Multiple Regression 3:
Model Selection

1. Some Perspective: Why do we want to choose a model?
2. Basic approaches
3. AIC Criteria

"Manual" Model Selection (Small # of Models)

4. Polynomial Regression Example
5. ANCOVA Example

Automated Model Selection (Larger # of Models)

6. Forward, Backward or Stepwise selection
7. Best (or All) Subsets selection
8. Discussion

Examples:

1. Glue Strength Example: AIC
2. Highway1: Stepwise Selection
3. Highway2: Best Subsets Selection
4. Highway3: "Bigger" data
5. Model Selection Simulation
6. Cement Example: Akaike weights

1. Some Perspective: Why do we want to choose a model?

Explanation or Prediction

Note: This discussion is taken almost verbatim from a presentation by Brian Ripley.

Model selection for **explanation** is like doing scientific research. For explanation, Occam's razor applies and we want to:

"Make everything as simple as possible, but not simpler."

(Albert Einstein.)

For explanation, we have a concept of a "true" model or at least a model that is an approximation of truth because:

"All models are false, but some are useful." (George Box)

Model selection for **prediction** purposes is more like doing engineering development. All that matters is that is works. And if the aim is prediction, model choice should be based on the quality of the predictions.

Workers in pattern recognition have long recognized this and used training data to choose between models and separate test data to assess the quality of predictions from the chosen model.

For either **explanation or prediction**, it is possible that there may be several (roughly) equally good models.

More discussion on model selection for explanation or prediction in the discussion section (toward the end of the notes).

A few more quotes about model selection:

“I would no more let an automatic routine select my model than I would let some best-fit procedure pack my suitcase.” Ronan Conroy

“The data analyst knows more than the computer”…. “failure to use that knowledge produces inadequate data analysis”. Henderson and Velleman

““Let the computer find out” is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting.” Burnham and Anderson

Some “warnings” about model selection methods:

1. There is NO reason that different model selection methods should arrive at the same model! If that was the case, there would be no need for multiple methods. However, it is nice when different methods DO select the same model.
2. In most cases, the model selected can have terms that are not statistically significant (ex: AIC).
3. Model selection methods do NOT replace:
   A. User judgment and interpretation
   B. Diagnostic plots!

2. Basic Model Selection Approaches

Backward Elimination: We start with the most complicated model (some judgement required here) and remove terms from the model one at a time.

Forward selection: We start with the simplest model and add terms to the model one at a time.

Compare Models based on some Criteria: Unlike the approaches above, this is not a sequential process. We fit some number of models (possibly "all" of them) and compare them based on some criteria. Most common criteria is AIC (or AICc or BIC).

Some things to consider:

1. "**Principle of Hierarchy**", which says that if your model includes a higher order term (ex: polynomial or interaction terms) it should also include the lower order terms that contribute. We will focus on model selection approaches that follow this principle. For example, if your model includes the interaction between X1*X2 it should also include X1 and X2.
2. **Categorical predictors:** Remember that when we include a categorical predictor, R creates a group of indicator variables (#levels -1). We will focus on model selection approaches that include or exclude all of the indicators corresponding to a single categorical predictor. For hypothesis testing approaches, we will use ANOVA F-tests to decide whether to include or exclude a categorical predictor.

3. AIC Criteria

Akaike Information Criterion (AIC) is based on information coding theory.

$$\text{AIC} = n \ln\left(\frac{SS\text{Resid}}{n}\right) + 2p$$
$$= -2\ln(L) + 2p$$
$$\text{where p} = \#\text{parameters}$$
$$\text{and L} = \text{likelihood}$$

1. **Choose the model with <u>smallest</u> AIC!**

2. **AIC deals with the trade-off between the goodness of fit of the model and the complexity of the model.** The left term is related to the error variance. The right term is a "penalty" for including parameters. We try to minimize the estimated error variance, subject to a cost (or penalty) for adding variables.

3. **p is the # parameters in the model.** This includes the intercept.

