COLUMBIA UNIVERSITY

Department of Economics

Completed in Partial Fulfillment of the Requirements of

ECON UN3412

Introduction to Econometrics

Spring 2026

HOMEWORK 3

Seyhan Erden

seyhan.erden@columbia.edu

Tyler Sotomayor

tyler.sotomayor@columbia.edu

Special thanks to the Teaching Assistants of ECON UN3412

Daniel Garcia

dg3342@columbia.edu

Chris Kye

mk4449@columbia.edu

Shivesh Babu Narayan

sb5211@columbia.edu

Luca Salvatori

ls4241@columbia.edu

Zijian Zhang

zz3263@columbia.edu

ECON UN3412 Spring 2026: Homework 3 Tyler Sotomayor

Introduction to Econometrics tyler.sotomayor@columbia.edu

Instructor: Seyhan Erden Due Date: 26 February 2026

Problem 1 [40 points]

Use Table 2 and the `GPA4.dta` data file to answer the following questions.

(a) [8 points] Table 2 presents the results of three regressions, one in each column. Estimate the indicated regressions and fill in the values (you may either handwrite or type the entries in, at your convenience). For example, to fill in column (1), estimate the regression with `colGPA` as the dependent variable and `hsGPA` and `skipped` as the independent variables, using the "robust" option, and fill in the estimated coefficients.

SOLUTION:

See Table 2 in the Appendix. ■

(b) [6 points] Write the regression in column (1) in "equation form," with the standard error below the respective regression coefficient.

SOLUTION:

$$\widehat{colGPA}_i = \underset{(0.325)}{1.579} + \underset{(0.094)}{0.459}\, hsGPA_i - \underset{(0.025)}{0.077}\, skipped_i$$

■

(c) [5 points] Explain in words what the coefficient on `hsGPA` means in regression (1), holding `skipped` unchanged.

SOLUTION:

In regression (1), the coefficient on `hsGPA` is 0.459. Holding `skipped` fixed, this means that a one-point increase in high school GPA is associated with an increase of approximately 0.459 points in college GPA, on average.

In other words, comparing two students who miss the same number of lectures per week, the student with a high school GPA that is one point higher is predicted to have a college GPA that is about 0.459 points higher. ■

(d) [5 points] Using regression (1), test the hypothesis that the coefficient on `skipped` is zero, against the alternative that it is nonzero, at the 5% significance level. In everyday words (not statistical terms), what precisely is the hypothesis that you are testing?

SOLUTION:

We test

$$H_0: \ \beta_{\texttt{skipped}} = 0 \quad \text{vs} \quad H_1: \ \beta_{\texttt{skipped}} \neq 0$$

using regression (1). The estimate is $\widehat{\beta}_{\texttt{skipped}} = -0.077$ with robust standard error 0.025, so the test statistic is

$$t \ = \ \frac{-0.077}{0.025} \ \approx \ -3.05.$$

The corresponding two-sided p-value is $p = 0.003$, so at the 5% significance level we reject H_0 and conclude that `skipped` has a statistically significant relationship with `colGPA` in regression (1).

Essentially, we are testing whether, holding high school GPA fixed, missing an extra lecture per week has *no impact at all* on a student's expected college GPA versus whether missing lectures *does* change expected college GPA (in either direction). ■

(e) [6 points] Test the hypothesis that the coefficient on `skipped` is zero in regressions (1), (2), and (3) at the 1% level using the p-value. Does your answer change depending on what other variables are included in the regression?

SOLUTION:

We test $H_0 : \beta_{\texttt{skipped}} = 0$ versus $H_1 : \beta_{\texttt{skipped}} \neq 0$ using the two-sided p-values from each regression.

- **Regression (1):** $p = 0.003 < 0.01 \Rightarrow$ reject H_0 at the 1% level.
- **Regression (2):** $p = 0.013 > 0.01 \Rightarrow$ fail to reject H_0 at the 1% level.
- **Regression (3):** $p = 0.007 < 0.01 \Rightarrow$ reject H_0 at the 1% level.

