MATH 423 A3

Alyzeh Jiwani

07/12/2019

Question 1

The number of rings is indicative of the age of the abalone. The research group believes that there is a linear relationship between the height of the abaolones and their age. The linear model is:

Age = beta_0 + beta_1*Height + epsilon

where beta_0 is the intercept of the linear regression line, beta _1 is the coefficient of the slope of the regression line, and epsilon is iid Normal (mean = 0, variance = sigma^2) we can estimate the model as follows:

```
names(abalone)
```

```
## [1] "Height" "Rings"
```

```
summarise_all(abalone, mean)
```

```
##      Height    Rings
## 1 0.1395164 9.933684
```

```
summarise_all(abalone, sd)
```

```
##       Height    Rings
## 1 0.04182706 3.224169
```

```
range_height <- max(abalone$Height)-min(abalone$Height)
range_age <- max(abalone$Rings)-min(abalone$Rings)
Ranges <- c(range_height, range_age)
Ranges
```

```
## [1]  1.13 28.00
```

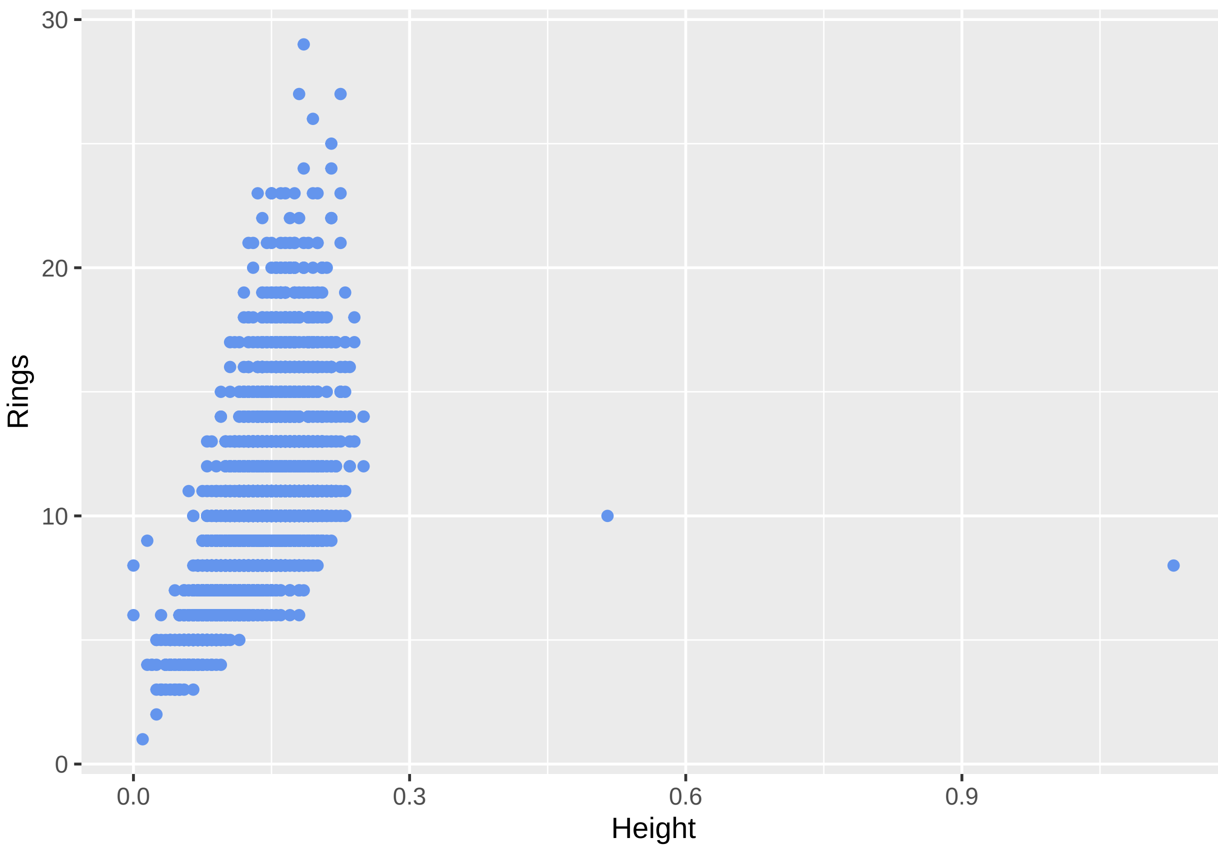
```
m<-lm(Rings~Height, data=abalone)

summary(m)
```

```
##
## Call:
## lm(formula = Rings ~ Height, data = abalone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.496  -1.657  -0.607   0.839  17.112
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9385     0.1443   27.30   <2e-16 ***
## Height       42.9714     0.9904   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.677 on 4175 degrees of freedom
## Multiple R-squared:  0.3108, Adjusted R-squared:  0.3106
## F-statistic:  1882 on 1 and 4175 DF,  p-value: < 2.2e-16
```

```
ggplot(abalone, aes(x=Height, y= Rings))+
  geom_point(col="cornflowerblue")
```



