# Solutions to Homework 1

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

## Multiple choice questions

1. In the weekly sales in a sports club, we have seen a seasonal pattern (a), exceptional high values during Christmas week and long weekends(b), and changes of deviation from seasonal pattern on the week-to-week basis (c). Out of these three [(a), (b), (c)] sources of variation in the data

   **A** (a) represents an assignable cause of variation, while (b) and (c) represent variation caused by chance;

   **B** (a) represents variation caused by chance, while (b) and (c) represent assignable cause of variation;

   **C** all three (a), (b), and (c) represent assignable causes of variation;

   ***D** (a) and (b) represent assignable causes of variation, while (b) represents variation caused by chance;

   **E** all three (a), (b), and (c) represent variation caused by chance.

   *Please, consult the definition of chance and assignable causes of variation given in Section 1.4, p. 19. Specifically, we can clearly consider both seasonal change and holidays as predictable factors that each individualy has an effect on the variable of interest (in this case, sales in the club). On the other hand week to week changes of deviations from seasonal pattern are unpredictable in values and individually not very influential on overall value.*

2. The process view of data

   ***A** displays data in a time ordered line plot;

   **B** displays data in histogram format;

   **C** processes data by summarising the main features in numerical form and displays the resulting summaries;

   **D** uses statistical data processing software to provide a professional level summary of the data;

   **E** summarises the data in the shape of a Normal curve.

   *The answer is obvious in view of the definition of the process view (please, consult Section 1.1 pages 3-6).*

3. There are available data on weekly precipitation in a town and weekly beer sales in a local bar over a period of ten years. To see if considering weather prediction in running the bar business is worth of an effort one could

   **A** draw histograms of these two data sets and examine if their shapes are similar;

***B** draw a scatterplot view of the data examine its shape and detect any form of dependence between the two variables;

**C** check during which week the most beer was sold and check how much rain there was during this week;

**D** write data into the table and examine if there are any relation between values;

**E** check if the histograms for both data sets resemble a Normal curve.

*The goal is to identify any dependence of the beer sales on amount of rain. From Section 1.5, we know that scatterplots are very convenient graphical tools for detecting dependences between two variables.*

4. Assignable causes of variation

**A** always follow the Normal model;

**B** follow the Normal model when there are no chance causes of variation;

**C** correspond to the most frequent classes in a histogram;

**D** are the many unpredictable but collectively influential factors that affect a process or system;

***E** are the few factors, individually influential and with predictable effect, that affect a process or system.

*This answer follows from the discussion of the answer to the first question.*

5. The Normal model for chance variation

***A** is a flexible family of frequency distributions with mean and standard deviation as parameters;

**B** is the usual model for the linear part of a simple linear regression equation;

**C** has a frequency distribution which is evenly spread across the scale;

**D** is the standard model for the assignable causes in a chance system;

**E** is the result of a series of orderly experiments.

*The normal family of distributions, i.e. the Normal model is discussed on pages 20-22, Section 1.4 of the textbook. In the particular, the model is characterized by the center expressed as the mean and by the spread that can be expressed in the terms of the standard deviation.*

6. Which one of the following is correct?

**A** response relationships reflect the flow in a process;

***B** a response variable is seen as responding to changes in an explanatory variable;

**C** an explanatory variable is one in which the explanation changes from time to time;

**D** an explanatory variable is one whose values are explained by the order of a process;

**E** explanatory variables may be correlated with the response variable but not with each other.

*The question deals with regression models in which we have two types variables: explanatory ones, and response. In the regression model, it is naturally to assume that there is a cause-effect relation between explanatory and response variables.*

7. A simple linear regression model

   **A** is one where the vertical axis moves closer to the horizontal axis;

   **B** is one where all the points fall on a straight line;

   **C** attempts to explain the variation in a response variable in terms of distance from a straight line;

   ***D** attempts to explain the variation in a response variable in terms of a linear relation with an explanatory variable and chance variation;

   **E** is another name for a scatter plot.

   *The mathematical formulation of the simple regression model is*

   $$Y = \alpha + \beta X + \epsilon,$$

   *where $X$ is explanatory variable, $Y$ is response variable, $\alpha + \beta X$ represents a linear relation in $X$ and $\epsilon$ stand for chance variation.*

