STA3010 Regression Analysis

Feng YIN

The Chinese University of Hong Kong (Shenzhen)

yinfeng@cuhk.edu.cn

March 11, 2020

Overview

1. Model Adequacy Checking via Residual Analysis
   - Definition of Residual
   - Scaling Residuals
   - Residual Plots
2. Test for Lack of Fit

Definition of Residual

Aim

We concentrate on the multiple linear regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. We aim to verify if the random errors $\varepsilon_i, i = 1, 2, ..., n$ are i.i.d. and follow Gaussian distribution with zero mean and variance σ^2. There is no model mismatch problem.

Recall that the residuals are defined as:

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

When the least-squares (LS) estimator is applied, we have

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\,\mathbf{y} = (\mathbf{I} - \mathbf{H})\,\varepsilon, \tag{2}$$

where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is the "hat matrix" with various nice properties.

More about the Hat Matrix **H**

Some interesting properties of the hat matrix **H** include:

1. **H** is symmetric and idempotent
2. $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent
3. The i-th diagonal entry of **H**, defined by h_{ii}, satisfies $0 \leq h_{ii} \leq 1$. Here, $h_{ii} = \mathbf{x}_i^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_i$, where it is redefined that $\mathbf{x}_i = [1, x_{i,1}, x_{i,2}, ..., x_{i,k}]^T$ and $\mathbf{x}_i^T$ represents then the i-th row of **X**.
4. For the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$,

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}, \quad i = 1, 2, ..., n. \tag{3}$$

5. $\mathbf{Hy} = \mathbf{H\hat{y}}$, which implies $\mathbf{He} = \mathbf{0}$ and moreover $\sum_{j=1}^n e_j h_{ij} = 0$, $\forall i = 1, 2, ..., n$. (Used in Appendix C.8)
6. The eigenvalues of **H** consists of p ones and $n - p$ zeros.

Definition of Residual

If the random errors $\varepsilon_i, i = 1, 2, ..., n$ are indeed i.i.d. and follow Gaussian distribution with zero mean and variance σ^2, then we have

$$E(\mathbf{e}) = \mathbf{0}, \tag{4}$$

$$Cov(\mathbf{e}) = \sigma^2 (\mathbf{I} - \mathbf{H}). \tag{5}$$

Moreover, $\mathbf{e} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 (\mathbf{I} - \mathbf{H})\right)$.

The diagonal terms of $Cov(\mathbf{e})$ are $var(e_i) = \sigma^2(1 - h_{ii})$ and the off-diagonal terms are $cov(e_i, e_j) = -\sigma^2 h_{ij}$.

Derivations will be shown on the WB.

Scaling Residuals

Four Methods for Scaling Residuals

1. Standardized Residuals
2. Studentized Residuals
3. Standardized PRESS Residuals
4. R-Student Residuals

Standardized Residuals:

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}, \quad i = 1, 2, ..., n \tag{6}$$

where it is naively assumed that

$$var(e_1) = var(e_2) = ... = var(e_n) \approx MS_{Res}. \tag{7}$$

Scaling Residuals

Four Methods for Scaling Residuals

1. Standardized Residuals
2. Studentized Residuals
3. Standardized PRESS Residuals
4. R-Student Residuals

Studentized Residuals:

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}, \quad i = 1, 2, ..., n, \tag{8}$$

where the denominator is constructed due to the fact that:
(1) $var(e_i) = \sigma^2(1-h_{ii})$ and (2) MS_{Res} is an estimator of σ^2.

Scaling Residuals

Four Methods for Scaling Residuals

1. Standardized Residuals
2. Studentized Residuals
3. Standardized PRESS Residuals
4. R-Student Residuals

Studentized Residuals:

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}}, \quad i = 1, 2, ..., n.$$

The studentized residuals have approximately (however depending on the goodness of MS_{Res}) constant variance $var(r_i) = 1$ regardless of the location of an input.

