2021/22 Frist Term

INM702 Programming and Mathematics for Artificial Intelligence

Report – Task 2

Analysis, interpretation, and diagnostics in linear regression

Focus on outlier and collinearity

By Alex Collins and Suen Chi Hang

[Alex.Collins@city.ac.uk](mailto:Alex.Collins@city.ac.uk)

[Chi.Suen@city.ac.uk](mailto:Chi.Suen@city.ac.uk)

**Outliers**

An outlier is an observation that lies outside the overall pattern of a distribution (Moore and McCabe 1999). It can be found by residual plots and scatter plot of x, y points.

Chart

Description automatically generated Chart, line chart

Description automatically generated

Fig. 1 Residual plots Fig. 2 Scatter plot of x, y points

In Fig. 1 outliers are found in the far negative area of the two residual plots. In the histogram of residual plot below, most residuals are around zero in a shape similar to normal distribution, except those in far negative area, hence those are outliers. If standardized, residuals beyond +/- 3 may be regarded as outliers. Residual scatter plot against predicted values of y can also display obvious outliers e.g. bottom left of Fig. 1.

In Fig. 2, scatter plot of x and y also visualizes outliers in the top left. However, note that outlier in scatter plot of two independent variables x1 and x2 (i.e. not showing y) may not affect the model if the predicted value y is not too far from actual value.

Impact of outliers

An outlier can affect the linear regression model greatly as the difference is squared for minimization. We will study how outliers affect the intercept, coefficients, residuals and R2 score of the model by holding the same underlying true linear model y = 1 + 2x1 + 3x2 + e (where e is random noise variable ~N(0,1)), but with varying magnitude, number and position of outliers. Simulations are done 1000 times for calculating the variance and mean.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Fig. 3 Coefficient (b1) with outliers

Position matters a lot - outliers at centre of feature space do not affect b1 at all! Evenly distributed outliers across the feature space may slightly change b1, but the mean of b1 remain the same as the true model and the variance is smaller than if outliers are concentrated all at the high end (or all at the low end) of the feature space. The mean of b1 is in linear relationship with magnitude and number of outliers that are not evenly distributed. The variance of b1 increases with number of outliers but only fluctuates with the magnitude.

The charts for coefficients b2 show the same properties as b1, see supplementary figure 1.

Intercept – it increases linearly with magnitude and number of outliers, regardless of the positions of outliers. Interestingly, while the variance of the intercept increases with the number of outliers, it only fluctuates with the magnitude of the outliers. See supplementary figure 2.

Residuals can be measured by sum of squares of residuals (ssr). The mean of ssr increases linearly with number of outliers and quadratically with the magnitude, regardless of position of outliers. The ssr variance has the same shape as the mean of ssr, though less smoothly. See supplementary figure 3.

R2 score = , interpreted as proportion of explained variance, is commonly used as one indicator of fitness of the model, though it may need adjustment when compared with models of different number of parameters. Its mean decreases with absolute magnitude and number of outliers, which implies that ssr increases more than the total variance. Its variance also have same shape except at the point of zero magnitude or zero number of outliers. See supplementary figure 4.

**Multicollinearity**

We built a subclass to study multicollinearity. We consider two related scenarios which give different effects. Scenario 1 is when 1 or more features is a linear combination of other features. Scenario 2 is when to features are correlated.

In scenario 1 we have a model with 5 features. Features 1, 2 and 3 are independent of each other and features 3 and 4 are linear combinations of 1 and 2. Specifically and are of the form . If then the rank of the matrix is 4 and we know that there are two collinearities. However, if (due to natural variance or measurement error) then the features matrix will be of full rank and we need to find the collinearities another way.

We can inspect the variance inflation factor (VIF) of each features. After normalizing features to mean 0 and unit length, we calculate the of the feature relative to the other features to get the VIF. If this is big then it suggests there are collinearities. In the model outlined above (see Jupyter Notebook for details of coefficients etc), we calculated the VIFs to be 102.9, 47.7, 1.0, 49.7 and 100.8 for the 5 features, with of 81%.

The fit of the model is good, but to explain the impact of features, we should try to remove correlated features. Removing the features successively with the highest VIF leaves us with a 3-feature model with an of 81% still, and inflation variance factors close to 1. We are left with an easier model to explain and no loss of fit.

In scenario 2 we have 3 features, the first two of which have correlation of 0.8 with each other The fit is good: 94% , but we have modestly high VIFs of 2.8, 2.8 and 1. If we remove a feature, as above, we end up with a 2-feature model without any correlation within features, but drops to 83%: we have lost important information.

In this case, we can replace one of the correlated features () by the residuals of regressed on the other features. This leaves us with a 3 feature model with inflation co-factors of almost 1 and of 94%, so no evidence of collinearity or loss of information.

One of the big problems with colinear or correlated data is that it is harder to predict coefficients for the correlated variables. Small fluctuations in observed values of y can have big impacts on the estimation of beta. We have the following equation which states that variance in beta estimates is proportional to the VIF.

Text, letter

Description automatically generated with medium confidenceRawlings, Pantula, Dickey, 1998, "Applied Regression Analysis"

We can see this empirically in our data. We can experiment with different correlations between and to see what impact it has on VIF and variance of beta coefficients. The charts below show the variance of beta when the data is generated with different random seeds. For high and low correlations, uncertainty about the coefficients is highest:

|  |  |
| --- | --- |
|  |  |
| *Figure: when correlation between and is high or low, variance of estimates is high.* | *Figure: VIF when correlation is low or high.* |

The equivalent charts are even more interesting for scenario 1: features which are linear combinations of other features.

*Summary:* correlated or co-linear features are a problem. Including all features in a model does a good job at explaining the dependent variable. The problem arises for both explaining the model – isn’t it better to reduce the number of features? And concretely for estimating coefficients for features. We have explored two techniques: removing features and transforming the features. The former is better for co-linear features, and the latter is better for correlated features.

Supplementary Figures

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

S-Fig. 1 Coefficient b2 with outliers S-Fig. 2 Intercept (b0\_mean) with outliers

Chart, line chart

Description automatically generated Chart

Description automatically generated

S-Fig. 3 Residuals - ssr with outliers S-Fig. 4 R2 score with outliers

|  |  |
| --- | --- |
|  |  |
| S-Fig. 4 – where is a linear combination of , the variance of gets much bigger as *b* gets bigger. () | S-fig. 5 – VIF beta gets low or high. Where we have |

References:

### Rawlings, Pantula, Dickey, (1998), *Applied Regression Analysis* 2nd edition, New York: Springer.

﻿<https://www.scikit-yb.org/en/latest/api/regressor/residuals.html>