# PES University, Bangalore

## Established under Karnataka Act No. 16 of 2013


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

Dept. of CSE, Ishita Bharadwaj - PES1UG20CS648

Collaborated with Hita - PES1UG20CS645


# Simple Linear Regression

Simple linear regression is a statistical technique for finding the existence of an association relationship between a dependent variable and an independent variable.

## Data reading

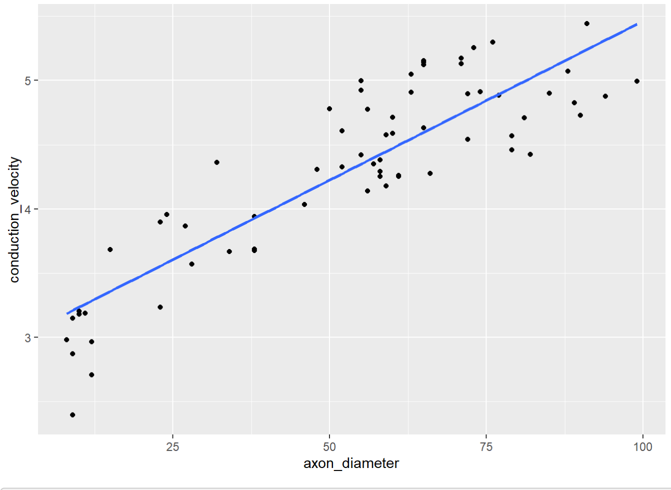
```
library(ggplot2)
dragon_neurons <- read.csv('dragon_neurons.csv')
head(dragon_neurons)
```

```
##   X axon_diameter conduction_velocity X.1
## 1 0            72            4.541130  NA
## 2 1            66            4.275300  NA
## 3 2            74            4.912093  NA
## 4 3             9            2.872806  NA
## 5 4             9            2.395194  NA
## 6 5            65            5.120160  NA
```

## Solution 1

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).

```
# lm - linear model
ols.lm<-lm(formula=conduction_velocity ~ axon_diameter, data=dragon_neurons)
ggplot(dragon_neurons, aes(x=axon_diameter, y=conduction_velocity)) +geom_point() + geom_smooth(method='lm', se=FALSE)
```




```
print(ols.lm)
```

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

**The OLS for the the linear model between the conduction velocity and axon diameter is plotted above.The slope for the model is 0.02475 and the intercept is 2.98761.**
**This Linear model has a slight non-uniformity in the spread of data points about the best fit line which can't be explained.**
**Thus, I have explored other functional forms in search of a model which produces uniform variance about the best fit line.**

## Solution 2

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).

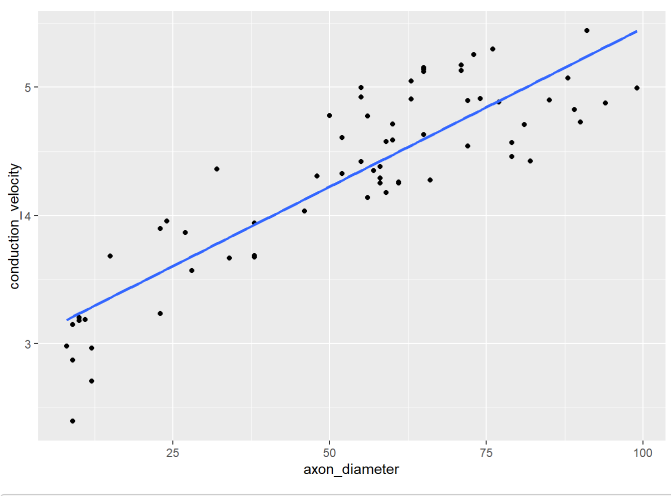
**Linear Model**

```
ols.lm<-lm(formula=conduction_velocity ~ axon_diameter, data=dragon_neurons)
ols.lm
```

