(1)(a) $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$ $\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$$= \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right) = \sum_{i=1}^{n} \left[\left(y_i - \left((\bar{y} - \hat{\beta}_1 \bar{x}_1) + \hat{\beta}_1 x_i\right)\right] = \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\right]$$

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

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

$$\sum_{i=1}^{n} (y_i - \bar{x}) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} = n\bar{y} - \sum_{i=1}^{n} \bar{y} = n\bar{y} - n\bar{y} = 0 = 0 - 0 = 0$$

(b) $$\sum_{i=1}^{n} x_i(y_i - \hat{y}_i) = 0 = \sum_{i=1}^{n} \left[x_i\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)\right] = \sum_{i=1}^{n} \left[x_i\left(y_i - \left((\bar{y} - \hat{\beta}_1 \bar{x}_1) + \hat{\beta}_1 x_i\right)\right]$$

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

$n\bar{y} - \sum_{i=1}^{n} \bar{y} = n\bar{y} - n\bar{y} = 0$

$0 \times x_i = 0x_i = 0$

$0 \times y_i = 0y_i = 0$

We can also say 0 multiplied by anything is still 0.

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

(c) $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$

Least square estimators β_0 must satisfy $\frac{\delta s}{\delta s} = 0$

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

$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$ $\quad \hat{\beta}_0 = y - \hat{\beta}_1 x$

(d) $\sum_{i=1}^{n} \hat{y} e_i = 0$, the previous question (c) proves this. However, we can also do

$\sum_{i=1}^{n} x_i (y_i - \bar{y} + \hat{\beta}_1 x - \hat{\beta}_1 x_i) = 0, \sum_{i=1}^{n} x(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - x)$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - x)} = \frac{\sum_{i=1}^{n} (x_i - x)(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - x)^2} = \frac{S_{xy}}{S_{xx}}$

②(b) Consider the models $y_i = \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$ where ϵ_i are assumed i.i.d $N(0, \sigma^2)$, is called the no-intercept simple linear regression model.

(i) Find the ordinary least square estimator of β_1.

$y_i = \beta_1 x_i + \epsilon_i$ Assumption: $\epsilon_i \sim N(0, \sigma^2) = y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Least square estimator $\hat{\beta}_1$ must satisfy $\frac{\delta S}{\delta \beta_1} = 0$

$$\frac{\delta S}{\delta \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_1 x_i)$$, setting this at 0 gives: $\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 x_i) = 0$

(ii) Find an estimator σ^2 of σ^2

We estimate σ^2 by: $\sigma^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$, $n-2$ is the degrees of freedom & $\sum_{i=1}^{n} e_i^2$ is called the radical sum of squares, denoted SSE.

$Var(\epsilon_i) = \sigma^2$, $Var(y_i) = \sigma^2$. 3

(2)(a) Consider the model $Y_i = \beta_0 + \epsilon_i$, $i = 1, \dots, n$ where ϵ_i are assumed i.i.d. $N(0, \sigma^2)$.

(i) Find the ordinary least square estimator of β_0.

Least squares estimator $\hat{\beta}_0$ must satisfy: $\frac{\delta S}{\delta \hat{\beta}_0} = 0$

$\frac{\delta S}{\delta \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0)$ Setting this to 0 at $\hat{\beta}_0$ gives: $\sum_{i=1}^{n} (Y_i - \hat{\beta}_0) = 0$

(ii) Find an estimator σ^2 of σ^2.

We estimate σ^2 by: $\sigma^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$, $n-2$ is the degrees of freedom & $\sum_{i=1}^{n} e_i^2$ is called the residual sum of squares, denoted SSE.

