HUDM 6026
Computational Statistics

Model Selection

Linear Model Selection

- Ordinary least squares (OLS) regression works well in many real-world applications. In OLS, we fit a linear model of the form

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

- by finding estimates for the betas that minimize the sum of squared errors (SSE):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \sum_{k=1}^{p} \hat{\beta}_k x_{ki})^2 = SSE$$

- The solution to the problem can be expressed by the normal equations:

$$\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

- where the design matrix **X** is of size *n* by (*p* + 1) because the first column is all ones for the intercept.

Linear Model Selection

- In some cases, though, there may be good reasons to use other fitting rules than the normal equations. Why?
  - If the sample size *n* is small compared to the number of predictors the least squares fit will have high variability.
  - If the sample size *n* is smaller than the number of predictors, the normal equations do not have a unique solution. For example, 20 variables and 20 cases yields the following output:

```
ALL 20 residuals are 0: no residual degrees of freedom!
Coefficients: (1 not defined because of singularities)
 Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.08494         NA       NA        NA
X1          -0.00754         NA       NA        NA
X2           0.03858         NA       NA        NA
X3           0.19541         NA       NA        NA
...            ...
```

  - If there are a lot of predictors (i.e., *p* is large), some of them may be *non-informative*. That is, they may not be related to the outcome variable. In that case, OLS will estimate small coefficients for the variables, but will not set the coefficients to be exactly zero.
  - Thus, we might seek out a method that sets the magnitude of some coefficients to exactly zero. This would result in *feature selection* or *variable selection*.

Three Alternatives to OLS for Linear Models

1. **Subset selection.** Instead of fitting OLS on all possible predictor variables, we first eliminate some by setting their coefficients to zero and then fit OLS on the remaining predictors.
2. **Regularization/shrinkage.** The model is fit on all *p* predictors, but the magnitudes of the coefficients are shrunk towards zero. These methods work by adding a penalty to the loss function based on the magnitude of the regression coefficients. One method of regularization, called *the lasso*, uses SSE plus a regularization term as the loss function:

SSE: usual error term for OLS regression

$$\left(Y_i - \hat{\beta}_0 - \sum_{k=1}^{p} \hat{\beta}_k x_{ki}\right)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|$$

Shrinkage penalty for the lasso.

3. **Dimension reduction.** Here we project the *p* predictors down into an *M*-dimensional space, where $M < p$, via linear combinations of variables. We then use the smaller set of *M* projections as the predictors to be fit by OLS.

Subset Selection

- *Best subset selection* involves fitting the OLS model for *every possible combination* of the *p* predictors and picking the one that is best.
- This is a *brute force* method where we must fit all possible models. Assume *p* predictors. The number of possible combinations (order of the predictors doesn't matter) are
- 0 variables $\binom{p}{0} = \frac{p}{0!(p)!}$
- 1 variable: $\binom{p}{1} = \frac{p}{1!(p-1)!}$
- 2 variables: $\binom{p}{2} = \frac{p}{2!(p-2)!}$
- 3 variables: $\binom{p}{3} = \frac{p}{3!(p-3)!}$

$\vdots$

$\vdots$

- p – 1 variables: $\binom{p}{p-1} = \frac{p}{1!(p-1)!}$
- p variables: $\binom{p}{p} = \frac{p}{0!(p)!}$

The sum of all these combinations is 2^p.

Best Subset Selection

- Algorithm for best subset selection:

1. Let M_0 represent the *null model* (i.e., the model with no predictors). This model will only estimate a constant intercept which will represent the sample mean.
2. For $k = 1, 2, \ldots, p$:
    a) Fit all p choose k models that contain exactly k predictors.
    b) Select the best-fitting model among all the p choose k models, as measured by SSE, R^2, or some other measure of model fit, and call it M_k.
3. Pick the single best model from $M_0, \ldots, M_\mathrm{p}$ using cross-validation prediction error, or some other measure of model fit such as the AIC or BIC.

Generate Some Data for an Example

```
### Generate data with 10 covariates and an outcome that
### is related to only five of them linearly
library(mvtnorm)
library(clusterGeneration)
set.seed(1790)
### Generate a random covariance matrix with package clusterGeneration
cov1 <- genPositiveDefMat(dim = 10, covMethod = "eigen")
### Generate a random vector of 20 means from norm(0, 10)
mns1 <- rnorm(10, 0, 10)
### Generate coefficients for the output for Y from norm(0, 1)
coef1 <- rnorm(11, 0, 1/4) # First one is the intercept
### Set ten of them equal to zero
coef1[sample(2:11, 5, replace = FALSE)] <- 0

datGen <- function(N) {
  ### Generate the X matrix
  X <- rmvnorm(n = N, mean = mns1, sigma = cov1$Sigma)
  ### Add column of 1s for the intercept
  X_aug <- cbind(1, X)
  ### Create the output Y
  Y <- X_aug %*% coef1 + rnorm(N, 0, 1)
  dfOut <- cbind(X,Y)
  dfOut
}
```

