Advanced Methods in Biostatistics II

Lecture 6

November 9, 2017

Linear model

- Consider the linear model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

  where $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

- Today we will discuss issues related to model selection.

Example

Recall the housing data set.

```
> Housing = read.table("housing.txt",header=TRUE)
> head(Housing)
  Taxes Bedrooms Baths  Price Size   Lot
1  1360        3   2.0 145000 1240 18000
2  1050        1   1.0  68000  370 25000
3  1010        3   1.5 115000 1130 25000
4   830        3   2.0  69000 1120 17000
5  2150        3   2.0 163000 1710 14000
6  1230        3   2.0  69900 1010  8000
```

Example

```
> results = lm(Price ~ Size + Lot + Taxes + Bedrooms + Baths, data=Housing)
> summary(results)

Call:
lm(formula = Price ~ Size + Lot + Taxes + Bedrooms + Baths, data = Housing)

Residuals:
   Min     1Q Median     3Q    Max
-89978 -16931  -1407  19077  73705

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6633.7997 15834.6177   0.419 0.676214
Size           33.5714     8.8904   3.776 0.000279 ***
Lot             1.6162     0.4948   3.266 0.001522 **
Taxes          20.6436     5.2558   3.928 0.000163 ***
Bedrooms    -6469.6862  5313.1550  -1.218 0.226396
Baths       11824.4881  7320.9445   1.615 0.109628
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 27980 on 94 degrees of freedom
Multiple R-squared:  0.7659,Adjusted R-squared:  0.7535
F-statistic: 61.52 on 5 and 94 DF,  p-value: < 2.2e-16
```

Model selection

- We often have data on a large number of explanatory variables and wish to build a model using some subset of them.
- This becomes a problem of choosing between different competing linear models.

Model selection

- If the model is too small we ‘underfit’ the data. This leads to poor predictions and high bias, but low variance.
- If the model is too big we ‘overfit’ the data. This leads to poor predictions and high variance, but low bias.
- When the model is ‘just right’, we balance bias and variance to get good predictions.

The principle of parsimony

- The Principle of Parsimony says that when two competing models have the same predictive power, the model with the lower number of parameters should be used.
- Occam's Razor - simple models are preferred over complicated ones.

Model selection

- One approach towards model selection is to consider all possible subsets of the pool of explanatory variables and find the ‘best’ model according to some criteria (i.e., perform an exhaustive search).
- Another approach is to use a search algorithm to find the ‘best’ model.
- The latter approach is usually more efficient when the number of variables is large.

Model selection

- In both approaches a number of different criteria may be used to select the best model.
- Popular choices include Adjusted R^2, Mallow's C_p, AIC, BIC, and the PRESS statistic.
- These criteria assign scores to each model and allow us to choose the model with the best score.

Sums of squares

- Recall, we can partition the data into sums of squares:

$$\text{SST}_p = \text{SSE}_p + \text{SSR}_p$$

where

$$\begin{aligned}
\text{SST}_p &= ||\mathbf{y} - \bar{y}\mathbf{J}_n||^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H}_\mathbf{J})\mathbf{y} \\
\text{SSE}_p &= ||\mathbf{y} - \hat{\mathbf{y}}||^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H}_\mathbf{X})\mathbf{y} \\
\text{SSR}_p &= ||\hat{\mathbf{y}} - \mathbf{J}_n\bar{y}||^2 = \mathbf{y}'(\mathbf{H}_\mathbf{X} - \mathbf{H}_\mathbf{J})\mathbf{y}
\end{aligned}$$

- Here the subscript p refers to the fact that the model includes p parameters.

R_p^2

- Recall the coefficient of multiple determination is

$$R_p^2 = \frac{\text{SSR}_p}{\text{SST}_p} = 1 - \frac{\text{SSE}_p}{\text{SST}_p}.$$

- This represents the proportion of the total variability explained by our model.
- This is guaranteed to be between 0 and 1.

R_p^2

- High values imply that the explanatory variables are useful in explaining the response and low values imply that the explanatory variables are not useful.
- Since R^2 increases with the size of the model, it is however not a good criterion for variable selection.
- It would always choose to include all variables.

Adjusted R_p^2

- The adjusted coefficient of multiple determination, uses the mean squares instead of the sums of square, i.e.

$$R_{a,p}^2 = 1 - \frac{\text{MSE}_p}{\text{MST}_p} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}_p}{\text{SST}_p}.$$

- Since the term includes the number of model parameters, p, it penalizes for model complexity.

Mallow's C_p

- Mallow's C_p is a criteria for assessing fits when models with different numbers of parameters are being compared.
- It can be expressed as:

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

- Here the MSE of the full model is used to estimate s^2.

Mallow's C_p

- If the model includes all important variables, then

$$E(SSE_p) = (n-p)\sigma^2.$$

- If s^2 provides a good estimate of σ^2 then

$$E(C_p) \approx \frac{(n-p)\sigma^2}{\sigma^2} - n + 2p = p.$$

Mallow's C_p

- Values close to the corresponding p indicate a good model.
- The best model has a small C_p value.

