# STAT 530 HW1

*Md Muhtasim Billah*

*1/28/2020*

## Question 1.

Write TRUE or FALSE with reasons for the following:

## Answer

**a.** Number on the jerseys of Basketball players is a numerical variable.

**FALSE.**

Reason: The number on the jerseys of basketball players are onlys used to identify the players. It has neither any mathematical meaning nor it can be manipulated, rather it has categorical value. Thus, it can be considered as a categorical variable, not numerical.

**b.** If y follows normal with mean 4 and variance 1, y^2 follows a chi-square distribution with 1 df

**TRUE.**

Reason:

It follows a chi-distribution with 1 dgrees of freedom, but it will be a non-central chi-square as y doesn't follow a standard normal distribution.

**c.** If y1 is a normal with mean 0 and variance 1, and y2 follows chi-square with n-1 degrees of freedom then y1/sqrt(y2/n-1) follows a F distribution with (n-1) degrees of freedom.

**FALSE.**

Reason:

It will be a student's t-distribution with (n-1) df.

**d.** If y3 follows a F distribution then square root of y3 follows t.

**FALSE.**

Reason:

The square of t-distribution follows F distribution but the opposite is not true.

## Question 2.

Consider the data set on lead concentration and mortality of midge flies given in our class website names Hw1-2020.xls. or Hw1-2020.csv (same data set in two formats). Answer the questions based on this data set:

a. Why is mortality considered the response variable? Why is lead concentration the explanatory variable?

b. Plot the relationship. What do you see from the plot?

c. Determine least squares' equation that can be used for predicting mortality.

d. Is the slope significantly greater than 0. Use alpha = .05.

# Answer:

## a.

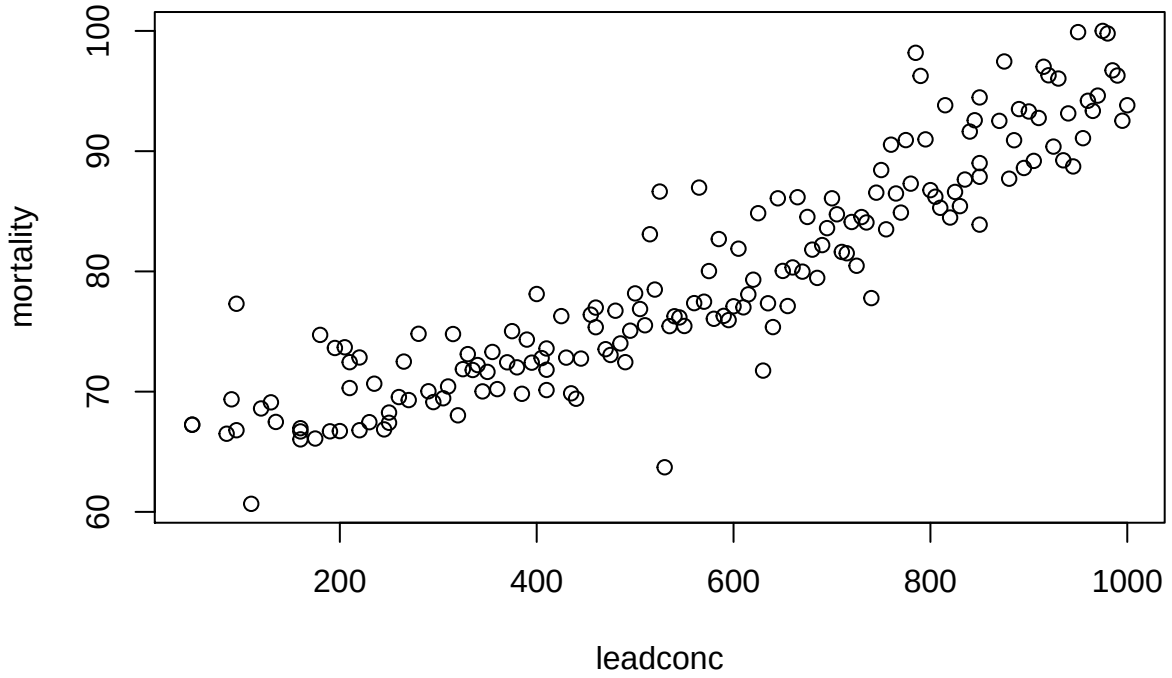
Lead concentration is considered the explanatory variable because it can be easily measured and the mortality rate is directly dependent on the lead concentration.

