



JOINT INSTITUTE
交大密西根学院

VE401 Probabilistic Methods in Eng. Solution Manual for RC 7

Chen Xiwen

April 27, 2020

Linear Regression in Practice

Residual Analysis

Q. How to check that the linear model is an appropriate model for the data? How to check equal variances?

In simple linear regression, the model assumes that

$$Y|x_i = \beta_0 + \beta_1 x_i + E_i$$

and residuals can be used to approximate E_i , which are also random variables, are given by

$$\widehat{E}_i = Y|x_i - \widehat{Y}|x_i = \beta_0 + \beta_1 x_i + E_i - (B_0 + B_1 x_i),$$

which follows a normal distribution. We can calculate the mean

$$E[\widehat{E}_i] = \beta_0 + \beta_1 x_i + E[E_i - B_0 - B_1 x_i] = 0,$$

and variance

$$\begin{aligned} \text{Var}[\widehat{E}_i] &= \text{Var}[E_i - B_1 x_i - (\beta_0 + \beta_1 \bar{x} + \bar{E} - B_1 \bar{x})] \\ &= \text{Var}[E_i - \bar{E} + B_1(\bar{x} - x_i)] \\ &= \text{Var}\left[E_i - \bar{E} + (\bar{x} - x_i) \cdot \left(\beta_1 + \frac{\sum (x_j - \bar{x}) E_j}{S_{xx}}\right)\right] \\ &= \text{Var}\left[E_i - \bar{E} + \frac{(\bar{x} - x_i) \sum (x_j - \bar{x}) E_j}{S_{xx}}\right] \\ &= \sigma^2 + \frac{\sigma^2}{n} + \frac{(\bar{x} - x_i)^2 \sigma^2}{S_{xx}} - 2\text{Cov}[E_i, \bar{E}] + 2\text{Cov}\left[E_i, \frac{(\bar{x} - x_i) \sum (x_j - \bar{x}) E_j}{S_{xx}}\right] \\ &\quad - 2\text{Cov}\left[\bar{E}, \frac{(\bar{x} - x_i) \sum (x_j - \bar{x}) E_j}{S_{xx}}\right], \end{aligned}$$

where

$$\text{Cov}[E_i, \bar{E}] = \frac{1}{n} \sum_{j=1}^n \text{Cov}[E_i, E_j] = \frac{1}{n} \sigma^2,$$

$$\text{Cov}\left[E_i, \frac{(\bar{x} - x_i) \sum (x_j - \bar{x}) E_j}{S_{xx}}\right] = (\bar{x} - x_i) \sum_{j=1}^n (x_j - \bar{x}) \cdot \text{Cov}\left[E_i, \frac{E_j}{S_{xx}}\right] = -\frac{(\bar{x} - x_i)^2}{S_{xx}} \sigma^2,$$

$$\text{Cov}\left[\bar{E}, \frac{(\bar{x} - x_i) \sum (x_j - \bar{x}) E_j}{S_{xx}}\right] = \frac{\bar{x} - x_i}{n} \sum_{j=1}^n (x_j - \bar{x}) \text{Cov}\left[E_j, \frac{E_j}{S_{xx}}\right] = \frac{(\bar{x} - x_i) \sigma^2}{n S_{xx}} \sum_{j=1}^n (x_j - \bar{x}) = 0.$$

