# 705604096_stats101a_hw3

Jade Gregory

2023-04-18

## Question 1

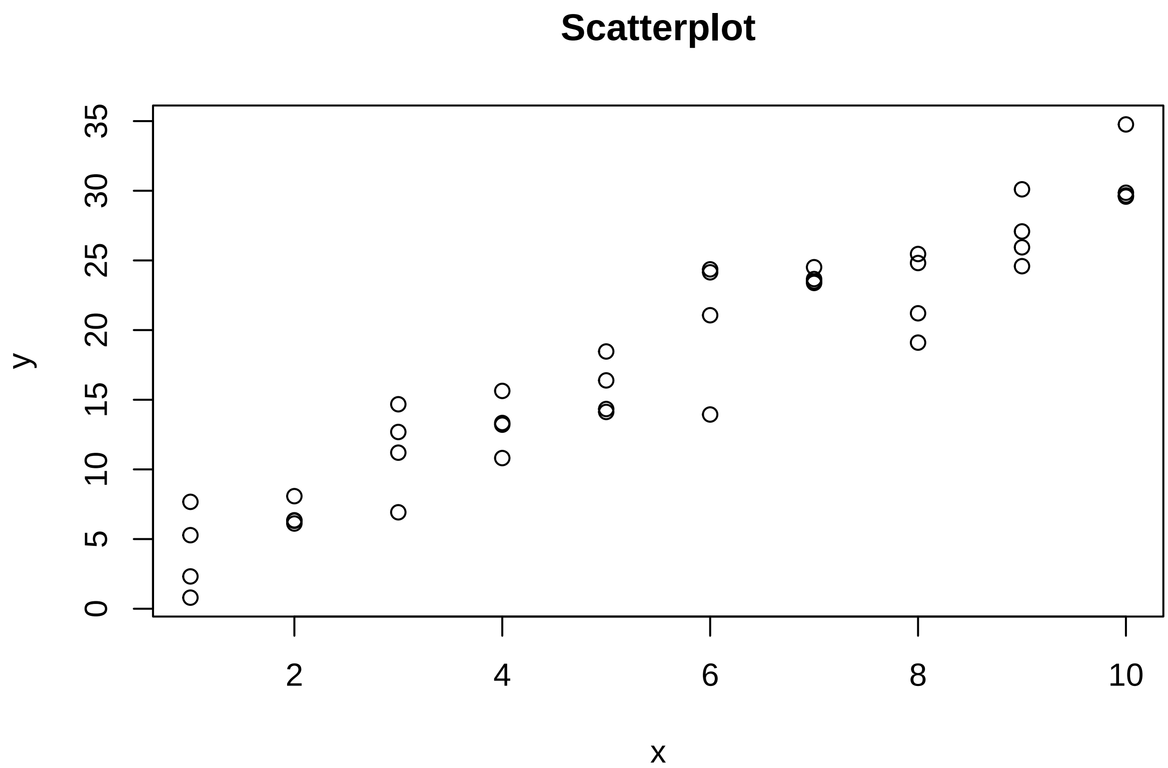
A)

a)

```
slr <- function(beta_0, beta_1, sigma, x, rns){
  set.seed(rns)
  eps <- rnorm(length(x), 0, sigma)
  beta_0 + (beta_1 * x) + eps
}
```

b)

```
x <- rep(1:10, by = .1, 4)
y <- slr(1, 3, 3, x, 123)
plot(x, y, main = "Scatterplot")
```




B)

```
cor(x, y)
```

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

The correlation coefficient is 0.9529631

C)

```
x2 <- rep(1:10, by = .1, 4)
y2 <- slr(1, 30, 3, x2, 123)
cor(x2, y2)
```

```
## [1] 0.9995272
```

