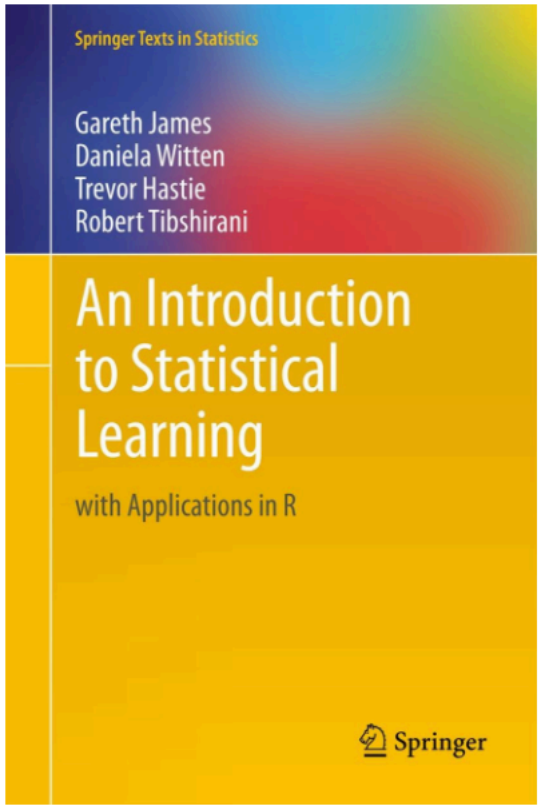
3. Model Selection

ESC Spring 2018 – Data Mining and Analysis

SeoHyeong Jeong



Textbook:

An Introduction to Statistical Learning

Lecture Slides:

Stanford Stats 202: Data Mining and Analysis

Spring 17’ ESC Statistical Data Analysis

Reading:

An Introduction to Statistical Learning

chapter 3.1.3 Potential Problems
chapter 2.2 Assessing Model Accuracy
chapter 6.1 Subset Selection
chapter 5.1 Cross-Validation (except 5.1.5)

