Regresión Lineal Múltiple: Otros Detalles

Clase 23

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

1. Remember, Remember...

2. Regresión con solo variables categóricas
   - Motivación
   - ANOVA como regresión

3. Relaciones no lineales
   - Motivación
   - Transformaciones de variables
   - Linealización de modelos

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?

Varias Variables Categóricas

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

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

Sabemos que, 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.

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

Motivación

Un caso particular de los modelos que hemos visto, sería cuando solo hay variables categóricas, y no hay ninguna continua.

Ejemplo

Suponga que tiene la variable categótica Estatura {Baja, Promedio, Alta}. Las dummies serían:

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

Motivación

Considere el modelo de regresión dado ahora por:

$$Y = \beta_0 + \gamma_1 B + \gamma_2 P + \epsilon$$

Para el ejemplo, con 6 individuos, siendo 2 altos, 2 bajos y 2 promedios, la matriz de diseño sería:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

La matriz de diseño son solo 0 y 1!

Motivación

Se puede demostrar que los coeficientes de regresión están asociados a los promedios muestrales

Ejemplo

En el ejemplo se tendría que:

$$\hat{Y} = \bar{Y}_{A.} + (\bar{Y}_{B.} - \bar{Y}_{A.})B + (\bar{Y}_{P.} - \bar{Y}_{A.})P$$

De forma intuitiva, esto se debe a que si queremos el mejor predictor con un único valor, entonces lo mejor con lo que podemos predecir es es con el promedio

Outline

ANOVA como regresión

¿Que relación tiene esto con el diseño de experimentos?

En DOE, nos interesaba estudiar una variable de interés, a partir de factores categóricos con distintos niveles...

Equivalencia regresión y ANOVA

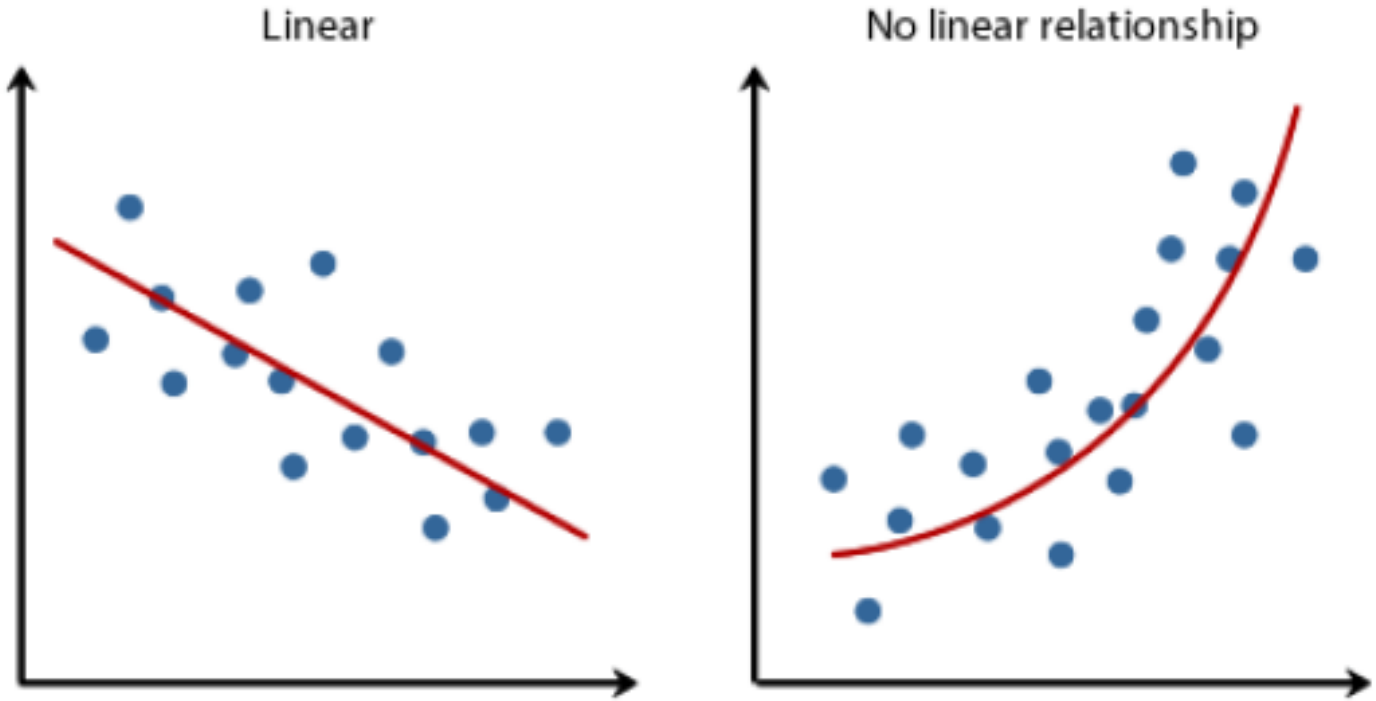
La relación es que en el caso en que solo se usan variables categóricas en una regresión, se está haciendo un procedimiento IDÉNTICO al del ANOVA. No existe diferencia!

