Faraway\_ch10

Harry Woo

2020 5 29

## Faraway Chapter 10 Variable Selection

### Problem 1

*Use the prostate data with lpsa as the response and the other variables as predictors. Implement the following variable selection methods to determine the “best” model:*

library(faraway)  
library(ggplot2)  
library(GGally)  
library(dplyr)  
library(knitr)  
library(MASS)  
library(car)  
library(tibble)  
library(gridExtra)  
library(leaps)

### Data loading and EDA

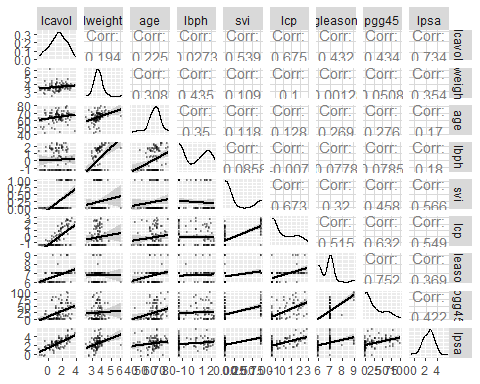
data("prostate")  
str(prostate)

## 'data.frame': 97 obs. of 9 variables:  
## $ lcavol : num -0.58 -0.994 -0.511 -1.204 0.751 ...  
## $ lweight: num 2.77 3.32 2.69 3.28 3.43 ...  
## $ age : int 50 58 74 58 62 50 64 58 47 63 ...  
## $ lbph : num -1.39 -1.39 -1.39 -1.39 -1.39 ...  
## $ svi : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ lcp : num -1.39 -1.39 -1.39 -1.39 -1.39 ...  
## $ gleason: int 6 6 7 6 6 6 6 6 6 6 ...  
## $ pgg45 : int 0 0 20 0 0 0 0 0 0 0 ...  
## $ lpsa : num -0.431 -0.163 -0.163 -0.163 0.372 ...

summary(prostate)

## lcavol lweight age lbph   
## Min. :-1.3471 Min. :2.375 Min. :41.00 Min. :-1.3863   
## 1st Qu.: 0.5128 1st Qu.:3.376 1st Qu.:60.00 1st Qu.:-1.3863   
## Median : 1.4469 Median :3.623 Median :65.00 Median : 0.3001   
## Mean : 1.3500 Mean :3.653 Mean :63.87 Mean : 0.1004   
## 3rd Qu.: 2.1270 3rd Qu.:3.878 3rd Qu.:68.00 3rd Qu.: 1.5581   
## Max. : 3.8210 Max. :6.108 Max. :79.00 Max. : 2.3263   
## svi lcp gleason pgg45   
## Min. :0.0000 Min. :-1.3863 Min. :6.000 Min. : 0.00   
## 1st Qu.:0.0000 1st Qu.:-1.3863 1st Qu.:6.000 1st Qu.: 0.00   
## Median :0.0000 Median :-0.7985 Median :7.000 Median : 15.00   
## Mean :0.2165 Mean :-0.1794 Mean :6.753 Mean : 24.38   
## 3rd Qu.:0.0000 3rd Qu.: 1.1786 3rd Qu.:7.000 3rd Qu.: 40.00   
## Max. :1.0000 Max. : 2.9042 Max. :9.000 Max. :100.00   
## lpsa   
## Min. :-0.4308   
## 1st Qu.: 1.7317   
## Median : 2.5915   
## Mean : 2.4784   
## 3rd Qu.: 3.0564   
## Max. : 5.5829

ggpairs(prostate,   
 lower = list(continuous = wrap("smooth", alpha = 0.3, size = 0.1)),  
 diag = list(discrete="barDiag",   
 continuous = wrap("densityDiag", alpha = 0.5 )),  
 upper = list(combo = wrap("box\_no\_facet", alpha = 0.5),  
 continuous = wrap("cor", size = 4,  
 alignPercent = 0.8))) +  
 theme(legend.position = "bottom")



