Modeling Movie Gross Revenue

Manasi Singh

12/16/2025

Introduction

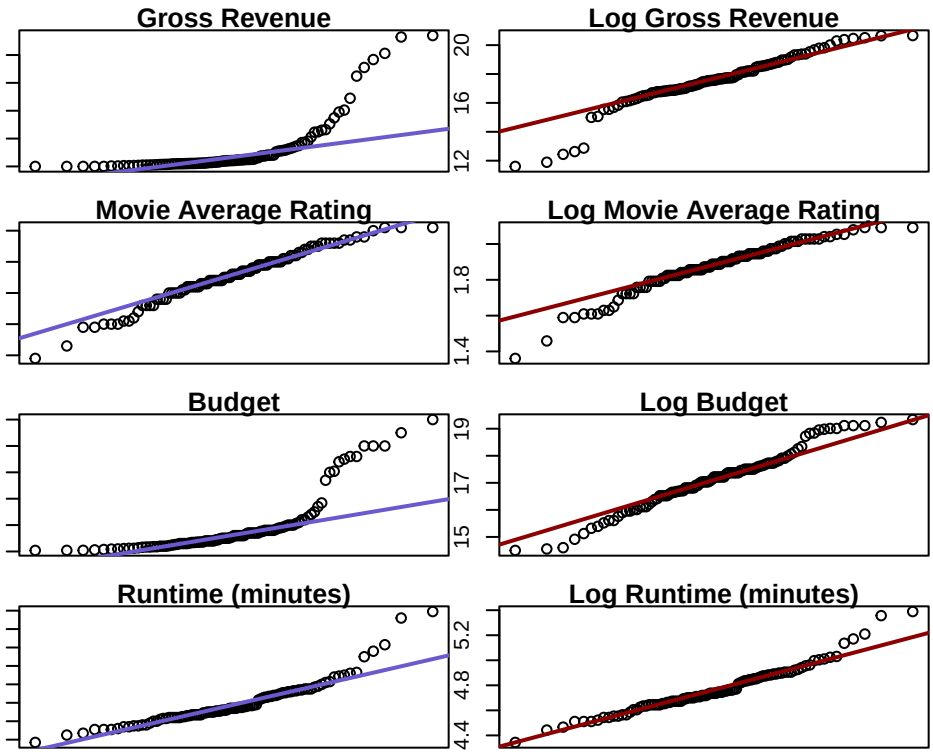
Cinema is a major industry in the United States, where individual films can generate hundreds of millions, and sometimes even billions, of dollars in revenue. At the same time, producing a major motion picture is expensive, with production budgets often reaching tens or hundreds of millions of dollars. Because of these high costs, studio executives must make careful decisions about which projects to fund and how much money to invest in them.

In this paper, we build a regression model to explore which factors influence a film's gross revenue using data from approximately 100 motion pictures directed by 30 different directors. The model focuses on a small set of interpretable variables, including production budget, director name, average audience rating, and genre indicators for Drama, Thriller, and Romance. The goal is to balance modeling accuracy with simplicity so that the results are meaningful and accessible to non-statistical decision-makers.

Exploratory Data Analysis - EDA

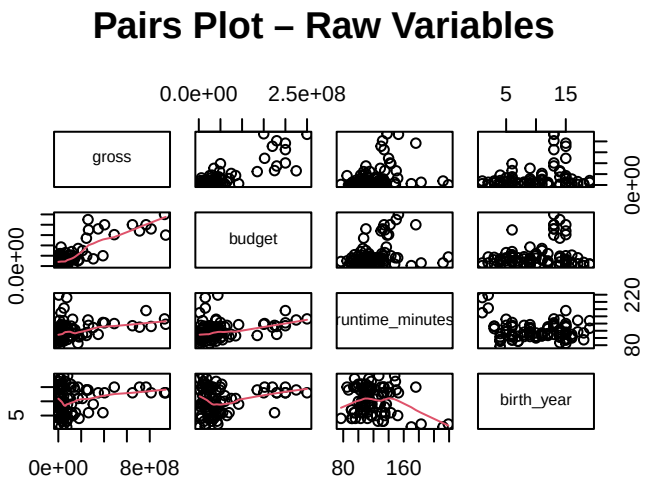
In our initial dataset, we are given information on movie title, production date, budget, director name, genre, director's birth year, runtime in minutes, and the movie's average rating, in addition to gross revenue. To construct an appropriate linear regression model, we first examined whether the assumptions of linearity were reasonable for the given data.

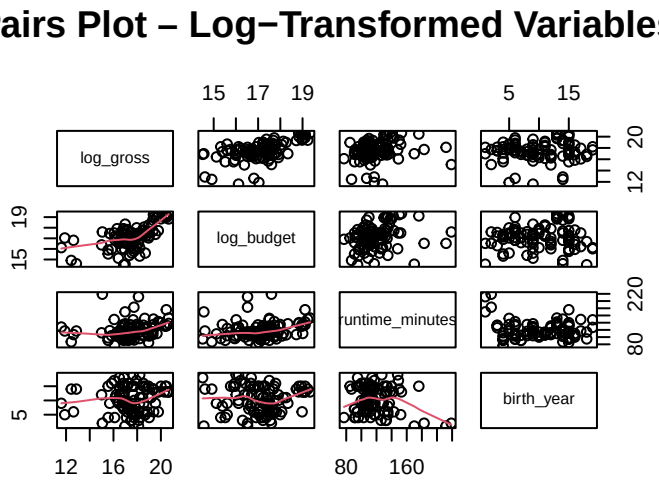
As an initial step, we performed distributional checks by comparing Q–Q plots of the raw variables to those of their log-transformed counterparts. This allowed us to identify skewed behavior in the data and assess whether transformations were necessary to better satisfy the assumptions of linear regression.



As shown in the plots above, Gross Revenue and Budget exhibit significant skewness, which is largely corrected by applying a log transformation. Runtime shows some curvature ("bowing") in its distribution, and while the log transformation provides slight improvement, the effect is minimal. This suggests that Gross Revenue and Budget should be log-transformed before inclusion in the model, whereas Average Rating and Runtime appear sufficiently linear and do not require transformation.

After assessing the linearity of individual variables, we performed pairwise comparisons for both the raw and log-transformed values to confirm the necessity of the log transformations for Gross Revenue and Budget.





In the pairwise plots, we see much less fanning and more consistent spread of the data after applying the log transformations. This is especially clear in plots involving Gross Revenue and Budget, which initially showed increasing variance as the values got larger. After the transformation, the points are more evenly spread, which better satisfies the assumptions of linear regression, such as constant variance and linear relationships. This confirms that Gross Revenue and Budget should be log-transformed, while Runtime and Average Rating do not need any transformation.

Overall, the EDA shows that Gross Revenue and Budget are highly skewed and benefit from log transformation, which helps reduce fanning and stabilizes variance in the pairwise plots. Runtime and Average Rating, on the other hand, appear fairly linear and do not require transformation. These observations suggest that applying log transformations to Gross Revenue and Budget before including them in the model will help satisfy the assumptions of linear regression and improve model performance.

Model Building

To begin the model building, we start with the following baseline model: log(gross) $\sim$ log(budget)+runtime+birth year From there, we compared several slightly modified models by adding one parameter at a time to evaluate improvements in model fit. Specifically, we compared the baseline model to four separate models that included genre indicators, director effects, average rating, and production year. Finally, we examined a full model that combined all baseline and additional parameters to determine which combination provided the best fit.

Using AIC, BIC, and likelihood ratio tests, we determined that the full model provides the best fit, as it has the lowest AIC value. After concluding this, we did three stepwise tests, to confirm that the full model is the best out of all possible combinations of the given parameters. The stepwise tests include a forwards stepwise test, a backwards stepwise test, and bi-directional stepwise test.

As seen above, the AIC test confirms that the backward stepwise procedure produced the best model by a slight margin, since it has the lowest AIC. The backward stepwise selection determined that not all of the top 10 genres were necessary, reducing them to the three most influential genres. The resulting main-effects model includes budget, average movie rating, the selected genres (Drama, Thriller, Romance), and director effects:

$$\log(\text{gross}) = \beta_0 + \beta_1 \log(\text{budget}) + \beta_2 \text{movie\_averageRating} + \beta_3 \text{Drama} + \beta_4 \text{Thriller} + \beta_5 \text{Romance} + \sum_{i=1}^{n} \gamma_i \text{Director}_i$$

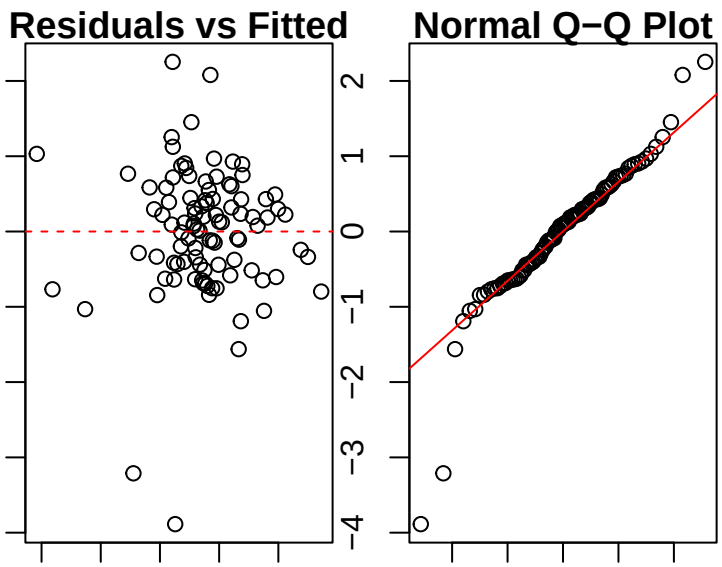
To further validate the model, we considered potential interactions among the categorical variables. This resulted in the following interaction model:

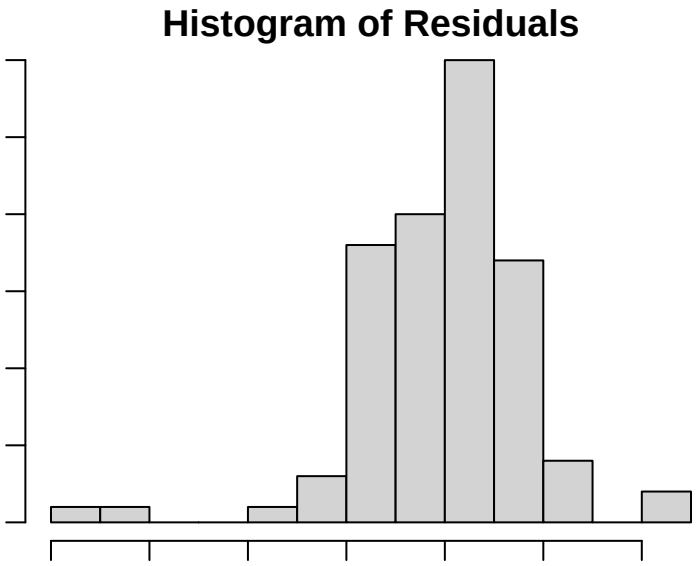
$$\begin{aligned}\log(\text{gross}) = {} & \beta_0 + \beta_1 \log(\text{budget}) + \beta_2 \text{movie\_averageRating} + \beta_3 \text{Drama} + \beta_4 \text{Thriller} + \beta_5 \text{Romance} \\ & + \beta_6(\log(\text{budget}) \times \text{Drama}) + \beta_7(\log(\text{budget}) \times \text{Thriller}) + \beta_8(\log(\text{budget}) \times \text{Romance}) \\ & + \beta_9(\text{movie\_averageRating} \times \text{Drama}) + \beta_{10}(\text{movie\_averageRating} \times \text{Thriller}) + \beta_{11}(\text{movie\_a} \\ & + \sum_{i=1}^{n} \gamma_i \text{Director}_i\end{aligned}$$

Comparing the AIC values for the main-effects and interaction models shows that the interaction terms do not significantly improve the model. Therefore, the main-effects mode, the original backward stepwise model, is concluded to be the best model for these data.

