Kurt Schmidheiny
Universität Basel

The Multiple Linear Regression Model

matrix-free

1 Introduction

The multiple linear regression model and its estimation using ordinary least squares (OLS) is doubtless the most widely used tool in econometrics. It allows to estimate the relation between a dependent variable and a set of explanatory variables. Prototypical examples in econometrics are:

- Wage of an employee as a function of her education and her work experience (the so-called Mincer equation).
- Price of a house as a function of its number of bedrooms and its age (an example of hedonic price regressions).

The dependent variable is an interval variable, i.e. its values represent a natural order and differences of two values are meaningful. The dependent variable can, in principle, take any real value between $-\infty$ and $+\infty$. In practice, this means that the variable needs to be observed with some precision and that all observed values are far from ranges which are theoretically excluded. Wages, for example, do strictly speaking not qualify as they cannot take values beyond two digits (cents) and values which are negative. In practice, monthly wages in dollars in a sample of full time workers is perfectly fine with OLS whereas wages measured in three wage categories (low, middle, high) for a sample that includes unemployed (with zero wages) ask for other estimation tools.

2 The Econometric Model

The multiple linear regression model assumes a linear (in parameters) relationship between a dependent variable y_i and a set of explanatory variables $x_{i0}, x_{i1}, ..., x_{iK}$. x_{ik} is also called an independent variable, a covariate or a regressor. The first regressor $x_{i0} = 1$ is a constant unless otherwise specified.

Consider a sample of N observations $i = 1, \ldots, N$. Every single observation i follows

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + u_i$$

where $\beta_0, \beta_1, ..., \beta_K$ are $K+1$ parameters and u_i is called the error term.

The *binary regression* model is a special case with only one regressor:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

The data generation process (dgp) is fully described by a set of assumptions. Several of the following assumptions are formulated in different alternatives. Different sets of assumptions will lead to different properties of the OLS estimator.

OLS1: Linearity

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + u_i \text{ and } E(u_i) = 0$$

OLS1 assumes that the functional relationship between dependent and explanatory variables is linear *in parameters*, that the error term enters additively and that the parameters are constant across individuals i.

OLS2: Independence

$$\{x_{i1}, ..., x_{iK}, y_i\}_{i=1}^N \text{ i.i.d. (independent and identically distributed)}$$

OLS2 means that the observations are independently and identically distributed. This assumption is in practice guaranteed by random sampling.

OLS3: Exogeneity

a) $u_i|x_{i1},...,x_{iK} \sim N(0,\sigma_i^2)$

b) $\forall k: u_i \perp x_{ik}$ (independent)

c) $E(u_i|x_{i1},...,x_{iK}) = 0$ (mean independent)

d) $\forall k: cov(x_{ik}, u_i) = 0$ (uncorrelated)

OLS3a assumes that the error term is normally distributed conditional on the explanatory variables. *OLS3b* means that the error term is independent of the explanatory variables. *OLS3c* states that the *mean* of the error term is independent of the explanatory variables. *OLS3d* means that the error term and the explanatory variables are uncorrelated. Either *OLS3a* or *OLS3b* imply *OLS3c* and *OLS3d*. *OLS3c* implies *OLS3d*.

OLS4: Identifiability

$(x_{i0}, x_{i1}, \cdots, x_{iK})$ are not linearly dependent and
$0 < V(x_{ik}) < \infty$ for all $k > 0$

The *OLS4* assumes that the regressors are not perfectly collinear, i.e. no variable is a linear combination of the others. For example, there can only be one constant. Intuitively, *OLS4* means that every explanatory variable adds additional information. *OLS4* also assumes that all regressors (but the constant) have non-zero variance and not too many extreme values.

OLS5: Error Structure

a) $V(u_i|x_{i1},...,x_{iK}) = \sigma^2 < \infty$ (homoscedasticity)

b) $V(u_i|x_{i1},...,x_{iK}) = \sigma_i^2 = g(x_{i1},...,x_{iK}) < \infty$
(conditional heteroscedasticity)

OLS5a (homoscedasticity) means that the variance of the error term is a constant. *OLS5b* (conditional heteroscedasticity) allows the variance of the error term to depend on the explanatory variables.

3 Estimation with OLS

Ordinary least squares (OLS) minimizes the squared distances between the observed and the predicted dependent variable y:

$$S(\beta_0, ..., \beta_K) = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK})]^2 \rightarrow \min_{\beta_0,...,\beta_K}$$

For the binary regression model, the resulting OLS estimators of β_0 and β_1 are:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

The general case with K explanatory variables is derived likewise but not easily expressed without matrix notation.

