Stat 206: Linear Models

Lecture 15

Nov. 20, 2019

Recap: Key Components for Model Selection

- **Criterion to compare models**:
  - R_a^2, C_p, AIC_p, BIC_p, $Press_p$, etc.
- **Procedure to search for good model(s):**
  - *Best subset selection*: Exhaustive search; When the number of potential X variables is not too big
  - *Stepwise regression*: Greedy search; The number of potential X variables can be large.

Surgical Unit: Model Selection Criteria

Consider X_1, X_2, X_3, X_4 (clotting, prognostic, enzyme, liver) as the potential pool of X variables. There are 16 sub-models.

p	intercept	X1	X2	X3	X4	sse	R^2	R^2_a	Cp	aic	bic	press
1	1	0	0	0	0	12.805	0.000	0.000	151.569	-75.716	-73.727	13.292
2	1	0	0	1	0	7.334	0.427	0.416	66.518	-103.811	-99.833	8.329
2	1	0	0	0	1	7.408	0.421	0.410	67.696	-103.268	-99.290	8.024
2	1	0	1	0	0	9.974	0.221	0.206	108.469	-87.205	-83.227	10.738
2	1	1	0	0	0	12.028	0.061	0.043	141.093	-77.096	-73.118	13.508
3	1	0	1	1	0	4.313	0.663	0.650	20.523	-130.479	-124.512	5.066
3	1	0	0	1	1	5.132	0.599	0.583	33.536	-121.089	-115.122	6.123
3	1	1	0	1	0	5.783	0.548	0.531	43.873	-114.644	-108.677	6.989
3	1	0	1	0	1	6.620	0.483	0.463	57.175	-107.342	-101.375	7.474
3	1	1	0	0	1	7.299	0.430	0.408	67.961	-102.070	-96.103	8.472
3	1	1	1	0	0	9.437	0.263	0.234	101.937	-88.194	-82.227	11.055
4	1	1	1	1	0	3.109	0.757	0.743*	3.388*	-146.161*	-138.205*	3.914*
4	1	0	1	1	1	3.615	0.718	0.701	11.434	-138.011	-130.055	4.598
4	1	1	0	1	1	4.970	0.612	0.589	32.960	-120.823	-112.867	6.209
4	1	1	1	0	1	6.568	0.487	0.456	58.358	-105.763	-97.807	7.902
5	1	1	1	1	1	3.084	0.759*	0.739	5.000	-144.587	-134.642	4.069

Within each subset size, models are sorted in ascending SSE. Consequently, within each subset size, R_p^2, $R_{a,p}^2$ are from the largest to the smallest and C_p, BIC_p, AIC_p are from the smallest to the largest. $Press_p$ may not be monotone with SSE.

AIC_p and BIC_p Criteria

- *Akaike's information criterion (AIC)*:

$$AIC_p = n \log \frac{SSE_p}{n} + 2p.$$

- *Bayesian information criterion (BIC)*:

$$BIC_p = n \log \frac{SSE_p}{n} + (\log n)p.$$

- **We should look for models with small AIC (BIC).**
  - Surgical unit. The model with X_1, X_2, X_3 has the smallest AIC and BIC among the models being considered.

- The first term: $n \log \frac{SSE_p}{n}$ reflects the of the model to the observed data.
  - It by adding more X variables into the model.
- The second term, $2p$ for AIC and $(\log n)p$ for BIC, reflects .
  - It by adding more X variables into the model.
  - If $n \geq 8$, then $\log n > 2$ and BIC puts penalty on model complexity and tends to choose models than AIC.

- The first term: $n \log \frac{SSE_p}{n}$ reflects the *goodness-of-fit* of the model to the **observed data**.
    - It decreases by adding more X variables into the model.
- The second term, $2p$ for AIC and $(\log n)p$ for BIC, reflects model complexity.
    - It increases by adding more X variables into the model.
    - If $n \geq 8$, then $\log n > 2$ and BIC puts more penalty on model complexity and tends to choose smaller models than AIC.

- Overly simplified models have model complexity (p), but they tend to have SSE (underfitting; high).
- Overly complicated models may have a SSE, but they have model complexity (overfitting, high).
- By minimizing AIC (or BIC), we are trying to find a model that between model complexity and the goodness-of-fit.

