96.75

STA 207 HW-5

Due Date: 10/22 by 10:20AM

Derin Gezgin | Camel ID: 00468038

Notes:

1. Make sure you label all plots in the title and then use them in the context.
2. Make sure you do not put R code in the middle of each problem, instead you present the entire solution as a report.

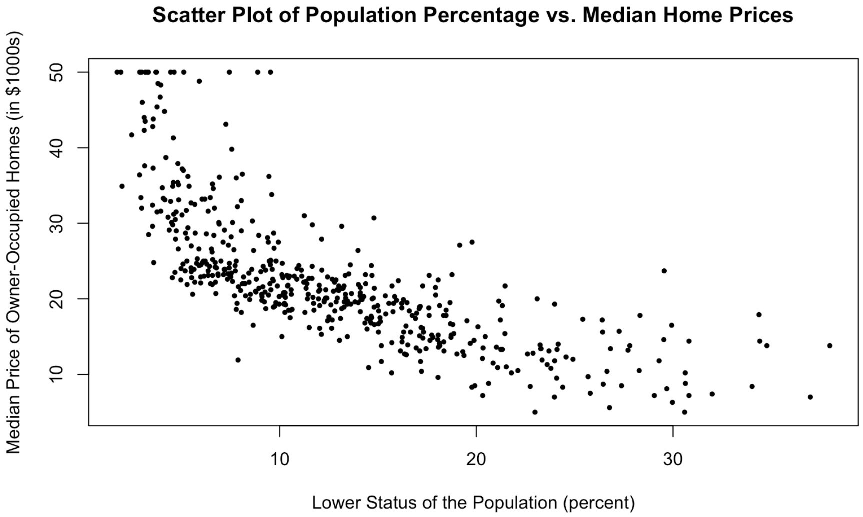
We will work with the data on median housing prices in neighborhoods in the suburbs of Boston.

install.packages("MASS") | library(MASS) | data(Boston)

Our goal is to analyze the relationship between neighborhood poverty and housing prices.

Using dataset Boston in the R package MASS answer the following:

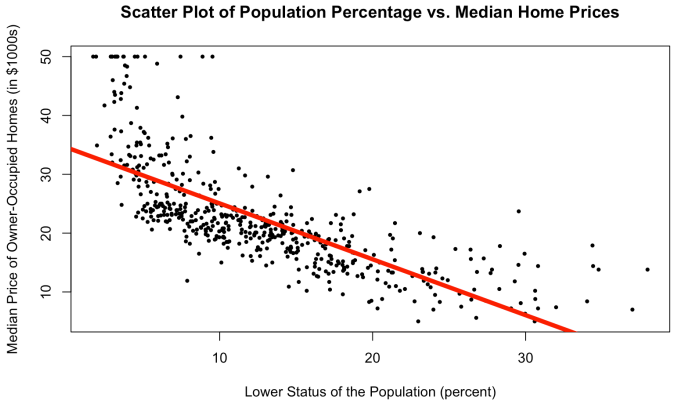
a.) [5 points] Make a scatterplot with x as % in Poverty using variable lstat and y as the Median home prices using variable medv. Comment on the relationship.



✓

From the scatterplot, we can see that there's a ***strong*** relationship in the ***negative*** direction. As the X value increases, the Y value decreases so we can conclude that the relationship is *negative.* At the same time, we can see a few outliers around X=20, 30, and towards the end. While the relationship is strong, even from this simple scatterplot, we can see that the relationship is not linear. ✓

b.) [5 points] Regress medv (Y) on lstat (X) and report the fitted model.



The fitted equation: $\boldsymbol{\widehat{Y} = -0.95X + 34.554}$

Where:

x = Lower Status of population in percentage

$\widehat{Y}$= **Estimated** median value of Owner-Occupied Homes

c.) [7.5 points] Share goodness of this fit.

A) Coefficient of Determination

R^2: Square of cor(X,Y) for SLR.

cor(X,Y): -0.7376627

R^2: 0.5441

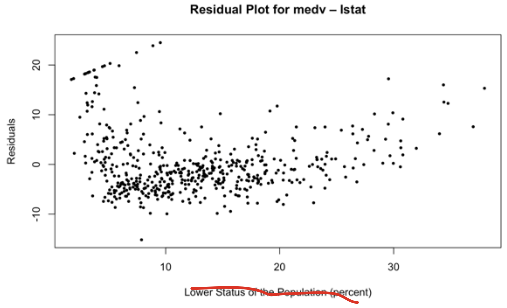
This means that **54%** of the variation in fitted Y values can be explained by X. This is a very low value as nearly half of the variation in fitted Y values can't be explained by X.

B) Variability of Errors

The residual standard error → **6.216**

This means that average variation in fitted values is approximately 6.216 which is significant considering this can be a difference of $6,000.

