hw7

Jiayi

2024-10-28

```
library(faraway)
library(leaps)
```

Part 1: Prostate Data Analysis

```
data(prostate)
str(prostate)
```

Load and Explore the Data

```
## '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
```

```r
model_full <- lm(lpsa ~ ., data = prostate)
model_backward <- step(model_full, direction = "backward", k = qchisq(0.20, 1, lower.tail = FALSE))
```

(a) Backward Elimination with Alpha = 0.20

```
## Start:  AIC=-61.54
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##
##           Df Sum of Sq    RSS     AIC
## - gleason  1    0.0412 44.204 -63.092
## - pgg45    1    0.5258 44.689 -62.035
## - lcp      1    0.6740 44.837 -61.714
## <none>                 44.163 -61.540
## - age      1    1.5503 45.713 -59.836
## - lbph     1    1.6835 45.847 -59.554
## - lweight  1    3.5861 47.749 -55.610
## - svi      1    4.9355 49.099 -52.907
## - lcavol   1   22.3721 66.535 -23.428
##
## Step:  AIC=-63.09
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - lcp      1    0.6623 44.867 -63.292
## <none>                 44.204 -63.092
## - pgg45    1    1.1920 45.396 -62.154
## - age      1    1.5166 45.721 -61.463
## - lbph     1    1.7053 45.910 -61.063
## - lweight  1    3.5462 47.750 -57.249
## - svi      1    4.8984 49.103 -54.541
## - lcavol   1   23.5039 67.708 -23.375
##
## Step:  AIC=-63.29
## lpsa ~ lcavol + lweight + age + lbph + svi + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - pgg45    1    0.6590 45.526 -63.520
## <none>                 44.867 -63.292
## - age      1    1.2649 46.131 -62.238
## - lbph     1    1.6465 46.513 -61.438
## - lweight  1    3.5647 48.431 -57.519
## - svi      1    4.2503 49.117 -56.155
## - lcavol   1   25.4189 70.285 -21.394
```

```
##
## Step:  AIC=-63.52
## lpsa ~ lcavol + lweight + age + lbph + svi
##
##           Df Sum of Sq    RSS     AIC
## <none>                 45.526 -63.520
## - age      1    0.9592 46.485 -63.140
## - lbph     1    1.8568 47.382 -61.285
## - lweight  1    3.2251 48.751 -58.523
## - svi      1    5.9517 51.477 -53.245
## - lcavol   1   28.7665 74.292 -17.659
```

```
summary(model_backward)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, 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 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## 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
```

```
leaps_result <- regsubsets(lpsa ~ ., data = prostate, nbest = 1)
summary_leaps <- summary(leaps_result)

# Number of observations
n <- nrow(prostate)

# Calculate AIC for each model
aic_values <- numeric(length(summary_leaps$rss))
for (i in 1:length(aic_values)) {
  k <- sum(summary_leaps$which[i, ])  # Number of predictors (including intercept)
  rss <- summary_leaps$rss[i]         # Residual Sum of Squares
  aic_values[i] <- n * log(rss / n) + 2 * k
}
```

```r
# Get the selected models' indices
aic_model_index <- which.min(aic_values)
bic_model_index <- which.min(summary_leaps$bic)
cp_model_index <- which.min(summary_leaps$cp)


# Extract the variables included in those models
aic_model_vars <- names(which(summary_leaps$which[aic_model_index, ]))
bic_model_vars <- names(which(summary_leaps$which[bic_model_index, ]))
cp_model_vars <- names(which(summary_leaps$which[cp_model_index, ]))

# Report the selected variables
list(
  AIC_Model_Vars = aic_model_vars,
  BIC_Model_Vars = bic_model_vars,
  Cp_Model_Vars = cp_model_vars
)
```

(b), (c), (d) AIC, BIC, and Mallows Cp using leaps package

```
## $AIC_Model_Vars
## [1] "(Intercept)" "lcavol"      "lweight"     "age"         "lbph"
## [6] "svi"
##
## $BIC_Model_Vars
## [1] "(Intercept)" "lcavol"      "lweight"     "svi"
##
## $Cp_Model_Vars
## [1] "(Intercept)" "lcavol"      "lweight"     "lbph"        "svi"
```

```r
model_stepwise <- step(model_full, direction = "both", trace = 0)
summary(model_stepwise)
```

(e) Stepwise Selection using AIC Criterion

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, 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 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
```

```
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## 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
```

The stepwise selection method, which uses AIC to determine the best model, selected lcavol (log cancer volume), lweight (log prostate weight), age, lbph (log benign prostatic hyperplasia), and svi (seminal vesicle invasion) as the most relevant predictors for lpsa (log prostate-specific antigen)

Part 2: TeenGamb Data Analysis

