CSC 423 - Chapter 3 Homework

Jasmine Dumas

September 29, 2015

page 103 #3.16

Dataset: LIQUIDSPILL

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/LIQUIDSPILL.Rdata")
head(LIQUIDSPILL, n = 12) # recreate table from the textbook
```

```
##    TIME MASS
## 1     0 6.64
## 2     1 6.34
## 3     2 6.04
## 4     4 5.47
## 5     6 4.94
## 6     8 4.44
## 7    10 3.98
## 8    12 3.55
## 9    14 3.15
## 10   16 2.79
## 11   18 2.45
## 12   20 2.14
```

```
tail(LIQUIDSPILL, n = 11)
```

```
##    TIME MASS
## 13   22 1.86
## 14   24 1.60
## 15   26 1.37
## 16   28 1.17
## 17   30 0.98
## 18   35 0.60
## 19   40 0.34
## 20   45 0.17
## 21   50 0.06
## 22   55 0.02
## 23   60 0.00
```

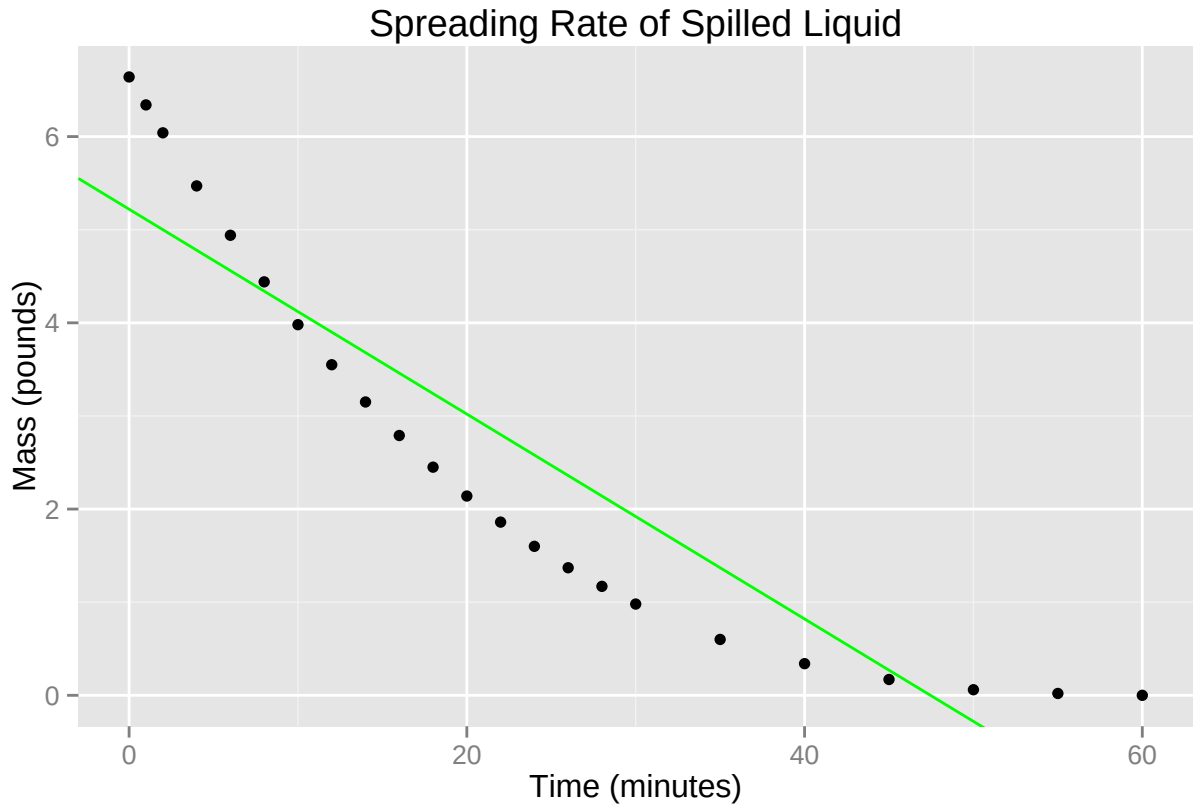
Yes, the data does suggest that mass of the spill tends to diminish as time increases.

The dependent variable is MASS (`LIQUIDSPILL$MASS`), and the independent variable is TIME (`LIQUIDSPILL$TIME`). If I want to determine how much mass deminishes each minute I will fit a `lm()` to the data. y = 5.22 - 0.11x is the equation which equates to a **reduction** in MASS of approximately 5.11 pounds for every minute during a spill.

