que recibe el medicamento? Realice una prueba de hipótesis apropiada e interprete.

v. Comente sobre el ajuste del modelo incluyendo la interpretación de cada uno de los coeficientes.
vi. Argumente en contra o a favor de la afirmación: "El medicamente funciona aumentando el número de anticuerpos para todos los pacientes entre 25 y 60 años". Se puede apoyar de pruebas de hipótesis o intervalos de confianza simultáneos.

6. (2 puntos)

Considere los datos del archivo *Ex6.csv*.

i. Considere un modelo de regresión lineal múltiple con las covariables X_1 a X_6 sin ninguna interacción. La variable dependiente es Y. Verifique los supuestos del modelo, en la parte de linealidad considere un análisis de forma global y para cada covariable. En caso de que alguno no se cumpla, realice transformaciones convenientes hasta que obtenga un modelo que parezca cumplir con los supuestos del modelo de regresión lineal múltiple.
ii. Con las variables transformadas en el inciso anterior, realice una selección de variables. Justifique su respuesta.