QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Solution to Exercise Sheet 3

1. (a) From the definitions we have that

$$\begin{aligned} SS_R &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \left(\hat{\beta}_0 + \widehat{\beta}_1 x_i - \bar{y}\right)^2 = \\ &= \sum_{i=1}^{n} \left(\bar{y} + \widehat{\beta}_1 (x_i - \bar{x}) - \bar{y}\right)^2 = \sum_{i=1}^{n} \left(\widehat{\beta}_1 (x_i - \bar{x})\right)^2 = \\ &= \sum_{i=1}^{n} \left(\frac{S_{xy}}{S_{xx}} (x_i - \bar{x})\right)^2 = \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} = \\ &= \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

(b) From the analysis of variance identity $SS_T = SS_R + SS_E$, we have

$$SS_E = SS_T - SS_R$$

Now, we have $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = S_{yy}$. Substituting the result for SS_R from above, we obtain the required expression.

2. (a) Since $\widehat{\beta}$ is a linear combination of the y_i, which are normally distributed, it is also normally distributed. Therefore

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

Hence we have

$$\frac{\widehat{\beta} - \beta}{\sigma / \sqrt{\sum x_i^2}} \sim \mathcal{N}(0, 1)$$

and so

$$\frac{\widehat{\beta} - \beta}{S / \sqrt{\sum x_i^2}} \sim t_{n-1}$$

where $S = \sum_i e_i^2 / (n-1)$. Note that it has $n-1$ degrees of freedom as there is only one unknown parameter so this is the divisor of S^2.

Then a $100(1-\alpha)\%$ confidence interval for β_1 will be given by

$$\widehat{\beta} \pm t_{n-1}\left(\frac{\alpha}{2}\right) \frac{S}{\sqrt{\sum_i x_i^2}}$$

(b) We have

$$E[e_i] = E[y_i] - x_i E\left[\widehat{\beta}\right] = \beta x_i - x_i \beta = 0$$

Now we move to the variance

$$\begin{aligned} Var[e_i] &= Var\left[y_i - x_i\widehat{\beta}\right] = Var\left[y_i - x_i \sum_{j=1}^{n} a_j y_j\right] = \\ &= Var\left[y_i\right] + x_i^2 \sum_j a_j^2 Var[y_j] - 2x_i \sum_j a_j Cov[y_i, y_j] = \\ &= \sigma^2 + x_i^2 \sigma^2 \frac{\sum_j x_j^2}{(x_i^2)^2} - 2x_i \frac{x_i}{\sum_j x_j^2}\sigma^2 = \\ &= \sigma^2 \left[1 - \frac{x_i^2}{\sum_j x_j^2}\right] \end{aligned}$$

Note that the covariance is zero expect for the case when $j = i$.

3. We install the package

```
> install.packages("datarium")#Packages need to be installed only once
> library(datarium)
> data("marketing", package = "datarium")
> attach(marketing)
```

Then we define the two variables of interest:

```
> y <- marketing$sales
> x <- marketing$youtube
```

(a) We run a simple linear regression and we plot the data

```
> plot(x, y)
> mody <- lm(y~x)
```

In Figure 1.1, we have the data plotted and the data against the fitted regression line. At a first look, we can see a positive relation between the sales and the advertising on Youtube. Thus an increasing impact of the advertising media such as Youtube on the sales is positive.

(b) We run a linear regression model on the data by using the following commands

```
> mody <- lm(y~x)
> summary(mody)

Call:
lm(formula = y ~ x)
```