```
x <- LIQUIDSPILL$TIME
y <- LIQUIDSPILL$MASS
reg <- lm(y~x)
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8861 -0.7593 -0.3024  0.6229  1.6207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.22070    0.29598   17.64 4.55e-14 ***
## x           -0.11402    0.01032  -11.05 3.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8573 on 21 degrees of freedom
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8464
## F-statistic: 122.2 on 1 and 21 DF,  p-value: 3.26e-10
```

```
library(ggplot2)
df = data.frame(x, y)
liquid.plot <- ggplot(df, aes(x, y)) +
               geom_point() + # scatterplot
               geom_abline(intercept =5.22, slope = -0.11, colour = "green") +
               labs(title = "Spreading Rate of Spilled Liquid",
                    x = "Time (minutes)", y = "Mass (pounds)")

print(liquid.plot)
```



page 109 #3.22

Dataset: HEAT

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/HEAT.Rdata")
head(HEAT) # view table
```

```
##   RATIO HEAT
## 1  1.93  4.4
## 2  2.00  5.2
## 3  1.95  5.3
## 4  1.77  4.7
## 5  1.78  4.5
## 6  1.62  4.2
```

*Given information: The dependent variable is the heat transfer enhancement ratio (**HEAT$HEAT**), y and the independent variable is the unflooded area ration (**HEAT$RATIO**), x.*

a) Fit a least squares line to the data. The equation is y = 0.21 + 2.43x

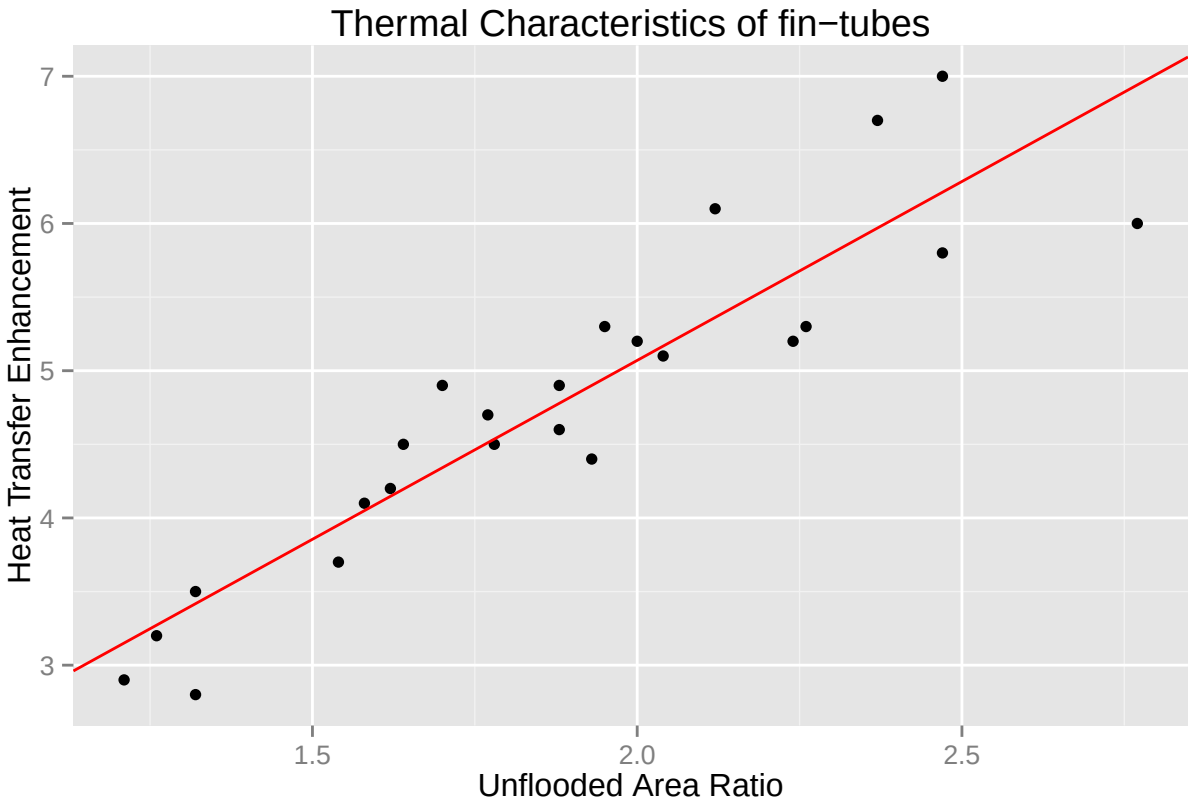
```
heat.transfer.y <- HEAT$HEAT
unflooded.x <- HEAT$RATIO

heated <- lm(heat.transfer.y ~unflooded.x)
summary(heated)
```

```
##
## Call:
## lm(formula = heat.transfer.y ~ unflooded.x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93449 -0.28678  0.01028  0.22076  0.79343
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2134     0.4390   0.486    0.632
## unflooded.x   2.4264     0.2283  10.630 3.92e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4538 on 22 degrees of freedom
## Multiple R-squared:  0.837,  Adjusted R-squared:  0.8296
## F-statistic:   113 on 1 and 22 DF,  p-value: 3.925e-10
```

