# Multivariate Regression

## Assumptions

There is no significant correlation between the regressed independent variables

We live in a dynamic world where every factor is probably interrelated in some way that is too complex to be predicted. We chose variables to be as distinct as possible to minimise the cross-talk between factors.

Multicollinearity is a statistical phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other. In other words, multicollinearity indicates a strong linear relationship among the predictor variables.

<https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/>

Multicollinearity is a problem because it undermines the statistical significance of an independent variable.

<https://link.springer.com/chapter/10.1007/978-0-585-25657-3_37#:~:text=Multicollinearity%20exists%20whenever%20an%20independent,significance%20of%20an%20independent%20variable>.

However, the good news is that you don't always have to find a way to fix multicollinearity.

<https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/#:~:text=Multicollinearity%20makes%20it%20hard%20to,a%20way%20to%20fix%20multicollinearity>.

## Model Development

Could use something analogous to K-fold cross validation. Split the data set into subsections and then iteratively regress across data chunks, then verify using MAE, and MSE on predicted data from model and test data in the fold. Then do

General linear models.

By convention, least-squares method is used to fit a line to the data. We MAY use LAD. Calculate p-value for R^2.

LAD method:

Briefly describe squared residual thing. Instead, we aim to minimise (see reference at the bottom):



LaTex:

\min\_w \lvert Xw - y \rvert

This choice aims to minimise the mean absolute error (MAE) and mean squared error (MSE) metrics of each fitted model on the test set.

Least-squares method:

Minimising the squared residuals. R^2 tells us what % of the variation in Y variable can be explained by X variable

Calculating R^2:

1. Measure, square and sum the distance from the data to the mean
2. Measure, square and sum the distance from the data to this 

SS(fit) is the sum of squares around the least-squares fit. Which is the sum of the distances between the data and the line, squared.

Doing a least-squares fit in higher dimensional data yields the corresponding hyperplane, instead of a line. Least-squares will estimate n different parameters, including an intercept. Predictions of our models are represented as the set of vectors that satisfy the hyperplane described by our least-squares fit.

The more silly parameters that are added to the equation, the more opportunities there are for random events to reduce the SS(fit) (due to unseen correlations (see assumption at the top)) and thus yield a better R^2. Thus, there is an adjusted R^2 value that scales by num of parameters.

Determining if R^2 is statistically significant => p-value (example 2 data points always yields R^2=1)

Consider function F, defined s.t.

F = variation in predicted variable explained by predictor variable/variation in predicted variable not explained by it



p\_fit= number of parameters estimated by the least-squares

P\_mean = number of parameters in the mean line ( 1)

N = number of data points

The more parameters that are in the equation, the more data is required to estimate the parameters.

If the fit is good, the variation explained by the extra parameters in the fit should be higher than the variation not explained by the extra parameters in the fit. Thus F should be a large number.

Generate a set of random data, calculate F. SEE STATQUEST FOR LINEAR REGRESSION

FOR MULTIVARIATE DATA (FOR DETERMINING THE USEFULNESS OF INCREASING NUMBER OF PARAMETERS)



Use this equation to find F, where simple is the values generated by the simple regression model.

If the difference in R^2 values between simple and multiple regression is large, and p-value is small, then adding an extra variable to the model is worth the cost in model complexity.

You can do this generally, with n variables, going to n+1 variables (multi to even more multivariate).

ALWAYS PLOT DATA IN A GRAPH.

## Advantages

## Disadvantages

Outlier Onset. Linear regression models are highly sensitive to outliers or high leverage points. If any of these variables undergo substantial changes that develop an outlier data point, it would disrupt the efficacy of the model.

In OLS, a large residual for a single data point makes an outsized contribution (to the loss function). This may cause predictions to be skewed (in context of ML).

Reference:

https://gurobi-optimization-gurobi-optimods.readthedocs-hosted.com/en/latest/mods/lad-regression.html

Furthermore, the linear assumption is not ideal because clearly not all of the relationships would necessarily be linear. A more in depth use of the model would include translations where necessary

## Sensitivity Analysis

# **Meyer Notes**

* For both - Firstly scale the x and y axis
* Linear multivariate regression
  + Uses sklearn’s LinearRegression function - minimises least squares
  + Produces equation in the form: y=ax1+bx2…+const
* General multivariate regression
  + Uses scipy’s curve\_fit function - minimises least squares
  + Produces equation where “ax” in regular equation is replaced with “ax1+bx2…”
  + E.g. Ae^(ax) -> Ae^(ax1 + bx2 …)
* I think the coefficients have some meaning but I can’t be bothered to search it up so figure that one out