MG-UY 4204 Project 2 Regression Analysis Project

In this project, we, as a group, conducted linear regression analysis and calculations through software, to interpret those results in this report. For the data collection process we accessed this website: https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html?fbclid=IwAR0gT8VDyGCyXr9kXIqU_we0zmVx4unqTV-W3jOSg_R6VtLx_CaqGEhCe38, and obtained a data set containing five quantitative continuous variables which seem to be linearly related. Additionally, the data set has 506 observations/instances. We first identified one variable as a dependent variable (y) and others as independent variables (x).

I. Proposal

To reiterate, the data set we collected was regarding multivariate, time series type of data with 506 instances from the US census Service collection of housing data from the area of Boston, Massachusetts; although, these comparisons were primarily completed externally, and not on Delve, suspicions are raised on the validity of the data. This data particularly is used to benchmark algorithms used for multiple regression analysis. Regarding the dataset, we called it boston, and there were two goals of the data: to determine the nitrous oxide level of an area, as well as determine the median value of a home given several independent variables.

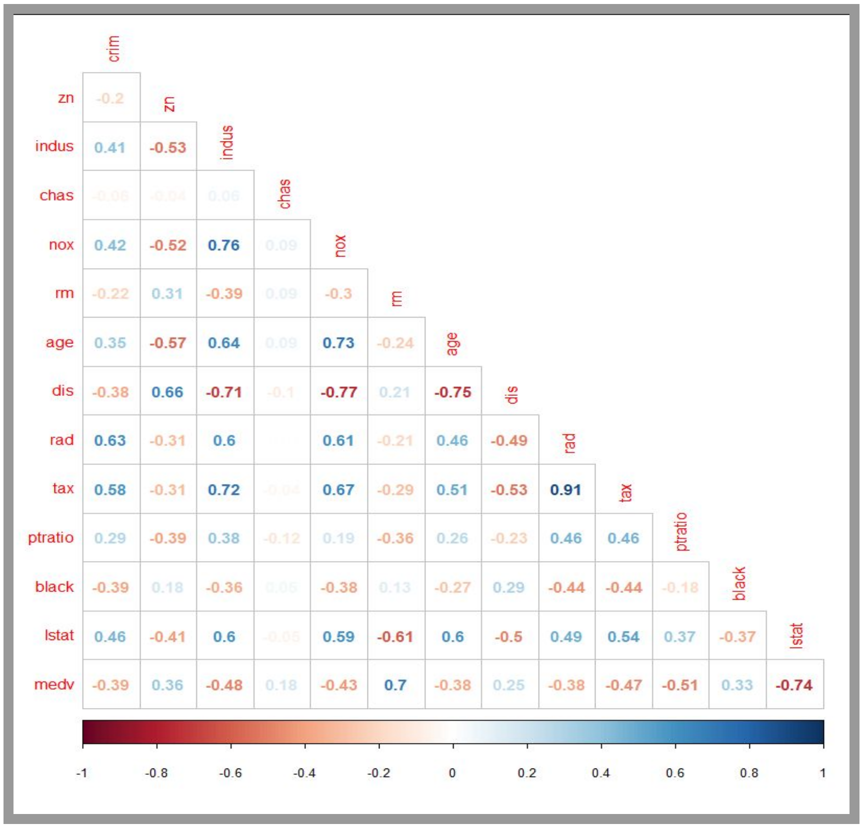
Some attributes/variables that are taken into account from the data are:

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX – nitrogen oxides concentration (parts per million).
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV – median value of owner-occupied homes in $1000s.

We will use this information, as well as Rstudio and Excel, to determine MEDV - Median value of owner-occupied homes in $1000's as our response variable. We chose not to classify NOX - nitric oxides concentration (parts per 10 million) to be our dependent variable, since to determine the NOX, the process is based more so on speculation, since it’s hard to definitively argue that the features (parameters) listed above can be used to determine NOX. When discussing MEDV, it was listed that it seems to be censored at a median price of $50,000, represented as 50.0. It is suggested that censoring of highest median price data has occurred, since exactly 16 cases report $50,000 median value, and another 15 cases result in prices between $40,000- $50,000, with prices rounded to the nearest hundred.

Our four most useful predictor variables are proportion of non-retail business acres per town (INDUS), average number of rooms per dwelling (RM), full-value property-tax rate per $10,000 (TAX), and % lower status of the population (LSTAT).

