POLS 6481. Research Design and Quantitative Methods II
Lecture 5. Multiple Regression

Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

Outline:

1. The 2–regressor linear model
2. Calculation of multiple regression coefficients
3. Interpretation of multiple regression coefficients
4. Illustration: *Air Pollution and Mortality*

POLS 6481. Research Design and Quantitative Methods II

Lecture 5. Multiple Regression

Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

0. Transition "all else being equal" [1.4 & 3.2c]

"The notion of **ceteris paribus** – which means "other factors being equal" – plays an important role in causal analysis…. Holding other factors fixed is critical for policy analysis…. The key question in most empirical studies is: Have enough other factors been held fixed to make a cause for causality?"

Example 1.1 "Economic Model of Crime"

The effects of the criminal justice policies (the probability of being caught committing crime, the conditional probability of being convicted if caught committing crime, the severity of sentences, etc.) on individuals' criminal activity…

… controlling for wages that can be earned in legal employment, age, etc.

Example 1.2 "Job Training and Worker Productivity"

The effect of a computer training program on a workers' wages…

… controlling for the worker's years of education and years of experience

POLS 6481. Research Design and Quantitative Methods II

Lecture 5. Multiple Regression

Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

1. The 2–regressor linear model

“explained” variable	y
“explanatory” variables	x_1, x_2
intercept parameter	β_0
slope parameters	β_1, β_2
population regression function	$E(y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
disturbance	u

3 conditions: [MLR.1] Linearity in Parameters

[MLR.2] Simple Random Sampling $\{(x_{1i}, x_{2i}, y_i): i = 1,\ldots, n\}$

Combine MLR.1 & MLR.2: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i\,, i = 1, \ldots, n$

[MLR.4] Zero Conditional Mean Disturbance $E(u \mid x_1, x_2) = 0$

See p. 70 in Wooldridge *5e*

POLS 6481. Research Design and Quantitative Methods II
Lecture 5. Multiple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

1. The 2–regressor linear model

Estimate model coefficients $\widehat{\beta_0}\,, \widehat{\beta_1}\,, \widehat{\beta_2}$

Calculate fitted values $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i}$ [3.9, 3.14]

Calculate residuals $\widehat{u_i} = y_i - \widehat{y_i}$

$= y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{1i} - \widehat{\beta_2} x_{2i}$ [3.10]

Check the following three properties: $\sum \widehat{u_i} = 0$ [3.13]

$\sum \widehat{u_i} x_{ki} = 0$ for $k = 1, 2$

(*regression* *plane* *passes through* $(\overline{x_1}, \overline{x_2}, \overline{y})$) $\overline{y} = E(y \mid x_1 = \overline{x_1},\ \ x_2 = \overline{x_2})$

$= \widehat{\beta_0} + \widehat{\beta_1} \cdot \overline{x_1} + \widehat{\beta_2} \cdot \overline{x_2}$

(Versions of two of these three properties are at the bottom of p. 77)

To show the property that $\sum \widehat{u_i}^{\,2}$ [3.12] is *minimized* requires calculus (see p. 73, 113-114).

POLS 6481. Research Design and Quantitative Methods II

Lecture 5. Multiple Regression

Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

2. Calculation of multiple regression coefficients

For all of the below equations, assume summation is over *i* = 1, … , *n*

Variances

$$var(x_1) = \frac{1}{n-1}\sum(x_{1i} - \overline{x_1})^2$$

$$var(x_2) = \frac{1}{n-1}\sum(x_{2i} - \overline{x_2})^2$$

$$var(y) = \frac{1}{n-1}\sum(y_i - \bar{y})^2$$

Covariances

$$cov(x_1, y) = \frac{1}{n-1}\sum(x_{1i} - \overline{x_1}) \cdot (y_i - \bar{y})$$

$$cov(x_2, y) = \frac{1}{n-1}\sum(x_{2i} - \overline{x_2}) \cdot (y_i - \bar{y})$$

$$cov(x_1, x_2) = \frac{1}{n-1}\sum(x_{1i} - \overline{x_1}) \cdot (x_{2i} - \overline{x_2})$$

The *Variance–Covariance Matrix* has variances on diagonal, covariances off diagonal.