- Some people (Burnham and Anderson) also include σ^2 in the count of parameters.
- R and SAS do NOT include σ^2 in the count of parameters.
- As long as we are consistent, it doesn't matter for the purpose of comparing AICs in a given problem. (See note 4.)
- For multiple regression, p = k +1 (# predictors + intercept)
- For other models, the ANOVA table can be helpful for counting the number of (non-zero) parameters. Specifically,

p = df Model + 1 (for the intercept)

4. **AIC is useful for comparing models on a fixed data set.** Comparing AIC values across data sets is meaningless!

5. Since n (sample size) is used in the calculation of AIC, be careful of missing values (which can reduce sample size). We want to compare AIC values on the same data with the same sample size.

6. All that matters is the difference between two AIC values. Negative AIC values are possible. In addition, software programs may calculate AIC differently (up to an additive constant), but the difference (Δ) in AIC should be the same. We will see this in the Glue Strength Example. Be careful about comparing AIC values from different software programs. This includes different functions/libraries in R or different Procs in SAS.

7. AIC is a general method of model selection, used in many types of models, from time series to categorical data, not just regression.

8. A Model that is a reduced version of a "full" model may be compared to the full model using the "Likelihood Ratio Test" (more about this later). When the models differ by only one parameter AIC will be lower for the full model if the Likelihood Ratio Test has p-value less that 0.157. Therefore, lowest AIC is generally a more liberal criterion for including parameters than hypothesis testing at the usual 0.05 level.

9. The model with the smallest AIC value can include terms that are not significant at the 0.05 level.

10. For mixed models, if REML estimation is used, we can only compare AIC values for models with the same fixed effects! More on this later.

4. Polynomial Regression Example

Polynomial regression is an example of a "complete" hierarchy, because each model extends the previous model.

(0) Null Model: $y = \beta_0 + \varepsilon$

(1) Linear: $y = \beta_0 + \beta_1 x + \varepsilon$

(2) Quadratic: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

(3) Cubic: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

In this example, we will consider backwards elimination and forward selection. But AIC criteria could also be used to compare models.

Backward elimination for Polynomial Regression:

1. Start with Cubic model (most complicated model considered), look at p-value corresponding to cubic (x^3) term (highest order term). If significant ($p < 0.05$), use Cubic model.

2. If previous test not significant ($p \geq 0.05$), move to Quadratic model. Look at p-value corresponding to quadratic (x^2) term (highest order term). If significant ($p < 0.05$), use Quadratic model.

3. If previous test not significant ($p \geq 0.05$), move to Linear model. Look at p-value corresponding to linear (x) term to determine statistical significance.

Forward Selection for Polynomial Regression:

1. Start with Linear model, look at p-value corresponding to the linear (x) term (highest order term). If not significant ($p \geq 0.05$), stop with Null model.
2. If previous test significant ($p < 0.05$), move to Quadratic model. Look at p-value corresponding to quadratic (x^2) term (highest order term). If not significant ($p < 0.05$), use Linear model.
3. If previous test significant ($p < 0.05$), move to the Cubic model, look at p-value corresponding to cubic (x^3) term (highest order term). If not significant ($p < 0.05$), use Quadratic model.
4. If previous test significant ($p < 0.05$), use the Cubic model.

1. Models are compared using t-test (from summary() output) or F-test (from ANOVA table) for the parameter being added or omitted. Recall that when considering only 1 parameter, t-test is equivalent to the F-test.
2. Backward elimination looks for lack of significance to move to a less complicated model. Forward selection looks for significance to move to a more complicated model.
3. Forward selection can stop too soon, e.g. stop at linear, when cubic would have been significant. Check diagnostic plots!
4. Backward elimination can leave you, by chance, with an unnecessarily complicated model.
5. Backward and Forward approaches can result in different models.
6. Use plots to help with your decisions.

5. ANCOVA Example

ANCOVA is an example of a "partial" hierarchy.

