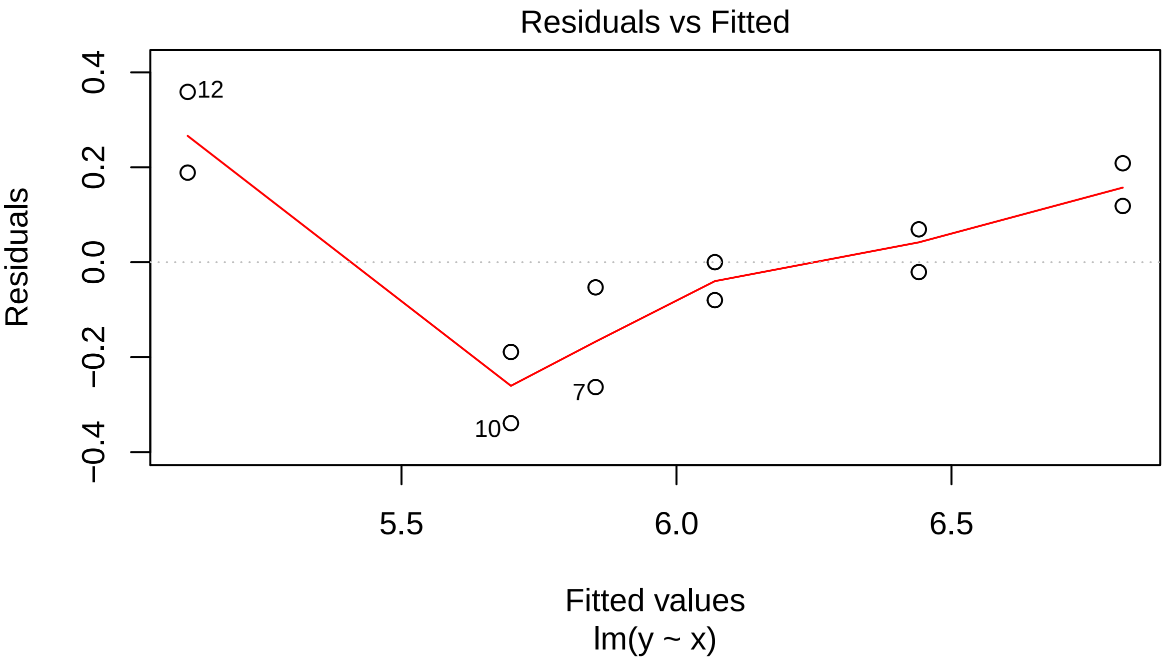
Hw 9 S139

Callin Switzer

November 9, 2014

#1



(a) As the residual plot shows, a simple linear model doesn't seem to fit the data very well. The residuals do not have equal spread.

- Here's the F-test to compare the separate means model to the equal means

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## as.factor(x)  5  3.916  0.7833   79.59 2.11e-05 ***
## Residuals     6  0.059  0.0098                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This output tests the null hypothesis that the best model is the equal means model. The p-value of << 0.05 means we reject the null hypothesis that the all the means are equal.

*Here are results from the F-test that compares equal-means model to the regression model.

```
##   value   numdf   dendf 
## 77.1764  1.0000 10.0000
```

```
## [1] 5.140138e-06
```

This output tests the null hypothesis that the equal means model is the best model (when compared to the regression model). The p-value of << 0.05 means we reject the null hypothesis that the means are all equal or that the slope == 0.

- Here is the lack of fit F-test:

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ as.factor(x)
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     10 0.45602
## 2      6 0.05905  4   0.39697 10.084 0.007841 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

