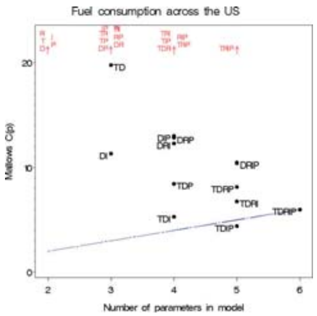
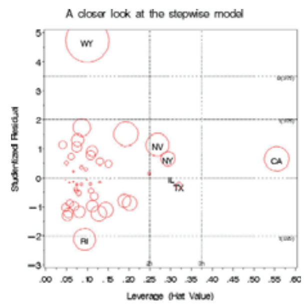


H_0 true (no bias)

$C_p \approx p$

$F_p \approx 1$



Model selection in regression

Michael Friendly
Psychology 6140

Selecting the "best" model

There are often:

- many variables to choose
- many models: subtly different configurations
- different costs
- different power
- not an unequivocal "best"

More realistic goal:

Select a "most-satisficing" model – gets you where you want to go, at reasonable cost

Box: "All models are wrong, but some are useful"

Selecting the "best" model

Criteria for model selection

- Sometimes quantifiable
- Sometimes subjective
- Sometimes biased by pre-conceived ideas
- Sometimes pre-conceived ideas are truly important
- How well do they apply in future samples?

Model selection: the task of selecting a (mathematical) model from a set of **potential** models, given **evidence** and some **goal**.

Regression: Opposing criteria

- Good fit, good in-sample prediction
  - Make R^2 large or MSE small
  - → Include many variables
- Parsimony:
  - Keep cost of data collection low, interpretation simple, standard errors small
  - → Include few variables

Statistical goals

- Descriptive/exploratory
  - Describe relations between response & predictors
  - → want precision (+ parsimony ?)
- Scientific explanation
  - Test hypothesis, possibly ‘causal’ relations
  - → Control/adjust for background variables
  - → Want precise tests for hypothesized predictors
- Prediction/selection
  - How well will my model predict/select in *future* samples?
  - → Cross-validation methods
- Data mining
  - Sometimes we have a huge # of possible predictors
  - Don’t care about explanation
  - Happy with a small % “lift” in prediction

Model selection criteria

- $R^2 = SSR_{model} / SS_{Total}$
  - Cannot decrease as more variables added
  - → look at ΔR^2 as new variables added
- Adjusted R^2 attempts to adjust for # predictors.

$$Adj\ R^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)$$

- This is on the right track, but antiquated (Wherry, 1931)

Model selection criteria: C_p

- Mallow’s C_p: measure of ‘total error of prediction’ using *p* parameters
  - est. of $$\frac{1}{\sigma^2}\sum\left(\underbrace{\text{var}(\hat{y})}_{\text{random error}} + \underbrace{(\hat{y}_{true} - \hat{y}_p)^2}_{\text{bias}}\right)$$
  - $C_p = (SSE_p / MSE_{all}) - (n-2p)$
  - Related to AIC and other measures favoring model parsimony

Model selection criteria: C_p

- Relation to incremental F test:

  $C_p = p + (m+1-p)\,(F_p - 1)$

  F_p = incremental F for omitted predictors, testing H_0: $\beta_{p+1} = \ldots = \beta_m = 0$ when there are *m* available predictors.

  p = # parameters, including intercept

H_0 true (no bias)	H_0 false (bias)
$C_p \approx p$	$C_p > p$
$F_p \approx 1$	$F_p > 1$

A “good” model should therefore have $Cp \approx p$

Model selection criteria: Parsimony

- Attempt to balance goodness of fit vs. # predictors
- Akaike Information Criterion (AIC)

$$AIC = \underbrace{n\ln\left(\frac{SSE}{n}\right)}_{\text{error}} + \underbrace{2p}_{\text{penalty}}$$

- Bayesian Information Criterion (BIC)

$$BIC = \underbrace{n\ln\left(\frac{SSE}{n}\right)}_{\text{error}} + \underbrace{2(p+2)q - 2q^2}_{\text{penalty}} \quad \text{where } q = \frac{n\hat{\sigma}^2}{SSE}$$

- AIC & BIC
  - Smaller = Better
  - Model comparison statistics, not test statistics– no *p*-values
  - Applicable to *all* statistical model comparisons– logistic regression, FA, mixed models, etc.

