STAT 217 Midterm Exam

Instructions:

This document contains the questions for the midterm exam. Note that you do not need to answer all of these questions, but you should spend some time before October 5 familiarizing yourself with the questions and organizing materials that you will need.

On October 5, you will be randomly assigned questions from the list below. See D2L for your randomized questions. Do not assume that you have been assigned the same questions as others you may frequently work with.

After preparing answers for your assigned questions, you are to record a video (no more than five minutes long) that addresses each question. You should start your video by giving your name. Then state the part, question number, and read the second sentence of the question before giving your answer and explanation. Make sure all assigned questions are addressed. End your video by addressing anyone that you worked with on this exam. Failure to acknowledge collaboration will result in a zero and may lead to a report to the Dean of Students.

TechSmith must be used to record videos. A link to your recorded video is due to Gradescope by October 7 at 11:00 PM. Your submission will be graded on four parts (two questions, addressing collaborative efforts, and time).

Resources: You are allowed all provided materials (videos, notes, textbook, etc.) and access to discussions with other students, the MLC, other tutors, and can ask questions of your instructor. You must document the discussions/help that you have (other than with your instructor) in preparing your answer.

See full set of instructions along with information for recording with TechSmith in Midterm Announcement posted on D2L.

Part I (10 points):

Choose the most appropriate analysis from the following list of options: two-sample mean test, One-Way ANOVA, Two-Way ANOVA, Chi-squared independence, Chi-squared homogeneity, or none of the above. In your recorded video, **explain your choice of analysis**.

1) A company is interested in exploring what characteristics of potential customers are related to eventual sales. In particular, they collect information on the yearly income category of the customer (less than $10,000, between $10,000 and $50,000, over $50,000 per year) and the size of sales that were made (none, less than $20, between $20 and $100, over $100). The data are collected from visitors to the website based on a random sample of all visitors over a week.

2) A researcher is developing a new instrument for measuring air pollution which is much cheaper to make than the current instrument. They set up the new instrument next to the current one and take measurements every minute for a month. In order to obtain less dependent measurements, they randomly sample a pair of measurements from both machines at the same time from each day of the month (so they ended up with n=30). Air pollution measurements for both instruments are obtained in parts per million. The researchers are interested in assessing the relationship between the measurements obtained from the two machines.

3) A researcher is studying different fertilization methods for fruit tree production and whether the different types of fertilizers work differently in different varieties of trees. In a particular orchard, they randomly assign 10 pear trees from each of three varieties (Bradford, Bosc, and Chojuro) to get one of three different types of fertilizer (brands A, B, and C), so n=90 trees in the study. They measure the total amount of fruit produced in a season in total pounds from each tree.

4) Researchers are interested in studying the pitch of a child's voice based on the pitch of the cry of the same child when they were a baby. They measured the pitch for each child in the study (n=15) when the child was four months old and then again when they were 5 years old and recorded the frequency (in Hertz). They categorized the measured pitches at four months into low (350 to 425 Hz), medium (425 to 500 Hz), and high (over 500 Hz), but they did not categorize their pitch at 5 years old.

5) We are all interested in the speed of our internet connections and how we might optimize its performance - and how much of a difference there might be in performance based on our choices. One aspect of that is the choice between 2.4 Ghz vs 5 Ghz wifi. In order to study this, one of your colleagues uses an internet speed test from the same computer sitting in the same location, at a randomly sampled time of day each day, every day for month (so n=30). At the selected time, a coin flip is used to decide between connecting to 2.4 or 5 Ghz. If it is heads, then 2.4 is used, if tails then 5 Ghz is selected. Then the speed test is performed, recording the upload speed in Mbps.

6) We are all interested in the speed of our internet connections and how we might optimize its performance - and how much of a difference there might be in performance based on our choices. Another aspect of that choice is the brand of wifi router used. Four different brand new routers are purchased that purport to have the same specifications but are made by one of four different companies. Each is configured on the same network. On each day of the study, a router is randomly selected to be used on that day for the test and just one test measurement is performed per day. Each router is used on 5 different days (so n=20). At the same time of day each day, the internet speed is measured that combines information from both upload and download speeds, with categories of slow, medium, fast, and extremely fast.

7) A researcher wants to explore sources of kidney damage from pesticides. A national data base is to be used that measured hundreds of attributes about the subjects based on a single blood and urine collection event from each participant. Using this information, they can classify kidney damage into one of four categories, from none, minor, moderate, or high. They also collect information on exposure to a common pesticide and categorize it as none, low, medium, and high.

8) A Chamber of Commerce is interested in understanding what drives purchases at local stores and comparing different ways to possibly stimulate local purchases. They purchase a list of addresses and randomly assign 500 addresses to receive one of three different advertisements: a "buy local" color door hanger placed on their front door, a letter explaining how important it is to buy local, or they get sent a coupon for 5 dollars off a purchase at a local store to go with the letter explaining how important it is to buy local. Each participant is sent a survey that asks for how much money they spent (USD) locally in the two weeks that followed the study initiation.

9) It is expensive to get exact body fat percentage measurements, with machines costing thousands of dollars used in laboratory research. Researchers are interested in building a model that they can use to predict body fat using simple and easy anthropomorphic measurements. Researchers obtain a sample of three hundred subjects and measure their body fat using a "BOD POD", and also measure their age (years), body mass index (BMI, kg/m^2), hip circumference (cm), waist circumference (cm), height (cm), and weight (kg). They want to use all the easy to obtain measures to explain the observed body fat percentages.

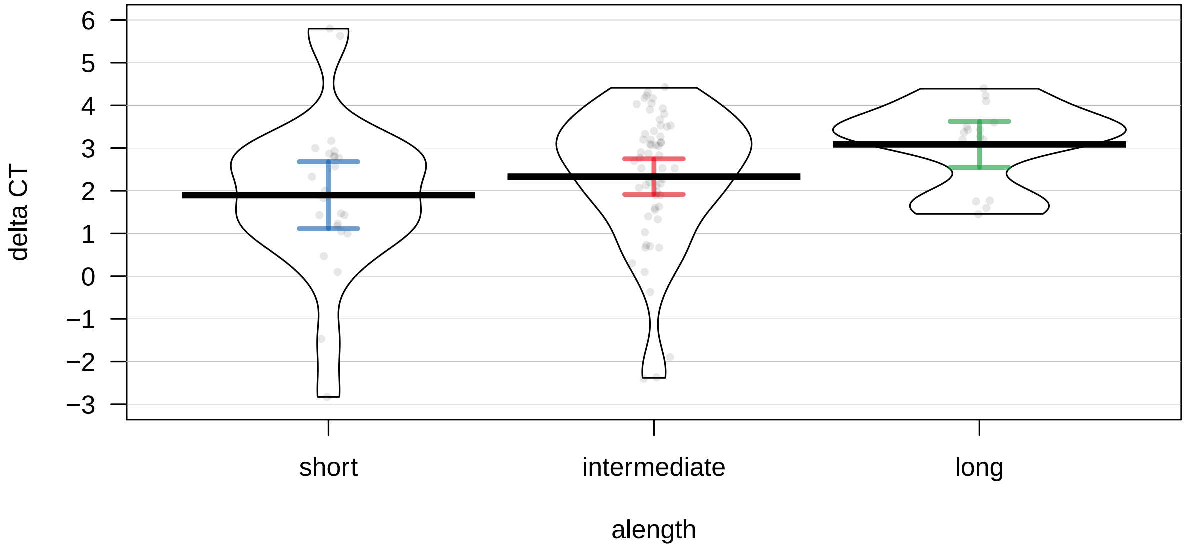
Part II (15 points):

Genetics of alcohol craving

Various studies have linked alcohol dependence phenotypes to chromosome 4. One potential gene is *NACP* (non-amyloid component of plaques), coding for alpha synuclein. Higher expression of alpha synuclein has been found in other studies to be associated with alcohol craving. In one study in Germany of alcohol-dependent subjects (volunteers) in Bonsch et al. (2005), the level of expression of alpha synuclein mRNA was measured (`elevel`, units of “delta CT” which relate to relative expression fold change between this an internal standard and the alpha synuclein target, you can just call it “delta CT points” for your answers) and the allele length of NACP (the potential gene of interest) was classified into short, intermediate, or long groups (`alength`) for each subject. The subjects were all part of a larger study of alcoholism in Germany and there was no indication of random sampling in the discussion of the selection of the subjects, so you can assume that these were subjects who volunteered that were part of the larger study. Use the following output to answer these questions.

