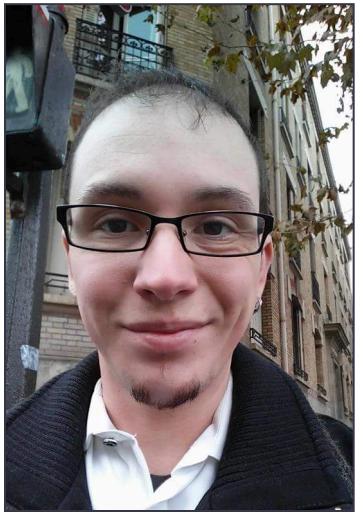
LINEAR MODEL SELECTION

March 10, 2016

R. Jordan Crouser

Guest Lecture in SDS 291: Multiple Regression

Introductions & Background



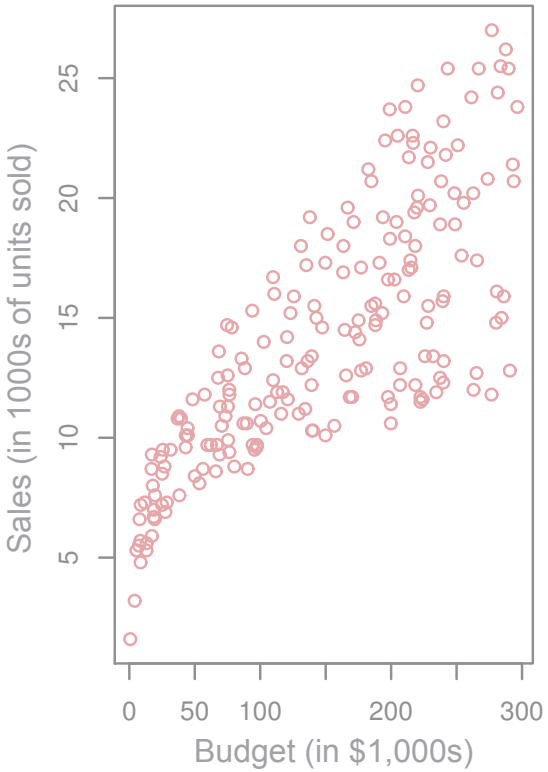
- 2015 to now: Assistant Prof. in SDS (Smith)
- 2013 – 2015: Research Scientist (MIT)
- 2010 – 2013: PhD in Visual Analytics (Tufts)
- 2008 – 2010: MSc in Educational Tech. (Tufts)
- 2004 – 2008: BA in CS and Math (Smith)

