STA 108 Discussion 9: Variable selection

Yuanyuan Li

Reference: Textbook Chapter 8, 9.

1. Regression Models for Quantitative and Qualitative Predictors

Let use the *iris* data as an example. *iris* is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species. Which are Quantitative variables, and which are predictor variables?

```
library(knitr)
kable(summary(iris))
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	

The response variable is always a quantitative variable (numeric variable) in linear regression model. Let use "Sepal.Length" as the response variable, and consider others as predictors.

1.1 Polynonrial Regression Models for quantitative predictors

If x_1 is a quantitative predictor, the second-order model is $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,1}^2 + \epsilon_i$. `I()` function in R protects the higher-order terms as separate terms.

```
lm(Sepal.Length~Sepal.Width+ I(Sepal.Width^2), data=iris)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width + I(Sepal.Width^2), data = iris)
##
## Coefficients:
##      (Intercept)       Sepal.Width  I(Sepal.Width^2)
##           2.4594            2.4312            -0.4246
```

1.2 Models for qualititave predictors

If x is a qualitative predictor with k categories, one can define $k-1$ indicator variables as follows:

$$x_1 = \begin{cases} 1, & \text{if category 1,} \\ 0, & \text{otherwise} \end{cases}, x_2 = \begin{cases} 1, & \text{if category 2,} \\ 0, & \text{otherwise} \end{cases}, \cdots, x_{k-1} = \begin{cases} 1, & \text{if category k-1,} \\ 0, & \text{otherwise} \end{cases}.$$

For example, "Species" has 3 categories: setosa, versicolor, virginica. R will use the first category as the baseline level, then created 2 indicator(dummy) variables. The model can be expressed as $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{k-1} x_{i,k-1} + \epsilon_i$.

```
lm(Sepal.Length~ Species, data=iris)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Species, data = iris)
##
## Coefficients:
##       (Intercept)  Speciesversicolor   Speciesvirginica
##             5.006              0.930              1.582
```

Interpretation of coefficients: Comparing with setosa, the expected/mean Sepal Length of versicolor will increase 0.930; Comparing with setosa, the expected/mean Sepal Length of virginica will increase 1.582.

1.3 Models with Interation terms

Interpretation of regression coefficients: Express the model as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$, it can be shown that the change in the mean response with a unit increase in X_1 when X_2 is held constant is: $\beta_1 + \beta_3 X_2$. Similarly, the change in the mean response with a unit increase in X_2 when X_1 is held constant is: $\beta_2 + \beta_3 X_1$. Hence, in regression model with interaction term both the effect of X_1 for given level of X_2 and the effect of X_2 for given level of X_1 depend on the level of the other predictor variable.

- 2 quantitative predictors:

```
#lm(Sepal.Length~ Sepal.Width+Petal.Length+Sepal.Width:Petal.Length, data=iris)
lm(Sepal.Length~ Sepal.Width*Petal.Length, data=iris)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width * Petal.Length, data = iris)
##
## Coefficients:
##              (Intercept)               Sepal.Width              Petal.Length
##                  1.40438                  0.84996                  0.71846
## Sepal.Width:Petal.Length
##                 -0.07701
```

- quantitative v.s. qualititive predictors:

```r
lm(Sepal.Length~ Sepal.Width*Species, data=iris)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width * Species, data = iris)
##
## Coefficients:
##                   (Intercept)                   Sepal.Width
##                        2.6390                        0.6905
##             Speciesversicolor                Speciesvirginica
##                        0.9007                        1.2678
## Sepal.Width:Speciesversicolor   Sepal.Width:Speciesvirginica
##                        0.1746                        0.2110
```

- 2 qualitative predictors:

```r
#create one more qualitative predictor
Sepal.Width_class=factor(ifelse(iris$Sepal.Width>3,1,0))#=1 when Sepal.Width>3; otherwise,0
lm(Sepal.Length~Sepal.Width_class*Species, data=iris)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width_class * Species, data = iris)
##
## Coefficients:
##                            (Intercept)                    Sepal.Width_class1
##                                 4.6375                                0.4387
##                      Speciesversicolor                       Speciesvirginica
##                                 1.1935                                1.8383
## Sepal.Width_class1:Speciesversicolor   Sepal.Width_class1:Speciesvirginica
##                                 0.2179                               -0.1086
```

2. Regression Variable Selection

```r
library(MASS)
colnames(swiss)
```

```
## [1] "Fertility"        "Agriculture"      "Examination"      "Education"
## [5] "Catholic"         "Infant.Mortality"
```

```r
mod1 <- lm(Fertility ~ Agriculture + Examination, data = swiss)
mod2 <- lm(Fertility ~ Agriculture + Examination+Catholic, data = swiss)
```

2.1 R^2

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Larger R^2 shows a more accurate fit to the data. However, R^2 can never decrease as additional X variables are included in the model. Hence, R^2 is only used to compare models with same number of variables (parameters).