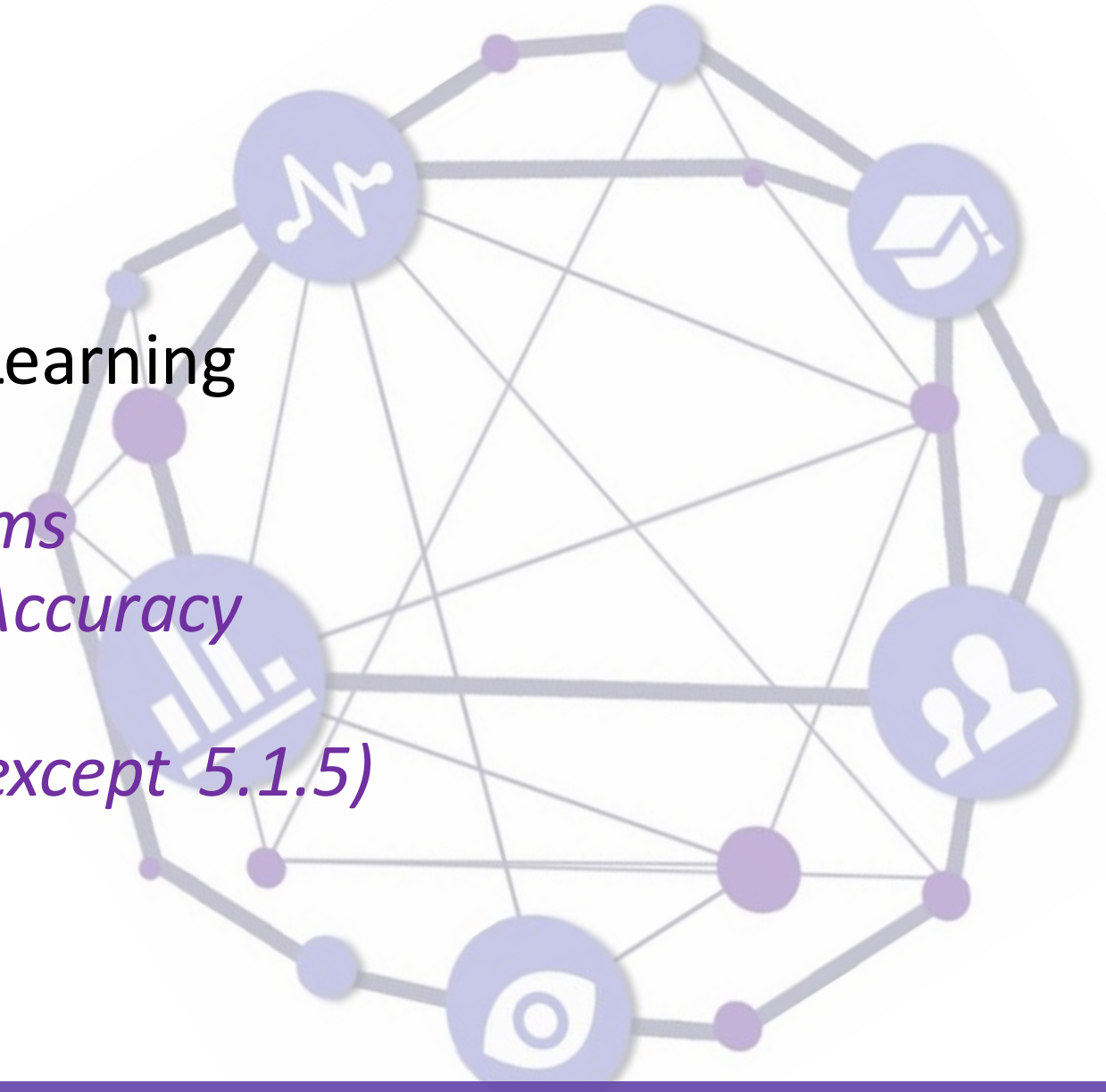


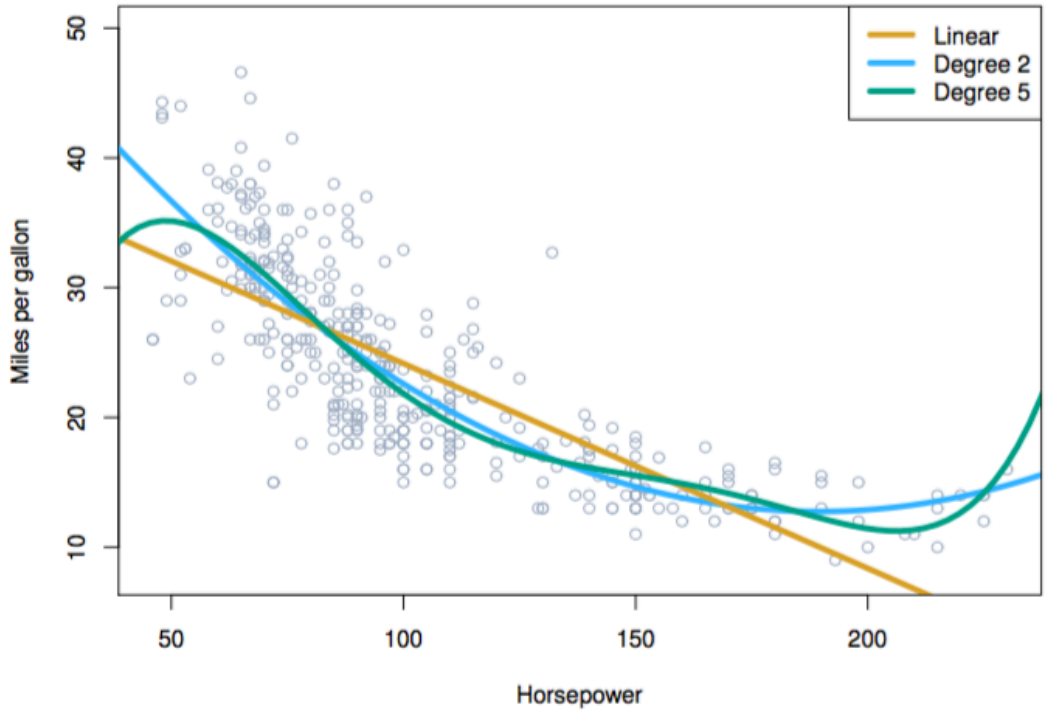
Table of Contents

Potential Issues in Linear Regression

1. Non-linear relationships
2. Correlation of error terms
3. Non-constant variance of error (heteroskedasticity)
4. Outliers
5. High leverage points
6. Collinearity

1. Non-linearity

Example: Auto dataset



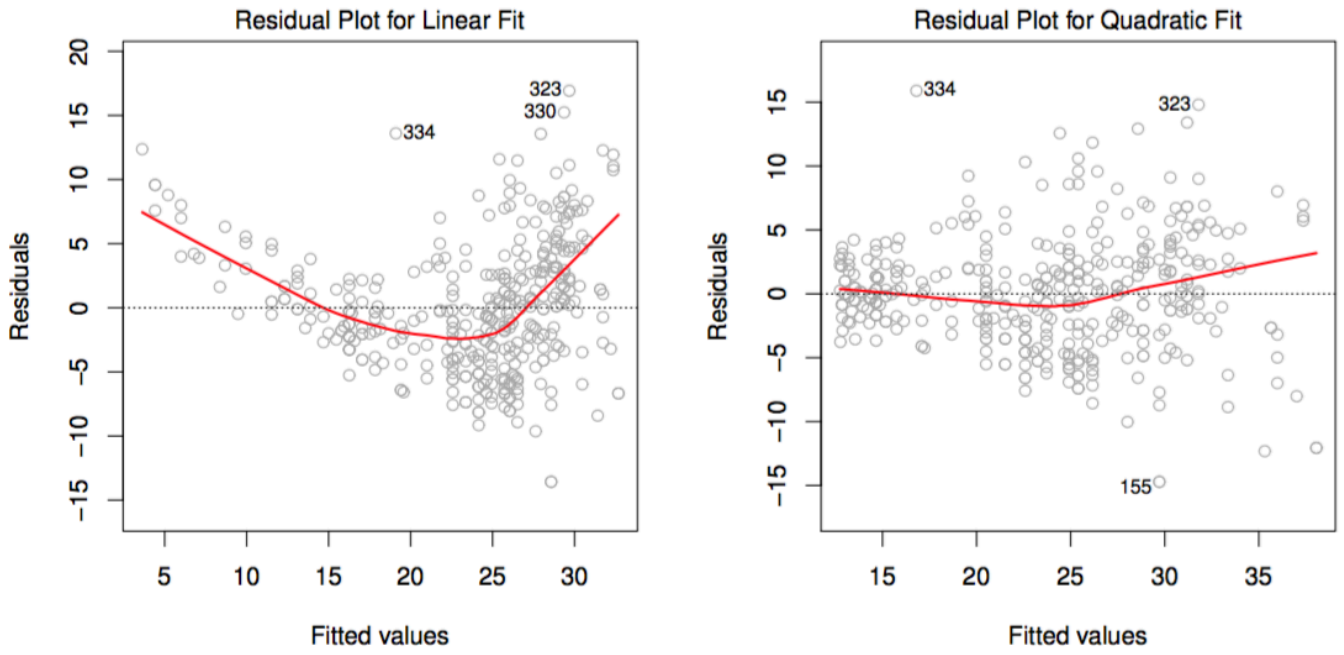
- A scatterplot between a predictor and the response may reveal a non-linear relationship.

Solution: include polynomial terms in the model.

$$\begin{aligned}\texttt{MPG} = \beta_0 + \beta_1 \times \texttt{horsepower} + \varepsilon \\ + \beta_2 \times \texttt{horsepower}^2 + \varepsilon \\ + \beta_3 \times \texttt{horsepower}^3 + \varepsilon \\ + \ldots + \varepsilon\end{aligned}$$

- In 2 or 3 dimensions, this is easy to visualize. What if there are more than 3 predictors?

