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

Description automatically generated Chart, line chart

Description automatically generated

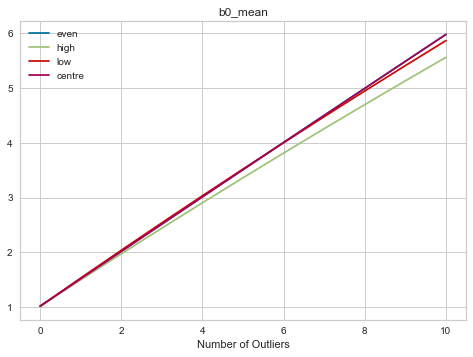
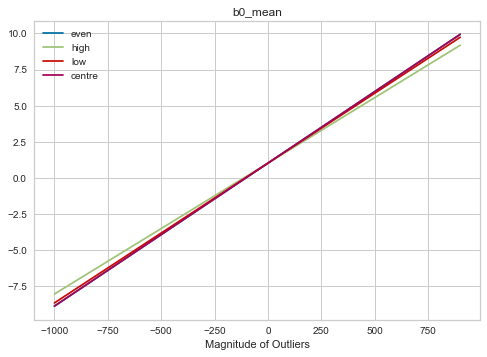
Fig. 1 Residual plots Fig. 2 Scatter plot of x, y points

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.



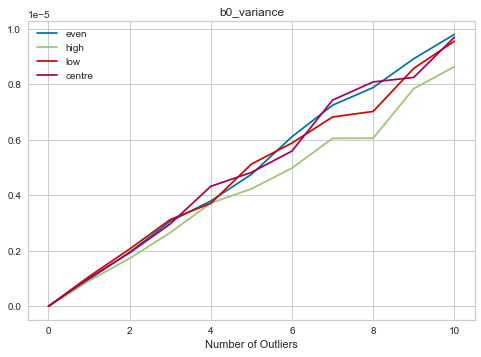
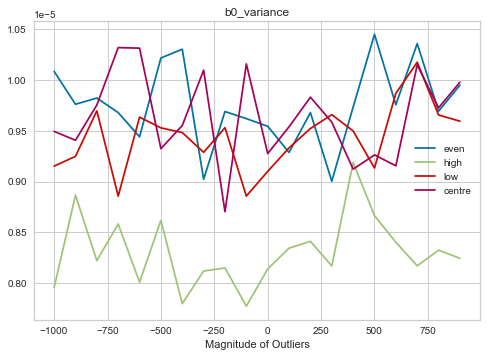


Fig. 3 Intercept (b0\_mean) with outliers

The fitted intercept 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.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Fig. 4 Coefficient (b1) with outliers

The charts for coefficients b2 are showing the same properties as b1, hence omitted here.

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.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Fig. 5 Residuals (measured by ssr) with outliers

We measure residuals 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.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Fig. 6 R2 score with outliers

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.

Supplementary Figures

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

S-Fig. 1 Coefficient b2 with outliers

Reference

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