C) Residual Plot



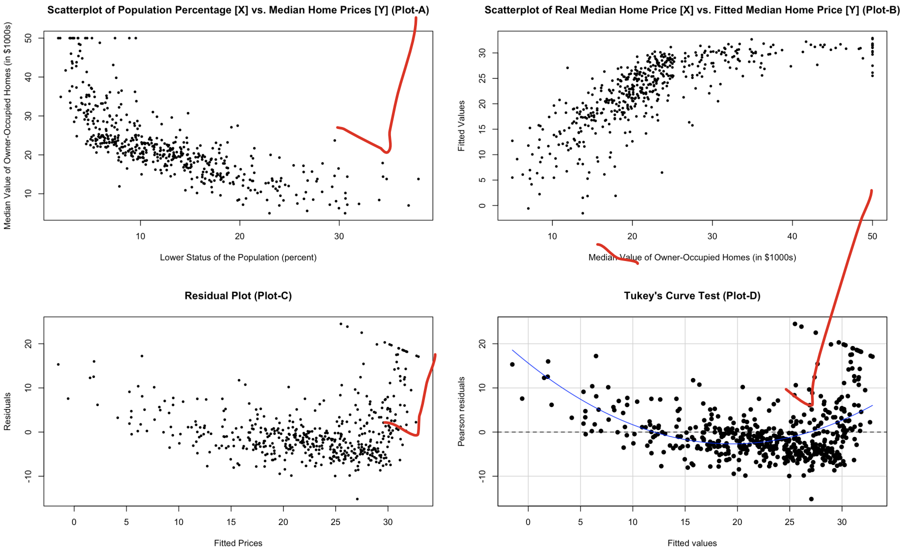
-1, points should be flipped as x should be fitted values

In the residual plot, we can see kind of a clear pattern that goes slightly upwards. Which is not a sign of a good fit.

In general; we cannot conclude that there is a good fit and we have enough evidence to show that this is not a very good fit at all.

d.) [15 points] Report if the LINE conditions are met or not for model in part (c).

L: There is a linear relationship between X and Y variables



From Plot-A and Plot-B, we can see that there's a slightly curved relationship which can't be completely classified as linear.

In the residual plot (Plot-C), we can see that there's a clear linear pattern towards the end of the X-axis.

From Plot-D, we can see that the blue-fit line shows a curved pattern, which shows that the residuals show a pattern.

Tukey Test

H_0 = The population percentage and the median home prices does have a linear relationship.

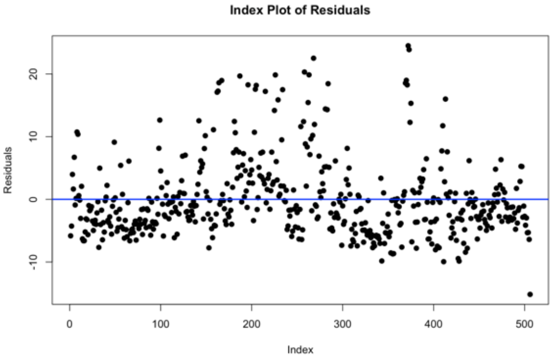
H_A = the population percentage and the median home prices does not have a linear relationship.

Level of Significance → 0.05 | Test Statistic → 11.63 | p-Value → 2.986e-31

We can reject the null hypothesis as the p-Value is less than the level of significance. Which means that at 95% level of confidence, we have enough evidence that the population percentage and the median home prices does not have a linear relationship.

Linearity assumption fails.

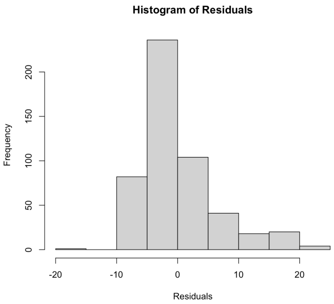
I: Errors are independent

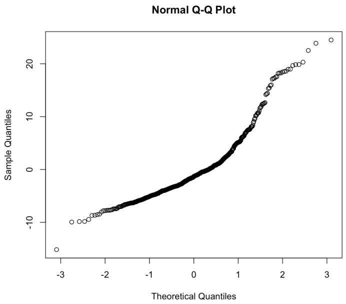


If we check the index plot of the residuals, we can see that while there's a *kind of* random distribution in the upper side of the plot, in the lower side of the plot, most of the points are cumulated around the line. At the same time, while it is not very clear, we can conclude that there's a repetitive pattern in residuals, as well. ***We cannot conclude that the errors are independent.***

-1, there is no obvious pattern so assumption holds

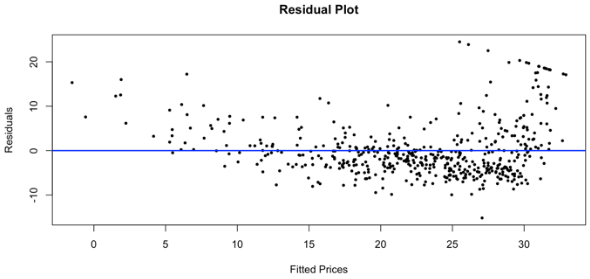
N: Errors are normally distributed





From the histogram, we can see that there's a very non-normal distribution. The middle value box is very saturated and high. The higher and the lower parts of the histogram are very low. Which is very far from the normal distribution.
If we check the Normal Q-Q Plot, there is not a linear relationship and we can see that there is a slight curve.
We cannot conclude the errors are normally distributed.

E: Errors are homoscedastic/have equal variance, Var(ϵ) = σ^2



While it's not extremely clear, we can see that the errors are not really randomly distributed and it expends a little bit towards the end of the plot.