The plot seems to indicate that there is a correlation between Rings and Height As we can see, the estimate of the intercept of the regression line is

beta_hat_0 = 3.9385 The estimate of the slope coefficient is beta_hat_1 = 42.9714

The estimated regression model is Age_hat = 3.9385 + 42.9714*Height As the slope coefficient is greater than zero, this implies that a larger ehight is associauted with an older age. the value of the slope coefficient suggests that for a one unit increase in the height of an abalone, the age increases by 42.9714 We test the following hypotheses to see if the linear relationship is significant and if the value of the slope coefficient is positive. H_0 : beta_1 = 0, i.e. the height is not a predictor of the age of the abalone H_a : beta_1 > 0, i.e. a larger height is associated with an older age alpha = 0.05

Question 2

Given that Y_i = Beta_1 x_1,1 + Beta_2x_2,1 + epsilon_i The least square estimate of Beta_hat_1 and Beta_hat_2 from the multiple regression will be the same as the sample seperate regression on x_1 and x_2 y_ix_1,i + epsilon_i; i=1,2,. . . n thus: Beta_hat_1 = (x_1, x_1)ˆ-1 x,y where y' = (y_1, . . . , y_n) and x_1' = (x_1,1,. . . , x_1,n) we have y_i = Beta_2x_2,i + epsilon_i thus, the least squares est. of beta_hat_2 is beta_hat_2 = (y_2'x_2)'x_2y , x_2' = (x_2,1,. . . , x_2,n) Multiple regression model: y_i = beta_1x_1,i +beta_2x_2,i _ epsilon_i sum(x_1,ix_2,i)=0 1<=i<=n (y_1, y_2, y_3)ˆT = (x_1, x_2)(Beta_1, Beta_2)ˆT + (epsilon_1, epsilon_2, epsilon_3)ˆT y x*Beta + epsilon where: y' = (y_1, . . . , y_n) x = (x_1, x_2) Beta' = (beta_1, beta_2) Thus we have: beta_hat = (x'x)ˆ(-1)x'y = ((x_1', x_2)ˆT* (x_1, x_2)) * matrix((x_1', x_2', y, y, ncol = 2)) = matrix(x_1'x_1, x_2'x_1, x_1'x_2, x_2'x_2, ncol=2)ˆ(-1) * matrix((x_1', x_2', y, y, ncol = 2)) = matrix(x_1'x_1, 0, 0, x_2'x_2, ncol=2)* matrix((x_1', x_2', y, y, ncol = 2) = (beta_hat_1, beta_hat_2)ˆT = beta_hat = matrix(x_1'x_1, x_2'x_2, x_1'y, x_2'y, ncol=2) We can see that the values are the same as in the seperated model.

