4/27/2017

Beer Sales

Team 26

**Dhuri Sayli**

**Jariwala Ishani**

**Joshi Aditya**

**Bhadauria Poornima**

**Pokale Chetan**

Contents

[**Linear Regression Analysis** 2](#_Toc481067823)

[**Significant Variables** 3](#_Toc481067824)

[**Non-Significant Variables** 3](#_Toc481067825)

[**Model Re-Run** 3](#_Toc481067826)

[**Test of regression assumptions** 4](#_Toc481067827)

[**Linearity** 4](#_Toc481067828)

[**Correlation of X variables** 5](#_Toc481067829)

[**Heteroscedasticity** 7](#_Toc481067830)

[**Error terms correlation** 8](#_Toc481067831)

[**Outliers** 8](#_Toc481067832)

[**Regression after correction** 9](#_Toc481067833)

[**Linearity** 9](#_Toc481067834)

[**Multicollinearity** 9](#_Toc481067835)

[**Heteroscedasticity** 9](#_Toc481067836)

[**Serial correlation** 9](#_Toc481067837)

[**Outliers** 9](#_Toc481067838)

[**3D Graph using R** 10](#_Toc481067839)

[**Geographic Representations** 11](#_Toc481067840)

[**Business Analysis** 13](#_Toc481067841)

[**Other Factors** 14](#_Toc481067842)

[**Excel File** 14](#_Toc481067843)

# **Linear Regression Analysis**

We performed linear regression analysis using Weekly Averages Sales Volume of beer as the dependent variable Y and the following as our explanatory variables(X):

1. Age9
2. Age60
3. % College Graduates
4. % With No Vehicles
5. Exponential of Log of Median Income
6. Average Household Size
7. % Working Women with full-time jobs
8. Mean Household Value (Approximated)
9. % of Singles
10. % of Retired
11. % of Unemployed
12. % of working women with children
13. % of non-working women with children
14. % of households with mortgages
15. % of population that is non-white
16. % of population with income under $15,000

All the variables do not have a significant effect on the dependent variable.

## **Significant Variables**

Based on higher co-efficient value (both positive and negative) and lower p-value we have shortlisted the below significant variables

1. Age60
2. % of Unemployed
3. % With No Vehicles
4. % of population with income under $15,000
5. % of working women with children
6. % of Retired
7. % Working Women with full-time jobs

## **Non-Significant Variables**

We have dropped the below variables as we feel they are non-significant because they have lower co-efficient value in our regression analysis and have minimum impact on the dependent variable.

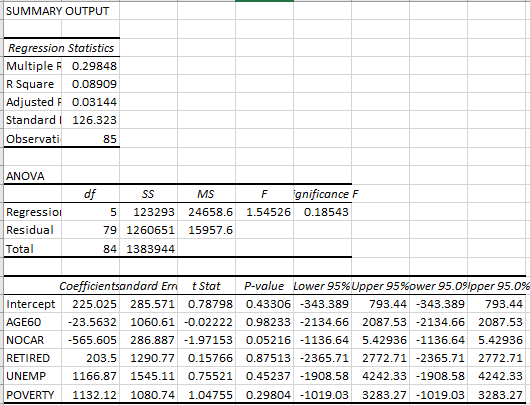
1. % of households with mortgages
2. Mean Household Value (Approximated)
3. % of Singles
4. % of population that is non-white
5. % of non-working women with children
6. % College Graduates
7. Age9
8. Average Household Size
9. Exponential of Log of Median Income

## **Model Re-Run**

We reran the model using five of the below significant variables.

1. Age60
2. % With No Vehicles
3. % of Retired
4. % of Unemployed
5. % of population with income under $15,000

We get an adjusted R square value of 0.03 which is very low and higher p-values as seen below indicating that model might not be linear.



