ECON 340
Economic Research Methods

Div Bhagia

Final Exam Review

Final Exam

- Thursday, Dec 14 at 1 pm.
- 90 minutes, 20 points
- Closed book, can use a calculator
- No formula sheet
- Not cumulative
- Study guide and sample exam
- Sample questions for last module

Topics Covered

Linear Regression Model (75-80%)

- Ordinary Least Squares & Goodness of Fit
- OLS Assumptions for Causal Inference
- Inference (p-values, t-stats, confidence intervals)
- Multiple Regression: Omitted variable bias, $Adjusted R^2$
- Categorical variables, interaction terms
- Quadratic and Log Functional Forms

Additional Topics (20-25%)

- Experiments & Quasi-experimental methods
- Differences-in-Differences
- Big Data & Machine Learning

Linear Regression Model

Start by assuming a linear relationship between X and Y:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad E(u_i) = 0$$

- Estimate using Ordinary Least Squares (OLS) method, which minimizes the sum of squared errors

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- Under the exogeneity assumption, $E(u|X) = E(u) = 0$, can interpret β_1 as the causal impact of X on Y

Test Scores and Class Size

```
=========================================
                  Dependent variable:
             ----------------------------
                       testscr
-----------------------------------------
str                   -2.280***
                       (0.480)

Constant              698.933***
                       (9.467)

-----------------------------------------
Observations             420
R2                      0.051
Adjusted R2             0.049
=========================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

- Predicted values/residuals from the fitted line:

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

- Interpret the output
    - Coefficients
    - Statistical significance (t-stats, p-values)
    - R^2

- Exogeneity assumption:

$$E(u_i|STR_i) = E(u_i) = 0$$

Omitted Variable Bias

Consider the following linear regression model:

$$Y = \beta_0 + \beta_1 X + u$$

- Here, u captures omitted factors that impact Y.
- If u is correlated with X, the exogeneity assumption fails and OLS estimates are biased.

$$\hat{\beta}_1 = \beta_1 + \frac{Cov(X, u)}{Var(X)}$$

- Strength and direction of bias depends on $Cov(X, u)$

Omitted Variable Bias

$$Y = \beta_0 + \beta_1 X + u$$

Note that omitted variable bias only occurs when both of the following are true:

(1) The omitted variable is correlated with X

(2) The omitted variable $\rightarrow$ Y

Omitted Variable Bias

In our example:

$$testscr = \beta_0 + \beta_1 str + u$$

Omitting $comp\_stu$ from this model will probably overestimate the impact of str.

This is because we expect $comp\_stu$ to positively impact $testscr$ and $Cov(comp\_stu, str) < 0$.

So $comp\_stu$ being omitted leads to $Cov(str, u) < 0$, hence from the OVB formula $\hat{\beta}_1 < \beta_1$.

