Descriptive statistics by urban/rural area

Maggie Walters

June 16, 2017

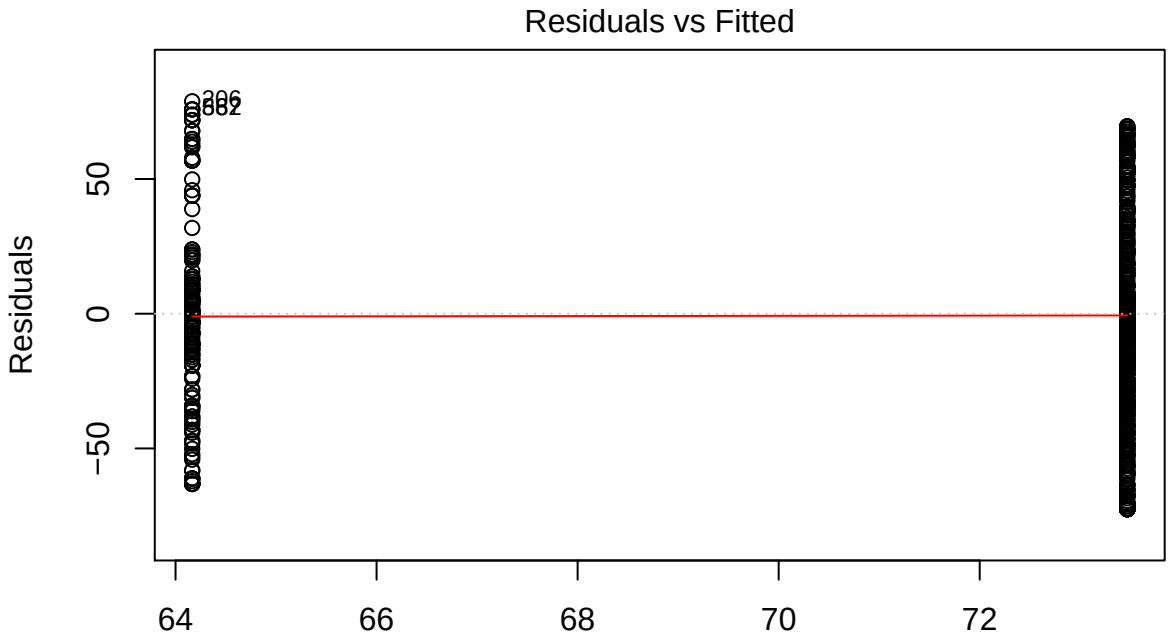
ANOVA with RMSP as a factor

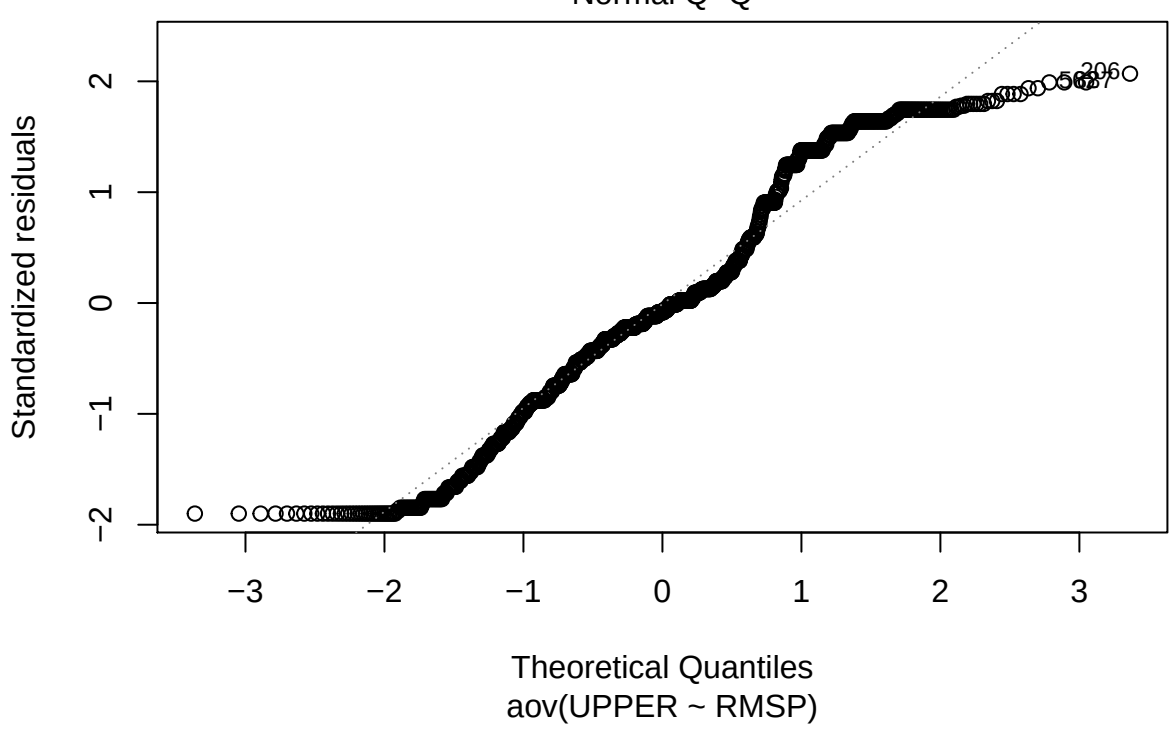
```
#list of each county
county_vec <- as.character(unique(data$COUNTY))
#create blank upper quartile matrix
upperq_matrix <- matrix(rep(NA, length(county_vec) * 12), nrow = 12, ncol = length(county_vec))
colnames(upperq_matrix) <- county_vec
months <- c("January", "Febuary", "March",
            "April", "May", "June",
            "July", "August", "September",
            "October", "November", "December")
row.names(upperq_matrix) <- months

#find upper quartile for each county in each month
for(i in 1:12){
  x <- subset(data, data$MONTH == i)
  for(j in 1:387){
    which.county_j <- which(x$COUNTY == county_vec[j])
    county_j_ages <- x$AGE[which.county_j]
    upperq_matrix[i,j] <- quantile(county_j_ages)[4]
  }
}
upper_matrix <- matrix(rep(NA, 4 * 12*length(county_vec)), ncol = 4)
colnames(upper_matrix) <- c("UPPER", "COUNTY", "MONTH", "RMSP")

#fill in upper
for(i in 1:12 * length(county_vec)){
  upper_matrix[,1] <- upperq_matrix[,]
}
#fill in months
upper_matrix[,3] <- rep(seq(1:12), length(county_vec))
#fill in counties
upper_matrix[seq(1,12),2] <- rep(county_vec[1], 12)
for(i in 1:49){
  x <- 12 * (i-1)
  upper_matrix[x + seq(1,12),2] <- rep(county_vec[i], 12)
}

#isolate RMSPs
x <- rep(NA, length(county_vec))
for(i in 1:length(county_vec)){
  which.i <- which(data$COUNTY == county_vec[i])
  y <- rep(NA, length(which.i))
  y <- data$RMSP[which.i]
  x[i] <- mean(y)
}
```

```r
rmsp_vec <- as.character(x)

#fill in RMSPs
upper_matrix[seq(1,12),4] <- rep(rmsp_vec[1], 12)
for(i in 1:length(county_vec)){
  x <- 12 * (i-1)
  upper_matrix[x + seq(1,12),4] <- rep(rmsp_vec[i], 12)
}

#convert to data frame
upper_matrix <- as.data.frame(upper_matrix)

#remove NAs
upper_matrix <- subset(upper_matrix, !is.na(UPPER))

#change RMSP to a factor
upper_matrix$RMSP <- as.factor(upper_matrix$RMSP)

#change UPPER to a number
upper_matrix$UPPER <- as.numeric(upper_matrix$UPPER)

#THIS WORKED, p = 0.000292
rmsp_aov_mod <- aov(UPPER ~ RMSP, data = upper_matrix)

#assess normality
shapiro.test((resid(rmsp_aov_mod)))
```

