Modelling the Gross of the Top 2010s Movies

FEATURING JULIET AGUH, ALBERT PUTRANEGORO, JODIE CHEN, AND SAUL MAGALLON

Coming soon to Professor Cha's Winter 2024 Stats 101A

RESEARCH QUESTION

- Can we predict the domestic revenue of a 2010s movie by its month of release, Rotten Tomatoes scores (critics' and audience's), U.S. inflation at month of release, budget, and combined net worth of its two main stars?

- We gathered data from Rotten Tomatoes and IMDb, then put together a model

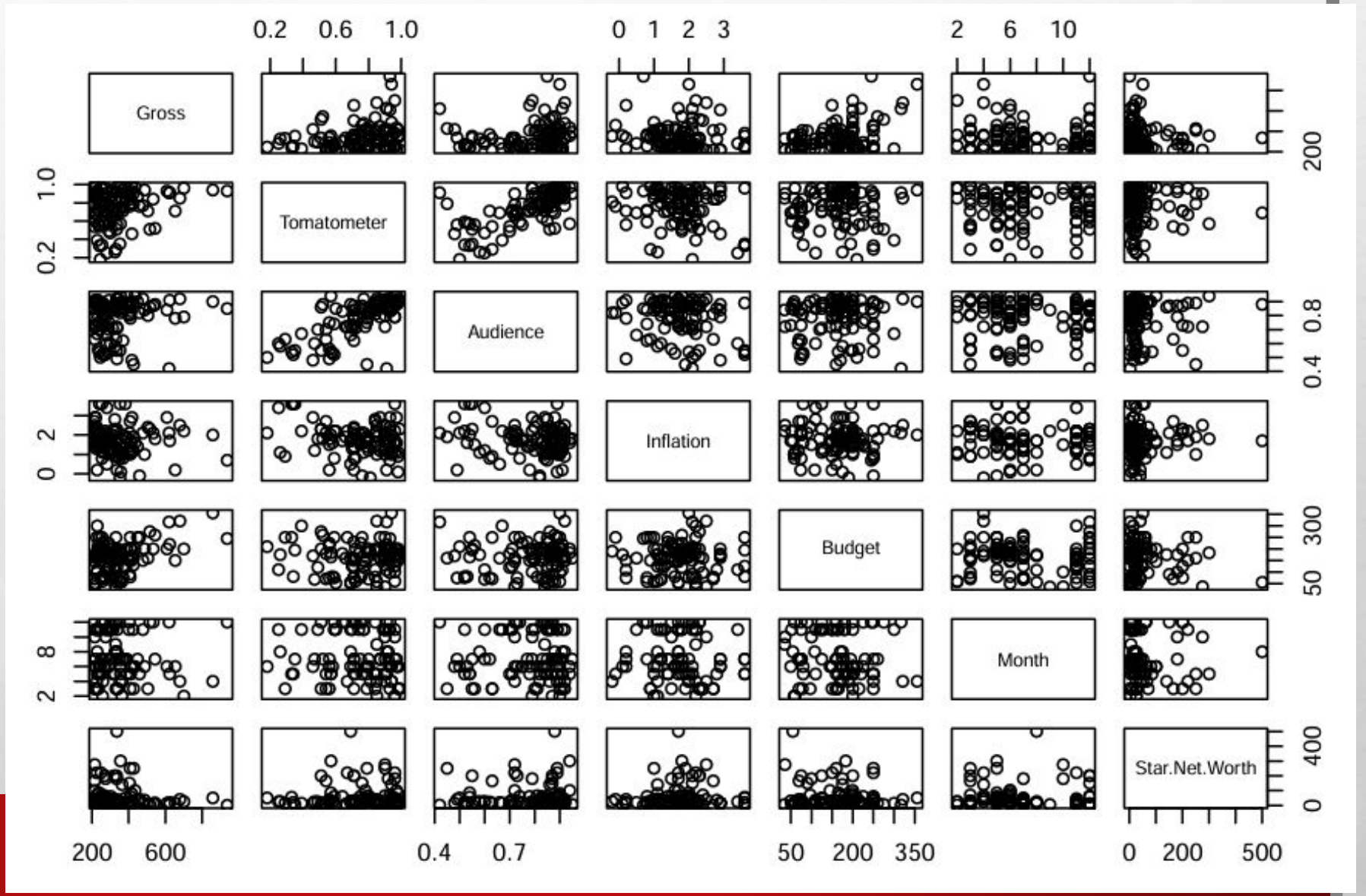
Y = Gross

X_1 = Tomatometer, X_2 = Audience, X_3 = Inflation, X_4 = Budget, X_5 = Month, X_6 = Star.Net.Worth

