Task 2: Analysis of factors affecting Linear Regression

Please refer to 'Analyzing factors affecting Linear Regression.ipynb' notebook for code and graphs supporting observations made in this report.

1.0 Introduction

The two factors affecting linear regression that were analyzed are outliers and multicollinearity. While generating sample data, the data for all covariates was set to follow a uniform distribution whereas the noise variable follows a normal distribution. This was kept consistent for all models to isolate any effects of changing data distributions. For model building, statsmodels' Ordinary Least Squares (OLS) model was used.

We start by establishing two baseline models to validate our assumptions for linear regression. The first, with one covariate to establish that OLS can to a high degree of accuracy capture the model parameters (i.e., covariate coefficients) when there's no presence of outliers. The second, with three independent variables where the scale of each of the covariates vary by an order of magnitude. Here we note that while this affects the standard deviation with which the model parameters are estimated, they still closely follow the true values when there's no collinearity amongst the variables.

2.0 Outliers

Outliers are datapoints that generally diverge from the overall pattern of data [1]. For analyzing their effects, a single variable model was used to allow for visual representation of the outliers on a 2D graph.

2.1 Studying effects of outliers

The presence of outliers in a data model tend to affect the accuracy with the model parameters are estimated. During analysis an increasing percentage of outliers were introduced to the dataset and metrics such as the estimated slope in relation to the true value and the standard deviation of the estimated slope were tracked. It was observed that both measures increase linearly as the percentage of outliers increases. A similar observation is made in relation to RMSE which implies the larger the number of outliers in a dataset the worse the model is in predicting the true value.

A comparable observation was also made while studying the effect of changing the magnitude of outliers on model parameter estimation and RMSE, where an increase in outlier magnitude results in a poorer parameter estimation and higher RMSE. When examining changing magnitudes from -2 to 3.5, it is interesting to note that there's symmetry of the graph when magnitude of outliers = 1 i.e., when there are no outliers. That is the single point at which the estimated coefficient error and standard deviation (and RMSE) are minimized.

The observations made are consistent with expectations as in a least squares model, the best fit line is obtained by trying to minimize the squared distance between the observed and estimated points. In having either many "small" outliers or having outliers with a large magnitude, the model skews the regression line away from the general pattern of data to better fit the outliers (i.e., minimize the squared error). With this method, when large error values are squared, the influence of those points on the regression line are amplified. The next section identifies ways to detect outliers.

The effects of normalization of covariates in the presence of outliers were also studied, however this improved model performance almost trivially. Since outliers, even post-normalization remain as outliers, there was marginal improvement in the estimated model parameter and their variance.

2.2 Techniques to detect outliers

In 2D graphs detecting outliers is straightforward as this can be done through visual inspection of scatter plots. For higher dimensional models, we employ tools such as influence plots. Influence plots plot each datapoint on a residual vs. a leverage graph, where the size of the datapoint is a measure of influence. To identify outliers, the plot can be used to identify points with high leverage (point 34) or high residuals (point 18, 41), or both (point 21). Datapoints with high influence can affect model parameters such as slope or coefficient of determination.

