MATH3823 - Solutions to Chapter 1 Exercises

Exercise 1.1 Again we consider the beetle data from the Lecture Notes. In the by-hand calculations we need all the parts of: $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ and $\hat{\beta} = s_{xy}/s_x^2$.

The basic quantities are: $n = 8$, $\sum x_i = 14.3474$, $\sum y_i = 4.816257$, hence $\bar{x} = 1.793425$ and $\bar{y} = 0.6020321$. Further, $\sum x_i^2 = 25.7628383$ and $\sum x_i y_i = 8.8072082$. These lead to $s_{xy} = (\sum x_i y_i - n\bar{x}\bar{y})/(n-1) = 0.0045504$ and $s_x^2 = (\sum x_i^2 - n\bar{x}^2)/(n-1) = 0.0242303$, giving $\hat{\beta} = 5.324937$ and then, $\hat{\alpha} =$ -8.947843.

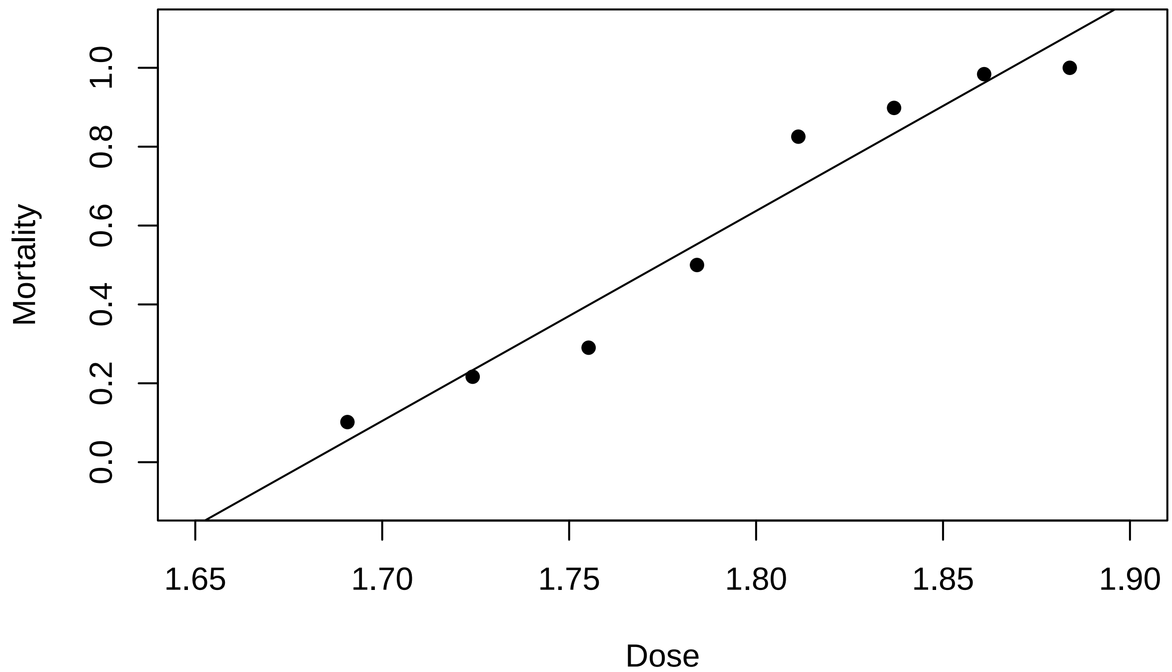
Checking these in R requires us to use it as a simple calculator to replicate every step in the hand calculation. An alternative, is to use the in-built R functions to give the final results and only replicating the intermediate values if needed – it is unlikely that you would get the final values correct without getting all the intermediate steps correct!

