Week 6 - Midterm Assignment: Simulation Project

STAT 420, Summer 2023, D. Unger

Directions

This is an **individual** project similar to homework assignments. However, because of the project structure of your submission, collaboration should be more limited than an homework assignment to protect yourself from duplication. Discussion of question intent, coding problems/issues, and project administration may be discussed on the message board on a limited basis. However, sharing, copying, or providing any part of this project to another student is an infraction of the University's rules on academic integrity. Any violation will be punished as severely as possible.

- Your project must be submitted through Coursera. You are required to upload one `.zip` file, named `yourNetID-sim-proj.zip` , which contains:
  - Your RMarkdown file which should be saved as `yourNetID-sim-proj.Rmd` .
  - The result of knitting your RMarkdown file as `yourNetID-sim-proj.html` .
  - Any outside data provided as a `.csv` file. (In this case, `study_1.csv` and `study_2.csv` .)
- Your `.Rmd` file should be written such that, when stored in a folder with any data you are asked to import, it will knit properly without modification. If your `.zip` file is organized properly, this should not be an issue.
- Include your name and NetID in the final document, not only in your filenames.

This project consists of **three** simulation studies. MCS-DS and other campus graduate students must complete all three studies. Campus undergraduate students need to complete on the first two studies. There is no extra credit for undergraduates who complete the third study.

Unlike a homework assignment, these "exercises" are not broken down into parts (e.g., a, b, c), and so your analysis and report submission will not be partitioned into parts. Instead, your document should be organized more like a true project report, and it should use the overall format:

- Simulation Study 1
- Simulation Study 2
- Simulation Study 3 (*MCS and graduates only*)

Within each of the simulation studies, you should use the format:

- Introduction
- Methods
- Results
- Discussion

The **introduction** section should relay what you are attempting to accomplish. It should provide enough background to your work such that a reader would not need this directions document to understand what you are doing. Basically, assume the reader is mostly familiar with the concepts from the course, but not this project. [For the Midterm Assignment, the Introduction section is allowed to simply be the exercise statements as I have typed them later in this file. Yes, a direct copy.]

The **methods** section should contain the majority of your “work.” This section will contain the bulk of the `R` code that is used to generate the results. Your `R` code is not expected to be perfect idiomatic `R`, but it is expected to be understood by a reader without too much effort. Use RMarkdown and code comments to your advantage to explain your code if needed. [For the Midterm Assignment, you may type a sentence or two explaining what you are attempting to do and then the code chunk that does it. Repeat as necessary to tell a coherent story about your method. That is, it would be clear to display your code in a few smaller chunks explaining along the way, as opposed to one giant chunk with a large paragraph trying to describe the whole thing at once.]

The **results** section should contain numerical or graphical summaries of your results as they pertain to the goal of each study. [For the Midterm Assignment, while the code chunks appear in Methods, the actual plots, numeric tables, etc. would appear in Results.]

The **discussion** section should contain discussion of your results. The discussion section should contain discussion of your results. Potential topics for discussion are suggested at the end of each simulation study section, but they are not meant to be an exhaustive list. These simulation studies are meant to be explorations into the principles of statistical modeling, so do not limit your responses to short, closed form answers as you do in homework assignments. Use the potential discussion questions as a starting point for your response. [For the Midterm Assignment, This is where you may give your summary of what you observe in the Results. The “Potential Topics” are meant to get your thoughts flowing and give you ideas about what to discuss. This section will be just your typed responses with no new code or results.]

