Regresión Lineal Múltiple: Variables Categóricas

Clase 21 y 22

Nicolás Mejía M.
n.mejia10@uniandes.edu.co

Probabilidad y Estadística II
Departamento de Ingeniería Industrial
Universidad de Los Andes, Bogotá, Colombia

2020-19

Outline

Motivación

Hasta el momento, todas las variables que hemos incluido en el modelo de regresión han sido variables continuas. Esto no siempre ocurre! Algunas son variables categóricas.

Ejemplo

- Sexo {Hombre, Mujer}
- Estrato socioeconomico {Estrato 1, ... , Estrato 6}
- Localidad {Usaquen, Fontibon, Bosa, etc}

La pregunta es ¿cómo incluir esto en el modelo de regresión?

Motivación

El problema con este tipo de variables, es que son variables cualitativas y no cuantitativas, luego no es posible asignar una relación como la que se ha estado planteando hasta el momento. Para esto debemos crear alguna forma de representar numéricamente las variables. La idea natural seria asignar un numero a cada categoría de la forma:

Ejemplo

$$\begin{aligned} \text{Usaquén} &\Leftrightarrow 1 \\ \text{Fontibón} &\Leftrightarrow 2 \\ \text{Bosa} &\Leftrightarrow 3 \\ &\vdots \end{aligned}$$

Motivación

De esta forma se crearía una variable numérica Z para incluir en la regresión

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

¡Pero la verdad esto no tiene sentido!

- La asignación 1,2,3 etc fue arbitraria. Se podríamos cambiar el orden de la enumeración, y daría otra cosa.
- La asignación 1,2,3 etc fue arbitraria. Se podríamos cambiar el orden de la enumeración, y daría otra cosa.
- Es ilógico plantear una relación lineal sobre una asignación numérica sin sentido

¿Que hacer entonces?

Outline

Variables Dummies

La estrategia para incluir la estructura de categorías consiste en crear variables que indiquen si el individuo pertenece a una categoría o no:

Ejemplo

Considere el ejemplo del sexo ({Hombre, Mujer}). Ahora definamos las variables indicadoras:

$$Z_1 = \begin{cases} 1 \text{ Si es Hombre} \\ 0 \text{ en caso contrario} \end{cases} \qquad Z_2 = \begin{cases} 1 \text{ Si es Mujer} \\ 0 \text{ en caso contrario} \end{cases}$$

Estas variables indicadoras reciben el nombre de variables dummies

Variables Dummies

Creando estas variables indicadoras,a se puede plantear el modelo de regresión:

$$Y = \beta_0 + \beta_1 X + \gamma_1 Z_1 + \gamma_2 Z_2 + \epsilon$$

de esta forma, es posible tener efectos distintos para cada categoría.

Pero hay un problema terrible!

¡ESTE MODELO NO PUEDE ESTIMARSE!

Outline

Dummie Trap

Ejemplo

Suponga que para estimar el modelo de regresión anterior tiene un muestra de 2 hombres y 3 mujeres. La matriz de diseño $\mathbf{X}$ sería:

$$\mathbf{X} = \begin{bmatrix} 1 & 10 & 1 & 0 \\ 1 & 7 & 1 & 0 \\ 1 & 5 & 0 & 1 \\ 1 & 9 & 0 & 1 \\ 1 & 3 & 0 & 1 \end{bmatrix}$$

Observe la primera, tercera y cuarta columna. ¿Que relación tienen?

La matriz de diseño tiene columnas linealmente dependientes, por lo tanto $\mathbf{X}^T\mathbf{X}$ NO ES INVERTIBLE

Dummie Trap

¿Que hacemos entonces?

Si recuerdan álgebra lineal, entonces recordarán que la dependencia lineal está asociada a la redundancia de información. Esto sucede porque podemos inferir la categoría sin conocer todas las variables Z.

Ejemplo

En el caso del sexo, la redundancia es clarísima pues $Z_1 = 1 - Z_2$. Como solo hay dos posibles opciones: hombre o mujer, entonces si la variable Z_1 toma el valor de 1, sabemos que es hombre, mientras que si toma el valor de 0, es un mujer. La variable Z_2 sobra.

