Homework 3

Jun Ryu, UID: 605574052

2023-04-21

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

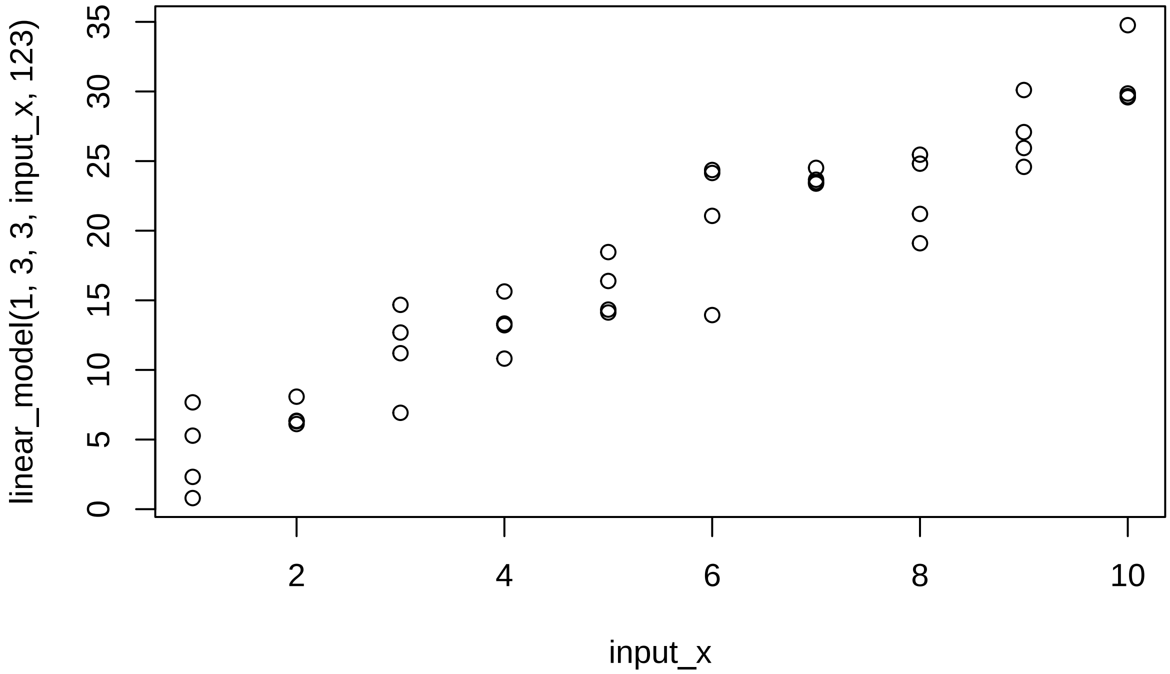
```
library(dplyr)
```

Question 1

a)

```
linear_model <- function(beta_0, beta_1, sigma, x, random.seed = 123) {
  set.seed(random.seed)
  epsilon <- rnorm(length(x), 0, sigma)
  beta_0 + beta_1*x + epsilon
}

input_x <- rep(1:10,by=.1,4)
plot(input_x, linear_model(1, 3, 3, input_x, 123))
```



b)

```r
cor(input_x, linear_model(1, 3, 3, input_x, 123))
```

```
## [1] 0.9529631
```

c)

```r
# only need to change the sigma value from 3 to 1 to minimize the errors
linear_model(1, 3, 1, input_x, 123)
```

```
##  [1]  3.439524  6.769823 11.558708 13.070508 16.129288 20.715065 22.460916
##  [8] 23.734939 27.313147 30.554338  5.224082  7.359814 10.400771 13.110683
## [15] 15.444159 20.786913 22.497850 23.033383 28.701356 30.527209  2.932176
## [22]  6.782025  8.973996 12.271109 15.374961 17.313307 22.837787 25.153373
## [29] 26.861863 32.253815  4.426464  6.704929 10.895126 13.878133 16.821581
## [36] 19.688640 22.553918 24.938088 27.694037 30.619529
```

```r
cor(input_x, linear_model(1, 3, 1, input_x, 123))
```

```
## [1] 0.9946951
```

