Chapter 3: Multiple Regression Analysis: Estimation

Introductory Econometrics: A Modern Approach

3.1 Definition of the multiple linear regression model

"Explains variable 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,...

Why use multiple regression model?

1. Incorporate more explanatory factors into the model
2. Explicitly hold fixed other factors that otherwise would be in the error term
3. Allow for more flexible functional forms

Examples

- 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…

$\boxed{\text{Hourly wage}}$ Years of education Years of labor market experience

- Average test scores and per student spending

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

Other factors

$\boxed{\text{Average standardized test score of school}}$

Per student spending at this school

Average family income of students at this school

- Per student spending is likely to be correlated with average family income at a given high school because of school financing.
- Omitting average family income in regression would lead to biased estimate of the effect of spending on average test scores.

Examples

- Family income and family consumption
    - Model has two explanatory variables: income and income squared

$$\underbrace{cons}_{\boxed{\text{Family consumption}}} = \beta_0 + \beta_1 \underbrace{inc}_{\text{Family income}} + \beta_2 \underbrace{inc^2}_{\text{Family income squared}} + \underbrace{u}_{\text{Other factors}}$$

By how much does consumption increase if income is increased by one unit?

$$\frac{\Delta\,\mathrm{cons}}{\Delta\,\mathrm{inc}} \approx \beta_1 + 2\beta_2\,\mathrm{inc}$$

Depends on how much income is already there.

- CEO salary, sales and CEO tenure

$$\underbrace{\log(salary)}_{\boxed{\text{Log of CEO salary}}} = \beta_0 + \beta_1 \underbrace{\log(sales)}_{\text{Log sales}} + \underbrace{\beta_2 ceoten + \beta_3 ceoten^2}_{\text{Quadratic function of CEO tenure with the firm}} + u$$

3.2 Mechanics and Interpretation of Ordinary Least Squares

a) Obtaining the OLS estimates

- Random sample

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, \ldots, n\}$$

- Regression residuals

$$\widehat{u}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \ldots - \widehat{\beta}_k x_{ik}$$

- Minimize sum of squared residuals

$$\min \sum_{i=1}^{n} \widehat{u}_i^2 \rightarrow \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$$

b) Interpreting the OLS Regression Equation

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \ldots + \widehat{\beta_k} x_k$$

By how much does the dependent variable change if the k-th independent variable is increased by one unit, holding all other independent variables constant?

$$\beta_k = \frac{\Delta y}{\Delta x_k}$$

- "Ceterius paribus" holding all other independent variables constant
- The multiple linear regression model manages to hold the values of other explanatory variables fixed even if they are correlated with the explanatory variable under consideration.
- It has still to be assumed that unobserved factors do not change if the explanatory variables are changed.

```r
data(gpa1, package='wooldridge')

GPAsingle<-feols(colGPA~ ACT, data = gpa1)
GPAres<-feols(colGPA~ hsGPA+ACT, data = gpa1)

models <- list( GPAsingle,GPAres)

modelsummary(models,output = "markdown")
```

	Model 1	Model 2
(Intercept)	2.403	1.286
	(0.264)	(0.341)
ACT	0.027	0.009
	(0.011)	(0.011)
hsGPA		0.453
		(0.096)
Num.Obs.	141	141
R2	0.043	0.176
R2 Adj.	0.036	0.164
R2 Within		

Example: 3.2 Hourly wage equation

```r
data(wage1, package='wooldridge')

summary(feols(lwage~educ+exper+tenure, data = wage1))
```

```
## OLS estimation, Dep. Var.: lwage
## Observations: 526
## Standard-errors: IID
##             Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 0.284360   0.104190  2.72923 6.5625e-03 **
## educ        0.092029   0.007330 12.55525  < 2.2e-16 ***
## exper       0.004121   0.001723  2.39144 1.7136e-02 *
## tenure      0.022067   0.003094  7.13307 3.2944e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.439183   Adj. R2: 0.312082
```

c) The Meaning of "Holding Other Factors Fixed" in Multiple Regression

- In example 3.1, we observed that the coefficient on `ACT` measures the predicted difference in `colGPA`, holding `hsGPA` fixed
- It may seem that we sampled people with the same `hsGPA` but different `ACT` scores
- **Data is random** , no restrictions were placed on the values of `hsGPA` and `ACT`
- If we could collect a sample of individuals with the same `hsGPA`, then we could preform a simple regression `ColGPA` and `ACT`
- Multiple regression allows us to mimic this situation without restricting the values of any independent variables

d) Changing More than One Independent Variable Simultaneously

$$\Delta \widehat{\log(\text{ wage })} = .0041\Delta exper + 0.22\Delta tenure$$

What is the effect on `wage` if `exper` and `tenure` both increase by *one year*?

