FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Quynh Trang Nguyen	Vietnam	nqtrang.2111@gmail.com	
Willytom Tanguy Rigaud	France	willytomrig@gmail.com	
Francis Kwami Dzikpe	USA	Fkdzikpe@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).	
Team member 1	Quynh Trang Nguyen
Team member 2	Willytom Rigaud
Team member 3	Francis Kwami Dzikpe

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed. Note: You may be required to provide proof of your outreach to non-contributing members upon request.

GWP1:
PROBLEM 1:

1a) No. This happens because when an important variable, such as z(i), is omitted, μ(i) will include the missing part: $\mu(i)=\delta z(i)+\varepsilon(i)$. So if the missing variable z(i) is somehow correlated with the remaining variables, then the new error will also be correlated with those variables. This is not allowed because it violates one of the basic rules (Wooldridge (p. 92) and Stock & Watson (p. 117)).

1b) OLS estimates for α and β can get biased and inconsistent, especially if the variable left out is related to the ones kept in the model. Also, we can't figure out δ. Basically, the expected value of β in the wrong model won't equal the real β, but instead, it'll be off by an amount that depends on how much the missing variable is related to the other variables in the model and how big δ is. This is a common problem in regressions and is explained pretty clearly in Wooldridge (pages 93–94) and Greene (pages 279–281).

1c) The estimates from model (2) would be the same as those from model (1) **if and only if z(i) is uncorrelated with all included regressors (x(i), w(i))**. In other words, there can't be any correlation between z(i) and x(i) or w(i). If that's true, then leaving out z(i) doesn't make the error term related to the other variables, so there's no omitted variable bias. (Stock and Watson, p.118).

1d) Methodology: We made a regression model like this:

- **True Model:** $Y(i) = a + bX(i) + cZ(i) + e(i)$ (with a=2, b=3, c=1.5)
- **Wrong Model:** $Y(i) = a + bX(i) + \mu(i)$ (missing Z)

Then we tested different levels of correlation between X and Z (from 0.0 to 0.9) and tried different sample sizes (from 50 to 5,000). We ran 1,000 simulations each time.

What We Found:

- **Effect of Correlation:** Basically, the higher the correlation, the more bias.
  - r = 0.0 → no bias (about 0)
  - r = 0.9 → big bias (up to 43.9% of the true coefficient)
- **Effect of Sample Size:** Bias was about the same no matter the sample size =>adding more data didn't fix the problem.
  - n = 50 → bias = 1.089
  - n = 5,000 → bias = 1.087
- **Theory vs. Simulation:** The difference was less than 1%, which was really close.
  - Theory said bias should be 0.8961.
  - Our simulation showed 0.9031.

What We Learned:

1. If we leave out Z, the estimate for X is biased.
2. The bias gets worse as X and Z are more related.
3. Even with more data, the bias doesn't go away.
4. This is called omitted variable bias, and can't fix it just by adding more data. We have to include all the important variables in the model.

PROBLEM 2:

2a) Why is OLS sensitive to outliers?

OLS (Ordinary Least Squares) works by squaring the errors (how far each point is from the line) and trying to make the total as small as possible.

=> But squaring makes big errors even bigger.

=> So one outlier can throw everything off and pull the line in its direction.

What Can Go Wrong When Outliers Show Up?

- Messy Coefficients: The slope and intercept can change a lot.
- Big Standard Errors: Outliers can make the data more spread out.
- High Leverage: Some outliers have really extreme X values that can "pull" the line and hide other weird points.
- Broken Assumptions: Outliers can mess up some conditions that OLS needs to work well.

What Should We Do About It?

Before running any regression, we need to check outliers to make sure that they correctly fit into the assumptions of the models we want to use. If the outliers aren't the results of measuring errors or input mistake and you can't remove them (because they're real data), we can try other methods like *robust regression* that aren't as sensitive. If our outliers showcase heteroskedasticity, we can use the Weighted Least Square method to apply less importance to outliers. Finally, in rare cases, we might be able to fix mistakes, if we are sure that the outstanding values indeed are so. Outliers are really important because in a lot of cases in finance, we are interested in what happens in the worst case scenario. Ignoring the outliers means ignoring important information on signals and results of black swan events for example.

2b) Simulation Design: A simple regression model: $Y=2+1.5X+\varepsilon$. We used normal errors ($\sigma = 2$) and 100 data points. Then we added 5 outliers on purpose to see how they mess things up. Some were high leverage (extreme X values) and some were high influence (changing both the slope and the intercept).

