SOLUTION LAB 4 - Stat 230 - MLR I (Intro) - 3.1-3.3

P.B. Matheson adapted from A. Wagaman

Introduction to Multiple Linear Regression

In order to practice fitting multiple linear regression models, we need to introduce a new data set that we can use for illustration of methods. Here, we are going to take a look at the RailTrail data set that is built into mosaic.

```
data(RailTrail)
RailTrail <- mutate(RailTrail, dayType = factor(dayType))
glimpse(RailTrail)
```

```
## Rows: 90
## Columns: 11
## $ hightemp   <int> 83, 73, 74, 95, 44, 69, 66, 66, 80, 79, 78, 65, 41, 59, 50,~
## $ lowtemp    <int> 50, 49, 52, 61, 52, 54, 39, 38, 55, 45, 55, 48, 49, 35, 35,~
## $ avgtemp    <dbl> 66.5, 61.0, 63.0, 78.0, 48.0, 61.5, 52.5, 52.0, 67.5, 62.0,~
## $ spring     <int> 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0,~
## $ summer     <int> 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0,~
## $ fall       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1,~
## $ cloudcover <dbl> 7.6, 6.3, 7.5, 2.6, 10.0, 6.6, 2.4, 0.0, 3.8, 4.1, 8.5, 7.2~
## $ precip     <dbl> 0.00, 0.29, 0.32, 0.00, 0.14, 0.02, 0.00, 0.00, 0.00, 0.00,~
## $ volume     <int> 501, 419, 397, 385, 200, 375, 417, 629, 533, 547, 432, 418,~
## $ weekday    <lgl> TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, FALSE, FALSE, TR~
## $ dayType    <fct> weekday, weekday, weekday, weekend, weekday, weekday, weekd~
```

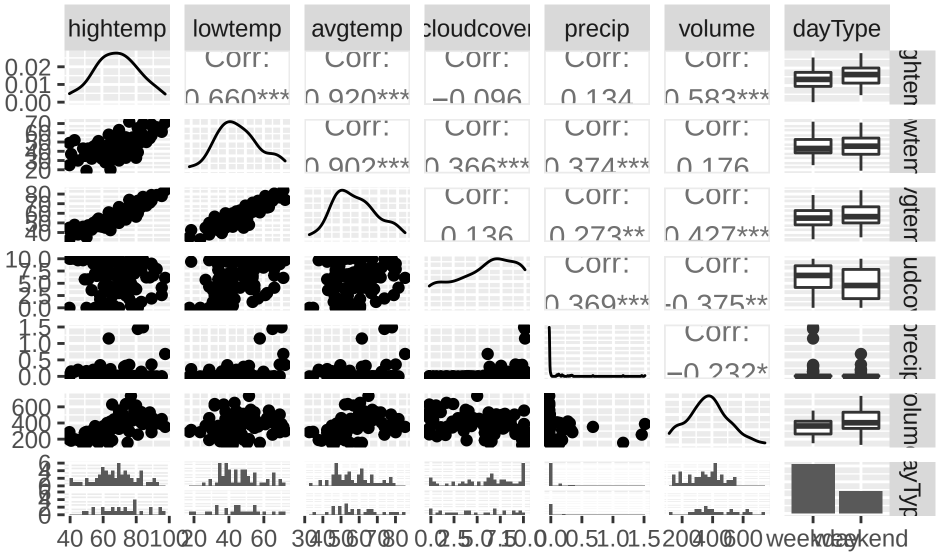
Remember that you can get help on a data set by using the help command:

```
#help(RailTrail) #remove the # sign in front to access the help, put it back to compile
#look at the types of variables in the data
```

This data is LOCAL REAL data collected in 2005 in order to assess traffic on the Northampton Rail Trail. We are interested in predicting the *volume* based on the other variables, and we'll start building up some possible models in this lab. To begin, let's use just a SLR with average temperature to predict volume. It is always a good idea to look at the distributions of the relevant variables first (you can check out the scatterplot matrix), but will we look ahead to modeling since that is what we are aiming to practice.

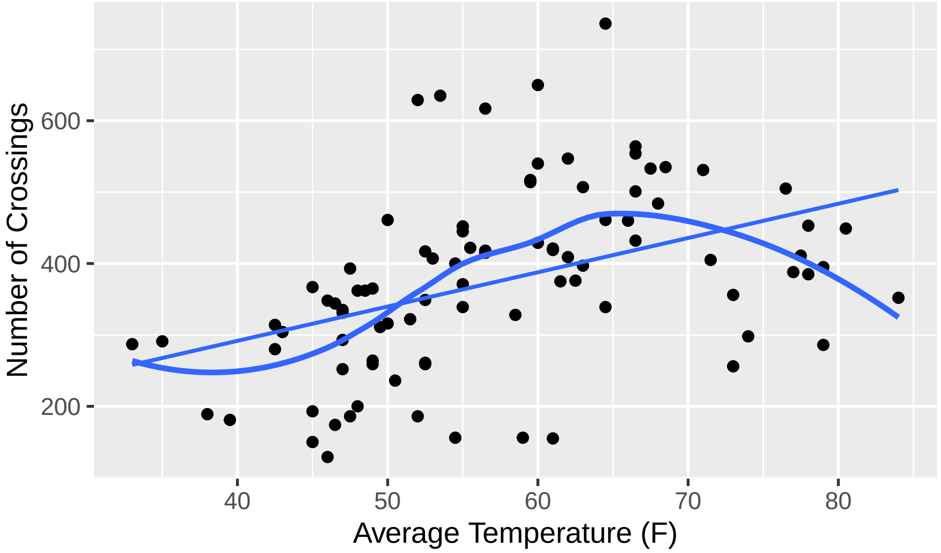
```
ggpairs(RailTrail, columns = c(1:3, 7:9, 11))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
gf_point(volume ~ avgtemp, main='graph 1 - volume by temp', xlab="Average Temperature (F)", ylab= "Numbe
    gf_lm() %>%
    gf_smooth
```

```
## `geom_smooth()` using method = 'loess'
```



1. Describe the form of the bivariate relationship between average temperature and volume.

   ANSWER: (is the relationship linear? how strong is the association?) There is a moderately strong positive relationship. It appears slightly curvilinear.

