STAT 380: Regression Diagnostics

Subhadip Pal

Simple Linear Regression: Model Selection Criterion

The Linear Fit

- The simple linear regression line is given by

$$Y = \alpha + \beta X + \varepsilon. \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

The parameters of the models are α, β, σ^2.

The corresponding estimates are $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$. We have seen expression of these estimators in terms of the observed data in one of the previous slide.

Let (X_i, Y_i) be the observed value of the i^{th} data points, $i = 1, \ldots, n$.

The corresponding predicted response is $\hat{Y}_{\mathrm{i}} = \hat{\alpha} + \hat{\beta} X_{\mathrm{i}}$

Measures of Goodness of Fit

R^2 Goodness of Fit

$\text{ESS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}}$

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{Y})^2$$

$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2$$

Result:

$$\text{TSS} = \text{ESS} + \text{SSR}.$$

$$R^2 = 1 - \frac{ESS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{Y})^2}$$

Adjusted-R^2 Goodness of Fit

$$\boxed{\text{ESS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

$$\boxed{\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{Y})^2}$$

$$\boxed{\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2}$$

$$R^2_{\text{adj}} = 1 - \frac{\frac{ESS}{n-p}}{\frac{TSS}{n-1}} = \boxed{1 - \frac{n-1}{n-p}(1 - R^2)}.$$

here $p = 2$ as we have only two regression coefficients, namely α, β.

F Statistic: Model Fitness

$$\text{ESS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{Y})^2$$

$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2$$

Result:

$$\text{TSS} = \text{ESS} + \text{SSR}.$$

$$F = \frac{\frac{SSR}{p-1}}{\frac{SSE}{n-p}} \sim \text{F}_{(p-1),(n-p)}\text{df}$$

p denotes the total number of regression coefficients including the intercept. Here $p = 2$ as we have only two regression coefficients α, β.

Model Selection Criterion: AIC, BIC

Likelighood of The Linear Statistical Model

$\underset{\sim}{\mathbf{y}} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}.$

$\underset{\sim}{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I_{n\times n})$

Likelihood of the data:= Probability Density of the Response

Likelihood:= $\boxed{\frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{(\underset{\sim}{\mathbf{y}}-X\underset{\sim}{\beta})^T(\underset{\sim}{\mathbf{y}}-X\underset{\sim}{\beta})}{2\sigma^2}}}$

Log Likelihood:= $\boxed{n\log\left(\sqrt{2\pi}\sigma\right) - \frac{(\underset{\sim}{\mathbf{y}}-X\underset{\sim}{\beta})^T(\underset{\sim}{\mathbf{y}}-X\underset{\sim}{\beta})}{2\sigma^2}}$

Maximum Likelihood Principle: **Larger Value for the Likelihood/ Log Likelihood indicates a better model fit.**

Model selection: Akaike's Information Criteria (AIC)

The AIC Criterion for a General model is:

$$\boxed{AIC = 2\,p - 2 \times \text{Log-Likelihood}\,(\text{Model})\,.}$$

Here p: denotes the model dimension/ model complexity.

Identify the one among the competing models that have the **smallest AIC.**

Model selection: Schwarz's Bayesian information criterion (BIC)

The BIC Criterion for a General model is: $\boxed{BIC = p\log(n) - 2 \times \text{Log-Likelihood (Model)}\,.}$ Here p: denotes the model dimension/ model complexity.

Identify the one among the competing models that have the **smallest BIC.**

Outlier, Influential Points and Leverages

Assumptions of SLR

The observed data: $\{Y_i, X_i\}_{i=1}^n$. According to the model: $Y_i = \alpha + \beta X_i + \varepsilon_i$

1. $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$, $\sigma^2 > 0$.
2. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are *statistically independent* (i.e. random and unrelated.)
3. σ^2 (likely to be unknown) is constant. It does not depend on the responses of the covariates.

Apart from checking the model assumptions, we also need to be cautious if there is any outlier/influential points in the data.

Assumptions of SLR

Need to check whether the data set does not violate any of the assumptions mentioned in the previous slide.

What else can go wrong?

1. Regression function can be wrong- missing predictors, nonlinearity
2. Outliers: both in predictors and observations.
3. Influential Points: these points have high influence on the regression function.

