StateLearning HW 0

Christian Conroy

Feb 2nd, 2018

1. (3 pts) Textbook #2.4.2

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 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

This is a regression problem as indicated by the quantitative response (CEO Salary) where we are interested in inference. Our goal here is not to predict the CEO salary but to understnd the relationships between record profit, number of employees, and industry and the CEO salary. The n is 500 and the p is 4.

(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.

This is a classification problem as indicated by the qualitative response (success or failure) where we are focused on prediction. Our goal here is to use what we have recorded on the product to predict whether the new product will be a success or failure. The n is 20 and the p is 14.

(c) We are interest in predicting the % change in the USD/Euroexchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

This is a regression problem as indicated by the quantitative response (% change in USD/Euroexchange) where we are interested in prediction. We do not care as much about the relationships between the different types or market % change and the % change in the USD Euroexchange as we do about using those market % change numbers to predict what the next % change in the USD/Euroexchange will be. The n is 52 and the p is 4.

2. (3 pts) Textbook #2.4.4

You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. Classification can be used in machine learning for image recognition. For example, if someone is designing an app that can tell you what flower you are looking at through the lens of your smartphone camera, they will need to first train the underlying model on millions of images of flowers, otherwise called labels or targets. The predictors are the pixels that each flower image is broken down into and the response is the flower genus. The goal of course is prediction because the app should be able to apply a label to a flower that you come across in the wild.

2. Classification can be used to identify the factors that result in certain individuals receiving less education than others. While there is a grey area in that one can treat education as quantitative (consecutive numbers for grades) or qualitative (primary school, secondary school, postsecondary school, etc.), often the latter is more interesting. One could build a multinomial probit model to evaluate the relationship between variables like the education level of one's parents, household income, number of siblings, number of schools within a certain distance, etc. as the predictors and the level of education that someone has as the response. The focus here of course is on inference.

3. Classification can be used to assess the factors that influence whether someone may or may not be accepted into a college or certain colleges. The predictors could include SAT scores, ethnicity, an index of extracurriculurs, household income, legacy status, education level of one's parents, etc. and the response would consist of a binary variable coded on admission or rejection. While one could feasibly be interested in prediction, the most interesting reason to conduct this type of analysis is infernece.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. A local government is deciding whether to provide free nutrition packets to households to alleviate child malnutrtion rates. They carry out a randomized control trial where a treatment group recieves the packets and a control group does not. The nutrition status of both groups is measured before distribution of the nutrition packets. The nutrition status of both groups is later measured at the conclusion of the period of the trial. A regression can be used to evaluate the causal impact of the nutrtion packets on alleviating malnutrtion rates. Here we'd be interested in inference because the focus is on whether this treatment (predictor) has a positive relationship with nutrition rates (response).

2. A company wants to assess how changes in different macroeconomic conditions in a country affect the market size for their product in that country. They put together a

large panel data set of macroeconomic indicators (predictors) for countries and the market size for their product (response) in those respective countries over time. The goal is to be able to predict the change in their market size when macroeconomic indicators shift so that they can change how much they pay for floor allocation at local retail outlets, how much production they demand from factories on the ground, and how much inputs they source into those local areas. The goal of any longitudinal regression model they employ will be prediction in this case.

3. A city wants to start a bike share program. It initiates a one year pilot program. The goal is to assess whether the bikeshare program resulted in less cars being on the road and reduced traffic. The predictor would be the treatment (i.e. the bikeshare program) and the response would be traffic density. Like the example above, there would be a control group area where the bikeshare program does not reach to. The goal here is inference.

(c) Describe three real-life applications in which cluster analysis might be useful.