  Plot the residuals against the *fitted values* and look for the pattern:



2. Correlation of Error Terms

- We assumed that the errors for each sample are independent:

$$y_i = f(x_i) + \varepsilon_i \quad ; \quad \varepsilon_i \sim \mathcal{N}(0, \sigma) \text{ i.i.d.}$$

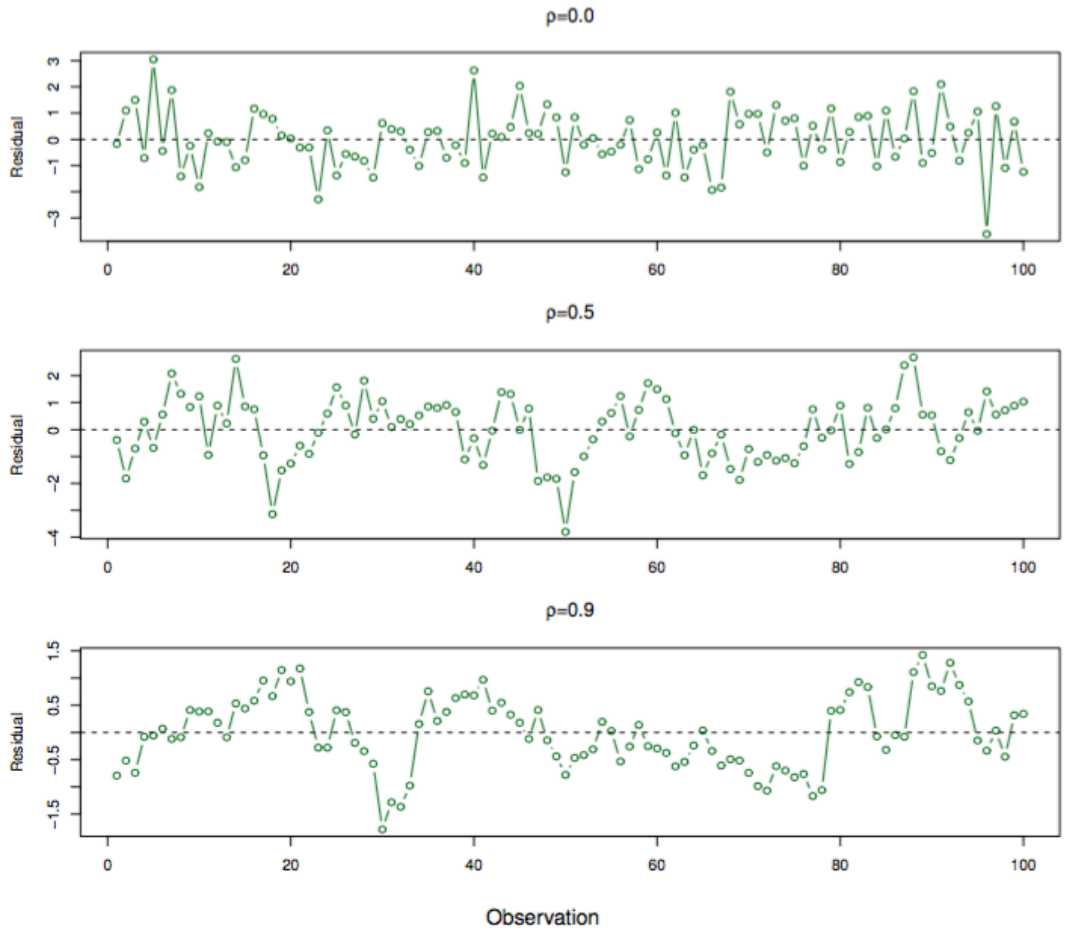
- What if this assumption breaks down?
- The main effect is that this invalidates any assertions about standard errors, confidence intervals, and hypothesis tests:

  Example: Suppose that by accident, we double the data (each sample twice).

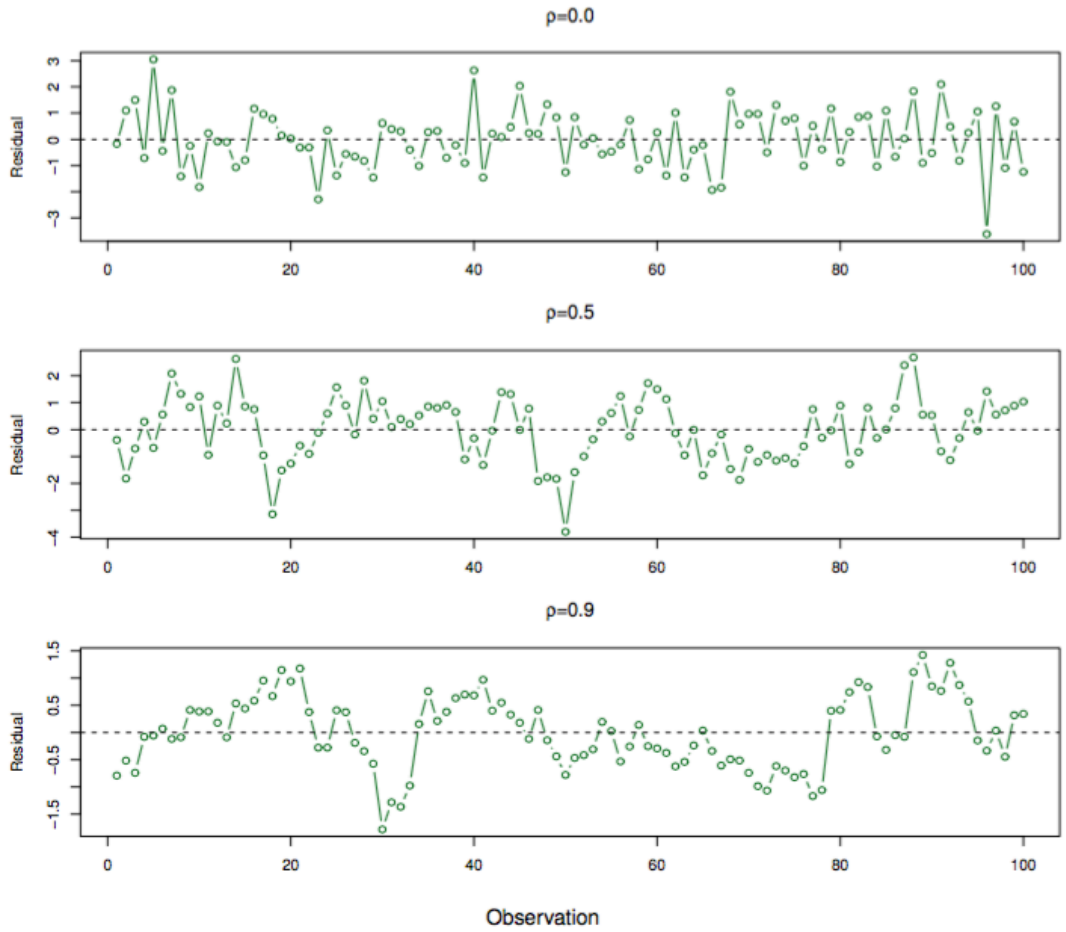
  Then, the standard errors would be artificially smaller by a factor of $\sqrt{2}$

