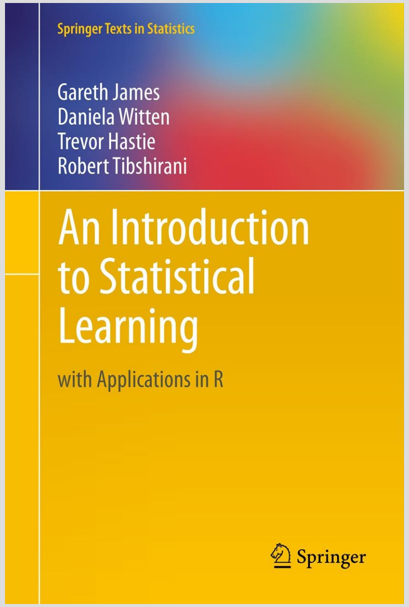
2.2.1 MEASURING THE QUALITY OF FIT



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

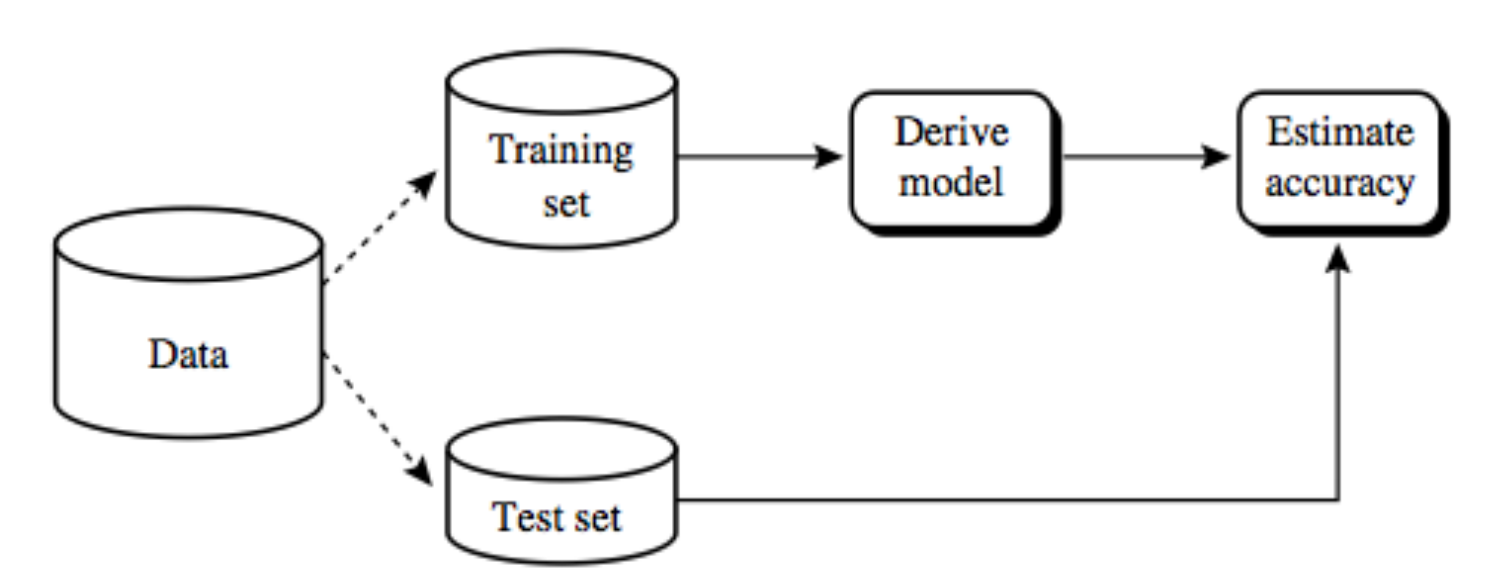
Springer

Section 2.2.1

http://faculty.marshall.usc.edu/gareth-james/ISL/

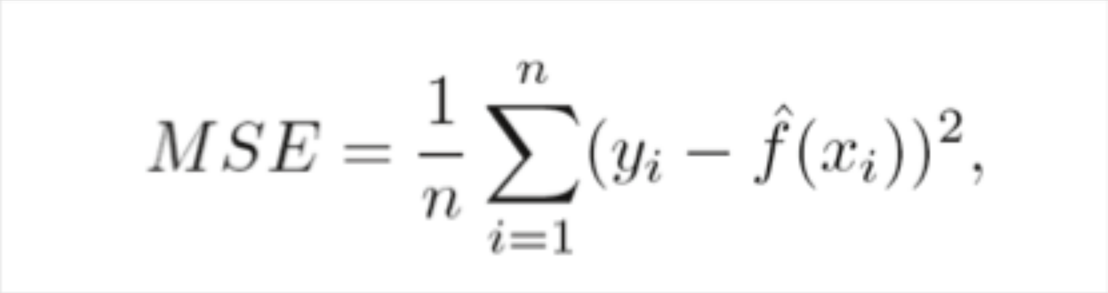
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

- **Mean Squared Error** measures how well y is predicted by function that we have learned from data
  - MSE is a criterion by which we fit regression models (e.g. least squares in linear regression)
  - For purpose of comparing models, one can also use sum of squared errors (= n*MSE) or **root-mean-square error (RMSE) =** sqrt(MSE) which is measured in the original units of the outcome variable y
- Does small MSE in **training data** it automatically generalize to new **test data ?**

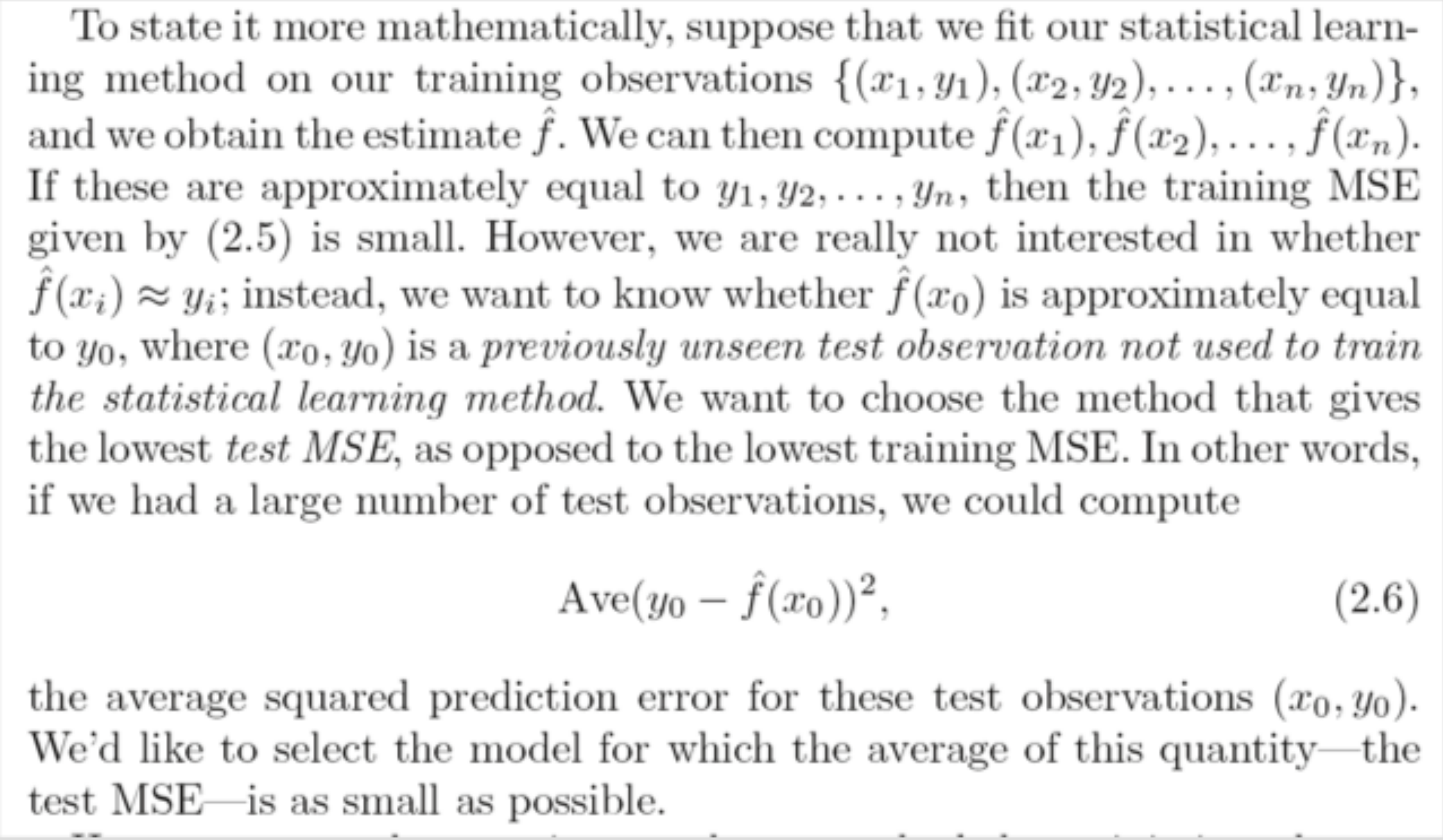


- Does small MSE in **training data** automatically generalize to new **test data ?**
  - Will MSE be equally small in test data that it is in training data?
  - In the end, we want to predict something we don't yet know
    - If all we can predict well is training data, that is not useful
    - We want to predict well also in unseen test data

Figure from
www.nosimpler.me/



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



To state it more mathematically, suppose that we fit our statistical learning method on our training observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, and we obtain the estimate $\hat{f}$. We can then compute $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n)$. If these are approximately equal to $y_1, y_2, \ldots, y_n$, then the training MSE given by (2.5) is small. However, we are really not interested in whether $\hat{f}(x_i) \approx y_i$; instead, 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 statistical learning method*. We want to choose the method that gives the lowest *test MSE*, as opposed to the lowest training MSE. In other words, if we had a large number of test observations, we could compute

$$\text{Ave}(y_0 - \hat{f}(x_0))^2, \tag{2.6}$$

the average squared prediction error for these test observations (x_0, y_0). We'd like to select the model for which the average of this quantity—the test MSE—is as small as possible.

How model flexibility affects training error and test error

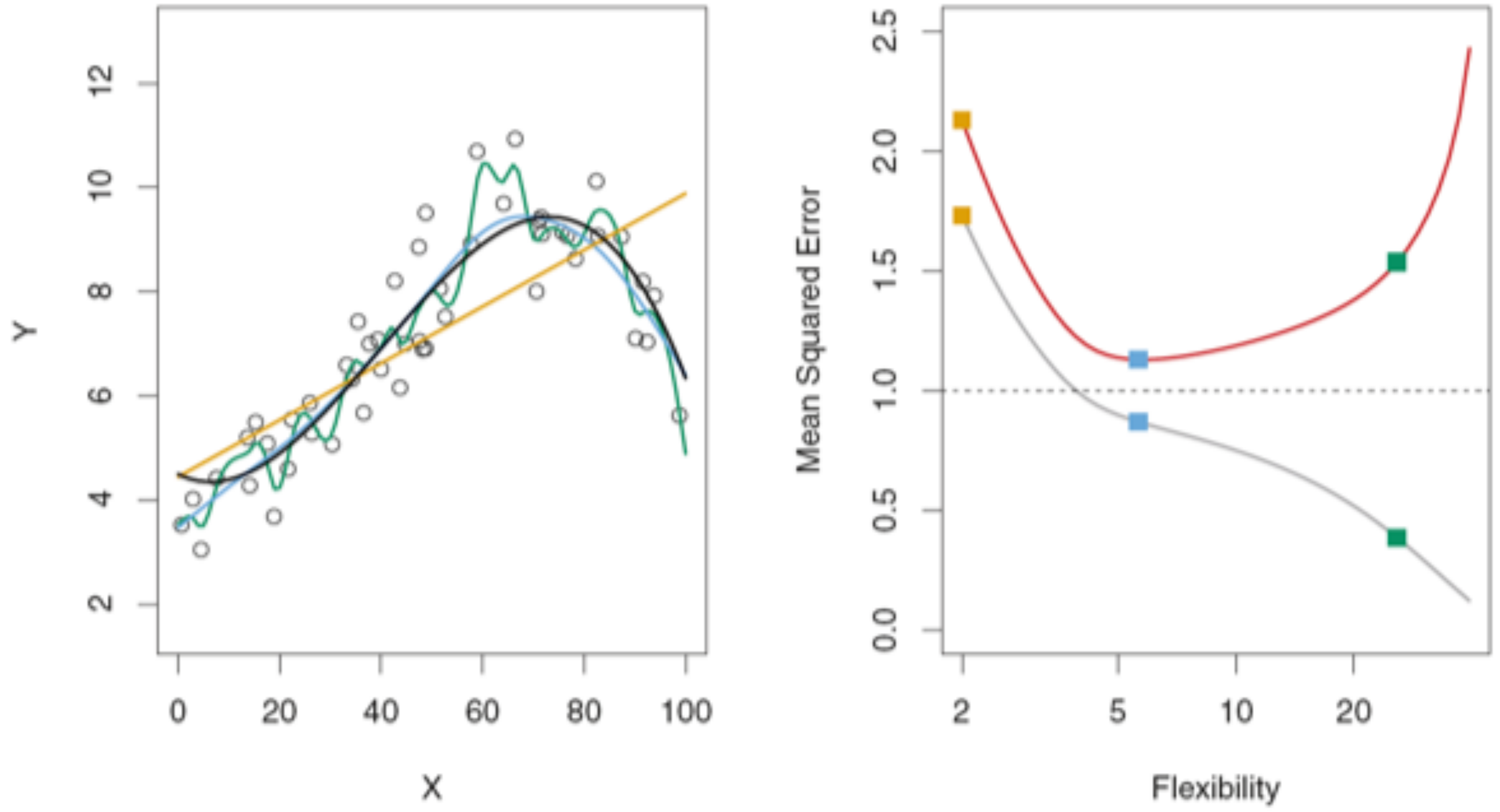


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

Training error is Monotonically decreasing

Test error has U-Shape

Overfitting:
Model has adapted to patterns in training data that are specific to training data and not generalizable to test data