Best Subset Selection

True coefficient values.

```
summary(lm(Y ~ X))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.086839   1.810540  -0.600 0.549842       -0.25
X1          -0.020904   0.048126  -0.434 0.665084        0.00
X2           0.063468   0.043645   1.454 0.149413      0.08 -
X3          -0.073515   0.043913  -1.674 0.097621 .      0.03
X4           0.009611   0.033618   0.286 0.775629        0.00
X5           0.252679   0.046435   5.442 4.61e-07 ***    0.31
X6          -0.011763   0.082320  -0.143 0.886700      0.00 -
X7          -0.262061   0.070673  -3.708 0.000363 ***  0.27 -
X8          -0.120030   0.065162  -1.842 0.068802 .      0.16
X9          -0.051375   0.043551  -1.180 0.241288        0.00
X10          0.002091   0.045053   0.046 0.963087        0.00

Residual standard error: 1.044 on 89 degrees of freedom
Multiple R-squared:  0.417,   Adjusted R-squared:  0.3515
F-statistic: 6.367 on 10 and 89 DF,  p-value: 2.598e-07
```

The outcome variable *Y* was generated as a linear combination of the *X* variables plus some random normal error.

Five of the *X* variables were assigned to have a coefficient of exactly 0. That is, they are uninformative, noisy variables.

The uninformative variables (i.e., those with no linear relationship with the outcome *Y*) are colored in red.

Best Subset Selection

CHOSEN WITH BIC:

```
    (Intercept)    X1    X2    X3    X4    X5    X6    X7    X8    X9   X10 logLikelihood      BIC
0          TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE   -25.4208998 50.84180
1          TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE   -18.1910797 40.98733
2          TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE    -6.4467824 22.10391
3          TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE    -3.4853247 20.78616
4*         TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE    -1.0415533 20.50379
5          TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE     0.5539902 21.91787
6          TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE     1.2836776 25.06367
7          TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE     1.5073631 29.22147
8          TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE     1.5469240 33.74751
9          TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE     1.5584188 38.32969
10         TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE     1.5596289 42.93244
```

CHOSEN WITH AIC:

```
    (Intercept)    X1    X2    X3    X4    X5    X6    X7    X8    X9   X10 logLikelihood       AIC
0          TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE   -25.4208998 50.841800
1          TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE   -18.1910797 38.382159
2          TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE    -6.4467824 16.893565
3          TRUE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE    -3.4853247 12.970649
4          TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE    -1.0415533 10.083107
5*         TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE     0.5539902  8.892020
6          TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE     1.2836776  9.432645
7          TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE     1.5073631 10.985274
8          TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE     1.5469240 12.906152
9          TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE     1.5584188 14.883162
10         TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE     1.5596289 16.880742
```

Best Subset Selection

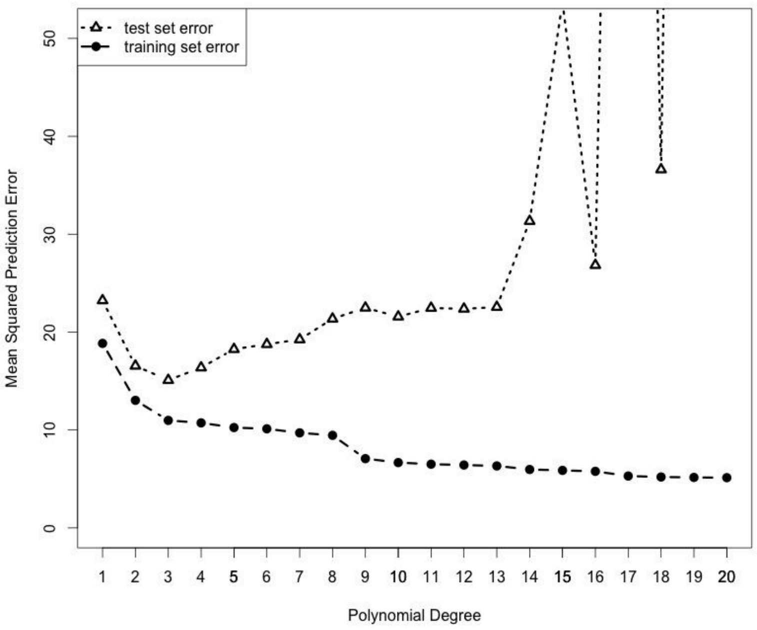
- Best subset selection with the BIC dropped all five non-informative variables but failed to retain X_2, one of the true linear predictors (albeit with the smallest coefficients.
- Best subset selection with the AIC dropped all five non-informative variables and retained all five true linear predictors.
- Can also choose based on cross-validated prediction accuracy via leave-one-out or K-fold.
- What are AIC and BIC?
  - AIC is Akaike’s Informaion Criterion
  - BIC is Bayesian Information Criterion

- The AIC and BIC are measures of relative fit of statistical models.
- Why not simply use mean squared prediction error (MSPE)?