#### Criterion-based procedures

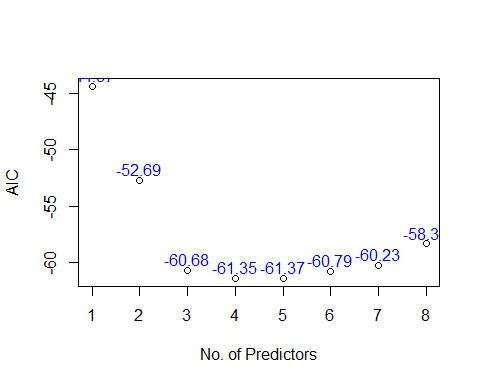
#### (a) AIC

pst\_rs <- regsubsets(lpsa ~ ., data = prostate)  
pst\_rss <- summary(pst\_rs)  
pst\_rss$which

## (Intercept) lcavol lweight age lbph svi lcp gleason pgg45  
## 1 TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 2 TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## 3 TRUE TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE  
## 4 TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE FALSE  
## 5 TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE  
## 6 TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE TRUE  
## 7 TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE  
## 8 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

leaps 패키지의 regsubsets 문을 활용하여 설명변수들의 모든 가능한 조합을 검토하여 주어진 모델 크기에서 최적의 RSS를 갖는 변수들을 찾아내어, 이를 pst\_rss로 저장하였다.

#linear regression 에서 -2 max log-likelihood는 nlog(RSS/n) + 2p(penalty term)  
  
pst\_n <- nrow(prostate)  
pst\_AIC <- pst\_n \* log(pst\_rss$rss / pst\_n) + (2:9) \* 2  
plot(pst\_AIC ~ I(1:8), ylab = "AIC", xlab = "No. of Predictors")  
text(1:8, pst\_AIC + 1, labels = round(pst\_AIC, 2), col = "blue", cex= 1)



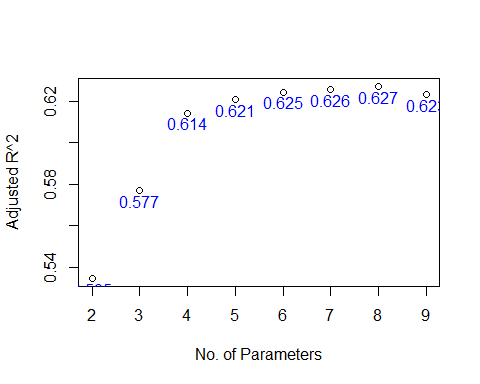
which.min(pst\_AIC)

## [1] 5

Akaike Information Crieteria(AIC)를 변수선택 기준으로 활용할 때는 여러 모형 중 AIC가 가장 낮은 모형이 최적 모형으로 선택된다. lpsa\_rss 자료를 기준으로, 설명변수의 조합에 따른 AIC 값을 계산하여 표로 나타낸 결과, 5개의 변수(lcavol, lweight, age, lbph, svi)를 사용했을 때 AIC 값이 최저임을 확인할 수 있다.

#### (b) Adjusted R2

plot(2:9, pst\_rss$adjr2, xlab = "No. of Parameters", ylab = "Adjusted R^2")  
text(2:9, pst\_rss$adjr2 - 0.005, labels = round(pst\_rss$adjr2, 3), col = "blue", cex= 1)



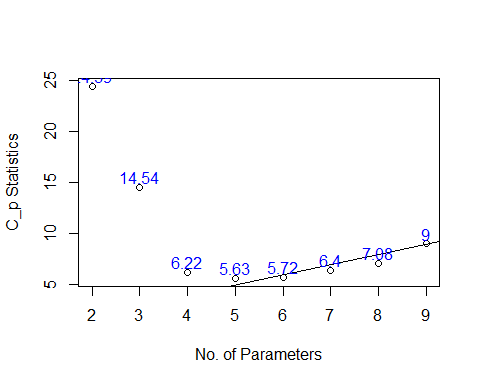
which.max(pst\_rss$adjr2) #결과 7 -> 변수 7개 조합을 의미

