Section B1: Linear Regression

B1.1. Introduction

Linear regression, also known as ordinary least square (OLS), is a method of analyzing linear relationships between variables. Particular methods used depend on the type of data set to be analyzed. Types of data sets used in regression analyses include the following:

- A *cross-sectional* data set comprises records for a sample of businesses, individuals, households, cities, states, etc. measured for a time period called a reporting or *reference period* (e.g., a day, week, or month). Differences between members of a cross-sectional dataset are treated as random.
- A *time series* data set comprises observations on a variable or collection of variables for a sequence of different reference periods, e.g., months or years.
- A *pooled cross-sectional* data set comprises records from two or more disjoint cross sectional data sets, where each cross sectional data set may have a different temporal reference period.
- A *panel* or *longitudinal* data set consists of a time series for a collection of cross-sectional members. The cross-sectional members, sometimes called a cohort, remain the same for all reference periods.

The basic OLS estimation and diagnostic techniques were originally developed for use with cross-sectional data sets. Methods were extended for use with other types of data. Time series data presents special problems, e.g., it may be difficult to tell whether or not any changes over time can safely be treated as random. Specialized techniques have been developed for time-series regression. These techniques exploit the across-time correlation in time series data and address concerns regarding "spurious regression" (the appearance of correlation between series that merely exhibit a similar trend). Specialized time-series regression techniques include co-integration analysis with error-correction models and dynamic linear regression models under the Bayesian paradigm.

B1.2. Estimating Regression Coefficients

We assume that a dependent (response) variable y varies with a set of independent explanatory variables $x_1, \ldots, x_m$. The variation in y has two components:

- A *systematic* component that can be modeled as a linear function of the x variables.
- A *random* component that is unexplained and unrelated to any variation in the x variables. The random component is represented by an error term ε.

The linear regression model has the general form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon, \tag{B1.2.1}$$

where $\beta_0, \beta_1, \ldots, \beta_m$ are fixed, unknown parameters, and ε is a normal (Gaussian) random variable representing the random component of the variation in y. Conditional on the explanatory variables x_j, we assume the following:

1. The expected value $E(\varepsilon_i) = 0$ for all i.
2. All of the ε_i have the same variance, i.e., $Var(\varepsilon_i) = \sigma^2$ for all i.
3. For $j \neq i$, the error terms ε_i and ε_j are independent, i.e., $E(\varepsilon_i, \varepsilon_j) = 0$.

We also assume that the linear functional form B1.2.1 is appropriate and that none of the dependent variables x_j may be expressed as a linear combination of the others.

We apply linear regression analysis to a set of n observed values of a dependent variable y and $m \geq 1$ associated independent variables x_j to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_m$. For $i \in \{1, \ldots, n\}$ we assume that the observed value y_i can be estimated as a linear function $\hat{y}_i$ of the independent variables and the estimated parameters:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_m x_{i,m}. \tag{B1.2.2}$$

When using the model for prediction, we compute $\hat{y}_k \notin \{\hat{y}_1, \ldots, \hat{y}_n\}$, where y_k is unknown and the covariates $x_{k,1}, \ldots, x_{k,m}$ are known or predicted by other means.

The vector of estimators $\widehat{\boldsymbol{\beta}}$ that minimizes the sum of squared prediction errors $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$. It can be shown that

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{B1.2.3}$$

where $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_m \end{bmatrix}$, and $\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,m} \end{bmatrix}$.

Although the matrix $\mathbf{X}'\mathbf{X}$ is always square, it need not have a unique inverse. When $\mathbf{X}'\mathbf{X}$ is not of full rank, we cannot use OLS to estimate the parameters. In this case, we must use generalized matrix inverses, and it may not be possible to estimate all the parameters.

B1.3. Regression Diagnostics

Before using the results of a regression analysis, we compute test statistics that indicate whether or not the data support the assumptions embedded in the OLS model. For $i \in \{1, \ldots, n\}$, let $\hat{y}_i$ be as defined in equation B1.2.2. The difference $\hat{\varepsilon}_i \equiv y_i - \hat{y}_i$ is known as the i^{th} *residual* and plays an important role in model evaluation. Some fundamental test statistics are based on the *partitioned sum of squares*. Let

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{B1.3.1}$$

It can be shown that the sum of squares $\sum_{i=1}^{n}(y_i - \bar{y})^2$ can be partitioned into two components representing Model Error ($\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$) and Residual Error ($\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$):

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2. \qquad \text{(B1.3.2)}$$

Total Error = Model Error + Residual Error

The Model Error term $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ is the sum of the squared residuals.

