DOES GENRE AFFECT A MOVIE'S EARNING POTENTIAL (DOMESTIC GROSS)?

KATHERINE PULLY

JULY 15, 2016

MOVIELENS DATASET

1. Animation
2. Drama
3. Adventure
4. Children
5. Sci-Fi
6. Horror
7. Comedy
8. Romance
9. Drama
10. Thriller
11. Mystery
12. Fantasy
13. Documentary
14. Musical
15. Western
16. War
17. Crime
18. Film-Noir

FIRST PASS RESULTS

Dep. Variable:	dgross	**R-squared:**	0.012
Model:	OLS	**Adj. R-squared:**	0.007
Method:	Least Squares	**F-statistic:**	2.249
Date:	Fri, 15 Jul 2016	**Prob (F-statistic):**	0.0169
Time:	00:04:04	**Log-Likelihood:**	-32264.
No. Observations:	1638	**AIC:**	6.455e+04
Df Residuals:	1628	**BIC:**	6.460e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	8.878e+07	5.31e+06	16.717	0.000	7.84e+07 9.92e+07
thriller	-1.419e+06	6.6e+06	-0.215	0.830	-1.44e+07 1.15e+07
comedy	-8.559e+05	5.61e+06	-0.153	0.879	-1.19e+07 1.02e+07
drama	-6.953e+06	5.39e+06	-1.291	0.197	-1.75e+07 3.61e+06
documentary	1.496e+06	1.33e+07	0.113	0.910	-2.45e+07 2.75e+07
action	4.727e+06	6.55e+06	0.722	0.471	-8.12e+06 1.76e+07
animation	2.249e+07	1.42e+07	1.589	0.112	-5.26e+06 5.02e+07
horror	3.11e+07	9.95e+06	3.127	0.002	1.16e+07 5.06e+07
fantasy	-1.118e+07	1.88e+07	-0.595	0.552	-4.8e+07 2.57e+07
romance	-3.577e+06	6.17e+06	-0.579	0.562	-1.57e+07 8.53e+06

FIRST PASS RESULTS

Dep. Variable:	dgross	**R-squared:**	0.012
Model:	OLS	**Adj. R-squared:**	0.007
Method:	Least Squares	**F-statistic:**	2.249
Date:	Fri, 15 Jul 2016	**Prob (F-statistic):**	0.0169
Time:	00:04:04	**Log-Likelihood:**	-32264.
No. Observations:	1638	**AIC:**	6.455e+04
Df Residuals:	1628	**BIC:**	6.460e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	8.878e+07	5.31e+06	16.717	0.000	7.84e+07 9.92e+07
thriller	-1.419e+06	6.6e+06	-0.215	0.830	-1.44e+07 1.15e+07
comedy	-8.559e+05	5.61e+06	-0.153	0.879	-1.19e+07 1.02e+07
drama	-6.953e+06	5.39e+06	-1.291	0.197	-1.75e+07 3.61e+06
documentary	1.496e+06	1.33e+07	0.113	0.910	-2.45e+07 2.75e+07
action	4.727e+06	6.55e+06	0.722	0.471	-8.12e+06 1.76e+07
animation	2.249e+07	1.42e+07	1.589	0.112	-5.26e+06 5.02e+07
horror	3.11e+07	9.95e+06	3.127	0.002	1.16e+07 5.06e+07
fantasy	-1.118e+07	1.88e+07	-0.595	0.552	-4.8e+07 2.57e+07
romance	-3.577e+06	6.17e+06	-0.579	0.562	-1.57e+07 8.53e+06

