Lab 3

Greg Ceccarelli

July 29, 2015

Part 1. Multiple Choice (4 points each)

1.
2.
3.
4.
5.
6.
7.
8.

Part 2. Test Selection (20 points)

9.
10.
11.
12.
13.

Part 3: Data Analysis and Short Answer (48 points)

14. Task 1: Conduct a chi-square test to determine if there is an association between marital status (marital) and political orientation (politics).

```
#set up working envrionment
##running on macbook air, set relevant directory
setwd('/Users/ceccarelli/MIDS/DATASCI_W203/Assignments/Labs/Lab 3/')
rm( list = ls() )

#Load relevant packages
library(ggplot2)
library(car)
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following object is masked from 'package:car':
##
##     logit
##
## The following object is masked from 'package:ggplot2':
##
##     %+%
```

```
library(gmodels)
library(MASS)

#load supplied R data file
load("GSS.Rdata")

#define function to omit row values where only certain columns are NA
data.complete <- function(data, desiredCols) {
  completeVec <- complete.cases(data[, desiredCols])
  return(data[completeVec, ])
}

#question 14
#inspect variable to check on categories
summary(GSS$marital)
```

```
##       married       widowed      divorced     separated never married
##           795           165           213            40           286
##            NA
##             1
```

```
##appears NA is a string in Marital, recode to true NA
is.na(GSS) <- GSS=="NA"

# Can fix the remaining "NA" after fix by recreating the factor
GSS$marital <- factor(GSS$marital)

# true NA's now
summary(GSS$marital)
```

```
##       married       widowed      divorced     separated never married
##           795           165           213            40           286
##          NA's
##             1
```

```
# check politics as well
summary(GSS$politics)
```

```
##      Liberal     Tend Lib     Moderate    Tend Cons Conservative
##          193           193           527           248           282
##         NA's
##           57
```

```r
#created NA subsetted dataset to run chi square test on
GSS.mar_pol <- data.complete(GSS[,c("marital","politics")], c("marital","politics"))


#review tabulated results after removing null
table(GSS.mar_pol)
```

```
##                politics
## marital          Liberal Tend Lib Moderate Tend Cons Conservative
##   married             93       92      271       140          173
##   widowed             15       16       57        24           37
##   divorced            22       36       79        38           29
##   separated            7        3       22         6            1
##   never married       55       46       98        40           42
```

```r
##leverage CrossTable function to compute Pearson's chisq test
CrossTable(GSS.mar_pol$marital, GSS.mar_pol$politics,
           prop.c = FALSE, prop.t = FALSE, prop.chisq = FALSE,
           chisq = TRUE, expected = TRUE, sresid = TRUE, format = "SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## |             Row Percent |
## |            Std Residual |
## |-------------------------|
##
## Total Observations in Table:  1442
##
##                      | GSS.mar_pol$politics
## GSS.mar_pol$marital  |      Liberal  |     Tend Lib  |     Moderate  |    Tend Cons  | Conservative
## ---------------------|---------------|---------------|---------------|---------------|---------------|-----
##              married |           93  |           92  |          271  |          140  |          173  |
##                      |      102.391  |      102.924  |      281.042  |      132.255  |      150.387  |
##                      |       12.094% |       11.964% |       35.241% |       18.205% |       22.497% |
##                      |       -0.928  |       -1.077  |       -0.599  |        0.673  |        1.844  |
## ---------------------|---------------|---------------|---------------|---------------|---------------|-----
##              widowed |           15  |           16  |           57  |           24  |           37  |
##                      |       19.839  |       19.942  |       54.454  |       25.626  |       29.139  |
##                      |       10.067% |       10.738% |       38.255% |       16.107% |       24.832% |
##                      |       -1.086  |       -0.883  |        0.345  |       -0.321  |        1.456  |
## ---------------------|---------------|---------------|---------------|---------------|---------------|-----
##             divorced |           22  |           36  |           79  |           38  |           29  |
##                      |       27.162  |       27.304  |       74.555  |       35.085  |       39.895  |
##                      |       10.784% |       17.647% |       38.725% |       18.627% |       14.216% |
##                      |       -0.991  |        1.664  |        0.515  |        0.492  |       -1.725  |
## ---------------------|---------------|---------------|---------------|---------------|---------------|-----
##            separated |            7  |            3  |           22  |            6  |            1  |
##                      |        5.193  |        5.220  |       14.253  |        6.707  |        7.627  |
##                      |       17.949% |        7.692% |       56.410% |       15.385% |        2.564% |
##                      |        0.793  |       -0.972  |        2.052  |       -0.273  |       -2.400  |
```

```
## --------------------|-------------|-------------|-------------|-------------|-------------|-----
##       never married |          55 |          46 |          98 |          40 |          42 |
##                     |      37.415 |      37.610 |     102.696 |      48.327 |      54.953 |
##                     |      19.573% |      16.370% |      34.875% |      14.235% |      14.947% |
##                     |       2.875 |       1.368 |      -0.463 |      -1.198 |      -1.747 |
## --------------------|-------------|-------------|-------------|-------------|-------------|-----
##        Column Total |         192 |         193 |         527 |         248 |         282 |
## --------------------|-------------|-------------|-------------|-------------|-------------|-----
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  44.2255     d.f. =  16     p =  0.0001822704
##
##
##
##        Minimum expected frequency: 5.192788
```

```
# also leverage a different function to compare results
cs <- chisq.test(GSS$marital, GSS$politics)

#since this is a test of indepedence, leverage cramer's v for effect size
cramers_v = function(cs)
{
  cv = sqrt(cs$statistic / (sum(cs$observed) * (min(dim(cs$observed))-1)))
  print.noquote("Cramer's V:")
    return(as.numeric(cv))
}

cramers_v(cs)
```

```
## [1] Cramer's V:

## [1] 0.08756363
```

