Can you predict domestic total gross of a movie based on past movie performance?

Lucia
April 22, 2016

Data Sources, Features and Assumptions

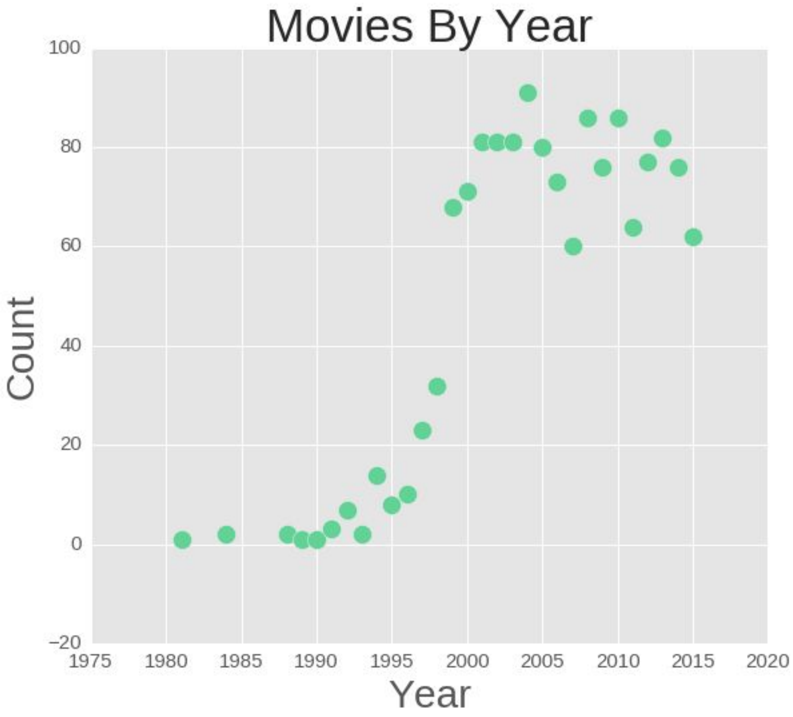
- BOX OFFICE MOJO:

  Domestic Total Gross, Production Budget, Release Date, Widest Release Theaters Count, Genre, Runtime, Rating, Distributor, Director, Producer(s), Actor(s), Writer(s), Composer(s), Cinematographer(s)

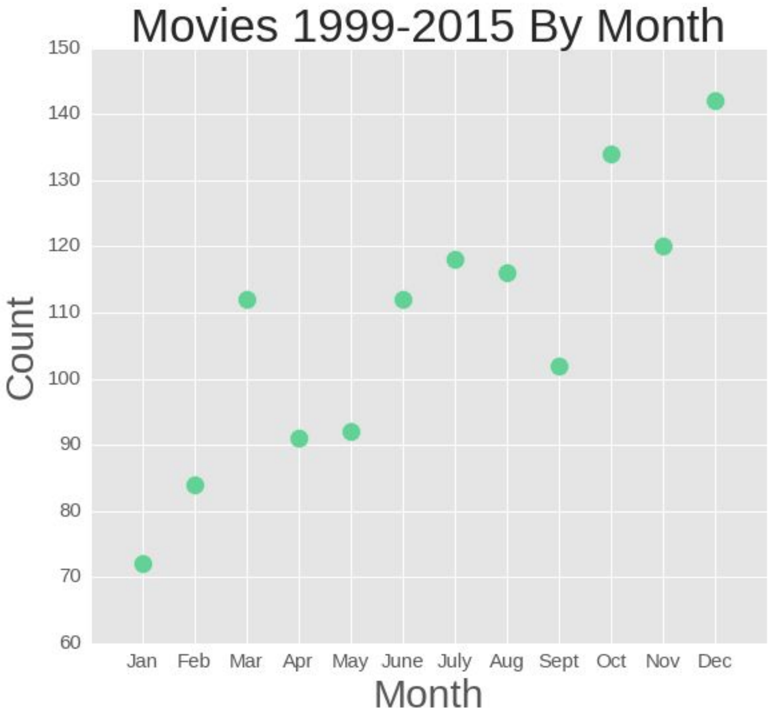
- METACRITIC:

