STA 207 HW-5
Due Date: 10/22 by 10:20AM

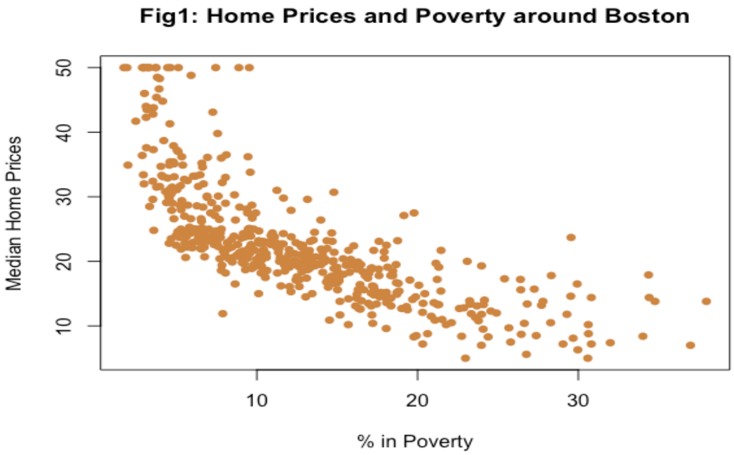
Notes:

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.

Problem 1: We will work with the data on median housing prices in neighborhoods in the suburbs of 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.) 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.

In Fig. 1, there is a negative relation between the median home prices and percentage of poverty in Boston. This is not surprising as one expects that where there is more poverty, there are lower home prices. It is also important to note that this relation does not appear linear as there is an obvious curve, meaning that every increase in poverty is not linearly associated with an equal decrease in house prices. There are no obvious outliers.



b.) Regress medv (Y) on lstat (X) and report the fitted model.

Fitted Model

$\hat{Y} = 34.55 - 0.95X$, where

X is the % in Poverty in Boston

Y is the Median home prices in Boston in 1000s of $.

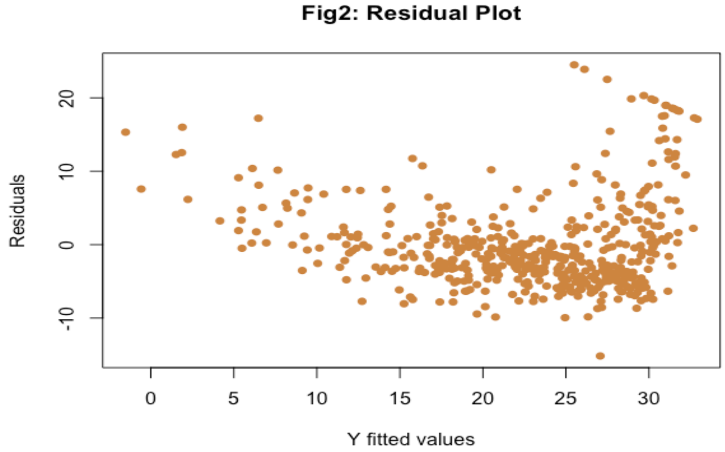
c.) Share goodness of this fit.

Goodness of fit:

i) $R^2 = 0.5441$ showing that approximately 54.4% change in median house prices in Boston is explained by the % of poverty.

ii) $\hat{\sigma} = 6.216$, meaning that on the average the fitted values of median house prices vary 6,216$ from the observed prices.

iii) Residual Plot: In Fig.2 there is a clear curve pattern in the residual plot, meaning that the model is not extracting all information present in the median house prices.



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

Testing Linearity Assumption: To test whether there is a linear relation between the median house prices and percentage of poverty in Boston

In Fig. 3, Plot a below shows a negative, non-linear relationship. There is no obvious outlier (as explained in part a above). Plot b below shows that the estimated value of median house prices (using model 1) agrees with the observed (sample) prices but there is a curve there too. Plot c below is the residual plot which shows clear curve pattern. Plot d) is the Tukey's curve which shows deviation from the flat line, indicating presence of non-linearity. All plots show that linearity assumption is failing. Next, we do the Tukey's curve test to confirm:

Step 1: Null hypothesis (Ho): Linearity assumption holds

Alternative Hypothesis (Ha): Linearity assumption fails

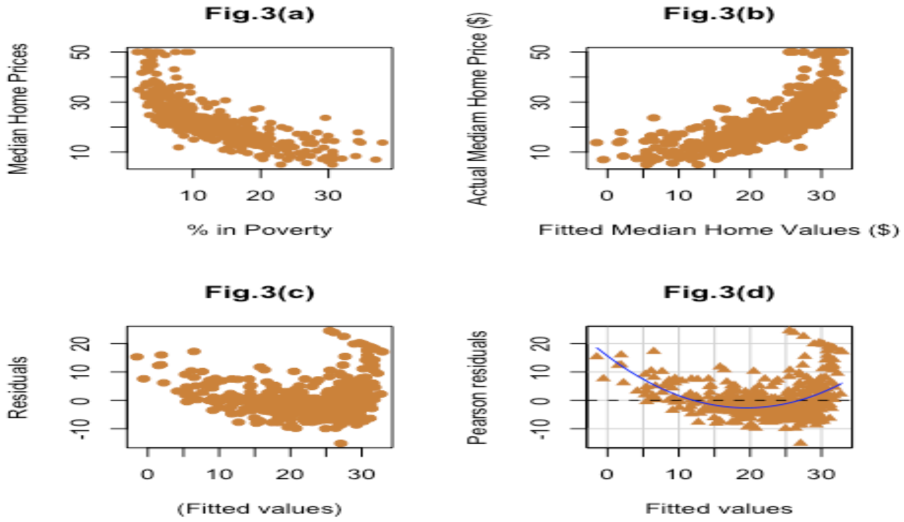
Step 2: Set Level of significance α as 0.05 (default value)

