

Chapter 3

Multiple Regression Analysis: Estimation

Multiple Regression Analysis: Estimation

- Definition of the **multiple** linear regression model

"Explains variable y in terms of variables x_1, x_2, \dots, x_k "

↑
 k : number of variables

Intercept

Slope parameters

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

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

Independent variables,
explanatory variables,
regressors,...

Error term,
disturbance,
unobservables,...

Multiple Regression Analysis: Estimation

- **Motivation for multiple regression**

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

- **Example: Wage equation**

Now measures effect of education explicitly holding experience fixed

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

Hourly wage

Years of education

Years of labor market experience

All other factors...

Multiple Regression Analysis: Estimation

- **Example: Average test scores and per student spending**

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

Average standardized
test score of school

Per student spending
at this school

Average family income
of students at this school

Other factors

- 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
- In a simple regression model, effect of per student spending would **partly include** the effect of family income on test scores

Multiple Regression Analysis: Estimation

- Example: Family income and family consumption**

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

Family consumption Family income Family income squared Other factors

- Model has two explanatory variables: income and income squared
- Consumption is explained as a **quadratic** function of income
- One has to be very careful when interpreting the coefficients:

By how much does consumption increase if income is increased by one unit? $\frac{\Delta \text{cons}}{\Delta \text{inc}} \approx \beta_1 + 2\beta_2 \text{inc}$ Depends on how much income is already there

Multiple Regression Analysis: Estimation

- **Example: CEO salary, sales, and CEO tenure**

$$\log(\text{salary}) = \beta_0 + \beta_1 \log(\text{sales}) + \beta_2 \text{ceoten} + \beta_3 \text{ceoten}^2 + u$$

Log of CEO salary

Log sales

Quadratic function of CEO tenure with the firm

- Model assumes a constant elasticity relationship between CEO salary and the sales of his or her firm
- Model assumes a quadratic relationship between CEO salary and his or her tenure with the firm
- **Meaning of “linear” regression**
 - The model has to be **linear in the parameters** (not in the variables)

Multiple Regression Analysis: Estimation

- **OLS Estimation of the multiple regression model**
- **Random sample**

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

- **Regression residuals**

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

- **Minimize sum of squared residuals (SSR)**

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



Minimization will be carried out by computer

Multiple Regression Analysis: Estimation

- **Interpretation of the multiple regression model**

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

By how much does the dependent variable change if the j-th independent variable is increased by one unit, holding all other independent variables and the error term constant

- The multiple linear regression model manages to hold the values of other explanatory variables fixed even if, in reality, they are correlated with the explanatory variable under consideration
- “**Ceteris paribus**”-interpretation
- It has still to be assumed that unobserved factors (error) do not change if the explanatory variables are changed

Multiple Regression Analysis: Estimation

- **Example: Determinants of college GPA**

$$\widehat{colGPA} = 1.29 + .453 \, hsGPA + .0094 \, ACT$$

Grade point average at college

High school grade point average

Achievement test score

- **Interpretation**

- Holding ACT fixed, another point on high school grade point average is associated with another .453 points 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 .453 higher than that of student B
- **Holding high school grade point average fixed**, another 10 points on ACT are associated with less than one point on college GPA

Multiple Regression Analysis: Estimation

- **Properties of OLS on any sample of data**
- **Fitted values and residuals**

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

↑
Fitted or predicted values

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

↑
Residuals

- **Algebraic properties of OLS regression**

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

↑
Deviations from regression line sum up to zero

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

↑
Covariance between deviations and regressors are zero

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \dots + \hat{\beta}_k \bar{x}_k$$

↑
Sample averages of y and of the regressors lie on regression line

Multiple Regression Analysis: Estimation

- **“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 this 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 dep. variable

Multiple Regression Analysis: Estimation

- **Goodness-of-Fit**
- **Decomposition of total variation**

$$STT = SSE + SSR$$

- **R-squared**

$$R^2 \equiv SSE/SST = 1 - SSR/SST$$

Notice that R-squared can only increase if another explanatory variable is added to the regression