THE SUMMARY STATISTICS: VARIABLE DISTRIBUTIONS

```
     Film               Month            Gross          Tomatometer
 Length:100         Min.   : 2.00   Min.   :214.5   Min.   :0.1800
 Class :character   1st Qu.: 5.00   1st Qu.:251.0   1st Qu.:0.6300
 Mode  :character   Median : 7.00   Median :324.8   Median :0.7900
                    Mean   : 7.22   Mean   :350.6   Mean   :0.7471
                    3rd Qu.:11.00   3rd Qu.:405.4   3rd Qu.:0.9025
                    Max.   :12.00   Max.   :936.7   Max.   :0.9900
    Audience        Inflation         Budget          Star
 Min.   :0.4200   Min.   :-0.20   Min.   : 35.0   Length:100
 1st Qu.:0.7175   1st Qu.: 1.20   1st Qu.:117.5   Class :character
 Median :0.8350   Median : 1.75   Median :167.5   Mode  :character
 Mean   :0.7805   Mean   : 1.75   Mean   :163.8
 3rd Qu.:0.8825   3rd Qu.: 2.20   3rd Qu.:200.0
 Max.   :0.9500   Max.   : 3.60   Max.   :356.0
   Co.star        Star.Net.Worth
 Length:100         Min.   :  0.73
 Class :character   1st Qu.: 14.63
 Mode  :character   Median : 25.67
                    Mean   : 57.67
                    3rd Qu.: 60.72
                    Max.   :500.84
```

THE CORRELATION MATRIX



VARIABLE CORRELATIONS

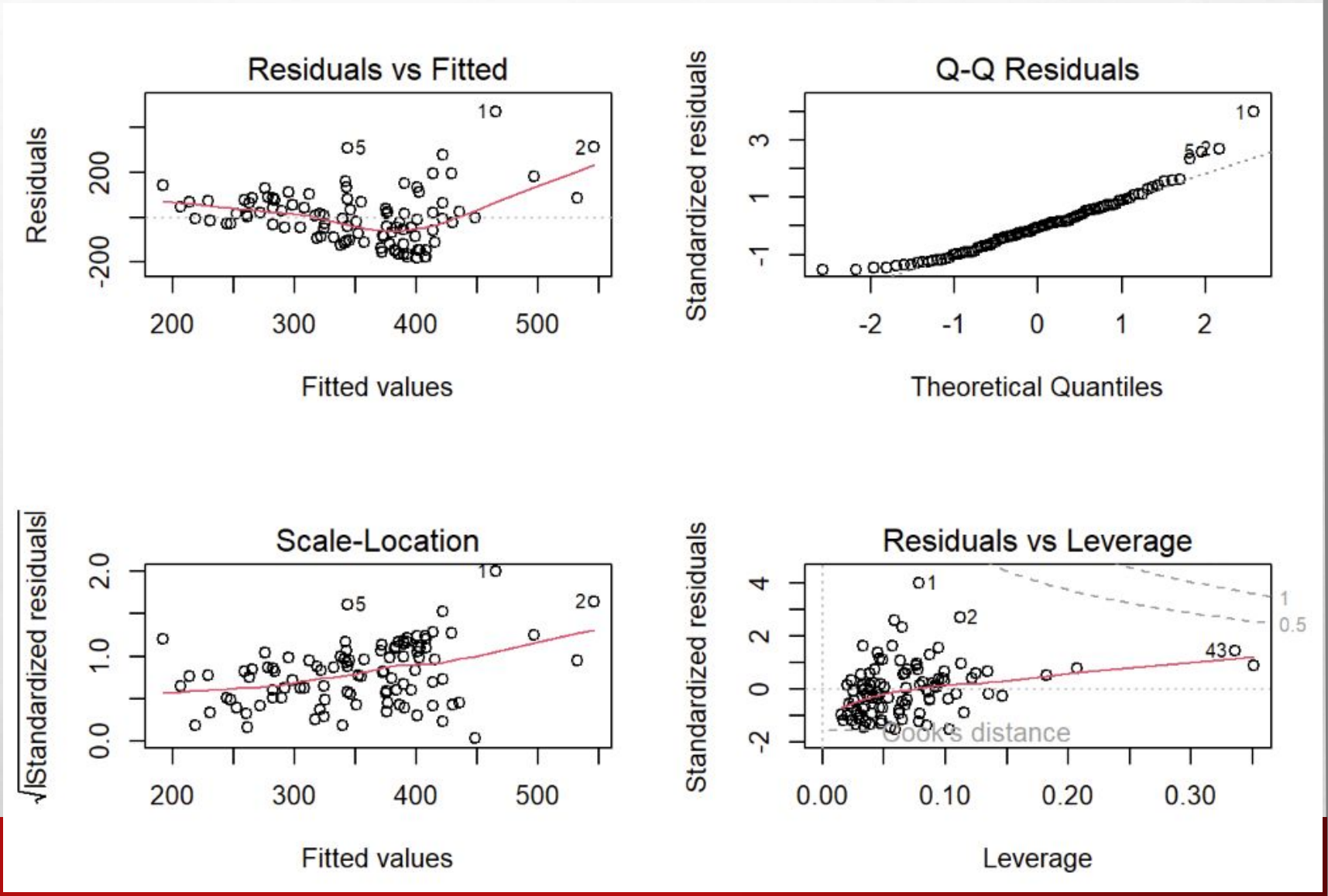
	Gross	Tomatometer	Audience	Inflation	Budget
Gross	1.00000000	0.24389290	0.17518906	-0.08996553	0.42362369
Tomatometer	0.24389290	1.00000000	0.69160543	-0.22403776	-0.01045194
Audience	0.17518906	0.69160543	1.00000000	-0.19518236	0.04922034
Inflation	-0.08996553	-0.22403776	-0.19518236	1.00000000	-0.01143342
Budget	0.42362369	-0.01045194	0.04922034	-0.01143342	1.00000000
Month	-0.04274638	-0.05535040	0.03615927	-0.03512832	-0.08841495
Star.Net.Worth	-0.11847343	0.03701571	0.02386478	0.06128619	-0.13179164

