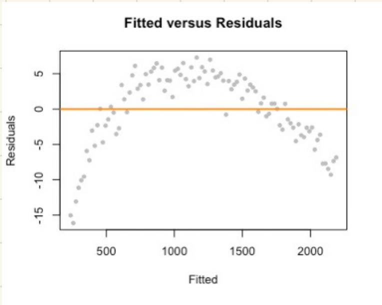
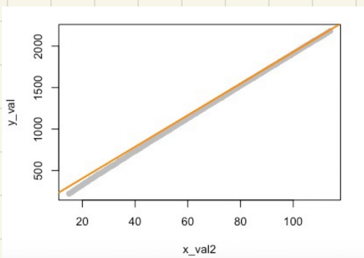
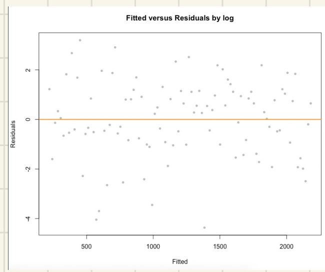
#1 Base on x_val and x_val2 extrme similar, I started using log transformation for this. And R code attached sepreatly.



This residual plot suggest that the SLR model is not consistant with this dataset.

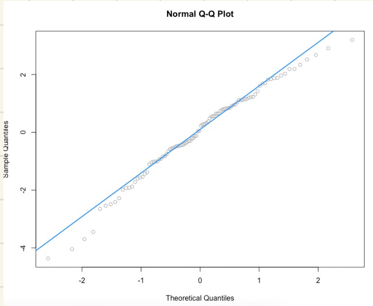




if we try the $Y \sim X_1 + \log(X_2)$ the plot will looks pretty good. and residuals vs fitted valued shows diramatic improvement.

```
        studentized Breusch-Pagan test

data:  Dataset_2_log_fit
BP = 2.0917, df = 2, p-value = 0.3514

>
> shapiro.test(resid(Dataset_2_log_fit))

        Shapiro-Wilk normality test

data:  resid(Dataset_2_log_fit)
W = 0.98138, p-value = 0.17
```

Then by using BP test, we can get p-value 0.3514 which is large enough to imply homoscedasticity. And by using shapiro test, the p-value is 0.17 that imply normality base on this test.



And Q-Q plot corresponding this model is also prety good.

```
> Dataset_2_log_fit = lm(y_val ~ x_val1 + log(x_val2), data = Dataset_1)
> summary(Dataset_2_log_fit)

Call:
lm(formula = y_val ~ x_val1 + log(x_val2), data = Dataset_1)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3677 -0.9193  0.1400  1.1144  3.1974 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.22991    3.61272   6.984 3.61e-10 ***
x_val1      19.02931    0.02202 864.251  < 2e-16 ***
log(x_val2) 37.32225    1.17827  31.675  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.562 on 97 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 6.635e+06 on 2 and 97 DF,  p-value: < 2.2e-16
```

Base we have nice R^2 and $\hat{R}^2$, low p-value, and residuals distribute nearly symmetric.

So over all, this model is the "right" one.

#2

(a)

$$Y_n = \beta_0 + \beta_1 X_n + \epsilon_n$$

$$E(\epsilon_n) = 0$$

$$E(Y_n) = E(\beta_0 + \beta_1 X_n + \epsilon_n)$$

$$= E(\beta_0) + E(\beta_1 X_n) + 0$$

$$= \beta_0 + \beta_1 X_n$$

$$E(\overline{Y}) = E\left[\frac{1}{N}\sum_{n=1}^{N} Y_n\right] = \frac{1}{N}\sum_{n=1}^{N} E(Y_n) = \frac{1}{N}\sum \beta_0 + \beta_1 X_n = \beta_0 + \beta_1 \overline{X}$$

$$E(\hat{\beta}_1) = E\left(\frac{\Sigma(X_n - \overline{X})\cdot Y_n}{\Sigma(X_n - \overline{X})\cdot X_n}\right)$$

$$= \frac{1}{\Sigma(X_n - \overline{X})X_n} \cdot \Sigma(X_n - \overline{X}) \cdot E(Y_n)$$

$$= \frac{1}{\Sigma(X_n - \overline{X})X_n} \cdot \Sigma(X_n - \overline{X}) \cdot [\beta_0 + \beta_1 \overline{X}]$$

$$= \frac{\beta_0 \, \cancel{\Sigma(X_n - \overline{X})}^{\,0}}{\Sigma(X_n - \overline{X})X_n} + \frac{\beta_1 \, \cancel{\Sigma(X_n - \overline{X})X_n}}{\cancel{\Sigma(X_n - \overline{X})X_n}}$$

