Multiple linear regression

In simple linear regression, we attempt to find a line of best fit such that $y = \beta_0 + \beta_1 x_1$, i.e. our only independent variable is x_1. Multple linear regression is a generalisation to include more variables (in total $m + 1$ variables). We then have:

$$y = \sum_{i=0}^{m} \beta_i x_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$$

But remember that we have n such equations, where n is the number of observations. All of this information can be simply encoded into matrices:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{Y}$ is an $n \times 1$ matrix of predicted results.
- $\mathbf{X}$ is an $n \times m$ matrix, where each column is a different variable and each row is a different observation.
- ϵ is an error term, which is an $n \times 1$ matrix.

We can find β by minimizing $(\mathbf{Y} - \mathbf{X}\beta)^2$, which will eventually give us:

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

where $\mathbf{X}'$ denotes the transpose of $\mathbf{X}$.

Assumptions

1. Linearity
2. Homoscedasticity - the variance around the regression line doesn't change.
3. Multivariate normality - essentially, this criterion is met if every x_i can be treated as a sample from some normal distribution.
4. Independence of errors
5. Lack of multicollinearity - sometimes, one of our features may be able to be linearly predicted from another. This will affect analyses of individual predictors, but not affect our model as a whole.

Dummy variables

Categorical variables are dealt with by introducing dummy variables which can take values of 0 or 1.

Note that if we have n categorical variables, we will only require $n - 1$ dummy variables. This is because the constant β_0 *takes into account* the other variable. This can be clearly seen for the case of two categories. If a datapoint belongs

to the first category (i.e $D_1 = 1$), we immediately know it doesn't belong to the other ($D_2 = 0$).

$$\begin{aligned} y &= \beta_0 + \beta_1 D_1 + \beta_2 D_2 \\ &= \beta_0 + \beta_1 D_1 + \beta_2 (1 - D_1) \\ &= (\beta_0 + \beta_2) + (\beta_1 - 1) D_1 \\ y &= \beta_0' + \beta_1' D_1 \end{aligned}$$

Building a model

Often, we will have some independent variables that are poor predictors of our dependent variable, which should be removed to improve our model. There are a few different methods for doing so. In most cases we first need to choose a significance level (SL), which we will set to 0.05.

Backward elimination

1. Fit the model with all possible predictors
2. While the worst predictor has p-value > SL, remove that predictor and re-fit the model.
3. Once all the predictors have p-values < SL, the model is ready.

Forward selection

1. Fit all single variable regression models, and select the one with the lowest p-value and discard the remainder.
2. Fit all possible models with one extra predictor on top of the first one. Keep the predictor with the lowest p-value, if that p-value < SL.
3. Repeat the above until it is impossible to add another predictor with a p-value < SL.

All possible models

The obvious solution is to fit all possible models. The only problem is that for m variables, the number of possible models is:

$$\binom{m}{1} + \binom{m}{2} + \ldots + \binom{m}{m} = 2^m - 1$$