RESPONSE: A. What are the null and alternative hypothesis for your test? We use the chi-sq test to test the relationship between marital status The null hypothesis is that marital status and political orientation are indepedent the alternative hypothesis is that marital status and political orientation are not indepedent (they are associated)

B. What test statistic and p-value do you get? The value of the chi-square statistic is 44.2255. This value is highly significant ($p < .0001$), indicating that marital status has a significant effect on political orientation.

C. Conduct an effect size calculation for your relationship. Using cramer's v returns an effect size of 0.08756363.

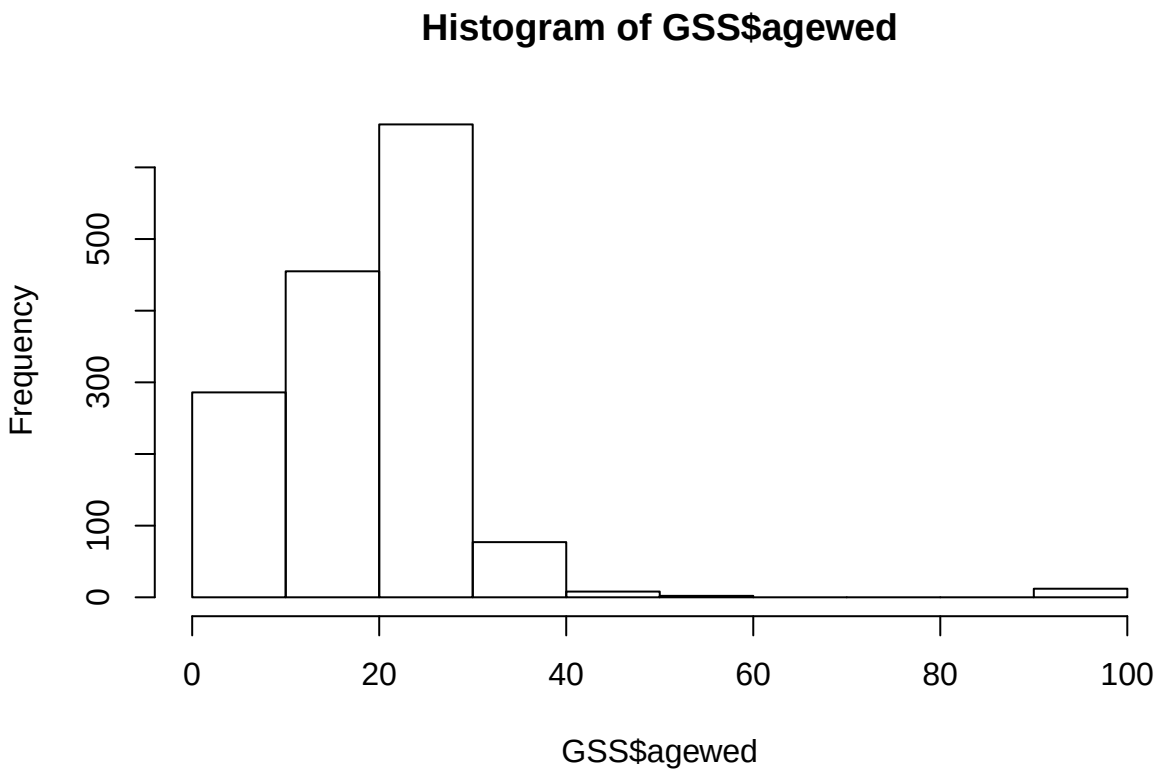
D. Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results? While the results from the chi-square test are statistically significant, it is hard to make a good conclusion given the lack of practical significance (anything under .2 is considered weak). Thus while the variable's are associated, their association is not very strong.

15. Task 2: Conduct a Pearson correlation analysis to examine the association between age when married (agewed) and hours of tv watched (tvhours).

```
#question 15
#check how linear the variables are
#inspect variable
describe(GSS$agewed)
```

```
##   vars    n  mean    sd median trimmed  mad min max range skew kurtosis
## 1    1 1500 19.06 12.33     21   19.04 4.45   0  99    99 1.71    12.62
##     se
## 1 0.32
```

