Lecture 3: Ordinary Least Squares

1. Ordinary Least Squares With a Single Independent Variable

Previously we have discussed linear regression in a fairly general sense, but now we will focus on a particular method of determining linear regression coefficients called Ordinary Least Squares (OLS). OLS is by far the most widely used tool in regression analysis. It is a benchmark of sorts, so that even when researchers go on to use other more sophisticated methods, the OLS estimates are commonly presented first for comparison.

Recall that the theoretical model of interest is the linear model

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \]

From this equation, we seek to use the information contained in a dataset of observations \((X_i, Y_i)\) to estimate the values of \(\beta_0\) and \(\beta_1\), which we call \(\hat{\beta}_1\) and \(\hat{\beta}_2\). The estimated equation is

\[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \]

and we defined the residuals as the difference between the fitted values \(\hat{Y}_i\) and the observed values \(Y_i\):

\[ e_i \equiv Y_i - \hat{Y}_i. \]

To estimate the regression coefficients in a consistent way, we must first formally define our loss function, the criteria by which we determine whether the fit is good or not. OLS is founded on minimizing the sum of squared residuals (SSR), defined as

\[ SSR = \sum_{i=1}^{n} e_i^2 = e_1 + e_2 + \cdots + e_n, \]

where \(n\) is the sample size. (Note that Studenmund calls this quantity the Residual Sum of Squares or RSS, but SSR is probably the more common term of the two.) Therefore, the OLS estimates of \(\beta_0\) and \(\beta_1\) are defined to be the values of \(\hat{\beta}_0\) and \(\hat{\beta}_1\) which, when plugged into the estimated regression equation, minimize the SSR. Note that these estimates will be different for a different sample, both for another sample of size \(n\) and for samples with more or less observations.

1.1. Why Ordinary Least Squares?

OLS is used so often for several reasons.

1. It is very straightforward to implement, both by hand and computationally, and it is simple to work with theoretically. The form of the OLS coefficients is easy to calculate by hand and in general, running OLS on a computer is simple computationally as it only involves basic matrix algebra, which computers handle well. The objective function, the sum of squared residuals, is a well-behaved function that lends itself to easier mathematical analysis than some other choices.
2. The criteria of minimizing the squared residuals is intuitive. There are several other possibilities we could consider as well, but each has practical or theoretical drawbacks, which as resulted in OLS being the most widely used method. For example, we might try simply minimizing the sum of the residuals ($\sum_1^n \varepsilon_i$), but that would weight negative and positive residuals differently. In most cases, we probably want to weight them the same. We might also try minimizing the sum of the absolute values of the residuals ($\sum_1^n |\varepsilon_i|$). This method, called Least Absolute Deviations (LAD), is very robust to outliers, but it is very difficult to implement. There is no closed form for the regression coefficients, like we have with OLS, and calculating them on a computer is very computationally complex.

3. Third reason?

1.2. The Ordinary Least Squares Estimator

The Ordinary Least Squares estimator of $\beta_0$ and $\beta_1$ in the linear model is defined to be the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimize the sum of squared residuals

$$\sum_{i=1}^n |Y_i - \hat{Y}_i|^2 = \sum_{i=1}^n |Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i|^2.$$ 

The general equations for the coefficients which minimize this quantity are well-known and can be written:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \cdot \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2},$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$ 

Equivalently, we can also write $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n [(X_i - \bar{X}) \cdot (Y_i - \bar{Y})]}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$ 

The first form for $\hat{\beta}_1$ is easier to calculate by hand while the second form is perhaps more intuitive, being the ratio of the sample covariance between $X$ and $Y$ to the sample variance of $X$.

Using the first set of equations for the coefficients, note that we only need to calculate the following sample properties to find numerical values for $\hat{\beta}_0$ and $\hat{\beta}_1$:

1. $\sum_{i=1}^n X_i$,
2. $\sum_{i=1}^n Y_i$,
3. $\sum_{i=1}^n X_i Y_i$,
4. $\sum_{i=1}^n X_i^2$.

Once we know these values, and the sample size ($n$), we can calculate the regression coefficients and draw the regression line.
Example 1. Recall our simple example dataset on stock price and trade volume, given in the table below.

