Tutorial 4 Regression part 2 and logistics regression

Introduction to Topic 4

Lecture

Regression Analysis and Logistic regression

Last week, we started on the topic of multivariate linear model. The two-hour lecture didn't get us very far, so in the fist half of this week's lecture. We finally completed the topic of multivariate linear model and moved into the area of classification problem.

There are several methods available for tackling classification problems. We begin by fitting a "linear" model to a binary classification task (y is binary, e.g. Yes/No). I put "linear" in quotation marks because, technically, we aren't fitting a linear model in the strictest sense; rather, we're fitting a logistic curve to the linear combination of all predictors, such that

$$Pr(y_i = 1|x_{i1}, \ldots, x_{ip}) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

The logit function ensures that the output is constrained between 0 and 1. Similar to a multivariate linear model, our goal with a given dataset is to estimate the value of $\boldsymbol{\beta}$), calculate the variance of $\hat{\boldsymbol{\beta}}$, assess the model's goodness of fit (note that R^2 is not applicable here!!), check underlying assumptions, and make inferences. However, because many data scientists lack a background in statistics, these crucial steps are often overlooked. Unfortunately to cover all these steps, it will take a few lectures, but if you are interested in knowing "HOW to do these", you may like to enrol in MA2405. In the scope of MA3405, we will focus on model fitting and interpretation of output.

Logistics regression is a member of Generalized Linear Model (GLM), its parameters can be estimated using the maximum likelihood approach. With the appropriate settings, the `(glm)` function in R will generate these estimates.

Textbook Reading

Prescribed reading

Author	Title	Relevant chapters
James et al	An Introduction to Statistical Learning with Applications in R	ch 3.3 ch 4.3
Faraway	Linear model with R	entire book

Tutorial

In the last week's tutorial, you fitted multivariate linear models to various regression problems. You used the `lm()` function in R to find the estimated value of coefficients (i.e. $\boldsymbol{\beta}$), determine variance of $\hat{\boldsymbol{\beta}}$, calculated R^2

of the model, and make inference using the R output. This week, we will concentrate on interpreting model outputs, particularly when dealing with categorical predictors, verifying model assumptions and checking for collinearity. We'll conclude the tutorial by addressing a binary classification problem using logistic regression.

Independendent Learning

In a group of two, discuss the following topics

- How can AIC, BIC and Mallow's CP be used in linear model?
- What do these criteria have in common? How do they differ?
- Explain how a categorical predictor (e.g with 3 levels) is coded in X matrix.
- If a predictor has three categories; explain why R only produces coefficients for two of those categories
- What is synergy effect?
- How can synergy effect be tested in linear model setting?
- What is collinearity?
- How can collinearity be tested?
- What is a odds ratio?
- What is Simpson's paradox?

Exercises - Part 1

Question 1:

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (three levels, postgraduate, college, high school), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars).

Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_{GPA} = 20, \hat{\beta}_{IQ} = 0.07, \hat{\beta}_{college} = 35, \hat{\beta}_{postgraduate} = 50, \hat{\beta}_{GPA:IQ} = 0.01, \hat{\beta}_{GPA:college} = -10$ and $\hat{\beta}_{GPA:postgraduate} = -11$

Using the result above,

1. Specify the predictive model used to predict the starting salary of High school graduate
2. Specify the predictive model used to predict the starting salary of College graduate
3. Specify the predictive model used to predict the starting salary for individuals with postgraduate degree
4. If the p-value for $\hat{\beta}_{GPA:college}$ is less than 0.05, what does it implies?

Question 2:

Last week, in Exercise 9(d), you used the `plot` function to produce diagnostic plots of the linear model. Four figures were produced using this function. Explain which assumption each of these figures is designed to assess.

Question 3:

Exercise 14, Chapter 3.7

Independendent Learning

Labs

(We will leave LDA and QDA until next week)

Exercises-part 2

- Question 13 a to d, Chapter 4.8
- Question 14 a, b and f, Chapter 4.8