SECOND ATTEMPT – LOG TRANSFORMATION

Dep. Variable:	log_gross	**R-squared:**	0.014
Model:	OLS	**Adj. R-squared:**	0.009
Method:	Least Squares	**F-statistic:**	2.596
Date:	Fri, 15 Jul 2016	**Prob (F-statistic):**	0.00566
Time:	00:11:45	**Log-Likelihood:**	-2578.9
No. Observations:	1638	**AIC:**	5178.
Df Residuals:	1628	**BIC:**	5232.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	17.7862	0.072	248.645	0.000	17.646 17.926
thriller	0.0282	0.089	0.318	0.751	-0.146 0.203
comedy	0.0290	0.076	0.383	0.702	-0.119 0.177
drama	-0.0512	0.073	-0.705	0.481	-0.194 0.091
documentary	0.1214	0.179	0.679	0.497	-0.229 0.472
action	0.1036	0.088	1.174	0.241	-0.069 0.277
animation	0.1550	0.191	0.813	0.416	-0.219 0.529
horror	0.4648	0.134	3.470	0.001	0.202 0.728
fantasy	-0.0239	0.253	-0.094	0.925	-0.520 0.472
romance	-0.1482	0.083	-1.783	0.075	-0.311 0.015

SECOND ATTEMPT – LOG TRANSFORMATION

Dep. Variable:	log_gross	**R-squared:**	0.014
Model:	OLS	**Adj. R-squared:**	0.009
Method:	Least Squares	**F-statistic:**	2.596
Date:	Fri, 15 Jul 2016	**Prob (F-statistic):**	0.00566
Time:	00:11:45	**Log-Likelihood:**	-2578.9
No. Observations:	1638	**AIC:**	5178.
Df Residuals:	1628	**BIC:**	5232.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	17.7862	0.072	248.645	0.000	17.646 17.926
thriller	0.0282	0.089	0.318	0.751	-0.146 0.203
comedy	0.0290	0.076	0.383	0.702	-0.119 0.177
drama	-0.0512	0.073	-0.705	0.481	-0.194 0.091
documentary	0.1214	0.179	0.679	0.497	-0.229 0.472
action	0.1036	0.088	1.174	0.241	-0.069 0.277
animation	0.1550	0.191	0.813	0.416	-0.219 0.529
horror	0.4648	0.134	3.470	0.001	0.202 0.728
fantasy	-0.0239	0.253	-0.094	0.925	-0.520 0.472
romance	-0.1482	0.083	-1.783	0.075	-0.311 0.015

LOG TRANSFORMATION + REGULARIZATION

L1

```
model_lasso = linear_model.Lasso()
model_lasso.fit(X, y)
model_lasso.coef_
```

```
array([ 0.,  0., -0., -0.,  0.,  0.,  0.,  0., -0., -0.])
```

```
model_lasso.score(X, y)
```

```
0.0
```

L2

```
model_ridge = linear_model.Ridge()
model_ridge.fit(X, y)
model_ridge.coef_
```

```
array([[ 0.        ,  0.02758377,  0.02763372, -0.05264566,  0.11724253,
         0.10228963,  0.15008299,  0.45839441, -0.02333363, -0.14812609]])
```

```
model_ridge.score(X, y)
```

```
0.014146680619607666
```

LOG TRANSFORMATION + REGULARIZATION

L1

```
model_lasso = linear_model.Lasso()
model_lasso.fit(X, y)
model_lasso.coef_
```

```
array([ 0.,  0., -0., -0.,  0.,  0.,  0.,  0., -0., -0.])
```

```
model_lasso.score(X, y)
```

```
0.0
```

L2

```
model_ridge = linear_model.Ridge()
model_ridge.fit(X, y)
model_ridge.coef_
```

```
array([[ 0.        ,  0.02758377,  0.02763372, -0.05264566,  0.11724253,
         0.10228963,  0.15008299,  0.45839441, -0.02333363, -0.14812609]])
```

```
model_ridge.score(X, y)
```

```
0.014146680619607666
```

NOW WHAT?

NOW WHAT?

