Simple Linear Regression

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

January 25, 2007

Phosphorous Example

- Researchers gathered data to evaluate the use of phosphorus (P) by nine corn plants.
- The data consist of $x$, the inorganic P in soil (ppm), and $y$, the plant-available P (ppm).

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>4</th>
<th>5</th>
<th>9</th>
<th>13</th>
<th>11</th>
<th>23</th>
<th>23</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>64</td>
<td>71</td>
<td>54</td>
<td>81</td>
<td>93</td>
<td>76</td>
<td>77</td>
<td>95</td>
<td>109</td>
</tr>
</tbody>
</table>

Exploratory data analysis

- Graphical summaries (scatter plot)
- Numerical summaries (means, sds, correlation)
- Use R!
Graphical exploration of two quantitative variables

\[ x = c(1, 4, 5, 9, 13, 11, 23, 23, 28) \]
\[ y = c(64, 71, 54, 81, 93, 76, 77, 95, 109) \]
\[ \text{plot}(x, y, pch = 16, las = 1) \]
\[ \text{fit} = \text{lm}(y \sim x) \]
\[ \text{abline}(\text{fit}) \]

Objectives of simple linear regression

**Description** To describe the relationship between inorganic P in soil and plant-available P

**Estimation** To estimate the population mean plant-available P level at a given level of inorganic P in soil

**Prediction** To predict the plant-available P level for an individual plant at a given level of inorganic P in soil

**Testing** To test if there is a relationship between inorganic P in soil and plant-available P
Simple Linear Regression Model

- \( y_i = \beta_0 + \beta_1 x_i + e_i \), \( e_i \sim \text{iid } N(0, \sigma^2), i = 1, \ldots, n \)
- \( y = \beta_0 + \beta_1 x \) is the “true regression line”
- \( \beta_0 \) is the intercept, \( \beta_1 \) is the slope
- \( x_i \) is the explanatory variable
- \( y_i \) is the response variable
- \( e_i \) is random error
- iid stands for \textit{independent and identically distributed}

Simple Linear Regression Assumptions

1. The model is correct: \( \mathbb{E}(y_i) = \beta_0 + \beta_1 x_i \).
2. Errors \( e_i \) are independent.
3. Errors \( e_i \) have homogeneous variance: \( \text{Var}(e_i) = \sigma^2 \).
4. Errors \( e_i \) have normal distribution: \( e_i \sim N(0, \sigma^2) \).
Estimating Model Parameters

- A well estimated line should be “close to the data points”.
- The least squares criterion says that best line is the one that minimizes \( \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \).
- The solution to this problem is:
  \[
  \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
  \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
  \]
- The fitted values are \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \)
- The estimated variance is \( \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \)

An Alternative Viewpoint

- The correlation coefficient \( r \) is a number between \(-1\) and \(1\) that measures the strength of the linear relationship between \(x\) and \(y\).
  \[
  r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
  \]
- The estimated \( y \) for an \( x \) that is \( z \) standard deviations from the mean is \( rz \) standard deviations from the mean.
- In other words, \( \hat{y} = \bar{y} + rz s_y \).
- The estimated slope and intercept are:
  \[
  \hat{\beta}_1 = r \frac{s_y}{s_x}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
  \]
- The regression line goes through the point \((\bar{x}, \bar{y})\).
Hypothesis Testing: $H_0 : \beta_1 = 0$

- The test statistic $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim T_{n-2}$ under $H_0 : \beta_1 = 0$.
- The p-value is the area under the $t$-distribution more extreme than the observed $t$.

ANOVA Approach

<table>
<thead>
<tr>
<th>Sum of squares</th>
<th>Expression</th>
<th>Degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total ($SS_{Tot}$)</td>
<td>$\sum_{i=1}^{n}(y_i - \bar{y})^2$</td>
<td>$n - 1$</td>
</tr>
<tr>
<td>Regression ($SS_{Reg}$)</td>
<td>$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$</td>
<td>$1$</td>
</tr>
<tr>
<td>Error ($SS_{Err}$)</td>
<td>$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$</td>
<td>$n - 2$</td>
</tr>
</tbody>
</table>

**Partition:** $SS_{Tot} = SS_{Reg} + SS_{Err}$.

**ANOVA Table**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSReg</td>
<td>SSReg/df</td>
<td>MSReg/MSErr</td>
</tr>
<tr>
<td>Error</td>
<td>n-2</td>
<td>SSErr</td>
<td>SSErr/df</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SSTot</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$F = \frac{MS_{Reg}}{MSE_{Err}} \sim F_{1, n-2}$ under $H_0 : \beta_1 = 0$
R Example

Simple Linear Regression in R

```r
> soilP = c(1, 4, 5, 9, 13, 11, 23, 23, 28)
> plantP = c(64, 71, 54, 81, 93, 76, 77, 95, 109)
> plot(soilP, plantP)
> fit = lm(plantP ~ soilP)
> coef(fit)
> summary(fit)
> anova(fit)
> plot(fitted(fit), residuals(fit))
```