Stat 230 - Lab 9 Two-way ANOVA with Interaction - Solution

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

This lab continues the bike example (with data wrangling part removed) to allow you to practice two-way ANOVA with interaction.

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 years 2011 and 2012 in the Capital (Washington D.C.) bikeshare system with the 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 automatic. 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 research. 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.

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

The variables included in the data 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

```
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 =
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), labels = c("Clear/Cl
```

In the previous lab, we examined differences in the average number of registered users depending on the weather situation and whether or not it is a holiday. A re-expression of the response was considered there.

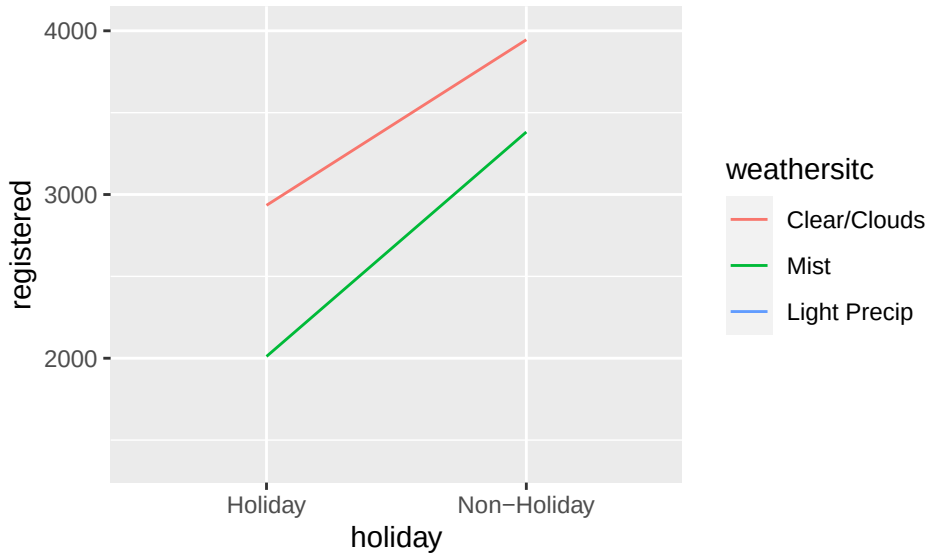
Two-Way Model with Interaction Now we need to assess an interaction component in the model we just considered, but with the original response. Note that usually, a statistician would start by determining if interaction was needed, rather than fitting the additive model first.

1. Make an interaction plot that generates lines for the weather levels. Does it appear that there is interaction between holiday and weather level? Why are there only 2 lines?

SOLUTION:

```
#gf_line(Response ~ Var1, color = ~ Var2, data = bike, group = ~ Var2, stat = "summary")
gf_line(registered ~ holiday, color = ~ weathersitc, data = bike, group = ~ weathersitc, stat = "summary
```

```
## No summary function supplied, defaulting to `mean_se()`
```




```
#adjust line above and be sure you get lines for the correct variable! Lines go with Var2!
```

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

```
##                weathersitc min      Q1 median      Q3  max    mean      sd   n
## 1     Holiday.Clear/Clouds 887 1894.00 2627.0 3836.00 5172 2934.07 1431.83  15
## 2 Non-Holiday.Clear/Clouds 416 2918.25 3898.0 5109.00 6946 3945.52 1527.29 448
## 3             Holiday.Mist 573  890.25 1513.5 2762.25 4604 2010.83 1563.18   6
## 4         Non-Holiday.Mist 491 2339.00 3399.0 4240.00 6844 3381.81 1448.73 241
## 5     Holiday.Light Precip  NA      NA     NA      NA   NA     NaN      NA   0
## 6 Non-Holiday.Light Precip  20  655.00 1672.0 2199.00 4324 1617.81 1068.29  21
## 7             Clear/Clouds 416 2878.50 3875.0 5092.00 6946 3912.76 1533.35 463
## 8                    Mist 491 2288.50 3352.0 4232.00 6844 3348.51 1463.57 247
## 9             Light Precip  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
## 7       0
## 8       0
## 9       0
```