With parameters

$$\beta_0 = 1$$

$$\beta_1 = 30$$

$$\sigma = 3$$

we get a correlation of 0.9995272

## Question 2

```
my_data <- read_csv('armspans2022_gender.csv')
```

```
## Rows: 46 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): compmother, compfather
## dbl (3): height, armspan, is.female
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

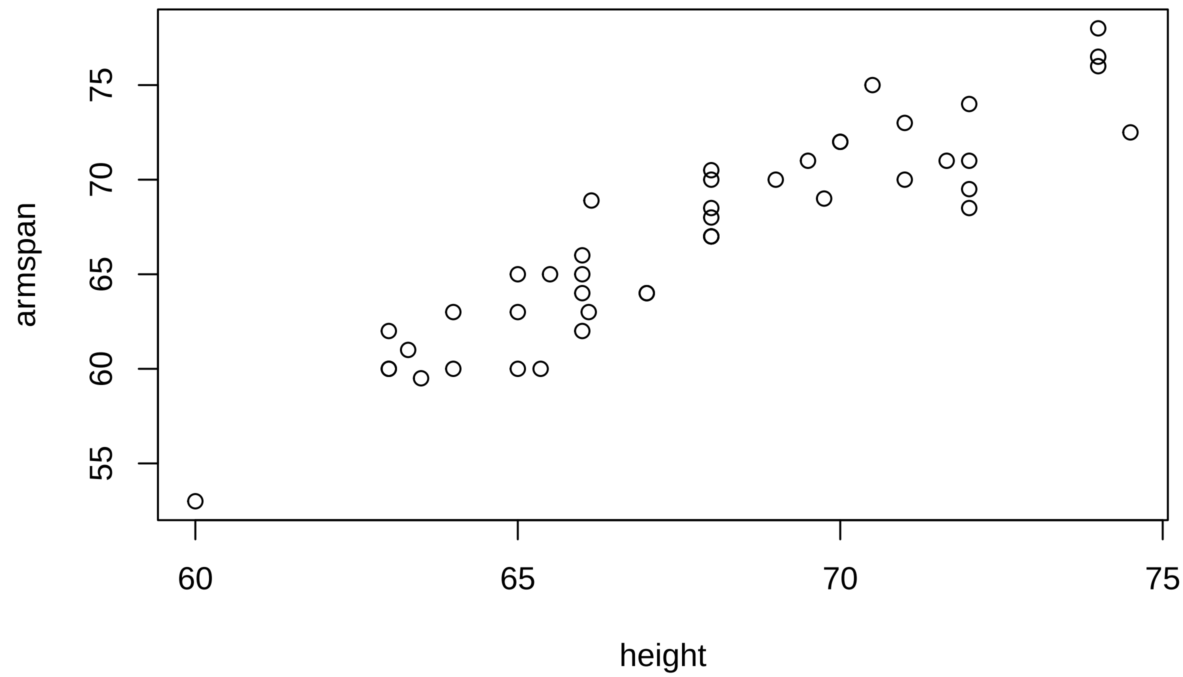
```
my_data['compmother'] <- tolower(my_data$compmother)
my_data['compfather'] <- tolower(my_data$compfather)
my_data$compmother <- str_replace_all(my_data$compmother, c(" " = "_", "'" =""))
my_data$compfather <- str_replace_all(my_data$compfather, c(" " = "_", "'" =""))
my_data
```

```
## # A tibble: 46 x 5
##    height armspan is.female compmother     compfather
##     <dbl>   <dbl>     <dbl> <chr>          <chr>
##  1   74      76           0 taller         taller
##  2   65      65           0 taller         about_the_same
##  3   60      53           1 shorter        shorter
##  4   69.8    69           0 taller         about_the_same
##  5   70      72           0 taller         about_the_same
##  6   68      70.5         0 taller         shorter
##  7   64      60           0 taller         taller
##  8   68      67           0 taller         about_the_same
##  9   68      67           1 taller         shorter
## 10   63      60           1 about_the_same shorter
## # ... with 36 more rows
```

When cleaning this data, I decided to make all of the elements in the columns lower case to make it easier to manipulate in the future. I also replaced spaces in between words with underscores for the same reason.

a)

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




We can observe a positive linear trend in this scatter plot. We can describe the strength as in between moderate and strong between the observations, since there is a clear linear pattern that is noticeable. There are not any unusual features we can discern from this scatter plot, except for the lowest value being farther from the bulk of the data but still in line with our linear pattern.

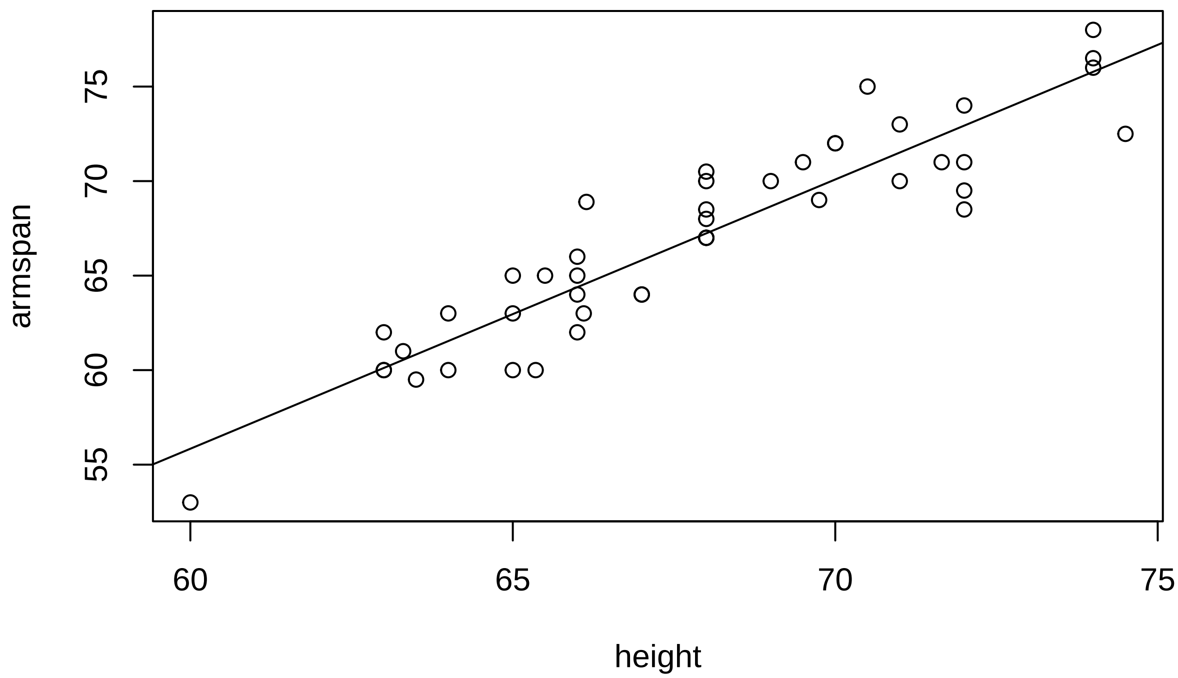
b)