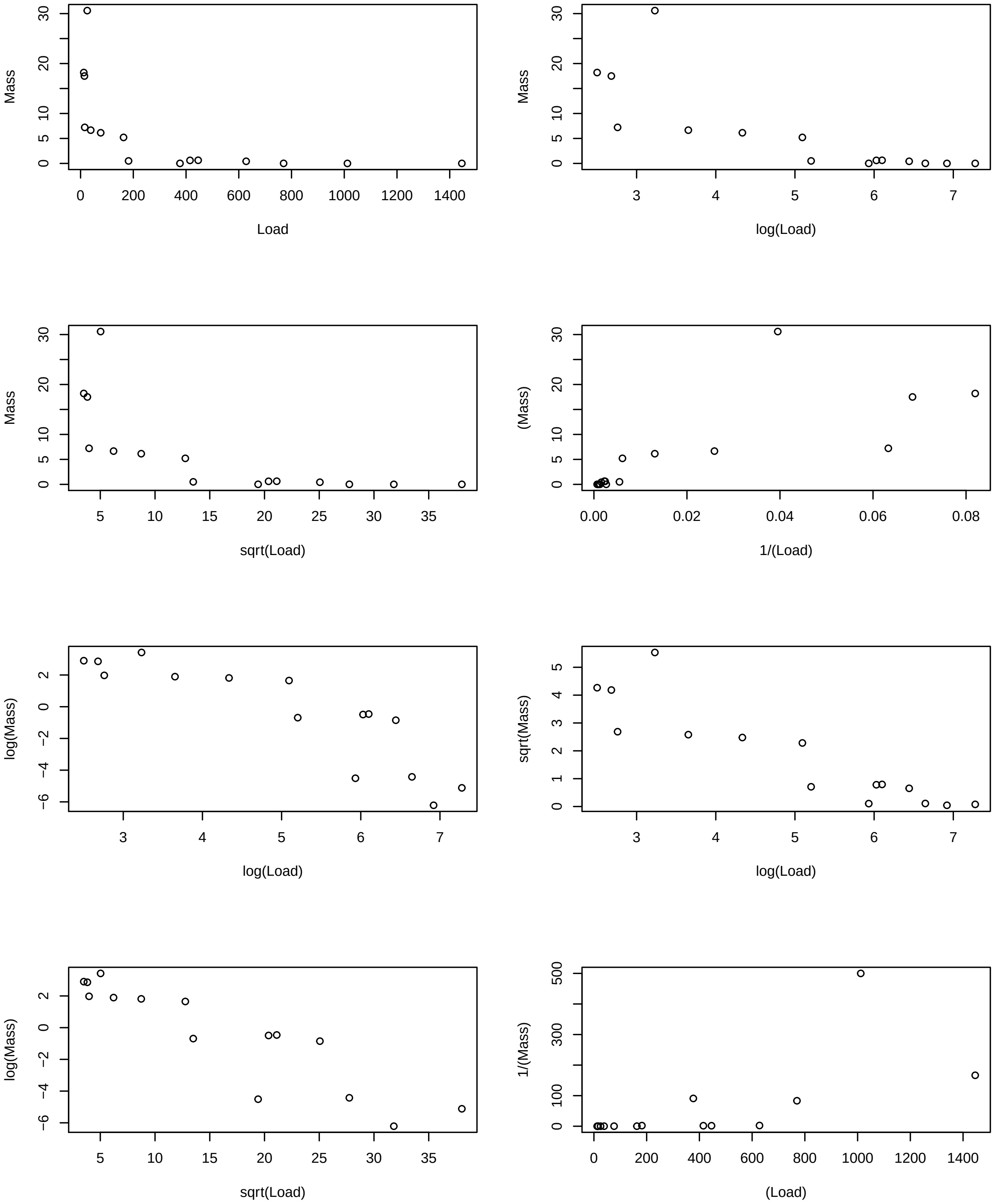
(b) The lack of fit F-test tests the H_0 that the linear regression model fits. The p-value from the lack of fit F-test on the steer data is 0.0078415. We reject the null hypothesis that the linear regression fits.

- In summary, we should probably use a separate means model, if we're using all 12 data points.

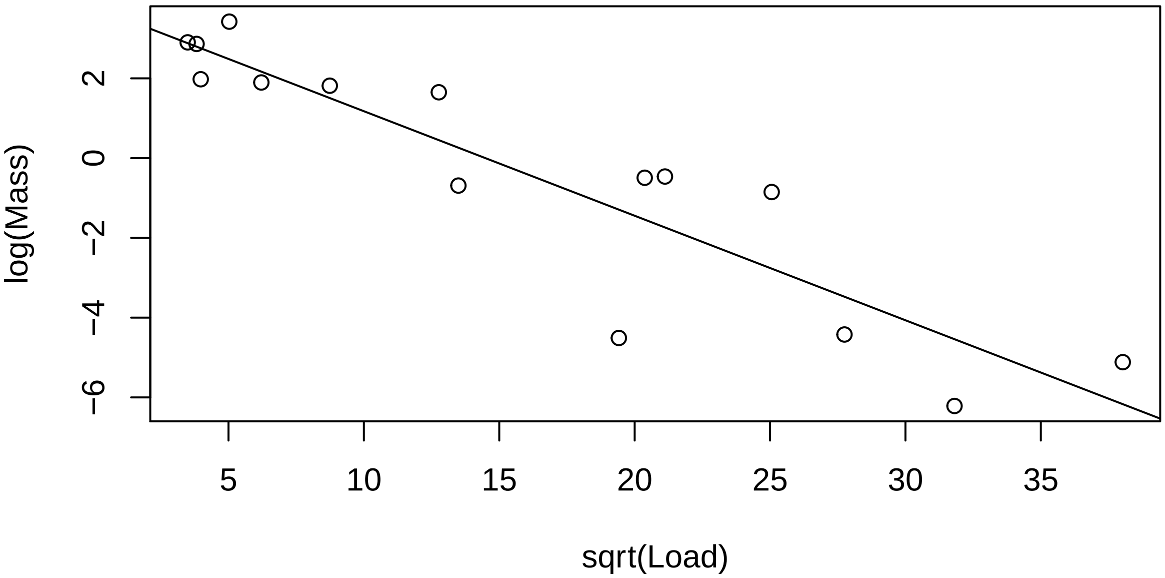
(c) There is justification for dropping the last two carcasses. The question of interest asks how much time is needed for the carcasses to reach a pH of 6. The relationship appears quite linear within the range of interest.

