DATA609 HW2

Bin Lin

2017-9-7

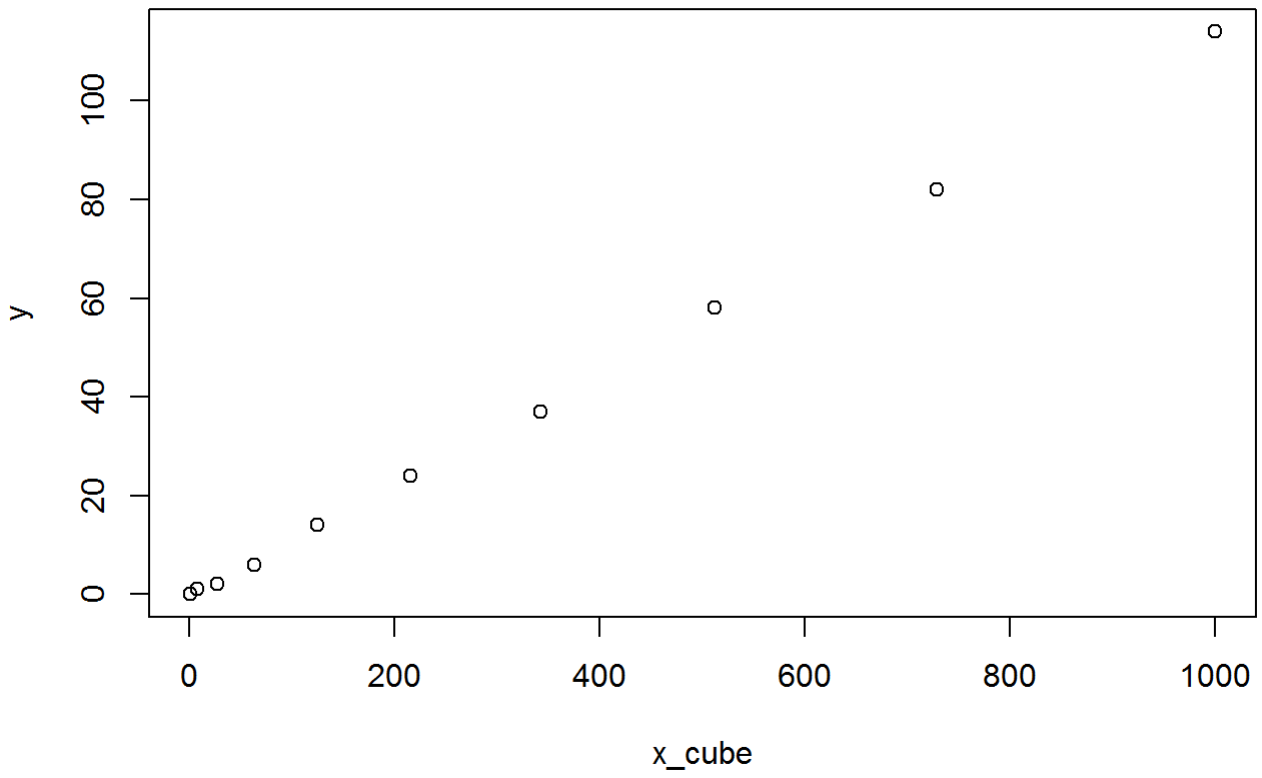
Page 79: #11. Determine whether the data set supports the stated proportionality model.

Force 10 20 30 40 50 60 70 80 90

Stretch 19 57 94 134 173 216 256 297 343

From the scatter plot, we can tell that y and x^3 has linear relationship which is about to pass the origin (0, 0). Then I created for the two variables and force the intercept to equal to 0. Their best fit line is: y = 0.113 * x^3, where k = 0.113.

```
y <- c(0, 1, 2, 6, 14, 24, 37, 58, 82, 114)
x <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
x_cube <- x ^ 3
plot(x_cube, y)
```




```
m1 <- lm(y ~ x_cube + 0)
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ x_cube + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7564 -0.8897 -0.2478  0.0438  1.0075
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## x_cube 0.1129925  0.0006307   179.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8872 on 9 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9997
## F-statistic: 3.209e+04 on 1 and 9 DF,  p-value: < 2.2e-16
```

The following graph shows the estimation of k value is valid. Therefore, the dataset support the proportionality model.