Given the OLS estimator, we can predict the dependent variable by $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + ... + \widehat{\beta}_K x_{iK}$ and the error term by $\widehat{u}_i = y_i - \widehat{y}_i$. $\widehat{u}_i$ is called the *residual*.

4 Goodness-of-fit

The goodness-of-fit of an OLS regression can be measured as

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

where $SST = \sum_{i=1}^N (y_i - \bar{y})^2$ is the total sum of squares, $SSE = \sum_{i=1}^N (\widehat{y}_i - \bar{y})^2$ the explained sum of squares and $SSR = \sum_{i=1}^N \widehat{u}_i^2$ the residual sum of squares. R^2 lies by definition between 0 and 1 and reports the fraction of the sample variation in y that is explained by the xs.

Note: R^2 is in general not meaningful in a regression without a constant. R^2 increases by construction with every (also irrelevant) additional

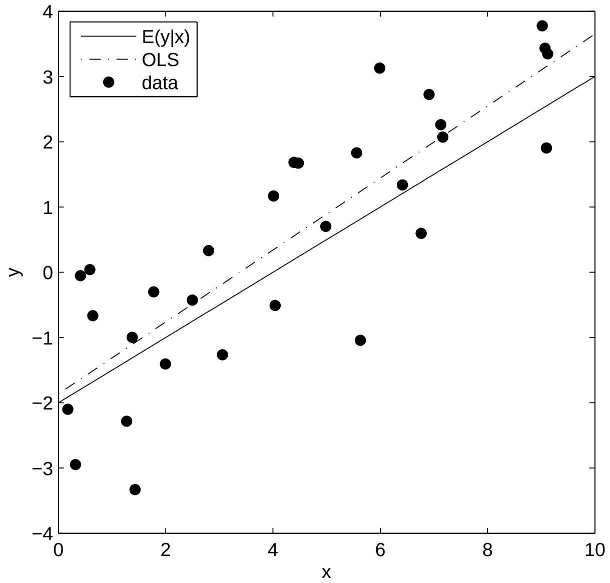

Figure 1: The linear regression model with one regressor. $\beta_0 = -2$, $\beta_1 = 0.5$, $\sigma^2 = 1$, $x \sim \text{uniform}(0, 10)$, $u \sim N(0, \sigma^2)$.

regressors and is therefore not a good criterium for the selection of regressors.

5 Small Sample Properties

Assuming *OLS1*, *OLS2*, *OLS3a*, *OLS4*, and *OLS5*, the following properties can be established for finite, i.e. even small, samples.

- The OLS estimator of β is *unbiased*:

$$E(\widehat{\beta}_k | x_{11}, ..., x_{NK}) = \beta_k$$

- The OLS estimator is (multivariate) normally distributed:

$$\widehat{\beta}_k | x_{11}, ..., x_{NK} \sim N\left(\beta_k, V(\widehat{\beta}_k)\right) .$$

  Under homoscedasticity (*OLS5a*) the variance $V(\widehat{\beta}_1)$ can be *unbiasedly* estimated. For the binary regression model, it is estimated as

$$\widehat{V}(\widehat{\beta}_1 | x_{11}, ..., x_{NK}) = \frac{\widehat{\sigma}^2}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

  with

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^N \widehat{u}_i^2}{N - K - 1} = \frac{\sum_{i=1}^N \widehat{u}_i^2}{N - 2} .$$

- Gauß-Markov-Theorem: under homoscedasticity (*OLS5a*),

$$\widehat{\beta}_k \text{ is BLUE (best linear unbiased estimator)}$$

Table 1 provides a systematic overview of assumptions and their implied properties.

Table 1: Properties of the OLS estimator.

Case	[1]	[2]	[3]	[4]	[5]	[6]
Assumptions						
OLS1: linearity	✓	✓	✓	✓	✓	✓
OLS2: independence	✓	✓	✓	✓	✓	✓
OLS3: exogeneity						
- OLS3a: normality	✓	×	×	×	✓	×
- OLS3b: independent	✓	✓	×	×	×	×
- OLS3c: mean indep.	✓	✓	✓	✓	✓	×
- OLS3d: uncorrelated	✓	✓	✓	✓	✓	✓
OLS4: identifiability	✓	✓	✓	✓	✓	✓
OLS5: error Structure						
- OLS5a: homoscedastic	✓	✓	✓			
- OLS5b: heteroscedastic				✓	✓	✓
Small sample properties of $\hat{\beta}$						
unbiased	✓	✓	✓	✓	✓	×
normally distributed	✓	×	×	×	✓	×
efficient	✓	✓	✓	×	×	×
t-test, F-test	✓	×	×	×	×	×
Large sample properties of $\hat{\beta}$						
consistent	✓	✓	✓	✓	✓	✓
approx. normal	✓	✓	✓	✓	✓	✓
asymptotically efficient	✓	✓	✓	×	×	×
z-test, Wald test	✓	✓	✓	✓*	✓*	✓*

Notes: ✓ = fullfiled, × = violated, * = corrected standard errors.