- **Alternative expression for R-squared**

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

R-squared is equal to the squared correlation coefficient between the **actual** and the **predicted** value of the dependent variable

$$R^2 = \widehat{\text{Cov}}(y, \hat{y})$$

Multiple Regression Analysis: Estimation

- **Example: Explaining arrest records**

Number of times
arrested 1986

Proportion prior arrests
that led to conviction

Months in prison 1986

Quarters employed 1986

$$\widehat{narr86} = .712 - .150 pcnv - .034 ptime86 - .104 qemp86$$

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

- **Interpretation:**

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

Multiple Regression Analysis: Estimation

- **Example: Explaining arrest records (cont.)**

- An additional explanatory variable is added:

$$\widehat{narr86} = .707 - .151 pcnv + .0074 \text{avgsen} - .037 ptime86 - .103 qemp86$$

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

Average sentence in prior convictions

R-squared increases only slightly

- **Interpretation:**

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

- **General remark on R-squared**

- Even if R-squared is small (as in the given example), regression may still provide good estimates of ceteris paribus effects

Multiple Regression Analysis: Estimation

- **Standard assumptions for the multiple regression model**
- **Assumption MLR.1 (Linear in parameters)**

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

In the population, the relationship between y and the explanatory variables is linear

- **Assumption MLR.2 (Random sampling)**

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

The data is a random sample drawn from the population

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

Each data point therefore follows the population equation



Multiple Regression Analysis: Estimation

- **Standard assumptions for the multiple regression model (cont.)**

- **Assumption 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.”

- **Remarks on MLR.3**

- 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)

Multiple Regression Analysis: Estimation

- **Example for perfect collinearity: small sample**

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

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**

$$voteA = \beta_0 + \beta_1 shareA + \beta_2 shareB + u$$

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

Multiple Regression Analysis: Estimation

- **Standard assumptions for the multiple regression model (cont.)**

- **Assumption MLR.4 (Zero conditional mean)**

$$E(u_i | x_{i1}, x_{i2}, \dots, x_{ik}) = 0$$

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

- In a multiple regression model, the zero conditional mean assumption is much more likely to hold because fewer things end up in the error
- **Example: Average test scores**

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

← If avginc was not included in the regression, it would **end up in the error term**; it would then be hard to defend that expend is uncorrelated with the error

Multiple Regression Analysis: Estimation

- **Discussion of the zero mean conditional assumption**
 - Explanatory variables that are correlated with the error term are called **endogenous**; endogeneity is a **violation of assumption MLR.4**
 - Explanatory variables that are uncorrelated with the error term are called **exogenous**; MLR.4 holds if all explanat. var. are exogenous
 - Exogeneity is the key assumption for a causal interpretation of the regression, and for unbiasedness of the OLS estimators

- **Theorem 3.1 (Unbiasedness of OLS)**

$$MLR.1 - MLR.4 \Rightarrow E(\hat{\beta}_j) = \beta_j, \quad j = 0, 1, \dots, k$$

- Unbiasedness is an **average property** in **repeated samples**; in a given sample, the estimates may still be far away from the true values

Multiple Regression Analysis: Estimation

- **Including irrelevant variables in a regression model**

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

No problem because $E(\hat{\beta}_3) = \beta_3 = 0$. ← = 0 in the population

However, including irrelevant variables may **increase sampling variance**.

- **Omitting relevant variables: the simple case**

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

← True model (contains x_1 and x_2)

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

← Estimated model (x_2 is omitted)

Multiple Regression Analysis: Estimation

- **Omitted variable bias**

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

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

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

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

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**

Multiple Regression Analysis: Estimation

- **Example: Omitting ability in a wage equation**

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

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

Will both be positive

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

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** or **uncorrelated**

Multiple Regression Analysis: Estimation

- **Omitted variable bias: more general cases**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \quad \leftarrow \text{True model (contains } x_1, x_2, \text{ and } x_3)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + w \quad \leftarrow \text{Estimated model (} x_3 \text{ is omitted)}$$

- No general statements possible about direction of bias
- Analysis as in simple case if one regressor uncorrelated with others

- **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.