Overfitting leads to small training error but large test error

Overfitting has happened when a less flexible model would have given a lower test error than observed

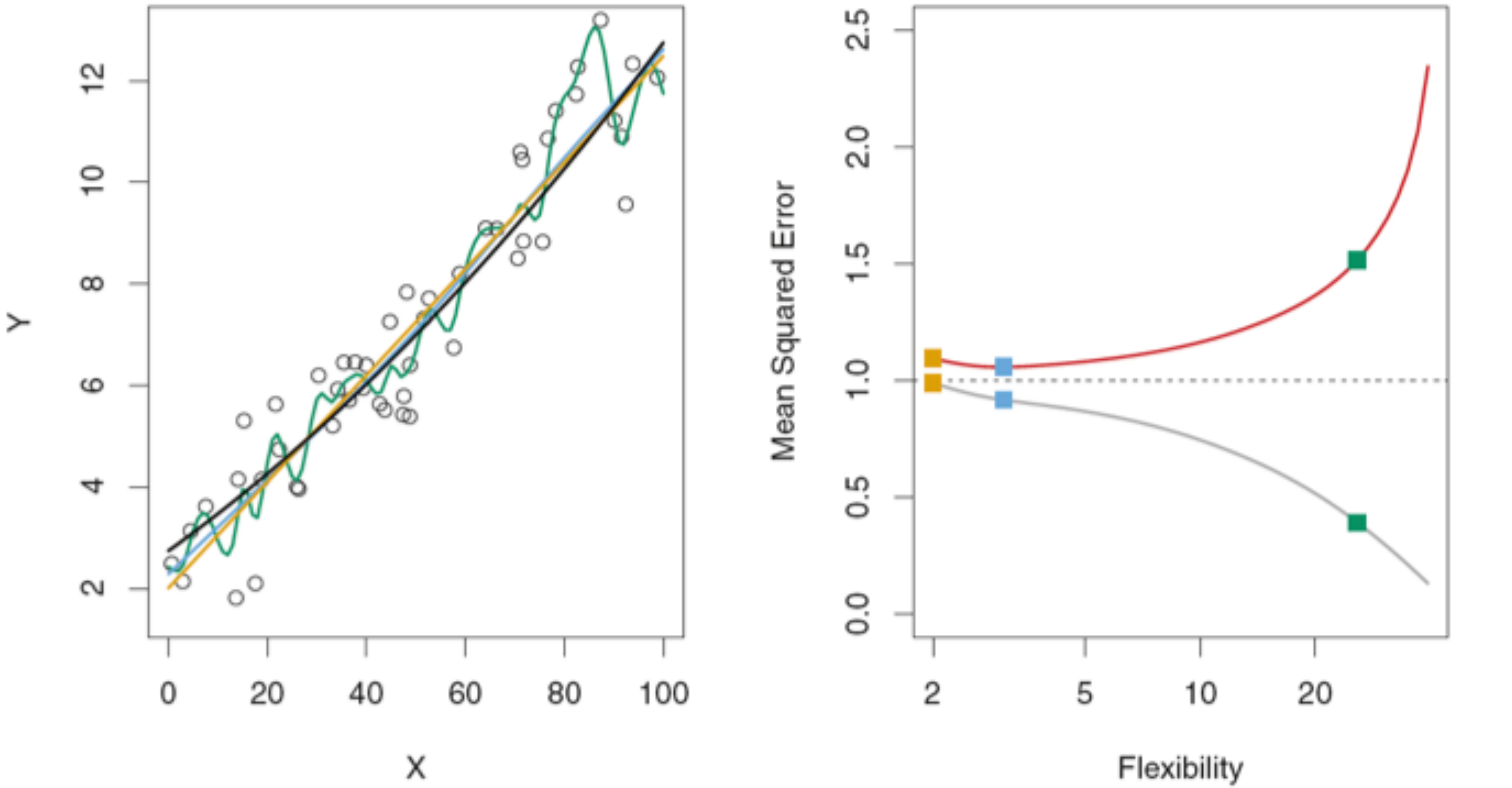


FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

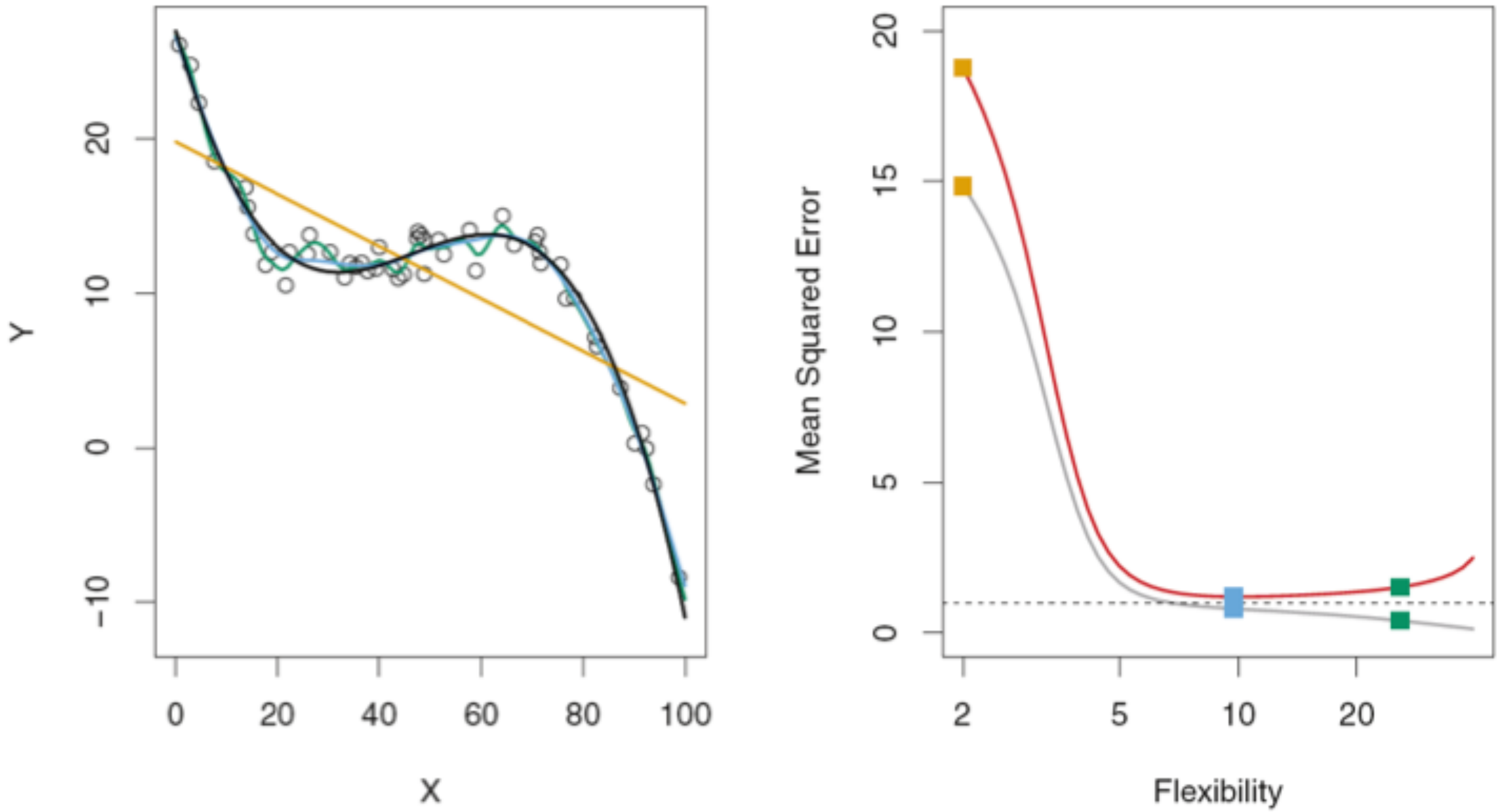
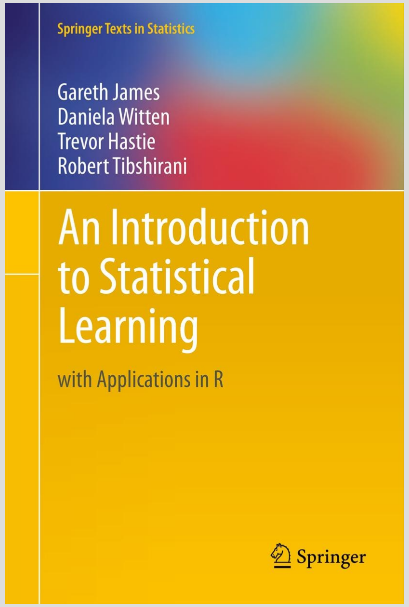


FIGURE 2.11. *Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.*

2.2.2 BIAS-VARIANCE TRADE-OFF



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

Section 2.2.2

http://faculty.marshall.usc.edu/gareth-james/ISL/

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

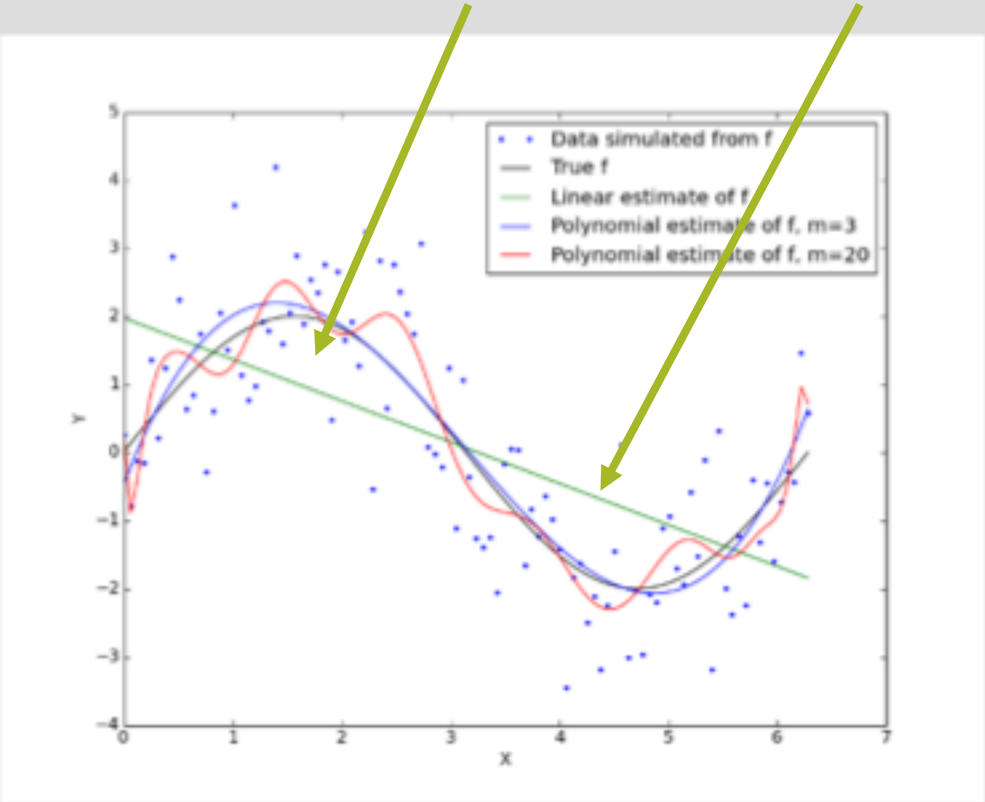
“variance” “bias” “irreducible error”