6 Tests in Small Samples

Assume *OLS1*, *OLS2*, *OLS3a*, *OLS4*, and *OLS5a*.

A simple null hypotheses of the form $H_0 : \beta_k = q$ is tested with the t-test. If the null hypotheses is true, the t-statistic

$$t = \frac{\widehat{\beta}_k - q}{\widehat{se}(\widehat{\beta}_k)} \sim t_{N-K-1}$$

follows a t-distribution with $N-K-1$ degrees of freedom. The standard error is $\widehat{se}(\widehat{\beta}_k) = \sqrt{\widehat{V}(\widehat{\beta}_k)}$. For example, to perform a two-sided test of

H_0 against the alternative hypotheses $H_a : \beta_k \neq q$ on the 5% significance level, we calculate the t-statistic and compare its absolute value to the 0.975-quantile of the t-distribution. With $N = 30$ and $K = 2$, H_0 is rejected if $|t| > 2.052$.

A null hypotheses of the form $H_0 : r_{j1}\beta_1 + ... + r_{jK}\beta_K = q_j$ with J linear restrictions $j = 1...J$ is jointly tested with the F-test. If the null hypotheses is true, the F-statistic follows an F distribution with J numerator degrees of freedom and $N-K-1$ denominator degrees of freedom:

$$F \sim F_{J,N-K-1}$$

This F-statistic cannot be easily expressed without matrix notation. For example, to perform a two-sided test of H_0 against the alternative hypotheses $H_a : r_{j1}\beta_1 + ... + r_{jK}\beta_K \neq q_j$ for all j at the 5% significance level, we calculate the F-statistic and compare it to the 0.95-quantile of the F-distribution. With $N = 30$, $K = 2$ and $J = 2$, H_0 is rejected if $F > 3.35$. We cannot perform one-sided F-tests.

Only under homoscedasticity (*OLS5a*), the F-statistic can also be computed as

$$F = \frac{(SSR_{restricted} - SSR)/J}{SSR/(N-K-1)} = \frac{(R^2 - R^2_{restricted})/J}{(1-R^2)/(N-K-1)} \sim F_{J,N-K-1}$$

where $SSR_{restricted}$ and $R^2_{restricted}$ are, respectively, estimated by restriced least squares which minimizes $S(\beta)$ s.t. $r_{j1}\beta_1 + ... + r_{jK}\beta_K \neq q_j$ for all j. Exclusionary restrictions of the form $H_0 : \beta_k = 0, \beta_m = 0, ...$ are a special case of $H_0 : r_{j1}\beta_1 + ... + r_{jK}\beta_K = q_j$ for all j. In this case, restricted least squares is simply estimated as a regression were the explanatory variables $k, m, ...$ are excluded.

7 Confidence Intervals in Small Samples

Assuming *OLS1*, *OLS2*, *OLS3a*, *OLS4*, and *OLS5a*, we can construct confidence intervals for a particular coefficient β_k. The $(1-\alpha)$ confidence interval is given by

$$\left(\beta_k - t_{(1-\alpha/2),(N-K-1)}\widehat{se}(\widehat{\beta}_k)\;,\;\beta_k + t_{(1-\alpha/2),(N-K-1)}\widehat{se}(\widehat{\beta}_k)\right)$$

where $t_{(1-\alpha/2),(N-K-1)}$ is the $(1-\alpha/2)$ quantile of the t-distribution with $N-K-1$ degrees of freedom. For example, the 95 % confidence interval with $N=30$ and $K=2$ is $\left(\beta_k - 2.052\widehat{se}(\widehat{\beta}_k)\;,\;\beta_k + 2.052\widehat{se}(\widehat{\beta}_k)\right)$.

8 Asymptotic Properties of the OLS Estimator

Assuming *OLS1*, *OLS2*, *OLS3d*, *OLS4*, and *OLS5a* or *OLS5b* the following properties can be established for large samples.