POLS 6481. Research Design and Quantitative Methods II

Lecture 5. Multiple Regression

Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

2. Calculation of multiple regression coefficients

Differentiate $\sum\left(\hat{y}_i - \widehat{\beta_0} - \widehat{\beta_1}x_{1i} - \widehat{\beta_2}x_{2i}\right)^2$ with respect to $\widehat{\beta_1}$ and $\widehat{\beta_2}$ (and $\widehat{\beta_0}$ too) and set the first derivatives equal to zero. Algebra reveals the 'normal equations' [3.13, sort of].

Or, you can start with the assumptions that $E(\hat{u} \cdot x_1) = 0$ and $E(\hat{u} \cdot x_2) = 0$ …

After calculus & algebra:

$$\widehat{\beta_1} = \frac{cov(x_1, y)}{var(x_1)} - \widehat{\beta_2} \cdot \frac{cov(x_1, x_2)}{var(x_1)}$$

$$\widehat{\beta_2} = \frac{cov(x_2, y)}{var(x_2)} - \widehat{\beta_1} \cdot \frac{cov(x_1, x_2)}{var(x_2)}$$

POLS 6481. Research Design and Quantitative Methods II
Lecture 5. Multiple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

2. Calculation of multiple regression coefficients

After many steps of algebra:

$$\widehat{\beta_1} = \frac{\frac{cov(x_1,y)}{var(x_1)} - \frac{cov(x_2,y)}{var(x_2)} \cdot \frac{cov(x_1,x_2)}{var(x_1)}}{1 - {r_{12}}^2}$$

$$\widehat{\beta_2} = \frac{\frac{cov(x_2,y)}{var(x_2)} - \frac{cov(x_1,y)}{var(x_1)} \cdot \frac{cov(x_1,x_2)}{var(x_2)}}{1 - {r_{12}}^2}$$

Where the correlation between x_1 and x_2 is:

$$r_{12} = \frac{cov(x_1,x_2)}{\sqrt{var(x_1) \cdot var(x_2)}}$$

Note on perfect multicollinearity: if $r_{12} = 1$, then we cannot compute $\widehat{\beta_1}$ or $\widehat{\beta_2}$.

Because the denominator would equal 0, statistical software drops x_1 or x_2 to estimate.

So, we add: [MLR.3] No perfect multicollinearity between regressors.

POLS 6481. Research Design and Quantitative Methods II

Lecture 5. Multiple Regression

Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

2. Calculation of multiple regression coefficients

Compare multiple regression…

$$\widehat{\beta_1} = \frac{\frac{cov(x_1,y)}{var(x_1)} - \frac{cov(x_2,y)}{var(x_2)} \cdot \frac{cov(x_1,x_2)}{var(x_1)}}{1 - {r_{12}}^2}$$

…to simple regression

$$\widetilde{\beta_1} = \frac{cov(x_1,y)}{var(x_1)}$$

$\widehat{\beta_1} = \widetilde{\beta_1}$ if (1) ${r_{12}}^2 = \frac{cov(x_1,x_2)}{\sqrt{var(x_1) \cdot var(x_2)}} = 0$ i.e., x_1 & x_2 are uncorrelated

(2) $\frac{cov(x_1,x_2)}{var(x_1)} = 0$ i.e., x_1 & x_2 do not covary (redundant!)

(3) $\frac{cov(x_2,y)}{var(x_2)} = 0$ i.e., x_2 & y do not covary

If assumptions (1) and (2) are satisfied, then $\widetilde{\beta_1}$ and $\widehat{\beta_1}$ are equally unbiased; however… if assumption (3) is not satisfied, then including x_2 in the specification may still be helpful for reducing standard errors and improving model fit.

POLS 6481. Research Design and Quantitative Methods II
Lecture 5. Multiple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

3. Interpretation of multiple regression coefficients

Example 1. wage = f(education, experience)

$$\widehat{wage} = \widehat{\beta_0} + \widehat{\beta_1} \cdot educ + \widehat{\beta_2} \cdot exper$$

Wooldridge uses the terms **hold fixed** (76-77) and ‘ceteris paribus’ (12, 74, 76).

Suppose we calculate a predicted wage, given a level of education and experience, and we wish to explore effects of different levels of education and/or experience. Then:

$$\Delta\widehat{wage} = \widehat{\beta_1} \cdot \Delta\, educ + \widehat{\beta_2} \cdot \Delta\, exper$$

To “hold fixed” education means Δ *educ* = 0: $\dfrac{\Delta\widehat{wage}}{\Delta educ} = \widehat{\beta_1}$

To “hold fixed” experience means Δ *exper* = 0 : $\dfrac{\Delta\widehat{wage}}{\Delta exper} = \widehat{\beta_2}$

The estimates $\widehat{\beta_1}$ and $\widehat{\beta_2}$ are have **partial effect** interpretations. “The power of multiple regression analysis is that it provides this **ceteris paribus** interpretation even though the data have not been collected in a ceteris paribus fashion” (p. 76).

