INF 511 Assignment 2

Jen Diehl (6179236), Sam Watson (6174574), Natasha Wesely (6180693)

2022-09-16

$\frac{S1}{S2}$

Question 1

(a)

```r
sapflow.df<-  readRDS(file="sapflow.rds")
str(sapflow.df)
```

summary?

-1

```
## 'data.frame':    16 obs. of  5 variables:
##  $ V1: num  1 2 1 2 1 2 1 2 1 2 ...
##  $ V2: Factor w/ 2 levels "lo","hi": 2 2 1 1 2 2 1 1 2 2 ...
##  $ V3: Factor w/ 2 levels "lo","hi": 1 1 1 1 2 2 2 2 1 1 ...
##  $ V4: Factor w/ 2 levels "lo","hi": 2 2 2 2 2 2 2 2 1 1 ...
##  $ V5: num  192 248 168 204 303 ...
```

V2 seems to be a binary variable that would most likely take on a binomial distribution because there are only two options.

(b)

```r
names(sapflow.df) = c("light","fertilizer","temperature","moisture","sapflow")
head(sapflow.df)
```

```
##   light fertilizer temperature moisture sapflow
## 1     1         hi          lo       hi   192.5
## 2     2         hi          lo       hi   248.3
## 3     1         lo          lo       hi   168.4
## 4     2         lo          lo       hi   204.2
## 5     1         hi          hi       hi   302.8
## 6     2         hi          hi       hi   341.2
```

(c)

```r
sapflow.df$light = factor(sapflow.df$light, labels = c("lo","hi"))
head(sapflow.df)
```

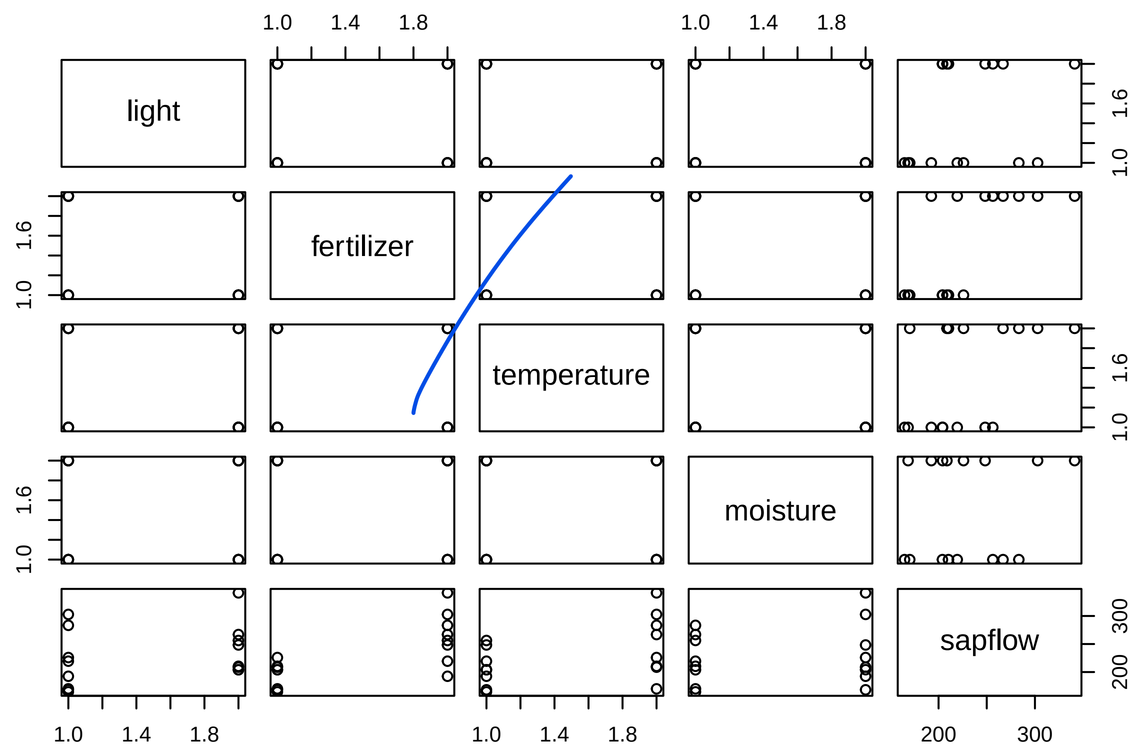
```
##   light fertilizer temperature moisture sapflow
## 1    lo         hi          lo       hi   192.5
## 2    hi         hi          lo       hi   248.3
## 3    lo         lo          lo       hi   168.4
## 4    hi         lo          lo       hi   204.2
## 5    lo         hi          hi       hi   302.8
## 6    hi         hi          hi       hi   341.2
```

```
str(sapflow.df)
```

```
## 'data.frame':    16 obs. of  5 variables:
##  $ light      : Factor w/ 2 levels "lo","hi": 1 2 1 2 1 2 1 2 1 2 ...
##  $ fertilizer : Factor w/ 2 levels "lo","hi": 2 2 1 1 2 2 1 1 2 2 ...
##  $ temperature: Factor w/ 2 levels "lo","hi": 1 1 1 1 2 2 2 2 1 1 ...
##  $ moisture   : Factor w/ 2 levels "lo","hi": 2 2 2 2 2 2 2 2 1 1 ...
##  $ sapflow    : num  192 248 168 204 303 ...
```

