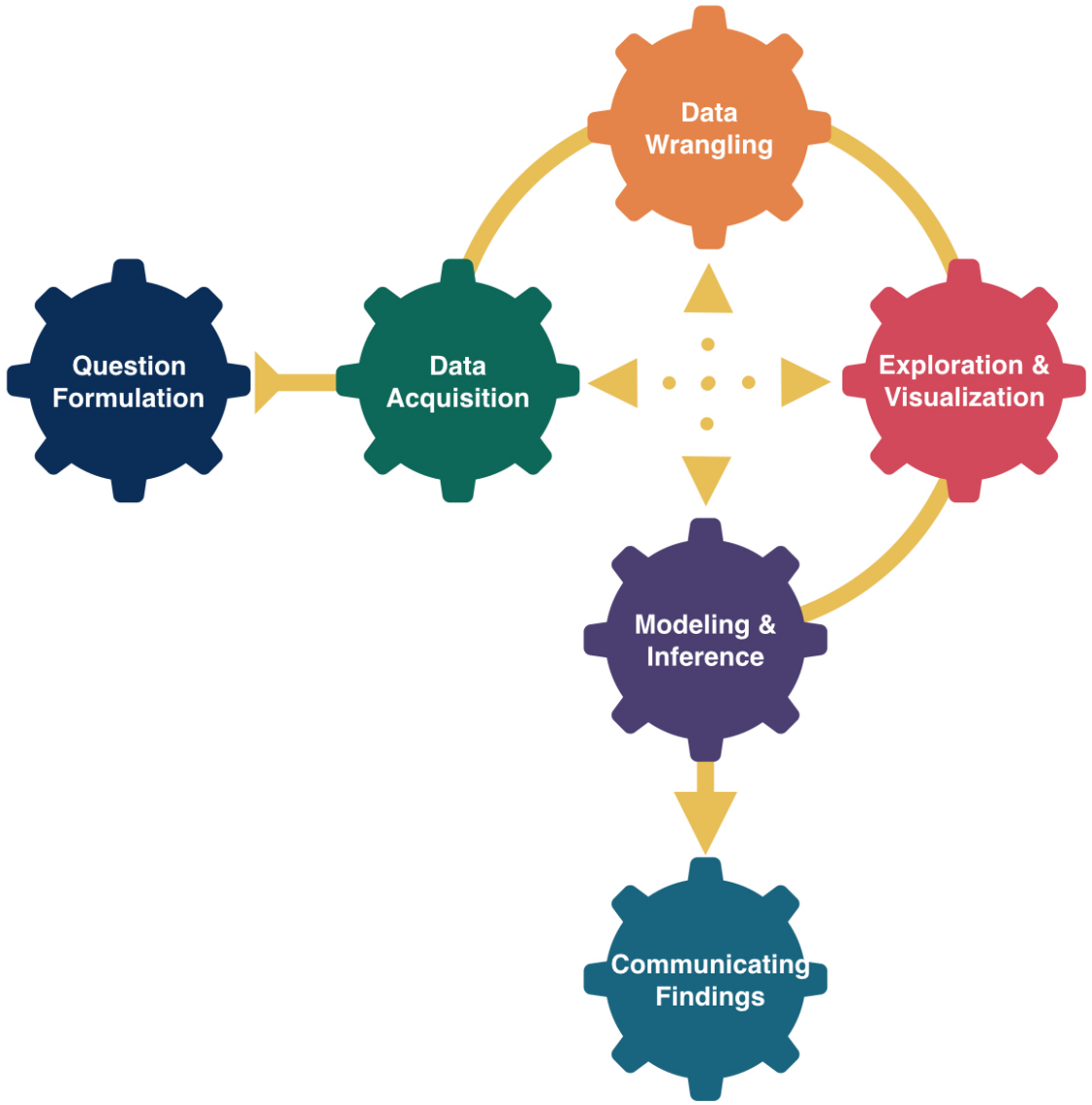


Multiple Linear Regression

Kelly McConville

Stat 100 | Week 6 | Spring 2023

Announcements

- Mid-Term Exam: Wednesday, March 8th - Friday, March 10th
  - Will post the oral exam sign-up sheet after lecture today!
  - Will get a review sheet on Wednesday.
  - Section this week will involve exam review.
- On your p-sets, make sure your answers are your own for **both the coding and narrative components.**

Goals for Today

- Broadening our idea of linear regression models
- Discuss the **multiple** linear regression model
- Explore interaction terms

The Importance of Problem 4 on P-Set 5

"The best thing about being a statistician is that you get to play in everyone's backyard." -- John Tukey

"Consider context. The bottom line for numbers is that they cannot speak for themselves."