## [1] 7

모형을 선택함에 있어서 MSE를 최소화한다는 것은 곧 수정결정계수를 최대화하는 것과 동일한 기준을 갖는다고 할 수 있다. 이에 따라, pst\_rss로부터 수정결정계수가 최대로 나타내는 모형을 찾아낸 결과, 7개의 변수(lcavol, lweight, age, lbph, svi, lcp, gleason, pgg45)를 선택했을 때 Adjusted R^2 값이 최대가 되는 최적모형임을 확인할 수 있다.

#### (c) Mallows Cp

# C\_p = SSE\_p / MSE + 2p - n  
  
plot(2:9, pst\_rss$cp, xlab = "No. of Parameters", ylab = "C\_p Statistics")  
abline(0, 1)  
text(2:9, pst\_rss$cp + 1, labels = round(pst\_rss$cp, 2), col = "blue", cex= 1)



pst\_rss$cp

## [1] 24.394559 14.541475 6.216935 5.626422 5.715016 6.401965 7.082184  
## [8] 9.000000

which.min(pst\_rss$cp)

## [1] 4

모든 변수를 포함했을 때의 full model을 기준으로, p개의 설명변수를 포함하는 모형의 fit을 검토하기 위해 MSE의 n개의 합을 이용하는 C\_p 통계량을 활용할 수 있다. 가급적 낮은 C\_p 값이 더 선호되며 C\_p가 parameter의 개수 p에 근사한 모델이 일반적으로 선택된다.

Mallows C\_p를 통해 검토한 결과, 변수 4개 모델과 변수 5개 모델 사이의 competition이 관찰된다. C\_p가 p 이상인 모델에서 변수 4개 모델(lcavol, lweight, lbph, svi)이 가장 낮은 C\_p 값을 가지며, age가 추가된 변수 5개 모델(lcavol, lweight, age, lbph, svi)은 C\_p가 p+1 보다 작고 거의 p에 근사하다는 점에서 두 모델 모두 최적모델로 고려할 수 있을 것으로 판단된다.

#### (d) Forward selection method

#step 문을 활용한 전진 선택법  
  
pst\_lm\_null <- lm(lpsa ~ 1, data = prostate) #  
pst\_lm\_full <- lm(lpsa ~ ., data = prostate)  
  
step(pst\_lm\_null, scope = list(lower = pst\_lm\_null, upper = pst\_lm\_full),  
 data = prostate, direction = "forward")

