Linear Regression with Multiple Regressions

SW Chapter 6

Multiple regression analysis

Omitted variable bias

Measures of fit

Least squares assumptions

Learning objectives

- Just go to town on some multiple linear regression - implementing and interpreting
- Deepen our understanding of omitted variable bias
- Calculate and interpret a new measure of fit, the adjusted R^2
- Update our knowledge of least square assumption and the sampling distribution of the OLS estimator in the case of multiple independent variables

Multiple regression analysis

Definition of the multiple linear regression model

Explains y in terms of variables $x_1, x_2, \ldots, x_k$

Intercept

Slope parameters

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Dependent variable,
explained variable,
response variable,...

Independent variables,
explanatory variables,
regressors,...

Error term,
disturbance,
unobservables,...

Motivation for multiple linear regression

- Incorporate more explanatory factors into the model
- Explicitly hold fixed other factors that otherwise would be in
- Allow for more flexible functional forms

Example: Wage equation

Now measures effect of education <u>explicitly holding experience fixed</u>

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

All other factors...

Hourly wage

Years of education

Years of labor market experience

Example: Average test scores and per student spending

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u$$

Other factors

Average standardized test score of school

Per student spending at this school

Average family income of students at this school

Why would we include average family income in this regression?

Example: Family income and family consumption

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$$

Other factors

Family consumption

Family income

Family income <u>squared</u>

- Why would we include average family income in this regression?
- Model has two explanatory variables: income and income squared
- Consumption is explained as a quadratic function of income
- Be careful when interpreting the coefficients!

$$\frac{\Delta cons}{\Delta inc} \approx \beta_1 + 2\beta_2 inc \quad (1)$$

OLS estimation of multiple regression model

Same idea: minimize sum of squared residuals

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - ... - \hat{\beta}_k xik \tag{2}$$

$$min \sum_{i=1}^{n} \hat{u_i}^2 \tag{3}$$

We will use statistical packages to carry out this calculation

Interpretation of the OLS model

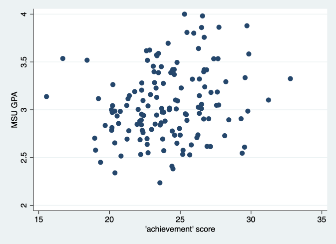
$$\beta_j = \frac{\Delta y}{\Delta x_j} \tag{4}$$

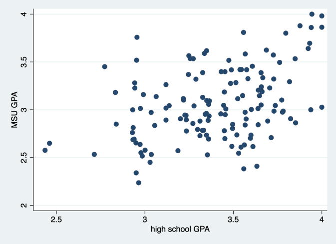
β_j is how much the dependent variable changes if the jth independent variable changes, **holding constant** (or **controlling for**) all other independent variables

- The multiple linear regression model holds the values of other explanatory variables fixed even if they are correlated with the other variables (*ceteris paribus*)
- β_j is the **partial effect** of X_j on Y, holding all other variable fixed.
- Still assume unobserved factors do not change if the explanatory variables are changed

Example: Determinants of college GPA

Let's look at the relationship between high school and college GPA, controlling for test scores.





What predicts college GPA?

Source: Christopher Lemmon, who surveyed 194 MSU students, in Fall 1994. (Wooldridge)

Example: Determinants of college GPA

We can set up the following **population multiple regression model**

$$colGPA_i = \beta_0 + \beta_1 hsGPAI + \beta_2 ACTI + u_i \quad (5)$$

Example: Determinants of college GPA

```
. regress colGPA hsGPA ACT

      Source |       SS           df       MS      Number of obs   =       141
-------------+----------------------------------   F(2, 138)       =     14.78
       Model |  3.42365506         2  1.71182753   Prob > F        =    0.0000
    Residual |  15.9824444       138  .115814814   R-squared       =    0.1764
-------------+----------------------------------   Adj R-squared   =    0.1645
       Total |  19.4060994       140  .138614996   Root MSE        =    .34032

------------------------------------------------------------------------------
      colGPA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsGPA |   .4534559   .0958129     4.73   0.000     .2640047    .6429071
         ACT |    .009426   .0107772     0.87   0.383    -.0118838    .0307358
       _cons |   1.286328   .3408221     3.77   0.000      .612419    1.960237
------------------------------------------------------------------------------
```

Example: Determinants of college GPA

Estimated equation (or **OLS regression line**):

$$\widehat{colGPA} = 1.29 + 0.452\widehat{hsGPA} + 0.0094ACT \tag{6}$$

- Interpretation: Holding ACT fixed, another point on high school grade point average is associated with another 0.453 points on college grade point average
- Or: If we compare two students with the same ACT, but the *hsGPA* of student A is one point higher, we predict student A to have a *colGPA* that is 0.453 points higher than that of student B

Omitted variable bias

Omitting relevant variables: the simple case

Let's work through some theory!

