STAT 408
Applied Regression Analysis

Nan Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

Model Selection

Model Selection Problem

- In practice, we often face different choices of model building
  - Which predictors should be included?
  - Should we apply any transformation on variables?

- We introduced F-test to compare two models with nested predictors and same response Y

- The F-test essentially measures if the decrease of RSS due to predictor increase is "large enough"

- However, the F-test only tells between two models which one is better; its doesn't tell which model is the "absolutely best one"

Criteria-Based Model Selection

- Criteria-Based model selection method defines some objective criteria and select the model that can maximize this criteria
- It provides an “absolute standard” to select a best combination of existing predictors
- Note that the predictor here is not limited to the original X, but includes transformations
- In the most general case, we can include the transformation of response Y
- Criteria-based model selection may or may not rely on statistical test

Step Methods

- Step methods use the p-value as criteria to decide if adding or removing one predictor

- Backward elimination
  1. Start with all the predictors in the model
  2. Remove the predictor with highest p-value greater than a cutoff c (e.g., 0.05)
  3. Refit the model and remove the remaining least significant predictor with p-value greater than the cutoff c
  4. Repeat the process until all “nonsignificant” predictors are removed

Step Methods

- Let’s illustrate the backward elimination method using “state” dataset, which includes the life expectancy and social-economic variables of 50 states in 1977

```
state <- read.csv('state.csv’)
```

	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area
UT	1203	4022	0.6	72.90	4.5	67.3	137	82096
AK	365	6315	1.5	69.31	11.3	66.7	152	566432
NV	590	5149	0.5	69.03	11.5	65.2	188	109889
CO	2541	4884	0.7	72.06	6.8	63.9	166	103766
WA	3559	4864	0.6	71.72	4.3	63.5	32	66570
WY	376	4566	0.6	70.29	6.9	62.9	173	97203

Step Methods

- We fit a model with life expectancy as response and all other variables as predictors

```
lm.model <- lm(Life.Exp~., data=state)
summary(lm.model)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.094e+01	1.748e+00	40.586	< 2e-16	***
Population	5.180e-05	2.919e-05	1.775	0.0832	.
Income	-2.180e-05	2.444e-04	-0.089	0.9293	
Illiteracy	3.382e-02	3.663e-01	0.092	0.9269	
Murder	-3.011e-01	4.662e-02	-6.459	8.68e-08	***
HS.Grad	4.893e-02	2.332e-02	2.098	0.0420	*
Frost	-5.735e-03	3.143e-03	-1.825	0.0752	.
Area	-7.383e-08	1.668e-06	-0.044	0.9649	

- We set up the cutoff p-value as 0.05
- Among all insignificant predictors, Area has the largest p-value and will be dropped first

Step Methods

- We fit the second model without Area

```
lm.model <- update(lm.model, .~.-Area)
summary(lm.model)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.099e+01	1.387e+00	51.165	< 2e-16	***
Population	5.188e-05	2.879e-05	1.802	0.0785	.
Income	-2.444e-05	2.343e-04	-0.104	0.9174	
Illiteracy	2.846e-02	3.416e-01	0.083	0.9340	
Murder	-3.018e-01	4.334e-02	-6.963	1.45e-08	***
HS.Grad	4.847e-02	2.067e-02	2.345	0.0237	*
Frost	-5.776e-03	2.970e-03	-1.945	0.0584	.

- Using the same cutoff p-value as 0.05, among current insignificant predictors, <u>Illiteracy</u> has the largest p-value and will be dropped second

Step Methods

- We fit the third model without Area and Illiteracy

lm.model <- update(lm.model, .~.-Illiteracy)

summary(lm.model)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.107e+01	1.029e+00	69.067	< 2e-16	***
Population	5.115e-05	2.709e-05	1.888	0.0657	.
Income	-2.477e-05	2.316e-04	-0.107	0.9153	
Murder	-3.000e-01	3.704e-02	-8.099	2.91e-10	***
HS.Grad	4.776e-02	1.859e-02	2.569	0.0137	*
Frost	-5.910e-03	2.468e-03	-2.395	0.0210	*

