Homework #10

Robert Campbell

15 May 2021

Chapter 11

Problem 02

a.

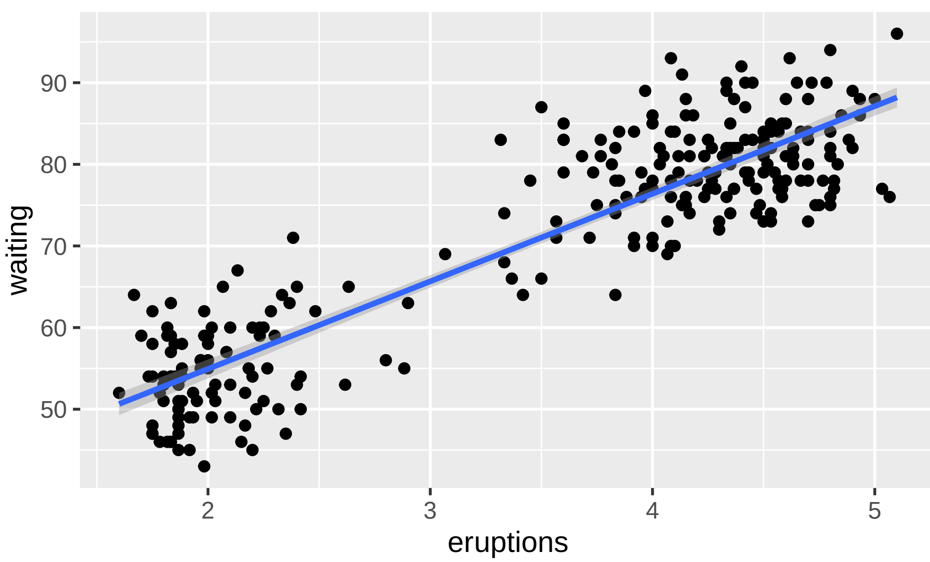
```
faith_mod<-lm(waiting ~ eruptions, data=faithful)
faith_mod
```

```
##
## Call:
## lm(formula = waiting ~ eruptions, data = faithful)
##
## Coefficients:
## (Intercept)    eruptions
##       33.47        10.73
```

b.

```
faithful %>%  ggplot(aes(x=eruptions, y=waiting)) + geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



c.

```
predict(faith_mod, newdata = data.frame(eruptions=4.3))
```

```
##        1
## 79.61186
```

Problem 03

a.

```
j <- ISwR::juul
j<-j %>% filter(tanner==5, age<20)
model <- lm(igf1 ~ age, data=j)
model
```

```
##
## Call:
## lm(formula = igf1 ~ age, data = j)
##
## Coefficients:
## (Intercept)          age
##     1135.49       -38.94
```

b. igf1 = (-38.94)*age + 1135.49
c.

```
predict(model, newdata=data.frame(age=16))
```

```
##       1
## 512.517
```

Problem 04

a. Greater than .03
b. Between -.03 and .03
c. Less than -.03

d. Greater than .03

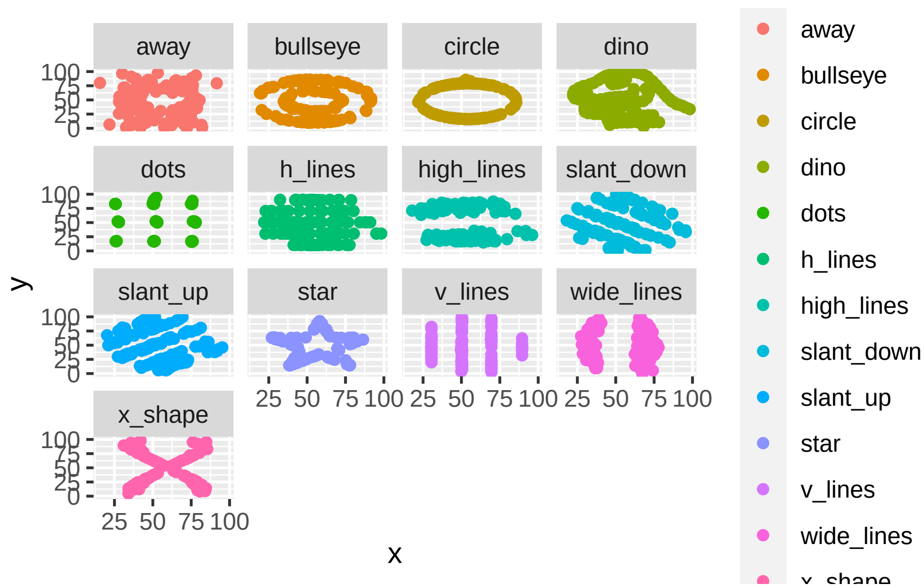
Problem 05

```
d<-datasauRus::datasaurus_dozen
d %>% group_by(dataset) %>% summarize(correlation = cor(x, y, use="complete.obs"),
        xbar=mean(x), sx = (sd(x)/sqrt(142)), ybar=mean(y), sy=(sd(y)/sqrt(142)))
```

```
## # A tibble: 13 x 6
##    dataset    correlation  xbar    sx  ybar    sy
##    <chr>            <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 away           -0.0641  54.3  1.41  47.8  2.26
##  2 bullseye       -0.0686  54.3  1.41  47.8  2.26
##  3 circle         -0.0683  54.3  1.41  47.8  2.26
##  4 dino           -0.0645  54.3  1.41  47.8  2.26
##  5 dots           -0.0603  54.3  1.41  47.8  2.26
##  6 h_lines        -0.0617  54.3  1.41  47.8  2.26
##  7 high_lines     -0.0685  54.3  1.41  47.8  2.26
##  8 slant_down     -0.0690  54.3  1.41  47.8  2.26
##  9 slant_up       -0.0686  54.3  1.41  47.8  2.26
## 10 star           -0.0630  54.3  1.41  47.8  2.26
## 11 v_lines        -0.0694  54.3  1.41  47.8  2.26
## 12 wide_lines     -0.0666  54.3  1.41  47.8  2.26
## 13 x_shape        -0.0656  54.3  1.41  47.8  2.26
```

They all seem to have pretty similar averages and stderr. X and Y seem to largely be not correlated.