In R, to plot the data we repeat the steps in the code chunk from Lecture Notes and then add the regression line using the parameters just calculated by hand.

```r
beetle = read.table("https://rgaykroyd.github.io/MATH3823/Datasets/beetle.txt", header=T)

dose = beetle$dose
mortality = beetle$died/beetle$total

plot(dose, mortality, pch=16,
     xlim=c(1.65, 1.90), xlab="Dose",
     ylim=c(-0.1, 1.1),  ylab="Mortality")
abline(-8.947843, 5.324937)
```



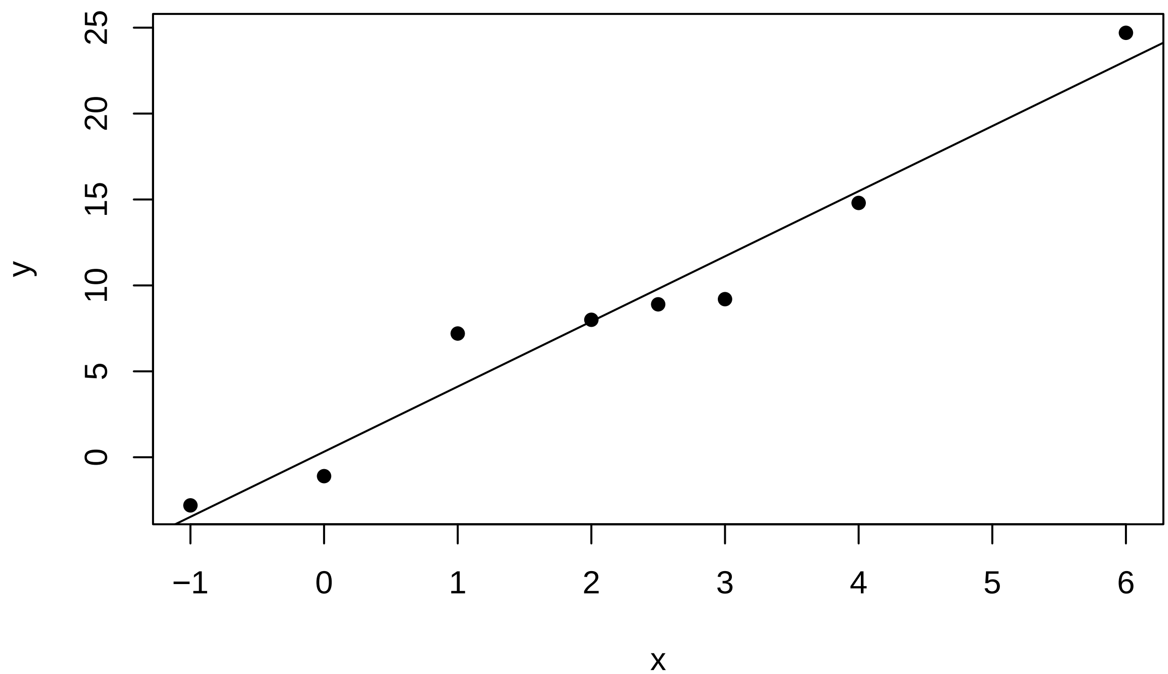
At first glance, this may seem to be a reasonable fit but we recall from the earlier discussion that the context of the problem means that it is a stupid model. This gives us a warning that we should always understand the background to any statistical analysis to ensure that the results are meaningful. It is bad practice to apply methods without thought.

Exercise 1.2 There is no prepared file containing these data and hence we create R vectors and then follow the *usual* steps in the regression modelling.

```
x = c(-1, 0, 1, 2, 2.5, 3, 4, 6)
y = c(-2.8, -1.1, 7.2, 8.0, 8.9, 9.2, 14.8, 24.7)

plot(x, y, pch=16)

my.fit = lm(y ~ x)
abline(my.fit)
```



The fit seems appropriate, but we have learnt from the previous question that we should know, at least a little, about the context of a problem before being satisfied – there is no background here and so we proceed with caution.

Predicting the y-value when $x = 5$ seems fine as there are data points above and below, whereas $x = 10$ is well beyond the highest data point and hence is likely to be less reliable. The first of these cases is referred to as *interpolation* (within the data range) and the second as *extrapolation* (outside the data range).

We can use R to predict values using the function `predict` – see the help page `?predict.lm` for details of how this works for linear models fitted using `lm` – to give the output:

```
predict(my.fit, newdata = data.frame(x=c(5,10)))
```

```
##        1        2
## 19.27185 38.22181
```

Here, having to use `data.fame` is cumbersome, but the same approach is necessary when dealing with multiple explanatory variables and hence please try to understand it in this simple case.

