Regression Diagnostics

Dr. Kiah Wah Ong

ILLINOIS TECH

Introduction

In the previous sections, we fitted regression models to data by assuming all the hypotheses of the models are satisfied.

Now, we will explore how to use *R* to assess the validity of our regression models.

Model Assumptions

Recall the major assumptions we have made in linear regression models

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \cdots, n$$

are

- The relationship between the response and regressors is linear.
- The error terms ϵ_i have mean zero.
- The error terms ϵ_i have constant variance σ^2 (homoscedasticity)
- The error terms ϵ_i are normally distributed.
- The error terms ϵ_i and ϵ_j are uncorrelated for $i \neq j$.
- The regressors $x_1, \cdots, x_k$ are nonrandom.
- The regressors $x_1, \cdots, x_k$ are measured without error.
- The regressors are linearly independent.

Diagnostic Plots

We can examine residual plots of the regression model to detect model deficiencies in the linearity and the homoscedasticity assumptions.

To do this, remember that when fitting the linear model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

to a set of data by the least squares method, we obtain the fitted values as:

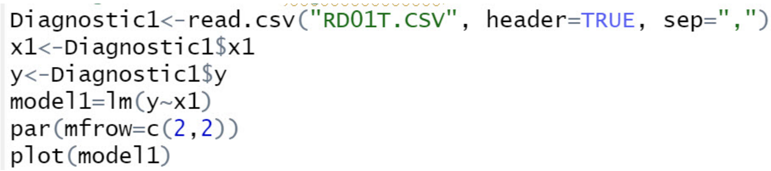
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}, \quad i = 1, 2, \cdots, n$$

Hence the corresponding ordinary least squares residue is given by

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n$$

Diagnostic Plots

Let us call upon the *R* built-in diagnostic plots for linear regression analysis. We do this by using the $\mathrm{Plot}()$ function as shown below:

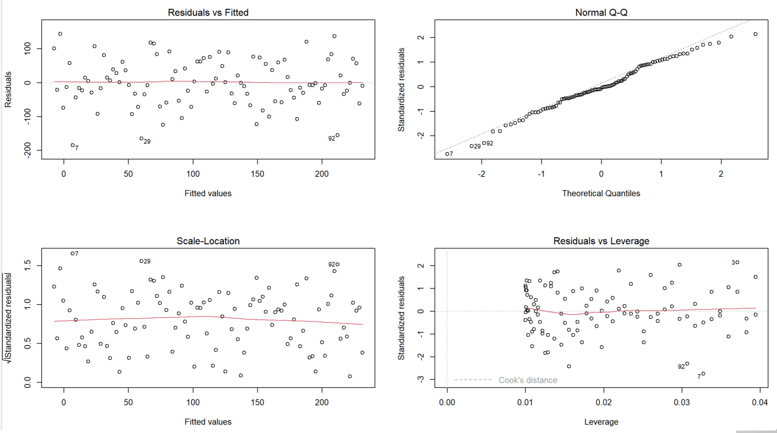


```
Diagnostic1<-read.csv("RD01T.CSV", header=TRUE, sep=",")
x1<-Diagnostic1$x1
y<-Diagnostic1$y
model1=lm(y~x1)
par(mfrow=c(2,2))
plot(model1)
```

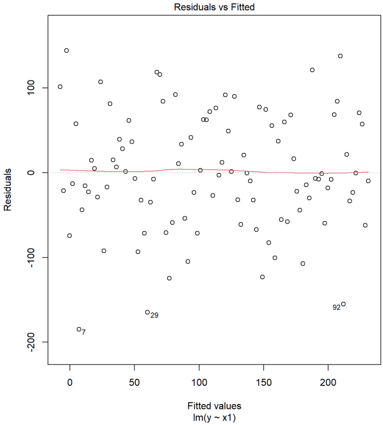
Note: The function $\mathrm{par}(\mathrm{mfrow}) = \mathrm{c}(2, 2)$ set up the regression model as an object and create a plotting environment of two rows and two columns.

Diagnostic Plots

These diagnostics include:
(i) Residuals vs. fitted values, (ii) Q-Q plots, (iii) Scale location plots, and (iv) Cook's distance plots.