The total effect (holding `educ` fixed) is:

$$\Delta \widehat{\log(\text{ wage })} = .0041\Delta exper + 0.22\Delta tenure = .0041 * 1 + .022 * 1 = .0261$$

Because `exper` and `tenure` each increase by one year, we just add the coefficients on `exper` and `tenure` and multiply by 100 to tun the effect into percentage (about 2.6 %).

e) OLS Fitted Values and Residuals

Fitted Values:

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_k x_{ik}$$

Residuals:

$$\widehat{u}_i = y_i - \hat{y}_i$$

Algebraic properties of OLS regression:

1. Deviations from regression line sum up to zero

$$\sum_{i=1}^{n} \widehat{u}_i = 0$$

2. Covariance between deviations and regressors are zero

$$\sum_{i=1}^{n} x_{ij}\widehat{u}_i = 0$$

3. Sample averages of y and of the regressors lie on regression line

$$\bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1\bar{x}_1 + \ldots + \widehat{\beta}_k\bar{x}_k$$

f) "Partialling Out" Interpretation of Multiple Regression

One can show that the estimated coefficient of an explanatory variable in a multiple regression can be obtained in two steps:

1. Regress the explanatory variable on all other explanatory variables

2. Regress y on the $residuals$ from previous regression Why does this procedure work?

    - The residuals from the first regression is the part of the explanatory variable that is uncorrelated with the other explanatory variables.
    - The slope coefficient of the second regression therefore represents the isolated effect of the explanatory variable on the dependent variable.

h) Goodness-of-fit

Decomposition of total variation

$$SST = SSE + SSR$$

R squared

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

R^2 can be views as the squared correlation coefficient between the actual y_i and fitted values $\widehat{y}_i$:

$$R^2 = \frac{\left(\sum_{i=1}^{n} (y_i - \bar{y}) \left(\hat{y}_i - \bar{\hat{y}}\right)\right)^2}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right) \left(\sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2\right)}$$

Examples

Explaining arrest records

Number of times arrested 1986

Proportion prior arrests that led to conviction

Months in prison 1986

Quarters employed 1986

$$\widehat{narr}86 = .712 - .150\, pcnv - .034\, ptime86 - .104\, qemp86$$

$$n = 2,725, \quad R^2 = .0413$$

Interpretation:

- If the proportion prior arrests increases by 0.5, the predicted fall in arrests is 7.5 arrests per 100 men, ceteris paribus.
- If the months in prison increase from 0 to 12, the predicted fall in arrests is 0.408 arrests for a particular man, ceteris paribus.
- If the quarters employed increase by 1, the predicted fall in arrests is 10.4 arrests per 100 men, ceteris paribus.

An additional explanatory variable is added.

$$\widehat{narr}86 = .707 - .151\, pcnv + .0074\, avgsen - .037\, ptime86 - .103\, qemp86$$

Average sentence in prior convictions

$$n = 2,725, \quad R^2 = .0422$$

R-squared increases only slightly

- Average prior sentence increases the number of arrests (?)
- Limited additional explanatory power as R-squared increases by little

3.3 The Expected Value of OLS Estimators

Standard assumptions for the multiple regression model:

MLR 1. Linear in parameters

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

MLR 2. Random sampling

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, \ldots, n\}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$$

Standard assumptions for the multiple regression model:

MLR 3. No perfect collinearity

In the sample (and therefore in the population), none of the independent variables is constant and there are no exact linear relationships among the independent variables.

- The assumption only rules out perfect collinearity/correlation between explanatory variables; **imperfect correlation** is allowed.

--

- If an explanatory variable is a perfect linear combination of other explanatory variables it is superfluous and may be eliminated.

--

- Constant variables are also ruled out (collinear with intercept).