What We Found:

Big Parameter Changes:

- Slope changed by 38.8%
- Intercept changed by -168.3%
- R^2 went way down with outliers

Different Outlier Types:

- High Leverage: messed up the slope a bit ($\Delta\beta \approx 0.08$), not so much the intercept ($\Delta\alpha \approx -0.74$)
- High Influence: messed up both slope and intercept a lot ($\Delta\alpha \approx 3.37$, $\Delta\beta \approx -0.40$), R^2 dropped to 0.4173
- Extreme Y Outliers: R^2 dropped to 0.3384

What This Means:

- Just a few outliers can mess up regression results a lot.
- Outliers made the residuals bigger than ±2.5, making them easy to spot.
- As we've seen, just 5% outliers can greatly affect our regression results, so it's really important to conduct thorough checks before any modelling.
- Especially in finance, where weird stuff happens a lot, we need to be extra careful with outliers so our forecasts and conclusions make sense.

PROBLEM 3:
Method 1a. : Feature Selection through AIC Minimizing with OLS
We tested every combination with itertools. For each model tested, the workflow is as follows : create the y and X dataset to use sm.OLS, and fit the model. We then test for error normality, heteroskedasticity,

autocorrelation. At each iteration, if the model does not pass one of the tests, we exclude it from the results dictionary and would proceed to estimate it using an adequate regression method. However, every model passed all assumptions tests.
From our AIC minimizing operation, we find that the model with the lowest AIC is

$$y = \beta 2z2 + \beta 3z3 + \beta 4z4 + \beta 5z5$$

Which gives an AIC of 260.6167.
(Method 1b,Method 1c are forward and backward selection : checks to see if we have the same results as with one, using more computationally efficient approaches)
Method 2 : Algorithm Selection to maximize R2
Though R2 can be really prone to overfitting for model selection, given that we are using the same amount of regressors for every model, this does not need to concern us at this point. We tested Ordinary Least Squares, Ridge Regression and Lasso Regression. OLS and LASSO R2 were really close, less than 1 basis point of difference, but the best model with regards to this metric is OLS, which beats Lasso by a mere 0.8 basis points
We therefore find that the best model given this data on what we tested is OLS on

$$y = \beta 2z2 + \beta 3z3 + \beta 4z4 + \beta 5z5$$

PROBLEM 4 :

Elasticity measures the percentage change in one variable in response to a percentage change in another variable. For OLS regression models, we estimate elasticity by regressing the natural log of the dependent variable (ln(y)) on the natural log of the independent variable (ln(x)). Mathematically, elasticity is give as

$$Elasticity = \frac{d\ln(y)}{d\ln(x)} = \frac{dy}{dx} \times \frac{x}{y}$$

(i) y = 2 + 0.8x

Model 1 is not a log-log model and hence 0.8 is not the elasticity measure. We can estimate elasticity as follows. First, from the regression model, dy/dx = 0.8, hence the elasticity of y in terms of x is

$$Elasticity = \frac{dy}{dx} \times \frac{x}{y} = 0.8 \times \frac{x}{y} = \frac{0.8x}{2 + 0.8x}$$

(ii) ln(y) = 0.1 + 0.4x

Model 2 is also not a log-log model and hence 0.4 is not the elasticity measure. We can estimate the elasticity as. First, from the regression model, dln(y)/dx = 0.4, hence the elasticity of y in terms of x is

$$Elasticity = \frac{d\ln(y)}{d\ln(x)} = \frac{d\ln(y)}{dx} \times \frac{dx}{d\ln(x)} = 0.4 \times x = 0.4x$$

iii) ln(y) = 0.1 + 0.25ln(x)

Model 3 is a log-log model and hence the coefficient of 0.25 is the elasticity of y in terms of x. That is,

$$Elasticity = \frac{d\ln(y)}{d\ln(x)} = 0.25$$

(iv) y = 0.15 + 1.2ln(x)

Model 4 is a linear-log model and hence 1.2 is not the elasticity measure. We can estimate the elasticity as. First, from the regression model, dy/dln(x) = 1.2. Hence the elasticity of y in terms of x is

$$Elasticity = \frac{d\ln(y)}{d\ln(x)} = \frac{d\ln(y)}{dy} \times \frac{dy}{d\ln(x)} = 1.2 \times \frac{1}{y} = \frac{1.2}{0.15+1.2\ln(x)}$$

PROBLEM 4B.

To estimate the elasticity of y with respect to x using the regression models above, we need to use the correct functional form which is the log-log model. From the regression modes, only model 3 - ln(y) = 0.1 + 0.25ln(x) is the correct functional form. Hence the elasticity of y with respect to x is 0.25.