The quantities we need to calculate are:

\[
\begin{align*}
\sum_{i=1}^{n} X_i &= 1 + 4 + 1 = 6 \\
\sum_{i=1}^{n} Y_i &= 2 + 2 + 3 = 7 \\
\sum_{i=1}^{n} X_i Y_i &= 2 + 8 + 3 = 13 \\
\sum_{i=1}^{n} X_i^2 &= 1 + 16 + 1 = 18
\end{align*}
\]

Substituting these values into the equations for the regression coefficients above, we have:

\[
\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \frac{3 \cdot 13 - 6 \cdot 7}{3 \cdot 18 - 6^2} = \frac{39 - 42}{54 - 36} = \frac{-3}{18} = -\frac{1}{6},
\]

and

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{7}{3} - \left(-\frac{1}{6} \cdot 2\right) = \frac{14 + 2}{6} = \frac{8}{3}.
\]

Therefore, the regression line is a line with intercept equal to 8/3 and slope −1/6:

\[
\hat{Y}_i = \frac{8}{3} - \frac{1}{6} X_i.
\]

Does the regression line pass through the middle of the two points (1, 2) and (1, 3)? To check, we can evaluate the fitted values of \(Y\) for each point in the dataset. The SSR is \(\sum_{i=1}^{n} e_i^2 =
\]

<table>
<thead>
<tr>
<th>(i)</th>
<th>(\hat{Y}_i)</th>
<th>(e_i)</th>
<th>(e_i^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.5</td>
<td>-0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2.5</td>
<td>0.5</td>
<td>0.25</td>
</tr>
</tbody>
</table>
0.25 + 0 + 0.25 = 0.5.

In particular, for the regression line at the point $X = 1$:

$$
\hat{Y} = \frac{8}{3} - \frac{1}{6} \cdot 1 = \frac{16 - 1}{6} = \frac{15}{6} = 2.5.
$$

So, yes, this line passes exactly through the middle of the points $(1,2)$ and $(1,3)$. It also passes directly through the point $(4,2)$ since

$$
\hat{Y} = \frac{8}{3} - \frac{1}{6} \cdot 4 = \frac{16 - 4}{6} = \frac{12}{6} = 2.
$$

**Example 2.** Using the dataset below, calculate $\hat{\beta}_0$ and $\hat{\beta}_1$, plot the data points, and draw the regression line.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$X_i$</th>
<th>$Y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>4</td>
</tr>
</tbody>
</table>

**Example 3.** Using the dataset below, calculate $\hat{\beta}_1$ and interpret your findings.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$X_i$</th>
<th>$Y_i$</th>
<th>$X_i Y_i$</th>
<th>$X_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>12</td>
<td>24</td>
<td>12</td>
</tr>
</tbody>
</table>

In this example, we have

$$
\hat{\beta}_1 = \frac{3 \cdot 24 - 6 \cdot 12}{3 \cdot 12 - 6^2} = \frac{72 - 72}{36 - 36} = \frac{0}{0}.
$$

This is undefined! So, what is the slope of the regression line? To see what is going wrong in this example, plot the data points. The sample variance of $X_i$ is exactly zero. In order to estimate the regression coefficients, which represent $Y_i$ as a function of $X_i$, we need to have some variation in $X_i$. If we only observe a single value of $X_i$, we can never estimate the slope of the regression line, which represents the effect of changes in $X_i$ on $Y_i$.

### 2. Interpretation of Regression Coefficients

In the regression line

$$
Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,
$$

the coefficient on $X_i$ represents the amount by which we predict $Y_i$ will increase when $X_i$ increases by one unit.
For example, suppose that $Y_i$ is individual $i$’s annual demand for housing in dollars and that $X_i$ is individual $i$’s annual income, also measured in dollars. Then $\hat{\beta}_1$ is the number of additional dollars individual $i$ is predicted to spend on housing when income increases by one dollar. The intercept, $\hat{\beta}_0$, is an individual’s predicted expenditure on housing when income is zero.

Note that when $\hat{\beta}_1$ is zero, $X_i$ does not influence $Y_i$. Many empirically relevant questions focus on whether or not a particular coefficient is approximately zero or not. For example, what is the effect of the number of young children in a household on the mother’s decision about how many hours to work outside the home. We could regress hours worked, $Y_i$, on the number of children under age six, $X_i$, and look at the coefficient on $X_i$ to determine whether there is any effect.

3. Anscombe’s Quartet: A Cautionary Tale

Now that we have formally defined Ordinary Least Squares, a particular type of linear regression founded on minimizing the sum of squared residuals, it is useful to take a step back and think about what the results tell us, and to look at some cases where they can be very misleading. In particular, there is a collection of four datasets referred to as Anscombe’s Quartet, displayed in Figure 1 that serve to highlight the importance of exploring the data carefully before blindly running a regression Anscombe (1973). Despite looking very different when plotted, these four datasets have many identical statistical properties. Despite the fundamental differences, the OLS coefficients, and therefore the estimated regression lines, are identical. Plotting the data first might suggest other functional forms (e.g., perhaps we should regress on $Z = X^2$) or might suggest that we should worry about outliers (e.g., perhaps we should be using a more advanced technique such as LAD, which is beyond the scope of this course).
References