$$= \beta_1$$

(b)

$$V(\hat{\beta}_1) = V\left(\frac{\sum_{n=1}^{N}(x_n - \bar{x}) \cdot y_n}{S_{xx}}\right)$$

$$= \left(\frac{1}{S_{xx}}\right)^2 \cdot \sum_{n=1}^{N} (x_n - \bar{x})^2 \cdot V(y_n)$$

$$= \left(\frac{1}{S_{xx}}\right)^{\cancel{2}} \cdot \left[\cancel{\sum_{n=1}^{N} (x_n - \bar{x})^2}\right] \sigma^2$$

$$= \frac{\sigma^2}{S_{xx}}$$

(c)

$$V(\hat{\beta}_0) = V(\bar{y} - \hat{\beta}_1 \bar{x})$$

$$= V(\bar{y}) + V(-\hat{\beta}_1 \bar{x}) + 2\,\mathrm{cov}(\bar{y}, -\bar{x}\hat{\beta}_1)$$

$$= V(\bar{y}) + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x}\,\mathrm{cov}(\bar{y}, \hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + \bar{x}^2\left(\frac{\sigma^2}{S_{xx}}\right) - 2\bar{x}\,\mathrm{cov}(\bar{y}, \hat{\beta}_1)$$

$$Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\sum_{i=1}^{n} \frac{1}{n} y_i , \sum_{j=1}^{n} \frac{x_j - \bar{x}}{S_{xx}} y_j \right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_i - \bar{x}}{n S_{xx}} Cov(y_i, y_j)$$

$$= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{n S_{xx}} \sigma^2 + 0$$

$$= 0$$

$$V(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}^2 \left(\frac{\sigma^2}{S_{xx}} \right)$$

$$= \sigma^2 \frac{S_{xx} + N\bar{x}^2}{NS_{xx}}$$

$$= \sigma^2 \frac{\sum_{n=1}^{N} (x_n - \bar{x})^2 + N\bar{x}^2}{NS_{xx}}$$

$$= \sigma^2 \frac{\sum_{n=1}^{N} (x_n^2 - 2\bar{x} x_n + \bar{x}^2) + N\bar{x}^2}{N S_{xx}}$$

$$= \sigma^2 \left(\frac{\sum_{n=1}^{N} x_n^2}{N \cdot S_{xx}} \right)$$

$$= \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}} \right)$$

(d) Yes. we can.

From the lecture 2&3 the Gauss-Markov theorem the $\hat{\beta}_1$ is independent normally-distributed. Since Y_n is normally distributed with β_1, a linear combination of independent normal random variables is normal. Since $E[\hat{\beta}_1] = \beta_1$. We can know $\hat{\beta}_1$ is linear combination of Y_n.

$\hat{\beta}_1 = \sum_{i=1}^{n} X_i Y_i$ estimator is the sum of independent random variable.

And from the CLT. the sum and means of independent random variable tend to be Normally distributed in large samples.

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim N(0,1)$$

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\hat{SE}[\hat{\beta}_1]} \sim N(0,1)$$

$$\Rightarrow \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

#3

I will use Logistic Regression Model. From the lecture 16 & 17.

Base on that, we will have $P(x) = P[Y=1 \mid X=x]$ & $P(Y=0 \mid X=x) = 1-P(x)$

And we can define the LRM which is $\log\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots \beta_M X_M$

Then we add a second index to note that it is being appiled to each observation that is $\log\left[\frac{P(X_i)}{1-P(X_i)}\right] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{mi}$. $i = 1 \ldots m$, then we apply the inverse logit transformation, using to following function to

$$P(X_i) = P[Y_i = 1 \mid X_i = x_i] = \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im}) / (1 + \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im}))$$

$$1 - P(X_i) = P[Y_i = 0 \mid X = x_i] = 1 / 1 + \exp(\beta_0 + \beta_1 X_{i1} + \cdots \beta_m X_{iM})$$

Code:

```
?mtcars
car = mtcars[, c("vs","wt","disp")]

# GLM model
car_model = glm(vs ~ wt + disp, data = car, family = "binomial")
summary(car_model)
coef(summary(car_model))

#Predicting probabilities for 0 and 1
x = predict(model,new_data =data.frame(wt = 2.8, disp = 160),type = "response")
y = as.data.frame(x)

y['Prob_0'] = 1 - y$x
colnames(y) = c("Prob_1","Prob_0")
y
```

estimate for wt, disp

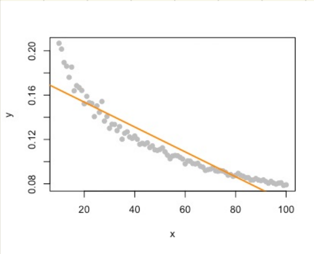
```
> coef(summary(car_model
               Estimate
(Intercept)  1.60859260
wt           1.62635325
disp        -0.03443373
```

