Econometrics I

Lecture 5: Extended Example: The Wage Equation

Paul T. Scott
NYU Stern

Fall 2018

Preliminaries

- I'm posting problem set solutions and grades on NYU Classes.
- Start thinking about your group project groups and topics, if you haven't already! I will distribute some topic suggestions by next week's lecture.

Mincerian Regression

- Recall the Mincerian regression (wage equation):

$$\ln wage_i = \beta_0 + \beta_{ed} Edu_i + \beta_{exp} Exp_i + \beta_{Fem} Fem_i + \cdots + \varepsilon_i$$

- Let's revisit estimating this with the Cornwell and Rupert data.

Baseline Results

```
> suppressMessages(library(tidyverse))
Warning messages:
1: package 'tibble' was built under R version 3.4.3
2: package 'tidyr' was built under R version 3.4.4
3: package 'purrr' was built under R version 3.4.4
4: package 'forcats' was built under R version 3.4.3
> data <- read.csv('cornwell-rupert.csv')
> #data <- cbind(data, EXP2=data$EXP^2)
> data <- data %>% mutate(EXP2 = EXP^2)
>
> reg_1 <- lm(LWAGE ~ ED + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+             + MS + FEM + UNION, data = data)
> summary(reg_1)
```

```
Call:
lm(formula = LWAGE ~ ED + EXP + EXP2 + WKS + OCC + SOUTH + SMSA +
    MS + FEM + UNION, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2034 -0.2379 -0.0071  0.2327  2.1380

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.245e+00  7.170e-02  73.153  < 2e-16 ***
ED           5.654e-02  2.612e-03  21.644  < 2e-16 ***
EXP          4.045e-02  2.174e-03  18.605  < 2e-16 ***
EXP2        -6.811e-04  4.783e-05 -14.242  < 2e-16 ***
WKS          4.485e-03  1.090e-03   4.115 3.94e-05 ***
OCC         -1.405e-01  1.472e-02  -9.544  < 2e-16 ***
SOUTH       -7.210e-02  1.249e-02  -5.773 8.37e-09 ***
SMSA         1.390e-01  1.207e-02  11.513  < 2e-16 ***
MS           6.736e-02  2.063e-02   3.265   0.0011 **
FEM         -3.892e-01  2.518e-02 -15.457  < 2e-16 ***
UNION        9.015e-02  1.289e-02   6.993 3.13e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3524 on 4154 degrees of freedom
Multiple R-squared:  0.4183,    Adjusted R-squared:  0.4169
F-statistic: 298.7 on 10 and 4154 DF,  p-value: < 2.2e-16
```

Relaxing Linear Effect of Education

```
> data <- data %>% mutate(NOHS = ifelse(ED <= 8, 1, 0),
+                         SOMEHS = ifelse((ED >= 9) & (ED <= 11), 1, 0),
+                         HS = ifelse(ED == 12, 1, 0),
+                         SOMECOL = ifelse((ED >= 13) & (ED <= 15), 1, 0),
+                         COL = ifelse(ED == 16, 1, 0),
+                         POST = ifelse(ED >= 17, 1, 0)
+ )
> data <- data %>% mutate(SUM = NOHS + SOMEHS +
+                             HS + SOMECOL + COL + POST)
>
> reg_2 <- lm(LWAGE ~ NOHS + SOMEHS + HS + SOMECOL + COL
+             + POST + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+             + MS + FEM + UNION, data = data)
> summary(reg_2)
```

- Note that we're missing a coefficient on one of the education categories.

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.188e+00  5.888e-02 105.112  < 2e-16 ***
NOHS        -5.337e-01  2.947e-02 -18.108  < 2e-16 ***
SOMEHS      -3.937e-01  2.496e-02 -15.776  < 2e-16 ***
HS          -2.855e-01  2.106e-02 -13.554  < 2e-16 ***
SOMECOL     -1.973e-01  2.214e-02  -8.912  < 2e-16 ***
COL         -2.711e-02  2.127e-02  -1.274 0.202570
POST                NA         NA      NA       NA
EXP          4.100e-02  2.184e-03  18.769  < 2e-16 ***
EXP2        -6.940e-04  4.799e-05 -14.461  < 2e-16 ***
WKS          4.599e-03  1.103e-03   4.168 3.14e-05 ***
OCC         -1.386e-01  1.509e-02  -9.184  < 2e-16 ***
SOUTH       -7.618e-02  1.259e-02  -6.052 1.56e-09 ***
SMSA         1.436e-01  1.211e-02  11.861  < 2e-16 ***
MS           6.919e-02  2.070e-02   3.343 0.000837 ***
FEM         -3.819e-01  2.532e-02 -15.080  < 2e-16 ***
UNION        9.402e-02  1.300e-02   7.235 5.52e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3529 on 4150 degrees of freedom
Multiple R-squared:  0.4174,    Adjusted R-squared:  0.4154
F-statistic: 212.4 on 14 and 4150 DF,  p-value: < 2.2e-16
```

Dropping a Category Dummy