- Test error at predictor values x_0 is sum of 3 components:
  - Variance of the regression function estimator
  - Squared bias of the regression function estimator
  - Irreducible error that the regression function cannot account for
- To make error small we want small variance AND small bias^2
  - We can’t do anything to the irreducible error (unless we got some new predictors that could explain it, whence it would not anymore be “irreducible”)

$$\text{Bias}(x_0) = \mathrm{E}(\hat{f}(x_0)) - f(x_0)$$

- Bias of estimator of true regression function f, at value x_0, tells how much, on average, the predicted value differs from the truth
- The expectation is taken when the model is fitted over many data sets and each provides a different estimate for the function
- Typically bias is high if the method used for estimating f is not flexible enough to fit true shape of f well
  - E.g. fitting a linear model to 3 degree polynomial leads to high bias

Linear model has
downward bias here & upward bias here

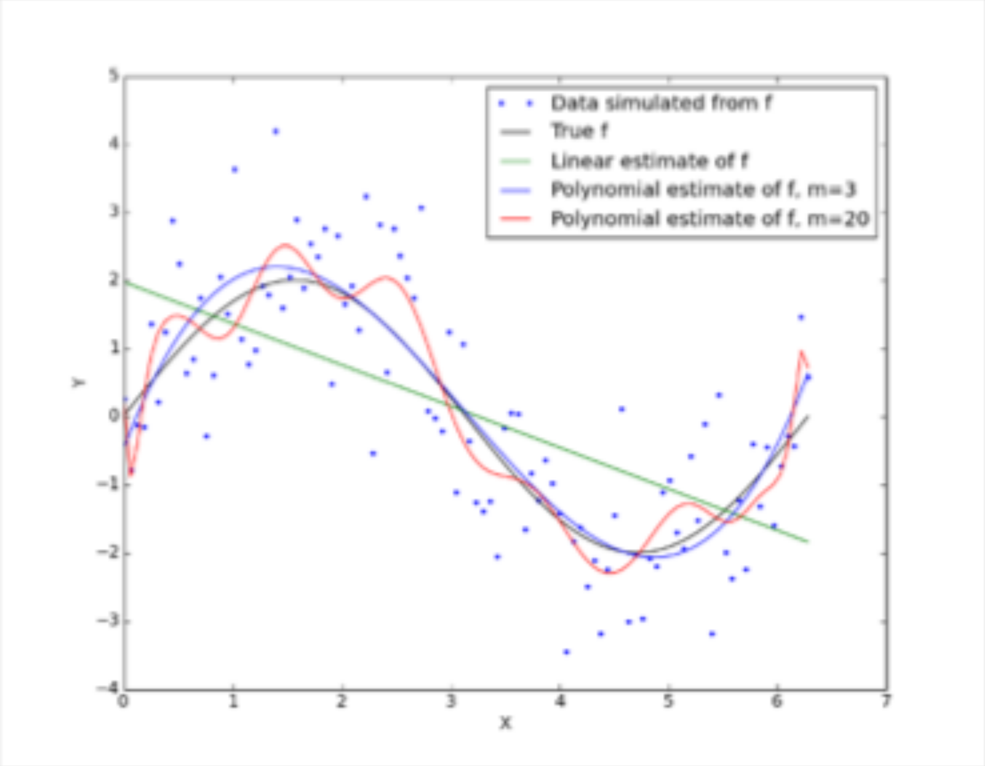


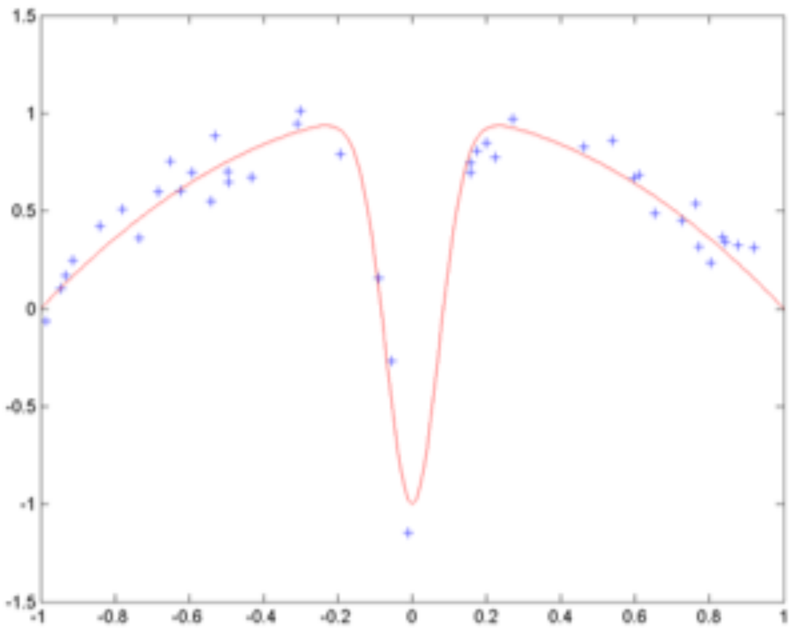
Michael Halls-Moore, Quantstart.com

$$\mathrm{Var}(\hat{f}(x_0)) = \mathrm{E}(\hat{f}(x_0) - \mathrm{E}(\hat{f}(x_0)))^2$$

Red polynomial fit is having
high variance as it would
change considerablt with resampling
a new data set

- Variance of estimator of true regression function f, at value x_0, tells how much, on average, the predicted value varies across data sets
- The expectation is taken when the model is fitted over many data sets and each provides a different estimate for the function
- Typically variance is high if the method used for estimating f is very flexible and adjusts to the specific properties of each observed data set that may change across data sets
  - E.g. fitting a 20 degree polynomial to data that originates from a cubic polynomial leads to high variance whereas a linear model fit there would have low variance (but higher bias)

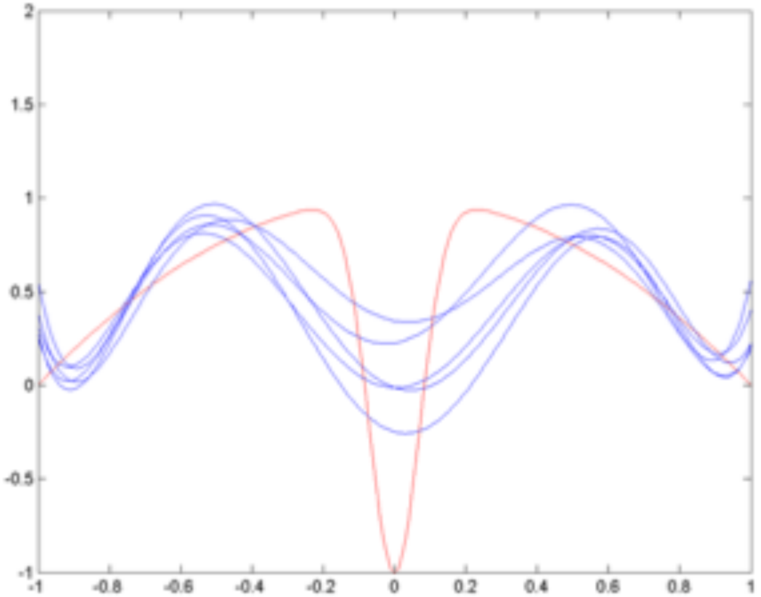


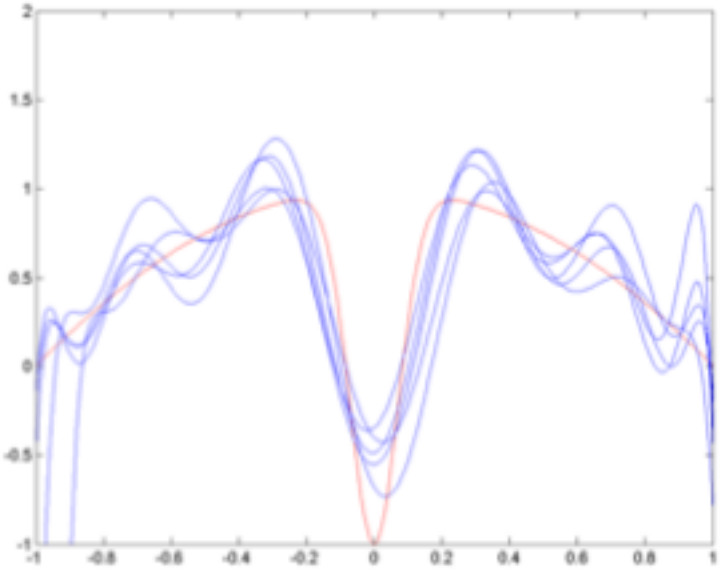


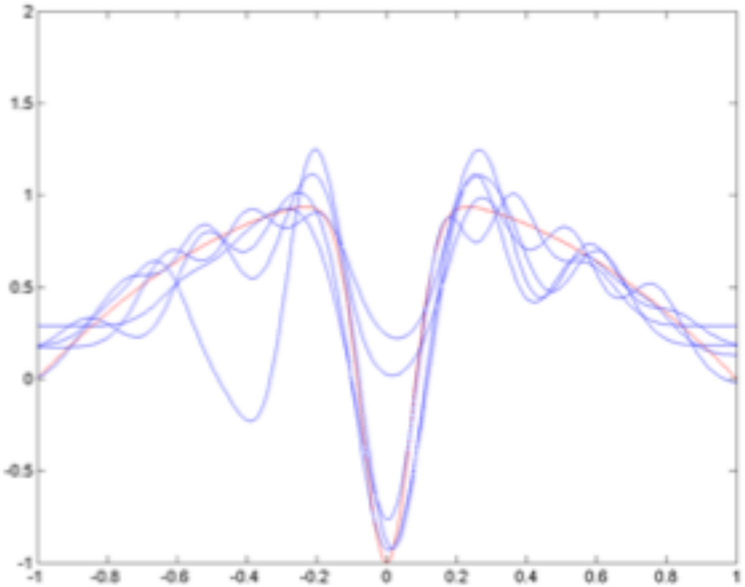
Data (blue) and
true function (red) on left

Which is the ranking of the 3 function estimators in terms of bias and variance?

