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:

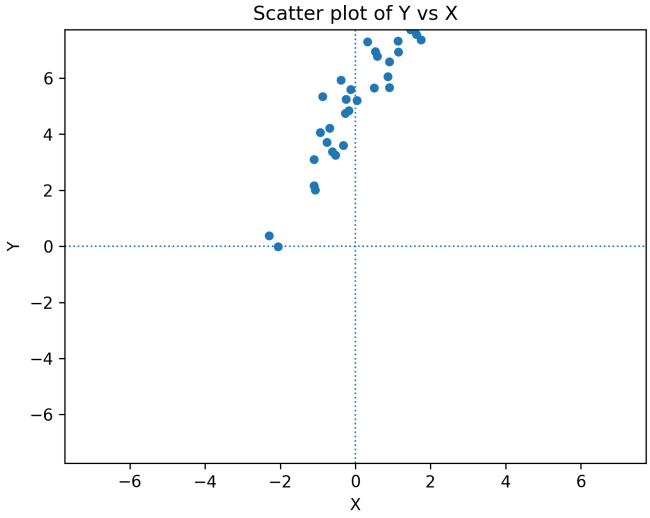


Figure 1: Scatter plot of generated data points (X_i, Y_i).

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:

Dep. Variable:		Y	**R-squared:**		0.840	
Model:		OLS	**Adj. R-squared:**		0.834	
	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	5.0677	0.155	32.787	0.000	4.751	5.384
X	1.8516	0.153	12.110	0.000	1.538	2.165

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Problem 1.3: Interpretation

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

Answer:
The estimated coefficients are very close to the true parameters. The confidence intervals for both coefficients include the true values of 5 (intercept) and 2 (slope), indicating that our estimates are accurate and reliable. The p-values for both coefficients are approximately 0, suggesting that we can reject the null hypothesis that the coefficients are equal to zero.

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.

Answer:

Dep. Variable:		Y	**R-squared:**		0.027
Model:		OLS	**Adj. R-squared:**		-0.008
	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	4.5388	0.500	9.082	0.000	3.515	5.563
X	0.3533	0.402	0.879	0.387	-0.470	1.177

Problem 1.5: Interpretation

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

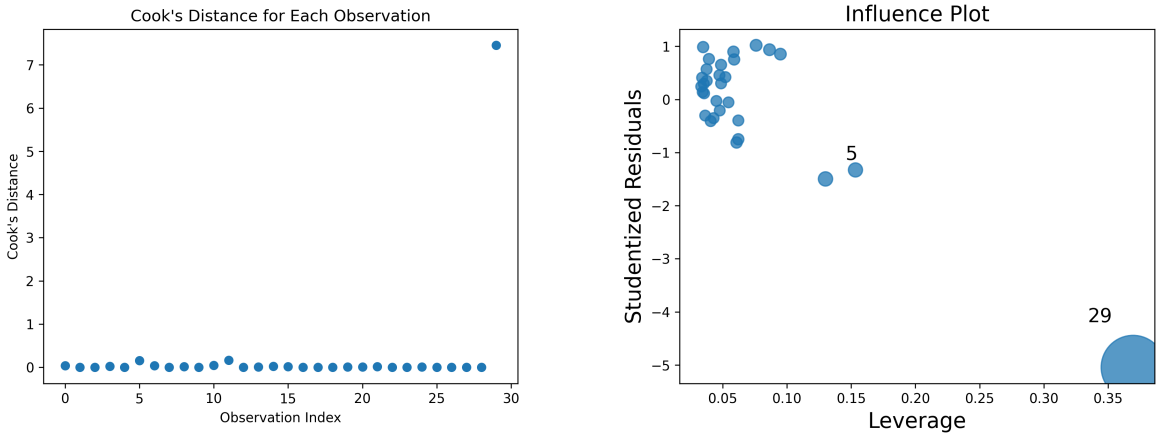
Answer:
The estimated slope coefficient has changed dramatically from approximately 1.85 to 0.35 due to the addition of the influential point.

The confidence interval for the slope now includes zero, indicating that we cannot reject the null hypothesis. The p-value indicates that the probability of observing such a slope under the null hypothesis is quite high (0.774).

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:



(a) Cook's Distance for Each Observation

(b) Influence Plot

Figure 2: Cook's Distance and Influence Plot for Modified Dataset

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:
Under $X = \begin{pmatrix} 1 & x_1 \end{pmatrix}^\top$ design matrix setup, the formula of leverage is $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$. So leverage is influenced by the distance of the predictor value from the mean of the predictor values. Since 0 is exactly the population mean of the X values, the leverage of the point (0, -5) would be lower than that of the point (4, -5).