```
plot(armspan ~ height, data = my_data )
my_lm <- lm(armspan ~ height, data = my_data)
summary(my_lm)
```

```
##
## Call:
## lm(formula = armspan ~ height, data = my_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4353 -1.5302  0.0369  1.4893  4.3080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -29.63530    6.22105  -4.764 2.19e-05 ***
## height        1.42459    0.09158  15.555  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.135 on 43 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8491, Adjusted R-squared:  0.8456
## F-statistic:   242 on 1 and 43 DF,  p-value: < 2.2e-16
```

```r
abline(my_lm)
```




The equation for our estimated line is

$$\hat{y}_i = -29.63530 + 1.42459x_i$$

c)

```r
my_func <- function(x){
  -29.6353 + 1.42459 * x
}
my_func(63)
```

```
## [1] 60.11387
```

```r
residual <- 61 - my_func(63)
residual
```

```
## [1] 0.88613
```

Based on our model, the predicted arm span for my height is 60.11387 inches. The residual is 0.88613 inches.

d)

```
my_func(76)
```

```
## [1] 78.63354
```

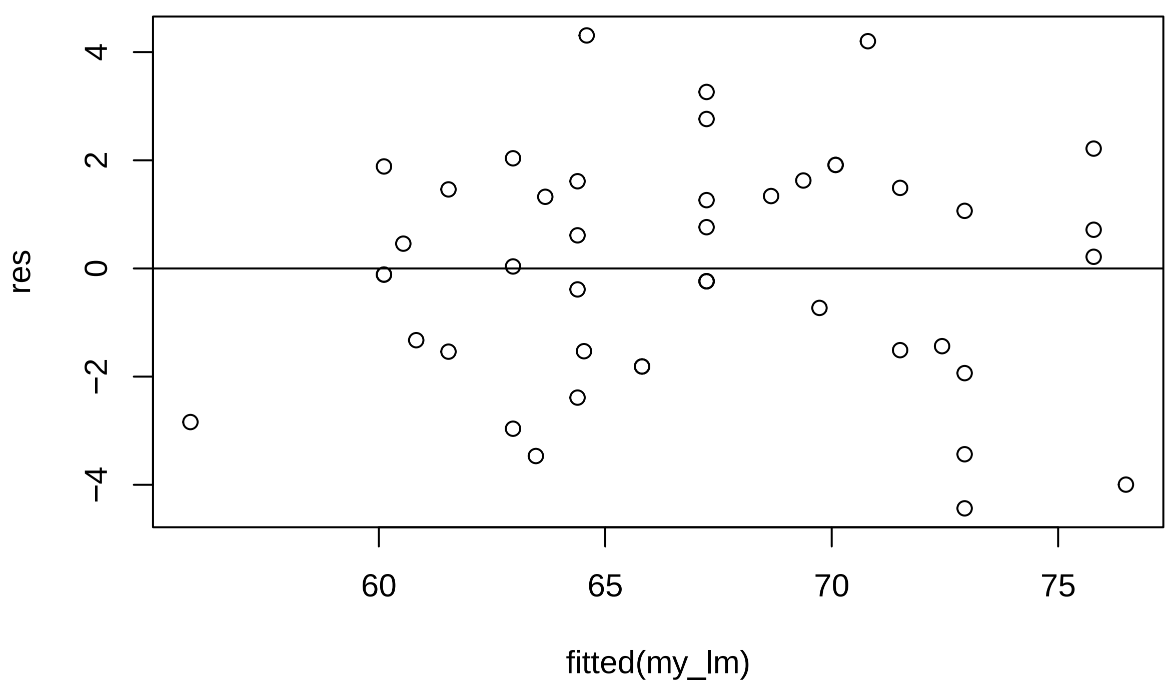
```
residual2 <- 79 - my_func(76)
residual2
```

```
## [1] 0.36646
```

No, this does not seem unusual as the residual is relatively small.

e)

