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)

Outlier

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, line chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Fig. 1 Scatter plot of X vs y and X1 vs X2 Fig. 2 Residual plots

Fig. 1 scatter plot of X and y visualizes outliers in the top, while scatter plot of X1 and X2 have some outliers hidden in the normal range of feature space.

In Fig. 2 Residual plots - outliers are found in the far negative area. In the histogram of residual plot, however, the outlier is too small to see as proportion by count is too small compared with 1000 samples. If standardized, residuals beyond +/- 3 may typically be regarded as outliers. Note that if outliers become too influential (say, because sample size not enough), coefficients may be distorted to fit the outlier such that the residuals measured by fitted model may not be obvious for the outlier. Check further by removing a suspected outlier (e.g. in Fig.1) to see if it greatly affects the model and check the accuracy of raw data.

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 using the same true linear model y = 1 + 2X1 + 3X2 + e (where e is random noise variable ~N(0,1)) with 1000 samples X1 , X2 generated from uniform distribution (-10,10). Then add outliers with specified magnitude, number and position. This is repeated 1000 times for calculating the variance and mean of key estimates. Further repeat above by varying the position, magnitude, number of outliers to study the impact.

|  |
| --- |
| Varying positions of outliers in feature space X  Centre: (0,0)  Even: evenly distributed  High: (10,10), i.e. max of normal feature range  Low: (-10,-10) i.e. min of normal feature range  Ultra-high: (100,100), far beyond high end of X  Ultra-low: (-100, -100), far beyond low end of X  Outliers are generated by adding magnitude k to y of points at specified position |

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 of unevenly distributed outliers. Number of outliers also have similar relationship except ultra-outliers (we define as those far beyond the normal range of feature space X1 , X2), which can change b1 sharply with just one or two outliers. b1 variance increases with number of outliers but only fluctuates with magnitude, except ultra-outliers, which increases b1 variance sharply for first couple of outliers and then decreases.

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

Intercept – its mean increases linearly with magnitude of outliers, regardless of the positions, though ultra-outliers have much less impact. It also increases with number of outliers, except the ultra, which doesn’t change its mean much after the first couple of outliers. Interestingly, while the intercept variance increases with the number of outliers (except ultra, which decreases its variance after the first outlier), 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 for all positions of outliers except for ultra, which increases ssr mean slightly and then almost flattens with more number of outliers. The ssr variance has the same shape as the mean of ssr, though less smoothly and ultra is an exception – it decreases ssr variance after the first one or two outliers. 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 generally 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. The exception is ultra – after the first outlier, the score improves a bit with more outliers. Coefficients are so much influenced to fit ultra-outliers that more ultra-outliers make it more plausible to keep model coefficients which already incline to fitting ultra-outliers. This gives us an alert that we cannot simply detect presence of outliers by R2 score, especially for ultra-ones. And for ultra-high outliers, the mean of score only slightly decreases with higher positive magnitude of outliers. See supplementary figure 4.

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 with outliers

Chart, line chart

Description automatically generated Chart, histogram

Description automatically generated

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

Reference

Mendenhall, W and Sincich, T (2003) *A Second Course in Statistics: Regression Analysis.* 6th edition. USA: Pearson Education.

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