DSA3361
WorkShop 4 – Review

Outline

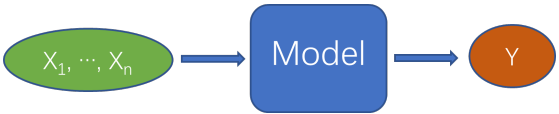
1. Linear Regression Model

2. Some practical issues

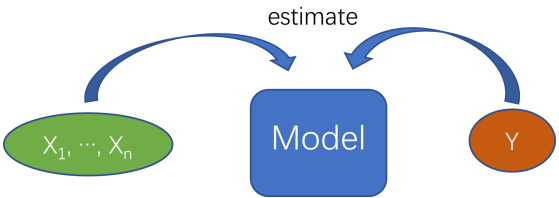
Linear Regression Model

What is regression?

- (From Wikipedia) In statistical modelling, **regression** analysis is a set of statistical processes for estimating the relationships between a dependent variable Y (often called the 'outcome' or 'response' variable) and one or more independent variables $X_1, \cdots, X_n$ (often called 'predictors', 'covariates', 'explanatory variables' or 'features').



- Given a set of observations (*Data*) of the predictors $X_1, \cdots, X_n$ and corresponding response variable Y, and a *hypothesized functional form* for the postulated relationship between X_is and Y, regression is then to **estimate parameters** of the function that best fit the data.



What is regression for?

(From Wikipedia) Regression analysis is primarily used for two conceptually distinct purposes:

- First, regression analysis is widely used for prediction and forecasting.
- Second, in some situations regression analysis can be used to infer relationships between the independent and dependent variables.

Linear regression model

- Linear regression models assume **linear** relationship between the (*continuous*) independent and dependent variables:
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n.$$

- Given a set of data points, we are to estimate parameters
$$\beta_0, \beta_1, \cdots, \beta_n$$
that best fit the given data.

- Several assumptions are to be checked to validate that the *linear* assumption is reasonable.

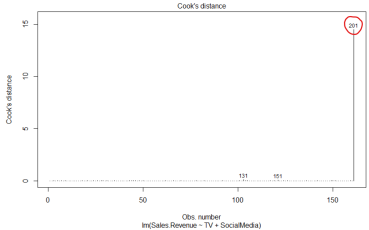
Evaluating the model

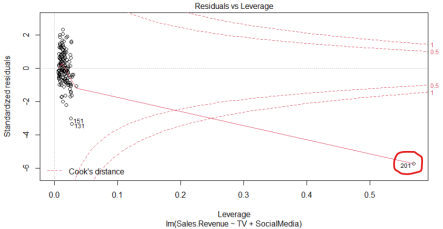
- Evaluate the model by comparing the difference between the actual and predicted response variables.
- Some commonly used measurements:
  - MSE (Mean Squared Error): $\sum(\text{actual}\,Y - \text{predicted}\,Y)^2$, the smaller the better
  - MAE (Mean Absolute Error): $\sum|\text{actual}\,Y - \text{predicted}\,Y|$, the smaller the better
  - (Adjusted) R^2: between 0 and 1, the larger the better (As a rule of thumb, an $R^2 \geq 0.7$ is interpreted as an acceptable model.)
- Use **training set** to fit the parameters, and use **test set** to evaluate the performance of the model on unseen data.

Some practical issues

Data issues

- Data cleaning, such as missing data, duplicates, etc.
- Outliers/influential points
  - Cook's distance: observations with large values are considered to be influential points.
  - Residual vs leverage plot: Any point falling outside of Cook's distance (the red dashed lines) is considered to be an influential observation.





Variable issues

- Multicolinearity between predictors: Variance Inflation Factor (VIF)
  - A VIF value of 1 indicates that a particular variable is uncorrelated with all the other predictors
  - A VIF value > 5 indicates a high multicollinearity.
- Applying appropriate variable transformation
- Introducing interaction between predictors
- Creating dummy variables for *categorical* predictors
- Variable selection: Forward/Backward step-wise selection, etc.

Model issues

The '**linear**' assumption may not be satisfied:

- Using variable transformation or interactive variable
- Using additive models, such as quadratic, cubic, etc. models.
- Using more complicated **nonlinear** models

Practical issues

- Be cautious of actions on data.
- Including more predictors generally help. However, too many predictors might lead to overfit.
- Expert knowledge is important, especially with limited data.
- **Randomly** split the training/test datasets.