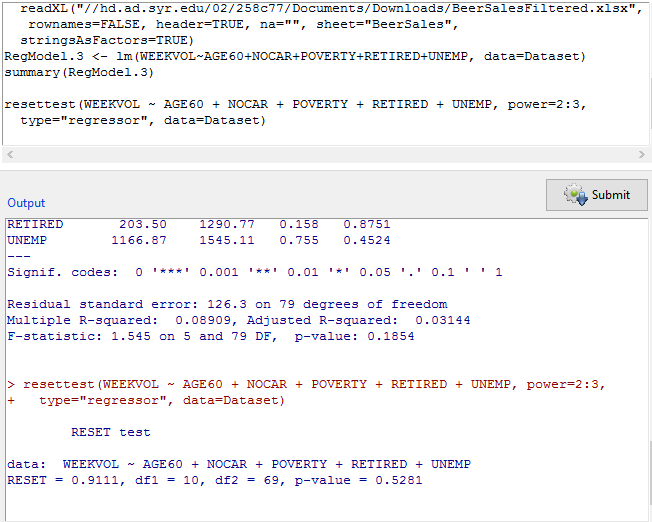
# 

# **Test of regression assumptions**

For performing linear regression there are certain assumptions that should hold true.

## **Linearity**

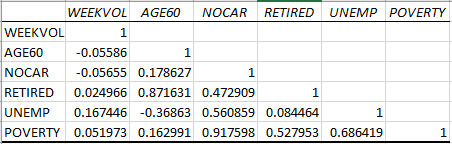
To perform regression analysis, the relationship between dependent and explanatory variables must be linear which is true in our case. This can be seen by performing Ramsey Regression Equation Specification Error Test (RESET) (1969) to test for linearity:



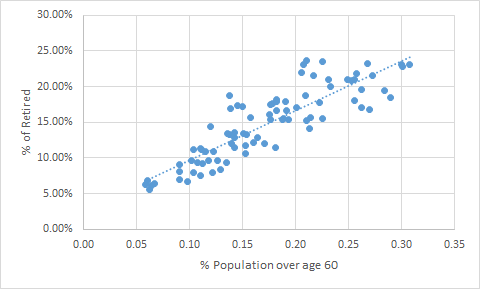
From the above tests it can be seen that p-value is 0.5281 which is greater than 0.05 and there is no linearity problem. So this assumption is not violated.

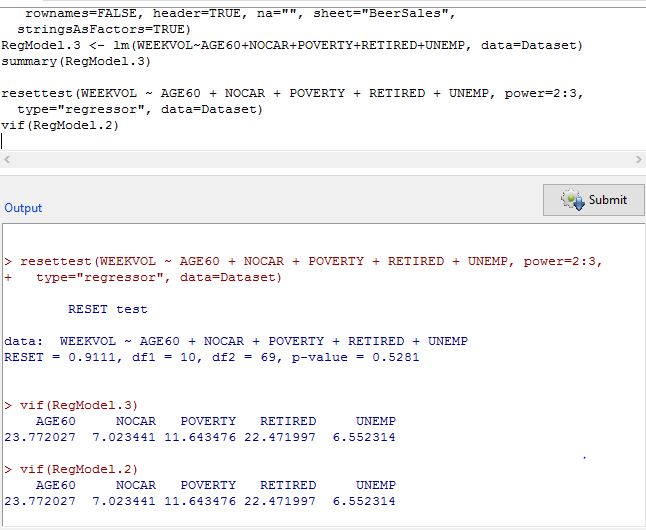
## **Correlation of X variables**

X variables or the explanatory variables should not be correlated with each other. However there is a relationship between some of our X variables and multi-collinearity exists as seen from the below plots and Variance Inflation Factor test of correlated explanatory variables tests.





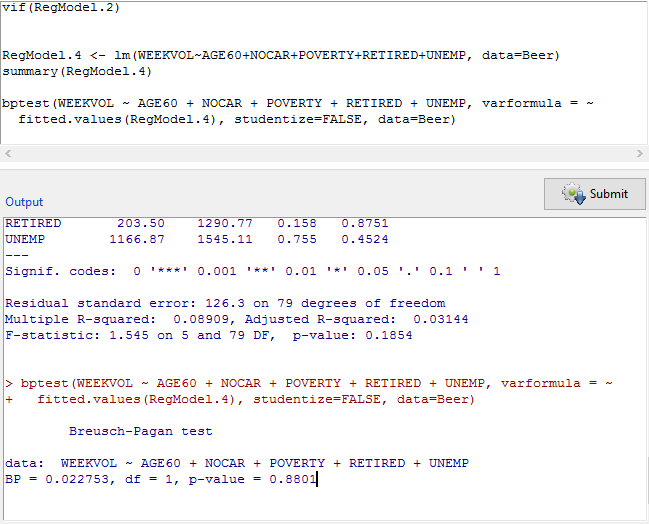




From the above tests it can be seen that variation influence factors are more than 10 for three variables Age60, Poverty and Retired which proves that multi collinearity exists and this assumption is violated.