For analysis of our response-predictor and predictor-predictor pairwise correlations:

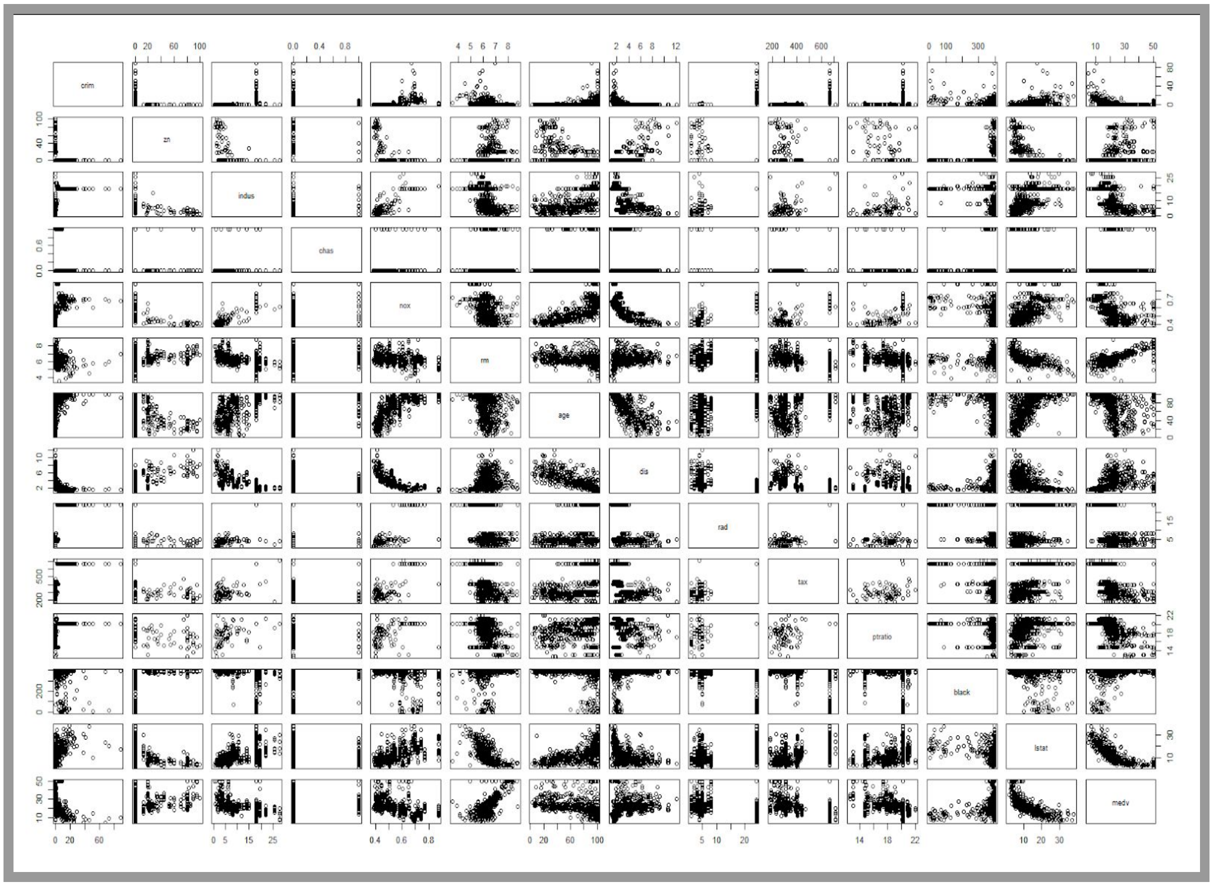


	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
zn	-0.2												
indus	0.41	-0.53											
chas	-0.06	-0.04	0.06										
nox	0.42	-0.52	0.76	0.09									
rm	-0.22	0.31	-0.39	0.09	-0.3								
age	0.35	-0.57	0.64	0.09	0.73	-0.24							
dis	-0.38	0.66	-0.71	-0.1	-0.77	0.21	-0.75						
rad	0.63	-0.31	0.6	[illegible]	0.61	-0.21	0.46	-0.49					
tax	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91				
ptratio	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46			
black	-0.39	0.18	-0.36	0.05	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18		
lstat	0.46	-0.41	0.6	-0.05	0.59	-0.61	0.6	-0.5	0.49	0.54	0.37	-0.37	
medv	-0.39	0.36	-0.48	0.18	-0.43	0.7	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74

Analysis: As stated before, our response variable is MEDV, and our predictor variables are INDUS, RM, TAX, and LSTAT. As seen in the response-predictor pairwise correlations, INDUS and MEDV have a -0.48, meaning that proportion of non-retail business acres per town is shown as a moderate, negative relationship with the median value of owner-occupied homes in $1000's. LSTAT and MEDV have a -0.74, meaning that % lower status of the population is shown as a strong negative relationship with the median value of owner-occupied homes in $1000's. For TAX and MEDV there is a -0.47, meaning that full-value property-tax rate per $10,000 is shown as a negative moderate relationship with the median value of owner-occupied homes in $1000's. Lastly, for RM and MEDV there is a -.70 meaning that average number of rooms per dwelling is shown as a strong negative relationship with the median value of owner-occupied homes in $1000's.

