STA302/1001 - Methods of Data Analysis I

(Week 06 lecture B)

Wei (Becky) Lin

June 19-23, 2017

UNIVERSITY OF TORONTO

Last Lecture

- Type I and Type III SS
- Use of Extra Sum of Squares in Tests for Regression coefficients
- Coefficient of Partial Determination
- Summary of Tests Concerning Regression Coefficients
- Multicollinearity and Its Effects

Week 6 Lectuer B- Learning objectives & Outcomes

- More on multicollinearity
- Model selection
- Final review.

How to detect multicollinearity

Some of the common methods used for detecting multicollinearity include:

- The analysis exhibits the signs of multicollinearity — such as, estimates of the coefficients vary from model to model.
- The t-tests for each of the individual slopes are non-significant ($P > 0.05$), but the overall F-test for testing all of the slopes are simultaneously 0 is significant ($P < 0.05$).
- The correlations among pairs of predictor variables are large. (Looking at pairwise correlation is limiting, e.g. $X_1 = 1 + 2X_2 + 5X_5$, a linear dependence exists among three or even more variables)
- Variance Inflation Factor

$$VIF_k = \frac{Var(b_k|X_1, \ldots, X_k, \ldots)}{Var(b_k|X_k)}$$

Reference: `https://onlinecourses.science.psu.edu/stat501/node/347`

Solutions to multicollinearity

- If interest is only in mean response estimation and prediction, multicollinearity can be ignored since it does not affect $\hat{Y}$ or its standard error (either $\widehat{Var}(\hat{Y})$ or $\widehat{Var}(\hat{Y} - Y)$).
    - True only if the x_h at which we want estimation or prediction is within the range of the data.
- If the wish is to establish association patters between y and the predictors, then analyst can:
    - Eliminate some predictors from the model.
    - Design an experiment in which the pattern of correlation is broken.
    - Using centered predictor variables in polynomial regression.
        - $x = 2, 3, 4, 5, 6$ and $x^2 = 4, 9, 16, 25, 36$, $cor(x, x^2) = 0.98$.
        - $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$.
        - $z = x - \bar{x} = -2, -1, 0, 1, 2; z^2 = 4, 1, 0, 1, 4$, $cor(z, z^2) = 0$
        - $E(Y) = \gamma_0 + \gamma_1 z + \gamma_2 z^2$.

Model Selection

In general

How to compare two non-nested models?

- Bias - variance trade-off
- C_p
- AIC
- Cross-validation
- BIC

How to search the space of possible models?

- Step-wise search
- Best subets.
- LASSO
- Bayesian methods

Model Selection - AIC

Definition of AIC: *Akaike Information Criterion (or AIC)* for a model M is defined as

$$AIC(M) = n\log(SSE_M/n) + 2(p_M + 1)$$

where p_M is the number of predictors in the model.

- Want to minimize the Kullback-Leibler distance ($p(y)$ is the true model for y)

$$K(p, \hat{p}_j) = \int p(y)\log\frac{p(y)}{\hat{p}_j(y;\theta)} = \int p(y)\log p(y)dy - \int p(y)\log\hat{p}_j(y;\theta)dy$$

  - same as maximizing $K_j = \int p(y)\log\hat{p}_j(y;\theta)dy$
  - a good estimate of K_j is $\bar{K}_j = \frac{1}{n}\sum_i^n \log P(y_i;\hat{\theta}_j) = \frac{1}{n}\ell_j(\hat{\theta}_j)$
  - Akaike showed that the bias of $\bar{K}_j$ is about $\approx d_j/n$ where d_j =dim(parameters),therefore

$$\hat{K}_j = \ell_j(\hat{\theta}_j)/n - 2d_j$$

- In R, the functions AIC() and extractAIC().In AIC, set k=log(n) gives BIC value.
- AIC is the most commonly used evaluator. It is used in the R command *step*()
- The model with the smaller AIC is considered better.