(3)

A simple linear regression model may be written as either:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{or} \quad Y_i = \alpha_0 + \alpha_1 (x_i - \bar{x}) + \epsilon_i$$

(a) Interpet the parameters $\beta_0, \beta_1, \alpha_0$ & α_1 & find the relationship between the α parameters & β parameters.

$$\beta_0 = \alpha_0, \quad \epsilon_i = \epsilon_i, \quad \beta_1 x_i = \alpha_1 (x_i - \bar{x}) \rightarrow$$

$$\beta_1 = \alpha_1 (-\bar{x})$$

(b) $\cancel{S(\beta_0, \beta_1)}$

$$S(\alpha_0, \alpha_1) = \sum_{i=1}^{n} (Y_i - \cancel{\beta}\alpha_0 - \alpha_1 (x_i - \bar{x}))^2 \qquad \frac{\delta}{\delta \alpha_0} = 0 \quad \frac{\delta}{\delta \alpha_1}$$

$$\frac{\delta S}{\delta \alpha_0} = -2 \sum_{i=1}^{n} \left(Y_i - \alpha_0 - \alpha_1 (x_i - \bar{x})\right)^2 = \sum_{i=1}^{n} (Y_i - \alpha_0 - \alpha_1 (x_i - \bar{x})) = 0$$

$$\frac{\delta S}{\delta \alpha_1} = -2 \sum_{i=1}^{n} x_i \left(Y_i - \alpha_0 - \alpha_1 (x_i - \bar{x})\right)^2 \quad \sum_{i=1}^{n} x_i (Y_i - \alpha_0 - \alpha_1 (x_i - \bar{x})) = 0$$

~~The expected values are positive~~ $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

$$Var(Y_0 - \hat{Y}_0) = Var(Y_0) + Var(\hat{Y}_0) = \sigma^2 + \sigma^2 \times h_{00} = \sigma^2 (1 + h_{00}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

(d) $Cov(\hat{\alpha}_0, \hat{\alpha}_1) = 0 = \mathbb{E}(\hat{\alpha}_0(\hat{\alpha}_1)) - \mathbb{E}(\hat{\alpha}_0)\mathbb{E}(\hat{\alpha}_1) = 0$

(e) Because it's a lot easier ~~to~~ for a number with a lot of decimal places to work on a computer with mean centering.

R code:

```
#title: "Assignment 1"

#output: pdf_document

#author: Colm Mooney 20325583


library(magrittr)

library(ggplot2)

library(readr)

Concrete <- read_csv("SharedFiles/ST303/data/Concrete.csv")

Life <- read_csv("SharedFiles/ST303/data/Life.csv")

#Q(4) (a)


plot(x = Concrete$`7Day`, y = Concrete$`28Day`, xlab ="7Day", ylab= "28Day")

abline(lm(Concrete$`7Day`~ Concrete$`28Day`))


par(mfrow=c(1,2))

plot(Concrete$`7Day`, ylab = "Strength", xlab = "Concrete Samples")

plot(Concrete$`28Day`, ylab = "Strength", xlab = "Concrete Samples")

# Does it seem appropriate to assume a linear relationship between the two variables?

#Look at attached image for answer.


#(b) & (c) & (d)


mylm <- lm(formula = Concrete$`7Day` ~ Concrete$`28Day`, data = Concrete)

summary(mylm)

anova(mylm)


mylm %>%

  broom::augment(Concrete) %>%

  head()
```

```
mylm %>%
  broom::augment(Concrete) %>%
  ggplot(aes(x = Concrete$`7Day`, y = Concrete$`28Day`, col = "red")) +
  geom_point() +
  geom_line(aes(x = Concrete$`7Day`, y = Concrete$`28Day`, col = "blue"))

y <- Concrete$`28Day`
x <- Concrete$`7Day`
n <- 6
fit <- lm(y ~ x)
coef(fit)
s <- sqrt(sum(residuals(fit)^2) / (n - 2))
s^2



#Q5
#Answer the following questions using R for industrialised countries only and excluding South Africa.
library(dplyr)
table(Life)

Part1=filter(Life, Life$country != "South_Africa" && Life$code == 1)
```

10/10

```
#(a) Draw a scatterplot of life expectancy versus per capita income
plot(Part1$life, Part1$income, xlab = "life", ylab ="Income")
```

20/20

```
#(b) Fit the simple linear regression of life expectancy on capita income.
mylm2 <- lm(Part1$life ~ Part1$income, data = Part1)
summary(mylm2)
anova(mylm2)
```

10/10

```
#(c) Draw the fitted line on the scatterplot.
abline(lm(Part1$income ~ Part1$life))
```

10/20

#(d) What are the estimates of the slope and intercept and what are their standard errors?