Yes, the answer changes depending on which controls are included: `skipped` is significant at the 1% level in (1) and (3), but not in (2). This happens because adding or removing regressors can change both the estimated coefficient and its standard error (and therefore the p-value) by altering what variation in `skipped` is used to identify its partial association with `colGPA`. ■

(f) [5 points] Using regression (3), consider the coefficient on `campus`. Does the sign and magnitude make sense? Explain.

SOLUTION:

In regression (3), the coefficient on `campus` is -0.124 (robust SE 0.079). Since `campus`= 1 indicates living on campus, the estimate implies that, holding `hsGPA`, `skipped`, `PC`, and `bgfriend` fixed, students living on campus have a predicted college GPA about 0.124 points lower than otherwise similar students not living on campus.

The negative sign is plausible: living on campus could be correlated with more social activity, distractions, or less structured study environments, which could lower GPA. The magnitude is also modest (about one-eighth of a GPA point), so it is not an implausibly large effect. However, the estimate is not statistically significant at conventional levels ($p = 0.116$), so we should treat it as suggestive rather than conclusive evidence of a difference. ■

(g) [5 points] Using regression (3), consider the coefficient on `bgfriend`. Does the sign and magnitude make sense? Explain. In regression (3), is the coefficient on `campus` statistically significant at the 1% significance level? Is the coefficient on `bgfriend` statistically significant at the 1% significance level?

SOLUTION:

In regression (3), the coefficient on `bgfriend` is 0.086 (robust SE 0.054). Since `bgfriend`= 1 indicates having a boyfriend/girlfriend, this suggests that, holding `hsGPA`, `skipped`, `PC`, and `campus` fixed, students with a boyfriend/girlfriend have a predicted college GPA about 0.086 points higher than otherwise similar students without one.

The positive sign can make sense if having a partner is associated with greater

emotional support or stability (or if it proxies for other traits correlated with academic performance). The magnitude is small (less than one-tenth of a GPA point), so it is not implausibly large, and could easily reflect selection rather than a causal effect.

At the 1% significance level, neither coefficient is statistically significant. In regression (3), `campus` has $p = 0.116 > 0.01$ and `bgfriend` has $p = 0.115 > 0.01$, so we fail to reject $H_0 : \beta = 0$ for both variables at the 1% level. ■

Problem 2 [25 points]

Nursing Home Utilization. This question considers nursing home data provided by the Wisconsin Department of Health and Family Services (DHFS) and its description is in posted dataset.

2.1 [20 points] Part 1: Use cost-report year 2000 data, and do the following analysis.

(a) **[6 points] Correlations**

(i) [3 points] Calculate the correlation between `TPY` and `LOGTPY`. Comment on your result.

SOLUTION:

Using the year 2000 sample, the correlation between `TPY` and `LOGTPY` is

$$\text{Corr}(\texttt{TPY}, \texttt{LOGTPY}) = 0.9372 \quad (n = 362).$$

This is very close to 1, so `LOGTPY` preserves almost all of the ordering and comovement in `TPY`. The correlation is not exactly 1 because the log transformation is nonlinear: it compresses large values of `TPY` relative to smaller values. ■

(ii) [3 points] Calculate the correlation among `TPY`, `NUMBED`, and `SQRFOOT`. Do these variables appear highly correlated?

SOLUTION:

Using the year 2000 sample, the pairwise correlations are:

$$\text{Corr}(\texttt{TPY}, \texttt{NUMBED}) = 0.9789,$$

$$\text{Corr}(\texttt{TPY}, \texttt{SQRFOOT}) = 0.8244,$$

$$\text{Corr}(\texttt{NUMBED}, \texttt{SQRFOOT}) = 0.8192.$$

based on $n = 357$ complete observations. Yes, these variables appear highly correlated, especially `TPY` and `NUMBED` (0.9789). This is consistent with larger facilities (more beds and more square footage) having higher output/utilization. ■

(b) **[4 points] Scatter plots.** Plot `TPY` versus `NUMBED` and `TPY` versus `SQRFOOT`. Comment on the plots.