The mortality rate is the response variable because it is dependent directly on the explanatory varible- the lead concentration and we can make an inference about it based on the explanatory variable.

## b.

Plotting the relationship between lead concentration and mortality rate in R.

```
setwd("/Users/muhtasim/Desktop/STAT530/HWs")
data1=read.csv("HW_1_2020.csv",header=TRUE,sep=",",na.strings=" ")
leadconc=data1[,1]
mortality=data1[,2]
plot(leadconc,mortality)
```




From the plot, it is evident that the mortality rate almost linearly increases with lead concentration.

### c.

The model for lest square method is:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Where, $\beta_0$ is the intercept, $\beta_1$ is the slope and $\varepsilon$ is the random error.
For determining the least squares' equation, we can use the linear regression model in R.

```
fit=lm(mortality~leadconc, data=data1)
coefficients(fit)
```

```
## (Intercept)    leadconc
## 61.23564557  0.03271988
```

From R, we find that the intercept is $\beta_0 = 61.24$ and the positive slope is $\beta_1 = 0.03$. So, the least squares' equation will be,

$$Y = 61.24 + 0.03x$$

### d.

To see if the slope is significantly greater than 0 or not, we can run a summary in R for the linear regression model.

```
summary(fit)
```

```
##
## Call:
## lm(formula = mortality ~ leadconc, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.867  -2.456  -0.684   2.538  12.966
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.23565    0.69579   88.01   <2e-16 ***
## leadconc     0.03272    0.00112   29.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.845 on 169 degrees of freedom
## Multiple R-squared:  0.8346, Adjusted R-squared:  0.8336
## F-statistic: 852.8 on 1 and 169 DF,  p-value: < 2.2e-16
```

Looking at the significance codes in the summary output, it can be concluded that the slope is significantly greater than 0 (using $\alpha = 0.05$). Here, the null hypothesis is, $H_0 : \beta_0 = 0$ and the alternative hypothesis is, $H_1 : \beta_0 \neq 0$. Since, the obtained $p - value = 2.2e - 6 < 0.05$, the null can be rejected. Which means that the slope is significantly greater than zero.

## Question 3.

Why is the regression estimates called the LEAST SQUARE estimates. Comment.

## Answer:

Regression estimates can largely be divided into two categories - least square method and maximum likelihood method. The model for estimating the parameters for both the methods is following:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

The error estimate from this equation is,

$$\varepsilon = Y - (\beta_0 + \beta_1 x)$$

But, $\sum \varepsilon$ will always be zero, so minimization of this quantity doesn't make any sense. So, instead we find the summation of the square of the error given by the following expression.

$$\sum \varepsilon^2 = \sum \{Y - (\beta_0 + \beta_1 x)\}^2$$

That is why the regression estimates are called the LEAST SQUARE estimates. This entitiy is also visible in case of maximum likelihood method. The likelihood function is given as below.

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left[\frac{\{y_i - (\beta_0 + \beta_1 x_i)\}^2}{2\sigma^2}\right]$$

On the right side of the expression, we can see the same entity which is the least squared error estimate and that is why the name LEAST SQUARE is used for the regression estimate.

## Question 4.

Give an example from your discipline where you think linear regression would be appropriate.

## Answer:

Linear regression can be applied in so many scientific fields. I am a mechanical engineering major and heat conduction through material is a common phenomena. As it turns out, with increased thermal conductivity of a material, the heat transfer rate is increased alsmost linearly and linear regression can be fruitfully used to predict the heat transfer rate in a material.