Data Based Evaluations of the Performances of a Regression Model

Jude Chukwura Obi
Department of Statistics,
Chukwuemeka Odumegwu Ojukwu University, Anambra State, Nigeria
jco.coou@gmail.com

Abstract
The regression model is parametric and thus has a number of assumptions that necessarily should be satisfied prior to using it. Aware that the real-world datasets hardily comply with all regression assumptions, this study has taken into consideration, a number of datasets often encountered in the real-world. They include linear dataset, non-linear dataset, dataset with failed normality assumptions, dataset with outliers and dataset with correlated observations. In all, it has been shown that optimal performance of the regression model is guaranteed on the linear dataset. Furthermore, the study importantly reveals that the use of one statistic, the Root Mean Square Error (RMSE), say, is not enough in assessing the performance of a regression model. Adding another statistic like the Coefficient of Determination (R^2) *will be adequate.*

Keywords: Regression, Supervised Learning, Statistical Learning, Variable Selection.

1 Introduction

The regression model is one instance of a supervised learning model (Nasteski, 2017), in that it is predictive and characterised by the presence of input and output. The output is basically the response variable whereas the input is the explanatory variable when only one variable is involved or explanatory variables when we have the presence of two or more variables. Symbolically, the regression model can be represented as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad (1.1)$$

where $\mathbf{y}$ is a vector of observation and X, a design matrix. Also, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\boldsymbol{\epsilon}$ a vector of independent error terms.

The regression model is parametric, hence the optimum performance of the model is dependent on meeting a number of underlying assumptions (Berry, 1993). These assumptions include:

- Linearity assumption, meaning that the regression model is linear in parameter. For instance, consider
$$Y = \beta_0 + (\beta_1 X_1) + (\beta_2 X_2^2),$$
and note that although X_2 is raised to power 2, the equation is still linear in parameter (Obi, 2020).
- $E(\epsilon_i) = 0$, meaning that the mean of residual is zero.
- Homoscedasticity of residuals or equal variances.
- The X variables and the residuals (ϵ_i) are uncorrelated, hence $E(\epsilon_i X_j) = 0$.
- $\epsilon \sim NID(0, \sigma^2)$.
- The number of observations (n), must be greater than the number of X_s.

A common belief is that where the assumptions of the regression model are met, the predictive accuracy of the model is optimum. In other words, the model performs optimally. On the other hand, what happens if the assumptions are not entirely met? Could the output of the model be trusted in such instances? Similarly, if all the assumptions are not entirely met, could one still be confident to use the regression model?

In the light of these questions, it has become imperative to carry out a data-based investigation on the performances of the regression model. Such investigation is anchored on the fact that different datasets have varying degrees of conformity with the regression assumptions. We would therefore like to know about the behaviour of the model on datasets with varying degree of conformity with the model's assumptions.

1.1 The Simple and Multiple Linear Regression

The simple linear regression is a term used to refer to a regression problem with one explanatory variable. The model symbolically can be written as

$$Y = \beta_0 + \beta X + \epsilon. \qquad (1.2)$$

One advantage working with a simple regression is that your data can be easily graphed, thereby affording one a first-hand impression on the conformity with the linearity assumption. The regression model may not perform optimally if a given dataset fails this assumption. On the other hand, the multiple regression is characterized by the presence of two or more explanatory variables in the regression model. It gives room to study the impact of several variables in determining the model's output. A multiple regression model can symbolically be represented as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon. \qquad (1.3)$$

The model presented in (1.3) shows that we have p explanatory variables, where $p \in \mathbb{R}$. One drawback working with the multiple regression model is that the regression graph here is $p + 1$ dimensional, and the moment the number of explanatory variables exceeds two, it is usually difficult to represent the response and explanatory variables graphically.

2 Aim and Objectives of the Study

- On the strength of data-based evidence (a post-evaluation outcome), this study aims to discover the strengths and weaknesses of the regression model.

The following objective have also been targeted:

- To determine which regression assumption will cause the most adverse effect on optimum performance of the regression model, when not particularly complied with.
- To determine the conditions that guarantee optimum performances of the regression model.
- To determine the factors that further contribute to very poor performances of the regression model based on the data used in the study.
- To determine the extent of variation in the dependent variable explained by the independent variables, given every dataset used in the study.