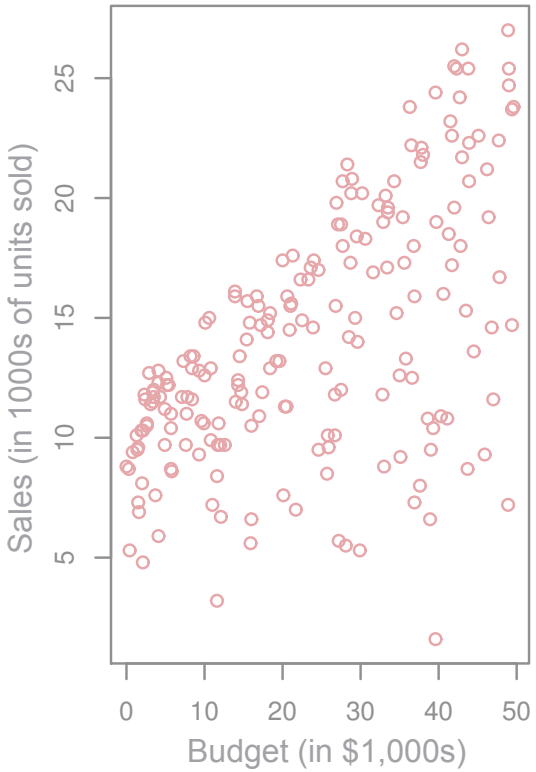
Outline

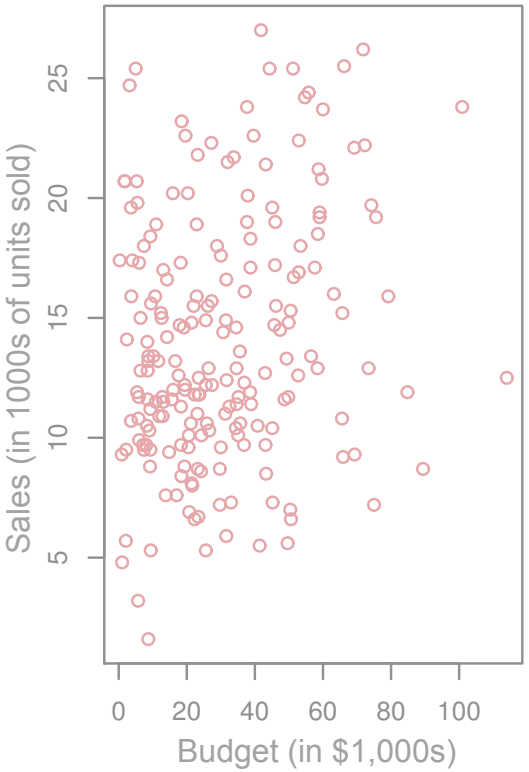
- Running example: advertising sales
- Subset selection
  - Best subset
  - Forward selection
  - Backward selection
  - Stepwise selection
  - Estimating error
- Lab

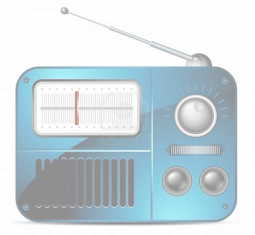
Running example

Last year's advertising budget

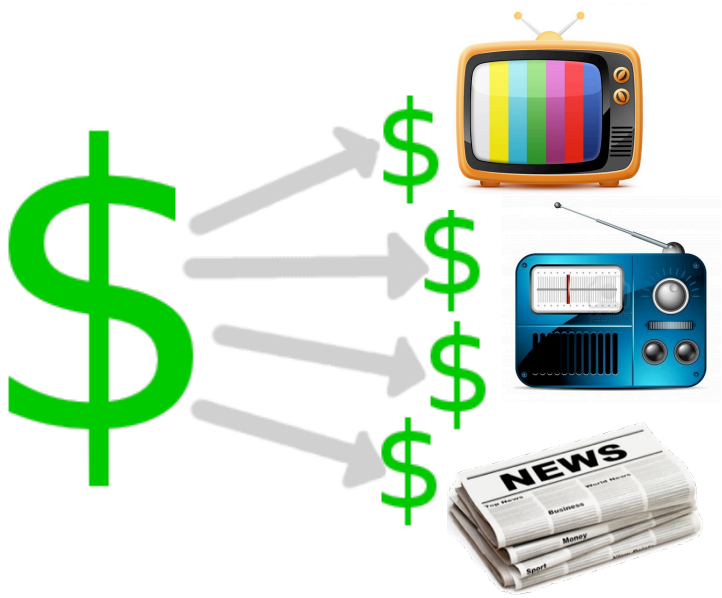








Your task

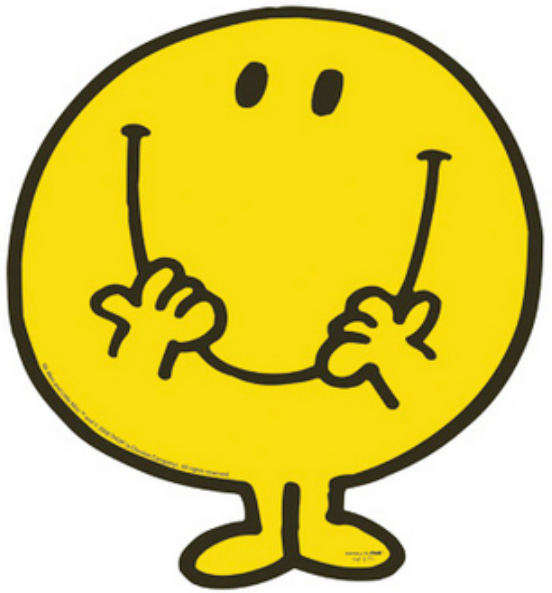


So what do we do?

Ahh, linear models…

Provides nice, interpretable results and is a good starting point for many applications:

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



Thanks, Roger Hargreaves!

Q1: is at least one predictor useful?

- In SLR, we tested to see if the slope was 0 (no effect)
- In MLR, we need to test whether **ALL** of the slopes are 0 to prove that there is no effect
- **Question:** how do we do that?

