Model selection:
  Given various possible models, which are appropriate?
  Graphical illustration: fit polynomial models to data generated by a cubic
  Almost always do not want to use all possible variables
    Overfit current data set, get worse predictions for new data
    Why: complex models fit the errors, not the true means
  Have already seen one way to answer this: hypothesis tests (T or F)
    Requirement: models are nested. Reduced is simpler than full
    Philosophical issue: null model treated differently from alternative
  Various alternative approaches:
    Stepwise variable selection, minimum MSE, maximum R^2: all have problems
    Generally recommended approach, AIC (or closely related AICc or BIC)

AIC model selection: widely used, well behaved
  Concept: find well-fitting model that's not too complex
    AIC = lack of fit + complexity,
    For standard assumptions (equal variance, normality): AIC = n log(SSE) + 2 k
    fit: n log(SSE), penalty for complexity: 2 k
  Want a model with small AIC (or more negative AIC)
    Only useful comparatively. AIC can be -500 or 20 or 3000. Don't care.
    Best model = smallest AIC.
  AICc: Same as AIC but with a correction for small sample sizes
    When given the opportunity, use AICc instead of AIC
    Software may only provide AIC
  BIC: same concept as AIC, larger penalty for complexity - depends on sample size
    BIC = n log(SSE) + (log n)k
  Both AIC and BIC can be used with other assumptions about data
    E.g., yes/no observations with logistic regression
    In general: AIC = -2 log likelihood + 2k, BIC = -2 log likelihood + (log n)k
  Only useful comparatively. AIC can be -500 or 20 or 3000. Don't care.

Comparing models:
  Basic advice: choose model with smallest AIC/AICc/BIC
  Better advice:
    Look at AIC values for multiple models
    Do any models have AIC values close to that of the best?
  General recommendations:
    within 2 of the best are reasonable alternatives
    more than 10 from the best is not reasonable

Numerical example: 101 observations
  Truth: cubic polynomial
  Consider linear, quadratic, $\cdots$ 10'th degree polynomial

	AIC		AICc		BIC	
Model	value	Δ	value	Δ	value	Δ
cubic	-139.5	0.00	-138.2	0.00	-129.8	0.00
4th	-138.5	1.04	-136.6	1.62	-126.9	2.97
5th	-137.8	1.67	-135.2	2.94	-124.3	5.53
10th	-134.6	4.94	-126.4	11.81	-111.4	18.46
quadratic	-101.0	38.47	-100.2	38.01	-93.3	36.54
linear	-91.4	48.08	-90.9	47.25	-85.6	44.21

Advice about strategy:
Easy to fit models with all possible combinations of variables
Multiple linear regression, considering all subsets of 30 variables takes 2 seconds with a good algorithm
Try not to if at all possible
use subject knowledge to choose small subset of models
e.g., fitting polynomials, don't consider models like
$y = \beta_0 + \beta_2 x^2$ or $y = \beta_0 + \beta_2 x^2 + \beta_5 x^5$
Only consider the sequence of increasing degree:
$y = \beta_0 + \beta_1 x$, $y = \beta_0 + \beta_1 x + \beta_2 x^2$, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$, $\cdots$

Practical advice and pitfalls:
Must use exactly same observations (Y) for all models
Possible problems:
1) Can't compare a model with Y to a model with log Y
Different observations
Can compare regression models with different variables
or X in one model and log X in another
2) Watch out for dropped observations because of missing X values
If X2 missing for some observations,
Y = b0 + b1 X1 + b2 X2 fit to different observations than Y = b0 + b1 X1
Different software can give different AIC values
Different functions in same software can give different AIC values
Usual culprit: AIC actually = n log(SSE) + 2 k + constant
Can compare AIC or BIC so long as the same constant used for all models
Constants don't depend on the model
Different software (or functions) often use different constants
Can't compare values! User beware, unless only using same function or software

How precise are predictions?
Assume your MLR goal is to predict new observations
Common (but bad) practice:
Fit a model to data,
SSE quantifies how well model fits these data
rMSE (approx. = se predicted obs) quantifies uncertainty in predictions

Does not tell you how well model predicts new observations
The problem is that you're using the same data twice
Once to fit the model; again to assess the precision
rMSE too small