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:

$$y = 2x$$

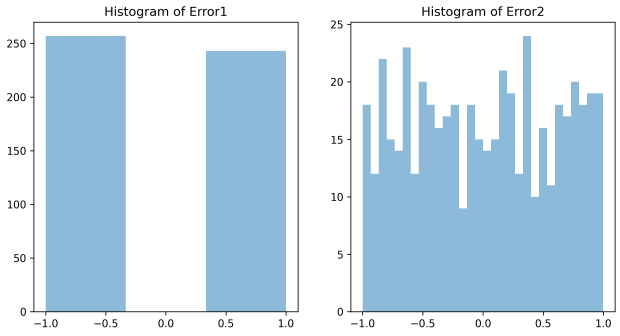
Where x is sampled from the uniform distribution $x \sim \text{Uniform}(0, 1)$. Note that we are not including an intercept, nor any noise.

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:

```
res_beta = {}
x= np.random.rand(1000)
for n in [10, 100, 1000]:
    x_temp = x[:n]
    temp = {'beta1': [], 'beta2': []}
    for i in range(1000):
        eps1 = 2 * np.random.binomial(1, 0.5, n) - 1
        eps2 = 2 * np.random.rand(n) - 1

        y1 = 2 * x_temp + eps1
        y2 = 2 * x_temp + eps2

        beta1 = sm.OLS(y1, x_temp).fit().params[-1]
        beta2 = sm.OLS(y2, x_temp).fit().params[-1]
        temp['beta1'].append(beta1)
        temp['beta2'].append(beta2)
    res_beta[n] = pd.DataFrame(temp)
```

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:

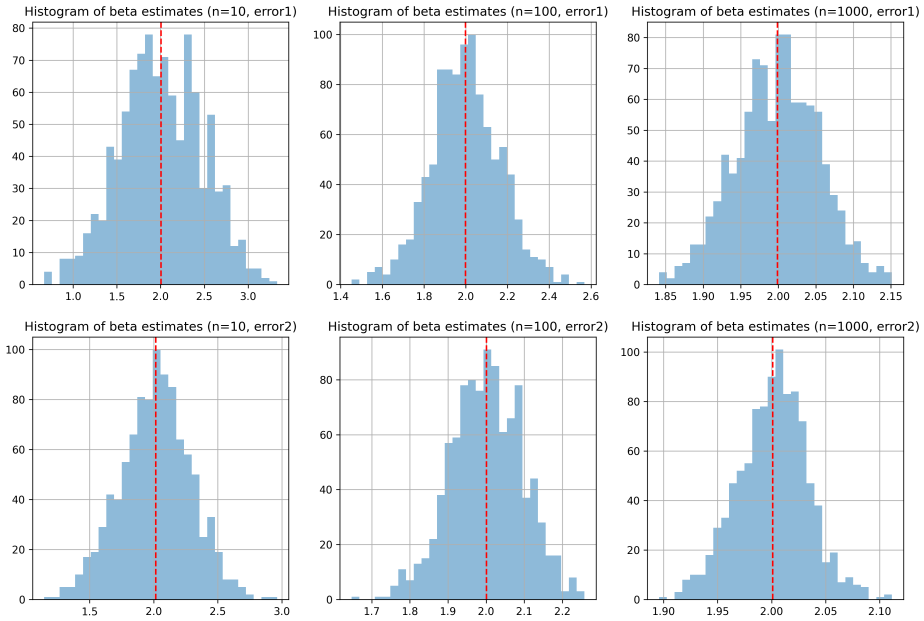


Figure 3: Histograms of $\hat{\beta}_{\text{OLS}}$ estimates for different sample sizes and noise distributions.

	10_beta1	10_beta2	100_beta1	100_beta2	1000_beta1	1000_beta2
shapiro_stat	0.994535	0.996154	0.999380	0.997112	0.998398	0.998662
shapiro_p	0.001090	0.014110	0.989955	0.068768	0.489748	0.662977

The null hypothesis of the Shapiro-Wilk test is that the data is normally distributed. From the p-values reported above, with smaller sammple (n=10), it is less likely that the estimates are normally distributed. This shows the asymptotic normality through Central Limit Theorem.