La tabla ANOVA de diseño de experimentos es la misma tabla ANOVA de la regresión.

Outline

Motivación

Hasta el momento hemos visto modelos que plantean una relación lineal entre Y y X, es decir, una linea recta. Pero la verdad esto es extremadamente limitado en algunos casos!



Outline

Transformaciones de Variables

Si bien hemos planteado una relación lineal entre Y y X, realmente nos hemos limitado a nosotros mismos ...

El supuesto de regresión lineal, es que este sea lineal en los COEFICIENTES, en ningún momento se ha hablado de las variables!

Ejemplo

El siguiente es un modelo de regresión lineal:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 log(X) + \beta_4 \frac{1}{X} + \epsilon$$

Mientras que este no:

$$Y = \beta_0 + \beta_1 X + \beta_2 X_3^{\beta} + \beta_4 log(\beta_5 X) + \epsilon$$

Transformaciones de variables

Al incluir transformaciones de variables, hacemos el modelo más flexible para representar comportamientos no lineales, pero hay sacrificios:

- Aumenta el costo de estimación! En el ejemplo anterior, agregamos varias transformaciones, pero realmente estamos trabajando con solo una X. Es decir, se gastan muchos grados de libertad por variable.
- La interpretación! al tener el modelo como una recta, era fácil interpretar la pendiente como una razón de cambio. Al transformar, ya no es clara la relación de Y y X
- Maldición de la dimensionalidad! En el ejemplo, solo se trabajo con una X al incluir más covariables el tema se complica mucho!

Existen formas más eficientes de incorporar relaciones no lineales, como con splines, pero esto es para sus futuras electivas en estadística. Lo mejor es solo incluir transformaciones de las variables, cuando sus motivos lo justifiquen.

Transformaciones de variables

Algunas relaciones interesantes, usadas especialmente en econometría, son:

LIN-LOG

$$Y = \beta_0 + \beta_1 log(X) + \epsilon$$

Donde un incremento de 1% en X, aumenta β_1 unidades en Y.

LOG-LIN

$$log(Y) = \beta_0 + \beta_1 X + \epsilon$$

Donde un incremento de 1 unidad en X, aumenta β_1 % unidades en Y.

LOG-LOG

$$log(Y) = \beta_0 + \beta_1 log(X) + \epsilon$$

Donde un incremento de 1% en X, aumenta β_1 % unidades en Y.

Outline

Linealización de modelos

Ya vimos que existen modelos de tipo regresión que no son lineales, pero haciendo trucos, algunos se pueden linealizar, y hacer todo como ya se ha venido trabajando:

Ejemplo: Regresión exponencial

Una relación exponencial se plantea de la forma:

$$Y = Ae^{rx}$$

donde A y r son constantes de ajuste del modelo. Es usual ver este modelo en el contexto de biología para explicar el número de bacterias a través del tiempo, donde A sería el número inicial de bacterias y r sería la tasa de crecimiento.

Claramente, este modelo no se puede plantear como una regresión lineal, pero es linealizable.

