Wroclaw University, Fall 2015
Applied Statistical Methods
Exam1

As we discussed, you will work in three groups. Please type the names of all members of your group on the top of the first page of the solutions to this exam.

In all problems, you need to follow steps if any are given, and answer all questions asked. Please be very concise and precise in what you write. Please do not include the same information twice- be gentle for the trees! **The total number of pages that will be graded will not exceed 5 pages.** Do all tests of significance on the significance level of 0.05, unless specified otherwise. Each part of every problem is worth 3 points. For all testing hypotheses problems, you must write down the hypotheses you test, the test statistic, the critical number or p-value, the decision you make and the answer to the problem. Write solutions of the problems in the order they were asked. Please be at your best handwriting or type. Remember, what I can not read, I can not grade.
The exam is due on Tuesday, November 3 at class time (6pm). No late work will be accepted.

ENJOY!

Problem 1. Use the data set reg_hwk3.*. It is in the same folder as this exam. Your answer must include printout of the regression models and other relevant information you get from MINITAB (no repetitions, please!).

1. Write the equation of the regression model for y as a response to three predictors: x1, x2, and x3.
2. Include diagnostic plots. Can we assume that the residuals have a normal distribution with constant variance? Why/why not?
3. For all measures of leverage, outliers and influence (standardized and deleted-t residuals, h-leverages, Cooks' D and DFFITS), find the "critical" values of those measures that separate OK values from the high ones. Use the table format below.

Measure	**Critical number**
Standardized residuals	
deleted t-residuals, use $t_{0.05}$	
leverages h_i	
Cook's distance	
DFITTS, use $F_{0.1}$	

4. Is there an influential observation? If YES, which one and why do you think it is influential?
5. Remove the influential observation and run regression again. The result will be your second model.
6. For the smaller data set find the "critical" values of all measures of leverage, outliers and influence.
7. Are there any influential observations?
8. Is the second model a reasonable regression model? Why or why not?
9. If you decided that the second model is reasonable, you are done with this problem. If you decided that you can develop a better model, please do so and argue why do you think your new model is better.

Problem 2. Generate (make it up any way you want) values for a response y and three explanatory variables (x1, x2, and x3) so that all of the following conditions are satisfied:

a) There are 10 observations;
b) x1 is not a significant predictor for y;
c) x2 and x3 are significant predictors for y;
d) 10^{th} observation is an influential one.

1. Describe how you constructed your data set. Explain how did you get each variable. Please be very concise and precise. Do not print the data set.
2. Show that x1 is not a significant predictor for y. That is perform/refer to a partial F (or t) test for the appropriate hypothesis. Show results of MINITAB computation for this question.
3. Show that x2 and x3 are significant predictors for y. That is perform/refer to an appropriate F test. Show results of MINITAB work for this question.
4. Show that 10^{th} observation is influential. That is compute appropriate statistics, show their values and explain why the observation is indeed influential.

If you can not generate a data set with all of the above properties, generate one with 10 obs and at least one of properties b) – d). This will give you partial credit and a data set to work on for some of the questions 1-4.

Problem 3. For this problem you will use the data set called HEIGHTS.*. The data are heights of 20 boys and 20 girls along with the heights of both parents. All heights are in inches. The data are from the U.S. Department of Health and Human Services, National Center for Health Statistics, Third National Health and Nutrition Examination Survey.

1. Develop a reasonable regression model for the children's heights (both genders together) using as few as possible predictors.
2. Justify your choice of the model.
3. Is the model reasonable from the prediction point of view? How about inference point of view?
4. Develop regression models for boys and girls' heights using predictor(s) in the data set and a dummy variable "Gender".
5. Are the regression models for heights different for the two genders? Discuss possible differences in slope and intercepts of the regression equations for boys and girls. Justify your answer. Include your best regression models by gender.
6. Are the models reasonable from the point of view of fit/prediction and inference?

Problem 4. Using the normal error linear regression model, in an engineering safety experiment, a researcher found for the first 10 cases that R^2 was zero.
(a) Is it possible that for the complete set of 30 cases R^2 will not be zero?
(b) Could R^2 not be zero for the first 10 cases, yet equal to zero for all 30 cases?
Explain. If you believe the answer to a question is "NO", justify why you think so. If you believe the answer is "YES", construct a data set with the needed properties and show in MINITAB that the properties hold.

Problem 5. Several measurements were recorded in a data set BEARS.* on 50 bears. Prediction of a bear's head width using other measurements was of interest. One model predicted a bear's head's width (HEADWTH) using its neck circumference (NECK), head length (HEADLEN) and (AGE). Another model used only one explanatory variable: NECK. Perform an appropriate test to decide if the more complex model is significantly better than the simpler one. State your (a) models, (b) hypotheses, (c) test statistic, (d) p-value, and (e) decision.