Question 3

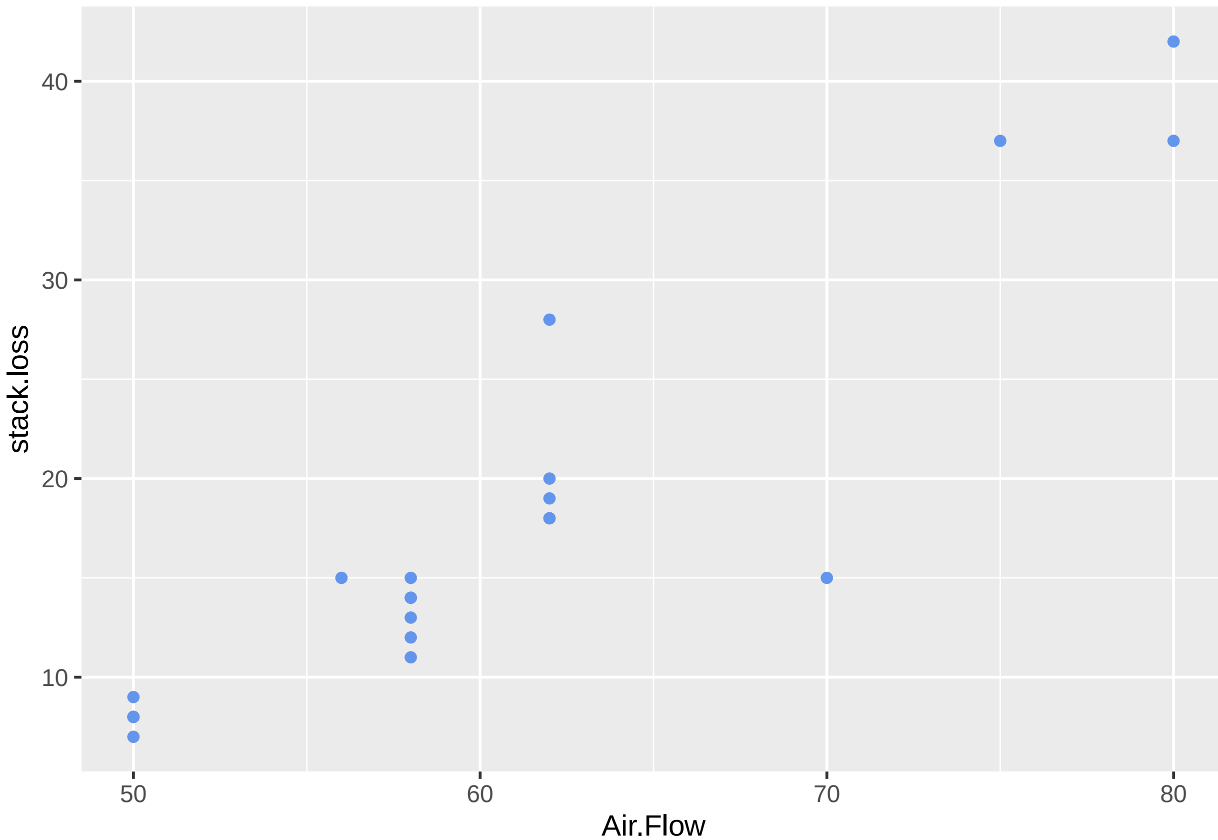
```
stackloss<-read.csv("stackloss.csv")
stackloss
```

```
##     X Air.Flow Water.Temp Acid.Conc. stack.loss
## 1   1       80         27         89         42
## 2   2       80         27         88         37
## 3   3       75         25         90         37
## 4   4       62         24         87         28
## 5   5       62         22         87         18
## 6   6       62         23         87         18
## 7   7       62         24         93         19
## 8   8       62         24         93         20
## 9   9       58         23         87         15
## 10 10       58         18         80         14
## 11 11       58         18         89         14
## 12 12       58         17         88         13
## 13 13       58         18         82         11
## 14 14       58         19         93         12
## 15 15       50         18         89          8
## 16 16       50         18         86          7
## 17 17       50         19         72          8
## 18 18       50         19         79          8
## 19 19       50         20         80          9
## 20 20       56         20         82         15
## 21 21       70         20         91         15
```