1. Market segmentation for a business deciding how to allocate marketing budget to different consumer groups
2. Human genetic clustering or DNA sequencing
3. Social network analysis

3. (3 pts) Extra #4

a) Modify the function myclosest() so that it uses exactly k neighbors instead of 100 to classify a test digit. The new function should have two arguments, namely mydigit and k.

b) Demonstrate the modified function by trying to classify a test digit of your choice. Find a value of k such that the classification is correct and another value of k < 1000 such that the classification of the same test digit is incorrect.

```r
data <- load("mnist_all.Rdata")

# predict a digit from the MNIST training set from the most frequent digit
# among the 100 closest neighbors in the training set
myclosest = function(mydigit, k){
digit.dist = function(j){
return(sqrt(mean((test$x[mydigit,] - train$x[j,])^2) ) )
}
mnist.distances = sapply(1:60000,FUN = digit.dist)
myclosest = head(order(mnist.distances),k)
mytable <- table(train$y[myclosest])
myindex = which.max(mytable)
return(as.numeric(names(mytable[myindex])))
}
# Try it. Prediction and actual value of digit 234 in the test set
```

```r
# Correct Classification
c(test$y[160], myclosest(160, 100))

## [1] 4 4

# Incorrect Classification
c(test$y[160], myclosest(160, 500))

## [1] 4 1

# Come back to
```

4. (5 pts) Textbook #2.4.8

(b) Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```r
data("College")
head(College)

##                              Private Apps Accept Enroll Top10perc
## Abilene Christian University     Yes 1660   1232    721        23
## Adelphi University               Yes 2186   1924    512        16
## Adrian College                   Yes 1428   1097    336        22
## Agnes Scott College              Yes  417    349    137        60
## Alaska Pacific University        Yes  193    146     55        16
## Albertson College                Yes  587    479    158        38
##                              Top25perc F.Undergrad P.Undergrad Outstate
## Abilene Christian University        52        2885         537     7440
## Adelphi University                  29        2683        1227    12280
## Adrian College                      50        1036          99    11250
## Agnes Scott College                 89         510          63    12960
## Alaska Pacific University           44         249         869     7560
## Albertson College                   62         678          41    13500
##                              Room.Board Books Personal PhD Terminal
## Abilene Christian University       3300   450     2200  70       78
## Adelphi University                 6450   750     1500  29       30
## Adrian College                     3750   400     1165  53       66
## Agnes Scott College                5450   450      875  92       97
## Alaska Pacific University          4120   800     1500  76       72
## Albertson College                  3335   500      675  67       73
##                              S.F.Ratio perc.alumni Expend Grad.Rate
## Abilene Christian University      18.1          12   7041        60
## Adelphi University                12.2          16  10527        56
## Adrian College                    12.9          30   8735        54
## Agnes Scott College                7.7          37  19016        59
## Alaska Pacific University         11.9           2  10922        15
## Albertson College                  9.4          11   9727        55
```

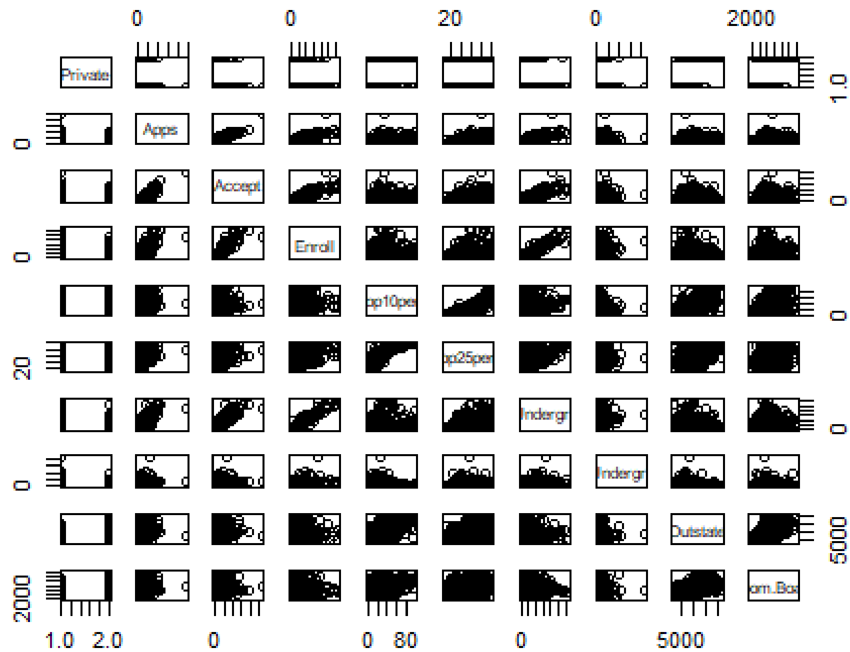
```r
write.csv(College, "College.csv")
college <- read.csv('College.csv', header = TRUE)