#2

(a) I tried a bunch of transformations (see below)



(b) I decided to go with this transformation: log(Mass)~sqrt(Load)



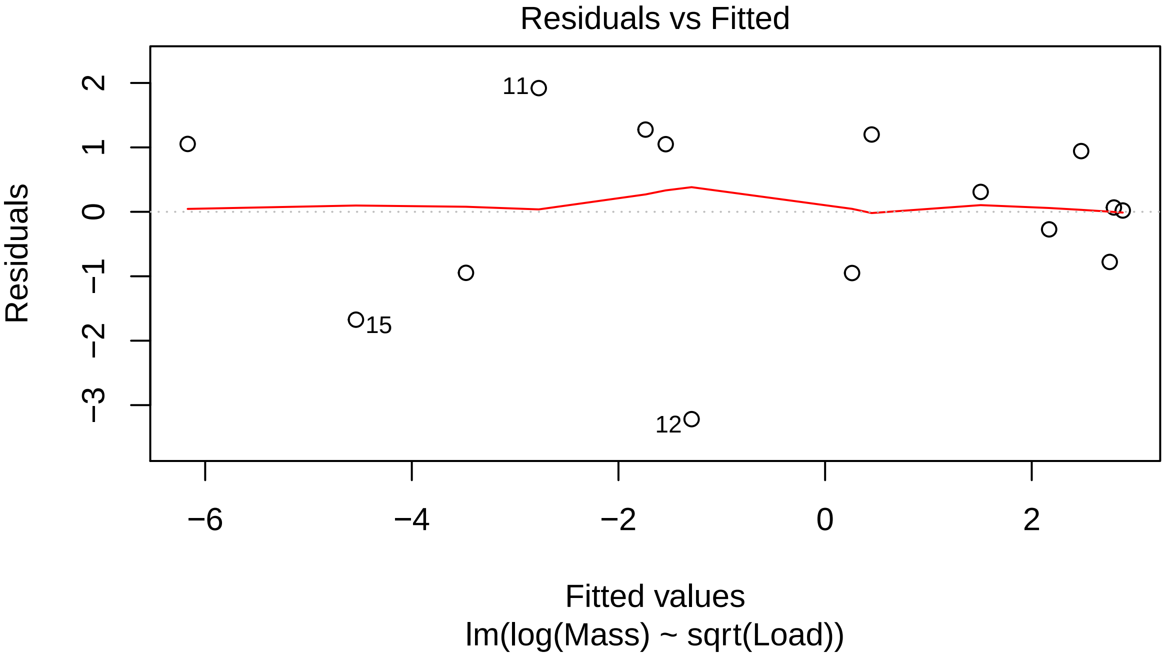
```
## [1] "Here are the residuals"
```

```
##           1           2           3           4           5           6 
##  0.02029732  0.06707645 -0.77792852  0.94270698 -0.27213957  0.30905397 
##           7           8           9          10          11          12 
##  1.20005319 -0.95004354  1.04976907  1.27620770  1.92018167 -3.21773398 
##          13          14          15 
## -0.94697964  1.05337196 -1.67389306
```

```
## [1] "Here are the fitted values"
```

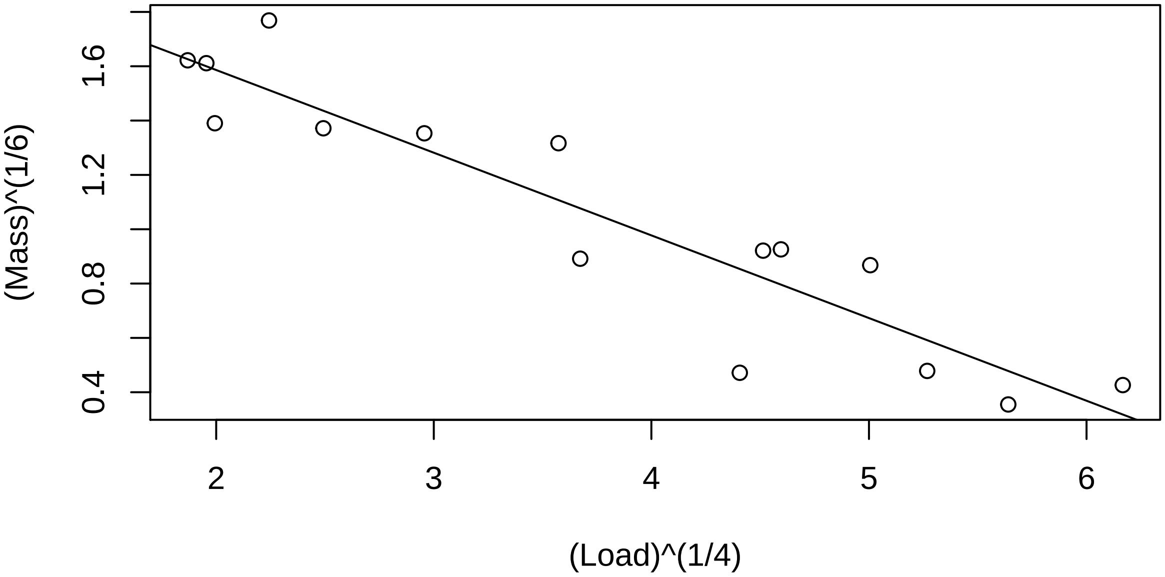
```
##         1         2         3         4         5         6 
##  2.8811243  2.7951244  2.7547835  2.4782930  2.1682591  1.5057708 
##         7         8         9        10        11        12 
##  0.4505267  0.2608884 -1.5424274 -1.7382432 -2.7711529 -1.2921260 
##        13        14        15 
## -3.4758690 -6.1693678 -4.5407150
```

