Model Diagnostics and Remedial Measures in SLR

1 Introduction

The regression model is based on a number of assumptions. Those assumptions are made so that we can apply commonly used probability distributions to we quantify the variability associated with our estimated regression model. This means that if the assumptions are not met for our regression model, then how we quantify the variability associated with our model is no longer reliable. All our analysis with statistical inference becomes questionable.

In this module, you will learn how to assess whether the regression assumptions are met. We will explore ways in which we can transform our variables after diagnosing which assumptions are not met so that we can still proceed to build our regression model.

2 Assumptions in Linear Regression

In module 3, we stated the SLR model as

$$y = f(x) + \epsilon \tag{1}$$

where $f(x) = \beta_0 + \beta_1 x$. We need to make some assumptions for the error term ϵ. Mathematically, the assumptions are expressed as

$$\epsilon_1, \ldots, \epsilon_n \; i.i.d. \sim N(0, \sigma^2) \tag{2}$$

Breaking down (2) the assumptions can be expressed as the following:

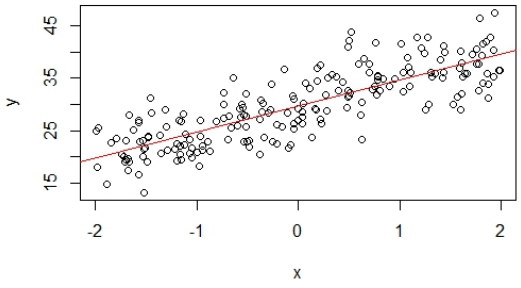
1. The errors have **mean 0**.
2. The errors have **constant variance denoted by** σ^2.
3. The errors are **independent**.
4. The errors are **normally distributed**.

Let's dig a little deeper into the meaning and implications of these 4 assumptions.

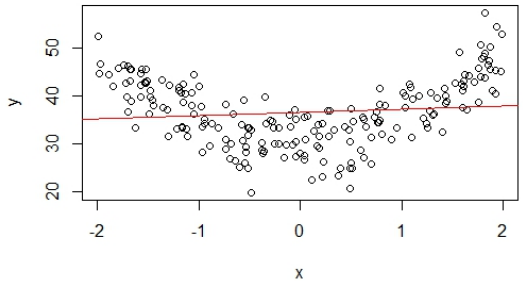
2.1 Assumption 1: Errors have mean 0.

For each value of the predictor, the errors have **mean 0**. A by-product of this statement is that the relationship between y and x, as expressed via $y \approx f(x)$, is correct. So, if $f(x) = \beta_0 + \beta_1 x$, then the relationship is approximately linear.

The plots in Figure 1 are based on simulated data. The scatterplot shown in Figure 1a is an example of when this assumption is met. As we move from left to right on the plot, the data points are generally evenly scattered on both sides of the regression line that is overlaid.



(a) Plot with Linear Relationship



(b) Plot with Non Linear Relationship

Figure 1: Assumption 1

The scatterplot shown in Figure 1b is an example of when this assumption is **not** met. As we move from left to right on the plot in Figure 1b, the data points are generally not evenly scattered on both sides of the regression line that is overlaid.

- When $-2 \leq x \leq -1.2$, the data points are generally above the regression line;
- then when $-1.2 < x < 1$, the data points are generally below the regression line;
- and then when $x \geq 1$, the data points are generally above the regression line.

2.1.1 Consequences of violating this assumption

Predictions will be biased. This means that predicted values will systematically over- or under- estimate the true values of the response variable. Of the 4 assumptions listed, this is **most crucial assumption**.

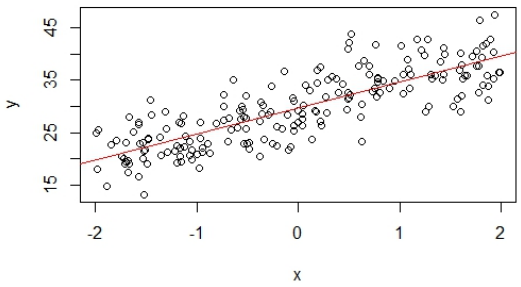
Using Figure 1b as an example, this implies that

- when $-2 \leq x \leq -1.2$, the regression line will systematically under-predict the response variable;
- then when $-1.2 < x < 1$, the regression line will systematically over-predict the response variable;
- and then when $x \geq 1$, the regression line will systematically under-predict the response variable.

