Supuestos Regresión Lineal: Multicolinealidad

Clase 26

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

Definición

Cuando planteamos el modelo de regresión lineal,

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_k X_k + \epsilon$$

y hallamos el estimador por mínimos cuadrados, encontramos que:

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

¿Qué supuestos hicimos para encontrar ese estimador?

Definición

En efecto, asumimos que la matriz $(\mathbf{X}^T\mathbf{X})$ es una matriz invertible, pero esto no siempre pasa, o al menos no a totalidad:

Ejemplo: Dummy Trap

Cuando trabajamos con variables categóricas, vimos que si incluíamos todas las variables dummies que se generaban para cada categoría, entonces la matriz $(\mathbf{X}^T\mathbf{X})$ no era invertible. Esto sucedía por la redundancia de información.

En la practica, puede suceder que nuestras covariables sean redundantes entre ellas hasta cierto punto.

Definición

Ejemplo

Suponga que se quiere explicar el rendimiento de un estudiante en la universidad (Y), en función de los resultados en Matemáticas (X_1) y Física (X_2) obtenidos en las pruebas Saber 11. Si bien son resultados diferentes,los X's tiene una clara asociación entre ellos!

Es decir, existen variables que aunque son indicadores de cosas distintas, pueden tener comportamientos extremadamente similares, incurriendo en la redundancia de información.

Cuando esto sucede de forma grave, decimos que existen problemas de multicolinealidad.

Definición

¿Por que decimos que es un problema?

Cuando decimos redundancia de información, hacemos referencia a redundancia LINEAL, esto es, que las columnas de $\mathbf{X}$ pueden expresarse como combinaciones lineales entre ellas.

Esto lleva a que la matriz $(\mathbf{X}^T\mathbf{X})$ no sea invertible, y por tanto gran cantidad de las cuentas que hacemos en regresión no se puedan hacer.

Pero el problema realmente va más a fondo!

Definición

Multicolinealidad

Decimos que hay problemas de multicolinealidad cuando los regresores X están relacionados linealmente entre ellos, o equivalentemente que están colacionados entre ellos.

En la practica, la relación de multicolinealidad no es perfecta, es decir, la combinación lineal entre las columnas no se da a totalidad debido a efectos de aleatoriedad.

Debido a esto, sucede que si bien hay casi una relación lineal entre las X, esta no se da a cabalidad y la matriz $(\mathbf{X}^T\mathbf{X})$ si será invertible, siendo esto el núcleo de los problemas.

Outline

Implicaciones

Si bien la matriz $(\mathbf{X}^T\mathbf{X})$ es invertible, realmente esta es mal condicionada y su inversa es inestable!. Esto se puede ver fácilmente por:

$$A^{-1} = \frac{1}{det(A)} Cofactor(A)^T$$

Si la matriz $\mathbf{A}$ está cerca de no ser invertible, entonces $det(A) \approx 0$, luego el cociente de arriba se indetermina.

En la practica, la inversa podrá calcularse numéricamente, pero su resultado es totalmente cuestionable. Por lo tanto, las cuentas posteriores también serán dudosas.

Implicaciones

Estimación de Coeficientes

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Los coeficientes son inestables. Se pueden presentar valores sin sentidos, incluso signos contrarios a lo pensando! La estimación se vuelve sensible al tamaño de la muestra.

Varianza de los Coeficientes

$$Var\left(\hat{\beta}\right) = \hat{\sigma}^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

La varianza de los coeficientes también es inestable y se infla, es decir, se tienen varianzas gigantes. Esto implica que todos los procesos de inferencia estadística se ven comprometidos.

Implicaciones

- Las pruebas de significancia individual no son creíbles: Por la varianza elevada, se tenderá a no rechazar la hipótesis nula y a declarar variables como no significativas.
- El valor estimado de las pendientes de las variables tampoco es creíble por la alta volatilidad en la estimación.
- En resumen, se pierde la interpretabilidad del modelo de regresión, lo cual es uno de los aspectos más importantes para hacer una regresión!
- A pesar de todo, los estimadores siguen siendo insesgados y eficientes dentro de la clase de estimadores lineales.
- Y aún menos intuitivo, los intervalos de confianza y de predicción para Y no se ven afectados! El problema no es la predicción, sino identificar de donde proviene dicho poder predictivo en las X.