- When could this happen in real life:
  - **Time series:** Each sample corresponds to a different point in time. The errors for samples that are close in time are correlated.
  - **Spatial data:** Each sample corresponds to a different location in space.
  - Study on predicting height from weight at birth. Suppose some of the subjects in the study are in the same family, their shared environment could make them deviate from $f(x)$ in similar ways.

Simulations of time series with increasing correlations between ε_i.



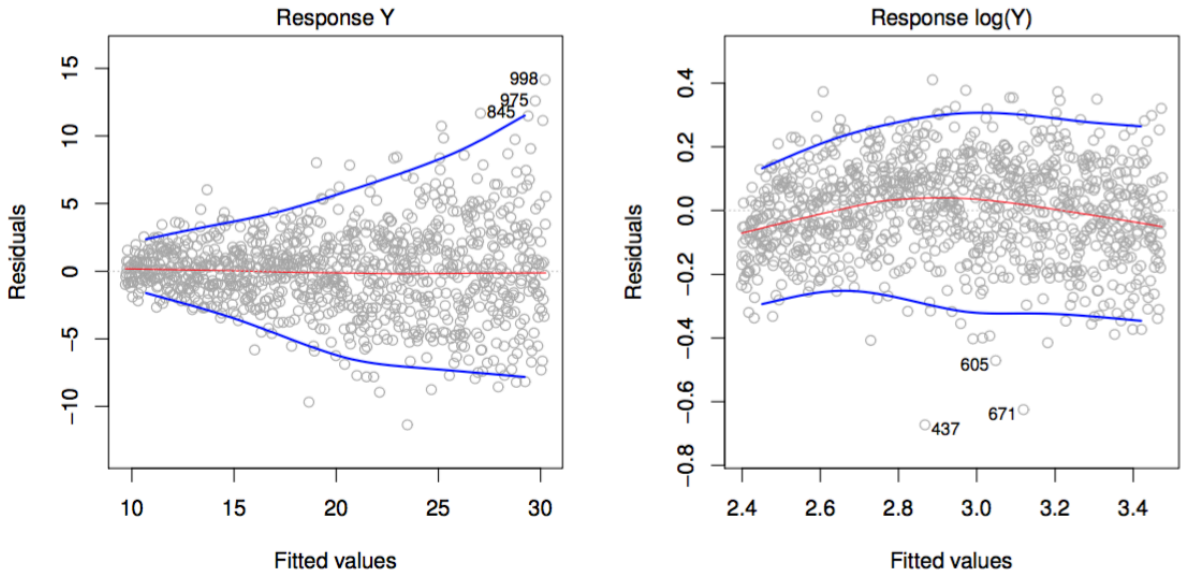
Simulations of time series with increasing correlations between ε_i.



3. Non-constant variance of error

- For example, the variance of the error depends on the input.

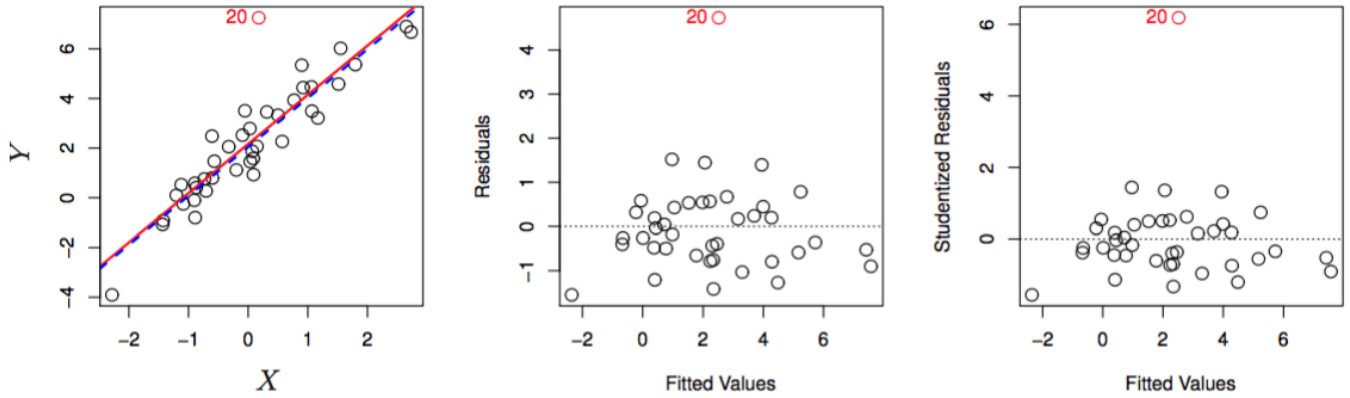
  To diagnose this, we can plot residual vs. fitted values:



Solution: If the trend in variance is relatively simple, we can transform the response using a logarithm, for example.

4. Outliers

- Outliers are points with very high errors.