In terms of predictor-predictor pairwise correlations: LSTAT and INDUS have a 0.6, meaning that % lower status of the population is shown as a moderate positive relationship with the proportion of non-retail business acres per town. LSTAT and TAX have a .54, meaning that % lower status of the population is shown as a moderate positive relationship with the full-value property-tax rate per $10,000. LSTAT and RM have a -.61, meaning that % lower status of the population is shown as a moderate negative relationship with the average number of rooms per dwelling. TAX and INDUS have a 0.72, meaning that proportion of non-retail business acres per town is shown as a strong positive relationship with the full-value property-tax rate per $10,000. TAX and RM have a -0.29, meaning that full-value property-tax rate per $10,000 is shown as a weak negative relationship with the average number of rooms per dwelling. Lastly, INDUS and RM have a -0.39, meaning that proportion of non-retail business acres per town is shown as a weak negative relationship with the average number of rooms per dwelling.

Here is our matrix scatter plot of the 14 variables:



The reason why this is so important is because it shows the correlation, as well as the strength of the correlations between the response variables, and the overall trends that occur between the response variables and the predictor variables. To reiterate: INDUS and MEDV have a moderate, negative relationship. LSTAT and MEDV have a strong negative relationship. For TAX and MEDV there is a negative moderate relationship. Lastly, for RM and MEDV there is a strong negative relationship. In terms of predictor-predictor pairwise correlations: LSTAT and INDUS have a moderate positive relationship. LSTAT and TAX have a moderate positive relationship. LSTAT and RM have a moderate negative relationship. TAX and INDUS have a strong positive relationship. TAX and RM have a weak negative relationship. Additionally, INDUS and RM have a weak negative relationship.

The reason why modeling this data set would be meaningful is because it could help provide insight about the real estate value of owner-occupied homes in Boston, Massachusetts. In our analysis we measured the proportion of non-retail business acres per town (INDUS), average number of rooms per dwelling (RM), full-value property-tax rate per $10,000 (TAX), and % lower status of the population (LSTAT). This information lends us the understanding of how strong the correlation is between an increase or decrease in these values, and the increase in median household values. This could also bring about debate of values of housing depending on the types of housing such as condos, summer housing, apartment studios, timeshares, etc. This could also take into account how renovations can add value to the house, as well as further distinguish how house value changes if there is renting vs. ownership.

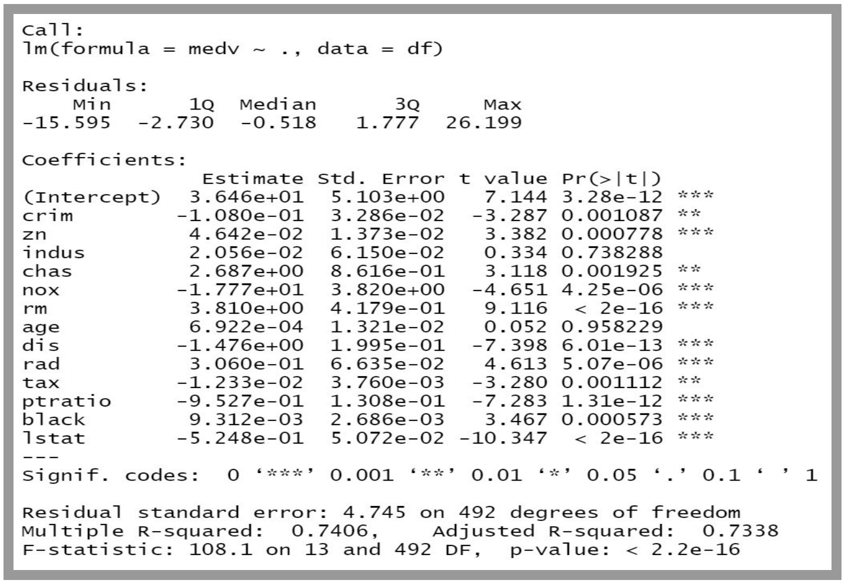
This brings about consideration of median income and accumulated wealth of citizens in the neighborhood, and how that correlates with housing prices. So does the proximity to large companies, taxation rates, real estate market inflation, and the zip code and/or neighborhood tabulation area the sample is taken. Lastly, you could determine the level of 311 calls (maybe about environmental or infrastructure issues) concentration as a determinant of household prices, while also understanding global pollution levels, and other climate change issues.

There are potential complications such as possible curvilinearity, multicollinearity, outliers, noise, skew, bias, and errors. This is very possible since this is multiple linear regression. In the cases of curvilinear data, this could mean that there are fluctuations in the data, causing a smooth curve, so it must be represented in polynomial form. In the case of multicollinearity, there might be two variables that have a similar linearity. With the case of outliers, this could be based on either errors when collecting data, the x-axis values are out of range, etc. Not to mention, an outlier could accidentally be mistaken for an extreme, significant point.