	Month	Star.Net.Worth
Gross	-0.04274638	-0.11847343
Tomatometer	-0.05535040	0.03701571
Audience	0.03615927	0.02386478
Inflation	-0.03512832	0.06128619
Budget	-0.08841495	-0.13179164
Month	1.00000000	-0.04720620
Star.Net.Worth	-0.04720620	1.00000000

THE FULL MODEL

```
Call:
lm(formula = Gross ~ Tomatometer + Audience + Inflation + Budget +
    Month + Star.Net.Worth, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-180.87  -85.75   -5.36   68.02  471.47

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    110.5885    93.8914   1.178   0.2419
Tomatometer    190.1122    89.2105   2.131   0.0357 *
Audience       -36.8368   128.4205  -0.287   0.7749
Inflation       -4.7490    16.2128  -0.293   0.7702
Budget           0.8540     0.1864   4.580 1.44e-05 ***
Month            0.2782     4.0316   0.069   0.9451
Star.Net.Worth  -0.1175     0.1521  -0.772   0.4419
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 123 on 93 degrees of freedom
Multiple R-squared:  0.2476, Adjusted R-squared:  0.1991
F-statistic: 5.102 on 6 and 93 DF,  p-value: 0.0001439
```

THE FULL MODEL: DIAGNOSTIC PLOTS