SOLUTION:

Figures 1 and 2 in the Appendix display the requested scatter plots.

The plot of `TPY` versus `NUMBED` shows an extremely tight and nearly linear relationship. The observations lie very close to the fitted line, indicating that total patient days are almost proportional to the number of beds. This is consistent with the very high correlation of 0.9789. Economically, this makes sense because the number of beds directly constrains patient capacity.

The plot of `TPY` versus `SQRFOOT` also shows a positive relationship, but with noticeably more dispersion around the fitted line. While larger facilities tend to have higher patient days, square footage is a noisier proxy for operational

capacity than the number of beds. This aligns with the lower (though still high) correlation of 0.8244. ■

(c) **[10 points] Basic linear regression, homoskedactic errors are fine here.**

(i) [2 points] Fit a basic linear regression model using `TPY` as the outcome variable and `NUMBED` as the explanatory variable. Summarize the fit by quoting the coefficient of determination, R^2, and the t-statistic for `NUMBED`.

SOLUTION:

Estimating the regression

$$TPY_i = \beta_0 + \beta_1 NUMBED_i + u_i,$$

yields an R^2 of 0.9586, indicating that approximately 95.9% of the variation in total patient days is explained by the number of beds.

The t-statistic for `NUMBED` is 91.35, which is extremely large in magnitude, indicating that the relationship is statistically very strong. ■

(ii) [3 points] Repeat c(i), using `SQRFOOT` instead of `NUMBED`. In terms of R^2, which model fits better?

SOLUTION:

Estimating

$$TPY_i = \beta_0 + \beta_1 SQRFOOT_i + u_i,$$

produces an R^2 of 0.6797.

Comparing the two models, the regression using `NUMBED` fits substantially better, as it explains about 95.9% of the variation in `TPY`, whereas `SQRFOOT` explains only about 68.0%. Thus, `NUMBED` is a much stronger predictor of total patient days. ■

(iii) [2 points] Repeat c(i), using `LOGTPY` for the outcome variable and `LOG(NUMBED)` as the explanatory variable.

SOLUTION:

Estimating the log–log model

$$\log(TPY_i) = \beta_0 + \beta_1 \log(NUMBED_i) + u_i$$

yields an R^2 of 0.9483 and a t-statistic of 81.23 for $\log(NUMBED)$.

The estimated coefficient on $\log(NUMBED)$ is approximately 1.012, indicating that a 1% increase in the number of beds is associated with roughly a 1.01% increase in total patient days. This suggests an approximately proportional (unit-elastic) relationship. ■

(iv) [3 points] Repeat c(iii) using `LOGTPY` for the outcome variable and `LOG(SQRFOOT)` as the explanatory variable.

SOLUTION:

Estimating

$$\log(TPY_i) = \beta_0 + \beta_1 \log(SQRFOOT_i) + u_i$$

produces an R^2 of 0.6765 and a t-statistic of 27.25.

The estimated elasticity is approximately 0.687, implying that a 1% increase in square footage is associated with a 0.69% increase in total patient days. While statistically significant, this model fits considerably worse than the model using NUMBED. ■

2.2 [5 points] Part 2: Run the same regression as in Part 1.c(i) using 2001 data. Are the patterns stable over time?

Part 2: Run the same regression as in Part 1.c(i) using 2001 data. Are the patterns stable over time?

SOLUTION:

Using the 2001 data, we estimate

$$TPY_i = \beta_0 + \beta_1 NUMBED_i + u_i.$$

The estimated slope coefficient on NUMBED is approximately 0.932, very close to the 2000 estimate of 0.921. The coefficient of determination is $R^2 = 0.9762$, which is even slightly higher than the 2000 value of 0.9586. The t-statistic on NUMBED is 120.39, indicating an extremely strong relationship.

Overall, the results are highly stable over time. In both years, total patient days are almost perfectly proportional to the number of beds, and the fit of the model remains extremely strong. ■

Problem 3 [35 points]

Consider the following Population Linear Regression Function (PLRF):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i \tag{1}$$

where, Y_i = average hourly earnings/wage in \$, X_1 = years of education, X_2 = years of potential experience, X_3 = years with current employer (tenure), $X_4 = 1$ if female, $X_5 = 1$ if nonwhite, and u_i = the usual error term of the model.

