Remedial Measures in SLR Tutorial 1

The linear regression model involves several assumptions. Among them are:

1. The errors, for each fixed value of x, have mean 0. This implies that the relationship as specified in the regression equation is appropriate.
2. The errors, for each fixed value of x, have constant variance. That is, the variation in the errors is theoretically the same regardless of the value of x (or $\hat{y}$).
3. The errors are independent.
4. The errors, for each fixed value of x, follow a normal distribution.

To assess assumptions 1 and 2, we can examine scatterplots of:

- y versus x.
- residuals versus fitted values, $\hat{y}$.

Assumption 3 is assessed based on knowledge of the data. An autocorrelation (ACF) plot of the residuals may also be used.

Assumption 4 is assessed with a normal probability plot, and is considered the least crucial of the assumptions.

We will see how to generate the relevant graphical displays to help us assess whether the assumptions are met, and if needed, carry out transformations on the variable(s) so the assumptions are met.

For this tutorial, we will go over a dataset involving prices of used cars (Mazdas). The two variables are the sales price of the used car, and the age of the car in years. Download the data file, `mazda.txt` and read the data:

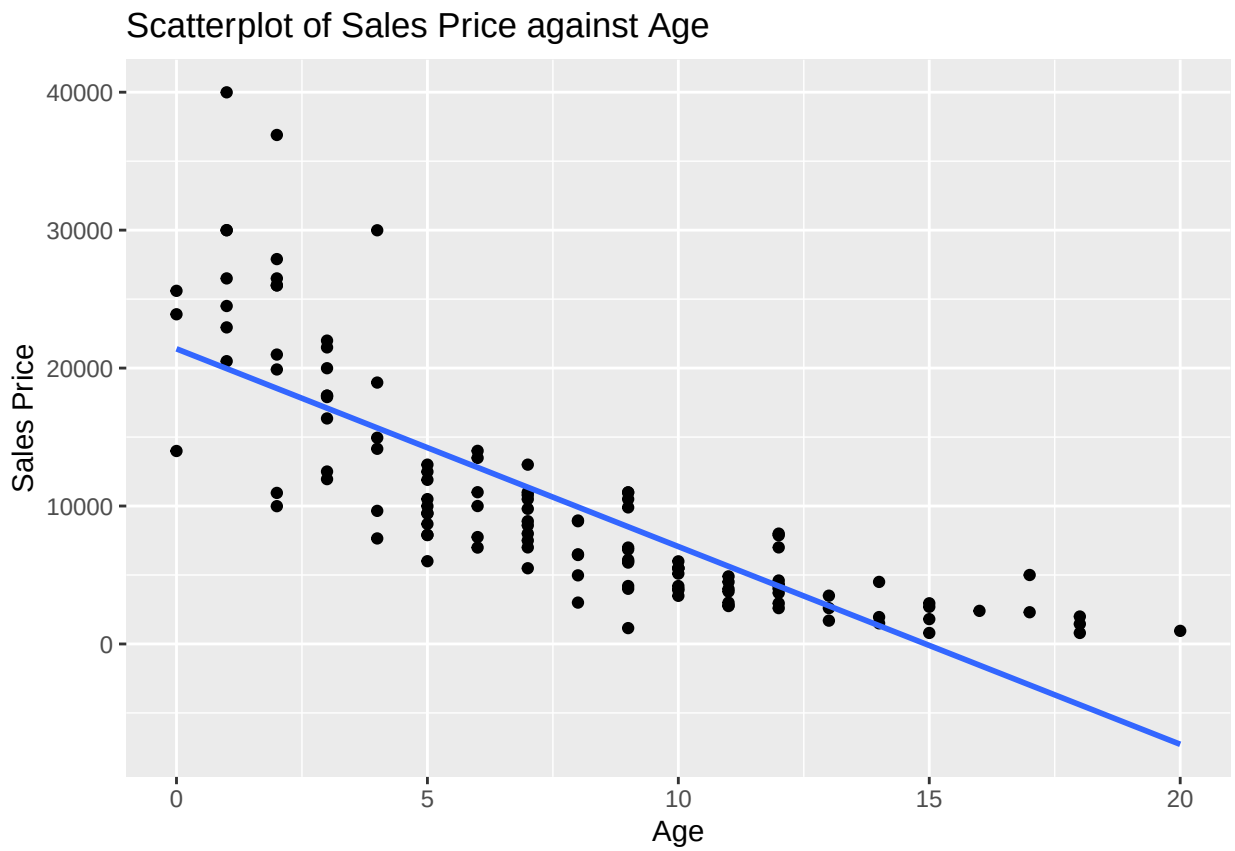
```
Data<-read.table("mazda.txt", header=TRUE, sep="")
```

