CEE322: Data Analytics for Urban Systems

Lab 5: Confirmatory Data Analysis: Linear Regression

Author: Davyd Tamrazov

Due date: Oct 26, 2020

Stanford University

1 Introduction

The following report outlines the performed confirmatory data analysis on the dataset containing Home Depot's quarterly revenue over several quarters as well as a number of macroeconomic, weather, and other variables. The aim of this analysis is to identify some relationships that would be useful in the improvement of Home Depot's revenue analysis and modeling processes. This is done by selecting three sets of variables and building a linear regression model for each. Namely, variables considered in the analysis include extreme weather events, macroeconomic indicators, as well as customer and employee satisfaction. Finally, throughout the analysis, AIC step methodology is used to narrow down the selection of variables for each set and each model.

2 Dataset

The dataset contains 19 entries, each representing the average value of the 224 factors collected quarterly. Home Depot's quarterly revenue is included as one of the factors and is used as the output for the linear regression analysis in the future sections. Only a limited set of the variables will be used as regressors to predict Home Depot's revenue, that can be further subdivided into the following categories:

1. *Extreme weather events* (see Table 7, Appendix 1) – include the extreme highest category snowfall as well as extreme average regional snowfall index specified for six regions in the US. Further, percentage of days with certain characteristics is provided US wide. Finally, another selection of variables includes data on hurricanes and tropical cyclones in North Atlantic.
2. *Macroeconomic indicators* (see Table 8, Appendix 1) – a selection of the key macroeconomic variables was made by the criteria of being indicators of the national economic performance. These were roughly divided into the economy, consumer, and housing related indicators.
3. *Customer/Employee satisfaction* (see Table 9, Appendix 1) – a number of quantitative ratings given by the consumer and employees as well as the customer confidence index.

Note that the summaries provided include only the non-NA values in each variable. However, due to a varying number of the NA values in the variables, dataset is re-filtered before every analysis to make sure that only valid observations are considered fitting the linear model.

3 Analysis

The following analyses aim to identify if there is a significant relation present between the Home Depot's quarterly revenue and three sets of variables: extreme weather events, macroeconomic parameters, and customer/employee satisfaction. In order to do that, a linear regression and AIC step feature selection is performed on each set. The latter procedure performs a stepwise selection of the model by iteratively adding or removing variables based on the AIC score until no improvement can be attained. For the purpose of this analysis, stepwise search is performed in both directions.

It is important to note that given a very small dataset size, overfitting becomes a major concern when the number of features approaches the total number of observations. As such, the number of regressors included in each linear regression analysis is limited to 6 and the following procedure is performed on each set:

1. A linear regression model is fit using AIC stepwise selection to subsets of features in the set, each containing less than or equal to 6 regressors;
2. Selected features in each of the valid models are recombined and AIC stepwise selection if performed to determine the final model;
3. Final model is validated using the distribution of residuals;

3.1 Extreme Weather Events

Initially, the analysis aims to scrutinize the hypothesis that there is a relationship between extreme weather events and the Home Depot's revenue. As mentioned previously, the extreme weather events can be roughly subdivided into the regional snowfall data, nation-wide percentages of days with certain parameters, and North Atlantic hurricane and tropical cyclones information. However, since regional data contains 12 variables, it requires further subdivision into Northeast/Northern Rockies/Ohio and South/Southeast/Midwest subsets. Thus, 4 linear regression submodels are fitted and rationalized using AIC stepwise approach, as summarized in Table 1. Additionally, the last row illustrates the final model (F) obtained by recombining all of the submodels.