  Movie Score

Box Office Mojo -- founded in 1999

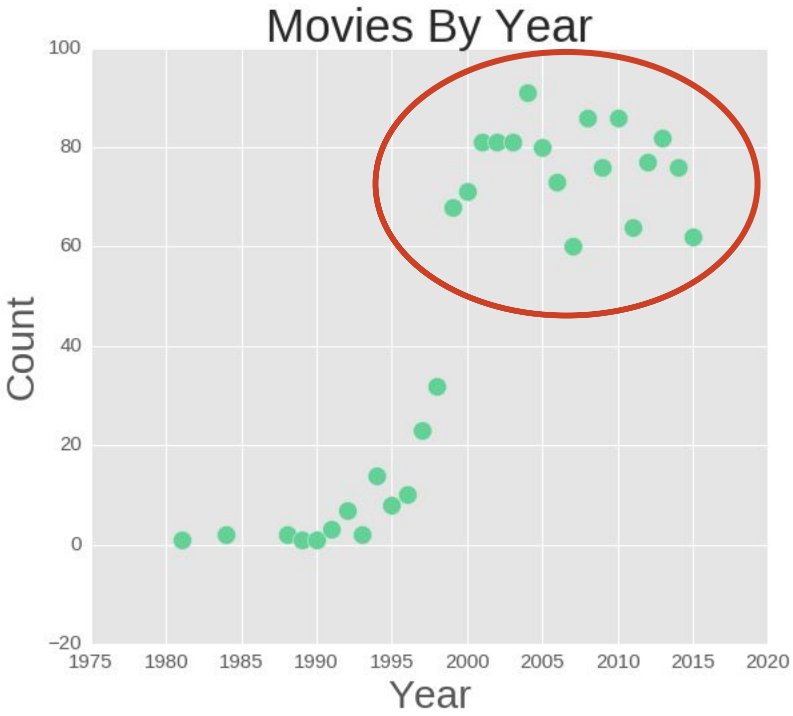


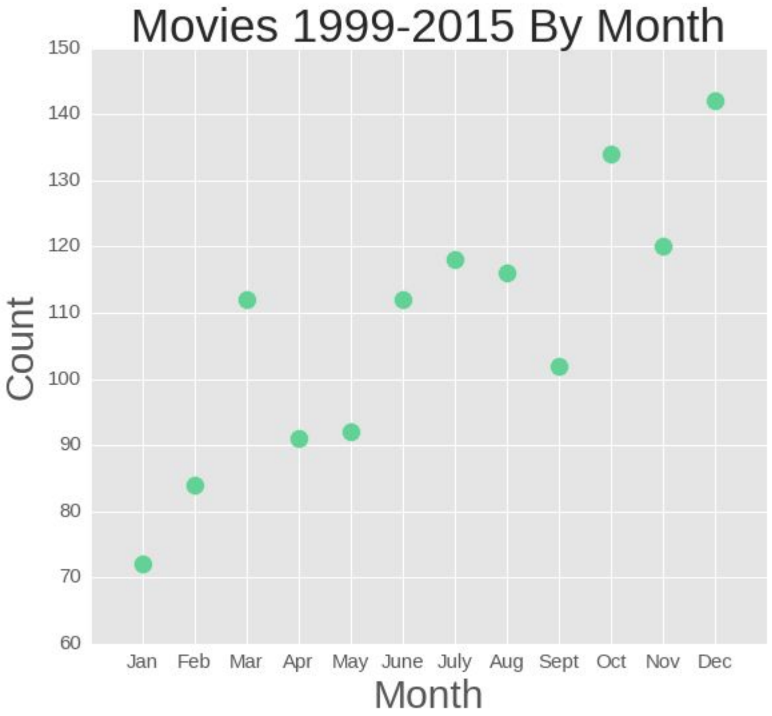
Movies By Year



Movies 1999-2015 By Month

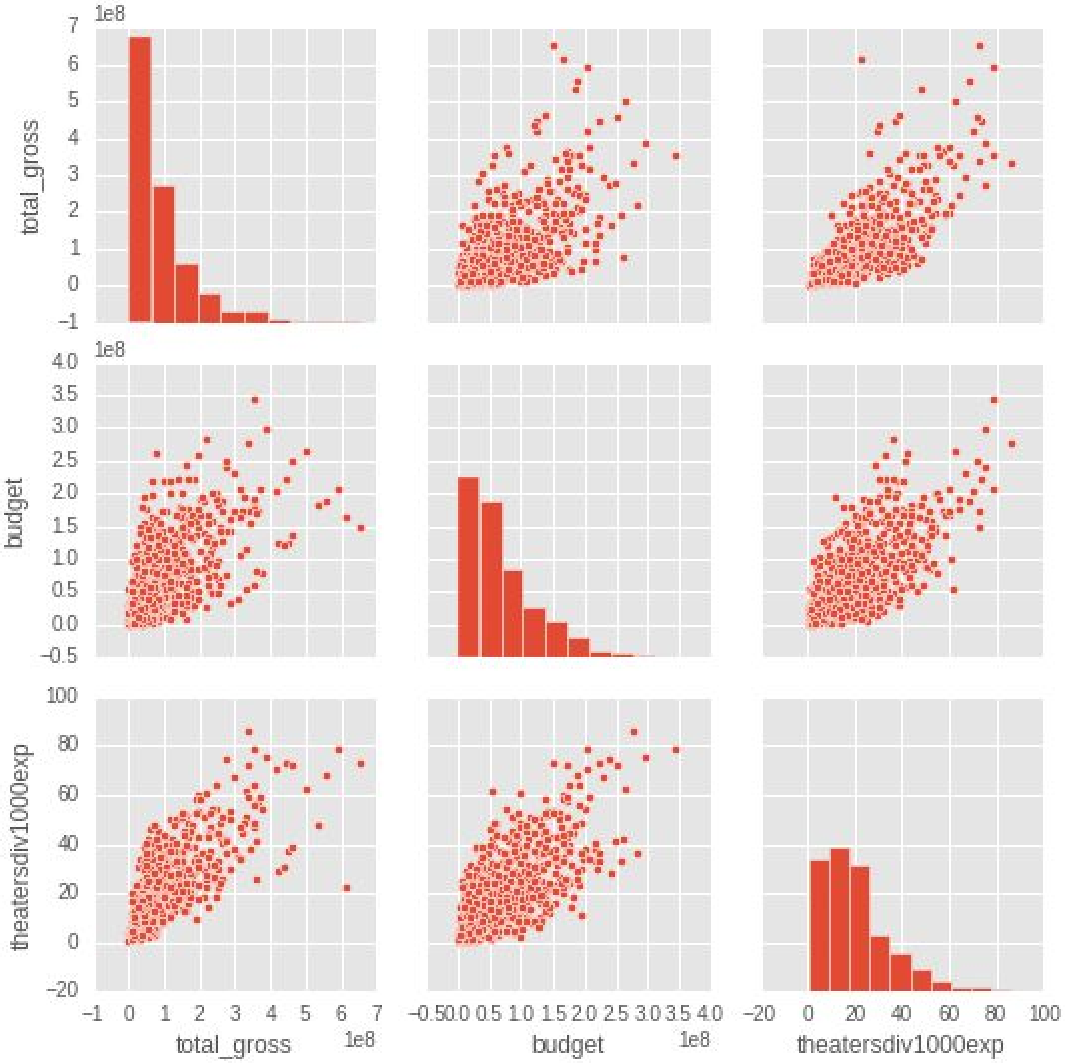
Box Office Mojo -- founded in 1999


Movies By Year


Movies 1999-2015 By Month

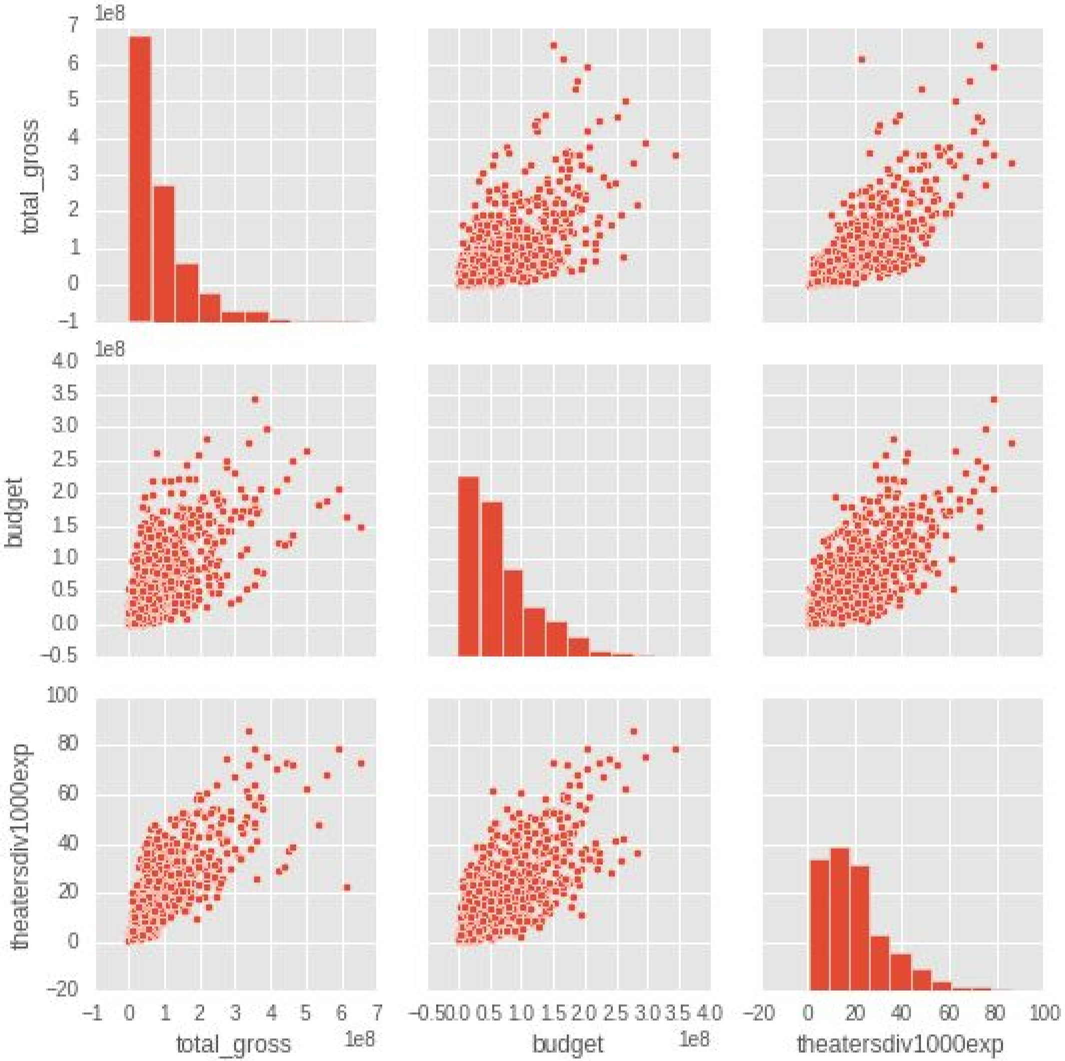
Correlations

	Domestic Total Gross	Production Budget
Production Budget	0.63	
Theaters	0.75	0.68



