# Chapter 3 Exercises

Adrien Osakwe

## Conceptual

### 1)

Table 3.4 shows basic stats from the linear model for the estimation of the number of units sold given the advertising budgets for (the predictors): *TV, radio and newspaper*. The null hypotheses the p-values correspond to are that there is no link between the budgets of each form of advertisement and the number of units sold ( == the coefficients of the predictors are equal to 0).

Based on this relationship, the p-values indicate that the null hypotheses for both the TV and radio budgets' relationship to sales can be rejected. However, the null hypothesis for newspaper must be accepted as its p-value is much larger than 0.05. We would therefore conclude that TV and radio budgets are two predictors worth using to estimate sales.

### 2)

The KNN classifier will be used to estimate a qualitative response, (a category). This functions as a conditional probability where we find the k nearest neighbors to our observation and determine what proportion of the neighbors are in a given class. i.e: If k-1/k of our observation's neig- hbors are in class Green, then we would classify our observation as green with a conditional probability of k-1/k.

The KNN regression method is used to estimate quantitative responses, where the estimation of f(x) will be the average response of the k nearest-neighbors . i.e: if the average of our observation's neighborhood response is 8, then our regression model with k neighbors will estimate the response as 8.

### 3)

The linear model can be written as follows:

y = 50 + 20GPA + 0.07IQ + 35Level + 0.01(GPA x IQ) -10(GPA x Level)

We can simplify the model as the following: Level = 1: College graduate y = 85 + 10GPA + 0.07IQ + 0.01(GPA x IQ) Level = 0: High School Graduate y = 50 + 20GPA + 0.07IQ +0.01(GPA X IQ)

–> These models differ at two coefficients: the intercept is 35 units larger for a college graduate and the GPA coefficient is twice as large for the high school graduate

A) Based on the different coefficients in the two instances in the model, the student with the higher salary is dependent on GPA. Because the intercept is larger for the college instance, the college graduate will have a higher salary for low GPAs. However, when GPA is 3.5, both students have the same salary (10GPA =35, which equals the difference in the intercept). Therefore, GPA > 3.5 will lead to a higher salary for a high school graduate.

   The correct answer is therefore (iii)

B) The salary will be 137100 dollars. What did they study? How is it so high?!! Who knows…

C) This is false. Evidence for an interaction effect would be based on the t-statistic and resulting p-value for the interaction's coefficient to know if we can accept the null hypothesis that there is no interaction effect.

### 4)

a. There is not enough information to know as the cubic regression may enable the model to overfit more to any noise in the data. On the other hand, since the linear model is closer to the truth, it may also have a lower RSS.
b. We expect it to be lower for the linear regression as its assumption are closest to the true trend, reducing reducible error. The cubic model is more likely to have overfit, increasing reducible error.
c. There is not enough information as the true trend may not be cubic either, in which case RSS will also be larger for the cubic regression. If anything, the size of the RSS for either may help give a clue of what kind of non-linear relationship is the truth.
d. As in c, little can be said here without knowing what the true relationship is. Therefore, it is most likely higher in the regression furthest from the truth.

### 5)

Note: I really need to review LaTex…

y(i) = x(i)sum(x(j)y(j))/sum(xˆ2(i'))

y(i) = sum(x(i)x(j)y(j))/sum(xˆ2(i'))

y(i) = sum( ((x(i)x(j))/sum(xˆ2(i')))y(j))

y(i) = sum(a(j)y(j))

a(j) = x(i)(j)/sum(xˆ2(i'))

### 6)

y = b0 + b1x

x = x(ˆ)