For this question, use the WAGE data set. Here is the description of the variables in the dataset for your consumption. We might be using this data set for the coming problem sets too.

Observations: 526 Variables: 21

1. wage average hourly earnings
2. educ years of education
3. exper years potential experience
4. tenure years with current employer
5. nonwhite =1 if nonwhite
6. female =1 if female
7. married =1 if married
8. numdep number of dependents
9. smsa =1 if live in SMSA
10. northcen =1 if live in north central U.S
11. south =1 if live in southern region

12. west =1 if live in western region
13. construc =1 if work in construc. Indus.
14. ndurman =1 if in nondur. Manuf. Indus.
15. trcommpu =1 if in trans, commun, pub ut
16. trade =1 if in wholesale or retail
17. services =1 if in services indus.
18. profserv =1 if in prof. serv. Indus.
19. profocc =1 if in profess. Occupation
20. clerocc =1 if in clerical occupation
21. servocc =1 if in service occupation

(a) [6 points] Consider the following restricted version of equation (1)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i.$$

Suppose that X_2 is omitted from the model by the researcher. For X_2 to cause omitted variable bias (OVB), what conditions should it satisfy?

SOLUTION:

Let the true model be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,$$

but suppose the researcher omits X_{2i} and estimates Y_i on X_{1i} only. Then X_2

generates omitted variable bias in the estimated coefficient on X_1 if both of the following hold:

(i) X_2 **matters for** Y**:** $\beta_2 \neq 0$ (equivalently, X_2 belongs in the true conditional mean of Y).

(ii) X_2 **is correlated with the included regressor** X_1**:** $\text{Cov}(X_1, X_2) \neq 0$ (equivalently, $E[X_2 \mid X_1]$ varies with X_1).

If either condition fails, omitting X_2 does not bias the OLS slope on X_1: if $\beta_2 = 0$, X_2 has no direct effect on Y; if $\text{Cov}(X_1, X_2) = 0$, the omitted component $\beta_2 X_2$ is uncorrelated with X_1 and is absorbed into the error term without creating endogeneity. ■

(b) [6 points] Run a regression of $Y_i = \beta_0 + \beta_4 X_{4i} + u_i$ and interpret the slope coefficient β_4. (Hint: X_4 is a binary explanatory variable.)

SOLUTION:

We estimate the regression

$$wage_i = \beta_0 + \beta_4 female_i + u_i.$$

The estimated equation is

$$\widehat{wage}_i = \underset{(0.210)}{7.099} - \underset{(0.303)}{2.512}\, female_i.$$

Because `female` is a binary variable equal to 1 for women and 0 for men:

- The intercept, 7.099, is the average hourly wage for men.
- The slope coefficient, -2.512, measures the difference in average wages between women and men.

Thus, women earn on average approximately \$2.51 less per hour than men in this sample. Equivalently, the average female wage is about \$4.59 per hour.

The negative sign indicates that women earn less than men on average. ■

(c) [6 points] First generate a dummy variable D_i, such that $D_i = 1$ if male and $D_i = 0$ if female. Then run a regression of

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_4 X_{4i} + \beta_6 D_i + u_i.$$

What do you notice in the result? Explain why? Show mathematically that if X_4 and D_i are related, this result is inevitable.

SOLUTION:

We generate the dummy variable D_i such that $D_i = 1$ if male and $D_i = 0$ if female. Since $X_{4i} = \texttt{female}_i$ equals 1 for females and 0 for males, we have the exact relationship

$$D_i = 1 - X_{4i}.$$

When we estimate

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_4 X_{4i} + \beta_6 D_i + u_i,$$

Stata reports that one of the dummy variables (here, `male`) is omitted because of perfect collinearity.