- Your resulting `.html` file will be considered a self-contained “report,” which is the material that will determine the majority of your grade. Be sure to visibly include all `R` code and output that is *relevant*. (You should not include irrelevant code you tried that resulted in error or did not answer the question correctly.)
- Grading will be based on a combination of completing the required tasks, discussion of results, `R` usage, RMarkdown usage, and neatness and organization. For full details see the provided rubric.
- At the beginning of *each* of the three simulation studies, set a seed equal to your birthday, as is done on homework. (It should be the first code run for each study.) These should be the only three times you set a seed.

```
birthday = 18760613
set.seed(birthday)
```

One Final Note: The simulations in this Midterm Assignment require combinations of several factors that result in a lot of computation. For example, in Simulation Study 1, the response vector that you generate will have $2(models) \times 3(sigmas) \times 2000(sims) = 12000$ simulated values. Expect that `R` may take longer to compile than your typical weekly Homework Assignment, especially your final report. I'll suggest two tips.

- Tip 1: Make a separate `R` script, notebook, or Rmd for each Simulation Study. Thus, while you are working through one study, you are not making `R` try to compile the code for the other studies that you are not actively working on.
- Tip 2: Start with a smaller number of simulations until you work out the bugs. For example, 200 simulations in Study 1 instead of 2000. The end Results will not be correct initially, but `R` will compile faster while you are still figuring out the code in the Methods. Once you've nailed that, update your code with the correct number of simulations.

Good luck!

Simulation Study 1: Significance of Regression

In this simulation study we will investigate the significance of regression test. We will simulate from two different models:

1. The **"significant"** model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and

- $\beta_0 = 3$,
- $\beta_1 = 1$,
- $\beta_2 = 1$,
- $\beta_3 = 1$.

2. The **"non-significant"** model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and

- $\beta_0 = 3$,
- $\beta_1 = 0$,
- $\beta_2 = 0$,
- $\beta_3 = 0$.

For both, we will consider a sample size of 25 and three possible levels of noise. That is, three values of σ.

- $n = 25$
- $\sigma \in (1, 5, 10)$

Use simulation to obtain an empirical distribution for each of the following values, for each of the three values of σ, for both models.

- The F **statistic** for the significance of regression test.
- The **p-value** for the significance of regression test
- R^2

For each model and σ combination, use 2000 simulations. For each simulation, fit a regression model of the same form used to perform the simulation.

Use the data found in `study_1.csv` (study_1.csv) for the values of the predictors. These should be kept constant for the entirety of this study. The `y` values in this data are a blank placeholder.

Done correctly, you will have simulated the `y` vector $2(models) \times 3(sigmas) \times 2000(sims) = 12000$ times.

Potential discussions:

- Do we know the true distribution of any of these values?
- How do the empirical distributions from the simulations compare to the true distributions? (You could consider adding a curve for the true distributions if you know them.)
- How are each of the F statistic, the p-value, and R^2 related to σ? Are any of those relationships the same for the significant and non-significant models?

An additional tip:

- Organize the plots in a grid for easy comparison. For example, a 1×3 row of F statistic plots as σ changes, then a 1×3 row of p-value plots as σ changes, followed by a similar row for the R^2 values. Consider a similar setup for the values attributed to the significant model and then again for the nonsignificant model.

Simulation Study 2: Using RMSE for Selection?

In homework we saw how Test RMSE can be used to select the “best” model. In this simulation study we will investigate how well this procedure works. Since splitting the data is random, we don’t expect it to work correctly each time. We could get unlucky. But averaged over many attempts, we should expect it to select the appropriate model.

We will simulate from the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and

- $\beta_0 = 0$,
- $\beta_1 = 3$,
- $\beta_2 = -4$,
- $\beta_3 = 1.6$,
- $\beta_4 = -1.1$,
- $\beta_5 = 0.7$,
- $\beta_6 = 0.5$.

We will consider a sample size of 500 and three possible levels of noise. That is, three values of σ.

- $n = 500$
- $\sigma \in (1, 2, 4)$

Use the data found in `study_2.csv` (study_2.csv) for the values of the predictors. These should be kept constant for the entirety of this study. The `y` values in this data are a blank placeholder.

Each time you simulate the data, randomly split the data into train and test sets of equal sizes (250 observations for training, 250 observations for testing).

