HUL 315

Assignment - 1

Group Members

Lokendra Singh Gohil	2022EE11164
Raunit Kumar Singh	2022TT12160
Yuvraj	2022MS11902
Lakshya Kumar	2022TT12183

Indian Institute of Technology, Delhi

November 13, 2025

Solutions

Problem 1

(a) The OLS method minimizes the sum of squared residuals:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Taking partial derivatives and setting them equal to zero:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

This gives us: $\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

This gives us: $\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$

Solving this system of equations: We get

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

where $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ and $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

(b) Computing $\mathbf{X'X}$:

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix}$$

Computing $\mathbf{X'Y}$:

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix}$$

Inverse of $\mathbf{X}'\mathbf{X}$:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \begin{bmatrix} \sum_{i=1}^{n} X_i^2 & -\sum_{i=1}^{n} X_i \\ -\sum_{i=1}^{n} X_i & n \end{bmatrix}$$

Therefore:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

gives the estimators:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}$$

Using:

$$n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2 = n\sum_{i=1}^{n}(X_i - \bar{X})^2$$

and

$$n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i = n\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

We get

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Thus the given vector expression is the same as the OLS estimators found in (a).

Problem 2

(a) We have the equations:

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \tag{1}$$

$$\ln\left(\frac{Y_i}{X_i}\right) = \alpha_1 + \alpha_2 \ln X_i + \nu_i \tag{2}$$

Equation (2) can be rewritten as:

$$\ln Y_i - \ln X_i = \alpha_1 + \alpha_2 \ln X_i + \nu_i \implies \ln Y_i = \alpha_1 + (\alpha_2 + 1)\ln X_i + \nu_i.$$

Comparing with equation (1), we get:

$$\beta_1 = \alpha_1, \quad \beta_2 = \alpha_2 + 1 \quad u_i = \nu_i$$

Yes, equation (2) can be expressed as a restricted version of equation (1), where the restriction is:

$$\beta_2 = \alpha_2 + 1$$

(b) Let $\hat{\beta}_1, \hat{\beta}_2$ be OLS estimates from (1) and $\hat{\alpha}_1, \hat{\alpha}_2$ from (2). Then:

$$\hat{\beta}_1 = \hat{\alpha}_1, \quad \hat{\beta}_2 = \hat{\alpha}_2 + 1.$$

(c) (Assuming the question meant $\hat{z} = \hat{y} - x$)

$$y = \ln Y, \qquad x = \ln X, \qquad z = \ln\left(\frac{Y}{X}\right) = y - x.$$

From the y-regression:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x, \qquad \hat{u} = y - \hat{y}.$$

From the z-regression:

$$\hat{z} = \hat{\alpha}_1 + \hat{\alpha}_2 x, \qquad \hat{\nu} = z - \hat{z}.$$

Since

$$z = y - x,$$

comparing the model forms gives

$$\alpha_1 = \beta_1, \qquad \alpha_2 = \beta_2 - 1.$$

Thus, for OLS estimates:

$$\hat{\alpha}_1 = \hat{\beta}_1, \qquad \hat{\alpha}_2 = \hat{\beta}_2 - 1.$$

$$\hat{z} = \hat{\alpha}_1 + \hat{\alpha}_2 x = \hat{\beta}_1 + (\hat{\beta}_2 - 1)x = (\hat{\beta}_1 + \hat{\beta}_2 x) - x = \hat{y} - x.$$

(d) The residuals from (2) are:

$$\nu_i = z_i - (\hat{\alpha}_1 + \hat{\alpha}_2 x_i) = (y_i - x_i) - (\hat{\beta}_1 + (\hat{\beta}_2 - 1)x_i) = y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i) = u_i.$$

Hence, the residuals from both regressions are identical.

(e) The regressors of the 2 equations are related as:

$$\hat{\alpha}_1 = \hat{\beta}_1, \qquad \hat{\alpha}_2 = \hat{\beta}_2 - 1.$$

Since the residuals are identical and the regressors in equation (2) are a linear transformation of the regressors in equation (1), the OLS formula for the variance of the estimators:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

implies that the estimated standard errors for $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\alpha}_1, \hat{\alpha}_2$ are identical. Therefore, the standard errors of the coefficients are the same.

(f) For a regression with dependent variable z_i and fitted values $\widehat{z}_i$,

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_i (z_i - \widehat{z}_i)^2}{\sum_i (z_i - \overline{z})^2},$$

where RSS is the residual sum of squares and TSS is the total sum of squares of the dependent variable.

The two regressions have different dependent variables because of which the total sum of squares $\sum_i (z_i - \overline{z})^2$ is different between the two regressions in general. Moreover, OLS produces different fitted values and different residual in general. Hence both the numerator (SSR) and the denominator (TSS) in the formula for R^2 will typically be different for the two regressions. Therefore R^2 would not be the same for the two regressions.

Problem 3

Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm

df = pd.read_csv("healthcare.csv")

# Part (a)
X = sm.add_constant(df['EDUC'])
y = df['HHNINC']
model = sm.OLS(y, X).fit()
beta0_hat = model.params['const']
beta1_hat = model.params['EDUC']

print("beta0_hat = ", beta0_hat , "; beta1_hat = ", beta1_hat)

plt.figure(figsize=(7,4))
plt.axvline(beta1_hat, color='red', linewidth=2)
plt.yticks([])
plt.xlabel(r'Estimates of $\beta_1$')
plt.title("$\hat{\\beta}_1$ for the complete sample as a vertical
   line")
plt.show()

