M09Guided

Alanna Hazlett

2024-03-30

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(leaps)
Data<-read.table("nfl.txt", header=TRUE, sep="")
```

Problem 1

Use the regsubsets() function from the leaps package to run all possible regressions. Set nbest=2. Identify the model (the predictors and the corresponding estimated coefficients) that is best in terms of Adjusted R2, Mallow's Cp, and BIC.

```r
allreg2 <- leaps::regsubsets(y ~., data=Data, nbest=2)
coef(allreg2,which.max(summary(allreg2)$adjr2))
```

```
##  (Intercept)           x2           x7           x8           x9
## -1.821703427  0.003818572  0.216894094 -0.004014887 -0.001634926
```

```r
coef(allreg2,which.min(summary(allreg2)$cp))
```

```
##  (Intercept)           x2           x7           x8
## -1.808372059  0.003598070  0.193960210 -0.004815494
```

```r
coef(allreg2,which.min(summary(allreg2)$bic))
```

```
##  (Intercept)           x2           x7           x8
## -1.808372059  0.003598070  0.193960210 -0.004815494
```

Problem 2

Run forward selection, starting with an intercept-only model. Report the predictors and the estimated coefficients of the model selected.

```r
##intercept only model
regnull <- lm(y~1, data=Data)
##model with all predictors
regfull <- lm(y~., data=Data)
step(regnull, scope=list(lower=regnull, upper=regfull), direction="forward")
```

```
## Start:  AIC=70.81
## y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + x8    1   178.092 148.87 50.785
## + x1    1   115.068 211.90 60.669
## + x7    1    97.238 229.73 62.931
## + x5    1    86.116 240.85 64.255
## + x2    1    76.193 250.77 65.385
## + x9    1    30.167 296.80 70.104
## <none>              326.96 70.814
## + x4    1    21.844 305.12 70.878
## + x6    1    16.411 310.55 71.372
## + x3    1     2.135 324.83 72.631
##
## Step:  AIC=50.78
## y ~ x8
##
##        Df Sum of Sq     RSS    AIC
## + x2    1    64.934  83.938 36.741
## + x5    1    11.607 137.265 50.512
## <none>              148.872 50.785
## + x1    1     6.636 142.236 51.508
## + x3    1     6.368 142.504 51.561
## + x4    1     6.345 142.527 51.565
## + x7    1     0.974 147.898 52.601
## + x6    1     0.487 148.385 52.693
## + x9    1     0.008 148.864 52.783
##
## Step:  AIC=36.74
## y ~ x8 + x2
##
##        Df Sum of Sq    RSS    AIC
## + x7    1   14.0682 69.870 33.604
## + x1    1   11.1905 72.748 34.734
## + x3    1    8.9010 75.037 35.602
## + x5    1    5.8147 78.124 36.730
## <none>              83.938 36.741
## + x9    1    2.0256 81.913 38.057
## + x6    1    1.3216 82.617 38.296
## + x4    1    0.0161 83.922 38.735
##
## Step:  AIC=33.6
## y ~ x8 + x2 + x7
##
##        Df Sum of Sq    RSS    AIC
## + x9    1    4.8657 65.004 33.583
## <none>              69.870 33.604
```

```
## + x3    1    1.3873 68.483 35.043
## + x4    1    0.9792 68.891 35.209
## + x1    1    0.9022 68.968 35.240
## + x6    1    0.4879 69.382 35.408
## + x5    1    0.2987 69.571 35.484
##
## Step:  AIC=33.58
## y ~ x8 + x2 + x7 + x9
##
##        Df Sum of Sq    RSS    AIC
## <none>              65.004 33.583
## + x1    1   1.86452 63.140 34.768
## + x4    1   1.74260 63.262 34.822
## + x3    1   0.70148 64.303 35.279
## + x6    1   0.45071 64.554 35.388
## + x5    1   0.32667 64.678 35.442
```

```
##
## Call:
## lm(formula = y ~ x8 + x2 + x7 + x9, data = Data)
##
## Coefficients:
## (Intercept)           x8           x2           x7           x9
##   -1.821703    -0.004015     0.003819     0.216894    -0.001635
```

The forward selection candidate model:

$$y = -1.822 + -0.004x_8 + 0.004x_2 + 0.217x_7 + -0.002x_9$$

Problem 3

Run backward elimination, starting with the model with all predictors. Report the predictors and the estimated coefficients of the model selected.