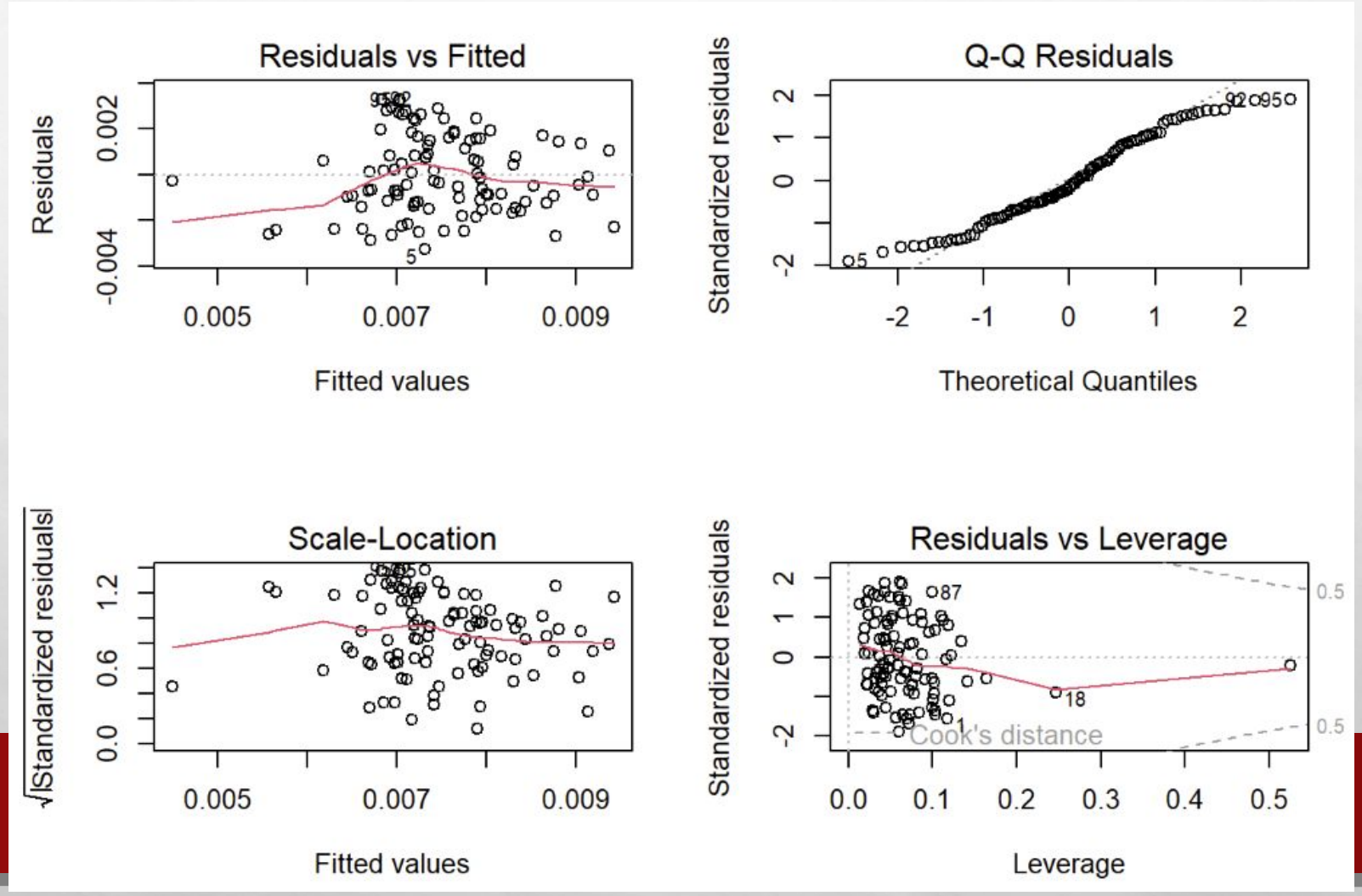
POWER TRANSFORMED MODEL

Power Transformations

- Gross^(-0.85),
- Tomatometer^(2.04),
- Audience^(-3.75),
- Inflation^(1),
- Budget^(0.84),
- Month^(0.5)
- Star.Net.Worth^(0.03)

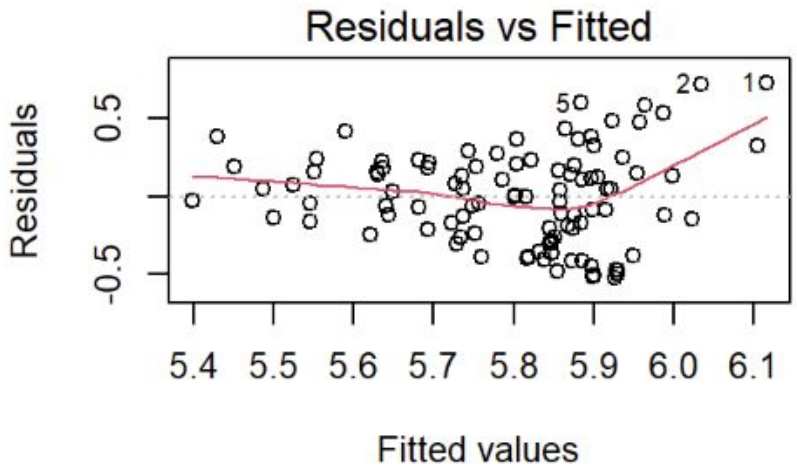
```
Call:
lm(formula = Gross ~ Tomatometer + Audience + Inflation + Budget +
    Month + Star.Net.Worth, data = movies_transformed)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0032690 -0.0012325 -0.0002581  0.0014905  0.0032843

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.257e-03  4.904e-03   0.664  0.50826
Tomatometer    -1.724e-03  7.811e-04  -2.206  0.02982 *
Audience       -4.502e-05  4.824e-05  -0.933  0.35317
Inflation       2.517e-04  2.332e-04   1.079  0.28322
Budget         -2.389e-05  7.075e-06  -3.377  0.00107 **
Month           4.688e-05  3.089e-04   0.152  0.87970
Star.Net.Worth  5.959e-03  4.166e-03   1.430  0.15594
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001775 on 93 degrees of freedom
Multiple R-squared:  0.1858, Adjusted R-squared:  0.1332
F-statistic: 3.536 on 6 and 93 DF,  p-value: 0.003381
```

