Data Analytics (UE18CS312)

Unit 2

Aronya Baksy

October 2020

1 Regression Analysis

- Regression is the task of finding the existence of an **association relationship** between a **dependent** variable (aka response or outcome variable, represented as Y) and an **independent** variable (aka explanatory or predictor variable, represented as X).
- Regression does not say that the current value of Y is dependent on the current value of X, or is **caused by** the current value of X.
- Regression only aims to find that there is an *association* between changes in Y and changes in X.
- Regression is a form of **supervised learning**, in that it requires knowledge of both dependent and independent variables in the dataset.

2 Simple Linear Regression

- In SLR, there is a *linear relationship* between the dependent variable Y and the regression coefficients β_0 and β_1.
- The simplest functional form of SLR model can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

  Where
  Y_i is the value of the dependent variable for the i^{th} observation in the data
  X_i is the value of the independent variable for the i^{th} observation in the data
  β_0 and β_1 are the regression coefficients
  ε_i is the error term or **residual** in prediction for the i^{th} observation.

- **Note**:
  Models like $Y_i = \beta_0 + \beta_1 log(X_i) + \varepsilon_i$ and $Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i$ are also **linear models**.
  But models like $Y_i = \beta_0 + \frac{1}{1+\beta_1} X_i + \varepsilon_i$ and $Y_i = \beta_0 + e^{\beta_1} X_i + \varepsilon_i$ are **non-linear models**.
- Determining the appropriate functional form is important to get a good fit with low error for the given data.

2.1 Parameter Estimation: OLS

- The **Ordinary Least Squares** method is used to estimate the regression parameters β_0 and β_1.
- OLS provides the **Best Linear Unbiased Estimate** (BLUE) of the parameters. The condition for an estimate to be BLUE is:

$$\mathbb{E}(\beta - \hat{\beta}) = 0$$

  Where $\hat{\beta}$ is the estimated value of the parameter β.

- OLS is guaranteed to provide the best fit line under the following assumptions:
  1. The model is linear wrt. the regression parameters β_0 and β_1
  2. The exploratory variable X is deterministic, not stochastic
  3. The conditional expected value of all residuals is 0, ie. $\mathbb{E}(\varepsilon_i|X_i) = 0$
  4. For time series data, residuals are uncorrelated, ie. $Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall \ i \neq j$
  5. The residuals ε_i follow a normal distribution
  6. The variance of the residuals ε_i does not depend on the value of X_i and it is a constant value, in other words, X is **homoscedastic**.
- The OLS estimate of the parameters β_0 and β_1 are:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{2}$$

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} X_i(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n} X_i(X_i - \overline{X})} \tag{3}$$

- This BLUE is obtained by minimizing the sum of squared errors (SSE).
- Sum of Squared Total variations (SST) is

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

- Sum of Squared Errors (SSE) is

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

- Sum of Squared error due to Regression (SSR) is

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 \tag{4}$$

- Therefore

$$SST = SSR + SSE$$

- And the **coefficient of determination**, r^2 is given as

$$r^2 = \frac{SSR}{SST} = \frac{\sum\limits_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum\limits_{i=1}^{n} (Y_i - \overline{Y})^2} = 1 - \frac{\sum\limits_{i=1}^{n} (Y_i - \hat{Y})^2}{\sum\limits_{i=1}^{n} (Y_i - \overline{Y})^2} \tag{5}$$

3 Hypothesis Tests for Regression Coefficients

3.1 t-Test for Regression Coefficient

- **Null Hypothesis** H_0: There is no linear relationship between X and Y
- **Alternate Hypothesis** H_1: There is a linear relationship between X and Y

- The standard error for the estimate $\hat{\beta}_1$ is:

$$S_e(\hat{\beta}_1) = \frac{\sqrt{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/(n-2)}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- The test statistic t with degrees of freedom $n-2$ is then calculated as:

$$t = \frac{\hat{\beta}_1}{S_e(\hat{\beta}_1)}$$

3.2 F-test for overall model: ANOVA

- The f-score is given as

$$f = \frac{SSR}{SSE/(n-2)} = \frac{R^2(n-2)}{(1-R^2)}$$

4 Outlier Analysis

The distance measures used in observing outliers are:

- **Z-score**: given by

$$z = \frac{\hat{Y} - \overline{Y}}{\sigma_Y}$$

- **Mahalanobis Distance**: Given by

$$D_m = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}$$

- **Minknowski Distance**: between $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$ is given by

$$D(X,Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

  For $p = 1$ this is the **Manhattan Distance**, and for $p = 2$ this is the **Euclidean Distance**.