Scientific explanation

- Need to include variable(s) whose effect you are testing
  - Does gasoline price affect consumption?
  - Does physical fitness decrease with age?
- Need to include *control variable(s)* that could affect the outcome
  - Omitted control variables can bias other estimates
  - E.g., per capita income might affect consumption
  - Weight might affect physical fitness
- Better to risk some reduced precision than bias by including more variables, even if *p*-values NS

Descriptive/Exploratory

- Generally only include variables with strong statistical support (low *p* values). Choose models with highest adjusted R^2 or lowest AIC)
  - Parsimony particularly valuable for making in-sample predictions
    - High precision
    - Fewer variables to measure
- Models with AIC close to best model are also supported by the data
  - If you need to choose just one, pick the simplest in this group
  - Better to report alternatives, perhaps in a footnote
- Examine whether statistically significant relationships have effects **sizes** & **signs** that are meaningful
  - Units of regression coefficients: units of Y/units of X

Example: US Fuel consumption

```
pop       Population (1000s)
tax       Motor fuel tax (cents/gal.)
nlic      Number licensed drivers (1000s)
inc       Per Capita Personal income ($)
road      Length Federal Highways (mi.)
drivers   Proportion licensed drivers
fuel      Fuel consumption (/person)
```

state	pop	tax	nlic	inc	road	drivers	fuel
AL	3510	7.0	1801	3333	6594	0.513	554
AR	1978	7.5	1081	3357	4121	0.547	628
AZ	1945	7.0	1173	4300	3635	0.603	632
CA	20468	7.0	12130	5002	9794	0.593	524
CO	2357	7.0	1475	4449	4639	0.626	587
CT	3082	10.0	1760	5342	1333	0.571	457
DE	565	8.0	340	4983	602	0.602	540
FL	7259	8.0	4084	4188	5975	0.563	574
...	...	...					

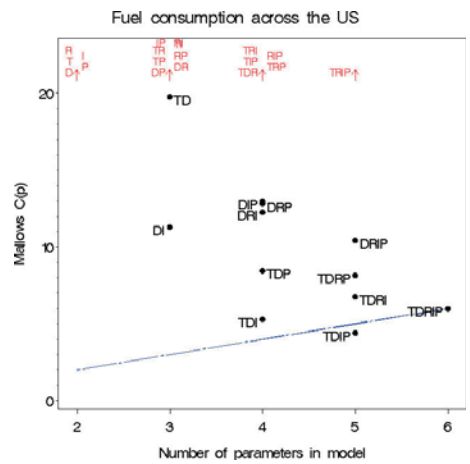
```
%include data(fuel);
proc reg data=fuel;
  id state;
  model fuel = pop tax inc road drivers /
          selection = rsquare cp aic best=4;  run;
```

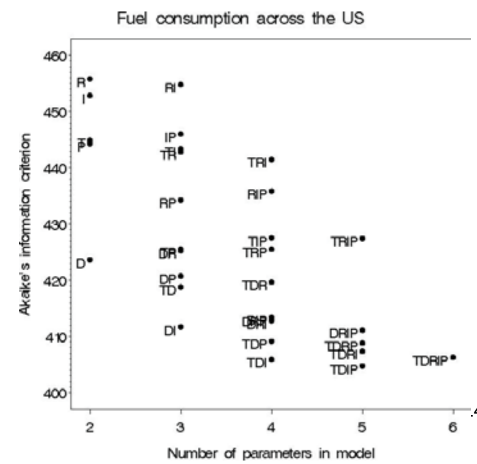
Number in Model	R-Square	C(p)	AIC	Variables in Model
1	0.4886	27.2658	423.6829	drivers
1	0.2141	65.5021	444.3002	pop
1	0.2037	66.9641	444.9368	tax
1	0.0600	86.9869	452.8996	inc
2	0.6175	11.2968	411.7369	inc drivers
2	0.5567	19.7727	418.8210	tax drivers
2	0.5382	22.3532	420.7854	pop drivers
2	0.4926	28.6951	425.2970	road drivers
3	0.6749	5.3057	405.9397	tax inc drivers
3	0.6522	8.4600	409.1703	pop tax drivers
3	0.6249	12.2636	412.7973	inc road drivers
3	0.6209	12.8280	413.3129	pop road drivers
4	0.6956	4.4172	404.7775	pop tax inc drivers
4	0.6787	6.7723	407.3712	tax inc road drivers
4	0.6687	8.1598	408.8362	pop tax road drivers
4	0.6524	10.4390	411.1495	pop inc road drivers
5	0.6986	6.0000	406.3030	pop tax inc road drivers