BP-Test

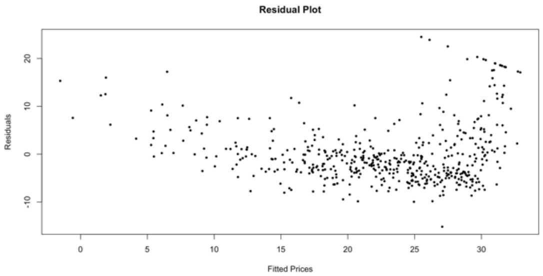
Ho: Error variance is constant, homoscedasticity.
Ha: Error variance is not constant, heteroscedasticity

Level of Significance = 0.05 | Test Statistics = 15.497 | P-value= 0.00008262

As the p-Value is less than our level of significance, we can say that we **don't** have enough evidence to conclude that the Error variance is not constant and there's a heteroscedastic distribution.

Homoscedasticity assumption also fails.

e.) [2.5 points] Using residual plot, can you tell which assumption(s) are failing?



Only looking at the residual plot, we can say that there's a slight heteroscedasticity in the distribution of errors. At the same time, we can see that there is a clear linear pattern towards the end of the graph.

I can conclude that: *Linearity and Homoscedasticity assumptions fail.*

f.) ***[40 points] Log transformation:*** ***To see if taking the log of poverty percentage (lstat) and median home values (medv) adjusts the relationship below.***

- ***[5 points] Regress log(medv) on log(lstat) and report the fitted model.***

  The fitted model: $\boldsymbol{log(\widehat{Y}) = -0.5598log(X) + 4.3618}$

  Where:

  x = Lower Status of population in percentage

  $\widehat{Y}$= **Estimated** median value of Owner-Occupied Homes

  [5 points] Regress log(medv) on lstat and report the fitted model.

  The fitted model: $\boldsymbol{log(\widehat{Y}) = -0.046X + 3.6176}$

  Where:

  x = Lower Status of population in percentage

  $\widehat{Y}$= **Estimated** median value of Owner-Occupied Homes

  [5 points] Regress medv on log(lstat) and report the fitted model.

  The fitted model: $\boldsymbol{\widehat{Y} = -12.4810log(X) + 52.1248}$

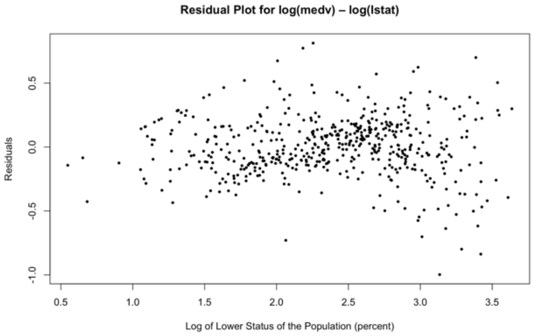
  Where:

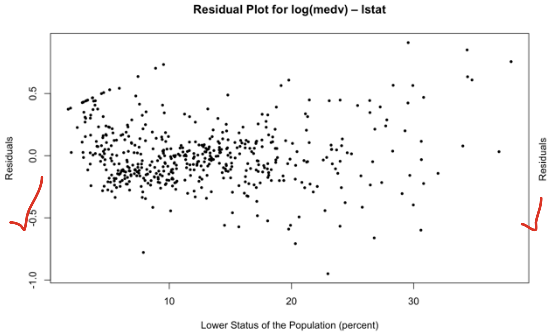
  x = Log of the Lower Status of population in percentage

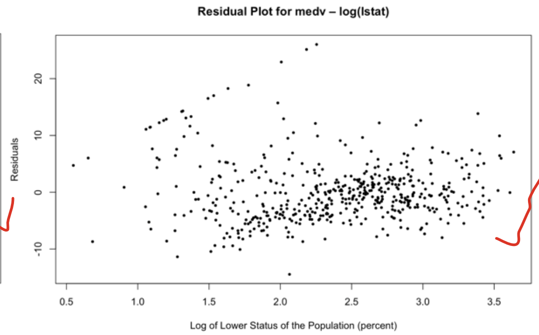
  $\widehat{Y}$= **Estimated** median value of Owner-Occupied Homes

- ***[5 points] Compare models (a), (b), and (c) using goodness of fit measures.***

Values	***Model (A) log(medv) – log(lstat)***	***Model (B) log(medv) – lstat***	***Model (C) medv – log(lstat)***
Coefficient of Determination	***0.6773***	***0.6481***	***0.6649***
Variability of Errors	***0.2324***	***0.2427***	***5.329***







In general, **coefficient of determination** shows the variation in fitted Y values can be explained by X. At the same time, **variability of errors** is the approximate average variation in fitted values. Also, we'd like to see a random distribution in the residual plot.

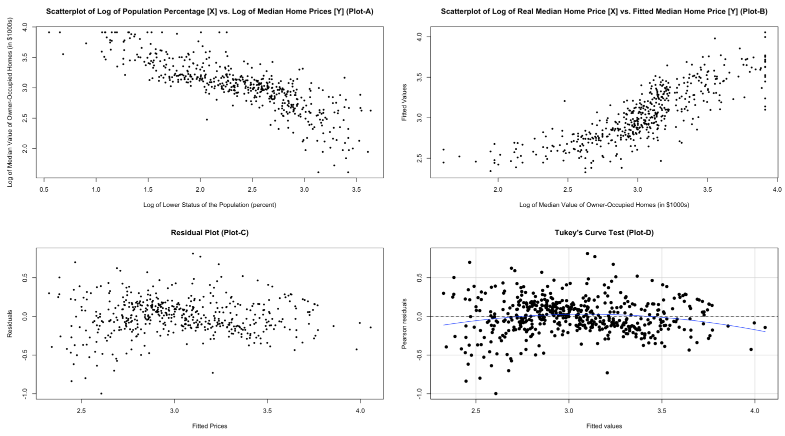
In an optimal model, we'd like to have a higher coefficient of determination to explain a larger variation of fitted Y values by X. On the other hand, we'd like to have a lower variability of errors which is important to have a better model.

