Superhero Movie Performance in China: How Profitable are They?

Alison Yao (yy2564)

Superhero movies are one of the highest-grossing movie franchises in the world. China, with a population of almost 1.4 billion, has generated extraordinary box office performance and huge gross for these superhero movies. However, in mainland China, all entertainment content such as TV and films are censored by National Radio and Television Administration (NRTA) and China Film Administration (CFA). China mainland allows merely 34 foreign films into the Chinese market every year, so both the movie production companies (such as Marvel Studios and DC Film), and the CFA make an effort to introduce superhero movies thanks to their profitability for both parties.

As a devoted fan, I will build on the last homework to investigate the relationship between the production budget of these superhero movies, IMDB ratings, Douban ratings, whether the superhero is Marvel or DC (predictors) and the total gross in China (target). I gathered a sample of 34 observations of superhero movies from *The Numbers* (www.the-numbers.com), *Box Office Mojo* (www.boxofficemojo.com), *IMDb* (https://www.imdb.com) and *Douban Movies* (https://movie.douban.com). Please note that lifetime gross is the summation of gross for several releases. Sometimes *The Numbers* lists "Gross" instead of "Lifetime Gross", but they mean the same if a movie is only released once. It is possible that *The Numbers* does not provide the gross in China because a movie might not have made the 34-foreign-film list and was not released in China mainland at all. Occasionally, *The Numbers* does not provide the budget information, so I turned to this list from *Box Office Mojo* (https://www.the-numbers.com/movies/franchise/Marvel-Cinematic-Universe#tab=summary) for budget information. I am including IMDb Rating as an indicator of audience acceptance internationally and mostly outside of China. *Douban Movies* is the Chinese *IMDb*, where Chinese viewers rate and comment on movies, so Douban ratings may be more predictive for the total gross in the China market. The indicator variable of whether the superheroes belong to Marvel or DC is easily ready based on common sense and some Internet Googling. 1 means the superhero(es) in a movie belong(s) to Marvel and 0 means the superhero(es) belong(s) to DC. I chose Marvel and DC because they are the most well-known and the biggest rivals in the superhero movie market. Arguably, Marvel is winning. Due to COVID-19 and the closing of cinemas, I only selected

movies released before 2020. The raw data in the Budget and the Chinese Gross columns are very large, so I scaled them from dollars to million dollars, but this should not change the form of the regression model or any plots or test statistics such as p-values compared to the original unit. The first few observations of the adjusted dataset are shown below:

Year	Movie Title	Budget (Mil$)	Marvel	Douban Rating	IMDb Rating	Chinese Gross (Mil$)
2002	Spider-Man	139	1	7.9	7.3	4.983142
2004	Spider-Man 2	200	1	7.6	7.3	6.102882
2007	Spider-Man 3	258	1	7.5	6.2	18.991487
2008	Iron Man	140	1	8.3	7.9	15.274332
2008	The Incredible Hulk	150	1	7	6.7	9.336251

Here are some basic statistics:

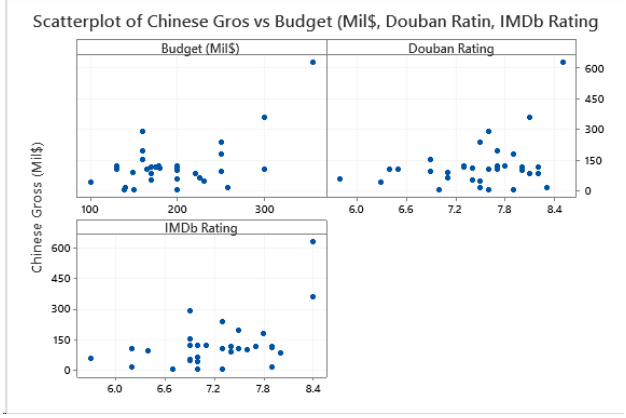
Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Budget (Mil$)	34	0	196.03	9.40	54.78	100.00	160.00	190.00	226.25
Marvel	34	0	0.8235	0.0664	0.3870	0.0000	1.0000	1.0000	1.0000
Douban Rating	34	0	7.482	0.105	0.611	5.800	7.100	7.600	7.925
IMDB Rating	34	0	7.256	0.105	0.612	5.700	6.900	7.300	7.725
Chinese Gross (Mil$)	34	0	120.5	20.5	119.4	5.0	53.7	105.2	121.2

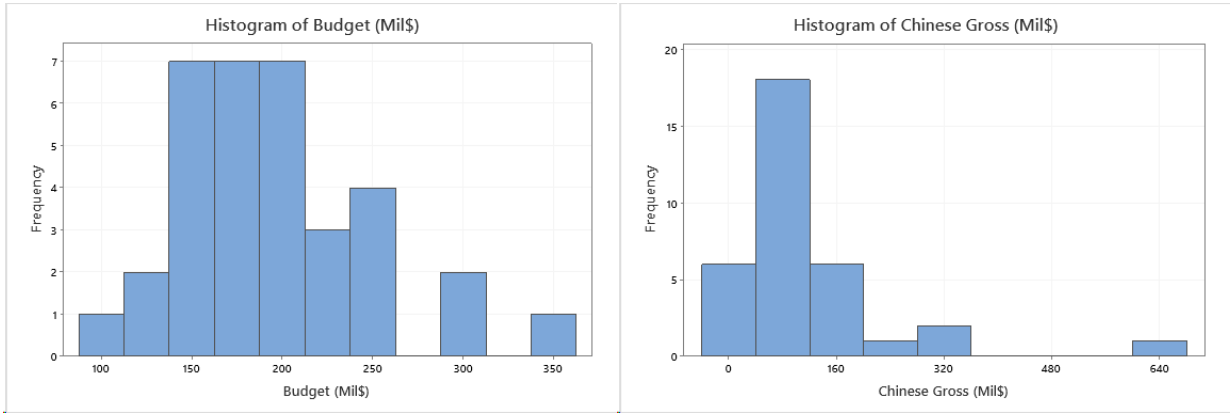
Variable	Maximum
Budget (Mil$)	356.00
Marvel	1.0000
Douban Rating	8.500
IMDB Rating	8.400
Chinese Gross (Mil$)	629.1

Here, the statistics of the indicator variable (Marvel) are less interpretable. A more interpretable statistic is that there are 6 0s and 28 1s because there are many more Marvel movies than DC movies. Other than that, Douban ratings are slightly higher than IMDb ratings, but the standard error of the mean and the standard deviation are almost identical.