(d)

```
pairs(sapflow.df)
```



The pairs function reports the factors as their numeric value of 1 or 2.

(e)

A single histogram may not be enough to assess if sapflow is normally distributed because it is just a visual assessment of normality which is inherently subjective. Creating a histogram is not formally or statistically assessing normality.

Question 2

(a)

```
X<- model.matrix(sapflow ~ light + fertilizer + temperature + moisture, data=sapflow.df)
y<- sapflow.df$sapflow
```

$E(Y_1 \mid x_1) = \beta_0 * X_{1,0} + \beta_1 * X_{1,1} + \beta_2 * X_{1,2} + \beta_3 * X_{1,3} + \beta_4 * X_{1,4}$

$E(Y_1 \mid x_1) = \beta_0 * 1 + \beta_1 * 0 + \beta_2 * 1 + \beta_3 * 0 + \beta_4 * 1$

$E(Y_1 \mid x_1) = \beta_0 + \beta_2 + \beta_4$

(b)

$E(Y_5 \mid x_5) = \beta_0 * X_{5,0} + \beta_1 * X_{5,1} + \beta_2 * X_{5,2} + \beta_3 * X_{5,3} + \beta_4 * X_{5,4}$

$E(Y_5 \mid x_5) = \beta_0 * 1 + \beta_1 * 0 + \beta_2 * 1 + \beta_3 * 1 + \beta_4 * 1$

$E(Y_5 \mid x_5) = \beta_0 + \beta_2 + \beta_3 + \beta_4$

$E\ (Y_5 \mid x_5) - E(Y_1 \mid x_1) = \beta_0 + \beta_2 + \beta_3 + \beta_4 - \beta_0 - \beta_2 - \beta_4$

$E\ (Y_5 \mid x_5) - E(Y_1 \mid x_1) = \beta_3$

The difference mean sapflow between high and low temperature and other inputs is equal to β_3.

(c)

```
model = lm(sapflow ~ light + fertilizer + temperature + moisture, data = sapflow.df)
summary(model)
```

```
##
## Call:
## lm(formula = sapflow ~ light + fertilizer + temperature + moisture,
##     data = sapflow.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.387 -16.863   5.125  16.069  34.787
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    151.96      14.73  10.318  5.4e-07 ***
## lighthi         26.60      13.17   2.019 0.068511 .
## fertilizerhi    69.30      13.17   5.261 0.000268 ***
```

```
## temperaturehi    43.92      13.17   3.334 0.006660 **
## moisturehi       14.63      13.17   1.110 0.290602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.35 on 11 degrees of freedom
## Multiple R-squared:  0.8004, Adjusted R-squared:  0.7278
## F-statistic: 11.03 on 4 and 11 DF,  p-value: 0.0007654
```

The model tells us that when moving from low to high temperature there is a 43.92 mL/h increase in sapflow.

