Handout 7: Multiple Regression

ANSWER KEY

EC 282: Introduction to Econometrics

Spring 2026

1 Setup: The Current Population Survey (CPS) Wage Data

Question 1.1: Examine the summary statistics.

From `summary()`:

- Average hourly wage: $\approx \$5.90$
- Average education: ≈ 12.56 years (roughly a high school diploma plus some college)
- Fraction female: $\approx 47.9\%$ (`mean(female)` $= 0.4791$)

The sample has 526 workers. Note that these are 1976 wages, so they are much lower than today's wages. Adjusting for inflation, \$5.90 in 1976 is roughly \$30 in 2024 dollars.

Question 1.2: Scatterplot.

The scatterplot shows a clear **positive relationship**: workers with more education tend to earn higher wages. However, there is substantial spread—at any given education level, wages vary considerably. The relationship appears roughly linear, though there is more spread at higher education levels (possible heteroskedasticity).

2 Starting Point: Simple Regression

Question 2.1: Simple regression output.

(a) $\widehat{\text{Wage}}_i = -0.9049 + 0.5414 \times \text{Educ}_i$

(b) Each additional year of education is associated with a \$0.54 increase in hourly wages. A worker with 16 years of education (college graduate) earns, on average, about $0.54 \times 4 =$ \$2.16 more per hour than a worker with 12 years (high school graduate).

(c) Yes, $\hat{\beta}_1$ is highly significant. The t-statistic is ≈ 10.17 and the p-value is essentially zero (< 0.0001). It is significant at the 1%, 5%, and 10% levels.

(d) $R^2 \approx 0.165$. Education alone explains about 16.5% of the variation in wages. This means 83.5% of wage variation is due to other factors—there is a lot of unexplained variation.

Question 2.2: Omitted variables.

Two important omitted variables:

1. Experience:

- Correlated with education? **Yes, negatively**. Workers who spend more years in school tend to have fewer years of labor market experience (for a given age, exper $\approx$ age $-$ educ $-$ 6).
- Affects wages? **Yes, positively**. More experienced workers earn more.
- Bias direction: $\text{Corr}(\text{educ}, \text{exper}) < 0$ and $\beta_{\text{exper}} > 0$, so the bias is **negative** (downward). The simple regression *underestimates* the return to education!

2. Gender (female):

- Correlated with education? Slightly—in this sample the correlation is small.
- Affects wages? **Yes**. Women earn less on average (gender wage gap).
- Bias direction: Depends on the exact correlations, but typically small.

Other valid examples: innate ability, family background, quality of schooling, occupation, industry.

3 Multiple Regression and Partial Effects

Question 3.1: Multiple regression with experience and tenure.

(a) $\widehat{\text{Wage}}_i = -2.873 + 0.599 \times \text{Educ}_i + 0.022 \times \text{Exper}_i + 0.169 \times \text{Tenure}_i$

(b) Each additional year of education is associated with a \$0.60 increase in hourly wages, **holding experience and tenure constant**. This is the *partial effect* or *ceteris paribus* effect of education. Unlike the simple regression, we are now comparing workers with the same experience and tenure but different education levels.

(c) The coefficient on tenure (≈ 0.169) means that each additional year with the current employer is associated with about \$0.17 higher wages, *holding education and experience constant*. In practice, this means we are comparing two workers with identical education and total experience, but one has been at their current job one year longer.

(d) The education coefficient **increased** from ≈ 0.541 (Model 1) to ≈ 0.599 (Model 2). This is surprising because we usually expect coefficients to shrink when adding controls. The reason is that the omitted variable bias was *negative* (see Question 3.2).

Question 3.2: OVB verification.

The omitted variable bias formula (for the two-variable case) is:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \times \hat{\alpha}_1$$

(a) $\hat{\alpha}_1 \approx -1.47$ (from regressing experience on education). This is **negative**: more educated workers have *less* experience. This makes sense—a person who spends 4 extra years in college has 4 fewer years of work experience, holding age constant.

(b) $\hat{\beta}_2 > 0$ (experience has a positive effect on wages). More experienced workers earn more.

(c) The bias is $\hat{\beta}_2 \times \hat{\alpha}_1 \approx (+)(-) < 0$. The bias is **negative**, so the simple regression $\tilde{\beta}_1$ is *smaller* than the true partial effect $\hat{\beta}_1$. In words: the simple regression underestimates the return to education because it doesn't account for the fact that more educated workers have less experience. Part of the "education effect" is offset by the "lost experience" penalty.

