STA674: Regression Analysis and Design of Experiments

Assignment #2

Submission:

You must format your assignments as a pdf. Handwritten assignments will not be accepted.

When are ready to submit your assignment, copy your R (or RStudio work) or SAS code and paste it at the end of your document. You may also use RMarkdown. *Don't forget to add comments to help the grader follow your work.* Collaboration during the process of solving the problems is not only allowed but encouraged; that said, the submissions are each expected to be an individual effort reflecting the individual's work. Identical submissions or even submissions found to be **"too close to be coincidental" will be flagged and given no credit** (1 warning, then enforcement.)

Each problem is worth 4 points, for 20 points total. Longer problem parts are worth more. Please submit it to your instructor **by the due date on Canvas and the syllabus** via electronic submission on Canvas.

Questions

1. Apex Corporation, a firm that produces corrugated paper for use in making boxes and other packing materials, has called in consulting help to improve its cost-control program. The consultant is analyzing manufacturing costs to understand more fully the important influences on these costs. She has assembled monthly data on a group of variables, and she is using regression analysis to help assess how these variables are related to total manufacturing cost. The variables she has selected to study, the data for which are contained in the file (on Canvas) MFGCOST, are

    y total manufacturing cost per month in thousands of dollars (COST)
    x_1 total production of paper per month in tons (PAPER)
    x_2 total machine hours used per month (MACHINE)
    x_3 total variable overhead costs per month in thousands of dollars (OVERHEAD)
    x_4 total direct labor hours used each month (LABOR)

    a. What is the equation that is determined using all four explanatory variables?

    b. In the cost accounting literature, the sample regression coefficient corresponding to x_k is regarded as an estimate of the *true marginal cost* of output associated with the variable x_k. Find a point estimate of the true marginal cost associated with total machine hours per month. Also, find a 95% confidence interval estimate of the true marginal cost associated with total machine hours.

    c. What percentage of the variation in y has been explained by the regression?

2. The relationship between exchange rates, prices, and agricultural exports is of interest to agricultural economists. One such export of interest is wheat. The file named WHEAT on the Canvas site contains data on the following variables:

   y U.S. wheat export shipments (SHIPMENT)
   x_1 the real index of weighted-average exchange rates of the U.S. dollar (EXCHRATE)
   x_2 the per-bushel real price of no. 1 red winter wheat (PRICE)

   The dependent variable is U.S. wheat export shipments. The explanatory variables are exchange rate and price.

   a. What is the estimated regression equation relating SHIPMENT to EXCHRATE and PRICE?

   b. Test the overall fit of the regression. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. Use a 5% level of significance. What conclusion can be drawn from the result of the test?

   c. After taking account of the effect of PRICE, are SHIPMENT and EXCHRATE related? Conduct a hypothesis test to answer this question and use a 5% level of significance. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. What conclusion can be drawn from the result of the test?

3. A company is interested in the relationship between profit (PROFIT, in $1000) on a number of projects and two explanatory variables. These variables are the expenditure on research and development for the project (RD, in $100) and a measure of risk assigned at the outset of the project (RISK). The file RDS on the Canvas site has these data, reproduced below:

RD	RISK	PROFIT
132.580	8.5	396
81.928	7.5	130
145.992	10.0	508
90.020	8.0	172
114.408	7.0	256
53.704	7.5	32
76.244	7.0	102
71.680	8.0	102
151.592	9.5	536
74.816	7.5	102
108.752	6.0	214
92.372	8.5	200
92.260	7.0	158
60.732	6.5	32
78.120	7.5	116
90.000	5.5	120
105.532	9.0	270
111.832	8.0	270

a. Plot PROFIT versus RISK and RD and perform the regression of PROFIT on these two explanatory variables. Interpret the results.

b. Add a variable to your dataset—the square of RD (call it RDSQR). Perform a regression of PROFIT on RISK, RD, and RDSQR. Congratulations! You've just done a polynomial regression. What about your results in a. suggested that this was appropriate?

c. Compare the outputs of a. and b. Check 3 things specifically: the overall F-test values, the R-squared values, and the residual plots. Comment on these.

4. Data for the following variables for 93 employees of a Chicago bank are available in a file named BANK on the Canvas site.

y	beginning salaries in dollars (SALARY)
x_1	years of schooling at the time of hire (EDUCAT)
x_2	number of months of previous work experience (EXPER)
x_3	number of months after January 1, 1969, that the individual was hired (MONTHS)
x_4	indicator variable coded 1 for males and O for females (MALES)

Investigators were concerned with whether there was evidence of discrimination against females in salaries; we will now be taking into account two other variables besides education.

a. Perform the regression of y on all four explanatory variables. Use the results to conduct the F-test for the overall fit of the regression. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. Use a 5% level of significance. What conclusion can be drawn?

b. Is there a difference in salaries, on average, for male and female workers after accounting for the effects of the three other explanatory variables? Use a 5% level of significance to answer this question. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. Is there evidence that Harris Bank discriminated against female employees?

c. What salary would you forecast, on average, for males with 12 years education, 10 years of experience, and with time hired equal to 15? A point forecast is sufficient. What salary would you forecast, on average, for females if all other factors are equal?

5. Using the same data as 4, suppose that legal counsel representing the bank suggests that an interaction exists between education and experience and that the introduction of this term into the regression may account for the difference in average salaries. Create the interaction term

$$EDUCEXPR = EDUCAT * EXPER$$

and add it to the data set.

a. What is the adjusted R^2 for this regression? Compare this value to the adjusted R^2 for the regression without the interaction variable (done in #4.) Which model appears to be the best choice based on the adjusted R^2?

b. Test to see whether the interaction term is important in this regression model. Use a 5% level of significance. State the hypotheses to be tested, the decision rule, the test statistic, and your decision.

c. Does the interaction variable seem to be important in explaining SALARY? How would you respond to the suggestion that introduction of the interaction term into the regression may account for the difference in average salaries?