- The OLS estimator is consistent:

$$\text{plim}\ \widehat{\beta}_k = \beta_k$$

- The OLS estimator is asymptotically normally distributed

$$\sqrt{N}(\widehat{\beta}_k - \beta_k) \xrightarrow{d} N\left(0, \varsigma^2\right)$$

- The OLS estimator is approximately normally distributed

$$\widehat{\beta}_k \overset{A}{\sim} N\left(\beta_k, Avar(\widehat{\beta}_k)\right)$$

  where the asymptotic variance is $Avar(\widehat{\beta}) = \varsigma^2/N$. For the binary regression under *OLS5a* (homoscedasticity) it can be consistently estimated as

$$\widehat{Avar}(\widehat{\beta}_1) = \frac{\widehat{\sigma}^2}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

  with

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^N \widehat{u}_i^2}{N}\,.$$

  Under *OLS5b* (heteroscedasticity), $Avar(\widehat{\beta})$ can be consistently estimated as the *robust* or *Eicker-Huber-White* estimator (see handout on "Heteroscedasticity in the linear Model"). For the binary regression, the robust variance estimator is calculated as

$$\widehat{Avar}(\widehat{\beta}_1) = \frac{\sum_{i=1}^N \widehat{u}_i^2 (x_i - \bar{x})^2}{\left[\sum_{i=1}^N (x_i - \bar{x})^2\right]^2}$$

Note: In practice we can almost never be sure that the errors are homoscedastic and should therefore always use robust standard errors.

Table 1 provides a systematic overview of assumptions and their implied properties.

9 Asymptotic Tests

Assume *OLS1*, *OLS2*, *OLS3d*, *OLS4*, and *OLS5a* or *OLS5b*.

A simple null hypotheses of the form $H_0: \beta_k = q$ is tested with the z-test. If the null hypotheses is true, the z-statistic

$$z = \frac{\widehat{\beta}_k - q}{\widehat{se}(\widehat{\beta}_k)} \overset{A}{\sim} N(0,1)$$

follows approximately the standard normal distribution. The standard error is $\widehat{se}(\widehat{\beta}_k) = \sqrt{\widehat{Avar}(\widehat{\beta}_k)}$. For example, to perform a two sided test of H_0 against the alternative hypotheses $H_a: \beta_k \neq q$ on the 5% significance level, we calculate the z-statistic and compare its absolute value to the 0.975-quantile of the standard normal distribution. H_0 is rejected if $|z| > 1.96$.

A null hypotheses of the form $H_0: r_{j1}\beta_1 + ... + r_{jK}\beta_K = q_j$ with J linear restrictions $j = 1...J$ is jointly tested with the Wald test. If the null hypotheses is true, the Wald-statistic follows approximately an χ^2 distribution with J degrees of freedom:

$$W \overset{A}{\sim} \chi^2_J$$

This Wald statistic cannot be easily expressed without matrix notation. For example, to perform a test of H_0 against the alternative hypotheses $H_a : r_{j1}\beta_1 + ... + r_{jK}\beta_K \neq q_j$ for all j on the 5% significance level, we calculate the Wald statistic and compare it to the 0.95-quantile of the χ^2-distribution. With $J = 2$, H_0 is rejected if $W > 5.99$. We cannot perform one-sided Wald tests.

Under *OLS5a* (homoscedasticity) only, the Wald statistic can also be computed as

$$W = \frac{(SSR_{restricted} - SSR)}{SSR/N} = \frac{(R^2 - R^2_{restricted})}{(1-R^2)/N} \overset{A}{\sim} \chi^2_J$$

where $SSR_{restricted}$ and $R^2_{restricted}$ are, respectively, estimated by restricted least squares which minimizes $S(\beta)$ s.t. $r_{j1}\beta_1 + ... + r_{jK}\beta_K \neq q_j$ for all j. Exclusionary restrictions of the form $H_0 : \beta_k = 0, \beta_m = 0, ...$ are a special case of $H_0 : r_{j1}\beta_1 + ... + r_{jK}\beta_K = q_j$ for all j. In this case, restricted least squares is simply estimated as a regression were the explanatory variables $k, m, ...$ are excluded.

Note: the Wald statistic can also be calculated as

$$W = J \cdot F \overset{A}{\sim} \chi^2_J$$

where F is the small sample F-statistic. This formulation differs by a factor $(N - K - 1)/N$ but has the same asymptotic distribution.

10 Confidence Intervals in Large Samples

Assuming *OLS1*, *OLS2*, *OLS3d*, *OLS4*, and *OLS5a* or *OLS5b*, we can construct confidence intervals for a particular coefficient β_k. The $(1-\alpha)$ confidence interval is given by

$$\Big(\beta_k - z_{(1-\alpha/2)}\widehat{se}(\widehat{\beta}_k)\,,\, \beta_k + z_{(1-\alpha/2)}\widehat{se}(\widehat{\beta}_k)\Big)$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. For example, the 95 % confidence interval is $\Big(\beta_k - 1.96\widehat{se}(\widehat{\beta}_k)\,,\, \beta_k + 1.96\widehat{se}(\widehat{\beta}_k)\Big)$.

