Statistical modeling in logistics

MMS075

Lecture 6b – Training error vs test error, Validation set, K-fold cross-validation

Acknowledgement: Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Department of Mechanics and Maritime Sciences
Division of Vehicle Safety

Outline

- Logistic regression (cont.)
  - Confounding
- Multinomial logistic regression
- Training error vs test error
  - Mean squared error
  - Overfitting
- Methods for estimating test error:
  - Validation set approach
  - K-fold cross-validation
- Feed-forward

Recommended resources

Reading in ISL: Section 2.2.1 and 5.1 for theory, 5.3.1-5.3.3 for R codes

The videos from the Statistical Learning course are available at this link. Relevant videos for the new material today:

- **Multivariate Logistic Regression and Confounding** (9:53)
- **Case-Control Sampling and Multiclass Logistic Regression** (7:28)
- **Assessing Model Accuracy and Bias-Variance Trade-off** (10:04)
- **Estimating Prediction Error and Validation Set Approach** (14:01)
- **K-fold Cross-Validation** (13:33)
- **Cross-Validation: The Right and Wrong Ways** (10:07)

Logistic regression (cont.)

Confounding

Coefficient of student

Outputs for student only & student+balance models:

```
Call:
glm(formula = default ~ student, family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2970  -0.2970  -0.2434  -0.2434   2.6585

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
```

```
Call:
glm(formula = default ~ student + balance, family = "binomial"
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4578  -0.1422  -0.0559  -0.0203   3.7435

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

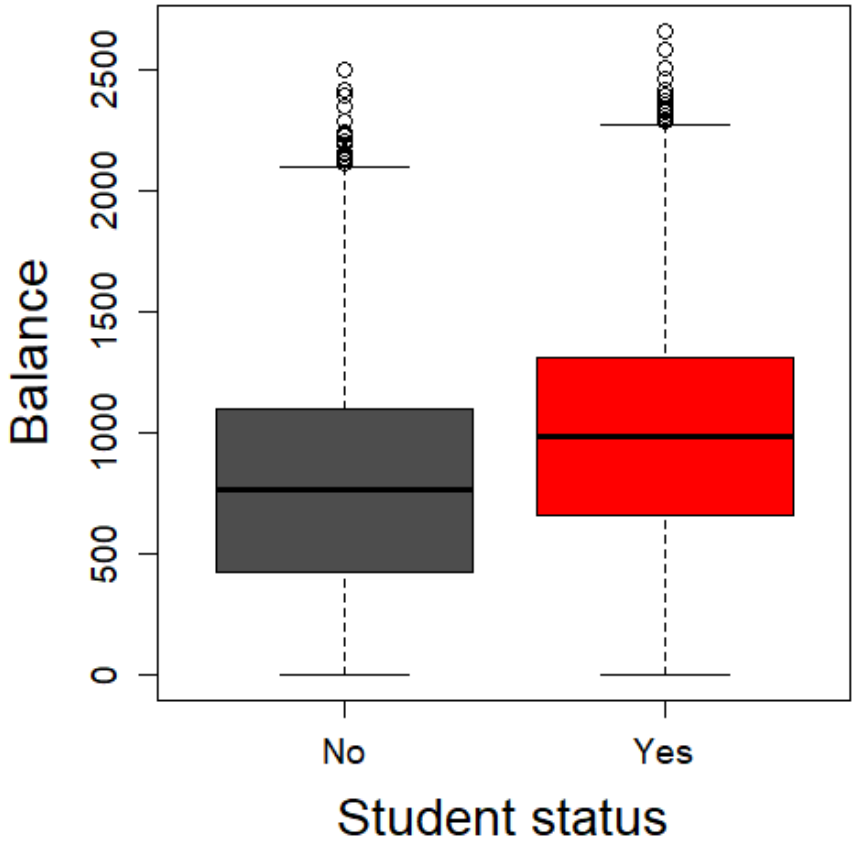
> 0 in single variable model,
< 0 in multivariate model

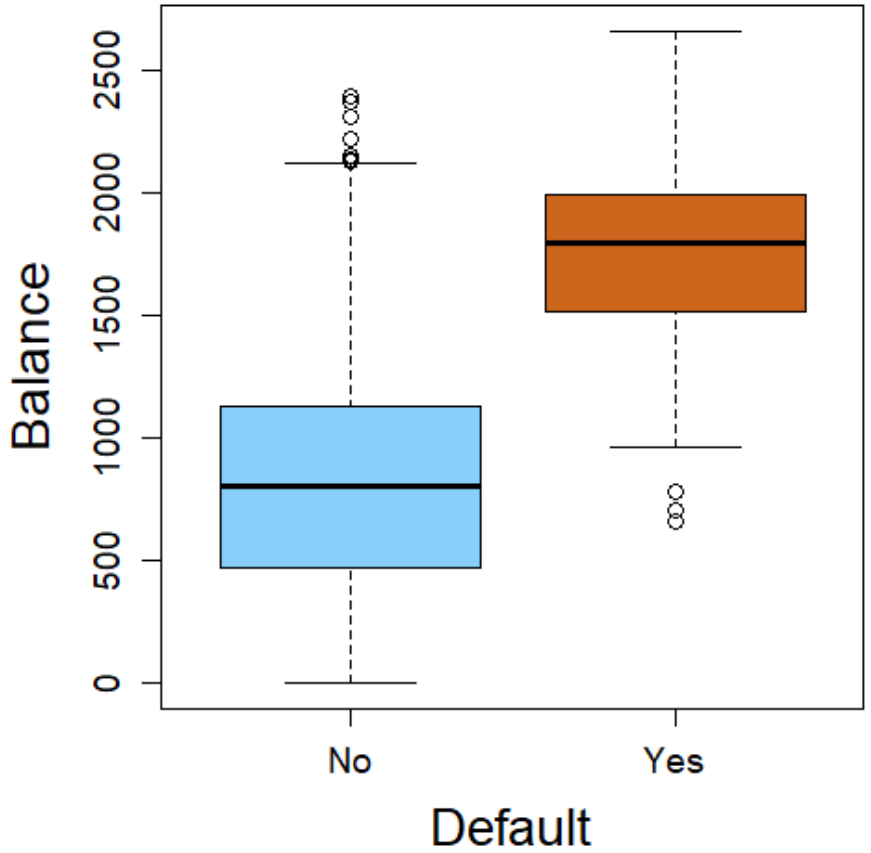
Are students more likely or less likely to default??

- Without further information → students have higher probability
- Among people with given balance → students have lower probability

How is this possible?