```
RS1=summary(mod1)$r.squared
RS2=summary(mod2)$r.squared
kable(data.frame(Model=1:2, Rsquare=c(RS1,RS2)))
```

Model	Rsquare
1	0.4326045
2	0.4460681

2.2 Adjusted R^2

$$R_a^2 = 1 - \frac{MSE}{MSTO},$$

where $MSTO = SSTO/(n-1)$, $MSE = SSE/(n-p)$. Select the model that maximizes R_a^2, or, equivalently, the model that minimizes MSE. Using this criterion, since when adding a new variable, both SSE and $n-p$ will decrease, only the variables that can bring a larger decrease of SSE(that can offset the decrease of $n-p$) will be selected into the model.

Relation with R^2: $R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} = 1 - \frac{SSE}{SSTO} \cdot \frac{n-1}{n-p} = 1 - (1-R^2) \cdot \frac{n-1}{n-p}$.

```
ARS1=summary(mod1)$adj.r.squared
ARS2=summary(mod2)$adj.r.squared
kable(data.frame(Model=1:2,Adj.Rsquare=c(ARS1,ARS2)))
```

Model	Adj.Rsquare
1	0.4068138
2	0.4074217

Preference : mod2 > mod1.

2.3 Mallows' C_p

If P regressors are selected from a set of $K > P$ potential predictors, the C_p statistic for that particular set of regressors is defined as:

$$C_p = \frac{SSE_p}{MSE_K} - (n - 2p),$$

where SSE_p is the SSE of fitting the regression with $x_1, \cdots, x_{p-1}$ being the predictors.

The best subset of predictors (model) corresponds to the one such that C_p is small and close to p.

2.4 AIC and BIC(SBC)

Both are based on balancing the model fitness and its complexity:

$$AIC = n\log(SSE_p/n) + 2p,$$
$$BIC = n\log(SSE_p/n) + \log(n)p.$$

Choose a subset of predictors (model) that minimizes AIC or BIC.

```
AIC1=AIC(mod1)
BIC1=BIC(mod1)
AIC2=AIC(mod2)
BIC2=BIC(mod2)
kable(data.frame(Model=1:2, AIC=c(AIC1,AIC2),BIC=c(BIC1,BIC2)))
```

Model	AIC	BIC
1	351.0906	358.4912
2	351.9619	361.2127

Both AIC and BIC pick "Model 1".

2.5 All possible(best) subset selection

If we have p predictors, a naive procedure would be to check all the possible models that can be constructed with them and then select the best one in terms of adjusted R^2/Cp/BIC/AIC. This exhaustive search is the so-called best subset selection. To automatically run the procedure, we can use the `regsubsets()` function in the R package `leaps`. `nbest` is the number of the best subsets of each size to save. If `nbest=1`, only the best model will be saved for each number of predictors. In the `summary()` output, the following objects that can be printed:

- `which`: A logical matrix indicating which predictors are in each model. 1 indicates a variable is included and 0 not.
- `rsq`: The r-squared for each model (higher, better)
- `adjr2`: Adjusted r-squared (higher, better)
- `cp`: Mallows' Cp (smaller, better)
- `bic`: Schwartz's Bayesian information criterion, BIC (lower, better)
- `rss`: Residual sum of squares(SSE) for each model (lower, better)

