Chapter 12.2: Estimating Model Parameters

Instructor: Dr. Arnab Maity
Simple Linear Regression Model

Observed data: \((y_1, x_1), \ldots, (y_n, x_n)\)

The simple linear regression model is

\[ y_i = \beta_0 + \beta_1 x_i + e_i. \]

where

- \(\beta_0\) is the intercept parameter
- \(\beta_1\) is the slope parameter
- \(e_i\) is the random error associated with the \(i\)-th observation of the response variable.

**Example:** Recall the example discussed in the previous lecture. We saw data on \(x = \text{pH}\) and \(y = \text{arsenic removed (\%) by a particular process. Data for this example is shown in the book (Example 12.2).}\)

![Graph of arsenic removed vs pH](image)

In this example we have

- \(y_i = \text{arsenic removal (\%) for the } i\text{-th water sample}\)
- \(x_i = \text{pH of the } i\text{-th water sample}\)
- \(i = 1, \ldots, 18\) (total 18 water samples)

**Goal:** Learn to estimate the unknown parameters \(\beta_0, \beta_1\) and \(\sigma^2\).
Looking at the regression model, the observed data points should be distributed around the true regression line randomly. So we should seek to fit a line that is “close” to the observed data points.

Which of the above plots do you think gives you a possible good regression line and why?

The above figure and discussion indicates that a “good” estimate of the line should provide, in some sense, a best fit to the observed data points. This is typically done by the principle of **least squares**. This principle tells us that

- A line provides a good fit to the data if the vertical distances from the observed data points to the fitted line are small
- The measure of the goodness of fit is the sum of the squares of these deviations
- The best-fit line is then the one having the smallest possible sum of squared deviations
**Principle of Least Squares**

The *vertical deviation* of the point \((y_i, x_i)\) from the line \(y = \beta_0 + \beta_1 x\) is

\[ y_i - \beta_0 - \beta_1 x_i. \]

Compute this deviation for each point in the observed data. Then the sum of squares of all such deviations is

\[ f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \]

Find out which values of \(\beta_0\) and \(\beta_1\) minimizes the above sum of squares and denote them by \(\hat{\beta}_0\) and \(\hat{\beta}_1\), respectively.

These estimates, \(\hat{\beta}_0\) and \(\hat{\beta}_1\), are called the **least squares estimates**.

The **estimated regression line** of least squares line is \(y = \hat{\beta}_0 + \hat{\beta}_1 x\)

**How do we minimize the sum of squares:** Recall that we have

\[ f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \]

We take derivative of \(f(\beta_0, \beta_1)\), set them to zero and solve them to obtain the minimizers. The equations are

\[ \frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0 \]

\[ \frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0 \]

We can further simplify and write

\[ n\beta_0 + \left( \sum_{i=1}^{n} x_i \right) \beta_1 = \sum_{i=1}^{n} y_i \]

\[ \left( \sum_{i=1}^{n} x_i \right) \beta_0 + \left( \sum_{i=1}^{n} x_i^2 \right) \beta_1 = \sum_{i=1}^{n} x_i y_i \]

These are called the **normal equations**.

We can easily solve them and obtain the minimizer to be

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \]
Estimation of error variance $\sigma^2$

As we have discussed before, the parameter $\sigma^2$ represent the amount of variability in the regression model. A small value of $\sigma^2$ means that the observed points are very close to the regression line. A large value of $\sigma^2$ indicates that the points are spread out. We need to estimate $\sigma^2$ in order to

- to construct confidence intervals of our point estimates/predictions
- to use in hypothesis testing about various parameters.

Define the **predicted** or **fitted values** of the response as

$$\hat{y}_i = \beta_0 + \hat{\beta}_1 x_i.$$  

The **residuals** are defined as the differences between the observed and the fitted values

$$y_i - \hat{y}_i$$

In theory, **sum of the residual should be zero**. In practice, the sum may slightly differ from zero due to rounding errors.

The **error sum of squares** or **residual sum of squares** is given by

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The estimate of $\sigma^2$ is computed as

$$\hat{\sigma}^2 = \frac{SSE}{n-2}.$$  

The denominator $(n-2)$ in the formula for $\hat{\sigma}^2$ is due to the fact that we need to estimate 2 parameters $\beta_0$ and $\beta_1$ out of $n$ sample size before estimating $\sigma^2$.

**pH versus Arsenic removal example continued**: Using the data set, we can obtain a regression line $y = 190.2683 - 18.0342 x$, with estimate error standard deviation $\hat{\sigma} = 6.126$. 

![Graph of pH versus Arsenic removal with regression line](image)
One can use the fitted line for two different purposes as described below.

- For a fixed value of $x = x^*$, to provide a point estimate of the expected value of $y$. For example, we may want to estimate the expected arsenic removal (%) when pH is set to 9.

- For a fixed value of $x = x^*$, to provide a point prediction of a single new observation made at $x = x^*$. For example, if we set pH to 9, we want to predict a new observation (not the expected value) at that setting.

- The value of both the point estimate and the point prediction at $x = x^*$ is the same: $\hat{\beta}_0 + \hat{\beta}_1 x^*$. In our example, this value is for $pH = 9$ is

- Caution: The least squares line should not be used to make a prediction outside the range of the data. Such an extrapolation is dangerous since the relationship may not be valid outside the range of the observed data.

**How well does the model fit the data?** This is a natural question to ask after any regression analysis. This can be answered by looking at previously defined SSE. We can think of SSE to the the unexplained variance from the model, that is, how much of the total variation in $y$ is left unaccounted even after fitting the model.

To see this, consider the case where the observed data falls exactly on a straight line. What is SSE in this case?

Suppose the data are very spread out. What type value would you expect of SSE in this case?

We measure the total amount of variation in the data using **total sum of squares**

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$  

The **coefficient of determination** is defined as

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

We interpret this as the proportion of the observed $y$ variation that can be explained by the simple linear regression model.

As SST is always more than SSE, $R^2$ can take values between only 0 and 1.

For the model to be a good fit, we expect to have a higher $R^2$ (close to 1). Smaller value of $R^2$ (close to zero) indicates that the simple linear regression may not be a good fit.