Fall 2013 - STAT542 - TEST #1 | Due Friday, October 18
October 11, 2013

Directions: There are 12 total questions on this exam. Each question is worth 8 points, except for problem 3(a) which is worth 12 points. You may use your textbook, class notes, and relevant software on this test. You may not consult other textbooks and may not discuss the problems with anyone except me. Please read all problems carefully, and when asked for explanations, respond in complete sentences. If you do not understand a question, come ask me. Good luck!

1. The rise in abundance of algae in coastal waters is thought to be due to increases in nutrients such as nitrates and other forms of nitrogen. It is theorized that the excessive amounts of nitrate are due to human influences. Human populations can affect nitrogen inputs to rivers through industrial and automobile emissions to the atmosphere (causing the nitrogen to enter the river through rainfall), through fertilizer runoff, through sewage discharge, and through watershed disturbance. Researchers gathered data from 42 rivers around the world to gauge the evidence that nitrates in the discharges of rivers around the world are associated with human population density. The data are on the course webpage under the filename **nitriver.txt**. Among the variables measured were:

   y = nitrate concentration (in μM/l)
   x_1 = discharge: the estimated annual average discharge of the river into the ocean (l/(sec $times$ k $\times m^2$))
   x_2 = runoff: the estimated annual average runoff from the watershed (in l/(sec $times$ k $\times m^2$))
   x_3 = precipitation (in cm/year)
   x_4 = area of watershed (in km^2)
   x_5 = human population density (in people/km^2)
   x_6 = deposition: the product of precipitation times nitrate concentration
   x_7 = nitrate precipitation: the concentration of nitrate in wet precipitation at sites located near the watersheds (in μmol NO_3/(sec $\times$ km^2))
   x_8 = nitrate export: the product of runoff times nitrate concentration.

   (a) Perform some exploratory data analysis on these variables and write a brief (no more than 2 paragraphs) summary of your findings. Since nitrate concentration is the response variable, your analysis should focus on the relationship between nitrate concentration and the other variables, as well as any relationships among the explanatory variables. Also, include in your discussion what problems you might expect to encounter in trying to determine a "best" model according to some selection criterion. [Do NOT actually

look for a "best" model - I am just looking for what problem(s) you might anticipate in the process of finding such a model based on your EDA.]

(b) Suppose after examining numerous models, you decide that the general linear model with explanatory variables (x_2, x_6, x_7, x_8) is the best model. Using this "best" model, give a set of Bonferroni joint confidence intervals for the model parameters. Explain in a sentence or two why it is important to use a Bonferroni correction here.

(c) In the "best" model of part (b), perform a permutation test of $H_0 : \beta_7 = 0$ where β_7 is the parameter corresponding to x_7. Report the permutation p-value and a clear conclusion to the test. Does this agree with the normal-based p-value reported in the **Coefficients** table?

2. As part of a study of the effects of predatory intertidal crab species on snail populations, researchers measured the mean closing forces and the propodus heights of the claws on several crabs of three species. Specifically, there were three crab species in the study (*Hemigrapsus nudus, Lophopanopeus bellus, Cancer productus*) with between 12-14 crabs of each type on which measurements were taken. The data are given below and can be found on the course webpage under the filename **crab.txt**:

H. nudus ($n = 14$)		*L. bellus* ($n = 12$)		*C. productus* ($n = 12$)	
Force	**Height**	**Force**	**Height**	**Force**	**Height**
3.2	5.0	2.1	5.1	5.0	6.7
6.4	6.0	8.7	5.9	7.8	7.1
2.0	6.4	2.9	6.6	14.6	11.2
2.0	6.5	6.9	7.2	16.8	11.4
4.9	6.6	8.7	8.6	17.7	9.4
3.0	7.0	15.1	7.9	19.8	10.7
2.9	7.9	14.6	8.1	19.6	13.1
9.5	7.9	17.6	9.6	22.5	9.4
4.0	8.0	20.6	10.2	23.6	11.6
3.4	8.2	19.6	10.5	24.4	10.2
7.4	8.3	27.4	8.2	26.0	12.5
2.4	8.8	29.4	11.0	29.4	11.8
4.0	12.1				
5.2	12.2				

Let:

$$
\begin{aligned}
y &= \text{the log mean closing force of a crab,} \\
x_1 &= \text{the log mean propodus height of a crab claw,} \\
x_2 &= \left\{ \begin{array}{ll} 1 & \text{if the crab species is } \textit{H. nudus} \\ 0 & \text{otherwise} \end{array} \right\}, \\
x_3 &= \left\{ \begin{array}{ll} 1 & \text{if the crab species is } \textit{L. bellus} \\ 0 & \text{otherwise} \end{array} \right\},
\end{aligned}
$$

(a) Consider the multiple linear regression model given by:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i} x_{2i} + \beta_5 x_{1i} x_{3i} + \epsilon_i.$$

Interpret the parameters β_1, β_2, & β_4 clearly in the context of the problem. For example, do not just say: "β_1 is the partial slope of y on x_1." Explain its meaning in terms of the mean response y.