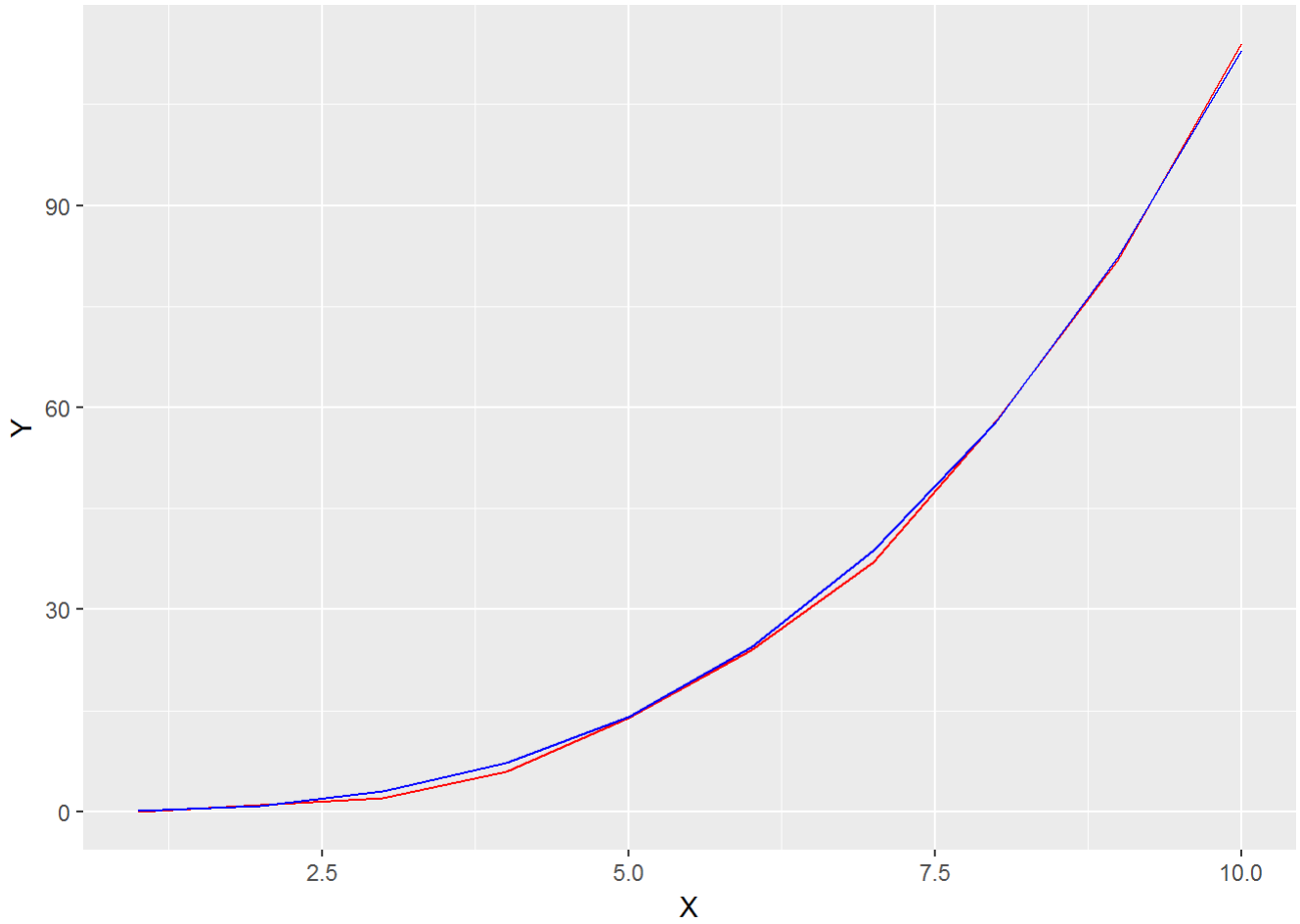
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
df1 <- data.frame(X = x, Y = y, Y_head = 0.113 * x_cube)
head(df1)
```