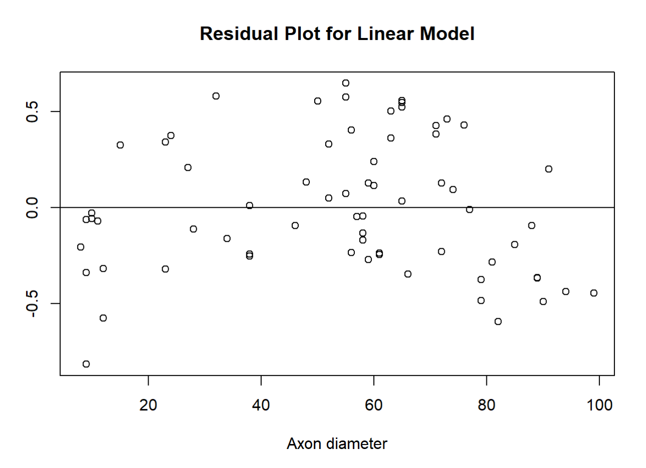
```
##
## Call:
## lm(formula = conduction_velocity ~ axon_diameter, data = dragon_neurons)
##
## Coefficients:
##   (Intercept)  axon_diameter
##       2.98761        0.02475
```

```
ggplot(dragon_neurons, aes(x=axon_diameter, y=conduction_velocity)) +geom_point() + geom_smooth(method='lm', se=FALSE)
```

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




```
ols.res<-resid(ols.lm)
# plot(dragon_neurons)
plot(dragon_neurons$axon_diameter, ols.res,ylab='Residuals', xlab='Axon diameter', main='Residual Plot for Linear Model')
abline(0,0)
```




```
summary(ols.lm)
```

```
##
## Call:
## lm(formula = conduction_velocity ~ axon_diameter, data = dragon_neurons)
##
## 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
```

**The Multiple R-squared value is 0.7656 for linear model. Since the coefficient of determination isn't high enough, the linear model doesn't fit the data best.**
**Also from the residual plot, here is slight non-uniform deviation of residuals about the mean(0). Thus, the residuals are not following a perfect normal distribution.**
**This variance in the residuals is unexplanable. Thus, there is presence of some white noise in the data.**

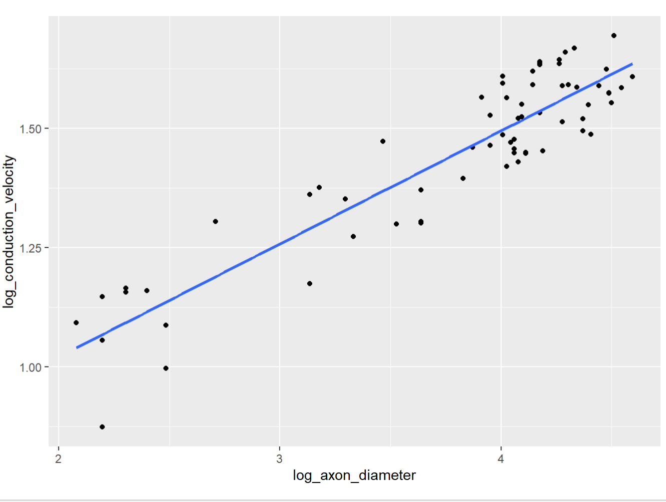
**Logarithmic Model**

```
dragon_neurons$log_axon_diameter<-NA
dragon_neurons$log_conduction_velocity<-NA
dragon_neurons$log_axon_diameter<-log(dragon_neurons$axon_diameter)
dragon_neurons$log_conduction_velocity<-log(dragon_neurons$conduction_velocity)

ols.lm.log<-lm(formula=log_conduction_velocity ~ log_axon_diameter, data=dragon_neurons)
ols.lm.log
```