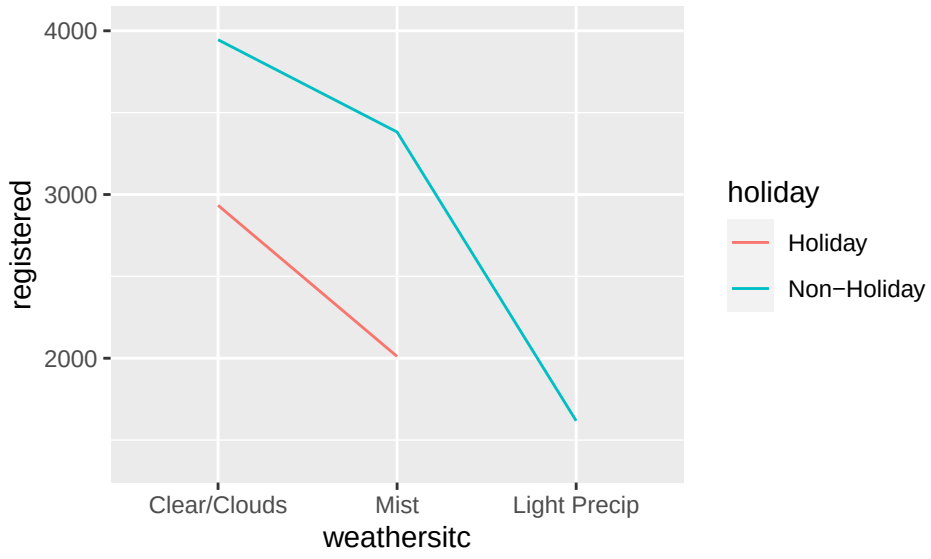
There may be some interaction, as the lines are not parallel. There are only two lines because there are no holidays with light precipitation to plot, so light precip is not plotted at all.

2. Make an interaction plot that generates lines for holidays and non-holidays. Does it appear that there is interaction between holiday and weather level?

SOLUTION:

```
gf_line(registered ~ weathersitc, color = ~ holiday, data = bike, group = ~ holiday, stat = "summary",
```

```
## No summary function supplied, defaulting to `mean_se()`
```



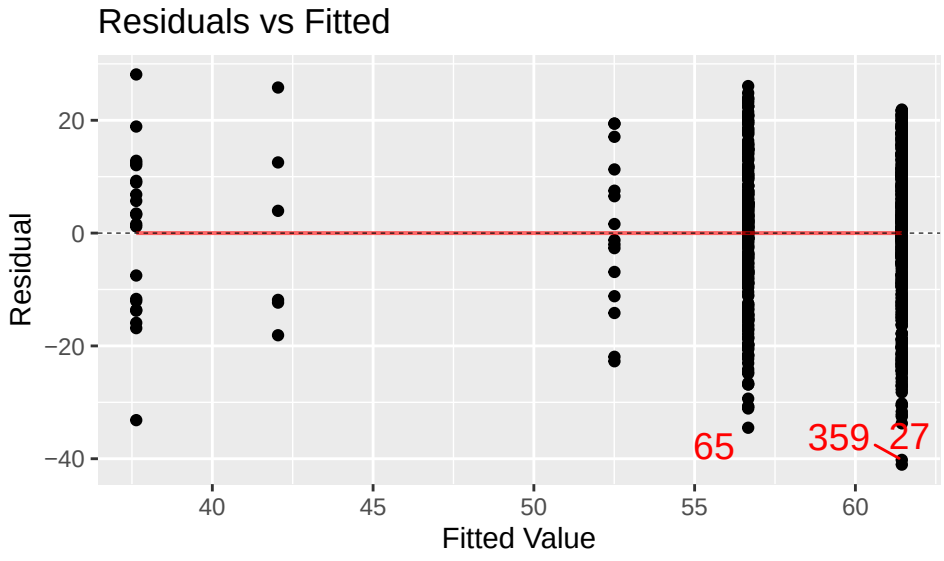
Yes, it looks like an interaction may be present. The issue is whether or not the effect is large enough that it is significant. The lines are clearly not parallel.

