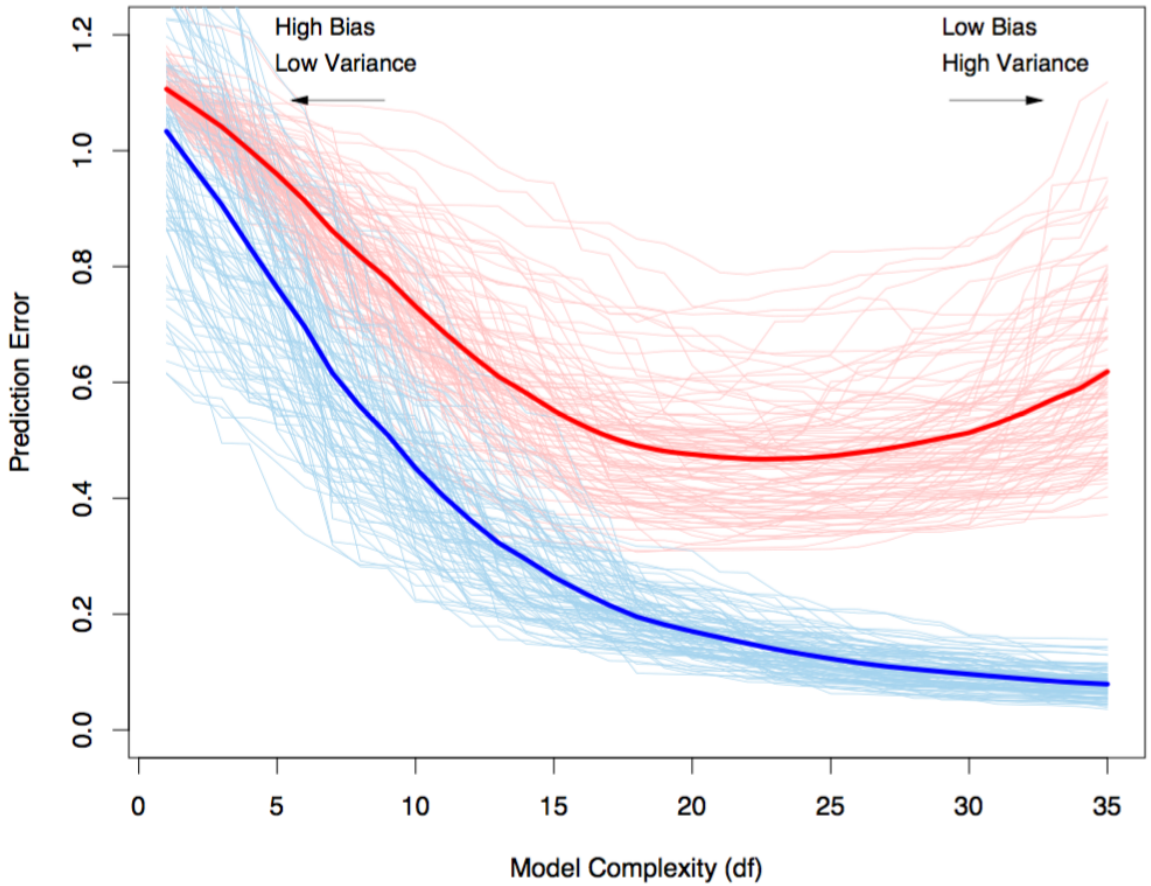
Day 5: Model Selection I

Lucas Leemann

Essex Summer School

Introduction to Statistical Learning

Fundamental Problem: Model Complexity



Red: Test error.
Blue: Training error.