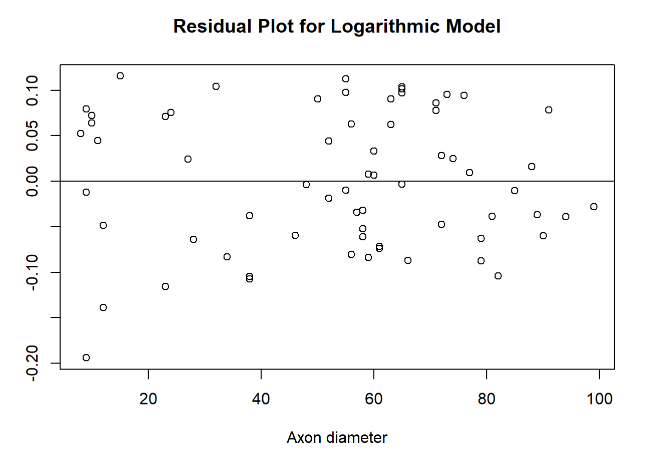
```
##
## Call:
## lm(formula = log_conduction_velocity ~ log_axon_diameter, data = dragon_neurons)
##
## Coefficients:
##       (Intercept)  log_axon_diameter
##            0.5467             0.2370
```

```
ggplot(dragon_neurons, aes(x=log_axon_diameter, y=log_conduction_velocity)) +geom_point() + geom_smooth(method='lm', se=FALSE)
```

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




```
ols.res.log<-resid(ols.lm.log)
# plot(dragon_neurons)
plot(dragon_neurons$axon_diameter, ols.res.log,ylab='Residuals', xlab='Axon diameter', main='Residual Plot for Logarithmic Model')
abline(0,0)
```




```
summary(ols.lm.log)
```

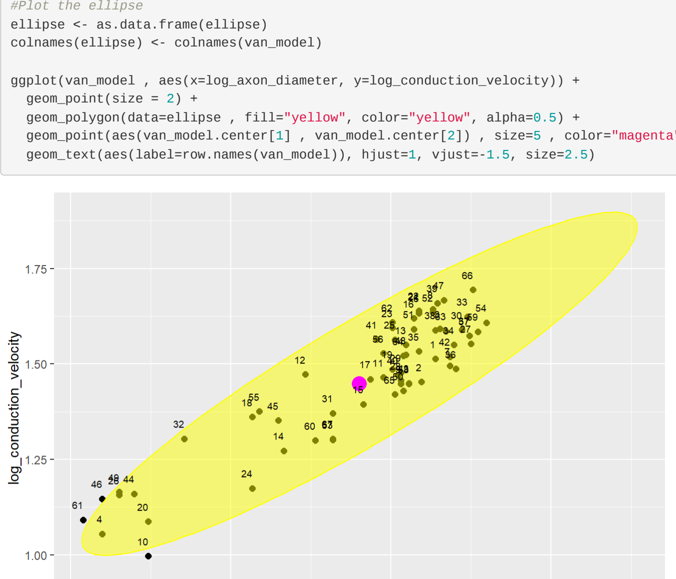
```
##
## Call:
## lm(formula = log_conduction_velocity ~ log_axon_diameter, data = dragon_neurons)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.193959 -0.059711 -0.003799  0.071776  0.115607
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (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
```

**The Multiple R-squared value is 0.8371 for logarithmic model. Thus this model better fits the data as compared to the linear model.**
**Moreover, from the above residual plot, it is evident that the log-vs-log model for the explanatory variable and response variable provide for uniform deviation of residuals from the mean 0 line.**
**There is an approximately constant variance in the residuals, i.e., The residuals follow a normal distribution. Hence, log-vs-log model will be taken up for further analysis.**

## Solution 3

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.

```
van_model <- dragon_neurons[c('log_axon_diameter' ,'log_conduction_velocity')]
# Find the center and covariance
van_model.center<-NA
van_model.center <- colMeans(van_model)


# ncol(van_model) ==2
van_model.cov <- cov(van_model)

# Find the radius of the ellipse
van_model.rad <- sqrt(qchisq(p=0.95, df=ncol(van_model)))

# Find the ellipse coordinates
ellipse <- car::ellipse(center=van_model.center, shape=van_model.cov, radius=van_model.rad, segments=150, draw=FALSE)

#Plot the ellipse
ellipse <- as.data.frame(ellipse)
colnames(ellipse) <- colnames(van_model)

ggplot(van_model , aes(x=log_axon_diameter, y=log_conduction_velocity)) +
  geom_point(size = 2) +
  geom_polygon(data=ellipse , fill="yellow", color="yellow", alpha=0.5) +
  geom_point(aes(van_model.center[1] , van_model.center[2]) , size=5 , color="magenta") +
  geom_text(aes(label=row.names(van_model)), hjust=1, vjust=-1.5, size=2.5)
```