Exercise 1.3 For this theoretical question, we start from

$$RSS(\alpha, \beta) = \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2 .$$

We wish to find the parameter values which best fit the data, that is which minimize the RSS. This can be started using the linear algebra *trick* of completing the squares, but here we will use the more general approach of (partial) differentiation as follows:

$$\frac{\partial RSS}{\partial \alpha} = -2\sum\left(y_i - (\alpha + \beta x_i)\right) = -2\left(\sum y_i - n\alpha - \beta\sum x_i\right).$$

Also,

$$\frac{\partial RSS}{\partial \beta} = -2\sum x_i\left(y_i - (\alpha + \beta x_i)\right) = -2\left(\sum x_i y_i - \alpha\sum x_i - \beta\sum_{i=1}^{n} x_i^2\right).$$

Simultaneously setting these equations equal to zero and solving for α and β gives the least squares estimates, $\hat{\alpha}$ and $\hat{\beta}$. Straight away from the first, after cancelling the -2 and diving by n, we get

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

With the second, substitute for $\hat{\alpha}$, and cancel the -2, to give

$$\sum_{i=1}^{n} x_i y_i - \left(\bar{y} - \hat{\beta}\bar{x}\right)\sum_{i=1}^{n} x_i - \hat{\beta}\sum_{i=1}^{n} x_i^2 = 0.$$

Which can be re-arranged to give

$$\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \hat{\beta}\left(\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right) = 0.$$

Then, using $\sum x_i = n\bar{x}$ and the definitions of s_{xy} and s_x^2, gives the result

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}.$$

Finally, we should check that this is indeed a minimum. For this we can apply the *second derivative test*[1], that is we need to show that the Hessian determinant is positive, where the Jacobian is the matrix of second derivatives, and that its diagonal elements are positive. Here, we have

$$H = \begin{bmatrix} 2n & 2\sum x_i \\ 2\sum x_i & 2\sum x_i^2 \end{bmatrix}$$

and hence $\det(H) = 4(n\sum x_i^2 - (\sum x_i)^2)$ which is proportional to the variance and hence is positive. Also, $2n > 0$ (and $2\sum x_i^2 > 0$). Therefore we have identified a valid minimum.

Next we consider unbiasedness, which requires $E[\hat{\alpha}] = \alpha$ and $E[\hat{\beta}] = \beta$.

As a preliminary, recalling that $y_i = \alpha + \beta x_i + \epsilon_i$ and hence $E[y_i] = \alpha + \beta x_i$, we note that

$$E[\bar{y}] = \frac{1}{n}\sum E[y_i] = \frac{1}{n}\sum(\alpha + \beta x_i) = \alpha + \beta\bar{x}.$$

Firstly, for unbiasedness, starting with $\hat{\beta}$, as this does not involve α,

$$E\left[\hat{\beta}\right] = E\left[\frac{s_{xy}}{s_x^2}\right] = \frac{1}{s_x^2}E[s_{xy}]$$

Hence, we must considered $E[s_{xy}]$, and in particular

$$E[s_{xy}] = E\left[\frac{1}{(n-1)}\sum(x_i - \bar{x})(y_i - \bar{y})\right] = \frac{1}{(n-1)}\sum(x_i - \bar{x})E[(y_i - \bar{y})]$$

[1] https://mathworld.wolfram.com/SecondDerivativeTest.html

and therefore

$$E[s_{xy}] = \frac{1}{(n-1)} \sum (x_i - \bar{x})(\alpha + \beta x_i - (\alpha + \beta \bar{x})) = \beta s_x^2$$

which gives

$$E\left[\hat{\beta}\right] = \frac{1}{s_x^2} E[s_{xy}] = \beta \frac{s_x^2}{s_x^2} = \beta$$

meaning that $\hat{\beta}$ is unbaised for β.

Next consider $\hat{\alpha}$,

$$E[\hat{\alpha}] = E[\bar{y} - \hat{\beta}\bar{x}] = E[\bar{y}] - E[\hat{\beta}]\bar{x}$$

Now, using the results above, we have

$$E[\hat{\alpha}] = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha$$

as required and hence $\hat{\alpha}$ is unbiased for α.