8. The prediction error associated with a simple linear prediction formula

   **A** is conventionally set at plus or minus twice the standard error of the mean;

   **B** is always less than the prediction itself;

   **C** is less than the prediction itself for large values and greater for small values;

   **D*** is conventionally set at plus or minus twice the standard deviation of chance variation;

   **E** improves as the degree of extrapolation increases.

   *The prediction formula is conventionally presented as*

   $$\hat{Y} = \alpha + \beta X \pm 2 * \sigma,$$

   *where $\hat{Y}$ is the prediction and $\sigma$ is standard deviation of $\epsilon$.*

9. The regression coefficients in a multiple linear regression

   ***A** are not readily interpretable when the explanatory variables are related;

   **B** are determined in advance through the process of problem formulation;

   **C** are only meaningful in an observational study;

   **D** are also referred to as indicator variables;

   **E** indicate by their values which version of the prediction formula is appropriate.

   *Compare with the comment made at the top of page 44. Read also the entire paragraph on interpreting regression coefficients pp. 43-44.*

10. Explanatory variables in a statistical prediction model

**A** explain why the prediction did not work;

**B** explain the prediction error involved in forecasting;

***C** are related to the variable being predicted;

**D** are the coefficients in a multiple regression equation;

**E** are variables whose values explain why prediction is necessary.

*In the regression model, the value being predicted is the response variable. We have commented previously that the response variable is dependent on explanatory variables in the cause-effect relation.*

## Problems

1. In the process of manufacturing of tennis balls three new presses have been acquired. After running several batches of tennis balls through the production process, it has been established that the mean values of the tennis ball core diameters for the three prersses are 95, 77, 90, respectively, and their standard deviations are 1.5, 4.1, 5.5. If the quality bands set by the Tennis Federation are: the lower bound is 70 while the upper one is 100, and the manufacturer sets the two standard deviations rule for the press, find out which of the three presses will pass the quality inspection. What is the answer to this question in the case when the manufacturer puts a more stringent three sigma rule as the quality requirement?

   *The two standard deviation rule (or two sigma rule) means that the band around mean given by*

   $$\mu \pm 2\sigma$$

   *will be within* $(70, 100)$ *interval. Here are the computation of* $2\sigma$ *bands for each of the presses*

   $$\begin{array}{lcl} 95 \pm 2 * 1.5 & = & 95 \pm 3 \\ & (92, 98) & \\ 77 \pm 2 * 4.1 & = & 77 \pm 8.2 \\ & (68.8, 85.2) & \\ 90 \pm 2 * 5.5 & = & 90 \pm 11 \\ & (79, 101) & \end{array}$$

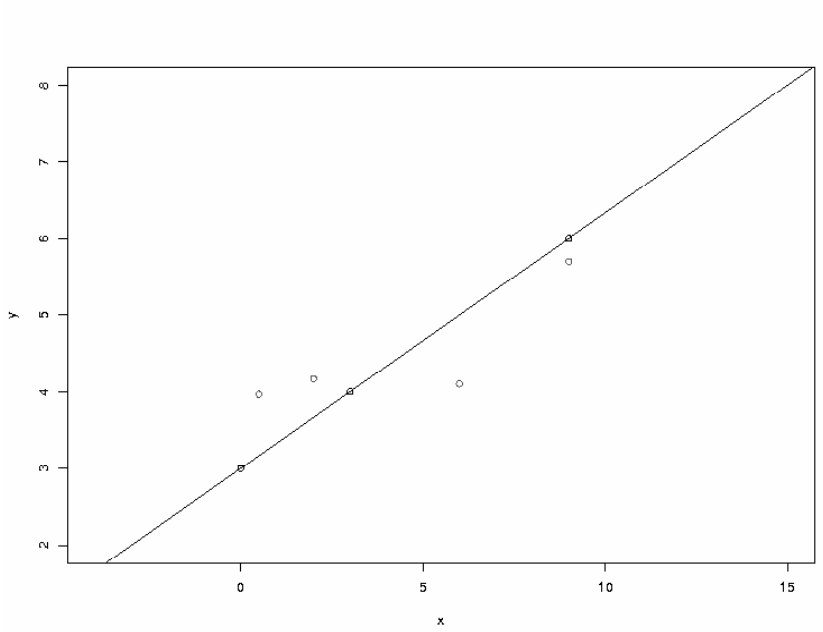
   *We conclude that only the press number one meets the quality standard set by the manufacturer. For the three sigma rule the first press yields the limits* $(90.5, 99.5)$ *which is still within quality bands required by the federation.*

