Allison Schlissel

Predicting Movie Worldwide Gross

An Exploration of Web Scraping and Linear Regression

July 15, 2016

The Client

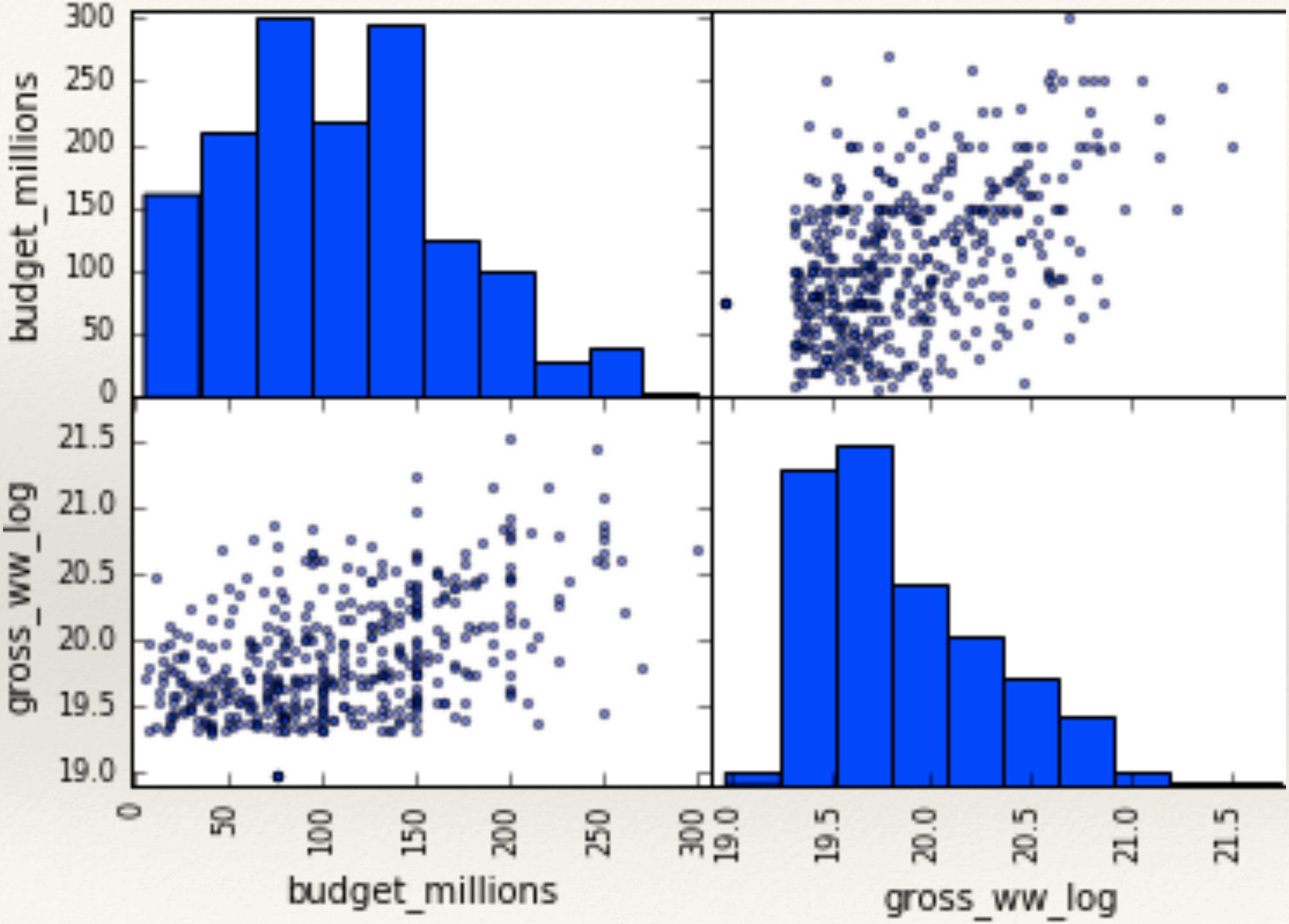
- A movie investor who wants to ensure that a movie will yield a large worldwide gross
- Movie Features
  - Budget
  - Genre
  - MPAA Rating