NB: C_p always = p for model with all predictors

cpplot macro

```
%cpplot(data=fuel,
      yvar=fuel, xvar=tax drivers road inc pop,
      gplot=CP AIC,
      plotchar=T D R I P, cpmax=20);
```





Variable selection methods

- All possible regressions (or best subsets)
  - proc reg; model … / selection=rsquare best=;
  - R: leaps package: **`regsubsets()`**
  - 2^p -1 : p=10 → 1023 models!
  - Useful overview, but beware of:
    - Effects of collinearity
    - Influential observations (*n*: small, moderate)
    - Lurking variables: unmeasured, but important
  - → Use R^2, C_p, AIC to select **candidate** models, to be explored in more detail, not for final selection

Variable selection methods

- Forward selection
  - proc reg; model … / selection=forward SLentry=.10;
  - At each step, find the variable X_k with the **largest** *partial* F_k value

$$F_k = \frac{MSR(X_k \mid \text{others})}{MSE(X_k + \text{others})}$$

  - If Pr(Fk)<SLentry: add to model; else STOP
  - Result depends on SLentry (liberal default)

Stepwise Selection: Step 1

Statistics for Entry

Variable	Tolerance	Model R-Square	F Value	Pr > F
tax	1.000000	0.2037	11.76	0.0013
drivers	1.000000	0.4886	43.94	<.0001
road	1.000000	0.0004	0.02	0.8978
inc	1.000000	0.0600	2.93	0.0935
pop	1.000000	0.2141	12.54	0.0009

Variable **drivers** Entered: R-Square = 0.4886 and C(p) = 27.2658

Stepwise Selection: Step 2

Statistics for Entry

Variable	Tolerance	Model R-Square	F Value	Pr > F
tax	0.917035	0.5567	6.92	0.0117
road	0.995887	0.4926	0.36	0.5497
inc	0.975329	0.6175	15.17	0.0003
pop	0.866451	0.5382	4.83	0.0331

Variable **inc** Entered: R-Square = 0.6175 and C(p) = 11.2968

Variable selection methods

- Backward elimination
  - proc reg; model … / selection=backward SLstay=.10;
  - Start with all variables in the model
  - At each step, find the X_k with the **smallest** *partial* F_k value
  - If Pr(Fk)>SLstay: remove from model; else STOP
  - Result depends on SLstay (liberal default)

Variable selection methods

- Stepwise regression
  - proc reg; model … / selection=stepwise SLentry=.10 SLstay=.10;
  - Start with 2 forward selection steps
  - Then alternate:
    - Forward step: Add X_k w/ highest F_k if Pr(Fk)<SLentry
    - Backward step: Del X_k w/ lowest F_k if Pr(Fk)>SLstay
    - Until: no variables entered or removed

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square
1	drivers		Proportion licensed drivers	1	0.4886
2	inc		Per Capita Personal income ($)	2	0.1290
3	tax		Motor fuel tax (cents/gal.)	3	0.0573
4	pop		Population (1000s)	4	0.0207

Summary of Stepwise Selection

Step	Model R-Square	C(p)	F Value	Pr > F
1	0.4886	27.2658	43.94	<.0001
2	0.6175	11.2968	15.17	0.0003
3	0.6749	5.3057	7.76	0.0078
4	0.6956	4.4172	2.93	0.0942

But:

- Does the model make sense?
- Have all important control variables been included?

