Kellogg School of Management
NORTHWESTERN UNIVERSITY

TEACHING NOTE

KARL SCHMEDDERS **KEL701**

Hollywood Rules

Case Synopsis

Wall Street hedge fund manager Kim Meyer is considering investing in an SFA (slate financing arrangement) in Hollywood, and Dave Griffith, a Hollywood producer, is pitching for the investment. In order to prepare for an upcoming meeting with Meyer, Griffith will be conducting a broad analysis of recent movie data to determine the drivers of a movie's success. The data set includes, among other pieces, information on the movies' respective titles, budgets, domestic and foreign gross, times of release, ratings, and so on. In order to convince Meyer to invest in an SFA, Griffith must anticipate possible questions to maximize his persuasiveness.

Teaching Objective

This case introduces students to various types of statistical analysis. Answering the questions and analyzing the factors driving a movie's revenue requires students to calculate point estimates, compute confidence intervals, conduct hypothesis tests, and develop regression models. In particular, the development of regression models requires both choosing the relevant set of independent variables as well as determining an appropriate functional form for the regression equation. The case also requires students to interpret the quantitative findings in the context of the application.

Positioning

The case is versatile and can be used in different classroom settings. It has been taught successfully in the core statistics courses of both the MBA and executive MBA programs at the Kellogg School of Management. The case is also suited for a market research class or, if the students have a good understanding of regression analysis, even an empirical finance class.

Prerequisites for Students

Ideally, students should be able to compute confidence intervals and conduct hypothesis tests both for a single population mean as well as for comparing two means. Moreover, students should have at least a rudimentary knowledge of linear regression analysis. In this teaching note, we report linear regressions for all regression questions, so an instructor could limit the discussion to linear regressions. We also show how to check the basic assumptions for regression models and

point out when linear regressions are inappropriate and log-linear regressions should be applied instead. Thus, this case may be best suited for students who are comfortable with such models.

Students do not need to have a specific knowledge of finance or marketing. Any relevant terminology is explained in detail in the case.

Software

A number of software packages are available for performing the different types of statistical analysis applied in this case. These packages range from simple Microsoft Excel add-in tools to sophisticated statistical software such as R or SPSS. This teaching note contains screenshots from the IBM SPSS 19 statistical analysis program, but any package can be used and should lead to the same results.

Supplemental Materials

The following materials accompany the case:

- *Hollywood Rules Exhibit 1 Spreadsheet (XLS):* Data from seventy-five movies first shown in U.S. movie theatres in 2006

Case Analysis

1. To obtain an initial overview of the data, calculate the minimum, average, and maximum values of the variables—opening gross, total U.S. gross, total non-U.S. gross, and opening theatres. How many of the movies in the data set are comedies and how many movies are R-rated?

The required descriptive statistics are as follows:

Descriptive Statistics

	N	Minimum	Maximum	Mean
Opening Gross	75	4120497	68033544	17468465,5
Total U.S. Gross	75	13090630	198000317	59620650,8
Total Non-U.S. Gross	75	0	456235122	59560982,5
Opening Theatres	75	852	3964	2766,28
Valid N (listwise)	75			

MPAA_D

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Non-R-Rated	60	80,0	80,0	80,0
	R-Rated	15	20,0	20,0	100,0
	Total	75	100,0	100,0	

Comedy_D

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Non-Comedy	52	69,3	69,3	69,3
	Comedy	23	30,7	30,7	100,0
	Total	75	100,0	100,0	

Among the seventy-five movies in the data set, fifteen are R-rated and twenty-three are comedies.

2. Michael London (of Sideways *fame) declared in* The Hollywood Reporter, *"The studio business historically returns around 12 percent a year." Griffith knew any investor would want justification for such a statement.*

a. Calculate the U.S. return on investment (ROI) (simply defined as the difference of total U.S. box-office gross and budget divided by budget, ignoring any form of discounting) for each movie in the data set.

This question only requires a simple calculation in the data sheet.

b. Provide a 95 percent confidence interval for the mean U.S. ROI of movies.

The 95 percent confidence interval for the mean U.S. ROI is [0.1348, 0.4510].

Descriptives

			Statistic	Std. Error
ROI_TotalU.S.Gross	Mean		,2929	,07935
	95% Confidence Interval for Mean	Lower Bound	,1348	
		Upper Bound	,4510	
	5% Trimmed Mean		,2554	
	Median		,1672	
	Variance		,472	
	Std. Deviation		,68723	
	Minimum		-,82	
	Maximum		2,56	
	Range		3,39	
	Interquartile Range		1,00	
	Skewness		,850	,277
	Kurtosis		,729	,548

c. Show that the mean U.S. ROI is significantly larger than the 12 percent London cited.

