Cinema Revenue Analysis

Megan Solly

2025-12-29

Introduction

The cinema industry involves substantial financial risks, with production budgets often being millions of dollars. Studio executives are tasked with deciding which projects to fund while facing considerable uncertainty about future revenue. The primary objective of this analysis is to identity which aspects of a film are most strongly associated with gross revenue. Using data on approximately 100 films from 30 directors, this study develops a model to predict movie revenue.

Exploratory Data Analysis

The dataset contains films with associated information including budget, gross revenue, runtime, genre, director birth year, production year, and ratings. The original dataset contained multiple genres for each movie, so as a part of data cleaning, I transformed the genres column into a set of binary indicator variables. Each column represents whether a movie belongs to that genre (1) or not (0). For example, if a movie was classified as adventure and comedy then is_Adventure = 1, is_Comedy = 1, and all other genre columns are set to 0. This resulted in a dataset that was easier to evaluate the effects of each genre on revenue. Additionally, I adjusted the monetary values for inflation by setting them to 2024 dollars. Finally, the dataset included director name and birth years, so I computed each director's age at the time of film production, which is informative because it gives insights

into the stage of their career and possibly their amount of experience. Figure 1 depicts the distribution of the director age at the time of production.

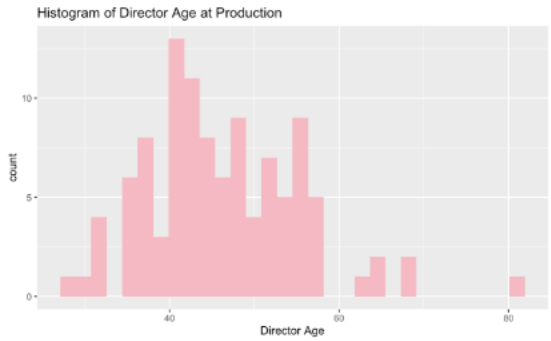
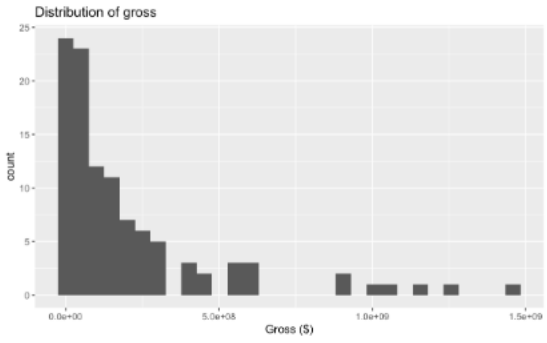


Figure 1: Distribution of the Producer's Age at the Time of Production

Before modeling, I conducted exploratory data analysis to understand the distributions of variables and identify if any transformations were necessary. From this, I found that gross (Figure 2) and budget (Figure 3) were strongly right-skewed, with some high-earning films stretching the upper tails. I also examined the residuals versus fitted values plot and there was clear heteroscedasticity, with the spread of residuals increasing for larger fitted values.



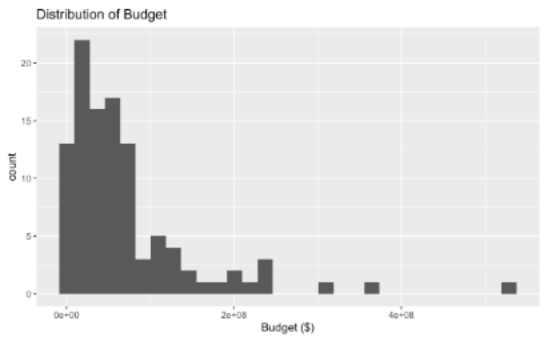


Figure 2: Distribution of the Gross Revenue

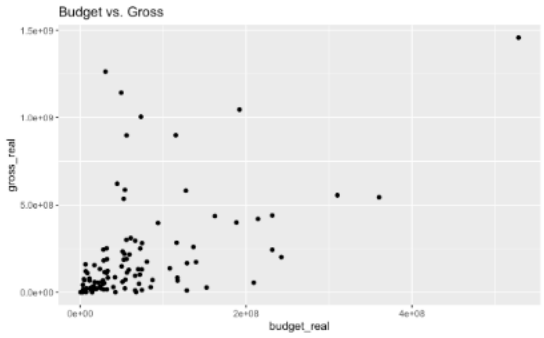
Figure 3: Distribution of the Budget

These patterns indicate that a transformation of the response variable may help stabilize the variance and improve the model. The observed skewness and heteroscedasticity suggest that a log transformation on both gross and budget would be appropriate. I confirmed this by

plotting both log-transformed variables, which demonstrated a more symmetric distribution. I also plotted the residuals and fitted values with the log transformed variables and the spread appeared roughly constant, indicating that the transformed data fixed the heteroscedasticity. Other continuous predictors, runtime and movie rating, were approximately symmetric and did not require transformation.

The most important bivariate distribution was between budget and gross. On the original scale, the scatterplot showed a nonlinear pattern (Figure 4). However, after transforming both variables, the relationship was linear which justified the transformations (Figure 5).

Gross revenue also displayed correlations with average rating and genre, suggesting that film quality and genre may play significant roles in explaining gross revenue differences between films.



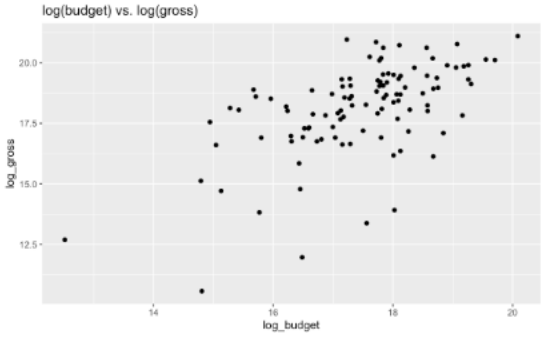


Figure 4: Budget vs. Gross (raw)

Figure 5: Budget vs. Gross (log-transformed)

Model Selection

To select an appropriate model, I utilized several model selection techniques including backward, forward, and stepwise AIC, and exhaustive search via regsubsets. The backward, forward, and stepwise AIC consistently selected the model that used log-transformed budget, average rating, director age at production, and indicator variables for the drama and fantasy genres as predictors of log-transformed gross revenue. The exhaustive search also

ranked the same model among the top models by AIC and BIC, further supporting its selection. Ultimately, all selection procedures pointed to the same model, so it was chosen as the final model. These results indicated that log(budget) was the strongest single predictor, while average rating, genre (fantasy and drama), director age at production, and production year also provided meaningful explanatory power.

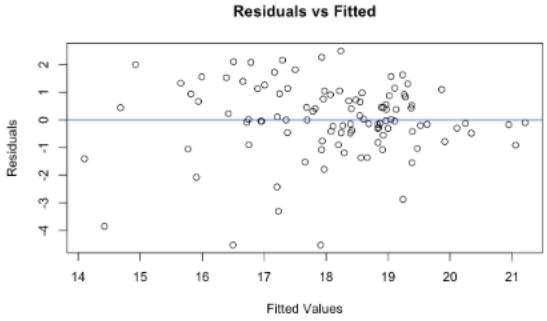
`The final model is:`

$$log(gross) = B_0 + B_1 log(budget) + B_2 movie\_averageRating + B_3 is\_Drama + B_4 is\_Fantasy + B_5 director\_age\_at\_prod + B_6 production\_year + e \text{ where } e \sim N(0, \sigma^2)$$

The assumptions include linearity between each predictor and the log of gross revenue, independence of residuals, constant variance of residuals, and normality of residuals.

Model Fit and Diagnostics

Model diagnostics were examined to ensure the assumptions of the regression model were satisfied. The residual versus fitted plot (Figure 6) showed no strong curvature, suggesting linearity of the log transformed variables is appropriate. Additionally, the normal QQ plot (Figure 7) demonstrated that the residuals are approximately normally distributed. Finally, the plot of residuals versus log-transformed budget, average movie rating, director age at production, and production year showed no fanning or patterns, suggesting the residuals have constant variance. Overall, these diagnostic plots demonstrate that the final model adequately fits the data and that the key regression assumptions are satisfied.



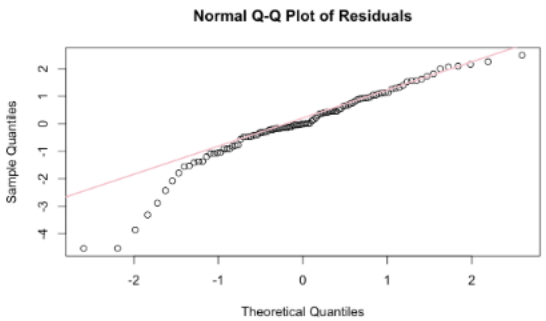


Figure 6: Residual vs Fitted Plot

Figure 7: QQ Plot

Interpretation and Analysis

Interpretation of the results from the regression model provide insights into the factors associated with a film's gross revenue. All monetary values throughout this analysis are in 2024 dollars due to the adjustment for inflation. The results of the statistical analysis show that the coefficient on log(budget) is 0.7219, meaning that a 1% increase in budget is associated with a 0.72% increase in expected gross revenue when holding other variables constant. This indicates that higher budgets are associated with higher revenues. The average rating coefficient is 0.5226, because the model is in log form, the percent change is given by $e^{0.5226}$˘1. This evaluates to around 68% which means that for each 1-unit increase in the average rating, there is approximately a 68% increase in expected revenue when controlling for other variables.

Genre effects are measured using binary indicator variables, with the reference category being action films. The reference was established by excluding the is_Action indicator from the model, so all genre coefficient represent differences relative to action movies. Therefore, each genre coefficient measures the expected difference in log-transformed gross revenue between that genre and an action film, holding other variables constant. The coefficient for drama is -0.9455, indicating that drama films, on average, earn $e^{-.9455} - 1$ = -68% less gross revenue than action films. In contrast, fantasy films, on average, earn $e^{1.0344} - 1 = 181\%$ more gross than action films. This demonstrates that genre has a significant role in predicting gross revenue, with fantasy films significantly outperforming action movies, and drama films significantly underperforming.

The coefficient for director age at the time of production is -0.0322, indicating that each additional year of a director's age is associated with an expected 3.2% decrease in gross revenue, holding other variables constant. This suggests that films directed by younger

directors tend to generate higher revenues, though this effect may reflect unobserved factors such as genre specialization.

Finally the coefficient for production year is -0.0188, implying that for each additional year of production, holding other factors constant, expected revenue decreases by around 1.9%.

A central question for a studio executive is whether increasing a film's budget is financially justified. In the final model, the coefficient for log(budget) was 0.7219, meaning that a 10% increase in budget is associated with a 7.2% increase in expected gross revenue when holding other factors constant. This suggests that larger budgets do tend to generate higher revenues, but not at 1:1 rate. While a larger budget generally pays off, the returns diminish as budgets grow. However, it is important to recognize that this estimate may be biased due to unobserved factors such as marketing, actors, franchise popularity, and many other possible factors.

Using the final regression model, I predicted the expected gross revenue for a hypothetical film released on June 15, 2013, with an average rating of 7.0, and a budget of $10 million. The final model includes log-transformed budget, average movie rating, genre indicators for fantasy and drama, director age at the time of production, and production year so the prediction is based on those variables. The film is a comedy, so both indicator variables were set to 0. For the director, I used the first director alphabetically in the dataset and computed their age at the time of production. The following monetary predictions are in 2024 dollars, the model expects that the film will earn approximately $49,937,901 in revenue. The 95% prediction interval ranges from about $2.96 million to $744 million which reflects the variability in film performance and large uncertainty in predicting individual movie revenues.

```
library(tidyverse)
library(readr)
library(leaps)
library(dplyr)
```

```r
library(priceR)
movies = read.csv("~/Downloads/moviedata.csv")
```

```r
# clean dataset
movies_clean = movies %>%
  mutate(
  budget = as.numeric(budget),
  gross = as.numeric(gross),
  production_date = mdy(production_date),
  production_year = year(production_date),
  production_year = ifelse(production_year > 2025, production_year - 100,
                           production_year),
    budget_real = adjust_for_inflation(
      budget,
      from_date = production_year,
      to_date = 2024,
      country = "US"
    ),
    gross_real = adjust_for_inflation(
      gross,
      from_date = production_year,
      to_date = 2024,
      country = "US"
    )
  )
```

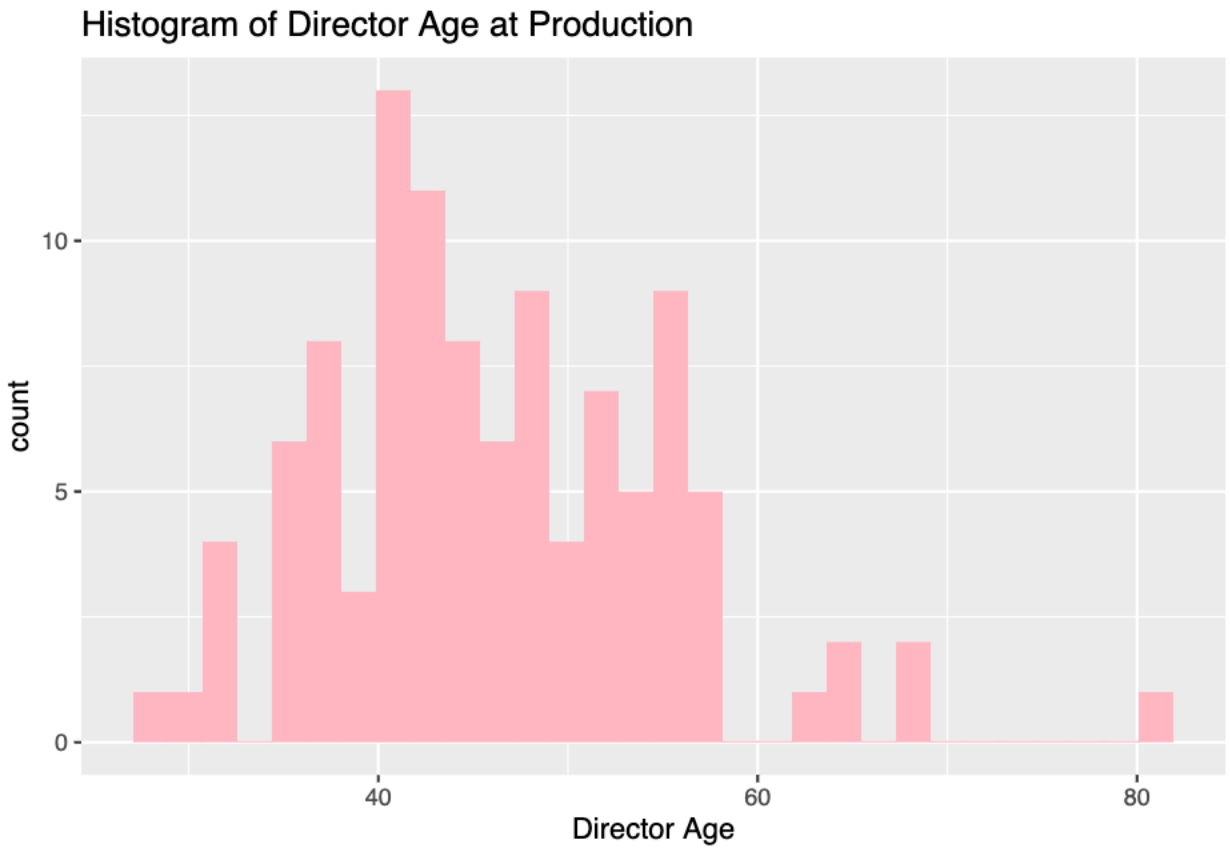
```
## Retrieving countries data
```

```
## Generating URL to request all 296 results
## Retrieving inflation data for US
## Generating URL to request all 65 results

## Retrieving countries data

## Generating URL to request all 296 results
## Retrieving inflation data for US
## Generating URL to request all 65 results
```

```r
# modify genre to create indicator variables
# creates a vector to hold all unique genres
all_genres = movies_clean %>% mutate(genre_list = str_split
  (genres, ",\\s*")) %>% unnest(genre_list) %>% distinct(genre_list) %>%
  pull(genre_list)

movies_clean = movies_clean %>%
  mutate(genre_list = str_split(genres, ",\\s*"))

# creates column names for each unique genre
for (g in all_genres) {
  colname = paste0("is_", g)
  movies_clean[[paste0("is_", g)]] = sapply(movies_clean$genre_list,
  function(x) g %in% x)
}

movies_clean = movies_clean %>% select(-genre_list)
```