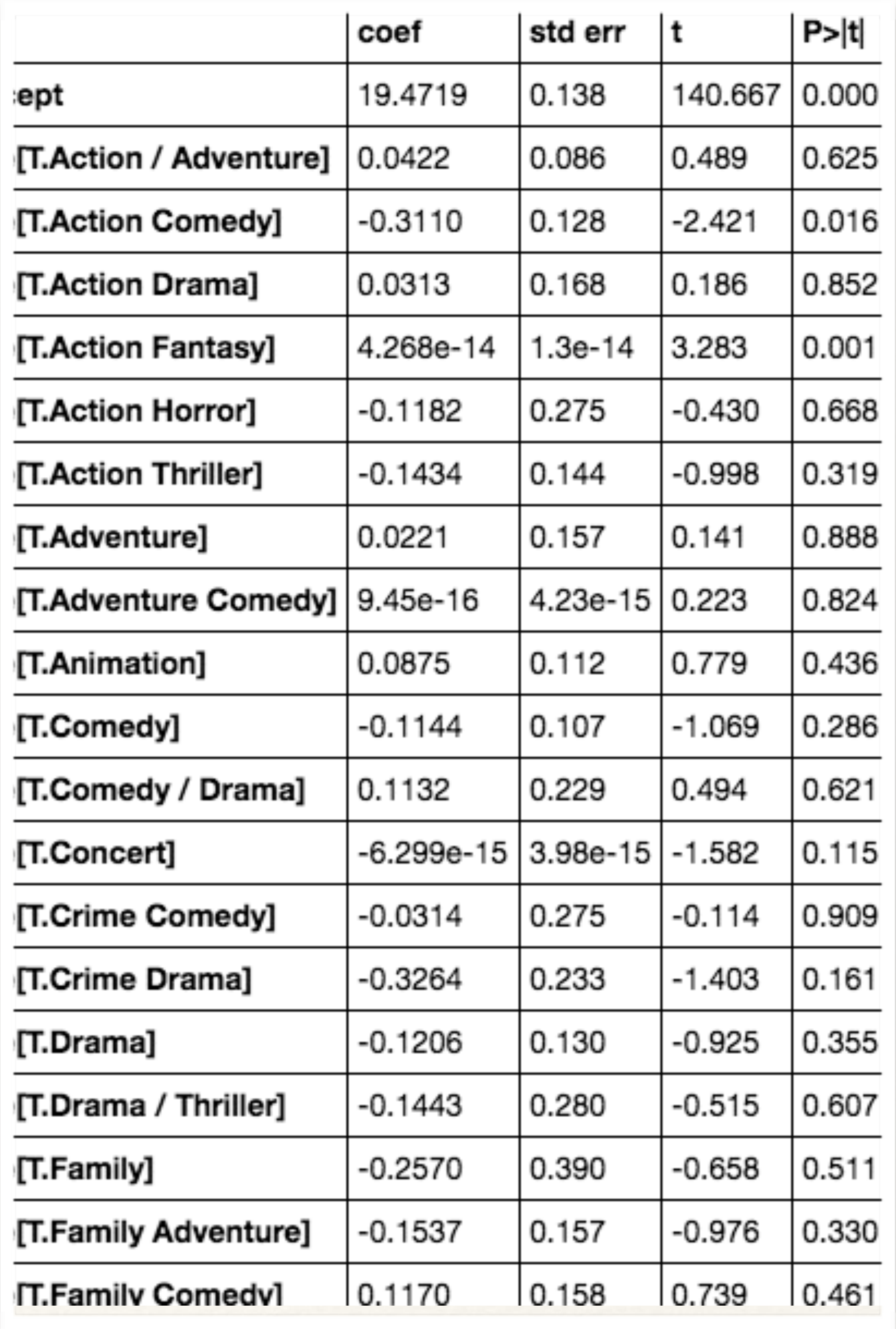
The Process

- Web scraping with BeautifulSoup
- DataFrame of web values for OLS
- Lasso Regularization