```
d %>% ggplot(aes(x=x, y=y, color=dataset)) + geom_point() + facet_wrap(vars(dataset))
```



These graphs are just the shapes described by their names.

Problem 06

General physical activity, income, average hours worked per day. Basically, do museum attendees self-select for lower risk of mortality among known risk factors.

Problem 10

f(b) = sum over i of (yi-B-xi). Take derivative of B with respect to B -2 sum (yi-B-xi) = -2(sum(yi-xi) - n*B) = -2(ybar - xbar -nB) = 0 ybar-xbar = nB, (ybar-xbar)/n = B

Problem 14

a. Yes. Intuitively, at least, denser urban areas seem to have lower school funding.
b.

```
a <- carData::Anscombe
cor(a$education, a$young)
```

```
## [1] 0.3114855
```

c. Slope is significant with p=0.026

```
model <- lm(education ~ young, data=a)
summary(model)$coefficients
```

```
##                Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -20.4247185 94.6640880 -0.2157599 0.83007051
## young         0.6039196  0.2631974  2.2945499 0.02608268
```

d. From residuals, AK is a massive outlier. Updated model shows that the slope is not significantly different from 0.

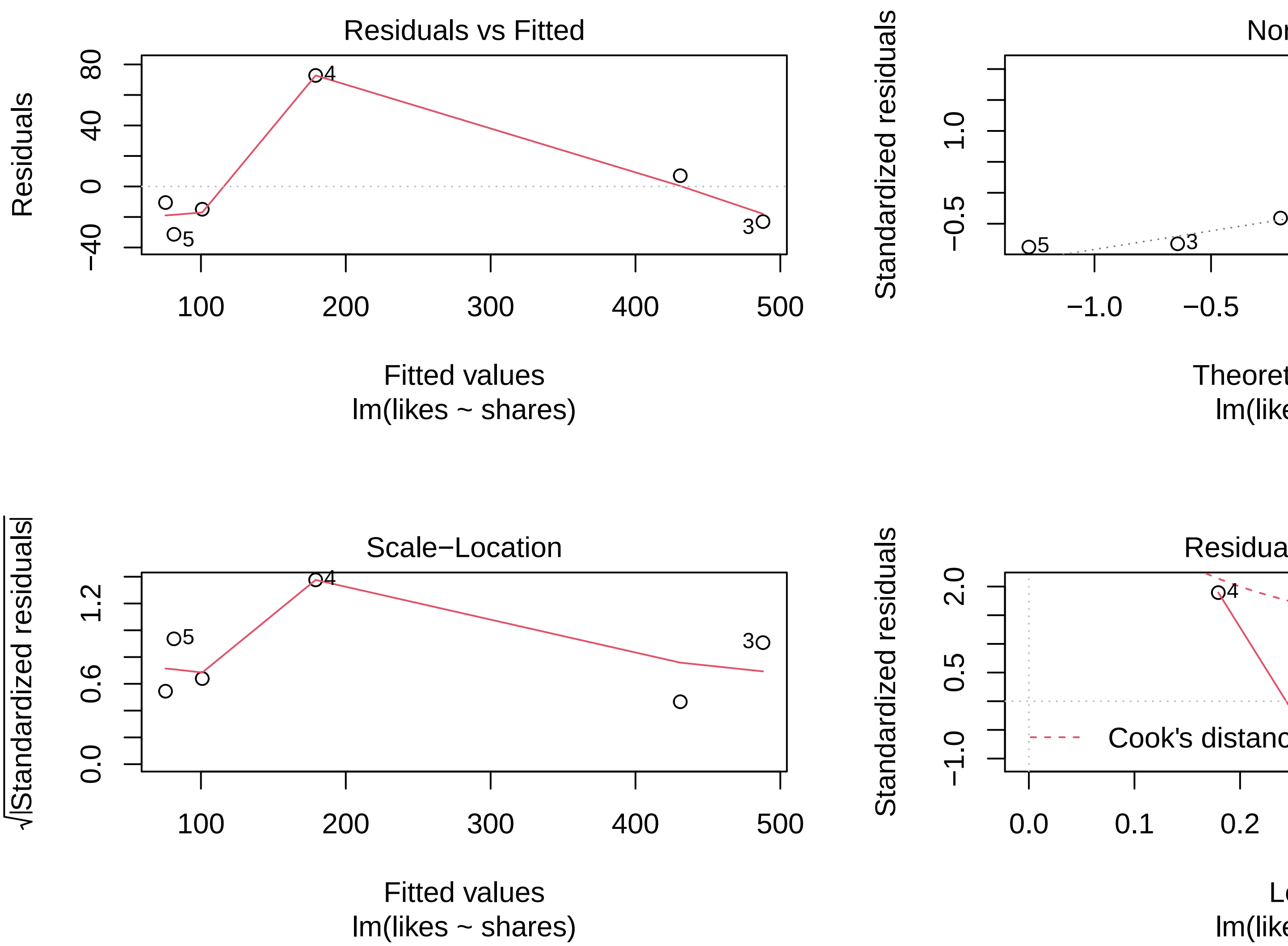
```
a<-a %>% filter(rownames(.)!="AK")
model <- lm(education ~ young, data=a)
summary(model)$coefficients
```

