LECTURE 10:

LINEAR MODEL SELECTION PT. 1

October 16, 2017

SDS 293: Machine Learning

Outline

- Model selection: alternatives to least-squares
- Subset selection
  - Best subset
  - Stepwise selection (forward and backward)
  - Estimating error
- Shrinkage methods
  - Ridge regression and the Lasso
  - Dimension reduction
- Labs for each part

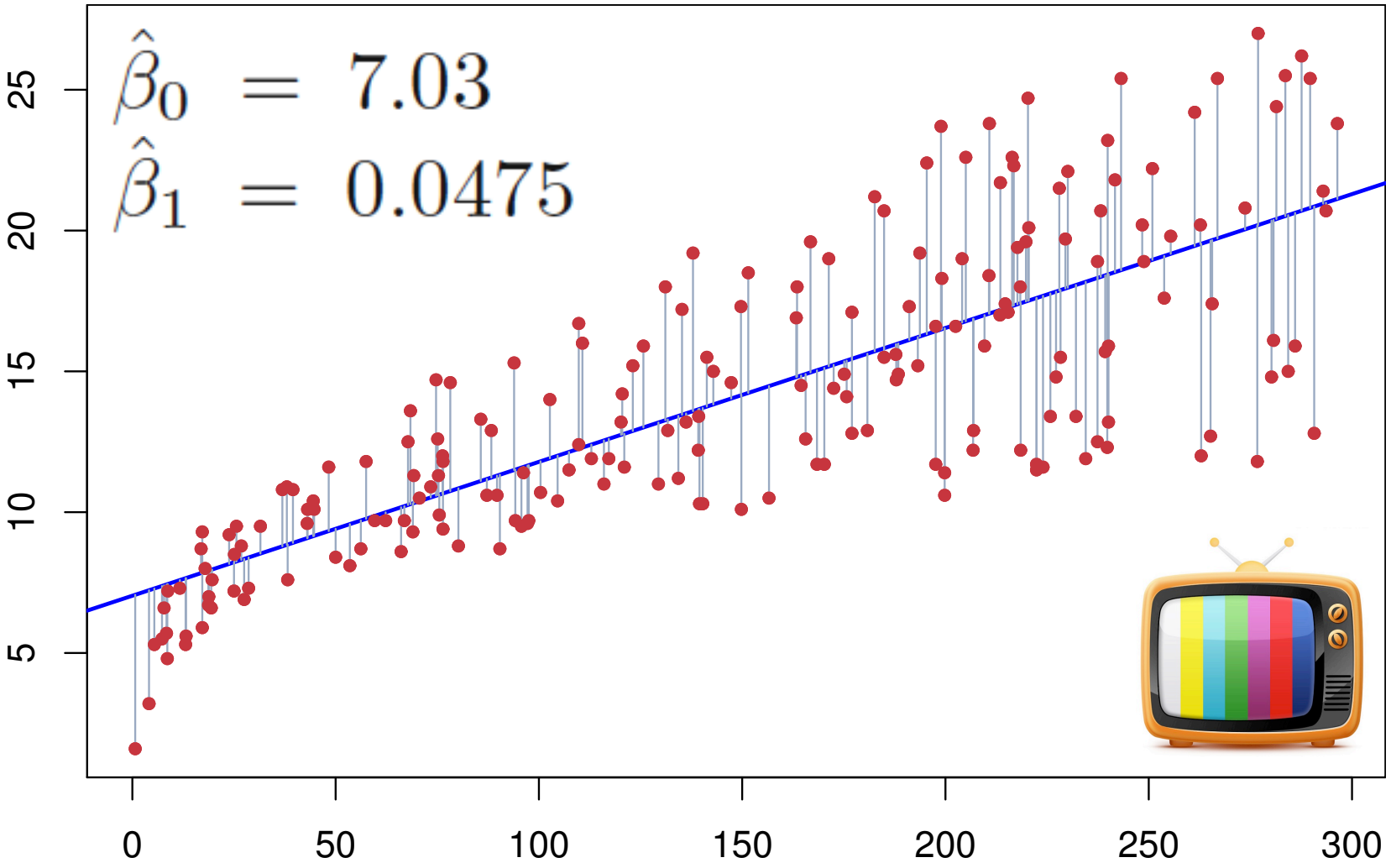
Back to the safety of linear models…

$$\mathtt{Y} \approx \beta_0 + \beta_1 \mathtt{X}_1 + \ldots + \beta_p \mathtt{X}_p$$

Back to the safety of linear models…

$$Y \approx \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Flashback: minimizing RSS



Discussion

Why do we **minimize RSS**?

(…have you ever questioned it?)