Now let's fit the SLR model.

```
mod1 <- lm(volume ~ avgtemp, data = RailTrail)
msummary(mod1)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     99.60      63.47    1.57     0.12
## avgtemp          4.80       1.08    4.43  2.7e-05 ***
##
```

```
## Residual standard error: 116 on 88 degrees of freedom
## Multiple R-squared:  0.182,  Adjusted R-squared:  0.173
## F-statistic: 19.6 on 1 and 88 DF,  p-value: 2.72e-05
```

```
predvol1 <- makeFun(mod1)
```

2. In the context of the problem, report and interpret the intercept from this model.

   ANSWER: The intercept of 99.6 corresponds to the predicted number of crossings when the average temperature was 0 degrees Fahrenheit.

3. Report and interpret the slope ($\hat{\beta}_1$) from this model.

   ANSWER: For every unit increase in average temperature, we would predict that the number of crossings would increase by 4.8, on average.

4. What is the predicted number of crossings on a day with average temperature equal to 40 degrees?

   ANSWER: (Yes you can use the code to get this)

```
predvol1(avgtemp = 40)
```

```
##       1
## 291.684
```

Either by manual calculation or using the function created by the model, we find the predicted number of crossings to be 291.7 when the temperature is 40 degrees.

5. What is the predicted number of crossings on a day with average temperature equal to 60 degrees?

   ANSWER: (Yes you can use the code to get this)

```
predvol1(avgtemp = 60)
```

```
##       1
## 387.725
```

Either by manual calculation or using the function created by the model, we find the predicted number of crossings to be 387.7 when the temperature is 60 degrees.

Surely we can add other variables besides average temperature to our model. Two variables are of particular interest: cloudcover and dayType.

6. What do you expect the relationship between cloudcover and volume to be? Think about the problem, no need to generate any plots.

   ANSWER: Fewer people might cross if there is more cloudcover. More cloudcover suggests it may not be the best day to be outside.

7. What do you expect the relationship between dayType and volume to be? Think about the problem, no need to generate any plots.

   ANSWER: I'd expect more crossings on the weekend/holidays than on weekdays.

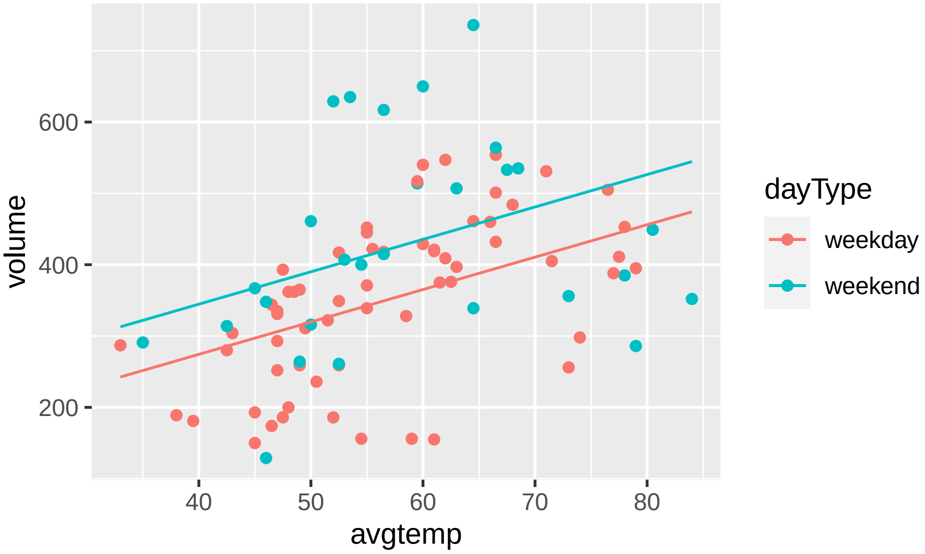
Now, we fit a new model with average temperature and dayType used to predict volume.

```
mod2 <- lm(volume ~ avgtemp + dayType, data = RailTrail)
summary(mod2)
```

```
##
## Call:
## lm(formula = volume ~ avgtemp + dayType, data = RailTrail)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -242.92  -75.72    3.15   66.90  280.07
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         92.70      61.28    1.51   0.1340
## avgtemp              4.54       1.05    4.32  4.1e-05 ***
## dayTypeweekend      70.32      25.57    2.75   0.0072 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 112 on 87 degrees of freedom
## Multiple R-squared:  0.248,  Adjusted R-squared:  0.23
## F-statistic: 14.3 on 2 and 87 DF,  p-value: 4.21e-06
```

```
#to make a plot of parallel lines we have to create a function and
# then a special plot to show the parallel lines
mod2fun <-makeFun(mod2)
gf_point(volume ~ avgtemp, main="Graph 2 - volume by daytype (no interaction term in the model)",
        data = RailTrail, color= ~ dayType) %>%
 gf_fun(mod2fun(volume, dayType="weekday") ~volume, color = ~"weekday") %>%
 gf_fun(mod2fun(volume, dayType="weekend") ~volume, color = ~"weekend")
```



```
predvol2 <- makeFun(mod2)
```

8. Looking at the model fit, is it a parallel slopes model or not?

   ANSWER: This is a parallel slopes model because it doesn't have an interaction term. The coefficient of daytype only allows a difference in intercepts.

