Lecture 5
Inference for Simple Linear Regression
STAT 432, Spring 2021

A **parameter** is a numerical characteristic of a population (e.g., the population mean height μ of all students at CSUEB)

Statistical inference refers to the process of using data collected from a sample to answer questions about population parameters.

- Point estimate: our best guess for the value of the population parameter (e.g., the sample mean height $\bar{x}$ of $n = 100$ randomly selected CSUEB students)
- Confidence interval: a plausible range of values for the population parameter
- Hypothesis test: is a specific value of the population parameter plausible?

Simple linear regression model for the population:

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_0 and β_1 are the population parameters (fixed and unkown)

Least squares line (estimated from the sample):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates (random, varies from sample to sample)

$1 - \alpha$ confidence interval for the slope β_1:

$$\hat{\beta}_1 \pm t_{\alpha/2;n-2} se(\hat{\beta}_1)$$

- $\hat{\beta}_1$ is the point estimate
- $t_{\alpha/2;n-2}$ is the t-critical value, with $n - 2$ degrees of freedom
- $se(\hat{\beta}_1)$ is the standard error
- $1 - \alpha$ is the confidence level (e.g., $\alpha = 0.05$ for a 95% confidence interval)

Hypothesis test for whether or not the slope β_1 is zero. We can also interpret this as a hypothesis test for whether or not there is a linear assoication between x and y.

$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

Test statistic:

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}; \quad \text{df=n-2}$$

Formulas for standard error computations (can use software for these computations):

- Residual standard error:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

- Standard error of $\hat{\beta}_1$:

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

- Recall the conditions for simple linear regression:
  - **Linearity**. The data should follow a linear trend.
  - **Constant variability**. The variability of points around the least squares line remains roughly constant.
  - **Normality**. The residuals should be approximately normally distributed with mean 0.
  - **Independence**. Values of the response variable are independent of each other.
- When computing confidence intervals or conducting hypothesis tests for linear regression models it is important that the conditions are adequately statistified.
- Residuals plots are a useful way to check the conditions (especially linearity and constant varibility). A histogram of the residuals can also be used to check normality.