**Finding outliers:**

```
#van_model$distances<-NA
#van_model$cutoff<-NA
print(van_model.center)
```

```
##       log_axon_diameter log_conduction_velocity
##                3.803262                1.448066
```

```
distances<-mahalanobis(x=van_model, center=van_model.center,
                       cov=van_model.cov)
cutoff<-qchisq(p=0.95, df=ncol(van_model))
van_model$distances<-distances
van_model$cutoff<-cutoff
van_model[van_model$distances>cutoff,]
```

```
##    log_axon_diameter log_conduction_velocity distances   cutoff
## 5           2.197225               0.8734643 11.993207 5.991465
## 10          2.484907               0.9968085  6.972770 5.991465
## 46          2.197225               1.1467301  6.284871 5.991465
## 61          2.079442               1.0918291  6.419079 5.991465
```

**There are 4 potential outliers - 4 values whose mahalanobis distance is greater than the cutoff. Their values are displayed in the output cell above.**

**Cutoff: 5.991465**
**Mahalanobis distance of 4 outliers:**
**11.993207**
**6.972770**
**6.284871**
**6.419079**

## Solution 4

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)

```
summary(ols.lm)
```

```
##
## Call:
## lm(formula = conduction_velocity ~ axon_diameter, data = dragon_neurons)
##
## 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
```

```
summary(ols.lm.log)
```

```
##
## Call:
## lm(formula = log_conduction_velocity ~ log_axon_diameter, data = dragon_neurons)
##
## 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
```

**For initial linear model - Multiple R-squared: 0.7656**
**For logarithmic linear model- Multiple R-squared: 0.8371**
**Coefficient of determination - A higher value of R-squared implies a better fit, a higher proportion of the residual value between the dependant(Y) and independant(X) variable can be explained.**
**R^2 = SSR/SST**
**SST = Sum of squares of total variation.**
**SSR = Sum of squares of variation explained by the regression model.**

**The logarithmic model seems to be a better fit as it has a higher coefficient of determination.**

## Solution 5

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).

```
qf(0.05,1,65,lower.tail = FALSE)
```

```
## [1] 3.98856
```

```
corr_log<-cor(dragon_neurons[c('log_axon_diameter')], dragon_neurons[c('log_conduction_velocity')], method = c("pearson"))
print(corr_log)
```

```
##                   log_conduction_velocity
## log_axon_diameter               0.9149064
```

```
corr<-cor(dragon_neurons[c('axon_diameter')], dragon_neurons[c('conduction_velocity')], method = c("pearson"))
print(corr)
```

```
##               conduction_velocity
## axon_diameter           0.8749965
```

**F statistic of linear model: 212.3**
**F statistic of log-log model: 333.9**
**F critical at 5% significant level, with 1 numerator dof, 65 denominator dof = 3.99856**

**If F statistic > F critical, we reject null hypothesis. F test is a right tailed test. Here, 212.3 and 333.9 >> 3.99856.**
**So the null hypothesis is rejected as the linear relationship is statistically significant for alpha = 0.05.**

**The relationship between dragons' axon diameters and conduction velocity is better fitted with the functional form log(y) = b0 + b1*log(x). The interpretation is given as an expected percentage change in Y when X increases by some percentage. The explanatory variable(Axon diameter) and the outcome variable(Conduction velocity) in the linear logarithmic model we've chosen is highly correlated(0.9149) which means their linear relationship is strong. As the axon diameter increases, they are able to send signals faster. This is because there is less resistance facing the ion flow. Hence, the conduction velocity increases with increase in axon diameter.**

**Thank You!**