

Supplementary Handout

Supplementary material that didn't fit on the poster.

Data Scraping

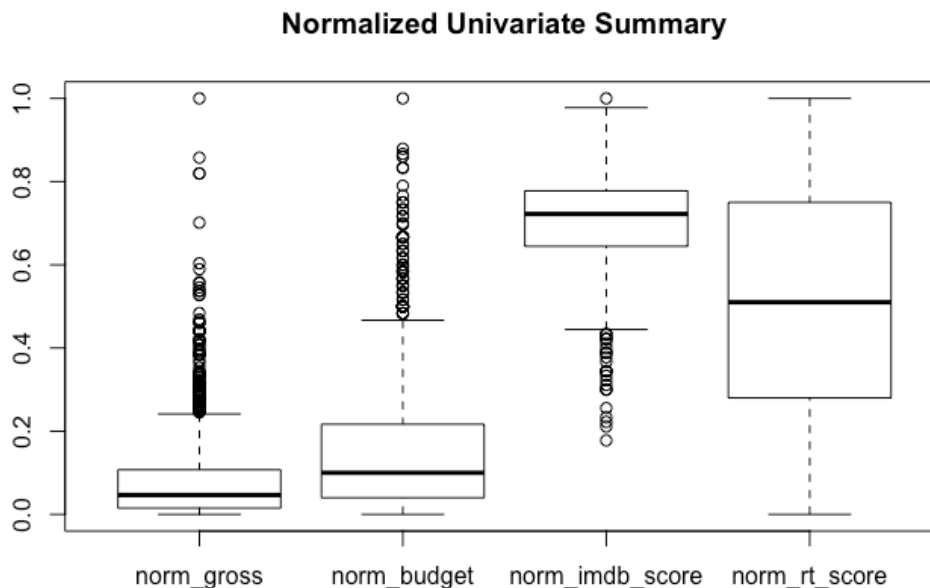
We scraped Rotten Tomatoes data using “urllib2” and “re”.

We tried a variety of different methods in order to scrape data from metacritic but there wasn't much success. A few of the methods we tried included:

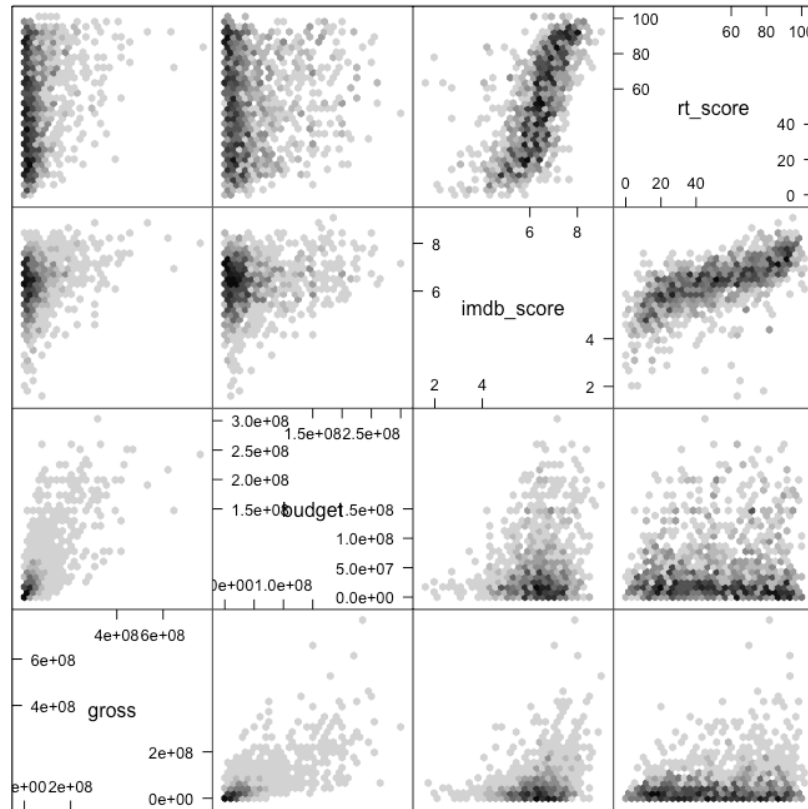
- Scraping directly from metacritic.
 - Failed because metacritic blocks requests made via urllib in python. Potentially there are other scraping methods that could circumvent this if we could set up the GET request properly in html (after all, it serves requests made from web browsers).
- Scraping from search engine results (google).
 - This didn't work because google blocks your IP after a certain number of requests. Potentially using a timer and scraping at random intervals (perhaps Poisson would be best) could circumvent this.

Exploratory Analysis

Univariate summaries.



Scatter plot matrix for our continuous random variables.