"Lacking this context for orientation, strangers in the data set run the risk of getting things entirely wrong or actually doing harm by filling in the missing information with their own biases and assumptions." -- Catherine D'Ignazio and Lauren F. Klein

Linear Regression

Linear regression is a flexible class of models that allow for:

- Both quantitative and categorical explanatory variables.
- Multiple explanatory variables.
- Curved relationships between the response variable and the explanatory variable.
- BUT the **response variable is quantitative**.

- Today's explorations:
  - Including more than one predictor
  - Interaction terms
  - Handling categorical variables with more than two categories

Multiple Linear Regression

Form of the Model:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

How does extending to more predictors change our process?

What doesn't change:

→ Still use Method of Least Squares to estimate coefficients

→ Still use `lm()` to fit the model and `predict()` for prediction

What does change:

→ Meaning of the coefficients are more complicated and depend on other variables in the model

→ Need to decide which variables to include and how (linear term, squared term...)

Multiple Linear Regression

- We are going to see a few examples today and next lecture.
- We will need to return to modeling later in the course to more definitively answer questions about model selection.

Example

Meadowfoam is a plant that grows in the Pacific Northwest and is harvested for its seed oil. In a randomized experiment, researchers at Oregon State University looked at how two light-related factors influenced the number of flowers per meadowfoam plant, the primary measure of productivity for this plant. The two light measures were light intensity (in mmol/m^2 /sec) and the timing of onset of the light (early or late in terms of photo periodic floral induction).

Response variable:

Explanatory variables:

Model Form:

Data Loading and Wrangling

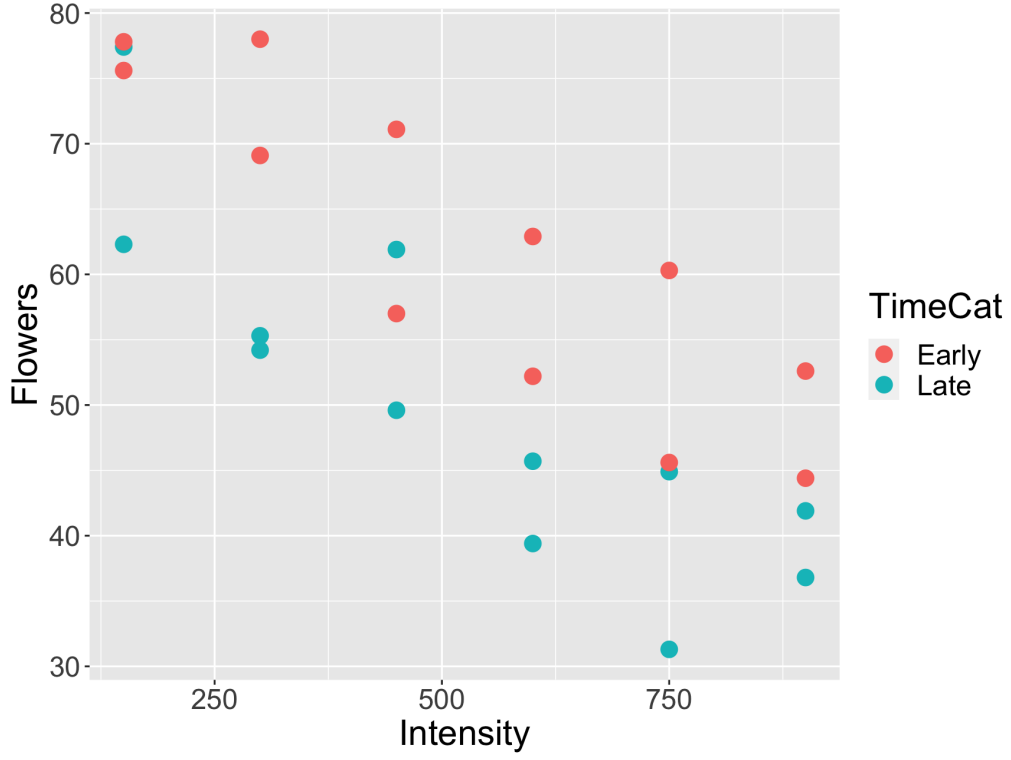
```
library(tidyverse)
library(Sleuth3)
data(case0901)

# Recode the timing variable
case0901 <- case0901 %>%
  mutate(TimeCat = case_when(
    Time == 1 ~ "Late",
    Time == 2 ~ "Early"
    ))
```

