Introduction to Machine Learning – Subset Selection

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

Plan for Today

- Validation / Error
- Linear Model selection and Regularization
  - Best Subset Selection
  - Stepwise Selection
  - Choosing the Optimal Model

Motivation

So far, we've thought about how to evaluate our models using error calculations.

Besides fit, what else might influence the effectiveness / accuracy of our models?

Motivation

So far, we've thought about how to evaluate our models using error calculations.

Besides fit, what else might influence the effectiveness / accuracy of our models?

- https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms?language=en
- http://gendershades.org/overview.html

Motivation

Moving back to linear models…

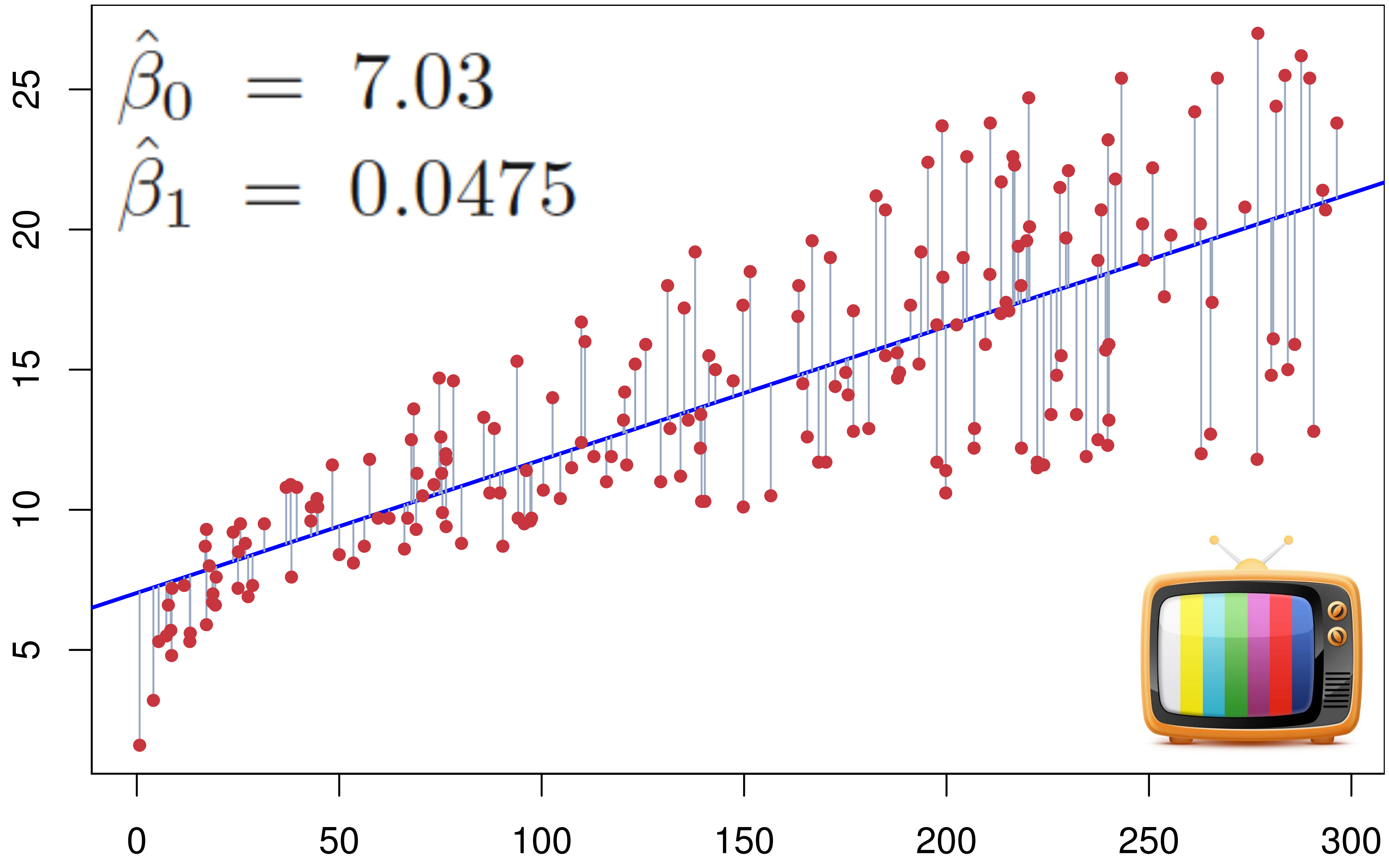
In regression, the standard model is:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

How did we find the coefficients for this model?