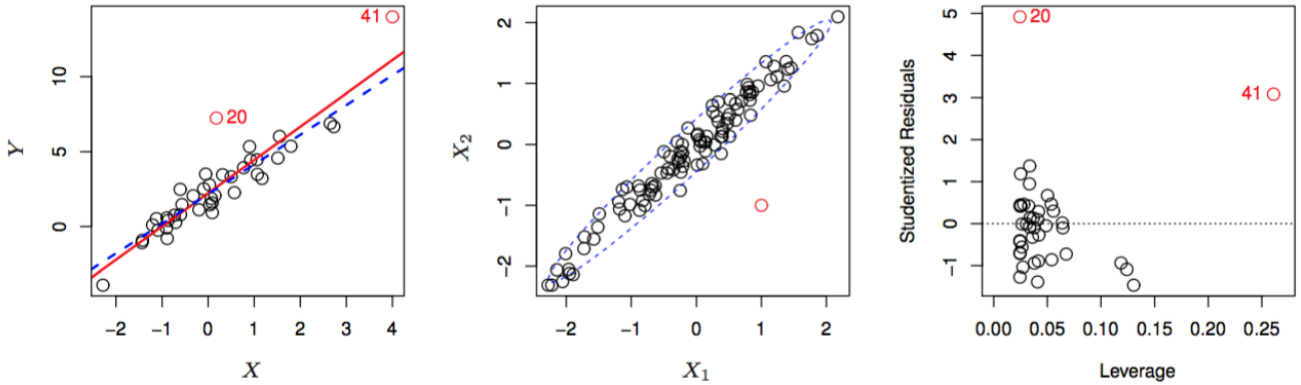


While they may or may not affect the fit, they affect out assessment of model quality.

- Possible solutions:
  - If we believe an outlier is due to an error in data collection, can remove it.
  - An outlier might be evidence of a missing predictor, or the need to specify a more complex model.

5. High Leverage Points

- **High leverage points** are observations with unusual input values. They can have an outsized effect on the fit $\hat{\beta}$**!**



Quantified with the **leverage statistic** or **self influence:**

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} = \underbrace{(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)}_{\text{Hat matrix}}{}_{i,i} \in [1/n, 1].$$

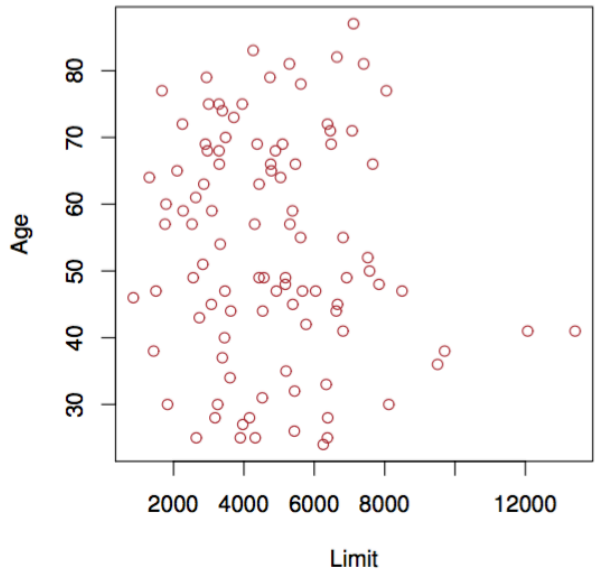
Hat matrix satisfies $\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y = \mathbf{H}y$

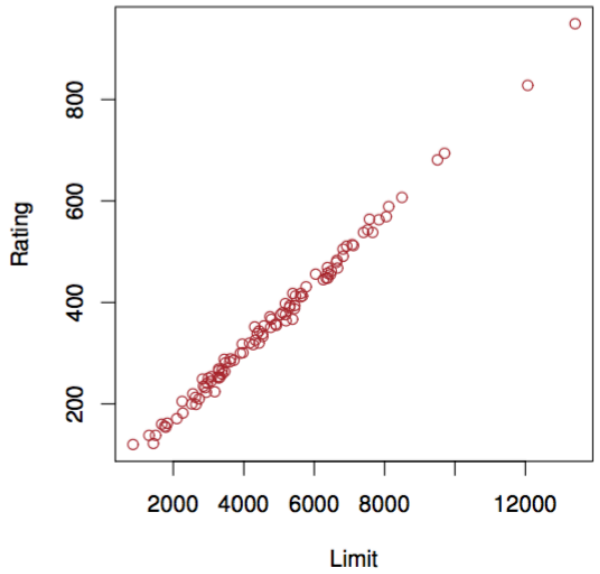
6. Collinearity

- Two predictors are collinear if they are highly correlated:

$$limit \approx a \times rating + b$$

i.e. if one variable is approximately a linear function of the other.
In that case they contain approximately the same information.

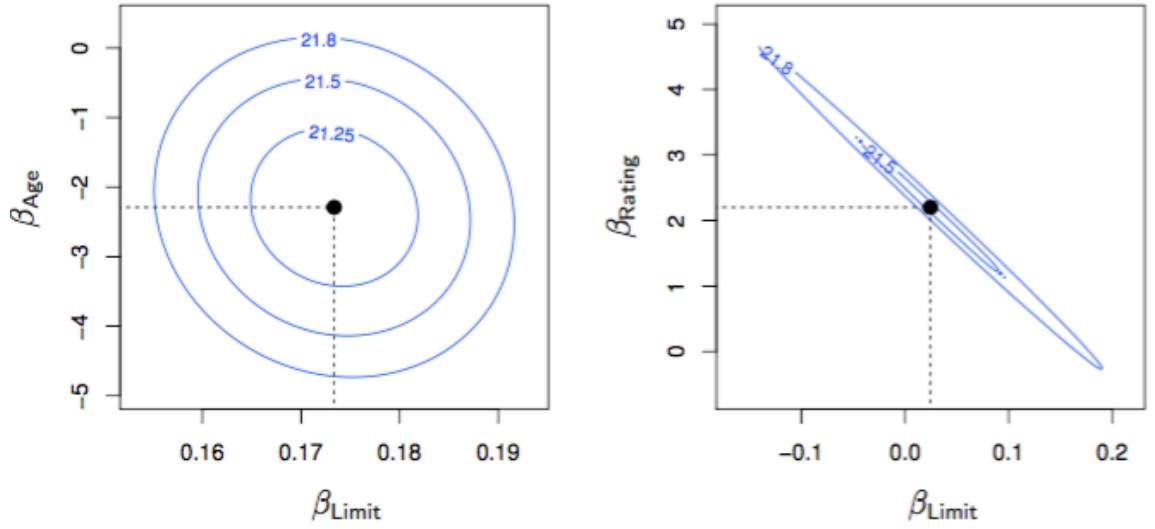




- **Problem:** Coefficient estimates become less certain and more variable (as training data changes). Consider the extreme case of using two identical predictors `limit`:

$$\begin{aligned}\texttt{balance} &= \beta_0 + \beta_1 \times \texttt{limit} + \beta_2 \times \texttt{limit} \\ &= \beta_0 + (\beta_1 + 100) \times \texttt{limit} + (\beta_2 - 100) \times \texttt{limit}\end{aligned}$$

The fit $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ is just as good as $(\hat{\beta}_0, \hat{\beta}_1 + 100, \hat{\beta}_2 - 100)$.



