Contents

More Types of Predictors (Ch. 8)

1. Polynomial Regression

- We can fit a quadratic, cubic, etc. relationship by defining squares, cubes, etc., of a single X in a data step and using them as additional explanatory variables.

- We can do this with more than one explanatory variable if needed.

- Issue: When we do this, we generally create a multicollinearity problem.

- Remedy: We can remove the correlation between explanatory variables and their powers by <u>centering</u>: subtract the mean before squaring etc. (However, whether center the variable or not does not affect the highest order of the polynomial term.)

- If a high-order polynomial term (e.g., X^2) of a predictor is in the model, the low order term (e.g., X) will remain in the model.

- See SPSS Lab 4 for examples.

2. Interaction

With several explanatory variables, we need to consider the possibility that the effect of one variable depends on the value of another variable. In linear regression, interaction between two predictors (X_1 and X_2) refers to the product term X_1X_2.

- **Two numerical variables**
  - Model: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \varepsilon$
  - Can be rewritten as follows:
    $Y = \beta_0 + (\beta_1 + \beta_3X_2)X_1 + \beta_2X_2 + \varepsilon$
    $Y = \beta_0 + \beta_1X_1 + (\beta_2 + \beta_3X_1) X_2 + \varepsilon$
  - The coefficient of one explanatory variable depends on the value of the other explanatory variable.

- **One binary variable and one numerical variable**
  - X_1 takes values 0 (Group A) and 1 (Group B) corresponding to two different groups.
  - X_2 is a continuous variable.

- Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$
- Regression lines for each group:
  Group A: $X_1 = 0$. Then, $Y = \beta_0 + \beta_2 X_2 + \varepsilon$
  Group B: $X_1 = 1$. Then, $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \varepsilon$
- Interpretation of the regression coefficients (parameters):
  β_0 is the intercept for Group A;
  $(\beta_0 + \beta_1)$ is the intercept for Group B.
  Similar interpretation for slopes β_2 and β_3.
- H_0: $\beta_1 = \beta_3 = 0$ tests the hypothesis that the regression lines are the same for both groups.
- H_0: $\beta_1 = 0$ tests equal intercepts
- H_0: $\beta_3 = 0$ tests equal slopes (Parallel lines, no interaction)
- See Lab for example.

➢ If the interaction term of two predictors (e.g. X_1X_2) is in the model, the first-order terms of the predictors (e.g., X_1 and X_2) will remain in the model.

3. Qualitative (Categorical) Predictors

Qualitative/Categorical predictors may be saved as numbers in the data set. However, such numerical coding may not be suitable for analysis.

➢ **The Dummy Coding**
Use k-1 indicators (dummy variables) to represent a k-level categorical predictor. One of the k-levels (typically the first one or the last one for convenience) is considered as the "control" (based-line) level. For example,
- Binary (2-level) predictor: Group A vs B (see Interaction)
- Consider a regression model of tool wear (Y) on tool speed (X_1) and tool model (X_2). The predictor X_2 has 4 values for 4 classes of tool models (M1, M2, M3 and M4). Create 3 indicator variables to replace the original X_2 (assume M4 is the "control")
  $X_2 = 1$ if M1, and 0 otherwise
  $X_3 = 1$ if M2, and 0 otherwise
  $X_4 = 1$ if M3, and 0 otherwise
  Hence,
  if a case is M1, $(X_2, X_3, X_4)=(1, 0, 0)$;
  if a case is M2, $(X_2, X_3, X_4)=(0, 1, 0)$;
  if a case is M2, $(X_2, X_3, X_4)=(0, 0, 1)$;
  if a case is M4, $(X_2, X_3, X_4)=(0, 0, 0)$.

➢ **The Interpretation**
- Each level (group) of the categorical variable can have its own regression line.

- Each slope parameter β for the indicator variable is the estimated difference between the corresponding level and the control (baseline) level, after adjusted to other predictors.
- H_0: $\beta_3 = 0$ tests if M2 is significantly different from M4, after adjusted to other predictors.
- H_0: $\beta_2 = \beta_3 = \beta_4 = 0$ tests whether, after the response variable is adjusted to other predictors, the categorical predictor is significant. In this case, we can test any tool type is different from another with respect to the tool wear, after considering other variables.

➢ **Interactions Between Categorical and Numerical Predictors**

- Treat the set of indicators as a whole and create interaction terms between the numerical predictor and each indicator.
- In the previous example, if there is an interaction between speed and model, we will need X_1X_2, X_1X_3, X_1X_4.
- Interactions between 2 categorical predicators need more terms. (Why?)