- Note: While not needed to answer the questions, the original paper is available at https://academic.oup.com/hmg/article/14/7/967/626665
- Bonsch, D., Lederer, T., Reulbach, U., Hothorn, T., Kornhuber, J., and Bleich, S. (2005) Joint analysis of the NACP-REP1 marker within the alpha synuclein gene concludes association with alcohol dependence, *Human Molecular Genetics*, 14(7), 967–971.

```
data("alpha", package = "coin")
alpha$alength <- factor(alpha$alength)
library(yarrr)
pirateplot(elevel ~ alength, data = alpha, inf.method="ci",
           inf.disp="line", theme=2, ylab= "delta CT")
```



```
library(mosaic)
favstats(elevel ~ alength, data = alpha)
```

```
##        alength   min     Q1 median     Q3   max     mean         sd   n missing
## 1        short -2.83 1.1400  1.915 2.8175 5.80 1.897917 1.8548268 24       0
## 2 intermediate -2.40 1.6075  2.735 3.3825 4.43 2.332069 1.5808245 58       0
## 3         long  1.45 2.4850  3.370 3.5500 4.40 3.086667 0.9738632 15       0
```

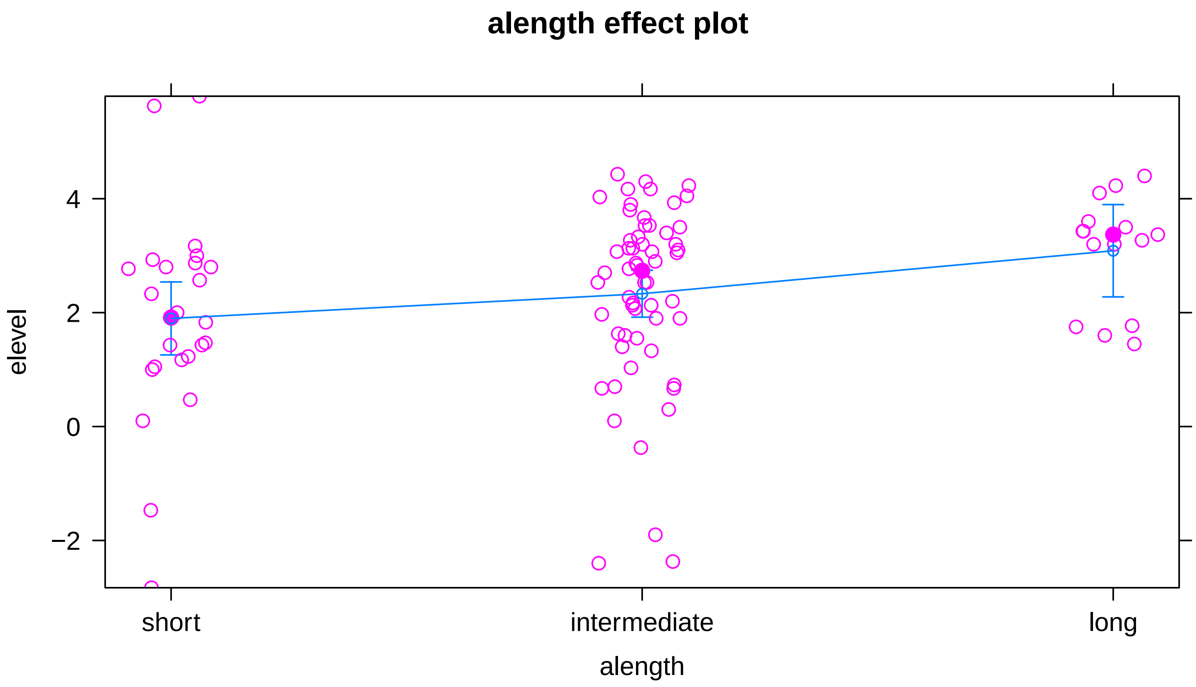
```
amod <- lm(elevel ~ alength, data = alpha)
summary(amod)
```

```
##
## Call:
## lm(formula = elevel ~ alength, data = alpha)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7321 -0.7321  0.2833  0.9721  3.9021
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.8979     0.3226   5.882 6.17e-08
## alengthintermediate   0.4342     0.3836   1.132   0.2606
## alengthlong           1.1888     0.5203   2.285   0.0246
##
## Residual standard error: 1.581 on 94 degrees of freedom
## Multiple R-squared:  0.05267,    Adjusted R-squared:  0.03251
## F-statistic: 2.613 on 2 and 94 DF,  p-value: 0.07863
```

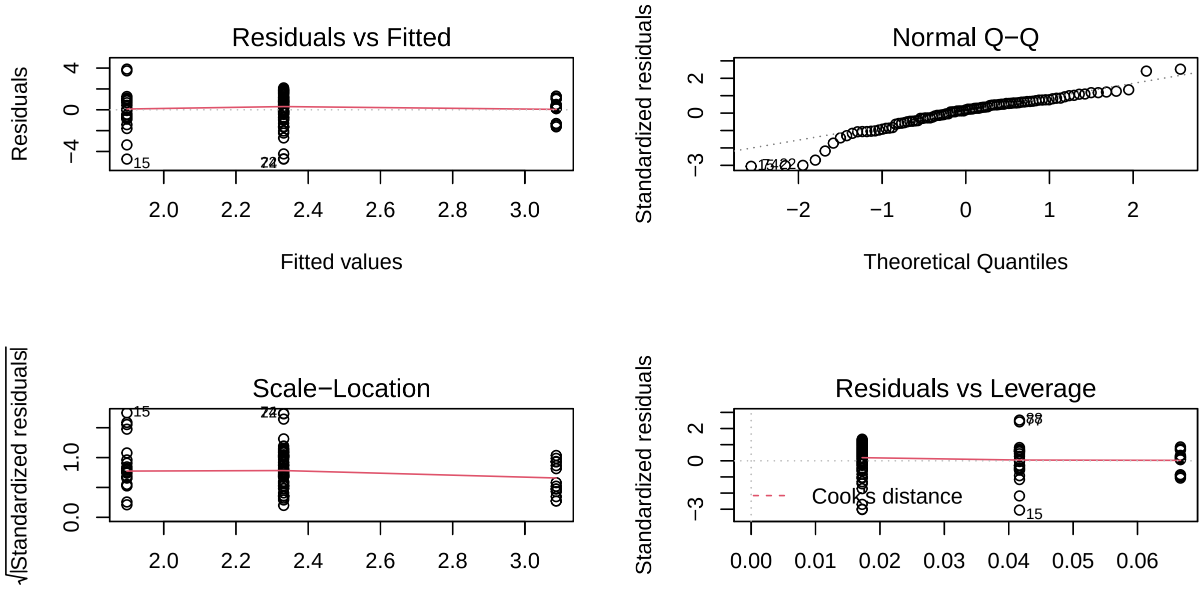
```
library(car)
Anova(amod)
```

```
## Anova Table (Type II tests)
##
## Response: elevel
##            Sum Sq Df F value  Pr(>F)
## alength    13.057  2   2.613 0.07863
## Residuals 234.850 94
```

```
library(effects)
plot(allEffects(amod, residuals=T))
```



```r
par(mfrow=c(2,2))
plot(amod)
```



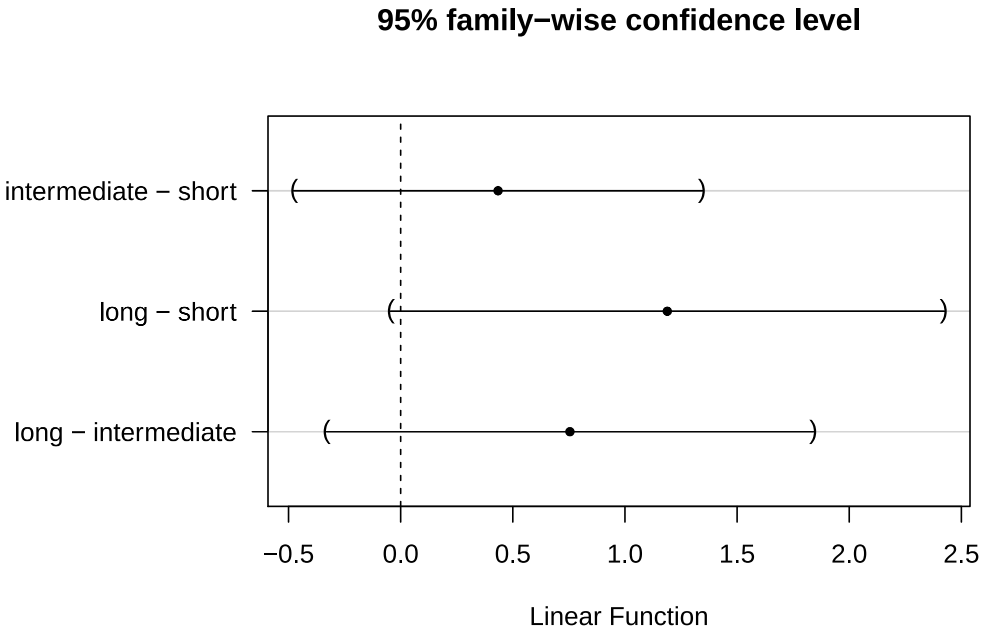
```r
par(mfrow=c(1,1))

library(multcomp)
amod_glht <- glht(amod, linfct = mcp(alength = "Tukey"))
summary(amod_glht)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = elevel ~ alength, data = alpha)
##
## Linear Hypotheses:
##                             Estimate Std. Error t value Pr(>|t|)
## intermediate - short == 0     0.4342     0.3836   1.132   0.4924
## long - short == 0             1.1888     0.5203   2.285   0.0614
## long - intermediate == 0      0.7546     0.4579   1.648   0.2270
## (Adjusted p values reported -- single-step method)
```

```r
confint(amod_glht)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = elevel ~ alength, data = alpha)
##
## Quantile = 2.3713
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                             Estimate lwr      upr
## intermediate - short == 0  0.43415 -0.47557  1.34388
## long - short == 0          1.18875 -0.04493  2.42243
## long - intermediate == 0   0.75460 -0.33114  1.84033
```

```r
old.par <- par(mai=c(1,2.5,1,1))
plot(amod_glht)
```



```
cld(amod_glht)
```

```
##        short intermediate         long 
##          "a"          "a"          "a" 
```

10) What method of analysis is most appropriate for these data? For this analysis, state the null and alternative hypotheses in words, with context. Then provide a complete conclusion in context. Finally, provide a scope of inference for this study in context.

