Group 4- Project 2

2326127C

09/03/2022

Introduction

Our group has been assigned to work with a database of films from IMDB which contains information about a number of films and rating out of 10 for each films. The variables in the database are:

- Film.id- a unique identifying number for the film
- Year of release
- Length of Film (in minutes)
- Budget of the Film (in $1000000s)
- Number of positive votes received by viewers
- Genre of the Film
- IMDB Rating of the Film

Our task is the find which properties of a film influence whether a film receives an IMBD rating greater than 7 or not. We will be performing logistic regression with different combinations of the explanatory variables to see which variables are the most significant predictors.

Exploratory Data Analysis

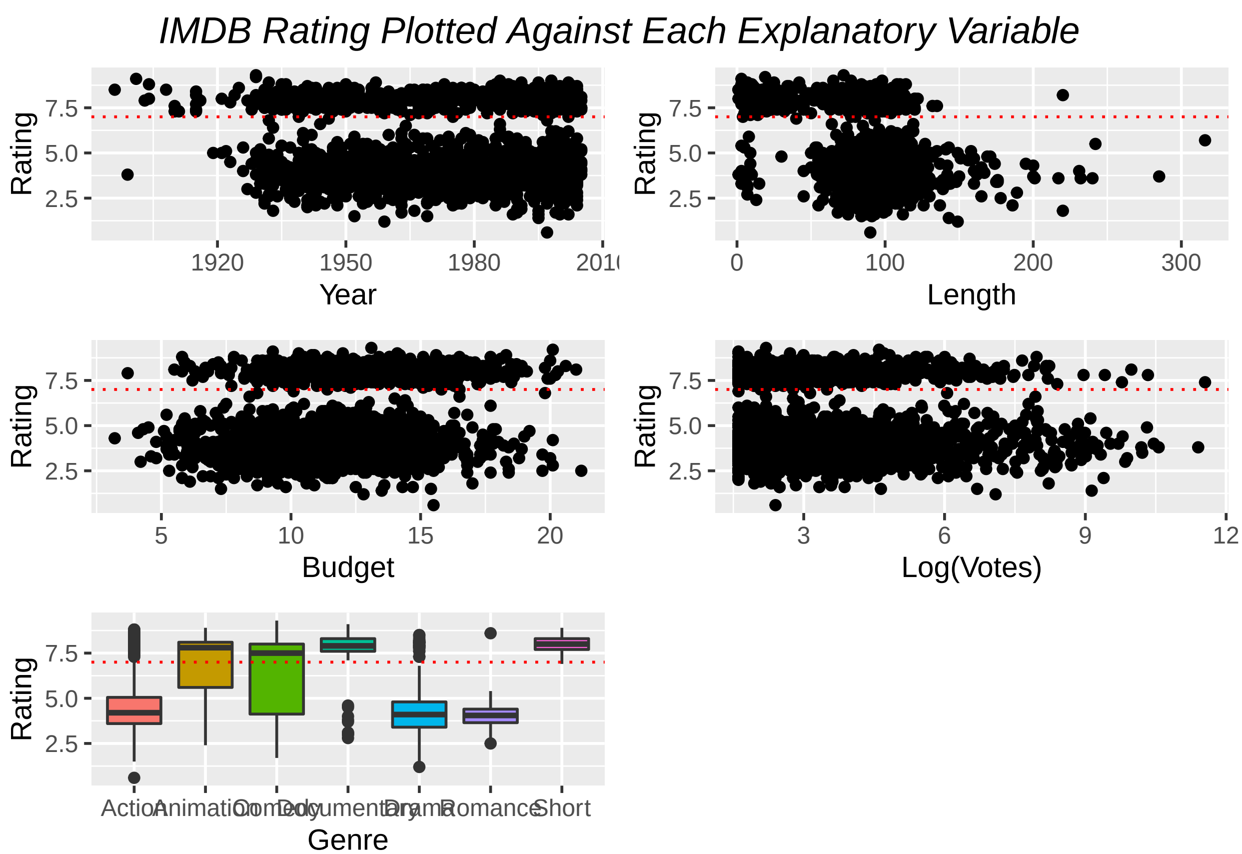
First we will plot the relationships between IMBD rating and each of the explanatory variables. Each plot has a red dotted line at rating equals 7 as we are interested in films that receive a rating over 7.