#They both roughly equal 0. using the summary() command, we can see the Residual standard error is 11.21

#They all give some variation of 6.449e+01. They are supposed to say 0. It is like this due to a rounding Error.

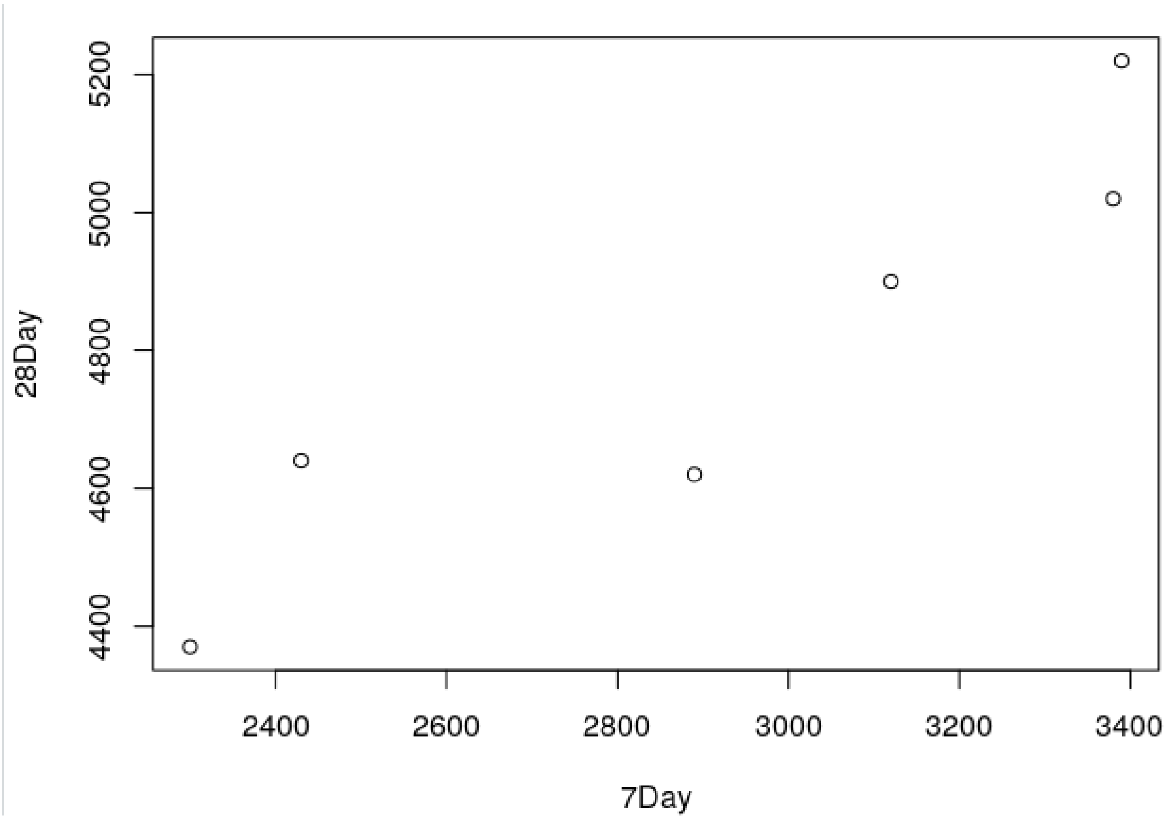
8/10

#(e) Yes, There very much is one.

