# FIE401: Financial Econometrics

## Multiple Regression

Darya Yuferova

Basics of Multiple Regression

Measures of Fit

Hypothesis Testing

Regression Specification

# Basics of Multiple Regression

# Why Multiple Regression?

- We can incorporate more explanatory factors into the model.

# Example: Test Scores

Does the class size matter for educational output?

Are test scores completely determined by class size?

- testscr: test scores
- str: student-to-teacher ratio
- el_pct: proportion of English learners
- meal_pct: proportion of students receiving subsidized meal

# Multiple Regression Analysis

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + u$$

- $y$ = dependent variable, explained variable, response variable, left-hand side variable [LHS].
- $x$ = independent variable, explanatory variable, regressor, right-hand side variable [RHS].
- $u$ = error term, disturbance, unobservables, residuals.
- $\beta_0$ = intercept.
- for $j = 1, ..., K$, $\beta_j = \frac{\triangle y}{\triangle x_j}$ = slope coefficient = effect on $y$ of a change in $x_j$, holding everything else constant.

# Ordinary Least Squares (OLS) Estimator

- Our data: $(y_i, x_{1i}, x_{2i}, ..., x_{Ki})$ for $i = 1, ..., N$.
- OLS estimate is an estimate that minimizes the sum of squared residuals:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - ... - \hat{\beta}_K x_{Ki}$$
$$\min \Sigma_{i=1}^{N} \hat{u}_i^2 \rightarrow \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_K$$

# Example: Test Scores and Class Size

$$\widehat{testscr} = 698.93 - 2.28str$$

$$\widehat{testscr} = 686.03 - 1.10str - 0.65el\_pct$$

```
#load data
load("M:/Projects/FIE401/data/data_lectures/School.Rdata");
#run simple regression
fit1<-lm(testscr~str,data=School);
#summary of the simple regession
summary(fit1);
#run multiple regression
fit2<-lm(testscr~str+el_pct,data=School);
#summary of the multiple regession
summary(fit2);
```

# Example: Test Scores and Class Size

```
##
## Call:
## lm(formula = testscr ~ str, data = School)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

# Example: Test Scores and Class Size

```
##
## Call:
## lm(formula = testscr ~ str + el_pct, data = School)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.845 -10.240  -0.308   9.815  43.461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
## str          -1.10130    0.38028  -2.896  0.00398 **
## el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.46 on 417 degrees of freedom
## Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237
## F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

# Measures of Fit

# Goodness-of-Fit

- $R^2$ = fraction of variance of $y$ explained by $x_1, x_2, ..., x_K$.
- *Adjusted* $R^2 = 1 - \frac{SSR/(N-K-1)}{SST/N-1} = R^2$ with a degrees-of-freedom correction that adjusts for estimation uncertainty.
- *Adjusted* $R^2 < R^2$.
- The $R^2$ always increases when you add another regressor - a bit of a problem for a measure of "fit".
- The *Adjusted* $R^2$ corrects this problem by "penalizing" you for including another regressor.
- The *Adjusted* $R^2$ does not necessarily increase when you add another regressor.

# Example: Test Scores and Class Size

$$\widehat{testscr} = 698.93 - 2.28str$$

$$R^2 = 5.1\% \ Adjusted \ R^2 = 4.9\%$$

and

$$\widehat{testscr} = 686.03 - 1.10str - 0.65el\_pct$$

$$R^2 = 42.6\% \ Adjusted \ R^2 = 42.4\%$$

# Example: Test scores and class size

```
##
## Call:
## lm(formula = testscr ~ str + el_pct, data = School)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.845 -10.240  -0.308   9.815  43.461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
## str          -1.10130    0.38028  -2.896  0.00398 **
## el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.46 on 417 degrees of freedom
## Multiple R-squared:  0.4264, Adjusted R-squared:  0.4237
## F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

# Hypothesis Testing

# Hypothesis Tests for Single Coefficient

- Hypothesis tests and confidence intervals for a single coefficient in multiple regression follow the same logic and recipe as for the slope coefficient in a simple linear regression model.
- Thus, hypotheses on $\beta_1, \beta_2, ..., \beta_K$ can be tested using the usual $t$-statistic.

# Tests of Joint Hypotheses (i)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

- H0: $\beta_1 = \beta_2 = 0$
- H1: either $\beta_1 \neq 0$ or $\beta_2 \neq 0$ or both
- A joint hypothesis test imposes a restriction on two or more coefficients.
- In general, a joint hypothesis will involve $q$ restrictions.
- In the example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.
- A "common sense" idea is to reject if either of the individual $t$-statistics exceeds 1.96 in absolute value.
- But this "one at a time" test is not valid: the resulting test rejects too often under the null hypothesis (more than 5%)!

# Tests of Joint Hypotheses (ii)

- The size of a test is the actual rejection rate under the null hypothesis.
- The size of the "common sense" test is not 5%!
- In fact, its size depends on the correlation between $t_1$ and $t_2$.
- Solution: use a different test statistic designed to test both $\beta_1$ and $\beta_2$ at once: the $F$-statistic.

# $F$-statistic: General idea

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

- H0: $\beta_1 = \beta_2 = 0$.
- H1: either $\beta_1 \neq 0$ or $\beta_2 \neq 0$ or both.
- Run two regressions, one under the null hypothesis (the "restricted" regression) and one under the alternative hypothesis (the "unrestricted" regression).
- Restricted regression: $y = \beta_0 + \beta_3 x_3 + u$
- Unrestricted regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$
- Compare the fits of the regressions - the $R^2$s - if the "unrestricted" model fits sufficiently better, reject the null.

# $F$-statistic

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(N - K_{unrestricted} - 1)}$$

- $F$-statistic rejects when adding the two variables increased the $R^2$ by "enough" – that is, when adding the two variables improves the fit of the regression by "enough".
- $F$-statistic has a large-sample distribution that is $\chi^2_q$.

# Example: Test Scores and Class Size

$$\widehat{testscr} = 686.03 - 1.10 str - 0.65 el\_pct$$

$F$-statistic:

$$F = 155.01 \; df = 2 \;\; p\text{-value} < 2.2e-16$$

# Example: Test Scores and Class Size