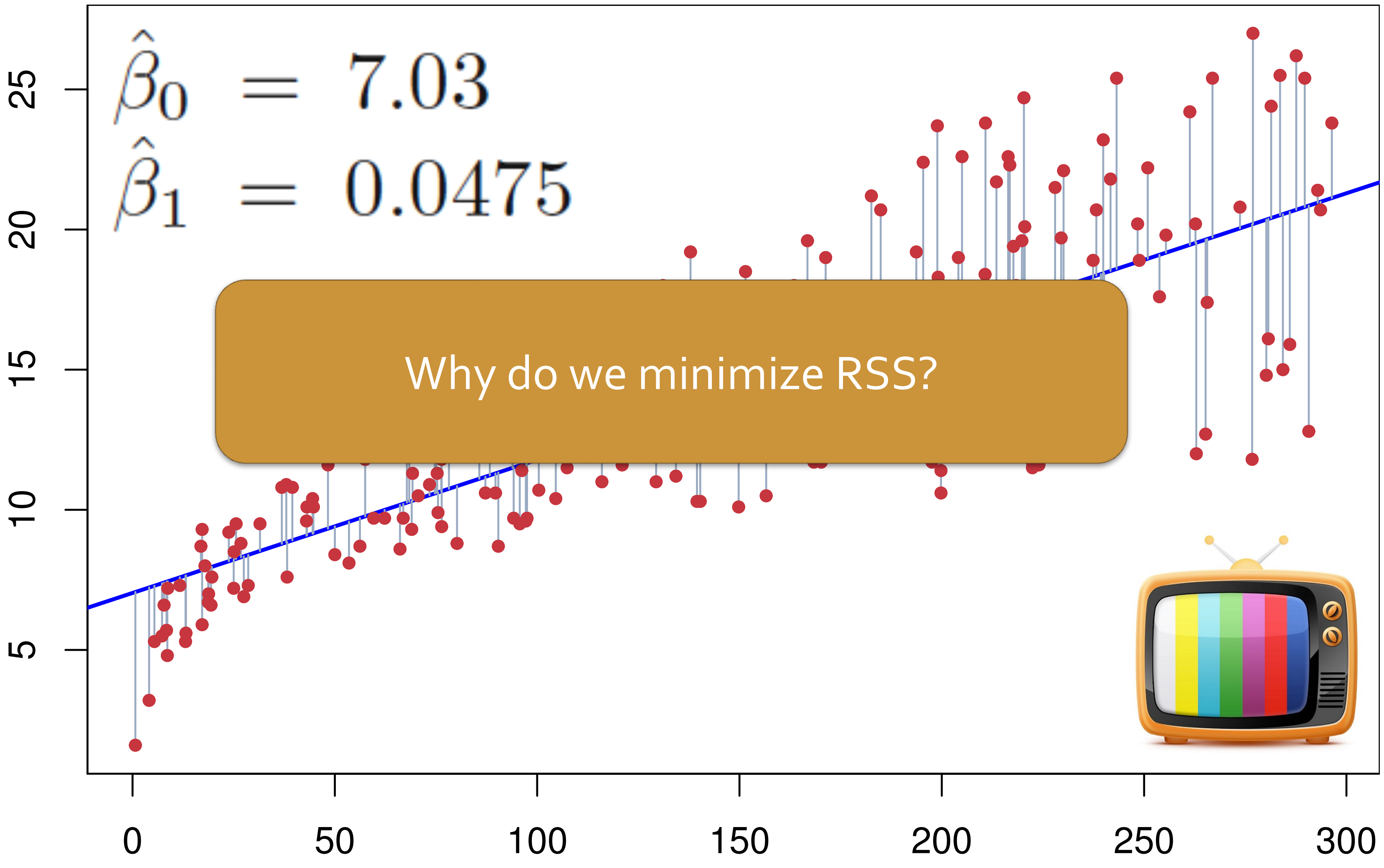
Motivation

Flashback: minimizing RSS



Motivation

Flashback: minimizing RSS



Why do we minimize RSS?

Motivation

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?