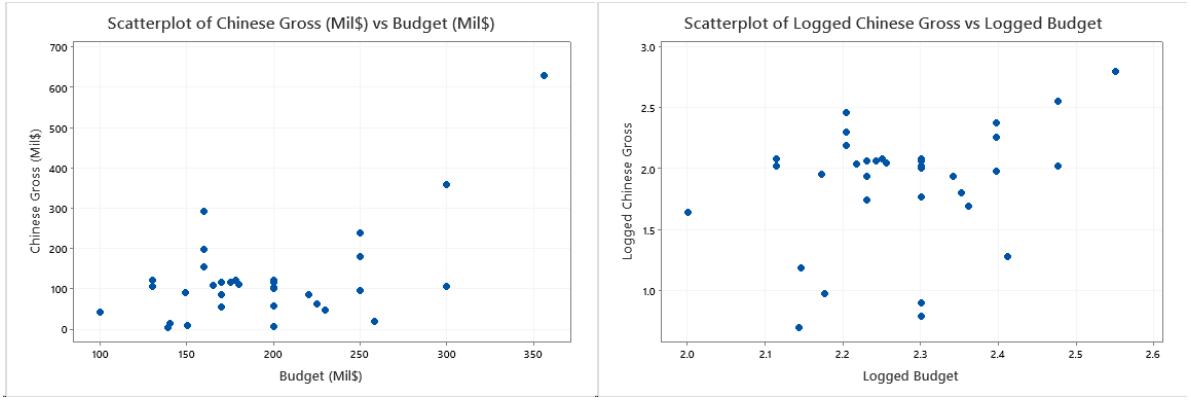
Then, let's look at the scatter plot of Chinese Gross vs Budget, Douban Rating and IMDb rating:



There seems to be a relatively weak relationship between all three predictor-target pairs. Since there are 2 “money data” columns, Budget and Chinese Gross, I plot the histograms here to determine if it would be better to take the logarithm of both.



There seems to be a longer right tail than the left one. Therefore, I will test out taking the log of both Budget and Chinese Gross. Here are the scatter plots of the original one and the logged one.



Visually, there does not seem to be a huge improvement in terms of any stronger linear relationship. The scatter plot of the logged one actually seems to be more scattered and separated into two clusters. Since the scatter plots are not too helpful for me to determine the better model right away, I will fit the data to both models and compare quantitively.

The first model to fit is the simplest model with no logarithm transformation and no consideration of the indicator variable:

$$Chinese\ Gross = \beta_0 + \beta_1 \cdot Budget + \beta_2 \cdot Douban\ Rating + \beta_3 \cdot IMDb\ Rating + random\ error$$

The output from Minitab is as follows:

Regression Equation

Chinese Gross (Mil$) = -630 + 1.159 Budget (Mil$) - 24.6 Douban Rating + 97.5 IMDb Rating

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-630	204	-3.09	0.004	
Budget (Mil$)	1.159	0.290	4.00	0.000	1.03
Douban Rating	-24.6	40.0	-0.62	0.542	2.43
IMDb Rating	97.5	39.6	2.46	0.020	2.39

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
90.0843	48.28%	43.11%	21.67%

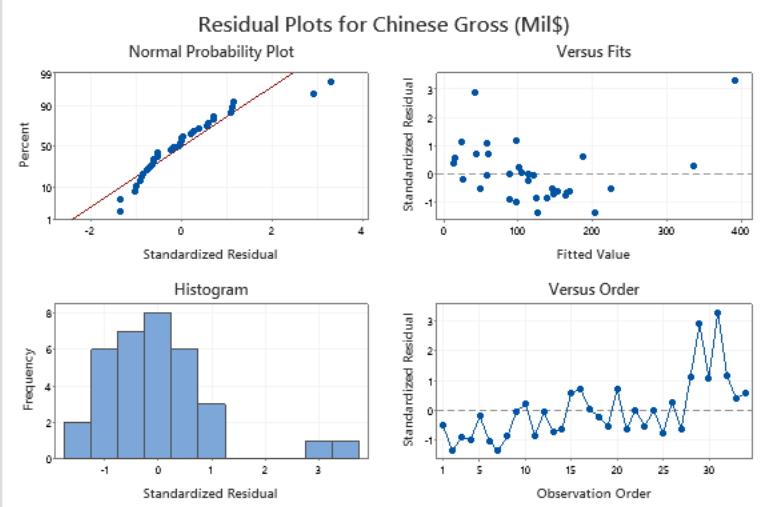
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	227240	75747	9.33	0.000
Budget (Mil$)	1	129760	129760	15.99	0.000
Douban Rating	1	3083	3083	0.38	0.542
IMDb Rating	1	49169	49169	6.06	0.020
Error	30	243455	8115		
Total	33	470696			

Fits and Diagnostics for Unusual Observations

Obs	Chinese Gross (Mil$)	Fit	Resid	Std Resid		
29	291.8	41.1	250.7	2.91	R	
31	629.1	392.5	236.6	3.30	R	X

R Large residual
X Unusual X



The regression is relatively strong, with nearly 50% of the variability accounted for by the multiple regression. Since the VIF values are all small, there is no collinearity problem. The overall F-test and the t-test for Budget are highly statistically significant. The t-test for IMDb

Rating is statistically significant, but the t-test for Douban Rating is not statistically significant, meaning that given Budget and IMDb rating, Douban rating does not add much to the predictive power for the total gross of superhero movies in China, which is not too surprising because we know from the basic statistics that Douban ratings seem to be only slightly higher than IMDb ratings with almost the same standard deviation. The coefficient of Budget means that given the other two variables are held fixed, a million-dollar increase in Budget is associated with an estimated expected increase in Chinese Gross of 1.159 million dollars; the coefficient of IMDb Rating means that holding the other two variables fixed, a one-point increase in IMDb Rating is associated with an estimated expected 97.5-million-dollar increase in Chinese Gross. The rough 95% prediction interval is about $\pm 2s = \pm 2 \times 90 = \pm 180$ million dollars, but it is not very useful in practice because the minimum Budget is 100 million dollars and negative budgets are meaningless.

From the four-in-one plot, there is an obvious violation of the assumptions: non-constant variance and nonnormality of the residuals. Also, we can see several unusual points: Avengers: Infinity War is both an outlier and a leverage point; Avengers: Infinity War is a leverage point; Aquaman is an outlier. However, it seems that excluding outliers and leverage points will not be very helpful to alleviate the assumption violations. For now, let's keep the unusual observations and deal with them later.

