Box Office Prediction Model

Predicting US Gross Box Office Receipts Along with Days in Release

Jonathan Presto
Statistics 685 Masters Project
Texas A&M University
March 24, 2015

Introduction

- **Problem Statement:** The goal of this project is to develop models which predict the following outcome variables:
  - 1. US **total gross** box office receipts of films released sometime between 2010 and 2013 (y1)
  - 2. The number of days **in release** each film was open in the US (y2)
- **Multiple Linear Regression** (MLR) Techniques
  - Weighted Least Squares (WLS)
  - Variable Selection Methods
- **Application:** MLR Techniques popular in forecasting future demands for other industries (music album sales, car sales, call center call flows, etc.)

Data Set

Initial raw data set

- This raw dataset contains the US **total gross** box office receipts of English language films released sometime between 2010 and 2013 (y1, one response variable) and the number of days **in release** each film was open in the US (y2, the other response variable) along with the corresponding information for each film:

VARIABLE	DESCRIPTION
Title –	Movie title
Academy Nominations –	Number of Academy Award nominations received
Academy Awards –	Number of Academy Awards received
Studio –	Name of the studio
Major Studio –	A dummy variable which is 1 if the studio is considered to be a "major studio"
Director –	Name of the director(s)
Runtime –	Run time of the movie (in minutes)
Genre –	Genre of the movie
Rating –	Movie rating
Production Budget –	Movie production budget (in millions)
Release Date –	Date on which film was released
Close Date –	Date on which film closed
Widest Release –	Largest number of theaters at which movie was shown
Opening Weekend Gross –	Gross Box office receipts for the opening weekend
Limited Opening –	A dummy variable which is 1 if the film had a limited opening
Opening Weekend Rank –	Ranking of opening weekend box office receipts for the relevant weekend
Opening Weekend Theaters –	Number of theaters at which the movie was shown on the opening weekend
Critic Rating –	% of critics giving the movie a positive review as reported by Rotten Tomatoes
Audience Rating –	% of audience giving the movie a positive review as reported by Rotten Tomatoes
Month of Release –	Month during which film was released
Year –	Year during which film was released

Data Set

Custom Variables Added

- **Genre indicators** – some movies call fall into multiple genres

GENRE INDICATORS			
Action__c	Crime__c	Historical__c	Romance__c
Adventure__c	Drama__c	Horror__c	Sci_Fi__c
Animation__c	Family__c	Music__c	Sports__c
Comedy__c	Fantasy__c	Period__c	Thriller__c
Concert__c			War__c

- **Movie Title indicators** – outliers identified from kitchen-sink model without the movie indicators. Helps minimize bias caused by bad leverage.

VARIABLE	DESCRIPTION
I_Dallas__c	Dallas Buyer's Club
I_Ramona__c	Ramona and Beezus
I_Winter__c	Winter's Bone
I_MidParis__c	Midnight in Paris
I_Winnie__c	Winnie the Pooh
I_XmasCand__c	The Christmas Candle
I_Airbend__c	The Last Airbender
I_Getlow__c	Get Low
I_Hugo__c	Hugo
I_Anna__c	Anna Karenina
I_August__c	August: Osage County

Data Set

Custom Variables Added

- **Miscellaneous** – custom class and indicator variables

VARIABLE	TYPE	DESCRIPTION
Rating__c	Class	G PG PG-13 R Other(NC-17 or Unrated)
Season__c	Class	Fall(Aug-Oct) Winter(Nov-Jan) Spring(Mar-May) Summer(Jun-Jul)
Novbkst__c	Indicator	Based on novels but also non-fiction, short stories, magazines stories or articles, poems, and comic book
Trueins__c	Indicator	True stories/inspirational movies
DisFlood__c	Indicator	Movie release between Apr-May 2011: Mississippi river floods, tornado outbreak, Joplin tornado
DisSandy__c	Indicator	Movie released around the time of Hurricane Sandy (late October 2012)
DisFire__c	Indicator	Movie released around the time of Yarnell Hill Fire (late June to July 2013)
AcademyNominee__c	Indicator	1 if movie was nominated for at least 1 academy award, 0 otherwise
AcademyAwardWinner__c	Indicator	1 if movie won at least 1 academy award, 0 otherwise
DirectorNominee__c	Indicator	1 if movie director was nominated for Best Director in Oscar's within the last 20 years

Data Analysis

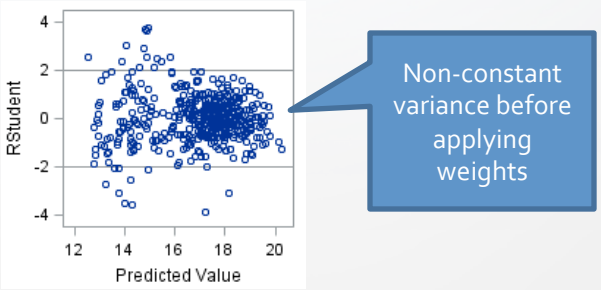
Steps Taken

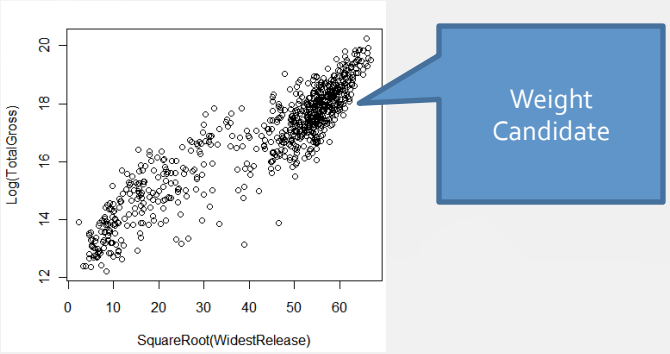
- Step 1 - Exploratory data analysis
  - Scatter plots, box-plots, add-variable plots, etc. – visual aids to assess linear relationship between response and predictor variables; transformation of response and/or predictors
  - Interaction effects
  - 737 total observations (491 used for training, 246 for test)
- Step 2 - Kitchen Model
  - Throw in all variables into the model using training set (randomly sampled from full set)
  - Evaluate if model assumptions are satisfied
  - Marginal model plots - add/remove variables until valid model is obtained
- Step 3 - Variable Selection
  - On training data, run variable selection procedures: stepwise, forward and backward
  - Determine best subset of predictors to simplify model using BIC
- Step 4 - Model Verification
  - Test validity of model from test data
- Step 5 - Report prediction performance for each variable selection method
- Step 6 - Model Interpretation

Data Analysis (Total Gross)

Step 1 - Exploratory

- Weighted Least Squares
  - Weight = SqrtWidestRelease





- Transformed predictors used in prediction models
  - Log transformations:
    - OpeningWeekendGross
    - ProductionBudget
  - Square root transformations
    - InRelease
- Interaction effects used in prediction models
    - MajorStudio*AudienceRating
    - MajorStudio*LogProductionBudget
    - CriticRating*AudienceRating
    - SqrtInRelease*Season

Data Analysis (Total Gross)

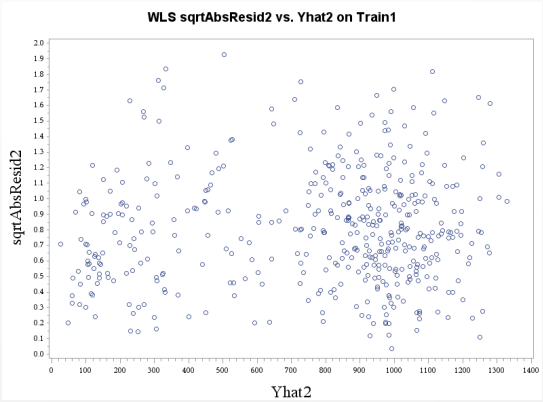
Step 2 – Kitchen Sink Model

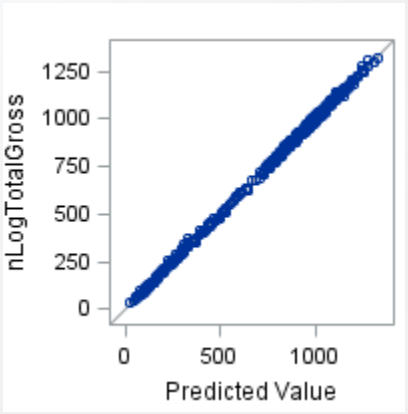
```
ods graphics on;
Proc GLM data = Train1New1  PLOTS=DIAGNOSTICS ;
class Rating__c Season__c Year_Char ;
title1 ls=1.5 "Kitchen Sink Model on Train1 WLS";

model nLogTotalGross =  Rating__c Season__c Year_Char
AcademyAwardWinner__c nAcademyAwards nAcademyNominations AcademyNominee__c
nAudienceRating nCriticRating DirectorNominee__c
Action__c Adventure__c Animation__c Comedy__c Concert__c
Crime__c Family__c Fantasy__c Historical__c Horror__c
Music__c Period__c Romance__c Sci_Fi__c Sports__c Thriller__c
MajorStudio__c nRuntime nRuntimeSq
nLogOpeningWeekendGross nOpeningWeekendRank nOpeningWeekendTheaters
nSqrtInRelease nLogProductionBudget
Novbkst__c Trueins__c DisFlood__c DisSandy__c DisFire__c
/*  interaction terms */
MajorStudio__c*nCriticRating  MajorStudio__c*nAudienceRating nCriticRating*nAudienceRating
MajorStudio__c*nAcademyNominations MajorStudio__c*nLogProductionBudget
MajorStudio__c*nOpeningWeekendRank
Season__c*Rating__c Season__c*nSqrtInRelease Season__c*nLogProductionBudget
Season__c*nCriticRating Season__c*nAudienceRating Season__c*nRuntime
nAcademyAwards*DirectorNominee__c nAcademyNominations*DirectorNominee__c
nAudienceRating*DirectorNominee__c nCriticRating*DirectorNominee__c nRuntime*DirectorNominee__c
nLogProductionBudget*DirectorNominee__c nLogOpeningWeekendGross*DirectorNominee__c
nOpeningWeekendRank*DirectorNominee__c nOpeningWeekendTheaters*DirectorNominee__c
/* movie indicators */
I_Dallas__c I_Ramona__c I_Winter__c I_MidParis__c I_Winnie__c
I_XmasCand__c I_Airbend__c I_Getlow__c I_Hugo__c I_Anna__c I_August__c
/ss3 solution tolerance;
output out = Train1New2 predicted=yhat2 residual=resid2 stdr=eresid2 cookd=cooksD2 rstudent=rstud2;
run; quit;
ods graphics off;
```