```
res <- resid(my_lm)
plot(fitted(my_lm), res)
abline(0,0)
```

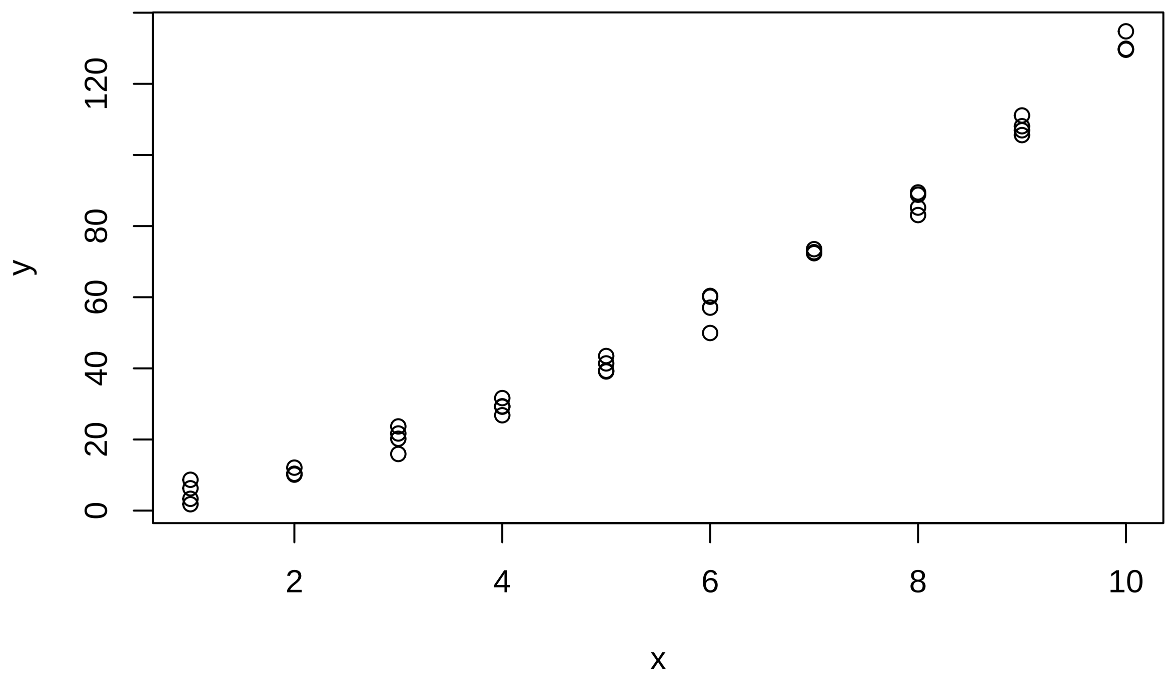



Because there does not seem to be any pattern in the residual points with respect to the line at x = 0, we can conclude that the linear model is the right fit for this data. We can also note that there is a fan shape that appears in our residual plot, with exception to the outlier as previously mentioned, that would indicate that the standard deviation increases as x increases.

### Question 3

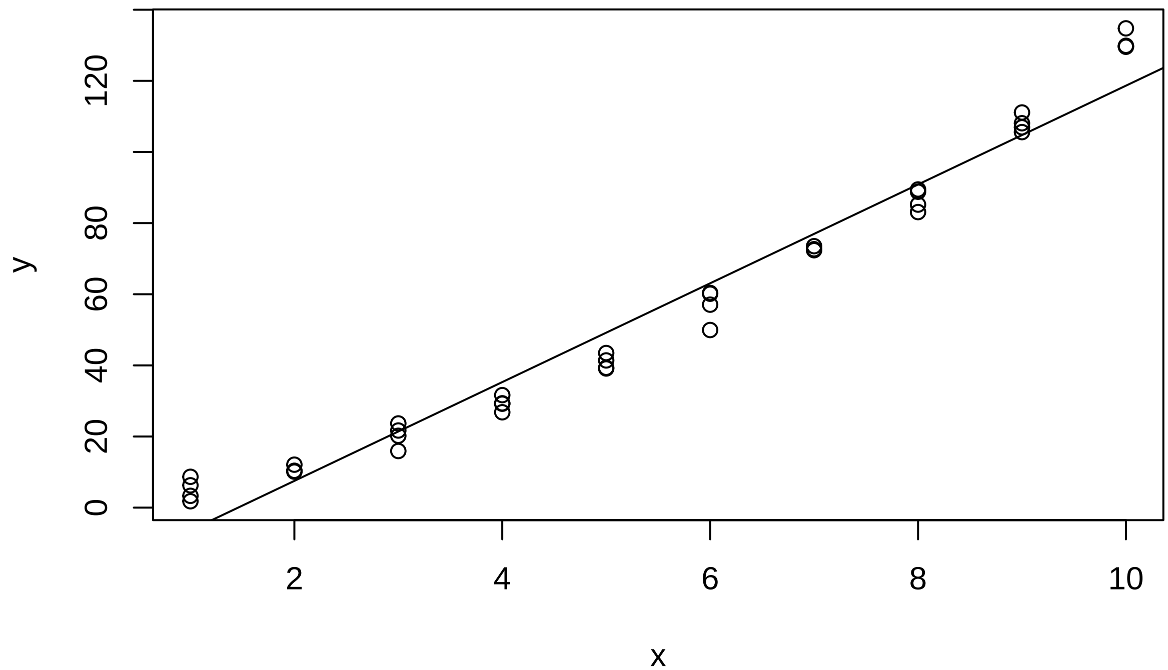
a)

