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.

TRUE.

Reason: A numerical or continuous variable is one that may take on any value within a finite or infinite interval. The number on the jerseys of Basketball players can be any numeric value and that's why it is a numerical variable.

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

TRUE.

Reason:

Though chi-square distribution requires standard normal distribution (mean=0 and variance=1), not being so won't affect the degree of freedom. So, for the given case, the df will be 1 but it will be a non-central chi-squre 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.

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

FALSE.

Reason:

It will not necessarily be a t-distribution.

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.

e. Use your model to predict the mortality rate percentage when lead concentration is 225 ppm.

f. Provide 95% confidence intervals for e.

g. Provide 95% prediction intervals for (e)

h. Comment on your findings in (f) and (g).

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.

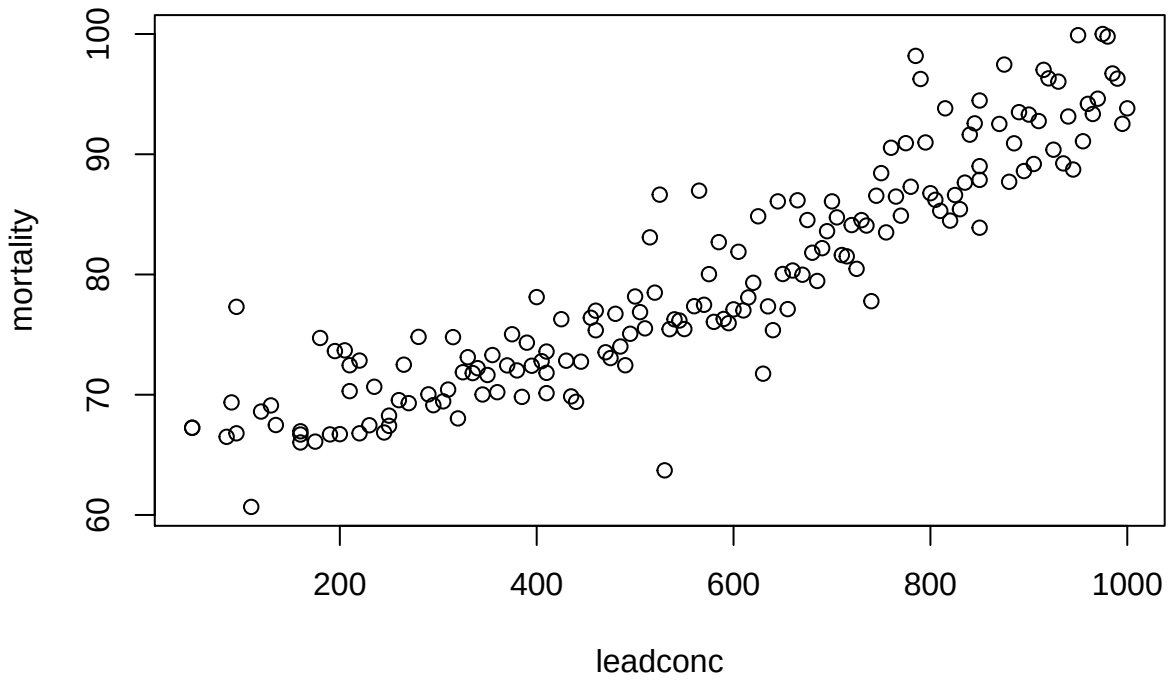
```
setwd("/Users/muhtasim/Desktop/STAT530/HWs")
getwd()
```

```
## [1] "/Users/muhtasim/Desktop/STAT530/HWs"
```

```
data1=read.csv("HW_1_2020.csv",header=TRUE,sep=",",na.strings=" ")
head(data1)
```

```
##   leadconc.in.ppm mortality.percent
## 1              50             67.26
## 2              50             67.24
## 3              85             66.50
## 4              90             69.36
## 5              95             66.79
## 6              95             77.31
```

```
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, β_0 is the intercept, β_1 is the slope and ε is the 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,

$$Y = 61.24 + 0.03x + \varepsilon$$

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 not significantly greater than 0 (using $\alpha = 0.05$)

e.

Using the linear model, the mortality rate percentage can be predicted for lead concentration of 225 ppm.