#	*Features*	*P-value*	F-stat.	Adj. R2	AIC
1	Northeast Region Extreme Highest Category Snowfall	0.105	2.018	0.203	304.6
	Northeast Region Extreme Average Regional Snowfall Index	0.162			
	Ohio Region Extreme Highest Category Snowfall	0.159			
	Ohio Region Extreme Average Regional Snowfall Index	0.042			
2	Midwest Region Extreme Average Regional Snowfall Index	0.018	4.377	0.297	301.1
	South Region Extreme Highest Category Snowfall	0.057			
3	Percentage of Days with Thunderstorm	0.005	5.343	0.338	323.9
	Percentage of Days When Temperature Is Higher than 100F	0.027			
4	—	—	—	—	—
F	Northeast Region Extreme Highest Category Snowfall	0.035	4.655	0.533	296.0
	Ohio Region Extreme Highest Category Snowfall	0.038			
	Ohio Region Extreme Average Regional Snowfall Index	0.067			
	Midwest Region Extreme Average Regional Snowfall Index	0.006			
	South Region Extreme Highest Category Snowfall	0.027			

Table 1: *Summary of the linear models obtained using extreme weather events features*

	Model #1	*Model #2*	*Model #3*	*Model #4*	*Final Model*
Q1	-839.6	-1045.0	-874.18	—	-606.4
Q2	-110.7	98.1	-208.0	—	-126.1
Q3	938.6	545.0	662.2	—	679.8

Table 2: *Summary of the residual quantiles*

The above table shows that, generally, the correlation between the Home Depot's revenue and the exteme weather events is not strong. This is particularly clear when looking at the performance of the submodels, where despite significant scores of some features, adjusted R2 value is small. From model #1, it may be observed that Northern Rockies region is excluded from the model, while snowfall in Northeast and Ohio regions are insignificant when considered by themselves. Similarly, only a small number of parameters representing percentage of days was selected, suggesting that extreme events other than snowfall have weak correlation with the output. Finally, model #4 reveals that North Atlantic hurricane and tropical cyclones have no influence on the Home Depot's revenues, what confirms the previously made observation.

It is important to validate the obtained model with the residuals distributions summarized in quantile form in Table 2. It has to be noted that due to a small dataset size, each residual has a very large effect on the distribution of the residuals, meaning that the range of acceptable deviation of the median (Q2) from 0 is large. Just from the table, it is clear that model #2 and model #3 are not valid with the Q1 and Q3 being too different while median significantly deviating from 0. However, for the final model, the quantile information has to be considered together with the visual representation of the residual distribution to make a meaningful conclusion about the validity of the model, as shown in Figure 1. Looking at the histogram further reveals that the distribution of the residuals is, in fact, skewed to the right. At the same time, Q-Q plot suggests that such skew is not significant and the distribution still closely aligns with the normal distribution, apart from slight deviations at either end, meaning that the model is valid.

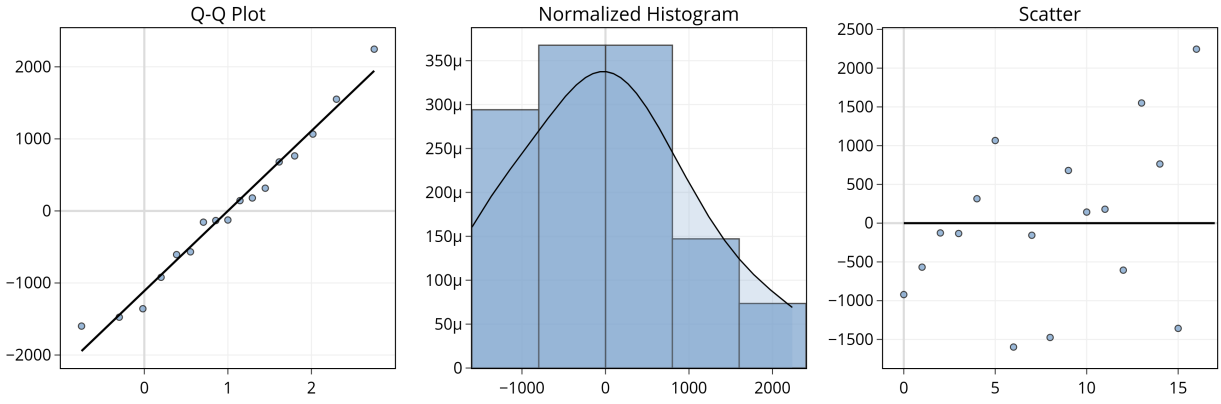

Figure 1: *Diagnostics of the residual for the final model*

With most of the parameters being statistically significant, this model achieves a fair predictive accuracy (adjusted R2 of 0.533), meaning that extreme weather events have a certain effect on the revenue. However,given that F-statistic is not large, the relationship between the extreme weather events and the Home Depot's revenues is most likely to be weak.