Model Selection - BIC

Definition of BIC: *Bayesian Information Criterion (or BIC)* for a model M is defined as

$$BIC(M) = n \log(SSE_M/n) + \log(n)(p_M + 1)$$

- We put a prior $\pi_j(\theta_j)$ on parameter θ_j, and a prior p_j that M_j is the true model.

$$P(M_j|Y_1, \ldots, Y_n) \propto P(Y_1, \ldots, Y_n|M_j)P_j = \int L(\theta_j)\pi_j(\theta_j)d\theta_j$$

- We choose j to maximize

$$\log \int L(\theta_j)\pi_j(\theta_j)d\theta_j + \log p_j$$

- Taylor series approximations show that

$$\log \int L(\theta_j)\pi_j(\theta_j)d\theta_j + \log p_j \approx \ell_j(\hat{\theta}_j) - \frac{d_j}{2}\log n = BIC_j$$

- In contrast to AIC, $BIC_M \geq AIC_M$ when $n > 7.3 (= \exp(2))$.
- Puts more penalty for having more predictors and chooses simpler model than AIC.
- The model with the smaller BIC is considered better..

Model selection - Mallows' C_p

If p predictor variables are selected from a set of $K > p$, the Mallows C_p statistic for that particular set of X's is defined as:

$$C_p = SSE_p + 2p\hat{\sigma}^2$$

- SSE_p is the residual sum of squares on a training set of data.
- p is the number of predictor variables.
- $\hat{\sigma}^2$ is the estimate of $\sigma^2 = Var(Y)$ using K predictor variables.
- Usual practice: plot C_p versus p, choose model with minimum.
- The model with the smaller C_p is considered better..

Model selection - Cross-validation

Idea: we divide up the data by training sample and a testing sample. The training sample is used to fit the regression model while the testing sample is used to test how accurate the model is.
K-fold CV

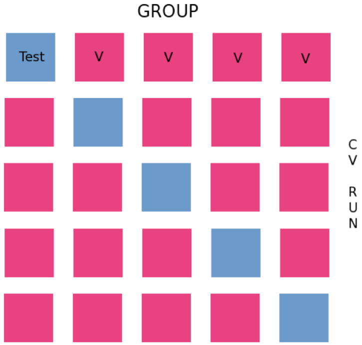
- Randomly divide your observations into K parts. Each part should have roughly n/K observations.
- For each part
  - Define this part to be your testing sample.
  - Define all other parts to be your training sample.
  - Fit the model using only the training sample
  - Compute the prediction MSE, denote as PMSE

$$PMSE = \frac{1}{\text{size of testing samples}} \sum_{i \in \text{testing sample}} (Y_i - \hat{Y}_i)^2$$

- Take the average of the K PMSE computed in the loop.

Model selection - Cross-validation (contd.)

- 5-fold CV



- For possible models in consideration, we compute the K-fold CV and obtain the PMSE for each model. We generally choose the model with the smallest $\overline{PMSE}$. In practice, K is chosen to be 5, although it can depend on the initial sample size.

Model selection procedure - All-subset selection

All-subset Selection

- Suppose we have p X's and we want to choose the best subset $X_1, \ldots X_p$ by a criterion.
- Try every subset of $X_1, \ldots X_p$. There will be a total of 2^p subsets because that each X can either in or out of the model.
- Using the information criterion outlined before and choose the subset with the lowest value.
- Usual practice: this procedure is only practical for small p.
- The R command regsubsets() in the leaps package implements this procedure.

Example: all-subset selection

```r
data=read.table("/Users/Wei/TA/Teaching/STA302-2016F/Week12-Nov28/BostonHousing.txt",
                sep=",",header=T)
n <- dim(data)[1]
head(data)
```

```
##      CRIM ZN INDUSTRY CHAR   NOX NROOMS  AGE    DIS RAD TAX PTRATIO      B
## 1 0.00632 18     2.31    0 0.538  6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0     7.07    0 0.469  6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0     7.07    0 0.469  7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0     2.18    0 0.458  6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0     2.18    0 0.458  7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0     2.18    0 0.458  6.430 58.7 6.0622   3 222    18.7 394.12
##   LSAT MEDV
## 1 4.98 24.0
## 2 9.14 21.6
## 3 4.03 34.7
## 4 2.94 33.4
## 5 5.33 36.2
## 6 5.21 28.7
```

```r
### Run All-Subsets ###
library(leaps)
model.allsubsets = regsubsets(log(MEDV) ~ INDUSTRY + NROOMS + AGE + TAX + CRIM,data=data)
```