(c) Here is the residual plot:



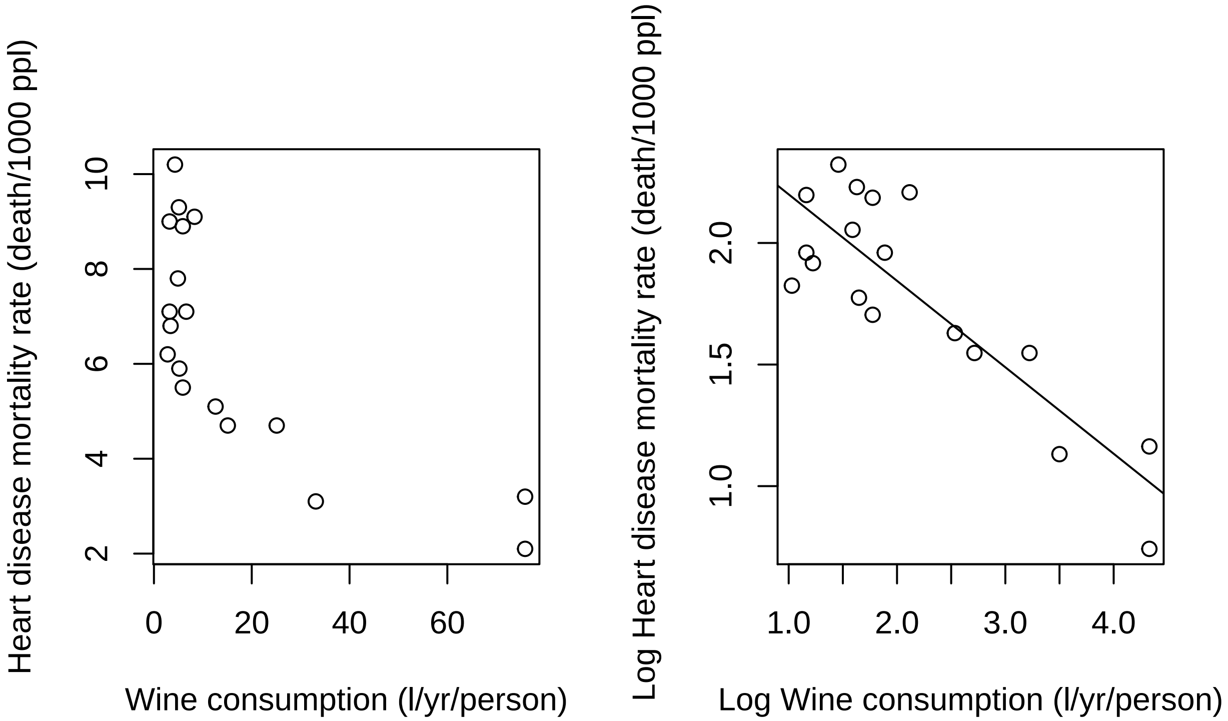
I would like to try some transformations that fit X and Y to exponents between 0 and 1. I played around with a bunch of different exponenets, and found another way to make the relationship approximately linear. I used $(\text{Mass})^{(1/6))\sim(\text{Load})}(1/4)$. I suggest that We use the transformation I proposed in part b, because it is simpler.

