Programming and Mathematics for AI Coursework

Task 2

By Muhammad Faaiz Ansari

Table of Contents

Task 2 Description:

Study the effect of 2 factors that affect the performance and interpretation of a Linear Regression Model.

-Effect of Outliers:

-What is an outlier: An outlier is a value that is an abnormal distance away from the majority of the other values in a random sample from a population. An outlier value may be caused due to variability in measurement of variables; could be due to sampling error; or it could just indicate a valid standout result that might be interesting to investigate in more detail as to why it so different. If an outlier is brought about due to an error, it may be excluded from a data set.

-How they affect the estimation of the coefficients:

An outlier skews a correlation coefficient. In most cases, an outlier generally decreases the value of a correlation coefficient and weakens the regression relationship. In some cases, however, an outlier could even unfairly or incorrectly increase a correlation coefficient and strengthen a regression relationship, more than they should be in a true or fair case.

In terms of linear regression relationships, outliers only serve to skew and reduce the coefficient of determination and make the fit worse. This may be a bit different for cases exploring non-linear relationships.

-How can we detect them and remove them:

-How to detect outliers in the covariate variables:

First, to clarify, A variable is a covariate if it is related to the dependent variable. A covariate is a possible predictive or explanatory variable of the dependent variable. In regression analysis, independent variables (the regressors) are called covariates.

Generally, outliers can be found in our independent variables. The simplest way to detect outliers can be to plot your data and visualise it in a graph. Most outliers can be seen as they are observably far away from the other illustrated values. A more direct way to identify outliers is to add and subtract 1.5x Interquartile Range (IQR) from the first quartile and 3rd quartile. Any values falling outside this value may be considered an outlier. We could alternatively use Standard deviationx2 and all points that fall outside of it may be considered outliers. Now that all the outliers have been identified, they can be isolated from the rest of the data and removed. In python this can be done by dropping the outliers from a pandas data frame of all values by using a query and .drop functionality.

-Can we detect outliers in the noise variables:

First, to clarify, statistical noise variables refer to variability within a sample, random unexpected errors in a regression equation, or estimation error. This noise is represented as a random variable. Noise can be defined as mislabelled examples (class noise) or errors in the values of attributes (attribute noise).

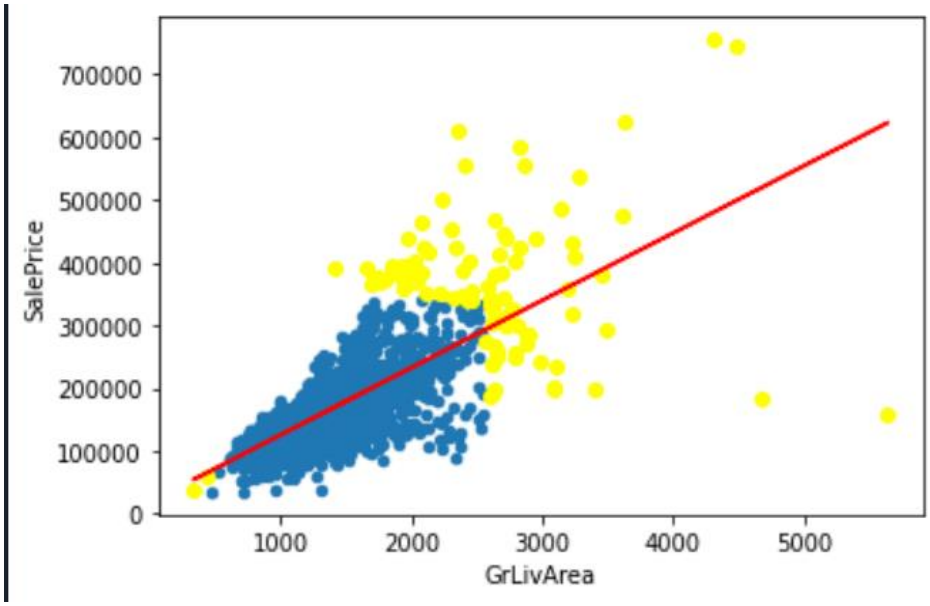
So can we detect outliers in the noise variables. Yes, but not specifically. Noise variables are just part of the attribute data (attribute noise). We can detect outliers from all the independent variables attributes. This includes noise and non-noise variables within an attribute in the data. Furthermore, outliers can include noise and valid discordant data points.