Visualizing the Data

```
ggplot(case0901,
       aes(x = Intensity,
           y = Flowers,
           color = TimeCat)) +
  geom_point(size = 4)
```

Why don't I have to include `data =` and `mapping =` in my `ggplot()` layer?



Building the Linear Regression Model

Full model form:

```
modFlowers <- lm(Flowers ~ Intensity + TimeCat, data = case0901)

library(moderndive)
get_regression_table(modFlowers)
```

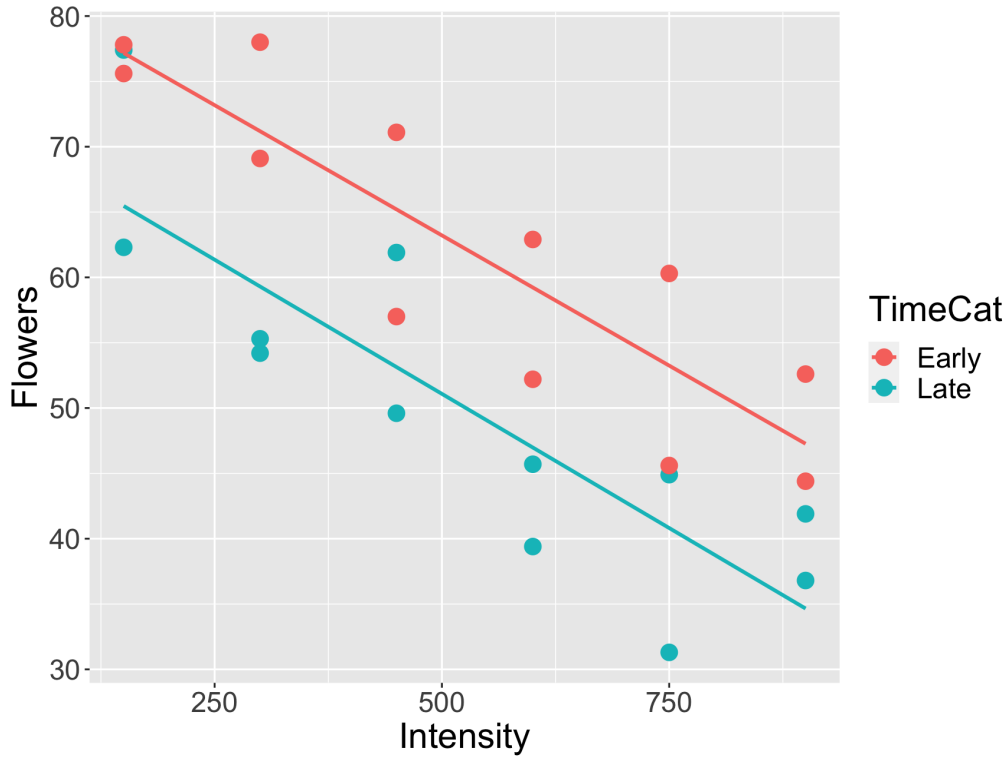
```
## # A tibble: 3 × 7
##   term          estim…¹ std_e…² stati…³ p_value lower…⁴ upper…⁵
##   <chr>           <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 intercept       83.5    3.27    25.5        0  76.7     90.3
## 2 Intensity       -0.04   0.005   -7.89       0  -0.051   -0.03
## 3 TimeCat: Late  -12.2    2.63    -4.62       0 -17.6     -6.69
## # … with abbreviated variable names ¹estimate, ²std_error,
## #   ³statistic, ⁴lower_ci, ⁵upper_ci
```

- Estimated regression line for $x_2 = 1$:

- Estimated regression line for $x_2 = 0$:

Appropriateness of Model Form

```
ggplot(case0901,
       aes(x = Intensity,
           y = Flowers,
           color = TimeCat)) +
  geom_point(size = 4) +
  geom_smooth(method = "lm", se = FALSE)
```



Is the assumption of equal slopes reasonable here?