(Hastie et al, 2008: 220)

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.
- We wish to choose a model with low test error, not a model with low training error. Recall that training error is usually a poor estimate of test error.
- Therefore, RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

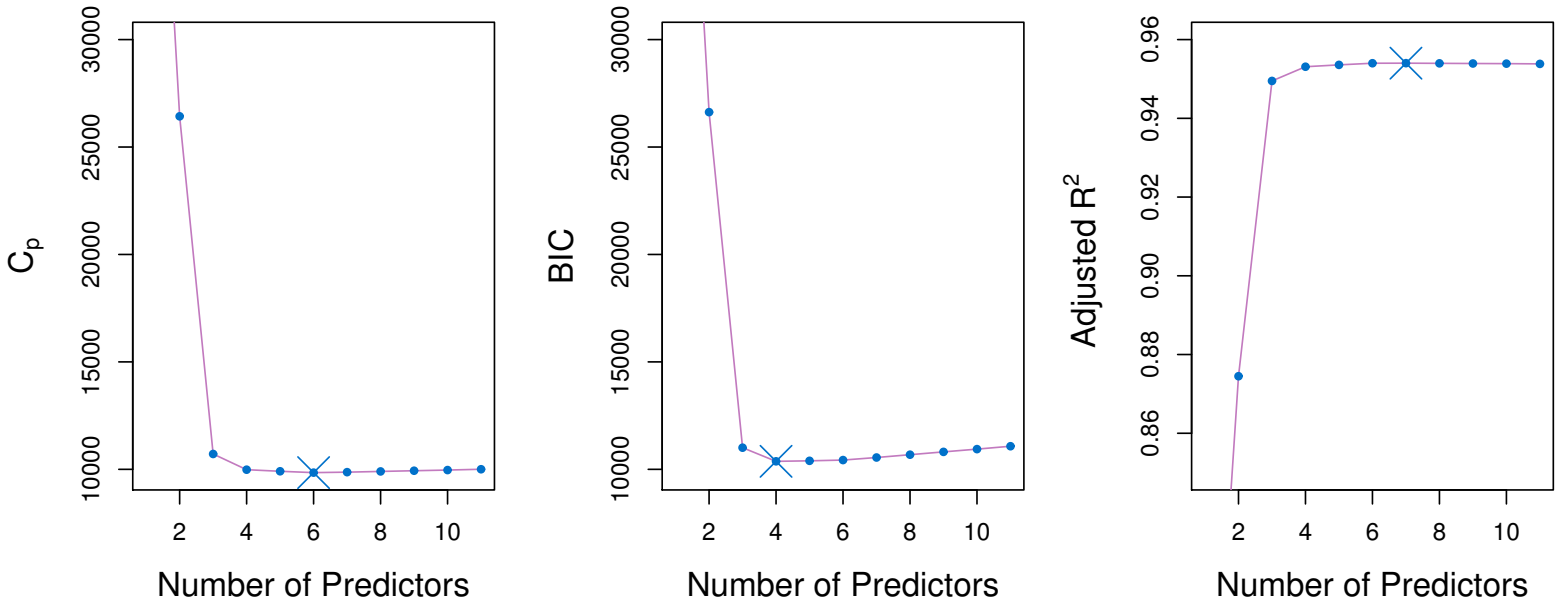
Estimating test error: two approaches

- We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.
- We can directly estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in previous lectures.

C_p, AIC, BIC, and Adjusted R^2

- These techniques adjust the training error for the model size, and can be used to select among a set of models with different numbers of variables.
- The next figure displays C_p, BIC, and adjusted R^2 for the best model of each size produced by best subset selection on the Credit data set.

Example: Credit data



Mallow's C_p & AIC

- Mallow's C_p:

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2),$$

where d is the total number of parameters used and $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement.

- The AIC criterion is defined for a large class of models fit by maximum likelihood:

$$AIC = 2logL + 2\dot{d}$$

where L is the maximized value of the likelihood function for the estimated model.

- In the case of the linear model with Gaussian errors, maximum likelihood and least squares are the same thing, and C_p and AIC are equivalent.

Details on BIC

$$BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2)$$

- Like C_p, the BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value.
- Notice that BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $log(n)d\hat{\sigma}^2$ term, where n is the number of observations.
- Since $log(n) > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p.