Structure of Statistical Hypothesis Testing

There is a **NULL Hypothesis** and **Alternative** Hypothesis

Null is Denoted as H_0, Alternative is denoted as H_1 or H_a

General Procedure for Statistical Hypothesis Testing:

Step 1 Test Statistics: We calculate the value of the Test Statistic from the particular Data.

Step 2 We obtain/know the 'Probability Distribution ' of the Test Statistic Assuming 'NULL' hypothesis to be True.

Step 3 Calculate corresponding p-Value of the Test that uses the Distribution in Step2 and calculated Statistic in Step1.

tep 4:

1. If p-value is small: We **have strong statistical evidence to Reject** the Null hypothesis.
2. If p-value is Large: We **do not have enough statistical evidence to Reject** the Null hypothesis.

Structure of Statistical Hypothesis Testing

There is a **NULL Hypothesis** and **Alternative** Hypothesis

Null is Denoted as H_0, Alternative is denoted as H_1 or H_a

General Procedure for Statistical Hypothesis Testing:

Step 1 Test Statistics: We calculate the value of the Test Statistic from the particular Data.

Step 2 We obtain/know the 'Probability Distribution ' of the Test Statistic Assuming 'NULL' hypothesis to be True.

Step 3 Calculate corresponding p-Value of the Test that uses the Distribution in Step2 and calculated Statistic in Step1.

tep 4:

1. If p-value is small: We **have strong statistical evidence to Reject** the Null hypothesis.
2. If p-value is Large: We **do not have enough statistical evidence to Reject** the Null hypothesis.

Structure of Statistical Hypothesis Testing

There is a **NULL Hypothesis** and **Alternative** Hypothesis

Null is Denoted as H_0, Alternative is denoted as H_1 or H_a

General Procedure for Statistical Hypothesis Testing:

tep 1 Test Statistics: We calculate the value of the Test Statistic from the particular Data.

tep 2 We obtain/know the 'Probability Distribution ' of the Test Statistic Assuming 'NULL' hypothesis to be True.

tep 3 Calculate corresponding p-Value of the Test that uses the Distribution in Step2 and calculated Statistic in Step1.

tep 4:

1. If p-value is small: We **have strong statistical evidence to Reject** the Null hypothesis.
2. If p-value is Large: We **do not have enough statistical evidence to Reject** the Null hypothesis.

Structure of Statistical Hypothesis Testing

There is a **NULL Hypothesis** and **Alternative** Hypothesis

Null is Denoted as H_0, Alternative is denoted as H_1 or H_a

General Procedure for Statistical Hypothesis Testing:

Step 1 Test Statistics: We calculate the value of the Test Statistic from the particular Data.

Step 2 We obtain/know the 'Probability Distribution ' of the Test Statistic Assuming 'NULL' hypothesis to be True.

Step 3 Calculate corresponding p-Value of the Test that uses the Distribution in Step2 and calculated Statistic in Step1.

tep 4:

1. If p-value is small: We **have strong statistical evidence to Reject** the Null hypothesis.
2. If p-value is Large: We **do not have enough statistical evidence to Reject** the Null hypothesis.

Structure of Statistical Hypothesis Testing

There is a **NULL Hypothesis** and **Alternative** Hypothesis

Null is Denoted as H_0, Alternative is denoted as H_1 or H_a

General Procedure for Statistical Hypothesis Testing:

Step 1 Test Statistics: We calculate the value of the Test Statistic from the particular Data.

Step 2 We obtain/know the 'Probability Distribution ' of the Test Statistic Assuming 'NULL' hypothesis to be True.

Step 3 Calculate corresponding p-Value of the Test that uses the Distribution in Step2 and calculated Statistic in Step1.

tep 4:

1. If p-value is small: We **have strong statistical evidence to Reject** the Null hypothesis.
2. If p-value is Large: We **do not have enough statistical evidence to Reject** the Null hypothesis.

Measure of Leverage to Detect Influential Points

Reminder: Confidence Interval for Predicted Responses

Leverage: A leverage point is an observation that has an unusual covariate/predictor value (very different from the majority of the observations).

Influence point : An influence point is an observation whose removal from the data set would cause a large change in the estimated regression model coefficients).