Prediction

```
flowersNew <- data.frame(Intensity = 700,
                         TimeCat = "Early")
predict(modFlowers, newdata = flowersNew)
```

```
##        1
## 55.13417
```

New Example

For this example, we will use data collected by the website pollster.com, which aggregated 102 presidential polls from August 29th, 2008 through the end of September. We want to determine the best model to explain the variable `Margin`, measured by the difference in preference between Barack Obama and John McCain. Our potential predictors are `Days` (the number of days after the Democratic Convention) and `Charlie` (indicator variable on whether poll was conducted before or after the first ABC interview of Sarah Palin with Charlie Gibson).

```
Pollster08 <-
  read_csv("data/Pollster08.csv")
glimpse(Pollster08)
```

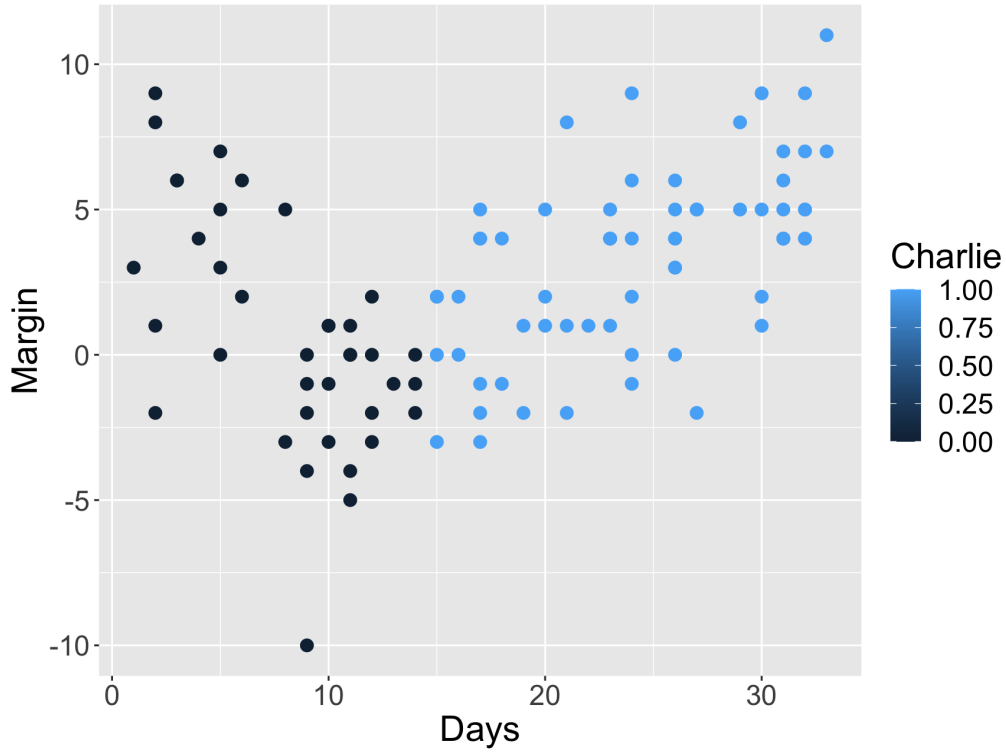
Response variable:

Explanatory variables:

```
## Rows: 102
## Columns: 11
## $ PollTaker <chr> "Rasmussen", "Zogby", "Diageo/Hotline", "CB…
## $ PollDates <chr> "8/28-30/08", "8/29-30/08", "8/29-31/08", "…
## $ MidDate   <chr> "8/29", "8/30", "8/30", "8/30", "8/30", "8/…
## $ Days      <dbl> 1, 2, 2, 2, 2, 3, 3, 4, 5, 5, 5, 5, 6, 6, 8…
## $ n         <dbl> 3000, 2020, 805, 781, 927, 3000, 1200, 1728…
## $ Pop       <chr> "LV", "LV", "RV", "RV", "RV", "LV", "LV", "…
## $ McCain    <dbl> 46, 47, 39, 40, 48, 45, 43, 36, 42, 39, 42,…
## $ Obama     <dbl> 49, 45, 48, 48, 49, 51, 49, 40, 49, 42, 42,…
## $ Margin    <dbl> 3, -2, 9, 8, 1, 6, 6, 4, 7, 3, 0, 5, 2, 6, …
## $ Charlie   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ Meltdown  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
```