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

AIC and BIC

- As we have seen, the MSPE (mean squared prediction error) measured on the training set is an underestimate of the test set MSPE.



- In particular, the training set MSPE will go down so long as more variables/flexibility are included in the model; however, the test set MSPE will not.

AIC and BIC

- The AIC and BIC are both based on -2 times the maximized value of the likelihood (-2*LL*). *k* is the number of parameters and *LL* is the log-likelihood value at the MLE.

$$AIC = 2 * k - 2LL \quad \text{and} \quad BIC = \ln(n) * k - 2LL$$

Here there is a penalty for magnitude of the *LL*; smaller is better

- The likelihood (assuming normally distributed errors) for multiple linear regression is

$$\left(\frac{1}{\sqrt{2ps}}\right)^n \exp\left(-\frac{1}{2s^2}|Y - Xb|^2\right)$$

Here there is a penalty for more complexity in the model.

- The log-likelihood is

$$-\frac{n}{2}\log 2p - n\log\hat{s} - \frac{1}{2\hat{s}^2}|Y - Xb|^2$$

This is the residual sum of squares

AIC and BIC

- As you see from their definitions, AIC and BIC are similarly constructed, with the essential difference being the term multiplied by *k*, the number of parameters estimated.
- As a result (2 vs. log(n)), the BIC tends to penalize model complexity more heavily than the AIC.
- There is debate about which information criteria (there are others aside from AIC and BIC) are "best" and under what circumstances.
- Because the AIC is somewhat more permissive of model complexity than the BIC, it may be preferred for *prediction*.
- When the intent is to create a model for *explanation*, BIC may be preferred because it will tend to produce a simpler (i.e., more easily interpretable) model.

Best Subset and Computational Efficiency

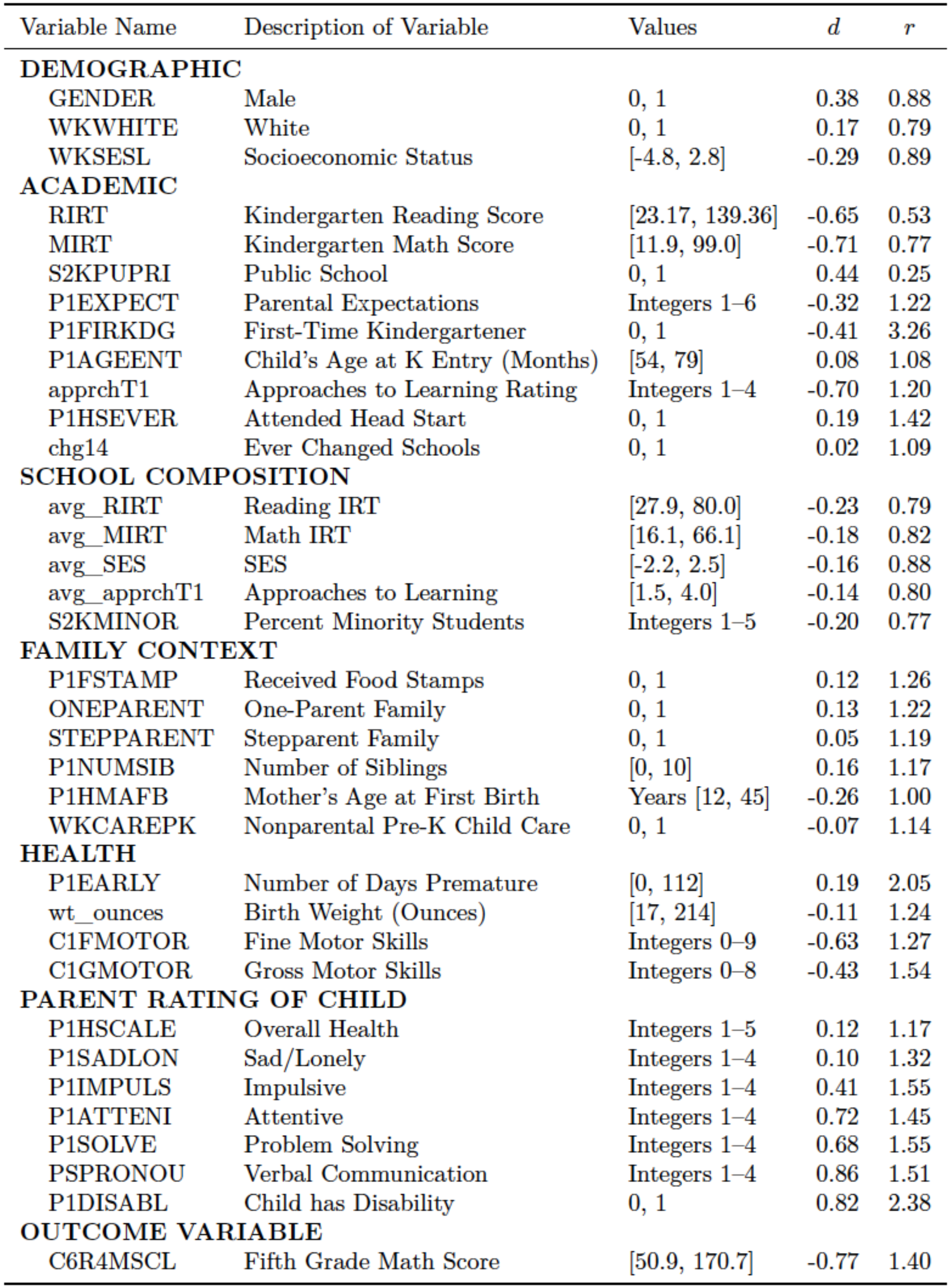
- Best subset selection is computationally demanding.
- For p = 30, for example, best subset will require fitting 2^{30} = 1,073,741,824 models.
- Whereas, forward selection (described next) will require fitting 1 + 1 + 2 + 3 + … + 29 + 30 = $p(p + 1)/2 + 1$ = 451 models.
- That said, the ease of fitting Gaussian linear models make even large (i.e., 30+ covariates) problems tractable within a few seconds. This is possible in part due to advances in computational algorithms over the last decades.
- The problem of computational efficiency is quickly made worse, however, by using other models that require iterative numerical methods for solving.
- For example, with Gaussian linear models (which can be solved analytically), a best subset search with 11 predictors takes less than a second.
- By contrast, with a logistic model (solved using Newton’s method), 11 predictors takes nearly 2 minutes, and adding each subsequent predictor causes the time to more than double.