THE NUMBERS®

Where Data and the Movie Business Meet

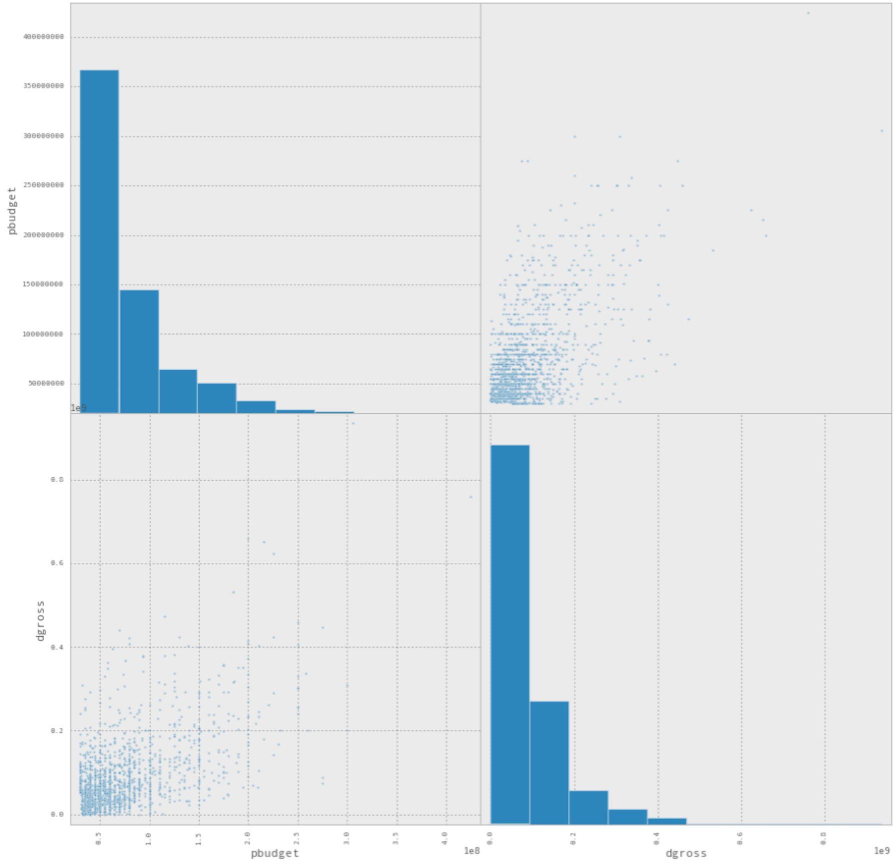
NOW WHAT?

THE NUMBERS®

Where Data and the Movie Business Meet

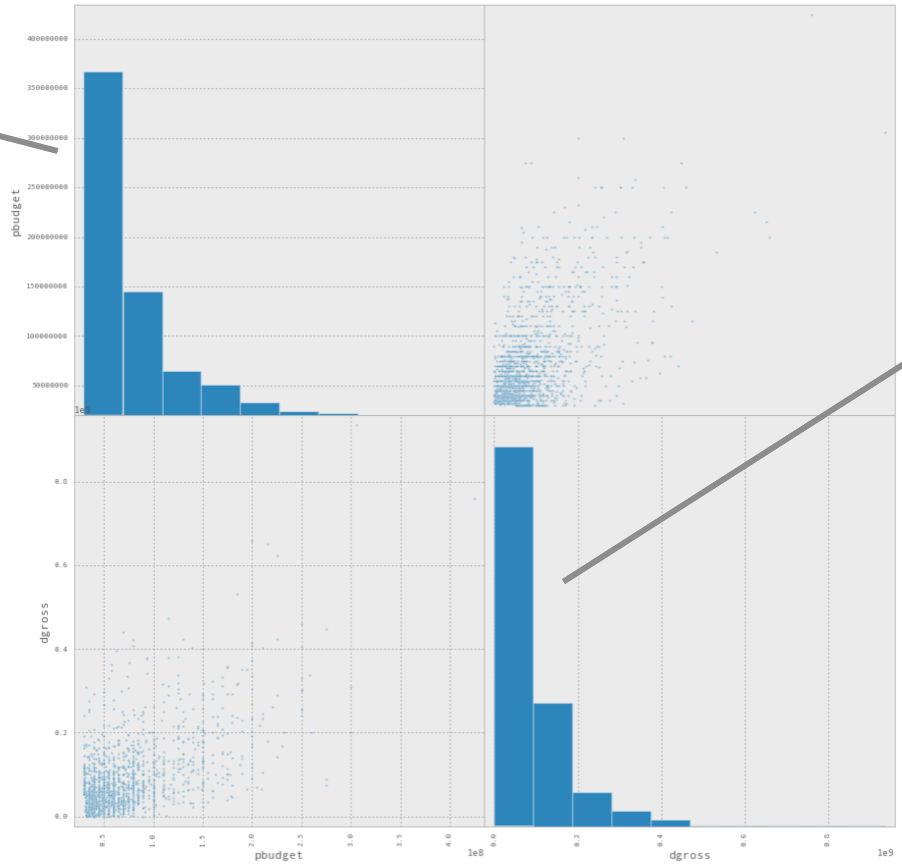
Does genre AND budget affect a movie's earning potential?