Model Verification

To verify the validity of the model, we plot a number of qq-plots to indicate linearity and show the fit.





As shown in the plots above, the model residuals exhibit approximate normality, constant variance, and linearity, indicating that the assumptions of linear regression are reasonably satisfied.

Model Interpretation

As mentioned before the final model was

$\log(\text{gross}) = \beta_0 + \beta_1 \log(\text{budget}) + \beta_2 \text{movie\_averageRating} + \beta_3 \text{Drama} + \beta_4 \text{Thriller} + \beta_5 \text{Romance} + \sum_{i=1}^{n} \gamma_i \text{Director}_i$

Now we can add in all of our coefficients, we found them by using the summary tool. After adding in the coefficients we have the following:

$$
\begin{aligned}
\log(\text{gross}) = &-0.24155 + 0.64553 \cdot \log(\text{budget}) \\
&- 0.77815 \cdot \mathbf{1}_{\text{Barbra Streisand}} - 0.87162 \cdot \mathbf{1}_{\text{Bennett Miller}} - 1.12104 \cdot \mathbf{1}_{\text{Billy Bob Thornton}} \\
&- 0.98183 \cdot \mathbf{1}_{\text{Boaz Yakin}} + 0.10757 \cdot \mathbf{1}_{\text{David E. Talbert}} + 0.13426 \cdot \mathbf{1}_{\text{David Yates}} \\
&+ 0.30288 \cdot \mathbf{1}_{\text{Dennie Gordon}} - 0.03747 \cdot \mathbf{1}_{\text{Forest Whitaker}} - 0.51406 \cdot \mathbf{1}_{\text{Gore Verbinski}} \\
&- 0.09791 \cdot \mathbf{1}_{\text{Hugh Wilson}} - 0.26540 \cdot \mathbf{1}_{\text{Jay Chandrasekhar}} - 4.76517 \cdot \mathbf{1}_{\text{Jeb Stuart}} \\
&- 1.43626 \cdot \mathbf{1}_{\text{John Crowley}} - 0.46926 \cdot \mathbf{1}_{\text{John Lee Hancock}} + 0.04088 \cdot \mathbf{1}_{\text{Jonathan Frakes}} \\
&- 0.94842 \cdot \mathbf{1}_{\text{Kevin Reynolds}} - 0.38707 \cdot \mathbf{1}_{\text{Matt Reeves}} - 2.42230 \cdot \mathbf{1}_{\text{Michael Cimino}} \\
&- 3.24487 \cdot \mathbf{1}_{\text{Michael Landon Jr.}} - 0.04800 \cdot \mathbf{1}_{\text{Michael Tollin}} - 0.60891 \cdot \mathbf{1}_{\text{Nicholas Hytner}} \\
&- 2.03208 \cdot \mathbf{1}_{\text{Richard Eyre}} - 2.26115 \cdot \mathbf{1}_{\text{Rod Lurie}} - 0.74596 \cdot \mathbf{1}_{\text{Rodrigo Cortés}} \\
&- 0.64029 \cdot \mathbf{1}_{\text{Sean McNamara}} - 0.29683 \cdot \mathbf{1}_{\text{Stephen Herek}} + 0.90339 \cdot \mathbf{1}_{\text{Tarsem Singh}} \\
&- 1.54847 \cdot \mathbf{1}_{\text{Taylor Hackford}} - 1.41649 \cdot \mathbf{1}_{\text{William Wyler}} \\
&+ 1.19373 \cdot \text{movie\_averageRating} - 0.78721 \cdot \text{Drama} \\
&+ 1.02090 \cdot \text{Thriller} + 0.57461 \cdot \text{Romance}
\end{aligned}
$$

Given the number of coefficients in the model, a general explanation of their meaning is provided below. All coefficients other than the intercept represent the expected change in the log of gross revenue associated with a one-unit increase in the corresponding predictor, holding all other variables constant. For example, the coefficient of log(budget) is 0.64553. This means that a 1-unit increase in the log of the budget is associated with an expected increase of 0.64553 in the log of gross revenue. Similarly, the coefficient of movie_averageRating is 1.19373, meaning that a 1-unit increase in the average rating is associated with an expected increase of 1.19373 in the log of gross revenue. For categorical variables such as directors or genres, the coefficients indicate the expected difference in log gross revenue compared to the reference category. For instance, if the director is Jeb Stuart, the expected log gross revenue decreases by 4.76517 compared to the reference director. In short, each coefficient quantifies the expected effect of its predictor on the log of gross revenue, controlling for all other variables in the model.

Prediction

```
##        1
## 18.51928
```

```
##         1
## 110362748
```

```
##         fit     lwr        upr
## 1 110362748 7536100 1616212102
```

In this section, we use the previously fitted regression model to predict the gross revenue of a new film based on its characteristics. The film has a production date of June 15, 2013, a runtime of 2 hours, an average rating of 7.0, belongs to the Comedy genre, has a budget of 10 million dollars, and is directed by the first director alphabetically in our dataset. Using these values, the model predicts a gross revenue of $110,362,748$ dollars. To account for uncertainty in the prediction, we also calculate a 95% prediction interval, which is $(7,536,100,\ 1,616,212,102)$ dollars. Our predicted value falls well within this range, reflecting the variability captured by the model.

Conclusion

Our model suggests that increasing the budget is generally associated with higher gross revenue, so spending more tends to lead to making more. However, this estimate isn't perfect. Other factors like marketing spend, star power, or a director's previous successes could also influence revenue, and our current model doesn't include them. To confidently say whether a bigger budget is truly justified, we'd need data on these additional factors to account for their effects and reduce bias. For now, the predicted boost in revenue from increasing the budget should be interpreted cautiously, as it may not fully capture the true impact.

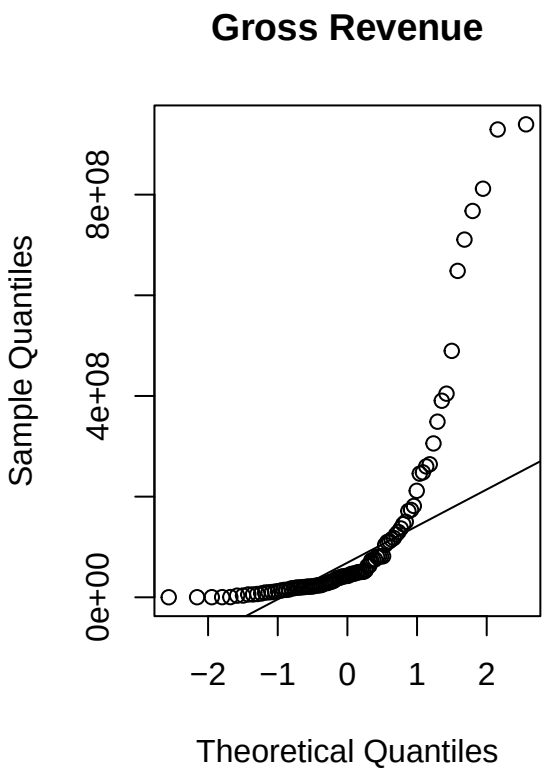
##Appendix

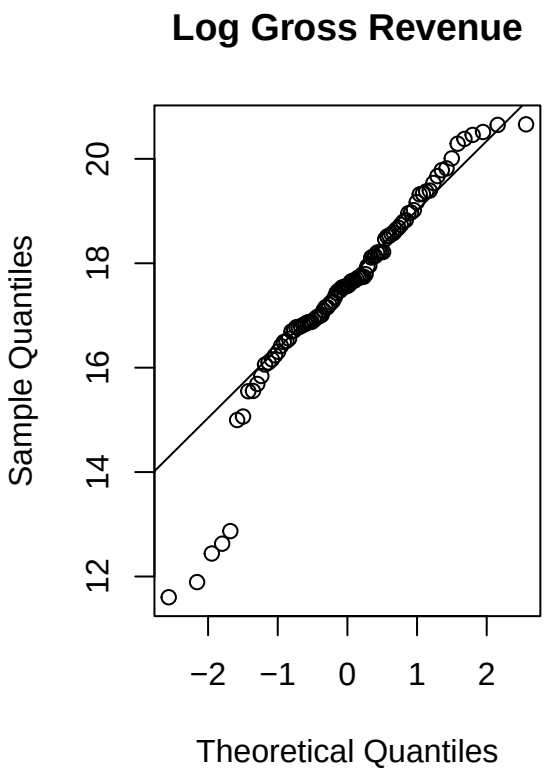
```
library(tidyverse)
library(stringr)
library(lubridate)
#Dataset
cinema <- read_csv("C:/Users/manas/Downloads/dataset42-1.csv")
```

```
## Rows: 97 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): movie_title, production_date, genres, director_name
## dbl (5): runtime_minutes, birth_year, movie_averageRating, budget, gross
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

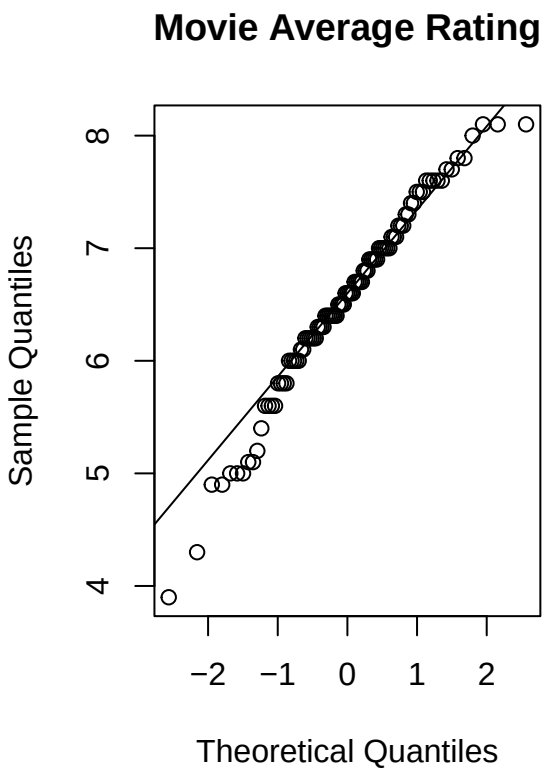
```
genre_counts <- cinema %>%
  separate_rows(genres, sep = ",") %>%
  mutate(genres = trimws(genres)) %>%
  count(genres, sort = TRUE) %>%
  slice_head(n = 10)

top_genres <- genre_counts$genres
```

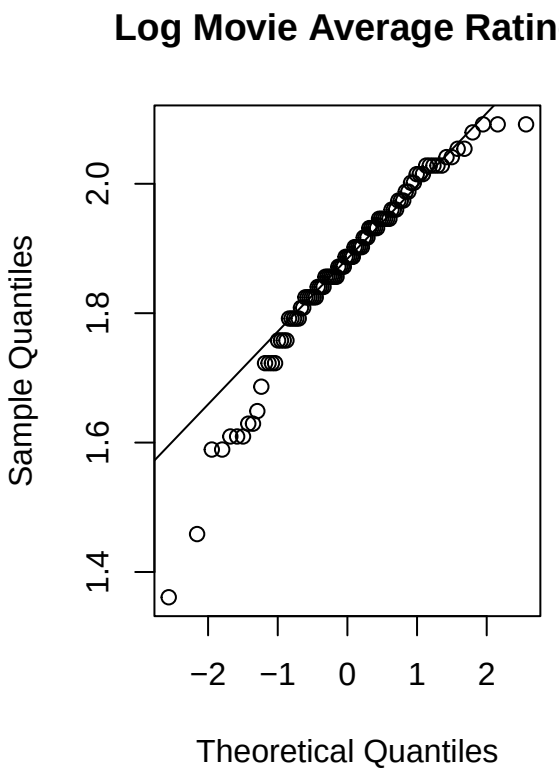
```r
cinema <- cinema %>%
  mutate(genres = trimws(genres)) %>%
  separate_rows(genres, sep = ",") %>%
  mutate(genres = trimws(genres)) %>%
  filter(genres %in% top_genres) %>%  # Keep only top 10
  mutate(value = 1L) %>%
  pivot_wider(names_from = genres, values_from = value, values_fill = 0)

if("Sci-Fi" %in% names(cinema)) {
  names(cinema)[names(cinema) == "Sci-Fi"] <- "SciFi"
}

cinema$production_year <- year(mdy(cinema$production_date))

cinema$director_name <- as.factor(cinema$director_name)
cinema$birth_year <- as.factor(cinema$birth_year)

