Computational Statistics Final Paper

HUDM 6026

Spring, 2023

The final project will be a write-up based on simulation with simple linear regression to study the properties of some estimators with computationally intensive methods. The primary goals are to create data generation functions motivated by a real data set and then to use those functions to evaluate statistical estimators via Monte Carlo simulation and resampling methods.

Technical elements of the final project.

1. **Select a motivating data set.** Select data to work with that have a numeric outcome variable and a numeric predictor variable. You can choose data from open access peer reviewed publications such as those in PLOS ONE or from machine learning repositories or Kaggle competitions that have released publicly available data. Try to find data that interest you.
2. **Data generation.** You will be simulating data from a simple linear regression model so you will need to specify an intercept, a slope, a distribution for the predictor, and the variance of the random normal error term. That is,
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$
where $\epsilon_i \sim N(0, \sigma^2)$. Put these elements together to create a function called `dat_gen()` that takes as argument the sample size n and produces as output a data frame containing the predictor, X, and the outcome, Y. Again, though, note that the data generation details should be motivated by the real data analysis from step (1).
3. **Estimators.** Write a function called `reg()` "from scratch" that takes as input a data frame generated from your `dat_gen()` function and produces as output estimates of the slope on X (beta1) and the error variance (sigma^2) from the following estimators.
    a. 1st estimator for beta1: least squares estimator $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
    b. 2nd estimator for beta1: alternative estimator $\beta_1^a = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i-\bar{y}}{x_i-\bar{x}}$
    c. 1st estimator for sigma^2: usual estimator $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$
    d. 2nd estimator for sigma^2: alternative estimator $\sigma_a^2 = \frac{SSE}{n}$
4. **Monte Carlo simulation.** Run a Monte Carlo simulation with sample size n = 40 by generating a large number R of replications (e.g., R = 1000 or more) from `dat_gen(n = 40)`. For each replication, apply your function `reg()` and save the results. Create histograms of sampling distributions with density plots for each of the four statistics reported by `reg()`. Calculate and report the mean and the standard error of the mean for each parameter and for each estimator based on the Monte Carlo replications. Estimate the bias, variance and MSE of each estimator.
5. **Resampling - bootstrap**. Now, instead of using the data generation function over and over, let's suppose that we only have a single data set. Set a seed and generate a single data set with n = 40 from `dat_gen()`. Use bootstrap resampling to

generate a large number B of bootstrap replications (e.g., B = 500 or more). For each bootstrap replication apply your function `reg()` and record results. As before, create histogram/density plots of parameter estimates. Estimate the bias, variance and MSE of each estimator based on the bootstrap replications.

6. **Resampling – jackknife**. Using the same data set that was generated for the bootstrap work above, use the jackknife to estimate the bias and variance of each estimator.

Structure of the write-up.

1. This is a group project. The final paper will be due by 11:59 pm on Tuesday May 9th. All members of the group should participate. When you submit the final paper, include on the cover page in addition to names of group members and title of report a brief summary of each group member's contributions to the technical part, coding part, and write-up.
2. My preference is that the write up be done in R Markdown. If you wish to use another format, please discuss with me via email.
3. Begin your write-up by walking the reader through the six technical points outlined above. In this part you should use a combination of written paragraphs, commented code, and plots and numerical summaries to communicate results.
4. Can you say anything about the theoretically-based bias, variance, or MSE of any of these estimators? If so, what and how do you know?
5. How do results from the Monte Carlo, bootstrap, and jackknife procedures compare with theory and with each other?
6. Comment on which of the three computational procedures is the most accurate in estimating bias, variance, and MSE of these estimators and explain why you believe so.
7. Conclude with some discussion about the pros and cons of Monte Carlo vs bootstrap vs jackknife for learning about the properties of estimators.