Homework 8

Here are the problems for the eighth homework assignment.

Multiple choice questions

1. A prediction interval based on a simple linear regression is

   A a time period during which it is safe to make predictions;

   B 2 standard errors from the regression coefficient;

   C a confidence interval for the regression coefficient;

   D the interval of Y values determined by the regression line from an interval of predicted X values;

   E a confidence interval for an anticipated value of Y given a specific value of X.

2. In the simple regression model, the two variables involved

   A play symmetric roles so they can be interchanged without affecting the results of analysis;

   B have asymmetric roles and one variable is called an independent variable and the other is called a response variable;

   C are called predicted variables as they predict future observations;

   D are called intercept and slope;

   E are estimated from the pairs of observations.

3. In the standard reporting of the results of a simple linear regression analysis,

   A the estimated regression coefficients provide a test of the statistical significance of the linear dependence in the data;

   B the t-ratios, which are the ratios of the regression coefficients to the corresponding standard errors, are used in tests of the statistical significance of the regression coefficients;

   C the value of s may be ignored, along with those of R squared;

   D the sampling distribution of the regression coefficients is the standard Normal distribution with s used to estimate σ;

   E the values of the estimated regression coefficients $\pm 2s$, where is s^2 is estimator of the error variance, provide confidence intervals for the true coefficient values.

4. R^2 coefficient

   A is an artifact of regression analysis and should not be considered when reporting the results of regression analysis;

   B is very important in determining the intercept coefficient;

   C can not be computed unless we know exact values of slope and intercept;

D is equal to the estimate of error term variance;

E reports what proportion of variation in the response variable has been explained by its relation to the explanatory variable.

5. The correlation coefficient

   A is closely related to the slope coefficient in simple linear regression; if one is 0 then so is the other;

   B provides an improvement on the simple linear regression slope coefficient in that its interpretation is not restricted to simple linear relationships;

   C attempts to explain the variation in a response relationship in terms of slope and intercept;

   D is statistically insignificant unless its value exceeds 0.5;

   E is 0 when all deviations of data points from a straight line are 0.

6. The correlation coefficient is positive

   A when all deviations of observations from the fitted straight line are positive;

   B when the values of the response variable tend to increase as the values of the explanatory variable increase;

   C when the slope coefficient in the corresponding simple linear regression is statistically significant;

   D when both the slope and intercept coefficients in the corresponding simple linear regression are statistically significant;

   E when the relationship between the Y and X variables is desirable.

7. Which of the following statements about the residuals is *not true*:

   A the residuals in simple regression represent deviations of the actual data from the fitted line;

   B the sum of squared residuals when divided by the sample size less two serves as the estimator of the error term variance;

   C sum of residuals is equal to zero;

   D if slope is significantly non-zero, then all residuals have to be equal to zero;

   E the sum of squared residuals represents the portion of the total variability of the response variable that is not explained by the explanatory variable.

Problems

1. In the textbook and lecture, the case of US mail has been fitted by simple regression model after excluding three values, namely two for the Christmas periods corresponding to period 7 of both fiscal years and one that was corresponding to the equipment malfunction in period 6 of Fiscal Year 1963. The fitted model with two sigma error band is

$$Y = 50 + 3.3X \pm 20.$$

- Given the Volume figures for periods 7 of Fiscal Year 1962, 6 and 7 of Fiscal Year 1963, what predictions, including prediction errors, would you make for the Manhours requirement?
- How do these predictions relate to the actual manhours used?
- Do the same for periods 6 of Fiscal Year 1962, 1, 5 of Fiscal Year 1963.
- Compare the results and comment.

Here are the values of the volume and corresponding manhour for the periods amentioned:

Period	Volume	Manhour
Period 6, 1962	184	671
Period 7, 1962	268	1053
Period 1, 1963	154	569
Period 5, 1963	191	700
Period 6, 1963	180	765
Period 7, 1963	270	1070

2. In a study of a wholesaler's distribution costs, undertaken with a view to controling cost, the volume of goods handled and the overall costs were recorded for one month in each of ten depots in a distribution network. The results are presented in the following table

	Volume	Costs
1	48	20
2	57	22
3	49	19
4	45	18
5	50	20
6	62	24
7	58	21
8	55	21
9	38	15
10	51	20

In the following you are asked to perform a full regression analysis of the cost (Y) on the volume (X).

- Plot a scatter plot of the data and after inspecting it answer the following two questions:
  - Does the data appear to follow a simple regression model?
  - Are there any "suspicious" points that should be excluded before fitting the model?

