Tutorial 1: Linear Regression

Alexander Wintzéus*

February 6, 2024

In this tutorial, we will implement Ordinary Least Squares (OLS) estimation in `Stata`. To this end, we will be relying on an empirical study of Acemoglu, Johnson, and Robinson (2001). The exercises are collected in the first part of this document. The second part of this document provides a theoretical summary of the linear regression model and the Ordinary Least Squares principle.

1 Application

The basic question asked by Acemoglu et al. (2001) is the following: What are the fundamental causes of the large differences in income per capita across countries? The authors point out that differences in institutions and property rights may explain this pattern. To investigate the extent to which institutions explain differences in income, the authors propose and estimate the following linear model:

$$y_i = \mu + \alpha R_i + \mathbf{x}_i'\gamma + \varepsilon_i \tag{1}$$

where y_i denotes the log of income per capita in country i, R_i is a measure of protection against expropriation of private property[1], $\mathbf{x}_i$ is a vector of covariates, and ε_i is a random error term. The parameter of interest is α, the effect of institutions on (log) income per capita.

To replicate the authors' findings, we will be relying on data for the so-called *base sample*. The data contains a subset of the variables that are used in the paper, but should suffice for our purposes. The *base sample* consists of 64 countries that were ex-colonies and for which information on settler mortality, protection against expropriation risk, and GDP per capita (in 1995 PPP) is available. As a first step, load the data file `AJR2001-AER.dta` into `Stata` and get to know its contents by using the `browse` and `describe` commands.

1. Plot the empirical relationship between `loggdp` (log income per capita in 1995 PPP) and `risk` (average protection against expropriation risk) using the `scatter` command. Is the linearity

*KU Leuven, Department of Economics. alexander.wintzeus@kuleuven.be

[1]This variable captures the average risk of expropriation of private foreign investment by the government between 1985 and 1995. The values range from 0 to 10 and higher values mean *less* risk. It used as a proxy for the quality of institutions in a given country.

assumption reasonable? What does the plot tell us about the relationship between income and institutions?

2. Estimate equation (1) by OLS using the `regress` command. You can disregard any control variables for now.

   (a) Interpret the OLS estimate of the effect of `risk` on `loggdp`.

   (b) Use the `predict` command to compute the fitted values ($\hat{y}_i$). Compute the implied residuals (e_i).

   (c) Use the `scatter` command to obtain a residual plot. What can this plot tell us about the conditional homoskedasticity assumption?

3. Can you think of variables that are not included in the previous regression but that may be correlated with `risk`? If such variables are relevant determinants of income per capita, not including them in the regression could give rise to *omitted variable bias* and may seriously hamper the causal interpretation of our estimate.[2]

   (a) Compute the correlation coefficient between the variables `latitude` (standardized measure of distance from the equator) and `risk` using the `correlate` or `pwcorr` command. Based on the obtained figure and your intuition, do you think our previous estimate of α may be suffering from omitted variable bias?

   (b) Estimate equation (1) by OLS using the `regress` command. This time include `latitude` as a control variable. Did the estimate of α change according to your expectations?

   (c) Estimate equation (1) by OLS using the `regress` command. This time include the full set of control variables – that is, `latitude` and continent dummies.[3] Why would it be problematic to include a dummy for each continent? Interpret the OLS estimate of the effect of `risk` on `loggdp`.

4. Compare the coefficient of determination (R^2) for each of the previous regressions.

   (a) How do we interpret the coefficient of determination?

   (b) Why should we be careful when interpreting the coefficient of determination as a measure of fit? And consequently, when comparing these values across specifications?

5. Can you think of reasons why our current estimate of α could not be causal? Can you think of mechanisms that could potentially create endogeneity issues?

[2] The existence of such variables implies that the strict exogeneity assumption is violated. Consequently, the OLS estimator will provide biased estimates and is generally inconsistent.

[3] There are four continent dummies included in the data set. One for each of Africa, Asia, and America. The final dummy collects all other continents.

2 Theoretical summary

In this section, we briefly recapitulate the linear regression model and Ordinary Least Squares (OLS) estimation.

2.1 Linear regression model

In the linear regression model, a variable of interest (the *dependent* variable or *regressand*) is assumed to be related to several other variables (the *independent* variables or *regressors*). As with any other model, the linear regression model simply comprises a set of restrictions on the joint distribution of the dependent and independent variables.