Measure of Leverage to Detect Influential Points

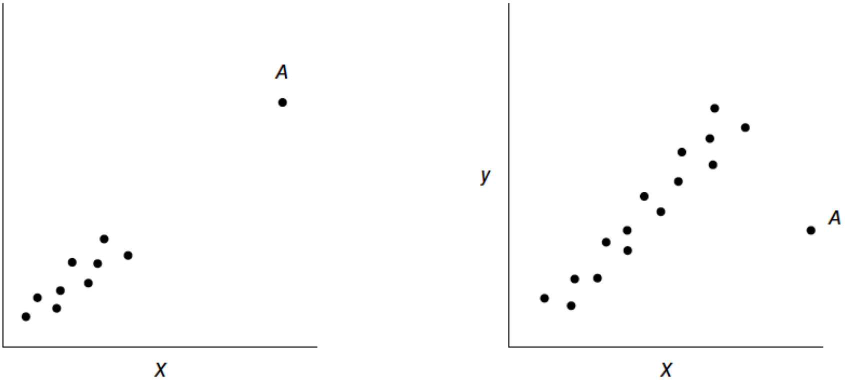
- A measure of Leverage for the i^{th} Data-point (X_i, Y_i) is given as:

$$\boxed{h_{i,i} := \frac{1}{n} + \frac{\left(X_{\mathrm{i}} - \bar{x}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$\frac{1}{n} \leq h_{i,i} < 1$$

A large value of the leverage means the corresponding data point is far away from the data center. It might be a point of significant influence to the regression coefficients.

Measure of Leverage to Detect Influential Points



Cook's Distance to Detect Influential Points

- Standardized Residual residual: $z_i := \frac{\hat{Y}_i - Y_i}{\hat{\sigma}}$: Based on the linear regression assumptions, we might expect the z_i's to resemble a sample from a N(0, 1) distribution.

- Studentized Residual: It is defined to be $e_i := \frac{\hat{Y}_i - Y_i}{\hat{\sigma}\sqrt{1 - h_{i,i}}}$. : approximately follows a t distribution with $n - p$ degrees of freedom (under the standard assumptions of the SLR). In large data sets (large n), the standardized and studentized residuals should not differ dramatically.

- Cook's Distance:

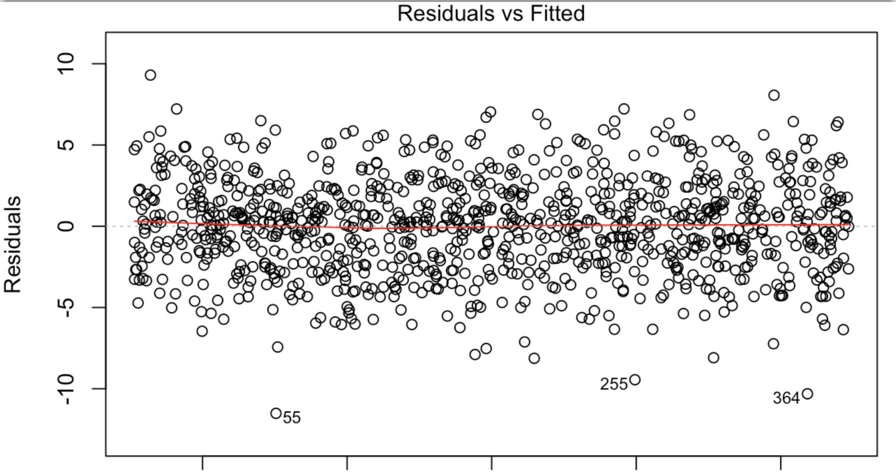
$$\mathrm{CD}_i := \frac{e_i^2}{p} \frac{h_{i,i}}{1 - h_{i,i}}.$$

It is suggested to examine the cases with $\mathrm{CD}_i > 0.5$ and that cases with $\mathrm{CD}_i > 1$ can be highly influential.

Residual Plots

A Typical Desirable Residual Plot

Residual Plot: Graph of Residuals against Predictor variable or against the fitted values is helpful to see if the variance of error terms are constant.



Residual Plots

■ The plots of the residuals, Studentized Residual and Standardized residuals can be utilized to visually identify key insights, not only in identifying influential points, but also many additional information. regarding the regression diagnostics.

