What is the effect of Walt Disney's movie budget on its box office?

Presented by: Andrew, Lotus, Benson

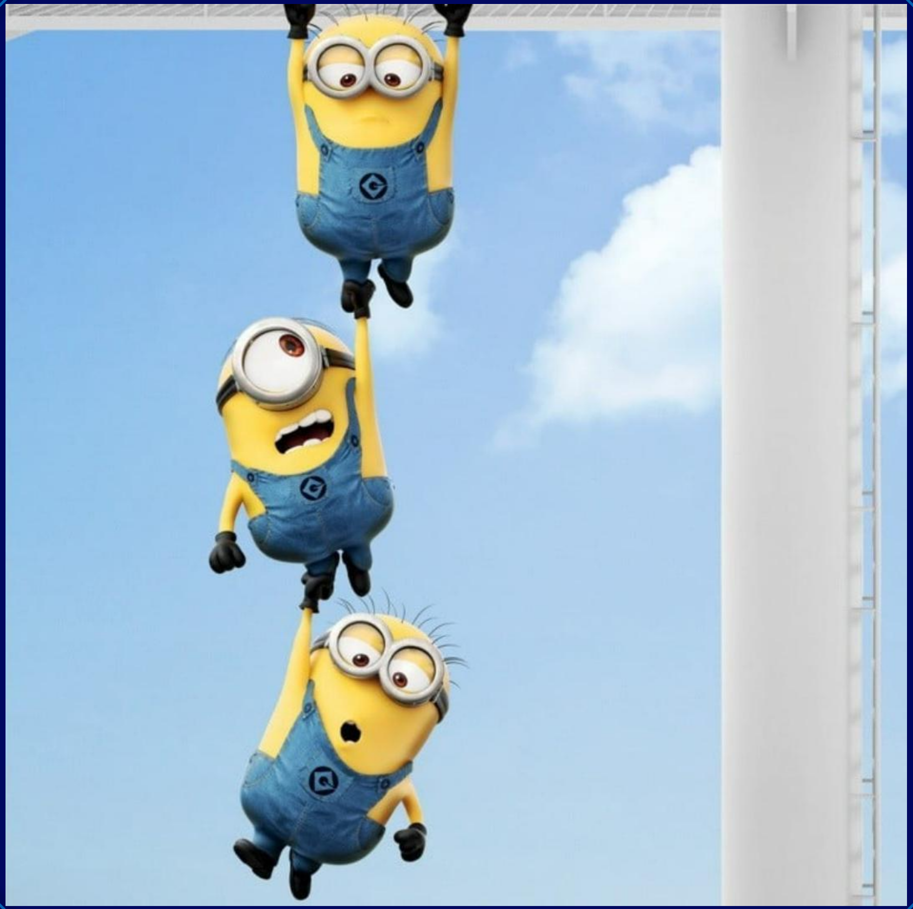
Contents

01 Descriptive Statistics

02 Baseline Regression Analysis

03 Internal and External Validity

04 Conclusion



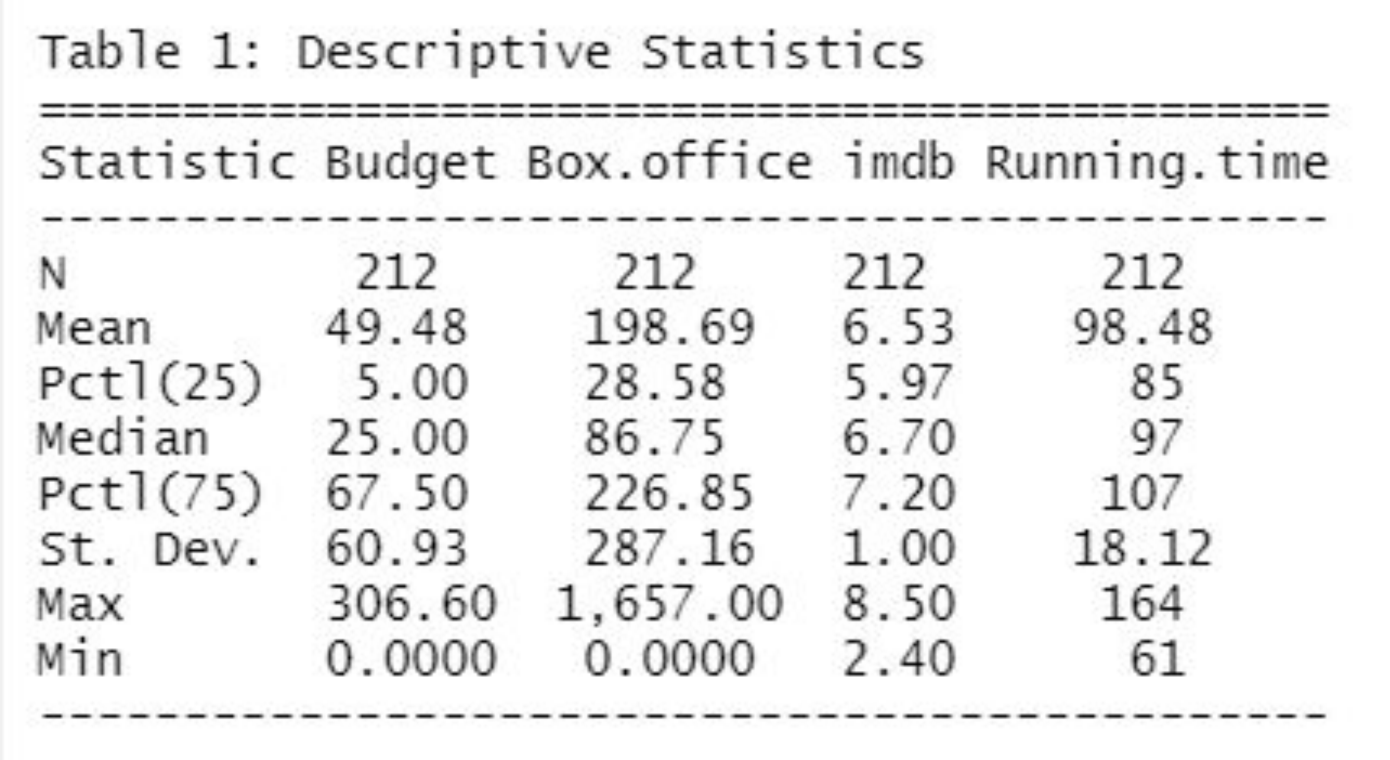
01 Descriptive Statistics

Columns overview

- **Title** : Movie Name
- **Production Company** : Production Company of the movie
- **Country** : Country of origin
- **Language** : Movie released in which language
- **Running time** : Duration of the movie (in minutes)
- **Budget** : Budget of the movie (in dollars)
- **Box Office** : Box office of the movie (in dollars)
- **Release date** : Released date (datetime)
- **imdb** : imdb ratings
- **metascore** : metascore (rating of the movie)
- **rotten_tomatoes** : Rotten Tomatoes score (quality of movie)
- **Directed by** : Director of movie
- **Produced by** : Producer of movie
- **Based on** : On which the movie is based on
- **Starring** : Main characters
- **Music by** : Music head of the movie
- **Distributed by** : buyers representing the theaters
- **Cinematography** : Director of Photography
- **Edited by** : Movie edited by
- **Screenplay by** : scripted by

