Stat 6021: Homework Set 1, Name: H. Diana McSpadden, UID: hdm5s

’’’

Name: H. Diana McSpadden

UID: hdm5s

Assignment: Homework Set 1

Question 1

(R required) We will use the dataset “copier.txt” for this question.

The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from **45 recent calls on users** to perform routine preventive maintenance service; for each call, Serviced is the number of copiers serviced and Minutes is the total number of minutes spent by the service person.

Knowns n = 45 Serviced == number of copy machines serviced on a call Minutes == number minutes spent by service person on the call

First, get the data

’’’

```
#getwd()
data <- read.table('copier.txt', sep='\t', header = TRUE)
head(data,5)
```

```
##   Minutes Serviced
## 1      20        2
## 2      60        4
## 3      46        3
## 4      41        2
## 5      12        1
```

’’’ ## 1a: What is the response variable in this analysis? What is predictor in this analysis?

Answer The response variable is **Minutes**. The predictor variable is **Serviced**. In other words, we believe the number of copiers serviced predicts the length, in minutes, of the service call.

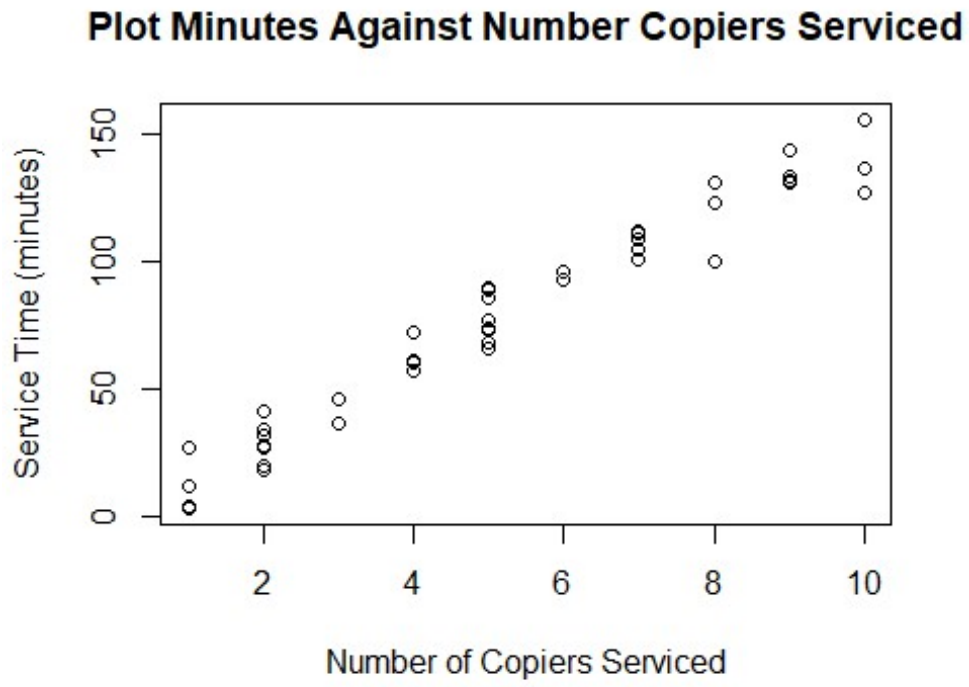
1b: Produce a scatterplot of the two variables.

How would you describe the relationship between the number of copiers serviced and the time spent by the service person?

```
'''
```

```
#attach the data frame
#attach(data)

#use plot(...) to create a scatterplot
plot(x=data$Serviced, y=data$Minutes, main='Plot Minutes Against Number Copie
rs Serviced', xlab = 'Number of Copiers Serviced', ylab = 'Service Time (minu
tes)')
```




```
'''
```

Relationship There is a strong linear relationship with positive correlation. I do not see any clusters or outliers.

1c: Use the lm() function to fit a linear regression for the two variables.

