Stat 230 - Lab 8 - SOLUTION Two-way ANOVA

P.B. Matheson adapted from A.S. Wagaman

This lab will lead you through some data management practice and a two-way ANOVA example. You will be providing your own code. Be sure to work with those around you to tackle the problems, and ask for assistance as needed!

Bike Rentals - Data and Background from UC-I Machine Learning Repository

Two-Way Additive Model This dataset contains the hourly and daily count of rental bikes between the years of 2011 and 2012 in Washington D.C. based on data from the bikeshare system with corresponding weather and seasonal information.

A little bit of background from UCI's page on the data set: Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automated and recorded. Through these systems, a user is able to easily rent a bike from a particular position and return it at another position. Currently, there are over 500 bike-sharing programs around the world. There is great interest in understanding these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the researcher. As opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns a bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring of these data.

We will be using this data set to explore 2-way ANOVA and get you to practice data management skills.

```
bike <- read.csv("https://pmatheson.people.amherst.edu/stat230/bikerental.csv")
```

The variables included in the dataset are:

1) season : season (1:spring, 2:summer, 3:fall, 4:winter)

2) yr : year (0:2011, 1:2012)

3) mnth : month (1 to 12)

4) hr : hour (0 to 23)

5) holiday : whether day is holiday (1) or not (0)

6) weekday : day of the week - coded 0 (Sunday) to 6 (Saturday)

7) workingday : if day is neither weekend nor holiday is 1, otherwise is 0.

8) weathersit : weather situation with the following levels:

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

9) temp : Normalized temperature in Celsius. The values are as a fraction of 41 (max)

10) atemp: Normalized feeling temperature in Celsius. The values are as a fraction of 50 (max)

11) hum: Normalized humidity. The values are as a fraction of 100 (max)

12) windspeed: Normalized wind speed. The values are as a fraction of 67 (max)

13) casual: count of casual users

14) registered: count of registered users

15) cnt: count of total rental bikes including both casual and registered

1. Are there any indicator variables present in the data set? If so, which variables?

SOLUTION: Yes. Year, holiday, and workingday can all be considered indicator variables.

2. Are there any numerically coded categorical variables that are not indicator variables in the data set? If so, which variables?

SOLUTION: Yes. Weathersit has four levels and weekday has 7 levels. Both are categorical but coded numerically.

We are interested in examining differences in average numbers of casual, registered, and total users across several variables including holiday, weekday, workingday, and weathersit.

3. We want to modify the data set in R to treat these four variables appropriately. The first two are provided for you, but you need to set up workingday and weathersit on your own.

IMPORTANT: call `weathersit` something different… like `weathersitcat`. It is a good idea to save the variable with a new name, in case you want to consider it as a quantitative variable or categorical for some reason. There are other reasons not to overwrite, as you'll see below.

```r
bike <- mutate(bike, holiday = ifelse(holiday == 1, "Holiday", "Non-Holiday"))
bike <- mutate(bike, weekday = cut(weekday, breaks = c(-.5, .5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5),
    labels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"),
    include.lowest = TRUE))
```

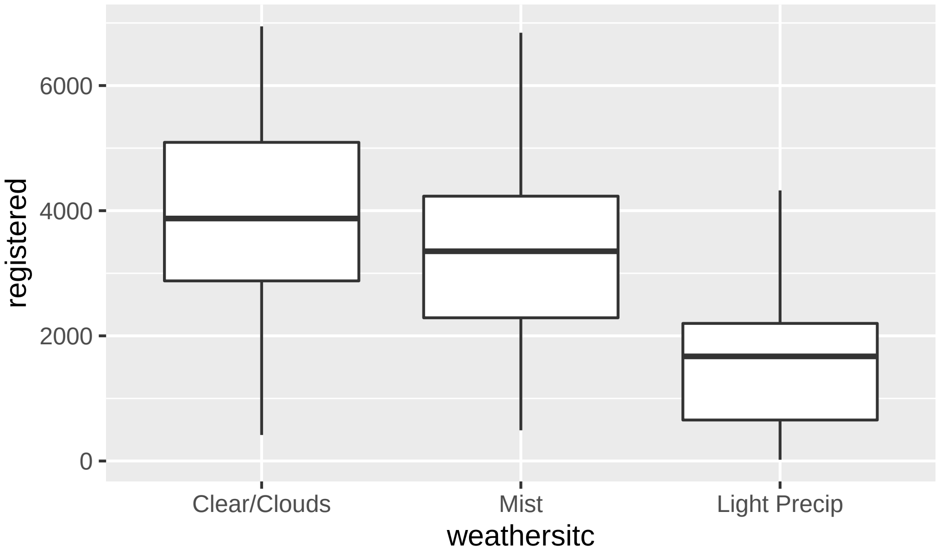
```r
# YOUR TURN - remember not to overwrite weathersit
bike <- mutate(bike, workingday = ifelse(workingday == 1, "Work Day", "Non-Work Day"))
bike <- mutate(bike, weathersitc = cut(weathersit, breaks = c(0.5, 1.5, 2.5, 3.5, 4.5),
        labels = c("Clear/Clouds", "Mist", "Light Precip", "Heavy Precip"),
        include.lowest = TRUE))
```