01 Descriptive Statistics

Descriptive Statistics



```
Table 1: Descriptive Statistics
============================================
Statistic Budget Box.office imdb Running.time
--------------------------------------------
N           212      212      212      212
Mean       49.48    198.69    6.53    98.48
Pctl(25)    5.00     28.58    5.97      85
Median     25.00     86.75    6.70      97
Pctl(75)   67.50    226.85    7.20     107
St. Dev.   60.93    287.16    1.00    18.12
Max       306.60  1,657.00    8.50     164
Min       0.0000    0.0000    2.40      61
--------------------------------------------
```

- The median budget and box office are lower than their average, indicating their distributions are positively skewed (right-skewed).

- The median imdb rating and running time are close to the average, so their distribution should be symmetrical.

- There is significant variation in budget, box office and running time. However, the variation in imdb is relatively small.

01 Descriptive Statistics

Scatterplot

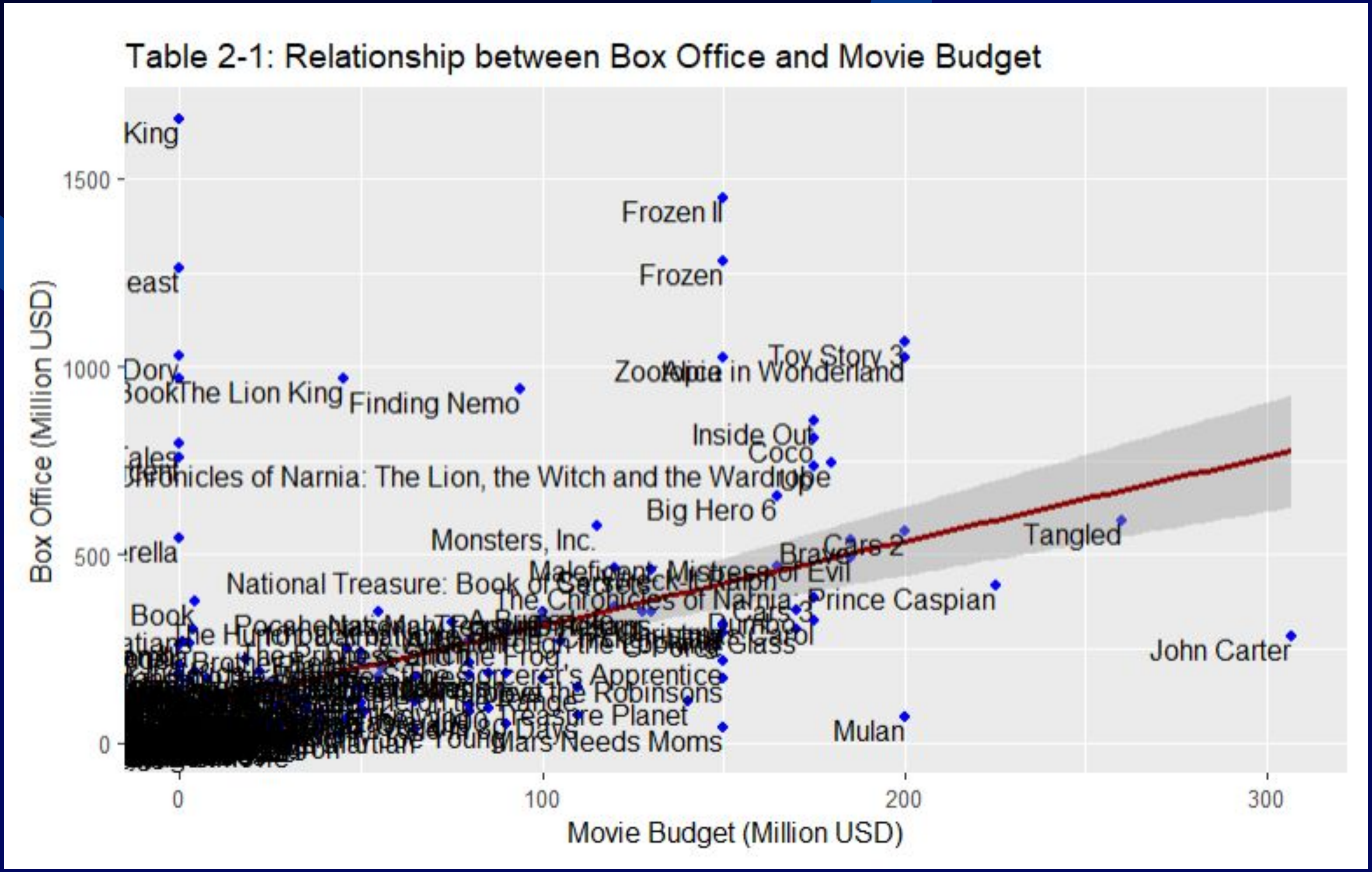


Table 2-1: Relationship between Box Office and Movie Budget

- The scatterplot suggests there is a positive correlation between box office and movie budget. In other words, when the company increased their budget, they would obtain a higher revenue.

- However, it seems like there are some outliers should be discussed later.