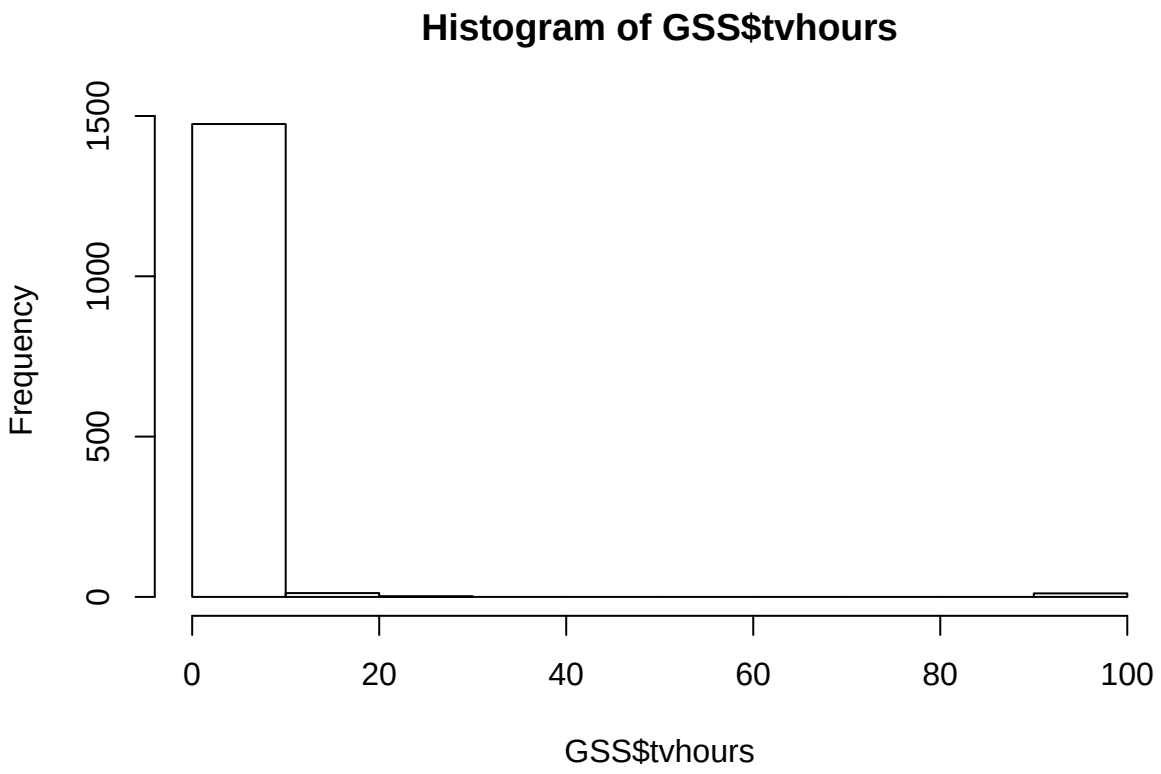
```
#agewed is is not normally distributed and exhibits leptokurtic positive skew
hist(GSS$agewed)
```



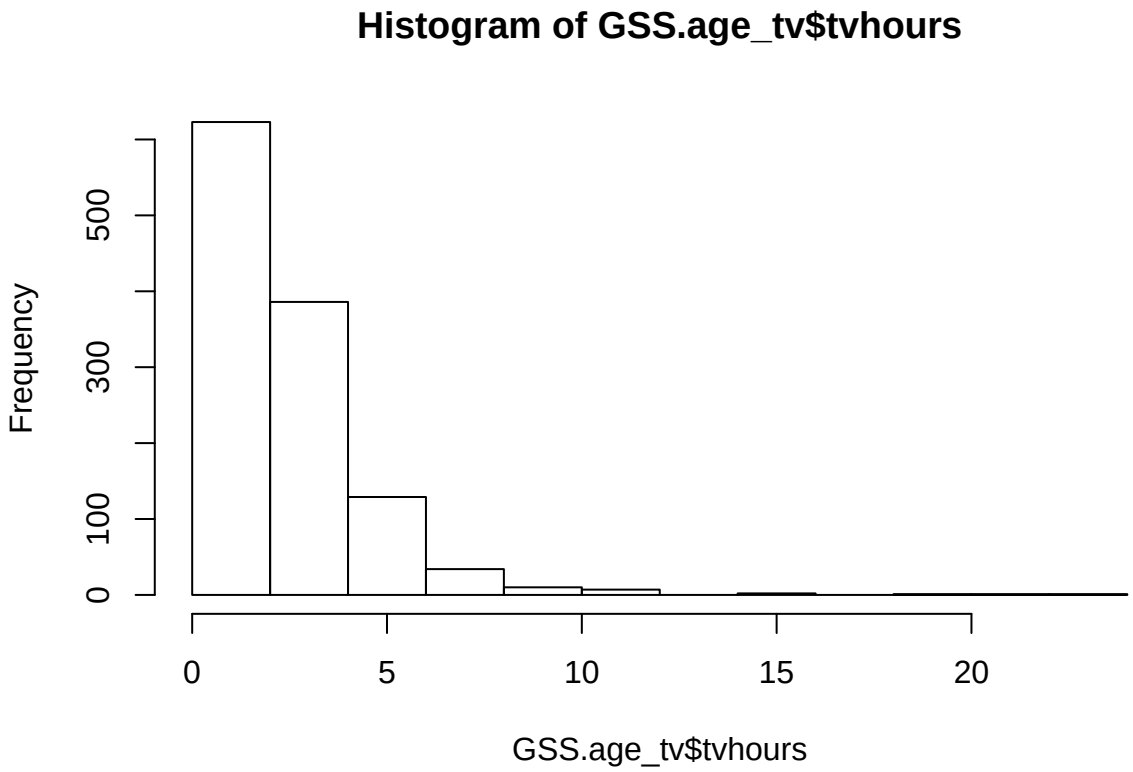

```
#inspect tvhours
describe(GSS$tvhours)
```

```
##   vars    n mean  sd median trimmed  mad min max range skew kurtosis   se
## 1    1 1500  3.6 8.5      2     2.6 1.48   0  99    99 10.4   113.43 0.22
```