3 Research Methodology

This section concerns how this study has been designed to ensure valid and reliable results that address the research aim and objectives. To discuss here are the datasets, statistical tools to be used in analysis and some inferential procedures.

3.1 Datasets

The datasets for this analysis will be simulated using the R statistical package (Team & others, 2013). The datasets will address the following cases:

- Linear separability or otherwise in datasets.
- Datasets with failed normality assumption.
- Datasets with outliers.
- Datasets with correlated observations.

A procedure for simulating the datasets are contained in the Appendices.

3.2 Tools for Statistical Analysis

The simple linear regression tool will be used extensively for the analysis that follows in section three. Also, graphical illustrations will be provided on the underlying relationship between the input and output variables for all datasets used in the analysis. For these reasons, in the subsections that follows, we shall discuss how to obtain the parameters of a simple linear regression.

3.2.1 Estimations of Parameter of a Simple Linear Regression

The model of a simple linear regression is given in **Error! Reference source not found.**. The parameters to be estimated as shown in the model are β_0 and β. Both parameters can be estimated using either the method of the least squares or the simultaneous equation method. Here, we shall use the least squares method.

Continuing, the random error term (ϵ) can be defined as:

$$\epsilon_i = y_i - \beta_0 - \beta x_i; i = 1,2,\cdots,n \tag{3.1}$$

The sum of squares can be written as:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta x_i)^2 \tag{3.2}$$

To obtain estimate for β_0,

$$\frac{\partial L}{\partial \beta_0} = 0 \Longrightarrow 2\sum_{i=1}^{n} (y_i - \beta_0 - \beta x_1) \times (-1) = 0$$

$$\sum_{i=1}^{n} y_i - n\beta_0 - \beta \sum_{i=1}^{n} x_i = 0$$

$$n\beta_0 = \sum_{i=1}^{n} y_i - \beta \sum_{i=1}^{n} x_i$$

It follows that

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta} \sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}\bar{x}. \tag{3.3}$$

Similarly,

$$\frac{\partial L}{\partial \beta} = 0; \Longrightarrow 2\sum_{i=1}^{n} (y_i - \beta_0 - \beta x_i) \times x_i = 0$$

$$\sum_{i=1}^{n}(x_i y_i - \beta_0 x_i - \beta x_i^2) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta\bar{x}) \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - (n\bar{x}\bar{y} - \beta n\bar{x}^2) - \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} + \beta n\bar{x}^2 - \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y} + \beta\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = 0$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{3.4}$$

Note that we can further express (3.4) as

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

3.2.2 Variances of the Parameters of a Simple Linear Regression

We state without proof that the variances are:

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \text{ and } \mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{3.5}$$

hence, we summarize that

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \tag{3.6}$$

3.2.3 Estimations of σ^2

The variance (σ^2) can be estimated using any of the following formulars:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{Y}_i)^2}{n-2},$$

Or

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta} \sum_{i=1}^{n} x_i y_i}{n-2}$$

Or

$$\begin{aligned}\hat{\sigma}^2 &= \text{MSE} \\ &= \frac{SSE}{n-2} \\ &= \frac{SS_{yy} - \hat{\beta} SS_{xy}}{n-2}\end{aligned}$$

It should be noted that the division by $n-2$ was because only two parameters β_0 and β were estimated.

3.2.4 Data Analysis Using a Simple Linear Regression Model

Using a simple linear regression model, we intend to predict the response variable *Y*, given one independent variable *X*. We are interested in knowing how well the model has performed, and this, inevitably, leads to the calculation of root mean square error or the R-squared. These indices are useful in assessing the performances of a regression model.