01 Descriptive Statistics

Scatterplot

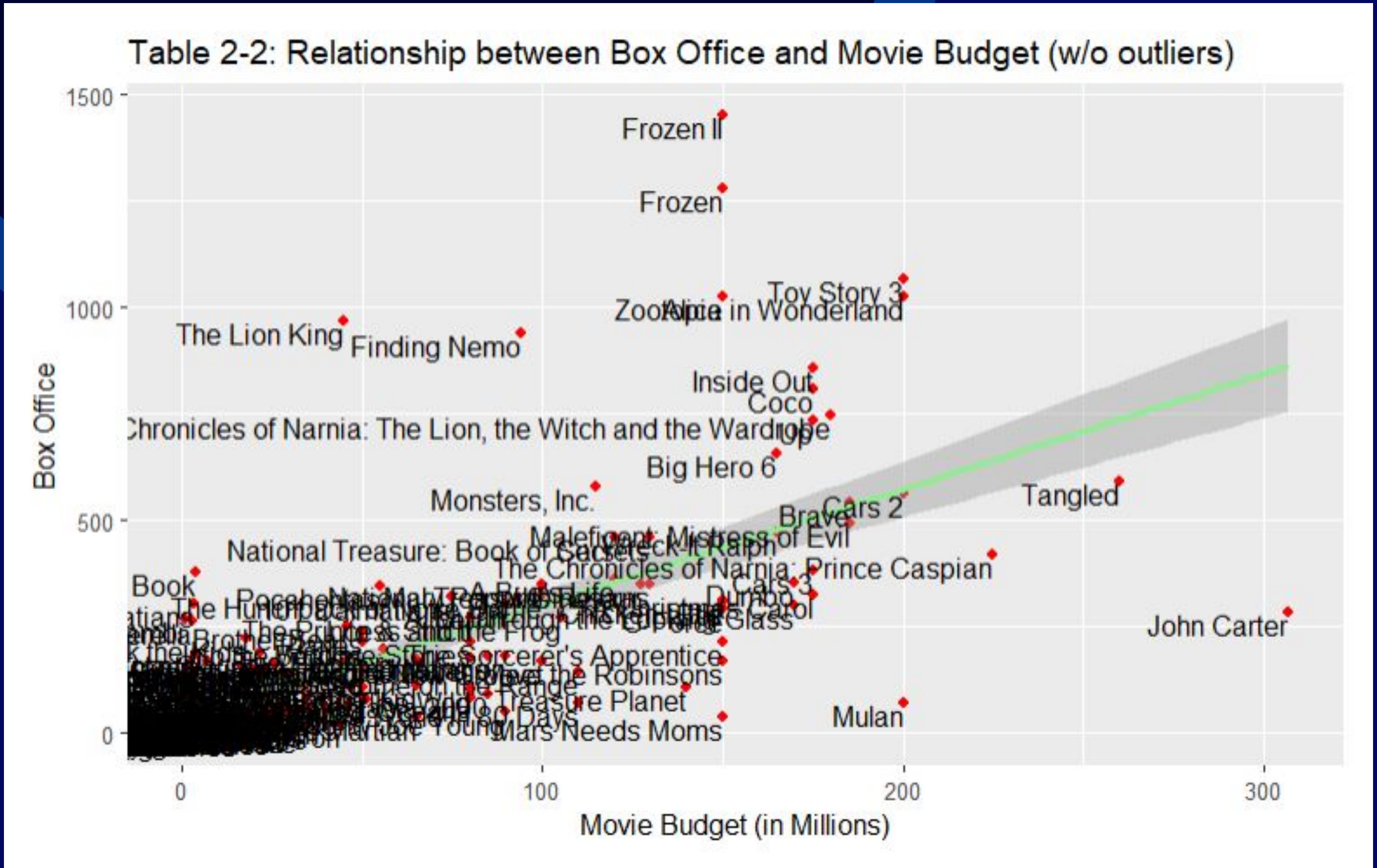
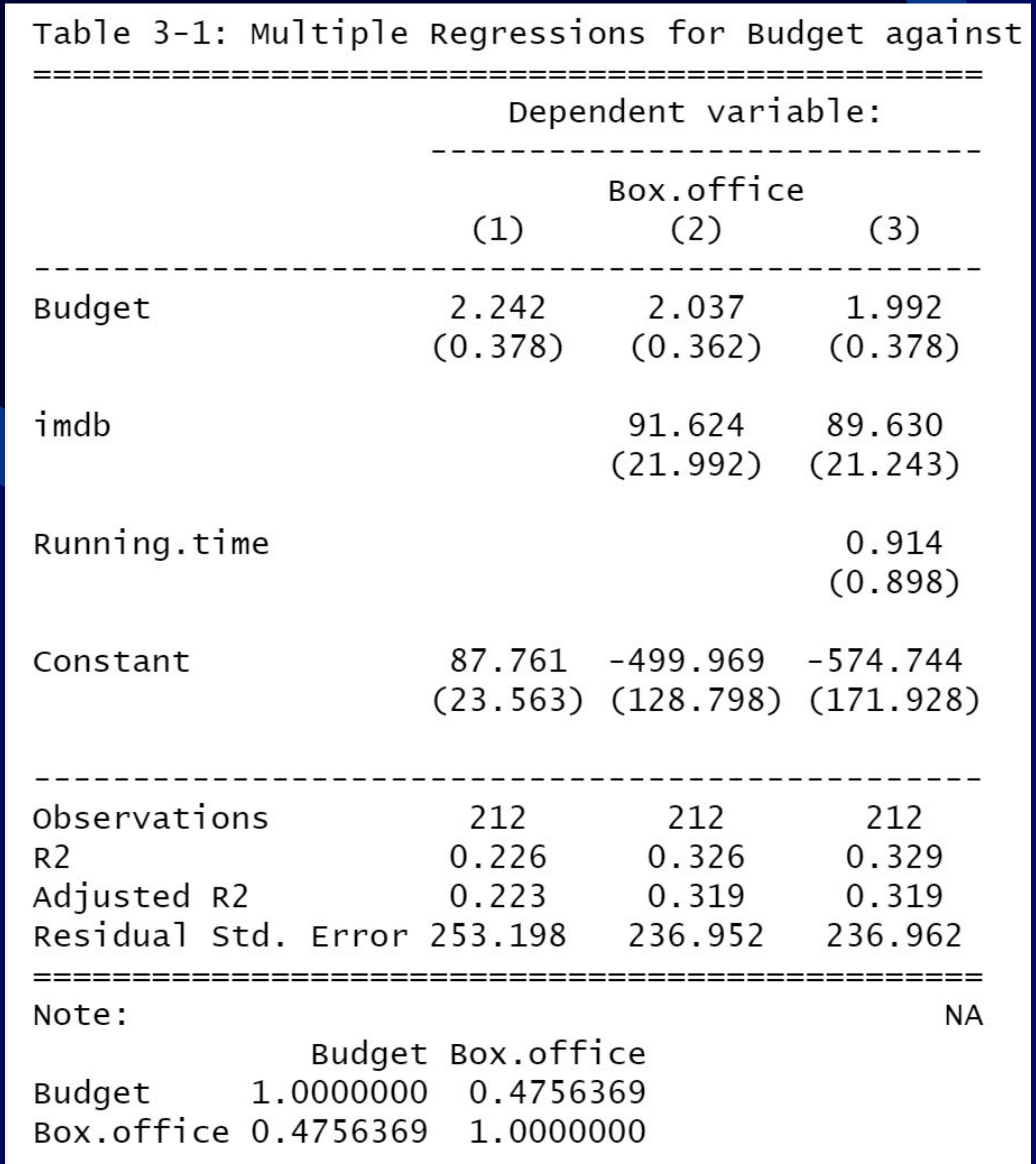


