Regresion y Prediccion

Walter Sosa Escudero

Universisad de San Andrés y CONICET

La herencia: estimacion e inferencia

La herencia econometrica

$$Y = X\beta + u; \qquad \hat{\beta} = (X'X)^{-1}X'Y$$

- Objetivo: estimar β. Inferencia acerca de β.
- Supuestos clasicos
- Teorema de Gauss/Markov: $\hat{\beta}$ es MELI. $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- $\hat{\beta}$ minimiza $\sum e_i^2$ o maximiza R^2.

Predictor optimo

X_i, vector columna igual a la i-esima fila de X

$$Y_i = X_i'\beta + u_i, \qquad i = 1, \ldots, n$$

- $\hat{Y}_i \equiv X_i'\hat{\beta}$
- $E(\hat{Y}_i) = X_i'\beta$
- $V(\hat{Y}_i) = X_i'V(\hat{\beta})X_i = \sigma^2 X_i'(X'X)^{-1}X_i$

Resultado: si $\hat{\beta}$ es insegado y de varianza minima, entonces $\hat{Y}_i$ es un *predictor* insegado y de varianza minima, ambos en la clase de estimadores/predictores lineales e insesgados.

Demostracion:

Si $\hat{\beta}$ es insesgado, $E(X_i'\hat{\beta}) = X_i'\beta$

Si $\hat{\beta}$ es eficiente, $V(\tilde{\beta}) - V(\hat{\beta})$ es una matriz semidefinida positiva, para cualquier $\tilde{\beta}$ lineal e insesgado:

$$\lambda' \left(V(\tilde{\beta}) - V(\hat{\beta}) \right) \lambda \geq 0$$

para cualquier $\lambda \in \Re^k$. En particular para cualquier X_i:

$$X_i' V(\tilde{\beta}) X_i - X_i' V(\hat{\beta}) X_i \geq 0$$

por lo que $X_i' V(\hat{\beta}) X_i$ es minimo en la clase de predictores basados en estimadores lineales e insesgados.

Prediccion vs estimacion?

Error cuadratico medio

Idea: predecir requiere estimar.

- $\hat{\beta}$, Objetivo: estimar β.
- *Error cuadratico medio*:

$$ECM(\hat{\beta}) \equiv E(\hat{\beta} - \beta)^2.$$

  $\hat{\beta}$ es una VA, β, no.

- Recordar: $V(\hat{\beta}) \equiv E\left(\hat{\beta} - E(\hat{\beta})\right)^2$ y $Sesgo(\hat{\theta}) \equiv E(\hat{\beta}) - \beta$

Descomposicion sesgo varianza:

$$ECM(\hat{\beta}) = Sesgo^2(\hat{\beta}) + V(\hat{\beta}).$$

Intuicion: cuan mal predice $\hat{\beta}$ depende de cuan descentrado esta en relacion a la verdad (sesgo) mas cuan disperso es relacion a su propio centro (varianza).

Demostracion: Sumar y restar $E(\hat{\beta})$ en la formula del ECM y obtener

$$E\left[\left(\hat{\beta} - E(\hat{\beta})\right) + \left(E(\hat{\beta}) - \beta\right)\right]^2$$

. Expandir los cuadrados y tomar esperanzas:

$$E\left(\hat{\beta} - E(\hat{\beta})\right)^2 + E\left(E(\hat{\beta}) - \beta\right)^2 + 2E\left[\left(\hat{\beta} - E(\hat{\beta})\right)\left(E(\hat{\beta}) - \beta\right)\right]$$

A partir de aca, los dejo solos.

Error de pronostico

- *Error de pronostico*: $Err(\hat{Y}) \equiv E\left(Y - \hat{Y}\right)^2$.
- Diferencia con ECM: compara variables aleatorias.
- Modelo: $Y = f(X) + u, \quad E(u) = 0, V(u) = \sigma^2$.
- $f(X)$ es la parte *sistematica* y u la *no sistematica*
- Resultado: m que mejor predice Y es $m = E(Y)$.
- Si $E(Y) = f(X)$, u no observable y $f(X)$ conocida, $f(X)$ es el mejor predictor.