# Part (b)

b1 = []

np.random.seed(123)
```

```
for i in range(1000):
    samp = df.sample(30, replace=False)
    X_samp = sm.add_constant(samp['EDUC'])
    y_samp = samp['HHNINC']
    m = sm.OLS(y_samp, X_samp).fit()
    b1.append(m.params['EDUC'])

plt.figure(figsize=(7,4))
plt.hist(b1, bins=30, density=True, alpha=0.6)
plt.axvline(beta1_hat, color='red', linewidth=2)
plt.xlabel(r'Estimates of $\beta_1$')
plt.title(r"Empirical Distribution of $\mathbf{b_1}$ (sample size =
   30)")
plt.show()
```

(a) We find $\hat{\beta}_0 = 0.12609026767785442$ *and* $\hat{\beta}_1 = 0.019962964033458944$
Code output:

```
beta0_hat =  0.12609026767785442 ; beta1_hat =  0.019962964033458944
```

Plot:

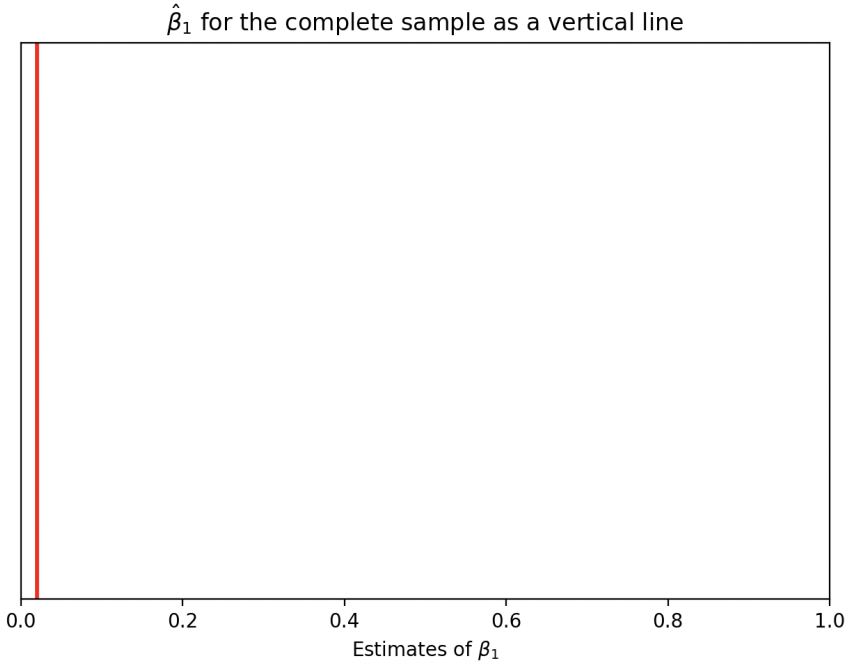

Figure 1: $\hat{\beta}_1$ as a vertical line

(b) Plot:

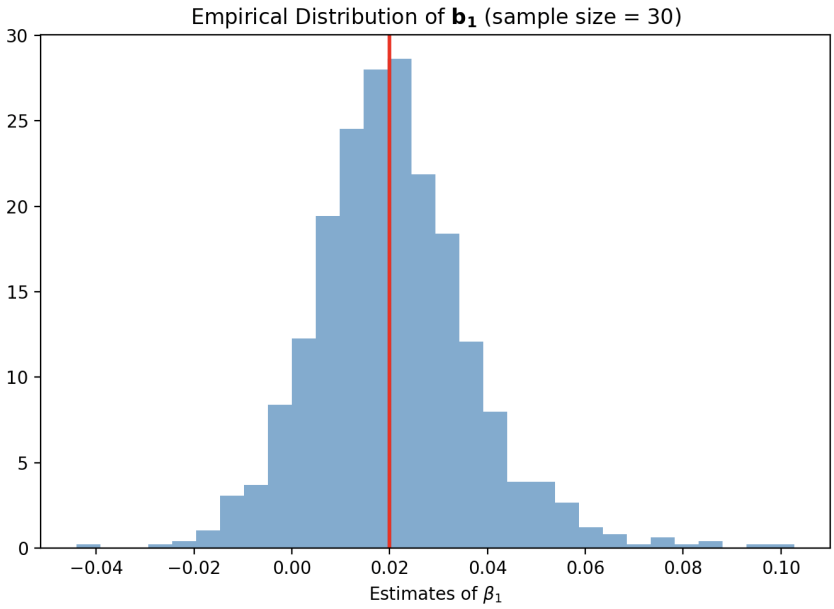

