Final Exam

Robert Campbell

Monday, May 17, 2021

Place your answers into this markdown document, knit it, and hand in the result as a PDF. There are 15 questions, each worth 10 points.

You may use R, the internet, and any reference material, but do not work together and do not get help (except from Dr. Clair).

Honor Pledge

The work I have submitted represents my own effort. While working on this exam, I did not communicate in any form with individuals other than the instructor.

Signed: Robert V Campbell

```
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(ggplot2))
```

Problem 1

Let X be a continuous uniform random variable on the interval [0,5].

a. What is $P(X > 4)$? 0.2

```
1 - punif(4, min=0, max=5)
```

```
## [1] 0.2
```

b. What is the conditional probability $P(X > 4|X > 3)$? = P(x>4 & x>3) / P(x>3) = p(x>4)/P(x>3) = 0.5

```
(1 - punif(4, min=0, max=5)) / (1-punif(3, 0, 5))
```

```
## [1] 0.5
```

c. What is $P(X^2 > 4)$? = P(x > 2) = 0.6

```
1 - punif(2, min=0, max=5)
```

```
## [1] 0.6
```

Problem 2

Let X be a discrete rv which only takes the values 1,2,3,4. The probability distribution for X is

$P(X = 1) = 0.1$, $P(X = 2) = 0.2$, $P(X = 3) = 0.3$, $P(X = 4) = 0.4$.

Find or estimate the mean and standard deviation of X. mean = 3 sd = 1

```
x = c(1,2,3,4)
pb = c(0.1,0.2,0.3,0.4)
y<-replicate(10000, sample(x,1, prob=pb,replace=TRUE))
mean(y)
```

```
## [1] 2.9899
```

```
sd(y)
```

```
## [1] 1.001648
```

Problem 3

As an experiment, you flip a coin three times. Here are some events:

- Event A: Your first flip is heads.
- Event B: Your last flip is heads.
- Event C: You flip more heads than tails.
- Event D: You flip an odd number of heads.

True or false:

1. A and B are independent TRUE
2. A and C are independent FALSE
3. A and D are independent FALSE
4. B and C are independent FALSE
5. B and D are independent FALSE
6. C and D are independent FALSE

Problem 4

Suppose a population has an exponential distribution with $\lambda = 0.5$, so we know it has mean 2. You take a sample of size 60 and test the null hypothesis that $\mu = 2$ with significance level 0.05.

a. Approximate the type I error rate if you use a one sample t-test. Approx 0.065

```
y<-replicate(10000,{x<-rexp(60,rate=.5)

t.test(x,mu=2)$p.value<0.05})
mean(y)
```

```
## [1] 0.0633
```

b. Approximate the type I error rate if you use a Wilcoxon signed rank test. Approx 0.25

```
y<-replicate(10000,{x<-rexp(60,rate=.5)

wilcox.test(x,mu=2)$p.value<0.05})
mean(y)
```

```
## [1] 0.2533
```

c. What should the type I error rate be for these tests? Which test is performing closer to its designed behavior? TypeI rate is supposed to be same as alpha, so approx 0.05. It is working for the t.test, but not the wilcoxon.

Problem 5

Scientists want to estimate the caloric content of almonds. They plan to put two groups of subjects on a controlled diet, with one group recieving an additional daily serving of almonds.

For each subject, measured energy content is approximately normal with a standard deviation of 13 kcal. The researchers hope to detect a difference of 20 kcal between the two groups.

a. With 9 subjects in each group, what power would a t test have at the 0.05 level of significance? 0.86

```
power.t.test(n=9, delta=20, sd=13, type="two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 9
##           delta = 20
##              sd = 13
##       sig.level = 0.05
##           power = 0.8648782
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

b. If they want to detect a smaller energy difference, how would that affect the power of the test? It would reduce the power of the test

```
power.t.test(n=9, delta=10, sd=13, type="two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 9
##           delta = 10
##              sd = 13
##       sig.level = 0.05
##           power = 0.3351694
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

c. If they increase the number of subjects in each group, what would happen to the power? It would increase the power.

```
power.t.test(n=18, delta=20, sd=13, type="two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 18
##           delta = 20
##              sd = 13
##       sig.level = 0.05
##           power = 0.9941588
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Problem 6

According to the Big Bang theory, more distant objects should be moving away from us more rapidly. The data `case0701` from `Sleuth3` has the outward Velocity and Distance of 24 nebulae outside the Milky Way.

a. Find the equation of the regression line for Distance as a function of Velocity. Distance = .001372Velocity + 0.39917

```
e<- Sleuth3::case0701
model<-lm(Distance ~ Velocity, data=e)
model
```

```
##
## Call:
## lm(formula = Distance ~ Velocity, data = e)
##
## Coefficients:
## (Intercept)     Velocity
##    0.399170     0.001372
```

b. Predict the distance from Earth of a nebula with outward Velocity 800 km/s.
1.497units