Finally, let us consider $\hat{\sigma}^2$, as an estimator of σ^2, with definition

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Descriptive arguments are:

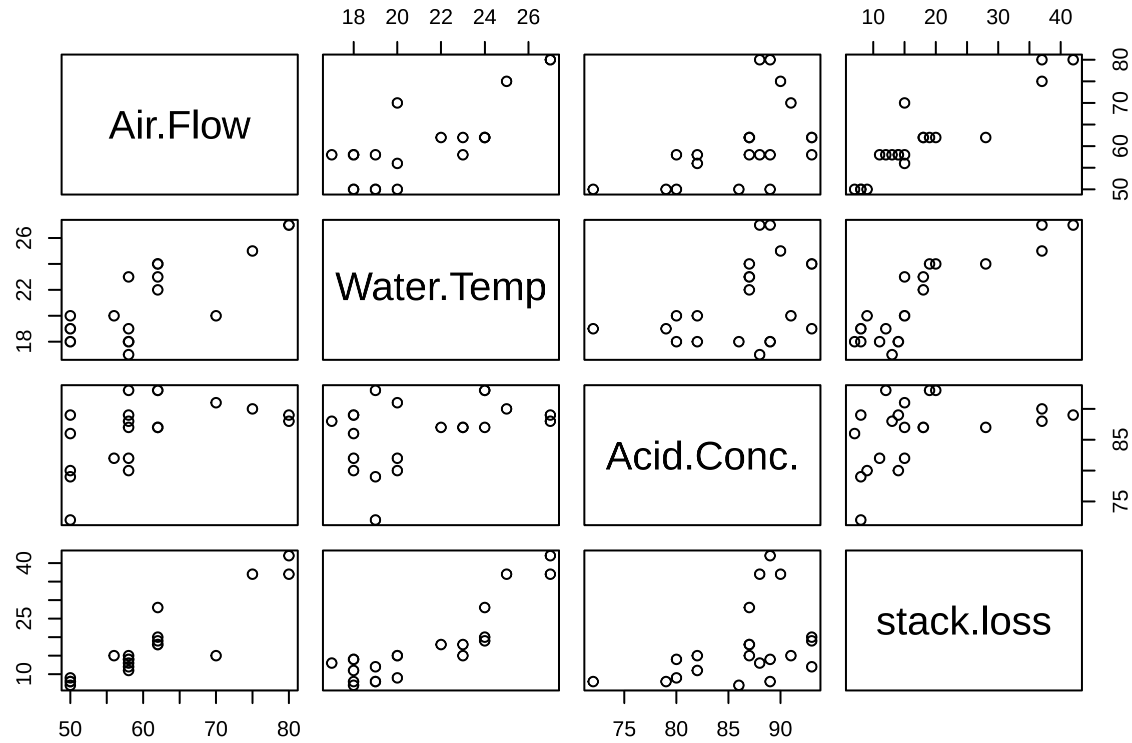
(1) By definition, $\sigma^2 = \frac{1}{n} \sum (\epsilon_i - \bar{\epsilon}_i)^2$ and $\hat{\epsilon}_i = y_i - \hat{y}_i$, the residuals are estimates of the true errors, and it can easily be shown that $\bar{\hat{\epsilon}} = 0$. Hence we can calculate $\sum (\hat{\epsilon}_i - \bar{\hat{\epsilon}}_i)^2$, but since we have estimated 2 model parameters, α and β, from the data the fitted model is closer to the data than the true model, which suggests we divide by $(n-2)$, the degrees of freedom, rather than by n or $n-1$.

(2) The estimator $\hat{\sigma}^2$ has a chi-squared random distribution with $n-2$ degrees of freedom, and hence expectation of $n-2$, multiplied by $\sigma^2/(n-2)$. This means that $E[\hat{\sigma}^2] = \sigma^2$, that is it is unbiased for σ^2.