##Checking for skewed data

par(mfrow = c(1, 2))
qqnorm(cinema$gross, main = "Gross Revenue"); qqline(cinema$gross)
qqnorm(log(cinema$gross), main = "Log Gross Revenue"); qqline(log(cinema$gross))
```



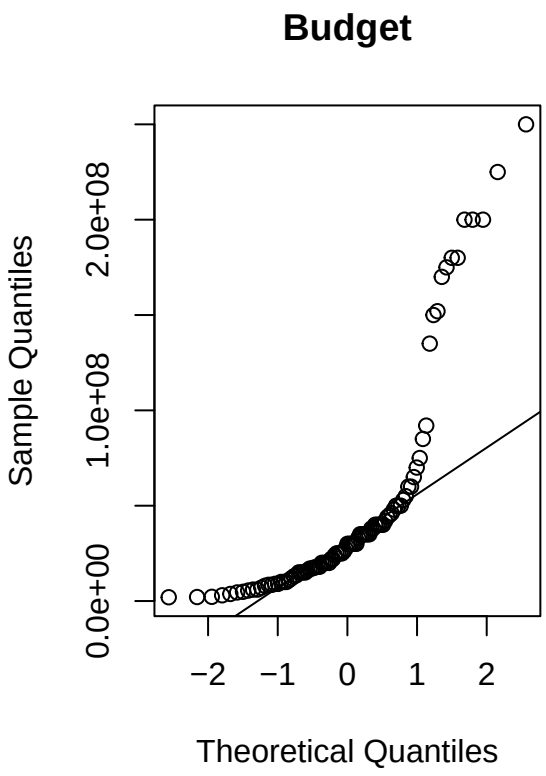


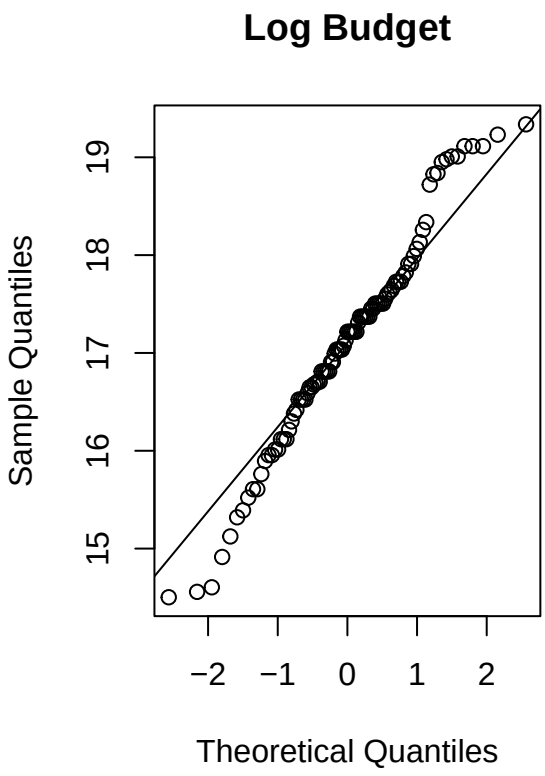

```r
par(mfrow = c(1, 2))
qqnorm(cinema$movie_averageRating, main = "Movie Average Rating")
qqline(cinema$movie_averageRating)
qqnorm(log(cinema$movie_averageRating), main = "Log Movie Average Rating")
qqline(log(cinema$movie_averageRating))
```



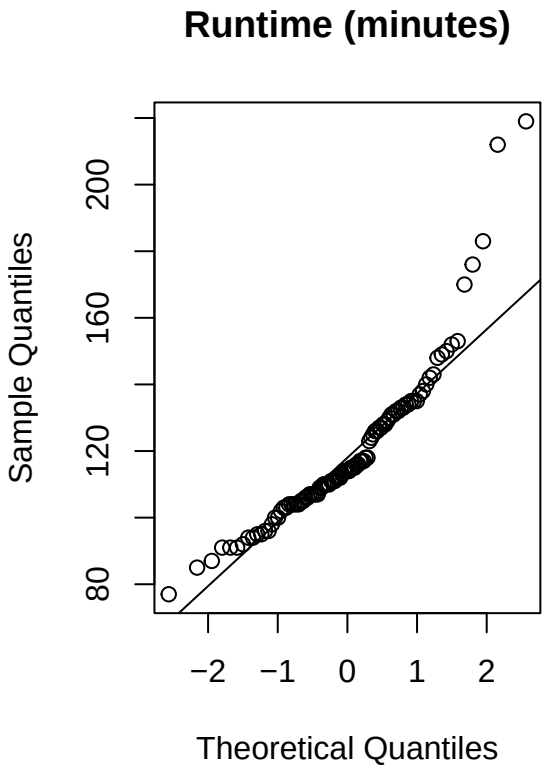


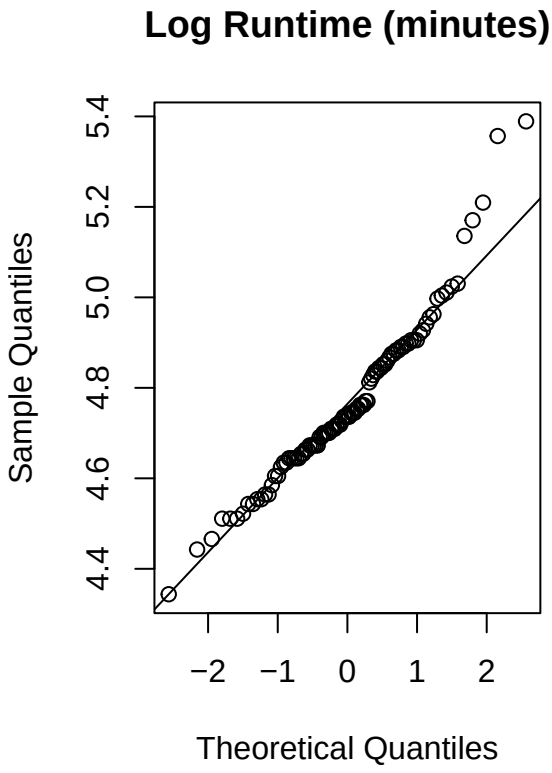

```r
par(mfrow = c(1, 2))
qqnorm(cinema$budget, main = "Budget"); qqline(cinema$budget)
qqnorm(log(cinema$budget), main = "Log Budget"); qqline(log(cinema$budget))
```



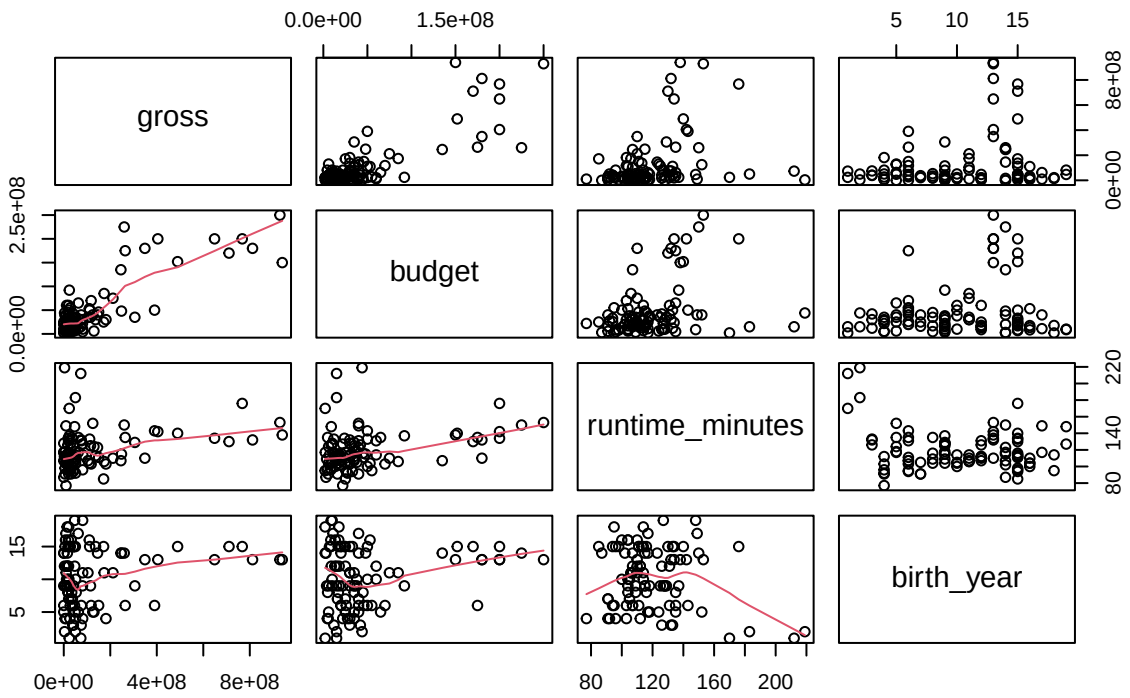



```
par(mfrow = c(1, 2))
qqnorm(cinema$runtime_minutes, main = "Runtime (minutes)"); qqline(cinema$runtime_minute
qqnorm(log(cinema$runtime_minutes),  main = "Log Runtime (minutes)"); qqline(log(cinema$
```

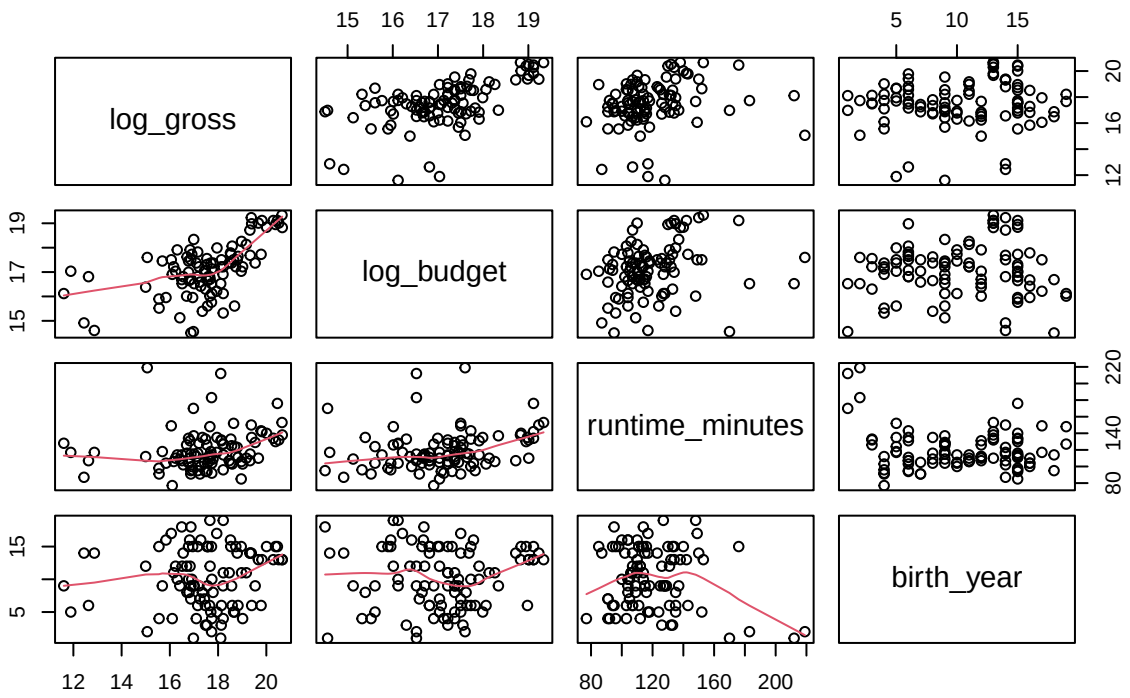