Data from the early childhood longitudinal study.

The outcome is 5th grade math score (C6R4MSCL).

The exposure is an indicator that is 1 if student received special education services at or before 3rd grade; 0 otherwise.

Interested in the effect of exposure to special education on math score, controlling for other variables.

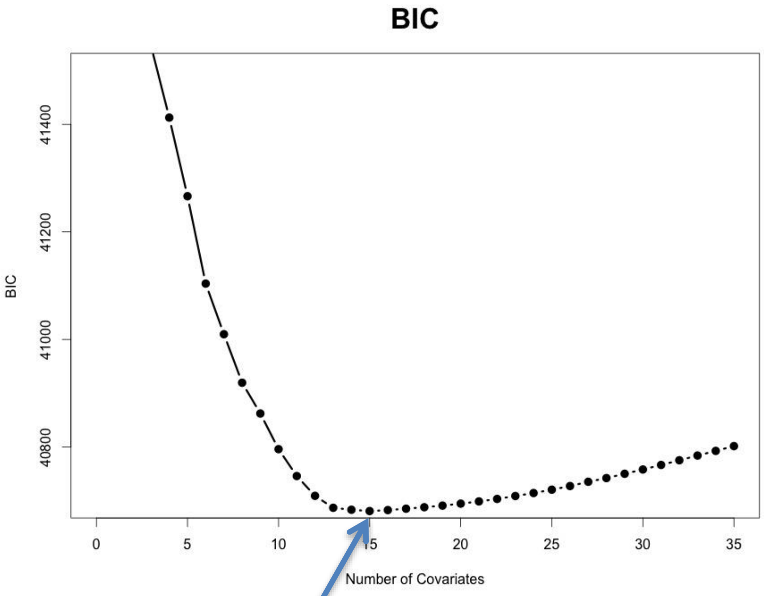


Variable Name	Description of Variable	Values	*d*	*r*
DEMOGRAPHIC				
GENDER	Male	0, 1	0.38	0.88
WKWHITE	White	0, 1	0.17	0.79
WKSESL	Socioeconomic Status	[-4.8, 2.8]	-0.29	0.89
ACADEMIC				
RIRT	Kindergarten Reading Score	[23.17, 139.36]	-0.65	0.53
MIRT	Kindergarten Math Score	[11.9, 99.0]	-0.71	0.77
S2KPUPRI	Public School	0, 1	0.44	0.25
P1EXPECT	Parental Expectations	Integers 1–6	-0.32	1.22
P1FIRKDG	First-Time Kindergartener	0, 1	-0.41	3.26
P1AGEENT	Child's Age at K Entry (Months)	[54, 79]	0.08	1.08
apprchT1	Approaches to Learning Rating	Integers 1–4	-0.70	1.20
P1HSEVER	Attended Head Start	0, 1	0.19	1.42
chg14	Ever Changed Schools	0, 1	0.02	1.09
SCHOOL COMPOSITION				
avg_RIRT	Reading IRT	[27.9, 80.0]	-0.23	0.79
avg_MIRT	Math IRT	[16.1, 66.1]	-0.18	0.82
avg_SES	SES	[-2.2, 2.5]	-0.16	0.88
avg_apprchT1	Approaches to Learning	[1.5, 4.0]	-0.14	0.80
S2KMINOR	Percent Minority Students	Integers 1–5	-0.20	0.77
FAMILY CONTEXT				
P1FSTAMP	Received Food Stamps	0, 1	0.12	1.26
ONEPARENT	One-Parent Family	0, 1	0.13	1.22