Step 3: Test statistic= 11.63

Step 4: P-value=$2.98\times 10^{-31}\sim 0$

Step 5: Since P-value$< \alpha$, so 5eject Ho

Conclusion: At 5% level of significance, linearity assumption fails.



Independence of errors assumption

Fig.4(c) below is the index plot of errors and there is no obvious pattern. Hence the independence of errors assumption holds based on the plot.

Normality of errors assumption

Fig.4(a) shows histogram of errors which is clearly right skewed with a single peak. Fig.4(b) shows the QQ plot, with a clear curve in the plot. Therefore, we can say that the normality of errors assumption fails.

Errors are homoscedastic assumption

In Fig.4(d), the residual plot with a mean of errors line. We can see that there is a curve and variance of errors is not constant. There is evidence of homoscedasticity failing.

Using the Breusch-Pagan test:

Step 1: H0: Errors are homoscedasticity)
Ha: Errors are heteroscedasticity

Step 2: $\alpha = 0.05$ (default value)

Step 3: Test statistic = 15.497

Step 4: P-value = 0.00008

Step 5: Since P-value < α, so we reject the H0.

Conclusion: At 5% level of significance, we have enough evidence that errors are not homoscedastic.

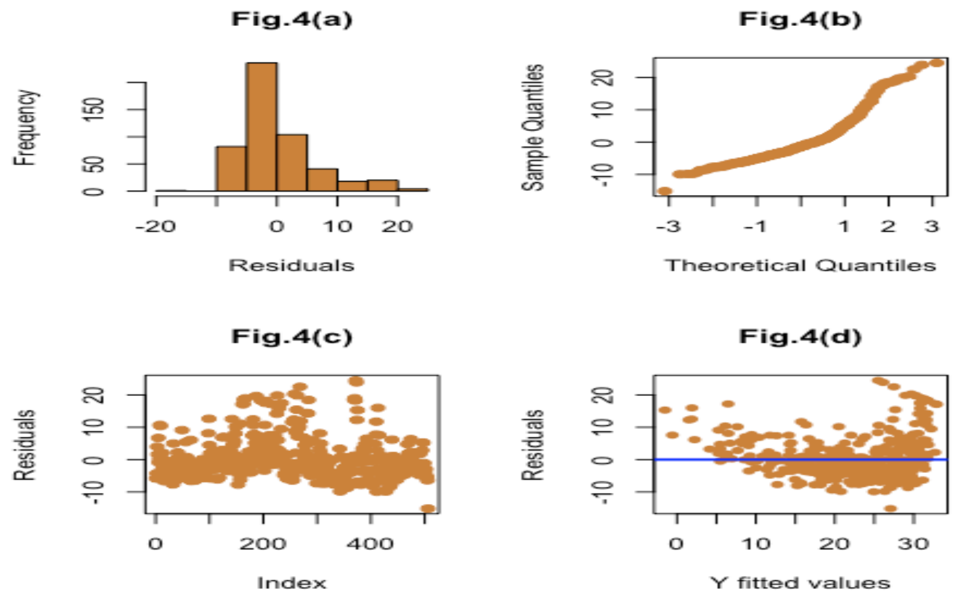


Fig.4(a)

Fig.4(b)

Fig.4(c)

Fig.4(d)

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

In Fig.2 there is a clear pattern in the residual plot, meaning that the model is not extracting all information present in the median house prices. We can see that linearity assumption fails and there is also a change in the variance of errors which indicates that homoscedasticity of residuals might fail.

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

- Regress log(medv) on log(lstat) and report the fitted model.
- Regress log(medv) on lstat and report the fitted model.

- Regress medv on log(lstat) and report the fitted model.
- Compare these three models using goodness of fit measures.
- Compare these three models using LINE conditions
- Did the transformation help with your analysis, why or why not?

