THE AUSTRALIAN NATIONAL UNIVERSITY

RESEARCH SCHOOL OF FINANCE, ACTUARIAL STUDIES AND STATISTICS

STAT3008/STAT7001
APPLIED STATISTICS

Assignment 2

Lecturer: Dr Tao Zou

Last Updated: Sun Oct 8 21:43:34 2017

This assignment is due at 12:00 pm, Oct 18, 2017.

This assignment is worth 10% of your final grade but is optional and redeemable. Students are expected to complete this assignment **individually**. Maximum points: 10.0. You cannot get partially correct for all the questions, since each question is only worth 0.5 points. **Assignments can only be submitted via the physical assignment box at the front of the reception on Level 4, CBE Building (26C). Hard copy submission is required.** Late submission will not be accepted and the weight will roll over to your final exam. Identical submissions are treated as cheating.

Please **exactly follow the instructions of questions** and write down the answers of the following questions in the **answer sheet** file on the Wattle. Note that you do not need to copy the questions in the answer sheet. Please only submit your finished answer sheet and do not paste any unrelated results. The data used in this assignment are on the Wattle or in the R package "Sleuth3", whose instruction manual is on the Wattle.

The significance level for all the questions is set to be 0.05.

Question 1 (Variable Selection and Multicollinearity, 2.5 points)

Consider the data used in Quesitons 1 - 3 in Assignment 1. Please answer the following questions in the answer sheet.

a) (0.5 points) If we only consider the response variable and explanatory variables used in Quesiton 3 e) in Assignment 1 (please use indicator variables instead of the original categorical variables), please paste the Cp plot among all subsets in the answer sheet (showing at most 2 subsets for each size).

b) (0.5 points) Based on the above Cp statistics, which variables should we choose to predict the logarithm of "WeeklyEarnings" by using the variable selection among all subsets?

c) (0.5 points) If we still consider the response variable and explanatory variables used in Quesiton 3 e) in Assignment 1 (please use indicator variables instead of the original categorical variables), please use R to obtain the variance inflation factors (VIF) for each of the explanatory variables, and paste them in the answer sheet. Based on the "rule of thumb" cut-off for VIF, does the multicollinearity problem exist if we regress the response on the explanatory variables?

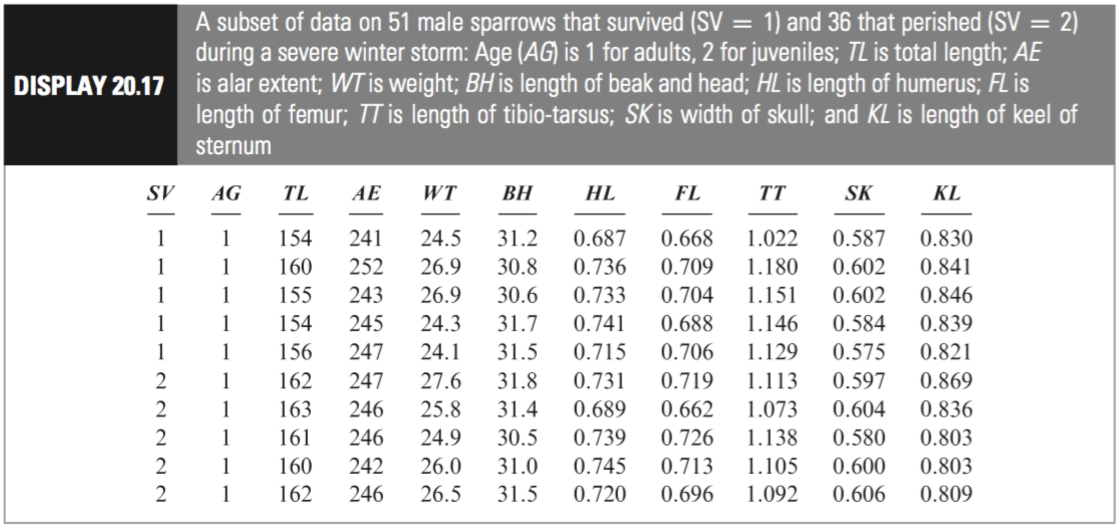
d) (0.5 points) Please use the backward elimination idea to solve this multicollinearity problem and report the variables we should use in the regression model such that there is no multicollinearity problem. (Hint: similar to the codes on page 56 of Lecture Notes 8.)

e) (0.5 points) Please paste the R codes for all the above analyses of Question 1 in the answer sheet.