```
## Model Internal Comparison

pairs(
  cinema[, c("gross", "budget", "runtime_minutes", "birth_year")],
  lower.panel = panel.smooth,
  main = "Pairs Plot - Raw Variables"
)
```




```r
pairs(
  data.frame(
    log_gross  = log(cinema$gross),
    log_budget = log(cinema$budget),
    runtime_minutes = cinema$runtime_minutes,
    birth_year = cinema$birth_year
  ),
  lower.panel = panel.smooth,
  main = "Pairs Plot - Log-Transformed Variables"
)
```




```
## Model Comparison
m_base <- lm(
  log(gross) ~ log(budget) + runtime_minutes + birth_year,
  data = cinema
)
genre_vars <- intersect(names(cinema)[sapply(cinema, function(x) all(x %in% c(0,1)))],

m_genre <- lm(
  as.formula(
    paste(
      "log(gross) ~ log(budget) + runtime_minutes + birth_year +",
      paste(genre_vars, collapse = " + ")
    )
  ),
  data = cinema
)

m_dir <- lm(
  log(gross) ~ log(budget) + runtime_minutes + birth_year +
    factor(director_name),
  data = cinema
```

```
)

m_avgRat <- lm(
  log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating,
  data = cinema
)

m_prodYear <- lm(
  log(gross) ~ log(budget) + runtime_minutes + birth_year + production_year,
  data = cinema
)

m_full <- lm(
  as.formula(
    paste(
      "log(gross) ~ log(budget) + runtime_minutes + birth_year + factor(director_name) +
      paste(genre_vars, collapse = " + ")
    )
  ),
  data = cinema
)

m_full_noProd <- lm(
  as.formula(
    paste(
      "log(gross) ~ log(budget) + runtime_minutes + birth_year + factor(director_name) +
      paste(genre_vars, collapse = " + ")
    )
  ),
  data = cinema
)

AIC(m_base, m_genre, m_dir, m_avgRat, m_prodYear, m_full, m_full_noProd)
```

```
##                df      AIC
## m_base         22 371.5545
## m_genre        32 385.1740
## m_dir          33 356.5082
## m_avgRat       23 348.8285
## m_prodYear     23 370.3857
## m_full         45 328.8171
## m_full_noProd  44 327.6422
```

```r
BIC(m_base, m_genre, m_dir, m_avgRat, m_prodYear, m_full, m_full_noProd)
```

```
##                df      BIC
## m_base         22 428.1981
## m_genre        32 467.5648
## m_dir          33 441.4736
## m_avgRat       23 408.0469
## m_prodYear     23 429.6041
## m_full         45 444.6791
## m_full_noProd  44 440.9295
```

```r
anova(m_dir, m_full)
```

```
## Analysis of Variance Table
##
## Model 1: log(gross) ~ log(budget) + runtime_minutes + birth_year + factor(director_na
## Model 2: log(gross) ~ log(budget) + runtime_minutes + birth_year + factor(director_na
##     movie_averageRating + production_year + Action + Adventure +
##     Family + Crime + Drama + Comedy + Thriller + Biography +
##     Mystery + Romance
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     65 113.496
## 2     53  66.611 12    46.885 3.1087 0.002161 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m_base, m_genre)
```

```
## Analysis of Variance Table
##
## Model 1: log(gross) ~ log(budget) + runtime_minutes + birth_year
## Model 2: log(gross) ~ log(budget) + runtime_minutes + birth_year + Action +
##     Adventure + Family + Crime + Drama + Comedy + Thriller +
##     Biography + Mystery + Romance
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     76 166.28
## 2     66 155.70 10    10.586 0.4487 0.9163
```

```r
anova(m_base, m_dir)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: log(gross) ~ log(budget) + runtime_minutes + birth_year
## Model 2: log(gross) ~ log(budget) + runtime_minutes + birth_year + factor(director_na
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     76 166.28
## 2     65 113.50 11    52.786 2.7483 0.005537 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m_avgRat, m_base)
```

```
## Analysis of Variance Table
##
## Model 1: log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRatin
## Model 2: log(gross) ~ log(budget) + runtime_minutes + birth_year
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     75 128.87
## 2     76 166.28 -1   -37.415 21.776 1.312e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m_prodYear, m_base)
```

```
## Analysis of Variance Table
##
## Model 1: log(gross) ~ log(budget) + runtime_minutes + birth_year + production_year
## Model 2: log(gross) ~ log(budget) + runtime_minutes + birth_year
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     75 160.94
## 2     76 166.28 -1   -5.3443 2.4905 0.1187
```

```
anova(m_full, m_base)
```

```
## Analysis of Variance Table
##
## Model 1: log(gross) ~ log(budget) + runtime_minutes + birth_year + factor(director_na
##     movie_averageRating + production_year + Action + Adventure +
##     Family + Crime + Drama + Comedy + Thriller + Biography +
##     Mystery + Romance
## Model 2: log(gross) ~ log(budget) + runtime_minutes + birth_year
##   Res.Df     RSS  Df Sum of Sq     F    Pr(>F)
## 1     53  66.611
```

```
## 2     76 166.282 -23   -99.671 3.448 0.0001009 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m_full_noProd, m_full)
```

```
## Analysis of Variance Table
##
## Model 1: log(gross) ~ log(budget) + runtime_minutes + birth_year + factor(director_na
##     movie_averageRating + Action + Adventure + Family + Crime +
##     Drama + Comedy + Thriller + Biography + Mystery + Romance
## Model 2: log(gross) ~ log(budget) + runtime_minutes + birth_year + factor(director_na
##     movie_averageRating + production_year + Action + Adventure +
##     Family + Crime + Drama + Comedy + Thriller + Biography +
##     Mystery + Romance
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     54 67.180
## 2     53 66.611  1   0.56901 0.4527  0.504
```

```r
## Shows that the full model is the best model
# Forward selection: Start from m_base, add variables one at a time
step_forward <- step(m_base,
                     scope = list(lower = m_base, upper = m_full_noProd),
                     direction = "forward",
                     trace = TRUE)
```

```
## Start:  AIC=94.28
## log(gross) ~ log(budget) + runtime_minutes + birth_year
##
##                          Df Sum of Sq    RSS    AIC
## + movie_averageRating     1    37.415 128.87 71.554
## + factor(director_name) 11    52.786 113.50 79.234
## <none>                              166.28 94.280
## + Drama                   1     2.947 163.34 94.546
## + Action                  1     2.092 164.19 95.052
## + Adventure               1     1.095 165.19 95.640
## + Biography               1     0.603 165.68 95.928
## + Comedy                  1     0.322 165.96 96.092
## + Crime                   1     0.076 166.21 96.236
## + Romance                 1     0.064 166.22 96.243
## + Thriller                1     0.061 166.22 96.245
## + Family                  1     0.033 166.25 96.261
## + Mystery                 1     0.004 166.28 96.278
```

```
##
## Step:  AIC=71.55
## log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating
##
##                          Df Sum of Sq     RSS    AIC
## + factor(director_name) 11    47.266  81.601 49.231
## + Drama                  1    12.806 116.061 63.402
## + Comedy                 1     6.462 122.405 68.564
## <none>                               128.867 71.554
## + Romance                1     1.920 126.947 72.098
## + Biography              1     1.077 127.790 72.741
## + Adventure              1     0.567 128.299 73.127
## + Family                 1     0.441 128.425 73.222
## + Action                 1     0.159 128.708 73.435
## + Thriller               1     0.046 128.821 73.520
## + Crime                  1     0.010 128.856 73.547
## + Mystery                1     0.004 128.863 73.552
##
## Step:  AIC=49.23
## log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating +
##     factor(director_name)
##
##             Df Sum of Sq    RSS    AIC
## + Drama      1    8.1939 73.407 40.966
## + Thriller   1    6.5684 75.032 43.091
## + Biography  1    3.1558 78.445 47.405
## + Crime      1    2.2829 79.318 48.479
## <none>                   81.601 49.231
## + Action     1    1.1623 80.438 49.840
## + Comedy     1    0.7176 80.883 50.374
## + Mystery    1    0.6884 80.912 50.409
## + Romance    1    0.6332 80.967 50.475
## + Family     1    0.1901 81.410 51.005
## + Adventure  1    0.0143 81.586 51.214
##
## Step:  AIC=40.97
## log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating +
##     factor(director_name) + Drama
##
##             Df Sum of Sq    RSS    AIC
## + Thriller   1    3.5273 69.879 38.190
## + Biography  1    2.2985 71.108 39.881
## <none>                   73.407 40.966
## + Romance    1    1.4394 71.967 41.045
## + Action     1    0.8060 72.601 41.896
```

```
## + Comedy     1    0.2368 73.170 42.653
## + Mystery    1    0.2269 73.180 42.666
## + Crime      1    0.1111 73.295 42.819
## + Adventure  1    0.0294 73.377 42.928
## + Family     1    0.0122 73.394 42.950
##
## Step:  AIC=38.19
## log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating +
##     factor(director_name) + Drama + Thriller
##
##              Df Sum of Sq    RSS    AIC
## + Romance    1   1.87076 68.009 37.558
## <none>                   69.879 38.190
## + Biography  1   1.20167 68.678 38.507
## + Comedy     1   0.14969 69.730 39.982
## + Adventure  1   0.12123 69.758 40.021
## + Family     1   0.09132 69.788 40.063
## + Mystery    1   0.05873 69.821 40.108
## + Crime      1   0.00969 69.870 40.176
## + Action     1   0.00594 69.873 40.182
##
## Step:  AIC=37.56
## log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating +
##     factor(director_name) + Drama + Thriller + Romance
##
##              Df Sum of Sq    RSS    AIC
## <none>                   68.009 37.558
## + Biography  1   0.51862 67.490 38.815
## + Adventure  1   0.28149 67.727 39.155
## + Comedy     1   0.07143 67.937 39.456
## + Mystery    1   0.05398 67.955 39.481
## + Family     1   0.03431 67.974 39.509
## + Action     1   0.02187 67.987 39.526
## + Crime      1   0.01227 67.996 39.540
```

```r
# Backward elimination: Start from m_full_noProd, remove variables
step_backward <- step(m_full_noProd,
                      direction = "backward",
                      trace = TRUE)
```

```
## Start:  AIC=50.37
## log(gross) ~ log(budget) + runtime_minutes + birth_year + factor(director_name) +
##     movie_averageRating + Action + Adventure + Family + Crime +
##     Drama + Comedy + Thriller + Biography + Mystery + Romance
```

```
##
##
## Step:  AIC=50.37
## log(gross) ~ log(budget) + runtime_minutes + factor(director_name) +
##     movie_averageRating + Action + Adventure + Family + Crime +
##     Drama + Comedy + Thriller + Biography + Mystery + Romance
##
##                          Df Sum of Sq     RSS    AIC
## - Action                  1     0.001  67.180 48.369
## - Family                  1     0.004  67.184 48.375
## - Comedy                  1     0.005  67.184 48.375
## - Crime                   1     0.041  67.221 48.427
## - Mystery                 1     0.069  67.249 48.467
## - Adventure               1     0.181  67.361 48.629
## - Biography               1     0.347  67.526 48.867
## - runtime_minutes         1     1.032  68.212 49.847
## - Romance                 1     1.283  68.463 50.203
## <none>                                 67.180 50.368
## - Thriller                1     2.502  69.682 51.916
## - factor(director_name) 29    57.455 124.635 52.316
## - Drama                   1     4.721  71.900 54.955
## - log(budget)             1     6.264  73.444 57.016
## - movie_averageRating     1    28.780  95.960 82.954
##
## Step:  AIC=48.37
## log(gross) ~ log(budget) + runtime_minutes + factor(director_name) +
##     movie_averageRating + Adventure + Family + Crime + Drama +
##     Comedy + Thriller + Biography + Mystery + Romance
##
##                          Df Sum of Sq     RSS    AIC
## - Family                  1     0.004  67.185 46.375
## - Comedy                  1     0.005  67.186 46.377
## - Crime                   1     0.041  67.221 46.428
## - Mystery                 1     0.069  67.249 46.468
## - Adventure               1     0.185  67.365 46.635
## - Biography               1     0.355  67.536 46.881
## - runtime_minutes         1     1.051  68.231 47.874
## - Romance                 1     1.287  68.467 48.210
## <none>                                 67.180 48.369
## - Thriller                1     3.055  70.235 50.682
## - factor(director_name) 29    57.930 125.110 50.685
## - Drama                   1     4.731  71.911 52.970
## - log(budget)             1     7.114  74.295 56.133
## - movie_averageRating     1    28.985  96.166 81.162
##
```