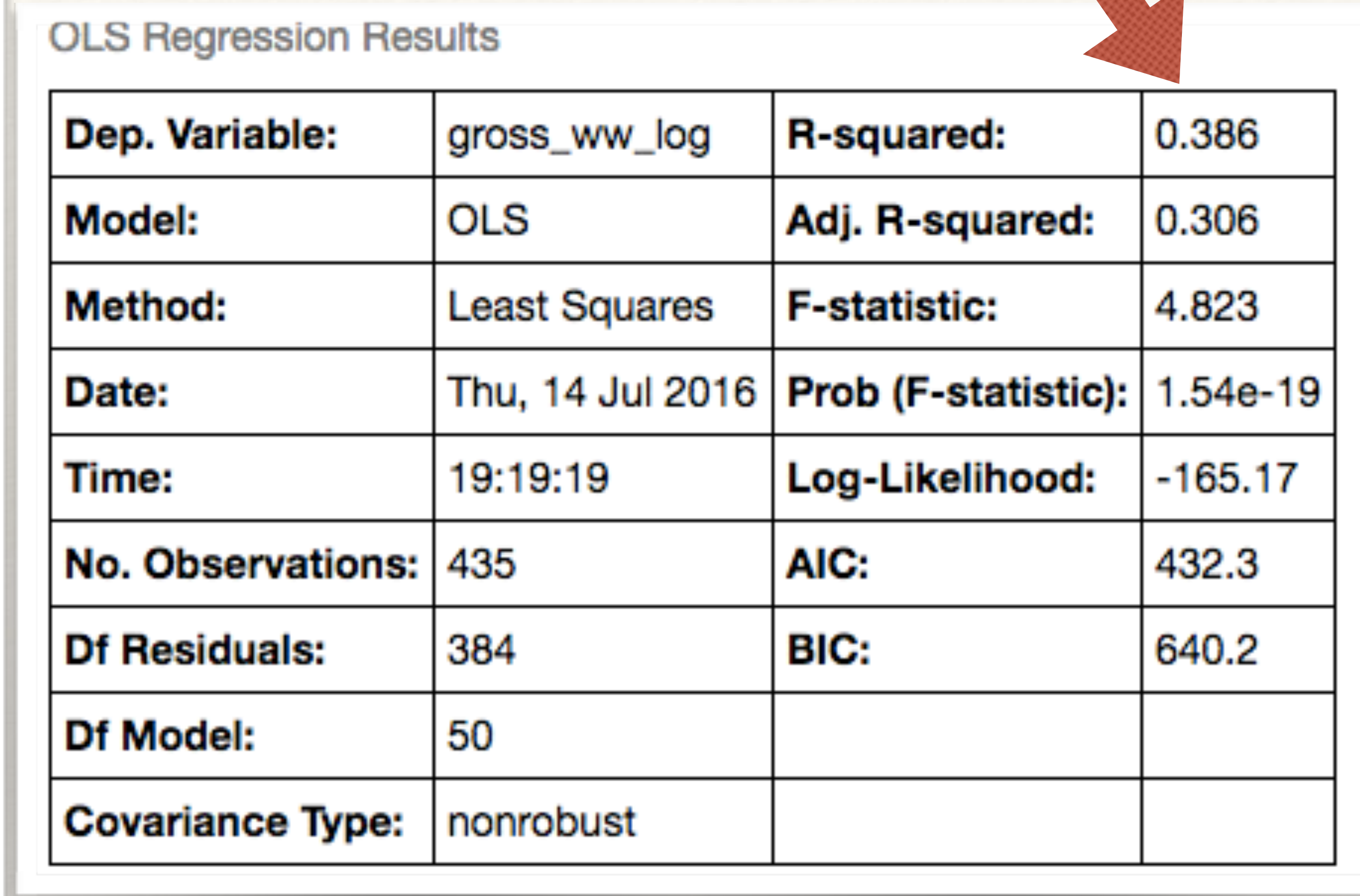
Web Scraping

- boxofficemojo.com
- BeautifulSoup
- Data: 500 top grossing movies
- dependent variable: worldwide gross
- independent variables: budget, genre, MPAA rating


budget_millions
300
250
200
150
100
50
0
gross_ww_log
21.5
21.0
20.5
20.0
19.5
19.0
0
50
100
150
200
250
300
budget_millions
19.0
19.5
20.0
20.5
21.0
21.5
gross_ww_log



	coef	std err	t	P>|t|
ept	19.4719	0.138	140.667	0.000
[T.Action / Adventure]	0.0422	0.086	0.489	0.625
[T.Action Comedy]	-0.3110	0.128	-2.421	0.016
[T.Action Drama]	0.0313	0.168	0.186	0.852
[T.Action Fantasy]	4.268e-14	1.3e-14	3.283	0.001
[T.Action Horror]	-0.1182	0.275	-0.430	0.668
[T.Action Thriller]	-0.1434	0.144	-0.998	0.319
[T.Adventure]	0.0221	0.157	0.141	0.888
[T.Adventure Comedy]	9.45e-16	4.23e-15	0.223	0.824
[T.Animation]	0.0875	0.112	0.779	0.436
[T.Comedy]	-0.1144	0.107	-1.069	0.286
[T.Comedy / Drama]	0.1132	0.229	0.494	0.621
[T.Concert]	-6.299e-15	3.98e-15	-1.582	0.115
[T.Crime Comedy]	-0.0314	0.275	-0.114	0.909
[T.Crime Drama]	-0.3264	0.233	-1.403	0.161
[T.Drama]	-0.1206	0.130	-0.925	0.355
[T.Drama / Thriller]	-0.1443	0.280	-0.515	0.607
[T.Family]	-0.2570	0.390	-0.658	0.511
[T.Family Adventure]	-0.1537	0.157	-0.976	0.330
[T.Family Comedy]	0.1170	0.158	0.739	0.461