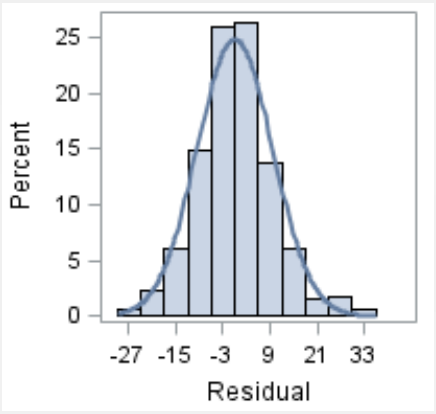
Data Analysis (Total Gross)

Step 2 – Kitchen Sink Model

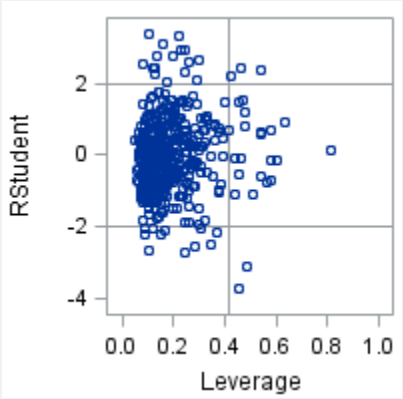
- Diagnostic plots

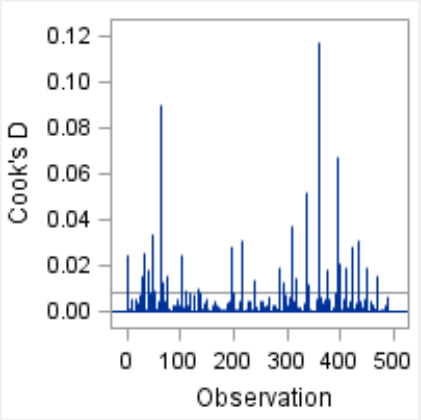






- ✓ Normality of error distribution
- ✓ Homoscedasticity
- ✓ Linear relationship between response and each predictor (holding others fixed)
- ✓ Independence of the errors

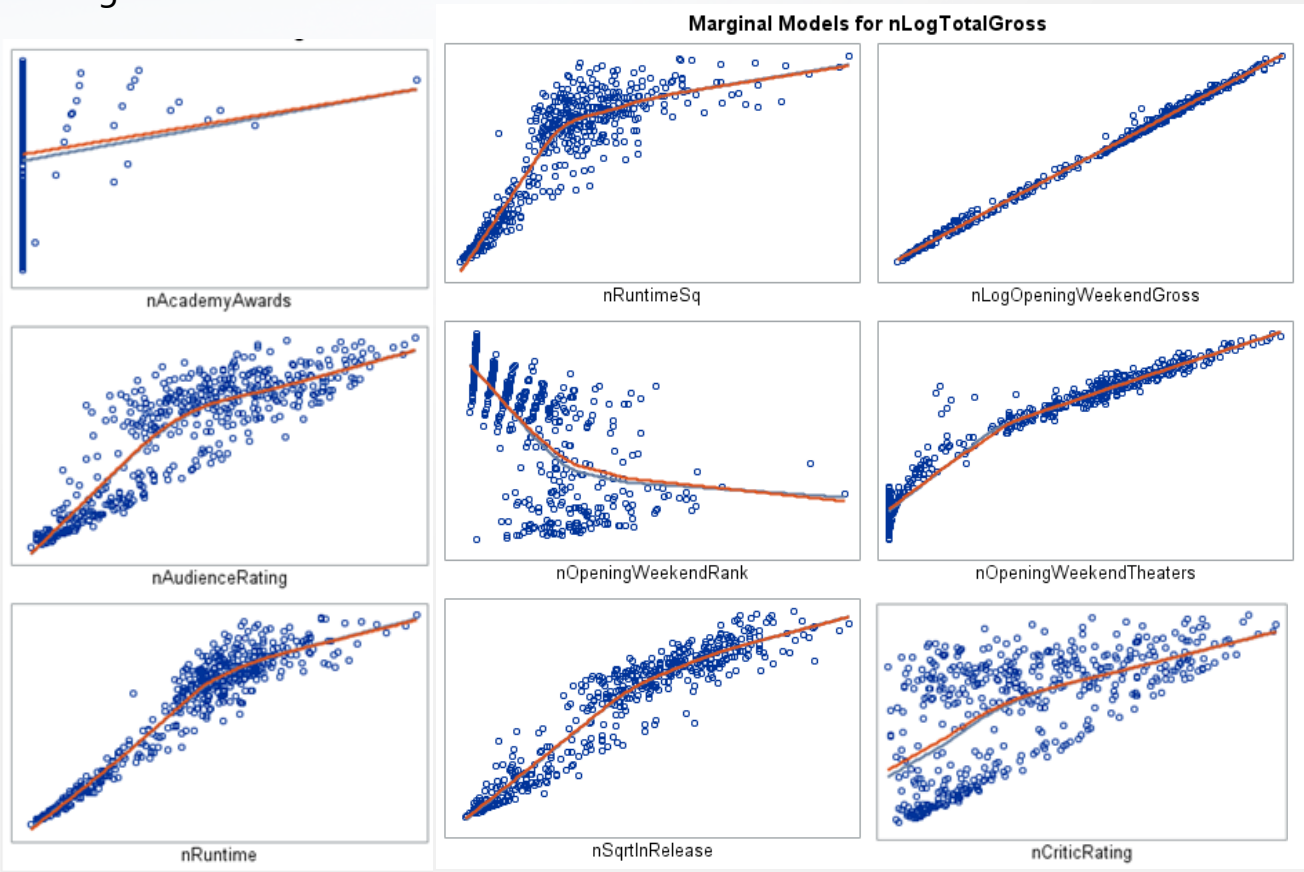




Data Analysis (Total Gross)

Step 2 – Kitchen Sink Model

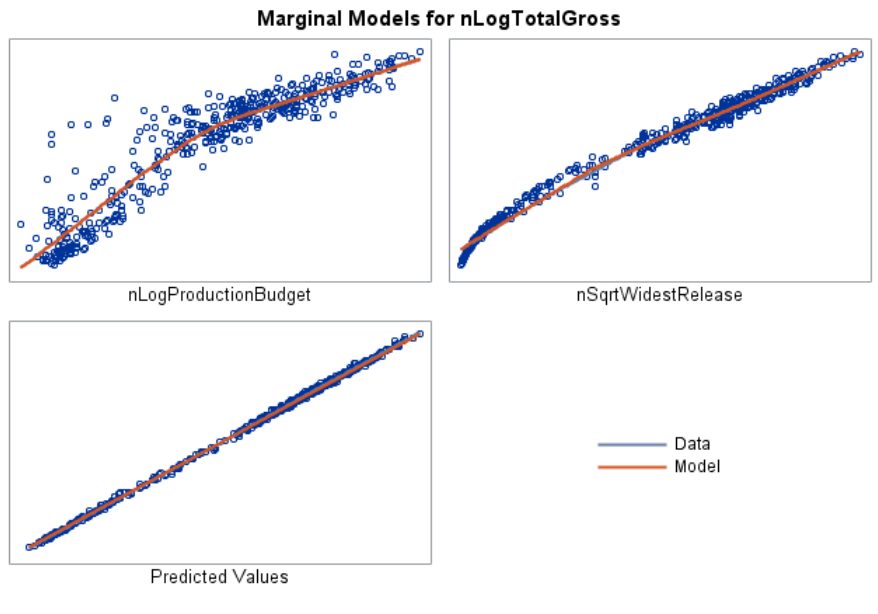
- Marginal Model Plots



Data Analysis (Total Gross)

Step 2 – Kitchen Sink Model

- Marginal Model Plots



✓ Blue lines align well with the red lines → model is a good fit

Data Analysis (Total Gross)