- If 2 variables are collinear, we can easily diagnose this using their correlation.

  A group of q variables is **multilinear** if one variable is approximately a linear function of the other variables. Pairwise correlations may not reveal multilinear variables.

  The Variance Inflation Factor (VIF) measures how linearly predictable a variable is from the other variables:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

  where $R^2_{X_j|X_{-j}}$ is the R^2 statistic for Multiple Linear regression of the predictor X_j onto the remaining predictors.

Training Error

- In order to evaluate the performance of a statistical model on a given data set, we use the *mean squared error* (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

  where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the ith observation
- The MSE will be small if the predicted responses are very close to the true responses, and will be large if for some of the observations, the predicted and true responses differ substantially.
- However, in general, we do not really care how well the model works on the training data. *We are interested in the accuracy of the predictions that we obtain when we apply our model to preciously unseen test data.*

Test(Prediction) Error

- We want to know whether $\hat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a previously *unseen test observation* not used to train the model.
- We want to choose the model that gives the lowest *test* MSE

$$\text{Ave}(y_0 - \hat{f}(x_0))^2,$$

as opposed to the lowest training MSE.

- The main challenge is that a low training MSE does not imply a low test MSE.

Model Selection

1. Best Subset

2. Stepwise Selection
   - Forward Stepwise Selection
   - Backward Stepwise Selection
   - Hybrid Approaches

3. Choosing the Optimal Model
   - Adjustment to the Training Error: $C_p, AIC, BIC, Adjusted\ R^2$
   - Directly Estimating the Test Error: Cross-Validation

What Do We Know So Far

- In linear regression, adding predictors always decreases the training error (or RSS).
- However, adding predictors does not necessarily improve the test error.
- When $n < p$, there is no least squares solution:

$$\hat{\beta} = \underbrace{(\mathbf{X}^T\mathbf{X})}_{\text{Singular}}^{-1}\mathbf{X}^T y.$$

So, we must find a way to select fewer predictors if we want to use least squares linear regression.

1. Best Subset

- Simple Idea: Let's compare all models with k predictors.
- There are $\binom{p}{k} = p!/[k!\,(p-k)!]$ possible models.

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

2. Stepwise Selection Methods

- Best subset selection has 2 problems:
  1. It is often very expensive computationally. We have to fit 2^p models.
  2. If for a fixed k, there are too many possibilities, we increase our chances of overfitting. The model selected has *high variance*.

- In order to mitigate these problems, we can restrict our search space for the best model.

- This reduces the variance of the selected model at the expense of an increase in bias.

1) Forward Selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

2) Backward Selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward vs. Backward Selection

- You cannot apply backward selection when $p > n$.

- Although it seems like they should, they need not produce the same sequence of models.

Example. $X_1, X_2 \sim \mathcal{N}(0, \sigma)$ independent.

$$X_3 = X_1 + 3X_2$$
$$Y = X_1 + 2X_2 + \epsilon$$

Regress Y onto X_1, X_2, X_3.

- Forward: $\{X_3\} \to \{X_3, X_2\} \to \{X_3, X_2, X_1\}$
- Backward: $\{X_1, X_2, X_3\} \to \{X_1, X_2\} \to \{X_2\}$

3) Hybrid Approaches

- **Mixed Stepwise Selection:** Do forward selection, but at every step, remove any variables that are no longer "necessary".
- **Forward Stagewise Selection:** Like forward selection, but when introducing a new variable, do not adjust the coefficients ot the previously added variables.
- **…**

3. Choosing the Optimal Model

- The model containing all of the predictors will always have the smallest RSS and the largest R^2, since these quantities are related to the training error.
- In order to select the best model with respect to test error, we need to estimate this test error. There are two common approaches:

  1. *Indirectly* estimate test error by making an *adjustment* to the training error to account for the bias due to overfitting.

  2. *Directly* estimate the test error, using either a validation set approach or a *cross-validation* approach.

$C_p, AIC, BIC, Adjusted\ R^2$

1. *Akaike Information Criterion* (AIC):

$$\frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2k\hat{\sigma}^2)$$

or C_p:

$$\frac{1}{n}(\text{RSS} + 2k\hat{\sigma}^2)$$

where $\hat{\sigma}^2$ is an estimate of irreducible error, typically $\hat{\sigma}^2 = \frac{1}{n-p-1} \times RSS$ of full model with all p predictors and where k is the number of predictors.

1. *Bayesian Information Criterion (BIC)*:

$$\frac{1}{n}(\text{RSS} + \log(n)k\hat{\sigma}^2)$$

2. *Adjusted* R^2:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $TSS = \sum_i (y_i - \bar{y})^2$

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}$$

3. Choosing the Optimal Model

- The model containing all of the predictors will always have the smallest RSS and the largest R^2, since these quantities are related to the training error.
- In order to select the best model with respect to test error, we need to estimate this test error. There are two common approaches:

  1. *Indirectly* estimate test error by making an *adjustment* to the training error to account for the bias due to overfitting.

  2. *Directly* estimate the test error, using either a validation set approach or a *cross-validation* approach.

