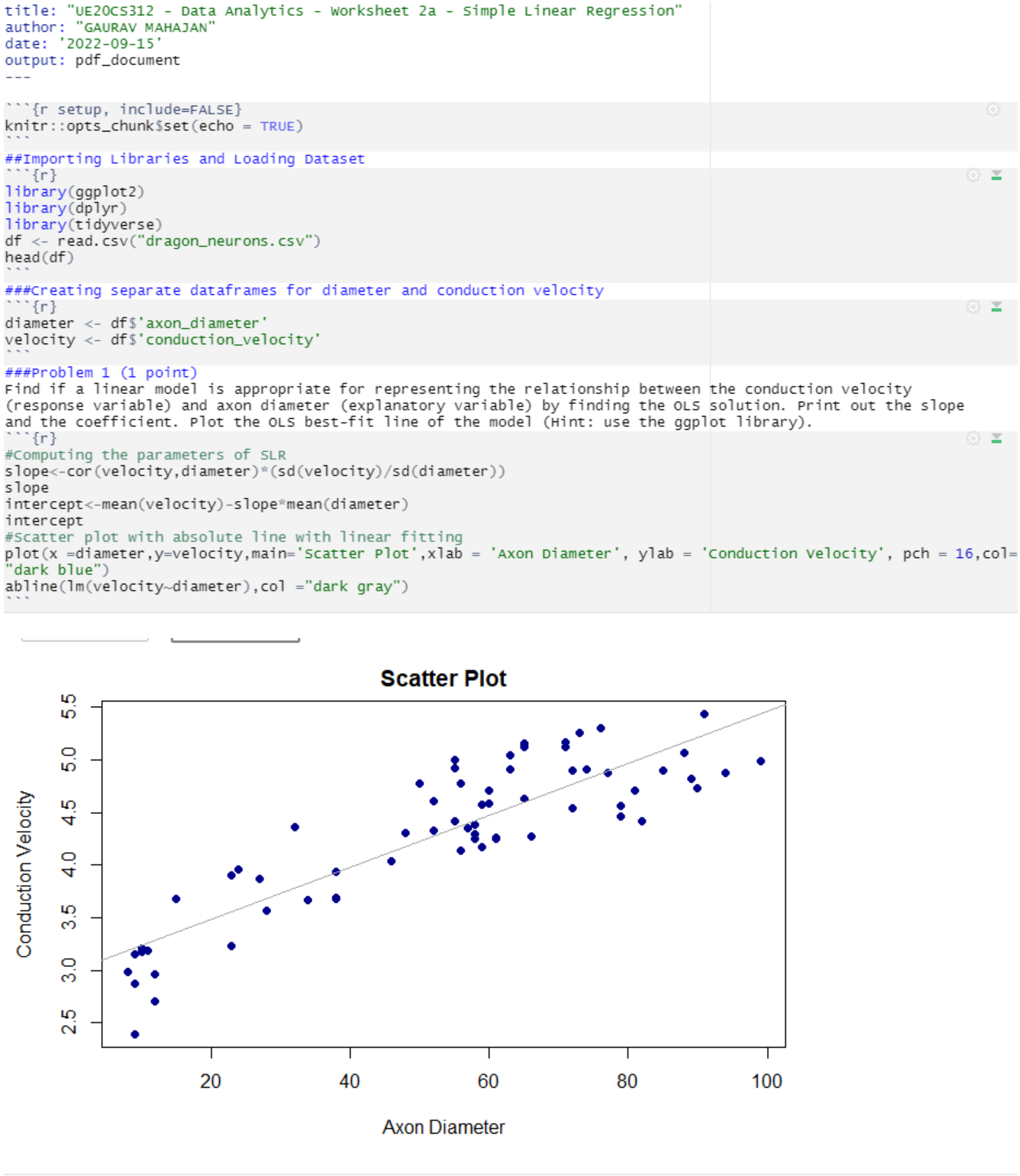
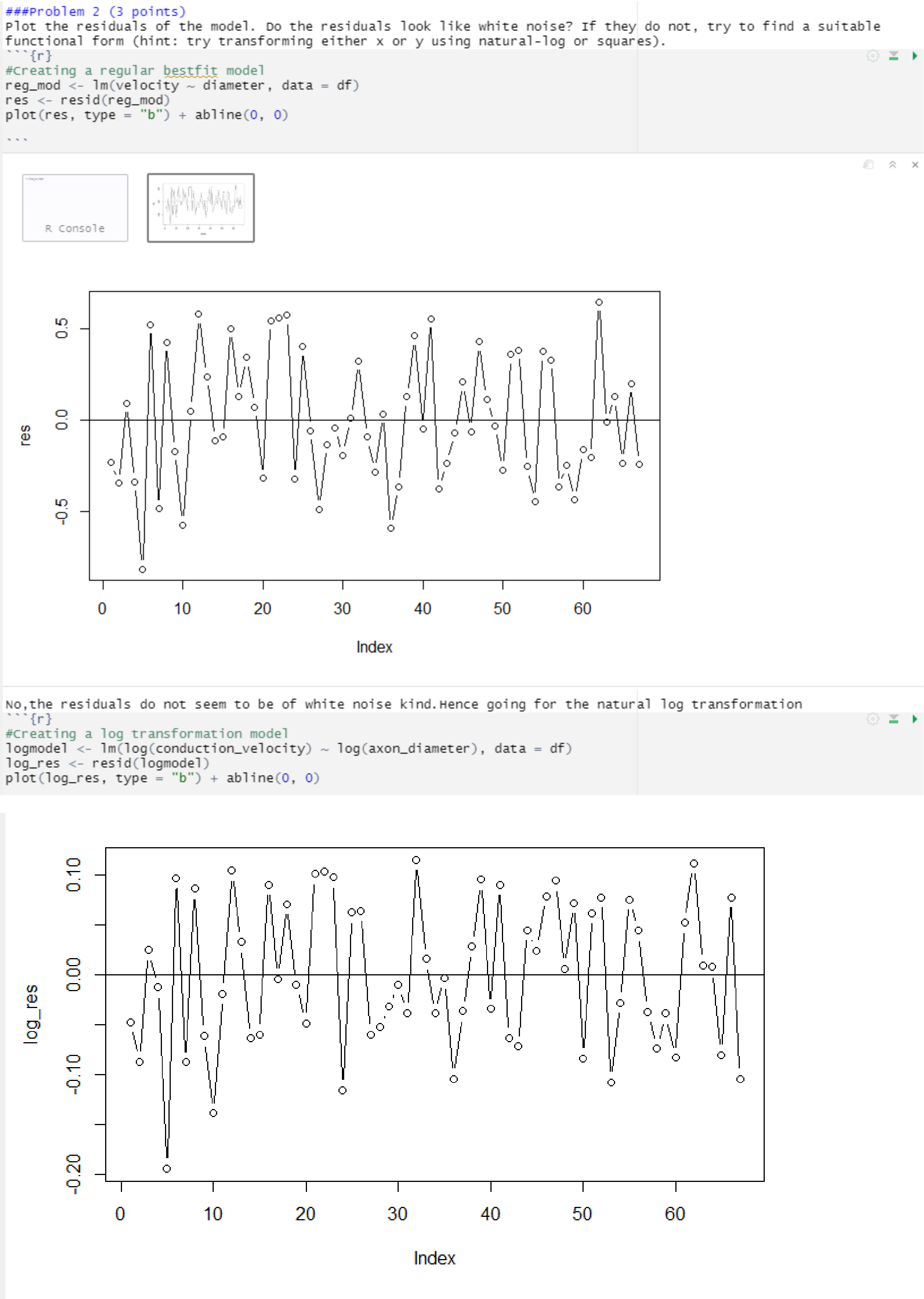
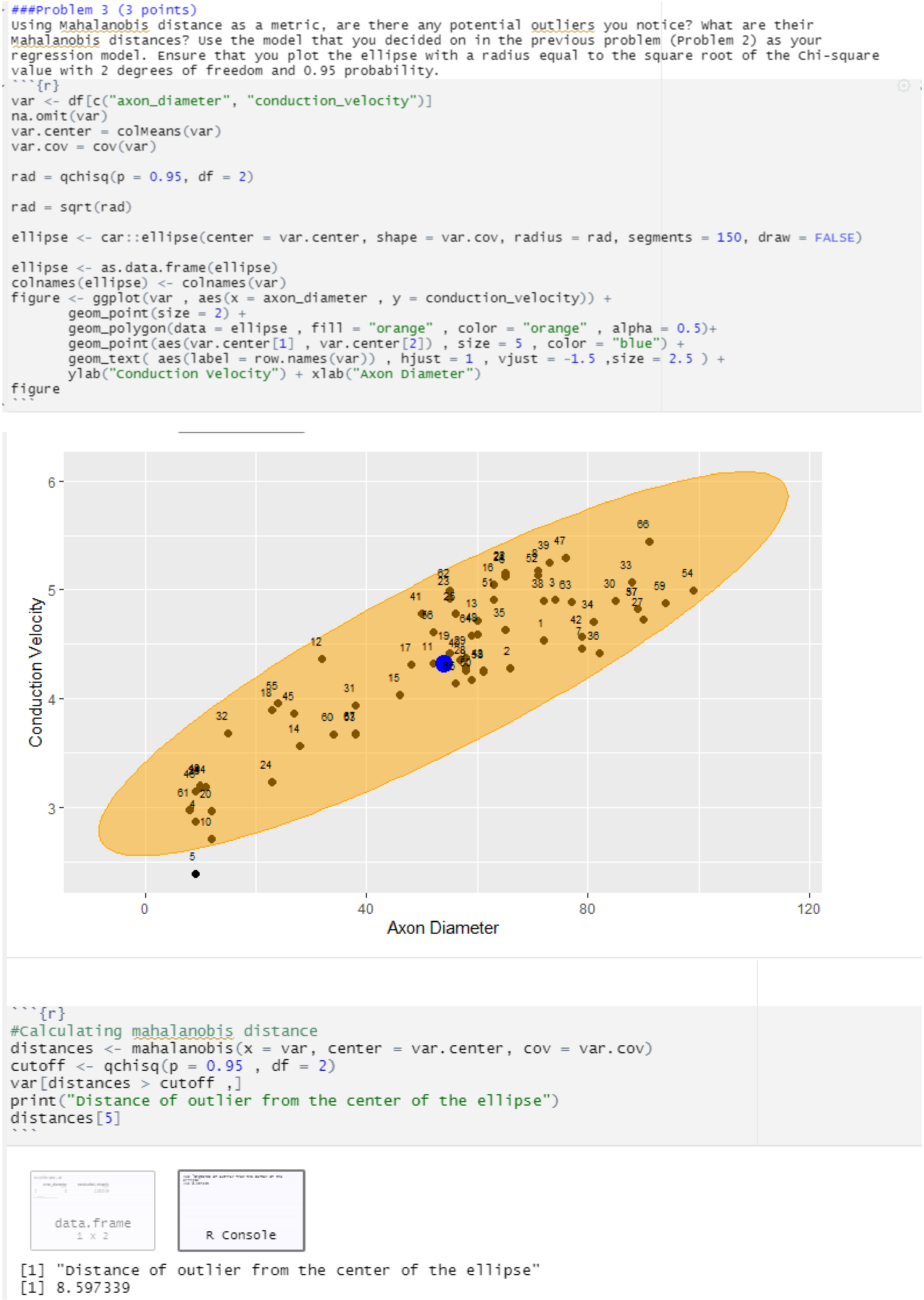


~~~
title: "UE20CS312 - Data Analytics - Worksheet 2a - Simple Linear Regression"
author: "GAURAV MAHAJAN"
date: '2022-09-15'
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~
##Importing Libraries and Loading Dataset
```{r}
library(ggplot2)
library(dplyr)
library(tidyverse)
df <- read.csv("dragon_neurons.csv")
head(df)
```
###Creating separate dataframes for diameter and conduction velocity
```{r}
diameter <- df$'axon_diameter'
velocity <- df$'conduction_velocity'
```
###Problem 1 (1 point)
Find if a linear model is appropriate for representing the relationship between the conduction velocity
(response variable) and axon diameter (explanatory variable) by finding the OLS solution. Print out the slope
and the coefficient. Plot the OLS best-fit line of the model (Hint: use the ggplot library).