Step 3 – Variable Selection

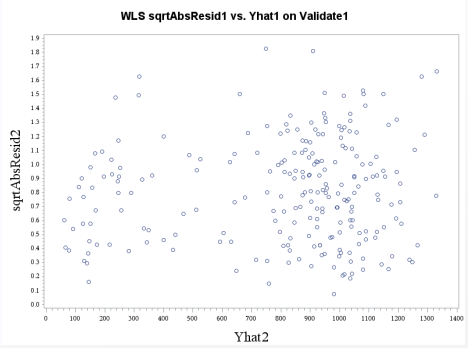
- SAS - **PROC GLMSelect** (select=BIC)
- Backward selection – included more variables, some overlap

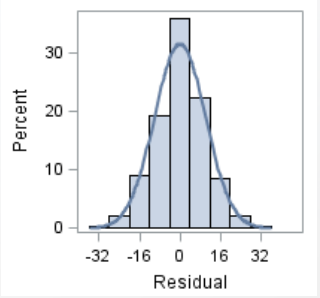
Stepwise Selection	Forward Selection	Backward Selection
Rating__c	Rating__c	
Year_Char	Year_Char	Year_Char
nAcademyNominations	nAcademyNominations	
nAudienceRating	nAudienceRating	nAudienceRating
nCriticRating	nCriticRating	nCriticRating
DirectorNominee__c	DirectorNominee__c	DirectorNominee__c
Action__c	Action__c	Action__c
Adventure__c	Adventure__c	Adventure__c
Comedy__c	Comedy__c	Comedy__c
		Historical__c
Horror__c	Horror__c	Horror__c
		Sports__c
Thriller__c	Thriller__c	Thriller__c
nLogOpeningWeekendGross	nLogOpeningWeekendGross	nLogOpeningWeekendGross
nOpeningWeekendRank	nOpeningWeekendRank	nOpeningWeekendRank
nOpeningWeekendTheaters	nOpeningWeekendTheaters	nOpeningWeekendTheaters
Novbkst__c	Novbkst__c	Novbkst__c
nCriticRating*nAudienceRating	nCriticRating*nAudienceRating	
MajorStudio__c*nOpeningWeekendRank	MajorStudio__c*nOpeningWeekendRank	
Season__c*nSqrtInRelease	Season__c*nSqrtInRelease	Season__c*nSqrtInRelease
Season__c*nLogProductionBudget	Season__c*nLogProductionBudget	
nLogOpeningWeekendGross*DirectorNominee__c	nLogOpeningWeekendGross*DirectorNominee__c	
nOpeningWeekendRank*DirectorNominee__c	nOpeningWeekendRank*DirectorNominee__c	nOpeningWeekendRank*DirectorNominee__c
	nOpeningWeekendTheaters*DirectorNominee__c	nOpeningWeekendTheaters*DirectorNominee__c
		AcademyAwardWinner__c
		nAcademyAwards
		MajorStudio__c
		nRuntime
		nRuntimeSq
		nSqrtInRelease
		Trueins__c
		MajorStudio__c*nAudienceRating
		MajorStudio__c*nCriticRating
		MajorStudio__c*nAcademyNominations
		MajorStudio__c*nLogProductionBudget
		MajorStudio__c*nCriticRating
		Season__c*nAudienceRating
		nAcademyAwards*DirectorNominee__c
		nAcademyNominations*DirectorNominee__c
		nAudienceRating*DirectorNominee__c

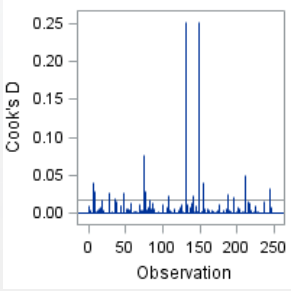
Data Analysis (Total Gross)

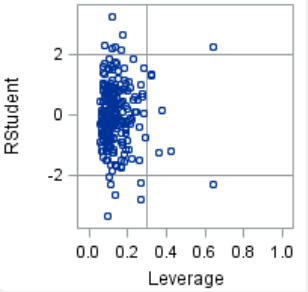
Step 4 – Model Verification

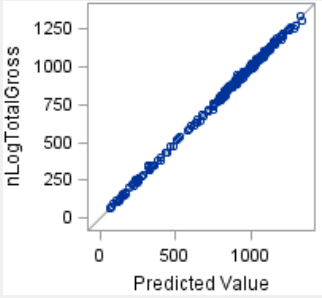
- Stepwise Selection: Reduced model using test data

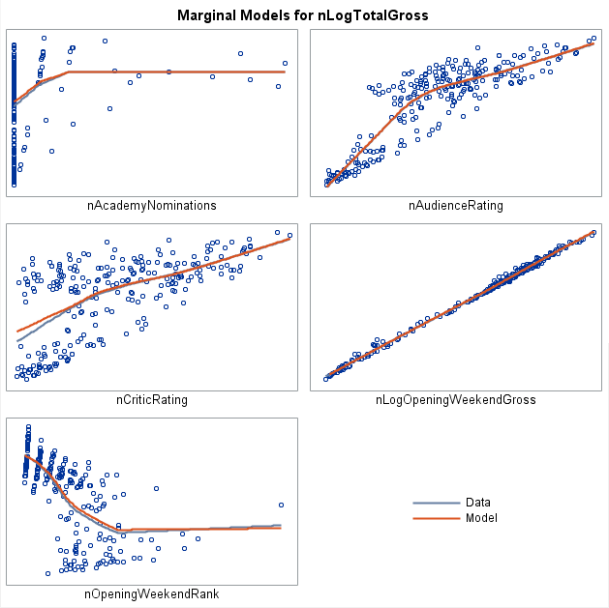


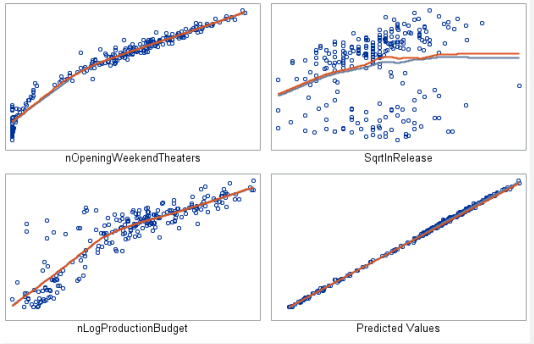












- ✓ Normality of error distribution
- ✓ Homoscedasticity
- ✓ Linear relationship between response and each predictor (holding others fixed)
- ✓ Independence of the errors
- ✓ Cook's D values small
- ✓ Marginal model plots indicate good fit

Data Analysis (Total Gross)

Step 4 – Model Verification

- Stepwise Selection: P-values and VIF on test data set
  - 14 of the 34 parameter estimates remain significant (p-value<=0.025)
  - Majority of VIF are less than 10

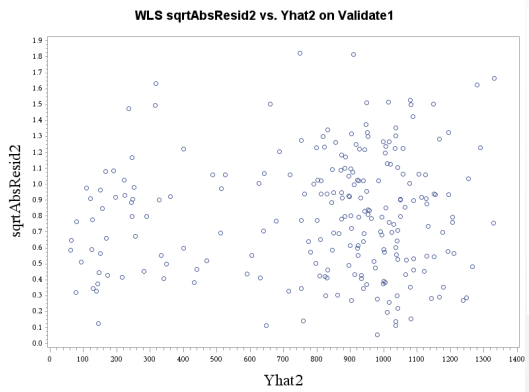
Parameter	Estimate	Pr > |t|	VIF
Intercept	-3.94	0.4015	0.00
Rating__c G	-15.22	0.1095	1.00
Rating__c PG	-3.85	0.2084	1.00
Rating__c PG-13	2.47	0.2114	1.14
Year_Char 2010	-0.83	0.7022	1.02
Year_Char 2011	0.08	0.9734	1.13
Year_Char 2012	-1.91	0.3533	1.43
nAcademyNominations	0.04	0.0036	1.02
nAudienceRating	0.01	0.0003	1.41
nCriticRating	0.01	<.0001	3.13
DirectorNominee__c	9.17	0.4947	1.37
Action__c	-4.15	0.0697	1.18
Adventure__c	-5.06	0.1886	1.20
Comedy__c	5.54	0.0049	1.09
Horror__c	2.29	0.4818	1.20
Thriller__c	1.53	0.5332	1.31
nLogOpeningWeekendGr	0.94	<.0001	4.59
nOpeningWeekendRank	0.04	<.0001	2.22
nOpeningWeekendTheat	0.00	0.1667	17.11

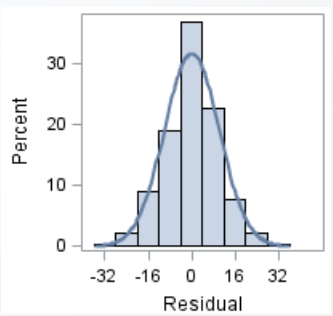
Parameter	Estimate	Pr > |t|	VIF
Novbkst__c	6.73	0.0409	1.16
nAudience*nCriticRat	0.00	<.0001	29.82
nOpeningW*MajorStudi	0.00	0.8527	1.40
nSqrtlnRel*Season__c Fall	0.09	<.0001	1.20
nSqrtlnRel*Season__c Spring	0.12	<.0001	1.41
nSqrtlnRel*Season__c Summer	0.12	<.0001	1.84
nSqrtlnRel*Season__c Winter	0.10	<.0001	9.44
nLogProduc*Season__c Fall	0.10	0.0033	12.96
nLogProduc*Season__c Spring	0.03	0.3821	17.07
nLogProduc*Season__c Summer	0.03	0.3564	14.12
nLogProduc*Season__c Winter	0.11	0.0005	11.11
DirectorN*nLogOpenin	0.00	0.6957	7.39
DirectorN*nOpeningWe	-0.04	0.1435	8.80
l_MidParis__c	-12.23	0.3333	1.32
l_Getlow__c	26.86	0.0313	1.28
l_Anna__c	28.05	0.0184	1.16

