Stat 601 Homework 6 Due 12/11/2015

1. This is a continuation of Problem 3 in Homework 5. Suppose that after appropriate variable transformations, you are going to

   (a) Develop a valid ridge regression model containing all seven potential predictor variables listed in Problem 3 in Homework 5. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots of standardized residuals, and any other relevant diagnostic plots;

   (b) Develop a valid principle component regression model containing all seven potential predictor variables listed in Problem 3 in Homework 5. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots of standardized residuals, and any other relevant diagnostic plots;

   (c) Develop a valid partial least squares regression model containing all seven potential predictor variables listed in Problem 3 in Homework 5. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots of standardized residuals, and any other relevant diagnostic plots;

   (d) Describe any weaknesses in each of the models in (a)-(c).

   (e) The golf fan wants to remove all predictors with insignificant t -values from the full model in a single step. Explain why you would not recommend this approach.

2. In this hypothetical problem, all variable names are masked, and you should simply treat the response variable as Y (the first column in the data matrix), and all covariates as $X_1,\ X_2, \ldots, X_8$ (the last 8 columns in the data matrix). The data is saved in HWK6q2.txt. Perform appropriate variable transformations and

   (a) Develop a valid ridge regression model containing all eight potential predictor variables. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots of standardized residuals, and any other relevant diagnostic plots;

   (b) Develop a valid principle component regression model containing all eight potential predictor variables. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots of standardized residuals, and any other relevant diagnostic plots;

   (c) Develop a valid partial least squares regression model containing all eight potential predictor variables. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots of standardized residuals, and any other relevant diagnostic plots;

   (d) Describe any weaknesses in each of the models in (a)-(c).

3. In an experiment, the number of *Ceriodaphnia* organisms are counted in a controlled environment in which reproduction is occurring among the organisms. The experimenter places into the containers a varying concentration of a particular component of jet fuel that impairs reproduction. Hence, it is expected that as the concentration of jet fuel grows, the mean number of counts should decrease. The experiment is done on two different strains and the results are available in the **ceriodaphnia.txt** file from the dataset. The columns indicate: number of organisms, the concentration of jet fuel in grams per liter and the strain of the organism. Build a generalized linear model (fit the models, provide diagnostic plots, tests etc) to study the effect of the fuel concentration and strain type on the number of Ceriodaphnia organisms.

```
> str(cerio <- with(read.table("ceriodaphnia.txt", sep=''')),
+ data.frame(numb=V1, conc=V2, strain=factor(V3))))
'data.frame': 70 obs. of 3 variables:
$ numb : int 82 58 106 58 63 62 99 58 101 73 ...
$ conc : num 0 0 0 0 0 0 0 0 0 0 ...
$ strain: Factor w/ 2 levels "1","2": 1 2 1 2 1 2 1 2 1 2 ...
```

4. The board of directors of a professional association conducted a random sample survey of 30 members to assess the effects of several possible amounts of dues increase. Two values are recorded: (1) The dollar increase in annual dues posited in the survey interview; (2) Whether or not the interviewee indicated that the membership will not be renewed at that amount of dues.

```
> str(renewal <- with(read.table("renewal.txt", sep=''')),
+ data.frame(renew=factor(V1, levels=1:0,
+ labels=c("N","Y")), amt=V2)))
'data.frame': 30 obs. of 2 variables:
$ renew: Factor w/ 2 levels "N","Y": 2 1 2 2 2 2 1 2 2 1 ...
$ amt : num 30 30 30 31 32 33 34 35 35 35 ...
```

Consider fitting a logistic regression model to these data treating renew as the response and amt as the covariate.

$$\Pr(renewjamt) = \text{logit}(\beta_0 + \beta_1(amt)).$$

(a) Find the maximum likelihood estimates of β_0 and β_1 of the logistic regression model.

(b) Create a scatter plot of the data with both the fitted logistic response from part (a) and a lowess smooth superimposed. You may want to use jittering to avoid overplotting of the data points. Does the fitted logistic response function appear to fit well?

(c) What is the interpretation of the $\hat{\beta}_1$?

(d) What is the estimated probability that association members will nor review their member- ship if the dues are increased by \$40?

(e) Estimate the amount of dues increase for which 75 percent of the members are expected not to renew their association membership?

(f) Obtain an approximate 90 percent cofidence interval for $\exp(\beta_1)$. Interpret your interval.