With X as the % in Poverty in Boston and Y as the Median home prices in Boston in 1000s of \$, the three fitted models based on transformations are as follows:

Model (a): $log(\widehat{Y}) = 4.36 - 0.56 \log(X)$,

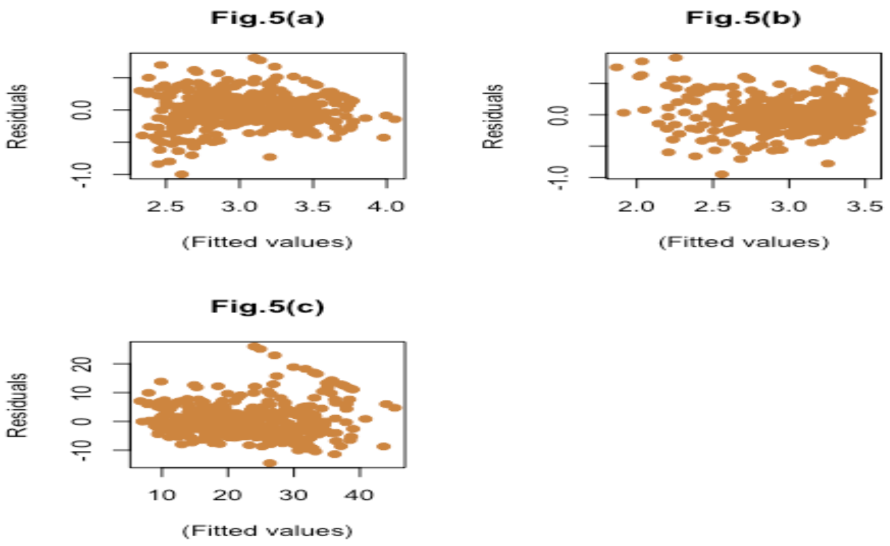
Model (b): $log(\widehat{Y}) = 3.62 - 0.05\ X$,

Model (c): $\hat{Y} = 52.12 - 12.48 \log(X)$.

Comparing the three models using goodness of fit measures.

Model	R^2	$\hat{\sigma}$
(a)	0.6773	0.2324
(b)	0.6481	0.2427
(c)	0.665	5.329

Fig. 5 shows that models first and third have more of a random spread of points as compared to the second model. As compared to the residual plot in Fig. 2 above using the model with original variables, we notice that residual plots(a)-(c) in Fig.5 below from transformed models don't show the obvious curve.



Comparing the above residual plots (Fig. 5 a-c) with residual plot in Fig. 2 (part-e), we see that the pattern in original plot (Fig. 2) is sparse but this is not the case with residual plots in Fig. 5 (all three). The curve in residual plot in Fig. 2 when X is between 15-35 is also not present in the other residual plots (Fig.5).

LINE conditions: Since there are three models (a), (b), (c), including log-log transformation, log(Y) transformation, and log(X) transformation (as fitted models given above), we will discuss the LINE conditions from all three models together below. Figures 6, 7, 8, 9, 10, and 11 will be presented after the discussion.

Model (a)

Testing Linearity Assumption: To test whether there is a linear relation between the log median house prices and log percentage of poverty in Boston we use Fig. 6. Plot a below shows a negative, non-linear relationship. There are no obvious outliers. Plot b below shows that the estimated value of log median house prices (using model a) agrees with the observed (sample) log median prices but there is a curve there too. Plot c below is the residual plot which shows no curve pattern. Plot d) is the Tukey's curve which shows some deviation from the flat line. Next, we do the Tukey's curve test to confirm if there is non-linearity:

Step 1: Null hypothesis (Ho): Linearity assumption holds
Alternative Hypothesis (Ha): Linearity assumption fails
Step 2: Set Level of significance α as 0.05 (default value)
Step 3: Test statistic = -3.215
Step 4: P-value = 0.00013
Step 5: Since P-value< α, so 5eject Ho
Conclusion: At 5% level of significance, linearity assumption fails.

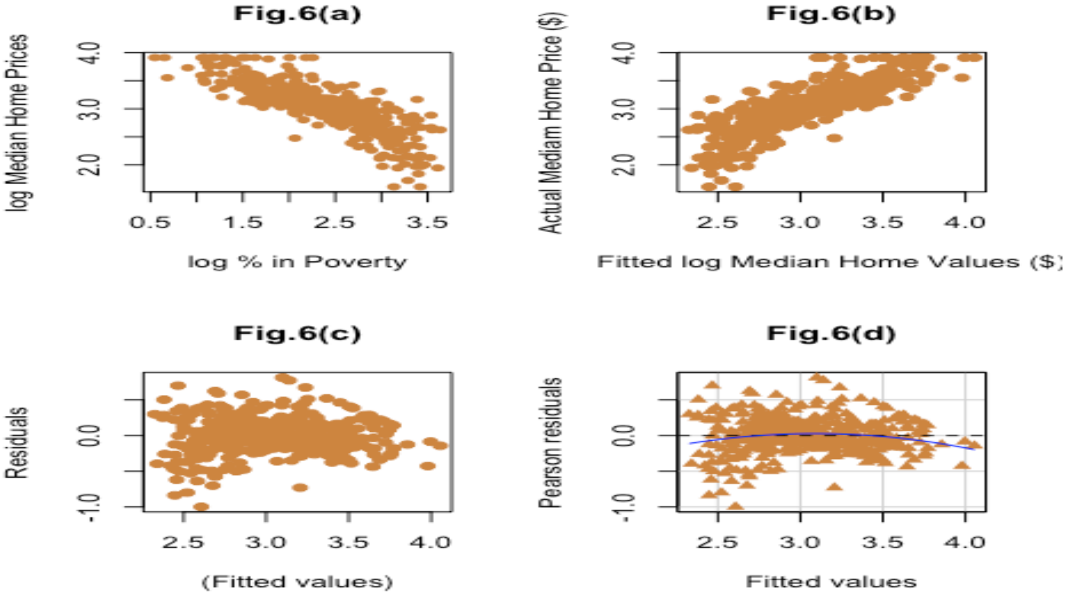


Fig.6(a)

Fig.6(b)

Fig.6(c)

Fig.6(d)

Independence of errors assumption

Fig.7(c) below is the index plot of errors and there is no obvious pattern. Hence the independence of errors assumption holds based on the plot.