Variable selection in R

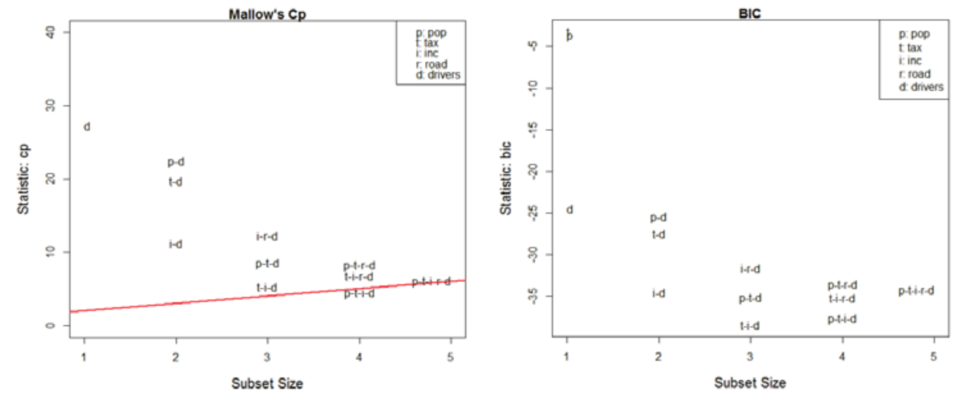
leaps package: **regsubsets()** does a variety of selection methods

```
library(leaps)
fuel.subsets <-
         regsubsets(fuel ~ pop + tax + inc + road + drivers,
                    data = fuel,
                    nbest = 3,         # best 3 models for each number of predictors
                    nvmax = NULL,      # no limit on number of variables
                    force.in = NULL,   # variables to force in
                    force.out = NULL,  # exclude from all models
                    method = "exhaustive")  #or, "forward", "backward", ...
fuel.subsets
```

```
Subset selection object
Call: regsubsets.formula(fuel ~ pop + tax + inc + road + drivers, data = fuel,
    nbest = 3, nvmax = NULL, force.in = NULL, force.out = NULL,
    method = "exhaustive")
5 Variables  (and intercept)
        Forced in Forced out
pop         FALSE      FALSE
tax         FALSE      FALSE
inc         FALSE      FALSE
road        FALSE      FALSE
drivers     FALSE      FALSE
3 subsets of each size up to 5
Selection Algorithm: exhaustive
```


car package: **subsets()** plots model selection criteria

```
subsets(fuel.subsets, statistic="cp", main="Mallow's Cp",
          legend="topright", ylim=c(0,40), cex.lab=1.25)
abline(a=1, b=1, col="red", lwd=2)

subsets(fuel.subsets, statistic="bic", main="BIC",
          legend="topright", cex.lab=1.25)
```



MASS package: **stepAIC()** does forward/backward/stepwise using AIC or BIC

```
library(MASS)
fuel.mod <- lm(fuel ~ pop + tax + inc + road + drivers,
                    data = fuel)
final.mod <- stepAIC(fuel.mod)
```

```
Start:  AIC=406.3
fuel ~ pop + tax + inc + road + drivers

          Df Sum of Sq    RSS    AIC
- road     1      1762 179106 404.78
<none>                 177344 406.30
- pop      1     11706 189050 407.37
- inc      1     17565 194909 408.84
- tax      1     27188 204533 411.15
- drivers  1    110017 287361 427.47

Step:  AIC=404.78
fuel ~ pop + tax + inc + drivers

          Df Sum of Sq    RSS    AIC
<none>                 179106 404.78
- pop      1     12197 191302 405.94
- inc      1     25516 204621 409.17
- tax      1     44659 223765 413.46
- drivers  1    121015 300121 427.56
```

In **stepAIC()** :

- `k = 2` defines penalty factor for AIC; use `k=log(n)` for BIC
- `scope = list(lower=~1, upper=…)` defines the scope of models considered

Dangers of ‘blind’ stepwise methods

- Gives R^2 values that are badly biased high
  - Substantial shrinkage in a future sample
- *F*, *t* (or χ^2) statistics for each variable don’t have the claimed distribution:
  - *p*-values are wrong, because they don’t take selection into account
- Confidence intervals for effects and predicted values are overly narrow
  - Based on one model, not selection from many
- Problems of collinearity: why X4, not X7?
  - Tiny difference in data might select X7

Dangers of ‘blind’ stepwise methods

- Based on methods (e.g. *F* tests for nested models) intended to test **pre-specified** hypotheses.
- Allows us to not **think** about the problem.
- Generates a lot of output, but most people just look at the final summary.
- *“Treat all claims based on stepwise algorithms as if they were made by Saddam Hussein on a bad day with a headache having a friendly chat with George Bush.”*
  (From: Stepwise regression = Voodo Regression web page)