3. Fit an appropriate model with interaction. Check conditions and re-express the response as necessary until you believe conditions are met.

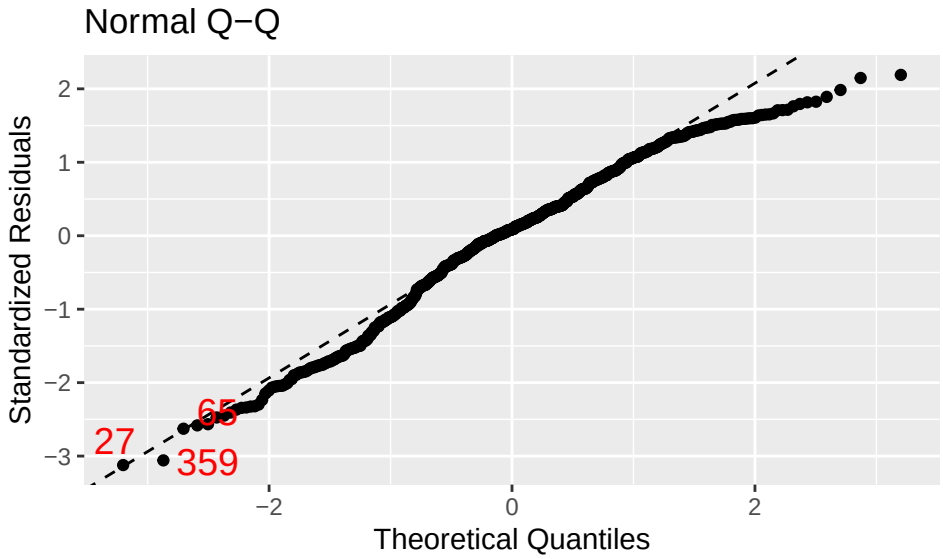
SOLUTION:

```
bike <- mutate(bike, sqrtregister = sqrt(registered))
mod <- lm(sqrtregister ~ holiday * weathersitc, data = bike)
mplot(mod, which = 1)
```

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




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



I chose to use a sqrt transform of the response again to deal with issues of non-normality. There are still slight issues in the upper tail but the log transform causes more issues than it fixes.

4. Determine whether or not the interaction between weather and holiday is significant. State your test statistic, p-value, and conclusion.

SOLUTION:

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: sqrtregister
##                       Df Sum Sq Mean Sq F value  Pr(>F)
## holiday                1   1880    1880  10.854 0.00103 **
## weathersitc            2  13900    6950  40.126 < 2e-16 ***
## holiday:weathersitc    1    135     135   0.778 0.37799
## Residuals            726 125747     173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction F test statistic is 0.78, with a p-value of 0.378. So, it is not significant. We could refit the model without it and use the results we previously obtained with the additive model. We do not have evidence of a significant interaction between holiday and weather situation in terms of the average value of sqrt registered users.

Choose Your Own Two-Way Model Now we want to investigate the behavior of *casual* bike users. You might expect some differences between how registered and casual bike users behave.

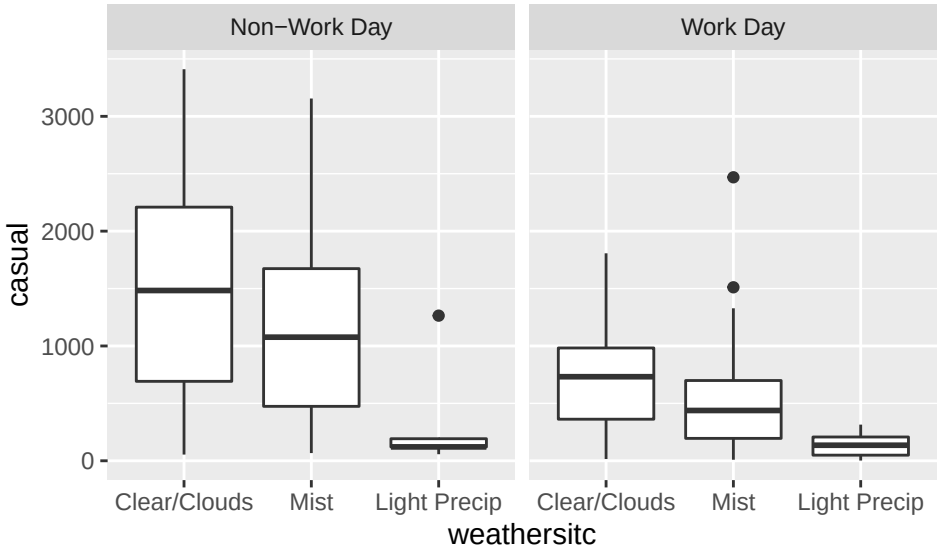
5. What are some differences you might expect in the behavior of casual users compared to registered users?

