Lecture 3:
Inference for Simple Linear Regression
STAT 452, 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 unknown)

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 association 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 rely on 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 $\hat{\beta}_1$:

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

Example

Going back to fitting a simple linear model for sale price, using above ground living area in square feet as a predictor.

```
> lm1 <- lm(Sale_Price ~ Gr_Liv_Area, data = ames)
> summary(lm1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13289.634   3269.703   4.064 4.94e-05 ***
Gr_Liv_Area   111.694      2.066  54.061  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56520 on 2928 degrees of freedom
Multiple R-squared:  0.4995,Adjusted R-squared:  0.4994
F-statistic:  2923 on 1 and 2928 DF,  p-value: < 2.2e-16
```

(a) Do the data provide strong evidence of a linear association between sale price and living area? State the null and alternative hypothesis, report the test statistic and p-value, and state your conclusion.

(b) Calculate a 95% confidence interval for the slope β_1. Note that there are $n = 2930$ properties (rows) in the data set.

In R, we can use the `confint()` function to compute confidence intervals for the regression parameters.

```
> confint(lm1, level = 0.95)
                2.5 %     97.5 %
(Intercept) 6878.4845 19700.7842
Gr_Liv_Area  107.6429   115.7451
```

Conditions for SLR

- **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.

- The linearity condition is the most important if we are primary concerned with prediction accuracy.
- The other three conditions are important if we are also concerned about making valid inferences (hypothesis testing and confidence intervals)
- Generally, simple linear regression is robust to mild violations of these conditions.

Does the scatter plot below show any violations of the SLR conditions?