multivariate_regression

scott cunningham

Introduction

In this week's section, we will open up the black box of multiple regression. Ordinary Least Squares (OLS) can handle many covariates at once, but what exactly happens when it does? By working through simple examples with real data from the 1980 National Longitudinal Survey (a random sample of U.S. workers), we'll see how OLS uses covariance algebra to isolate the independent contribution of each variable—what it means to "hold something constant."

Our goal is to understand not just how to *run* a multivariate regression, but how the math behind it enforces this idea of control: separating overlapping relationships so we can interpret coefficients as partial effects.

```r
# Load the only package needed to read Stata data
library(haven)

# Import the dataset
nls80 <- read_dta("https://github.com/scunning1975/mixtape/raw/master/nls80.dta")

# Take a quick look
summary(nls80[c("lwage", "educ", "kww")])
```

```
     lwage             educ             kww       
 Min.   :4.745   Min.   : 9.00   Min.   :12.00  
 1st Qu.:6.506   1st Qu.:12.00   1st Qu.:31.00  
 Median :6.808   Median :12.00   Median :37.00  
 Mean   :6.779   Mean   :13.47   Mean   :35.74  
 3rd Qu.:7.056   3rd Qu.:16.00   3rd Qu.:41.00  
 Max.   :8.032   Max.   :18.00   Max.   :56.00  
```

Question 1 (6 points)

We'll use data from the **1980 National Longitudinal Survey (NLS80)**, a nationally representative sample of U.S. workers collected in 1980. It includes information on wages, education, and a test score measuring general knowledge and reasoning ability (the "Knowledge of the World of Work" or **KWW** score). Our goal is to study how schooling affects wages and to see how our estimates change once we control for ability.

```r
# Model 1: log wages on education (no controls)
model1 <- lm(lwage ~ educ, data = nls80)
summary(model1)
```

```

Call:
lm(formula = lwage ~ educ, data = nls80)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.94620 -0.24832  0.03507  0.27440  1.28106 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.973063   0.081374   73.40   <2e-16 ***
educ        0.059839   0.005963   10.04   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4003 on 933 degrees of freedom
Multiple R-squared:  0.09742,	Adjusted R-squared:  0.09645 
F-statistic: 100.7 on 1 and 933 DF,  p-value: < 2.2e-16
```

Prompt.

1. Regress `lwage` onto `educ` using OLS. How much is each additional year associated with changes in earnings?

The outcome is the natural log of wages, and therefore each additional year of schooling is associated with approximately 6% higher wages.

2. If the worker's unobserved ability is positively correlated with both wages and schooling, then do you think the estimated coefficient is overstating or understating the effect of schooling on earnings?

I would expect that the effect of schooling we just estimated is too large.

Question 2 (6 points)

The three variables we'll focus on are **lwage** — the natural logarithm of the respondent's hourly wage. Using the log helps interpret changes in terms of percentage differences; **educ** — years of completed schooling. This is our main explanatory variable, representing education; **kww** — the *Knowledge of the World of Work* test score, a measure often used as a proxy for cognitive ability or general knowledge. Together, these let us test how education relates to wages and whether part of that relationship reflects underlying differences in ability rather than schooling itself.

```r
# Correlation between education and ability
cor(nls80$educ, nls80$kww, use = "complete.obs")
```

```
[1] 0.3881342
```

```r
# Model 2: add KWW (proxy for ability)
model2 <- lm(lwage ~ educ + kww, data = nls80)
summary(model2)
```

```
Call:
lm(formula = lwage ~ educ + kww, data = nls80)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.09326 -0.22741  0.03595  0.26587  1.42661 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.761978   0.085755  67.191  < 2e-16 ***
educ        0.043620   0.006327   6.894  1.0e-11 ***
kww         0.012017   0.001820   6.604  6.7e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3915 on 932 degrees of freedom
Multiple R-squared:  0.1378,    Adjusted R-squared:  0.1359 
F-statistic: 74.46 on 2 and 932 DF,  p-value: < 2.2e-16
```

Prompt.

1. Check whether schooling (`educ`) is positively or negatively correlated with knowledge of the work of the world (`kww`).

They are positively correlated. This means that individuals with more knowledge of the world of the world also have more schooling.

2. Does this imply that the original OLS coefficient on `educ` from question 1 is probably too high or too small?

This implies that the original OLS coefficient is probably too large.r.

