hw_01

lisa liubovich

2024-05-21

1. 2.4.1 Flexible vs. Inflexible

For each part, indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. **Justify your answer**.

(a) The sample size n is extremely large, and the number of predictors p is small

   (a) a flexible method will be better because of how large n is. Even though p is small, flexible methods are less likely to overfit because of the ratio of p to n.

(b) The number of predictors p is extremely large, and the number of observations n is small.

   (a) a flexible method will be worse because of how small n is because flexible methods tend to overfit when p is large relative to n.

(c) The relationship between the predictors and response is highly non-linear.

   (a) flexible method will be better because inflexible models like linear regression would not effectively capture the patterns in the data. More flexible methods are designed to capture non-linear relationships due to their complexity.

(d) The variance of the error terms is extremely high.

   (a) a flexible method will be worse because it will likely overfit the data to the noise associated with high variance in the error term.

2. Exercise 2.4.2.(a,b) Classification or Regression

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(a) We collect a set of data on the top 300 publicly-traded firms in the US. For each firm we record profit, number of employees, industry, average stock price over the last year, and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (a) this is a regression problem because the outcome variable is numerical (CEO salary)
   (b) we are more interested in inference than prediction because we are interested in understanding the association between the predictors and the outcome
   (c) n = 300, p = 5

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   (a) this is a classification problem because the outcome variable is categorical (success or failure)

   (b) this is a prediction problem because rather than looking at associations between the predictors and outcome variable, we are looking to predict whether the product launch will be a success or failure

   (c) n = 20, p = 14

3 Exercise 2.4.4 (modified) Examples of SML.

Identify some examples of potential real-life applications for statistical learning. In each example, describe the response and the predictors. Also state the goal - inference or prediction.

(a) Describe one real-life applications in which classification might be useful.

   (a) You are trying to model the probability of winning or losing a football game. Your outcome variable (win or loss) is categorical and you have predictors like time spent in practice, average number of completed passes per game, etc. Your goal is inference, or trying to understand the relationship between game outcome and predictors like completed passes per game or average points scored per game.

(b) Describe one real-life applications in which regression might be useful.

   (a) You are trying to predict the average earnings of a data scientist. Your outcome variable (average earnings) is numerical and you have predictors like age, gender, educational attainment, years of experience, etc. Your goal is prediction because instead of trying to understand the relationship between earnings and the predictors, you want to use factors like age, gender, etc to accurately predict average earnings.

4 Exercise 3.7.4. Training and Test Residual Sums of Squares

Residual Sums of Squares (RSS) is also referred to as the "Sum of Squares of Errors' " (SSE). I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, to include a separate cubic regression, i.e., Y =B0 +B1X+B2X2 +B3X3 +E

(a) Suppose the true relationship between X and Y is linear, i.e., Y = B0 + B1X. Consider the training Residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

   (a) The training RSS for the cubic regression would be equal to or lower than the training RSS for the linear regression because the cubic model can capture the linear relationship and potentially fit the noise in the data, which reduces error.

(b) Answer (a) using test rather than training RSS.

   (a) The test RSS for the linear regression is expected to be lower than the test RSS for the cubic regression. This is because the cubic model is likely to overfit the training data and thus not predict as accurately to new, unseen data due to it being more flexible.

(c) Suppose the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(a) The training RSS for the cubic regression would be lower than the training RSS for the linear regression because the cubic model is more flexible and thus can better capture the non-linear patterns in the data.

(d) Answer (c) using test rather than training RSS.

(a) There is not enough information to tell. The test RSS for the cubic regression could be lower than the test RSS for the linear regression if the non-linear relationship is such that the cubic terms significantly help in capturing the true pattern. However, if the cubic model overfits the training data, the linear model could have a lower test RSS.

