QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Solution to Exercise Sheet 2

1.

(a)
```
> x<- c(3.4,1.8,4.6,2.3,3.1,5.5,0.7,3.0,
2.6,4.3,2.1,1.1,6.1,4.8,3.8)
> y<- c(26.2,17.8,31.3,23.1,27.5,36.0,14.1,
22.3,19.6,31.3,24.0,17.3,43.2,36.4,26.1)
> plot(x,y, main="Plot of Y versus X")
> fire<-lm(y~x)
> summary(fire)
Call:
lm(formula = y ~ x)
Residuals:
    Min      1Q  Median      3Q     Max
-3.4682 -1.4705 -0.1311  1.7915  3.3915
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
x             4.9193     0.3927  12.525 1.25e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.316 on 13 degrees of freedom
Multiple R-squared:  0.9235,Adjusted R-squared:  0.9176
F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08

> plot(x,y, main="Fitted Line Plot")
> (abline(fire))
NULL
> stdres<- rstandard(fire)
> print(stdres)
          1           2           3           4           5
-0.35920557 -0.61671822 -0.73812880  0.68389288  0.88172686
          6           7           8           9          10
-0.64739155  0.18972092 -1.22407120 -1.56097482 -0.05952374
         11          12          13          14          15
 1.54912299  0.77909572  1.49866138  1.16347978 -1.28850409
> plot(x,stdres, main="Std residuals versus x")
> fits<- fitted(fire)
> plot(fits,stdres, main="Std residuals versus fits")
> qqnorm(stdres, main="Q-Q Plot")
> qqline(stdres)
```

The plot of the data, the standardized residuals are shown in Figure 1.1, while the plots of the standardized residuals versus the fits and of the QQ-Norm in Figure 1.2. In particular, the relationship does seem linear. There are a couple of points off the Q-Q line but not enough to cause much concern. The variance seems constant.

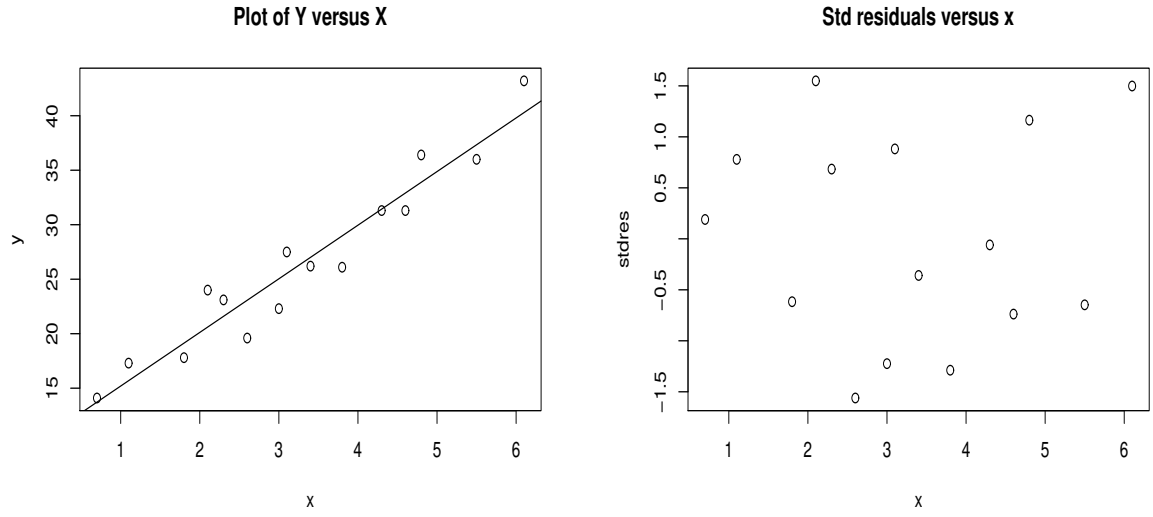

Figure 1.1: Plot of the data versus the fitted regression line (left) and standardized residuals versus x (right).