Visualizing the Data

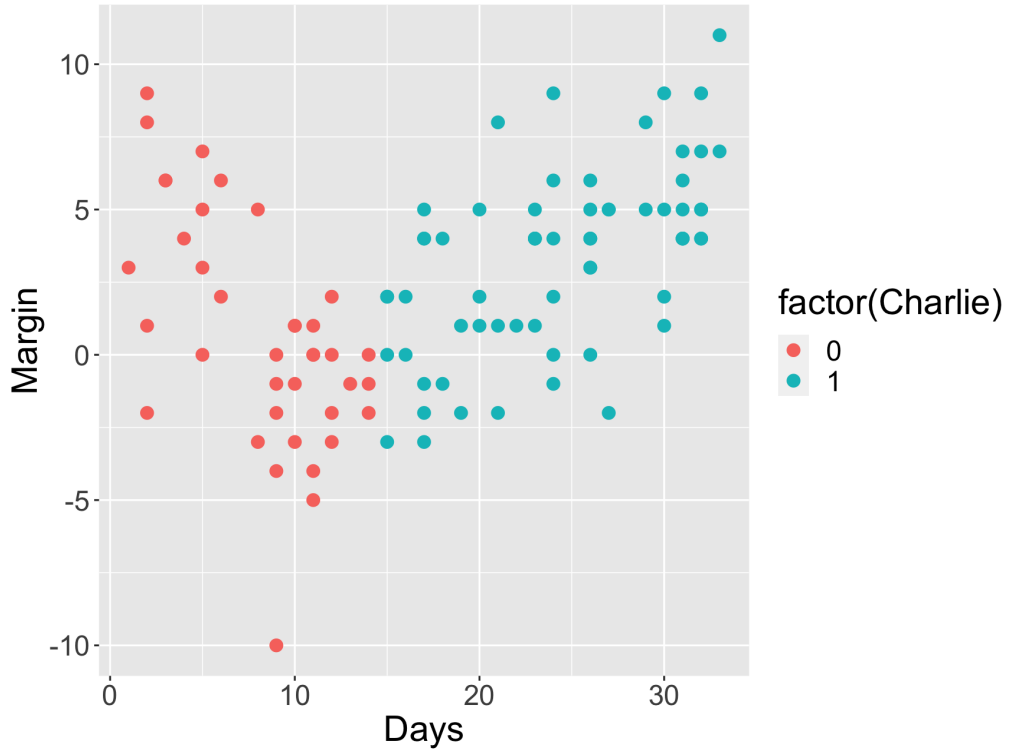
```
ggplot(Pollster08,
       aes(x = Days,
           y = Margin,
           color = Charlie)) +
  geom_point(size = 3)
```



What is wrong with how one of the variables is mapped in the graph?

Visualizing the Data

```
ggplot(Pollster08,
       aes(x = Days,
           y = Margin,
           color = factor(Charlie))) +
  geom_point(size = 3)
```



Is the assumption of equal slopes reasonable here?

Model Forms

Same Slopes Model:

Different Slopes Model:

- Line for $x_2 = 1$:

- Line for $x_2 = 0$:

Fitting the Linear Regression Model

```
modPoll <- lm(Margin ~ Days*factor(Charlie), data = Pollster08)

get_regression_table(modPoll)
```