Mean Squared Error

Hypothesis tests on the regression parameters use the mean squared error (MSE or s^2), which is the sum of the squared residuals divided by the model's degrees of freedom:

$$s^2=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-m-1}=\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}}{n-m-1}. \qquad \text{(B1.3.3)}$$

The model's degrees of freedom is the sample size n minus the number of parameters estimated, $m+1$. To compute the MSE, we must have $n>m+1$. It can be shown that the MSE is an unbiased estimator of σ^2, the variance of the ε's.

Coefficient of Variation R^2

The coefficient of determination or R^2 statistic for a regression model is

$$R^2\equiv 1-\frac{\text{Model Error}}{\text{Total Error}}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \qquad \text{(B1.3.4)}$$

R^2 (sometimes called multiple R^2) indicates the proportion of the total error that the independent variables explain. Because the R^2 statistic may be inflated by the use of a large number of explanatory variables, the adjusted R^2 (R_A^2), which corrects for the number of degrees of freedom, is generally preferred:

$$R_A^2\equiv 1-\frac{\sum_{i=1}^{n}\left[(y_i-\hat{y}_i)^2\Big/(n-m-1)\right]}{\sum_{i=1}^{n}\left[(y_i-\bar{y})^2\Big/(n-1)\right]}=1-\frac{(n-1)s^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \qquad \text{(B1.3.5)}$$

Variances of the Regression Coefficients β_j

In order to compute test statistics, we estimate the variances and standard errors of the regression coefficients. When there is only one x variable, the variances of β_0 and β_1 are estimated as follows:

$$s_{\hat{\beta}_0}^2=\frac{s^2\sum_{i=1}^{n}x_{i,1}^2}{n\sum_{i=1}^{n}\left(x_{i,1}-\bar{x}_1\right)^2},\ \text{where } \bar{x}_1=\frac{1}{n}\sum_{i=1}^{n}x_{i,1},. \qquad \text{(B1.3.6)}$$

and

$$s_{\hat{\beta}_1}^2 = \frac{s^2}{\sum_{i=1}^{n}\left(x_{i,1} - \bar{x}_1\right)^2}. \tag{B1.3.7}$$

When there are multiple explanatory variables, we estimate the covariance matrix of the vector $\widehat{\boldsymbol{\beta}}$ as

$$\boldsymbol{s}_{\widehat{\boldsymbol{\beta}}}^{\boldsymbol{2}} = s^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{B1.3.8}$$

Test for the significance of a parameter β_j
For $j \in \{0,1,\ldots,m\}$, suppose we wish to test the hypothesis

$$H_0\colon\ \beta_j = 0 \quad \text{vs.} \quad H_A\colon\ \beta_j \neq 0. \tag{B1.3.9}$$

For this two-sided hypothesis test, we may use the test statistic

$$t_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}. \tag{B1.3.10}$$

For large samples, under the null hypothesis, the density of t_j is approximately normal (Gaussian) with mean 0 and variance 1. For $Z \sim N(0,1)$, $P(|Z| > 1.96) \approx 0.05$. Thus when testing B1.3.9 with $\alpha = 0.05$, we generally reject H_0 when $t_j \geq 2$. This means that the explanatory variable x_j explains a significant amount of the variation in the dependent variable y.

Weighted Least Squares and Robust Standard Errors
The linear regression setup is based on several assumptions. We assume, for example, that the linear functional form is appropriate and that the model's systematic component incorporates all of the important explanatory variables. Also, we assume that, for all $i \in \{1,\ldots,n\}$,

a) $E(\varepsilon_i) = 0$ and
b) $Var(\varepsilon_i) = \sigma_i^2 = \sigma^2$.

The assumption of constant variances (assumption b above), is called *homoscedasticity*. When this assumption is violated, the data are *heteroscedastic,* and the following problems occur:

- The least squares estimator (B1.2.3) is not a Best Linear Unbiased Estimator (BLUE). It is still unbiased, but there is another estimator with lower variance.
- The standard errors of the estimated coefficients, calculated by formula B1.3.8, may be incorrect; this can lead to erroneous conclusions from hypothesis tests.

The Breusch-Pagen (BP) test is a χ^2 test that tests the null hypothesis of homoscedasticity against the alternative hypothesis that the data are heteroscedastic. Other homoscedasticity tests

are also discussed in the statistics literature. For the BP test, we use an auxiliary regression equation to regress the residuals ($\hat{\varepsilon}_i$) on the independent variables and compute the R^2 statistic $\tilde{R}^2$ from this regression. Under the null hypothesis, the test statistic $n\tilde{R}^2$ has a χ^2 distribution with m degrees of freedom. If the p-value from the χ^2 test falls below 0.05, we reject the null hypothesis and conclude that heteroscedasticity is present in the data.