- Overly simplified models have small model complexity (p), but they tend to have large SSE (underfitting; high bias).
- Overly complicated models may have a small SSE, but they have large model complexity (overfitting, high variance).
- By minimizing AIC (or BIC), we are trying to find a model that balances between model complexity and the goodness-of-fit.

$Press_p$ Criterion

Predicted residual sum of squares ($Press_p$):

$$Press_p = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{i(i)})^2.$$

- Y_i is the observed response of the *ith* case.
- $\widehat{Y}_{i(i)}$ is the predicted value for the ith case obtained by fitting the model only using $n-1$ cases excluding case i.
- $Press_p$ is also known as *leave-one-out-cross-validation (LOOCV)*.
- Models with small $Press_p$ are considered good in terms of predictive ability.
  - Surgical unit: the model with X_1, X_2, X_3 has $Press_p = 3.914$ which is the smallest among all models being considered here.

Calculate $Press_p$

$Press_p$ can be calculated without actually performing n regressions.

- This is because the *deleted residual* for the *ith* case:

$$d_i := Y_i - \widehat{Y}_{i(i)} = \qquad\qquad\qquad\qquad , \quad i = 1, \cdots, n.$$

  where $e_i = Y_i - \widehat{Y}_i$ is the residual of the *ith* case and h_{ii} is the *ith* diagonal element of the hat matrix $\mathbf{H}$, both from the regression fit using .
- So

Calculate $Press_p$

$Press_p$ can be calculated without actually performing n regressions.

- This is because the *deleted residual* for the *ith* case:

$$d_i := Y_i - \widehat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}, \quad i = 1, \cdots, n.$$

  where $e_i = Y_i - \widehat{Y}_i$ is the residual of the *ith* case and h_{ii} is the *ith* diagonal element of the hat matrix $\mathbf{H}$, both from the regression fit using **all** n cases.
- So

$$Press_p = \sum_{i=1}^{n} \frac{(Y_i - \widehat{Y}_i)^2}{(1 - h_{ii})^2}.$$

Derive the Deleted Residuals

Optional Reading.

- Define $\tilde{\mathbf{Y}}$ by replacing the ith element of the response vector $\mathbf{Y}$ with the leave-i-out predicted value $\hat{Y}_{i(i)}$ of the ith case:

$$\tilde{\mathbf{Y}} = (Y_1, \cdots, Y_{i-1}, \hat{Y}_{i(i)}, Y_{i+1}, \cdots, Y_n)^T.$$

- Let $\hat{\beta}_{(i)}$ be the leave-i-out LS fitted regression coefficients. Then $\hat{\beta}_{(i)}$ is also the LS fitted regression coefficients by using $\tilde{\mathbf{Y}}$ as the response vector, i.e. $\hat{\beta}_{(i)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\tilde{\mathbf{Y}}$. *Why?*
- The leave-i-out fitted values are:

$$\hat{\mathbf{Y}}_{(i)} = \mathbf{X}\hat{\beta}_{(i)} = H\tilde{\mathbf{Y}} = H(\mathbf{d}_{(i)}+\mathbf{Y}), \quad \mathbf{d}_{(i)} = \tilde{\mathbf{Y}}-\mathbf{Y} = (0, \cdots, -d_i, \cdots, 0)^T.$$

- Subtracting the ith element from Y_i on both sides gives:

$$d_i = h_{ii}d_i + e_i \Longrightarrow d_i = \frac{e_i}{1-h_{ii}}.$$

Surgical Unit: Full Model X_1, X_2, X_3, X_4

```
> fit.f =lm(log(Y)~X1+X2+X3+X4, data=data.o)
> summary(fit.f)
Call:
lm(formula = log(Y) ~ X1 + X2 + X3 + X4, data = data.o)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.851933   0.266263  14.467  < 2e-16 ***
X1          0.083739   0.028834   2.904  0.00551 **
X2          0.012671   0.002315   5.474 1.50e-06 ***
X3          0.015627   0.002100   7.440 1.38e-09 ***
X4          0.032056   0.051466   0.623  0.53627
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
Residual standard error: 0.2509 on 49 degrees of freedom
Multiple R-squared: 0.7591,     Adjusted R-squared: 0.7395
F-statistic: 38.61 on 4 and 49 DF,  p-value: 1.398e-14
> anova(fit.f)
Analysis of Variance Table

Response: log(Y)
Df Sum Sq Mean Sq  F value    Pr(>F)
X1          1 0.7770  0.7770  12.3443 0.0009618 ***
X2          1 2.5904  2.5904  41.1565 5.341e-08 ***
X3          1 6.3286  6.3286 100.5490 1.838e-13 ***
X4          1 0.0244  0.0244   0.3879 0.5362698
Residuals 49 3.0841  0.0629
```