rownames(college) = college[,1]
fix(college)

college = college[,-1]
fix(college)
```

i. Use the summary() function to produce a numerical summary of the variables in the data set.

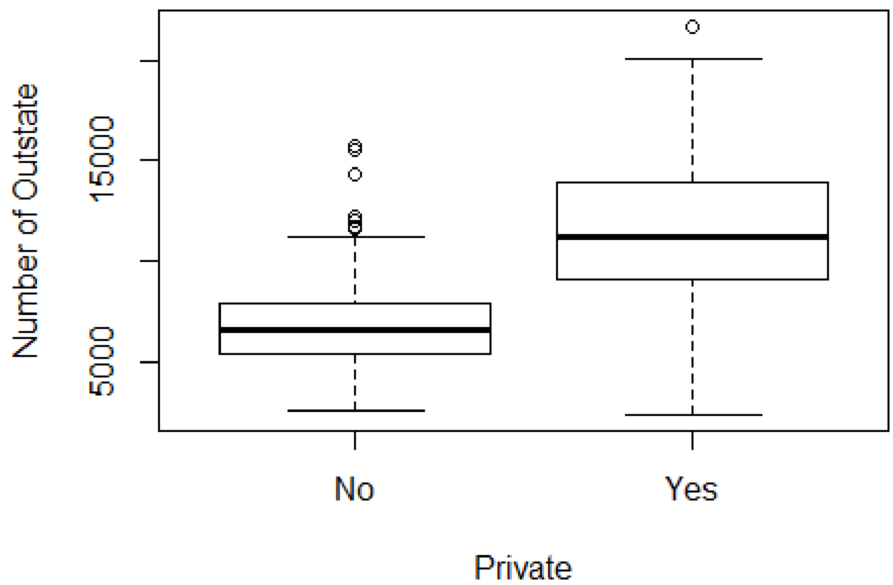
```r
summary(college)

##  Private        Apps           Accept          Enroll       Top10perc   
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.0  
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.0  
##            Median : 1558   Median : 1110   Median : 434   Median :23.0  
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.6  
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.0  
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.0  
##    Top25perc      F.Undergrad     P.Undergrad       Outstate    
##  Min.   :  9.0   Min.   :  139   Min.   :    1   Min.   : 2340  
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95   1st Qu.: 7320  
##  Median : 54.0   Median : 1707   Median :  353   Median : 9990  
##  Mean   : 55.8   Mean   : 3700   Mean   :  855   Mean   :10441  
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967   3rd Qu.:12925  
##  Max.   :100.0   Max.   :31643   Max.   :21836   Max.   :21700  
##    Room.Board       Books         Personal         PhD       
##  Min.   :1780   Min.   :  96   Min.   : 250   Min.   :  8.0  
##  1st Qu.:3597   1st Qu.: 470   1st Qu.: 850   1st Qu.: 62.0  
##  Median :4200   Median : 500   Median :1200   Median : 75.0  
##  Mean   :4358   Mean   : 549   Mean   :1341   Mean   : 72.7  
##  3rd Qu.:5050   3rd Qu.: 600   3rd Qu.:1700   3rd Qu.: 85.0  
##  Max.   :8124   Max.   :2340   Max.   :6800   Max.   :103.0  
##     Terminal       S.F.Ratio     perc.alumni        Expend     
##  Min.   : 24.0   Min.   : 2.5   Min.   : 0.0   Min.   : 3186  
##  1st Qu.: 71.0   1st Qu.:11.5   1st Qu.:13.0   1st Qu.: 6751  
##  Median : 82.0   Median :13.6   Median :21.0   Median : 8377  
##  Mean   : 79.7   Mean   :14.1   Mean   :22.7   Mean   : 9660  
##  3rd Qu.: 92.0   3rd Qu.:16.5   3rd Qu.:31.0   3rd Qu.:10830  
##  Max.   :100.0   Max.   :39.8   Max.   :64.0   Max.   :56233  
##    Grad.Rate    
##  Min.   : 10.0  
##  1st Qu.: 53.0  
##  Median : 65.0  
##  Mean   : 65.5  
##  3rd Qu.: 78.0  
##  Max.   :118.0  
```

ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