Test Scores and Class Size

	Dependent variable:	
	testscr	
	(1)	(2)
str	-2.280***	-1.593***
	(0.480)	(0.493)
comp_stu		65.160***
		(14.351)
Observations	420	420
R2	0.051	0.096
Adjusted R2	0.049	0.092

Multiple Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

- Assumptions: (1) random sample, (2) no large outliers, (3) no perfect multicollinearity, (4) $E(u|X_1, X_2) = 0$
- Under these assumptions, β_1 captures the causal effect of X_1 keeping X_2 constant, and β_2 captures the causal effect of X_2 keeping X_1 constant.

Control Variables

- While there are cases where we might want to evaluate the effect of both the variables, it is hard to find exogenous variables
- A really good use of the multiple regression model is to instead *control* for omitted variable W while trying to estimate the causal effect of X

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

Control Variables

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

- So instead of assumption (4), we can assume *conditional mean independence*

$$E(u|X, W) = E(u|W)$$

- The idea is that once you control for the W, X becomes independent of u
- Under this modified assumption, we can interpret β_1 as the causal effect of X while *controlling* for W

Adjusted R^2

R^2 never decreases when an explanatory variable is added

An alternative measure called Adjusted R^2

$$AdjustedR^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

where k is the number of variables.

$AdjustedR^2$ only rises if RSS declines by a larger percentage than the degrees of freedom $(n-k-1)$.

Dummy Variables

What if the independent variable is a binary variable that takes two values 1 and 0?

$$Y = \beta_0 + \beta_1 D + u$$

Taking conditional expectation (assuming exogeneity):

$$E[Y|D=1] = \beta_0 + \beta_1 \cdot 1 = \beta_0 + \beta_1$$
$$E[Y|D=0] = \beta_0 + \beta_1 \cdot 0 = \beta_0$$

So,

$$\beta_1 = E[Y|D=1] - E[Y|D=0]$$

ACS Data: Gender Wage Gap

	Wages
Intercept	67,220.17***
	(439.87)
Female	−14,661.12***
	(637.27)
Observations	17,578
R^2	0.03
Note:	*p<0.1; **p<0.05; ***p<0.01

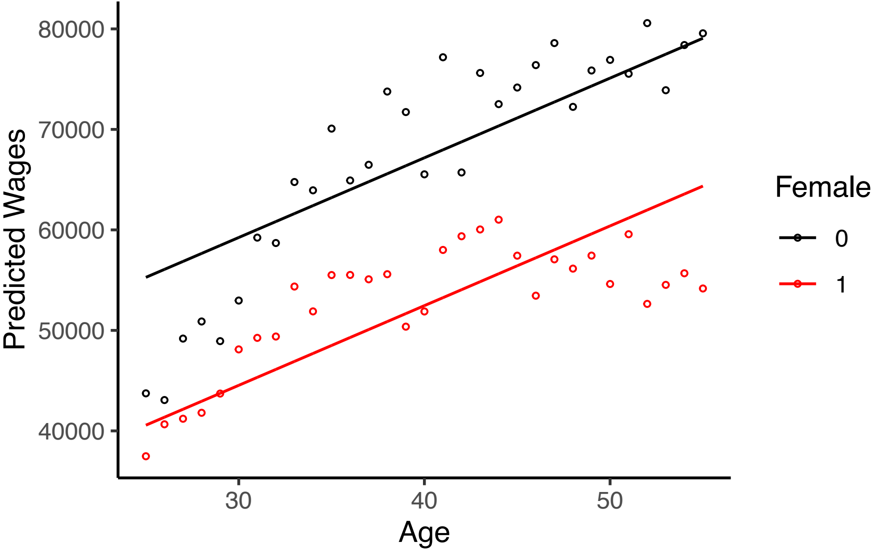
Dummy Variables in Multiple Regression

$$Wages = \beta_0 + \beta_1 Age + \beta_2 Female + u$$

Taking conditional expectation (assuming exogeneity):

$$E[Wages|Age, Female = 1] = (\beta_0 + \beta_2) + \beta_1 Age$$
$$E[Wages|Age, Female = 0] = \beta_0 + \beta_1 Age$$

ACS Data: Wages and Age



Interaction Terms

We can also include interaction terms in our model as follows:

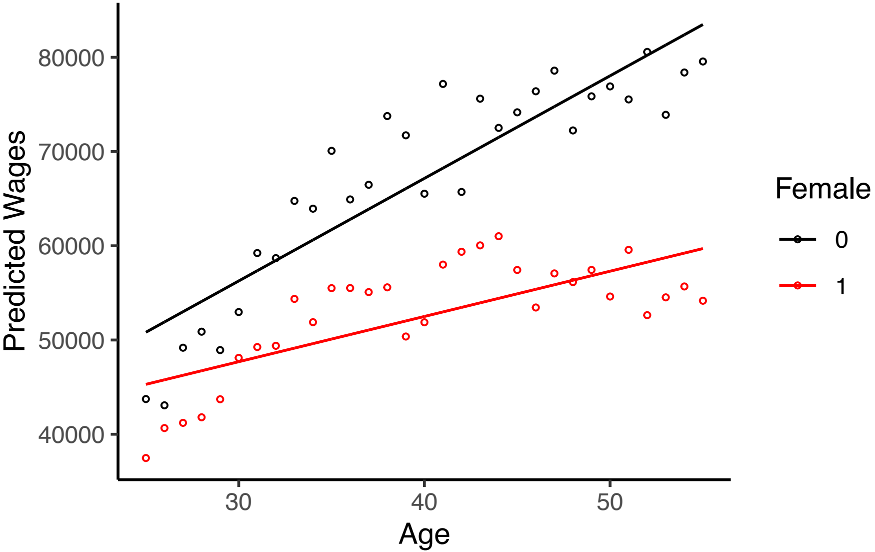
$$Wages = \beta_0 + \beta_1 Age + \beta_2 Female + \beta_3 Female \times Age + u$$

Taking conditional expectation (assuming exogeneity):

$$E[Wages|Age, Female = 1] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)Age$$
$$E[Wages|Age, Female = 0] = \beta_0 + \beta_1 Age$$