Normality of errors assumption

Fig.7(a) shows histogram of errors which is bell shaped. Fig.7(b) shows the QQ plot, with agreement between the observed and estimate log median prices. Therefore, we can say that the normality of errors assumption holds.

Errors are homoscedastic assumption

In Fig.7(d), the residual plot with a mean of errors line. We can see some change in the spread of residuals, but it not clear if there are obvious signs of hetroscedasticity. Using the Breusch-Pagan test:

Step 1: H_0: Errors are homoscedasticity)
H_a: Errors are heteroscedasticity
Step 2: $\alpha = 0.05$ (default value)
Step 3: Test statistic = 16.316
Step 4: P-value = 0.000054
Step 5: Since P-value < α, so we reject the H0.

Conclusion: At 5% level of significance, we have enough evidence that errors are not homoscedastic.

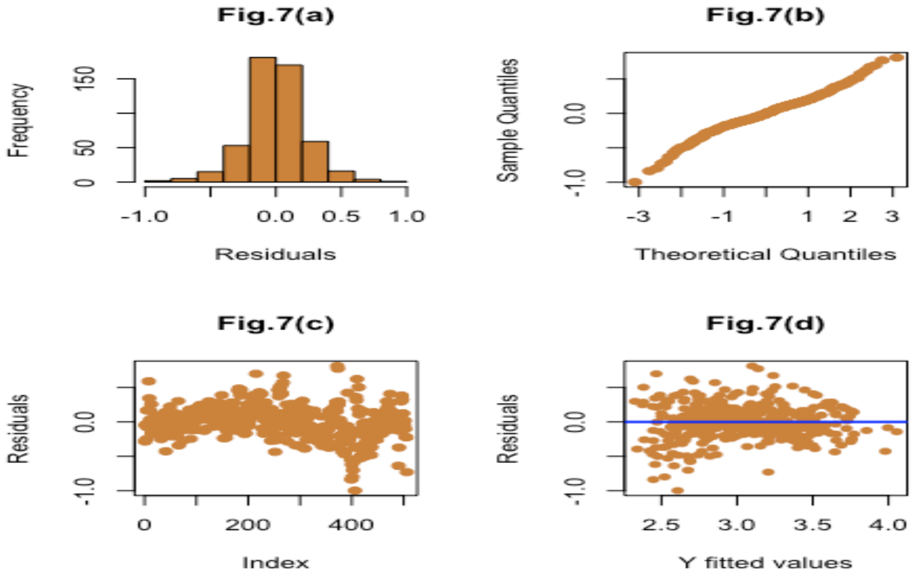


Fig.7(a)
Fig.7(b)
Fig.7(c)
Fig.7(d)

Model (b):

Testing Linearity Assumption: To test whether there is a linear relation between the log median house prices and percentage of poverty in Boston we use Fig. 8. Plot a below shows a negative, non-linear relationship. There are no obvious outliers. Plot b below shows that the estimated value of log median house prices (using model b) agrees with the observed (sample) log median prices but there is a curve there too. Plot c below is the residual plot which shows no pattern. Plot d is the Tukey's curve which shows some deviation from the flat line. Next, we do the Tukey's curve test to confirm if there is non-linearity:

Step 1: Null hypothesis (Ho): Linearity assumption holds
Alternative Hypothesis (Ha): Linearity assumption fails
Step 2: Set Level of significance α as 0.05 (default value)
Step 3: Test statistic = 7.074
Step 4: P-value = 0.00000000000015
Step 5: Since P-value< α, so 5eject Ho
Conclusion: At 5% level of significance, linearity assumption fails.