Linealización de modelos

Ejemplo: Regresión Exponencial

Aplicando logaritmo a ambos lados de la igualdad se tiene que:

$$Y = Ae^{rx}$$

$$ln(Y) = ln\left(Ae^{rx}\right)$$

$$ln(Y) = ln(A) + rX$$

haciendo los cambios de nombres $\tilde{Y} = ln(Y)$, $\beta_0 = ln(A)$, $\beta_1 = r$ y agregando un choque aleatorio, se puede plantear el siguiente modelo de regresión lineal para la estimación:

$$\tilde{Y} = \beta_0 + \beta_1 X + \epsilon$$

Linealización de modelos

Ejemplo: Regresión exponencial

Los parámetros cambian su interpretación por las transformaciones! (e.g β_1 no es la relación de cambio de Y con X, sino la de $ln(Y)$ con X)

Todo el tema de inferencia estadística y estimación se mantiene, salvo algunos detalles:

- Para obetener los parámetros originales debemos despejar (e.g $A = e_0^{\beta}$
- Lo mismo para los intervalos de confianza (e.g si $IC_{1-\alpha/2}(\beta_0) = [L_{\beta 0}, U_{\beta 0}]$, entonces $IC_{1-\alpha/2}(A) = [e^{L_{\beta 0}}, e^{U_{\beta 0}}]$)
- Nuevamente para los intervalos de predicción (e.g si $IP_{1-\alpha/2}\left(\tilde{Y}\right) = [L_{\tilde{Y}}, U_{\tilde{Y}}]$, entonces $IP_{1-\alpha/2}(Y) = [e^{L_{\tilde{Y}}}, e^{U_{\tilde{Y}}}]$)

Linealización de modelos

Ejemplo: Curva Logística

Una relación logística viene del planteamiento de la ecuación diferencial, y respectiva solución dadas por:

$$\frac{dY}{dX} = rY\left(1 - \frac{Y}{K}\right) \quad Y = \frac{K}{1 + Ae^{-rX}}$$

donde A, K y r son constantes de ajuste del modelo. Es usual ver este modelo en el contexto de crecimiento poblacional, donde A se asocia a unas características iniciales de la población, K es la capacidad de carga (conocida), y r sería la tasa de crecimiento.

Nuevamente, este modelo no se puede plantear como una regresión lineal, pero es linealizable.

Linealización de modelos

Ejemplo: Curva logística

Aplicando logaritmo a ambos lados de la igualdad se tiene que:

$$Y = \frac{K}{1 + Ae^{-rX}}$$

$$\frac{K}{Y} = 1 + Ae^{-rX}$$

$$ln\left(\frac{K}{Y} - 1\right) = ln\,(A) - rX$$

haciendo los cambios de nombres $\tilde{Y} = ln\left(\frac{K}{Y} - 1\right)$, $\beta_0 = ln(A)$, $\beta_1 = -r$ y agregando un choque aleatorio, se puede plantear el siguiente modelo de regresión lineal para la estimación:

$$\tilde{Y} = \beta_0 + \beta_1 X + \epsilon$$

Linealización de modelos

Ejemplo: Curva Logística

Los parámetros cambian su interpretación por las transformaciones! (e.g β_1 no es la relación de cambio de Y con X, sino la de $ln\left(\frac{K}{Y}-1\right)$ con X)

Todo el tema de inferencia estadística y estimación se mantiene, salvo algunos detalles:

- Para obetener los parámetros originales debemos despejar (e.g $A = e_0^{\beta}$
- Lo mismo para los intervalos de confianza (e.g si $IC_{1-\alpha/2}(\beta_0) = [L_{\beta 0}, U_{\beta 0}]$, entonces $IC_{1-\alpha/2}(A) = [e^{L_{\beta 0}}, e^{U_{\beta 0}}]$)
- Nuevamente para los intervalos de predicción (e.g si $IP_{1-\alpha/2}\left(\tilde{Y}\right) = [L_{\tilde{Y}}, U_{\tilde{Y}}]$, entonces $IP_{1-\alpha/2}(Y) = [\frac{K}{1+e^{U_{\tilde{Y}}}}, \frac{K}{1+e^{l_{\tilde{Y}}}}]$