Introducing Multiple Linear Regression

Dr. Maria Tackett

10.17.19

Click for PDF of slides

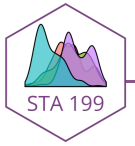
Announcements

- Complete Reading 05 (if you haven't already done so)

Data & packages

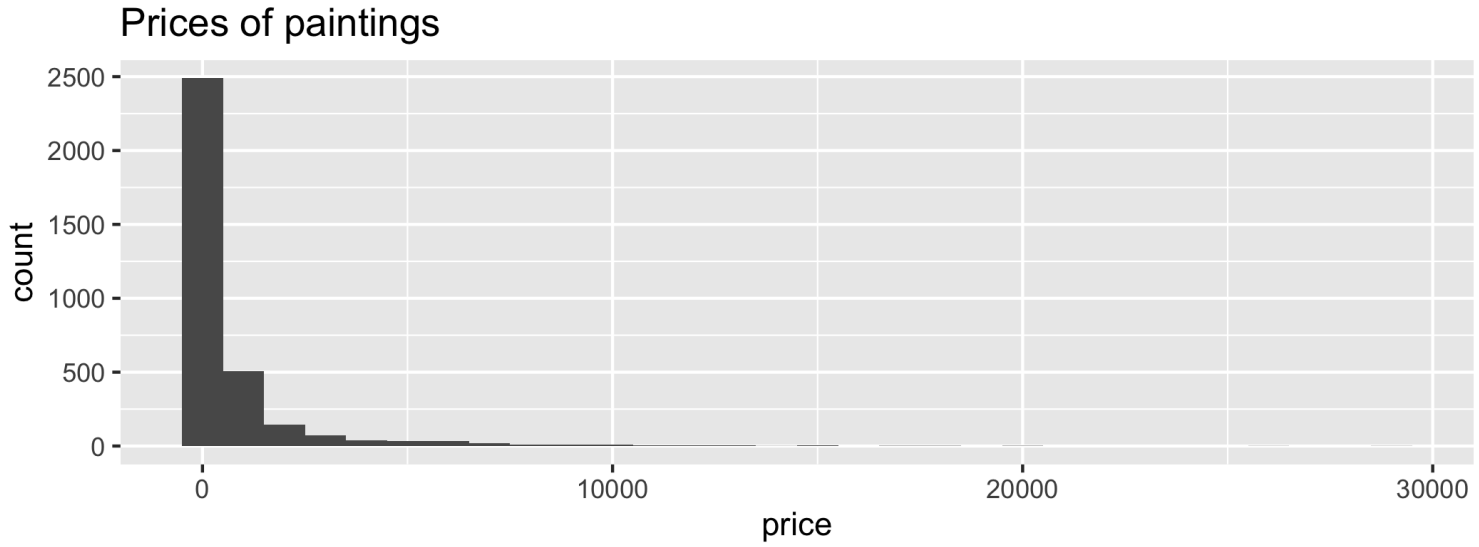
```
library(tidyverse)
library(broom)
```

```
pp <- read_csv("data/paris_paintings.csv",
               na = c("n/a", "", "NA"))
```