(d) The formula should hold exactly in the two-variable case:

$$0.5414 \approx 0.6443 + 0.0701 \times (-1.4682) = 0.6443 - 0.1029 = 0.5414 \checkmark$$

Note: These are the coefficients from `wage ~ educ + exper` (two regressors), not from Model 2 which has three regressors. With three regressors, the formula is an approximation.

Question 3.3: Adding demographic controls.

(a) $\hat{\beta}_{\text{female}} \approx -1.81$: Women earn about \$1.81 less per hour than men, *holding education, experience, tenure, and race constant.* This is a large gap—it means a woman with identical human capital characteristics to a man earns roughly \$1.81/hour less, which amounts to about \$3,770 per year (assuming 2,080 working hours).

(b) $\hat{\beta}_{\text{nonwhite}} \approx -0.12$ with a p-value of ≈ 0.79. This is **not statistically significant**—we cannot reject the null that the wage gap between white and nonwhite workers is zero, after controlling for education, experience, tenure, and gender. This does not necessarily mean there is no racial wage gap—it may mean that racial differences in wages are largely explained by differences in education, experience, and tenure.

(c) The education coefficient decreased slightly from ≈ 0.599 (Model 2) to ≈ 0.570 (Model 3). Adding gender as a control absorbed some of what was previously attributed to education.

4 Measures of Fit: SER, RMSE, and Adjusted R^2

Question 4.1: Measures of fit comparison.

Model	R^2	Adj. R^2	SER	RMSE
(1) Educ only	0.1648	0.1632	3.3784	3.3720
(2) + Exper, Tenure	0.3064	0.3024	3.0845	3.0727
(3) + Female, Nonwhite	0.3636	0.3575	2.9602	2.9433

(a) R^2 increases with every added variable: $0.1648 \to 0.3064 \to 0.3636$. This is guaranteed—adding any variable can only increase (or maintain) R^2, because the OLS algorithm can always set the new coefficient to zero.

(b) Adjusted R^2 also increases here ($0.1632 \to 0.3024 \to 0.3575$), but it does **not** have to. Adjusted R^2 penalizes for adding regressors via the factor $\frac{n-1}{n-k-1}$. It only increases if the new variable improves fit enough to justify the penalty. This makes it better for model comparison because it prevents "overfitting" by adding irrelevant variables.

(c) The SER in Model 3 is $\approx \$2.96$. This means our typical prediction error is about \$2.96 per hour—if we predict a worker's wage using their education, experience, tenure, gender, and race, we'd typically be off by about \$3. Given the mean wage is \$5.90, this is still a large error (about 50% of the mean).

Question 4.2: Irrelevant variable.

When we add `numdep` (number of dependents):

- R^2 increases slightly: $0.3064 \to 0.3098$
- Adjusted R^2 barely changes: $0.3024 \to 0.3045$

R^2 went up because it *always* goes up when you add a variable. But Adjusted R^2 hardly changed, correctly signaling that `numdep` adds very little explanatory power. If we added a truly irrelevant variable (like a column of random numbers), R^2 would still increase but Adjusted R^2 could actually *decrease*. This demonstrates why Adjusted R^2 is preferred for model comparison: it doesn't reward you for adding noise.

5 Dummy Variables and the Dummy Variable Trap

Question 5.1: The dummy variable trap.

(a) R reports `NA` for the coefficient on `male`, with the message "1 not defined because of singularities." R automatically detects the perfect multicollinearity and drops one of the redundant variables.

(b) Since $\text{Male}_i = 1 - \text{Female}_i$ for every observation, the model:

$$Y_i = \beta_0 + \beta_1 \text{Educ}_i + \beta_2 \text{Exper}_i + \beta_3 \text{Female}_i + \beta_4 \text{Male}_i + u_i$$

can be rewritten as:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \text{Educ}_i + \beta_2 \text{Exper}_i + \beta_3 \text{Female}_i + \beta_4 (1 - \text{Female}_i) + u_i \\ &= (\beta_0 + \beta_4) + \beta_1 \text{Educ}_i + \beta_2 \text{Exper}_i + (\beta_3 - \beta_4) \text{Female}_i + u_i \end{aligned}$$

We cannot separately identify β_0, β_3, and β_4—only the combinations $(\beta_0 + \beta_4)$ and $(\beta_3 - \beta_4)$. The system of equations is under-determined.