(b) Construct a scatterplot of log force vs. log height, with a different symbol for each crab species. In a few sentences, describe the associations between the three variables (log force, log height, crab species).

(c) Fit the multiple linear regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i} x_{2i} + \beta_5 x_{1i} x_{3i} + \epsilon_i$. Report the **ANOVA** and **Coefficients** table. Test simultaneously whether or not either of the interactions are significant [As always, give the hypotheses, test statistic, p-value, and a clear conclusion based on this p-value.] Do the results of this test confirm or refute what is seen in the scatterplot? Explain.

(d) Predict the mean closing force using a 95% prediction interval for a crab of species *L. bellus* with a mean propodus height of 11.5. Would you expect the prediction interval to be wider or narrower if the height had been 8? Explain in a sentence or two.

(e) Still considering the full model, test for a difference in the slopes between log force and log height for the *H. nudus* and *C. productus* crab species. That is, test whether or not the slopes resulting from a regression of log force on log height are different for the two crab species. [Again, state your hypotheses clearly, give the test statistic, give the p-value, and state your conclusion based on this p-value.]

(f) Another way we might test for differences in the slopes of the regression lines between log force and log height for the three crab species is to run separate regressions for the three species, and compare the slopes using 3 independent t-tests. Discuss in a short paragraph which of these two methods (3 independent t-tests, multiple linear regression) would be better and why? Think about degrees of freedom, and the mean squared error. [Do NOT perform the tests to answer this question.]

3. A study was undertaken in major US cities to examine the effects of pollution levels on mortality, adjusting for climate and socioeconomic information. Specifically, data were collected at each city on the following variables:

y = mortality rate (deaths per 100,000 people over a 3-year period),

x_1 = **precip** = mean annual precipitation (inches),

x_2 = **education** = the mean number of school years completed, for persons of age 25 years or older

x_3 = **nonwhite** = the percentage of the population that is non-white,

x_4 = **NOx** = the relative pollution potential of oxides and nitrogen,

x_5 = **SO2** = the relative pollution potential of sulfur dioxide.

The primary goal of the study was to investigate whether or not mortality is associated with either of the two pollution variables after accounting for the effects of the climate and socioeconomic variables on mortality. Suppose a sequence of models was fit with this in mind, resulting in the output tables given below.

Coefficients Table

Predictor	$\widehat{\beta}_k$	SE($\widehat{\beta}_k$)	t	p-val
Intercept	1000.1021	92.3981		3.85e-15
x_1	1.3792	0.7000		0.053943
x_2	-15.0790	7.0706		0.037519
x_3	3.1602		5.026	5.84e-06
x_4	-0.1076	0.1359		0.432066
x_5		0.0914		

Analysis of Variance Table

Source	df	SS	MS	F	pval
Regression					0.000
Error		75385			
Total					

Sequential Sums of Squares

Source	df	Seq. SS
x_1	1	59256
x_2	1	20492
x_3	1	51163
x_4	1	867
x_5	1	21110

(a) The R^2-statistic for the full model was $R^2 = 0.6698$ and the residual standard error was 37.36. Using this information and the information given in the tables above, fill in all of the missing information in the **Coefficients** and **ANOVA** tables. If you cannot figure out what some value is, just make up a value, let me know what value you made up, and complete the problem with your values.

(b) The researcher's goal was to test the significance of the two pollution variables simultaneously after accounting for the effects of the three climate and socioeconomic variables (x_1, x_2, x_3). Conduct such a test using the information in these tables. State your hypotheses clearly, give the test statistic and p-value, and state a conclusion in context of the problem.

(c) Test the model with only x_1 against the model with variables (x_1, x_2, x_3). Again, state the hypotheses, give the test statistic & p-value, and give a clear conclusion in context of the problem.