Scaling Residuals

Four Methods for Scaling Residuals

1. Standardized Residuals
2. Studentized Residuals
3. Standardized PRESS Residuals
4. R-Student Residuals

For big data, we have in general Studentized Residual and Standardized Residual approximately equal, i.e.,

$$d_i \approx r_i. \tag{9}$$

This is because h_{ii} becomes small (close to zero) when n is very large, thus

$$MS_{Res} \approx MS_{Res}(1 - h_{ii}), \quad \forall i \in \{1, 2, ..., n\} \tag{10}$$

Scaling Residuals

Four Methods for Scaling Residuals

1. Standardized Residuals
2. Studentized Residuals
3. Standardized PRESS Residuals
4. R-Student Residuals

We define PREdiction Error Sum of Squares (PRESS) residuals:

$$e_{(i)} = y_i - \hat{y}_{(i)}, \quad i = 1, 2, ..., n, \tag{11}$$

where $\hat{y}_{(i)}$ is the fitted value of the i-th output based on all data points except for the i-th one.

It can be shown (on WB) that

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}, \; i = 1, 2, ..., n, \tag{12}$$

which is just the ordinary residuals scaled by $1 - h_{ii}$, and for larger h_{ii}, the corresponding PRESS residual is larger.

Scaling Residuals

Intermezzo I: PRESS Statistic is defined to be the sum of the squared PRESS residuals, concretely,

$$\mathrm{PRESS} \triangleq \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2 . \tag{13}$$

PRESS is generally regarded as a measure of how well a regression model will perform in predicting new data. We desire small value of PRESS.

Intermezzo II: The PRESS statistic can be used to compute an R^2-like statistic for prediction/overall model adequacy,

$$R^2_{\mathrm{predict}} = 1 - \frac{\mathrm{PRESS}}{SS_T} . \tag{14}$$

Scaling Residuals

Four Methods for Scaling Residuals

1. Standardized Residuals
2. Studentized Residuals
3. Standardized PRESS Residuals
4. R-Student Residuals

A standardized PRESS residual is defined as:

$$\frac{e_{(i)}}{\sqrt{var(e_{(i)})}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}} \tag{15}$$

equivalent to the studentized residual, when replacing σ^2 with MS_{Res}.

Scaling Residuals

Four Methods for Scaling Residuals

1. Standardized Residuals
2. Studentized Residuals
3. Standardized PRESS Residuals
4. R-Student Residuals

Instead of using MS_{Res} as the estimator of σ^2, we alternatively use

$$S_{(i)}^2 \triangleq \frac{\sum_{j\neq i}^{n}(y_j - \mathbf{x}_j^T\hat{\beta}_{(i)})^2}{n-p-1} = \frac{(n-p)MS_{Res} - e_i^2/(1-h_{ii})}{n-p-1}, \tag{16}$$

which is a "robust" estimator of σ^2 based on the dataset with the i-th data point removed.

The resulting externally studentized residual, or R-student residual, is

$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2(1-h_{ii})}}, \quad i = 1, 2, ..., n. \tag{17}$$

Scaling Residuals

Four Methods for Scaling Residuals

1. Standardized Residuals
2. Studentized Residuals
3. Standardized PRESS Residuals
4. R-Student Residuals

Under the assumption that the random errors are Gaussian i.i.d., it can be shown that the externally studentized residual, or R-student residual

$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2 (1 - h_{ii})}} \sim t_{n-p-1}. \tag{18}$$

[Optional Reading]: In our textbook, appendix C.9 established a formal hypothesis-testing procedure for outlier detection using R-student residuals.

Residual Plots

We mainly study the following three residual plots:

1. Normal probability plot
2. Plot of residuals against fitted values
3. Plot of residuals against regressor

Normal Probability Plot

First, take a look at the student-t pdfs and cdfs:

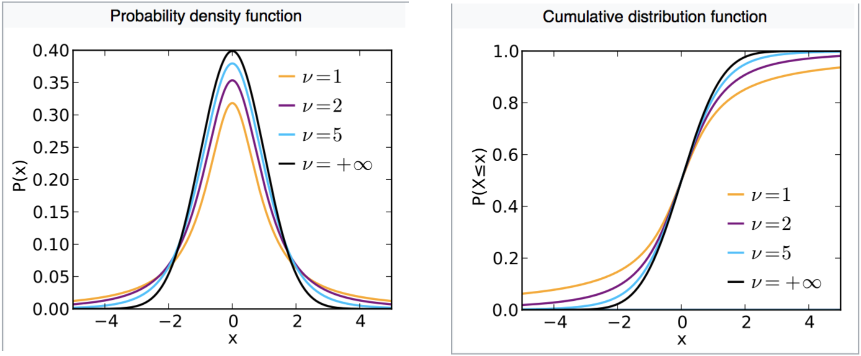


image source:wikipedia

Observation: a nearly straight line between around 0.33 to 0.67 cumulative probability points.

