Chapter 7 Linear Regression

04/05/2016

Huamei Dong

1. **Review Least square regression line**
2. **Example of linear regression analysis with R**

1. Review of linear regression analysis

In last lecture, we learned correlation coefficient R :

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

We also learned how to find linear regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

from summary statistics: $\bar{x}, \bar{y}, s_x, s_y, R$

by finding $\hat{\beta}_1$ using formula $\hat{\beta}_1 = \frac{s_y}{s_x} R$

Actually if we substitute R and s_x , s_y into $\hat{\beta}_1 = \frac{s_y}{s_x} R$, $\hat{\beta}_1$ is also equal to

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

2. Example of linear regression analysis with R

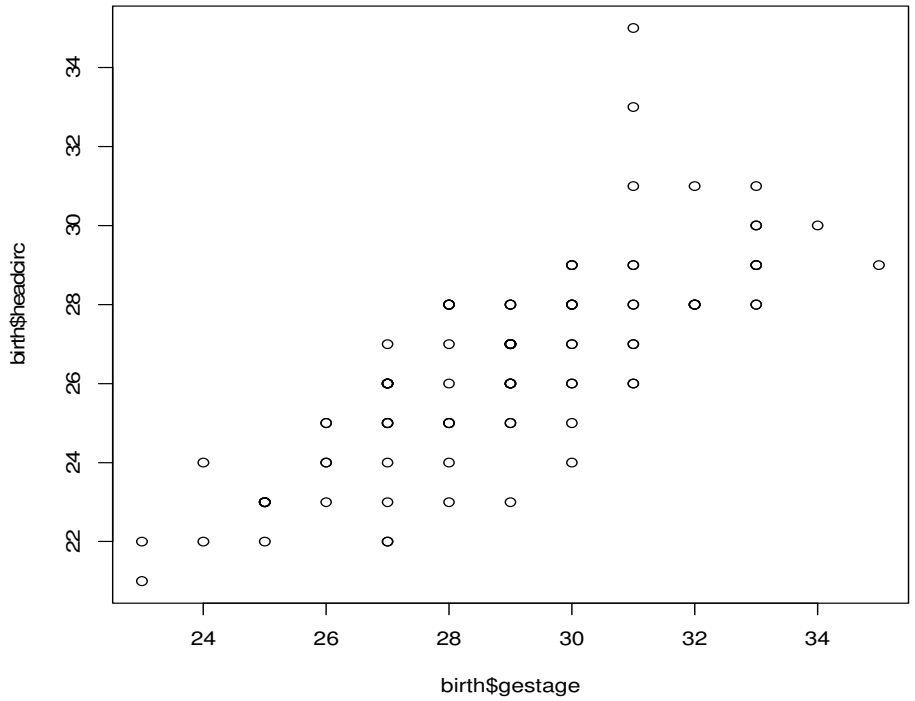
Example 1 The data ("low_birth_weight_infants.txt") is a sample of 100 low birth infants born in Boston.

(1) Is there any correlation between gestational age and head circumference from a scatter plot? If there is, what is the correlation coefficient?
(2) Find least squares regression line using summary statistics.
(3) Find least squares regression line using R. Are they the same?
(4) Plot the residuals. Do the residuals satisfy the conditions: **linearity , nearly normal and constant variability?**

Answer: (1)

```
> birth<-read.table("low_birth_weight_infants.txt",as.is=T,header=T,sep="\t")
> head(birth)
 headcirc length gestage birthwt momage toxemia
1       27     41      29    1360     37       0
2       29     40      31    1490     34       0
3       30     38      33    1490     32       0
4       28     38      31    1180     37       0
5       29     38      30    1200     29       1
6       23     32      25     680     19       0
```

>plot(birth$headcirc~birth$gestage)



From the scatter plot, there is a strong correlation.

(2)

>**cor(birth$gestage, birth$headcirc)**

The correlation coefficient is 0.78.

