Handout 15: Linear Regression

Setup Today we want to expand the regression set-up that we saw last time. Specifically, for some fixed positive integer p, consider a set of fixed real numbers $x_{i,j}$ for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. Then, consider observing a independent random sample of size n denoted by $Y_1, \ldots, Y_n$ where

$$Y_i \sim N(\sum_j x_{i,j} \cdot b_j, \sigma^2)$$

For some unknown constants $b_1, \ldots, b_p$, and σ^2. We can write the expected values of all of the observations as a single equation as follows:

$$\mathbb{E} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & \ddots & & x_{2,p} \\ \vdots & & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}.$$

Or, significantly more compactly, in a matrix format:

$$\mathbb{E}Y = Xb$$

Here, we now have a random vector Y on the left and a matrix multiplied by a vector of unknown parameters on the right.

Interpretation The parameter b_j can be interpreted as the average change in Y expected in a unit change of x_j where all other variables are held fixed.[1] These can be thought of as analogous to partial derivatives. Note that we do not have an explicit intercept term in the model because we could integrate one by setting $x_{i,1}$ to 1 for all i.

[1] We use x_j to indicate the feature underlying the individual values $x_{i,j}$ associated with each observation.

MLE Just as we saw last time, the MLE estimators for the b_j parameters of linear regression come from minimizing the sum of squared differences between the Y_i's and their expected means. In matrix form, this means minimizing $||Y - Xb||_2^2$.[2] To do this, we take the gradient with respect to b, which can be done as follows:

[2] Neither multivarate calculus nor linear algebra are prerequisites for this class, so it's okay if some of the details are hazy here. I won't ask any of this on an exam and am actually moving quicker than usual.

$$\begin{aligned} \nabla_b \left[||Y - Xb||_2^2 \right] &= \nabla_b \left[Y^t Y + b^t X^t X b - 2 Y^t X b \right] \\ &= 2X^t X b - 2X^t Y. \end{aligned}$$

Then, setting it to zero, we get:

$$\widehat{b}_{MLE} = (X^t X)^{-1} X^t Y.$$

This result is call the normal equation (or normal equations). Similarly, the estimator of the variance is given by:

$$\widehat{\sigma^2} = \frac{1}{n-p} ||Y - Xb||_2^2.$$

<u>Inference</u> Looking at the normal equation, you can see that the MLE estimator of each b_j is a linear combination of the values of Y_i. Therefore, each will be normally distributed. Specifically, we have:

$$\widehat{b}_j \sim N(b_j, \sigma^2 \cdot (X^t X)^{-1}_{j,j}).$$

From here, using the same methods we used the first several weeks of the course, we can show that for any $j \in \{1, \ldots, p\}$, the following is a pivot statistic with a T-distribution having $n - p$ degrees of freedom:

$$T = \frac{\widehat{b}_j - b_j}{\sqrt{\widehat{\sigma^2} \cdot (X^t X)^{-1}_{j,j}}}$$

We can use this to compute confidence intervals and hypothesis tests for individual parameters b_j.

<u>Extensions</u> This has been a very quick introduction to linear regression, a topic best covered through a semester-long course following this one (we hope to offer such a course at some point, but likely not until most of you have graduated). I hope that several of you will be showing some common extensions to the core model for your final project.