SLIIT UNI
THE KNOWLEDGE UNIVERSITY

Sri Lanka Institute of Information Technology

Faculty of Computing

IT2120 - Probability and Statistics

Ms. K. G. M. Lakmali

Year 02 and Semester 01

Lecture 12

REGRESSION ANALYSIS (Part 2)

Multiple Linear Regression (MLR)??

Multiple Linear Regression

- In majority of real-world applications, the response of an experiment can be predicted more precisely not on the basis of a single independent variable but on a collection of such variables.
- Let say response variable **Y** can be predicted based on **k** independent variables. Then the equation to predict **Y** can be written as follows:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e$$

Multiple Linear Regression

- Accordingly, let's assume sample size is n.
- Then **multiple linear regression model** to predict response variable **Y** based on **k** independent variables can be written in matrix form as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Multiple Linear Regression

Here notations, in the equation can be defined as follows:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- $\mathbf{Y}$ is an $n \times 1$, $\mathbf{X}$ an $n \times p$, β a $p \times 1$, and $\mathbf{e}$ an $n \times 1$ matrix where $p \equiv k + 1$

Significance of Regression Coefficients

- All the regression coefficients needs to be significant in the fitted model.
- Significance of regression coefficients can be tested using p values.
- Also, significance of the fitted regression model can be tested using the p value for the fitted regression model.

Coefficient of Determination (R^2)

- This explains how well the model fitted for data.
- This refers to the proportion of the total variation that is explained by the linear regression of y on x variables. In other words, R^2 is percentage of variation of Y explained by the X variables in the fitted model.

$$R^2 = \frac{SSR * 100}{SST}$$

Coefficient of Determination (R^2)

- The R^2 value always increases when you add more independent variables to the model, even if those variables are completely irrelevant.
- This is a major weakness of R^2 as it can lead to overfitting of the model. To address this issue, **Adjusted R^2** can be used.
- Thus, Adjusted R^2 will be used to evaluate the true explanatory power of the fitted model.

Assumptions for MLR

- The model is **linear in parameters.**
- **No multicollinearity** among predictor variables (All the predictor variables are independent).
- **The residuals (errors)** are assumed to be **normally distributed** with **zero mean** and **constant variance (σ^2).**
- **No Autocorrelation** among residuals (The residuals (ε_i) are independent).

Exercise

Exercise

A real estate company wants to predict house prices based on house size (Size), number of bedrooms (Bedrooms), age of the house (Age) and distance to nearest city (Distance). They have considered 100 randomly selected houses. Following is the R output for the regression model:

```
Call:
lm(formula = Price ~ Size + Bedrooms + Age + Distance, data = house_data)

Residuals:
    Min      1Q  Median      3Q     Max
-65.908 -19.891  -0.155  19.479  67.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.346063   9.328443   5.611 1.78e-07 ***
Size         0.148019   0.008327  17.776  < 2e-16 ***
Bedrooms    14.892659   2.671995   5.574 2.12e-07 ***
Age         -0.795337   0.192962  -4.122 7.75e-05 ***
Distance    -2.501393   0.417094  -5.997 2.71e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.63 on 95 degrees of freedom
Multiple R-squared:  0.8316,    Adjusted R-squared:  0.8245
F-statistic: 117.3 on 4 and 95 DF,  p-value: < 2.2e-16
```

1. Write the equation for the fitted regression model.
2. Check the significance of the regression coefficients in the fitted model.
3. What is the coefficient of determination? Interpret the value.
4. Is the fitted regression model significant? Justify your answer.

Thanks!

Any questions?