```
> # regression with categories, dropping one
> reg_3 <- lm(LWAGE ~ SOMEHS + HS + SOMECOL + COL
+             + POST + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+             + MS + FEM + UNION, data = data)
> summary(reg_3)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.655e+00  6.342e-02  89.170  < 2e-16 ***
SOMEHS       1.400e-01  2.485e-02   5.632 1.90e-08 ***
HS           2.482e-01  2.292e-02  10.827  < 2e-16 ***
SOMECOL      3.364e-01  2.679e-02  12.555  < 2e-16 ***
COL          5.066e-01  2.835e-02  17.868  < 2e-16 ***
POST         5.337e-01  2.947e-02  18.108  < 2e-16 ***
EXP          4.100e-02  2.184e-03  18.769  < 2e-16 ***
EXP2        -6.940e-04  4.799e-05 -14.461  < 2e-16 ***
WKS          4.599e-03  1.103e-03   4.168 3.14e-05 ***
OCC         -1.386e-01  1.509e-02  -9.184  < 2e-16 ***
SOUTH       -7.618e-02  1.259e-02  -6.052 1.56e-09 ***
SMSA         1.436e-01  1.211e-02  11.861  < 2e-16 ***
MS           6.919e-02  2.070e-02   3.343 0.000837 ***
FEM         -3.819e-01  2.532e-02 -15.080  < 2e-16 ***
UNION        9.402e-02  1.300e-02   7.235 5.52e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3529 on 4150 degrees of freedom
Multiple R-squared:  0.4174,	Adjusted R-squared:  0.4154 
F-statistic: 212.4 on 14 and 4150 DF,  p-value: < 2.2e-16
```

Dropping the Constant Term

```
> reg_4 <- lm(LWAGE ~ NOHS + SOMEHS + HS + SOMECOL + COL
+            + POST + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+            + MS + FEM + UNION -1, data = data)
> summary(reg_4)
```

```
Coefficients:
         Estimate Std. Error t value Pr(>|t|)    
NOHS     5.655e+00  6.342e-02  89.170  < 2e-16 ***
SOMEHS   5.795e+00  6.240e-02  92.864  < 2e-16 ***
HS       5.903e+00  6.095e-02  96.855  < 2e-16 ***
SOMECOL  5.991e+00  6.096e-02  98.276  < 2e-16 ***
COL      6.161e+00  5.966e-02 103.268  < 2e-16 ***
POST     6.188e+00  5.888e-02 105.112  < 2e-16 ***
EXP      4.100e-02  2.184e-03  18.769  < 2e-16 ***
EXP2    -6.940e-04  4.799e-05 -14.461  < 2e-16 ***
WKS      4.599e-03  1.103e-03   4.168 3.14e-05 ***
OCC     -1.386e-01  1.509e-02  -9.184  < 2e-16 ***
SOUTH   -7.618e-02  1.259e-02  -6.052 1.56e-09 ***
SMSA     1.436e-01  1.211e-02  11.861  < 2e-16 ***
MS       6.919e-02  2.070e-02   3.343 0.000837 ***
FEM     -3.819e-01  2.532e-02 -15.080  < 2e-16 ***
UNION    9.402e-02  1.300e-02   7.235 5.52e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3529 on 4150 degrees of freedom
Multiple R-squared:  0.9972,    Adjusted R-squared:  0.9972 
F-statistic: 9.96e+04 on 15 and 4150 DF,  p-value: < 2.2e-16
```

Two Ways of Testing Hypotheses