Table 1: Summary statistics on number of films which are rating larger than 7).

Variable	n	Mean	SD	Min	Q1	Median	Q3	Max	IQR
year	641	1974.91	26.41	1896.0	1951.0	1984.0	1999.0	2005	15.0
length	641	56.12	39.76	1.0	12.0	71.5	91.0	220	19.5
budget	641	13.08	2.84	3.7	11.1	13.0	15.1	21	2.1
votes	641	438.65	4459.97	5.0	10.0	23.0	66.0	103854	43.0

Table 2: Summary statistics on number of films which are rating smaller than 7).

Variable	n	Mean	SD	Min	Q1	Median	Q3	Max	IQR
year	1296	1976.85	21.81	1899.0	1960.0	1981.0	1997.00	2005.0	16.00
length	1296	96.02	25.43	1.0	85.0	94.0	105.00	316.0	11.00
budget	1296	11.51	2.82	3.2	9.5	11.5	13.50	21.2	2.00
votes	1296	665.56	3581.20	5.0	14.0	37.0	158.25	89722.0	121.25

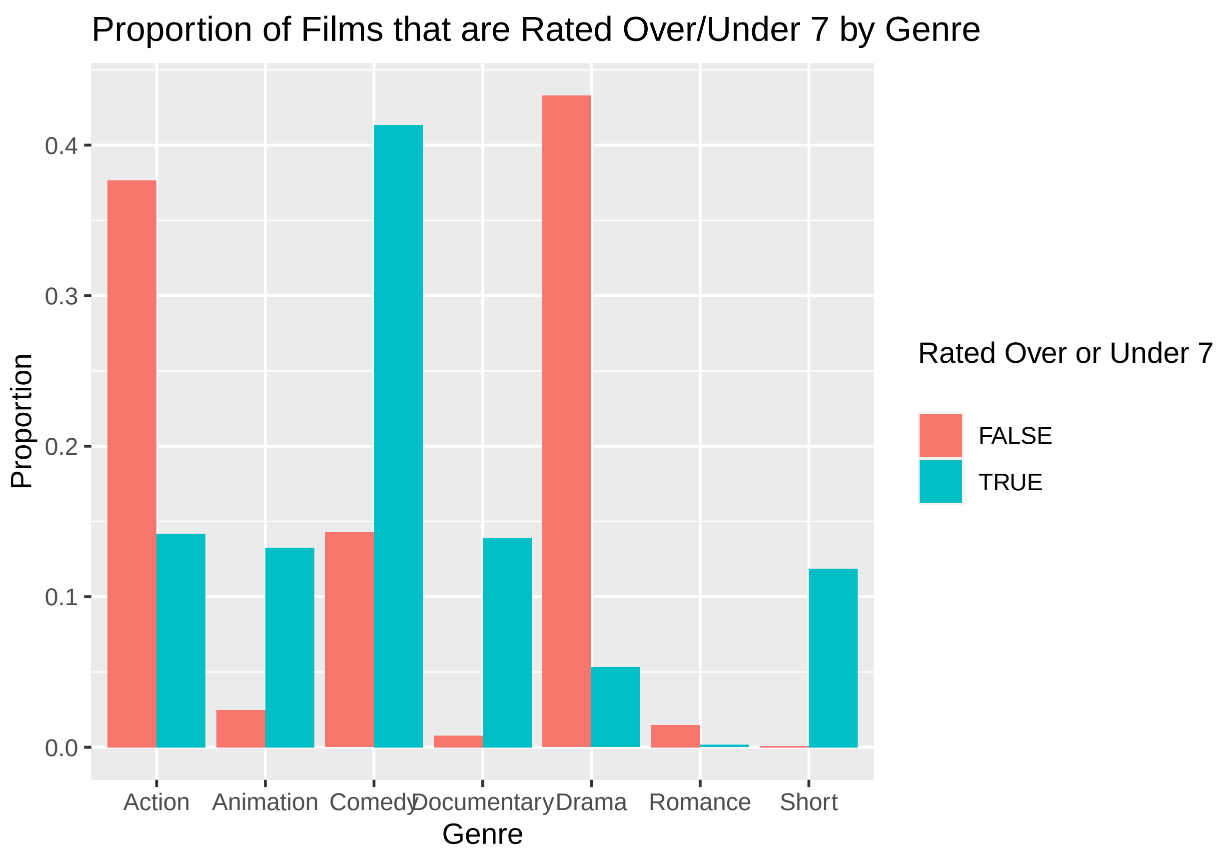


Summary statistics were presented in the following table for each factor separately.