Table 2-2: Relationship between Box Office and Movie Budget (w/o outliers)

- We removed the outliers which the movie budget and box office are lower than $1000.
- The relationship between box office and movie budget is still positive. Yet, the slope of the regression line has become steeper than the previous one, indicating that an increase in movie budget would lead to a higher box office.

Multiple Regression Model



```
Table 3-1: Multiple Regressions for Budget against
================================================
                        Dependent variable:
                    ----------------------------
                              Box.office
                       (1)       (2)       (3)
------------------------------------------------
Budget                2.242     2.037     1.992
                     (0.378)   (0.362)   (0.378)

imdb                           91.624    89.630
                              (21.992)  (21.243)

Running.time                              0.914
                                         (0.898)

Constant              87.761  -499.969  -574.744
                     (23.563) (128.798) (171.928)

------------------------------------------------
Observations           212       212       212
R2                    0.226     0.326     0.329
Adjusted R2           0.223     0.319     0.319
Residual Std. Error 253.198   236.952   236.962
================================================
Note:                                         NA
              Budget Box.office
Budget     1.0000000  0.4756369
Box.office 0.4756369  1.0000000
```

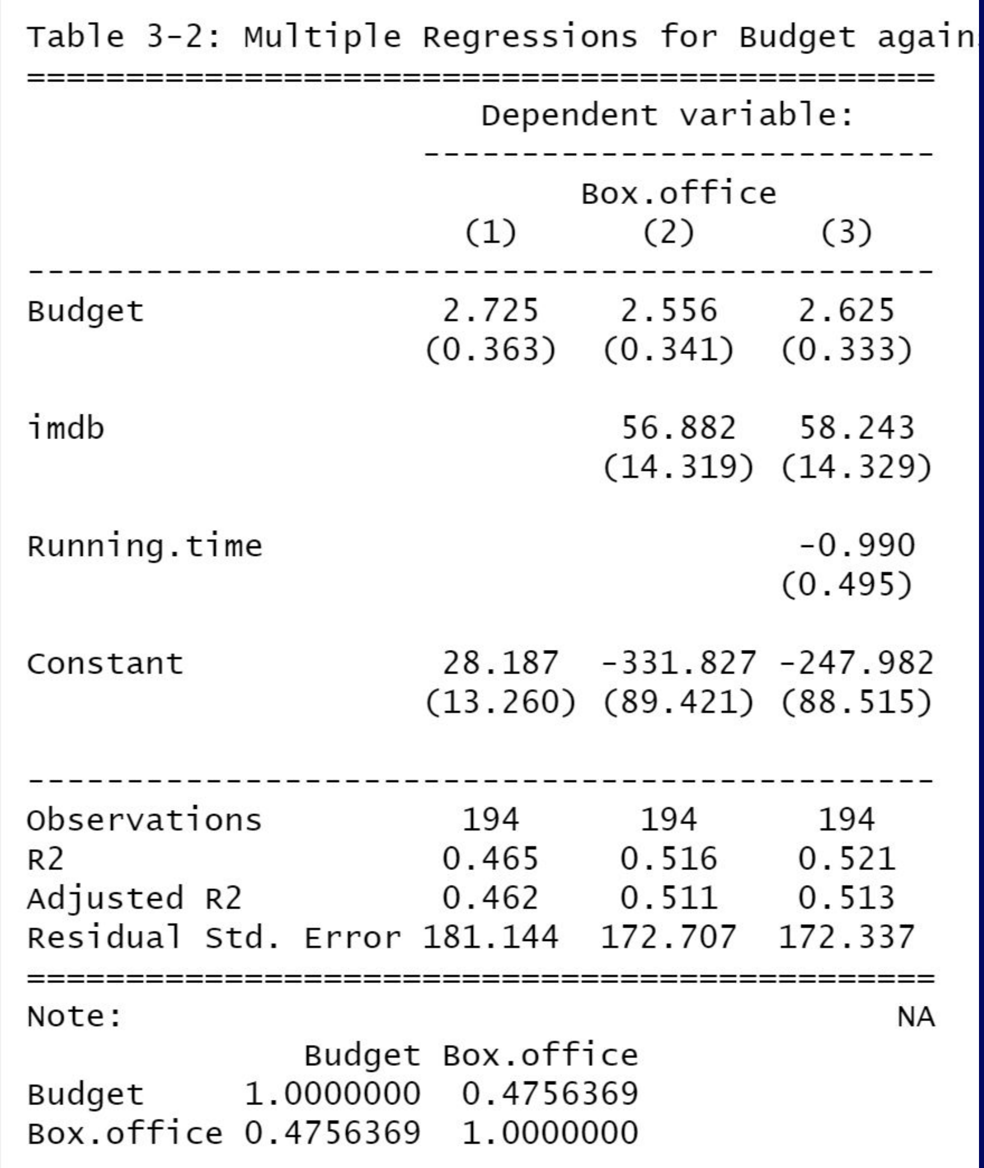
- This model includes 212 records. The independent variables include movie budget, imdb rating, and running time of the movies.

- We observed that Budget and imdb are statistically significant at 1%. However, running time is not statistically significant.

- The highest adjusted R2 among these model is 32%, which can only explain 32% of the variation.