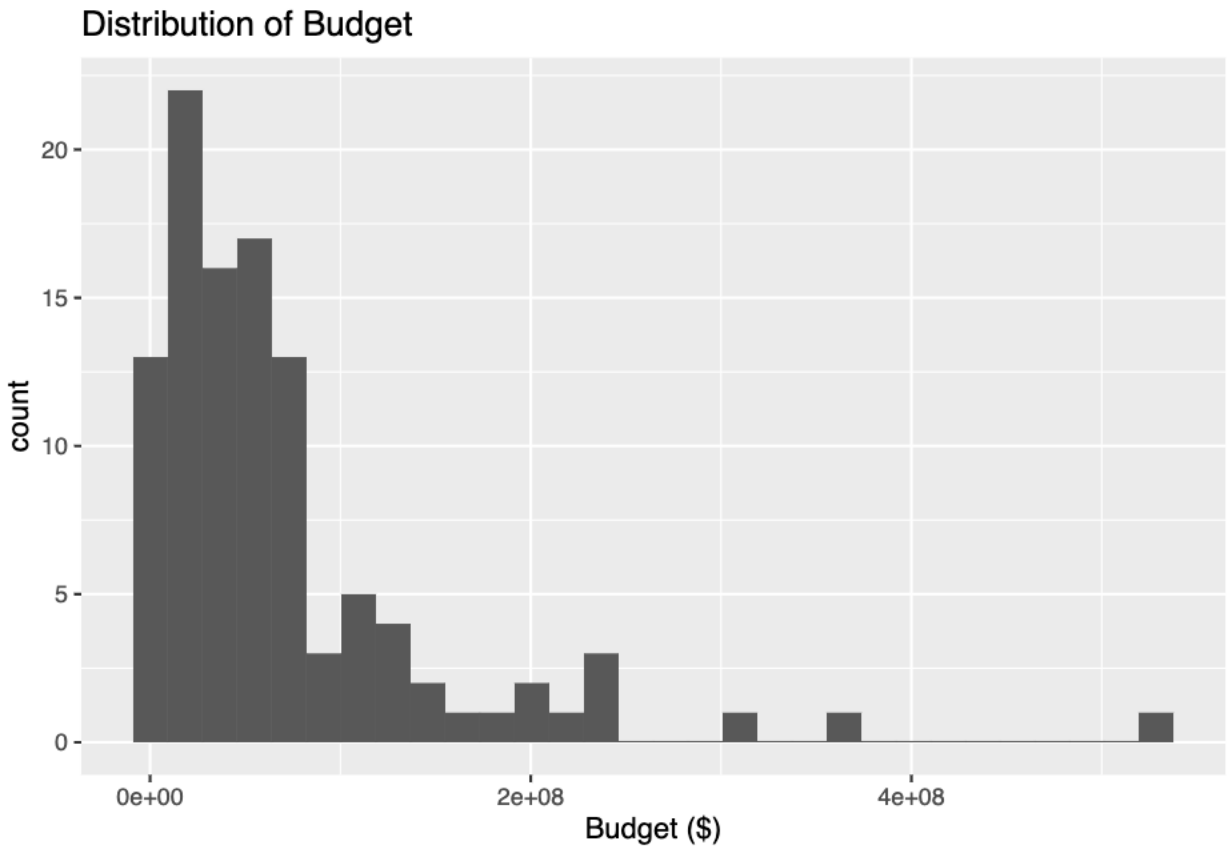
```r
# cleans the sci fi column name
if ("is_Sci-Fi" %in% names(movies_clean)) {
  names(movies_clean)[names(movies_clean) == "is_Sci-Fi"] = "is_Sci"
}
```

```r
# create director age at production variable
movies_clean = movies_clean %>%
  mutate(director_age_at_prod = production_year - birth_year)

# plots distribution of director age at production
ggplot(movies_clean, aes(x = director_age_at_prod)) +
  geom_histogram (bins = 30, fill = "lightpink") +
  labs(title = "Histogram of Director Age at Production", x = "Director Age")
```



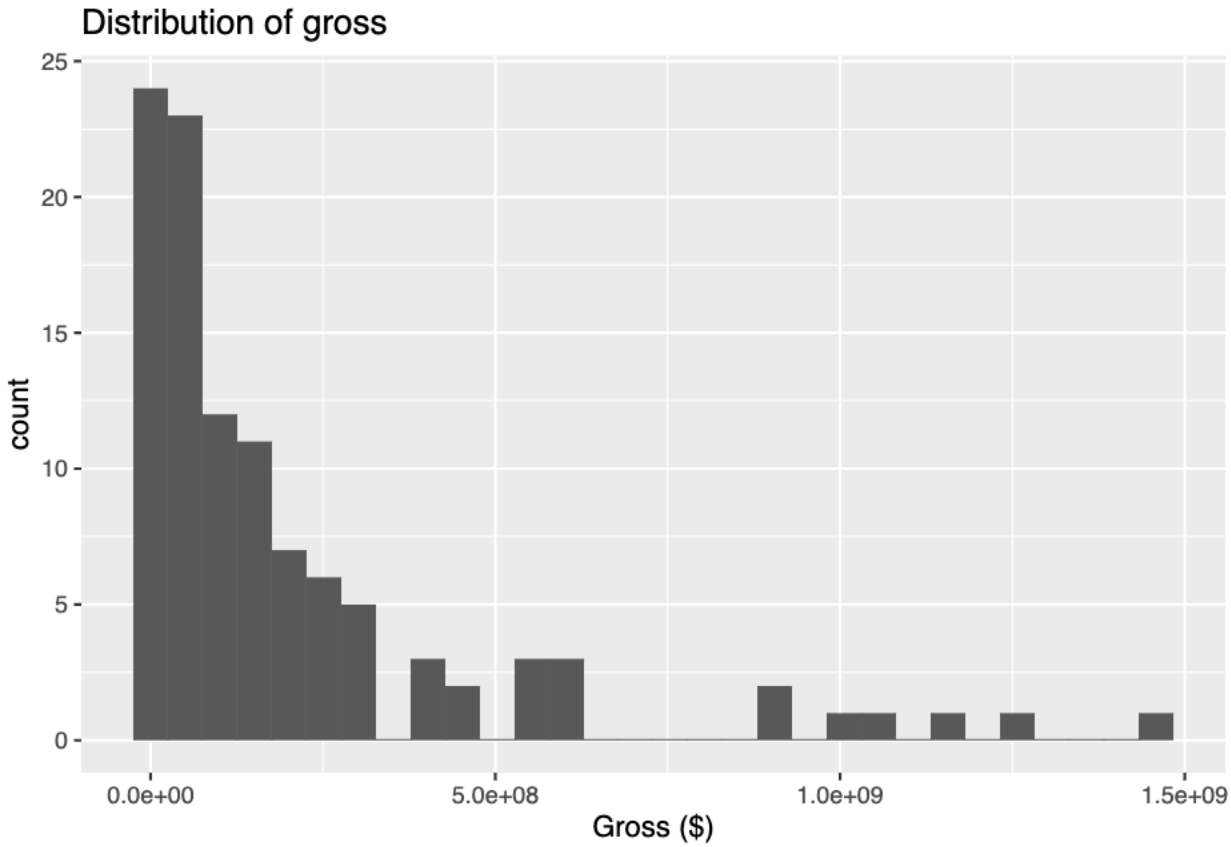
Univariate EDA

```r
# Plots budget distribution
movies_clean %>%
  ggplot(aes(budget_real)) +
  geom_histogram(bins = 30) +
  labs(title = "Distribution of Budget", x = "Budget ($)")
```



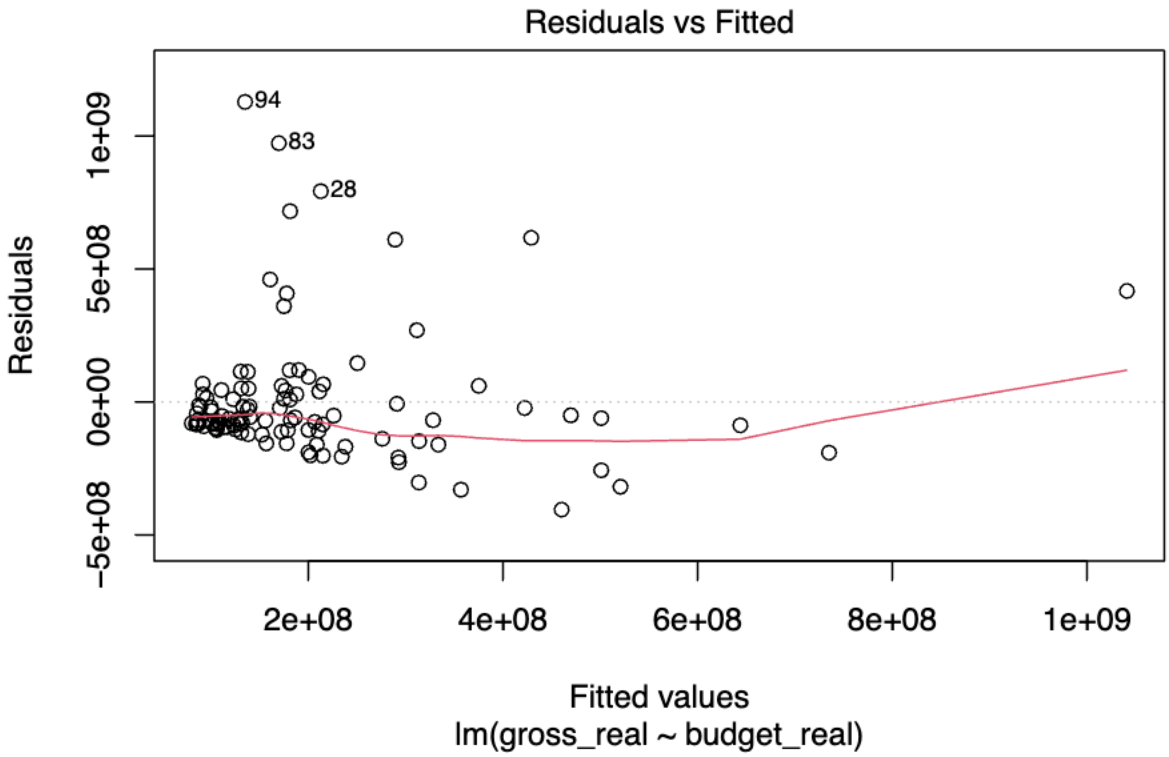
Budget is extremely right-skewed, log transformation should be considered.

```r
# plots gross distribution
movies_clean %>%
  ggplot(aes(gross_real)) +
  geom_histogram(bins = 30) +
  labs(title = "Distribution of gross", x = "Gross ($)")
```



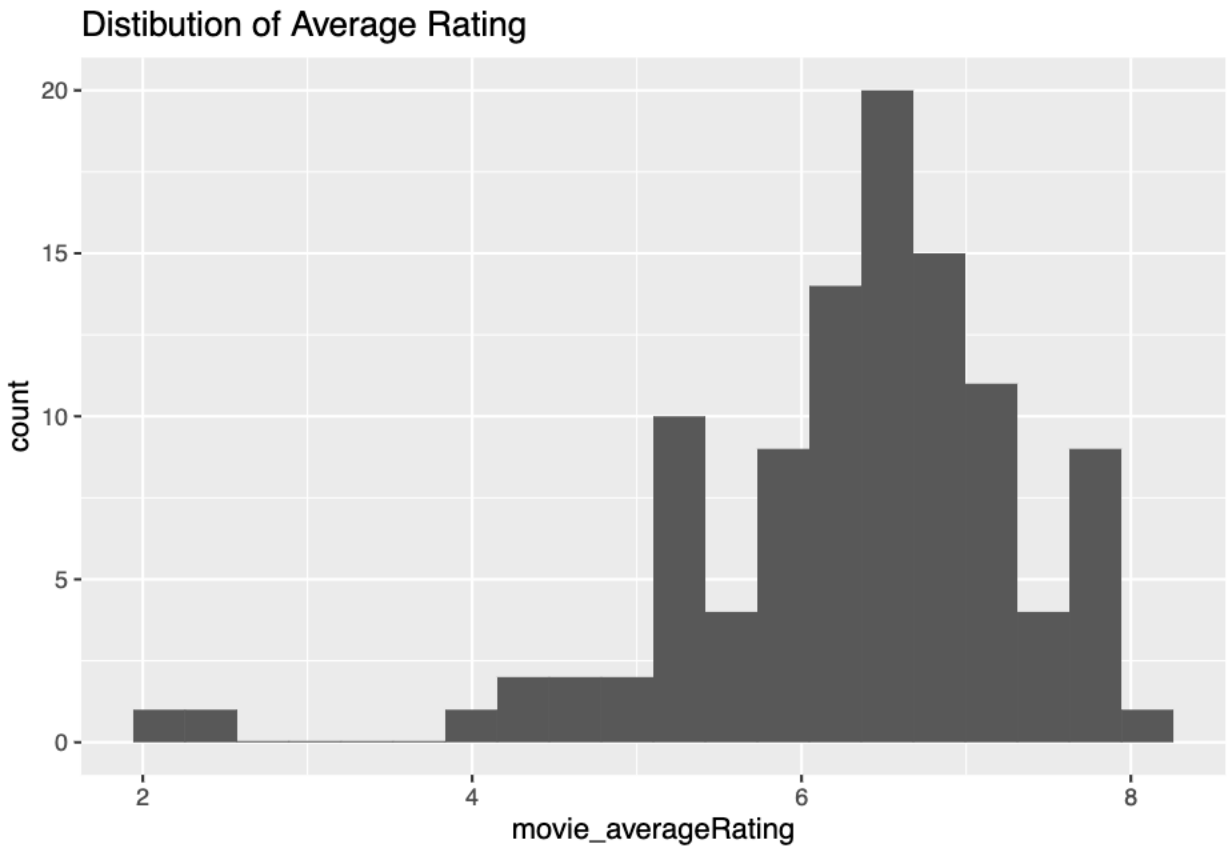
Gross is extremely left-skewed, log transformation should be considered.

```
# creates lm with budget predicting gross
m = lm(gross_real ~ budget_real, data = movies_clean)
# creates a resid vs fitted plot
graphics::plot(m, which = 1)
```