Adjusted R^2

- For a least squares model with d variables, the adjusted R^2 statistic is calculated as

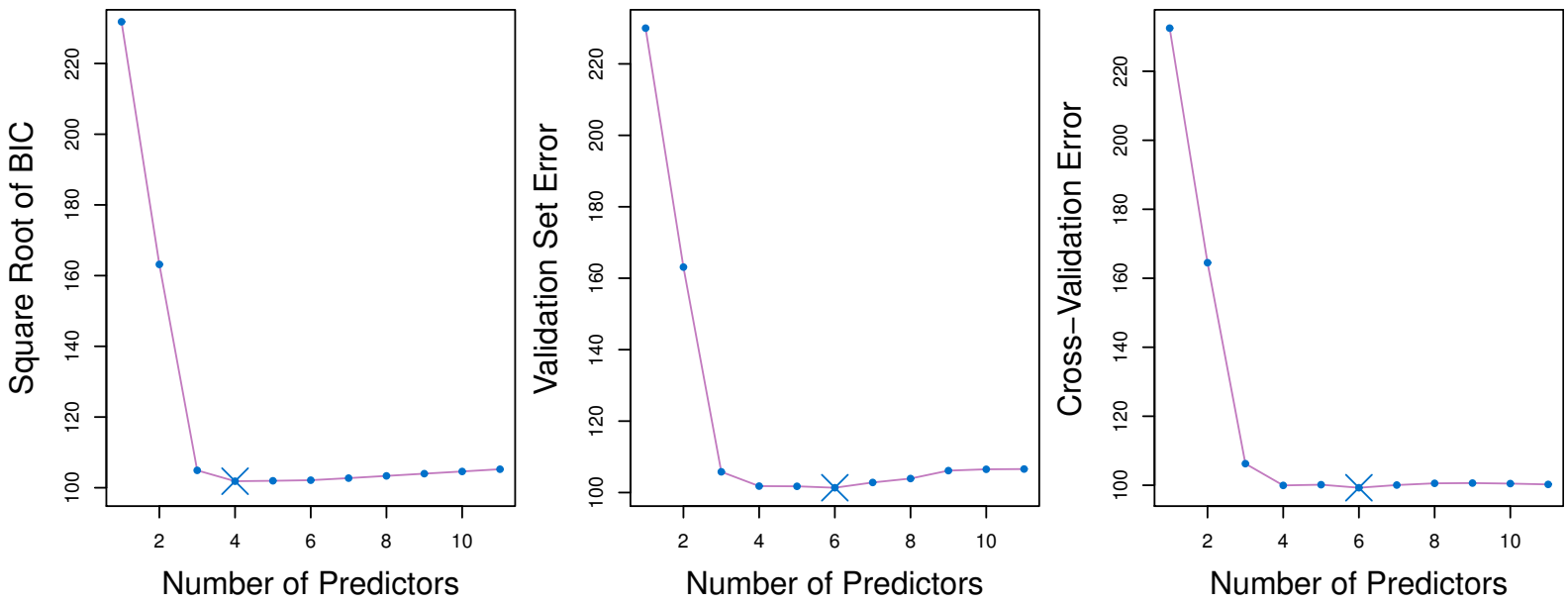
$$\textit{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

  where TSS is the total sum of squares.
- Unlike C_p, AIC, and BIC, for which a small value indicates a model with a low test error, a large value of adjusted R^2 indicates a model with a small test error.
- Maximizing the adjusted R^2 is equivalent to minimizing $\frac{RSS}{n-d-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{RSS}{n-d-1}$ may increase or decrease, due to the presence of d in the denominator.
- Unlike the R^2 statistic, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model.

Validation and Cross-Validation

- Each of the procedures returns a sequence of models $\mathcal{M}_k$ indexed by model size $k = 0, 1, 2, \ldots$. Our job here is to select $\hat{k}$. Once selected, we will return model $\mathcal{M}_{\hat{k}}$
- We compute the validation set error or the cross-validation error for each model $\mathcal{M}_k$ under consideration, and then select the k for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to AIC, BIC, C_p, and adjusted R2, in that it provides a direct estimate of the test error.
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

Example: Credit data



Explaining the example above

- The validation errors were calculated by randomly selecting three-quarters of the observations as the training set, and the remainder as the validation set.
- The cross-validation errors were computed using $k = 10$ folds. In this case, the validation and cross-validation methods both result in a six-variable model.
- However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.
- In this setting, we can select a model using the one-standard-error rule. We first calculate the standard error of the estimated test MSE for each model size, and then select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve.