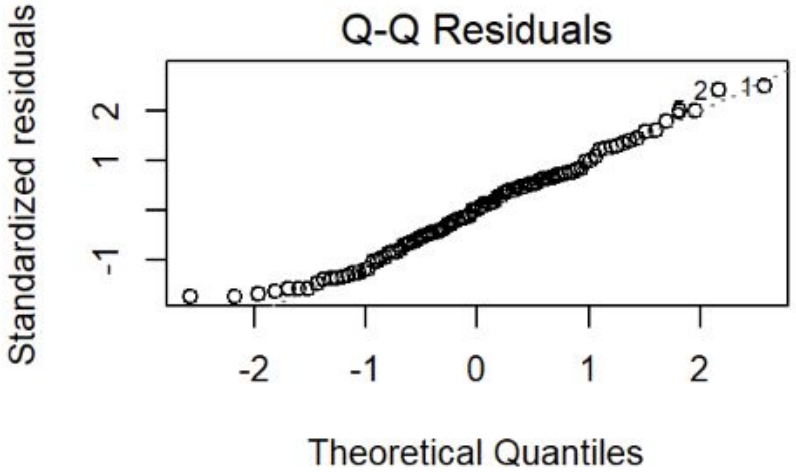
POWER TRANSFORMED MODEL: DIAGNOSTIC PLOTS

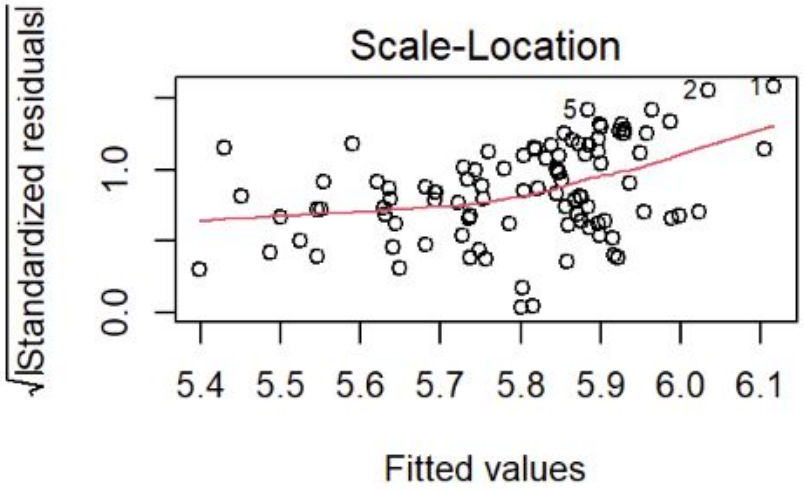


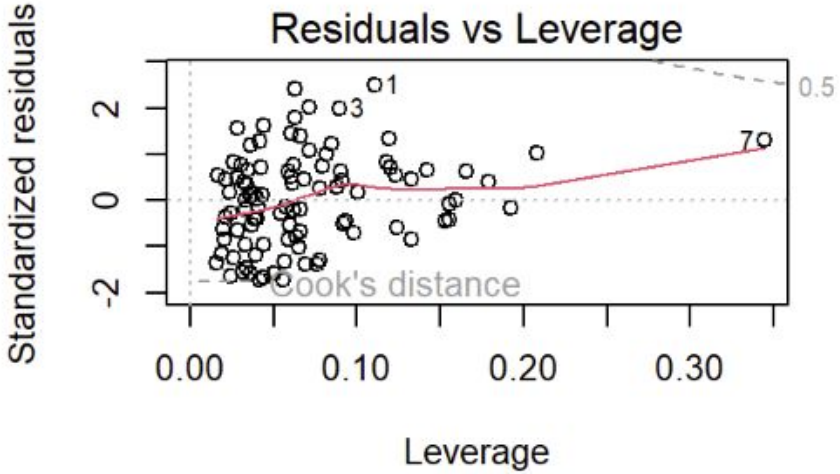
LOG MODEL

```
Call:
lm(formula = log(Gross) ~ log(Tomatometer) + log(Audience) +
    log(movies$Inflation + 1) + log(Budget) + log(Month) + log(Star.Net.Worth),
    data = movies_adjusted)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52211 -0.21421  0.00422  0.19256  0.72631

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.03009    0.37785  13.312  < 2e-16 ***
log(Tomatometer)           0.22259    0.12433   1.790 0.076654 .
log(Audience)             -0.02623    0.21328  -0.123 0.902376
log(movies$Inflation + 1) -0.07228    0.09644  -0.749 0.455499
log(Budget)                0.22419    0.06304   3.556 0.000594 ***
log(Month)                -0.03777    0.06626  -0.570 0.570019
log(Star.Net.Worth)       -0.04327    0.02383  -1.815 0.072689 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3085 on 93 degrees of freedom
Multiple R-squared:  0.1945, Adjusted R-squared:  0.1425
F-statistic: 3.743 on 6 and 93 DF,  p-value: 0.002221
```