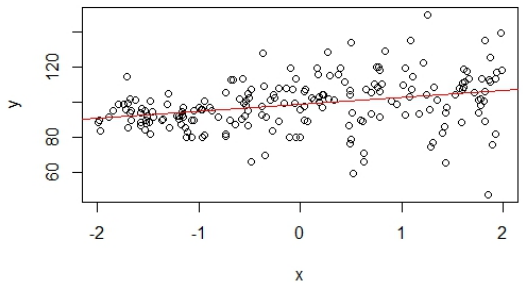
2.2 Assumption 2: Errors have constant variance

For each value of the predictor, the error terms have **constant variance**, denoted by σ^2. This implies that when looking at a scatterplot, the vertical variation of data points around the regression equation has the same magnitude everywhere.

The plots in Figure 2 are based on simulated data. The scatterplot shown in Figure 2a is an example of when this assumption is met (this figure is actually the same as Figure 1a, so the data that produced these plots satisfy both assumptions). As we move from left to right on the plot, the vertical variation of the data points about the regression line is approximately constant.



(a) Plot with Constant Variance



(b) Plot with Increasing Variance

Figure 2: Assumption 2

The scatterplot shown in Figure 2b is an example of when this assumption is **not** met. As we move from left to right on the plot in Figure 2b, the vertical variation of the data points about the regression line becomes larger as the value of the response variable gets larger, so the variance is not constant.

2.2.1 Consequences of violating this assumption

Statistical inference will no longer be reliable. This means that the results from any hypothesis test, confidence interval, or prediction interval are no longer reliable.

Interestingly, for the scatterplot in Figure 2b, we can say that assumption 1 is met, since the the data points are generally evenly scattered on both sides of the regression line. Predictions will still be unbiased; the predicted response, $\hat{y}$, do not systematically over- or under-predict the response variable. So if our goal is to assess if the relationship is approximately linear, this scatterplot is fine. We do lose the utility from hypothesis tests, CIs, and PIs.

2.3 Assumption 3: Errors are independent

A by-product of this assumption is that the values of the response variable, y_i, are independent from each other. Any y_i does not depend on other values of the response variable.

2.3.1 Consequences of violating this assumption

Statistical inference will no longer be reliable. This means that the results from any hypothesis test, confidence interval, or prediction interval are no longer reliable.

2.4 Assumption 4: Errors are normally distributed

If we were to create a density plot of the errors, the errors should follow a normal distribution.

2.4.1 Consequences of violating this assumption

The regression model is fairly robust to the assumption that the errors are normally distributed. In other words, violation of this particular assumption is not very consequential. **Of the 4 assumptions, this is the least crucial to satisfy.**

3 Assessing Regression Assumptions

There are a few visualizations that help in detecting violations of the regression assumptions. These visualizations are:

- Scatterplot of y against x (assumptions 1 and 2).
- Residual plot (assumptions 1 and 2).
- Autocorrelation function (ACF) plot of residuals (assumption 3).
- Normal probability plot of residuals (often called QQ plot) (assumption 4).

3.1 Scatterplot

We can examine the scatterplot of y against x to check for assumptions 1 and 2. We want to see the following in the scatterplot:

- **No nonlinear pattern** (assumption 1).
- Data points **evenly scattered** (for each value on the x-axis) around fitted line (assumption 1).
- Vertical variation of data points constant (assumption 2).

We have used Figure 2a as an example of a scatterplot that meets these assumptions. Let us take a look at another example that we have worked with. This scatterplot is from the `elmhurst` dataset from the `openintro` package that we have been seeing in tutorials. We are regressing the amount of gift aid a student receives based on the student's family income. The corresponding scatterplot is shown in in Figure 3.

