Problem Set 5

Cem Kozanoglu

November 18, 2025

Collaborators: Wooyong Park, Roberto Gonzalez-Tellez, Hanniel Ho and Aileen Wu.

1 Problem 1

1.1 Part A

1.1.1 Model

We estimate the following model:

$$\text{lwage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{state\_avg\_educ}_i + \epsilon_i$$

Where lwage_i is the log of wage for individual i, educ_i is the years of education for individual i, and state_avg_educ_i is the average years of education in the state where individual i resides. The error term ϵ_i captures unobserved factors affecting wages.

1.1.2 Regression Results

	(1)
const	4.9952***
	(0.004)
educ	0.0709***
	(0.000)
R-squared	0.117
Observations	329,509

Table 1: OLS Regression Results

These results show that the coefficient on education is positive and statistically significant at the 1% level. This suggests that, on average, an additional year of education is associated with a 7.09% increase in wages. The R-squared value of 0.117 indicates that approximately 11.7% of the variation in log wages can be explained by years of education alone.

1.2 Part B, C

1.2.1 Standard Error Formulas

The formula for the Homoskedastic OLS standard error is:

$$SE(\hat{\beta}) = \sqrt{\frac{1}{N} \cdot \frac{\sum(\epsilon_i - \bar{\epsilon})^2}{\sum(X_i - \bar{X})^2}}$$

Where σ^2 is the variance of the error term, and $\sum(X_i - \bar{X})^2$ is the sum of squared deviations of the independent variable from its mean.

Whereas the formula for the clustered standard error is:

$$\text{SE}(\hat{\beta}_j) = \sqrt{(X'X)^{-1}\left(\sum_c X_c'\hat{\Omega}_c X_c\right)(X'X)^{-1}\frac{C}{C-1}\frac{n-1}{n-k}}$$

Where C is the number of clusters, n is the total number of observations, and k is the number of parameters estimated. The term $\hat{\Omega}_c$ represents the covariance matrix of the residuals within cluster c.

1.2.2 Empirical results

Standard Error Type	Value
Homoskedastic	0.0003386
Clustered	0.0018

Table 2: Standard Error Comparison

As we can see from the table, the clustered standard error is significantly larger than the homoskedastic standard error. This suggests that there is likely some correlation of errors within clusters (states), which violates the homoskedasticity assumption and leads to underestimated standard errors when using the homoskedastic formula. Despite this, we still observe that the coefficient on education is statistically significant at the 1% level, even when using the larger clustered standard error.

1.3 Part D

1.3.1 Bootstrap algorithms

For the OLS standard error, I run a simple Parametric bootstrap by resampling the residuals from the OLS regression and recalculating the OLS estimates for each bootstrap sample.

For the clustered standard error, I first run a Clustered bootstrap by resampling clusters (states) with replacement and then sample observations within these clusters to create our full bootstrap sample. I calculate the coefficients for each bootstrap sample and compute the standard error across these bootstrap estimates.

1.3.2 Empirical Results

The empirical results from the bootstrap methods after 10,000 simulations for both are as follows:

Bootstrap Method	Standard Error
OLS	0.0003356
Clustered	0.0017811

Table 3: Bootstrap Standard Error Results

As we can see, the values are very close to the standard errors calculated using the formulas, which provides validation for our bootstrap approach.

1.4 Part E

1.4.1 Model

In this section we calculate a slightly different model by including the state average education as an additional regressor. The model is as follows:

$$\text{lwage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{state\_avg\_educ}_i + \epsilon_i$$

Where lwage_i is the log of wage for individual i, educ_i is the years of education for individual i, and state_avg_educ_i is the average years of education in the state where individual i resides. The error term ϵ_i captures unobserved factors affecting wages.

The economic reason that we might want to include the average state education is that it could capture the overall educational environment of the state, which may have an impact on individual wages. For example, living in a state with higher average education levels might provide better job opportunities, networking effects, or a more skilled workforce, all of which could positively influence an individual's wage. In our case, we don't know how good a proxy state of birth is for state of residence, but we can assume that it is a reasonable proxy for the sake of this analysis. Even if it isn't it probably is a good proxy for state where the individual was educated, where we might further expect better educated states to have better education systems in general, so we have good reason to expect that one or both of these channels would be affecting an individual's wages.

1.4.2 Regression Resuls

We get the following results:

	(1)
const	4.1542***
	(0.019)
educ	0.0671***
	(0.000)
state_avg_educ	0.0696***
	(0.001)
R-squared	0.123
Observations	$329,509$

Table 4: OLS Regression Results with State Average Education

We observe that all three coefficients are statistically significant at the 1% level. The coefficient on education is slightly smaller than in the previous model, which suggests that some of the effect of education on wages is captured by the state average education variable. The coefficient on state average education is positive and significant, indicating that living in a state with higher average education levels is associated with higher wages. Interestingly though, the R-squared only shows a marginal increase, meaning that the state average education is not actually giving us that much more new information.