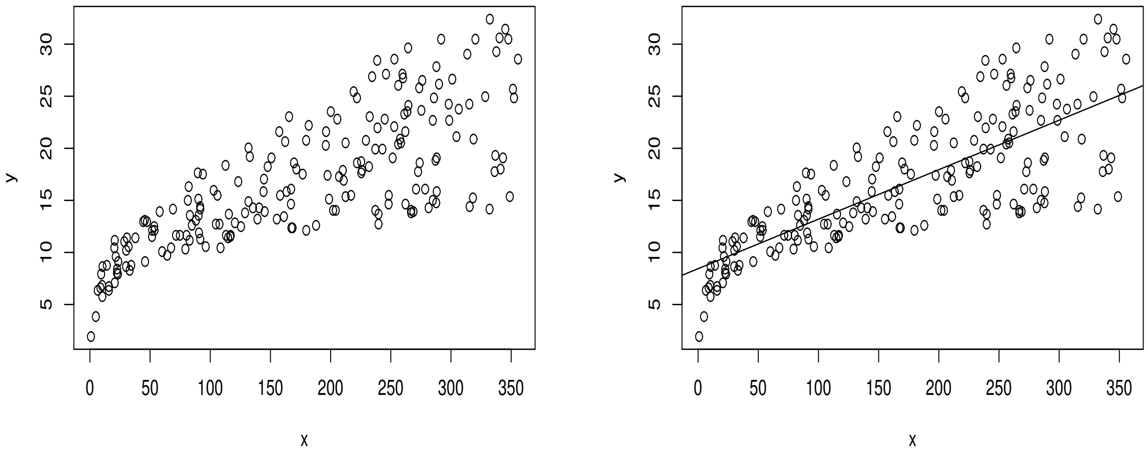

Figure 1.1: Plot of the data (left) and of the data versus the fitted regression line (right).

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.0632  -2.3454  -0.2295   2.4805   8.6548

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.439112   0.549412   15.36   <2e-16 ***
x           0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.91 on 198 degrees of freedom
Multiple R-squared:  0.6119,Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16

> anova(mody)
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1 4773.1  4773.1  312.14 < 2.2e-16 ***
Residuals 198 3027.6    15.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can see that both the parameters (the intercept and the slope) are significant and that both are positive. Although the slope parameter seems to be really close to zero. Regarding the R^2, we can see that it is greater than 50%, in particular is aroung 60%, thus explaining an high variability of the model. In the future, we will try to include other variables in the analysis to look at the best model.

(c) In conclusion, we can carry out a Shapiro-Wilk test on the standardized residuals or looking directly at the plots.

```
> qqnorm(stdres1, main="Q-Q Plot")
> qqline(stdres1)
> shapiro.test(stdres1)

Shapiro-Wilk normality test

data:  stdres1
W = 0.99052, p-value = 0.2126
```

In this scenario, the p-value of the Shapiro-Wilk test is higher than 0.05. Thus the assumption of normality of the standardized residuals is correct. This is also confirmed by the QQ plot reported below

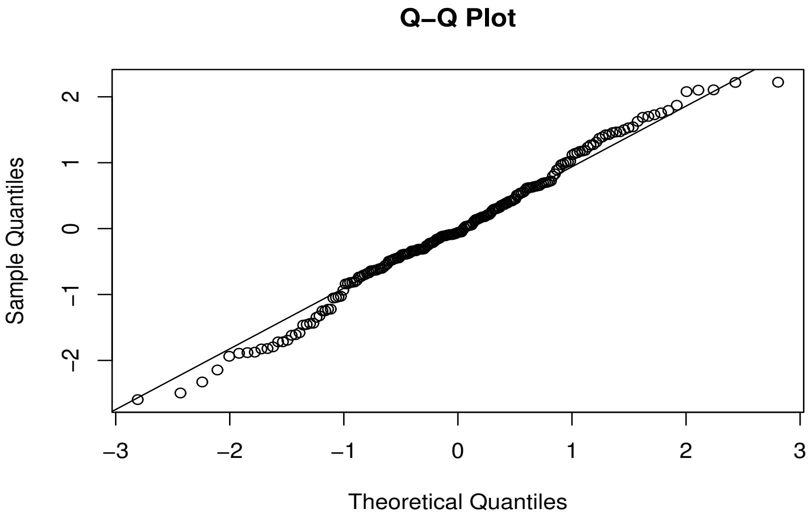

Figure 1.2: QQ Plot of the standardized residuals.