```
quad_func <- function(a, b, c, sigma, x = rep(1:10,by=.1,4), random.seed){
  set.seed(random.seed)
  a + (b * x) + (c * (x^2)) + rnorm(length(x), 0, sigma)
}
x <- rep(1:10,by=.1,4)
y <- quad_func(1, 3, 1, 3, x, 123)
plot(x, y)
```

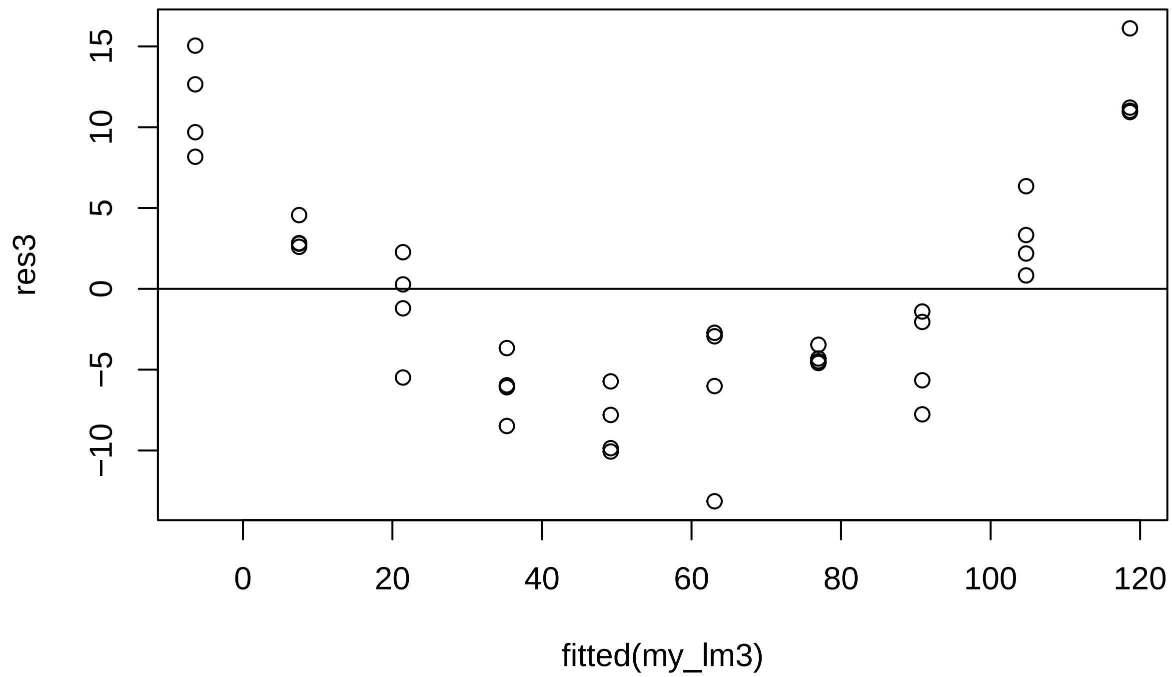



b)

```
plot(x, y)
my_lm3 <- lm(y ~ x)
abline(my_lm3)
```




```
res3 <- resid(my_lm3)
plot(fitted(my_lm3), res3)
abline(0,0)
```

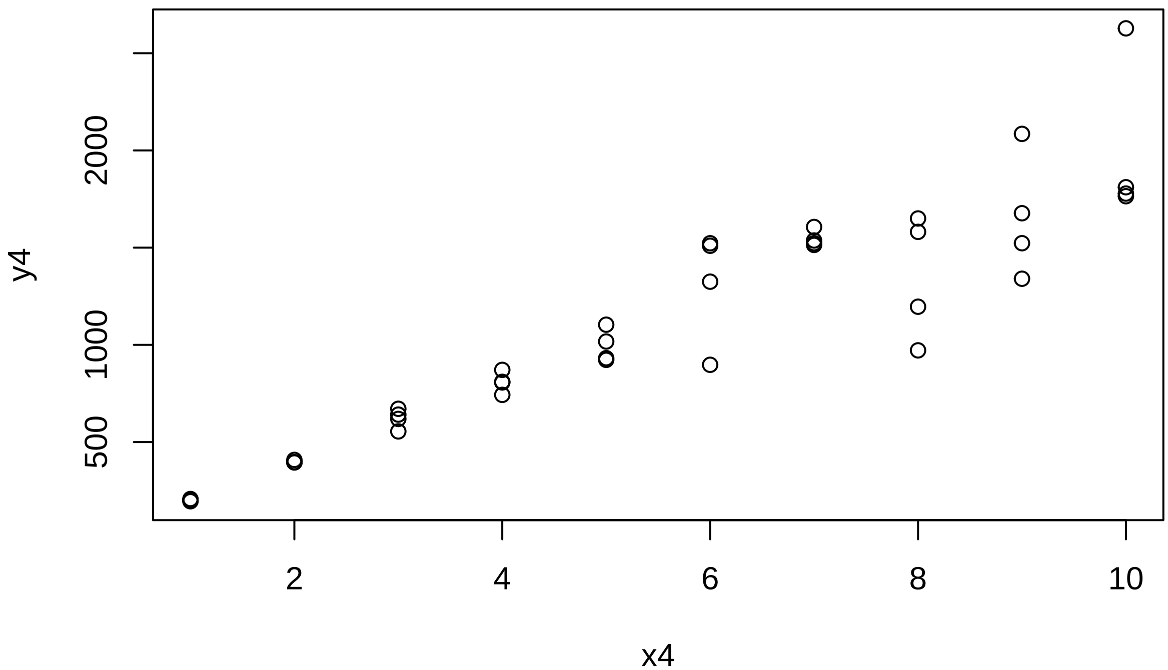



In this residual plot, we can observe a quadratic shape or a parabola appear.

c) If the residual plot shows features like parabolic shapes, we can conclude that the trend is non-linear.

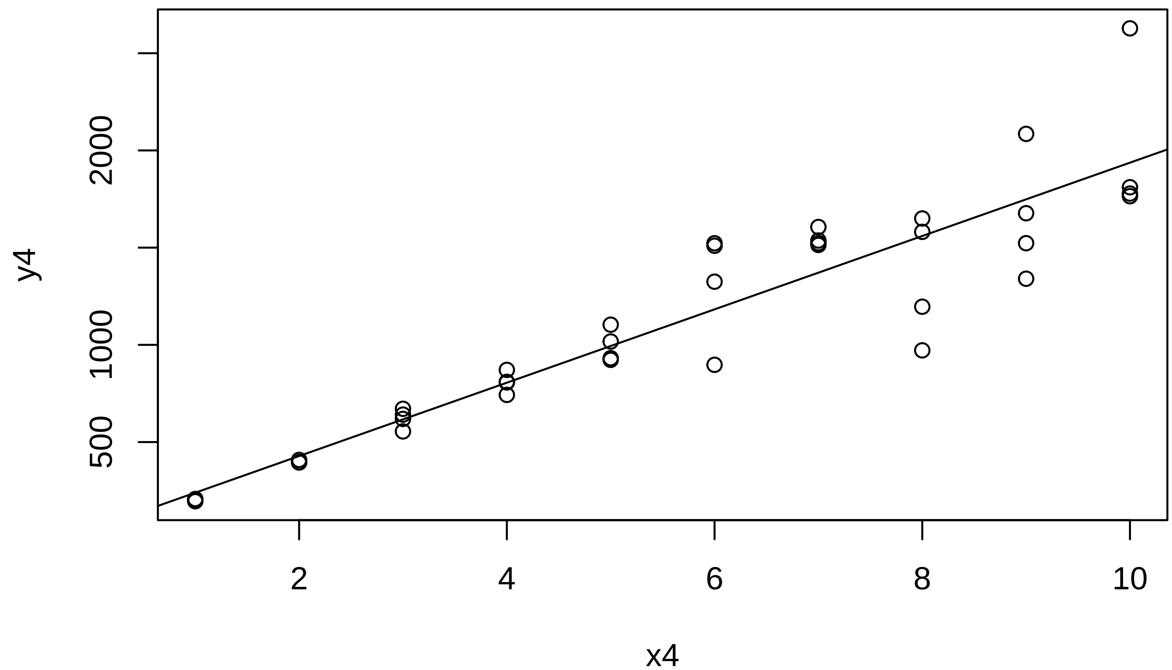
d)