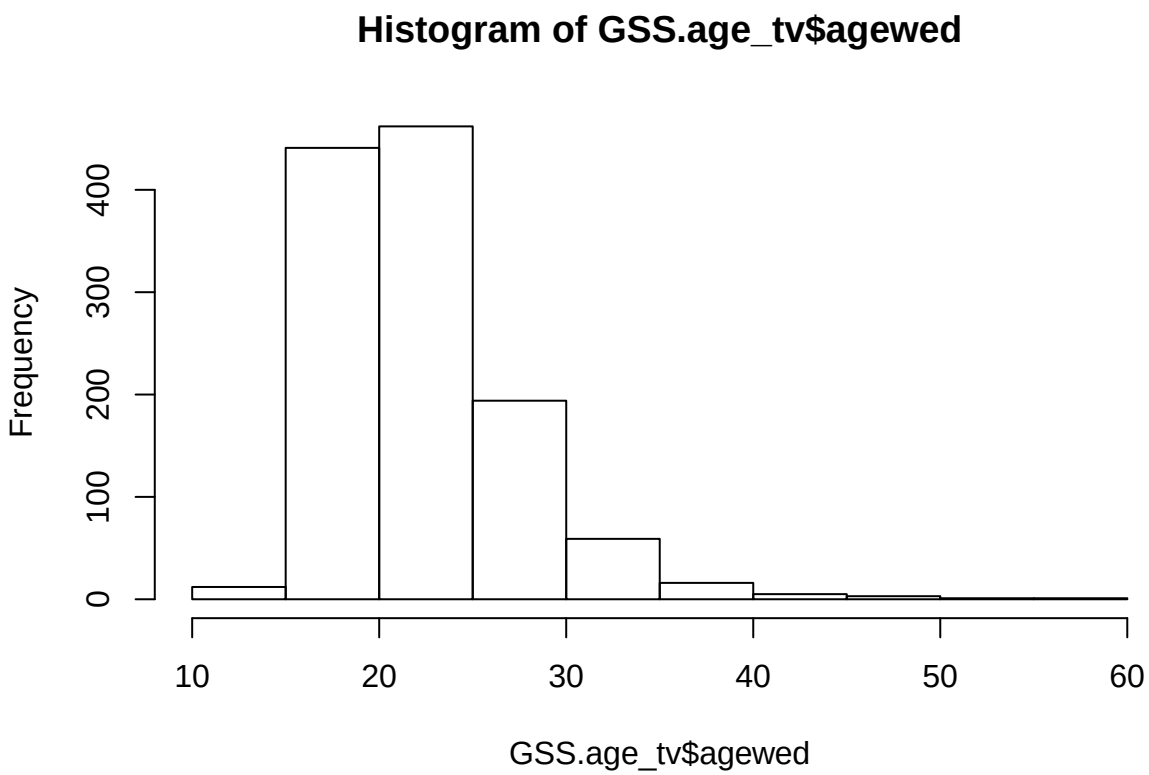
```
#tvhours are even more non-normally distributed and have some outliers
hist(GSS$tvhours)
```