Examine differences in the average number of registered users depending on the weather situation and whether or not it is a holiday.

4. Obtain appropriate graphs and descriptive statistics in order to describe how the number of registered users depends on weather situation.

SOLUTION:

```
gf_boxplot(registered ~ weathersitc, data = bike)
```



```
favstats(registered ~ weathersitc, data = bike)
```

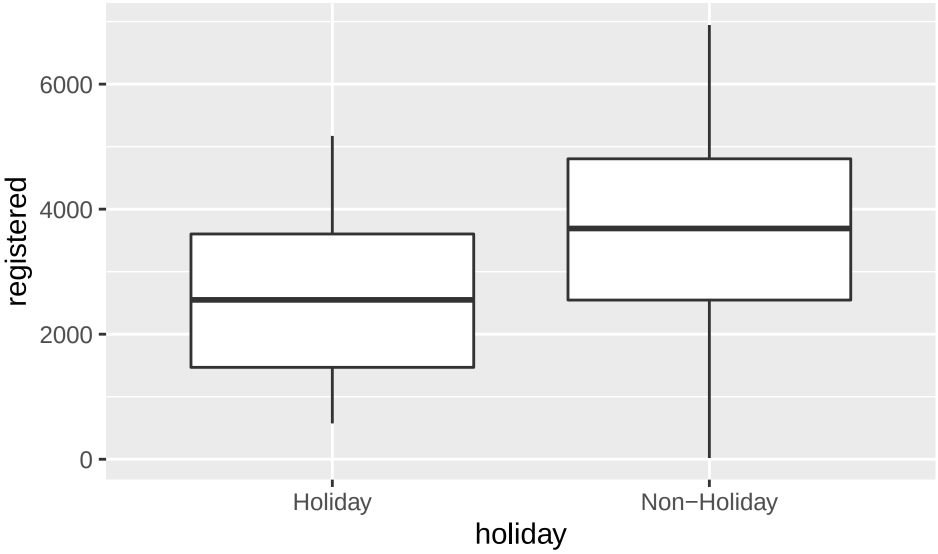
```
##    weathersitc min     Q1 median   Q3  max    mean      sd   n missing
## 1 Clear/Clouds 416 2878.5   3875 5092 6946 3912.76 1533.35 463       0
## 2         Mist 491 2288.5   3352 4232 6844 3348.51 1463.57 247       0
## 3 Light Precip  20  655.0   1672 2199 4324 1617.81 1068.29  21       0
## 4 Heavy Precip  NA     NA     NA   NA   NA     NaN      NA   0       0
```

It looks like the number of registered users decreases as the weather gets worse. We also note there are no heavy precipitation days. There are only 21 days with light precipitation.

5. Obtain appropriate graphs and descriptive statistics in order to describe how the number of registered users depends on whether or not it is a holiday.

SOLUTION:

```
gf_boxplot(registered ~ holiday, data = bike)
```



```
favstats(registered ~ holiday, data = bike)
```

```
##       holiday min   Q1 median     Q3  max    mean      sd   n missing
## 1     Holiday 573 1470   2549 3603.0 5172 2670.29 1492.86  21       0
## 2 Non-Holiday  20 2546   3691 4805.5 6946 3685.33 1553.70 710       0
```

We see that there appear to be more registered users on non-holidays than on holidays, though there are only 21 observations for holidays.