Q1: is at least one predictor useful?

- **Answer:** use the F-statistic:

$$F = \frac{(TSS - RSS)/p}{(RSS)/(n - p - 1)}$$

where p is the # of predictors and n is the sample size

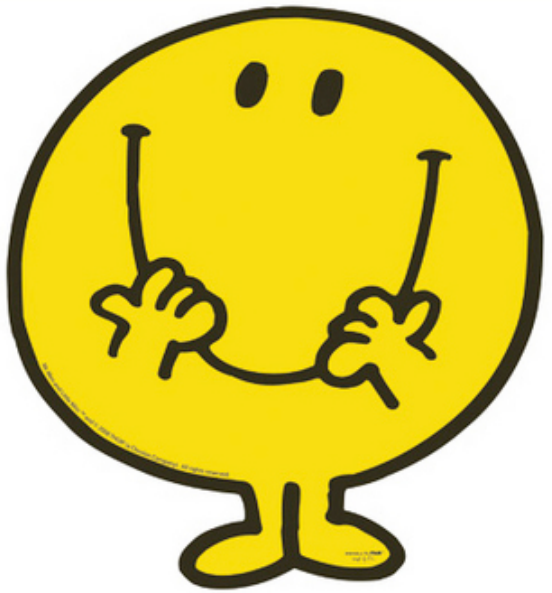
- Value close to 1 → no effect

- **Question**: why look at the F-statistic and not just at the p-values for each predictor in turn? (*hint*: lots of predictors?)

Ahh, linear models…

- The linear regression model provides nice, interpretable results and is a good starting point for many applications