```
## Step:  AIC=46.38
## log(gross) ~ log(budget) + runtime_minutes + factor(director_name) +
##     movie_averageRating + Adventure + Crime + Drama + Comedy +
##     Thriller + Biography + Mystery + Romance
##
##                          Df Sum of Sq     RSS    AIC
## - Comedy                  1     0.005  67.190 44.382
## - Crime                   1     0.045  67.230 44.440
## - Mystery                 1     0.068  67.252 44.473
## - Adventure               1     0.183  67.368 44.639
## - Biography               1     0.382  67.566 44.925
## - runtime_minutes         1     1.062  68.247 45.897
## - Romance                 1     1.292  68.476 46.222
## <none>                                 67.185 46.375
## - Thriller                1     3.066  70.251 48.704
## - factor(director_name) 29    58.456 125.641 49.095
## - Drama                   1     4.946  72.131 51.265
## - log(budget)             1     7.116  74.301 54.141
## - movie_averageRating     1    29.051  96.236 79.233
##
## Step:  AIC=44.38
## log(gross) ~ log(budget) + runtime_minutes + factor(director_name) +
##     movie_averageRating + Adventure + Crime + Drama + Thriller +
##     Biography + Mystery + Romance
##
##                          Df Sum of Sq     RSS    AIC
## - Crime                   1     0.052  67.241 42.457
## - Mystery                 1     0.075  67.264 42.490
## - Adventure               1     0.179  67.368 42.640
## - Biography               1     0.442  67.632 43.019
## - runtime_minutes         1     1.060  68.249 43.900
## - Romance                 1     1.289  68.479 44.226
## <none>                                 67.190 44.382
## - Thriller                1     3.064  70.254 46.708
## - Drama                   1     5.110  72.300 49.493
## - factor(director_name) 29    63.733 130.922 51.090
## - log(budget)             1     7.794  74.984 53.028
## - movie_averageRating     1    38.532 105.722 86.352
##
## Step:  AIC=42.46
## log(gross) ~ log(budget) + runtime_minutes + factor(director_name) +
##     movie_averageRating + Adventure + Drama + Thriller + Biography +
##     Mystery + Romance
##
##                          Df Sum of Sq     RSS    AIC
```

```
## - Mystery                1     0.049  67.291 40.528
## - Adventure              1     0.221  67.462 40.775
## - Biography              1     0.401  67.642 41.034
## - runtime_minutes        1     1.037  68.278 41.941
## - Romance                1     1.335  68.576 42.363
## <none>                               67.241 42.457
## - Thriller               1     3.013  70.254 44.708
## - Drama                  1     5.461  72.702 48.031
## - factor(director_name) 29    64.009 131.250 49.332
## - log(budget)            1     7.744  74.985 51.030
## - movie_averageRating    1    38.520 105.761 84.388
##
## Step:  AIC=40.53
## log(gross) ~ log(budget) + runtime_minutes + factor(director_name) +
##     movie_averageRating + Adventure + Drama + Thriller + Biography +
##     Romance
##
##                         Df Sum of Sq     RSS    AIC
## - Adventure              1     0.199  67.490 38.815
## - Biography              1     0.436  67.727 39.155
## - runtime_minutes        1     1.034  68.325 40.007
## - Romance                1     1.316  68.607 40.408
## <none>                               67.291 40.528
## - Thriller               1     3.095  70.386 42.890
## - Drama                  1     5.584  72.875 46.261
## - factor(director_name) 29    64.863 132.154 47.998
## - log(budget)            1     7.708  74.999 49.048
## - movie_averageRating    1    38.692 105.983 82.591
##
## Step:  AIC=38.82
## log(gross) ~ log(budget) + runtime_minutes + factor(director_name) +
##     movie_averageRating + Drama + Thriller + Biography + Romance
##
##                         Df Sum of Sq     RSS    AIC
## - Biography              1     0.519  68.009 37.558
## - runtime_minutes        1     1.031  68.521 38.285
## - Romance                1     1.188  68.678 38.507
## <none>                               67.490 38.815
## - Thriller               1     2.940  70.430 40.952
## - Drama                  1     5.553  73.043 44.485
## - factor(director_name) 29    64.754 132.244 46.064
## - log(budget)            1     8.634  76.124 48.492
## - movie_averageRating    1    38.625 106.115 80.712
##
## Step:  AIC=37.56
```

```
## log(gross) ~ log(budget) + runtime_minutes + factor(director_name) +
##     movie_averageRating + Drama + Thriller + Romance
##
##                          Df Sum of Sq     RSS    AIC
## - runtime_minutes         1     1.293  69.302 37.385
## <none>                                 68.009 37.558
## - Romance                 1     1.871  69.879 38.190
## - Thriller                1     3.959  71.967 41.045
## - Drama                   1     5.795  73.804 43.490
## - factor(director_name) 29    64.921 132.930 44.566
## - log(budget)             1     8.539  76.547 47.030
## - movie_averageRating     1    38.251 106.260 78.844
##
## Step:  AIC=37.38
## log(gross) ~ log(budget) + factor(director_name) + movie_averageRating +
##     Drama + Thriller + Romance
##
##                          Df Sum of Sq     RSS    AIC
## <none>                                 69.302 37.385
## - Romance                 1     1.490  70.792 37.448
## - Thriller                1     3.396  72.698 40.026
## - Drama                   1     5.172  74.474 42.367
## - factor(director_name) 29    67.032 136.334 45.019
## - log(budget)             1    18.284  87.586 58.097
## - movie_averageRating     1    45.326 114.627 84.197
```

```
# Both directions (most thorough)
step_both <- step(m_base,
                  scope = list(lower = m_base, upper = m_full_noProd),
                  direction = "both",
                  trace = TRUE)
```

```
## Start:  AIC=94.28
## log(gross) ~ log(budget) + runtime_minutes + birth_year
##
##                          Df Sum of Sq    RSS    AIC
## + movie_averageRating     1    37.415 128.87 71.554
## + factor(director_name) 11    52.786 113.50 79.234
## <none>                               166.28 94.280
## + Drama                   1     2.947 163.34 94.546
## + Action                  1     2.092 164.19 95.052
## + Adventure               1     1.095 165.19 95.640
## + Biography               1     0.603 165.68 95.928
## + Comedy                  1     0.322 165.96 96.092
```

```
## + Crime                   1     0.076 166.21 96.236
## + Romance                 1     0.064 166.22 96.243
## + Thriller                1     0.061 166.22 96.245
## + Family                  1     0.033 166.25 96.261
## + Mystery                 1     0.004 166.28 96.278
##
## Step:  AIC=71.55
## log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating
##
##                          Df Sum of Sq     RSS    AIC
## + factor(director_name) 11    47.266  81.601 49.231
## + Drama                   1    12.806 116.061 63.402
## + Comedy                  1     6.462 122.405 68.564
## <none>                                128.867 71.554
## + Romance                 1     1.920 126.947 72.098
## + Biography               1     1.077 127.790 72.741
## + Adventure               1     0.567 128.299 73.127
## + Family                  1     0.441 128.425 73.222
## + Action                  1     0.159 128.708 73.435
## + Thriller                1     0.046 128.821 73.520
## + Crime                   1     0.010 128.856 73.547
## + Mystery                 1     0.004 128.863 73.552
## - movie_averageRating     1    37.415 166.282 94.280
##
## Step:  AIC=49.23
## log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating +
##     factor(director_name)
##
##                          Df Sum of Sq     RSS    AIC
## + Drama                   1     8.194  73.407 40.966
## + Thriller                1     6.568  75.032 43.091
## + Biography               1     3.156  78.445 47.405
## + Crime                   1     2.283  79.318 48.479
## <none>                                81.601 49.231
## + Action                  1     1.162  80.438 49.840
## + Comedy                  1     0.718  80.883 50.374
## + Mystery                 1     0.688  80.912 50.409
## + Romance                 1     0.633  80.967 50.475
## + Family                  1     0.190  81.410 51.005
## + Adventure               1     0.014  81.586 51.214
## - factor(director_name) 11    47.266 128.867 71.554
## - movie_averageRating     1    31.895 113.496 79.234
##
## Step:  AIC=40.97
## log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating +
```

```
##     factor(director_name) + Drama
##
##                          Df Sum of Sq     RSS    AIC
## + Thriller                1     3.527  69.879 38.190
## + Biography               1     2.298  71.108 39.881
## <none>                                 73.407 40.966
## + Romance                 1     1.439  71.967 41.045
## + Action                  1     0.806  72.601 41.896
## + Comedy                  1     0.237  73.170 42.653
## + Mystery                 1     0.227  73.180 42.666
## + Crime                   1     0.111  73.295 42.819
## + Adventure               1     0.029  73.377 42.928
## + Family                  1     0.012  73.394 42.950
## - Drama                   1     8.194  81.601 49.231
## - factor(director_name) 11    42.654 116.061 63.402
## - movie_averageRating     1    38.439 111.846 79.814
##
## Step:  AIC=38.19
## log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating +
##     factor(director_name) + Drama + Thriller
##
##                          Df Sum of Sq     RSS    AIC
## + Romance                 1     1.871  68.009 37.558
## <none>                                 69.879 38.190
## + Biography               1     1.202  68.678 38.507
## + Comedy                  1     0.150  69.730 39.982
## + Adventure               1     0.121  69.758 40.021
## + Family                  1     0.091  69.788 40.063
## + Mystery                 1     0.059  69.821 40.108
## + Crime                   1     0.010  69.870 40.176
## + Action                  1     0.006  69.873 40.182
## - Thriller                1     3.527  73.407 40.966
## - Drama                   1     5.153  75.032 43.091
## - factor(director_name) 11    45.511 115.390 64.840
## - movie_averageRating     1    36.503 106.383 76.956
##
## Step:  AIC=37.56
## log(gross) ~ log(budget) + runtime_minutes + birth_year + movie_averageRating +
##     factor(director_name) + Drama + Thriller + Romance
##
##                          Df Sum of Sq     RSS    AIC
## <none>                                 68.009 37.558
## - Romance                 1     1.871  69.879 38.190
## + Biography               1     0.519  67.490 38.815
## + Adventure               1     0.281  67.727 39.155
```

```
## + Comedy                  1     0.071  67.937 39.456
## + Mystery                 1     0.054  67.955 39.481
## + Family                  1     0.034  67.974 39.509
## + Action                  1     0.022  67.987 39.526
## + Crime                   1     0.012  67.996 39.540
## - Thriller                1     3.959  71.967 41.045
## - Drama                   1     5.795  73.804 43.490
## - factor(director_name) 11    43.641 111.650 63.643
## - movie_averageRating     1    38.251 106.260 78.844
```