The second model we can test is the log-log model, which takes the logarithm (base 10) of Budget and Chinese Gross:

$$Logged\ Chinese\ Gross = \beta_0 + \beta_1 \cdot Logged\ Budget + \beta_2 \cdot Douban\ Rating + \beta_3 \cdot IMDb\ Rating + random\ error$$

Regression Equation

Logged Chinese Gross = -2.55 + 1.454 Logged Budget - 0.312 Douban Rating + 0.476 IMDb Rating

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-2.55	1.73	-1.47	0.152	
Logged Budget	1.454	0.676	2.15	0.040	1.03
Douban Rating	-0.312	0.200	-1.56	0.128	2.46
IMDb Rating	0.476	0.197	2.41	0.022	2.41

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.447572	25.33%	17.86%	10.59%

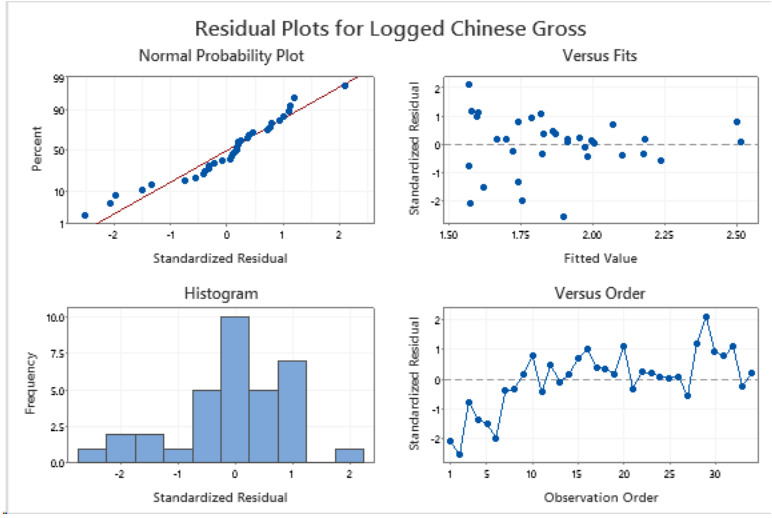
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	2.0383	0.6794	3.39	0.031
Logged Budget	1	0.9252	0.9252	4.62	0.040
Douban Rating	1	0.4897	0.4897	2.44	0.128
IMDb Rating	1	1.1625	1.1625	5.80	0.022
Error	30	6.0096	0.2003		
Total	33	8.0479			

Fits and Diagnostics for Unusual Observations

Obs	Logged Chinese Gross	Fit	Resid	Std Resid	
1	0.698	1.573	-0.876	-2.08	R
2	0.786	1.897	-1.111	-2.52	R
29	2.465	1.565	0.900	2.10	R

R Large residual



The model does not improve because it only accounts for about 25% of the variability now. There is no collinearity problem because the VIF values are all quite small. The overall F-test is less statistically significant than the first model, although still statistically significant. The t-tests for Logged Budget and IMDb Rating are also statistically significant, but the t-test for Douban Rating is not. The coefficient of Logged Budget means that holding IMDb Rating and Douban Rating fixed, a 1% higher Budget is associated with 1.454% higher Chinese Gross. The coefficient of IMDb rating means that if the other two variables are held fixed, adding 1 to the IMDb rating is associated with multiplying Chinese Gross by $10^{0.476} = 2.99$. Non-constant variance and nonnormality of the residuals persist.

It is a little disappointing that taking log does not yield a better result, probably because the long right tails were not as typical in the first place. I also fitted the data to the two semi-log

models for further testing, but the first model with no logarithm remains the best. Therefore, I am going to stick to models without logarithm going forward.

We may remove Douban Rating variable since it is not statistically significant, but for now, let's try using linear restrictions to see if the model will improve after using the sum of IMDb ratings and Douban ratings as a predictor.

The third model is of the form:

$$Chinese\ Gross = \gamma_0 + \gamma_1 \cdot Budget + \gamma_2 \cdot (Douban\ Rating + IMDb\ Rating) + random\ error$$

Recall that the first model has the form of:

$$Chinese\ Gross = \beta_0 + \beta_1 \cdot Budget + \beta_2 \cdot Douban\ Rating + \beta_3 \cdot IMDb\ Rating + random\ error$$

The third model (subset model) is a special case of the first model where $\beta_2 = \beta_3 = \gamma_2$. Thus, the partial F-statistics tests the null hypothesis:

$$H_0: \beta_2 = \beta_3 \text{ versus } H_a: \beta_2 \neq \beta_3$$

The partial F-statistics has the form:

$$F = \frac{(Residual\ SS_{subset} - Residual\ SS_{first})/1}{Residual\ SS_{first}/(n-3-1)}$$

on (1, n - 4) degrees of freedom. Minitab output of the model is:

Regression Equation

Chinese Gross (Mil$) = -640 + 1.118 Budget (Mil$) + 36.8 Total Rating

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-640	209	-3.06	0.005	
Budget (Mil$)	1.118	0.297	3.77	0.001	1.02
Total Rating	36.8	14.1	2.60	0.014	1.02

Model Summary

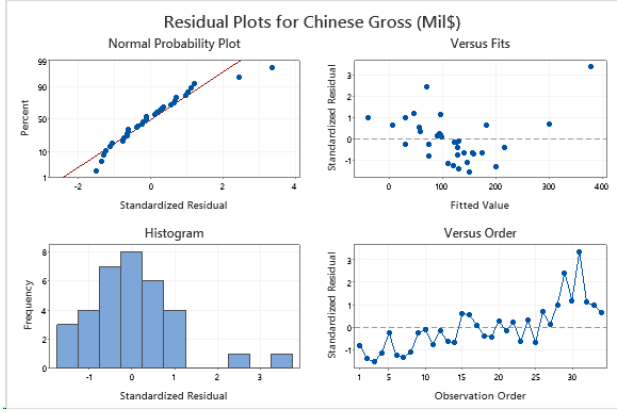
S	R-sq	R-sq(adj)	R-sq(pred)
92.4880	43.66%	40.03%	17.50%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	205521	102760	12.01	0.000
Budget (Mil$)	1	121643	121643	14.22	0.001
Total Rating	1	57724	57724	6.75	0.014
Error	31	265175	8554		
Lack-of-Fit	30	258551	8618	1.30	0.612
Pure Error	1	6624	6624		
Total	33	470696			

Fits and Diagnostics for Unusual Observations

Obs	Chinese Gross (Mil$)	Fit	Resid	Std Resid		
25	106.1	158.2	-52.1	-0.66		X
29	291.8	71.4	220.4	2.44	R	
31	629.1	378.8	250.3	3.37	R	X
34	59.1	5.9	53.2	0.68		X



R-sq, Adj R-sq and Pred R-sq all drop slightly compared with the previous model. The F-test is still highly statistically significant; the t-test for Budget is highly statistically significant as well while the t-test for Total Rating is statistically significant. Let's calculate the F-statistics now:

$$F = \frac{(265175 - 243455)/1}{243455/(34 - 3 - 1)} = 2.6765$$

which has a tail probability of 0.11, so we can see little evidence to reject the null hypothesis. We still have non-constant variance and nonnormality of the residuals. Summing up IMDb rating and Douban Rating as a new variable gives us a slightly worse result (based on R-sqs) and we cannot reject $\beta_2 = \beta_3$, so let us fall back to the model without linear restrictions.