(c) **Rule:** For a categorical variable with k categories, include $k - 1$ dummy variables. The omitted category becomes the **reference group** (baseline), and each included dummy measures the difference relative to that baseline.

Question 5.2: Region dummies.

(a) The **reference category is Northeast**. We know this because the three included dummies are `northcen`, `south`, and `west`—Northeast is the one that's missing. The intercept represents the expected wage for a Northeast worker (with all other variables at zero).

(b) $\hat{\beta}_{\text{south}} \approx -0.63$: Workers in the South earn about \$0.63 less per hour than workers in the Northeast, *holding all other variables constant* (education, experience, tenure, gender, race, marital status). This is a comparison between otherwise identical workers who differ only in their region.

(c) At the 5% level, **none** of the region dummies are individually significant:

- `northcen`: $p \approx 0.087$
- `south`: $p \approx 0.073$
- `west`: $p \approx 0.181$

However, as we'll see in Question 7.2, they are *jointly* significant according to the F-test!

6 Multicollinearity

Question 6.1: Perfect multicollinearity.

R reports `NA` for the coefficient on `age` with the message "1 not defined because of singularities." This is **perfect multicollinearity**: since Age = Educ + 6 + Exper exactly, age is a perfect linear function of the other two regressors. OLS cannot separate the individual effects of education, experience, and age because any change in education *automatically* changes age (holding experience constant).

Algebraically: $\text{Age}_i = 6 + \text{Educ}_i + \text{Exper}_i$, so including all three plus a constant creates a linear dependency among the columns of the X matrix. The matrix $X'X$ is singular and cannot be inverted.

Question 6.2: Imperfect multicollinearity.

(a) The standard errors are dramatically inflated:

	Model 2 (no MC)	Noisy age model
SE(educ)	0.0513	0.0925
SE(exper)	0.0121	0.0731

The SE for education nearly doubled, and the SE for experience increased by a factor of 6! This is the hallmark of multicollinearity: the estimates become very **imprecise**.

(b) Education is still significant ($t \approx 8.37$), but the coefficient on experience, while still significant ($p \approx 0.008$), has a much larger standard error. With even slightly more noise or collinearity, experience could easily become insignificant.

(c) The coefficients are distorted: education jumps to ≈ 0.77 (vs. ≈ 0.60 in Model 2) and experience jumps to ≈ 0.19 (vs. ≈ 0.02), while `age_noisy` gets a negative coefficient (≈ -0.12). These magnitudes are unreliable—multicollinearity means OLS can't reliably decompose the effects among highly correlated variables.

Question 6.3: VIF computation.

Variable	Model 2 VIF	Noisy age VIF
educ	1.11	3.26
exper	1.48	**48.83**
tenure	1.35	—
age_noisy	—	**44.81**

In Model 2, all VIFs are below 2—no multicollinearity concern. In the noisy age model, `exper` and `age_noisy` have VIFs of ≈ 49 and ≈ 45, far exceeding the threshold of 5 (or the more conservative threshold of 10). This confirms severe multicollinearity.

We should **not** keep `age_noisy` in the model. Since age $\approx$ educ+6+exper, age is essentially

redundant information. Including it adds no new information but massively inflates standard errors. The solution: drop the redundant variable.

7 Hypothesis Testing and F-Tests

Question 7.1: Testing education significance.

(a) $H_0 : \beta_{\text{educ}} = 0$ (education has no effect on wages, holding all else constant)
$H_A : \beta_{\text{educ}} \neq 0$ (education has some effect)

(b) $t = \hat{\beta}_{\text{educ}}/SE(\hat{\beta}_{\text{educ}}) = 0.5479/0.0499 \approx 10.98$

$|t| = 10.98 > 1.96$ (5% critical value) $\Rightarrow$ Reject H_0 at 5% level.

$|t| = 10.98 > 2.576$ (1% critical value) $\Rightarrow$ Reject H_0 at 1% level.

Education is overwhelmingly statistically significant.