8/10

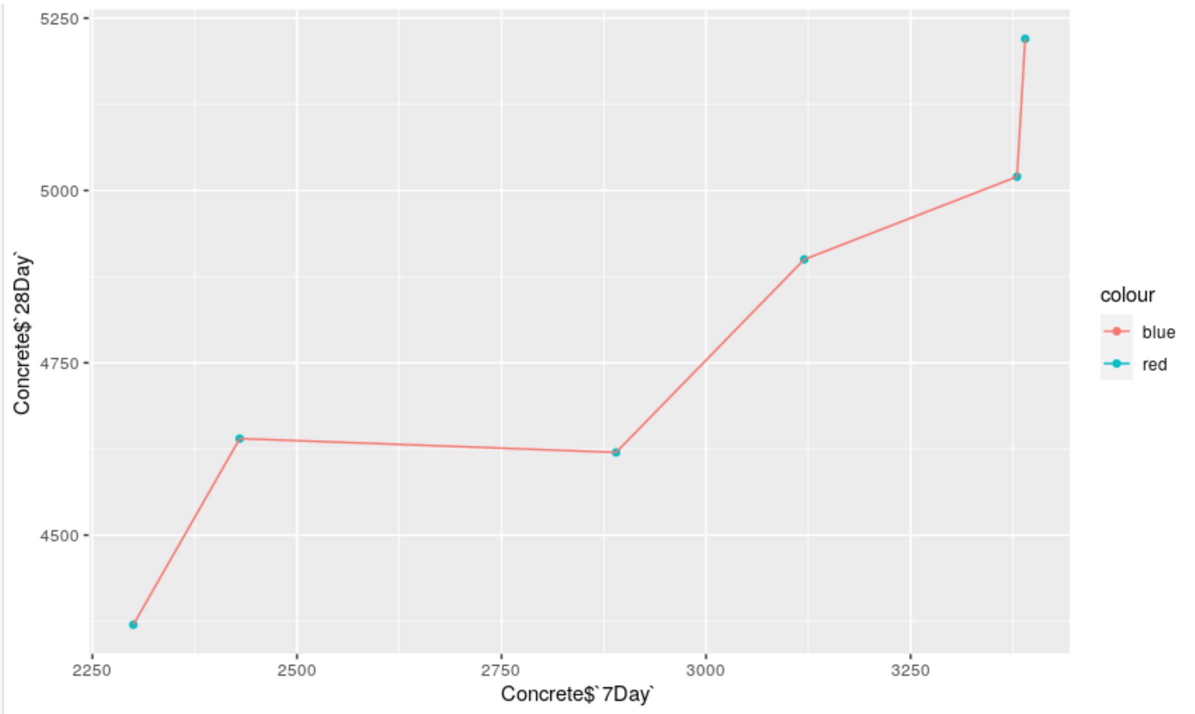
#(f)

```
y2 <- Part1$life

x2 <- Part1$income

n <- 19

fit2 <- lm(y2 ~ x2)

coef(fit2)

s2 <- sqrt(sum(residuals(fit)^2) / (n - 2))

s2^2
```



```
> summary(mylm)

Call:
lm(formula = Concrete$`7Day` ~ Concrete$`28Day`, data = Concrete)

Residuals:
      1       2       3       4       5       6 
 -23.00 -271.21  216.81   54.58  146.49 -123.67 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)   
(Intercept)      -3798.4822  1393.7055  -2.725  0.05269 . 
Concrete$`28Day`     1.4008     0.2902   4.828  0.00848 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 200.5 on 4 degrees of freedom
Multiple R-squared:  0.8535,	Adjusted R-squared:  0.8169 
F-statistic: 23.31 on 1 and 4 DF,  p-value: 0.008475

> anova(mylm)
Analysis of Variance Table

Response: Concrete$`7Day`
                 Df Sum Sq Mean Sq F value   Pr(>F)   
Concrete$`28Day`  1 937062  937062  23.307 0.008475 **
Residuals         4 160821   40205                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