1.5 Part F, G, H

1.5.1 Standard Error Results

Using the same general formulas for the standard errors as in Part B and C, we get the following results:

Standard Error Type	educ	state_avg_educ
Homoskedastic	0.0003375	0.0014547
Clustered	0.0013508	0.0119502

Table 5: Standard Error Comparison for Part E

I used the same bootstrap algorithms as in Part D to calculate the standard errors for the new model. The results are as follows after 10,000 simulations:

Standard error type	Intercept	Education	State Avg Education
OLS	0.0186	0.0003	0.0015
Clustered	0.1612	0.0014	0.0122

Table 6: Bootstrap Standard Errors for Part E

We see quick convergence to the observed standard errors again, giving extra confidence in our standard error estimates.

Here we can make two observations regarding clustering. First, the clustered standard error for the state average education variable is much larger than the homoskedastic standard error, which suggests that there is significant within-cluster correlation for this variable. This makes sense because individuals within the same state are likely to share similar unobserved characteristics that affect their wages, leading to correlated errors.

An additional economic observations is that while the OLS standard error of the educ variable has not changed much from the first mode, the clustered standard error has decreased significantly. This is likely because the state average education variable captures some of the within-state variation in education levels and also some of the unobserved state-level factors, which reduces the correlation of errors within clusters for the educ variable. In other words, by including state average education, we are controlling for some of the unobserved factors that were previously causing correlated errors within states, leading to a more accurate estimate of the standard error for the educ variable.

2 Problem 2

We define the error covariance structure Ω as:

$$\Omega_{ij} = \begin{cases} \sigma_\eta^2 + \sigma_\nu^2 & \text{if } i = j \\ \sigma_\eta^2 & \text{if } i \neq j,\ S_i = S_j \\ 0 & \text{otherwise} \end{cases}$$

Then we can rewrite Ω in matrix notation as:

$$\Omega = \sigma_\nu^2 I_N + \sigma_\eta^2 Z Z^\top$$

Furthermore, we can use the properties of $ZZ^\top$ and X to get the following:

$$\begin{aligned} \Omega X &= (\sigma_\nu^2 I_N + \sigma_\eta^2 Z Z^\top) X \\ &= \sigma_\nu^2 X + \sigma_\eta^2 Z Z^\top X \end{aligned}$$

Analysis of the term $ZZ^\top X$: Since X is constant within clusters, let x_c be the row vector of regressor values for cluster c. We can visualize the multiplication of the block-diagonal matrix $ZZ^\top$ and X as follows:

$$ZZ^\top X = \begin{bmatrix} J_{N_1} & 0 & \cdots \\ 0 & J_{N_2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \mathbf{1}_{N_1} x_1 \\ \mathbf{1}_{N_2} x_2 \\ \vdots \end{bmatrix}$$

Where J_{N_c} is an $N_c \times N_c$ matrix of ones, and $\mathbf{1}_{N_c}$ is a column vector of ones. For any specific cluster c:

$$J_{N_c}(\mathbf{1}_{N_c}x_c) = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_c \\ \vdots \\ x_c \end{bmatrix} = \begin{bmatrix} N_c x_c \\ \vdots \\ N_c x_c \end{bmatrix} = N_c(\mathbf{1}_{N_c}x_c)$$

Assuming $N_c = N/C$ is constant across all clusters:

$$ZZ^\top X = N_c X$$

Substituting this back into the expression for ΩX:

$$\Omega X = \sigma_\nu^2 X + \sigma_\eta^2(N_c)X$$
$$\Omega X = (\sigma_\nu^2 + N_c\sigma_\eta^2)X$$

Let $\lambda = (\sigma_\nu^2 + N_c\sigma_\eta^2)$. Thus, X is an eigenvector of Ω with eigenvalue λ.

Calculating the Variance of $\hat{\beta}_{OLS}$:

$$\begin{aligned} V(\hat{\beta}_{OLS}) &= (X^\top X)^{-1}(X^\top \Omega X)(X^\top X)^{-1} \\ &= (X^\top X)^{-1}X^\top(\lambda X)(X^\top X)^{-1} \\ &= \lambda(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1} \\ &= (\sigma_\nu^2 + N_c\sigma_\eta^2)(X^\top X)^{-1} \end{aligned}$$

Calculating the Variance of $\hat{\beta}_{GLS}$: Since $\Omega X = \lambda X$, it implies $\Omega^{-1}X = \frac{1}{\lambda}X$.

$$\begin{aligned} V(\hat{\beta}_{GLS}) &= (X^\top \Omega^{-1} X)^{-1} \\ &= \left(X^\top \frac{1}{\lambda} X\right)^{-1} \\ &= \lambda(X^\top X)^{-1} \\ &= (\sigma_\nu^2 + N_c\sigma_\eta^2)(X^\top X)^{-1} \end{aligned}$$

Conclusion:

$$V(\hat{\beta}_{OLS}) = V(\hat{\beta}_{GLS}) = (\sigma_\nu^2 + N_c\sigma_\eta^2)(X^\top X)^{-1}$$

Therefore we've shown that the variance of the OLS estimator is equal to the variance of the GLS estimator under the given error covariance structure.

2.1 Relevance

In the slides, we saw that the variance of the OLS estimator is greater than or equal to the variance of the GLS estimator. In this problem, we've shown one case, when the regressors are constant within clusters, where the variance of the OLS estimator is actually equal to the variance of the GLS estimator. This illustrates that while GLS can be more efficient than OLS in general, there are specific cases where OLS can perform just as well as GLS, particularly when the regressors do not vary within clusters. This is an important insight because it highlights that the efficiency gains from using GLS depend on the structure of the regressors and the error covariance.