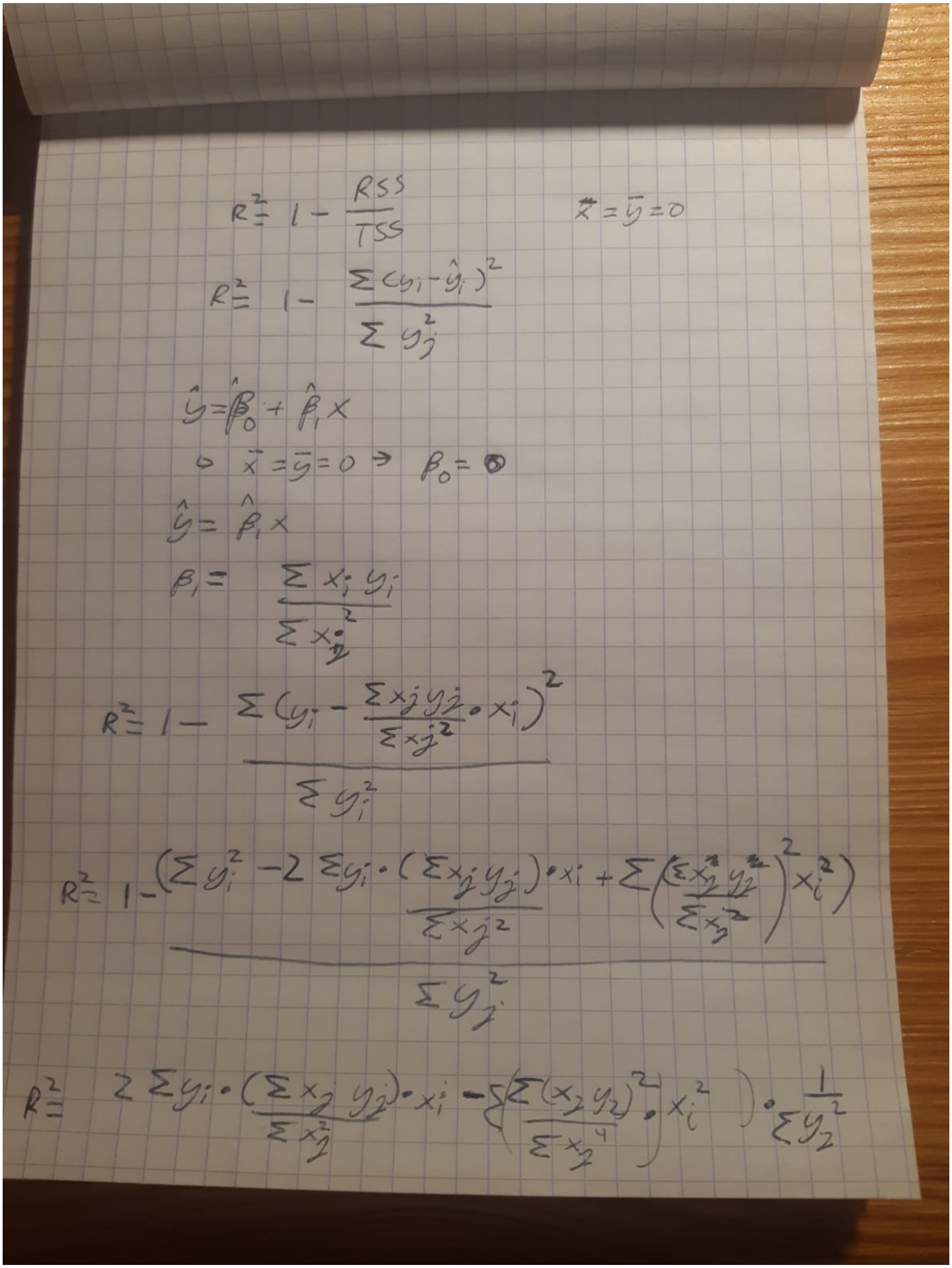
y = b0 + b1x(ˆ)

b0 = y(ˆ) - b1x(ˆ)

y = y(ˆ)

### 7)

```
setwd(getwd())
```



$$R^2 = 1 - \frac{RSS}{TSS} \qquad \bar{x} = \bar{y} = 0$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum y_j^2}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\circ \ \bar{x} = \bar{y} = 0 \Rightarrow \beta_0 = 0$$

