STAT 40001/STAT 50001 Statistical Computing

Lecture 17

Department of Mathematics and Statistics
Purdue University Northwest

Model Selection

The Principle of Parsimony:

A model should contain the smallest number of variables necessary to fit the data.

When there are many possible explanatory variables, often times several models are nearly equally good at explaining variation in the response variable.

- R^2 and adjusted R^2 measure closeness of fit, but are poor criteria for variable selection.
- AIC and BIC are sometimes used as objective criteria for model selection.
- Stepwise regression searches for best models, but does not always find them.
- Models selected by AIC or BIC are often over fit.
- Tests after model selection are not valid, typically.
- Parameter interpretation is complex.

R^2 and Adjusted R^2

The coefficient of multiple determination, is denoted by R^2 , and is defined by

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}$$

and the adjusted coefficient of multiple determination, denoted by $R^2_{Adj.}$ is given by

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

which can be written as

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

Remark: In general the value of R^2 never decreases when a regressor is added to the model, regardless of the value of the contribution of that variable. Therefore, it is difficult to judge whether an increase in R^2 is really telling us anything important.
But $R^2_{Adj.}$ will only increase on adding a variable to the model if the addition of the variable reduces the residual mean squares.

Akaike's Information Criterion (AIC)

Akaike's Information Criterion (AIC) is based on maximum likelihood and a penalty for each parameter.
The expression for AIC for a general model is

$$AIC = -2 \log L + 2p$$

where L is the likelihood and p is the number of parameters.
In multiple linear regression model with with $p-1$ regressor variables and n observations it becomes

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

where SSE is the residual sum of squares.
A model having minimum AIC is considered to be a better model.

Bayesian Information Criterion (BIC)

Schwartz's Bayesian Information Criterion (BIC) is similar to AIC but penalizes additional parameters more.
The general form is

$$BIC = -2 \log L + (\log n)p$$

where n is the number of observations. L is the likelihood and p is the number of parameters.
In multiple linear with $p-1$ regressor variables it becomes

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

where SSE is the residual sum of squares .
A model having minimum BIC is considered to be a better model.
Observe that for both the AIC and BIC the first term is the same which decreases as p increases and the second term increases with the number of parameters, p. If $n \geq 8$ the penalty term for BIC is larger than that for AIC , hence BIC criteria tends to favor more parsimonious models.

Mallow's C_p Statistics

The Mallow's C_p statistic is based on the normalized expected total error of estimation. After substituting the sample estimator s^2 for σ^2, we have

$$C_p = \frac{SSE}{s^2} + 2p - n$$

A value of C_p near p suggests that the model bias is small which means there is no significant over-fitting or under-fitting of the model. Values of c_p near or below p is generally desirable.
If the p-term model has negligible bias then SSE=0. Consequently, $E(SSE) = (n-p)\sigma^2$ and

$$E(C_p|Bias = 0) = \frac{(n-p)\sigma^2}{\sigma^2} - n + 2p = p$$

PRESS Statistic

For a specified model the PRESS (prediction sum of squares) statistic is formed by predicting each observation based on a model developed by using the other observations. Let us define the PRESS residual as

$$e_{(i)} = y_i - \hat{y}_{(i)} \quad i = 1, 2, 3, \cdots n$$

where $\hat{y}_{(i)}$ is the fitted value of the i^{th} response based on all $n-1$ observations except the i^{th} one. Note that large PRESS residuals are potentially useful in identifying observations where the model does not fit the data well or observations for which the model is likely to provide poor future predictions. We define the PRESS statistic as

$$PRESS = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2$$

Note that models with smaller PRESS statistics are preferred.
PRESS can be computed in R using library MPV :

```
library(MPV)
PRESS(model)
```

Variable Selection

Some testing based procedures for variable selection are

- Forward Selection
- Backward Elimination
- Stepwise Regression

Forward selection assumes a model with an intercept only and adds the most significant (smallest p-value) variables one at a time. The function *add*1() in R is used to find out which variable will be added in the model
Backward elimination starts with all the variables in the model and eliminates variables with the largest (least significant) p-values:
The function *drop*1() in R is used to find out which variable will be added in the model
This is a combination of backward elimination and forward selection. This allows to reenter the variable that has been dropped in the backward elimination method.
In order to perform the stepwise selection we need library named *MASS*.