Motivation

Least Squares

Assumption 1: we're fitting a linear model

Assumption 2: the true relationship between the predictors and the response is linear

Consider:

- Case 1: the number of observations in the training data is much larger than the number of predictors (n >>p)
  - Variance should be low; our model will perform well on test observations

Motivation

Least Squares

Assumption 1: we’re fitting a linear model

Assumption 2: the true relationship between the predictors and the response is linear

Consider:

- Case 2: the number of observations is not much larger than the number of predictors ($n \approx p$)
  - Variance will get (pretty) high; performance of the model will suffer

Motivation

Least Squares

Assumption 1: we're fitting a linear model

Assumption 2: the true relationship between the predictors and the response is linear

Consider:

- Case 3: the number of observations smaller than the number of predictors ($n < p$)
  - Variance will be infinite; we will no longer get a unique least squares estimate

Motivation

So what can we do in cases where $n < p$ or $n \approx p$?

Ideas?

Subset selection

So what can we do in cases where $n < p$ or $n \approx p$?

Subset selection

- If we have too many predictors, we can get rid of some to improve model performance

- But how will we choose?

Ideas?

Subset selection

So what can we do in cases where $n < p$ or $n \approx p$?

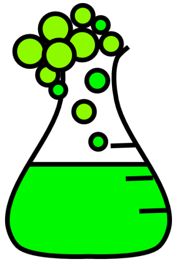
Subset selection

- If we have too many predictors, we can get rid of some to improve model performance

- But how will we choose?
  - For each possible subset of predictors
    - Fit least squares
  - Choose the "best" model from the collection