The answer to this question requires a one-tailed hypothesis test:

H_0: $\mu_{US\text{-}ROI} \leq 0.12$

H_A: $\mu_{US\text{-}ROI} > 0.12$

One-Sample Test

	Test Value = 0.12					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
ROI_TotalU.S.Gross	2,179	74	,032	,17293	,0148	,3310

Note: SPSS always issues the t-value for a two-sided t-test. Thus, to obtain the correct p-value for our one-tailed hypothesis test, we need to divide the reported value by 2. Therefore, the relevant p-value is 0.032 / 2 = 0.016.

Therefore, we can conclude that the mean U.S. ROI is significantly larger than 12 percent, both at the 5 percent as well as the 10 percent significance level (because 0.016 < 0.05 < 0.1).

3. While any genre can produce a blockbuster, Griffith suspected that some categories are more likely to do so than others. If he could stack the deck in his favor through storyline selection, he did not want to pass up the opportunity.

a. Compare the total U.S. box-office gross of movies from the comedy genre with movies from other genres. Is there a statistically significant difference between the total U.S. gross of comedies and non-comedy movies?

The answer to this question requires a two-tailed hypothesis test:

H_0: $\mu_{\text{Non-Comedy}} = \mu_{\text{Comedy}}$

H_A: $\mu_{\text{Non-Comedy}} \neq \mu_{\text{Comedy}}$

The test statistic is as follows:

$t = (\text{x-bar}_1 - \text{x-bar}_2) / \text{sqrt}\ (s_1^2/n_1 + s_2^2/n_2)$

$= (68'743'100.4 - 55'585'721.1) / \text{sqrt}\ (41'419'962.5^2/52 + 36'797'919.9^2/23)$

$= 1.373$

The resulting p-value is 0.174 (assuming $(n_1 + n_2 - 2)$ degrees of freedom); therefore, we cannot establish a statistically significant difference between the U.S. box-office gross of comedy and non-comedy movies.

Group Statistics

	Comedy_D	N	Mean	Std. Deviation	Std. Error Mean
Total U.S. Gross	Comedy	23	68743100,4	36797919,9	7672896,71
	Non-Comedy	52	55585721,1	41419962,5	5743915,33

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Total U.S. Gross	Equal variances assumed	,017	,896	1,311	73	,194	13157379,2	10037534,5	-6847398,1	33162156,6
	Equal variances not assumed			1,373	47,176	,176	13157379,2	9584670,43	-6122595,6	32437354,1

Note: Another common approach to this problem is to set up a regression in which the U.S. box-office gross is the dependent variable and the dummy variable Comedy/Non-Comedy acts as the independent variable. Using a linear regression approach rests on the assumption that the variances of total U.S. gross of the two groups of movies are identical. Therefore, this approach yields the same p-value as the t-test with "equal variances assumed."

b. Griffith was not so sure about the results, because they were contrary to his gut feelings. Maybe higher revenue accompanied higher investments? Calculate additionally the difference of U.S. ROIs from movies of the comedy genre and of other movie genres. Is there a statistically significant difference between the U.S. ROIs?

We can determine whether higher revenue is accompanied by higher investments by looking at the correlations.

Correlations

		Total U.S. Gross	Budget
Total U.S. Gross	Pearson Correlation	1	,496**
	Sig. (2-tailed)		,000
	N	75	75
Budget	Pearson Correlation	,496**	1
	Sig. (2-tailed)	,000	
	N	75	75

**. Correlation is significant at the 0.01 level (2-tailed).

There appears to be a statistically significant correlation between the total U.S. gross and the budget of movies.

The difference between the U.S. ROI from movies of the comedy genre and other movie genres can again be tested using a regular two sample t-test:

H_0: $\mu_{ROI\text{-}Non\text{-}Comedy} = \mu_{ROI\text{-}Comedy}$

H_A: $\mu_{ROI\text{-}Non\text{-}Comedy} \neq \mu_{ROI\text{-}Comedy}$

Group Statistics

	Comedy_D	N	Mean	Std. Deviation	Std. Error Mean
ROI_TotalU.S.Gross	Comedy	23	,5402	,71387	,14885
	Non-Comedy	52	,1836	,65244	,09048

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
ROI_TotalU.S.Gross	Equal variances assumed	,191	,663	2,120	73	,037	,35660	,16817	,02144	,69175
	Equal variances not assumed			2,047	38,965	,047	,35660	,17419	,00425	,70894

The test statistic is as follows:

$t = (\text{x-bar}_1 - \text{x-bar}_2) / \text{sqrt}\ (s_1^2/n_1 + s_2^2/n_2)$

$= (0.5402 - 0.1836) / \text{sqrt}\ (0.71387^2/23 + 0.65244^2/52)$

$= 2.04716$

The p-value is 0.0375, so the difference of U.S. ROIs from movies of the comedy genre and other movie genres is indeed significant at the 5 percent level.

4. Prevailing wisdom maintained that R-rated movies performed better than other movies.

a. Is there a statistically significant difference between the total U.S. gross of R-rated movies and movies with other ratings?

