EC2020 Live Session Exercise 2

Question 1

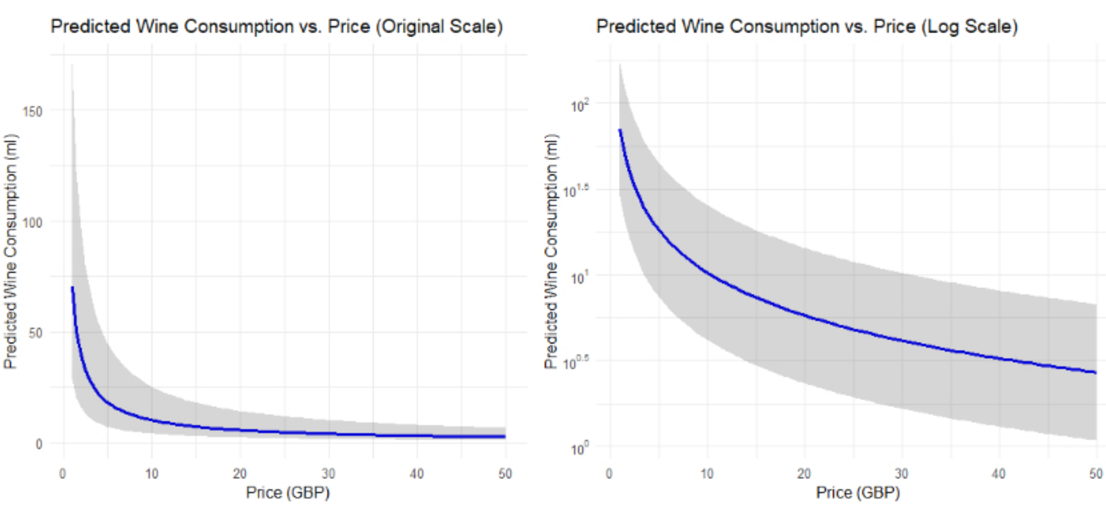
To investigate the relationship between the price of wine and consumption of wine, an economist estimates the following regression using a sample of 32 individuals for one week in 2013:

$$\widehat{\log(\text{ wine })} = \underset{(0.8911)}{4.2514} - \underset{(0.0031)}{0.8328} \log(\text{ price }),$$
$$n = 32, R^2 = 0.89.$$

wine denotes the amount of wine consumed per week in milliliters (a medium glass contains 175 ml), and price denotes the average price of a medium glass of wine of a selection of wines during the week in GBP (£). The numbers in parentheses are the standard errors.

(a) Discuss what would happen to the parameter estimate of the slope coefficient if we had measured the amount of wine consumed per week in number of medium glasses instead of millilitres. Explain your answer.

(b) Discuss what would happen to the parameter estimate of the intercept if we had measured the amount of wine consumed per week in number of medium glasses instead of millilitres. Explain your answer.



When working with regressions, the interpretation of coefficients depends on the units of measurement for the variables. In this problem, the dependent variable is the logarithm of the quantity of wine consumed, and the independent variable is the logarithm of the price. The regression model is in a **log-log** form, which means the coefficients represent elasticities (i.e., the percentage change in the dependent variable for a 1% change in the independent variable).

- The coefficient of $\log(\text{price})$ is -0.8328, indicating that **a 1% increase in price is associated with a 0.8328% decrease in the quantity of wine consumed**, on average.
- The intercept 4.2514 represents the expected value of $\log(\text{wine})$ when $\log(\text{price})$ is zero (i.e., when the price is 1 GBP).

> a) The slope coefficient measures the **elasticity** of wine consumptions with respect to price. If we change the units of the dependend variable, from milliliters to glasses, we would change the scale of the variable but not the relative percentage changes. Therefore, the slope coefficient will remain 0.8328%.

