UE20CS312 - Data Analytics - Worksheet 2a - Simple Linear Regression

PES University

Reshmi Pradeep, Dept. of CSE - PES2UG20CS270

2022-09-10

```
###PROBLEM 1
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## 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()
```

```
df<-read_csv('dragon_neurons.csv')
```

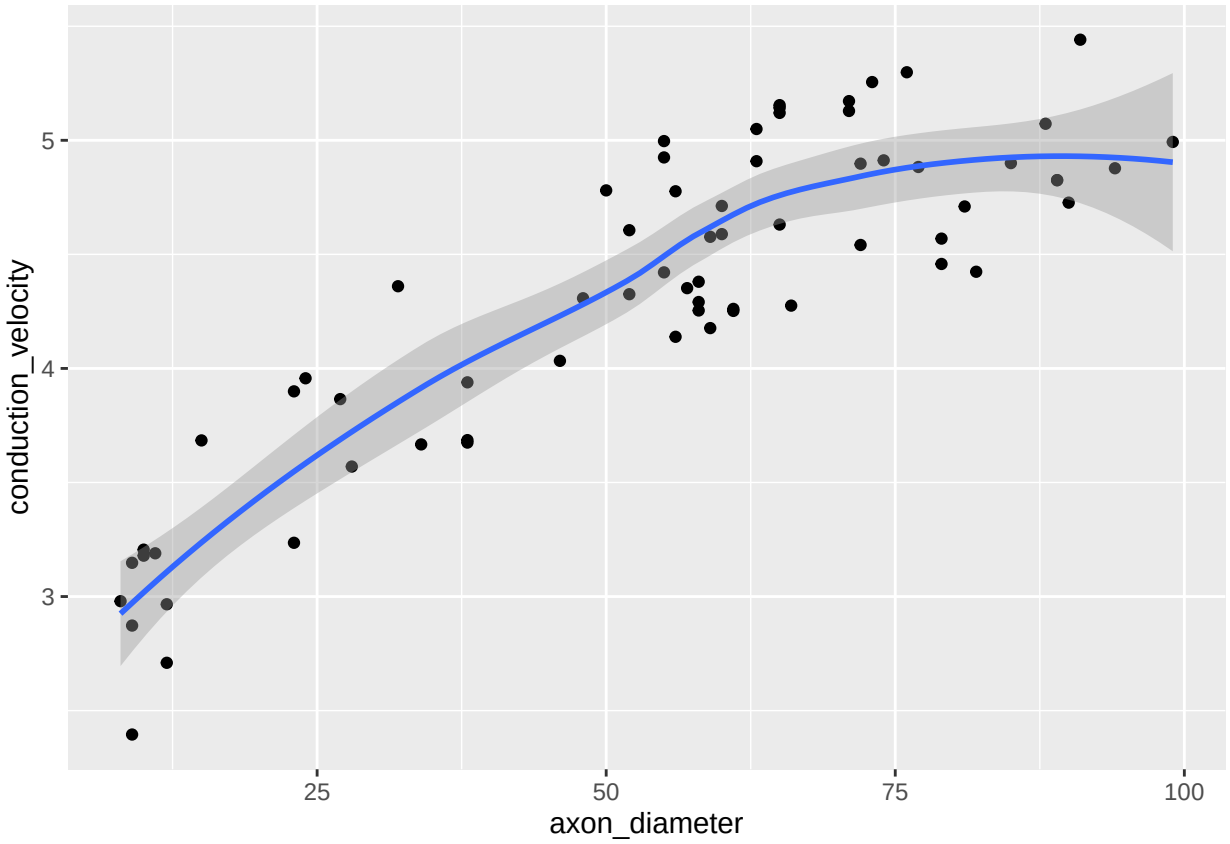
```
## New names:
## Rows: 67 Columns: 4
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (3): ...1, axon_diameter, conduction_velocity lgl (1): ...4
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
## * `` -> `...4`
```

```
head(df)
```

```
## # A tibble: 6 x 4
##    ...1 axon_diameter conduction_velocity ...4
##   <dbl>         <dbl>               <dbl> <lgl>
## 1     0            72                4.54 NA
## 2     1            66                4.28 NA
## 3     2            74                4.91 NA
## 4     3             9                2.87 NA
## 5     4             9                2.40 NA
## 6     5            65                5.12 NA
```

```
ggplot(data=df,mapping=aes(x=axon_diameter,y=conduction_velocity))+ geom_point()+ stat_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
#graph shows linear relationship

cor(df$axon_diameter,df$conduction_velocity)
```

```
## [1] 0.8749965
```