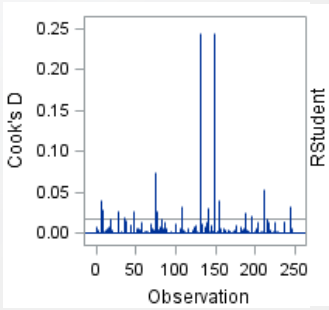
Data Analysis (Total Gross)

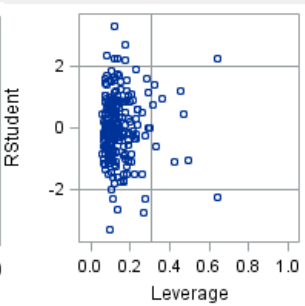
Step 4 – Model Verification

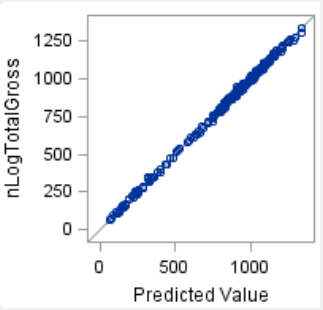
- Forward Selection: Reduced model using test data

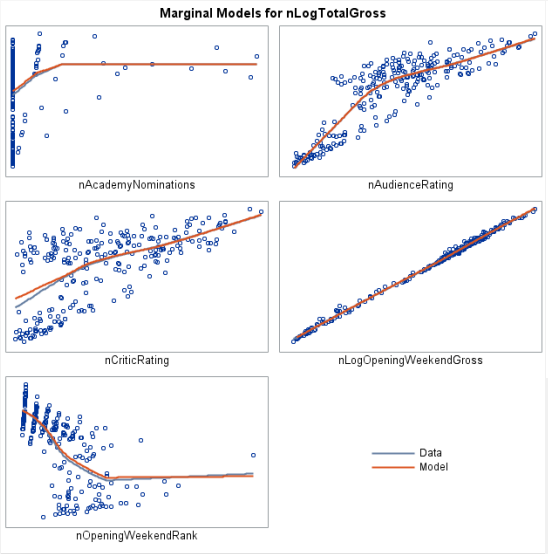


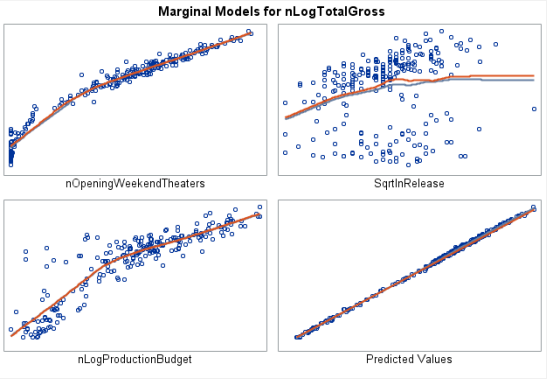












- ✓ Normality of error distribution
- ✓ Homoscedasticity
- ✓ Linear relationship between response and each predictor (holding others fixed)
- ✓ Independence of the errors
- ✓ Cook's D values small
- ✓ Marginal model plots indicate good fit

Data Analysis (Total Gross)

Step 4 – Model Verification

- Forward Selection: P-values and VIF on test data set
  - 14 of the 35 parameter estimates remain significant (p-value<=0.025)
  - Majority of VIF are less than 10

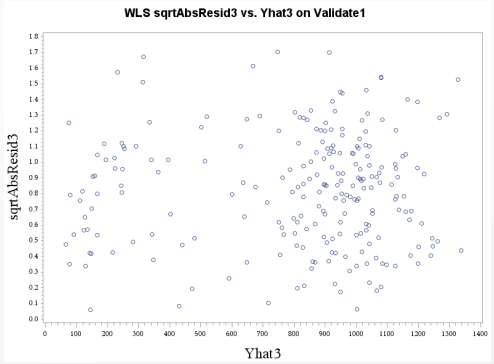
Parameter	Estimate	Pr > |t|	VIF
Intercept	-3.61	0.4461	0.00
Rating__c G	-15.35	0.1073	1.00
Rating__c PG	-3.85	0.2089	1.00
Rating__c PG-13	2.55	0.1991	1.14
Year_Char 2010	-0.73	0.7375	1.02
Year_Char 2011	0.13	0.9533	1.13
Year_Char 2012	-1.98	0.3384	1.43
nAcademyNominations	0.04	0.0163	1.02
nAudienceRating	0.01	0.0003	1.41
nCriticRating	0.01	<.0001	3.13
DirectorNominee__c	8.04	0.5545	1.37
Action__c	-4.16	0.0697	1.18
Adventure__c	-5.06	0.1894	1.20
Comedy__c	5.43	0.0061	1.09
Horror__c	2.36	0.4701	1.20
Thriller__c	1.69	0.4941	1.31
nLogOpeningWeekendGr	0.94	<.0001	4.59
nOpeningWeekendRank	0.04	<.0001	2.22
nOpeningWeekendTheat	0.00	0.1393	17.11

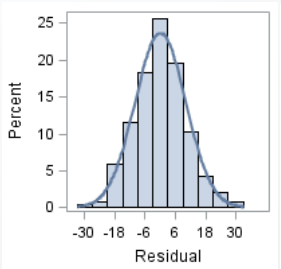
Parameter	Estimate	Pr > |t|	VIF
Novbkst__c	6.72	0.0415	1.16
nAudience*nCriticRat	0.00	<.0001	29.82
nOpeningW*MajorStudi	0.00	0.8248	1.40
nSqrtlnRel*Season__c Fall	0.09	<.0001	1.20
nSqrtlnRel*Season__c Spring	0.11	<.0001	1.41
nSqrtlnRel*Season__c Summer	0.12	<.0001	1.84
nSqrtlnRel*Season__c Winter	0.10	<.0001	9.44
nLogProduc*Season__c Fall	0.10	0.0035	12.96
nLogProduc*Season__c Spring	0.03	0.3952	17.07
nLogProduc*Season__c Summer	0.03	0.3555	14.12
nLogProduc*Season__c Winter	0.10	0.0006	11.11
DirectorN*nLogOpenin	0.01	0.6914	7.39
DirectorN*nOpeningWe	-0.05	0.1184	8.80
DirectorN*nOpeningWe	0.00	0.5664	8.80
l_MidParis__c	-14.52	0.2743	1.45
l_Getlow__c	26.80	0.032	1.28
l_Anna__c	28.64	0.0167	1.17

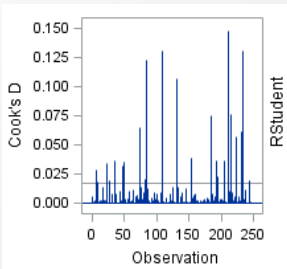
Data Analysis (Total Gross)

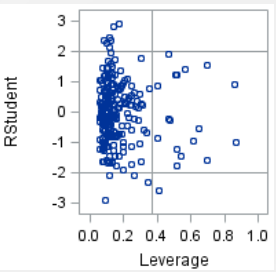
Step 4 – Model Verification

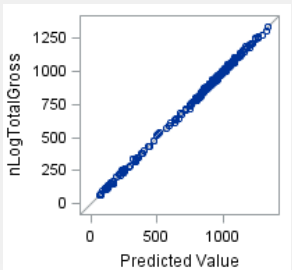
- Backward Selection: Reduced model using test data

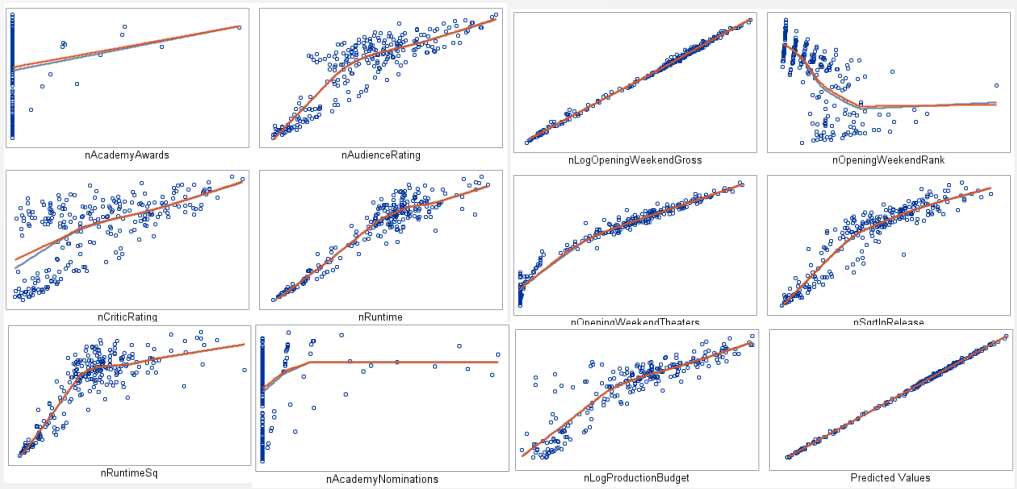












- ✓ Normality of error distribution
- ✓ Homoscedasticity
- ✓ Linear relationship between response and each predictor (holding others fixed)
- ✓ Independence of the errors
- ✓ Cook's D values small
- ✓ Marginal model plots indicate good fit