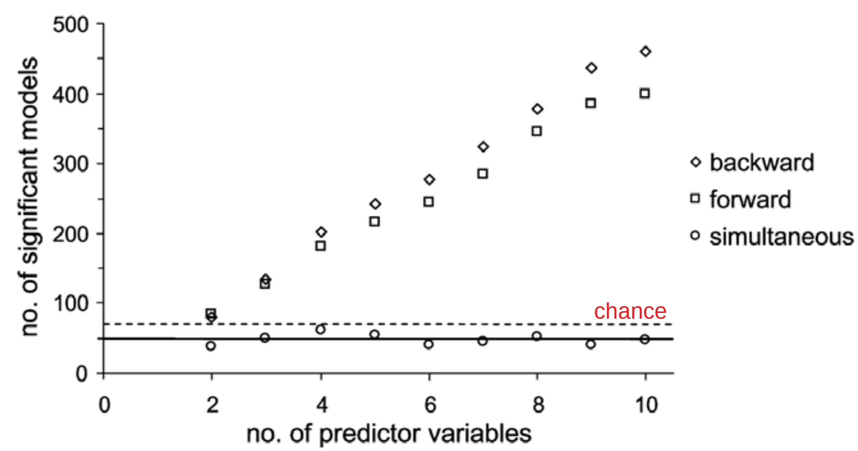
Stepwise dangers: demo

```
title 'Stepwise demo: add 15 random predictors to Fitness data';
%include data(fitnessd);

data fitness;
   set fitness;
   array x{15} x1-x15;
   *-- generate the 'artificial' predictors;
   do i=1 to 15;
      x(i) = normal(7654321);
      end;

proc reg data=fitness;
   model oxy=runtime age weight runpulse maxpulse rstpulse
           x1-x15 / selection=rsquare mse best=4 stop=6;
   run;
```


Number in Model	R-Square	MSE	Variables in Model
1	0.7434	7.53384	runtime
1	0.1616	24.61437	x15
1	0.1595	24.67582	rstpulse
1	0.1584	24.70817	runpulse
2	0.7771	6.77903	runtime x10
2	0.7753	6.83234	runtime x5
2	0.7650	7.14558	runtime x4
2	0.7642	7.16842	runtime age
3	0.8111	5.95669	runtime age runpulse
3	0.8100	5.99157	runtime runpulse maxpulse
3	0.8070	6.08587	runtime age x5
3	0.7986	6.35153	runtime x5 x10
4	0.8662	4.38138	runtime age runpulse x5
4	0.8399	5.24242	runtime age maxpulse x5
4	0.8368	5.34346	runtime age runpulse maxpulse
4	0.8321	5.49727	runtime runpulse maxpulse x9
5	0.8795	4.10420	runtime age runpulse x5 x13
5	0.8780	4.15627	runtime age runpulse x5 x7
5	0.8767	4.19971	runtime age runpulse maxpulse x5
5	0.8742	4.28573	runtime age runpulse x5 x11
6	0.8927	3.80744	runtime age runpulse x1 x5 x13
6	0.8888	3.94390	runtime age runpulse x5 x7 x13
6	0.8868	4.01488	runtime age runpulse maxpulse x5 x11
6	0.8861	4.04208	runtime age runpulse x5 x11 x13

Note how often random predictors occur among ‘best’ models!

Numbers of significant multiple regressions (out of 1,000) based on random data for which the null hypothesis is, by definition, true.



From: **Roger Mundry & Charles L. Nunn, “Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution.” *The American Naturalist,* Vol. 173, No. 1 (January 2009), pp. 119-123.**

Guided selection : One way out

Divide predictors into three subsets:

- **Key variables**: ought to be included in any model
  - Necessary control variables
  - → force inclusion
- **Promising variables**: deserve special attention
  - → examine all subsets | key variables (C_p, R^2, BIC)
  - Do the good ones make sense?

```
proc reg;
  model Y = Age IQ Test1-Test5 /
     include=2 /* force Age, IQ */
     selection=rsquare CP AIC
     stop = 4;
```

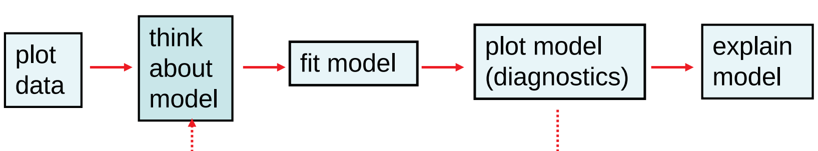

Guided selection

- **The haystack**: motley collection that remains
  - → stepwise (haystack | key + selected promising)
  - Any here worth including?

```
proc reg;
  model Y = Age IQ Test2 Test5 X1-X15 /
     include=4
     selection=stepwise
     SLentry = .15 SLstay=.15;
```


Model building & selection

- The single most important 'tool' is substantive knowledge of the area and properties of variables
  - Expected sign & magnitude of coefficients?
  - Necessary control variables?
  - What hasn't been measured?
- Never let a computer do your thinking for you.