Correlations

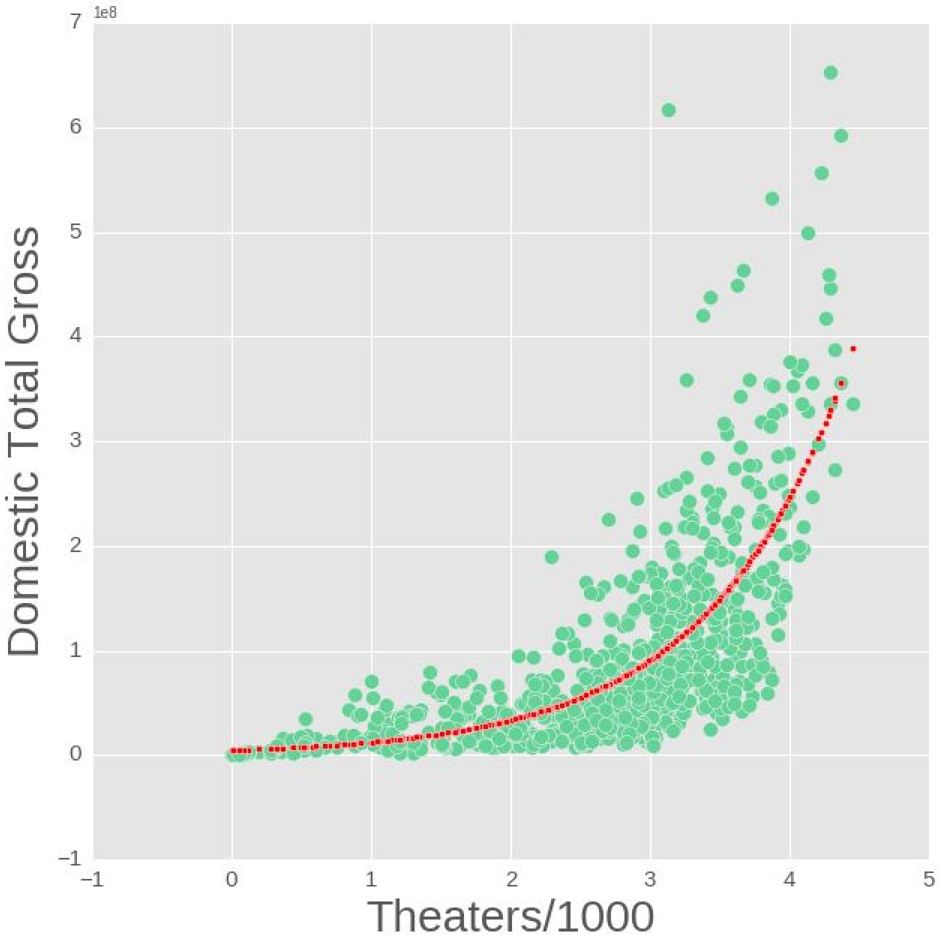
	Domestic Total Gross	Production Budget
Production Budget	0.63	
Theaters	0.75	0.68



Correlations

	Domestic Total Gross	Production Budget
Production Budget	0.63	
Theaters	0.75	0.68

$$\textbf{Total Gross} = 4{,}517{,}793\ \mathbf{e}^{\textbf{Theaters/1000}}$$



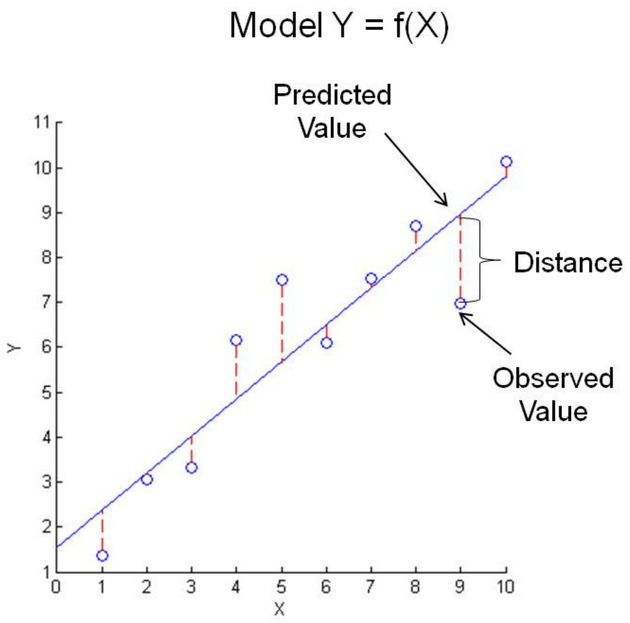
Ordinary Least Squares Models

$$\begin{tikzcd}
 & \mathit{minimize} \sum (y - y')^2 \arrow[dl] \arrow[dr] & \\
\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \arrow[dr] & & \mathbf{b} = \mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{yx}} \arrow[dl] \\
 & \mathbf{Y}' = \mathbf{bX} &
\end{tikzcd}$$

$$\mathbf{Y} = \mathbf{Y}' + \varepsilon$$

Covariance matrix between Xs and Y

Covariance matrix of Xs



Ordinary Least Squares Models

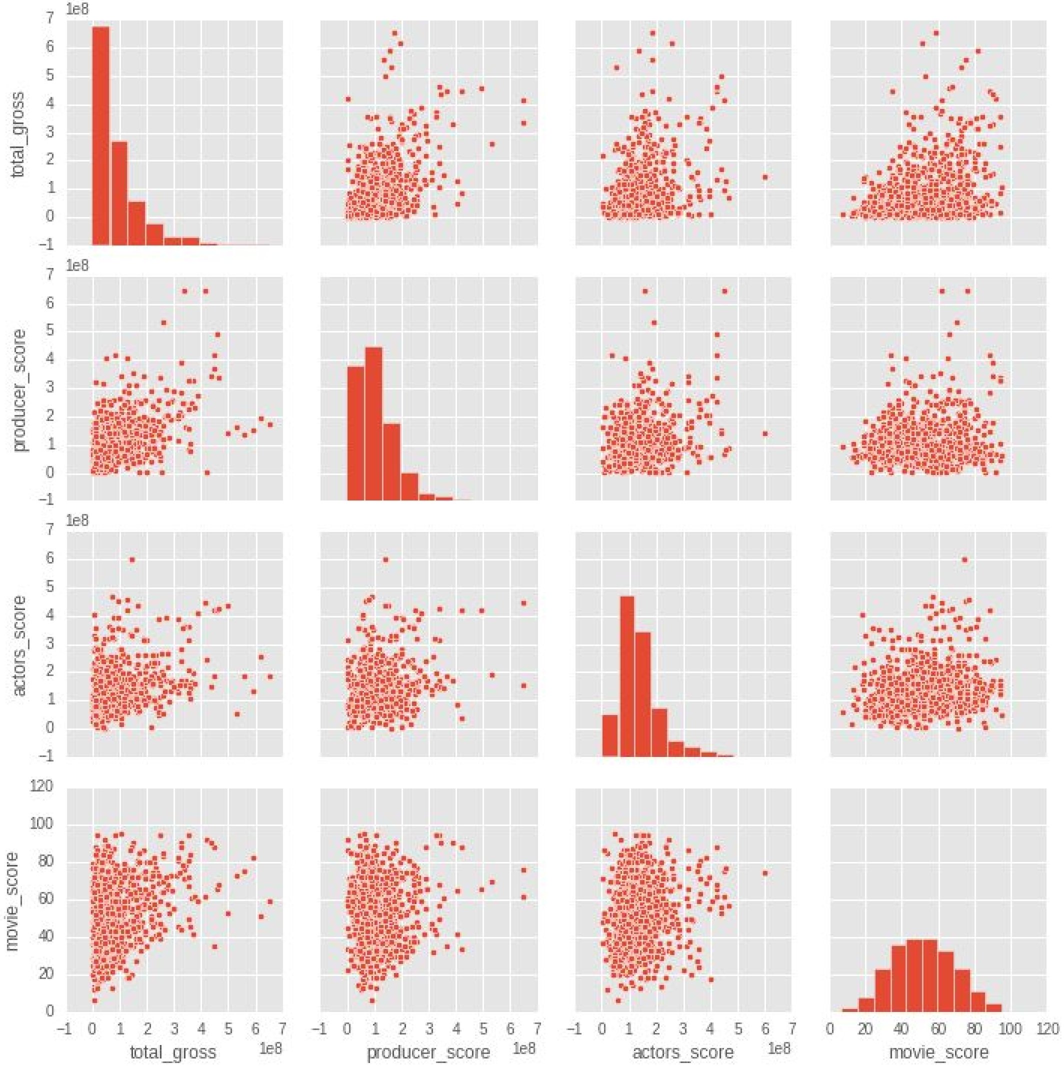
Domestic Total Gross ~	**Adjusted R^2: training set**	**Adjusted R^2: validation set**	**Durbin-Watson Autocorrelation Test**
Exp(Theaters/1000)	0.77	0.73	1.997