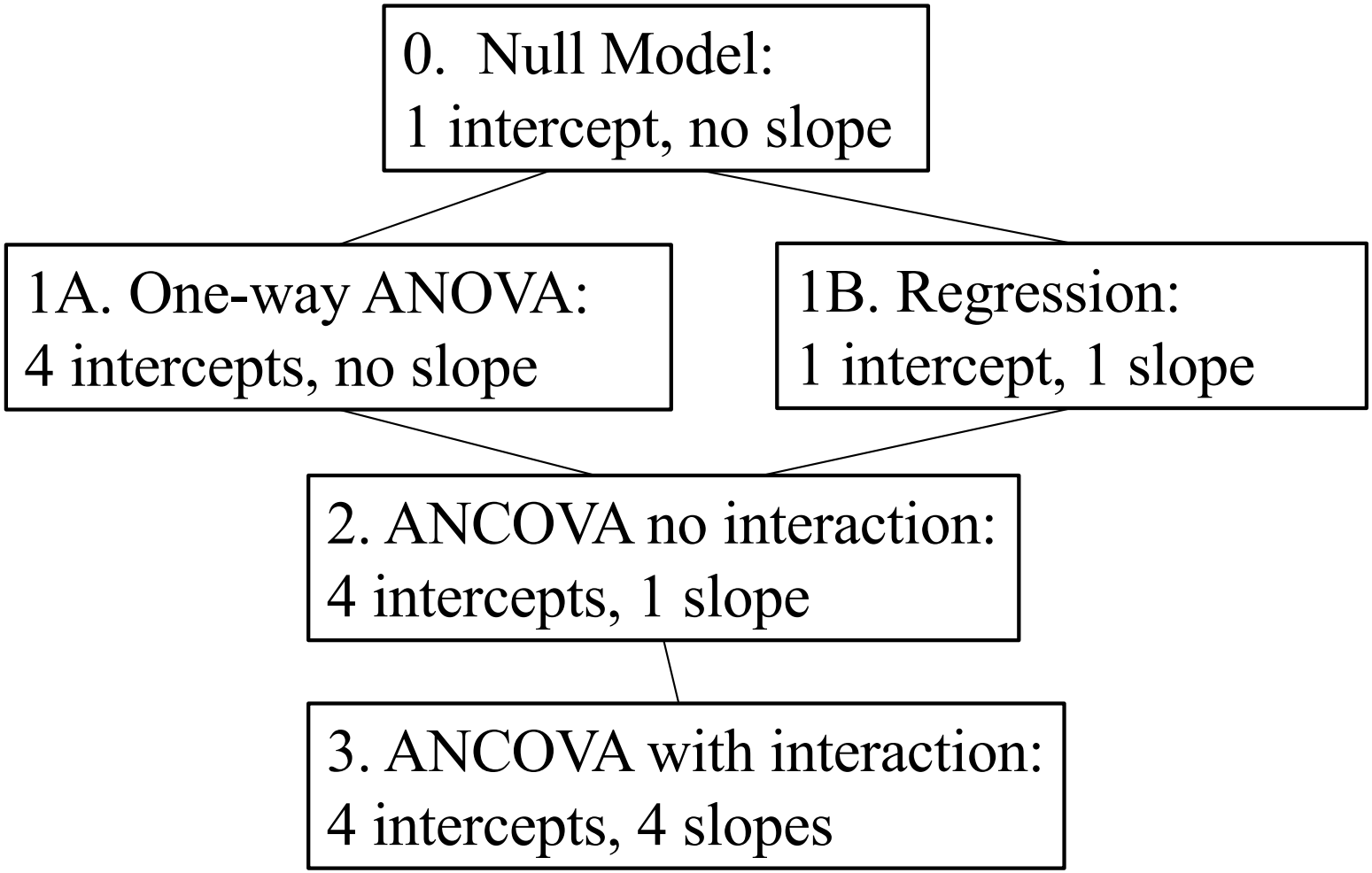
Recall the Glue Strength example: Response is strength. Predictors are thickness (#) and glue (A, B, C, D).