What do we know about least-squares?

- Assumption 1: we're fitting a **linear** model
- Assumption 2: the **true relationship** between the predictors and the response is **linear**

What can we say about the **bias** of our least-squares estimates?

What do we know about least-squares?

- Assumption 1: we're fitting a **linear** model
- Assumption 2: the **true relationship** between the predictors and the response is **linear**
- Case 1: the number of observations is much larger than the number of predictors (**n>>p**)

What can we say about the **variance** of our least-squares estimates?

What do we know about least-squares?

- Assumption 1: we're fitting a **linear** model
- Assumption 2: the **true relationship** between the predictors and the response is **linear**
- Case 2: the number of observations is **not much larger** than the number of predictors (**n≈p**)

What can we say about the **variance** of our least-squares estimates?

What do we know about least-squares?

- Assumption 1: we're fitting a **linear** model
- Assumption 2: the **true relationship** between the predictors and the response is **linear**
- Case 3: the number of observations is **smaller** than the number of predictors (**n<p**)

What can we say about the **variance** of our least-squares estimates?

Bias vs. variance

Discussion

How could we

reduce the variance?

Subset selection

- **Big idea:** if having too many predictors is the problem maybe we can get rid of some

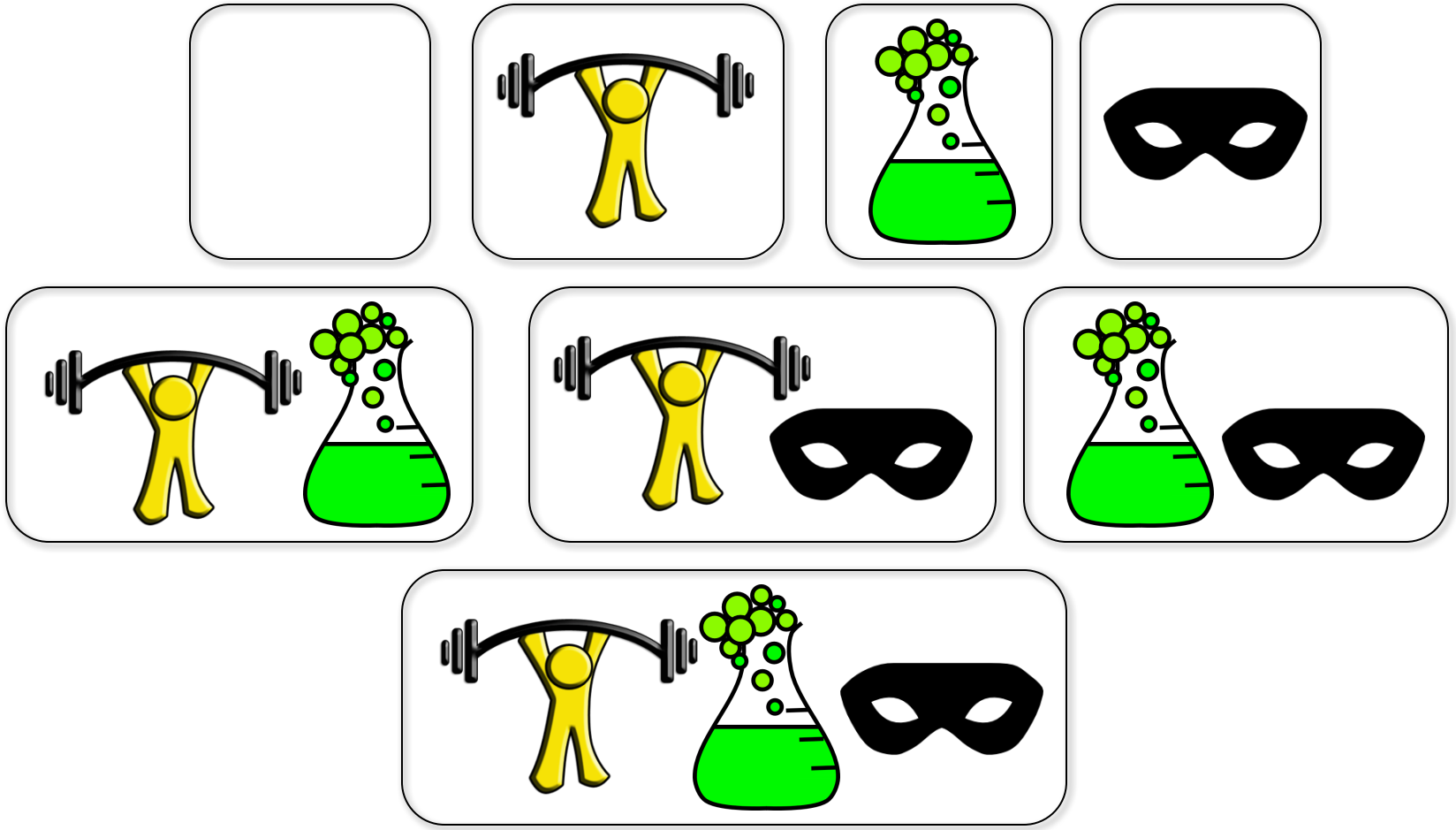
- **Problem:** how do we choose?