Normal Probability Plot

Normal probability plot is graphical analysis of residuals.

The procedure is as follow:

1. Sort the (preferred) externally studentized residuals in an increasing order such that $t_1 < t_2 < t_3 < ... < t_n$
2. Plot against the cumulative probability $P_i = \frac{i-0.5}{n}, i = 1, 2, ..., n$

Conclusion

Assuming that the multiple linear regression model is subject to zero mean Gaussian i.i.d. random errors, if we plot the externally studentized residuals against the cumulative probability $P_i = \frac{i-0.5}{n}, i = 1, 2, ..., n$ on the normal probability plot, the resulting points should lie approximately on a straight line between around 0.33 to 0.67 points.

Normal Probability Plot

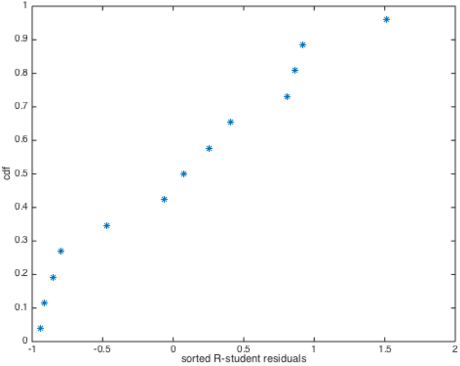
Rule-of-thumb:

- Small sample sizes ($n < 16$) often produce normal probability plots that deviate substantially from linearity.
- For larger sample sizes ($n > 32$) the plots are much better behaved.
- Usually about $n = 30$ points are required to produce normal probability plots that are stable enough to be easily interpreted.

Let us have a look at some examples!