Surgical Unit: Full Model

- Full model has $P = 5$ and

  $SSE = 3.0841, \; MSE = 0.0629, \; R^2 = 0.7591, \; R_a^2 = 0.7395.$

- By definition, for the full model, $C_P = P = 5$.
- Sample size $n = 54$, so for the full model:
  $AIC_P = 54\log(3.0841/54) + 2 \times 5 = -144.5871$ and
  $BIC_P = 54\log(3.0841/54) + \log(54) \times 5 = -134.6422$.
- $Press_p = 4.069$.

```
> e.f=fit.f$residuals  ## residuals
> h.f=influence(fit.f)$hat  ## diagonals of hat matrix
> press.f= sum(e.f^2/(1-h.f)^2)  ## calculate press
```

Model Search Procedures

- The number of possible models, 2^{P-1}, grows very fast with the number potential X variables $P - 1$.
- Evaluating every possible model can be computationally infeasible even for moderate P.
- A variety of search procedures have been developed to efficiently search for the "best" model(s) in the model space.
    - **Stepwise regression procedures**
    - Best subsets algorithms: Not applicable when the pool of potential X variables is large.

Stepwise Regression Procedures

- Applicable to situations with a large number of potential X variables.
- Use “greedy” search strategies by developing a sequence of models, at each step adding or deleting only one X variable according to a pre-specified criterion (e.g., *AIC*).
- May end up with a *suboptimal model* rather than the global “best” model.
- Commonly used stepwise procedures include: *forward stepwise*, *forward selection*, *backward stepwise* and *backward elimination*.

Forward Stepwise Procedure

Need to specify:

- A model selection criterion, e.g., *AIC*.
- An initial model $\mathbb{M}_0$, usually a small model, e.g., the null-model with no X variable.
- The pool of potential X variables $\mathcal{X}$.
- The set of X variables that will always be in the model $\mathcal{X}_0$, e.g., the intercept term.

Starting from the initial model $\mathbb{M}_0$, at each step:

(a) Consider the X variables in the potential pool $\mathcal{X}$ that are not currently in the model. Examine the change in the criterion by adding each such variable into the current model.

(b) Consider the X variables that are already in the model but not in the set $\mathcal{X}_0$. Examine the change in the criterion by dropping each such variable out of the current model.

- Choose the operation that improves the criterion the most and update the current model accordingly. Repeat steps (a) and (b) for the updated model.
- If there is no operation that can improve the criterion anymore, then stop the search procedure and return the current model as the selected model.

Forward Selection and Backward Elimination

- Forward selection is a simplified version of forward stepwise procedure, omitting the considerations of dropping a variable currently in the model at each step.
- Backward elimination is the opposite of the forward selection.
  - It starts with a "big" initial model, e.g., the full model.
  - At each step, it examines the change of the criterion by dropping a variable currently in the model.
- Backward stepwise procedure. *Guess what is it?*
- Another commonly used strategy is to perform one pass of forward selection followed by one pass of backward elimination.

`stepAIC ()` Function

We can use the `stepAIC()` function in the `MASS` library to perform various stepwise regression.

- `direction=‘‘both"` corresponds to forward stepwise procedure or backward stepwise procedure (depending on the initial model); `direction=‘‘forward"` corresponds to froward selection; `direction=‘‘backward"` corresponds to backward elimination.
- The option `scope` specifies the potential pool of X variables (`upper`) and the X variables that should always be included in the model (`lower`).
- `k=2` corresponds to *AIC* criterion; `k=log(n)` corresponds to *BIC* criterion.