Example-Survival Time

A hospital studied the survival time(y) of patients who had undergone a liver operation. The regressor variables for the data includes

x_1: blood clotting score

x_2: prognostic index

x_3: enzyme function test score

x_4: liver function test score

x_5: age, in years

x_6: gender,(0=male, 1=female)

x_7 : moderate use of alcohol

x_8: severe use of alcohol

Sample of the data is as below (Complete data in the Brightspace)

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y
6.7	62	81	2.59	50	0	1	0	695
5.1	59	66	1.70	39	0	0	0	403
7.4	57	83	2.16	55	0	0	0	710
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6.4	59	85	2.33	63	0	1	0	550
8.8	78	72	3.20	56	0	0	0	651

Example-Survival Time- Forward Selection

```
> add1(lm(y~1), scope=(~.+x1+x2+x3+x4+x5+x6+x7+x8),test="F")
Single term additions

Model:
y ~ 1
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              8369521 647.36
x1      1   1005152 7364369 642.45  7.0974  0.010255 *
x2      1   1479767 6889754 638.85 11.1685  0.001547 **
x3      1   2798310 5571211 627.38 26.1186 4.671e-06 ***
x4      1   3804272 4565248 616.63 43.3322 2.289e-08 ***
x5      1    118863 8250658 648.59  0.7491  0.390726
x6      1    251809 8117712 647.71  1.6130  0.209721
x7      1    271062 8098458 647.58  1.7405  0.192857
x8      1   1454057 6915463 639.06 10.9336  0.001717 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that x4 has the smallest p-value so it will be added in the model.

Example-Survival Time-Forward Selection

```
> add1(lm(y~1+x4), scope=(~.+x1+x2+x3+x5+x6+x7+x8),test="F")
Single term additions
Model:
y ~ 1 + x4
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              4565248 616.63
x1      1       685 4564563 618.62  0.0077 0.9306162
x2      1    285592 4279656 615.14  3.4034 0.0708749 .
x3      1    896004 3669244 606.83 12.4539 0.0008935 ***
x5      1      3726 4561522 618.59  0.0417 0.8390782
x6      1      6162 4559086 618.56  0.0689 0.7939538
x7      1    225396 4339852 615.90  2.6488 0.1097942
x8      1    939077 3626171 606.19 13.2076 0.0006480 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that x8 has the smallest p-value so it will be added in the model.

The process continues until we obtain the p values for each of the regressor variables greater than desired level of α.

Example-Survival Time-Backward Elimination

```
> drop1(lm(y~x1+x2+x3+x4+x5+x6+x7+x8), test="F")
Single term deletions
Model:
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>               1825906 581.14
x1      1    263780 2089686 586.43  6.5009 0.0142584 *
x2      1    930187 2756093 601.38 22.9247 1.857e-05 ***
x3      1   1307757 3133663 608.31 32.2301 9.390e-07 ***
x4      1     51016 1876922 580.63  1.2573 0.2681093
x5      1      5231 1831137 579.30  0.1289 0.7212314
x6      1      2990 1828896 579.23  0.0737 0.7872685
x7      1       572 1826478 579.16  0.0141 0.9060073
x8      1    576636 2402542 593.96 14.2114 0.0004736 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that x7 has the largest p-value so it is eliminated from the model.

Example-Survival Time-Backward Elimination

```
> drop1(lm(y~x1+x2+x3+x4+x5+x6+x8), test="F")
Single term deletions

Model:
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x8
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>               1826478 579.16
x1      1    263219 2089697 584.43  6.6292   0.01331 *
x2      1    936544 2763022 599.51 23.5869 1.420e-05 ***
x3      1   1309433 3135911 606.35 32.9782 7.027e-07 ***
x4      1     51951 1878429 578.68  1.3084   0.25860
x5      1      4878 1831356 577.31  0.1229   0.72755
x6      1      2958 1829436 577.25  0.0745   0.78613
x8      1    726329 2552807 595.24 18.2926 9.472e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Note that x6 has the largest p-value so it is eliminated from the model.
The process continues until we obtain the p values for each of the regressor variables less than desired level of α.