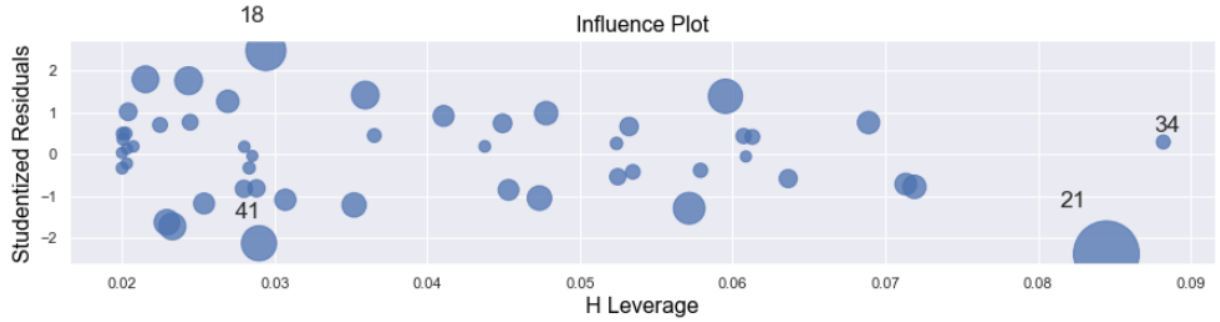

Figure 2.1: Influence Plots

A residual-vs-fitted plot may also be used to identify outliers in the covariates. The graph below on the left-hand side shows the characteristics of a well-behaved plot (i.e., constant variance around the regression line/ residual =0 line). Whereas the right-hand side plot shows the same plot for a dataset with outliers. Here, the trail of outliers can be easily identified. In the same plot, we also note that that one of the assumptions for linear regression, homoscedasticity is violated.

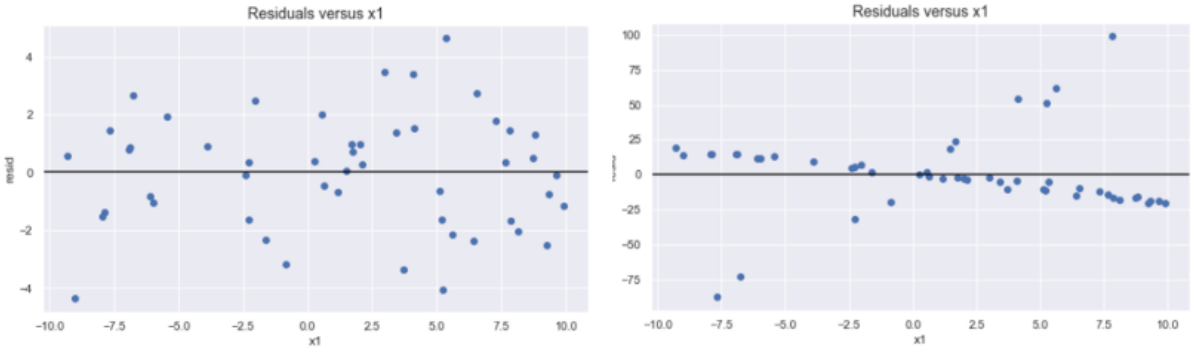

Figure 2.2 Residuals vs Fitted Plots

2.3 Handling outliers

Before removing outlier from a model, it is important to evaluate the nature of the datapoint in the context of the problem. For example, in experiments, if the datapoint is erroneous in nature (i.e., human/testing error) discarding the value can improve how well the model mimics the real-life system. However, when the outlier deviates from the general trend but is a valid datapoint, it

should be retained and alternative solutions such as transformation of variables should be considered.

3.0 Multicollinearity

Multicollinearity is a condition when two or more variables in a multiple regression model are highly correlated. This violates one of the assumptions in linear regression where the model assumes all covariates are independent. In this section, we'll study the effects of multicollinearity, how to identify its presence and how to mitigate its effects when present.

3.1 Examining effects of multicollinearity

In regression models, in addition to building a model that can accurately predict the dependent variable, we also aim to identify how the dependent variable reacts to changes in each independent variable. When there's multicollinearity present, one of the challenges in isolating the impact of each variable is that the stronger the correlation, the harder it is to isolate its impact.

A consequence of having multicollinearity in regression models is poor estimation of model parameters. We note that covariate coefficients become sensitive to small changes in the model and can swing wildly based on which variables are in the model. This is the result of highly correlated covariates, where small changes in one covariate propagate the effects of the change to all other correlated covariates. Nonetheless, it can be noted this phenomenon only affects variables that are correlated. Estimation of covariate coefficients for independent variables in the model that aren't collinear are unaffected and the effects of that variable are captured with good accuracy.

Another consequence of multicollinearity is that it can affect the reliability of the p-values of the coefficients. Meaning the p-values alone can't be used to identify independent variables that are statistically significant.

3.2 Techniques to detect multicollinearity

To assess multicollinearity in regression models, the Variance Inflation Factor (VIF) can be used to identify the strength of correlation between independent variables. This is done by first modelling each variable in the model as the independent variable and all other variables as the dependent variables and calculating the coefficient of determination. The R-squared value is then plugged into the following equation $VIF_i = \frac{1}{1-R_i^2}$ to obtain VIF and the process repeated for each variable. VIF values can range from 1 to infinity, where smaller values represent a lower level of correlation. Generally, values above a threshold of 5, represents a high degree of correlation where the model parameters and statistical significance of the model are affected[3].