1. Model Diagnostics with Scatterplots

We can use a scatterplot of the response variable against the predictor to assess assumptions 1 and 2:

```
library(tidyverse)

##scatterplot, and overlay regression line
ggplot2::ggplot(Data, aes(x=Age,y=Price))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE)+
  labs(x="Age", y="Sales Price", title="Scatterplot of Sales Price against Age")
```



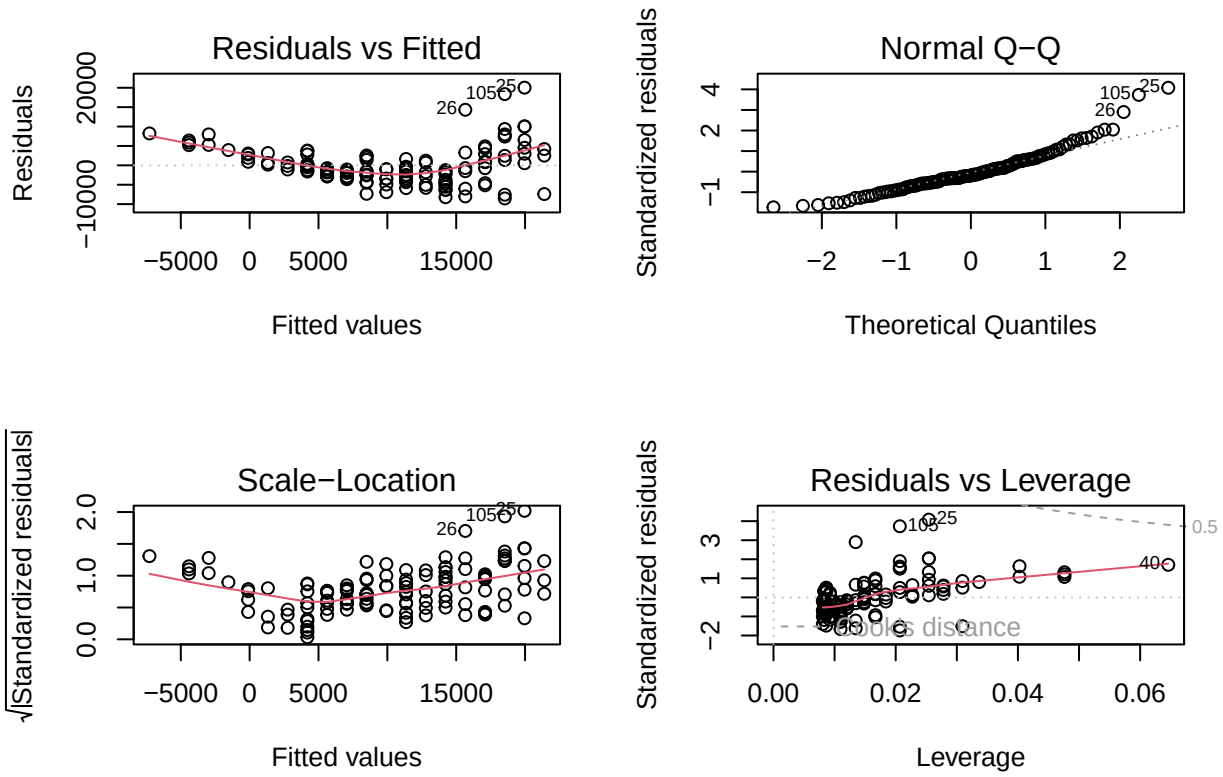
To assess assumption 1, the data points should be evenly scattered on both sides of the regression line, as we move from left to right. We do not see this in the scatterplot, so assumption 1 is not met. When age is between 0 and 2, the data points are mostly above the line. When age is between 5 and 11, the data points are mostly below the line, and when age is greater than 13, the data points are above the line.