```
predict(model, newdata=data.frame(Velocity=800))
```

```
##        1
## 1.497096
```

The next two problems use the `msleep` data set from the `ggplot2` library. This has information on sleep and size for 83 species of mammals.

Problem 7

a. Change the `msleep` data so it has a new variable `logbody` which is the logarithm of body weight.

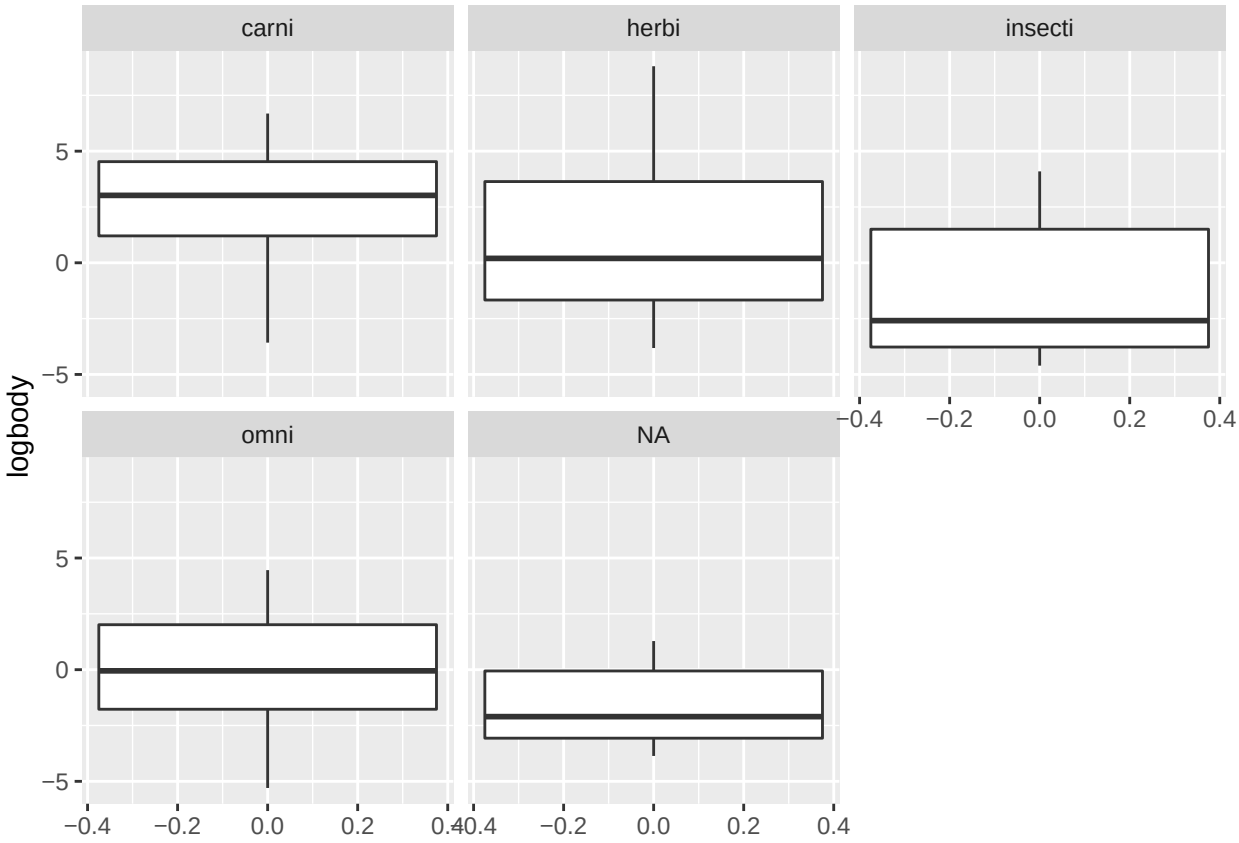
```
sl<-ggplot2::msleep
sl<- sl %>% mutate(logbody=log(bodywt))
```

b. Use `ggplot` to make a boxplot of `logbody` showing with one box for each `vore` type in the data. Don't include the NA values.

Problem 8

Perform analysis of variance to test for a difference in the log of body weight across the different levels of `vore`.

```
sl %>% filter(!is.na(logbody)) %>%  ggplot(aes(y=logbody)) + geom_boxplot() + facet_wrap(vars(vore))
```



Explain your findings with a p-value.

The type of diet of an animal does significantly affect the log of the weight of that animal with a pvalue of 0.034

```
sl %>% filter(!is.na(logbody)) %>% lm(logbody~vore, data=.) %>% anova()
```

```
## Analysis of Variance Table
##
## Response: logbody
##           Df Sum Sq Mean Sq F value  Pr(>F)
## vore       3  90.70 30.2327  3.0605 0.03355 *
## Residuals 72 711.23  9.8782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The next three problems use the data set `cigs` from the `fosdata` library.

Problem 9

a. Which brand has the highest nicotine content (`nic`)?

```
d <- fosdata::cigs
d %>% filter(nic==max(.$nic,na.rm=TRUE))
```

```
##       brand_name flavor co nic tar size filter pack menthol
## 1 English Ovals         15   2  26 King     NF   HP      no
```

b. Make a table showing the number of cigarette brands for each possible `size`.

```
d %>% group_by(size) %>% summarize(numBrand=n())
```

```
## # A tibble: 5 x 2
##   size  numBrand
##   <fct>    <int>
## 1 100        570
## 2 120         28
## 3 80           1
## 4 King       678
## 5 Reg         17
```

c. What is the mean `tar` content for filtered cigarette brands?

```
d %>% filter(filter=="F") %>%  summarize(meanTar= mean(tar,na.rm=TRUE))
```

```
##    meanTar
## 1 10.58326
```

Problem 10

Use t tests to answer these questions, with a significance level of 0.05.