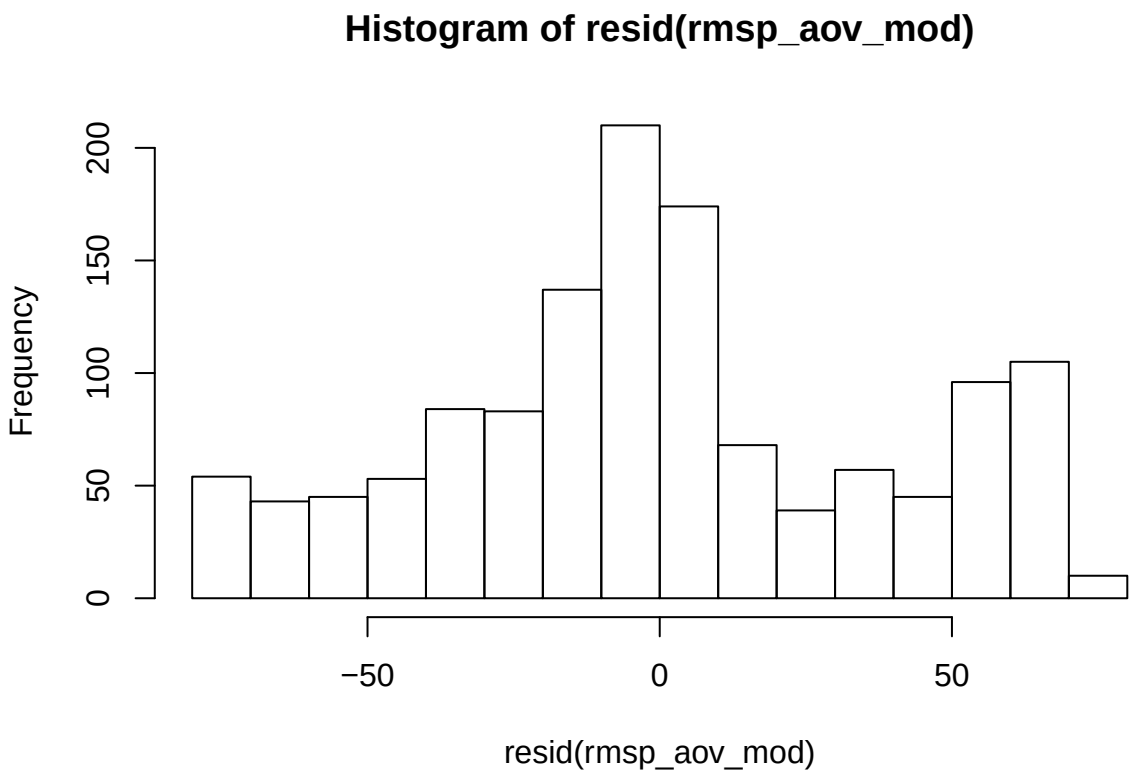
```
##
##  Shapiro-Wilk normality test
##
## data:  (resid(rmsp_aov_mod))
## W = 0.96439, p-value < 2.2e-16
```

```r
plot(rmsp_aov_mod, which = c(1,2))
```






```
hist(resid(rmsp_aov_mod))
```




```
#NON-PARAMETRIC, p = 0.0008138
upper_rmsp_np <- kruskal.test(UPPER ~ RMSP, data = upper_matrix)
```

Preliminary analysis

- rmsp_aov_mod yielded significant results with a p-value of 0.000292.
  - Unfortunately, the residuals are not normally distributed. Having a hard time with transformations so going to try a non-parametric test.
- A non-parametric test (Kruskall-Wallis) was completed with the following null and alternative hypotheses:
  - H_O : RMSP does not influence the upper quartile value for age of onset.
  - H_A : RMSP influences the upper quartile value for age of onset.
- The KW test revealed a p-value of 0.0008138 which allows for the rejection of H_O and acceptance of the alternative hypothesis, that RMSP influences the upper quartile value for age of onset.