```
##                Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) 146.5563817 95.9488273 1.5274432 0.1332142
## young         0.1294361  0.2680985 0.4827929 0.6314375
```

Problem 15

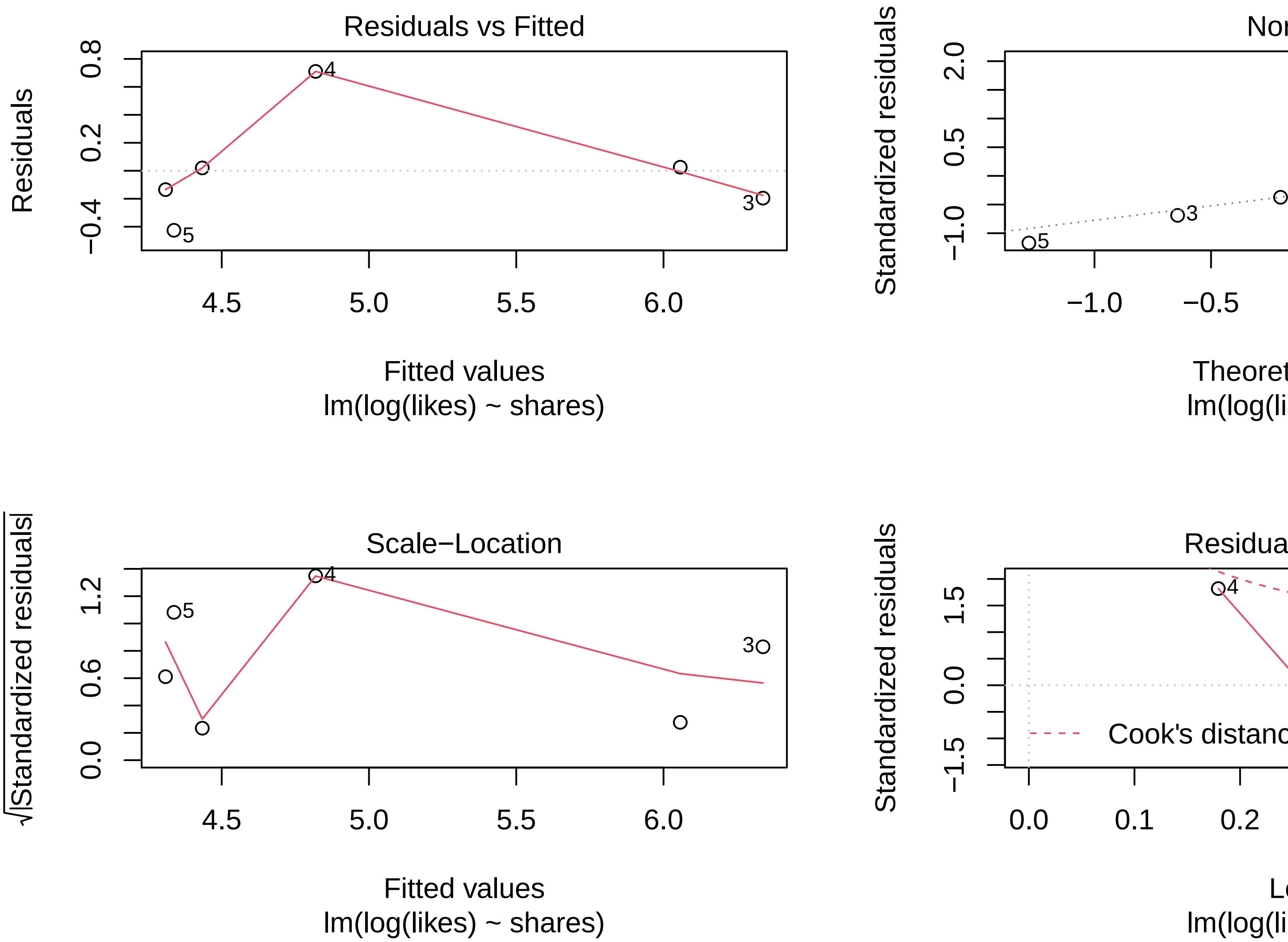
a.

```
c <- fosdata::cern
c<-c %>% filter(platform=="Twitter") %>% head()
model<-lm(likes ~ shares, data=c)
plot(model)
```



b.

```
l_model<-lm(log(likes) ~ shares, data=c)
plot(l_model)
```

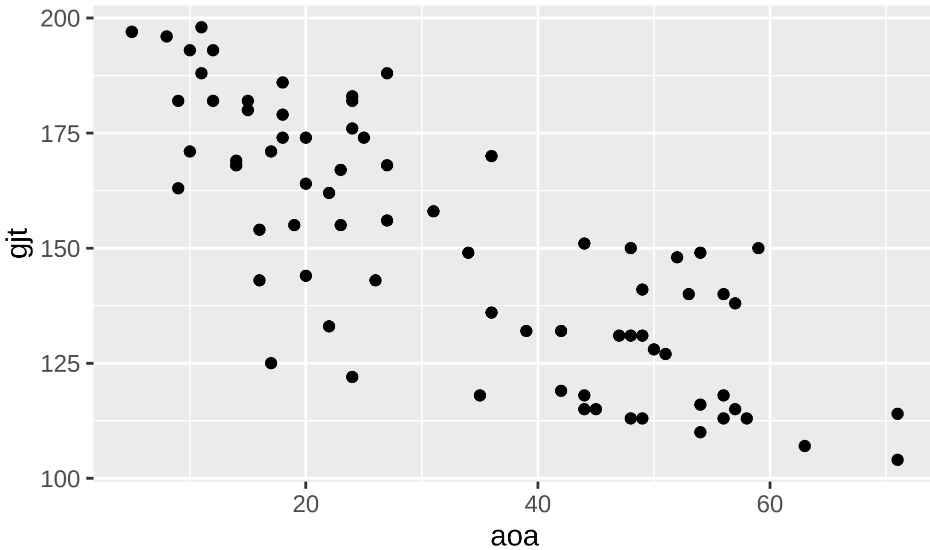


c. The log model looks a little better, but it is still pretty bad.

Problem 16

a.