```
> suppressMessages(library(car))
> suppressMessages(library(sandwich))
>
> # separate male and female categories
> data <- data %>% mutate(MALE = ifelse(FEM == 1, 0, 1))
> reg_5 <- lm(LWAGE ~ ED + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+             + MS + FEM + MALE + UNION -1, data = data)
> linearHypothesis(reg_5, c("FEM = MALE"),
+                  vcov = vcovHC(reg_5, type = "HC1"))
Linear hypothesis test

Hypothesis:
FEM - MALE = 0

Model 1: restricted model
Model 2: LWAGE ~ ED + EXP + EXP2 + WKS + OCC + SOUTH + SMSA + MS + FEM +
    MALE + UNION - 1

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1   4155
2   4154  1 263.33 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
> # now with intercept and different (but equivalent) hypothesis test
> data <- data %>% mutate(MALE = ifelse(FEM == 1, 0, 1))
> reg_6 <- lm(LWAGE ~ ED + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+             + MS + FEM + UNION , data = data)
> linearHypothesis(reg_6, c("FEM = 0"), vcov = vcovHC(reg_6, type = "HC1"))
Linear hypothesis test

Hypothesis:
FEM = 0

Model 1: restricted model
Model 2: LWAGE ~ ED + EXP + EXP2 + WKS + OCC + SOUTH + SMSA + MS + FEM +
    UNION

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1   4155
2   4154  1 263.33 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Delta Method I

- We know how to compute standard errors on our coefficients, but sometimes we are interested in *functions of those statistics*
- For example, if we have linear and quadratic terms of experience ($\beta_{exp} Exp_i + \beta_{exp2} Exp_i^2$), then the model doesn't just have a simple "effect of experience".
- We might be interested in the effect of experience for somebody with 10 years of experience:

$$\left. \frac{d \ln Wage_i}{dExp_i} \right|_{Exp_i = 10} = \beta_{exp} + 2\beta_{exp2} exp_i = \beta_{exp} + 20\beta_{exp2}$$

Delta Method II

- Suppose we have an asymptotic distribution for an estimator:

$$\sqrt{n}\left(\mathbf{b}-\boldsymbol{\beta}\right) \Rightarrow_d \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right).$$

- Then the asymptotic distribution of a function of the estimator is

$$\sqrt{n}\left(g\left(\mathbf{b}\right)-g\left(\boldsymbol{\beta}\right)\right) \Rightarrow_d \mathcal{N}\left(\mathbf{0}, \left(\nabla g\left(\boldsymbol{\beta}\right)\right)' \boldsymbol{\Sigma} \nabla g\left(\boldsymbol{\beta}\right)\right),$$

  where $\nabla g\left(\boldsymbol{\beta}\right)$ is the gradient of $g\left(\boldsymbol{\beta}\right)$:

$$\nabla g\left(\boldsymbol{\beta}\right) = \begin{pmatrix} \frac{\partial g(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial g(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial g(\boldsymbol{\beta})}{\partial \beta_K} \end{pmatrix}.$$

- Note that we can estimate $\nabla g\left(\boldsymbol{\beta}\right)$ with $\nabla g\left(\mathbf{b}\right)$.

Delta Method in R

```
> library(alr3)
> deltaMethod(reg_1, "EXP + 20*EXP2")
                    Estimate          SE      2.5 %     97.5 %
EXP + 20 * EXP2 0.02682765 0.001267014 0.02434435 0.02931095
```

Numerical Bootstrap

- Given the the asymptotic distribution of a parameter estimate

$$\mathbf{b} \sim_d \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

  we have an estimated density function $\hat{f}$. Let $\hat{f}$ be the multivariate normal density with mean $\boldsymbol{\beta}$ and variance $\boldsymbol{\Sigma}$.
- We can simulate the asymptotic distribution of $g(\mathbf{b})$ by
  - Simulating draws $\mathbf{b}_m$ for $m = 1, 2, \ldots M$ from $\hat{f}$
  - Computing $g(\mathbf{b}_m)$ for each draw
  - Then $(g(\mathbf{b}_1), g(\mathbf{b}_2), \ldots, g(\mathbf{b}_M))$ will be a simulated asymptotic distribution for
- This can be useful when you have code to compute $g(\cdot)$, but computing the derivative $g(\cdot)$ would be difficult. For example, when $g(\cdot)$ represents an complex behavioral (or equilibrium) model.