```
> anova(mylm)
Analysis of Variance Table

Response: Concrete$`7Day`
                  Df Sum Sq Mean Sq F value   Pr(>F)   
Concrete$`28Day`   1 937062  937062  23.307 0.008475 **
Residuals          4 160821   40205                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mylm %>%
+   broom::augment(Concrete) %>%
+   head()
# A tibble: 6 x 8
  `7Day` `28Day` .fitted .resid  .hat .sigma .cooksd .std.resid
   <dbl>   <dbl>   <dbl>  <dbl> <dbl>  <dbl>   <dbl>      <dbl>
1   2300    4370   2323.  -23.0 0.545   231.  0.0173     -0.170
2   2430    4640   2701. -271.  0.217   149.  0.324      -1.53 
3   2890    4620   2673.  217.  0.231   182.  0.228       1.23 
4   3120    4900   3065.   54.6 0.190   229.  0.0107      0.302
5   3380    5020   3234.  146.  0.273   209.  0.138       0.857
6   3390    5220   3514. -124.  0.545   206.  0.500      -0.914
> mylm %>%
+   broom::augment(Concrete) %>%
+   ggplot(aes(x = Concrete$`7Day`, y = Concrete$`28Day`, col = "red")) +
+   geom_point() +
+   geom_line(aes(x = Concrete$`7Day`, y = Concrete$`28Day`, col = "blue"))
> 
```



(d)

	7-Day	28-Day	Differences
1	2300	4370	2070
2	2430	4640	2210
3	2890	4620	1730
4	3120	4900	1780
5	3380	5020	1640
6	3390	5220	1830

$\bar{x} = 2918.3$ $\bar{y} = 4795$

A linear relationship describes a straight line between 2 variables. It's safe to assume there isn't one in this case. The differences are too random for this to be the case. Best seen with 5&6 despite the 7-Days being only 10 apart, their 28 days are far apart from each other. ~~En~~ When plotted, there isn't a straight line.

(e) The regression coefficients are given by the R command coef(fit). Verify (by hand) the slope & intercept parameters & also estimate (again, by hand) the model's variance parameter.

$$Slope = \frac{Rise}{run} = (Y2-Y1)/(X2-X1) = 270/130 = 2.076923077 = \frac{27}{13}$$

$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x}) \times (y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
-618.3	-425	262777.5	382294.89	180625
-488.3	-155	75685.5	238436.89	24025
-28.3	-175	4952.5	800.89	30625
201.7	105	21178.5	40682.89	11025
461.7	225	103882.5	213166.89	50625
471.7	425	200472.5	222500.89	180625
~~Σ~~	Σ	Σ = 668949	1097878	Σ477550

$$1097878 + 5.34 = \Sigma 1097883.34$$

