MA 575 Linear Models:

Cedric E. Ginestet, Boston University

Ordinary Least Squares Estimation

Week 2, Lecture 2

1 System of Linear Equations

1.1 Under- and Over-determined Systems

One can think of the problem of estimating the parameters of a regression model as solving a **system of linear equations**, where we have n **cases**; and p **parameters**. So far, we have seen the case of $p = 2$, where β_0 and β_1 are the parameters of interest. Such a system of equation may be represented as follows:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 \\ y_2 &= \beta_0 + \beta_1 x_2 \\ \vdots \quad & \quad \vdots \\ y_n &= \beta_0 + \beta_1 x_n. \end{aligned}$$

In a system of linear equations, each equation corresponds to a **line** in a p-dimensional space. Thus, the parameters become the quantities that vary. As a result, in order to apply the theory of systems of linear equations to regression, we need to change our viewpoint, by treating the *points* in the context of linear regression as *lines* in the context of a system of linear equations. We can best adopt this viewpoint by inverting the ordering of the terms in these equations, such that

$$\begin{aligned} \beta_0 &= y_1 - x_1 \beta_1 \\ \beta_0 &= y_2 - x_2 \beta_1 \\ \vdots \quad & \quad \vdots \\ \beta_0 &= y_n - x_n \beta_1. \end{aligned}$$

This leads to a **dual** representation of the set of points that we are considering. This dual representation is provided in figure 1, on page 2. The **solution set** of a system of linear equations depends on the number of unknowns and the number of equations in the system. As before,

i. Let p denote the number of **unknowns** or **parameters**,

ii. Let n denote the number of **equations** or **data points**.

Depending on the relationship between n and p, the solution set may fall into one of three situations:

i. **Under-determined** system ($n < p$): The solution set is *generally* infinite.

ii. **Determined** system ($n = p$): There *generally* exists a unique solution.

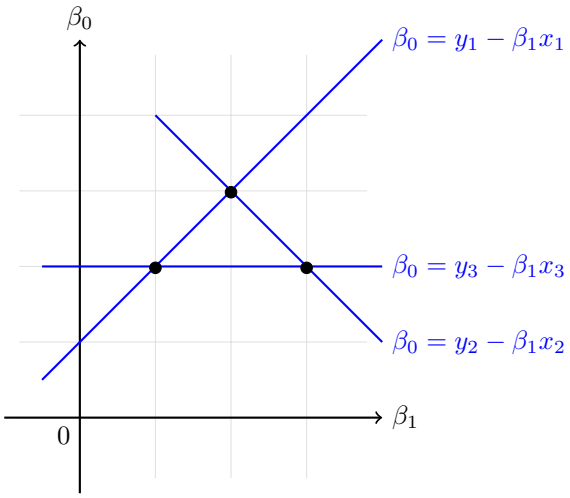

Figure 1. Representation of a regression problem with two parameters (β_0, β_1) and three points (x_1, y_1), (x_2, y_2), and (x_3, y_3). In this dual representation, the points determine the equations of three lines, and the Cartesian coordinates are given by the parameters (β_0, β_1). This particular system of linear equations is **over-determined**, since there are more equations than parameters. Moreover, the three lines do not intersect at a single point, and therefore this system has no solution. See also example 1.

iii. **Over-determined** system ($n > p$): There is *generally* no solution.

The description of the solution sets in these three cases only applies to general situations. For a system to be determined, in case (ii) for instance, the equations in the system must form a **consistent** system. An inconsistent system of equations is one for which the lines determined by the equations are **parallel** (e.g. $\beta_0 = 1 - 2\beta_1$ and $\beta_0 = 2 - 2\beta_1$, for instance). In such a situation, a determined system would not have any solution. Similarly, if some over-determined systems is composed of equations that are **linearly dependent**, then this system may have a single solution. When the equations of a system are linearly dependent, removing one of them increases the size of the solution set.

1.2 Homogeneous and Non-homogeneous

In linear regression, we are assuming that we have more data points than parameters, or **more equations than unknowns**. Therefore, we are considering an **over-determined system** of linear equations. Moreover, recall that a system of linear equations may be either homogeneous or non-homogeneous:

i. **Homogeneous**: There exists an *all-zero* solution –i.e. $\boldsymbol{\beta} = \mathbf{0}$.

ii. **Non-homogeneous**: There does not exist an *all-zero* solution.

Of course, within the context of regression, we are necessarily dealing with non-homogeneous systems of linear equations, since the constant terms are the y_i's, which cannot be all zeros; as this would constitute a trivial regression problem. Altogether, estimating the parameters in a regression model is therefore equivalent to solving a **non-homogeneous, overdetermined** system of linear equations.