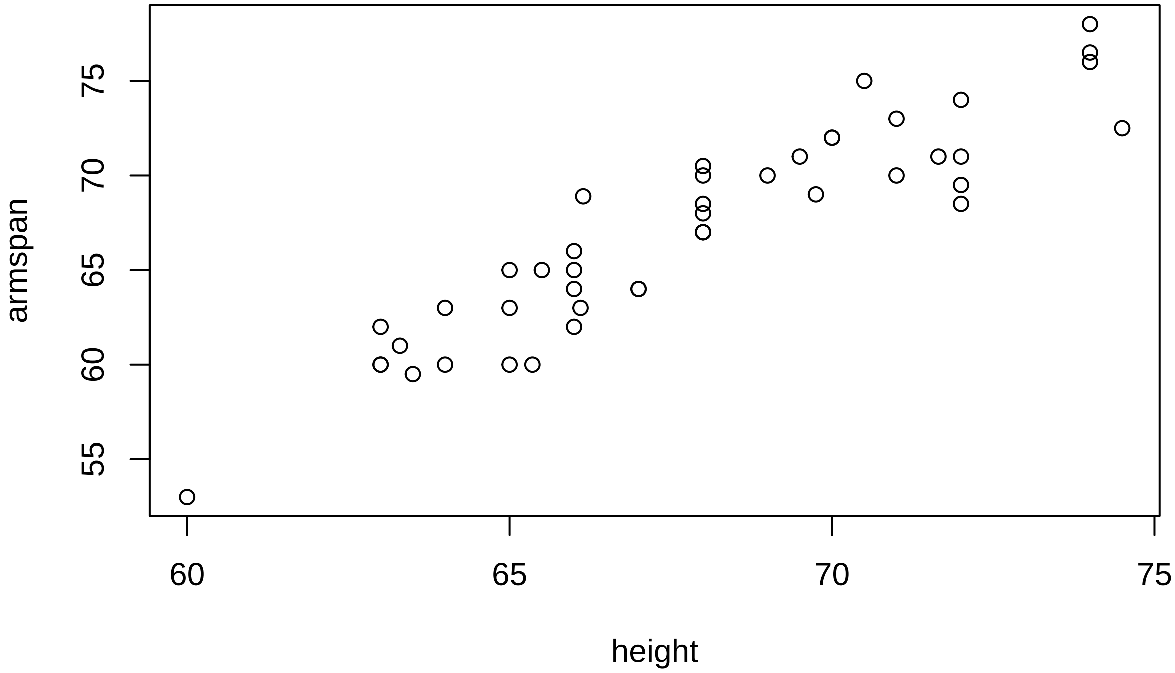
Question 2

```
armspan22 <- read.csv("armspans2022_gender.csv")
armspan22 <- na.omit(armspan22) # remove the observations with NA values
head(armspan22)
```

```
##   height armspan is.female compmother     compfather
## 1  74.00    76.0         0     Taller         Taller
## 2  65.00    65.0         0     Taller About the same
## 3  60.00    53.0         1    Shorter        Shorter
## 4  69.75    69.0         0     Taller About the same
## 5  70.00    72.0         0     Taller About the same
## 6  68.00    70.5         0     Taller        Shorter
```

a)

```
plot(armspan~height, data=armspan22)
```



```
cor(armspan22$height, armspan22$armspan)
```

```
## [1] 0.92147
```