Flashback: superhero example

$$height = \beta_1 \left(\quad \right) + \beta_2 \left(\quad \right) + \beta_3 \left(\quad \right)$$

Image credit: Ming Malaykham

Best subset selection: try them all!



Finding the “best” subset

Start with the null model M_0 (containing no predictors)

1. For $k = 1,2,\ldots,p$:
   a. Fit all (p choose k) models that contain exactly p predictors.
   b. Keep only the one that has the smallest RSS (or equivalently the largest R^2). Call it M_k.
2. Select a single “best” model from among $M_0 \ldots M_p$ using cross-validated prediction error or something similar.

Discussion

Question 1: why not just use the one with the lowest RSS?

Answer: because you'll always wind up choosing the model with the highest number of predictors (why?)

Discussion

Question 2: why not just calculate the cross-validated prediction error on all of them?

Answer: so… many... models...

A sense of scale…

- We do a lot of work in groups in this class
- How many different possible groupings are there?
- Let's break it down:

47 individual people

1,081 different groups of two

16,215 different groups of three…

Model overload

- Number of possible models on a set of p predictors:

$$\sum_{k=1}^{p} \binom{p}{k} = 2^p$$

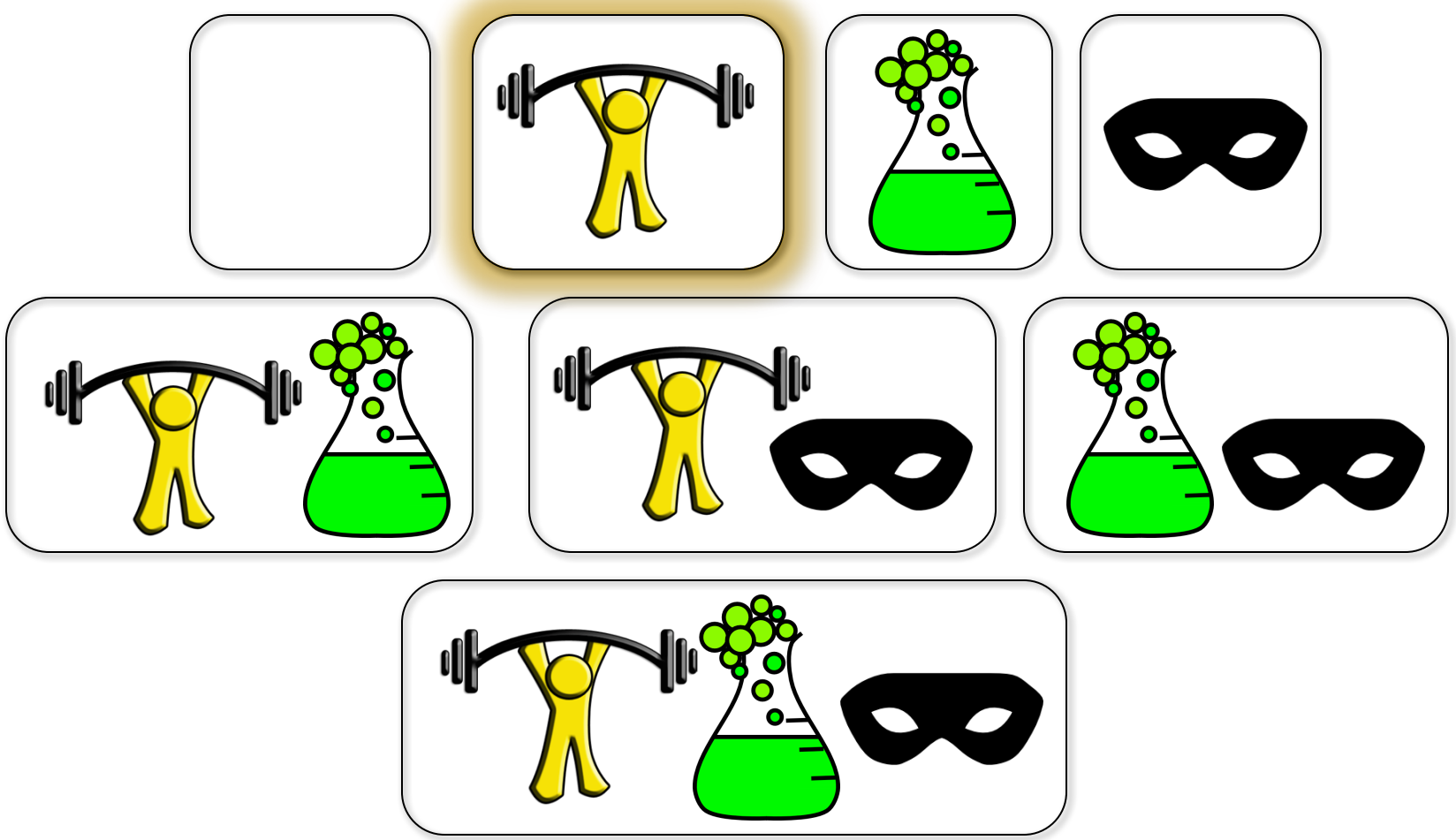
- On 10 predictors: **1,024** models
- On 20 predictors: **1,048,576** models