The hypotheses to test the conventional wisdom are:

H_0: $\mu_{U.S.Gross\text{-}R\text{-}Rated} = \mu_{U.S.Gross\text{-}NON\text{-}R\text{-}Rated}$

H_A: $\mu_{U.S.Gross\text{-}R\text{-}Rated} \neq \mu_{U.S.Gross\text{-}NON\text{-}R\text{-}Rated}$

Group Statistics

	MPAA_D	N	Mean	Std. Deviation	Std. Error Mean
Total U.S. Gross	R-Rated	15	53330311,8	28378855,6	7327389,01
	Non-R-Rated	60	61193235,5	42790360,9	5524211,84

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Total U.S. Gross	Equal variances assumed	2,540	,115	-,674	73	,503	-7862923,7	11670168,2	-31121535,	15395688,0
	Equal variances not assumed			-,857	31,986	,398	-7862923,7	9176466,97	-26555104,	10829256,9

The test statistic

$t = (\text{x-bar}_1 - \text{x-bar}_2) / \text{sqrt}\ (s_1^2/n_1 + s_2^2/n_2)$

$= (61'193'235.5 - 53'330'311.8) / \text{sqrt}\ (42'790'360.9^2/60 + 28'378'855.6^2/15)$

$= 0.857$

results in a p-value of 0.398. Hence, we cannot prove the conventional wisdom to be correct.

5. Believed to be among the preproduction factors driving success were budget (which expresses both the cost of the film and the quality of the actors as expressed by their fee), genre (comedy vs. non-comedy), MPAA rating (R-rated vs. other rating), and audiences' familiarity with the story (whether the film is a sequel or an adaptation of a known story).

For the following exercises the provided data set needs to be expanded. We need to define a dummy variable (1 = Comedy, 0 = Non-Comedy) distinguishing between comedies and movies from other genres. The dummy variable MPAA-D for the MPAA rating R is already part of the original data set.

a. Based on the described beliefs, determine a sound regression model predicting total U.S. box-office gross of movies prior to production.

As a first step, we run a standard linear regression and check the validity of the basic assumptions that have to be met for the linear regression to be sound.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,596[a]	,355	,308	33494159,5	1,603

a. Predictors: (Constant), Known Story, Budget, Comedy_D, Sequel, MPAA_D
b. Dependent Variable: Total U.S. Gross

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4,264E16	5	8,528E15	7,601	,000[a]
	Residual	7,741E16	69	1,122E15		
	Total	1,200E17	74			

a. Predictors: (Constant), Known Story, Budget, Comedy_D, Sequel, MPAA_D
b. Dependent Variable: Total U.S. Gross

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	12225019,8	10520788,0		1,162	,249		
	Budget	,897	,165	,527	5,424	,000	,991	1,009
	Comedy_D	14758022,2	8919589,78	,170	1,655	,103	,884	1,131
	MPAA_D	-4156466,8	10310130,8	-,042	-,403	,688	,879	1,137
	Sequel	29166706,7	12774202,6	,225	2,283	,025	,962	1,040
	Known Story	-9977764,1	8245399,99	-,120	-1,210	,230	,955	1,047

a. Dependent Variable: Total U.S. Gross

The Breusch-Pagan heteroskedasticity test returns a p-value of 0.002, showing that the data is heteroskedastic. Therefore, the linear model should be discarded.

We next run a log-linear regression with the dependent variable LN(Total U.S. Gross).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	16,875	,170		99,035	,000
	Budget	1,408E-8	,000	,509	5,258	,000
	Comedy_D	,337	,144	,239	2,336	,022
	MPAA_D	,068	,167	,042	,410	,683
	Sequel	,533	,207	,253	2,578	,012
	Known Story	-,127	,134	-,094	-,948	,347

a. Dependent Variable: LN_TotalU.S.Gross

The Breusch-Pagan test has a p-value of 0.43; that is, we cannot reject the hypothesis of homoskedasticity. The residual plot may indicate some curvature but for simplicity we assume that the current specification is satisfying.

The Durbin-Watson statistic does not suggest autocorrelation. All variance inflation factors are small, so we do not face a multicollinearity problem.

The resulting regression is:

$$\text{LN(Total U.S.Gross)} = 16.875 + 1.408*10^{-8}\ \text{Budget} + 0.337\ \text{Comedy\_D} + 0.068\ \text{MPAA\_D} + 0.533\ \text{Sequel} - 0.127\ \text{KnownStory}$$

b. Drop all variables from the regression that are not significant at a 10 percent level of significance. Report the final regression.

Only the variables Budget, Sequel, and Comedy_D are significant at a 10 percent level. The new regression model is then as follows:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	16,842	,156		107,917	,000
	Budget	,000000014	,000	,509	5,296	,000
	Sequel	,570	,202	,271	2,819	,006
	Comedy_D	,329	,135	,233	2,431	,018

a. Dependent Variable: LN_TotalU.S.Gross

The final regression equation is:

$$LN(\text{Total U.S.Gross}) = 16.842 + 1.4*10^{-8}\ \text{Budget} + 0.570\ \text{Sequel} + 0.329\ \text{Comedy\_D}$$

c. Holding all other explanatory variables in your regression fixed, which movies have higher total U.S. gross, those that are a sequel or those that are not?

