ECON3150/4150 Spring 2016

Lecture 5
Multiple regression model

Siv-Elisabeth Skjelbred

University of Oslo

February 1st
Last updated: January 29, 2016

Outline

- Continue from slide 34 on lecture 4.
- Regressions when X is a binary variable
- Omitted variable bias
- Introduction to multiple linear regression model and OLS

Reminder

Interpretation and prediction:

```
                                                      Statistics/Data Analysis

  1 . reg ahe age

        Source |       SS       df       MS              Number of obs =    7711
  -------------+------------------------------           F(  1,  7709) =  230.43
         Model |  23005.7375     1  23005.7375           Prob > F      =  0.0000
      Residual |  769645.718  7709  99.8372964           R-squared     =  0.0290
  -------------+------------------------------           Adj R-squared =  0.0289
         Total |  792651.456  7710   102.80823           Root MSE      =  9.9919

  ------------------------------------------------------------------------------
           ahe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
           age |   .6049863   .0398542    15.18   0.000     .5268613    .6831113
         _cons |   1.082275   1.184255     0.91   0.361    -1.239187    3.403737
  ------------------------------------------------------------------------------
```

The regression result gives:

$$\hat{Y} = 1.08 + 0.60 age$$

Predictions:

- A 26 year old worker is predicted to have an average hourly wage of: \$ 16.68 (1.08+0.6*26).
- For each year of education you are predicted to earn \$ 0.6 more.

Regression when X is a binary variable

- A lot of information relevant for econometric analysis is qualitative.
- This information can be summarized with one or multiple binary variables.
- In econometrics binary variables are typically called dummy variables.
- In defining a dummy variable we must decide which event is assigned the value one and which is assigned the value 0.
- The name typically indicates the event with value one.
    - Female (1=female, 0=male)
    - Higher_educ (1=college or more, 0=less than college)
    - Public_transport (1=use public transport to work, 0=do not use public transport)
    - Drug (1=received the drug, 0= received placebo)

Regression when X is a binary variable

The population regression model with the binary variable D_i (D=1 if female, D=0 if male) is:

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

when i is a male (D=0) we get:

$$Y_i = \beta_0 + u_i \rightarrow E(Y_i|D = 0) = \beta_0$$

while if i is a female (D=1) we get:

$$Y_i = \beta_0 + \beta_1 + u_i \rightarrow E(Y_i|D = 1) = \beta_0 + \beta_1$$

Thus $\beta_1 = E(Y_i|Female) - E(Y_i|male)$

Dummy variables

- The group with an indicator of 0 is the base group, the group against which comparisons are made.
- It does not matter how we choose the base group, but it is important to keep track of which group is the base group.
- If the two groups do not differ then β_1 is zero.

Example

Data from additional E4.1

- Data from on average hourly earnings from a sample of full-time workers.
- Female $= 1$ the person is female, female $= 0$ the person is male.

```
1 . reg ahe female

      Source |       SS       df       MS              Number of obs =    7711
-------------+------------------------------           F(  1,  7709) =  129.46
       Model |  13091.0876     1  13091.0876           Prob > F      =  0.0000
    Residual |  779560.368  7709   101.12341           R-squared     =  0.0165
-------------+------------------------------           Adj R-squared =  0.0164
       Total |  792651.456  7710   102.80823           Root MSE      =  10.056

------------------------------------------------------------------------------
         ahe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -2.629912   .2311422   -11.38   0.000    -3.083013    -2.17681
       _cons |   20.11387   .1520326   132.30   0.000     19.81584    20.41189
------------------------------------------------------------------------------
```

Proportions and percentages as dependent variables

- The **proportional change** is the change in a variable relative to its initial value, mathematically, the change divided by the initial value.
- The **percentage change** is the proportional change in a variable, multiplied by 100.
- The **percentage point change** is the difference between two percentages.

Proportions and percentages as dependent variables

In a dataset on CEO's where y is annual salary in thousands of dollars and X is the average return on equity (roe) the following OLS regression line can be obtained:

$$salary = \beta_0 + \beta_1 roe + u$$

- ROE is defined in terms of net income as a percentage of common equity, thus if roe=10, the average return on equity is 10%.
- The slope parameter β_1 measures the change in annual salary, in thousands of dollars, when return on equity increase by one percentage point.

Homoskedasticity

The dummy variable example can shed light on what is meant by homoskedasticity:

- The definition of homoskedasticity requires the error term to be independent of X, i.e it must not depend on female in our example.
- For women the error term (u_i) is the deviation of the i^{th} woman's earning from the population mean earnings for women.
- For men the error term (u_i) is the deviation of the i^{th} man's earning from the population mean earnings for men.
- Thus the variance of earnings must be the same for men as it is for women.