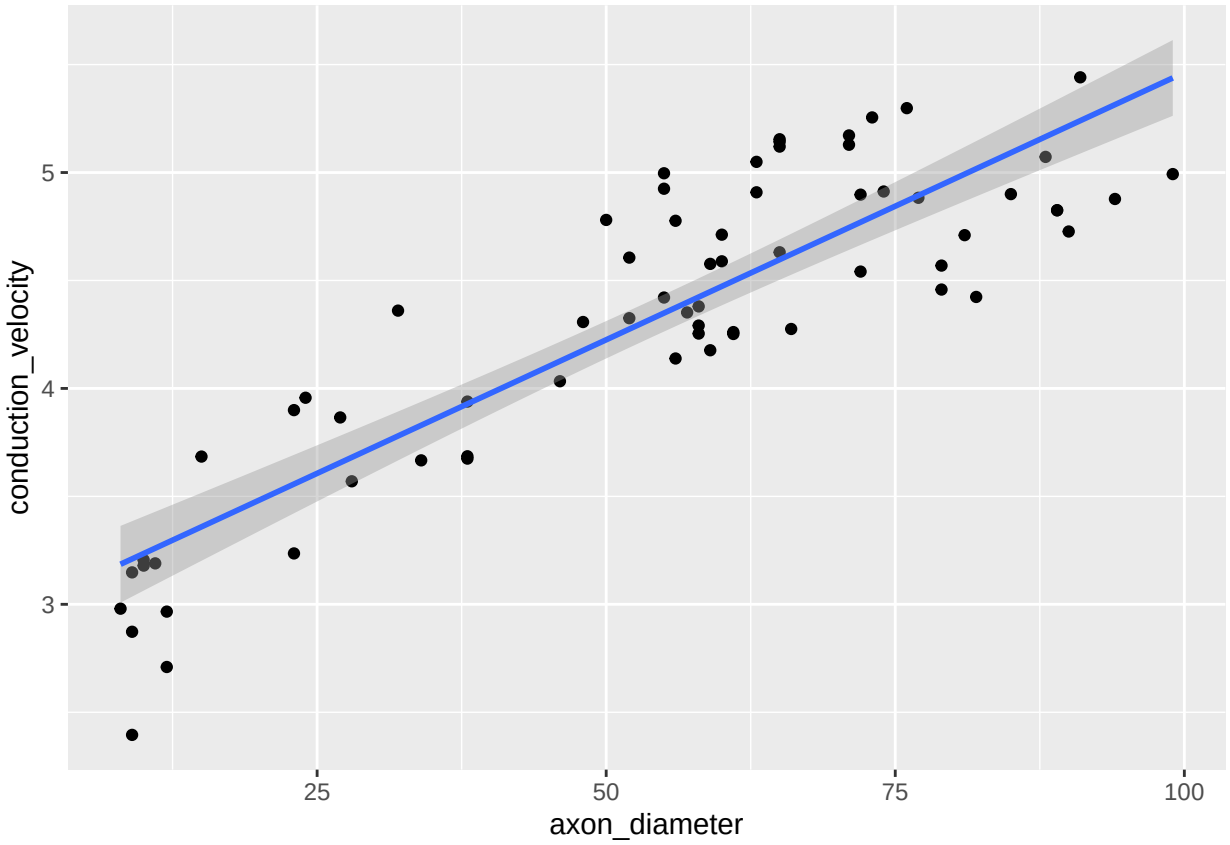
```
#correlation also indicates linear relationship

model<-lm(conduction_velocity ~ axon_diameter,data =df)
print(model)
```