```
y=mortality
x=leadconc
#prediction of mortality for leadconc of 225 ppm
new=data.frame(x = seq(225,0.05))
predict(lm(y ~ x), new)
```

```
##        1        2        3        4        5        6        7        8
## 68.59762 68.56490 68.53218 68.49946 68.46674 68.43402 68.40130 68.36858
##        9       10       11       12       13       14       15       16
## 68.33586 68.30314 68.27042 68.23770 68.20498 68.17226 68.13954 68.10682
##       17       18       19       20       21       22       23       24
## 68.07410 68.04138 68.00866 67.97594 67.94322 67.91050 67.87778 67.84506
##       25       26       27       28       29       30       31       32
## 67.81234 67.77962 67.74690 67.71418 67.68146 67.64874 67.61602 67.58330
##       33       34       35       36       37       38       39       40
## 67.55058 67.51786 67.48514 67.45242 67.41970 67.38698 67.35426 67.32154
##       41       42       43       44       45       46       47       48
## 67.28882 67.25610 67.22338 67.19066 67.15794 67.12522 67.09250 67.05978
##       49       50       51       52       53       54       55       56
## 67.02706 66.99434 66.96162 66.92890 66.89618 66.86346 66.83074 66.79802
##       57       58       59       60       61       62       63       64
## 66.76530 66.73258 66.69987 66.66715 66.63443 66.60171 66.56899 66.53627
##       65       66       67       68       69       70       71       72
## 66.50355 66.47083 66.43811 66.40539 66.37267 66.33995 66.30723 66.27451
##       73       74       75       76       77       78       79       80
## 66.24179 66.20907 66.17635 66.14363 66.11091 66.07819 66.04547 66.01275
##       81       82       83       84       85       86       87       88
## 65.98003 65.94731 65.91459 65.88187 65.84915 65.81643 65.78371 65.75099
##       89       90       91       92       93       94       95       96
```

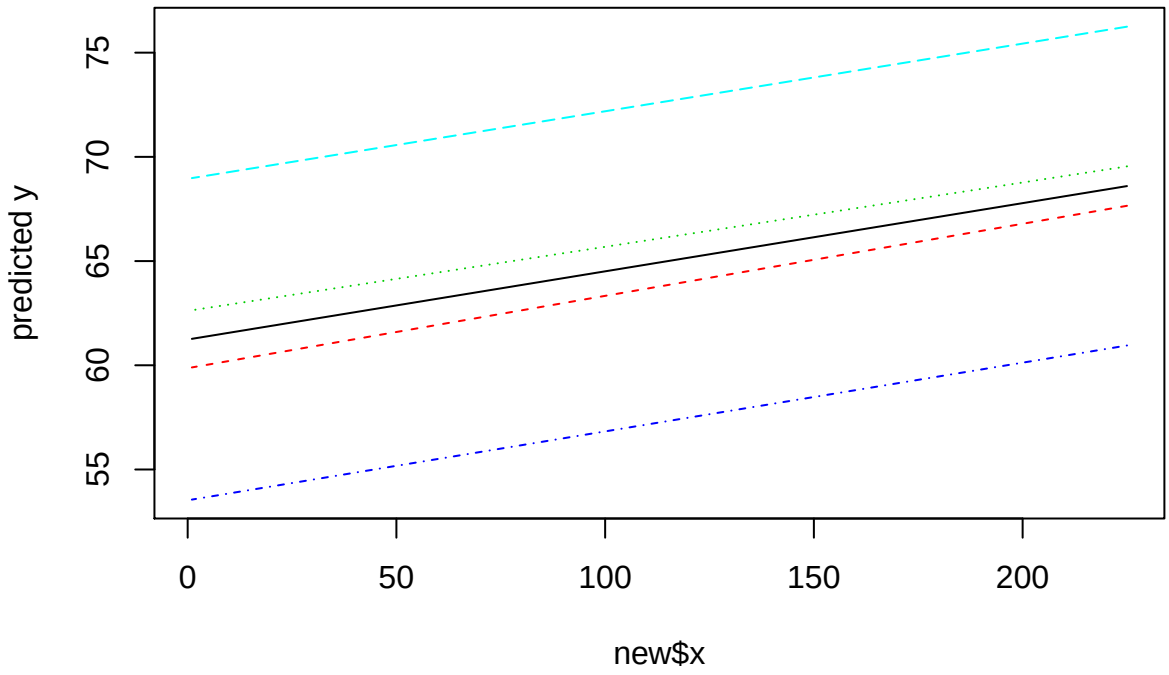
```
## 65.71827 65.68555 65.65283 65.62011 65.58739 65.55467 65.52195 65.48923
##       97       98       99      100      101      102      103      104
## 65.45651 65.42379 65.39107 65.35835 65.32563 65.29291 65.26019 65.22747
##      105      106      107      108      109      110      111      112
## 65.19475 65.16203 65.12931 65.09659 65.06387 65.03115 64.99843 64.96571
##      113      114      115      116      117      118      119      120
## 64.93299 64.90027 64.86755 64.83483 64.80211 64.76939 64.73667 64.70395
##      121      122      123      124      125      126      127      128
## 64.67123 64.63851 64.60579 64.57307 64.54035 64.50763 