Biostatistics Preparatory Course: Methods and Computing

Lecture 6

Simulations

Recap / Warm-up: Linear Regression

In the group exercise 2, we were given the following model:

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

where Y was birthweight, X_1 was smoking status, and X_2 was mother's weight gain.

- Why might β_3 be of interest?

Recap / Warm-up: Linear Regression

In the group exercise 2, we were given the following model:

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

where Y was birthweight, X_1 was smoking status, and X_2 was mother's weight gain.

- Why might β_3 be of interest?
  - If you believe that the effect of mother's weight gain varies within levels of smoking status
- What are the interpretations of β_1 and β_2?

Recap / Warm-up: Linear Regression

In the group exercise 2, we were given the following model:

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

where Y was birthweight, X_1 was smoking status, and X_2 was mother's weight gain.

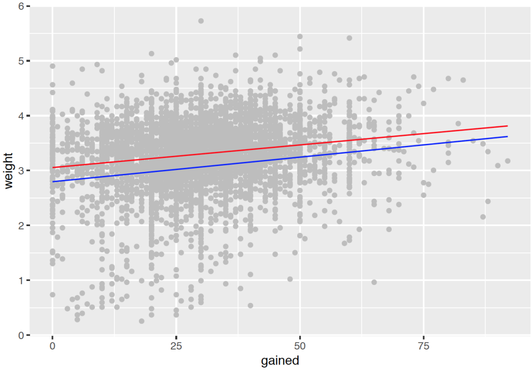
- Why might β_3 be of interest?
  - If you believe that the effect of mother's weight gain varies within levels of smoking status
- What are the interpretations of β_1 and β_2?
  - The mean change in birthweight comparing smokers to non-smokers among mother's who did not gain weight
  - The mean change in birthweight corresponding to a one unit change in mother's weight gain among non-smokers

Recap / Warm-up: Linear Regression

$$E[Y|X_1 = 1, X_2] = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2$$
$$E[Y|X_1 = 0, X_2] = \hat{\beta}_0 + \hat{\beta}_2 X_2$$

Recap / Warm-up: Linear Regression

$$E[Y|X_1 = 1, X_2] = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2$$
$$E[Y|X_1 = 0, X_2] = \hat{\beta}_0 + \hat{\beta}_2 X_2$$



Simulations studies

1. What is a simulation?
   - Numerical technique to conduct experiments on a computer
   - In statistics, we typically care about 'Monte Carlo' (MC) simulations which involve random sampling from probability distributions
2. Why bother?
   - When developing a new method, it is important to establish its properties so that it can be used in practice
   - **Case I**: Analytical derivations of properties are not always possible
     - It is often feasible to obtain large sample approximations, but evaluation of the approximation in finite samples is necessary
   - **Case II**: If you can derive analytic results, they usually require assumptions
     - What are the properties of the method when various conditions are violated?

Important terms

- An **unbiased estimator** for some parameter means that the expected value of the estimator is equal to the parameter
- A confidence interval has **nominal coverage** if it covers the true value of the parameter the correct proportion of times
- The **size** of a hypothesis test is equal to the probability of rejecting the null hypothesis given that the null is true
- The **power** of a hypothesis test is equal to the probability of rejecting the null hypothesis given that the null is false

MC Simulations: The usual questions

- Under what conditions is an estimator unbiased?
  ex. Suppose the data is generated according to

$$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

  but I fit the model $y = \alpha_0 + \alpha_1 x_1$. When is $\hat{\beta}_1$ unbiased for α_1?

MC Simulations: The usual questions

- Under what conditions is an estimator unbiased?
  ex. Suppose the data is generated according to
  $$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$
  but I fit the model $y = \alpha_0 + \alpha_1 x_1$. When is $\hat{\beta}_1$ unbiased for α_1?
- How does the estimator compare to other estimators? What is its sampling variability?
  ex. Suppose the data is generated according to
  $$y \sim \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \epsilon$$
  with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 I$. How do the OLS estimators compare to
  $$\hat{\alpha}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i^* - \bar{x}^*)}{\sum_{i=1}^n (x_i - \bar{x})(x_i^* - \bar{x}^*)} \quad \text{and} \quad \hat{\alpha}_0 = \frac{\sum_{i=1}^n y_i/x_i}{\sum_{i=1}^n 1/x_i} - \hat{\alpha}_1 \frac{n}{\sum_{i=1}^n 1/x_i}$$
  where $\bar{x}^*$ is mean of $x_i^* = 1/x_i$?