Considering these conditions, Model (A) [log(medv) – log(lstat)] is the best model as it has the highest coefficient of determination and a lower variability of errors.

- *[15 points] Compare models (a), (b), and (c) using LINE conditions*

Testing the LINE conditions for Model (A) – log(medv) on log(lstat)

L: There is a linear relationship between X and Y variables



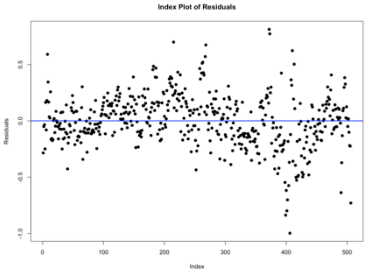
✓

From Plot-A and Plot-B, we can see that the relationship is much linear than what we had initially. It can still be said that the relationship is very slightly curved but its linearity is definitely much significant.
Similarly, in Plot-C the clear pattern in our initial model is disappeared and there is a significantly more random distribution in the residuals. We still cannot say that it is completely random as it seems like there is a pattern. ✓
In Plot-D, we can see that the huge curve in the initial graph is disappeared but there is a slight curve anyways.
Tukey Test
H_0 = The population percentage and the median home prices does have a linear relationship.
H_A = the population percentage and the median home prices does not have a linear relationship. ✓
Level of Significance → 0.05 | Test Statistic → -3.2155 | p-Value → 0.0013

We can reject the null hypothesis as the p-Value is less than the level of significance. Which means that at 95% level of confidence, we have enough evidence that the population percentage and the median home prices does not have a linear relationship.

In general, despite seeing better plots overall, all the plots still cannot conclude there is a linearity as they have slight defects. The Tukey test also supports this claim as it fails. ***Linearity assumption fails.*** ✓

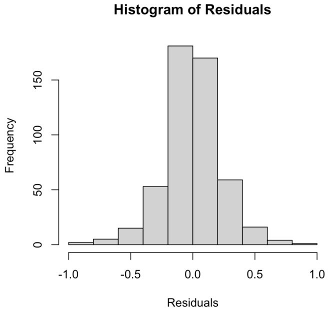
I: Errors are independent

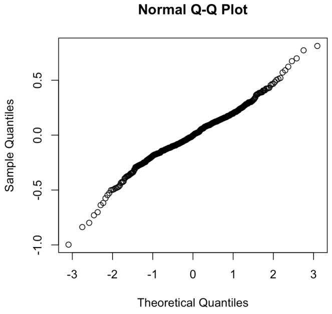


From the index plot of residuals, we can see the cumulative pattern in the initial model is disappeared. At the same time, there a new wave pattern in the residuals which can classified as a cyclic pattern. **We still *cannot* conclude that the errors are independent.**

-0.25, no pattern so assumption holds

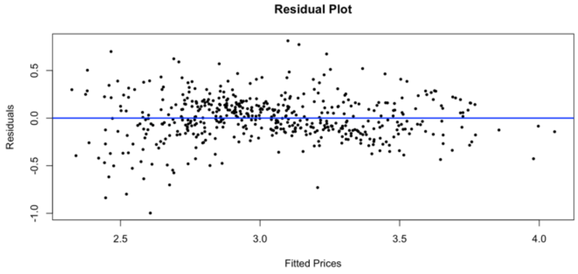
N: Errors are normally distributed


Histogram of Residuals


Normal Q-Q Plot

From the histogram, we can see that the initial skew in the middle is disappeared and slightly distributed to its neighbors. But still, we don't have the *bell curve* of a normal distribution as the middle values are over-saturated.
Similarly, in the Q-Q Plot, while the relationship became more linear, it still is not completely linear and we can say that there is a slight curve.
We cannot conclude the errors are normally distributed.

E: Errors are homoscedastic/have equal variance, $\mathrm{Var}(\epsilon) = \sigma^2$


Residual Plot

We can clearly see that while the residuals were very equally distributed in the first part of the plot, as the X value increases the residuals start to get closer to each other which fails the homoscedasticity assumption.

BP-Test

Ho: Error variance is constant, homoscedasticity.
Ha: Error variance is not constant, heteroscedasticity

