Homework 8

Here we present solutions to the problems posted in the eighth homework assignment. The solutions and related commentary are put in italics. Remember that problems can have several different but still correct ways of solving them.

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.

   Solution: *The regression line allows us to make predictions about the response variable Y given a considered value of X. A prediction interval tells us the range of possible values for Y in which, with high confidence, we can locate the actual value of Y when it is observed.*

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.

   Solution: *The regression model is about two variables observed in pairs, with one called an independent variable (also known as explanatory variable or predictor) and the other called a response variable (also known as dependent or predicted variable). Their role is not interchangable both in analysis and in interperetation as one is used to explain variability of the other.*

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.

Solution: *The t-ratios are defined as $\hat{\alpha}/SE(\hat{\alpha})$ and $\hat{\beta}/SE(\hat{\beta})$ and are used for testing if either intercept α or slope β are non-zero.*

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.

Solution: *If the total variation of Y variable is measured by the sum of squares of its deviations from its mean, then R^2 is proportion of this variation that equals to the sum of squared deviations of predictions of Y from its mean, i.e. the proportion of variation due to the relation with X.*

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.

Solution: *The correlation coefficient is another way of measuring linear relation between variables and thus is closely related to the slope coefficient. Namely, the relation has the form*

$$\hat{\beta} = r\frac{S_Y}{S_X},$$

where r is the correlation coefficient and $\hat{\beta}$ is the estimate of slope.

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.

Solution: *The correlation coefficient is always a number between -1 and 1. Its value indicates the strength of linear relation between two variables, with positive values indicating direct proportional relation, i.e. larger values of the explanatory variables generally correspond to larger values of the response variable, while negative values indicating inverse proportional relation.*

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.

   Solution: *The residuals represent deviations of Y from the fitted line so they are not dependent on if slope is significant or not. All residuals are equal to zero only if the data lie perfectly on a straight line.*

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

Solution:

- *The model tells us the following predictions for three periods:*

$$\widehat{Y} = 50 + 3.3 * 268 \pm 20 = 934.4 \pm 20$$
$$\widehat{Y} = 50 + 3.3 * 180 \pm 20 = 644 \pm 20$$
$$\widehat{Y} = 50 + 3.3 * 270 \pm 20 = 941 \pm 20$$

- *The actual values are 1053, 765, and 1070. In all three cases, the actual manpower have been underestimated from the model by about* 120 ± 20*.*
- *The model tells us the following predictions for three periods:*

$$\widehat{Y} = 50 + 3.3 * 184 \pm 20 = 657.2 \pm 20$$
$$\widehat{Y} = 50 + 3.3 * 154 \pm 20 = 558.2 \pm 20$$
$$\widehat{Y} = 50 + 3.3 * 191 \pm 20 = 680.3 \pm 20$$

  The actual observed values are 671, 569, 700. All observed values are well within the prediction error bands around the prediction values.
- *When comparing the fitted model predictions for the values that have not been used for the fit with the prediction for the values based on which the model has been fitted, we observe that the model underperform for the first ones and is quite accurate for the second ones. One could suggest another model for the Christmas period if more than just two year period data were available.*

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:

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

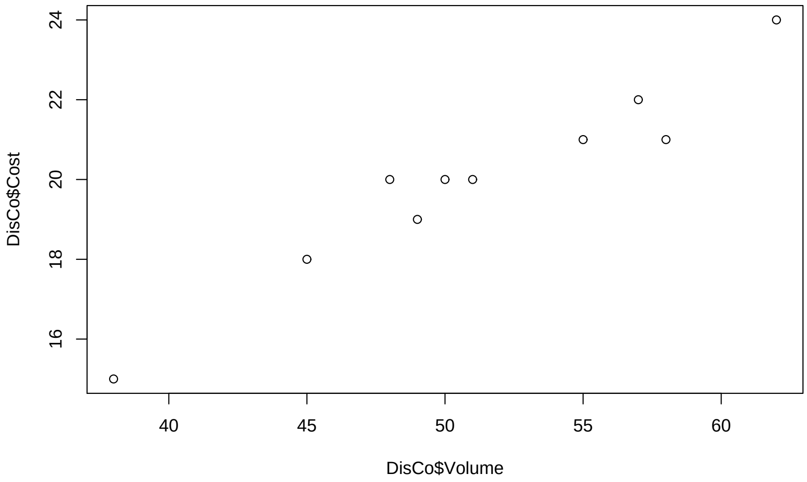
  Evaluate these formulas and plot the regression line on the graph. The following results of calculations can help you in this task:

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

- 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?