Con esto en mente, podemos omitir una de las variables dummies sin repercusiones, y así solucionar la no invertibilidad de la matriz

Dummie Trap

La pregunta ahora es ¿cual omitir?

Ejemplo

Ya se mostró que que Z_1 es suficiente para recuperar la información. Si hacemos la misma lógica, tendremos que Z_2 también es suficiente para recuperar toda la información.

La verdad es que se puede omitir cualquiera de las categoría sin ningún problema. Más aún, podríamos quitar el intercepto. El único cambio es en la interpretación.

La variable que se deja omitida se le conoce como la variable base, y todas las interpretaciones se harán con respecto a esta.

Outline

Interpretación

Para facilidad, asuma que se omite la variable Z_2. De esta forma se estima el modelo de regresión:

$$Y = \beta_0 + \beta_1 X + \gamma_1 Z_1 + \epsilon$$

Reemplazando Z_1 por 0 y 1 según corresponda, se tiene que:

$$\begin{cases} Y = (\beta_0 + \gamma_1) + \beta_1 X + \epsilon \Leftrightarrow Z_1 = 1 \ \text{Hombre} \\ Y = \beta_0 + \beta_1 X + \epsilon \Leftrightarrow Z_1 = 0 \ \text{Mujer} \end{cases}$$

Observe que β_0 es el intercepto para las mujeres (la base), mientras que $\beta_0 + \gamma_1$ es el intercepto de los hombres. El coeficiente γ_1 es el efecto adicional sobre Y que se dan en el intercepto por ser hombre, en comparación a ser una mujer.

Interpretación

En ese caso, si el coeficiente γ_1 es igual a 0, el intercepto entre ambas categorías sería igual, luego no es necesario hacer dicha distinción. De lo contrario, existe una diferencia significativa y es necesario diferencias las categorías.

Bajo esa formulación, se observa que la pendiente de las rectas de cada categoría son iguales. Es decir, el efecto de X sobre Y es igual para ambas categorías. ¡Esto no necesariamente es el caso!

Para esto se deben considerar interacciones entre la variable X y las variables dicotómicas.

Outline

Interacciones

Considere el modelo de regresión dado por:

$$Y = \beta_0 + \beta_1 X + \gamma_1 Z_1 + \eta_1 (X * Z_1) + \epsilon$$

Donde $(X * Z_1)$ es el producto componente a componente de la variable X con la variables Z_1. A este término lo denominamos interacción entre X y la variable categórica. Para el ejemplo. la matriz de diseño sería:

$$\mathbf{X} = \begin{bmatrix} 1 & 10 & 1 & 10 \\ 1 & 7 & 1 & 7 \\ 1 & 5 & 0 & 0 \\ 1 & 9 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix}$$

Interacciones

Considere el modelo de regresión dado por:

$$Y = \beta_0 + \beta_1 X + \gamma_1 Z_1 + \eta_1 (X * Z_1) + \epsilon$$

Reemplazando Z_1 por 0 y 1 según corresponda, se tiene que:

$$\begin{cases} Y = (\beta_0 + \gamma_1) + (\beta_1 + \eta_1)X + \epsilon \Leftrightarrow Z_1 = 1 \ \text{Hombre} \\ Y = \beta_0 + \beta_1 X + \epsilon \Leftrightarrow Z_1 = 0 \ \text{Mujer} \end{cases}$$

Además de lo mencionado para intercepto, observe ahora que β_1 es la pendiente de X para las mujeres (la base), mientras que $\beta_1 + \eta_1$ es la pendiente de X para los hombres.

El coeficiente η_1 es el efecto adicional sobre Y que hace X por ser hombre, en comparación al ser una mujer.

Interpretación

Si el coeficiente η_1 es igual a 0, la pendiente entre ambas categorías sería iguañ, luego no es necesario hacer dicha distinción

Prueba de Interés

Vale la pena diferencias entre hombres y mujeres?

$$H_0 : \gamma_1 = \eta_1 = 0$$

Es el efecto de X sobre Y igual entre hombres y mujeres?

$$H_0 : \eta_1 = 0$$

Hay diferencia entre los interceptos de hombres y mujeres?

$$H_0 : \gamma_1 = 0$$

Outline

Varias Variables Categóricas

Puede que tengamos más de una variable categórica. El tratamiento es el mismo, solo que ahora se tendrán varios grupos de dummmies para cada variable categórica.