02 Baseline Regression Analysis

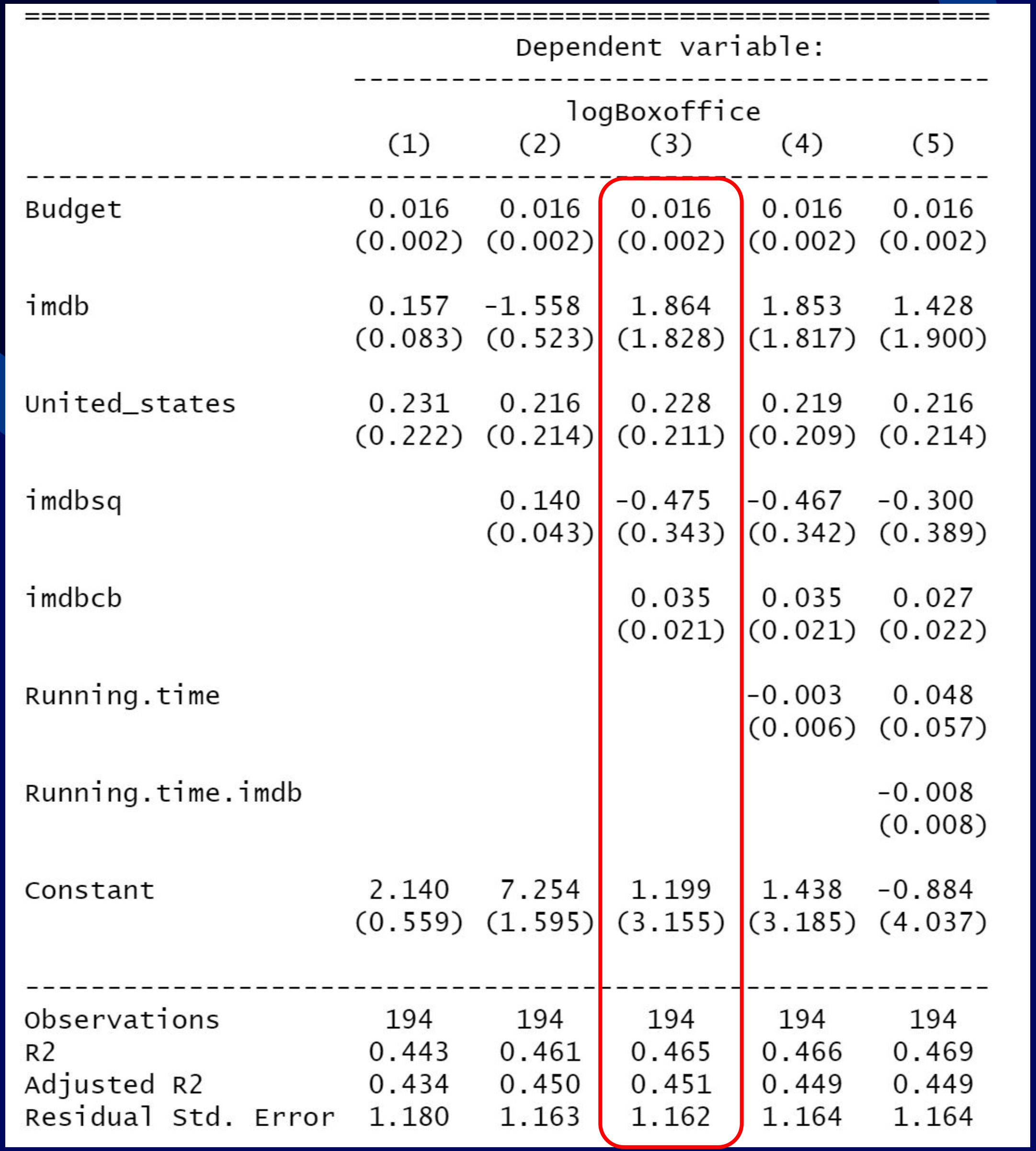
Multiple Regression Model



```
Table 3-2: Multiple Regressions for Budget again
==============================================
                    Dependent variable:
                    --------------------------
                         Box.office
                      (1)      (2)      (3)
----------------------------------------------
Budget               2.725    2.556    2.625
                    (0.363)  (0.341)  (0.333)

imdb                         56.882   58.243
                            (14.319) (14.329)

Running.time                          -0.990
                                     (0.495)

Constant            28.187  -331.827 -247.982
                   (13.260) (89.421) (88.515)

----------------------------------------------
Observations          194      194      194
R2                   0.465    0.516    0.521
Adjusted R2          0.462    0.511    0.513
Residual Std. Error 181.144  172.707  172.337
==============================================
Note:                                       NA
              Budget Box.office
Budget     1.0000000  0.4756369
Box.office 0.4756369  1.0000000
```

- After removing the outliers, the adjusted R2 has improved nearly 20%.
- Compare to model 1 and 2, budget suffers from upward bias.
  Both signs are positive as expected. The higher the budget and imdb rating, the higher the box office will get.
- In model 3, budget and imdb are statistically significant at 1%. Even though the running time is statistically significant at 5%, there is no obvious improvement in adjusted R2.