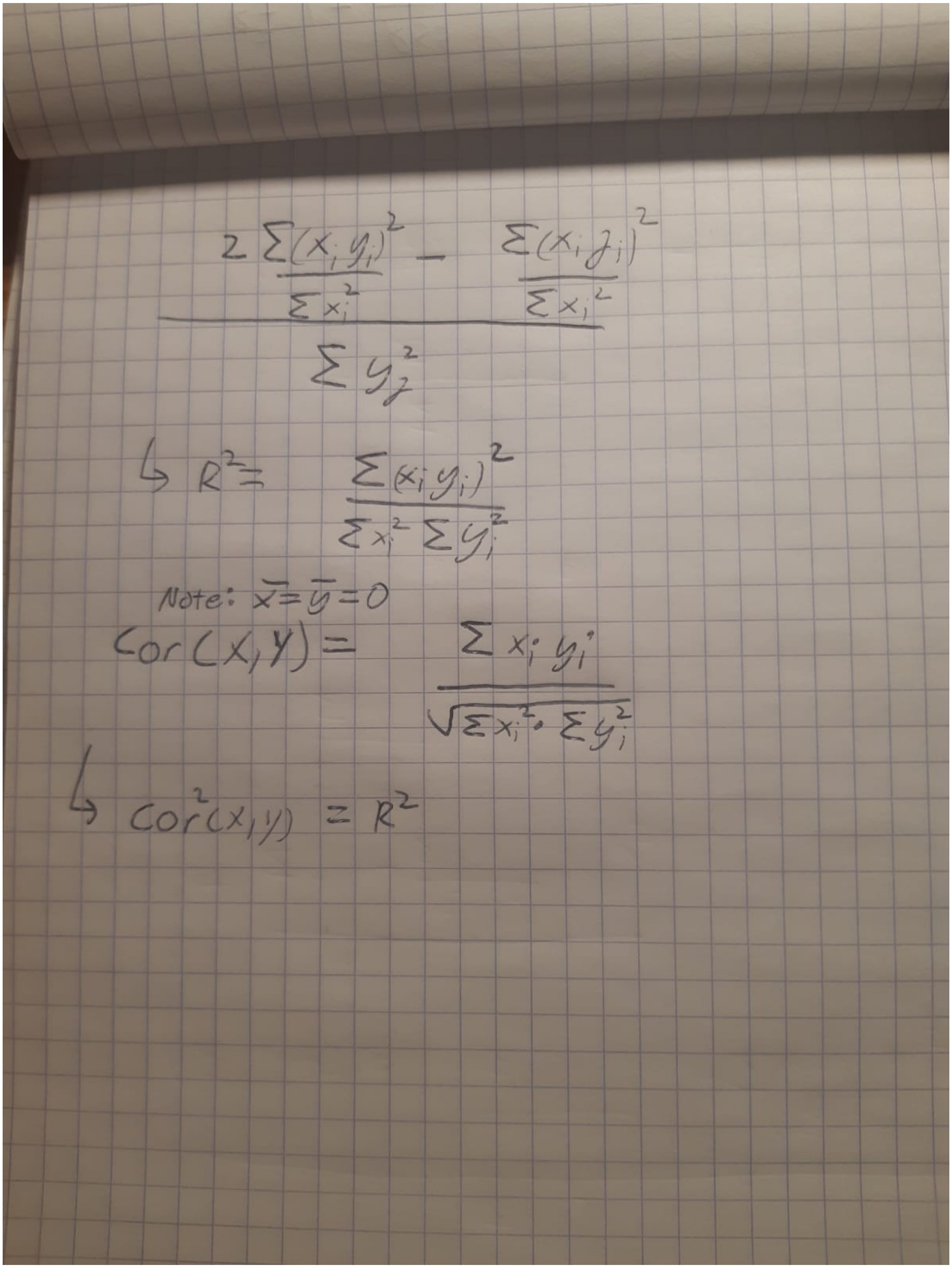
$$\hat{y} = \hat{\beta}_1 x$$

$$\beta_1 = \frac{\sum x_i y_i}{\sum x_j^2}$$

$$R^2 = 1 - \frac{\sum \left(y_i - \frac{\sum x_j y_j}{\sum x_j^2} \cdot x_i\right)^2}{\sum y_i^2}$$

$$R^2 = 1 - \frac{\left(\sum y_i^2 - 2 \sum y_i \cdot \frac{(\sum x_j y_j)}{\sum x_j^2} \cdot x_i + \sum \left(\frac{\sum x_j^2 y_j^2}{\sum x_j^2}\right)^2 x_i^2\right)}{\sum y_j^2}$$

$$R^2 = \left(2 \sum y_i \cdot \frac{(\sum x_j y_j)}{\sum x_j^2} \cdot x_i - \sum \left(\frac{\sum (x_j y_j)^2}{\sum x_j^4} \cdot x_i^2\right)\right) \cdot \frac{1}{\sum y_j^2}$$



$$\frac{2\frac{\sum(x_i y_i)^2}{\sum x_i^2} - \frac{\sum(x_i y_i)^2}{\sum x_i^2}}{\sum y_j^2}$$

↳ $R^2 = \dfrac{\sum(x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$

Note: $\bar{x} = \bar{y} = 0$

$$Cor(x,y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}}$$

↳ $cor^2(x,y) = R^2$

## Applied