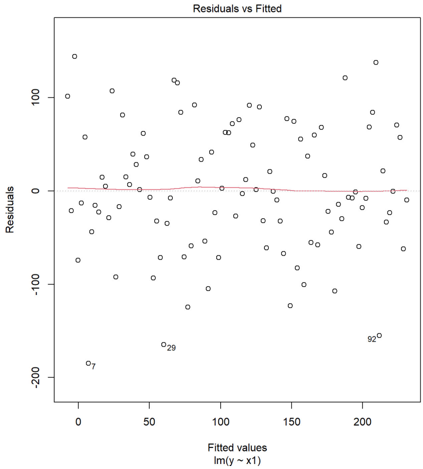
Residuals vs. Fitted Values



The plot is useful for checking the assumption of

- linearity
- homoscedasticity

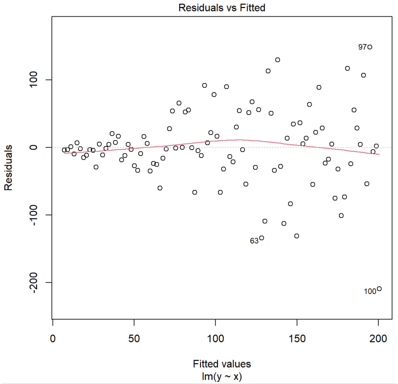
Residuals vs. Fitted Values



If the plot have non-linear patterns. There could be a non-linear relationship between regressors and respond.

If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships and homoscedasticity condition is satisfied.

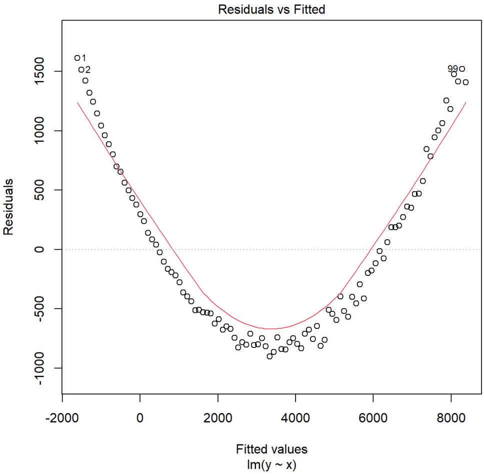
Residuals vs. Fitted Values



As oppose to a good fit plot we saw in the previous slide, the patterns you see on the left indicate that the variance of the errors is not constant (heteroscedasticity).

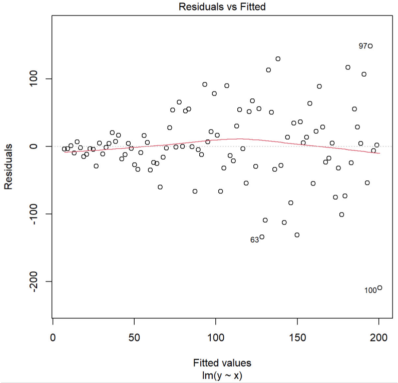
The outward - opening funnel pattern implies that the variance is an increasing function of y.

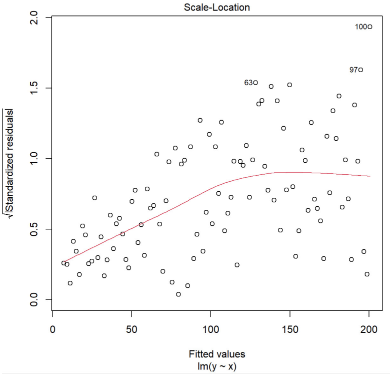
Residuals vs. Fitted Values



The patterns you see above indicate nonlinearity. Nonlinear terms are needed.