5 Application: Predict the number of college application.

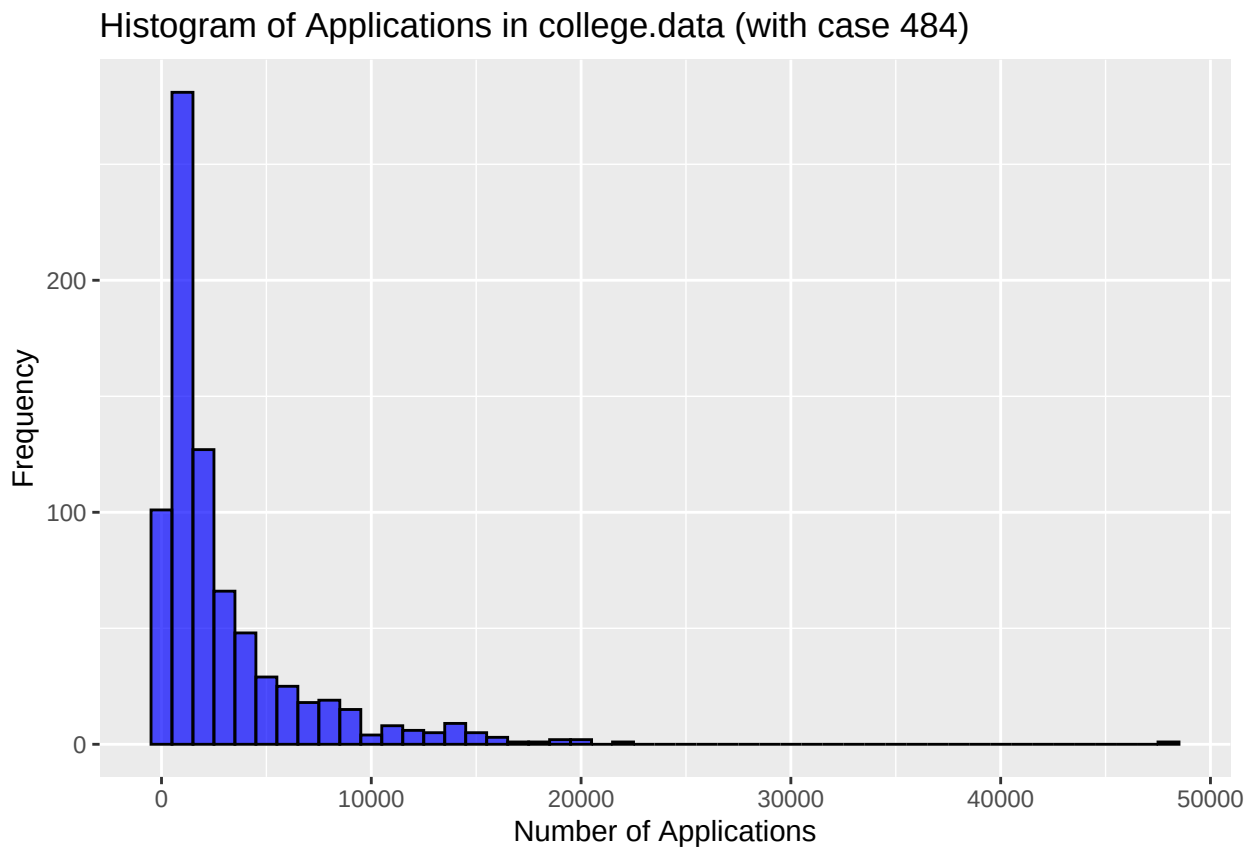
Predict the number of applications received based on the other variables in the College data set in {ISLR2}. Use the following code to split the data set randomly. Only use the subset college.data in this exercise.

```
library(ISLR2)
data("College")
my.college <- College[-484, ] # remove an extreme case.
my.college <- College[College$Apps <=16000, ] # remove several extreme case.
train.pct <- 0.78
set.seed(2024)
Z <- sample(nrow(my.college), floor(train.pct*nrow(my.college)))
college.data <- my.college[Z, ]
holdout.data <- my.college[-Z, ]
```