SOLUTION: Casual users might be more sensitive to weather conditions and might be more likely to use the bikes on holidays than non-holidays, or non-working days compared to working days.

6. Investigate appropriate graphs and descriptive statistics to shed light on how the number of casual users is related to weather situation and whether or not it is a workday. Do you see any possible concerns with performing a two-way ANOVA?

SOLUTION:

```
gf_boxplot(casual ~ weathersitc | workingday, data = bike)
```




```
favstats(casual ~ weathersitc + workingday, data = bike)
```

```
##      weathersitc.workingday min  Q1 median      Q3  max     mean       sd   n
## 1 Clear/Clouds.Non-Work Day  54 692 1483.0 2209.00 3410 1483.776 873.5824 156
## 2         Mist.Non-Work Day  67 474 1076.5 1673.50 3155 1192.986 821.5411  70
## 3 Light Precip.Non-Work Day  57 120  121.0  192.00 1264  350.800 512.7248   5
## 4     Clear/Clouds.Work Day  15 362  732.0  982.50 1807  699.925 382.8250 307
## 5             Mist.Work Day   9 195  438.0  699.00 2469  487.384 359.4374 177
## 6     Light Precip.Work Day   2  49  135.5  207.25  315  133.812  93.5402  16
##   missing
## 1       0
## 2       0
## 3       0
## 4       0
## 5       0
## 6       0
```

It looks like casual users are out and about more on bikes when it is a non-work day than a work day. The values also decrease as the weather gets worse. Serious issues with fitting an ANOVA due to drastic differences in spread. The ratio of SDs is 873.6/93.5 which is much larger than 2. It's obvious from the IQRs on the boxplot as well.

7. Assess with plots whether or not you'd like to fit an interaction term in the model.

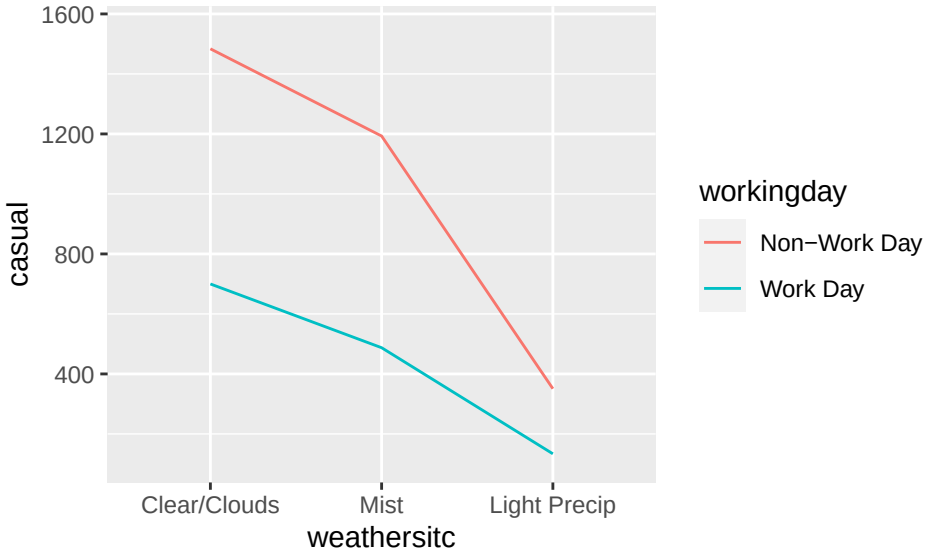
SOLUTION:

```
gf_line(casual ~ workingday, color = ~ weathersitc, data = bike, group = ~ weathersitc, stat = "summary
```

```
## No summary function supplied, defaulting to `mean_se()`
```


```
gf_line(casual ~ weathersitc, color = ~ workingday, data = bike, group = ~ workingday, stat = "summary"
```

```
## No summary function supplied, defaulting to `mean_se()`
```