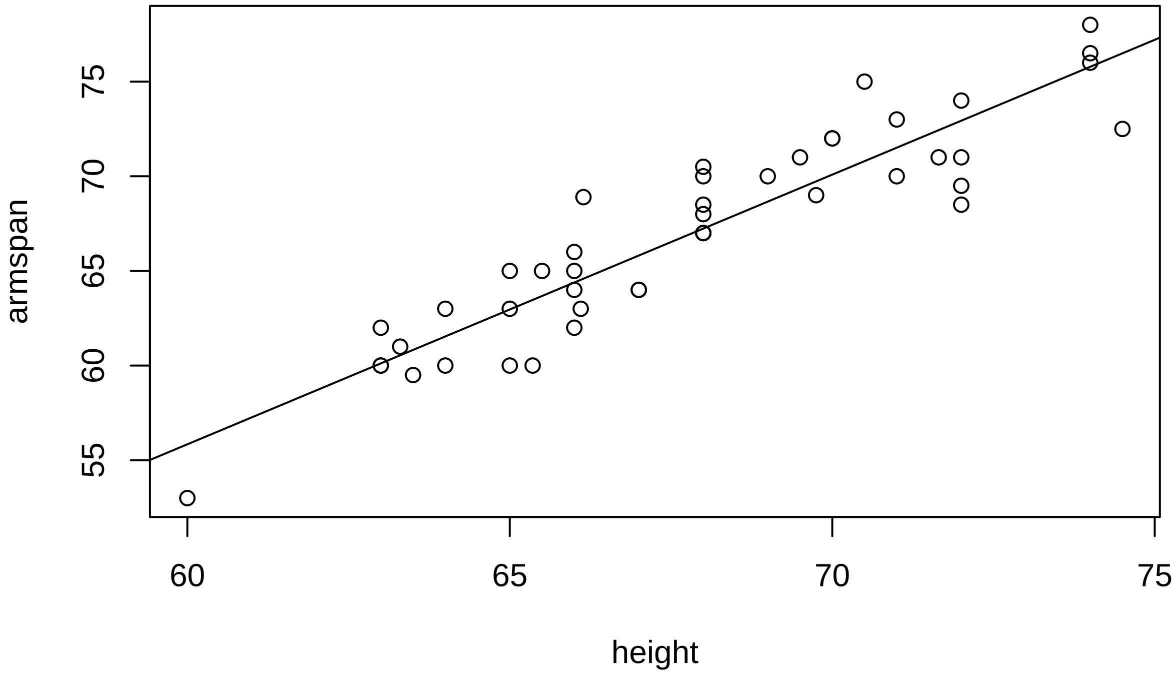
According to the plot, the two variables seem to have a strong positive correlation. Indeed, checking the correlation coefficient, we get 0.92147 (a value very close to 1), indicating a strong positive correlation.

b)

```
model <- lm(armspan~height, data=armspan22)
model # yields the equation: armspan = 1.425*height - 29.635
```

```
##
## Call:
## lm(formula = armspan ~ height, data = armspan22)
##
## Coefficients:
## (Intercept)       height
##     -29.635        1.425
```

```
plot(armspan~height, data=armspan22)
abline(model)
```



c)

```
# for my height 5'7'' (67 inches):
predict(model, data.frame(height = 67)) # prediction: 65.8 inches
```

```
##        1
## 65.81231
```

```
# my actual armspan is 68 inches
residual <- 68 - predict(model, data.frame(height = 67))
residual # (actual - predicted) = 2.188 inches
```

```
##        1
## 2.187694
```

d)

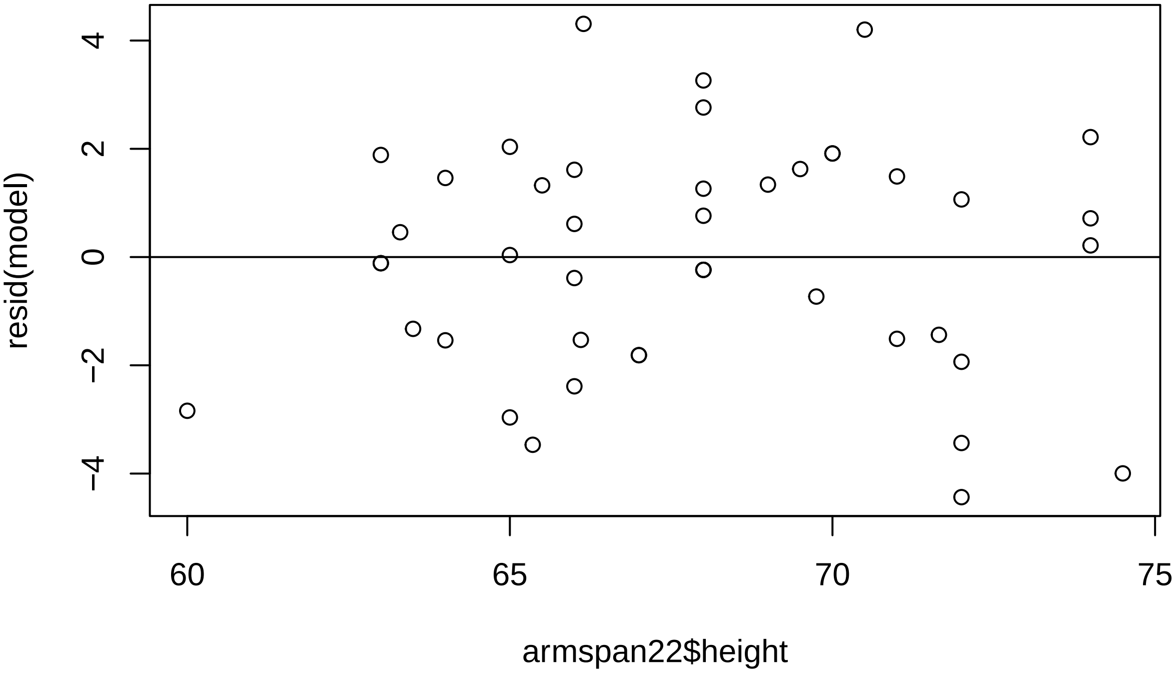
```
# for Michael Phelps (76 inches tall):
predict(model, data.frame(height = 76))
```

```
##        1
## 78.63363
```