STEPPARENT	Stepparent Family	0, 1	0.05	1.19
P1NUMSIB	Number of Siblings	[0, 10]	0.16	1.17
P1HMAFB	Mother's Age at First Birth	Years [12, 45]	-0.26	1.00
WKCAREPK	Nonparental Pre-K Child Care	0, 1	-0.07	1.14
HEALTH				
P1EARLY	Number of Days Premature	[0, 112]	0.19	2.05
wt_ounces	Birth Weight (Ounces)	[17, 214]	-0.11	1.24
C1FMOTOR	Fine Motor Skills	Integers 0–9	-0.63	1.27
C1GMOTOR	Gross Motor Skills	Integers 0–8	-0.43	1.54
PARENT RATING OF CHILD				
P1HSCALE	Overall Health	Integers 1–5	0.12	1.17
P1SADLON	Sad/Lonely	Integers 1–4	0.10	1.32
P1IMPULS	Impulsive	Integers 1–4	0.41	1.55
P1ATTENI	Attentive	Integers 1–4	0.72	1.45
P1SOLVE	Problem Solving	Integers 1–4	0.68	1.55
PSPRONOU	Verbal Communication	Integers 1–4	0.86	1.51
P1DISABL	Child has Disability	0, 1	0.82	2.38
OUTCOME VARIABLE				
C6R4MSCL	Fifth Grade Math Score	[50.9, 170.7]	-0.77	1.40

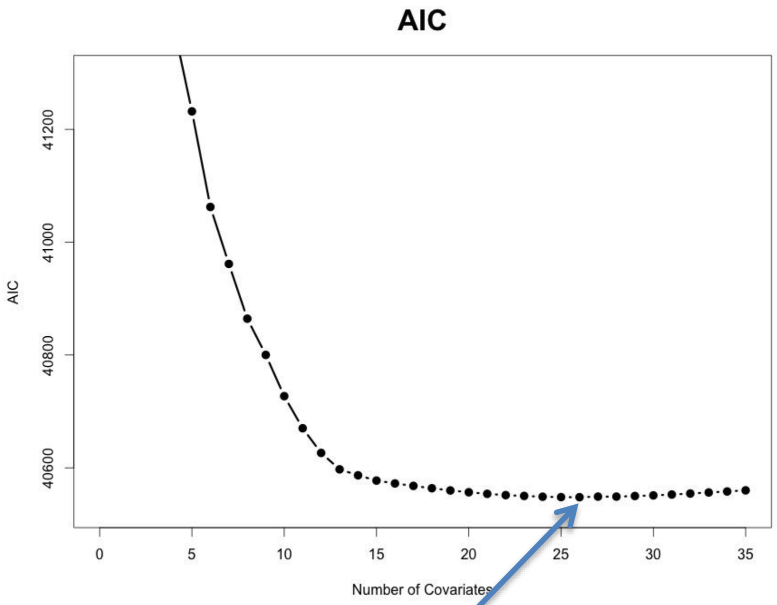
ECLSK Example – Best Subset Selection

```
### Outcome is C6R4MSCL (5th grade math score)
### 36 predictors examined
bs3 <- bestglm(Xy = eclsk1, family = gaussian, IC = "BIC")
bs4 <- bestglm(Xy = eclsk1, family = gaussian, IC = "AIC")
```

Best subset selection works here because the model is Gaussian. If we were to try a logistic model for binomial data with 36 predictors we would get an error message.



BIC identified 15 covariates as the optimal number.