Now a overview of the detailed version:

Exercise 1.4 To plot all pairs of variables we can use the `pairs` function in R:

```r
pairs(stackloss)
```

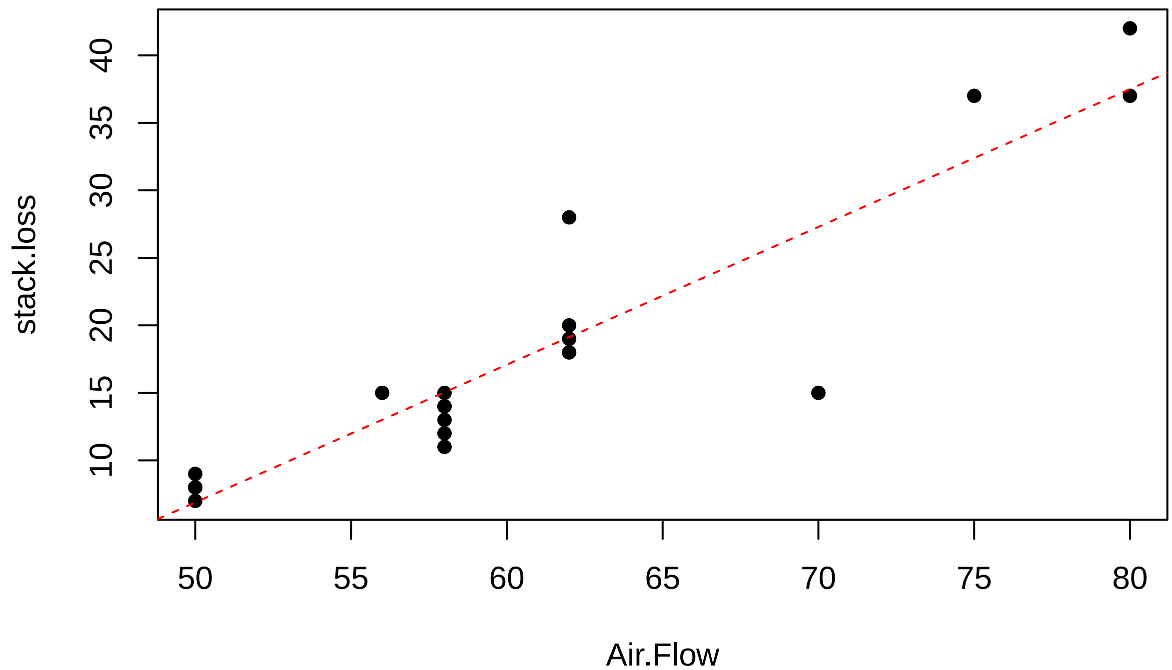


In order for a linear regression to be appropriate the relationship between the response `stack.loss` and an explanatory variable must be well described by a linear equation. This seems to be most true for `Air.Flow`, whereas the other two variables appear to have non-linear relationships with `stack.loss`.

```r
attach(stackloss, warn.conflicts=FALSE)

plot(Air.Flow, stack.loss, pch=16)

myresults = lm(stack.loss ~ Air.Flow)
abline(myresults, lty=2, col="red")
```



The fitted line appears to describe the relationship well but it would be wise to perform a statistical test to confirm.

Exercise 1.5 Following the usual approach: read in the data, use `lm` to fit the model. Then, `plot` and `abline` to plot data and fitted model.

First, read the data and have a look at the first few lines of the data.

```
physics = read.csv("https://rgaykroyd.github.io/MATH3823/Datasets/physics_from_data.csv", header=T)

attach(physics, warn.conflicts=FALSE)

head(physics)
```

```
##                Ball Radius..m. Mass..kg. Density..kg.m. Max.vel...m.s. Max.Re
## 1         Golf ball   0.021963  0.045359     1022.06643          26.63 175000
## 2          Baseball   0.035412  0.141747      762.03752          26.61 283000
## 3       Tennis ball   0.033025  0.056699      375.81325          21.95 218000
## 4        Volleyball   0.105000        NA             NA          22.09 696000
## 5  Blue basketball   0.119366  0.510291       71.62838          24.80 888000
## 6 Green basketball   0.116581  0.453592       68.34291          25.06 877000
```

Next, the plotting and fitting using each of the potential explanatory variables in turn.