```
##   X  Y Y_head
## 1 1  0  0.113
## 2 2  1  0.904
## 3 3  2  3.051
## 4 4  6  7.232
## 5 5 14 14.125
## 6 6 24 24.408
```

```
ggplot() + geom_line(data = df1, aes(x = X, y = Y), color = "red") + geom_line(data = df1, aes(x
 = X, y = Y_head), color = "blue")
```



Page 94: #4. Lumber Cutters-Lumber cutters wish to use readily available measurements to estimate the number of board feet of lumber in a tree. Assume they measure the diameter of the tree in inches at waist height. Develop a model that predicts board feet as a function of diameter in inches.

Use the following data for your test:

x 17 19 20 23 25 28 32 38 39 41

y 19 25 32 57 71 113 123 252 259 294

The variable x is the diameter of a ponderosa pine in inches, and y is the number of board feet divided by 10.

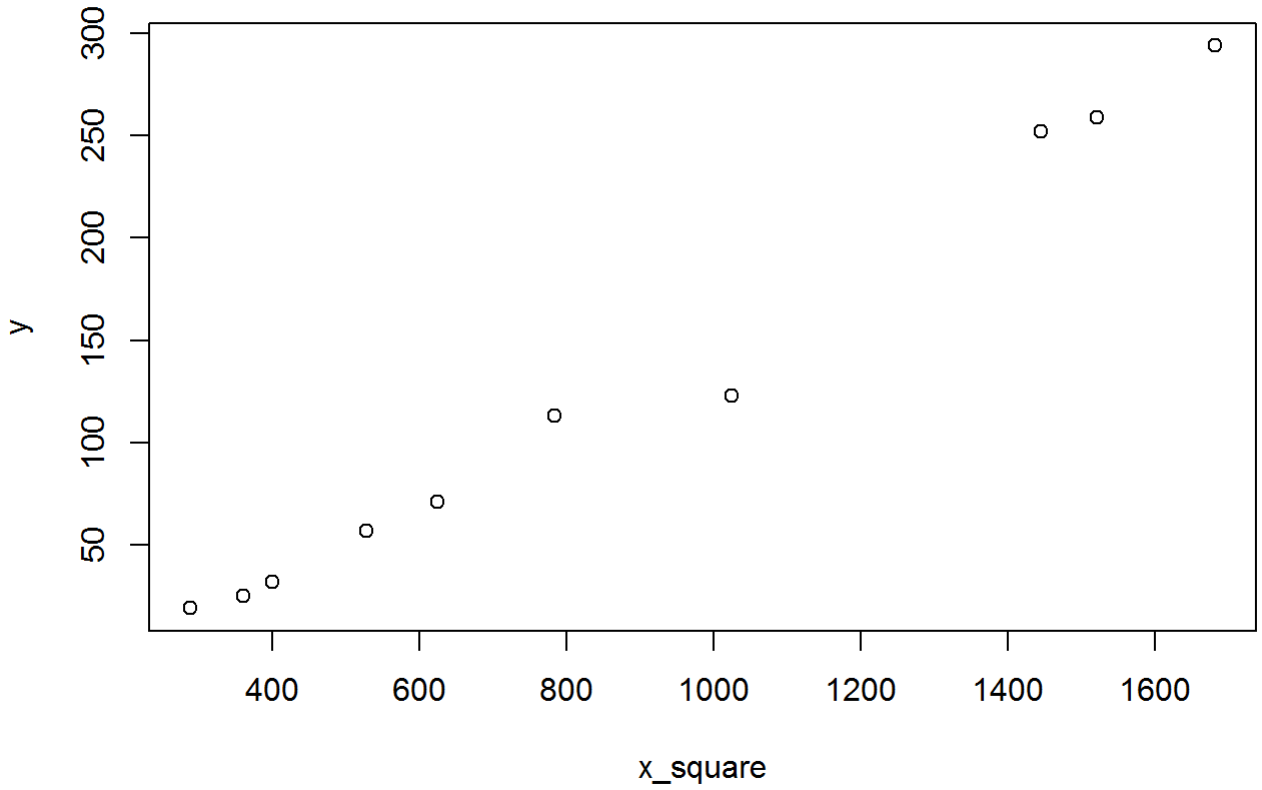
a. Consider two separate assumptions, allowing each to lead to a model. Completely analyze each model.

b. Assume that all trees are right-circular cylinders and are approximately the same height.

If all trees have the same height, so we can consider h to be a constant. Based on the formula for right-circular cylinder $V = \pi r^2 h$, we are able to tell that $V \propto r^2$ or $V \propto d^2$. In this case, we have $y \propto x^2$. Based on the linear model (intercept was forced to be 0), the best fit line is y = 0.158 * x^2, where k = 0.158.