Example: all-subset selection (contd.)

```
summary(model.allsubsets) # Provides a summary of which X is in the model
# Output #
# Each row corresponds to the best subset of Xs for p number of independent variables
# Selection Algorithm: exhaustive
#          INDUSTRY NROOMS AGE TAX CRIM
# 1  ( 1 ) " "      "*"    " " " " " "
# 2  ( 1 ) " "      "*"    " " " " "*"
# 3  ( 1 ) " "      "*"    " " "*" "*"
# 4  ( 1 ) " "      "*"    "*" "*" "*"
# 5  ( 1 ) "*"      "*"    "*" "*" "*"

summary(model.allsubsets)$cp #Provides Mallow's Cp for each p.
# Output #
# Each value corresponds to the Cp value for each p.
# Choose the value with the lowest Cp, in our case, p =4.
# [1] 283.905095  77.475928  23.309122   4.494921   6.000000

summary(model.allsubsets)$bic #Provides BIC for each p
# Output #
# Each value corresponds to the BIC value for p.
# Choose the value with the lowest BIC, in our case p= 4
# [1] -245.5618 -395.2689 -440.8181 -455.2089 -449.4830

# More details about summary(model.allsubsets) can be found by typing
?summary.regsubsets
```

Model selection procedure - Forward Selection

1. Start with the most parsimonious model $Y_i = \beta_0 + \epsilon_i$
2. For the current model
   - For each X_k that is left, add it to the model and perform an F test comparing the current model (i.e. the reduced model) with the model that includes X_k (i.e. the full model)
   - Choose the X_j with the lowest p-value (or largest F observed value)
   - If this p-value is lower than a pre-specified significance level (e.g. $\alpha = 0.05$), include it into the model, declare this as your current model, and repeat the procedure. Otherwise, declare the current model as your final model.

Example: Forward Selection

```r
data=read.table("/Users/Wei/TA/Teaching/STA302-2016F/Week12-Nov28/BostonHousing.txt",
                sep=",",header=T)

### Run forward-selection ###
currentmod = lm(log(MEDV)~1,data=data )
add1(currentmod,~INDUSTRY + NROOMS + AGE + TAX + CRIM,test="F",data=data)
```

```
## Single term additions
##
## Model:
## log(MEDV) ~ 1
##          Df Sum of Sq    RSS      AIC F value    Pr(>F)
## <none>                84.376  -904.37
## INDUSTRY  1    24.746 59.630 -1078.02  209.16 < 2.2e-16 ***
## NROOMS    1    33.704 50.672 -1160.39  335.23 < 2.2e-16 ***
## AGE       1    17.347 67.029 -1018.83  130.43 < 2.2e-16 ***
## TAX       1    26.599 57.777 -1093.99  232.03 < 2.2e-16 ***
## CRIM      1    23.518 60.858 -1067.70  194.76 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# and repeat the procedure
# currentmod = lm(log(MEDV) ~ NROOMS)
# add1( currentmod,~INDUSTRY  + AGE + TAX + CRIM,test="F",data=data)
#...
```

Example: Forward Selection (contd.)

```
data=read.table("/Users/Wei/TA/Teaching/STA302-2016F/Week12-Nov28/BostonHousing.txt",
                sep=",",header=T)
### Run automatic forward-selection ###
# no predictor in the model
nullmod= lm (log(MEDV)~1, data=data)
# with all predictors in the model
fullmod = lm( log(MEDV)~INDUSTRY  + AGE + TAX,data=data)
# forward selection method
forwards = step(nullmod ,scope=list(lower=formula(nullmod),
                upper=formula(fullmod)), direction="forward")
formula(forwards)
# will NOT get you the same results since steps() automatically uses AIC, not F tests!
```

Model selection procedure - Backward Elimination

1. Start with the least parsimonious model
   $Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon_i$
2. For the current model
   - For each X_k that's in the current model, drop it from the model and perform an F test comparing the current model (i.e. the full model) with the model that includes X_k (i.e. the reduced model)
   - Choose the X_j with the largest p-value (or smallest F observed value)
   - If this p-value is larger than a pre-specified significance level (e.g. $\alpha = 0.05$), remove X_j it into the model, declare this as your current model, and repeat the procedure. Otherwise, declare the current model as your final model.