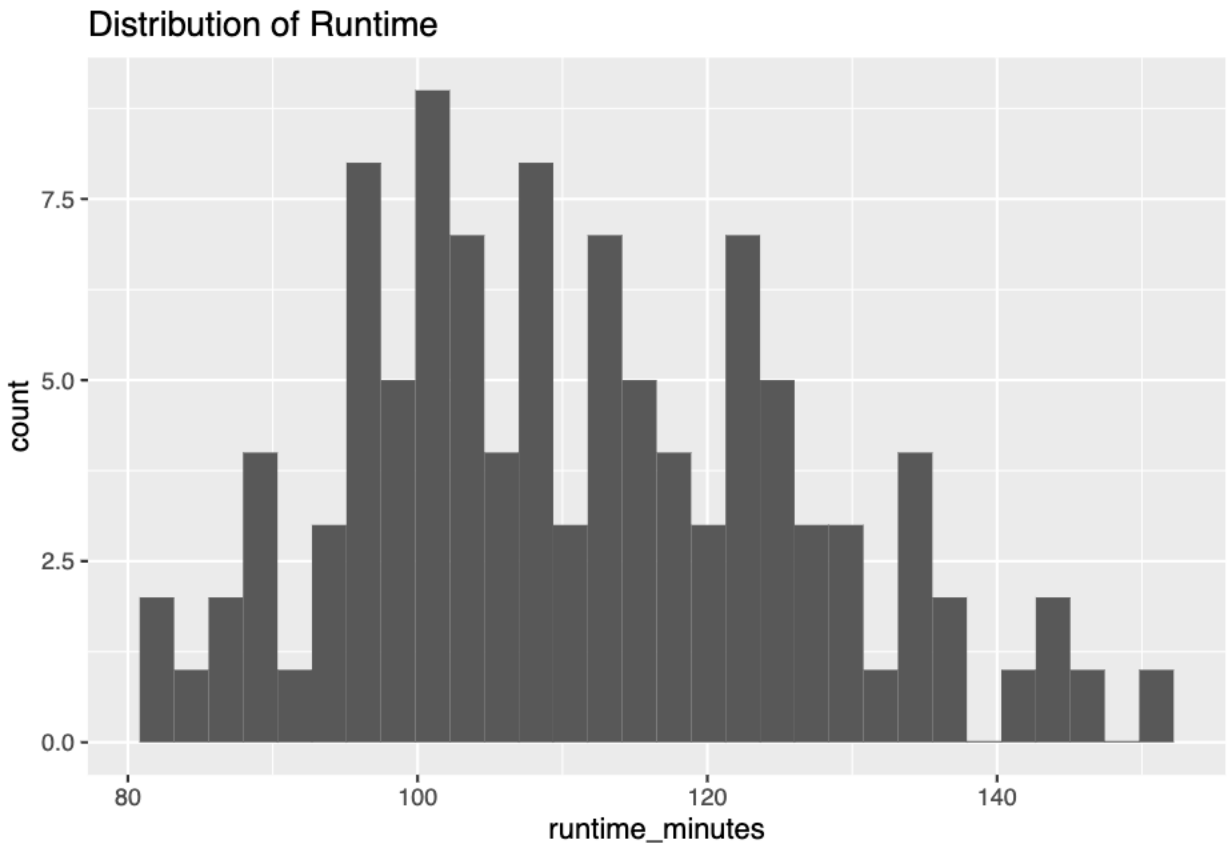
The residual plot shows a funnel pattern which also supports a log transformation.

```
# plots average rating distribution
movies_clean %>%
  ggplot(aes(movie_averageRating)) +
  geom_histogram(bins = 20) + labs(title = "Distibution of Average Rating")
```



Average rating is approximately normally distributed, so no transformation is needed.

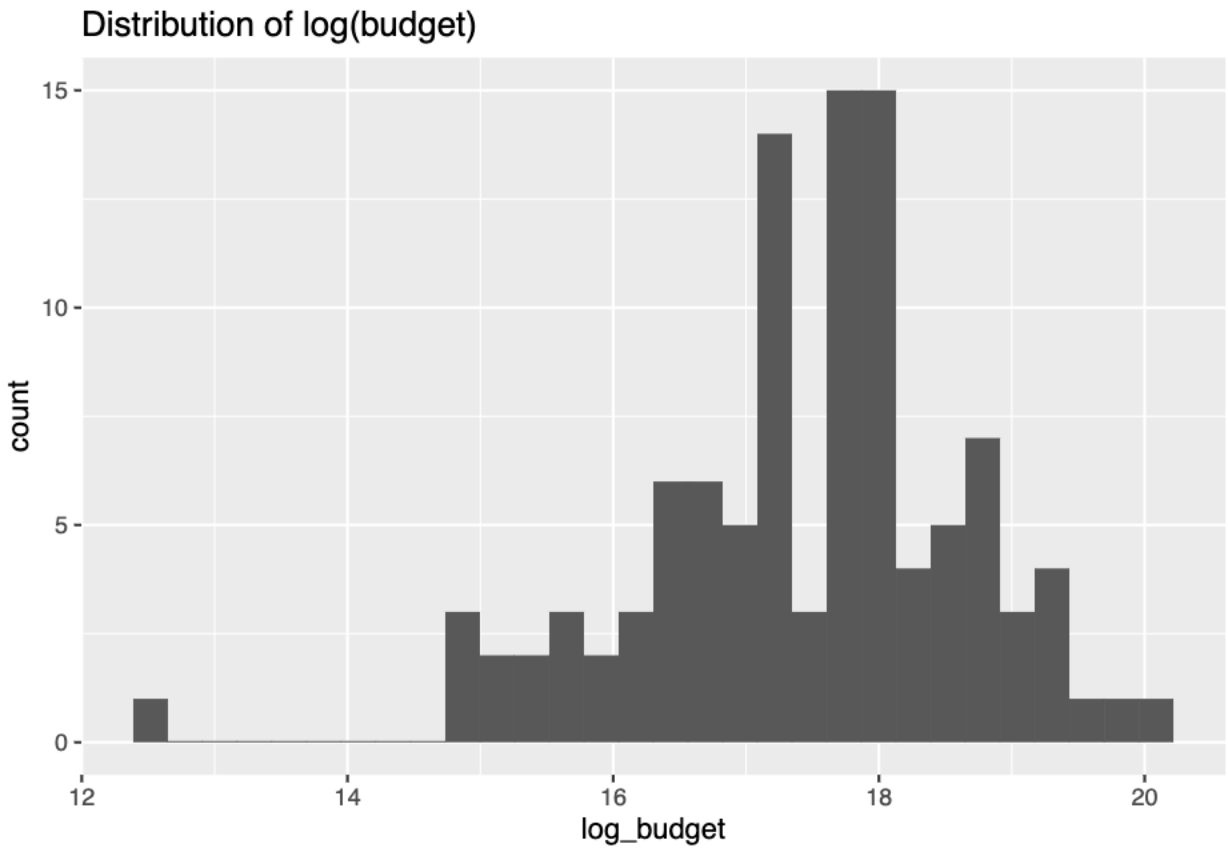
```
# plots runtime distribution
movies_clean %>%
  ggplot(aes(runtime_minutes)) +
  geom_histogram(bins = 30) + labs (title = "Distribution of Runtime")
```



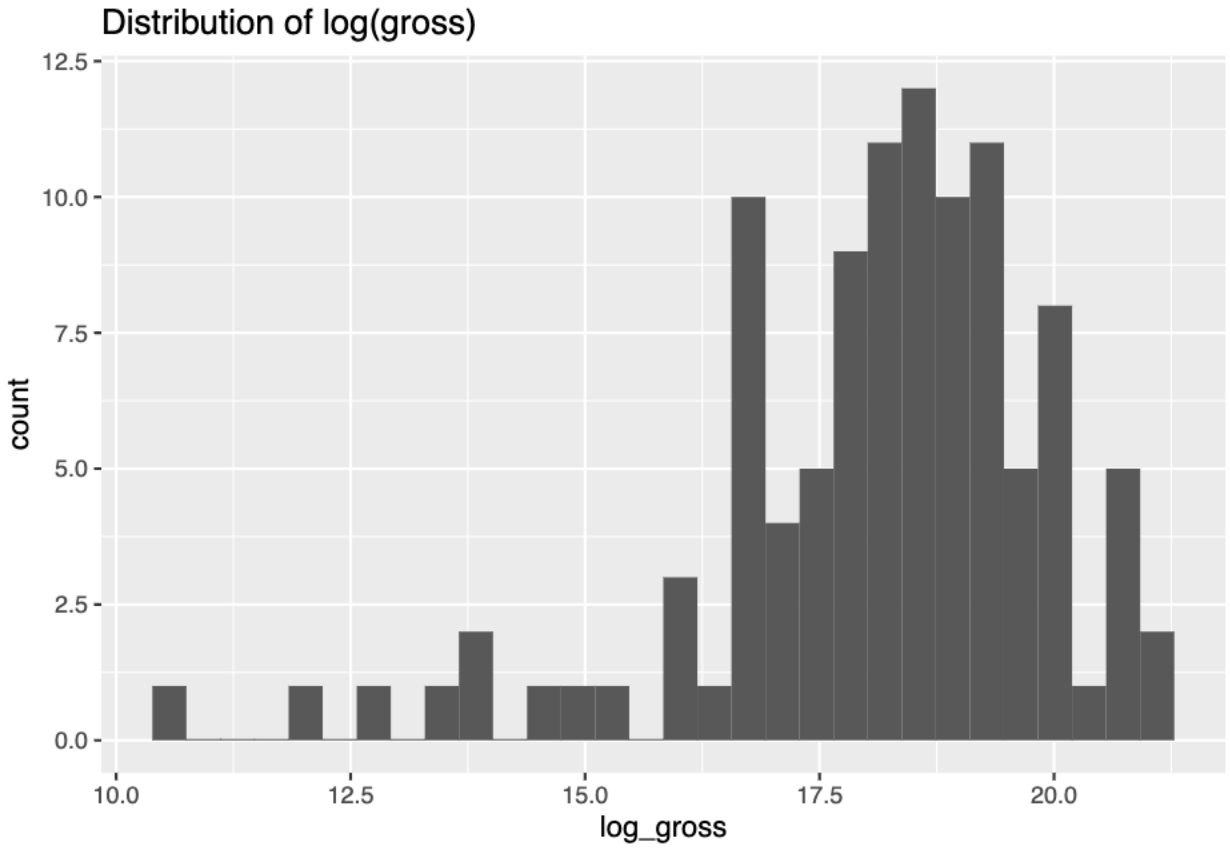
Runtime is approximately normally distributed, so no transformation is needed.

```
# log transforms gross and budget
movies_clean <- movies_clean %>%
  mutate(
    log_gross = log(gross_real),
    log_budget = log(budget_real)
  )
```

```
# plots distribution of log(budget)
movies_clean %>%
  ggplot(aes(x = log_budget)) +
  geom_histogram(bins = 30) + labs(title = "Distribution of log(budget)")
```

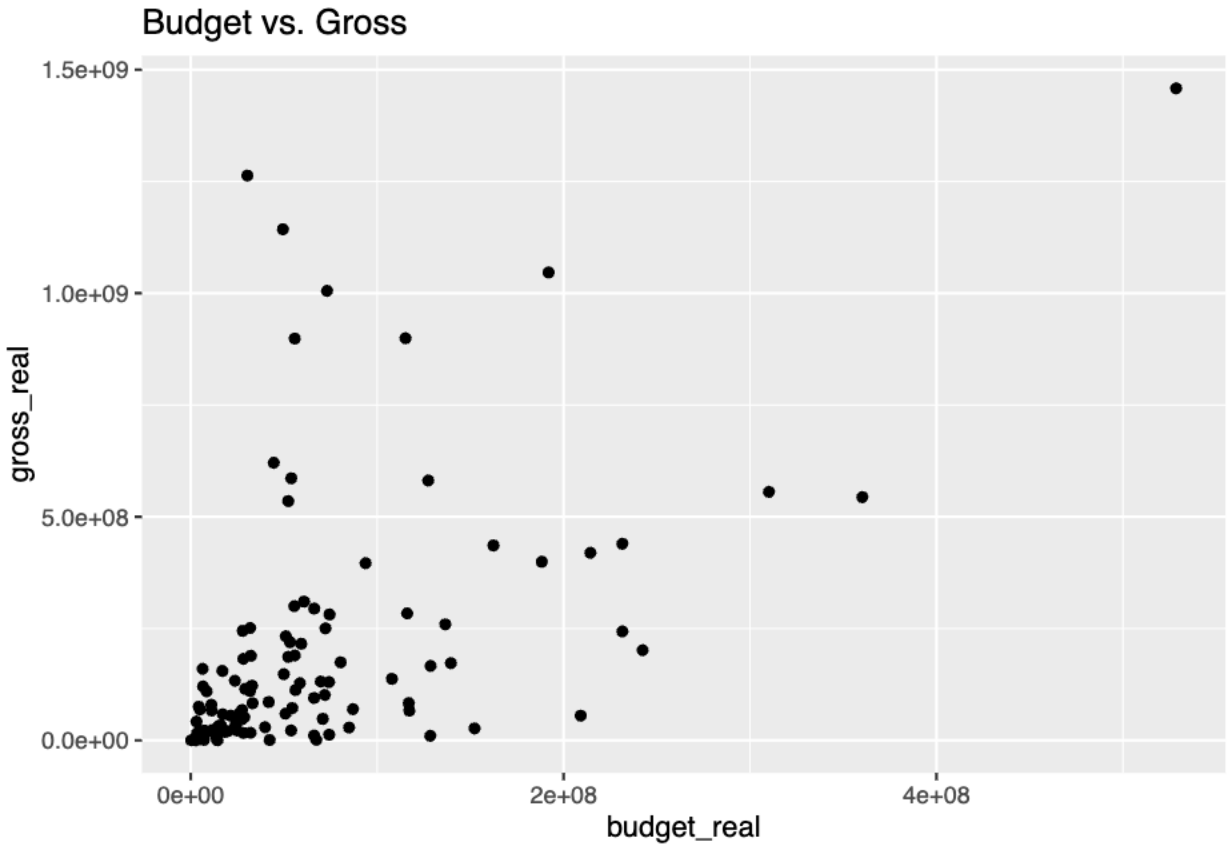



```
# plots distribution of log(gross)
movies_clean %>%
  ggplot(aes(x = log_gross)) +
  geom_histogram(bins = 30) + labs(title = "Distribution of log(gross)")
```