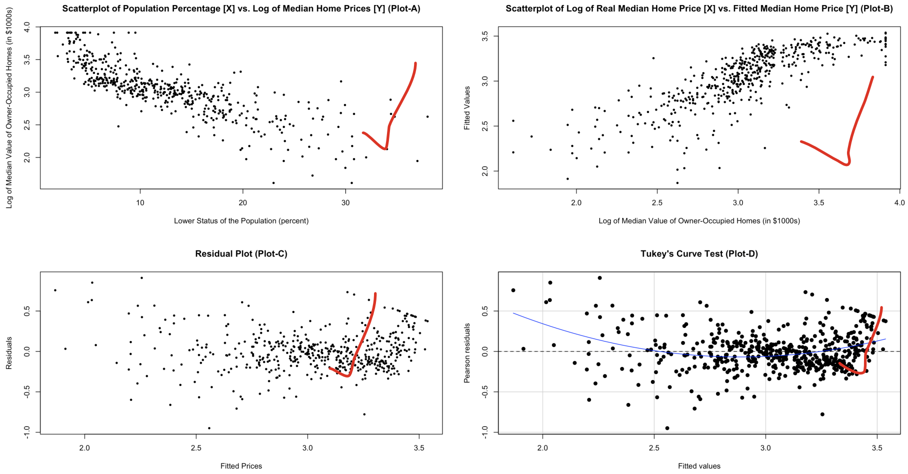
Level of Significance = 0.05 | Test Statistics = 16.316| P-value = 0.00005361

As the p-Value is less than our level of significance, we can say that we **don't** have enough evidence to conclude that the Error variance is not constant and there's a heteroscedastic distribution.

Homoscedasticity assumption also fails.

Testing the LINE conditions for Model (B) – log(medv) on lstat

L: There is a linear relationship between X and Y variables



From Plot-A and Plot-B, we can see that there is a slightly curved but also slightly linear relationship between the axis. At the same time, as the values spread out a lot, the relationship disappears.

Similarly, in Plot-C, rather than having a consistent random distribution (like the beginning of the plot), the random distribution cumulates towards the end of the X-axis.

Lastly, from the Tukey's Curve Test, we can see that the blue-fit-line is curved, which means that residuals still show a pattern.

Tukey Test

H_0 = The population percentage and the median home prices does have a linear relationship.

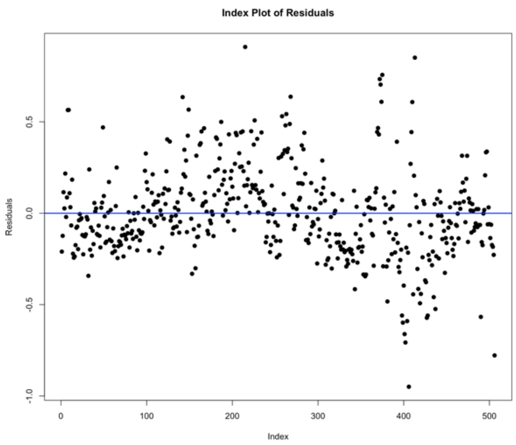
H_A = the population percentage and the median home prices does not have a linear relationship.

Level of Significance → 0.05 | Test Statistic → 7.0738| p-Value → 0.0000000000015

We can reject the null hypothesis as the p-Value is less than the level of significance. Which means that at 95% level of confidence, we have enough evidence that the population percentage and the median home prices does not have a linear relationship.

Linearity assumption fails.

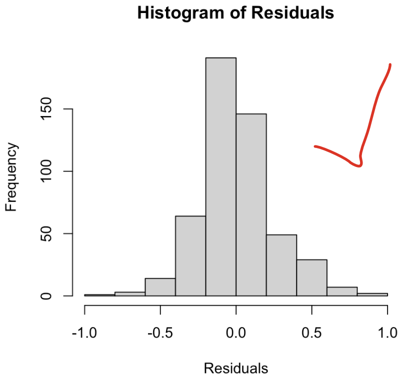
I: Errors are independent

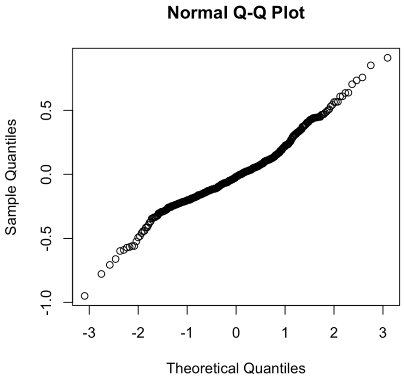


If we check the index plot of the residuals, we can see a very slight curved pattern that goes like a "s" throughout the graph. This pattern can also be classified as cyclic. ***We cannot conclude that the errors are independent.***

-0.25, no pattern so assumption holds

N: Errors are normally distributed



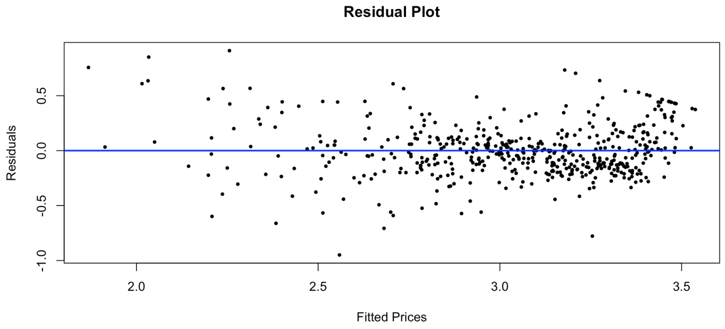


Similar to Model (A), we can see that the bell-curve pattern in the residuals is more prominent. On the other hand, the middle values are over-saturated that we cannot classify this as a normal distribution.

The Q-Q Plot is also much linear compared to our initial Q-Q plot but the line is still not linear.

We cannot conclude the errors are normally distributed.

E: Errors are homoscedastic/have equal variance, Var(ϵ) = σ^2



It is very clear in the residual plot that while the residuals start with a very wide spread, they start to get closer and closer. Which proves that the distribution of errors is not homoscedastic.