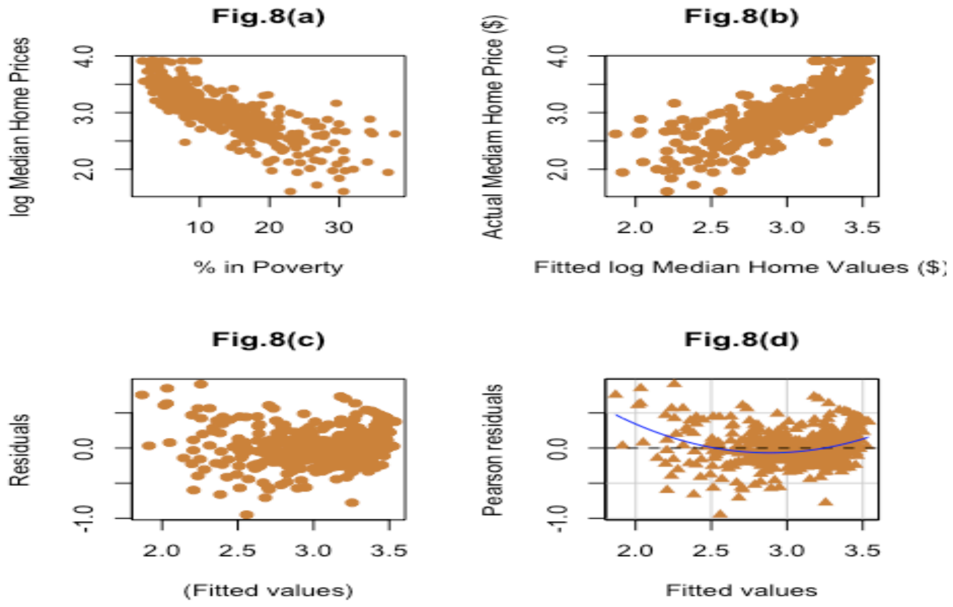


Fig.8(a)

Fig.8(b)

Fig.8(c)

Fig.8(d)

Independence of errors assumption

Fig.9(c) below is the index plot of errors and there is no obvious pattern. Hence the independence of errors assumption holds based on the plot.

Normality of errors assumption

Fig.9(a) shows histogram of errors which is bell shaped. Fig.9(b) shows the QQ plot, with agreement between the observed and estimate log median prices. Therefore, we can say that the normality of errors assumption holds.

Errors are homoscedastic assumption

In Fig.9(d), the residual plot with a mean of errors line. We can see some change in the spread of residuals, so there are signs of hetroscedasticity. Using the Breusch-Pagan test:

Step 1: H0: Errors are homoscedasticity)
Ha: Errors are heteroscedasticity

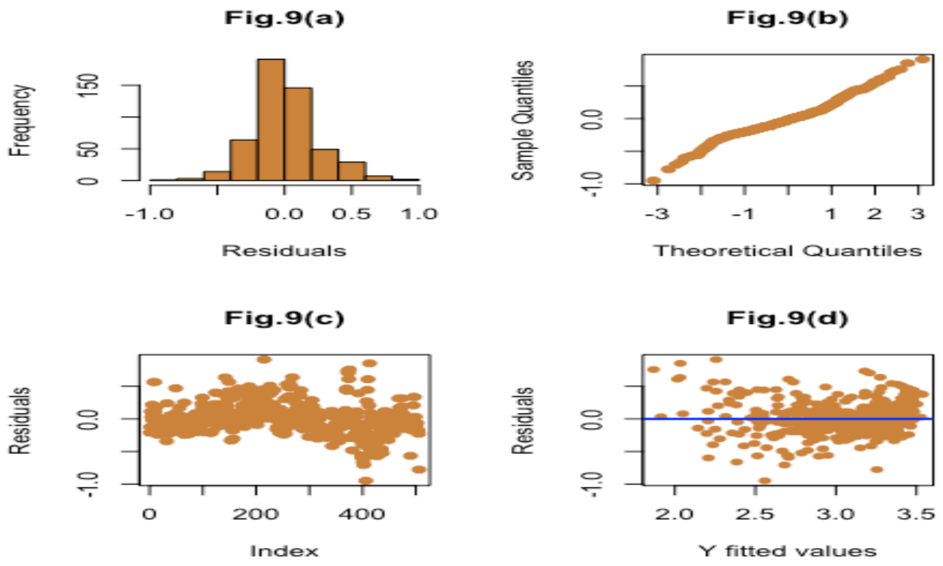
Step 2: $\alpha = 0.05$ (default value)

Step 3: Test statistic = 29.583

Step 4: P-value = 0.000000053

Step 5: Since P-value < α, so we reject the H0.

Conclusion: At 5% level of significance, we have enough evidence that errors are not homoscedastic.



Model (c)

Testing Linearity Assumption: To test whether there is a linear relation between the median house prices and log percentage of poverty in Boston we use Fig. 10. Plot a below shows a negative and linear relationship. There are no obvious outliers. Plot b below shows that the estimated value of median house prices (using model c) agrees with the observed (sample) median prices. Plot c below is the residual plot which shows no clear pattern. Plot d is the Tukey's curve which shows some deviation from the flat line. Next, we do the Tukey's curve test to confirm if there is non-linearity:

Step 1: Null hypothesis (Ho): Linearity assumption holds

Alternative Hypothesis (Ha): Linearity assumption fails

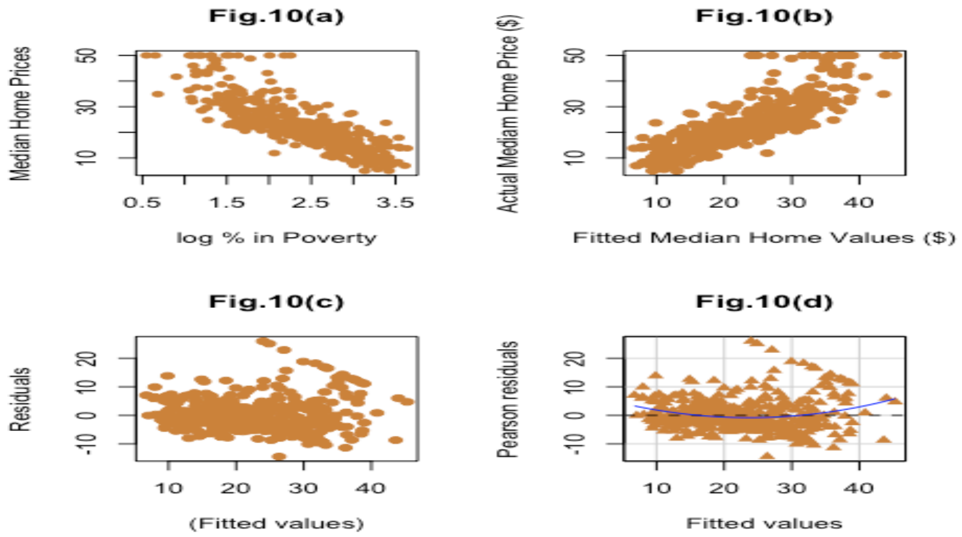
Step 2: Set Level of significance α as 0.05 (default value)

Step 3: Test statistic = 4.150

Step 4: P-value = 0.0000332

Step 5: Since P-value$< \alpha$, so 5eject Ho

Conclusion: At 5% level of significance, linearity assumption fails.



Independence of errors assumption

Fig.11(c) below is the index plot of errors and there is no obvious pattern. Hence the independence of errors assumption holds based on the plot.

Normality of errors assumption

Fig.11(a) shows histogram of errors which is right skewed. Fig.11(b) shows the QQ plot, with agreement between the observed and estimate median prices but there is a curve. Therefore, we can say that the normality of errors assumption fails.

Errors are homoscedastic assumption

In Fig.11(d), the residual plot with a mean of errors line. It is difficult to see signs of hetroscedasticity. Using the Breusch-Pagan test:

Step 1: H0: Errors are homoscedasticity)
Ha: Errors are heteroscedasticity

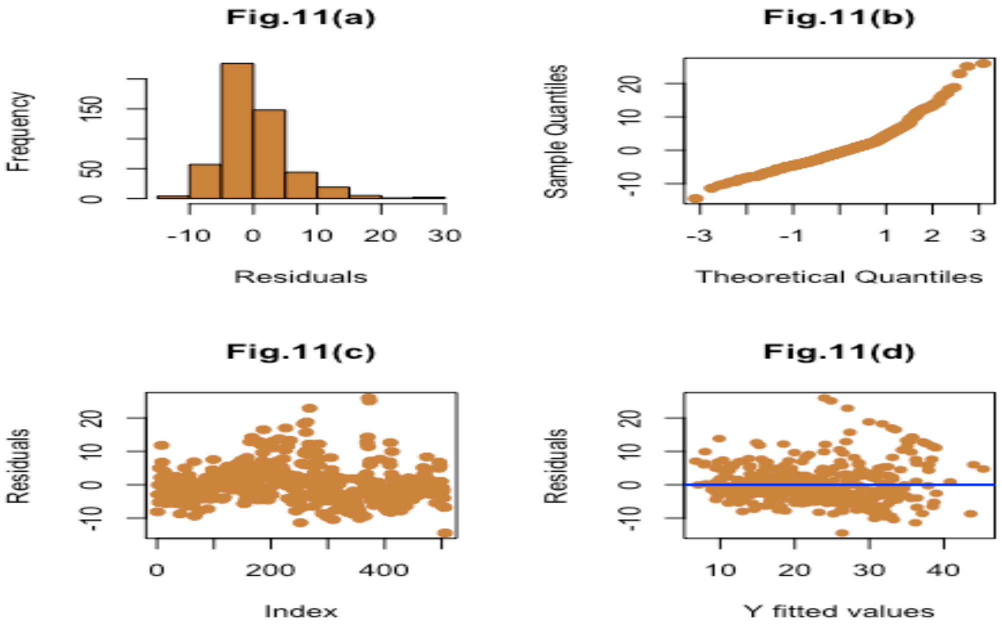
Step 2: $\alpha = 0.05$ (default value)

Step 3: Test statistic = 29.583

Step 4: P-value = 0.000000053

Step 5: Since P-value < α, so we reject the H0.

Conclusion: At 5% level of significance, we have enough evidence that errors are not homoscedastic.



The proposed transformations in the three models have not helped much with the assumptions as same assumptions failed in the three models.

g.) Polynomial model: 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 (using what you reported in e)? Did the transformation help with your analysis, why or why not?

Fitted Model:

$\hat{Y} = 42.86 - 2.33X + 0.04X^2$, where

X is the % in Poverty in Boston

Y is the Median home prices in Boston in 1000s of $.

Goodness of Fit:

i) $R^2 = 0.6407$ showing that approximately 64.1% change in median house prices in Boston is explained by the % of poverty and its square term.

ii) $\hat{\sigma} = 5.524$, meaning that on an average the fitted values of median house prices vary 5,524$ from the observed prices.

iii) Residual Plot: In Fig.12 below, we notice that the problem of linearity from Fig. 2 is addressed as the clear curve pattern is no longer present. The hetroscedasticity problem we observed in Fig.2 has also been addressed.