3.2.5 The Root Mean Square Error (RMSE)

The RMSE, otherwise called the prediction error, presents a standard way to measure the error of a model in predicting quantitative data. It is given as:

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

where $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$ are the predicted values, whereas $y_1, y_2, \cdots, y_n$ are the observed values and n is the number of observations. To understand the RMSE better, we take the expected value of the error squared and see where it takes us to as in the following illustration:

$$\begin{aligned}E\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right] &= E\left(\sum_{i=1}^{n} \epsilon_i^2\right) \\ &= \sum_{i=1}^{n} E(\epsilon_i^2) \\ &= Var(\epsilon) + E(\epsilon)^2 \\ &= \sigma^2 + \mu^2.\end{aligned}$$

On how we arrived that

$$\sum_{i=1}^{n} E(\epsilon_i^2) = Var(\epsilon) + E(\epsilon)^2,$$

Recall that

$$\begin{aligned}E(y_i - \hat{y}_i)^2 &= Var(y_i - \hat{y}_i) + E(y_i - \hat{y}_i)^2 \\ &= \text{Var}(\epsilon) + \text{E}(\epsilon)^2 \\ &= \sigma^2 + \mu^2.\end{aligned}$$

Thus far,

$$E\left(\sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}\right) = \sqrt{\frac{\sigma^2 + \mu^2}{n}},$$

But in linear regression, we assume that $\epsilon \sim NID\ (0, \sigma^2)$. Therefore,

$$E\left(\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}}\right)=\sqrt{\frac{\sigma^2+0}{n}}$$
$$=\sqrt{\frac{\sigma^2}{n}}$$
$$=\frac{\sigma}{\sqrt{n}}.$$

In line with the central limit theorem, as n gets larger, the quantity $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}=\frac{\sum_{i=1}^{n}E(\epsilon)^2}{n}$ converges towards n. It further tells us that while $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}$ is a good estimator for $E\left(\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}\right)=\sigma^2$, the RMSE is a good estimator for the standard deviation (σ) of the error for a given single observation, and not that of the total error. The division by n keeps this measure of the error consistent as one moves from one observation to another. As the sample size increases, the estimate of σ becomes more accurate.

3.2.6 The Coefficient of Determination (R^2)

It is a statistic used to measure in a regression model, the proportion of variability in the dependent variable that is explained by the independent variable. Put differently, the coefficient of variation tells how well the data fits the model. Some researchers have argued that although the coefficient of determination provides some useful information regarding the regression model, one does not need to rely exclusively on it in assessing the model's performance. This is owed to the fact that it does not disclose information about the causation relationship between the dependent and independent variable. The coefficient of determination can mathematically be given as:

$$R^2=1-\frac{SSR}{SST},$$

where SSR is the sum of square regression and SST is the sum of square total.

Recall that the SST measures the variation in the observed data (data used in regression modelling), whereas the sum of square regression measures how well the regression model represents the data.

Interpretation of Coefficient of Determination

A very common interpretation of the coefficient of determination is how well the data fits the model. For instance, a coefficient of determination of 70% for instance, tells that 70% of the observed data fits the model. In general, a higher coefficient of variation indicates a better fit to the data.

4 Data Analysis/Results

The datasets for this analysis will be simulated using the R statistical package. The datasets will address the following cases:

- Linearity or otherwise in datasets.
- Datasets with failed normality assumption.
- Datasets with outliers.
- Datasets with correlated observations.

4.1 Linearity or Otherwise in Dataset

The datasets that will help us to understand the performances of a linear model when datasets are linear or otherwise, have been simulated using the R statistical software. The simulated datasets are contained in the Appendices as LinearDat and NonlinearDat respectively for linear and nonlinear datasets. Figures 4.1 and 4.2 present graphical display of the two datasets.

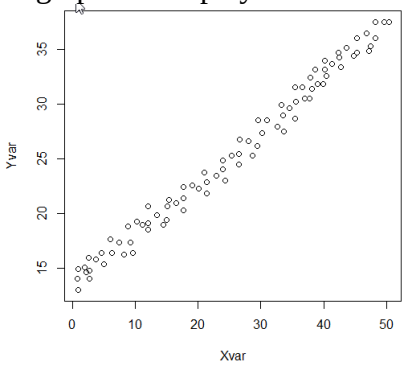

Figure 4.1 Linear separable dataset.