BP-Test

Ho: Error variance is constant, homoscedasticity.
Ha: Error variance is not constant, heteroscedasticity

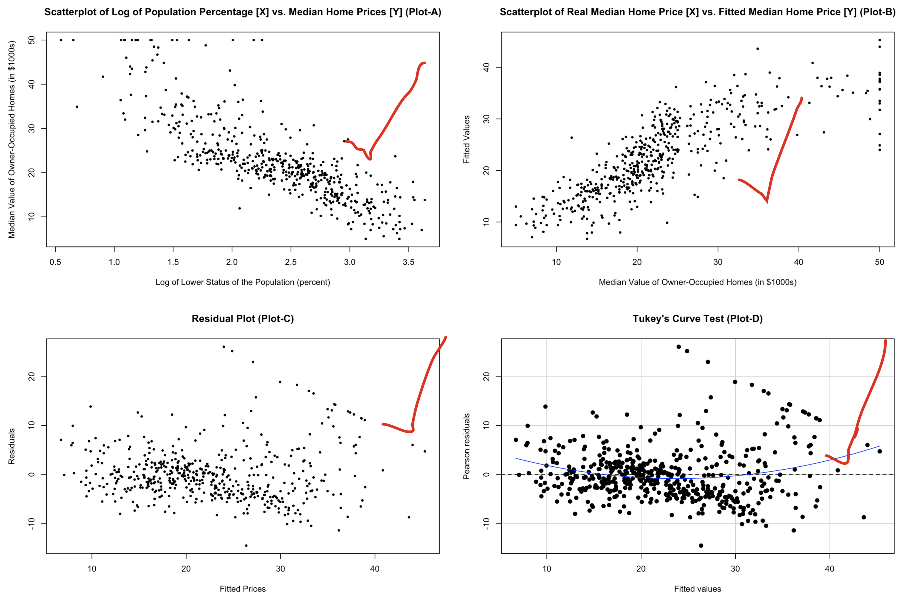
Level of Significance = 0.05 | Test Statistics = 29.583 | P-value= 0.000000054

As the p-Value is less than our level of significance, we can say that we **don't** have enough evidence to conclude that the Error variance is not constant and there's a heteroscedastic distribution.

Homoscedasticity assumption also fails.

Testing the LINE conditions for Model (C) –medv on log(lstat)

L: There is a linear relationship between X and Y variables



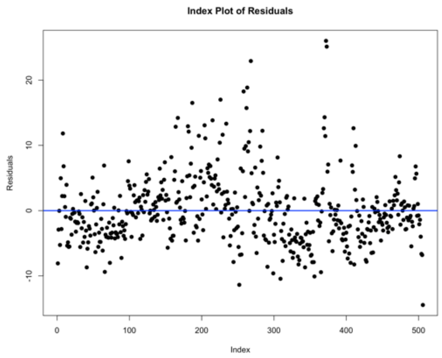
From Plot-A and Plot-B, we can see a linear pattern but the issue is, the values are spread very further from the line so it cannot be classified as a very strong linear relationship.
From the residual plot, we can see that the residuals have kind of a random relationship but we can still conclude that there are patterns in the center and top right of the plot.
In Plot-D, we can see the blue-fit line curved, which shows that residuals show a pattern.

Tukey Test

H_0 = The population percentage and the median home prices does have a linear relationship.
H_A = the population percentage and the median home prices does not have a linear relationship.
Level of Significance → 0.05 | Test Statistic → 4.15 | p-Value → 0.0000332

We can reject the null hypothesis as the p-Value is less than the level of significance. Which means that at 95% level of confidence, we have enough evidence that the population percentage and the median home prices does not have a linear relationship.
Linearity assumption fails.

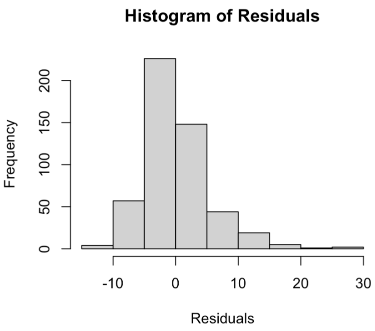
I: Errors are independent

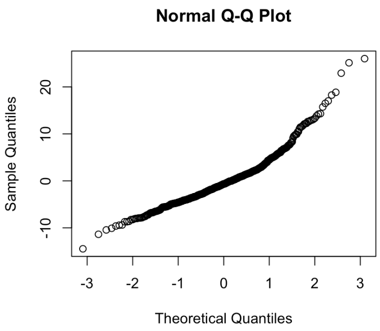


From the index-plot of the residuals, it is so hard to see a clear cyclic pattern but we can see that -like the initial graph, there is a lot of outlier residuals in the upper half of the line, while the lower half is cumulated around the line which is a repetitive pattern over the graph. ***We cannot conclude that the errors are independent.***