2. On a graph of $y$ against $x$, plot the points $(3, 4)$, $(0, 3)$, $(9, 6)$ and the line with equation $y = 3 + \frac{1}{3}x$. If the simple linear model is represented by

   $$Y = 3 + \frac{1}{3}X + \epsilon$$

   and for values of $X = 0.5, 2, 9, 6$ the error term $\epsilon$ took values $0.1, 0.05, -0.1, -0.2$, respectively, plot on the graph the observed values of $Y$.

   *The answer is represented in the following graph*




```
The following is the R-code to produce the graph.
postscript("lineplot.eps")
x=c(3,0,9)
y=c(4,3,6)
plot(x,y,xlim=c(-4,15),ylim=c(2,8),xlab="x",ylab="y")
abline(3,1/3)
xx=c(0.5, 2, 9, 6)
eps=c(0.8,0.5,-0.3,-0.9)
yy=3+xx/3+eps
par(new=T)
plot(xx,yy,col="red",xlim=c(-4,15),ylim=c(2,8),axes=F,xlab="",ylab="")
dev.off()
```

3. The simple linear regression model that was proposed in the textbook for the relation between manpower and volume of processed mail has been given by

$$\hat{Y} = 50 + 3.3X + \hat{\epsilon},$$

where the prediction error $\hat{\epsilon}$ is in $\pm 20$ range (this according to the two sigma rule so the standard deviation of chance variation is 10). Use the above prediction formula to estimate the loss incurred through equipment breakdown in period 6, fiscal 1963, and to predict the *extra* manpower requirement during the Christmas period, based on the experience of period 7, fiscal 1962 and 1963. We refer here to Table 1.3, p.24.

*For period 6, fiscal 1962, the volume of mail* $X = 180$. *Thus according to the prediction formula*

$$Y = 50 + 3.3 * 180 \pm 20 = 644 \pm 20$$

*The actual value observed during this period is* 765 *which a way more then two sigma band permits. We can estimate the loss in terms of manhour as* $765 - 644 \pm 20 = 121 \pm 20$, *i.e. between* 101 *and* 141.

*As for the Christmas period, we observe the volumes of mail 268 in 1962 and 270 in 1963. The volume prognosis for the Christmas periods are*

$$Y = 50 + 3.3 * 268 \pm 20 = 934.4 \pm 20,$$

*and*

$$Y = 50 + 3.3 * 270 \pm 20 = 941 \pm 20.$$

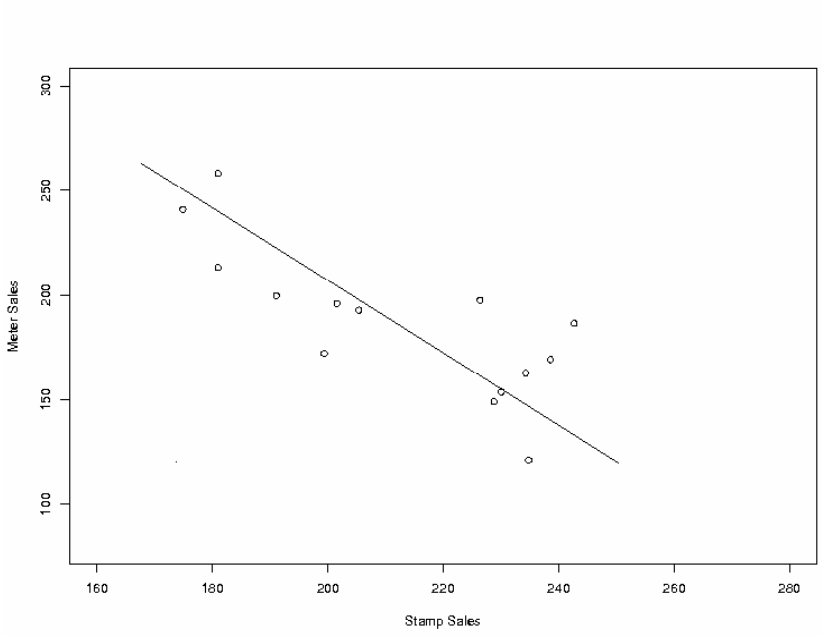
*However, the observed values of manhours are* 1053 *and* 1070*, thus we observe that there was need for extra* $1053 - 934.4 \pm 20 = 118.6 \pm 20$ *and* $1070 - 941 \pm 20 = 129 \pm 20$ *manhour, respectively. Thus it would be save to say that the extra manhour needed for Christmas period is roughly somewhere between* 96.6 *and* 149*.*