```
crit <- fosdata::crit_period
crit <- crit %>% filter(locale=="North America")
crit %>% ggplot(aes(x=aoa, y=gjt)) + geom_point()
```



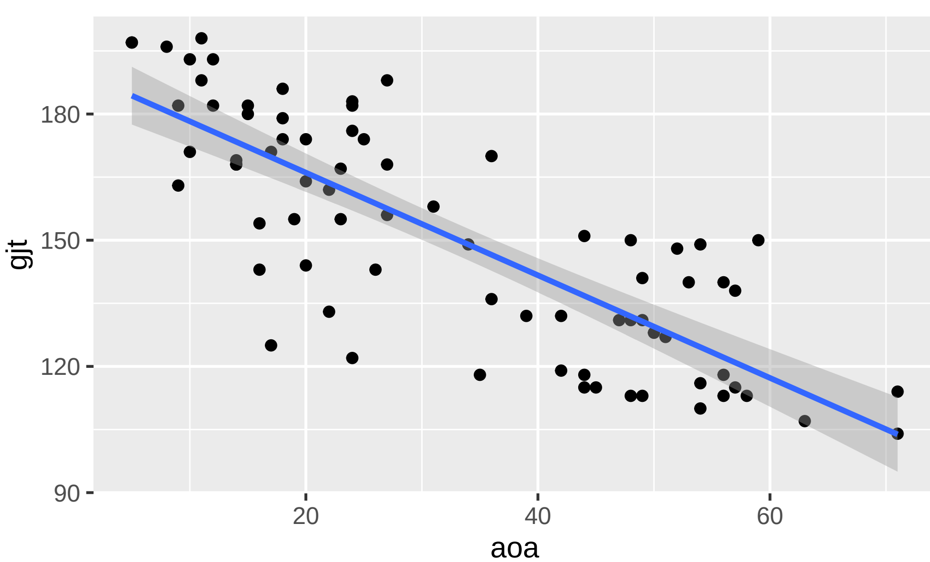
b.

```
model <- lm(gjt ~ aoa, data=crit)
model
```

```
##
## Call:
## lm(formula = gjt ~ aoa, data = crit)
##
## Coefficients:
## (Intercept)          aoa
##      190.46        -1.22
```

```
crit %>% ggplot(aes(x=aoa, y=gjt)) + geom_point() + geom_smooth(method="lm")
```

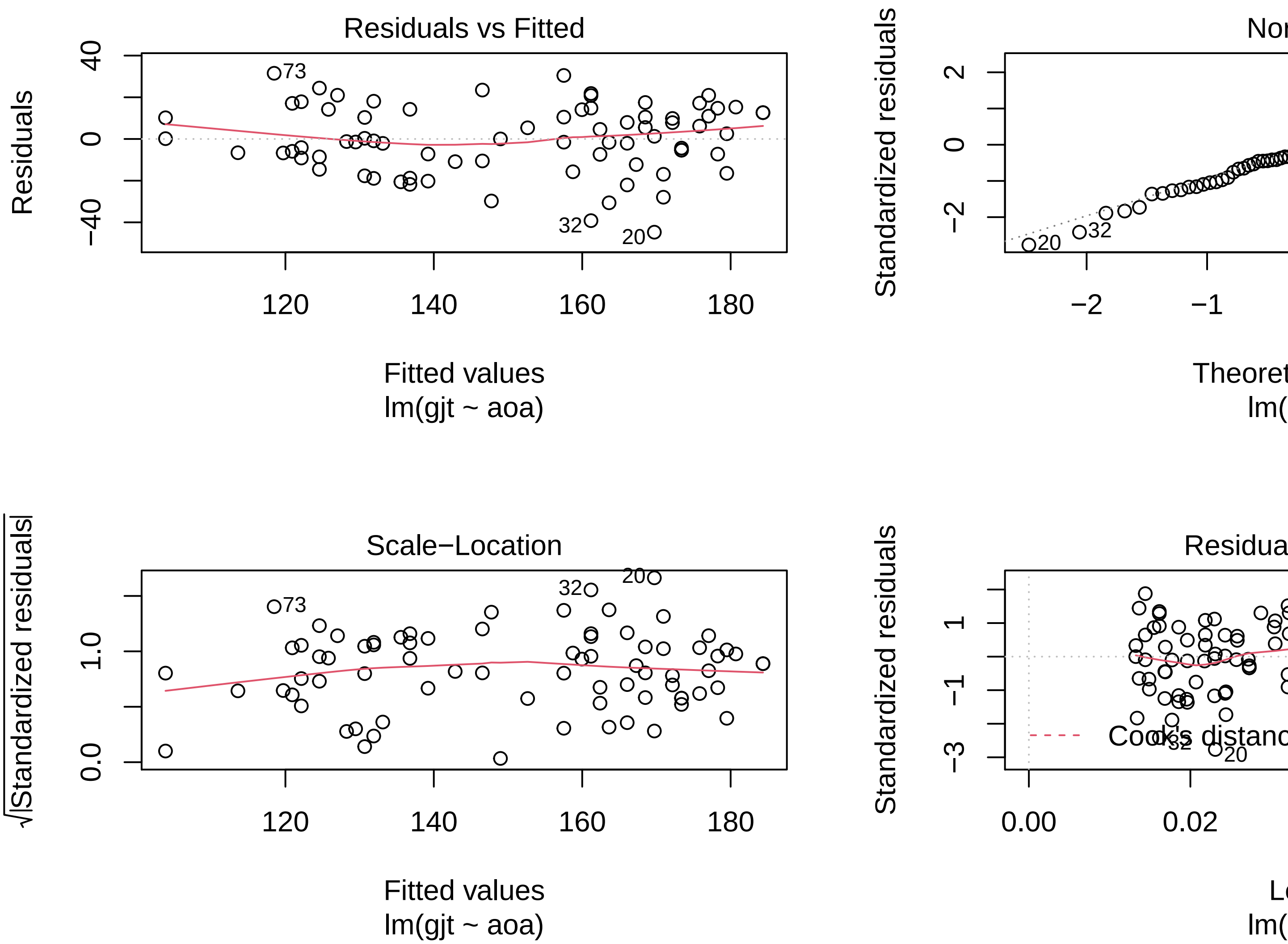
```
## `geom_smooth()` using formula 'y ~ x'
```



c. The slope represent the change is score based on the age someone learns a second language

d. No, there is no distinct bend in the graph.