a. Does nicotine content (`nic`) depend on whether a brand is filtered? Yes, there is a correlation with pvalue = 6.2683e-28

```
t.test(nic ~ filter, data=d)$p.value
```

```
## [1] 6.2683e-28
```

b. Does tar content (`tar`) depend on whether a brand is filtered? Yes, there is a correlation with pvalue = 1.70833e-50

```
t.test(tar ~ filter, data=d)$p.value
```

```
## [1] 1.70833e-50
```

c. Does nicotine content (`nic`) depend on whether a brand is menthol? No, reject the alternate hypothesis of correlation with pvalue: 0.956

```
t.test(nic ~ menthol, data=d)$p.value
```

```
## [1] 0.9558782
```

d. Does tar content (`tar`) depend on whether a brand is menthol? No, reject the alternate hypothesis of correlation with pvalue: 0.64

```
t.test(tar ~ menthol, data=d)$p.value
```

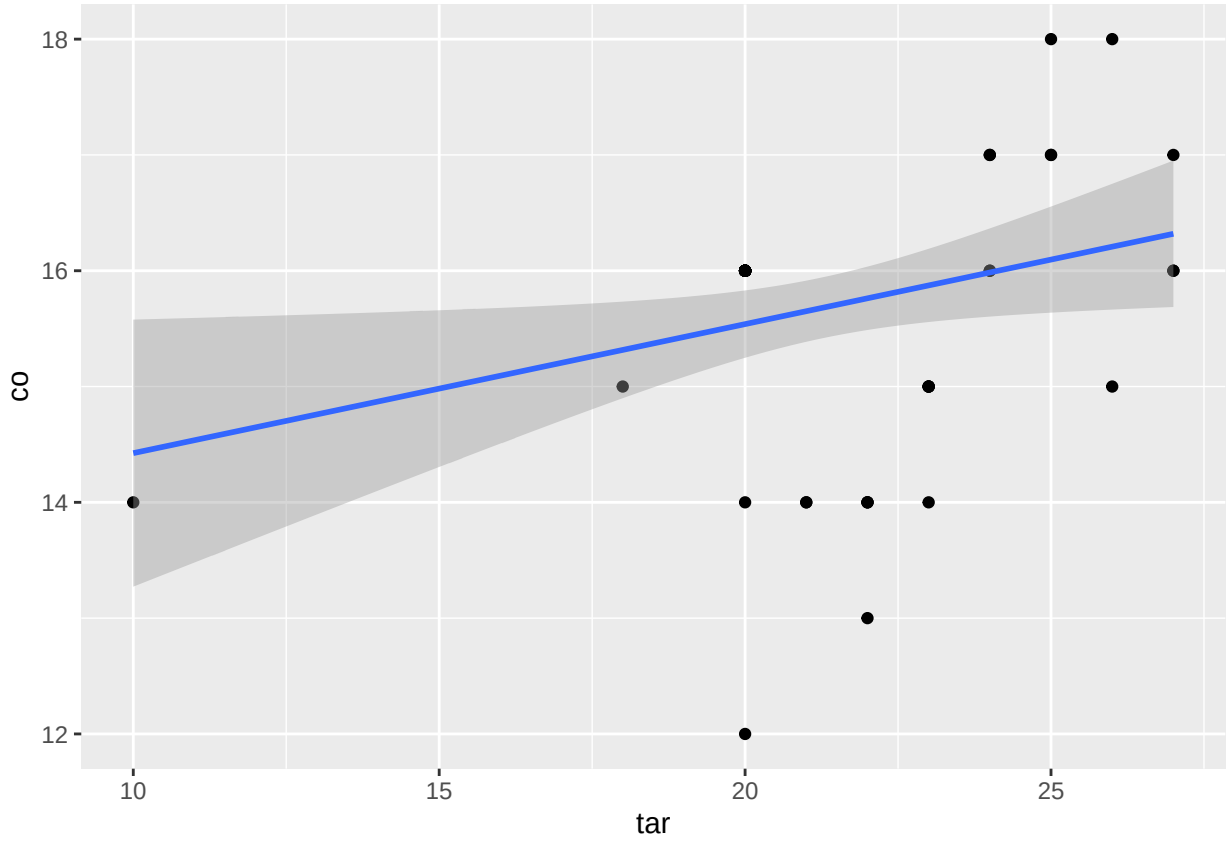
```
## [1] 0.6403627
```

Problem 11

Make a plot of carbon monoxide (`co`) as a function of tar for just the unfiltered cigarettes (`filter == "NF"`). Add the regression line to your plot.

```
d %>% filter(!is.na(co), filter=="NF") %>% ggplot(aes(x=tar,y=co)) +
  geom_point() + geom_smooth(method="lm")
```