Subset Selection

Subset Selection: Which Variables?

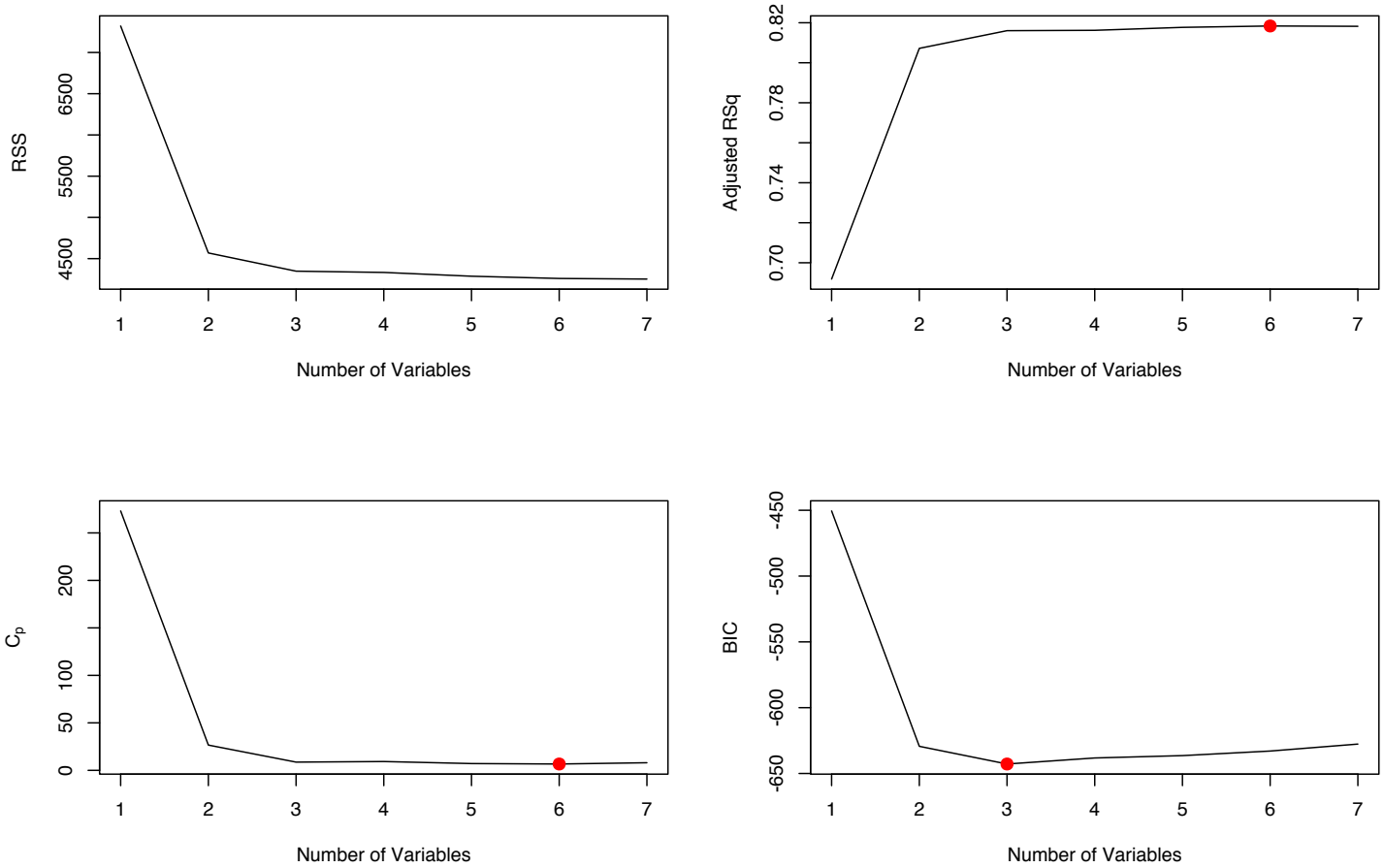
Algorithm:

1. Generate an empty model and call it $\mathcal{M}_0$
2. For each $k = 1....p$:
   i) Generate all $\binom{p}{k}$ possible models with k explanatory variables
   ii) determine the model with the best criteria value (e.g. R^2) and call it $\mathcal{M}_k$
3. Determine best model within the set of these models: $\mathcal{M}_0$, ..., $\mathcal{M}_p$
   - rely on CV or a criteria like AIC, BIC, R^2, or C_p

Example 1 (1)

```
> regfit.full <- regsubsets(mpg ~ ., Auto[,-9])
> summary(regfit.full)
Subset selection object
Call: regsubsets.formula(mpg ~ ., Auto[, -9])
7 Variables  (and intercept)
             Forced in Forced out
cylinders        FALSE      FALSE
displacement     FALSE      FALSE
horsepower       FALSE      FALSE
weight           FALSE      FALSE
acceleration     FALSE      FALSE
year             FALSE      FALSE
origin           FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         cylinders displacement horsepower weight acceleration year origin
1  ( 1 ) " "       " "          " "        "*"    " "          " "  " "
2  ( 1 ) " "       " "          " "        "*"    " "          "*"  " "
3  ( 1 ) " "       " "          " "        "*"    " "          "*"  "*"
4  ( 1 ) " "       "*"          " "        "*"    " "          "*"  "*"
5  ( 1 ) " "       "*"          "*"        "*"    " "          "*"  "*"
6  ( 1 ) "*"       "*"          "*"        "*"    " "          "*"  "*"
7  ( 1 ) "*"       "*"          "*"        "*"    "*"          "*"  "*"
```

Example 1 (2)


RSS
4500
5500
6500
1
2
3
4
5
6
7
Number of Variables
Adjusted RSq
0.70
0.74
0.78
0.82
Number of Variables
C_p
0
50
100
200
Number of Variables
BIC
-650
-600
-550
-500
-450
Number of Variables

Subset Selection

- Subset selection can be very challenging when p is large since we are then looking at $\binom{p}{k}$ possibilities in the k^{th} step. For $p = 10$ we have about 1000 models and for $p = 20$ we are already facing more than 1 million models.
- What if $p >> n$?
- Different approaches: stepwise selection

Stepwise Selection

Forward Stepwise Selection (1)

Algorithm:

1. Generate an empty model and call it $\mathcal{M}_0$
2. For $k = 0....p-1$:
   i) Consider all $p-k$ possible models that have one predictor more than $\mathcal{M}_k$
   ii) determine the *best* model among all models in (i) and call it $\mathcal{M}_{k+1}$

   (Here: best refers to highest R^2 or smallest MSE since k constant within each step)
3. Determine best model within the set of these models: $\mathcal{M}_0$, ..., $\mathcal{M}_p$ - rely on CV or on a criteria like AIC, BIC, R^2, or C_p

Forward Stepwise Selection (2)

- Best subset selection involves looking at 2^p models, whereas *forward stepwise selection* only uses $1 + p(p+1)/2$ models.
- Can be used when $n < p$ (at least for $\mathcal{M}_0$ up to $\mathcal{M}_{n-1}$).
- Forward stepwise selection usually does well but it is not guaranteed to find best model:

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

(James et al. 2013: 209)

Backwards Stepwise Selection (1)

Algorithm:

1. Let $\mathcal{M}_p$ denote the full model with p predictors
2. For $k = p,\ p-1,\ p-2,\ .....,\ 1$:
   i) Consider all k possible models that have $k-1$ predictors (one less than $\mathcal{M}_k$)
   ii) determine the *best* model among the k models in (i) and call it $\mathcal{M}_{k-1}$

   (Here: best refers to highest R^2 or smallest MSE since k constant within each step)
3. Determine best model within the set of these models: $\mathcal{M}_0,\ ....,\ \mathcal{M}_p$ - rely on CV or on a criteria like AIC, BIC, R^2, or C_p

Backwards Stepwise Selection (2)

- As forward stepwise selection *backward stepwise selection* only needs to estimate $1 + p(p + 1)/2$ models.
- BSS cannot be used when $p > n$.