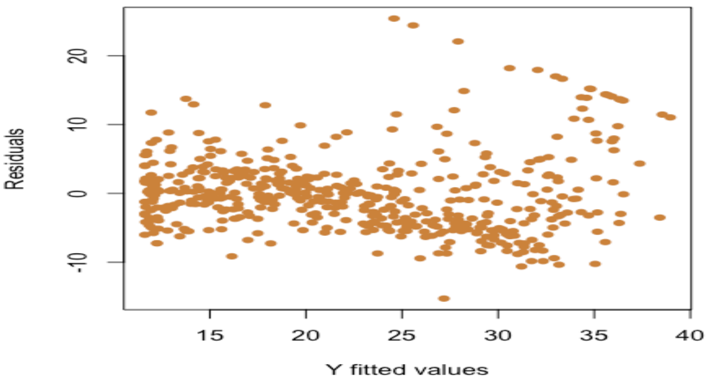


Fig12: Residual Plot

LINE conditions for the polynomial model

Testing Linearity Assumption: To test whether there is a linear relation between the median house prices and second order polynomial in percentage of poverty in Boston. In Fig. 13, Plot a below shows a negative, non-linear relationship. There is no obvious outlier (as explained in part a above).

It is important to point out that this plot can vary in polynomial settings, as some people might look at Y vs X (which we already when we started the assignment in part a, some might use just $Y\ vs\ X^2$ to see additional information that squared X brings, some might see $Y\ vs\ X + X^2$, and some might use $Y\ vs\hat{Y}$ (this is our plot b). It is ideal to see Y vs X, $Y\ vs\ X^2$ to understand relation each power of X has with Y.

Plot b below shows that the estimated value of median house prices (using model 1) agrees with the observed (sample) prices and the curve we observed earlier is fixed. Plot c below is the residual plot which shows no obvious pattern. Plot d) is the Tukey's curve which shows very small deviation from the flat line, indicating absence of non-linearity.

All plots show that linearity assumption is failing.

Next, we do the Tukey's curve test to confirm:

Step 1: Null hypothesis (Ho): Linearity assumption holds

Alternative Hypothesis (Ha): Linearity assumption fails

Step 2: Set Level of significance α as 0.05 (default value)

Step 3: Test statistic= 6.196

Step 4: P-value=0.0000000006 ~0

Step 5: Since P-value< α, so we reject Ho

Conclusion: At 5% level of significance, linearity assumption still fails.

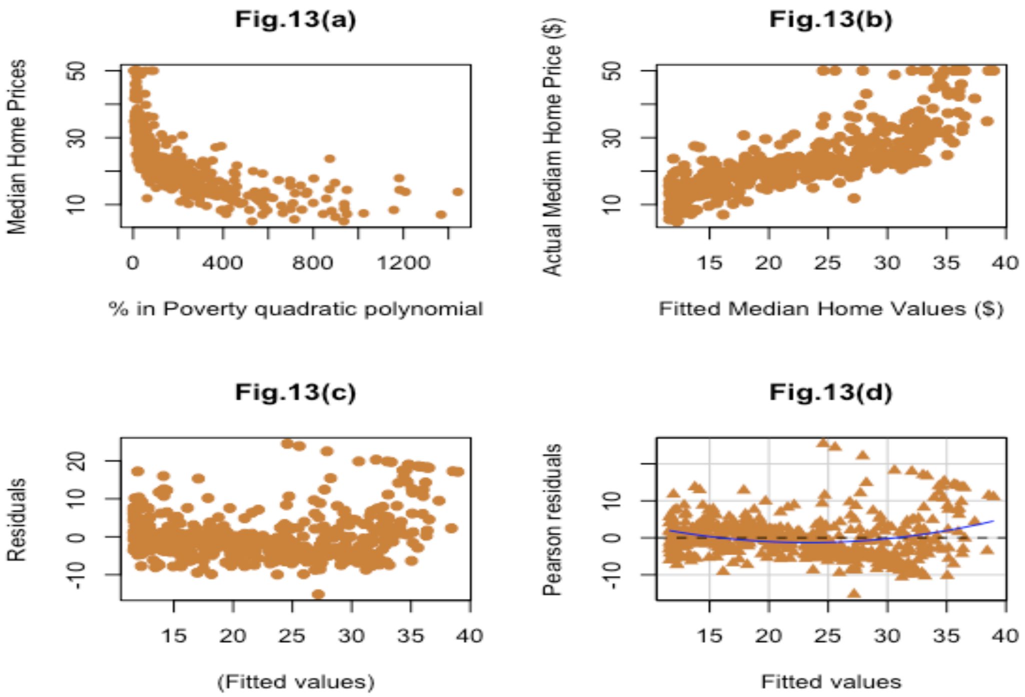


Fig.13(a)

Fig.13(b)

Fig.13(c)

Fig.13(d)

Independence of errors assumption

Fig.14(c) below is the index plot of errors and there is no obvious pattern. Hence the independence of errors assumption holds based on the plot.

Normality of errors assumption

Fig.14(a) shows histogram of errors which is bell shared but there is small right skew. Fig.14(b) shows the QQ plot, which shows an agreement in the sample and normal quantiles, but there is a clear curve in the plot. Therefore, we can say that the normality of errors assumption might fail.