11) What method of analysis is most appropriate for these data? Describe the validity conditions for this type of analysis. Assess each condition using the information about the study, numerical summaries, and the provided plots. Should a parametric, nonparametric, or neither approach be used here? Explain your choice. Finally, provide a scope of inference for this study in context.

12) What method of analysis is most appropriate for these data? Locate the estimated model components from the fitted model (named `amod`) in the `summary` function output. Discuss the type of model that has been fit here and what each of the values in the Estimate column represent. Interpret the value in the `alengthlong` line of this output in context. Then explain the process of using this output to calculate each group's estimated mean and report your numerical answers. Finally, provide a scope of inference for this study in context.

13) What method of analysis is most appropriate for these data? Locate the output for Tukey's HSD (numeric intervals, plotted intervals, and cld). Examining the intervals first, are any pairs detectably different from each other? If so, which? Explain how you determined this using the intervals and/or the plot. Then, summarize what the compact letter display tells us about these three groups. Interpret the 95% confidence interval from Tukey's HSD associated with the comparison of the "long" and "short" groups in context. Finally, provide a scope of inference for this study in context.

Alzheimer's disease and smoking

In a study of Alzheimer's disease and smoking (Salib and Hillier, 1997), a group of 198 female and male Alzheimer's patients were compared to patients who had other dementias or other diagnoses but no dementia. Each subject was selected from all available patients of each type at a geriatric unit that served an elderly population and then their gender and smoking status were measured, with `smoking` having levels of "None", "<10", "10-20" and ">20" (cigarettes per day). The researchers are interested in whether cigarette smoking differs based on disease status.

Specifically, the authors' state that the following about the subjects:

> *All patients [Alzheimer's] who had been referred to the psychogeriatrician, seen either at their homes, in residential settings, in other hospitals or at the Psychogeriatric Unit in Warrington and "newly" diagnosed as Alzheimer's disease during the duration of the study were included... The control groups [other dementias, other diagnoses but not dementias] included all patients referred to and seen by a psychogeriatrician either in hospital or in the community during the same period of the study with a diagnosis other than Alzheimer's disease.*

- Note: Original paper is behind a paywall (but available through the library) at https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1099-1166(199703)12:3%3C295::AID-GPS476%3E3.0.CO;2-3. The paper is also available upon request from your instructor, but should not be needed.
- Salib, E. and Hillier, V. (1997) A CASE-CONTROL STUDY OF SMOKING AND ALZHEIMER'S DISEASE. Int. J. Geriat. Psychiatry, 12: 295-300.

```r
#output for disease statue and smoking habits
data("alzheimer", package = "coin")
summary(alzheimer)
```

```
##   smoking                disease       gender
##  None :309   Alzheimer      :198   Female:338
##  <10  : 28   Other dementias:164   Male  :200
##  10-20:110   Other diagnoses:176
##  >20  : 91
```

```r
library(mosaic)
table1 <- tally(~smoking+disease, data=alzheimer)
table1
```

```
##        disease
## smoking Alzheimer Other dementias Other diagnoses
##   None        126              79             104
##   <10          15               8               5
##   10-20        30              33              47
##   >20          27              44              20
```

```r
chisq.test(table1)$expected
```

```
##        disease
## smoking Alzheimer Other dementias Other diagnoses
##   None  113.72119       94.193309      101.085502
##   <10    10.30483        8.535316        9.159851
##   10-20  40.48327       33.531599       35.985130
##   >20    33.49071       27.739777       29.769517
```

```r
chisq.test(table1)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  table1
## X-squared = 28.012, df = 6, p-value = 9.346e-05
```

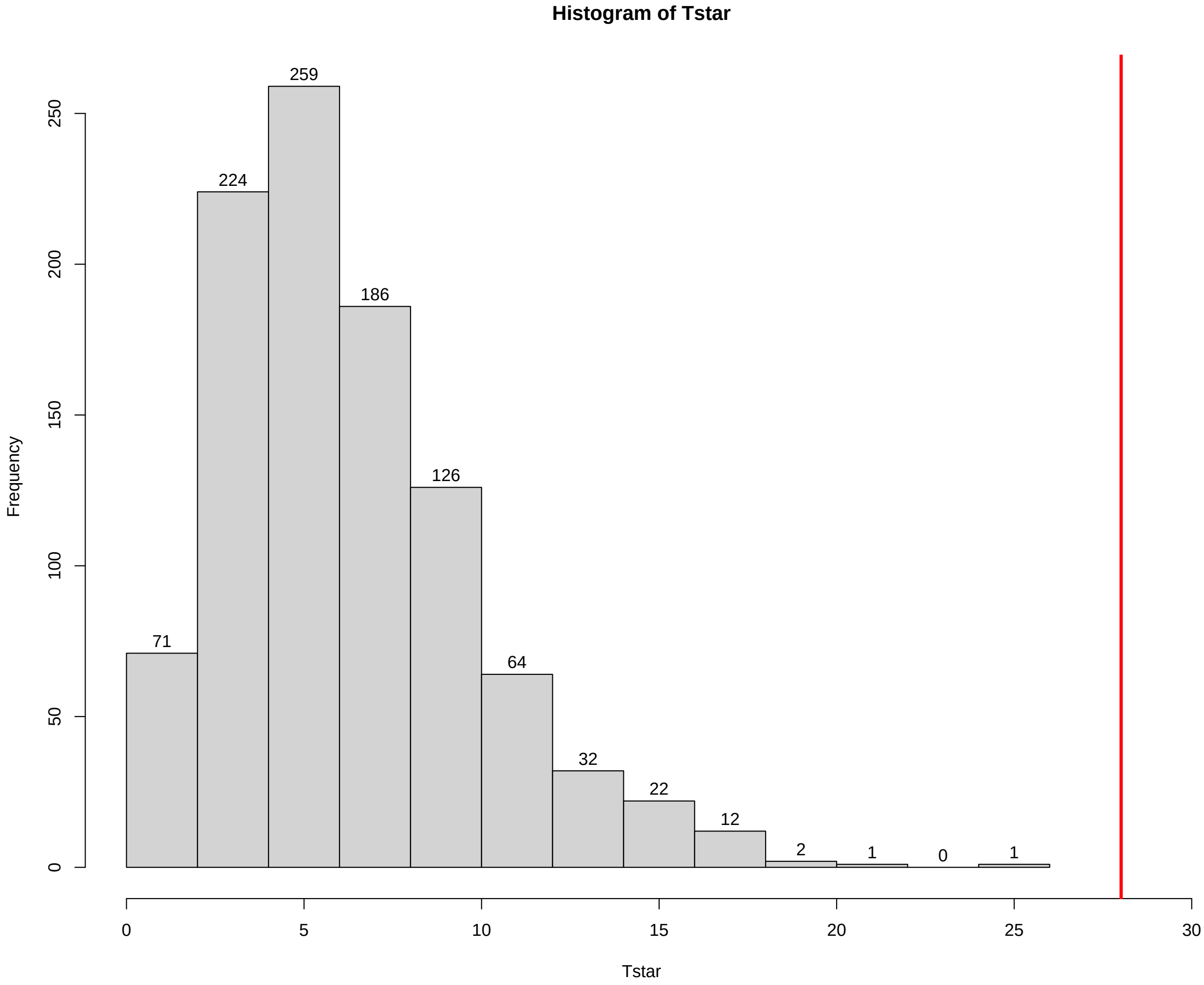
```
Tobs <- chisq.test(table1)$statistic; Tobs
```

```
## X-squared
##  28.01243
```