11 Small Sample vs. Asymptotic Properties

The t-test, F-test and confidence interval for small samples depend on the normality assumption *OLS3a* (see Table 1). This assumption is strong and unlikely to be satisfied. The asymptotic z-test, Wald test and the confidence interval for large samples rely on much weaker assumptions. Although most statistical software packages report the small sample results by default, we would typically prefer the large sample approximations. In practice, small sample and asymptotic tests and confidence intervals are very similar already for relatively small samples, i.e. for $(N - K) > 30$. Large sample tests also have the advantage that they can be based on heteroscedasticity robust standard errors.

12 Implementation in Stata 12

The multiple linear regression model is estimated by OLS with the `regress` command For example,

```
webuse auto.dta
regress mpg weight displacement
```

regresses the mileage of a car (`mpg`) on `weight` and `displacement`. A constant is automatically added if not suppressed by the option `noconst`

```
regress mpg weight displacement, noconst
```

Estimation based on a subsample is performed as

```
regress mpg weight displacement if weight>3000
```

where only cars heavier than 3000 lb are considered. Transformations of variables are included with new variables

```
generate logmpg = log(mpg)
generate weight2 = weight^2
regress logmpg weight weight2 displacement
```

The Eicker-Huber-White covariance is reported with the option `robust`

```
regress mpg weight displacement, vce(robust)
```

F-tests for one or more restrictions are calculated with the post-estimation command `test`. For example

```
test weight
```

tests $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$, and

```
test weight displacement
```

tests $H_0 : \beta_1 = 0$ and $\beta_2 = 0$ against $H_A : \beta_1 \neq 0$ or $\beta_2 \neq 0$

New variables with residuals and fitted values are generated by

```
predict uhat if e(sample), resid
predict pricehat if e(sample)
```


13 Implementation in R 2.13

The multiple linear regression model is estimated by OLS with the `lm` function. For example,

```
> library(foreign)
> auto <- read.dta("http://www.stata-press.com/data/r11/auto.dta")
> fm <- lm(mpg~weight+displacement, data=auto)
> summary(fm)
```

regresses the mileage of a car (`mpg`) on `weight` and `displacement`. A constant is automatically added if not suppressed by `-1`

```
> lm(mpg~weight+displacement-1, data=auto)
```

Estimation based on a subsample is performed as

```
> lm(mpg~weight+displacement, subset=(weight>3000), data=auto)
```

where only cars heavier than 3000 lb are considered. Tranformations of variables are directly included with the `I()` function

```
> lm(I(log(mpg))~weight+I(weight^2)+ displacement, data=auto)
```

The Eicker-Huber-White covariance is reported after estimation with

```
> library(sandwich)
> library(lmtest)
> coeftest(fm, vcov=sandwich)
```

F-tests for one or more restrictions are calculated with the command `waldtest` which also uses the two packages `sandwich` and `lmtest`

```
> waldtest(fm, "weight", vcov=sandwich)
```

tests $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$ with Eicker-Huber-White, and

```
> waldtest(fm, .~.-weight-displacement, vcov=sandwich)
```

tests $H_0 : \beta_1 = 0$ and $\beta_2 = 0$ against $H_A : \beta_1 \neq 0$ or $\beta_2 \neq 0$.

New variables with residuals and fitted values are generated by

```
> auto$uhat <- resid(fm)
> auto$mpghat <- fitted(fm)
```

References

Introductory textbooks

Stock, James H. and Mark W. Watson (2007), Introduction to Econometrics, 2nd ed., Pearson Addison-Wesley. Chapters 4 - 9.

Wooldridge, Jeffrey M. (2009), Introductory Econometrics: A Modern Approach, 4th ed., South-Western Cengage Learning. Chapters 2 - 8.

Advanced textbooks

Cameron, A. Colin and Pravin K. Trivedi (2005), Microeconometrics: Methods and Applications, Cambridge University Press. Sections 4.1-4.4.

Wooldridge, Jeffrey M. (2002), Econometric Analysis of Cross Section and Panel Data, MIT Press. Chapters 4.1 - 4.23.

Companion textbooks

Angrist, Joshua D. and Jörn-Steffen Pischke (2009), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press. Chapter 3.

Kennedy, Peter (2008), A Guide to Econometrics, 6th ed., Blackwell Publishing. Chapters 3 - 11, 14.