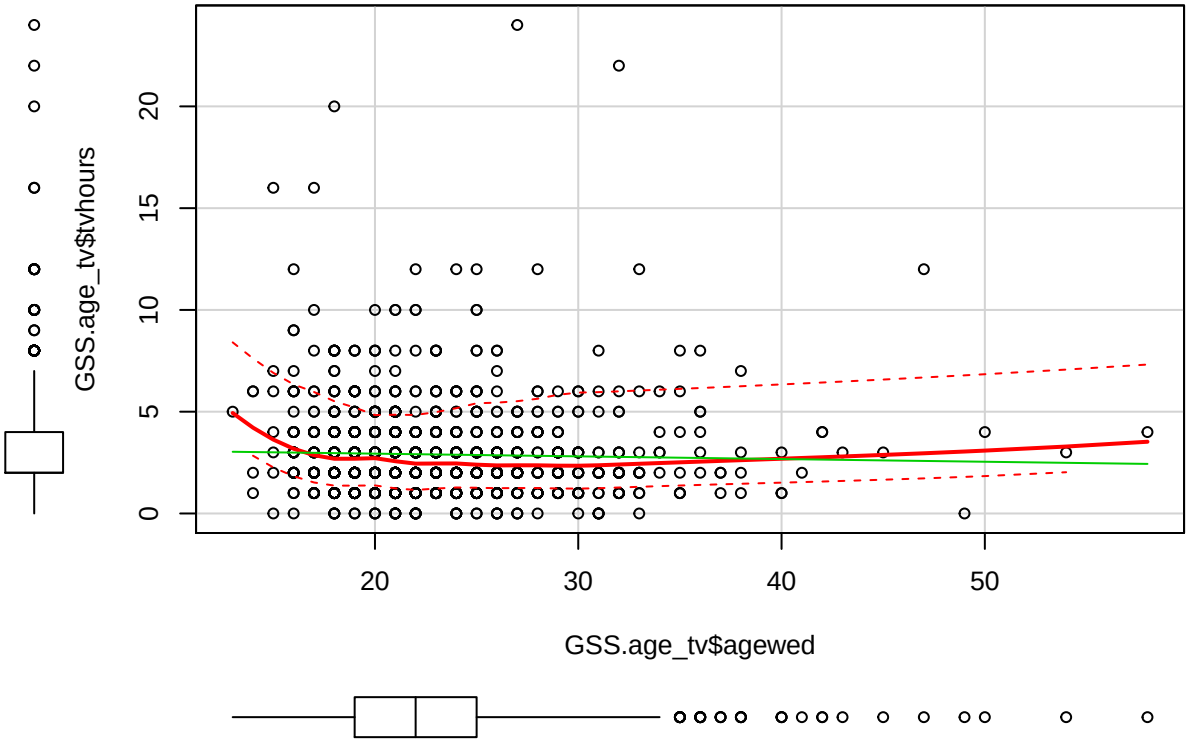

```r
#subset dataset because tvhours variable  cannot exceed 24
GSS.age_tv <- subset(GSS, tvhours <= 24, select = c(agewed,tvhours))

#subset again because 0, 98 and 99 are invalid for agewed
GSS.age_tv <- subset(GSS.age_tv, agewed > 0 & agewed<=90)

#review histogram again
hist(GSS.age_tv$tvhours)
```




```
hist(GSS.age_tv$agewed)
```



```r
# Use a scatterplot to see how linear the relationship looks on the new
# variables
scatterplot(GSS.age_tv$agewed, GSS.age_tv$tvhours)
```




```r
#check the correlation
cor.results <- cor.test(GSS.age_tv$agewed, GSS.age_tv$tvhours,
                        use = "complete.obs", method = "pearson")

cor.results
```

```
##
##  Pearson's product-moment correlation
##
## data:  GSS.age_tv$agewed and GSS.age_tv$tvhours
## t = -1.0349, df = 1192, p-value = 0.3009
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08654554  0.02681630
## sample estimates:
##         cor
## -0.02996096
```

```r
# compute r^2
cor.results$estimate**2
```

```
##          cor
## 0.0008976593
```

```
## compute r^2 another way with a correlation matrix
cor(GSS.age_tv[,c("agewed", "tvhours")], use = "complete.obs")**2
```

```
##                agewed      tvhours
## agewed  1.0000000000 0.0008976593
## tvhours 0.0008976593 1.0000000000
```

RESPONSE: A. What are the null and alternative hypotheses for your test? The null hypothesis is that the linear relationship between agewed and tvhours is 0 The alternative hypothesis is that the linear relationship is not 0

B. What test statistic and p-value do you get? Test statistic of -1.0349 and a p-value of .3009.

Also, an R value of -.0299. 95 percent confidence interval -0.08654554 and 0.02681630. R^2 value of 0.0008976593.

C. Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results? Given the very high pvalue, there is not enough evidence to reject the null hypothesis that the linear relationship is 0 (aka there is not one). This is confirmed via practical signifance and the confidence interval that spans 0.

16. Task 3: Create a new binary/dummy variable, "married", that denotes whether an individual is currently married or not currently married. Next, we want to consider just the subpopulation of 23-year olds in this sample. Conduct a Wilcox rank-sum test to determine whether your new "married" variable is associated with the number of children (childs) for respondents who are 23 years old.

```
#question 16
#create dummy variable for married using if else
GSS$marital_dummy <- ifelse(GSS$marital == "married",1,0)

#check summary stats for this variable
summary(GSS$marital_dummy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  1.0000  0.5304  1.0000  1.0000       1
```

```
#create subset for variable
GSS.marital_dummy <- subset(GSS, age == 23, select = c(age,marital_dummy,childs
                                                       ))
#View(GSS.marital_dummy)

##look at mean/proportion of married in new subset
mean(GSS.marital_dummy$marital_dummy)
```

```
## [1] 0.2857143
```