Each blue curve is one fitted estimate of the red function



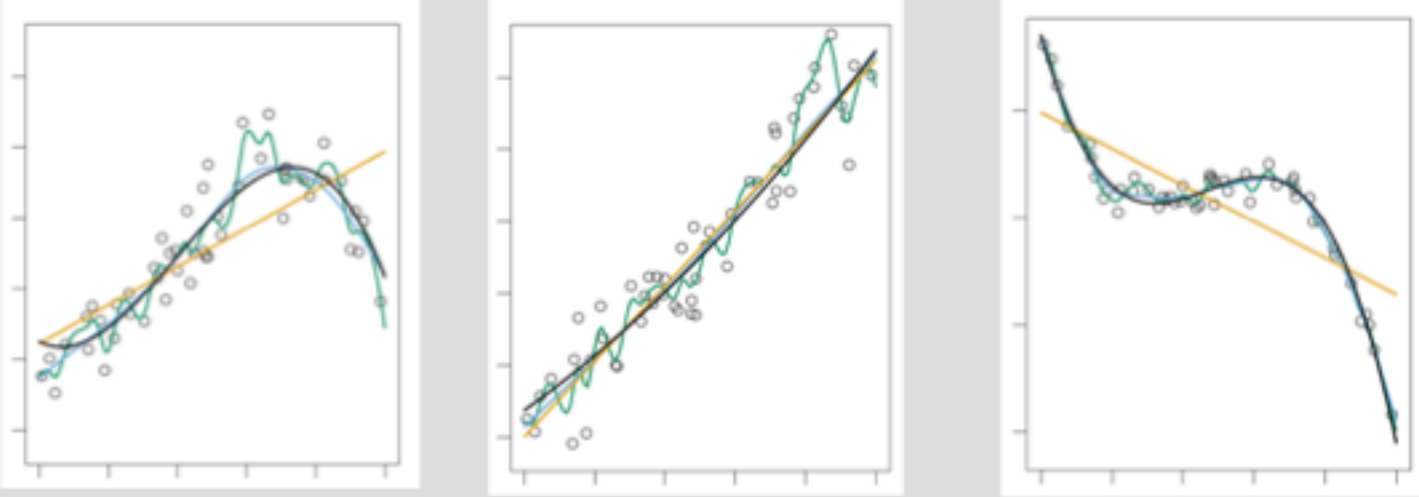




Figs by Andres Sandberg, Wikipedia

TRADE-OFF

- As we use more flexible methods, the variance will increase and the bias will decrease having opposite effects on test error
  - As we increase the flexibility methods, the bias tends to initially decrease faster than the variance increases and MSE declines
  - At some point increasing flexibility has little impact on the bias but starts to significantly increase the variance and MSE increases
  - This is the reason for U-shaped test error curves as a function of model flexibility



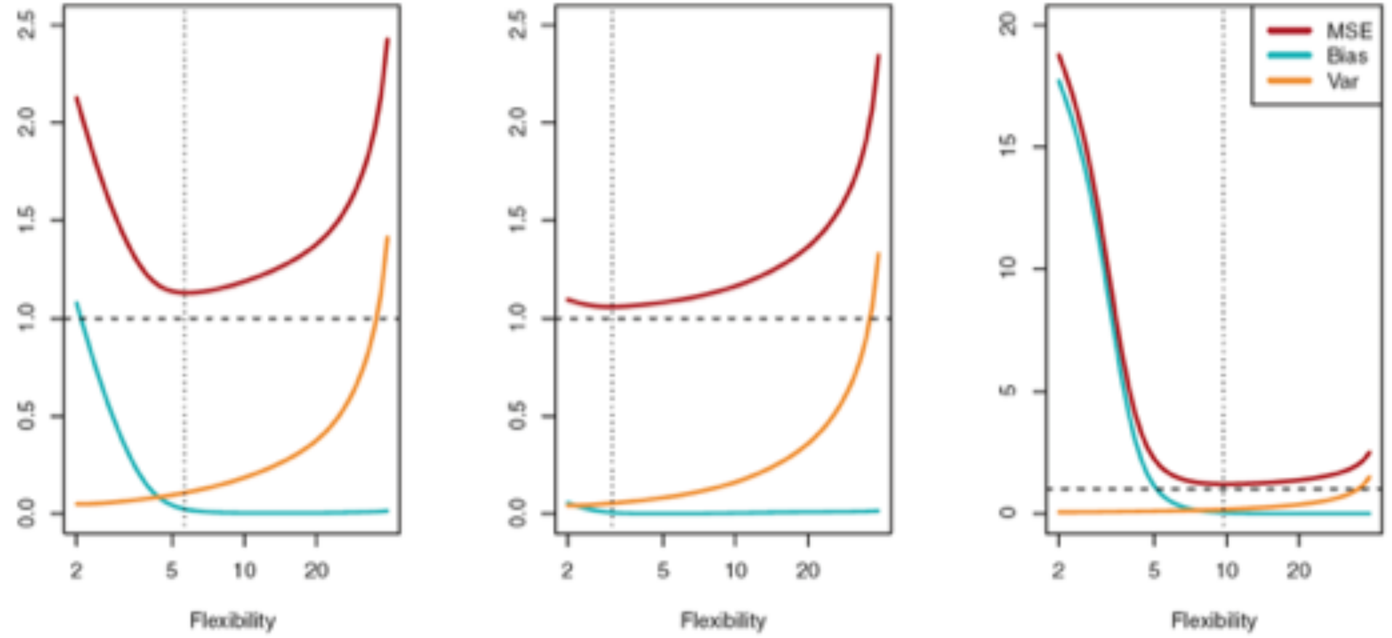


FIGURE 2.12. *Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.*

K-NEAREST NEIGHBORS CLASSIFIER

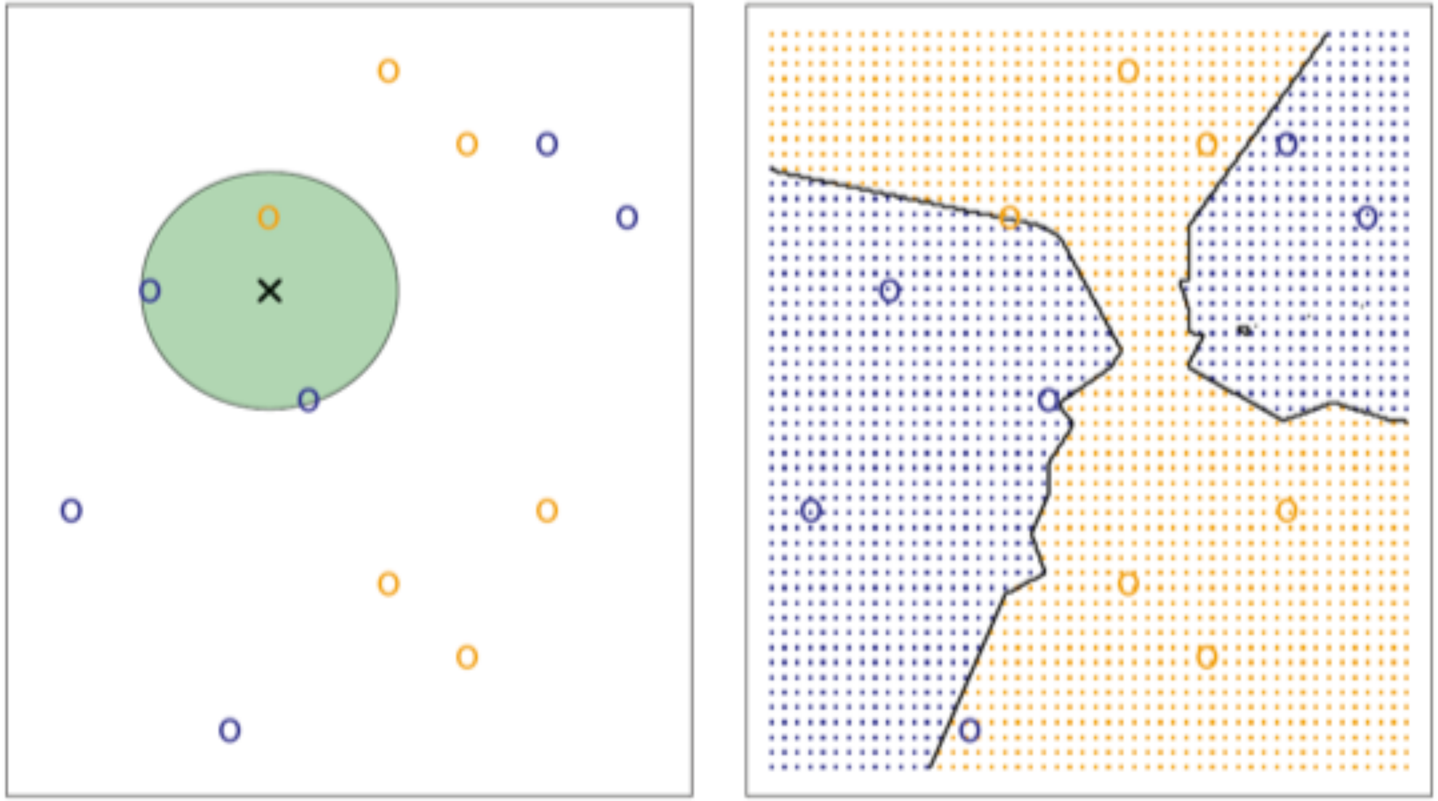


FIGURE 2.14. *The KNN approach, using $K = 3$, is illustrated in a simple situation with six blue observations and six orange observations.* Left: *a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue.* Right: *The KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.*