```
plot(model)
```

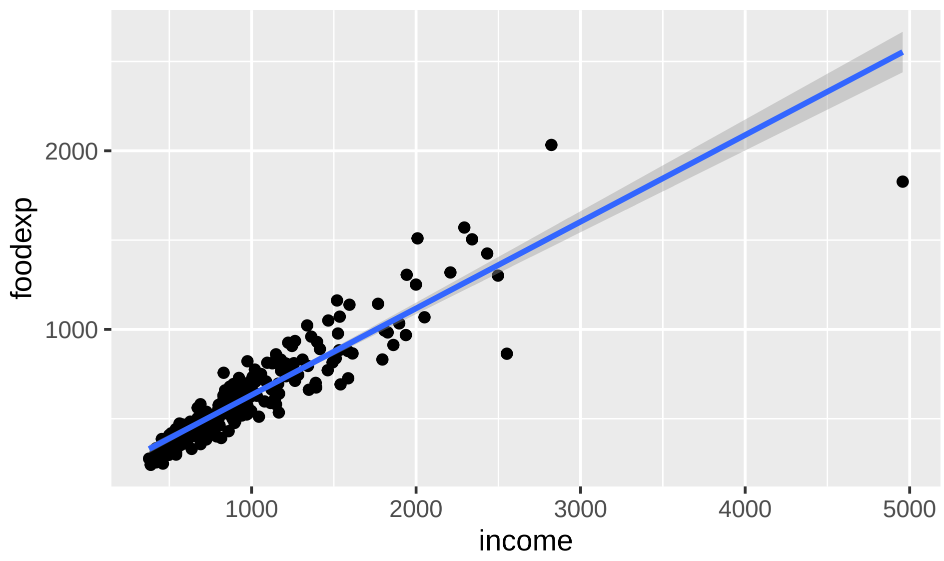


Problem 17

a.

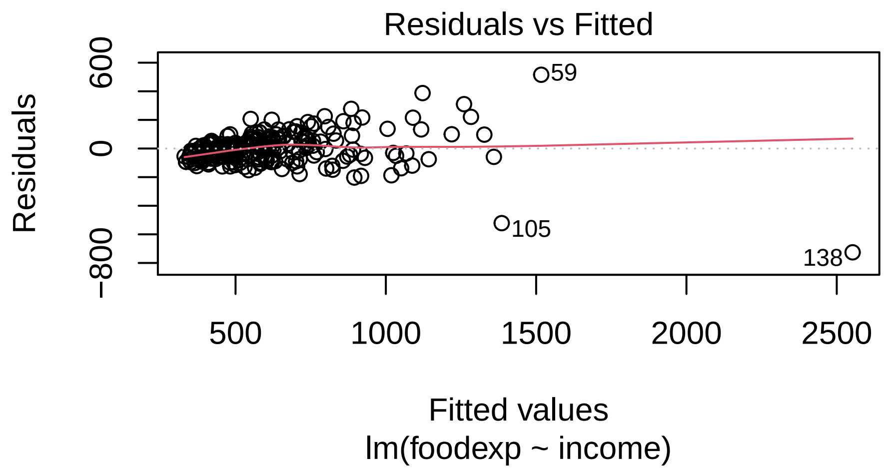
```
data(engel)
e<-engel
e %>%  ggplot(aes(x=income,y=foodexp)) + geom_point() + geom_smooth(method="lm")
```

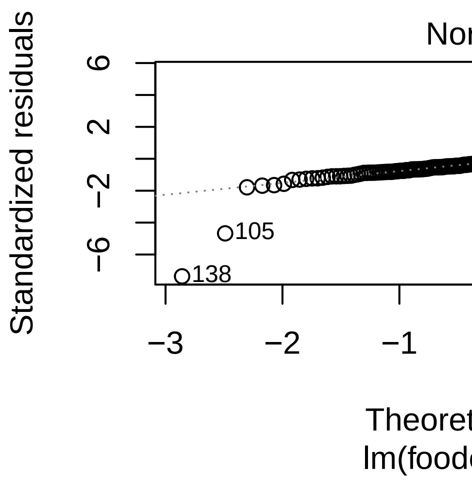
```
## `geom_smooth()` using formula 'y ~ x'
```

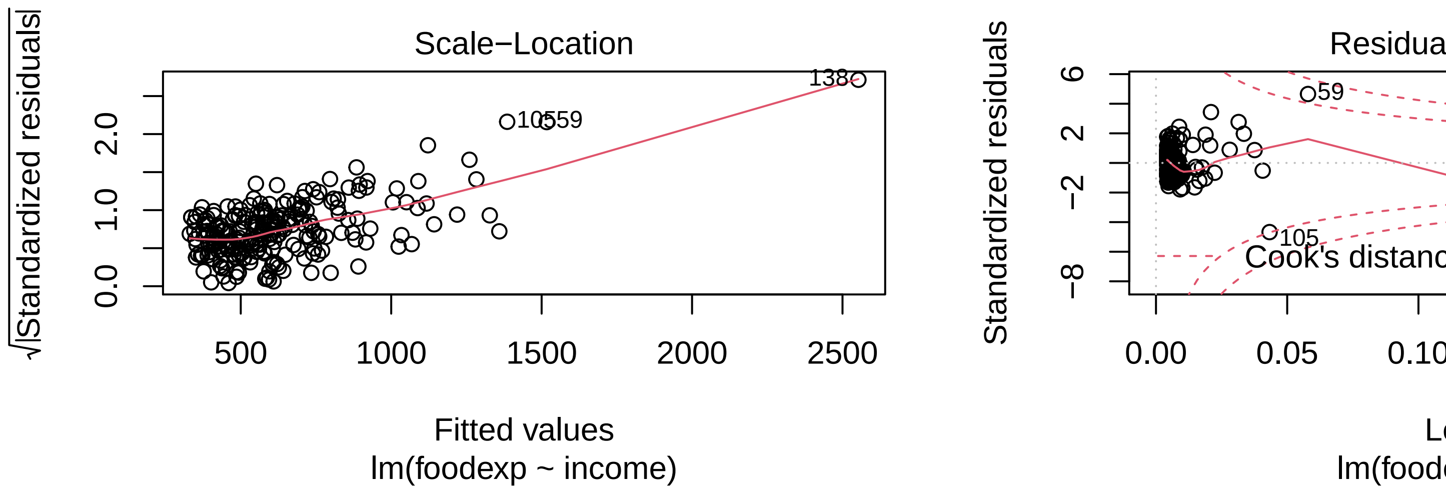


b.