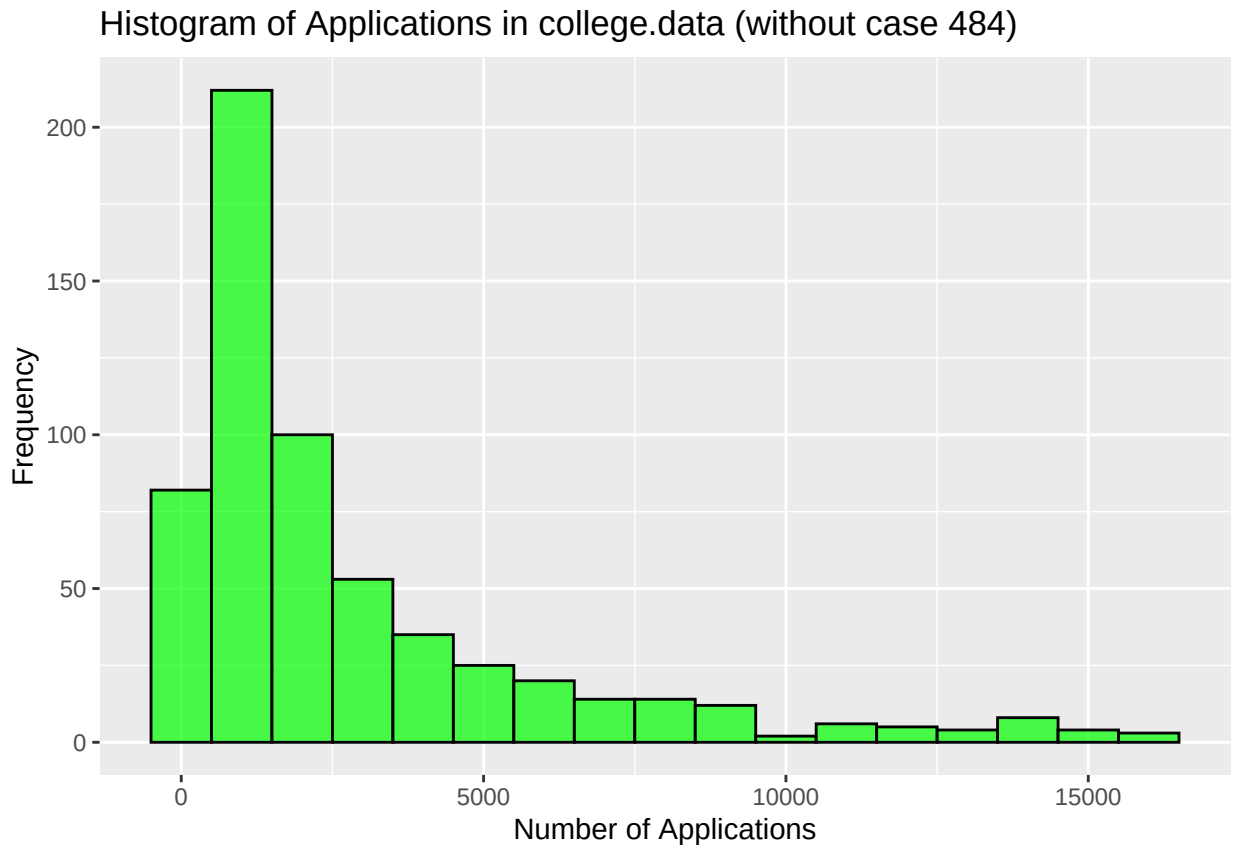
(a) (2 pts) Read the help file of the data file and determine the response variable for this study. Prepare a histograms of the response with and without case 484 in the original College data set. Why would I remove it from the analysis for this excise?

The response variable is Apps (number of applications received)

```
# with case 484
ggplot(College, aes(x = Apps)) +
  geom_histogram(binwidth = 1000, fill = "blue", color = "black", alpha = 0.7) +
  ggtitle("Histogram of Applications in college.data (with case 484)") +
  xlab("Number of Applications") +
  ylab("Frequency")
```




```
# without case 484
ggplot(college.data, aes(x = Apps)) +
  geom_histogram(binwidth = 1000, fill = "green", color = "black", alpha = 0.7) +
  ggtitle("Histogram of Applications in college.data (without case 484)") +
  xlab("Number of Applications") +
  ylab("Frequency")
```



You would remove it from the data for this exercise because it skews the data right.

b. (2 pts) Fit a first order linear regression with all available predictors. Be sure to use college.data data frame. (In lm() and glm() function, y ~ . fit a model between y and all other variables in the data. y ~ 1 fits a model with only intercept without any predictor.)

```
lm_college <- lm(Apps ~ Private + Accept + Enroll + Top10perc + Top25perc + F.Undergrad + P.Undergrad +
tidy(lm_college)
```

```
## # A tibble: 18 x 5
##    term          estimate std.error statistic  p.value
##    <chr>            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept) -655.       431.        -1.52  1.29e- 1
##  2 PrivateYes  -644.       147.        -4.37  1.45e- 5
##  3 Accept         1.24       0.0556    22.3   4.65e-80
##  4 Enroll         0.416      0.224      1.85  6.46e- 2
##  5 Top10perc     45.9        5.88       7.80  2.79e-14
##  6 Top25perc    -14.0        4.66      -3.01  2.75e- 3
##  7 F.Undergrad   -0.0462     0.0387    -1.19  2.33e- 1
##  8 P.Undergrad    0.0430     0.0330     1.30  1.93e- 1
##  9 Outstate      -0.0512     0.0203    -2.52  1.19e- 2
## 10 Room.Board     0.186      0.0507     3.66  2.71e- 4
## 11 Books          0.0584     0.264      0.221 8.25e- 1
## 12 Personal      -0.00990    0.0662    -0.150 8.81e- 1
## 13 PhD           -6.16       5.01      -1.23  2.20e- 1
## 14 Terminal      -4.48       5.37      -0.835 4.04e- 1
```

```
## 15 S.F.Ratio      8.82      13.3        0.662 5.08e- 1
## 16 perc.alumni   -4.30       4.46      -0.963 3.36e- 1
## 17 Expend         0.0725     0.0121     5.98  4.01e- 9
## 18 Grad.Rate     10.9        3.06       3.55  4.20e- 4
```

