Estimation and Inference for Ordinary Least Squares Regression

- A *simple linear regression* is the special case of an OLS regression model with a single predictor variable.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

- For the ith observation we will denote the regression model by

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{2}$$

- For the random sample Y_1, Y_2, ..., Y_n we can estimate the parameters β_0 and β_1 by minimizing the sum of the squared errors,

$$\text{Min} \sum_{i=1}^{n} \epsilon_i^2 \tag{3}$$

which is equivalent to minimizing

$$\text{Min} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2. \tag{4}$$

Estimators and Estimates for Simple Linear Regression

- The estimators for β_0 and β_1 can be computed analytically and are given by

$$\widehat{\beta}_1 = \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{(X_i - \bar{X})^2} = \frac{\mathsf{Cov}(Y, X)}{\mathsf{Var}(X)} \tag{5}$$

and

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} \tag{6}$$

- The regression line always goes through the *centroid* $(\bar{X}, \bar{Y})$.

- We refer to the formulas for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as the *estimators* and the values that these formulas can take for a given random sample as the *estimates*. In statistics we put *hats* on all estimators and estimates.

- Given $\widehat{\beta}_0$ and $\widehat{\beta}_1$ the *predicted value* or *fitted value* is given by

$$\hat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X. \tag{7}$$

Estimation - The General Case

- We seldom build regression models with a single predictor variable. Typically we have multiple predictor variables denoted by X_1, X_2, ..., X_k, and hence the standard regression case is sometimes referred to as *multiple regression* in introductory regression texts.

- We can still think about the estimation of β_0, β_1, β_2, ..., β_k in the same manner as the simple linear regression case

$$\text{Min} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \ldots - \beta_k X_{ki})^2, \tag{8}$$

but the computations will be performed as matrix computations.

General Estimation - Matrix Notation

Before we set up the matrix formulation for the OLS model, let's begin by defining some matrix notation.

- The error vector $\boldsymbol{\epsilon} = [\epsilon_1 \cdots \epsilon_n]^T$.
- The response vector $\mathbf{Y} = [Y_1 \cdots Y_n]^T$.
- The design matrix or predictor matrix $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_1\ \mathbf{X}_2 \cdots \mathbf{X}_k]$.
- The parameter vector $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2 \cdots \beta_k]^T$.
- All vectors are column vectors, and the superscript T denotes the vector or matrix *transpose*.

- We minimize the sum of the squared error by minimizing $S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ which can be re-expressed as
$$S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{9}$$

- Taking the matrix derivative of $S(\boldsymbol{\beta})$, we get
$$S_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}) = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}}. \tag{10}$$

- Setting the matrix derivative to zero, we can write the expression for the *least squares normal equations*
$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}, \tag{11}$$
which yield the estimator
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{12}$$

- The estimator form $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ assumes that the inverse matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ exists and can be computed. In practice your statistical software will directly solve the normal equations using a QR factorization.

Statistical Inference with the t-Test

- In OLS regression the statistical inference for the individual regression coefficients can be performed by using a t-test.

- When performing a t-test there are three primary components: (1) Stating the null and alternate hypotheses, (2) Computing the value of the test statistic, and (3) Deriving a statistical conclusion based on a desired significance level.

- Step 1: The *null* and *alternate* hypotheses for β_i are given by

$$H_0 : \beta_i = 0 \quad \text{versus} \quad H_1 : \beta_i \neq 0. \tag{13}$$

Statistical Inference with the t-Test - Continued

- Step 2: The t statistic for β_i is computed by

$$t_i = \frac{\widehat{\beta}_i}{\mathsf{SE}(\widehat{\beta}_i)} \tag{14}$$

  and has a degrees of freedom equal to the sample size minus the number of model parameters, i.e. df = n - dim(Model). For example if you had a regression model with two predictor variables and an intercept estimated on a sample of size 50, then the t statistic would have 47 degrees of freedom.

- Step 3: *Reject* H_0 or *Fail to Reject* H_0 based on the value of your t statistic and your significance level. This decision can be made by using the p-value of your t statistic or by using the critical value for your significance level.

Confidence Intervals for Parameter Estimates

An alternative to performing a formal hypothesis test is to use a confidence interval for your parameter estimate. There is a duality between confidence intervals and formal hypothesis testing for regression parameters.

- The confidence interval for $\widehat{\beta}_i$ is given by
$$\widehat{\beta}_i \pm t(df, \alpha/2) * \mathsf{SE}(\widehat{\beta}_i), \tag{15}$$
where $t(df, \alpha/2)$ is a t value from a theoretical t distribution, not a t statistic value.

- If the confidence interval does not contain zero, then this is equivalent to rejecting the null hypothesis $H_0 : \beta_i = 0$.

The phrase *predicted value* is used in statistics to refer to the in-sample *fitted values* from the estimated model or to refer to the out-of-sample *forecasted values*. The dual use of this phrase can be confusing. A better habit is use the phrases *in-sample fitted values* and the *out-of-sample predicted values* to clearly reference these different values.

- Given $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ the vector of fitted values can be computed by $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{Y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. The matrix $\mathbf{H}$ is called the *hat matrix* since it puts the hat on $\mathbf{Y}$.

- The point estimate $\widehat{Y}_0$ at the point $\mathbf{x}_0$ can be computed by $\widehat{Y}_0 = \mathbf{x}_0^T\widehat{\boldsymbol{\beta}}$.

Statistical Intervals for Predicted Values - Continued

- The *confidence interval for an in-sample point* $\mathbf{x}_0$ *on the estimated regression function* is given by

$$\mathbf{x}_0^T \widehat{\boldsymbol{\beta}} \pm \widehat{\sigma} \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}. \tag{16}$$

- The *prediction interval for the point estimate* $\widehat{Y}_0$ for an out-of-sample $\mathbf{x}_0$ is given by

$$\mathbf{x}_0^T \widehat{\boldsymbol{\beta}} \pm \widehat{\sigma} \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}. \tag{17}$$

- Note that the out-of-sample prediction interval is always wider than the in-sample confidence interval.

Further Notation and Details

In order to compute the t statistic you need the standard error for the parameter estimate. Most statistical software packages should provide this estimate and compute this t statistic for you. However, it is always a good idea to know from where this number comes. Here are the details needed to compute the standard error for $\widehat{\beta}_i$.

- The estimated parameter vector $\widehat{\boldsymbol{\beta}}$ has the covariance matrix given by

$$\text{Cov}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2 \mathbf{X}^T \mathbf{X}, \tag{18}$$

  where

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-k-1}. \tag{19}$$

- The variance of $\widehat{\boldsymbol{\beta}}_i$ is the ith diagonal element of the covariance matrix

$$\text{Var}(\widehat{\boldsymbol{\beta}}_i) = \widehat{\sigma}^2 (\mathbf{X}^T \mathbf{X})_{ii}. \tag{20}$$