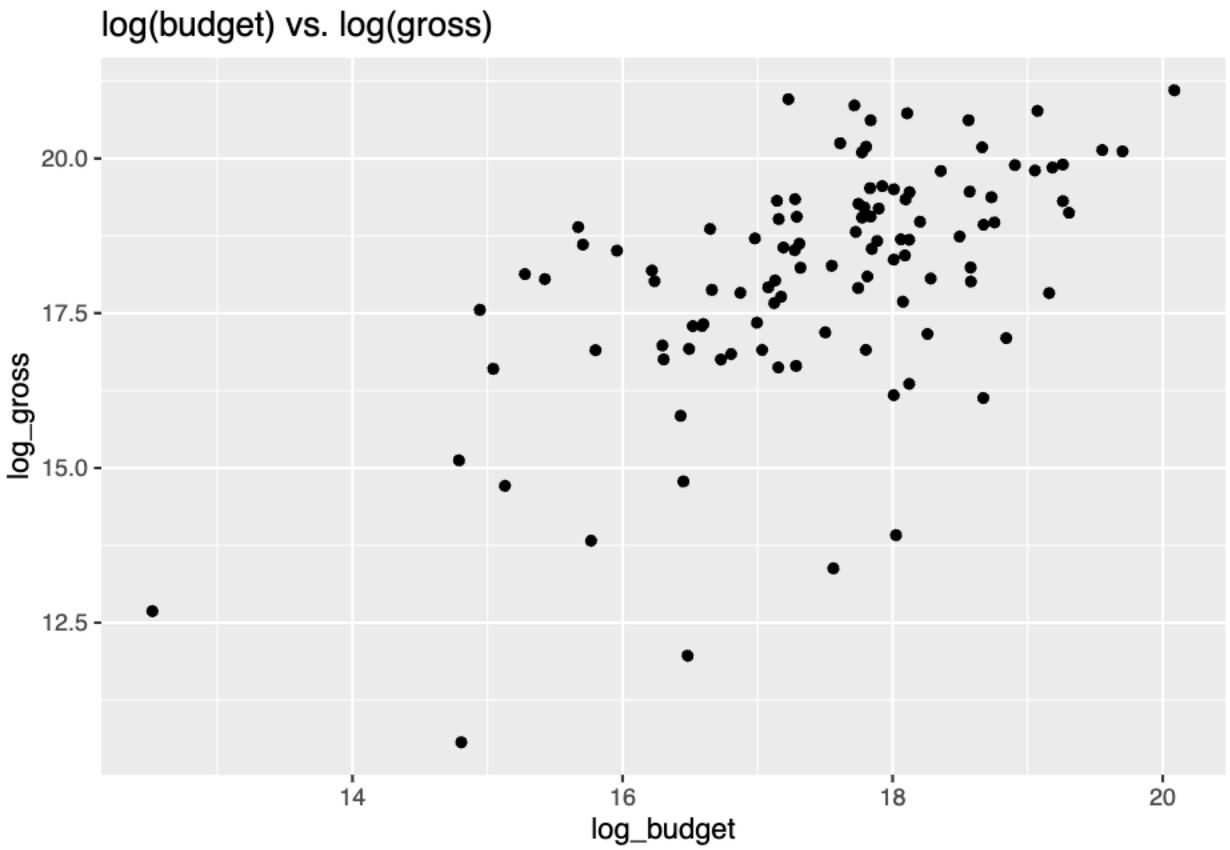


Bivariate EDA

```
# plots budget vs gross
ggplot(movies_clean, aes(budget_real, gross_real)) +
  geom_point() + labs(title = "Budget vs. Gross")
```



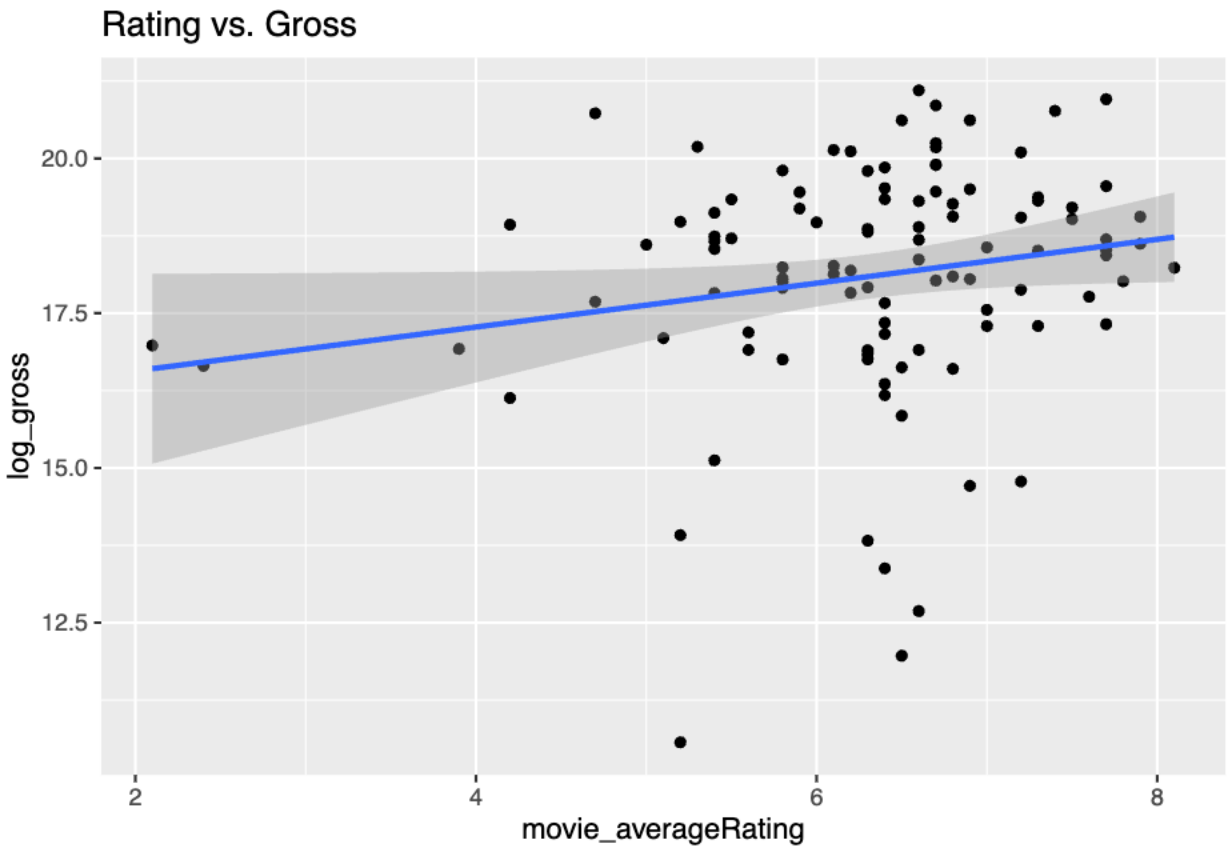

```
# plots log(budget() vs log(gross)

ggplot(movies_clean, aes(log_budget, log_gross)) +

  geom_point() + labs(title = "log(budget) vs. log(gross)")
```



Budget and gross are positively associated.

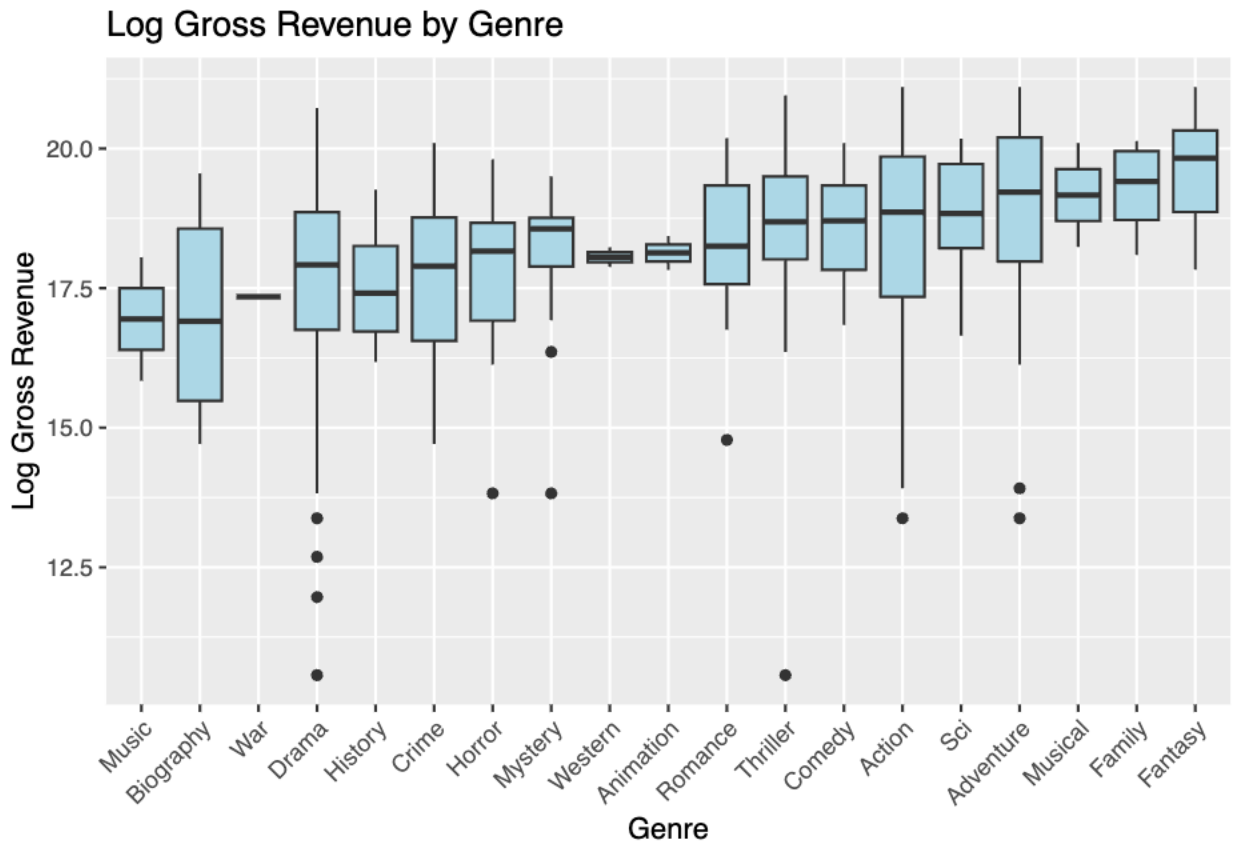
```
# plots rating vs gross
ggplot(movies_clean, aes(movie_averageRating, log_gross)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE) + labs(title = "Rating vs. Gross")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The association is weak, but it is positively associated.

```
# converts genre indicators to rows and removes "is"
movies_long = movies_clean %>% pivot_longer(cols = starts_with("is_"),
  names_to = "genre", values_to = "present") %>%
  filter(present == TRUE) %>% mutate(genre = sub("is_", "", genre))


# plots log(gross) by genre
movies_long %>%  mutate(genre = reorder(genre, log_gross, FUN = mean)) %>%
  ggplot(aes(x = genre, y = log_gross)) + geom_boxplot(fill = "lightblue") +
  labs(title = "Log Gross Revenue by Genre", x = "Genre", y = "Log Gross Revenue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
genre_vars = grep("^is_", names(movies_clean), value = TRUE)


# sets action as reference

genre_vars = setdiff(genre_vars, "is_Action")
```

```
movies_clean[genre_vars] = lapply(movies_clean[genre_vars], as.numeric)


formula_string = paste(
  "log_gross ~ log_budget + movie_averageRating + runtime_minutes +
  director_age_at_prod + production_year +", paste(genre_vars, collapse = " + "))


# Convert to formula

full_form = as.formula(formula_string)
```

```r
# Fit full LM
lm_full = lm(full_form, data = movies_clean)

# exhaustive search via regsubsets
model = regsubsets(full_form, data = movies_clean, nvmax = 10, method="exhaustive")

summary(model)
```

```
## Subset selection object
## Call: regsubsets.formula(full_form, data = movies_clean, nvmax = 10,
##     method = "exhaustive")
## 23 Variables  (and intercept)
##                       Forced in Forced out
## log_budget                FALSE      FALSE
## movie_averageRating       FALSE      FALSE
## runtime_minutes           FALSE      FALSE
## director_age_at_prod      FALSE      FALSE
## production_year           FALSE      FALSE
## is_Adventure              FALSE      FALSE
## is_Fantasy                FALSE      FALSE
## is_Family                 FALSE      FALSE
## is_Sci                    FALSE      FALSE
## is_Comedy                 FALSE      FALSE
## is_Animation              FALSE      FALSE
## is_Horror                 FALSE      FALSE
## is_Thriller               FALSE      FALSE
## is_Drama                  FALSE      FALSE
```

```
## is_Romance               FALSE      FALSE
## is_Musical               FALSE      FALSE
## is_Crime                 FALSE      FALSE
## is_Mystery               FALSE      FALSE
## is_History               FALSE      FALSE
## is_Biography             FALSE      FALSE
## is_Music                 FALSE      FALSE
## is_War                   FALSE      FALSE
## is_Western               FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: exhaustive
##          log_budget movie_averageRating runtime_minutes director_age_at_prod
## 1  ( 1 ) "*"        " "                 " "             " "
## 2  ( 1 ) "*"        "*"                 " "             " "
## 3  ( 1 ) "*"        "*"                 " "             " "
## 4  ( 1 ) "*"        "*"                 " "             " "
## 5  ( 1 ) "*"        "*"                 " "             "*"
## 6  ( 1 ) "*"        "*"                 " "             "*"
## 7  ( 1 ) "*"        "*"                 " "             "*"
## 8  ( 1 ) "*"        "*"                 " "             "*"
## 9  ( 1 ) "*"        "*"                 " "             "*"
## 10  ( 1 ) "*"       "*"                 " "             "*"
##          production_year is_Adventure is_Fantasy is_Family is_Sci is_Comedy
## 1  ( 1 ) " "             " "          " "        " "       " "    " "
## 2  ( 1 ) " "             " "          " "        " "       " "    " "
## 3  ( 1 ) " "             " "          " "        " "       " "    " "
## 4  ( 1 ) " "             " "          "*"        " "       " "    " "
```

```
## 5  ( 1 )  " "            " "          "*"        " "       " "    " "
## 6  ( 1 )  "*"            " "          "*"        " "       " "    " "
## 7  ( 1 )  "*"            " "          "*"        " "       " "    "*"
## 8  ( 1 )  "*"            " "          "*"        " "       " "    "*"
## 9  ( 1 )  "*"            " "          "*"        " "       " "    "*"
## 10  ( 1 ) "*"            " "          "*"        " "       " "    "*"
##           is_Animation is_Horror is_Thriller is_Drama is_Romance is_Musical
## 1  ( 1 )  " "          " "       " "         " "      " "        " "
## 2  ( 1 )  " "          " "       " "         " "      " "        " "
## 3  ( 1 )  " "          " "       " "         "*"      " "        " "
## 4  ( 1 )  " "          " "       " "         "*"      " "        " "
## 5  ( 1 )  " "          " "       " "         "*"      " "        " "
## 6  ( 1 )  " "          " "       " "         "*"      " "        " "
## 7  ( 1 )  " "          " "       " "         "*"      " "        " "
## 8  ( 1 )  " "          "*"       " "         "*"      " "        " "
## 9  ( 1 )  "*"          "*"       " "         "*"      " "        " "
## 10  ( 1 ) "*"          "*"       " "         "*"      " "        "*"
##           is_Crime is_Mystery is_History is_Biography is_Music is_War
## 1  ( 1 )  " "      " "        " "        " "          " "      " "
## 2  ( 1 )  " "      " "        " "        " "          " "      " "
## 3  ( 1 )  " "      " "        " "        " "          " "      " "
## 4  ( 1 )  " "      " "        " "        " "          " "      " "
## 5  ( 1 )  " "      " "        " "        " "          " "      " "
## 6  ( 1 )  " "      " "        " "        " "          " "      " "
## 7  ( 1 )  " "      " "        " "        " "          " "      " "
## 8  ( 1 )  " "      " "        " "        " "          " "      " "
## 9  ( 1 )  " "      " "        " "        " "          " "      " "
```