The ideal analysis

- The aim of regression is often to identify causality.
- In an ideal randomized controlled experiment the only difference between the "treatment" and "control" group is the variable you study.
- In observational data there may be a systematic difference the "treatment" group and the "control group" in one or more variables.
- If those variables are not included in the regression we have omitted variables.

Omitted variable bias -ZCM assumption

- In the last lecture you saw that $E(u|X) = 0$ is important in order for the OLS estimator to be unbiased.
- The omitted variable is thus important if the omission leads to a violation of the ZCM assumption.
- The bias that arise from such an omission is called omitted variable bias.

Omitted variable bias

Omitted variable bias

For omitted variable bias to occur, the omitted variable "Z" must satisfy two conditions:

- The omitted variable is correlated with the included regressor (i.e. $corr(Z, X) \neq 0$)
- The omitted variable is a determinant of the dependent variable (i.e. Z is part of u)

OVB example

We estimate:

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

while the true model is:

$$y_i = \beta_0 + \beta_1 X + \beta_2 Z + v$$

The exclusion of Z leads to a bias in β_1 whenever Z is a determinant of Y and correlated with X.

Example: $Corr(Z, X) \neq 0$

The omitted variable (Z) is correlated with X, example

$$\text{wages} = \beta_0 + \beta_1 \text{educ} + \underbrace{u_i}_{\delta_1 pinc + v_i}$$

- Parents income is likely to be correlated with education, college is expensive and the alternative funding is loan or scholarship which is harder to acquire.

Example: Z is a determinant of Y

The omitted variable is a determinant of the dependent variable,

$$\text{wages} = \beta_0 + \beta_1 \text{educ} + \underbrace{u_i}_{\delta_2 MS + v_i}$$

- Market situation is likely to determine wages, workers in firms that are doing well are likely to have higher wages.

Example: Omitted variable bias

The omitted variable is both determinant of the dependent variable, i.e. $corr(X_2, Y) \neq 0$ and correlated with the included regressor

$$\text{wages} = \beta_0 + \beta_1 \text{educ} + \underbrace{u_i}_{\delta_3 ability + v_i}$$

- Ability - the higher your ability the "easier" education is for you and the more likely you are to have high education.
- Ability - the higher your ability the better you are at your job and the higher wages you get.

How to overcome omitted variable bias

1. Run a ideal randomized controlled experiment
2. Do cross tabulation
3. Include the omitted variable in the regression

Cross tabulation

One can address omitted variable bias by splitting the data into subgroups. For example:

	College graduates	High school graduates
High family income	$\bar{Y}_{HFI,C}$	$\bar{Y}_{HFI,H}$
Medium family income	$\bar{Y}_{MFI,C}$	$\bar{Y}_{MFI,H}$
Low family income	$\bar{Y}_{LFI,C}$	$\bar{Y}_{LFI,H}$

Cross tabulation

- Cross tabulation only provides a difference of means analysis, but it does not provide a useful estimate of the ceteris paribus effect.
- To quantify the partial effect on Y_i on the change in one variable (X_{1i}) holding the other independent variables constant we need to include the variables we want to hold constant in the model.
- When dealing with multiple independent variables we need the multiple linear regression model.

Multiple linear regression model

Multiple linear regression model

- Multiple linear regression models contain more than one independent variable.
- Multiple variables is necessary if:
  - You are interested in the ceteris paribus effect of multiple parameters.
  - Y is a polynomial function of X (more in chapter 8)
  - You fear violation omitted variable bias.

Y	X	Other variables
Wages	Education	Experience, Ability
Crop Yield	Fertilizer	Soil quality, location (sun etc)
Test score	STR	Average family income

Multiple linear regression model

The general multiple linear regression model for the population can be written in the as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

- Where the subscript i indicates the i^{th} of the n observations in the sample.
- The first subscript, 1,2,...,k, denotes the independent variable number.
- The intercept β_0 is the expected value of Y when all the X's equal zero.
- The intercept can be thought of as the coefficient on a regressor, X_{0i}, that equals one for all i.
- The coefficient β_1 is the coefficient of X_{1i}, β_2 the coefficient on X_{2i} etc.

Multiple linear regression model

The average relationship between the k independent variables and the dependent variable is given by:

$$E(Y_i|X_{1i} = x_1, X2i = x_2, ..., X_{ki} = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

- β_1 is thus the effect on Y of a unit change in X_1 holding all other independent variables constant.
- The error term includes all other factors than the X's that influence Y.