b) plot of the data and the regression line on the plot

```
df.heat = data.frame(unflooded.x, heat.transfer.y)
heated.plot <- ggplot(df.heat, aes(unflooded.x, heat.transfer.y)) +
               geom_point() + # scatterplot
               geom_abline(intercept = 0.21, slope = 2.43, colour = "red") +
               labs(title = "Thermal Characteristics of fin-tubes",
```

```
                    x = "Unflooded Area Ratio", y = "Heat Transfer Enhancement")

print(heated.plot)
```



c) Calculate SSE (Sum of Squared Errors) and s^2 (MSE/Mean of Squared Error)

```
SSE = sum(resid(heated)^2)
print(SSE)
```

```
## [1] 4.531079
```

```
s.squared = SSE / nrow(HEAT) - 2
print(s.squared)
```

```
## [1] -1.811205
```

d) Calculate s and interpret its value. **The standard deviation for this data set is small, considering the x-axis and y-axis range all below 8, and the small sample size.**

```
RMSE = sqrt(deviance(heated)/df.residual(heated)) # standard deviation
print(RMSE)
```

```
## [1] 0.4538261
```

page 114 #3.28

Dataset: BOXING2

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/BOXING2.Rdata")
print(BOXING2)
```

```
##    LACTATE RECOVERY
## 1      3.8        7
## 2      4.2        7
## 3      4.8       11
## 4      4.1       12
## 5      5.0       12
## 6      5.3       12
## 7      4.2       13
## 8      2.4       17
## 9      3.7       17
## 10     5.3       17
## 11     5.8       18
## 12     6.0       18
## 13     5.9       21
## 14     6.3       21
## 15     5.5       20
## 16     6.5       24
```

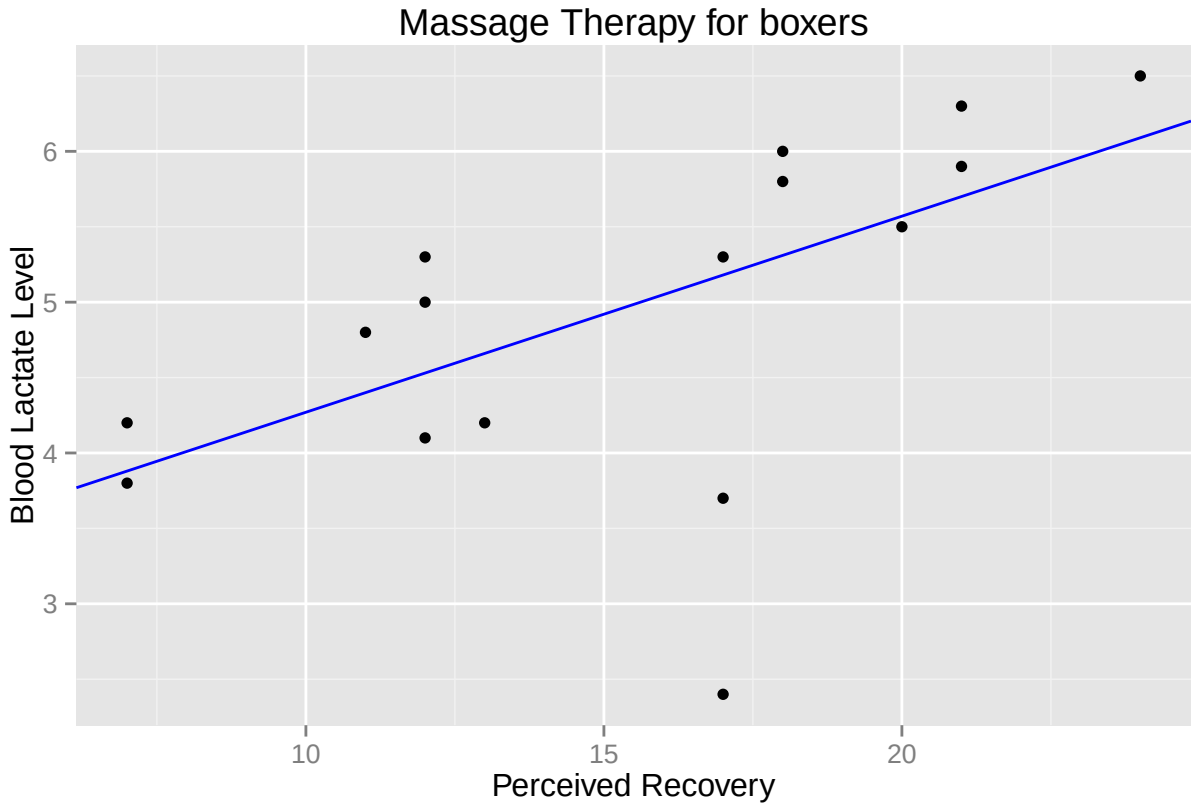
Conduct a test to determine whether blood lactate level (y) is lineraly related to perceived recovery (x). Use $\alpha = 0.10$