## **Heteroscedasticity**

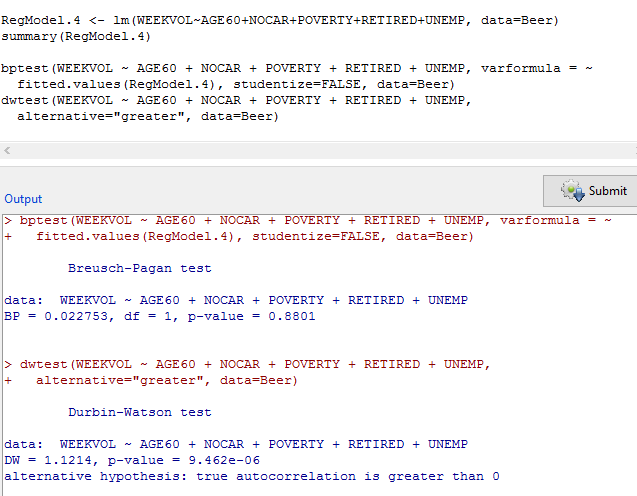
The error term must have constant variance over a range of X values. In our case the size of error term does not depend on any explanatory variable and there is no heteroscedasticity as seen from the below Breusch-Pagan test below:



As the p-value is 0.8801 which is more than 0.05, there is no problem with heteroscedasticity and the assumption is not violated.

## **Error terms correlation**

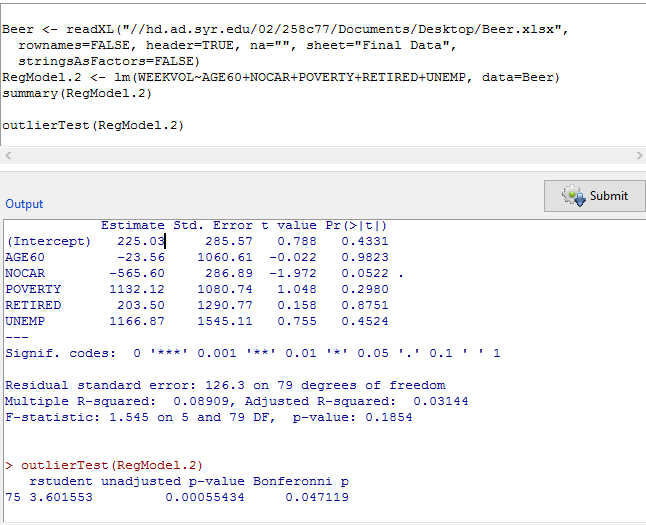
Error terms must not be correlated over the time period. For our regression analysis there is a problem with serial correlation as seen from the Durbin-Watson test below because p-value is 9.462e-06 which is less than 0.05:



Hence this assumption is violated.

## **Outliers**

Ideally there should not be any outliers for performing linear regression. However outliers exists in our model which can be seen in the linear scatter plots above and can be proved from the Bonferroni outlier test below.



It can be seen that there is one outlier which is row number 75 and this assumption is violated.

# **Regression after correction**

Some corrections have been made for the assumptions that were violated.

## **Linearity**

As there was no linearity problem and the assumption was not violated Box-Cox was not performed and there is no need for transformation of Y.

## **Multicollinearity**

As multicollinearity existed between three variables Age60, Poverty and Retired, average was taken for these three and was converted into single variable named AGE\_POV\_RET as seen in the Rerun Regression tab in the excel.

## **Heteroscedasticity**

Assumption was not violated and hence no correction was performed.