0. Null Model: 1 intercept, no slope.
1A. One-way ANOVA (glue only): 4 intercepts, no slope.
1B. Simple Linear Regression (thick only): 1 intercept, 1 slope.
2. ANCOVA no interaction (glue and thick): 4 intercepts, 1 slope.
3. ANCOVA with interaction (glue, thick, glue:thick): 4 intercepts, 4 slopes.

Since we have a categorical predictor, use F-tests from Type 3 ANOVA table.

ANCOVA Example

Glue Strength Example (X = thickness, Glue=A,B,C,D)



Backward elimination for ANCOVA (General Plan):

1. Start with ANCOVA with interaction (most complicated model considered), look at F-test p-value corresponding to interaction term (highest order term). If significant ($p < 0.05$), use this model.
2. If previous test not significant ($p \geq 0.05$), move to ANCOVA no interaction model. Look at F-test p-values corresponding to both main effects. If both significant ($p < 0.05$), use this model.
3. If at least one of the previous tests is not significant, drop the least significant (highest p-value) term.

Backward elimination for Glue Strength Example:

1. Start with ANCOVA with interaction (most complicated model considered). Looking at F-test p-value corresponding to interaction (glue:thick) term (highest order term), we find glue:thick F = 0.7729, p-value = 0.53.

2. Since previous test not significant ($p \geq 0.05$), move to ANCOVA no interaction model. Looking at p-values corresponding to both main effects, we find:

   glue F = 4.6632, p-value = 0.017

   thick F = 42.6236, p-value < 0.001

   Since both significant ($p < 0.05$), use this model.

Comments on Glue Strength Example:

1. Since the research question was about comparing glues and thickness was not of primary research interest, the regression model with just thick would likely be of no interest and probably not formally considered.

2. Likewise, when the numerical predictor is a covariate, it would be standard to focus on the ANCOVA no interaction model (unless there was some compelling reason to do otherwise).

3. Hence, in practice, the researcher might just consider the ANCOVA no interaction vs one-way ANOVA.

4. The big-picture idea is that you are allowed to use personal/scientific judgement when comparing models.

Example: AIC model selection for Glue Strength (n=20)

We will compare 2 ways to calculate AIC: `extractAIC()` and `dredge()`. Note that `extractAIC()` matches hand calculation.

		extractAIC()			dredge()	
Model	p	SSResid	AIC	Δ AIC*	AIC	Δ AIC*
Thick	2	40.51	18.12	7.18	76.88	7.18
Glue	4	80.53	35.86	24.92	94.61	24.92
Thick + Glue	5	20.96	10.94	0.00	69.70	0.00
Thick + Glue+Think:Glue	8	17.57	13.41	2.47	72.16	2.47

Δ AIC = AIC_i – min AIC. So for this example, comparisons are to the ANCOVA NO interaction model (because that model has smallest AIC).

Discussion of the Glue Strength AIC example:

1. The "best" AIC model (with smallest AIC) is the ANCOVA No interaction model. This matches the result using a hypothesis testing approach.

2. AIC values from `extractAIC()` and `dredge()` do NOT match, but ΔAIC value do! In other words, the conclusions are the same using either function.

2. Both `extractAIC()` and `dredge()` show the number of parameters. Other ways to count parameters include: ANOVA table or summary output.

Comments about `MuMin`:

1. `MuMIn` accepts categorical predictors.
2. It follows the rule of hierarchy. So, it does not consider a model with interaction but without main effects.
3. By default, `MuMin` will rank models by AICc. This is a completely reasonable choice. Other options include AIC, BIC and QAIC. See next slide.
4. The options(na.action = "na.fail") line is required before using `dredge()`. This forces the function to fail if there is even a single missing value in the data set! This allows AIC comparisons with fixed sample size.
5. R^2 can be added to the summary table using the option `extra = c("R^2")`

AICc and SBC/BIC

When the number of parameters is large, or the sample size (n) is small, a "corrected" version of AIC is preferred:

$$\boxed{\mathrm{AIC}_C = \mathrm{AIC} + \frac{2p(p+1)}{n-p-1}}$$

When AICc is computed the value of p is increased by 1 to account for σ^2.