Figure 2: Empirical distribution of $\mathbf{b}_1$

(c) The distribution of estimates of $\hat{\beta}_1$ is centered around the true value 0.0199 and is roughly bell-shaped. This shows that $\hat{\beta}_1$ is an unbiased estimator, with some natural sampling variance.

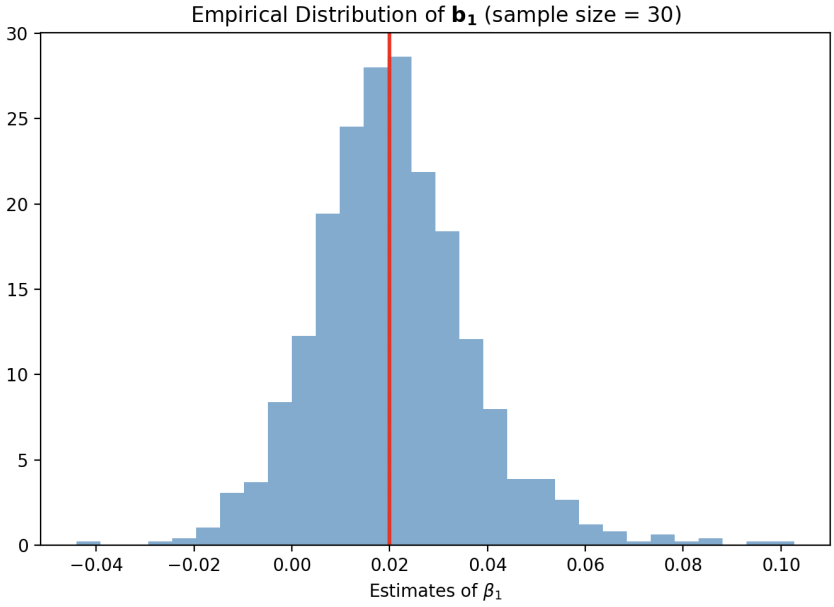

Figure 3: Empirical distribution of $\hat{\beta}_1$.

Problem 4

(a) The model in Equation (3) was estimated using R. The console output for the model summary is shown in Figure 4.

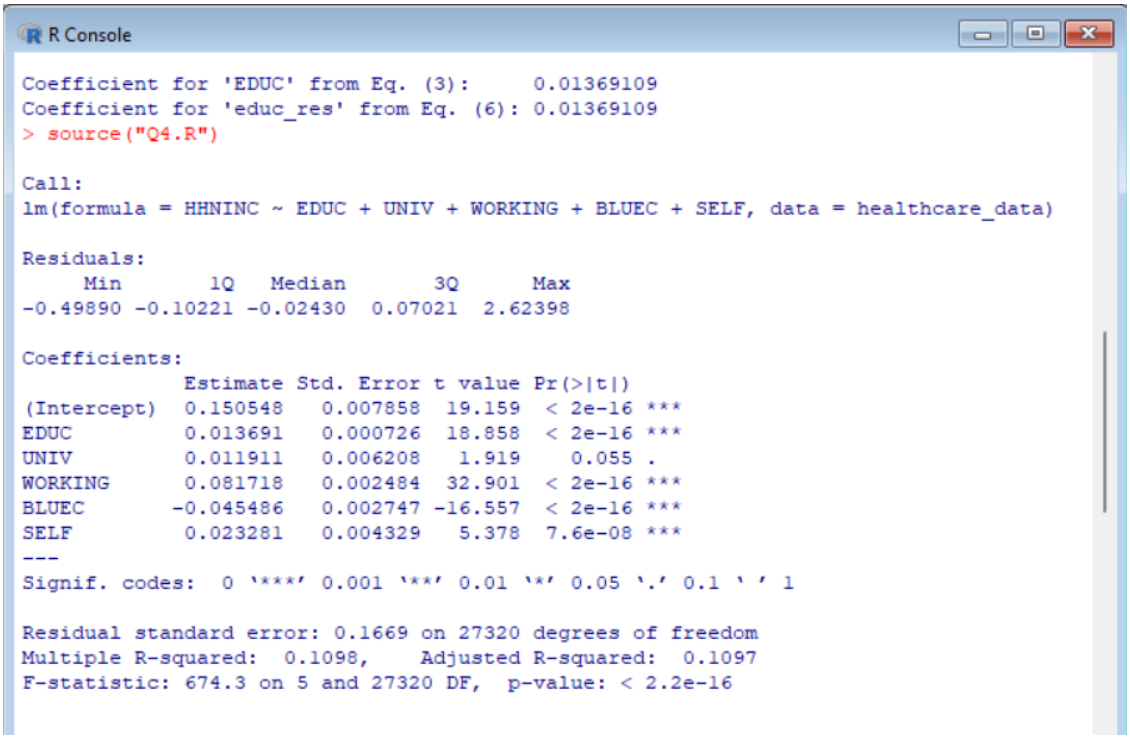

Figure 4: R Console Output for Equation (3)

Based on this output, the point estimate for `EDUC` is $\hat{\beta}_1 = \mathbf{0.013691}$.

(b) The models in Equations (4) and (5) were estimated to generate the residuals. The console summaries for these intermediate regressions are shown in Figures 5 and 6. The residuals `hhninc_res` and `educ_res` were stored.