For each, fit **nine** models, with forms:

- `y ~ x1`
- `y ~ x1 + x2`
- `y ~ x1 + x2 + x3`
- `y ~ x1 + x2 + x3 + x4`
- `y ~ x1 + x2 + x3 + x4 + x5`
- `y ~ x1 + x2 + x3 + x4 + x5 + x6`, the correct form of the model as noted above
- `y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7`
- `y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8`
- `y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9`

For each model, calculate Train and Test RMSE.

$$\text{RMSE}(\text{model}, \text{data}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Repeat this process with 1000 simulations for each of the 3 values of σ. For each value of σ, create a plot that shows how average Train RMSE and average Test RMSE changes as a function of model size. Also show the number of times the model of each size was chosen for each value of σ.

Done correctly, you will have simulated the y vector $3 \times 1000 = 3000$ times. You will have fit $9 \times 3 \times 1000 = 27000$ models. A minimal result would use 3 plots. Additional plots may also be useful.

Potential discussions:

- Does the method **always** select the correct model? On average, does is select the correct model?
- How does the level of noise affect the results?

An additional tip:

- To address the second discussion topic, consider making a line graph for the RMSE values at each level of σ. Within a single plot for a given σ, one line could correspond to the training data and the other to the test data.

Simulation Study 3: Power

In this simulation study we will investigate the **power** of the significance of regression test for simple linear regression.

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

Recall, we had defined the *significance* level, α, to be the probability of a Type I error.

$$\alpha = P[\text{Reject } H_0 \mid H_0 \text{ True}] = P[\text{Type I Error}]$$

Similarly, the probability of a Type II error is often denoted using β; however, this should not be confused with a regression parameter.

$$\beta = P[\text{Fail to Reject } H_0 \mid H_1 \text{ True}] = P[\text{Type II Error}]$$

Power is the probability of rejecting the null hypothesis when the null is not true, that is, the alternative is true and β_1 is non-zero.

$$\text{Power} = 1 - \beta = P[\text{Reject } H_0 \mid H_1 \text{ True}]$$

Essentially, power is the probability that a signal of a particular strength will be detected. Many things affect the power of a test. In this case, some of those are:

- Sample Size, n
- Signal Strength, β_1
- Noise Level, σ
- Significance Level, α

We'll investigate the first three.

To do so we will simulate from the model

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

where $\epsilon_i \sim N(0, \sigma^2)$.

For simplicity, we will let $\beta_0 = 0$, thus β_1 is essentially controlling the amount of “signal.” We will then consider different signals, noises, and sample sizes:

- $\beta_1 \in (-2, -1.9, -1.8, \ldots, -0.1, 0, 0.1, 0.2, 0.3, \ldots 1.9, 2)$
- $\sigma \in (1, 2, 4)$
- $n \in (10, 20, 30)$

We will hold the significance level constant at $\alpha = 0.05$.

Use the following code to generate the predictor values, `x` : values for different sample sizes.

```
x_values = seq(0, 5, length = n)
```

For each possible β_1 and σ combination, simulate from the true model at least 1000 times. Each time, perform the significance of the regression test. To estimate the power with these simulations, and some α, use

$$\hat{\text{Power}} = \hat{P}[\text{Reject } H_0 \mid H_1 \text{ True}] = \frac{\text{\# Tests Rejected}}{\text{\# Simulations}}$$

It is *possible* to derive an expression for power mathematically, but often this is difficult, so instead, we rely on simulation.

Create three plots, one for each value of σ. Within each of these plots, add a “power curve” for each value of n that shows how power is affected by signal strength, β_1.

Potential discussions:

- How do n, β_1, and σ affect power? Consider additional plots to demonstrate these effects.
- Are 1000 simulations sufficient?

An additional tip:

- Search online for examples of power curves to give you inspiration for how you might construct your own plots here. You’ll find both two-sided and one-sided power curves. Based on the way you’re asked to construct the β_1 vector, you should be able to figure out which type is appropriate here.