It definitely looks like we want an interaction term if we can solve the problem of nonconstant spread. The lines get closer together on work days, and the size of the difference between work and non-work days for light precipitation in particular appears to be different than the corresponding differences in clear/clouds and mist settings.

8. Fit an appropriate model to determine whether there are differences in the average number of (possibly re-expressed) casual users depending on the weather situation and whether or not it is a workday, building in interaction if you feel it is necessary. Be sure you check your model conditions.

SOLUTION:

```
bike <- mutate(bike, sqrtcasual = sqrt(casual), logcasual = log(casual))
mod2 <- lm(sqrtcasual ~ workingday * weathersitc, data = bike)
favstats(sqrtcasual ~ weathersitc + workingday, data = bike)
```

```
##      weathersitc.workingday     min       Q1  median      Q3     max    mean
## 1 Clear/Clouds.Non-Work Day 7.34847 26.30581 38.5092 46.9999 58.3952 36.3854
```

```
## 2         Mist.Non-Work Day 8.18535 21.77140 32.8034 40.9077 56.1694 32.0677
## 3 Light Precip.Non-Work Day 7.54983 10.95445 11.0000 13.8564 35.5528 15.7827
## 4     Clear/Clouds.Work Day 3.87298 19.02630 27.0555 31.3449 42.5088 25.1726
## 5             Mist.Work Day 3.00000 13.96424 20.9284 26.4386 49.6890 20.5631
## 6     Light Precip.Work Day 1.41421  6.99888 11.6333 14.3949 17.7482 10.6134
##         sd   n missing
## 1 12.68508 156       0
## 2 12.92418  70       0
## 3 11.27534   5       0
## 4  8.15379 307       0
## 5  8.05654 177       0
## 6  4.75174  16       0
```

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

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

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 25.173
```

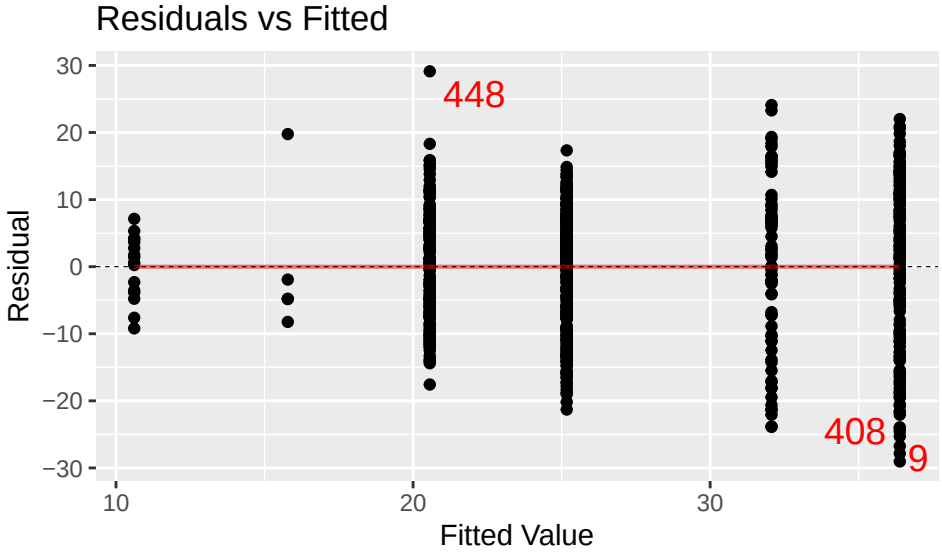
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 6.8951
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 2.1436e-15
```