```r
# Compare results
summary(step_forward)
```

```
##
## Call:
## lm(formula = log(gross) ~ log(budget) + runtime_minutes + birth_year +
##     movie_averageRating + factor(director_name) + Drama + Thriller +
##     Romance, data = cinema)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7669 -0.4795  0.0626  0.5046  2.1563
##
## Coefficients: (18 not defined because of singularities)
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -1.71043    3.17702  -0.538 0.592277
## log(budget)                            0.53071    0.19177   2.767 0.007470
## runtime_minutes                        0.01207    0.01121   1.077 0.285709
## birth_year1939                        -0.95309    1.11473  -0.855 0.395900
## birth_year1942                         1.42291    1.27153   1.119 0.267506
## birth_year1943                         0.58722    1.56500   0.375 0.708798
## birth_year1944                         0.68170    1.23933   0.550 0.584287
## birth_year1952                         1.45101    1.35176   1.073 0.287310
## birth_year1953                         2.92079    1.58587   1.842 0.070373
## birth_year1955                         2.43695    1.51466   1.609 0.112802
## birth_year1956                         1.73362    1.31549   1.318 0.192481
## birth_year1958                         2.24535    1.50007   1.497 0.139595
## birth_year1961                         3.50153    1.56917   2.231 0.029337
## birth_year1962                         1.79913    1.48935   1.208 0.231709
## birth_year1963                         2.55774    1.42895   1.790 0.078426
## birth_year1964                        -0.91225    1.45741  -0.626 0.533690
## birth_year1966                         2.03741    1.33962   1.521 0.133455
## birth_year1968                         2.18062    1.44542   1.509 0.136553
## birth_year1969                         0.76738    1.31497   0.584 0.561661
```

```
## birth_year1973                              1.62824    1.47218   1.106 0.273069
## birth_year1986                              2.11002    1.28133   1.647 0.104754
## movie_averageRating                         1.13789    0.19427   5.857 2.03e-07
## factor(director_name)Barbra Streisand            NA         NA      NA       NA
## factor(director_name)Bennett Miller        -0.57921    0.78234  -0.740 0.461927
## factor(director_name)Billy Bob Thornton    -1.30707    1.09403  -1.195 0.236814
## factor(director_name)Boaz Yakin            -0.47082    0.79514  -0.592 0.555958
## factor(director_name)David E. Talbert       0.42575    0.96869   0.440 0.661846
## factor(director_name)David Yates                 NA         NA      NA       NA
## factor(director_name)Dennie Gordon               NA         NA      NA       NA
## factor(director_name)Forest Whitaker       -1.13549    1.09096  -1.041 0.302070
## factor(director_name)Gore Verbinski         2.87563    1.09561   2.625 0.010945
## factor(director_name)Hugh Wilson            1.83751    1.00743   1.824 0.073058
## factor(director_name)Jay Chandrasekhar           NA         NA      NA       NA
## factor(director_name)Jeb Stuart            -4.24654    1.05460  -4.027 0.000159
## factor(director_name)John Crowley                NA         NA      NA       NA
## factor(director_name)John Lee Hancock       0.16039    0.82690   0.194 0.846845
## factor(director_name)Jonathan Frakes        1.21113    0.76625   1.581 0.119142
## factor(director_name)Kevin Reynolds              NA         NA      NA       NA
## factor(director_name)Matt Reeves                 NA         NA      NA       NA
## factor(director_name)Michael Cimino              NA         NA      NA       NA
## factor(director_name)Michael Landon Jr.          NA         NA      NA       NA
## factor(director_name)Michael Tollin              NA         NA      NA       NA
## factor(director_name)Nicholas Hytner             NA         NA      NA       NA
## factor(director_name)Richard Eyre                NA         NA      NA       NA
## factor(director_name)Rod Lurie             -1.75264    1.05641  -1.659 0.102239
## factor(director_name)Rodrigo Cortés              NA         NA      NA       NA
## factor(director_name)Sean McNamara               NA         NA      NA       NA
## factor(director_name)Stephen Herek               NA         NA      NA       NA
## factor(director_name)Tarsem Singh                NA         NA      NA       NA
## factor(director_name)Taylor Hackford             NA         NA      NA       NA
## factor(director_name)William Wyler               NA         NA      NA       NA
## Drama                                      -0.84106    0.36889  -2.280 0.026119
## Thriller                                    1.11422    0.59131   1.884 0.064290
## Romance                                     0.65028    0.50201   1.295 0.200075
##
## (Intercept)
## log(budget)                                **
## runtime_minutes
## birth_year1939
## birth_year1942
## birth_year1943
## birth_year1944
## birth_year1952
## birth_year1953                             .
```

```
## birth_year1955
## birth_year1956
## birth_year1958
## birth_year1961                          *
## birth_year1962
## birth_year1963                          .
## birth_year1964
## birth_year1966
## birth_year1968
## birth_year1969
## birth_year1973
## birth_year1986
## movie_averageRating                     ***
## factor(director_name)Barbra Streisand
## factor(director_name)Bennett Miller
## factor(director_name)Billy Bob Thornton
## factor(director_name)Boaz Yakin
## factor(director_name)David E. Talbert
## factor(director_name)David Yates
## factor(director_name)Dennie Gordon
## factor(director_name)Forest Whitaker
## factor(director_name)Gore Verbinski     *
## factor(director_name)Hugh Wilson        .
## factor(director_name)Jay Chandrasekhar
## factor(director_name)Jeb Stuart         ***
## factor(director_name)John Crowley
## factor(director_name)John Lee Hancock
## factor(director_name)Jonathan Frakes
## factor(director_name)Kevin Reynolds
## factor(director_name)Matt Reeves
## factor(director_name)Michael Cimino
## factor(director_name)Michael Landon Jr.
## factor(director_name)Michael Tollin
## factor(director_name)Nicholas Hytner
## factor(director_name)Richard Eyre
## factor(director_name)Rod Lurie
## factor(director_name)Rodrigo Cortés
## factor(director_name)Sean McNamara
## factor(director_name)Stephen Herek
## factor(director_name)Tarsem Singh
## factor(director_name)Taylor Hackford
## factor(director_name)William Wyler
## Drama                                   *
## Thriller                                .
## Romance
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 61 degrees of freedom
## Multiple R-squared:  0.7751, Adjusted R-squared:  0.6461
## F-statistic: 6.008 on 35 and 61 DF,  p-value: 6.186e-10
```

```
summary(step_backward)
```

```
##
## Call:
## lm(formula = log(gross) ~ log(budget) + factor(director_name) +
##     movie_averageRating + Drama + Thriller + Romance, data = cinema)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8866 -0.4418  0.0912  0.4454  2.2530
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -0.24155    3.07590  -0.079 0.937659
## log(budget)                           0.64553    0.15961   4.044 0.000148
## factor(director_name)Barbra Streisand -0.77815    1.09990  -0.707 0.481926
## factor(director_name)Bennett Miller   -0.87162    0.97322  -0.896 0.373924
## factor(director_name)Billy Bob Thornton -1.12104  1.08877  -1.030 0.307177
## factor(director_name)Boaz Yakin       -0.98183    0.99607  -0.986 0.328111
## factor(director_name)David E. Talbert  0.10757    1.17399   0.092 0.927291
## factor(director_name)David Yates       0.13426    1.03231   0.130 0.896940
## factor(director_name)Dennie Gordon     0.30288    1.01474   0.298 0.766335
## factor(director_name)Forest Whitaker  -0.03747    1.19306  -0.031 0.975049
## factor(director_name)Gore Verbinski   -0.51406    0.97684  -0.526 0.600592
## factor(director_name)Hugh Wilson      -0.09791    0.99004  -0.099 0.921540
## factor(director_name)Jay Chandrasekhar -0.26540   0.99356  -0.267 0.790268
## factor(director_name)Jeb Stuart       -4.76517    1.10269  -4.321 5.71e-05
## factor(director_name)John Crowley     -1.43626    1.09670  -1.310 0.195154
## factor(director_name)John Lee Hancock -0.46926    0.88525  -0.530 0.597942
## factor(director_name)Jonathan Frakes   0.04088    0.99244   0.041 0.967272
## factor(director_name)Kevin Reynolds   -0.94842    0.91968  -1.031 0.306430
## factor(director_name)Matt Reeves      -0.38707    0.92184  -0.420 0.676020
## factor(director_name)Michael Cimino   -2.42230    1.07044  -2.263 0.027154
## factor(director_name)Michael Landon Jr. -3.24487  1.09791  -2.956 0.004409
## factor(director_name)Michael Tollin   -0.04800    1.10788  -0.043 0.965577
## factor(director_name)Nicholas Hytner  -0.60891    1.01320  -0.601 0.550049
## factor(director_name)Richard Eyre     -2.03208    1.17314  -1.732 0.088214
```

```
## factor(director_name)Rod Lurie          -2.26115    1.02375  -2.209 0.030901
## factor(director_name)Rodrigo Cortés     -0.74596    1.10827  -0.673 0.503395
## factor(director_name)Sean McNamara      -0.64029    1.07758  -0.594 0.554547
## factor(director_name)Stephen Herek      -0.29683    0.96431  -0.308 0.759255
## factor(director_name)Tarsem Singh        0.90339    1.02663   0.880 0.382283
## factor(director_name)Taylor Hackford    -1.54847    0.93432  -1.657 0.102505
## factor(director_name)William Wyler      -1.41649    1.10920  -1.277 0.206352
## movie_averageRating                      1.19373    0.18746   6.368 2.64e-08
## Drama                                   -0.78721    0.36596  -2.151 0.035374
## Thriller                                 1.02090    0.58569   1.743 0.086275
## Romance                                  0.57461    0.49771   1.155 0.252723
##
## (Intercept)
## log(budget)                             ***
## factor(director_name)Barbra Streisand
## factor(director_name)Bennett Miller
## factor(director_name)Billy Bob Thornton
## factor(director_name)Boaz Yakin
## factor(director_name)David E. Talbert
## factor(director_name)David Yates
## factor(director_name)Dennie Gordon
## factor(director_name)Forest Whitaker
## factor(director_name)Gore Verbinski
## factor(director_name)Hugh Wilson
## factor(director_name)Jay Chandrasekhar
## factor(director_name)Jeb Stuart         ***
## factor(director_name)John Crowley
## factor(director_name)John Lee Hancock
## factor(director_name)Jonathan Frakes
## factor(director_name)Kevin Reynolds
## factor(director_name)Matt Reeves
## factor(director_name)Michael Cimino     *
## factor(director_name)Michael Landon Jr. **
## factor(director_name)Michael Tollin
## factor(director_name)Nicholas Hytner
## factor(director_name)Richard Eyre       .
## factor(director_name)Rod Lurie          *
## factor(director_name)Rodrigo Cortés
## factor(director_name)Sean McNamara
## factor(director_name)Stephen Herek
## factor(director_name)Tarsem Singh
## factor(director_name)Taylor Hackford
## factor(director_name)William Wyler
## movie_averageRating                     ***
## Drama                                   *
```

```
## Thriller                                  .
## Romance
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.057 on 62 degrees of freedom
## Multiple R-squared:  0.7709, Adjusted R-squared:  0.6452
## F-statistic: 6.134 on 34 and 62 DF,  p-value: 3.988e-10
```