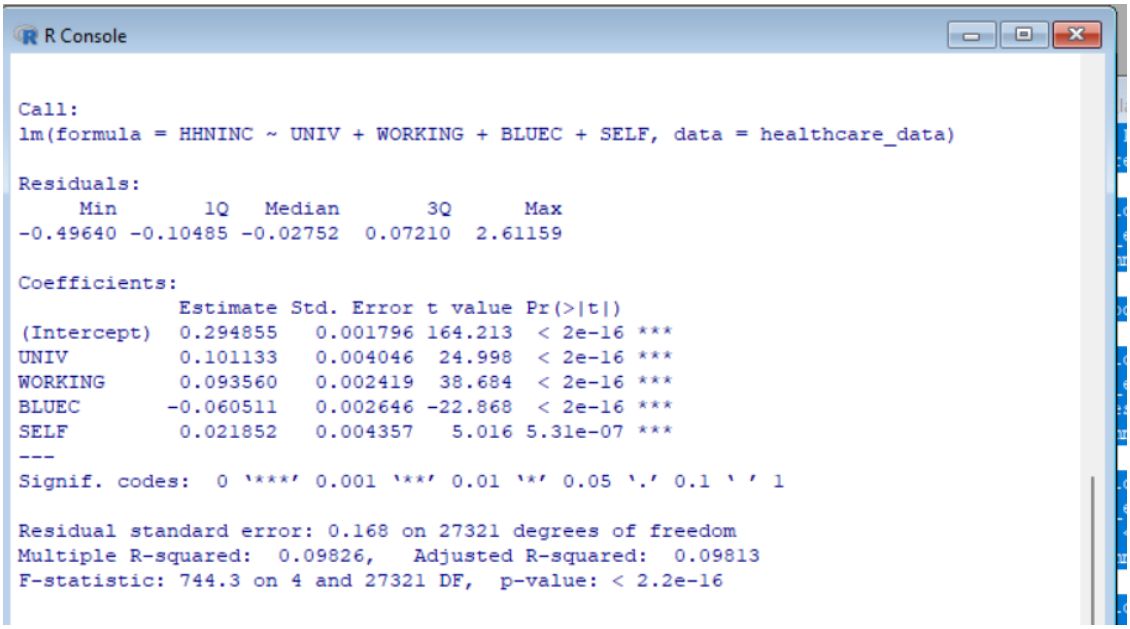

Figure 5: R Console Output for Equation (4)

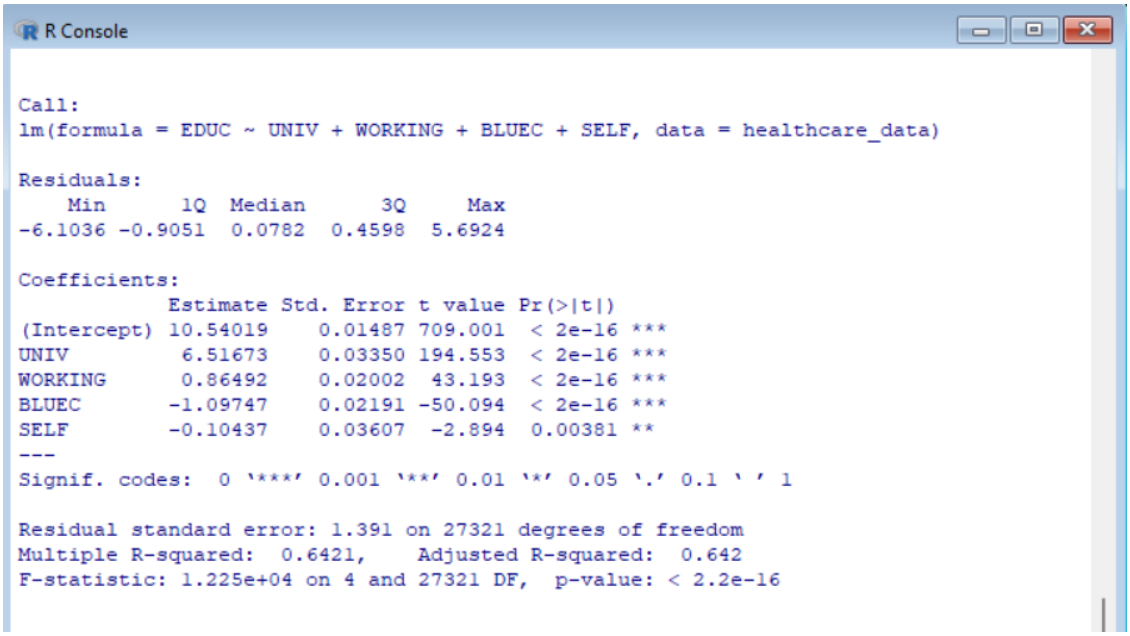

Figure 6: R Console Output for Equation (5)

(c) The model in Equation (6), which regresses `hhninc_res` on `educ_res`, was estimated. The R console output for this regression is shown in Figure 7.

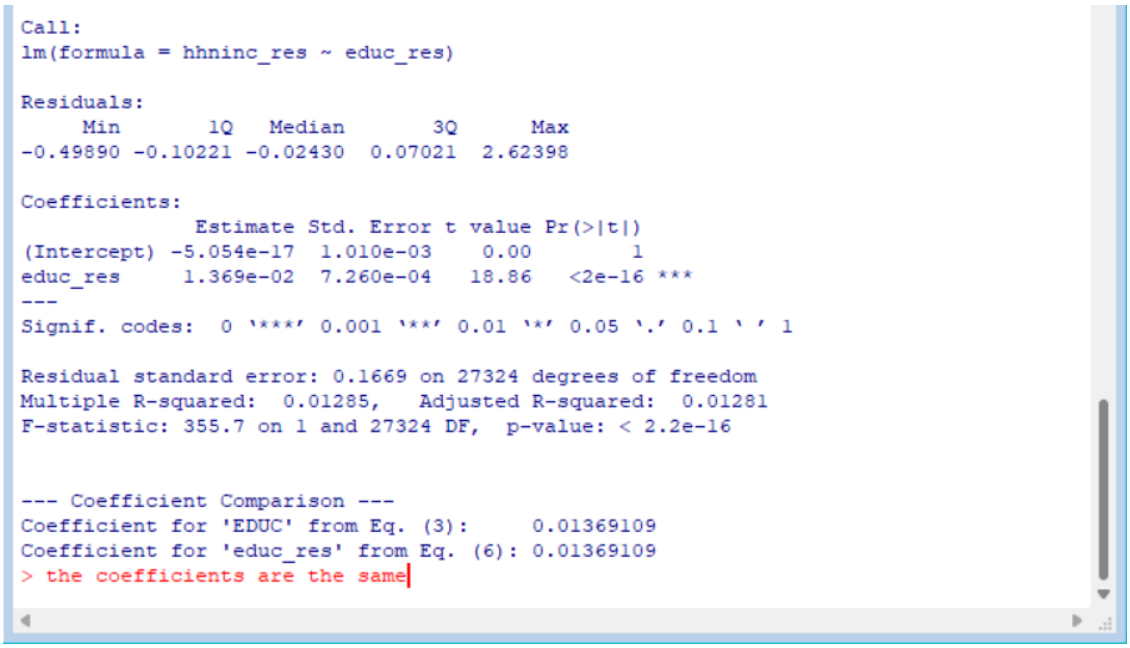

Figure 7: R Console Output for Equation (6) and Final Comparison

The point estimate for `educ_res` is **1.369e − 02** (which is **0.01369109**). The intercept is -5.054×10^{-17}, which is effectively zero, as predicted by theory.

(d) *Comparison:*

As shown in the R console's final `--- Coefficient Comparison ---` output (at the bottom of Figure 7):