5 Multiple Linear Regression

- The most basic functional form of an MLR model is given as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i \tag{6}$$

- Let X be a matrix with the first column all 1s, and the next column being the values of the first feature X_1 and so on. Then the matrix form of MLR is

$$Y = X\beta + \varepsilon \tag{7}$$

- The OLS estimate of β is then given by

$$\hat{\beta} = (X^T X)^{-1}(X^T Y) \tag{8}$$

5.1 Auto-Correlation

- Auto correlation is the correlation between successive error terms in a time-series regression problem. Given a time series model
$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$
- If there is auto correlation, the standard error estimate of beta coefficient will be underestimated, which leads to a low p value. Hence a variable that has no statistically significant relationship with the response variable y may be accepted because of auto correlation.
- Auto correlation can be tested using the **Durbin-Watson Test**

5.2 Durbin-Watson's Test

- If ρ be the correlation between the residual terms ε_t and ε_{t-1}, then:
  - **Null Hypothesis** H_0: $\rho = 0$
  - **Alternate Hypothesis** H_1: $\rho \neq 0$
- The test statistic D is given as
$$D = 2\left(1 - \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}\right)$$
- Given the upper and lower limits D_U and D_L on the test statistic D, we have
  - If $D < D_L$ then errors are +vely autocorrelated
  - If $D > D_L$ there is no evidence for +ve autocorrelation
  - If $(4 - D) < D_L$ then errors are -vely autocorrelated
  - If $(4 - D) > D_U$ then no evidence for -ve autocorrelation
  - If $D_L < 4 - D < D_U$ or $D_L < D < D_U$ then test is inconclusive

5.3 DFFIT and DFBETA

- DFFIT and DFBETA are the values of the repsonse variable and beta coefficient when one observation i is removed from the data.
- DFFIT is given by
$$DFFIT = \hat{Y}_i - \hat{Y_{i(i)}}$$
$$DFBETA_i(j) = \hat{\beta}_j - \beta Y_{j(i)}$$
- $DFBETA_i(j)$ represents the change in the coefficient of variable X_j when observation i is removed.
- The standardized versions SDFFIT and SDFBETA are also used, standardized by the standard error $S_e(\hat{\beta}_j)$

5.4 Bias-Variance Tradeoff in MLR

- The **bias** is the difference between the actual population value of an estimator and its expected value. It measures the accuracy of the estimates.
$$Bias(\hat{\beta}) = \mathbb{E}(\hat{\beta}) - \beta \tag{9}$$

- The **variance** measures the spread, or uncertainty, in these estimates.

$$Var(\hat{\beta}) = \frac{E^{'}E}{n-m} \tag{10}$$

where E is the matrix of residuals given as $y - X\hat{\beta}$, n is number of observations and m is number of independent variables.

- The OLS estimator that is used for estimating regression coefficients is unbiased, but it can have high variance in the cases where there are lots of predictor variables (X), the predictors are highly correlated with one another, or when $n - m$ tends towards 0.

- Solution to the above is to reduce variance at the penalty of introducing some bias. This is called **regularization**.

5.4.1 LASSO Regression

- Least Absolute Shrinkage and Selection. Uses L1 norm or the 'absolute value' of coefficients scaled by shrinkage.

- LASSO tends to zero out smaller (unimportant) coefficients (and helps with feature selection)

- With LASSO, the MLR objective function is now

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{11}$$

- Disadvantages of LASSO:

  1. In small-n-large-m dataset the LASSO selects at most n variables before it saturates.
  2. If there are grouped variables (highly correlated between each other) LASSO tends to select one variable from each group ignoring the others

5.4.2 Ridge Regression

- Uses L2 norm or the squared value of coefficients scaled by shrinkage. It is used when number of predictor variables in a set exceeds the number of observations, or when a data set has correlations between predictor variables.

- We shrink the estimated association of each variable.

- With LASSO, the MLR objective function is now

$$y = \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{12}$$

- Coefficients produced by OLS are scale invariant but that is not the case with Ridge Regression, so we must remember to scale the input

6 Logistic Regression

- A binary logistic regression model is given as

$$P(Y = 1) = \frac{e^Z}{1 + e^Z}$$

Where

$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

- The above equation can be transformed into

$$Z = ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) \tag{13}$$

- The quantity $\frac{P(Y=1)}{1-P(Y=1)}$ is called the **odds**, and the natural logarithm of the odds is called the **log-odds**.
- There is no closed form solution when OLS is used. Hence numerical methods like **gradient descent** are used to estimate the parameters of the Logistic Regression Model.

6.1 Contingency Table and Metrics

- A confusion matrix for a binary classification problem is as follows

	Predicted True	Predicted False
Actual True	TP	FN
Actual False	FP	TN

- The following metrics are defined from these quantities:

$$\begin{aligned}
Accuracy &= \frac{TP}{TP + TN + FP + FN} \\
Precision &= \frac{TP}{TP + FP} \\
Recall &= \frac{TP}{TP + FN} \text{ also known as sensitivity} \\
Specificity &= \frac{TN}{TN + FP} \\
F_1 &= \frac{2 * Precision * Recall}{Precision + Recall} \\
J &= Specificity + Recall - 1
\end{aligned}$$

- J is the **Youden's J Statistic**