```{r}
#Computing the parameters of SLR
slope<-cor(velocity,diameter)*(sd(velocity)/sd(diameter))
slope
intercept<-mean(velocity)-slope*mean(diameter)
intercept
#Scatter plot with absolute line with linear fitting
plot(x =diameter,y=velocity,main='Scatter Plot',xlab = 'Axon Diameter', ylab = 'Conduction Velocity', pch = 16,col=
"dark blue")
abline(lm(velocity~diameter),col ="dark gray")
```
```
```



~~~
###Problem 2 (3 points)
Plot the residuals of the model. Do the residuals look like white noise? If they do not, try to find a suitable
functional form (hint: try transforming either x or y using natural-log or squares).
```{r}
#Creating a regular bestfit model
reg_mod <- lm(velocity ~ diameter, data = df)
res <- resid(reg_mod)
plot(res, type = "b") + abline(0, 0)

```
~~~

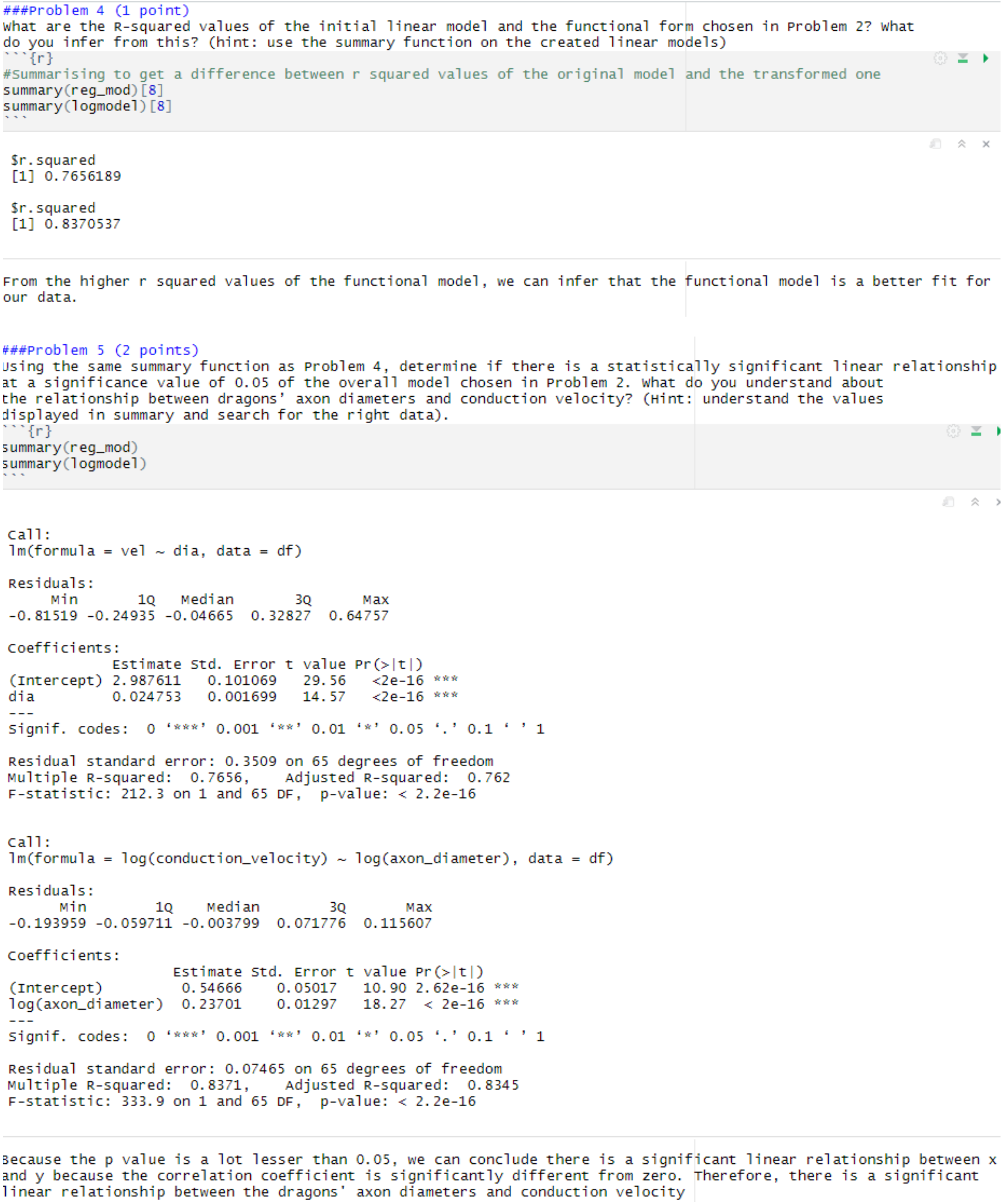
~~~
No,the residuals do not seem to be of white noise kind.Hence going for the natural log transformation
```{r}
#Creating a log transformation model
logmodel <- lm(log(conduction_velocity) ~ log(axon_diameter), data = df)
log_res <- resid(logmodel)
plot(log_res, type = "b") + abline(0, 0)
~~~



~~~
###Problem 3 (3 points)
Using Mahalanobis distance as a metric, are there any potential outliers you notice? What are their
Mahalanobis distances? Use the model that you decided on in the previous problem (Problem 2) as your
regression model. Ensure that you plot the ellipse with a radius equal to the square root of the Chi-square
value with 2 degrees of freedom and 0.95 probability.