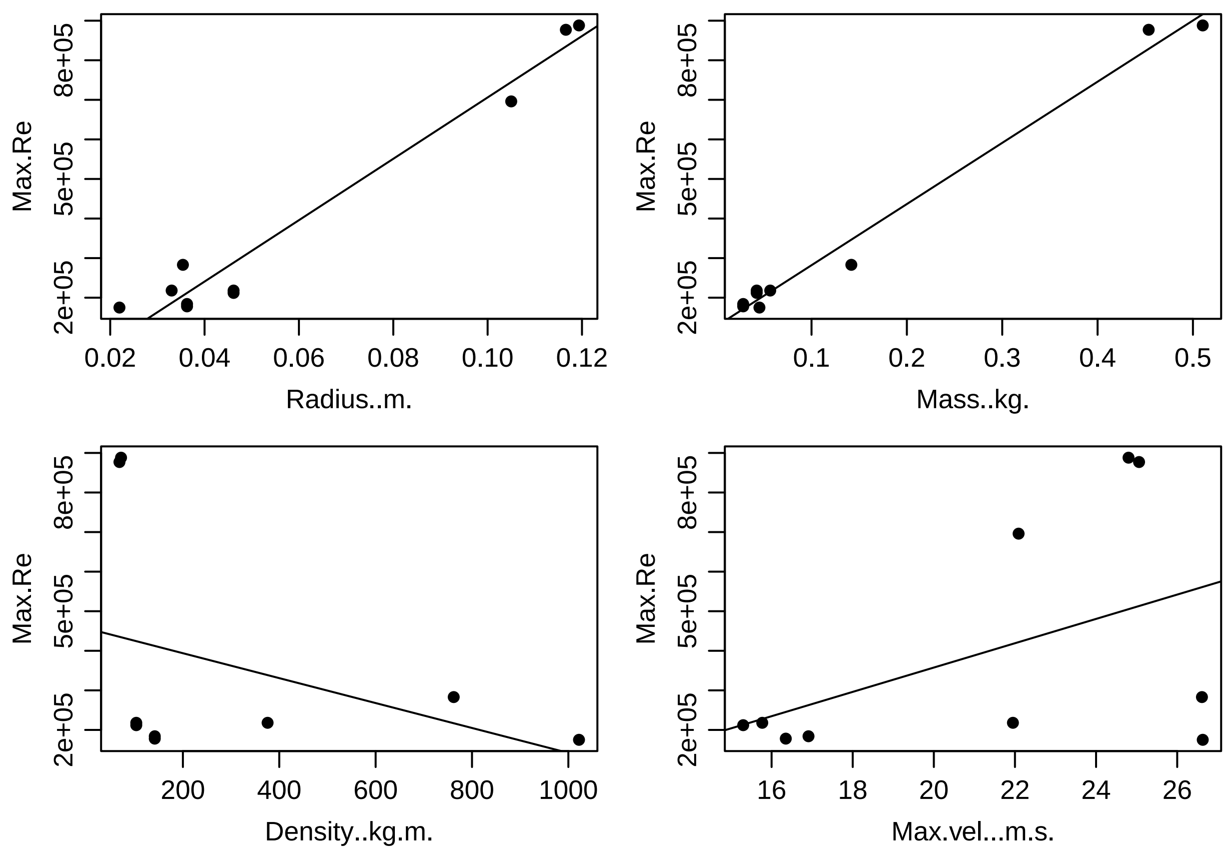
```
par(mfrow=c(2,2), mar=c(3,3,1,1), mgp=c(2,0.7,0) )
plot(Radius..m., Max.Re, pch=16)
```

```
fit1 = lm(Max.Re ~ Radius..m.)
abline(fit1)

plot(Mass..kg., Max.Re, pch=16)
fit2 = lm(Max.Re ~ Mass..kg.)
abline(fit2)

plot(Density..kg.m., Max.Re, pch=16)
fit3 = lm(Max.Re ~ Density..kg.m.)
abline(fit3)

plot(Max.vel...m.s., Max.Re, pch=16)
fit4 = lm(Max.Re ~ Max.vel...m.s.)
abline(fit4)
```



Perhaps none of these appear to show linear relationships. The best are, arguably, `Radius` and `Mass`, but there is a lack of data for medium sized balls with moderate mass – which makes it harder to estimate the relationship in that region. Perhaps we should seek further information, and possibly more data, before we can make any meaningful conclusions. Sometimes, we have to be prepared to say that the data is not sufficient to answer the question – this is a matter of professional ethics.

End of Solutions to Chapter 1 Exercises