Ordinary Least Squares Models

Domestic Total Gross ~	**Adjusted R^2: training set**	**Adjusted R^2: validation set**	**Durbin-Watson Autocorrelation Test**
Exp(Theaters/1000)	0.77	0.73	1.997
Budget	0.68	0.70	2.004

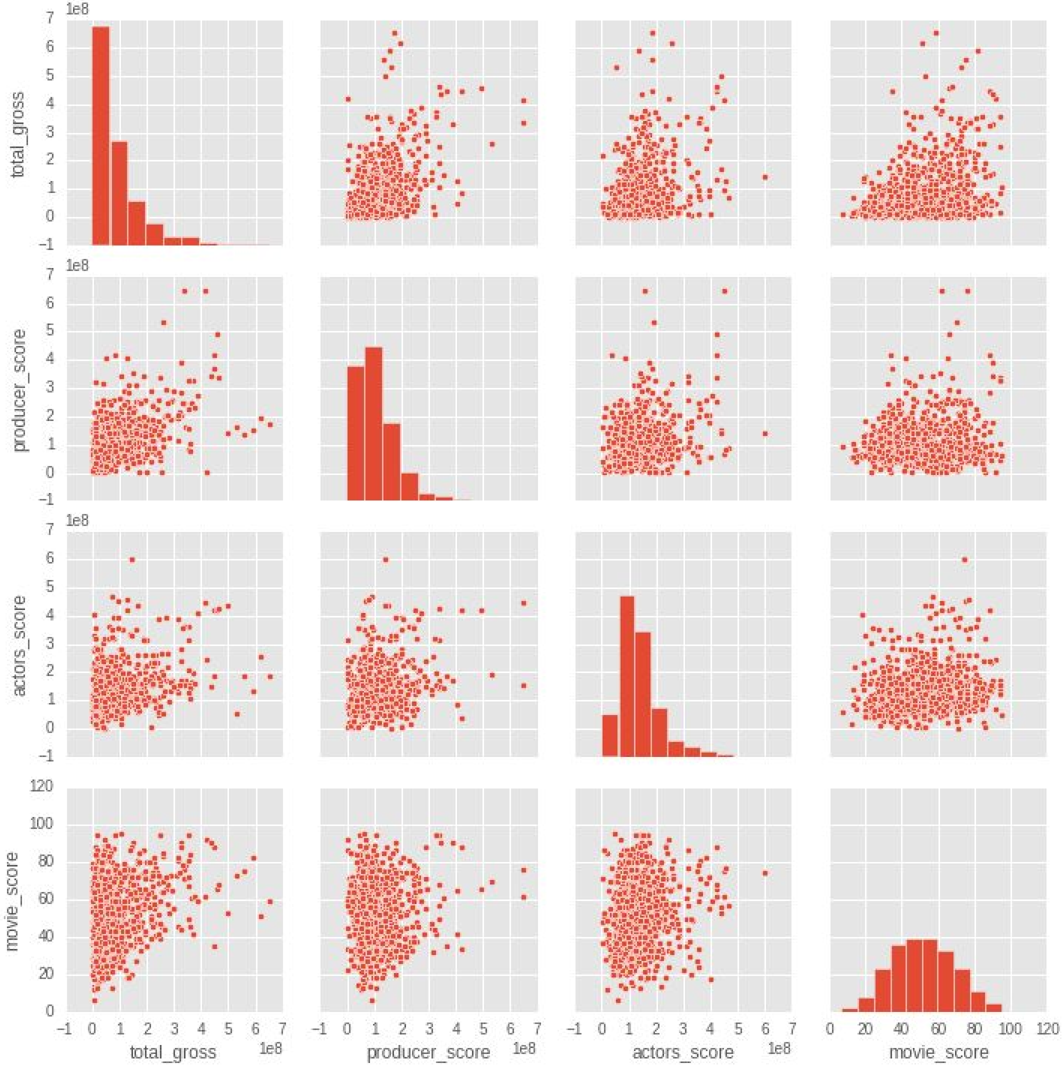
Correlations

	Domestic Total Gross	Producer Score
Producer Score	0.49	
Actors Score	0.34	0.27
Movie Score	0.31	0.13