Here is a plot, using the transformation mentioned above

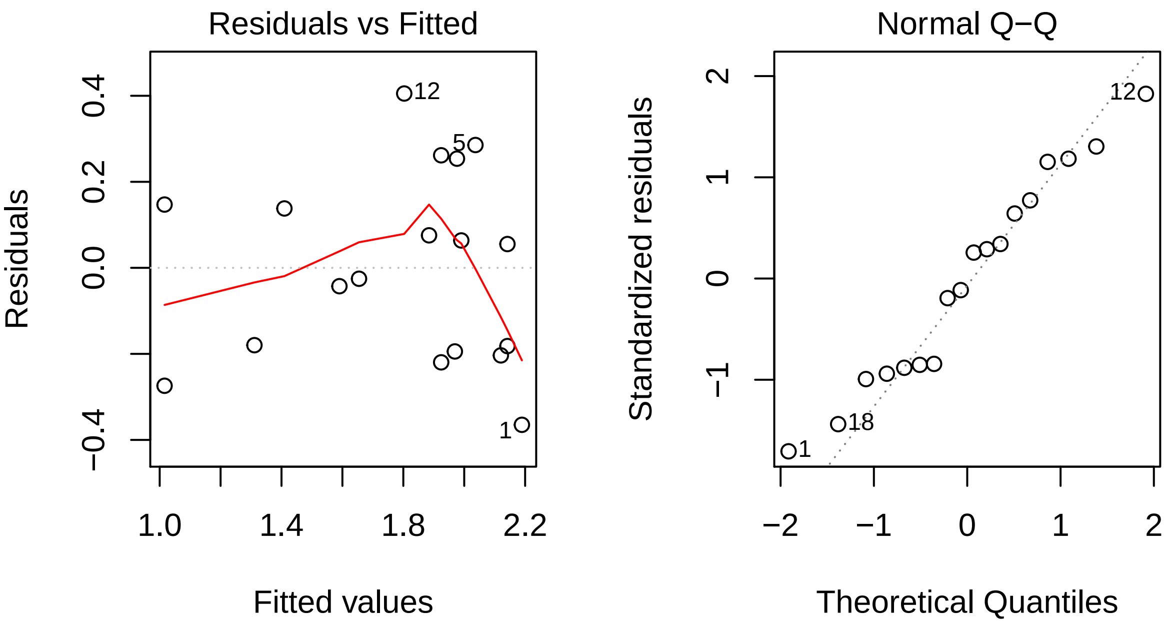


#3

Here is a plot of the raw data and a plot of the transformed data



```
## [1] "Here are some diagnostic plots for the transformed data"
```



- I choose to use linear regression with both X and Y variables log-transformed. The assumptions are as follows:

1. Linearity

2. Constant Variance along the line
3. Normality of each subpopulation of responses (log(y)) at the different values of log(x)
4. Independence:Location in relation to the mean cannot be predicted with knowledge of other responses
5. Random sample

The transformation makes the data linear and makes it so it has fairly constant variance along the line (see residual plot). Also, the transformed data don't look perfectly normal, but I think they're good enough (see Q-Q plot). The data may not be independent; for instance, many countries that grow a lot of grapes drink a lot of wine and are geographically near each other. Also, the data do not come from a random sample.

The R^2 value is 0.738433. The correlation between log(wine) and log(mortality) is -0.8593213. If I square the correlation, I get the R^2, or 0.738433.

After reading the article, "Ecological Inference and the Ecological Fallacy", I don't think it would be right to advise people to increase their wine coinsumption if they want to decrease their risk of death from heart disease. This would be incorrect because of possible confounding in the observational study (maybe wine consumption is correlated with different diets or less stress) and aggregation bias (the response of the individual may not be the same as the response of the aggregate).

Statistical Report

1. Summary of Statistical Findings

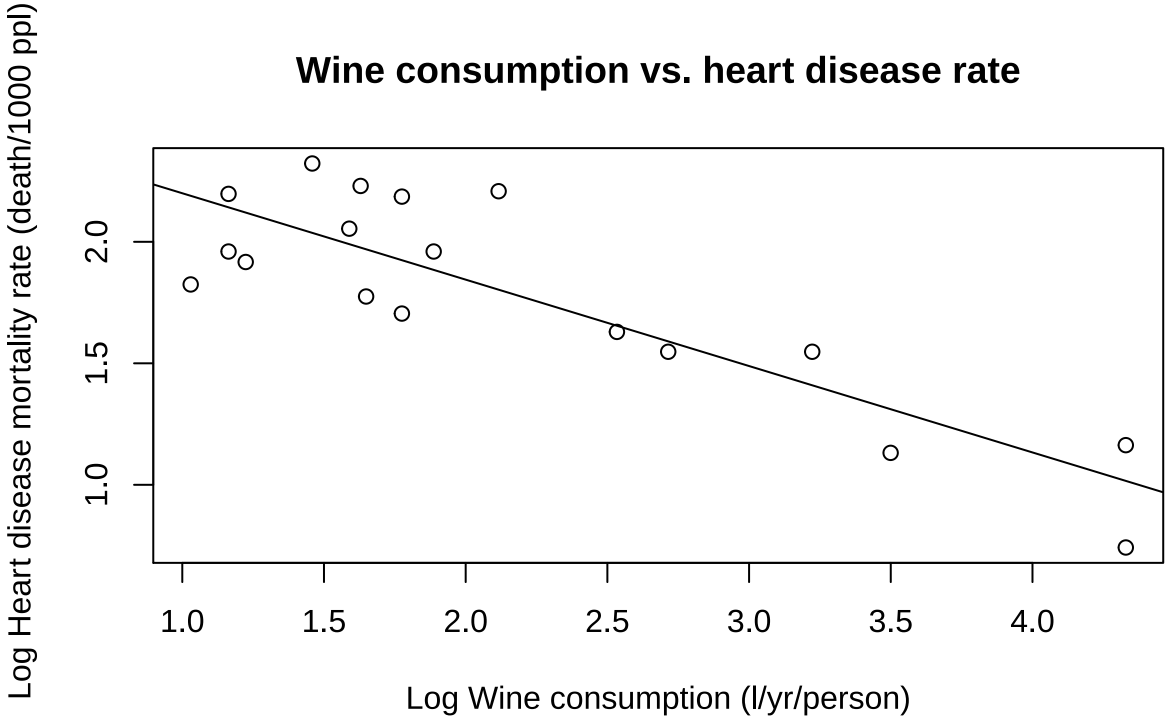
There is a negative association of wine consumption and death rate by heart disease. As shown by the table below (p << 0.05). The figure below shows the relationship graphically. Since the data are on a log-log scale, Median{D|W} = $exp(\beta_0)X^{\beta_1}$. A doubling in wine comsumption is associated with a change of 2^{β_1} in the median of heart disease mortality. β_0 is the intercept, and β_1 is the slope of the line from the table below. The 95% confidence interval for the slope is (-0.467, -0.234). A confidence interval for the multiplicative factor in the median is $2^{-0.467}$ to $2^{-0.234}$ or 0.7234674 to 0.8502742. The mean decrease is about $2^{-0.356}$. This suggests taht the estimated median nuber of heart disease deaths decreases by 22% when wine consumption is doubled.