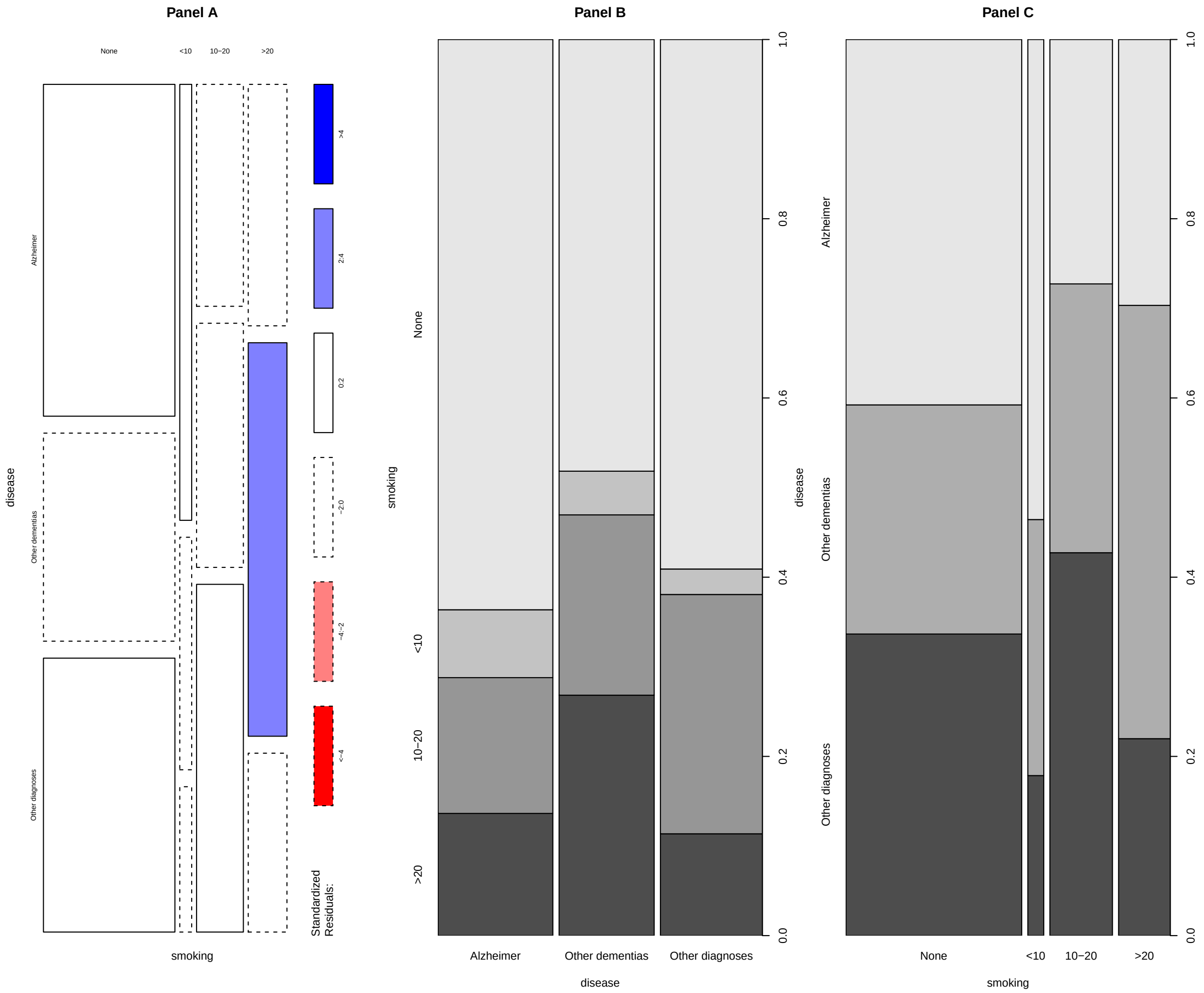
```
B <- 1000
Tstar <- matrix(NA, nrow=B)
for (b in (1:B)){
  Tstar[b] <- chisq.test(tally(~shuffle(smoking)+disease,
                               data=alzheimer))$statistic
}
pdata(Tstar, Tobs, lower.tail=F)[[1]]
```

```
## [1] 0
```

```
hist(Tstar, xlim=c(0,Tobs+1), labels=T)
abline(v=Tobs, col="red",lwd=3)
```



```
par(mfrow=c(1,3))
mosaicplot(table1,shade=T, main="Panel A")
plot(smoking~disease, data=alzheimer, main="Panel B")
plot(disease~smoking, data=alzheimer, main="Panel C")
```



14) For the smoking and disease variables, what type of test is being performed? State the null hypothesis for this test in context. Given this, which of the three provided plots is the most appropriate way to display these data (Panel A, B, or C)? Explain your choice. Based on that plot, what can you say about the research question? Don't use hypothesis test results in your discussion, instead describe features of the chosen plot. Finally, provide a scope of inference for these results for smoking and disease in context.

15) For the smoking and disease variables, what type of test is being performed? Assess the assumptions for this type of test, referencing information about the study and the provided tables. Based on your assessment, should a parametric, nonparametric, or neither approach be used here? Explain your choice. Given that, provide a complete conclusion in context from the provided output. Finally, provide a scope of inference for these results for smoking and disease in context.

16) For the smoking and disease variables, what type of test is being performed? Provide a complete conclusion for this test in context (you may reference either the parametric or nonparametric results, do not worry about assumptions). Calculate and then interpret a standardized residual that contributed to this conclusion in context. Finally, provide a scope of inference for these results for smoking and disease in context.

The researchers were also interested in assessing smoking and gender. The following results consider those two variables.

```
#output for smoking habits and gender
data("alzheimer", package = "coin")
summary(alzheimer)
```

```
##    smoking               disease       gender   
##  None :309   Alzheimer      :198   Female:338  
##  <10  : 28   Other dementias:164   Male  :200  
##  10-20:110   Other diagnoses:176               
##  >20  : 91                                     
```

```
table2 <- tally(~smoking+gender, data=alzheimer)
table2
```

```
##        gender
## smoking Female Male
##   None     226   83
##   <10       17   11
##   10-20     56   54
##   >20       39   52
```

```
chisq.test(table2)$expected
```

```
##        gender
## smoking    Female      Male
##   None  194.13011 114.86989
##   <10    17.59108  10.40892
##   10-20  69.10781  40.89219
##   >20    57.17100  33.82900
```

```
chisq.test(table2)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  table2
## X-squared = 36.351, df = 3, p-value = 6.311e-08
```

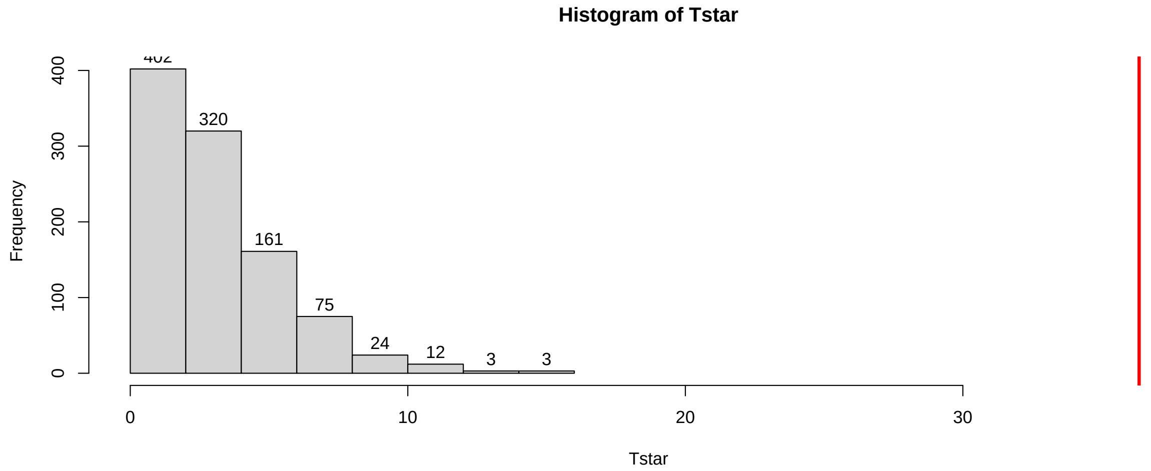
```
Tobs <- chisq.test(table2)$statistic; Tobs
```

```
## X-squared 
##  36.35117
```