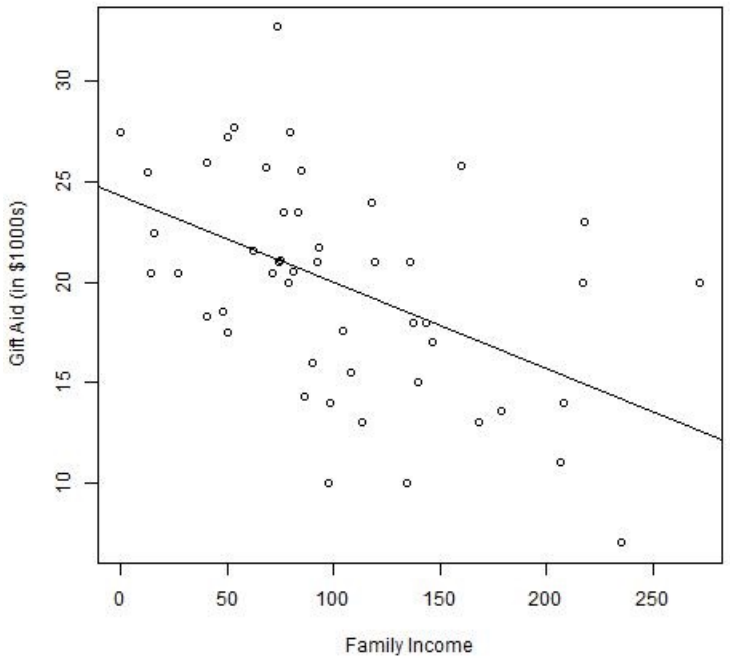


Figure 3: Scatterplot of Gift Aid Against Family Income

In Figure 3, we see that the data points are evenly scattered around the fitted line. We also see the vertical variation of the data points is fairly constant. So assumptions that the errors have 0 mean and constant variance appear to be met.

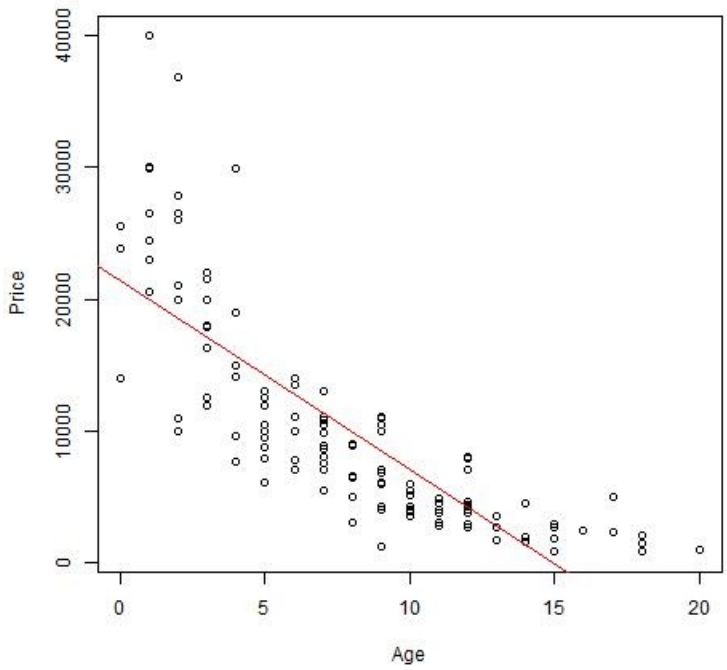


Figure 4: Scatterplot of Sale Price Against Age

3.1.1 Practice question

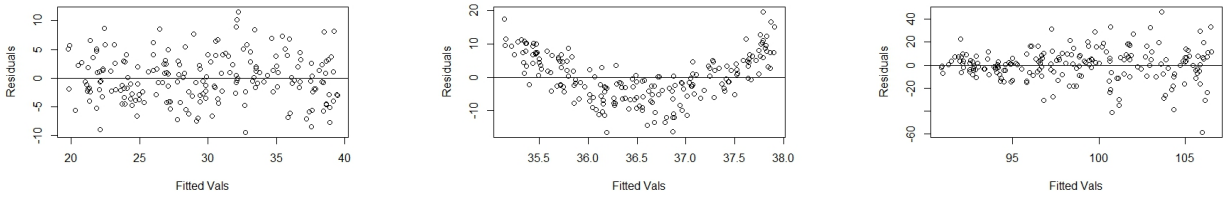
The data are about the prices of used cars. We are regressing the sale price of the car against the age of the car. The corresponding scatterplot is shown in Figure 4 Based on Figure 4, which of assumptions 1 or 2 (or both, or neither), is met? We will go over this in the tutorial.