A bigger problem

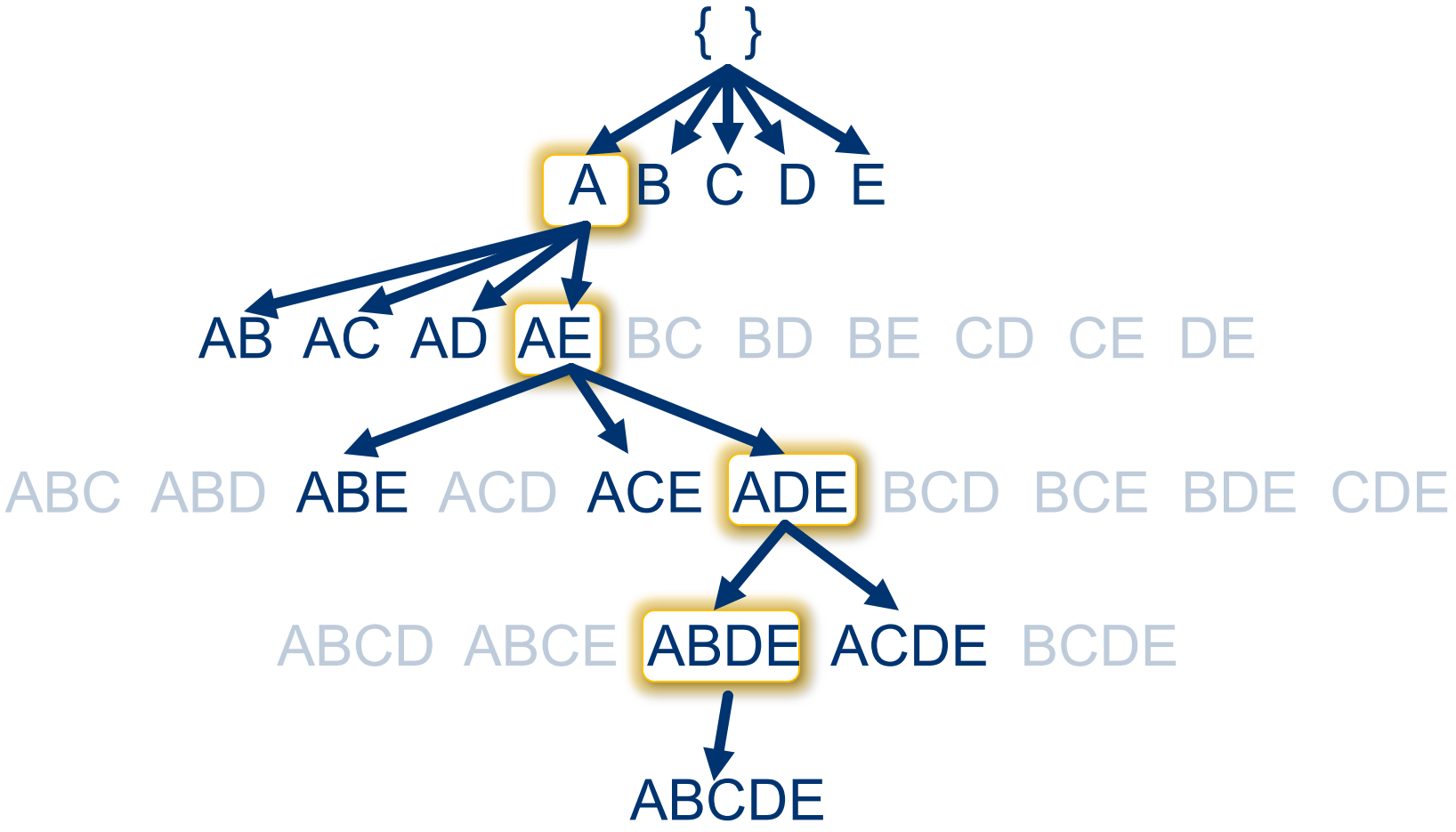
Question: what happens to our estimated coefficients as we fit more and more models?