-0.25, no pattern so assumption holds

N: Errors are normally distributed

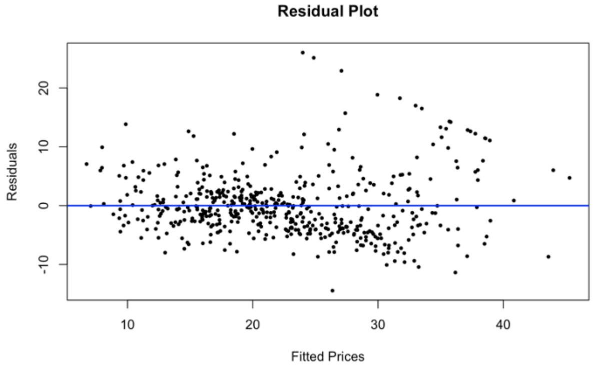




✓

When we check the histogram we can see that the center values in the graph is over-saturated and the neighboring values are very low. While it would be a bell-curve distribution if we had a more balanced distribution of values, right now we cannot conclude that.
Similarly, in the Q-Q plot, we can see that there is a clear curve pattern rather than a completely linear relationship.
We cannot conclude the errors are normally distributed. ✓

E: Errors are homoscedastic/have equal variance, Var(ϵ) = σ^2



While this is the best residual we saw in terms of homoscedasticity, there's a slight spread towards the end of the plot which shows heteroscedasticity.

BP-Test

Ho: Error variance is constant, homoscedasticity.
Ha: Error variance is not constant, heteroscedasticity

Level of Significance = 0.05 | Test Statistics = 28.667| P-value= 0.000000086 ✓

As the p-Value is less than our level of significance, we can say that we **don't** have enough evidence to conclude that the Error variance is not constant and there's a heteroscedastic distribution.

Homoscedasticity assumption also fails.

- ***[5 points] Compare what changed in residual plot reported in part (e)***

Original Residual Plot vs. Model (A)

We can definitely see that the residuals started to have a more random distribution in this plot. There are still signs of slight homoscedasticity but it's very obvious that the distribution is more random.

Original Residual Plot vs. Model (B)

We can see that the pile of data-points in the right on the original plot is shifted towards the left in this residual plot. We can still see that the errors are spreading open in this residual plot.

Original Residual Plot vs. Model (C)

Among all the plots, this plot is the most similar to the original residual plot though the points are a bit more randomly spread out in this plot.

In general;

We can say that the all the plots definitely show a more random spread than the original residual plot. As we saw in the goodness of fit part, Model (A) shows the most random and spread fit.

- ***[5 points] Did the transformation help with your analysis, why or why not?***

The transformation slightly helped with the analysis as general problems like non-linearity in graphs, residual distribution, etc. became less severe. But we weren't able to pass any of the tests so we cannot conclude that the transformation helped us to find the best model for this problem. I think that the reason it wasn't able to help us really much is the problems are in a large magnitude which is not possible to fix with transformations like this.

g.) <u>[20 points] Polynomial model:</u> Regress medv on lstat and lstat squared, that is a polynomial model. Report the fitted model. Share goodness of this fit. Check LINE conditions for this model. What changed in the residual plot reported in (e)? Did the transformation help with your analysis, why or why not?

<u>Reporting the Fitted Model</u>

The fitted model: $\hat{Y} = \mathbf{42.862 - 2.333}X \mathbf{+ 0.044}X^2$

Where:

x = Lower Status of population in percentage

$\hat{Y}$= **Estimated** median value of Owner-Occupied Homes

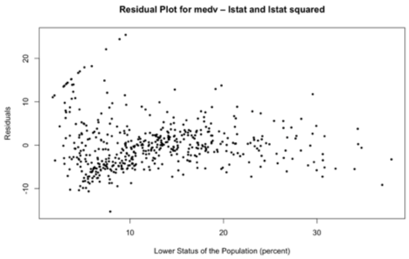
<u>Goodness of Fit</u>

R-Squared Value: 0.6407

This means that **64%** of the variation in fitted Y values can be explained by X. This is a higher value than our initial linear model. But it's still not very good.

Residual Standard Error: 5.524

This means that average variation in fitted values is approximately 5.524 which is significant considering this can be a difference of $5,500. But it's lower than our original linear model so we can say that it is better.



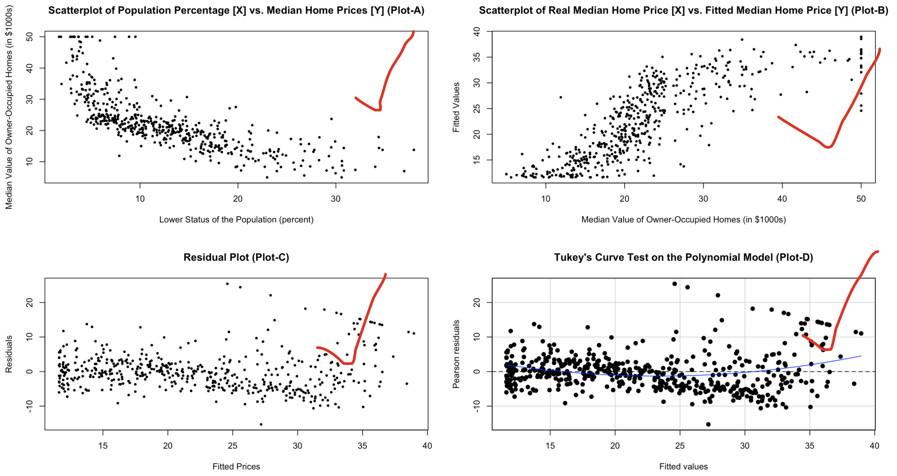
Residual Plot for medv – lstat and lstat squared

If we look at the residual plot, we can see that residuals are more equally distributed than our initial linear model.