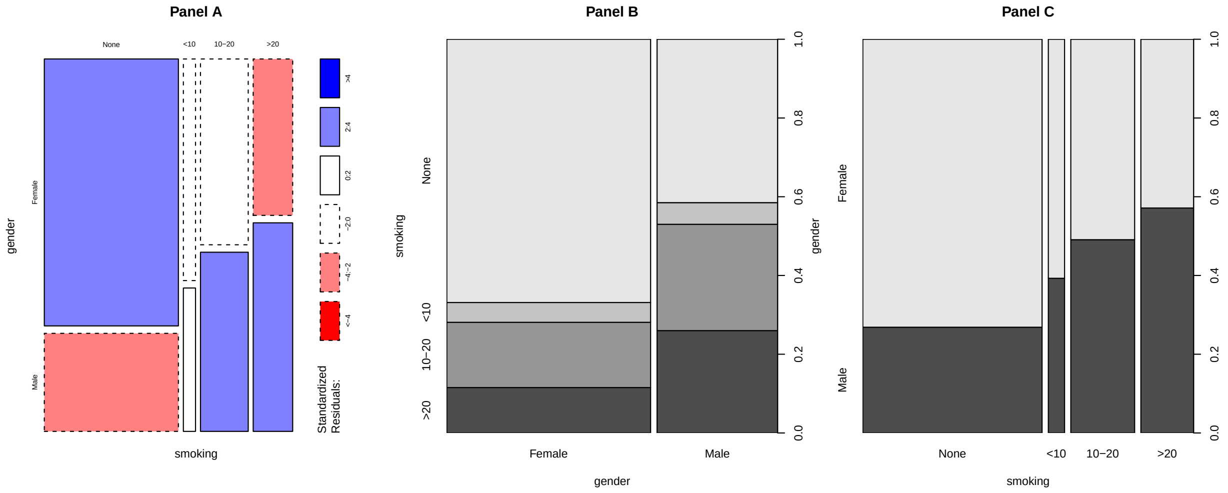
```
B <- 1000
Tstar <- matrix(NA, nrow=B)
for (b in (1:B)){
  Tstar[b] <- chisq.test(tally(~shuffle(smoking)+gender,
                               data=alzheimer))$statistic
}
pdata(Tstar, Tobs, lower.tail=F)[[1]]
```

```
## [1] 0
```

```
hist(Tstar, xlim=c(0,Tobs+1), labels=T)
abline(v=Tobs, col="red",lwd=3)
```



```r
par(mfrow=c(1,3))
mosaicplot(table2,shade=T, main="Panel A")
plot(smoking~gender, data=alzheimer, main="Panel B")
plot(gender~smoking, data=alzheimer, main="Panel C")
```



17) For the `gender` and `smoking` variables, what type of test is being performed? State the null hypothesis for this test in context. Given this, which of the three provided plots is the most appropriate way to display these data (Panel A, B, or C)? Explain your choice. Based on that plot, what can you say about the research question? Don't use hypothesis test results in your discussion, instead describe features of the chosen plot. Finally, provide a scope of inference for these results for `gender` and `smoking` in context.

18) For the `gender` and `smoking` variables, what type of test is being performed? Assess the assumptions for this type of test, referencing information about the study and the provided tables. Based on your assessment, should a parametric, nonparametric, or neither approach be used here? Explain your choice. Given that, provide a complete conclusion in context from the provided output. Finally, provide a scope of inference for these results for `gender` and `smoking` in context.

19) For the `gender` and `smoking` variables, what type of test is being performed? Provide a complete conclusion for this test in context (you may reference either the parametric or nonparametric results, do not worry about assumptions). Calculate and then interpret a standardized residual that contributed

to this conclusion in context. Finally, provide a scope of inference for these results for `gender` and `smoking` in context.

Iron concentrations in milk

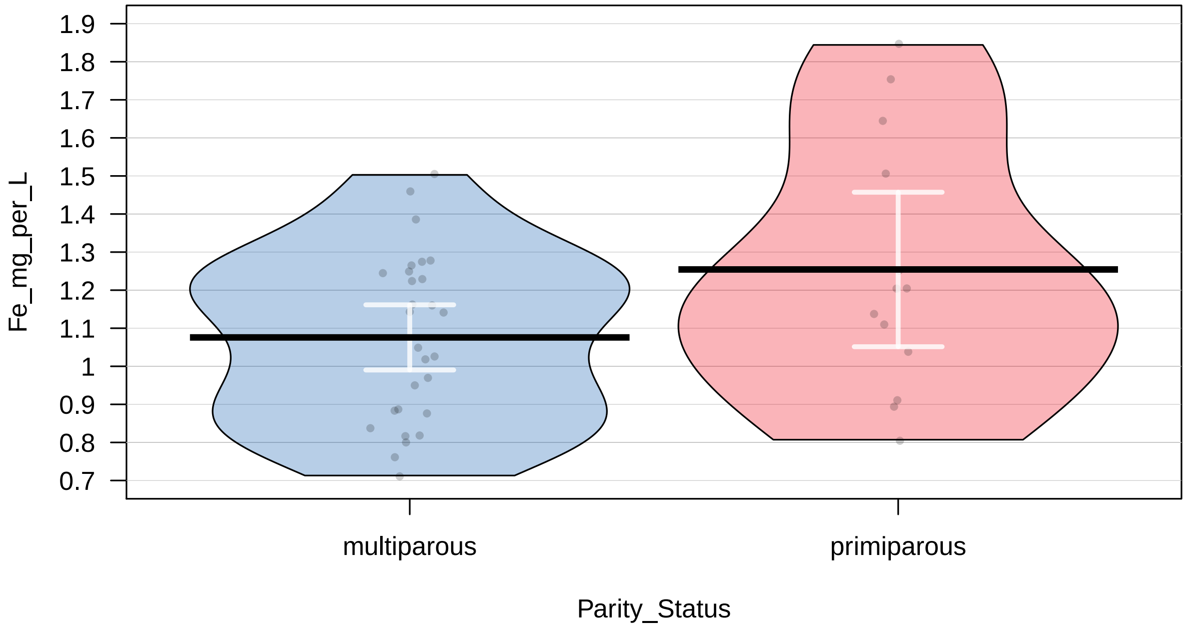
Researchers collected human milk samples from four different countries and published their results in Klein et al. (2017). Among many other research questions, the researchers are interested in comparing concentrations of trace elements among the countries studied.

They note that the "Samples for this project were collected as part of a larger investigation of the composition of human milk across diverse populations." Also from their paper, they state that "Women living in the Boston area provided human milk samples from June to August 2013 and represent an urban W.E.I.R.D. (Westernized, educated, industrial, rich, democratic) population. . . . Argentinean samples were collected from indigenous Qom (formerly Toba) women in northeastern Argentina from September 2012 to March 2013. Traditionally, the Qom people were hunter-gathers, but today many have migrated to poor peri-urban barrios where they have access to free governmental healthcare, but often share outdoor water taps and lack indoor toilets." There is no indication of random sampling in either of these instances for obtaining lactating mothers to participate in the studies, so you can assume that these were subjects who volunteered.

- Klein LD, Breakey AA, Scelza B, Valeggia C, Jasienska G, Hinde K (2017) Concentrations of trace elements in human milk: Comparisons among women in Argentina, Namibia, Poland, and the United States. PLoS ONE 12(8): e0183367. https://doi.org/10.1371/journal.pone.0183367 (Available at https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0183367)

We will be focused on a subset of observations from the paper just from the USA and Argentina and just on the iron concentrations in the milk, measured in mg/L. For `Parity_Status`, "primiparous" means having given birth to just one child and "multiparous" usually means giving birth two, three, or four times (or probably just more than one here).

```r
#output for iron concentrations and parity status
library(readr)
milkconcentrations <- read_csv("milkconcentrations.csv")
milkconcentrations$Fe_mg_per_L <- milkconcentrations$Fe_ug_per_L/1000

milkc <- subset(milkconcentrations, Population %in% c("USA", "Argentina"))

library(yarrr)
pirateplot(Fe_mg_per_L ~ Parity_Status, data=milkc, inf.method="ci", inf.disp="line")
```



```r
favstats(Fe_mg_per_L ~ Parity_Status, data=milkc)
```

```
##   Parity_Status       min        Q1   median       Q3      max     mean
## 1   multiparous 0.7108984 0.8816581 1.094996 1.245820 1.505108 1.075857
## 2   primiparous 0.8043542 1.0381718 1.204151 1.506118 1.847060 1.254352
##          sd  n missing
## 1 0.2207452 28       0
## 2 0.3358250 13       0
```

```r
lm_2 <- lm(Fe_mg_per_L ~ Parity_Status, data=milkc)
summary(lm_2)
```

```
## 
## Call:
## lm(formula = Fe_mg_per_L ~ Parity_Status, data = milkc)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.45000 -0.19967 -0.04985  0.17322  0.59271 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               1.07586    0.04944  21.762   <2e-16
## Parity_Statusprimiparous  0.17850    0.08780   2.033   0.0489
## 
## Residual standard error: 0.2616 on 39 degrees of freedom
## Multiple R-squared:  0.09582,    Adjusted R-squared:  0.07264 
## F-statistic: 4.133 on 1 and 39 DF,  p-value: 0.0489
```