$$S_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}} = \sqrt{\frac{477550}{5}} = 309.046922$$

$$S_x = \sqrt{\frac{1097883.34}{5}}$$

$$r = \frac{668949}{1097883.34 \times 477550}$$

calculator

$= \cancel{1.2759...}$

$$r \frac{309.046922}{468.5400853}$$

$Slope = 0.60931$

$$Y\ intercept = 4795 - .60931(2918.3) = 3016.850627$$

$$Variance = \frac{1097883.34}{5} = 219576.668 = \cancel{\ }$$

(f) Interpret Parameter estimates: $Y = \beta_0 + \beta_1 X + \varepsilon$. β_1 is the change in the mean of Y for a 1 unit increase in X.

This model is too random.

Intercept:
Slope:

Std error:

(h) No, there
& stre

(i) Coeff

~~Inter~~
Concre

Plea

Coe

Intercept: 3016.8337559
Slope: 0.60931

~~Std error~~ = ~~0.80777~~
= 0.01778

662742.6341 ~~Residual Std error: 0.1542~~
133796.2586

Std error: 1393.7055 = Intercept
0.2902 = Slope

.462 × 3016.8337559 = std error
.4763 × .0.60931 = 0.2902

(a) No, there isn't evidence of a linear association between mean 28-Day strength & strength at 7 days.

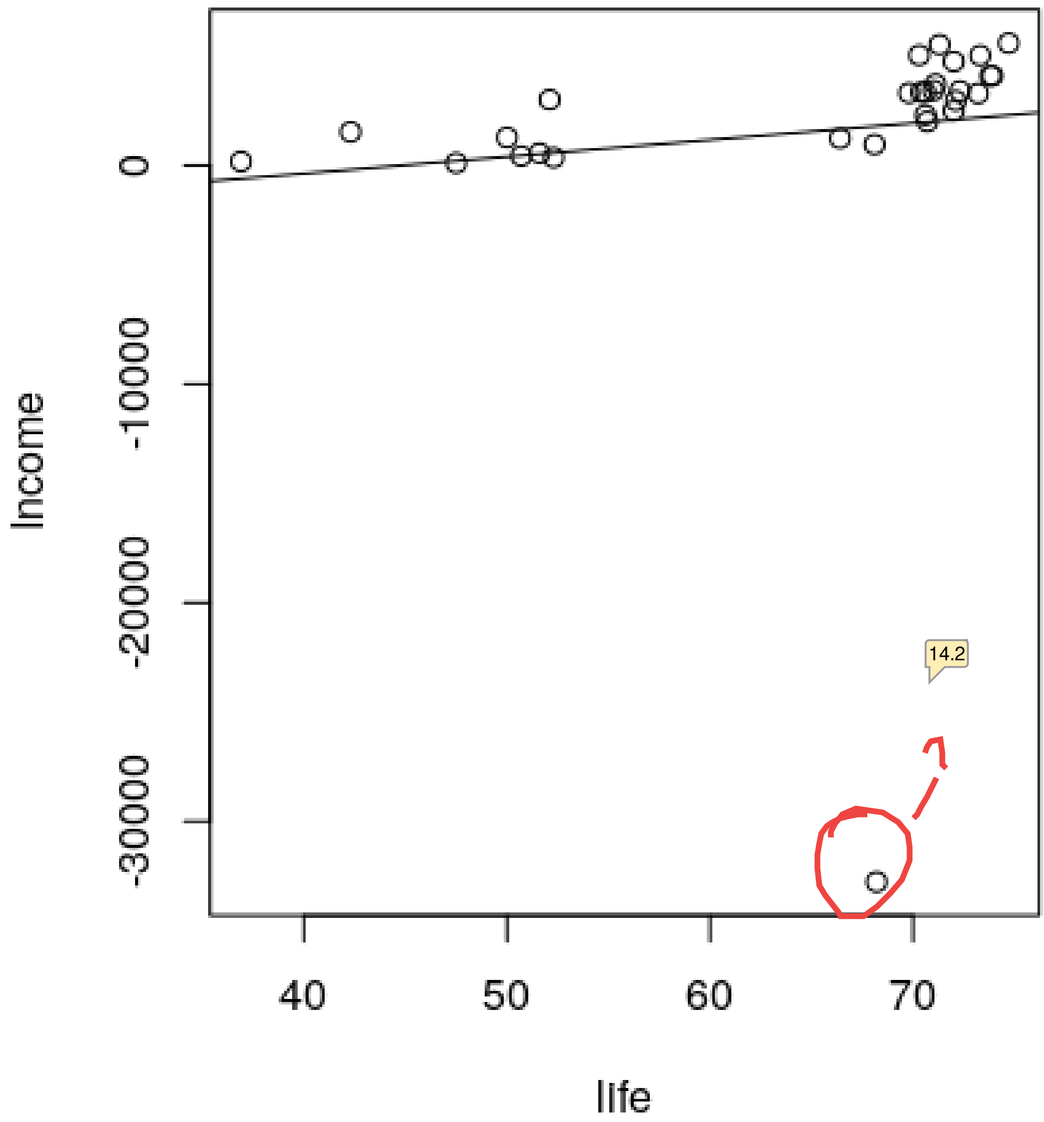
(c) Coefficients: -3798.4822 & 1.4008

	Estimate	Std. Error	t value	Pr(>|t|)
~~Intercept~~	~~-3798.4822~~	~~0.80777~~	~~19.754~~	~~1.09e-06~~
Concrete$28Day	~~-0.2774~~	~~0.01746~~	~~-7.328~~	~~0.00033~~

~~Please don't make me do this~~

Coefficient of determinant: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{477550}{1097883.34} = 0.4349733552$

```
Call:
lm(formula = Part1$life ~ Part1$income, data = Part1)

Residuals:
    Min      1Q  Median      3Q     Max 
-27.622 -12.259   5.645   7.009  10.415 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.449e+01  2.136e+00  30.194   <2e-16 ***
Part1$income 2.045e-04  3.114e-04   0.657    0.517    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.21 on 27 degrees of freedom
Multiple R-squared:  0.01572,	Adjusted R-squared:  -0.02073 
F-statistic: 0.4312 on 1 and 27 DF,  p-value: 0.5169
```



Index of comments

8.1 Just one single &. When you did that you didnt filtered the data properly

9.1 Wrong values due to wrong model specifications.

9.2 How, which evidence?

9.3 Which value?

14.1 Wrong values due to wrong model specifications.

14.2 This shouldn't be here with the correct model specifications. You filtered trhe data wrongly