STAT 350 Regression Analysis Final Exam/Project Fall 2022

Due by 11:59pm, Monday December 5, 2022

You must work **independently** (violation of academic integrity will be reported according to Northwestern University regulations). Only the instructor may be asked to clarify questions.

1. The final project (total 100 points) consists two parts:
    a. Part 1 (20 points) has four short answer questions which do not involve data analysis using R programming or a calculator
    b. Part 2 (80 points) is the project. You will perform data analysis and draw conclusions using the provided dataset.

2. Submission requirement:
    a. Your solution should **be one file** (either .docx or .pdf should be fine)
    b. The exam is between 12pm and 11:59pm. Please submit by 11:59pm, December 5,2022.
    c. Arrange your answers (including **graphs, tables, and needed software outputs**) in the order of the problems.
    d. If you use R markdown, please show the R code; if you do not use R markdown, please attach the R code at the end of your solution.

Note: Please check your grades on homework and midterm exam on Canvas. If there is any problem, please let me know by email by Friday December 9, 2022. Thanks.

Part 1: Short questions (20 points total; 5 points each)

1) A researcher computed the pairwise correlations among the predictor variables in his dataset and found out that they are all small. Then he concluded that the predictors do not interact with each other. Is he right? Why?

2) Consider the regression of tool wear (Y) on tool speed (X_1) and tool model which has four classes (M1, M2, M3, M4). Suppose the indicator variables had been defined as follows: $X_2 = 1$ if tool model M2 and 0 otherwise; $X_3 = 1$ if tool model M3 and 0 otherwise; $X_4 = 1$ if tool model M4 and 0 otherwise. Now we fit the following first-order regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon.$$

Please specify the null and alternative hypotheses for testing if the response function is the same for all four tool models.

3) (Continuation of part 2). Please specify the null and alternative hypotheses for testing if the expected tool wear is the same for tool model M3 and tool model M4 while holding tool speed (X_1) fixed.

4) A researcher decides to develop a regression model to predict Y using three predictors (X_1, X_2, and X_3). To find out what is the appropriate form for X_1, she performed two regression analyses. In the first analysis shed used Y as the response variable, X_2 and X_3 as predictors, and computed the residuals (Residual 1). For the second analysis, shed used X_1 as the response variable, X_2 and X_3 as predictors, and computed the residuals (Residual 2). Then she plotted Residual 1 versus Residual 2 and found a linear pattern. Therefore, she decided to include the first order term of X_1 in the regression function. Was her plan sound reasonable? Why?

Part 2 (80 points; 10 points each) A tax assessor was interested in predicting the resident home sales prices. She has collected data on 522 home sales in a mid-city during the year 2003. Each line of the data set has an identification number (Variable 1) and information on sales price (Variable 2) and various characteristics of the home and surrounding property (Variables 3 to 13). The variables are listed in the following table.

Variable Number	Variable Name	Description
1	Identification number	1–522
2	Sales price	Sales price of residence (dollars)
3	Finished square feet	Finished area of residence (square feet)
4	Number of bedrooms	Total number of bedrooms in residence
5	Number of bathrooms	Total number of bathrooms in residence
6	Air conditioning	Presence or absence of air conditioning: 1 if yes; 0 otherwise
7	Garage size	Number of cars that garage will hold
8	Pool	Presence or absence of swimming pool: 1 if yes; 0 otherwise
9	Year built	Year property was originally constructed
10	Quality	Index for quality of construction: 1 indicates high quality; 2 indicates medium quality; 3 indicates low quality
11	Style	Qualitative indicator of architectural style
12	Lot size	Lot size (square feet)
13	Adjacent to highway	Presence or absence of adjacency to highway: 1 if yes; 0 otherwise

The following shows what the data look like for two home sales. You can find the complete dataset from "HomeSales.txt". Please note that variable 1, "Identification Number", should not be considered as predictor.

1	2	3	4	5	6	7	8	9	10	11	12	13
1	360000	3032	4	4	1	2	0	1972	2	1	22221	0
2	340000	2058	4	2	1	2	0	1976	2	1	22912	0

For the initial analysis of parts 1) to 6), she only considers **two variables, Finished square feet (variable 3) and Presence or absence of swimming pool (variable 8).**

1) She suspects that there is an interaction effect between these two variables, therefore will fit the regression function to predict the selling price by considering the potential interaction effect. Please write out the **statistical model** for this purpose. Please write out the estimated regression function and interpret the association between selling price and finished square feet for houses with and without swimming pool respectively.

2) For the model you used in part 1), test whether there is a regression relation between sales price and these two predictor variables (i.e., whether the model is useful in predicant the selling price) using α=0.05. State the null and alternative hypotheses and conclusion. What is the P-value of the test?

3) Is there an interaction effect between Finished square feet (variable 3) and Pool (variable 8)? Please write out the null and alternative hypothesis and draw your decision using significance level 0.05.

4) For a house which has a swimming pool and has 2061 finished square feet, please predict this house's selling price using your model and give the 95% prediction interval. Does the prediction involve extrapolation? Why?

5) Are there any houses in the data set which have (i) unusually high or (ii) unusually low sales price? Why? If the answer is yes, please point out which houses.

6) Which observation has the largest Cook's distance? Is it an influential data point? Why?

For the following analysis, i.e., parts 7) to 8), she will consider **all the variables, i.e., variables 3 to 13 to predict the sales price.**

7) Please use stepwise selection method to select the variables which are useful in predicting the sales price. Which variables will be considered as qualitative variables? If you consider a variable, say variableX, as a qualitative variable, you can use as.factor(variableX) instead of variableX in R lm model statement. Please clearly state which variables are selected.

8) For the variables you selected in part 7), please fit the first order regression model and plot the residuals against the predicted values. Also prepare a normal probability plot. Interpret your plots and summarize your findings. Please discuss what would you do next?