I created a dataset using np.random.normal (default mean and standard deviations of 0 and 1) and np.random.unform (default min of 0 and max of 1). Then I created some coefficients based on data from (<https://digitalcommons.coastal.edu/cgi/viewcontent.cgi?article=1220&context=etd>), which explored how the pandemic influenced enrollment in two-year colleges. I utilized the coefficients for their constant, gender, enrollment intensity, campus setting, and tuition increase, I chose to square enrollment intensity as I imagine there could be a potential quadratic relationship between intensity and enrollment numbers. I assigned the gender and campus setting variables to random uniform, and enrollment intensity and tuition increase to the random normal distribution. The reasoning was that these variables would align with these distributions well.

Then I utilized train\_test\_split to create a .30 proportion of testing, Using the training and testing data, I calculate the MSE for both. The MSE was slightly higher for the test data (), compared to training data () but they were still very close. This provides some evidence that the model selection performed well with our artificial “new” data.

For the bootstrapping section, I created a For loop, calculating a loop of 10 creating new samples with replacement using the full 10,000 cases. I ran OLS regression and placed the coefficients in a data frame. The average across all the samples were computed, along with the standard deviation.

To answer question 8, the results were very close to the estimates in our original OLS validation training and testing. I have placed them side by side in Table 1 below. Overall, this makes us more confident in our original results provided from validation and indicates that our model is less susceptible to sampling variation.

**Discussion Prompts:**

1. **Model Selection vs. Model Evaluation**  
   Explain, in your own words, the difference between *model selection* and *model evaluation*. Why is it important not to use the same data (or the same folds) for both?

Model selection is choosing the best model for your goals. This can mean deciding between different models, like doing multiple regression or a decision tree, but it can also mean just deciding what values for the parameters. Alternatively, model evaluation is literally evaluating the model you have chosen, to see how well it works with new data! The important part is the “new data” aspect. If you use the same exact data or folds, your model will appear to work very well, leading to overfitting. You will likely get a low error, but that’s simply because the model was trained on that data ! If order to truly evaluate the model, you want to see how it works with new data.

1. **Bootstrap Philosophy**  
   The bootstrap method might feel strange at first: we create synthetic datasets by resampling the one dataset we already have. Reflect on this: Why do you think the bootstrap works? What assumptions does it rely on? When might those assumptions break down?

At first, I struggle to understand how this is different from using the same data to evaluate our model. However, the bootstrap works because of the randomness itself, using the sample we have as a proxy for the population. The replacement aspect is what truly allows for simulation of variability, we are pretending that we have different datasets! In terms of assumptions, we treat the data as an empirical distribution as if we were resampling from the population. However, if we have a unique sample that was not originally randomly chosen from the population, the bootstrap may not work as well, and instead, reinforce bias. Additionally, we need a large sample to be able to make reasonable conclusions, if the sample size is small we may struggle.

1. **Signal vs. Noise**  
   Share a real-world example where you think randomness or noise might dominate over signal. How would this affect your approach to modeling?

I think in the social sciences, we have a lot of randomness or noise that makes it difficult to model and evaluate outcomes. In particular, think of a situation where we are trying to evaluate the long-term outcomes of a new educational policy in elementary school. Perhaps we want to see if eliminating homework in elementary school has a relationship with graduate rates in high school, and if the cost is worth implementing. In this situation there is a lot of potential factors influencing graduate rates in high school, it would be really difficult to tie these two together. With modeling, we

Take one dataset (real or synthetic), and:

1.       Generate 100 bootstrap resamples.

2.       Fit the same model to each resample.

3.       Track the variability in the coefficients or predictions.

**Questions to consider:**

·        How stable are your estimates across bootstrapped samples? Estimates are relatively stable

·        Are some features consistently important?

·        What does this tell you about uncertainty in your model?

The estimates were relatively stable overall, although the confidence intervals were larger for

For this practice, I ended up importing data from a study I did last year on how trust in the government in the United States can predict political participation, when controlling for political spectrum placement and income, utilizing publicly available 2014 data from ISSP. My goal was to evaluate if my original estimates would remain stable across bootstrapping. After removing missing data, my sample size 972 which appears to be moderate enough for bootstrapping purposes (citation). First, I inputted the data using *pandas*, and then I built a regression model with the follow features: trust in the government (continuous), political spectrum (continuous) and income (categorical), to predict the outcome: political participation (continuous).

Overall, the final average coefficients and standard deviation were very similar to what I originally found. Both in direction and magnitude. When evaluating the individual features, they all seemed consistently important, but particularly income. This tells me that I can be more certain that my multiple regression was an adequate model to choose and that my original multiple regression is not overly sensitive to sampling variation. I decided to visualize the sampling variability using histograms. I found this code: “(<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html>”, which I adapted to build my own For loop across my different predictors. The following histograms were generated. As you can see they all look relatively normal, with not too large variation.