START Mon 3/28/22 (week 10, lecture 26)

w/ rcode bridgeselect.R

until 45 min mark

* Also talked about k-fold CV.

Stat 608 Chapter 7

Variable Selection

STARTED monday 3/21/22 (week 9, lecture 23)

* starts w/ Rcode related to ch. 6: statcsp.R, circulation.R & bridge.R (Diagnostic stuff for multiple Reg)

START Wednesday 3/23/22 (week 11, lecture 24)

Introduction

- Overspecified model (or contains irrelevant predictors):
  - MSE: fewer degrees of freedom.
  - Standard errors for regression coefficients inflated.
  - Thus: larger p-values and wider confidence intervals.

$$t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$$

will be smaller

→ p-value larger.

- Underspecified model (too few predictors):
  - Regression coefficients and thus predictions are biased.
  - Arguably worse than overspecified model.

Introduction

Problems with multicollinearity:

- Even when the model is significant, it’s possible that no individual predictors are significant.
- Slopes may have the wrong sign.
- Predictors that explain substantial variation in y may be insignificant.

Example: Bridge data

Bridge model, predicting log(Time):

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.28590	0.61926	3.691	0.000681	***
log(DArea)	-0.04564	0.12675	-0.360	0.720705	
log(CCost)	0.19609	0.14445	1.358	0.182426	
log(Dwgs)	0.85879	0.22362	3.840	0.000440	***
log(Length)	-0.03844	0.15487	-0.248	0.805296	
log(Spans)	0.23119	0.14068	1.643	0.108349	

correlation btwn DArea & Length w/ log(time) is positive, but coef estimates are neg ⇒ multicollinearity.

Confidence interval: deck area

```
log(DArea)   -0.04564    0.12675  -0.360 0.720705
qt(0.975, 39)
[1] 2.022691
```

CI: $\hat{\beta}_1 \pm t_{.975,39} * SE(\hat{\beta}_1)$

$= -0.04564 \pm 2.022691 \times 0.12675$

$= (-0.302, 0.211)$

- I am 95% confident that a 1% increase DArea is associated w/ a on average 30.2% decrease or 21.1% increase in time, holding all other variables constant.

Hypothesis test: deck area

```
log(DArea)  -0.04564    0.12675  -0.360 0.720705
```

$H_0: \beta_1 = 0$, $H_a: \beta_1 \neq 0$

$$t = \frac{-0.04564 - 0}{0.12675} = -0.360$$

p-value = 0.720705

$\Rightarrow$ fail to reject.

Introduction

- Goal: Choose the best model using variable selection methods.
- Start by considering the full model containing all m potential predictor variables:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + e$$

- Variable selection methods choose the subset of predictors that is “best”.
- **Overfitting**: including too many predictors (model performs as well or worse than simpler models at predicting new data)
- **Underfitting**: including too few predictors (model doesn’t perform as well as models with more predictors)

Introduction

– explanatory.

- If the goal is interpretation, simpler models are usually preferred. Use a method that chooses fewer models.

- If the goal is prediction, more variables may be acceptable.

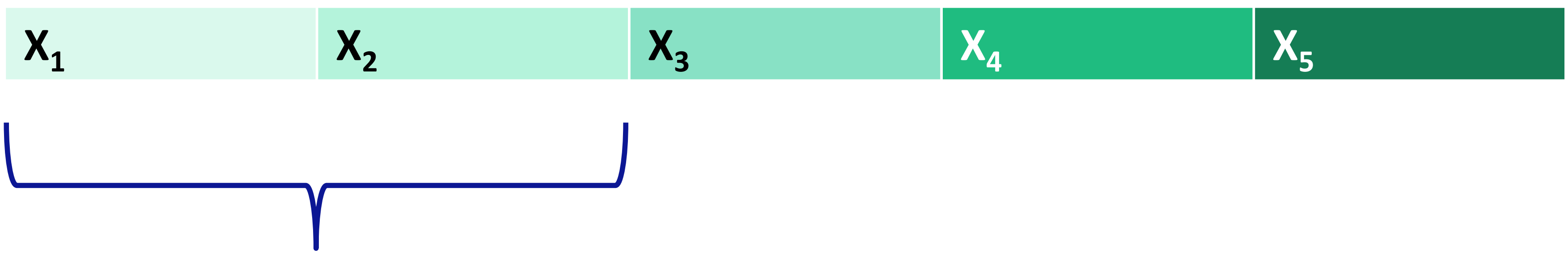
Forward, Backward, and Stepwise Subsets

- If there are m variables, there are 2^m possible regression equations. If m is small enough, run all of them (all possible subsets).
- Backward, Forward, and Stepwise selection procedures examine only *some* of the 2^m possible regression equations.
- **Backward elimination**:
  1. All variables are included in the model. The predictor with the largest p-value is deleted (as long as it isn't significant).
  2. The remaining m-1 variables are now in the model. Again, the predictor with the largest p-value is deleted (as long as it isn't significant).
  3. Variables are deleted until all remaining variables are significant.