To assess assumption 2, the vertical spread of the data points should be constant as we move from left to right. The spread seems to be decreasing as we move from left to right (or in other words, the spread is increasing as the response increases), so assumption 2 is not met.

2. Model Diagnostics with Residual Plots

Sometimes, a residual plot is easier to visualize than a scatterplot. We fit our SLR model using `lm()` as usual. Applying the `plot()` function to an object created with `lm()` actually produces a four diagnostic plots. To display the four diagnostic plots in a 2 by 2 array, we specify `par(mfrow = c(2, 2))` so the plotting window is split into a 2 by 2 array:

```
result<-lm(Price~Age, data=Data)
par(mfrow = c(2, 2))
plot(result)
```



- The first plot (top left) is the residual plot, with residuals on the y-axis and fitted values on the x-axis. The residual plot can be used to address assumptions 1 and 2. A red line is overlayed to represent the average value of the residuals for differing values along the x-axis. This line should be along the x-axis without any apparent curvature to indicate the form of our model is reasonable. This is not what we see, as we see a clear curved pattern. So assumption 1 is not met. For assumption 2, we want to see the vertical spread of the residuals to be fairly constant as we move from left to right. We do not see this in the residual plot; the vertical spread increases as we move from left to right, so assumption 2 is not met.
- The second plot (top right) is the normal probability plot (also called a QQ plot), and addresses assumption 4. If the residuals are normal, the residuals should fall along the 45 degree line. The regression model is fairly robust to this assumption though; the normality assumption is the least crucial of the four.
- The third plot (bottom left) is a plot of the square root of the absolute value of the standardized residuals against the fitted values (scale-location). This plot should be used to assess assumption 2, the constant variance assumption. A red line is overlayed to represent the average value on the vertical axis for differing values along the x-axis. If the variance is constant, the red line should be horizontal and the vertical spread of the plot should be constant. This plot should be used to assess assumption 2, if we have a small sample size. Otherwise, this plot should tell a similar story to the first plot (top left) when assessing assumption 2.
- The last plot (bottom right) is a plot to identify influential outliers. Data points that lie in the contour lines with large Cook's distance are influential. None of our data points have Cook's distance greater than 0.5. As a general rule of thumb, observations with Cook's distance greater than 1 are flagged as influential. We will talk more about influential observations in a future module.

Now that we know that both assumptions 1 and 2 are not met. We need to transform the response variable first, to stabilize the variance.

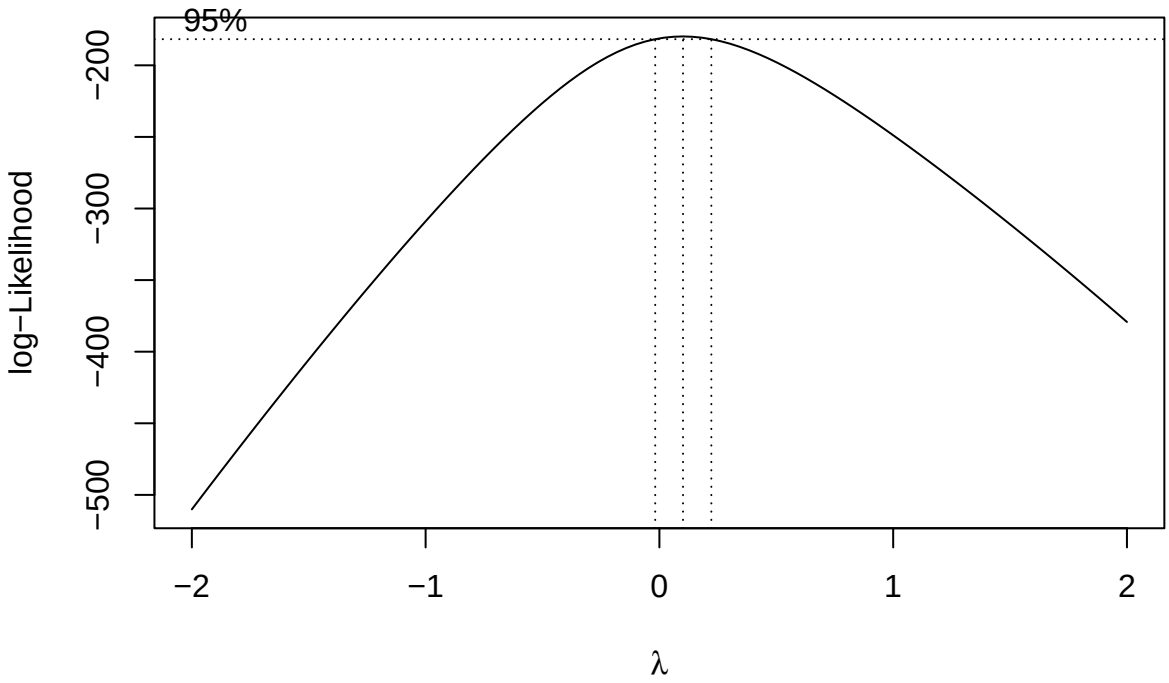
Based on the residual plot, we see that the variance of the residuals increases as we move from left to right. So we know we need to transform the response variable using $y^* = y^\lambda$ with $\lambda < 1$. A log transform should be considered since we can still interpret regression coefficients.

