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, such as comparing multiple regression with a decision tree, but it can also mean simply tuning hyperparameters. Alternatively, model evaluation is evaluating said chosen model, to see how well it works with new data! The important part to focus on 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. I like to think of it like cheating on a test. The model will likely have low error, but that’s simply because the model cheated, it’s not actually learning! To get a true idea of how well a model generalizes, model evaluation should always be done on unseen data, separate from what was used in model selection.

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 really struggled to understand how bootstrapping could model the central limit theorem or sampling variability. However, through these class practices I realized the bootstrap works *because* of the randomness itself, using the sample as a proxy for the population to resample with replacement. In this case, we can use the idea of variability and randomness to our advantage. The key aspect of bootstrapping is replacement; this is what truly allows for simulation of variability and different datasets!

In terms of assumptions, we treat the data as an empirical distribution. We must assume it is representative of the population, the data is independent and identically distributed, and the sample is large enough (Solomon, 2020). For example, if we have a convenience sample that was not randomly chosen from the population, the bootstrap may not work as well, and instead, reinforce bias. Additionally, if the sample size is small, we may struggle to get the variation needed.

Solomon, D. (2020, November 26). Bootstrapping: The basics. Medium. [https://drew-solomon.medium.com/bootstrapping-the-basics-4dbd7ca965f1](https://drew-solomon.medium.com/bootstrapping-the-basics-4dbd7ca965f1?utm_source=chatgpt.com)

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?

Particularly in the social sciences, we have a lot of randomness or noise that makes it difficult to model and evaluate outcomes. For example. I work on a survey team, and we have a common issue called survey fatigue where participants may respond without thinking, choose random answers, or rush through questions, especially in long surveys. In this situation there is a lot of random noise that can overshadow true relationships, making it difficult to make valid inferences or predictions. To rectify this (aside from changing the survey itself), we often analyze response time, use statistical tools more robust to noise (E.g., robust standard errors), and analyze data longitudinally (e.g., mixed models) to uncover patterns overtime. Another thing we sometimes do is utilize latent models, like factor analysis, which can help separate measurement error from true signal.

**Experiment 3: Bootstrapping for Uncertainty:** 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 the International Social Survey Programme (ISSP). My goal was to see if my original estimates would remain stable when bootstrapping. After removing missing data, my sample had 972 rows which is considered large enough for bootstrapping purposes (Hayden, 2019). First, I inputted the data using *pandas*, and then I built a regression model with the following features: trust in the government (continuous 1-5), political spectrum (continuous 1-10) and income (categorical 0/1), to predict the outcome: political participation (continuous 0-9).

I decided to visualize the sampling variability using histograms. I adapted example code from the SciPy documentation (Virtanen et al., 2020) to build my own for-loop across the three predictors “(<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html>”, (Virtanen, et al., 2020).

**Table 1.**

*Sample Versus Bootstrapped Descriptives.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Original Mean** | **Bootstrap Mean** | **Original SE** | **Bootstrap SD** |
| **Intercept** | 3.848 | 3.833 | 0.275 | 0.250 |
| **Government Trust** | -0.185 | -0.185 | 0.069 | 0.067 |
| **Political Spectrum** | -0.090 | -0.087 | 0.033 | 0.032 |
| **Income** | 1.208 | 1.196 | 0.149 | 0.125 |

The following bootstrapped histograms (and confidence intervals) were generated. These approximate a normal distribution, with no extreme skewness or heavy tails. The overall spread is moderate, further supporting the stability of the regression estimates across resamples.

A graph of a distribution of a number of points

AI-generated content may be incorrect.A graph of a diagram

AI-generated content may be incorrect.A graph of a tall tower

AI-generated content may be incorrect.A graph of a tall blue line

AI-generated content may be incorrect.

**How stable are your estimates across bootstrapped samples?** Overall, the final average coefficients and standard deviation were extremely similar to my original findings, both in direction and magnitude. When evaluating the confidence intervals and min/max values across bootstraps, the estimates were very stable: none changed direction or crossed zero, and the ranges, relative to their original measurement scales, were reasonable. In some cases, such as income, the bootstrap standard deviations were slightly smaller than my original OLS standard errors, indicating the estimates were more stable than expected.

**Are some features consistently important?** When evaluating the individual features, they all seemed consistently important, particularly because none of the confidence intervals contained zero, indicating consistent statistical effect. However, income had the largest relationship with political participation, across samples. Government trust was more moderate. Political spectrum placement had a smaller effect on average, with its histogram centered closer to zero and its confidence interval lying nearer to zero, indicating less importance. This aligned with my original interpretation of the data.

**What does this tell you about uncertainty in your model?** These results indicate that my regression coefficients and estimates are stable and robust to sampling variation. I can be more certain that my multiple regression was an adequate model to choose for the data, and that my original multiple regression is not overly sensitive to the specific sample I chose.

Virtanen, P., Gommers, R., Oliphant, T.E. et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python*. Nature Methods,* 17(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>

Hayden, R. W. (2019). Questionable claims for simple versions of the bootstrap. *Journal of Statistics Education, 27*(3), 208–215. <https://doi.org/10.1080/10691898.2019.1669507>