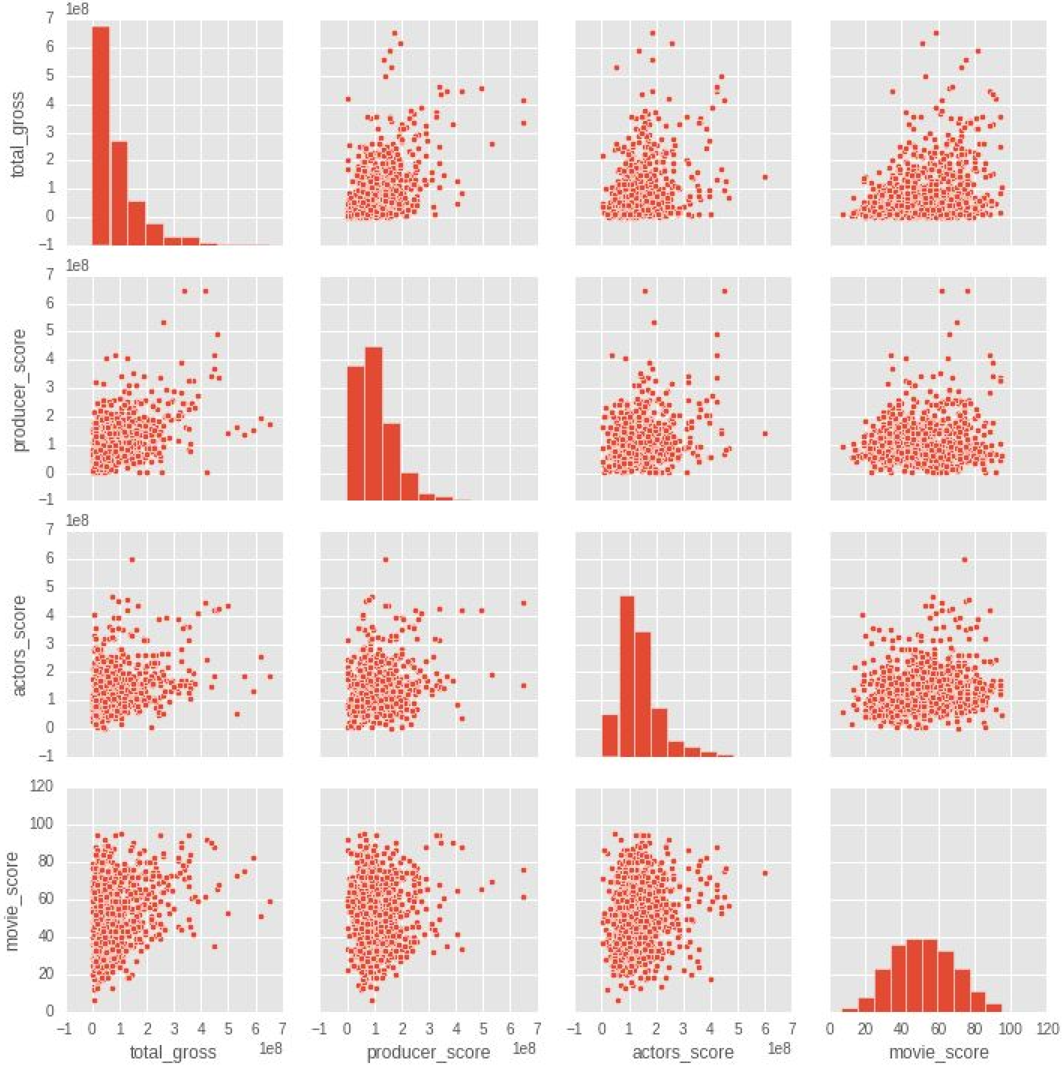
Correlations

	Domestic Total Gross	Producer Score
Producer Score	0.49	
Actors Score	0.34	0.27
Movie Score	0.31	0.13



Correlations

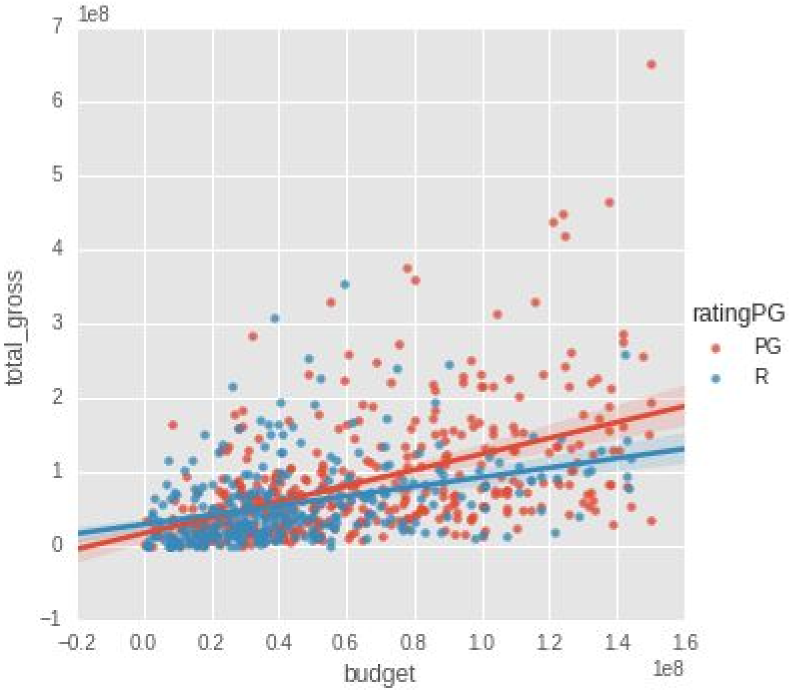
	Domestic Total Gross	Producer Score
Producer Score	**0.49**	
Actors Score	**0.34**	**0.27**
Movie Score	0.31	0.13

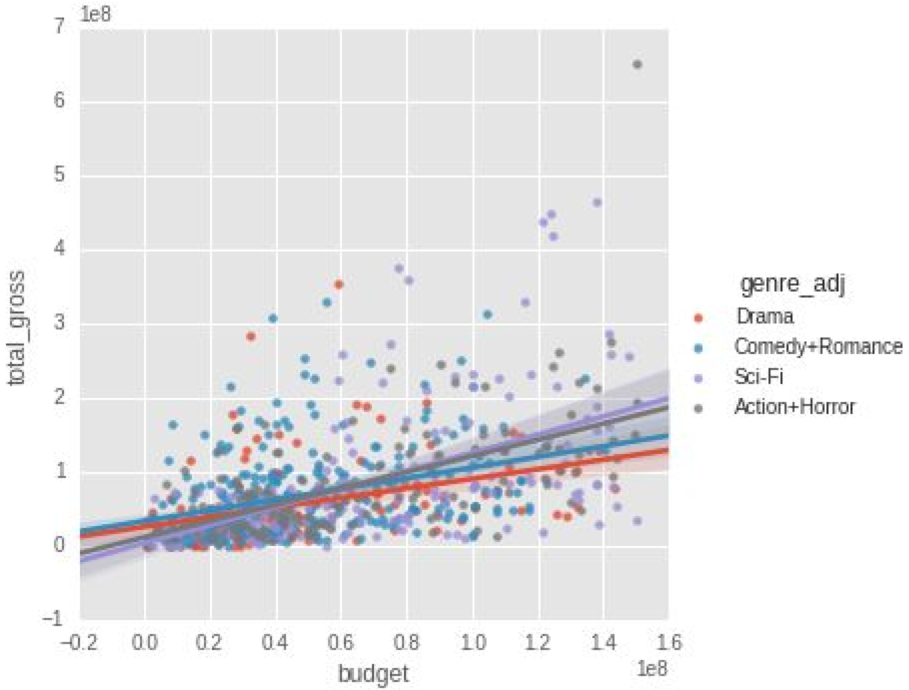


Ordinary Least Squares Models

Domestic Total Gross ~	Adjusted R^2: training set	Adjusted R^2: validation set	Durbin-Watson Autocorrelation Test
Exp(Theaters/1000)	0.77	0.73	1.997
Budget	0.68	0.70	2.004
Budget + Producers	0.71	0.72	2.036
Budget + Producers + Actors	0.71	0.73	2.017

Movie Rating, Genre ?





Ordinary Least Squares Models

Domestic Total Gross ~	Adjusted R^2: training set	Adjusted R^2: validation set	Durbin-Watson Autocorrelation Test
Exp(Theaters/1000)	0.77	0.73	1.997
Budget	0.68	0.70	2.004
Budget + Producers	0.71	0.72	2.036
Budget + Producers + Actors	0.71	0.73	2.017
Budget + Producers + Actors + Budget*Rating + Budget*Genre	0.72	0.73	1.998