3. Box Cox Transformation on y

The Box Cox plot can be used to decide how to transform the response variable. The transformation takes the form $y^* = y^\lambda$, with the value of λ to be chosen. If $\lambda = 0$, we perform a log transformation.

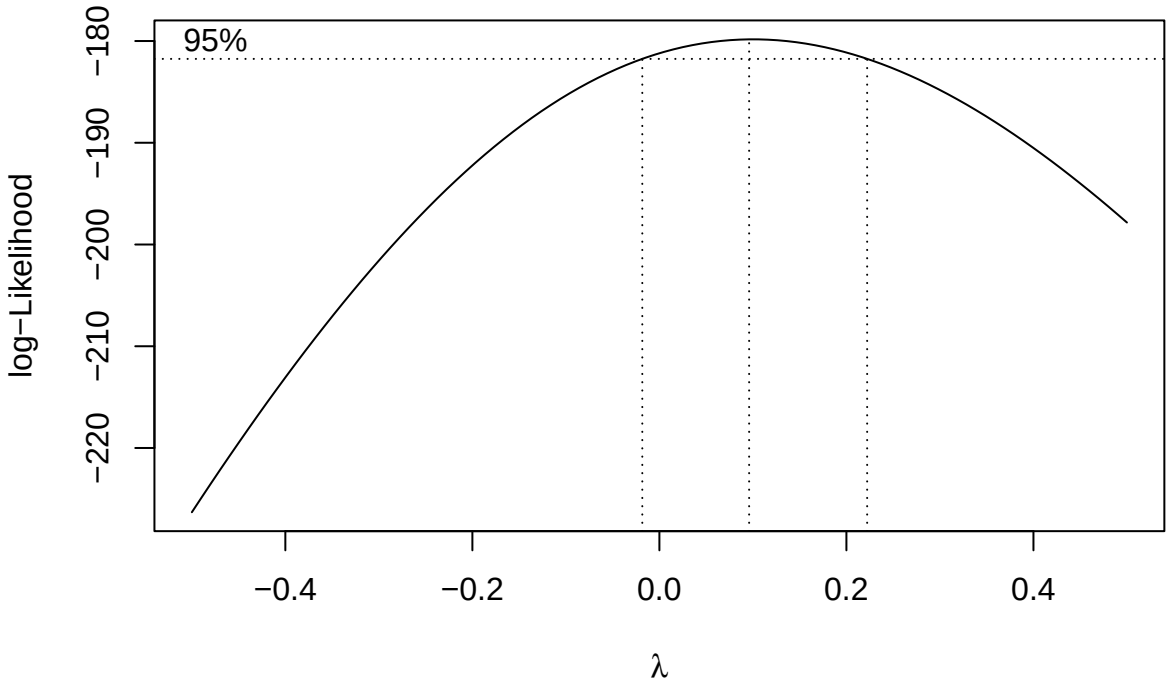
We will use the `boxcox()` function from the `MASS` package:

```
library(MASS) ##to use boxcox function
MASS::boxcox(result)
```