Example for PERFECT collinearity:

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

In a small sample, `avginc` may accidentally be an exact multiple of expend; it will not be possible to disentangle their separate effects because there is exact covariation

- Example for perfect collinearity: relationships between regressors

$$vote\ A = \beta_0 + \beta_1 share\ A + \beta_2 share\ B + u$$

Either *share A* or *share B* will have to be dropped from the regression because there is an exact linear relationship between them: $share\ A + share\ B = 1$

MLR 4. Zero conditional mean

The value of the explanatory variables must contain no information about the mean of the unobserved factors:

$$E\left(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}\right)=0$$

In a multiple regression model, the zero conditional mean assumption is much more likely to hold because fewer things end up in the error term

$avgscore = \beta_0 + \beta_1 expend + \beta_2 \, \text{avginc} + u$

- If `avgin` was not included in the regression, it would end up in the error term
- I would then be harder to defend that `expend` is uncorrelated with the error term

Theorem: Unbiasedness of OLS: Under assumptions MLR.1 and MLR.4 $E(\widehat{\beta_j}) = \beta_j$ for j=0,1, … k for any values of the population parameter β_j.

In other words, the OLS estimators are unbiased estimators of the population parameters

Including irrelevant variables in the regression model

Overspecifying the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

No problem because $E(\widehat{\beta_3}) = \beta_3 = 0$

However, including irrelevant variables may increase sampling variance (We will see in the next section)

Ommitting relevant variables: the simple case

What is omitted variable bias?

True model contains x_1 and x_2

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

Estimated model (x_2 is omitted)

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{u}$$

- If x_1 and x_2 are correlated, assume a linear regression relationship between them

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \left(\delta_0 + \delta_1 x_1 + v\right) + u$$

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

If y is only regressed on x_1 this will be the estimated intercept

If y is only regressed on x_1, this will be the estimated slope on x_1

error term

Conclusion: All estimated coefficients will be biased

Ommitted variable bias

Example: Omitting ability in a wage equation

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

Will both be positive

$$abil = \delta_0 + \delta_1 educ + v$$

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

The return to education β_1 will be <u>overestimated</u> 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 or uncorrelated

Ommitted variable bias

Example: Omitting ability in a wage equation

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

If `exper` is approximately uncorrelated with `educ` and `abil`, then the direction of the omitted variable bias can be as analyzed in the simple two variable case.

Table 3.2. Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

?	$\mathrm{Corr}(x_1, x_2) > 0$	$\mathrm{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

MLR.5 Homoskedasticity

The value of the explanatory variables must contain no information about the variance of the unobserved factors

$$\mathrm{Var}(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2$$

$$\mathrm{Var}(u_i | \text{ educ }_i, \text{ exper }_i, \text{ tenure }_i) = \sigma^2$$

This assumption may also be hard to justify in many cases Short hand notation:

$$\mathrm{Var}(u_i \mid \mathbf{x}_i) = \sigma^2 \quad \text{with} \quad \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$$

3.4 The Variance of the OLS Estimators

Theorem: Sampling variances of the OLS slope estimators

Under assumptions MLR.1 through MLR.5, conditional on the sample values of the independet variables:

Variance of the error term

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \quad j = 1, \dots, k$$

Total sample variation in explanatory variable x_j:

$$\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

R-squared from a regression of explanatory variable x_j on all other independent variables (including a constant)

for $j = 1, 2, \dots k$

Note: R_j^2 is the R-squared from regression x_j on all other independent variables (and including an intercept)

Components of OLS variances

The error variance (σ^2)

- A high error variance increases the sampling variance because there is more “noise” in the equation.
- A large error variance does not necessarily make estimates imprecise.
- The error variance does not decrease with sample size. **The Total Sample Variation in x_j**

$$\mathrm{SST}_j = \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2$$

- More sample variation leads to more precise estimates.
- Total sample variation automatically increases with the sample size.
- Increasing the sample size is thus a way to get more precise estimates.

Components of OLS variances

Linear relationships among the independent variables

- Regress x_j on all other independent variables (including constant)
- The R-squared of this regression will be the higher when x_j can be better explained by the other independent variables.
- The sampling variance of the slope estimator for x_j will be higher when x_j can be better explained by the other independent variables.
- Under perfect multicollinearity, the variance of the slope estimator will approach infinity.

Multicollinearity

Average standardized test score of school

Expenditures for teachers

Expenditures for instructional materials

Other expenditures

$$avgscore = \beta_0 + \beta_1 teachexp + \beta_2 matexp + \beta_3 othexp + \ldots$$

- The different expenditure categories will be strongly correlated because if a school has a lot of resources it will spend a lot on everything.
- It will be hard to estimate the differential effects of different expenditure categories because all expenditures are either high or low. For precise estimates of the differential effects, one would need information about situations where expenditure categories change differentially.
- As a consequence, sampling variance of the estimated effects will be large.
- In the above example, it would probably be better to lump all expenditure categories together because effects cannot be disentangled.
- In other cases, dropping some independent variables may reduce multicollinearity (but this may lead to omitted variable bias).