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

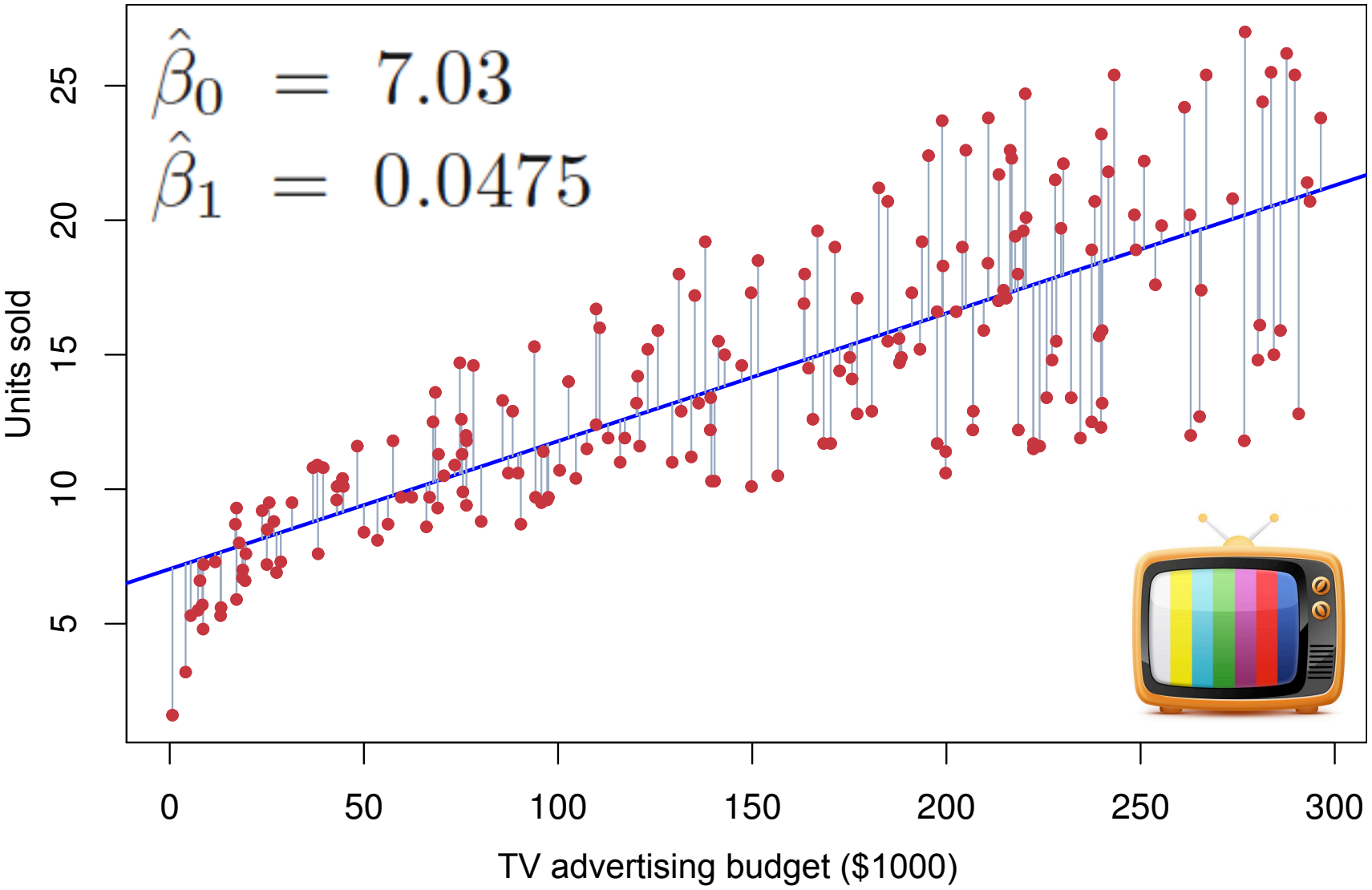


Thanks, Roger Hargreaves!

But wait…

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

Answer: minimizing RSS



Discussion

What happens when 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?

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?

Case 2: the number of observations is **not much larger** than the number of predictors ($n \approx p$)

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

What do we know about least-squares?

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

So how do we reduce the variance in our estimate?

Q2: do we need them all?

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

- We know that at least one predictor has an effect: **which one(s) is it**?

- Determining **which** predictors are associated with the response is referred to as *model selection*

Method 1: 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 k 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: why not just pick the one that minimizes RSS?

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

Thought exercise: how many groups?

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

10 individual people

45 different groups of two

120 different groups of three…

Model overload

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

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

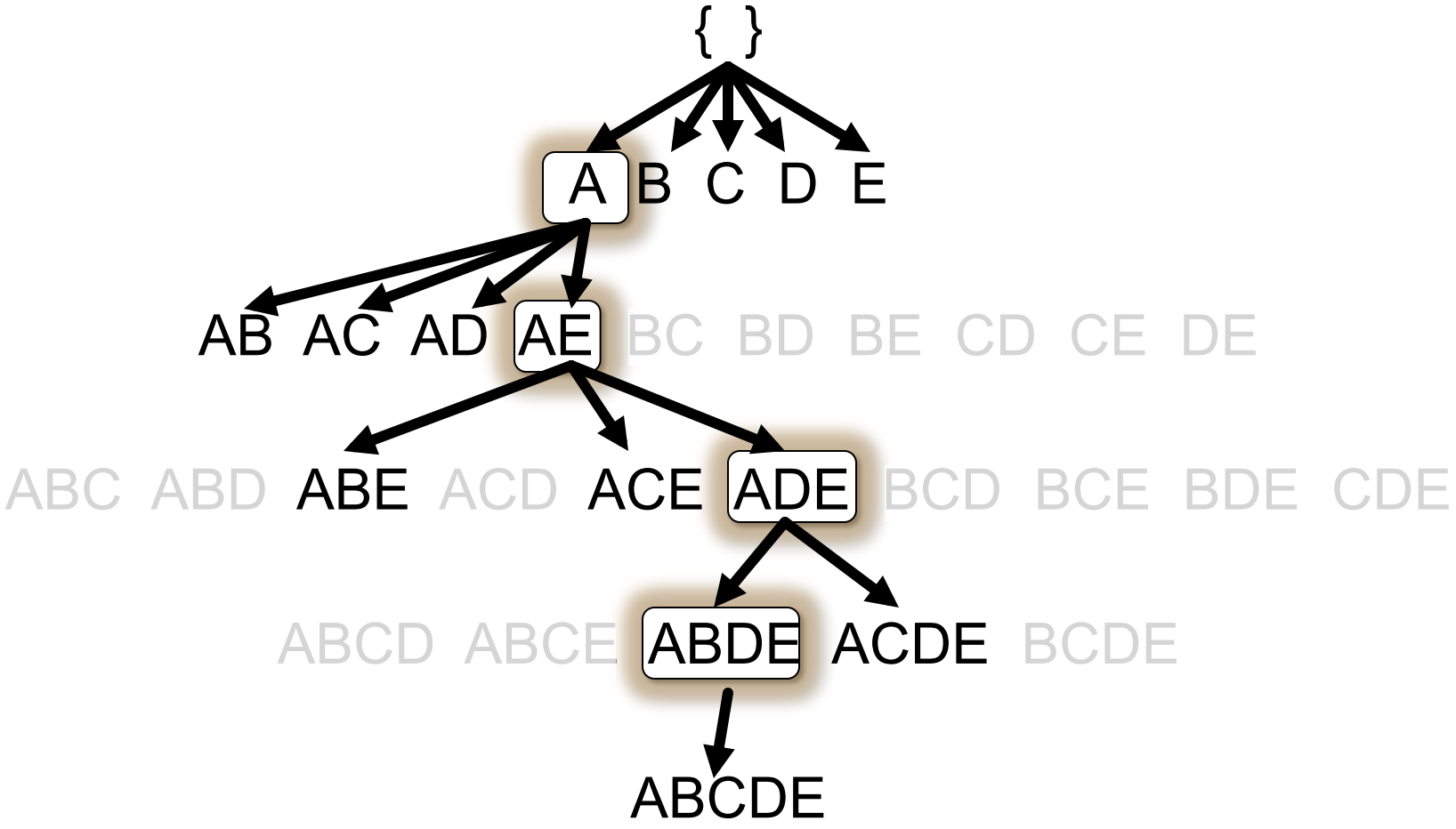
- On 10 predictors: 1024 models
- On 20 predictors: over 1 million