Solution:

- *A scatterplot of the data is presented below*



After examining the plot, we conclude that there is a strong linear relation between the two variables and we do not detect any unusual deviations from the simple linear regression model.

- *Using the second formula for the slope we obtain*

$$\widehat{\beta} = \frac{10406 - 10 * (513/10) * (200/10)}{26757 - 10 * (513/10)^2} \approx 0.33$$

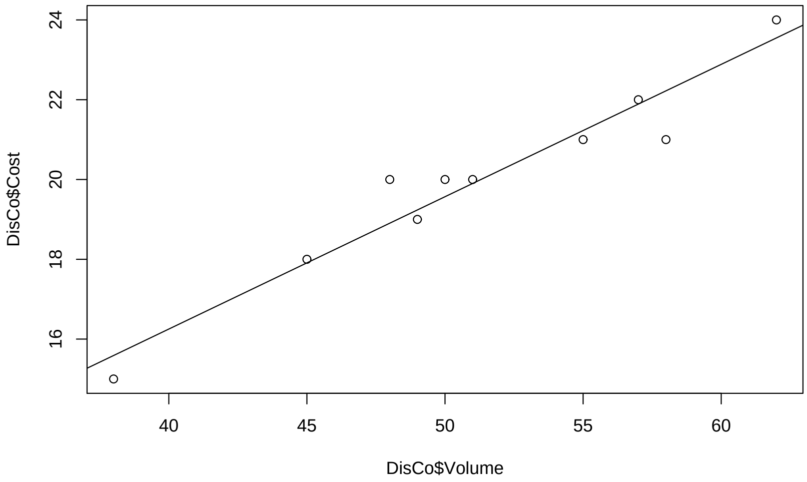
and the intercept can be obtained as follows

$$\widehat{\alpha} = 20 - 0.33 * 51.3 = 2.98.$$

Consequently, the fitted line has the equation

$$\widehat{Y} = 2.98 + 0.33 * X.$$

The line together with the scatterplot of the data is seen below.



- *In order, to evaluate the residuals one has first to evaluate the cost predictions at the given volume data. These are obtained by evaluating* $\widehat{Y}$ *from the regression equation* $\widehat{Y} = 2.98 + 0.3317 * X$ *at* X *values:* $48, 57, 49, 45, 50, 62, 58, 55, 38, 51$. *We get*

  $$18.9, 21.9, 19.2, 17.9, 19.6, 23.6, 22.2, 21.2, 15.6, 19.9$$

  The residuals are then obtained by subtracting from the observed values Y*:* $20, 22, 19, 18, 20, 24, 21, 21, 15, 20$ *the above predictions* $\widehat{Y}$*. We obtain*