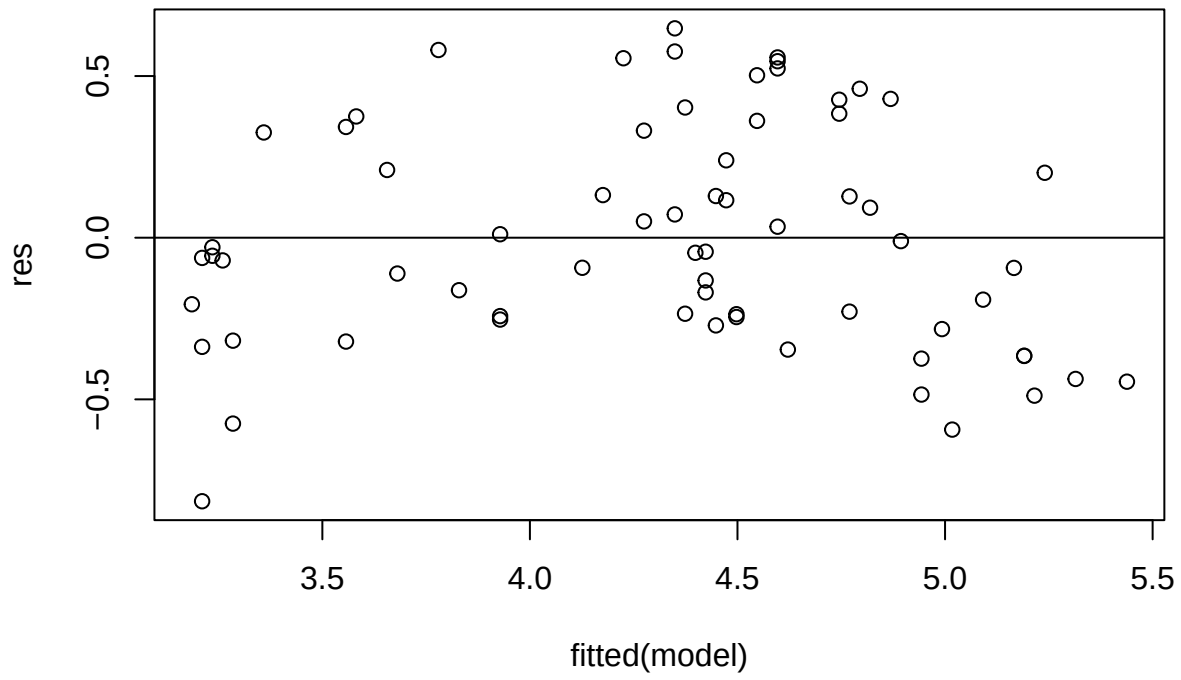
```
##
## Call:
## lm(formula = conduction_velocity ~ axon_diameter, data = df)
##
## Coefficients:
##   (Intercept)  axon_diameter
##       2.98761        0.02475
```