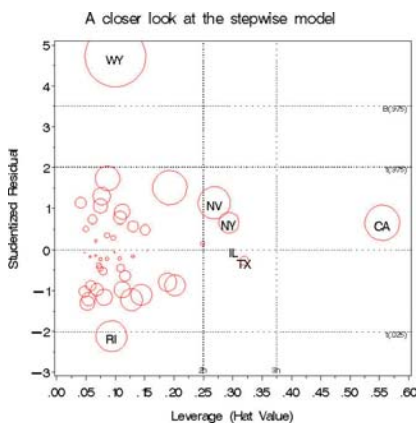


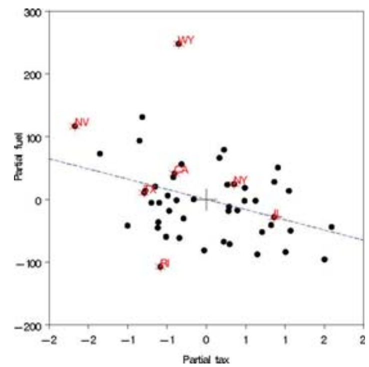
Are we done yet? Model diagnostics

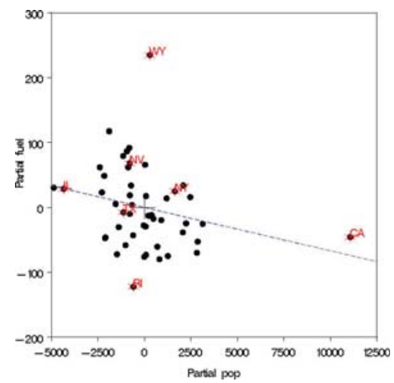
- Examine model diagnostics for selected models
  - (Just a preview; explained more next week)
  - Influential observations?
  - Partial relations? Outliers?

```
proc reg data=fuel;
   id state;
   model fuel = tax drivers inc pop/ r influence partial;
   plot r. * p. = state
        rstudent. * h. = state ;
   title 'A closer look at the stepwise model';
run;
%inflplot(data=fuel,
   y=fuel,   x=tax drivers road inc pop,
   id=state);
%partial(data=fuel,
   yvar=fuel, xvar=tax drivers road inc pop,
   id=state);
```

- Some consistently unusual observations (WY, CA, RI)
- What is going on here?
- Could there be a variable we have missed?







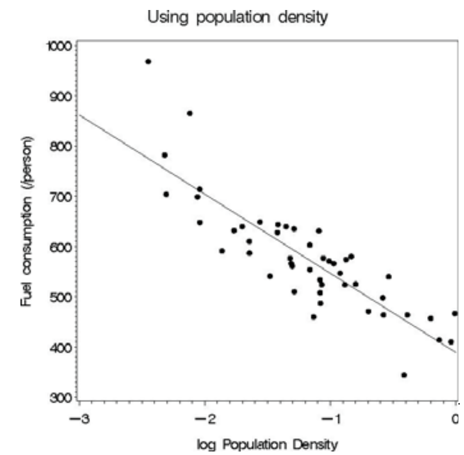

```
title 'Using population density';
data fuel;
   set fuel;
   popden = log10 ( pop / area);
proc gplot;
   plot fuel * popden  / vaxis=axis1;
   symbol i=rl v=dot c=black;
run;
proc reg;
   model fuel = popden tax drivers road inc
        / selection = stepwise;
```

Model with *just* popden is better than the previous stepwise model!

Do we need anything else?

Summary of Stepwise Selection

Step	Variable Entered	Number Vars In	Partial R-Square
1	popden	1	0.7270
2	drivers	2	0.0666
3	tax	3	0.0106



Cross-validation & shrinkage

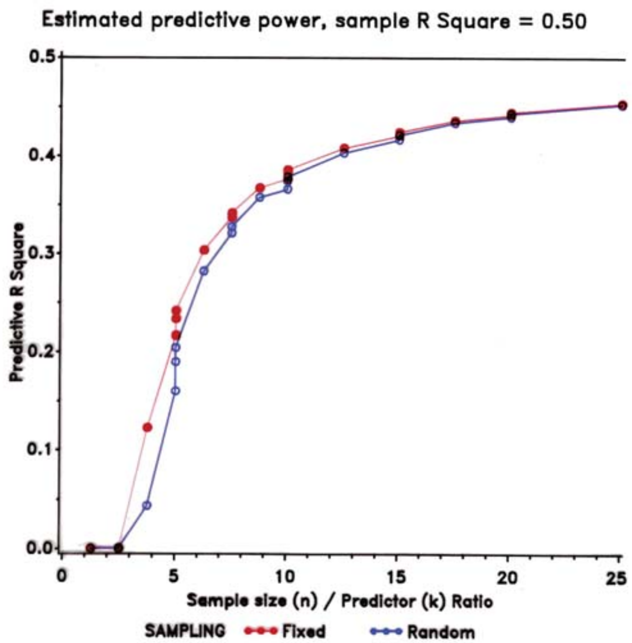
- Fit model in prediction sample [y_{ps} | X_{ps}]
  - Get R^2, $\mathbf{b}_{ps}$
- Apply coefficients $\mathbf{b}_{ps}$ to validation sample [y_{vs} | X_{vs}]
- Cross-validated R^2 : using $\mathbf{b}_{ps}$ in new sample