Data Analysis (Total Gross)

Step 4 – Model Verification

- Backward Selection: P-values and VIF on test data set
  - 12 of the 42 parameter estimates remain significant (p-value<=0.025)
  - Majority of VIF are < 10, six have VIFs >20

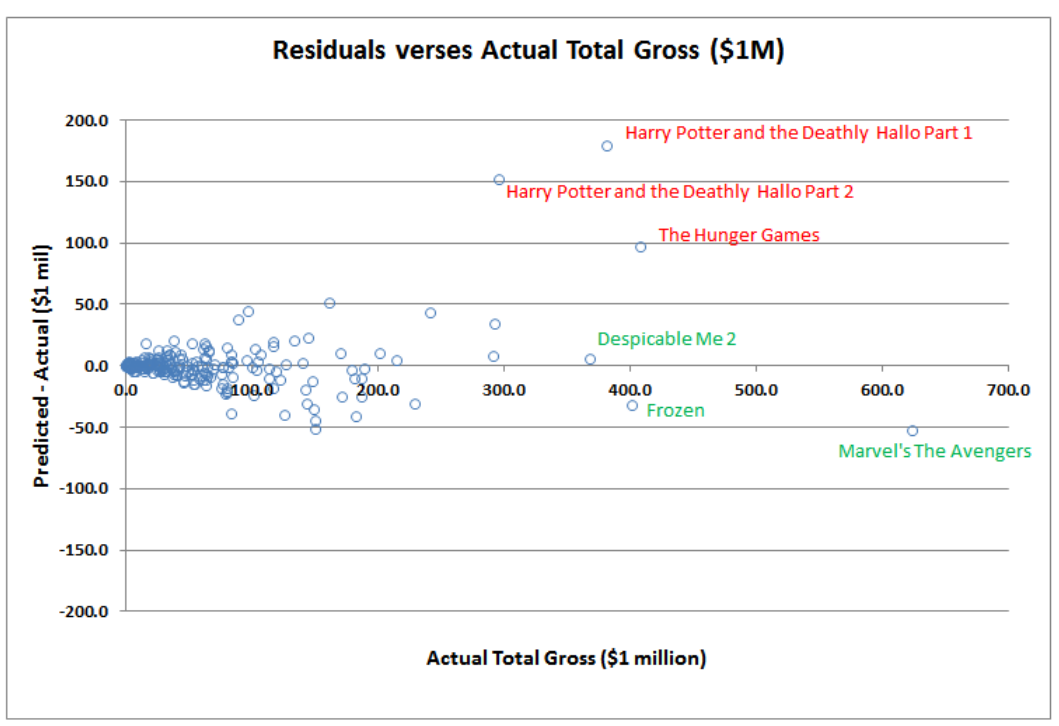
Parameter	Estimate	Pr > |t|	VIF
Intercept	-3.45	0.4633	0.00
Year_Char 2010	-0.83	0.7037	1.00
Year_Char 2011	-0.19	0.9339	1.10
Year_Char 2012	-2.50	0.242	1.42
AcademyAwardWinner__	-1.31	0.8715	1.00
nAcademyAwards	0.08	0.3385	3.06
nAudienceRating	0.01	0.0041	1.07
nCriticRating	0.00	0.0761	2.82
DirectorNominee__c	15.67	0.2466	1.11
Action__c	-2.31	0.3079	1.10
Adventure__c	-6.09	0.1269	1.16
Comedy__c	6.46	0.0014	1.07
Historical__c	10.90	0.2613	1.09
Horror__c	3.05	0.3614	1.11
Sports__c	-3.14	0.7068	1.03
Thriller__c	1.73	0.4814	1.22
MajorStudio__c	2.37	0.6282	1.44
nRuntime	0.02	0.0129	3.72
nRuntimeSq	0.00	0.0194	15.70
nLogOpeningWeekendGr	0.91	<.0001	63.17
nOpeningWeekendRank	0.03	<.0001	2.39
nOpeningWeekendTheat	0.00	0.6067	18.52
nSqrtlnRelease	0.08	<.0001	8.30

Parameter	Estimate	Pr > |t|	VIF
Novbkst__c	7.02	0.0481	1.21
Trueins__c	-4.76	0.5783	1.22
nAudience*MajorStudi	0.00	0.4432	20.51
nCriticRa*MajorStudi	0.00	0.4346	27.53
MajorStud*nAcademyNo	-0.03	0.3784	3.21
MajorStud*nLogProduc	0.05	0.0173	10.96
nSqrtlnRel*Season__c Fall	0.02	0.2895	1.32
nSqrtlnRel*Season__c Spring	0.01	0.6723	1.48
nSqrtlnRel*Season__c Summer	0.05	0.0154	2.19
nAudienceR*Season__c Fall	-0.01	0.0921	28.15
nAudienceR*Season__c Spring	0.00	0.2979	28.80
nAudienceR*Season__c Summer	-0.01	0.0056	31.87
nAcademyA*DirectorNo	-0.23	0.0343	2.34
DirectorN*nAcademyNo	0.09	0.043	17.00
nAudience*DirectorNo	0.00	0.6105	10.93
DirectorN*nOpeningWe	-0.06	0.0821	8.30
DirectorN*nOpeningWe	0.00	0.8787	8.30
l_MidParis__c	6.46	0.6469	1.56
l_Getlow__c	32.38	0.0081	1.17
l_Anna__c	31.60	0.0225	1.51

Data Analysis (Total Gross)

Step 5 – Prediction performance

- Stepwise and Forward Selection performed slightly better than Backward Selection
- Graphical representation of performance (stepwise selection)



Prediction Power	Stepwise	Forward	Backward
Within ±5%	23%	23%	21%
Within ±10%	38%	38%	36%
Within ±15%	53%	53%	53%
Within ±20%	62%	62%	64%
Within ±25%	71%	71%	70%
Adjusted R2	0.99	0.99	0.99

- Predicted some relatively high grossing films within 8% of actual
  - Marvel's The Avengers
  - Frozen
  - Despicable Me 2
- Over-predicted high grossing films
  - Harry Potter and the Deathly Hallo Part 1.
  - Harry Potter and the Deathly Hallo Part 2.
  - The Hunger Games

Data Analysis (Total Gross)

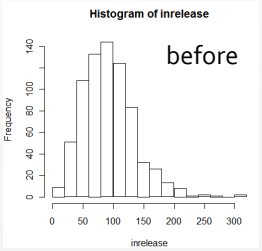
Step 6 – Final Model Interpretation

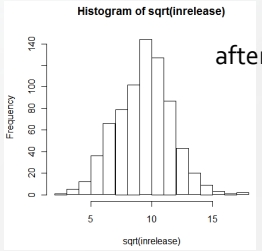
- Model indicates the strongest predictor is *Comedy* genre, followed by *nLogOpeningWeekendGross*
  - 1% increase in OpeningWeekendGross is associated with a 0.9% increase in LogTotalGross
- Other good predictors:
  - Model also predicts movies released around the *spring* and *summer* seasons (interacted with *nSqrtInRelease*) are likely to gross more than those released in fall or winter
    - Consistent with high movie demand during spring and summer breaks
  - Custom indicator, *Novbkst__c*, was close to being significant (p-value = 0.0409), suggests that the model predicts movies based on novels, books or stories are likely to gross more

Data Analysis (In Release)

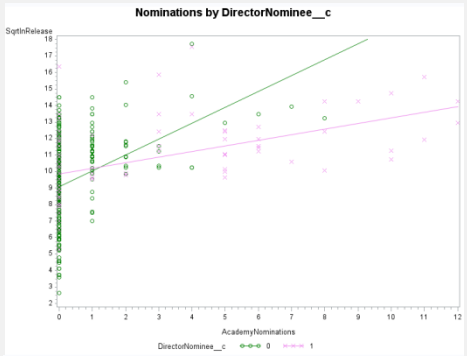
Step 1 - Exploratory

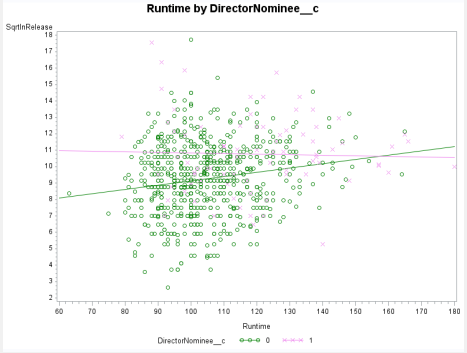
- Ordinary Least Squares
  - No weight was needed, homoscedasticity satisfied
  - Response variable was **square root** transformed to satisfy normality





- Transformed predictors used in prediction models
  - Log transformations:
    - TotalGross
    - OpeningWeekendGross
  - Square root transformations
    - WidestRelease
- Interaction effects used in prediction models
    - DirectorNominee__c*AcademyNominations
    - DirectorNominee__c*OpeningWeekendRank
    - DirectorNominee__c*CriticRating
    - DirectorNominee__c*Runtime
    - Season__c*LogTotalGross
    - Season__c*LogOpeningWeekendGross





Data Analysis (In Release)