In the case of noise, this is any data or in fact, a parameter that either leads to an outlier in the data set, or useless, non-conclusive data. With skew, this could mean the data is heavily distributed around one side more than the other. In the case of bias, this could mean that the expected value differs highly from the actual data, making our residual values high. In the case of errors, these include bias, outliers, data-collection errors, etc. We aren't sure if the data has noise added, or if any errors there might have been made in the data collection process, including censoring of median household value. Also, this goes without saying, but the data was collected in 1978, so this data is slightly outdated. Potentially, pollution levels increases, tax rates changed, proportion of non-retail business acres per town increased, etc.

II. Preliminary Multiple Linear Regression Model Analysis

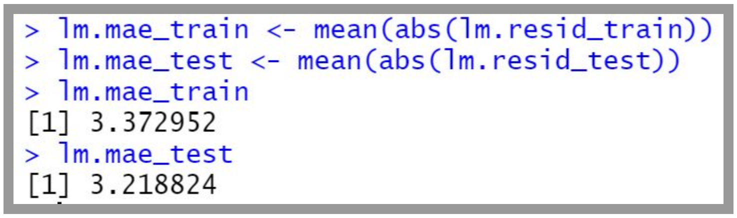
We initially partitioned our dataset into a training and testing set, with 80% of the data falling in the training set, and the remainder in the test. This is so that we can check the generalizability of our models on a held-out set. Our preliminary linear model (shown below) simply attempted to predict the Boston housing prices based upon all of the other features in the data set.



```
Call:
lm(formula = medv ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.595  -2.730  -0.518   1.777  26.199 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 ** 
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288    
chas         2.687e+00  8.616e-01   3.118 0.001925 ** 
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229    
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 ** 
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338 
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

When looking at the significance values in this summary table, we can see that most variables are significant with the exception of “INDUS” (proportion of non-retail business acres per town) and “AGE”. Furthermore, the “CRIM” (crime rate) and “TAX” variables are significant at the p < 0.01 threshold, while the rest of the features are significant at the p < 0.001 threshold. When adjusting for multiple hypothesis testing these two features may not pass a significance test, so we should be cautious about their impact on the final model. The adjusted R-squared produced -- 0.7338 -- is a good baseline to establish before further modeling. It’s important to use the adjusted R-squared, as it accounts for the number of features in the model.

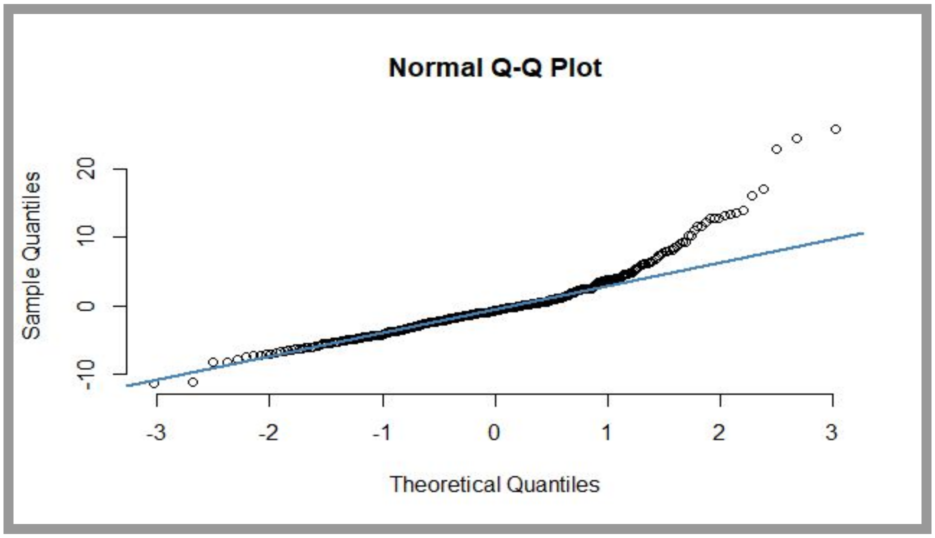