To find the BLUE in the presence of heteroscedasticity, we may use weighted least squares (WLS), which is an example of generalized least squares (GLS). For $i \in \{1, \dots, n\}$, let $\mathbf{x}_i$ denote the values of the explanatory variables for observation i, and let $w_i(\mathbf{x}_i)$ be a function of $\mathbf{x}_i$ such that

$$Var(\varepsilon_i|\mathbf{x}_i) = \sigma^2 w_i(\mathbf{x}_i). \tag{B1.3.11}$$

Then

$$E\left(\frac{\varepsilon_i}{\sqrt{w_i(\mathbf{x}_i)}}\middle|\mathbf{x}_i\right) = 0 \text{ and } Var\left(\frac{\varepsilon_i}{\sqrt{w_i(\mathbf{x}_i)}}\middle|\mathbf{x}_i\right) = \sigma^2. \tag{B1.3.12}$$

Setting $w_i = w_i(\mathbf{x}_i)$, we divide equation B1.2.1 by $\sqrt{w_i}$ to obtain, for $i \in \{1, \dots, n\}$,

$$\frac{y_i}{\sqrt{w_i}} = \frac{\beta_0}{\sqrt{w_i}} + \beta_1\left(\frac{x_{i,1}}{\sqrt{w_i}}\right) + \cdots + \beta_m\left(\frac{x_{i,m}}{\sqrt{w_i}}\right) + \frac{\varepsilon_i}{\sqrt{w_i}}, \tag{B1.3.13}$$

or

$$\tilde{y}_i = \beta_0\tilde{x}_{i,0} + \beta_1\tilde{x}_{i,1} + \cdots + \beta_m\tilde{x}_{i,m} + \tilde{\varepsilon}_i, \tag{B1.3.14}$$

where $\tilde{y}_i = \frac{y_i}{\sqrt{w_i}}$, $\tilde{x}_{i,0} = \frac{1}{\sqrt{w_i}}$, $\tilde{\varepsilon}_i = \frac{\varepsilon_i}{\sqrt{w_i}}$, and, for $j \in \{1, \dots, m\}$, $\tilde{x}_{i,j} = \frac{x_{i,j}}{\sqrt{w_i}}$. For the transformed equation B1.3.14, we have $Var(\tilde{\varepsilon}_i) = \sigma^2$ for all $i \in \{1, \dots, n\}$, so the homoscedasticity assumption is valid. The regression results, however, must be interpreted using the original model B1.2.1.

The use of WLS assumes the availability of an appropriate set of weights w_i. An alternative is to use robust standard errors (also called heteroscedasticity-consistent standard errors or Huber-White standard errors) when testing B1.3.9. The robust variance is estimated as

$$s^2_{H,\hat{\beta}_j} = \frac{\sum_{i=1}^n \hat{r}^2_{i,j}\hat{\varepsilon}^2_i}{SSR^2_j}, \tag{B1.3.15}$$

where $\hat{r}_{i,j}$ is the i^{th} residual from the auxiliary regression with x_j as the dependent variable and all the other x's as independent variables, and SSR^2_j is the sum of the squared residuals ($\hat{\varepsilon}_i$'s) from this auxiliary regression. For example, in a model with only one independent variable x, the robust variance estimate for $\hat{\beta}_1$ is

$$s^2_{H,\hat{\beta}_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{B1.3.16}$$

Diagnostics such as t-statistics computed using the robust standard errors provide accurate results in the presence of heteroscedasticity.

B1.4. Instrumental Variables and Two-stage Lease Squares (2SLS)Regression Diagnostics

In order for OLS estimator $\widehat{\boldsymbol{\beta}}$ (shown in B1.2.3) to be consistent, the explanatory variables $x_1, \ldots, x_m$ must be *exogenous*, i.e., they must be uncorrelated with the error term ε. When one or more of the explanatory variables is correlated with ε, they are jointly determined with the dependent variable y, and the use of two-stage least squares (2SLS) provides consistent parameter estimates for the model B1.2.1. The 2SLS technique, which employs *instrumental variables* to compensate for deficiencies in the OLS model, is also useful when important variables determining y cannot be accurately measured.

Consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \tag{B1.4.1}$$

and suppose that the explanatory variable x_2 is endogenous or cannot be accurately measured. If we perform OLS with x_2 omitted, it will be incorporated into the error term ε, and the coefficients β_0 and β_1 (computed as in B1.2.3) may be biased. A new variable z that is correlated with x_1 but not with ε may be used as an instrumental variable (IV) for x_1. We test the correlation between x_1 and z by fitting the model

$$x_1 = \gamma_0 + \gamma_1 z + \zeta, \tag{B1.4.2}$$

where ζ is a normally distributed random error term, and testing the hypothesis H_0: $\gamma_1 = 0$ vs. H_A: $\gamma_1 \neq 0$.

The OLS model with x_2 omitted is

$$y = \beta_0 + \beta_1 x_1 + \varepsilon. \tag{B1.4.3}$$

We compute the covariance of each term in B1.4.3 with the instrumental variable z:

$$Cov(y, z) = \beta_1 Cov(x_1, z) + Cov(\varepsilon, z), \tag{B1.4.4}$$

where we have used the fact that the covariance of x_1 with the constant β_0 is zero. Because we assume that $Cov(\varepsilon, z) = 0$, equation B1.4.4 yields the IV estimator of β_1:

$$\hat{\beta}_1 = \frac{Cov(y, z)}{Cov(x_1, z)}, \tag{B1.4.5}$$

which we can estimate from the data. In the special case that $x_1 = z$, the $\hat{\beta}_1$ in B1.4.5 reduces to the usual OLS estimator. The IV technique may be extended to the multiple regression setting in a straightforward manner.

B1.5. The Two-stage Least Squares (2SLS) Procedure

In 2SLS, we assume that at least one instrumental variable is available for each endogenous explanatory variable in the OLS model (B1.2.1). Consider the model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \gamma_1 \ddot{y}_1 + \cdots + \gamma_p \ddot{y}_p + \varepsilon, \quad \text{(B1.5.1)}$$

where the variables $x_1, \ldots, x_m$ are exogenous, and the variables $\ddot{y}_1, \ldots, \ddot{y}_p$ are endogenous. For each endogenous explanatory variable $\ddot{y}$, we combine the available IV's into a single IV by running an auxiliary regression of the form

$$\ddot{y} = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \gamma_q z_q + \zeta, \quad \text{(B1.5.2)}$$

where q is the number of IV's for $\ddot{y}$ and ζ is a normally-distributed error term. In practice, we may include the exogenous explanatory variables in the OLS model B1.5.1, as well as all additional IV's, as independent variables (z's) in equation B1.5.2. Fitting the model B1.5.2 for each endogenous explanatory variable is the first stage of 2SLS.

After fitting model B1.5.1, we compute the model's fitted values for each observation:

$$\hat{\ddot{y}} = \hat{\gamma}_0 + \hat{\gamma}_1 z_1 + \hat{\gamma}_2 z_2 + \cdots + \hat{\gamma}_p z_p, \quad \text{(B1.5.3)}$$

where we have suppressed the subscript indicating an individual observation. In the second stage of 2SLS, we fit model B1.5.1 with the estimated values $\hat{\ddot{y}}$ substituted for the observed $\ddot{y}$ values of the exogenous explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \gamma_1 \hat{\ddot{y}}_1 + \cdots + \gamma_p \hat{\ddot{y}}_p + \varepsilon. \quad \text{(B1.5.4)}$$

Although this is the second stage of 2SLS, the procedure should not be performed "by hand" in two independent stages, because the second stage model fitting will yield residuals and variance estimates that don't account for the variance of the error term ζ in equation B1.5.2. Statistical software packages with 2SLS functions automatically correct these residuals and the associated diagnostic statistics.

Because of the additional random error term in the 2SLS procedure, the variances of 2SLS coefficients always exceed those of the OLS coefficients computed from formula B1.2.3. In the case of endogenous explanatory variables, however, 2SLS provides consistent parameter estimates, whereas the OLS estimates are inconsistent.

B1.6. Regression with Time Series Data

A time series is a set of observations $x_1, \ldots, x_n$, where, for $t \in \{1, \ldots, n\}$, x_t is the value or set of values observed for time period t. We assume that the data are generated by a stochastic process, and thus contain random errors, and that the time intervals are equal and consecutive. Regression on time series data differs from regression on cross-sectional data because of the following considerations:

1. Unlike cross-sectional data, time series data are ordered.
2. While we assume that each cross-sectional observation is uncorrelated with the other observations, time series data are generally autocorrelated.