To fix ideas, suppose that we have at our disposal a sample of n observations of the dependent and independent variables. Let y_i denote the value of the *dependent* variable for observation i. Similarly, let x_{ik} be the value for the i-th observation of *independent* variable k ($k = 1, \ldots, K$). We can collect the values of the independent variables for observation i into a single column vector $\mathbf{x}_i$. It is worthwhile to introduce more vector and matrix notation to ease the rest of the exposition. Let $\mathbf{y}$ be the $n \times 1$ vector of values for the dependent variable. Furthermore, let $\mathbf{X}$ denote $n \times K$ matrix of values for the independent variables. Both $\mathbf{y}$ and $\mathbf{X}$ contain as many rows as there are observations in the sample, with each row corresponding to a particular observation i. For this reason, they are sometimes referred to as the *data vector* and the *data or design matrix*, respectively.

The linear regression model consists of three core assumptions:

1. **Linearity:** The relationship between the dependent variable and the independent variables is linear. That is, for all observations i,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{2}$$

   where $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters and ε_i an unobserved, idiosyncratic error term. Using matrix notation, this assumption can be compactly summarized as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3}$$

2. **Strict exogeneity:** For all $i = 1, \ldots, n$

$$E[\varepsilon_i|\mathbf{X}] = 0 \tag{4}$$

   where E denotes the expectations operator. Note that this assumption implies that **(i)** the *un*conditional mean of the error term equals 0, and **(ii)** the error terms are orthogonal to the independent variables or regressors for *all* observations.

3. **No multicollinearity:** With probability 1,

$$\text{rank}(\mathbf{X}) = K. \tag{5}$$

This assumption implies that the columns of the design matrix $\mathbf{X}$ must be linearly independent. That is, it is not possible to write any of the independent variables as a linear combination of the others. As such, it further implies that there must be at least as many observations as regressors $n \geq K$. Essentially, the assumption is a rank condition for identification, because if it is not satisfied, it would not be possible to uniquely recover (or *identify*) all individual parameters from the data.

It can readily be verified that under these assumptions (strictly, assumptions 1 and 2 are sufficient)

$$E[y_i|\mathbf{X}] = \mathbf{x}_i'\boldsymbol{\beta} \qquad \text{for } i = 1, \ldots, n. \tag{6}$$

Often, an additional assumption regarding the second moments of the error terms is made:

4. **Spherical error variance:** For all $i, j = 1, \ldots, n$

$$E[\varepsilon_i^2|\mathbf{X}] = \sigma^2 \tag{7}$$

$$E[\varepsilon_i\varepsilon_j|\mathbf{X}] = 0. \tag{8}$$

The assumption consists of two parts. The first part states that the conditional second moment of the error term does not depend on $\mathbf{X}$ and equals $\sigma^2 > 0$. It is more commonly known as the (conditional) homoskedasticity assumption. The second part implies that the errors are uncorrelated across observations, conditional on $\mathbf{X}$. Using matrix notation, assumption 4 can be written more compactly as:

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{I}_n \tag{9}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Assumptions 1 through 4 constitute what is known as the *classical* linear regression model. We can relax assumption 4 so as to allow for (conditional) heteroskedasticity. This is discussed in section 2.3.

2.2 Ordinary Least Squares

To estimate the parameters of interest in the linear model, we can resort to the principle of Ordinary Least Squares (OLS). Concretely, OLS estimation seeks to find the parameter vector $\boldsymbol{\beta}$ that minimizes the sum of squared residuals. That is,

$$\boldsymbol{\beta}_{OLS} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \tag{P$_{OLS}$}$$

where

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{10}$$

$$= \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2. \tag{11}$$

The first-order condition characterizing the solution to P_{OLS} provides us with the so-called normal equations:

$$(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}. \tag{12}$$

The solution to these equations is the Ordinary Least Squares estimator of the parameter vector:[1]

$$\boldsymbol{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{13}$$

Algebraically, problem P_{OLS} can be envisaged as finding the vector $\mathbf{X}\boldsymbol{\beta}$ in the subspace spanned by the columns of the design matrix $\mathbf{X}$ that is at a minimum distance – in the Euclidean sense – from the data vector $\mathbf{y}$. The vector in the column space of $\mathbf{X}$ for which this is achieved is simply the orthogonal projection of the vector $\mathbf{y}$ onto that space. In other words, OLS regression boils down to decomposing the data vector $\mathbf{y}$ into two mutually orthogonal components: The systematic component $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}_{OLS}$ representing the OLS fitted or predicted values and the orthogonal component $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ representing the OLS residual.