Example: Backward selection

```r
data=read.table("/Users/Wei/TA/Teaching/STA302-2016F/Week12-Nov28/BostonHousing.txt",
                sep=",",header=T)

### Run backward-selection ###
currentmod =lm(log(MEDV) ~ INDUSTRY + NROOMS + AGE + TAX,data=data )
drop1(currentmod,test="F",data=data)
```

```
## Single term deletions
##
## Model:
## log(MEDV) ~ INDUSTRY + NROOMS + AGE + TAX
##          Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## <none>                35.864 -1329.3
## INDUSTRY  1    0.0001 35.864 -1331.3   0.0021    0.9636
## NROOMS    1   17.4624 53.327 -1130.5 243.9374 < 2.2e-16 ***
## AGE       1    1.3351 37.199 -1312.8  18.6507 1.893e-05 ***
## TAX       1    4.4342 40.299 -1272.3  61.9429 2.200e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Choose the X (INDUSTRY) with the highest p-value or the lowest F value.
# without INDUSTRY and repeat the procedure
# model.current = lm(log(MEDV) ~  NROOMS + AGE + TAX,data=data)
# drop1( model.current,test="F",data=data)
# ...
```

Example: Backward Elimination (contd.)

```
data=read.table("/Users/Wei/TA/Teaching/STA302-2016F/Week12-Nov28/BostonHousing.txt",
                sep=",",header=T)

### Run automatic backforward-elimination ###
# no predictor in the model
nullmod= lm (log(MEDV)~1, data=data)
# with all predictors in the model
fullmod = lm( log(MEDV)~INDUSTRY  + AGE + TAX + CRIM,data=data)
# forward selection method
backward = step(fullmod ,scope=list(lower=formula(nullmod),
                upper=formula(fullmod)), direction="backward")
```

```
## Start:  AIC=-1177.23
## log(MEDV) ~ INDUSTRY + AGE + TAX + CRIM
##
##             Df Sum of Sq    RSS     AIC
## <none>                   48.436 -1177.2
## - AGE        1    0.7017 49.137 -1172.0
## - TAX        1    0.8831 49.319 -1170.1
## - INDUSTRY   1    1.5685 50.004 -1163.1
## - CRIM       1    4.8912 53.327 -1130.5
```

```
formula(backward)
```

```
## log(MEDV) ~ INDUSTRY + AGE + TAX + CRIM
```

```
# will NOT get you the same results since steps() automatically uses AIC, not F tests!
```

Model selection procedure - Stepwise Selection

1. Start with the some model. In R, this usually is the least parsimonious model $Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon_i$
2. For the current model, compute the information criterion.
    - Consider all the variables that can be removed. For each of the variables removed, compute the information criterion.
    - Consider all the variables that can be added. For each of the variables added, compute the information criterion.
    - From the models formulated by removing or adding predictors, choose the model with the smallest information criterion.
    - If the information criterion associated with this model has a smaller value than current model, replace the current model with this one. Otherwise, declare the current model as the final model.