I don't have a "rule of thumb" for "small n" or "large p". See *Model Selection and Inference*, by Burnham and Anderson.

Shwarz Bayesian Information Criteria:

$$\boxed{BIC = SBC = n\ln(SS\mathrm{Resid}/n) + p\ln(n)}$$

Closely related to AIC, but with a different penalty term.

Automated Model Selection (no natural hierarchy)

Example: Highway data, 1973 (Weisberg) 39 segments of Minnesota state highway:

Y	RATE = accident rate (per million veh. miles)
X1	LEN = length of segment (miles)
X2	ADT = av. daily truck traffic (thousands)
X3	TRKS = truck volume (percent of total volume)
X4	SLIM = speed limit (1973, before 55mph limit)
X5	LWID = lane width (feet)
X6	SHLD = width of shoulder (feet)
X7	ITG = nr of fwy type interchanges (per m.)
X8	SIGS = nr of signal interchanges (per mile)
X9	ACPT = number of access points (per mile)
X10	LANE = total nr of traffic lanes (both direct.)
X11	FAI = 1 if Federal aid interstate, 0 otherwise
X12	PA = 1 if principal arterial highway, else 0
X13	MA =1 if minor arterial highway, else 0

Objective: Fit a model that predicts RATE as a function of the other variables. Of particular interest is the SLIM variable, because policy makers want to know if altering speed limits will save lives.

Start by fitting the **full model** : Y vs all 13 other vars.

R-square=0.76 Model F = 6.11 (p=0.0001), Yet none of the individual variables is significant ($p<0.05$) in the full model. Seems surprising!!

Explanation: Some of the predictor variables may be, in fact, poor predictors of RATE. They are adding noise to the fit. But many of the potentially good predictors are correlated among themselves. When predictors are correlated, the individual predictors do not add much to the model, given that the other predictors are in the equation.

We want to consider smaller models, but how do we choose?

6. Forward, backward and stepwise selection

Backwards elimination: Start with the full model. Eliminate terms one at a time based on some criteria. Once a term has been eliminated, it is not reconsidered for re-entry.

Forward selection: Start with the null model. Add terms one at a time based on some criteria. Once a term has been added, it is not considered for elimination.

Stepwise selection: This is a hybrid between forward selection and backward elimination. It starts like forward selection. However, after the second variable has been added, we check to see if the first variable can now be dropped out.

*Forwards/Backwards/Stepwise can select different models.

Traditional stepwise selection is based on hypothesis testing criteria. When adding/dropping terms the following are equivalent:

1. smallest p-value (either t or F-test)
2. largest test statistic (F or abs(t))
3. greatest increase in R^2

Can be done manually using `drop1()` or `add1()` or using `fastbw()` from the `rms` package.

AIC stepwise selection is based on AIC criteria. This can be done using the `step()` function in R.

*Traditional/AIC stepwise can select different models.

Comments on Stepwise selection:

1. In software programs other than R, "traditional" stepwise selection is almost always offered.
2. Stepwise selection methods do not examine all possible models. Good models can be missed by these procedures.
3. Stepwise methods were popular when computing of all possible models was not feasible. They are less popular now
4. Stepwise selection methods are still useful, particularly in cases where there are a large number of predictors and no "a priori" list of candidate models.

7. Best (or All) Subsets Selection

Instead of stepping through some of the possible models, why not just compute all of them, then select the most suitable model based on some criterion?

Highway Example: With k=13 potential predictor variables, there are $2^{13} = 8{,}192$ possible models.

With k=20, there would be $2^{20} = 1{,}048{,}576$ models.

This does not take long to run with modern software/computers!

AIC, AICc and BIC are common choices for ranking models.

Note: R^2 cannot be used to compare models of different sizes. R^2 increases (or stays the same) when you add a term to the model. Do NOT use R^2 for model selection!