Some properties of the OLS estimator $\boldsymbol{\beta}_{OLS}$:

1. **Unbiasedness:** Under assumptions 1 to 3:

$$E[\boldsymbol{\beta}_{OLS}|\mathbf{X}] = \boldsymbol{\beta}_0. \tag{14}$$

2. **Gauss-Markov theorem:** Under assumptions 1 to 4, the variance of the OLS estimator is given by:

$$\mathrm{Var}(\boldsymbol{\beta}_{OLS}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{15}$$

   Furthermore, within the class of linear unbiased estimators, the OLS estimator is *efficient*, i.e., it has the least variance (**BLUE**).

3. **OLS estimate σ^2:** An unbiased estimate for the error variance σ^2 is given by

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - K}. \tag{16}$$

[1] Note that we obtain a unique solution for the normal equations given that assumption 3 holds.

2.3 Heteroskedasticity and Generalized Least Squares

Assumption 4 posits that the $n \times n$ matrix of conditional second moments $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]$ is spherical, that is, proportional to the identity matrix. Relaxing this assumption would imply that each element of this matrix is, in general, a nonlinear function of the data $\mathbf{X}$. If the error terms are not (conditionally) homoskedastic, the diagonal elements of the matrix are generally not the same. Furthermore, if there would be correlation in the errors across observations, the off-diagonal elements are non-zero (e.g., if there is serial correlation in a time-series model). In this case, the OLS estimator of the parameter vector $\boldsymbol{\beta}$ loses some of its desirable properties. Although the OLS estimator is still unbiased, it no longer has the least variance within the class of linear unbiased estimators. Furthermore, the estimate for the variance of the OLS estimator is no longer a consistent estimate of the true underlying variance.

If the error terms are not conditionally homoskedastic – so that the variance of the errors need not be the same across observations – but are independent, then an important alternative estimator for the variance of the OLS estimator is given by White (1980):

$$\widehat{\text{Var}}_{HCE}(\beta_{OLS}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\big(\mathbf{X}'\,\text{diag}(\mathbf{e}^2)\mathbf{X}\big)(\mathbf{X}'\mathbf{X})^{-1} \tag{17}$$

where $\mathbf{e}^2$ denotes the $n \times 1$ vector of squared OLS residuals. This estimator is consistent, however, it may be biased if the data are effectively homoskedastic.

Suppose that in contrast to assumption 4, we believe that

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{V}(\mathbf{X}) \tag{18}$$

where the $n \times n$ matrix $\mathbf{V}(\mathbf{X})$ is assumed to be nonsingular and known. As mentioned above, in this case, the OLS estimator of the parameter vector $\boldsymbol{\beta}$ is no longer the **BLUE**. However, an alternative estimator, known as the Generalized Least Squares (GLS) estimator is. Without going through the derivations, the GLS estimator of the parameter vector and its variance are given by:

$$\boldsymbol{\beta}_{GLS} = \big(\mathbf{X}'\mathbf{V}(\mathbf{X})^{-1}\mathbf{X}\big)^{-1}\mathbf{X}'\mathbf{V}(\mathbf{X})^{-1}\mathbf{y}, \tag{19}$$

$$\text{Var}(\boldsymbol{\beta}_{GLS}|\mathbf{X}) = \sigma^2\big(\mathbf{X}'\mathbf{V}(\mathbf{X})^{-1}\mathbf{X}\big)^{-1}. \tag{20}$$

An important caveat of the GLS estimator is that generally $\mathbf{V}(\mathbf{X})$ is not known. However, there exists an implementable version of GLS known as Feasible Generalized Least Squares (FGLS). FGLS proceeds in two steps. In a first step, the model is estimated using some consistent but inefficient estimator (e.g., OLS). The residuals from this first step are then used to build a consistent estimator of $\mathbf{V}(\mathbf{X})$. In the second step, one can estimate the model again, but using GLS.

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, *91*(5), 1369-1401.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817-838.