3.2 Macroeconomic Indicators

This part aims to address the hypothesis that Home Depot's revenue largely depends on the national economy. To start with, it may be observed from Table 8 that standard deviation of the interest rate is 0, meaning that it is constant for each observation in the dataset and should not be considered as a feature. Additionally, some of the economic indicators tend to be strongly collinear, so it is important to consider correlation between variables, as shown in Figure 2. There, it is clear that the main macroeconomic indicators, namely, gross national product, GDP growth rate, unemployment rate, and consumer price index, are positively collinear. Similarly, these indicators are all negatively correlated with the average monthly disposable personal income. This means, that only one variable is necessary to accurately represent the effect of these features on the Home Depot's revenue. The same observation can be made about housing indices, where housing price index is very strongly correlated with both price and affordability as well as the main macroeconomic indicators. Thus, housing price index can be dropped from the analysis with minimal loss.

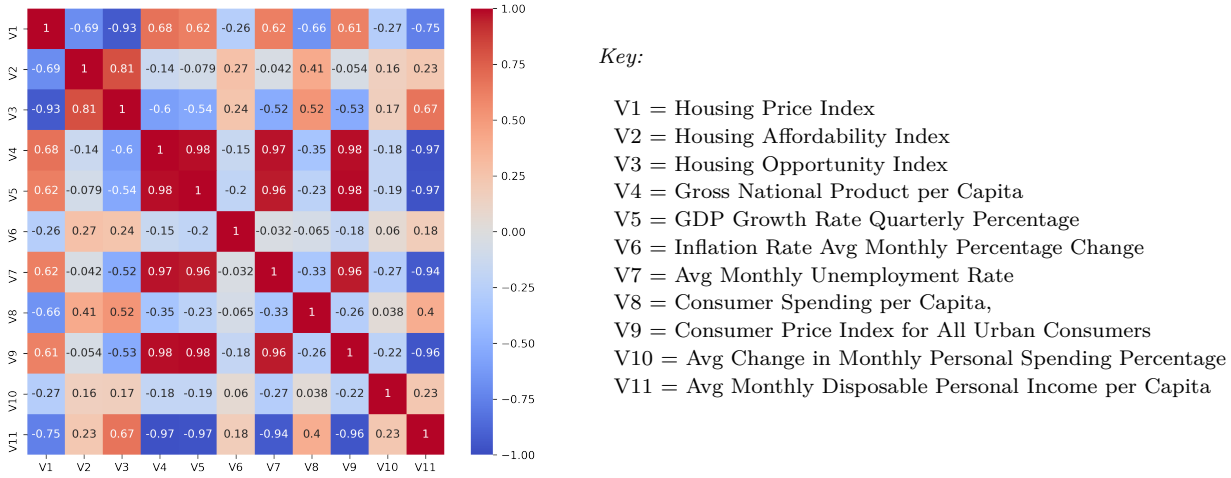

Figure 2: *Correlation table*

Omitting the collinear terms results in the total of 6 variables, so the final model can be determined without further subdivision of the variables, as summarized in Table 3.