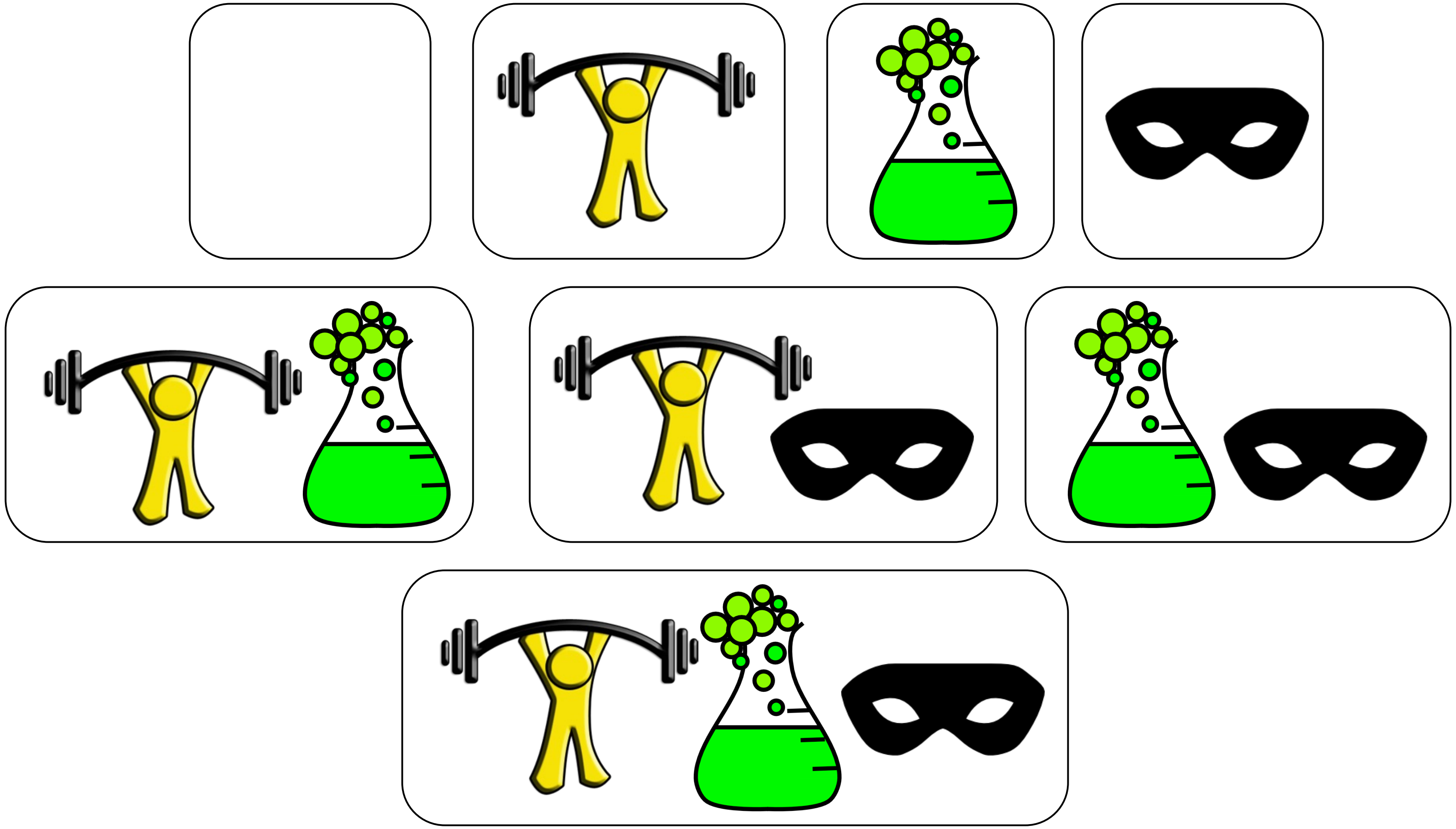
Subset selection

Superhero Example

$height = \beta_1 \big($  $\big) + \beta_2 \big($  $\big) + \beta_3 \big($  $\big)$