FIRST LOOK



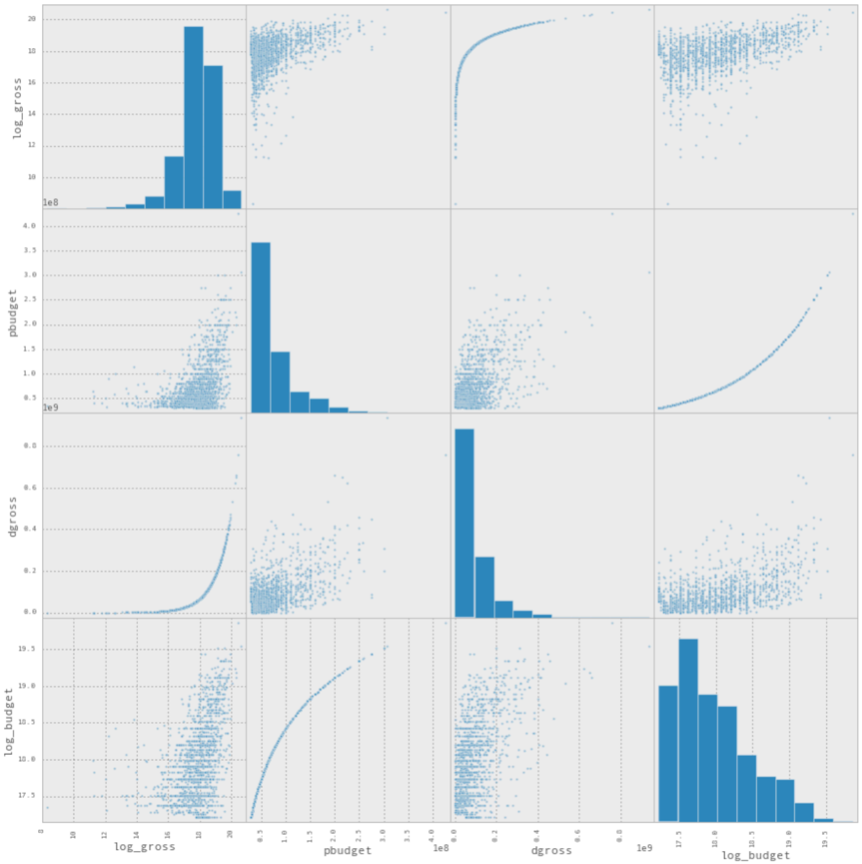
FIRST LOOK

Left skew