From the linear model and the corresponding prediction of 78.63 inches, his armspan length (79 inches) is not unusual.

e)

```
plot(armspan22$height, resid(model))
abline(0,0)
```



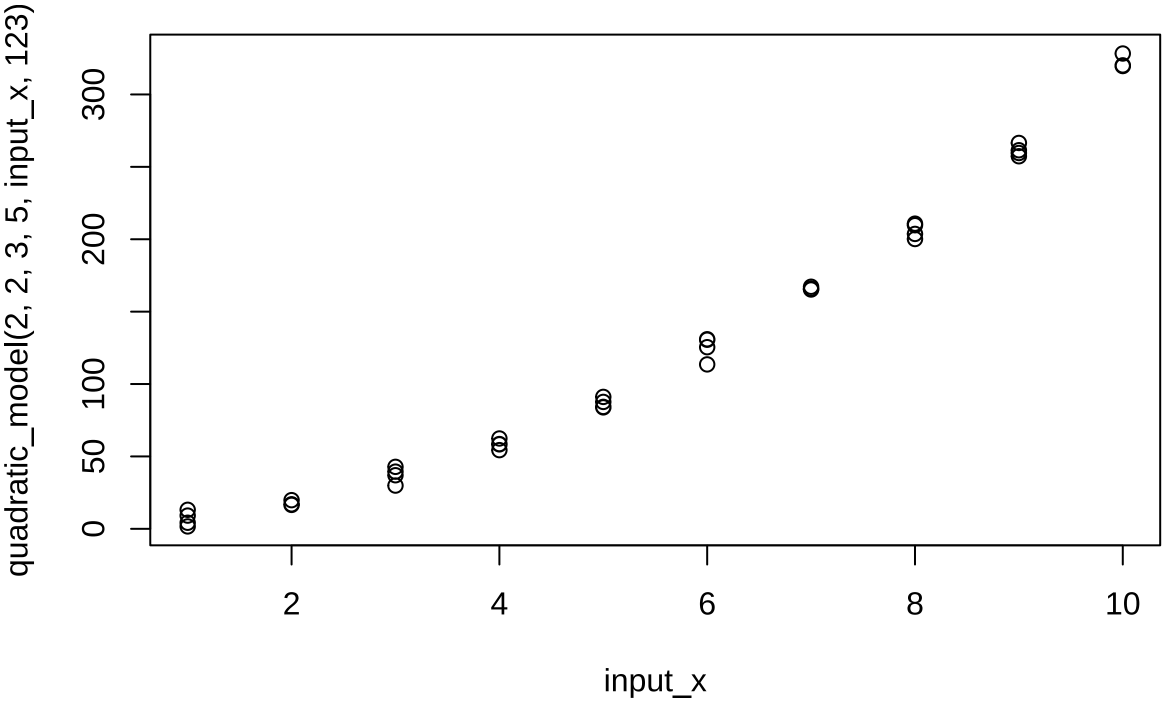
Looking at the residual plot, we see a good constant spread of the residual up and down the midline (indicating 0 error) across the independent observations. Therefore, this indicates a good linear fit between the two variables of armspan and height.