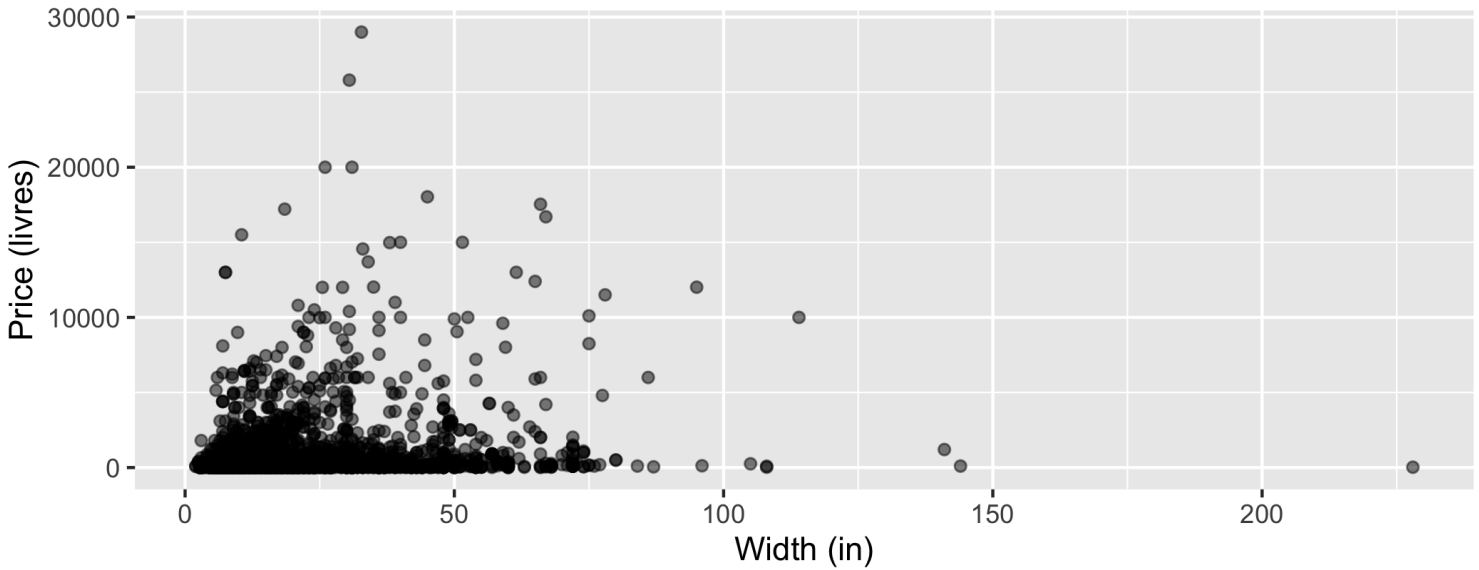
Exploring linearity

Data: Paris Paintings



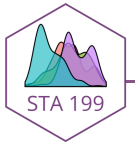
Price vs. width

Describe the relationship between price and width of painting.



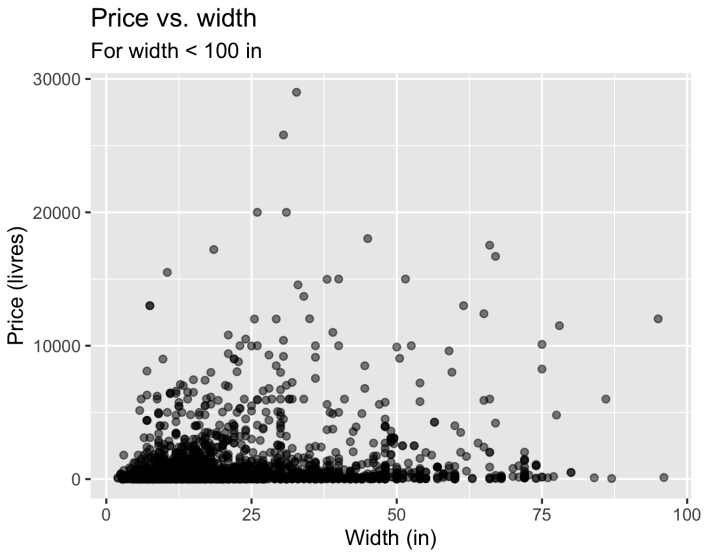
Let's focus on paintings with `Width_in < 100`

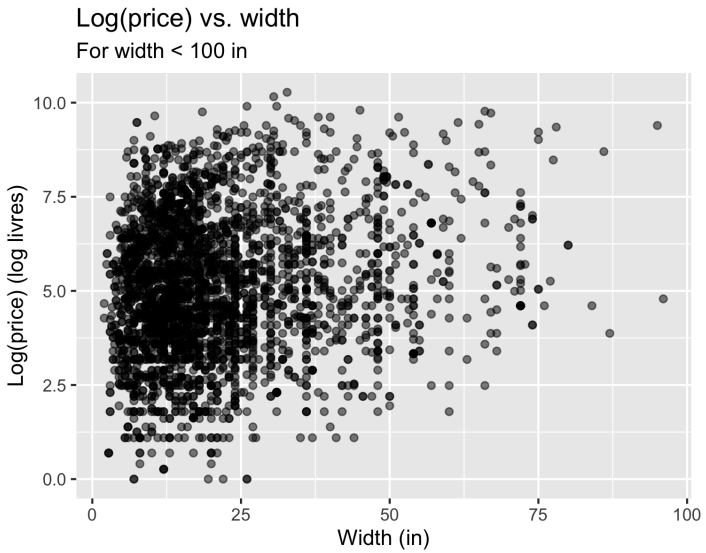
```
pp_wt_lt_100 <- pp %>%
  filter(Width_in < 100)
```