```
--- Coefficient Comparison ---
Coefficient for 'EDUC' from Eq. (3):     0.01369109
Coefficient for 'educ_res' from Eq. (6): 0.01369109
```

The point estimates are numerically identical. This result is a direct demonstration of the **Frisch-Waugh-Lovell (FWL) Theorem**.

The FWL theorem states that the coefficient for any single variable in a multiple linear regression (like $\hat{\beta}_1$ for `EDUC` in Eq. 3) can be obtained through a three-step "partialling out" process:

1. Regress the dependent variable (Y, or `HHNINC`) on all other covariates (W, or `UNIV`, `WORKING`, `BLUEC`, `SELF`). This is our Equation (4). The residuals (`hhninc_res`) represent the portion of `HHNINC` that *cannot* be explained by the other controls.

2. Regress the variable of interest (X_1, or `EDUC`) on all other covariates (W). This is our Equation (5). The residuals (`educ_res`) represent the portion of `EDUC` that is *orthogonal to* (i.e., uncorrelated with) the other controls.

3. Regress the residuals from Step 1 on the residuals from Step 2. This is our Equation (6).

The FWL theorem proves that the slope coefficient from this final residual regression (Step 3) is mathematically identical to the coefficient for the variable of interest (X_1) in the original, full multiple regression (Eq. 3). Our R output confirms this.

Listing 1: R Code for Question 4