Outline

Identificación

Para saber si hay problemas de multicolinealidad, no existen como tal pruebas estadísticas, pero hay indicios que pueden generar sospecha:

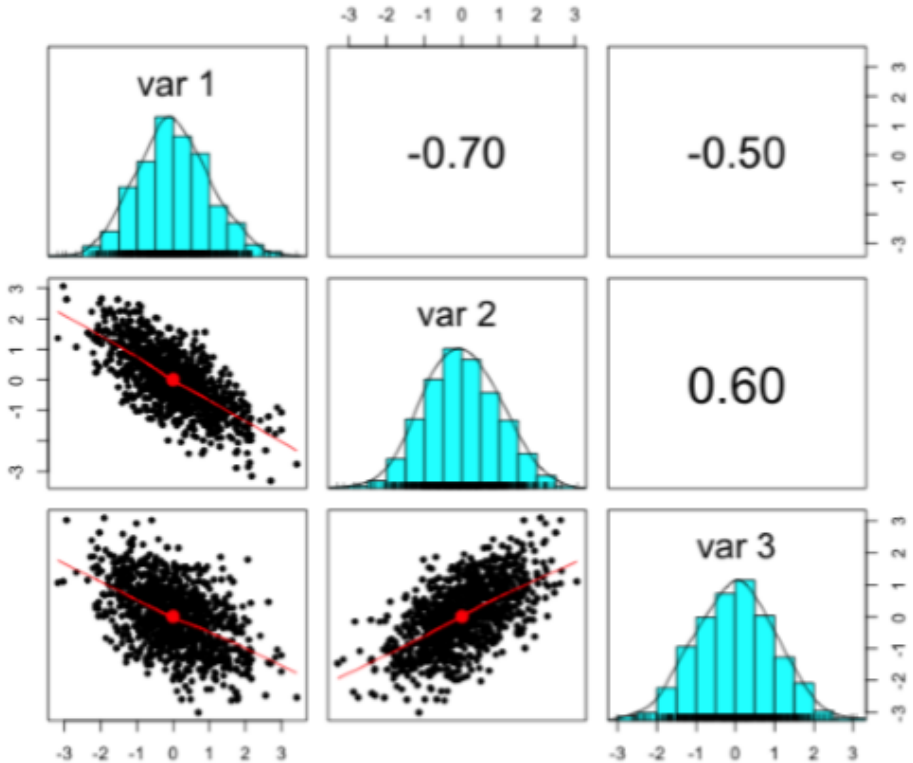
- Valores de R^2 cercanos a 1. Es difícil que un modelo sea tan bueno!
- Modelo globalmente significativo, pero muy pocas o ninguna variables individualmente significativa.
- Agregar o quitar VARIABLES al modelo causan cambios drásticos en la estimación de los coeficientes, así como en la significancia de las variables.
- Agregar o quitar DATOS causan cambios drásticos en la estimación de los coeficientes, así como en la significancia de las variables.

Pero cuidado, estos solo son indicios, no son pruebas determinantes.

Identificación

Así mismo, diferentes estrategias descriptivas y gráficas se pueden utilizar y ser de más utilidad si se dese identificar las variables que están relacionadas.

- Matrices de correlación entre las X (valores grandes en magnitud).
- Gráficos de dispersión entre los X (tendencias entre las variables).



Identificación

Los análisis anteriores, solo permiten identificar la multicolinealidad si una variable está asociada directamente a otra, más no el caso en que la asociación se de en combinaciones lineales.

Así mismo, los procedimientos son extremadamente informales y subjetivos, y no cuantifican la magnitud del problema.

Para esto ya tenemos la respuesta!

Ejemplo

Suponga que quiere verificar si hay problemas de multicolinealidad en el modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Identificación

Ejemplo

Ahora considere las siguientes regresiones (una X contra las demás):

$$X_1 = \gamma_0 + \gamma_2 X_2 + \gamma_3 X_3 + \epsilon$$

$$X_2 = \gamma_0 + \gamma_1 X_1 + \gamma_3 X_3 + \epsilon$$

$$X_3 = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \epsilon$$

Si alguna de las regresiones es globalmente significativa, se estaría diciendo que una de las X se puede expresar como combinación lineal de las otras, es decir, hay presencia de multicolinealidad.

Identificación

Una medición extremadamente popular que nace a partir de estas regresiones auxiliares, es el factor de inflación de la varianza, o por sus siglas en ingles VIF. El VIF para la j-ésima variable se define como:

$$VIF_j = \frac{1}{1 - R_j^2}$$

donde R_j^2 es el R^2 que se obtiene de la regresión de X_j vs el resto de las X. El VIF_j se interpreta como un factor que muestra el incremento en la varianza de $\hat{\beta}_j$, debido a la multicolinealidad.

Ejemplo

Si $VIF_2 = 20$, quiere decir que la varianza de $\hat{\beta}_2$ es 20 veces más grande a la que se obtendría si se tuvieran variables no correlacionadas.