```
blood.y <- BOXING2$LACTATE
recovery.x <- BOXING2$RECOVERY

h.swank <- lm( blood.y ~ recovery.x)

summary(h.swank)
```

```
##
## Call:
## lm(formula = blood.y ~ recovery.x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7229 -0.1396  0.3071  0.5204  0.8104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.96960    0.78959   3.761  0.00211 **
## recovery.x   0.12667    0.04878   2.597  0.02110 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9507 on 14 degrees of freedom
## Multiple R-squared:  0.3251, Adjusted R-squared:  0.2769
## F-statistic: 6.744 on 1 and 14 DF,  p-value: 0.0211
```

With a p-value of 0.0211 which is smaller than the α of 0.10 you could reject the H_0 which means percieved recovery (x) possibly contributes information to blood lactate level (y).

```
df.boxing = data.frame(recovery.x, blood.y)
boxing.plot <- ggplot(df.boxing, aes(recovery.x, blood.y)) +
               geom_point() + # scatterplot
               geom_abline(intercept = 2.97, slope = 0.13, colour = "blue") +
               labs(title = "Massage Therapy for boxers",
                    y = "Blood Lactate Level", x = "Perceived Recovery")

print(boxing.plot)
```



page 127 #3.46

Dataset: SNOWGEESE

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/SNOWGEESE.Rdata")
head(SNOWGEESE, n = 5)
```

