ANSWER SHEET

01:960:486 COMPUTING AND GRAPHICS IN APPLIED STATISTICS May 7, 2020 **Test #2**

Please submit your answers into Canvas today by 3:15 pm. I will be online for emergency situations (for students who cannot access the test, cannot submit the test, etc.). Assume I am proctoring the test and I have no knowledge regarding the questions on the test. Just state any assumption you make if you have difficulty understanding a question and/or parts of a question. You do not need to carry through calculations. It is OK to leave results as 3/8 or (20 + 3)/(100 + 50).

There are 13 questions on 4 pages with room on each page for your answers. You must submit these 4 pages and any additional pages you need into Canvas. Just be sure to clearly label your answers if you use additional pages.

Reminder: The test is open book, open notes, open online resources.

You are to work alone. Rutgers Honors Pledge is in effect.

1. (15 pts) Given the following results from an All Possible Regressions of Y on X1, X2, X3, X4, X5:

Model Number	Variables in Model	Adjusted R-square
1	X3	0.97622
2	X2	0.66077
3	X4	0.37346
4	X1	0.07996
5	X5	-0.00359
6	X3 X2	0.97777
7	X4 X3	0.97719
8	X3 X1	0.97637
9	X5 X3	0.97634
10	X5 X4	0.95772
11	X5 X2	0.85562
12	X4 X2	0.71647
13	X2 X1	0.67041
14	X4 X1	0.45987
15	X5 X1	0.08471
16	X5 X4 X3	0.98511
17	X5 X3 X2	0.97872
18	X3 X2 X1	0.97790
19	X4 X3 X2	0.97769
20	X4 X3 X1	0.97766
21	X5 X3 X1	0.97663
22	X5 X4 X2	0.95849
23	X5 X4 X1	0.95773
24	X5 X2 X1	0.86274
25	X4 X2 X1	0.71549
26	X5 X4 X3 X1	0.98521
27	X5 X4 X3 X2	0.98508
28	X5 X3 X2 X1	0.97863
29	X4 X3 X2 X1	0.97782
30	X5 X4 X2 X1	0.95833
31	X5 X4 X3 X2 X1	0.98515

a) If you were to perform a backward elimination stepwise regression using the adjusted R-square criterion arriving at a model with only the intercept, what is the sequence to remove all 5 variables?

b) If you were to perform a forward stepwise regression using the adjusted R-square criterion arriving at a model with all 5 variables, what is the sequence to add all 5 variables?

c) If you were to perform a sequential (bi-directional) stepwise regression using the adjusted R-square criterion arriving at a final model, what is the sequence to add/remove variables as you arrive at the final model? Be sure to state your final model.

a) $X_2 \to X_1 \to X_5 \to X_4 \to X_3$

b) $X_3 \to X_2 \to X_5 \to X_4 \to X_1$

c) $X_3 \to X_2 \to X_5 \to X_4 \to$ Drop X_2

$R^2_{adj} = .98508$ $R^2_{adj} = .98511$

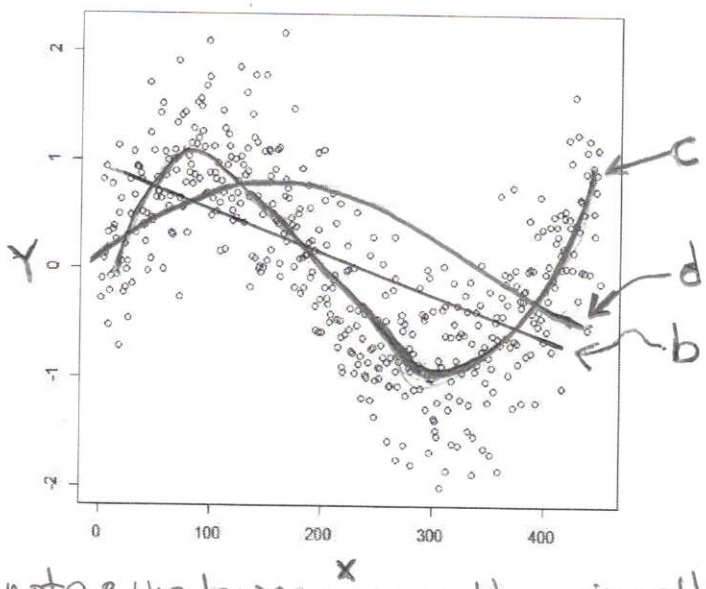
$\to$ Add X_1 (X_1, X_3, X_4, X_5)