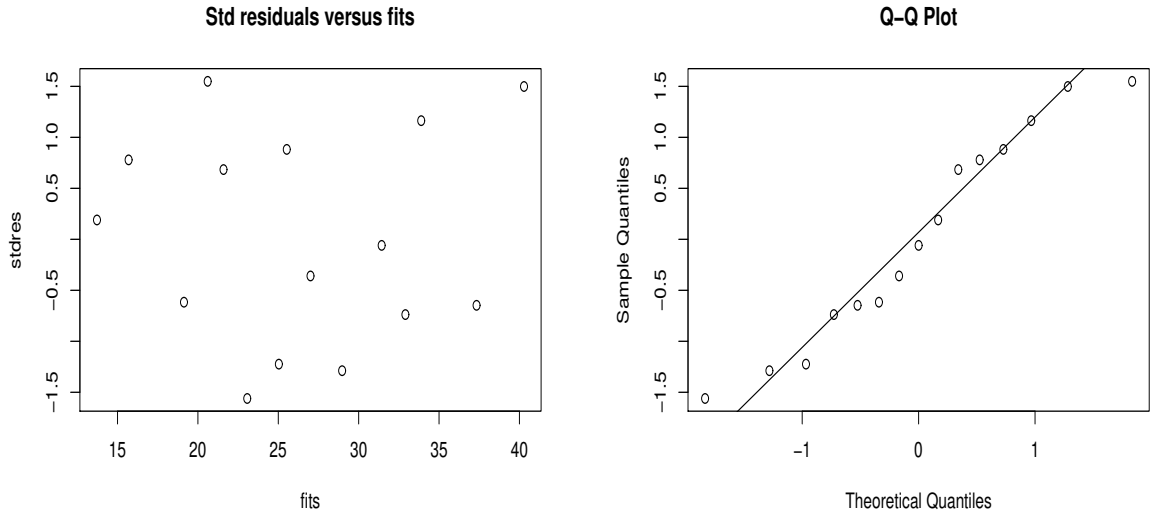

Figure 1.2: Plot of the data versus the standardized residuals versus fitted values (left) and QQ Norm (right).

(b) Report should cover the following points:

- The values of the intercept (10.28) and slope (4.92) are highly significant. A possible interpretation the intercept is the average cost of a fire which hap-

pened next door to a fire station. The fire would do quite a bit of damage before the alarm was raised and it would take some time for the fire fighters to get all their equipment ready.

- The slope is the extra average cost due to being 1km away from the fire station. This represents the extra time it would take to get to the fire and the extra damage done in that time.
- The relationship does seem to be linear and there is no reason to doubt the assumptions of linearity or constant variance. The first and last points on the QQ plot are a little away from the line but there does not seem much reason to doubt the normality assumption.

2. (a) The Least Squares Estimator $\widehat{\beta}$ minimizes the sum of squares of errors, which for the no-intercept model is,

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta x_i)^2.$$

By differentiating the function with respect to β we obtain

$$\frac{\partial S(\beta)}{\partial \beta} = -2 \sum_{i=1}^{n} (Y_i - \beta x_i) x_i = -2 \sum_{i=1}^{n} (Y_i x_i - \beta x_i^2).$$

Hence,

$$\frac{\partial S(\beta)}{\partial \beta} = 0 \text{ is equivalent to } \sum_{i=1}^{n} Y_i x_i = \widehat{\beta} \sum_{i=1}^{n} x_i^2.$$

This gives

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{1}{a} \sum_{i=1}^{n} Y_i x_i, \quad \text{where} \quad a = \sum_{i=1}^{n} x_i^2.$$

To confirm that the function $S(\beta)$ attains minimum at the solution $\widehat{\beta}$, we will check the sign of the second derivative of the function. We have

$$\frac{\partial^2 S(\beta)}{\partial \beta^2} = 2 \sum_{i=1}^{n} x_i^2 > 0 \quad \text{for all } \beta.$$

Hence the second derivative is also positive for $\widehat{\beta}$. It means the function $S(\beta)$ attains minimum at $\widehat{\beta}$.