The previous attempt at linear restrictions is built upon the previous model where we use both ratings as predictors, but there was evidence that we might be able to leave Douban Rating out and acquire a better model. To select relevant predictors more systematically, let's perform model selection on the subsets of predictions in Minitab and see if leaving out Douban Rating will be a good idea.

Response is Chinese Gross (Mil$)

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	Vars	IMDb Rating	Douban Rating	Budget (Mil$)
1	31.4	29.3	449454.8	4.5	9.8	100.45	414.685	418.464	1.000	1			X
1	20.7	18.2	459044.4	2.5	16.0	108.01	419.617	423.396	1.000	1	X		
2	47.6	44.2	359805.3	23.6	2.4	89.179	408.090	412.816	1.213	2	X		X
2	37.8	33.8	426203.7	9.5	8.1	97.157	413.917	418.643	1.368	2		X	X
3	48.3	43.1	368714.2	21.7	4.0	90.084	410.426	415.915	7.661	3	X	X	X

There are only a few combinations because we are only selecting from 3 predictors, so it is difficult to see where R-sq levels off. But leaving out Douban and keeping IMDb and Budget does prove to be the best model among them because it maximizes Adj R-sq and Pred R-sq and minimizes Mallows' Cp (here I prefer this version of Cp because Cp≈p+1 does not hold well) and AICc. It may even be the model where R-sq levels off because the R-sq of 48.3 is not too much bigger than 47.6. Either way, removing Douban Rating seems to be a good idea.

Let's take a closer look at model 4 (pooled model), where we only have Budget and IMDb Rating as our predictors. The model is of the form

$$Chinese\ Gross_{ij} = \beta_0 + \beta_1 \cdot Budget_{ij} + \beta_2 \cdot IMDb\ Rating_{ij} + random\ error_{ij}$$

where i represents Marvel or DC, while j represents the number of superhero movie within the type.

Regression Equation

Chinese Gross (Mil$) = -675 + 79.0 IMDb Rating + 1.137 Budget (Mil$)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-675	189	-3.58	0.001	
IMDb Rating	79.0	25.5	3.10	0.004	1.01
Budget (Mil$)	1.137	0.285	3.99	0.000	1.01

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
89.1787	47.62%	44.24%	23.56%

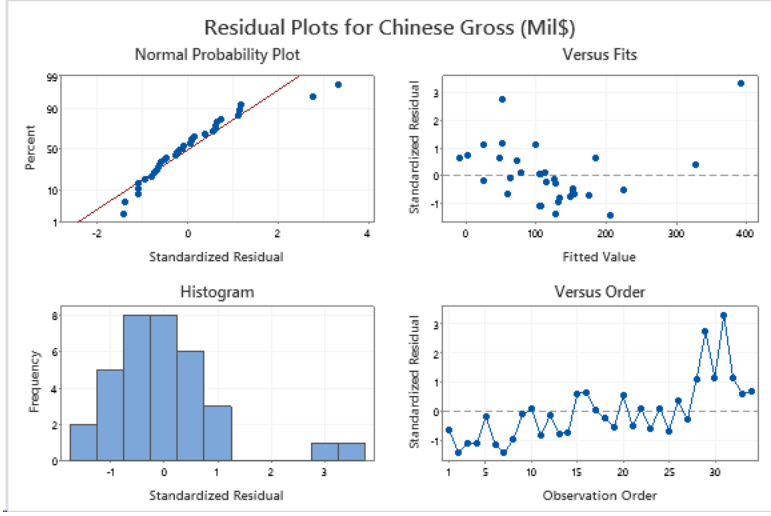
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	224158	112079	14.09	0.000
IMDb Rating	1	76361	76361	9.60	0.004
Budget (Mil$)	1	126773	126773	15.94	0.000
Error	31	246538	7953		
Lack-of-Fit	29	232157	8005	1.11	0.582
Pure Error	2	14381	7191		
Total	33	470696			

Fits and Diagnostics for Unusual Observations

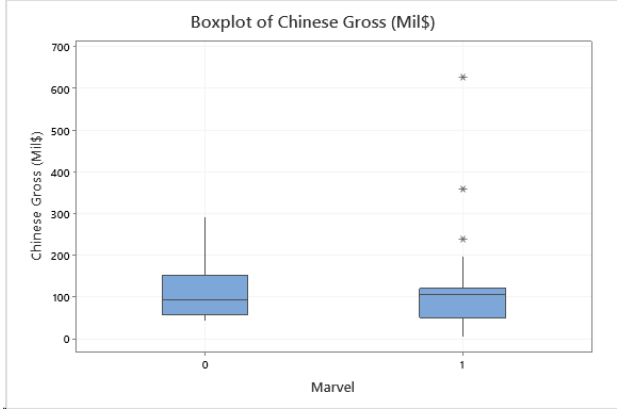
Obs	Chinese Gross (Mil$)	Fit	Resid	Std Resid		
29	291.8	51.4	240.4	2.77	R	
31	629.1	392.6	236.5	3.33	R	X

R Large residual
X Unusual X



There is already a relatively strong relationship even when we are not considering the indicator variable, with nearly 50% of the variability accounted for by the regression model. There are still no collinearity problems because the VIFs are small. The F-test is highly statistically significant, as well as the t-tests for Budget and IMDb Rating, meaning the relationship is strong and the predictors remained in this model are predictive. The coefficient of Budget means that given IMDb Rating is held fixed, a million-dollar increase in Budget is associated with an estimated expected increase in Chinese Gross of 1.137 million dollars; the coefficient of IMDb Rating means that holding Budget fixed, a one-point increase in IMDb Rating is associated with a 79.0-million-dollar increase in Chinese Gross. The rough 95% prediction interval is about $\pm 2s = \pm 2 \times 89.2 = \pm 178.4$ million dollars, but it is not of much practical use. Judging from the four-in-one plot, we still have the nonconstant variance problem and residual nonnormality problem. Despite this, the model has a potential flaw of ignoring the type of the superheroes. There may not be a need for any attempts with linear restrictions now that we only have one rating, but it is time to take the indicator variable into consideration.