#	*Features*	*P-value*	F-stat.	Adj. R2	AIC
F	Housing Opportunity Index	0.045	8.706	0.694	312.0
	Housing Affordability Index	0.011			
	Consumer Spending per Capita	0.005			
	GDP Growth Rate Quarterly Percentage	0.040			
	Average Change in Monthly Personal Spending Percentage	0.128			

Table 3: *Summary of the linear models obtained using macroeconomic indicator features*

The 25%, 50%, and 75% quantiles of the residuals in the above model are -328.4, 169.4, and 552.3 respectively. Further, Figure 3 reveals that given the range of the residuals and the total number of points, distribution looks somewhat normal. Although it may still be insufficient to consider this model as valid for predictive analysis purposes, for the purpose of confirmatory analysis this model may be assumed to be valid. F-statistic in Table 3 shows that there is certainly

a relationship between macroeconomic indicators and the Home Depot's revenue. In particular, housing affordability index and consumer spending per capita appear to have the most significant contribution to explaining a large portion of the variation in the revenue. On the other hand, average change in monthly personal spending is the only statistically insignificant variable in the developed model.

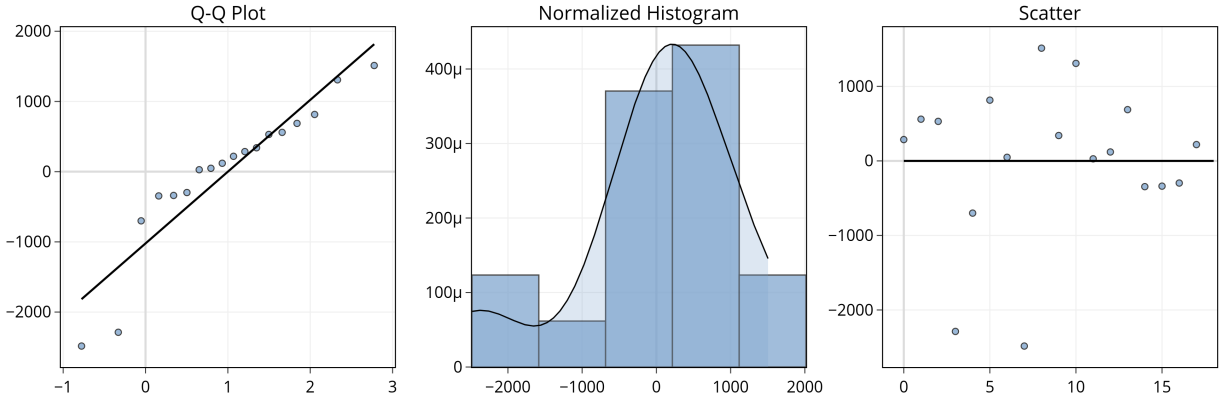

Figure 3: *Diagnostics of the residual for the final model*

3.3 Customer/Employee Satisfaction

Here, the goal is to answer whether customer and employee satisfaction correlates with higher sales. Firstly, Home Depot average culture values rating is only available for 7 observations; thus, it is omitted as a part of this analysis. Next, a set of variables described in Table 9 is split into two groups: customer related and employee related. Performing AIC stepwise model selection results in the linear models described in Table 4 with residual quantiles shown in Table 5.

#	*Features*	*P-value*	F-stat.	Adj. R2	AIC
1	Consumer Experience Rating Homedepot	0.016	7.307	0.271	324.8
2	Home Depot Average Seniormanagement Rating by Employees Home Depot Average Compbenefits Rating by Employees	0.133 0.019	3.436	0.223	326.8
F	Consumer Experience Rating Homedepot	0.016	7.307	0.271	324.8

Table 4: *Summary of the linear models obtained using customer/employee satisfaction features*

	Model #1	*Model #2*	*Final Model*
Q1	-952.6	-940.2	-952.6
Q2	-375.1	5.765	-375.1
Q3	817.1	706.6	817.1

Table 5: *Summary of the residual quantiles*

From the above results, it is evident that the relationship between the customer/employee satisfaction features and Home Depot's revenues is extremely weak. Although the second model appears to be valid, the F-statistics is small, suggesting weak correlation between variables. Thus, the hypothesis that customer and employee satisfaction has a significant influence on the Home Depot's revenue can be rejected.

3.4 Overall Model

The performed analysis identified that the strongest relationship with the Home Depot's revenue comes from the macroeconomic indicators, followed by the extreme weather events. Combining both of these features results in a well performing 9D model, which, however, very likely overfits the data. Thus, it is necessary to be able to balance the dimensionality of the problem with the model's accuracy as well as its validity. As such, the final proposed model is showed in Table 6 and the corresponding residual plots in Figure 4.