6. Obtain appropriate output to describe how weather situation and whether or not it is a holiday are related to one another.

SOLUTION:

```
tally(~ weathersitc + holiday, data = bike)
```

```
##                holiday
## weathersitc     Holiday Non-Holiday
##   Clear/Clouds      15         448
##   Mist               6         241
##   Light Precip       0          21
##   Heavy Precip       0           0
```

We see that there are no holidays with light precipitation, and no heavy precipitation days.

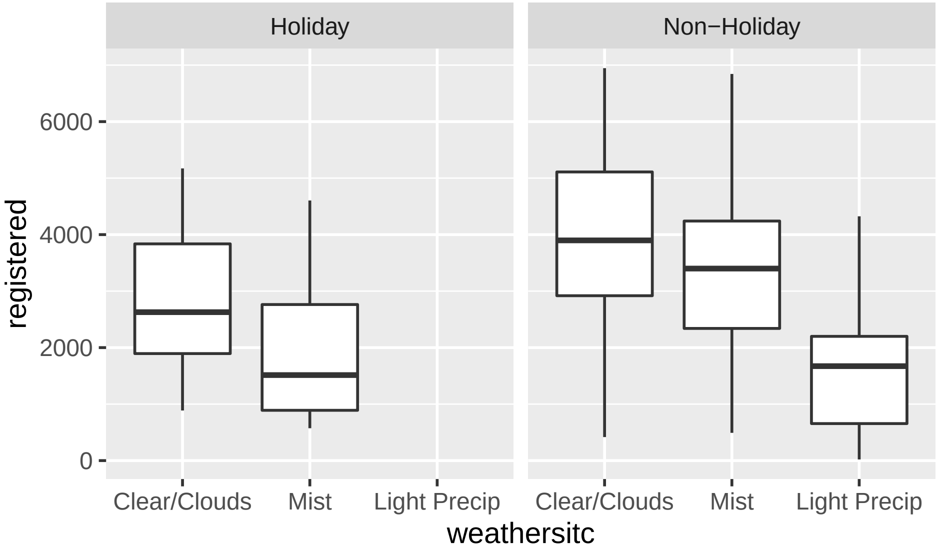
Note: You may discover something along the way here that may want to make you adjust the levels of variables previously defined. You should make appropriate adjustments before continuing. This is why you save a mutated variable with a different name. If you didn't, it won't let you adjust the command. You'd need to reload the data set and execute each transformation line again.

```
bike <- mutate(bike, weathersitc = cut(weathersit, breaks = c(0.5, 1.5, 2.5, 3.5),
                                       labels = c("Clear/Clouds", "Mist", "Light Precip"), include.lowe
```

7. Obtain appropriate graphs and descriptive statistics in order to describe how the number of registered users depends on BOTH weather situation and whether or not it is a holiday.

SOLUTION:

```
gf_boxplot(registered ~ weathersitc | holiday, data = bike)
```



```
favstats(registered ~ weathersitc + holiday, data = bike)
```

```
##        weathersitc.holiday min      Q1 median      Q3  max    mean      sd   n
## 1      Clear/Clouds.Holiday 887 1894.00 2627.0 3836.00 5172 2934.07 1431.83  15
## 2              Mist.Holiday 573  890.25 1513.5 2762.25 4604 2010.83 1563.18   6
## 3     Light Precip.Holiday  NA      NA     NA      NA   NA     NaN      NA   0
## 4 Clear/Clouds.Non-Holiday 416 2918.25 3898.0 5109.00 6946 3945.52 1527.29 448
## 5         Mist.Non-Holiday 491 2339.00 3399.0 4240.00 6844 3381.81 1448.73 241
## 6 Light Precip.Non-Holiday  20  655.00 1672.0 2199.00 4324 1617.81 1068.29  21
##   missing
## 1       0
## 2       0
## 3       0
## 4       0
## 5       0
## 6       0
```