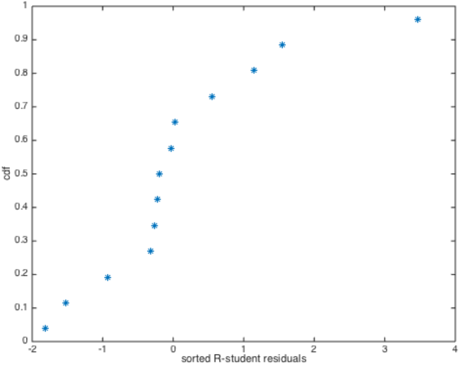
Normal Probability Plot

Example I-a: t_{10} distributed (close to normal), $n = 13$ data points



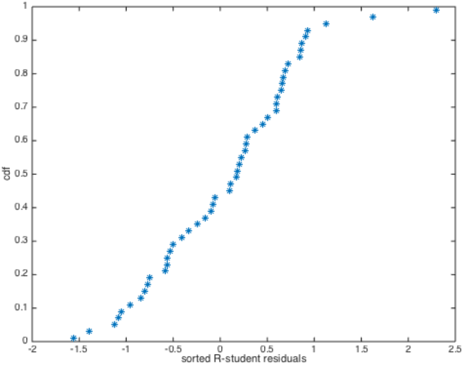
Normal Probability Plot

Example I-b: t_{10} distributed (close to normal), $n = 13$ data points



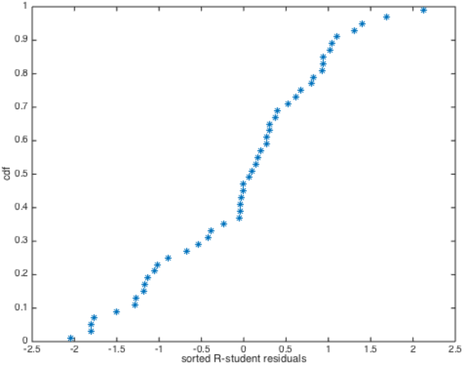
Normal Probability Plot

Example II-a: t_{20} distributed (close to normal), $n = 50$ data points



Normal Probability Plot

Example II-b: t_{20} distributed (close to normal), $n = 50$ data points



Normal Probability Plot

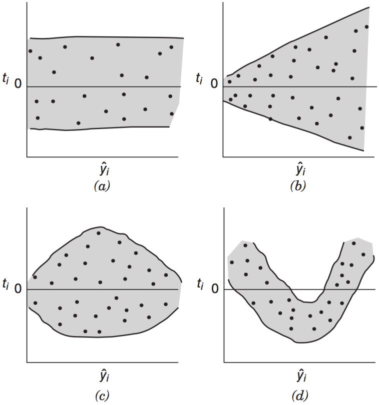
Cited from our textbook: Andrews [1979] and Gnanadesikan [1977] note that normal probability plots often exhibit no unusual behavior even if the errors ε_i are not normally distributed. This problem occurs because the residuals are not a simple random sample; they are the remnants of a parameter estimation process. The residuals are actually linear combinations of the model errors (the ε_i). Thus, fitting the parameters tends to destroy the evidence of non-normality in the residuals, and consequently we cannot always rely on the normal probability plot to detect departures from normality.

So, we'd better also plot the residuals against the fitted values.

Plot of Residuals against Fitted Values

Plot of the scaled residuals (preferred t_i) versus the corresponding fitted values is useful for detecting several common types of model inadequacies.

For instance:



Source: textbook

Plot of Residuals against Fitted Values

Note that the residuals should better be plotted versus the fitted values $\hat{y}_i$ and not the observed values y_i because the e_i and the $\hat{y}_i$ are uncorrelated while the e_i and the y_i are usually correlated.

Plot of Residuals against Fitted Values

Explanations to the above figures are as follows:

- Panel (a) indicates that the residuals are contained in a horizontal band, then there are no obvious model defects.
- Panels (b) and (c) indicate that the variance of the errors is not constant. Approaches for dealing with this issue: (1) apply a suitable transformation to either the input and/or output; (2) use the method of weighted least squares.
- Panel (d) indicates nonlinearity between the output and some input.

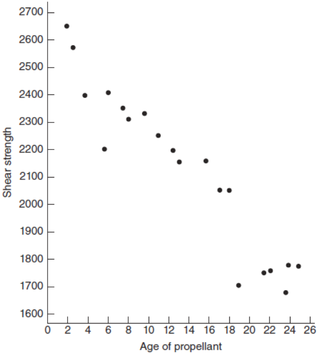
Plot of Residuals against Regressor

Plotting the residuals against the corresponding values of each regressor/input variable can also be helpful.

- Just replace fitted values with each input values x_{ij}, $i = 1, 2, ..., n$, $j \in \{1, 2, ..., k\}$.
- The conclusions drawn above for the fitted value plot hold in general.

Residual Plots: Real Example-I

Input: $x = \text{age of propellant}$, Output: $y = \text{shear strength}$
Data contains $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, $n = 20$ samples

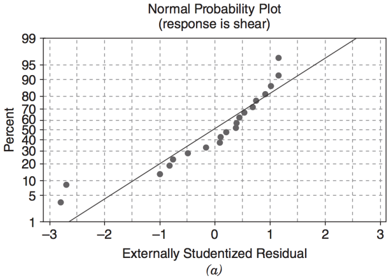


Scatter plot of shear strength versus propellant age

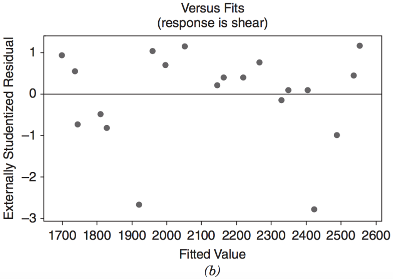
Two potential outliers $(x = 5.5, y = 2207)$, $(x = 19, y = 1708)$.

Residual Plots: Real Example-I

Input: x = age of propellant, Output: y = shear strength
Data contains $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, $n = 20$ samples
The least-squares fit is $\hat{y} = 2627.82 - 37.15x$



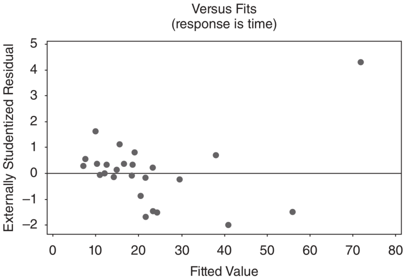
Normal probability plot



Residuals versus fitted $\hat{y}_i$

Residual Plots: Real Example-II

Figure below presents the plot of the externally studentized residuals versus the fitted values of delivery time.



