Akaike's Information Criterion
(Statistical Modelling I)

Lubna Shaheen

Week 9, Lecture 2

Outline

1. What is Backwards Elimination Method
2. What is Forward Elimination Method
3. Akaike's Information Criteria

Automatic Methods of model selection

Remark

If we have small number of explanatory variables: then calculating the

1. MSE
2. R^2
3. Adjusted R^2
4. Mallow's CP

for each of the candidate of models and then making a selection is feasible.

However as the number of potential variables grows, these all calculations for all models becomes more challenging. In response to this a number of **Automatic regression model selection procedures have been introduced.**

Each of these have their advantages and disadvantages.

Automatic Methods of model selection

1. Backwards Elimination Method
2. Forwards Elimination Method
3. Akaike's Information Criteria (AIC)

Backwards Elimination Method

What Is Backward Elimination Technique in Multiple Linear Regression?

The backward elimination technique is used to find the best subset of features from a given set of features. It works by iteratively removing features that are not predictive of the target variable or have the least predictive power.

Backward elimination process iteratively remove the least important variables until only the most important ones remain.

The backward elimination process begins by fitting a multiple linear regression model with all the independent variables. The variable with the highest **p-value** (or small F value) is removed from the model, and a new model fits. This process is repeated until all variables in the model have a **p-value** below some threshold, typically 0.05.

Process of Backwards Elimination

Summary

1. **Step 1**: Fit the multiple linear regression model that uses all the explanatory variables
2. **Step 2**: As the number of variables increases and we risk including variables that are essentially themselves linear combinations of other variables in the model, we run into the problem of multicollinearity which we will look at in the section below
3. **Step 3**: Calculate the F statistic (or the t statistic) for the exclusion of each variable
4. **Step 4**: Find the variable with the lowest F statistic and eliminate this if the statistic is smaller than some predetermined value
5. **Step 5**: This leaves a model with one fewer variable. Now fit this model and re-run the process above
6. **Step 6**: Stop when a variable is not omitted (because the smallest F statistic is no longer smaller than the predetermined value).

Determine the least significant variable to remove at each step

How backward elimination works, we need to know

- The least significant variable at each step
- The stopping rule.

The least significant variable is a variable that:

1. Has the highest p-value in the model, or
2. Its elimination from the model causes the lowest drop in R^2, or
3. Its elimination from the model causes the lowest increase in RSS (Residuals Sum of Squares) compared to other predictors

Choose a stopping rule

The stopping rule is satisfied when all remaining variables in the model have a p-value smaller than some pre-specified threshold.

Example of Backwards Elimination

- **Backward selection of variable;**

Stopping rule: if no variable is greater than 0.05(p>0.05)

Start with full model

Full_model=lm(y~x1+x2+x3+x4)

summary(full_model)

Variable	P value
X1	0.0000447613
X2	0.0010443454
X3	0.18160235
X4	0.8763012799

- Eliminate x4 as it has highest p value

Model=Y = b0+b1x1+b2x2+b3x3+ei

Queen Mary University of London

Example of Backwards Elimination

- Now to eliminate another variable, run the model again

M1=lm(y~x1+x2+x3)

summary(M1)

Variable	P value
X1	0.00002536951
X2	0.0005643223
X3	0.149285

➢ Eliminate x3 as only it has highest p value and p value>0.05

Model = Y=b0+b1x1+b2x2+ei

Example of Backwards Elimination

- Run model with remaing variables

M2=lm(y~x1+x2)

Summary(M2)

Variable	P value
X1	0.000016
X2	0.000519705

- We will not eliminate any variable as no variable has p value>0.05 (stopping rules implies.

Final backward Model = Y = b0+b1x1+b2x2+ei

Queen Mary University of London

Advantages and Disadvantages of Backward Elimination

Advantages

Where backward stepwise is better

(1) Starting with the full model has the advantage of considering the effects of all variables simultaneously.
This is especially important in case of **collinearity** because backward stepwise may be forced to keep them all in the model unlike forward selection where none of them might be entered.