```
#plotting best-fit
ggplot(df,aes(axon_diameter,conduction_velocity))+geom_point()+stat_smooth(method=lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

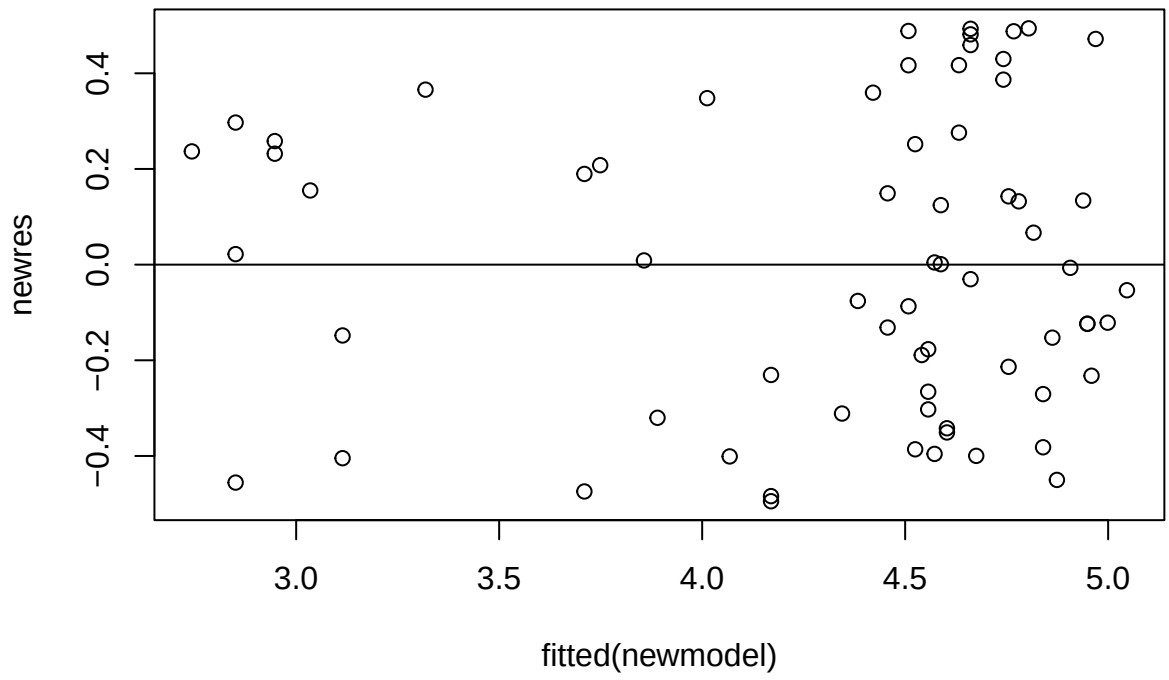