Before jumping into model fitting, let's first see the box plot of the two groups. The differences may not be too noticeable.



We continue to fit the model to model 5, the constant shift model, by adding Marvel to the equation:

$$Chinese\ Gross_{ij} = \beta_0 + \beta_1 \cdot Budget_{ij} + \beta_2 \cdot IMDb\ Rating_{ij} + \beta_3 \cdot Marvel_{ij} + random\ error_{ij}$$

We can see that when Marvel = 0 (Group 1: DC movies), we have:

$$\begin{aligned} Chinese\ Gross_{1j} &= \beta_0 + \beta_1 \cdot Budget_{1j} + \beta_2 \cdot IMDb\ Rating_{1j} + random\ error_{1j} \\ &= \beta_{10} + \beta_1 \cdot Budget_{1j} + \beta_2 \cdot IMDb\ Rating_{1j} + random\ error_{1j} \end{aligned}$$

where $\beta_0 = \beta_{10}$. When Marvel = 1 (Group 2: Marvel movies), we have:

$$\begin{aligned} Chinese\ Gross_{2j} &= \beta_0 + \beta_3 + \beta_1 \cdot Budget_{1j} + \beta_2 \cdot IMDb\ Rating_{1j} + random\ error_{1j} \\ &= \beta_{20} + \beta_1 \cdot Budget_{2j} + \beta_2 \cdot IMDb\ Rating_{2j} + random\ error_{2j} \end{aligned}$$

where $\beta_3 = \beta_{20} - \beta_{10}$.

The output from Minitab is as follows:

Regression Equation

Chinese Gross (Mil$) = -701 + 87.2 IMDb Rating + 1.125 Budget (Mil$) - 38.4 Marvel

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-701	191	-3.67	0.001	
IMDb Rating	87.2	27.2	3.21	0.003	1.14
Budget (Mil$)	1.125	0.286	3.93	0.000	1.01
Marvel	-38.4	42.8	-0.90	0.377	1.13

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
89.4630	48.99%	43.89%	18.70%

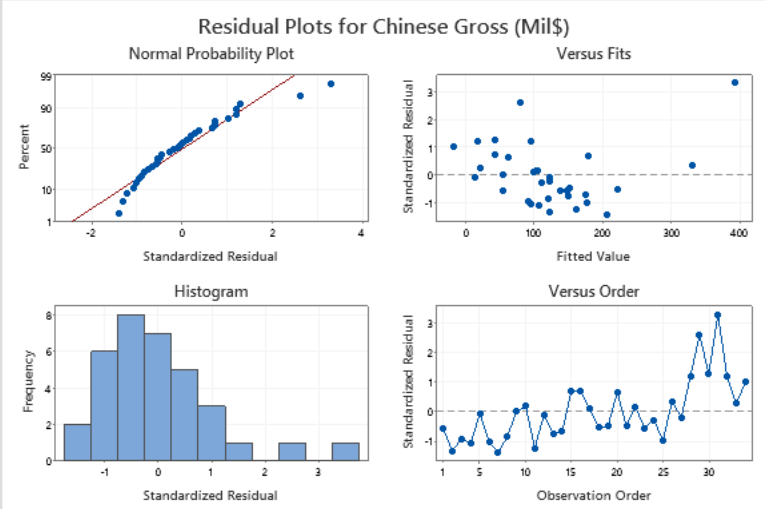
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	230587	76862	9.60	0.000
IMDb Rating	1	82448	82448	10.30	0.003
Budget (Mil$)	1	123855	123855	15.47	0.000
Marvel	1	6429	6429	0.80	0.377
Error	30	240109	8004		
Lack-of-Fit	29	235212	8111	1.66	0.557
Pure Error	1	4896	4896		
Total	33	470696			

Fits and Diagnostics for Unusual Observations

Obs	Chinese Gross (Mil$)	Fit	Resid	Std Resid	
29	291.8	80.5	211.3	2.61	R
31	629.1	393.4	235.7	3.31	R X

R Large residual
X Unusual X



The result does not seem to change too much. The F-test is still highly statistically significant, but the t-test for Marvel shows that it is not statistically significant. This t-statistics tests the hypothesis:

$$H_0: \beta_{10} = \beta_{20} \text{ versus } H_a: \beta_{10} \neq \beta_{20}$$

and the p-value of 0.377 means that we fail to reject the null hypothesis. This is surprising for me because I would expect Marvel and DC to be two very distinct groups and the coefficient of the indicator variable to be at least marginally significant because Marvel movies are so much better than DC movies in my opinion.

There is a possibility that the constant shift model is not much better than the pooled model, but the full model might be. To further generalize the model, let's consider model 6, the full model, to model the relationship as two separate lines with different coefficients:

$$Chinese\ Gross_{ij} = \beta_0 + \beta_1 \cdot Budget_{ij} + \beta_2 \cdot IMDb\ Rating_{ij} + \beta_3 \cdot Marvel_{ij}$$
$$+\beta_4 \cdot BudgetMarvel_{ij} + \beta_5 \cdot IMDbMarvel_{ij} + random\ error_{ij}$$

where BudgetMarvel is the product of Budget and Marvel and IMDbMarvel is the product of IMDb Rating and Marvel. We can see that when Marvel = 0 (Group 1: DC movies), we have:

$$Chinese\ Gross_{1j} = \beta_0 + \beta_1 \cdot Budget_{1j} + \beta_2 \cdot IMDb\ Rating_{1j} + random\ error_{1j}$$
$$= \beta_{10} + \beta_{11} \cdot Budget_{1j} + \beta_{12} \cdot IMDb\ Rating_{1j} + random\ error_{1j}$$

where $\beta_0 = \beta_{10}, \beta_1 = \beta_{11}$ and $\beta_2 = \beta_{12}$. When Marvel = 1 (Group 2: Marvel movies), we have:

$$Chinese\ Gross_{2j} = \beta_0 + \beta_1 \cdot Budget_{2j} + \beta_2 \cdot IMDb\ Rating_{2j} + \beta_3 + \beta_4 \cdot Budget_{2j} + \beta_5 \cdot IMDb\ Rating_{2j} + random\ error_{2j}$$
$$= (\beta_0 + \beta_3) + (\beta_1 + \beta_4) Budget_{2j} + (\beta_2 + \beta_5) IMDbRating_{2j} + random\ error_{2j}$$
$$= \beta_{20} + \beta_{21} \cdot Budget_{2j} + \beta_{22} \cdot IMDb\ Rating_{2j} + random\ error_{2j}$$

where $\beta_3 = \beta_{20} - \beta_{10}$, $\beta_4 = \beta_{21} - \beta_{11}$ and $\beta_5 = \beta_{22} - \beta_{12}$.

