# VE401 Probabilistic Methods in Eng.
# Solution Manual for RC 7

Chen Xiwen

April 26, 2020

**University of Michigan–Shanghai Jiao Tong University Joint Institute**

# 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

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

and variance

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

where

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

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

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

Therefore,

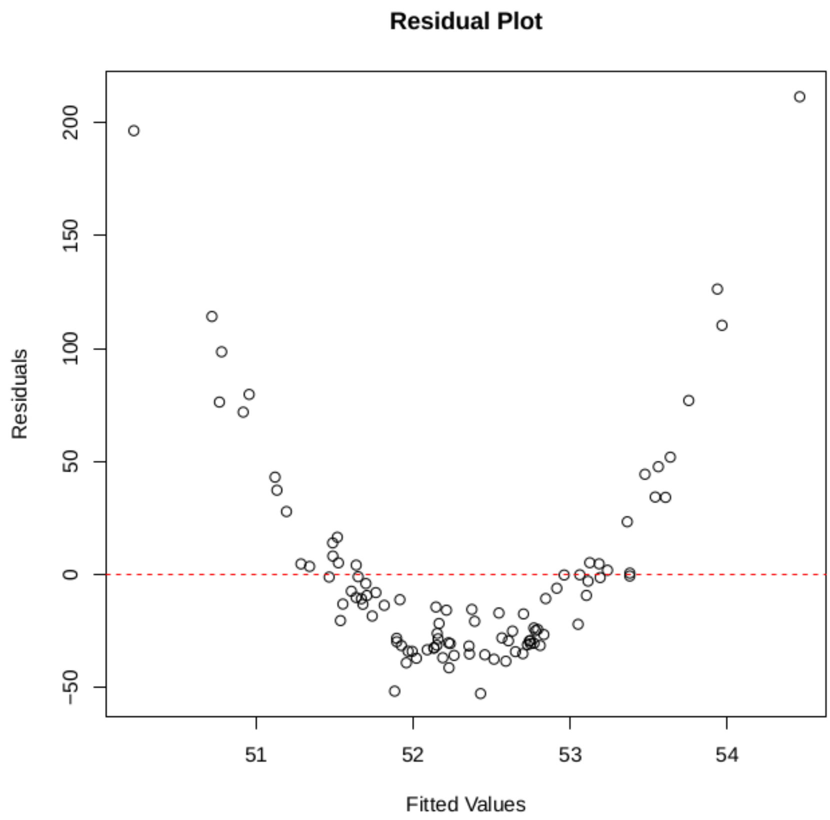
$$\mathrm{Var}[\widehat{E}_i] = \mathrm{Var}\left[Y|x_i - \widehat{Y|x_i}\right] = \left(1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}}\right)\sigma^2.$$

Note that this is different from the distribution of prediction

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

where in prediction, we are not choosing $x$ to be equal to any of the $x_1, \ldots, x_n$ that are used to estimate $\beta_1$ and $\beta_0$. 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 $\overline{x}$, 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.

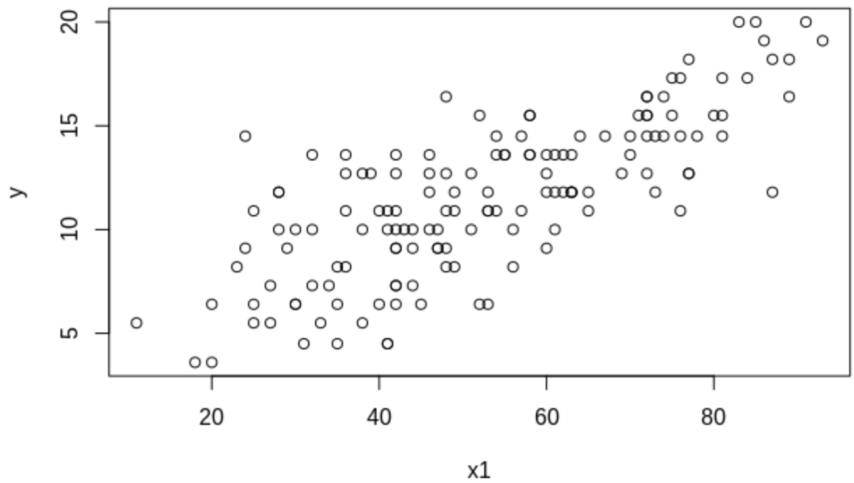



## Linear Regression using R

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)
```

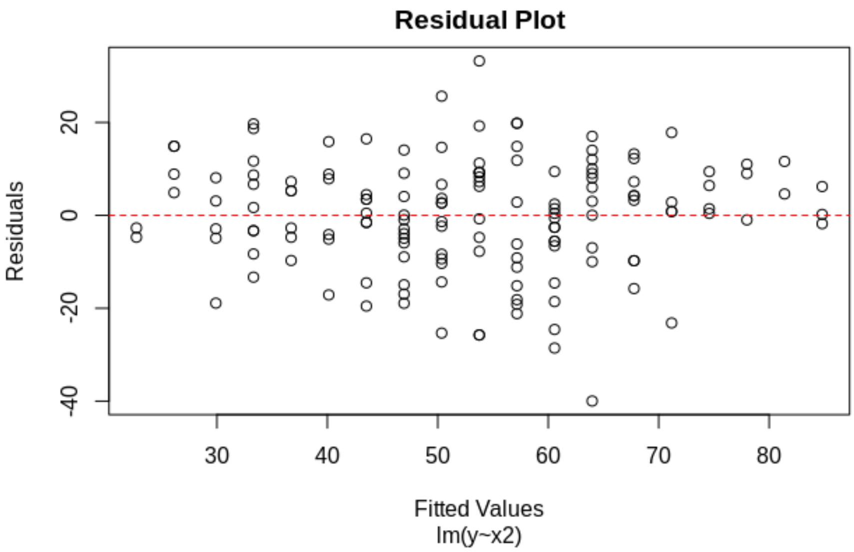
```
# plot data
plot(
    rc.df$resp,
    rc.df$var2,
    type = "p", xlab = "x1", ylab = "y"
)
```




```r
# 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
```

```r
# 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")
```


Residual Plot
Residuals
20
0
-20
-40
30
40
50
60
70
80
Fitted Values
lm(y~x2)