(2) Easy to conduct

(3) It improves model generalizability

Disadvantages

Where backward stepwise is NOT better

(1) It does not consider all possible combination of potential predictors

(2) It outputs biased regression coefficients, confidence intervals, p-values, and R^2

(3) It produces an unstable selection of variables

(4) It does not consider the causal relationship between variables

Forward Elimination Method

What Is Forward Elimination Technique in Multiple Linear Regression?

Forward variable selection is a statistical method used in regression analysis to select the best subset of predictors from a larger set of potential predictors

Forward selection method begins with a model that contains no variables called the Null Model. Then starts adding the most significant variables one after the other until a pre-specified stopping rule is reached or until all the variables under consideration are included in the model.

Process of Forwards Elimination

Summary

Forward Regression Elimination works in the opposite direction to Backward Elimination. This process can be summarised as

1. **Step 1**: Start with the null model $\beta_0 + \epsilon_i$
2. **Step 2**: Fit simple linear regression models for each of the explanatory variables under consideration
3. **Step 3**: Select the explanatory variable whose simple linear regression model has the largest F statistic (or t statistic)
4. **Step 4**: Now add the explanatory variable with the next highest F statistic
5. **Step 5**: Test (via subset deletion) whether either of the two variables can be omitted according to some predetermined F value. [Sometimes the process may have a higher value needed for omission of an existing variable than for the newest variable just added]
6. **Step 6**: Continue until no more variables are added or omitted.

Determine the most significant variable to remove at each step

How backward elimination works, we need to know

- The most significant variable at each step
- The stopping rule.

The most significant variable is a variable that:

1. It has the smallest p-value
2. It provides the highest increase in R^2
3. It provides the highest drop in model RSS (Residuals Sum of Squares) compared to other predictors under consideration.

Choose a stopping rule

The stopping rule is satisfied when all remaining variables to consider have a p-value larger than some specified threshold, if added to the model. When we reach this state, forward selection will terminate and return a model that only contains variables with p-values $<$ threshold.

Example of Forward Elimination Criteria

- Stopping rule: If no variables have p value less than 0.05
- First start with intercept model
- Y=bo+ei
- Now add variables one by one
- Now fit model with one predictor variable at one time and check their P value independently

Apply lm command on every variable individually

M1=lm(y~x1)

summary(M1)

M2=lm(y~x2)

summary(M2)

M3=lm(y~x3)

summary(M3)

Example of Forward Elimination Criteria

```
M3=lm(y~x3)
summary(M3)
```

Model	P value
Y=b0+b1x1+ei	0.000779
Y=b0+b2x2+ei	0.009272
Y=bo+b3x3+ei	0.683
Y=b0+b4x4+ei	0.184

- Choose x1 variable as minimum p value

Model=Y=bo+b1x1+ei

Example of Forward Elimination Criteria

- Now to choose 2nd variable to be added in model

M1= lm(y~x1+x2)

summary(M1)

M2=lm(y~x1+x3)

summary(M2)

M3= lm(y~x1+x4)

summary(M3

Model	P value
M1=Y= bo+b1x1+b2x2+ei	0.000520
M2=Y=bo+b1x1+b3x3+ei	0.177711
M3=Y=bo+b1x1+b4x4+ei	0.181487

➢ Choose M1 model

Model=Y=b0+b1x1+b2x2+ei

Example of Forward Elimination Criteria

- Now to choose third variable;

M1=lm(y~x1+x2+x3)

summary(M1)

M2=lm(y~x1+x2+x4)

Summary(M2)

Model	P value
M1=Y =b0+b1x1+b2x2+b3x3+ei	0.149285
M2=Y =b0+b1x1+b2x2+b4x4+ei	0.6235

- We choose no model as both variables have P value>0.05 (stopping rules implies)

Final Forward Model = Y =b0+b1x1+b2x2+ei

Advantages and Disadvantages of Forward Elimination

- Unlike backward elimination, forward stepwise selection can used when the number of variables under consideration is very large, even larger than the sample size!
- This is because forward selection starts with a null model (with no predictors) and proceeds to add variables one at a time, and so unlike backward selection, it DOES NOT have to consider the full model (which includes all the predictors).