Why this occurs:

Substituting $D_i = 1 - X_{4i}$ into the regression gives

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_4 X_{4i} + \beta_6(1 - X_{4i}) + u_i,$$

which simplifies to

$$wage_i = (\beta_0 + \beta_6) + \beta_1 educ_i + (\beta_4 - \beta_6) X_{4i} + u_i.$$

Thus, only the combinations $(\beta_0 + \beta_6)$ and $(\beta_4 - \beta_6)$ affect the fitted values. The three parameters β_0, β_4, and β_6 cannot be separately identified.

Since D_i and X_{4i} satisfy the exact linear relationship

$$D_i + X_{4i} = 1,$$

the regressors are perfectly collinear. Therefore, the matrix $X'X$ is singular, and OLS must drop one of the variables.

This result is inevitable whenever two dummy variables partition the same category (male/female) while an intercept is included in the regression. ■

(d) [6 points] Run, first, a simple regression of $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$, then

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i.$$

Explain what happened to β_1 (before and after) and why it happened.

SOLUTION:

In the simple regression

$$wage_i = \beta_0 + \beta_1 educ_i + u_i,$$

the estimated coefficient on education is $\widehat{\beta}_1 = 0.541$.

After adding experience,

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + u_i,$$

the estimated education coefficient increases to $\widehat{\beta}_1 = 0.644$ (and $\widehat{\beta}_2 = 0.070$).

This change occurs because in the simple regression, `exper` is omitted and absorbed into the error term. If experience affects wages ($\beta_2 \neq 0$) and is correlated with education, then the simple regression suffers omitted variable bias. A useful heuristic is

$$\widehat{\beta}_1^{\text{simple}} \approx \beta_1 + \beta_2 \cdot \frac{\text{Cov}(educ,\, exper)}{\text{Var}(educ)}.$$

Here, $\widehat{\beta}_2 > 0$, and the fact that $\widehat{\beta}_1$ rises when `exper` is included implies that $\text{Cov}(educ, exper) < 0$ in this sample: more educated workers tend to have less potential experience (e.g., they spent more years in school). Omitting `exper` therefore biases the simple education slope downward, and controlling for experience removes that bias and raises the estimated return to education. ■

(e) [5 points] Now run the full model (1), using both homoskedasticity-only and heteroskedasticity-robust standard errors, and interpret and compare the results of both regressions. Why

do we care about heteroskedasticity problem that might exist in the data?

SOLUTION:

We estimate the full model

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + \beta_4 female_i + \beta_5 nonwhite_i + u_i$$

using both classical (homoskedasticity-only) and heteroskedasticity-robust standard errors. The regression results are reported in Table 3 in the Appendix.

Interpretation of coefficients:

Holding other factors fixed:

- Each additional year of education increases hourly wages by about \$0.57.
- Each additional year of experience increases wages by about \$0.025.
- Each additional year of tenure increases wages by about \$0.14.
- Women earn about \$1.81 less per hour than comparable men.
- The coefficient on `nonwhite` is small and statistically insignificant.

Comparison of classical and robust results:

The estimated coefficients and R^2 are identical across the two regressions, as expected. However, the standard errors differ. In several cases (e.g., education and tenure), the robust standard errors are larger than the classical ones, which slightly reduces the magnitude of the t-statistics. The intercept also becomes less statistically significant under robust standard errors.

Why we care about heteroskedasticity:

If heteroskedasticity is present, the usual OLS standard errors are biased and lead to invalid inference (incorrect t-tests and confidence intervals). Although OLS coefficients remain unbiased and consistent, the classical variance formula assumes constant error variance. Robust standard errors correct for potential heteroskedasticity and provide valid inference even when the variance of u_i is not constant.

Thus, robust standard errors are generally preferred in applied work because they protect against incorrect inference due to heteroskedasticity. ■

(f) [6 points] Based on the regression result of the latter (i.e., heteroskedasticity-robust standard errors), conduct the following hypothesis testing:

(i) H_0: $\beta_i = 0$ vs H_1: $\beta_i \neq 0$ where $i = 1, 2, \ldots, 5$

(ii) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ vs H_1: At least one $\beta_i \neq 0$

SOLUTION:

(i) Here we test $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$ for $i = 1, \ldots, 5$ using the robust regression results.