Note: D = average number of deathes per year per 1000 people by heart disease, and W = the number of liters consumed by one person per year on average.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.555552	0.126897	20.138801	0.0000000000
log(Wine)	-0.355596	0.052909	-6.720848	0.0000049136

Table 1: Summary for linear model of log(wine consumption) vs. log(mortality via heart disease)

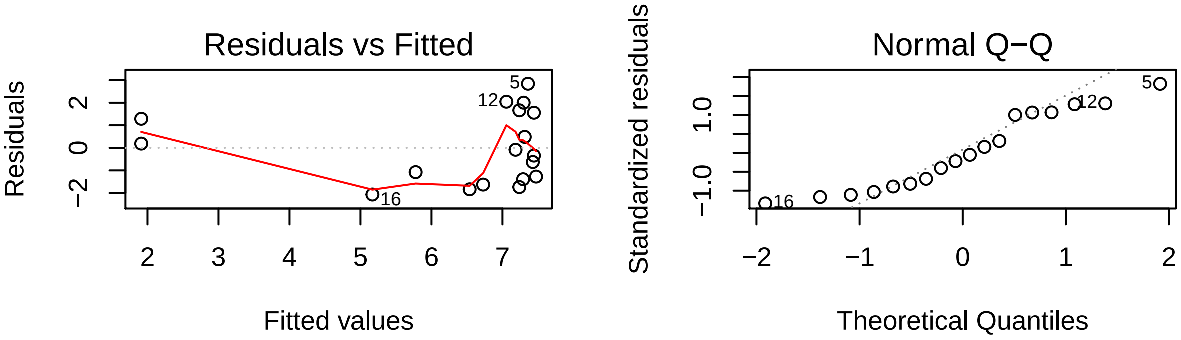
Graphical Display



Methods

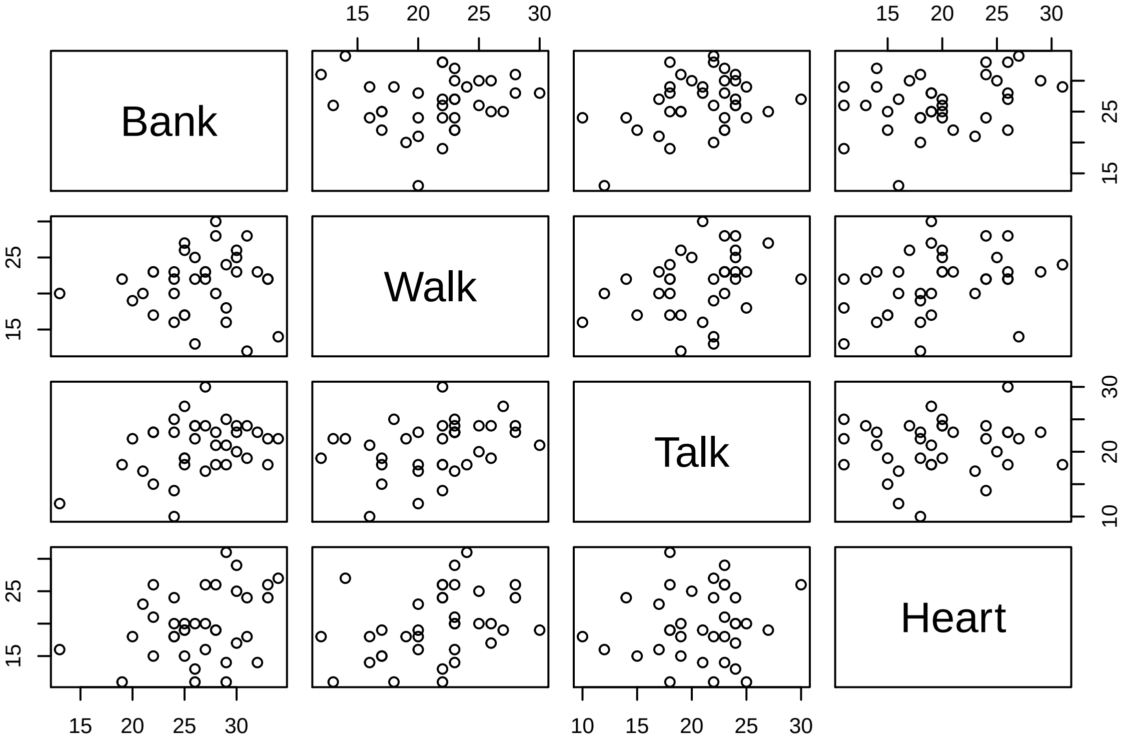
I tried a bunch of transformations to make the data fit the normality and linearity assumptions. I chose a log-log transformation because it seemed to make the data fit the assumptions best. I used an F-test to determine if the slope was significant from zero. This test has a null hypothesis that the slope is 0, when used in simple linear regression.