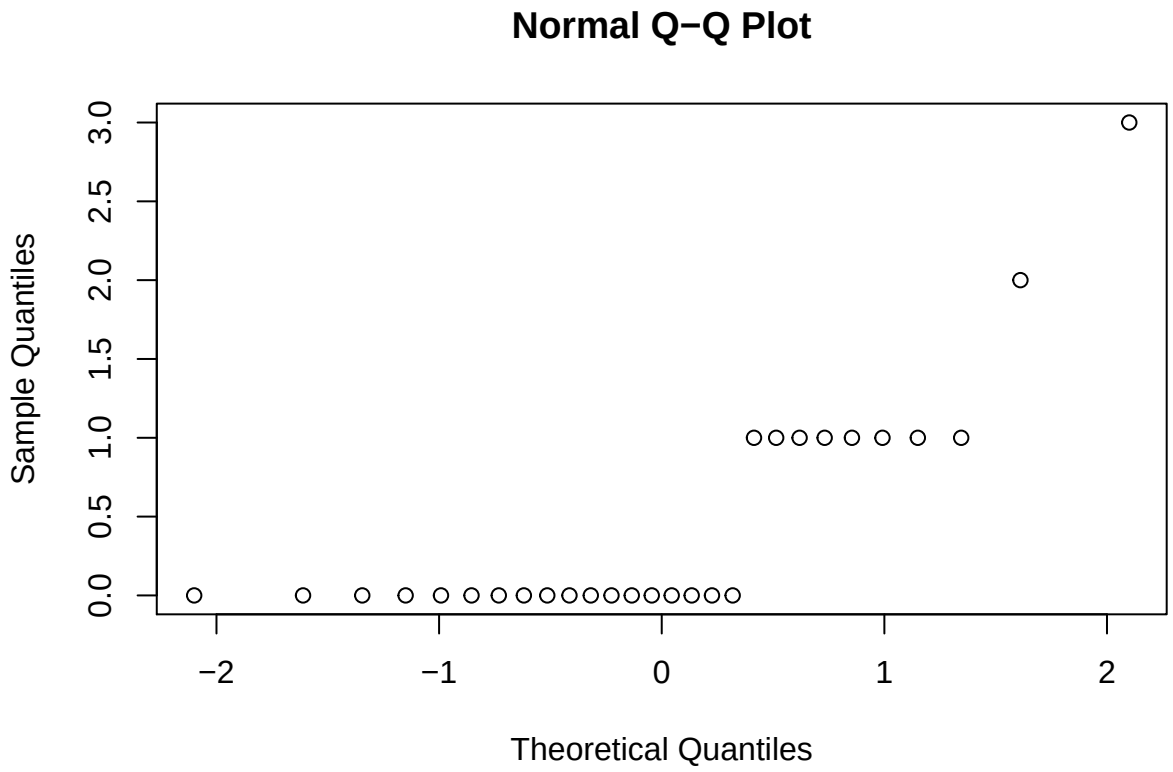
```
summary(GSS.marital_dummy)
```

```
##       age     marital_dummy        childs
##  Min.   :23   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:23   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :23   Median :0.0000   Median :0.0000
##  Mean   :23   Mean   :0.2857   Mean   :0.4643
##  3rd Qu.:23   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :23   Max.   :1.0000   Max.   :3.0000
```

```r
# We may ask whether the married 23 years olds have more or less children
by(GSS.marital_dummy$childs,GSS.marital_dummy$marital_dummy, mean, na.rm = TRUE)
```

```
## GSS.marital_dummy$marital_dummy: 0
## [1] 0.15
## ------------------------------------------------------------
## GSS.marital_dummy$marital_dummy: 1
## [1] 1.25
```

```r
# Notice that number of children is not at all normal.
qqnorm(GSS.marital_dummy$childs)
```




```r
# Use the Wilcoxon rank-sum test to compare means
wilcox.test(GSS.marital_dummy$childs ~ GSS.marital_dummy$marital_dummy)
```

```
## Warning in wilcox.test.default(x = c(0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  GSS.marital_dummy$childs by GSS.marital_dummy$marital_dummy
## W = 19, p-value = 0.0002656
## alternative hypothesis: true location shift is not equal to 0
```

```r
# leverage cohensd function from previous work to compute effect size
cohens_d <- function(x, y) {
  # this function takes two vectors as inputs, and compares
    # their means

    # first, compute the pooled standard error
  lx = length(subset(x,!is.na(x)))
  ly = length(subset(y,!is.na(y)))
    # numerator of the pooled variance:
    num = (lx-1)*var(x, na.rm=T) + (ly-1)*var(y, na.rm=T)
    pooled_var = num / (lx + ly - 2) # variance
    pooled_sd = sqrt(pooled_var)

    # finally, compute cohen's d
    cd = abs(mean(x, na.rm=T) - mean(y, na.rm=T)) / pooled_sd
    return(cd)
}


#need vectors to plug into cohens d
children_m = GSS.marital_dummy$childs[GSS.marital_dummy$marital_dummy==1]
children_nm = GSS.marital_dummy$childs[GSS.marital_dummy$marital_dummy==0]

# plug them into our cohen's d function
cohens_d(children_m, children_nm)
```

```
## [1] 1.976885
```

RESPONSE: A. What is the mean of your new "married" variable among 23-year-olds (e.g., the proportion of cases in the category coded "1")? .2857 or 28%

B. What is the null and alternative hypotheses for your test? The null hypothesis is the mean number of children had by individuals (married or not) is identical The alternative hypothesis is that the mean number of children had by individuals (married or not) is not identical

C. What test statistic and p-value do you get? W = 19 and p = .0002656

D. Conduct an effect size calculation for your relationship. Using cohens d, the effect size is extraordinarily big! 1.976885

E. Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results? In light of evaluating the extremely significant p value and the large effect size, we can safely conclude that the means between these two groups are different.