```
## 10  ( 1 ) " "      " "         " "         " "           " "       " "
##           is_Western
## 1  ( 1 )  " "
## 2  ( 1 )  " "
## 3  ( 1 )  " "
## 4  ( 1 )  " "
## 5  ( 1 )  " "
## 6  ( 1 )  " "
## 7  ( 1 )  " "
## 8  ( 1 )  " "
## 9  ( 1 )  " "
## 10  ( 1 ) " "
```

```
# backward AIC
backward_model = step(lm_full, direction = "backward")
```

```
## Start:  AIC=98.46
## log_gross ~ log_budget + movie_averageRating + runtime_minutes +
##     director_age_at_prod + production_year + is_Adventure + is_Fantasy +
##     is_Family + is_Sci + is_Comedy + is_Animation + is_Horror +
##     is_Thriller + is_Drama + is_Romance + is_Musical + is_Crime +
##     is_Mystery + is_History + is_Biography + is_Music + is_War +
##     is_Western
##
##                       Df Sum of Sq    RSS     AIC
## - is_Western           1     0.003 170.62  96.458
## - is_Music             1     0.057 170.68  96.491
## - is_Family            1     0.060 170.68  96.493
```

```
## - is_Crime              1     0.079 170.70  96.505
## - is_War                1     0.098 170.72  96.517
## - runtime_minutes       1     0.099 170.72  96.517
## - is_Romance            1     0.324 170.94  96.657
## - is_Animation          1     0.374 170.99  96.688
## - is_Mystery            1     0.529 171.15  96.784
## - is_Biography          1     0.595 171.22  96.825
## - is_History            1     0.776 171.40  96.937
## - is_Musical            1     1.078 171.70  97.124
## - is_Sci                1     1.287 171.91  97.253
## - is_Adventure          1     1.316 171.94  97.271
## - is_Thriller           1     1.457 172.08  97.357
## - is_Horror             1     2.661 173.28  98.096
## - is_Drama              1     2.725 173.34  98.135
## <none>                            170.62  98.456
## - is_Comedy             1     4.578 175.20  99.263
## - director_age_at_prod  1     5.992 176.61 100.115
## - production_year       1     7.897 178.52 101.252
## - is_Fantasy            1    11.022 181.64 103.092
## - movie_averageRating   1    17.212 187.83 106.643
## - log_budget            1    32.703 203.32 115.044
##
## Step:  AIC=96.46
## log_gross ~ log_budget + movie_averageRating + runtime_minutes +
##     director_age_at_prod + production_year + is_Adventure + is_Fantasy +
##     is_Family + is_Sci + is_Comedy + is_Animation + is_Horror +
##     is_Thriller + is_Drama + is_Romance + is_Musical + is_Crime +
```

```
##     is_Mystery + is_History + is_Biography + is_Music + is_War
##
##                         Df Sum of Sq    RSS     AIC
## - is_Music               1     0.056 170.68  94.493
## - is_Family              1     0.061 170.68  94.496
## - is_Crime               1     0.076 170.70  94.505
## - is_War                 1     0.095 170.72  94.517
## - runtime_minutes        1     0.097 170.72  94.518
## - is_Romance             1     0.322 170.95  94.658
## - is_Animation           1     0.375 171.00  94.691
## - is_Mystery             1     0.526 171.15  94.784
## - is_Biography           1     0.593 171.22  94.826
## - is_History             1     0.773 171.40  94.937
## - is_Musical             1     1.075 171.70  95.124
## - is_Sci                 1     1.285 171.91  95.254
## - is_Adventure           1     1.378 172.00  95.311
## - is_Thriller            1     1.465 172.09  95.365
## - is_Horror              1     2.658 173.28  96.097
## - is_Drama               1     2.729 173.35  96.140
## <none>                             170.62  96.458
## - is_Comedy              1     4.589 175.21  97.271
## - director_age_at_prod   1     6.298 176.92  98.300
## - production_year        1     9.599 180.22 100.260
## - is_Fantasy             1    11.023 181.65 101.094
## - movie_averageRating    1    17.351 187.97 104.724
## - log_budget             1    33.222 203.84 113.316
##
```

```
## Step:  AIC=94.49
## log_gross ~ log_budget + movie_averageRating + runtime_minutes +
##     director_age_at_prod + production_year + is_Adventure + is_Fantasy +
##     is_Family + is_Sci + is_Comedy + is_Animation + is_Horror +
##     is_Thriller + is_Drama + is_Romance + is_Musical + is_Crime +
##     is_Mystery + is_History + is_Biography + is_War
##
##                         Df Sum of Sq    RSS     AIC
## - is_Crime               1     0.057 170.74  92.528
## - is_Family              1     0.060 170.74  92.530
## - is_War                 1     0.089 170.77  92.548
## - runtime_minutes        1     0.105 170.78  92.558
## - is_Romance             1     0.288 170.97  92.671
## - is_Animation           1     0.367 171.05  92.721
## - is_Mystery             1     0.517 171.20  92.814
## - is_Biography           1     0.727 171.41  92.944
## - is_History             1     0.743 171.42  92.953
## - is_Musical             1     1.080 171.76  93.162
## - is_Sci                 1     1.289 171.97  93.290
## - is_Adventure           1     1.350 172.03  93.328
## - is_Thriller            1     1.448 172.13  93.388
## - is_Horror              1     2.607 173.29  94.100
## - is_Drama               1     2.684 173.36  94.147
## <none>                             170.68  94.493
## - is_Comedy              1     4.802 175.48  95.434
## - director_age_at_prod   1     6.531 177.21  96.473
## - production_year        1     9.608 180.29  98.298
```

```
## - is_Fantasy            1    11.002 181.68  99.115
## - movie_averageRating   1    17.306 187.99 102.730
## - log_budget            1    33.206 203.89 111.337
##
## Step:  AIC=92.53
## log_gross ~ log_budget + movie_averageRating + runtime_minutes +
##     director_age_at_prod + production_year + is_Adventure + is_Fantasy +
##     is_Family + is_Sci + is_Comedy + is_Animation + is_Horror +
##     is_Thriller + is_Drama + is_Romance + is_Musical + is_Mystery +
##     is_History + is_Biography + is_War
##
##                        Df Sum of Sq    RSS     AIC
## - is_Family             1     0.059 170.79  90.564
## - is_War                1     0.076 170.81  90.575
## - runtime_minutes       1     0.126 170.86  90.606
## - is_Romance            1     0.241 170.98  90.677
## - is_Animation          1     0.383 171.12  90.766
## - is_Mystery            1     0.502 171.24  90.839
## - is_History            1     0.755 171.49  90.996
## - is_Biography          1     0.809 171.54  91.029
## - is_Musical            1     1.160 171.90  91.246
## - is_Sci                1     1.246 171.98  91.299
## - is_Adventure          1     1.318 172.05  91.343
## - is_Thriller           1     1.423 172.16  91.408
## - is_Horror             1     2.605 173.34  92.133
## - is_Drama              1     2.887 173.62  92.306
## <none>                              170.74  92.528
```

```
## - is_Comedy             1     4.751 175.49  93.437
## - director_age_at_prod  1     6.476 177.21  94.474
## - production_year       1     9.564 180.30  96.305
## - is_Fantasy            1    11.146 181.88  97.232
## - movie_averageRating   1    17.249 187.99 100.730
## - log_budget            1    33.232 203.97 109.379
##
## Step:  AIC=90.56
## log_gross ~ log_budget + movie_averageRating + runtime_minutes +
##     director_age_at_prod + production_year + is_Adventure + is_Fantasy +
##     is_Sci + is_Comedy + is_Animation + is_Horror + is_Thriller +
##     is_Drama + is_Romance + is_Musical + is_Mystery + is_History +
##     is_Biography + is_War
##
##                        Df Sum of Sq    RSS     AIC
## - is_War                1     0.078 170.87  88.613
## - runtime_minutes       1     0.100 170.90  88.627
## - is_Romance            1     0.218 171.01  88.700
## - is_Animation          1     0.446 171.24  88.841
## - is_Mystery            1     0.506 171.30  88.878
## - is_History            1     0.746 171.54  89.026
## - is_Biography          1     0.809 171.60  89.065
## - is_Musical            1     1.154 171.95  89.278
## - is_Sci                1     1.198 171.99  89.305
## - is_Thriller           1     1.397 172.19  89.428
## - is_Adventure          1     1.474 172.27  89.476
## - is_Horror             1     2.554 173.35  90.138
```

```
## - is_Drama              1     2.873 173.67  90.333
## <none>                            170.79  90.564
## - is_Comedy             1     4.756 175.55  91.476
## - director_age_at_prod  1     6.445 177.24  92.491
## - production_year       1     9.527 180.32  94.318
## - is_Fantasy            1    11.557 182.35  95.505
## - movie_averageRating   1    17.392 188.19  98.843
## - log_budget            1    34.260 205.05 107.943
##
## Step:  AIC=88.61
## log_gross ~ log_budget + movie_averageRating + runtime_minutes +
##     director_age_at_prod + production_year + is_Adventure + is_Fantasy +
##     is_Sci + is_Comedy + is_Animation + is_Horror + is_Thriller +
##     is_Drama + is_Romance + is_Musical + is_Mystery + is_History +
##     is_Biography
##
##                        Df Sum of Sq    RSS     AIC
## - runtime_minutes       1     0.111 170.98  86.682
## - is_Romance            1     0.194 171.07  86.733
## - is_Animation          1     0.441 171.31  86.886
## - is_Mystery            1     0.493 171.37  86.918
## - is_History            1     0.719 171.59  87.058
## - is_Biography          1     0.771 171.64  87.090
## - is_Musical            1     1.132 172.00  87.313
## - is_Sci                1     1.162 172.03  87.331
## - is_Thriller           1     1.350 172.22  87.447
## - is_Adventure          1     1.405 172.28  87.481
```

```
## - is_Horror             1     2.504 173.38  88.155
## - is_Drama              1     2.890 173.76  88.391
## <none>                              170.87  88.613
## - is_Comedy             1     4.725 175.60  89.504
## - director_age_at_prod  1     6.379 177.25  90.498
## - production_year       1    10.001 180.87  92.642
## - is_Fantasy            1    11.482 182.35  93.507
## - movie_averageRating   1    17.328 188.20  96.851
## - log_budget            1    34.327 205.20 106.018
##
## Step:  AIC=86.68
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Adventure + is_Fantasy + is_Sci + is_Comedy +
##     is_Animation + is_Horror + is_Thriller + is_Drama + is_Romance +
##     is_Musical + is_Mystery + is_History + is_Biography
##
##                        Df Sum of Sq    RSS     AIC
## - is_Romance            1     0.196 171.18  84.803
## - is_Mystery            1     0.483 171.47  84.981
## - is_Animation          1     0.627 171.61  85.070
## - is_History            1     0.795 171.78  85.173
## - is_Biography          1     0.807 171.79  85.181
## - is_Sci                1     1.158 172.14  85.398
## - is_Musical            1     1.161 172.15  85.399
## - is_Thriller           1     1.467 172.45  85.588
## - is_Adventure          1     1.540 172.52  85.632
## - is_Horror             1     2.521 173.50  86.233
```

```
## - is_Drama              1     2.797 173.78  86.402
## <none>                            170.98  86.682
## - is_Comedy             1     4.615 175.60  87.505
## - director_age_at_prod  1     6.268 177.25  88.498
## - production_year       1     9.940 180.92  90.671
## - is_Fantasy            1    12.656 183.64  92.251
## - movie_averageRating   1    21.972 192.96  97.496
## - log_budget            1    47.691 218.68 110.760
##
## Step:  AIC=84.8
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Adventure + is_Fantasy + is_Sci + is_Comedy +
##     is_Animation + is_Horror + is_Thriller + is_Drama + is_Musical +
##     is_Mystery + is_History + is_Biography
##
##                        Df Sum of Sq    RSS     AIC
## - is_Mystery            1     0.408 171.59  83.056
## - is_Animation          1     0.687 171.87  83.228
## - is_History            1     0.689 171.87  83.229
## - is_Biography          1     0.715 171.90  83.245
## - is_Sci                1     1.006 172.19  83.424
## - is_Musical            1     1.166 172.35  83.523
## - is_Thriller           1     1.301 172.48  83.606
## - is_Adventure          1     1.345 172.53  83.633
## - is_Horror             1     2.325 173.50  84.233
## - is_Drama              1     3.028 174.21  84.662
## <none>                            171.18  84.803
```

```
## - is_Comedy              1     5.048 176.23  85.884
## - director_age_at_prod   1     6.406 177.59  86.698
## - production_year        1    10.006 181.19  88.825
## - is_Fantasy             1    12.480 183.66  90.262
## - movie_averageRating    1    22.049 193.23  95.646
## - log_budget             1    49.842 221.02 109.891
##
## Step:  AIC=83.06
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Adventure + is_Fantasy + is_Sci + is_Comedy +
##     is_Animation + is_Horror + is_Thriller + is_Drama + is_Musical +
##     is_History + is_Biography
##
##                        Df Sum of Sq    RSS     AIC
## - is_Biography          1     0.592 172.18  81.421
## - is_History            1     0.628 172.22  81.443
## - is_Animation          1     0.740 172.33  81.512
## - is_Sci                1     0.975 172.56  81.657
## - is_Adventure          1     1.085 172.67  81.724
## - is_Musical            1     1.097 172.69  81.732
## - is_Thriller           1     1.315 172.90  81.865
## - is_Horror             1     3.103 174.69  82.955
## <none>                              171.59  83.056
## - is_Drama              1     3.719 175.31  83.329
## - is_Comedy             1     4.691 176.28  83.915
## - director_age_at_prod  1     6.606 178.19  85.060
## - production_year       1     9.699 181.29  86.884
```

```
## - is_Fantasy            1    12.304 183.89  88.396
## - movie_averageRating   1    22.963 194.55  94.369
## - log_budget            1    49.436 221.03 107.893
##
## Step:  AIC=81.42
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Adventure + is_Fantasy + is_Sci + is_Comedy +
##     is_Animation + is_Horror + is_Thriller + is_Drama + is_Musical +
##     is_History
##
##                        Df Sum of Sq    RSS     AIC
## - is_History            1     0.727 172.91  79.867
## - is_Animation          1     0.815 173.00  79.921
## - is_Sci                1     0.883 173.06  79.963
## - is_Adventure          1     0.918 173.10  79.985
## - is_Musical            1     1.033 173.21  80.055
## - is_Thriller           1     1.127 173.31  80.113
## - is_Horror             1     2.789 174.97  81.124
## <none>                              172.18  81.421
## - is_Drama              1     4.055 176.24  81.888
## - is_Comedy             1     4.217 176.40  81.985
## - director_age_at_prod  1     6.228 178.41  83.187
## - production_year       1     9.383 181.56  85.045
## - is_Fantasy            1    11.994 184.17  86.559
## - movie_averageRating   1    23.933 196.11  93.217
## - log_budget            1    48.982 221.16 105.958
##
```