To compare the full model with the pooled model, we need to test if $\beta_3 = 0$ and $\beta_4 = 0$ and $\beta_5 = 0$, which requires a partial F-test of the form:

$$F = \frac{(Residual\ SS_{pooled} - Residual\ SS_{full})/3}{Residual\ SS_{full}/(n - 5 - 1)}$$

on (3, n - 6) degrees of freedom.

To compare the full model with the constant shift model, we need to test if $\gamma_4 = 0$ and $\gamma_5 = 0$, which requires a partial F-test of the form:

$$F = \frac{(Residual\ SS_{constantShift} - Residual\ SS_{full})/2}{Residual\ SS_{full}/(n - 5 - 1)}$$

on (2, n - 6) degrees of freedom.

To get the value of $Residual\ SS_{full}$, let's fit the full model in Minitab:

Regression Equation

Chinese Gross (Mil$) = 506 - 49 IMDb Rating - 0.298 Budget (Mil$) - 1213 Marvel + 1.764 BudgetMarvel + 122 IMDbMarvel

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	506	1129	0.45	0.658	
IMDb Rating	-49	145	-0.34	0.740	34.80
Budget (Mil$)	-0.298	0.864	-0.35	0.733	9.90
Marvel	-1213	1147	-1.06	0.299	870.68
BudgetMarvel	1.764	0.927	1.90	0.067	30.04
IMDbMarvel	122	148	0.83	0.414	809.68

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
86.4067	55.59%	47.66%	3.55%

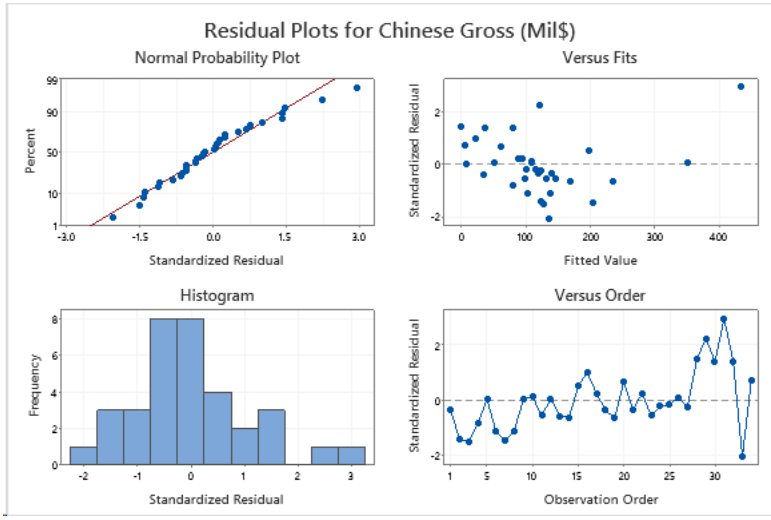
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	261644	52328.9	7.01	0.000
IMDb Rating	1	840	840.2	0.11	0.740
Budget (Mil$)	1	890	889.6	0.12	0.733
Marvel	1	8353	8353.1	1.12	0.299
BudgetMarvel	1	27041	27041.5	3.62	0.067
IMDbMarvel	1	5126	5125.6	0.69	0.414
Error	28	209051	7466.1		
Lack-of-Fit	27	204155	7561.3	1.54	0.572
Pure Error	1	4896	4896.5		
Total	33	470696			

Fits and Diagnostics for Unusual Observations

Obs	Chinese Gross (Mil$)	Fit	Resid	Std Resid		
24	90.5	101.3	-10.8	-0.20		X
25	106.1	114.6	-8.6	-0.16		X
29	291.8	122.3	169.5	2.25	R	
31	629.1	434.0	195.1	2.97	R	
33	43.8	135.3	-91.5	-2.04	R	X

R Large residual
X Unusual X



Therefore, we can calculate the two F-statistics. The first one is:

$$F = \frac{(246538 - 209051)/3}{209051/(34 - 5 - 1)} = 1.6737$$

which has a tail probability of 0.1953. This is not statistically significant to reject the null hypothesis, so we still prefer the pooled model.

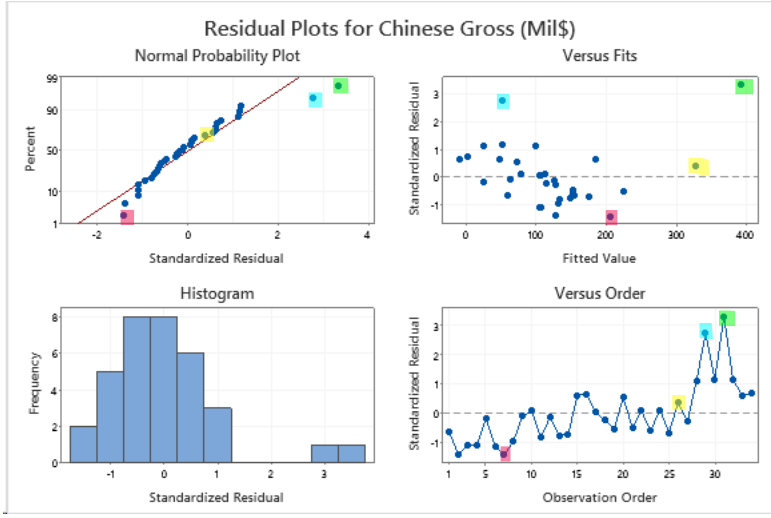
The second F-statistics is:

$$F = \frac{(240109 - 209051)/2}{209051/(34 - 5 - 1)} = 2.0799$$

which has a tail probability of 0.1438. This is not statistically significant either, so we stick to the pooled model.