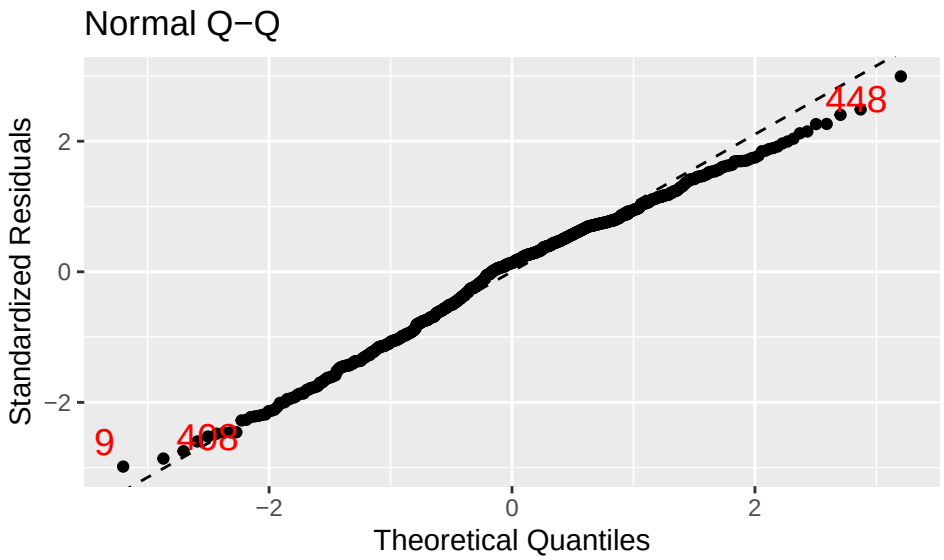
```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## 25.173
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 6.8951
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 2.1436e-15
```



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



If we use a sqrt transform, we fix issues with normality but still violate the rule of 2, but not as badly as before. If we go to a log transform, we fix the issues with spread but add major issues with normality. So I'm using a sqrt xform.

