Residuals

Remember that the predicted values are
\[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_m x_{mi}, \quad i = 1, \ldots, n. \]

The residuals are \( e_1, \ldots, e_n \), where
\[ e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n. \]

Plots to consider:

1) Construct a histogram, boxplot or normal probability plot of residuals to check on normality assumption.

2) Plot residuals against the predicted values. This is a good plot for checking the equal variances assumption.

3) If the independent variables are not highly related, plot residuals against each independent variable.
4) If the data are collected over time, plot the residuals against time. If time does not affect the response, this plot should show no pattern. *Durbin-Watson* test can be used to test for time effect. The Durbin-Watson statistic can be gotten in SPSS via *Regression → Linear → Statistics → Durbin-Watson*. Values of the statistic larger than 2.5 or less than 1.5 are indicative of a time effect.

**Outliers**

As in simple regression, outliers that occur near the boundary of the $x$-region may not show up in a residual plot. So, methods besides residuals are needed to spot outliers.

Define

$$ DFFITS(i) = \frac{\hat{y}_i - \hat{y}(i)}{\text{scale factor}}, $$

where $\hat{y}_i$ is as usual and $\hat{y}(i)$ is the $i$th predicted value obtained *after* removing the $i$th observation from the data set.
A large value of $DFFITS(i)$ indicates that the $i$th observation may be an outlier. Values bigger than 2 in absolute value indicate potential outliers.

The $DFFITS$ statistics are obtained in SPSS as follows:  $Regression \rightarrow Linear \rightarrow Save \rightarrow Standardized DfFit$.

Plot $DFFITS(i)$ against $i$ or one of the independent variables to check for outliers.

Always plot both residuals and $DFFITS$.

- Residuals may miss outliers near boundary of $x$-region.

- $DFFITS$ may miss outliers in ”middle” of $x$-region.
What should one do with outliers?

- After spotting an outlier, check to see if an error was made in recording the data. If an error was made, correct it and re-estimate the model using all the data.

- If no errors were made, there are at least two courses of action:
  - Throw out the outlier(s) and estimate the model with the remaining data. Consult a statistician if you want to predict the response at values of $x$ near the ones thrown out.
  - Use an alternative to least squares analysis, such as robust regression.