We have a true population model, which really needs x_1 and x_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{7}$$

But, we estimate an OLS regression line using *only* x_1

$$\widetilde{y} = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_1 + \widetilde{u} \tag{8}$$

Omitted variable bias

Assume a linear relationship between x_1 and x_2

$$x_2 = \delta_0 + \delta_1 x_1 + v \tag{9}$$

Plug in x_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{10}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u \tag{11}$$

$$= (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) x_1 + (\beta_2 v + u) \tag{12}$$

Conclusion: All estimated coefficents will be biased!

Example: Omitting ability in a wage equation

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u \tag{13}$$

$$abil = \delta_0 + \delta_1 educ + v \tag{14}$$

$$wage = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)educ + (\beta_2 v + u) \tag{15}$$

The return to education β_1 will be *overestimated* because $\beta_2\delta_1 > 0$. It will look as if people with many years of education earn very high wages, but this is partly due to the fact that people with more education are also more able on average.

When is there no omitted variable bias?

- If the omitted variable is irrelevant ($\beta_2 = 0$)
- If the omitted variable is uncorrelated ($\delta_1 = 0$)

Signing the direction of the bias

- With one omitted variable, we can sign the bias if we know the direction of β_2 and δ_1
- Conditional on x_1 and x_2, we can compute $E[\widetilde{\beta}_1]$

$$E[\widetilde{\beta}_1] = \beta_1 + \beta_2 \widetilde{\delta}_1 \tag{16}$$

- Note that the sign of $\widetilde{\delta}_1$ is the same as the sign of $Cov(x_{i1}, x_{i2})$.

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

OBV: More general cases

We can extend this intution when we add more independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \tag{17}$$

$$\widetilde{y} = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_1 + \widetilde{\beta}_2 x_2 \tag{18}$$

- No general statements possible about direction of bias
- Can assume one regressor uncorrelated with others to make analysis tractable

Measures of fit

SER and RMSE

As in regression with a single regressor, the *SER* and *RMSE* are measures of the spread of the Y around the regression line

Standard error of the regression

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^{\,2}} = \sqrt{\frac{SSR}{n-k-1}} \tag{19}$$

Root mean squared error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^{\,2}} = \sqrt{\frac{SSR}{n}} \tag{20}$$

R^2 and adjusted R^2

The R^2 is the fraction of the variance explained – same definition as in regression with a single regressor:

The problem with R^2

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \tag{21}$$

Recall that $ESS = \sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2$; $SSR = \sum_{i=1}^{n} \hat{u}_i{}^2$; and $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

But, the R^2 always increases when you add another regressor!

Meet your friend, the adjusted R^2

The **adjusted R^2, $\bar{R}^2$** corrects this problem by "penalizing" you for adding another regressor.

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} \tag{22}$$

Note that $\bar{R}^2 < R^2$, however, the two will become very close together if n is large.

Least squares assumptions

Least squares assumptions

We add one more assumption as we upgrade to the multiple regression model

- Zero conditional mean assumption: $E(u_i|X_{1i}, X_{2i}, ..., X_{ki}) = 0$
- all the Xs and Ys are independently and identically distributed draws from their joint distribution
- Large outliers are unlikely: all have nonzero finite fourth moments
- There is no perfect multicollinearity

Assumption 1: Zero conditional mean

- This has the same interpretation as in regression with a single regressor.
- Failure of this condition leads to **omitted variable bias**, specifically, if an omitted variable belongs in the equation (so is in u) and is correlated with an included X, then this condition fails and there is OVB.
- The best solution, if possible, is to include the omitted variable in the regression.
- A second, related solution is to include a variable that *controls* for the omitted variable (discussed in Ch. 7)

Assumptions 2 and 3

- **Assumption 2 (Xs and Y are i.i.d)** is satisfied automatically if the data are collected by simple random sampling
- **Assumption 3: Large outliers are rare** is the same we had before. Check your data (scatterplots!) to make sure no crazy values

Assumption 4: No multicollinearity

Perfect multicollinearity: When one of the regressors is an exact linear function of another regressor

Perfect multicollinearity means that you cannot estimate your models ... but Stata will fix this for you automatically by excluding any perfectly collinear variable!