```
x <- c(17, 19, 20, 23, 25, 28, 32, 38, 39, 41)
y <- c(19, 25, 32, 57, 71, 113, 123, 252, 259, 294)

x_square <- x ^ 2
plot(x_square, y)
```



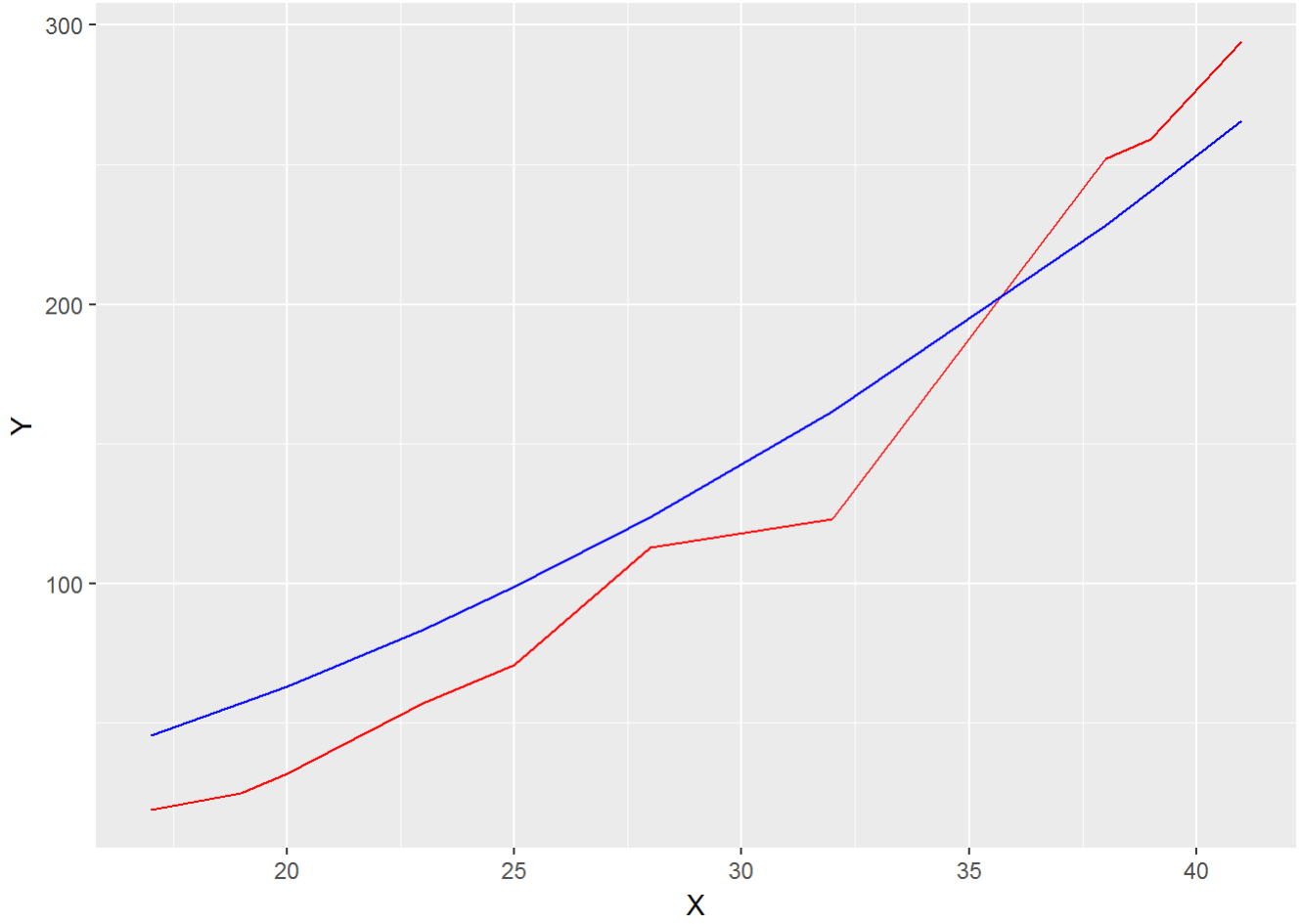

