01:960:463 REGRESSION METHODS **May 12, 2020** **Final Exam**

Please submit your answers into Canvas tonight by 11:15 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 11 questions on 5 pages. Based on experience gained from administering my other course exam, the suggested method for completing the exam is to write your answers on your own paper and take pictures of the pages to submit into Canvas. You only need to submit pages with answers. Just be sure to clearly label your answers. Also, do not wait until 11:15 pm to assemble your answers, Canvas will stop accepting submissions promptly at 11:15 pm.

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

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

1. (12 pts) Given the following results from an All Possible Regressions of Y, a continuous variable, 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 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.

2. (14 pts) A researcher ran a logistic regression of Y on 4 predictors X1, X2, X3, X4 in R with the following results:

```
Coefficients:
             Estimate Std. Error
(Intercept)  1.001311   1.177228
X1          -0.042552   0.038687
X2          -0.010615   0.006916
X3           0.373837   0.367666
X4           1.452990   0.463061
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 202.82  on 167  degrees of freedom
Residual deviance: 184.21  on 163  degrees of freedom
AIC: 194.21

Number of Fisher Scoring iterations: 4

Analysis of Deviance Table

Model: binomial, link: logit

Response: Y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                   167     202.82
X1    1   0.9201       166     201.90 0.337444
X2    1   4.8749       165     197.03 0.027249
X3    1   2.7651       164     194.26 0.096338
X4    1  10.0529       163     184.21 0.001521
---
```

```
Logistic regression predicting Y

                crude OR(95%CI)    adj. OR(95%CI)    P(LR-test)
X1 (cont. var.) 0.97 (0.9,1.04)    0.96 (0.89,1.03)  0.266
X2 (cont. var.) 0.99 (0.97,1)      0.99 (0.98,1)     0.106
X3: 1 vs 0      1.84 (0.94,3.61)   1.45 (0.71,2.99)  0.311
X4: 1 vs 0      4.76 (2.01,11.26)  4.28 (1.73,10.6)  0.002

Log-likelihood = -92.1045
No. of observations = 168
AIC value = 194.2091
```

a) What is the Wald test statistic for testing H0: β_4 =0?

b) What is the LRT statistic for testing
H0: all variables can be dropped from the model?

c) Test the hypothesis below at the 0.05 level
H0: β_1= β_2; β_4=0 . Specify any model(s) that need to be run if not already displayed in the output.

d) X4 is an indicator variable taking on the values 0 or 1. Interpret the Odds Ratio shown for X4 from the overall model for the researcher.

3. (7 pts) A researcher fit the following logistic model
P(Y=1) = E(Y) = exp(βX)/(1 + exp(βX)) to the data (n=4): (X,Y) =(0,0), (0.1,1), (1,1), (10,1)
Based on the experimental application, there are only two possible values for β, 0 or 1. Find the maximum likelihood estimator for β.

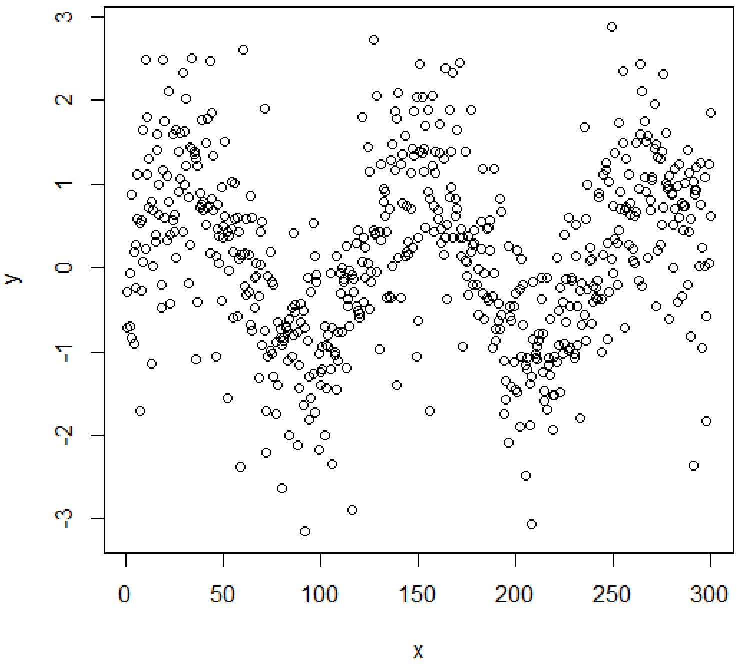
4. (5pts) A researcher performed an OLS regression of Y on X1 and X2 where the predicter variable levels have no duplicates, that is, each (x1i, x2i) is unique for i=1, 2, . . ., n. The researcher wants to check the assumption of constancy of error variance. Describe a graphical technique that can be used by the researcher to determine if weighted least squares should be performed.