$R^2 = .98521$

largest R^2_{adj}

STOP

2. (15 pts) You are given the following data (n=450) shown in the plot below.



note: the lowess smooth using all data points should have the lowest number of peaks/valleys among all lowess smooths

Here are the first 20 observations:

x	y	
1	-0.3	
2	-0.1	
3	0.8	
4	0.1	
5	0.9	
6	0.3	weight = ?
8	-0.5	
9	-0.2	
10	-0.1	weight= ?
11	0.8	
12	0.4	
13	-0.1	
16	1.1	
17	0.5	
18	-0.7	
19	0.6	
20	0.1	

weight = 0

weight = 1

width = 6

half-width = 3

a) Suppose that you decide to perform a lowess smooth using a span ~ 0.013 which results in **6** data points. Suppose you are interested in performing the smooth at x=10. Use the tri-cube weight function found in the document **Loess_Chapter.pdf from Lecture 18** to list the two requested weights in the table above. Just write the formula filled in with the appropriate values, no need to carry through the calculation.

b) On the plot above, draw as best as you can a straight line smooth of the data and label it with a 'b'. (smoothed y=mx+b)

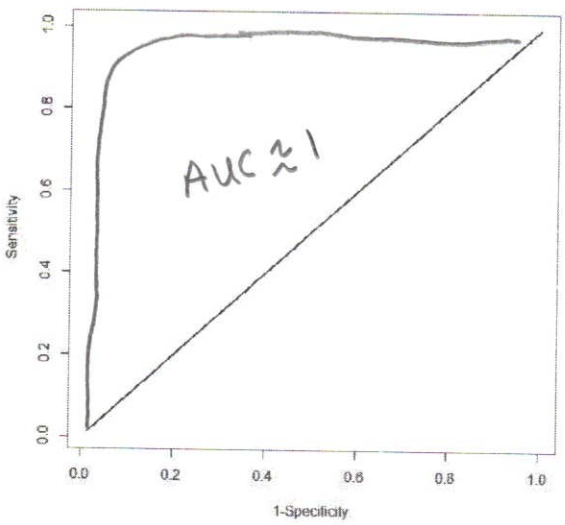
c) On the plot above, draw as best as you can what a lowess smooth with a span equating to 10 data points might look like and label it with a 'c'. No calculations are needed.

d) On the plot above, draw as best as you can what a lowess smooth with a span equating to all n=450 data points might look like and label it with a 'd'. No calculations are needed.

e) In a lowess smooth with n data points, what is the number of regressions that need to be performed?

# of distinct X-levels

3. (10 pts) Suppose a simple test is developed to detect whether a person has COVID-19 disease.



a) Using the figure to the left, plot the ideal ROC curve you would want for this simple test.

b) Suppose the following 2x2 table resulted from the simple test and a "gold standard" test applied to a sample of 120 individuals with high risk for COVID-19 disease.

		Gold Standard Test Result: Positive	Gold Standard Test Result: Negative	
Simple Test Result	Positive	35	10	45
	Negative	5	70	75
		40	80	

Compute the accuracy, sensitivity, specificity, positive predictive power, and negative predictive power for the simple test.

Accuracy
$\frac{105}{120}$

Sensitivity
$\frac{35}{40}$

Specificity
$\frac{70}{80}$

Positive Predictive Power
$\frac{35}{45}$

Negative Predictive Power
$\frac{70}{75}$

4. (6 pts) Briefly describe an experimental situation where a Bayes Regression with 1 predictor is likely to perform better than an OLS Regression with 1 predictor.