Descriptive statistics for upper quartile

```
upper_mean_is <- subset(upper_matrix, upper_matrix$RMSP == 1)
y1 <- mean(upper_mean_is$UPPER)
x1 <- sd(upper_mean_is$UPPER)

upper_mean_not <- subset(upper_matrix, upper_matrix$RMSP == 0)
y2 <- mean(upper_mean_not$UPPER)
x2 <- sd(upper_mean_not$UPPER)

#sort by month for both
```

```
x <- order(upper_mean_is$MONTH)
upper_mean_is <- upper_mean_is[order(upper_mean_is$MONTH),]
```

Prliminary analysis

- In order to provide some sort of descriptive statistics here, the mean of those within the regional metropolitan area (RMSP = 1) is 64.165493 and the standard deviation is 28.9415899.
- The mean of those not within the regional metropolitan area (RMSP = 0) is 73.468106 and the standard deviation is 40.3572421.

TO DO:

- Sort by month regardless of county, so that progress of the epidemic within each county can be accounted for.
- Plot upper_mean_is (black) and upper_mean_not (red) in order to look for a visual trend/sign that there is a difference between counties within the regional metropolitan area and not within the regional metropolitan area.

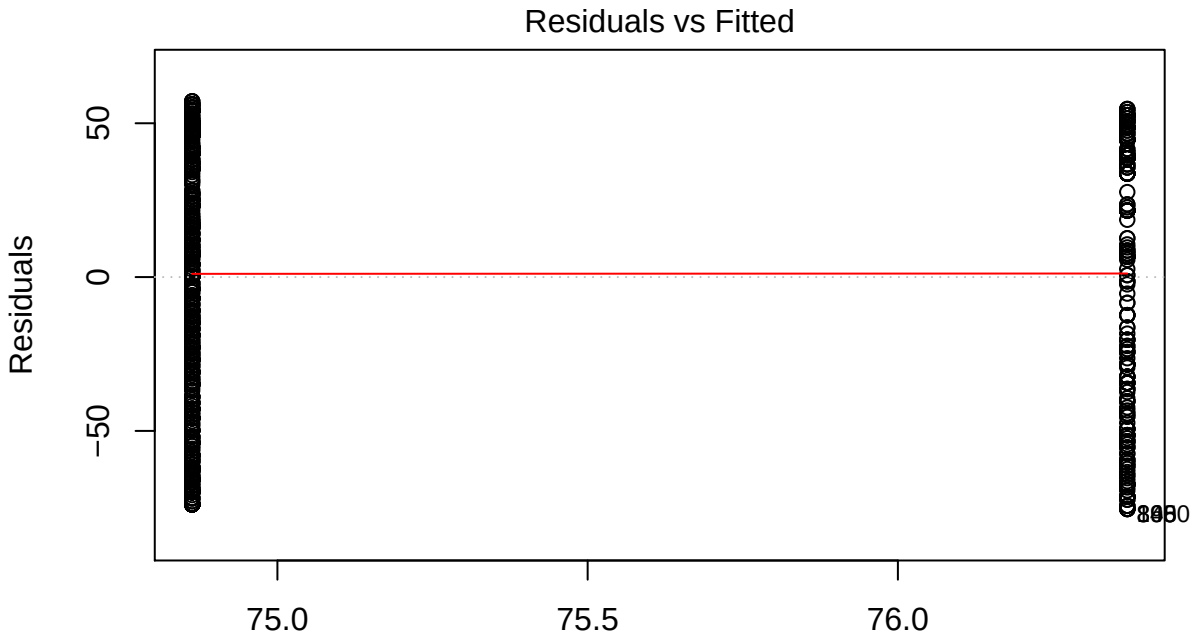
RMSP and Lower quartile

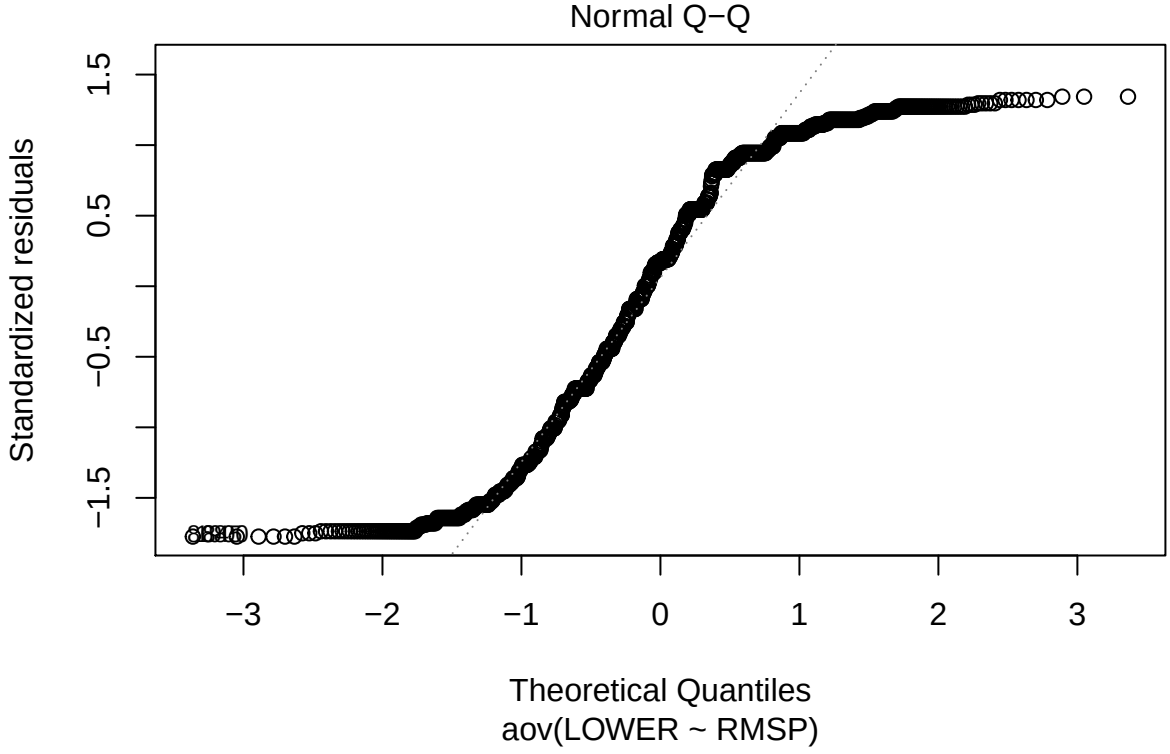
```
#list of each county
county_vec <- as.character(unique(data$COUNTY))
#create blank upper quartile matrix
lowerq_matrix <- matrix(rep(NA, length(county_vec) * 12), nrow = 12, ncol = length(county_vec))
colnames(lowerq_matrix) <- county_vec
months <- c("January", "Febuary", "March",
            "April", "May", "June",
            "July", "August", "September",
            "October", "November", "December")
row.names(lowerq_matrix) <- months

#find upper quartile for each county in each month
for(i in 1:12){
  x <- subset(data, data$MONTH == i)
  for(j in 1:387){
    which.county_j <- which(x$COUNTY == county_vec[j])
    county_j_ages <- x$AGE[which.county_j]
    lowerq_matrix[i,j] <- quantile(county_j_ages)[2]
  }
}
lower_matrix <- matrix(rep(NA, 4 * 12*length(county_vec)), ncol = 4)
colnames(lower_matrix) <- c("LOWER", "COUNTY", "MONTH", "RMSP")

#fill in upper
for(i in 1:12 * length(county_vec)){
  lower_matrix[,1] <- lowerq_matrix[,]
}
#fill in months
lower_matrix[,3] <- rep(seq(1:12), length(county_vec))
#fill in counties
lower_matrix[seq(1,12),2] <- rep(county_vec[1], 12)
for(i in 1:49){
```

```
  x <- 12 * (i-1)
  lower_matrix[x + seq(1,12),2] <- rep(county_vec[i], 12)
}

#isolate RMSPs