## Start: AIC=28.84  
## lpsa ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + lcavol 1 69.003 58.915 -44.366  
## + svi 1 41.011 86.907 -6.658  
## + lcp 1 38.528 89.389 -3.926  
## + pgg45 1 22.814 105.103 11.783  
## + gleason 1 17.416 110.501 16.641  
## + lweight 1 16.041 111.876 17.840  
## + lbph 1 4.136 123.782 27.650  
## + age 1 3.679 124.238 28.007  
## <none> 127.918 28.837  
##   
## Step: AIC=-44.37  
## lpsa ~ lcavol  
##   
## Df Sum of Sq RSS AIC  
## + lweight 1 5.9485 52.966 -52.690  
## + svi 1 5.2375 53.677 -51.397  
## + lbph 1 3.2658 55.649 -47.898  
## + pgg45 1 1.6980 57.217 -45.203  
## <none> 58.915 -44.366  
## + lcp 1 0.6562 58.259 -43.453  
## + gleason 1 0.4156 58.499 -43.053  
## + age 1 0.0025 58.912 -42.370  
##   
## Step: AIC=-52.69  
## lpsa ~ lcavol + lweight  
##   
## Df Sum of Sq RSS AIC  
## + svi 1 5.1814 47.785 -60.676  
## + pgg45 1 1.9489 51.017 -54.327  
## <none> 52.966 -52.690  
## + lcp 1 0.8371 52.129 -52.236  
## + gleason 1 0.7810 52.185 -52.131  
## + lbph 1 0.6751 52.291 -51.935  
## + age 1 0.4200 52.546 -51.463  
##   
## Step: AIC=-60.68  
## lpsa ~ lcavol + lweight + svi  
##   
## Df Sum of Sq RSS AIC  
## + lbph 1 1.30006 46.485 -61.352  
## <none> 47.785 -60.676  
## + pgg45 1 0.57347 47.211 -59.847  
## + age 1 0.40251 47.382 -59.497  
## + gleason 1 0.38901 47.396 -59.469  
## + lcp 1 0.06412 47.721 -58.806  
##   
## Step: AIC=-61.35  
## lpsa ~ lcavol + lweight + svi + lbph  
##   
## Df Sum of Sq RSS AIC  
## + age 1 0.95924 45.526 -61.374  
## <none> 46.485 -61.352  
## + pgg45 1 0.35332 46.131 -60.092  
## + gleason 1 0.21256 46.272 -59.796  
## + lcp 1 0.10230 46.383 -59.565  
##   
## Step: AIC=-61.37  
## lpsa ~ lcavol + lweight + svi + lbph + age  
##   
## Df Sum of Sq RSS AIC  
## <none> 45.526 -61.374  
## + pgg45 1 0.65896 44.867 -60.789  
## + gleason 1 0.45601 45.070 -60.351  
## + lcp 1 0.12927 45.396 -59.650

##   
## Call:  
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age, data = prostate)  
##   
## Coefficients:  
## (Intercept) lcavol lweight svi lbph age   
## 0.95100 0.56561 0.42369 0.72095 0.11184 -0.01489

summary(lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age, data = prostate))

##   
## Call:  
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age, data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.83505 -0.39396 0.00414 0.46336 1.57888   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.95100 0.83175 1.143 0.255882   
## lcavol 0.56561 0.07459 7.583 2.77e-11 \*\*\*  
## lweight 0.42369 0.16687 2.539 0.012814 \*   
## svi 0.72095 0.20902 3.449 0.000854 \*\*\*  
## lbph 0.11184 0.05805 1.927 0.057160 .   
## age -0.01489 0.01075 -1.385 0.169528   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7073 on 91 degrees of freedom  
## Multiple R-squared: 0.6441, Adjusted R-squared: 0.6245   
## F-statistic: 32.94 on 5 and 91 DF, p-value: < 2.2e-16

step 문을 활용한 전진선택법을 실시하였다. 절편만을 포함한 모형 pst\_lm\_null에서 시작하여, 설명변수를 하나씩 추가하였을 때 AIC 값이 가장 작은 설명변수를 선택하고, 검정결과가 유의할 때까지 설명변수를 추가한 결과, 5개의 변수(lcavol, lweight, svi, lbph, age)를 포함하는 모형이 최종 모델로 선택되었다.

#### (e) Stepwise Selection method

step(pst\_lm\_null, scope = list(upper = pst\_lm\_full), data = prostate,  
 direction = "both")