Example-Survival Time-Stepwise Regression

```
> library(MASS)
> step <- stepAIC(model, direction="both")
Start:  AIC=581.14
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8
       Df Sum of Sq     RSS    AIC
- x7    1       572 1826478 579.16
- x6    1      2990 1828896 579.23
- x5    1      5231 1831137 579.30
- x4    1     51016 1876922 580.63
<none>              1825906 581.14
- x1    1    263780 2089686 586.43
- x8    1    576636 2402542 593.96
- x2    1    930187 2756093 601.38
- x3    1   1307757 3133663 608.31
Step:  AIC=579.16
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x8
       Df Sum of Sq     RSS    AIC
- x6    1      2958 1829436 577.25
- x5    1      4878 1831356 577.31
- x4    1     51951 1878429 578.68
<none>              1826478 579.16
+ x7    1       572 1825906 581.14
- x1    1    263219 2089697 584.43
- x8    1    726329 2552807 595.24
- x2    1    936544 2763022 599.51
- x3    1   1309433 3135911 606.35
Step:  AIC=577.25
y ~ x1 + x2 + x3 + x4 + x5 + x8
       Df Sum of Sq     RSS    AIC
- x5    1      4281 1833716 575.38
- x4    1     63596 1893032 577.09
<none>              1829436 577.25
+ x6    1      2958 1826478 579.16
+ x7    1       540 1828896 579.23
- x1    1    260399 2089835 582.44
- x8    1    723371 2552807 593.24
- x2    1    934511 2763947 597.53
- x3    1   1306483 3135918 604.35
```

Model Selection using regsubsets

The R package *leaps* is needed for the function *regsubsets*() for $R^2_{Adj.}$

```
> library(leaps)
> subsets=regsubsets(y~x1+x2+x3+x4+x5+x6+x7+x8, data=data)
> subsets
Subset selection object
Call: regsubsets.formula(y ~x1+x2+x3+x4+x5+x6+x7+x8,data=data)
8 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
x5     FALSE      FALSE
x6     FALSE      FALSE
x7     FALSE      FALSE
x8     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
```

We can add the option *nbest*= "." to choose the number of best models.

To get some additional information about the data we must use

Model Selection using regsubsets

```
> summary(subsets)
         x1  x2  x3  x4  x5  x6  x7  x8
1  ( 1 ) " " " " " " "*" " " " " " " " "
2  ( 1 ) " " " " " " "*" " " " " " " "*"
3  ( 1 ) "*" "*" "*" " " " " " " " " " "
4  ( 1 ) "*" "*" "*" " " " " " " " " "*"
5  ( 1 ) "*" "*" "*" "*" " " " " " " "*"
6  ( 1 ) "*" "*" "*" "*" "*" " " " " "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" " " "*"
8  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*"
```

The useful part of this is the last, a "matrix" indicating which variables are included in each of the models. Models are listed in order of size (the first column), and within a size, in order of fit (best model first). The included variables are indicated by asterisks in quotes and variables not in a model have empty quotes.

Model Selection using regsubsets

An option is available which makes this matrix perhaps more readable:

```
> summary(subsets, matrix.logical=TRUE)
```

The subsets obtained above can be plotted using the code below in R

```
> plot(subsets)
```

Model Selection using AIC /BIC/Cp

AIC and BIC:

There is a function that gives AIC or Schwartz' BIC, from an lm object:

```
> model=lm(y~x1+x2+x3+x4+x5+x6+x7+x8)
> AIC(model)
[1] 736.3899
```

The Mallow's c_p and adjusted $R^2_{Adj.}$ can be obtained as

```
> library(leaps)
> subsets=regsubsets(y~x1+x2+x3+x4+x5+x6+x7+x8, data=data)
> summary(subsets)$cp
> summary(subsets)$adjr2
```

Model building steps

- Fit the largest model possible to the data
- Perform a through analysis of the model
- Determine if a transformation of the data is necessary
- Determine if all possible regressions is feasible
  - If all possible regressions is feasible perform all possible regressions using such criteria as Mallow's C_p, Adjusted R^2 and the PRESS statistic to rank the best subset models.
  - if all possible regression is not feasible , use stepwise selection techniques to generate the largest model
- Compare and contrast the best models recommended by each criteria
- Perform a through analysis of the "best" models
- Explore the need for further transformation
- Make the recommendation.

Model Validation

The final step in the model building process is the validation of the selected regression model. Model validation involves checking a candidate model against independent data. Three types of model validation procedures are

- Collection of new (or fresh) data to check the model and its predictive ability
- Comparison of the results with theoretical expectations, earlier empirical results and simulation results.
- Data Splitting, that is, set aside some of the original data and use these observations to investigate the model's predictive performance.