$\mu_Y \equiv E(Y)$ **es el mejor predictor de** Y

Sumar y restar μ_Y in $E(Y-m)^2$, distribuir el cuadrado y obtener:

$$E(Y-\mu_y)^2 + E(\mu_y - m)^2 + 2E\left[(Y-\mu_Y)(\mu_Y - m)\right]$$

Mostraremos que el tercer sumando es cero. Expandiendo el producto:

$$2E\left[Y\mu_Y - Ym - \mu_Y^2 + \mu_Y m\right]$$

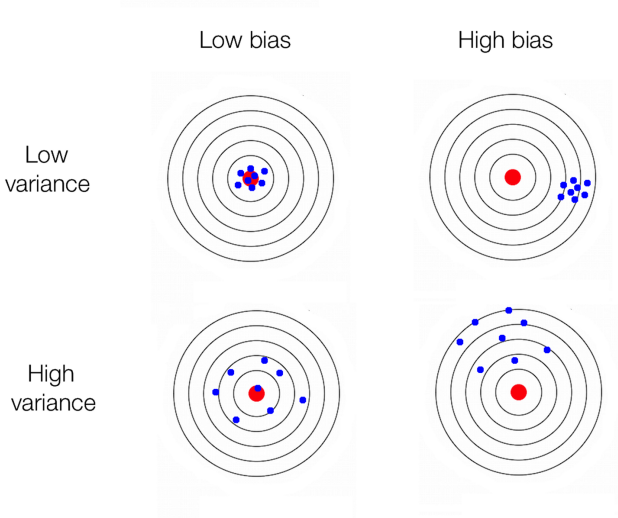
Tomamos esperanzas, para obtener

$$2\left[\mu_Y^2 - \mu_Y m - \mu_Y^2 + \mu_Y m\right] = 0$$

. Nos queda:

$$E(Y-\mu_y)^2 + E(\mu_Y - m)^2$$

El primer termino no depende de m, no podemos hacer nada para achicarlo. El segundo sumanod es no negativo, podemos anularlo si hacemos$m = \mu_Y$. Entonces, tomando $m = \mu_Y$ minimizamos $E(Y-m)^2$



Fuente: https://medium.com/@aymantidy/bias-variance-decomposition-the-story-behind

ECM y error de pronostico

En la practica, reemplazar $f(X)$ con $\hat{f}(X)$ (una VA).

- $Err(\hat{Y}) = E\left(Y - \hat{f}\right)^2$
- Importante: $Err\left(Y - \hat{f}\right) = ECM(\hat{f}) + \sigma^2$

$$\begin{aligned} E\left(Y - \hat{f}\right)^2 &= E\left((Y - f) + (f - \hat{f})\right)^2 \\ &= E(f - \hat{f})^2 + E(u^2) + 2\,E\left(u(f - \hat{f})\right) \\ &= \sigma^2 + ECM(\hat{f}) \end{aligned}$$

- Error reducible (ECM) mas irreducible (σ^2)
- Link prediccion y estimacion. Predecir correctamente requiere estimar correctamente.

- Usando la descomposicion

$$Err\left(Y-\hat{f}\right)=Sesgo^2(\hat{f})+V(\hat{f})+\sigma^2$$

- Econometria: $\hat{f}$ insesgado, minimizar Err es minimizar varianza. Preferencia lexicografica por la insesgadez.
- Machine learning: estrategias *sesgadas* pueden implicar una reduccion drastica en la varianza.
- Puede ser que el minimo ECM ocurra para predictores sesgados.

Prediccion dentro y fuera de la muestra

Supongamos que tenemos $(Y_i, X_i), \quad i = 1 \ldots, n$ para $Y_i = f(X_i) + u_i$. El error de pronostico estimado es:

$$\widehat{Err(\hat{Y})} \equiv \sum_{i=1} (Y_i - \hat{Y}_i)^2$$

- Alternativas: promedio $(1/n)$ o raiz cuadrada.
- En econometria, $\widehat{Err(\hat{Y})} = \sum_{i=1}^{n} e_i^2$
- $R^2 = 1 - \widehat{Err(\hat{Y})}/SCT$ (bondad del ajuste)
- MCO minimiza el error de pronostico y maximiza R^2: minimizar el error de pronostico es maximixar el ajuste.

- Problema: el desafio de machine learning es minimizar el error de pronostico *fuera* de la muestra
- MCO: minimiza *dentro* de la muestra.
- Importante: que $\hat{f}$ funcione bien dentro de la muestra no implica que lo haga fuera.
- Overfit (lo veremos mas adelante): estimadores que funcionan muy bien *dentro* funcionan muy mal *fuera* de la muestra.

- Muestra original: MO
- Dos conjuntos disjuntos: muestra de entrenamiento (ME) y muestra de evaluacion o test (MT), de modo que $MO \cup MT = MU$
- Error de pronostico en la muestra de test/evaluacion:

$$\widehat{Err_{MT}}(\hat{Y}) = \sum_{i \in MT} (Y_i - \hat{Y}_i)^2$$

- Ejemplo: $N = 100$, las primeras 70 observaciones son usadas para estimar (entrenar) y las 30 restantes para evaluar el modelo:

$$\widehat{Err_{MT}}(\hat{Y}) = \sum_{i \in 71}^{100} (Y_i - \hat{Y}_i)^2$$

- No hay una forma obvia de partir en MT y MU.
- En algunos casos si (Netflix game)
- Econometria: $ME = \emptyset$. Por que?
- Trade off en la particion ME, MT
- Como resolverlo: cross validation (mas adelante)