MC Simulations: The usual questions

- Does a confidence interval procedure attain nominal coverage?
  ex.
  - The sum of n independent Bernoulli trials with common success probability is distributed according to $Bin(n, \pi)$
  - The MLE for π is $\hat{\pi} = \frac{X}{n}$ where X is the observed number of successes
  - The *Wald 95% Confidence Interval* for π is given by:

$$\left(\hat{\pi} - z_{0.975}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{0.975}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)$$

  - The *Score 95% Confidence Interval* for π is given by:

$$\hat{\pi}\left(\frac{n}{n+z_{0.975}^2}\right) + \frac{1}{2}\left(\frac{z_{0.975}^2}{n+z_{0.975}^2}\right) \pm$$

$$z_{0.975}\sqrt{\frac{1}{n+z_{0.975}^2}\left[\hat{\pi}(1-\hat{\pi})\left(\frac{n}{n+z_{0.975}^2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z_{0.975}^2}{n+z_{0.975}^2}\right)\right]}$$

How does the coverage compare for both intervals as we increase n and vary p?

MC Simulations: The usual questions

- Does a hypothesis testing procedure achieve the specified size? If so, what is the power like? How does it compare to alternative procedures?

  ex. Consider the one sample t-test for

  $$H_0 : \mu = 0 \quad \text{vs.} \quad H_A : \mu \neq 0$$

  How does the power vary when the data is generated under some alternative hypothesis $\mu \neq \mu_0$?

How does Monte Carlo simulation help to answer these questions?

MC Simulations: Intuition

- An estimator/test statistic has a true sampling distribution under some set of conditions
- We'd like to know the true sampling distribution so we can answer the questions on the previous slide but...
  1. The (finite sample) derivation is difficult

     and/or

  2. We'd like to see how well the method holds up when assumptions are violated

So, we approximate the sampling distribution of an estimator/test statistic under a particular set of conditions through simulation

How to Approximate the Sampling Distribution

- Generate B independent data sets according to the data generating process
- Compute the value of the estimator/test statistic $T(data)$ for each data set $\rightarrow \{T_1, \ldots, T_B\}$

If b is large enough, summary statistics using $\{T_1, \ldots, T_b\}$ should be good approximations to the true sampling properties of the estimator/test statistic under the specified conditions

ex. T_b is the value of T from the b^{th} data set, $b = 1, \ldots, B$

- The empirical mean computed with the B data sets is an estimate of the true mean of the sampling distribution of the estimator
- The empirical standard error computed with the B data sets is an estimate of the true standard deviation of the sampling distribution of the estimator

How can you assess the following?

- An **unbiased estimator** for some parameter means that the expected value of the estimator is equal to the parameter

How can you assess the following?

- An **unbiased estimator** for some parameter means that the expected value of the estimator is equal to the parameter
  - In numerous samples generated from the truth, take the mean of the estimated parameters. Is it close to the true value of the parameter?

How can you assess the following?

- An **unbiased estimator** for some parameter means that the expected value of the estimator is equal to the parameter
  - In numerous samples generated from the truth, take the mean of the estimated parameters. Is it close to the true value of the parameter?
- A confidence interval has **nominal coverage** if it covers the true value of the parameter the correct proportion of times

How can you assess the following?

- An **unbiased estimator** for some parameter means that the expected value of the estimator is equal to the parameter
  - In numerous samples generated from the truth, take the mean of the estimated parameters. Is it close to the true value of the parameter?
- A confidence interval has **nominal coverage** if it covers the true value of the parameter the correct proportion of times
  - In numerous samples generated from the truth, calculate the confidence interval, how often does it cover the true value of the parameter?

How can you assess the following?

- An **unbiased estimator** for some parameter means that the expected value of the estimator is equal to the parameter
  - In numerous samples generated from the truth, take the mean of the estimated parameters. Is it close to the true value of the parameter?
- A confidence interval has **nominal coverage** if it covers the true value of the parameter the correct proportion of times
  - In numerous samples generated from the truth, calculate the confidence interval, how often does it cover the true value of the parameter?
- The **size** of a hypothesis test is equal to the probability of rejecting the null hypothesis given that the null is true

How can you assess the following?