x <- rep(NA, length(county_vec))
for(i in 1:length(county_vec)){
  which.i <- which(data$COUNTY == county_vec[i])
  y <- rep(NA, length(which.i))
  y <- data$RMSP[which.i]
  x[i] <- mean(y)
}
rmsp_vec <- as.character(x)

#fill in RMSPs
lower_matrix[seq(1,12),4] <- rep(rmsp_vec[1], 12)
for(i in 1:length(county_vec)){
  x <- 12 * (i-1)
  lower_matrix[x + seq(1,12),4] <- rep(rmsp_vec[i], 12)
}

#convert to data frame
lower_matrix <- as.data.frame(lower_matrix)

#remove NAs
lower_matrix <- subset(lower_matrix, !is.na(LOWER))

#change RMSP to a factor
lower_matrix$RMSP <- as.factor(lower_matrix$RMSP)

#change UPPER to a number
lower_matrix$LOWER <- as.numeric(lower_matrix$LOWER)

#p = 0.598
rmsp_aov_mod <- aov(LOWER ~ RMSP, data = lower_matrix)

#assess normality, really not normal
shapiro.test((resid(rmsp_aov_mod)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  (resid(rmsp_aov_mod))
## W = 0.90779, p-value < 2.2e-16
```

```
plot(rmsp_aov_mod, which = c(1,2))
```






```
hist(resid(rmsp_aov_mod))
```


```
#NON-PARAMETRIC, p = 0.2159
lower_rmsp_np <- kruskal.test(LOWER ~ RMSP, data = lower_matrix)
```

Preliminary analysis

- Anova yielded very insignificant results (p = 0.598), but residuals were not normally distributed. A non-parametric test (Kruskall-Wallis) was then ran with the following null and alternative hypotheses:
  - H_O : RMSP does not influence the lower quartile value for age of onset.
  - H_A : RMSP influences the lower quartile value for age of onset.
- The KW test yielded a p-value of 0.2159. This leads to the failure to reject the null hypothesis that RMSP does not influence the lower quartile value for age of onset.

TO BE CONSIDERED:

- What is implied if the upper quartile is influenced by RMSP but the lower quartile is not?
  - The distribution of older individuals differs between rural and urban areas but not the distribution of younger individuals. This then lends itself to the question of how the the this distribution differs.
    - i.e. Are individuals being infected in urban areas older or younger than those infected in rural areas?
  - These questions can most likelt be answered using descriptive statistics.

Divide counties by RMSP

```
#list of each county
county_vec <- as.character(unique(data$COUNTY))

#are not in regional metropolitan area
```

```
rural <- subset(data, data$RMSP == 0)

#are in the regional metropolitan area
urban <- subset(data, data$RMSP == 1)
```