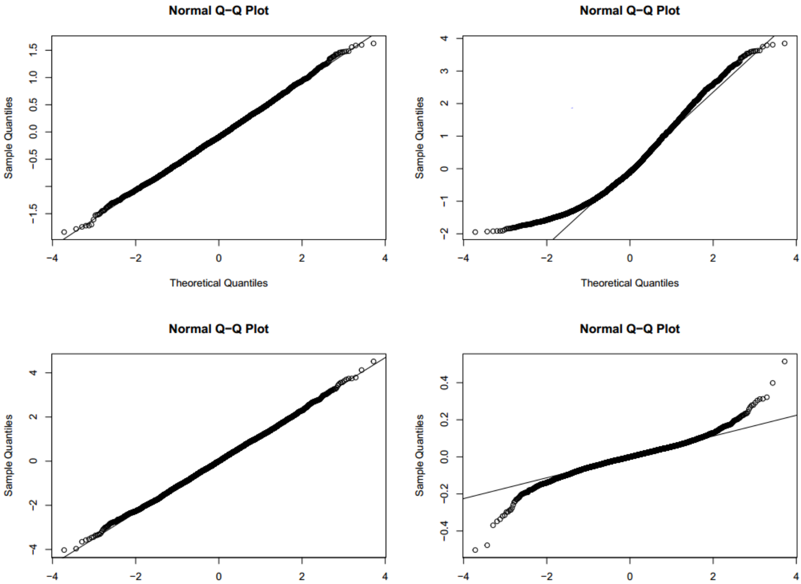
Detection for Lack of Normality of Residuals: Shapiro-Wilk's Test for Residuals

Normal Q-Q Plot

Normal Q-Q Plot: Theoretical quantiles (percentiles) from the standard normal distribution are plotted against the empirical quantiles of the standardized residuals.

In ideal scenario, when there is no model assumptions violation on normality of the residuals, all the points on a Normal Q-Q plot for the standardized residuals should be on the (very close to the) X=Y straight line.

Examples: Q-Q Plots



Statistical Test of Normality for Residuals: Shapiro-Wilk test

It tests for whether the errors/residuals are Normally distributed or nor.

H_0 : Residuals are Normaly Distributed

vs

H_1 : Residuals are not Normaly Distributed

Statistical Test of Normality for Residuals: Shapiro-Wilk test

It tests for whether the errors/residuals are Normally distributed or nor.

H_0 : Residuals are Normaly Distributed

vs

H_1 : Residuals are not Normaly Distributed

The test statistic is Denoted as W. $0 < W \leq 1$. W is a fraction.

Decision: We have strong statistical evidence to Reject the Null Hypothesis if the corresponding p-value is less than .05.

If Assumption of normality are true : the p-value is **more then .05** then there is no statistical evidence to believe that the residuals are not Normally distributed.

A Possible Remedy to Deal with Non-Normality of the Residuals.

Box-Cox Power Transformation of the Responses

The power transformation is parametrized by λ (a real value can be positive negative or zero.)

$$\boxed{\tilde{Y} = \begin{cases} \log(Y) & \text{if } \lambda = 0 \\ \frac{Y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \end{cases}}$$

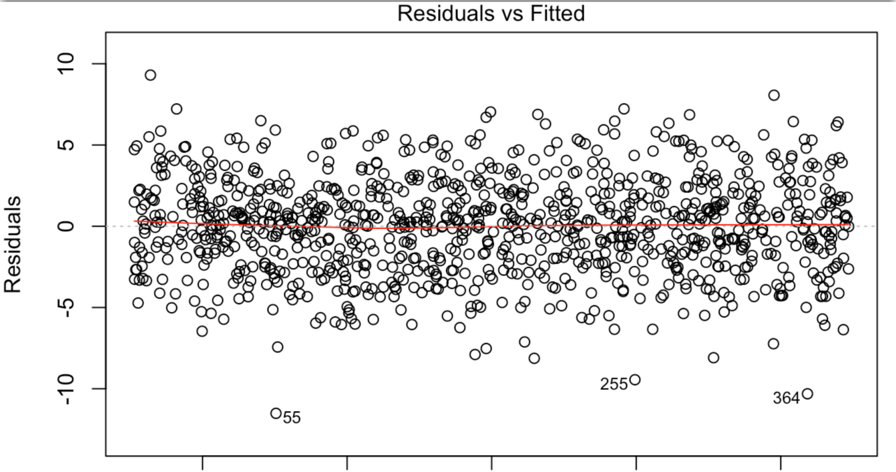
Optimal λ is often estimated by a cross validation procedure. Standard Statistical software typically provide a procedure to identify its optimal value.

Box-Cox Transformation may resolve the issues due to the non normality of the residuals and heteroscedasticity.