```
my_nl_func <- function(a, b, x, sigma, random.seed){
  set.seed(random.seed)
  a + (b * x) + rnorm(length(x), 0, (sigma * (x^2)))
}
x4 <- rep(1:10, by = .1, 4)
y4 <- my_nl_func(1, 200, x4, 5, 123)
plot(y4 ~ x4)
```

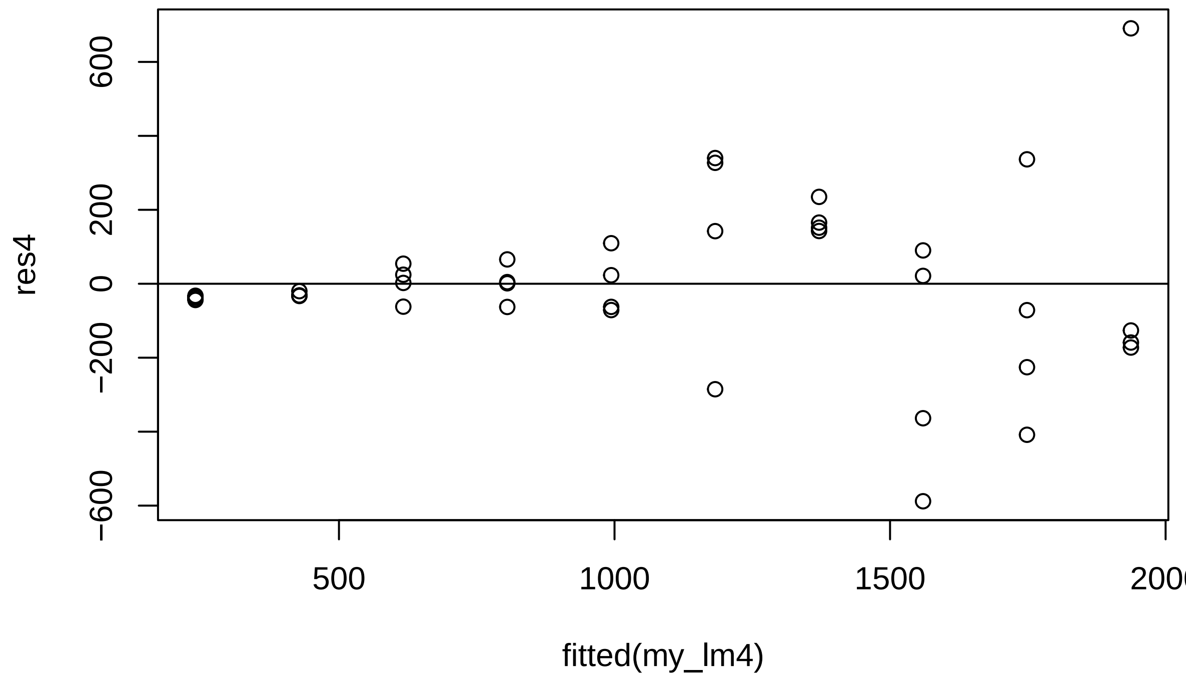



e)