(d)

i. $\hat{\beta}_3 = 43.92$

ii. SE $\hat{\beta}_3 = 13.17$

iii. t-value $\hat{\beta}_3 = 3.334$

iv. 11 degrees of freedom

v. p-value $= 0.006660$

vi. Yes, the effect of temperature is statistically significant because the t-value is far into the tails of the t distribution, therefore the p-value is small and less than alpha.

(e)

```
xtxi = solve(t(X) %*% X)
betahat = xtxi %*% t(X) %*% y
betahat
```

```
##                     [,1]
## (Intercept)   151.9625
## lighthi        26.6000
## fertilizerhi   69.3000
## temperaturehi  43.9250
## moisturehi     14.6250
```

```
coef(model)
```

```
##   (Intercept)       lighthi  fertilizerhi temperaturehi    moisturehi
##      151.9625       26.6000       69.3000       43.9250       14.6250
```

The beta hats are very similar from the manual calculations and the model created with lm.

(f)

```
S = t(X)%*%X
Xinv = solve(S)
H = X %*% Xinv %*% t(X)
H
```

```
##          1       2       3       4       5       6       7       8       9
## 1   0.3125  0.1875  0.1875  0.0625  0.1875  0.0625  0.0625 -0.0625  0.1875
## 2   0.1875  0.3125  0.0625  0.1875  0.0625  0.1875 -0.0625  0.0625  0.0625
## 3   0.1875  0.0625  0.3125  0.1875  0.0625 -0.0625  0.1875  0.0625  0.0625
## 4   0.0625  0.1875  0.1875  0.3125 -0.0625  0.0625  0.0625  0.1875 -0.0625
## 5   0.1875  0.0625  0.0625 -0.0625  0.3125  0.1875  0.1875  0.0625  0.0625
## 6   0.0625  0.1875 -0.0625  0.0625  0.1875  0.3125  0.0625  0.1875 -0.0625
## 7   0.0625 -0.0625  0.1875  0.0625  0.1875  0.0625  0.3125  0.1875 -0.0625
## 8  -0.0625  0.0625  0.0625  0.1875  0.0625  0.1875  0.1875  0.3125 -0.1875
## 9   0.1875  0.0625  0.0625 -0.0625  0.0625 -0.0625 -0.0625 -0.1875  0.3125
## 10  0.0625  0.1875 -0.0625  0.0625 -0.0625  0.0625 -0.1875 -0.0625  0.1875
## 11  0.0625 -0.0625  0.1875  0.0625 -0.0625 -0.1875  0.0625 -0.0625  0.1875
## 12 -0.0625  0.0625  0.0625  0.1875 -0.1875 -0.0625 -0.0625  0.0625  0.0625
## 13  0.0625 -0.0625 -0.0625 -0.1875  0.1875  0.0625  0.0625 -0.0625  0.1875
## 14 -0.0625  0.0625 -0.1875 -0.0625  0.0625  0.1875 -0.0625  0.0625  0.0625
## 15 -0.0625 -0.1875  0.0625 -0.0625  0.0625 -0.0625  0.1875  0.0625  0.0625
## 16 -0.1875 -0.0625 -0.0625  0.0625 -0.0625  0.0625  0.0625  0.1875 -0.0625
##         10      11      12      13      14      15      16
## 1   0.0625  0.0625 -0.0625  0.0625 -0.0625 -0.0625 -0.1875
## 2   0.1875 -0.0625  0.0625 -0.0625  0.0625 -0.1875 -0.0625
## 3  -0.0625  0.1875  0.0625 -0.0625 -0.1875  0.0625 -0.0625
## 4   0.0625  0.0625  0.1875 -0.1875 -0.0625 -0.0625  0.0625
## 5  -0.0625 -0.0625 -0.1875  0.1875  0.0625  0.0625 -0.0625
## 6   0.0625 -0.1875 -0.0625  0.0625  0.1875 -0.0625  0.0625
## 7  -0.1875  0.0625 -0.0625  0.0625 -0.0625  0.1875  0.0625
## 8  -0.0625 -0.0625  0.0625 -0.0625  0.0625  0.0625  0.1875
## 9   0.1875  0.1875  0.0625  0.1875  0.0625  0.0625 -0.0625
## 10  0.3125  0.0625  0.1875  0.0625  0.1875 -0.0625  0.0625
## 11  0.0625  0.3125  0.1875  0.0625 -0.0625  0.1875  0.0625
## 12  0.1875  0.1875  0.3125 -0.0625  0.0625  0.0625  0.1875
## 13  0.0625  0.0625 -0.0625  0.3125  0.1875  0.1875  0.0625
## 14  0.1875 -0.0625  0.0625  0.1875  0.3125  0.0625  0.1875
## 15 -0.0625  0.1875  0.0625  0.1875  0.0625  0.3125  0.1875
## 16  0.0625  0.0625  0.1875  0.0625  0.1875  0.1875  0.3125
```