Where are the values of B1, B0, R2, and sigma-hat2 for this linear regression? ”’

```
lmodel = lm(data)
summary(lmodel)
```

```
##
## Call:
## lm(formula = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## Serviced     15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

```
lmodel_sigma = sigma(lmodel)
lmodel_sigma
```

```
## [1] 8.913508
```

”’

B1: 15.0352

B0: -0.5802

R2: 0.9575

sigma-hat2: 8.914

1d: Interpret the values of B1, B0 contextually.

Does the value of B0 make sense in this context?

Answer B1 Interpretation: The change in service call time, in minutes, will be +15.0352 minutes for each additional copier serviced, after the first copier.

B0 Interpretation: B0 tells us that a service call for 0 minutes would take negative 0.5802 minutes, which does not makes sense, and a service call for 0 copiers also does not make sense.

1e: Use the anova() function to produce the ANOVA table for this linear regression.

What is the value of the ANOVA F statistic? What null and alternative hypotheses are being tested here? What is a relevant conclusion based on this ANOVA F statistic? ”’

```
anova(lmodel)
```

```
## Analysis of Variance Table
##
## Response: Minutes
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Serviced   1  76960   76960  968.66 < 2.2e-16 ***
## Residuals 43   3416      79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

”’

ANOVA F STAT: 968.66

H0: B1 == 0, or that the slope of the simple linear regression equation equals 0.

Ha: B1 != 0, or that the slope of the simple linear regression equation does not equal 0.

The F statistic is large (given the data set) are tells us that we should reject the null hypothesis and that our linear model predicts service call minutes based on number of copiers serviced better than using the average number of service call minutes, or any other number within the range of service minutes.

Question 2. (Do not use R in this question)

Suppose that for n = 6 students, we want to predict their scores on the second quiz using scores from the first quiz. The estimated regression line is:

y-hat = 20 + 0.8x

2.a For each individual observation, calculate its predicted score on the second quiz y-hat-i and the residual e-i.

y-hat-i = 20 + (0.8)(x-i)

e-i = (y-i) - (y-hat)

element	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6

x-i	70	75	80	80	85	90
y-i	75	82	80	86	90	91
y-hat-i	76	80	84	84	88	92
e-i	-1	2	-4	2	2	-1

2.b Complete the ANOVA table for this dataset below.

Note: Cells with *** in them are typically left

SS-R = SUM[((y-i)- (y-bar))^2]

y-bar = (75+82+80+86+90+91) / 6 == **84**

SS-res = SUM[((y-i) - (y-hat))^2]

	DF	SS	MS	F-stat	p-value
Regression	1	190	190 / 1 == **190**	(190 / 7.5) == **25.3333**	0.0099
Residual	n-2	30	30 / (6-2) == **7.5**	***	***
Total	n-1	190 + 30 == **220**	***	***	***

2.c Calculate the sample estimate of the variance sigma^2 for the regression model.

sigma^2 == SS-res / (n-2) == 30 / (6-2) == **7.5**

2.d What is the value of R2 here?

R2 == SS-R / SS-T == 190 / 220 == **.8636**

2.e Carry out the ANOVA F test.

From Table A.4 of our textbook F-.01,1,4 == 21.20

Is F0 > F-.01,1,4 ?

Answer: Yes, 25.3333 > 21.20

What is an appropriate conclusion?

The appropriate conclusion is that we reject the null hypothesis, that B1 = 0, and we state that there is a linear relationship between the first quiz score and the second quiz score.

Question 3.

Please see images 3a, 3bc, and 3d - please see below interpretations

Also, give a one-sentence interpretation of what the equalities (6) to (9) mean.

Interpretation 6: the sum of the residuals is 0.

Interpretation 7: The sum of the actual y values is equal to the sum of the predicted y values.

Interpretation 8: The sum of the predictor variable, x, multiplied/weighted by the residual error is 0.

Interpretation 9: The sum of the predicted values, y-hat, multiplied/weighted by the residual error is 0.