Other Criteria for Subsets Selection

$$\text{adjusted } R^2 = 1 - \frac{(n-1)}{(n-p)}(1-R^2)$$

$$R^2 = 1 - \frac{\text{MSResid}}{\text{MSTotal}}$$

1. "Best" model is the one with largest adjusted R^2.
2. For Regression, p = k + 1 = # predictor variables +1
3. The model with the largest adjusted R^2 will be the model with the smallest MSResid. MSTotal is the sample variance of the Y's, which is the same for all models.
4. Adjusted R^2 will increase whenever a term is added with an F-statistic > 1 (i.e. a |t-statistic| > 1)

$$\text{Mallow's } C_p = \frac{\text{SSResid}_{\text{reduced}}}{\text{MSResid}_{\text{full}}} - n + 2p$$

1. "Best" model is the one with <u>smallest</u> C_p.
2. C_p is calculated for a particular model (the reduced model) compared to the model with all the variables (the full model).
3. C_p will go down whenever terms are deleted from the full model that have an F-statistic < 2 (or a |t-statistic| <$\sqrt{2}$).
4. Often gives results very similar to AIC.

PRESS statistic (Prediction Error Sum of Squares):

$$\text{PRESS} = \sum \left(y_i - \hat{y}_{i(-i)} \right)^2 \text{ where } \hat{y}_{i(-i)} \text{ is the prediction using all but the i}^{\text{th}} \text{ point.}$$

"Best" model is the one with smallest PRESS.

Sometimes used to summarize model fit.

Subsets Selection using R:

1. The dredge() function from MuMin allows models to be ranked by AIC, AICc, BIC. See earlier discussion of dredge().

2. The regsubsets() function from leaps allows models to be ranked by BIC, Adjusted R^2 or Cp.

Summary of Highway Selection Example:

A summary of the results from Highway Examples 1 (Stepwise selection) and 2 (Best Subsets Selection).

Method	ACPT	LEN	SLIM	SIGS	PA
p: Forward, Backward	X	X	X		
AIC: Forward, Backward Stepwise	X	X	X	X	X
AIC Best Subsets	X	X	X	X	X
Cp, Adj R2 Best Subsets	X	X	X	X	X

In this case there is good agreement between the methods. In general, the methods do NOT have to agree. That includes forward, backward and stepwise based on the same criteria (ex: AIC).

8. Discussion of model selection methods

- The methods we have discussed (stepwise, AIC, etc) are for model selection. They have no effect on estimation!
- "a priori" list of candidate models and then choosing among those models is a great idea, but does not apply to all situations (ex: large predictive modeling scenarios). Still important to check diagnostic plots. (curve in plots?)
- Forward, backward, and stepwise methods were developed when it was not possible to do all subsets. It is possible that these methods could miss a good model, but they are still commonly used. A current trend is to use an AIC (instead of p-value based) stepwise selection approach.

- Remember: every time a term is added or deleted from the model, it changes the meaning (and estimate and test) of all the other term in the model. These are partial regression coefficients: "Change in Y for a unit change in X1, holding other variables in the model constant" almost always depends on which variables are in the model.
- When the sample size is large, many terms will be significant (due to high power), and you may want to simplify the model, dropping terms that are significant, but have small practical effect or involve variables that are difficult or expensive to collect. The same problem occurs when using AIC criteria. See **Highway Example 3**.

When the goal of model selection is **prediction**:

- The "best" model is the one that makes the best predictions. Testing of individual parameters may be of little or no interest.
- Note that R^2 measures how well the model fits, but using the data that was used to fit the model.
- If you have a large enough sample size, consider dividing the data into training and test sets. This should be done BEFORE starting analysis. The training data is used for model development and test data is used for evaluating performance.
- Cross Validation (CV) is another approach to measuring the predictive performance of a model. Two common approaches are K-fold CV (original sample is randomly partitioned into K subsamples and one is left out in each iteration) and leave-one-out CV (ex: PRESS statistic).
- For the purposes of prediction, model averaging may be of interest.

When the goal of model selection is **explanation**:

- Example: Finding a model that describes a process in order to understand it better.

- It makes sense to be more conservative about including terms than we are for prediction, because we are making claims about whether terms belong in the model.