Example

To make it more tractable consider a model with two independent variables. Then the population model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u$$

Example:

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exp_i + u_i$$

$$wage_i = \beta_0 + \beta_1 exp_i + \beta_2 IQ_i^2 + u_i$$

Interpretation of the coefficient

In the two variable case the predicted value is given by:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Thus the predicted change in y given the changes in X_1 and X_2 are given by:

$$\Delta \hat{Y} = \hat{\beta}_1 \Delta X_1 + \hat{\beta}_2 \Delta X_2$$

Thus if x_2 is held fixed then:

$$\Delta \hat{Y} = \hat{\beta}_1 \Delta X_1$$

Interpretation of the coefficient

Using data on 526 observations on wage, education and experience the following output was obtained:

```
1 . reg wage educ exper

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   75.99
       Model |  1612.2545     2  806.127251            Prob > F      =  0.0000
    Residual | 5548.15979   523  10.6083361            R-squared     =  0.2252
-------------+------------------------------           Adj R-squared =  0.2222
       Total | 7160.41429   525  13.6388844            Root MSE      =   3.257

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6442721   .0538061    11.97   0.000     .5385695    .7499747
       exper |   .0700954   .0109776     6.39   0.000     .0485297    .0916611
       _cons |  -3.390539   .7665661    -4.42   0.000    -4.896466   -1.884613
------------------------------------------------------------------------------
```

Holding experience fixed another year of education is predicted to increase your wage by 0.64 dollars.

Interpretation of the coefficient

If we want to change more than one independent variable we simply add the two effects.

Example:

$$\hat{wage} = -3.39 + 0.64educ + 0.07exp$$

If you increase education by one year and decrease experience by one year the predicted increase in wage is 0.57 dollars. (0.64-0.07)

Example: Smoking and birthweight

Using the data set birthweight_smoking.dta you can estimate the following regression:

$$\widehat{birthweight} = 3432.06 - 253.2\text{Smoker}$$

If we include the number of prenatal visits:

$$\widehat{birthweight} = 3050.5 - 218.8\text{Smoker} + 34.1\text{nprevist}$$

Example education

The relationship between years of education of male workers and the years of education of the parents.

```
8 . reg educ meduc feduc, robust

Linear regression                               Number of obs =       1129
                                                    F(  2,  1126) =     159.83
                                                    Prob > F      =     0.0000
                                                    R-squared     =     0.2689
                                                    Root MSE      =     2.2595

------------------------------------------------------------------------------
             |               Robust
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       meduc |   .1844065   .0223369     8.26   0.000     .1405798    .2282332
       feduc |   .2208784   .0259207     8.52   0.000     .1700201    .2717368
       _cons |   8.860898   .2352065    37.67   0.000     8.399405     9.32239
------------------------------------------------------------------------------
```

- Interpret the coefficient on mother's education.
- What is the predicted difference in education for a person where both parents have 12 years of education and a person where both parents have 16 years of education?

Example education

From stata:

```
. display _cons+_b[meduc]*12+_b[feduc]*12
5.8634189

. display _cons+_b[meduc]*16+_b[feduc]*16
7.4845585

.
. display 7.484-5.863
1.621

.
. *or
.
. display _b[meduc]*4+_b[feduc]*4
1.6211396
```

Or by hand:

$$0.1844 * (16 - 12) + 0.2209 * (16 - 12) = 1.6212$$

Multiple linear regression model

Advantages of the MLRM over the SLRM:

- By adding more independent variables (control variables) we can explicitly control for other factors affecting y.
- More likely that the zero conditional mean assumption holds and thus more likely to have an unbiased estimator.
- By controlling for more factors, we can explain more of the variation in y, thus better predictions.
- Can incorporate more general functional forms.

Assumptions of the MLRM

1. (The model is linear in parameters)
2. Random sampling
3. Large outliers are unlikely
4. Zero conditional mean, i.e the error u has an expected value of zero given any value of the independent variables

$$E(u|X_1, x_2, ...X_k) = 0$$

5. (There is sampling variation in X) **and there are no exact linear relationships among the independent variables.**

Under these assumptions the OLS estimators are unbiased estimators of the population parameters. In addition there is the homoskedasticity assumption which is necessary for OLS to be BLUE.

No exact linear relationships

Perfect collinearity

A situation in which one of the regressors is an exact linear function of the other regressors.

- This is required to be able to compute the estimators.
- The variables can be correlated, but not perfectly correlated.
- Typically perfect collinearity arise because of specification mistakes.
  - Mistakenly put in the same variable measured in different units
  - The dummy variable trap: Including the intercept plus a binary variable for each group.
  - Sample size is to small compared to parameters (need at least k+1 observations to estimate k+1 parameters)