The coefficient of the variable Sequel is positive and statistically significant at the 1 percent significance level. Therefore, sequels have on average a higher total U.S. gross than non-sequels, holding the budget and the comedy dummy fixed. In fact, sequels have, on average, a higher total U.S. gross of EXP(0.570) – 1 = 0.768, that is, of about 77 percent, holding the other two variables fixed.

6. Griffith knew the age-old Hollywood wisdom that the opening weekend is absolutely critical for the overall commercial success of a movie. Therefore, both the release date (whether during the summer, on a U.S. holiday, or around Christmas) and the number of movie theatres in which a movie is shown during opening weekend are assumed to be very important. These factors are believed to strongly influence revenue during the opening weekend and, thereby, to have a strong impact on the overall commercial success of a movie.

a. Determine a sound regression model predicting opening weekend box-office gross revenue. Consider both the preproduction success factors as well as the factors describing the opening weekend.

We begin with a linear regression model:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,724[a]	,524	,458	8156928,33	1,998

a. Predictors: (Constant), Opening Theatres, Holiday, Summer, Known Story, MPAA_D, Budget, Sequel, Comedy_D, Christmas
b. Dependent Variable: Opening Gross

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-7743523,8	4459475,11		-1,736	,087		
	Budget	,135	,043	,289	3,133	,003	,862	1,161
	Known Story	-2605925,8	2043045,94	-,114	-1,276	,207	,922	1,084
	Sequel	9297737,02	3288201,95	,261	2,828	,006	,861	1,161
	MPAA_D	618549,538	2604048,42	,022	,238	,813	,818	1,223
	Comedy_D	1133908,57	2289730,10	,047	,495	,622	,796	1,257
	Summer	-4125785,3	2239719,80	-,173	-1,842	,070	,832	1,202
	Holiday	147197,179	3555280,00	,004	,041	,967	,829	1,206
	Christmas	-3849947,6	3493935,90	-,108	-1,102	,275	,763	1,311
	Opening Theatres	7115,745	1559,379	,451	4,563	,000	,750	1,334

a. Dependent Variable: Opening Gross

The Breusch-Pagan heteroskedasticity test yields a p-value of 0.000, which is a strong indication of heteroskedasticity. Therefore, we discard the linear model and run a log-linear regression with the dependent variable LN(Opening Gross):

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	14,967	,219		68,198	,000		
	Budget	,000000005	,000000002	,227	2,619	,011	,862	1,161
	Opening Theatres	,000473811	,000076743	,573	6,174	,000	,750	1,334
	Known Story	-,089	,101	-,074	-,881	,382	,922	1,084
	Sequel	,354	,162	,189	2,186	,032	,861	1,161
	Summer	-,256	,110	-,205	-2,324	,023	,832	1,202
	Holiday	,046	,175	,023	,265	,792	,829	1,206
	Christmas	-,189	,172	-,101	-1,101	,275	,763	1,311
	MPAA_D	,113	,128	,078	,879	,383	,818	1,223
	Comedy_D	,071	,113	,057	,631	,530	,796	1,257

a. Dependent Variable: LN_OpeningGross

The Breusch-Pagan test yields a p-value of 0.38, so the null hypothesis of homoskedasticity cannot be rejected. The residual plot provides no indication of curvature.

The Durbin-Watson statistic does not suggest autocorrelation. All variance inflation factors are small, so we do not face a multicollinearity problem.

The regression equation is as follows:

$$\text{LN(US\_OpeningGross)} = 14.967 + 5*10^{-9}\,\text{Budget} + 4.738*10^{-4}\,\text{OpeningTheaters} - 0.089\,\text{KnownStory} + 0.354\,\text{Sequel} - 0.256\,\text{Summer} + 0.046\,\text{Holiday} - 0.189\,\text{Christmas} + 0.113\,\text{MPAA\_D} + 0.071\,\text{Comedy\_D}$$

b. Drop all variables from the regression that are not significant at a 10 percent level of significance. Report the final regression.

After sequentially dropping all insignificant independent variables, we obtain the following regression:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	14,935	,197		75,630	,000		
	Budget	,0000000051	,0000000020	,206	2,487	,015	,913	1,095
	Opening Theatres	,0004903720	,0000708329	,593	6,923	,000	,856	1,168
	Sequel	,362	,156	,194	2,321	,023	,902	1,109
	Summer	-,234	,100	-,187	-2,338	,022	,984	1,016

a. Dependent Variable: LN_OpeningGross

At this point we should check the assumptions underlying the regression model once again. There appears to be no curvature and no heteroskedasticity problem.

$$\text{LN(US\_OpeningGross)} = 14.935 + 5.1*10^{-9}\,\text{Budget} + 4.904*10^{-4}\,\text{OpeningTheaters} + 0.362\,\text{Sequel} - 0.234\,\text{Summer}$$

c. Carefully interpret the slope coefficient of each variable in the regression.