The linear time-series regression model has the general form

$$y_t = \beta_0 + \beta_1 x_{t,1} + \cdots + \beta_m x_{t,m} + \varepsilon_t, \tag{B1.6.1}$$

where, for $t \in \{1, \ldots, n\}$, the error term ε_t is independent of the explanatory variables x_t for all t, and $\varepsilon_t \sim N(0, \sigma^2)$, conditional on the explanatory variables $x_{t,j}$.

In time series regression, all of the assumptions detailed in section B1.2 apply, and we add a further assumption regarding the error terms ε_t: we assume that, conditional on the sample data, error terms associated with different time periods are uncorrelated:

$$Corr(\varepsilon_t, \varepsilon_s) = 0 \text{ for all } s \neq t. \tag{B1.6.2}$$

Time series regression also requires additional attention to independent variables that may be correlated with the error term. For an explanatory variable x_t to be considered exogenous, it cannot react to any past or future changes in the dependent variable y_t.

Under these assumptions, we may apply OLS to time series data, estimating coefficients and diagnostic statistics as described in section B1.

Dummy Variables

Time series regression models often include binary variables, known as dummy variables, that represent events. A dummy variable takes a value of 1 if the event occurred and a value of 0 otherwise. Common uses of dummy variables include the following:

1. Reducing outlier effects by including a dummy variable that takes a value of 1 only in the time period coinciding with a series outlier.
2. Accounting for seasonality by including monthly or quarterly dummy variables.
3. Representing permanent changes, e.g., policy changes, that influence the dependent variable. In this case, the value of the dummy variable is 0 for all time periods prior to the change and 1 for all time period after the change.

Time Trends

Many time series contain trends that can cause regression diagnostics to erroneously indicate causal relationships between series. Two unrelated series with similar trends can appear

correlated, leading to the "spurious regression problem." In some cases we may alleviate this problem by including a trend term in the model. Adding a trend term αt to equation B1.6.1 provides the model

$$y_t = \alpha t + \beta_0 + \beta_1 x_{t,1} + \cdots + \beta_m x_{t,m} + \varepsilon_t, \tag{B1.6.3}$$

where α is a trend parameter. When the series y_t contains both upward and downward trends, a quadratic trend term may be added to the model:

$$y_t = \alpha_1 t + \alpha_2 t^2 + \beta_0 + \beta_1 x_{t,1} + \cdots + \beta_m x_{t,m} + \varepsilon_t. \tag{B1.6.4}$$

The addition of cubic and higher-ordered trend terms is allowed but usually not recommended, because too many polynomial trend terms may obscure the significance of the explanatory variables $x_{t,j}$.

Autocorrelation in Time-series Regression Residuals

To test that assumption B3.2.2 holds for a particular regression model, some regression software packages compute the Durbin-Watson (D-W) statistic:

$$d = \frac{\sum_{i=2}^{n} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}. \tag{B1.6.5}$$

The D-W statistic compares the squared differences of adjacent residuals to the squared residuals. If the squared differences are small relative to the squared residuals, resulting in a low value of d, there is evidence of first-order autocorrelation. Critical values for the D-W statistic depend on n and m. The test is for first-order autocorrelation only, i.e., it tests only for correlation between residuals from adjacent time periods.

One way to reduce the autocorrelation of time-series regression residuals is to include a lagged dependent variable as an independent variable in the model. Model B1.6.1 then becomes

$$y_t = \beta_0 + \beta_{-1} y_{t-1} + \beta_1 x_{t,1} + \cdots + \beta_m x_{t,m} + \varepsilon_t. \tag{B1.6.6}$$

For models that include a lagged dependent variable as an independent variable, the D-W statistic underestimates autocorrelation. For these models, Durbin proposed the h-statistic, which is a bias-adjusted version of the D-W statistic. It follows a normal distribution for large samples. Let β_{-1} be the regression coefficient of y_{t-1}. When the variance $s_{\beta_{-1}}^2 < \frac{1}{n}$,

$$h = \left(1 - \frac{d}{2}\right) \sqrt{\frac{n}{1 - n s_{\beta_{-1}}^2}}. \tag{B1.6.7}$$

Because the D-W statistic is difficult to interpret, Wooldridge (2018) suggests the alternative of performing a t-test on the estimated first-order correlation coefficient of the OLS residuals $\hat{\varepsilon}_t$. That is, we first run the OLS regression B1.6.1 and then run a second regression of $\hat{\varepsilon}_t$ on $\hat{\varepsilon}_{t-1}$.

The t-statistic B1.3.10 of the coefficient of $\hat{\varepsilon}_{t-1}$ will indicate the significance of first-order autocorrelation in the OLS residuals. The t-statistic is asymptotically normal and may be interpreted in the usual manner.