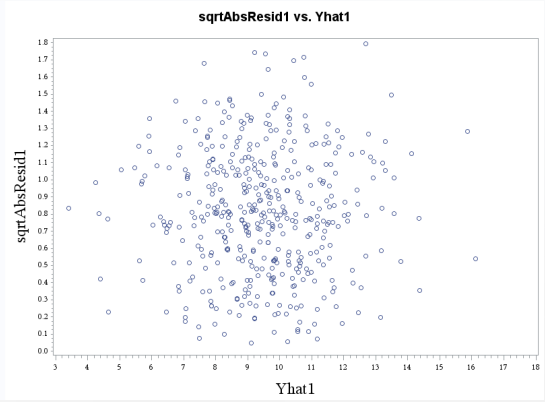
Step 2 – Kitchen Sink Model

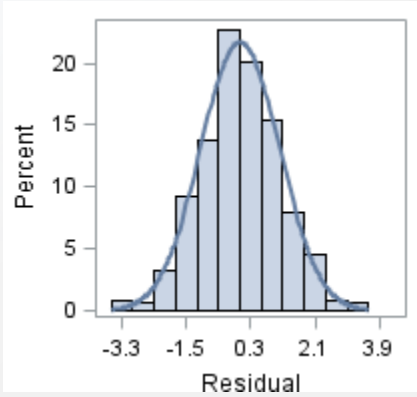
```
ods graphics on;
Proc GLM data = Train1
PLOTS=DIAGNOSTICS;
class Rating__c Season__c Year_Char ;

title1 ls=1.5 "Kitchen Sink Model on Train1";
model SqrtInRelease =
Rating__c Season__c Year_Char
AcademyNominations AcademyNominee__c AcademyAwards AcademyAwardWinner__c
MajorStudio__c Runtime RuntimeSq
LogProductionBudget SqrtWidestRelease CriticRating AudienceRating
LogTotalGross LogOpeningWeekendGross OpeningWeekendRank
Novbkst__c Trueins__c DisFlood__c DisSandy__c DisFire__c
/*  interaction terms */
MajorStudio__c*AcademyAwards MajorStudio__c*AcademyNominations MajorStudio__c*AudienceRating
MajorStudio__c*Runtime MajorStudio__c*RuntimeSq MajorStudio__c*LogProductionBudget
Season__c*Rating__c Season__c*LogProductionBudget
Season__c*CriticRating Season__c*AudienceRating Season__c*Runtime Season__c*RuntimeSq
Season__c*LogTotalGross Season__c*LogOpeningWeekendGross
AcademyAwards*DirectorNominee__c AcademyNominations*DirectorNominee__c
AudienceRating*DirectorNominee__c
CriticRating*DirectorNominee__c Runtime*DirectorNominee__c RuntimeSq*DirectorNominee__c
LogProductionBudget*DirectorNominee__c
LogOpeningWeekendGross*DirectorNominee__c OpeningWeekendRank*DirectorNominee__c
OpeningWeekendTheaters*DirectorNominee__c
/* movie indicators */
I_Dallas__c I_Ramona__c I_Winter__c I_MidParis__c I_Winnie__c
I_XmasCand__c I_Airbend__c I_Getlow__c I_Hugo__c I_Anna__c I_August__c
/ss3 solution tolerance;

output out=Train1New1 predicted=yhat1 residual=resid1 stdr=eresid1 cookd=cooksD1 rstudent=rstud1;
run; quit;
ods graphics off;
```

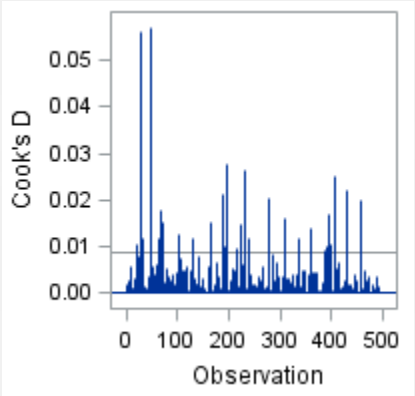
Data Analysis (In Release)

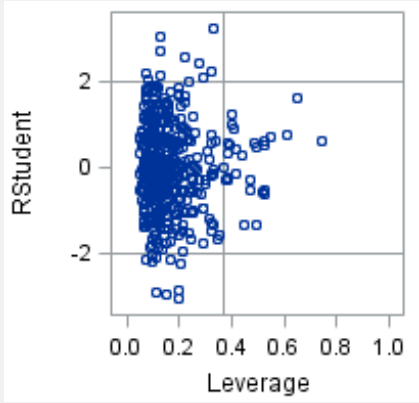
Step 2 – Kitchen Sink Model

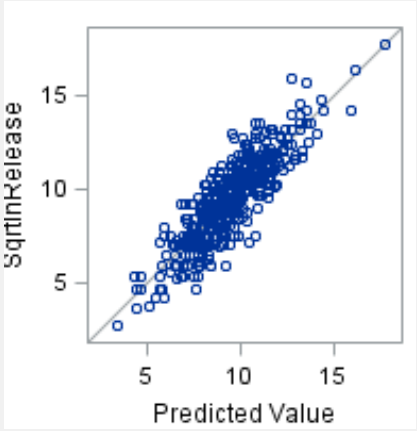
- Diagnostic plots









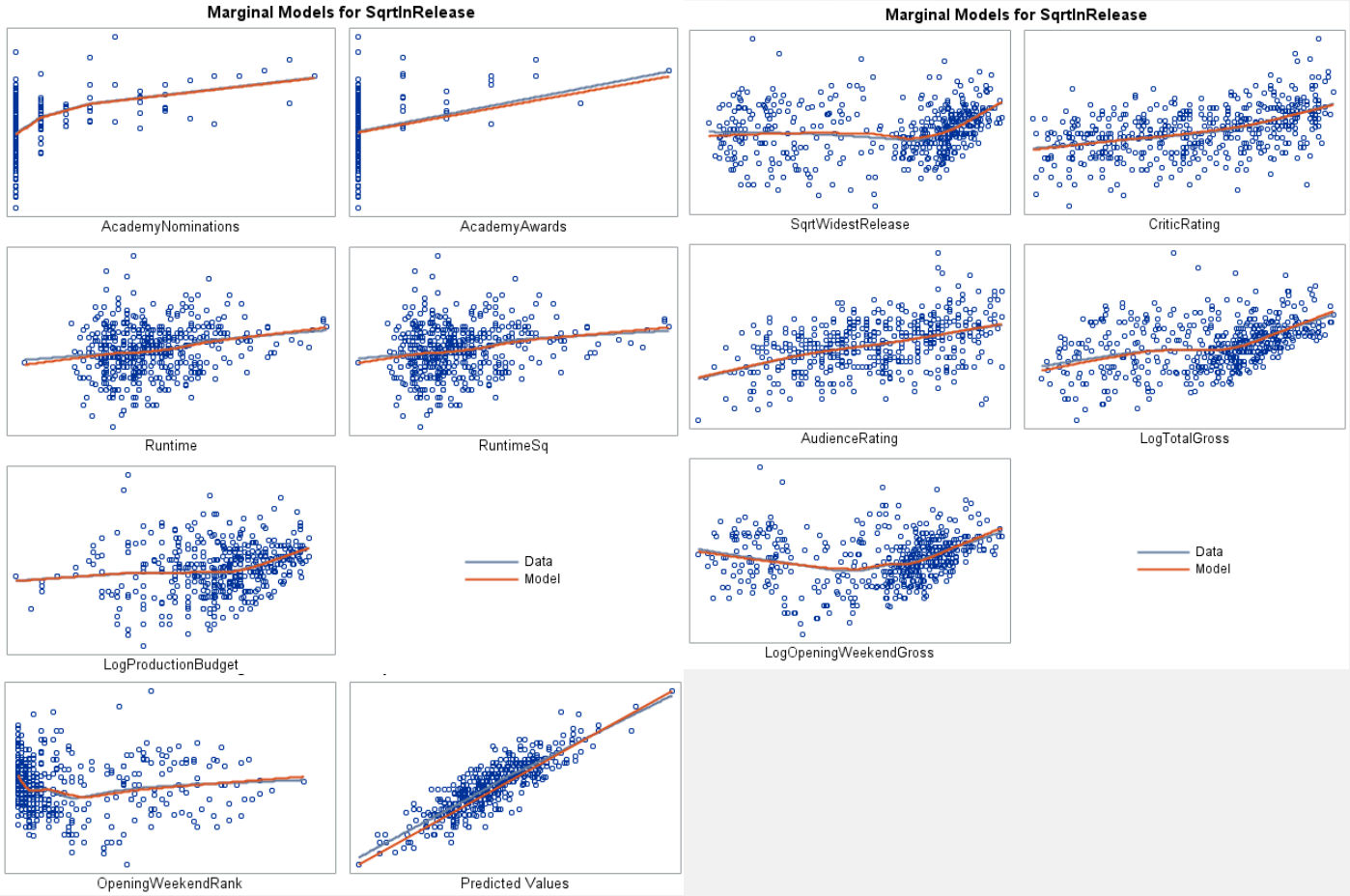


- ✓ Normality of error distribution
- ✓ Homoscedasticity
- ✓ Linear relationship between response and each predictor (holding others fixed)
- ✓ Independence of the errors

Data Analysis (In Release)

Step 2 – Kitchen Sink Model

- Marginal Model Plots
  - ✓ Blue lines align well with the red lines → model is a good fit



Data Analysis (In Release)