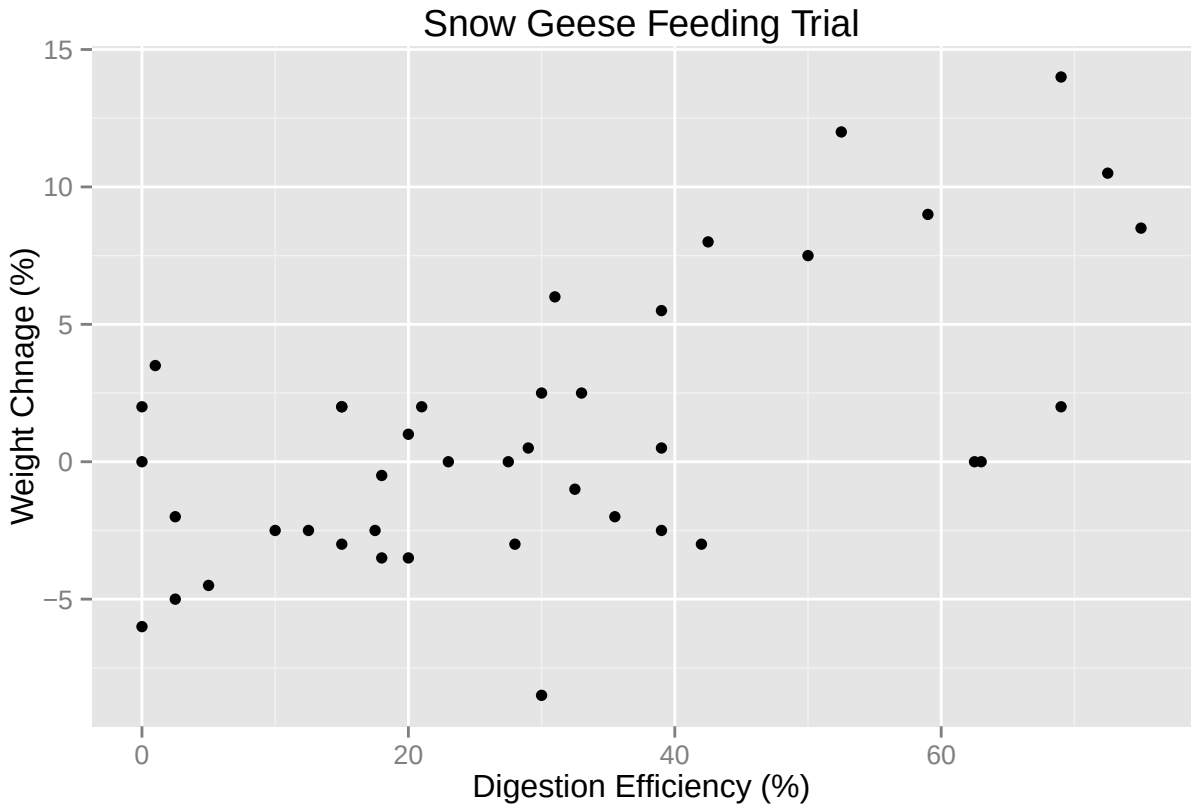
```
##   TRIAL             DIET WTCHNG DIGEFF ADFIBER
## 1     1 Plants         -6.0    0.0    28.5
## 2     2 Plants         -5.0    2.5    27.5
## 3     3 Plants         -4.5    5.0    27.5
## 4     4 Plants          0.0    0.0    32.5
## 5     5 Plants          2.0    0.0    32.0
```

```
tail(SNOWGEESE, n = 5)
```

```
##    TRIAL             DIET WTCHNG DIGEFF ADFIBER
## 38    38 Chow          9.0   59.0     8.5
## 39    39 Chow         12.0   52.5     8.0
## 40    40 Chow          8.5   75.0     6.0
## 41    41 Chow         10.5   72.5     6.5
## 42    42 Chow         14.0   69.0     7.0
```

a) Plot data. weight change (y) & digestion efficiency (x). **There is a moderate positive trend between the independent and dependent variables.**

```
digest.x = SNOWGEESE$DIGEFF
weight.y = SNOWGEESE$WTCHNG
down.coat <- data.frame(digest.x, weight.y)
sg <- ggplot(down.coat, aes(digest.x, weight.y)) +
            geom_point() +
            labs(title = "Snow Geese Feeding Trial",
                y = "Weight Chnage (%)", x = "Digestion Efficiency (%)")
print(sg)
```



b) Find the coefficient of correlation, r. **r = 0.6122317 which suggests a moderate positive linear relationship between weight change and digestion efficiency.**

```
cor(digest.x, weight.y)
```

```
## [1] 0.6122317
```

c) conduct a test with $\alpha = 0.01$. **With a p-value of 1.642e-05 which is smaller than the α of 0.01 means that you could reject the H_0 and accept the alternative.**

```
snow.model <- lm(weight.y ~ digest.x)
summary(snow.model)
```

```
##
## Call:
## lm(formula = weight.y ~ digest.x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5733 -2.7288 -0.2575  3.1273  7.7436
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.17067    1.06784  -2.969  0.00503 **
## digest.x     0.14147    0.02889   4.897 1.64e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.003 on 40 degrees of freedom
## Multiple R-squared:  0.3748, Adjusted R-squared:  0.3592
## F-statistic: 23.98 on 1 and 40 DF,  p-value: 1.642e-05
```

d) repeat of parts b and c but only data from plants and not duck chow. **The coefficient of correlation is low, but still a positive correlation. The p-value is greater than the α or 0.01 which means you can fail to reject the H_0 which means that could be an effect of x on y but more data is needed.**

```
plant.food <- subset(SNOWGEESE, DIET == "Plants          ", # white space in df included
                     select = c(WTCHNG, DIGEFF))

cor(plant.food$DIGEFF, plant.food$WTCHNG) # coefficient of correlation
```

```
## [1] 0.3094942
```