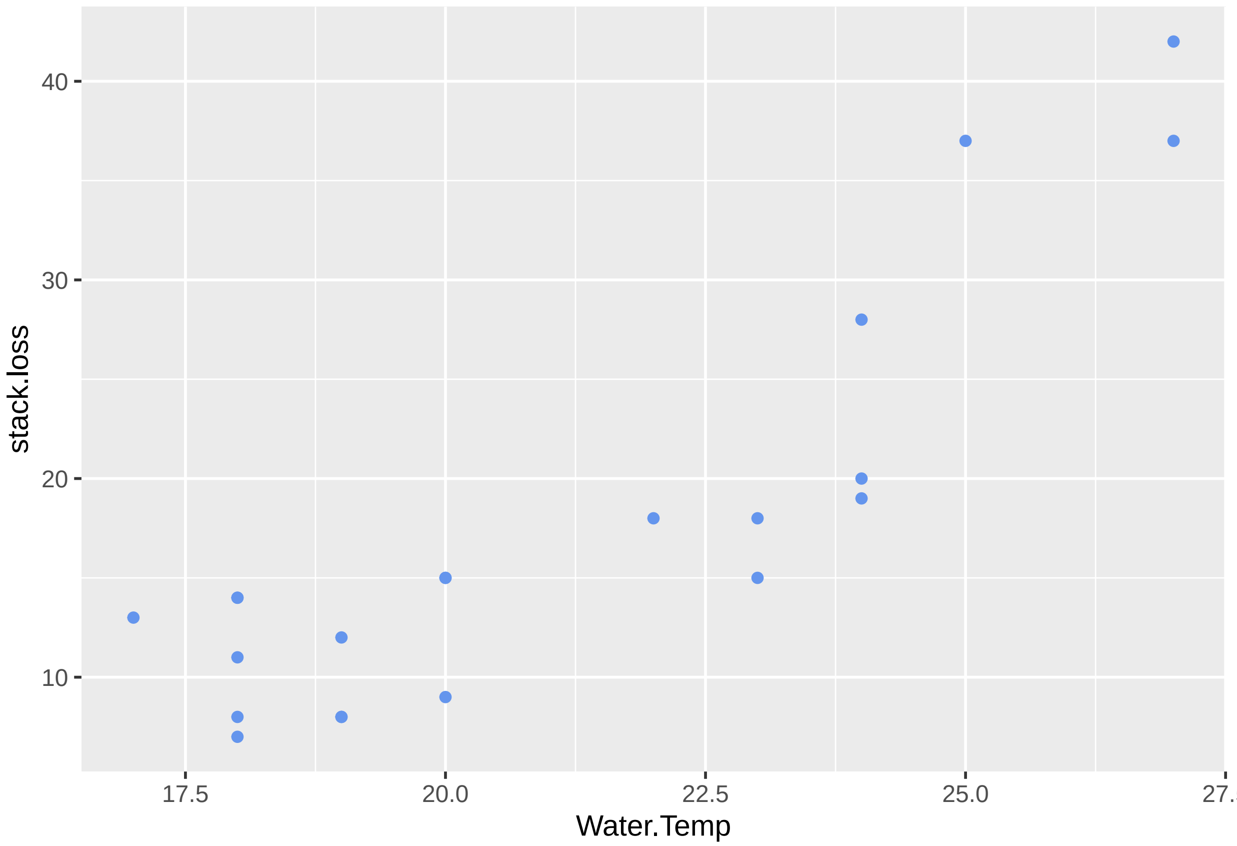
```
colnames(stackloss)
```

```
## [1] "X"          "Air.Flow"   "Water.Temp" "Acid.Conc." "stack.loss"
```