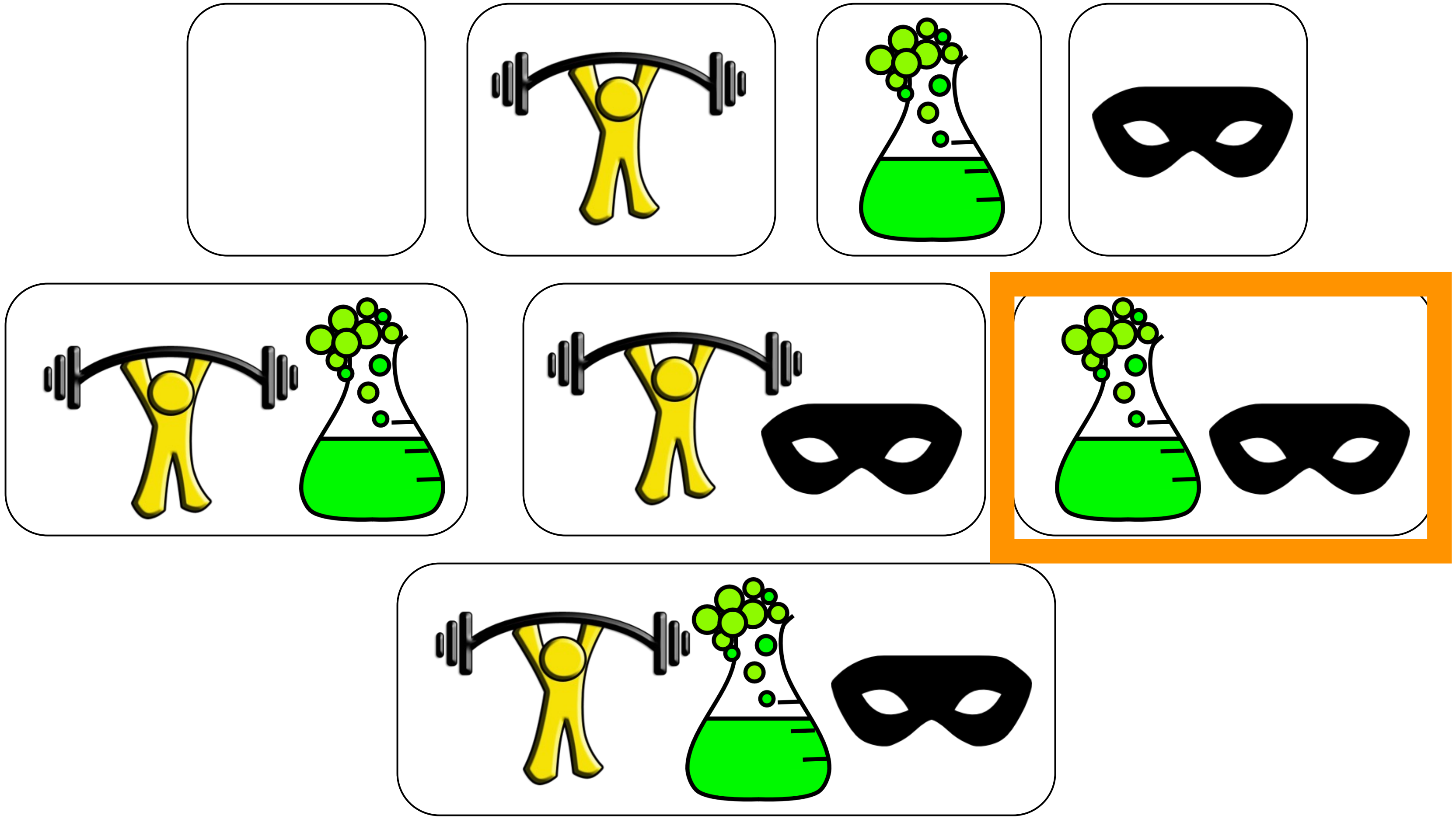
Superhero Example

Subset selection



Superhero Example

Subset selection



Subset selection

Algorithm

Start with the null model, M_0 (no predictors)

For $k = 1, 2, \ldots, p$

Fit all $\binom{p}{k}$ models that contain exactly k predictors

Keep only the one that has the smallest RSS. Call it M_k

Select the "best" model from $M_0, M_1, \ldots M_p$

Subset selection

Algorithm

Start with the null model, M_0 (no predictors)

For $k = 1, 2, \ldots, p$

Fit all $\binom{p}{k}$ models that contain exactly k predictors

Keep only the one that has the smallest RSS. Call it M_k

Select the "best" model from $M_0, M_1, \ldots M_p$

Can we do this for models that are not least-squares regression models?

Subset selection

Algorithm

Start with the null model, M_0 (no predictors)

For $k = 1, 2, \ldots, p$

Fit all $\binom{p}{k}$ models that contain exactly k predictors

Keep only the one that has the smallest RSS. Call it M_k

Select the "best" model from $M_0, M_1, \ldots M_p$

Subset selection

Algorithm

Start with the null model, M_0 (no predictors)

For $k = 1, 2, \dots, p$

Fit all $\binom{p}{k}$ models that contain exactly k predictors

Keep only the one that has the smallest RSS. Call it M_k

Select the "best" model from $M_0, M_1, \dots M_p$

What drawbacks do you see to this approach?

Subset selection

Model overload

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

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

How many possible models do we get for 10 predictors? For 20?

