2.1.7 Polynomial Regression

In many practical situations, the simple straight line

$$y = \beta_0 + \beta_1 x$$

is not appropriate. Instead, a model including powers of x

$$x^2, x^3, \ldots, x^k$$

should be considered. For example

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x^j = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$$

The **Polynomial Regression Model**

$$Y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k + \epsilon$$

where ϵ is a random error term as before can be used to model data.

Two immediate problems:

1. How to choose k
2. How to carry out inference
   - estimation
   - testing
   - prediction

We begin by addressing 2. The estimation of parameters can be again carried out using **Least Squares** provided that the model assumptions listed before are valid. Consider $k = 2$.

We choose $\underset{\sim}{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathsf{T}}$ to minimize the **sum of squared errors**

$$SSE(\underset{\sim}{\beta}) = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

that is the fitted values for parameters $\underset{\sim}{\beta}$ are

$$\widehat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

$\underset{\sim}{\widehat{\beta}}$ can be found to minimize SSE using calculus techniques (differentiating with respect to the elements of $\underset{\sim}{\beta}$) to give the minimum SSE

$$SSE(\underset{\sim}{\beta}) = \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i - \widehat{\beta}_2 x_i^2)^2$$

We can also compute the estimated **standard errors**

$$s_{\widehat{\beta}_0}, s_{\widehat{\beta}_1}, s_{\widehat{\beta}_2}$$

which allow tests of parameters to be carried out, and confidence intervals calculated.

We can also compute prediction intervals.

The best estimate of the residual error variance σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-3} SSE(\widehat{\underset{\sim}{\beta}})$$

p is the number of parameters estimated equal to three, so we divide by $n - 3$.

We can also compute

- Residuals
  - can be used to assess the fit of the model.
  - the residuals should be *patternless* if the model fit is good.
- R^2, Adjusted R^2 statistics
  - used to assess the global fit of the model.
  - used to compare the quality of fit with other models.

Example: Hooker Pressure Data.

For the Hooker pressure data, a **quadratic** polynomial ($k = 2$) might be suitable.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

We need to estimate β_0, β_1 and β_2 for these data to see if the model fits better than the straight line model we fitted previously. This can be achieved using SPSS.

It transpires that the quadratic model produces a set of residuals that are patternless, which the straight line model when fitted does not.

See Handout for full details.

Note: It is common to use the **Standardized Residuals**

$$\widehat{z}_i = \frac{\widehat{e}_i}{\widehat{\sigma}} = \frac{y_i - \widehat{y}_i}{\widehat{\sigma}}$$

where $\widehat{\sigma}^2$ is the estimate of σ^2 defined previously, as

$$\text{Var}[\widehat{z}_i] \approx 1$$

if the model fit is good, whereas

$$\text{Var}[\widehat{e}_i] \approx \sigma^2$$

which clearly depends on σ. This makes it hard to compare $\widehat{e}_i$ across different models when inspecting residuals.