Price vs. width

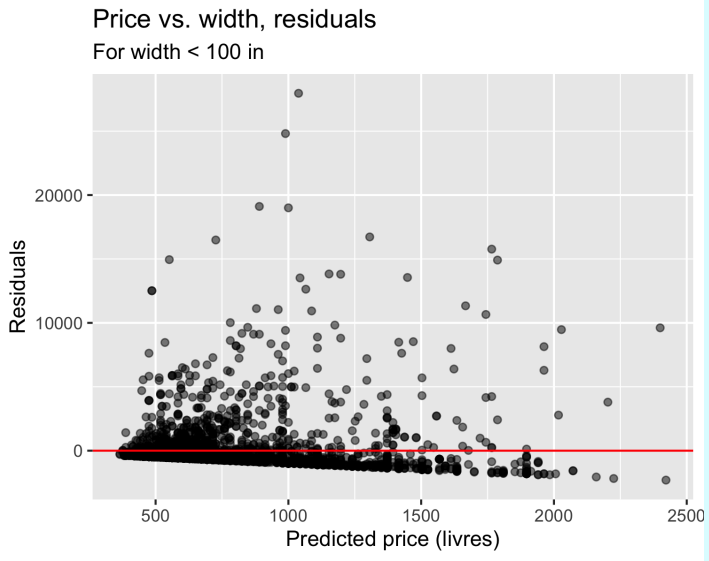
Which plot shows a more linear relationship?

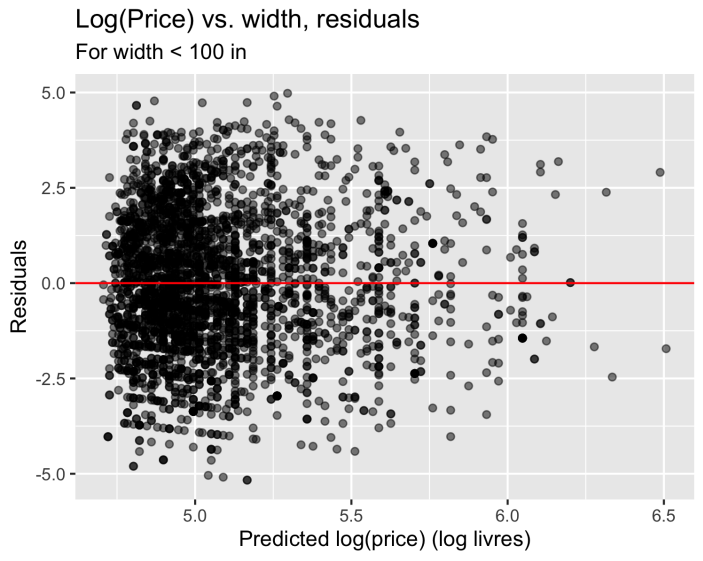




Price vs. width, residuals

Which plot shows a residuals that are uncorrelated with predicted values from the model?





What's the unit of residuals?

Transforming the data

- We saw that **`price`** has a right-skewed distribution, and the relationship between price and width of painting is non-linear.
- We also observed signs of the model violation, non-constant variance.
- In these situations a transformation applied to the response variable (y) may be useful.
    - The most common transformation is the log transformation $(\log(y) = ln(y))$
- This is beyond the scope of the course, but I'm happy to provide guidance if you want to try modeling a response that requires transformation in your final project