  $$1.1, 0.1, -0.2, 0.1, 0.4, 0.5, -1.2, -0.2, -0.6, 0.1$$

  Finally, the estimator of the error term variance σ^2 *is given by the formula*

  $$S^2 = \frac{\sum_{i=1}^n e_i^2}{n-2},$$

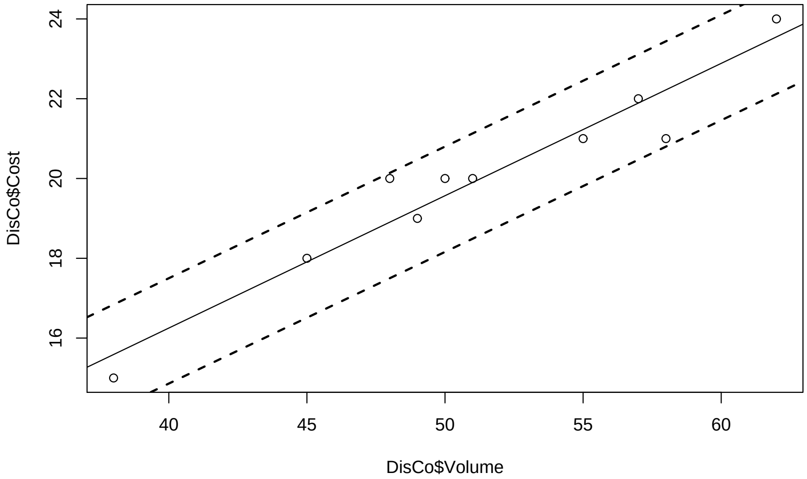
  where e_i*'s are the residuals. For our data, we get*

  $$(1.1^2+0.1^2+0.2^2+0.1^2+0.4^2+0.5^2+1.2^2+0.2^2+0.6^2+0.1^2)/8 = 3.53/8 \approx 0.44$$

- *The estimated value of* σ *is given by* $\sqrt{0.44} \approx 0.66$. *Thus two sigma control band is approximately given by*

  $$Y = 2.98 + 0.33 * X \pm 1.33$$

  and the plot of the control bands is given below



As all the points are within the control limits and we do not observe any unusual patterns, we conclude the cost process as a function of the volume is under control.

- *The R^2 coefficient is computed from the formula*

$$R^2 = \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n Y_i^2 - n\overline{Y}^2}$$

with the last expression being more convienient as we have already computed that the sum of squared residuals is 3.53 and sums of Y and its squares are given in the problem. Consequently,

$$R^2 = 1 - 3.53/(4052 - 10 * 20^2) \approx 1 - 0.0679 = 0.9321.$$

We conclude that slightly above 93% variation in costs has been explained by their relation to volume.

- *The correlation coefficient is given by*

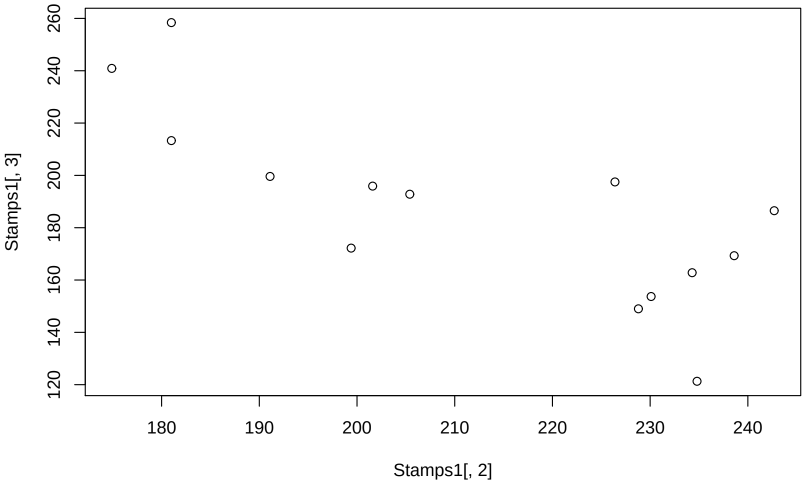
$$r = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}} = \frac{\sum_{i=1}^n X_i Y_i - n\overline{XY}}{\sqrt{\left(\sum_{i=1}^n X_i^2 - n\overline{X}^2\right)\left(\sum_{i=1}^n Y_i^2 - n\overline{Y}^2\right)}}.$$