If there is strong, reliable prior information on B_0, B_1.

5. (6 pts) A researcher has predictor variables X1 though X40 to include in a regression model. In addition to Model Selection, the researcher is also exploring the possibility to use Principal Component Regression (PCR).

a) State one advantage for using the Model Selection approach.

The ability to explore various subsets and its simplicity to explain results to the researcher.

b) State one advantage for using the PCR approach.

The ability to perform dimension reduction.

6. (10 pts) **Power and Sample Size Questions**

a) The probability of incorrectly rejecting the null hypothesis is called Type I Error rate.

b) The probability of correctly rejecting the null hypothesis is called Power of the test.

c) Briefly describe the impact to the Effect Size in a study with two treatments if the sample sizes for each treatment is increased.

stays the same Effect size is independent of sample size

d) Briefly describe the impact to Power in a study with two treatments if the sample sizes for each treatment is increased.

Power is increased.

e) Briefly describe the impact to the Effect Size in a study with two treatments if the variance in the sampled population is decreased. (You can assume a common variance for the study.)

Effect size is increased.

7. (8 pts) a) Briefly describe an experimental situation where Ridge regression is preferred over LASSO regression.

when all variables must be retained in the model - see fat content example.

b) Briefly describe an experimental situation where LASSO regression is preferred over Ridge regression.

In presence of high multicollinearity when only a few of the features are preferred.

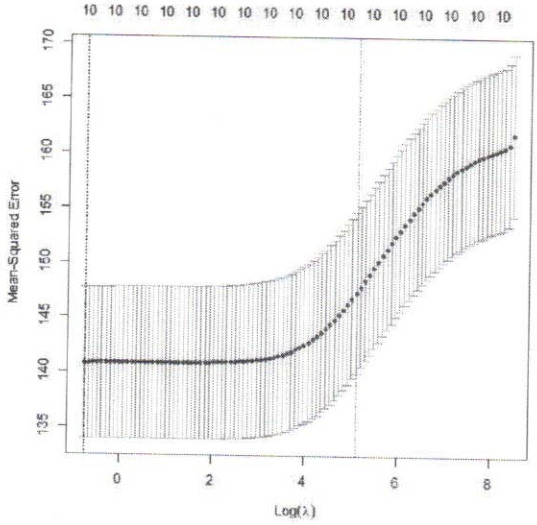
8. (5 pts) Select the statement below that is most true when Ridge regression is compared to OLS regression.

a) Ridge has larger bias, larger variance. **b)** Ridge has larger bias, smaller variance.

c) Ridge has smaller bias, larger variance. **d)** Ridge has smaller bias, smaller variance.

Answer is b) larger bias, smaller variance

9. (5 pts) A Ridge regression was performed on 10 variables, V1-V10. The plot below was generated. Briefly explain how the information from this plot is used in the next step of conducting a Ridge regression analysis.



Note the minimum MSE occurs at lambda=0.47, log(0.47)= - 0.755
MSE at lambda=0.47 is equal to 140.8

lambda = 0.47 allows the next step to compute regression coefficients

10. (5 pts) Briefly explain why backward elimination selection of variables is computationally slower than forward selection of variables if only a few variables are useful in the model.

Backward elimination performs k regressions to start the process and then grinds through the procedure to eliminate variables one at a time.

11. (5 pts) Briefly describe an experimental situation where a backward elimination of variables may not even be possible to execute.

If linear combinations of predictors are exactly equal to another predictor then $(X'X)^{-1}$ does not exist.

12. (5 pts) Briefly explain why Cross Validation metrics tend to be worse in training data sets than in test data sets.

Overfitting tends to be a concern when working with training data sets.

13. (5 pts) Briefly describe an experimental situation where a Naïve Bayes classifier is likely to perform better than a Logistic Regression.

see medium.com
Search "Naive Bayes vs Logistic Regression"
"If the data set follows the bias, then Naive Bayes will perform better than logistic.
Also, if features are conditionally independent.

ENJOY YOUR SUMMER AND BE SURE TO TAKE CARE OF YOURSELF!