Perfect multicollinearity

```
.       gen ACT_36 = ACT/36

.   regress colGPA hsGPA ACT ACT_36
note: ACT_36 omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =       141
-------------+----------------------------------   F(2, 138)       =     14.78
       Model |  3.42365506         2  1.71182753   Prob > F        =    0.0000
    Residual |  15.9824444       138  .115814814   R-squared       =    0.1764
-------------+----------------------------------   Adj R-squared   =    0.1645
       Total |  19.4060994       140  .138614996   Root MSE        =    .34032

------------------------------------------------------------------------------
      colGPA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsGPA |   .4534559   .0958129     4.73   0.000     .2640047    .6429071
         ACT |    .009426   .0107772     0.87   0.383    -.0118838    .0307358
      ACT_36 |          0  (omitted)
       _cons |   1.286328   .3408221     3.77   0.000      .612419    1.960237
------------------------------------------------------------------------------
```

Perfect multicollinearity

```
. gen lowhsGPA = hsGPA < 2

. regress colGPA hsGPA lowhsGPA ACT
note: lowhsGPA omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =       141
-------------+----------------------------------   F(2, 138)       =     14.78
       Model |  3.42365506         2  1.71182753   Prob > F        =    0.0000
    Residual |  15.9824444       138  .115814814   R-squared       =    0.1764
-------------+----------------------------------   Adj R-squared   =    0.1645
       Total |  19.4060994       140  .138614996   Root MSE        =    .34032

------------------------------------------------------------------------------
      colGPA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsGPA |   .4534559   .0958129     4.73   0.000     .2640047    .6429071
    lowhsGPA |          0  (omitted)
         ACT |    .009426   .0107772     0.87   0.383    -.0118838    .0307358
       _cons |   1.286328   .3408221     3.77   0.000      .612419    1.960237
------------------------------------------------------------------------------
```

Perfect multicollinearity - dummy variable trap

Here we have a dummy variable trap

$$colGPA = \beta_0 + \beta_1 fresh + \beta_2 soph + \beta_3 junior + \beta_4 senior + \beta_5 hsGPS + u \tag{23}$$

Perfect multicollinearity - dummy variable trap

```
.  regress colGPA fresh soph jun senior hsGPA,robust
note: fresh omitted because of collinearity

Linear regression                               Number of obs     =        141
                                                F(4, 136)         =       6.66
                                                Prob > F          =     0.0001
                                                R-squared         =     0.1734
                                                Root MSE          =     .34344
```

colGPA	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
fresh	0	(omitted)				
soph	.0714571	.3010712	0.24	0.813	-.5239295	.6668436
junior	-.0086131	.0914072	-0.09	0.925	-.1893764	.1721503
senior	-.0224848	.0881555	-0.26	0.799	-.1968178	.1518482
hsGPA	.4739247	.1003441	4.72	0.000	.2754881	.6723613
_cons	1.457486	.3277041	4.45	0.000	.8094308	2.105541

Imperfect multicollinearity

- Imperfect multicollinearity occurs when two or more regressors are very highly correlated.
- Their scatterplot will pretty much look like a straight line – almost "co-linear" – but unless the correlation is exactly ± 1, that collinearity is imperfect.
- The idea: the coefficient on X_1 is the effect of X_1 holding X_2 constant; but if X_1 and X_2 are highly correlated, there is very little variation in X_1 once X_2 is held constant – so the data don't contain much information about what happens when X_1 changes but X_2 doesn't. If so, the variance of the OLS estimator of the coefficient on X_1 will be large.
- Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.
- Think carefully about what controls you need when buildling your regression

Sampling distribution of OLS estimators, multiple regression

- Under the four Least Squares Assumptions,
- The sampling distribution of $\hat{\beta}_1$ has mean β_1 (unbiased!)
- $var(\hat{\beta}_1)$ is inversely proportional to n
- Other than its mean and variance, the exact (finite-n) distribution of $\hat{\beta}_1$ is very complicated; but for large $n \ldots$
  - $\hat{\beta}_1$ is consistent: $\hat{\beta}_1 \xrightarrow{p} \beta_1$
  - The OLS estimators are jointly normally distributed
  - Each $\dfrac{\hat{\beta}_j - E(\hat{\beta}_j)}{\sqrt{var(\hat{\beta}_j}}$ is distributed approximately $N(0, 1)$
  - These hold statements for all our $\hat{\beta}_j$

Conceptually, there is nothing new here!

Conclusion

Multiple regression analysis

Omitted variable bias

Measures of fit

Least squares assumptions