**Exercise 1**

Open the file **World Happiness 2019.csv** again. This dataset has a “happiness score” which was measured by simply asking the people in the country how happy they are on a scale of 0 to 10 and then calculating the average per country. It also contains many other variables which might be related to happiness. The dataset was compiled by the Sustainable Development Solutions Network (2020).

1. Run simple regressions with Happiness.Score as the dependent variable and only independent variable at a time. Summarize the results.

All of the variables are statistically significant and positive except for generosity

1. Check if there is any multicollinearity between the variables (except for Happiness\_Score).

None of the VIF values are higher than 5 so there should not be any serious multicollinearity problems.

1. Run a multiple regression with Happiness.Score as the dependent variable and all of the independent variables in a single model. Are there any differences compared to when you ran the simple regressions?

The coefficients have becomes smaller. Perceptions of corruption is no longer significant.

**Exercise 2**

Open the file **winequality-red.xls**. You don’t have to do any data cleaning. This dataset contains data of the Portuguese “Vinho Verde” wine which also used in the study by Cortez et al (2009).

1. Check if there is multicollinearity and deal with it in case it is a problem.

Fixed\_acidity and density have VIF scores higher than 5. If you remove one of them the problem is solved. I removed density, but you could also choose to remove the other one.

1. Carry out a regression analysis in which you try to predict the quality of the wine using all of the other variables. What do you see?

From the model it can be seen that free sulfur dioxide, sulphates, and alcohol have a positive effect, volatile acidity, chlorides, total sulfur dioxide, and pH have a negative effect. The other variables are not significant so we can’t say anything about their effects. The R-squared is 0.36 indicating that 36% of the variance in the quality can be explained by these factors.

1. Run a new model but this time exclude the variables which were not significant in the first model. Are there any changes to the coefficients and R-squared?

There are hardly any changes, all of the effects are in the same direction and remain significant. This shows that variables which are not significant often have a negligible effect on the final outcome, although it’s always important to test them anyway. When presenting the final model in a paper it’s best to present a model which does not contain variables which are not relevant, simplicity is usually preferred. Of course you should mention that you at least tested the other variables but found no effect.

**Exercise 3 (more advanced)**

Open the file **GBvideos.xls**. This dataset contains a large amount of YouTube videos which have been trending in the UK. You want to see if you can predict how well received a video is by looking at the number of views and the number of comments.

1. Why does it not make sense to simply predict the number of likes based on the number of views and comments?

The goal is to look at how well-received a video is. The number of likes will always be positively correlated to the number of views and comments simply because videos with more likes receive more attention. So this won’t tell use much. Instead, we need to take into account the number of dislikes as well.

1. Filter the data so that you only include videos with at least 1 like and 1 dislike. Also only use videos which have not blocked their rating, not blocked comments, and have not been removed.

The answers are in the script.

1. Create a variable, like\_percentage that measures the total number of likes divided by the total number of likes and dislikes

The answers are in the R script.

1. Check if there is multicollinearity between the variables.

There appears to be none, computation is in the script.

1. Run the regression model where you predict the like\_percentage using views and comments. What do you think these results could mean?

Views has a positive relationship while comments has a negative relationship. So apparently videos that are well received get more views but videos which are less well received get more comments. Perhaps very controversial videos create a lot of discussion?

**References**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Sustainable Solutions Development Network (2020. *World Happiness Report 2020*. Retrieved from: https://worldhappiness.report/ed/2020/