9. In the context of the problem, report and interpret the intercept ($\hat{\beta}_0$).

   ANSWER: We have to keep in mind the reference level is a weekday in order to interpret the intercept. For a weekday with an average temperature of 0 degrees F, we would predict 92.7 crossings.

10. Report and interpret the third regression parameter from this model ($\hat{\beta}_2$).

    ANSWER: The third regression parameter is the coefficient of daytype. We predict that weekends/holidays would have 70.3 more crossings than weekdays at the same average temperature.

11. What is the predicted number of crossings on a weekday with average temperature equal to 60 degrees? How about on a weekend/holiday?

    ANSWER:

```
predvol2(avgtemp = 60, dayType = "weekday") # predicted volume for a 60 degree weekday
```

```
##       1
## 365.179
```

```
predvol2(avgtemp = 60, dayType = "weekend") # predicted volume for a 60 degree weekend/holiday
```

```
##       1
## 435.499
```

For a warmer day of average temperature 60, we get 365.2 crossings on a weekday and 435.5 on a weekend/holiday. Subtract 365.2 from 435.5 and you get the coefficient ($\hat{\beta}_2$) for daytype. If you look at the graph of parallel slopes you can see that the lines are different by 70.32 crossings regardless of temp.

Because you did not put in an interaction term the parallel lines, by definition, the lines have the same slope 4.541. For every additional degree increase in temperature you predict 4.5 more crossings regardless of daytype.

Here is some R code that may be of use in understanding the coefficients.

```
# difference in volume for an additional degree in temp on weekend (coefficient for avgtemp(slope))
predvol2(avgtemp = 61, dayType = "weekend") - predvol2(avgtemp = 60, dayType = "weekend")
```

```
##       1
## 4.54129
```

```
# difference in volume for an additional degree in temp on weekday (coefficient for avgtemp(slope))
predvol2(avgtemp = 61, dayType = "weekday") - predvol2(avgtemp = 60, dayType = "weekday")
```

```
##       1
## 4.54129
```

```
# difference in predicted volume for weekend-weekday (coefficient of dayTypeweekend)
predvol2(avgtemp = 60, dayType = "weekend") - predvol2(avgtemp = 60, dayType = "weekday")
```

```
##       1
## 70.3203
```

12. If we decide to add an interaction term to the model, will we be using a parallel slopes model or not?

    ANSWER: Adding the interaction term will allow both the intercept and slope to vary for each group, so it will no longer be a parallel slopes model. We are asking if the relationship between volume and temp changes or is different depending on if its a weekday vs. weekend. If the answer is yes, then the lines will have different slopes. Graph 3 below shows the lines when we color them by daytype. The model (mod3) tests whether those differences in slope are significant using an interaction term (avgtemp:dayType).

Let's try it! We can also visualize this new model.

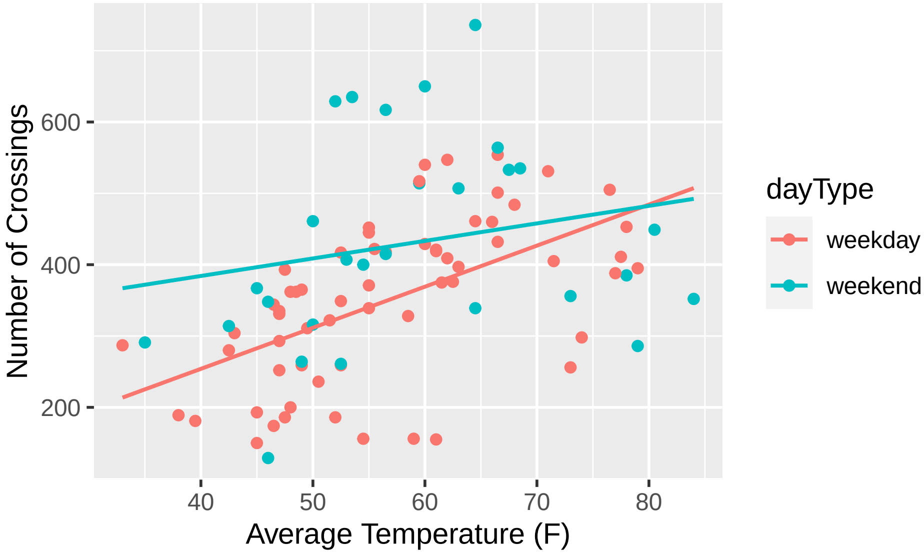
```
mod3 <- lm(volume ~ avgtemp + dayType + avgtemp*dayType, data = RailTrail) # avgtemp:dayType is equival
msummary(mod3)
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              23.69      75.77    0.31    0.755
## avgtemp                   5.76       1.31    4.39  3.2e-05 ***
## dayTypeweekend          262.19     128.18    2.05    0.044 *
## avgtemp:dayTypeweekend   -3.30       2.16   -1.53    0.130
```

```
##
## Residual standard error: 111 on 86 degrees of freedom
## Multiple R-squared:  0.267,  Adjusted R-squared:  0.242
## F-statistic: 10.5 on 3 and 86 DF,  p-value: 6.12e-06
```

```
predvol3 <- makeFun(mod3)

gf_point(volume ~ avgtemp, main="Graph 3 - volume by temp with interaction term for daytype",
    xlab = "Average Temperature (F)", ylab = "Number of Crossings", data = RailTrail, color = ~ dayType
    gf_lm()
```



13. Report and interpret the second regression parameter from this model ($\hat{\beta}_1$).

    ANSWER: This is the slope of the line for weekdays. On a weekday (reference level! -R shows the category being compared to the reference aka “dayTypeweekend), we would predict an additional 5.76 crossings for each unit increase in average temperature.