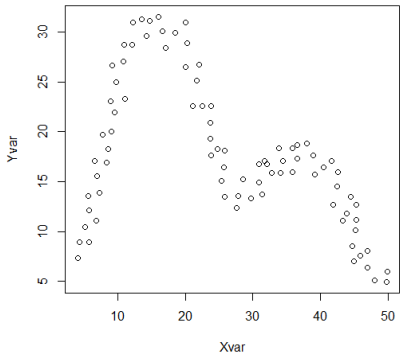

Figure 4.2 Nonlinear dataset.

4.2 Dataset with Failed Normality Assumption

This dataset was generated from the gamma distribution using the R statistical software. The R codes that were used to generate the dataset is contained in the Appendices. Figure 4.3 contains a graphical display of the dataset.

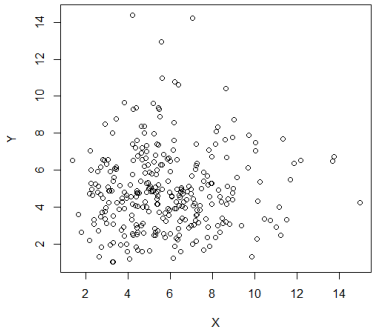

Figure 4.3 An instance of a dataset that does not follow a normal distribution.

4.3 Dataset with Outliers

The dataset with outliers were equally generated using the R statistical software. The full dataset is contained in the Appendices and Figures 4.4 and 4.5 contain datasets with outliers and another without outliers respectively.

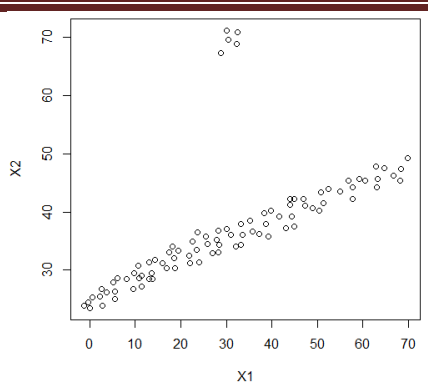

Figure 4.4 Dataset with five points outliers.

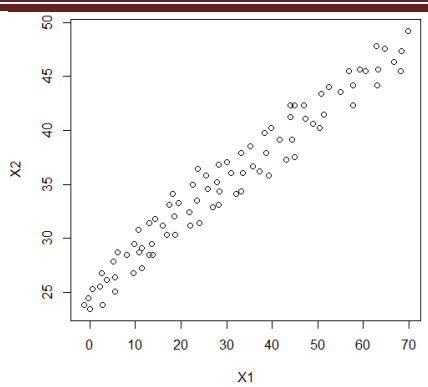

Figure 4.5 Dataset with removed five points outliers.

4.4 Correlated Dataset

The dataset that will be used to assess the performance of a linear model on correlated dataset will be drawn from TH.data package in R (Hothorn & Hothorn, 2019). The name of the dataset is bodyfat and we shall show interest in age and anthro4 because the two datasets are correlated. Here, age is the independent variable.

4.5 Analysis of Data

S/No	Dataset Involved	RMSE	R^2
1.	Linear separable dataset	0.8853	0.9847
2.	Non-linear dataset	6.1730	0.2396
3.	Dataset with failed normality assumption	2.1685	0.0018
4.	Dataset with five points outliers	7.9349	0.3971
5.	Dataset with removed five points outliers	1.4593	0.9536
6.	Correlated dataset	0.5941	0.1187

Table 4.1: A table showing the RMSE and R^2 of the various dataset used in the analysis.

Displayed on Table 4.1 are the RMSE and the coefficient of determination (R^2) for all the dataset used in the analysis. We are interested in the dataset with the smallest RMSE and very high coefficient of determination at the same time. The reason is owed to the fact that the regression model (see (1.2)) is deemed to have performed optimally when the RMSE is very small, with very high coefficient of determination. So far, only the linear dataset has very small RMSE and high coefficient of determination. As contained in the Table, the dataset has 0.8853 as RMSE and 0.9847 as value for R^2. The regression model performed badly given the non-linear dataset because here, we recorded a RMSE of 6.1730 and R^2 value of 0.2396. This is followed by dataset with five-points outliers with a high RMSE (7.9349) and small R^2(0.3971). The regression model performed badly on the dataset with failed normality assumption, because the value of RMSE obtained is 2.1685 and extremely small value for R^2 (0.0018).