Errors are homoscedastic assumption

In Fig.14(d), the residual plot with a mean of errors line. We can see that the variance is stable for most part except near the end. It is difficult to see whether the homoscedasticity holds or not just from this graph. Using the Breusch-Pagan test:

Step 1: H0: Errors are homoscedasticity)
Ha: Errors are heteroscedasticity

Step 2: $\alpha = 0.05$ (default value)

Step 3: Test statistic = 48.74

Step 4: P-value = 0.00000000003~0

Step 5: Since P-value < α, so we reject the H0.

Conclusion: At 5% level of significance, we have enough evidence that errors are not homoscedastic.

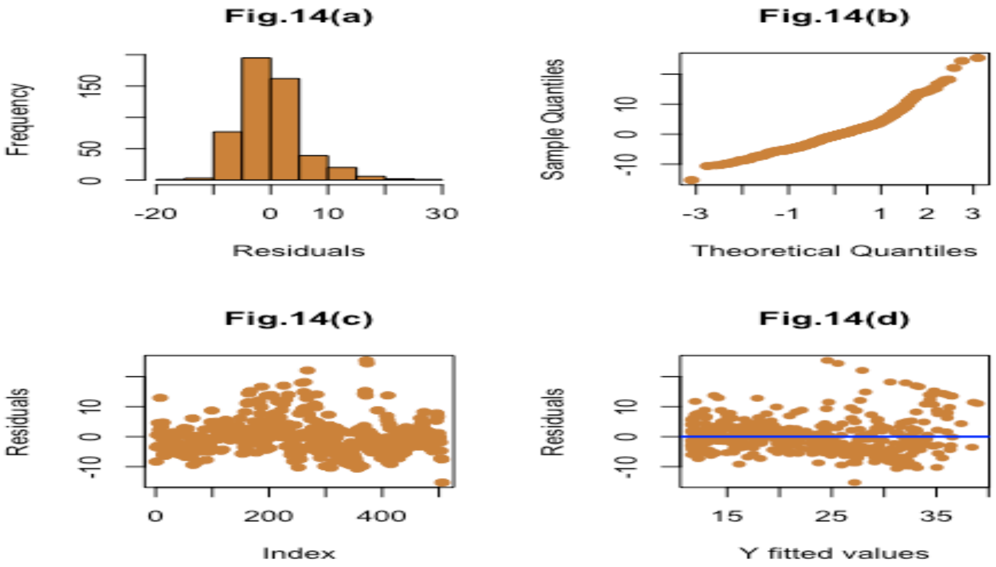


Fig.14(a)

Fig.14(b)

Fig.14(c)

Fig.14(d)

What changed in the residual plot reported in (e).

We are asked to compare Fig. 2 and 12, so bringing them below, we see that the spread in Fig. 2 points are more fewer/sparse in the beginning and there was a curve when fitted values are between 15-35. In Fig. 12, the points are not sparse in the beginning and the curve has been removed.

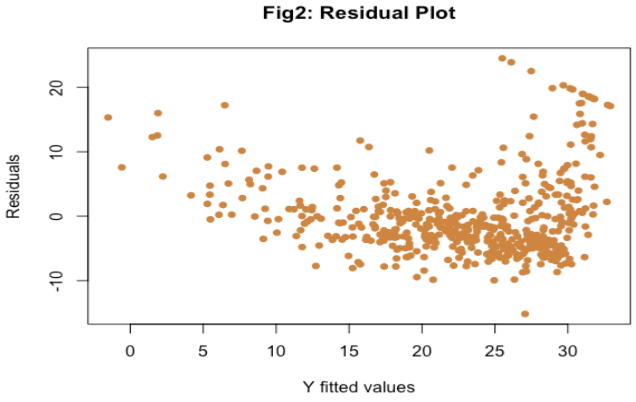


Fig2: Residual Plot

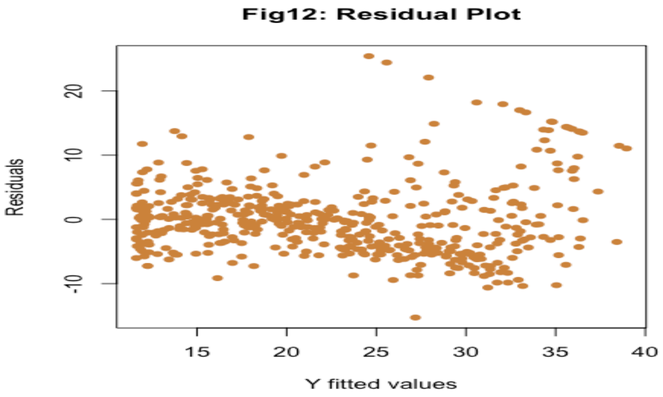


Fig12: Residual Plot

Did including the squared lstat term improved the model? Justify.

Yes, overall, this additional squared lstat term helped as coefficient of determination increased from 54.4% to 64.7% (whereas original model is nested in the polynomial model so this comparison might not be accurate) and residual standard error reduced. The residual plot from linear model had non-linearity, which got addressed to a good extent.