Stat 200A, Fall 2018
A. Adhikari, UC Berkeley

Multiple Regression Model and Estimation

The Model

There are n individuals, on each of whom you have measured several variables. The goal is to estimate one of the variables by the least squares linear function of the others, under the following assumptions:

For $1 \le i \le n$,

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

where:

- the intercept β_0 and slopes $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are unobservable constants
- $x_{i,1}, x_{i,2}, \ldots, x_{i,p-1}$ are the observed constant values of $p-1$ predictor variables for individual i
- ϵ_i has the normal $(0, \sigma^2)$ distribution for some unobservable σ^2, and $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are i.i.d.

Thus Y_i is a random variable. The model assumes that we observe Y_i for each individual. Based on these observations and the known values of the predictor variables, the goal is to find the best estimates of the intercept and slopes, and also of the error variance σ^2.

These estimates can then be used to make predictions for new individuals, assuming that the model holds for the new individuals as well.

The model is more compactly written as

$$Y \;=\; X\beta + \epsilon$$

where:

- $\beta = [\beta_0, \beta_1, \ldots, \beta_{p-1}]^T$ is a $p \times 1$ vector of the coefficients
- X is an $n \times p$ matrix with Column 0 a vector of 1's and Column j for $1 \le j \le p-1$ consisting of the n observations on the jth predictor variable; the ith row has the values of all the predictor variables for individual i
- ϵ is an $n \times 1$ multivariate normal $(0, \sigma^2 I)$ vector; the mean vector is an n-vector of 0's and I is the $n \times n$ identity matrix

The Goal

Any linear function of the predictor variables can be written as $X\gamma$ where γ is a $p \times 1$ vector of coefficients. Think of $X\gamma$ as an estimate of Y. The goal is to find the vector γ that minimises the mean squared error

$$MSE(\gamma) \;=\; \frac{1}{n}\sum_{i=1}^{n}(Y_i - (X\gamma)_i)^2$$

This is the same as minimizing the sum of squared errors

$$SSE(\gamma) \;=\; \sum_{i=1}^{n}(Y_i - (X\gamma)_i)^2$$

Again for compactness it will help to use matrix notation. For an $n \times 1$ vector w,

$$\sum_{i=1}^{n} w_i^2 \;=\; w^T w \;=\; w \cdot w \;=\; \|w\|^2$$

Then the goal is to find the $p \times 1$ vector $\hat{\beta}$ that minimizes $\|Y - X\gamma\|^2$ over all vectors γ.

Typically you will also have to estimate σ^2 but we will not cover that; take Stat 200B.

The Candidate

The claim is that the vector $\hat{\beta}$ defined as follows will do the job.

$$\hat{\beta} \;=\; (X^T X)^{-1} X^T Y$$

Before proving this, notice that $\hat{\beta}$ is a linear function of Y. This makes it straightforward to identify its distribution, which you will do in exercises.

Also note that the estimated Y is

$$\hat{Y} \;=\; X\hat{\beta} \;=\; X(X^T X)^{-1} X^T Y$$

which is also a linear function of Y.

Proof: Step 1

Define the $n \times 1$ vector of observed errors, or *residuals*, as

$$e \;=\; Y - \hat{Y}$$

As we have seen repeatedly, the key to least squares is that the error is orthogonal to the space of allowed functions. Our space of allowed functions is all linear functions of X. So we will show:

The residuals are orthogonal to each column of X.

To see this, calculate the $p \times 1$ vector $X^T e$. Each of its elements is the dot product of e and one column of X.

$$X^T e \;=\; X^T(Y - \hat{Y}) \;=\; X^T Y - X^T \hat{Y} \;=\; X^T Y - X^T X (X^T X)^{-1} X^T Y \;=\; X^T Y - X^T Y \;=\; 0$$

Proof: Step 2

In the calculation below, keep in mind that $e = Y - X\hat{\beta}$ and $SSE(\hat{\beta}) = \|Y - X\hat{\beta}\|^2$.

Let γ be any $p \times 1$ vector. Then

$$
\begin{aligned}
SSE(\gamma) &= \|Y - X\gamma\|^2 \\
&= \|(Y - X\hat{\beta}) + (X\hat{\beta} - X\gamma)\|^2 \\
&= \|Y - X\hat{\beta}\|^2 \ + \ \|X\hat{\beta} - X\gamma\|^2 + 2(X\hat{\beta} - X\gamma)^T(Y - X\hat{\beta}) \\
&= SSE(\hat{\beta}) \ + \ \|X\hat{\beta} - X\gamma\|^2 + 2(X\hat{\beta} - X\gamma)^T e \\
&= SSE(\hat{\beta}) \ + \ \|X\hat{\beta} - X\gamma\|^2 + 2(\hat{\beta} - \gamma)^T X^T e \\
&= SSE(\hat{\beta}) \ + \ \|X\hat{\beta} - X\gamma\|^2 \qquad \text{by Step 1} \\
&\geq SSE(\hat{\beta})
\end{aligned}
$$

In exercises you will find the distributions of $\hat{\beta}$ of $\hat{Y}$. Both will depend on the unknown σ^2. It should come as no surprise that the best estimate of σ^2 has a chi-squared distribution. There is a bit of work involved in establishing that the estimate is

$$S^2 \ = \ \frac{1}{n-p}\|e\|^2$$

A bit more work establishes that $\frac{n-p}{\sigma^2}S^2$ has the chi-squared $(n-p)$ distribution. We'll leave that work for 200B.

But you do know the result in the particular case $p = 1$. That's the case when you don't have any predictor variables at all. The matrix X has just one column, consisting of all 1's. In other words, you are trying to find the best constant by which to estimate Y.

You know that the best constant is $\bar{Y}$. You also know that

$$S^2 \ = \ \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

is the best (least squares and unbiased) estimate of σ^2 and that $\frac{n-1}{\sigma^2}S^2$ has the chi-squared $n-1$ distribution under the assumption of normality. That's the special case of the stated new result when $p = 1$.