4. Consider the data in Table 1.4, pp. 33. Using a pencil and paper construct a scatter plot of the stamp sales vs. meter sales in the sixties. Do you observe any relation between these two variables? If yes, draw (by hand) a straight line that in your opinion the best represents this relation. After that write an approximate equation for this line in the form

$$Y = a + bX,$$

i.e. find approximate values of coefficients $a$ and $b$ from your graph. Finally, using this equation evaluate prediction errors for the data from the sixties.

*The following figure has been created by a computer software but the line has been drawn by hand.*




*From the figure we see that the slope is approximately* $(125 - 265)/(250 - 170) = -7/4$*, so the line can be written as*

$$y = -7/4x + b,$$

*where* $b = 125 + 250 * 7/4 = 562.5$*. Thus the line is given as* $y = -1.75x + 565.5$*. First, we evaluate the predictions of meter sales for the years 1964 through 1977 by plugging values of stamp sales for the* $x$*-variable in the above equation. We obtain*

```
154.600 165.100 162.825 155.475 147.950
140.775 169.300 216.550 206.050 212.700
231.075 248.750 259.425 248.750
```

*Then the prediction errors can be obtained by substracting from the predicted values the observed values of meter sales which yields*

```
 33.300  16.100   9.125  -7.325 -21.350
-45.725 -28.200  44.350  13.250  16.800
 31.475  35.450  18.525  -9.650
```

*The following is R-code that assisted the above solution.*

```
Stamps=read.csv("Stamps.csv")
SixSev=Stamps[Stamps$Year<1978,]
SixSev=SixSev[SixSev$Year>1963,]
postscript("SixSev.eps")
plot(SixSev$Stamp.Sales,SixSev$Meter.Sales, xlim=c(160,280),ylim=c(80,300),xlab="Stamp Sales", ylab="Meter Sales")
dev.off()

y=-1.75*SixSev$Stamp.Sales+565.5

y
 [1] 154.600 165.100 162.825 155.475 147.950 140.775 169.300 216.550 206.050
[10] 212.700 231.075 248.750 259.425 248.750

y-SixSev$Meter.Sales

[1]  33.300  16.100   9.125  -7.325 -21.350 -45.725 -28.200  44.350  13.250
[10]  16.800  31.475  35.450  18.525  -9.650
```

5. The prediction formula for stamp sales that was found in the textbook, see page 41, is given by
$$Y = 340 - 0.0316X_1 - 70.2X_2 \pm 8,$$
(where the bounds for error are given by the two sigma rule so the standard deviation $\sigma$ is equal to 4). Here $X_1$ denotes Gross National Product (GNP) and $X_2$ represents Real Letter Price (RLP), the latter being defined as the standard stamp value divided by the Consumer Price Index (CPI).

   Suppose that the progonosis for 1984 of GNP is 1484.5 and of RLP is 1.835. For 1985, the Central Bank predicts 1.5% increase of GNP and the inflation rate of 5.5% (inflation is increase of CPI expressed as the precentage of the previous year CPI). Use these values to obtain the predicted stamp sales for 1985.