3. Estimate the effect of `educ` on `lwage` controlling for `kww` using `lm()`. How does your estimate once you control for `KWW` compare to what you found in the previous question?

When we estimate the effect of education on wages controlling for `kww`, the effect falls from 6% to 4.4% higher wages for every additional year of schooling. This is consistent with what we predicted in the previous question which is that part of the returns to schooling was reflecting unobserved ability.

Question 3 (6 points)

Now we want to better understand the way in which OLS calculated those coefficients, so we will do it manually ourselves. Recall from our slides that the formulas are:

$$\beta_1 = \frac{\text{Cov}(X_1,Y)\text{Var}(X_2) - \text{Cov}(X_2,Y)\text{Cov}(X_1,X_2)}{\text{Var}(X_1)\text{Var}(X_2) - \text{Cov}(X_1,X_2)^2}$$

$$\beta_2 = \frac{\text{Cov}(X_2,Y)\text{Var}(X_1) - \text{Cov}(X_1,Y)\text{Cov}(X_1,X_2)}{\text{Var}(X_1)\text{Var}(X_2) - \text{Cov}(X_1,X_2)^2}$$

and then the intercept:

$$\beta_0 = \bar{Y} - \beta_1\bar{X}_1 - \beta_2\bar{X}_2$$

```
# Define variables
Y  <- nls80$lwage
X1 <- nls80$educ
X2 <- nls80$kww

# Remove missing values
complete <- complete.cases(Y, X1, X2)
Y  <- Y[complete]; X1 <- X1[complete]; X2 <- X2[complete]

# Compute covariances and variances
Cov_X1Y  <- mean(X1*Y)  - mean(X1)*mean(Y)
Cov_X2Y  <- mean(X2*Y)  - mean(X2)*mean(Y)
Var_X1   <- mean(X1^2)  - mean(X1)^2
```

```r
Var_X2   <- mean(X2^2)  - mean(X2)^2
Cov_X1X2 <- mean(X1*X2) - mean(X1)*mean(X2)

# Compute coefficients manually
b1 <- (Cov_X1Y*Var_X2 - Cov_X2Y*Cov_X1X2) / (Var_X1*Var_X2 - Cov_X1X2^2)
b2 <- (Cov_X2Y*Var_X1 - Cov_X1Y*Cov_X1X2) / (Var_X1*Var_X2 - Cov_X1X2^2)
b0 <- mean(Y) - b1*mean(X1) - b2*mean(X2)

# Display manual coefficients
manual_results <- c(b0 = b0, b1 = b1, b2 = b2)
manual_results
```

```
        b0         b1         b2 
5.76197800 0.04361977 0.01201686 
```

```r
# Compare to lm() results
summary(lm(lwage ~ educ + kww, data = nls80))
```

```

Call:
lm(formula = lwage ~ educ + kww, data = nls80)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.09326 -0.22741  0.03595  0.26587  1.42661 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.761978   0.085755  67.191  < 2e-16 ***
educ        0.043620   0.006327   6.894  1.0e-11 ***
kww         0.012017   0.001820   6.604  6.7e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3915 on 932 degrees of freedom
Multiple R-squared:  0.1378,	Adjusted R-squared:  0.1359 
F-statistic: 74.46 on 2 and 932 DF,  p-value: < 2.2e-16
```

Prompt

4. Using these formulae, calculate the effect of schooling on `lwage` controlling for `kww`, the coefficient on `kww` itself, as well as the intercept term. Then compare your results with the coefficients you get by running the regression of `lwage` onto `kww` and `educ` using `lm()`.

We are able to exactly reproduce the multivariate regression coefficients from `lm()` using these covariance and variance formulas.

Question 4 (16 points)

The **FWL theorem** shows that multivariate regression is, at its core, a sequence of simple regressions. It tells us that when we estimate a model like

$$\text{lwage} = \beta_0 + \beta_1 \, \text{educ} + \beta_2 \, \text{kww} + \varepsilon$$

that the coefficient on `educ`, $\widehat{\beta_1}$ is the same as if we first removed from `educ` the part that can be explained by `kww`, and then regressed `lwage` on this "cleaned" version of education.