This table shows the year of production, length of films, film budget ($1 million), and the number of positive audience votes for all films rated 7.0 or higher. We are unable to show the genres of movies in this table and the following table, the genre analysis will be shown in the histogram below.

By comparing the two tables, we can find that the number of movies with a rating greater than 7.0 is significantly smaller than the number of movies with a rating less than 7.0. Movies with a rating greater than 7.0 generally have shorter movie durations and higher budgets. But in terms of voting. Movies rated less than 7.0 received more votes.

```
##         genre       FALSE        TRUE
##        Action 84.3% (488) 15.7%  (91)
##     Animation 27.4%  (32) 72.6%  (85)
##        Comedy 41.1% (185) 58.9% (265)
##   Documentary 10.1%  (10) 89.9%  (89)
##         Drama 94.3% (561)  5.7%  (34)
##       Romance 95.0%  (19)  5.0%   (1)
##         Short  1.3%   (1) 98.7%  (76)
```



This histogram shows the genre of all movies with a rating greater than 7.0. Through this figure, we can find that comedy movies occupy a very large proportion of movies with a score greater than 7.0, while romance movies have almost no high rating.

Formal Data Analysis

We have created a new variable named over7 which is a binary variable which indicates whether the rating a film received is over 7 or not. If a film has a rating over 7 it will have the value 1 in this variable. The explanatory variables we will use to model over 7 are- genre, votes, length, budget and year. There are 31 one unique ways to choose different combinations of these five explanatory variables so to start with we will fit all of these models and generate a table of the objective criteria of each model. This table can be found below:

Table 3: Objective Criteria for Each Possible Model

Formula	AIC	BIC	Deviance
over7 ~ year+length+budget+genre	956.88	1012.02	936.88
over7 ~ year+length+budget+votes+genre	957.43	1018.09	935.43
over7 ~ length+budget+genre	962.64	1012.27	944.64
over7 ~ length+budget+votes+genre	962.71	1017.85	942.71
over7 ~ year+length+votes+genre	1235.54	1290.68	1215.54
over7 ~ year+length+genre	1235.83	1285.46	1217.83
over7 ~ length+votes+genre	1238.81	1288.43	1220.81
over7 ~ length+genre	1239.55	1283.67	1223.55
over7 ~ budget+genre	1281.22	1325.34	1265.22
over7 ~ year+budget+genre	1281.40	1331.03	1263.40
over7 ~ budget+votes+genre	1281.52	1331.15	1263.52
over7 ~ year+budget+votes+genre	1282.05	1337.19	1262.05
over7 ~ year+genre	1508.67	1552.79	1492.67
over7 ~ genre	1508.75	1547.35	1494.75
over7 ~ votes+genre	1509.49	1553.60	1493.49
over7 ~ year+votes+genre	1509.74	1559.37	1491.74
over7 ~ year+length+budget	1589.53	1611.59	1581.53
over7 ~ year+length+budget+votes	1589.77	1617.34	1579.77
over7 ~ length+budget+votes	1600.21	1622.26	1592.21
over7 ~ length+budget	1600.56	1617.10	1594.56
over7 ~ year+length+votes	1754.20	1776.26	1746.20
over7 ~ year+length	1754.94	1771.48	1748.94
over7 ~ length+votes	1763.22	1779.76	1757.22
over7 ~ length	1764.54	1775.57	1760.54
over7 ~ year+budget+votes	2221.98	2244.04	2213.98
over7 ~ budget+votes	2222.36	2238.90	2216.36
over7 ~ year+budget	2223.02	2239.57	2217.02
over7 ~ budget	2224.03	2235.06	2220.03
over7 ~ year+votes	2332.24	2348.78	2326.24
over7 ~ year	2332.50	2343.53	2328.50
over7 ~ votes	2332.66	2343.69	2328.66

From this table we can see that there is a wide range of AIC, BIC and deviation values. When taking these values into consideration to help choose our model we can see that the models that include votes tend to perform worse than the others. Furthermore, we can see that based on the AIC and BIC values the model with year, length, budget and genre performs the best. If we were prepared to make a small compromise in performance it could be argued that the best model to choose would be the model which only use length, budget and genre as it is simpler and has close to the best AIC and BIC models. The best model which includes two variables uses length and genre. The best single explanatory variable model is genre.