```
plot(x4, y4)
my_lm4 <- lm(y4 ~ x4)
abline(my_lm4)
```




```
res4 <- resid(my_lm4)
plot(fitted(my_lm4), res4)
abline(0,0)
```




The cone shape of the residual plot indicates that the standard deviation increases as x increases, and therefore the constant standard deviation assumption is violated.

## Question 4

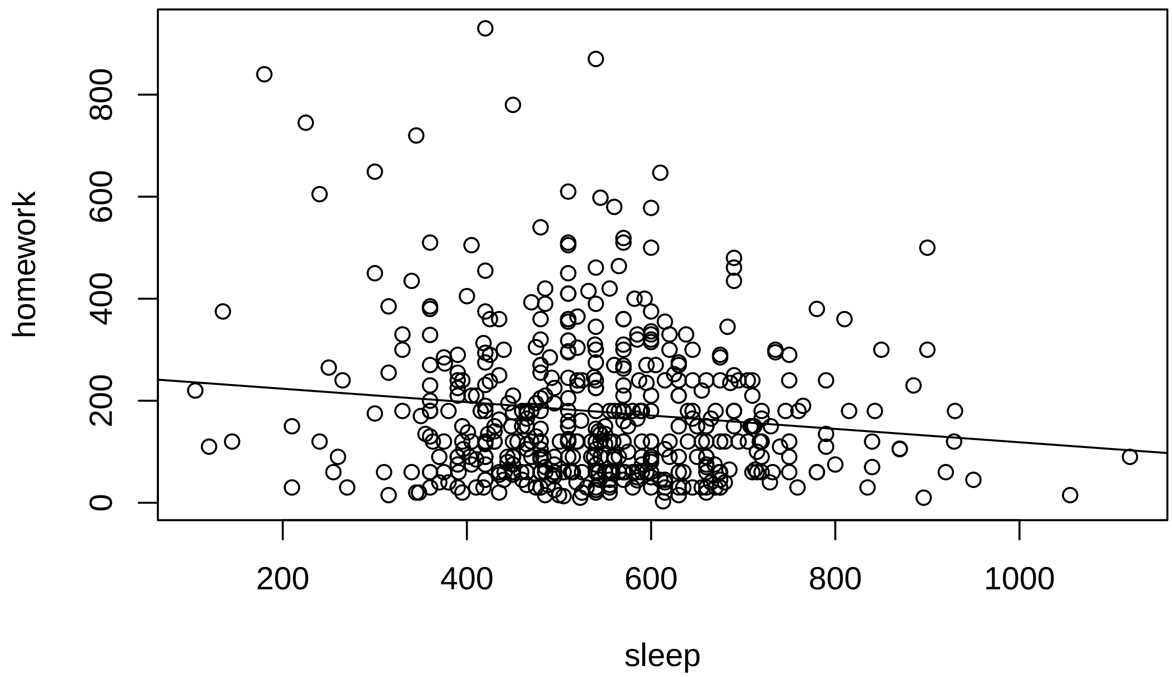
```
atus <- read.csv('atus.csv')
atus1 <- subset(atus, homework > 0)
head(atus1)
```

```
##           caseid        state age gender            citizen  marital_stat
## 24  2.01201e+13      Florida  20   Male Native, Born in USA Never married
## 32  2.01201e+13    Wisconsin  67 Female Native, Born in USA Never married
## 50  2.01201e+13      Florida  18   Male Native, Born in USA Never married
## 65  2.01201e+13     New York  43 Female Native, Born in USA       Married
## 103 2.01201e+13 Pennsylvania  30   Male Native, Born in USA Never married
## 105 2.01201e+13  Connecticut  20 Female Native, Born in USA Never married
##         veteran active_armedforces          emp_status multi_jobs
## 24  Non-Veteran                 No Unemployed, Looking  No answer
## 32      Veteran                 No  Not in labor force  No answer
## 50  Non-Veteran                 No            Employed         No
## 65  Non-Veteran                 No            Employed         No
## 103 Non-Veteran                 No            Employed         No
## 105 Non-Veteran                 No  Not in labor force  No answer
##                         work_class   retired fulltime_emp hours_worked
```