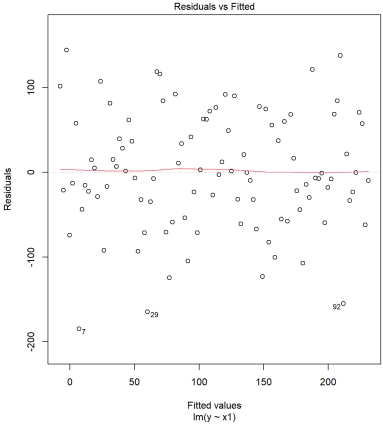
Scale-Location Plot

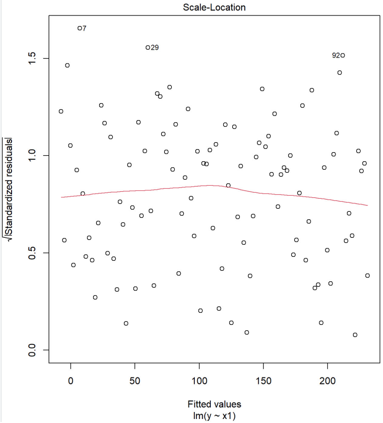




The scale-location plot can be used for evaluating homo/heteroscedasticity. It's good if you see a horizontal line with equally (randomly) spread points. The example above show the the two plots for a case where the variance are not constant. (See the non-horizontal line on the Scale-Location plot).

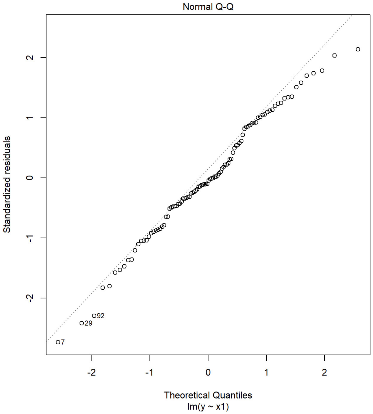
Scale-Location Plot





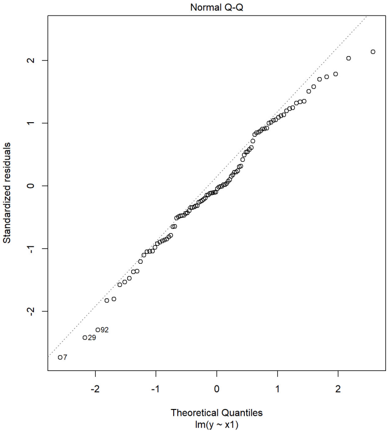
The scale-location plot above suggests that the variances of the error terms are equal (homoscedasticity.)

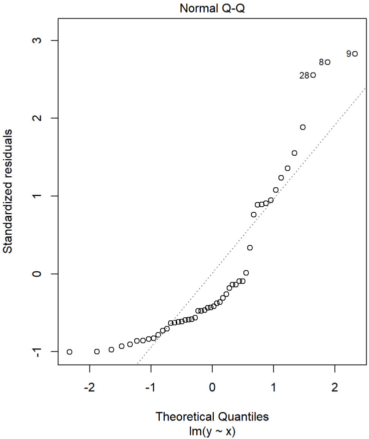
QQ Plot



The quantile-quantile (QQ) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution. When we run a regression analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption.

QQ Plot

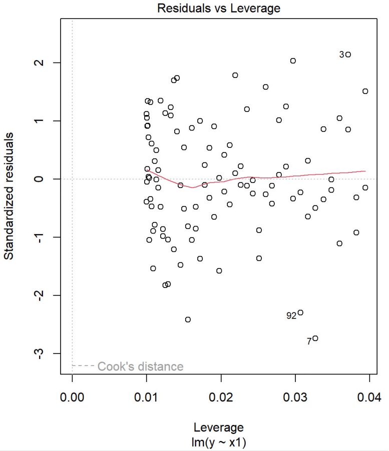


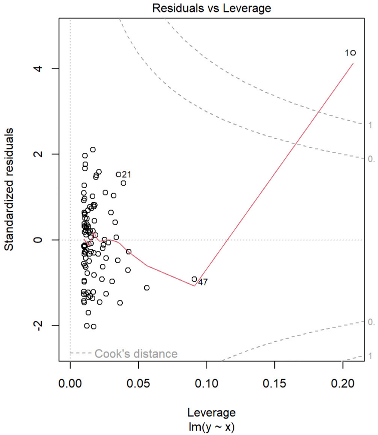


It's good if residuals are lined well on the straight dashed line (see left figure).

Note: QQ plot gives a visual check, not an air-tight proof and is also somewhat subjective.

Residuals vs Leverage





This plot helps us to find influential point (an observation that changes the slope of the line). We look for outlying values at the upper right corner or at the lower right corner (cases outside of the dashed lines). Those spots are the places where cases can be influential against a regression line.