```
model<-lm(foodexp ~ income, data=e)
plot(model)
```







c. Scale-location should be flat instead of trending upwards.

Problem 20

P = 1.33*10^19, slope is significant, reject the null.

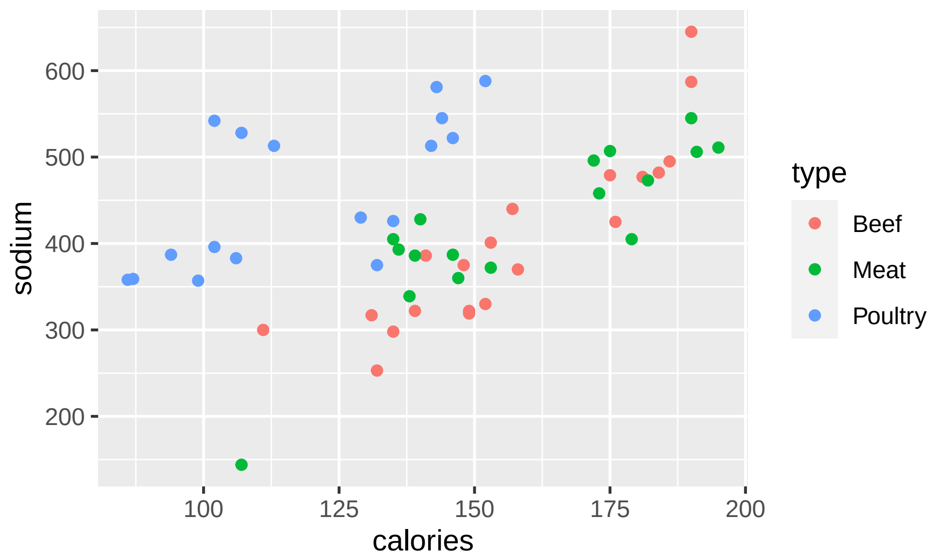
```
pen <- palmerpenguins::penguins
pen<-pen %>% filter(species=="Gentoo")
model<-lm(body_mass_g ~ flipper_length_mm, data=pen)
summary(model)$coefficients
```

```
##                    Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)       -6787.2806 1092.551940 -6.212318 7.649742e-09
## flipper_length_mm    54.6225    5.028244 10.863137 1.330279e-19
```

Problem 21

a.

```
hot <- fosdata::hot_dogs
hot %>% ggplot(aes(x=calories,y=sodium, color=type)) + geom_point()
```



b. Remove poultry the data looks to have 4 separate subgroups. Compare meat and beef instead. sodium = 3.613(calories) - 160.58

```
hot<-hot %>% filter(type!="Poultry")
model<-lm(sodium ~ calories, data=hot)
model
```

```
##
## Call:
## lm(formula = sodium ~ calories, data = hot)
##
## Coefficients:
## (Intercept)     calories
##    -160.580        3.613
```

c.

```
predict(model, newdata=data.frame(calories=140))
```

```
##        1
## 345.1826
```

d.

```
predict(model, newdata=data.frame(calories=140), interval="predict")
```

```
##        fit      lwr      upr
## 1 345.1826 244.4656 445.8995
```

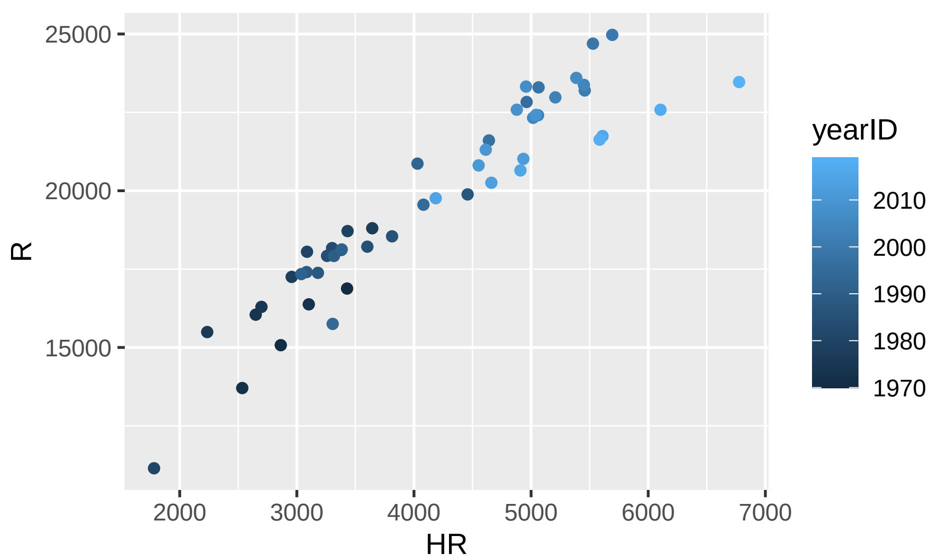
Problem 22

a&b.

```
bat<-Lahman::Batting
dat<-bat %>% group_by(yearID) %>% summarize(HR=sum(HR), R=sum(R))
dat %>% ggplot(aes(x=HR, y=R, color=yearID)) + geom_point()
```

c.