AIC identified 26 covariates as the optimal number.

ECLSK Example - Results

BIC Results

```
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 92.52027    3.30622  27.984  < 2e-16 ***
GENDER        6.20534    0.38073  16.299  < 2e-16 ***
WKWHITE       3.19381    0.46598   6.854 7.76e-12 ***
WKSESL        2.14455    0.34933   6.139 8.73e-10 ***
MIRT          1.17759    0.02432  48.427  < 2e-16 ***
S2KPUPRI      5.52234    0.49119  11.243  < 2e-16 ***
P1FIRKDG     12.04908    1.07558  11.202  < 2e-16 ***
P1AGEENT     -0.72637    0.04860 -14.946  < 2e-16 ***
apprchT1      2.68547    0.32903   8.162 3.85e-16 ***
P1HSEVER     -3.60635    0.61973  -5.819 6.16e-09 ***
ONEPARENT    -1.90534    0.52685  -3.616 0.000301 ***
P1HMAFB       0.21746    0.04114   5.286 1.29e-07 ***
C1FMOTOR      1.67406    0.10746  15.579  < 2e-16 ***
P1SOLVE      -1.16148    0.34532  -3.364 0.000773 ***
avg_SES       3.14823    0.53396   5.896 3.89e-09 ***
F5SPECS      -7.17163    0.81787  -8.769  < 2e-16 ***
```

AIC Results

```
 Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.88775    4.24281  22.836  < 2e-16 ***
GENDER        6.15373    0.38889  15.824  < 2e-16 ***
WKWHITE       2.27241    0.55271   4.111 3.98e-05 ***
WKSESL        1.92348    0.35772   5.377 7.81e-08 ***
MIRT          1.16513    0.02451  47.540  < 2e-16 ***
S2KPUPRI      5.57930    0.49124  11.358  < 2e-16 ***
P1EXPECT      0.41871    0.19160   2.185 0.028893 *
P1FIRKDG     12.05349    1.07801  11.181  < 2e-16 ***
P1AGEENT     -0.72955    0.04909 -14.861  < 2e-16 ***
apprchT1      3.03527    0.38177   7.951 2.14e-15 ***
P1HSEVER     -2.98581    0.64604  -4.622 3.87e-06 ***
P1FSTAMP     -1.58801    0.67575  -2.350 0.018799 *
S2KMINOR     -0.49038    0.16931  -2.896 0.003787 **
ONEPARENT    -1.22370    0.55536  -2.203 0.027595 *
P1HMAFB       0.21800    0.04131   5.278 1.35e-07 ***
WKCAREPK     -1.09543    0.52133  -2.101 0.035657 *
P1EARLY       0.04088    0.01806   2.264 0.023633 *
wt_ounces     0.03510    0.01065   3.295 0.000987 ***
C1FMOTOR      1.68800    0.10862  15.541  < 2e-16 ***
C1GMOTOR     -0.20252    0.10899  -1.858 0.063187 .
P1ATTENI     -0.56043    0.33260  -1.685 0.092030 .
P1SOLVE      -0.82948    0.36751  -2.257 0.024037 *
P1PRONOU     -0.64512    0.32960  -1.957 0.050353 .
P1DISABL      0.85005    0.58371   1.456 0.145356
avg_SES       2.92792    0.55637   5.263 1.46e-07 ***
avg_apprchT1 -1.75277    0.70086  -2.501 0.012410 *
F5SPECS      -7.14141    0.82992  -8.605  < 2e-16 ***
```

Special Ed indicator is significant in both models and the treatment effect estimate is about the same (-7.17 vs. -7.14).

Forward Stepwise Selection

- Forward and backward stepwise approaches to model selection approximate the best subset solution by working with a (much!) more restricted set of models.
- Forward stepwise selection algorithm:
  1. Let M_0 denote the null model (i.e., no predictors).
  2. For $k = 0, \ldots, p - 1$:
     a) Consider all $p - k$ models that increase the predictors in M_k with one additional parameter.
     b) Choose the best among the $p - k$ models and call it $M_{k+1.}$ Note that "best" is defined in terms of smallest SSE.
  3. Select a single best model from $\{M_0, \ldots, M_k\}$ via CV, AIC, BIC, etc.

- In total, this will involve fitting 1 null model and $p - k$ models in the kth iteration, for $k = 0, 1, \ldots, p - 1$.
- This is in contrast to 2^p for best subset selection.

Stepwise Selection in R

There are many functions and R packages for computing stepwise regression. These include:

- `stepAIC()` [MASS package], which choose the best model by AIC. It has an option named direction, which can take the following values: i) “both” (for stepwise regression, both forward and backward selection); “backward” (for backward selection) and “forward” (for forward selection). It return the best final model.