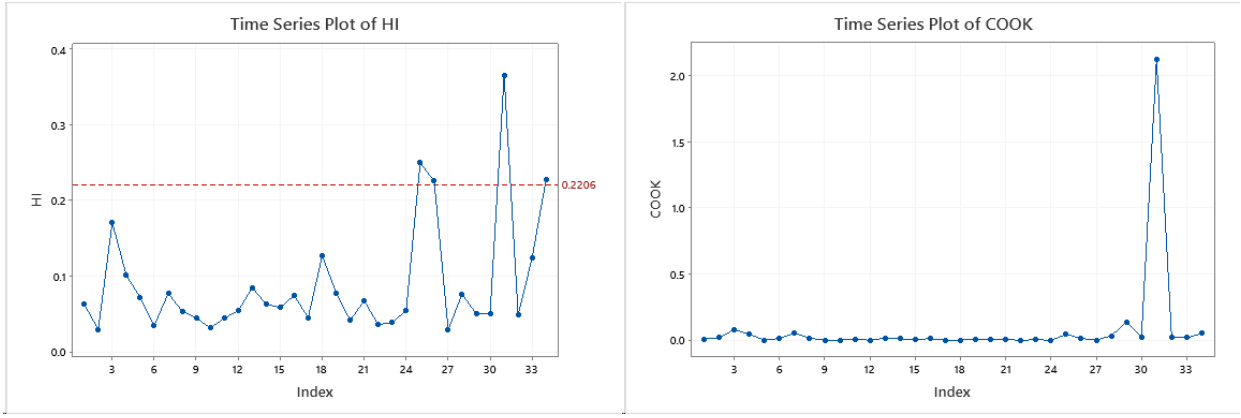
It is a little disappointing that all the work on testing the nested models with the indicator variable has not been able to yield a better model (and even more disappointing that Marvel movies are not too distinctive from DC movies because I am a big Marvel fan), so I will go on to run diagnostics on the observations and check unusual points to see if the pooled model will better fit a subset of the sample. A recap on the pooled model:

$$Chinese\ Gross_{ij} = \beta_0 + \beta_1 \cdot Budget_{ij} + \beta_2 \cdot IMDb\ Rating_{ij} + random\ error_{ij}$$



Avengers: Endgame (highlighted green) and Aquaman (highlighted blue) remain to be unusual observations (recall that they were also identified as unusual in the very first model we tested). The residual of The Avengers (highlighted purple) seems to be a little unusual while Avengers: Infinity War (highlighted yellow) is another leverage point.

Now, let's run diagnostics quantitively by calculating the Standardized Residual, leverage values and Cook's distances. The reference line in the HI plot is 2.5*((2+1)/34) = 0.2206.



The spike in Cook plot is Avengers: Endgame, which also has a high HI value of 0.364921 and a standardized residual of 3.32711. Its predictor value and its target value are unusually large. It is the most commercially successful Marvel movie there is. It builds on its previous Avenger movie, Avengers: Infinity War, where half of the population in the universe is disintegrated, leaving the biggest cliffhanger in Marvel history and attracting more people to its sequel Avengers: Endgame. In addition, Avengers: Endgame assembles all superheroes in the Marvel Cinematic Universe, so the celebrity cast also attracts more people to the cinema. Most importantly, this movie marks the end of a decade-long Marvel era with the irreversible death of possibly their most popular superhero -- Iron Man, highlighting the end of a decade-long run. Because of all these reasons, Avengers: Endgame was extremely popular worldwide and also in the Chinese market. The other three points above the reference line are Justice League, Avengers: Infinity War and X-men: Dark Phenix, but they are actually still very close to the reference line.

I will exclude Avengers: Endgame, which is both an outlier and a leverage point, and model on the new sample with 33 observations. I repeated all the models from the very beginning and arrived at the same pooled model with Budget and IMDb Rating as predictors. Since it is verbose to rehash everything again, here I will record the last few test result of the pooled model, the constant shift model and the full model here.

Output of the pooled model:

Regression Equation

Chinese Gross (Mil$) = -360 + 49.4 IMDb Rating + 0.567 Budget (Mil$)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-360	172	-2.09	0.045	
IMDb Rating	49.4	22.0	2.25	0.032	1.01
Budget (Mil$)	0.567	0.271	2.09	0.045	1.01

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
72.6872	22.37%	17.19%	0.00%

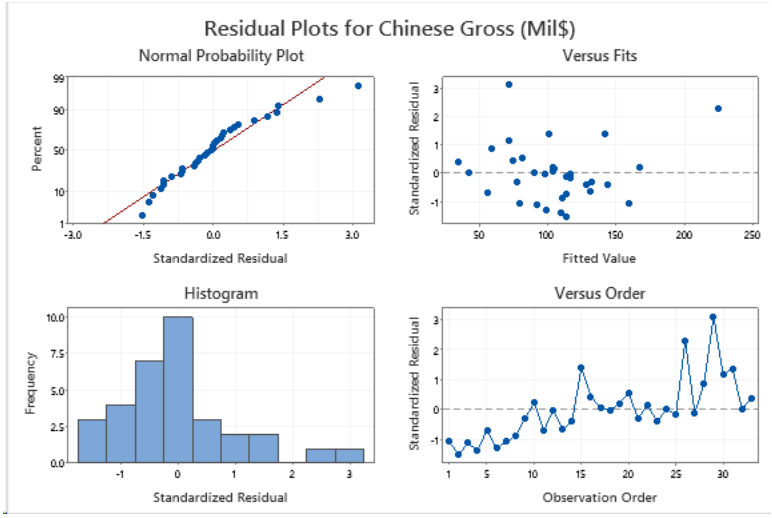
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	45667	22833	4.32	0.022
IMDb Rating	1	26666	26666	5.05	0.032
Budget (Mil$)	1	23175	23175	4.39	0.045
Error	30	158503	5283		
Lack-of-Fit	28	144122	5147	0.72	0.736
Pure Error	2	14381	7191		
Total	32	204169			

Fits and Diagnostics for Unusual Observations

Obs	Chinese Gross (Mil$)	Fit	Resid	Std Resid	
26	359.5	225.0	134.5	2.29	R X
29	291.8	71.5	220.3	3.12	R

R Large residual
X Unusual X



The constant shift model:

Regression Equation

Chinese Gross (Mil$) = -386 + 57.4 IMDb Rating + 0.558 Budget (Mil$) - 36.6 Marvel

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-386	173	-2.23	0.034	
IMDb Rating	57.4	23.2	2.47	0.020	1.13
Budget (Mil$)	0.558	0.270	2.06	0.048	1.01
Marvel	-36.6	34.7	-1.05	0.300	1.12

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
72.5520	25.23%	17.50%	0.00%