The linear model with multiple predictors

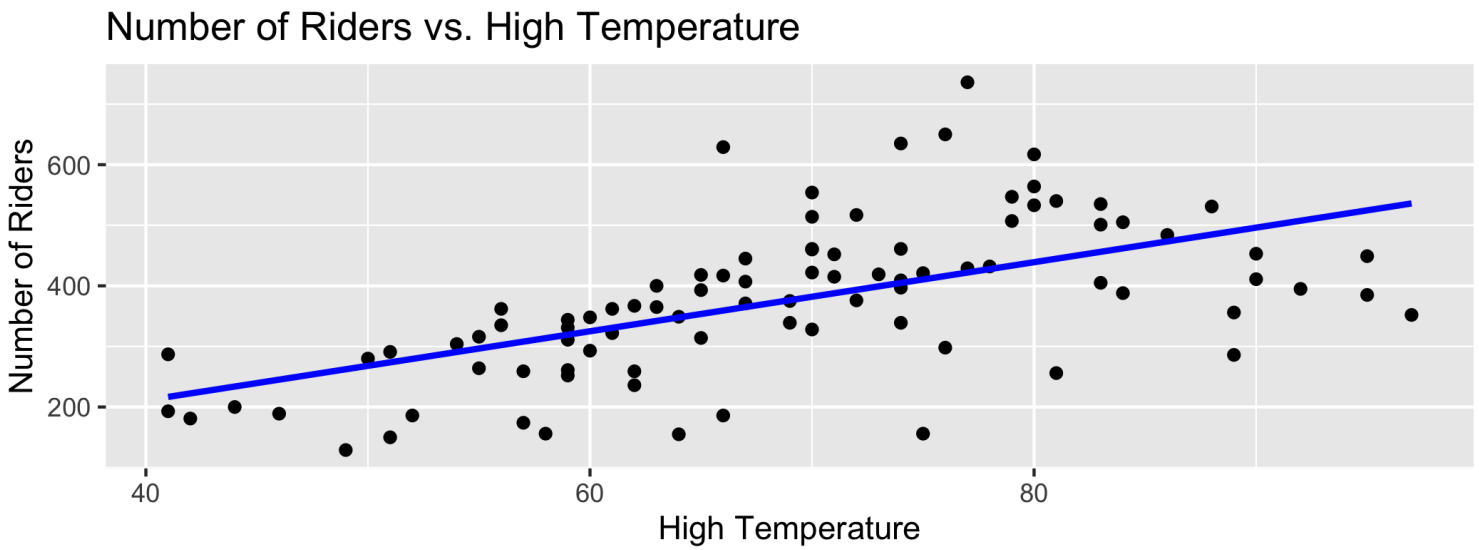
Riders in Florence, MA

The Pioneer Valley Planning Commission collected data in Florence, MA for 90 days from April 5 to November 15, 2005 using a laser sensor, with breaks in the laser beam recording when a rail-trail user passed the data collection station.

- **`hightemp`**: daily high temperature (in degrees Fahrenheit)
- **`volume`**: estimated number of trail users that day (number of breaks recorded)
- **`dayType`**: weekday or weekend

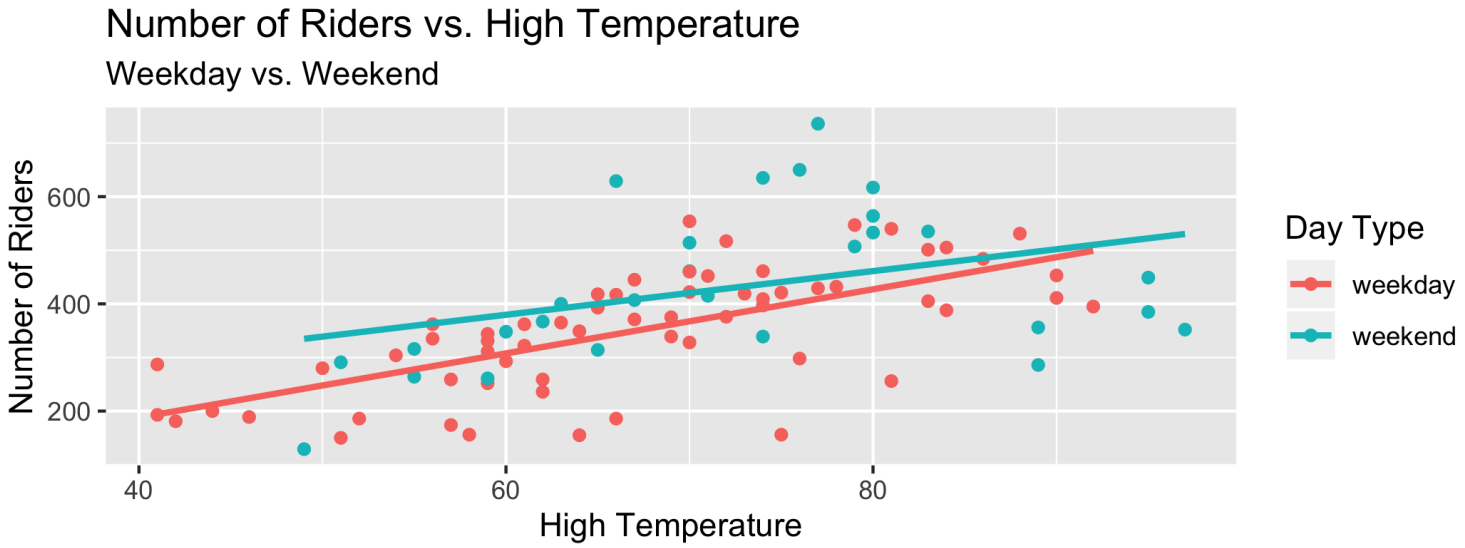
```
library(mosaicData)
data(RailTrail)
```

