**IN 3063: Programming and Mathematics for AI**

**Task 2 Report**

**By Amir Harvey & Krzysztof Szymanski**

# Outliers

## Overview:

Outliers are points that are non-representative of the distribution, for example because there is a lot of noise at that data point, so you know that there shouldn’t be data there as it doesn’t match the other trends.

## How they affect the estimation of the coefficients?

Outliers affect the estimation of the coefficients as it increases the range of variables it could be to a larger window than necessary so by dismissing them you can have a more accurate representation of what the variable coefficients should be.

## How can we detect them and remove them?

### How to detect outliers in the covariate variables

With data, you can determine the interquartile range by looking at the difference between the 25th percentile and 75th percentile. From there, if you multiplied that range by 1.5 you can use that value to detect outliers. Subtract that value from the 25th percentile and if there is any data still below that, they can be considered outliers. Similarly, you can add that value to the 75th percentile and if there is any data still above that, they can also be considered outliers. If any outliers were detected, you can safely remove them from the data and decrement the number of values considered per outlier removed.

### Can we detect outliers in the noise variables?

Outliers in the noise variables can’t be detected as noise in attribute values can make the data look more randomized or unusual, therefore it is possible that some instances in noisy data will appear as outliers.

## How do they affect the normalization of covariates?

Outliers affect the normalization of covariates as it would transform the data in the wrong way. As it would keep the outliers but change the. Range. For example, if the data points were: 0,2,5,10,15,20,22,24,990,1000, we know that 990 and 1000 are the outliers here. But using Min-max Normalisation, these data-points will transform to 0, 0.002, 0.005, 0.01, 0.015, 0.02, 0.022, 0.024, 0.99, 1.

## Generate data; add a certain proportion of outliers manually. Vary the number and magnitude of outliers.

See code

# Nested Models:

## Overview:

Model A: *y = a1x1+ a2x2+ a3x3+c*

Model B: *y = a1x1+ a3x3 +c*

To quickly explain the concept of Nested Loops, lets consider Model A as our “complete” model. Model B would then be our “reduced” model as it is a subset of Model A. The purpose of having reduced models is to observe data more efficiently, as we can potentially remove terms that don’t affect the data we are observing as we analyse it.

## How to compare nested models?

We can compare tests for a complete model and a reduced model on the same dataset to assess if the results we get are the same/what we want, to then check if we can use the reduced model for efficiency. For this, a popular test to execute is known as the t-test. This is a simple test where we can say the coefficient in the complete model that isn’t present in the reduced model is equal to zero/null, but only if the reduced model is deemed sufficient.

If we have multiple different reduced models from the same complete model, we can compare tests to see which reduced model would be the best to use for accurate results as well as efficiency. For this we would use what is called an F-test, which would give us the statistical differences between a reduced model and the complete model. We ideally want to pick a reduced model with a small difference, while still having few parameters to keep it efficient.

*F statistic = ((RSSR – RSSC) / kR – kC) / (RSSC / (n-kC))*

* *RSSR*: The residual sum of squares of a reduced model.
* *RSSC*: The residual sum of squares of the complete model.
* *kR*: The number of coefficients in the reduced model.
* *kC*: The number of coefficients in the complete model.
* *n*: The total amount of observations in the dataset.

## How to decide how many variables we need as covariates in our model?

You decide how many variables we need as covariates in our model, by performing a correlation analysis to discriminate the collinearity of the variables, usually less than five variables are ideal, since when you use more than five variables the model loses power of explanation and is more likely to produce some unwanted inefficiency.

## How can we use nested models to quantify the importance of a particular covariate?

We can compare results of each covariate for a model. If there were to be a small statistical difference from a model with a covariate when compared to a model without that covariate, then we can discard that covariate as it can be deemed too insignificant to keep. If there were to be a large difference however, then it would have to be deemed as important to keep as it would mean the result is more dependent on that covariate.

## Can we compare the coefficients obtain for different nested models?

If coefficients are deemed as too significant to dismiss, we can also assess the degree of change in the final result for every unit of change in the covariate. If a simple t-test illustrates the coefficient to be significant, we can still further assess it. Depending on if the coefficient is a positive or negative value, I can conclude whether increasing the respective covariate increases or decreases the final result with it being positive increasing the final value, and it being negative decreasing the final value.

### Is there some relation between the *a*’s, *b*’s and *c*?

The models we are given are:

Model A: *y = a1x1+ a2x2+ anx3+εy*

Model B: *y = b1x1+ b3x3 +ε’y*

Model C: *y = c3x3+ε”y*

### How the dependencies between the covariates affects these relations?

## Discuss the difference between interpretating the importance of each covariate in terms of the magnitude of its coefficient or using nested models

The difference between interpreting the importance of each covariate in terms of the magnitude of its coefficient or using nested models is that interpreting is when, the coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant. This property of holding the other variables constant is crucial because it allows you to assess the effect of each variable in isolation from the others.