Tutorial 6

YANG YANG

The Australian National University

Week 7, 2017

Overview

A1 solution

Please see the solution to Assignment 1 on Wattle. Compare your answer to the solution and you will understand why you lost a half or one mark in particular questions. Ask me if you have any questions.

I am not authorized to remark any assignment. If there is any mistakes in adding up your total results, please let me know.

Introduction to multiple linear regression

- Need one dependent variable (continuous) and at least two independent variables (continuous)
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$
- Use least squares method to estimate parameters, minimise the error (residual) sum of squares
- **In R/RStudio, use "**$lm(Y \sim X_1 + X_2 + \cdots + X_n)$**"**

Matrix notation

$$Y = X\beta + \epsilon,$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}$$

Assumptions for MLR models

- We assume uncorrelated (independence) and homoscedastic (constant variance) errors.
- We generally assume that the ϵ_i's are normally distributed with zero mean and constant variance.
- We assume that the underlying true relationship between the response and the predictors is a linear one. $\rightarrow$ "**linear in the parameters**"

Linear relationship?

- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
- $Y = \frac{x_1 x_2}{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$

Linear relationship?

- Taking reciprocals reveals that the model has a linear relationship in parameters
- $\frac{1}{Y} = \frac{\beta_0 + \beta_1 x_1 + \beta_2 x_2}{x_1 x_2}$
- $\frac{1}{Y} = \beta_0 \frac{1}{x_1 x_2} + \beta_1 \frac{1}{x_2} + \beta_2 \frac{1}{x_1}$
- Notice that the roles of the β's appear to have changed and this model has no "intercept" term.

Interpreting a partial regression coefficient

Imagine a case with two predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

β_1 represents the expected change in Y when X_1 is increased by one unit, but X_2 is held constant or otherwise controlled.

Intercept

- The interpretation of the intercept is the expected value of Y when all the X variables are equal to 0.
- We need to be careful here. Sometimes the β_0 is not directly interpretable (extrapolation).

Partial regression

To explore a complex dataset, we can plot y against **each** x_i. However, the other predictors often affect the relationship between a given predictor and the response. In RStudio, this step can be quickly done using pairs(dataset). But there may exist confounding variables.

Alternatively, partial regression plots can help isolate the effect of x_i on y.

Partial regression

- We regress y on all predictors except for x_j, and get residuals $e_{Y|X_{-j}}$. Use it as the new response.
- We regress x_j on all predictors except for x_j, and get residuals $e_{x_j|X_{-j}}$. Use it as the new predictor.
- Partial regression plots (added variable plots) shows any relationship between $e_{Y|X_{-j}}$ and $e_{x_j|X_{-j}}$ without being "contaminated" by any of the possible (linear) confounding effects of the other variables. (Page 18 of Chapter 2; Lecture Notes)