Week 3 Statistical computing

MS 276

September, 2017

The Bechdel test

Today we're going to look at data in the `fivethirtyeight` package (via Albert Kim), which contains data sets relevant to articles on fivethirtyeight.com. One particular article will motivate today's class is found at https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/, titled *The Dollar-And-Cents Case Against Hollywood's Exclusion of Women* and written by Walt Hickey. In it, Hickey writes

```
One of the most enduring tools to measure Hollywood's gender bias
is a test originally promoted by cartoonist Alison Bechdel in a 1985
strip from her "Dykes To Watch Out For" series. Bechdel said that if
a movie can satisfy three criteria - there are at least two named women
in the picture, they have a conversation with each other at some point,
and that conversation isn't about a male character - then it passes
"The Rule," whereby female characters are allocated a bare minimum of
depth. You can see a copy of that strip here.
```

In today's class, we'll replicate part of this project.

Let's take a look:

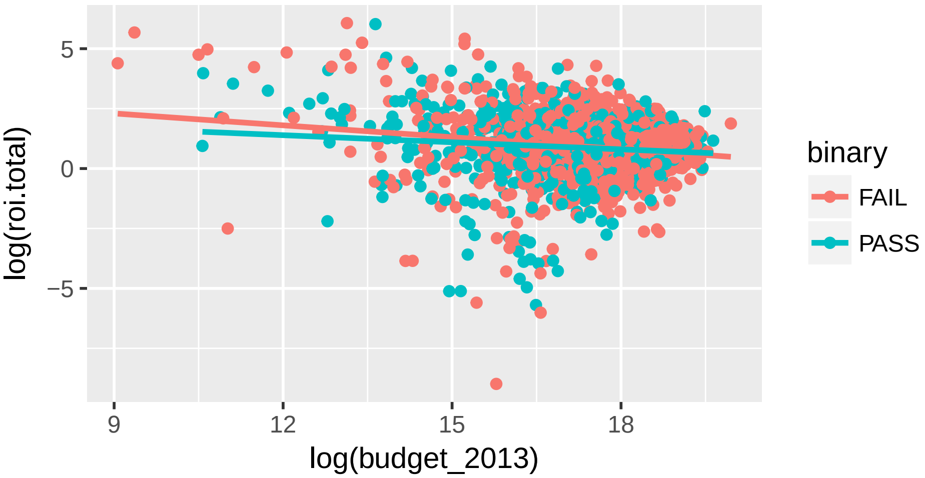
```r
library(fivethirtyeight)
glimpse(bechdel)
```

```
## Observations: 1,794
## Variables: 15
## $ year          <int> 2013, 2012, 2013, 2013, 2013, 2013, 2013, 2013, ...
## $ imdb          <chr> "tt1711425", "tt1343727", "tt2024544", "tt127287...
## $ title         <chr> "21 & Over", "Dredd 3D", "12 Years a Slave", "2 ...
## $ test          <chr> "notalk", "ok-disagree", "notalk-disagree", "not...
## $ clean_test    <fctr> notalk, ok, notalk, notalk, men, men, notalk, o...
## $ binary        <chr> "FAIL", "PASS", "FAIL", "FAIL", "FAIL", "FAIL", ...
## $ budget        <int> 13000000, 45000000, 20000000, 61000000, 40000000...
## $ domgross      <dbl> 25682380, 13414714, 53107035, 75612460, 95020213...
## $ intgross      <dbl> 42195766, 40868994, 158607035, 132493015, 950202...
## $ code          <chr> "2013FAIL", "2012PASS", "2013FAIL", "2013FAIL", ...
## $ budget_2013   <int> 13000000, 45658735, 20000000, 61000000, 40000000...
## $ domgross_2013 <dbl> 25682380, 13611086, 53107035, 75612460, 95020213...
## $ intgross_2013 <dbl> 42195766, 41467257, 158607035, 132493015, 950202...
## $ period_code   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ decade_code   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
```

Referring to the article, make a list of the tranformation steps that the article makes, and think about how you could use `dplyr` to make this code yourself.

Goals for class

1. Double check that there are 1794 films between 1970 and 2013.
2. Create a new variable, `generous`, which is TRUE if either the film passed (`clean_test` == ok) or came close to passing (`clean_test` == dubious), and check that 53 percent of films passed the Bechdel test using this variable.
3. Replicate the (rough) findings of `The Bechdel Test Over Time` figure, at least as far as percentages. In other words, use one of the time variables to estimate that the fraction of movies has increased over time, but that this number has not increased in the last few years.
4. Calculate the median return on investment for each `generous` group. Note that return on investment for a movie is defined as `intgross_2013 / budget_2013`, where use inflation-adjusted numbers for gross and budget.
5. Why is median a better metric than mean?
6. Calculate the median budget for films since 1990, replicating the numbers shown in the chart `Median budget for films since 1990`.
7. What does the following chart show? And why is log(budget) and log(roi) use in terms of the original variables?



7. Answer any question of your own using the data.