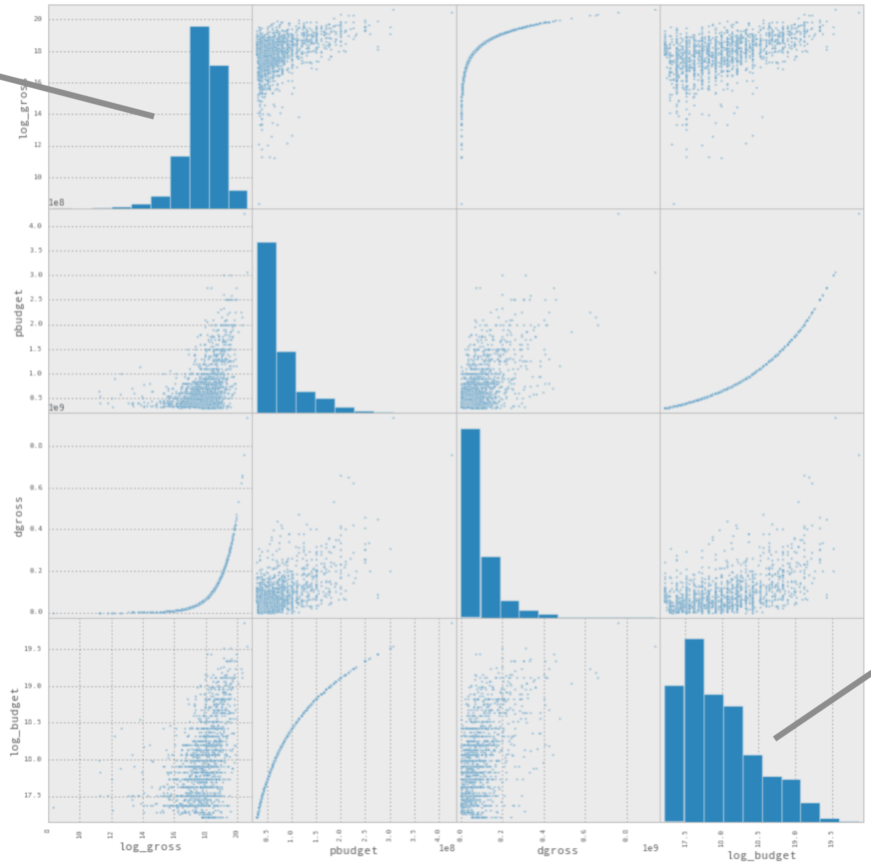
Left skew

LOG TRANSFORMATION



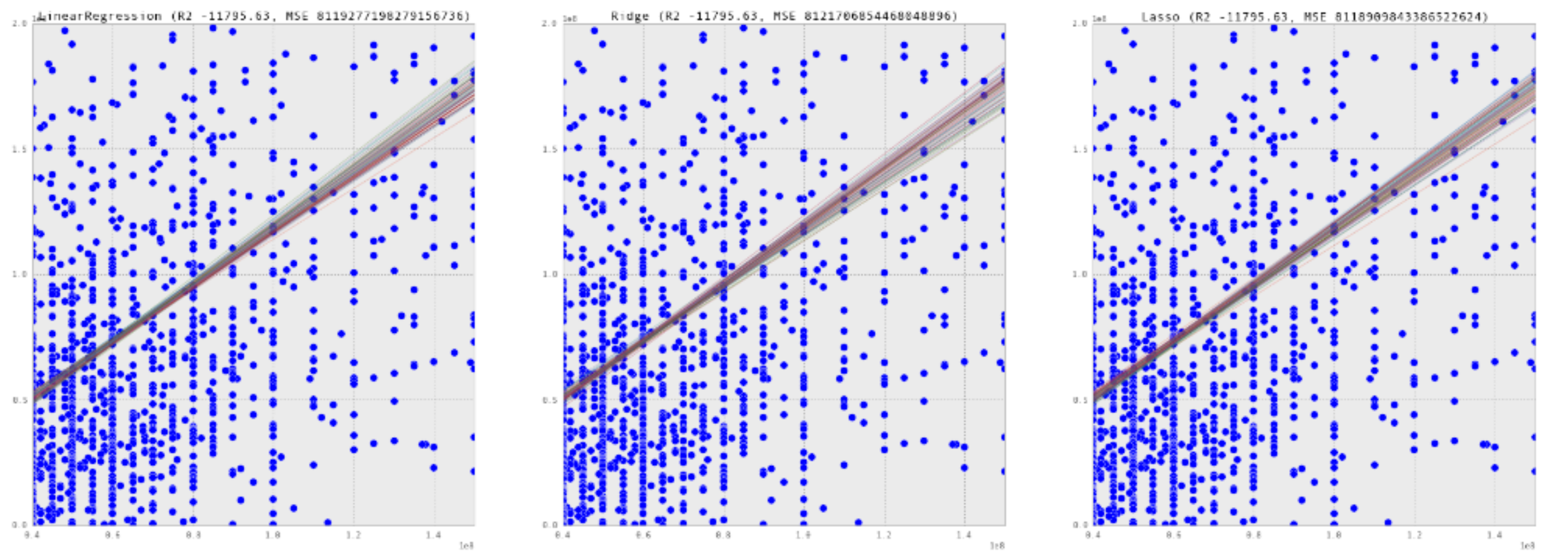