Multiple Regression Analysis: Estimation

- **Standard assumptions for the multiple regression model (cont.)**

- **Assumption MLR.5 (Homoskedasticity)**

$$Var(u_i | x_{i1}, x_{i2}, \dots, x_{ik}) = \sigma^2$$

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

- **Example: Wage equation**

$$Var(u_i | educ_i, exper_i, tenure_i) = \sigma^2$$

This assumption may also be hard to justify in many cases

- **Short hand notation**

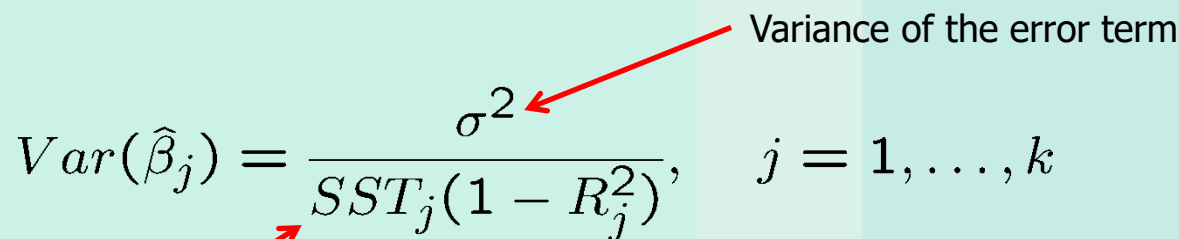
$$Var(u_i | \mathbf{x}_i) = \sigma^2 \quad \text{with} \quad \mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$$

All explanatory variables are collected in a random vector

Multiple Regression Analysis: Estimation

- Theorem 3.2 (Sampling variances of the OLS slope estimators)**

Under assumptions MLR.1 – MLR.5:

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


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)



Multiple Regression Analysis: Estimation

- **Components of OLS Variances:**
- **1) The error variance**
 - A high error variance increases the sampling variance because there is more “**noise**” in the equation
 - A large error variance necessarily makes estimates imprecise
 - The error variance **does not decrease with sample size**
- **2) The total sample variation in the explanatory variable**
 - **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

Multiple Regression Analysis: Estimation

- **3) Linear relationships among the independent variables**

Regress x_j on all other independent variables (including a constant)



The R-squared of this regression will be the higher the better x_j can be linearly explained by the other independent variables

- Sampling variance of $\hat{\beta}_j$ will be the higher the better explanatory variable x_j can be linearly explained by other independent variables
- The problem of **almost** linearly dependent explanatory variables is called **multicollinearity** (i.e. $R_j \rightarrow 1$ for some j)

Multiple Regression Analysis: Estimation

- An example for **multicollinearity**

Average standardized test score of school

Expenditures for teachers

Expenditures for instructional materials

Other expenditures

$$avgscore = \beta_0 + \beta_1 teachexp + \beta_2 matexp + \beta_3 otherexp + \dots$$

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.



Multiple Regression Analysis: Estimation

- **Discussion of the multicollinearity problem**
 - 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)

Multiple Regression Analysis: Estimation

- 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**”

$$VIF_j = 1/(1 - R_j^2)$$

As an (arbitrary) rule of thumb, the variance inflation factor should not be larger than 10

Multiple Regression Analysis: Estimation

- **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**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \leftarrow \text{True population model}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \leftarrow \text{Estimated model 1}$$

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \leftarrow \text{Estimated model 2}$$

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

Multiple Regression Analysis: Estimation

- **Variances in misspecified models (cont.)**

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

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

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

- **Case 1:**

Conclusion: Do not include irrelevant regressors


$$\beta_2 = 0 \Rightarrow E(\hat{\beta}_1) = \beta_1, E(\tilde{\beta}_1) = \beta_1, Var(\tilde{\beta}_1) < Var(\hat{\beta}_1)$$

- **Case 2:** Trade off bias and variance; Caution: bias will not vanish even in large samples

$$\beta_2 \neq 0 \Rightarrow E(\hat{\beta}_1) = \beta_1, E(\tilde{\beta}_1) \neq \beta_1, Var(\tilde{\beta}_1) < Var(\hat{\beta}_1)$$

Multiple Regression Analysis: Estimation

- **Estimating the error variance**

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


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.

$$MLR.1 - MLR.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2$$

- **Theorem 3.3 (Unbiased estimator of the error variance)**

Multiple Regression Analysis: Estimation

- **Estimation of the sampling variances of the OLS estimators**

The true sampling variation of the estimated β_j

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

Plug in $\hat{\sigma}^2$ for the unknown σ^2

The estimated sampling variation of the estimated β_j

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

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

Multiple Regression Analysis: Estimation

- **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

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

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

Multiple Regression Analysis: Estimation

- **Theorem 3.4 (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.

$$Var(\hat{\beta}_j) \leq Var(\tilde{\beta}_j) \quad j = 0, 1, \dots, k$$

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

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