Volume vs. Temperature



Volume, Temperature, and Day

Does the relationship between and **`volume`** and **`hightemp`** differ by whether or not it's a weekday or weekend.



Modeling with main effects

```
m_main <- lm(volume ~ hightemp + dayType, data = RailTrail)
m_main %>% tidy() %>% select(term, estimate) %>%
  kable(format = "markdown", digits =3) #knitr package
```

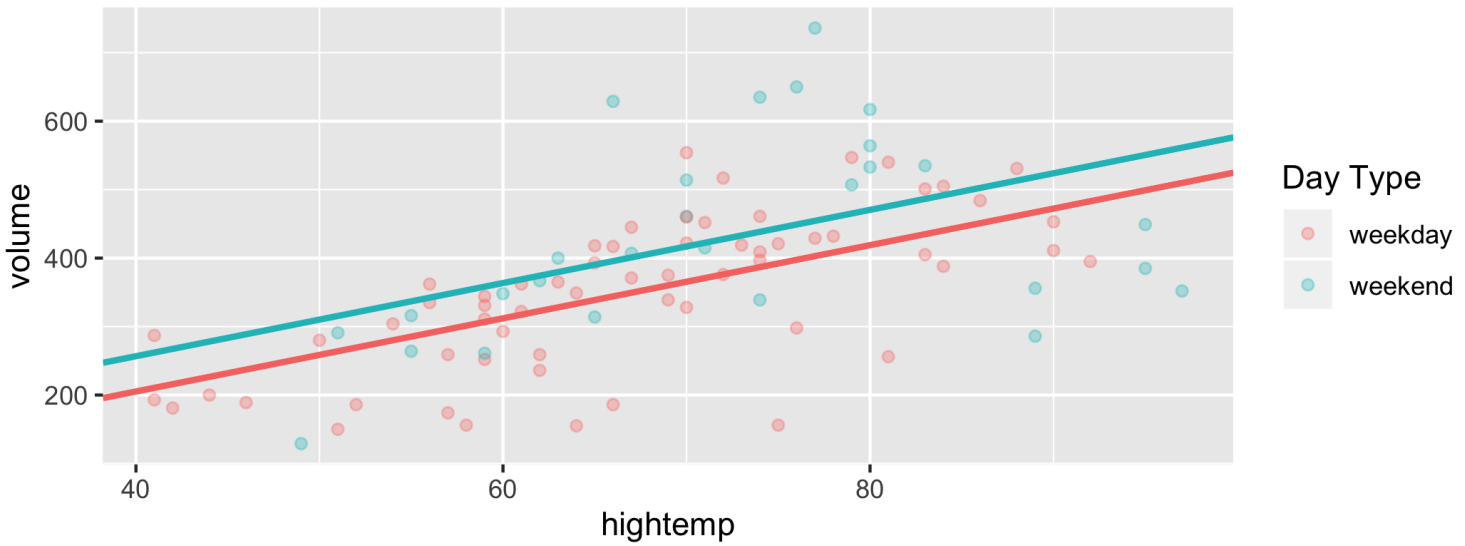
term	estimate
(Intercept)	-8.747
hightemp	5.348
dayTypeweekend	51.553

Linear model:

$$\widehat{volume} = -8.747 + 5.348\ hightemp + 51.553\ dayTypeweekend$$

- Plug in 0 for **dayTypeweekend** to get the linear model for weekdays, i.e. days that aren't the weekend.
- Plug in 1 for **dayTypeweekend** to get the linear model for days on the weekend.