```
## Step:  AIC=79.87
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Adventure + is_Fantasy + is_Sci + is_Comedy +
##     is_Animation + is_Horror + is_Thriller + is_Drama + is_Musical
##
##                       Df Sum of Sq    RSS     AIC
## - is_Adventure         1     0.695 173.60  78.292
## - is_Sci               1     0.715 173.62  78.305
## - is_Thriller          1     0.873 173.78  78.401
## - is_Musical           1     0.900 173.81  78.417
## - is_Animation         1     0.926 173.83  78.434
## - is_Horror            1     2.548 175.46  79.418
## <none>                             172.91  79.867
## - is_Comedy            1     3.764 176.67  80.150
## - is_Drama             1     3.954 176.86  80.264
## - director_age_at_prod  1     6.307 179.21  81.665
## - production_year      1     8.775 181.68  83.114
## - is_Fantasy           1    11.357 184.26  84.610
## - movie_averageRating  1    23.861 196.77  91.570
## - log_budget           1    51.492 224.40 105.498
##
## Step:  AIC=78.29
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Fantasy + is_Sci + is_Comedy + is_Animation +
##     is_Horror + is_Thriller + is_Drama + is_Musical
##
##                       Df Sum of Sq    RSS     AIC
```

```
## - is_Sci                1     0.524 174.12  76.612
## - is_Thriller           1     0.631 174.23  76.677
## - is_Musical            1     0.678 174.28  76.706
## - is_Animation          1     0.786 174.39  76.772
## - is_Horror             1     2.089 175.69  77.561
## - is_Comedy             1     3.289 176.89  78.282
## <none>                              173.60  78.292
## - is_Drama              1     5.417 179.02  79.550
## - director_age_at_prod  1     5.830 179.43  79.794
## - production_year       1     9.517 183.12  81.950
## - is_Fantasy            1    11.096 184.70  82.860
## - movie_averageRating   1    23.167 196.77  89.571
## - log_budget            1    65.631 239.23 110.284
##
## Step:  AIC=76.61
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Fantasy + is_Comedy + is_Animation +
##     is_Horror + is_Thriller + is_Drama + is_Musical
##
##                         Df Sum of Sq    RSS     AIC
## - is_Thriller           1     0.422 174.55  74.868
## - is_Musical            1     0.598 174.72  74.975
## - is_Animation          1     1.065 175.19  75.258
## - is_Horror             1     2.064 176.19  75.861
## - is_Comedy             1     2.845 176.97  76.330
## <none>                              174.12  76.612
## - director_age_at_prod  1     5.818 179.94  78.095
```

```
## - is_Drama              1     6.847 180.97  78.700
## - production_year       1     9.250 183.38  80.098
## - is_Fantasy            1    10.575 184.70  80.862
## - movie_averageRating   1    22.893 197.02  87.705
## - log_budget            1    70.120 244.25 110.482
##
## Step:  AIC=74.87
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Fantasy + is_Comedy + is_Animation +
##     is_Horror + is_Drama + is_Musical
##
##                        Df Sum of Sq    RSS     AIC
## - is_Musical            1     0.550 175.10  73.202
## - is_Animation          1     1.384 175.93  73.705
## - is_Horror             1     2.032 176.58  74.095
## - is_Comedy             1     2.428 176.97  74.332
## <none>                              174.55  74.868
## - director_age_at_prod  1     6.556 181.10  76.776
## - production_year       1     9.015 183.56  78.206
## - is_Drama              1     9.461 184.01  78.463
## - is_Fantasy            1    10.168 184.72  78.870
## - movie_averageRating   1    25.853 200.40  87.509
## - log_budget            1    69.739 244.29 108.500
##
## Step:  AIC=73.2
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Fantasy + is_Comedy + is_Animation +
```

```
##     is_Horror + is_Drama
##
##                         Df Sum of Sq    RSS     AIC
## - is_Animation           1     1.436 176.53  72.068
## - is_Horror              1     1.957 177.06  72.380
## - is_Comedy              1     2.665 177.76  72.803
## <none>                               175.10  73.202
## - director_age_at_prod   1     6.529 181.63  75.083
## - production_year        1     8.925 184.02  76.472
## - is_Drama               1     9.277 184.38  76.674
## - is_Fantasy             1     9.873 184.97  77.016
## - movie_averageRating    1    25.768 200.87  85.755
## - log_budget             1    72.000 247.10 107.713
##
## Step:  AIC=72.07
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Fantasy + is_Comedy + is_Horror + is_Drama
##
##                         Df Sum of Sq    RSS     AIC
## - is_Horror              1     2.324 178.86  71.454
## - is_Comedy              1     3.347 179.88  72.059
## <none>                               176.53  72.068
## - director_age_at_prod   1     6.788 183.32  74.067
## - is_Drama               1     8.278 184.81  74.925
## - production_year        1    10.075 186.61  75.951
## - is_Fantasy             1    11.144 187.68  76.556
## - movie_averageRating    1    25.158 201.69  84.190
```

```
## - log_budget             1    70.922 247.46 105.866
##
## Step:  AIC=71.45
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Fantasy + is_Comedy + is_Drama
##
##                         Df Sum of Sq    RSS     AIC
## - is_Comedy              1     2.761 181.62  71.078
## <none>                               178.86  71.454
## - director_age_at_prod   1     7.337 186.19  73.715
## - production_year        1     9.250 188.11  74.799
## - is_Fantasy             1    10.525 189.38  75.515
## - is_Drama               1    12.348 191.21  76.530
## - movie_averageRating    1    22.881 201.74  82.215
## - log_budget             1    69.399 248.26 104.209
##
## Step:  AIC=71.08
## log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Fantasy + is_Drama
##
##                         Df Sum of Sq    RSS     AIC
## <none>                               181.62  71.078
## - production_year        1     8.342 189.96  73.838
## - director_age_at_prod   1     8.378 190.00  73.858
## - is_Fantasy             1     9.453 191.07  74.456
## - is_Drama               1    18.776 200.40  79.506
## - movie_averageRating    1    25.531 207.15  83.020
```

```
## - log_budget             1    66.806 248.43 102.280
```

```
summary(backward_model)
```

```
##
## Call:
## lm(formula = log_gross ~ log_budget + movie_averageRating + director_age_at_prod +
##     production_year + is_Fantasy + is_Drama, data = movies_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5341 -0.4799 -0.0085  0.8997  2.4942
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           41.752391  18.139863   2.302 0.023448 *
## log_budget             0.721868   0.119622   6.035 2.78e-08 ***
## movie_averageRating    0.522601   0.140088   3.731 0.000319 ***
## director_age_at_prod -0.032209   0.015071  -2.137 0.035056 *
## production_year      -0.018838   0.008834  -2.132 0.035447 *
## is_Fantasy             1.034404   0.455694   2.270 0.025378 *
## is_Drama              -0.945456   0.295528  -3.199 0.001852 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.354 on 99 degrees of freedom
## Multiple R-squared:  0.5209, Adjusted R-squared:  0.4919
## F-statistic: 17.94 on 6 and 99 DF,  p-value: 5.525e-14
```

```r
# forward AIC
null_model = lm(log_gross ~ 1, data = movies_clean)
forward_model = step (null_model, scope = formula(lm_full), direction = "forward")
```

```
## Start:  AIC=137.09
## log_gross ~ 1
##
##                         Df Sum of Sq    RSS     AIC
## + log_budget             1   130.258 248.86  94.467
## + is_Drama               1    50.484 328.64 123.941
## + runtime_minutes        1    36.240 342.88 128.438
## + is_Fantasy             1    30.466 348.66 130.209
## + is_Adventure           1    28.274 350.85 130.873
## + director_age_at_prod   1    16.574 362.55 134.350
## + movie_averageRating    1    13.888 365.23 135.133
## + is_Biography           1     8.423 370.70 136.707
## <none>                               379.12 137.088
## + is_Sci                 1     6.179 372.94 137.347
## + production_year        1     6.074 373.05 137.376
## + is_Comedy              1     5.683 373.44 137.488
## + is_Family              1     5.635 373.49 137.501
## + is_Crime               1     4.281 374.84 137.885
## + is_Music               1     2.703 376.42 138.330
## + is_Musical             1     2.331 376.79 138.435
## + is_Horror              1     2.139 376.98 138.489
## + is_Thriller            1     2.044 377.08 138.515
## + is_History             1     1.177 377.95 138.759
```

```
## + is_War                 1     0.572 378.55 138.929
## + is_Romance             1     0.252 378.87 139.018
## + is_Mystery             1     0.166 378.96 139.042
## + is_Western             1     0.004 379.12 139.087
## + is_Animation           1     0.002 379.12 139.088
##
## Step:  AIC=94.47
## log_gross ~ log_budget
##
##                         Df Sum of Sq    RSS    AIC
## + movie_averageRating    1   20.7693 228.09 87.230
## + is_Drama               1   13.3540 235.51 90.621
## + director_age_at_prod   1   11.8459 237.02 91.298
## + is_Comedy              1   10.6039 238.26 91.852
## + production_year        1   10.5835 238.28 91.861
## <none>                               248.86 94.467
## + is_Thriller            1    3.5608 245.30 94.940
## + is_Fantasy             1    2.7356 246.13 95.296
## + is_Animation           1    2.1535 246.71 95.546
## + is_Mystery             1    2.0788 246.78 95.578
## + is_History             1    1.2002 247.66 95.955
## + runtime_minutes        1    0.6184 248.25 96.204
## + is_Crime               1    0.6035 248.26 96.210
## + is_Musical             1    0.3705 248.49 96.310
## + is_Western             1    0.3000 248.56 96.340
## + is_Music               1    0.1139 248.75 96.419
## + is_War                 1    0.1114 248.75 96.420
```

```
## + is_Horror              1    0.1064 248.76 96.422
## + is_Adventure           1    0.1013 248.76 96.424
## + is_Biography           1    0.0422 248.82 96.449
## + is_Sci                 1    0.0374 248.83 96.452
## + is_Romance             1    0.0321 248.83 96.454
## + is_Family              1    0.0008 248.86 96.467
##
## Step:  AIC=87.23
## log_gross ~ log_budget + movie_averageRating
##
##                         Df Sum of Sq    RSS    AIC
## + is_Drama               1   25.5083 202.59 76.659
## + director_age_at_prod   1   10.8703 217.22 84.054
## + is_Comedy              1    9.4781 218.62 84.731
## + is_Horror              1    6.6080 221.49 86.114
## + is_Fantasy             1    6.2675 221.83 86.277
## + production_year        1    4.7438 223.35 87.002
## <none>                               228.09 87.230
## + is_Mystery             1    3.5122 224.58 87.585
## + is_Animation           1    2.7393 225.35 87.949
## + is_Crime               1    2.6440 225.45 87.994
## + is_History             1    1.6303 226.46 88.470
## + runtime_minutes        1    1.5769 226.52 88.495
## + is_Thriller            1    1.5665 226.53 88.500
## + is_Biography           1    0.7076 227.39 88.901
## + is_Sci                 1    0.3238 227.77 89.079
## + is_Musical             1    0.2223 227.87 89.127
```

```
## + is_War                1    0.1266 227.97 89.171
## + is_Adventure          1    0.1180 227.98 89.175
## + is_Romance            1    0.1005 227.99 89.183
## + is_Family             1    0.0877 228.01 89.189
## + is_Western            1    0.0695 228.03 89.198
## + is_Music              1    0.0247 228.07 89.219
##
## Step:  AIC=76.66
## log_gross ~ log_budget + movie_averageRating + is_Drama
##
##                        Df Sum of Sq    RSS    AIC
## + is_Fantasy            1    6.2946 196.29 75.313
## + director_age_at_prod  1    5.2536 197.33 75.874
## + is_Animation          1    4.8776 197.71 76.076
## + production_year       1    4.7920 197.79 76.122
## <none>                              202.59 76.659
## + is_Comedy             1    1.8446 200.74 77.689
## + is_Horror             1    1.2290 201.36 78.014
## + is_Crime              1    1.0347 201.55 78.116
## + runtime_minutes       1    0.7751 201.81 78.253
## + is_Music              1    0.4025 202.18 78.448
## + is_Musical            1    0.3978 202.19 78.451
## + is_Mystery            1    0.2859 202.30 78.509
## + is_Romance            1    0.0860 202.50 78.614
## + is_Family             1    0.0741 202.51 78.620
## + is_Biography          1    0.0682 202.52 78.623
## + is_Adventure          1    0.0608 202.53 78.627
```

```
## + is_Western            1    0.0539 202.53 78.631
## + is_War                1    0.0241 202.56 78.646
## + is_History            1    0.0233 202.56 78.647
## + is_Sci                1    0.0142 202.57 78.652
## + is_Thriller           1    0.0000 202.59 78.659
##
## Step:  AIC=75.31
## log_gross ~ log_budget + movie_averageRating + is_Drama + is_Fantasy
##
##                        Df Sum of Sq    RSS    AIC
## + director_age_at_prod  1    6.3312 189.96 73.838
## + production_year       1    6.2947 190.00 73.858
## + is_Animation          1    3.8961 192.40 75.188
## <none>                              196.29 75.313
## + is_Comedy             1    2.7109 193.58 75.839
## + is_Horror             1    1.5111 194.78 76.494
## + is_Musical            1    0.6915 195.60 76.939
## + is_Crime              1    0.6279 195.66 76.974
## + is_Mystery            1    0.3471 195.94 77.126
## + is_Music              1    0.3448 195.95 77.127
## + is_Romance            1    0.2698 196.02 77.167
## + is_Thriller           1    0.1581 196.13 77.228
## + is_Adventure          1    0.1120 196.18 77.253
## + is_Biography          1    0.0830 196.21 77.268
## + is_Western            1    0.0561 196.24 77.283
## + is_Sci                1    0.0535 196.24 77.284
## + runtime_minutes       1    0.0519 196.24 77.285
```

```
## + is_War                1    0.0456 196.25 77.289
## + is_Family             1    0.0317 196.26 77.296
## + is_History            1    0.0008 196.29 77.313
##
## Step:  AIC=73.84
## log_gross ~ log_budget + movie_averageRating + is_Drama + is_Fantasy +
##     director_age_at_prod
##
##                   Df Sum of Sq    RSS    AIC
## + production_year  1    8.3420 181.62 71.078
## <none>                         189.96 73.838
## + is_Animation     1    3.4687 186.49 73.885
## + is_Comedy        1    1.8527 188.11 74.799
## + is_Horror        1    1.1454 188.81 75.197
## + is_Western       1    0.6910 189.27 75.452
## + is_Musical       1    0.6674 189.29 75.465
## + is_Crime         1    0.3860 189.57 75.622
## + runtime_minutes  1    0.3768 189.58 75.628
## + is_War           1    0.1886 189.77 75.733
## + is_Music         1    0.1302 189.83 75.765
## + is_Sci           1    0.1267 189.83 75.767
## + is_Romance       1    0.0798 189.88 75.793
## + is_Mystery       1    0.0775 189.88 75.795
## + is_Adventure     1    0.0422 189.92 75.814
## + is_Thriller      1    0.0078 189.95 75.834
## + is_Family        1    0.0049 189.96 75.835
## + is_History       1    0.0037 189.96 75.836
```

