

ST 320

Lab assignment 8

Question 1

For the following problem. JMP will not be used. All the calculation should be carried out by hand or using the calculator. We have the following data:

X	-1.0	1.5	1.6	2.7	2.8	3.4	4.4	4.7	5.7	6	1.5
Y	1.9	1.5	3.0	4.2	3.9	5.4	6.2	7.5	7.8	9.0	-6.0

The aim is to see if there is a linear relation between y and x . We suppose that the regression equation is expressed as the following:

$$Y = \beta_0 + \beta_1 X + e$$

We know that in this case:

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

$$\beta_0 = \bar{Y} - b_1 \bar{X}$$

- 1-What is the estimated regression equation relating Y to X ?
- 2-Make a scatterplot of Y versus X ?
- 3-Define the terms outlier and Leverage points.
- 4-Use the scatterplot to identify visually the one point that is the most likely to be an outlier and the one point that is the most likely to be a leverage point.
- 5-Is X important? Describe all the step of your test statistics and your decision.

$$\text{Use the expression : } se(\beta_j) = \frac{s_Y}{s_X \sqrt{n-1}}$$

- 6-Can we predict $X^* = 9.5$. Calculate the prediction interval of $X^* = 9.5$

Question 2

Data used here are: MLS

In this problem, data are expressed by 7 indicator variables which are: the number of bedrooms ($BED2 = 2$, an indicator variable as to whether a house has 2 bedrooms, $BED3 = 3$, whether a house has 3 bedrooms, $BED4 = 4$, whether a house has 4 bedrooms, $BED5 = 5$ whether a house has 5 bedrooms) and the neighborhood ($W13 = 13$, is an indicator variable as to whether the house is in neighborhood 13, $W14 = 14$, whether the house is in neighborhood 14, $W15 = 15$, whether the house is in neighborhood 15).

After some preliminary analysis, it was found that a new variable called $EXSQFT$ which is the exposed square feet, the number of bedrooms, and the neighborhood variables may seem to affect the selling price of a house.

1. Provide the added variable plot of $PRICE$ versus $EXSQFT$, after controlling for $W13$, $W14$, $BED3$, $BED4$ and $BED5$
2. Explain how to construct the added variable plot.
3. Calculate the correlation coefficient produced from the residuals of the added variable plot. What do we call this coefficient? Can we calculate this coefficient using a simpler method. Provide the formula and the number calculated using that formula.
4. Fit the regression model of $PRICE$ on $EXSQFT$, $W13$, $W14$, $BED3$, $BED4$ and $BED5$
5. From the regression model fitted described in Part (4), several observations appear to be unusually large. identify the two unusually large residuals.
6. What proportion of the sum of squares error is represented by the two observations that you identify in part (4).
7. What proportion of the variance is represented by the two outliers.
8. Calculate the average leverage for the data set. Identify lower and upper bounds of the leverage point. Provide a rule for identifying a high leverage point.
9. From the regression model fitted, calculate the leverage for the observation number 4. What is your conclusion

Question 3

Data used here are: Loan

A study was conducted at a local savings and loan to understand the relationship between the size of a bank loan and various characteristics. A sample of 35 loans was drawn from approximately 24,000 home mortgage loans originated from 1984 to 1988. For each sample loan, we have information concerning the initial loan amount (LNMT) and the monthly income (INCOME). Financial characteristics of the borrower include the total outstanding debt (net of the mortgage, denoted by LOGNETWH). Characteristics of the loan include the loan to value ratio (LTV). You decide to see first if we can relate LNMT and the other variables.

1. Perform a backward stepwise regression of LNMT to the eight variables and provide the model suggested by the backward regression. We decided to include INCOME along with the variables remaining from the stepwise regression.
2. Fit a regression model relating LNMT and the five variables produced by the stepwise regression.
3. Provide the VIF of the five variables remained to explain LNMT. Which variable/variables should be removed from this analysis?
4. It has been hypothesized that having a higher loan to value ratio qualifies a borrower for a higher loan. Use the output above to test the null hypothesis that the regression coefficient associated with LTV is 0 versus the alternative that it is not equal. State your null hypothesis, alternative hypothesis and all components of the decision making rule. Use a 5% level of significance.
5. It has been hypothesized that the regression coefficient associated with INCOME is 4. Use the model fit in part (2) to test this assumption. Use 5% level of significance.
6. You are working for a client who would like to understand the size of a typical loan with the client's characteristics. Suppose that the client enjoys \$150,000 in net worth, has a monthly INCOME of \$12,000 and would like to work with a loan value ratio of 0.90. To provide a benchmark figure, calculate the loan amount, in dollars, for your client under the model fit in part (3). (Note that the net worth is modeled using natural logarithm units).
7. Provide its mean response and individual response 95% confidence interval

Question 4

Data used here are called COMP.