PROBLEM 5A.

A time series is "stationary" if its statistical properties, including the joint probability distributions, are stable over time (Woolridge, 2020). Specifically, a stationary series has a constant mean and variance and its autocovariances are independent of time. A time series that is not stationary lacks the above properties and may lead to spurious regressions and hence affecting confidence interval estimation and inferences. Notable causes non-stationarity includes the presence of a deterministic trend or a stochastic trend. Deterministic trends may occur due to the underlying data generation process. They are usually fixed, non-random and deviations from trend are temporal. On the other hand, stochastic trends are random, varies over time, induced by random unpredictable shocks, and have permanent effects. A typical example of a stochastic process are random walk models. A random walk AR(1) model with characteristic equation equal to 1 is what economist generally refers to as a "Unit root process". Since random walks are non-stationary, the presence of unit root indicates non-stationarity.

Because non-stationarity may lead to spurious regression results, it is a general practice to test unit roots before any analysis. A visual approach to test for unit roots is to create a scatterplot of the time series and determine if the observations oscillate around a constant (no unit root). A more formal procedure, is to use a statistical approachs such as the Augmented Dicky Fuller (ADF) and Philip Perrone (PP) test. Once unit root is determined, we can difference the series to remove the unit root. Differencing transforms the series into a stationary process. Sometimes, we can also use log transformations before differencing particularly if first and second differencing is not enough to remove the unit root.

PROBLEM 5B.

To illustrate the unit root process in time series, we use the adjusted close price of Apple (AAPL) and Walmart (WMT). We obtained the dataset from Yahoo Finance. The data covers the period, Jan 2nd, 2010, to December 31st, 2024. We clean the data by removing all NA's and plot the series to observe the nature of the trend. To test for unit roots, we use the Augmented Dickey Fuller (ADF) test and the Philip Perrone (PP) test. Both the ADF and PP test, test the null hypothesis of the presence of a unit root in the time series. Because both AAPL and WMT exhibit trends, we test the unit root with trend and without trend. In all cases (both with trend and without trend) our analysis shows p-values greater than 0.05 implying we fail to reject the null hypothesis of a unit root. The full result of the unit root test is shown in the table below.

Variable	Type of Unit root test	Test statistic	P-value
AAPL	ADF with constant + trend	-0.629	0.977
	ADF with constant	1.764	0.998
	PP test with constant + trend	-0.662	0.976
	PP test with constant	1.805	0.998
WMT	ADF with constant + trend	1.273	1.000
	ADF with constant	3.025	1.000
	PP test with constant + trend	1.621	1.000
	PP test with constant	3.415	1.000

PROBLEM 5C:

As mentioned earlier, economists and finance professionals are usually more concerned about the presence of unit root in time series because it may lead to spurious regressions and misleading confidence intervals and forecasting. Aside this, unit roots are realistic, very common in many real-life economics and financial time series data and hence faced in day-to-day activities. A characteristic root of say 1.5 however is unrealistic, and rare in many time series data. A series with a root of 1.5 is explosive and may amplify any shock in the time series. Since roots beyond 1 are uncommon and not faced in real life, many economists and finance professionals are less concerned about potentially having to deal with them than they are for unit roots.

PROBLEM 6:

A simple way to test for a structural break is to use a dummy variable. By conditioning the dummy variable on being equal to 1 if t>10, and running the regression :

$$Y_t = \alpha + \beta_1 x_t + \beta_2 D_t x_t + \varepsilon_t$$

where D_t= 1 if t>10, 0 otherwise.
We then simply need to conduct a parameter significativity test and check if $\beta 2$ is significant.

As we can see when we run this in python, the x2 here is the coefficient associated to the dummy variable. We see that it is indeed significant at any threshold (10%, 5%, 1%). The model was created by setting beta at 1.5 when T<=10 and 5 when T>10. The coefficient for the dummy indeed tells us that.

REFERENCES

1. Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley.
2. Greene, W. H. (2018). *Econometric Analysis* (8th ed.). Pearson.
3. Huber, P. J., & Ronchetti, E. M. (2009). *Robust Statistics* (2nd ed.). Wiley.
4. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis* (6th ed.). Wiley.
5. Stock, J. H., & Watson, M. W. (2019). *Introduction to Econometrics* (4th ed.). Pearson.
6. Wooldridge, J. M. (2020). *Introductory Econometrics: A Modern Approach* (7th ed.). Cengage.