Surgical Unit: Forward Stepwise

Reading material:
Start with the null-model.

```
> library(MASS)
> fit.0 =lm(log(Y)~1, data=data.o)  ## initial model: null-model with only intercept term
> step.0=stepAIC(fit.0,scope=list(upper=~X1+X2+X3+X4+X5+X6+X7+X8, lower=~1), direction="both", k=2)
Start:  AIC=-75.72
log(Y) ~ 1
Df Sum of Sq     RSS      AIC
+ X3     1    5.4708  7.3337 -103.811
+ X4     1    5.3967  7.4079 -103.268
+ X2     1    2.8303  9.9742  -87.205
+ X8     1    1.7808 11.0238  -81.802
+ X1     1    0.7770 12.0275  -77.096
+ X6     1    0.6889 12.1156  -76.703
<none>               12.8045  -75.716
+ X5     1    0.2694 12.5351  -74.864
+ X7     1    0.2067 12.5978  -74.595

Step:  AIC=-103.81
log(Y) ~ X3
Df Sum of Sq     RSS      AIC
+ X2     1    3.0209  4.3129 -130.479
+ X4     1    2.2018  5.1319 -121.089
+ X1     1    1.5512  5.7825 -114.644
+ X8     1    1.1386  6.1951 -110.922
<none>                7.3337 -103.811
+ X6     1    0.2582  7.0755 -103.747
+ X5     1    0.2390  7.0947 -103.600
+ X7     1    0.0659  7.2679 -102.298
- X3     1    5.4708 12.8045  -75.716
```

Surgical Unit: Forward Stepwise (Cont'd)

```
Step:  AIC=-130.48
log(Y) ~ X3 + X2

Df Sum of Sq     RSS      AIC
+ X8    1     1.4709 2.8420 -151.002
+ X1    1     1.2044 3.1085 -146.161
+ X4    1     0.6979 3.6150 -138.011
+ X7    1     0.2280 4.0849 -131.412
+ X5    1     0.1648 4.1481 -130.583
<none>               4.3129 -130.479
+ X6    1     0.0822 4.2306 -129.518
- X2    1     3.0209 7.3337 -103.811
- X3    1     5.6613 9.9742  -87.205

Step:  AIC=-151
log(Y) ~ X3 + X2 + X8
Df Sum of Sq     RSS      AIC
+ X1    1     0.6642 2.1778 -163.376
+ X4    1     0.4658 2.3761 -158.669
+ X6    1     0.1372 2.7048 -151.674
<none>               2.8420 -151.002
+ X5    1     0.0709 2.7711 -150.367
+ X7    1     0.0241 2.8179 -149.462
- X8    1     1.4709 4.3129 -130.479
- X2    1     3.3531 6.1951 -110.922
- X3    1     4.9403 7.7823  -98.605
```

Surgical Unit: Forward Stepwise (Cont'd)

```
Step:  AIC=-163.38
log(Y) ~ X3 + X2 + X8 + X1

Df Sum of Sq    RSS      AIC
+ X6     1    0.0966 2.0812 -163.826
<none>               2.1778 -163.376
+ X5     1    0.0760 2.1018 -163.293
+ X4     1    0.0415 2.1363 -162.415
+ X7     1    0.0224 2.1554 -161.935
- X1     1    0.6642 2.8420 -151.002
- X8     1    0.9307 3.1085 -146.161
- X2     1    2.9891 5.1670 -118.722
- X3     1    5.4459 7.6237  -97.717

Step:  AIC=-163.83
log(Y) ~ X3 + X2 + X8 + X1 + X6
Df Sum of Sq    RSS     AIC
+ X5     1    0.0769 2.0043 -163.86
<none>               2.0812 -163.83
- X6     1    0.0966 2.1778 -163.38
+ X7     1    0.0219 2.0593 -162.40
+ X4     1    0.0163 2.0649 -162.25
- X1     1    0.6236 2.7048 -151.67
- X8     1    0.9754 3.0567 -145.07
- X2     1    2.8287 4.9099 -119.48
- X3     1    5.0742 7.1554  -99.14
```

Surgical Unit: Forward Stepwise (Cont'd)

```
Step:  AIC=-163.86
log(Y) ~ X3 + X2 + X8 + X1 + X6 + X5
Df Sum of Sq     RSS      AIC
<none>                2.0043 -163.858
- X5     1     0.0769 2.0812 -163.826
- X6     1     0.0975 2.1018 -163.293
+ X7     1     0.0326 1.9718 -162.743
+ X4     1     0.0022 2.0021 -161.919
- X1     1     0.6284 2.6327 -151.133
- X8     1     0.9011 2.9054 -145.810
- X2     1     2.7644 4.7688 -119.052
- X3     1     5.0752 7.0795  -97.716

> step.0$anova
Stepwise Model Path
Analysis of Deviance Table
Initial Model:
log(Y) ~ 1
Final Model:
log(Y) ~ X3 + X2 + X8 + X1 + X6 + X5
Step Df   Deviance Resid. Df Resid. Dev        AIC
1                         53  12.804509  -75.71608
2 + X3  1 5.47078352      52   7.333726 -103.81102
3 + X2  1 3.02085553      51   4.312870 -130.47855
4 + X8  1 1.47089284      50   2.841977 -151.00214
5 + X1  1 0.66416961      49   2.177808 -163.37593
6 + X6  1 0.09659084      48   2.081217 -163.82569
7 + X5  1 0.07688125      47   2.004335 -163.85826
```

