Handout 14: Simple Linear Regression

Setup Let $x_1, \ldots, x_n$ be a set of fixed real numbers. Consider observing an independent random sample of size n denoted by $Y_1, \ldots, Y_n$ where

$$Y_i \sim N(b_0 + b_1 \cdot x_i, \sigma^2)$$

For some unknown constants b_0, b_1, and σ^2. So, while the observations are independent, they are not identically distributed because they have different means. This model is called a **(simple) linear regression**.[1] One can visualize the model as trying to fit a line with intercept b_0 and slope b_1 to a plot with the x_i's on the x-axis and the Y_i's on the y-axis. It is often called a *best-fit line* in a non-technical context.

[1] We will see the more general case of a linear regression in the next set of notes.

Interpretation The slope parameter b_1 has an important interpretation: it gives the amount that we expect Y to change on average with a unit change in x. For example, if x are the number of hours you prepare for the SAT and Y is your score on the exam, b_1 would be the average increase in SAT score for every extra hour studied.

Point Estimators As you might imagine, the most popular method for estimating the unknown parameters is maximum likelihood. On today's worksheet, we will derive the MLE in the special case that b_0 is equal to zero. In the full version, the MLE values for the intercept and slope can be computed as

$$\begin{aligned}\widehat{b}_1 &= \frac{\sum_i (x_i - \bar{x}) \cdot (Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2} \\ \widehat{b}_0 &= \bar{Y} - \widehat{b}_1 \cdot \bar{x}\end{aligned}$$

If we define $\widehat{Y}_i$ to be $\widehat{b}_0 + \widehat{b}_1 \cdot x_i$, the **fitted value** of Y, then the MLE of the variance is

$$\widehat{\sigma^2} = \frac{1}{n-2} \sum_i (Y_i - \widehat{Y}_i)^2.$$

There should be some intuition from our previous results about the denominator in the previous equation from the fact that we have n data points and a model for the mean with 2 parameters. The difference in their dimensionality gives the expected amount that the predicted values of Y should vary about their actual values.

Inference The form of the estimators for the slope and intercept allow us to compute explicit T-statistics much like we did for the first unit of the course. You will see how to do this is the simple case on the worksheet and in the more general case on the next handout.