We have investigated a handful of models in detail and we will share our discoveries below.

Model 1 The first model is investigating the relationship between the year a film was released and whether or not the film received a rating over 7. The equation for this model is:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{year} \tag{1}$$

where,

- p is the probability that the film is ranked over 7,
- year is the year the film was released,
- α is the intercept value
- β is the regression coefficient.

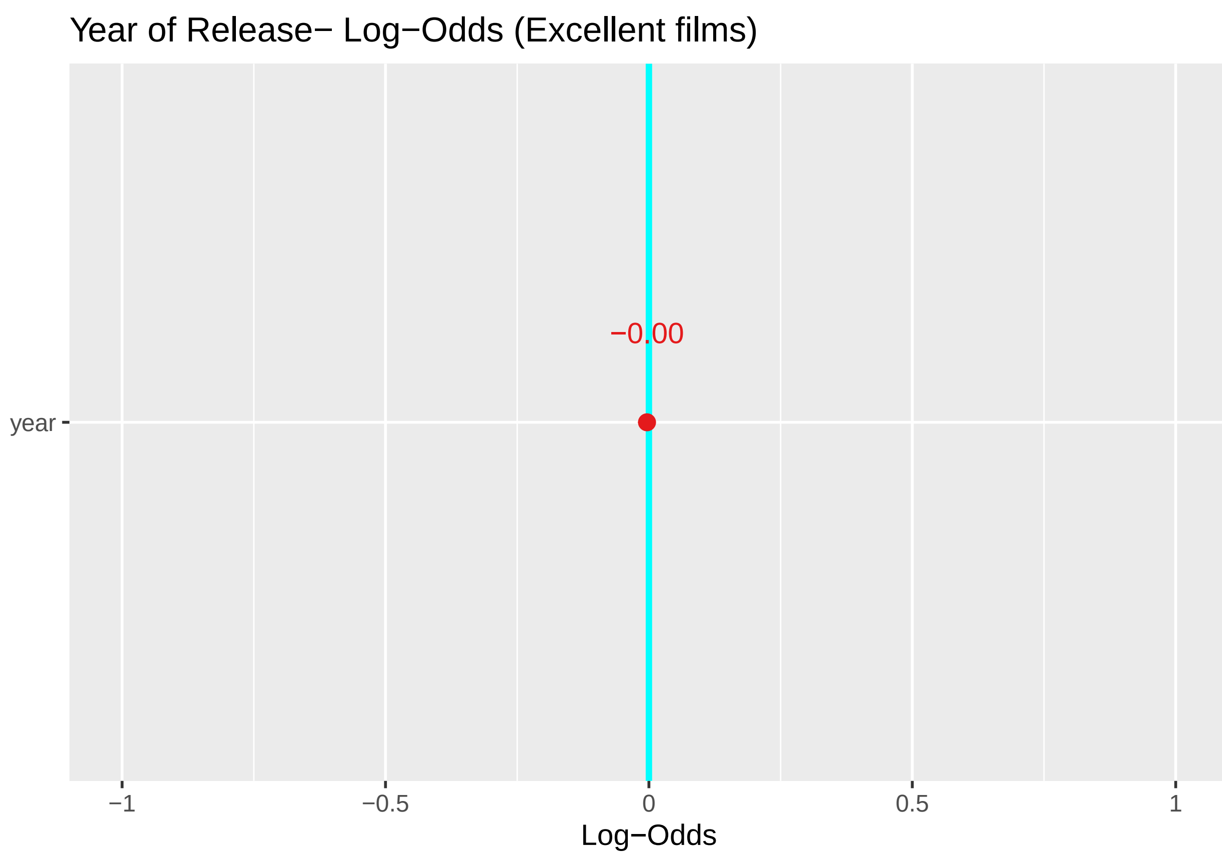
```
# create logistic regression model between over7 and the year of film release and show summary
model1 = glm(over7 ~ year, data = films,
             family = binomial(link = "logit"))

model1 %>%
  summary()
```

```
##
## Call:
## glm(formula = over7 ~ year, family = binomial(link = "logit"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0043  -0.9033  -0.8702   1.4448   1.5326
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.438996   4.122032   1.562   0.1183
## year        -0.003613   0.002087  -1.731   0.0834 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.5  on 1833  degrees of freedom
## Residual deviance: 2328.5  on 1832  degrees of freedom
## AIC: 2332.5
##
## Number of Fisher Scoring iterations: 4
```