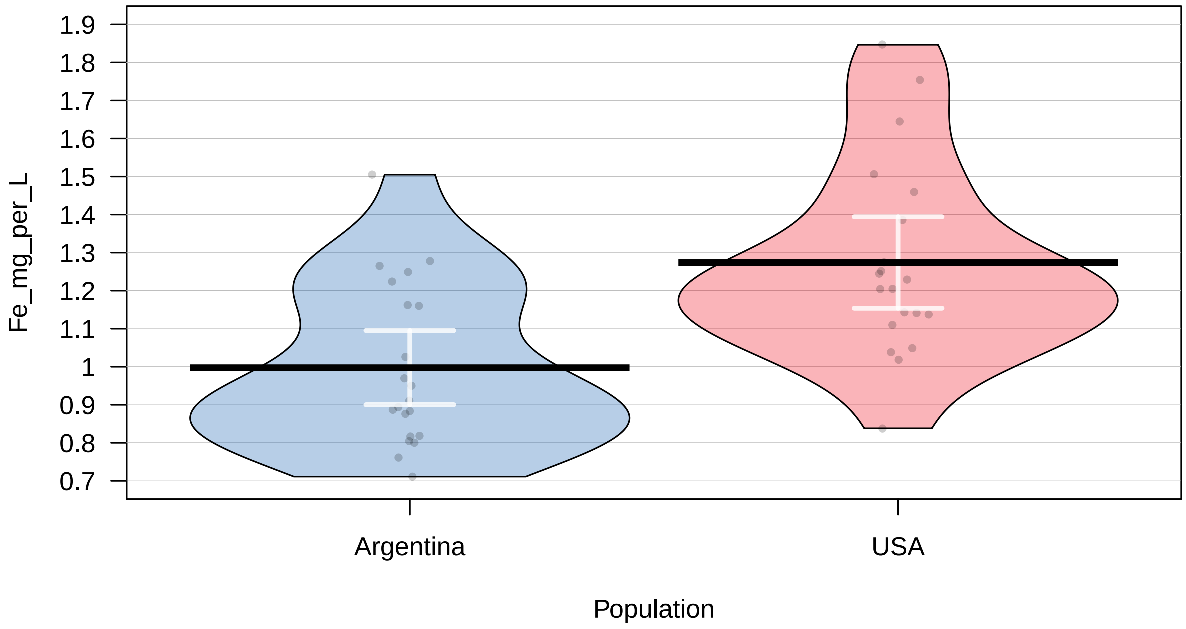
20) For the study on iron concentrations and `Parity_Status`, what method of analysis is most appropriate

for these data? For this analysis, state the null and alternative hypotheses in words, with context. Then provide a complete conclusion in context. Finally, provide a scope of inference for this study in context.

21) For the study on iron concentrations and `Parity_Status`, what method of analysis is most appropriate for these data? Describe the validity conditions for this type of analysis. Assess each condition using the information about the study, numerical summaries, and the provided plots. Should a parametric, nonparametric, or neither approach be used here? Explain your choice. Finally, provide a scope of inference for this study in context.

22) For the study on iron concentrations and `Parity_Status`, what method of analysis is most appropriate for these data? Locate the estimated model components from the fitted model (named `lm_2`) in the `summary` function output. Discuss what each of the values in the Estimate column represent. Then explain the process of using this output to calculate each group's estimated mean and report your numerical answers. Finally, provide a scope of inference for this study in context.

```
pirateplot(Fe_mg_per_L ~ Population, data=milkc, inf.method="ci", inf.disp="line")
```



```
favstats(Fe_mg_per_L ~ Population, data=milkc)
```

```
##   Population       min        Q1    median       Q3      max      mean
## 1  Argentina 0.7108984 0.8182358 0.9108702 1.162134 1.505108 0.9976656
## 2        USA 0.8374397 1.1305675 1.2167896 1.404179 1.847060 1.2739803
##          sd  n missing
## 1 0.2141309 21       0
## 2 0.2567723 20       0
```

```
lm_3 <- lm(Fe_mg_per_L ~ Population, data=milkc)
summary(lm_3)
```

```
##
## Call:
## lm(formula = Fe_mg_per_L ~ Population, data = milkc)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43654 -0.16411 -0.06948  0.16447  0.57308
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.99767    0.05147  19.383  < 2e-16
## PopulationUSA  0.27631    0.07370   3.749 0.000575
##
## Residual standard error: 0.2359 on 39 degrees of freedom
## Multiple R-squared:  0.265,  Adjusted R-squared:  0.2461
## F-statistic: 14.06 on 1 and 39 DF,  p-value: 0.0005747
```

23) For the study on iron concentrations and `Population`, what method of analysis is most appropriate for these data? For this analysis, state the null and alternative hypotheses in words, with context. Then provide a complete conclusion in context. Finally, provide a scope of inference for this study in context.

24) For the study on iron concentrations and `Population`, what method of analysis is most appropriate for these data? Describe the validity conditions for this type of analysis. Assess each condition using the information about the study, numerical summaries, and the provided plots. Should a parametric, nonparametric, or neither approach be used here? Explain your choice. Finally, provide a scope of inference for this study in context.

25) For the study on iron concentrations and `Population`, what method of analysis is most appropriate for these data? Locate the estimated model components from the fitted model (named `lm_3`) in the `summary` function output. Discuss what each of the values in the Estimate column represent. Then explain the process of using this output to calculate each group's estimated mean and report your numerical answers. Finally, provide a scope of inference for this study in context.

Mercury from fish

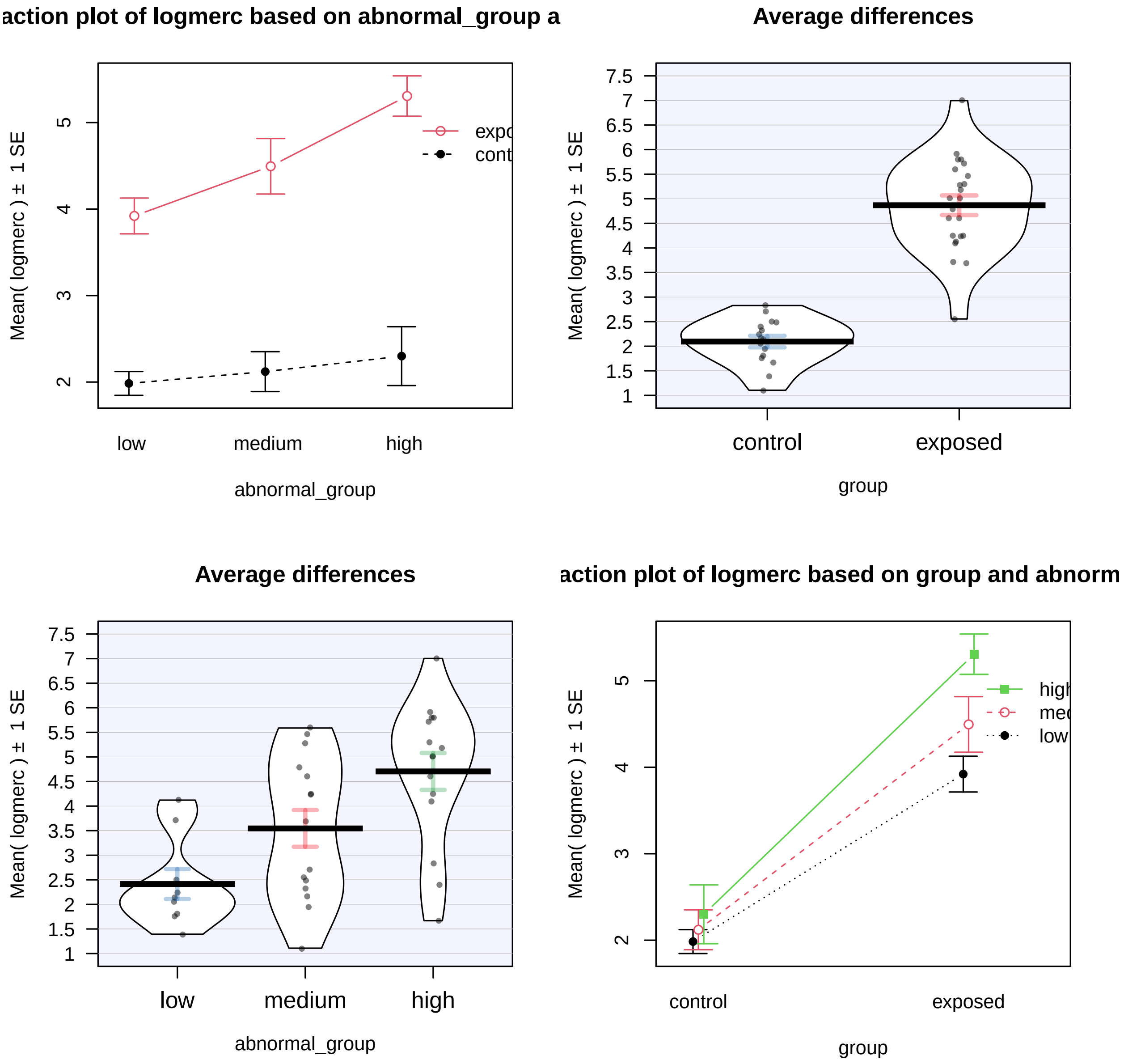
Skerfving et al. (1974) studied two different groups of subjects in Sweden. They state that "The exposed group consisted of 23 subjects (5 females and 18 males). They lived in different areas in Sweden... Nine persons were fishermen, five fishermen's wives, six workmen, two farmers, and one clerk. Sixteen control subjects (4 females and 11 males) from the Stockholm metropolitan area were also studied. In this group, five subjects were clerks, four craftsmen, three porters, three workmen, and one a glass washer. The age distribution in the control group was similar to that in the exposed group, but the mean age of the controls was somewhat higher than that of the exposed persons. Samples from both groups were obtained during 1968-1972 (all seasons)... All subjects in the exposed group had had more than three meals a week of contaminated fish (0.5-7 mg mercury as methylmercury/kg fish) for more than 3 years... None of the control subjects had a history indicating regular consumption of contaminated fish. They all had eaten fish caught at sea (0.05 mg mercury/kg fish or less) once a week or less."