```
summary(step_both)
```

```
##
## Call:
## lm(formula = log(gross) ~ log(budget) + runtime_minutes + birth_year +
##     movie_averageRating + factor(director_name) + Drama + Thriller +
##     Romance, data = cinema)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7669 -0.4795  0.0626  0.5046  2.1563
##
## Coefficients: (18 not defined because of singularities)
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            -1.71043    3.17702  -0.538 0.592277
## log(budget)                             0.53071    0.19177   2.767 0.007470
## runtime_minutes                         0.01207    0.01121   1.077 0.285709
## birth_year1939                         -0.95309    1.11473  -0.855 0.395900
## birth_year1942                          1.42291    1.27153   1.119 0.267506
## birth_year1943                          0.58722    1.56500   0.375 0.708798
## birth_year1944                          0.68170    1.23933   0.550 0.584287
## birth_year1952                          1.45101    1.35176   1.073 0.287310
## birth_year1953                          2.92079    1.58587   1.842 0.070373
## birth_year1955                          2.43695    1.51466   1.609 0.112802
## birth_year1956                          1.73362    1.31549   1.318 0.192481
## birth_year1958                          2.24535    1.50007   1.497 0.139595
## birth_year1961                          3.50153    1.56917   2.231 0.029337
## birth_year1962                          1.79913    1.48935   1.208 0.231709
## birth_year1963                          2.55774    1.42895   1.790 0.078426
## birth_year1964                         -0.91225    1.45741  -0.626 0.533690
## birth_year1966                          2.03741    1.33962   1.521 0.133455
## birth_year1968                          2.18062    1.44542   1.509 0.136553
## birth_year1969                          0.76738    1.31497   0.584 0.561661
## birth_year1973                          1.62824    1.47218   1.106 0.273069
## birth_year1986                          2.11002    1.28133   1.647 0.104754
```

```
## movie_averageRating                        1.13789    0.19427   5.857 2.03e-07
## factor(director_name)Barbra Streisand          NA         NA      NA       NA
## factor(director_name)Bennett Miller      -0.57921    0.78234  -0.740 0.461927
## factor(director_name)Billy Bob Thornton  -1.30707    1.09403  -1.195 0.236814
## factor(director_name)Boaz Yakin          -0.47082    0.79514  -0.592 0.555958
## factor(director_name)David E. Talbert     0.42575    0.96869   0.440 0.661846
## factor(director_name)David Yates               NA         NA      NA       NA
## factor(director_name)Dennie Gordon             NA         NA      NA       NA
## factor(director_name)Forest Whitaker     -1.13549    1.09096  -1.041 0.302070
## factor(director_name)Gore Verbinski       2.87563    1.09561   2.625 0.010945
## factor(director_name)Hugh Wilson          1.83751    1.00743   1.824 0.073058
## factor(director_name)Jay Chandrasekhar         NA         NA      NA       NA
## factor(director_name)Jeb Stuart          -4.24654    1.05460  -4.027 0.000159
## factor(director_name)John Crowley              NA         NA      NA       NA
## factor(director_name)John Lee Hancock     0.16039    0.82690   0.194 0.846845
## factor(director_name)Jonathan Frakes      1.21113    0.76625   1.581 0.119142
## factor(director_name)Kevin Reynolds            NA         NA      NA       NA
## factor(director_name)Matt Reeves               NA         NA      NA       NA
## factor(director_name)Michael Cimino            NA         NA      NA       NA
## factor(director_name)Michael Landon Jr.        NA         NA      NA       NA
## factor(director_name)Michael Tollin            NA         NA      NA       NA
## factor(director_name)Nicholas Hytner           NA         NA      NA       NA
## factor(director_name)Richard Eyre              NA         NA      NA       NA
## factor(director_name)Rod Lurie           -1.75264    1.05641  -1.659 0.102239
## factor(director_name)Rodrigo Cortés            NA         NA      NA       NA
## factor(director_name)Sean McNamara             NA         NA      NA       NA
## factor(director_name)Stephen Herek             NA         NA      NA       NA
## factor(director_name)Tarsem Singh              NA         NA      NA       NA
## factor(director_name)Taylor Hackford           NA         NA      NA       NA
## factor(director_name)William Wyler             NA         NA      NA       NA
## Drama                                    -0.84106    0.36889  -2.280 0.026119
## Thriller                                  1.11422    0.59131   1.884 0.064290
## Romance                                   0.65028    0.50201   1.295 0.200075
##
## (Intercept)
## log(budget)                              **
## runtime_minutes
## birth_year1939
## birth_year1942
## birth_year1943
## birth_year1944
## birth_year1952
## birth_year1953                           .
## birth_year1955
## birth_year1956
```

```
## birth_year1958
## birth_year1961                          *
## birth_year1962
## birth_year1963                          .
## birth_year1964
## birth_year1966
## birth_year1968
## birth_year1969
## birth_year1973
## birth_year1986
## movie_averageRating                     ***
## factor(director_name)Barbra Streisand
## factor(director_name)Bennett Miller
## factor(director_name)Billy Bob Thornton
## factor(director_name)Boaz Yakin
## factor(director_name)David E. Talbert
## factor(director_name)David Yates
## factor(director_name)Dennie Gordon
## factor(director_name)Forest Whitaker
## factor(director_name)Gore Verbinski     *
## factor(director_name)Hugh Wilson        .
## factor(director_name)Jay Chandrasekhar
## factor(director_name)Jeb Stuart         ***
## factor(director_name)John Crowley
## factor(director_name)John Lee Hancock
## factor(director_name)Jonathan Frakes
## factor(director_name)Kevin Reynolds
## factor(director_name)Matt Reeves
## factor(director_name)Michael Cimino
## factor(director_name)Michael Landon Jr.
## factor(director_name)Michael Tollin
## factor(director_name)Nicholas Hytner
## factor(director_name)Richard Eyre
## factor(director_name)Rod Lurie
## factor(director_name)Rodrigo Cortés
## factor(director_name)Sean McNamara
## factor(director_name)Stephen Herek
## factor(director_name)Tarsem Singh
## factor(director_name)Taylor Hackford
## factor(director_name)William Wyler
## Drama                                   *
## Thriller                                .
## Romance
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.056 on 61 degrees of freedom
## Multiple R-squared:  0.7751, Adjusted R-squared:  0.6461
## F-statistic: 6.008 on 35 and 61 DF,  p-value: 6.186e-10
```

```
AIC(step_forward, step_backward, step_both, m_full)
```

```
##               df      AIC
## step_forward  37 314.8316
## step_backward 36 314.6589
## step_both     37 314.8316
## m_full        45 328.8171
```

```
summary(step_backward)
```

```
##
## Call:
## lm(formula = log(gross) ~ log(budget) + factor(director_name) +
##     movie_averageRating + Drama + Thriller + Romance, data = cinema)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8866 -0.4418  0.0912  0.4454  2.2530
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -0.24155    3.07590  -0.079 0.937659
## log(budget)                           0.64553    0.15961   4.044 0.000148
## factor(director_name)Barbra Streisand -0.77815    1.09990  -0.707 0.481926
## factor(director_name)Bennett Miller   -0.87162    0.97322  -0.896 0.373924
## factor(director_name)Billy Bob Thornton -1.12104  1.08877  -1.030 0.307177
## factor(director_name)Boaz Yakin       -0.98183    0.99607  -0.986 0.328111
## factor(director_name)David E. Talbert  0.10757    1.17399   0.092 0.927291
## factor(director_name)David Yates       0.13426    1.03231   0.130 0.896940
## factor(director_name)Dennie Gordon     0.30288    1.01474   0.298 0.766335
## factor(director_name)Forest Whitaker  -0.03747    1.19306  -0.031 0.975049
## factor(director_name)Gore Verbinski   -0.51406    0.97684  -0.526 0.600592
## factor(director_name)Hugh Wilson      -0.09791    0.99004  -0.099 0.921540
## factor(director_name)Jay Chandrasekhar -0.26540   0.99356  -0.267 0.790268
## factor(director_name)Jeb Stuart       -4.76517    1.10269  -4.321 5.71e-05
## factor(director_name)John Crowley     -1.43626    1.09670  -1.310 0.195154
## factor(director_name)John Lee Hancock -0.46926    0.88525  -0.530 0.597942
## factor(director_name)Jonathan Frakes   0.04088    0.99244   0.041 0.967272
```

```
## factor(director_name)Kevin Reynolds     -0.94842    0.91968  -1.031 0.306430
## factor(director_name)Matt Reeves        -0.38707    0.92184  -0.420 0.676020
## factor(director_name)Michael Cimino     -2.42230    1.07044  -2.263 0.027154
## factor(director_name)Michael Landon Jr. -3.24487    1.09791  -2.956 0.004409
## factor(director_name)Michael Tollin     -0.04800    1.10788  -0.043 0.965577
## factor(director_name)Nicholas Hytner    -0.60891    1.01320  -0.601 0.550049
## factor(director_name)Richard Eyre       -2.03208    1.17314  -1.732 0.088214
## factor(director_name)Rod Lurie          -2.26115    1.02375  -2.209 0.030901
## factor(director_name)Rodrigo Cortés     -0.74596    1.10827  -0.673 0.503395
## factor(director_name)Sean McNamara      -0.64029    1.07758  -0.594 0.554547
## factor(director_name)Stephen Herek      -0.29683    0.96431  -0.308 0.759255
## factor(director_name)Tarsem Singh        0.90339    1.02663   0.880 0.382283
## factor(director_name)Taylor Hackford    -1.54847    0.93432  -1.657 0.102505
## factor(director_name)William Wyler      -1.41649    1.10920  -1.277 0.206352
## movie_averageRating                      1.19373    0.18746   6.368 2.64e-08
## Drama                                   -0.78721    0.36596  -2.151 0.035374
## Thriller                                 1.02090    0.58569   1.743 0.086275
## Romance                                  0.57461    0.49771   1.155 0.252723
##
## (Intercept)
## log(budget)                             ***
## factor(director_name)Barbra Streisand
## factor(director_name)Bennett Miller
## factor(director_name)Billy Bob Thornton
## factor(director_name)Boaz Yakin
## factor(director_name)David E. Talbert
## factor(director_name)David Yates
## factor(director_name)Dennie Gordon
## factor(director_name)Forest Whitaker
## factor(director_name)Gore Verbinski
## factor(director_name)Hugh Wilson
## factor(director_name)Jay Chandrasekhar
## factor(director_name)Jeb Stuart         ***
## factor(director_name)John Crowley
## factor(director_name)John Lee Hancock
## factor(director_name)Jonathan Frakes
## factor(director_name)Kevin Reynolds
## factor(director_name)Matt Reeves
## factor(director_name)Michael Cimino     *
## factor(director_name)Michael Landon Jr. **
## factor(director_name)Michael Tollin
## factor(director_name)Nicholas Hytner
## factor(director_name)Richard Eyre       .
## factor(director_name)Rod Lurie          *
## factor(director_name)Rodrigo Cortés
```

```
## factor(director_name)Sean McNamara
## factor(director_name)Stephen Herek
## factor(director_name)Tarsem Singh
## factor(director_name)Taylor Hackford
## factor(director_name)William Wyler
## movie_averageRating                      ***
## Drama                                    *
## Thriller                                 .
## Romance
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.057 on 62 degrees of freedom
## Multiple R-squared:  0.7709, Adjusted R-squared:  0.6452
## F-statistic: 6.134 on 34 and 62 DF,  p-value: 3.988e-10
```

```
## The step backward model is the best model

m_with_interactions <- lm(
  log(gross) ~ (log(budget) + movie_averageRating) *
               (Drama + Thriller + Romance) +
               log(budget):movie_averageRating +
               factor(director_name),
  data = cinema
)


main_effects <- step_backward
step_interactions <- step(main_effects,
                          scope = list(lower = main_effects,
                                       upper = m_with_interactions),
                          direction = "both",
                          trace = TRUE)
```

```
## Start:  AIC=37.38
## log(gross) ~ log(budget) + factor(director_name) + movie_averageRating +
##     Drama + Thriller + Romance
##
##                                    Df Sum of Sq    RSS    AIC
## + movie_averageRating:Drama         1   2.12849 67.173 36.359
## <none>                                          69.302 37.385
## + log(budget):movie_averageRating   1   1.07043 68.231 37.875
## + log(budget):Romance               1   0.78364 68.518 38.282
## + movie_averageRating:Thriller      1   0.48570 68.816 38.703
```

