Plan for the week

- **M: Simple Linear Regression**
  - Theory: Assumptions
  - Practice: Testing the Assumptions

- **W: Multiple Regression**
  - Data Transformations
  - Categorical Data
  - Model Checking: Cross-validation

- **F: Logistic Regression** (Classification!)
iClicker Question

How much do you care about theory vs. practice re: data science?
A. I love the theory
B. Everything in moderation
C. I only care about the practice
Properties of Estimators
What is an Estimator?

- A point estimator is a function that takes data (i.e., a sample) as input, and produces point estimates as output.
  - The sample mean function outputs the mean of its input.
  - Likewise, the sample variance function outputs the variance of its input.

- Note the nomenclature: a point estimator is a rule for generating point estimates.
  - “Average all the values in the sample” is a rule/function.
  - The average of all the values in a particular sample is an estimate.
Example: Normal RVs

- Assume $n$ i.i.d. (independent and identically distributed) normally-distributed random variables $X_1, X_2, \ldots, X_n$ with mean $\mu$ and standard deviation $\sigma$

- $\bar{X} = (1/n) \sum X_i$ (i.e., the sample mean function) is an estimator of the mean

- $\bar{x} = (1/n) \sum x_i$ (i.e., a sample mean) is an estimate of $\mu$
Example: Bernoulli RVs

- Assume $n$ i.i.d. (independent and identically distributed) Bernoulli random variables $X_1, X_2, \ldots, X_n$ with parameter $p$
- $\bar{X} = \frac{1}{n} \sum X_i$ (i.e., the sample proportion function) is an estimator of $p$
- $\bar{x} = \frac{1}{n} \sum x_i$ (i.e., a sample proportion) is an estimate of $p$
Evaluating Estimators (and Estimates)

- Any function of the data is an estimator!
- So how do we know we’ve got a good one?
- Desiderata:
  - In the limit, as the sample size tends to $\infty$, a consistent estimator converges to the model parameter it is estimating
  - An estimator is called unbiased if its expected value is the model parameter it is estimating
  - The efficiency of an estimator measures the quantity of data necessary to produce a certain quality estimate
Consistency

- An estimator is consistent if its value approaches its true value as the sample size tends to $\infty$.
- Consistent estimators become more accurate as the sample size increases.
- Is the sample mean a consistent estimator? Why or why not?
Bias

- Suppose $\theta^*$ is our model parameter, and $\theta$ is our estimator.
- The function $\theta$ applied to data $x \sim X$ yields a point estimate.
- $E_X[\theta(x)]$ is the expected value of the estimator, where the randomness comes from the randomness in the data.
- $\text{Bias}[\theta] = E_X[\theta(x)] - \theta^*$
- If $\text{Bias}[\theta] = 0$, then $\theta$ is called unbiased.
- If an estimator is unbiased, then on average it yields an accurate prediction of the model parameter.
Example

- Let $\bar{X} = \frac{1}{n} \sum X_i$ represent the sample mean estimator.
- $\text{Bias}[\bar{X}] = E_X[\bar{X}] - \mu = E_X[(1/n) \sum X_i] - \mu = (1/n) \sum E[X_i] - \mu = (1/n) \sum \mu = (1/n) n \mu - \mu = \mu - \mu = 0.$
- The sample mean estimator is unbiased.

- Let $\bar{X} = \frac{1}{n} \sum X_i$ represent the sample proportion estimator.
- $\text{Bias}[\bar{X}] = E_X[\bar{X}] - p = E_X[(1/n) \sum X_i] - p = (1/n) \sum E[X_i] - p = (1/n) \sum p = (1/n) np - p = p - p = 0.$
- The sample proportion estimator is unbiased.
Example, continued

- But $X_1$ and $X_2$ and so on are also unbiased.
- So why is $\bar{X}$ a better estimator than $X_1$ (or $X_2$, and so on)?
- Given two unbiased estimators, the preferred one is the one with lower variance (i.e., the more efficient one):
  - $\text{Var}(X_1) = \sigma$
  - $\text{Var}(\bar{X}) = \text{Var}(1/n \sum X_i)$
    - $= 1/n^2 \sum \text{Var}(X_i)$
    - $= \sigma/n$
  - $\text{Var}(\bar{X}) < \text{Var}(X_1)$
  - $\text{Var}(X_1) = p(1 - p)$
  - $\text{Var}(\bar{X}) = \text{Var}(1/n \sum X_i)$
    - $= 1/n^2 \sum \text{Var}(X_i)$
    - $= p(1 - p)/n$
  - $\text{Var}(\bar{X}) < \text{Var}(X_1)$
Best Linear Unbiased Estimators (BLUE)

- The sample mean is the most efficient estimator of the population mean, among all other weighted average that are also unbiased estimators.
- This result follows from the Gauss-Markov theorem, which states that the OLS estimators $b_0, b_1$ are the most efficient among all linear unbiased estimators, under standard assumptions.
Linear Regression, Revisited
Linear Model

- The distribution of $X$ is arbitrary.
- The distribution of $Y$ depends on that of $X = x$ in a linear fashion:
  - $Y$ is distributed with mean $\beta_0 + \beta_1 x$.
- Find $\beta_0$ and $\beta_1$ that minimize the mean squared error:
  - $(\beta_0, \beta_1)$ s.t $E[(Y - \beta_0 + \beta_1 x)^2 \mid X = x]$ is minimized
Linear Model (cont’d)

- The distribution of $X$ is arbitrary.
- The distribution of $Y$ depends on that of $X = x$ in a linear fashion:
  - $Y$ is distributed with mean $\beta_0 + \beta_1 x$.
- Find $\beta_0$ and $\beta_1$ that minimize the mean squared error:
  - $(\beta_0, \beta_1)$ s.t $\text{E}[(Y - \beta_0 + \beta_1 x)^2 | X = x]$ is minimized
- Solve as usual with calculus:
  - Take partial derivatives, and set them equal to zero.
- Out pops:
  - $\beta_0 = \text{E}[Y] - \beta_1 \text{E}[X]$
  - $\beta_1 = \text{Cov}[X, Y] / \text{Var}[X] = \text{Corr}[X, Y] \sigma_Y / \sigma_X$
  - $b_0 = \bar{y} - b_1 \bar{x}$
  - $b_1 = r_{xy} (s_{yy} / s_{xx})$
- The same answer as before—in expectation!
Linear Model: Additional Assumptions

- The distribution of $X$ is arbitrary.
- The distribution of $Y$ depends on that of $X = x$ in a linear fashion:
  - $Y$ is distributed with mean $\beta_0 + \beta_1 x$.
- The noise in the distribution can be described by random variables $\varepsilon_i$.
  - Linear model: given $X = x$, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, for all $1 \leq i \leq n$.
  - The conditional expectation of the noise terms is 0: $E[\varepsilon_i \mid X = x] = 0$.
  - The conditional variance of the noise terms is constant: $\text{Var}[\varepsilon_i \mid X = x] = \sigma^2$.
  - The noise terms are uncorrelated with themselves (i.e., no time-series data): $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0$, for all $i \neq j$.
- Under these assumptions, $b_0$ and $b_1$ are unbiased and consistent estimators.
  - Unbiased, because the conditional expectation of the noise terms is 0.
  - Consistent, by the law of large numbers, and other assumptions of the model.