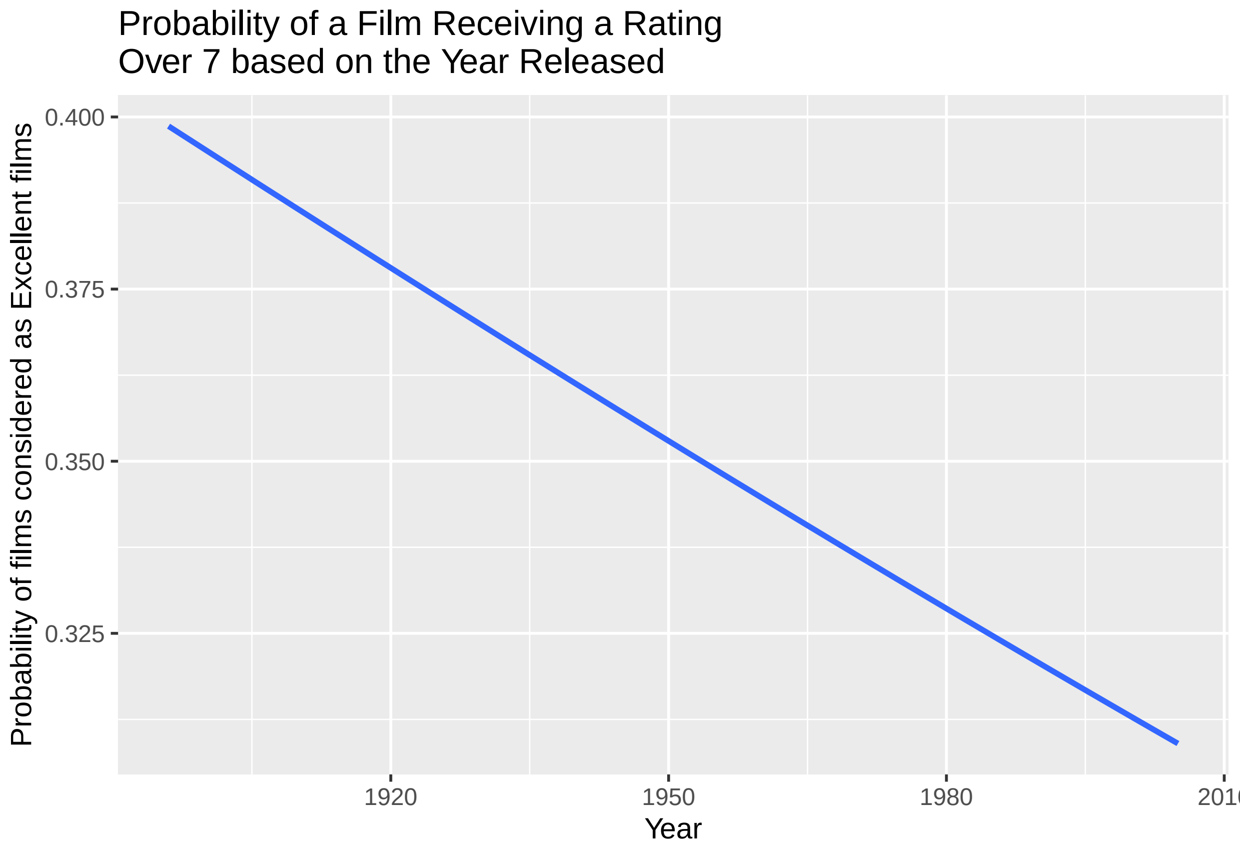
```r
# calculate confindence intervals
confint(model1) %>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	-1.6600637	14.5056305
year	-0.0076969	0.0004861