14. Report and interpret the fourth regression parameter from this model ($\hat{\beta}_3$).

    ANSWER: This is the slope of the line for weekends relative to the slope for weekdays. We would predict a decrease of 3.30 crossings on average for each one unit increase in average temperature on a weekend/holiday relative to the slope on a weekday.

15. What is the predicted number of crossings on a weekday with average temperature equal to 40 degrees? How about on a weekend/holiday?

    ANSWER:

```
predvol3(avgtemp = 40, dayType = "weekday")
```

```
##       1
## 253.985
```

```
predvol3(avgtemp = 40, dayType = "weekend")
```

```
##       1
## 384.164
```

254 crossings on a weekday and 384.2 on a weekend/holiday with an average temperature of 40. Be sure you are comfortable getting these by hand from the regression equation!

The regression equation is:
$(vol\hat{u}me) = 23.695 + 5.757$(avgtemp) + 262.193(dayTypeweekend) - 3.3(avgtemp*dayTypeweekend)

Let's tease out these coefficients a bit.

```r
mod3 <- lm(volume ~ avgtemp + dayType + avgtemp*dayType, data = RailTrail)
predvol3 <- makeFun(mod3)

# Let's look at intercept related info

# This is the difference in volume comparing weekends to weekdays when it's cold out (temp =40degrees).
# Think about this as the different volume
# between weekends and weekdays in the line at that moment on the line (40deg).
# you can see from the plot that when temp is cold (40 degrees) the lines diverge.
predvol3(avgtemp = 40, dayType = "weekend") - predvol3(avgtemp = 40, dayType = "weekday")
```

```
##       1
## 130.179
```

```r
# This is the difference in volume comparing weekends to weekdays when it warm out (temp =79degrees).
# Think about this as the different volume
# between weekends and weekdays in the line at that moment on the line (79deg).
# you can see from the plot that when temp is warm (79degrees) the lines predict nearly the same value.
predvol3(avgtemp = 79, dayType = "weekend") - predvol3(avgtemp = 79, dayType = "weekday")
```

```
##       1
## 1.46595
```

```r
#Let's look at slope related info

# this is the change in volume when temp goes up by 1 degree for weekdays (the slope for weekdays)
predvol3(avgtemp = 41, dayType = "weekday") - predvol3(avgtemp = 40, dayType = "weekday")
```

```
##       1
## 5.75726
```

```r
# this is the change in volume when temp goes up by 1 degree for weekends (the slope for weekends)
predvol3(avgtemp = 41, dayType = "weekend") - predvol3(avgtemp = 40, dayType = "weekend")
```

```
##       1
## 2.45692
```

Because we have an interaction, `daytype` is interacting with (influencing) the relationship between `volume` and `temp`, the size of the differences between our predictions for weekdays vs. weekends will change depending on temp.

In real words, temperature matters more on weekdays. When it's cold out, volume is much lower on weekdays than on weekends. When it's warm out the volume is less effected by whether or not it's a weekend vs. weekday.

In other words, volume on the railtrail is associated with temperatures on weekdays but less so on weekends. Why would that be? Imagine you only get to ride you bike on weekends, so you will go regardless of the weather. If you can ride any day of the week or you are a bike commuter, you'll take the car when it's cold on weekdays.

You can use the regression equation above to solve for any temperatures and the day type. It will work as given. That said, there really are two simplified equations that can be stated.

For weekdays: $(vol\hat{u}me) = 23.695 + 5.757$temp ::: we don't need any terms with weekend in them because weekend $= 0$

For weekends: $(\widehat{volume}) = 285.888 + 2.457(\text{temp})$ ::: we've simplified the model by adding terms

The intercept of 285.888 is the addition of the intercept for weekdays (23.695) and the additional volume associated with weekends (262.193).
The slope of 2.457 for temp is the addition of the slope for weekdays (5.757) and the reduction in the strength of the relationship between temp and volume seen on weekends (-3.3).

Thus the interpretation of the coefficient for the interaction term (which is -3.3). We predict a decrease of 3.30 crossings on average for each one unit increase in average temperature on a weekend/holiday relative to the slope on a weekday.

Model Comparison

16. Which of these three models do you currently prefer for predicting volume? Discuss with those around you. Explain your choice.

    ANSWER: To me, the final model (mod3 with the interaction term) is doing the best job, but the second one (parallel slopes with dayType) is also not too bad. The major challenge is the relationship for weekend/holidays appears to be very curved and we aren't capturing that in the model, so the R-squareds are not suggesting very good fits. I think allowing for non-parallel slopes makes more intuitive sense though.