Estimating precision for new predictions
Training / test set methodology
divide data set into two parts
training: used to develop the model, often 80% of obs
test set: assess quality of predictions on these obs (the other 20%)
look at bias: systematically wrong predictions
and precision: rMSEP, root Mean Square Error of Prediction
$= \sqrt{\sum (Y_j - \hat{Y}_j)^2 / n_{test}}$, where j is each obs. in test set
or overall accuracy: MAPE, mean absolute prediction error
MAPE $= \sum \mid Y_j - \hat{Y}_j \mid / n_{test}$
Cross-validation: out-of-sample error using all observations
Divide data into chunks (e.g., each 10% of data set)
Remove chunk one, fit model to remaining 90%
assess quality on the left-out 10%
Put back chunk 1, remove chunk 2, fit/assess
Continue for all chunks
Chunk is often 1/10'th = 10-fold cross-validation
or 1/5'th = 5-fold cross-validation
Leave-one-out cross-validation = loo
Each chunk is a single observation
Predict Y_i from all observations **except** Y_i
Requires N fits, but often can be done very quickly (matrix algebra tricks)
PRESS statistic, Prediction Residual Error Sums-of-Squares
loo idea, quantifying overall accuracy predicting new observations
PRESS $= \Sigma_{obs}(Y_i - \hat{Y}_{-i})^2$
$\hat{Y}_{-i}$ is prediction of Y_i from model fit without Y_i
Almost always larger than SSE $= \Sigma_{obs}(Y_i - \hat{Y}_i)^2$
Because PRESS prediction of Y_i not based on Y_i
Training / Validation / Test
Variation on Training / Test approach, but with 3 groups of observations
Used when modeling approach requires choosing tuning parameters
that control the algorithm (e.g., whether to use AIC, AICc, or BIC)
Training data used to find best model given each choice of tuning parameter
Validation data used to chose the best algorithm
Test data used at the very end to calculate prediction accuracy

Uses of model selection:
Prediction: what set of variables $\rightarrow$ good predictions?
use AIC or BIC

with training / test or cross-validation to assess
Choosing variables to adjust / control for in an observational study
Want to control for important variables
Best: use subject-specific info. to choose the important variables
When no subject-specific info, use model selection on possibly useful covariates
Leave out the variable of interest (e.g., sex in Case 12.2, bank salary)
Usually AIC to choose a good model (a few more X's less bad than too few X's)
Add variable of interest back to the best covariate model
Evaluate sensitivity to choice of covariate model
by adding sex to 2 or 3 good covariate models
Think carefully about the potentially important covariates
If you omit an important variable, it's ignored and your conclusions may be biased

Uses that require lots of careful thought:
Identifying important causal variables
Goal: what would be the "effect" of increasing a focus X
Example: expend (per student expenditure) in the SAT case study
Method I: Use all variables for model selection
Example: expend is in "the best" model to predict SAT scores
Bad logic: selected variables are biologically important
and omitted variables are biologically irrelevant
Claim that increasing expenditure on public schools will increase SAT scores
WRONG for two reasons:
causal inference from an observational study
model selection may select a correlated variable, not the true one
If expend is correlated with some other X in the data set
Model sel. will sometimes pick expend, sometimes pick correlated X
Method II: Use control/adjust for logic described above
Omit expend, do model sel. on all other variables
Add expend to selected model (or multiple "close" models)
Deals with correlated variables in the data set
Can not deal with unmeasured variables correlated with expend

How large a data set do you need?
Depends on how many variables you want to consider
Can do model selection with 100 variables and 20 observations
DON'T. Almost certainly overfitting specifics of this data set
General guideline: 6-10 observations per potential variable
more than 6-10 is even better!
SAT: 7 variables (with both takers and log takers): 49 observations, fine
bank salary: 14 variables, 93 observations. ok (just)
tractor sales: 10 obs, 8 variables, NO
(even though the company asked the undergraduate intern to do the analysis)

Things to keep in mind:

Multiple strident opinions about model selection
  "Data dredging is strongly discouraged and can result in spurious
    (and irrelevant or worse, wrong) results and inference."
My response: this is a reaction to bad interpretations of model selection results
  not something inherently wrong about model selection
Never, never, forget your subject-matter knowledge or intuition
  Including subject-matter knowledge is more likely to produce a useful model
One highly-recommended strategy
  Identify a small number (5?, 10?) models based on subject-matter knowledge
  Use model selection on this set, not all subsets
Remember that model selection starts with a "full" model:
  With variables that enter linearly (usually) and additively (almost always)
  Reality may be non-linear, include interactions, or depend on omitted covariates
    Can add polynomial terms and interactions to the "full" model
    But now have many, many more X variables, remember 6-10 obs per variable
If the goal is prediction, always use out-of-sample error, not in-sample

Alternatives to model selection
  Model averaging:
    Instead of making conclusions from one selected model
    Combine information from multiple models
    Increasingly popular, for very good reasons
  Combining variable selection with estimation
    Methods that allow a parameter estimate to be 0 or non-zero
      without searching all subsets
    LASSO and elastic net: two very useful methods to do this
  Letting the data specify the form of the regression model
    Classification and regression trees (CART)
      Allows arbitrary forms of interaction
    Random Forests
      Extension of CART - averages many imprecise predictions
      My current choice for a "black-box" prediction engine
  All of these require lots of data for successful use