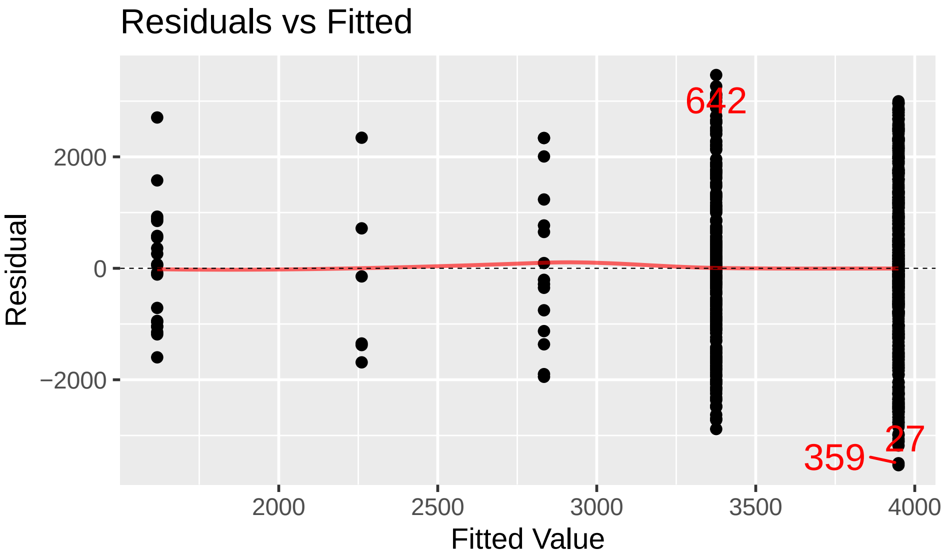
It looks like values for registered are higher on non-holidays than the corresponding holidays based on weather situation. We do see issues with really small sample sizes for holidays. We also see many of these distributions have long upper tails. You might also do an interaction plot here, which we'll come back to in a few minutes.

8. Fit a two-way additive ANOVA model to determine whether or not there are differences in the average number of registered users depending on the weather situation and whether or not it is a holiday. Check the conditions for the model.

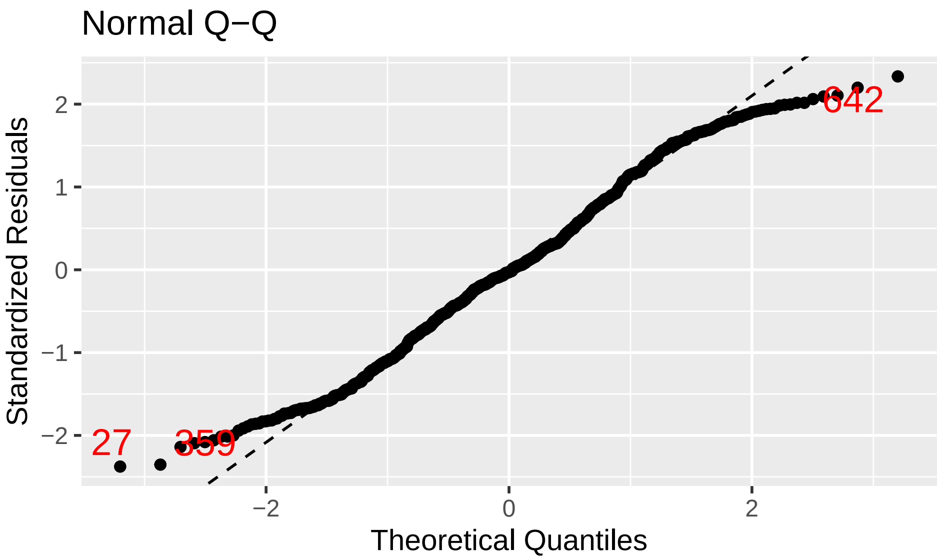
SOLUTION:

```
mod <- lm(registered ~ holiday + weathersitc, data = bike)
mplot(mod, which = 1)
```

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



```
mplot(mod, which = 2)
```



The residuals vs. fitted plot shows a possible problem with differences in spread. However, the ratio rule for SDs is not violated: 1563/1068 < 2, so we are probably satisfied with the constant variance condition. The QQ plot shows some issues in both tails, so there are issues with normality. For independence and randomization, we have to assume this is a representative sample of the bike rental data, and we are uncertain of the independence of the errors.