Another 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?



Method 2: 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.

Forward 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

Discussion: 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.

Method 2: 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.

Method 3: backward selection

Start with the full model M_p (containing all predictors)

1. For $k = p, p\text{-}1, \ldots 1$:
   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.

Discussion: backward selection

Question: what potential problems do you see?

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

Method 4: stepwise selection

1. Start with the null model.
2. As with forward selection, iteratively add the variable that provides the best fit.
3. If the p-value for one of the variables in the model rises above a certain threshold, remove that variable.
4. Iterate until all variables *in the model* have a sufficiently low p-value , and all variables *outside the model* would have a large p-value if added.

Discussion: stepwise selection

Question: what potential problems do you see?

Answer: what did the ASA just say??

> *"Widespread use of 'statistical significance' (generally interpreted as 'p < 0.05') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process…"*

Choosing the optimal model

- Each of these methods produced a set of models (one of each size) and then said:

Select a single "best" model from among $M_0 \ldots M_p$

- **Question:** how do we do that?

Choosing the optimal model

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

Adjusted R^2

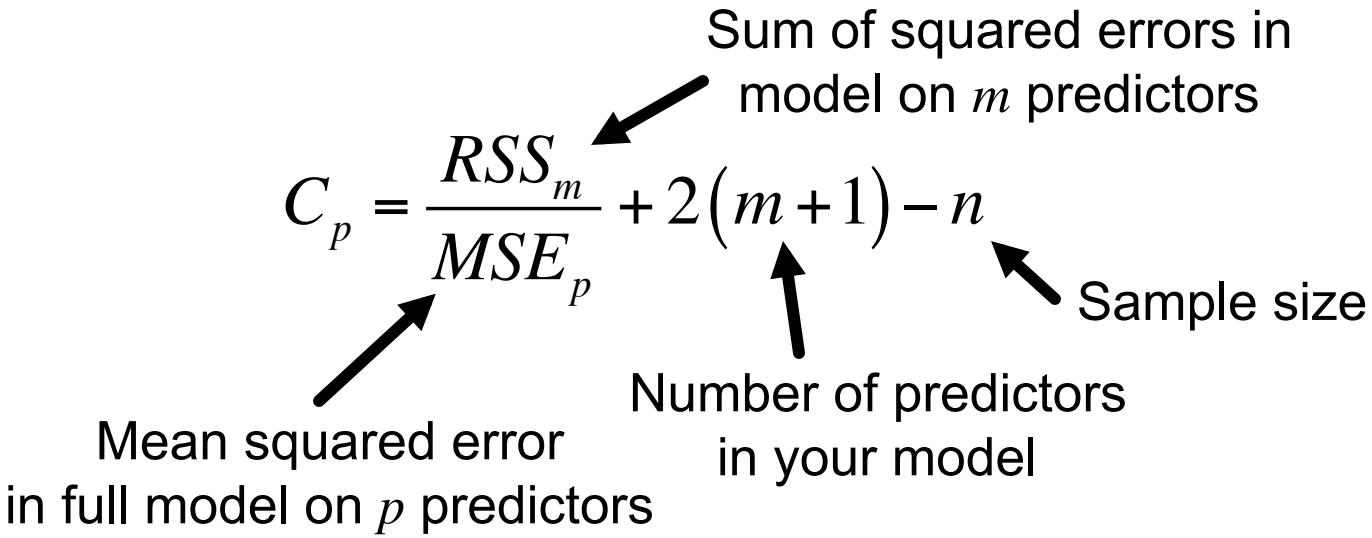
- **Intuition:** once all of the useful variables have been included in the model, adding additional junk variables will lead to only a small decrease in RSS