(c) 95% CI: $0.5479 \pm 1.96 \times 0.0499 = [0.450,\ 0.646]$

We are 95% confident that the true partial effect of an additional year of education on hourly wages is between \$0.45 and \$0.65. The interval does not include zero, consistent with rejecting H_0 at 5%.

Question 7.2: F-test for joint significance of region dummies.

(a) The F-statistic using the R^2 formula is:

$$F = \frac{(R^2_{\text{unr}} - R^2_{\text{res}})/q}{(1 - R^2_{\text{unr}})/(n - k - 1)} = \frac{(0.3835 - 0.3682)/3}{(1 - 0.3835)/(526 - 9 - 1)} = \frac{0.0153/3}{0.6165/516} = \frac{0.00510}{0.001195} \approx 4.27$$

(b) Critical value for $q = 3$ at 5%: $F^*_{3,516} \approx 2.62$. Since $4.27 > 2.62$, we **reject** H_0. The p-value is ≈ 0.006, which is less than 0.05. The region dummies are jointly statistically significant at the 5% level.

(c) This is a key insight about joint testing:

- Individual t-tests check each coefficient *one at a time*
- The F-test checks all three *simultaneously*
- Individual t-tests can miss joint significance for two reasons:
  (a) **Multiple testing problem**: With 3 tests at 5%, the probability of incorrectly rejecting at least once is $\approx 14\%$, not 5%. The F-test maintains the correct 5% significance level.
  (b) **Correlated estimates**: The region dummies are correlated with each other, so their individual t-statistics don't capture the combined information. The F-test accounts for these correlations.
- In this case, no single region is significantly different from Northeast, but taken together, regional location does matter for wages.

8 Putting It All Together

Question 8.1: Model comparison table.

	Model 1 Educ only	Model 2 + Exp, Ten	Model 3 + Fem, NW	Model 4 Full
$\hat{\beta}_{\text{educ}}$	0.5414	0.5990	0.5703	0.5479
$SE(\hat{\beta}_{\text{educ}})$	0.0533	0.0513	0.0496	0.0499
R^2	0.1648	0.3064	0.3636	0.3835
Adj. R^2	0.1632	0.3024	0.3575	0.3727
SER	3.378	3.085	2.960	2.925

Question 8.2: Discussion.

(a) The education coefficient follows an interesting pattern:

- Model 1 → Model 2: **Increases** (0.54 → 0.60). Adding experience revealed that the simple regression *underestimated* the return to education (negative OVB from the negative education-experience correlation).
- Model 2 → Model 3: **Decreases** (0.60 → 0.57). Adding gender absorbed some of the education effect.
- Model 3 → Model 4: **Decreases** slightly (0.57 → 0.55). Adding marital status and region further adjusts the estimate.

The overall message: the simple regression gives a misleading estimate. Controlling for relevant variables changes the coefficient substantially and in both directions.

(b) Model 3 or Model 4 are reasonable "preferred" specifications. Model 4 has the highest Adjusted R^2 (0.3727 vs. 0.3575), and the region dummies are jointly significant. However, Model 3 is more parsimonious and the gain from adding regions is modest. A good argument can be made for either.

(c) **No, we cannot claim causation** even with all these controls. Remaining concerns include:

- **Ability bias**: More able people get more education *and* earn more. If we can't measure ability, $\hat{\beta}_{\text{educ}}$ captures both the education effect and the ability effect.
- **Family background**: Wealthier families invest more in education and provide better labor market connections.
- **School quality**: A year at a top university is not the same as a year at any institution.
- **Selection**: People choose education levels based on unobserved characteristics (motivation, career goals).

To establish causation, we would need a research design such as an instrumental variable (e.g., distance to college), a natural experiment, or a randomized experiment.

Question 8.3: Gender wage gap visualization.

The plot shows two parallel (or nearly parallel) regression lines—one for men and one for women. The male line is **higher** at every education level, visually confirming the gender wage gap. The gap appears roughly constant across education levels (approximately \$1.80 per hour, matching our regression estimate of $\hat{\beta}_{\text{female}} \approx -1.81$), suggesting that the gender wage gap does not widen or narrow significantly with more education.

This pattern is consistent with a model where gender shifts the intercept but not the slope—in other words, women's wages are lower by a roughly fixed amount regardless of their education level. To test whether the gap actually varies with education, one would need to add an interaction term (`educ` $\times$ `female`), which is a topic for future study.