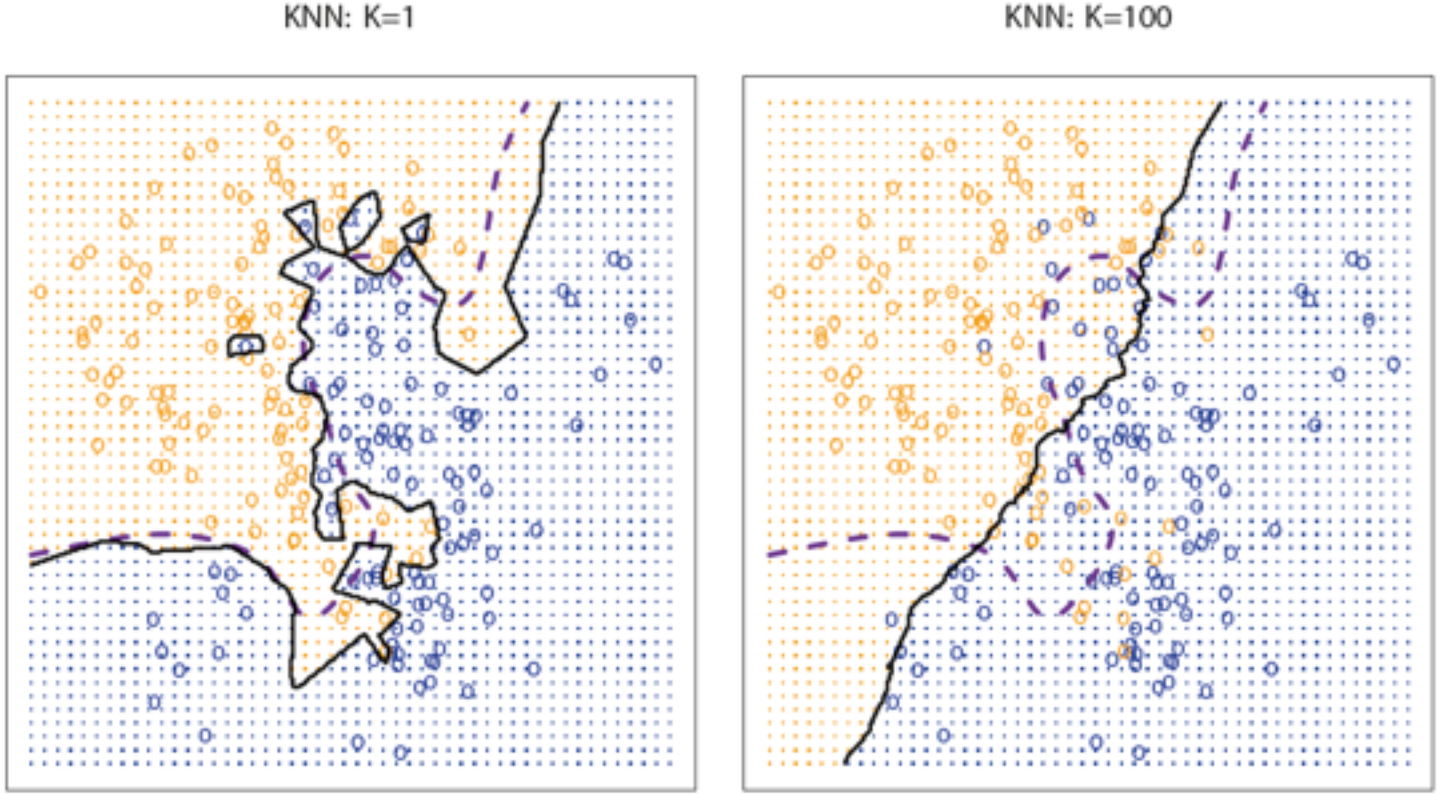


FIGURE 2.16. *A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.*

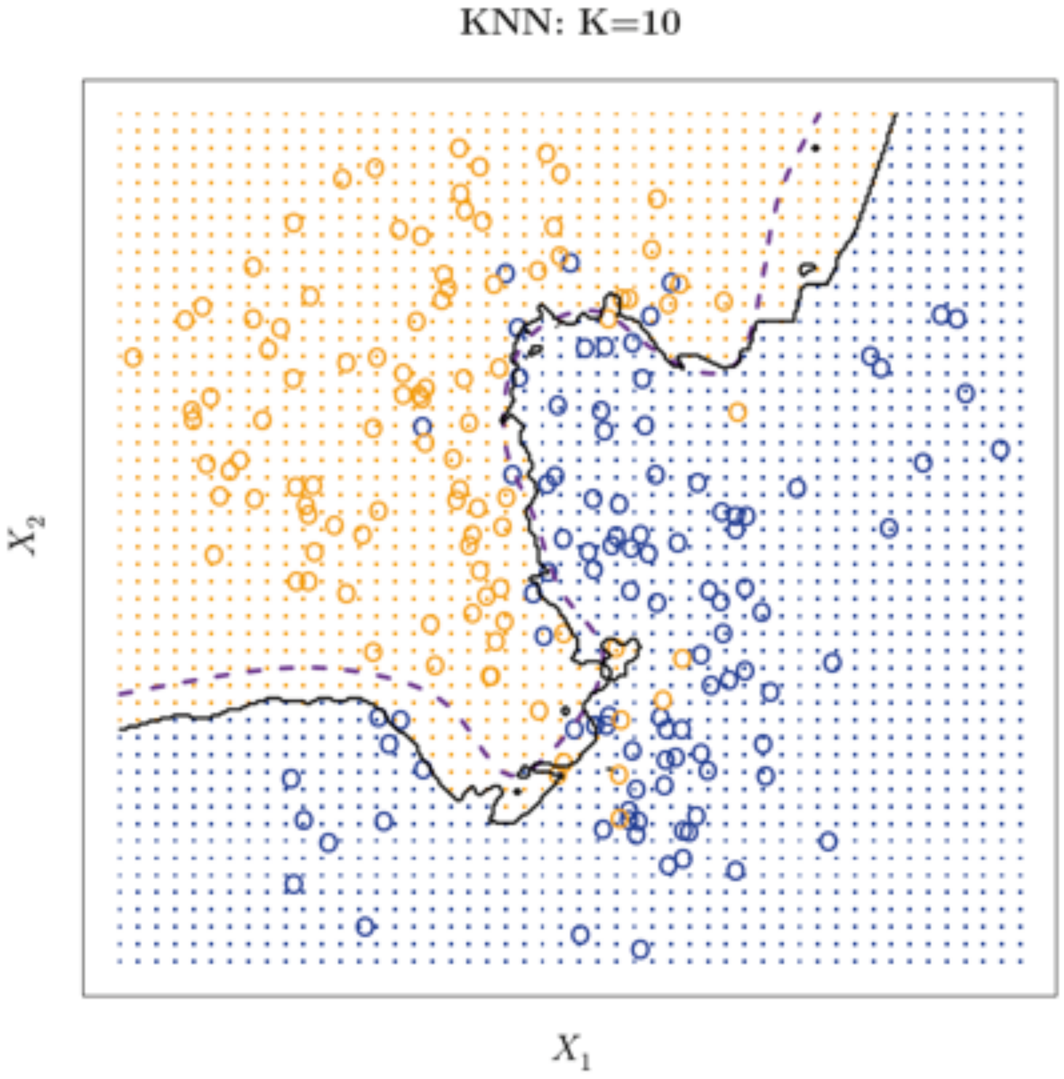


FIGURE 2.15. *The black curve indicates the KNN decision boundary on the data from Figure 2.13, using $K = 10$. The Bayes decision boundary is shown as a purple dashed line. The KNN and Bayes decision boundaries are very similar.*

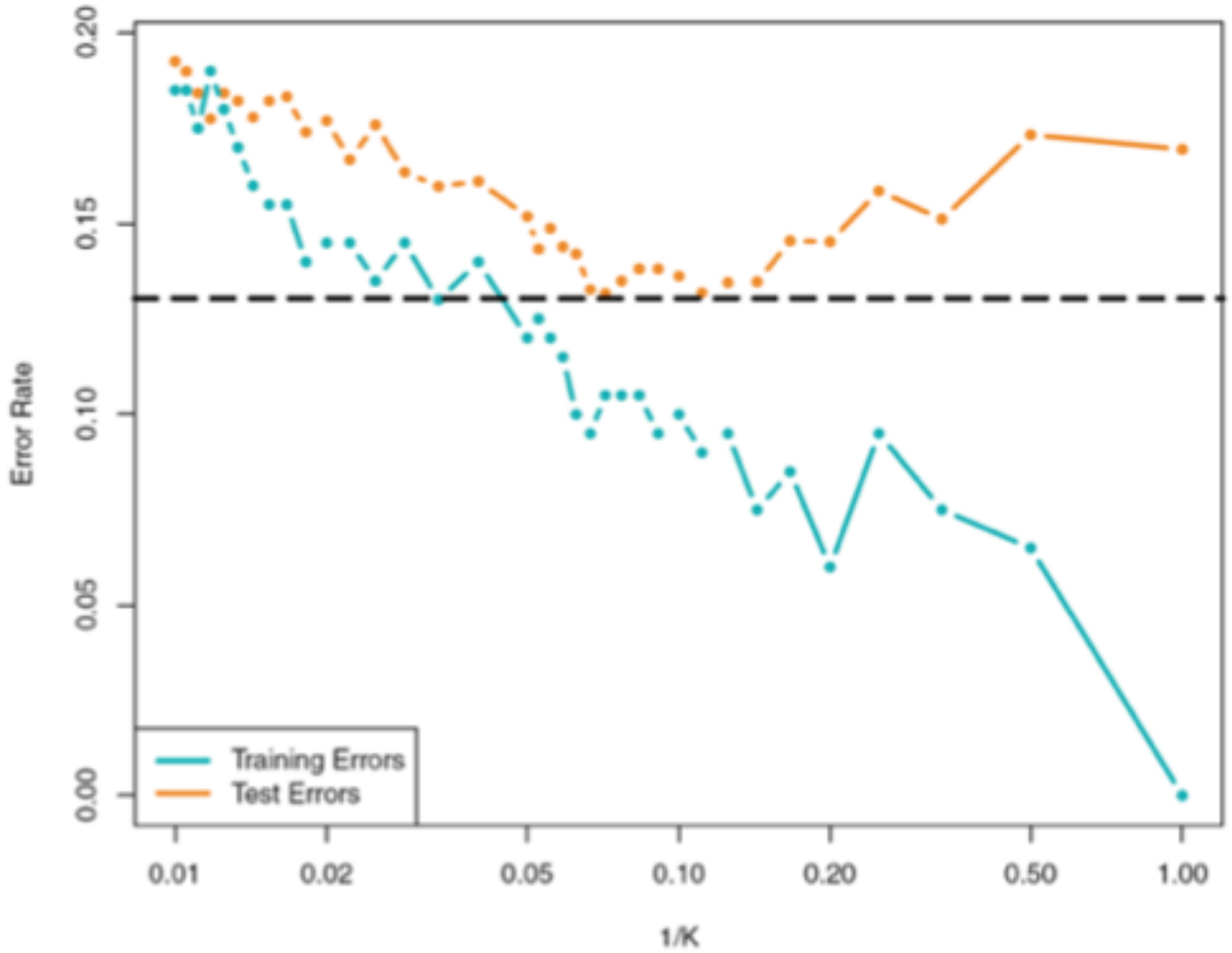
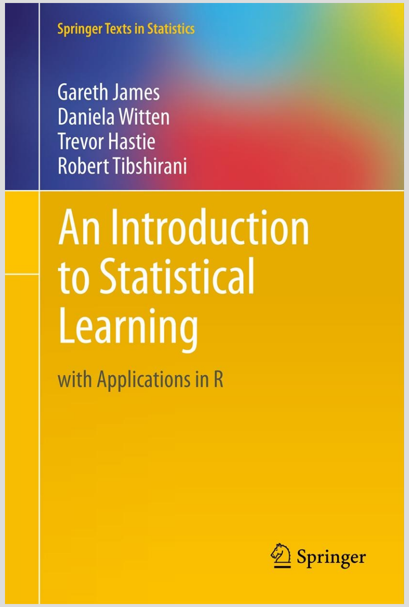


FIGURE 2.17. *The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using $1/K$) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training data set.*

6.2.1 BEST SUBSET SELECTION & 6.2.2 STEPWISE SELECTION



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

Section 6.2.1 – 6.2.2

http://faculty.marshall.usc.edu/gareth-james/ISL/

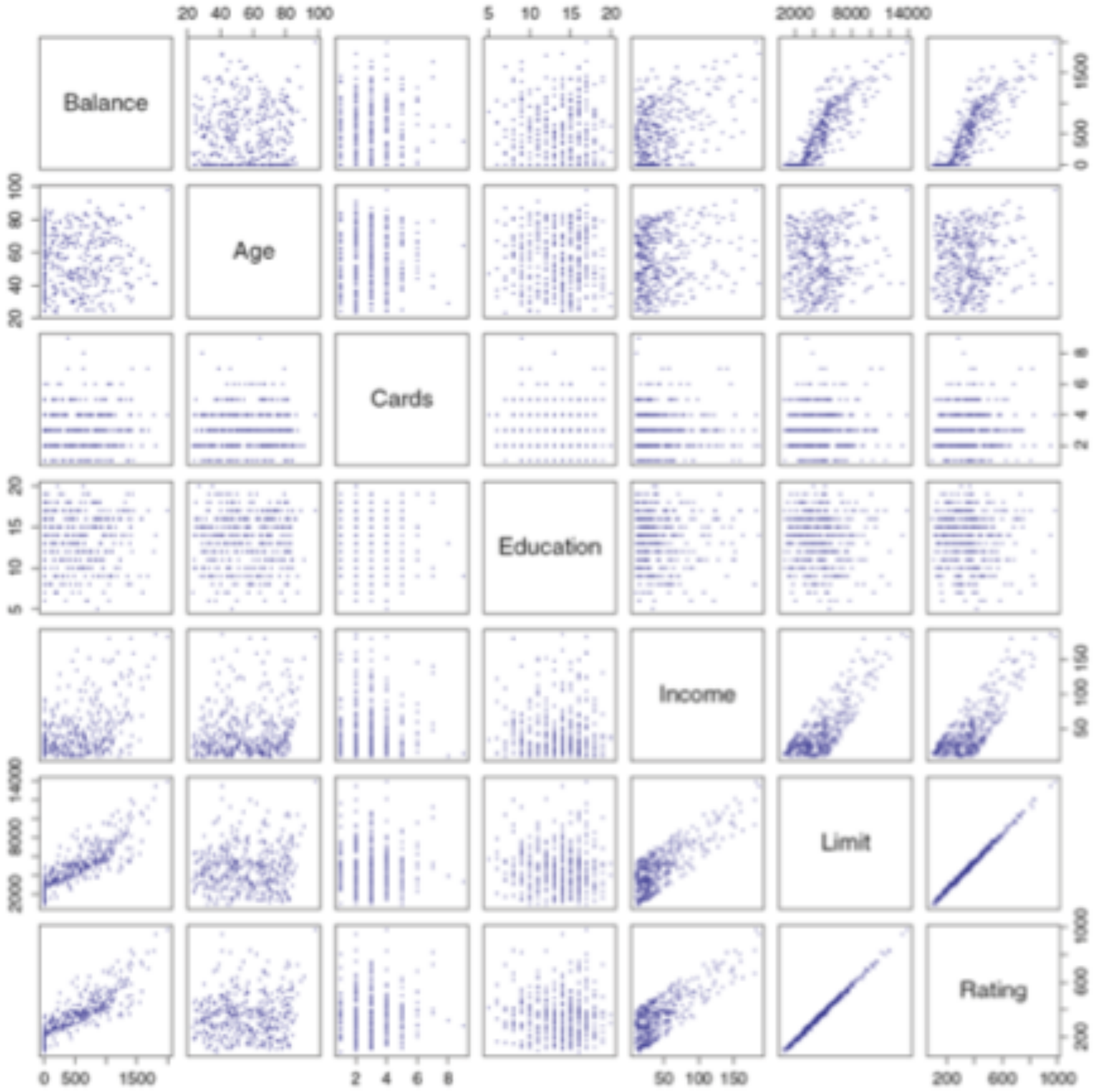
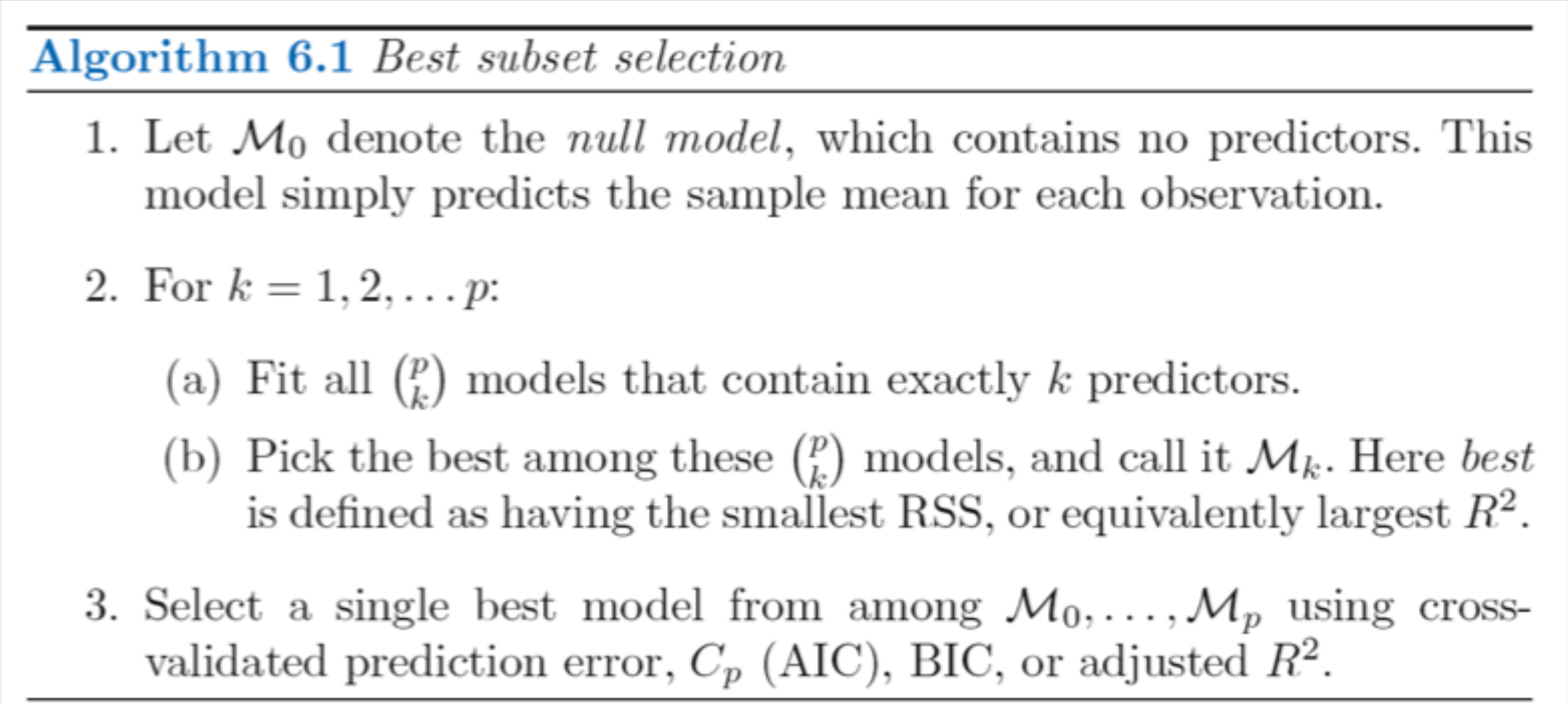


