Lecture 9

Regression

1) Origin
   - Sir Francis Galton
   - Ref: Shifts and Contrivances Available in Wild Places
   - Father’s height \((x)\) vs. son’s height \((y)\)
   - Direct proportion?
   - i.e., \(y = kx\), where \(k\) is a constant?

Father’s height vs. son’s height
2) Regression

- Regression to the mean
- Linear regression
- Problem: Mothers height is not included.
- Multiple regression
**Example 1**  Dosages of a new allergy drug vs. duration of relief

<table>
<thead>
<tr>
<th>Dosage ($x$)</th>
<th>3 3 4 5 6 6 7 8 8 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration of relief (hours, $y$)</td>
<td>9 5 12 9 14 16 22 18 24 22</td>
</tr>
</tbody>
</table>

Scatter plot
3) Statistical Model

1. **Statistical model for a linear regression**

   \[ y = \beta_0 + \beta_1 x + \epsilon \]

2. **Assumptions**

   (a) The relationship is linear and \( \text{E}(\epsilon_i) = 0 \) for all \( i \).
   (b) The errors have the same variance.
   (c) \( \epsilon_i \sim N(0, \sigma^2) \) for all \( i \).

3. **Least squares estimation**

   - The best fitting straight line is given by
     \[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \]
   - The least-square estimates of \( \beta_1 \) and \( \beta_0 \) are
     \[ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]
     where \( S_{xx} = \sum x_i^2 - n(\bar{x})^2 \) and \( S_{xy} = \sum x_i y_i - n(\bar{x})(\bar{y}) \).
   - Residuals: \( \hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \), \( i = 1, \ldots, n \)
   - Residual sum of square (error sum of square):
     \[ \text{SSE} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}, \quad \text{where} \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 \]
   - Estimate of the error variance \( \sigma^2 \):
     \[ S^2 = \frac{\text{SSE}}{n - 2} \]

   - Confidence interval
   - Testing
4) Regression Equation

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
<th>x²</th>
<th>y²</th>
<th>xy</th>
<th>(\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x)</th>
<th>(\hat{\epsilon})</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>9</td>
<td>9</td>
<td>81</td>
<td>27</td>
<td>7.15</td>
<td>1.85</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>9</td>
<td>25</td>
<td>15</td>
<td>7.15</td>
<td>-2.15</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>16</td>
<td>144</td>
<td>48</td>
<td>9.89</td>
<td>2.11</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>25</td>
<td>81</td>
<td>45</td>
<td>12.63</td>
<td>-3.63</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>36</td>
<td>196</td>
<td>84</td>
<td>15.37</td>
<td>-1.37</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>36</td>
<td>256</td>
<td>96</td>
<td>15.37</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>22</td>
<td>49</td>
<td>484</td>
<td>154</td>
<td>18.11</td>
<td>3.89</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>18</td>
<td>64</td>
<td>324</td>
<td>144</td>
<td>20.85</td>
<td>-2.85</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>24</td>
<td>64</td>
<td>576</td>
<td>192</td>
<td>20.85</td>
<td>3.15</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>22</td>
<td>81</td>
<td>484</td>
<td>198</td>
<td>23.59</td>
<td>-1.59</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>151</td>
<td>389</td>
<td>2651</td>
<td>1003</td>
<td>0.04</td>
<td></td>
</tr>
</tbody>
</table>

\[ \bar{x} = 5.9 \quad \bar{y} = 15.1 \]

\[ S_{xx} = 389 - 10(5.9)^2 = 40.9 \]
\[ S_{xy} = 1003 - 10(5.9)(15.1) = 112.1 \]
\[ S_{yy} = 2651 - 10(15.1)^2 = 370.9 \]
\[ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{112.1}{40.9} = 2.74 \]
\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 15.1 - (2.74)(5.9) = -1.07 \]
\[ \hat{y} = -1.07 + 2.74x \]

The mean of the duration of relief increases by 2.74 if the dosage increases by 1.
5) Correlation

- Coefficient of determination:

\[ R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} \]

- Sample correlation coefficient:

\[ r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \]

- Example 1 (continued)

\[ R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{(112.1)^2}{(40.9)(370.9)} = 0.83 \]

83% of the variability in y is explained by linear regression.