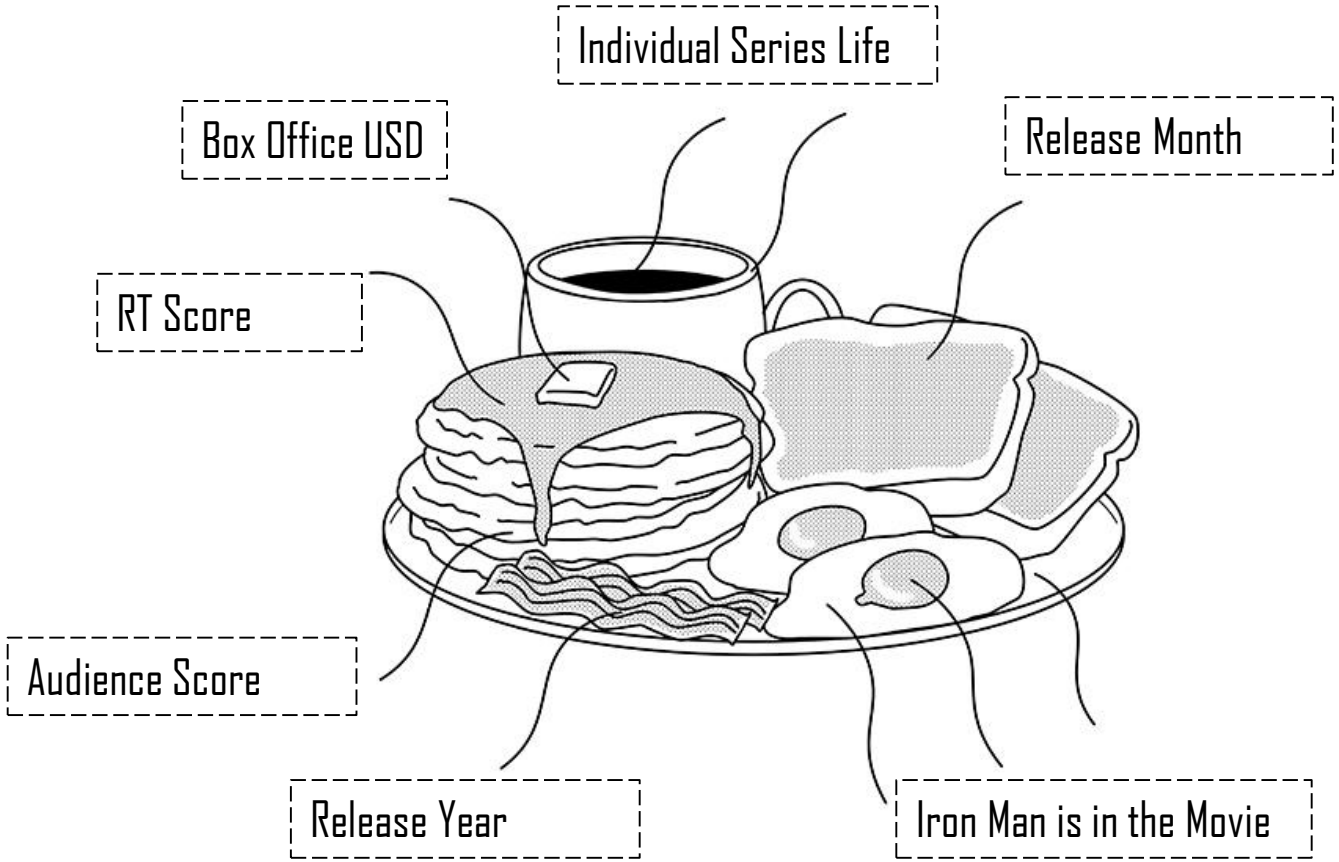


Fig. 002 Marvel Cinematic Universal Multiple Linear Regression

Marvel Cinematic Universal Multiple Linear Regression

This recipe card is a quick guide to the topics covered in this post. The goal is to fit a regression model to Box Office USD for Marvel Cinematic Movie releases.
**At the time of cooking Ant-man and the Wasp did not have finalized Box Office USD data (This movie was excluded.)*

INGREDIENTS:

Box Office USD
Release Month
Rotten Tomatoes Audience Score
Months Since Last MCU Release
Individual Series Life
Rotten Tomatoes Critics Score
Release Year
Is Iron Man in the Move?