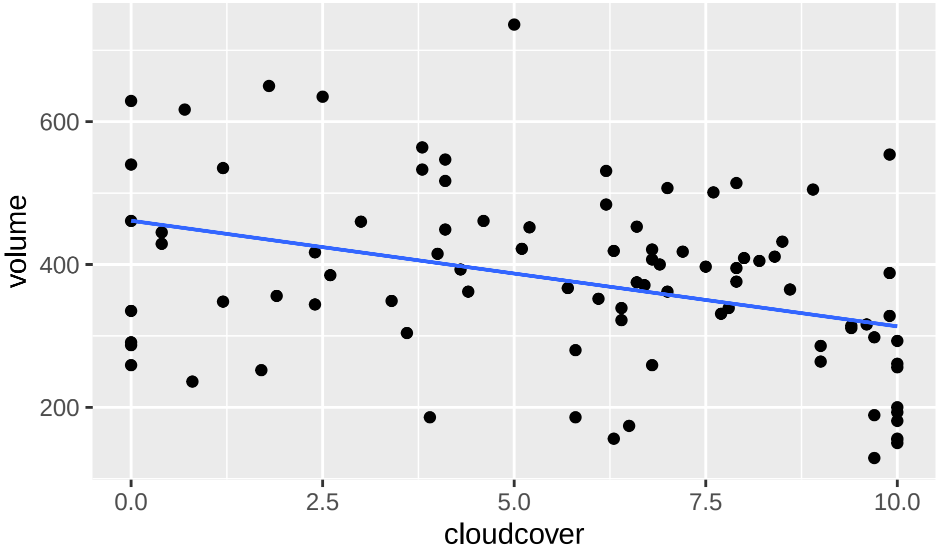
THE PARTS BELOW SHOULD BE ATTEMPTED ONCE YOU ARE COMFORTABLE WITH THE ABOVE MATERIAL, and may extend beyond class time (and you can check your answers via the solution).

Adding Cloudcover

We can add cloudcover to the model in two different ways. It's a quantitative variable, so we can just add it as a predictor, or we could consider creating groups BASED on it. In this section, you'll be taking a look at models where cloudcover is treated quantitatively versus as a categorical variable with three levels.

First, let's examine the relationship between cloudcover and volume.

```r
gf_point(volume ~ cloudcover, data = RailTrail) %>% gf_lm()
```



17. How would you characterize this relationship?

    ANSWER: There appears to be a weak negative relationship between the variables of cloudcover and volume. It might be linear looking, but there is an outlying point with volume over 700 with a cloudcover value near 5.

Fit a model that uses *avgtemp*, *dayType*, and *cloudcover* to predict *volume* , including an interaction between *avgtemp* and *dayType* (cloudcover will be treated quantitatively).

```
mod4 <- lm(volume ~ avgtemp * dayType + cloudcover, data = RailTrail)
msummary(mod4)
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                57.72      66.84    0.86   0.3903
## avgtemp                     7.04       1.18    5.98  5.2e-08 ***
## dayTypeweekend            321.12     113.10    2.84   0.0057 **
## cloudcover                -17.20       3.33   -5.16  1.6e-06 ***
## avgtemp:dayTypeweekend     -4.73       1.92   -2.47   0.0157 *
##
## Residual standard error: 97.4 on 85 degrees of freedom
## Multiple R-squared:  0.442,  Adjusted R-squared:  0.416
## F-statistic: 16.8 on 4 and 85 DF,  p-value: 3.34e-10
```

```
predvol4 <- makeFun(mod4)
```

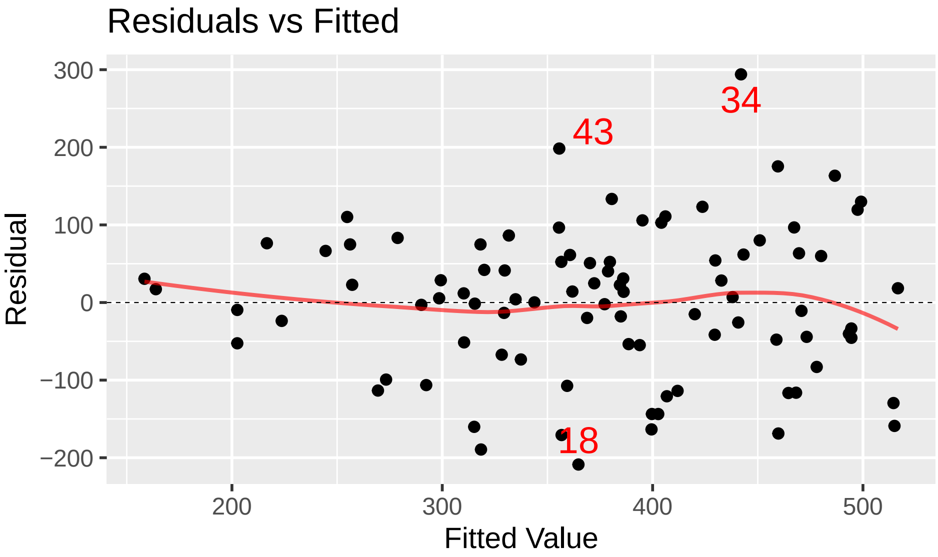
18. Interpret the slope coefficient for cloudcover in this model.

    ANSWER: With average temperature and dayType both accounted for, we predict a decrease of 17.2 crossings on average for each 1 unit increase in cloudcover.