This leads us to

$$r = \frac{10406 - 10 * 51.3 * 20}{\sqrt{(26757 - 10 * 51.3^2) * (4052 - 10 * 20^2)}} \approx 0.965$$

and since the value is close to one we conclude a strong linear relation between volume and costs. In particular, the reduction in the prediction error *of Y given some value of X is by factor $\sqrt{1-r^2} \approx 0.26$, i.e. if we predict Y just using $\overline{Y}$, the prediction error based on two sigma rule is $\pm 2 * S_Y = \pm 2 * \sqrt{(405.2-20^2)10/9} \approx \pm 2 * 2.40 = \pm 4.8$, while if we predict Y using the regression model $Y = +X$, then the approximated prediction error is $\pm 2 * \sqrt{1-r^2} * S_Y \approx \pm 2 * 0.26 * 2.40 = \pm 1.24$[1]. We note a quite dramatic improvement in accuracy.*

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
```

[1] The exact value of the reduction term should be $\sqrt{(n-1)/(n-2)} \cdot \sqrt{1-r^2} \approx 0.278$, which yields the regression prediction error 1.32 in better agreement with earlier discussion on the prediction error.

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?

Solution:

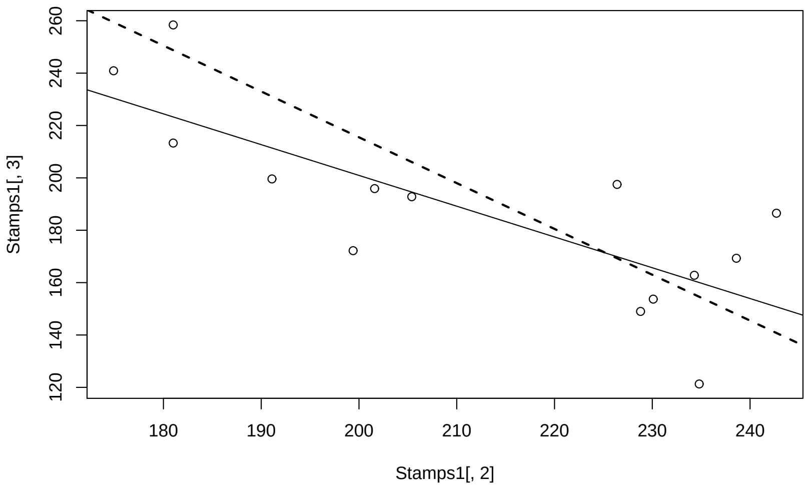
- *The equation for the least square fit is*

  $$y = -1.18x + 436.0.$$

  Here we have used the rounded up coefficients listed in the computer printout. The equation of the line obtained in Problem 4, Homework 1 is

  $$y = -1.75x + 565.5.$$

- *On the graph below the dotted line represents the visually fitted line while the solid line corresponds to the least square fit*



We observe that the lines differ quite a lot although they both capture negative slope and general inverse proportional dependence of the metered sales on the stamps sales.

- *Using the computer output we get the standard deviation to be 23.56 and thus the two sigma control limits for the fitted model are* $y = -1.1752x + 435.9786 \pm 47.12$.
- *We observe in the printout that the standardized statistic for the slope coefficient is given by*

  $$T = \frac{\widehat{\beta}}{\widehat{SE}(\widehat{\beta})} = \frac{-1.1752}{0.2716} = -4.3277$$

  and comparing this with the critical value ± 2.18 *based on the Student-t distribution with* $n - 2 = 12$ *degrees of freedom at the significance level 5%, we have to reject the hypothesis that the slope is zero. In fact, the p-value as reported in the printout is listed at 0.001 (1%).*
- R-squared *is the measure of variability explained by the linear regression relation between the two variables and it is listed in the printout at about 61%.*
- *The correlation between variables can be computed from the formula*

  $$r = \widehat{\beta}\frac{S_X}{S_Y} = -1.1752\frac{24.1}{36.2} \approx -0.78.$$

- *Using the observed value of Stamp Sales was 188.2, we get the prediction* $\widehat{y} = -1.1752 * 188.2 + 435.98 \pm 47.12 \approx 214.8 \pm 47.12$.
- *The two sigma control limits around the prediction are 167.68 and 261.92, and they include between them the observed value 240.8 of metered sales for this year.*