In a log-linear model, an increase of 1 unit in x will result in an expected multiplicative change of $EXP(\beta_x)$ in y, holding all other variables constant. We express such changes in percentage terms using $EXP(\beta_x) - 1$.

- *$5.1*10^{-9}$ Budget:* (Due to the scale of the given data, it makes more sense to speak of a budget increase of one million U.S. dollars than of a single dollar increase.) A $1 million increase of budget will result, on average, in an increase of U.S. opening gross by $EXP(5.1*10^{-9} \times 10^{6}) - 1 = 0.5\%$, holding all other variables constant.
- *$4.9*10^{-4}$ OpeningTheaters:* For each additional theater showing a movie on the opening weekend, the movie's U.S. opening gross will, on average, increase by EXP(0.00049) – 1 = 0.05%, holding all other variables constant.
- *0.362 Sequel:* Sequels have, on average, a U.S. opening gross that exceeds that of non-sequels by EXP(0.362) – 1 = 43.6%, holding all other variables constant.
- *-0.234 Summer:* Movies released in the summer have, on average, a U.S. opening gross that is 20.9 percent smaller than the opening gross of movies released at other times (because EXP(-0.234) – 1 = -20.9%), holding all other variables constant.

d. Suppose the number of movie theatres showing a movie on the opening weekend increases by one hundred. Provide a point estimate and a 95 percent confidence interval for the expected change in the opening weekend box-office revenue.

The confidence interval can be calculated by:

$(4.904 \times 10^{-4} \pm 1.9944 \times 7.083 \times 10^{-5}) \times 100 = [0.0349, 0.0632]$

(using the Excel command TINV(0.05,70) = 1.9944). Therefore, the expected increase in opening gross is estimated to fall in the range of 3.55 percent and 6.52 percent (because EXP(0.0349) – 1 = 0.0355, EXP(0.0632) – 1 = 0.0652).

7. Griffith also knew the even stronger version of that age-old Hollywood wisdom which stated that 25 percent of a movie's U.S. box-office gross revenue came in during the opening weekend. All this conventional wisdom made him curious to examine the relationship between total U.S. box-office gross and opening weekend box-office gross.

a. Run a simple linear regression predicting total U.S. box-office gross from opening weekend box-office gross.

The resulting regression therefore is:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	5108220,42	4502659,66		1,134	,260		
	Opening Gross	3,121	,218	,859	14,310	,000	1,000	1,000

a. Dependent Variable: Total U.S. Gross

Total U.S. Gross = 5'108'220.42 + 3.121 OpeningGross

b. If the stronger version of that age-old wisdom were true, that is, if indeed 25 percent of a movie's U.S. box-office gross revenue came in during the opening weekend, what would the value of the slope coefficient in the linear regression model have to be?

For the stronger version of the age-old wisdom to hold true, the value of the slope coefficient in the linear regression model has to be 4. If 25 percent of a movie's U.S. box-office gross revenue comes in during the opening weekend, then the total gross is four times larger than the opening gross.

c. Can the age-old wisdom be rejected based on the simple linear regression?

The linear regression provides a slope coefficient of 3.121 instead of the "needed" value of 4. We can test whether the value of 3.121 is statistically significantly different from 4:

H_0: $\beta_{OpeningGross} = 4$

H_A: $\beta_{OpeningGross} \neq 4$

$t = (3.121 - 4) / 0.218 = -4.032$

The corresponding p-value is 0.00013. Therefore, the null hypothesis can be rejected at all common significance levels. We conclude that the age-old wisdom appears to be wrong.

d. Critique the statistical analysis in part (c).

The Breusch-Pagan test reports a p-value of 0.00, which is a strong indication of heteroskedasticity. Therefore, we should not use the reported standard error of the coefficient for a hypothesis test as we did in the answer to the previous question.

At this point the instructor could introduce heteroskedasticity-consistent standard errors.

e. Determine a sound regression model predicting total U.S. box-office gross from opening weekend box-office gross.

For a log-linear regression, the Breusch-Pagan test reports a p-value of 0.69, so the null hypothesis of homoskedasticity cannot be rejected. The residual plot provides no indication of curvature. The resulting regression model is as follows:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	16,873	,086		197,063	,000		
	Opening Gross	,000000047	,000000004	,800	11,384	,000	1,000	1,000

a. Dependent Variable: LN_TotalU.S.Gross

$$LN(\text{Total-U.S.Gross}) = 16.873 + 4.7*10^{-8}\ \text{OpeningGross}$$

f. Examine the validity of the age-old wisdom using the new regression.

Obviously, for the log-linear regression the coefficient of the independent variable cannot be 4, even if the age-old wisdom were true. In fact, as the comparison of a few fitted values shows, the ratio of predicted total U.S. gross to opening gross varies considerably for different values of the opening gross.

g. What proportion of the variation in total U.S. box-office gross revenue can be explained by variation in the opening weekend box-office gross revenue?