Surgical Unit: Forward Stepwise (Cont'd)

- The selected model is $X_1, X_2, X_3, X_5, X_6, X_8$ ($p = 7$) with $AIC_p = -163.858$.
- In this case, the forward selection procedure also selects the same model.

```
## forward selection
> step.0.f=stepAIC(fit.0,  scope=list(upper=~X1+X2+X3+X4+X5+X6+X7+X8, lower=~1),
+ direction="forward", k=2)
```

Surgical Unit: Backward Elimination

Start with the full model with all eight predictors.

```
> fit.f =lm(log(Y)~., data=data.o)
> step.b=stepAIC(fit.f, scope= list(upper=~., lower=~1), direction="backward", k=2)
Start:  AIC=-160.78
log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8

Df Sum of Sq     RSS     AIC
- X4     1    0.00126 1.9718 -162.74
- X7     1    0.03159 2.0021 -161.92
- X5     1    0.07359 2.0441 -160.80
<none>                1.9705 -160.78
- X6     1    0.08403 2.0545 -160.52
- X1     1    0.31845 2.2890 -154.69
- X8     1    0.84489 2.8154 -143.51
- X2     1    2.09285 4.0634 -123.70
- X3     1    2.98863 4.9591 -112.94
```

Surgical Unit: Backward Elimination (Cont'd)

```
Step:  AIC=-162.74
log(Y) ~ X1 + X2 + X3 + X5 + X6 + X7 + X8
Df Sum of Sq     RSS       AIC
- X7     1     0.0326 2.0043 -163.858
<none>               1.9718 -162.743
- X5     1     0.0876 2.0593 -162.396
- X6     1     0.0969 2.0687 -162.152
- X1     1     0.6269 2.5987 -149.835
- X8     1     0.8438 2.8156 -145.506
- X2     1     2.6755 4.6473 -118.446
- X3     1     5.0934 7.0652  -95.825
Step:  AIC=-163.86
log(Y) ~ X1 + X2 + X3 + X5 + X6 + X8
Df Sum of Sq     RSS       AIC
<none>               2.0043 -163.858
- X5     1     0.0769 2.0812 -163.826
- X6     1     0.0975 2.1018 -163.293
- X1     1     0.6284 2.6327 -151.133
- X8     1     0.9011 2.9054 -145.810
- X2     1     2.7644 4.7688 -119.052
- X3     1     5.0752 7.0795  -97.716

## backward stepwise
> step.bs=stepAIC(fit.f, scope= list(upper=~., lower=~1),
+ direction="both", k=2)
```

Again the model $X_1, X_2, X_3, X_5, X_6, X_8$ is selected. Backward stepwise also selects the same model.

Stepwise Procedures: Comments

- Forward stepwise procedure often works better than forward selection when there is .
- Backward procedures are not good when the number of potential X variables, $P - 1$, is . Particularly, they are not feasible when P n, since then the full model can not be fitted.
- A potential disadvantage of forward procedures is the MSE and thus the standard errors of the LS estimators tend to be in the initial steps due to

Stepwise Procedures: Comments

- Forward stepwise procedure often works better than forward selection when there is high multicollinearity.
- Backward procedures are not good when the number of potential X variables, $P - 1$, is large. Particularly, they are not feasible when $P > n$, since then the full model can not be fitted.
- A potential disadvantage of forward procedures is the MSE and thus the standard errors of the LS estimators tend to be overestimated in the initial steps due to underfitting since important X variables are likely to be omitted in those steps.

Model Building: Comments

For the sake of interpretability:

- It is often appropriate to select all the indicator variables corresponding to a qualitative variable as a group (i.e., to be in or out of the model simultaneously).
- **Hierarchical principle**: If higher-order terms (e.g., interactions, powers) are selected, it is often appropriate to include the related lower-order terms as well.