```
## + log(budget):Thriller              1   0.04415 69.258 39.323
## + movie_averageRating:Romance       1   0.02788 69.274 39.346
## + log(budget):Drama                 1   0.02292 69.279 39.353
##
## Step:  AIC=36.36
## log(gross) ~ log(budget) + factor(director_name) + movie_averageRating +
##     Drama + Thriller + Romance + movie_averageRating:Drama
##
##                                    Df Sum of Sq    RSS    AIC
## <none>                                          67.173 36.359
## + movie_averageRating:Thriller      1   0.86760 66.306 37.098
## + log(budget):movie_averageRating   1   0.68480 66.489 37.365
## - movie_averageRating:Drama         1   2.12849 69.302 37.385
## + log(budget):Romance               1   0.63717 66.536 37.434
## + movie_averageRating:Romance       1   0.28420 66.889 37.948
## + log(budget):Drama                 1   0.05134 67.122 38.285
## + log(budget):Thriller              1   0.00568 67.168 38.351
```

```
# Compare
AIC(main_effects, step_interactions)
```

```
##                   df      AIC
## main_effects      36 314.6589
## step_interactions 37 313.6330
```

```
summary(step_interactions)
```

```
##
## Call:
## lm(formula = log(gross) ~ log(budget) + factor(director_name) +
##     movie_averageRating + Drama + Thriller + Romance + movie_averageRating:Drama,
##     data = cinema)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8131 -0.4326  0.0755  0.4295  2.2617
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             1.22792    3.23080   0.380 0.705216
## log(budget)                             0.65406    0.15854   4.125 0.000114
## factor(director_name)Barbra Streisand  -0.60811    1.09855  -0.554 0.581903
## factor(director_name)Bennett Miller    -0.89871    0.96617  -0.930 0.355945
```

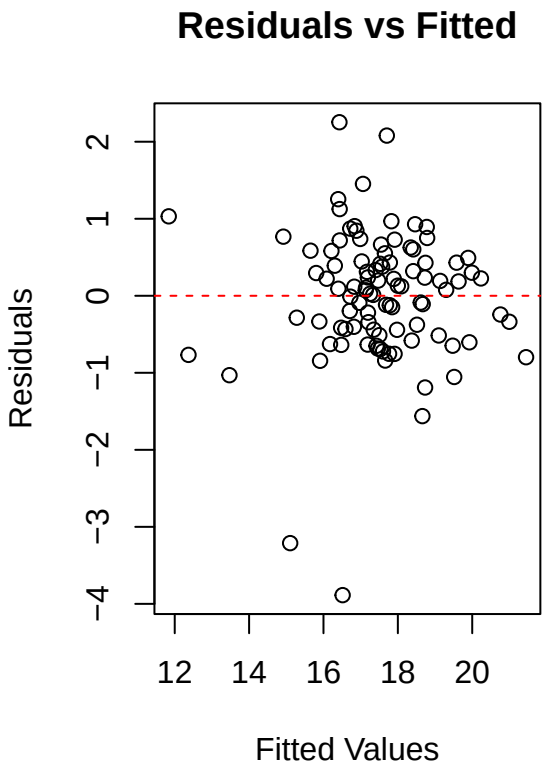
```
## factor(director_name)Billy Bob Thornton -1.02884    1.08270  -0.950 0.345733
## factor(director_name)Boaz Yakin         -0.92414    0.98953  -0.934 0.354028
## factor(director_name)David E. Talbert    0.21869    1.16800   0.187 0.852099
## factor(director_name)David Yates         0.48503    1.05524   0.460 0.647407
## factor(director_name)Dennie Gordon       0.40815    1.01004   0.404 0.687555
## factor(director_name)Forest Whitaker     0.37117    1.22012   0.304 0.762005
## factor(director_name)Gore Verbinski     -0.24656    0.98849  -0.249 0.803867
## factor(director_name)Hugh Wilson        -0.02371    0.98412  -0.024 0.980858
## factor(director_name)Jay Chandrasekhar  -0.22779    0.98654  -0.231 0.818166
## factor(director_name)Jeb Stuart         -4.61371    1.09989  -4.195    9e-05
## factor(director_name)John Crowley       -1.35736    1.09002  -1.245 0.217797
## factor(director_name)John Lee Hancock   -0.42134    0.87934  -0.479 0.633545
## factor(director_name)Jonathan Frakes    -0.05275    0.98736  -0.053 0.957569
## factor(director_name)Kevin Reynolds     -0.86345    0.91488  -0.944 0.349002
## factor(director_name)Matt Reeves        -0.38722    0.91499  -0.423 0.673640
## factor(director_name)Michael Cimino     -2.47198    1.06307  -2.325 0.023398
## factor(director_name)Michael Landon Jr. -3.02780    1.10087  -2.750 0.007824
## factor(director_name)Michael Tollin      0.23554    1.11839   0.211 0.833896
## factor(director_name)Nicholas Hytner    -0.45820    1.01149  -0.453 0.652156
## factor(director_name)Richard Eyre       -1.96592    1.16538  -1.687 0.096723
## factor(director_name)Rod Lurie          -2.18890    1.01746  -2.151 0.035419
## factor(director_name)Rodrigo Cortés     -0.65167    1.10212  -0.591 0.556515
## factor(director_name)Sean McNamara      -0.59203    1.07013  -0.553 0.582128
## factor(director_name)Stephen Herek      -0.11923    0.96562  -0.123 0.902138
## factor(director_name)Tarsem Singh        1.15896    1.03544   1.119 0.267404
## factor(director_name)Taylor Hackford    -1.44364    0.93043  -1.552 0.125934
## factor(director_name)William Wyler      -1.56316    1.10599  -1.413 0.162634
## movie_averageRating                      0.90560    0.27852   3.252 0.001871
## Drama                                   -3.87633    2.25143  -1.722 0.090189
## Thriller                                 1.08743    0.58329   1.864 0.067094
## Romance                                  0.50623    0.49645   1.020 0.311891
## movie_averageRating:Drama                0.49405    0.35536   1.390 0.169496
##
## (Intercept)
## log(budget)                             ***
## factor(director_name)Barbra Streisand
## factor(director_name)Bennett Miller
## factor(director_name)Billy Bob Thornton
## factor(director_name)Boaz Yakin
## factor(director_name)David E. Talbert
## factor(director_name)David Yates
## factor(director_name)Dennie Gordon
## factor(director_name)Forest Whitaker
## factor(director_name)Gore Verbinski
## factor(director_name)Hugh Wilson
```

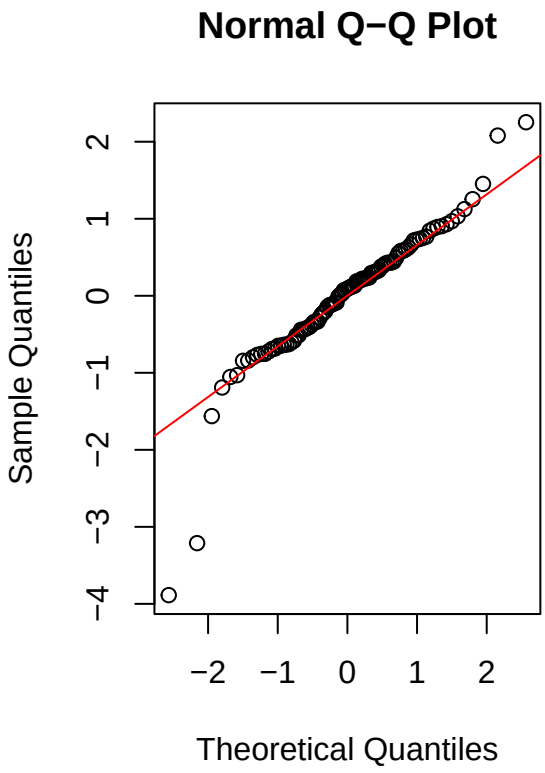
```
## factor(director_name)Jay Chandrasekhar
## factor(director_name)Jeb Stuart         ***
## factor(director_name)John Crowley
## factor(director_name)John Lee Hancock
## factor(director_name)Jonathan Frakes
## factor(director_name)Kevin Reynolds
## factor(director_name)Matt Reeves
## factor(director_name)Michael Cimino     *
## factor(director_name)Michael Landon Jr. **
## factor(director_name)Michael Tollin
## factor(director_name)Nicholas Hytner
## factor(director_name)Richard Eyre       .
## factor(director_name)Rod Lurie          *
## factor(director_name)Rodrigo Cortés
## factor(director_name)Sean McNamara
## factor(director_name)Stephen Herek
## factor(director_name)Tarsem Singh
## factor(director_name)Taylor Hackford
## factor(director_name)William Wyler
## movie_averageRating                     **
## Drama                                   .
## Thriller                                .
## Romance
## movie_averageRating:Drama
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.049 on 61 degrees of freedom
## Multiple R-squared:  0.7779, Adjusted R-squared:  0.6505
## F-statistic: 6.104 on 35 and 61 DF,  p-value: 4.488e-10
```

```
# Test if the interaction significantly improves fit
anova(main_effects, step_interactions)
```

```
## Analysis of Variance Table
##
## Model 1: log(gross) ~ log(budget) + factor(director_name) + movie_averageRating +
##     Drama + Thriller + Romance
## Model 2: log(gross) ~ log(budget) + factor(director_name) + movie_averageRating +
##     Drama + Thriller + Romance + movie_averageRating:Drama
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     62 69.302
## 2     61 67.173  1    2.1285 1.9329 0.1695
```

```r
##Given that the interactions model is not statistically significant, we can just use

# 1. Residuals vs Fitted (Linearity check)
plot(fitted(main_effects), residuals(main_effects),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs Fitted")
abline(h = 0, col = "red", lty = 2)

# 2. Normal Q-Q plot (Normality check)
qqnorm(residuals(main_effects))
qqline(residuals(main_effects), col = "red")
```






```r
# 3. Histogram of residuals
hist(residuals(main_effects), breaks = 20,
     main = "Histogram of Residuals",
     xlab = "Residuals")

#Final model:  log(gross) ~ log(budget) + factor(director_name) + movie_averageRating

#Predict value:
test_director <- levels(cinema$director_name)[1]
```

```r
test_movie <- data.frame(
  budget = 10000000,
  movie_averageRating = 7.0,
  Drama = 0,
  Thriller = 0,
  Romance = 0,
  director_name = factor(test_director, levels = levels(cinema$director_name))
)
log_pred <- predict(main_effects, newdata = test_movie)
log_pred
```

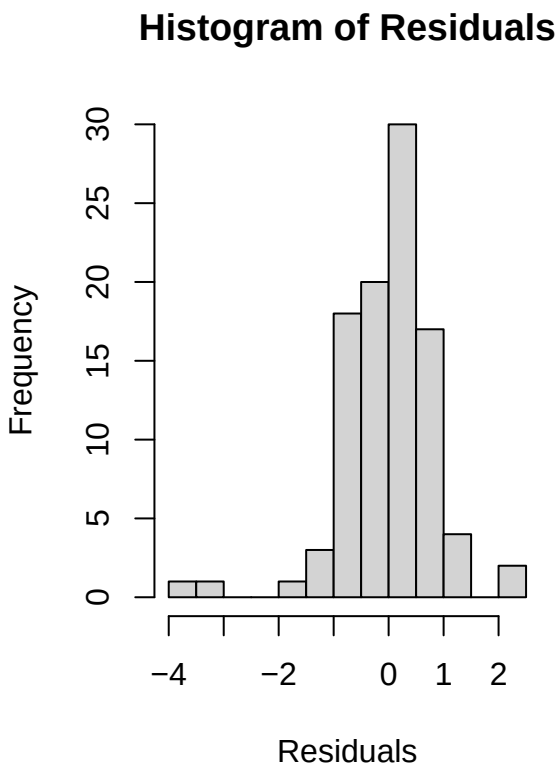
```
##        1
## 18.51928
```

```r
gross_pred <- exp(log_pred)
gross_pred
```

```
##         1
## 110362748
```

```r
pred_int <- predict(
  main_effects,
  newdata = test_movie,
  interval = "prediction"
)

exp(pred_int)
```

```
##         fit     lwr        upr
## 1 110362748 7536100 1616212102
```


Histogram of Residuals
Frequency
0
5
10
15
20
25
30
−4
−2
0
1
2
Residuals