Step 3 – Variable Selection

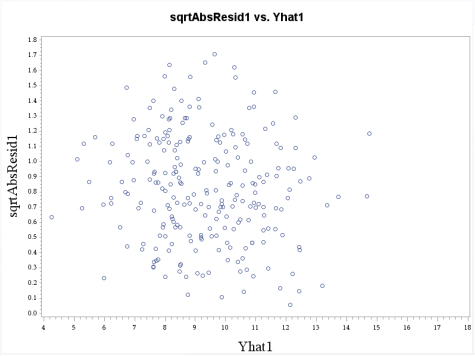
- SAS - **PROC GLMSelect** (select=BIC)
- Stepwise and Forward Selection outputted the same variable sets
- Backward selection – included more variables, some overlap

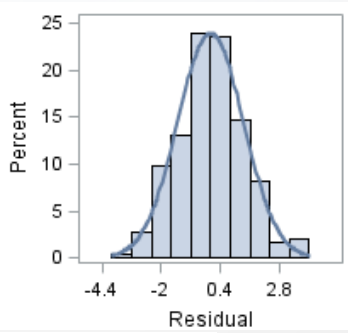
Stepwise Selection	Forward Selection	Backward Selection
Rating__c	Rating__c	Rating__c
		Season__c
		Year_Char
AcademyAwards	AcademyAwards	
SqrtWidestRelease	SqrtWidestRelease	SqrtWidestRelease
CriticRating	CriticRating	CriticRating
LogTotalGross	LogTotalGross	LogTotalGross
LogOpeningWeekendGross	LogOpeningWeekendGross	
OpeningWeekendRank	OpeningWeekendRank	OpeningWeekendRank
OpeningWeekendRank*DirectorNominee__c	OpeningWeekendRank*DirectorNominee__c	
		DisSandy__c
		DisFire__c
		Season__c*CriticRating
		Season__c*AudienceRating
		Season__c*LogTotalGross
		Season__c*LogOpeningWeekendGross
		LogProductionBudget*DirectorNominee__c
		LogOpeningWeekendGross*DirectorNominee__c

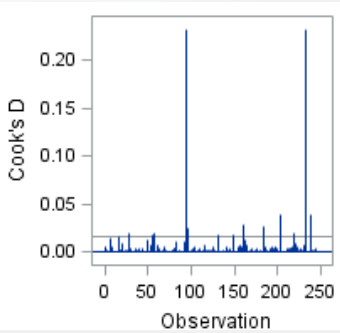
Data Analysis (In Release)

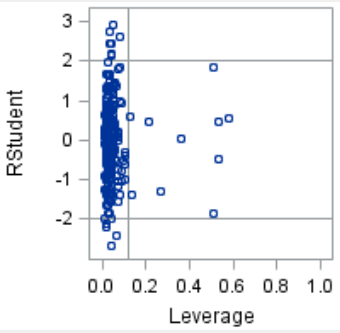
Step 4 – Model Verification

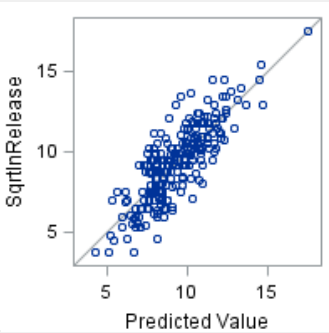
- Stepwise and Forward Selection: Reduced model using test data

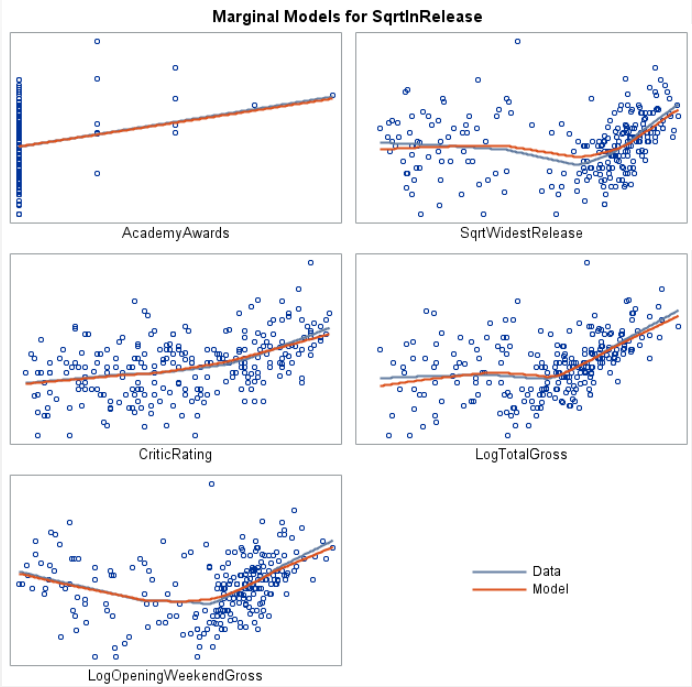












- ✓ Normality of error distribution
- ✓ Homoscedasticity
- ✓ Linear relationship between response and each predictor (holding others fixed)
- ✓ Independence of the errors
- ✓ Cook's D values small
- ✓ Marginal model plots indicate satisfactory fit, minor separation in predicted values and academy awards plot

Data Analysis (In Release)

Step 4 – Model Verification

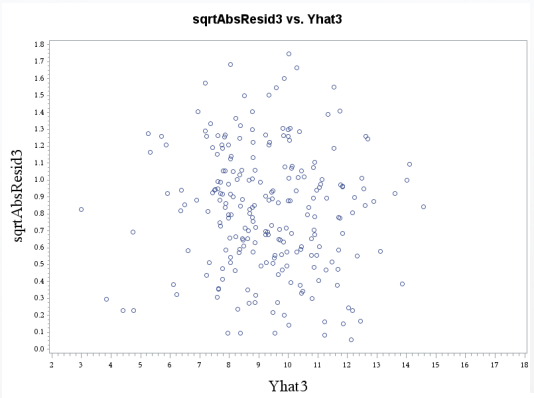
- Stepwise and Forward Selection: P-values and VIF on test data set
  - 5 of the 14 parameter estimates remain significant (p-value<=0.025)
  - VIFs acceptable, majority are less than 10

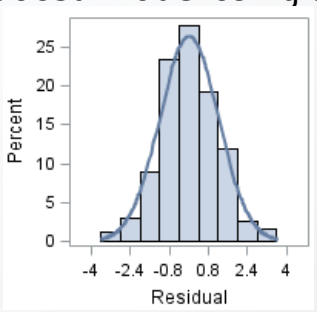
Parameter	Estimate	Pr > |t|	VIF
Intercept	-15.02	<.0001	0.00
Rating__c G	1.35	0.184	1.00
Rating__c Other	2.37	0.0288	1.00
Rating__c PG	1.52	<.0001	1.00
Rating__c PG-13	0.29	0.1681	1.15
AcademyAwards	0.28	0.1967	1.02
SqrtWidestRelease	-0.07	<.0001	1.18
CriticRating	0.01	0.0557	1.09
LogTotalGross	2.56	<.0001	8.73
LogOpeningWeekendGro	-1.09	<.0001	15.35
OpeningWeekendRank	0.02	0.3715	12.87
OpeningWe*DirectorNo	0.03	0.1422	1.07
I_MidParis__c	4.31	0.0029	1.07
I_Getlow__c	0.85	0.559	1.11
I_Anna__c	-2.65	0.0704	1.12

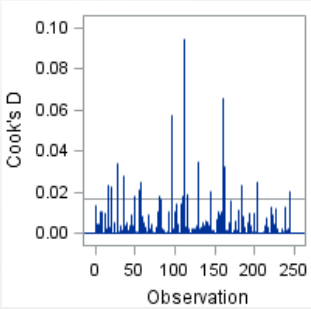
Data Analysis (In Release)

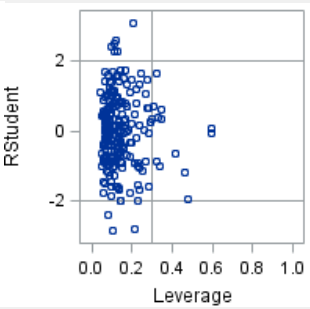
Step 4 – Model Verification

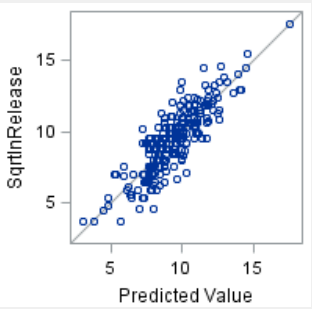
- Backward Selection: Reduced model using test data