Nonlinear Regression Model



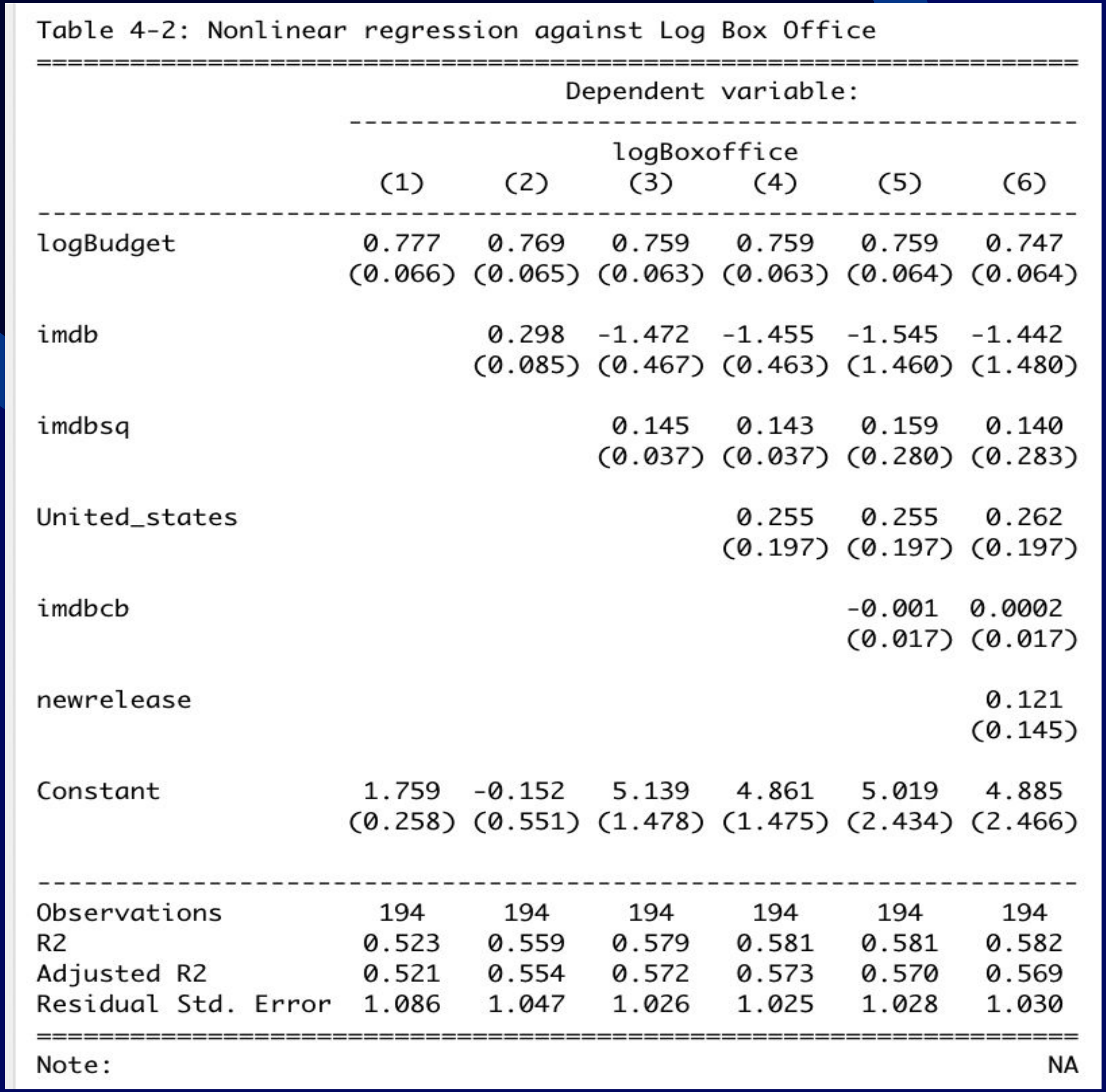
```
=============================================================
                              Dependent variable:
                    -----------------------------------------
                                   logBoxoffice
                      (1)     (2)     (3)     (4)     (5)
-------------------------------------------------------------
Budget               0.016   0.016   0.016   0.016   0.016
                    (0.002) (0.002) (0.002) (0.002) (0.002)

imdb                 0.157  -1.558   1.864   1.853   1.428
                    (0.083) (0.523) (1.828) (1.817) (1.900)

United_states        0.231   0.216   0.228   0.219   0.216
                    (0.222) (0.214) (0.211) (0.209) (0.214)

imdbsq                       0.140  -0.475  -0.467  -0.300
                            (0.043) (0.343) (0.342) (0.389)

imdbcb                               0.035   0.035   0.027
                                    (0.021) (0.021) (0.022)

Running.time                                -0.003   0.048
                                            (0.006) (0.057)

Running.time.imdb                                   -0.008
                                                    (0.008)

Constant             2.140   7.254   1.199   1.438  -0.884
                    (0.559) (1.595) (3.155) (3.185) (4.037)

-------------------------------------------------------------
Observations          194     194     194     194     194
R2                   0.443   0.461   0.465   0.466   0.469
Adjusted R2          0.434   0.450   0.451   0.449   0.449
Residual Std. Error  1.180   1.163   1.162   1.164   1.164
```

- We ran an non-linear regression model by adding imdbsq, imdbcb, United States (1 if the movies were made in US), running time, and the interaction between running time & imdb.

- Among these regression model, we concluded that model (3) is the best one since it has the highest adjusted R2 (0.45), which can explain 45% of the variation.

- According to model (3), the only estimated coefficient which is statistically significant at 1% is Budget. The others are all statistically insignificant.

Nonlinear Regression Model

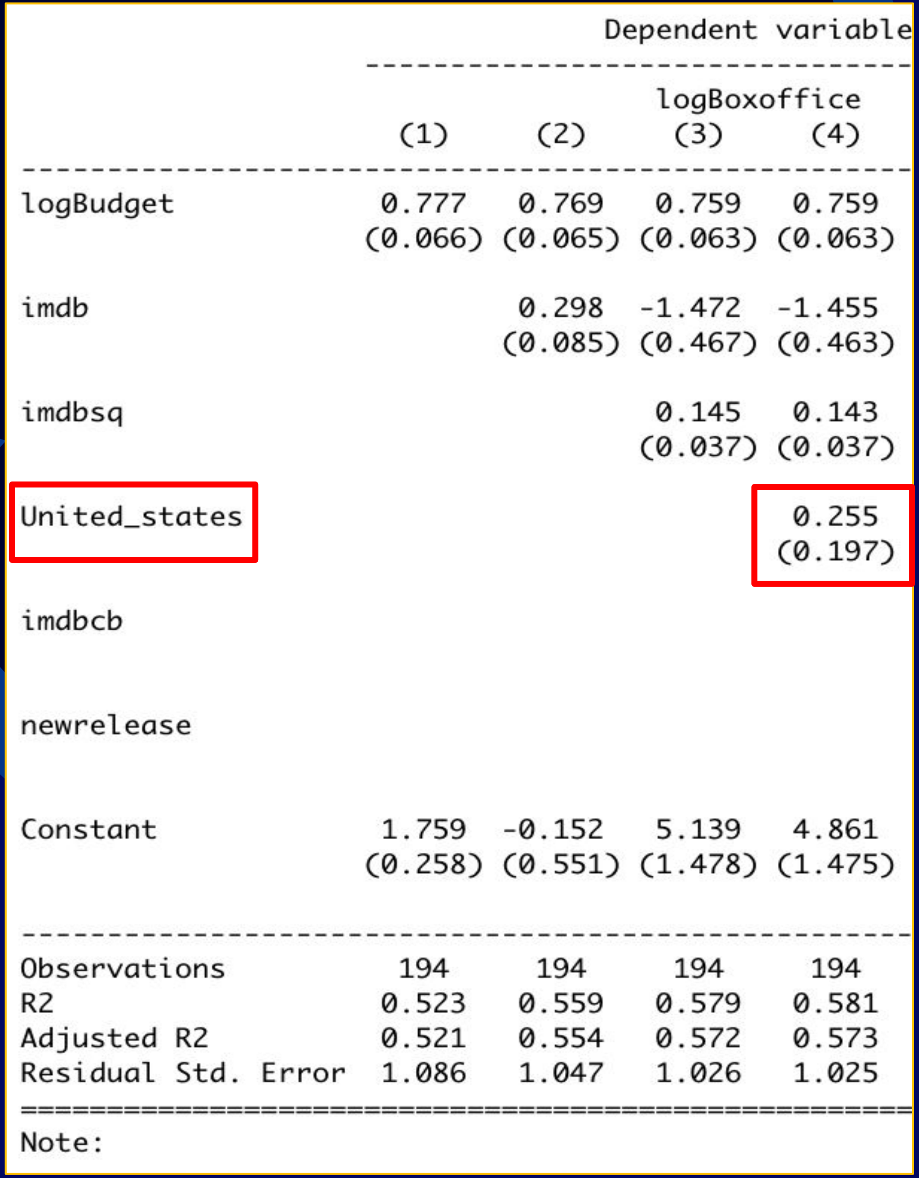


```
Table 4-2: Nonlinear regression against Log Box Office
====================================================================
                                  Dependent variable:
                    ------------------------------------------------
                                     logBoxoffice
                       (1)     (2)     (3)     (4)     (5)     (6)
--------------------------------------------------------------------
logBudget             0.777   0.769   0.759   0.759   0.759   0.747
                    (0.066) (0.065) (0.063) (0.063) (0.064) (0.064)

imdb                          0.298  -1.472  -1.455  -1.545  -1.442
                            (0.085) (0.467) (0.463) (1.460) (1.480)

imdbsq                                0.145   0.143   0.159   0.140
                                    (0.037) (0.037) (0.280) (0.283)

United_states                                 0.255   0.255   0.262
                                            (0.197) (0.197) (0.197)

imdbcb                                               -0.001  0.0002
                                                    (0.017) (0.017)

newrelease                                                    0.121
                                                            (0.145)

Constant              1.759  -0.152   5.139   4.861   5.019   4.885
                    (0.258) (0.551) (1.478) (1.475) (2.434) (2.466)

--------------------------------------------------------------------
Observations           194     194     194     194     194     194
R2                    0.523   0.559   0.579   0.581   0.581   0.582
Adjusted R2           0.521   0.554   0.572   0.573   0.570   0.569
Residual Std. Error   1.086   1.047   1.026   1.025   1.028   1.030
====================================================================
Note:                                                             NA
```