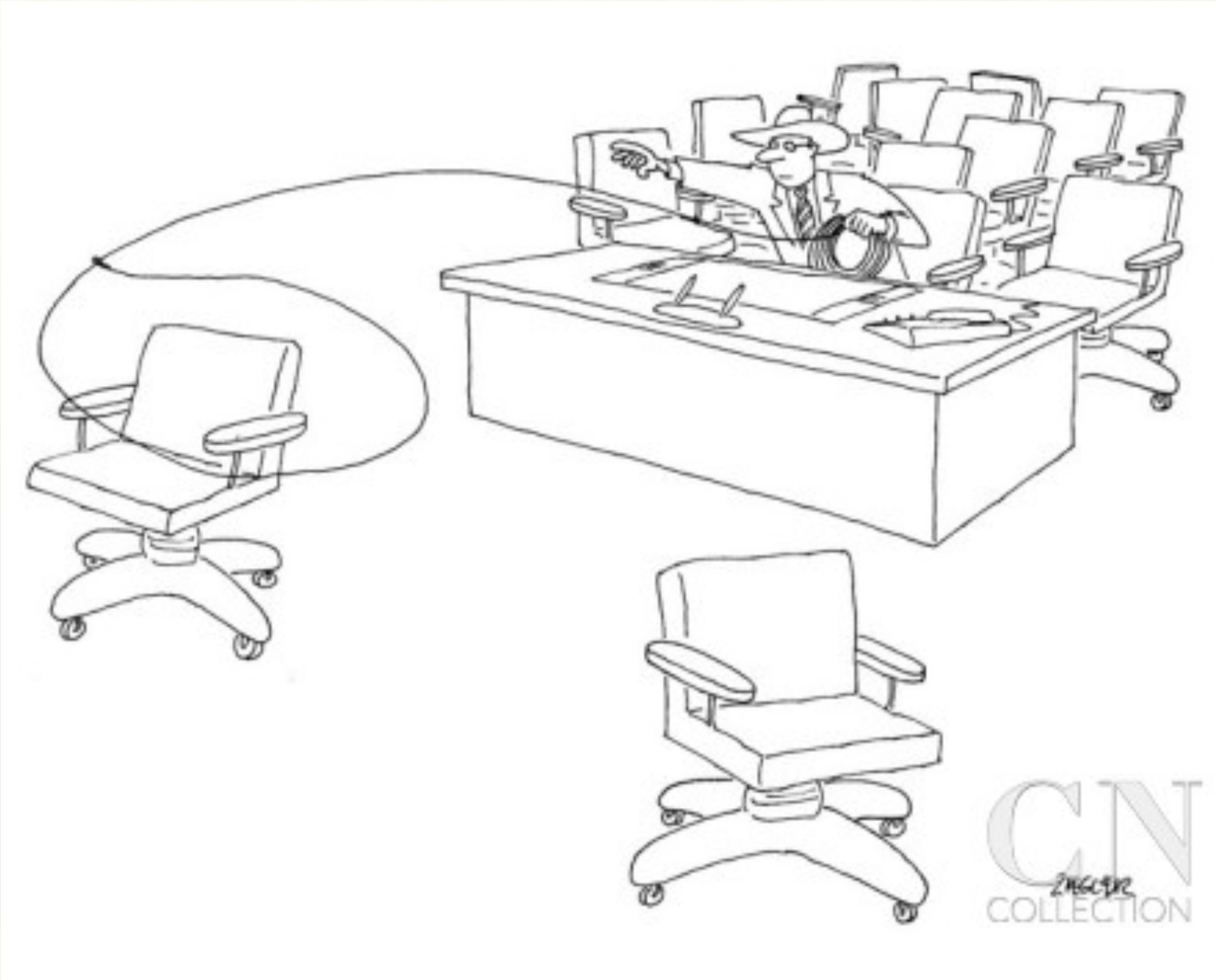


OLS Regression Results

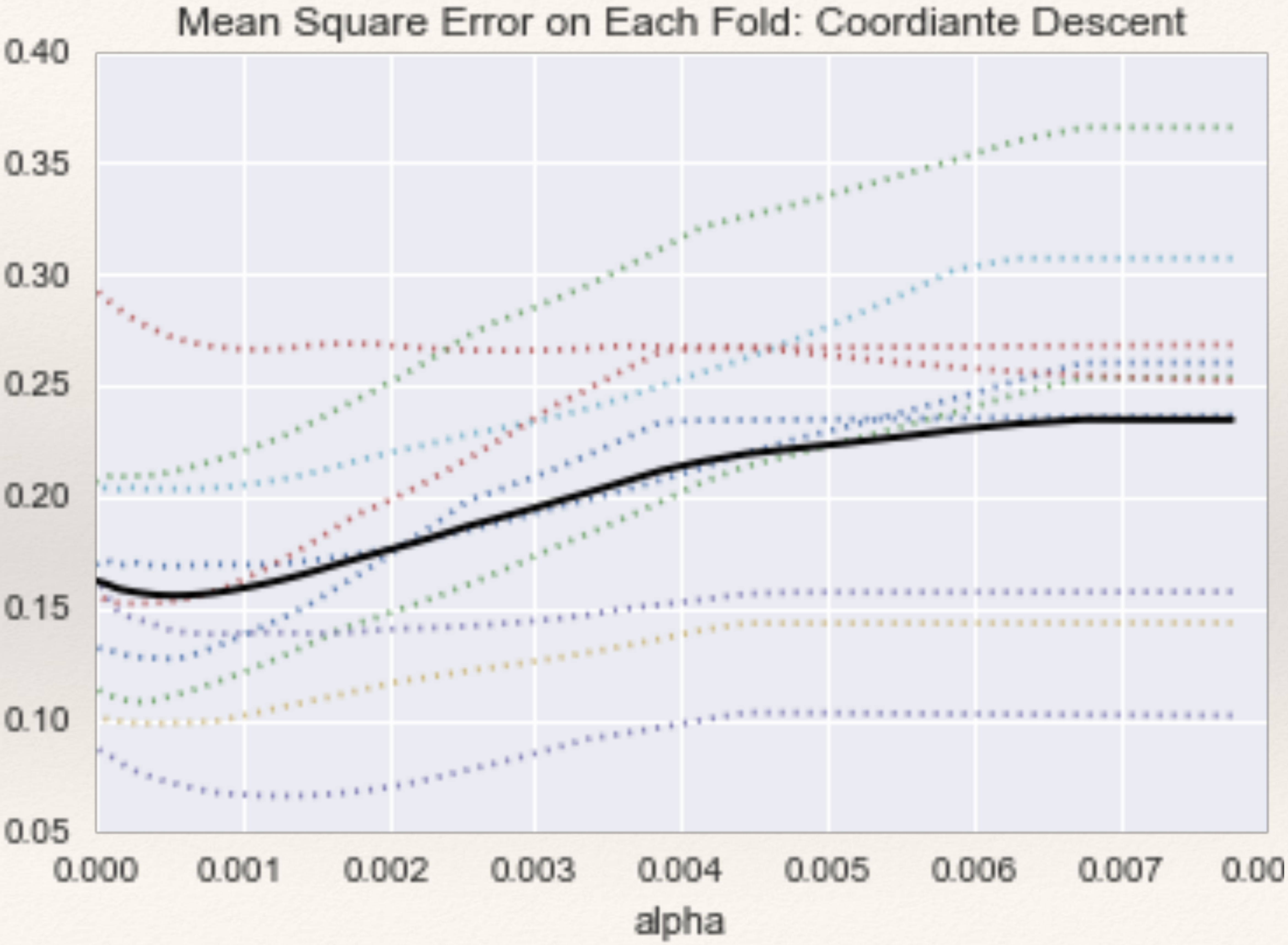
Dep. Variable:	gross_ww_log	**R-squared:**	0.386
Model:	OLS	**Adj. R-squared:**	0.306
Method:	Least Squares	**F-statistic:**	4.823
Date:	Thu, 14 Jul 2016	**Prob (F-statistic):**	1.54e-19
Time:	19:19:19	**Log-Likelihood:**	-165.17
No. Observations:	435	**AIC:**	432.3
Df Residuals:	384	**BIC:**	640.2
Df Model:	50		
Covariance Type:	nonrobust		

Lasso Regularization

- LassoCV from sklearn
- Test Size: 20%
- Mean Squared Error Train: 0.133459747906
- Mean Squared Error Test: 0.146652316896
- alpha= 0.000547619453713