Answer: the larger the search space, the larger the variance. We're overfitting!

What if we could eliminate some?



A slightly larger example ($p = 5$)



Best subset selection

Start with the null model M_0 (containing no predictors)

1. For $k = 1, 2, \ldots, p$:
   a. Fit all $(p \; choose \; k)$ models that contain exactly p predictors.
   b. Keep only the one that has the smallest RSS (or equivalently the largest R^2). Call it M_k.
2. Select a single "best" model from among $M_0 \ldots M_p$ using cross-validated prediction error or something similar.

Forward selection

Start with the null model M_0 (containing no predictors)

1. For $k = 1, 2, \ldots, p$:
   a. Fit all $(p - k)$ models that augment M_{k-1} with exactly 1 predictor.
   b. Keep only the one that has the smallest RSS (or equivalently the largest R^2). Call it M_k.
2. Select a single "best" model from among $M_0 \ldots M_p$ using cross-validated prediction error or something similar.

Stepwise selection: way fewer models

- Number of models we have to consider:

$$\sum_{k=1}^{p}\binom{p}{k} = 2^p \rightarrow \sum_{k=0}^{p-1}(p-k) = 1 + \frac{p(p+1)}{2}$$

- On 10 predictors: 1024 models → **51 models**
- On 20 predictors: over 1 million models → **211 models**

Forward selection

Question: what potential problems do you see?

Answer: there's a risk we might prune an important predictor too early. While this method usually does well in practice, it is not guaranteed to give the optimal solution.

Forward selection

Start with the null model M_0 (containing no predictors)

1. For $k = 1,2,\ldots,p$:
   a. Fit all $(p-k)$ models that augment M_{k-1} with exactly 1 predictor.
   b. Keep only the one that has the smallest RSS (or equivalently the largest R^2). Call it M_k.
2. Select a single "best" model from among $M_0 \ldots M_p$ using cross-validated prediction error or something similar.

Backward selection

Start with the full model M_p (containing all predictors)

1. For $k = p, (p-1), \ldots, 1$:
   a. Fit all k models that reduce M_{k+1} by exactly 1 predictor.
   b. Keep only the one that has the smallest RSS (or equivalently the largest R^2). Call it M_k.
2. Select a single "best" model from among $M_0 \ldots M_p$ using cross-validated prediction error or something similar.

Forward selection

Question: what potential problems do you see?

Answer: if we have more predictors than we have observations, this method won't work (why?)

Choosing the optimal model

- Flashback: measures of **training** error (RSS and R^2) aren't good predictors of **test** error (what we care about)
- Two options:
  1. We can **directly** estimate the test error, using either a validation set approach or cross-validation
  2. We can **indirectly** estimate test error by making an adjustment to the training error to account for the bias

Adjusted R^2

- **Intuition:** once all of the useful variables have been included in the model, adding additional junk variables will lead to only a small decrease in RSS

$$R^2 = 1 - \frac{RSS}{TSS} \rightarrow R^2_{Adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

- Adjusted R^2 pays a penalty for unnecessary variables in the model by dividing RSS by $(n\text{-}d\text{-}1)$ in the numerator

AIC, BIC, and C_p

- Some other ways of penalizing RSS

Estimate of the variance of the error terms

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

$$AIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + 2d\hat{\sigma}^2\right)$$

Proportional for least-squares models

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

More severe penalty for large models

Adjust or validate?

Question: what are the benefits and drawbacks of each?

	Adjusted measures	Validation
Pros	Relatively **inexpensive** to compute	More **direct** estimate (makes fewer assumptions)
Cons	Makes more **assumptions** about the model – more opportunities to be wrong	More **expensive**: requires either cross validation or a test set

Lab: subset selection

- To do today's lab in R: **`leaps`**
- To do today's lab in python: **`itertools`**, **`time`**
- Instructions and code:

  [course website]/labs/lab8-r.html

  [course website]/labs/lab8-py.html
- Full version can be found beginning on p. 244 of ISLR

Coming up

- Estimating error with cross-validation
- A3 due tonight by 11:59pm
- A4 out, due Oct. 20th