- Skerfving, S. Hansson, K., Mangs, C., Lindsten, J., and Ryman, N. (1974) Methylmercury-induced chromosome damage in man, *Environmental Research*,7(1) 83-98. https://www.sciencedirect.com/science/article/abs/pii/0013935174900784

We are interested in whether the relationship between exposure/not to contaminated fish changes the relationship between abnormal cells and the concentration of mercury in the blood (ng/g), which was log-transformed for the following analyses. Abnormal cells (`abnormal`) was a measurement of the percentage of the cells with structural abnormalities and was grouped in low, medium, or high levels.

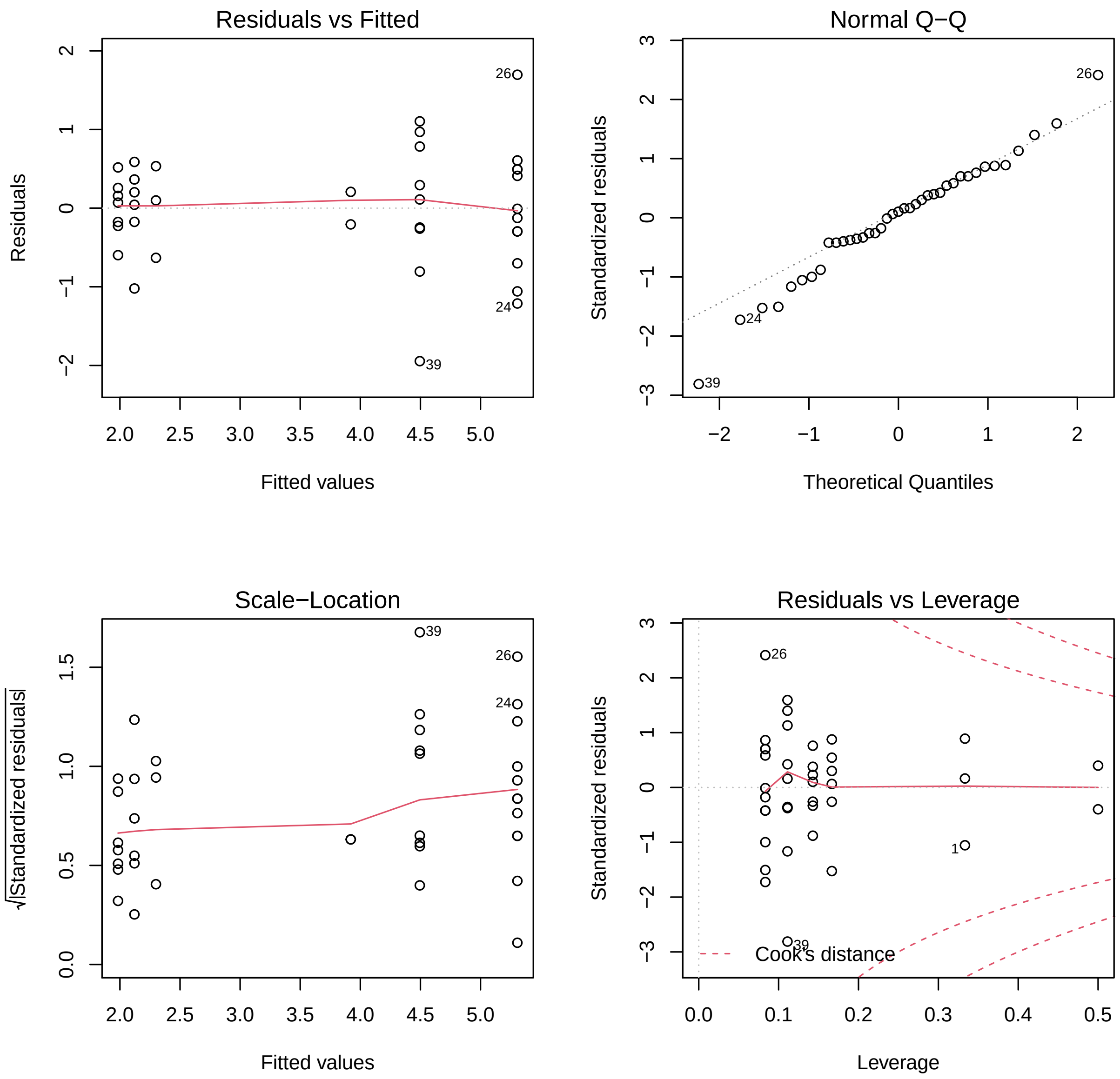
```r
source("http://www.math.montana.edu/courses/s217/documents/intplotfunctions_v3.R")
library(coin)
data(mercuryfish)
```

```
library(car)

mercuryfish$abnormal_group <- factor(cut(mercuryfish$abnormal, breaks=c(0,3,8,22),
                                         include.lowest=T))

levels(mercuryfish$abnormal_group) <- c("low", "medium", "high")

mercuryfish$logmerc <- log(mercuryfish$mercury)

intplotarray(logmerc~group*abnormal_group, data=mercuryfish)
```



```
lm_fish <-lm(logmerc~group*abnormal_group, data=mercuryfish)
```

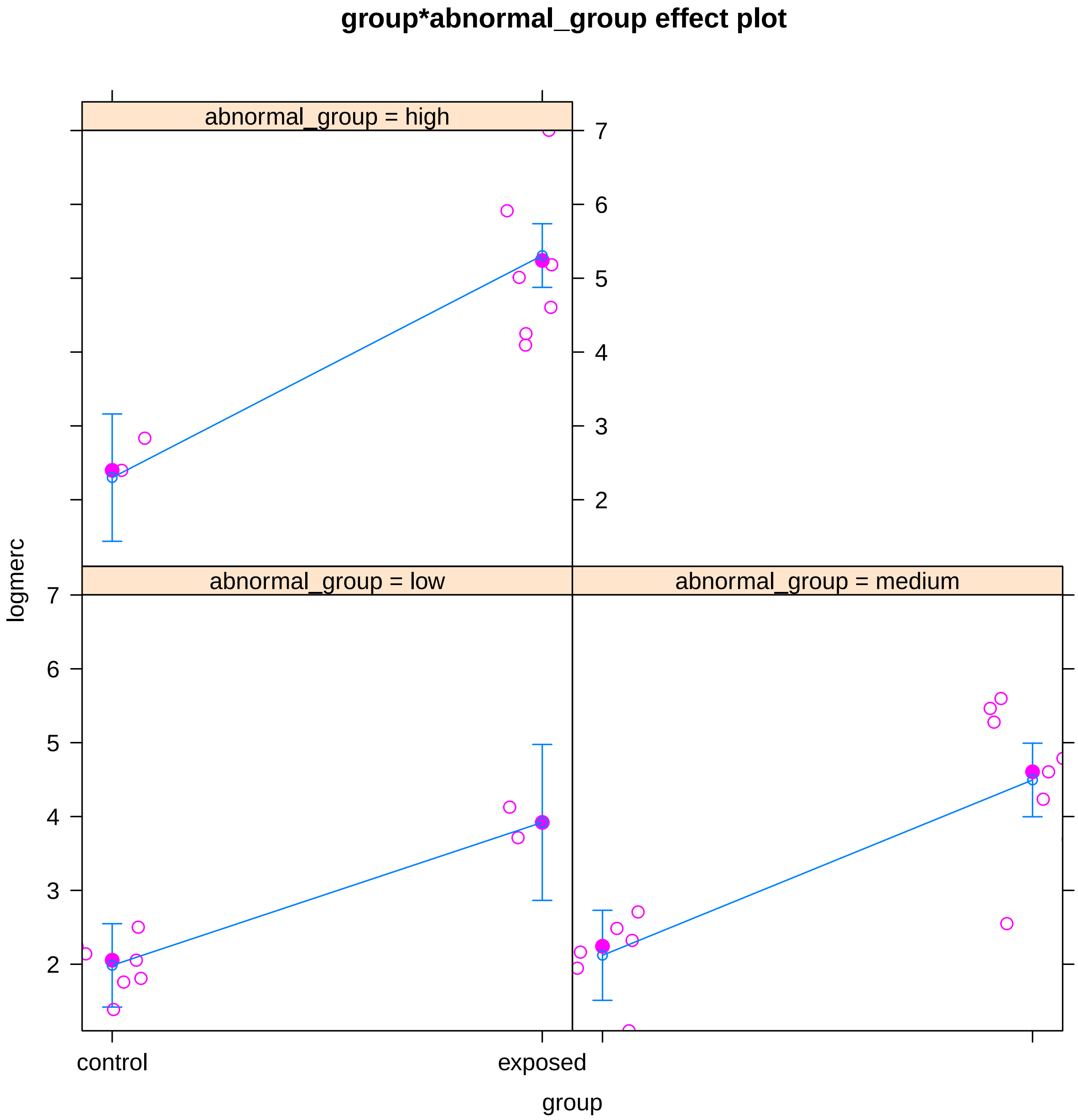
```
par(mfrow=c(2,2))
plot(lm_fish)
```



```
Anova(lm_fish)
```

```
## Anova Table (Type II tests)
##
## Response: logmerc
##                       Sum Sq Df F value    Pr(>F)
## group                 46.663  1 86.6101 9.488e-11
## abnormal_group         4.408  2  4.0906   0.02588
## group:abnormal_group   1.167  2  1.0829   0.35035
## Residuals             17.780 33
```

```r
plot(allEffects(lm_fish, resid=T))
```



26) Describe the method of analysis that is most appropriate for these data and describe the type of model that has been stored as `lm_fish`. State the null hypothesis for this model that you should test in context. Then discuss what the provided interaction plots suggest about these hypotheses. Describe the features of this plot in your answer. Finally, provide a scope of inference for this study in context.