Results



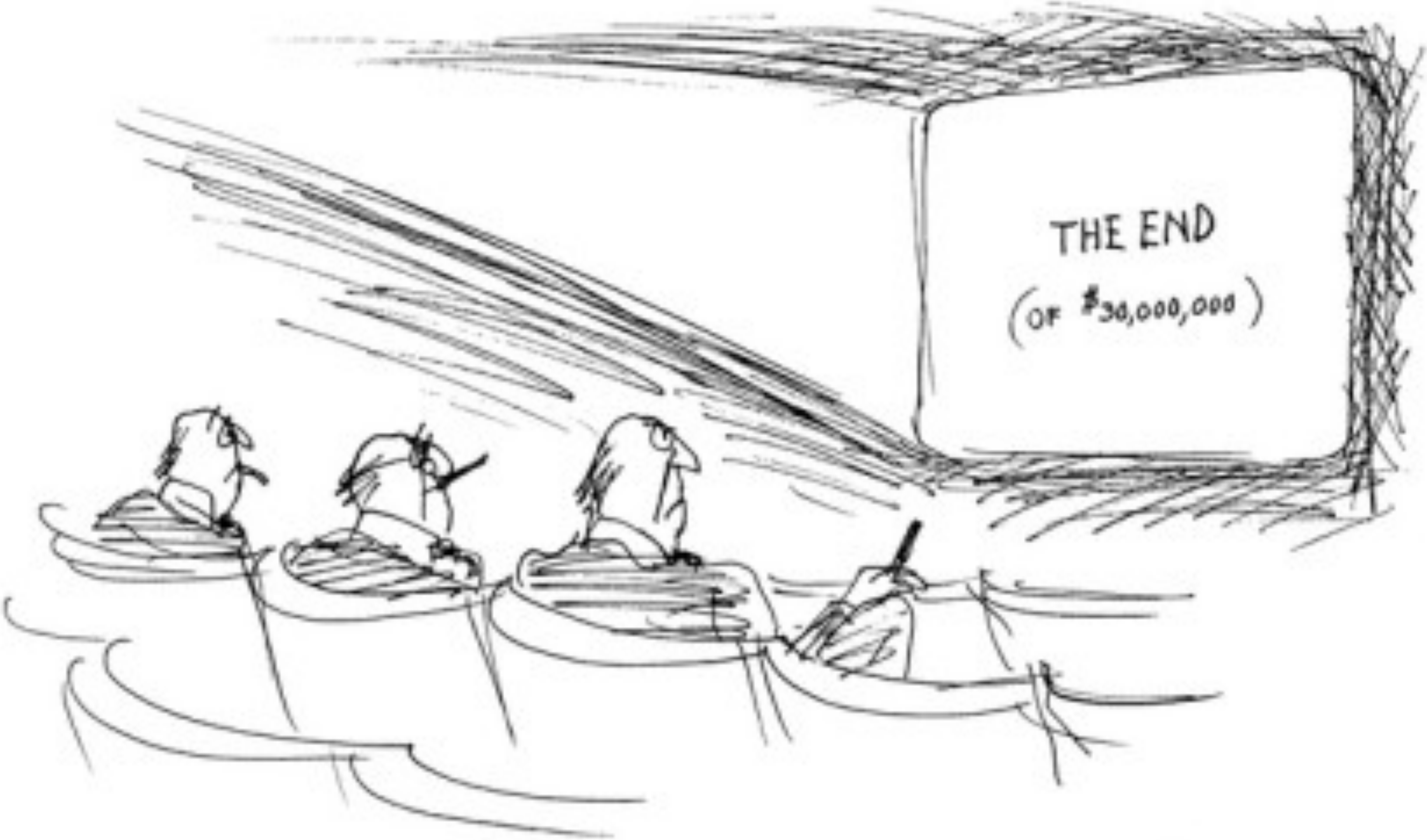
Mean Square Error on Each Fold: Coordiante Descent
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.00
alpha

Results

- Most important indicators ($P < 0.1$):
  - Budget
  - Genres: Action/Comedy, Action/Fantasy, Concert, Fantasy, Foreign, Period/Horror, Sci-Fi/Adventure, Sci-Fi/Fantasy, Sci-Fi/Horror
  - MPAA rating not significant

Future Relationship

- Star Power
- Release Date
- Movie Franchises
- Publicity Scandals
- Gender / Ethnic Diversity


THE END
(OF $30,000,000)