Regression

Rather than looking for associations between two variables, we may want to model the mean of an outcome variable as a function of a predictor or explanatory variable (an extension of the ANOVA model).

Example: How do the expenses per admission depend on the length of stay in the hospital?

Questions that regression can address:
- If the length of stay goes up by a day, how do costs change?
- If a patient’s length of stay is 9 days, what would we expect their costs to be (with 95% confidence)?

Least Squares Line

Interpretation of:
- Slope
- Intercept

Scatter Plot

Simplest model to start with is that as LOS increases, Expenses per Admission increase linearly:

Guesstimate of the slope of the line?
Y-intercept?

Ordinary Least Squares

Find the line $\hat{y} = \hat{\alpha} + \hat{\beta} x$
that minimizes the sum of the squared deviations of the observed points $y$ to the line

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

$\hat{\beta} = \frac{S_y}{S_x}$

$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$
Assumptions for Simple Linear Regression

- The mean of Y given X is linear \( \mu_{Y|X} = \alpha + \beta x \)
- The standard deviation of Y given X \( \sigma_{Y|X} \) is constant for all X (homoscedasticity)
- The outcomes Y are independent
- For testing and confidence intervals, need normality:
  \[ Y \sim N(\mu_{Y|X}, \sigma_{Y|X}^2) \]

S-Plus Output

Coefficients:

| Value   | Std. Error | t value | Pr(>|t|) |
|---------|------------|---------|----------|
| (Intercept) | 121.9595 | 60.8104 | 2.0101 0.0420 |
| lms      | 191.5630  | 80.4654 | 2.3807 0.0212 |

Residual standard error: 57.7 on 49 degrees of freedom
Multiple R-Squared: 0.1037
F-statistic: 5.666 on 1 and 49 degrees of freedom, p-value 0.02121

Analysis of Variance Table

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>F Value</th>
<th>Pr(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lms</td>
<td>1890785</td>
<td>1890785</td>
<td>5.66764</td>
<td>0.021219</td>
</tr>
<tr>
<td>Residuals</td>
<td>49</td>
<td>16246819</td>
<td>333609</td>
<td></td>
</tr>
</tbody>
</table>