Interpretation of main effects



- Weekday (`dayTypeweekend == 0`)

$$\widehat{volume} = -8.747 + 5.438\ hightemp + 51.553 \times 0$$
$$= -8.747 + 5.438\ hightemp$$

- Weekend (`dayTypeweekend == 1`):

$$\widehat{volume} = -8.747 + 5.438\ hightemp + 51.553 \times 1$$
$$= 42.806 + 5.438\ hightemp$$

- Weekday (`dayTypeweekend == 0`)

$$\widehat{volume} = -8.747 + 5.438\ hightemp + 51.553 \times 0$$
$$= -8.747 + 5.438\ hightemp$$

- Weekend (`dayTypeweekend == 1`):

$$\widehat{volume} = -8.747 + 5.438\ hightemp + 51.553 \times 1$$
$$= 42.806 + 5.438\ hightemp$$

- Rate of change in volume as the high temperature increases is the same on days that are weekdays and weekends (same slope).
- Given the same high temperature, days on the weekend are expected to have higher volume than days that are weekdays (different intercept).

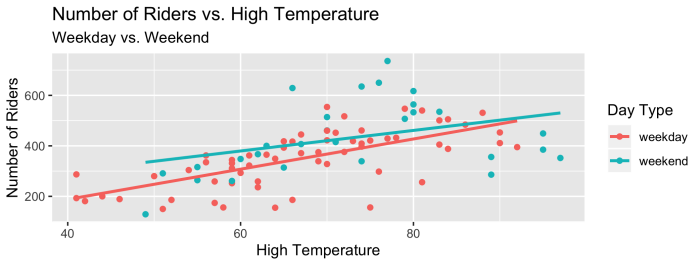
Main effects, numerical and categorical predictors

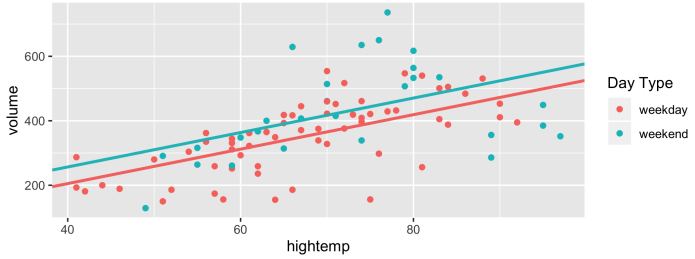
term	estimate
(Intercept)	-8.747
hightemp	5.348
dayTypeweekend	51.553

- For each additional degree Fahrenheit in the day's high temperature, there are predicted to be, on average, 5.3478168 (about 5) additional riders on the trail, holding all else constant.
- Days on the weekend are predicted to have, on average, 51.553496 (about 52) more riders on the trail than days that are weekdays, holding all else constant.
- Weekdays that have a high temperature of 0 degrees Fahrenheit are predicted to have -8.7469229 (about -9) riders, on average.

What went wrong?

Why is our linear regression model different from what we got from **`geom_smooth(method = "lm")`**?





What went wrong? (cont.)