```r
step(regfull, scope=list(lower=regnull, upper=regfull), direction="backward")
```

```
## Start:  AIC=41.48
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x5    1     0.000  60.293 39.476
## - x1    1     0.549  60.842 39.730
## - x3    1     0.746  61.039 39.821
## - x6    1     0.803  61.096 39.847
## - x4    1     1.968  62.261 40.376
## - x7    1     3.451  63.744 41.035
## <none>               60.293 41.476
## - x9    1     5.348  65.642 41.856
## - x8    1    12.072  72.365 44.587
## - x2    1    62.448 122.741 59.380
##
## Step:  AIC=39.48
## y ~ x1 + x2 + x3 + x4 + x6 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
```

```
## - x1    1     0.553  60.846 37.732
## - x3    1     0.750  61.043 37.822
## - x6    1     0.818  61.111 37.854
## - x4    1     2.053  62.346 38.414
## - x7    1     3.859  64.152 39.213
## <none>               60.293 39.476
## - x9    1     5.351  65.644 39.857
## - x8    1    12.086  72.379 42.592
## - x2    1    66.979 127.272 58.395
##
## Step:  AIC=37.73
## y ~ x2 + x3 + x4 + x6 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x6    1     0.690  61.536 36.048
## - x3    1     1.715  62.561 36.510
## - x4    1     3.051  63.897 37.102
## <none>               60.846 37.732
## - x9    1     4.852  65.698 37.880
## - x7    1     8.961  69.807 39.579
## - x8    1    16.599  77.445 42.486
## - x2    1    67.010 127.856 56.524
##
## Step:  AIC=36.05
## y ~ x2 + x3 + x4 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x3    1     1.726  63.262 34.822
## - x4    1     2.767  64.303 35.279
## <none>               61.536 36.048
## - x9    1     4.831  66.367 36.164
## - x7    1     9.390  70.926 38.024
## - x8    1    18.314  79.851 41.343
## - x2    1    66.447 127.984 54.552
##
## Step:  AIC=34.82
## y ~ x2 + x4 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x4    1     1.743  65.004 33.583
## <none>               63.262 34.822
## - x9    1     5.629  68.891 35.209
## - x8    1    17.701  80.962 39.730
## - x7    1    18.583  81.845 40.033
## - x2    1    75.598 138.860 54.835
##
## Step:  AIC=33.58
## y ~ x2 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## <none>               65.004 33.583
## - x9    1     4.866  69.870 33.604
## - x7    1    16.908  81.913 38.057
## - x8    1    23.299  88.303 40.160
```

```
## - x2    1    82.892 147.897 54.601
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8 + x9, data = Data)
##
## Coefficients:
## (Intercept)           x2           x7           x8           x9
##   -1.821703     0.003819     0.216894    -0.004015    -0.001635
```

The backward elimination candidate model:

$$y = -1.822 + 0.004x_2 + 0.217x_7 + -0.004x_8 + -0.002x_9$$

Problem 4

Run stepwise regression, starting with an intercept-only model. Report the predictors and the estimated coefficients of the model selected.

```r
step(regnull, scope=list(lower=regnull, upper=regfull), direction="both")
```

```
## Start:  AIC=70.81
## y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + x8    1   178.092 148.87 50.785
## + x1    1   115.068 211.90 60.669
## + x7    1    97.238 229.73 62.931
## + x5    1    86.116 240.85 64.255
## + x2    1    76.193 250.77 65.385
## + x9    1    30.167 296.80 70.104
## <none>              326.96 70.814
## + x4    1    21.844 305.12 70.878
## + x6    1    16.411 310.55 71.372
## + x3    1     2.135 324.83 72.631
##
## Step:  AIC=50.78
## y ~ x8
##
##        Df Sum of Sq    RSS    AIC
## + x2    1    64.934  83.94 36.741
## + x5    1    11.607 137.27 50.512
## <none>              148.87 50.785
## + x1    1     6.636 142.24 51.508
## + x3    1     6.368 142.50 51.561
## + x4    1     6.345 142.53 51.565
## + x7    1     0.974 147.90 52.601
## + x6    1     0.487 148.39 52.693
## + x9    1     0.008 148.86 52.783
## - x8    1   178.092 326.96 70.814
##
## Step:  AIC=36.74
## y ~ x8 + x2
##
##        Df Sum of Sq     RSS    AIC
```