3.2 Residual plot

While using the scatterplot is an intuitive way of assessing regression assumptions, it has a limitation. It cannot be used if we have multiple predictors in our regression, which we will encounter (and happens more often than just having one predictor). Another visualization that we can use to assess assumptions 1 and 2 is a **residual plot**. This is a scatterplot of residuals, e, against fitted values, $\hat{y}$. We want to observe the following in a residual plot.

- Residuals should be **evenly scattered** across the horizontal axis (assumption 1).
- The residuals should have **similar vertical variation** across the plot (assumption 2).
- Some writers combine these two points into the following statement: the residuals should fall in a **horizontal band around 0** with no apparent pattern (assumption 1, 2).

The residual plots in Figure 5 are based on simulated data from Figures 1a, 1b, and 2b.



(a) Assumptions 1 & 2 Met (b) Assumption 1 Not Met (c) Assumption 2 Not Met

Figure 5: Residual Plots from Fig 1a, 1b, 2b Respectively

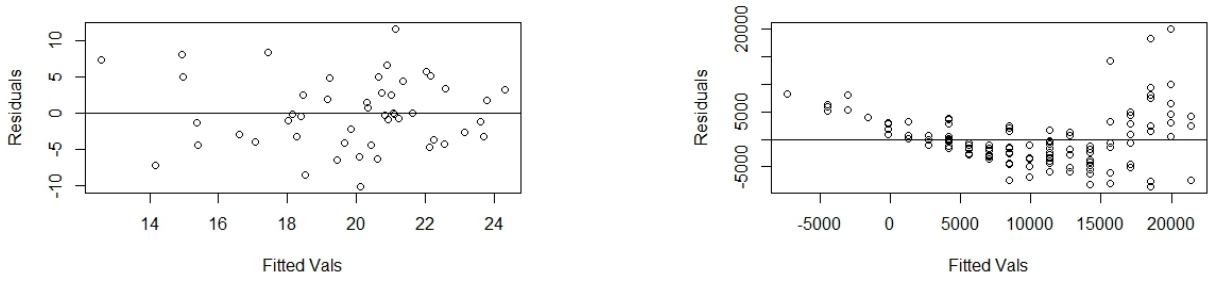
We make the following observations:

- From Figure 5a, we see that the residuals are evenly scattered across the horizontal axis, and their vertical variation is fairly constant across the plot. So both assumptions are met.
- From Figure 5b, we see that the residuals are **not** evenly scattered across the horizontal axis, although their vertical variation is fairly constant across the plot. So only assumption 1 is not met.
- From Figure 5c, we see that the residuals are evenly scattered across the horizontal axis, but their vertical variation is **not constant** across the plot. In fact, the vertical variation is increasing as we move from left to right. So only assumption 2 is not met.

If you compare the conclusions from the residuals plots and scatterplots, they are the same. In SLR, the takeaways should be consistent.

3.2.1 Practice questions

1. The residual plot in Figure 6a comes from regressing gift aid against family income for the `elmhurst` dataset. Based on this residual plot, which assumptions are met?
2. The residual plot in Figure 6b comes from regressing price of cars against age for the used cars dataset. Based on this residual plot, which assumptions are met?



