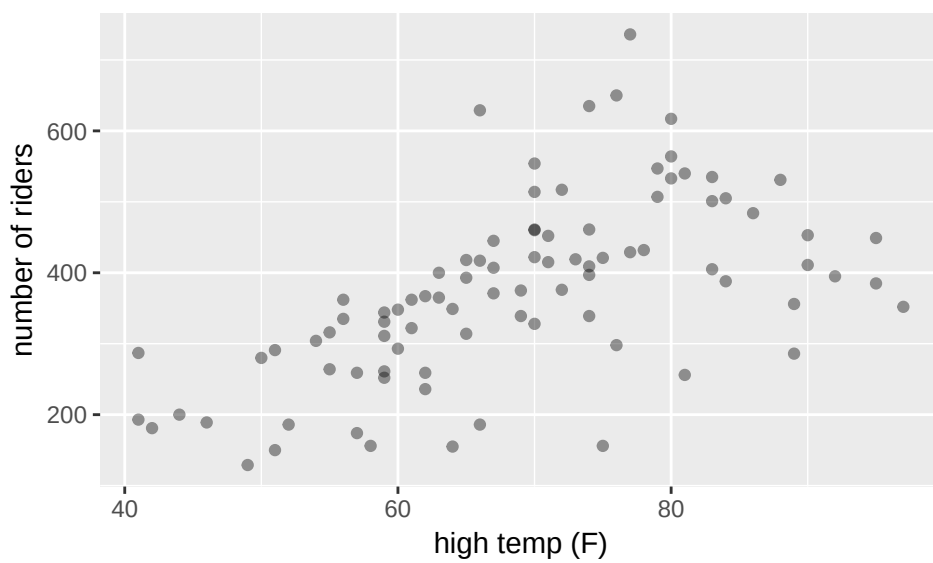
Rail Trail, interaction

The variables used in the following analysis are `hightemp`, `volume`, `precip` and `weekday`. A description of the data is given at:

```
library(mosiacData)
?RailTrail
```

1. It is *always* a good idea to graph your data and look at numerical summaries. Sometimes you'll find out important artifacts or mistakes.

```
RailTrail %>%
  ggplot(aes(x=hightemp, y=volume)) + geom_point(alpha=0.4) +
  xlab("high temp (F)") + ylab("number of riders")
```




```
RailTrail %>%
  select(hightemp, volume, weekday) %>%
  skim()

## Skim summary statistics
##  n obs: 90
##  n variables: 3
##
## Variable type: factor
##  variable missing complete  n n_unique          top_counts ordered
##   weekday       0       90 90        2 1: 62, 0: 28, NA: 0   FALSE
##
## Variable type: integer
##  variable missing complete  n   mean     sd  p0    p25 median    p75 p100
##  hightemp       0       90 90  68.83  13.02  41  59.25   69.5  77.75   97
##    volume       0       90 90 375.4  127.46 129 291.5   373   451.25  736
```

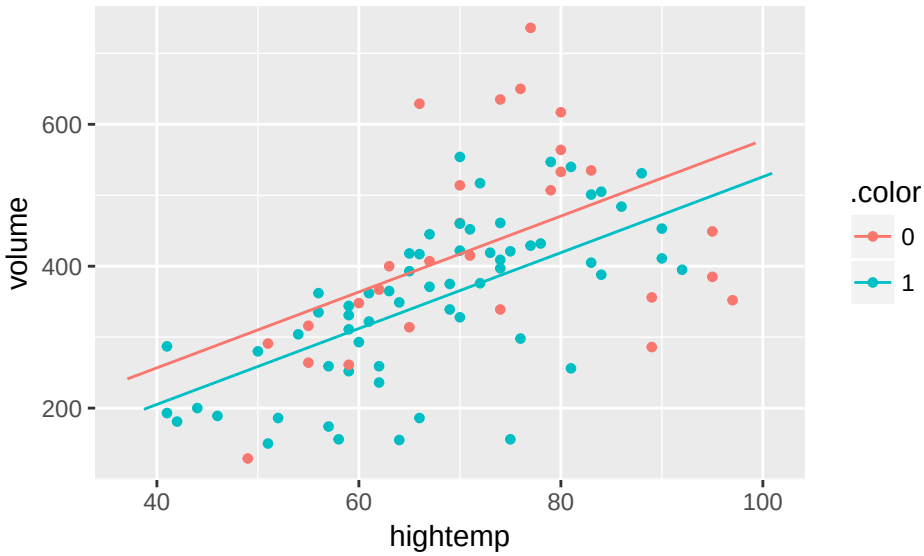
2. We're interested in predicting the volume of riders from the high temperature (in F) in a given day.

```
tidy(lm(volume ~ hightemp, data=RailTrail))

##          term estimate std.error statistic  p.value
## 1 (Intercept)    -17.1    59.395    -0.288 7.74e-01
## 2    hightemp      5.7     0.848     6.724 1.71e-09

summary(lm(volume ~ hightemp, data=RailTrail))

##
## Call:
## lm(formula = volume ~ hightemp, data = RailTrail)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -254.56  -57.80    8.74   57.35  314.03
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.079     59.395   -0.29     0.77
## hightemp       5.702      0.848    6.72  1.7e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104 on 88 degrees of freedom
## Multiple R-squared:  0.339,Adjusted R-squared:  0.332
## F-statistic: 45.2 on 1 and 88 DF,  p-value: 1.71e-09
```