Find upper quartile for rural

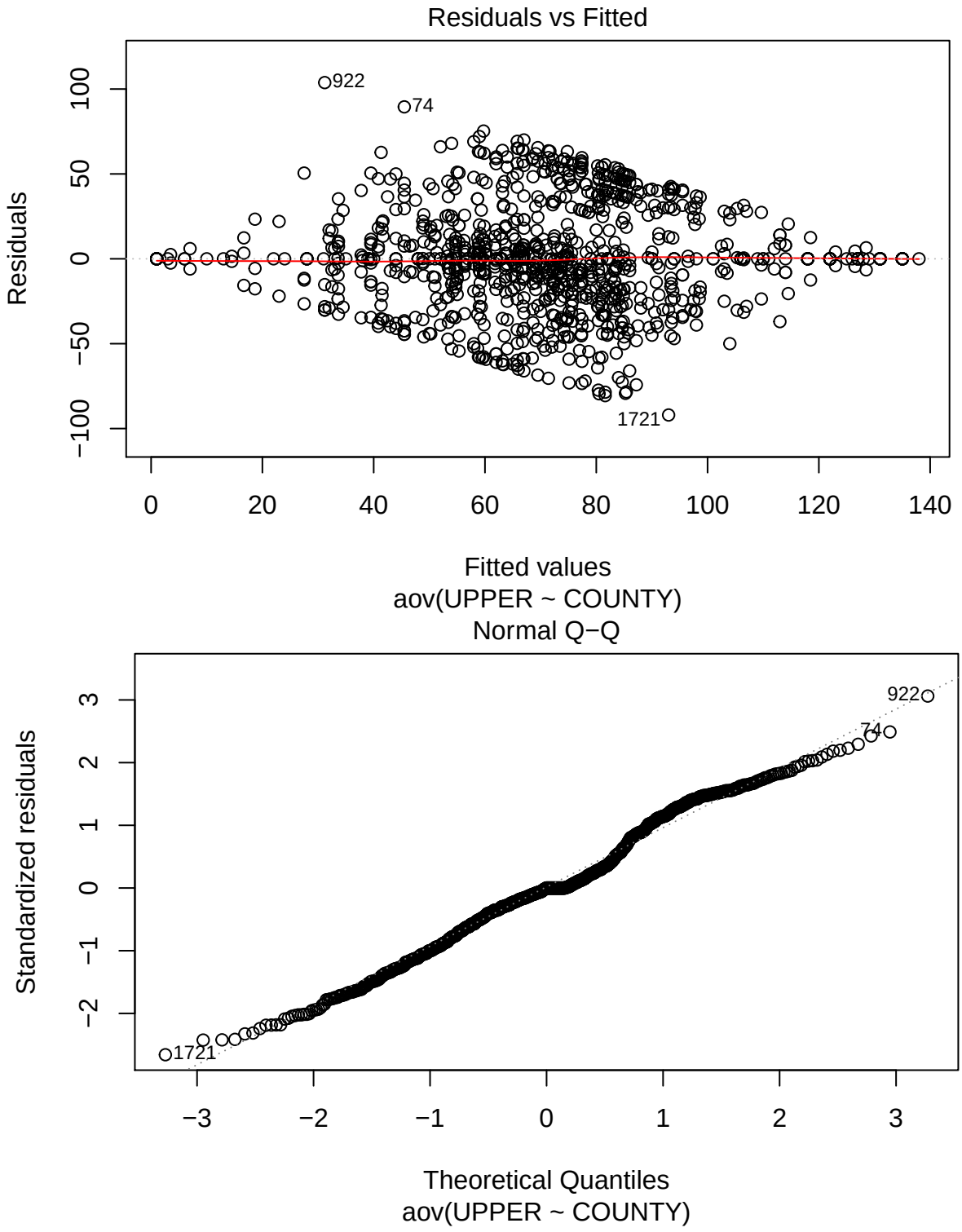
Run ANOVA for county/month: UPPER

```
upper_rural$UPPER <- as.numeric(upper_rural$UPPER)

####################################
#COUNTY
####################################
#p = 0.0322
upper_county_aov <- aov(UPPER ~ COUNTY, data = upper_rural)
#p = 6.897e-10, not normal
shapiro.test((resid(upper_county_aov)))
```

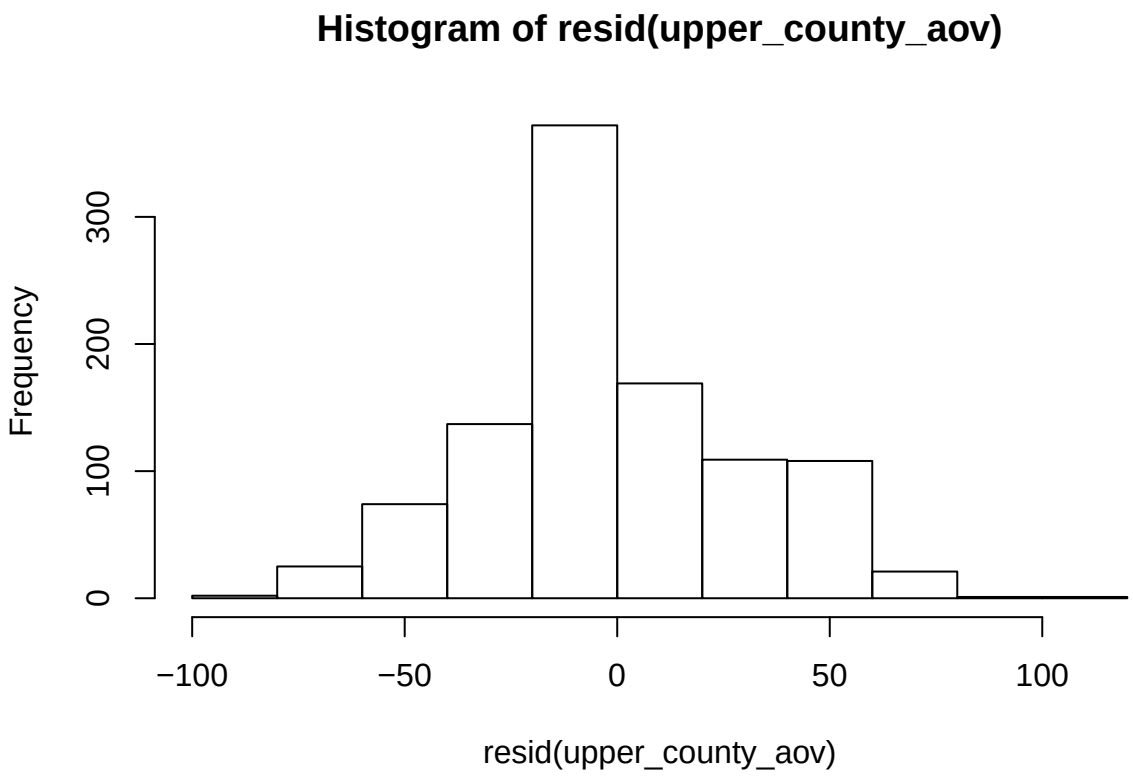
```
##
##  Shapiro-Wilk normality test
##
## data:  (resid(upper_county_aov))
## W = 0.98204, p-value = 6.897e-10
```

```
plot(upper_county_aov, which = c(1,2))
```

```
## Warning: not plotting observations with leverage one:
##   187, 197, 201, 219, 243, 269, 307, 308, 335, 336, 362, 387, 395, 413, 416, 425, 426, 436, 439, 440
```