Scatter Plot Matrix

Initial Regression Result

Call:

```
lm(formula = gross ~ ., data = data_initial)
```

Residuals:

Min	1Q	Median	3Q	Max
-236630412	-26076779	-7207365	16453032	460557534

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5.838e+07	1.291e+07	-4.521	6.72e-06 ***
budget	1.097e+00	4.062e-02	27.001	< 2e-16 ***
imdb_score	8.658e+06	2.159e+06	4.011	6.40e-05 ***
Adventure	-7.292e+06	5.172e+06	-1.410	0.15884
Animation	5.613e+06	7.699e+06	0.729	0.46610
Biography	5.138e+06	7.348e+06	0.699	0.48450
Comedy	8.833e+06	4.102e+06	2.153	0.03147 *
Crime	-6.446e+06	4.799e+06	-1.343	0.17940
Documentary	-1.827e+07	1.093e+07	-1.672	0.09470 .
Drama	-1.121e+07	3.901e+06	-2.872	0.00414 **
Family	-4.552e+06	6.022e+06	-0.756	0.44989
Fantasy	-7.261e+05	4.776e+06	-0.152	0.87918
History	-7.774e+06	1.020e+07	-0.762	0.44618
Horror	6.477e+06	5.873e+06	1.103	0.27030

Music	1.009e+07	7.389e+06	1.365	0.17239
Musical	-2.715e+06	1.039e+07	-0.261	0.79394
Mystery	-1.456e+06	5.668e+06	-0.257	0.79738
Romance	8.988e+05	4.117e+06	0.218	0.82724
Sci.Fi	6.495e+06	5.093e+06	1.275	0.20244
Sport	-8.764e+06	7.761e+06	-1.129	0.25900
Thriller	-8.084e+05	4.422e+06	-0.183	0.85498
War	7.136e+06	9.634e+06	0.741	0.45901
Western	-1.459e+07	1.529e+07	-0.954	0.34037
rt_score	3.673e+05	7.802e+04	4.707	2.78e-06 ***
2007	-3.844e+06	6.807e+06	-0.565	0.57239
2008	-3.283e+06	6.523e+06	-0.503	0.61488
2009	1.060e+06	6.516e+06	0.163	0.87085
2010	-3.376e+06	6.608e+06	-0.511	0.60949
2011	-7.857e+06	6.692e+06	-1.174	0.24055
2012	3.803e+06	6.644e+06	0.572	0.56719
2013	-5.225e+06	6.712e+06	-0.779	0.43638
2014	3.953e+06	6.754e+06	0.585	0.55845
2015	2.968e+06	7.045e+06	0.421	0.67357

Residual standard error: 54160000 on 1296 degrees of freedom
Multiple R-squared: 0.5759, Adjusted R-squared: 0.5655
F-statistic: 55 on 32 and 1296 DF, p-value: < 2.2e-16

Transformations

bcPower Transformations to Multinormality

	Est.Power	Std.Err.	Wald Lower Bound	Wald Upper Bound
Y1	0.2952	0.0109	0.2739	0.3165
Y2	0.2543	0.0139	0.2270	0.2816
Y3	2.4339	0.1108	2.2167	2.6510
Y4	0.9676	0.0467	0.8761	1.0590

Likelihood ratio tests about transformation parameters

	LRT	df	pval
LR test, lambda = (0 0 0 0)	2995.474884	4	0.0000000
LR test, lambda = (1 1 1 1)	3899.397058	4	0.0000000
LR test, lambda = (0.3 0.25 2.43 1)	0.480411	4	0.9753797

So we used lambda = (0.25, 0.25, 2, 1), where Y1 is gross, Y2 is budget, Y3 is imdb score and Y4 is rotten tomatoes score.

Censor Results

Call:

```
lm(formula = data_c6$t_gross ~ ., data = as.data.frame(data_c6))
```

Residuals:

Min	1Q	Median	3Q	Max
-335.62	-45.60	0.31	46.17	338.66

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.289e-12	7.322e+01	0.000	1.000000
t_budget	9.866e-01	2.627e-02	37.556	< 2e-16 ***
t_imdb_score	-7.399e+01	1.415e+01	-5.230	1.97e-07 ***
t_rt_score	1.394e+01	3.484e+00	4.002	6.64e-05 ***
t_imdb_score2	6.939e+00	1.219e+00	5.692	1.55e-08 ***
year2006	1.862e+02	8.455e+01	2.202	0.027817 *
year2007	1.774e+02	8.465e+01	2.096	0.036285 *
year2008	1.803e+02	8.458e+01	2.132	0.033191 *
year2009	1.819e+02	8.471e+01	2.148	0.031921 *
year2010	1.847e+02	8.476e+01	2.179	0.029490 *
year2011	1.831e+02	8.467e+01	2.162	0.030791 *
year2012	1.990e+02	8.481e+01	2.347	0.019074 *
year2013	1.991e+02	8.477e+01	2.349	0.018978 *
year2014	2.073e+02	8.474e+01	2.447	0.014543 *
year2015	1.905e+02	8.484e+01	2.246	0.024896 *
c6_1	1.653e+01	4.904e+00	3.371	0.000772 ***
c6_2	-4.686e+00	4.761e+00	-0.984	0.325241
c6_3	-2.049e+01	6.748e+00	-3.036	0.002445 **
c6_4	-4.309e+00	5.179e+00	-0.832	0.405524
c6_5	-6.840e+00	5.433e+00	-1.259	0.208275
c6_6	2.546e+01	6.101e+00	4.172	3.21e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 73.22 on 1309 degrees of freedom
Multiple R-squared: 0.625, Adjusted R-squared: 0.6192
F-statistic: 109.1 on 20 and 1309 DF, p-value: < 2.2e-16

Call:

```
censReg(formula = data_c6$t_gross ~ data_c6$t_budget + data_c6$t_imdb_score +
  data_c6$t_rt_score + data_c6$t_imdb_score2 + data_c6$year +
  data_c6$c6_1 + data_c6$c6_2 + data_c6$c6_3 + data_c6$c6_4 +
  data_c6$c6_5 + data_c6$c6_6)
```

Observations:

Total	Left-censored	Uncensored	Right-censored
1330	1	1329	0

(Note that we augmented with one zero variable since we didn't actually have any movies that had exactly zero gross)

Coefficients:

	Estimate	Std. error	t value	Pr(> t)
(Intercept)	-231.20984	809.64485	-0.286	0.775208
data_c6\$t_budget	0.98660	0.02607	37.842	< 2e-16 ***
data_c6\$t_imdb_score	-73.96865	14.03893	-5.269	1.37e-07 ***
data_c6\$t_rt_score	13.94225	3.45777	4.032	5.53e-05 ***
data_c6\$t_imdb_score2	6.93755	1.20982	5.734	9.79e-09 ***
data_c6\$year2006	417.36904	810.73131	0.515	0.606689
data_c6\$year2007	408.58067	810.74172	0.504	0.614290
data_c6\$year2008	411.47693	810.73384	0.508	0.611779
data_c6\$year2009	413.09014	810.74759	0.510	0.610389
data_c6\$year2010	415.86186	810.75224	0.513	0.607998

```

data_c6$year2011    414.22537  810.74352  0.511 0.609407
data_c6$year2012    430.19563  810.75734  0.531 0.595689
data_c6$year2013    430.27326  810.75374  0.531 0.595621
data_c6$year2014    438.50428  810.75073  0.541 0.588603
data_c6$year2015    421.68073  810.76120  0.520 0.602991
data_c6$c6_1        16.52826   4.86652   3.396 0.000683 ***
data_c6$c6_2        -4.68529   4.72530  -0.992 0.321426
data_c6$c6_3       -20.48595   6.69707  -3.059 0.002221 **
data_c6$c6_4        -4.30921   5.14010  -0.838 0.401834
data_c6$c6_5        -6.84014   5.39199  -1.269 0.204592
data_c6$c6_6        25.45859   6.05539   4.204 2.62e-05 ***
logSigma            4.28595   0.01940 220.966 < 2e-16 ***

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Newton-Raphson maximisation, 16 iterations

Return code 2: successive function values within tolerance limit

Log-likelihood: -7581.796 on 22 Df

The only coefficients that really differ are the year factors and the intercept. All of the covariates of interest don't change in value.

Regression Results Without Outliers and Influential Points

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5.218e+03	1.438e+03	-3.629	0.000295 ***
t_budget	9.918e-01	2.584e-02	38.384	< 2e-16 ***
t_imdb_score	-8.160e+01	1.434e+01	-5.689	1.57e-08 ***
t_rt_score	1.253e+01	3.434e+00	3.649	0.000273 ***
t_imdb_score2	7.616e+00	1.226e+00	6.210	7.11e-10 ***
Year	2.698e+00	7.161e-01	3.768	0.000172 ***
c6_1	1.751e+01	4.827e+00	3.628	0.000297 ***
c6_2	-3.781e+00	4.689e+00	-0.806	0.420210
c6_3	-1.970e+01	6.609e+00	-2.980	0.002933 **
c6_4	-3.737e+00	5.093e+00	-0.734	0.463251
c6_5	-7.705e+00	5.331e+00	-1.445	0.148611
c6_6	2.547e+01	6.009e+00	4.239	2.40e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72.1 on 1314 degrees of freedom

Multiple R-squared: 0.6317, Adjusted R-squared: 0.6286

F-statistic: 204.9 on 11 and 1314 DF, p-value: < 2.2e-16