Question 2 (Binary Logistic Regression, 2.0 points)

Bumpus Natural Selection Data (Revised based on ex 16 of Chapter 20 in "The Statistical Sleuth"). Hermon Bumpus analyzed various characteristics of some house sparrows that were found on the ground after a severe winter storm in 1898. Some of the sparrows survived and some perished. The data on male sparrows in Display 20.17 are survival status (1 = survived, 2 = perished), age (1 = adult, 2 = juvenile), the length from tip of beak to tip of tail (in mm), the alar extent (length from tip to tip of the extended wings, in mm), the weight in grams, the length of the head in mm, the length of the humerus (arm bone, in inches), the length of the femur (thigh bones, in inches), the length of the tibio-tarsus (leg bone, in inches), the breadth of the skull in inches, and the length of the sternum in inches. The dataset is stored in the object "ex2016" of the R library "Sleuth3".

Analyze the data to see whether the probability of survival is associated with physical characteristics of the birds. This would be consistent, according to Bumpus, with the theory of natural selection: those that survived did so because of some superior physical traits. Realize that the sampling is from a population of grounded sparrows.



DISPLAY 20.17 A subset of data on 51 male sparrows that survived (SV = 1) and 36 that perished (SV = 2) during a severe winter storm: Age (*AG*) is 1 for adults, 2 for juveniles; *TL* is total length; *AE* is alar extent; *WT* is weight; *BH* is length of beak and head; *HL* is length of humerus; *FL* is length of femur; *TT* is length of tibio-tarsus; *SK* is width of skull; and *KL* is length of keel of sternum

SV	*AG*	*TL*	*AE*	*WT*	*BH*	*HL*	*FL*	*TT*	*SK*	*KL*
1	1	154	241	24.5	31.2	0.687	0.668	1.022	0.587	0.830
1	1	160	252	26.9	30.8	0.736	0.709	1.180	0.602	0.841
1	1	155	243	26.9	30.6	0.733	0.704	1.151	0.602	0.846
1	1	154	245	24.3	31.7	0.741	0.688	1.146	0.584	0.839
1	1	156	247	24.1	31.5	0.715	0.706	1.129	0.575	0.821
2	1	162	247	27.6	31.8	0.731	0.719	1.113	0.597	0.869
2	1	163	246	25.8	31.4	0.689	0.662	1.073	0.604	0.836
2	1	161	246	24.9	30.5	0.739	0.726	1.138	0.580	0.803
2	1	160	242	26.0	31.0	0.745	0.713	1.105	0.600	0.803
2	1	162	246	26.5	31.5	0.720	0.696	1.092	0.606	0.809

Display taken from class text: "The Statistical Sleuth".

In order to investigate the above problem, please use R to answer the following questions in the answer sheet.

a) (0.5 points) Please use the R function “factor()” to transform the vector “AG” in “ex2016” into a vector of factor values in the data frame. Then use R to regress survival status on all the other variables, including all the interactions between “AG” and the other continuous explanatory variables, in order to answer the question whether the probability of survival is associated with physical characteristics of the birds. Please do not use the indicator variable of “AG”, but instead, use the factor values of “AG” directly in the fitting of the regression. (Hint: Similar to page 12 of Lecture Notes 11). Based on the “summary” function output of this fitted model, can we use the “Null deviance” and “Residual deviance” in the output to construct a drop-in-deviance χ^2-test? If we can, what are the null hypothesis and the alternative hypothesis of this test?

b) (0.5 points) If we can construct a drop-in-deviance χ^2-test in the above question, please use R to accomplish this χ^2-test. What is the value of the test statistic? What conclusion can you obtain for this χ^2-test? If we cannot construct a drop-in-deviance χ^2-test in the above question, please state your reasons.

c) (0.5 points) Consider all the variables involved in Question 2 a) and use R to perform the forward selection based on BIC. Which variables should we choose to predict the probability of survival by using this variable selection method?

d) (0.5 points) Please paste the R codes for all the above analyses of Question 2 in the answer sheet.

Question 3 (Multicategory Response Regression, 4.0 points)

(Revised based on ex 16 of Chapter 3 in "Analysis of Categorical Data with R".)

Researchers at Penn State University performed a study to determine the optimal fat content for ice cream. Details of the study and corresponding data analysis are available at https://onlinecourses.science.psu.edu/stat504/node/187. In summary, 496 individuals were asked to taste and then rate a particular type of ice cream on a 9-point scale (1 to 9 with 1 equating to not liking and 9 equating to really liking). The ice cream given to the individuals had a fat proportion level of 0, 0.04, . . . , or 0.28 (fat). We treat "fat" as continuous variable in this question.