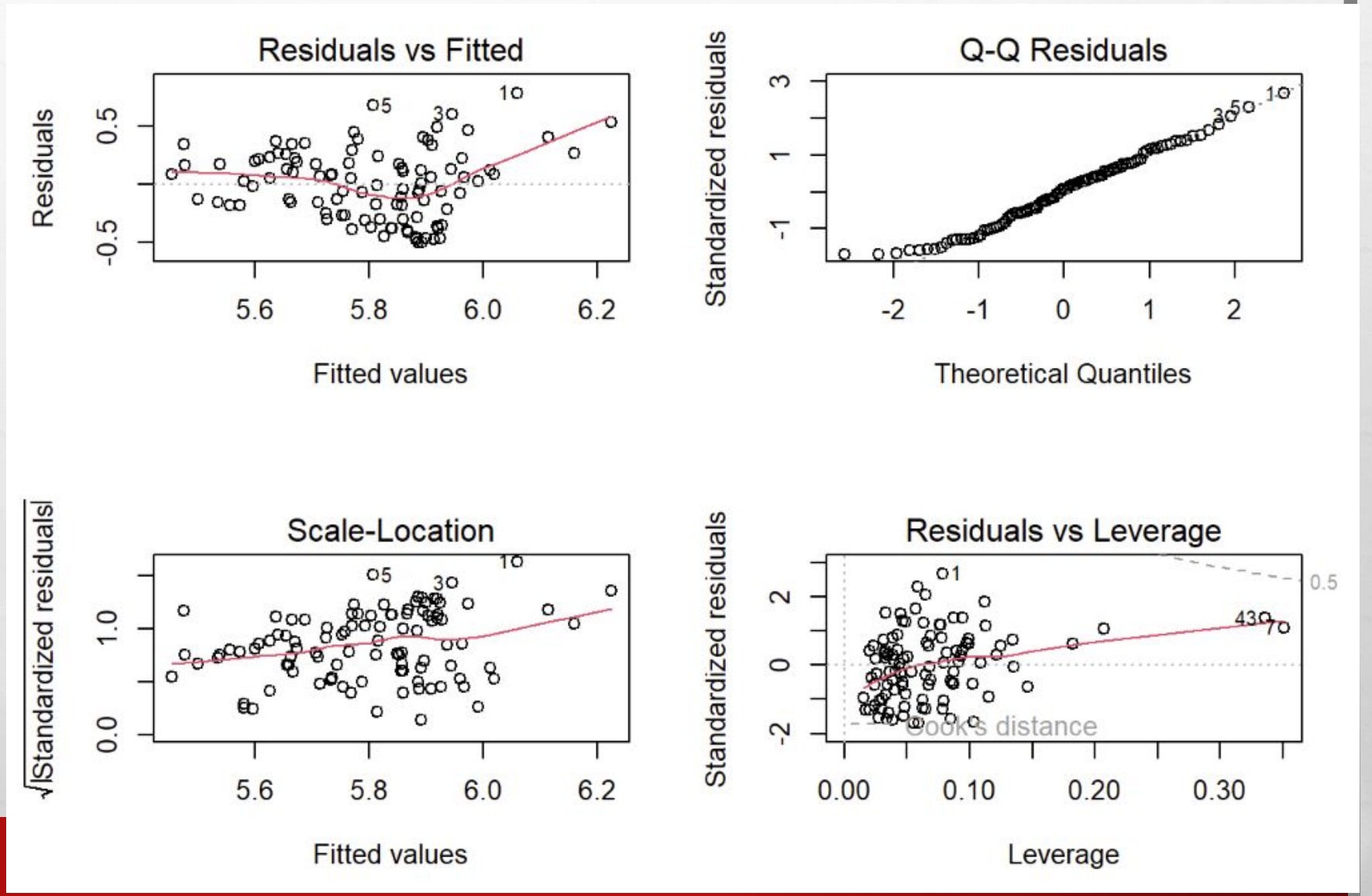


LOG Y MODEL: DIAGNOSTIC PLOTS

LOG Y MODEL

```
Call:
lm(formula = log(Gross) ~ Tomatometer + Audience + Inflation +
    Budget + Month + Star.Net.Worth, data = movies)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50384 -0.22652  0.01961  0.19930  0.78304

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.2737260  0.2337499  22.561  < 2e-16 ***
Tomatometer     0.3921748  0.2220963   1.766 0.080713 .
Audience       -0.0222226  0.3197127  -0.070 0.944734
Inflation      -0.0241508  0.0403631  -0.598 0.551067
Budget          0.0018593  0.0004642   4.006 0.000124 ***
Month           0.0001110  0.0100370   0.011 0.991199
Star.Net.Worth -0.0002244  0.0003788  -0.593 0.554919
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3061 on 93 degrees of freedom