This process—called *partialing out*—is what it really means to **control for** another variable. FWL shows that OLS doesn't perform magic; it simply measures how much of `lwage` changes with the variation in `educ` that is *independent of* `kww`. We can demonstrate this in three steps:

1. Regress `lwage` on those residuals.
2. Compare the slope from this regression to the coefficient on `educ` in the full model—it will be identical.

```r
# Step 1: regress educ on kww
reg1 <- lm(educ ~ kww, data = nls80)
summary(reg1)
```

```

Call:
lm(formula = educ ~ kww, data = nls80)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8318 -1.6086 -0.3899  1.7217  5.8424 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 9.478871   0.317128   29.89   <2e-16 ***
kww         0.111614   0.008676   12.86   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.026 on 933 degrees of freedom
Multiple R-squared:  0.1506,	Adjusted R-squared:  0.1497 
F-statistic: 165.5 on 1 and 933 DF,  p-value: < 2.2e-16
```

```r
# Step 2: get residuals (the variation in educ not explained by kww)
nls80$resid_educ <- resid(reg1)

# Step 3: regress lwage on these residuals
reg2 <- lm(lwage ~ resid_educ, data = nls80)

summary(reg2)
```

```
Call:
lm(formula = lwage ~ resid_educ, data = nls80)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.92009 -0.26490  0.03783  0.27714  1.29925 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.779004   0.013474  503.12  < 2e-16 ***
resid_educ  0.043620   0.006659    6.55 9.47e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.412 on 933 degrees of freedom
Multiple R-squared:  0.04397,	Adjusted R-squared:  0.04294 
F-statistic: 42.91 on 1 and 933 DF,  p-value: 9.475e-11
```

```r
# Compare to the coefficient from full model
summary(lm(lwage ~ educ + kww, data = nls80))
```

```
Call:
lm(formula = lwage ~ educ + kww, data = nls80)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.09326 -0.22741  0.03595  0.26587  1.42661 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.761978   0.085755  67.191  < 2e-16 ***
educ        0.043620   0.006327   6.894  1.0e-11 ***
kww         0.012017   0.001820   6.604  6.7e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3915 on 932 degrees of freedom
Multiple R-squared:  0.1378,	Adjusted R-squared:  0.1359 
F-statistic: 74.46 on 2 and 932 DF,  p-value: < 2.2e-16
```

Prompt

1. Regress `educ` on `kww` and save the residuals. These residuals represent the portion of schooling unrelated to ability.

This auxilary regression shows that a 1-unit increase in `kww` is associated with 0.112 more years of schooling. We then calculated the residuals from this regression.

2. Regress `lwage` on those residuals

We then regressed `lwage` onto the residuals, which we named `resid_educ`. The coefficient on `resid_educ` if 0.043620.

3. Compare the slope from this regression to the coefficient on `educ` in the full model `lwage ~ educ + kww` using the `lm()` command. What do you find when you compare the two coefficients?

It is identical.

4. In your own words, how would you explain how OLS "controls for" a covariate in a multivariate regression?

It removes from the covariate of interest (e.g., `educ`) that part of the variation that can be attributed to `kww`. Once that is removed, it regresses the outcome of interest (e.g., `lwage`) onto that remaining variation in `educ` which here is called `educ_residuals`. Note that this means we have removed from the original `educ` that part of the variable that is associated with `kww`. This is why we say that multivariate regression "partials out" the part of the variation that is caused by `kww`.

Question 5 (16 points)

To see that the Frisch-Waugh-Lovell theorem applies to any regression model, we'll now use a different dataset: Cattaneo's study on the effects of maternal smoking on infant birthweight which we've been using the last few weeks. We will focus just on three variables unlike previous weeks: `bweight` is the birthweight of the baby in grams, `mbsmoke` is an indicator of whether the mother smoked during pregnancy and `mage` is the mother's age. We will focus on estimating the relationship between maternal smoking and birthweight controlling for alcohol use.