$$R^2_{cv} = r^2(y_{vs}, \hat{y}_{pv}) \quad \text{where } \hat{y}_{pv} = X_{vs} b_{ps}$$

- How much we can expect to lose depends on n/k (k=# predictors)
- Recall goals: most important in prediction; gives realistic assessment for scientific explanation & data mining

TABLE 3.10
Estimated Predictive Power Using the Herzberg Formulas for Small to Fairly Large Subject/Variable Ratios

Subject/Variable Ratio	*Herzberg Estimate*	*Comment*	n/k
Small (5:1) $n=50$, $k=10$ $R^2 = .50$ ②	X fixed: $\hat{\rho}_c^2 = 1 - (n-1/n-k-1)(n+k+1/n)(1-R^2)$ $= 1 - 49/39\ (61/50)\ (.50)$ $= .234$ ① X random: *Correlation Model* $\hat{\rho}_c^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)(1-R^2)$ $= 1 - 49/39\ (48/38)\ (51/50)\ (.5)$ $= .191$	The estimated amount of shrinkage is great, i.e., on the average we expect the predictive power to be reduced by over 50%.	5
Moderate (10:1) $n=100$, $k=10$ $R^2 = .50$	$\hat{\rho}_c^2 = 1 - 99/89\ (111/100)\ (.50) = .383$ *Correlation Model* $\hat{\rho}_c^2 = 1 - 99/89\ (98/88)\ (101/100)\ (.5)$ $= .374$	The shrinkage is still fairly substantial. 25%	10
Fairly Large (15:1) $n=150$, $k=10$ $R^2 = .50$	$\hat{\rho}_c^2 = 1 - 149/139\ (161/150)\ (.50)$ $= .427$ *Correlation Model* $\hat{\rho}_c^2 = 1 - 149/139\ (148/138)\ (151/150)\ (.5)$ $= .421$	We finally reach a point where the expected amount of shrinkage is fairly small, i.e., about 12%.	15



Empirical cross-validation

- Ideal: do CV via replication, but if not…
  - Hold back a portion of the data as the CV sample
  - Fit model to prediction (“training”) subset
  - Evaluate R^2 in hold-back (“validation”) subset
  - You “waste” some data, but gain in prediction knowledge
- Can do ‘manually’ with any software via coding tricks
- Generalized CV methods:
  - Do this several times for different subsets & average
    - K-fold CV: Repeat { omit 1/K; validate on omitted 1/K}
  - There are now a wide variety of methods and algorithms
    - Jackknife, bootstrap, lasso, …
    - Modern methods use these for model selection!

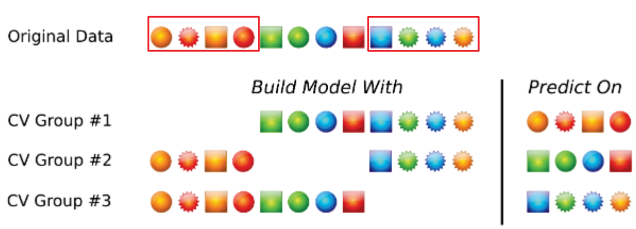
K–Fold Cross–Validation

Here, we randomly split the data into K distinct blocks of roughly equal size.

1. We leave out the first block of data and fit a model.
2. This model is used to predict the held-out block
3. We continue this process until we've predicted all K held–out blocks

The final performance is based on the hold-out predictions

K is usually taken to be 5 or 10 and leave one out cross–validation has each sample as a block



Example: Simple cross-validation

```
%include data(fitness);
title      'Cross-validation of a regression model';
/*
o Hold back a portion of the data from the original fit.
o Evaluate how well the model does on the cross-validation sample.
*/
 data fit2;
   set fitness;
   if uniform(125741) < .667   /* generate model on 2/3 of data */
      then oxy2 = oxy;
      else oxy2 = . ;          /* generate prediction on CV sample */

proc reg data=fit2;
   title2 'Model generation (2/3) sample';
   model oxy2 = age weight runtime runpulse / p;
   output out=stats p=predict r=resid;

proc corr data=stats nosimple;
   where (oxy2 = .);        /* select CV 1/3 sample */
   var predict oxy;         /* correlate y, yhat    */
   title 'Cross-validation (1/3) sample';
```