Information Criteria

- Many methods are based on combining a term based on the log-likelihood with one based on model complexity.
- This provides a means to balance model fit with model complexity when assessing the best fitting models.
- This allows us to penalize unnecessarily complicated models.

Information Criteria

- To illustrate, let $f(\mathbf{y}|\mathbf{X}, \beta)$ be the density of the response $\mathbf{y}$.
- For a sample of n observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots n$, the log-likelihood is given by

$$\log(L) = \sum_{i=1}^{n} log(f(y_i|\mathbf{x}_i, \beta).$$

- Let $\ell(\hat{\beta}_p)$ be the log-likelihood evaluated at the MLE under the model with p parameters.

Information Criteria

- Standard information criteria are of the form:

$$-2\ell(\hat{\beta}) - \phi(n, p).$$

- Here the first term represents model fit, and the second a penalized model complexity.
- In the linear model setting with Guassianity assumptions

$$\ell(\hat{\beta}) \propto -n/2 \log(SEE_p/n).$$

AIC

- Akaike's Information Criterion (AIC) tries to balance the conflicting demands of model accuracy and parsimony.
- For the linear model it can be expressed as:

$$AIC_p = nlog(SSE/n) + 2p.$$

- Low values indicate a better model.

BIC

- Several modifications of AIC have been suggested.
- For example, the Bayesian Information Criterion (BIC) is defined as:

$$BIC_p = nlog(SSE/n) + log(n)p.$$

- Again, low values indicate a better model.

AIC vs. BIC

- The difference between AIC and BIC lies in the severity of the penalty.
- The penalty is larger for BIC when $n > 8$.
- Hence, BIC tends to favor more parsimonious models compared to AIC which has a tendency to overfit (i.e., include too many explanatory variables).

PRESS

- The prediction sum of squares (PRESS) criterion measures how well the fitted values for a subset model can predict the observed response.
- The PRESS statistic is defined as:

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{(i),i})^2$$

PRESS

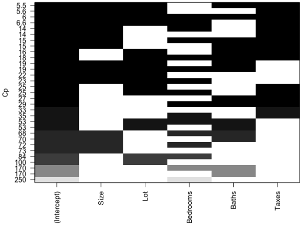
- Thus, the PRESS statistic is the sum of squared deleted residuals.
- The model with the smallest PRESS statistic is considered ‘best’.
- Leaving one item out at a time is known as leave-one-out cross-validation.
- This allows us to predict the performance of the model on holdout data.

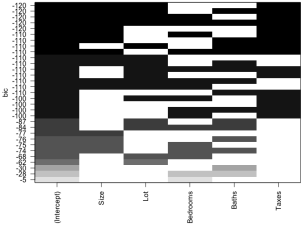
Exhaustive search

- One approach towards model selection is to consider all possible subsets of the pool of explanatory variables and find the ‘best’ model according to some criteria.
- However, if have 15 predictors there are 2^{15} different models (even before considering interactions, transformations, etc.)
- ‘Leaps and bounds’ is an efficient algorithm to perform such a search.

Example

```
> library(leaps)
> leaps=regsubsets(Price~Size+Lot+Bedrooms+Baths+Taxes,
data=Housing, nbest=10)
> plot(leaps, scale="Cp")
> plot(leaps, scale="bic")
```





Step-wise methods

- When the number of explanatory variables is large it is not feasible to fit all possible models.
- Instead, it is more efficient to use a search algorithm to find the best model.
- A number of such algorithms exist, including forward selection, backward elimination and stepwise regression.

Step-wise methods

- Let us begin by setting up the problem.
- Assume we are choosing from a set of P possible explanatory variables $v_k,\ k = 1, \ldots, P$.
- In each algorithm our goal is to find the subset of v_k that best balances model fit and parsimony.

Forward Selection

1. Fit the P simple linear regression models:

$$y_i = \beta_0 + \beta_1 v_{ki} + \epsilon_i \;\; k = 1, \ldots P.$$

2. Set $x_1 = v_k$, where v_k is the variable that has the most significant coefficient (i.e., the smallest p-value).

3. Lock in the variable x_1, and repeat the procedure with models that include two explanatory variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 v_{ki} + \epsilon_i, \;\; k = 1, \ldots P, \;\; v_k \neq x_1.$$

   Set $x_2 = v_k$, where v_k is the variable that has the most significant coefficient, or stop if no variable is significant.

4. Continue until no remaining v_k generate a p-value smaller than the preset significance level α.

Forward Selection

- The criteria for choosing whether to include a new variable can vary.
- As an alternative to using p-values one can instead use a criteria such as the AIC or BIC.
- In each step choose the variable whose inclusion lowers the AIC the most.
- If no variables lower the AIC than stop the algorithm.

