College Study

Mia Yu

1 Research Project

Data Overview

We explore the application of nonlinear models to analyze the **"College" dataset**, comprising statistics from 565 US colleges as reported in a past issue of US News and World Report. The predictors are

predictors	Explanation
Apps	umber of applications received
Accept	Number of applications accepted
Enroll	Number of new students enrolled
Top10perc	Pct. new students from top 10% of H.S. class
Top25perc	Pct. new students from top 25% of H.S. class
F.Undergrad	Number of fulltime undergraduates
P.Undergrad	Number of parttime undergraduates
Room.Board	Room and board costs
Books	Estimated book costs
Personal	Estimated personal spending
PhD	Pct. of faculty with Ph.D.'s
Terminal	Pct. of faculty with terminal degree
S.F.Ratio	Student/faculty ratio
perc.alumni	Pct. alumni who donate
Expend	Instructional expenditure per student
Grad.Rate	Graduation rate

For more information, you can go to U.S.News.

```
data(College)
skimr::skim(College)
```

Table 2: Data summary

Name	College
Number of rows	777
Number of columns	18
_____	
Column type frequency:	
factor	1
numeric	17
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Private	0	1	FALSE	2	Yes: 565, No: 212

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Apps	0	1	3001.64	3870.20	81.0	776.0	1558.0	3624.0	48094.0	
Accept	0	1	2018.80	2451.11	72.0	604.0	1110.0	2424.0	26330.0	
Enroll	0	1	779.97	929.18	35.0	242.0	434.0	902.0	6392.0	
Top10perc	0	1	27.56	17.64	1.0	15.0	23.0	35.0	96.0	
Top25perc	0	1	55.80	19.80	9.0	41.0	54.0	69.0	100.0	
F.Undergrad	0	1	3699.91	4850.42	139.0	992.0	1707.0	4005.0	31643.0	
P.Undergrad	0	1	855.30	1522.43	1.0	95.0	353.0	967.0	21836.0	
Outstate	0	1	10440.67	4023.02	2340.0	7320.0	9990.0	12925.0	21700.0	
Room.Board	0	1	4357.53	1096.70	1780.0	3597.0	4200.0	5050.0	8124.0	
Books	0	1	549.38	165.11	96.0	470.0	500.0	600.0	2340.0	
Personal	0	1	1340.64	677.07	250.0	850.0	1200.0	1700.0	6800.0	
PhD	0	1	72.66	16.33	8.0	62.0	75.0	85.0	103.0	
Terminal	0	1	79.70	14.72	24.0	71.0	82.0	92.0	100.0	
S.F.Ratio	0	1	14.09	3.96	2.5	11.5	13.6	16.5	39.8	
perc.alumni	0	1	22.74	12.39	0.0	13.0	21.0	31.0	64.0	
Expend	0	1	9660.17	5221.77	3186.0	6751.0	8377.0	10830.0	56233.0	
Grad.Rate	0	1	65.46	17.18	10.0	53.0	65.0	78.0	118.0	

Partition the dataset into two parts: *training data (80%) and test data (20%).*

```
set.seed(123)
total_rows <- nrow(College)
train_indices <- sample(1:total_rows, size = 0.8 * total_rows)
train_data <- College[train_indices, ]
test_data <- College[-train_indices, ]
```