- An **unbiased estimator** for some parameter means that the expected value of the estimator is equal to the parameter
  - In numerous samples generated from the truth, take the mean of the estimated parameters. Is it close to the true value of the parameter?
- A confidence interval has **nominal coverage** if it covers the true value of the parameter the correct proportion of times
  - In numerous samples generated from the truth, calculate the confidence interval, how often does it cover the true value of the parameter?
- The **size** of a hypothesis test is equal to the probability of rejecting the null hypothesis given that the null is true
  - In numerous samples generated from the truth, conduct the hypothesis test, how often does it incorrectly reject the null?

How can you assess the following?

- An **unbiased estimator** for some parameter means that the expected value of the estimator is equal to the parameter
  - In numerous samples generated from the truth, take the mean of the estimated parameters. Is it close to the true value of the parameter?
- A confidence interval has **nominal coverage** if it covers the true value of the parameter the correct proportion of times
  - In numerous samples generated from the truth, calculate the confidence interval, how often does it cover the true value of the parameter?
- The **size** of a hypothesis test is equal to the probability of rejecting the null hypothesis given that the null is true
  - In numerous samples generated from the truth, conduct the hypothesis test, how often does it incorrectly reject the null?
- The **power** of a hypothesis test is equal to the probability of rejecting the null hypothesis given that the null is false

How can you assess the following?

- An **unbiased estimator** for some parameter means that the expected value of the estimator is equal to the parameter
    - In numerous samples generated from the truth, take the mean of the estimated parameters. Is it close to the true value of the parameter?
- A confidence interval has **nominal coverage** if it covers the true value of the parameter the correct proportion of times
    - In numerous samples generated from the truth, calculate the confidence interval, how often does it cover the true value of the parameter?
- The **size** of a hypothesis test is equal to the probability of rejecting the null hypothesis given that the null is true
    - In numerous samples generated from the truth, conduct the hypothesis test, how often does it incorrectly reject the null?
- The **power** of a hypothesis test is equal to the probability of rejecting the null hypothesis given that the null is false
    - In numerous samples generated from the truth, conduct the hypothesis test, how often does it correctly reject the null?

Commonly reported quantities

Your simulation study has B replicates for some estimator T of θ.

- Simulation bias

$$\text{bias}(T) = \frac{1}{b}\sum_{b=1}^{B} T_b - \theta$$

- Simulation relative bias

$$\text{relative bias}(T) = \frac{\text{bias(T)}}{\theta}$$

- Simulation standard deviation

$$\text{sd}(T) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(T_b - \overline{T})^2}$$

- Simulation mean squared error

$$\text{MSE}(T) = \text{bias(T)}^2 + \text{sd}(T)^2$$

Although omitted, you may also be interested in reporting the empirical coverage for confidence interval, power, or size.

Tips for Running Your Own Simulation Studies

1. Setting parameter values:
   - First run your code under a favorable setting (make sure it works)
   - Then choose parameter values that will challenge your method
2. Don't make B too large to start (≈ 500)
3. Save all the estimates and not just the summary statistics
4. Set the seed
5. Document the code (i.e. comments)
6. Keep track of the versions of the code you use (i.e. use GitHub)
7. If you use Rmarkdown, use the cache=TRUE preamble
   - Your code will only be knitted/run the first time or anytime after it updated. Saves time!

Tips for Presenting Results

1. Only present what is interesting
   - ex. If the bias is small, just make a comment in the text rather than making a table
   - ex. If two parameter settings are similar, you don't need to include both
   - In homework assignments, you will typically be told what to report
2. Make the results easy for the reader to understand
   - Columns meant to be compared should be side-by-side
   - Make a graph if possible

Example: Population Mean

- Suppose you are interested in comparing the properties of the following 3 estimators for the mean μ for n iid draws $X_1, \ldots X_n$ with $X_i \sim f(x)$
  1. Sample mean, T^1
  2. Sample 15% trimmed mean, T^2
  3. Sample median, T^3

How would you expect the estimators to compare if $X_i \sim N(1, 16)$?