Question 3

a)

```
quadratic_model <- function(a, b, c, sigma, x, random.seed = 123) {
  set.seed(random.seed)
  epsilon <- rnorm(length(x), 0, sigma)
  a + b*x + c*x^2 + epsilon
}

input_x <- rep(1:10,by=.1,4)
plot(input_x, quadratic_model(2, 2, 3, 5, input_x, 123))
```

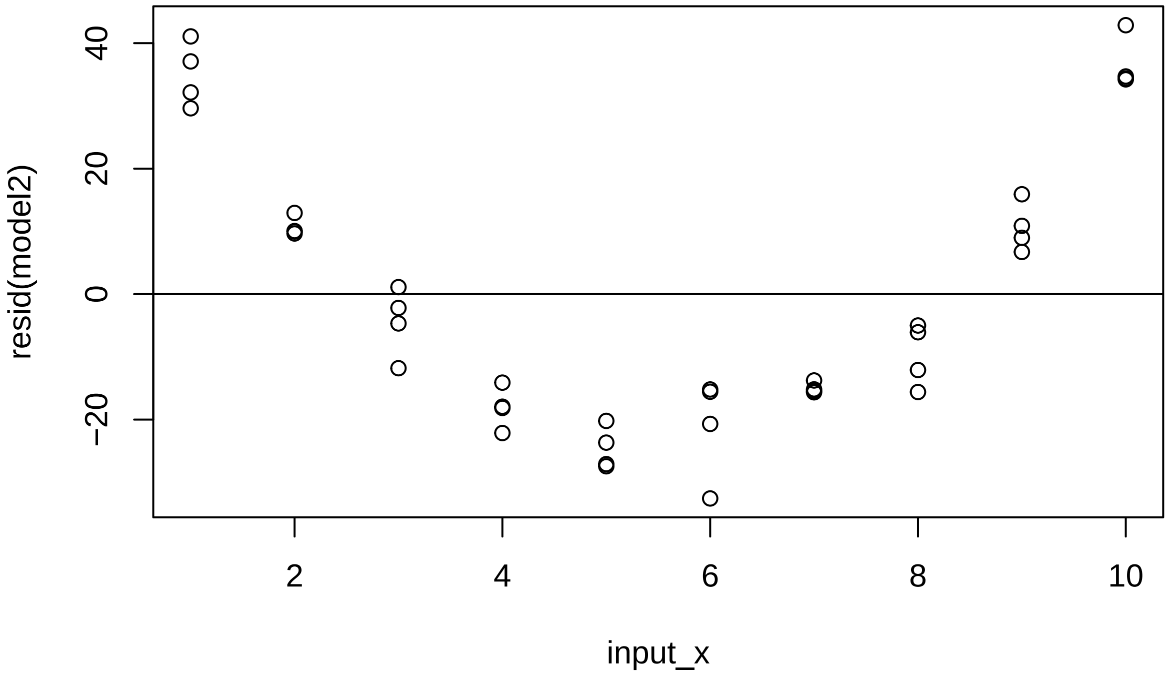


b)

```
model2 <- lm(quadratic_model(2, 2, 3, 5, input_x, 123) ~ input_x)
model2
```

```
##
## Call:
## lm(formula = quadratic_model(2, 2, 3, 5, input_x, 123) ~ input_x)
##
## Coefficients:
## (Intercept)      input_x
##      -62.77        34.82
```

```
plot(input_x, resid(model2))
abline(0,0)
```



From the above residual plot, we observe that the residuals are not constantly spread across the x variable. For example, in the region where the x values are $[4, 8]$, the residuals are all negative, while in the region outside of that bound, the residuals are mostly positive.