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:

$$y = 2x$$

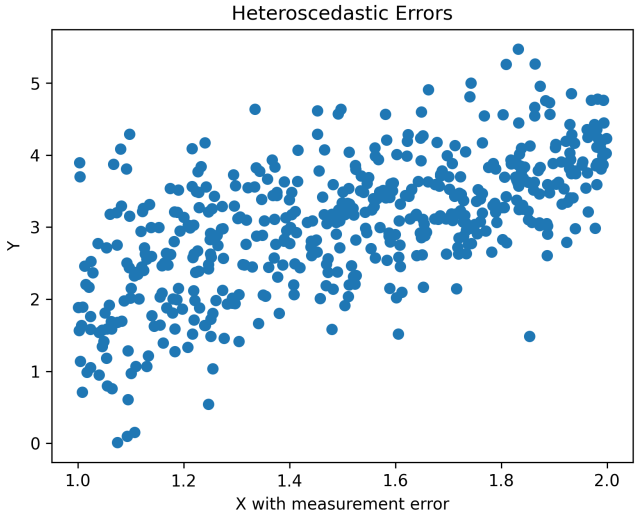
Where x is sampled from the uniform distribution $x \sim \text{Uniform}(1, 2)$.

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 $(x_i, y_i + \epsilon_i)$.

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:

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.0723	0.169	0.428	0.669	-0.260	0.404
x1	1.9801	0.110	17.937	0.000	1.763	2.197

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:

Case 1 Understanding this Problem 3 as having measurement error in X, this relates to the attenuation bias (larger variance of x, so β is biased towards zero). But this is not about the heteroscedasticity, so WLS does not help to solve this bias issue.

Case 2 When heteroscedasticity exists, naive OLS beta has larger variance compared to WLS. This is because WLS is the efficient estimator under heteroscedasticity.

(Proof sketch) Let $\text{Var}(\varepsilon \mid X) = \Omega$ where $\Omega = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2) \succ 0$. Then

$$\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top y \qquad \text{Var}(\hat{\beta}_{\text{OLS}} \mid X) = (X^\top X)^{-1} X^\top \Omega X (X^\top X)^{-1}$$
$$\hat{\beta}_{\text{WLS}} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y \quad \text{Var}(\hat{\beta}_{\text{WLS}} \mid X) = (X^\top \Omega^{-1} X)^{-1}$$

Let $A_{\text{OLS}} = (X^\top X)^{-1} X^\top$ and $A_{\text{WLS}} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1}$, which satisfy $AX = I$. Then you can derive that

$$\text{Var}(\hat{\beta}_{\text{OLS}} \mid X) = \text{Var}(\hat{\beta}_{\text{WLS}} \mid X) + (A_{\text{OLS}} - A_{\text{WLS}}) \Omega (A_{\text{OLS}} - A_{\text{WLS}})^\top$$

Then $\Omega \succ 0$ implies

$$\text{Var}(\hat{\beta}_{\text{OLS}} \mid X) - \text{Var}(\hat{\beta}_{\text{WLS}} \mid X) \succeq 0$$

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:

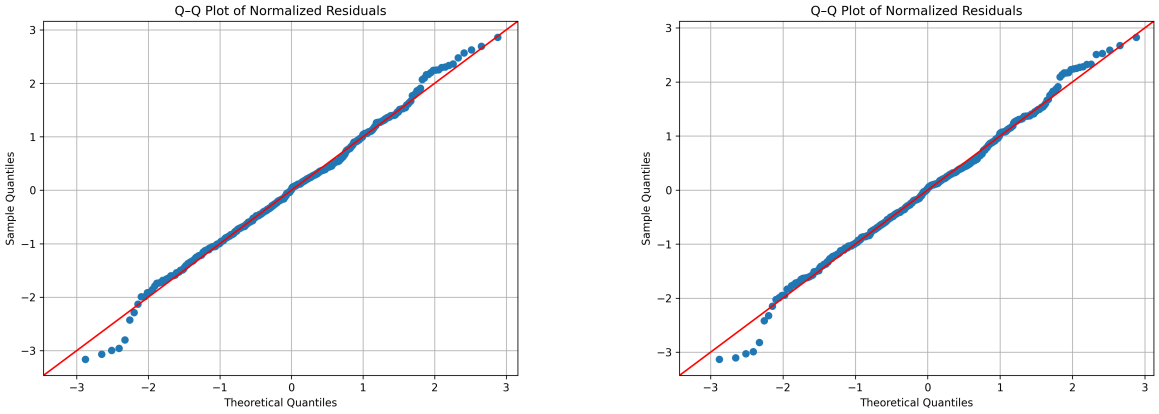
	coef	std err	t	P> |t|	[0.025	0.975]
const	0.1693	0.180	0.939	0.348	-0.185	0.523
x1	1.9187	0.110	17.393	0.000	1.702	2.135

Problem 3.4: Visualization

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

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?



(a) Naive OLS (b) WLS

Figure 4: Comparison of Q-Q Plots: Naive OLS vs WLS

From the OLS Q-Q plot, we can see that the residuals slightly deviate from the normal distribution, especially in the tails. It is improved in the WLS Q-Q plot, however, there are still some deviations from normality, indicating that while WLS has improved the fit, it may not fully address all issues related to the residuals' distribution.