POLS 6481. Research Design and Quantitative Methods II

Lecture 5. Multiple Regression

Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

3. Interpretation of multiple regression coefficients

Example 2. consumption = f(income, income^2*)*

$$\widehat{cons} = \widehat{\beta_0} + \widehat{\beta_1} \cdot income + \widehat{\beta_2} \cdot income^2$$

$$\frac{\Delta \widehat{cons}}{\Delta income} = \widehat{\beta_1} + 2\widehat{\beta_2} \cdot income$$ ← requires calculus

Three notes:

1. Solving the marginal effects is tricky and requires calculus
2. It's nonsensical to talk about "all else equal" between these two variables
3. *income* and *income*^{2} are highly correlated, so this increases "multicollinearity"

(Don't worry, we'll come back to this later in the semester…)

POLS 6481. Research Design and Quantitative Methods II

Lecture 5. Multiple Regression

Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

3. Interpretation of multiple regression coefficients

Example 3. mortality = f(pollution, socioeconomics)

$$\widehat{mortality} = \widehat{\beta_0} + \widehat{\beta_1} \cdot pollution + \widehat{\beta_2} \cdot socioecon$$

Suppose we calculate a predicted mortality rate, given a pollution level and aggregate education level, and we wish to explore effects of different levels each variable. Then:

$$\Delta\widehat{mortality} = \widehat{\beta_1} \cdot \Delta\, pollution + \widehat{\beta_2} \cdot \Delta\, socioecon$$

To "hold fixed" education means $\Delta\, socioecon$ = 0: $\dfrac{\Delta\widehat{mortality}}{\Delta pollution} = \widehat{\beta_1}$

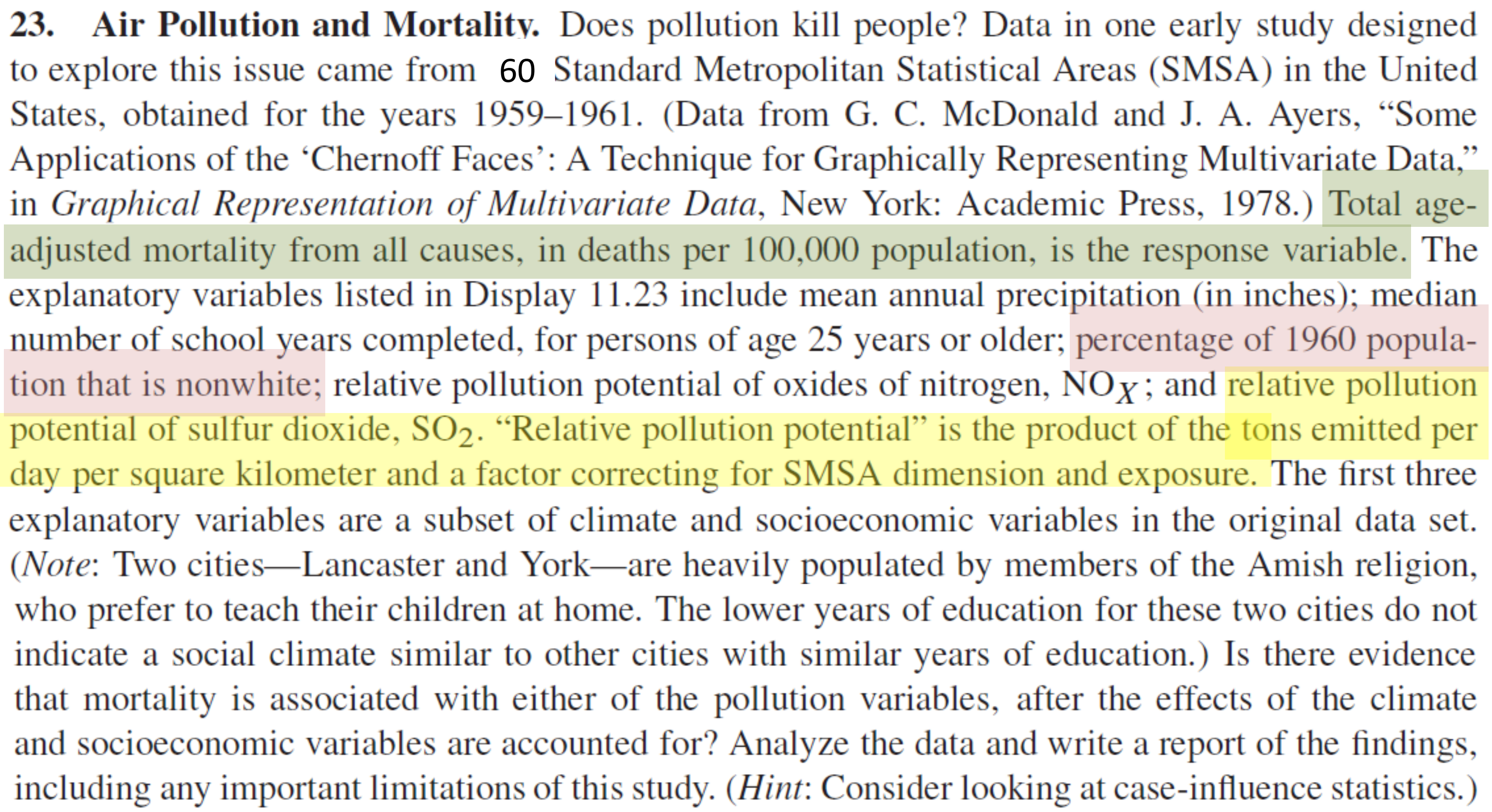
To "hold fixed" pollution means $\Delta\, pollution$ = 0 : $\dfrac{\Delta\widehat{mortality}}{\Delta socioecon} = \widehat{\beta_2}$