```
dat<-dat %>% filter(yearID>1969)
dat %>% ggplot(aes(x=HR, y=R, color=yearID)) + geom_point()
```



d. The slope predicts 2.54 runs for each homerun. The slope is significant with pval: 1.359*10^-23.

```
model<-lm(R ~ HR, data=dat)
model
```

```
##
## Call:
```

```
## lm(formula = R ~ HR, data = dat)
##
## Coefficients:
## (Intercept)           HR
##     9153.54         2.54
```

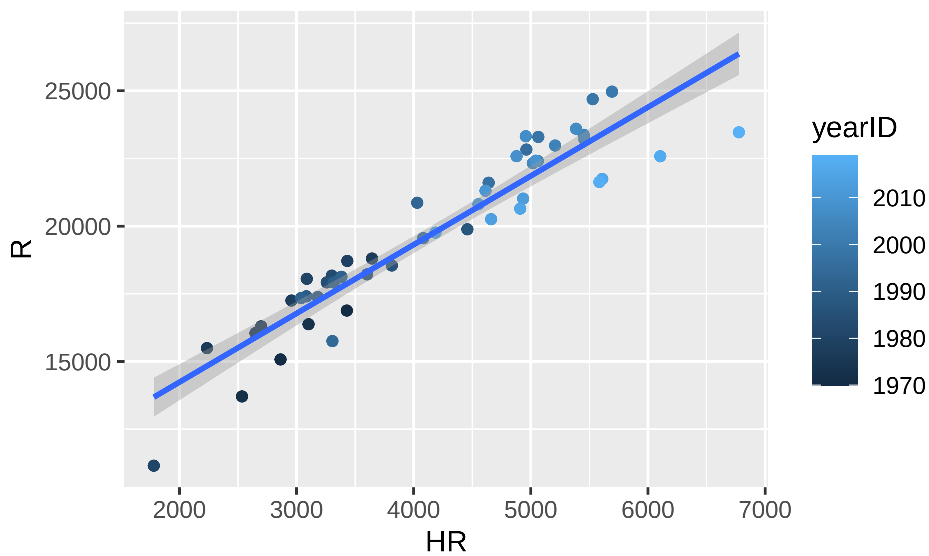
```
summary(model)$coefficients
```

```
##                Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 9153.538545  587.92491 15.56923 2.159845e-20
## HR             2.540337    0.13649 18.61190 1.359170e-23
```

e.

```
dat %>% ggplot(aes(x=HR, y=R, color=yearID)) + geom_point() + geom_smooth(method="lm")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



f. ~19315, it would not be valid for predicting 1870 data. As we saw, there was a significant change in the relationship between HR and R around 1970.

```
predict(model, newdata=data.frame(HR=4000))
```

```
##        1
## 19314.89
```

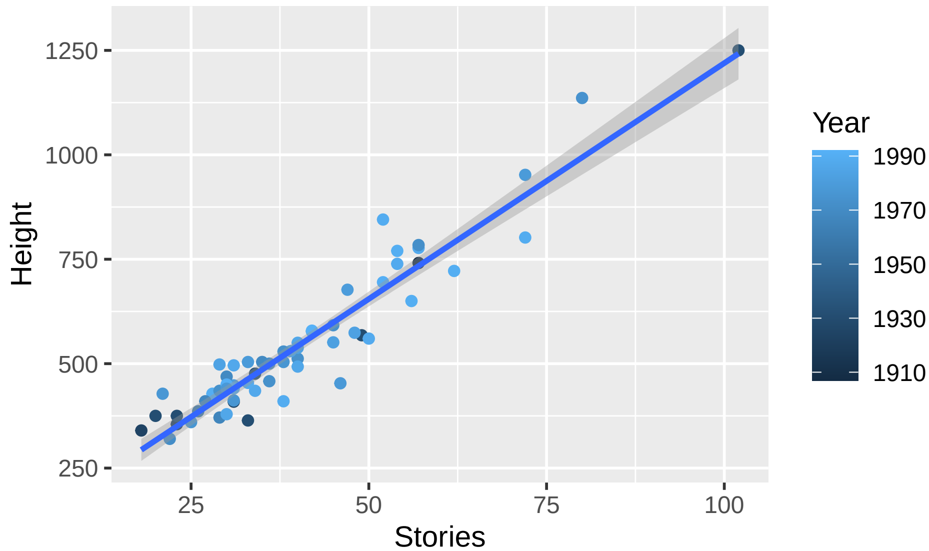
Problem 23

a.
b. Yes, there's a few. One Especially noticeable is one at ~51 stories that is close to 200 (height units?) taller than another 51 story building.

c. Not at a glance. They all seem to follow the same general trend.

```
e <- Sleuth2::ex0728
e %>% ggplot(aes(x=Stories, y=Height, color=Year)) + geom_point() + geom_smooth(method="lm")
```

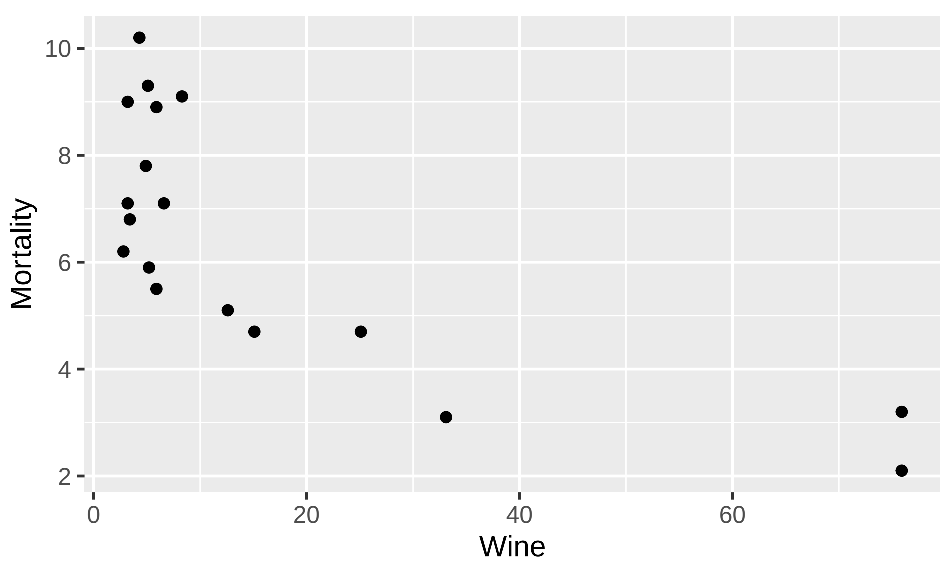
```
## `geom_smooth()` using formula 'y ~ x'
```