```
m2 <- lm(y ~ x_square + 0)
summary(m2)
```

```
##
## Call:
## lm(formula = y ~ x_square + 0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38.71 -30.30 -26.59  11.40  28.54
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## x_square 0.157919   0.009181    17.2 3.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.93 on 9 degrees of freedom
## Multiple R-squared:  0.9705, Adjusted R-squared:  0.9672
## F-statistic: 295.8 on 1 and 9 DF,  p-value: 3.418e-08
```

```
df2 <- data.frame(X = x, Y = y, Y_head = 0.158 * x_square)
head(df2)
```

```
##    X   Y  Y_head
## 1 17  19  45.662
## 2 19  25  57.038
## 3 20  32  63.200
## 4 23  57  83.582
## 5 25  71  98.750
## 6 28 113 123.872
```

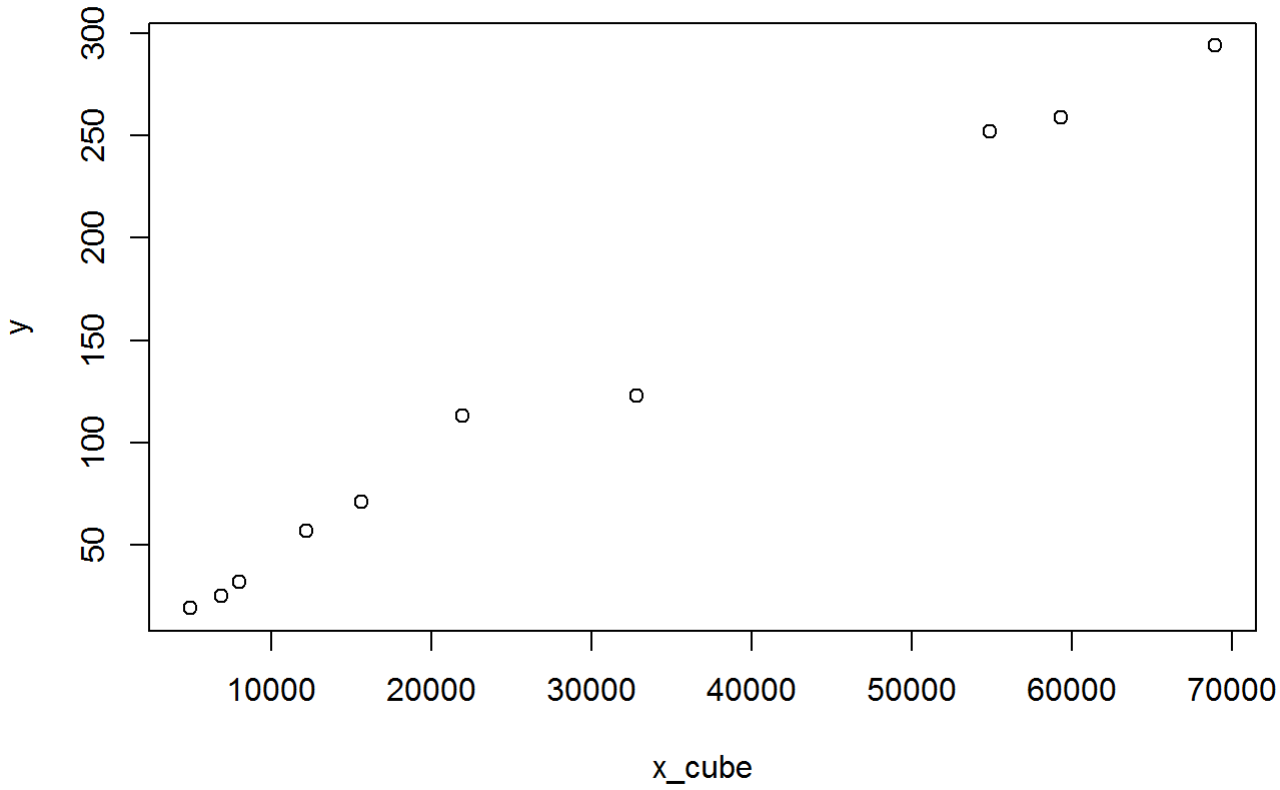
```
ggplot() + geom_line(data = df2, aes(x = X, y = Y), color = "red") + geom_line(data = df2, aes(x
 = X, y = Y_head), color = "blue")
```



ii. Assume that all trees are right-circular cylinders and that the height of the tree is proportional to the diameter.

Since $h \propto d$, we know $ hr$. According to the same formular $V = \pi r^2 h$, $V \propto r^3$. In the other words, $y^{ 3 } $. The best fit line will become y = 0.004362 * x^3, where k = 0.004362.

