Categorical Predictors and Interactions

Lecture 11 Handout

Statistics 139

Topics

- Categorical Predictors
- Extra-sums-of-squares (ESS) F-test
- Interaction Terms

The material in this handout corresponds to the Lecture 11 Notes.

In this lab we will be using the 'movies17.csv' data set to investigate the use of categorical variables and interaction terms in regression. This data set contains several measurements for many of the mainstream movies from 2017. We will use this data set to answer the following questions:

1. How does movie revenue relate to the production cost of movies?
2. Which of the major studios has the most grossing revenue?
3. Which studio takes advantage of production cost the most efficiently?

Variables useful for today's handout include (there are many more in the data set):

- `totalgross`: the total domestic gross revenue for the movie in the US (in US $).
- `budget`: the total production cost of the movie (in US $).
- `studio`: the studio that produced the movie. 6 categories: Fox, Paramount, Sony, Universal, WarnerBros, and Other.
- `sequel`: a binary variable to indicate whether the movie is a sequel or part of a franchise of movies.
- `month`: a categorical variable for the month when the movie was released.

Question 1: Explorations

a) Begin by changing the scale on the financial measurements from dollars to millions of dollars (divide by 1000000). Why might this be a better choice?

```r
# read in the data set and transform data
movies = read.csv("data/movies17.csv")
```

b) Investigate the distribution of movies across studios. What do you notice?

```
# visualize and/or summarize across groups
```

c) Investigate the distributions of each of the financial variables and describe what you see. Should you transform any of these variables?

```
# look at distributions
```

Question 2: Basic Modeling

a) Fit a linear model to predict `log(totalgross)` from `log(budget)` and provide a visual to support this model's results. Interpret the results.

```
# fit a model and interpret
```

b) Change the 'order' of the categories for the variable studio so that the 'other' group is the reference group (the `relevel` command may be useful). Why might this be a reasonable choice?

```
# use relevel
```

c) Fit a linear model to predict `log(totalgross)` from `studio` and provide a visual to support this model's results. Interpret the results.

```
# fit a model and visualize
```

d) Fit a linear model to predict `log(totalgross)` from `month` and provide a visual to support this model's results. Interpret the results.

```
# fit a model and visualize
```

e) Fit a linear model to predict `log(totalgross)` from `sequel` and provide a visual to support this model's results. Interpret the results.

```
# fit a model and visualize
```

f) Compare the residuals (both visually and quantitatively) from the model in part (a) across the various movie studios. Interpret the results.

```
# compare residuals across studios
```

g) Fit a linear model to predict `log(totalgross)` from `studio` and `log(budget)`. Interpret the results and describe what a visual to illustrate the model results would look like.

```
# fit a model and visualize
```

h) How do the previous two parts agree? In what ways are they different? Hint: what does model in part (g) take into account that the approach in part (f) does not?

Question 3: Interactions

a) Within the subset of 'other' studios, fit a regression model to predict `log(totalgross)` from `log(budget)`. Calculate this same model but for movies from the 'Fox' studio only.

```
# fit 2 models
```

b) Perform an appropriate test to determine whether the slopes in the two subgroup models in the previous part are equal or not.

```
# perform a test
```

c) Fit a regression model to predict `log(totalgross)` from `log(budget)`, `studio`, and the interaction between the two (for all studios in the original data set), and briefly interpret the results.

```
# fit a model and interpret
```

d) Create a plot to illustrate the results from the previous part.

```
# create a visual
```

e) Compare the results of the model in the previous part(s) to the models calculated within the subsets in part (a). Do they agree? Which approach is preferred? Why?

f) Formally test whether the inclusion of the interaction terms contribute a statistically significant amount to R^2.

```
# perform a test
```