

Simple Linear Regression

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

January 28, 2008

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).

x	1	4	5	9	13	11	23	23	28
y	64	71	54	81	93	76	77	95	109

- 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
```

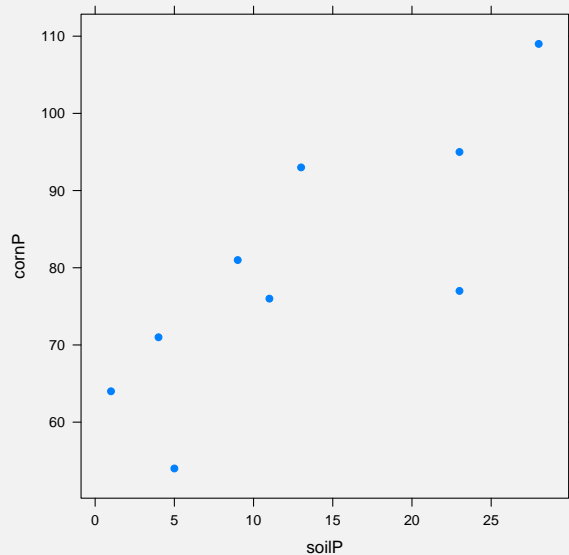
Creating a Data Frame in Excel

- Create a spread sheet 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



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

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

- $y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim \text{iid } N(0, \sigma^2)$, $i = 1, \dots, n$
- $y = \beta_0 + \beta_1 x$ is the “true regression line”
- β_0 is the intercept, β_1 is the slope
- x_i is the explanatory variable
- y_i is the response variable
- e_i is random error
- iid stands for *independent and identically distributed*

Simple Linear Regression Assumptions

- 1 The model is correct: $E(y_i) = \beta_0 + \beta_1 x_i$.
- 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, $\hat{y} = \bar{y} + rzs_y$.
- The estimated slope and intercept are:

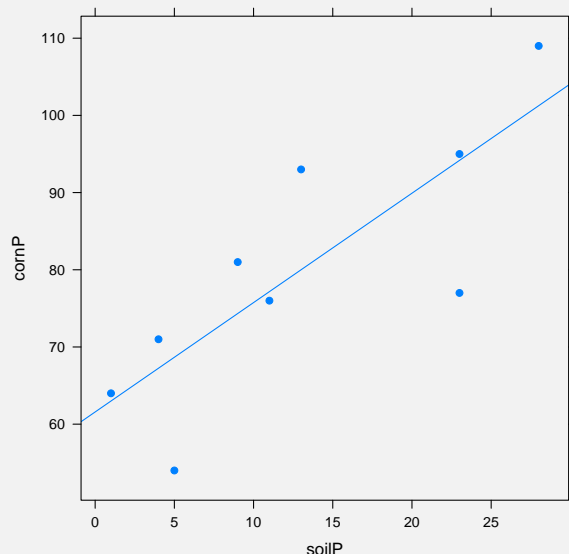
$$\hat{\beta}_1 = r \frac{s_y}{s_x}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
- The regression line goes through the point (\bar{x}, \bar{y}) .

Simple Linear Regression in 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
```

Simple Linear Regression in R



```
> plot(xyplot(cornP ~ soilP,
+ data = phos, pch = 16,
+ type = c("p", "r")))

```

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