- To answer if one contry had an unusually high or low mortality based on their wine consumption, I looked at the log-log regression residuals. From the plot, we notice that countries #1, #5 and #12 seem to be particularly far from the line. If we look at the outliers of the regression on the original scale (see below), we see that #5, 12, and 16 are all somewhat far from the line. Overall, I think that countries #5 and #12 (Canada and Sweden) tend to be different from the other countries with similar wine consumptions, in that they have higher death rates that is associated with other countries that drink similar amounts of wine.



#5

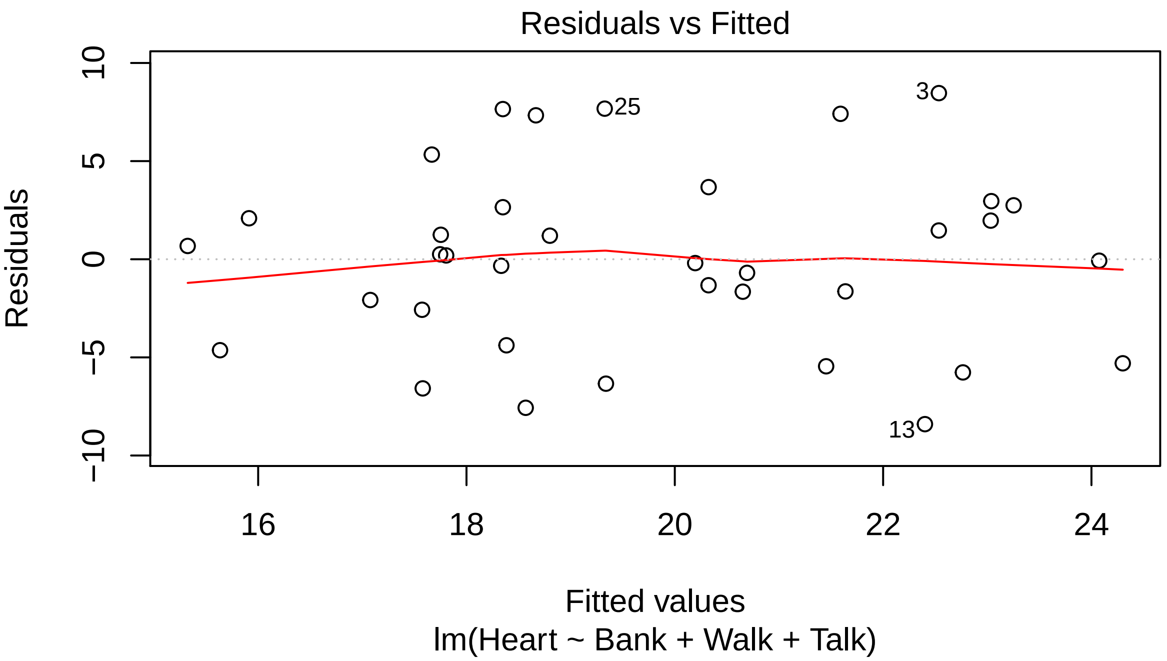
(a)



(b)

(c) Overall, the spread of the residuals may increase with increasing variance, but the fanning behavior is not really distinct. R has identified the top three outliers, but they are not especially far from 0.

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
##   \hline
## (Intercept) & 3.1787 & 6.3369 & 0.50 & 0.6194 \\
##   Bank & 0.4052 & 0.1971 & 2.06 & 0.0480 \\
##   Walk & 0.4516 & 0.2009 & 2.25 & 0.0316 \\
##   Talk & -0.1796 & 0.2222 & -0.81 & 0.4249 \\
##    \hline
## \end{tabular}
## \caption{Least Squares Fit for linear regression of
##              heart on bank, walk, and talk}
## \end{table}
```



(d) Here is a summary of the least squares fit

```
##
## Call:
## lm(formula = Heart ~ Bank + Walk + Talk, data = pace)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4014 -3.0263  0.0602  2.6748  8.4646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1787     6.3369   0.502   0.6194
## Bank          0.4052     0.1971   2.056   0.0480 *
## Walk          0.4516     0.2009   2.248   0.0316 *
## Talk         -0.1796     0.2222  -0.808   0.4249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.805 on 32 degrees of freedom
## Multiple R-squared:  0.2236, Adjusted R-squared:  0.1509
## F-statistic: 3.073 on 3 and 32 DF,  p-value: 0.04162
```

$\{\ \mu|bank, walk, talk\ \} = \beta_0 + \beta_1\dot{B}ank + \beta_2\dot{W}alk + \beta_3\dot{T}alk$. These are labeled as Beta[0], Beta[1], Beta[2],and Beta[3] below.

Here is a table of the coefficients and their standard errors:

	Beta[0]	Beta[1]	Beta[2]	Beta[3]
Estimate	3.18	0.41	0.45	-0.18
Std. Error	6.34	0.20	0.20	0.22

Code