Prediction sample:

```
Dependent Variable: OXY2
                                   Analysis of Variance

                            Sum of         Mean
      Source         DF    Squares       Square      F Value        Prob>F

      Model           4  689.58501    172.39625       24.336        0.0001
      Error          19  134.59670      7.08404
      C Total        23  824.18171

          Root MSE       2.66159     R-square       0.8367
          Dep Mean      47.57392     Adj R-sq       0.8023
```

$R^2 = 0.84$

Validation sample:

```
Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 7
                                         PREDICT                OXY

        PREDICT                          1.00000             0.62949
        Predicted Value of OXY2           0.0                 0.1298

        OXY                              0.62949             1.00000
                                          0.1298              0.0
```

$R^2 = 0.63^2 = 0.40$

This simply demonstrates shrinkage of R^2. In practice, we could average over all folds to get cross-validated estimates of coefficients

PROC GLMSELECT

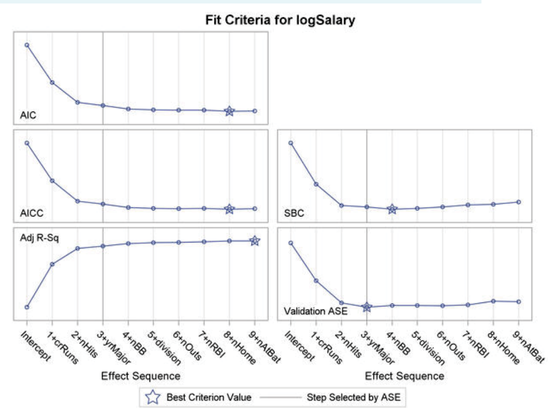
- supports all regression & ANOVA models
- partition data: training, validation & testing roles
- selection from large # of effects, variety of criteria
  - Forward, backward, stepwise, least angle regression, lasso
- leave-one-out and *k*-fold cross validation
- Extensive graphics via ODS Graphics
- In R:
  - DAAG package: **cv.lm()**
  - bootstrap package: **crossval()**, …
  - caret package: extensive facilities for "training", model selection and model averaging

GLMSELECT Example: Baseball data

```
ods graphics on;
   proc glmselect data=baseball plots=(CriterionPanel ASE) seed=1;
    partition fraction(validate=0.3 test=0.2);
    class league division;
    model logSalary = nAtBat nHits nHome nRuns nRBI nBB
                   yrMajor crAtBat crHits crHome crRuns crRbi
                   crBB league division nOuts nAssts nError
                 / selection=forward(choose=validate stop=10);
  run;
```

These use 50% for training, 30% for validation, and 20% for testing

- Validation ASE chooses smaller model.
- Regard these as candidate models



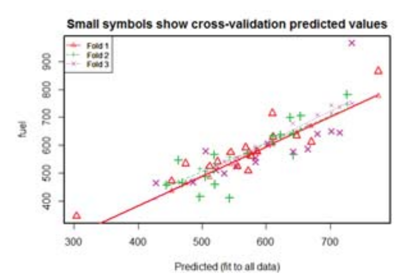
DAAG::cv.lm

```
library(DAAG)
fuel.cv <- cv.lm(data=fuel, final.mod, m=3)
```

m=3 folds of 16 in each test set

```
Analysis of Variance Table

Response: fuel
          Df Sum Sq Mean Sq F value  Pr(>F)
pop        1 125996  125996   30.25 1.9e-06 ***
tax        1 162056  162056   38.91 1.6e-07 ***
inc        1    194     194    0.05    0.83
drivers    1 121015  121015   29.05 2.8e-06 ***
Residuals 43 179106    4165
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cross validated tests for model effects



Summary

- Opposing criteria: goodness of fit vs. parsimony
  - Penalized measures (C_p, AIC, BIC) better
- Different goals
  - Description, explanation, prediction, data mining
  - Require different views of a “good” / “best” model
- Selection methods are tools, not gospel truth
  - Dangers of “blind” stepwise methods
  - Guided selection puts **you** in the modeling process
- Criticize & validate
  - Regression diagnostics to find/correct problems
  - Cross-validation to check replicability