Example

```
> null=lm(Price~1, data=Housing)
> null

Call:
lm(formula = Price ~ 1, data = Housing)

Coefficients:
(Intercept)
     126698

> full=lm(Price~., data=Housing)
> full

Call:
lm(formula = Price ~ ., data = Housing)

Coefficients:
(Intercept)        Taxes    Bedrooms       Baths          Size           Lot
   6633.800       20.644   -6469.686   11824.488        33.571         1.616
```

Example

```
> step(null, scope=list(lower=null, upper=full), direction="forward")
Start:  AIC=2188.89
Price ~ 1

                Df    Sum of Sq         RSS       AIC
+ Taxes          1   2.1337e+11  1.0107e+11    2077.4
+ Size           1   1.8222e+11  1.3221e+11    2104.2
+ Lot            1   1.6020e+11  1.5424e+11    2119.7
+ Baths          1   1.0258e+11  2.1186e+11    2151.4
+ Bedrooms       1   4.1519e+10  2.7291e+11    2176.7
<none>                           3.1443e+11    2188.9

Step:  AIC=2077.39
```

Example

```
......

Step:  AIC=2053.23
Price ~ Taxes + Size + Lot + Baths

            Df   Sum of Sq        RSS      AIC
<none>                       7.4755e+10   2053.2
+ Bedrooms  1   1160850856   7.3594e+10   2053.7

Call:
lm(formula = Price ~ Taxes + Size + Lot + Baths, data = Housing)

Coefficients:
(Intercept)       Taxes        Size        Lot      Baths
  -5363.254      20.517      29.484      1.689  10606.892
```

Backward Elimination

1. Start by fitting a model that includes all possible variables:

$$y_i = \beta_0 + \beta_1 v_{1i} + \cdots + \beta_P v_{Pi} + \epsilon_i.$$

2. Find the variable v_k which has the least significant coefficient (i.e. the largest p-value). If its p-value is smaller than some preset significance level, stop the algorithm, otherwise drop the variable.

3. Fit the largest model excluding v_k. Find the variable which has the least significant regression coefficient. If its p-value is smaller than some preset significance level, stop the algorithm, otherwise drop the variable.

4. Continue until the algorithm stops.

Backward Elimination

- Alternatively, AIC or BIC can be used as a criteria for determining whether to drop variables.
- Start with a full model. In each step choose the variable whose exclusion lowers the AIC the most.
- If the exclusion of any variable does not lower the AIC than stop the algorithm.

Example

```
> step(full, data=Housing, direction="backward")
Start:  AIC=2053.67
Price ~ Taxes + Bedrooms + Baths + Size + Lot

              Df   Sum of Sq        RSS       AIC
- Bedrooms     1   1.1609e+09   7.4755e+10   2053.2
<none>                          7.3594e+10   2053.7
- Baths        1   2.0424e+09   7.5636e+10   2054.4
- Lot          1   8.3521e+09   8.1946e+10   2062.4
- Size         1   1.1164e+10   8.4758e+10   2065.8
- Taxes        1   1.2078e+10   8.5672e+10   2066.9

Step:  AIC=2053.23
```

Example

```
Price ~ Taxes + Baths + Size + Lot

         Df  Sum of Sq         RSS     AIC
<none>                  7.4755e+10  2053.2
- Baths   1  1.6747e+09 7.6430e+10  2053.4
- Lot     1  9.2489e+09 8.4004e+10  2062.9
- Size    1  1.0042e+10 8.4797e+10  2063.8
- Taxes   1  1.1935e+10 8.6690e+10  2066.0

Call:
lm(formula = Price ~ Taxes + Baths + Size + Lot, data = Housing)

Coefficients:
(Intercept)        Taxes        Baths         Size          Lot
  -5363.254       20.517    10606.892       29.484        1.689
```

Stepwise Regression

1. Start in the same manner as in forward selection and add the most significant variable from a series of P simple linear regressions.
2. Once a new variable has been included, check the other variables already included in the model for their partial significance. Remove the least significant variable whose p-value is greater than the preset significance level.
3. Continue until no variables can be added and none removed, according to the specified criteria.

Again, note that AIC can be used instead of p-values.

Example

```
> step(null, scope = list(upper=full.lm), data=Housing,
direction="both")

....


Call:
lm(formula = Price ~ Taxes + Size + Lot + Baths, data = Housing)

Coefficients:
(Intercept)        Taxes         Size          Lot        Baths
  -5363.254       20.517       29.484        1.689    10606.892
```

Shrinkage Methods

- The subset selection procedure is a discrete process, as individual variables are either in or out.
- This method can have high variance in the sense that a different dataset from the same source can result in a totally different model.
- Shrinkage methods allow a variable to be partly included in the model.
- That is, the variable is included but with a shrunken co-efficient.

Shrinkage Methods

- Popular methods include ridge regression, the Lasso, and Elastic Net.
- For these approaches we seek to minimize the penalized sums of squares.
- We will revisit these methods in a later lecture.