Lesson 18. New Predictors from Old – Part 1

1 Overview

- Suppose we have three quantitative variables, Y, X_1, and X_2
- The model
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
allows us to fit linear relationships between Y, X_1, and X_2
  - In 3D: a flat surface (plane) through a cloud of observations
- But... what if that's not the pattern in the data?
- In this lesson, we will learn about new forms of predictors to
  - make the model more flexible, and
  - address non-linear patterns (especially if the linearity conditions are violated)

2 Polynomial terms

- We can include new predictors that take a quantitative predictor variable and raise it to some power
- Model examples:

- **Quadratic terms** allow us to curve the surface we are fitting to the data
- For a single quantitative variable X, a **polynomial regression model of degree** k has the form

3 Interactions

- In some situations, the slope with respect to one predictor might change for different values of the second predictor
- This is called an **interaction between the two predictors**
- In the previous lesson, we saw an interaction between a quantitative variable and an indicator variable
- Now we will consider interactions between two quantitative variables
- The **regression model with interaction for predictors** X_1 **and** X_2:

- The interaction term allows us to twist the surface we are fitting to the data

4 Complete second-order model

- The **complete second-order model for predictors X_1 and X_2**:

- For two predictors, a complete second-order model includes
  - linear and quadratic terms for both predictors, along with
  - the interaction term
- This extends to more than two predictor variables by including all linear terms, all quadratic terms, and all pairwise interactions

5 Notes

- A major indication that we should try including some of these new terms:
  - How do we check for this?
- It is important not to **overfit**: make the model too complicated so that it fits the sampled data well, but doesn't translate to the population
  - We want the simplest model that captures the structure in the data
  - We want a *parsimonious* model
- If a higher-order term (interaction, cubic, etc.) is significant, we will also leave the associated lower-order terms in the model (even if they aren't significant)
  - If a higher-order term is not significant, we should consider dropping it
- Two ways to guard against including unnecessary complexity:
  - Examine t-tests for the individual terms
  - Check how much additional variability is explained by these terms
- If linearity is met, we can make good point predictions, and we also have a reasonable summary of the general relationships among the variables
  - However, unless the other modeling conditions are met as well, we should not do formal inference (hypothesis testing, intervals)
  - For our purposes, we will only use p-values as a rough guide if we are in this case