(a) Residual Plot Regressing Gift Aid Against Family Income

(b) Residual Plot Regressing Price Against Age for Used Cars

Figure 6: Residual Plots for Practice Questions

3.3 ACF plot

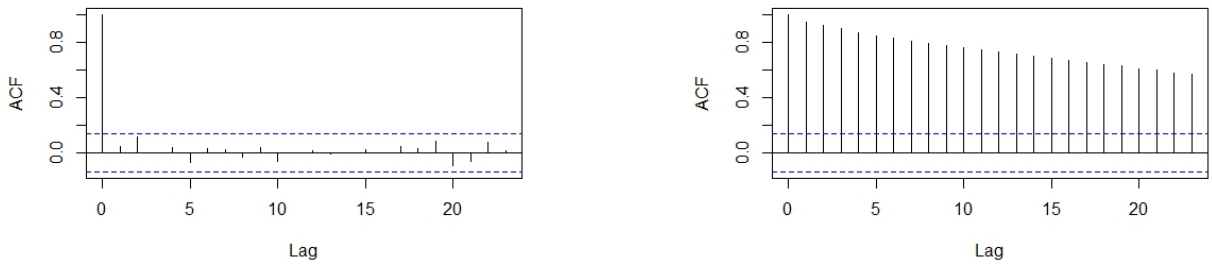
Assumption 3 states that the errors are **independent**. This assumption implies that the values of the response variable are independent from each other. This assumption is typically assessed via knowing the nature of the data.

- If the observations were obtained from a random sample, it is likely that the observations will be independent from each other. This is the very nature of a random sample and why random samples are preferred over convenience samples.
- If the data has some inherent sequence, it is likely the observations will not be independent, and are dependent. For example, if I record the value of a stock at the end of each day, the value at day 2 is likely to be related to its value at day 1. So the values of stock prices at the end of each day are not independent.

An autocorrelation function (ACF) plot of the residuals may be a used to help assess if the assumption that the errors are independent is met. However, the plot is not a substitute for using your understanding about the nature of the data and should only be used as a confirmation.

The ACF plot measures the correlation between a vector of observations and the lagged versions of the observations. If the observations are uncorrelated, the correlations between the vector of observations and lagged versions of these observations are theoretically 0. We may create an ACF plot for the residuals from our regression.

The ACF plot in Figure 7a and is based on simulated data that were independently generated.



(a) ACF of Residuals that are Uncorrelated

(b) ACF of Residuals when Response was Ordered

Figure 7: Assumption 3

A few notes about the ACF plot:

- The ACF at lag 0 is always 1. The correlation of any vector with itself is always 1.
- The dashed horizontal lines represent critical values. An ACF at any lag beyond the critical value indicates an ACF that is significant. We have evidence of correlation (and hence dependence) in our residuals.
- If the observed values for the response variable are independent, then we would expect the ACFs at lags greater than 0 to be insignificant. Do note that because we are conducting multiple hypothesis tests, do not be too alarmed if the ACFs are slightly beyond the critical values at an isolated lag or 2.

Based on Figure 7a, we see that the ACFs at all lags greater than 0 are insignificant. We do not have evidence the residuals are correlated with each other, so we do not have evidence that assumption 3 is not met.

Sometimes, the dataframe can be sorted in some manner (e.g. increasing order for response variable), and if so, we would actually expect to see significant correlations in the ACF plot. The ACF plot in Figure

7b is such an example. The residuals are from the same simulated dataset, only with the data sorted by the response variable. If we had just looked at the ACF plot in Figure 7b without understanding the data were simulated independently and then sorted, we would have erroneously concluded that the residuals are not independent and the regression assumption is not met.

3.4 QQ plot

A normal probability plot (also called a QQ plot) is used to assess if the distribution of a variable is normal. It typically plots the residuals against their theoretical residual if they followed a normal distribution. A QQ line is typically overlaid. If the plots fall closely to the QQ line, we have evidence that the observations follow a normal distribution. Figure 8 shows a QQ plot that comes from a normally distributed variable.