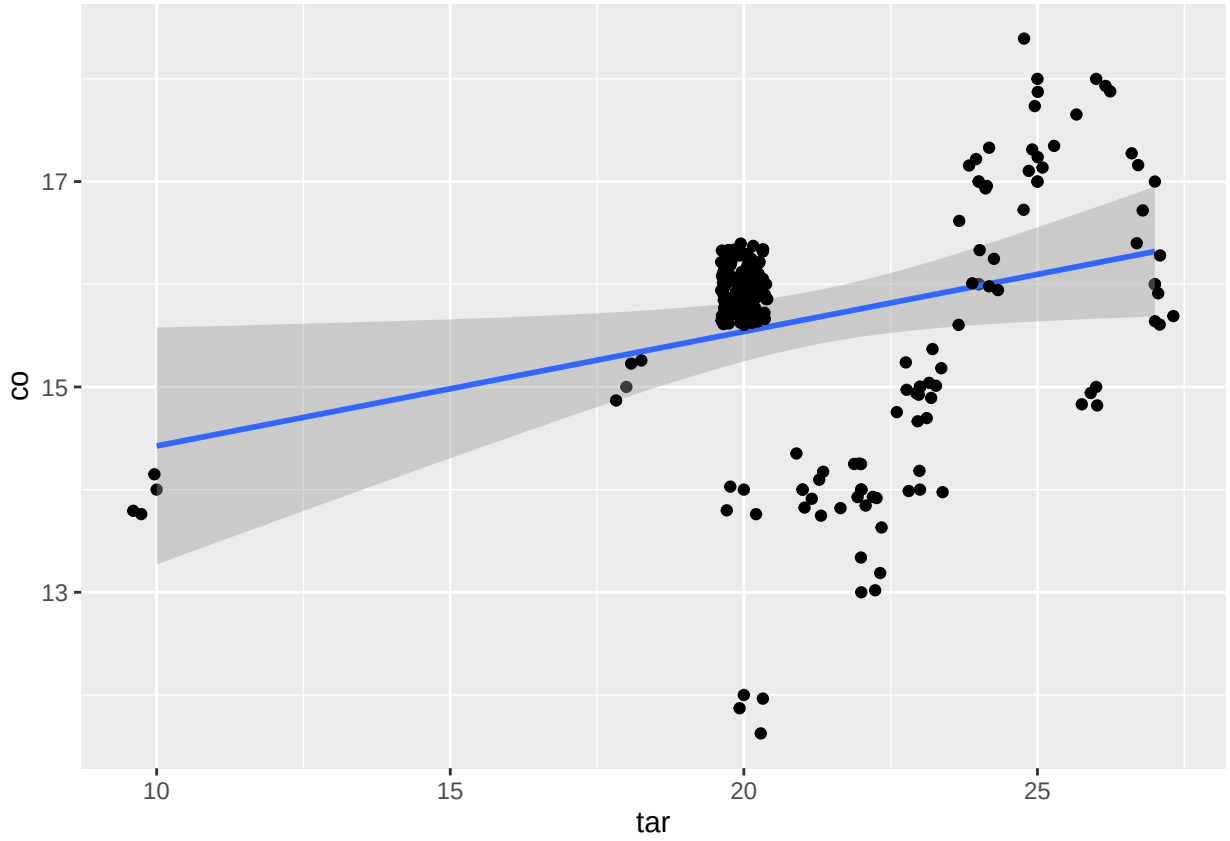
```
## `geom_smooth()` using formula 'y ~ x'
```



The line appears to be a poor fit to the data. Explain what's going on. (Hint: The variables are rounded off. Try `geom_jitter` to randomly move the points a little bit) With the rounding done, there is a large cluster of data around (20,16) which is throwing off the data

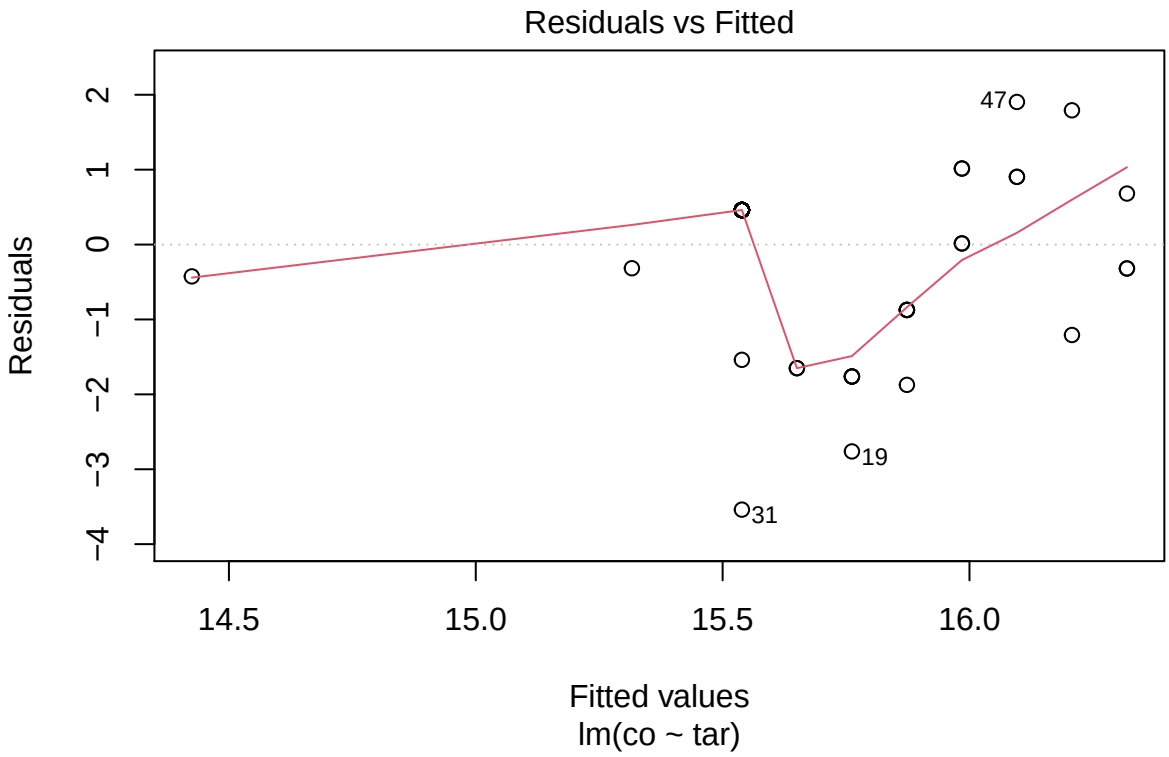
```
d %>% filter(!is.na(co), filter=="NF") %>% ggplot(aes(x=tar,y=co)) +
  geom_point() + geom_smooth(method="lm") + geom_jitter() + geom_jitter() + geom_jitter()
```