- Again, using the same cutoff p-value as 0.05, among current insignificant predictors, <u>Income</u> has the largest p-value and will be dropped third

Step Methods

- We repeat the same process by dropping two more predictors: income and population

```
lm.model <- update(lm.model, .~.-Income)
summary(lm.model)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.103e+01	9.529e-01	74.542	< 2e-16	***
Population	5.014e-05	2.512e-05	1.996	0.05201	.
Murder	-3.001e-01	3.661e-02	-8.199	1.77e-10	***
HS.Grad	4.658e-02	1.483e-02	3.142	0.00297	**
Frost	-5.943e-03	2.421e-03	-2.455	0.01802	*

```
lm.model <- update(lm.model, .~.-Population)
summary(lm.model)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	71.036379	0.983262	72.246	< 2e-16	***
Murder	-0.283065	0.036731	-7.706	8.04e-10	***
HS.Grad	0.049949	0.015201	3.286	0.00195	**
Frost	-0.006912	0.002447	-2.824	0.00699	**

- The final model includes three predictors, and they are all significant

Step Methods

- Forward Selection just reverses the backward elimination
    1. Start with no variables in the model (null model)
    2. For all predictors not in the model, choose the one with lowest p-value less than cutoff c (e.g., 0.05)
    3. Continue until no new predictors below the cutoff c can be added

Information Criteria Method

- Information criteria method tries to find a model that minimize the RSS and simultaneously constrains the model complexity

- Akaike information criteria (AIC) for linear model is defined as:

$$AIC = n * \log(RSS/n) + 2 * p$$

- We want to select a model to minimizes the AIC, which will make the first term smaller to improve the model fit, but at the same time would not be too complicated because of the second term

- AIC provides a balance between fit and simplicity in model selection

Information Criteria Method

- Bayes information criteria (BIC) is a modification of AIC by penalizing large model more

$$BIC = n * \log(RSS/n) + p * \log(n)$$

- The $\log(n)$ in the second term adds more penalty to more complex model - the larger data size will induce more penalty

Information Criteria Method

- Find the best predictor combinations that minimizes AIC or BIC can be computational expansive

- In state dataset, the number of models need to be compared is

$$C_7^1 + C_7^2 + C_7^3 + C_7^4 + C_7^5 + C_7^6 + C_7^7 = 127$$

- If there are 15 predictors, the number of models need to be compared is 32767

Step-Wise Criterion Method

- To avoid expansive full search, we can use a stepwise method
- Each time we remove one predict that will decrease AIC most
- We repeat the process until AIC no longer improves
- This is a “greedy” method – each time we look for a local optimum to get close to the global optimum

```
lm.model <- lm(Life.Exp~., data=state)
step(lm.model)
```

Step-Wise Criterion Method

```
Start:  AIC=-22.18
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost + Area

             Df Sum of Sq    RSS     AIC
- Area        1    0.0011 23.298 -24.182
- Income      1    0.0044 23.302 -24.175
- Illiteracy  1    0.0047 23.302 -24.174
<none>                    23.297 -22.185
- Population  1    1.7472 25.044 -20.569
- Frost       1    1.8466 25.144 -20.371
- HS.Grad     1    2.4413 25.738 -19.202
- Murder      1   23.1411 46.438  10.305

Step:  AIC=-24.18
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost

             Df Sum of Sq    RSS     AIC
- Illiteracy  1    0.0038 23.302 -26.174
- Income      1    0.0059 23.304 -26.170
<none>                    23.298 -24.182
- Population  1    1.7599 25.058 -22.541
- Frost       1    2.0488 25.347 -21.968
- HS.Grad     1    2.9804 26.279 -20.163
- Murder      1   26.2721 49.570  11.569
```

```
Step:  AIC=-26.17
Life.Exp ~ Population + Income + Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
- Income      1     0.006 23.308 -28.161
<none>                    23.302 -26.174
- Population  1     1.887 25.189 -24.280
- Frost       1     3.037 26.339 -22.048
- HS.Grad     1     3.495 26.797 -21.187
- Murder      1    34.739 58.041  17.456

Step:  AIC=-28.16
Life.Exp ~ Population + Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
<none>                    23.308 -28.161
- Population  1     2.064 25.372 -25.920
- Frost       1     3.122 26.430 -23.877
- HS.Grad     1     5.112 28.420 -20.246
- Murder      1    34.816 58.124  15.528

Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
    data = statedata)

Coefficients:
(Intercept)   Population       Murder      HS.Grad        Frost
  7.103e+01    5.014e-05   -3.001e-01    4.658e-02   -5.943e-03
```