Advantages and Disadvantages of Forward Elimination

Remark: One problem with this approach is the estimation of σ^2 in the F tests.

- At start the MSE model estimate of σ^2 is likely to be higher in the early rounds of this process and then fall over time as more variables are added.
- A potential solutions would be use the full model MSE as the estimator of $sigma^2$ in all the F tests beginning with the first explanatory variable to be added to the null model.
- Another variation of the stepwise process is one that only has addition and not omission of existing variables, that is once an explanatory variable is added it cannot then be omitted later in the process.

Advantages of Backwards(Forwards) Elimination

1. The ability to manage large amounts of potential predictor variables, fine tuning the model to choose the best predictor variables from the available options.
2. It's faster than other model-selection methods.
3. Watching the order in which variables are removed or added can provide valuable information about the quality of the predictor variables.

Disadvantages of Backwards(Forwards) Elimination

1. If two predictor variables in the model are highly correlated, only one may make it into the model.
2. R^2 values are usually too high.
3. Adjusted R^2 values might be high, and then dip sharply as the model progresses. If this happens, identify the variables that were added or removed when this happens and adjust the model.
4. Predicted values and confidence intervals are too narrow.
5. Regression coefficients are bias and coefficients for other variables are too high.
6. Collinearity is usually a major issue. Excessive collinearity may cause the program to dump predictor variables into the model.
7. Some variables (especially dummy variables) may be removed from the model, when they are deemed important to be included. These can be manually added back in.

False positives in F tests

One of the main issues with automatic methods that rely on F (or t) tests

Risk that we fail to reject H_0: $\beta_j = 0$ for some j when we should have rejected it

Means we would include (or fail to eliminate) an explanatory variable whose parameter value was really zero (and therefore a variable with no statistically significant explanatory power)

Akaike's Information Criterion (AIC) can help with this

AIC uses some of the Maximum Likelihood methods of week 6

Akaike's Information Criteria (AIC)

$$AIC = 2(p + 1) - 2logL$$

Where:

- p = the number of regression parameters (so p – 1 explanatory variables)
- L is the Likelihood function evaluated at the maximum likelihood estimates of each of the parameters

Using AIC in model selection

We seek the regression model that minimises AIC

- because of the $-2logL$ this is equivalent to the model that maximises likelihood balanced against the number of parameters

- We can look to do this through backward or forward type processes

Backwards Elimination using AIC

1. Construct the full model and calculate its AIC
2. Construct all possible models that omit one variable
3. Calculate the AIC for each of these models
4. If the full model has lowest AIC use that and stop
5. If another has lowest AIC move on to that model and repeat
6. Stop once AIC cannot be lowered by removing a variable

Backwards using AIC in R

This process can be automated in R programming using the `step()` function

If the full model is constructed with `lm()` and stored as `full_model` (say)

A backwards route to the reduced model with lowest AIC is given by

```
reduced_model <- step(full_model, direction = "backward")
```

Forward regression alterative using AIC

We start with the null model

In R this is found by `null_model <- lm(y ~ 1)`

say we have six explanatory variables to consider `x1 x2 x3 x4 x5 x6`

This is done in R with

```
forward_model <- step(null_model, scope =
x1+x2+x3+x4+x5+x6, direction = "forward")
```

Different results

Backwards and Forwards methods using AIC may lead to different recommended models

A third alternative is to set `direction = "both"` inside `step()`

- This has the effect of adding additional variables from the null model and then later allowing deletion of one or more of those variables once others are added

Example

Hald's real data

Y	x_1	x_2	x_3	x_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

The real data set in this question first appeared in Hald (1952). The data are given in Table and can be found on the book web site in the file Haldcement.txt. Throughout this question we shall assume that the full model below is a valid model for the data $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$ Identify the optimal model based on R^2, AIC from the approach based on all possible subsets.

