Individual Assignment3- Simple Linear Regression

1. Problem statement and data used:

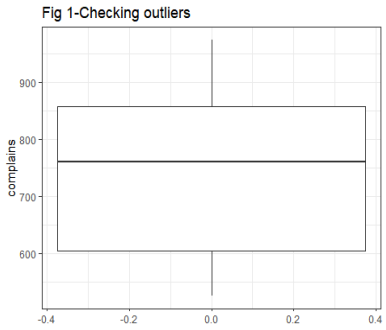
We are considering the data of `Bank of America` to estimate the potential complains for year 2022. Here we are considering 2 products i.e. checking or savings account, bank account or service and all 8 sub-products and so all the predictions will be limited to these products. The data is available from year 2012 to year 2021 up to current (3rd) month.

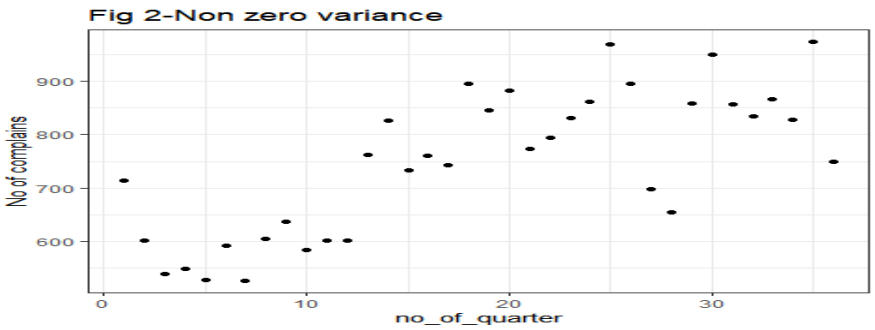
2. Planning:

Data is wrangled [Appx-1] and Year & Month wise breakdown of complains is also provided in [Appx-2, 3]. Please note that complains for 2021 will be very less as compared to previous because we have data only for 3 months. So for the sake of analysis we are considering complains per quarter, starting from 2^{nd} quarter of year 2012 to 1^{st} quarter of year 2021. Since at least 20 cases per independent variable in the analysis are required and we are considering only 1 predictor variable, our sample size (36) is sufficient.

Assumption check

- **The relationship between IVs and DV is linear:** The predictor variable (`quarter`) is quantitative & the outcome (`no of complains`) is quantitative, continuous [Appx-4] we can consider it to be unbounded. Since linear regression is sensitive to outliers, (fig-1) shows that we have no outlier.
- **Non zero variance:** The data varies and same can be visualized with the help of scatter plot (fig-2).





- **No or little multicollinearity:** Here we are considering single predictor so we need not to check multicollinearity.Remaining assumptions can be checked after linear regression.

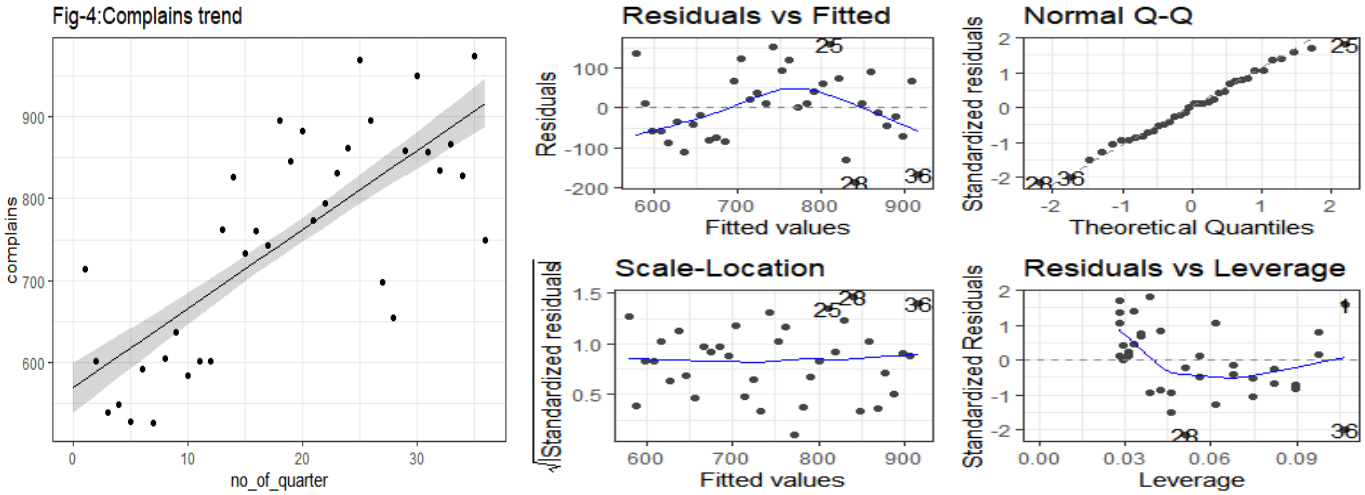
3. Analysis:

We are trying to predict the no of potential cases for year 2022 depending on complains in past so let us build the regression model [Appx-5]: $\mathbf{complains}_i = \boldsymbol{\beta_0} + \boldsymbol{\beta}_{\mathbf{quarter}}\mathbf{quarter}_i + \boldsymbol{\epsilon_i}$

- **Model assessment:** $\text{complains}_i = 568.96 + 9.67 * \text{quarter}_i + \epsilon_i$
- **Coefficients significance:** Both the intercept and `no_of_quarter` coefficient are significantly different than zero ($p < 2 \times 10^{-16}$).A statistically significant coefficient indicates that there is an association between the predictor (x) and the outcome (y) variable.This can also be confirmed by looking at the confidence interval of the coefficients [Appx-6]. The adjusted R^2 value in the above model is 0.5593, which explains 55.93% of the variance by the model. A large F-statistic will corresponds to a statistically significant p-value ($p < 0.05$). In our example, the F-statistic equal 45.42 producing a p-value of 9.589e-08, which is highly significant.

Assumption check:

- **Normality:** From QQ plot (graph-2), data under analysis is almost aligned diagonally with some minor deviation or histogram [appx-12] follows bell shape so the assumption of normality is met.
- **Linearity:** From Residual vs Fitted (graph-1), residuals are distributed equally about the zero & also the black line is approximately horizontal at zero. Thus, linearity assumption is followed.
- **Homoscedasticity:** From Scale-Location (graph-3), as we can see most of the residuals are spread equally along the range of predictor. Also, line is quite horizontal which shows that homoscedasticity is followed. Our plot of standardized residuals vs standardized predicted values [Appx-14] showed no obvious signs of funneling, suggesting the assumption of homoscedasticity has been met.
- **Auto-correlation:** Using Durbin Watson test [Appx-17] value of d is 1.14, p=0.006 which is not (1.5 < d < 2.5) show that there is auto-correlation in the data and hence residuals are not independent from each other. In order to handle this violation the bootstrapping can be done [Appx-18].
- **Influential cases:** Residuals vs Leverage (graph-4) about outliers and influential points.In this plot, the large values marked by cook's distance (26, 36, 38) might require further investigation (Such influential points tends to have a sizable impact of the regression line).



- **Model accuracy:** Residual standard error (RSE) = 89.41, meaning that observed no of complains from the true regression line by approximately 89.41 units in average.However, we can calculate the percentage error [Appx-7], which is 11.96% (quite low). Value of R^2 =0.57 shows that model fits the data well with F-statistic equal 45.42 producing a p-value of 9.58e-08, which is highly significant.
- **Predicting number of cases for year 2022:** From the equation of linear regression we can predict the future (quarter-wise) values as given in table below[Appx-15]:

Q2-2021	Q3-2021	Q4-2021	Q1-2022	Q2-2022	Q3-2022	Q4-2022
926.66	936.32	945.99	955.66	965.33	974.99	984.66

4. Conclusion:

From table given above and quarter-wise analysis of no of complains for checking or savings account, Bank account or service (fig-4) we can expect to receive 3881 complains in year 2022(with approx less or more 89 complains per quarter than as given in table). However, since some pre-requisite (Assumptions) are not met, we can better predict no of complains using other options like accounting for other factors (Multiple linear regression) or considering more instances (bootstrapping).