The correlated dataset has a very small RMSE (0.5941), the smallest of all and a small value for R^2 (0.1187). This shows that the model performed very poorly given the dataset. Nonetheless, it is pertinent to observe that if we had assessed the performance of the regression model based on the RMSE a lone, we would have concluded that the model performed very well given the dataset. Now, by considering the

R^2 as additional tool for assessment, it becomes obvious that correlation in a dataset inhibits an optimal performance of a linear regression model.

5 Conclusions/Recommendations

Thus far, the performances of the linear regression have been examined in the light of different datasets namely; Linear Dataset, Non-Linear Dataset, Dataset with Failed Normality Assumption, Dataset with Five Points Outliers, Dataset with Removed Five Points Outliers and Correlated Datasets. It has been observed that on linear datasets, the regression model performed optimally, whereas this is not true with other types of dataset.

In particular, the model performed badly on non-linear dataset, followed by dataset with five points outliers. There is another bad performance of the model on dataset with failed normality assumption. One important observation here is that on correlated datasets, the model performed very well when assessed based on RMSE. The inclusion of coefficient of determination as an assessment tool helped to clarify that a regression model performed badly when datasets are correlated. This observation has underlined the fact that the assessment of the regression model on the basis of one assessment tool may not always give the desired result.

5.1 Suggestions for Further Studies

It is being suggested that further studies can be carried out to discover if any transformation procedure can be carried out on the other datasets, to improve the performance of the linear regression model on them. Where transformation could not produce the desired result, the study will explore other ways of manipulating the datasets to ensure optimal performance of the regression model on them.

References

Berry, W. D. (1993). *Sage university papers series. Quantitative applications in the social sciences, Vol. 92. Understanding regression assumptions.* Sage Publications, Inc.

Hothorn, T., & Hothorn, M. T. (2019). *Package 'TH.data'*, R package version 1.0-10. https://CRAN.R-project.org/package=TH.data.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*, 51–62.

Obi, J. C. (2020). *A Foundation Course In Statistics With Applications in R*. Favour Fountains Concepts. pp299.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

Appendices

B. R Programmes Used for Data Analysis

B1. Linear Separable Dataset

```
> rm(list=ls())
```

```
> dat = read.table(file = "clipboard", header = T)
> mod = lm(Yvar ~ Xvar, data = dat)
> summary(mod)

Call:
lm(formula = Yvar ~ Xvar, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91313 -0.57263 -0.01024  0.71985  1.52737

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.343719   0.191381   69.72   <2e-16 ***
Xvar         0.477164   0.006519   73.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8959 on 83 degrees of freedom
Multiple R-squared:  0.9847,     Adjusted R-squared:  0.9846
F-statistic:  5358 on 1 and 83 DF,  p-value: < 2.2e-16

> anova(mod)
Analysis of Variance Table

Response: Yvar
          Df Sum Sq Mean Sq F value    Pr(>F)
Xvar       1 4300.6  4300.6  5357.6 < 2.2e-16 ***
Residuals 83   66.6     0.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> RMSE = sqrt(sum(mod$residuals^2)/nrow(dat)); RMSE
[1] 0.8853337
```

B2. Non-linear Dataset

```
> rm(list=ls())
> dat = read.table(file = "clipboard", header = T)
> mod = lm(Yvar ~ Xvar, data = dat)
> summary(mod)

Call:
lm(formula = Yvar ~ Xvar, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-15.8319  -3.7728   0.2459   3.5997  11.8070

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.14454    1.41881  17.017   <2e-16 ***
Xvar        -0.24759    0.04842  -5.113    2e-06 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.247 on 83 degrees of freedom
Multiple R-squared:  0.2396,     Adjusted R-squared:  0.2304
F-statistic: 26.15 on 1 and 83 DF,  p-value: 1.999e-06

> anova(mod)
Analysis of Variance Table

Response: Yvar
          Df Sum Sq Mean Sq F value    Pr(>F)
Xvar       1 1020.3 1020.32  26.146 1.999e-06 ***
Residuals 83 3239.0   39.02
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> RMSE = sqrt(sum(mod$residuals^2)/nrow(dat)); RMSE
[1] 6.172989
```