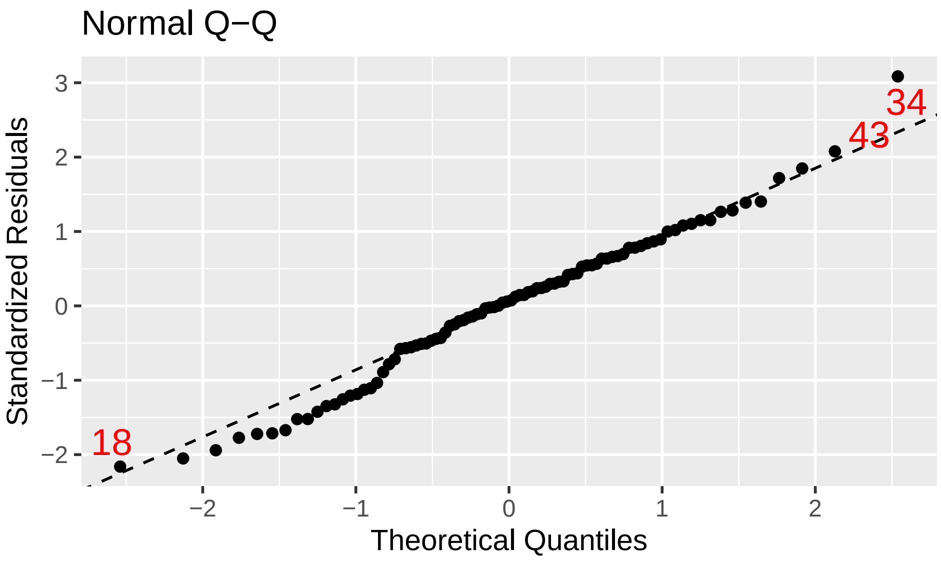
Check the regression conditions. Do you see any problems?

```
mplot(mod4, which = 1)
```

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



```
mplot(mod4, which = 2)
```



ANSWER: The QQ plot shows no serious issues though an outlier (observation #34) is noted that is particularly far from the other points. On the other hand, the residuals vs. fitted plot seems to suffer from slight increasing variation (heteroskedasticity). No other patterns are present (no issues with linearity) other than the one outlier (#34) which is beyond 2SEs from 0.

19. Assuming there are no serious problems with the conditions, does the ANOVA (F) test suggest the regression model is useful for predicting volume?

    ANSWER: Ignoring the heteroskedasticity issue, we see a very small p-value for the ANOVA, which suggests at least one predictor is significant for predicting volume.

20. Which individual predictors are significant according to the output?

    ANSWER: Again, ignoring the heteroskedasticity, it looks like all predictors, including the interaction term, are significant at the 0.02 level. However, we'd want to investigate ways of addressing the non-constant variance before going forward.

21. We can also treat cloudcover as a categorical variable if we break it into groups. Suppose we decide to break it into three groups. Code to accomplish this and relabel the groups follows:

```
#this introduces the new function cut; can you figure out what cut does?
RailTrail <- mutate(RailTrail, cloudgrp = cut(cloudcover, breaks = c(0, 3.333, 6.667, 10),
                     labels = c("low", "medium", "high"), include.lowest = TRUE))
#this introduces the new function cut; can you figure out what cut does?
tally(~ cloudgrp, data = RailTrail)
```

```
## cloudgrp
##    low medium   high
##     21     28     41
```

We have now generated low, medium, and high cloudcover groups in the variable *cloudgrp*. Again, we should always consider relationships between the variables, especially now that we have a new variable, but we forgo this in the interest of time and modeling here.

Now, let's fit a multiple linear regression model to the data using *cloudgrp*, *avgtemp*, *dayType*, and an interaction between *avgtemp* and *dayType* in order to predict *volume*.

```
#mod5 <- lm(volume ~ cloudgrp + dayType + avgtemp + avgtemp:dayType, data = RailTrail) #equivalent
mod5 <- lm(volume ~ cloudgrp + avgtemp*dayType, data = RailTrail)
#R is including the main effects of avgtemp and dayType as well
msummary(mod5)
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                31.37      71.39    0.44  0.66150
## cloudgrpmedium            -32.72      30.72   -1.07  0.28990
## cloudgrphigh             -104.50      28.67   -3.65  0.00046 ***
## avgtemp                     6.71       1.24    5.40  6.2e-07 ***
## dayTypeweekend            331.29     121.40    2.73  0.00774 **
## avgtemp:dayTypeweekend     -4.72       2.05   -2.30  0.02380 *
##
## Residual standard error: 103 on 84 degrees of freedom
## Multiple R-squared:  0.383,  Adjusted R-squared:  0.346
## F-statistic: 10.4 on 5 and 84 DF,  p-value: 8.41e-08
```

22. Does the overall model appear to be useful in predicting volume? Explain how you know.

    ANSWER: Assuming appropriate conditions are met, the ANOVA p-value is significant indicating the model is useful in predicting volume.

23. Pick a coefficient in the model. Interpret the coefficient in the regression model. Is the associated predictor significant in the model (while assuming conditions for inference hold)? How do you know?

    ANSWER: Answers will vary depending on what coefficient you choose.

$(\hat{\beta}_0)$ - The intercept says for a weekday with low cloudcover, at 0 average temperature, we'd predict an average of 31.4 crossings. This coefficient is for both reference categories (cloudgrplow and weekday). Because it's an intercept avgtemp=0.

$(\hat{\beta}_1)$ - On a medium cloudcover day, once average temperature and daytype are accounted for, in general, we'd predict 32.7 fewer crossings than on a day with low cloudcover. We know that more cloud cover will reduce volume, regardless of temp and daytype, so we'll need to reduce our estimate by 32.7 as compared to when it's low cloud cover (the reference category).