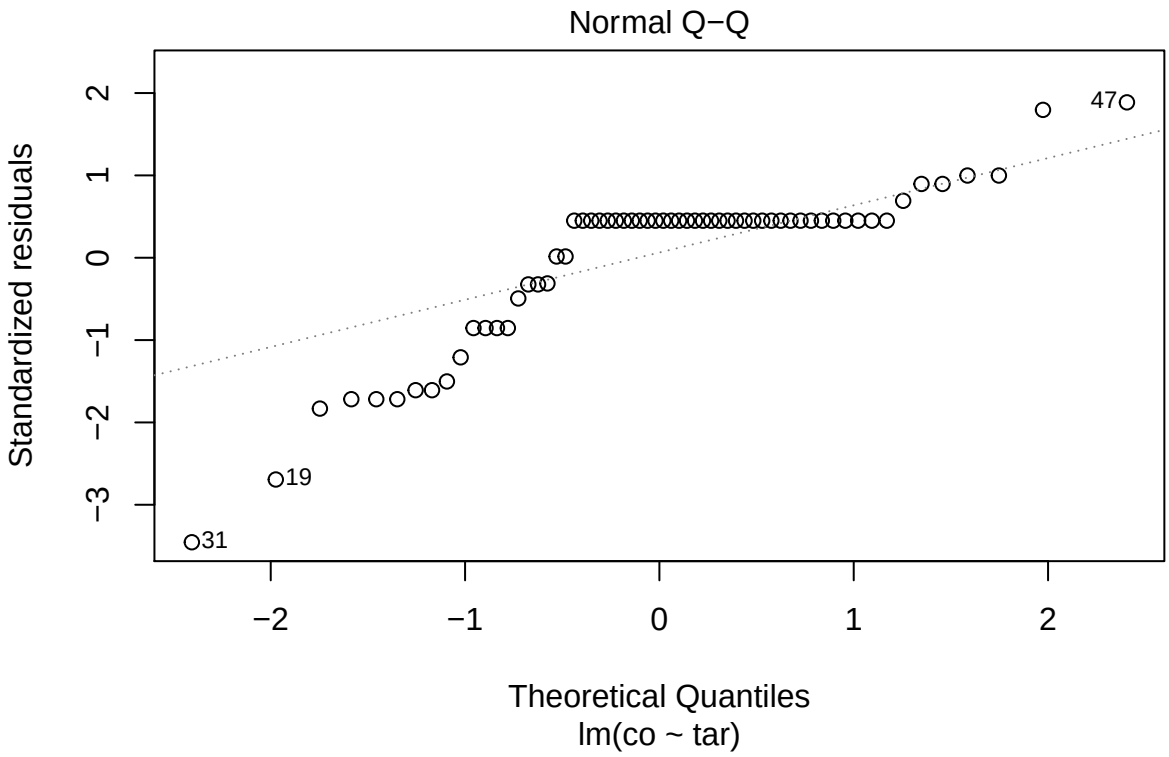
```
## `geom_smooth()` using formula 'y ~ x'
```