```
###PROBLEM 2
res<-resid(model)
plot(fitted(model), res)
abline(0,0)
```




```
#linear model is not appropriate for modeling this data as the points are not scattered randomly around

df$ad=log(df$axon_diameter) #new functional form
newmodel<-lm(conduction_velocity ~ ad ,data=df)
newres<-resid(newmodel)
plot(fitted(newmodel), newres)
abline(0,0)
```

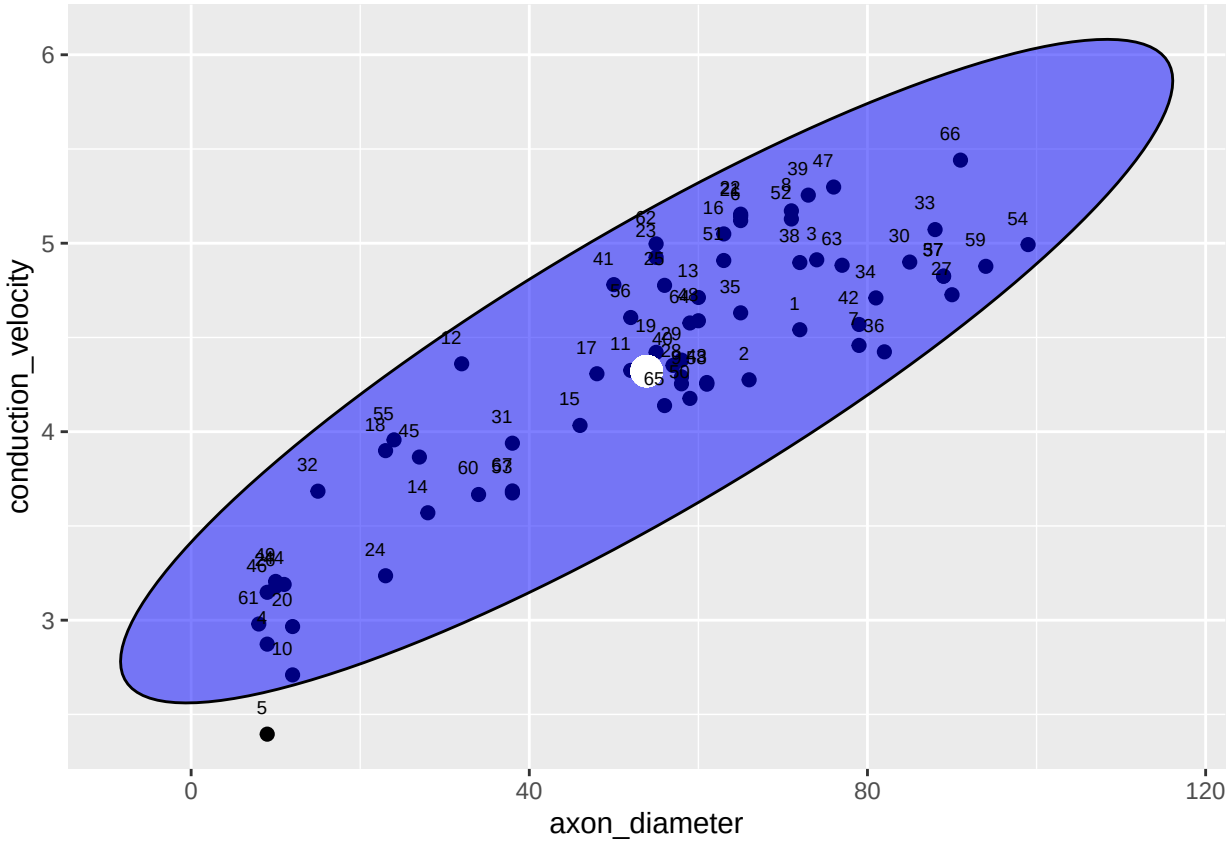



```
###Problem 3
newdf=df[c("axon_diameter","conduction_velocity")]
newdf=na.omit(newdf)
newdf.center=colMeans(newdf)
newdf.cov=cov(newdf)

el_radius=qchisq(p=0.95, df=ncol(newdf))
el_radius=sqrt(el_radius)
ellipse<-car::ellipse(center=newdf.center, shape=newdf.cov, radius=el_radius, segments=150, draw=FALSE)
ellipse<-as.data.frame(ellipse)
colnames(ellipse)<-colnames(newdf)

fig<-ggplot(newdf, aes(x=axon_diameter, y=conduction_velocity)) +geom_point(size=2) +geom_polygon(data=

print(fig)
```




```
# we can see from the ellipse plotted that there is one outlier - 5

###Problem 4
summary(model)
```

```
##
## Call:
## lm(formula = conduction_velocity ~ axon_diameter, 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 ***
## axon_diameter 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
```

```r
summary(newmodel)
```

```
##
## Call:
## lm(formula = conduction_velocity ~ ad, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49467 -0.26822 -0.00671  0.25506  0.49396
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.83911    0.21037   3.989 0.000171 ***
## ad           0.91559    0.05439  16.833  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3131 on 65 degrees of freedom
## Multiple R-squared:  0.8134, Adjusted R-squared:  0.8105
## F-statistic: 283.3 on 1 and 65 DF,  p-value: < 2.2e-16
```

```r
#r-squared shows how well the regression model explains observed data.
#Since the r-squared value for the second model is close to 1, a large proportion of the variability ha
#Hence, we can infer that the second model is better than the first model.

###Problem 5
#As the p-value is much less than 0.05, we reject the null hypothesis,
#there isn't a statistically significant linear relationship at a significance value of 0.05
#and Axon diameter has a significant impact on conduction velocity.
```