## Start: AIC=28.84  
## lpsa ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + lcavol 1 69.003 58.915 -44.366  
## + svi 1 41.011 86.907 -6.658  
## + lcp 1 38.528 89.389 -3.926  
## + pgg45 1 22.814 105.103 11.783  
## + gleason 1 17.416 110.501 16.641  
## + lweight 1 16.041 111.876 17.840  
## + lbph 1 4.136 123.782 27.650  
## + age 1 3.679 124.238 28.007  
## <none> 127.918 28.837  
##   
## Step: AIC=-44.37  
## lpsa ~ lcavol  
##   
## Df Sum of Sq RSS AIC  
## + lweight 1 5.949 52.966 -52.690  
## + svi 1 5.237 53.677 -51.397  
## + lbph 1 3.266 55.649 -47.898  
## + pgg45 1 1.698 57.217 -45.203  
## <none> 58.915 -44.366  
## + lcp 1 0.656 58.259 -43.453  
## + gleason 1 0.416 58.499 -43.053  
## + age 1 0.003 58.912 -42.370  
## - lcavol 1 69.003 127.918 28.837  
##   
## Step: AIC=-52.69  
## lpsa ~ lcavol + lweight  
##   
## Df Sum of Sq RSS AIC  
## + svi 1 5.181 47.785 -60.676  
## + pgg45 1 1.949 51.017 -54.327  
## <none> 52.966 -52.690  
## + lcp 1 0.837 52.129 -52.236  
## + gleason 1 0.781 52.185 -52.131  
## + lbph 1 0.675 52.291 -51.935  
## + age 1 0.420 52.546 -51.463  
## - lweight 1 5.949 58.915 -44.366  
## - lcavol 1 58.910 111.876 17.840  
##   
## Step: AIC=-60.68  
## lpsa ~ lcavol + lweight + svi  
##   
## Df Sum of Sq RSS AIC  
## + lbph 1 1.3001 46.485 -61.352  
## <none> 47.785 -60.676  
## + pgg45 1 0.5735 47.211 -59.847  
## + age 1 0.4025 47.382 -59.497  
## + gleason 1 0.3890 47.396 -59.469  
## + lcp 1 0.0641 47.721 -58.806  
## - svi 1 5.1814 52.966 -52.690  
## - lweight 1 5.8924 53.677 -51.397  
## - lcavol 1 28.0445 75.829 -17.884  
##   
## Step: AIC=-61.35  
## lpsa ~ lcavol + lweight + svi + lbph  
##   
## Df Sum of Sq RSS AIC  
## + age 1 0.9592 45.526 -61.374  
## <none> 46.485 -61.352  
## - lbph 1 1.3001 47.785 -60.676  
## + pgg45 1 0.3533 46.131 -60.092  
## + gleason 1 0.2126 46.272 -59.796  
## + lcp 1 0.1023 46.383 -59.565  
## - lweight 1 2.8014 49.286 -57.676  
## - svi 1 5.8063 52.291 -51.935  
## - lcavol 1 27.8298 74.315 -17.841  
##   
## Step: AIC=-61.37  
## lpsa ~ lcavol + lweight + svi + lbph + age  
##   
## Df Sum of Sq RSS AIC  
## <none> 45.526 -61.374  
## - age 1 0.9592 46.485 -61.352  
## + pgg45 1 0.6590 44.867 -60.789  
## + gleason 1 0.4560 45.070 -60.351  
## + lcp 1 0.1293 45.396 -59.650  
## - lbph 1 1.8568 47.382 -59.497  
## - lweight 1 3.2251 48.751 -56.735  
## - svi 1 5.9517 51.477 -51.456  
## - lcavol 1 28.7665 74.292 -15.871

##   
## Call:  
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age, data = prostate)  
##   
## Coefficients:  
## (Intercept) lcavol lweight svi lbph age   
## 0.95100 0.56561 0.42369 0.72095 0.11184 -0.01489

step 문을 활용한 단계별(stepwise) 선택법을 실시하였다. 전진선택법에서와 같이 절편만을 포함한 모형 pst\_lm\_null에서 시작하여, 설명변수를 하나씩 추가하였을 때 AIC 값이 가장 작은 설명변수를 선택하되, 검정결과가 유의하지 않을 경우 해당 설명변수를 제거하고 다시 설명변수의 추가를 검토하는 과정을 반복한 결과, 5개의 변수(lcavol, lweight, svi, lbph, age)를 포함하는 모형이 최종 모델로 선택되었다.