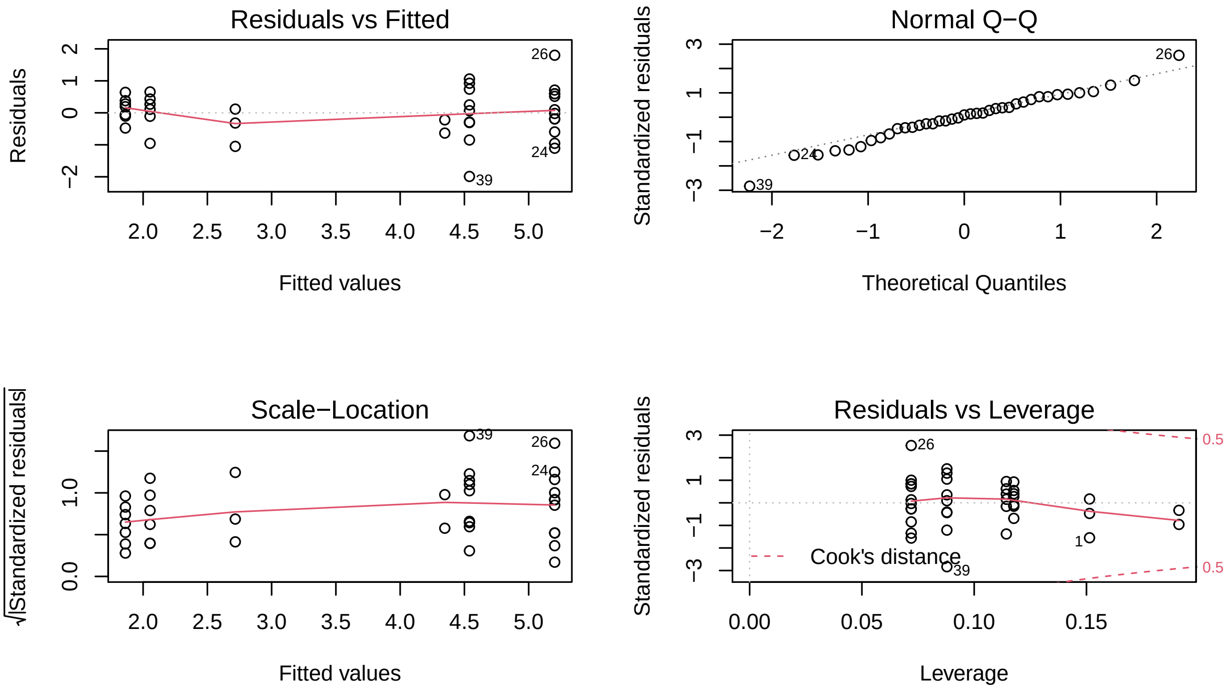
27) Describe the method of analysis that is most appropriate for these data and describe the type of model that has been stored as `lm_fish`. State the null hypothesis for this model that you should test in context. Then provide a complete conclusion in context. Finally, provide a scope of inference for this study in context.

28) Describe the method of analysis that is most appropriate for these data and describe the type of model that has been stored as `lm_fish`. Describe the validity conditions for this type of analysis. Assess each condition using the information about the study, numerical summaries, and the provided plots. Is it

reasonable to report parametric test results, given your assessment of these conditions? Clearly state "yes" or "no" and explain your reasoning. Finally, provide a scope of inference for this study in context.

We are also interested in exploring a different model as shown below in `lm_fish2`.

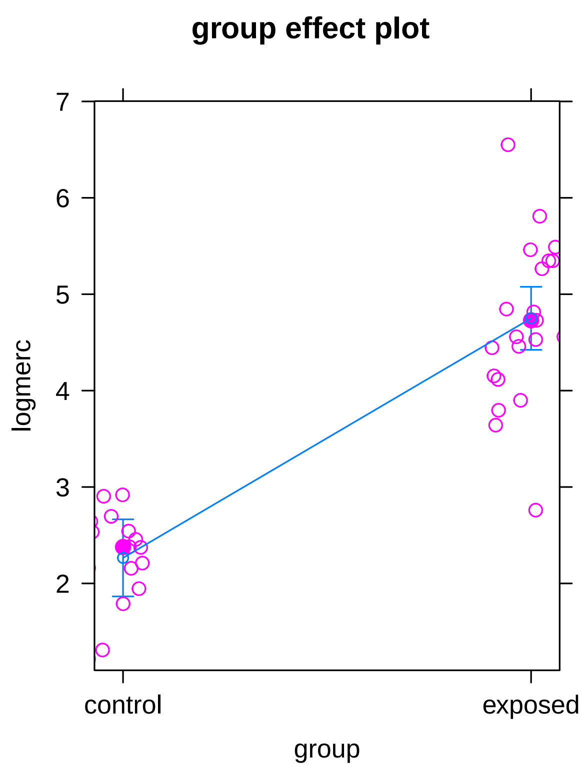
```
lm_fish2 <-lm(logmerc~group + abnormal_group, data=mercuryfish)
par(mfrow=c(2,2))
plot(lm_fish2)
```

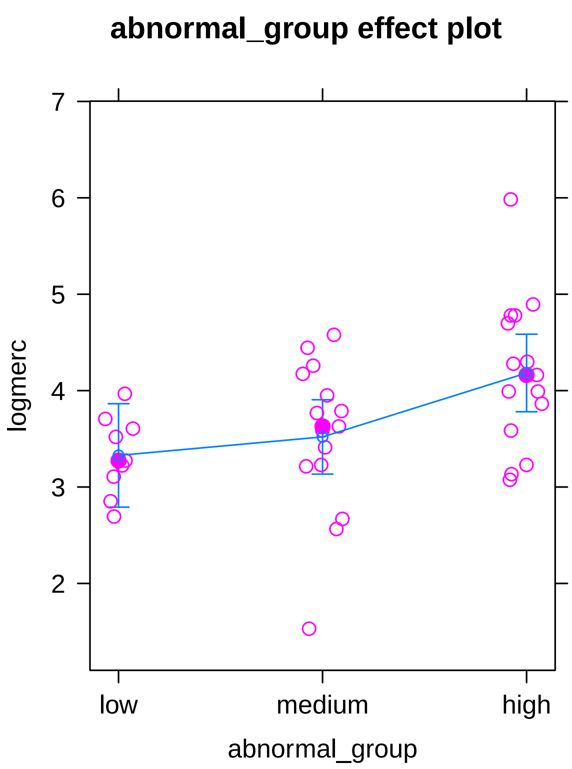


```
Anova(lm_fish2)
```

```
## Anova Table (Type II tests)
##
## Response: logmerc
##                Sum Sq Df F value    Pr(>F)
## group          46.663  1 86.2018 5.701e-11
## abnormal_group  4.408  2  4.0714   0.02572
## Residuals      18.946 35
```

```
plot(allEffects(lm_fish2, resid=T))
```





29) Describe the method of analysis that is most appropriate for these data and describe the type of model that has been stored as `lm_fish2`. State the null and alternative hypotheses associated with "group" for this model in context. Then provide a complete conclusion in context. Finally, provide a scope of inference for this study in context.

30) Describe the method of analysis that is most appropriate for these data and describe the type of model that has been stored as `lm_fish2`. Describe the validity conditions for this type of analysis. Assess each condition using the information about the study, numerical summaries, and the provided plots. Is it reasonable to report parametric test results given your assessment of these conditions? Clearly state "yes" or "no" and explain your reasoning. Finally, provide a scope of inference for this study in context.

Part III (5 points): Collaboration

- Discuss any support, discussions, or other feedback you used in preparing your answer. Failure to address this in your recording will result in losing all 5 points. You are allowed to discuss and get feedback on your answers but you must be the speaker in the video and your answers must be your own words.

Part IV (5 points): Time limit

- You must stay within the allotted 5 minutes for your recording. Any exceedance of that time may result in a loss of points.