We also checked the model assumptions in a number of ways. Firstly, we took a look at the mean absolute error of the model, which is the average error yielded from the model’s predictions shown below:



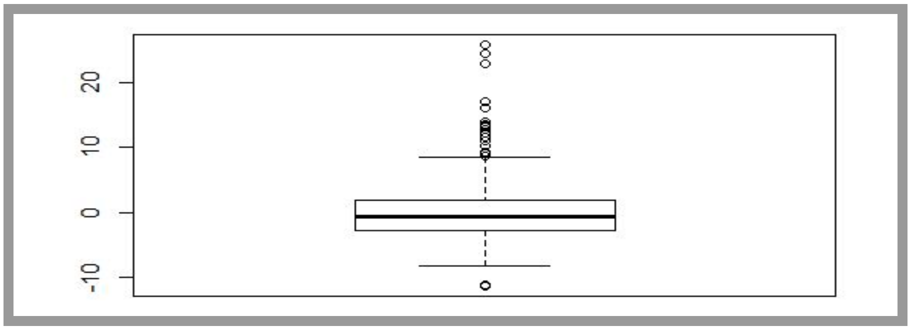
```
> lm.mae_train <- mean(abs(lm.resid_train))
> lm.mae_test <- mean(abs(lm.resid_test))
> lm.mae_train
[1] 3.372952
> lm.mae_test
[1] 3.218824
```

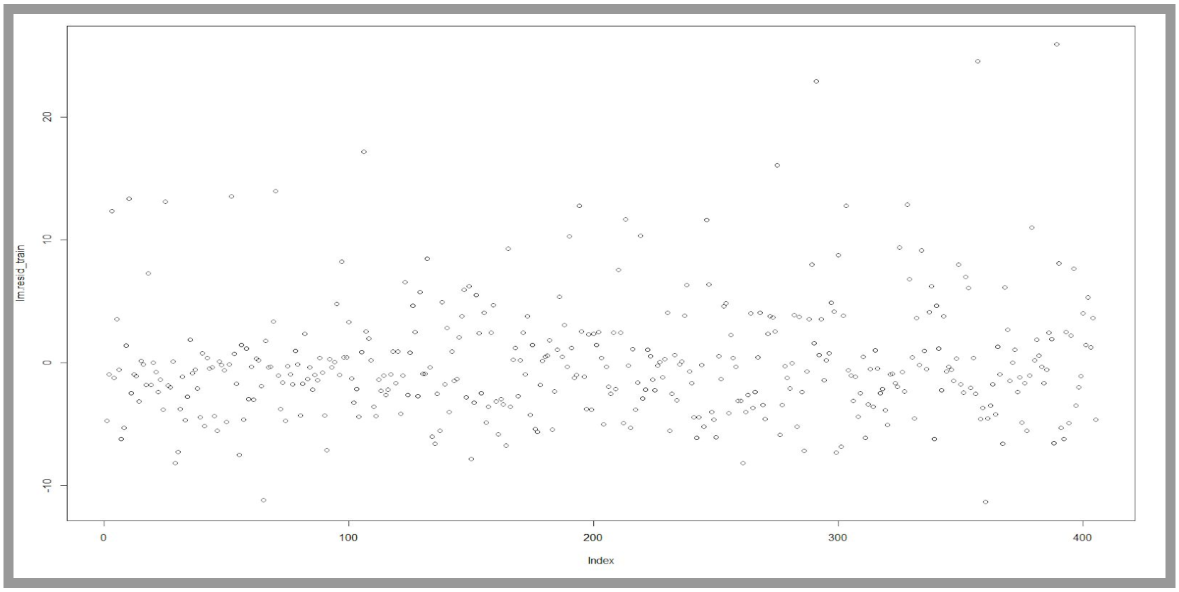
If the average error is smaller on the training set versus the test set, there is a good chance that the model is overfitting and not generalizing well on the data that it has not yet seen. In this case though, the train error -- 3.372952 -- and test error -- 3.218824 -- are very similar, which is a good sign.

We then created a QQ-Normal and QQ-LinePlot combined, as well as a barplot to not only show the trend when we test the theoretical quantiles vs sample quantiles.



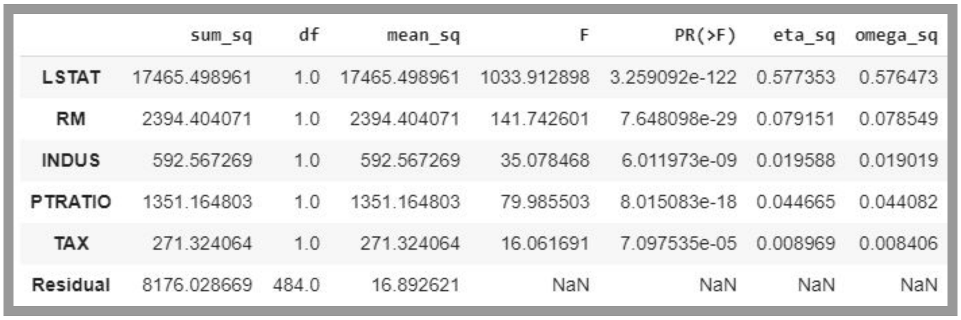
As shown in the Normal Q-Q plot, there is a short-tail on the left, but a long tail on the right, indicating a right skew. Next we take a look at the distribution of residuals to check the normality and constant variance assumptions shown below:





Assumingly, from all plots, the residuals seem to indicate that the variance is fixed across the residual distribution, and that the residuals are roughly normally distributed around a mean of 0. Though this visually shows a normal distribution, we will need to carry out a series of tests in order to confirm this normality, which we will do later on. In the box plot, there is more distribution around the 5 to 20 range of values.

Next, what we did was we conducted an Analysis of Variance (ANOVA) table:

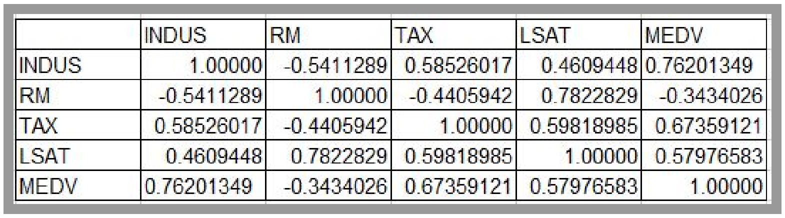


	sum_sq	df	mean_sq	F	PR(>F)	eta_sq	omega_sq
LSTAT	17465.498961	1.0	17465.498961	1033.912898	3.259092e-122	0.577353	0.576473
RM	2394.404071	1.0	2394.404071	141.742601	7.648098e-29	0.079151	0.078549
INDUS	592.567269	1.0	592.567269	35.078468	6.011973e-09	0.019588	0.019019
PTRATIO	1351.164803	1.0	1351.164803	79.985503	8.015083e-18	0.044665	0.044082
TAX	271.324064	1.0	271.324064	16.061691	7.097535e-05	0.008969	0.008406
Residual	8176.028669	484.0	16.892621	NaN	NaN	NaN	NaN

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

These values were tested with MEDV as the response variable. These predictors were chosen due to their pairwise correlation value. To be considered they needed to achieve a scoring of $|x| \geq 0.5$.

We also completed a Pearson Correlation test for linear correlation between variables:



	INDUS	RM	TAX	LSAT	MEDV
INDUS	1.00000	-0.5411289	0.58526017	0.4609448	0.76201349
RM	-0.5411289	1.00000	-0.4405942	0.7822829	-0.3434026
TAX	0.58526017	-0.4405942	1.00000	0.59818985	0.67359121
LSAT	0.4609448	0.7822829	0.59818985	1.00000	0.57976583