```
pairs(college[,1:10])
```



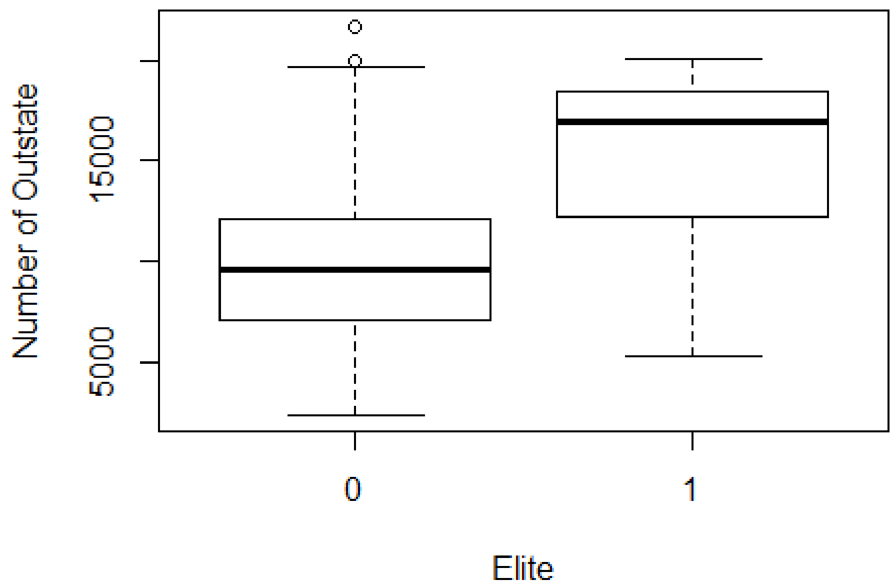
iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.



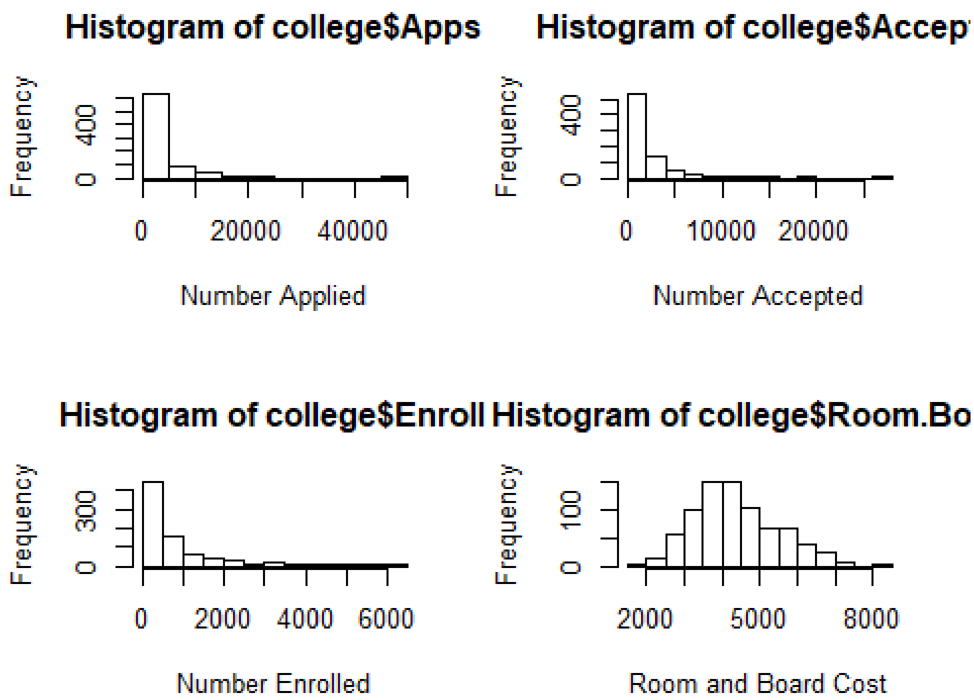
iv. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %.

Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

```
college$Elite <- ifelse((college$Top10perc > 50), 1, 0)
summary(factor(college$Elite))

##   0   1
## 699  78
```



v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.



vi. Continue exploring the data, and provide a brief summary of what you discover