Multiple R-squared:  0.2069, Adjusted R-squared:  0.1558
F-statistic: 4.044 on 6 and 93 DF,  p-value: 0.001204
```

LOG Y MODEL: DIAGNOSTIC PLOTS



VARIABLE SELECTION ON LOG Y MODEL

Step: Backward

```
Step:  AIC=-237.19
log(Gross) ~ Tomatometer + Budget

              Df Sum of Sq     RSS     AIC
<none>                      8.7871 -237.19
- Tomatometer  1   0.60326  9.3903 -232.55
- Budget       1   1.61920 10.4063 -222.28
```

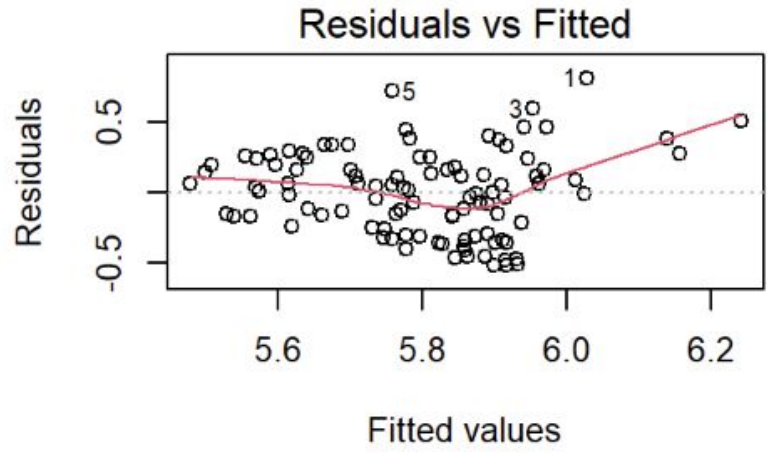
Step: Forward

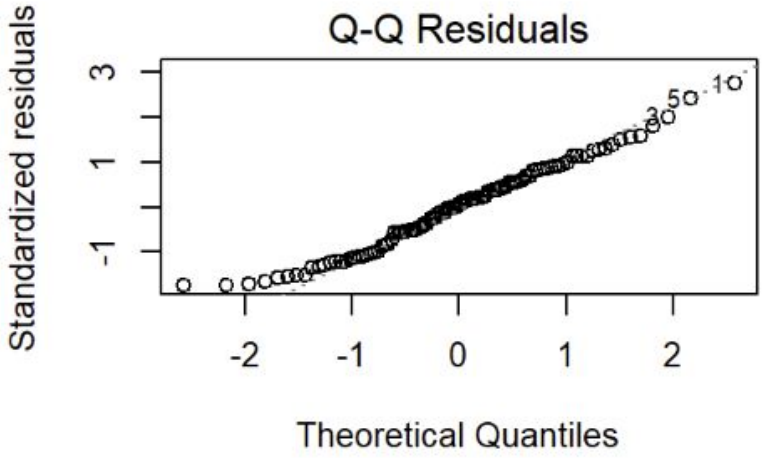
```
Step:  AIC=-237.19
log(Gross) ~ Budget + Tomatometer

                 Df Sum of Sq    RSS     AIC
<none>                       8.7871 -237.19
+ Inflation       1  0.038345 8.7487 -235.63
+ Star.Net.Worth  1  0.038193 8.7489 -235.62
+ Month           1  0.000433 8.7866 -235.19
+ Audience        1  0.000148 8.7869 -235.19
```