64.47491 64.44219
##      129      130      131      132      133      134      135      136
## 64.40947 64.37675 64.34403 64.31131 64.27859 64.24587 64.21315 64.18043
##      137      138      139      140      141      142      143      144
## 64.14771 64.11499 64.08227 64.04956 64.01684 63.98412 63.95140 63.91868
##      145      146      147      148      149      150      151      152
## 63.88596 63.85324 63.82052 63.78780 63.75508 63.72236 63.68964 63.65692
##      153      154      155      156      157      158      159      160
## 63.62420 63.59148 63.55876 63.52604 63.49332 63.46060 63.42788 63.39516
##      161      162      163      164      165      166      167      168
## 63.36244 63.32972 63.29700 63.26428 63.23156 63.19884 63.16612 63.13340
##      169      170      171      172      173      174      175      176
## 63.10068 63.06796 63.03524 63.00252 62.96980 62.93708 62.90436 62.87164
##      177      178      179      180      181      182      183      184
## 62.83892 62.80620 62.77348 62.74076 62.70804 62.67532 62.64260 62.60988
##      185      186      187      188      189      190      191      192
## 62.57716 62.54444 62.51172 62.47900 62.44628 62.41356 62.38084 62.34812
##      193      194      195      196      197      198      199      200
## 62.31540 62.28268 62.24996 62.21724 62.18452 62.15180 62.11908 62.08636
##      201      202      203      204      205      206      207      208
## 62.05364 62.02092 61.98820 61.95548 61.92276 61.89004 61.85732 61.82460
##      209      210      211      212      213      214      215      216
## 61.79188 61.75916 61.72644 61.69372 61.66100 61.62828 61.59556 61.56284
##      217      218      219      220      221      222      223      224
## 61.53012 61.49740 61.46468 61.43196 61.39924 61.36653 61.33381 61.30109
##      225
## 61.26837
```

From the generated data, the predicted mortality rate at lead concentration of 225 ppm is 61.27%.

f, g.

The plot is given below for the pediction with 95% confidence intervals and 95% prediction intervals.

```
pred.w.plim=predict(lm(y ~ x), new, interval = "prediction")
pred.w.clim=predict(lm(y ~ x), new, interval = "confidence")
#prediction plots
matplot(new$x, cbind(pred.w.clim, pred.w.plim[,-1]),  type = "l", ylab = "predicted y")
```



h.

The plot specifies the computation of confidence or prediction (tolerance) intervals at the specified level (95%). It is also referred to as narrow vs. wide intervals. This provides an idea about how good the prediction is.

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.