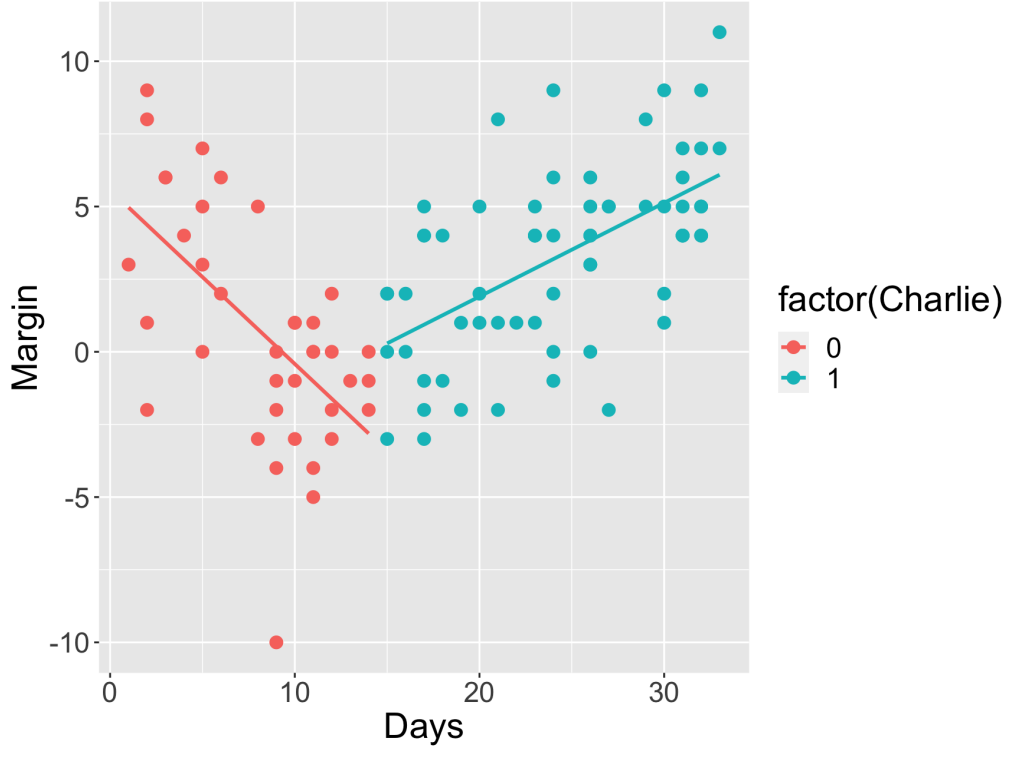
```
## # A tibble: 4 × 7
##   term          estim…¹ std_e…² stati…³ p_value lower…⁴ upper…⁵
##   <chr>           <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 intercept        5.57    1.09     5.11       0   3.40    7.73
## 2 Days            -0.598   0.121   -4.96       0  -0.838  -0.359
## 3 factor(Charl… -10.1     1.92    -5.25       0 -13.9    -6.29
## 4 Days:factor(…   0.921   0.136    6.75       0   0.65    1.19
## # … with abbreviated variable names ¹estimate, ²std_error,
## #   ³statistic, ⁴lower_ci, ⁵upper_ci
```

- Estimated regression line for $x_2 = 1$:

- Estimated regression line for $x_2 = 0$:

Adding the Regression Model to the Plot

```
ggplot(Pollster08,
       aes(x = Days,
           y = Margin,
           color = factor(Charlie))) +
  geom_point(size = 3) +
  stat_smooth(method = lm, se = FALSE)
```



Is our modeling goal here predictive or descriptive?

New Example: Movies

Let's model a movie's critic rating using the audience rating and the movie's genre.

```
library(tidyverse)
movies <- read_csv("https://www.lock5stat.com/datasets2e/HollywoodMovies.csv")

# Restrict our attention to dramas, horrors, and actions
movies2 <- movies %>%
  filter(Genre %in% c("Drama", "Horror", "Action")) %>%
  drop_na(Genre, AudienceScore, RottenTomatoes)
glimpse(movies2)
```

```
## Rows: 313
## Columns: 16
## $ Movie            <chr> "Spider-Man 3", "Transformers", "Pir…
## $ LeadStudio       <chr> "Sony", "Paramount", "Disney", "Warn…
## $ RottenTomatoes   <dbl> 61, 57, 45, 60, 20, 79, 35, 28, 41, …
## $ AudienceScore    <dbl> 54, 89, 74, 90, 68, 86, 55, 56, 81, …
## $ Story            <chr> "Metamorphosis", "Monster Force", "R…
## $ Genre            <chr> "Action", "Action", "Action", "Actio…
## $ TheatersOpenWeek <dbl> 4252, 4011, 4362, 3103, 3778, 3408, …
## $ OpeningWeekend   <dbl> 151.1, 70.5, 114.7, 70.9, 49.1, 33.4…
## $ BOAvgOpenWeekend <dbl> 35540, 17577, 26302, 22844, 12996, 9…
## $ DomesticGross    <dbl> 336.53, 319.25, 309.42, 210.61, 140.…
## $ ForeignGross     <dbl> 554.34, 390.46, 654.00, 245.45, 117.…
```

How should we encode a categorical variable with more than 2 categories?