Multicollinearity

- Only the sampling variance of the variables involved in multicollinearity will be inflated; the estimates of other effects may be very precise.
- Note that multicollinearity is not a violation of MLR.3 in the strict sense.
- Multicollinearity may be detected through “variance inflation factors.”
- As an (arbitrary) rule of thumb, the variance inflation factor (VIF) should not be larger than 10.

$$VIF_j = \frac{1}{1 - R_j^2}$$

Variances in misspecified models

The choice of whether to include a particular variable in a regression can be made by analyzing the tradeoff between bias and variance.

True population model

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

Estimated model 1

$$widehaty = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$$

Estimated model 2

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

It might be the case that the likely omitted variable bias in the misspecified model 2 is overcompensated by a smaller variance.

Variances in misspecified models

Conditional on x_1 and x_2, the variance of model 2 is always smaller than that in model 1

$$\mathrm{Var}\left(\widehat{\beta}_1\right) = \sigma^2 / \left[SST_1 \left(1 - R_1^2\right)\right]$$

$$\mathrm{Var}\left(\tilde{\beta}_1\right) = \sigma^2 / SST_1$$

- Case 1

$$\beta_2 = 0 \Rightarrow E\left(\widehat{\beta}_1\right) = \beta_1, E\left(\tilde{\beta}_1\right) = \beta_1, \mathrm{Var}\left(\tilde{\beta}_1\right) < \mathrm{Var}\left(\widehat{\beta}_1\right)$$

Conclusion: Do not include irrelevant regressors

- Case 2

  $$\beta_2 \neq 0 \Rightarrow E\left(\widehat{\beta}_1\right) = \beta_1, E\left(\tilde{\beta}_1\right) \neq \beta_1, \mathrm{Var}\left(\tilde{\beta}_1\right) < \mathrm{Var}\left(\widehat{\beta}_1\right)$$

  Conclusion: Trade off bias and variance; Caution: bias will not vanish even in large samples

Estimating the error variance

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} \widehat{u}_i^2\right) / [n - k - 1]$$

- An unbiased estimate of the error variance can be obtained by subtracting the number of estimated regression coefficients from the number of observations.
- The number of observations minus the number of estimated parameters is also called the degrees of freedom.
- The `n` estimated squared residuals in the sum are not completely independent but related through the `k+1` equations that define the first order conditions of the minimization problem.

Estimation of the sampling variances of the OLS estimators

The true sampling variation of the estimated β_j

$$sd\left(\widehat{\beta}_j\right) = \sqrt{\mathrm{Var}\left(\widehat{\beta}_j\right)} = \sqrt{\sigma^2 / \left[SST_j\left(1 - R_j^2\right)\right]}$$

The estimated sampling variation of the estimated β_j

Plug $\hat{\sigma}^2$ for σ^2

$$\mathrm{se}\left(\widehat{\beta}_j\right) = \sqrt{\widehat{\mathrm{Var}}\left(\widehat{\beta}_j\right)} = \sqrt{\hat{\sigma}^2 \Big/ \left[SST_j\left(1 - R_j^2\right)\right]}$$

Note: that these formulas are only valid under assumptions MLR.1-MLR.5 (in particular, there has to be homoskedasticity)

Efficiency of OLS: The Gauss-Markov Theorem

- Under assumptions MLR.1 - MLR.5, OLS is unbiased. However, under these assumptions there may be many other estimators that are unbiased.
- Which one is the unbiased estimator with the smallest variance?
  - In order to answer this question one usually limits oneself to linear estimators, i.e. estimators linear in the dependent variable.
- May be an arbitrary function of the sample values of all the explanatory variables; the OLS estimator can be shown to be of this form:

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$$

Efficiency of OLS: The Gauss-Markov Theorem

- Under assumptions MLR.1 - MLR.5, the OLS estimators are the best linear unbiased estimators (BLUEs) of the regression coefficients, i.e.

$$\mathrm{Var}\left(\widehat{\beta}_j\right) \leq \mathrm{Var}\left(\tilde{\beta}_j\right) \quad j = 0, 1, \ldots, k$$

for all $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$ for which $E\left(\tilde{\beta}_j\right) = \beta_j, j = 0, \ldots, k$.

- OLS is only the best estimator if MLR.1 – MLR.5 hold
- If there is heteroskedasticity for example, there are better estimators.