(b) The estimator is a linear combination of the random variables Y_i, that is

$$\widehat{\beta} = \frac{1}{a} \sum_{i=1}^{n} Y_i x_i = \sum_{i=1}^{n} c_i Y_i,$$

where $c_i = \dfrac{x_i}{a}$.

We assume that the variables Y_i are normally distributed. Hence, $\widehat{\beta}$, as a linear combination of normally distributed random variables, is also normally distributed.

Next we find its expectation and variance. We have

$$\mathrm{E}(\widehat{\beta}) = \mathrm{E}\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i \,\mathrm{E}(Y_i) = \sum_{i=1}^{n} c_i \beta x_i = \beta \sum_{i=1}^{n} c_i x_i = \beta$$

as $\sum_{i=1}^{n} c_i x_i = 1$. The estimator is unbiased.

Also, by the properties of the variance function and by independence of Y_i, we can write

$$\mathrm{Var}(\widehat{\beta}) = \mathrm{Var}\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i^2 \,\mathrm{Var}(Y_i).$$

Furthermore, $\mathrm{Var}(Y_i) = \sigma^2$. This gives

$$\mathrm{Var}(\widehat{\beta}) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i}{a}\right)^2 = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{a^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}.$$

Hence, we conclude that

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right).$$

3. Suppose Y_i is an i.i.d sequences of random variables, such as $Y_i \sim \mathcal{N}(\theta, 1)$ for $i = 1, \ldots, n$. Let us consider the following estimators:

$$\widehat{\theta}_1 = \frac{1}{n-1} \sum_{i=1}^{n} y_i, \qquad \widehat{\theta}_2 = \frac{1}{2}(y_1 + y_n)$$

   (a) For the first estimator we have

$$E[\widehat{\theta}_1] = E\left[\frac{1}{n-1} \sum_{i=1}^{n} y_i\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n} y_i\right] = \frac{1}{n-1} \sum_{i=1}^{n} E\left[y_i\right] =$$
$$= \frac{1}{n-1} \sum_{i=1}^{n} \theta = \frac{n\theta}{n-1}$$

   For the second estimator we have

$$E[\widehat{\theta}_2] = E\left[\frac{1}{2}(y_1 + y_n)\right] = \frac{1}{2} E[y_1 + y_n] = \frac{1}{2}\left(E[y_1] + E[y_n]\right)$$
$$= \frac{1}{2}\left(\theta + \theta\right) = \frac{2\theta}{2} = \theta$$

(b) Regarding the variances of the estimators. We start with the first estimator and due to the i.i.d. assumption of the y_i, we have

$$Var\left(\widehat{\theta}_1\right) = Var\left[\frac{1}{n-1}\sum_{i=1}^{n} y_i\right] = \frac{1}{(n-1)^2} Var\left[\sum_{i=1}^{n} y_i\right] = \frac{1}{(n-1)^2}\sum_{i=1}^{n} Var\left[y_i\right] =$$
$$= \frac{1}{(n-1)^2}\sum_{i=1}^{n} 1 = \frac{n}{(n-1)^2}$$

Analogously for the second estimator we have:

$$Var\left(\widehat{\theta}_2\right) = Var\left[\frac{1}{2}(y_1 + y_n)\right] = \frac{1}{4} Var\left(y_1 + y_n\right) = \frac{1}{4}\left(Var(y_1) + Var(y_n)\right) =$$
$$= \frac{1}{4} 2 = \frac{1}{2}$$

(c) From point (a), we have that the expected values of the $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are

$$E[\widehat{\theta}_1] = \frac{n}{n-1}, \qquad E[\widehat{\theta}_2] = \theta$$

Thus the first estimator differs from θ and it is biased, while the second estimator is unbiased.