Let's start with what NOT to do.

Equal Slopes Model:

How should we encode a categorical variable with more than 2 categories?

What we should do instead.

Equal Slopes Model:

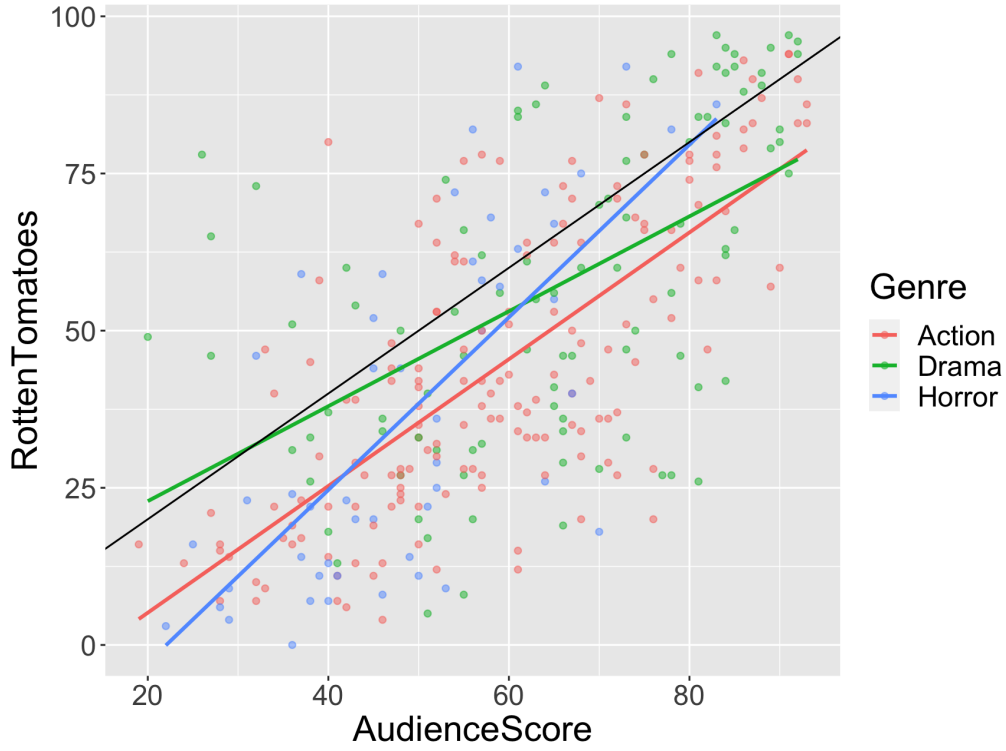
How should we encode a categorical variable with more than 2 categories?

Different Slopes Model:

Exploring the Data

```
ggplot(data = movies2,
       mapping = aes(x = AudienceScore,
                     y = RottenTomatoes,
                     color = Genre)) +
  geom_point(alpha = 0.5) +
  stat_smooth(method = lm, se = FALSE) +
  geom_abline(slope = 1, intercept = 0)
```

- Trends?
- Should we include interaction terms in the model?



