Chapter 7 Introduction to linear regression

Huamei Dong

03/31/2016

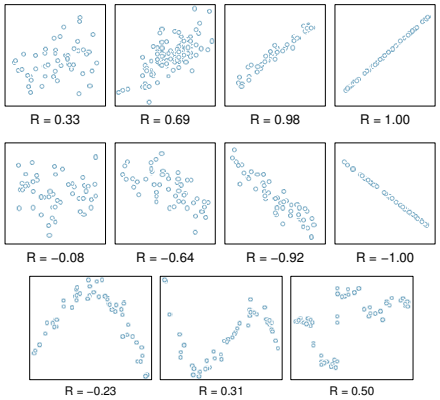
1. Correlation between two variables
2. Regression line
3. Residuals
4. Least Squares Regression line

1. Correlation

Suppose we would like to investigate the relationship between two continuous random variables, for example, cholesterol level and blood pressure level, we can create a two-way scatter plot. Simply by examining the graph, we can often determine whether a relationship exists between two variables.

The correlation quantifies the strength of the linear relationship between two variables. The estimator of the population correlation is known as correlation coefficient R.

$$R=\frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}$$

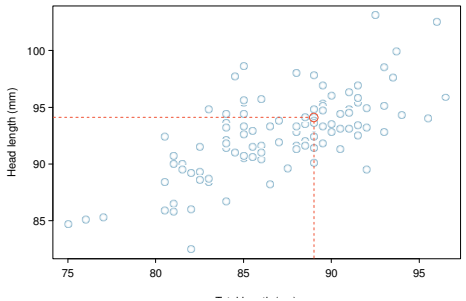


Correlation: strength of a linear relationship

Correlation, which always takes values between −1 and 1, describes the strength of the linear relationship between two variables. We denote the correlation by R.

```
>possum<-read.table('possum.txt', as.is=T, sep="\t", header=T)
> nrow(possum)
[1] 104
>head(possum)
site pop sex age headL skullW totalL tailL
1   1 Vic   m   8 94.1  60.4  89.0 36.0
2   1 Vic   f   6 92.5  57.6  91.5 36.5
3   1 Vic   f   6 94.0  60.0  95.5 39.0
4   1 Vic   f   6 93.2  57.1  92.0 38.0
5   1 Vic   f   2 91.5  56.3  85.5 36.0
6   1 Vic   f   1 93.1  54.8  90.5 35.5
```

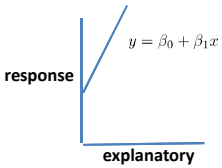
- plot(possum$totalL,possum$headL)
- cor(possum$totalL,possum$headL)



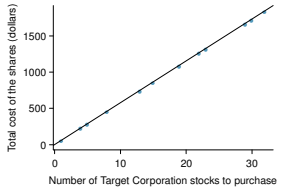
A scatterplot showing head length against total length for 104 brushtail possums.

2. Regression line

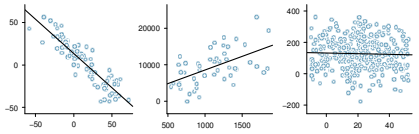
Suppose we would like to investigate the change in one variable, called the response variable, corresponding to a given change in the other, called the explanatory variable, we need another analysis: *simple linear regression.*



Here β_0 and β_1 represents two parameters of linear model. X is the explanatory or predictor variable. Y is the response variable.



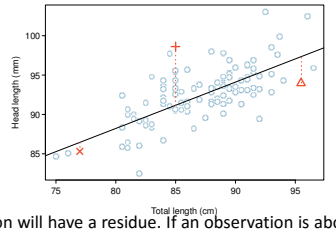
Sometimes the data don't fall exactly on a line. If we believe the relationship is linear, we Try to find a best fitting line.



3. Residuals

Assume we use this linear model to describe the relationship between head length and total length variable, that is, x (total length) is predictor and y(head length) is the response variable.

$$\hat{y} = 41 + 0.59x$$



Each observation will have a residue. If an observation is above the regression line, the residue is positive. If an observation is below the line, the residue is negative. Three observations are noted specially.

Residual: The difference between observed and expected

$$e_i = y_i - \hat{y}_i$$

We typically identify $\hat{y}_i$ by plugging x_i into the model.

Example1: The linear fit is given as $\hat{y} = 41 + 0.59x$. Based on this line, compute the residual of the observation (77.0, 85.3).

Answer: Denote this observation as (x_1, y_1) The predicted value is

$$\hat{y}_1 = 41 + (0.59)(77.0) = 86.4$$

$$\text{and the residual is}$$

$$e_1 = y_1 - \hat{y}_1 = 85.3 - 86.4 = -1.1$$

4 Least squares regression line

We want a line that has small residuals. Since some residuals are positive and some are negative, we choose a line that minimizes the sum of the squared residuals:

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

The line that minimizes this least squares criterion is called least squares line.

The conditions for least squares line:

(1) Linearity: The data should show a linear trend.

(2) Nearly normal residuals: Residuals should be nearly normal.

(3) Constant variability: The variability of points around the least squares line remains roughly constant. You can also look at the residual plot.

To identify the least squares line from summary statistics:

(1) Estimate the slope parameter using

$$\hat{\beta}_1 = \frac{s_y}{s_x} R$$

(1) Using point $(\bar{x}, \bar{y})$ and slope $\hat{\beta}_1$ in the point-slope equation:

$$y - \bar{y} = \hat{\beta}_1 (x - \bar{x})$$

(1) Simplify equation and you can find $\hat{\beta}_0$

Example 2: Using data from chapter 7 exercise data summary to

(1) Compute the slope for the least squares line.

(2) Find the least squared line.

(3) Interpret the parameters you get.

Homework: Finish Example 2 (due 04/07/16)