### Problem 2

*Using the trees data, fit a model with log (Volume) as the response and a second-order polynomial (including the interaction term) in Girth and Height. Determine whether the model may be reasonably simplified.*

### Data loading and EDA

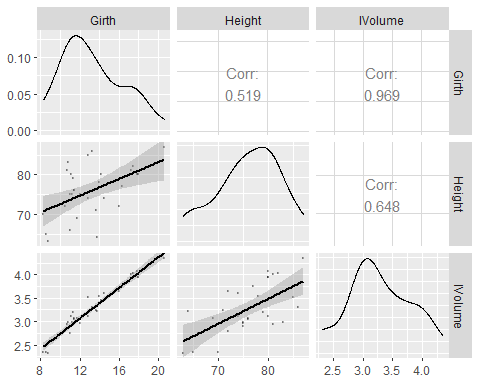
data("trees")  
str(trees)

## 'data.frame': 31 obs. of 3 variables:  
## $ Girth : num 8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...  
## $ Height: num 70 65 63 72 81 83 66 75 80 75 ...  
## $ Volume: num 10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...

summary(trees)

## Girth Height Volume   
## Min. : 8.30 Min. :63 Min. :10.20   
## 1st Qu.:11.05 1st Qu.:72 1st Qu.:19.40   
## Median :12.90 Median :76 Median :24.20   
## Mean :13.25 Mean :76 Mean :30.17   
## 3rd Qu.:15.25 3rd Qu.:80 3rd Qu.:37.30   
## Max. :20.60 Max. :87 Max. :77.00

trees %>% mutate(lVolume = log(Volume)) %>% dplyr::select(Girth, Height, lVolume) %>%   
 ggpairs(.,   
 lower = list(continuous = wrap("smooth", alpha = 0.3, size = 0.1)),  
 diag = list(discrete="barDiag",   
 continuous = wrap("densityDiag", alpha = 0.5 )),  
 upper = list(combo = wrap("box\_no\_facet", alpha = 0.5),  
 continuous = wrap("cor", size = 4,  
 alignPercent = 0.8))) +  
 theme(legend.position = "bottom")



trees\_lm <- lm(log(Volume) ~ Girth + Height + I(Girth^2) + I(Height^2) + Girth:Height, data=trees)  
summary(trees\_lm)

##   
## Call:  
## lm(formula = log(Volume) ~ Girth + Height + I(Girth^2) + I(Height^2) +   
## Girth:Height, data = trees)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.159718 -0.041905 -0.003371 0.055167 0.133780   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.9660208 2.0066922 -0.980 0.33660   
## Girth 0.2808126 0.0786856 3.569 0.00149 \*\*  
## Height 0.0484196 0.0567321 0.853 0.40150   
## I(Girth^2) -0.0042410 0.0032183 -1.318 0.19953   
## I(Height^2) -0.0002022 0.0004186 -0.483 0.63326   
## Girth:Height -0.0001975 0.0018089 -0.109 0.91395   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08469 on 25 degrees of freedom  
## Multiple R-squared: 0.9784, Adjusted R-squared: 0.9741   
## F-statistic: 226.7 on 5 and 25 DF, p-value: < 2.2e-16

문제에서 주어진 내용대로 log(Volume)을 종속변수로, Girth와 Height 및 2차식, 교호작용까지 고려한 회귀모형을 trees\_lm으로 적합하였다. 적합 결과, 결정계수는 0.9784로 매우 높으나 설명변수 중 Girth 하나만이 유의수준 0.05에서 유의한 것으로 나타나 좋은 모델이라고 할 수 없다.