9. Why are there only 5 groups in the residuals vs. fitted plot?

SOLUTION: With 2*3 levels, we'd expect 6 groups in the plot, but we know there are no light precipitation holidays, therefore we only see 5 groups in the plot.

10. Try a transformation to deal with problems with the conditions. Does a model with a transformed response seem more appropriate? Be careful, you can go too far down the ladder here.

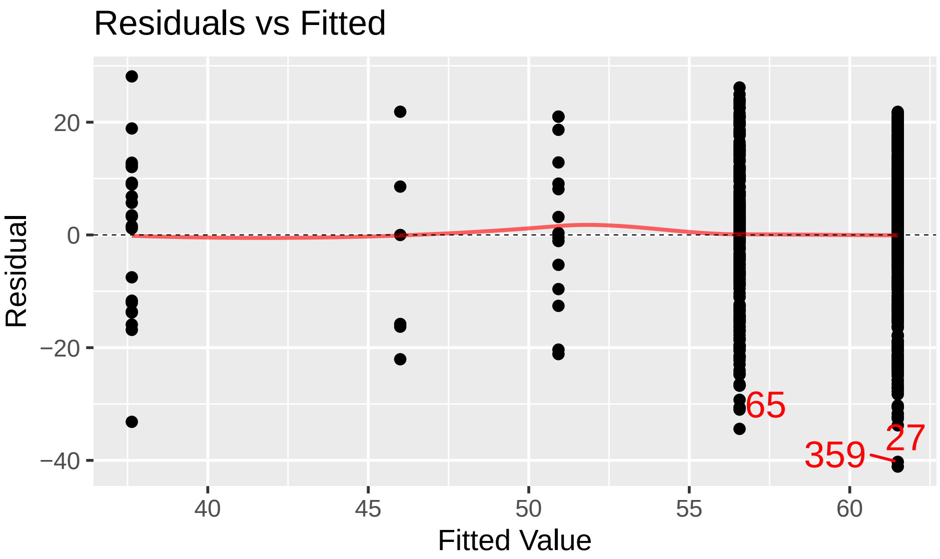
SOLUTION:

```
bike <- mutate(bike, sqrtregister = sqrt(registered))
mod2 <- lm(sqrtregister ~ holiday + weathersitc, data = bike)
favstats(sqrtregister ~ weathersitc + holiday, data = bike)
```

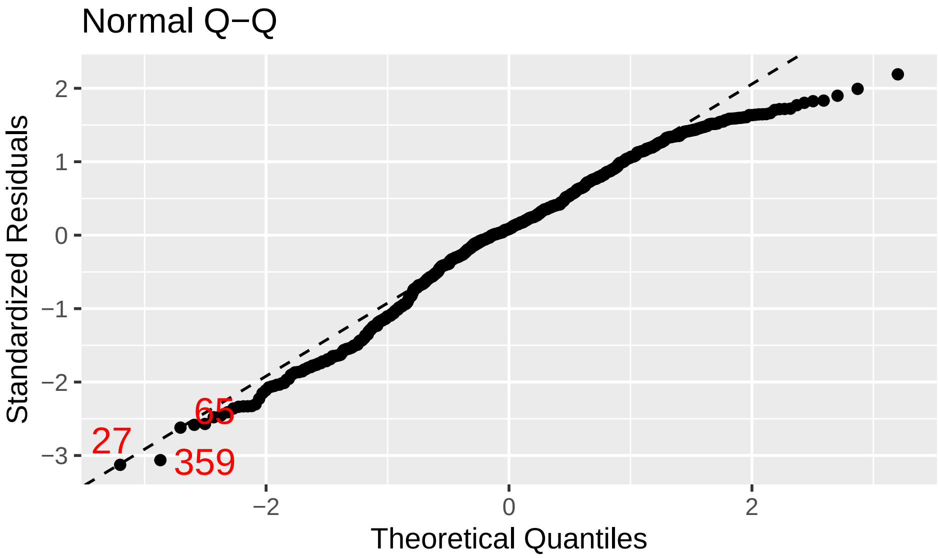
```
##        weathersitc.holiday      min      Q1  median      Q3     max    mean
## 1     Clear/Clouds.Holiday 29.78255 43.4669 51.2543 61.9069 71.9166 52.5063
## 2             Mist.Holiday 23.93742 29.8363 38.0942 52.4256 67.8528 42.0442
## 3     Light Precip.Holiday       NA      NA      NA      NA      NA     NaN
## 4 Clear/Clouds.Non-Holiday 20.39608 54.0208 62.4340 71.4773 83.3427 61.4439
## 5         Mist.Non-Holiday 22.15852 48.3632 58.3009 65.1153 82.7285 56.6659
## 6 Light Precip.Non-Holiday  4.47214 25.5930 40.8901 46.8935 65.7571 37.6336
##        sd   n missing
## 1 13.7773  15       0
## 2 17.0805   6       0
## 3      NA   0       0
## 4 13.0597 448       0
## 5 13.0957 241       0
## 6 14.5464  21       0
```

```
mplot(mod2, which = 1)
```