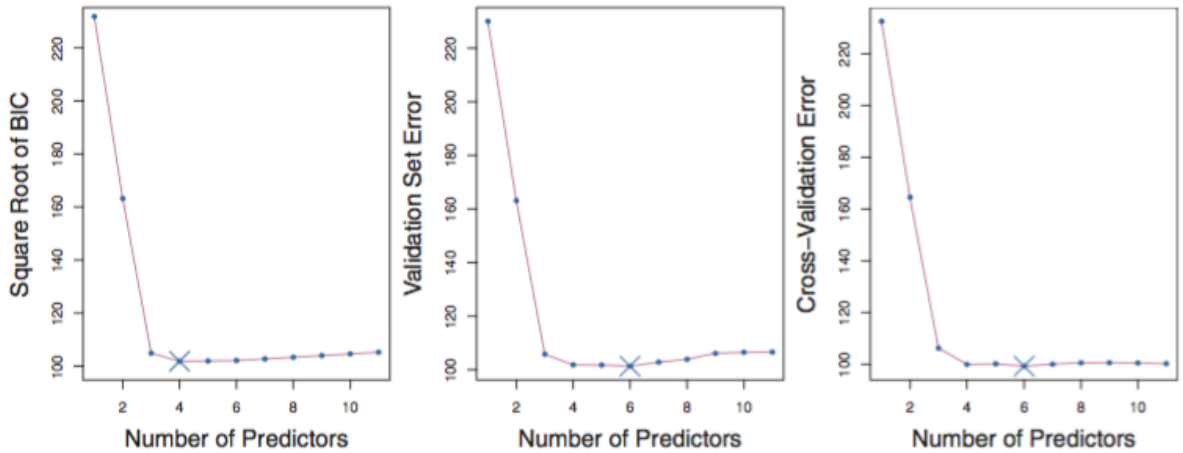
Example 2

```
> regfit.full <- regsubsets(Salary ~ .,data=Hitters, nvmax=19)
> regfit.for <- regsubsets(Salary ~ .,data=Hitters, nvmax=19, method="forward")
> regfit.back <- regsubsets(Salary ~ .,data=Hitters, nvmax=19, method = "backward")
>
> coef(regfit.full, 7)
 (Intercept)         Hits        Walks       CAtBat        CHits       CHmRun    DivisionW      PutOuts
  79.4509472    1.2833513    3.2274264   -0.3752350    1.4957073    1.4420538 -129.9866432    0.2366813
>
> coef(regfit.for, 7)
 (Intercept)        AtBat         Hits        Walks         CRBI       CWalks    DivisionW      PutOuts
 109.7873062   -1.9588851    7.4498772    4.9131401    0.8537622   -0.3053070 -127.1223928    0.2533404
>
> coef(regfit.back, 7)
 (Intercept)        AtBat         Hits        Walks        CRuns       CWalks    DivisionW      PutOuts
 105.6487488   -1.9762838    6.7574914    6.0558691    1.1293095   -0.7163346 -116.1692169    0.3028847
>
```

Models 1-6 identical, but models with seven variables are different according to the three methods.

Cross-Validation vs. Criteria

- We can either look at the *test error* or make an adjustment to the *training error*.
- Given recent advancements in computation power there is little to say against CV.
- *One-standard-deviation* rule: When comparing MSE we should also compute the standard error and chose a model within one standard error of the best model (here 3 variables)



(James et al. 2013: 214)

Lab

- We will apply various selection methods
- Write a function to select best subset (weekend project)