Prompt

1. Load the data into R using the `haven` package.

```r
library(haven)
data <- read_dta("https://github.com/scunning1975/mixtape/raw/master/cattaneo2.dta")
```

2. Estimate the effect `mbsmoke` on `bweight` controlling for `alcohol` using OLS.

```r
reg3 <- lm(bweight ~ mbsmoke + alcohol, data = data)
summary(reg3)
```

```
Call:
lm(formula = bweight ~ mbsmoke + alcohol, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-3074.42  -309.97    30.03   356.58  2085.58 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3414.415      9.297 367.260   <2e-16 ***
mbsmoke     -269.440     21.728 -12.401   <2e-16 ***
alcohol      -80.003     47.824  -1.673   0.0944 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 568.8 on 4639 degrees of freedom
Multiple R-squared:  0.03485,   Adjusted R-squared:  0.03443 
F-statistic: 83.74 on 2 and 4639 DF,  p-value: < 2.2e-16
```

3. Using the covariance and variance formulas from earlier, manually calculate the intercept, the coefficient on `mbsmoke` and the coefficient on `alcohol`. Confirm that your manual calculations are the same as what you found in part 2.

```r
# Define variables
Y  <- data$bweight
X1 <- data$mbsmoke
X2 <- data$alcohol

# Remove missing values
complete <- complete.cases(Y, X1, X2)
Y  <- Y[complete]; X1 <- X1[complete]; X2 <- X2[complete]

# Compute covariances and variances
Cov_X1Y  <- mean(X1*Y)  - mean(X1)*mean(Y)
Cov_X2Y  <- mean(X2*Y)  - mean(X2)*mean(Y)
Var_X1   <- mean(X1^2)  - mean(X1)^2
Var_X2   <- mean(X2^2)  - mean(X2)^2
Cov_X1X2 <- mean(X1*X2) - mean(X1)*mean(X2)

# Compute coefficients manually
b1 <- (Cov_X1Y*Var_X2 - Cov_X2Y*Cov_X1X2) / (Var_X1*Var_X2 - Cov_X1X2^2)
b2 <- (Cov_X2Y*Var_X1 - Cov_X1Y*Cov_X1X2) / (Var_X1*Var_X2 - Cov_X1X2^2)
b0 <- mean(Y) - b1*mean(X1) - b2*mean(X2)

# Display manual coefficients
manual_results <- c(b0 = b0, b1 = b1, b2 = b2)
manual_results
```

```
        b0          b1          b2 
3414.41509 -269.44026  -80.00322 
```

```r
# Compare to lm() results
summary(lm(bweight ~ mbsmoke + alcohol, data = data))
```

```

Call:
lm(formula = bweight ~ mbsmoke + alcohol, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
```

```
-3074.42  -309.97    30.03   356.58  2085.58

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3414.415      9.297 367.260   <2e-16 ***
mbsmoke     -269.440     21.728 -12.401   <2e-16 ***
alcohol      -80.003     47.824  -1.673   0.0944 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 568.8 on 4639 degrees of freedom
Multiple R-squared:  0.03485,	Adjusted R-squared:  0.03443 
F-statistic: 83.74 on 2 and 4639 DF,  p-value: < 2.2e-16
```

4. Using the FWL theorem, replicate the coefficient on `mbsmoke` step by step. What does the coefficient on residualized `mbsmoke` tell us?

```r
# Step 1: regress educ on kww
fwl1 <- lm(mbsmoke ~ alcohol, data = data)

# Step 2: get residuals (the variation in educ not explained by kww)
data$resid_mbsmoke <- resid(fwl1)

# Step 3: regress lwage on these residuals
fwl2 <- lm(bweight ~ resid_mbsmoke, data = data)

summary(fwl2)
```

```
Call:
lm(formula = bweight ~ resid_mbsmoke, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-3068.77  -318.77    21.23   362.23  2091.23 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3361.680      8.359  402.14   <2e-16 ***
resid_mbsmoke -269.440     21.758  -12.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 569.5 on 4640 degrees of freedom
Multiple R-squared:  0.03199,	Adjusted R-squared:  0.03178 
F-statistic: 153.4 on 1 and 4640 DF,  p-value: < 2.2e-16
```

```r
# Compare to the coefficient from full model
summary(lm(bweight ~ mbsmoke + alcohol, data = data))
```

```

Call:
lm(formula = bweight ~ mbsmoke + alcohol, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-3074.42  -309.97    30.03   356.58  2085.58 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3414.415      9.297 367.260   <2e-16 ***
mbsmoke     -269.440     21.728 -12.401   <2e-16 ***
alcohol      -80.003     47.824  -1.673   0.0944 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 568.8 on 4639 degrees of freedom
Multiple R-squared:  0.03485,	Adjusted R-squared:  0.03443 
F-statistic: 83.74 on 2 and 4639 DF,  p-value: < 2.2e-16
```