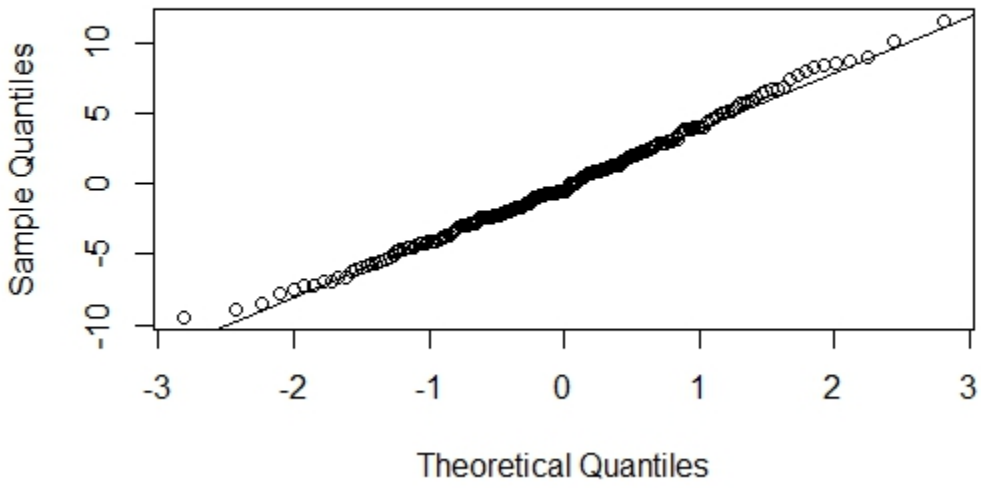


Figure 8: QQ Plot

3.5 Remedial Measures

We now know how to assess if specific regression assumptions are not met. The remedial measures involve transforming either the predictor variable and / or the response variable. These transformations are chosen to handle violations to assumptions 1 and / or 2 respectively. The general strategy on selecting which variable to transform:

- Transforming the response variable, y, affects both assumptions 1 and 2.
  - Visually, we can think of transforming y in terms of stretching or squeezing the scatterplot of y against x vertically. Thus, transforming y affects the shape of the relationship and the vertical spread of the data points.
  - However, the **choice on how we transform y is based on handling assumption 2.**
- Transforming the predictor variable, x affects assumption 1 and does not theoretically affect assumption 2.

  - Visually, we can think of transforming x in terms of stretching or squeezing the scatterplot of y against x horizontally. Thus, transforming x affects the shape of the relationship but not the vertical spread of the data points.
  - Therefore, **transforming x is based on handling assumption 1.**

- If assumption 2 is not met, we transform y to stabilize the variance and make it constant.
- If assumption 1 is not met, we transform x to find the appropriate shape to relate the variables.
- If both assumptions are not met, we transform y first to stabilize the variance. Once assumption 2 is solved, check if assumption 1 is not met. If not met, transform x.

Assumption 1 deals with whether the way we have expressed how y and x are related, through $f(x)$, is appropriate. Assumption 2 deals with the vertical variation of the data points in the scatterplot.

4 Remedial Measures: Variance Stabilizing Transformations

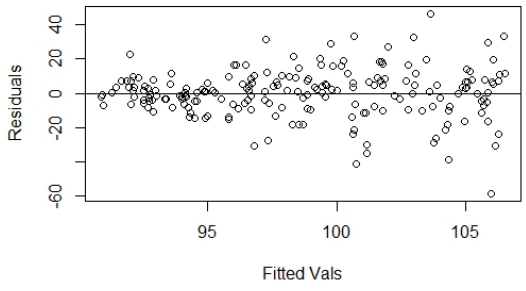
We transform the response variable to stabilize the variance (assumption 2). There are a couple of ways to decide the appropriate transformation:

1. Pattern seen in residual plot can guide choice in how to transform the response variable.
2. Box-Cox plot.

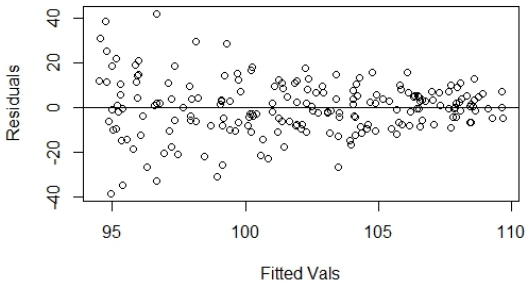
4.1 Use Pattern in Residual Plot

We can stabilize the variance of the errors based on the residual plot, if we see either of the following scenarios:

- vertical variation of residuals **increasing** as fitted response increases, or as we move from left to right, as in Figure 9a, or
- vertical variation of residuals **decreasing** as fitted response increases, or as we move from left to right, as in Figure 9b.



(a) Variance Increasing with Fitted Y



(b) Variance Decreasing with Fitted Y

Figure 9: Non Constant Variance in Residual Plot

Note that increasing variance as fitted response increases is much more common with real data. Generally, larger values of a variable are associated with larger spread.