3. What happens when `weekday` is included as a binary indicator variable?

```
library(mosaic)
tidy(lm(volume ~ hightemp + weekday, data=RailTrail))

##          term estimate std.error statistic  p.value
## 1 (Intercept)    42.81    64.344     0.665 5.08e-01
## 2    hightemp     5.35     0.846     6.319 1.09e-08
## 3    weekday1   -51.55    23.674    -2.178 3.21e-02

summary(lm(volume ~ hightemp + weekday, data=RailTrail))

##
## Call:
## lm(formula = volume ~ hightemp + weekday, data = RailTrail)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -236.3  -59.9   12.4   60.9  281.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42.807     64.344    0.67    0.508
## hightemp       5.348      0.846    6.32  1.1e-08 ***
## weekday1     -51.553     23.674   -2.18    0.032 *
```

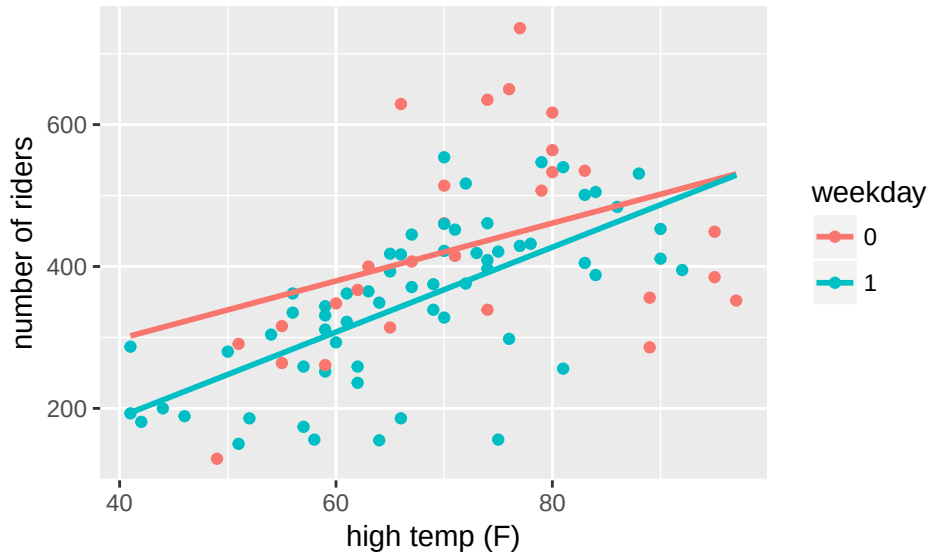
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102 on 87 degrees of freedom
## Multiple R-squared:  0.374,Adjusted R-squared:  0.359
## F-statistic: 25.9 on 2 and 87 DF,  p-value: 1.46e-09

plotModel(lm(volume ~ hightemp + weekday, data=RailTrail), system="ggplot2")
```



- Note that the F p-value is no longer equal to the p-value(s) associated with the t-test for any of the coefficients. Also, the degrees of freedom are now (2, 87) because the model estimates 3 parameters.
- Write out the estimated regression model separately for weekdays and weekends, and sketch the lines onto the scatterplot.
- How are the new coefficients (b_0, b_1, b_2) interpreted?
- How did the coefficient on `hightemp` change?
- How does R^2 change? MSE change?
- Why does it say `weekday1` instead of `weekday`?

4. What if `hightemp` and `weekday` *interact*?

```
RailTrail %>%
  ggplot(aes(x=hightemp, y=volume, color=weekday)) + geom_point() +
  xlab("high temp (F)") + ylab("number of riders") +
  geom_smooth(method="lm", se=FALSE, fullrange=TRUE)
```




```
tidy(lm(volume ~ hightemp * weekday, data=RailTrail))

##                term estimate std.error statistic p.value
## 1       (Intercept)   135.15    108.21      1.25 0.21505
## 2          hightemp     4.07      1.47      2.78 0.00676
## 3          weekday1  -186.38    129.25     -1.44 0.15293
## 4 hightemp:weekday1     1.91      1.80      1.06 0.29163

summary(lm(volume ~ hightemp * weekday, data=RailTrail))