Heterogeneous Effects

- When we have a model of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

we're implicitly saying that the effect of X_1 is the same for all individuals.

- Often we would like to relax this, allowing different groups to have different slopes with respect to X_1.
- This is easy as long as the group membership is observed in the data. We simply interact the regressor with dummy variables:

$$Y_i = \beta_0 + \beta_{0F} D_{Fi} + \beta_1 X_{1i} + \beta_2 X_{1i} D_{Fi} + \varepsilon_i$$

where D_{Fi} is a dummy variable for whether individual i is female. Note that we have allowed for the intercepts and slopes to vary by sex here.

Heterogeneous Effects in R

```
> data <- cbind(data, EDFEM=data$ED*data$FEM)
> reg_9 <- lm(LWAGE ~ ED + EDFEM + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+             + MS + FEM + UNION, data = data)
> summary(reg_9)
```

- Here, we construct interactions manually, allowing education to have a different effect for males and females.
- Does education have significantly different effects for males and females?

```
Call:
lm(formula = LWAGE ~ ED + EDFEM + EXP + EXP2 + WKS + OCC + SOUTH +
    SMSA + MS + FEM + UNION, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19425 -0.23540 -0.00569  0.23005  2.13574

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.277e+00  7.212e-02  73.173  < 2e-16 ***
ED           5.413e-02  2.689e-03  20.126  < 2e-16 ***
EDFEM        2.520e-02  6.868e-03   3.669 0.000246 ***
EXP          4.053e-02  2.171e-03  18.668  < 2e-16 ***
EXP2        -6.842e-04  4.776e-05 -14.324  < 2e-16 ***
WKS          4.518e-03  1.088e-03   4.151 3.37e-05 ***
OCC         -1.383e-01  1.472e-02  -9.396  < 2e-16 ***
SOUTH       -7.375e-02  1.248e-02  -5.910 3.70e-09 ***
SMSA         1.402e-01  1.206e-02  11.626  < 2e-16 ***
MS           6.539e-02  2.061e-02   3.173 0.001520 **
FEM         -7.153e-01  9.235e-02  -7.745 1.19e-14 ***
UNION        8.476e-02  1.296e-02   6.542 6.81e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3519 on 4153 degrees of freedom
Multiple R-squared:  0.4201,    Adjusted R-squared:  0.4186
F-statistic: 273.5 on 11 and 4153 DF,  p-value: < 2.2e-16
```

Interactions with the : Operator

```
> reg_10 <- lm(LWAGE ~ ED + ED:FEM + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+              + MS + FEM + UNION, data = data)
> summary(reg_10)
```

- We can avoid creating the interactions mandually with the : operator.

```
Call:
lm(formula = LWAGE ~ ED + ED:FEM + EXP + EXP2 + WKS + OCC + SOUTH +
    SMSA + MS + FEM + UNION, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19425 -0.23540 -0.00569  0.23005  2.13574

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.277e+00  7.212e-02  73.173  < 2e-16 ***
ED           5.413e-02  2.689e-03  20.126  < 2e-16 ***
EXP          4.053e-02  2.171e-03  18.668  < 2e-16 ***
EXP2        -6.842e-04  4.776e-05 -14.324  < 2e-16 ***
WKS          4.518e-03  1.088e-03   4.151 3.37e-05 ***
OCC         -1.383e-01  1.472e-02  -9.396  < 2e-16 ***
SOUTH       -7.375e-02  1.248e-02  -5.910 3.70e-09 ***
SMSA         1.402e-01  1.206e-02  11.626  < 2e-16 ***
MS           6.539e-02  2.061e-02   3.173 0.001520 **
FEM         -7.153e-01  9.235e-02  -7.745 1.19e-14 ***
UNION        8.476e-02  1.296e-02   6.542 6.81e-11 ***
ED:FEM       2.520e-02  6.868e-03   3.669 0.000246 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3519 on 4153 degrees of freedom
Multiple R-squared:  0.4201,	Adjusted R-squared:  0.4186
F-statistic: 273.5 on 11 and 4153 DF,  p-value: < 2.2e-16
```

Interactions with the * Operator

```
> reg_11 <- lm(LWAGE ~ ED*FEM + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+              + MS  + UNION, data = data)
> summary(reg_11)
```