Ordinary Least Squares Models

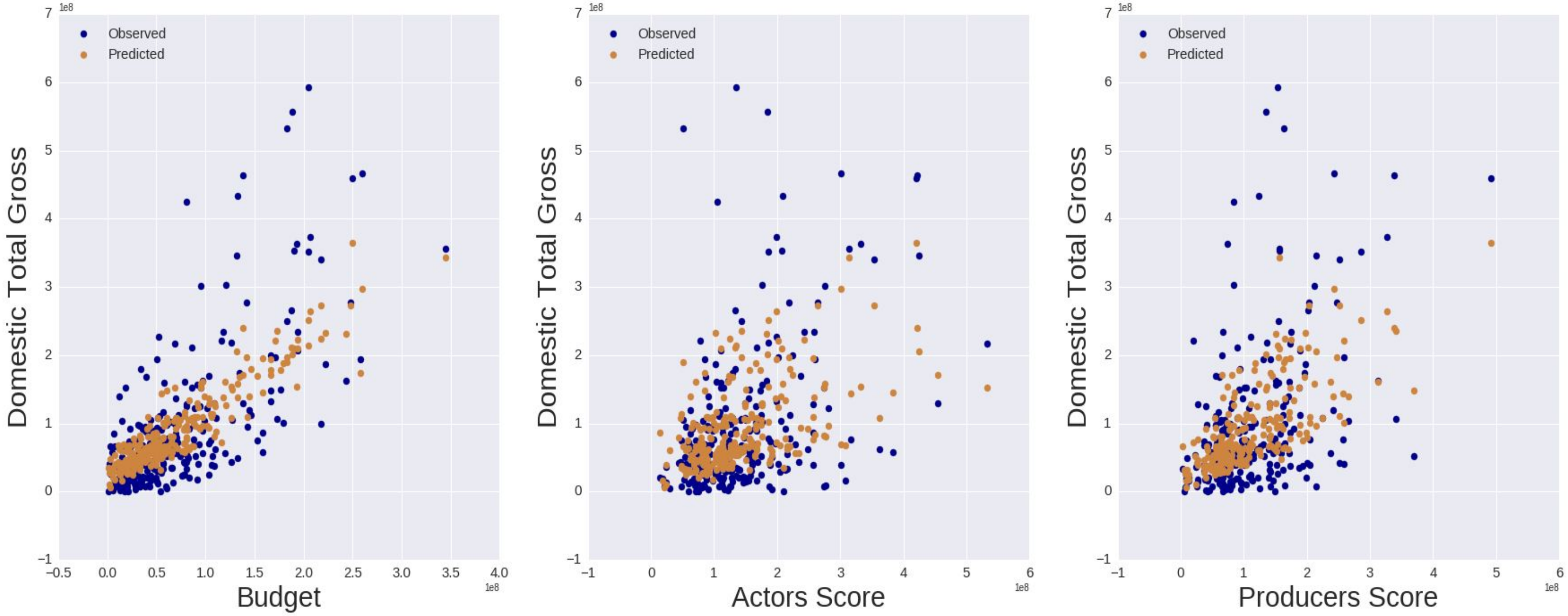
Domestic Total Gross ~	**Adjusted R^2: training set**	**Adjusted R^2: validation set**	**Durbin-Watson Autocorrelation Test**
Exp(Theaters/1000)	0.77	0.73	1.997
Budget	0.68	0.70	2.004
Budget + Producers	0.71	0.72	2.036
Budget + Producers + Actors	0.71	0.73	2.017
Budget + Producers + Actors + Budget*Rating + Budget*Genre	0.72	0.73	1.998
Movie Score + Runtime + Year + Exp(Theaters)	0.81	0.76	1.998

Ordinary Least Squares - Predictive Model

Domestic Total Gross ~	Adjusted R^2: training set	Adjusted R^2: validation set	Durbin-Watson Autocorrelation Test
Exp(Theaters/1000)	0.77	0.73	1.997
Budget	0.68	0.70	2.004
Budget + Producers	0.71	0.72	2.036
Budget + Producers + Actors	0.71	0.73	2.017
Budget + Producers + Actors + Budget*Rating + Budget*Genre	0.72	0.73	1.998
Movie Score + Runtime + Year + Exp(Theaters)	0.81	0.76	1.998

Predictions

$R^2 = 0.78$

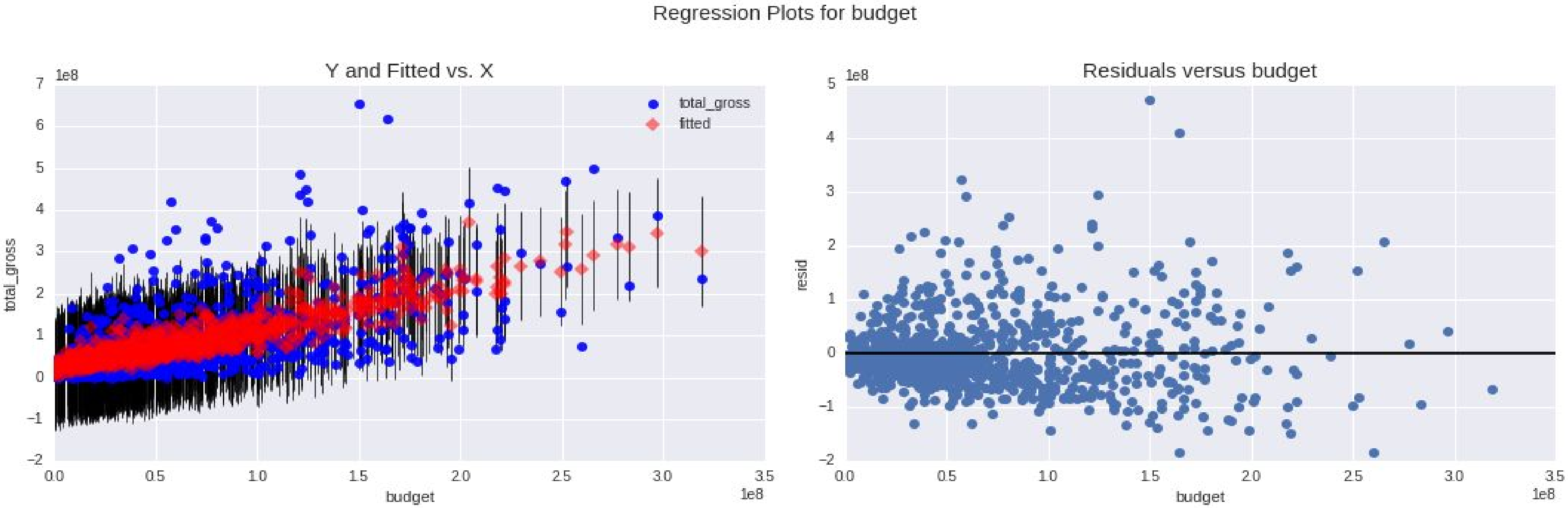


Model Coefficients

B_{budget} = 0.85 with 95%CI [0.76,0.94] “width” 0.18

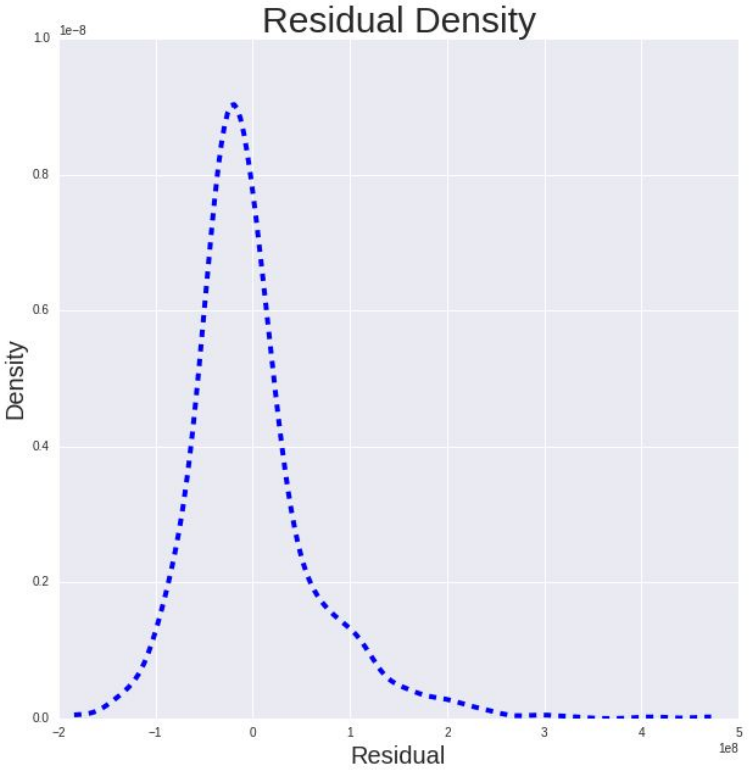
B_{actors} = 0.11 with 95%CI [0.06,0.16] “width” 0.10