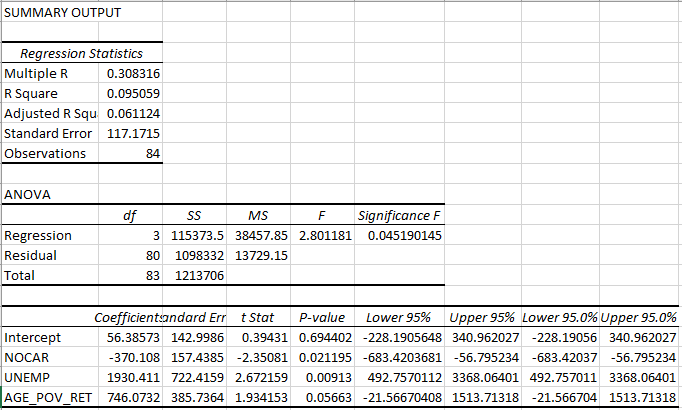
## **Serial correlation**

Though the assumption was violated, it is out of scope for this assignment.

## **Outliers**

One outlier was detected as per Bonferroni outlier test which was the record with weekly sales volume as 875. That particular record has been dropped.

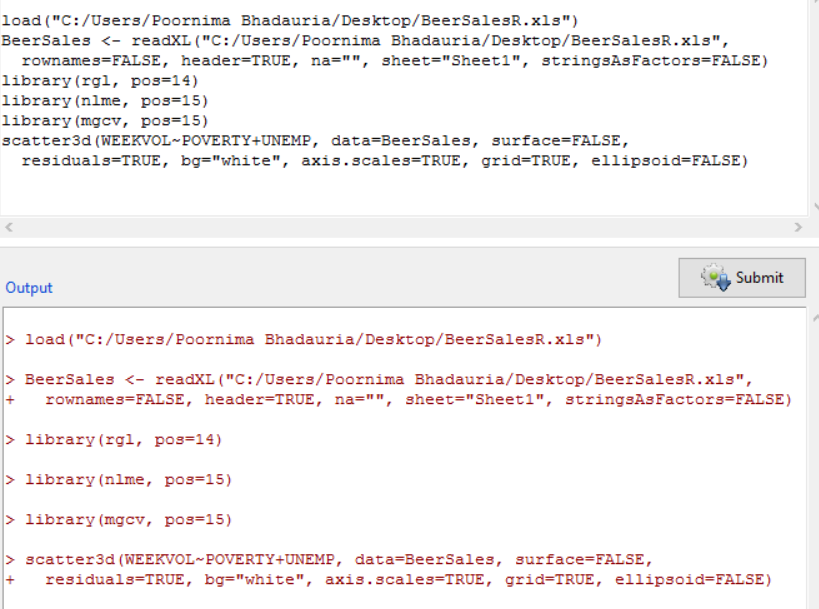
After making the above corrections, linear regression was rerun and the following output was obtained.