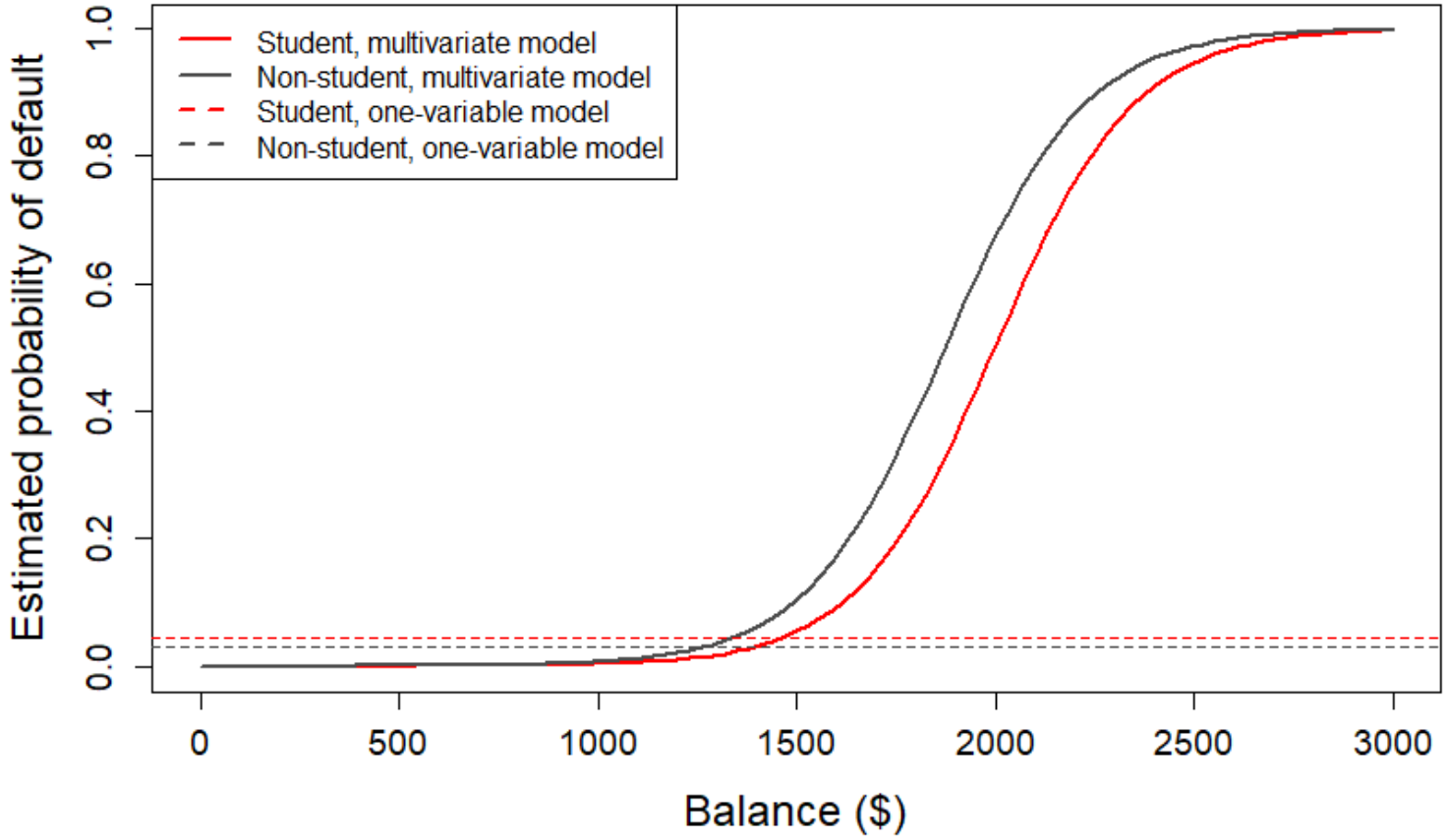
Being a student is associated with higher credit card balance which is associated with higher probability of default





Definition of confounding

Confounding is a phenomenon when the effect of a predictor on the response is turned around by another variable (→ results in multivariate model are very different from one-variable model)



Multinomial logistic regression

More than 2 response classes

We may want to classify a response variable with >2 classes:

- Select most likely diagnosis based on symptoms (epileptic seizure / drug overdose / stroke)
- Predict most likely outcome of car crash at given speed (fatality / serious injury / slight injury / no injury)
- Select preferable transport mode for goods (air transport / road transport / rail transport / water transport)

Multinomial logistic regression equation

- There are different forms of this model

- In the equation below (also used in the R package glmnet), each class gets its own linear model:

$$\Pr(Y = k|X) = \frac{e^{\beta_{0,k}+\beta_{1,k}X_1+\beta_{2,k}X_2+\cdots+\beta_{p,k}X_p}}{\sum_{m=1}^{K} e^{\beta_{0,m}+\beta_{1,m}X_1+\beta_{2,m}X_2+\cdots+\beta_{p,m}X_p}}$$

- We will see an example in the computer lab

Alternative formulation

- The multinomial model can be re-written to compare the probabilities of different response classes:

$$\log\left(\frac{\Pr(Y=k|X)}{\Pr(Y=s|X)}\right) = \beta_{0,k,s}+\beta_{1,k,s}X_1+\beta_{2,k,s}X_2+\cdots+\beta_{p,k,s}X_p$$

- Note: for binary response, this is exactly the binomial logistic regression model!

Training error vs test error

Mean squared error

Overfitting

What is a good model?

- We have n observations in form of predictor-response pairs:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- Based (**trained**) on these observations, we define a model $\hat{f}$ and hope that the model approximates the true connection between predictor and response, i.e. $\hat{f}(X) \approx Y$

- **Training mean squared error (training MSE)** measures how well this holds for training points, i.e. measures quality of fit:

$$MSE = \frac{(y_1 - \hat{f}(x_1))^2 + (y_2 - \hat{f}(x_2))^2 + \ldots (y_n - \hat{f}(x_n))^2}{n}$$

What do we really want?

- Why do we want a model with good quality of fit, i.e. small MSE?

- Why do we need a model at all? We KNOW the response values for the training set → why should we estimate them?

- Because we hope that the model gives useful information for **new data (test data)**: if $(x_1^{\text{new}}, y_1^{\text{new}}), (x_2^{\text{new}}, y_2^{\text{new}}), \ldots$ are new, previously unseen observations that were not used to train (i.e. define) the model, we want a model with small average prediction error

→ We want to minimize $\text{Average}\left[(y^{\text{new}} - \hat{f}(x^{\text{new}}))^2\right]$

Training error vs test error

Mean squared error

Overfitting

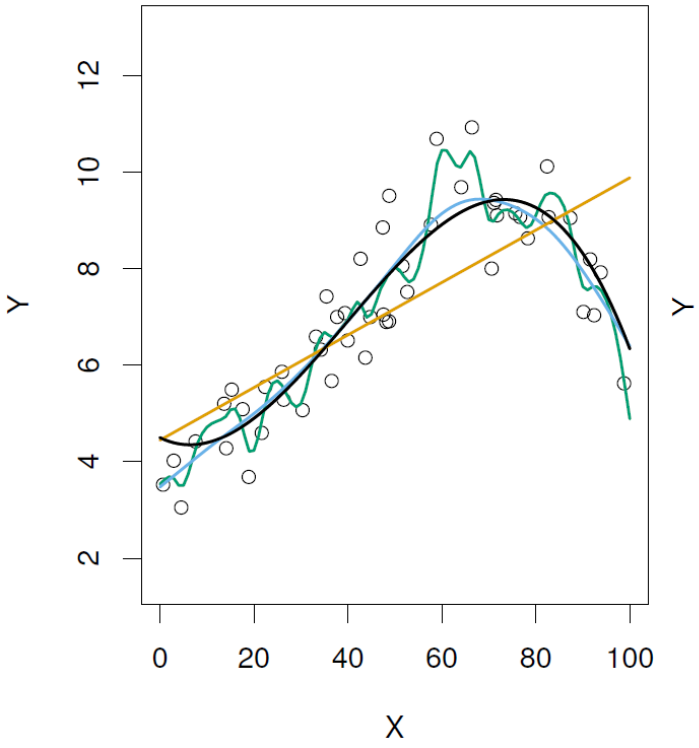
Is it best to minimize training MSE?