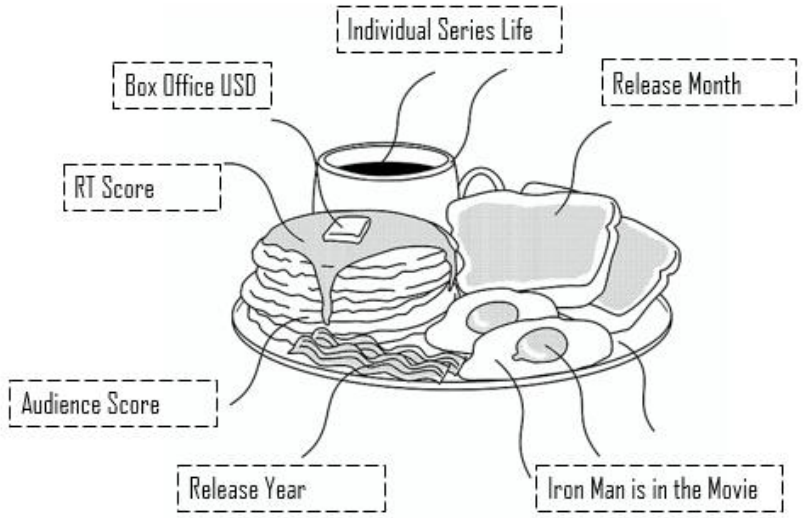


Fig. 002 Marvel Cinematic Universal Multiple Linear Regression

DIRECTIONS:

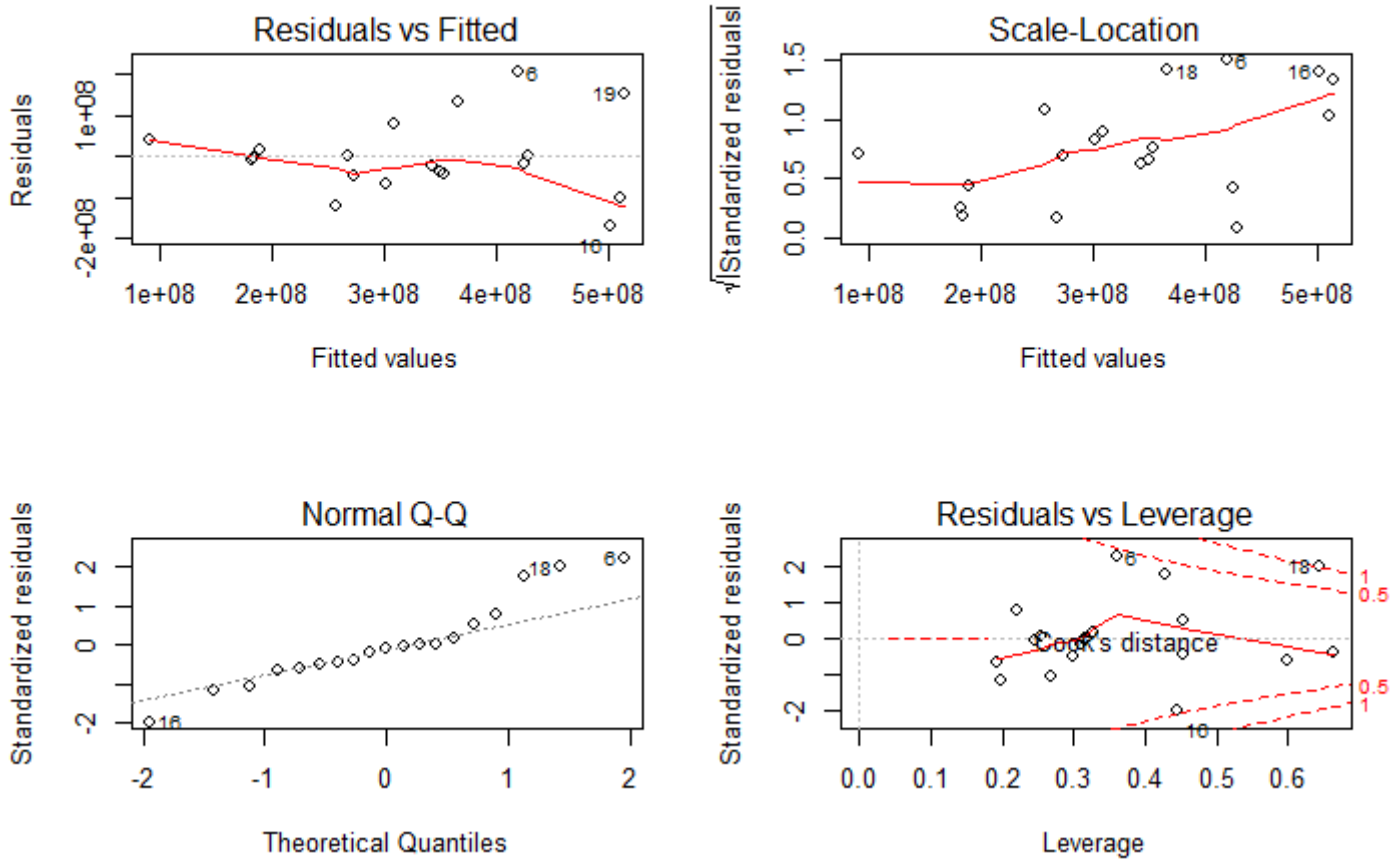
1. MCU Data extraction, transformation and load into R Studio
2. Evaluate data and understand confounding factors
3. Evaluate fit summary and plots
4. If necessary remodel (based on plots, how does your line fit to your data?)
5. Instead of going out for Shawarma, grab a stack of pancakes with the Avengers

```
Residuals:
       Min         1Q     Median         3Q        Max
-126182362  -34315276   -6868506   32875705  210246499

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.840e+10  2.209e+10  -1.738   0.1100
x1           1.624e+08  4.700e+08   0.346   0.7362
x2           1.156e+08  5.872e+08   0.197   0.8476
x3           1.912e+07  1.107e+07   1.727   0.1122
x4          -1.715e+07  1.140e+07  -1.504   0.1607
x5           1.153e+06  5.638e+06   0.204   0.8417
x6           1.435e+07  4.079e+07   0.352   0.7317
x7           1.408e+08  6.631e+07   2.124   0.0572 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.08e+08 on 11 degrees of freedom
Multiple R-squared:  0.6894,    Adjusted R-squared:  0.4918
F-statistic: 3.488 on 7 and 11 DF,  p-value: 0.03185
```

Time to view our plots perhaps?

There was an idea, to bring together a group of extraordinarily Marvel Cinematic Universe data points to build an algorithm when no other data would do.
Now, I know this isn't the same as getting to play director of shield but let's have some fun (Warner Brothers take note ☺).
I fit into a multi-linear regression the following variables: Box Office USD (dependent variable), Rotten Tomatoes Critics Score, Rotten Tomatoes Audience Score, Movie Release Year, Movie Release Month, Months Since the previous MCU movie was released, Release within individual franchise and my favorite variable "Was Iron Man in the Movie?" .