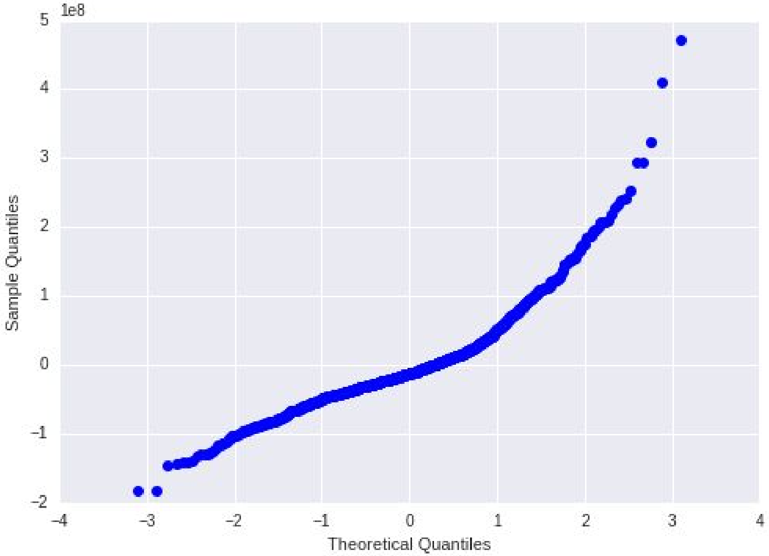
$B_{producer}$ = 0.22 with 95%CI [0.16,0.28] “width” 0.12



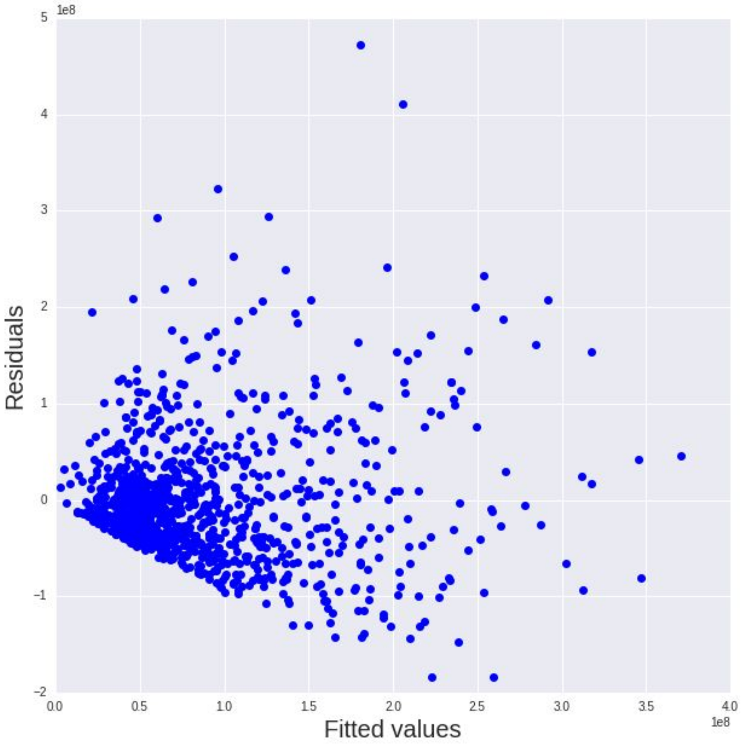
Movie Title	Domestic Total Gross	Budget	Release Date	Movie Score
Jurassic World	6.522706e+08	150000000.0	2015-06-12	59
Star Wars: Episode I - The Phantom Menace	6.164563e+08	164450000.0	1999-05-19	51
The Dark Knight	5.920133e+08	205350000.0	2008-07-18	82
Shrek 2	5.559451e+08	189000000.0	2004-05-19	75
Spider-Man	5.328924e+08	183480000.0	2002-05-03	73
Pirates of the Caribbean: Dead Man's Chest	4.995127e+08	265500000.0	2006-07-07	53

Assumptions: Normal distribution of errors





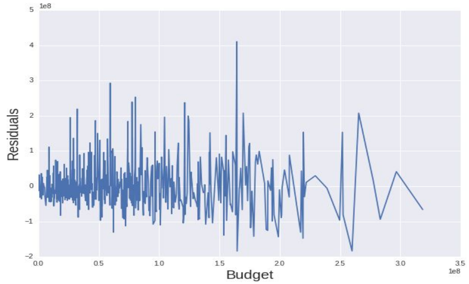
Assumptions: Homoskedasticity

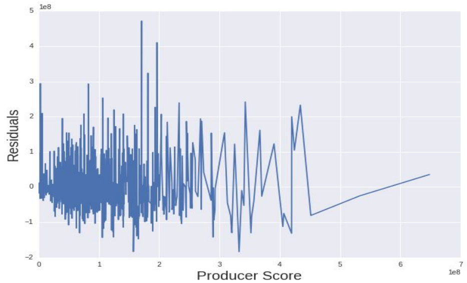


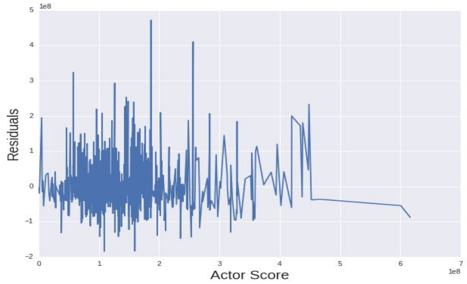
White Test:

OLS(Errors ~ X)