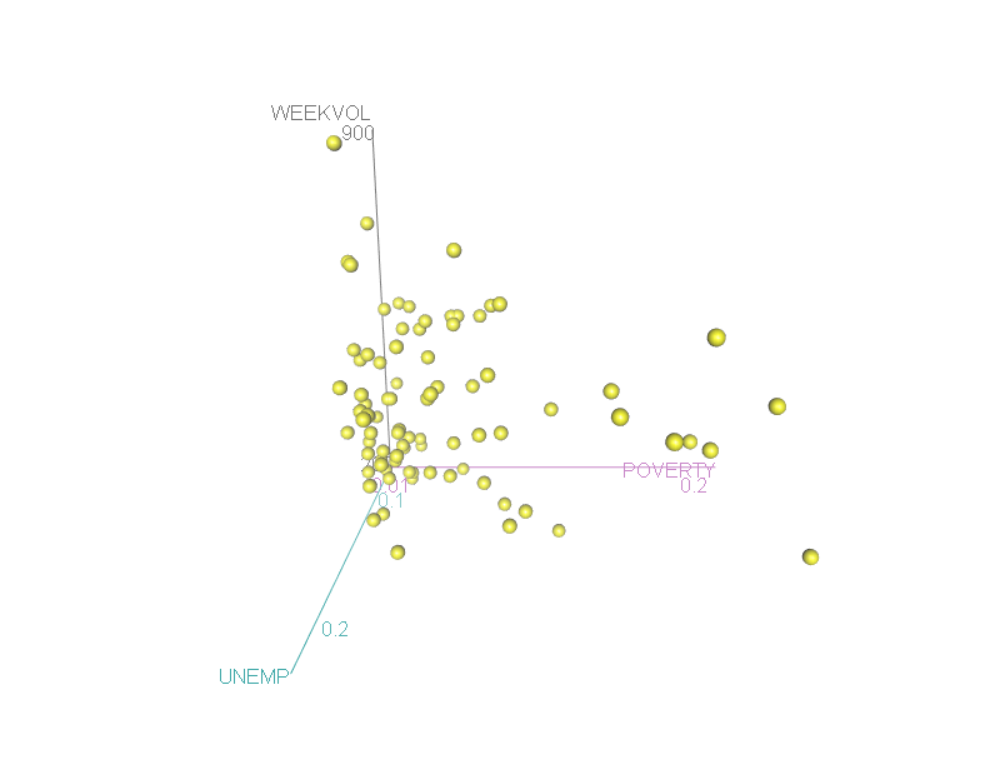


Adjusted R square improved slightly but is still on the lower side because of the small size of the data and real world fluctuations.

# 

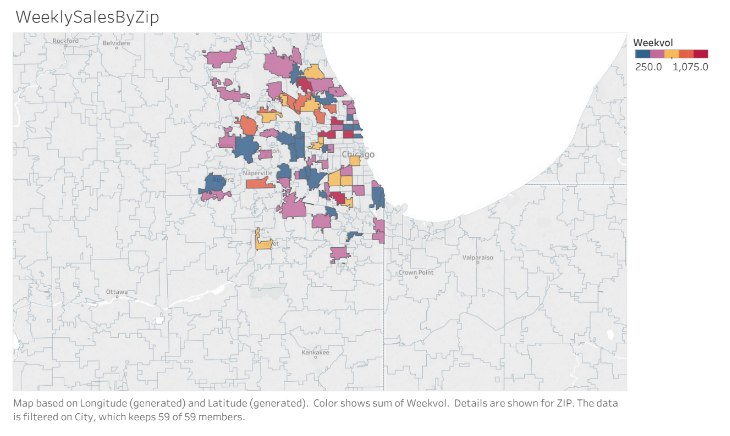
# **3D Graph using R**



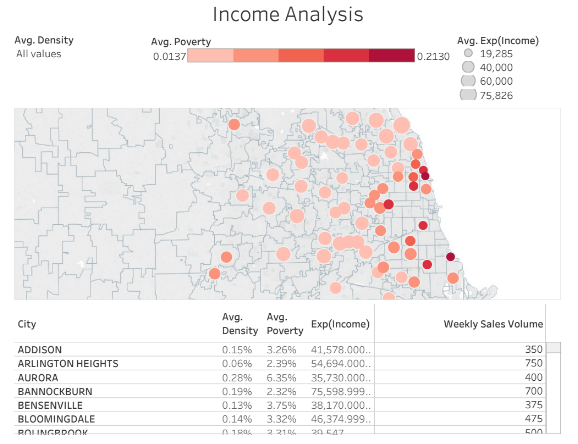


# **Geographic Representations**

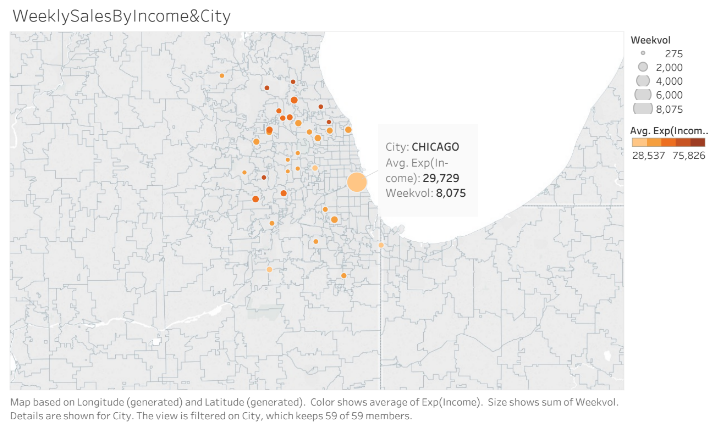
**Plot 1**: This plot describes weekly average sales volume by location. The highest sales is highlighted by red and the lowest sales is highlighted by blue as seen in the map