Prob: "1" "0"

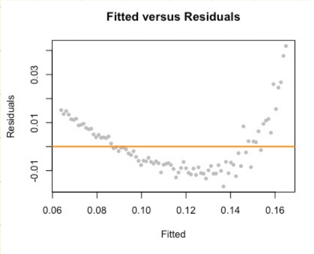
```
> final_pred
```

	Prob_1	Prob_0
Mazda RX4	0.589098973	0.41090103
Mazda RX4 Wag	0.684593276	0.31540672
Datsun 710	0.840625523	0.15937448
Hornet 4 Drive	0.114398085	0.88560192
Hornet Sportabout	0.005525208	0.99447479
Valiant	0.374768453	0.62523155
Duster 360	0.006817186	0.99318281
Merc 240D	0.851350376	0.14864962
Merc 230	0.867993906	0.13200609
Merc 280	0.807236894	0.19276311
Merc 280C	0.807236894	0.19276311
Merc 450SE	0.219433362	0.78056664
Merc 450SL	0.139202255	0.86079774
Merc 450SLC	0.149234967	0.85076503
Cadillac Fleetwood	0.002224998	0.99777500
Lincoln Continental	0.004453588	0.99554641
Chrysler Imperial	0.007772280	0.99222772
Fiat 128	0.922487560	0.07751244
Honda Civic	0.835966790	0.16403321
Toyota Corolla	0.895173677	0.10482632
Toyota Corona	0.814883948	0.18511605
Dodge Challenger	0.026171375	0.97382862
AMC Javelin	0.036518408	0.96348159
Camaro Z28	0.014802949	0.98519705
Pontiac Firebird	0.002700619	0.99729938
Fiat X1-9	0.884456037	0.11554396
Porsche 914-2	0.720433157	0.27956684
Lotus Europa	0.688821969	0.31117803
Ford Pantera L	0.004858739	0.99514126
Ferrari Dino	0.754118670	0.24588133
Maserati Bora	0.049742275	0.95025772
Volvo 142E	0.876897600	0.12310240

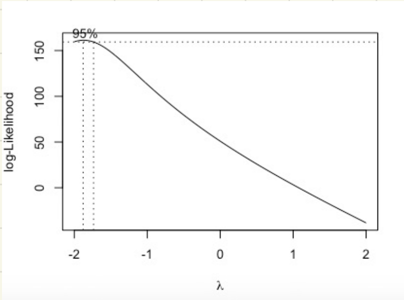
#4



By plot the original model, the SLR model clearly does not look like a good fit and the residuals vs fitted value plot looks significantly worse and poor fit. I have tried seveal transmations and decided to use boxcox as following:

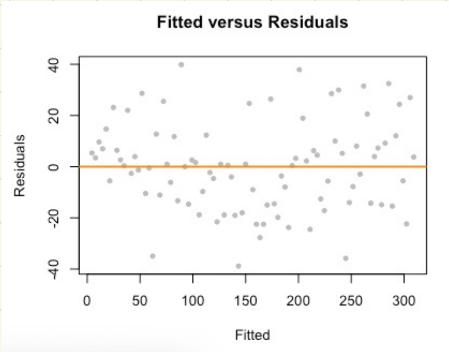


$\Downarrow$



As we see the max value in the log-likelihood $\sim \lambda$ is -1.8787

$\Downarrow$



We can clearly noticed this plot is much better than the original plot. which is even and symmetric

```
        studentized Breusch-Pagan test

data:  m
BP = 1.9594, df = 1, p-value = 0.1616

>
> shapiro.test(resid(m))

        Shapiro-Wilk normality test

data:  resid(m)
W = 0.98434, p-value = 0.3466
```

Then by using BP test, we can get p-value 0.16 which is large enough to imply homoscedasticity. And by using shapiro test, the p-value is 0.35 that imply normality base on this test.

```
Call:
lm(formula = z ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.839 -13.660   0.425   9.405  39.870 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -29.42716    4.21956  -6.974 5.22e-10 ***
x             3.38592    0.06923  48.909  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.35 on 89 degrees of freedom
Multiple R-squared:  0.9641,	Adjusted R-squared:  0.9637 
F-statistic:  2392 on 1 and 89 DF,  p-value: < 2.2e-16
```

Bcse we have nice R^2 and $\hat{R}^2$, low p-value, and residuals distribute nearly symmetric.

So over all. this model is the "right" one.