- `regsubsets()` [leaps package], which has the tuning parameter nvmax specifying the maximal number of predictors to incorporate in the model. It returns multiple models with different size up to nvmax. You need to compare the performance of the different models for choosing the best one. regsubsets() has the option method, which can take the values “backward”, “forward” and “seqrep” (seqrep = sequential replacement, combination of forward and backward selections).

ECLSK Example – Forward Stepwise Selection

- Several packages in R can run forward selection. The “stepAIC” function in package MASS is what we will use.

```
library(MASS)
# Smallest model with hust intercept:
min.model <- lm(C6R4MSCL ~ 1, data = eclsk1)
# Largest model with all predictors:
max.model <- lm(C6R4MSCL ~ ., data = eclsk1)
# Scope of the search for forward selection
scp <- list(lower = min.model, upper = max.model)
# Forward selection
fwd <- stepAIC(min.model, direction = 'forward', scope = scp)
fwd$coefficients
(Intercept)          MIRT       WKSESL     C1FMOTOR       GENDER     P1AGEENT
 96.88775211   1.16513347   1.92348365   1.68800475   6.15372996  -0.72954903
    P1FIRKDG     apprchT1      WKWHITE     S2KPUPRI      F5SPECS      avg_SES
 12.05348816   3.03526705   2.27240622   5.57929998  -7.14141261   2.92792434
    P1HSEVER      P1HMAFB    ONEPARENT      P1SOLVE     S2KMINOR    wt_ounces
 -2.98581062   0.21799665  -1.22369547  -0.82947802  -0.49037825   0.03510179
avg_apprchT1     P1EXPECT      P1EARLY     P1FSTAMP     WKCAREPK     P1PRONOU
 -1.75276672   0.41870822   0.04087811  -1.58800536  -1.09542548  -0.64512367
    C1GMOTOR     P1ATTENI     P1DISABL
 -0.20252002  -0.56042548   0.85004828
```

- Note that AIC via best subset also identified the same 26 covariates.

ECLSK Example – Forward Stepwise Selection

- To use the BIC we modify the argument *k* (default is 2 for AIC).

```
fwd2 <- stepAIC(min.model,
                direction = 'forward',
                scope = scp,
                k = log(nrow(eclsk1)))
fwd2$coefficients
(Intercept)        MIRT      WKSESL    C1FMOTOR      GENDER    P1AGEENT    P1FIRKDG
 92.5202705   1.1775938   2.1445523   1.6740648   6.2053449  -0.7263671  12.0490844
   apprchT1     WKWHITE    S2KPUPRI     F5SPECS     avg_SES    P1HSEVER     P1HMAFB
  2.6854711   3.1938133   5.5223400  -7.1716267   3.1482347  -3.6063459   0.2174640
  ONEPARENT     P1SOLVE
 -1.9053391  -1.1614815
```

- Note that BIC via best subset also identified the same 15 covariates.
- The two methods will not always agree.

Backward Stepwise Selection

- Backward stepwise selection is also a much more efficient alternative to best subset selection.
- Whereas forward stepwise selection begins with a null model and moves forward one predictor at a time, backward stepwise selection begins with the full model (i.e., containing all *p* predictors) and then removes the least useful predictors one at a time.
- Backward stepwise selection algorithm (ISLR, p. 209):

1. Let M_p denote the *full* model, containing all *p* predictors.
2. For $k = p, p – 1, p – 2, …, 1$:
   a) Consider all *k* models that contain all but one of the predictors in M_k, for a total of $k – 1$ predictors.
   b) Select the best of the *k* models and label it $M_{k\text{-}1}$. At this stage, *best* is defined by smallest residual sum of squares.
3. Select the single best model from M_0, … , M_p via cross-validation prediction error, AIC, BIC, etc.

Backward Subset Selection

```
bwd <- stepAIC(max.model,
               direction = 'backward',
               scope = scp)
bwd$coefficients
```

Note we now use max.model here instead of min.model.

```
(Intercept)        GENDER      WKWHITE       WKSESL         MIRT     S2KPUPRI
 96.16423579   6.15303750   2.34469806   1.92016994   1.17493417   5.48863902
    P1EXPECT     P1FIRKDG     P1AGEENT     apprchT1     P1HSEVER     P1FSTAMP
  0.42597045  11.96839592  -0.71978192   2.94808419  -2.96083960  -1.56001714
    S2KMINOR    ONEPARENT      P1HMAFB     WKCAREPK      P1EARLY    wt_ounces
 -0.56437253  -1.23950171   0.21721213  -1.06344788   0.04315746   0.03569199
    C1FMOTOR     C1GMOTOR     P1ATTENI      P1SOLVE     P1PRONOU     avg_RIRT
  1.68755103  -0.20388550  -0.54930933  -0.80559386  -0.53292879   0.08605165
    avg_MIRT      avg_SES avg_apprchT1      F5SPECS
 -0.12525895   3.04132473  -1.57149298  -6.96682986
```

- avg_MIRT and avg_RIRT were selected by backward subset selection, but not by best or forward.
- P1DISABL was selected by best and forward but not by backward.

