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 μ and standard deviation σ.
- The function $\bar{X}$ that maps a sample $x_1, x_2, \ldots, x_n$ drawn i.i.d. from $X_1, X_2, \ldots, X_n$ to $(1/n) \sum x_i$ (i.e., the sample mean function) is an estimator of the mean μ.

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*.
- The sum of these Bernoulli RVs is a binomial RV with mean *np*.
- The function *p* that maps a sample $x_1, x_2, \ldots, x_n$ drawn i.i.d. from $X_1, X_2, \ldots, X_n$ to $(1/n) \sum x_i$ (i.e., the sample proportion function)
  is an estimator of $np/n = p$.

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 ∞, 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 ∞.
- Consistent estimators become more accurate as the sample size increases.
- Is the sample mean a consistent estimator? Why or why not?

Bias

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

Example: Sample mean

- Let $\bar{X} = (1/n) \sum x_i$ represent the sample mean estimator.
- $\mathrm{Bias}[\bar{X}, \mu] = \mathrm{E}_{x \sim X}[\bar{X}] - \mu = \mathrm{E}_{x \sim X}[(1/n) \sum x_i] - \mu = (1/n) \sum \mathrm{E}_{x \sim X}[x_i] - \mu = (1/n) \sum \mu = (1/n)\, n\mu - \mu = \mu - \mu = 0$.
- Since μ was arbitrary, the sample mean estimator is unbiased.

Example: Sample proportion

- Let $\bar{X} = (1/n) \sum x_i$ represent the sample proportion estimator.
- $\text{Bias}[\bar{X}, p] = E_{x \sim X}[\bar{X}] - p = E_{x \sim X}[(1/n) \sum x_i] - p = (1/n) \sum E_{x \sim X}[x_i] - p = (1/n) \sum p = (1/n)\, np - p = p - p = 0.$
- Since p was arbitrary, the sample proportion estimator is unbiased.

Example: Sample variance

- The sample variance is not unbiased.
- But we can make it unbiased by dividing by n-1 instead of n.
    - Proof

Examples, 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):
  - $\text{Var}(X_1) = \sigma^2$
  - $\text{Var}(\bar{X}) = \text{Var}(1/n \sum X_i)$
    $= (1/n^2) \sum \text{Var}(X_i)$
    $= (1/n^2)(n\sigma^2)$
    $= \sigma^2/n$
  - $\text{Var}(\bar{X}) < \text{Var}(X_1)$
  - $\text{Var}(X_1) = p(1 - p)$
  - $\text{Var}(\bar{X}) = \text{Var}(1/n \sum X_i)$
    $= (1/n^2) \sum \text{Var}(X_i)$
    $= (1/n^2)(np(1 - p))$
    $= (p(1 - p))/n$
  - $\text{Var}(\bar{X}) < \text{Var}(X_1)$

Best Linear Unbiased Estimators (BLUE)

- The sample mean is the most efficient estimator of the population mean, among all other weighted averages 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.

Extras

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 β_0 and β_1 that minimize the mean squared error:
  - (β_0, β_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 β_0 and β_1 that minimize the mean squared error:
  - (β_0, β_1) s.t $E[(Y - \beta_0 + \beta_1 x)^2 \mid X = x]$ is minimized
- Solve as usual with calculus:
  - Take partial derivatives, and set them equal to zero.
- Out pops:
  - $\beta_0 = E[Y] - \beta_1 E[X]$
  - $\beta_1 = Cov[X, Y] / Var[X] = Corr[X, Y]\ \sigma_Y / \sigma_X$
  - $b_0 = \overline{y} - b_1 \overline{x}$
  - $b_1 = r_{XY}\ (s_{YY} / s_{XX})$
- The same answer as before—in expectation!

The Noise

- Given $X = x$, Y is distributed with mean $\beta_0 + \beta_1 x$.
  - Given $X = x_i$, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, for all $1 \leq i \leq n$.
  - This noise is described by the random variables ε_i.
  - It represents aspects of Y that are not determined by X.

- Assumptions
  - The conditional expectation of the noise terms is 0: $E[\varepsilon_i \mid X = x_i] = 0$
    (because any non-zero conditional expectation could be built into the model).
  - The conditional variance of the noise terms is constant: $Var[\varepsilon_i \mid X = x_i] = \sigma^2$.
  - The noise terms are uncorrelated with one another: $Cov[\varepsilon_i = \varepsilon_j] = 0$, 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.

The Noise (cont'd)

- Given $X = x$, Y is distributed with mean $\beta_0 + \beta_1 x$.
  - Given $X = x_i$, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, for all $1 \leq i \leq n$.
  - This noise is described by normal random variables ε_i.
  - It represents aspects of Y that are not determined by X.
- Assumptions
  - The conditional expectation of the noise terms is 0: $E[\varepsilon_i \mid X = x_i] = 0$ (because any non-zero conditional expectation could be built into the model).
  - The conditional variance of the noise terms is constant: $\mathrm{Var}[\varepsilon_i \mid X = x_i] = \sigma^2$.
  - The noise terms are independent of one another.
- Under these assumptions, the b_0 and b_1 are MLE estimators.