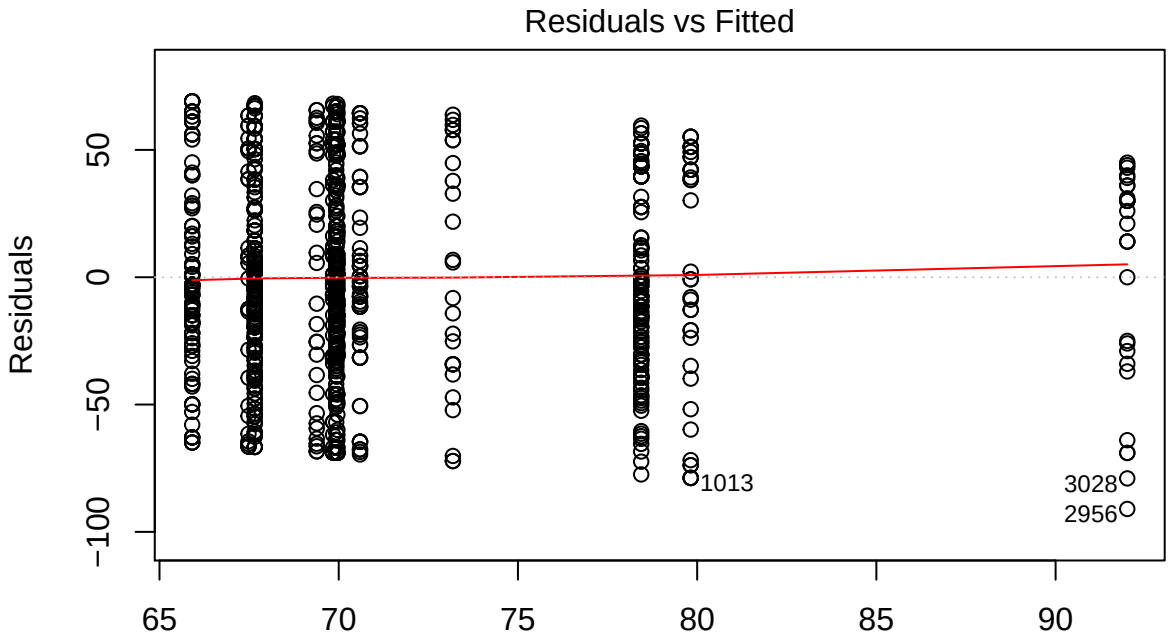

```
hist(resid(upper_county_aov))
```

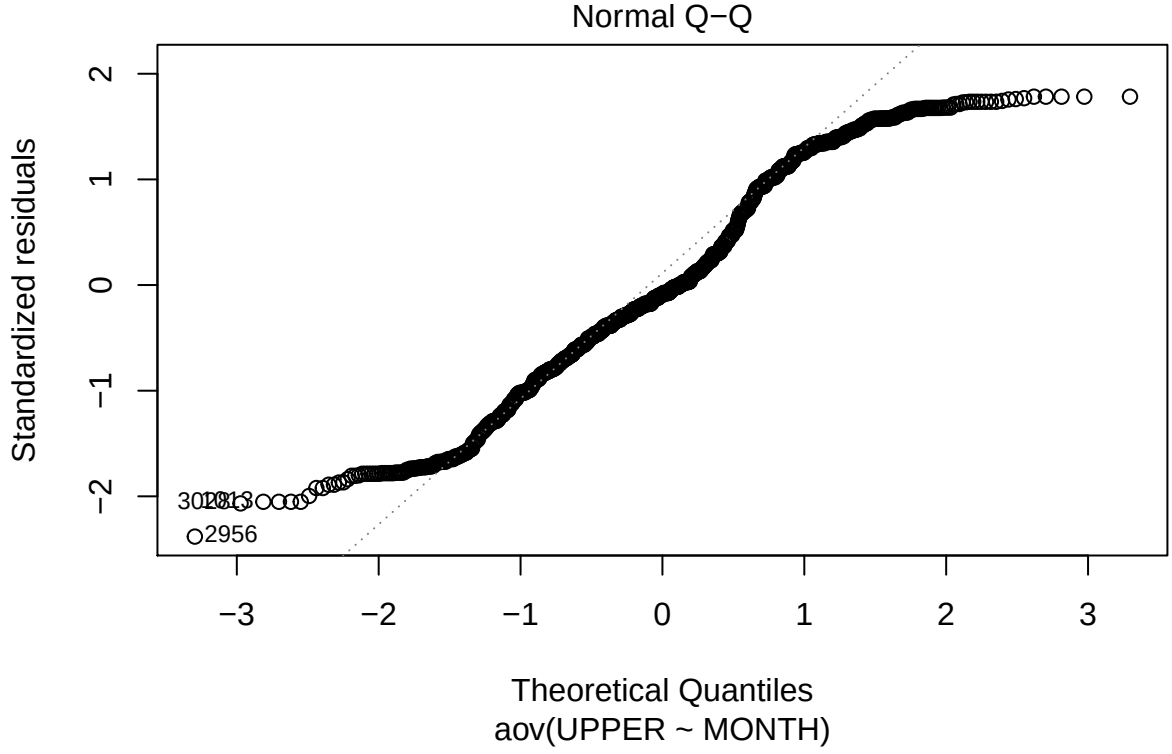



```
#########################
#COUNTY: NON-PARAMETRIC
#########################
#KW-test: p = 0.0747
upper_county_kw <- kruskal.test(UPPER ~ COUNTY, data = upper_rural)




###################################
#MONTH
###################################
#p = 0.0555
upper_month_aov <- aov(UPPER ~ MONTH, data = upper_rural)
#p = 2.242e-14, not normal
shapiro.test((resid(upper_month_aov)))
```