```
## + is_Biography     1    0.0013 189.96 75.837
##
## Step:  AIC=71.08
## log_gross ~ log_budget + movie_averageRating + is_Drama + is_Fantasy +
##     director_age_at_prod + production_year
##
##                   Df Sum of Sq    RSS    AIC
## <none>                         181.62 71.078
## + is_Comedy        1   2.76116 178.86 71.454
## + is_Animation     1   2.42503 179.19 71.653
## + is_Horror        1   1.73773 179.88 72.059
## + is_Musical       1   0.78773 180.83 72.617
## + is_Mystery       1   0.35181 181.27 72.872
## + runtime_minutes  1   0.32129 181.30 72.890
## + is_Sci           1   0.16792 181.45 72.980
## + is_History       1   0.15174 181.47 72.989
## + is_Music         1   0.12628 181.49 73.004
## + is_Western       1   0.09982 181.52 73.020
## + is_Adventure     1   0.05246 181.57 73.047
## + is_Biography     1   0.05110 181.57 73.048
## + is_Crime         1   0.04498 181.57 73.052
## + is_Romance       1   0.03726 181.58 73.056
## + is_Family        1   0.01447 181.60 73.069
## + is_Thriller      1   0.00840 181.61 73.073
## + is_War           1   0.00007 181.62 73.078
```

```r
summary(forward_model)
```

```
##
## Call:
## lm(formula = log_gross ~ log_budget + movie_averageRating + is_Drama +
##     is_Fantasy + director_age_at_prod + production_year, data = movies_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5341 -0.4799 -0.0085  0.8997  2.4942
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          41.752391  18.139863   2.302 0.023448 *
## log_budget            0.721868   0.119622   6.035 2.78e-08 ***
## movie_averageRating   0.522601   0.140088   3.731 0.000319 ***
## is_Drama             -0.945456   0.295528  -3.199 0.001852 **
## is_Fantasy            1.034404   0.455694   2.270 0.025378 *
## director_age_at_prod -0.032209   0.015071  -2.137 0.035056 *
## production_year      -0.018838   0.008834  -2.132 0.035447 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.354 on 99 degrees of freedom
## Multiple R-squared:  0.5209, Adjusted R-squared:  0.4919
## F-statistic: 17.94 on 6 and 99 DF,  p-value: 5.525e-14
```

```r
# stepwise AIC
empty_model = lm(log_gross ~ 1, data = movies_clean)


stepwise_model = step(
  empty_model,
  scope = formula(lm_full),
  direction = "both")
```

```
## Start:  AIC=137.09
## log_gross ~ 1
##
##                         Df Sum of Sq    RSS     AIC
## + log_budget             1   130.258 248.86  94.467
## + is_Drama               1    50.484 328.64 123.941
## + runtime_minutes        1    36.240 342.88 128.438
## + is_Fantasy             1    30.466 348.66 130.209
## + is_Adventure           1    28.274 350.85 130.873
## + director_age_at_prod   1    16.574 362.55 134.350
## + movie_averageRating    1    13.888 365.23 135.133
## + is_Biography           1     8.423 370.70 136.707
## <none>                               379.12 137.088
## + is_Sci                 1     6.179 372.94 137.347
## + production_year        1     6.074 373.05 137.376
## + is_Comedy              1     5.683 373.44 137.488
## + is_Family              1     5.635 373.49 137.501
## + is_Crime               1     4.281 374.84 137.885
## + is_Music               1     2.703 376.42 138.330
```

```
## + is_Musical            1     2.331 376.79 138.435
## + is_Horror             1     2.139 376.98 138.489
## + is_Thriller           1     2.044 377.08 138.515
## + is_History            1     1.177 377.95 138.759
## + is_War                1     0.572 378.55 138.929
## + is_Romance            1     0.252 378.87 139.018
## + is_Mystery            1     0.166 378.96 139.042
## + is_Western            1     0.004 379.12 139.087
## + is_Animation          1     0.002 379.12 139.088
##
## Step:  AIC=94.47
## log_gross ~ log_budget
##
##                        Df Sum of Sq    RSS     AIC
## + movie_averageRating   1    20.769 228.09  87.230
## + is_Drama              1    13.354 235.51  90.621
## + director_age_at_prod  1    11.846 237.02  91.298
## + is_Comedy             1    10.604 238.26  91.852
## + production_year       1    10.583 238.28  91.861
## <none>                               248.86  94.467
## + is_Thriller           1     3.561 245.30  94.940
## + is_Fantasy            1     2.736 246.13  95.296
## + is_Animation          1     2.154 246.71  95.546
## + is_Mystery            1     2.079 246.79  95.578
## + is_History            1     1.200 247.66  95.955
## + runtime_minutes       1     0.618 248.25  96.204
## + is_Crime              1     0.604 248.26  96.210
```

```
## + is_Musical            1     0.370 248.49  96.310
## + is_Western            1     0.300 248.56  96.340
## + is_Music              1     0.114 248.75  96.419
## + is_War                1     0.111 248.75  96.420
## + is_Horror             1     0.106 248.76  96.422
## + is_Adventure          1     0.101 248.76  96.424
## + is_Biography          1     0.042 248.82  96.449
## + is_Sci                1     0.037 248.83  96.452
## + is_Romance            1     0.032 248.83  96.454
## + is_Family             1     0.001 248.86  96.467
## - log_budget            1   130.258 379.12 137.088
##
## Step:  AIC=87.23
## log_gross ~ log_budget + movie_averageRating
##
##                        Df Sum of Sq    RSS     AIC
## + is_Drama              1    25.508 202.59  76.659
## + director_age_at_prod  1    10.870 217.22  84.054
## + is_Comedy             1     9.478 218.62  84.731
## + is_Horror             1     6.608 221.49  86.114
## + is_Fantasy            1     6.268 221.83  86.277
## + production_year       1     4.744 223.35  87.002
## <none>                              228.09  87.230
## + is_Mystery            1     3.512 224.58  87.585
## + is_Animation          1     2.739 225.36  87.949
## + is_Crime              1     2.644 225.45  87.994
## + is_History            1     1.630 226.46  88.470
```

```
## + runtime_minutes       1     1.577 226.52  88.495
## + is_Thriller           1     1.566 226.53  88.500
## + is_Biography          1     0.708 227.39  88.901
## + is_Sci                1     0.324 227.77  89.079
## + is_Musical            1     0.222 227.87  89.127
## + is_War                1     0.127 227.97  89.171
## + is_Adventure          1     0.118 227.98  89.175
## + is_Romance            1     0.101 227.99  89.183
## + is_Family             1     0.088 228.01  89.189
## + is_Western            1     0.069 228.03  89.198
## + is_Music              1     0.025 228.07  89.219
## - movie_averageRating   1    20.769 248.86  94.467
## - log_budget            1   137.140 365.23 135.133
##
## Step:  AIC=76.66
## log_gross ~ log_budget + movie_averageRating + is_Drama
##
##                        Df Sum of Sq    RSS     AIC
## + is_Fantasy            1     6.295 196.29  75.313
## + director_age_at_prod  1     5.254 197.33  75.874
## + is_Animation          1     4.878 197.71  76.076
## + production_year       1     4.792 197.79  76.122
## <none>                              202.59  76.659
## + is_Comedy             1     1.845 200.74  77.689
## + is_Horror             1     1.229 201.36  78.014
## + is_Crime              1     1.035 201.55  78.116
## + runtime_minutes       1     0.775 201.81  78.253
```

```
## + is_Music              1     0.403 202.18  78.448
## + is_Musical            1     0.398 202.19  78.451
## + is_Mystery            1     0.286 202.30  78.509
## + is_Romance            1     0.086 202.50  78.614
## + is_Family             1     0.074 202.51  78.620
## + is_Biography          1     0.068 202.52  78.623
## + is_Adventure          1     0.061 202.53  78.627
## + is_Western            1     0.054 202.53  78.631
## + is_War                1     0.024 202.56  78.646
## + is_History            1     0.023 202.56  78.647
## + is_Sci                1     0.014 202.57  78.652
## + is_Thriller           1     0.000 202.59  78.659
## - is_Drama              1    25.508 228.09  87.230
## - movie_averageRating   1    32.924 235.51  90.621
## - log_budget            1    91.179 293.76 114.050
##
## Step:  AIC=75.31
## log_gross ~ log_budget + movie_averageRating + is_Drama + is_Fantasy
##
##                        Df Sum of Sq    RSS     AIC
## + director_age_at_prod  1     6.331 189.96  73.838
## + production_year       1     6.295 190.00  73.858
## + is_Animation          1     3.896 192.40  75.188
## <none>                              196.29  75.313
## + is_Comedy             1     2.711 193.58  75.839
## + is_Horror             1     1.511 194.78  76.494
## - is_Fantasy            1     6.295 202.59  76.659
```

```
## + is_Musical            1     0.692 195.60  76.939
## + is_Crime              1     0.628 195.66  76.974
## + is_Mystery            1     0.347 195.94  77.126
## + is_Music              1     0.345 195.95  77.127
## + is_Romance            1     0.270 196.02  77.167
## + is_Thriller           1     0.158 196.13  77.228
## + is_Adventure          1     0.112 196.18  77.253
## + is_Biography          1     0.083 196.21  77.268
## + is_Western            1     0.056 196.24  77.283
## + is_Sci                1     0.054 196.24  77.284
## + runtime_minutes       1     0.052 196.24  77.285
## + is_War                1     0.046 196.25  77.289
## + is_Family             1     0.032 196.26  77.296
## + is_History            1     0.001 196.29  77.313
## - is_Drama              1    25.535 221.83  86.277
## - movie_averageRating   1    37.009 233.30  91.622
## - log_budget            1    67.365 263.66 104.588
##
## Step:  AIC=73.84
## log_gross ~ log_budget + movie_averageRating + is_Drama + is_Fantasy +
##     director_age_at_prod
##
##                        Df Sum of Sq    RSS     AIC
## + production_year       1     8.342 181.62  71.078
## <none>                              189.96  73.838
## + is_Animation          1     3.469 186.49  73.885
## + is_Comedy             1     1.853 188.11  74.799
```

```
## + is_Horror             1     1.145 188.81  75.197
## - director_age_at_prod  1     6.331 196.29  75.313
## + is_Western            1     0.691 189.27  75.452
## + is_Musical            1     0.667 189.29  75.465
## + is_Crime              1     0.386 189.57  75.622
## + runtime_minutes       1     0.377 189.58  75.628
## + is_War                1     0.189 189.77  75.733
## + is_Music              1     0.130 189.83  75.765
## + is_Sci                1     0.127 189.83  75.767
## + is_Romance            1     0.080 189.88  75.793
## + is_Mystery            1     0.078 189.88  75.795
## + is_Adventure          1     0.042 189.92  75.814
## + is_Thriller           1     0.008 189.95  75.834
## + is_Family             1     0.005 189.96  75.835
## + is_History            1     0.004 189.96  75.836
## + is_Biography          1     0.001 189.96  75.837
## - is_Fantasy            1     7.372 197.33  75.874
## - is_Drama              1    19.478 209.44  82.185
## - movie_averageRating   1    34.628 224.59  89.588
## - log_budget            1    66.598 256.56 103.695
##
## Step:  AIC=71.08
## log_gross ~ log_budget + movie_averageRating + is_Drama + is_Fantasy +
##     director_age_at_prod + production_year
##
##                         Df Sum of Sq    RSS     AIC
## <none>                                181.62  71.078
```

```
## + is_Comedy             1     2.761 178.86  71.454
## + is_Animation          1     2.425 179.19  71.653
## + is_Horror             1     1.738 179.88  72.059
## + is_Musical            1     0.788 180.83  72.617
## + is_Mystery            1     0.352 181.27  72.872
## + runtime_minutes       1     0.321 181.30  72.890
## + is_Sci                1     0.168 181.45  72.980
## + is_History            1     0.152 181.47  72.989
## + is_Music              1     0.126 181.49  73.004
## + is_Western            1     0.100 181.52  73.020
## + is_Adventure          1     0.052 181.57  73.047
## + is_Biography          1     0.051 181.57  73.048
## + is_Crime              1     0.045 181.57  73.052
## + is_Romance            1     0.037 181.58  73.056
## + is_Family             1     0.014 181.60  73.069
## + is_Thriller           1     0.008 181.61  73.073
## + is_War                1     0.000 181.62  73.078
## - production_year       1     8.342 189.96  73.838
## - director_age_at_prod  1     8.378 190.00  73.858
## - is_Fantasy            1     9.453 191.07  74.456
## - is_Drama              1    18.776 200.40  79.506
## - movie_averageRating   1    25.531 207.15  83.020
## - log_budget            1    66.806 248.43 102.280
```

```
summary(stepwise_model)
```

```
##
## Call:
```

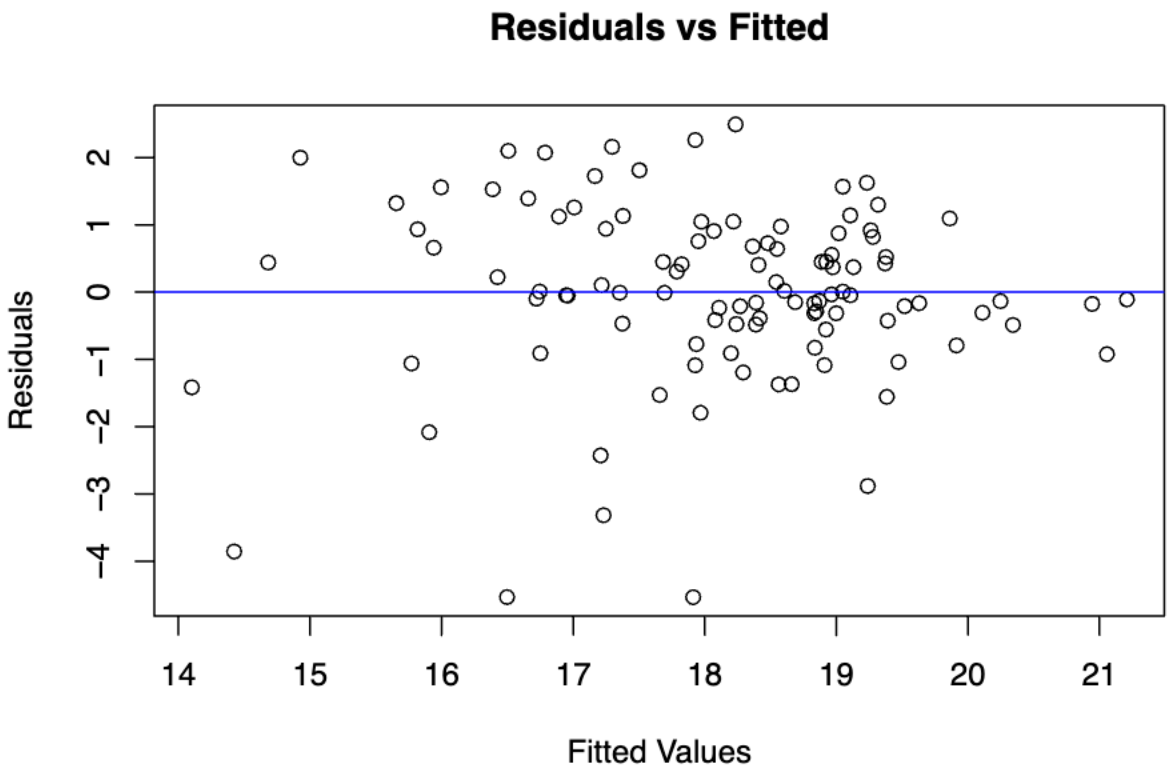
```
## lm(formula = log_gross ~ log_budget + movie_averageRating + is_Drama +
##     is_Fantasy + director_age_at_prod + production_year, data = movies_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5341 -0.4799 -0.0085  0.8997  2.4942
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           41.752391  18.139863   2.302 0.023448 *
## log_budget             0.721868   0.119622   6.035 2.78e-08 ***
## movie_averageRating    0.522601   0.140088   3.731 0.000319 ***
## is_Drama              -0.945456   0.295528  -3.199 0.001852 **
## is_Fantasy             1.034404   0.455694   2.270 0.025378 *
## director_age_at_prod -0.032209   0.015071  -2.137 0.035056 *
## production_year      -0.018838   0.008834  -2.132 0.035447 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.354 on 99 degrees of freedom
## Multiple R-squared:  0.5209, Adjusted R-squared:  0.4919
## F-statistic: 17.94 on 6 and 99 DF,  p-value: 5.525e-14
```