이에, 후진제거법을 활용하여 모델을 개선, 단순화해보았다.

step(trees\_lm, data = trees, direction = "backward")

## Start: AIC=-147.73  
## log(Volume) ~ Girth + Height + I(Girth^2) + I(Height^2) + Girth:Height  
##   
## Df Sum of Sq RSS AIC  
## - Girth:Height 1 0.0000855 0.17941 -149.72  
## - I(Height^2) 1 0.0016737 0.18100 -149.44  
## <none> 0.17932 -147.73  
## - I(Girth^2) 1 0.0124560 0.19178 -147.65  
##   
## Step: AIC=-149.71  
## log(Volume) ~ Girth + Height + I(Girth^2) + I(Height^2)  
##   
## Df Sum of Sq RSS AIC  
## - I(Height^2) 1 0.002487 0.18189 -151.29  
## - Height 1 0.005377 0.18478 -150.80  
## <none> 0.17941 -149.72  
## - I(Girth^2) 1 0.055232 0.23464 -143.39  
## - Girth 1 0.248603 0.42801 -124.76  
##   
## Step: AIC=-151.29  
## log(Volume) ~ Girth + Height + I(Girth^2)  
##   
## Df Sum of Sq RSS AIC  
## <none> 0.18189 -151.29  
## - I(Girth^2) 1 0.08025 0.26214 -141.96  
## - Height 1 0.21815 0.40004 -128.85  
## - Girth 1 0.32692 0.50881 -121.40

##   
## Call:  
## lm(formula = log(Volume) ~ Girth + Height + I(Girth^2), data = trees)  
##   
## Coefficients:  
## (Intercept) Girth Height I(Girth^2)   
## -0.783931 0.285333 0.015701 -0.004954

summary(lm(formula = log(Volume) ~ Girth + Height + I(Girth^2), data = trees))

##   
## Call:  
## lm(formula = log(Volume) ~ Girth + Height + I(Girth^2), data = trees)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.174348 -0.043284 -0.000147 0.059198 0.138282   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.783931 0.315182 -2.487 0.01935 \*   
## Girth 0.285333 0.040960 6.966 1.74e-07 \*\*\*  
## Height 0.015701 0.002759 5.690 4.80e-06 \*\*\*  
## I(Girth^2) -0.004954 0.001435 -3.451 0.00185 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08208 on 27 degrees of freedom  
## Multiple R-squared: 0.9781, Adjusted R-squared: 0.9757   
## F-statistic: 402.1 on 3 and 27 DF, p-value: < 2.2e-16

후진제거법을 통해 AIC 값을 최소화한 결과, Height^2 및 두 설명변수의 교호작용 항이 제거되었다.

trees\_final <- lm(formula = log(Volume) ~ Girth + Height + I(Girth^2), data = trees)  
summary(trees\_final)

##   
## Call:  
## lm(formula = log(Volume) ~ Girth + Height + I(Girth^2), data = trees)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.174348 -0.043284 -0.000147 0.059198 0.138282   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.783931 0.315182 -2.487 0.01935 \*   
## Girth 0.285333 0.040960 6.966 1.74e-07 \*\*\*  
## Height 0.015701 0.002759 5.690 4.80e-06 \*\*\*  
## I(Girth^2) -0.004954 0.001435 -3.451 0.00185 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08208 on 27 degrees of freedom  
## Multiple R-squared: 0.9781, Adjusted R-squared: 0.9757   
## F-statistic: 402.1 on 3 and 27 DF, p-value: < 2.2e-16

anova(trees\_lm, trees\_final)

## Analysis of Variance Table  
##   
## Model 1: log(Volume) ~ Girth + Height + I(Girth^2) + I(Height^2) + Girth:Height  
## Model 2: log(Volume) ~ Girth + Height + I(Girth^2)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 25 0.17932   
## 2 27 0.18189 -2 -0.0025722 0.1793 0.8369

최종 모델 trees\_final의 summary 확인 결과, 3개 변수(Girth, Height, Girth^2)가 모두 유의하며 결정계수 또한 큰 차이가 없는 것으로 나타났다. anova 결과에서도 p-value가 0.05 이상으로 귀무가설을 기각할 수 없는 바, 최종 모델 trees\_final이 기존 모형보다 더 개선, 단순화되었음을 확인할 수 있다.