c. (4 pt) Compute the variance inflation factor and comment on the severity of the collinearity of the data. Why is "collinearity" a concern, even if the model is correct?

```
vif(lm_college)
```

```
##     Private      Accept      Enroll   Top10perc   Top25perc F.Undergrad
##    2.737054    8.123611   23.668077    7.024986    5.566977   20.127900
## P.Undergrad    Outstate  Room.Board       Books    Personal         PhD
##    1.733848    4.260564    1.930562    1.144628    1.368765    4.127543
##    Terminal   S.F.Ratio perc.alumni      Expend   Grad.Rate
##    3.902255    1.802307    1.928466    2.785951    1.851383
```

There are several variables like Enroll, Top10Perc, Top25Perc, and F.Undergrad who have VIF higher than 5 and even higher than 10, indicating some significant collinearity. Multicollinearity is a concern because it can make interpretation difficult and inflate standard errors, which makes our inferences and predictions less accurate by increasing the width of confidence intervals and reduced statistical power.

d. (8 pts) Find the best least squares regression model(s), using adjusted R2, BIC, and Cp criteria with best subset algorithm. Also use the stepwise variable selection algorithm using AIC criterion. Note that the "best" model(s) may be different depending on the criterion. (Hint: In the regsubsets() function, set nvmax = 17 to consider models with up to 17 predictors.)

```
library(leaps)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:ISLR2':
##
##     Boston

## The following object is masked from 'package:dplyr':
##
##     select
```

```
# Fit regression models using best subset algorithm
best_subset <- regsubsets(Apps ~ ., data = college.data, nvmax = 17)

# Extract summary information
summary_best_subset <- summary(best_subset)

# Find the best model(s) based on adjusted R^2
best_adj_r_squared <- which.max(summary_best_subset$adjr2)

# Find the best model(s) based on BIC
best_bic <- which.min(summary_best_subset$bic)
```

```r
# Find the best model(s) based on Cp
best_cp <- which.min(summary_best_subset$cp)

# Extract the predictors selected by the best model(s)
best_predictors <- colnames(summary_best_subset$which)[best_adj_r_squared]

# Construct the formula for the best model(s)
best_formula <- as.formula(paste("Apps ~", paste(best_predictors, collapse = " + ")))

# Fit the best model(s) based on adjusted R^2 using lm
lm_best_adj_r_squared <- lm(best_formula, data = college.data)

# Fit regression model using stepwise variable selection algorithm
stepwise_model <- stepAIC(lm_best_adj_r_squared, direction = "both")
```

```
## Start:  AIC=9546.56
## Apps ~ PhD
##
##        Df  Sum of Sq        RSS    AIC
## <none>               4967022218 9546.6
## - PhD   1 1226461762 6193483980 9676.7
```

```r
# Get the AIC criterion for the stepwise model
stepwise_aic <- AIC(stepwise_model)

# Print the results
print(list(best_adj_r_squared = best_adj_r_squared,
           best_bic = best_bic,
           best_cp = best_cp,
           stepwise_aic = stepwise_aic))
```

```
## $best_adj_r_squared
## [1] 13
##
## $best_bic
## [1] 9
##
## $best_cp
## [1] 10
##
## $stepwise_aic
## [1] 11248.45
```

1. **best_adj_r_squared**: The number of predictors in the best model based on adjusted R^2 In this case, the best model contains 13 predictors.
2. **best_bic**: The number of predictors in the best model based on BIC (Bayesian Information Criterion). In this case, the best model contains 9 predictors.
3. **best_cp**: The number of predictors in the best model based on Cp (Mallows' Cp). In this case, the best model contains 10 predictors.
4. **$stepwise_aic**: The final AIC value after the stepwise variable selection process. In this case, the final AIC value is 11248.45.