#	*Features*	*P-value*	F-stat.	Adj. R2	AIC
F	Ohio Region Extreme Highest Category Snowfall	0.012	9.404	0.724	287.1
	Housing Opportunity Index	0.067			
	Housing Affordability Index	0.035			
	Consumer Spending per Capita	0.008			
	Average Change in Monthly Personal Spending Percentage	0.005			

Table 6: *Summary of final proposed linear model*

The above model minimizes AIC score, while satisfying the constraint of 6 input variables. At the same time, a noticeably larger F-statistic and adjusted R2 value may be observed, suggesting stronger relationship between variables. From the plot below, it is clear that, although distribution of the residuals does not perfectly fit the normal distribution, it is centered around 0 with 25% and 75% being reasonable close (-556.9 and 709.7). Better distribution of the residuals could potentially be achieved with more observations. Despite that, given that p-values are low for most of the variables, there can be a high confidence in the proposed model for the purpose of future development of a more accurate revenue analysis. Noticeably, all parameters but the housing opportunity index have statistically significant parameters, with consumer spending and change in monthly personal spending having the smallest p-values. This means that the adjusted R2 value should be very representative of the true value for this linear fit.

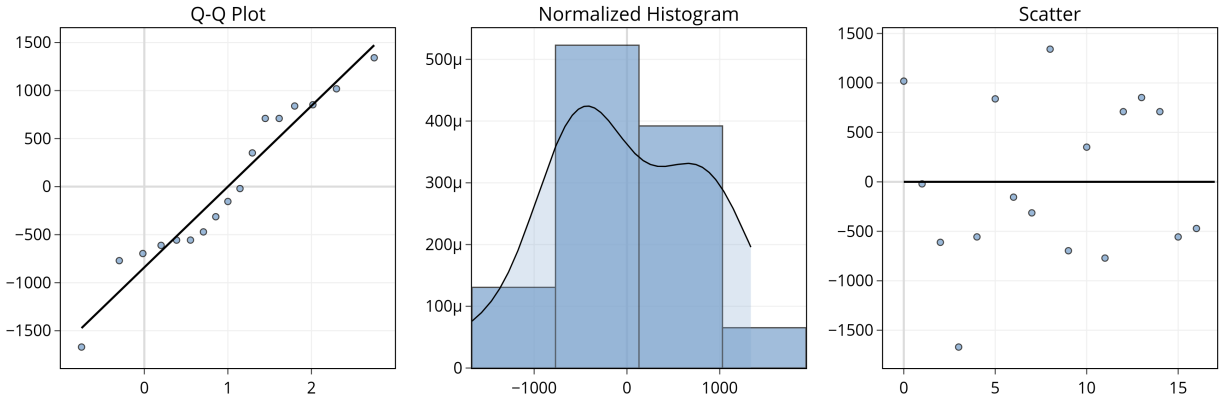

Figure 4: *Diagnostics of the residual for the final model*

4 Conclusion

To summarize, a number of linear regression models were fit to the selected variables from the data to identify any relationships with the Home Depot's revenue. As such, it was found that macroeconomic indicators are well correlated with the output, followed by the extreme weather events. On the other hand, customer and employee satisfaction did not have a statistically significant effect on the Home Depot's revenue. Finally, a model was proposed for further work on improving the revenue analysis.