Example 1. As an example, one may consider two data points $(y_1, x_1) = (7, 2)$ and $(y_2, x_2) = (5, 1)$, which yield a system of two equations: $\beta_0 = 7 - 2\beta_1$ and $\beta_0 = 5 - 1\beta_1$. This system is determined since $n = p$, and it has a unique solution at the intersection of the two lines, which is $(\beta_0, \beta_1) = (3, 2)$.

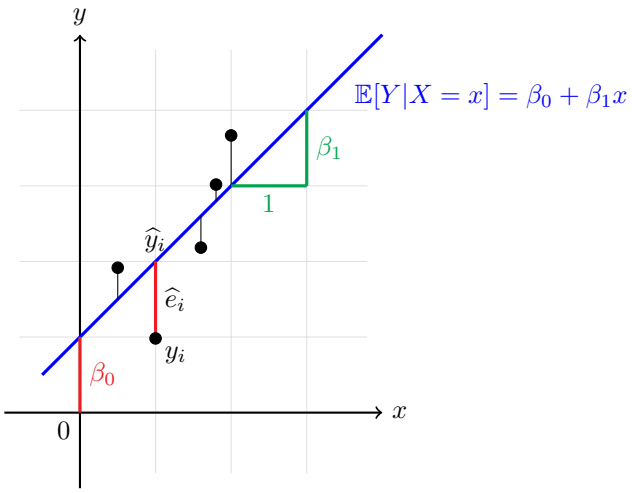

Figure 2. Schematic representation of a simple linear regression model. Here, the y_i's could be heights, whereas the x_i's could be ages, and we wish to predict heights using the ages of the subjects. Each data point is a **pair** of the form (y_i, x_i), with $i = 1, \ldots, n$. Here, I have also represented the i^{th} predicted value $\widehat{y}_i$ for a particular point y_i, and the corresponding error term for that prediction, $\widehat{e}_i$. Observe that both $\widehat{y}_i$ and its error $\widehat{e}_i$ have a hat, since they are both estimated quantities.

2 OLS Estimation

2.1 Mean and Variance Functions

In **simple regression**, we are given two sequences of data points, and we wish to understand how they are related to each other. Each pair of observations will be here referred to as a **case**.

$$(y_i, x_i), \qquad i = 1, \ldots, n.$$

In regression, as opposed to correlation, one of these variables is treated as the **outcome variable** or **dependent variable**, generally denoted by the y_i's. We will then use the other variables for predicting that outcome. As a result, the other variables are referred to as **predictors**, or **independent variables**, and are denoted by x_i's.

The univariate linear regression model consists of a **mean function**, expressed as a conditional expectation,

$$\mathbb{E}[Y|X = x_i] = \beta_0 + \beta_1 x_i,$$

and a **variance function**, expressed as a conditional variance operator,

$$\mathbb{V}\text{ar}[Y|X = x_i] = \sigma^2.$$

The (unknown) **parameters** in this model are $(\beta_0, \beta_1, \sigma^2)$, where:

i. β_0 is the y**-intercept** of $\mathbb{E}[Y|X]$, when $X = 0$. Thus, we have

$$\mathbb{E}[Y|X = 0] = \beta_0.$$

ii. β_1 is the **rate of change** of $\mathbb{E}[Y|X]$, such that

$$\mathbb{E}[Y|X = x + 1] - \mathbb{E}[Y|X = x] = \beta_1.$$

iii. σ^2 is the **variance** of Y, given X. It is strictly positive,

$$\sigma^2 > 0.$$

2.2 Statistical Errors

Because we are dealing with an **over-determined** system of linear equations, which has (generally) no solution, it follows that there does not exist a mean function, such that $\mathbb{E}[Y|X = x_i] = y_i$, for every $i = 1, \ldots, n$. The **observables** or **observed data**, denoted by y_i's, differ from the expected value of Y, given X. We account for this difference with the following equation,

$$y_i = \mathbb{E}[Y|X = x_i] + e_i, \qquad i = 1, \ldots, n,$$

where the e_i's are the **statistical errors**. In effect, we are here literally adding noise to the model. Thus, the e_i's are collectively referred to as **additive noise**.

The e_i's are defined as the difference between the observables and the conditional expected values –that is,

$$e_i = y_i - \mathbb{E}[Y|X = x_i],$$

or alternatively, if we are treating them as random variables,

$$E_i = Y_i - \mathbb{E}[Y|X = x_i],$$

Geometrically, the errors correspond to the **vertical distances** between each y_i and each conditional expectation or mean function, $\mathbb{E}[Y|X = x_i]$. Note that the error terms are not observable, since they depend on the unknown parameters (β_0, β_1). Thus, the errors should also be treated as **random variables**.

2.3 Minimization of the Statistical Criterion

Estimating the parameters is therefore conducted by finding the straight line that minimizes the sum of squared distances between the observed data points and their projections on the regression line. That is, we wish to minimize the **sum of squares of the residuals**. We proceed by minimizing this quantity, which we call a **statistical criterion**. Thus, we choose the pair, (β_0, β_1), that minimizes the following **residual sum of squares** (RSS),

$$\mathrm{RSS}(\beta_0, \beta_1) := \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i)\Big)^2.$$

This minimization can be re-written using the **argmin operator**, as follows,

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right) := \underset{\beta_0, \beta_1 \in \mathbb{R}^2}{\operatorname{argmin}} \, \mathrm{RSS}(\beta_0, \beta_1).$$

Importantly, this minimization can be accomplished in **closed-form**, through direct optimization, by taking the first partial derivatives of the RSS and setting these derivatives to zero:

$$\frac{\partial}{\partial \beta_0} \mathrm{RSS}(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0,$$
$$\frac{\partial}{\partial \beta_1} \mathrm{RSS}(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0.$$

The pair of estimated parameters that we obtain, $(\widehat{\beta}_0, \widehat{\beta}_1)$ can then be re-plugged inside the definition of the RSS to give the following,

$$\begin{aligned}\text{RSS}(\widehat{\beta}_0, \widehat{\beta}_1) &:= \sum_{i=1}^{n} \Big(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\Big)^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} \widehat{e}_i^2,\end{aligned}$$

where the $\widehat{y}_i$'s are the **fitted** or **predicted values**, defined as

$$\widehat{y}_i = \widehat{\mathbb{E}}[Y|X = x_i] = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \qquad i = 1, \ldots, n.$$

Importantly, the RSS can also be seen as the sum of squares of the residuals, $\widehat{e}_i$'s, which are defined as the difference between the observed and predicted values,

$$\widehat{e}_i = y_i - \widehat{y}_i, \qquad i = 1, \ldots, n.$$

Observe that this last equation could be re-written as $\widehat{E}_i = Y_i - \widehat{Y}_i$, since all these quantities are dependent on the Y_i's, and therefore constitute genuine random quantities. For similar reasons, note that both β_0 and β_1 are also random quantities, because they depend on the values taken by the Y_i's.

2.4 Parameter Estimates

The estimate of the y-**intercept** is given by

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

whereas the estimate of the **slope** of the regression line is given by

$$\widehat{\beta}_1 = \widehat{r}_{xy} \frac{\widehat{\sigma}_y}{\widehat{\sigma}_x}.$$

Here, $\widehat{r}_{xy}$ is the **estimated correlation coefficient** between X and Y. This quantity is modified by the ratio of the **estimated standard deviations** of Y and X, denoted $\widehat{\sigma}_y$ and $\widehat{\sigma}_x$, respectively.

Finally, it remains to estimate the variance parameter, σ^2. This is done by *normalizing* the RSS, estimated at $\widehat{\beta}_0$ and $\widehat{\beta}_1$,

$$\widehat{\sigma}^2 := \frac{1}{n-2} \text{RSS}(\widehat{\beta}_0, \widehat{\beta}_1),$$

where the denominator is $n - 2$, because we have already estimated two parameters. That is, from the n data points that we have started with, we have used two **degrees of freedom**, and we therefore need to correct for this. We are here incurring a penalty for using some of the data points, which translates into a larger variance estimate. This estimated quantity, σ^2, is also referred to as the **residual mean square**.

2.5 Gauss-Markov Assumptions on the Error Terms

We make three crucial assumptions on the joint moments of the error terms. These assumptions are required for the Gauss-Markov theorem to hold. (Note that this theorem also assumes that the fitted model is **linear** in the parameters. We will return to a formal statement of the Gauss-Markov theorem, later in the course.)

i. Firstly, we assume that the expectations of all the error terms are **centered at zero**, such that

$$\mathbb{E}[E_i|x_i] = 0, \qquad i = 1, \ldots, n.$$

ii. Secondly, we also assume that the variances of the error terms are constant for every $i = 1, \ldots, n$. This assumption is referred to as **homoscedasticity**.

$$\mathbb{V}\text{ar}[E_i|x_i] = \sigma^2, \qquad i = 1, \ldots, n.$$

iii. Thirdly, we assume that the error terms are **uncorrelated**,

$$\mathbb{C}\text{ov}[E_i, E_j|x_i, x_j] = 0, \qquad \forall\ i \neq j.$$