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).

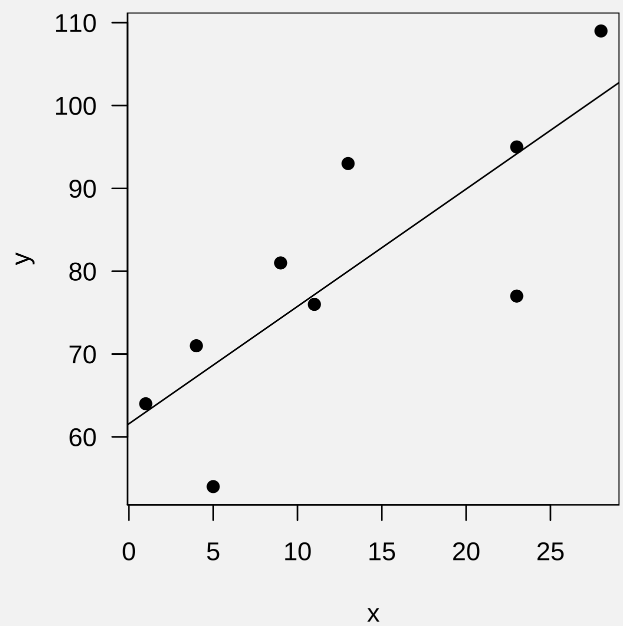
x	1	4	5	9	13	11	23	23	28
y	64	71	54	81	93	76	77	95	109

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)
> plot(x, y, pch = 16, las = 1)
> fit = lm(y ~ x)
> abline(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

- $y_i = \beta_0 + \beta_1 x_i + e_i, \qquad e_i \sim \mathrm{iid}\ \mathrm{N}(0, \sigma^2), i = 1, \ldots, n$
- $y = \beta_0 + \beta_1 x$ 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: $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$.
2. Errors e_i are independent.
3. Errors e_i have homogeneous variance: $Var(e_i) = \sigma^2$.
4. Errors e_i have normal distribution: $e_i \sim \mathrm{N}(0, \sigma^2)$.

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.
- $r = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(\dfrac{x_i - \bar{x}}{s_x}\right)\left(\dfrac{y_i - \bar{y}}{s_y}\right)$
- 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:
  $\hat{\beta}_1 = r\frac{s_y}{s_x}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- The regression line goes through the point $(\bar{x}, \bar{y})$.

Hypothesis Testing: $H_0: \beta_1 = 0$

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

ANOVA Approach

Sum of squares	Expression	Degrees of freedom
Total (SSTot)	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$
Regression (SSReg)	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1
Error (SSErr)	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$

Partition: $\mathrm{SSTot} = \mathrm{SSReg} + \mathrm{SSErr}$.

ANOVA Table

Source	df	SS	MS	F
Regression	1	SSReg	SSReg/df	MSReg/MSErr
Error	n-2	SSErr	SSErr/df	
Total	n-1	SSTot		

$F = \frac{\mathrm{MSReg}}{\mathrm{MSErr}} \sim F_{1,n-2}$ under $H_0 : \beta_1 = 0$

Simple Linear Regression in 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))
```