```{r}
var <- df[c("axon_diameter", "conduction_velocity")]
na.omit(var)
var.center = colMeans(var)
var.cov = cov(var)

rad = qchisq(p = 0.95, df = 2)

rad = sqrt(rad)

ellipse <- car::ellipse(center = var.center, shape = var.cov, radius = rad, segments = 150, draw = FALSE)

ellipse <- as.data.frame(ellipse)
colnames(ellipse) <- colnames(var)
figure <- ggplot(var , aes(x = axon_diameter , y = conduction_velocity)) +
       geom_point(size = 2) +
       geom_polygon(data = ellipse , fill = "orange" , color = "orange" , alpha = 0.5)+
       geom_point(aes(var.center[1] , var.center[2]) , size = 5 , color = "blue") +
       geom_text( aes(label = row.names(var)) , hjust = 1 , vjust = -1.5 ,size = 2.5 ) +
       ylab("Conduction Velocity") + xlab("Axon Diameter")
figure
```
~~~

~~~
```{r}
#Calculating mahalanobis distance
distances <- mahalanobis(x = var, center = var.center, cov = var.cov)
cutoff <- qchisq(p = 0.95 , df = 2)
var[distances > cutoff ,]
print("Distance of outlier from the center of the ellipse")
distances[5]
```
~~~

```
[1] "Distance of outlier from the center of the ellipse"
[1] 8.597339
```
]



~~~
###Problem 4 (1 point)
What are the R-squared values of the initial linear model and the functional form chosen in Problem 2? What
do you infer from this? (hint: use the summary function on the created linear models)
```{r}
#Summarising to get a difference between r squared values of the original model and the transformed one
summary(reg_mod)[8]
summary(logmodel)[8]
~~~
```

```
$r.squared
[1] 0.7656189

$r.squared
[1] 0.8370537
```

From the higher r squared values of the functional model, we can infer that the functional model is a better fit for our data.

```
###Problem 5 (2 points)
Using the same summary function as Problem 4, determine if there is a statistically significant linear relationship
at a significance value of 0.05 of the overall model chosen in Problem 2. What do you understand about
the relationship between dragons' axon diameters and conduction velocity? (Hint: understand the values
displayed in summary and search for the right data).
```{r}
summary(reg_mod)
summary(logmodel)
```
```

```
Call:
lm(formula = vel ~ dia, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.81519 -0.24935 -0.04665  0.32827  0.64757

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.987611   0.101069   29.56   <2e-16 ***
dia         0.024753   0.001699   14.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3509 on 65 degrees of freedom
Multiple R-squared:  0.7656,	Adjusted R-squared:  0.762 
F-statistic: 212.3 on 1 and 65 DF,  p-value: < 2.2e-16

Call:
lm(formula = log(conduction_velocity) ~ log(axon_diameter), data = df)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.193959 -0.059711 -0.003799  0.071776  0.115607

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         0.54666    0.05017   10.90 2.62e-16 ***
log(axon_diameter)  0.23701    0.01297   18.27  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07465 on 65 degrees of freedom
Multiple R-squared:  0.8371,	Adjusted R-squared:  0.8345 
F-statistic: 333.9 on 1 and 65 DF,  p-value: < 2.2e-16
```

Because the p value is a lot lesser than 0.05, we can conclude there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero. Therefore, there is a significant linear relationship between the dragons' axon diameters and conduction velocity
```