- Model with a good fit on training data is a natural aim
- **Overfitting**: if model follows training points too closely & reproduces noise effects (i.e. pick up patterns caused by chance rather than a meaningful relationship) that may not be present in new data
- This is caused by overly flexible models. Examples on next slides
- However, among models of given complexity (e.g. linear models), the one that fits training points best should be chosen

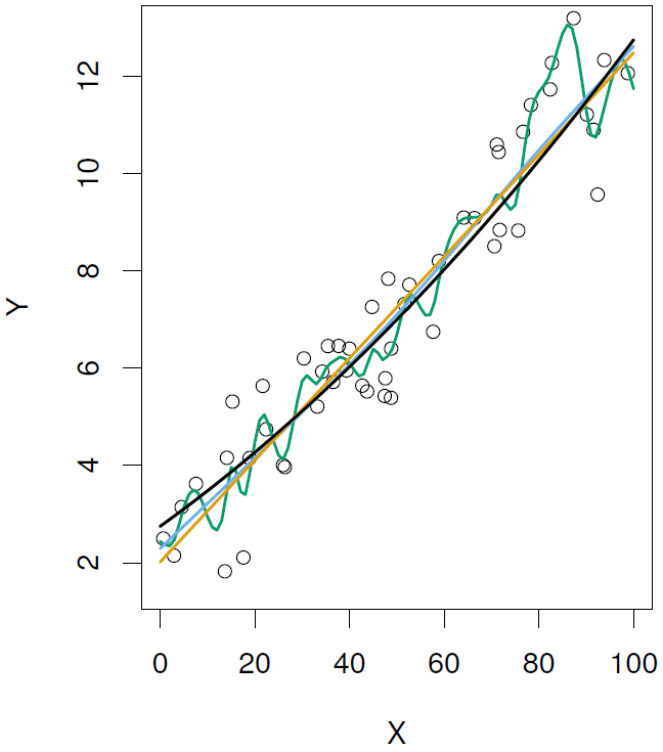
How flexible model do we need in general?

Structure (shape) of points looks close to...

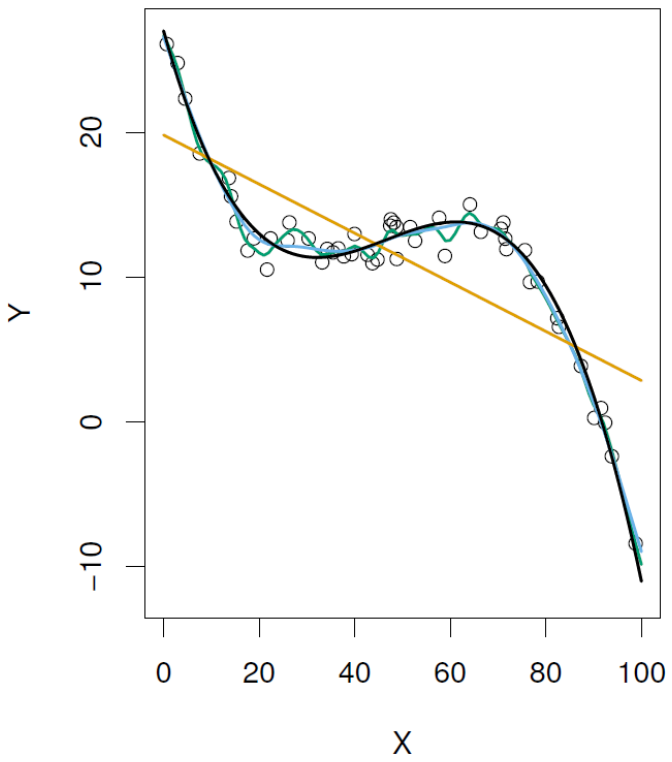