The final intended use of the model often indicates the appropriate validation methodology. Thus, validation of a model for prediction purpose should be as accurate as possible. If possible all the validation techniques must be used. The best means of model validation is through the collection of the new (fresh) data. Sometimes these new observations are called "conformation runs".

Diagnostics for Leverage and Influence

An outlying cases is defined as a particular observation $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$ that differs from the majority of the cases in the data set. Since several variables are involved in each case, one must distinguish among outliers in the y(response) dimension and outliers in the x(covariate) dimension. Note that the detection of outlier will be tricky if more than two dimensions are involved.
Outliers in the y are linked to the regression model as one tries to explain the response as a function of the covariates. Outliers in the y dimension may be due to many different reasons

- The random component in the regression may be unusually large.
- The response y or the one of the response variable may have been recorded incorrectly.
- Both x and y variables are correct but there is another covariate that is missing from the model and that can explain the "strange" observation.

Several procedures for detecting outliers has been developed. A simplest step is to compute the studentized residuals d_i and it has approximately standard normal distribution as long as the model assumptions are satisfied. Large value of d_i are unexpected. For example $|d_i| > 2.5$ is indicative of an outlier.

Identifying Influential Cases

After identifying cases that are outlying with respect to their y values and /or their x values, we want to know whether they are influential. A case is said to be influential if its exclusion from the model causes major changes in the fitted regression model. We can identify the influential cases using three different measures:

- DFFITS Measure
- Cook's Distance
- DFBETAS Measure

We can use R to calculate DFFITS, Cook's D and DFBETAS as below:

```
dffits(model)
cooks.distance(model)
dfbetas(model)
```

In fact, we can use a sigla code `influence.measures(model)` to obtain all of the DFFITS, Cook's D and DFBETAS values.

- If the absolute value of DFFITS exceeds 1 for small data set and $2\sqrt{p/n}$ for large data set then it must be considered as an influential case.
- We usually consider points for which $D_i > 1$ to be influential.
- If $|DFBETAS| > 1$ for small to medium data sets and $|DFBETAS| > 2/\sqrt{n}$ for large data set.

olsrr package

The olsrr package provides following tools for OLS regression using R:

- comprehensive regression output
- residual diagnostics
- measures of influence
- heteroskedasticity tests
- collinearity diagnostics
- model fit assessment
- variable contribution assessment
- variable selection procedures

Example

```
> library(olsrr)
> data(mtcars)
> model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
> ols_regress(mpg ~ disp + hp + wt + qsec, data = mtcars)
> ols_plot_resid_fit(model)
> ols_plot_cooksd_bar(model, threshold = 0.125)
> ols_plot_cooksd_chart(model)
> ols_plot_dfbetas(model)
> ols_plot_dffits(model)
> ols_plot_resid_lev(model)
```

Example

This data refer to the dataset `swiss` available in R. The following variables were collected (primarily from military records) for each of the 47 French-speaking provinces in Switzerland:

`Fertility:` (standardized)

`Agriculture :` Percent of males involved in agriculture as an Occupation

`Examination:` Percent of draftees who received the highest mark on their army examination

`Education:` Percent of draftees educated beyond primary school

`Catholic:` Percent Catholic (as opposed to Protestant)

`InfantMortality:` Percent of live births who live less than one year

```
> data(swiss)
> Fertility=swiss$Fertility
> Agriculture =swiss$Agriculture
> Exam=swiss$Examination
> Education=swiss$Education
> Catholic=swiss$Catholic
> Mortality=swiss$Infant.Mortality
> model=lm(Fertility~Agriculture+Exam+Education+Catholic+Mortality)
>influence.measures(model)
>extractAIC(model)
>extractAIC(model,k=log(n)) ## BIC
```

Example-Swiss Data

```
> library(leaps)
> subset=regsubsets(Fertility~ Agriculture+ Exam+ Education+ Catholic+
        + Mortality, data=swiss)
> summary(subset)
> plot(subset,scale="Cp")
> plot(subset,scale="r2")
> plot(subset,scale="adjr2")
> subset=regsubsets(Fertility~ Agriculture+ Exam+ Education+ Catholic+
        + Mortality, data=swiss, nbest=10)
> plot(subset)
```