B3. Dataset with Failed Normality Assumptions

```
> rm(list=ls())
> dat = read.table(file = "clipboard", header = T)
> mod = lm(Yvar ~ Xvar, data = dat)
> summary(mod)

Call:
lm(formula = Yvar ~ Xvar, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7842 -1.6277 -0.1216  1.2730  9.5700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.67514    0.33617  13.907   <2e-16 ***
Xvar         0.03889    0.05377   0.723     0.47
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.176 on 298 degrees of freedom
Multiple R-squared:  0.001752,   Adjusted R-squared:  -0.001598
F-statistic: 0.523 on 1 and 298 DF,  p-value: 0.4701

> anova(mod)
Analysis of Variance Table

Response: Yvar
           Df  Sum Sq Mean Sq F value Pr(>F)
Xvar        1    2.48  2.4758   0.523 0.4701
Residuals 298 1410.77  4.7341
> RMSE = sqrt(sum(mod$residuals^2)/nrow(dat)); RMSE
```

```
[1] 2.168537
```

B4. Dataset With Five Points Outliers

```
> rm(list=ls())
> dat = read.table(file = "clipboard", header = T)
> mod = lm(Yvar ~ Xvar, data = dat)
> summary(mod)

Call:
lm(formula = Yvar ~ Xvar, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-4.587 -2.985 -1.845 -0.381 34.081

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.23032    1.58148  17.218  < 2e-16 ***
Xvar         0.33095    0.04347   7.613 2.84e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.025 on 88 degrees of freedom
Multiple R-squared:  0.3971,     Adjusted R-squared:  0.3902
F-statistic: 57.96 on 1 and 88 DF,  p-value: 2.837e-11

> anova(mod)
Analysis of Variance Table

Response: Yvar
          Df Sum Sq Mean Sq F value    Pr(>F)
Xvar       1 3732.3  3732.3   57.96 2.837e-11 ***
Residuals 88 5666.7    64.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> RMSE = sqrt(sum(mod$residuals^2)/nrow(dat)); RMSE
[1] 7.93491
```

B5. Dataset With Removed Five Points Outliers

```
> rm(list=ls())
> dat = read.table(file = "clipboard", header = T)
> mod = lm(Yvar ~ Xvar, data = dat)
> summary(mod)

Call:
lm(formula = Yvar ~ Xvar, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6887 -1.1514 -0.0933  1.3261  3.2984
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.350975   0.293482   86.38   <2e-16 ***
Xvar         0.330534   0.008001   41.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.477 on 83 degrees of freedom
Multiple R-squared:  0.9536,     Adjusted R-squared:  0.9531
F-statistic:  1707 on 1 and 83 DF,  p-value: < 2.2e-16

> anova(mod)
Analysis of Variance Table

Response: Yvar
          Df Sum Sq Mean Sq F value    Pr(>F)
Xvar       1 3721.9  3721.9  1706.6 < 2.2e-16 ***
Residuals 83  181.0     2.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> RMSE = sqrt(sum(mod$residuals^2)/nrow(dat)); RMSE
[1] 1.459313
```

B6. Correlated Dataset

```
> rm(list=ls())
> dat = read.table(file = "clipboard", header = T)
> mod = lm(anthro4 ~ age, data = dat)
> summary(mod)

Call:
lm(formula = anthro4 ~ age, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-2.25318 -0.30863  0.08458  0.41974  1.09384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.563045   0.283091  16.119  < 2e-16 ***
age         0.016418   0.005386   3.048  0.00326 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6027 on 69 degrees of freedom
Multiple R-squared:  0.1187,     Adjusted R-squared:  0.1059
F-statistic: 9.293 on 1 and 69 DF,  p-value: 0.00326

> anova(mod)
Analysis of Variance Table
```

```
Response: anthro4
          Df  Sum Sq Mean Sq F value  Pr(>F)
age        1  3.3753  3.3753  9.2929 0.00326 **
Residuals 69 25.0616  0.3632
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> RMSE = sqrt(sum(mod$residuals^2)/nrow(dat)); RMSE
[1] 0.5941219
```