```
final_model = lm(log_gross ~ log_budget + movie_averageRating + is_Drama + is_Fantasy +
                   director_age_at_prod + production_year, data = movies_clean)
final_model
```

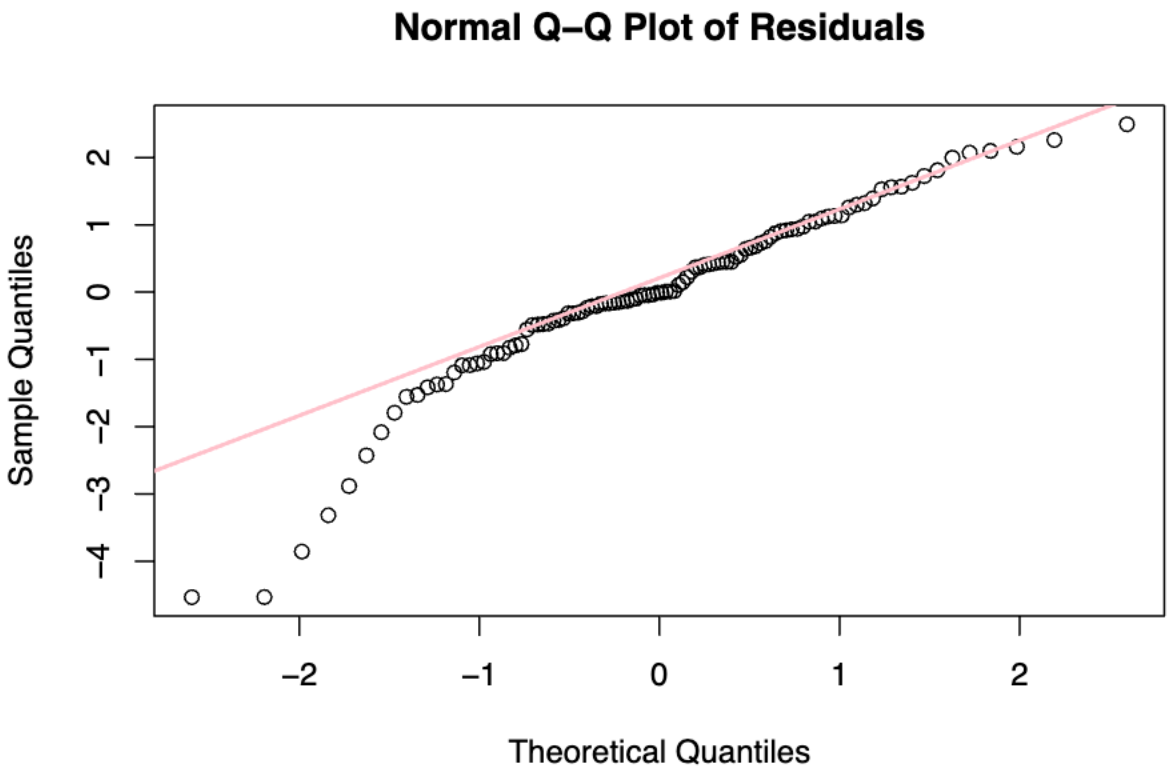
```
##
```

```
## Call:
## lm(formula = log_gross ~ log_budget + movie_averageRating + is_Drama +
##     is_Fantasy + director_age_at_prod + production_year, data = movies_clean)
##
## Coefficients:
##          (Intercept)            log_budget   movie_averageRating
##              41.75239               0.72187               0.52260
##             is_Drama            is_Fantasy  director_age_at_prod
##             -0.94546               1.03440              -0.03221
##      production_year
##             -0.01884
```

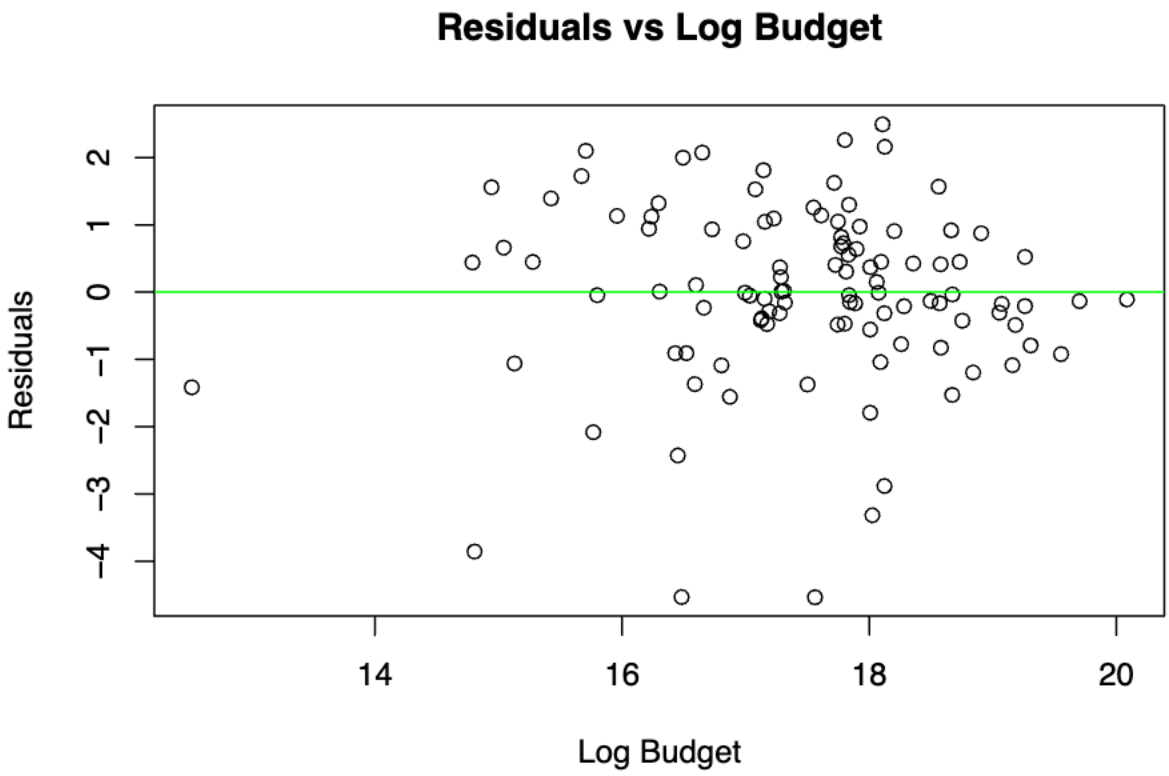
```r
# plots residuals vs fitted plot for final model
graphics::plot(final_model$fitted.values, resid(final_model),
  xlab = "Fitted Values", ylab = "Residuals", main = "Residuals vs Fitted")
abline(h = 0, col = "blue")
```



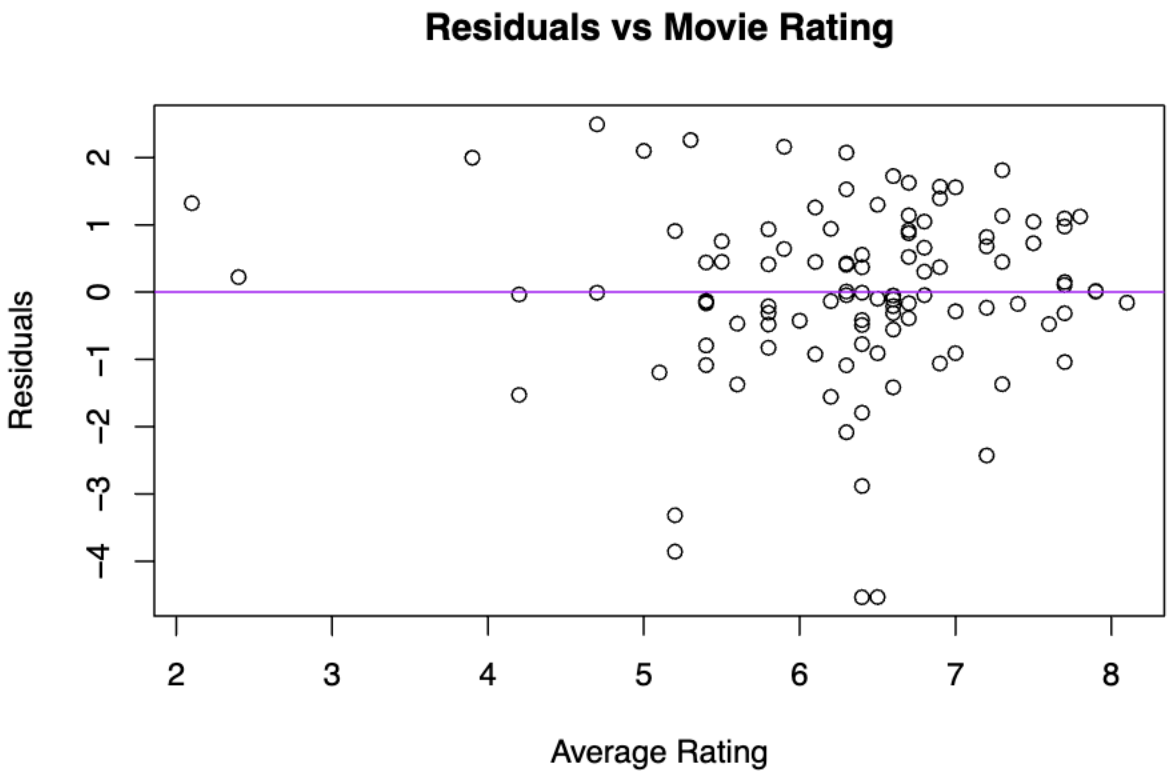

```r
# QQ plot for final model
qqnorm(resid(final_model), main = "Normal Q-Q Plot of Residuals")
qqline(resid(final_model), col = "pink", lwd = 2)
```




```r
# residuals vs log(budget) for final model
graphics::plot(movies_clean$log_budget,
     resid(final_model),
     xlab = "Log Budget",
     ylab = "Residuals",
     main = "Residuals vs Log Budget")
abline(h = 0, col = "green")
```

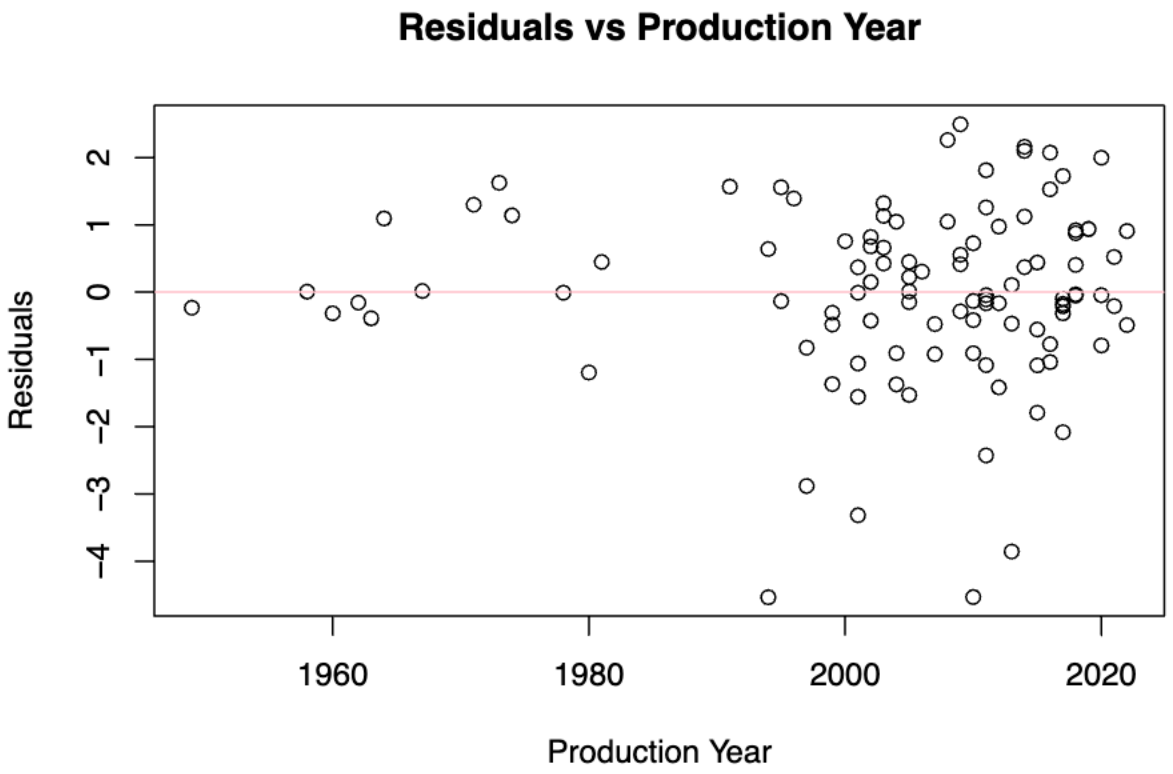



```r
# residuals vs movie_averageRating for final model
graphics::plot(movies_clean$movie_averageRating,
     resid(final_model),
     xlab = "Average Rating",
     ylab = "Residuals",
     main = "Residuals vs Movie Rating")
abline(h = 0, col = "purple")
```




```r
# residuals vs director_age_at_prod for final model
graphics::plot(movies_clean$director_age_at_prod,
      resid (final_model),
      xlab = "Director Age at Production",
      ylab = "Residuals",
      main = "Residuals vs Director Age")
abline(h = 0, col = "blue")
```


```r
# residuals vs production_year for final model
graphics::plot(movies_clean$production_year,
      resid (final_model),
      xlab = "Production Year",
      ylab = "Residuals",
      main = "Residuals vs Production Year")
abline(h = 0, col = "pink")
```




```
# predicting a new movie's revenue given data
# finds 1st director alphabetically
director1 = sort(unique(movies_clean$director_name))[1]
director1_birth_year = movies_clean$birth_year[movies_clean$director_name == director1][
age = 2013 - director1_birth_year
budget_24 = adjust_for_inflation(10e6, from_date = 2013, to_date = 2024,
                                 country = "US")
```

```
## Retrieving countries data

## Generating URL to request all 296 results
## Retrieving inflation data for US
## Generating URL to request all 65 results
```

```r
new_movie = data.frame(log_budget = log(budget_24), movie_averageRating = 7,
  is_Drama = as.numeric(0), is_Fantasy = as.numeric(0), director_age_at_prod = age,
  production_year = 2013)



log_pred = predict(final_model, newdata = new_movie, interval = "prediction")
exp(log_pred)
```

```
##        fit     lwr       upr
## 1 46937901 2961322 743980640
```