LOG TRANSFORMATION

Better?

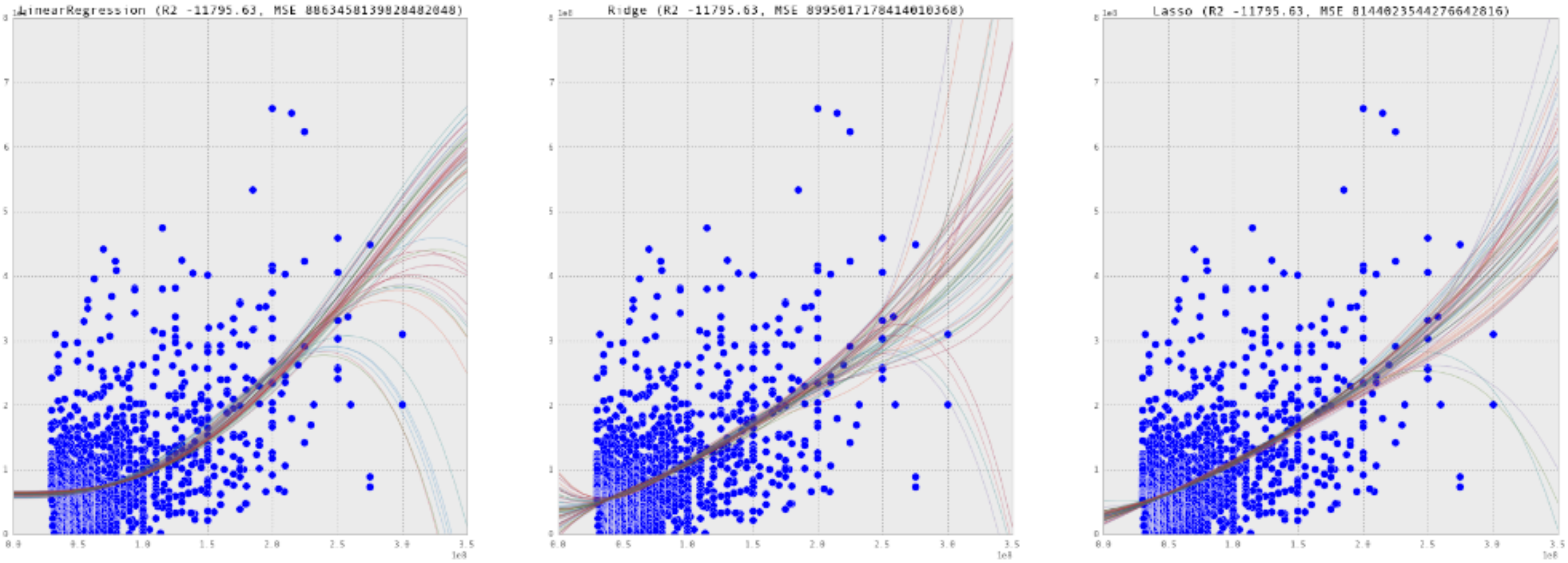


Better?