Ejemplo

Suponga que tiene las variables categóricas: Sexo {Hombre, Mujer} y Estatura {Baja, Promedio,Alta}. Las dummies serían:

$$H = \begin{cases} 1 \text{ Si es Hombre} \\ 0 \text{ d.l.c.} \end{cases} \qquad M = \begin{cases} 1 \text{ Si es Mujer} \\ 0 \text{ d.l.c.} \end{cases}$$

$$B = \begin{cases} 1 \text{ Si es Bajo} \\ 0 \text{ d.l.c.} \end{cases} \qquad P = \begin{cases} 1 \text{ Si es Promedio} \\ 0 \text{ d.l.c.} \end{cases} \qquad A = \begin{cases} 1 \text{ Si es Alto} \\ 0 \text{ d.l.c.} \end{cases}$$

Varias Variables Categóricas

Al igual que el caso anterior, no se necesitan todas las dummies de una variable categórica: con saber el valor de todas las dummies, salvo una, podemos hallar el valor de la restante. Luego podemos quitar una de las dummies para cada uno de los grupos de variables categóricas.

La elección es arbitraria. Las categorías que se quiten son la base.

$$H = \begin{cases} 1 \text{ Si es Hombre} \\ 0 \text{ d.l.c.} \end{cases}$$

$$B = \begin{cases} 1 \text{ Si es Bajo} \\ 0 \text{ d.l.c.} \end{cases} \qquad P = \begin{cases} 1 \text{ Si es Promedio} \\ 0 \text{ d.l.c.} \end{cases}$$

Varias Variables Categóricas

Considere el modelo de regresión dado ahora por:

$$Y = \beta_0 + \beta_1 X + \gamma_1 H + \eta_1 (X * H) + \gamma_2 B + \eta_2 (X * B) + \gamma_3 P + \eta_3 (X * P) + \epsilon$$

Recuerde que las interacciones permiten evidenciar efectos diferentes de la variable continua. Para el ejemplo, con 3 hombres y 3 mujeres, siendo 2 altos, 2 bajos y 2 promedios, la matriz de diseño sería:

$$\mathbf{X} = \begin{bmatrix} 1 & 10 & 1 & 10 & 1 & 0 & 10 & 0 \\ 1 & 7 & 1 & 7 & 1 & 0 & 7 & 0 \\ 1 & 5 & 0 & 0 & 0 & 1 & 0 & 5 \\ 1 & 9 & 0 & 0 & 0 & 1 & 0 & 9 \\ 1 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 6 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Varias Variables Categóricas

Considere el modelo de regresión dado ahora por:

$$Y = \beta_0 + \beta_1 X + \gamma_1 H + \eta_1 (X * H) + \gamma_2 B + \eta_2 (X * B) + \gamma_3 P + \eta_3 (X * P) + \epsilon$$

Reemplazando los valores por las dummies se tiene qué:

	Hombre ($H = 1$)	Mujer ($H = 0$)
Alto ($B = 0, P = 0$)	$(\beta_0 + \gamma_1) + (\beta_1 + \eta_1)X + \epsilon$	$\beta_0 + \beta_1 X + \epsilon$
Promedio ($B = 0, P = 1$)	$(\beta_0 + \gamma_1 + \gamma_3) + (\beta_1 + \eta_1 + \eta_3)X + \epsilon$	$(\beta_0 + \gamma_3) + (\beta_1 + \eta_3)X + \epsilon$
Bajo ($B = 1, P = 0$)	$(\beta_0 + \gamma_1 + \gamma_2) + (\beta_1 + \eta_1 + \eta_2)X + \epsilon$	$(\beta_0 + \gamma_2) + (\beta_1 + \eta_2)X + \epsilon$

La interpretación y preguntas de interés son iguales que antes.

