Individual Assignment 3, Dion Chang - 20812576

1) Problem statement and data used: The public consumer complaints data used for analysis had over 2 million rows of complaints ranging from the December 01, 2011 up to March 14, 2021. The objective was to determine how many potential complaints we would get from consumers in 2022.

2) Planning: For the data wrangling and cleaning, an aggregate of the total count of monthly complaints was performed. This was accomplished by counting the number of rows a month and year was repeated. As an example: for the month and year of December 2011, the total number complaints was determined to be 2536. This was for the purpose of building a linear regression model. As for assumption tests, a Durbin-Watson test was used to check for autocorrelation of residuals. Residual plots was used for diagnosing homoscedascity and linearity.

Furthermore, the dataframe was further filtered out by removing the last 12 observations in order to build a more linear model. As shown in Figure 1, there was a huge spike in complaints after March 2020 (piecewise regression may be performed, but not for the purpose of this report). Therefore, the final data set used for the linear regression analysis ranges from December 2011 to March 2020 as shown in Figure 2. Notice that in Figure 2, the dates have been encoded to numeric variables in order to perform linear regression.

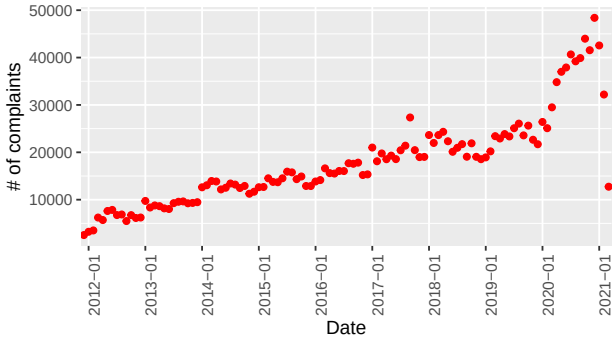

Figure 1. Plot of the number of complaints vs. Date (year−month).

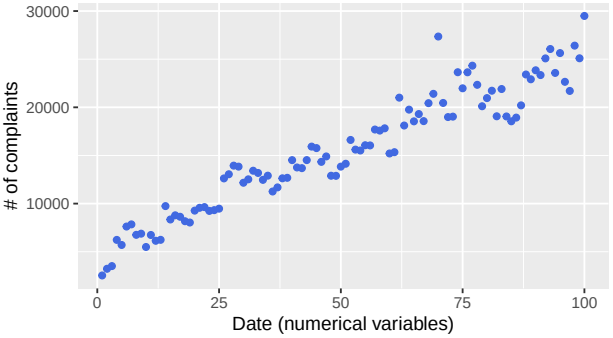

Figure 2. Cleaned plot of the number of complaints vs. Date converted to numeric variables.

A sample table (Table 1) has been provided to show how Figure 2 was plotted. The column **date.num** corresponds to the **year.month**, and it is the x-axis. The dependent variable was the number of complaints **(complaints.count)**:

Table 1. Sample table of the dataset after wrangling and cleaning in order to perform linear regression to determine the number of potential complaints from consumers in 2022.

date.num	year.month	complaints.count
1	2011-12	2536
2	2012-01	3230
3	2012-02	3509
4	2012-03	6230
5	2012-04	5703

3) Analysis: From the regression, it was determined that both intercept and **date.num** coefficient are significantly different from zero ($p < .001$). The R^2 is .92, meaning that 92% of the variance is explained. The number of consumer complaints is equal to 5106.368 + 206.746 (date.num), where date.num is coded as 1 = Dec 2011, 2 = Jan 2012, and onward monthly.

The **Durbin-Watson test** for independent errors was significant at the 5% level of significance ($d = 1.09$, $p < .001$). Therefore, the null hypothesis is rejected which indicates that the residuals are positively autocorrelated.

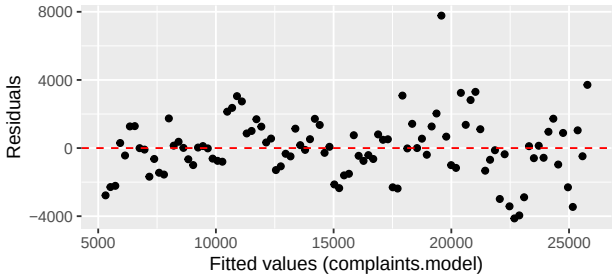

Figure 3. Residuals vs. fitted values

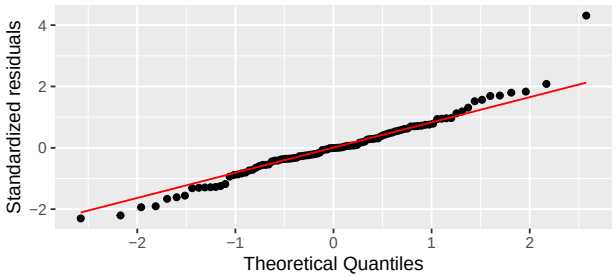

Figure 4. Q−Q Plot for normality

As shown in Figure 3, the residuals versus fitted values plot of the linear regression model shows that it is overall well-behaved. The horizontal band formed by the residuals suggested that the variances are constant (homoscedastic). This also shows that the residuals are scattered randomly, meaning linearity can be assumed.

The Q-Q plot shown in Figure 4 shows that it is safe to assume normal distribution.

To estimate the total number of potential complaints from consumers in 2022, a table of all the months in 2022 with their total number of complaints was predicted as shown in Table 2:

Table 2. Number of consumer complaints predicted for each month in the year 2022. "date.num" was used as the predictor variable which represented the actual dates.

date.num	Date	Predicted.Complaints
122	Jan 2022	30329.40
123	Feb 2022	30536.15
124	Mar 2022	30742.89
125	Apr 2022	30949.64
126	May 2022	31156.39
127	Jun 2022	31363.13
128	Jul 2022	31569.88
129	Aug 2022	31776.63
130	Sep 2022	31983.37
131	Oct 2022	32190.12
132	Nov 2022	32396.86
133	Dec 2022	32603.61

Therefore, the total sum of the complaints from Table 2 was determined to be 377598.1. This meant that in 2022, there could be approximately 377598 complaints from consumers.

4) Conclusion: The linear regression model determined that there will be **377598** potential complaints from consumers in 2022. The linear model was determined to be the following:

$$y = 206.7462x + 5106.369 \quad (1)$$

where y = Number of consumer complaints and x = the date in Month-Year format encoded as numerical values.

From the analysis, the model overall met residual assumptions for linear regression. The residuals are homoscedastic, meaning that it has constant variance. Linearity can also be assumed as shown in the residual plot in Figure 3. From the Q-Q plot (Figure 4), the data is normally distributed which is what we want. From the Durbin-Watson test, there was a strong positive autocorrelation but not unexpected since this is common when dealing with time series data. The visualization (scatter plot in Figure 2) overall shows a linear relationship which further shows a positive relationship between number of complaints and as time goes by.