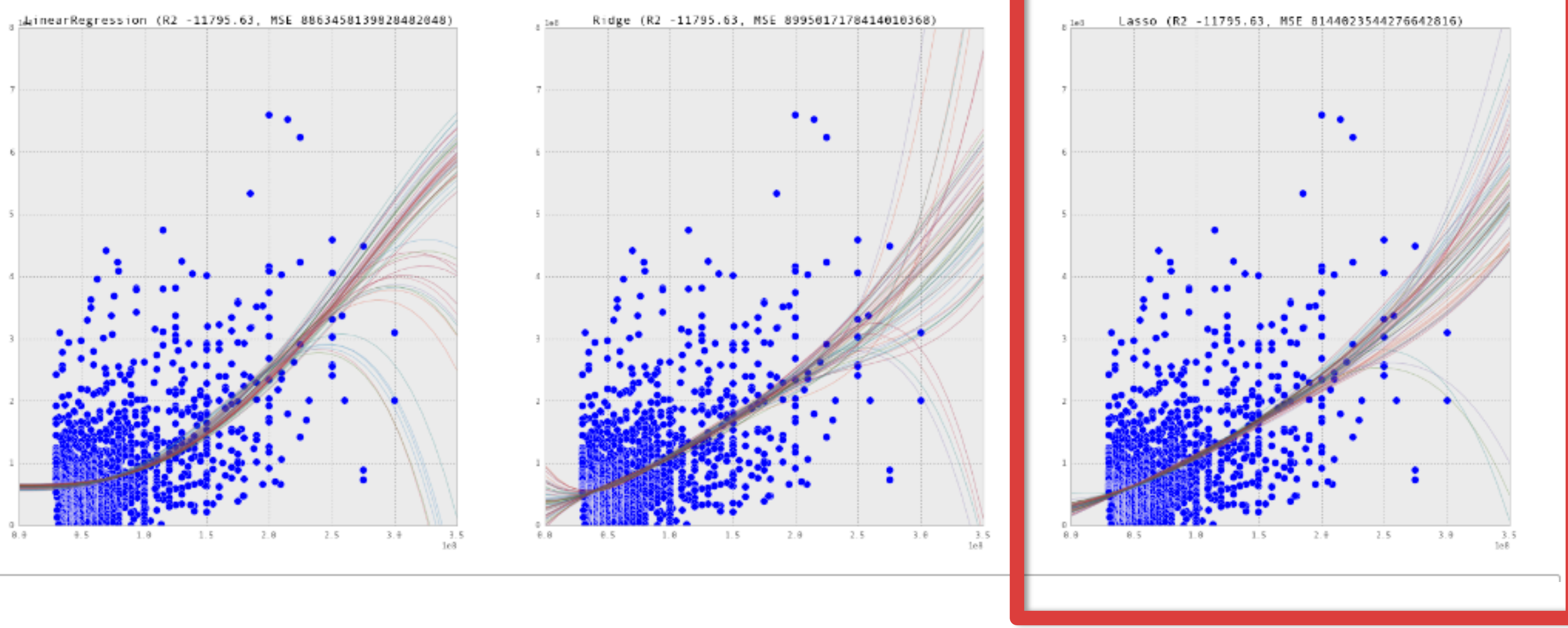
REGULARIZATION + CROSS-VALIDATION



REGULARIZATION + CROSS-VALIDATION



REGULARIZATION + CROSS-VALIDATION



FINAL MODEL

```
y, X = dmatrices('log_gross ~ log_budget + thriller + comedy + drama + documentary + action + animation + horror + fantasy + romanc
```

```
import sklearn
```

```
#X = sklearn.preprocesnormalize(X, axis=0)
#y = sklearn.preprocesnormalize(y, axis=0)
x_train, x_test, y_train, y_test = cv.train_test_split(X, y, test_size=0.20, random_state=1234)
model_lasso1 = linear_model.LassoCV(eps=0.001, n_alphas=100, cv=10, normalize=True).fit(x_train, sklearn.utils.column_or_1d(y_trair

#model_lasso1.predict(x_test,y_test)
print(metrics.mean_squared_error(y_train, model_lasso1.predict(x_train)))
print(metrics.mean_squared_error(y_test, model_lasso1.predict(x_test)))

print('alpha=', model_lasso1.alpha_)
m_alphas = model_lasso1.alphas_
print model_lasso1.coef_
```

```
1.14681260853
0.988424954023
('alpha=', 0.00071078618545511182)
[ 0.          0.94557742 -0.         -0.         -0.00890781  0.          0.1060979
  0.          0.1477817  -0.         -0.03913985]
```