Subset selection

Model overload

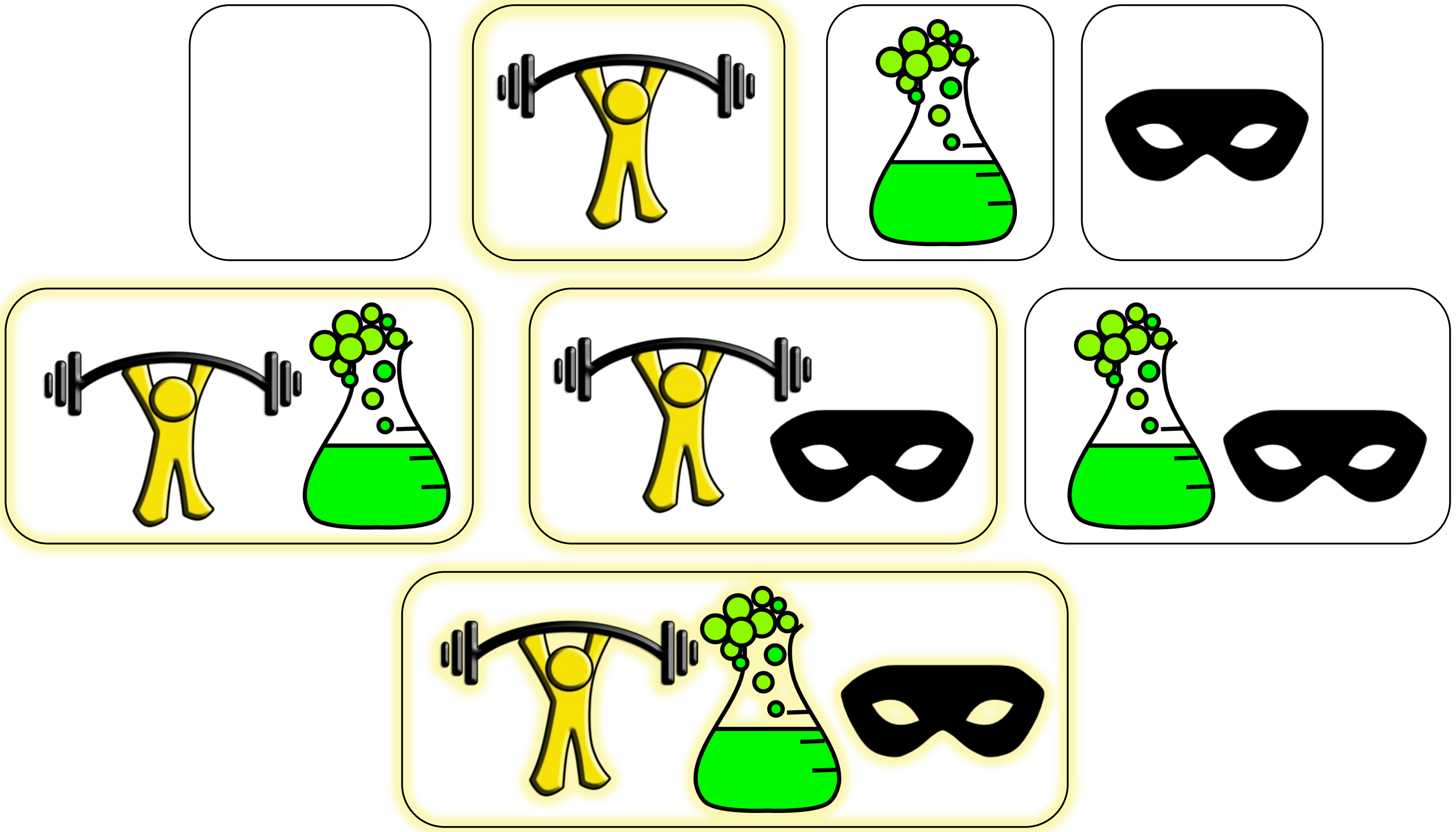
- Number of possible models on a set of p predictors is

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

As we fit more and more models, what happens to our estimated coefficients?

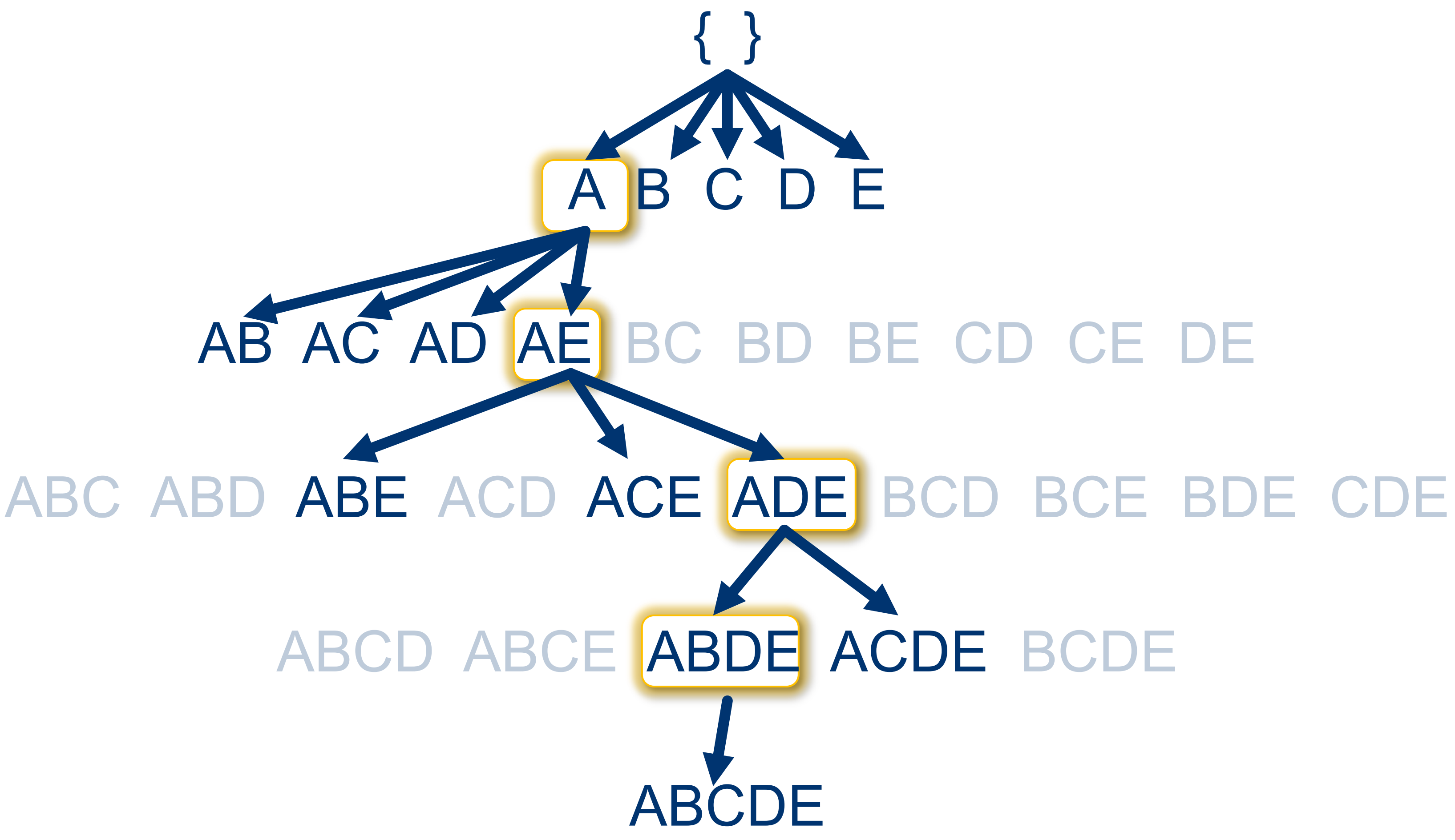
What if we could eliminate some of our model options?

