Module 3: Intro to Simple Linear Regression

Jeffrey Woo

MSDS Program, University of Virginia

Welcome

- You can download this set of slides. Find with the materials for the live session in Module 3.8.
- Remind me to record the live session!
- Recommended: put yourself on mute unless you want to speak.
- There is a "raise hand" button for you. Click on "Reactions" in the panel at the bottom.

Agenda

- Q&A
- Some comments on Module 3.
- Small group discussion of guided question set
- Large group discussion of guided question set plus other questions that pop up

Q&A

- Any questions?
- Logistical or from the reading for module 3?

Big Picture

- Models have the following generic form: $y = f(x) + \epsilon$, where ϵ is independent from x with mean 0.
- In module 3, with SLR, we have $y = \beta_0 + \beta_1 x + \epsilon$.
  - We are implicitly assuming that x and y have a linear relationship (if any relationship exists), and they are both **quantitative**.
  - Relationship can be verified by a **scatterplot**.
  - $\epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

  i.i.d. means independent and identically distributed.

Big Picture

- In module 5, we will explore what to do if the relationship is **no longer linear**.
- In module 6, we will see what happens when we consider **more than 1 predictor** (multiple linear regression).
- In module 8, we will consider **categorical predictors**.
- In module 11, we will consider **binary response variable** (logistic regression).

The Term: Linear Models

- SLR model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- In matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  where
  - $\boldsymbol{y} = (y_1, \cdots, y_n)'$,
  - $\boldsymbol{X}$ is the $n \times 2$ **design matrix**,
  - $\boldsymbol{\beta} = (\beta_0, \beta_1)'$,
  - $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)'$

- Linear models are **linear in the parameters** β, not predictor(s).

ANOVA F test in SLR

- The hypothesis test assumes that if any relationship exists, it is linear (or the form of $f(x)$ is reasonable).
- If relationship is not linear (from scatterplot), should not perform this test.
- $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$.
- H_0 : there is no linear relationship (since regression line is flat), H_a : there is a linear relationship (since regression line is not flat).

ANOVA F test in SLR

- $F_0 = \frac{SS_R/1}{SS_{res}/n-2}$
- Notice that F_0 measures squared distances, so there is no direction associated with the test.
- If H_0 is true, F_0 follows $F_{1,n-2}$ distribution.
- Compare F_0 with critical value F^*, or p-value with α.
- F^*: type qf(1 - α, 1, n-2)
- p-value: type 1-pf(F_0, 1, n-2)
- In SLR, the ANOVA F test gives the same result as a 2-sided t test for the slope (see in Module 4).

Sum of Squares

- **Total sum of squares**, SS_T: $\sum_{i=1}^{n}(y_i - \bar{y})^2$. Total variation in response variable.
- **Regression sum of squares**, SS_R: $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$. Variance in response variable explained by predictors.
- **Residual sum of squares**, SS_{res}: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Variance in response variable not explained by predictors.

Sum of Squares

- $SS_T = SS_R + SS_{res}$
- Higher SS_R indicates better fit.
- Lower SS_{res} indicates better fit.
- With the same response variable, SS_T is constant. **Why?**

Degrees of Freedom

Degrees of freedom, df: number of values in the calculation of a sample statistic that are free to vary.

- For SS_T, its df is $n - 1$.
- For SS_{res}, its df is $n - 2$.
- For SS_R, its df is 1.

Mean Squares

Dividing each sum of squares with its corresponding degrees of freedom gives the corresponding mean squares.

- $MS_{res} = \frac{SS_{res}}{n-2}$
- $MS_R = \frac{SS_R}{1}$

It turns out the estimator of σ^2, the variance of the error terms (and the regression model), is MS_{res}.

Small Group Discussion

- Materials can be found under Module 3 Live session.
- Have the guided question set and corresponding data set open.
- Have R open.
- Recommended: have easy access to your notes, textbook, as well as the tutorial.
- You can see who your group members are. As well as some roles you will have in your small group. Roles will rotate each session.
- You will be evaluating your group members based on how prepared they are and their contribution to the small group discussion.

Large Group Discussion

What is due...

- Homework 3
- Submit via assignments
- Get prepared for next Tuesday's live session.

WARNING: Collab does not work well when you open multiple tabs. If you upload your HW with multiple tabs open, your submission is likely to screw up! You can always check if the submission went through correctly, and you can resubmit if needed.

Getting help...

- Office hours on Zoom: Mondays and Thursdays (GTA).
- Discussion forums.
- Email is to be used for more personal questions.
- We can also set up time to meet on Zoom if a discussion is needed.