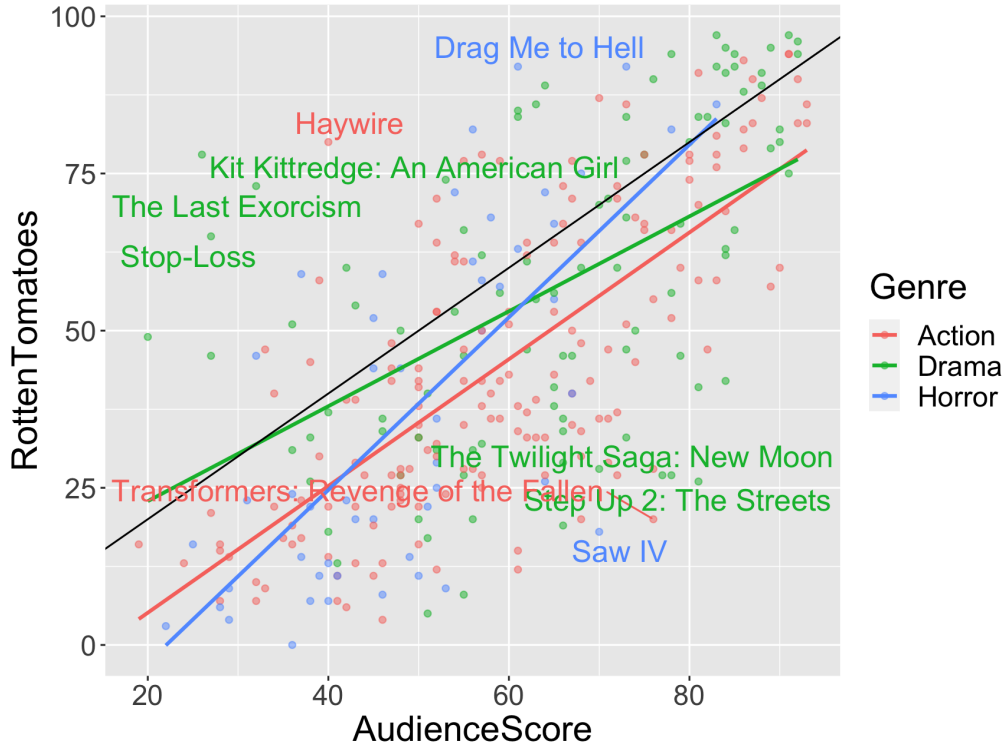
Side-bar: Identify Outliers on a Graph

```
outliers <- movies2 %>%
  mutate(DiffScore = AudienceScore - RottenTomatoes) %>%
  filter(DiffScore > 50 | DiffScore < -30) %>%
  select(Movie, DiffScore, AudienceScore, RottenTomatoes, Genre)
outliers
```

```
## # A tibble: 9 × 5
##   Movie                              DiffS…¹ Audie…² Rotte…³ Genre
##   <chr>                                <dbl>   <dbl>   <dbl> <chr>
## 1 Saw IV                                  52      70      18 Horr…
## 2 Step Up 2: The Streets                  55      81      26 Drama
## 3 Kit Kittredge: An American Girl        -52      26      78 Drama
## 4 Stop-Loss                              -38      27      65 Drama
## 5 Transformers: Revenge of the F…         56      76      20 Acti…
## 6 The Twilight Saga: New Moon             51      78      27 Drama
## 7 Drag Me to Hell                        -31      61      92 Horr…
## 8 The Last Exorcism                      -41      32      73 Drama
## 9 Haywire                                -40      40      80 Acti…
## # … with abbreviated variable names ¹DiffScore,
## #   ²AudienceScore, ³RottenTomatoes
```

Side-bar: Identify Outliers on a Graph

```
library(ggrepel)
ggplot(data = movies2,
       mapping = aes(x = AudienceScore,
                     y = RottenTomatoes,
                     color = Genre)) +
  geom_point(alpha = 0.5) +
  stat_smooth(method = lm, se = FALSE) +
  geom_abline(slope = 1, intercept = 0) +
  geom_text_repel(data = outliers,
                  mapping = aes(label =
                                  Movie),
                  force = 10,
                  show.legend = FALSE,
                  size = 6)
```



Building the Model:

Full model form:

```
mod <- lm(RottenTomatoes ~ AudienceScore*Genre, data = movies2)

library(moderndive)
get_regression_table(mod)
```

```
## # A tibble: 6 × 7
##   term          estim…¹ std_e…² stati…³ p_value lower…⁴ upper…⁵
##   <chr>           <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 intercept     -15.0     5.27    -2.85   0.005 -25.4    -4.67
## 2 AudienceScore   1.01    0.085   11.8    0       0.84    1.18
## 3 Genre: Drama   22.8     8.94     2.55   0.011   5.23   40.4
## 4 Genre: Horror -15.2    11.0     -1.39   0.165 -36.8     6.32
## 5 AudienceScor…  -0.253   0.136   -1.86   0.065  -0.522   0.015
## 6 AudienceScor…   0.365   0.206    1.77   0.078  -0.04    0.771
## # … with abbreviated variable names ¹estimate, ²std_error,
## #   ³statistic, ⁴lower_ci, ⁵upper_ci
```

Estimated model for Dramas:

Will consider adding curvative and get some practice on Wednesday!

Reminders:

- Mid-Term Exam: Wednesday, March 8th - Friday, March 10th
  - Will post the oral exam sign-up sheet after lecture today!