Subset size	Predictors	R^2_{adj}	AIC
1	X4	0.6450	58.8516
2	X1, X2	0.9744	25.4200
3	X1, X2, X4	0.9764	24.9739
4	X1, X2, X3, X4	0.9736	26.9443

Example

Backward elimination based on AIC

```
Start: AIC= 26.94
Y ~ x1 + x2 + x3 + x4
           DF     Sum of Sq        RSS        AIC
- x3        1         0.109     47.973     24.974
- x4        1         0.247     48.111     25.011
- x2        1         2.972     50.836     25.728
<none>                         47.864     26.944
- x1        1        25.951     73.815     30.576

Step: AIC= 24.97
Y ~ x1 + x2 + x4
           Df     Sum of Sq        RSS        AIC
<none>                         47.97      24.97
- x4        1          9.93     57.90      25.42
- x2        1         26.79     74.76      28.74
- x1        1        820.91     868.88     60.63
```

Example

```
Start: AIC= 29.77
Y ~ x1 + x2 + x3 + x4
           Df     Sum of Sq         RSS         AIC
- x3        1         0.109      47.973      27.234
- x4        1         0.247      48.111      27.271
- x2        1         2.972      50.836      27.987
<none>                          47.864      29.769
- x1        1        25.951      73.815      32.836

Step: AIC= 27.23
Y ~ x1 + x2 + x4
           Df     Sum of Sq         RSS         AIC
- x4        1          9.93       57.90       27.11
<none>                          47.97       27.23
- x2        1         26.79       74.76       30.44
- x1        1        820.91      868.88       62.32

Step: AIC= 27.11
Y ~ x1 + x2
           Df     Sum of Sq         RSS         AIC
<none>                          57.90       27.11
- x1        1        848.43      906.34       60.31
- x2        1       1207.78     1265.69       64.65
```

Example

Forward selection based on AIC

```
Start: AIC= 71.44
Y ~ 1
           Df     Sum of Sq         RSS         AIC
+ x4        1       1831.90      883.87       58.85
+ x2        1       1809.43      906.34       59.18
+ x1        1       1450.08     1265.69       63.52
+ x3        1        776.36     1939.40       69.07
<none>                         2715.76       71.44

Step: AIC= 58.85
Y ~ x4
           Df     Sum of Sq         RSS         AIC
+ x1        1        809.10       74.76       28.74
+ x3        1        708.13      175.74       39.85
<none>                          883.87       58.85
+ x2        1         14.99      868.88       60.63
```

Example

```
Step: AIC= 28.74
Y ~ x4 + x1
           Df     Sum of Sq         RSS         AIC
+ x2        1        26.789      47.973      24.974
+ x3        1        23.926      50.836      25.728
<none>                           74.762      28.742


Step: AIC= 24.97
Y ~ x4 + x1 + x2
           Df     Sum of Sq         RSS         AIC
<none>                           47.973      24.974
+ x3        1         0.109      47.864      26.944
```

Exams Style Questions

We use the Bridge.txt dataset available on QMPlus, where information from 45 bridge projects are compiled. The response and predictor variables are as follows:

- Y: Time is the design time in person-days;
- X_1: DArea is the deck area of bridge (000 sq ft);
- X_2: CCost is the construction cost ($000);
- X_3: Dwgs is the number of structural drawings;
- X_4: Length is the length of bridge (ft);
- X_5: Spans is the number of spans.

Take the logarithm transformation of all the variables.