```r
# i have kept the healthcare.csv in data folder. Loading the data
   into healthcare_data dataframe
healthcare_data <- read.csv("../data/healthcare.csv")

#estimation of the equation 3
model_in_eq_3 <- lm(HHNINC ~ EDUC + UNIV + WORKING + BLUEC + SELF,
   data = healthcare_data)
print(summary(model_in_eq_3))

beta_1_model_3 <- coef(model_in_eq_3)["EDUC"]

#estimation of the equation 4
model_in_eq_4 <- lm(HHNINC ~ UNIV + WORKING + BLUEC + SELF, data =
   healthcare_data)
hhninc_res <- resid(model_in_eq_4)
print(summary(model_in_eq_4))

#estimation of the equation 5
model_in_eq_5 <- lm(EDUC ~ UNIV + WORKING + BLUEC + SELF, data =
   healthcare_data)
educ_res <- resid(model_in_eq_5)
print(summary(model_in_eq_5))

#estimation of the equation 6
model_in_eq_6 <- lm(hhninc_res ~ educ_res)
print(summary(model_in_eq_6))

beta_1_model_6 <- coef(model_in_eq_6)["educ_res"]

#display the results.
cat("\n--- Coefficient Comparison ---\n")
cat("Coefficient for 'EDUC' from Eq. (3):    ", beta_1_model_3, "\n"
   )
cat("Coefficient for 'educ_res' from Eq. (6):", beta_1_model_6, "\n"
   )
```

Problem 5

Consider the model

$$\ln(wage)_{it} = \beta_0 + \beta_1 educ_i + \beta_2 exper_{it} + \beta_3 exper_{it}^2 + c_i + u_{it},$$

for individuals $i = 1, \ldots, n$ observed over $t = 1, \ldots, T$. Here c_i is an unobserved

time-invariant individual effect and u_{it} is idiosyncratic noise.

(a) Pooled OLS would estimate the model ignoring c_i. If c_i is correlated with any regressors included in the model (in particular with $educ_i$), then the omission of c_i creates an omitted variable bias. Formally, the OLS estimate of β_1 will converge to

$$\beta_1^{OLS} = \beta_1 + \frac{\text{Cov}(educ_i, c_i)}{\text{Var}(educ_i)} \quad \text{(plus small-sample/noise terms)},$$

so if $\text{Cov}(educ_i, c_i) \neq 0$ the pooled OLS is biased. Intuitively, c_i may capture intrinsic ability or family background that affects both education and wages; omitting it will pick up part of the effect of these unobservables and attribute it to education.

Kind of omitted-variable problem: Omitted time-invariant individual heterogeneity (fixed effect) that is correlated with a regressor (here $educ_i$). This is a classical omitted-variable bias problem.

(b Define individual averages over time

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} \ln(wage)_{it}, \qquad \bar{exper}_i = \frac{1}{T}\sum_{t=1}^{T} exper_{it}, \qquad \overline{exper^2}_i = \frac{1}{T}\sum_{t=1}^{T} exper_{it}^2,$$

and $\bar{u}_i = \frac{1}{T}\sum_t u_{it}$. Taking individual means of the structural equation:

$$\bar{y}_i = \beta_0 + \beta_1 educ_i + \beta_2 \bar{exper}_i + \beta_3 \overline{exper^2}_i + c_i + \bar{u}_i.$$

Now subtract the mean equation from the time-t equation:

$$\begin{aligned}\ln(wage)_{it} - \bar{y}_i = \beta_0 - \beta_0 + \beta_1(educ_i - \overline{educ}_i) + \beta_2(exper_{it} - \bar{exper}_i) \\ + \beta_3(exper_{it}^2 - \overline{exper^2}_i) + (c_i - c_i) + (u_{it} - \bar{u}_i).\end{aligned}$$

Because $educ_i$ is time-invariant, $\overline{educ}_i = educ_i$ and thus $educ_i - \overline{educ}_i = 0$. Therefore the within-transformed equation simplifies to

$$\ln(wage)_{it} - \bar{y}_i = \beta_2(exper_{it} - \bar{exper}_i) + \beta_3(exper_{it}^2 - \overline{exper^2}_i) + (u_{it} - \bar{u}_i).$$

(c) Including individual-specific intercepts (dummy variables for each i) in the regression is equivalent to the within transformation. The model becomes:

$$\ln(\text{wage})_{it} = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{exper}_{it}^2 + \sum_{i=1}^{n} \gamma_i D_i + u_{it},$$

where D_i is a dummy equal to 1 for individual i. The γ_i absorb $c_i + \beta_1 \text{educ}_i$ (time-invariant components), so OLS on this augmented model projects out the individual means, yielding the same estimates as the demeaning approach. This is computationally equivalent but uses more parameters (n dummies vs. implicit demeaning), making demeaning more efficient for large n.

(d) Including time dummies δ_t for each period t (e.g., year fixed effects) controls for unobserved time-varying heterogeneity common to all individuals, such as aggregate economic shocks (e.g., recessions affecting all wages in year t) or policy changes. The model becomes:

$$\ln(\text{wage})_{it} = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{exper}_{it}^2 + \sum_{t=1}^{T} \delta_t D_t + c_i + u_{it}.$$

This is useful in empirical work because it removes common trends or shocks that could bias coefficients (e.g., inflation in a given year confounding wage growth), improving identification of individual-specific effects like experience.

Problem 6

Structural model:

$$y_i = \alpha_0 + \alpha_s s_i + u_i,$$

where y_i is earnings and s_i is veteran status (actual service). Instrument: draft eligibility $d_i \in \{0, 1\}$, first stage

$$s_i = \beta_0 + \beta_s d_i + \nu_i.$$

We are to show that the instrumental variables (IV) estimate for α_s equals

$$\hat{\alpha}_s = \frac{\overline{y}_{d=1} - \overline{y}_{d=0}}{\overline{s}_{d=1} - \overline{s}_{d=0}},$$

where $\overline{y}_{d=1}$ and $\overline{y}_{d=0}$ denote mean earnings among draft-eligible and draft-ineligible individuals respectively, and similarly $\overline{s}_{d=1}, \overline{s}_{d=0}$ are mean service (proportions who served) in each group.

Derivation (Wald estimator): The IV estimator (for a single instrument and single endogenous regressor) can be written as

$$\hat{\alpha}_s^{IV} = \frac{\widehat{\text{Cov}}(y, d)}{\widehat{\text{Cov}}(s, d)}.$$

When d is binary, a simple manipulation shows the covariance form equals a difference in group means. Let $p = \Pr(d = 1)$. Then sample covariance can be expressed (population notation) as

$$\text{Cov}(y, d) = \Pr(d = 1)\big(E[y \mid d = 1] - E[y]\big) + \Pr(d = 0)\big(E[y \mid d = 0] - E[y]\big).$$

But because $E[y] = \Pr(d = 1)E[y \mid d = 1] + \Pr(d = 0)E[y \mid d = 0]$, algebra simplifies to

$$\text{Cov}(y, d) = \Pr(d = 1)\Pr(d = 0)\big(E[y \mid d = 1] - E[y \mid d = 0]\big).$$

Similarly,

$$\text{Cov}(s, d) = \Pr(d = 1)\Pr(d = 0)\big(E[s \mid d = 1] - E[s \mid d = 0]\big).$$

Taking the ratio cancels $\Pr(d=1)\Pr(d=0)$, yielding the Wald formula

$$\hat{\alpha}_s^{IV} = \frac{E[y \mid d=1] - E[y \mid d=0]}{E[s \mid d=1] - E[s \mid d=0]},$$