No perfect collinearity

Solving the three 1oc for the model with two independent variables gives:

$$\hat{\beta}_1 = \frac{\hat{\sigma}^2_{X_2}\hat{\sigma}_{Y,X_1} - \hat{\sigma}_{Y,X_2}\hat{\sigma}_{X_1,X_2}}{\hat{\sigma}^2_{X_1}\hat{\sigma}^2_{X_2} - \hat{\sigma}_{X_1,X_2}}$$

where $\hat{\sigma}^2_{X_j}$ $(j = 1, 2)$, $\hat{\sigma}^2_{Y,X_j}$ and $\hat{\sigma}^2_{X_1,X_2}$ are empirical variances and covariances. Thus we require that:

$$\hat{\sigma}^2_{X_1}\hat{\sigma}^2_{X_2} - \hat{\sigma}_{X_1,X_2} = \hat{\sigma}^2_{X_1}\hat{\sigma}^2_{X_2}(1 - r^2_{X_1,X_2}) \neq 0$$

Thus must have that $\hat{\sigma}^2_{X_1} > 0$, $\hat{\sigma}^2_{X_2} > 0$ and $r^2_{X_1,X_2} \neq 1$. Thus the sample correlation coefficient between X_1 and X_2 cannot be one or minus one.

Imperfect collinearity

- Occurs when two or more of the regressors are highly correlated (but not perfectly correlated).
- High correlation makes it hard to estimate the effect of the one variable holding the other constant.
- For the model with two independent variables and homoskedastic errors:

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\left(\frac{1}{1-\rho^2_{X_1,X_2}}\right)\frac{\sigma^2_u}{\sigma^2_{X_1}}$$

- The two variable case illustrates that the higher the correlation between X_1 and X_2 the higher the variance of $\hat{\beta}_1$.
- Thus, when multiple regressors are imperfectly collinear, the coefficients on one or more of these regressors will be imprecisely estimated.

Omitted variable bias

The direction of bias is illustrated in the the following formula:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X} \tag{1}$$

where $\rho_{Xu} = corr(X_i, u_i)$. The formula indicates that:

- Omitted variable bias exist even when n is large.
- The larger the correlation between X and the error term the larger the bias.
- The direction of the bias depends on whether X and u are negatively or positively correlated.

Example bias

Comparing estimates from simple and multiple regression. What is the return to education? Simple regression:

```
1 . reg wage educ, robust

  Linear regression                                Number of obs =        935
                                                      F(  1,   933) =     95.65
                                                      Prob > F      =    0.0000
                                                      R-squared     =    0.1070
                                                      Root MSE      =    382.32
```

wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	60.21428	6.156956	9.78	0.000	48.1312	72.29737
_cons	146.9524	80.26953	1.83	0.067	-10.57731	304.4822

Can we give this regression a causal interpretation? What happens if we include IQ in the regression?

forth

Example bias - two independent variables

Call the simple regression of Y on X_1 (think of regressing wage on education)

$$\tilde{Y} = \tilde{\beta_0} + \tilde{\beta_1} X_1 + v$$

while the true population model is:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$$

The relationship between $\tilde{\beta_1}$ and β_1 is:

$$\tilde{\beta_1} = \beta_1 + \beta_2 \tilde{\delta}_1$$

where $\tilde{\delta}_1$ comes from the regression $\hat{X}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 X_1$

Example bias - two independent variables

Thus the bias that arise from the omitted variable (in the model with two independent variables) is given by $\beta_2\tilde{\delta}_1$ and the direction of the bias can be summarized by the following table:

	$corr(x_1, x_2) > 0$	$corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Comparing estimates from simple and multiple regression

wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	42.05762	6.810074	6.18	0.000	28.69276	55.42247
IQ	5.137958	.9266458	5.54	0.000	3.319404	6.956512
_cons	-128.8899	93.09396	-1.38	0.167	-311.5879	53.80818

IQ	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	3.533829	.1839282	19.21	0.000	3.172868	3.89479
_cons	53.68715	2.545285	21.09	0.000	48.69201	58.6823

$$\tilde{\beta}_1 = 60.214 \approx 42.047 + 3.533 * 5.137$$

back

Bias - multiple independent variables

- Deriving the sign of omitted variable bias when there are more than two independent variables in the model is more difficult.
- Note that correlation between a single explanatory variable and the error generally results in all OLS estimators being biased.
- Suppose the true population model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

- But we estimate

$$\tilde{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_2$$

- If $Corr(X_1, X_3) \neq 0$ while $Corr(X_2, X_3) = 0$ $\tilde{\beta}_2$ will also be biased unless $corr(X_1, X_2) = 0$.

Bias - multiple independent variables

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

- People with higher ability tend to have higher education
- People with higher education tend to have less experience
- Even if we assume that ability and experience are uncorrelated β_2 is biased.
- We cannot conclude the direction of bias without further assumptions

Causation

- Regression analysis can refute a causal relationship, since correlation is necessary for causation..
- ..but cannot confirm or discover a causal relationship by statistical analysis alone.
- The true population parameter measures the ceteris paribus effect which holds all other (relevant) factors equal.
- However, it is rarely possible to literally hold all else equal:
  - "natural experiments" or "quasi-experiments".
  - Use instrument on unobserved factors.