Instructions

- Homework 2 is due December 7th at 16:00 Chicago Time.
  - We will not accept any submissions past 16:00:00, even if they are only one second late.
- You **must** upload the following files to the class Canvas:
  - `LASTNAME_FIRSTNAME.pdf`
  - `LASTNAME_FIRSTNAME.ipynb`
- Your code notebook **must** be runnable using my environment outlines in class 1 (Python 3.14, and the `requirements.txt`).
- You **must** use this template file and fill out your solutions for the written portion.
- Please note that your last name and first name should match what you appear on Canvas as.
- Include code snippets where required, as well as math and equations.
- Be *concise* where possible, all of the homework probelms can be answered in a few lines of math, code, and words.

Problem 1: Hands-On OLS

Problem 1.1: Setup

Set your random seet using `np.random.seed(1)`. Generate $n = 30$ observations where:

- The predictor X is drawn from a standard normal distribution, $X \sim N(0, 1)$.
- The error term ϵ is drawn from a normal distribution with mean 0 and standard deviation 1, $\epsilon \sim N(0, 1)$.
- The response variable is generated by the true relationship: $Y = 5 + 2X + \epsilon$.

Display a scatter plot of the generated data points (X_i, Y_i).

Answer:

Problem 1.2: A First Fit

Using the data generated above, fit an OLS model. You should report:

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ estimates.
2. Your confidence intervals and p-values for both coefficients.
3. The R^2 value of your fit.

Answer:

Problem 1.3: Interpretation

How do your estimated coefficients and confidence intervals compare to the true parameters?

Answer:

Problem 1.4: An Influential Point

Now, modify your dataset by overriding the last obvervation to be the point $(X_{30}, Y_{30}) = (4, -5)$. Refit your OLS model to this modified dataset, and report:

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ estimates.
2. Your confidence intervals and p-values for both coefficients.
3. The R^2 value of your fit.

Problem 1.5: Interpretation

How do your estimated coefficients and confidence intervals compare to the true parameters?

Answer:

Problem 1.6: Cook's Distance

Calculate Cook's distance for all observations in the modified dataset from part (d). Plot the Cook's distance values, and highlight the 31st observation.

Answer:

Problem 1.7: What if?

Suppose instead that the influential point you just added was $(X_{30}, Y_{30}) = (0, -5)$.

What would you expect the *leverage* of this point to be relative to the $(4, -5)$ point you added before?

Answer:

Problem 2: Central Limit Theorem

We found in Class 2, that if $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then as $n \to \infty$, then $\beta_{\text{OLS}} \to \mathcal{N}(\beta_{\text{OLS}}, \sigma^2_{\text{OLS}})$, where $\sigma^2_{\text{OLS}} = \sigma^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2$.

The goal of this problem is for you to empirically verify that this results holds *even if* ϵ_i are not normally distributed.

Problem 2.1: Setup

First, define the true model as:

$$y = 2x$$

Where x is sampled from the uniform distribution $x \sim \text{Uniform}(0, 1)$. Note that we are not including an intercept, nor any noise.

Second, define the noise ϵ_i to be sampled from 2 different distributions:

- A Bernoulli distribution with $p = 0.5$, and values $\{-1, 1\}$ (the Radamacher distribution).
- A uniform distribution with $\epsilon_i \sim \text{Uniform}(-1, 1)$.

For $n = 500$, please display histograms of the 2 noise distributions you defined above.

Answer:

Problem 2.2: Empirical Verification

Note: You should not re-sample your x values, only the ϵ_i values. That is, you should have a fixed set of x values for this entire problem.

For $n \in 10, 100, 1000$, and for each noise distribution defined above, do the following:

1. Sample n values of ϵ_i from the noise distribution.
2. Generate the target values as $y_i = 2x_i + \epsilon_i$.
3. Fit an OLS regression to the data (x_i, y_i), and obtain the estimate $\hat{\beta}_{\text{OLS}}$.
4. Repeat (1)-(3) 1,000 times, and save all of the $\hat{\beta}_{\text{OLS}}$ estimates.

You should now have 6 sets of 1,000 $\hat{\beta}_{\text{OLS}}$ estimates (2 noise distributions $\times$ 3 values of n).

Include your simulation code below:

Answer:

Problem 2.3: Visualization

For each of the 6 sets of $\hat{\beta}_{\text{OLS}}$ estimates obtained above, plot a histogram of the estimates. Additionally, conduct a Shapiro-Wilk test for normality on each set of estimates, report your p-values, and briefly discuss your results.

Answer:

Problem 3: Weighted Least Squares

This problem focuses on verifying the findings of Class 2 regarding Weighted Least Squares (WLS).

Problem 3.1: Setup

Similar to before, define the true model as:

$$y = 2x$$

Where x is sampled from the uniform distribution $x \sim \text{Uniform}(1, 2)$.

Second, define the noise ϵ_i to be sampled from a normal distribution with mean 0, and variance σ_i^2 that depends on x_i as follows:

$$\sigma_i^2 = \frac{1}{x_i^2}$$

For $n = 500$, please plot the generated data points $(x_i + \epsilon_i, y_i)$.

Answer:

Problem 3.2: Naive OLS

Fit a standard OLS regression to the data generated above.

Report your estimate $\hat{\beta}_{\text{OLS}}$, and the confidence interval for the estimate.

Answer:

Problem 3.3: Interpretation

Briefly discuss whether naive OLS will be under or over confident in its estimate of $\hat{\beta}_{\text{OLS}}$, and why.

Answer:

Problem 3.4: Visualization

Plot a Q-Q plot of the residuals from the naive OLS regression. What do you observe?

Answer:

Problem 3.5: Weighted Least Squares

Fit a Weighted Least Squares regression to the data generated above, using weights $w_i = x_i^2$. Report your estimate $\hat{\beta}_{\text{WLS}}$, and the confidence interval for the estimate.

Answer:

Problem 3.6: Visualization

Plot a Q-Q plot of the weighted residuals from the WLS regression. What do you observe compared to the Q-Q plot from the naive OLS regression?

Answer: