Wednesday September 29

Review of Monday’s class: Least Squares principle

Minimizing

$$\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

over all possible values of $\beta_0$ and $\beta_1$ gives

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

and

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

Note $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$, the average of the $Y_i$

The mathematical calculation requires being able to find the minimum of a function of two variables using differentiation
Exercise for today

1. $n = 2$, $X_1 = 1$, $X_2 = 2$, then $\sum_{i=1}^{n} X_i = 3$,
   \[ \sum_{i=1}^{n} X_i^2 = 1^2 + 2^2 = 5, \]
   and $\left( \sum_{i=1}^{n} X_i \right)^2 = 3^2 = 9$.

2. $\sum_{i=1}^{4} i = 1 + 2 + 3 + 4 = 10$.

3. Needed: $\sum_{i=1}^{3} X_i Y_i$, $\sum_{i=1}^{3} X_i$, $\sum_{i=1}^{3} Y_i$, $\sum_{i=1}^{3} X_i^2$
   $\Rightarrow$ make a table
Table for Q3:

<table>
<thead>
<tr>
<th>$Y_i$</th>
<th>$X_i$</th>
<th>$X_iY_i$</th>
<th>$X_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>20</td>
<td>40</td>
<td>400</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>120</td>
<td>900</td>
</tr>
<tr>
<td>7</td>
<td>60</td>
<td>170</td>
<td>1400</td>
</tr>
</tbody>
</table>

So,

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

$$= \frac{3 \cdot 170 - 60 \cdot 7}{2 \cdot 1400 - 60^2} = \frac{510 - 420}{4200 - 3600} = 0.15$$

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

$$= \frac{7}{3} - 0.15 \frac{60}{3} = -2/3 \approx -0.6667$$
Computer software: Eviews will calculate the regression line for us

Other packages that can do this:

- Excel, or other spreadsheet program
- SAS, SPSS
- Stata, Eviews, etc.
Monday's in-class example

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$Y_i$</th>
<th>$X_iY_i$</th>
<th>$X_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>13</td>
<td>18</td>
</tr>
</tbody>
</table>

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

\[
= \frac{3 \cdot 13 - 7 \cdot 6}{3 \cdot 18 - 6^2} = \frac{39 - 42}{54 - 36} = \frac{-3}{18}
\]

\[
= -\frac{1}{6} \approx 0.167
\]

Also

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{7}{3} - (-\frac{1}{6})(\frac{6}{3}) = \frac{8}{3} \approx 2.67
\]

Regression line: $y = (\frac{8}{3}) - (\frac{1}{6})x$
Q: Does the regression line \( y = \frac{8}{3} - \frac{1}{6}x \) pass through the middle of the points \((1, 2)\) and \((1, 3)\)?

A: Yes, if \( x = 1 \), the corresponding \( y \) value for the regression line is

\[
\left(\frac{8}{3}\right) - \left(\frac{1}{6}\right) \cdot 1 = \left(\frac{16}{6}\right) - \left(\frac{1}{6}\right)
\]

\[
= \frac{15}{6} = 2.5
\]
Other example

\[
\begin{array}{|c|c|c|c|}
\hline
Y_i & X_i & X_iY_i & X_i^2 \\
\hline
3 & 2 & 6 & 4 \\
4 & 2 & 8 & 4 \\
5 & 2 & 10 & 4 \\
12 & 6 & 24 & 12 \\
\hline
\end{array}
\]

So,

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

\[
= \frac{3 \cdot 24 - 6 \cdot 12}{3 \cdot 12 - 6^2} = \frac{72 - 72}{36 - 36} = \frac{0}{0} = ???
\]
Concepts / names

Predictions:

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

Residuals:

\[ \hat{e}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = Y_i - \hat{Y}_i \]

Residual sum of squares:

\[ RSS = \sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \]
Monday’s in-class example

Recall $\hat{\beta}_0 = 8/3; \hat{\beta}_1 = -1/6$

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$Y_i$</th>
<th>$X_iY_i$</th>
<th>$X_i^2$</th>
<th>$\hat{Y}_i = 8/3 - (1/6)X_i$</th>
<th>$e_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2.5</td>
<td>-0.5</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>8</td>
<td>16</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2.5</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>13</td>
<td>18</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Also

$$RSS = \sum_{i=1}^{n} e_i^2 = (-0.5)^2 + 0^2 + (0.5)^2 = 0.5$$
Interpretation of coefficients

We obtain the regression line

\[ y = \hat{\beta}_0 + \hat{\beta}_1 x \]

\(Y_i\): demand for housing of individual \(\neq i\), in dollars annually

\(X_i\): income in dollars annually

\(\hat{\beta}_1\) is how many dollars an individual is expected to spend on housing when income increases by \$1

\(\hat{\beta}_0\) is how many dollars an individual is expected to spend on housing when income equals \$0
Obvious and frivolous comment

\[ y = \hat{\beta}_0 + \hat{\beta}_1 x \]

If \( \hat{\beta}_1 \approx 0 \), then \( X_i \) does not “influence” \( Y_i \)

Earlier example: if \( \hat{\beta}_1 \approx 0 \), then expenditure on housing is not “influenced” by one’s income

Often, important questions correspond to a coefficient of 0

- Does being a woman negatively impact earnings potential?

- Is there a correlation between percentage of foreigners in a neighborhood and the crime rate?

- Is there a relationship between number of cans of soda sold in a stadium and temperature?