We can conclude that this is a better model than the initial one.

Reporting the LINE conditions for the model

L: There is a linear relationship between X and Y variables



Scatterplot of Population Percentage [X] vs. Median Home Prices [Y] (Plot-A)

Scatterplot of Real Median Home Price [X] vs. Fitted Median Home Price [Y] (Plot-B)

Residual Plot (Plot-C)

Tukey's Curve Test on the Polynomial Model (Plot-D)

From Plot-A we can see that there's a slightly curved relationship which can't be completely classified as linear.

From Plot-B, we can see that the relationship is more linear and which can be considered to satisfy linearity compared to our initial linear model.

In the residual plot (Plot-C), we can see that there's a more random distribution of the residuals than the initial model. But, at the same time, there's a slight line of fit that can be considered as not completely random.

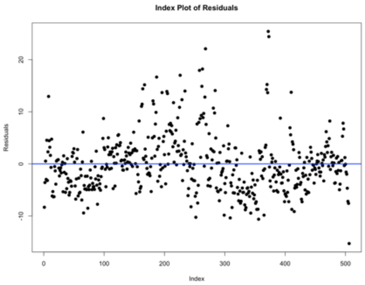
From Plot-D, we can see that the blue-fit line shows a much less curved pattern and a closer line to 0 so we can say that the residuals does not show a clear pattern. Towards the end of the X-axis, there are residuals that are clearly skews the blue line.

Tukey Test
H_0 = The population percentage and the median home prices does have a linear relationship.
H_A = the population percentage and the median home prices does not have a linear relationship.
Level of Significance → 0.05 | Test Statistic → 6.196 | p-Value → 0.000000000578

We can reject the null hypothesis as the p-Value is less than the level of significance. Which means that at 95% level of confidence, we have enough evidence that the population percentage and the median home prices does not have a linear relationship.

Linearity assumption fails.

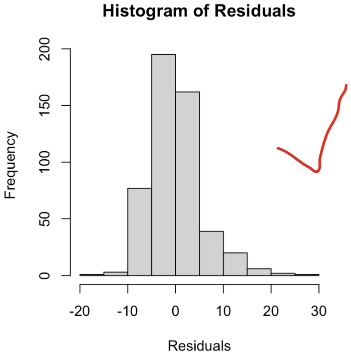
I: Errors are independent



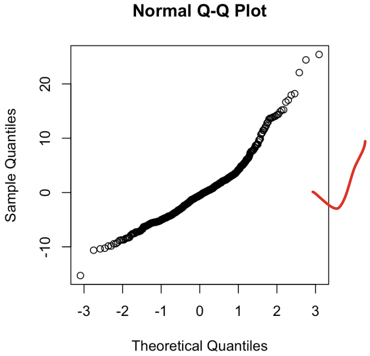
From the index plot of residuals, we can see that there is a slightly cyclic pattern in the residuals like a wave.

We can conclude the independence of errors assumption also fails. -0.25, no pattern so assumption holds

N: Errors are normally distributed



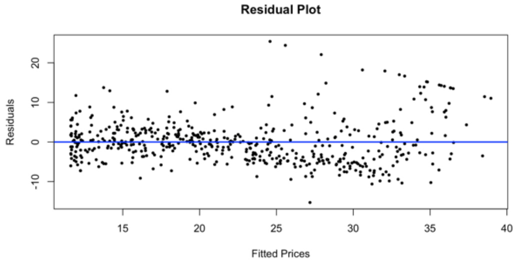
✓



✓

From the histogram, we can see a similar pattern as the previous questions. The middle values are very saturated while boxes neighboring the center has a very low frequency. This shows that the distribution is not normal.
Similarly, in the Normal Q-Q plot, we can definitely see a curved pattern rather than a linear one.
We cannot conclude that the errors are normally distributed. -0.25, assumption might fail

E: Errors are homoscedastic/have equal variance, Var(ϵ) = σ^2



From the residual plot, we can see that the errors are definitely more homoscedastic than the previous questions but the errors still get more further and further as the X-value in the plot increases. Which violates homoscedasticity.

BP-Test

Ho: Error variance is constant, homoscedasticity.
Ha: Error variance is not constant, heteroscedasticity

Level of Significance = 0.05 | Test Statistics = 48.74| P-value= 0.000000000026 ✓

As the p-Value is less than our level of significance, we can say that we **don't** have enough evidence to conclude that the Error variance is not constant and there's a heteroscedastic distribution.

Homoscedasticity assumption also fails.

What changed in the residual plot reported in (e)? Did the transformation help with your analysis, why or why not?

When we look at the residual plot in (e), we can see a much-spread distribution. On the other hand, in the residual plot of this model, we can see that the values are more concentrated around the same line. But, in both of the residual plots, towards the end of the plot, there are some outliers which creates an issue. We can say that the residual plot is improved but it is not perfect. ✓

The transformation slightly helped with my model as my R-Squared value increased from **0.54** to **0.64**. Similarly, my variability of errors dropped to **5.524** from **6.216**. The reason I said slightly is it did not make a huge difference that our log transformations is not able to do. Moreover, Model (A) in the log transformations was able to get an R-Squared value of **0.6773** and variability of errors of 0.2324 which is a significant difference. My final verdict would be: it helped slightly but I think that there is a better model (which we did not learn) that can fit into this model really well. ✓