```
cor(college[-c(1, 19)])

##                Apps  Accept  Enroll Top10perc Top25perc F.Undergrad
## Apps         1.0000  0.9435  0.8468    0.3388    0.3516      0.8145
## Accept       0.9435  1.0000  0.9116    0.1924    0.2475      0.8742
## Enroll       0.8468  0.9116  1.0000    0.1813    0.2267      0.9646
## Top10perc    0.3388  0.1924  0.1813    1.0000    0.8920      0.1413
## Top25perc    0.3516  0.2475  0.2267    0.8920    1.0000      0.1994
## F.Undergrad  0.8145  0.8742  0.9646    0.1413    0.1994      1.0000
## P.Undergrad  0.3983  0.4413  0.5131   -0.1054   -0.0536      0.5705
## Outstate     0.0502 -0.0258 -0.1555    0.5623    0.4894     -0.2157
## Room.Board   0.1649  0.0909 -0.0402    0.3715    0.3315     -0.0689
## Books        0.1326  0.1135  0.1127    0.1189    0.1155      0.1155
## Personal     0.1787  0.2010  0.2809   -0.0933   -0.0808      0.3172
## PhD          0.3907  0.3558  0.3315    0.5318    0.5459      0.3183
## Terminal     0.3695  0.3376  0.3083    0.4911    0.5247      0.3000
## S.F.Ratio    0.0956  0.1762  0.2373   -0.3849   -0.2946      0.2797
## perc.alumni -0.0902 -0.1600 -0.1808    0.4555    0.4179     -0.2295
## Expend       0.2596  0.1247  0.0642    0.6609    0.5274      0.0187
## Grad.Rate    0.1468  0.0673 -0.0223    0.4950    0.4773     -0.0788
##             P.Undergrad Outstate Room.Board   Books Personal    PhD
## Apps             0.3983   0.0502     0.1649 0.13256   0.1787 0.3907
## Accept           0.4413  -0.0258     0.0909 0.11353   0.2010 0.3558
## Enroll           0.5131  -0.1555    -0.0402 0.11271   0.2809 0.3315
## Top10perc       -0.1054   0.5623     0.3715 0.11886  -0.0933 0.5318
## Top25perc       -0.0536   0.4894     0.3315 0.11553  -0.0808 0.5459
```