+ Backward selection

α=0.05

Model 1: Full model



P-value < α

Backward selection

α=0.05

Model 2: Full model minus variable with largest p-value

X_1	X_2	X_3	X_4	X_5
X_1	X_2	X_3	X_4	

P-value < α

Backward selection

α=0.05

X_1	X_2	X_3	X_4	X_5
X_1	X_2	X_3	X_4	
	X_2	X_3	X_4	

P-value < α

Model 3: Final model: x_2, x_3, x_4.

Forward, Backward, and Stepwise Subsets

- **Forward selection**:
  1. No variables are in the model. All m models with only one predictor are run. The predictor with the smallest p-value is entered in the model (as long as it is significant). Call this variable x_1.
  2. All models with predictors x_1 and only one other predictor are run; of the remaining predictors $x_2, \ldots, x_m$, the one with the smallest p-value is entered (as long as it is significant).
  3. Variables are entered until no more predictors are significant, given the others already in the model.

Forward selection

α=0.05

Step 1: Enter the first variable

X_1	Model 1
X_2	Model 2
X_3	Model 3: Smallest p-value
X_4	Model 4
X_5	Model 5

Forward selection

α=0.05

Step 2: Enter the second variable

X_3,	**X_1**	Model 1: X_1 has smallest p-value
X_3,	**X_2**	Model 2
X_3,	**X_4**	Model 3
X_3,	**X_5**	Model 4

Forward selection

α=0.05

Step 3: No more variables are significant.

X_3, X_1, X_2 Model 1

X_3, X_1, X_4 Model 2

X_3, X_1, X_5 Model 3

Stepwise Subsets

- **Stepwise Selection Procedure:**
  1. Choose α_E and α_R, significance levels to Enter and Remove predictors.
  2. Forward step: No variables are in the model. All models with one predictor are run. The predictor with the smallest p-value is entered into the model, as long as the p-value is less than α_E. Call this variable x_1.
  3. Forward step: All models with predictors x_1 and only one other predictor are run; of the remaining predictors $x_2, \ldots, x_p$, the one with the smallest p-value is entered, as long as the p-value is less than α_E.
  4. Backward step: Check to see that the p-value for variable x_1 is smaller than α_R. If not, remove it. If so, leave it in.
  5. Take another forward step, attempting to add a third variable.
  6. Continue taking backward and forward steps until adding an additional predictor does not yield a p-value below α_E.

Could α_E be larger than α_R? Vice versa? $\alpha_E \leq \alpha_R$

Stepwise is a forward selection procedure, except that a variable can be removed once it is in.

Forward, Backward, and Stepwise Subsets

- These procedures only consider some of the predictors, so they do not necessarily find the model that fits the data the best among all possible subsets.
- Forward, backward, and stepwise may not produce the same final model, though they often do.
- If covariance of the predictors = 0, all three produce the same final model.
- These methods are prone to overfitting, but stiff criteria for adding or deleting variables can mitigate this problem.
- Shouldn't we just remove the insignificant terms all at once?

  - Chapter 5: F-Test for model reduction
  - Chapter 7: Algorithms (not hypothesis tests)

+

Selection Criteria: (1) R^2 -Adjusted

- Adding irrelevant predictor variables to the regression model often increases R^2.

- To compensate, we adjust for the number of predictors:

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

- Choose the subset of the predictors that has the highest value of R^2_{adj}. This is equivalent to choosing the subset of the predictors with the lowest value of MSE (mean square error).

Selection Criteria: (2) AIC (Akaike's Information Criterion)

- Based on maximum likelihood estimation
- R uses the calculation:

$$AIC = n \log\left(\frac{RSS}{n}\right) + 2p$$

- Choose the model which makes AIC as small as possible. (By small, we mean close to -∞).
- Only meant to compare sub-models to one another or to the full model, not models with different transformations.