- **Education (β_1):** $F(1, 520) = 86.49$, $p < 0.001$. Reject H_0. Education is statistically significant.
- **Experience (β_2):** $F(1, 520) = 6.67$, $p = 0.0101$. Reject H_0 at the 5% level.
- **Tenure (β_3):** $F(1, 520) = 25.40$, $p < 0.001$. Reject H_0.
- **Female (β_4):** $F(1, 520) = 50.68$, $p < 0.001$. Reject H_0.

- **Nonwhite (β_5):** $F(1, 520) = 0.09$, $p = 0.7680$. Fail to reject H_0; this coefficient is not statistically significant.

(ii) For the joint hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \quad \text{vs} \quad H_1 : \text{at least one } \beta_i \neq 0,$$

the robust joint test yields

$$F(5, 520) = 35.62, \quad p < 0.001.$$

We reject the null hypothesis. The regressors are jointly statistically significant, meaning that at least one of the explanatory variables has a nonzero effect on wages.

The robust hypothesis tests clarify the distinction between individual and joint significance. Education, experience, tenure, and gender are each individually statistically significant predictors of wages, while race (nonwhite) is not statistically distinguishable from zero in this specification. However, the strong rejection of the joint null hypothesis demonstrates that the regressors collectively have substantial explanatory power. This highlights an important econometric principle: even if some individual variables are insignificant, the model as a whole can be highly informative. Using heteroskedasticity-robust standard errors ensures that these conclusions about statistical significance remain valid even if the error variance is not constant. ■

Appendix: Tables and Figures

Table 2: College GPA Results (Dependent variable: `colGPA`)

	(1)	(2)	(3)
hsGPA	0.459***	0.455***	0.461***
	(0.094)	(0.093)	(0.090)
skipped	−0.077***	−0.065**	−0.071***
	(0.025)	(0.026)	(0.026)
PC		0.129**	0.137**
		(0.060)	(0.059)
bgfriend			0.086
			(0.054)
campus			−0.124
			(0.079)
Intercept	1.579***	1.527***	1.490***
	(0.325)	(0.321)	(0.317)
n	141	141	141
R-squared	0.223	0.250	0.278
Regression RMSE	0.331	0.326	0.322