```
x_cube <- x ^ 3
plot(x_cube, y)
```



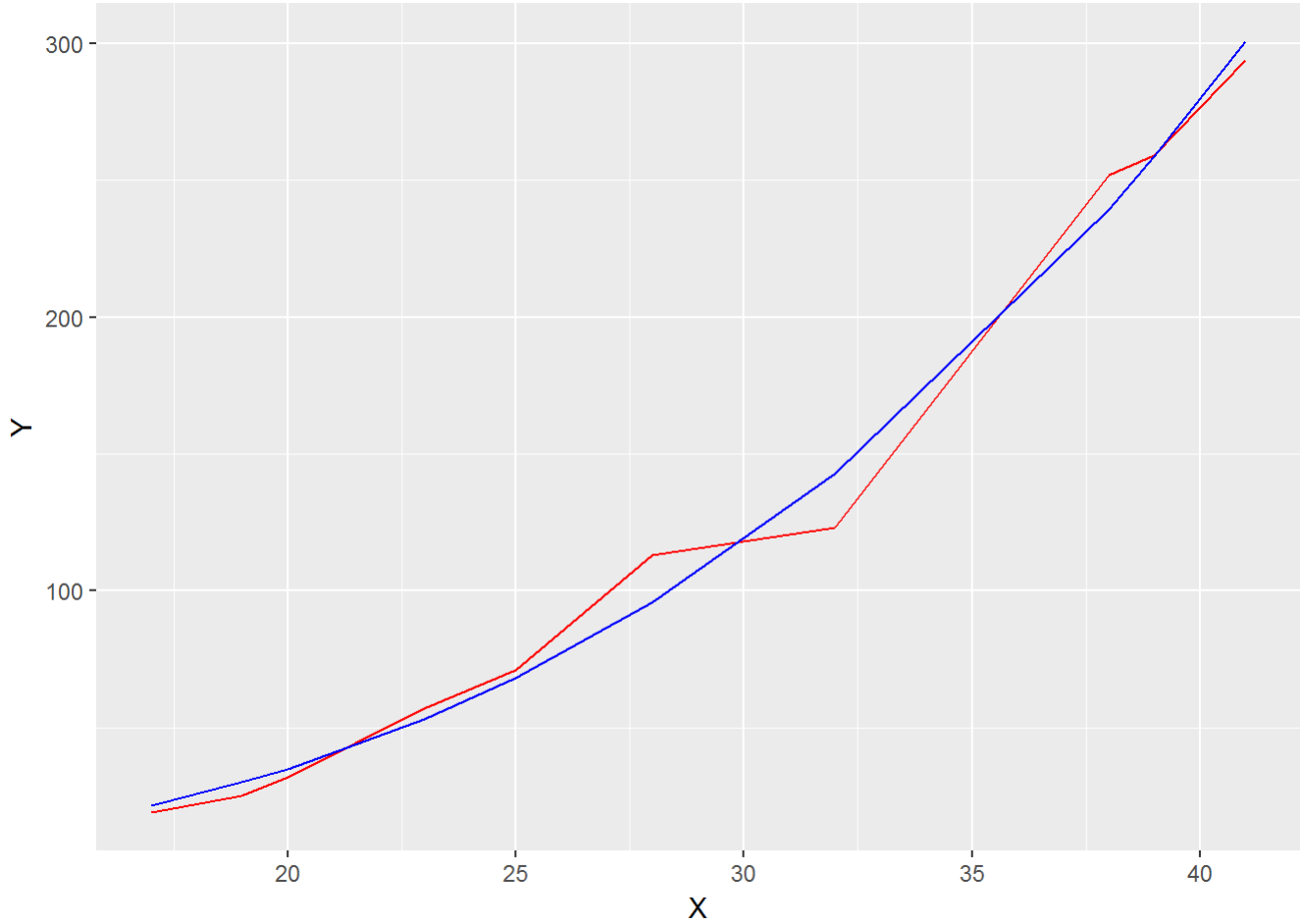

```
m3 <- lm(y ~ x_cube + 0)
summary(m3)
```

```
##
## Call:
## lm(formula = y ~ x_cube + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.935  -4.413  -1.091   3.656  17.245
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## x_cube 4.362e-03  8.938e-05    48.8 3.19e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.33 on 9 degrees of freedom
## Multiple R-squared:  0.9962, Adjusted R-squared:  0.9958
## F-statistic:  2382 on 1 and 9 DF,  p-value: 3.193e-12
```

```
df3 <- data.frame(X = x, Y = y, Y_head = 0.004362 * x_cube)
head(df3)
```

```
##    X   Y   Y_head
## 1 17  19 21.43051
## 2 19  25 29.91896
## 3 20  32 34.89600
## 4 23  57 53.07245
## 5 25  71 68.15625
## 6 28 113 95.75462
```

```
ggplot() + geom_line(data = df3, aes(x = X, y = Y), color = "red") + geom_line(data = df3, aes(x
 = X, y = Y_head), color = "blue")
```



b. Which model appears to be better? Why? Justify your conclusions.

The second model appears to be better, because the estimated line more closed matches the actual line. The reason for that I think, the second assumption makes much more sense than the first one. In reality, the height of trees should be proportional to the diameter of the trees.

Page 99: #3. Discuss several factors that were completely ignored in our analysis of the gasoline mileage problem.

I think the analysis ignore some factors such as the power for ignition and the power for maintain air conditioning. If drivers have heavy brake frequently, it will waste energy too. These factors, which are ignored in the analysis, will all affect fuel mileage