- Each time we remove one predict that will decrease AIC most; the removing sequence is Area, Illiteracy, Income

One General Consideration

- It is always preferred to consider removing higher order terms first before removing lower order terms

- Suppose we fit a polynomial model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

and believe that β_1 is not significant but β_2 is

- If we remove x from the model, then we have

$$y = \beta_0 + \beta_2 x^2 + \varepsilon$$

One General Consideration

- Suppose we make a scale change $x \rightarrow x + a$, then the model with only x^2 would become

$$y = \beta_0 + \beta_2 a^2 + 2\beta_2 a x + \beta_2 x^2 + \varepsilon$$

- The first order term x reappears in the model, which means it is not robust to a small-scale change of our data
    - It also makes the model interpretation depend on the scale of the data

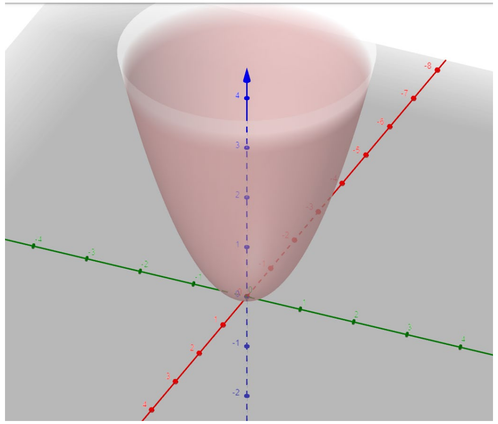
- Another issue of only keeping x^2 is that it forces the model to be symmetric at $x=0$, and response Y maximizes/minimizes at $x=0$
- Generally, there is no reason to make such strong model assumption

One General Consideration

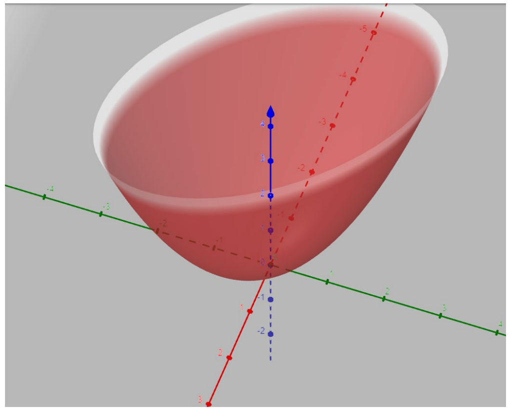
- Another consideration is to remove the quadratic term before interaction term in a second-order model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

- In this model, the interaction term $x_1 x_2$ shows the interactive impact of X_1 and X_2 on Y
- Removing the interaction term before the second order x_1 and x_2 will miss this information and force the surface aligned with the coordinate axes



$y = x_1^2 + x_2^2$



$y = x_1^2 + x_2^2 + x_1 x_1$

Summary

- F-test based model selection uses hypothesis test to compare two nested models
- F-test is a classical and “relative” comparison

- Criteria-based model selection select models that can optimize pre-defined criteria
- Criteria-based model selection is not restricted to nest models – it can compare any two models

- It is possible that different model selection methods provides different conclusions
- In that case, we need to consider the domain knowledge, interpretation, and prediction accuracy