... a quadratic (or low-degree) curve



... a line



... a cubic (or high-degree) curve

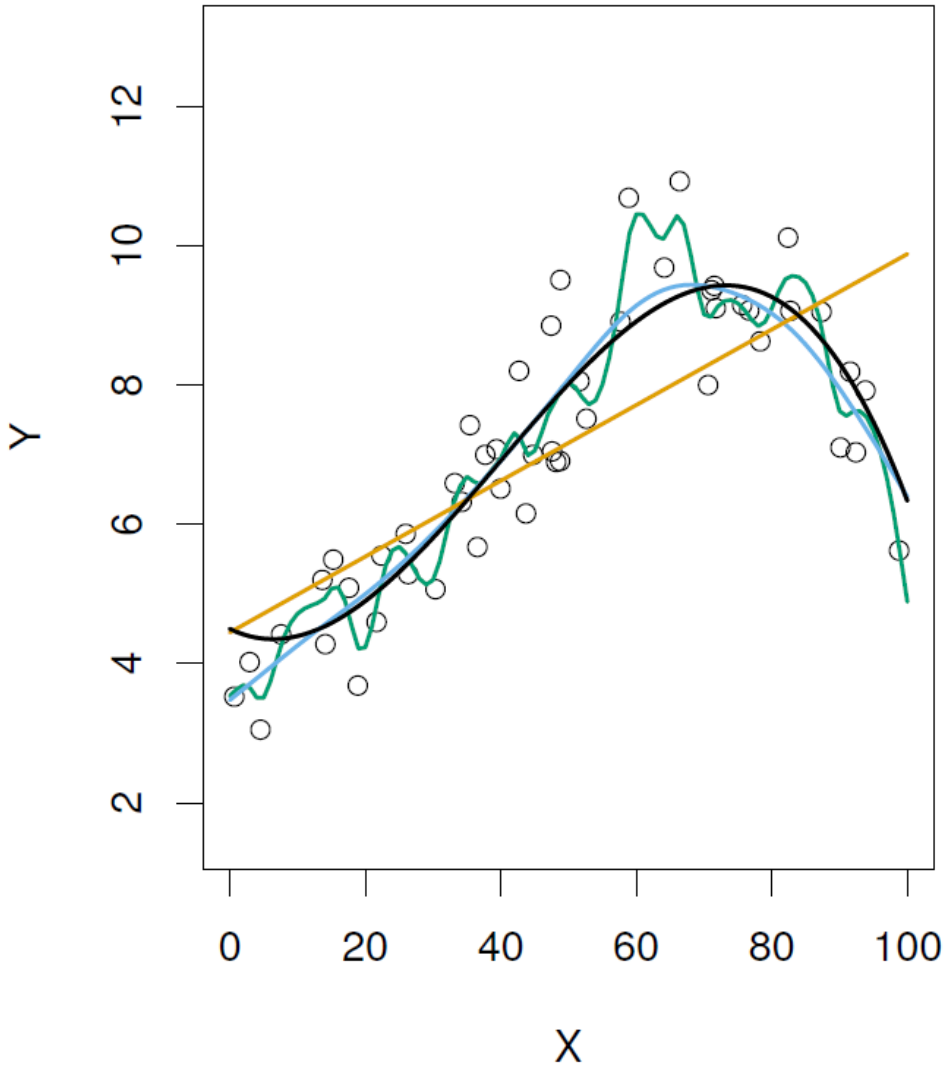


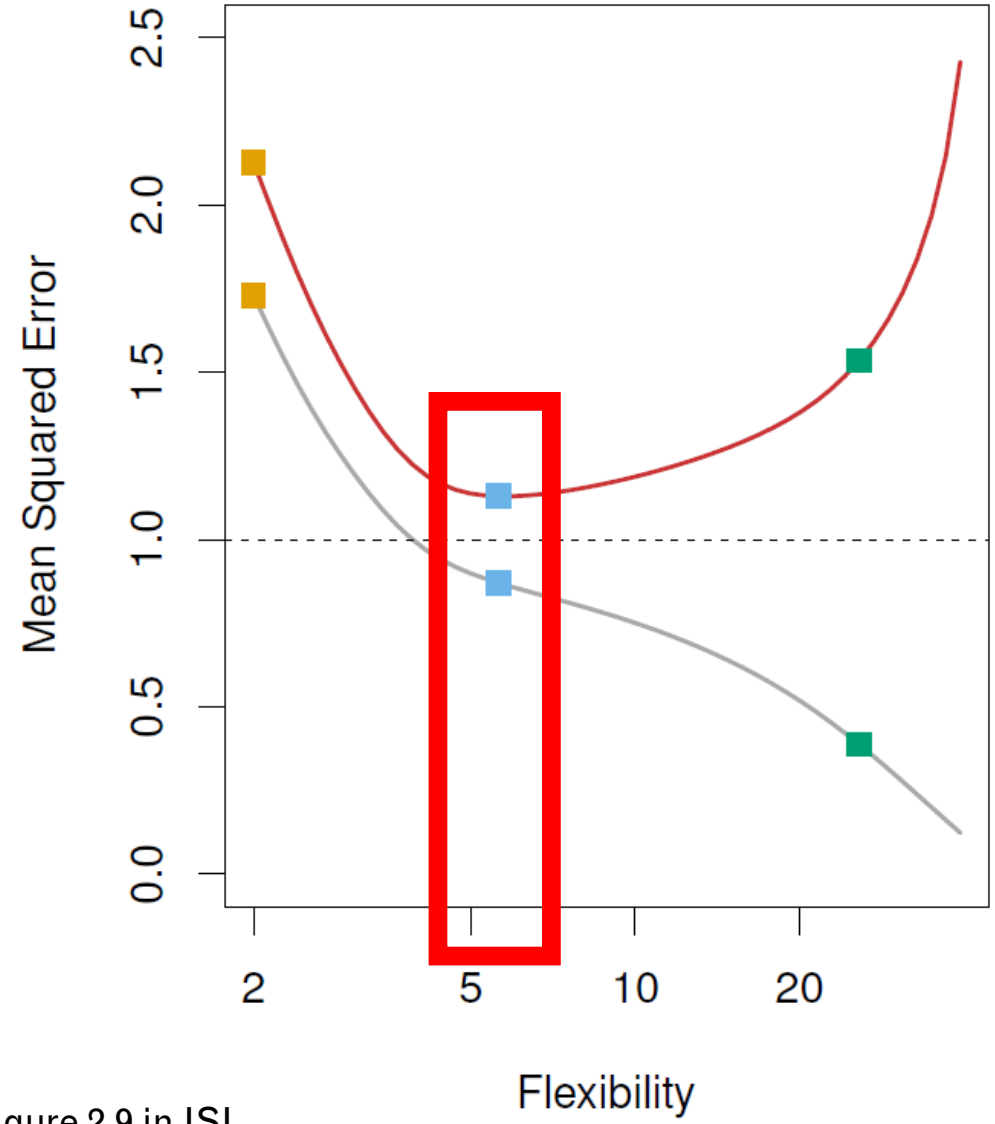
Source: Figures 2.9, 2.10 and 2.11 in ISL

What do we expect?

- Models of the corresponding flexibilility should work best:
  - Less complex models cannot reproduce observed shape
  - More complex models are too wiggly and can/will follow noise in data
- This was simulated data → can create a large test set with the true distribution & check test error values for various models
- Next slides show test MSE (red curves) and training MSE (grey curves) for models of different flexibility.