```
## 24                      No answer No answer    No answer           NA
## 32                      No answer No answer    No answer           NA
## 50            Private, for profit No answer    Part time           15
## 65  Self-employed, unincorporated No answer    Part time           30
## 103           Private, for profit No answer    Full time           40
## 105                     No answer No answer    No answer           NA
##               fam_income household_size household_kids household_child
## 24    $12,500 to $14,999              7              2             Yes
## 32    $12,500 to $14,999              1              0              No
## 50  $100,000 to $149,999              2              0              No
## 65    $60,000 to $74,999              5              3             Yes
## 103   $50,000 to $59,999              1              0              No
## 105    $150,000 and over              6              1             Yes
##     phys_challenge travel phone volunteer religion sports social food gov_civic
## 24  Has difficulty     30     0         0      120      0    590   10         0
## 32  Has difficulty     60     0         0        0      0    225   80         0
## 50   No difficulty     30     0         0        0      0    290   30         0
## 65   No difficulty    107     0         0        0      0     90   45         0
## 103  No difficulty     70    60         0        0      0      0   95         0
## 105  No difficulty      2     0         0        0      0    150  119         0
##     household pro_services purchasing education work care_nonhousehold
## 24          0            0          0        30    0                 0
## 32          0            0          0       189    0                 0
## 50          0            0         10       300    0                 0
## 65          0          280        105        30    0                 0
## 103         0            0          0       190  515                 0
## 105         0            0          0       461    0                 0
##     care_household household_chores personal_care sleep groom health_related
## 24               0                0           660   630    30              0
## 32               0              315           451   420    30              1
## 50               0                0           780   735    45              0
## 65               3               45           735   670    65              0
## 103              0                0           510   480    30              0
## 105              0                5           700   690    10              0
##     eating class homework socializing holiday      day year   month       date
## 24      10     0       30           0      No   Sunday 2012 January 2012-01-22
## 32      80    70      119           0      No   Friday 2012 January 2012-01-27
## 50      30     0      300           0      No   Sunday 2012 January 2012-01-29
## 65      45     0       30          82      No   Sunday 2012 January 2012-01-22
## 103     95   100       90           0      No Thursday 2012 January 2012-01-26
## 105    119     0      461           0      No   Sunday 2012 January 2012-01-29
```

a)

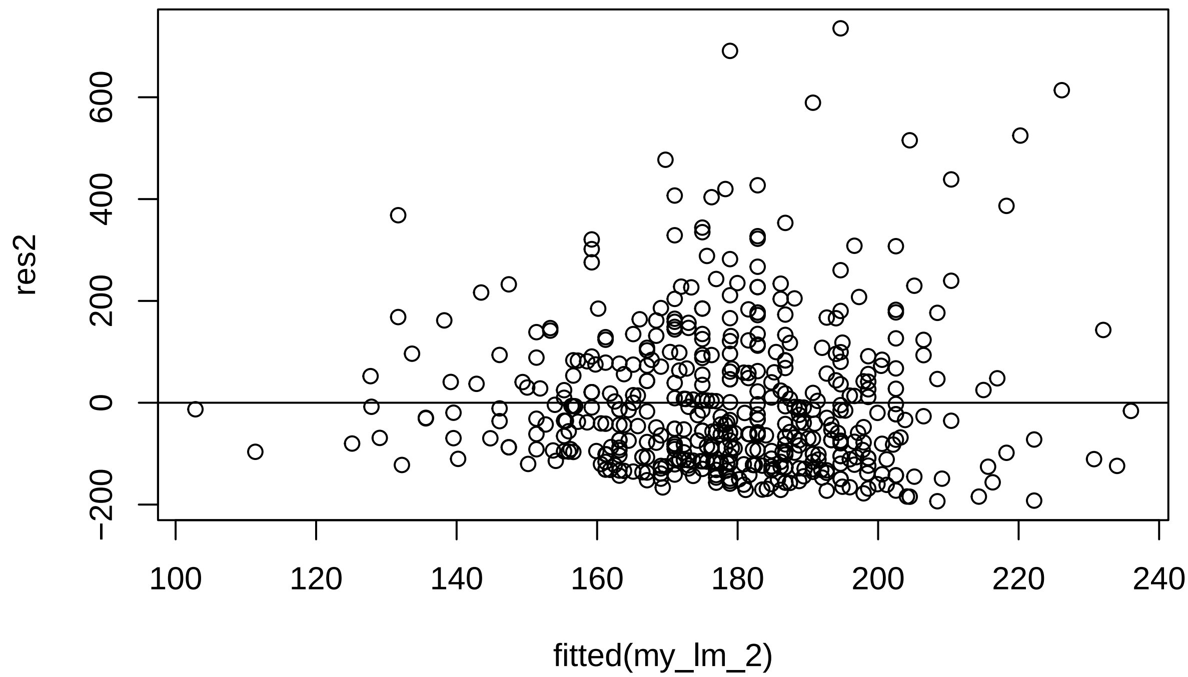
```
plot(homework ~ sleep, data = atus1)
my_lm_2 <- lm(homework ~ sleep, data = atus1)
abline(my_lm_2)
```




The linear model is not an appropriate model to describe the relationship between sleep times and homework times.

b)

```
res2 <- resid(my_lm_2)
plot(fitted(my_lm_2), res2)
abline(0,0)
```




From the residual plot, we can tell that this data is negatively correlated. We can also tell that since it is not evenly distributed, a linear model would not be the best model to approximate this data.

## Question 5

a)

```
t.test(atus1$household_chores ~ atus1$gender, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  atus1$household_chores by atus1$gender
## t = 6.3978, df = 446.68, p-value = 1.993e-10
## alternative hypothesis: true difference in means between group Female and group Male is greater than
## 95 percent confidence interval:
##  34.62905      Inf
## sample estimates:
## mean in group Female   mean in group Male
##             77.17730             30.53052
```

Let

$$\bar{x}_f$$

represent the average time persons identifying as female spent doing chores Let

$$\bar{x}_m$$

represent the average time persons identifying as male spent doing chores

$$H_0 : \bar{x}_f - \bar{x}_m = 0$$

$$H_a : \bar{x}_f - \bar{x}_m \neq 0$$

The test statistic is 6.3978. the observed value of the statistic is 446.68. The p-value is 1.993e-10. With a 5% significance level, we reject the null hypothesis.

b) We must assume the population distribution is normal or our sample size is sufficiently large to provide us with a good approximation. We can assume that these conditions are met because we have a large sample size observed in our data frame, which is sufficient.