5. (6 pts) We discussed using Logistic Regression as a classifier.

a) Draw the ideal ROC curve for a Logistic classifier. Be sure to properly label the axes.

b) Explain with adequate details how to generate a point on this ideal ROC curve that is based on the underlying data corresponding to the ideal ROC curve.

6. (10 pts) A researcher plans to perform lowess smoothing of time series data (n=600 with 2 observations for every timepoint) shown in the plot below. Instead of using the tri-cubic weight function we learned in class (see Loess_Chapter.pdf from Lecture 10), the researcher decides to use the weight function: f(z) = (1 - |z|) for |z| <= 1
f(z) =0 for |z| > 1.



Here are the first 20 data points:

x	y
1	-0.3
1	-0.7
2	-0.1
2	-0.7
3	0.9
3	-0.8
4	0.2
4	-0.9
5	0.3
5	-0.2
6	1.1
6	0.6
8	-0.3
8	0.6
9	0.1
9	1.6
10	0.2
10	2.5
11	1.1
11	1.8

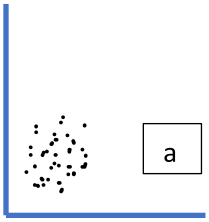
The researcher decided to use a span of .02 which results in including 12 data points (600*.02 = 12) for each regression performed.

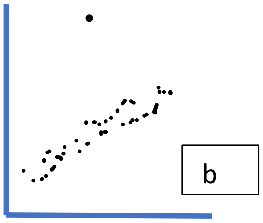
a) For the smooth to be performed at x=6, compute the researcher defined weights for x=6 and at x=2.

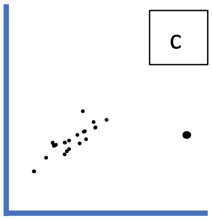
b) On your answer sheet draw the smooth you would expect if a span of 1 (all data) is used.

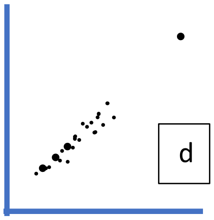
c) How many regressions need to be performed for this data?

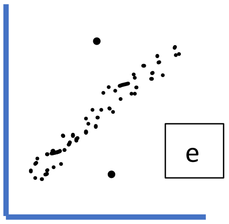
7. (12 pts) Given the following scatter plots (a,b,c,d,e):











Using a simple linear regression, choose **THE** (select only one) scatter plot above that **best** shows:

a) the deletion of a point with leverage will have no effect on the fitted regression. _______________

b) the deletion of a point with leverage will have a large effect on the fitted regression. ____________

c) the deletion of a point with small leverage will have a large effect on the fitted regression. __________

d) the variation in the predictor is small and hence the slope estimate may be unreliable. ___________

8. (6 pts) **Forecasting car sales.** Forecasts of automotive vehicle sales in the United States provide the basis for financial and strategic planning of large automotive corporations. The following forecasting model was developed for y, total monthly passenger car and light truck sales (in thousands):

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$

where

x_1 = Average monthly retail price of regular gasoline

x_2 = Annual percentage change in GNP per quarter

x_3 = Monthly consumer confidence index

x_4 = Total number of vehicles scrapped (millions) per month

x_5 = Vehicle seasonality

The model was fitted to monthly data collected over a 12-year period (i.e., n = 144 months) with the following results:

yˆ = −676.42 − 1.93x1 + 6.54x2 + 2.02x3 + .08x4 + 9.82x5

R2 = .856

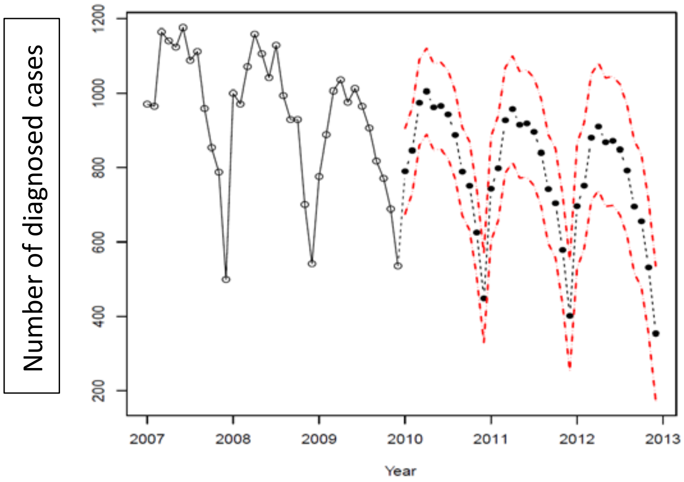
Durbin–Watson d = 1.01

(a) Is there sufficient evidence to indicate that the regression errors are positively correlated. Test using α = .05.