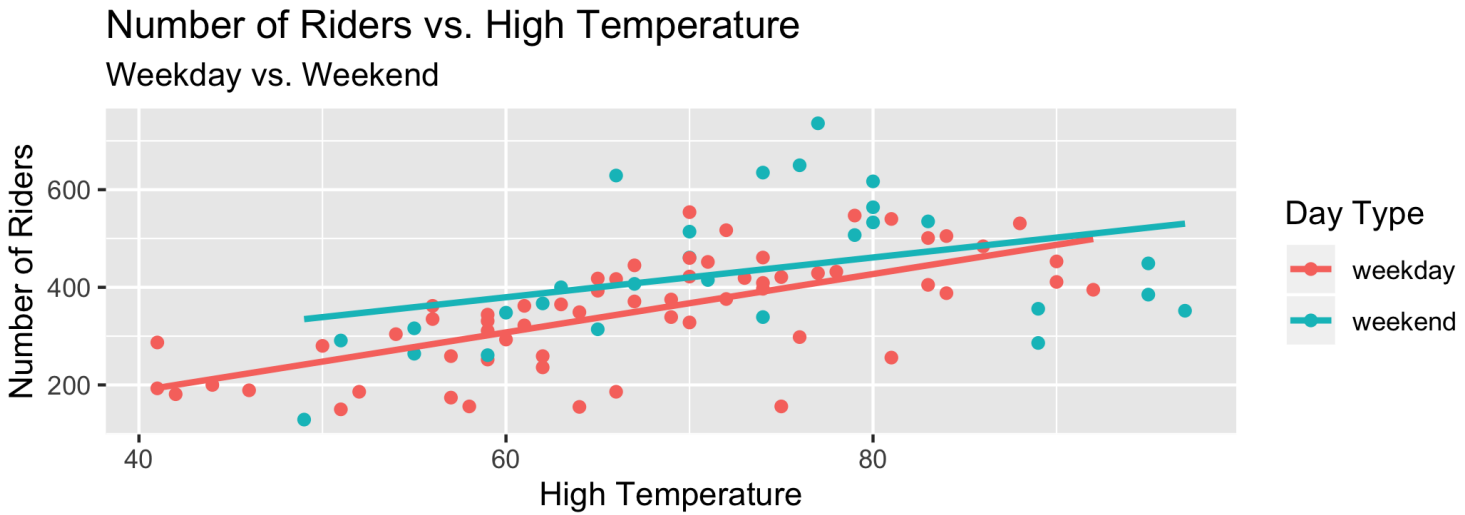
- The way we specified our model only lets **`dayTypeweekend`** affect the intercept.
- Model implicitly assumes that days on the weekend and the weekdays have the *same slope* and only allows for *different intercepts*.
- What seems more appropriate in this case?
  - Same slope and same intercept for both colors
  - Same slope and different intercept for both colors
  - Different slope and different intercept for both colors?

Interacting explanatory variables

- Including an interaction effect in the model allows for different slopes, i.e. nonparallel lines.
- This implies that the regression coefficient for an explanatory variable would change as another explanatory variable changes.
- This can be accomplished by adding an interaction variable - the product of two explanatory variables.

Price vs. hightemp and dayType interaction

```
ggplot(data = RailTrail, aes(x = hightemp, y = volume, color = dayType)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "High Temperature", y = "Number of Riders",
       title = "Number of Riders vs. High Temperature",
       subtitle = "Weekday vs. Weekend",
       color = "Day Type")
```



Modeling with interaction effects

```
m_int <- lm(volume ~ hightemp + dayType + hightemp*dayType,
            data = RailTrail)
kable(tidy(m_int) %>% select(term, estimate), format = "html", digits = 3)
```

term	estimate
(Intercept)	-51.224
hightemp	5.980
dayTypeweekend	186.377
hightemp:dayTypeweekend	-1.906

$$\widehat{volume} =$$
$$-51.224 + 5.980\ hightemp + 186.377\ dayTypeweekend - 1.906\ hightemp \times dayT$$

Interpretation of interaction effects

Weekdays

$$\begin{aligned}\widehat{volume} &= -51.224 + 5.980\ hightemp + 186.377 \times 0 - 1.906\ hightemp \times 0 \\ &= -51.224 + 5.980\ hightemp\end{aligned}$$

Weekends

$$\begin{aligned}\widehat{volume} &= -51.224 + 5.980\ hightemp + 186.377 \times 1 - 1.906\ hightemp \times 1 \\ &= -51.224 + 5.980\ hightemp + 186.377 - 1.906\ hightemp \\ &= 135.153 + 4.074\ hightemp\end{aligned}$$

Interpretation of interaction effects

Weekdays:

$$\widehat{volume} - 51.224 + 5.980\ hightemp$$

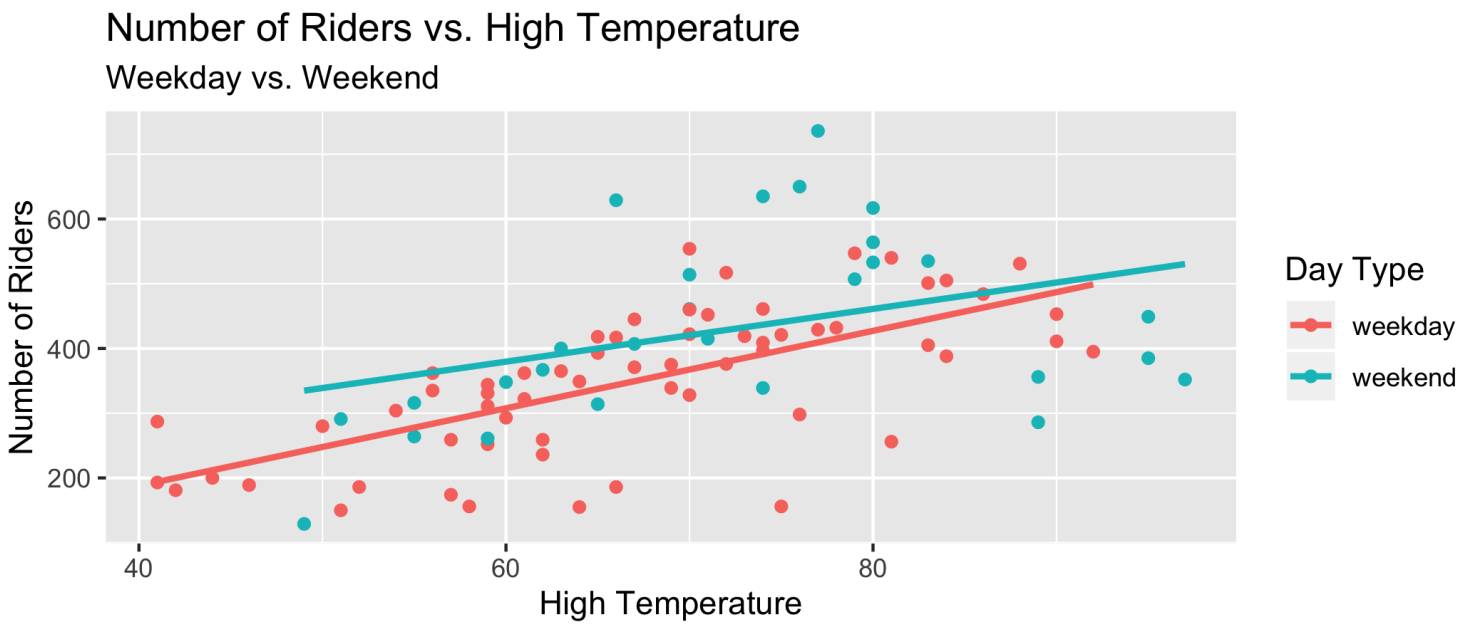
Weekends:

$$\widehat{volume} = 135.153 + 4.074\ hightemp$$

- Rate of change in volume as the high temperature increases is different on days that are weekdays versus those that are weekends (**different slope**)
- Given the same high temperature, days on the weekend are expected to have higher volume than days that are weekdays (**different intercept**).

Interpretation of interaction effects

- Weekdays: $\widehat{volume} - 51.224 + 5.980\ hightemp$
- Weekends: $\widehat{volume} = 135.153 + 4.074\ hightemp$



Third order interactions

- Can you? Yes
- Should you? Probably not if you want to interpret these interactions in context of the data.

Practice

Suppose you wish to fit a model using **`hightemp`** and **`summer`** to predict the number of riders on a trail.

- **`summer`**: 1 if the day is during the summer, 0 otherwise

term	estimate
(Intercept)	-232.432
hightemp	9.294
summer1	576.081
hightemp:summer1	-8.349

1. Interpret the coefficient of **`summer1`**.
2. Interpret the intercept. Is this interpretation meaningful?
3. Write the model equation for days that are not during the summer.
4. Write the model equation for days that are during the summer.

Exam 01

Exam 01

- Exam grades will be released after class
- Code for the exam may be found in the **Resources** folder on Sakai
- Review the code solutions and feedback in Gradescope!
  - Attend office hours if you have any questions about your exam grading.
- Be careful about joins!
  - Only use a join when needed!
  - Joins can be computationally intensive, so make sure you're using the correct one!

Writing better code

- We want to make sure the code we right is not only "correct" but also efficient
- This means
  - Not unnecessarily saving output
  - Not unnecessarily repeating processes
  - Using the simplest code possible!
- Writing code is an iterative process!
  - The first draft isn't always the best draft!

Practice: Writing better code

- Copy the **Writing better code** project in RStudio Cloud
- The file contains two pieces of code that perform the correct calculations but needs revision!
- Rewrite the code so that it performs the same calculations but does so in a simpler and less computationally intensive way.