- Training MSE always decreases with larger flexibility; test MSE is U-shaped, as anticipated, indicating a trade-off

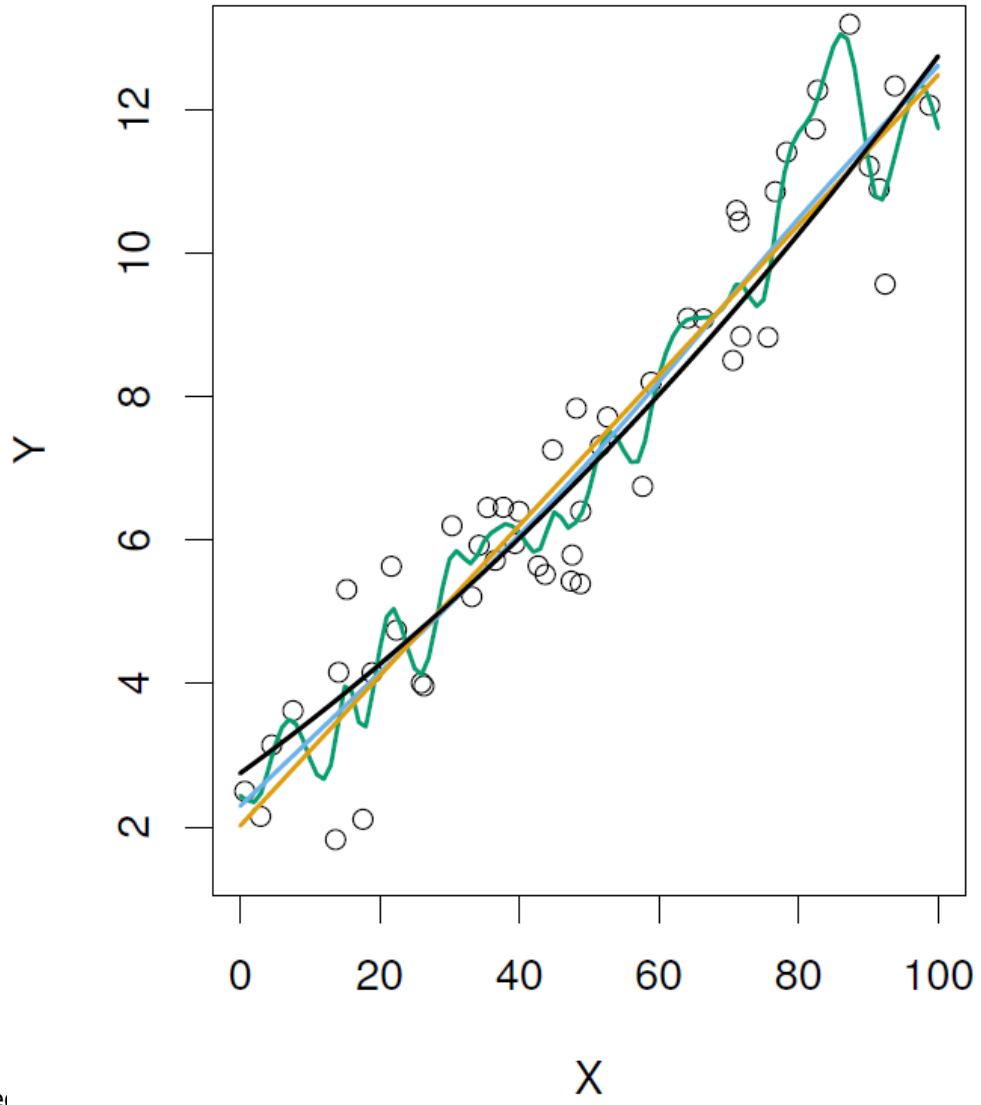
Test error curves by model complexity (1)

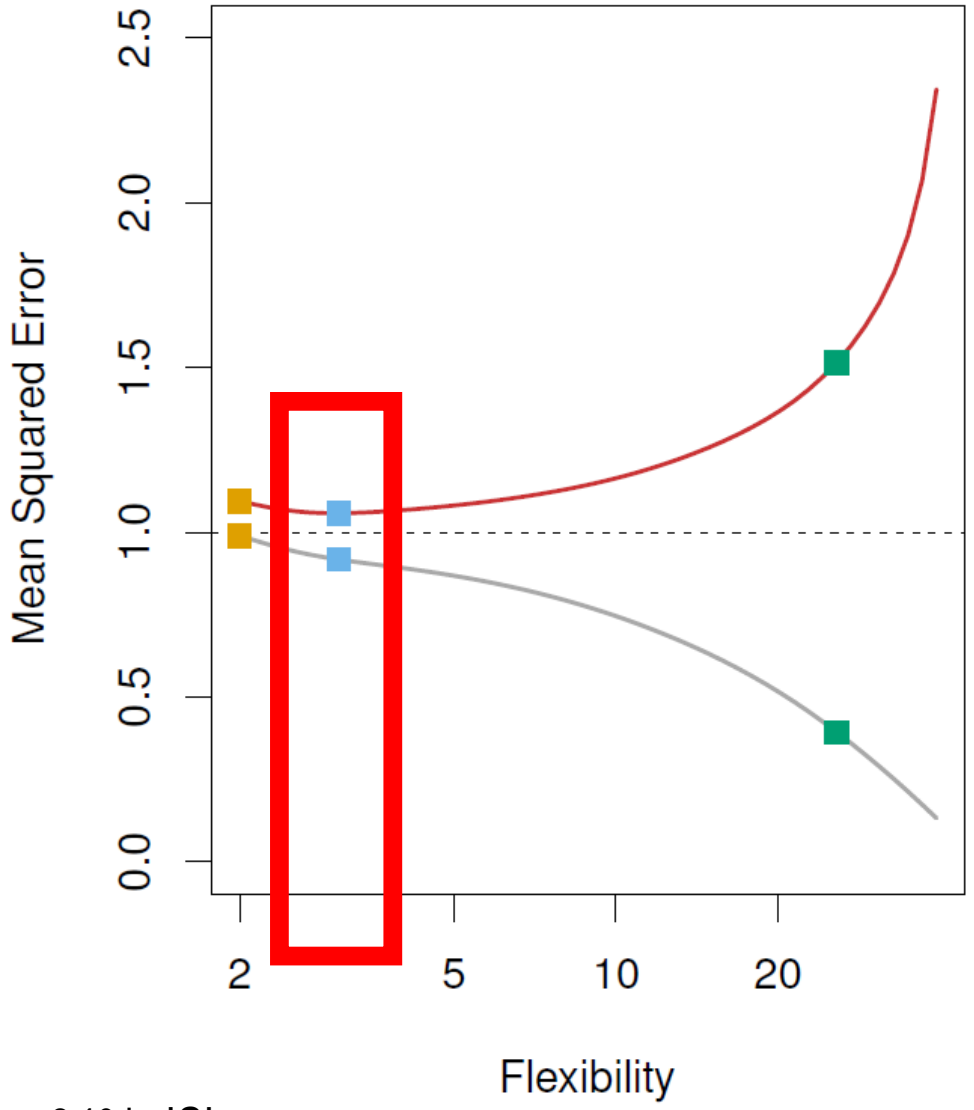




Source: Figure 2.9 in ISL

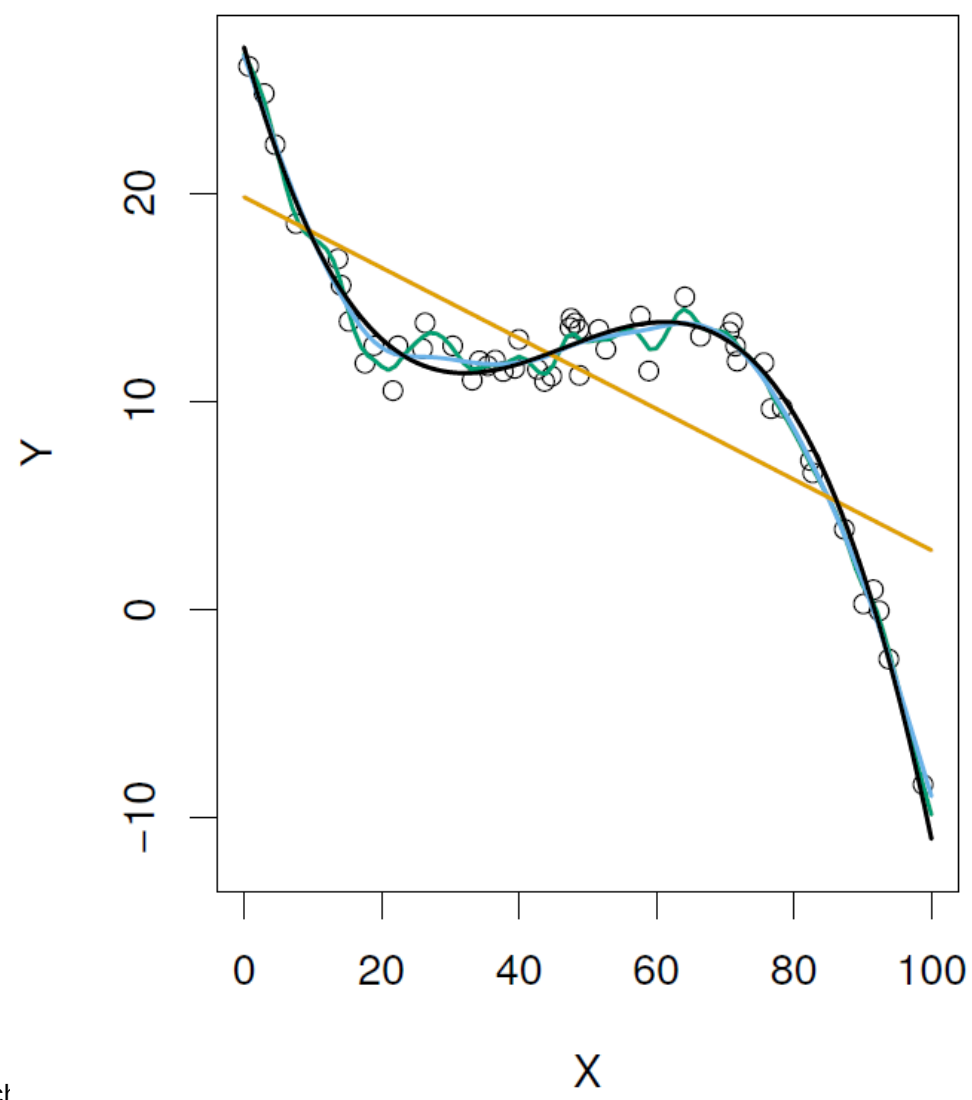
Test error curves by model complexity (2)

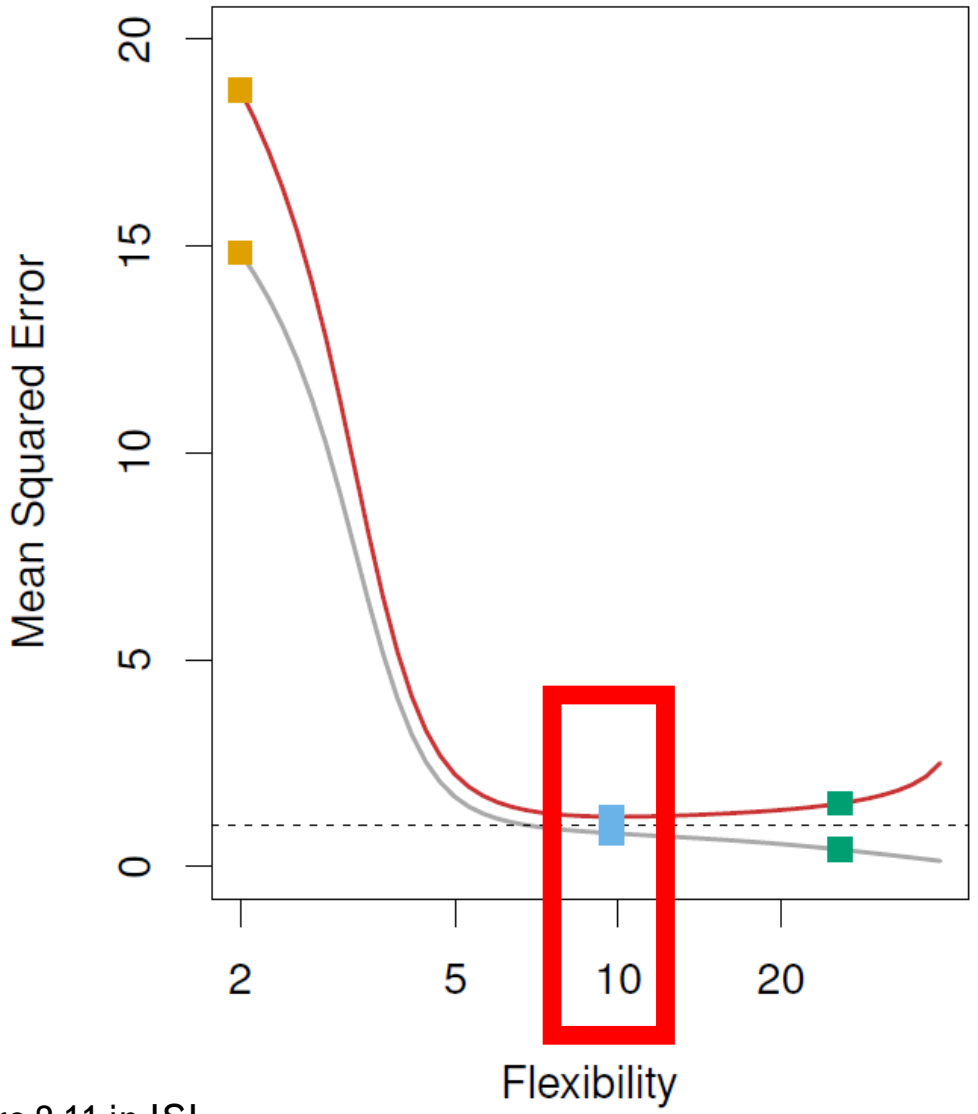




Source: Figure 2.10 in ISL

Test error curves by model complexity (3)





Source: Figure 2.11 in ISL

Simulated data is uncommon

- If data is simulated → as much test data can be produced as needed → we can have a full understanding of test error

- This is not a usual case! We typically have only one set of observations – how can we then estimate test error?