APPENDIX

DATA SOURCES

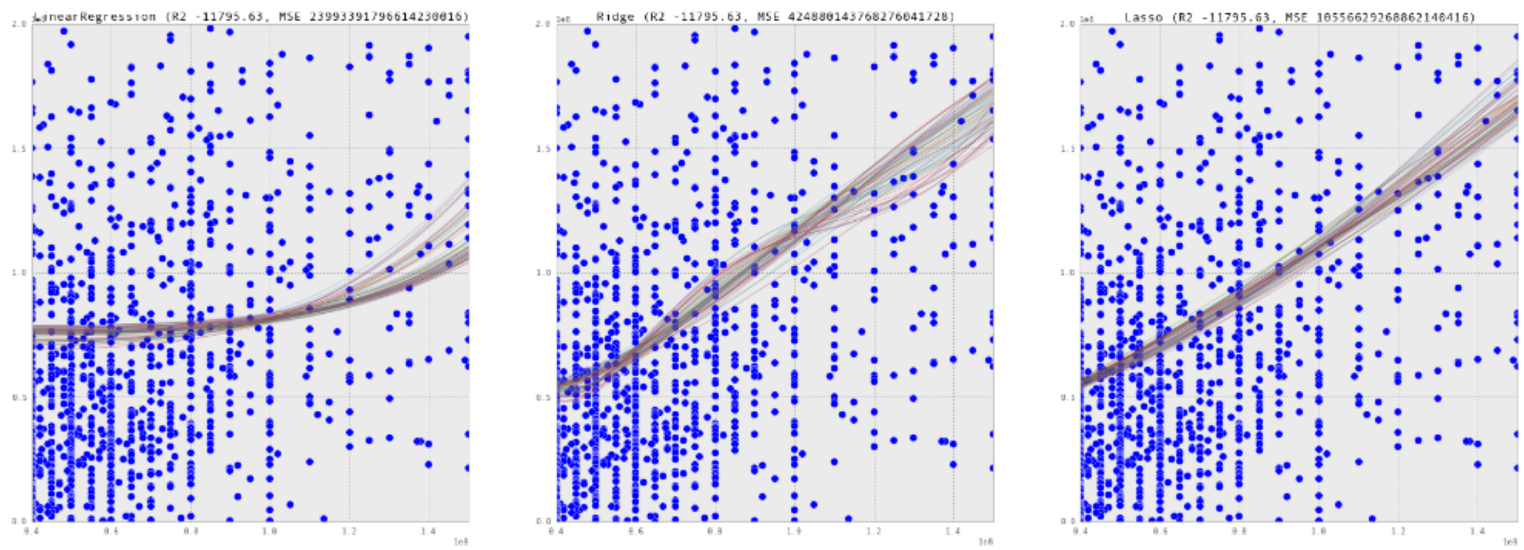
Numbers

Movielens

```
In [141]: len(set1.union(set2))
Out[141]: 6743
```

FEATURES

1. **Genres**
    1. Drama
    2. Comedy
    3. Thriller
2. **Production budget**
    1. For top 6000 movies


LinearRegression (R2 -11795.63, MSE 23993391796614230016)
Ridge (R2 -11795.63, MSE 424880143768276041728)
Lasso (R2 -11795.63, MSE 10556629268862140416)

MOVIELENS DATASET

1. Animation
2. Drama
3. Adventure
4. Children
5. Sci-Fi
6. Horror
7. Comedy
8. ~~Romance~~
9. Drama
10. Thriller
11. ~~Mystery~~
12. ~~Fantasy~~
13. ~~Documentary~~
14. ~~Musical~~
15. ~~Western~~
16. ~~War~~
17. ~~Crime~~
18. ~~Film-Noir~~