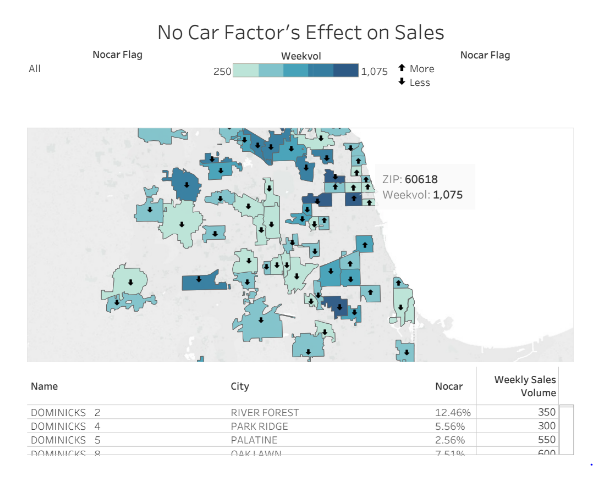
**Plot 2:** This plot tries to analyze how Income, Density and Poverty affects Weekly sales of stores. As seen the region “Aurora” has higher average Poverty and lower average income, hence the sales is low in that region.



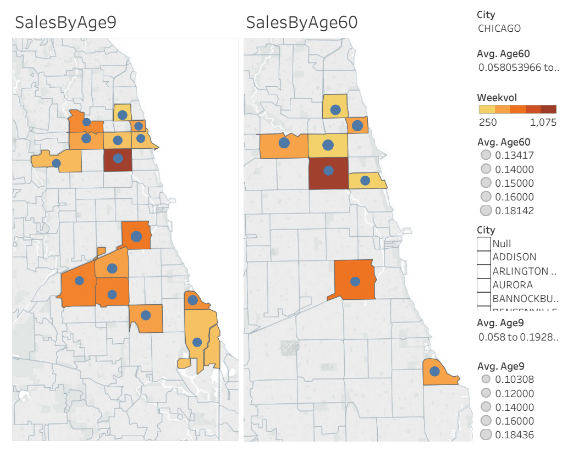
**Plot 3:** This plot shows weekly sales by Income and city. As highlighted, Chicago has highest sales with low Income.



**Plot 4**: This plot shows how No Car factor affects the sales per each store in particular region. As highlighted, River Forest has less sales, as there are only 12.46% people who don’t have car. At the same time, Park Ridge has almost the same sales as River Forest, but it has low percentage of people having no car. Probably the location of store Dominick’s 4 is located in the highly populated region and the store is within the walking distance.



**Plot 5:** This plot shows the comparison of beer bought by Age groups 9 and 60.



# **Business Analysis**

Looking at analytical patterns from a business perspective, we arrive at the following conclusions

* People with vehicles lead to a higher volume of sales

People with vehicles have an easy and convenient access to the beer stores thereby increasing the sales.

* People with income lower than 15000USD tend to make a bulk of the sales

This might indicate the fact people with higher income might go for costly liquor like Whiskey, Scotch or Vodka and people with lower income might be opting for beer as it is comparatively cheaper.

* Sales increase slightly with an increase in the amount of retired customer

Retired customers have more leisure and free time than people who are working. This might be the reason that they are more inclined towards purchasing beer.

* Sales tend to rise with the increase in the number of unemployed customers

Unemployed customers have both free time and monetary issues. So beer seems to be the most affordable option.

* Sales tend to increase when the percentage of population below 60 years is less.

If the percentage op population below 60 years is less it is highly likely that people are retired and have more free time.

# **Other Factors**

1. **Percentage of Unemployed Males**: Unemployed males tend to be a dedicated client base for beer purchases. Having this data would allow us to find ways and run campaigns to better target them.
2. **Holiday Sales Data**: Specific sales data for various holidays, such as Christmas and new years, etc would allow us to figure out which seasonal campaign was successful and when improvement is needed.
3. **Promotional event sales**: Sales data for any promotional campaign usually results in spikes of sales. Having this data would allow us to better realize the marketability and hence sales prediction of the beer.
4. **Unfulfilled request count:** Amount of times a customer was unable to find the beer of his choice would enable us to keep a smarter stock and help us make better predictions
5. **Sales timings**: Knowing what time attracts the most customers and comparing them with regular store timings can help us predict how the sales are going to be.

# **Excel File**