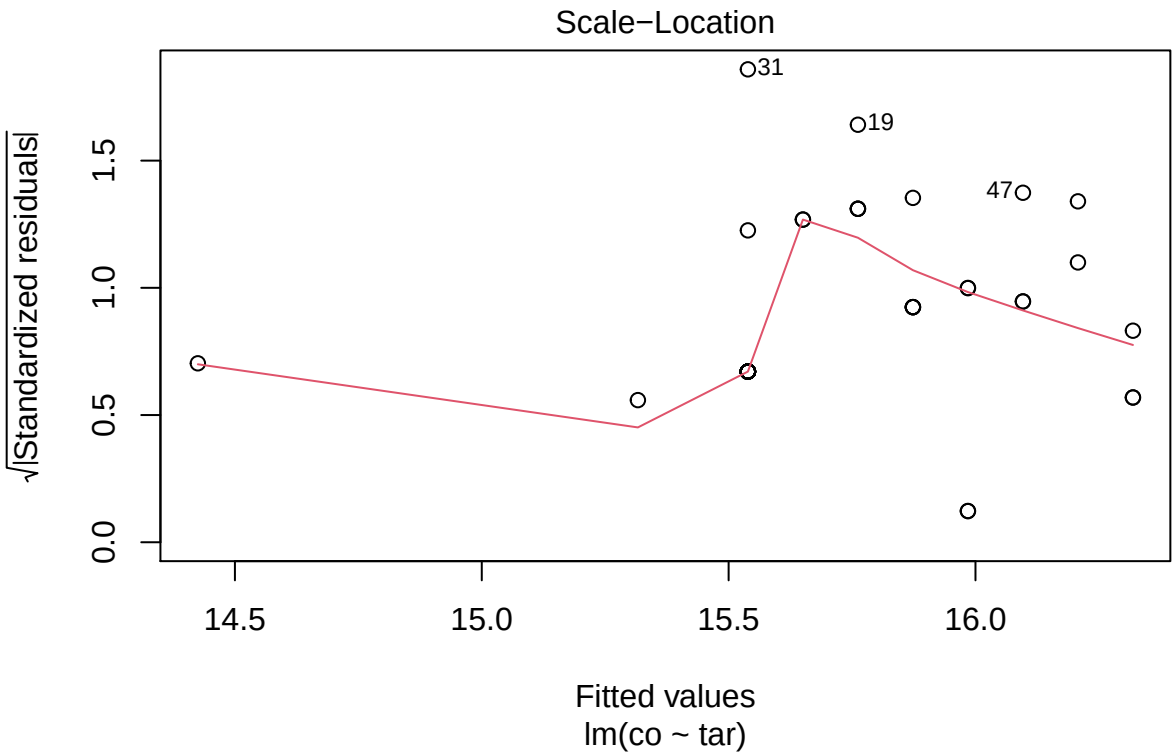


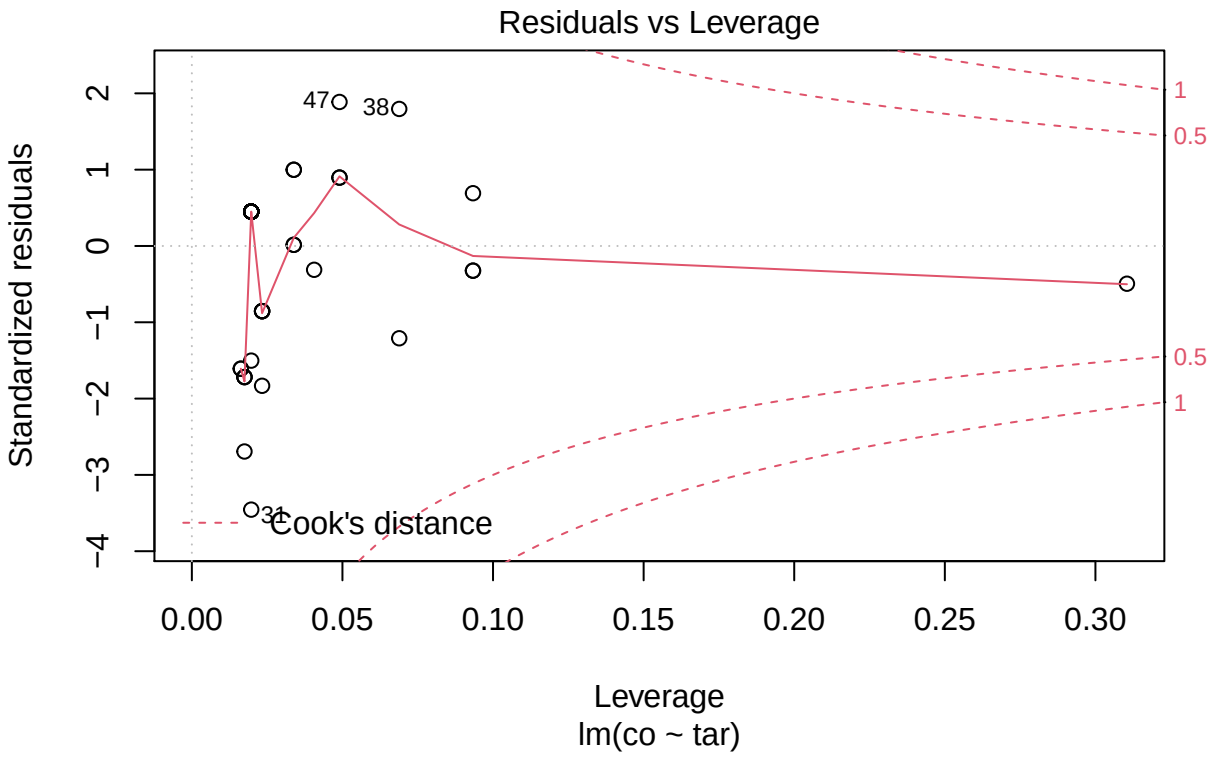
Looking at the residuals, none of the four plots follow the straight lines that they should if the the data would be a good fit for a linear model.

```
model<- d %>% filter(!is.na(co), filter=="NF") %>% lm(co ~ tar, data=.)
plot(model)
```


Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
−4
47
19
31
14.5
15.0
15.5
16.0
Fitted values
lm(co ~ tar)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(co ~ tar)
47
19
31
−3
−2
−1
0
1
2


Scale−Location
31
19
47
√|Standardized residuals|
0.0
0.5
1.0
1.5
14.5
15.0
15.5
16.0
Fitted values
lm(co ~ tar)