```
steer <- 1:12
timeAfterSlaughter <- c(1,1,2,2,4,4,6,6,8,8, 24, 24)
pH <- c(7.02, 6.93, 6.42, 6.51, 6.07, 5.99, 5.59, 5.80, 5.51, 5.36, 5.30, 5.47)

x <- log(timeAfterSlaughter)
y <- pH

#plot(x,y)
mod1 <- lm(y~x)

plot(mod1, which = 1)

############# #1a
modaa <- aov(y~as.factor(x))
summary(modaa)

modab <- lm(y~x)
mm <- summary(modab)
mm$fstatistic # F-statistic that compares equal means vs. regression model
mm$coefficients[2,4] # p-value for F-test

# lack of fit F-test
mod1a <- anova(mod1)  # for the lm

mod2 <- aov(y~as.factor(x))
mod2a <- summary(mod2)

ssrLR <- mod1a[2,2]
dfLR <- mod1a$Df[2]
ssrSM <- sum(mod2$residuals^2)
dfSM <- mod2$rank
sigmaHatSM <- ssrSM/dfSM

FStat <- ((ssrLR - ssrSM)/(dfLR-dfSM))/sigmaHatSM

pvalue <- pf(FStat, df1 = dfLR - dfSM, df2 = dfSM, lower.tail = F)


## Check for lack of fit F-test
AOV_modelSeparateMeans<- aov(y ~ as.factor(x))
regmodel <- lm(y ~ x)
AOV_RegModel <- aov(regmodel)
lackOfFit <- anova(AOV_RegModel, AOV_modelSeparateMeans, test="F") # same as above
lackOfFit

############# 2a
pest <- read.csv("data/ex0817.csv")
par(mfrow = c(4,2))
with(pest, plot(Mass~Load))
with(pest, plot(log(Load),Mass))
with(pest, plot(sqrt(Load),Mass))
```

```r
with(pest, plot(1/(Load),(Mass)))
with(pest, plot(log(Load),log(Mass)))
with(pest, plot(log(Load),sqrt(Mass)))#
with(pest, plot(sqrt(Load),log(Mass)))## This looks best
with(pest, plot((Load),1/(Mass)))
par(mfrow = c(1,1))

############# 2b
modP <- lm(log(Mass)~sqrt(Load), data = pest)
plot(log(Mass)~sqrt(Load), data = pest)
abline(modP)
print("Here are the residuals")
modP$residuals
print("Here are the fitted values")
modP$fitted.values



############# 2c
plot(modP, which = 1)

with(pest, plot((Load)^(1/4),(Mass)^(1/6))) ## this might work, too
abline(lm(I(pest$Mass^(1/6))~I(pest$Load^(1/4))))



############# #3
wine <- read.csv("data/ex0823.csv")

par(mfrow = c(1,2))
plot(wine$Mortality~wine$Wine, xlab  = "Wine consumption (l/yr/person)",
     ylab = "Heart disease mortality rate (death/1000 ppl)")
#abline(lm(wine$Mortality~wine$Wine))


with(wine, plot(log(Mortality)~log(Wine), xlab = "Log Wine consumption (l/yr/person)",
                ylab = " Log Heart disease mortality rate (death/1000 ppl)"))
wineMod <- lm(log(wine$Mortality)~log(Wine), data = wine)
abline(wineMod)

print("Here are some diagnostic plots for the transformed data")


plot(wineMod, which = 1:2)
par(mfrow = c(1,1))

#
wm <- summary(wineMod)

#confint(wineMod, 2, level = 0.95 )

corxy <- cor(x = log(wine$Mortality), y = log(wine$Wine))
#corxy^2
```

```
options(xtable.comment = FALSE)
library(xtable)
w <- xtable(wineMod, caption = "Summary for linear model of log(wine consumption) vs. log(mortality via
digits(w) <- c(6,6,6,6,10)
print(w)

with(wine, plot(log(Mortality)~log(Wine), xlab = "Log Wine consumption (l/yr/person)",
                ylab = " Log Heart disease mortality rate (death/1000 ppl)",
                main = "Wine consumption vs. heart disease rate"))
abline(wineMod)


par(mfrow = c(2,2))
plot(lm(wine$Mortality~wine$Wine), which = 1:2)
par(mfrow = c(1,1))


############# 5a
pace <- read.csv("data/ex0914.csv")
library(ggplot2)
pairs(pace)

############# 5b
pmod <- lm(Heart~Bank + Walk + Talk, data = pace)
px <- xtable(pmod, caption = "Least Squares Fit for linear regression of
             heart on bank, walk, and talk")

#############5c
print(px)
plot(pmod, which = 1)

############# 5d
summary(pmod)

rs <- summary(pmod)
atb <- t(rs$coefficients[1:4, 1:2])
colnames(atb) <- c("Beta[0]", 'Beta[1]', 'Beta[2]', "Beta[3]")
print(xtable(atb), floating = F)
```