Example: Stepwise Selection

```
data=read.table("/Users/Wei/TA/Teaching/STA302-2016F/Week12-Nov28/BostonHousing.txt",
                sep=",",header=T)

### Run automatic backforward-elimination ###
# no predictor in the model
nullmod= lm (log(MEDV)~1, data=data)
# with all predictors in the model
fullmod = lm( log(MEDV)~INDUSTRY  + AGE + TAX,data=data)
# stepwise selection method using AIC
stepwisemod = step(fullmod ,scope=list(lower=formula(nullmod),
                upper=formula(fullmod)), direction="both")
```

```
## Start:  AIC=-1130.55
## log(MEDV) ~ INDUSTRY + AGE + TAX
##
##             Df Sum of Sq    RSS     AIC
## <none>                   53.327 -1130.5
## - AGE        1    1.1608 54.488 -1121.7
## - INDUSTRY   1    1.2069 54.534 -1121.2
## - TAX        1    4.7344 58.061 -1089.5
```

```
formula(stepwisemod)
```

```
## log(MEDV) ~ INDUSTRY + AGE + TAX
```

```
# will NOT get you the same results since steps() automatically uses AIC, not F tests!
```

Example: Stepwise Selection (contd.)

```r
data=read.table("/Users/Wei/TA/Teaching/STA302-2016F/Week12-Nov28/BostonHousing.txt",
                sep=",",header=T)

### Run automatic Stepwise-Selection ###
nullmod= lm (log(MEDV)~1, data=data)
fullmod = lm( log(MEDV)~INDUSTRY  + AGE + TAX,data=data)
# Stepwise selection method using F-test
stepwisemod = step(fullmod ,scope=list(lower=formula(nullmod),
                upper=formula(fullmod)), direction="both",test="F")
```

```
## Start:  AIC=-1130.55
## log(MEDV) ~ INDUSTRY + AGE + TAX
##
##            Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                  53.327 -1130.5
## - AGE       1    1.1608 54.488 -1121.7  10.928 0.0010151 **
## - INDUSTRY  1    1.2069 54.534 -1121.2  11.361 0.0008077 ***
## - TAX       1    4.7344 58.061 -1089.5  44.568 6.522e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
formula(stepwisemod)
```

```
## log(MEDV) ~ INDUSTRY + AGE + TAX
```

Example: Stepwise Selection (contd.)

```
data=read.table("/Users/Wei/TA/Teaching/STA302-2016F/Week12-Nov28/BostonHousing.txt",
                sep=",",header=T)

### Run automatic Stepwise-Selection ###
nullmod= lm (log(MEDV)~1, data=data)
fullmod = lm( log(MEDV)~INDUSTRY  + AGE + TAX,data=data)
# Stepwise selection method using BIC
stepwisemod = step(fullmod ,scope=list(lower=formula(nullmod),
                upper=formula(fullmod)), direction="both",k=log(n))
```

```
## Start:  AIC=-1113.64
## log(MEDV) ~ INDUSTRY + AGE + TAX
##
##            Df Sum of Sq    RSS     AIC
## <none>                  53.327 -1113.6
## - AGE       1    1.1608 54.488 -1109.0
## - INDUSTRY  1    1.2069 54.534 -1108.5
## - TAX       1    4.7344 58.061 -1076.8
```

```
formula(stepwisemod)
```

```
## log(MEDV) ~ INDUSTRY + AGE + TAX
```

That's all ! You are..

Final Exam

- Cover page and Formula page (check out portal)
- Coverage on entire term from lecture 1 to 12 (very little on lecture 12)
- Questions type: multiple choice, short answer, proofs, data analysis
- 25% on proofs (more on MLR)

Final Exam

Suggestions:

- Practise all proofs in slides and extra assigned questions: All of them.
- Know how to read and interpret R output
  - Summary, anova output
  - Diagnostics plots
  - Other output that you have seen from slides.
- Review 3 assignments and midterm paper (both sections)
- Do some old exams might help you to see which topics are important

Final Exam: topics that you could skip

- Week 1 -Lect B:
  - different criterion of regression: reverse/orthogonal/reduced major axis regression.
  - How to find MLE
  - Review on inference
- Week 2 - Lect B: Normal correlation model
- Week 4 - lect B: Review on matrices (slide 6-16): skip only if you have good knowledge of matrices
- Week 5 - Lect B: Geometric perspective of LS regression(slides 4-6).
- Week 6 - Lect A: Type III SS.
- Week 6 - Lect B: Model selection.

GOOD LUCK!