Image 3a

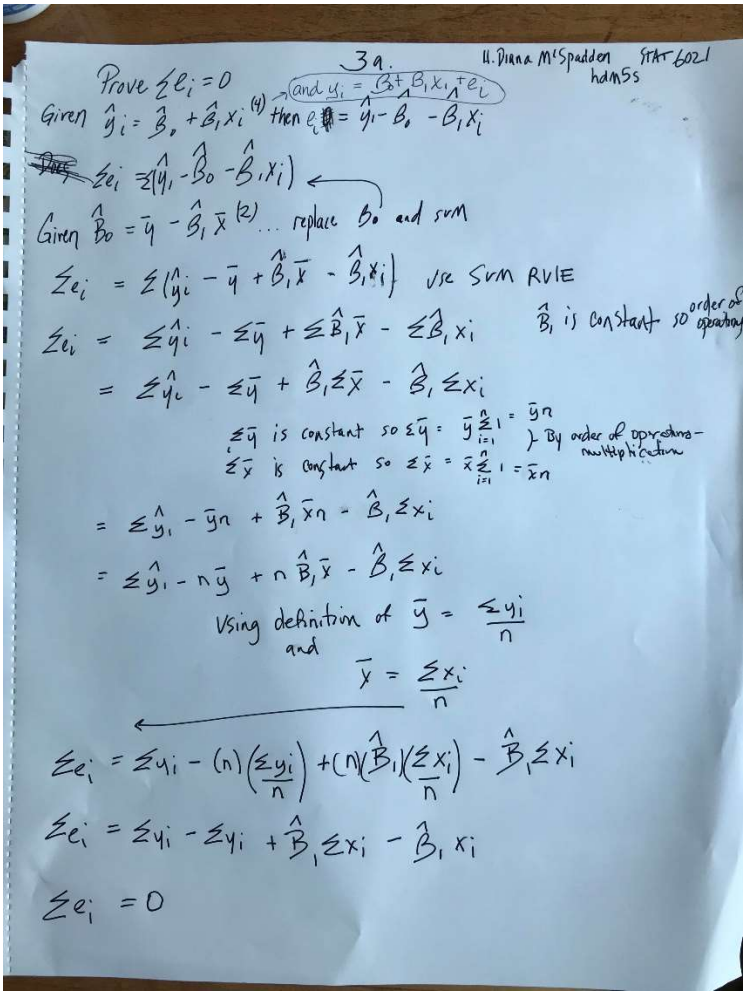