> b) The intercept represents the log of the expected value of wine consumption when price is 1 GBP. If you measure wine in glasses rather than milliliters, the intercept will adjusted by
>
> $$4.2514 - \log(175) = \log(\frac{17783.9}{175}) = 2.0069$$
>
> reflecting the conversion factor between milliliters and glasses.

Question 2

Suppose that average worker productivity at manufacturing firms (*avgprod*) depends on two factors, average hours of training (*avgtrain*) and average worker ability (*avgabil*):

$$\text{avgprod } = \beta_0 + \beta_1 \text{ avgtrain } + \beta_2 \text{ avgabil } + u$$

Assume that this equation satisfies the Gauss-Markov assumptions.

(a) Provide a clear interpretation of the parameter β_1.

(b) Suppose firms whose workers have less than average ability receive grants to provide training, so that *avgtrain* and *avgabil* are negatively correlated. Discuss the effect of omitting *avgabil* on the OLS estimator of β_1. Will the bias be upward or downward? Explain your answer.

(c) You are told that you can obtain the estimator of β_1 by running the regression

$$\text{avgprod}_i = \beta_1 \text{avgtrain}_i^* + e_i, \quad i = 1, \ldots, n$$

where $avgtrain_i^*$ is obtained by running a regression of *avgtrain* on an intercept and *avgabil*. Discuss this statement.

The model given is:

$$\text{avgprod} = \beta_0 + \beta_1 \text{avgtrain} + \beta_2 \text{avgabil} + u$$

where:

- avgprod is the average worker productivity,
- avgtrain is the average hours of training,
- avgabil is the average worker ability,
- u is the error term, which captures other factors affecting productivity that are not included in the model.

Assume the Gauss-Markov assumptions hold, which means the OLS estimators are unbiased, efficient, and consistent.

> a) β_1 measures the change in average worker productivity associated with a one-unit increase in average hours of training, assuming no change in worker ability.

When a relevant variable (avgabil) is left out of the regression, the Ordinary Least Squares (OLS) estimator of the included variable (avgtrain) can become biased if the omitted variable is correlated with the included variable. This bias is called **Omitted Variable Bias (OVB)**.

OVB Formula: The bias in the estimated coefficient $\hat{\beta}_1$ when avgabil is omitted can be expressed as:

$$\text{Bias}(\hat{\beta}_{1,\text{short}}) = \beta_2 \cdot \hat{\delta}_1$$

Where β_2 is the true effect of average ability (avgabil) on productivity. If we assume that higher *avgabil* is associated with higher *avgprod*, then;

$$\beta_2 > 0$$

The Ordinary Least Squares (OLS) estimate $\hat{\delta}_1$ can be expressed in terms of the covariance between avgtrain and avgabil as follows:

$$\hat{\delta}_1 = \frac{\text{Cov}(\text{avgtrain}, \text{avgabil})}{\text{Var}(\text{avgabil})}$$

$\hat{\delta}_1$ is the estimated coefficient from regressing avgtrain on avgabil.

Since avgtrain and avgabil are negatively correlated (as stated in the problem), we can denote this correlation as:

$$\text{Cov}(\text{avgtrain}, \text{avgabil}) < 0 \quad \longrightarrow \quad \hat{\delta}_1 < 0$$

and thus $\hat{\delta}_1$ will be **negative**.

Therefore,

$$\text{Bias}(\hat{\beta}_{1,\text{short}}) < 0$$

> b) This means that the OLS estimator $\hat{\beta}_1$ **underestimates** the true effect of training on productivity. In other words, $\hat{\beta}_1$ is smaller than the true β_1. The negative bias implies that, on average, the estimated effect of training $\hat{\beta}_1$ will be **smaller** than the true effect. In some cases, the estimate might even be negative, even if the true β_1 is positive.

Why Does This Bias Occur? When you omit *avgabil*, the variation in worker ability that is correlated with training becomes part of the error term. Since *avgabil* is negatively correlated with *avgtrain*, this correlation leads to bias in the estimated coefficient of *avgtrain*.