A firm CEO leads by developing and implementing a strategic policy for the firm. CEO salaries in the United States are of interest because of their relationship to salaries in international firms and to salaries of people that do not belong to Corporate America. The data for this study were drawn in 1992. The goal of this report is to study CEO and firm characteristics to determine the important factors influencing CEO compensation.

99 observations were selected from 800 companies representing the largest publicly traded companies in the United States. Firm characteristics studied in this report include:

- COMP: Sum of the salary, bonus and other compensation,
- SALES: Sales revenues, it measures the size and profitability of the firm,
- EDUC: Educational level,
- PCNTOWN: Percentage of the company's market value owned by the CEO,
- VAL: The market value of the CEO's stock in natural log,
- EXPER: Number of years as the firm CEO,
- TENURE: Number of years employed by the firm,
- PROF: Profits of the firm, before taxes,
- AGE: age of the CEO
- PERCENT5: Indicates that the CEO does own more than 5% of the firm's stock

1. Create the new variable PERCENT5
2. Plot the histogram of the variable "COMP". What do you observe? Is it symmetric?
3. We may act here by using a logarithm transformation LN to COMP?
4. Can we recommend the following linear regression:

$$LN_COMP = \beta_0 + \beta_1 LN_SALES + \beta_2 EXPER + \beta_3 EDUC + \beta_4 PERCENT5 + \beta_5 PROF + e$$

(we decided to remove the variable PCTOWN for the set of the chosen explanatory variables)

5. What does a negative slope associated with PERCENT5 mean?
6. What does a negative slope associated with EDUC mean?
7. What does a positive slope associated with LN_SALES mean?
8. What does a positive sign associated with EXPER mean?
9. Estimate the compensation and the expected 95% confidence interval for the following executive who was a CEO of Ameritech who: (a) has 8 years of experiences as CEO, (b) has a bachelor, (c) owned approximately 0.05% of the Ameritech stock, (d) had profits of \$1,166 millions on sales of \$10,818 millions.
10. What percentage does the model explain of the variation in compensation?
11. Is each explanatory variable important? why?
12. Is there any collinearity within the variables?
13. How many unusual observations do we have? which one is a leverage point and which one is an outlier?

14. Does the model improve when removing those unusual observations?

Question 5

Consider a situation where experimental design can be used. The effect of different selling approaches on sales of computers is to be studied. Three different selling approaches are to be compared. The object is to determine whether there is a difference in the effectiveness of the selling approaches. We will judge differences in effectiveness by looking at the average sales of the three approaches. An approach with significantly higher average sales will be judged more effective. Fifteen salespeople are chosen to participate in the study. Five salespeople each are randomly assigned to use one of the three approaches for the next month. At the end of the month, sales figures will be computed for each salesperson. These data will be analyzed to determine whether the sales approaches produce the same or different average sales.

In this situation, the salespeople were randomly assigned to the factor levels or treatments. This random assignment is possible when an experiment is designed to help answer a particular question. The type of experimental design used in this case is called a *completely randomized design*. Analysis of variance (ANOVA model) can be used in this situation.

Salesperson	Approach A	Approach B	Approach C
1	15	19	28
2	17	17	25
3	21	17	22
4	13	25	31
5	12	30	34

The hypotheses to be tested are $H_0: \mu_A = \mu_B = \mu_C$ versus H_a : At least one of means is not equal to the other means, where μ_i is the population average sales for selling approach i .

1. Give the summary output of the one-way factor analysis using JMP.
2. To test the hypotheses, the F statistics is used. What is the f_{value} calculated using the F table? Use the F statistics and the f_{value} to make a decision.
3. What is the decision, which means is the population average sales differ depending on what selling approach is used or not?
4. Note that the rejection of the null hypothesis simply says that the population means are not all equal. It doesn't say that they are all different or tell which ones are different from the others. In a case such as this, the researcher probably wants to know which population averages are different or whether two particular ones differ. A confidence interval estimate of the difference between two means, μ_i and μ_j can be constructed as follows:

$$(\bar{y}_i - \bar{y}_j) \pm t_{(1-\alpha/2)} \times s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $t_{(1-\alpha/2)}$ is calculated from the t-table with *probability* $(1 - \alpha/2)$ where α is the confidence level, and $n - c$ degree of freedom (c is the number of level), s is the square root of the MSE, and \bar{y}_i and \bar{y}_j are the sample means for samples i and j (or factors i and j). Use this equation to calculate the 95% confidence interval estimate of the difference between selling approaches A and C.

5. Can we say the following: *Because the 95% confidence interval estimate of the difference between these two populations means does not contain zero, this suggests that the population means for methods A and C are not equal.*
6. Do the same thing to compare:

$$\mu_A - \mu_B \quad \mu_B - \mu_C$$