UNIVERSITY OF GOTHENBURG

CHALMERS

MSG 500
Mini Analysis Report

Fionn Delahunty
Applied Data Science Program
Gusdelfi@student.gu.se

- **Mini 1**

For the king country house price dataset, I choose to look at square foot of interior housing space above ground as a predictor of house prices. The initial plot was heavily concentrated towards to the lower left hand side, and did not display a linear trend. I applied three transformations (Log10, square root and exponential) to the square foot of living. I decided that the log 10 transformation gave the best linear fit. In retrospect I did not experiment with transforming price, which was a hindsight on our part.

Following that transformation, I investigated if the model fitted the remaining four assumptions. A residual plot suggested that the model did not have equal errors around zero, with disproportionate positive amount. This also effected our constant error variance, where I did not see an even spread along the fitted line. Finally, I noted some outliers in the upper left side, but given the sample size I didn't feel these where practically influential.

Our final set of summary statistics suggested that square foot of interior housing did positively contribute to the model, the overall model had was able to significantly account for 34% of the variance in house prices.

The take home message was although I did find violations of the assumptions, they where not seriously worrying. Depending on the application of the model I could probably accept that it was a sufficient model for most purposes.

- **Mini 2**

Following an initial investigation of building the prediction model with all fitted variables I quite quickly identified that there was a strong presence of multicollinearity. It was also at this plotting stage that I performed a number of log transformations to create a better fit for the data. In this case almost all the variables where log transformed.

I then performed a standard backward selection criterion and came to the conclusion of selecting four variables for our final model. This model included

- Log (sqft_living)
- Log (sqft_living15)
- Log (bedrooms)
- Log (sqft_lot15)

This model itself had a significant prediction value of 51.23%. However, when dealing with multiple regression (or any model building set up) a much better estimate of the prediction power is to test the model on a hidden or test dataset.

- **Mini 3**

For mini analysis three, our chosen subject was outliers. Here I tested the effect of outliers position in the dataset on the overall adjusted R^2 and how I could indefinite these outliers via the cook distance plot. I worked with a small section of the king country house price dataset and manually added in outliers.

In figure 1 below I summaries the effect of outlier position on Adjusted R^2 . The important take home messages are that extrapolation increases the Adjusted R^2 although in the case of a single variable this might be infatuated value and a misrepresentation of the actual prediction power of the model. In the case of outliers to either the upper sides of the graphs it can be seen that these reduce the Adjusted R^2 to different levels. Importantly a cluster of outliers will have less effect then a scatter of outliers, which makes sense given the nature of how a regression line is fitted. In the case of a cluster of outliers in real data I would often consider if this was a group in the dataset.

Location	Adjusted R^2
No outlier	41%
Single extrapolation to the upper right side	47%
Multiple extrapolation to the upper right side	65%
Multiple extreme outliers scattered (upper)	18%
Multiple extreme outliers clustered (upper)	32%
Multiple extreme outliers clustered (lower)	1%

Figure 1. Summary of outlier location on adjusted R^2

Apart from plotting our data, our main identification of influential outliers is a cook distance chart, in which I look for values that are extremely outlying compared with the other lines in the plot. In all case the single outlier is easier to detect then a cluster of outliers. In general cluster can be harder to detect from this chart if there is a gradual increase in values then a single extreme. Depending on it's location it can almost appear if all values are slowly increasing. Our take home message here is that experience of comparing cook distance charts to actual scatterplots of data is required for a good understanding of how they truly represent data.

In the above cause, I also briefly considered the question of what is an outlying value. For the purpose of our little experiment, I considered a outlier was a value that was at least 2 values (sqft_living) higher than any other value on the dataset.

- **Mini 4**

For this mini analysis I choose to work on my own dataset, the same one was I was for my individual project. I attempted to build a projection model that would look that predicting the number of cigarettes an individual smoke per week.

I choose five variables to include in the model, my first major task was dealing with sorting the variables into levels and group. Our prediction variable smoking, had values from 0 to 420 cigarettes smoked per week. With the majority of those values being zero. To experiment I started working only on those who had smoked (above zero). I then grouped the data into five levels. I moved ahead with this.

Next, I employed a regression tree to compare how it selected the variables. The model the regression tree suggested had a higher significant (Adjusted R^2) value then a standard backward selection model. However, it's prediction value was considerable lower. This was a question that me and my group to then considered, was a higher significant value for a poorer prediction model better than a lower significant for a model that predicted things better? Regardless of that question, this model introduced a number of outliers into our graphs, so I disregarded it anyway.

My third comparison was trying to build the model using principal component analysis. Before I began comparing I noticed a number of violation of the assumptions, including outliers, poorly fitted residuals and poorly spread residuals. I performed some transformations I fixed some issues, but overall, I still had a poor fit with the variables this model suggested.

In conclusion our best model was selected by backward selection. Regression trees seemed promising while principal component analysis did not work.