The desired proportion is given by the R^2-statistic:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,800[a]	,640	,635	,39532	1,725

a. Predictors: (Constant), Opening Gross
b. Dependent Variable: LN_TotalU.S.Gross

The result indicates that 64 percent of the variation in U.S. box-office gross can be explained by the opening weekend box-office gross.

8. Investors often wonder just how much influence press reviews have on box-office admissions. If Meyer turned out to be a Flags of Our Fathers *fan who blamed the failure of his favorite film on the evil critics, how would Griffith respond?*

a. Determine a sound regression model predicting total U.S. box-office gross revenue. Consider all factors known after the opening weekend, including those known before production and those known only before opening weekend, as well as the opening box-office gross and the critics' opinion score.

For the initial linear regression, the Breusch-Pagan test indicates the presence of heteroskedasticity. Therefore, we discard the linear model and run a log-linear regression with the dependent variable LN(Total U.S. Gross).

b. Drop all variables from the regression that are not significant at a 10 percent level of significance. Report the final regression.

The final regression is as follows:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	15,761	,241		65,502	,000		
	Opening Gross	,000000030	,000000005	,520	5,900	,000	,443	2,259
	Budget	,000000004	,000000002	,168	2,365	,021	,681	1,469
	Opening Theatres	,000195910	,000075432	,210	2,597	,012	,525	1,905
	Known Story	-,056	,085	-,041	-,660	,511	,880	1,137
	Sequel	,098	,144	,047	,681	,498	,738	1,354
	Summer	-,239	,094	-,170	-2,546	,013	,775	1,290
	Holiday	,086	,145	,038	,592	,556	,822	1,217
	Christmas	-,033	,145	-,016	-,227	,821	,726	1,377
	MPAA_D	-,052	,109	-,032	-,480	,633	,765	1,307
	Critics Opinion	,013	,003	,325	4,363	,000	,621	1,610
	Comedy_D	,167	,094	,118	1,776	,080	,775	1,290

a. Dependent Variable: LN_TotalU.S.Gross

LN(Total U.S. Gross) = 15.763 + $3.2*10^{-8}$ OpeningGross + $4*10^{-9}$ Budget + $1.983*10^{-4}$ OpeningTheaters – 0.200 Summer + 0.012 CriticsOpinion + 0.174 Comedy_D

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	15,763	,228		69,272	,000
	Opening Gross	,000000032	,000000004	,555	7,140	,000
	Budget	,000000004	,000000001	,154	2,346	,022
	Opening Theatres	,000198275	,000073200	,213	2,709	,009
	Summer	-,200	,083	-,142	-2,425	,018
	Critics Opinion	,012	,003	,298	4,488	,000
	Comedy_D	,174	,085	,124	2,042	,045

a. Dependent Variable: LN_TotalU.S.Gross

c. Consider a movie with the characteristics of Flags of Our Fathers*. Using the regression from part (b), provide a point estimate and a 95 percent prediction interval for the total U.S. gross revenue of a movie with such characteristics.*

To obtain the requested fitted value, we have to plug in the specific values for all remaining independent variables for the specific movie.

The value for the movie *Flags of Our Fathers* is:

$$\text{LN(Total U.S. Gross)} = 15.763 + 3.2*10^{-8}*10'245'190 + 4*10^{-9}*90'000'000 + 1.98*10^{-4}*1'876 - 0.2*1 + 0.174*0 + 0.012*79 = 17.57$$

$$\text{Total U.S. Gross} = e^{17.57} = \$42'712'400$$

The corresponding prediction interval is:

[EXP(16.97450), EXP(18.40531)] = [23'546'843, 98'474'290]

d. Advise Griffith on how much he should be willing to invest in order to influence the critics to gain an extra ten points in the opinion score of a movie with the characteristics of Flags of Our Fathers*, thereby earning such a film a score of eighty-nine points instead of seventy-nine points.*

The ten-point increase in the variable CriticsOpinion leads, on average, to an expected change in the dependent variable of

$$\Delta\text{LN(Total U.S. Gross)} = 0.0120895*10 = 0.120895,$$

holding all other variables constant. Thus, the expected multiplicative change in total U.S. gross is:

$$e^{0.120895} = 1.128506$$

Thus, total U.S. gross increases by approximately 13 percent. If we knew the percentage of the gross going to the producer of a movie, then we would have an estimate for the amount Griffith should have been willing to spend for a ten-point increase of the critics' opinion score. (To account for the uncertainty in this estimate, we should calculate a confidence interval for the expected change.)

9. Griffith surmised that poor reviews affected the total U.S. box-office gross of comedies less strongly than the total U.S. gross of movies from other genres. In particular, he theorized that the critics' opinion score had a significantly smaller influence on total U.S. box-office gross for comedies than for non-comedies.

a. Modify your regression from Question 8 to examine Griffith's claim. Can you prove his theory?