Image 3b

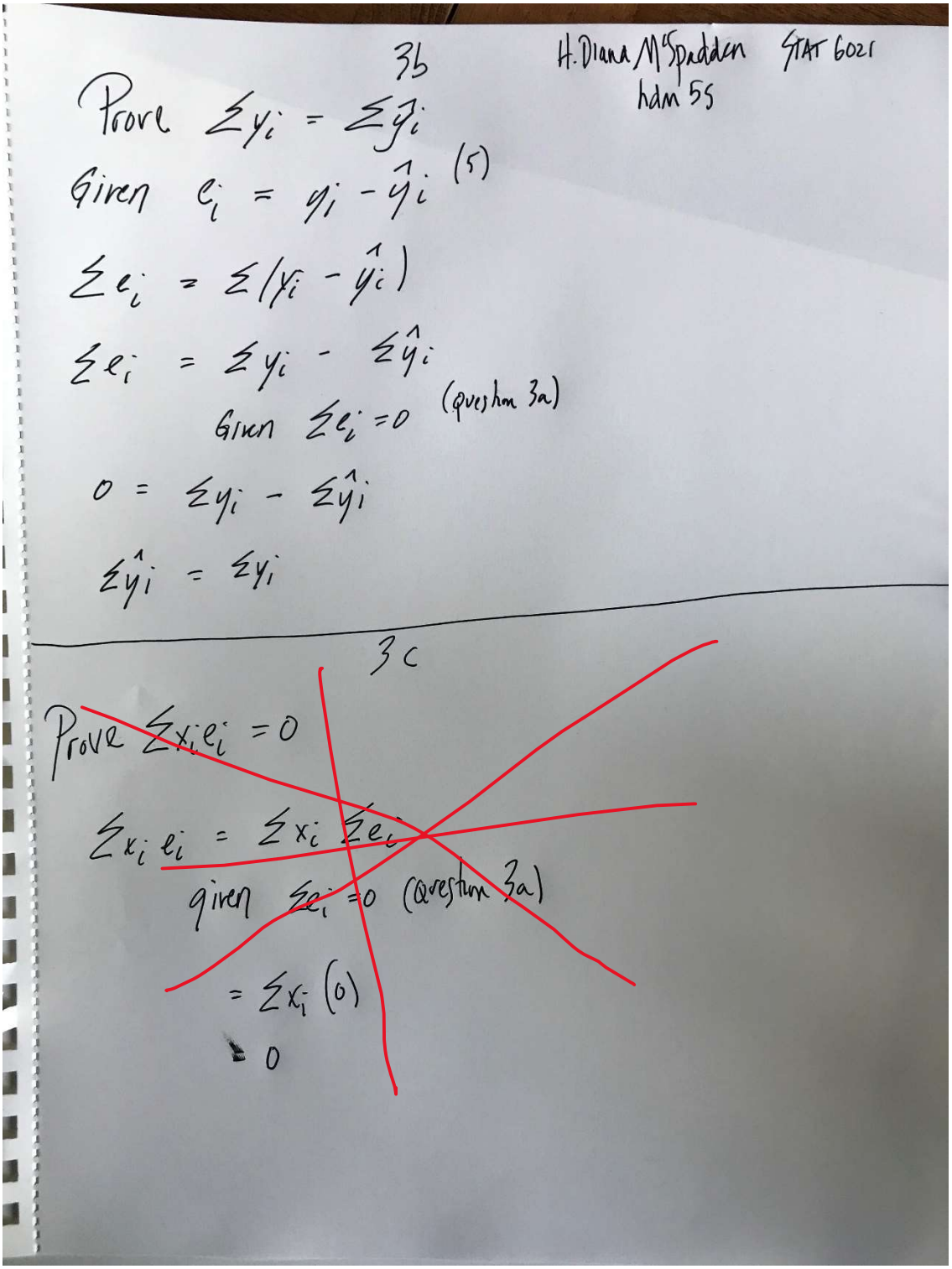