We transform y using $y^* = y^\lambda$, with λ chosen based on whether the variance of the residuals is increasing or decreasing with fitted response:

- For Figure 9a, choose $\lambda < 1$.
  - If $\lambda = 0$, it means we use a logarithmic transformation with base e, i.e. $y^* = \log(y)$.
  - Note that a logarithm with no base means a natural log, or ln.
- For Figure 9b, choose $\lambda > 1$.

So based on the residual plot, we have a range of values for λ.

4.2 Box-Cox Plot

We can use a Box-Cox plot to help us narrow the range of λ to use. It is a plot of the log-likelihood function against λ, and we choose λ that maximizes this log-likelihood function. For example, Figure 10 shows the Box Cox plot generated for the regression associated with the residual plot in Figure 9b

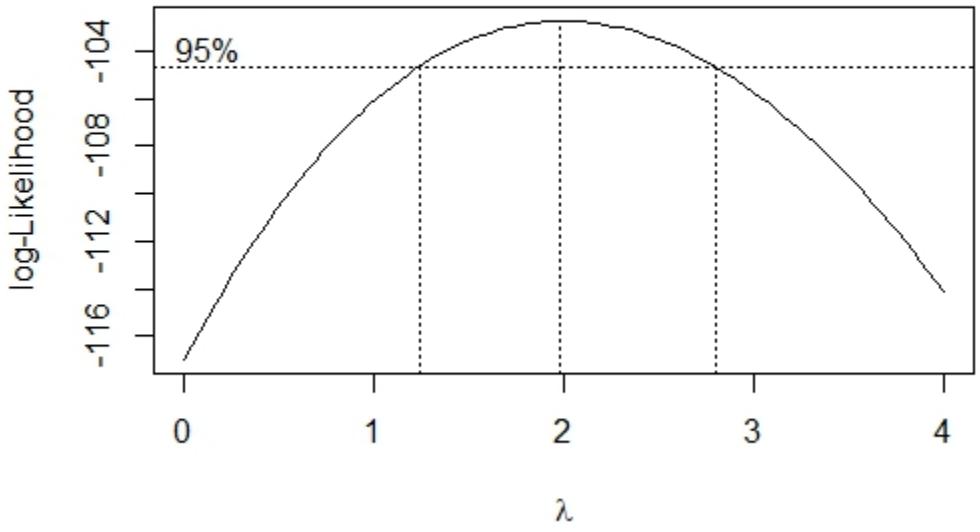


Figure 10: Box-Cox Plot based on Figure 9b

Notice an approximate 95% CI is provided for λ. A few comments on how to use the Box-Cox plot:

- Three vertical dashed lines are displayed: the middle line corresponds to the optimal value of λ; the other two lines are the lower and upper bounds of a 95% CI for λ.
- We choose λ within the CI (or even close to it) that is easy to understand. We do not have to choose the optimal value, especially if its value is difficult to interpret. In this example, I will choose $\lambda = 2$, so a square transformation for y. Transform response with $y^* = y^2$. Regress y^* against x.
- If 1 lies in the CI, **no transformation** on y may be needed.
- If a transformation is needed, a **log transformation** is preferred, since we can still interpret the estimated coefficients. It is difficult to interpret with any other type of transformation.
- View the Box-Cox procedure as a guide for selecting a transformation, rather than being definitive.
- Need to recheck the residuals after every transformation to assess if the transformation worked.

4.3 Interpretation with Log Transformed Response

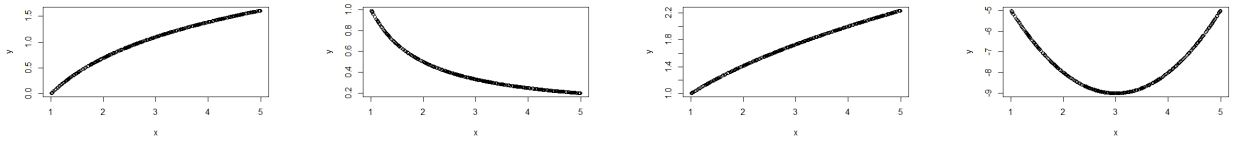
A log transformation on the response is preferred over any other transformation, as we can still interpret regression coefficients. A couple of ways to interpret the estimated slope $\hat{\beta}_1$:

- The predicted response variable is **multiplied by a factor** of $\exp(\hat{\beta}_1)$ for a one-unit increase in the predictor.
- We can also subtract 1 from $\exp(\hat{\beta}_1)$ to express the change as a percentage.
  - If $\hat{\beta}_1$ is positive, we have a percent **increase**. The predicted response variable increases by $(\exp(\hat{\beta}_1) - 1) \times 100$ percent for a one-unit increase in the predictor.
  - If $\hat{\beta}_1$ is negative, we have a percent **decrease**. The predicted response variable decreases by $(1 - \exp(\hat{\beta}_1)) \times 100$ percent for a one-unit increase in the predictor.

5 Remedial Measures: Linearization Transformations

We first ensure the variance has been stabilized and assumption 2 is met. If $f(x)$ does not accurately capture the relationship between the variables, we transform the predictor variable to meet assumption 1. Some writers call this a linearization transformation, as we seek to make the transformed version of the predictor variable, x^*, to be approximately linear with the response variable (or transformed y), i.e. $y = \beta_0 + \beta_1 x^* + \epsilon$. We do not consider transforming the response variable to deal with assumption 1, as transforming the response variable is likely to reintroduce violation of assumption 2.

The general strategy on how to transform the predictor is via a scatterplot of y (or y^*) against x. We use the pattern seen in the plot to decide how to transform the predictor. Some examples are shown in Figure 11 below.



(a) $x^* = \log(x)$ (b) $x^* = \frac{1}{x}$ (c) $x^* = \sqrt{x}$ (d) $x^* = x^2$

Figure 11: Transformations for x

5.1 Hierarchical Principle

One thing to be aware of is the **hierarchical principle**: if the relationship between the response and predictor is of a higher order polynomial (e.g. quadratic, cubic), the hierarchical principle states that the lower order terms should remain in the model. For example, if the relationship is of order h, fit $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_h x^h + \epsilon$ via a multiple linear regression framework. We will see how to do this in the next module.

5.2 Interpretation with Log Transformed Predictor

A log transformation on the predictor is preferred over any other transformation, as we can still interpret the regression coefficient, $\hat{\beta}_1$, in a couple of ways:

- For an a% increase in the predictor, the predicted response **increases by** $\hat{\beta}_1 \log(1 + \frac{a}{100})$**.**
- $\log(1 + \frac{1}{100}) \approx \frac{1}{100}$ (Taylor series). So an alternative interepretation is: for a 1% increase in the predictor, the predicted response increases by approximately $\frac{\hat{\beta}_1}{100}$.

5.3 Interpretation with Log Transformed Response and Predictor

If both response and predictor variables are log transformed, the regression coefficient, $\hat{\beta}_1$, can be interpreted in a couple of ways:

- For an $a\%$ increase in the predictor, the predicted response is **multiplied by** $(1 + \frac{a}{100})^{\hat{\beta}_1}$**.**
- $(1+\frac{1}{100})^{\hat{\beta}_1} \approx 1+\frac{\hat{\beta}_1}{100}$ (Taylor series). So an alternative interpretation is: for a 1% increase in the predictor, the predicted response **increases by approximately** $\hat{\beta}_1$ **percent.** Note that this approximation works better when $\hat{\beta}_1$ is small in magnitude.

5.4 Some General Comments about Assessing Assumptions and Transformations

- When assessing the assumptions with a residual plot, we are assessing if the assumptions are reasonably / approximately met.
- With real data, assumptions are rarely met 100%.
- If unsure, proceed with model building, and test how model performs on new data. If poor performance, go back to residual plot to assess what transformation will be appropriate.
- Assess the plots to decide which variables need to be transformed, and how. The choice of transformation should be guided by what you see in the plots, and not by trial and error.
- A residual plot should always be produced after each transformation. A Box Cox plot could also be produced. The plots should be assessed if the transformation helped in the way you intended.
- Solve assumption 2 first, then assumption 1.