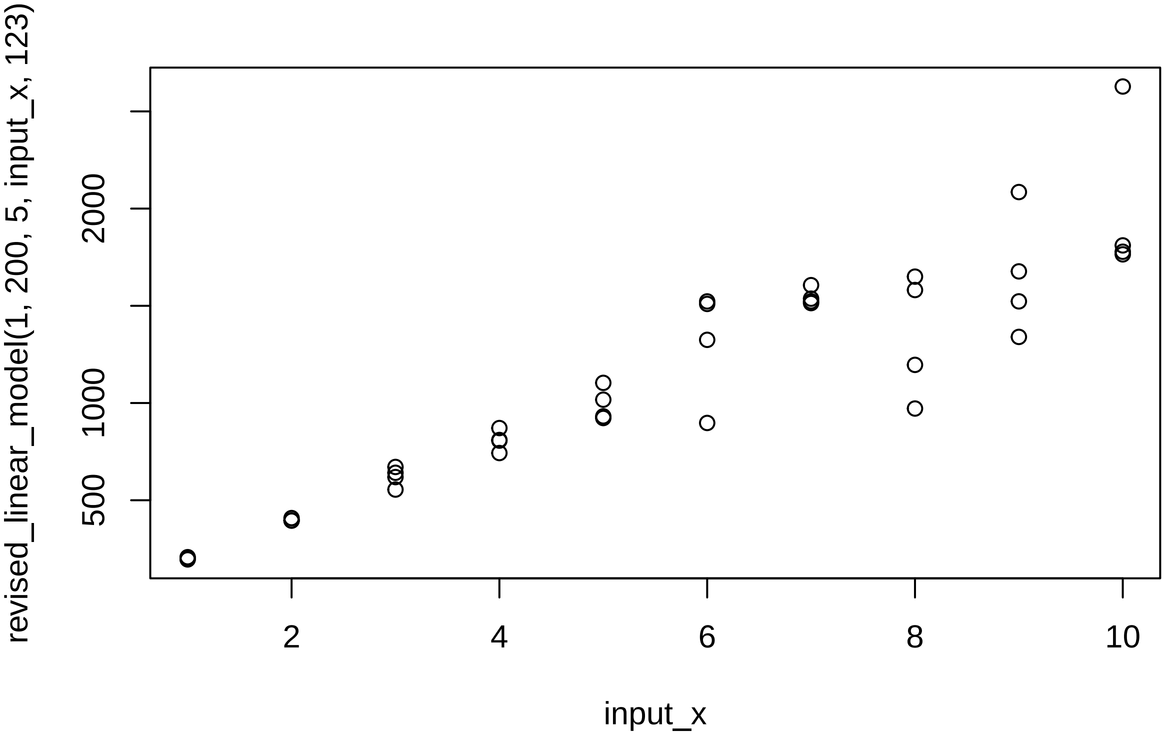
c)

We can use the residual plot to see if the residuals show a nature of randomness (independent, normal, and constant standard deviation). If the residuals seem to violate one of these three properties, then we can tell the trend is probably non-linear.

d)

```r
revised_linear_model <- function(a, b, sigma, x, random.seed = 123) {
  set.seed(random.seed)
  epsilon <- rnorm(length(x), 0, sigma*x^2)
  a + b*x + epsilon
}

input_x <- rep(1:10,by=.1,4)
plot(input_x, revised_linear_model(1, 200, 5, input_x, 123))
```



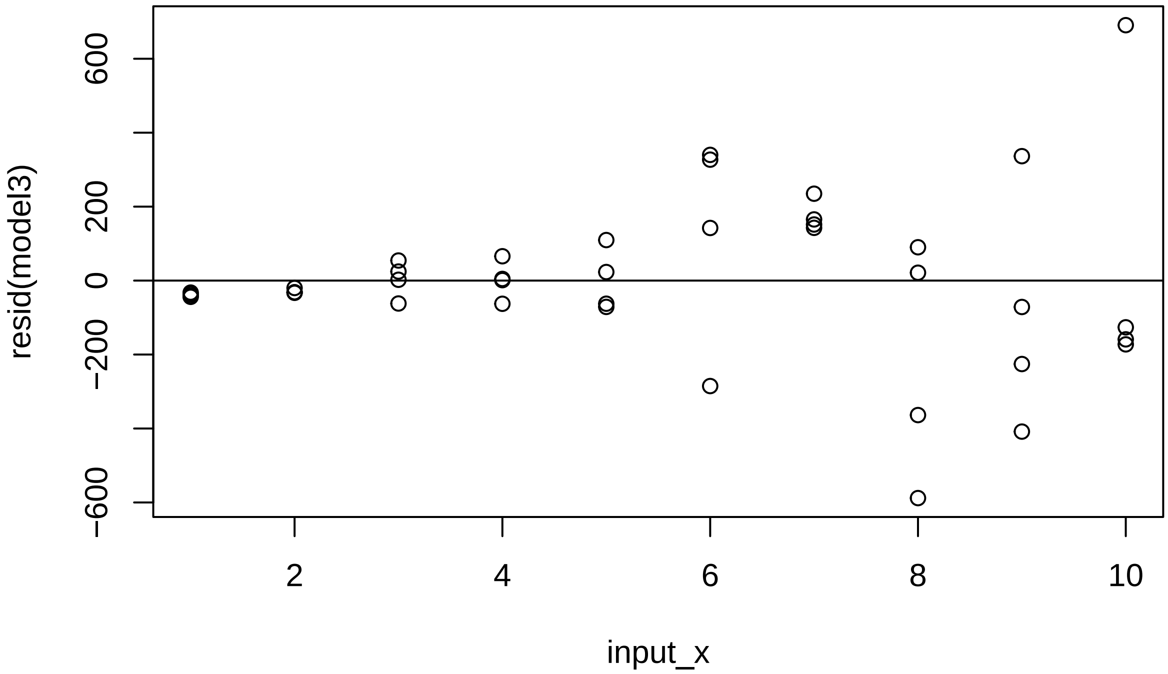
The plot does not seem to show a clear linear trend as the deviations cause some of the points to be further away from the linear fit.

