Chapter 8 – Polynomial Models and Qualitative Predictors

Example: In an environmental engineering setting, the concentrations of 18 separately prepared solutions were recorded at different times (three measured at each of 6 times).

\[ X_i = \text{time} \quad Y_i = \text{concentration (mg/ml)} \]

The data are included in the SAS classnotes.

Suppose we want to fit a polynomial equation to this data, for example

\[ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon \]

How should we do this?

Should we use a quadratic (cubic) equation?

(1) Generally — start with a lower order model and try to add terms to possibly get a better model.
Don't get carried away! Don't start with too high a power in a polynomial.

Note that we are including the intercept in this fit.

SAS Example:
Cubic model is the best fit.
\[ R^2 = 0.9868 \]

The overall ANOVA and all three coefficients are significant.
Two Predictor Polynomials:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \]

*First Order Model*

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \epsilon \]

*Second Order Model*
Categorical Predictors in Regression

Qualitative (categorical) variables often play an important role in modelling a response.

Example: Suppose you want to model

\[ Y = \text{person's weight (lbs)} \]

as a function of

\[ X_1 = \text{person's height (inches)} \]

In the model, it is also quite useful to include the person's gender

\[ X_2 = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases} \]
Consider the model
\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \]
for males it becomes
\[ E(Y) = \beta_0 + \beta_1 X_1 \]
and for females it is
\[ E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1 \]
You may not believe that these lines should be parallel. If so, you should introduce an interaction term into the model:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \]

For males the model becomes

\[ E(Y) = \beta_0 + \beta_1 X_1 \]

and for females it becomes
Why combine the data among men and women? Why not just model them separately?

(a) Suppose observe \( n_1 \) males and \( n_2 \) females.

- males \( \text{SSE} \ n_1 - 2 \ \text{df} \)
- females \( \text{SSE} \ n_2 - 2 \ \text{df} \)

But when combined we get

\( \text{SSE} \ n_1 + n_2 - 4 \ \text{df} \).

The larger \( \text{df} \) is an advantage as long as \( \sigma_M^2 = \sigma_F^2 \).
Several Categories

Example: \( Y = \) breath rate per minute while sleeping

\[ X_1 = \text{age of person} \]

\[ \text{Snoring} \begin{cases} \\ \text{Never} \\ \text{Occasional} \\ \text{Frequent} \end{cases} \]

How do we include snoring in our model?

This variable is an ordinal categorical variable, i.e. the categories have a natural ordering.

This model creates a rigid structure on the relationship between \( E(Y) \) and the predictors.
Never \[ E(Y) = \beta_0 + \beta_1 X_1 \]
Occasional \[ E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1 \]
Frequent \[ E(Y) = (\beta_0 + 2\beta_2) + \beta_1 X_1 \]

This model creates:
(a) evenly spaced lines
(b) the "occasional" category must be in the middle.
Never \[ E(Y) = \beta_0 + \beta_1 X_1 \]
Occasional \[ E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1 \]
Frequent \[ E(Y) = (\beta_0 + \beta_3) + \beta_1 X_1 \]

Now (a) the lines are not necessarily evenly spaced
(b) the "Occasional" category is not necessarily in the middle.
If a predictor is categorical and not ordinal, we should always use indicator variables.

If a categorical variable has c categories we use (c-1) indicator variables to describe it.

**Several Categorical Predictors**

**Example:** Model

\[ Y = \text{course point total} \]

\[ X_1 = \text{GPA entering course} \]

plus we want take into account a person's gender and whether that person is an undergrad or a grad student.

(A) \[ X_2 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases} \]

\[ X_3 = \begin{cases} 1 & \text{undergrad} \\ 0 & \text{grad} \end{cases} \]

Model:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \]
Male Grad \[ E(Y) = \beta_0 + \beta_1 X_1 \]
Male Undergrad \[ E(Y) = (\beta_0 + \beta_3) + \beta_1 X_1 \]
Female Grad \[ E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1 \]
Female Undergrad \[ E(Y) = (\beta_0 + \beta_2 + \beta_3) + \beta_1 X_1 \]

Note the structure imposed: the difference between male + female is the same for Undergrads and grads.