MEDV	0.76201349	-0.3434026	0.67359121	0.57976583	1.00000

The Correlation test also notably shows that there might be an identifiability problem due to the high correlation between the LSAT and RM variables. However, this is still ok because it won't hurt the model's predicatibility.

Confidence Interval and prediction interval, we calculated 95% CI (**α=95)** and PI for Normality Test: H0 = error variance is constant and H1 = error variance is not constant.

For our f-value test, we calculated the qf(), or the critical values of the normal distribution of the linear model. We found that since F*=85.96 > (F^{calc}=3.864637) , we reject H_0, which means the regression model is significant, and not reject H_1.

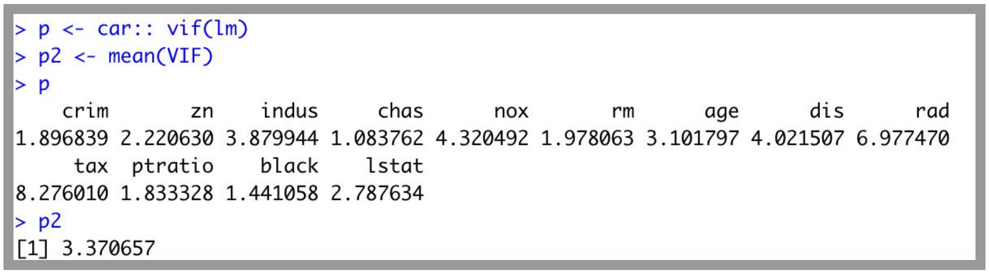
To find ρ^, for the Modified Levene Test, we divided the data into two sets by the median value. Our rationale is that if variance is different for the two sets, then error variance is non-existent. We then calculated the Pooled standard error as 4.91187 which is the error we encounter when we estimate variance of several different populations when the mean of each population may be different, but the variance of each population might be the same. Our ρ^ value is 0.80798 which means our mean of all sample proportions in the categories of interest.

We then took note of the R^2 (Coefficient of Determination): SSR/SST which explains the fraction of variability that is explained by the model, and mentions how well the regression model fits the data. In terms of 1-(SSE/SST), SSE/SST is the unexplained error. Our R^2 = .7406 which means that our correlation regression is strong positive linear regression.

We could then do adjusted R^2, by adjusting penalties with 1 - ((SSE/(n-p)) / (SST/(n-1))) which is: So if 0 <= Ra^2 <= R^2, Ra^2 and R^2 differ greatly, indicating one or more variable is not explaining much. However, Ra^2 can decrease when a variable is added, and R^2 never decreases when a variable is added. Our adjusted R^2 is 0.7338.

We then did a Regression significance test:
  H0: beta1 = beta2 = beta_p-1 =0 (reduced model: y_hat = beta_0 +epsilon_i)
  H1: at least beta_p-1 not zero
Which means some of our predictions might be helpful, but we cannot say accurately which one. Significance of Predictors: H0: beta_k =0. H1: beta_k != 0. if p < alpha, reject H0 if |t*| > t(1-(alpha/2), n-p) reject H0.

For our Mixed Linear Regression model, we calculated the Variance Inflection Factor(VIF), which is the number which is used to measure multilinearity, or how predictor variables affect each other. The mean VIF value is 3.370657, which is < 5, so we don't avoid the model, and VIF >/> (close to) 1 then there is not a serious multicollinearity between the variables.