This problem requires the introduction of an interaction variable (slope dummy). The new variable, Critics_x_Comedy, is the product of the variable Critics Opinion and the dummy Comedy. Including this slope dummy in the regression model leads to the following log-linear regression:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	15,663	,243		64,483	,000
	Opening Gross	,000000032	,000000004	,553	7,132	,000
	Budget	,000000004	,000000001	,148	2,248	,028
	Opening Theatres	,000204796	,000073240	,220	2,796	,007
	Summer	-,200	,082	-,142	-2,425	,018
	Critics Opinion	,014	,003	,341	4,487	,000
	Comedy_D	,472	,271	,335	1,738	,087
	Critics_x_Comedy	-,006	,006	-,216	-1,154	,253

a. Dependent Variable: LN_TotalU.S.Gross

$$\text{LN(Total-U.S.Gross)} = 15.663 + 3.2*10^{-8}\ \text{OpeningGross} + 4*10^{-9}\ \text{Budget} + 2.048*10^{-4}\ \text{OpeningTheaters} - 0.2\ \text{Summer} + 0.014\ \text{CriticsOpinion} + 0.472\ \text{Comedy\_D} - 0.006\ \text{Critics\_x\_Comedy}$$

The coefficient of the slope dummy is -0.006, which shows that, on average, the impact of critics' opinion scores is smaller for comedies than for non-comedies. However, the two-tailed p-value for the coefficient of the slope dummy is 0.253. Because the question asks for a one-tailed hypothesis test with the hypotheses

H_0: $\beta_{\text{Critics\_x\_Comedy}} \geq 0$

H_A: $\beta_{\text{Critics\_x\_Comedy}} < 0$,

the relevant p-value is 0.1265. As a result, we cannot reject the null hypothesis at a significance level of 10 percent. The critics' opinion score does not have a significantly smaller influence on the total U.S. gross for comedies than for non-comedies.

10. Standard paychecks for A-list stars such as George Clooney or Brad Pitt are routinely on the order of $15 million or more. Producers hope that famous faces in a film will guarantee packed movie theatres. Griffith concluded that it is not really a large budget per se that has a strong positive effect on total U.S. box-office gross revenue. Instead the number of star movie actors in a film drives up total U.S. gross. He regretted that he did not have any data on the number of movie stars in his data set to examine his claim.

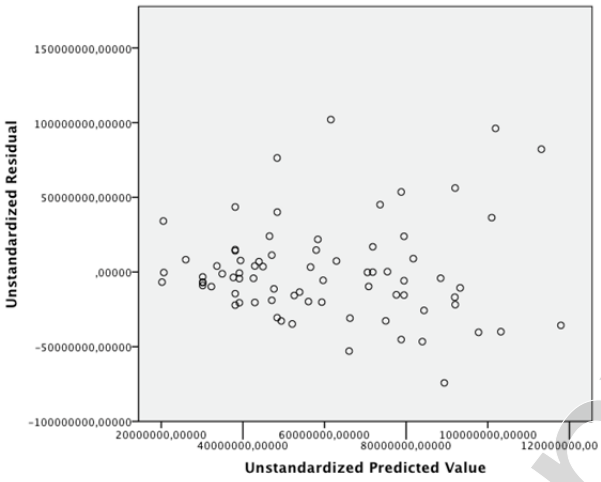
a. Consider a variable called "star power" that reports the number of A-list stars for a movie. If you had data for this variable and added it to your regression from Question 9, what would have to be true for the slope coefficient of star power and how would the slope coefficient of the budget variable have to change for Griffith's conclusion to be correct?

If Griffith's conclusion were correct, then his regression would suffer from an omitted variable bias, as the missing variable Stars would have a statistically significant positive effect on the dependent variable Total U.S. Gross. According to his beliefs, this effect is currently represented by the statistically significant positive coefficient of the variable Budget. Adding the variable Stars to the regression would remove this bias. The coefficient of Budget should be much smaller in a new regression than in his most recent regression—perhaps even insignificant. The coefficient of Stars would have to be positive as well as statistically significant. This conclusion

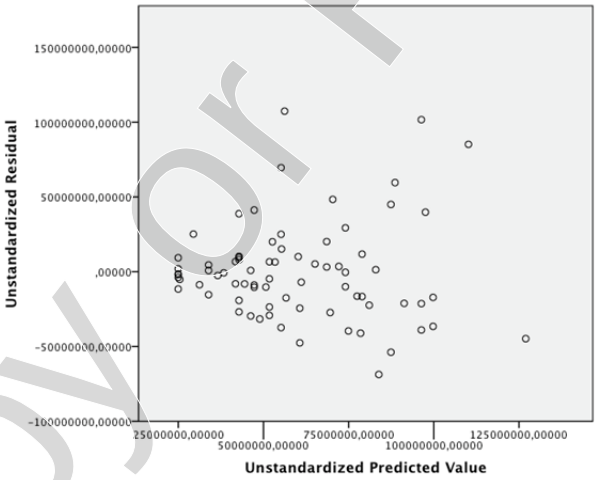
rests on Griffith's assumption of a strong positive correlation between the variables Budget and Stars.

