Multiple Linear Regression

The Gauss-Markov Theorem

Math 392

The Gauss-Markov Theorem makes the claim that **$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS}}$ are the best linear unbiased estimates (BLUE) of $\boldsymbol{\beta}$.**

What do we mean by linear?

Linear refers to estimates that are a linear function of the random variable, which in the regression setting is Y.

$$\hat{\beta}_j = c_{1,j}Y_1 + c_{2,j}Y_2 + \ldots c_{n,j}Y_n$$

Or in matrix form:

$$\hat{\beta} = CY$$

In the least squares case, $C = (X'X)^{-1}X'$.

What do we mean by unbiased?

The same thing that usually mean: $E(\hat{\beta}) = \beta$.

What do we mean by best?

Here best is defined as the estimator that has the lowest MSE. Since we're working with the class of estimates that are all unbiased, the one with the lowest MSE will be the one with the lowest variance. Here, though, we're thinking about the variance and covariance of a whole vector, so how can we compare them? We say that $\hat{\beta}_{LS}$ is best if $Var(\hat{\beta}_{LS}) - Var(\tilde{\beta})$ (which is a $p \times p$ matrix) is positive semi-definite for all other estimates $\tilde{\beta}$. A $p \times p$ matrix M that is positive semi-definite is defined by the property that

$$a'Ma \geq 0$$

for all non-zero vectors $a \in \mathbb{R}^p$.

Vaguely speaking, this means that you can think about M as you would a positive number. The vector a when multiplied by M will be rotated and rescaled, but never in a way that reflects it about the origin (changes its sign).

Note that while working through the proof of the claim of Gauss-Markov Theorem, there are three assumptions about the true data generating model that it relies upon.

1. $Y = X\beta + \epsilon$
2. $E(\epsilon) = 0$
3. $Var(\epsilon) = \sigma^2 I$

Proof

Let $\tilde{\beta} = CY$, $C = (X'X)^{-1}X' + D$ for any $p \times n$ non-zero matrix D. Note that this is an expression that characterizes *any* linear estimator based on different choices of D.

First consider the expected value of this estimator.

$$\begin{aligned}
E(\tilde{\beta}) = E(CY) &= CE(Y) && (1)\\
&= CE(X\beta + \epsilon) && (2)\\
&= CE(X\beta) + CE(\epsilon) && (3)\\
&= CX\beta && (4)\\
&= ((X'X)^{-1}X' + D)X\beta && (5)\\
&= (X'X)^{-1}X'X\beta + DX\beta && (6)\\
&= \beta + DX\beta && (7)
\end{aligned}$$

Since we're interested in the class of unbiased estimators, DX must be 0.

$$\begin{aligned}
Var(\tilde{\beta}) &= Var(CY) && (8)\\
&= CVar(Y)C' && (9)\\
&= C\sigma^2 IC' && (10)\\
&= \sigma^2((X'X)^{-1}X' + D)[(X'X)^{-1}X' + D]' && (11)\\
&= \sigma^2((X'X)^{-1}X' + D)(X(X'X)^{-1} + D') && (12)\\
&= \sigma^2[(X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'D' + DX(X'X)^{-1} + DD'] && (13)\\
&= \sigma^2(X'X)^{-1} + \sigma^2 DD' && (14)\\
&= Var(\hat{\beta}_{LS}) + \sigma^2 DD' && (15)
\end{aligned}$$

were many of the terms cancel because if we're working with unbiased estimators, $DX = 0$. Since $\sigma^2 DD'$ is a positive semi-definite matrix, this shows that *any* other linear unbiased estimator of β will have a variance at least as large as the least squares estimates.

Sidenote

We know that $\sigma^2 DD'$ is positive semi-definite by putting it in place of M in our definition of positive semi-definite

$$a'\sigma^2 DD'a = \sigma^2(D'a)'(D'a).$$

σ^2 is a scalar, $(D'a)'$ is an $1 \times n$ vector, and $(D'a)$ is the same vector but transposed. The resulting vector multiplication is just the sum of the squares of all of those values, so as long as the values of D are not all zero, the scalar will be positive.

This also accords with our informal sense of a positive semi-definite matrix acting like a positive number: any non-zero number squared (read: DD') is a positive number.

Postscript

So now we can chalk up several reasons for why statisticians have been pleased with $\hat{\beta}_{LS}$ for hundreds of years now.

- As a loss function, RSS has some intuitive appeal.
- $\hat{\beta}_{LS}$ has a closed form.
- $\hat{\beta}_{LS}$ are BLUE (Gauss-Markov).