We can obtain the estimator of β_1 by running the regression

$$\text{avgprod}_i = \beta_1 \text{avgtrain}_i^* + e_i, \quad i = 1, \ldots, n$$

where $avgtrain_i^*$ is obtained by running a regression of *avgtrain* on an intercept and *avgabil*.

Two-Stage Least Squares (2SLS) Approach:

- We regress avgtrain on avgabil. This gives us $\widehat{\text{avgtrain}}_i$, which represents the part of avgtrain that is predicted by avgabil.
- The residuals from this regression, $avgtrain_i^*$, represent the variation in avgtrain that is **uncorrelated** with avgabil. Essentially, $avgtrain_i^*$ captures the pure effect of training without being contaminated by the ability factor.
- Regressing avgprod on $avgtrain_i^*$ isolates the effect of training on productivity, free from the influence of ability. The resulting coefficient, $\hat{\beta}_1$, will provide an **unbiased estimate** of β_1.

> c) By using the residuals $avgtrain_i^*$, we are effectively controlling for the omitted variable bias that would result from not including avgabil. This is because $avgtrain_i^*$ is uncorrelated with avgabil, allowing us to estimate the effect of training on productivity without the bias introduced by omitting ability. Therefore running the regression with $avgtrain_i^*$ removes the bias and provides an unbiased estimate of β_1.

Question 3

In this question we are interested to see whether fast-food restaurants charge higher prices in areas with a larger concentration of ethnic minorities. ZIP code-level data on prices for various items along with characteristics of the ZIP code population in New Jersey and Pennsylvania are used.

Let us consider a model to explain the price of soda, *psoda*, in terms of the proportion of the population from ethnic minorities, *prpem*, and median income, *income*. Price and income are measured in US$. We obtain the following result

$$\widehat{\text{psoda}} = 0.956 + 0.115 \text{ prpem } + 0.0000016 \text{ income} \qquad (3.1)$$

(a) Interpret the parameter on *prpem*.

(b) What would happen to the parameter estimates if we use the percentage of the population that is from ethnic minorities (*pctem*) instead of the decimal equivalent *prpem*? What would happen to the parameter estimates if we measured income in $10,000 ?

(c) When adding a further measure of income to (3.1), the proportion of the population that is in poverty (*prppov*), the coefficient on *prpem* falls to .089. Discuss the following statement "Because log(*income*) and *prppov* are highly correlated, it is inappropriate to include both".

(d) A model with constant price elasticity with respect to income gives

$$\widehat{\log(\text{ psoda })} = -0.793 + 0.122 \text{ prpem } + 0.077 \log(\text{ income }) \qquad (3.2)$$

Interpret the parameter on *prpem*, and discuss what would happen to the parameter estimates if we use *pctem* instead of *prpem*.

> a) The parameter "0.115 prpem" indicates that a 1% increase in *prpem* would result in an expected increase of $0.00115 in the price of soda, all other factors remaining equal.

> b) If we replace *prpem* with *pctem* = 100 × *prpem*, its coefficient would equal to 0.0015, to ensure it has the same interpretation.

> c) Since the goal of the study is to understand whether pricing differences are due to ethnic discrimination, it is worthwhile to include multiple economic variables (when available) even if they are highly correlated. By including different variables, we better control for factors that might influence prices, allowing us to more accurately isolate the effect of ethnic minorities (prpem) on pricing.

> d) The parameter "0.122 prpem" indicates that a 1% increase in *prpem* would result in an expected increase of 0.122% increase in the price of soda, all other factors remaining equal.
>
> If we replace *prpem* with *pctem* = 100 × *prpem*, its coefficient would equal to 0.00122, to ensure it has the same interpretation.

