Handout 7: Multiple Regression

EC 282: Introduction to Econometrics

Spring 2026

Instructions: Run the provided R code and answer the questions. Show your work for calculations. This handout introduces a **new dataset**—we move from cross-country comparisons to individual-level U.S. wage data.

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

How much does an extra year of education increase your hourly wage? The answer seems simple—until you realize that education is correlated with experience, gender, and many other factors. We use data from the 1976 Current Population Survey (CPS) to explore these issues.

```r
# Install the wooldridge package (run once)
# install.packages("wooldridge")

library(wooldridge)
library(ggplot2)

options(scipen = 999)

# Load the wage data
data(wage1)

cat("Number of workers:", nrow(wage1), "\n")
cat("Variables:", paste(names(wage1)[1:12], collapse = ", "), "\n")

# Key variables:
# wage     = average hourly earnings ($)
# educ     = years of education
# exper    = years of potential experience
# tenure   = years with current employer
# female   = 1 if female, 0 if male
# nonwhite = 1 if nonwhite, 0 if white
# married  = 1 if married
# northcen, south, west = region dummies (northeast is omitted)

summary(wage1[, c("wage", "educ", "exper", "tenure",
                  "female", "nonwhite", "married")])
```

Question 1.1: Examine the summary statistics. What is the average hourly wage? What is the average education level? What fraction of the sample is female?

Question 1.2: Create a scatterplot of wages against education. What pattern do you see?

```
ggplot(wage1, aes(x = educ, y = wage)) +
  geom_jitter(color = "steelblue", alpha = 0.5, width = 0.3) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Hourly Wage vs. Years of Education",
       x = "Years of Education",
       y = "Hourly Wage ($)") +
  theme_minimal()
```

2 Starting Point: Simple Regression

We begin with a simple regression of wages on education—the "naive" estimate of the return to schooling.

Question 2.1: Run the simple regression and examine the output.

```
reg1 <- lm(wage ~ educ, data = wage1)
summary(reg1)
```

(a) Write down the estimated equation: $\widehat{\text{Wage}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Educ}_i$

(b) Interpret $\hat{\beta}_1$ in a sentence. What does it mean economically?

(c) Is $\hat{\beta}_1$ statistically significant? At what level?

(d) What is the R^2? What fraction of wage variation is explained by education alone?

Question 2.2: Think about what is in the error term u_i. Name at least two variables that are:

1. Correlated with education, AND
2. Likely affect wages

For each, predict the *direction* of the omitted variable bias. Does it make $\hat{\beta}_1$ too large or too small?

3 Multiple Regression and Partial Effects

Now we add control variables to address omitted variable bias. The key idea: $\hat{\beta}_1$ in a multiple regression measures the effect of education on wages **holding other variables constant**.

Question 3.1: Add experience and tenure to the regression.

```
reg2 <- lm(wage ~ educ + exper + tenure, data = wage1)
summary(reg2)
```

(a) Write down the estimated equation.

(b) Interpret the coefficient on `educ`. How does the interpretation differ from the simple regression?

(c) Interpret the coefficient on `tenure`. What does "holding education and experience constant" mean in practice?

(d) Compare the education coefficient to Model 1. Did it go up or down? Were you expecting this direction?

Question 3.2: The coefficient on education *increased* when we added controls (from ≈ 0.54 to ≈ 0.60). This is surprising—usually we expect omitted variable bias to make the simple regression coefficient too large. Verify the omitted variable bias formula for the two-variable case:

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

where $\hat{\alpha}_1$ is the slope from regressing `exper` on `educ`.

```r
# Step 1: Run the auxiliary regression
aux_reg <- lm(exper ~ educ, data = wage1)
alpha1_hat <- coef(aux_reg)["educ"]
cat("alpha1_hat (exper ~ educ):", round(alpha1_hat, 4), "\n")

# Step 2: Check the correlation
cat("Correlation(educ, exper):", round(cor(wage1$educ,
    wage1$exper), 4), "\n")

# Step 3: Verify the OVB formula (for the 2-variable case)
# Simple regression: wage ~ educ
beta1_tilde <- coef(reg1)["educ"]

# Two-variable regression: wage ~ educ + exper
reg_two <- lm(wage ~ educ + exper, data = wage1)
beta1_hat <- coef(reg_two)["educ"]
beta2_hat <- coef(reg_two)["exper"]

cat("\nbeta1_tilde (simple):", round(beta1_tilde, 4), "\n")
cat("beta1_hat (multiple):", round(beta1_hat, 4), "\n")
cat("beta2_hat * alpha1_hat:", round(beta2_hat * alpha1_hat, 4),
    "\n")
cat("beta1_hat + beta2_hat * alpha1_hat:",
    round(beta1_hat + beta2_hat * alpha1_hat, 4), "\n")
```

(a) What is the sign of $\hat{\alpha}_1$? What does this tell you about the relationship between education and experience?

(b) What is the sign of $\hat{\beta}_2$ (the effect of experience on wages)?

(c) Using the OVB formula, explain *why* the simple regression *underestimates* the return to education.

(d) Does the formula hold exactly? Verify numerically.

Question 3.3: Now add demographic controls.

```r
reg3 <- lm(wage ~ educ + exper + tenure + female + nonwhite,
           data = wage1)
summary(reg3)
```

(a) Interpret the coefficient on `female`. What does it mean in economic terms?

(b) Is the coefficient on `nonwhite` statistically significant? What does this tell us?

(c) How did the education coefficient change compared to Model 2? Why might this be?

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

As we add variables, we need better tools to evaluate model fit than R^2 alone.

Question 4.1: Compute measures of fit for each model.

```
models <- list(reg1, reg2, reg3)
model_names <- c("(1) Educ only", "(2) + Exper, Tenure",
                 "(3) + Female, Nonwhite")

cat("Model Comparison:\n")
cat(sprintf("%-25s %8s %8s %8s %8s\n",
    "Model", "R2", "Adj R2", "SER", "RMSE"))
cat(paste(rep("-", 60), collapse = ""), "\n")

for (i in 1:3) {
  s <- summary(models[[i]])
  n <- nrow(models[[i]]$model)
  k <- length(coef(models[[i]])) - 1
  RSS <- sum(residuals(models[[i]])^2)
  SER <- sqrt(RSS / (n - k - 1))
  RMSE <- sqrt(RSS / n)
  cat(sprintf("%-25s %8.4f %8.4f %8.4f %8.4f\n",
      model_names[i], s$r.squared, s$adj.r.squared, SER, RMSE))
}
```

(a) As we add variables, R^2 always increases. Verify this from the table.

(b) Does Adjusted R^2 also always increase? Why is Adjusted R^2 a better measure than R^2 for comparing models with different numbers of regressors?

(c) The SER tells you the typical size of prediction errors, in the same units as Y (dollars per hour). How large is it in Model 3?

Question 4.2: Now add an irrelevant variable (`numdep` = number of dependents) that shouldn't affect wages directly.

```
reg_irrel <- lm(wage ~ educ + exper + tenure + numdep,
                data = wage1)

cat("Model 2:     R2 =", round(summary(reg2)$r.squared, 4),
    "  Adj R2 =", round(summary(reg2)$adj.r.squared, 4), "\n")
cat("+ numdep:    R2 =",
    round(summary(reg_irrel)$r.squared, 4),
    "  Adj R2 =",
    round(summary(reg_irrel)$adj.r.squared, 4), "\n")
```

What happens to R^2 vs. Adjusted R^2 when you add a variable that doesn't belong? Why does this demonstrate the advantage of Adjusted R^2?

5 Dummy Variables and the Dummy Variable Trap

Question 5.1: Create a `male` dummy and try to include both `female` and `male` in the regression.

```
# Create male dummy
wage1$male <- 1 - wage1$female

# Try to include both
reg_trap <- lm(wage ~ educ + exper + female + male, data = wage1)
summary(reg_trap)
```

(a) What does R report for the coefficient on `male`? Why?

(b) Write down the mathematical reason: if $\text{Female}_i + \text{Male}_i = 1$ for all i, show algebraically why this creates a problem.

