SIMPLE REGRESSION

HOT NOTES FOR STATISTICS

Abstract. Fit a straight line through bivariate predictor-response data to model the conditional mean, and test for significance.

1. Ordinary Least Squares Estimate

The ordinary least square (OLS) estimate for the line modeling $E[Y|x]$ is the line which minimizes the sum of squared errors for the $y_i$ values

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

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

Minimizing $SSE$ with respect to the coefficients gives the estimators

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

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

$$\hat{y}_i = E[Y|x_i] = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

For example, fit a straight line to this sample

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$x_i^2$</th>
<th>$y_i^2$</th>
<th>$x_i y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>24</td>
<td>100</td>
<td>576</td>
<td>240</td>
</tr>
<tr>
<td>15</td>
<td>30</td>
<td>225</td>
<td>900</td>
<td>450</td>
</tr>
<tr>
<td>15</td>
<td>32</td>
<td>225</td>
<td>1024</td>
<td>480</td>
</tr>
<tr>
<td>20</td>
<td>42</td>
<td>400</td>
<td>1764</td>
<td>840</td>
</tr>
</tbody>
</table>

$$n = 4 \quad \bar{x} = 15 \quad \bar{y} = 32$$

$$\hat{\beta}_1 = \frac{2010 - 4 \times 15 \times 32}{950 - 4 \times 15^2} = 1.8$$

$$\hat{\beta}_0 = 32 - 1.8 \times 15 = 5$$
2. Regression ANOVA

The regression model can be tested using ANOVA. The sums of squares used in the table can be found from the data and the model:

\[ \begin{align*}
SST &= \sum_i y_i^2 - n \bar{y}^2 \\
SSE &= \sum_i y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i
\end{align*} \]

The ANOVA table compares explained variation (SSR) to unexplained variation (SSE):

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>regression</td>
<td>SST - SSE</td>
<td>p - 1</td>
<td>SSR/p - 1</td>
<td>MSR/MSE</td>
</tr>
<tr>
<td>error</td>
<td>SSE</td>
<td>n - p</td>
<td>SSE/n - p</td>
<td>MSE</td>
</tr>
<tr>
<td>Total</td>
<td>SST</td>
<td>n - 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For the example

\[ \begin{align*}
SST &= 4264 - 4 \times 32^2 = 168 \\
SSE &= 4264 - 5 \times 128 - 1.8 \times 2010 = 6
\end{align*} \]

and the ANOVA table is

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>regression</td>
<td>162</td>
<td>1</td>
<td>162</td>
<td>54</td>
</tr>
<tr>
<td>error</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>168</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

At the \( \alpha = 0.05 \) level, the critical value is \( F_{1,2,0.05} = 18.51 \). Since the \( F \) statistic exceeds the critical value, we have strong evidence favoring the fitted line as a good model for \( \mathbb{E}[Y|x] \).

3. Coefficient of Determination

The coefficient of determination is a summary statistic for the quality of fit. It is the proportion of variation in the response “explained” by variation in the predictor.

\[ r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \]

For the example

\[ r^2 = \frac{162}{168} = 0.964 \]

\[ \text{Department of Management Science and Statistics, UTSA} \]
\[ E-mail \text{ address: Michael.Anderson@utsa.edu} \]

---

\(^1\)For simple regression, this is a test of the slope, \( \hat{\beta}_1 \).