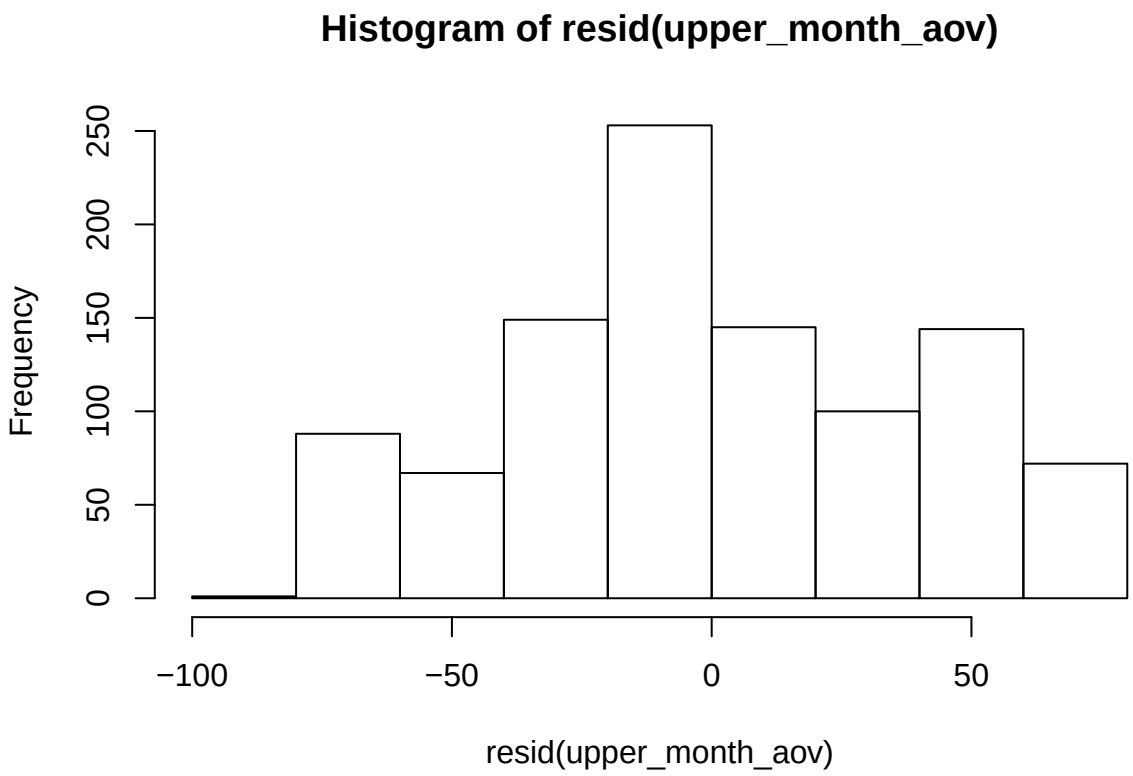
```
##
##  Shapiro-Wilk normality test
##
## data:  (resid(upper_month_aov))
## W = 0.96726, p-value = 2.242e-14
```

```
plot(upper_month_aov, which = c(1,2))
```






```
hist(resid(upper_month_aov))
```