- This version gives us the interacted and uninteracted terms with one term.

```
Call:
lm(formula = LWAGE ~ ED * FEM + EXP + EXP2 + WKS + OCC + SOUTH +
    SMSA + MS + UNION, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19425 -0.23540 -0.00569  0.23005  2.13574

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.277e+00  7.212e-02  73.173  < 2e-16 ***
ED           5.413e-02  2.689e-03  20.126  < 2e-16 ***
FEM         -7.153e-01  9.235e-02  -7.745 1.19e-14 ***
EXP          4.053e-02  2.171e-03  18.668  < 2e-16 ***
EXP2        -6.842e-04  4.776e-05 -14.324  < 2e-16 ***
WKS          4.518e-03  1.088e-03   4.151 3.37e-05 ***
OCC         -1.383e-01  1.472e-02  -9.396  < 2e-16 ***
SOUTH       -7.375e-02  1.248e-02  -5.910 3.70e-09 ***
SMSA         1.402e-01  1.206e-02  11.626  < 2e-16 ***
MS           6.539e-02  2.061e-02   3.173 0.001520 **
UNION        8.476e-02  1.296e-02   6.542 6.81e-11 ***
ED:FEM       2.520e-02  6.868e-03   3.669 0.000246 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3519 on 4153 degrees of freedom
Multiple R-squared:  0.4201,    Adjusted R-squared:  0.4186
F-statistic: 273.5 on 11 and 4153 DF,  p-value: < 2.2e-16
```

Minncerian Regression: Sample Selection

- What happens when some of the data is missing in a non-random way?
- For example, let's imagine that the low-wage individuals drop out of the labor market.
- Note: this may already be happening in the data, but let's make it happen more.

```
> reg_7 <- lm(LWAGE ~ ED + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+              + MS + FEM + UNION, data = subset(data, LWAGE>=6))
> summary(reg_7)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.609e+00  7.399e-02  75.805  < 2e-16 ***
ED           4.882e-02  2.569e-03  19.005  < 2e-16 ***
EXP          3.199e-02  2.177e-03  14.694  < 2e-16 ***
EXP2        -5.169e-04  4.801e-05 -10.766  < 2e-16 ***
WKS          2.827e-03  1.106e-03   2.556   0.0106 *
OCC         -9.138e-02  1.435e-02  -6.369 2.12e-10 ***
SOUTH       -5.565e-02  1.213e-02  -4.588 4.63e-06 ***
SMSA         1.118e-01  1.165e-02   9.596  < 2e-16 ***
MS           9.364e-03  2.049e-02   0.457   0.6477
FEM         -3.528e-01  2.625e-02 -13.439  < 2e-16 ***
UNION        2.012e-02  1.255e-02   1.603   0.1090
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3257 on 3833 degrees of freedom
Multiple R-squared:  0.3148,	Adjusted R-squared:  0.3131
F-statistic: 176.1 on 10 and 3833 DF,  p-value: < 2.2e-16
```

- Recall that the coefficient on ED in the original regression was 5.654e-01

Mincerian Regression: Measurement Error

- What happens if one of the variables of interest is measured with error?
- Let's say the the recorded education might be one year more or less than the person's actual education.
- Note: this may already be happening in the data, but let's make it happen more.

```
> noise <- sample(-1:1,dim(data)[1],replace=T)
> data <- cbind(data, ED_NOISY=data$ED + noise)
>
> reg_8 <- lm(LWAGE ~ ED_NOISY + EXP + EXP2 + WKS + OCC + SOUTH + SMSA
+            + MS + FEM + UNION, data = data)
> summary(reg_8)
```

```
Call:
lm(formula = LWAGE ~ ED_NOISY + EXP + EXP2 + WKS + OCC + SOUTH +
    SMSA + MS + FEM + UNION, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.23660 -0.23773 -0.00609  0.24132  2.09688

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.341e+00  7.094e-02  75.283  < 2e-16 ***
ED_NOISY     4.967e-02  2.451e-03  20.262  < 2e-16 ***
EXP          4.049e-02  2.188e-03  18.501  < 2e-16 ***
EXP2        -6.862e-04  4.813e-05 -14.257  < 2e-16 ***
WKS          4.618e-03  1.097e-03   4.209 2.62e-05 ***
OCC         -1.600e-01  1.457e-02 -10.977  < 2e-16 ***
SOUTH       -7.690e-02  1.255e-02  -6.127 9.79e-10 ***
SMSA         1.435e-01  1.214e-02  11.825  < 2e-16 ***
MS           6.591e-02  2.077e-02   3.174  0.00151 **
FEM         -3.972e-01  2.533e-02 -15.683  < 2e-16 ***
UNION        8.402e-02  1.295e-02   6.486 9.85e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3547 on 4154 degrees of freedom
Multiple R-squared:  0.4109,    Adjusted R-squared:  0.4095
F-statistic: 289.7 on 10 and 4154 DF,  p-value: < 2.2e-16
```