FIGURE 3.6. *The* Credit *data set contains information about* balance, age, cards, education, income, limit, *and* rating *for a number of potential customers.*

How to choose which predictors to include?

Here p<10, but how to extend the methods to cases with p in 100s?



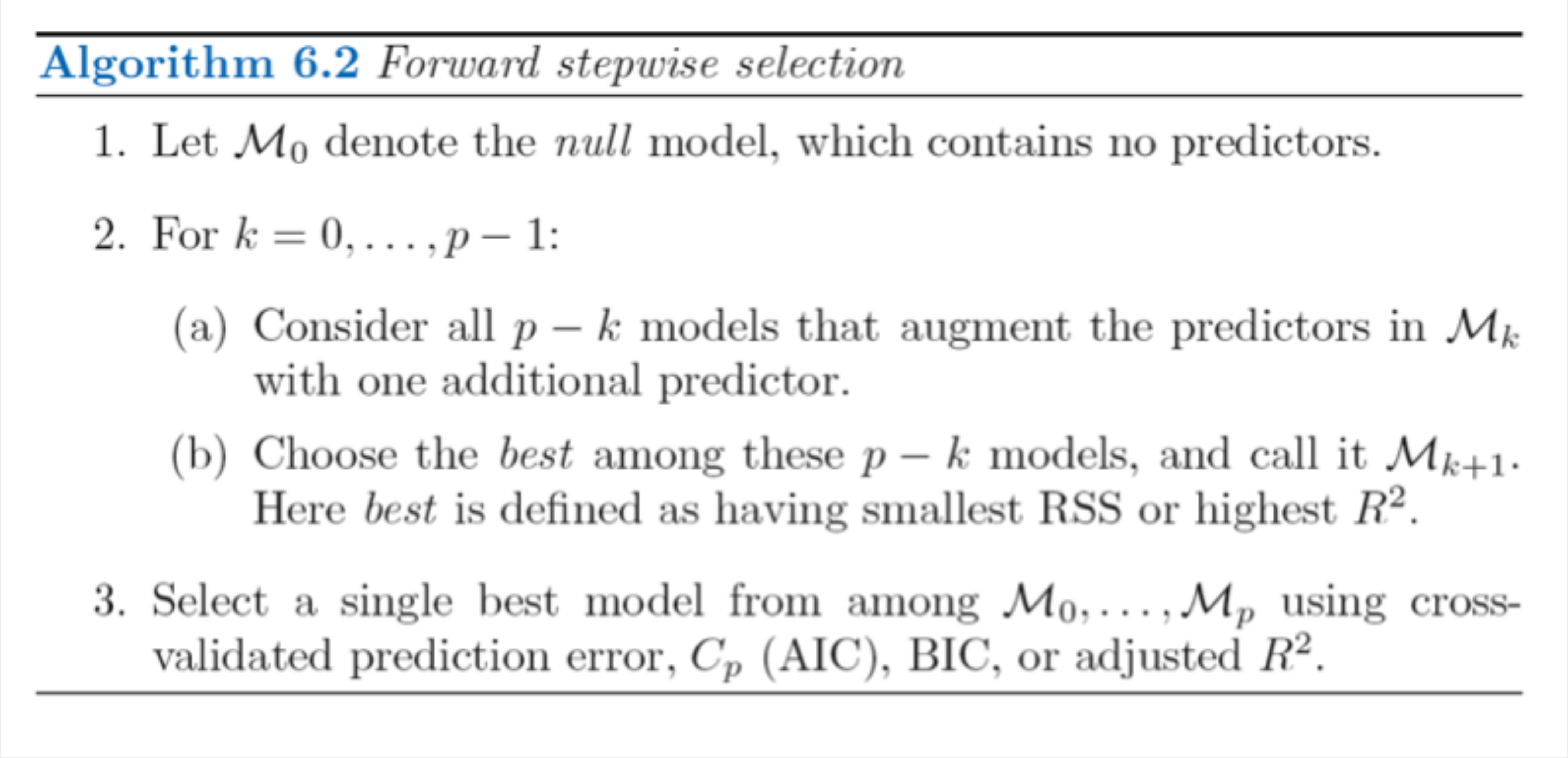
Algorithm 6.1 *Best subset selection*

```
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 $R^2$.
```

(We will look cross-validation soon. C_p is our AIC.)

2^p possible submodels grows too quickly for practical use when p is large

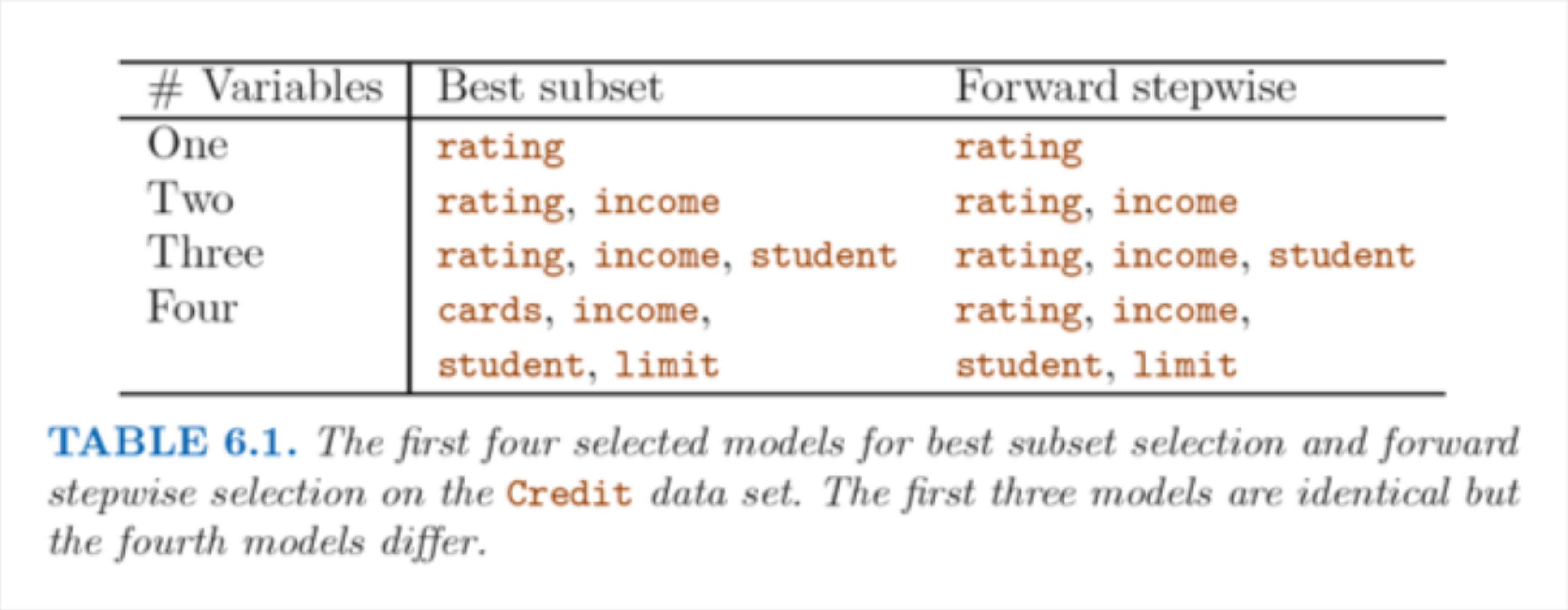
p=10: ~10^3 models, p=20: ~10^6, p=30: ~10^9 …



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.

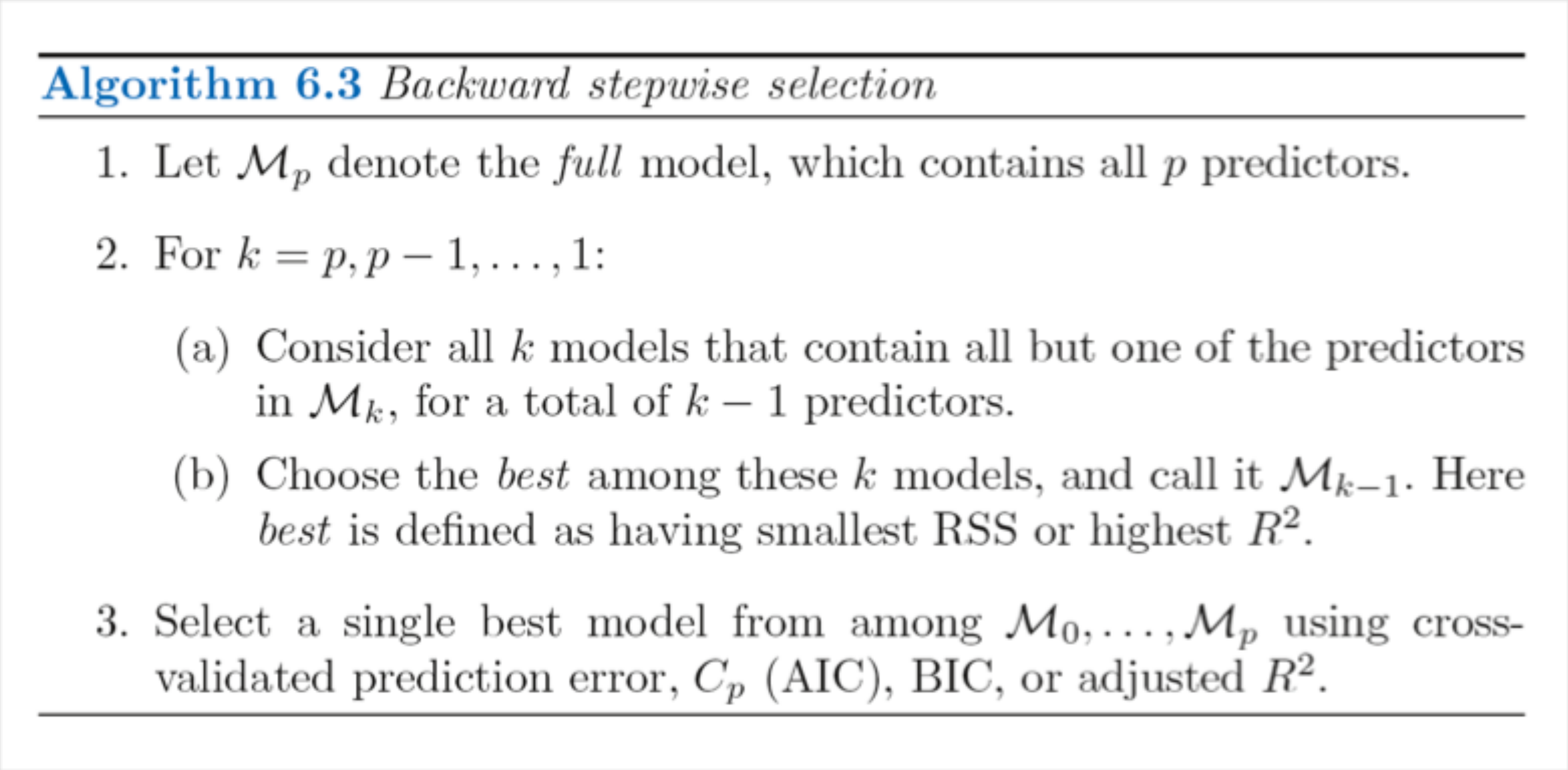
Now we have at most $(p^2+p)/2$ models to fit. Much better than 2^p.



# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

TABLE 6.1. *The first four selected models for best subset selection and forward stepwise selection on the* Credit *data set. The first three models are identical but the fourth models differ.*

Forward selection does not always pick the "best" model because the best model with k+1variables is not necessarily a superset of best model with k variables.

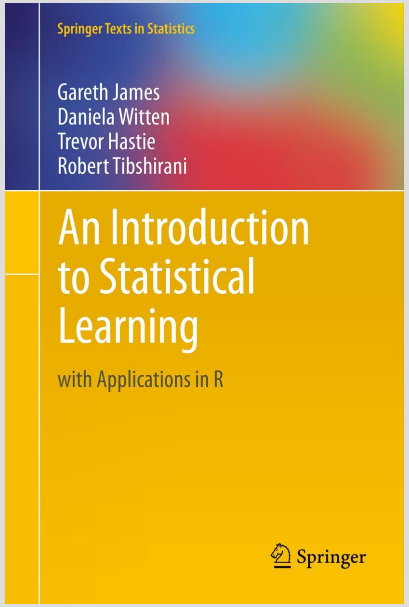


Algorithm 6.3 *Backward stepwise selection*

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.

Backward selection only possible when n>p, otherwise we cannot fit the full model with p variables.

CROSS-VALIDATION



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

Sections 5.1.1-5.1.4

http://faculty.marshall.usc.edu/gareth-james/ISL/

VALIDATION SET

- Goal is to estimate **test error** (i.e. the error that would be expected in a new unseen data) using existing data
- We can split the existing data into two parts: **training** and **validation** sets
  - Fit the model in training data
  - Estimate the error in validation set, that mimics an unseen test data set

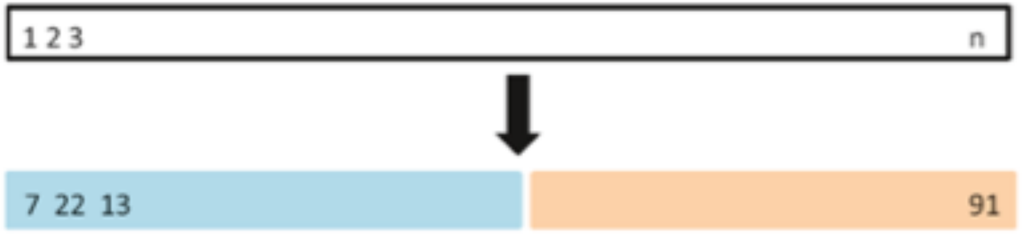


FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning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

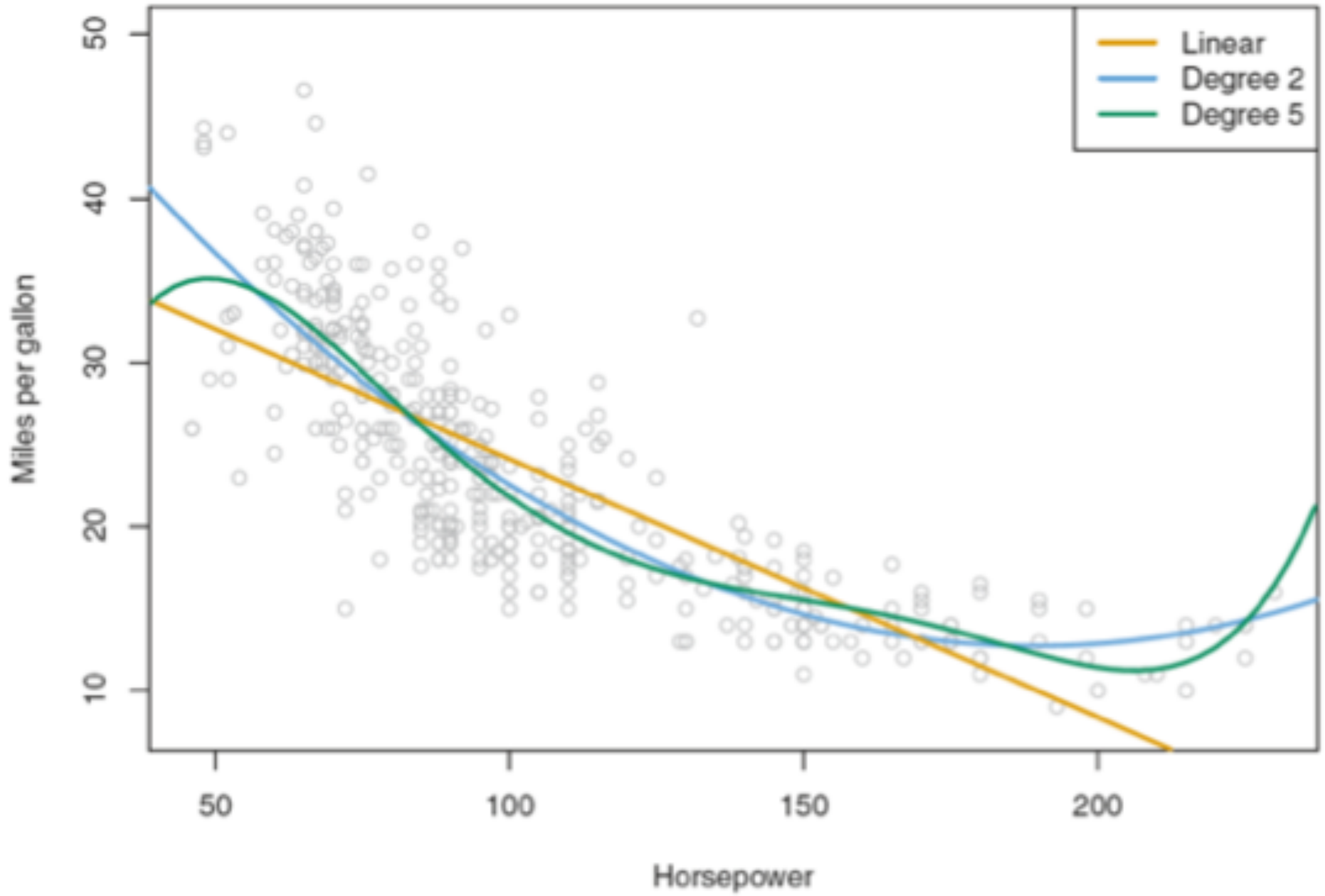


FIGURE 3.8. *The* Auto *data set. For a number of cars,* mpg *and* horsepower *are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes* horsepower^2 *is shown as a blue curve. The linear regression fit for a model that includes all polynomials of* horsepower *up to fifth-degree is shown in green.*

Fit polynomials of horsepower to explain mpg using linear model

Want to see which fits best.

We could look

1. P-values
2. AIC and BIC

But here we do

3. Validation set approach

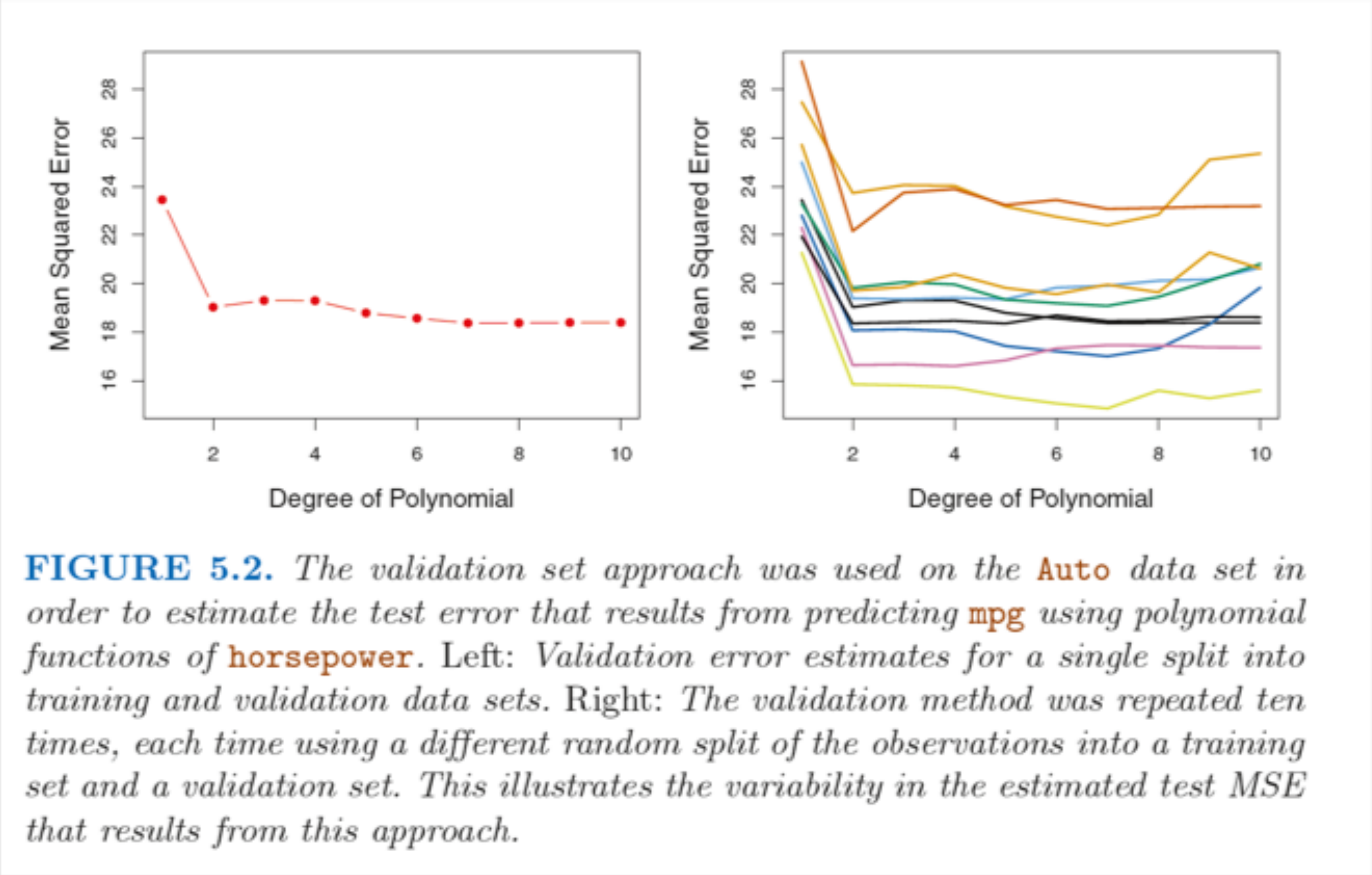


FIGURE 5.2. *The validation set approach was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *Validation error estimates for a single split into training and validation data sets.* Right: *The validation 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.*