- In the lecture and in the book the following formulas have been given for the least square fit of the slope and the intercept:

$$\widehat{\beta} = \frac{\sum_{i=1}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{Y})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - \overline{X}^2},$$
$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{X}$$

  Evaluate these formulas and plot the regression line on the graph. The following results of calculations can help you in this task:

$$48 + 57 + 49 + 45 + 50 + 62 + 58 + 55 + 38 + 51 = 513$$
$$20 + 22 + 19 + 18 + 20 + 24 + 21 + 21 + 15 + 20 = 200$$
$$48^2 + 57^2 + 49^2 + 45^2 + 50^2 + 62^2 + 58^2 + 55^2 + 38^2 + 51^2 = 26757$$
$$20^2 + 22^2 + 19^2 + 18^2 + 20^2 + 24^2 + 21^2 + 21^2 + 15^2 + 20^2 = 4052$$
$$48 * 20 + 57 * 22 + 49 * 19 + 45 * 18 + 50 * 20 + 62 * 24 + 58 * 21 +$$
$$+55 * 21 + 38 * 15 + 51 * 20 = 10406$$

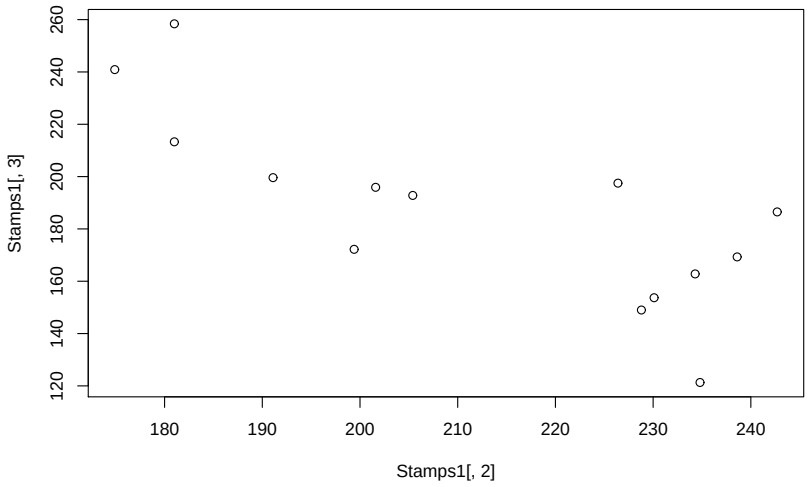
- Evaluate residuals to the fitted models and compute the estimator of the error term variance.

- Add to your graph the two-sigma control limits based on the obtained estimator of variance and comment if the cost data appear to be under control.

- Evaluate R^2 coefficient and comment about the percentage of variation of the data that is explained by the linear regression model.

- Compute correlation coefficient between volume and cost. What is the reduction in the error of prediction of a manhour value by accounting on the regression on the mail volume?

3. In Problem 4, Homework 1, you were asked to make a visual fit of a straight line to the stamp vs. metered mail data starting from year 1964 and ending in year 1977 (inclusive). A scatter plot of these data is given below



while the data themselves are as follows

	Year	Stamp.Sales	Meter.Sales
1	1964	234.8	121.3
2	1965	228.8	149.0
3	1966	230.1	153.7
4	1967	234.3	162.8
5	1968	238.6	169.3
6	1969	242.7	186.5
7	1970	226.4	197.5
8	1971	199.4	172.2
9	1972	205.4	192.8
10	1973	201.6	195.9
11	1974	191.1	199.6
12	1975	181.0	213.3
13	1976	174.9	240.9
14	1977	181.0	258.4

The following is an output from a computer program that represent results of regression analysis performed on the above data.

```
Residual Standard Error=23.5639
R-Square=0.6095
F-statistic (df=1, 12)=18.7292
p-value=0.001

          Estimate Std.Err t-value Pr(>|t|)
Intercept 435.9786 57.9535  7.5229    0.000
X          -1.1752  0.2716 -4.3277    0.001
```

- Write down the equation for the least square fit to the data (use the computer printout to get the coefficient for this line). Then write down the line you have got in Problem 4, Homework 1.
- On the included scatter plot, draw both the lines. Comment on the similarities and differences.
- Write down the fitted model with two sigma control limits.
- Is the slope of the fit statistically significant? Why?
- How much of variation in the model is accounted by the linear regression relation between the two variables?
- The standard deviation for the stamp sales $S_X = 24.1$ while for the metered mail is $S_Y = 36.2$. What is the correlation between the variables?
- In 1978, the observed value of Stamp Sales was 188.2. Based on this value and the fitted model, predict the value of Metered Sales.
- The actual value of Metered Sales was reported as 240.8. Is this value within two sigma control limits of the prediction?