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

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

Mallow's C_p

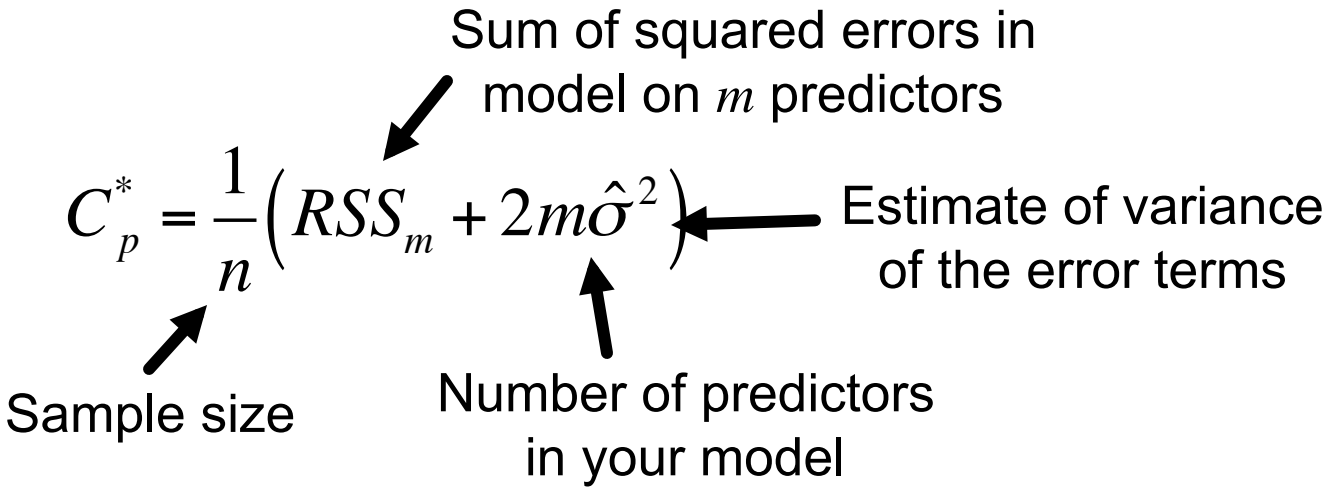
Your book tends to favor:



Small $C_p \Rightarrow$ better model

Mallow's C_p: alternate definition

My preferred way to think about this:



Small C_p ⇒ better model

Model with smallest C_p = model with smallest C_p^*

Akaike Information Criterion (AIC)

Default criteria in R:

In more complex models, this gets replaced with a max-likelihood function

$$AIC = 2m + n\ln(RSS)$$

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

Small AIC $\Rightarrow$ better model

Bayesian Information Criterion (BIC)

Amelia's favorite:

Each additional term incurs a harsher penalty

$$BIC = m\ln(n) + n\ln\left(\frac{RSS}{n}\right)$$

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

Small BIC $\Rightarrow$ better model

Comparing AIC, BIC, and C_p

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

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

Proportional for least-squares models

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

More severe penalty for large models

Caveats

- Automated methods are not substitutes for careful analysis:
  - Are your assumptions met?
  - Are measurements efficient?
  - Is it worth it to squeeze a little bit more performance out of R^2?
  - What about transformations?

Lab: model selection

- Instructions and code:

http://www.science.smith.edu/~amcnamara/sds291/lab_stepwise.html

Thanks, and enjoy your spring break!