Validation Set Approach

- **Goal**: Estimate the test error for a supervised model (Linear Regression, Classification etc.)
- **Strategy**:
  - Split the data in two parts.
  - Train the model in the first part.
  - Compute the error on the second part

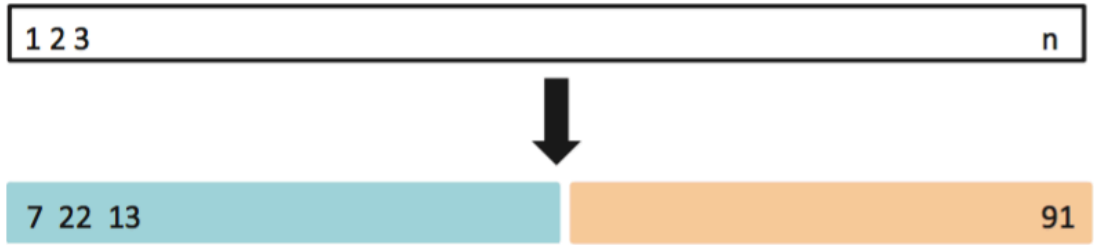


FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

- Polynomial regression to estimate `mpg` from `horsepower` in the Auto dataset.

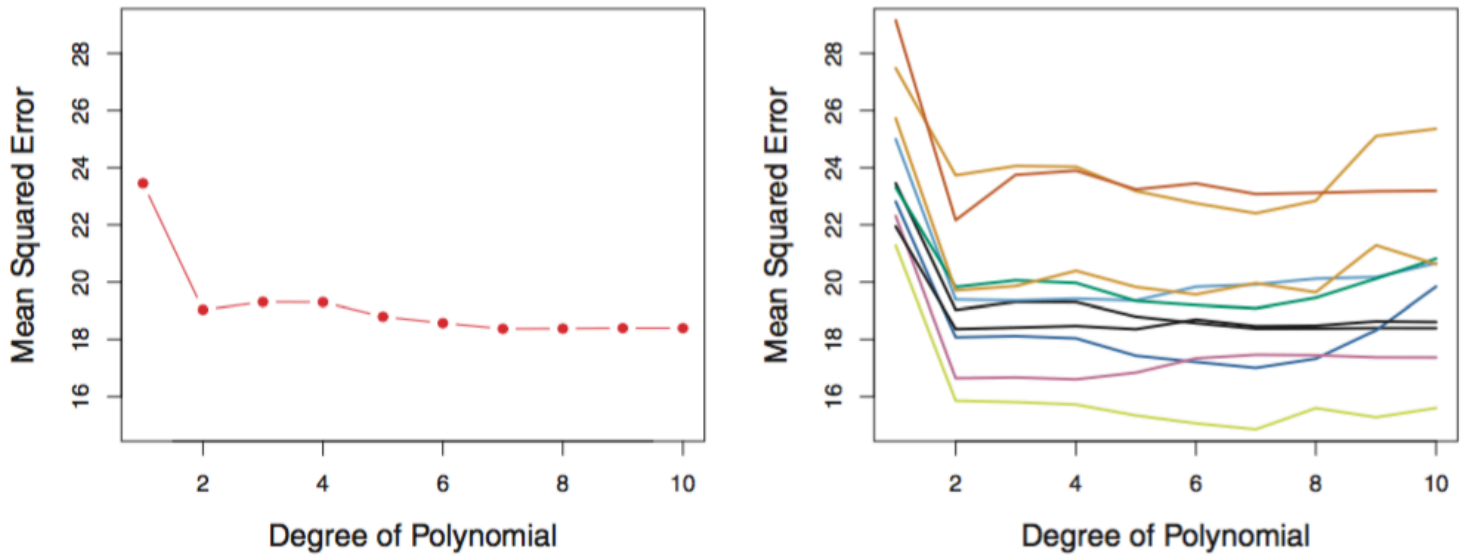


FIGURE 5.2. *The validation set approach was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *Validation error estimates for a single split into training and validation data sets.* Right: *The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach.*

Leave One Out Cross-Validation

- For every $i = 1, \ldots, n$:
  - Train the model on every point except i
  - Compute the test error on the held out point
  - Average the test errors.

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2$$

For a regression problem

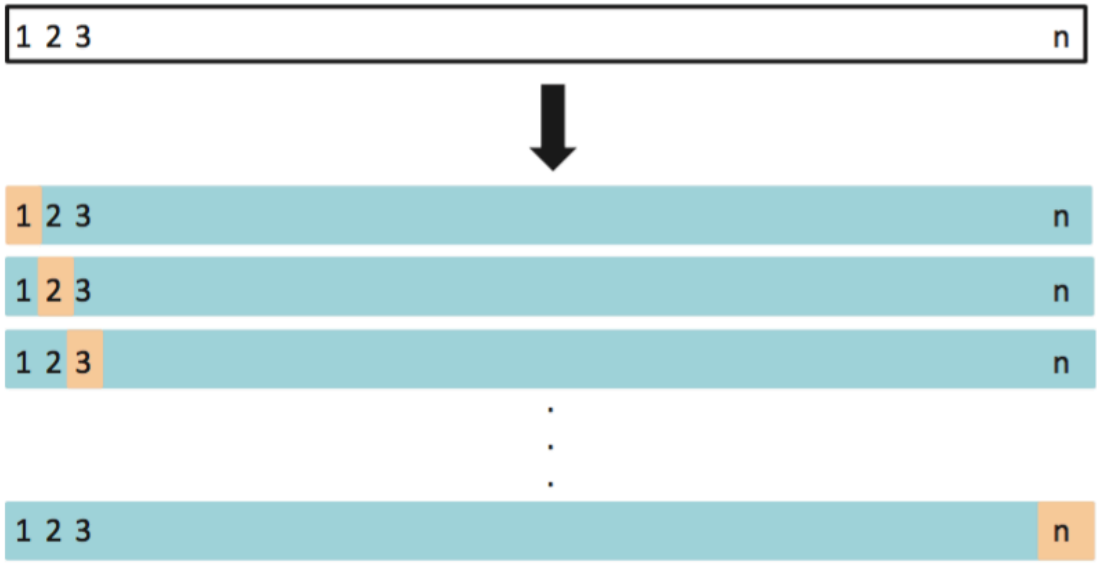


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

- Computing $CV_{(n)}$ can be computationally expensive, since it involves fitting the model n times.

  For linear regression, there is a shortcut:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}\right)^2$$

  where h_{ii} is the leverage statistic.