-How do they affect the normalization of covariates: The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

Outliers can cause normalization to be skewed. This is because outliers can affect the size of the ranges of values. For example, most of our values for a specific variable fall within 1-100 range but we have an outlier at 300, if this is included when the data is rescaled, the normalisation will treat it as though our data ranged from 1-300. Once the data has been normalised, it will consequently, skew the difference between most values for that variable, making them look comparatively much smaller than they were originally.

A way this can be overcome is using more robust scaling techniques using percentile data for numeric variables

-Output from code showing outliers based on method to identify them I mentioned above, below. yellow dots are outliers:



-Effect of Normalization of covariates:

-what is normalisation: Normalization is a transformation technique which computes the mean and standard deviation of the covariate and computes a new covariate by subtracting each covariate value from the mean and dividing by the standard deviation to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information. It also changes the distribution of the data to be more normal.

-Is the probability distribution of all covariates the same:

The probability distribution of covariates is changed by normalisation to be more similar to normal (gaussian) distribution in shape after normalisation.

-Is the scale of the covariates the same:

One of the main outcomes of normalisation of covariates is to transform them to all use a common scale, but without distorting differences in ranges of values or losing information. This means all the covariates are adapted to fit a new commonly shared scale.

-How normalising the covariates affects the value of the coefficients:

Whether we use normalized covariates or non-normalised covariates doesn't affect the statistical significance of coefficients. The Idea is to replace y and each x variable with a normalised version. Coefficient reflects standard deviation of y for a one standard deviation change in x.

The coefficients describe the nature of the relationship between two variables regardless of the range and the measurement units of them.

-Can we compare the magnitude of the coefficients of different covariates?

Yes. Once the covariates have been normalised, we can compare the magnitudes of the normalised coefficients and conclude which variable is most important or etc. The normalised coefficients are easily comparable as they are independent of units of measurement. An issue to be aware of is that confounders may affect results more if they are involved in rescaling of other covariates.

-What is the difference between using standardisation as a form of normalisation or

other options (e.g. Quantile-based normalisation). compare effect and pros and cons of different options of normalisation. How are types of normalisations affected by outliers:

-Standardisation:

-Mean and standard deviation is used for scaling.

-It is used when we want to ensure zero mean and unit standard deviation

-It is not bounded to a certain range.

-It is much less affected by outliers, so can be run without removing them.

-It is useful when the feature distribution is Normal (Gaussian).

-Other Normalisation (Scaling normalisation):

-Minimum and maximum value of features are used for scaling

-It is used when features are of different scales.

-Scales values between [0, 1] or [-1, 1].

-Strongly affected by outliers so they must be removed to use it.

-Useful when distribution isn't known and changes distribution to be more like normal (Gaussian)

-It is also more affected by the presence of potentially confounders, resulting in higher false-positive and false-negative rates.

Overall, scaling normalisation is much more widely usable as it can be used without knowing the covariate distribution and can make it more, and it can be used on covariates of different scales. Its main drawback is it is more affected by confounders and outliers.

Reference Links:

-https://online.stat.psu.edu/stat800/lesson/cautions-about-correlation-and-regression#:~:text=Influence%20Outliers,correlation%20value%20and%20improve%20regression.

-https://medium.com/@aatl2012/the-basic-difference-between-noise-and-outliers-in-data-cd3ff32343e0

-https://medium.com/@isalindgren313/transformations-scaling-and-normalization-420b2be12300

-http://www2.kobe-u.ac.jp/~kawabat/ch06.pdf

-https://towardsdatascience.com/normalization-vs-standardization-quantitative-analysis-a91e8a79cebf

-a dataset and my code learned from a couple principles of datascience IN3061 labs used as base for code section and plot.