Worksheet 15 (Solutions)

1. Consider a simple linear regression where the first $n/2$ values of x_i are zero and the second $n/2$ values of x_i are 1. We can use this to model the mean of a variable that depends on whether the corresponding x_i is in group 0 or group 1. Specifically, how do the means of Y_i in these two groups correspond to the parameters b_0 and b_1?

Solution: The parameter b_0 is the mean of Y_i in group 0 and the parameter b_1 gives the increase of the mean in group 1. So, group 1 has mean $b_0 + b_1$. Notice that checking the hypothesis $H_0 : b_1 = 0$ is the same as checking where there is no difference in mean across the two groups.

2. Let $\bar{y}_A$ be the mean of the first $n/2$ values of Y_i and let $\bar{y}_B$ be the mean of the second $n/2$ values of Y_i. Consider the following form of the MLE for $\widehat{b}_1$:[1]

[1] It can be shown that it is equivalent to the form on Worksheet 14.

$$\widehat{b}_1 = \frac{\sum_i (y_i - \bar{y})(x_i)}{\sum_i (x_i - \bar{x})^2}.$$

Find a simple formula for $\widehat{b}_1$ in terms of $\bar{y}_A$ and $\bar{y}_B$.

Solution: Note that $\bar{y} = \frac{1}{2} \cdot (\bar{y}_A + \bar{y}_B)$, an important identity that we will use below. Then, plugging in the definition, we have:

$$\begin{aligned}
\widehat{b}_1 &= \frac{\sum_i (y_i - \bar{y})(x_i)}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=n/2+1}^{n} (y_i - \bar{y})}{n/4} \\
&= \frac{4}{n} \times \sum_{i=n/2+1}^{n} (y_i - \bar{y}) \\
&= \frac{4}{n} \times \left[\frac{n}{2}\bar{y}_b - \frac{n}{2}\bar{y} \right] \\
&= 2\left[\bar{y}_b - \bar{y} \right] \\
&= 2\left[\bar{y}_b - \frac{1}{2} \cdot (\bar{y}_A + \bar{y}_B) \right] \\
&= \bar{y}_B - \bar{y}_A.
\end{aligned}$$

3. Continuing from the previous question, find a simple formula for $\widehat{b}_0$ in terms of $\bar{y}_A$ and $\bar{y}_B$.

Solution: Here we have, by plugging from the previous equation, the following:

$$
\begin{aligned}
\widehat{b}_0 &= \bar{y} - (\bar{y}_B - \bar{y}_A) \cdot \bar{x} \\
&= \frac{1}{2} \cdot (\bar{y}_A + \bar{y}_B) - (\bar{y}_B - \bar{y}_A) \cdot \frac{1}{2} \\
&= \frac{1}{2} \cdot (\bar{y}_A + \bar{y}_B - \bar{y}_B - \bar{y}_A) \\
&= \frac{1}{2} \cdot (2\bar{y}_A) \\
&= \bar{y}_A
\end{aligned}
$$

4. What is the connection between the linear regression here and a two-sample T-test for the means across the two groups?

Solution: They are exactly equivalent techniques! Just one reason that some people claim regressions are really all that you need.