- Researchers may be more interested in the confidence intervals than the significance of the terms.

- For the purposes of explanation, Akaike weights may be of interest. See the **Cement example**.

- Stepwise methods involve many tests at each step. Each test is also conditional on the result of tests that preceded it. This creates a multiple testing problem that is generally not accounted for in the modeling process.
- The multiple testing problem also exists when a model is selected by one of the selection criteria (ex: AIC).
  See the **Model Selection Simulation** example.
- Possible approaches to dealing with multiple testing problem include (1) reduce the number of models considered or (2) divide data into training and test sets.
- In very large scale model selection scenarios (hundreds or thousands of potential predictors?), consider filtering before model selection. In a regression setting, you could remove predictors that show low correlation with the response. Also consider LASSO and other more advanced methods.

Highway Example 3: "Bigger" data (For Illustration)

We will now explore the effect of sample size on model selection methods by pretending the sample size for the highway data is four times the actual sample size. We want to mimic a data set that has the same properties as the highway data, but is bigger.

Results:

- Top AIC model originally selected 5 predictors (acpt, len, slim, sigs, pa). With larger data 7 predictors (+trks, ma) included.
- For stepwise approaches (both traditional and AIC) terms would be added (or removed) from the model in the same order as before. (Not shown.)

Comparing like models, the parameter estimates are the same, but the p-values are much smaller for the "bigger" data.

	Original Data		Big Data	
	Estimate	p-value	Estimate	p-value
ACPT	0.06428	0.04126	0.06428	<0.001
LEN	-0.0741	0.00484	-0.0741	<0.001
SLIM	-0.1051	0.01585	-0.1051	<0.001
SIGS	0.79736	0.03791	0.79736	<0.001
PA	-0.7744	0.06816	-0.7744	<0.001

Model Selection Simulation (For Illustration)

500 data sets (reps), n = 50 per data set
Y and X1, X2, ….,X10 generated as independent random normals.
Since the data are random, the "true model" is the "null model":

$$Y_i = \beta_0 + \varepsilon_i$$

For each data set, find the model with the lowest AIC and lowest AICC, then count the number of predictors (k). Any thing besides the null model (no predictors) is a "false positive".

Results: More than 80% of the time the true (null) model is <u>not</u> selected. That's bad, right?

Cement Example: Akaike weights

This example is from Burnham and Anderson.

A small set of data (n=13) with four predictor variables (X1, X2, X3, X4) thought to be related to the heat evolved during the hardening (Y) of Portland cement.

Given the small sample size, AICC is appropriate here.

We consider all possible models (2^4-1 =15 models) and calculate AICC for each.

The model with X1 and X2 has the smallest AICC value.

We can now calculate the difference in AICC for each model i and the model with the lowest AICC:

$$\Delta_i = \text{AICC}_i - \text{AICC}_{min}$$

The Akaike weights are then calculated as:

$$w_i = \frac{\exp(-0.5\Delta_i)}{\sum \exp(-0.5\Delta_i)}$$

This value provides a relative weight of evidence for each model.

The weights also provide a way to estimate the relative importance of a predictor variable. This measure of relative importance can be calculated as the sum of the Akaike weights over all of the models in which the parameter of interest appears.

For the Cement data, we find that the relative support for the 4 variables is X1 (99%), X2 (81%), X3 (21%) and X4 (32%).

Another use of the Akaike weights is model averaging.

Another approach to calculating weights is to bootstrap the data and then tally the percentage of times that each variable occurred in the AIC selected model.

Cement Example Write-up:

Analysis was done using R. A multiple regression model was fit to the data with hardening as the response. Predictor variables considered for inclusion in the model were X1, X2, X3, and X4 (obviously you would use informative names in a real write-up!). Model selection was performed using AICc based all subsets selection with the MuMIn package (REF). Residual diagnostic plots were used to assess model assumptions (only if true).

After model selection, the predictors included in the model are X1 and X2 ($R^2 = 0.98$).

Could also provide estimated values (not shown here).