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



```
mplot(mod2, which = 2)
```



Using a log makes the normal plot worse. Using a sqrt, we can see we still don't have major issues with the constant variance condition (17.1/13.1 for the ratio), and we've corrected issues in the LOWER tail of the QQ plot, but not the upper tail. This is likely due to the strong right-skewness of several of the groups. However, going to a log makes it worse in the other direction, so stop with the sqrt.

11. Using appropriate output, determine whether or not there are differences in the average number of (possibly transformed) registered users depending on the weather situation and whether or not it is a holiday. Be sure to state which test statistics and p-values you are using. (You should be able to state the hypotheses in words or notation too!)

SOLUTION:

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: sqrtregister
##              Df Sum Sq Mean Sq F value  Pr(>F)
## holiday       1   1880    1880   10.86 0.00103 **
## weathersitc   2  13900    6950   40.14 < 2e-16 ***
## Residuals   727 125881     173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the ANOVA output, we see that we do have evidence of differences on average in the sqrt of the number of registered users on holidays vs. non-holidays based on an F of 10.9 and p-value of 0.001. We also see that we do have evidence of differences on average in the sqrt of the number of registered users for weather situation based on an F of 40.1 and p-value as small as R can report.

12. If you determined that differences were present, perform appropriate procedures to identify the differences.

SOLUTION:

```
# Hint: TukeyHSD is easy to apply!
TukeyHSD(mod2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $holiday
##                        diff    lwr     upr    p adj
## Non-Holiday-Holiday 9.60071 3.8806 15.3208 0.001032
##
## $weathersitc
##                                diff       lwr      upr p adj
## Mist-Clear/Clouds          -4.92139  -7.35638  -2.4864 7e-06
## Light Precip-Clear/Clouds -23.83176 -30.72669 -16.9368 0e+00
## Light Precip-Mist         -18.91037 -25.93489 -11.8858 0e+00
```

```
# PostHocTests are also possible!
```

The TukeyHSD output shows that ALL the differences examined are significant. That is to say, holiday vs. non-holiday sqrt registered usage is different, and so is tge sqrt registered users for all three possible differences in weather levels.

13. Obtain the regression model that is equivalent to the ANOVA output. Interpret the intercept in the model.

SOLUTION:

```
msummary(mod2)
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                50.93       2.89   17.64  < 2e-16 ***
## holidayNon-Holiday         10.57       2.92    3.63  0.00031 ***
## weathersitcMist            -4.93       1.04   -4.75  2.4e-06 ***
## weathersitcLight Precip   -23.86       2.94   -8.12  1.9e-15 ***
##
## Residual standard error: 13.2 on 727 degrees of freedom
## Multiple R-squared:  0.111,  Adjusted R-squared:  0.108
## F-statistic: 30.4 on 3 and 727 DF,  p-value: <2e-16
```

The intercept is 50.93. This is the average value for sqrt registered users on holidays at the lowest weather setting (clear-clouds).

14. Interpret the coefficient of holidayNon-Holiday in the model.

SOLUTION: The coefficient is 10.57. This is the increase in the average value for sqrt registered users on non-holidays relative to holidays when weather conditions have been taken into account.

15. Interpret the coefficient of the worst weather level in the model.

SOLUTION: The coefficient is -23.86. This is the decrease in the average value for sqrt registered users on light precipitation days relative to clear/slightly cloudy days when holiday status has been taken into account.

Be sure you are comfortable explaining how the ANOVA is fit as a regression.