R^2=0



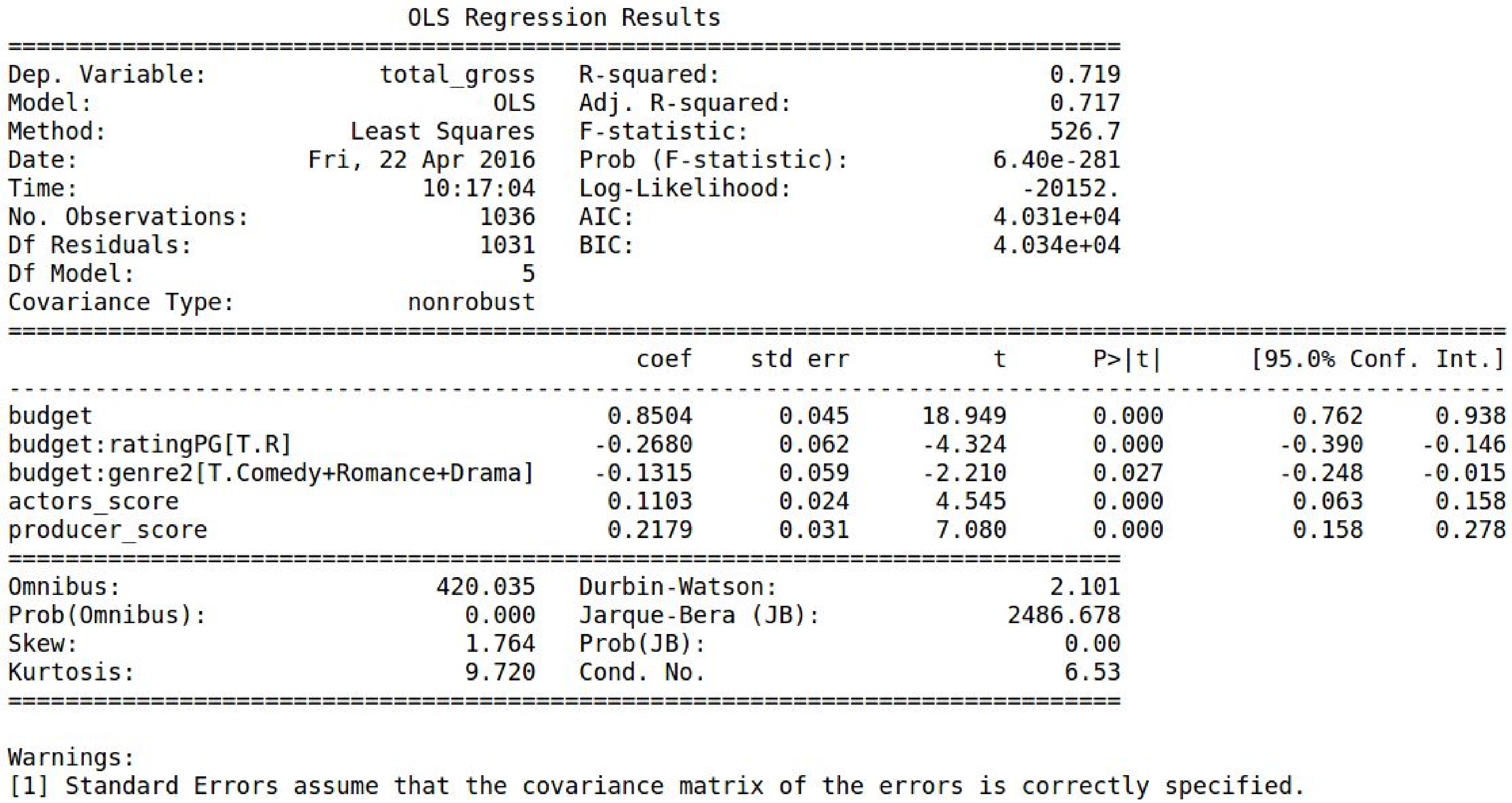




Model: BLUE

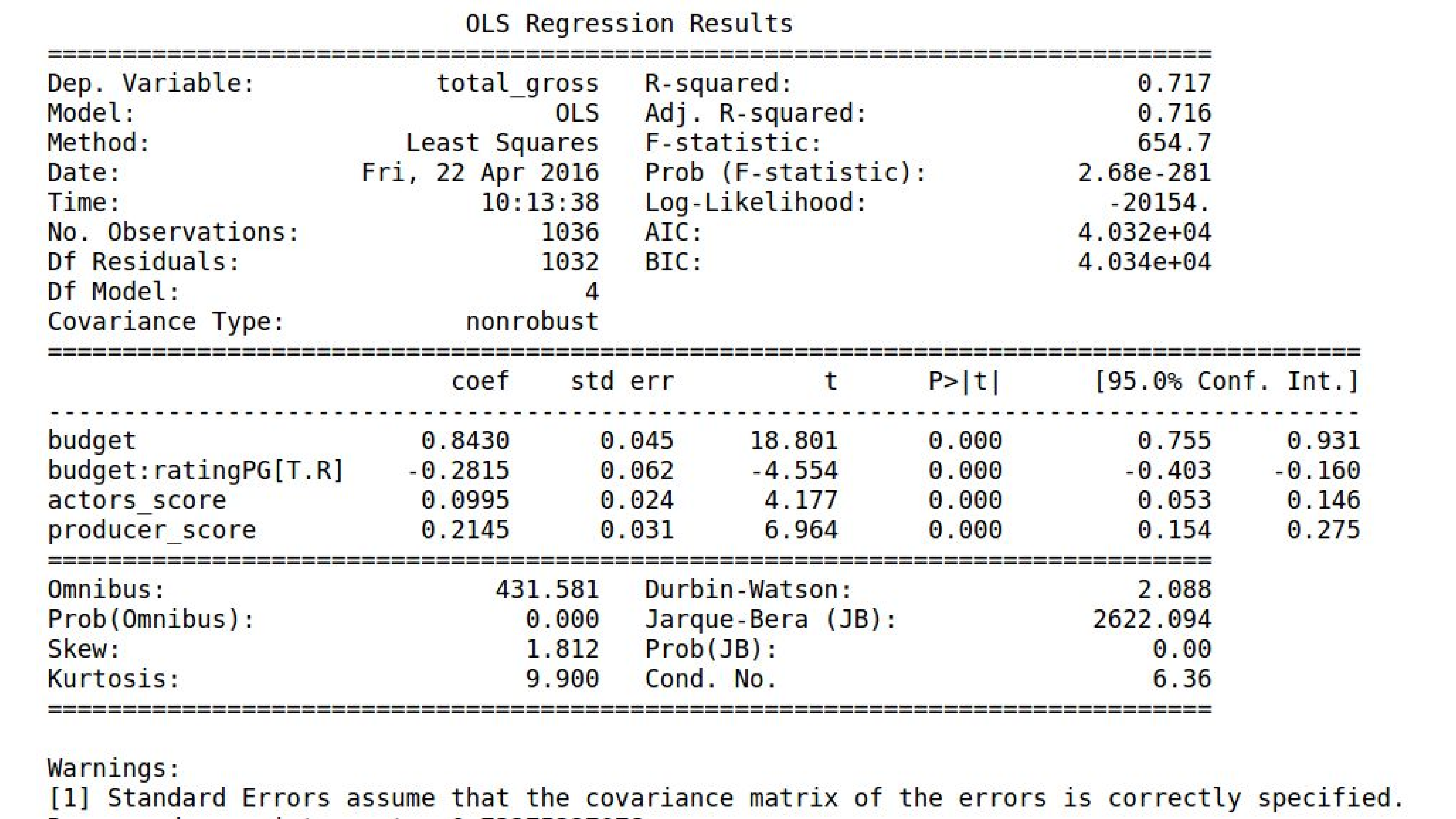
- Linear in parameters
- No exact multicollinearity
- Fixed covariates (X’s)
- Number of observations (1036) > Number of parameters (4)
- Mean residual error: nonzero, negative = systematically underestimating values (by -1,193,706)
- Homoskedasticity, Var(error$_i$) = const
- No autocorrelation, Cov(error$_i$, error$_j$) = 0, Durbin-Watson test close to 2

Thank you !



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            total_gross   R-squared:                       0.719
Model:                            OLS   Adj. R-squared:                  0.717
Method:                 Least Squares   F-statistic:                     526.7
Date:                Fri, 22 Apr 2016   Prob (F-statistic):          6.40e-281
Time:                        10:17:04   Log-Likelihood:                -20152.
No. Observations:                1036   AIC:                         4.031e+04
Df Residuals:                    1031   BIC:                         4.034e+04
Df Model:                           5
Covariance Type:            nonrobust
=========================================================================================================
                                           coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------------------------------
budget                                   0.8504      0.045     18.949      0.000         0.762     0.938
budget:ratingPG[T.R]                    -0.2680      0.062     -4.324      0.000        -0.390    -0.146
budget:genre2[T.Comedy+Romance+Drama]   -0.1315      0.059     -2.210      0.027        -0.248    -0.015
actors_score                             0.1103      0.024      4.545      0.000         0.063     0.158
producer_score                           0.2179      0.031      7.080      0.000         0.158     0.278
==============================================================================
Omnibus:                      420.035   Durbin-Watson:                   2.101
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2486.678
Skew:                           1.764   Prob(JB):                         0.00
Kurtosis:                       9.720   Cond. No.                         6.53
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            total_gross   R-squared:                       0.717
Model:                            OLS   Adj. R-squared:                  0.716
Method:                 Least Squares   F-statistic:                     654.7
Date:                Fri, 22 Apr 2016   Prob (F-statistic):         2.68e-281
Time:                        10:13:38   Log-Likelihood:                -20154.
No. Observations:                1036   AIC:                         4.032e+04
Df Residuals:                    1032   BIC:                         4.034e+04
Df Model:                           4
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------------
budget                   0.8430      0.045     18.801      0.000         0.755     0.931
budget:ratingPG[T.R]    -0.2815      0.062     -4.554      0.000        -0.403    -0.160
actors_score             0.0995      0.024      4.177      0.000         0.053     0.146
producer_score           0.2145      0.031      6.964      0.000         0.154     0.275
==============================================================================
Omnibus:                      431.581   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2622.094
Skew:                           1.812   Prob(JB):                         0.00
Kurtosis:                       9.900   Cond. No.                         6.36
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```