e)

```
model3 <- lm(revised_linear_model(1, 200, 5, input_x, 123) ~ input_x)
model3
```

```
##
## Call:
## lm(formula = revised_linear_model(1, 200, 5, input_x, 123) ~
##     input_x)
##
## Coefficients:
## (Intercept)      input_x
##       50.91       188.62
```

```
plot(input_x, resid(model3))
abline(0,0)
```



We notice that as the x values get larger, the residuals also become more extreme (in both positive and negative directions). This violates the constant standard deviation part as the residuals get further and further away.

Question 4

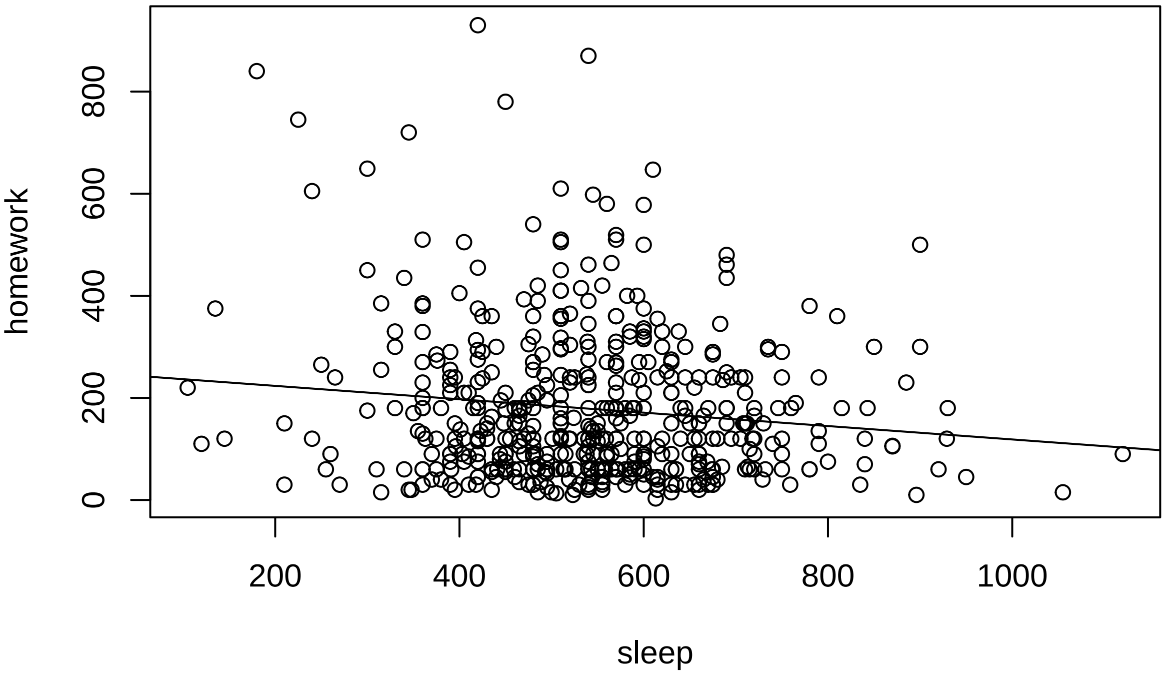
a)

```
atus <- read.csv("atus.csv")
atus1 <- subset(atus,homework>0)

plot(homework ~ sleep, data=atus1)
model4 <- lm(homework ~ sleep, data=atus1)
model4
```

```
##
## Call:
## lm(formula = homework ~ sleep, data = atus1)
##
## Coefficients:
## (Intercept)        sleep
##    249.7618      -0.1312
```