```
## + x7    1    14.068  69.870 33.604
## + x1    1    11.190  72.748 34.734
## + x3    1     8.901  75.037 35.602
## + x5    1     5.815  78.124 36.730
## <none>               83.938 36.741
## + x9    1     2.026  81.913 38.057
## + x6    1     1.322  82.617 38.296
## + x4    1     0.016  83.922 38.735
## - x2    1    64.934 148.872 50.785
## - x8    1   166.833 250.771 65.385
##
## Step:  AIC=33.6
## y ~ x8 + x2 + x7
##
##        Df Sum of Sq     RSS    AIC
## + x9    1     4.866  65.004 33.583
## <none>               69.870 33.604
## + x3    1     1.387  68.483 35.043
## + x4    1     0.979  68.891 35.209
## + x1    1     0.902  68.968 35.240
## + x6    1     0.488  69.382 35.408
## + x5    1     0.299  69.571 35.484
## - x7    1    14.068  83.938 36.741
## - x8    1    41.400 111.270 44.633
## - x2    1    78.028 147.898 52.601
##
## Step:  AIC=33.58
## y ~ x8 + x2 + x7 + x9
##
##        Df Sum of Sq     RSS    AIC
## <none>               65.004 33.583
## - x9    1     4.866  69.870 33.604
## + x1    1     1.865  63.140 34.768
## + x4    1     1.743  63.262 34.822
## + x3    1     0.701  64.303 35.279
## + x6    1     0.451  64.554 35.388
## + x5    1     0.327  64.678 35.442
## - x7    1    16.908  81.913 38.057
## - x8    1    23.299  88.303 40.160
## - x2    1    82.892 147.897 54.601
```

```
##
## Call:
## lm(formula = y ~ x8 + x2 + x7 + x9, data = Data)
##
## Coefficients:
## (Intercept)           x8           x2           x7           x9
##   -1.821703    -0.004015     0.003819     0.216894    -0.001635
```

The stepwise regression candidate model:

$$y = -1.822 + -0.004x_8 + 0.004x_2 + 0.217x_7 + -0.002x_9$$

Problem 5

The PRESS statistic can be used in model validation as well as a criteria for model selection. Unfortunately, the regsubsets() function from the leaps package does not compute the PRESS statistic. Write a function that computes the PRESS statistic for a regression model. Hint: the diagonal elements from the hat matrix can be found using the lm.influence() function.
https://stevencarlislewalker.wordpress.com/2013/06/18/calculating-the-press-statistic-in-r/

- Residuals of the model
  r<-resid(model)

- Predictively adjusted residuals
  pr<-resid(model)/(1-lm.influence(model)$hat)

- SSres
  SSres<-sum(r^2) or
  SSres <- sum((fitted(model) - Data$y)^2)

- PRESS
  sum(pr^2)
  press<-sum((resid(model)/(1-lm.influence(model)$hat))^2)

```
press<-function(model) {
  pr<-resid(model)/(1-lm.influence(model)$hat)
  press<-sum(pr^2)
  return (press)
}
```

Problem 6

Using the function you wrote in part 5, calculate the PRESS statistic for your regression model with x2, x7, x8 as predictors. Calculate the R2 Prediction for this model, and compare this value with its R2. What comments can you make about the likely predictive performance of this model?
https://www.statology.org/sst-ssr-sse-in-r/

```
model<-lm(y~x2+x7+x8,data=Data)
press<-sum((resid(model)/(1-lm.influence(model)$hat))^2)
sprintf("PRESS: %s", round(press,4))
```

```
## [1] "PRESS: 87.4612"
```

```
SSres <- sum((fitted(model) - Data$y)^2)
SSR <- sum((fitted(model) - mean(Data$y))^2)
SSt<-SSR + SSres
#SST<-sum(anova(model)$"Sum Sq")

