Simple Linear Regression Inference

In the previous hypothesis tests, we were trying to determine if we had different ‘sets’ of data, i.e., different means, proportions, whatever. Now, with bivariate data, we know there are different ‘sets’, we want to know whether they are linearly related or not. We know the x’s are different, what we want to know is whether they are each related to the different y’s in the same way—using a line with the same intercept and slope.

Again, there are various things we must assume:
1. There is a true or population line (or equation): \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \), where \( \beta_0 \) is the y-intercept and \( \beta_1 \) is the slope, which defines the linear relationship between the independent variable, \( x \), and the dependent, \( y \). The random deviations, \( \epsilon_i \)'s, allow the points to vary about the true line. (The estimated line is: \( \hat{y}_i = b_0 + b_1 x_i \).
2. The \( \epsilon_i \)'s have mean zero, \( \mu_\epsilon = 0 \).
3. The standard deviation of the \( \epsilon_i \)'s is constant, \( \sigma_\epsilon \) is not dependent on the \( x \)'s.
4. The \( \epsilon_i \)'s are independent of each other.
5. The \( \epsilon_i \)'s are normally distributed.

Combined, this say each of the \( \epsilon_i \)'s are independently, identically distributed \( \sim \text{N} (0, \sigma^2) \) or \( \epsilon \text{ iid} \sim \text{N}(0, \sigma^2) \). This means that the y's are also normal, and each \( y \sim \text{N}(\beta_0 + \beta_1 x, \sigma^2) \).

The method of Least Squares chooses estimates for \( \beta_0 \) and \( \beta_1 \), \( b_0 \) and \( b_1 \) respectively, which provide a line that minimizes the vertical distance between the points and the line. These distances are called residuals or errors and denoted by \( e_i \) where \( e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i) \) and the Method of Least Squares \( \text{min} \sum e_i^2 \).

**Simple Linear Regression (SLR) Inference** is just a hypothesis test to determine whether we have a statistically significant linear relationship between the x’s and y’s. The first step, of course, is to decide whether the assumptions have been met. The residuals plot help us in this (see the ‘lsinfer.doc’ for graphs):

1. If the plot is not just random scatter, i.e., there is a line or a curve or something that looks ‘predictable’, then the 1st or the 4th assumption has been violated.
2. If the plot is not centered around 0 (which probably means there’s an outlier), then the 2nd and maybe the 5th have been violated.
3. If the plot shows a wedge or fan shape, then the 3rd has been violated.

Now we can determine how good of a fit we have by running a hypothesis test. We test if the true slope is 0 or not (\( H_0: \beta_1 = 0 \) vs. \( H_A: \beta_1 \neq 0 \)). But this is just testing if the x’s are useful since a slope of 0 multiplied by the x’s means that the x’s fall out of the equation and we’re really just using the average \( y \), \( \bar{y} \), to predict the y’s. If we get a small p-value, we reject \( H_0 \) and conclude the line, with the x’s, is useful for predicting the y’s.

**Other Ways** to determine a ‘good’ fit:
1. Obviously, the correlation coefficient, \( r \), will also provide information about how linearly related the x’s and y’s are (see Bivariate Data handout). SLR instead uses the Coefficient of Determination, \( R^2 \), which is the proportion of the total variation of y that is explained with the least squares line(equation).

\[
R^2 = r^2 \quad \text{for Simple Linear Regression (this doesn’t hold for Multiple Regression)}
\]

2. **Standard Deviation about the Least Squares Line**, \( s_\epsilon \), is the typical amount by which an observation deviates from the least squares line.

Note: \( s_\epsilon \) is related to \( s \), and inversely related to \( R^2 \)

**NOTATION:**

- Sum of Squares Total= \( \text{SST} = \sum(y_i - \bar{y})^2 \) Note: if we divide by \( (n - 1) \), we would have the total variance of y.
- Sum of Squared Residuals = \( \text{SSR} \text{ (or SSE)} = \sum e_i^2 = \sum(y_i - \hat{y}_i)^2 \)
- Sum of Squares Model = \( \text{SSM} = \sum(\hat{y}_i - \bar{y})^2 \)
- \( \text{SST} = \text{SSE} + \text{SSM} \), so the better the fit the smaller the SSE and the closer SSM is to SST.

\[s_\epsilon = \sqrt{\frac{\text{SSE}}{df}} \text{ and } R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSM}}{\text{SST}}, \text{ so the better the fit, the smaller the } s_\epsilon \text{ and the larger the } R^2.\]