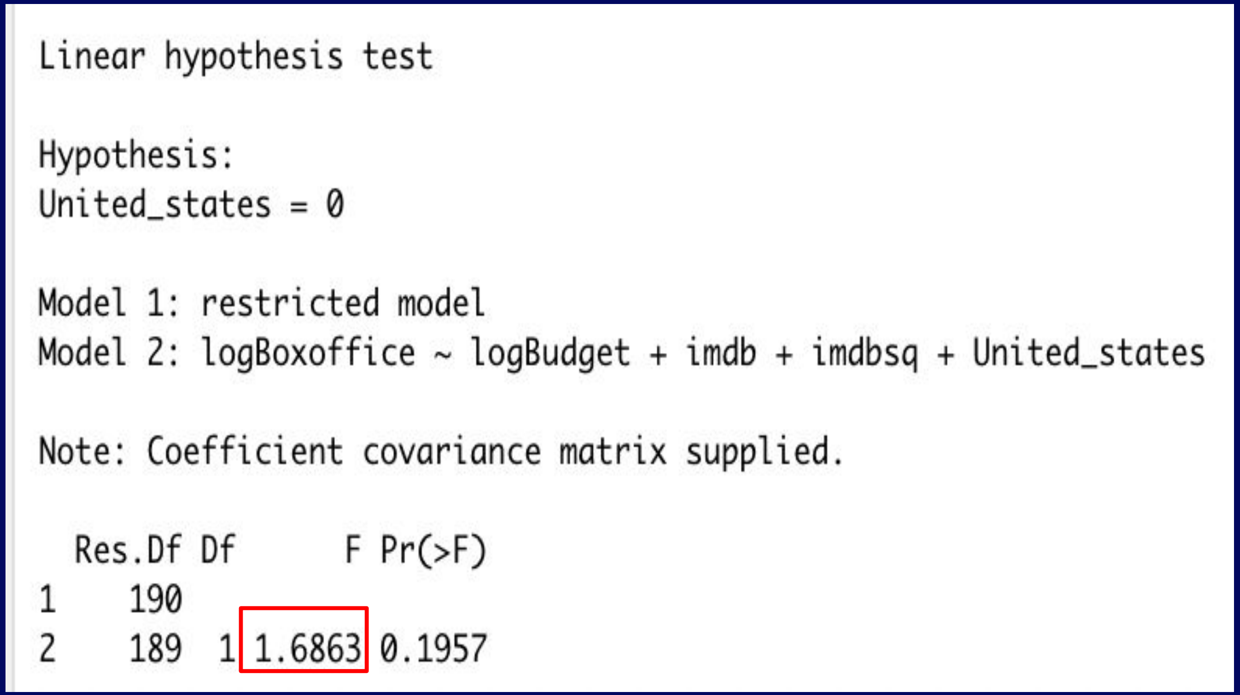
- We separated Box Office and Log Box Office to see the differences in interpretation.
- Column 1, as the log budget increases by 1%, box office revenue will increase by 0.78%.
- Column 6, as the log budget increases by 1%, box office revenue will increase by 0.75% holding everything constant.
- Column 4 seems to be the best fit model as the adjusted R2 is at .573. This model regression represents 57% of the variation in log box office. The spread of the residuals around the fitted line is 1.025 million dollars.

02 Baseline Regression Analysis

Measure of fit (F-test)