K-fold Cross-Validation

- Split the data into k subsets or *folds*.
- For every $i = 1, \ldots, k$:
  - Train the model on every fold except the ith fold
  - Compute the test error on the ith fold
  - Average the test errors.

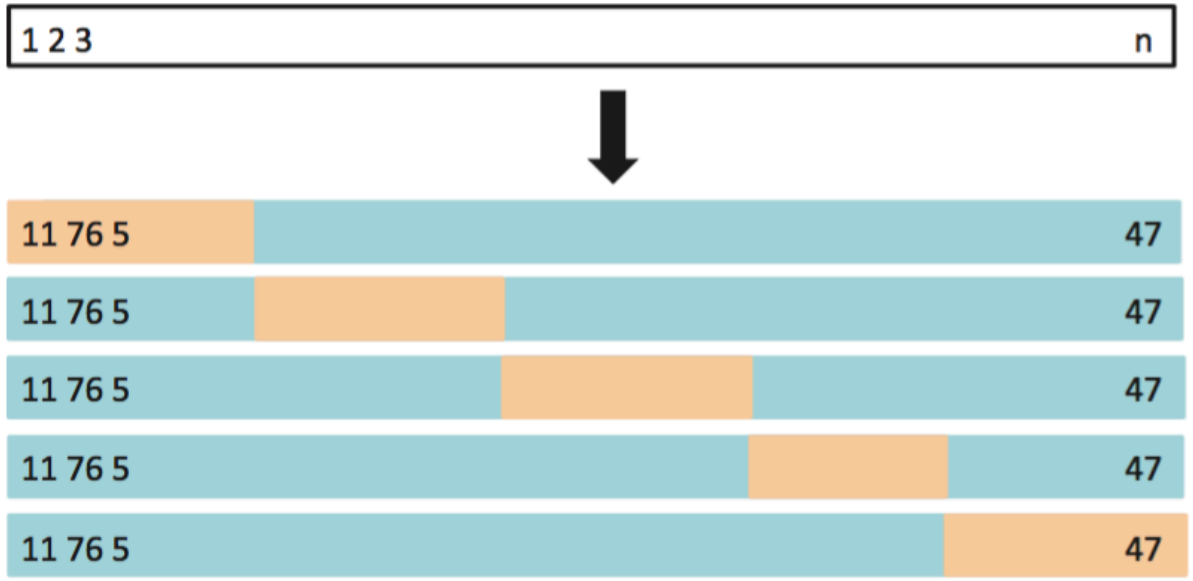
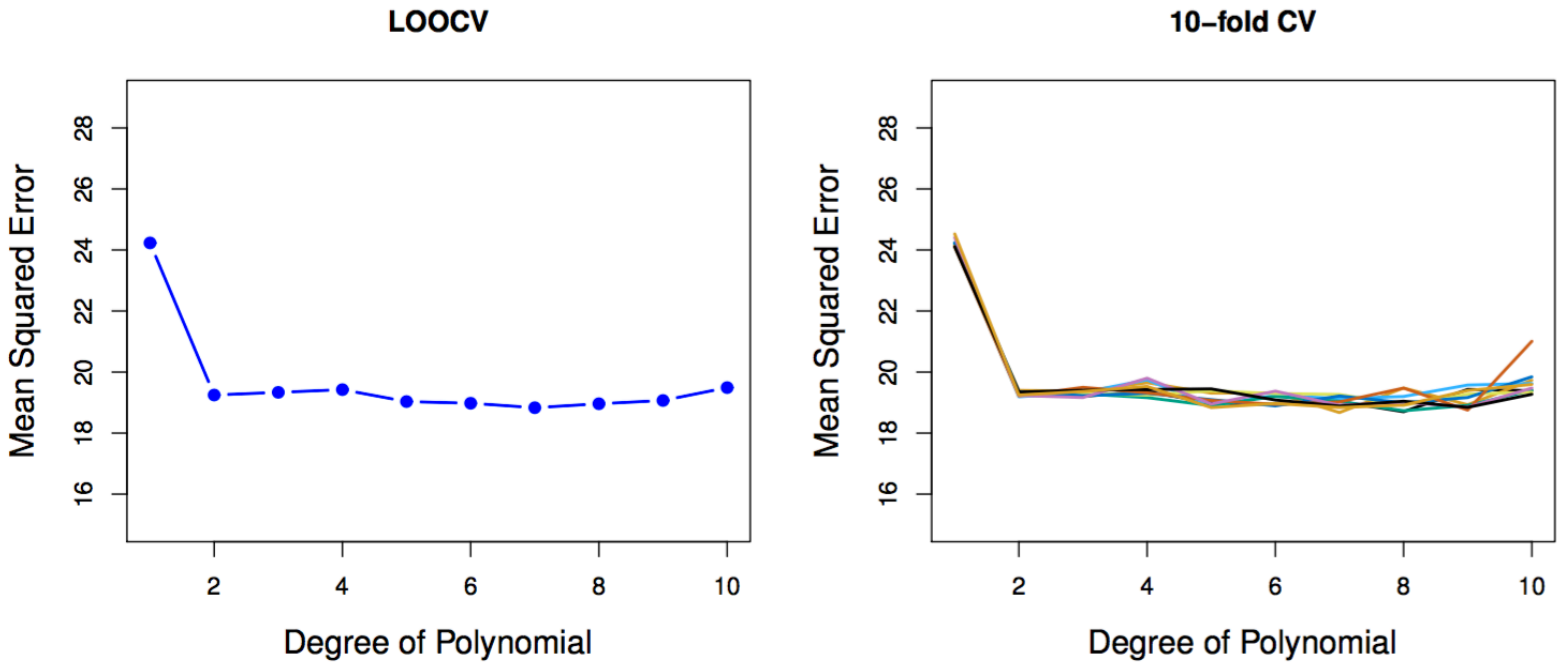


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

LOOCV vs. k-fold Cross-Validation



- k-fold CV depends on the chosen split.
- In k-fold CV, we train the model on less data than what is available. This introduces **bias** into the estimates of test error.
- In LOOCV, the training samples highly resemble each other. This increases the **variance** of the test error estimate.