The estimates $\widehat{\beta_1}$ and $\widehat{\beta_2}$ are have **partial effect** interpretations. "The power of multiple regression analysis is that it provides this **ceteris paribus** interpretation even though the data have not been collected in a ceteris paribus fashion" (p. 76).

POLS 6481. Research Design and Quantitative Methods II
Lecture 5. Multiple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

4. Illustration: *Pollution Levels and Mortality*

Examine how mortality rates (deaths per 100k) depends on levels of *SO2* while controlling for a socioeconomic factor – the nonwhite percent of population



23. Air Pollution and Mortality. Does pollution kill people? Data in one early study designed to explore this issue came from 60 Standard Metropolitan Statistical Areas (SMSA) in the United States, obtained for the years 1959–1961. (Data from G. C. McDonald and J. A. Ayers, “Some Applications of the ‘Chernoff Faces’: A Technique for Graphically Representing Multivariate Data,” in *Graphical Representation of Multivariate Data*, New York: Academic Press, 1978.) Total age-adjusted mortality from all causes, in deaths per 100,000 population, is the response variable. The explanatory variables listed in Display 11.23 include mean annual precipitation (in inches); median number of school years completed, for persons of age 25 years or older; percentage of 1960 population that is nonwhite; relative pollution potential of oxides of nitrogen, NO_X; and relative pollution potential of sulfur dioxide, SO_2. “Relative pollution potential” is the product of the tons emitted per day per square kilometer and a factor correcting for SMSA dimension and exposure. The first three explanatory variables are a subset of climate and socioeconomic variables in the original data set. (*Note*: Two cities—Lancaster and York—are heavily populated by members of the Amish religion, who prefer to teach their children at home. The lower years of education for these two cities do not indicate a social climate similar to other cities with similar years of education.) Is there evidence that mortality is associated with either of the pollution variables, after the effects of the climate and socioeconomic variables are accounted for? Analyze the data and write a report of the findings, including any important limitations of this study. (*Hint*: Consider looking at case-influence statistics.)