17. Task 4: Conduct an analysis of variance to determine if there is an association between religious affiliation (relig) and age when married (agewed).

```r
#question 17

library(ggplot2)
##inspect religion variable
summary(GSS$relig)
```

```
## Protestant   Catholic     Jewish       None      Other         DK
```

```
##        953        333         31        140         35          1
##         NA       NA's
##          0          7
```

```
#check out levels
levels(GSS$relig)
```

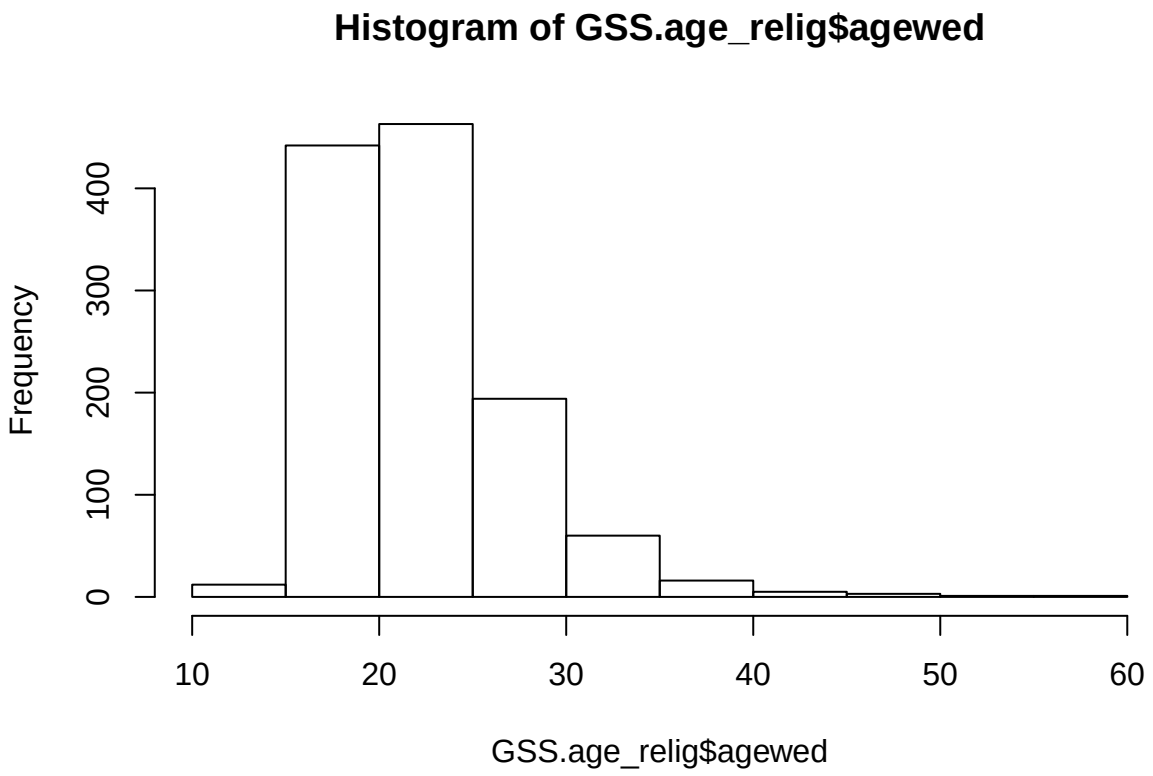
```
## [1] "Protestant" "Catholic"   "Jewish"     "None"       "Other"
## [6] "DK"         "NA"
```

```
#wrapper function to easily recode factors
changelevels <- function(f, ...) {
    f <- as.factor(f)
    levels(f) <- list(...)
    f
}

GSS$relig_clean <- changelevels(GSS$relig, Protestant=c("Protestant"), Catholi=c("Catholic"), Jewish=c(
```