```
#load necessary package
require(car);
#load data
load("M:/Projects/FIE401/data/data_lectures/School.Rdata");
#run multiple regression
fit2<-lm(testscr~str+el_pct,data=School);
#define H0
myH0<-c("str=0","el_pct=0");
#test
linearHypothesis(fit2,myH0);
```

# Example: Test Scores and Class Size

```
#define H0
myH0<-c("str=0","el_pct=0");
#test
linearHypothesis(fit2,myH0);
```

```
## Linear hypothesis test
##
## Hypothesis:
## str = 0
## el_pct = 0
##
## Model 1: restricted model
## Model 2: testscr ~ str + el_pct
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    419 152110
## 2    417  87245  2     64864 155.01 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Testing Single Restrictions on Multiple Coefficients

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

- H0: $\beta_1 = \beta_2$.
- H1: either $\beta_1 \neq \beta_2$.
- This null imposes a single restriction ($q = 1$) on multiple coefficients.
- It is not a joint hypothesis with multiple restrictions!

# Example: Test Scores and Class Size

$$\widehat{testscr} = 686.03 - 1.10 str - 0.65 el\_pct$$

$F$-statistic:

$$F = 1.3432 \; df = 1 \quad p\text{-value} = 0.2471$$

# Example: Test Scores and Class Size

```
#load necessary package
require(car);
#load data
load("M:/Projects/FIE401/data/data_lectures/School.Rdata");
#run multiple regression
fit2<-lm(testscr~str+el_pct,data=School);
#define H0
myH0<-c("str=el_pct");
#test
linearHypothesis(fit2,myH0);
```

# Example: Test Scores and Class Size

```
#define H0
myH0<-c("str=el_pct");
#test
linearHypothesis(fit2,myH0);
```

```
## Linear hypothesis test
##
## Hypothesis:
## str - el_pct = 0
##
## Model 1: restricted model
## Model 2: testscr ~ str + el_pct
##
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    418 87526
## 2    417 87245  1    281.04 1.3432 0.2471
```

# Regression Specification

# Regression Specification (i)

- We want to get an unbiased causal estimate of the effect on test scores of class size, holding constant other factors – percentage of English learners, ability, etc.
- If we could run an experiment, we would randomly assign students to classes of different size.
- Then test scores would be dependent only on the class sise and would be independent of all other things.
- But with observational data, test scores depend on these additional factors!

# Regression Specification (ii)

- If you can observe those factors (e.g., percentage of English learners), then include them in the regression.
- But often you cannot observe all these omitted causal factors (e.g., ability, help from the parents, etc.). In this case, you can include "control variables" which are correlated with these omitted causal factors, but which themselves are not causal.
- An effective control variable is one which, when included in the regression, makes the variable of interest as if it was randomly assigned.

# Example: Test Scores and Class size

$$\widehat{testscr} = 700.15 - 1.00str - 0.12el\_pct - 0.55meal\_pct$$

```
#load data
load("M:/Projects/FIE401/data/data_lectures/School.Rdata");
#run multiple regression
fit4<-lm(testscr~str+el_pct+meal_pct,data=School);
#summary of multiple regression
summary(fit4);
```

# Example: Test Scores and Class Size

```
##
## Call:
## lm(formula = testscr ~ str + el_pct + meal_pct, data = School)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.849  -5.151  -0.308   5.243  31.501
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 700.14996    4.68569 149.423  < 2e-16 ***
## str          -0.99831    0.23875  -4.181 3.54e-05 ***
## el_pct       -0.12157    0.03232  -3.762 0.000193 ***
## meal_pct     -0.54735    0.02160 -25.341  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.08 on 416 degrees of freedom
## Multiple R-squared:  0.7745, Adjusted R-squared:  0.7729
## F-statistic: 476.3 on 3 and 416 DF,  p-value: < 2.2e-16
```

# Example: Test Scores and Class Size

- How should we interpret coefficient in front of proportion of students receiving subsidized meal?
- What proportion of students receiving subsidized meal is proxing for?

# Implications for Variable Selection (i)

- Identify the variable of interest
- Think of the omitted causal effects that could result in omitted variable bias
- Include those omitted causal effects if you can or, if you cannot, include variables correlated with them that serve as control variables.
- This results in a "base" or "benchmark" model.

# Implications for Variable Selection (ii)

- Also specify a range of plausible alternative models, which include additional candidate variables.
- Estimate your base model and plausible alternative specifications (robustness checks).
- Does a candidate variable change the coefficient of interest ($\beta_1$)?
- Is a candidate variable statistically significant?
- Use judgment, not a mechanical recipe!
- Do not just try to maximize $R^2$!

# Digression about Measures of Fit

- It is easy to fall into the trap of maximizing the $R^2$ and *Adjusted* $R^2$, but this loses sight of our real objective, an unbiased estimator.
- A high $R^2$ (*Adjusted* $R^2$) means that the regressors explain the a lot of variation in $y$.
- A high $R^2$ (*Adjusted* $R^2$) does not mean that you have an unbiased estimator of a causal effect ($\beta_1$).
- A high $R^2$ (*Adjusted* $R^2$) does not mean that the included variables are statistically significant – this must be determined using hypotheses tests.

# Textbook

- Multiple regression: Chapters 6 and 7, Stock and Watson, Introduction to Econometrics, Global Edition, 4th edition