$(\hat{\beta}_2)$ - On a high cloudcover day, once average temperature and daytype are accounted for, we'd predict 104.5 fewer crossings than on a day with low cloudcover. Now it's even cloudier and we reduce volume predictions by 104.5 compared to a low cloud cover day. Look at the final graph and you'll see all the predictions of volume reduce as we move from low to high cloud cover. The most dramatic reduction is for high cloud cover.

$(\hat{\beta}_3)$ - On a weekday with low cloud cover (reference levels!), we would predict an additional 6.7 crossings for each unit increase in average temperature. This is the slope for average temp. It's describing the relationship between temp and volume assuming it's a low cloud cover weekday. The terms below avgtemp (i.e., dayTypeweekend and avgtemp:dayTypeweekend) aren't yet under consideration.

$(\hat{\beta}_4)$ - On a day with average temperature equal to 0, we would predict 331.294 more crossings on a weekday/holiday compared to a weekday, once cloudcover is accounted for. Think of this as the intercept for weekends. What amount of volume needs to be added to our prediction if people have the day off?

$(\hat{\beta}_5)$ - We would predict an decrease of 4.7 crossings on average for each one unit increase in average temperature on a weekend/holiday relative to the slope on a weekday, once cloudcover is accounted for. This is the toughest one to interpret. It isn't as intuitive. Adding cloudcover has changed this from a positive to a negative effect. We can see this coefficient is about slope since we are talking about change in volume based on temperature. Because this is the coefficient for the interaction term, it estimates how the relationship between temp and volume changes when it's a weekend vs. a weekday. When you look at the graph with parallel slops from mod3, weekends have a different slope; it's much less steep. Meaning that while weekends tend to have more volume than weekdays that effect diminishes as the temp rises.

FOR ALL OF THESE, except the intercept, the associated predictor is significant given the presence of the other variables in the model. Note that cloudgrpmedium is NOT significant, so we might be able to collapse medium and low into one category, but the HIGH category is significant, so we should still include a factor variable here.

The regression equation is:

$(\widehat{volume})$ = 31.369 + -32.716(cloudgrpmedium) -104.497(cloudgrphigh) +6.708(avgtemp) +33.294(dayTypeweekend) - 4.719(avgtemp*dayTypeweekend)

Let's tease apart some of these equations for certain predictions:

24. PREDICTION 1: What would the regression equation be to predict volume on a day with low cloud cover (reference category so no terms for medium or high needed) and not a weekday (aka, dayTypeweekend=1)?

    ANSWER:

pred_volume = 362.663 + 1.989(avgtemp)

where the following info was combined:
pred_volume = (31.369 + 331.294) + (6.708 - 4.719)(avgtemp)

For the intercept: we've added the intercept for our reference categories (lowcloudcover and weekday = 31.369) to the coefficient for dayTypeweekend (331.294) since we are being asked to predict a non weekday. Remember 331.294 is how many additional crossings we add to volume because it's a weekday having already accounted for cloudcover.

For the slope: we've added the slope for avgtemp (6.708) to the slope for the interaction term avgtemp:dayTypeweekend (-4.719). When we are predicting a weekend we are adjusting the change in volume downward because the relationship between temp and volume is not as strong on weekends as we see it on weekdays.

25. PREDICTION 2: What would the regression equation be to predict volume on a day with medium cloud cover that is a weekday? (medium cloud cover has a term in the model we'll need and weekday is a reference category so we won't hav eto worry about terms relating to dayTypeweekend i.e., dayTypeweekend and avgtmp::dayTypeweekend)

    ANSWER:

pred_volume = -1.347 + 6.708(avgtemp)

where the following info was combined: pred_volume = (31.369 - 32.716) + 6.708(avgtemp)

For the intercept: we've added the intercept for our reference group weekday (31.369) with the coefficient for cloudgrpmedium (-32.716) since we are being asked to predict volume for a medicum cloud cover weekday.

For the slope: we can just use the slope for avgtemp (6.708). We don't need to consider other terms (cloudgrphigh, dayTypeweekend, avgtemp:dayTypeweekend) in the equation because they don't relate to weekdays.

JUST FOR FUN! SAY you didn't know how to combine terms. What would happen if you punted and solved the full equation for PREDICTION 2?

$(\widehat{volume})$ = 31.369 + -32.716(cloudgrpmedium) -104.497(cloudgrphigh) +6.708(avgtemp) +33.294(dayTypeweekend) - 4.719(avgtemp*dayTypeweekend)

Plug in the values where cloudgrpmedium=1 and dayTypeweekend=0; $(\widehat{volume})$ = 31.369 + -32.716(1) -104.497(0) +6.708(avgtemp) +33.294(0) - 4.719(0)

If you solve the simplified equation and this equation you should get the same prediction. Then you would know you've correctly done the equation for that scenario.

These are only two possible scenarios for prediction. You could need to predict volume for a highcloudcover weekend. That would be a separate equation using the relevant terms in the equation for that scenario.