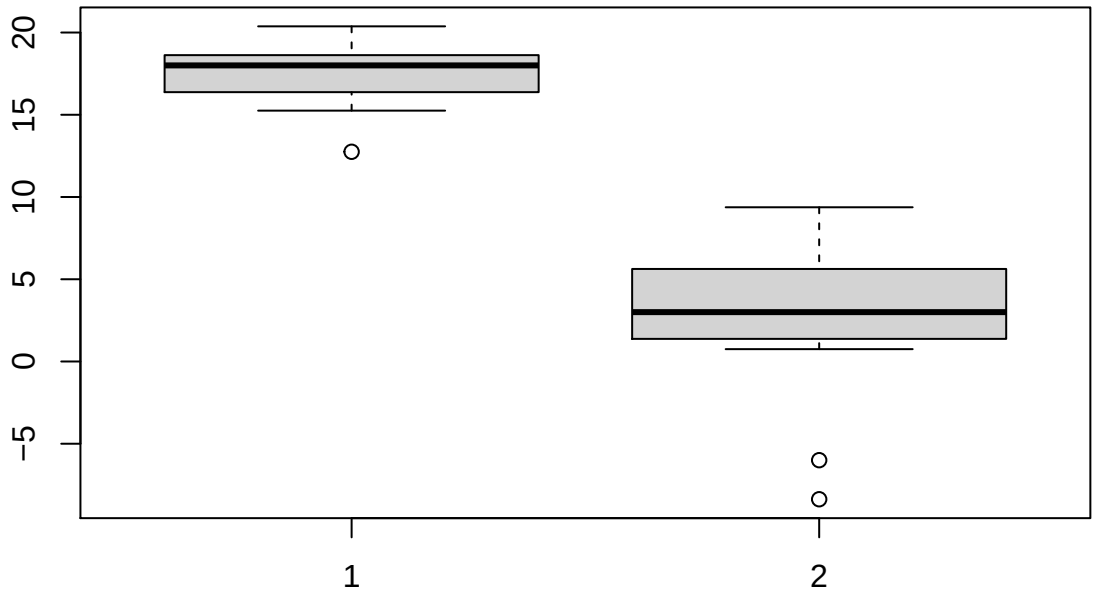
Problem 12

This problem uses the data `ZeaMays` which is in the `HistData` library. Data is from an experiment by Darwin, in 1876.

The variables `cross` and `self` record the heights of pairs of corn plants grown together, one produced by cross fertilization and one produced by self-fertilization.

a. Make a boxplot to compare the heights of self and cross fertilized corn. Self appears taller than diff

```
z <- HistData::ZeaMays
boxplot(z$self,z$diff)
```



b. Is there a significant difference (at the 0.05 level) between heights of cross and self fertilized plants? Choose an appropriate test and report your conclusions with a p value. There is a significant difference with p-value = 2.679*10^-7

```
t.test(z$self,z$diff, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  z$self and z$diff
## t = 9.1766, df = 14, p-value = 2.679e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11.46220 18.45446
## sample estimates:
## mean of the differences
##                14.95833
```

Problem 13

What type of random variable would best model X in each scenario? Normal, uniform, binomial, geometric, poisson, or exponential?

a. X is the total number of calls that arrive at a hotel reception desk between 10am and 10:15am on a Tuesday morning. Binomial

b. X is the total weight of raisins in a family size box of Raisin Bran cereal. Normal

c. A class of 30 first graders is checked for scoliosis (a sideways curve in the spine). X is the number that have scoliosis. Poisson

d. Matt Brockman wants to throw a playing card across the room and have it land in a revolving paper clip. X is the number of tries it takes. Geometric

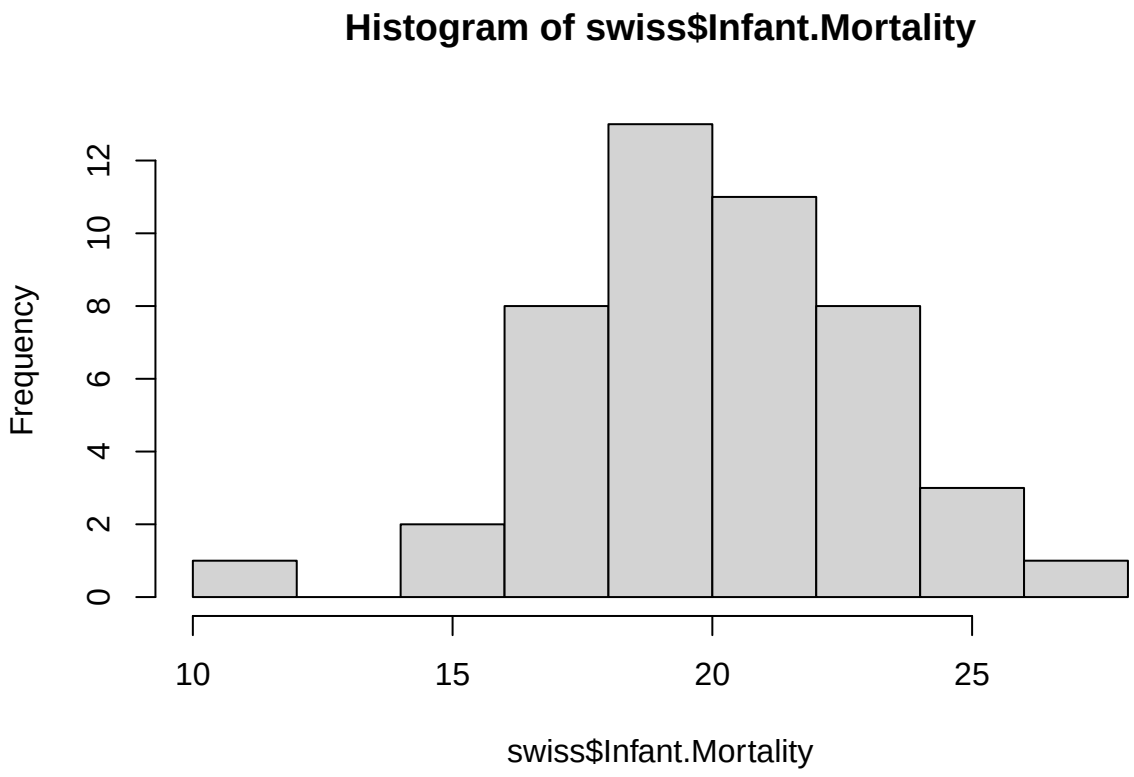
e. X is the blood volume (in ml) of a 1 year old laboratory mouse. Normal