Identificación

Ejemplo

En nuestro ejemplo, el calculo del VIF sería visualmente como:

$$X_1 = \gamma_0 + \gamma_2 X_2 + \gamma_3 X_3 + \epsilon \rightarrow VIF_1 = \frac{1}{1 - R_1^2}$$

$$X_2 = \gamma_0 + \gamma_1 X_1 + \gamma_3 X_3 + \epsilon \rightarrow VIF_2 = \frac{1}{1 - R_2^2}$$

$$X_3 = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \epsilon \rightarrow VIF_3 = \frac{1}{1 - R_3^2}$$

Una regla empírica del pulgar, establece que si se tiene un VIF mayor a 10, entonces hay problemas de multicolinealidad. Si este es menor a 10, entonces no hay problemas, o al menos no tan graves.

Identificación

- Hay medidas equivalentes al VIF que nacen a partir de las mismas definiciones. Algunos paquetes estadísticos usan unas en particular:

$$Tolerancia_j = \frac{1}{VIF_j} = 1 - R_j^2$$

- A la hora de determinar si existe o no multicolinealidad, lo mejor es analizar todos los gráficos y estadísticas descriptivas que puedan. No es bueno quedarse con un único criterio.
- Si encuentran indicios de problemas, lo mejor es hacer algo y no fingir que todo está bien.

Outline

Correción

Existen varios enfoques para superar los problemas de multicolinealidad:

- **Quitar variables redundantes**: Como la información está repetida en las variables X, entonces podemos dejar de incluir algunas dado que su aporte informativo se encuentra en las demás.

  En la práctica es usual quitar la variable X que tenga el VIF más grande, pues es la que mayor relación lineal tiene con las demás. Si el problema persiste después de quitar dicha variable, podría continuar con el proceso y quitar más variables hasta que se solucione.

  El problema con este enfoque es que se pierde la posibilidad de entender el efecto en Y de la variables X que se quiten, y de alguna forma, también se excluye información para el modelo

Correción

Existen varios enfoques para superar los problemas de multicolinealidad:

- **Crear índices**: En vez de quitar variables, se podrían identificar agrupaciones de variables X que estén colacionadas entre ellas, para así combinarlas y trabajar con índices.

Índices

Una forma de crear índices es con promedios ponderados de las variables:

$$W = \sum_{j=1}^{m} w_j X_j$$

Una forma óptima de selección de los pesos individuales (w_j) es por medio de componentes principales, pero eso no lo vemos aquí. Por ahora una selección simple es utilizar un promedio simple (i.e el mismo w_j para todos).

Correción

Ejemplo

Suponga que quiere utilizar como regresores X las notas obtenidas en Cálculo Diferencial, Cálculo Integral, Cálculo Vectorial, Español, Historia de Colombia, Biología Celular, Biología de Organismos y Química.

Claramente hay correlaciones entre grupos de variables! Podemos definir índices con promedios simples:

$$W_1 = \frac{1}{3}(\text{Cal. Dif.} + \text{Cal. Int.} + \text{Cal. Vect.}) \rightarrow \text{Habilidad Matemática}$$

$$W_2 = \frac{1}{2}(\text{Español} + \text{Hist. Col.}) \rightarrow \text{Habilidad Lingüística}$$

$$W_3 = \frac{1}{3}(\text{Bio. Cel.} + \text{Bio. Org.} + \text{Química}) \rightarrow \text{Habilidad en Ciencias}$$

Correción

Ejemplo

Luego el modelo de regresión no se hace sobre las variable originales, sino sobre los índices:

$$Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \epsilon$$

Si definieron bien los índices, estos no deberían estar correlacionados, y por tanto ya no hay problemas de multicolinealidad.

El problema con esto es que la regresión se hace sobre otras variables completamente diferentes a las originales! Dependiendo de sus objetivos puede ser de utilidad

Correción

Existen varios enfoques para superar los problemas de multicolinealidad:

- **Modelos alternos de regresión**: Existen otros modelos que pueden estimar la regresión en condiciones de multicolinealidad casi sin ninguna dificultad. Estos modelos suelen hacer la estimación de mínimos cuadrados, agregando una restricción a la inflación de la varianza de los coeficientes (métodos de Shinkage).

  Las más populares son las metodologías de Ridge regression y Lasso regression, que son más conocidas en el contextos de modelos predictivos, aprendizaje estadístico y Analytics.

  La ventaja de estos métodos con respecto a los anteriores, es que hacen la regresión con las variables originales sin necesidad de quitarlas o combinarlas, aunque hacen el sacrificio de crear sesgos en la estimación.