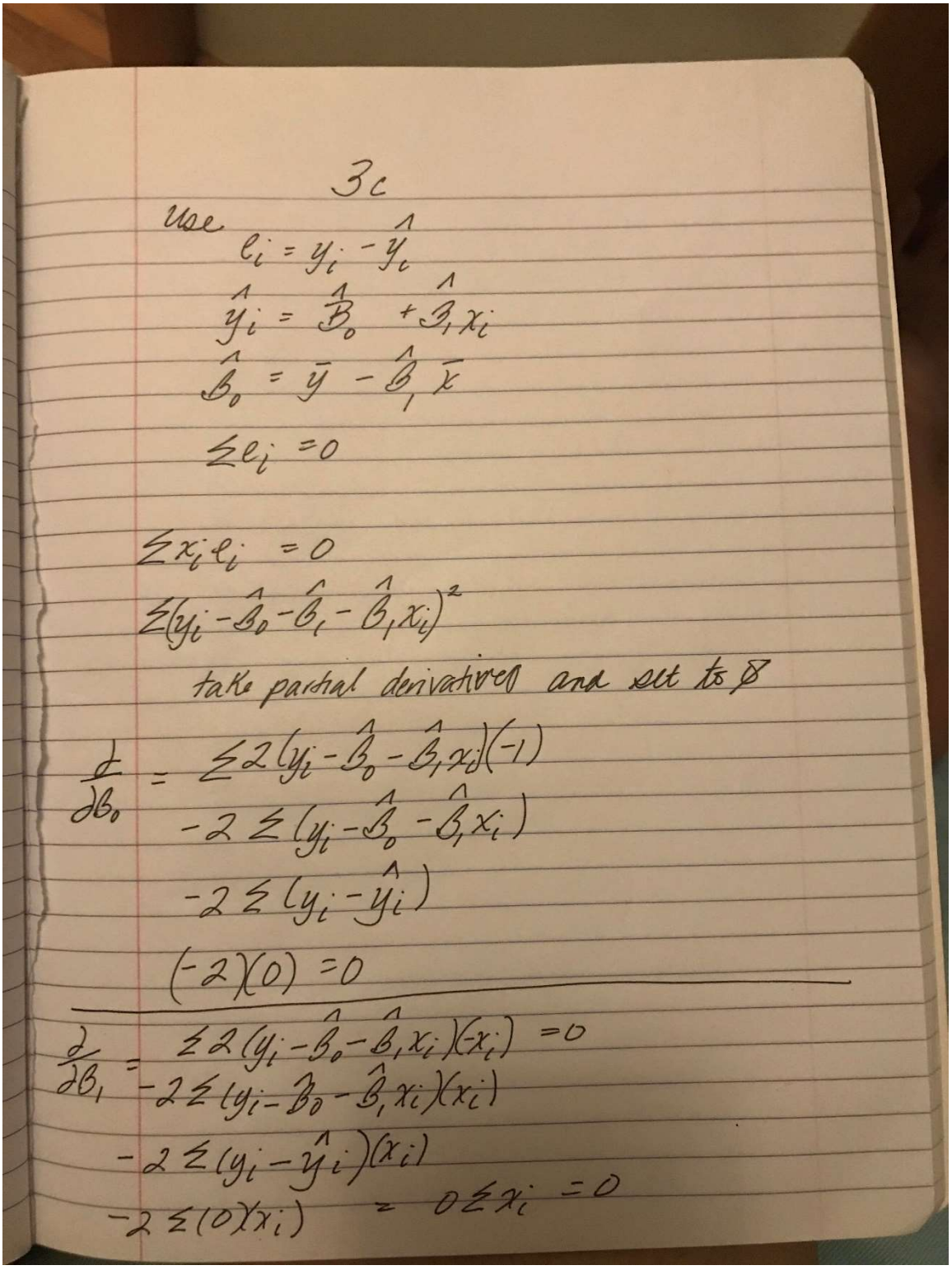