```
## F.Undergrad      0.5705  -0.2157    -0.0689  0.11555   0.3172  0.3183
## P.Undergrad      1.0000  -0.2535    -0.0613  0.08120   0.3199  0.1491
## Outstate        -0.2535   1.0000     0.6543  0.03885  -0.2991  0.3830
## Room.Board      -0.0613   0.6543     1.0000  0.12796  -0.1994  0.3292
## Books            0.0812   0.0389     0.1280  1.00000   0.1793  0.0269
## Personal         0.3199  -0.2991    -0.1994  0.17929   1.0000 -0.0109
## PhD              0.1491   0.3830     0.3292  0.02691  -0.0109  1.0000
## Terminal         0.1419   0.4080     0.3745  0.09995  -0.0306  0.8496
## S.F.Ratio        0.2325  -0.5548    -0.3626 -0.03193   0.1363 -0.1305
## perc.alumni     -0.2808   0.5663     0.2724 -0.04021  -0.2860  0.2490
## Expend          -0.0836   0.6728     0.5017  0.11241  -0.0979  0.4328
## Grad.Rate       -0.2570   0.5713     0.4249  0.00106  -0.2693  0.3050
##             Terminal S.F.Ratio perc.alumni  Expend Grad.Rate
## Apps          0.3695    0.0956     -0.0902  0.2596   0.14675
## Accept        0.3376    0.1762     -0.1600  0.1247   0.06731
## Enroll        0.3083    0.2373     -0.1808  0.0642  -0.02234
## Top10perc     0.4911   -0.3849      0.4555  0.6609   0.49499
## Top25perc     0.5247   -0.2946      0.4179  0.5274   0.47728
## F.Undergrad   0.3000    0.2797     -0.2295  0.0187  -0.07877
## P.Undergrad   0.1419    0.2325     -0.2808 -0.0836  -0.25700
## Outstate      0.4080   -0.5548      0.5663  0.6728   0.57129
## Room.Board    0.3745   -0.3626      0.2724  0.5017   0.42494
## Books         0.1000   -0.0319     -0.0402  0.1124   0.00106
## Personal     -0.0306    0.1363     -0.2860 -0.0979  -0.26934
## PhD           0.8496   -0.1305      0.2490  0.4328   0.30504
## Terminal      1.0000   -0.1601      0.2671  0.4388   0.28953
## S.F.Ratio    -0.1601    1.0000     -0.4029 -0.5838  -0.30671
## perc.alumni   0.2671   -0.4029      1.0000  0.4177   0.49090
## Expend        0.4388   -0.5838      0.4177  1.0000   0.39034
## Grad.Rate     0.2895   -0.3067      0.4909  0.3903   1.00000
```

There are some interesting correlations within the data that might be worth looking further into. While it makes sense that schools that receive more applicants would accept more (0.9435), it is surprising that there is such a strong relationship between out of state and expenditures. Perhaps schools that are getting more out of state tuitions and thereby charging more out of state tuition end up making things more expensive because they feel the out of state student population that already pays high tuition can afford it. It is equally interesting to see areas where the correlations are not strong when one would expect them to be. For example, one would expect that people might be deterred from going to a university with a high room and board cost. However, not only is this not the case, but the correlation indicates that there is a slight positive correlation, though not one strong enough to necessarily require further analysis.

5. (5 pts) Extra #6

This problem uses the Shiny app at https://keeganhines.shinyapps.io/bias_variance/ . Before working on this problem, load the app, read the explanation, play with the slider and the "Generate New Data" button, and answer the questions at the bottom of the page ("Check your understanding") for yourself or discuss them with others.

Model complexity = degree of the polynomial that is being fitted.

Check your understanding: What the pros and cons of using functions of high and low complexity?

Functions of high complexity make the model more flexible and able to deal with potential non-linear patterns in the data. However, the risk is that we end up fitting the model perfectly to the sample or training dataset we are using while simultaneously making it less likely to fit another sample or training dataset we might draw from the same population. The benefit of low complexity therefore is that it will be more generalizable across different samples drawn from the same population.

Would an Order-1 model have high or low bias? High or low variance?

An order-1 model would have high bias but low variance.

Would an Order-15 model have high or low bias? High or low variance?

An order-15 model with have low bias but high variance.

a) Make 10 different simulations with model complexity = 1. Compute the average Residual SSE and find the approximate range of the highest order coefficient for these 10 simulations. This is a measure for the baseline variance for a low complexity model.

Residual SSE Calcs Simulation 1: 43.12 Simulation 2: 98.79 simulation 3: 123.02 Simulation 4: 136.62 Simulation 5: 83.12 simulation 6: 96.14 Simulation 7: 132.32 Simulation 8: 66.42 simulation 9: 120.58 Simulation 10: 115.82

b) Make 10 different simulations with model complexity = 10. Compute the average Residual SSE. Which coefficient has the largest range in this case? What is that range? This is a measure for the variance for a high complexity model.

Residual SSE Calcs and Range of Highest Order Coefficient Simulation 1: 27.23 Simulation 2: 10.57 simulation 3: 21.61 Simulation 4: 21.34 Simulation 5: 24.57 simulation 6: 36.87 Simulation 7: 14.5 Simulation 8: 30.9 simulation 9: 3.68 Simulation 10: 10.33

The fifth order coefficient appears to has the largest range, with a minimum of -10000 and maximum of 8000, or a range of 18000

c) How do your results illustrate the bias - variance trade-off? The answer should be a short paragraph.

The results illustrate the bias-variance trade-off because although the residual SSE tends to be lower in the model with 10-order complexity, indicating lower bias, there is also a much higher range for coefficient values across the simulations, indicating a higher variance.

d) For which model complexity between 1 and 15 do you typically obtain a curve which is most similar and overall close to the unknown curve that is to be estimated? Try multiple simulation for several different model complexities, summarize what you see, and explain your answer. Pictures or numerical results are not required.

Model complexity =3 appears to opbtain the curve that is most similar to the unknown curve. Even as we generate new data, the blue curve continues to appear alongside the black curve with relative proximity.