STAT 302 - Chapter 3 : Multiple Regression - Part 3

Harsha Perera

Key Topics

- New Predictors from Old
- Interaction
- Polynomial Regression
- Complete Second-Order Model

New Predictors from Old

- In chapter 1 we saw someways to address nonlinear pattern in a simple linear regression setting by considering transformations such as square root or logarithm.
- In multiple linear regression model, we can include more than one function of one or more predictors in the model at the same time.
- In this section, we illustrate two common methods of combining predictors.
- **Interaction** - Using a product of two predictors.
- **Polynomial Regression** - Using one or more powers of a predictor.
- At the end we combine these two methods to produce a **Complete Second-Order** model.

Interaction

- In section 3.3 we discussed about the interaction between a quantitative predictor and a categorical predictor, which allowed the model to fit different slopes.
- An interaction term can also be defined with two continuous predictors.
- For two predictor variables, the full interaction model has the form,

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_1 \cdot X_2 + \in$$

- The interaction product term, $X_1 \cdot X_2$, allows the slope with respect to one predictor to change for values of the second predictor.

Interaction - Example 3.12 : Guessing IQ

Can you tell how smart someone is just by looking at a photo of the person ?

- The dataset contains Guess IQ, True IQ, and Age for 40 women.
- We divided Age into two groups (younger and older) and added separate regression lines. For younger women there is a negative relationship between True IQ and Guess IQ, but for older women the relationship is positive.
- Age is a quantitative variable and we don't have to limit ourselves to an interaction model with only two groups. Instead, we can use age as a continuous variable and fit an interaction model that allows the relationship between Guess IQ and Actual IQ to change continuously as age changes.

Interaction - Example 3.12 : Guessing IQ

- We create the interaction term between Guess IQ and Age and include it in the regression model. The interaction model is :

  $$TrueIQ = \beta_0 + \beta_1 \cdot GuessIQ + \beta_2 \cdot Age + \beta_3 \cdot GuessIQ \cdot Age + \in$$

- After fitting the model using R :

  $$\widehat{TrueIQ} = 834.76 - 5.70 \cdot GuessIQ - 33.02 \cdot Age + 0.26 \cdot GuessIQ \cdot Age$$

Example 3.12 : Guessing IQ - Inference

- T test for individual predictors: suggest all three predictors are significant including the interaction term.
- R^2: Only 15% of variation in True IQ is explained by the model. Without interaction term it is just 4%.
- Model without interaction: none of the predictors are significant.
- When dealing with two predictors always begin with full model. If interaction term is not significant we may proceed to additive model.

Polynomial Regression

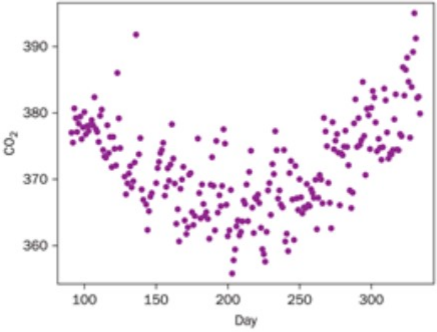
- In chapter 1, we saw methods for dealing with data that showed curved rather than linear relationship by transforming one or both of the variables.
- In multiple linear regression model, we can deal with the curvature by adding powers to one or more predictor variables.
- For a single quantitative predictor X, a quadratic regression model has the form :

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \in$$

Example 3.13 - Daily Carbon Dioxide

- CO_2 levels (parts per million) in the atmosphere for each day from April, 2001 to (Day 91) through November, 2001 (day 334)

- CO_2 level is lower in summer months.



- We want to find a model that captures the main trend in this scatterplot.

Example 3.13 - Daily Carbon Dioxide

- Fitting a multiple linear model to predict CO_2 based on Days and DaysSq is :

$$CO_2 = \beta_0 + \beta_1 \cdot Day + \beta_2 \cdot Day^2 + \in$$

- The fitted quadratic relationship from R is :

$$\widehat{CO_2} = 414.97 - 0.4760 \cdot Day + 0.0012 \cdot Day^2$$

- Inference - Refer the R output !!!

Polynomial Regression Model

- We can generalize the idea of quadratic regression to include additional powers of a single quantitative predictor variable. Even though, additional polynomial terms may not improve the model much.

- For a single quantitative predictor X, a polynomial regression model of degree k has the form :

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 X^2 + \ldots + \beta_K X^K + \in$$

- We added a $Day3 = Day^3$ predictor to the quadratic model to predict CO_2 levels based on a third degree polynomial regression model.

- Refer the R output !!!

Complete Second-Order Model

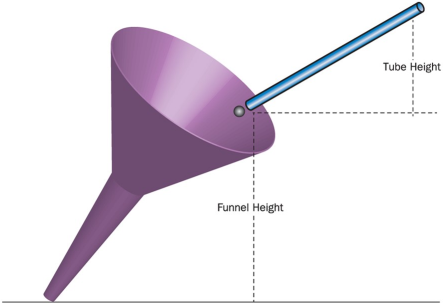
- A model with both interaction terms and polynomials of predictors.
- We should take care that we don't over parameterize the model, making it more complicated than needed, with terms that aren't important for explaining the structure of the data.
- For two predictors, X_1 and X_2, a complete second order model includes linear and quadratic terms for both predictors along with the interaction term :

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_1^2 + \beta_4 \cdot X_2^2 + \beta_5 \cdot X_1 \cdot X_2 + \in$$

- This extends to more than two predictors by including all linear, quadratic, and pairwise interactions.

Example 3.14 - Funnel Swirling

The diagram below shows the setup for an experiment where a steel ball is rolled through a tube at an angle into a funnel. The ball swirls around as it travels down the funnel until it comes out the bottom end and hits the table. We are interested in maximizing the time it takes for the ball to make this trip. We can adjust the relative steepness by changing the angle of the funnel (as reflected in the funnel height) or the height of the drop tube.



Example 3.14 - Funnel Swirling

- Four different funnel heights: 8, 11, 14 and 16 inches.
- Three different tube heights: 8, 11 and 14 inches.
- 10 runs from each funnel height and tube height combination which produced total 120 drop times.
- Objective is to maximise the drop time.

Example 3.14 - Funnel Swirling

- Scatter plots of the drop time vs both funnel height and tube height suggest longer drop times tend to occur at the middle heights, while the times tend to be somewhat smaller at the more extreme (high or low) heights.

- This curvature suggests adding possible quadratic terms for both funnel and tube heights.

- Angle of the tube relative to the funnel could have some impact on drop times. This introduces interaction term.

- Therefore the complete second order model : - Complete second order model

$$Time = \beta_0 + \beta_1 \cdot Funnel + \beta_2 \cdot Tube + \beta_3 \cdot Funnel^2 + \beta_4 \cdot Tube^2$$

$$+\beta_5 \cdot Funnel \cdot Tube + \in$$

Funnel Swirling - Inference

- No concerns with linearity, constant variance or normality of the residuals.
- F test: suggest at least one of the predictors is associated with drop time.
- All the individual t-tests are significant at 5% level. This is not always the case. In general if second order terms or interaction term is significant we need to keep the main effects even they are not significant.
- R^2: 29.3% of variation in drop time is explained by the model which is not that big. But Without the second order terms - R^2 is lower.