Limitations of Stepwise Selection

- **Limitation 1: Inflation of Type I error rate when testing the significance of predictors.**
- Stepwise routines fit many models along the way to the final formulation, testing many (or all in the case of best subset selection) possible combinations of covariates.
- Because so many combinations are tested, some are bound to be significant by chance.
- Thus, hypothesis testing of regression coefficients after running a stepwise selection routine will typically show that nearly every variable retained is a "significant" predictor of the outcome.
- The problem here lies in the fact that so many models were fit. It is, therefore, not appropriate to interpret the statistical significance of regression coefficients selected by stepwise selection routines at face value.
- For the purpose of making good *predictions*, on the other hand, stepwise selection methods are very useful because they can eliminate non-informative variables.

Limitations of Stepwise Selection

- We can demonstrate the potential to make Type I errors through simulation. Generate Y *completely unrelated* to X1 through X20.

```
### Generate random noise and use stepwise approach
noise <- function(N = 400, p = 20) {
  X <- rmvnorm(N, sigma = diag(p))
  Y <- rnorm(N, sd = 4)
  df <- data.frame(cbind(X,Y))
  names(df) <- c(paste0("X", 1:p), "Y")
  df
}

set.seed(1355)
df6 <- noise()
summary(lm(Y ~ X, df6))
bestglm(Xy = df6, family = gaussian, IC = "AIC")
```

```
              Estimate Std. Error   t value    Pr(>|t|)
(Intercept) -0.3564621  0.2014851 -1.769174 0.077638465
X7          -0.3251008  0.2075722 -1.566206 0.118103234
X11         -0.6132137  0.2182780 -2.809324 0.005211912
X14         -0.3750686  0.2065032 -1.816285 0.070086522
X16         -0.3229027  0.2140893 -1.508262 0.132288968
X20         -0.3001960  0.1912377 -1.569753 0.117275226
```

Best subset selection using the AIC identifies X11 as a significant predictor of Y even though Y and Xs are independent.

Using the BIC also identifies X11 as significant.

Furthermore, the linear regression of Y on all Xs yielded no significant relationships at alpha = 0.05.

Limitations of Stepwise Selection

- **Limitation 2: functional form assumptions are strong.**
- Linear and generalized linear stepwise procedures assume the functional form of the model is a subset of the most complex model you specify.
- For example, if you do not specify any squared terms or interactions in the "upper" model, the assumption is that all predictors are linearly related to the outcome.
- Simulate data so that X1 – X4 are true linear predictors, X5 is a quadratic predictor, and X6 – X7 are non-informative.

```
N <- 200
set.seed(7915)
X2 <- rmvnorm(N, sigma = diag(8))
colnames(X2) <- paste0("X", 1:8)
head(X2)
X2 <- data.frame(X2)
Y <- .4*X2$X1 + .2*X2$X2 + -.6*X2$X3 + -.1*X2$X4 + X2$X5^2 + rnorm(N)
```

- All predictors are uncorrelated for simplicity.

Limitations of Stepwise Selection

```
### Forward selection with AIC
min.mod <- lm(Y ~ 1, data = Xd2)
max.mod <- lm(Y ~ ., data = Xd2)
scp2 = list(lower = min.mod, upper = max.mod)
fwd3 <- stepAIC(min.mod,
                direction = 'forward',
                scope = scp2)
fwd3$coefficients
(Intercept)          X3          X1
  0.9249439  -0.6433981   0.3145278
```

No quadratic term for X5; X5 is dropped.

- One option is to include quadratic terms in your model.

```
### Forward selection with AIC and quadratics
max.mod2 <- lm(Y ~ . + I(X1^2) + I(X2^2) + I(X3^2) + I(X4^2) +
                       I(X5^2) + I(X6^2) + I(X7^2) + I(X7^2), data = Xd2)
scp3 = list(lower = min.mod, upper = max.mod2)
fwd4 <- stepAIC(min.mod,
                direction = 'forward',
                scope = scp3)
fwd4$coefficients

(Intercept)     I(X5^2)          X3          X1     I(X2^2)     I(X3^2)          X2
 0.09638271  0.99442203 -0.62727031  0.38415482 -0.16479006 -0.08016367  0.13308717
```

Quadratic term for X5; X5^2 is included.