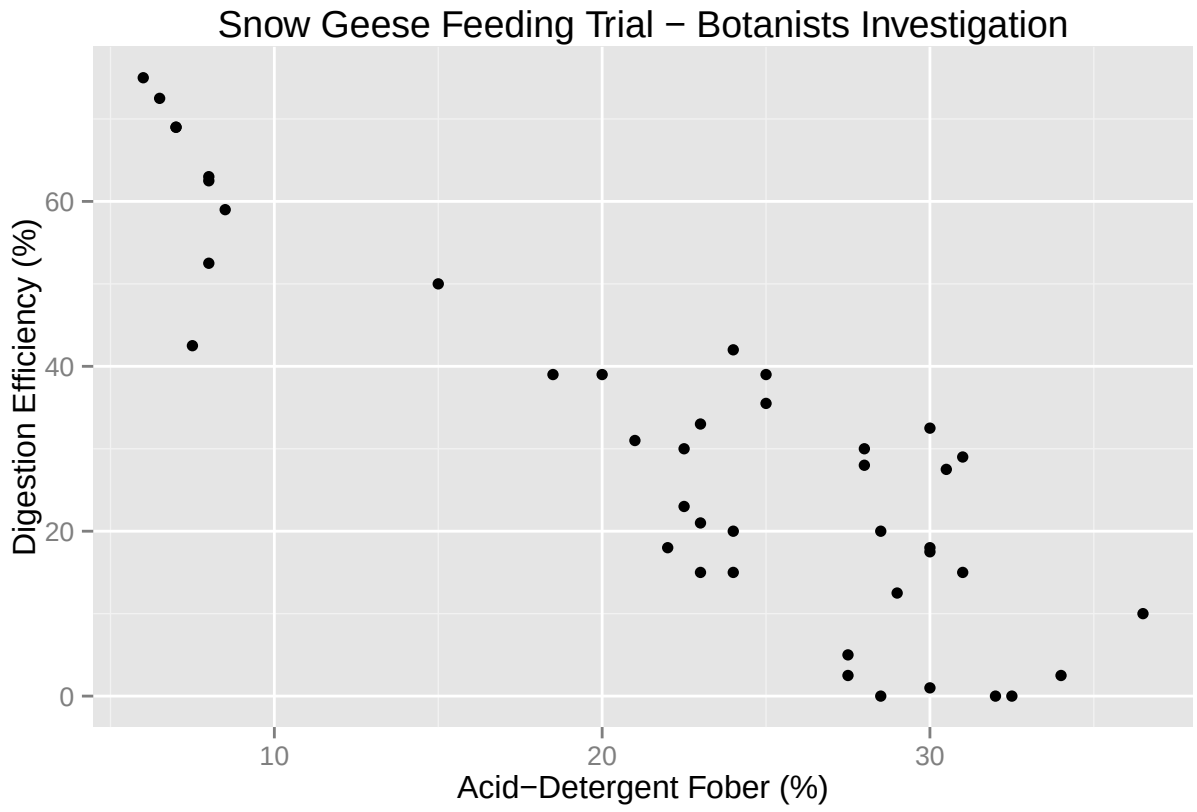
```
plant.model <- lm(plant.food$WTCHNG ~ plant.food$DIGEFF)
summary(plant.model) # linear model
```

```
##
## Call:
## lm(formula = plant.food$WTCHNG ~ plant.food$DIGEFF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6392 -2.6814  0.0144  2.3608  5.7946
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.21019    1.09193  -2.024   0.0516 .
## plant.food$DIGEFF  0.07831    0.04321   1.812   0.0797 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.391 on 31 degrees of freedom
## Multiple R-squared:  0.09579,    Adjusted R-squared:  0.06662
## F-statistic: 3.284 on 1 and 31 DF,  p-value: 0.07966
```