(b) Comment on the validity of the inference concerning model adequacy considering the result of part a.

Hint: Critical Values for the Durbin Watson Test can be found here.
http://berument.bilkent.edu.tr/DW.pdf
You can copy the URL above into your browser or use the tables in the textbook.

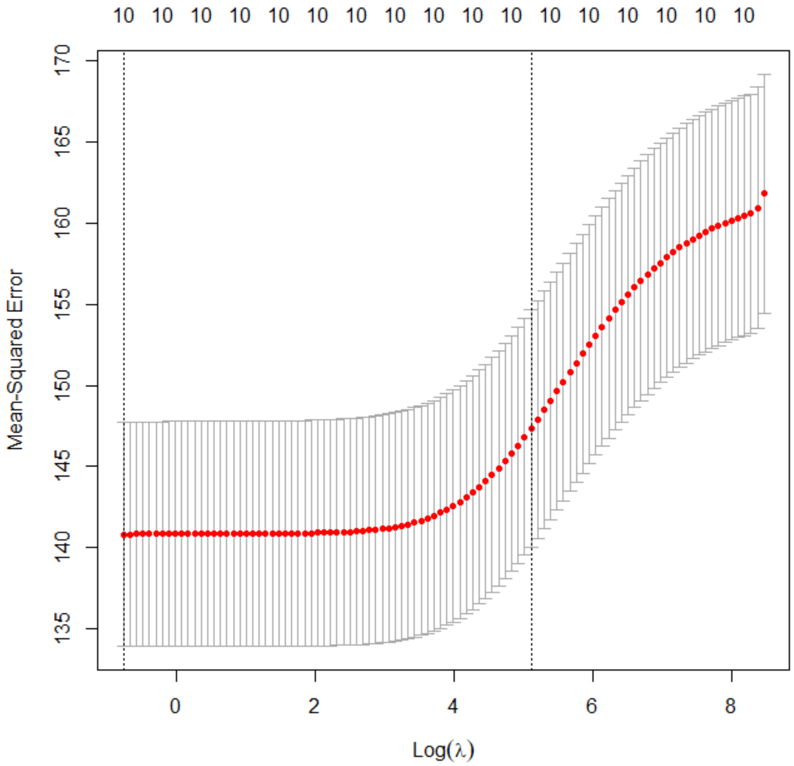
9. (12 pts) The figure below displays the results of fitting a **first-order autoregressive model** to 2007-2010 data on the number of diagnosed cases of a disease. A 95% prediction interval for the number of diagnosed cases in June 2010 is (830, 1120).



a. Provide an interpretation of the prediction interval.
b. What assumption on R_t is crucial for the prediction interval to be valid?
c. Is the model stationary? Explain.
d. Provide a simple graphical technique to estimate ϕ.

10. (6pts) Suppose you are asked to generate a Poisson regression analysis of Y = number of diagnosed cases with a disease using a predictor X = total family income. You were provided with 6 data points (x,y) and asked to fit a saturated model.
a) Write the saturated model equation you would use.
b) What would the estimate of σ^2 be? Provide a short explanation for your answer.

11. (10 pts) A researcher wishes to build a regression model with a continuous response Y and 10 predictors X1, X2, . . ., X10. The researcher realizes there is a high degree of collinearity and so initiates a Ridge regression analysis and generates the plot below.



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

a) Recommend the next step to be taken in the Ridge regression analysis based on the plot above.
b) Describe an experimental situation where the researcher may decide against using a variable selection technique such as LASSO even in the presence of a high degree of collinearity.
c) If you were to design a single software package that guides a researcher in choosing among analyses by OLS, Ridge, and LASSO, state one feature you would build into the package.

TAKE CARE OF YOURSELF AND ENJOY YOUR SUMMER! CONGRATULATIONS TO GRADUATING SENIORS.