26. Interpret the R-squared for your model5. How well does your model fit the data?

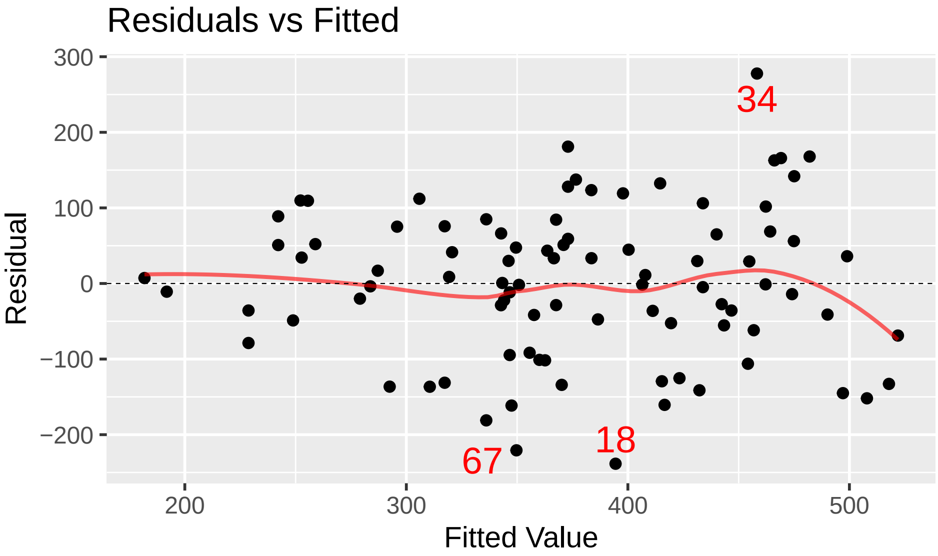
ANSWER: Our model explains 38.29% of the variability in number of crossings. This does not indicate a great fit.

27. Go check the conditions. Do you have any concerns about violations of conditions?

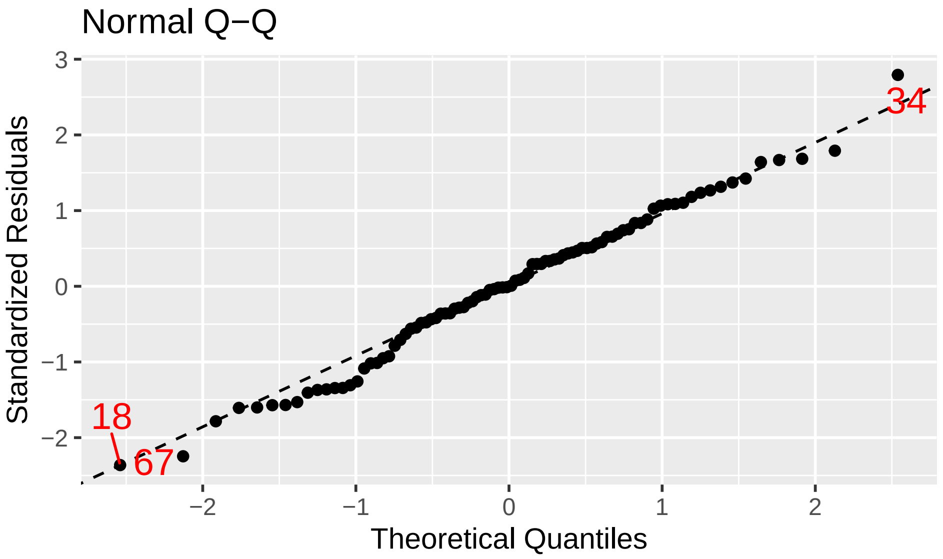
    ANSWER:

```
mplot(mod5, which = 1)
```

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



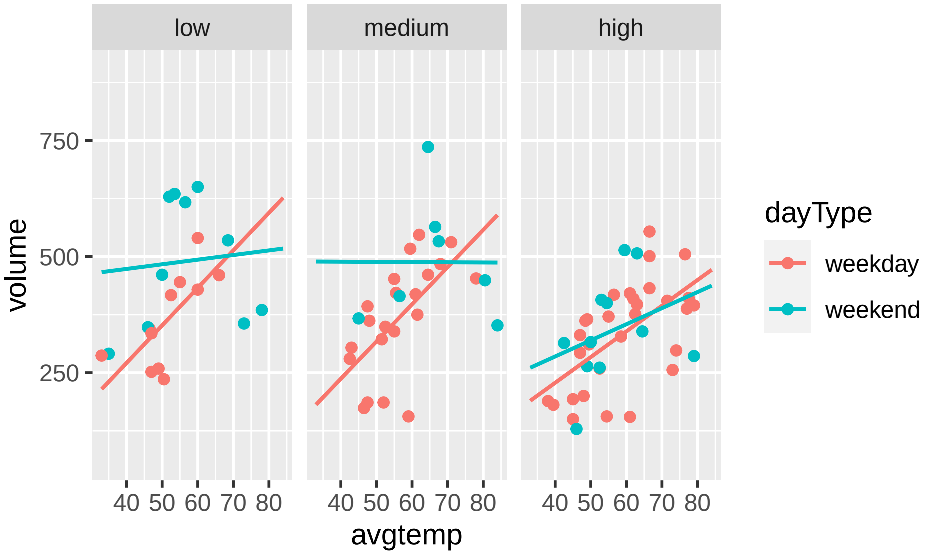
```
mplot(mod5, which = 2)
```



The same problems arise that arose with model 4. We see an unusual point #34, and we also see problems with increasing variability in the residuals vs. fitted plot. We may return to try to solve these problems later.

Finally, here is a plot that demonstrates how to give a visualization of this model. It works particularly well in this case because we are plotting the two quantitative variables and we can use color and paneling/faceting to represent the two categorical variables.

```
gf_point(volume ~ avgtemp | cloudgrp, data = RailTrail, color = ~ dayType) %>% gf_lm()
```



Favorite Model

28. After considering all of these models, which one is your favorite for predicting volume?

    ANSWER: The final model has the highest R^2, and takes into account variables that seem reasonable to me for predicting volume. I like it the best, though we still have issues with it to address!

In future classes, we'll see how to test between models to see if it's really worth adding/removing variables.