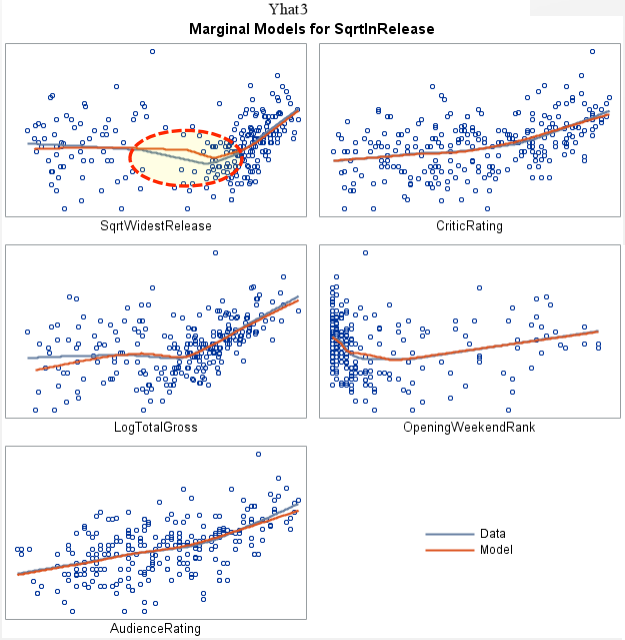




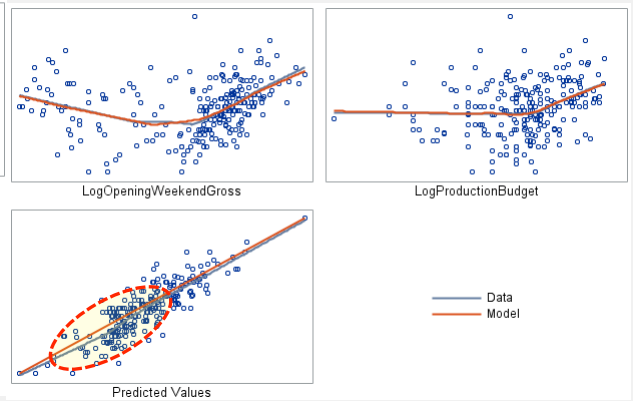








- ✓ Normality of error distribution
- ✓ Homoscedasticity
- ✓ Linear relationship between response and each predictor (holding others fixed)
- ✓ Independence of the errors
- ✓ Cook's D values small
- ✓ Some separation in marginal model plots, most plots fit well



Data Analysis (In Release)

Step 4 – Model Verification

- Backward Selection: P-values and VIF on test data set
  - 9 of the 34 parameter estimates remain significant (p-value<=0.025)
  - 7 of the 34 have high VIFs → makes me want to prefer selection model instead

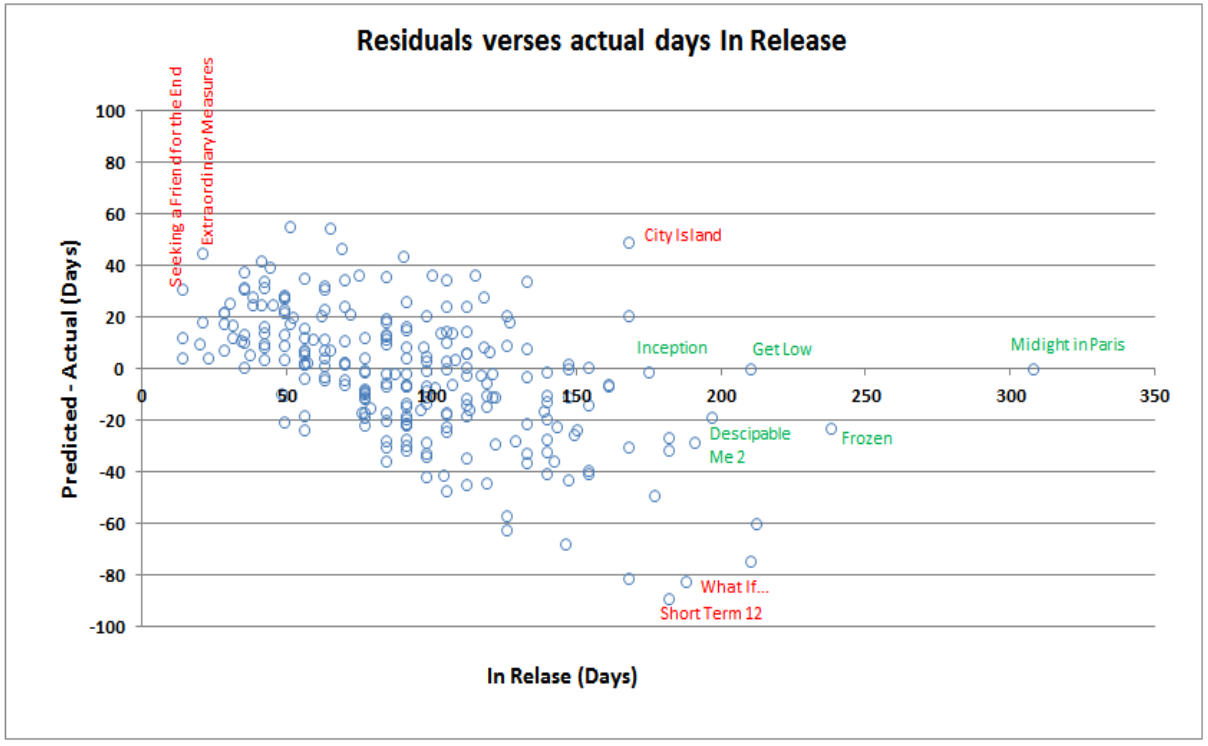
Parameter	Estimate	Pr > |t|	VIF
Intercept	-7.56	0.0562	0.00
Rating__c G	1.34	0.2035	1.00
Rating__c PG	1.30	<.0001	1.00
Rating__c PG-13	0.22	0.3171	1.14
Season__c Fall	-8.95	0.0126	1.01
Season__c Spring	-4.74	0.1606	1.13
Season__c Summer	-9.46	0.0041	1.53
Year_Char 2010	-0.37	0.1461	1.02
Year_Char 2011	-0.13	0.64	1.13
Year_Char 2012	-0.37	0.1477	1.44
SqrtWidestRelease	-0.06	0.0024	1.17
CriticRating	-0.01	0.2686	1.07
LogTotalGross	2.59	<.0001	7.66
OpeningWeekendRank	0.01	0.6464	5.20
DisSandy__c	0.52	0.4777	1.21
DisFire__c	0.09	0.8792	1.21
CriticRati*Season__c Fall	0.03	0.0839	5.92
CriticRati*Season__c Spring	0.02	0.1292	7.18
CriticRati*Season__c Summer	0.01	0.5166	9.70

Parameter	Estimate	Pr > |t|	VIF
AudienceRa*Season__c Fall	0.00	0.9212	25.04
AudienceRa*Season__c Spring	-0.01	0.7039	25.51
AudienceRa*Season__c Summer	0.01	0.3491	21.51
AudienceRa*Season__c Winter	0.01	0.6193	17.42
LogTotalGr*Season__c Fall	-0.14	0.8074	181.45
LogTotalGr*Season__c Spring	-0.22	0.6729	127.49
LogTotalGr*Season__c Summer	0.70	0.1909	227.13
LogOpening*Season__c Fall	-0.90	0.0283	697.36
LogOpening*Season__c Spring	-1.01	0.0061	610.07
LogOpening*Season__c Summer	-1.78	<.0001	599.65
LogOpening*Season__c Winter	-1.58	<.0001	838.84
LogProduc*DirectorNo	-0.51	0.2356	1.27
LogOpenin*DirectorNo	0.15	0.1275	27.84
I_MidParis__c	3.82	0.0096	1.24
I_Getlow__c	-0.52	0.7461	1.50
I_Anna__c	-1.91	0.2518	1.61

Data Analysis (In Release)

Step 5 – Prediction performance

- Backward model performed better
  - Prefer results from the Stepwise/Forward models
  - High VIFs in backward selection model
- Graphical representation of performance (Stepwise and Forward)



Prediction Power	Stepwise	Forward	Backward
Within ±5%	16%	16%	18%
Within ±10%	30%	30%	31%
Within ±15%	42%	42%	48%
Within ±20%	53%	53%	55%
Within ±25%	63%	63%	66%
Adjusted R2	0.65	0.65	0.69

- Predicted some long in release films
  - Inception
  - Despicable Me 2
  - Frozen
  - Midnight in Paris (had indicator)
  - Get Low (had indicator)
- Over-predicted
  - Seeking a Friend for the End
  - Extraordinary Measures
  - City Island
- Under-predicted
  - What If…
  - Short Term 12

Data Analysis (In Release)

Step 6 – Final Model Interpretation

- From stepwise selection, the model's strongest predictor for days in release is *LogTotalGross*
  - Intuitively this makes sense, major studios may have an incentive to want to keep high grossing films available longer in theaters.
- Model also predicts films rated-PG increases days in release
  - Suggests films targeted to children or a younger audience are held in theaters longer; perhaps due to many parents co-managing their children's activities outside movie time, so theaters keep the opportunity longer to watch such movies.
  - Wider audience for PG movies since they are less restrictive in content, so its to the studio's incentive to make them available in box office longer

Conclusion

- Overall the Total Gross model (M1) performed better predictions than the In Release model (M2)
- Stepwise and forward selection methods resulted in more simple models than backward selection
  - M1: Stepwise and forward performed slightly better than backward
  - M2: Backward performed better prediction, but stepwise and forward didn't contain parameters with very high VIFs.
- Majority of my custom Indicator variables initially thought to influence total gross (M1) actually did not result in p-values< 0.05 except for *Novbkst__c*
  - Suggests that movies based on novels, books or stories are likely to bring in more revenue to the box office.
- M1: strongest predictor is *Comedy* genre, followed by *nLogOpeningWeekendGross*
- M2: Strongest predictor was *LogTotalGross*
- Considerations to other models:
  - include indicator for A list actor/actresses
  - Some movies are seen multiple times by one individual (could affect total gross) – e.g. Titanic
  - Include indicator for sequal movies

End of presentation

- Questions?