Appendix 1

Descriptor	*mean*	*std*	*min*	*median*	*max*
Northeast Region Extreme Highest Category Snowfall	1.682	1.186	0.360	1.510	4.350
Northeast Region Extreme Average Regional Snowfall Index (SI)	1.118	1.317	0.000	1.000	4.000
Northern Rockies Region Extreme Highest Category Snowfall	1.664	0.896	0.530	1.450	3.330
Northern Rockies Region Extreme Average Regional SI	1.000	1.118	0.000	1.000	3.000
Ohio Region Extreme Highest Category Snowfall	1.437	0.976	0.000	1.030	3.020
Ohio Region Extreme Average Regional SI	0.941	1.298	0.000	1.000	5.000
South Region Extreme Highest Category Snowfall	3.334	2.226	0.000	3.500	7.390
South Region Extreme Average Regional SI	1.471	1.807	0.000	1.000	5.000
Southeast Region Extreme Highest Category Snowfall	1.049	1.205	0.000	0.660	2.970
Southeast Region Extreme Average Regional SI	1.492	1.832	0.000	0.500	5.390
Midwest Region Extreme Highest Category Snowfall	0.588	1.326	0.000	0.000	4.000
Midwest Region Extreme Average Regional SI	0.706	1.263	0.000	0.000	4.000
Percentage of Days When Temperature Is Higher than 100F	0.03%	0.05%	0.00%	0.00%	0.14%
Percentage of Days When Temperature Is Less than 32F	9.45%	10.9%	0.34%	3.89%	31.5%
Percentage of Days When Humidity Greater 95 Percent	0.89%	0.47%	0.31%	0.87%	1.69%
Percentage of Days When Mean Wind Speed Greater than 25mph	8.33%	5.92%	2.22%	7.35%	16.4%
Percentage of Days with Thunderstorm	0.02%	0.01%	0.00%	0.02%	0.04%
Percentage of Days with Tornado	0.05%	0.04%	0.01%	0.04%	0.12%
No. of Hurricane & Tropical Cyclones (H&TC) in North Atlantic	3.944	5.104	0.000	2.000	16.00
Average of Max Winds in North Atlantic, mph	49.11	38.30	0.000	60.00	105.0
Total No. of Deaths Caused by H&TC in North Atlantic	24.00	44.86	0.000	0.000	148.0
Damage Caused by H& TC in North Atlantic, $ million	4435	17627	0.000	0.000	75002

Table 7: *Summary of the extreme weather features*

Descriptor	*mean*	*std*	*min*	*median*	*max*
Housing Price Index	176.7	9.476	165.4	174.8	196.8
Housing Affordability Index	178.5	14.79	152.0	178.2	202.5
Housing Opportunity Index	71.51	4.334	62.60	72.75	77.5
Gross National Product per Capita, $	49435	955.6	47604	49600	50781
GDP Growth Rate Quarterly Percentage	2.19%	1.36%	-2.10%	2.00%	4.50%
Inflation Rate Average Monthly Percentage Change	2.00%	0.82%	1.13%	1.74%	3.77%
Interest Rate Monthly Percentage	0.25%	0.00%	0.25%	0.25%	0.25%
Average Monthly Unemployment Rate	8.30%	1.05%	6.43%	8.22%	9.77%
Consumer Spending per Capita, $	33242	696.4	31930	33238	34312
Consumer Price Index for All Urban Consumers	227.9	6.447	217.3	228.8	237.7
Average Change in Monthly Personal Spending Percentage	0.32%	0.14%	0.10%	0.33%	0.54%
Average Monthly Disposable Personal Income per Capita, $	38265	1424	35579	38494	40421

Table 8: *Summary of the macroeconomic indicator features*

Descriptor	*mean*	*std*	*min*	*median*	*max*
Consumer Experience Rating Homedepot	3.14	0.13	2.87	3.16	3.41
Consumer Confidence Index	178	43.3	80.5	187	243
Home Depot Average Overall Rating by Employees	3.20	0.12	2.96	3.25	3.38
Home Depot Average Culturevalues Rating by Employees	3.46	0.06	3.35	3.45	3.56
Home Depot Average Worklifebalance Rating by Employees	3.04	0.16	2.82	3.07	3.32
Home Depot Average Seniormanagement Rating by Employees	2.86	0.13	2.66	2.86	3.26
Home Depot Average Compbenefits Rating by Employees	3.24	0.13	3.03	3.24	3.49
Home Depot Average Career Opportunities Rating by Employees	3.10	0.14	2.90	3.09	3.35

Table 9: *Summary of the customer and employee satisfaction features*