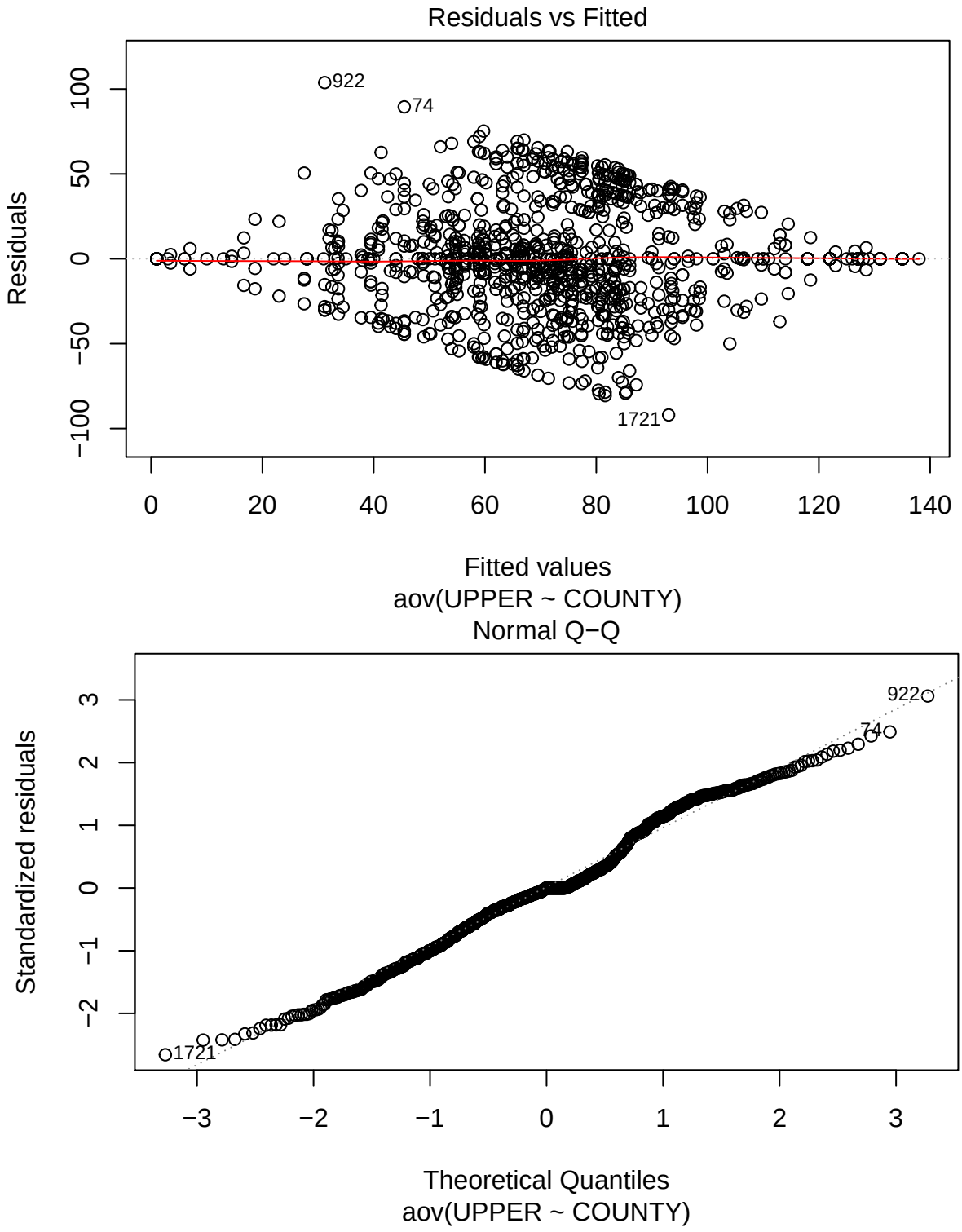



```
#########################
#MONTH: NON-PARAMETRIC
#########################
#KW-test: p = 0.07786
upper_month_kw <- kruskal.test(UPPER ~ MONTH, data = upper_rural)




###################################
#TWO-WAY
###################################
#p = 0.0322
upper_twoway_aov <- aov(UPPER ~ COUNTY * MONTH, data = upper_rural)
#p = 6.897e-10, not normal
shapiro.test((resid(upper_county_aov)))
```

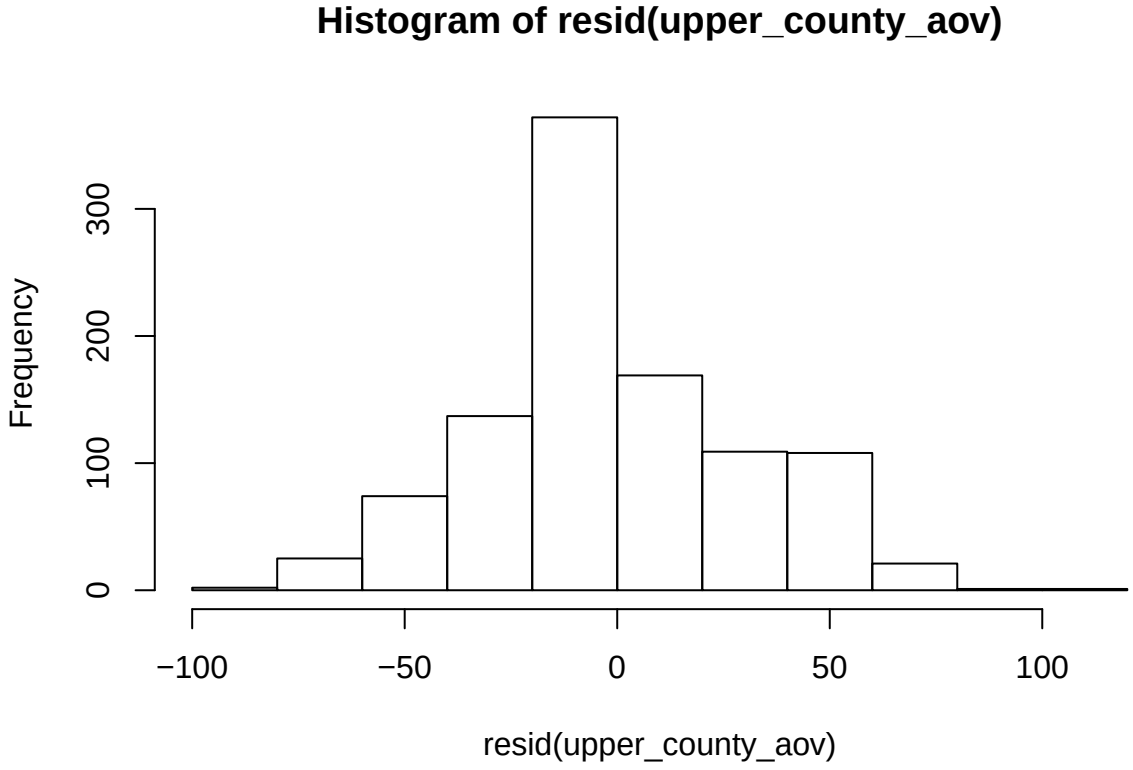
```
##
##  Shapiro-Wilk normality test
##
## data:  (resid(upper_county_aov))
## W = 0.98204, p-value = 6.897e-10
```

```
plot(upper_county_aov, which = c(1,2))
```

```
## Warning: not plotting observations with leverage one:
##   187, 197, 201, 219, 243, 269, 307, 308, 335, 336, 362, 387, 395, 413, 416, 425, 426, 436, 439, 440
```




```
hist(resid(upper_county_aov))
```



Preliminary results

- Tried many transformations, but could not normalize data. Attempted a KW test for month and county.
- Just realized that I should do a ANOVA with upper quartile and then the factor being RMSP. Need to use all data.