```
> p <- car:: vif(lm)
> p2 <- mean(VIF)
> p
    crim       zn    indus     chas      nox       rm      age      dis      rad 
1.896839 2.220630 3.879944 1.083762 4.320492 1.978063 3.101797 4.021507 6.977470 
     tax  ptratio    black    lstat 
8.276010 1.833328 1.441058 2.787634 
> p2
[1] 3.370657
```

According to the first rule if VIF_k > 5 then we should avoid the model (which in our case is true because RAD and TAX are both greater than 5) but if we take the mean of VIF we see that 3.370657 is not >> 1. (much greater than 1) so we should use the model. This does also affirm that we should use INDUS, RM, and LSTAT for our predictor variables.

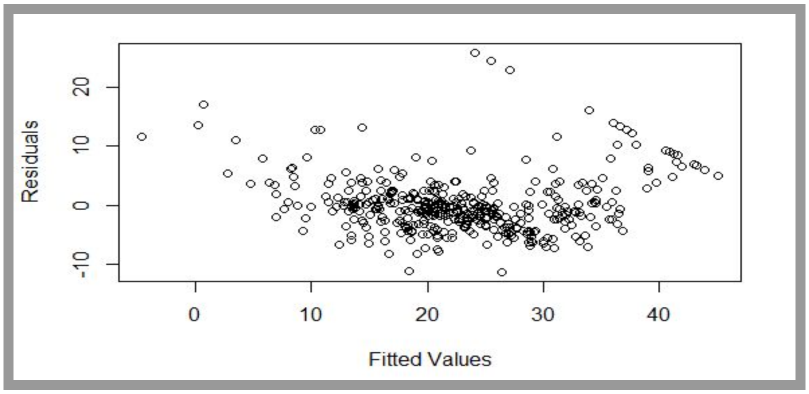
For analysis of our outliers, we calculated diagonal elements of hat_matrix (leverage values). We found that an x-outliers exists if hii > (2p)/n where p-1 is # of independent variables, and n = number of observations. For x-outliers we found that there are 33 outliers based on the equation above. This might mean that within the training set of 405 observations, these data points are not only extreme points that diverge in a big way from the overall trend due to variability in the measurement or perhaps experimental error. X-outliers in particular indicate far x-axis range.

To find the Y-outliers, we calculated the studentized deleted residuals, or t_i = (d_i) / (sqrt(MSE_i)(1-hii) where d_i = deleted residual where MSE: Mean Squared Error for the model without i.

For the Bonferoni Outlier test, we considered H_0 = observation is y-outlier and H_1 = observation is not y-outlier. If |t_i| > t(alpha/(2n)); n-p-1) then we can conclude that the observation is y-outlier which means we fail to reject H_0, and it has an influence on the regression. For y-outliers we found that there were 3 outliers based on the equations above.Our t-cutoff was 3.878. To determine the influence of the outliers, we used the Difference fitted value(DFFITS). We let y_hat_i(i) be the ith fitted value computed with observation i omitted. So, if (DFFITS = (y_hat_i - y_hat_i(i)) / (sqrt(MSE_i * hii)) or |ti*(sqrt(hii / (1-tii))) > 2*sqrt(p/n), then we flag that point [this is for n>=30]. Y_i is used to predict each corresponding $\widehat{Y}_i$ value. The value of this contribution, $h_{i,i}$, can be measured, and is used to determine the influence of individual points. It also is referred to as the leverage of the *ith* element. The leverage value always satisfies 0< $h_{i,i}$ <1. A large leverage (close to 1) suggests that the point is farther from the center of X elements, and therefore will have a larger influence on the accuracy of the regression and the specific parameters. Using Excel, we found 23 observations of flagged Difference fitted values.

We then measured Cooks Distance, the measure of the combined impact of observation i on all the LSE (Least Squares Estimation), so if D_i = e_i^2/ (p * MSE) > F(0.5,p,n-p), then obs_n is considered influential. Lastly, we combined diagonal elements of hat matrix, studentized deleted residuals, difference fitted values, and Cooks Distance into a dataframe, then exported the data.

Here is our residual vs y_hat plot, which shows the concentration of residuals of fitted values. According to the chart, there seems to be some outliers, but those might be significant points.



After flagging the values of studentized deleted residuals, difference fitted values, and Cooks Distance using Excel, we removed the values. When we flagged the difference fitted values, we found some 1's, which indicate an outlier, however, when flagging the Cooks Distance, at first we only got 0's as a result. To better find outliers, we modified our equation using ɑ/(2n). Upon conducting the test again, we came to the conclusion that the observations being flagged can possibly be really close to the boundary that we used and for that reason the observations are not influential.

III. Final Multiple Linear Regression Model

Here is our final model:

$\widehat{Y} =$ 31.812644 - 0.093538x1 + 0.040817x2 + 0.024657x3 - 14.686620x4 + 4.095199x5 - 0.005811x6 - 0.005811x7 - 1.438457x8 + 0.322527x9 - 0.013215x10 - 0.865237x11 + 0.010713x12 - 0.550524x13

For Confidence Interval (C.I) calculated at one xh of interest:

```
      fit          lwr          upr
1 7.328765   3.114705   11.54282
```

Based on our regression model we are 95% confident that if the per capita crime rate by town is **20**, the proportion of residential land zoned for lots over 25,000 sq.ft. is **15**, the proportion of non-retail business acres per town is **18**, the tract doesn't bound the river (**0**), the nitric oxides concentration is **.5** (parts per 10 million, the average number of rooms per dwelling is **5**, proportion of owner-occupied units built prior to 1940 is **40**, the weighted distances to five Boston employment centres is **8**, the index of accessibility to radial highways is **12**, the full-value property-tax rate per $10,000 is **215**, the pupil-teacher ratio by town is **14**, the proportion (1000(Bk – 0.63)^2) of blacks by town is **100**, and the lower status of the population is **22**% that the median value of owner-occupied homes in $1000's is between 3.114705 ($3,114.71) and 11.54282($11,542.82).

For Prediction Interval (P.I) calculated at one xh of interest:

```
      fit           lwr            upr
1 7.328765    -3.20762     17.86515
```

Based also on the criteria above with 95% confidence we can predict that the median value of owner-occupied homes in $1000's will lie between -3.20762 ($-3,207.62) and 17.86515

($17,865.15). The prediction interval contains a negative value and has a larger range because it takes into consideration the error variance in the model.

Confidence Interval and prediction interval, we calculated 95% CI and PI for for Normality Test: H0 = error variance is constant and H1 = error variance is not constant. For the value of qf(.05,14,391) we got 0.9544389.

IV. Final Conclusion

In our project, we decided to use thePareto Principle method of counting 80% of the data of 506 observations as our training set, but the remaining 20% as our testing set. Although stated in the instructions that we only needed 100 observations to develop our model, we decided to use the method stated in class, to maximize our efficiency of the model. We hope this can be used as a benchmark for future MLR projects, especially model accuracy and algorithm performance.

Based on the model and analysis that was done, we found that the higher housing prices will be attributed to areas with low crime rates and low pupil-teacher ratios. Additionally, housing prices are high if the houses are closer to the Charles river. Given the data, we have found that the lower the distance from the employment centers, the higher the potential to encounter air pollution of Nitrogen Oxide, and other greenhouse gases, or potentially even noise pollution. However, according to our data, a person might sacrifice living close to their employment center, if they knew the pollution levels were lower. Overall, this proves that hedonic pricing (the monetary value given to the house based upon surrounding areas) is more important than focusing on the house itself.

Some variables that we could have worked with included CRIM (crime rate in the area) and the ZN (proportion of residential land zoned for lots over 25,000 sq.ft), since those could be good determinants within estimating housing prices. We should also take into account zip code and/or neighborhood tabulation areas, distance from public parks and recreational centers, inflation rate of housing prices, etc. We would also want to consider the property transfers, deeds of ownership and tax assessments, renovations, depreciations, and vacancies. We might also want to use the number of floors of houses, number of balconies and whether there is a backyard, pool, garden, etc. Lastly, we might want to have a binary variable for whether or not the homeowners have insurance, whether they pay rent on time, or the overall mortgage.

We *Mark Holtje, Ian George, Graciela Casanova, Apoorva Nori, Yash Tekwani* did not give or receive any assistance on this project and the report submitted is wholly by us.”

Sources:

D. Lucas (ddlucas .at. alum.mit.edu), Lawrence Livermore National Laboratory.

http://lib.stat.cmu.edu/datasets/boston

Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978

https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-multiple-regression/mlr-with-interactions.htBelsley D.A., Kuh, E. and Welsch, R.E. (1980)Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Faraway, J. Linear Models with R. Chapman & Hall/CRC Texts in Statistical Science (Book 63).