Problem 14

The data set `swiss` is built in to R. It has data on birth and socioeconomic indicators for 47 provinces of Switzerland in 1888.

a. Plot a histogram of infant mortality rates for all the Swiss provinces. Does the data seem reasonably normal? Yes, seems reasonably normal

```
hist(swiss$Infant.Mortality)
```



b. Find the 95% confidence interval for the mean infant mortality percentage in Swiss provinces. 95% conf interval is [19.087%, 20.80%]

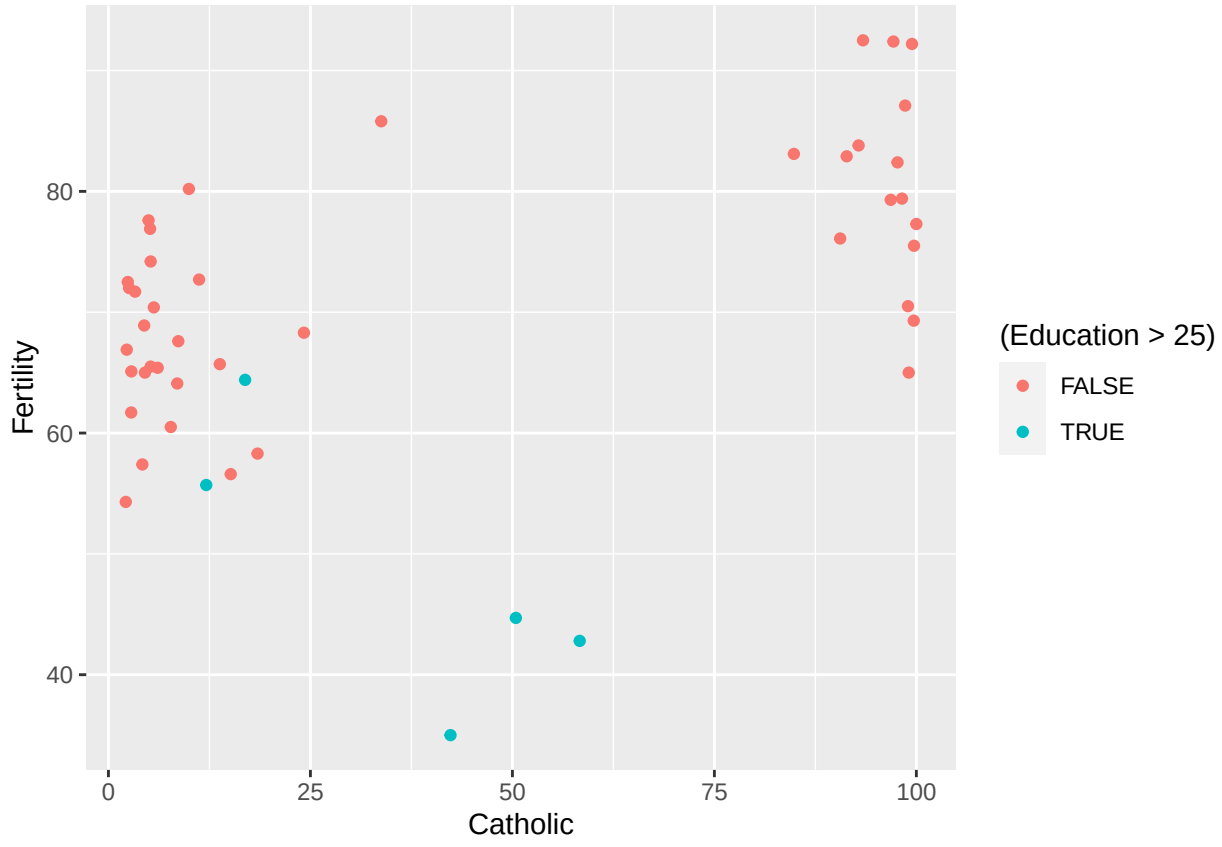
```
t.test(swiss$Infant,conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  swiss$Infant
## t = 46.939, df = 46, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  19.08735 20.79775
## sample estimates:
## mean of x
##  19.94255
```

Problem 15

a. Make a plot that shows how fertility varies in Swiss provinces as a function of the percentage of Catholic residents. Color the points two colors depending on whether the province has an Education rate higher than 25%.

```
swiss %>%  ggplot(aes(x=Catholic, y=Fertility, color=(Education>25))) + geom_point()
```



b. What does this plot tell you about the religious makeup of 19th century Switzerland? That provinces were largely split between whether or not they were catholic

c. How did religion and education interact with fertility in 19th century Switzerland? Religion seems to have affected it some, with fertility rates being somewhat higher in the fully catholic regions but not much more so than the non-catholic regions. Education seems to have played a much larger role with lower education, though there are not many datapoints of <25% education highlighted.

Bonus Problem

In the story "The Boy Who Cried Wolf", the villagers make both Type I and Type II errors. Explain.

There was not a wolf, but the villagers did belive the boy saying there was = Type II. There was a wolf, but the villagers did not believe the boy saying there was. = Type I.