- Recall that the coefficient on ED in the original regression was 5.654e-01

Omitted Variables Bias, Revisited

- Suppose the econometrician only observes regressors $\mathbf{X}$, but the true model is
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\gamma + \varepsilon,$$
- The OLS estimator will equal
$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{z}\gamma + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon$$
- The last term is mean zero given the strict exogeneity assumption.
- Note that the second term will not be zero if $\mathbf{X}$ and $\mathbf{z}$ are correlated; i.e. if $\mathbf{X}'\mathbf{z} \neq 0$.
- Implication: correlation between omitted variables and the observed regressors makes OLS biased.

Omitted Variables Bias II

- Using the Frisch-Waugh theorem, we can show that

$$E\left[b_{OLS,k}|\mathbf{X},\mathbf{z}\right] = \beta_k + \gamma\left(\frac{Cov\left(z, x_k|\mathbf{X}_{-k}\right)}{Var\left(x_k|\mathbf{X}_{-k}\right)}\right)$$

  where $\mathbf{X}_{-k}$ refers to all the regressors besides x_k.
- Suppose positive correlation between regressor x_k and omitted variable z.
- Also suppose $\beta_k > 0$ and $\gamma > 0$ so both variables have positive effects.
- Let's compare the average value of the dependent variable for $x_k = 0$ and $x_k = 1$. Two things change between these points:
  - Dependent variable Y increases by β_k because of direct effect of x_k.
  - Value of z should be higher because of the positive correlation between x_k and z. Higher values of z also contribute to a higher dependent variable because $\gamma > 0$.

Omitted Variables Bias in Mincerian Regression

- What sort of variables might the wage equation omit, and how would you expect them to affect the estimated coefficients?

```
Call:
lm(formula = LWAGE ~ ED + EXP + EXP2 + WKS + OCC + SOUTH + SMSA +
    MS + FEM + UNION, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2034 -0.2379 -0.0071  0.2327  2.1380

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.245e+00  7.170e-02  73.153  < 2e-16 ***
ED           5.654e-02  2.612e-03  21.644  < 2e-16 ***
EXP          4.045e-02  2.174e-03  18.605  < 2e-16 ***
EXP2        -6.811e-04  4.783e-05 -14.242  < 2e-16 ***
WKS          4.485e-03  1.090e-03   4.115 3.94e-05 ***
OCC         -1.405e-01  1.472e-02  -9.544  < 2e-16 ***
SOUTH       -7.210e-02  1.249e-02  -5.773 8.37e-09 ***
SMSA         1.390e-01  1.207e-02  11.513  < 2e-16 ***
MS           6.736e-02  2.063e-02   3.265   0.0011 **
FEM         -3.892e-01  2.518e-02 -15.457  < 2e-16 ***
UNION        9.015e-02  1.289e-02   6.993 3.13e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3524 on 4154 degrees of freedom
Multiple R-squared:  0.4183,    Adjusted R-squared:  0.4169
F-statistic: 298.7 on 10 and 4154 DF,  p-value: < 2.2e-16
```

Outliers and Influential Observations

- **Outliers** refer to observations that are "far away" from the rest of the data. They can be due to errors in the data. There is no standard formal definition.
- **Influential Observations** refer to observations that have a large result on the estimated coefficients. Again, no standard definition.
- What to do? Greene:"*It is difficult to draw firm general conclusions... It remains likely that in very small samples, some caution and close scrutiny of the data are called for.*" I'd say that's true even in large samples, but there isn't a generally accepted way of quantifying what counts as appropriate "caution and close scrutiny."
- Removing clear outliers from datasets is generally considered good practice (especially when such observations are likely errors).