(c) What is the general rule for including dummy variables for a categorical variable with k categories?

Question 5.2: The dataset has region indicators: `northcen`, `south`, and `west` (the omitted category is Northeast). Run the full model with region dummies.

```
reg4 <- lm(wage ~ educ + exper + tenure + female + nonwhite +
           married + northcen + south + west, data = wage1)
summary(reg4)
```

(a) Which region is the **reference category**? How do you know?

(b) Interpret the coefficient on `south`. What comparison is being made?

(c) Are any of the region dummies individually statistically significant at the 5% level?

6 Multicollinearity

What happens when regressors are highly correlated with each other? Let's explore with two experiments.

Question 6.1: Perfect multicollinearity. In this dataset, a worker's age is approximately: Age = Educ + 6 + Exper (assuming school starts at age 6). Create this variable and try to include it in the regression.

```
# Construct age (exact linear function of educ and exper)
wage1$age <- wage1$educ + 6 + wage1$exper

# Try to include all three
reg_perfect_mc <- lm(wage ~ educ + exper + age, data = wage1)
summary(reg_perfect_mc)
```

What happens? Why can't R estimate all three coefficients?

Question 6.2: Imperfect multicollinearity. Now add random noise to age (simulating measurement error in age). This breaks the exact linear relationship but keeps the variables highly correlated.

```
set.seed(42)  # For reproducibility
wage1$age_noisy <- wage1$age + rnorm(nrow(wage1), 0, 2)

cat("Correlation(age_noisy, age):",
    round(cor(wage1$age_noisy, wage1$age), 4), "\n\n")

# Run the regression with the noisy age variable
reg_noisy <- lm(wage ~ educ + exper + age_noisy, data = wage1)
summary(reg_noisy)
```

(a) R can now estimate all coefficients. But look at the standard errors on `educ` and `exper`—how do they compare to Model 2 (which had no multicollinearity)?

(b) Are the coefficients on `educ` and `exper` still statistically significant?

(c) What happened to the magnitudes of the coefficients? Are they reasonable?

Question 6.3: Compute the Variance Inflation Factor (VIF) to formally diagnose multicollinearity. Recall: $VIF(\hat{\beta}_j) = \frac{1}{1-R_j^2}$, where R_j^2 is the R^2 from regressing X_j on all other regressors.

```r
# Manual VIF function
compute_vif <- function(model) {
  vars <- attr(terms(model), "term.labels")
  dat  <- model$model
  vifs <- numeric(length(vars))
  names(vifs) <- vars
  for (i in seq_along(vars)) {
    f <- as.formula(paste(vars[i], "~",
         paste(vars[-i], collapse = " + ")))
    r2 <- summary(lm(f, data = dat))$r.squared
    vifs[i] <- 1 / (1 - r2)
  }
  return(round(vifs, 2))
}

cat("VIF for Model 2 (no multicollinearity):\n")
print(compute_vif(reg2))

cat("\nVIF for noisy age model (severe multicollinearity):\n")
print(compute_vif(reg_noisy))
```

Which variables have VIF > 5? What does this tell you? Would you keep `age_noisy` in the model?

7 Hypothesis Testing and F-Tests

Question 7.1: Using Model 4 (the full model), test whether the coefficient on education is statistically significant.

```
# Extract info for education coefficient
beta_educ <- coef(reg4)["educ"]
se_educ   <- summary(reg4)$coefficients["educ", "Std. Error"]
t_stat    <- beta_educ / se_educ
p_value   <- 2 * pnorm(-abs(t_stat))

cat("beta_hat (educ):", round(beta_educ, 4), "\n")
cat("SE:", round(se_educ, 4), "\n")
cat("t-statistic:", round(t_stat, 3), "\n")
cat("p-value:", p_value, "\n")
cat("95% CI: [", round(beta_educ - 1.96 * se_educ, 4), ",",
    round(beta_educ + 1.96 * se_educ, 4), "]\n")
```

(a) State H_0 and H_A.

(b) Is education significant at the 5% level? At the 1% level?

(c) Interpret the 95% confidence interval. Does it include zero?