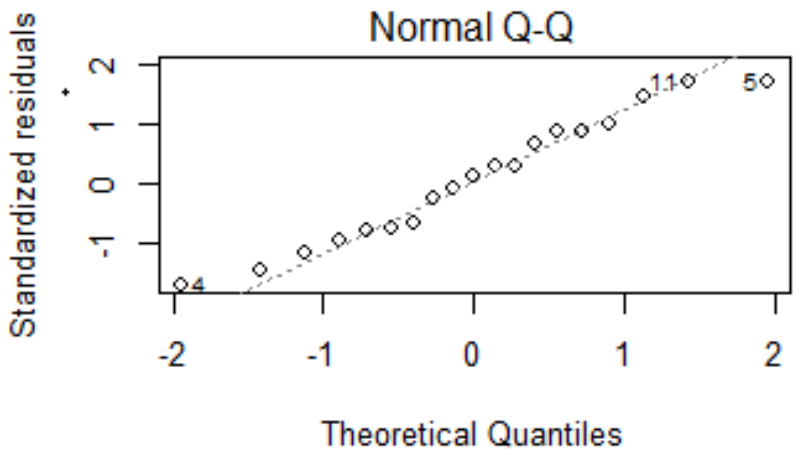
Notable insights from my fit summary is "Was Iron Man in the Movie?" shows the most significance (lowest P value) and this model isn't better than flipping a coin (Adjusted R-squared: .4918).



As I analyze the plots output it's clear this model is not fitting well. For determining fit I tend lean towards the Q plot, and the curve it is showing leads to believe there isn't a true linear relationship between Box Office USD and the variables I have on hand.

My step will be to attempt a quadratic approach and put some weight on variables I want to add some extra value too. Since this is movie sales and the MCU has stretched over multiple years, I can't help but wonder what economic factors are at play which may be confounding my data?

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -366990732  407189639  -0.901    0.434
x1            188320424  556574616   0.338    0.757
x2            523501437  653230744   0.801    0.481
poly(x3, 8)1  200489169  138018705   1.453    0.242
poly(x3, 8)2  215479578  151036126   1.427    0.249
poly(x3, 8)3  216805603  117628064   1.843    0.163
poly(x3, 8)4  229992909  150294221   1.530    0.223
poly(x3, 8)5 -106212514  104457074  -1.017    0.384
poly(x3, 8)6  125829768  136272074   0.923    0.424
poly(x3, 8)7   47238551  154883575   0.305    0.780
poly(x3, 8)8 -120182902  106391011  -1.130    0.341
x4               265995   16079895   0.017    0.988
x5              9007400    9645861   0.934    0.419
poly(x6, 2)1   84069437  141483420   0.594    0.594
poly(x6, 2)2  -85081845  147477456  -0.577    0.604
x7            100512589   69425353   1.448    0.243

Residual standard error: 86380000 on 3 degrees of freedom
Multiple R-squared:  0.9458,    Adjusted R-squared:  0.6751
F-statistic: 3.493 on 15 and 3 DF,  p-value: 0.1654
```



I've decided to add weight to the Movie Release Year and the placement within the movie's own individual franchise. I'm making the assumption the last movie in a trilogy on average earns more than the introductory movie.

The model is now more complex and there's a higher value on the Movie Release Year. A causality of this method (sorry RDJ) is the variable "Was Iron Man in the Movie?" has lost some significance.

Reviewing my Q plot, and I see this regression model fits better to our data (line through our snake, Hail Hydra!). How about R-squared? Our Multiple R-squared is nearly .95, but for a more practical use we'll use the adjusted R-squared of 0.6751. A quick note on P values and significance as a whole, it's really more of guide and point of view, up to you the analyst to decide on value.

Movie Release YR: 0.163
Iron Man P Value: 0.243

Adjusted R-Squared:
0.68

This case study brings to the light the improvement of storying telling within the MCU and RDJ reaching the Hugh Jackman staple level.

Before we finish our pancakes let's discuss some practical uses of this regression model:
The variables used can be used in forecasting
A case study later on when Iron Man makes his MCU exit
Can be used to assess "Super Hero Fatigue"?

Thank you for stopping by, I hope you enjoy the movies Disney and Marvel Studios are making and be sure to check out more Geeky Analytics from Pancake Analytics.

As a courtesy from me to you, be sure to check out the below Marvel themed pancake recipe.