```
data(teengamb)
str(teengamb)
```

Load and Explore the Data

```
## 'data.frame':    47 obs. of  5 variables:
##  $ sex   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ status: int  51 28 37 28 65 61 28 27 43 18 ...
##  $ income: num  2 2.5 2 7 2 3.47 5.5 6.42 2 6 ...
##  $ verbal: int  8 8 6 4 8 6 7 5 6 7 ...
##  $ gamble: num  0 0 0 7.3 19.6 0.1 1.45 6.6 1.7 0.1 ...
```

```
summary(teengamb)
```

```
##       sex            status          income          verbal
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
##  Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00
##  Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66
##  3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
##  Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00
##      gamble
##  Min.   :  0.0
##  1st Qu.:  1.1
##  Median :  6.0
##  Mean   : 19.3
##  3rd Qu.: 19.4
##  Max.   :156.0
```

```
model_full_gamble <- lm(gamble ~ ., data = teengamb)
model_backward_gamble <- step(model_full_gamble, direction = "backward", k = qchisq(0.20, 1, lower.tail
```

(a) Backward Elimination with Alpha = 0.20

```
## Start:  AIC=296.39
## gamble ~ sex + status + income + verbal
##
##          Df Sum of Sq   RSS    AIC
## - status  1      17.8 21642 294.78
## <none>                21624 296.39
## - verbal  1     955.7 22580 296.78
## - sex     1    3735.8 25360 302.24
## - income  1   12056.2 33680 315.57
##
## Step:  AIC=294.78
## gamble ~ sex + income + verbal
##
##          Df Sum of Sq   RSS    AIC
## <none>                21642 294.78
## - verbal  1    1139.8 22781 295.55
## - sex     1    5787.9 27429 304.28
## - income  1   13236.1 34878 315.57
```

```
summary(model_backward_gamble)
```

```
##
## Call:
## lm(formula = gamble ~ sex + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.639 -11.765  -1.594   9.305  93.867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.1390    14.7686   1.634   0.1095
## sex         -22.9602     6.7706  -3.391   0.0015 **
## income        4.8981     0.9551   5.128 6.64e-06 ***
## verbal       -2.7468     1.8253  -1.505   0.1397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.43 on 43 degrees of freedom
## Multiple R-squared:  0.5263, Adjusted R-squared:  0.4933
## F-statistic: 15.93 on 3 and 43 DF,  p-value: 4.148e-07
```

```
leaps_result_gamble <- regsubsets(gamble ~ ., data = teengamb, nbest = 1)
summary_leaps_gamble <- summary(leaps_result_gamble)

# Number of observations
n <- nrow(teengamb)

# Calculate AIC for each model
aic_values_gamble <- numeric(length(summary_leaps_gamble$rss))
for (i in 1:length(aic_values_gamble)) {
```

```
  k <- sum(summary_leaps_gamble$which[i, ])  # Number of predictors (including intercept)
  rss <- summary_leaps_gamble$rss[i]         # Residual Sum of Squares
  aic_values_gamble[i] <- n * log(rss / n) + 2 * k
}

# Get the model indices that minimize AIC, BIC, and Mallows Cp
aic_model_index_gamble <- which.min(aic_values_gamble)
bic_model_index_gamble <- which.min(summary_leaps_gamble$bic)
cp_model_index_gamble <- which.min(summary_leaps_gamble$cp)

# Extract the selected variables (removing intercept for clarity)
aic_model_vars_gamble <- names(which(summary_leaps_gamble$which[aic_model_index_gamble, ]))[-1]
bic_model_vars_gamble <- names(which(summary_leaps_gamble$which[bic_model_index_gamble, ]))[-1]
cp_model_vars_gamble <- names(which(summary_leaps_gamble$which[cp_model_index_gamble, ]))[-1]

# Print the selected variables for AIC, BIC, and Cp models
cat("AIC Model Variables:", paste(aic_model_vars_gamble, collapse = ", "), "\n")
```

(b), (c), (d) AIC, BIC, and Mallows Cp using leaps package

```
## AIC Model Variables: sex, income, verbal
```

```
cat("BIC Model Variables:", paste(bic_model_vars_gamble, collapse = ", "), "\n")
```

```
## BIC Model Variables: sex, income
```

```
cat("Cp Model Variables:", paste(cp_model_vars_gamble, collapse = ", "), "\n")
```

```
## Cp Model Variables: sex, income, verbal
```

```
model_stepwise_gamble <- step(model_full_gamble, direction = "both", trace = 0)
summary(model_stepwise_gamble)
```

(e) Stepwise Selection using AIC Criterion

```
##
## Call:
## lm(formula = gamble ~ sex + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.639 -11.765  -1.594   9.305  93.867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.1390    14.7686   1.634   0.1095
## sex         -22.9602     6.7706  -3.391   0.0015 **
## income        4.8981     0.9551   5.128 6.64e-06 ***
```

```
## verbal       -2.7468     1.8253  -1.505   0.1397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.43 on 43 degrees of freedom
## Multiple R-squared:  0.5263, Adjusted R-squared:  0.4933
## F-statistic: 15.93 on 3 and 43 DF,  p-value: 4.148e-07
```

Three varibales: sex, income, verbal