9. If you added an interaction term, is it significant? If you find it is not, refit your model without it.

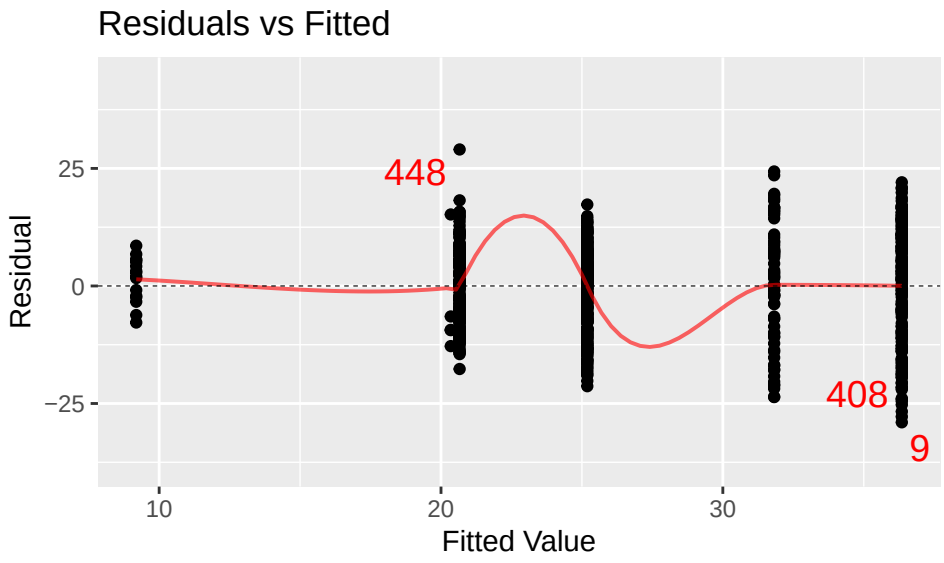
SOLUTION:

```
anova(mod2)
```

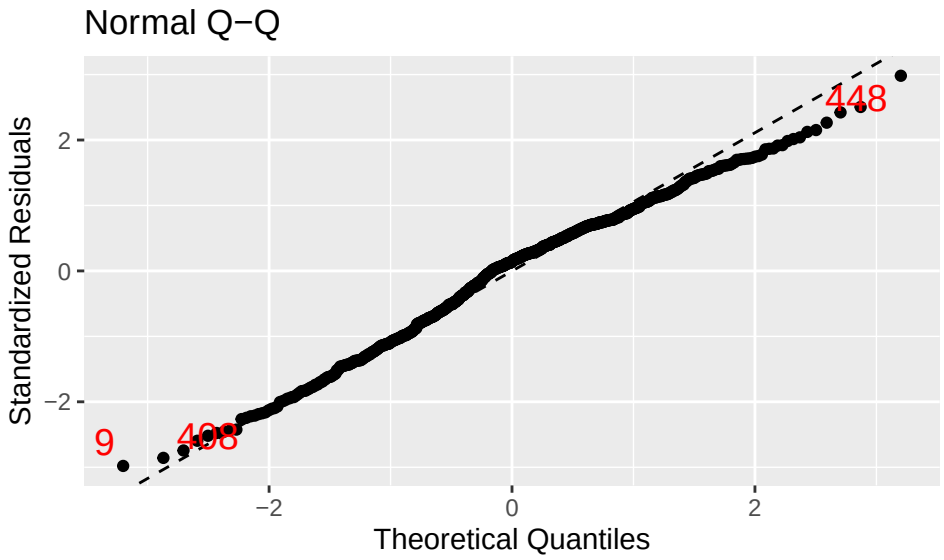
```
## Analysis of Variance Table
##
## Response: sqrtcasual
##                          Df Sum Sq Mean Sq F value Pr(>F)
## workingday                1  21100   21100  221.44 <2e-16 ***
## weathersitc               2   7526    3763   39.49 <2e-16 ***
## workingday:weathersitc    2    143      71    0.75  0.473
## Residuals               725  69082      95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(sqrtcasual ~ workingday + weathersitc, data = bike)
mplot(mod2, which = 1)
```

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




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



10. Interpret the main effects, if appropriate. If not appropriate, explain why not.

SOLUTION:

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: sqrtcasual
##              Df Sum Sq Mean Sq F value Pr(>F)
## workingday    1  21100   21100  221.60 <2e-16 ***
## weathersitc   2   7526    3763   39.52 <2e-16 ***
## Residuals   727  69225      95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
msummary(mod2)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                36.350      0.687   52.93  < 2e-16 ***
## workingdayWork Day        -11.160      0.778  -14.35  < 2e-16 ***
## weathersitcMist            -4.530      0.770   -5.88  6.2e-09 ***
## weathersitcLight Precip   -16.003      2.178   -7.35  5.5e-13 ***
##
## Residual standard error: 9.76 on 727 degrees of freedom
## Multiple R-squared:  0.293,  Adjusted R-squared:  0.29
## F-statistic:  100 on 3 and 727 DF,  p-value: <2e-16
```

We have evidence that there are differences in sqrt casual users between workingday and non-working day averages as well as differences in sqrt casual users between levels of weather. The regression output suggests that the numbers of casual users decrease on average on work days compared to non-work days, and that the numbers decrease as the weather gets worse. Both of which make sense. We still don't know if the differences are significant though (the regression p-values suggest they are).

11. If differences are present, use appropriate procedures to identify them.

SOLUTION:

```
TukeyHSD(mod2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $workingday
##                             diff      lwr      upr p adj
## Work Day-Non-Work Day -11.5561 -13.0802 -10.0321     0
##
## $weathersitc
##                                  diff       lwr       upr p adj
## Mist-Clear/Clouds           -4.50835  -6.31406  -2.70264 0e+00
## Light Precip-Clear/Clouds -15.96414 -21.07719 -10.85109 0e+00
## Light Precip-Mist         -11.45579 -16.66494  -6.24664 1e-06
```

All possible differences in average square root of the number of casual users by workingday and weather type are significant.