Movie Ratings

2024-03-28

Rotten Tomatoes Movies

The dataset "movieratings.csv" contains critic ratings (Tomato Meter) and audience ratings (Audience) for 200 random movies found in the Rotten Tomato movie database.

Our goals are to:

1. understand the relationship between critic and audience scores,
2. fit a (linear) model to summarize this relationship, and
3. use it to predict audience ratings for movies which have evaluated by critics.

Throughout this .Rmd file, fill in the blank code chunks with the relevant code to complete the task described in the comment.

```
# read in the movieratings.csv file
```

```
#look at the first 6 rows
```

Exploratory Data Analysis / Visualization / Cleaning

```
#Give a 5 number summary for
#1. the critic ratings and
#2. the audience ratings.
```

```
#Plot a histogram for
#1. the critic ratings and
#2. the audience ratings.
```

Let's try to plot critic rating (predictor, on the x-axis) vs. audience rating (response, on the y-axis) for each of our movies.

```
#load ggplot2 and tidyverse



#use ggplot to make a scatter plot of critic vs. audience ratings

#make sure to rename the x- and y-axes appropriately
```

Before we do anything else, first notice the warning! 3 of our points couldn't be plotted. Normally this happens when there is missing data or the points are outside the range of the plot. In this case its the former:

```
#use the complete.cases function to find the rows with missing data
# for help, use ?complete.cases

#How many incomplete cases are there? Which movies are they?
```

```
#create another df which only contains the complete cases
```

Let's go back to our plot. What kind of pattern is there?

```
#make scatterplot and describe the relationship between critic and audience scores.

#Answer the following questions:
#1. Is the relationship  between critic scores and audience scores positive or negative?
#2. Is the trend linear or non-linear? Is it appropriate to use linear regression to model this data?
#3. How strong is the relationship between critic scores and audience scores?


#To help with number 3, you can
#calculate the "correlation" between critic scores and audience scores
#using the cor() function.

#We will talk about what this means in class.
```

Making a Model

We want to fit a line to approximate the relationship between scores, i.e.:

```
#Use geom_smooth to add the least squares line to the plot
```

This is linear regression! To get to this point, we have to first define what we mean by a regression model.

Def: A regression model is a function m that describes the relationship between the predictor (x) and the response variable (y).

Idea:

reality = model + error

General regression model:

$$y = m(x) + \epsilon$$

In the linear regression context, we make the assumption that the functional form m should be a simple line:

$$m(x) = ax + b$$

so altogether we have:

$$y = ax + b + \epsilon$$

This is the **simple linear regression model**. The values $a, b $ are fixed, unknown parameters to be estimated, while y and ϵ are random variables.

How does this line up with the Rotten Tomatoes plot?

```
#fit linear model using the lm function
```

```
#call the summary of the lm object
```

```
#add additional columns to the complete ratings df which contain 1. the predicted values and 2. the err
```

```
#Add red lines for the residuals onto the existing scatterplot
```

reality = model + error If a, b are unknown, how do we estimate them? We're looking for the best choices of their estimators: $\hat{a}, \hat{b}$. Then we could predict the responses as $\hat{y} = \hat{a}_0 + \hat{b}_1 x$. But which values of $\hat{a}_0, \hat{b}_1$ are best?

Intuitively we want to pick the line that is closest to the datapoints as possible... let's recall how we defined the residual to solidify this idea.

Def: The **residual** is the difference between the observed response and the predicted response under the model.

$$e = y - \hat{y}$$

The least squares principle says we want to choose the estimators which minimize the sum of squared residuals over the observed sample, where i denotes which data point in the sample we're considering and $i = 1, ..., n$:

$$SSE = e_1^2 + e_2^2 + ... + e_n^2 = \sum_{i=1}^{n} e_i^2$$

Since the residuals depend on $\hat{y}$ and that depends on the choices of a and b, we can consider the sum of squared errors as a function of a and b:

$$SSE = Q(a, b) = \sum_{i=1}^{n} (y_i - ax - b)^2$$

The choices of a and b that we want are the ones that minimize this quantity. We will call these minimizers $\hat{a}$ and $\hat{b}$ since they are estimating the true a and b.

For now we'll just let R solve this problem– I'll show you a derivation later if you want!

```
# What are the choices of a and b which minimize SSE?
```

Now, write out the model in equation form:

$$... = ...$$

Interpreting the linear model

Fill in the proper interpretations:

Slope: For every increase in _____, we expect _____, on average.

Intercept: If the _____ is 0 points, we expect _____.

The slope is generally interpretable most of the time. The intercept however can be tricky.

Only interpret the intercept if: - the predictor can actually take values equal to or near zero, - there are values near zero in the observed data, and - the result makes practical/physical sense.

Making a prediction

"Barbie" received a Tomatometer score of 88. What is the predicted audience score, using our fitted linear regression model?

If the IRL audience score was actually 83, what's the residual for our prediction?

Now use the `predict()` function to provide predictions for 5 movies which have critic scores of: 88, 55, 22, 33 and 66.