>**mean(birth$gestage)**
>**mean(birth$headcirc)**
>**sd(birth$gestage)**
>**sd(birth$headcirc)**

Using the results : $R=0.78,\ \bar{x}=28.89,\ s_x=2.534,\ \bar{y}=26.45,\ s_y=2.532$
we get

$$\hat{\beta}_1=\frac{s_y}{s_x}R=\frac{2.532}{2.534}(0.78)=0.78$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=26.45-0.78(28.89)=3.916$$

The regression line from summary statistics is $\hat{y}=3.916+0.78x$

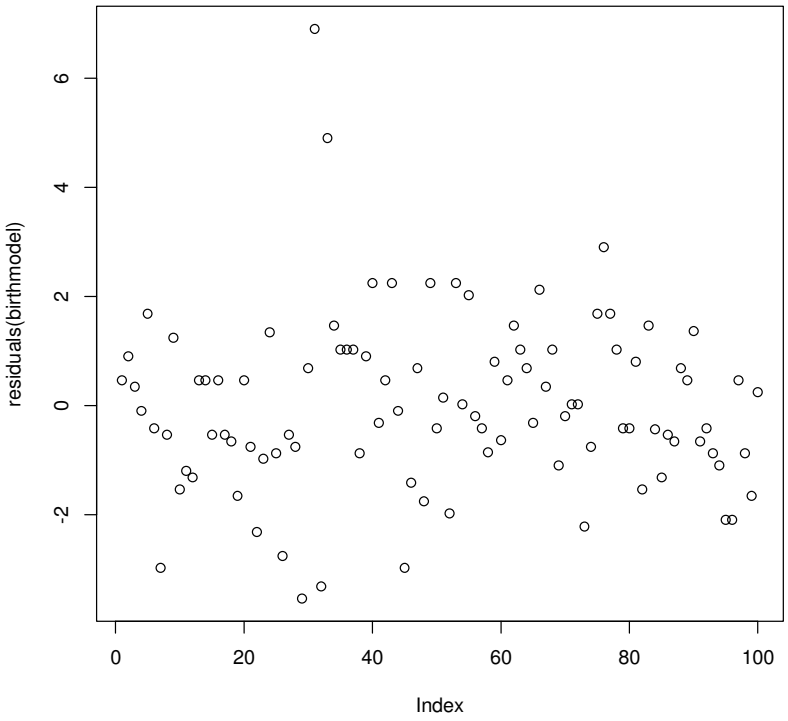
(3)

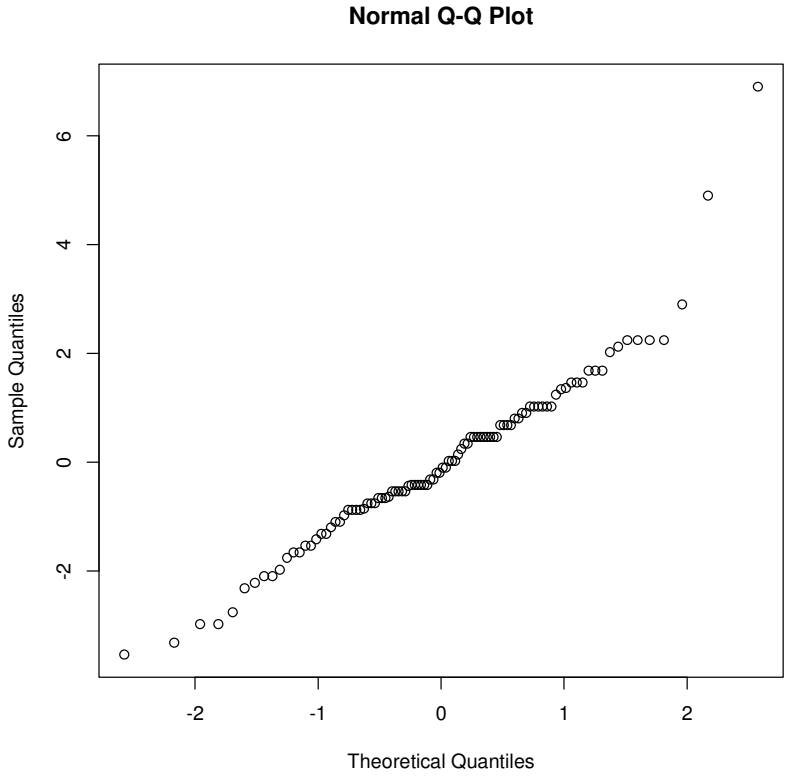
>**birthmodel<-lm(birth$headcirc~birth$gestage)**
>**summary(birthmodel)**

They are the same.

(4)

>**residuals(birthmodel)**
>**plot(residuals(birthmodel))**
>**qqnorm(residuals(birthmodel))**





The residuals qq plot looks fairly like a straight line. Also the residuals plot looks like constant variability. So the linear model fits quite well.

We can also use r to plot the least squares regression line.

>**plot(birth$headcirc, birth$gestage)**

>**abline(birthmodel)**