# R2 Prediction
R2Pre<-1-(press/SSt)
sprintf("R^2 Prediciton: %s", round(R2Pre,4))
```

```
## [1] "R^2 Prediciton: 0.7325"
```

```
# R2
R2<-SSR/SSt
```

```r
sprintf("R^2: %s", round(R2,4))
```

```
## [1] "R^2: 0.7863"
```

```r
#R2<-anova(m)$"Sum Sq"[1]/SST
```

- R2 Prediction is the proportion of variability in the response of the new observations that can be explained by our model.

- R2 (Coefficient of determination) is an indication of how well the data fits our model, proportion of variance in the response variable that is explained by the predictor. The closer this value is to 1 the better the fit.

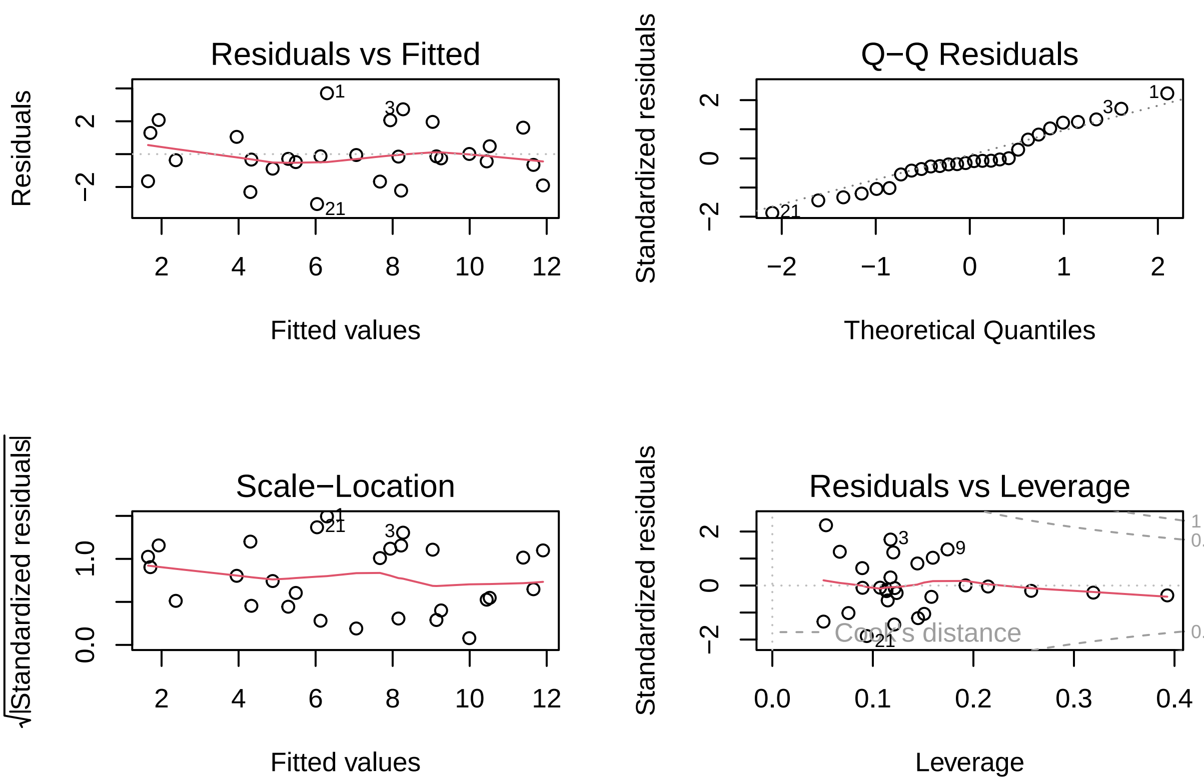
  These values are pretty similar, as we would expect the R2 Prediction is less than R2, as R2 does not have a penalty for additional parameters added to the model. They both indicate a good value for how well the the data fits the model.

  Professor's note: The model might be able to explain 73.25% of the variability in the new observations. The R2is 0.7863. Both values are fairly high and close to each other, so the model has good predictive ability.

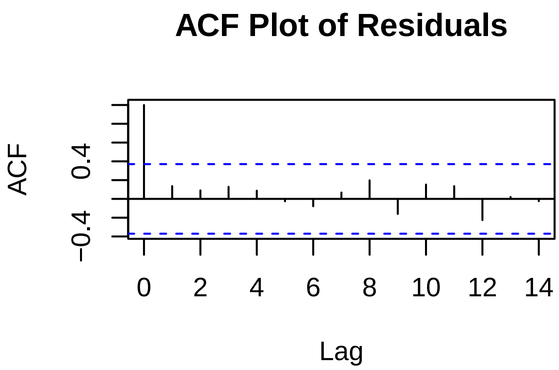
Problem 7

Create diagnostic plots for the model with x2, x7, x8 as predictors. What are these plots telling us?

```r
par(mfrow=c(2,2))
plot(model)
```



```r
acf(model$residuals, main="ACF Plot of Residuals")
#boxcox(model)
```



Assumption 1: Do the errors have mean of 0 for each value of the predictor -Yes
Assumption 2: Do the errors have constant variance for each value of the predictor -Yes
Assumption 3: Are the errors independent (acf plot) - Yes
Assumption 4: Are the errors normally distributed? -Yes
There is a linear relationship between the predictors x2 (Passing yards-Season), x7 (Percent rushing), x8 (Opponent's rushing yards) with our response variable (Games won).