Kevin Carone

Upgrad ML/AI

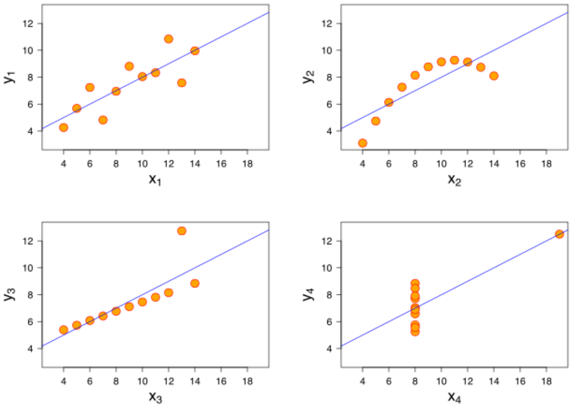
Multiple Linear Regression Subjective Questions

1. The summer months definitely had more bike rentals, while the colder months had the least. Same with the summer season having more, and the winter season having less. However some days in the winter weren't so cold, and on these days that was irrelevant. The primary variable was numerical, temperature. The weather situation had a dramatic effect top, obviously most people don't want to ride a bike in the rain unless they have to. The workingday and weekday categorical variables seemed to have no effect at all from just the exploratory data analysis. The year one and year two had a dramatic effect as well, being that it is a new technology it is growing - the second year had much more users.
2. It removes the first dummy variable which is redundant. The lack of the other variables can tell you the effect of the first variable. It also reduces clutter which can be important when you are creating dummy columns for 5+ categorical columns
3. The highest correlation on the pairplot with the target variable count, was registered and casual users. However these don't really count because they are all dependent on eachother. The highest correlation variable which was independent of count was temperature.
4. A error distribution chart was created which showed a normal distribution of the errors. From the start, the data was split into training and test data 70/30, so that the model created on the training data could be tested on the test data. The final plot had a very beautiful linear relationship.
5. In the final model the temperature was the most important variable with a 0.548 correlation, followed by an inverse relationship with the weather at -0.248, followed by the year at 0.233. For business purposes the year is not a very useful variable because the year is the same no matter what location we choose for the next bike rental city. In which case the third most important variable would be the wind with a -0.15 correlation.

# General Questions:

1. The goal of linear regression is to find the equation of a line which will go through a scatter plot whereby the sum of squares of residuals above and below that line are minimized. The residuals are the y coordinates of the actual data subtracted by the predicted y coordinates of the linear model.

2. Anscombes quartet is a plot of four graphs which have the same linear regression model applied to it, however its obvious in the visuals that these data are very different. The first is a good linear model, the second is a polynomial model with a linear model applied to it, the third is a linear model applied even with some major outliers, and the fourth is a linear model applied to categorical data. It's plausible that if you use linear regression in the wrong scenario you could end up with any of the other three plots. You should only use ordinary linear regression when the data is not like the other three.



https://en.wikipedia.org/wiki/Anscombe%27s_quartet#/media/File:Anscombe's_quartet_3.svg

3. Pearson's R is the pearson coefficient which is a number between -1 and 1 which tells you the direction and strength of a correlation. The more scattered the data is around a linear model, the lower its correlation. With a random plot of points having zero correlation and a straight line having 1 correlation. If the data isn't normally distributed you could try the spearman's coefficient.
4. Scaling is when a data set is scaled that all the variables we are working with can be viewed on the same scale. Especially useful when comparing numerical values with categorical values. Normalized scaling is when the numerical data is scaled between 0 and 1. Standardized scaling is when the data is scaled to a mean of zero and a standard deviation of 1
5. This could mean that two columns being analyzed are actually identical or close to it. It means that there is perfect correlation between two columns being analyzed
6. A Q-Q plot is a quantile-quantile plot, two datasets and their quantiles are plotted against eachother. First the data needs to be ranked in non-decreasing order. If

the relationship is linear then the data can be modeled by the target distribution. In a way it compares the histograms of two datasets. The Q-Q plot is a parametric curve from 0 to 1 in the R^2 plane.