(a) Use RStudio to find the best reduced model using the AIC procedure and state which is the best reduced model;

Exams Style Questions

```
> m1 <- lm(Y ~ X1 + X2 + X3 + X4 + X5)
> summary(m1)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68394 -0.17167 -0.02604  0.23157  0.67307

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.28590    0.61926   3.691 0.000681 ***
X1          -0.04564    0.12675  -0.360 0.720705
```

Exams Style Questions

```
X2           0.19609    0.14445   1.358 0.182426
X3           0.85879    0.22362   3.840 0.000440 ***
X4          -0.03844    0.15487  -0.248 0.805296
X5           0.23119    0.14068   1.643 0.108349
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3139 on 39 degrees of freedom
Multiple R-squared:  0.7762,Adjusted R-squared:  0.7475
F-statistic: 27.05 on 5 and 39 DF,  p-value: 1.043e-11
```

Exams Style Questions

```
> reduced.model <- step(m1, direction="backward")
Start:  AIC=-98.71
Y ~ X1 + X2 + X3 + X4 + X5

       Df Sum of Sq    RSS      AIC
- X4    1   0.00607 3.8497 -100.640
- X1    1   0.01278 3.8564 -100.562
<none>              3.8436  -98.711
- X2    1   0.18162 4.0252  -98.634
- X5    1   0.26616 4.1098  -97.698
- X3    1   1.45358 5.2972  -86.277

Step:  AIC=-100.64
Y ~ X1 + X2 + X3 + X5

       Df Sum of Sq    RSS      AIC
- X1    1   0.01958 3.8693 -102.412
<none>              3.8497 -100.640
- X2    1   0.18064 4.0303 -100.577
- X5    1   0.31501 4.1647  -99.101
- X3    1   1.44946 5.2991  -88.260
```

Exams Style Questions

```
Step:  AIC=-102.41
Y ~ X2 + X3 + X5

       Df Sum of Sq    RSS      AIC
<none>              3.8693 -102.412
- X2    1   0.17960 4.0488 -102.370
- X5    1   0.29656 4.1658 -101.089
- X3    1   1.44544 5.3147  -90.128
```

Thus, backward elimination based on AIC chooses the model with the three predictors X_2, X_3 and X_5, which are the logarithm of the construction cost; of the number of structural drawings and of the number of spans.

Based on the forward selection based on the AIC, arrives at the same model as backward elimination based on AIC.

Exams Style Questions

```
> modyn <- lm(Y ~ 1)
> aic.forward.model <- step(modyn, scope=~X1 + X2 + X3 + X4 + X5,
 direction="forward")
Start:  AIC=-41.35
Y ~ 1

       Df Sum of Sq     RSS     AIC
+ X3    1   12.1765  4.9975 -94.898
+ X2    1   11.6147  5.5593 -90.104
+ X1    1   10.2943  6.8797 -80.514
+ X4    1   10.0120  7.1620 -78.704
+ X5    1    8.7262  8.4478 -71.274
<none>              17.1740 -41.347

Step:  AIC=-94.9
Y ~ X3
```

Exams Style Questions

```
       Df Sum of Sq    RSS      AIC
+ X5    1   0.94866 4.0488 -102.370
+ X2    1   0.83170 4.1658 -101.089
+ X4    1   0.66914 4.3284  -99.366
+ X1    1   0.47568 4.5218  -97.399
<none>              4.9975  -94.898

Step:  AIC=-102.37
Y ~ X3 + X5

       Df Sum of Sq    RSS     AIC
+ X2    1  0.179598 3.8693 -102.41
<none>              4.0488 -102.37
+ X1    1  0.018535 4.0303 -100.58
+ X4    1  0.016924 4.0319 -100.56
```

Exams Style Questions

```
Step:  AIC=-102.41
Y ~ X3 + X5 + X2

       Df Sum of Sq    RSS     AIC
<none>              3.8693 -102.41
+ X1    1  0.019578 3.8497 -100.64
+ X4    1  0.012868 3.8564 -100.56
```

Thus in conclusion the best model is

$$Y = \beta_0 + \beta_1 X_3 + \beta_2 X_5 + \beta_3 X_2 + \varepsilon$$

where the variables are taken in logarithm.