which in sample form is

$$\hat{\alpha}_s = \frac{\overline{y}_{d=1} - \overline{y}_{d=0}}{\overline{s}_{d=1} - \overline{s}_{d=0}}.$$

Problem 7

(a) The model is

$$y_{it} = 1\{y_{it}^* > 0\}, \qquad y_{it}^* = x_{it}'\beta + u_{it}, \qquad u_{it} \sim \text{Logistic}(0,1).$$

The probability of attendance is

$$\Pr(y_{it}=1 \mid x_{it}) = \Pr(y_{it}^* > 0 \mid x_{it}) = \Pr(u_{it} > -x_{it}'\beta).$$

Using the logistic CDF

$$\Lambda(z) = \frac{1}{1+e^{-z}},$$

we obtain

$$\Pr(y_{it}=1 \mid x_{it}) = \Lambda(x_{it}'\beta) = \frac{1}{1+\exp(-x_{it}'\beta)}.$$

(b) The probability from part (a) is

$$P \equiv \Pr(y_{it}=1 \mid x_{it}) = \Lambda(x_{it}'\beta) = \frac{1}{1+\exp(-x_{it}'\beta)}.$$

The odds of attending are

$$\frac{P}{1-P} = \frac{\Lambda(x_{it}'\beta)}{1-\Lambda(x_{it}'\beta)}.$$

Using the logistic CDF's algebra,

$$\frac{P}{1-P} = \frac{1/(1+e^{-x_{it}'\beta})}{1-1/(1+e^{-x_{it}'\beta})} = \frac{1}{e^{-x_{it}'\beta}} = e^{x_{it}'\beta}.$$

Taking logs gives the log-odds (logit):

$$\log\left(\frac{\Pr(y_{it}=1 \mid x_{it})}{1-\Pr(y_{it}=1 \mid x_{it})}\right) = \log\left(e^{x_{it}'\beta}\right) = x_{it}'\beta.$$

Thus the log-odds are linear in the covariates and equal $x_{it}'\beta$.

(c) In the logistic latent-variable model

$$y_{it}^* = \beta_0 + \beta_1 \text{PreMidterm}_t + \beta_2 \text{AnnouncedQuiz}_t + \beta_3 \text{DistLHC}_i + u_{it},$$

the log-odds of attendance equal $x_{it}'\beta$ (see part (b)). Thus the coefficients β_1 and β_2 have the following interpretations.

Interpretation of β_1. β_1 is the change in the log-odds of attending a lecture when the lecture is held *before the midterm* ($\text{PreMidterm}_t = 1$) relative to the baseline ($\text{PreMidterm}_t = 0$), holding the other covariates fixed. Equivalently, a one-unit change in PreMidterm (i.e. from 0 to 1) multiplies the *odds* of attendance by e^{β_1}.

$$\frac{\Pr(y_{it} = 1 \mid \text{PreMidterm}_t = 1, \cdot)}{\Pr(y_{it} = 0 \mid \text{PreMidterm}_t = 1, \cdot)} \Bigg/ \frac{\Pr(y_{it} = 1 \mid \text{PreMidterm}_t = 0, \cdot)}{\Pr(y_{it} = 0 \mid \text{PreMidterm}_t = 0, \cdot)} = e^{\beta_1}.$$

So $\beta_1 > 0$ implies that, all else equal, attendance is more likely (higher odds) in pre-midterm lectures.

Interpretation of β_2. β_2 is the change in the log-odds of attendance when a quiz is *pre-announced* ($\text{AnnouncedQuiz}_t = 1$) relative to when it is not ($\text{AnnouncedQuiz}_t = 0$), holding other covariates fixed. A one-unit increase in AnnouncedQuiz multiplies the odds of attendance by e^{β_2}. Thus $\beta_2 > 0$ implies that announcing quizzes in advance increases the likelihood (odds) that students attend that lecture.

(d)

The variable DistLHC_i is a time-invariant student-specific characteristic that is likely negatively correlated with attendance. Omitting it causes omitted variable bias in $\hat{\beta}_1$ and $\hat{\beta}_2$.

Proposed method: Estimate the logit model with individual student fixed effects:

$$y_{it}^* = \alpha_i + \beta_1 \text{PreMidterm}_t + \beta_2 \text{AnnouncedQuiz}_t + u_{it}$$

where α_i absorbs DistLHC_i and all other time-invariant student heterogeneity (laziness, motivation, hostel location, etc.).

Since DistLHC_i is constant within each student, the fixed effects completely eliminate bias from this (and any other) unobserved time-invariant confounder. The coefficients β_1 and β_2 are identified from within-student variation: how the same student changes attendance behavior across lectures with different midterm status and quiz announcement policies.

This approach (implemented via conditional logit or linear probability model with student dummies) is the standard and most credible way to address unobserved time-invariant heterogeneity in panel data with binary outcomes.

(e) The probability from (a) is

$$P_{it} \equiv \Pr(y_{it} = 1 \mid x_{it}) = \Lambda(x_{it}'\beta), \qquad \Lambda(z) = \frac{1}{1 + e^{-z}}.$$

The marginal effect of Dist LHC_i on the attendance probability is

$$\frac{\partial \Pr(y_{it} = 1 \mid x_{it})}{\partial \text{Dist LHC}_i} = \frac{\partial \Lambda(x_{it}'\beta)}{\partial x_{it}'\beta} \cdot \frac{\partial (x_{it}'\beta)}{\partial \text{Dist LHC}_i} = \Lambda'(x_{it}'\beta)\, \beta_3,$$

where

$$\Lambda'(z) = \Lambda(z)\{1 - \Lambda(z)\} = \frac{e^{-z}}{(1+e^{-z})^2}.$$

The factor $\Lambda'(x'_{it}\beta)$ depends on the index $x'_{it}\beta$. Since $\Lambda'(z)$ is maximal near $z = 0$ and smaller for very negative or very positive z, the same β_3 produces different absolute probability changes for different values of the covariates x_{it}. Thus the marginal effect of distance is *heterogeneous* across students and lectures because it is scaled by $\Lambda'(x'_{it}\beta)$.

if Dist LHC$_i$ is observed and β_3 is identified/estimated. The sample average marginal effect (AME) for Dist LHC is

$$\text{AME}_{\text{Dist}} \;=\; \frac{1}{N}\sum_{i,t} \Lambda'(x'_{it}\hat{\beta})\,\hat{\beta}_3,$$

which can be compute using estimated $\hat{\beta}$ and the observed x_{it}

(f) The provided rollcall file (one row per student, columns = lecture dates) was reshaped to panel form with one observation per (student i, lecture t). I set

$$\text{PreMidterm}_t = 1\{\text{date}_t \le \text{Sep13, 2025}\}, \qquad \text{AnnouncedQuiz}_t = 1 - \text{PreMidterm}_t,$$

consistent with the experimental design (pre-midterm: unannounced quizzes; post-midterm: pre-announced).

Dist LHC$_i$ is unobserved and time-invariant. To remove bias from this omitted time-invariant heterogeneity we use the within (student fixed-effects) estimator. The (within) linear probability specification is:

$$y_{it} = \alpha_i + \gamma_1 \text{PreMidterm}_t + \gamma_2 \text{AnnouncedQuiz}_t + v_{it},$$

and we estimate the demeaned (within) equation

$$y_{it} - \bar{y}_i = \gamma_1(\text{PreMidterm}_t - \overline{\text{PreMidterm}_i}) + \gamma_2(\text{AnnouncedQuiz}_t - \overline{\text{AnnouncedQuiz}_i}) + \tilde{v}_{it},$$

using OLS (heteroskedasticity-robust SEs).

Estimated coefficients (student FE / within LPM, robust SEs):

Variable	$\widehat{\gamma}$	Std. Err.	p-value
PreMidterm (within)	0.05306	0.00904	< 0.001
AnnouncedQuiz (within)	-0.05306	0.00904	< 0.001

Notes:

- Because $\text{AnnouncedQuiz}_t = 1 - \text{PreMidterm}_t$, the two demeaned regressors are perfectly collinear in the within-transformation (so effectively one degree of freedom). The table shows the mechanically equal and opposite coefficients; interpret the PreMidterm effect.
- $\widehat{\gamma}_{\text{PreMidterm}} \approx 0.053$ implies that, after removing student fixed effects, lecture attendance is about $+\mathbf{5.3}$ percentage points higher for pre-midterm lectures than for post-midterm lectures (statistically significant at conventional levels).

Pooled logit (for comparison). We also estimated a pooled logit $\Pr(y_{it} = 1 \mid x_{it}) = \Lambda(x'_{it}\beta)$ (no student FE). That model produced unstable std errors / convergence warnings when attempting to include student dummies, the pooled logit without FE produced coefficient signs consistent with the within result (pre-midterm positive), but the logit with student FE (by adding many dummies) did not converge.

Note to the Dean Academics. *Finding:* After controlling for time-invariant student heterogeneity (student fixed effects), lectures held before the midterm have a higher attendance rate by ≈ 5.3 percentage points than lectures held after the midterm. This effect is statistically significant.
The timing of the midterm significantly affects attendance. If the Dean wants to raise attendance in post-midterm lectures, consider policies that replicate pre-midterm incentives (i.e. maintain unpredictability of quizzes)

Listing 2: Python Code

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from datetime import datetime

df = pd.read_csv('/content/attendance_rollcall.csv')

# Detect student ID column
def detect_id_col(df):
    for c in df.columns:
        low = c.lower()
        if "user" in low or "student" in low or (low.endswith("id")
           and df[c].nunique() > 1):
            return c
    return df.columns[0]

id_col = detect_id_col(df)
date_cols = [c for c in df.columns if c != id_col]

#Converting wide into long panel
df_long = df.melt(
    id_vars=[id_col],
    value_vars=date_cols,
    var_name='date',
    value_name='attended'
).rename(columns={id_col: 'student_id'})

# attendance to binary
df_long['attended'] = df_long['attended'].astype(str).str.strip().
   str.lower()
```

```
df_long['y'] = df_long['attended'].map({'yes':1,'no':0,'present':1,'
    absent':0})

# drop unparseable
df_long = df_long.dropna(subset=['y'])
df_long['y'] = df_long['y'].astype(int)

#Parsing date
def try_parse(s):
    for fmt in ('%m/%d/%Y','%Y-%m-%d','%d/%m/%Y','%d-%m-%Y'):
        try:
            return pd.to_datetime(s, format=fmt)
        except:
            pass
    return pd.to_datetime(s, errors='coerce')

df_long['date_parsed'] = df_long['date'].apply(try_parse)
df_long = df_long.dropna(subset=['date_parsed'])

#creating variables
midterm_cutoff = pd.to_datetime('2025-09-13')
df_long['pre_midterm'] = (df_long['date_parsed'] <= midterm_cutoff).
    astype(int)
df_long['announced_quiz'] = (df_long['pre_midterm'] == 0).astype(int
    )

#Within (student FE) LPM
grp = df_long.groupby('student_id')

df_long['y_dm'] = df_long['y'] - grp['y'].transform('mean')
df_long['pre_dm'] = df_long['pre_midterm'] - grp['pre_midterm'].
    transform('mean')
df_long['ann_dm'] = df_long['announced_quiz'] - grp['announced_quiz'
    ].transform('mean')

X_within = sm.add_constant(df_long[['pre_dm','ann_dm']])
ols_within = sm.OLS(df_long['y_dm'], X_within).fit(cov_type='HC1')

#Pooled logit
X_pool = sm.add_constant(df_long[['pre_midterm','announced_quiz']])
logit_pool = sm.Logit(df_long['y'], X_pool).fit(disp=False, maxiter
    =100)



print(" WITHIN (STUDENT FE) LPM RESULTS ")
print(ols_within.summary())
```

```
print(" POOLED LOGIT RESULTS (NO FE) ")
print(logit_pool.summary())
```

```
 WITHIN (STUDENT FE) LPM RESULTS
                            OLS Regression Results
==============================================================================
Dep. Variable:                    y_dm   R-squared:                       0.010
Model:                            OLS   Adj. R-squared:                  0.010
Method:                 Least Squares   F-statistic:                     34.47
Date:                Sun, 16 Nov 2025   Prob (F-statistic):           4.80e-09
Time:                        17:27:12   Log-Likelihood:                -1629.2
No. Observations:                2980   AIC:                             3262.
Df Residuals:                    2978   BIC:                             3274.
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.978e-16      0.008  -2.58e-14      1.000      -0.015       0.015
pre_dm         0.0531      0.009      5.871      0.000       0.035       0.071
ann_dm        -0.0531      0.009     -5.871      0.000      -0.071      -0.035
==============================================================================
Omnibus:                      270.156   Durbin-Watson:                   1.720
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              136.599
Skew:                          -0.360   Prob(JB):                     2.18e-30
Kurtosis:                       2.237   Cond. No.                     6.13e+15
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The smallest eigenvalue is 7.93e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
 POOLED LOGIT RESULTS (NO FE)
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 2980
Model:                          Logit   Df Residuals:                     2977
Method:                           MLE   Df Model:                            2
Date:                Sun, 16 Nov 2025   Pseudo R-squ.:                0.005383
Time:                        17:27:12   Log-Likelihood:                -2016.5
converged:                      False   LL-Null:                       -2027.4
Covariance Type:            nonrobust   LLR p-value:                 1.820e-05
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.1300   6.95e+06   1.87e-08      1.000   -1.36e+07    1.36e+07
pre_midterm        0.2800   6.95e+06   4.03e-08      1.000   -1.36e+07    1.36e+07
announced_quiz    -0.1501   6.95e+06  -2.16e-08      1.000   -1.36e+07    1.36e+07
==============================================================================
```

Figure 8: Output