Outline

Otros Detalles

- Recuerde, al cambiar la base, se estima otro modelo de regresión, pero los resultados numéricos de predicciones y pruebas de hipótesis SON IGUALES.
- Si bien parece que agregar las interacciones es útil, realmente lo mejor es evitar agregar demasiadas. Solo las que sean necesarias !
- Las expresiones deducidas en clases pasadas de intervalos de confianza y demás procedimientos estadísticos aplican igual. Solo basta reemplazar el valor de X por el que corresponde.
- Importante, quitar variables dummies NO TIENE SENTIDO. Lo que tiene sentido es o no quitar todo el grupo de dummies de la categoría.
- La codificación 1 y 0 puede ser diferente! lo importante es entender el significado detrás para ver los modelos para cada categoría.

Outline

Ejemplo

Para determinar el número de espectadores que tendrá una película en el primer mes de lanzamiento, se tiene en cuenta la duración de la película y el género y se estima un modelo de regresión

$$E = \beta_0 + \beta_1 DP + \beta_2 Género + \epsilon$$

Donde:

- E: Número de espectadores (Millones)
- Dp: Duración de la película (Horas)
- Género: Género de la película Terror (T) Drama (D) Comedia (C)

Ejemplo

Despues de tomar una muestra de 50 películas se obtiene los siguientes modelos:

$$\hat{E} = 950.78 - 37.9DP + 237.95C - 123.6D$$

a. Encuentre los modelos auxiliares para cuando la película es de terror, comedia y drama.

b. Si ahora se utiliza como base el género de comedia, ¿cómo cambian los estimadores? y ¿si la base es drama?

c. Si se desean incluir todas las variables de género de la película ¿Qué forma tendría el modelo y cuales serían sus coeficientes?

Ejemplo

Maria Alejandra y Nicolás acaban de comprar un nuevo apartamento. Después de amoblarlo y pasar todas sus cosas, se sentían muy solos en él. Por esta razón decidieron adoptar un par de mascotas: Tobías, (un perrito) y Maracuyá (una gatica). Ambos animalitos se llevan muy bien, pero al ser tan pequeños suelen ser muy traviesos y mientras se la pasan jugando, ya han dañado una gran cantidad de muebles en la casa. A la pareja le preocupa que, al crecer, las mascotas sigan haciendo daños, así que deciden estudiar el patrón de comportamiento de los dos.

Ejemplo

Para esto, estiman el siguiente modelo de regresión lineal:

$$Daños = \beta_0 + \beta_1 Tobías + \beta_2 Tiempo + \epsilon$$

Con lo cuál obtienen los siguientes resultados:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7548     0.6481  10.423 9.33e-10 ***
Tobías       -4.8098     0.5911  -8.137 6.26e-08 ***
Tiempo        2.7079     1.1264   2.404   0.0255 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.236 on 21 degrees of freedom
Multiple R-squared:  0.7707,    Adjusted R-squared:  0.7489
F-statistic: 35.29 on 2 and 21 DF,  p-value: 1.923e-07
```

a. Plantee el modelo que representa los daños hechos por maracuyá
b. Haga el procedimiento adecuado para contestar si el número de daños que hace Tobías es igual al número de daños que hace maracuyá.
c. ¿Cuál de las dos mascotas hace más daños?

Ejemplo

Después de instalar camaras en el apartamento para vigilar a las mascotas mientras no estan, Alejandra y Nicolás se han dado cuenta que los patrones de comportamiento de Tobías y Maracuyá no son tan simples como habían anticipado. Se han dado cuenta ques existen nuevas variables a considerar y que tienen efectos distintos dependiendo de la mascota. Despues de observarlos por mucho tiempo plantean el nuevo modelo:

$$Daños = \beta_0 + \beta_1 Tobías + \beta_2 TiempoSolos + \beta_3 Edad$$

$$+\beta_4(Tobías * TiempoSolos) + \beta_5 MomentoDía$$

$$+\beta_6(MomentoDía * TiempoSolos) + \epsilon$$

Donde MomentoDía es una variable indicadora que toma el valor de 1 si es en la mañana y 0 si es en la tarde.

Ejemplo

a. Escriba todos los modelos auxiliares.
b. ¿La variable TiempoSolos afecta el número de daños que hacen las mascotas?
c. ¿Es el efecto de la variable TiempoSolos igual para Tobías y para Maracuyá?
d. ¿Son los modelos paralelos?
e. ¿Hace Tobías el mismo numero de daños en la mañana que maracuyá en la tarde?
f. ¿Es necesaria hacer la distinción por momento del día?