Detection of heteroscedasticity (Non constant σ^2)

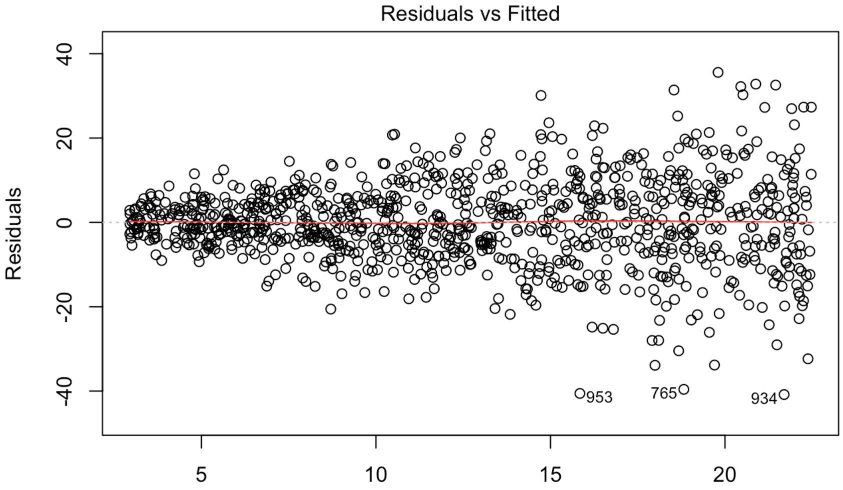
Identification of non-constant Error Variance

Residual Plot: Graph of Residuals against Predictor variable or against the fitted values is helpful to see if the variance of error terms are constant.



Identification of non-constant Error Variance

Residual Plot: Variability of the residuals appears to be increasing.

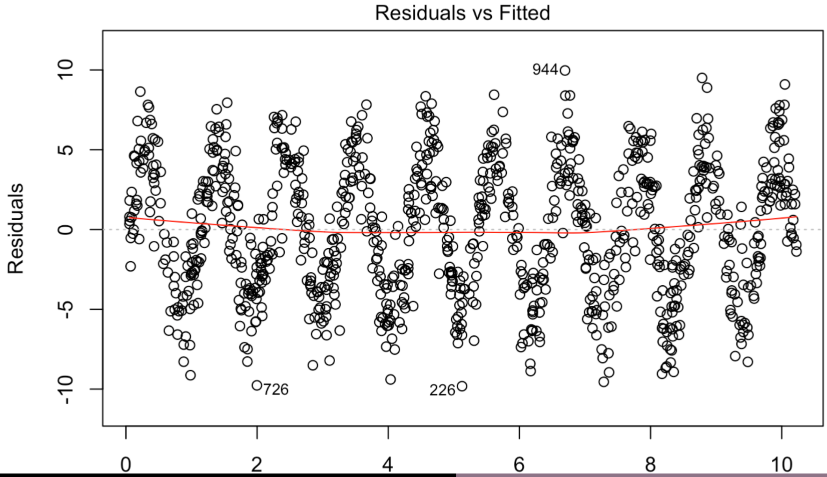


- A Possible remedy to deal with non constant variability is to utilize weighted Least Square.
- Use appropriate transformation on the Responses (Y).

Detection of Correlated Model Errors (ε_i): Durbin-Watson Test

Identification of Non Linearity

Residual Plot: Violation of the Linearity Assumption. The errors are not statistically independent/uncorrelated. There seem to be a seasonality.



Darbin Watson Test (DW Test) for identifying Correlated Model Errors

It tests for whether the errors/residuals are auto-correlated or not. Let ρ denotes the lag 1 auto correlation between the errors. i.e. If corr$(\varepsilon_i, \varepsilon_{i-1}) = \rho$ then it tests for

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0$$

Define: $e_i^* := \hat{Y}_i - Y_i$.

$$TestStatistic = d = \frac{\sum_{i=1}^{n} \left(e_i^* - e_{i-1}^*\right)^2}{\sum_{i=1}^{n} e^{*2}}$$

Reject Null Hypothesis if *d* is large or small. ‘significantly large’ value of d indicates negative correlation, ‘significantly small’ values of d is indicative of positive correlation

Assumption are met (errors are not correlated): If the corresponding p-value of the test is LARGER than .05

□ A Possible remedy is to use utilize Time series models such as AR, ARMA, ARIMA, GARCH model. Uf there is seasonality in the model try to include some periodic function along with the linear functions.

Thank You