1. When the split between training and validation sets is changed, MSE estimate also varies

2. For any one validation set, only a subset of data points are used in training, leading to inefficient use of data

Leave one out cross validation (LOOCV)

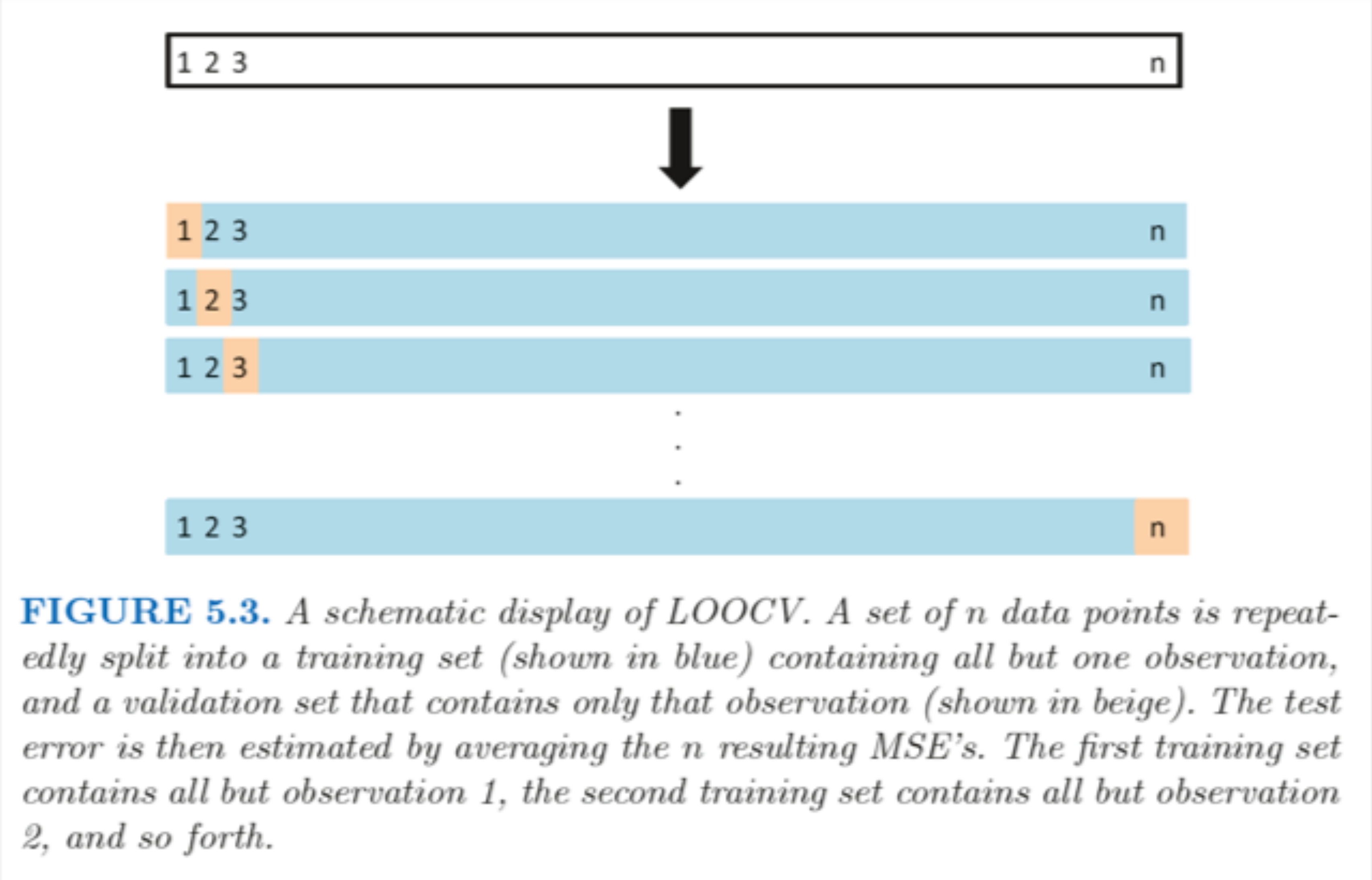


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's.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

Because training data is large in each of n steps of LOOCV, it tends to give more accurate estimate for test MSE than a single split to training and validation (i.e. LOOCV has less bias)

LOOCV may be costly to do since it requires fitting model n times

K-fold cross-validation

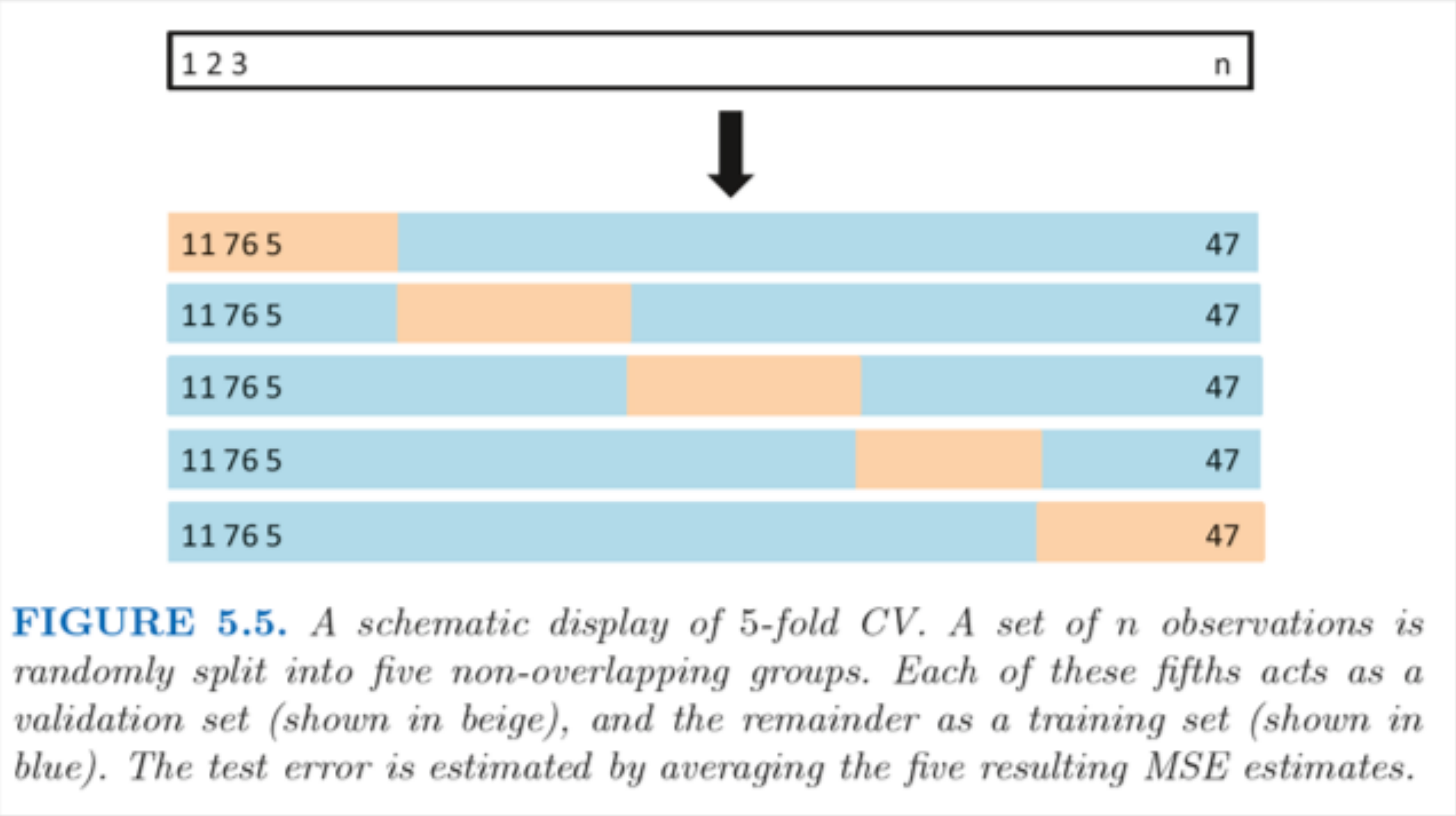


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE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

Less computational costly than LOOCV

Has more bias than LOOCV but less variance (because MSE estimated using different validation folds are less correlated than in LOOCV where difference Is only two data points between validation sets)

Often K=5 or K=10 is used.

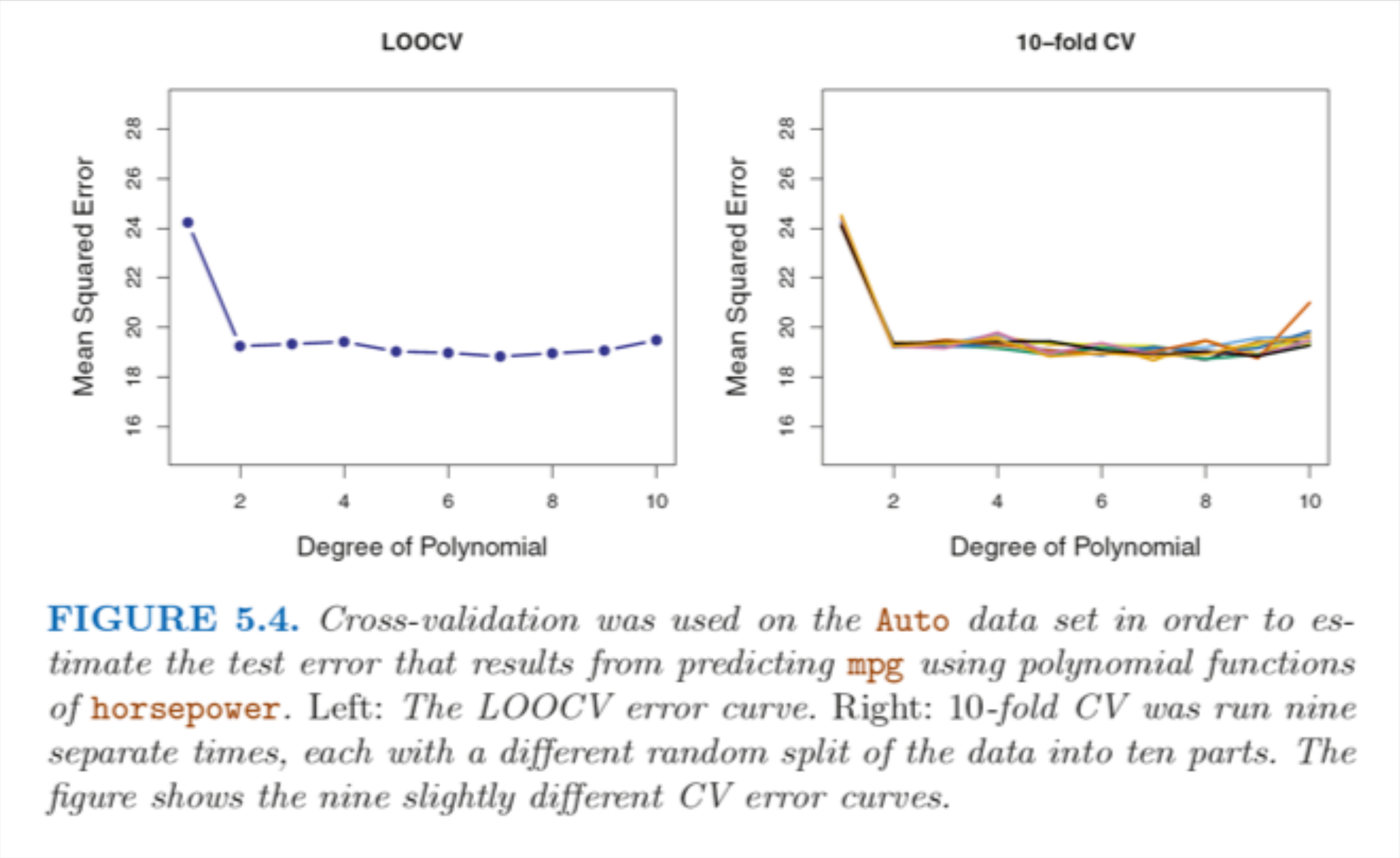


FIGURE 5.4. *Cross-validation was used on the* Auto *data set in order to estimate the test error that results from predicting* mpg *using polynomial functions of* horsepower. Left: *The LOOCV error curve.* Right: *10-fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.*

Some variability remains across sets of 10-fold CV, but much less than across a set of single validation set approaches