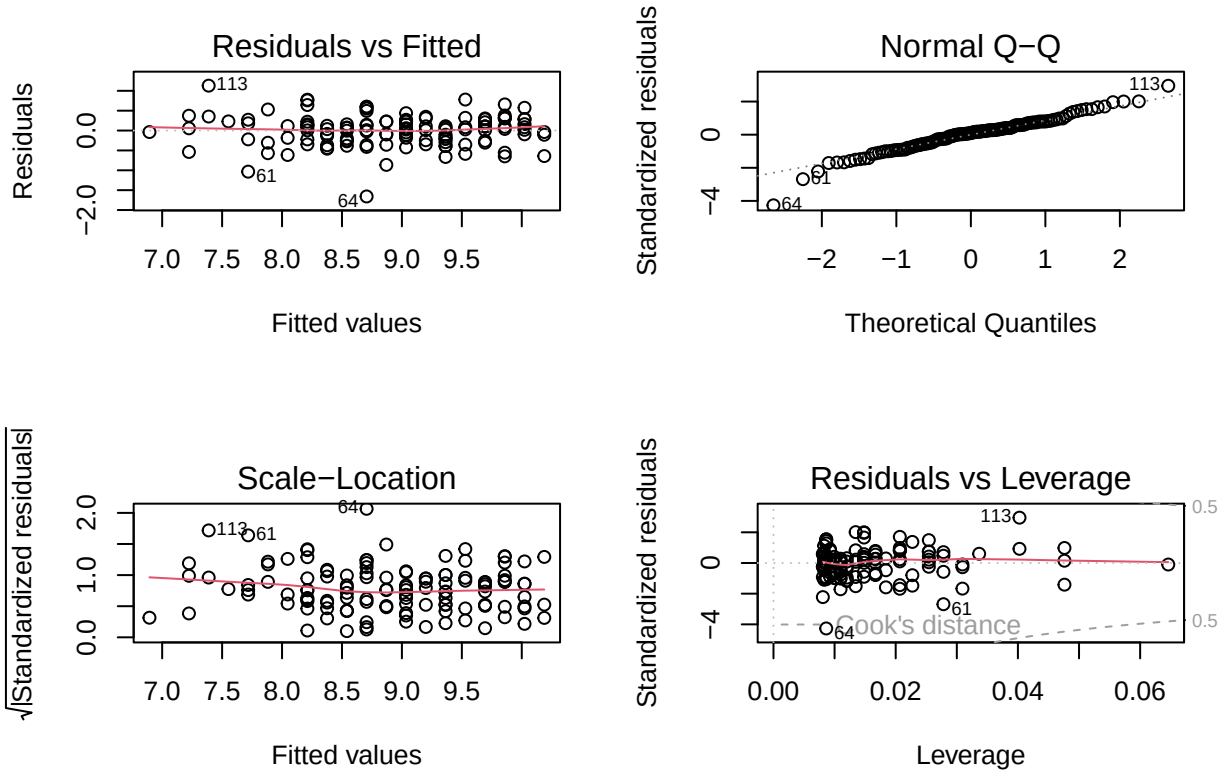
We can "zoom in" on the plot to have a better idea about the value of λ we can use, by specifying the range of `lambda` inside the function:

```
##adjust lambda for better visualization. Choose lambda between -0.5 and 0.5
MASS::boxcox(result, lambda = seq(-0.5, 0.5, 1/10))
```



We can choose any value of λ within the CI. A log transformation is preferred if possible, since we can still interpret coefficients. Since 0 lies in the CI, we choose $\lambda = 0$, to log transform the response variable to get $y^* = \log(y)$. We regress y^* against x, and check the resulting residual plot:

```
##transform y and then regress ystar on x
ystar<-log(Data$Price)
Data<-data.frame(Data,ystar)
result.ystar<-lm(ystar~Age, data=Data)
par(mfrow = c(2, 2))
plot(result.ystar)
```



We need to reassess assumptions 1 and 2 after the transformation.

- For assumption 2, we see that the vertical spread of the residuals in the residual plot (top left) is fairly constant, as we move from left to right. So assumption 2 is met. The log transformation worked.
- We also notice that the residuals are now evenly scattered across the horizontal axis in the residual plot (top left). So assumption 1 is now met.