e) repeat parts a-d between digestion efficiency (y) and acid-detergent fiber (x). **There seems to be a negative trend between the independent and dependent variables. The coefficient of correlation is strongly negative which matches the plot. The p-value 1.636e-14 which is smaller than the α of 0.01 means that you could reject the H_0 and accept the alternative. With the exclusion of the Chow food, the p-value is 4.913e-05 which is still smaller than α which means you can similarily reject the H_0**

```
# part a
adfib.x = SNOWGEESE$ADFIBER
digest.y = SNOWGEESE$DIGEFF
botanists <- data.frame(adfib.x, digest.y)
sg <- ggplot(botanists, aes(adfib.x, digest.y)) +
```

```
            geom_point() +
            labs(title = "Snow Geese Feeding Trial - Botanists Investigation",
                x = "Acid-Detergent Fober (%)", y = "Digestion Efficiency (%)")
print(sg)
```




```
# part b
cor(adfib.x, digest.y)
```

```
## [1] -0.8800539
```

```
# part c
botanists.model <- lm(digest.y ~ adfib.x)
summary(botanists.model)
```

```
##
## Call:
## lm(formula = digest.y ~ adfib.x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.238  -8.143   2.067   8.040  18.250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  77.5673     4.3520   17.82  < 2e-16 ***
## adfib.x      -2.1106     0.1801  -11.72 1.64e-14 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.4 on 40 degrees of freedom
## Multiple R-squared:  0.7745, Adjusted R-squared:  0.7689
## F-statistic: 137.4 on 1 and 40 DF,  p-value: 1.636e-14
```

```
# part d
botanists.food <- subset(SNOWGEESE, DIET == "Plants          ", # white space in df included
                     select = c(DIGEFF, ADFIBER))

cor(botanists.food$ADFIBER, botanists.food$DIGEFF) # coefficient of correlation
```

```
## [1] -0.6459071
```

```
botanists.model2 <- lm(botanists.food$DIGEFF ~ botanists.food$ADFIBER)
summary(botanists.model2) # linear model
```

```
##
## Call:
## lm(formula = botanists.food$DIGEFF ~ botanists.food$ADFIBER)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.647 -10.142   2.044   7.364  17.667
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              71.1224    10.7497   6.616 2.15e-07 ***
## botanists.food$ADFIBER  -1.8763     0.3983  -4.711 4.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.76 on 31 degrees of freedom
## Multiple R-squared:  0.4172, Adjusted R-squared:  0.3984
## F-statistic: 22.19 on 1 and 31 DF,  p-value: 4.913e-05
```