Below we display the residual plots for the regressions in Questions 5–8:

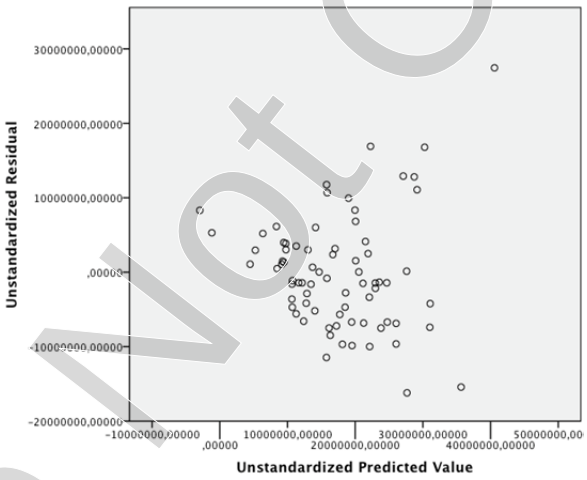
Question 5a:



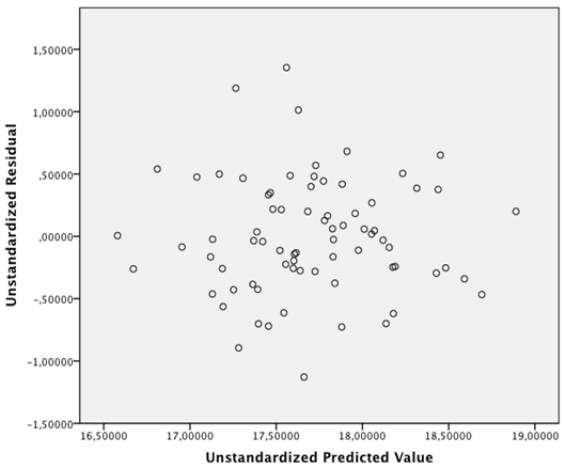
Question 5b:



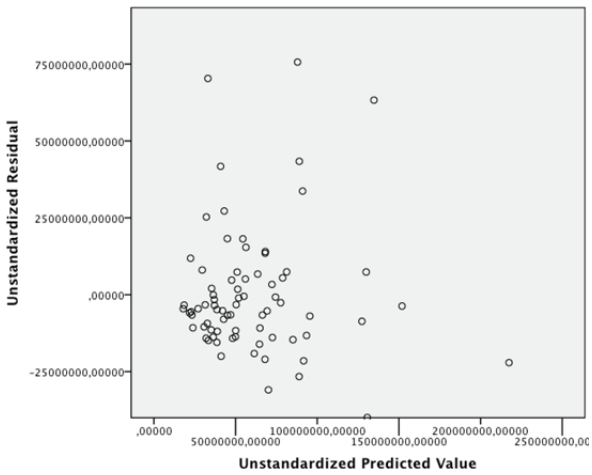
Question 6a:



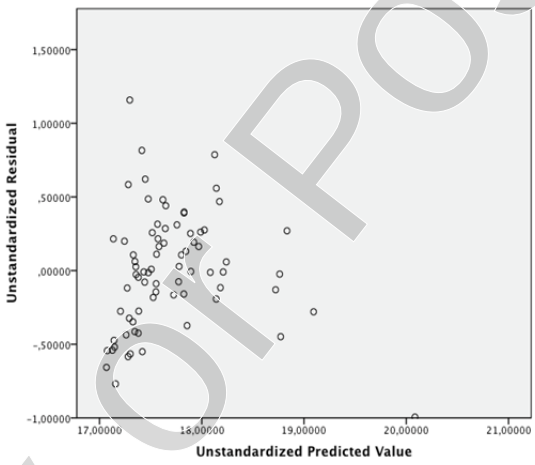
Question 6b:



Question 7d:

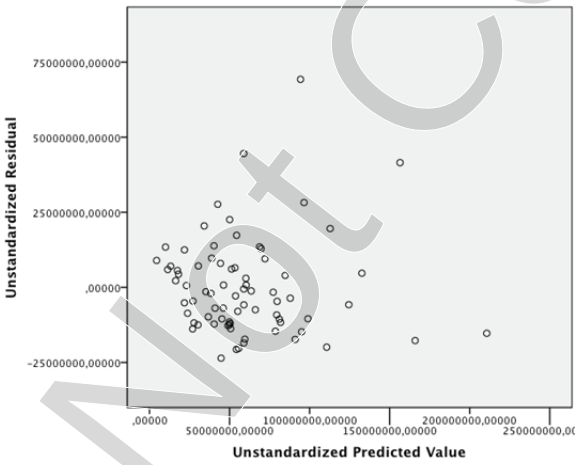


Question 7e:



Question 8a:

Residual plot for the original regression before dropping any variables:



Residual plot for the final regression (after dropping variables):