Robust standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

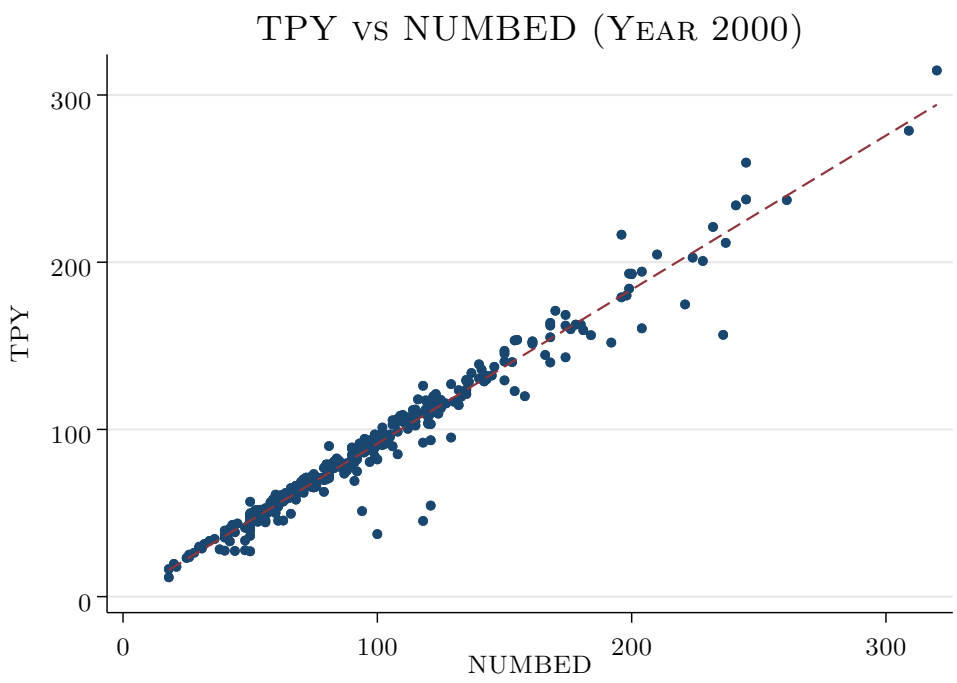

Figure 1: TPY vs. NUMBED (Cost-Report Year 2000). Scatter plot of total patient days (TPY) against number of beds. The relationship is nearly linear and extremely strong, consistent with the high correlation (0.979) and the R^2 of 0.959 from the simple regression.

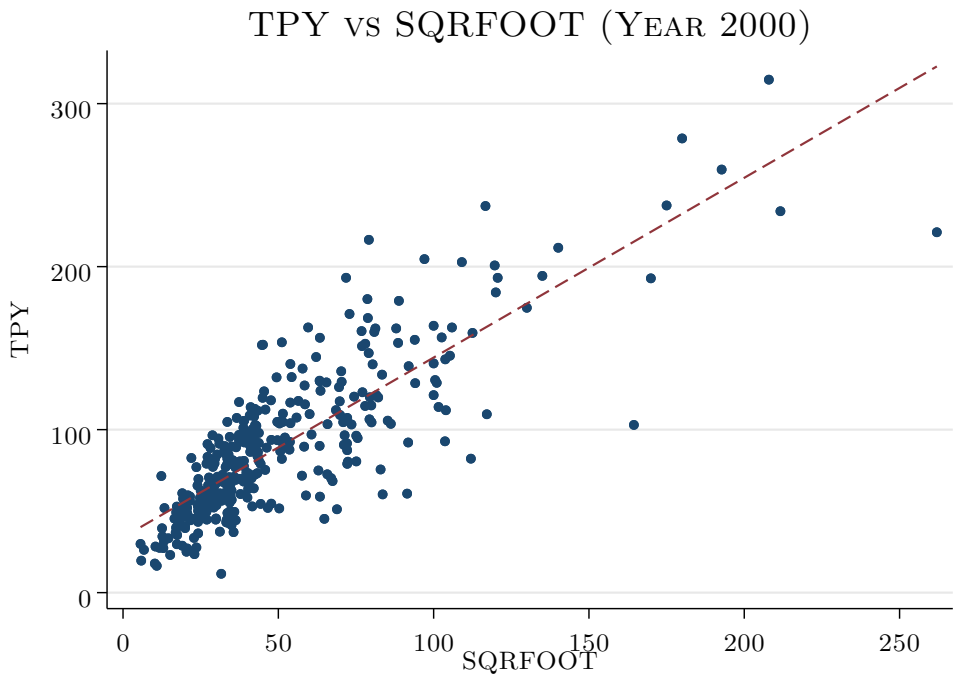

Figure 2: TPY vs. SQRFOOT (Cost-Report Year 2000). Scatter plot of total patient days (TPY) against facility square footage. While the relationship is positive, the dispersion is substantially larger than in Figure 1, consistent with the lower R^2 (0.680).

Table 3: Full Wage Regression: Classical vs Robust Standard Errors

	(1)	(2)
educ	0.570***	0.570***
	(0.050)	(0.061)
exper	0.025**	0.025**
	(0.012)	(0.010)
tenure	0.141***	0.141***
	(0.021)	(0.028)
female	−1.812***	−1.812***
	(0.265)	(0.255)
nonwhite	−0.116	−0.116
	(0.427)	(0.393)
Intercept	−1.540**	−1.540*
	(0.732)	(0.831)
n	526	526
R-squared	0.364	0.364
Regression RMSE	2.960	2.960

Column (2) reports heteroskedasticity-robust standard errors.

Replication Files

All replication materials for `homework_03` are available on GitHub:

- Course repository: https://github.com/tylersotomayor/econun3412-spring2026
- Homework 03 folder: https://github.com/tylersotomayor/econun3412-spring2026/tree/main/homework_03