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></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>$x$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<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>

- We wish to use the inorganic phosphorous level in the soil to predict the plant-available phosphorous in the corn plants.
- It is good practice to put the data into an R data frame.
- I will show two ways to accomplish this.
Creating a Data Frame in R

> soilP = c(1, 4, 5, 9, 13, 11, 23, 23, 28)
> cornP = c(64, 71, 54, 81, 93, 76, 77, 95, 109)
> phos = data.frame(soilP, cornP)
> rm(soilP, cornP)
> str(phos)

'data.frame': 9 obs. of 2 variables:
$ soilP: num 1 4 5 9 13 11 23 23 28
$ cornP: num 64 71 54 81 93 76 77 95 109

Creating a Data Frame in Excel

- Create a spreadsheet with a header row with variable names and one row per observation.
- Save the file as a comma-separated-variable file (CSV).
- Read using `read.table()` with the `sep=","` argument.

> phos2 = read.table("phos.csv", sep = ",", header = T)
> str(phos2)

'data.frame': 9 obs. of 2 variables:
$ soilP: int 1 4 5 9 13 11 23 23 28
$ cornP: int 64 71 54 81 93 76 77 95 109
Graphical exploration of two quantitative variables

```r
> library(lattice)
> plot(xyplot(cornP ~ soilP,
+       data = phos, pch = 16))
```

Objectives of simple linear regression

<table>
<thead>
<tr>
<th>Description</th>
<th>To describe the relationship between inorganic P in soil and plant-available P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation</td>
<td>To estimate the population mean plant-available P level at a given level of inorganic P in soil</td>
</tr>
<tr>
<td>Prediction</td>
<td>To predict the plant-available P level for an individual plant at a given level of inorganic P in soil</td>
</tr>
<tr>
<td>Testing</td>
<td>To test if there is a relationship between inorganic P in soil and plant-available P</td>
</tr>
</tbody>
</table>
Simple Linear Regression Model

- \( y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \sim \text{iid } \mathcal{N}(0, \sigma^2), i = 1, \ldots, n \)
- \( y = \beta_0 + \beta_1 x \) 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: \( e_i \sim \mathcal{N}(0, \sigma^2). \)
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 $\sum_{i=1}^{n}(y_i - (b_0 + b_1x_i))^2$.
- 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, $\hat{y} = \bar{y} + rzs_y$.
- 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 $((\bar{x}, \bar{y}))$. 
R Example

Simple Linear Regression in R

```r
> library(arm)
> fit = lm(cornP ~ soilP, data = phos)
> display(fit)

lm(formula = cornP ~ soilP, data = phos)

coef.est coef.se
(Intercept) 61.58 6.25
soilP 1.42 0.39

---

n = 9, k = 2
residual sd = 10.69, R-Squared = 0.65
```

The argument `type=c("p","r")` tells `xyplot()` to plot both points and a regression line.