The data for 496 individuals is randomly split into two parts. The training dataset (including 471 individuals) and the test dataset (including 25 individulas) are available in "ice_cream1.csv" and "ice_cream2.csv" on the Wattle, respectively. Using these data, please use R to answer the following questions in the answer sheet.

a) (0.5 points) Obviously, the ratings 1-9 are ordinal data. Please use the ordinal response regression model to regress the ordinal response "rating" on "fat" by utilizing the training dataset only. How many unknown regression parameters in the above ordinal response regression model? (Hint: similar to Question 2, please use the R function "factor()" to transform the vector "rating" in the data into a vector of factor values in the data frame first.)

b) (0.5 points) What is the 95% confidence interval for the coefficient of "fat" (rounded to four decimal places) based on the above fitted ordinal response regression model?

c) (0.5 points) If we are interested in testing whether or not "fat" is needed based on the above fitted ordinal response regression model, please construct an appropriate test. What is the p-value for your test (rounded to four decimal places)? What conclusion can you obtain based on the p-value?

d) (0.5 points) Suppose we ignore the order of the ratings 1-9 in this part and treat "rating" as a nominal response. Please use the nominal response regression model to regress the nominal response "rating" on "fat"" by utilizing the training dataset only. How many unknown regression parameters in the above nominal response regression model?

e) (0.5 points) If we are interested in testing whether or not "fat" is needed based on the above fitted nominal response regression model, please construct an appropriate test. What is the p-value for your test (rounded to four decimal places)? What conclusion can you obtain based on the p-value?

f) (0.5 points) Now we consider the test dataset. Let Y_ℓ be the "rating" for observation $\ell = 1, \cdots, n_{\text{test}}$ in the test dataset. Suppose $\hat{Y}_\ell$ to be the corresponding prediction of response based on either of the above two fitted models from the training dataset. Define an indicator variable

$$I\{Y_\ell = \hat{Y}_\ell\} = \begin{cases} 1, & \text{if } Y_\ell = \hat{Y}_\ell; \\ 0, & \text{otherwise.} \end{cases}$$

Then the percentage of correct forecast (PCF) can be defined by

$$\text{PCF} = \frac{1}{n_{\text{test}}} \sum_{\ell=1}^{n_{\text{test}}} I\{Y_\ell = \hat{Y}_\ell\}.$$

Based on the definition, what is the PCF for the ordinal response regression model?

g) (0.5 points) What is the PCF for the nominal response regression model? Based on these two PCFs, which model is better?

h) (0.5 points) Please paste the R codes for all the above analyses of Question 3 in the answer sheet.

Question 4 (Simulation for Binary Logistic Regression, 1.5 points)

Consider the binary logistic regression model $\text{logit}(\mu\{Y|X\}) = \beta_0 + \beta_1 X$ for the observations $\{Y_i, X_i\}_{i=1}^n$, and the maximum likelihood estimation (MLE) $\hat{\beta}_0$ and $\hat{\beta}_1$ for the coefficients β_0 and β_1 can be obtained.

Lily wants to use R to generate random samples based on the binary logistic regression model assumptions, in order to understand the "roughly" unbiased property for MLE, as well as "approximate" normality for the sampling distribution of MLE. She follows the steps below.

STEP 1: Specify $\beta_0 = 2$ and $\beta_1 = 1$.

STEP 2: Suppose the observations $X_1, \cdots, X_n$ are $0.001, 0.002, 0.003, \cdots, 1.000$, so the number of observations $n = 1000$.

STEP 3: Compute

$$\pi_i = \frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}} \text{ for } i = 1, \cdots, n.$$

STEP 4: Generate Y_i independently from the Bernoulli distribution with $\text{P}(Y_i = 1|X_i) = \pi_i$ for $i = 1, \cdots, n$. (Hint: R function "rbinom(1,1,π_i)" returns one random number of Y_i from the Bernoulli distribution with $\text{P}(Y_i = 1|X_i) = \pi_i$.)

STEP 5: Repeat Step 4 1,000 times and obtain 1,000 different datasets of $\{Y_i, X_i\}_{i=1}^n$.

Lei Li is a friend of Lily. Lily hands over the above 1,000 datasets to him but she does not tell him the true values of β_0 and β_1. Based on each dataset, Lei Li computes the MLE $\hat{\beta}_0$ and $\hat{\beta}_1$. Ultimately, he obtains 1,000 different MLEs.

Then Lily tells Lei Li the true value of β_1 and both Lily and Lei Li compare this true value to the sample average of these 1,000 different MLEs $\hat{\beta}_1$, as well as the histogram of these 1,000 estimates $\hat{\beta}_1$.

Please answer the following questions in the answer sheet.

a) (0.5 points) Suppose you play both roles of Lily and Lei Li and realise the above steps in R. Please paste the complete R codes for all the above procedures in the answer sheet. (Hint: similar to the codes on page 7 of Lecture Notes 2.)

b) (0.5 points) What is the sample average of 1,000 estimates $\hat{\beta}_1$ (rounded to four decimal places). Is it close to the true value of β_1? Please answer this question in the answer sheet.

c) (0.5 points) Please paste the histogram plot of 1,000 estimates $\hat{\beta}_1$ in the answer sheet. Is it close to the normal distribution?