3b
H. Diana McSpadden STAT 6021
hdm 55
Prove $\sum y_i = \sum \hat{y}_i$
Given $e_i = y_i - \hat{y}_i$ (5)
$\sum e_i = \sum (y_i - \hat{y}_i)$
$\sum e_i = \sum y_i - \sum \hat{y}_i$
Given $\sum e_i = 0$ (question 3a)
$0 = \sum y_i - \sum \hat{y}_i$
$\sum \hat{y}_i = \sum y_i$
3c
Prove $\sum x_i e_i = 0$
$\sum x_i e_i = \sum x_i \sum e_i$
given $\sum e_i = 0$ (question 3a)
$= \sum x_i (0)$
$= 0$

Image 3c



3c

use $e_i = y_i - \hat{y}_i$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\sum e_i = 0$

$\sum x_i e_i = 0$

$\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

take partial derivatives and set to $\emptyset$

$\frac{d}{d\beta_0} = \sum 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1)$

$-2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

$-2 \sum (y_i - \hat{y}_i)$

$(-2)(0) = 0$

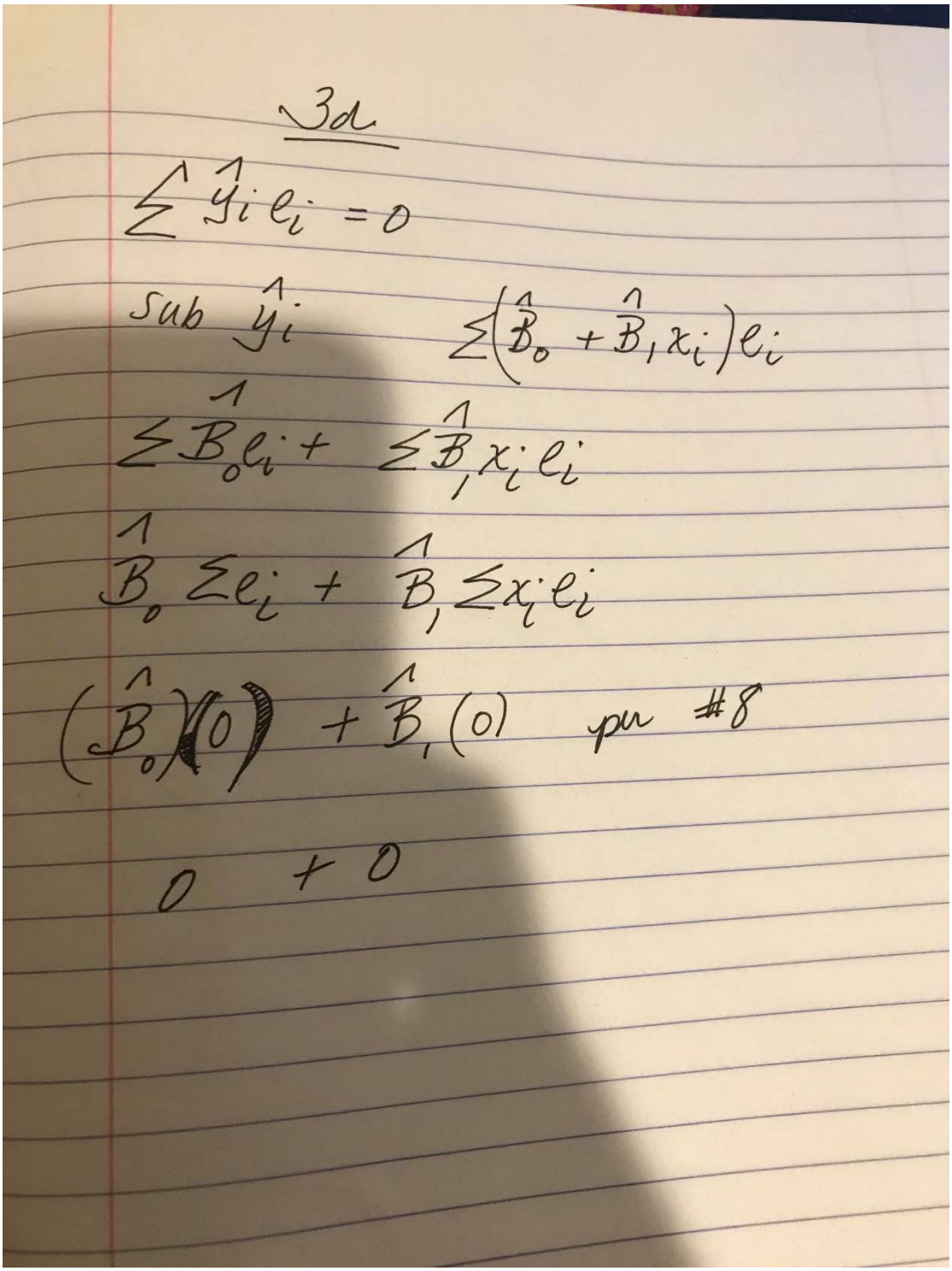
$\frac{d}{d\beta_1} = \sum 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$

$-2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i)$

$-2 \sum (y_i - \hat{y}_i)(x_i)$

$-2 \sum (0) x_i) = 0 \sum x_i = 0$

Image 3d


3d
$\sum \hat{y}_i e_i = 0$
sub $\hat{y}_i$
$\sum (\hat{B}_0 + \hat{B}_1 x_i) e_i$
$\sum \hat{B}_0 e_i + \sum \hat{B}_1 x_i e_i$
$\hat{B}_0 \sum e_i + \hat{B}_1 \sum x_i e_i$
$(\hat{B}_0)(0) + \hat{B}_1 (0)$ pr #8
$0 + 0$