Subset selection



Subset selection

Ex. when $p = 5$



Forward selection

Algorithm

Start with the null model, M_0 (no predictors)

For $k = 1, 2, \ldots, p$

Fit all $(p-k)$ models that augment M_{k-1} with exactly one predictor

Keep only the one that has the smallest RSS. Call it M_k

Select the "best" model from $M_0, M_1, \ldots M_p$

Stepwise selection

Far fewer models than best subset

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

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

How many possible models do we get for 10 predictors? For 20?

Forward selection

Algorithm

Start with the null model, M_0 (no predictors)

For $k = 1, 2, \dots, p$

Fit all $(p - k)$ models that augment M_{k-1} with exactly one predictor

Keep only the one that has the smallest RSS. Call it M_k

Select the "best" model from $M_0, M_1, \dots M_p$

What potential downsides do you see?

Backward selection

Algorithm

Start with the full model, M_p (all predictors)

For $k = p, p-1, \dots, 1$

Fit all k models that reduce M_{k+1} by exactly one predictor

Keep only the one that has the smallest RSS. Call it M_k

Select the "best" model from $M_0, M_1, \dots M_p$

Backward selection

Algorithm

Start with the full model, M_p (all predictors)

For $k = p, p-1, \dots, 1$

Fit all k models that reduce M_{k+1} by exactly one predictor

Keep only the one that has the smallest RSS. Call it M_k

Select the "best" model from $M_0, M_1, \dots M_p$

Choosing the Optimal Model

Recall: measures of **training** error (RSS and R^2) are not good predictors of **test** error.

In addition, they are not good ways of comparing models with different numbers of predictors.

Why?

Choosing the Optimal Model

Recall: measures of **training** error (RSS and R^2) are not good predictors of **test** error

We have two options for estimating test error:

- Directly estimate using a validation set or CV
- **Indirectly estimate by making an adjustment to the training error to account for 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

$$R^2 = 1 - \frac{RSS}{TSS}, (TSS = \sum(y_i - \bar{y})^2)$$

$$R^2_{Adj} = 1 - \frac{\frac{RSS}{n-d-1}}{\frac{TSS}{n-1}}, d = num\_predictors$$

To maximize R^2_{Adj} (get the best fit), we need to minimize $\frac{RSS}{n-d-1}$

As d increases, what will happen to this term?

AIC, BIC, and C_p

Other ways to penalize RSS when more predictors are added:

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

As d increases, what will happen to C_p?