POLS 6481. Research Design and Quantitative Methods II
Lecture 5. Multiple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

4. Illustration: *Pollution Levels and Mortality*

R script: Lecture 05 pollution.R
dataset: ex1123.csv

<u>Dependent variable</u>
Mort : deaths per 100,000 residents

<u>Explanatory variables</u>
SO2 : relative pollution potential of sulfur dioxide (tons per km^2)

Educ : median number of school years completed (for residents age 25 and older)

NonWhite : percentage of population that is nonwhite in 1960

Precip : average rainfall in inches

Sample is 60 cities in 1959-1961

Data found in Fred Ramsey and Daniel Schafer, *The Statistical Sleuth*, 3e

POLS 6481. Research Design and Quantitative Methods II
Lecture 5. Multiple Regression
Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

4. Illustration: *Pollution Levels and Mortality*

R script: Lecture 05 pollution.R
dataset: ex1123.csv

Dependent variable
Mort : deaths per 100,000 residents

Explanatory variables	Correlation with *Mort* :
SO2 : relative pollution potential of sulfur dioxide	.426
NonWhite : percentage of population that is nonwhite	.644
Precip : average rainfall	.509

Note: *SO2* & *NonWhite* correlate at $r = .159$
SO2 & *Precip* correlate at $r = –.107$

POLS 6481. Research Design and Quantitative Methods II

Lecture 5. Multiple Regression

Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

4. Illustration: *Pollution Levels and Mortality*

R script: Lecture 05 pollution.R

dataset: ex1123.csv

```
> data <- read.csv("C:/ex1123.csv"); names(data)
[1] "City" "Mort" "Precip" "Educ" "NonWhite" "NOX" "SO2"
```

Simple regression of *Mortality* on *SO2*

```
> simple <- lm(Mort ~ SO2, data)
> summary(simple)$coef
```

$$\widetilde{\beta_1} = \frac{cov(SO2,Mort)}{var(SO2)} = \frac{1680}{4018} = .418$$

Multiple regression of *Mortality* on *SO2* & *Nonwhite*

```
> multiple <- lm(Mort ~ SO2 + Nonwhite, data)
> summary(multiple)$coef
```

$$\widehat{\beta_1} = \frac{\frac{cov(SO2,Mort)}{var(SO2)} - \frac{cov(NonWhite,Mort)}{var(NonWhite)} \cdot \frac{cov(SO2,NonWhite)}{var(SO2)}}{1 - \frac{cov(SO2,NonWhite)^2}{var(SO2)\cdot var(NonWhite)}} = \frac{\frac{1680}{4018} - \frac{357}{80} \cdot \frac{90}{4018}}{1 - \frac{(90)^2}{(4018)\cdot(80)}}$$

$$= \frac{.418 - (4.463)\cdot(.0223)}{1 - .025} = \frac{.418 - .100}{.975}$$

$$= .326$$

```
> cov(data[c(2,5,7)])
              Mort   NonWhite        SO2
Mort     3869.0745 357.21555 1679.53476
NonWhite  357.2156  79.58688   90.08271
SO2      1679.5348  90.08271 4018.35141
```

POLS 6481. Research Design and Quantitative Methods II

Lecture 5. Multiple Regression

Readings: Wooldridge, *Introductory Econometrics 5e*, 1.4 + 3.1 + 3.2 + 3.3a

4. Illustration: *Pollution Levels and Mortality*

R script: Lecture 05 pollution.R

dataset: ex1123.csv

```
> data <- read.csv("C:/ex1123.csv"); names(data)
[1] "City" "Mort" "Precip" "Educ" "NonWhite" "NOX" "SO2"
```

Simple regression of *Mortality* on *SO2*

```
> simple <- lm(Mort ~ SO2, data)
> summary(simple)$coef
```

$$\widetilde{\beta_1} = \frac{cov(SO2,Mort)}{var(SO2)} = \frac{1680}{4018} = .418$$

Multiple regression of *Mortality* on *SO2* & *Precip*

```
> multiple <- lm(Mort ~ SO2 + Precip, data)
> summary(multiple)$coef
```

$$\widehat{\beta_1} = \frac{\frac{cov(SO2,Mort)}{var(SO2)} - \frac{cov(Precip,Mort)}{var(Precip)} \cdot \frac{cov(SO2,Precip)}{var(SO2)}}{1 - \frac{cov(SO2,Precip)^2}{var(SO2) \cdot var(Precip)}}$$

$$= \frac{\frac{1680}{4018} - \frac{316}{100} \cdot \frac{-68}{4018}}{1 - \frac{(-68)^2}{(4018) \cdot (100)}}$$

$$= \frac{.418 + (3.16) \cdot (.017)}{1 - .012} = \frac{.418 + .053}{.988}$$

$$= .477$$

```
> cov(data[c(2,3,7)])
              Mort    Precip        SO2
Mort     3869.0745 357.21555 1679.53476
Precip    316.4283  99.69379  -67.67571
SO2      1679.5348 -67.67571 4018.35141
```