(g)

```
yhat = H%*%y
yhat
```

```
##        [,1]
## 1  235.8875
## 2  262.4875
```

```
## 3  166.5875
## 4  193.1875
## 5  279.8125
## 6  306.4125
## 7  210.5125
## 8  237.1125
## 9  221.2625
## 10 247.8625
## 11 151.9625
## 12 178.5625
## 13 265.1875
## 14 291.7875
## 15 195.8875
## 16 222.4875
```

```
fitted(model)
```

```
##        1        2        3        4        5        6        7        8
## 235.8875 262.4875 166.5875 193.1875 279.8125 306.4125 210.5125 237.1125
##        9       10       11       12       13       14       15       16
## 221.2625 247.8625 151.9625 178.5625 265.1875 291.7875 195.8875 222.4875
```

```
all(round(yhat - fitted(model), 10)==0)
```

```
## [1] TRUE
```

Yes, the yhats calculated from the matrix-vector computations match the yhats from the model created with lm().

(h)

```
resids = y-yhat
all(round(resids - residuals(model), 10)==0)
```

```
## [1] TRUE
```

Yes, the residuals from the matrix-vector computation match the residuals from the model created with lm().

(i)

```
sd(resids)
```

```
## [1] 22.56238
```

From the model the residual standard error: 26.35

The standard deviation of errors from our matrix-vector computation is slightly different from the model made with lm(). The difference in the estimates is reasonable considering there are only 16 observations.

(j)

```
xtxi <- solve(t(X) %*% X)
varHat = sum(((y-yhat)^2) / 11 )
varB = xtxi * varHat
all(round(varB - vcov(model), 10)==0)
```

```
## [1] TRUE
```

Yes, our matrix computation of $Var(\hat{\beta})$ is the same as vcov(model).

```
sigHat = sqrt(varHat)
(seHatB3 = sqrt(diag(xtxi)) * sigHat)[4]
```

```
## temperaturehi
##      13.17359
```

```
summary(model)$coefficients[4,2]
```

```
## [1] 13.17359
```

Yes, the matrix computation of the $\hat{se}(\hat{\beta}_3)$ matches the $\hat{se}(\hat{\beta}_3)$ from the lm() object.

(k)

```
(R2 = cor(y, yhat)^2)
```

```
##           [,1]
## [1,] 0.8003646
```

```
summary(model)$r.squared
```

```
## [1] 0.8003646
```

Yes, our computed value of R2 agrees with the output from our previously computed lm() model.