

Stat 203 Final Exam

Due: Tuesday, March 16 at noon (12 P.M.) Please submit paper copy to my office.

Instructions: There are 4 problems, please submit each problem on separate sheet(s) and put your name on each sheet. Please turn in all of your R code and R output. **Submit only the plots and output that support your analysis. Errors in additional plots/analyses can be penalized.** For each question, explain clearly (i) your objectives, (ii) any hypotheses that you are testing, (iii) the statistical procedure, and (iv) the statistical support for your conclusions.

Honor Code: Please respect the honor code in completing this exam. You can use books and computers, but not other people.

1. Management of a company that develops websites was interested in determining which variables have the greatest impact on the number of websites developed and delivered to customers per quarter. Data were collected on website production output for 13 three-person website development teams, from January 2001 through August 2002. Each line of the data file `WebsiteDeveloper.txt` has 8 columns:

Variable name	Description
Identification number	1-73
Completed	Number of websites completed and delivered to customers during the quarter
Backlog	Number of website orders in backlog
TeamID	1-13
Experience	Number of months team has been together
ProcessChange	Whether a change occurred in the website development process during the second quarter of 2002
Year	2001 or 2002
Quarter	1,2,3 or 4.

- (a) The company is interested in understanding the effect of four covariates on performance: (1) the change in the website development process, (2) the size of backlog orders, (3) the team effect, and (4) the number of months experience of each team. Develop a best subset model for predicting production output. Justify your choice of model.
- (b) This data was collected over a two-year period, and the performance of each team may be correlated through time. starting with the best subsets model, test whether there is positive autocorrelation. If autocorrelation is present, correct for it and adjust your model choice accordingly.

2. Suppose that X and Y are conditionally independent given Z , and that X and Z are marginally independent.
 - (a) Show that X is jointly independent of Y and Z .
 - (b) Show that X and Y are marginally independent.
 - (c) Show that if X and Z are conditionally (rather than marginally) independent, then X and Y are still marginally independent.

3. The following table refers to the effect of gender and race on political party identification.

Gender	Race	Party Identification		
		Democrat	Republican	Independent
Male	White	132	176	127
	Black	42	6	12
Female	White	172	129	130
	Black	56	4	15

- (a) Analyze this data using a multinomial logit model treating party identification as a response. How many different models are possible?
 - (b) Determine the best fitting model. State your model selection criterion.
 - (c) Is there a Poisson model which assumes the same type of independence as the model you found in (b)? If there is, show how the parameters in the logistic model relate to the parameters in the poisson model.
 - (d) Interpret the model you chose for part (a), and use it to quantify the effect of race and gender on party identification. Be specific, use log odds ratios, and give confidence intervals. Do you believe in these confidence intervals? Why or why not?

4. Serum prostate-specific antigen (PSA) is a well-established screening test for prostate cancer. Oncologists want to examine the correlation between level of PSA and a number of other clinical variables. Data (`ProstateCancer.txt`) were collected on 97 men who were about to undergo radical prostectomies. Use multiple linear regression to analyze this data set. Answer the following questions:
 - (a) Are any variable transformations necessary?
 - (b) Use a stepwise method to greedily search the model space. Do AIC, BIC, and Chi-square test return the same model?
 - (c) Use LASSO to explore the model space. With minimizing prediction error as the goal, do the models given by LASSO and greedy search agree?