```
library(leaps)
#just like lm(), you can use "Fertility~Agriculture*Examination" to include interaction terms;
#"Fertility~." means regressing on all predictors in the full model
all<-regsubsets(Fertility~., data=swiss,nbest=1)
#Choose the best 3 models out of all subset models that can minimize Mallows'Cp
best3 = order(summary(all)$cp)[1:3]
# predictors and Cp of best 3 models; =1 if chosen, 0 not chosen
kable(cbind(summary(all)$which[best3,], Cp=summary(all)$cp[best3]))
```

	(Intercept)	Agriculture	Examination	Education	Catholic	Infant.Mortality	Cp
4	1	1	0	1	1	1	5.032800
5	1	1	1	1	1	1	6.000000
3	1	0	0	1	1	1	8.178162

2.6 Stepwise selection

The problem of all-subset selection is that there are 2^p possible models to inspect! For example, you need to fit and calculate the criterion values for $2^{10} = 1024$ models if you have 10 predictors.

`MASS::stepAIC` function helps us navigating this ocean of models by implementing stepwise model selection. Stepwise selection will iteratively add predictors that decrease an information criterion and/or remove those that increase it, depending on the mode of stepwise search that is performed.

- Forward: starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.

```
min.model=lm(Fertility ~1, data = swiss)
max.model <- lm(Fertility ~., data = swiss)
frd.model=stepAIC(min.model,direction = "forward",
                  scope = list(lower = min.model, upper = max.model))
```

```
## Start:  AIC=238.35
## Fertility ~ 1
##
##                    Df Sum of Sq    RSS    AIC
## + Education         1    3162.7 4015.2 213.04
## + Examination       1    2994.4 4183.6 214.97
## + Catholic          1    1543.3 5634.7 228.97
## + Infant.Mortality  1    1245.5 5932.4 231.39
## + Agriculture       1     894.8 6283.1 234.09
## <none>                         7178.0 238.34
##
## Step:  AIC=213.04
## Fertility ~ Education
##
##                    Df Sum of Sq    RSS    AIC
## + Catholic          1    961.07 3054.2 202.18
## + Infant.Mortality  1    891.25 3124.0 203.25
## + Examination       1    465.63 3549.6 209.25
## <none>                         4015.2 213.04
## + Agriculture       1     61.97 3953.3 214.31
##
## Step:  AIC=202.18
## Fertility ~ Education + Catholic
##
##                    Df Sum of Sq    RSS    AIC
## + Infant.Mortality  1    631.92 2422.2 193.29
## + Agriculture       1    486.28 2567.9 196.03
## <none>                         3054.2 202.18
## + Examination       1      2.46 3051.7 204.15
##
## Step:  AIC=193.29
## Fertility ~ Education + Catholic + Infant.Mortality
##
##                Df Sum of Sq    RSS    AIC
## + Agriculture   1   264.176 2158.1 189.86
## <none>                      2422.2 193.29
## + Examination   1     9.486 2412.8 195.10
```

```
##
## Step:  AIC=189.86
## Fertility ~ Education + Catholic + Infant.Mortality + Agriculture
##
##               Df Sum of Sq    RSS    AIC
## <none>                     2158.1 189.86
## + Examination  1    53.027 2105.0 190.69
```

- Backward: starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.

```
brd.model=stepAIC(max.model,direction = "backward")
```

```
## Start:  AIC=190.69
## Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## - Examination       1     53.03 2158.1 189.86
## <none>                          2105.0 190.69
## - Agriculture       1    307.72 2412.8 195.10
## - Infant.Mortality  1    408.75 2513.8 197.03
## - Catholic          1    447.71 2552.8 197.75
## - Education         1   1162.56 3267.6 209.36
##
## Step:  AIC=189.86
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## <none>                          2158.1 189.86
## - Agriculture       1    264.18 2422.2 193.29
## - Infant.Mortality  1    409.81 2567.9 196.03
## - Catholic          1    956.57 3114.6 205.10
## - Education         1   2249.97 4408.0 221.43
```

- Stepwise(or sequential replacement): a combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection).

```
step.model <- stepAIC(max.model, direction = "both")
```

```
## Start:  AIC=190.69
## Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## - Examination       1     53.03 2158.1 189.86
## <none>                          2105.0 190.69
## - Agriculture       1    307.72 2412.8 195.10
## - Infant.Mortality  1    408.75 2513.8 197.03
## - Catholic          1    447.71 2552.8 197.75
```

```
## - Education         1   1162.56 3267.6 209.36
##
## Step:  AIC=189.86
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##
##                     Df Sum of Sq    RSS    AIC
## <none>                           2158.1 189.86
## + Examination       1     53.03 2105.0 190.69
## - Agriculture       1    264.18 2422.2 193.29
## - Infant.Mortality  1    409.81 2567.9 196.03
## - Catholic          1    956.57 3114.6 205.10
## - Education         1   2249.97 4408.0 221.43
```