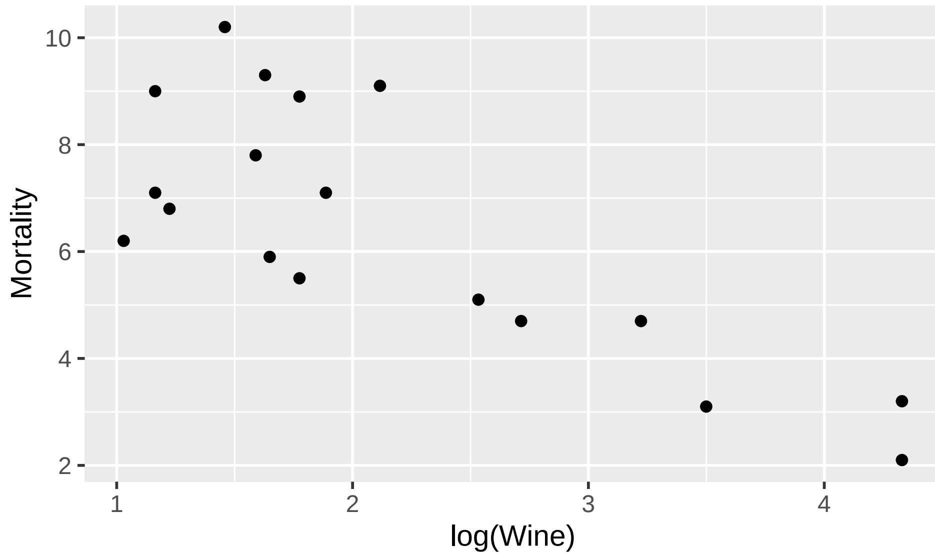
Problem 24

a. The shape is roughly exponential, so the log of wine might help

```
e <- Sleuth3::ex0823
e %>% ggplot(aes(x=Wine, y=Mortality)) + geom_point()
```



```r
e %>% ggplot(aes(x=log(Wine), y=Mortality)) + geom_point()
```



b. Yes, the data suggests there is a correlation between the two

```r
model <- lm(Mortality ~ log(Wine), data=e)
summary(model)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 10.279524  0.8316433 12.360497 1.338392e-09
## log(Wine)   -1.771155  0.3467517 -5.107847 1.053553e-04
```

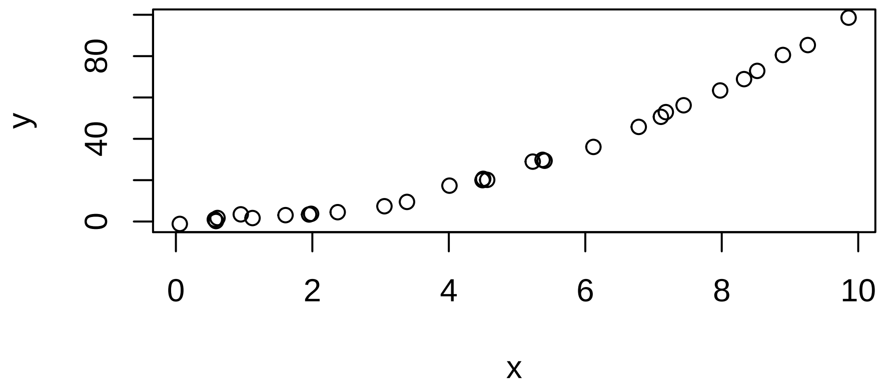
```r
cor(log(e$Wine), e$Mortality)
```

```
## [1] -0.7873139
```

c. No, because correlation does not imply causation.

Problem 25

```r
x <- runif(30, 0, 10)
epsilon <- rnorm(30)  # std normal, mean 0 sd 1
y <- x^2 + epsilon
plot(x,y)
```




```
mod <- lm(y ~ x)

# predict at x=5
pi <- predict(mod, data.frame(x=5), interval="predict")
rate<-replicate(10000000, {xnew <- 5;
ynew <- xnew^2 + rnorm(1);
# check the prediction
(pi[2] < ynew & pi[3] > ynew)})
mean(rate)
```

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

```
# predict at x=10
pi <- predict(mod, data.frame(x=10), interval="predict")
rate<-replicate(10000000, {xnew <- 10;
ynew <- xnew^2 + rnorm(1);
# check the prediction
(pi[2] < ynew & pi[3] > ynew)})
mean(rate)
```

```
## [1] 0.2393947
```

Problem 28

a&b.

```
x = seq(0,10,length.out = 21)

test<- replicate(10000, {epsilon <- rnorm(21, sd=3);
y = 1 + 2*x + epsilon;
model<-lm(y~x);
y[x==5.5] - predict(model, newdata=data.frame(x=5.5))})

sd(test)
```

```
## [1] 2.915049
```

c.

```r
x = seq(0,10,length.out = 21)

test<- replicate(10000, {epsilon <- rnorm(21, sd=3);
y = 1 + 2*x + epsilon;
model<-lm(y~x);
y[x==10] - predict(model, newdata=data.frame(x=10))})

sd(test)
```

```
## [1] 2.725244
```