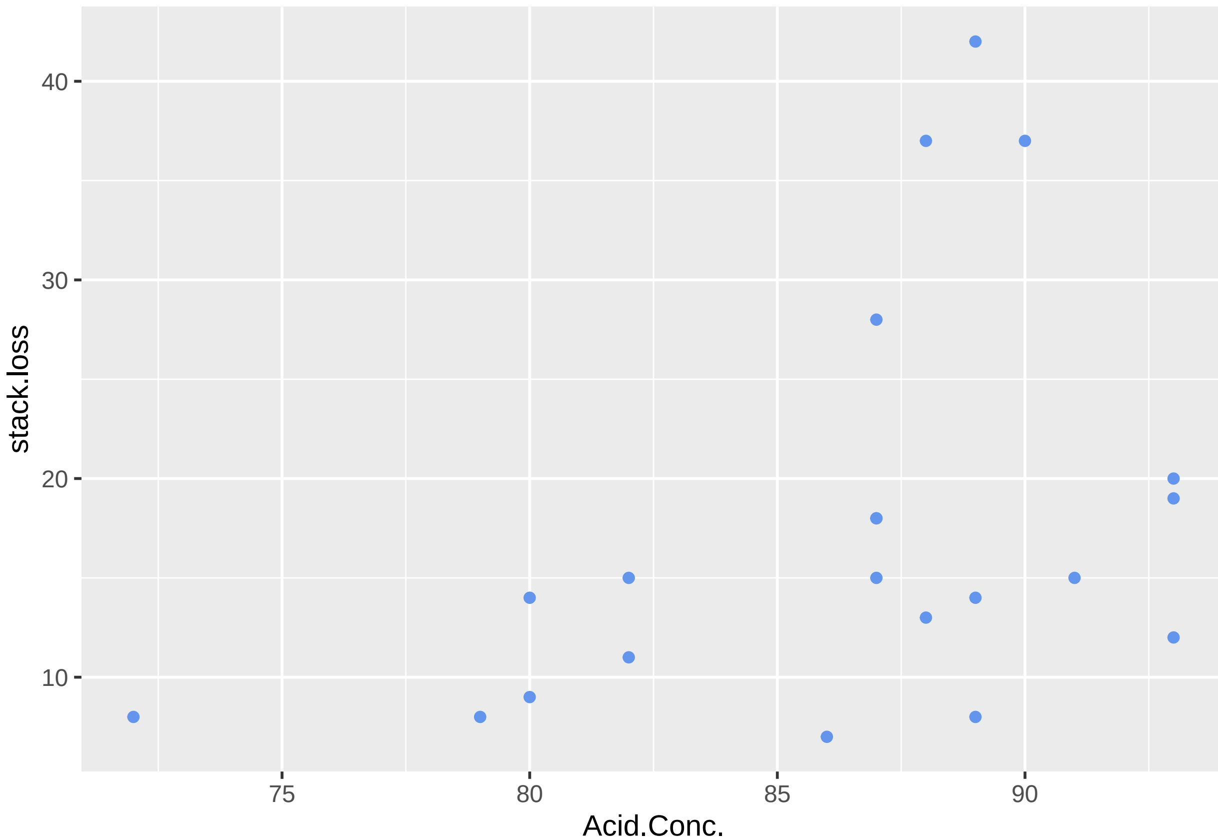
```
par(mfrow=c(2,2))
ggplot(stackloss, aes(x=Air.Flow, y= stack.loss))+
  geom_point(col="cornflowerblue")
```




```
ggplot(stackloss, aes(x=Water.Temp, y= stack.loss))+
  geom_point(col="cornflowerblue")
```




```r
ggplot(stackloss, aes(x=Acid.Conc., y= stack.loss))+
  geom_point(col="cornflowerblue")
```




```r
reg_model <- lm(stack.loss~., data=stackloss)
summary(reg_model)
```

```
##
## Call:
## lm(formula = stack.loss ~ ., data = stackloss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.974 -2.282  0.373  1.369  4.400
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.8549    14.0712  -1.411   0.1774
## X            -0.3779     0.1713  -2.206   0.0423 *
## Air.Flow      0.6656     0.1238   5.376 6.18e-05 ***
## Water.Temp    0.8694     0.3842   2.263   0.0379 *
## Acid.Conc.   -0.1973     0.1425  -1.384   0.1853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.927 on 16 degrees of freedom
## Multiple R-squared:  0.9337, Adjusted R-squared:  0.9172
## F-statistic: 56.36 on 4 and 16 DF,  p-value: 3.149e-09
```

this explains 91% of the variation in the data by the independent variables. Both Airflow and Water temp are significant in explaining the variation in the data. We can see taht acid conc. is not too significant as its p-value is greater than 0.05.

```r
confint(reg_model, level=0.9)
```

```
##                      5 %        95 %
## (Intercept) -44.4214632  4.71175789
## X            -0.6769469 -0.07885415
## Air.Flow      0.4494467  0.88178063
## Water.Temp    0.1986802  1.54016631
## Acid.Conc.   -0.4462103  0.05153996
```

```r
predict( reg_model, data = data.frame(Air.Flow = 58, Water.Temp = 20, Acid.Conc. = 86), interval = "pre
```

```
## Warning in predict.lm(reg_model, data = data.frame(Air.Flow = 58, Water.Temp = 20, : predictions on
```

```
##          fit        lwr      upr
## 1  38.927939 29.1708391 48.68504
## 2  38.747374 28.9292272 48.56552
## 3  32.907888 23.6207650 42.19501
## 4  23.599592 14.3555020 32.84368
## 5  21.482844 12.4040576 30.56163
## 6  21.974367 13.0069976 30.94174
## 7  21.281879 11.8395922 30.72417
## 8  20.903978 11.4409125 30.36704
## 9  18.178211  9.0480937 27.30833
## 10 14.834540  5.0633986 24.60568
## 11 12.680623  3.3274310 22.03382
## 12 11.630634  2.0144664 21.24680
## 13 13.306168  4.0338887 22.57845
## 14 11.627004  2.2211471 21.03286
## 15  5.844111 -3.4890042 15.17723
## 16  6.058216 -3.0353738 15.15181
## 17  9.312432 -0.8518983 19.47676
## 18  7.553185 -1.7352360 16.84161
## 19  7.847372 -1.7046626 17.39941
## 20 11.068484  1.5766449 20.56032
## 21 18.233158  7.1870655 29.27925
```