

Lecture 6b: Multiple Linear Regression miscellanea

Math 586

February 24, 2009

Objectives (revisited):

1. Use observations (\mathbf{x}_i, Y_i) to estimate β 's.
2. Assuming $\sigma^2 = \text{Var}(\varepsilon_i)$ is constant, estimate σ^2 .
3. Decide if terms can be dropped (strive for simpler models).

Assume that: (i) $\text{Var}(\varepsilon_i) = \sigma^2$ for all i ,
(ii) ε_i 's are independent.

Then, to estimate σ^2 , consider (p is the total length of vector β)

$$\hat{\sigma}^2 = s^2 := SSE/(N - p)$$

Can prove that s^2 is unbiased for σ^2 .

If we further assume that

(iii) ε_i 's are Normal

then can do F-tests about β 's, confidence intervals etc. (we will omit this)

Measures of fit:

- Residual variance s^2 : smaller variance is better
- Coefficient of determination $R^2 = 1 - SSE/SST$: higher R^2 is better

Example: compare s^2 and R^2 for linear vs quadratic model of grain size.

However, adding more predictors to your model will *always* decrease s^2 and increase R^2 . (Why?)

Adjusted R^2 :

$$R_{adj}^2 := 1 - \frac{N - 1}{N - p} (1 - R^2)$$

includes penalty for putting too many terms into the model.

Residual plots

They offer insight into any “left-over” structure in the data: non-stationarity of errors, non-linear structures, outliers etc.

Are the residuals independent? Will return to this later...

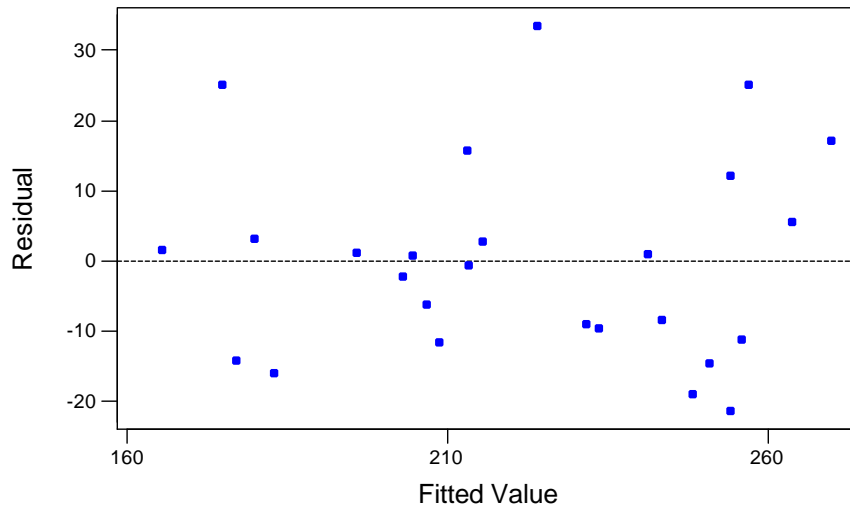


Figure 1: A little bit of non-linearity left.

Misc. notes:

- higher-order polynomial models tend to overfit (Extreme example $N = p$)
- higher-order models not good for extrapolation (overall, be careful about extending your predictions outside the scope of the data)
- check the residuals
- centering simplifies the calculations and makes some parameter estimates independent
- edge effects: the predicted fit is best in the center of region, worsens at the edges (especially severe for 3rd and higher order models).