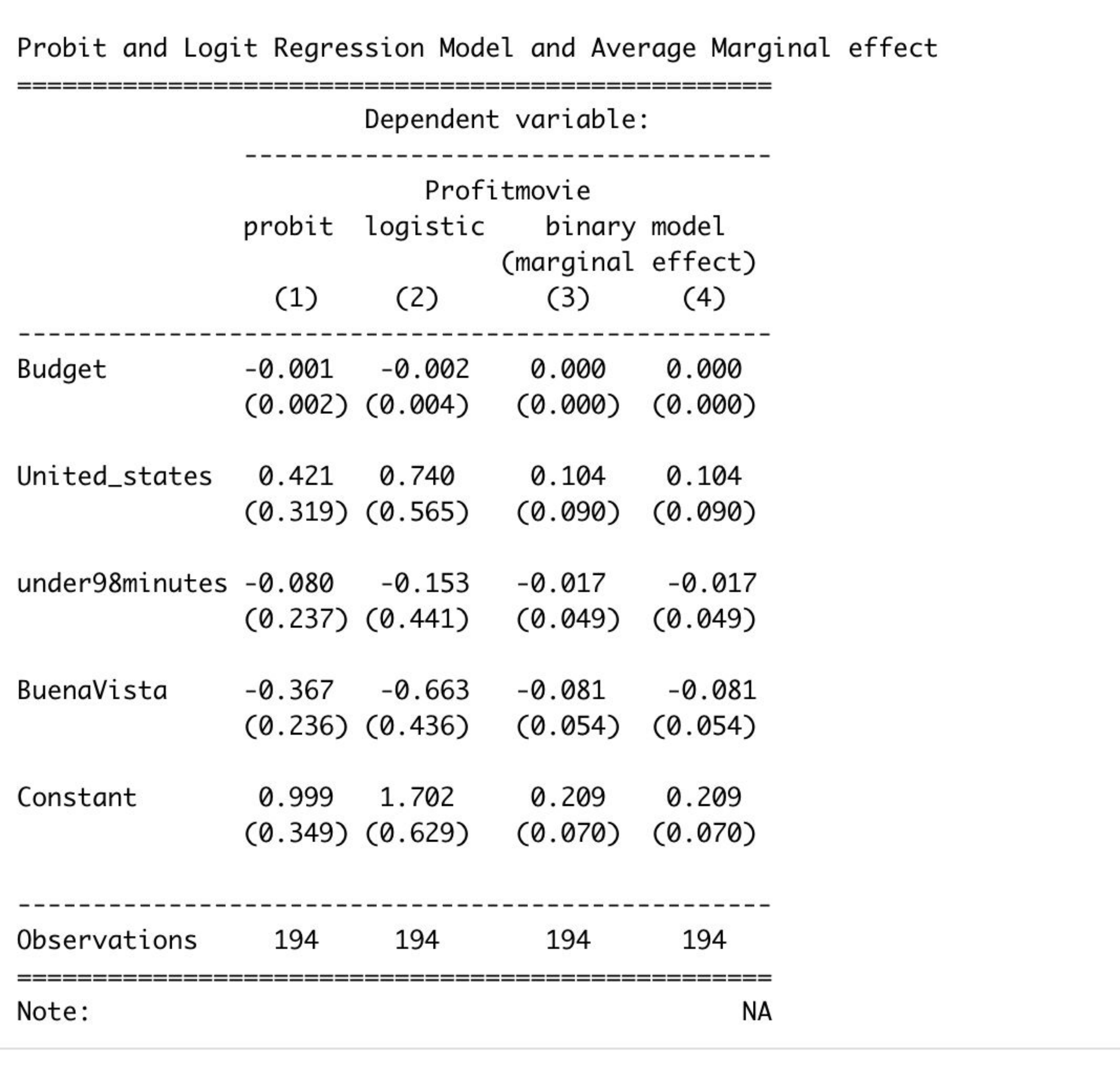
```
                           Dependent variable
                     ----------------------------------
                                logBoxoffice
                       (1)     (2)     (3)     (4)
-----------------------------------------------------
logBudget             0.777   0.769   0.759   0.759
                     (0.066) (0.065) (0.063) (0.063)

imdb                          0.298  -1.472  -1.455
                             (0.085) (0.467) (0.463)

imdbsq                                0.145   0.143
                                     (0.037) (0.037)

United_states                                 0.255
                                             (0.197)

imdbcb


newrelease


Constant              1.759  -0.152   5.139   4.861
                     (0.258) (0.551) (1.478) (1.475)

-----------------------------------------------------
Observations           194     194     194     194
R2                    0.523   0.559   0.579   0.581
Adjusted R2           0.521   0.554   0.572   0.573
Residual Std. Error   1.086   1.047   1.026   1.025
=====================================================
Note:
```



```
Linear hypothesis test

Hypothesis:
United_states = 0

Model 1: restricted model
Model 2: logBoxoffice ~ logBudget + imdb + imdbsq + United_states

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1    190
2    189  1 1.6863 0.1957
```

- From Regression 4, we use F test to see if we need to have United States Dummy Variable into our multiple regression model. F Statistic is 1.686 so we cannot reject the null hypothesis.
- We conclude that using Column 3 is the best regression model.

Probit and Logit Regression



```
Probit and Logit Regression Model and Average Marginal effect
==================================================
                       Dependent variable:
               -----------------------------------
                           Profitmovie
               probit  logistic    binary model
                                (marginal effect)
                 (1)     (2)       (3)      (4)
--------------------------------------------------
Budget         -0.001   -0.002    0.000    0.000
               (0.002) (0.004)   (0.000)  (0.000)

United_states   0.421   0.740     0.104    0.104
               (0.319) (0.565)   (0.090)  (0.090)

under98minutes -0.080   -0.153   -0.017    -0.017
               (0.237) (0.441)   (0.049)  (0.049)

BuenaVista     -0.367   -0.663   -0.081    -0.081
               (0.236) (0.436)   (0.054)  (0.054)

Constant        0.999   1.702     0.209    0.209
               (0.349) (0.629)   (0.070)  (0.070)

--------------------------------------------------
Observations     194     194       194      194
==================================================
Note:                                           NA
```

- We use average marginal effect of Budget, United_states, BuenaVista on probability of making money(profit).

- In the regression models, we found that these factor do not affect the probability of being profitable for a movie.

03 Internal and External Validity

Internal and External Validity

Internal Factors

- Budget might suffer from omitted variables, all these factors listed below might affect the box office as well.
- Voice Actor
- Social Media Advertising
- Cultural or Economic trends
- Movie Theme or message to audience
- Genre
- Live Action or Cartoon Animation
- Seasonality
- Remakes of original movies (ex: Mulan, Lion King, etc.)

External Factors

- Data sample is represented by most united states made movies.
- The sample does not include the data in 2021 and 2022.

04 Conclusion

$$\log(BoxOffice) = 5.14 + 0.76*\log(Budget) + 0.145*imdb^2 - 1.472*imdb$$

Through linear/nonlinear regression model building, hypothesis testing and data analysis, our team found that budget has a causal effect on box office. The higher the budget for a movie, the more box office it can get.

Imdb has positive correlation on box office.

Other factors,such as Running time, where the movie was made and whether the movie was released before or after 2010, whether the movie was distributed by Buenavista, do not have much effect on box office.

THANKS !

We appreciate your comments and suggestions!