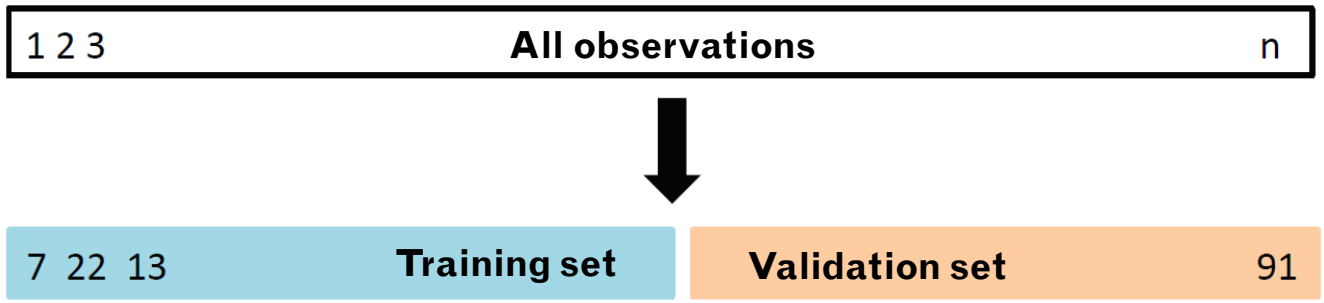
Methods estimating test error

Validation set approach

K-fold cross-validation

Validation set approch: basic idea

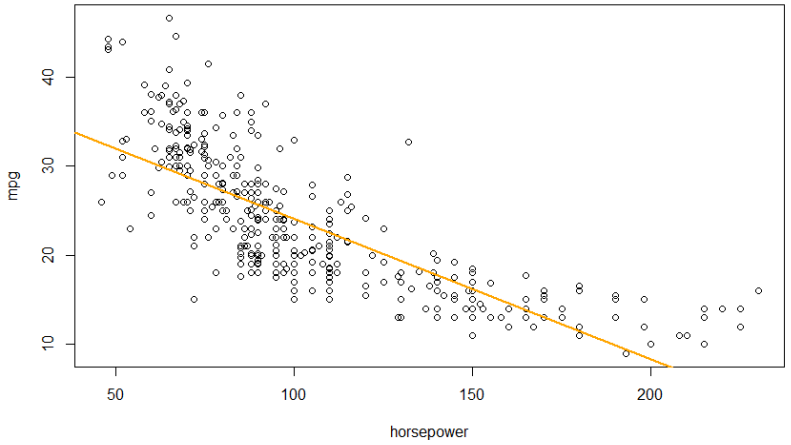
- Reserve some part of observations that will not be used in model building
- This set of points (called **hold out set** or **validation set**) is unseen by the model while the model is defined → it can play the role of test data to see how well the model can predict unseen points
- The set that was used for model definition (e.g. get coefficient estimates) is the **training set**



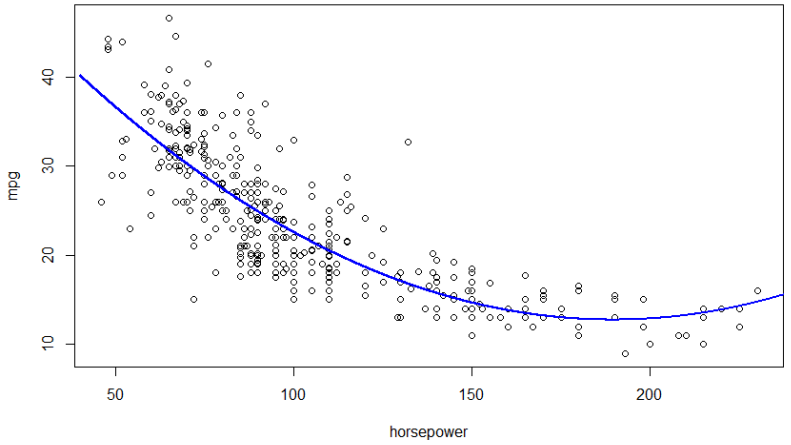
Validation set approach, source: Figure 5.1 in ISL, labels added

Recall mpg vs horsepower models

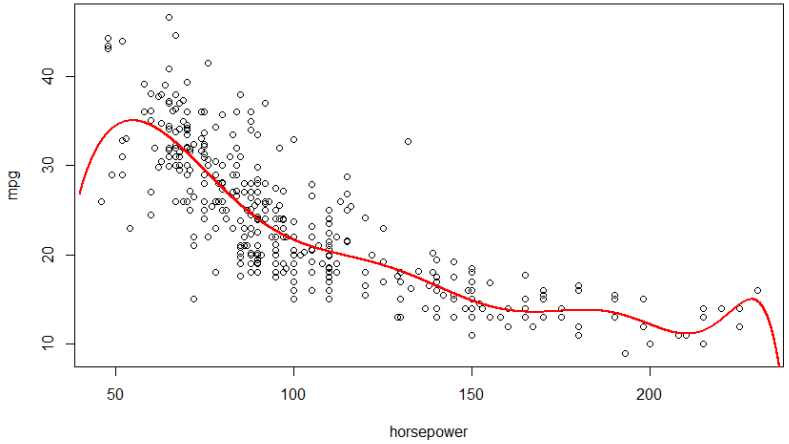
Linear model



Quadratic model (degree 2 polynomial)



High-degree model (degree 10 polynomial)



Plots and p-values suggest:

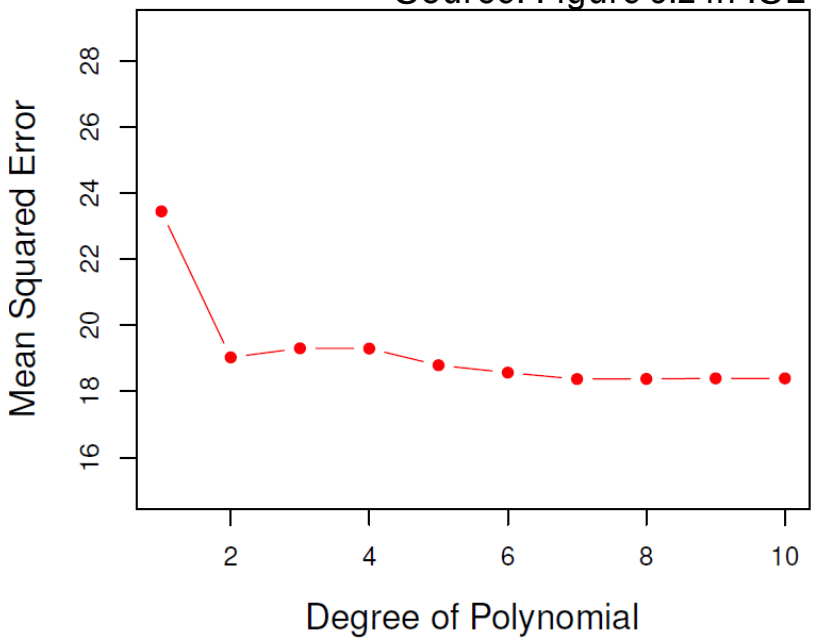
- Quadratic model is better than linear
- High-degree model may be overfitting

What can we conclude with the validation set approach?