Textbook Example 4.3: Plot of externally studentized residuals versus predicted for the delivery time data

The pattern indicates certain nonlinearity in the model. (The textbook thinks that the plot does not exhibit any strong unusual pattern,)

Test for Lack of Fit

Comparing the true/underlying model:

$$y = f(\mathbf{x}; \boldsymbol{\beta}) + \varepsilon \tag{19}$$

and the assumed multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon, \tag{20}$$

the following two error sources are identified to exist:

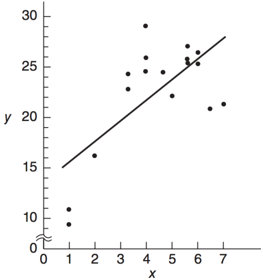
1. model mismatch error/lack of fit error between the true model and the assumed model (ignored in the previous lectures!)
2. random error ε

Test for Lack of Fit

Aim of the Test

We aim to test if the assumed multiple linear regression model is sufficient or more inputs and/or higher order terms may be added to improve the model fit.

Example:



There is some indication that the straight-line fit is unsatisfactory. Perhaps, a quadratic term x^2 or another input x_2 should be added

Test for Lack of Fit

Assumption

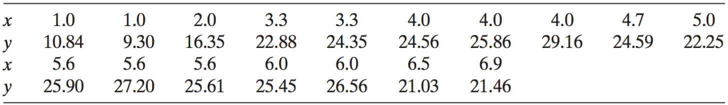
We still assume the random error terms are independently and identically Gaussian distributed with zero mean and covariance matrix $cov(\varepsilon) = \sigma^2\mathbf{I}$.

Requirement

The lack-of-fit test requires that we have replicate observations on the response y for at least one level of x. We emphasize that these should be true replications, not just duplicate readings or measurements of y. These replicated observations are used to obtain a model independent estimate of σ^2 (will be seen later).

Test for Lack of Fit

Suppose that we have n_i observations on the output at the i-th level of the input $x_i, i = 1, 2, ..., m$. Let y_{ij} denote the j-th observation on the output at $x_i, i = 1, 2, ..., m$ and $j = 1, 2, ..., n_i$. In total $n = \sum_{i=1}^{m} n_i$ observations.



x	1.0	1.0	2.0	3.3	3.3	4.0	4.0	4.0	4.7	5.0
y	10.84	9.30	16.35	22.88	24.35	24.56	25.86	29.16	24.59	22.25
x	5.6	5.6	5.6	6.0	6.0	6.5	6.9			
y	25.90	27.20	25.61	25.45	26.56	21.03	21.46			

The test procedure involves partitioning the residual sum of squares into two components, concretely,

$$SS_{Res} = SS_{PE} + SS_{LOF} \tag{21}$$

where SS_{PE} is the sum of squares due to pure error and SS_{LOF} is the sum of squares due to lack of fit.

Test for Lack of Fit

To develop this partitioning of SS_{Res}, note that the (ij)-th residual is

$$y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i) \tag{22}$$

where $\bar{y}_i$ is the average of the n_i observations at x_i. Squaring both sides and summing over i and j variables yields

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2. \tag{23}$$

where

$$SS_{PE} \triangleq \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2, \quad SS_{LOF} \triangleq \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2 \tag{24}$$

Test for Lack of Fit

It can be proven that if the multiple linear regression model is correct,

- $\frac{SS_{LOF}}{\sigma^2} \sim \chi^2_{m-p}$
- $\frac{SS_{PE}}{\sigma^2} \sim \chi^2_{n-m}$
- SS_{PE} and SS_{LOF} are independent

The test statistic is constructed by

$$F_0 = \frac{SS_{LOF}/(m-p)}{SS_{PE}/(n-m)} \sim F_{m-p,n-m} \tag{25}$$

To test for lack of fit, we would compute the test statistic F_0 and conclude that the regression function is insufficient if $F_0 > F_{\alpha,m-p,n-m}$. Alternatively, if $F_0 < F_{\alpha,m-p,n-m}$, there is no strong evidence of lack of fit.

Summary

To summarize with some keywords:

- Model adequacy checking
- Scaling residuals (four types)
- Press statistics
- Residual plots (three types)
- Test for lack of fit (F-test)
- Properties of Hat matrix