Selection Criteria: (3) AIC_C (AIC Corrected)

- Corrects for bias when n small or p large compared to n. (AIC tends to overfit; the penalty for model complexity is not strong enough.)
- Converges to AIC as n increases.

$$AIC_C = AIC + \frac{2(p+2)(p+3)}{n-p-3}$$

- Choose the model which makes AIC_C as small as possible.
- IMPORTANT NOTE: The formula above is correct; the textbook is incorrect on page 231. See www.stat.tamu.edu/~sheather/book/docs/Errata.pdf.

STOP Wed 3/23/22 (Week 4, lecture 24)

START Friday 3/25/22 (week 9, lecture 25)

Selection Criteria: (4) BIC (Bayesian Information Criterion, aka SBC)

- Based on posterior probability of model, but often used in a frequentist sense.

$$BIC = n \log\left(\frac{RSS}{n}\right) + (p+2)\log(n)$$

- Choose the model which makes BIC as small as possible.
- BIC is similar to AIC except with 2p replaced by p log(n). When n ≥ 8, log(n) ≥ 2, so the penalty term for BIC is larger than the penalty term for AIC. BIC favors simpler models than AIC. *

Selection Criteria: (5) Mallows' C_p

- Uses unbiasedness as a criterion for choosing a model; assumes the full model is unbiased.

$$C_p = \frac{RSS_p}{MSE_{full}} - n + 2p$$

- Choose a model whose C_p value is close to the # of parameters. in the model counting the intercept. (Err on the side of a smaller value of C_p.)

- Don't use C_p to choose the full model; C_p always equals p in that case.

- If the full model contains a large number of insignificant variables, MSE_{full} will be inflated (MSE involves the df). Then C_p is not an appropriate model for choosing the best model.

Comparison of Selection Criteria

• Typically use BIC

- Using p-values tends toward extreme over-fitting. (After doing 3 hypothesis tests, overall alpha increases from 0.05 to about 0.1…)
- R^2_{adj} and C_p tend toward over-fitting. *
- C_p is equivalent to AIC for linear models with normal errors. *
- AIC chooses models too complex when n is large. BIC chooses models too simple when n is small. $n \approx 10$ *
- Pro of AIC and AIC_C: They are "efficient." Asymptotically, the error in prediction from the model using AIC and AIC_C is no different from the error from the best model. Not true of BIC.
- Pro of BIC: The probability it selects the correct model is asymptotically 1. *
  Not true of AIC.
  as $n \to \infty$

Comparison of Selection Procedures

- All possible subsets:
  - If the number of predictors in the model is of fixed size p, all four criteria R^2_{adj}, AIC, AIC_C, and BIC choose the same model.
  - When comparing models with different numbers of predictors, we can get different answers.

- Forward, Backward, and Stepwise:
  - Using other information criteria (AIC, BIC) to select a model is equivalent to using p-values to add and remove variables; the difference is where the algorithm stops.

Reminders

- The regression coefficients obtained after variable selection are ✗ biased.
- P-values from these models are generally much smaller than their true values.
- Software treats each column of the design matrix as being completely separate, ignoring relationships in polynomial models and models with interaction terms. Package ‘glmulti’ considers only models with main effects corresponding to their included interaction terms.

Bridge Data

Subset Size	Predictors	R2adj	AIC	AICC	BIC
1	log(Dwgs)	0.702	-94.90	-94.31	-91.28
2	log(Dwgs), log(Spans)	0.753	-102.37	**-101.37**	**-96.95**
3	log(Dwgs), log(Spans), log(Ccost)	**0.758**	**-102.41**	-100.87	-95.19
4	log(Dwgs), log(Spans), log(Ccost), log(Darea)	0.753	-100.64	-98.43	-91.61
5	log(Dwgs), log(Spans), log(Ccost), log(Darea), log(Length)	0.748	-98.71	-95.68	-87.87

LASSO

- LASSO: Least Absolute Shrinkage and Selection Operator, performs variable selection and parameter estimation simultaneously.

- Constrained Least Squares:

$$\min \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_{1i} + \ldots \beta_p x_{pi}])^2, \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq s$$

for some number s non-negative.

- When some variables have larger scales (standard deviations, say), they appear more important to this method; standardize (z-scores) or normalize (transform to [0,1] scale) to mitigate this effect.

- We can again use a version of AIC, BIC, or C(p) to choose the best LASSO model.

- When s is very large, this is equivalent to the usual least squares estimates for the model.