Question 7.2: The F-Test. None of the three region dummies is individually significant at the 5% level. But are they *jointly* significant? That is, do we gain anything by including all three region dummies together?

$$H_0 : \beta_{\text{northcen}} = \beta_{\text{south}} = \beta_{\text{west}} = 0 \qquad \text{(regions don't matter)}$$

$$H_A : \text{At least one } \beta_j \neq 0 \qquad \text{(at least one region matters)}$$

```
# Restricted model (without region dummies)
reg_restricted <- lm(wage ~ educ + exper + tenure + female +
                     nonwhite + married, data = wage1)

# Unrestricted model (with region dummies)
reg_unrestricted <- reg4  # already estimated above

# R-squared values
R2_unr <- summary(reg_unrestricted)$r.squared
R2_res <- summary(reg_restricted)$r.squared

cat("R2 (unrestricted):", round(R2_unr, 4), "\n")
cat("R2 (restricted):", round(R2_res, 4), "\n")

```

```r
# F-test by hand
q <- 3        # number of restrictions
k <- 9        # regressors in unrestricted model
n <- nrow(wage1)

F_stat <- ((R2_unr - R2_res) / q) / ((1 - R2_unr) / (n - k - 1))
p_value_F <- 1 - pf(F_stat, q, n - k - 1)

cat("\nF-statistic (by hand):", round(F_stat, 4), "\n")
cat("Critical value (5%, q=3):", round(qf(0.95, q, n-k-1), 4),
    "\n")
cat("p-value:", round(p_value_F, 4), "\n")

# Verify with R's built-in anova()
cat("\nanova() output:\n")
print(anova(reg_restricted, reg_unrestricted))
```

(a) Compute the F-statistic by hand using the R^2 formula. Show your work.

(b) Compare it to the critical value. Do you reject H_0 at the 5% level?

(c) Explain the paradox: no individual region dummy is significant at 5%, yet the F-test rejects at 5%. Why can individual t-tests miss joint significance?

8 Putting It All Together

Question 8.1: Fill in the comparison table for our four main models.

	Model 1 Educ only	Model 2 + Exper, Tenure	Model 3 + Female, Nonwhite	Model 4 Full
$\hat{\beta}_{\text{educ}}$				
$SE(\hat{\beta}_{\text{educ}})$				
R^2				
Adjusted R^2				
SER				

```r
cat("Side-by-side comparison:\n\n")
models_all <- list(reg1, reg2, reg3, reg4)
names_all <- c("(1) Educ", "(2)+Exp,Ten",
               "(3)+Fem,NW", "(4) Full")

for (i in 1:4) {
  s <- summary(models_all[[i]])
  n <- nrow(models_all[[i]]$model)
  k <- length(coef(models_all[[i]])) - 1
  RSS <- sum(residuals(models_all[[i]])^2)
  cat(names_all[i], "\n")
  cat("  educ coef:", round(coef(models_all[[i]])["educ"], 4),"\n")
  cat("  SE(educ):",
      round(s$coefficients["educ", "Std. Error"], 4), "\n")
  cat("  R2:", round(s$r.squared, 4), "\n")
  cat("  Adj R2:", round(s$adj.r.squared, 4), "\n")
  cat("  SER:", round(sqrt(RSS / (n - k - 1)), 4), "\n\n")
}
```

Question 8.2: Based on the table, discuss:

(a) How does the education coefficient change as we add controls? What does this tell you about omitted variable bias in the simple regression?

(b) Which model would you choose as your "preferred" specification? Why?

(c) Even with all these controls, can we claim that education *causes* higher wages? What concerns remain?

Question 8.3: Visualize the gender wage gap, controlling for education.

```r
ggplot(wage1, aes(x = educ, y = wage, color = factor(female,
       labels = c("Male", "Female")))) +
  geom_jitter(alpha = 0.4, width = 0.3) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 1.2) +
  labs(title = "Wage vs. Education by Gender",
       x = "Years of Education", y = "Hourly Wage ($)",
       color = "Gender") +
  theme_minimal()
```

What does this plot reveal about the gender wage gap? Is it constant across education levels, or does it change?