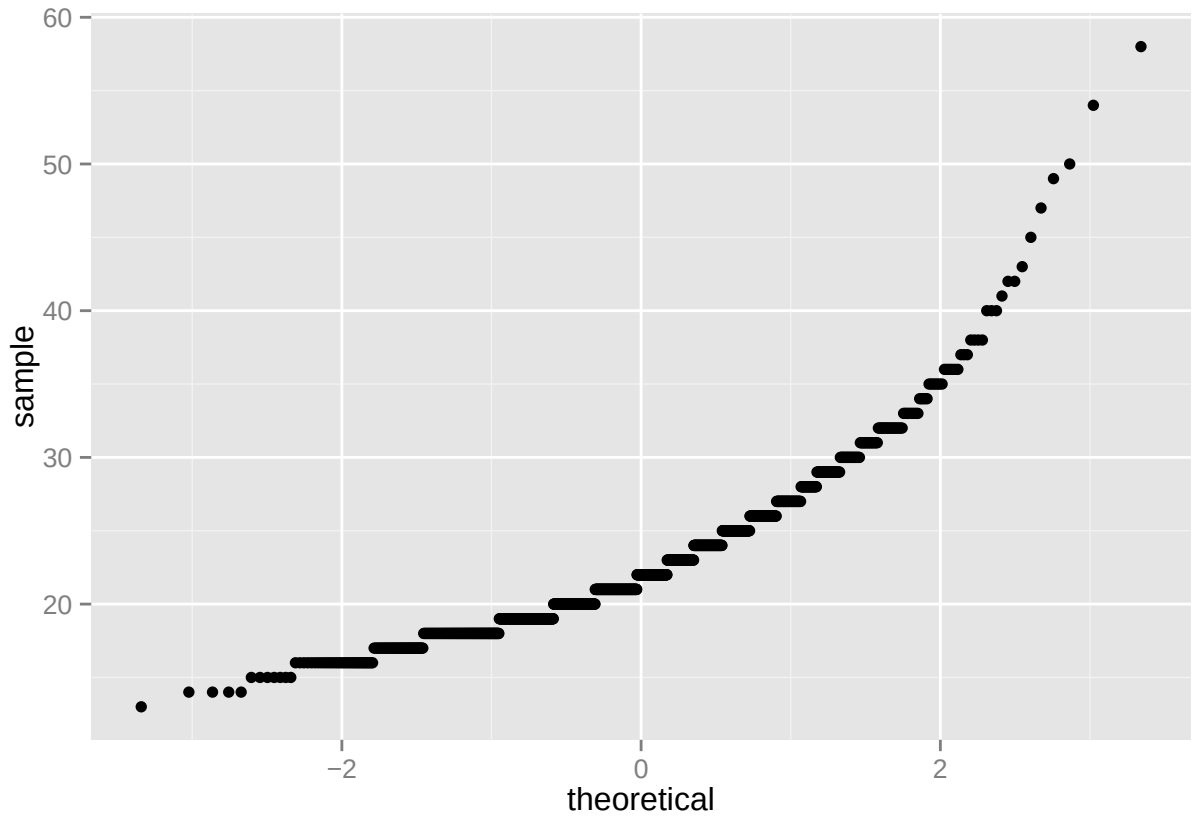
```
#recheck out levels
levels(GSS$relig_clean)
```

```
## [1] "Protestant" "Catholi"    "Jewish"     "None"       "Other"
```

```
#remove spurious agewed variables
GSS.age_relig <- subset(GSS, agewed > 0 & agewed<=90)

#remove NA's
GSS.age_relig <- data.complete(GSS.age_relig ,c("agewed","relig_clean"))

# check the agewed variable for normality
hist(GSS.age_relig$agewed)
```




```
qqplot = qplot(sample = GSS.age_relig$agewed, stat="qq")
qqplot
```



```r
# Normality is not great but ANOVA is a robust-test
# also data is interval

# Let's look at the means, by each category and overall
by(GSS.age_relig$agewed, GSS.age_relig$relig_clean, mean, na.rm=T)
```

```
## GSS.age_relig$relig_clean: Protestant
## [1] 22.25286
## ------------------------------------------------------------
## GSS.age_relig$relig_clean: Catholi
## [1] 23.63396
## ------------------------------------------------------------
## GSS.age_relig$relig_clean: Jewish
## [1] 23.77119
## ------------------------------------------------------------
## GSS.age_relig$relig_clean: None
## [1] NA
## ------------------------------------------------------------
## GSS.age_relig$relig_clean: Other
## [1] 25.37037
```

```r
mean(GSS.age_relig$agewed, na.rm=T)
```

```
## [1] 22.77861
```

```r
# Perform the analysis of variance and check the significance
aovm = aov(agewed ~ relig_clean, GSS.age_relig)
summary(aovm)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## relig_clean    3    709  236.35   9.543 3.14e-06 ***
## Residuals   1193  29547   24.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# perform a post hoc test to check for pairwise differences in means... use
# bonferroni correction
pw = pairwise.t.test(GSS.age_relig$agewed, GSS.age_relig$relig_clean, p.adjust.method = "bonferroni")

pw
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  GSS.age_relig$agewed and GSS.age_relig$relig_clean
##
##         Protestant Catholi Jewish
## Catholi 0.00059    -       -
## Jewish  0.01227    1.00000 -
## Other   0.00845    0.50641 0.79360
##
## P value adjustment method: bonferroni
```

```
#make output clearer using a function
sig_stars = function(p)
{
  stars = symnum(p, na = F, cutpoints = c(0, .001, .01, .05, .1, 1), symbols=c("***","**", "*", ".", " "
    return( paste(round(p, 3), stars) )
}

# apply our new function to every element in our matrix
t_table = noquote( apply(pw$p.value, c(1,2), sig_stars) )
t_table
```

```
##         Protestant Catholi Jewish
## Catholi 0.001 ***  NA      NA
## Jewish  0.012 *    1       NA
## Other   0.008 **   0.506   0.794
```

RESPONSE:

A. What is the null and alternative hypotheses for your test? The null hypothesis is that there are no differences in the mean wed age between religious groups The alternative hypothesis is that there are differences in the mean wed age between the religious groups

B. What test statistic and p-value do you get? F=8.199 and p value 1.6e-06. The overall (omnibus) model is very significant.

C. Are there statistically significant differences between individual pairs of groups, and if so, how do you know? Yes, using a post hoc test (pairwise ttest with the bonferroni correction), shows that there are statistically significant differences in wed age between Protestant and Catholics (largest significant difference), Protestant and Jewish and Protestant and Other groups.

D. Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results?

While there is an overall difference in mean age when wed, the most interesting statistically significant differences occur between Protestants and all groups except (None)