Now the impact of X on Y varies with D.

ACS Data: Wages and Age



Interaction of Two Dummy Variables

$$wages = \beta_0 + \beta_1 Female + \beta_2 Hispanic + \beta_3 Female \times Hispanic + u$$

Average wages for Non-Hispanic Males:

$$E(wages|Hispanic = 0, Female = 0) = \beta_0$$

Average wages for Non-Hispanic Females:

$$E(wages|Hispanic = 0, Female = 1) = \beta_0 + {\color{red}\beta_1}$$

Interaction of Two Dummy Variables

$$wages = \beta_0 + \beta_1 Female + \beta_2 Hispanic + \beta_3 Female \times Hispanic + u$$

Average wages for Hispanic Males:

$$E(wages|Hispanic = 1, Female = 0) = \beta_0 + \beta_2$$

Average wages for Hispanic Females:

$$E(wages|Hispanic = 1, Female = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

ACS Data: Gender and Ethnicity

	Wages
Intercept	70,179.09***
	(473.52)
Female	−16,046.81***
	(683.42)
Hispanic	−19,367.71***
	(1,211.46)
Female X Hispanic	8,163.75***
	(1,788.04)
Observations	17,578

Fitting a Line

Linear relationship:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Take the derivative:

$$\frac{d\hat{Y}}{dX} = \hat{\beta}_1 \rightarrow d\hat{Y} = \hat{\beta}_1 dX$$

Can think of d as 'change in': One unit change in X, associated with β_1 units change in Y.

Impact of X on Y constant with X.

Fitting a Curve

Quadratic relationship:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

Take the derivative:

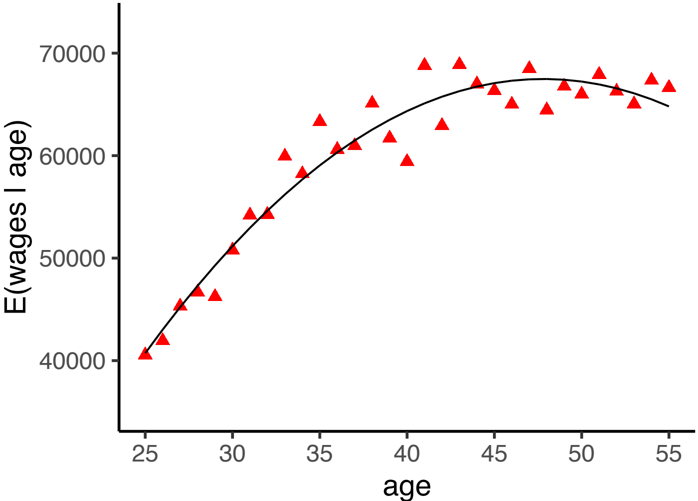
$$\frac{d\hat{Y}}{dX} = \hat{\beta}_1 + 2\hat{\beta}_2 X$$

Now the impact of X on Y changes with X.

Remember: Derivative captures the slope of the tangent line.

ACS Data: Wages and Age

$$\hat{wage} = -52207 + 4775.64 \cdot age - 49.493 \cdot age^2$$



Log Functional Forms

- Log-transformation leads to interpretation of regression coefficients in % changes than unit changes which can sometimes be more informative

- Can think of change in log of X as the relative change in X with respect to its original value

$$\frac{d}{dX}\log(X) = \frac{1}{X} \rightarrow d\log(X) = \frac{dX}{X}$$

In which case $100 \times d\log(X)$ represents % change in X

Log Functional Forms: Interpretation

Three possible models:

1. Level-Log: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 log(X)$
2. Log-Level: $\hat{log}(Y) = \hat{\beta}_0 + \hat{\beta}_1 X$
3. Log-Log: $\log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 \log(X)$

Log-Level Model

	Log Wages
Intercept	10.31***
	(0.02)
Age	0.01***
	(0.001)
Observations	17,578
R^2	0.03
Note:	*p<0.1; **p<0.05; ***p<0.01

1 year increase in age leads to 1% increase in predicted wages.

Log-Log Model

	Log Wages
Intercept	8.99***
	(0.08)
Log Age	0.49***
	(0.02)
Observations	17,578
R^2	0.03
Note:	*p<0.1; **p<0.05; ***p<0.01

1% increase in age leads to 0.49% increase in predicted wages.

A Few Last Words

Good luck and take care!

Thanks for a great semester!

Have a great break, and don't be a stranger!