   *In view of the prediction formula, our goal is to evaluate GNP* $X_1$ *for* 1985 *and RLP* $X_2$ *for the same year. Using the information given in the problem,* $X_1 = 1484.5+0.015*1484.5 = 1506.767$ *and* $X_2 = 1.835/1.055 = 1.739336$. *Consequently, the predicted stamp sales for 1985 are*
$$Y = 340 - 0.0316 * 1506.767 - 70.2 * 1.739336 \pm 8 = 170.2848 \pm 8.$$

6. Predict the effects on stamp sales in 1984 and 1985 of 5% and 10% increases in stamp prices and 5% decreases in stamp prices.

   *A change of stamp prices affects RLP* $X_2$. *In the year* 1984, *the values of* $X_2$ *for* 5%, 10% *increases and* 5% *are* $1.835 * 1.05 = 1.92675$, $1.835 * 1.1 = 2.0185$, $1.835 * 0.95 = 1.74325$, *respectively. Consequently the predicted sales for this year would be* $Y = 340 - 0.0316 * 1484.5 - 70.2 * 1.92675 \pm 8 = 157.8319 \pm 8$, $Y =$

$340 - 0.0316 * 1484.5 - 70.2 * 2.0185 \pm 8 = 151.3911 \pm 8$, *and* $Y = 340 - 0.0316 * 1484.5 - 70.2 * 1.74325 \pm 8 = 170.7136 \pm 8$, *respectively.*

*Similarly, for the year 1985, the new values of RLP are* $1.92675/1.055 = 1.826303$, $2.0185/1.055 = 1.91327$, $1.74325/1.055 = 1.652370$, *respectively, and the resulting predicted sales are* $340 - 0.0316 * 1506.767 - 70.2 * 1.826303 \pm 8 = 164.1797 \pm 8$, $340 - 0.0316 * 1506.767 - 70.2 * 1.91327 \pm 8 = 158.0746 \pm 8$, $340 - 0.0316 * 1506.767 - 70.2 * 1.652370 \pm 8 = 176.3898 \pm 8$, *respectively.*

7. In the analysis of housing completions in the "good years" 1993-2000 of Irish economy, the following model has been proposed for the quarterly data

$$Y = 3937 + 259t,$$

where $t$ represents time period and $Y$ represents predicted completions (see pages 47-50 of the textbook). To account for difference in completions in each of the quarter due to seasonal effect, the following more general model has been proposed

$$Y = 3248Q1 + 3901Q2 + 4174Q3 + 5031Q4 + 250t \pm 500,$$

where $Q1, Q2, Q3, Q4$ are dummies variables equal either zero or one depending from which quarter data are considered (see also Table 1.7, p. 49). Write down separate prediction formulas for each of the four quarters. Evaluate predictions for each quarter of 2001 and of 2002. Do you think that you can use the same approach to predict completion for the year 2009? Support your answer with proper arguments.

*The prediction formulas for the quarters will be given as follows*

$$\begin{aligned} Y &= 3248 + 250 * t \pm 500, \quad t = 1, 5, 9, ..., 29, ...; \\ Y &= 3901 + 250 * t \pm 500, \quad t = 2, 6, 10, ..., 30, ...; \\ Y &= 4174 + 250 * t \pm 500, \quad t = 3, 7, 11, ..., 31, ...; \\ Y &= 5031 + 250 * t \pm 500, \quad t = 4, 8, 12, ..., 32, ... \end{aligned}$$

*From these formulas, the predictions for 2001 are:* $3248+250*33\pm500 = 11498\pm500$, $3901 + 250 * 34 \pm 500 = 12401 \pm 500$, $4174 + 250 * 35 \pm 500 = 12924 \pm 500$, *and* $5031 + 250 * 36 \pm 500 = 14031 \pm 500$, *respectively for each quarter. Similarly, for 2002:* $3248 + 250 * 37 \pm 500 = 12498 \pm 500$, $3901 + 250 * 38 \pm 500 = 13401 \pm 500$, $4174 + 250 * 39 \pm 500 = 13924 \pm 500$, *and* $5031 + 250 * 40 \pm 500 = 15031 \pm 500$, *respectively for each quarter.*

*At present, the construction business is in a deep recession thus the linear growth represented by the above model is no longer valid. It is also supported by historical data, see Figure 1.30, p. 46, that the linear model is only justified for certain periods and for different such periods different coefficients should be considered.*