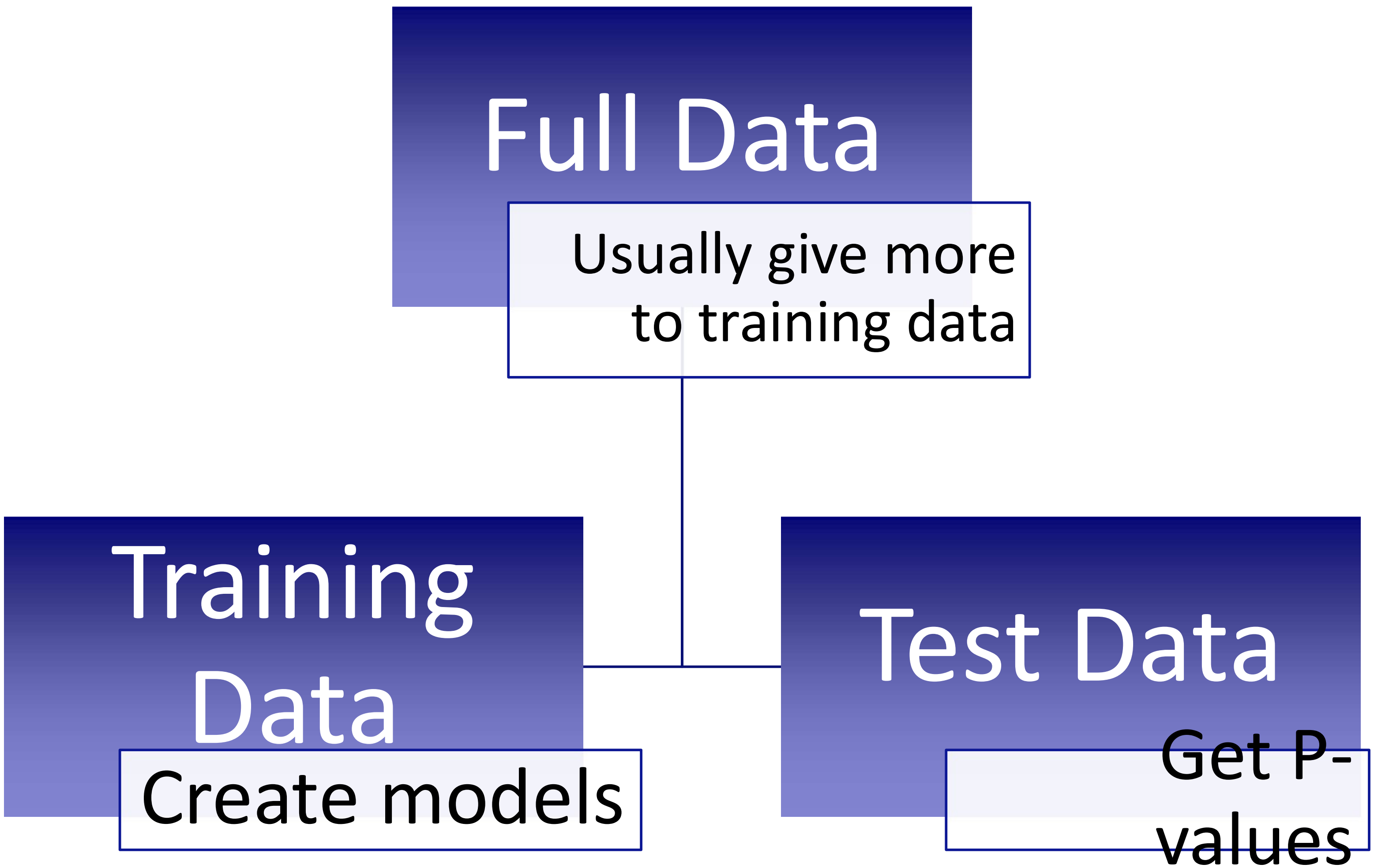
- When s is small, some of the coefficients are 0, effectively removing them from the model.

Assessing the Predictive Ability of Regression Models

- Since regression coefficients are biased and p-values are generally much smaller than their true values, we need another approach:

- Split the data, and see how well models built on one part predict the other part not being used to build the model.

Assessing the Predictive Ability of Regression Models



Assessing the Predictive Ability of Regression Models

- Ideally, the training and test data sets will be similar with respect to:
  - Univariate distributions of each of the predictors and response
  - Multivariate distributions of all variables
  - Means, variances, other moments
  - Outliers

- Usually, splitting the data is done randomly. However, especially in small data sets, the above criteria are not always met.

@28min mark

STOP Friday 3/25/22 (Week 9, lecture 25)

- finish friday w/ R example bridge select, [illegible]