Question 4

We are interested in investigating the factors governing the precision of regression coefficients. Consider the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

with OLS parameter estimates $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$. Under the Gauss Markov assumptions, we have

$$\text{Var}\left(\hat{\beta}_2 \mid X\right) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r^2_{X_2 X_3}}$$

where σ_ε^2 is the variance of ε and $r_{X_2 X_3}$ is the sample correlation between X_2 and X_3.

(a) Provide at least four factors that help us obtain more precise parameter estimates of β_2.

(b) In light of your answer to (a), discuss the concept of near multicollinearity. Provide a real life example where this problem is likely to occur.

(c) Are the following statements true or false? Give an explanation.

- In multiple regression, multicollinearity implies that the least squares estimators of the coefficients are biased and standard errors invalid.
- If the coefficient estimates in an equation have high standard errors, this is evidence of high multicollinearity.

The variance of $\hat{\beta}_2$, the estimated coefficient for X_2, is given by:

$$\text{Var}\left(\hat{\beta}_2\right) = \frac{\sigma_s^2}{\sum_{i=1}^n (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r^2_{X_2 X_3}}$$

This formula tells us how precise our estimate of $\hat{\beta}_2$ will be. We aim for lower variance.

We may note that;

$$\text{SampleVar } (X_2) = \frac{1}{n} \times \sum_{i=1}^n (X_{2i} - \bar{X}_2)^2$$

Thus, by subtitution, the expression of variance can be rewritten as;

$$\text{Var}\left(\hat{\beta}_2\right) = \frac{\sigma_s^2}{n \times \text{ SampleVar } (X_2) \times \left(1 - r^2_{X_2 X_3}\right)}$$

The four factors that can improve the precision of our parameter are;

- **Larger Sample Size** n - More data points provide a better estimate of the relationship between the variables, making the coefficient estimate more precise.
- **Larger** $\text{SampleVar}(X_2)$ - More spread out data points allow us to better distinguish the effect of X_2 on Y. Thus larger sample variability results in more accurate estimates.
- **Smaller** σ_ε^2 - The variance of the error term ε measures how much the actual values of Y deviate from the values predicted by the regression model. A smaller error variance indicates that the model is better at predicting Y. Less unexplained variation in Y (i.e., smaller σ_ε^2) results in more precise estimates of the coefficients.
- **Smaller** $r^2_{X_2 X_3}$ - The square of the correlation between X_2 and X_3 measures how strongly X_2 and X_3 are related. A higher correlation (closer to 1) means they move together more closely. If X_2 and X_3 are highly correlated (near multicollinearity), it becomes difficult to separate their individual effects on Y, leading to imprecise estimates. A lower correlation makes it easier to estimate $\hat{\beta}_2$ more precisely.

> **Near Multicollinearity:** This refers to a situation where two or more regressors are highly correlated, but not perfectly so. Near multicollinearity means that the variables move together closely, making it hard to separate their effects on the dependent variable.

When two variables are highly correlated, the OLS estimates of their coefficients become imprecise (i.e., they have large standard errors). This is because it's difficult to determine how much of the change in Y is due to X_2 and how much is due to X_3. The correlation between them makes it hard to distinguish their individual contributions.

A real-life example of near multicollinearity can be found in models predicting car prices, where some variables, such as **age** and **milage** may be highly correlated to each other, leading to redundancy in the model.

> *Statement: "In multiple regression, multicollinearity implies that the least squares estimators of the coefficients are biased and standard errors invalid."* **FALSE**

Multicollinearity increases the standard errors of the coefficients, reducing accuracy. However, **multicollinearity does not violate the Gauss-Markov assumptions**. As long as these assumptions hold, the OLS estimators remain unbiased.

> *Statement: "If the coefficient estimates in an equation have high standard errors, this is evidence of high multicollinearity."* **FALSE**

High standard errors can be a symptom of multicollinearity, but they aren't definitive proof of it. Other causes include; small sample sizes, and a high variance of the error term (the model doesn't predict the outcome well).