Lecture 15

Simple Linear Regression - I

In this lecture, we discuss the theory behind simple linear regression. This theory will extend naturally to multivariate regression but it is easier to explain and to picture with one predictor variable.

Simple Linear Regression – Theory

Suppose we have two variables, X and Y, measured on the same individuals, and we want to use X to predict the value for Y. Examples might include predicting whether a fetus has Down Syndrome or not from ultrasound measurements and the femur length of the fetus from its head diameter measurement. In the former situation, Y is binary and this is a problem for logistic regression. The latter situation is more typical of the one we will be considering here - both X and Y are continuous. The simplest model is linear

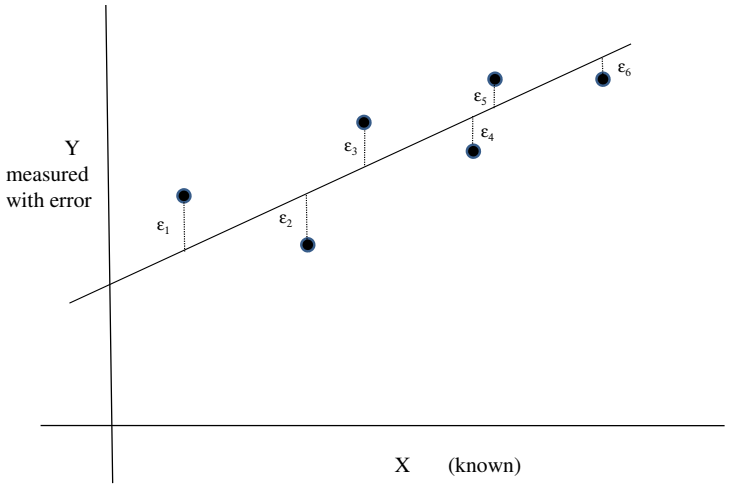
$$Y = a + bX = \beta_0 + \beta_1 X$$

That is the ideal situation. For statistics, we say $\mathrm{E}Y = a + bX = \beta_0 + \beta_1 X$ and

$$Y_i = a + bX_i + \epsilon_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$. Note that σ^2 is constant (along the line - for any value of X) - that is the *homogeneity of variance* assumption.

Given data, what are the best guesses for β_0, β_1 (a, b)? The definition of "best" here is the values that minimize the vertical distances from the observed data, Y_i, to their predicted values on the regression line, $a + bX_i$. That is, we minimize $\sum \epsilon_i^2$. (Note: $\sum \epsilon_i = 0$. Why?)



The mathematics of finding a and b is as follows: Note that $\sum \epsilon_i^2 = \sum (Y_i - a - bX_i)^2$ and let

$$f(a,b) = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

To minimize f, we take the derivative with respect to a and b, set them equal to zero, and solve for a and b:

$$\begin{aligned} \frac{\partial f}{\partial a}(a,b) &= \sum_{i=1}^{n} -2(Y_i - a - bX_i) = 0 \\ \frac{\partial f}{\partial b}(a,b) &= \sum_{i=1}^{n} -2X_i(Y_i - a - bX_i) = 0 \end{aligned}$$

Remember, X_i and Y_i are known data values and we are trying to solve for a and b. A little algebra yields the solution to the two equations above are:

$$\begin{aligned} \hat{a} &= \frac{(\sum X_i^2)(\sum Y_i) - (\sum X_i)(\sum X_i Y_i)}{n\sum X_i^2 - (\sum X_i)^2} \\ \hat{b} &= \frac{n(\sum X_i Y_i) - (\sum X_i)(\sum Y_i)}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \end{aligned}$$

The hats are because the values are estimated from the data and are estimates of the real values (statistics notation). Again, these formulas are programmed for you in Minitab or any other statistical software package you might use.

There is another point of view for finding a and b which is the method of maximum likelihood. We mention the point of view here because it is going to come up again - especially for logistic regression. Since the residuals are normally distributed, the probability density on them is

$$f(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(Y_i - a - bX_i)^2/(2\sigma^2)}$$

and the likelihood function for a and b given the data is the product of these over all the data (i):

$$\prod f(\epsilon_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^{n}(Y_i - a - bX_i)^2/(2\sigma^2)}$$

If we find the values of a and b that maximize this likelihood function, it is the same as minimizing the sum in the exponential and so we get the same answer as above. But maximizing the likelihood function generalizes to more situations that minimizing the sum of squares of the residuals does since some models do not have meaningful residuals.

After you find the best guess for a and b (β_0 and β_1), you want to address question such as: could the slope (b or β_1) be zero? If the slope is zero, then there is no significant linear relationship between X and Y. To answer questions like this, we need to know the distribution for $\hat{\beta}_0$ and $\hat{\beta}_1$. Both of these estimators are unbiased. That is, $E\hat{\beta}_0 = \beta_0$ and $E\hat{\beta}_1 = \beta_1$. Further, the distributions are both t distributions with $n-2$ degrees of freedom when the estimator is properly normalized by its standard error. The $n-2$ degrees of freedom comes from the fact that we are using the data to predict two parameters a and b. The standard errors are:

$$\text{S.E.}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}} \qquad \text{S.E.}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_x^2}}$$

Why do the above formulas work? Consider

$$\hat{\beta}_1 = \hat{b} = \frac{(\sum X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

View X as fixed and Y are random. Then

$$\text{Var}(\hat{b}) = \frac{\sum (X_i - \bar{X})^2 \text{Var}\left(Y_i - \bar{Y}\right)}{\left(\sum (X_i - \bar{X})^2\right)^2} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma^2}{(n-1)\sigma_x^2}$$

The example we will use is some data I was asked to analyze for a medical doctor in Oregon. He was trying to reproduce results showing that ultrasound measurements of fetuses are useful in predicting whether the fetus has Down syndrome or not. We will simply look at the problem of predicting femur length from biparietal diameter. The data are available at:

http://mypage.iu.edu/ ehouswor/Fall2004/Math467/L15.html

Exercises for Lecture 15

1. –

2. –