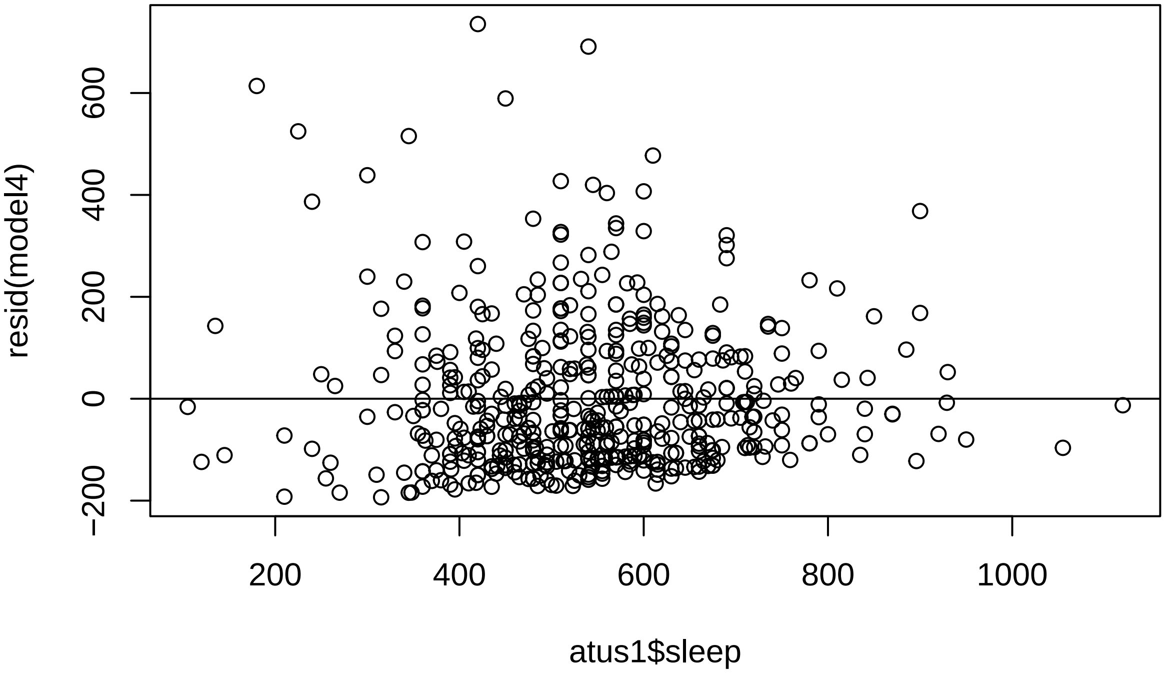
```
abline(model4)
```



The linear fit shows a negative correlation between the amount of time spent on sleep vs. homework, but looking at the plot in general, the model does not seem to be a good fit as we have numerous points that are widely detached from the linear model.

b)

```
plot(atus1$sleep, resid(model4))
abline(0,0)
```



From the residual plot, we can deduce a similar observation where a lot of the residuals are way above the midline, indicating a non-normal distribution of deviations. Thus, this linear model is not a great fit for this dataset.

Question 5

a)

Let $\mu_1 =$ (Average time doing household chores for those who identified as female) and $\mu_2 =$ (Average time doing household chores for those who identified as male). Our null is $H_0 : \mu_1 = \mu_2$ and our alternative is $H_a : \mu_1 \neq \mu_2$.

```
t.test(atus1$household_chores[atus1$gender == "Female"],
       atus1$household_chores[atus1$gender == "Male"])
```

```
##
##  Welch Two Sample t-test
##
## data:  atus1$household_chores[atus1$gender == "Female"] and atus1$household_chores[atus1$gender == "
## t = 6.3978, df = 446.68, p-value = 3.986e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  32.31764 60.97594
## sample estimates:
## mean of x mean of y
##  77.17730  30.53052
```

The test statistic here is the t-test value, which is 6.3978. The p-value is 3.986^{-10} assuming unequal variance. Using a 5% significance level, since our p-value is less than 0.05, we have sufficient evidence to reject the null hypothesis and conclude that there is a significant difference between the average time doing household chores for female vs. male.

b)

Some conditions that must be met include normality of population, independence between observations, random sampling, and homogeneity of variance. In our dataset, these are satisfied because we have a random sample with a sufficient sample size and independent observations.