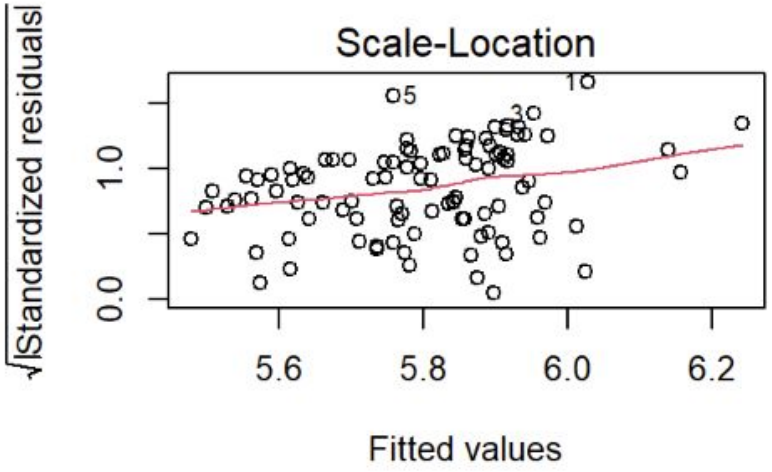
FINAL MODEL: THE REDUCED MODEL

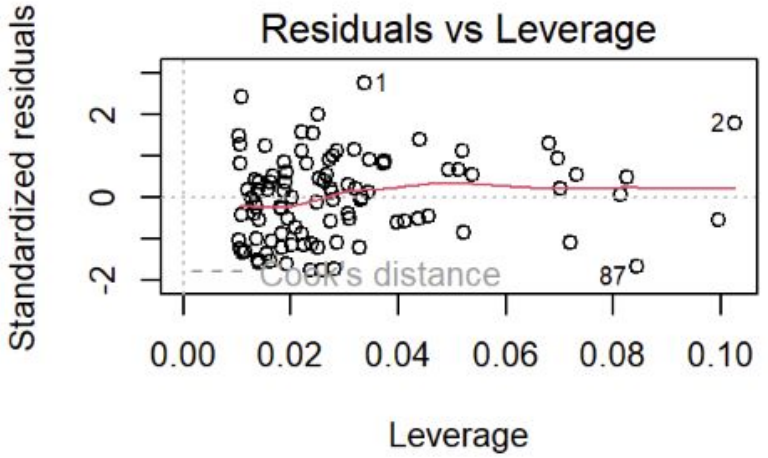
```
Call:
lm(formula = log(Gross) ~ Tomatometer + Budget, data = movies)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52124 -0.24851  0.01225  0.20980  0.81568

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.1901646  0.1410018  36.809  < 2e-16 ***
Tomatometer 0.3998904  0.1549618   2.581   0.0114 *
Budget      0.0018963  0.0004485   4.228 5.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.301 on 97 degrees of freedom
Multiple R-squared:  0.2004, Adjusted R-squared:  0.1839
F-statistic: 12.15 on 2 and 97 DF,  p-value: 1.949e-05
```

FINAL MODEL: DIAGNOSTIC PLOTS









SUMMARY AND REAL-WORLD APPLICATION

- After analyzing the potential effects of multiple variables on the domestic gross of the 100 highest grossing movies of the 2010s, we found that only critics' opinions and movie budgets were statistically significant predictors.

- Both coefficients are positive, which makes sense because as the quality of a film's content increases and the budget for a film increases, we'd expect the revenue to also increase.

SUPPORTING LITERATURE

Chang, B.-H., & Ki, E.-J. (2005). Devising a Practical Model for Predicting Theatrical Movie Success: Focusing on the Experience Good Property. *Journal of Media Economics*, *18(4)*, 247–269. https://doi.org/10.1207/s15327736me1804_2

- "…critics' evaluation…showed a significant relation with the total box office. (Chang & Ki, 2005, p. 264)."
- "Contrary to our expectations, neither brand power of actors nor directors was strong enough to affect the box office success of movies (Chang & Ki, 2005, p. 265)."
- "…budget was significant in all three models (Chang & Ki, 2005, p. 266)."

LIMITATIONS

- Relatively small sample (n = 100).
- Some non-constant variance.
- Only ~20% of the variance is explained by the model.

Potential fix: A larger sample size could improve the validity of the model by increasing constant variance.


The End