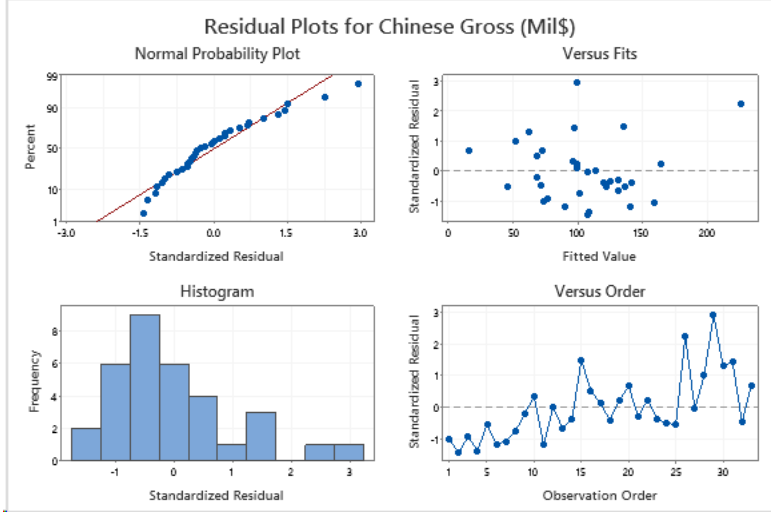
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	51519	17173	3.26	0.036
IMDb Rating	1	32158	32158	6.11	0.020
Budget (Mil$)	1	22374	22374	4.25	0.048
Marvel	1	5853	5853	1.11	0.300
Error	29	152650	5264		
Lack-of-Fit	28	147753	5277	1.08	0.656
Pure Error	1	4896	4896		
Total	32	204169			

Fits and Diagnostics for Unusual Observations

Obs	Chinese Gross (Mil$)	Fit	Resid	Std Resid	
26	359.5	226.8	132.8	2.27	R
29	291.8	99.2	192.6	2.94	R

R Large residual



The full model:

Regression Equation

Chinese Gross (Mil$) = 506 - 49 IMDb Rating - 0.298 Budget (Mil$) - 951 Marvel + 104 IMDbMarvel + 1.073 BudgetMarvel

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	506	951	0.53	0.600	
IMDb Rating	-49	122	-0.40	0.694	31.01
Budget (Mil$)	-0.298	0.728	-0.41	0.685	7.26
Marvel	-951	969	-0.98	0.335	870.14
IMDbMarvel	104	125	0.83	0.413	794.55
BudgetMarvel	1.073	0.805	1.33	0.194	27.15

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
72.8127	29.89%	16.91%	0.00%

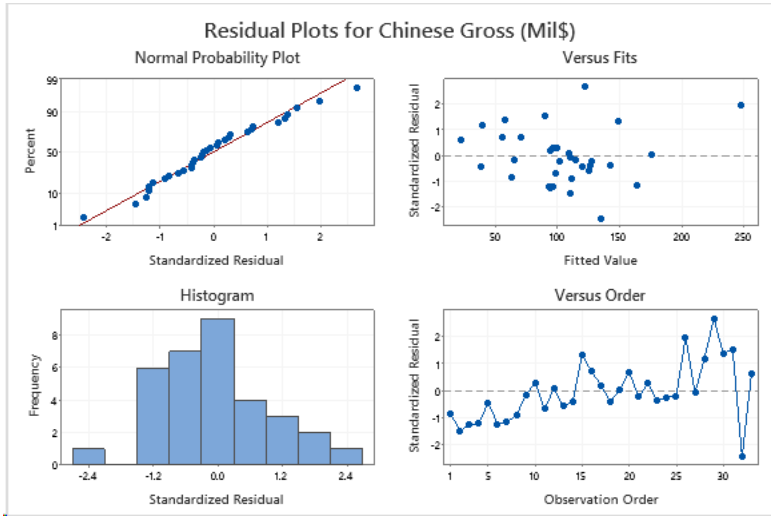
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	61024	12204.8	2.30	0.073
IMDb Rating	1	840	840.2	0.16	0.694
Budget (Mil$)	1	890	889.6	0.17	0.685
Marvel	1	5104	5103.5	0.96	0.335
IMDbMarvel	1	3659	3658.8	0.69	0.413
BudgetMarvel	1	9408	9408.2	1.77	0.194
Error	27	143146	5301.7		
Lack-of-Fit	26	138249	5317.3	1.09	0.654
Pure Error	1	4896	4896.5		
Total	32	204169			

Fits and Diagnostics for Unusual Observations

Obs	Chinese Gross (Mil$)	Fit	Resid	Std Resid		
24	90.5	101.3	-10.8	-0.24		X
25	106.1	114.6	-8.6	-0.19		X
29	291.8	122.3	169.5	2.67	R	
32	43.8	135.3	-91.5	-2.43	R	X

R Large residual
X Unusual X



Again, the R-sq does not improve significantly after considering the indicator variable. To compare the pooled model and the constant shift model, we check the t-test of Marvel, which has a p-value of about 0.3, so it is not statistically significant. To compare the pooled model with the full model, we perform the partial F-test and get the F-statistics:

$$F = \frac{(158503 - 143146)/3}{143146/(33 - 5 - 1)} = 0.9655$$

which has a tail probability of 0.4233. This is not statistically significant at all, so we fail to reject the null hypothesis and prefer the pooled model.

To compare the constant shift model with the full model, we perform the partial F-test and get the F-statistics:

$$F = \frac{(152650 - 143146)/2}{143146/(33 - 5 - 1)} = 0.8963$$

which has a tail probability of 0.4199. For the same reason, we prefer the pooled model.

The pooled model fitted with the dataset without Avengers: Endgame has a much smaller R-sq, which means this model only accounts for about 22% of the variability now. There is no problem with collinearity. The overall F-test and the t-tests are still statistically significant. The coefficient of Budget means that fixing IMDb Rating, a million-dollar increase in Budget is associated with an estimated expected increase in Chinese Gross of 0.567 million dollars; the coefficient of IMDb Rating means that holding Budget fixed, a one-point increase in IMDb Rating is associated with an estimated expected 49.4-million-dollar increase in the total gross in China. The model has a prediction interval of about ± 145.4 million dollars, roughly 95% of the time. When checking assumptions, we can see from the four-in-one plot that we still have non-constant variance and nonnormality of the residuals.

We can keep using fewer observations as new datasets by excluding more outliers and leverage points, but the model building, model selection and model interpretation generally follow the same processes as above. During my various attempts of excluding unusual observations, I have yet to find any model that performs any better, perhaps because linear regression models are simply not the most suitable for this dataset. Leaving out unusual points gives me models with smaller R-sq while the same violation of assumptions persists.

In conclusion, the simple pooled model with two predictors and no log transformation or indicator variable is better. Although movie ratings are subjective and movie studios are not obliged to tell the public their real production budget, Budget and IMDb Rating prove to be

significant predictors in our case. It is not too surprising that Douban Rating was left out the model because it is very similar to IMDb Rating. What is surprising to me is that the indicator variable of Marvel/DC did not contribute much to the model.

There is currently no hold-out dataset because there are only these many Marvel and DC superhero movies passing the Chinese government censorship and being released in the Chinese market. But superhero movies are indeed hits in China, so in the next decade or so, there will be more data to collect and analyze. And perhaps by then, the superiority of Marvel movies will be more prominent, and the indicator variable will be significant.