Therefore,

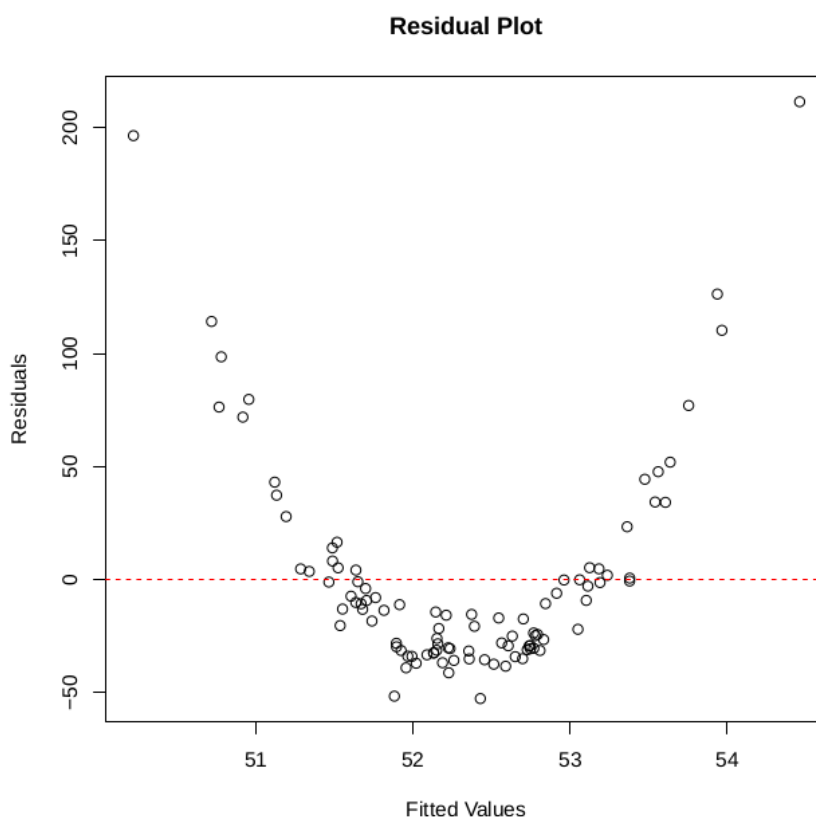
$$\text{Var}[\widehat{E}_i] = \text{Var}[Y|x_i - \widehat{Y}|x_i] = \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right) \sigma^2.$$

Note that this is different from the distribution of prediction

$$\text{Var}[\widehat{Y}|x - Y|x] = \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right) \sigma^2,$$

where in prediction, we are not choosing x to be equal to any of the x_1, \dots, x_n that are used to estimate β_1 and β_0 . When we assume that x_i is among those we use for estimation, there are additional covariance terms.

Therefore, if we plot the residuals, we can see when the point x_i is far from mean \bar{x} , the variance is smaller. In addition, the points should uniformly reside above the line $y = 0$ and below this line. This can be used to check for equal variance and linearity of our model. For instance, the following shows a residual plot with problems.

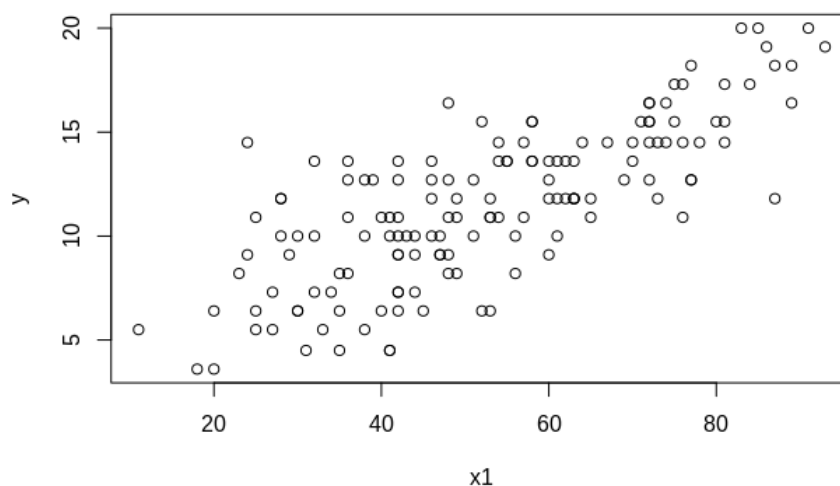


Linear Regression using R

In addition to Mathematica we use in the lecture, R is widely used for studies of data analysis. Here is an example of using R to perform a simple linear regression.

```
# load data
rc.df = read.table("data.txt", header = TRUE)

# plot data
plot(
  rc.df$resp,
  rc.df$var2,
  type = "p", xlab = "x1", ylab = "y"
)
```



```
# fit model and view model summary
rc.lm = lm(resp~var2, data = rc.df)
summary(rc.lm)
```

```
Call:
lm(formula = resp ~ var2, data = rc.df)

Residuals:
    Min       1Q   Median       3Q      Max
-39.980  -6.471   0.826   8.575  33.242

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.0845     3.2204   2.821  0.00547 **
var2          3.7859     0.2647  14.301 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.05 on 144 degrees of freedom
Multiple R-squared:  0.5868,    Adjusted R-squared:  0.5839
F-statistic: 204.5 on 1 and 144 DF,  p-value: < 2.2e-16
```

```
# residual plot
plot(fitted.values(rc.lm),
     residuals(rc.lm),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residual Plot",
     sub = "lm(y~x2)")

abline(a = 0, b = 0, lty = 2, col = "red")
```

