ASSESSING MODEL FIT

\[
RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.
\]

\[
RSE = \sqrt{\frac{1}{n - p - 1} RSS},
\]

\[
R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}
\]

where TSS = \(\sum (y_i - \bar{y})^2\) is the total sum of squares.

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2,
\]
• Randomly divide dataset into training and test subsets
• Create model on training set
• Calculate test MSE on test set
• Problem: May be highly variable; different selections of training sets can produce very different test MSEs for the same model

VALIDATION SET
• For each data point in the dataset, do the following:
  • Compute the model leaving the point \((x_i, y_i)\) out
  • Compute \(MSE_i = (y_i - \hat{y}_i)^2\)
• Then, compute an overall average test MSE:

\[
CV(n) = \frac{1}{n} \sum_{i=1}^{n} MSE_i
\]
• Divide the dataset in $k$ folds, and for each fold:
  • Compute the model leaving the fold out
  • Compute $MSE_i = \frac{1}{|F|} \sum (y_i - \hat{y}_i)^2$, the MSE for the fold
  • Then, compute an overall average test MSE:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

K-FOLD CROSS VALIDATION
• Test MSE is made up of two parts: bias and variance
• LOOCV tends to overestimate test MSE variance
  • The models are highly correlated
• K-fold tends to overestimate the bias
• Generally, K-fold is more accurate

K-FOLD VERSUS LOOCV
LINKS USED IN THIS LECTURE