We do not need to perform any other transformations.

4. Interpreting Coefficients with Log Transformed Response

So our regression equation is

```
result.ystar
```

```
##
## Call:
## lm(formula = ystar ~ Age, data = Data)
##
## Coefficients:
## (Intercept)          Age
##     10.1878      -0.1647
```

$\hat{y^*} = 10.1878 - 0.1647x$, where $y^* = \log(y)$. To interpret the slope:

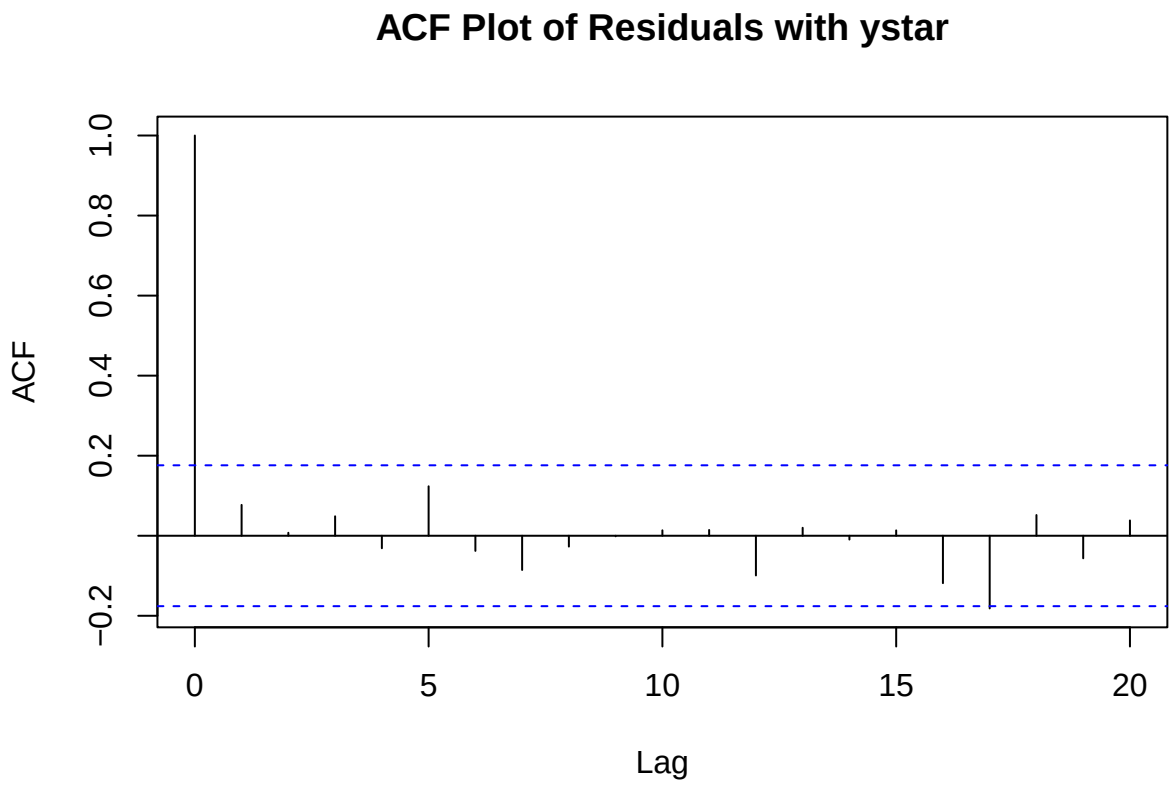
- The price of used Mazdas is multiplied by $\exp(-0.1647) = 0.8481481$ for each year older the car is.
- The price of used Mazdas decreases by $(1 - 0.8481481) \times 100$ percent, or 15.18519 percent, for each year older the car is.

5. ACF Plot of Residuals

We have yet to assess the assumption that the observed prices are independent from each other. Assuming that these prices are from different cars and not the same car measured repeatedly over time, there is no reason to think the prices are dependent on each other.

We an also produce an ACF plot to confirm our thought:

```
acf(result.ystar$residuals, main="ACF Plot of Residuals with ystar")
```



None of the ACFs beyond lag 0 are significant, so we don't have evidence that the observations are dependent on each other.