2.3 Mitigating effects of multicollinearity

In this section we'll address two ways to handle multicollinearity in regression model – using a subset of covariates and using principal component analysis (PCA)

The first method involves selection of a subset of covariates to include in the model by inspection of VIF. Figure 3.1 below shows effects of multicollinearity when using different sets of independent variables. Variables x1, x2 and x3 appear to be relatively free from any multicollinearity. On introduction of x4, we observe a moderate level of correlation. On adding variable x5 and x6 we can clearly note the effects of multicollinearity and can distinguish independent from correlated variables in the model. Therefore, if we're keen on removing multicollinearity using model 1 with

3 covariates is a good choice, however if hypothetically we were interested in studying the effect of x4 on the response, as the VIF is below the threshold, we may still do so at the expense of having some multicollinearity in the model.

	variables	VIF
0	x1	1.011815
1	x2	1.011702
2	x3	1.000293

	variables	VIF
0	x1	1.012717
1	x2	2.152754
2	x3	2.262910
3	x4	3.421330

	variables	VIF
0	x1	1.012717
1	x2	inf
2	x3	inf
3	x4	3.421330
4	x5	inf

	variables	VIF
0	x1	1.012717
1	x2	inf
2	x3	inf
3	x4	inf
4	x5	inf
5	x6	inf

Figure 3.1 – Variance Inflation Factor

Alternatively, we can also use principal component analysis. This technique is typically used extract features and model variables to a lower dimensional space. However, for our purposes, we may also use it to form independent covariates in the model. A downside in using this technique is the loss in interpretability of the features. The technique here is to first run the algorithm on the full set of independent variables and then find the cumulative sum of the percent of explained variance (see Figure 3.2 below). We observe that in using 3 components, 99% of the data variability is captured. To verify the validity of this method, we can run the model with any number or permutations of the 3 principal components and since the components are all independent, we notice that the coefficients of the estimated parameters are consistent across all models.

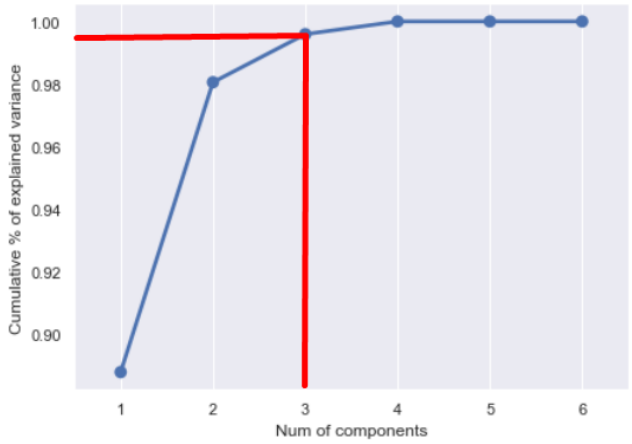

Figure 3.2: Cumulative Percentage of explained Variance vs. Number of Principal Components

It is important to note that handling multicollinearity is only necessary when trying to study the effects of independent variables on the dependent variable. This phenomenon doesn't affect how well the regression line is fit or accuracy of the predicted response, therefore based on what our motivation is in building the model its presence may or may not warrant corrective measures.

5.0 References

[1] Star Trek : Teach yourself statistics. *Influential points*. Influential Points in Regression. Retrieved January 1, 2022, from https://stattrek.com/regression/influential-points.aspx

[2] *Assumptions of linear regression*. Statistics Solutions. (2021, August 11). Retrieved January 2, 2022, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/

[3] Frost, J. (2021, September 24). *Multicollinearity in regression analysis: Problems, detection, and solutions*. Statistics By Jim. Retrieved January 2, 2022, from https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

[4] Kumar, S. (2020, December 19). *How to remove multicollinearity in dataset using PCA?* Medium. Retrieved January 1, 2022, from https://towardsdatascience.com/how-to-remove-multicollinearity-in-dataset-using-pca-4b4561c28d0b

Percentage of code borrowed: 30%