Model Fitting

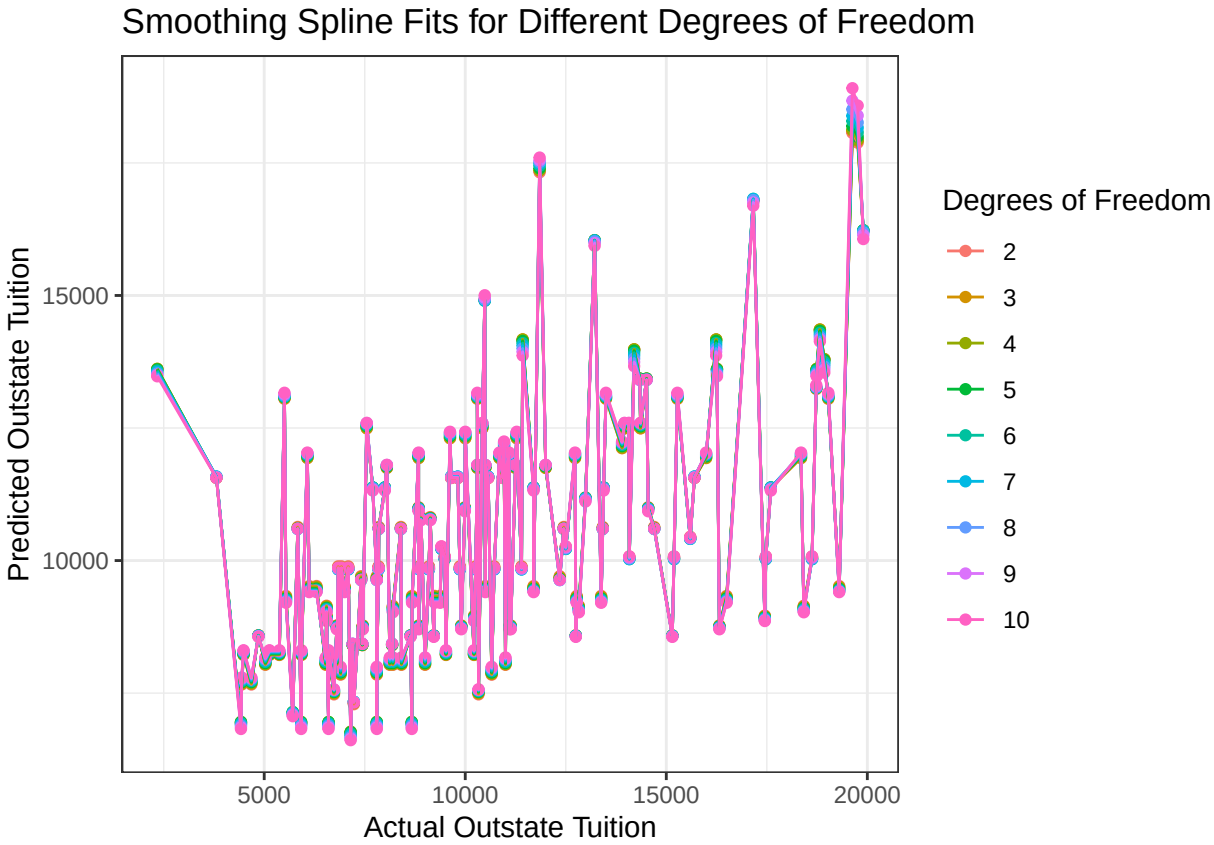
Fit smoothing spline models to predict out-of-state tuition (Outstate) using the percentage of alumni who donate (perc.alumni) as the only predictor, across a range of degrees of freedom.

```
df_results <- data.frame()

for(df in seq(2, 10, by=1)){
  fit.ss <- smooth.spline(train_data$perc.alumni, train_data$Outstate, df=df)

  pred.ss <- predict(fit.ss, x = test_data$perc.alumni)

  pred.ss.df <- data.frame(real = test_data$Outstate, pred = pred.ss$y, df = df)
  df_results <- rbind(df_results, pred.ss.df)
}
```

Plot the model fits for each degree of freedom.

```
ggplot(df_results, aes(x = real, y = pred, colour = factor(df))) +
  geom_point() +
  geom_line(aes(group = df)) +
```

```
  theme_bw() +
  labs(title = "Smoothing Spline Fits for Different Degrees of Freedom",
       x = "Actual Outstate Tuition",
       y = "Predicted Outstate Tuition",
       colour = "Degrees of Freedom")
```



The obtained result of 2.000214 suggests that the optimal degree of freedom is approximately 2, indicating that the model is relatively linear. As the degree of freedom increases, the model transitions from linear to more curved, but too high a degree of freedom may lead to overfitting, meaning the model complexity is too high, reflecting the random noise in the data rather than the true trend. This degree of freedom is recommended by generalized cross-validation (GCV) and is automatically chosen by the smooth.spline() function, to demonstrate the relationship between the percentage of alumni who donate and the out-of-state tuition.

2 Answering Questions

- How would you list all files in the current directory, including hidden ones?
  - use `list.files(all.files = TRUE)`
- What command would you use to find the number of lines in a file named data.txt?
  - Using the `wc -l data.txt` command in a terminal
  - Using `system("wc -l data.txt")` in R
- How can you search for the string "error" in all *.log files in the current directory?
  - Using `grep "error" *.log` in a terminal
- Describe how you would change the permissions of a file named script.sh to make it executable.
  - use the command in the terminal
    1. `chmod +x script.sh`
    2. `ls -l script.sh`
    3. `./script.sh`
- How would you display the last 20 lines of a file named output.log?
  - Using `tail -n 20 output.log` in terminal
- Explain how to combine the contents of file1.txt and file2.txt into a new file named combined.txt.
  - Using `cat file1.txt file2.txt > combined.txt` in the terminal.
- How would you check for the presence of the word "Completed" in a file named status.txt and display the line containing it?
  - Using `grep "Completed" status.txt` in the terminal.
- What command can you use to sort the lines in a file named unsorted.txt in alphabetical order and save the result to a new file named sorted.txt?
  - Using `sort unsorted.txt > sorted.txt` in the terminal.