(g) Conduct a Wald test to determine whether dollar increase in dues is related to the proba- bility of membership renewal by using $\alpha = 0.1$. State the null and alternative hypotheses, and conclusion.

(h) Conduct a likelihood ratio test to determine whether dollar increase in dues is related to the probability of membership renewal by using $\alpha = 0.1$. State the full and reduced models, and conclusion. How does the result here compare to that obtained from the Wald test in part (b)?

5. In order to asses the appropriateness of large sample inferences for Problem 4, employ the following parametric bootstrap procedure.

   - For each of the 30 cases, generate a Bernoulli outcome (0, 1) using the estimated probability $\hat{\pi}_i$ for the original X_i level according to the fitted model.
   - Fit the logistic regression model to the bootstrap sample and obtain the bootstrap estimates $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$.
   - Repeat the above two steps 500 times. Compute the mean and standard deviation of the 500 bootstrap estimates $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$.

   (a) Plot separate histograms of the bootstrap distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$. Are these distributions approximately normal?

   (b) Compare the points estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and their estimated standard deviations obtained in the original fit to the means and standard deviations of the bootstrap distributions.

   (c) Comment on the appropriateness of large sample inferences in this dataset.

6. A health insurance company collected information on 788 of its subscribers who had made claims resulting from ischemic (coronary) heart disease. Data were obtained on totals costs of services provided for these 788 subscribers and the nature of the various services for the period of January 1, 1998 through December 31, 1999. Each line in the dataset has identification number and provides information on 9 other variables for each subscriber. The 10 variables are(in the order of columns that appear in the dataset):

   - Identification number: 1-788.
   - Total cost: Total cost of claims by subscriber (dollars).
   - Age: Age of subscriber (years).
   - Gender: Gender of subscriber: 1 = male, 0 = female.
   - Interventions: Total number of interventions or procedures carried out.
   - Drugs: Number of tracked drugs prescribed.
   - Emergency room visits: Number of emergency room visits.
   - Complications: Number of other complications that arose during heart disease treatment.

- Comorbidities: Number of other diseases that the subscriber had during period.
- Duration: Number of days of duration of treatment condition.

```
> Isch <- with(read.table("Ischemic.txt", sep="")),
+ data.frame(id=V1, cost=V2, age=V3,
+ gender=factor(V4, labels=c("F","M")),
+ int=V5, ndrug=V6, emerg=V7, compl=V8,
+ comorb=V9, duration=V10))
> str(Isch)
'data.frame': 788 obs. of 10 variables:
$ id : int 1 2 3 4 5 6 7 8 9 10 ...
$ cost : num 179 319 9311 281 18727 ...
$ age : int 63 59 62 60 55 66 64 45 68 64 ...
$ gender : Factor w/ 2 levels "F","M": 1 1 1 2 1 1 2 2 1 2 ...
$ int : int 2 2 17 9 5 1 2 3 6 3 ...
$ ndrug : int 1 0 0 0 2 0 0 0 2 0 ...
$ emerg : int 4 6 2 7 7 3 3 5 5 2 ...
$ compl : int 0 0 0 0 0 0 0 0 0 0 ...
$ comorb : int 3 0 5 2 0 4 1 1 4 0 ...
$ duration: int 300 120 353 332 18 296 247 82 334 85 ...
```

Consider modeling the number of emergency room visits as a function of other variables in the dataset

(a) Obtain the fitted Poisson model using all the variables. State the estimated coeficients, their estimated standard errors, and the estimated response function.

(b) Comment on the adequacy of the fit of the Poisson regression model based on the deviance residuals from the fit.

(c) Can any of the predictors be dropped from the model? Explain.

(d) Perform variable selection on this fit. State your final model.

7. Explicitly derive the maximum likelihood equations for fitting a logistic regression model. *Note: In the lecture, we derived these in their most generic form. In this question, you will fill in the gaps for logistic regression.*

8. Suppose the distribution of Y_i belongs to the following exponential family:

$$f(y_i, \theta_i) = \exp\{w_i(y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi)\}.$$

Define the moment generating function m_{Y_i} by

$$m_{Y_i}(s) = E(e^{sY_i}).$$

Show that

(a) $m_{Y_i}(0) = 1$,
(b) $m'_{Y_i}(0) = E(Y_i)$,
(c) $m''_{Y_i}(0) = E(Y_i^2)$,
(d) $\phi = \frac{w_i Var(Y_i)}{b''(\theta_i)}$.