Validation set error estimates

1. Divide the observations into a training set and a validation set
2. Using the points in the training set, fit a linear, quadratic and higher degree models
3. Using the points in the validation set, compute MSE for all these models and plot the error estimates:

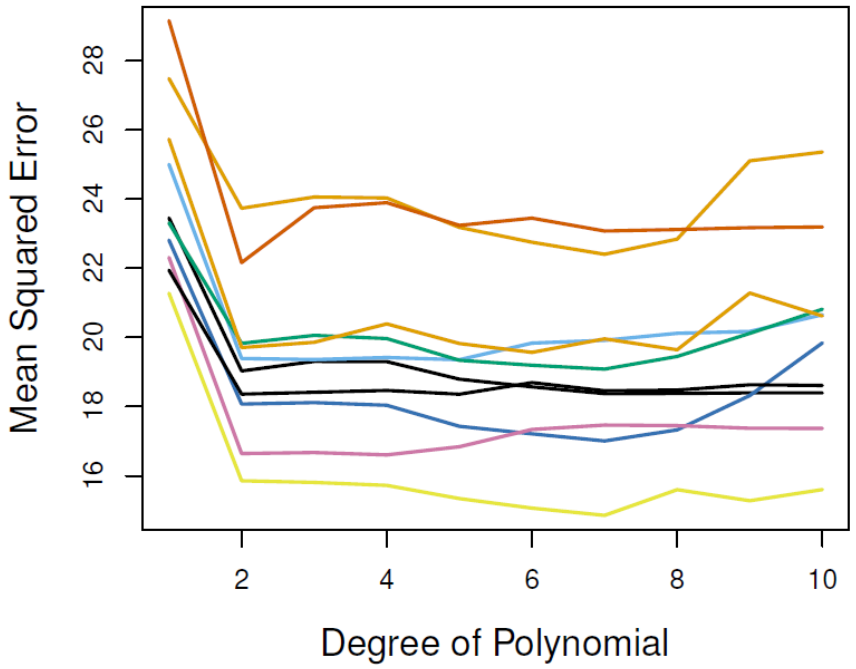
Source: Figure 5.2 in ISL



Issue with validation set approach

- Results depend on how the division of observations into training set and validation set was made

→ Validation set estimate of test error is highly variable



Validation set test error estimates with 10 different splits, source: Figure 5.2 in ISL

Note: while the value of the MSE varies wildly, all divisions show some similar patterns:

- Degree 2 polynomial is better than degree 1 (i.e. quadratic model is better than linear)
- No large difference between error estimates for different degrees when using ≥2 degrees

→ These results support using quadratic model

Methods estimating test error

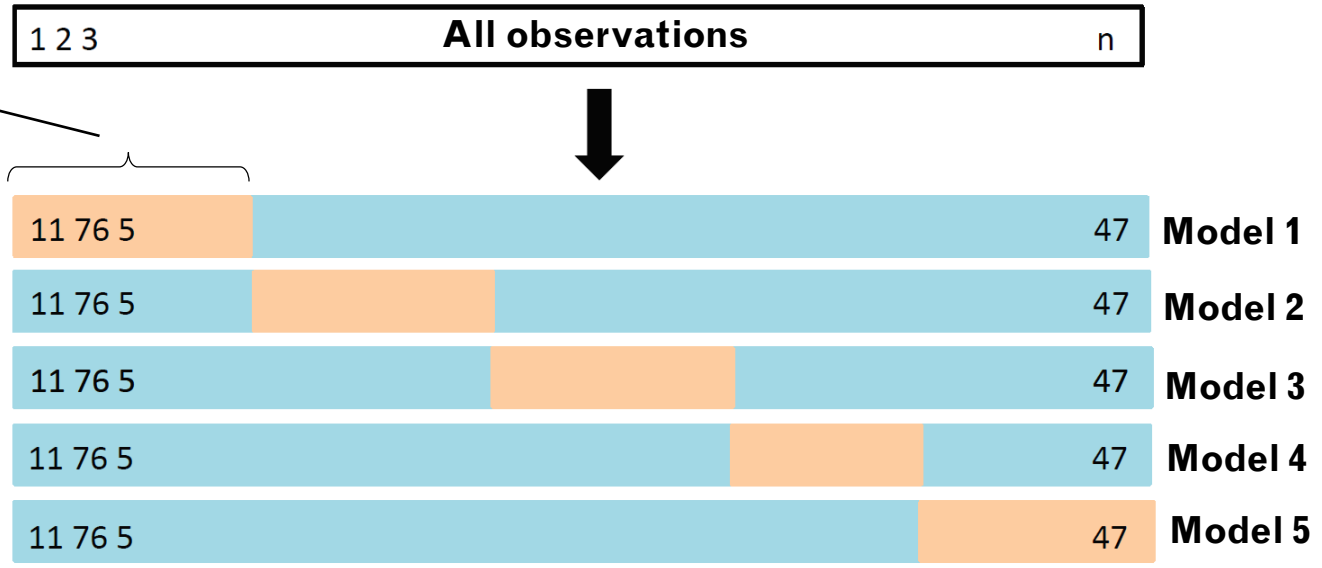
Validation set approach

K-fold cross-validation

K-fold cross-validation: basic idea

- Divide the n observations into K equal parts (as equal as possible)
- Consider K different models, considering each part once as validation set and the other K-1 parts combined as training set

These observations are in:
- Validation set in Model 1
- Training set in Model 2
- Training set in Model 3
- Training set in Model 4
- Training set in Model 5



5-fold CV, source: Figure 5.5 in ISL, labels added

K-fold CV error estimates for mpg example

1. Divide the observations into K equal sets

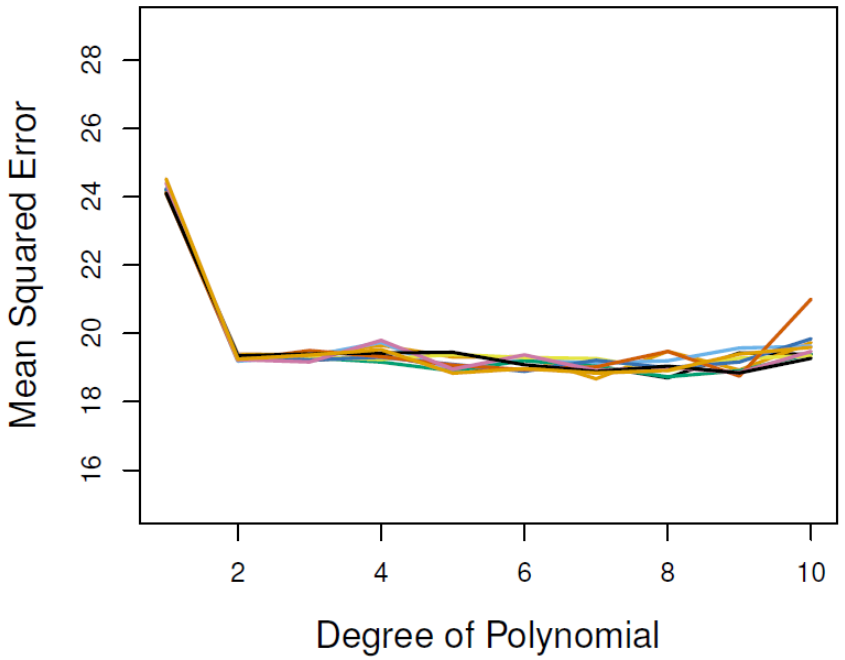
2. Changing the role of training set as shown on previous slide, fit a linear, quadratic and higher degree models K times

3. Compute MSE for each of the K linear models, K quadratic models, K higher degree models on the corresponding validation set (which is different for each of the K models)

4. Plot the average of the K error estimates, for each type of model

Less variablility in test error estimate

- Results depend somewhat on how the K folds were defined
- The variability in the estimate of test error is much smaller than it was with the validation set approach



10-fold CV test error estimates with
10 different splits, source: Figure 5.4 in ISL

In this case, the same patterns are even clearer:

- Degree 2 polynomial is better than degree 1 (i.e. quadratic model is better than linear)
- No large difference between error estimates for different degrees when using ≥2 degrees

→ 10-fold CV supports quadratic model

Feed-forward

Feed-forward quiz

We will review the course material in w7. Which parts should we emphasize more?

1. Go to www.menti.com
2. Enter the code 38 50 70
3. Answer the questions or enter other comments related to the course or today's lecture