##
## Call:
## lm(formula = volume ~ hightemp * weekday, data = RailTrail)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -241.30  -62.48    8.91   55.77  287.10
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         135.15     108.21    1.25   0.2150
## hightemp              4.07       1.47    2.78   0.0068 **
## weekday1           -186.38     129.25   -1.44   0.1529
## hightemp:weekday1     1.91       1.80    1.06   0.2916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102 on 86 degrees of freedom
## Multiple R-squared:  0.382,Adjusted R-squared:  0.36
## F-statistic: 17.7 on 3 and 86 DF,  p-value: 4.96e-09
```

- Note again that the F p-value is no longer equal to the t-stat p-value(s). Now the degrees of freedom are (3, 86) because the model estimates 4 parameters.

- Write out the estimated regression model separately for smokers and non-smokers, and sketch the lines onto the scatterplot.
- How do you interpret your new coefficients (b_0, b_1, b_2, b_3)?
- What happened to the significance? How did the coefficient on `weekday` change?
- How does R^2 change? MSE change?

5. What happens to the model with an additional quantitative variable?

```
tidy(lm(volume ~ hightemp + weekday + precip, data=RailTrail))

##          term estimate std.error statistic  p.value
## 1 (Intercept)     19.3    60.339      0.32 7.50e-01
## 2    hightemp      5.8     0.799      7.26 1.59e-10
## 3    weekday1    -43.1    22.194     -1.94 5.52e-02
## 4      precip   -145.6    38.894     -3.74 3.27e-04

summary(lm(volume ~ hightemp + weekday + precip, data=RailTrail))

##
## Call:
## lm(formula = volume ~ hightemp + weekday + precip, data = RailTrail)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -255.23  -49.70    6.47   44.12  270.03
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.319     60.339    0.32  0.74961
## hightemp       5.801      0.799    7.26  1.6e-10 ***
## weekday1     -43.144     22.194   -1.94  0.05517 .
## precip      -145.609     38.894   -3.74  0.00033 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 95.2 on 86 degrees of freedom
## Multiple R-squared:  0.461,Adjusted R-squared:  0.443
## F-statistic: 24.6 on 3 and 86 DF,  p-value: 1.44e-11
```

Note the p-values, parameter estimates, R^2, MSE, F-stat, df, and F-stat p-values.

6. Just for fun, let's investigate volume by season and how the coefficients covary.

```
tidy(lm(volume ~ spring + fall, data=RailTrail))

##          term estimate std.error statistic  p.value
## 1 (Intercept)    421.9      24.6     17.17 1.13e-29
## 2      spring    -50.2      29.8     -1.68 9.57e-02
## 3        fall   -126.8      43.2     -2.94 4.23e-03

vcov(lm(volume ~ spring + fall, data=RailTrail))

##             (Intercept) spring fall
## (Intercept)         604   -604 -604
## spring             -604    889  604
## fall               -604    604 1862
```

Fires, quadratic terms

US Wildfires from 1960 - 2012:

We can see that the quadratic term alone (without the linear term) doesn't help the model fit because the vertex is in the plot. By making the linear part zero, we force the vertex to be at X=0 which doesn't make sense for the model fit. Indeed, typically a quadratic term without a linear term is a good idea if (a) there is a curved relationship with constant errors and the vertex is not in the plot, and/or (b) you really believe that there is a reason why Y should be a (linear) function of X^2.

```
fires <- read.table("http://pages.pomona.edu/~jsh04747/courses/math158/fires.txt",
                    header=TRUE)
lmod = lm(Acres~Year, data=fires)
qmod1 = lm(Acres ~I(Year^2), data=fires)  # math squared term
qmod2 = lm(Acres~Year + I(Year^2), data=fires)  # math squared & linear terms
qmod3 = lm(Acres ~ Year^2, data=fires)  # second order (self) interaction term

tidy(lmod)

##          term  estimate std.error statistic p.value
## 1 (Intercept) -1.13e+08  37566682     -3.00 0.00415
## 2        Year  5.90e+04     18915      3.12 0.00299

tidy(qmod1)

##          term  estimate std.error statistic p.value
## 1 (Intercept) -5.45e+07  1.88e+07     -2.90 0.00547
## 2   I(Year^2)  1.49e+01  4.76e+00      3.13 0.00285

tidy(qmod2)

##          term  estimate std.error statistic  p.value
## 1 (Intercept)  2.19e+10  4.54e+09      4.83 1.33e-05
## 2        Year -2.22e+07  4.57e+06     -4.84 1.27e-05
## 3   I(Year^2)  5.59e+03  1.15e+03      4.86 1.22e-05

tidy(qmod3)

##          term  estimate std.error statistic p.value
## 1 (Intercept) -1.13e+08  37566682     -3.00 0.00415
## 2        Year  5.90e+04     18915      3.12 0.00299
```

Residual Plots