AIC, BIC, and C_p

Other ways to penalize RSS when more predictors are added:

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

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

AIC, BIC, and C_p

Other ways to penalize RSS when more predictors are added:

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

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

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

As d increases, what will happen to BIC? How will this differ if d is very large vs small?

AIC, BIC, and C_p

Other ways to penalize RSS when more predictors are added:

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

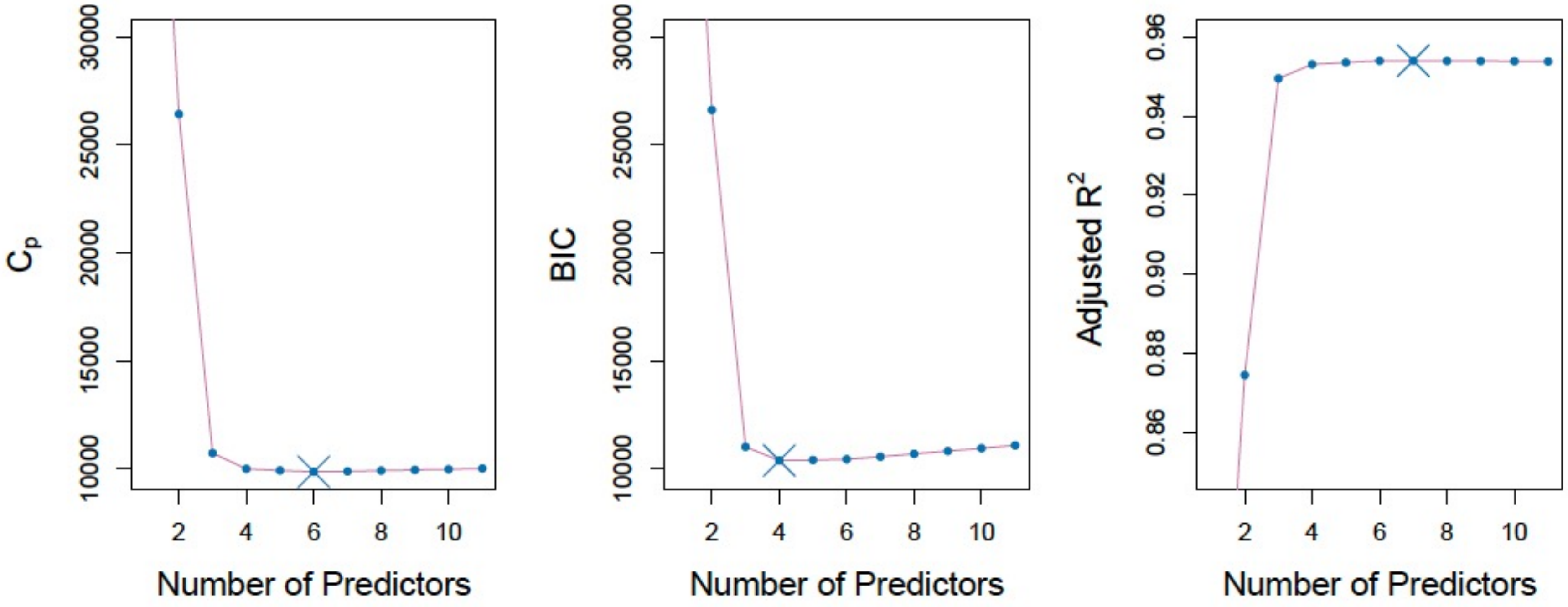
$$AIC = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

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

All of these measures are supported by statistical theory. AIC and BIC are also defined for more general models beyond least squares.

AIC, BIC, and C_p

Credit dataset



Choosing the optimal model

We have two options for estimating test error:

- Directly estimate using a validation set or CV
- Indirectly estimate by making an adjustment to the training error to account for bias

When do we use one vs the other?

- CV used to be very computationally expensive, so indirect measurements were developed
- With the computers we have today, CV is no longer as expensive and is the preferred method

Choosing the optimal model

We have two options for estimating test error:

- Directly estimate using a validation set or CV
- Indirectly estimate by making an adjustment to the training error to account for bias

When do we use one vs the other?

- CV used to be very computationally expensive, so indirect measurements were developed
- With the computers we have today, CV is no longer as expensive and is the preferred method

What if all models have about the same error (within one SE of each other)?

- Go with the simplest!