page 134 #3.56

Dataset: WHITESPRUCE

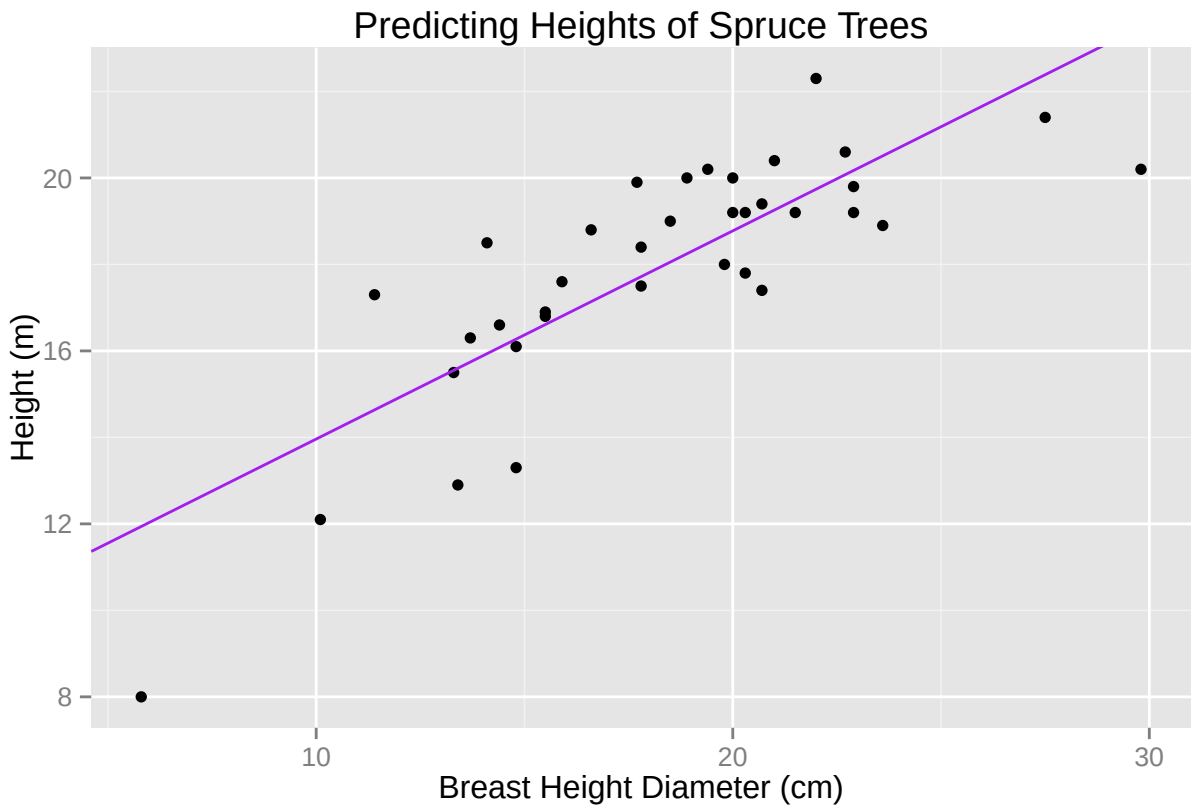
```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/WHITESPRUCE.Rdata")
WHITESPRUCE
```

```
##    DIAMETER HEIGHT
## 1      18.9   20.0
## 2      15.5   16.8
## 3      19.4   20.2
## 4      20.0   20.0
## 5      29.8   20.2
## 6      19.8   18.0
## 7      20.3   17.8
## 8      20.0   19.2
## 9      22.0   22.3
## 10     23.6   18.9
## 11     14.8   13.3
## 12     22.7   20.6
## 13     18.5   19.0
## 14     21.5   19.2
## 15     14.8   16.1
## 16     17.7   19.9
## 17     21.0   20.4
## 18     15.9   17.6
## 19     16.6   18.8
## 20     15.5   16.9
## 21     13.7   16.3
## 22     27.5   21.4
## 23     20.3   19.2
## 24     22.9   19.8
## 25     14.1   18.5
## 26     10.1   12.1
## 27      5.8    8.0
## 28     20.7   17.4
## 29     17.8   18.4
## 30     11.4   17.3
## 31     14.4   16.6
## 32     13.4   12.9
## 33     17.8   17.5
## 34     20.7   19.4
## 35     13.3   15.5
## 36     22.9   19.2
```

a) Scatterplot c) the least squares line on the scatter plot

```
big.trees = data.frame(WHITESPRUCE$DIAMETER, WHITESPRUCE$HEIGHT)

bt <- ggplot(big.trees, aes(WHITESPRUCE$DIAMETER, WHITESPRUCE$HEIGHT)) +
            geom_point() +
            geom_abline(intercept = 9.14684, slope = 0.48147, colour = "purple") +
            labs(title = "Predicting Heights of Spruce Trees",
```

```
                x = "Breast Height Diameter (cm)", y = "Height (m)")
print(bt)
```



b) least squares method to scatterplot. Fit a linear model to the data. **The y-intercept is 9.14684 and the slope is 0.48147**

```
HEIGHT <- WHITESPRUCE$HEIGHT
DIAMETER <- WHITESPRUCE$DIAMETER
spruce <- lm(HEIGHT ~ DIAMETER)
summary(spruce)
```

```
##
## Call:
## lm(formula = HEIGHT ~ DIAMETER)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9394 -0.9763  0.2829  0.9950  2.6644
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.14684    1.12131   8.157 1.63e-09 ***
## DIAMETER     0.48147    0.05967   8.069 2.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.678 on 34 degrees of freedom
## Multiple R-squared:  0.6569, Adjusted R-squared:  0.6468
## F-statistic:  65.1 on 1 and 34 DF,  p-value: 2.089e-09
```

d) The p-value is 2.089e-09 and $\alpha = 0.05$. Since the p-value is less than the α you can reject the H_0 which implies no effect of x on y and accept the alternative. The data provides sufficient evidence that the breast height diameter does contribute information about the prediction of the tree height.

e) find a confidence interval for average height of white spruce trees with a breast height diameter of 20 cm. **The 90 % confidence interval for a mean breast height of 20 cm is 18.26972 19.28293.**

```r
# DIAMETER has to be the same variable name as the one used in the model
predict(spruce, data.frame(DIAMETER = 20), interval="confidence", level=.90)
```

```
##        fit      lwr      upr
## 1 18.77632 18.26972 19.28293
```