```r
# calculate confindence intervals
plot_model(model1, show.values = TRUE, transform = NULL,
           title = "Year of Release- Log-Odds (Excellent films)", show.p = FALSE, vline.color = "cyan")
```



This tells us $\alpha = 6.44$ and that $\beta = -0.0036$. We can see that the p-values of the coefficients are not significant even at the 5% level. Both 95% confidence intervals contain zero. We can conclude that this is a poor performing model.



Model 2 The next model is investigating the relationship between the length of a film and whether or not the film received a rating over 7. The equation for this model is:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{length} \tag{2}$$

where,

- p is the probability that the film is ranked over 7,
- length is the length of the film in minutes,
- α is the intercept value
- β is the regression coefficient.

```
# create logistic regression model between over7 and the length of film in minutes and show summary
model2 = glm(over7 ~ length, data = films,
             family = binomial(link = "logit"))
model2 %>%
  summary()
```

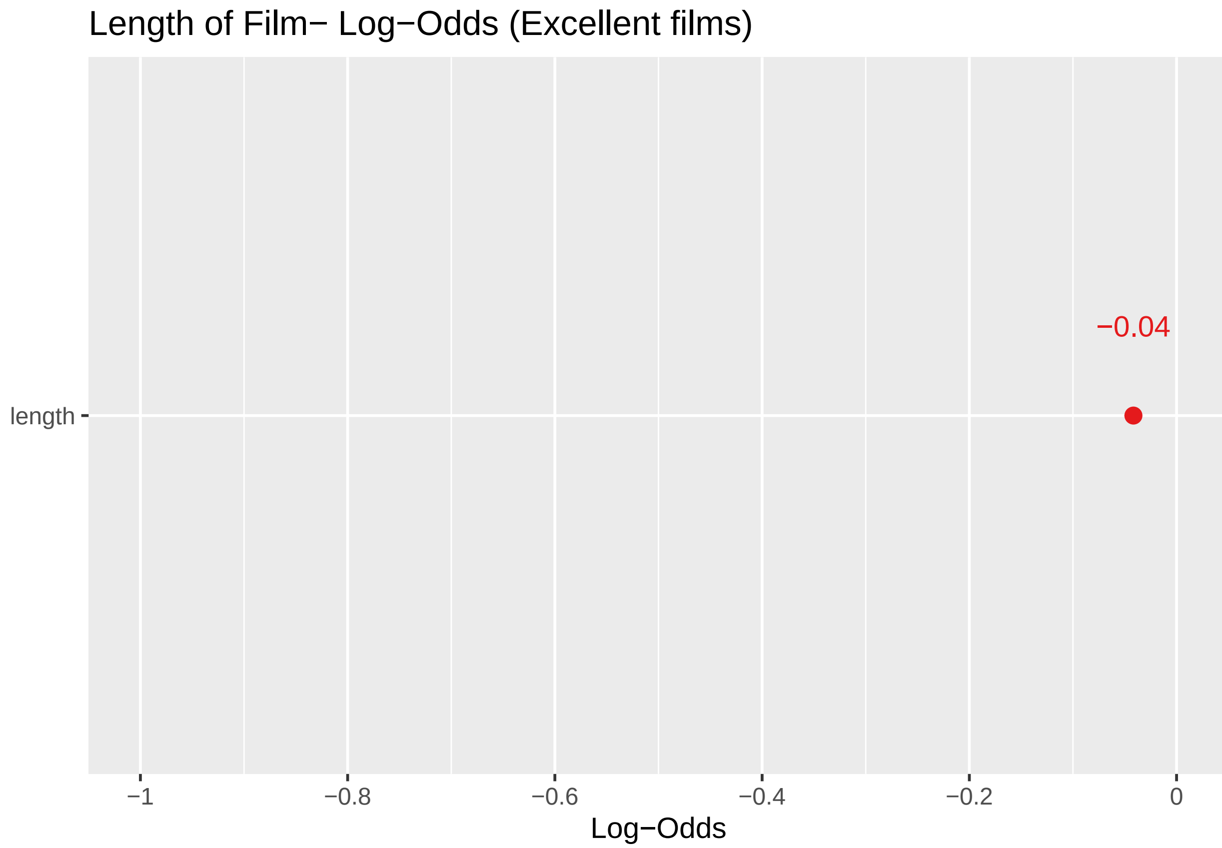
```
##
## Call:
## glm(formula = over7 ~ length, family = binomial(link = "logit"),
##     data = films)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2994  -0.7461  -0.5631   0.4780   3.6200
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.611624   0.192444   13.57   <2e-16 ***
## length      -0.041647   0.002252  -18.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.5  on 1833  degrees of freedom
## Residual deviance: 1760.5  on 1832  degrees of freedom
## AIC: 1764.5
##
## Number of Fisher Scoring iterations: 5
```

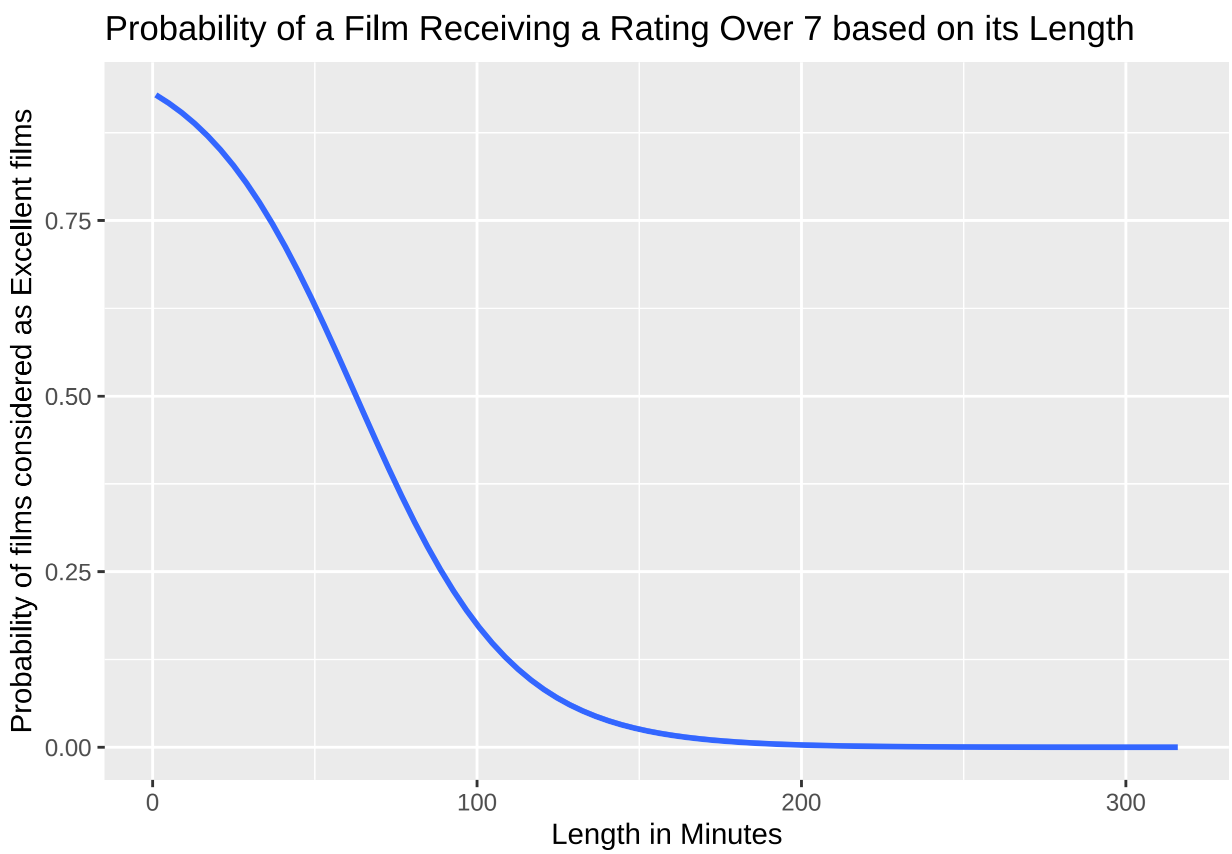
```r
# find coeff confidence intervals and plot the model
confint(model2) %>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	2.246300	3.0016387
length	-0.046199	-0.0373616

```r
plot_model(model2, show.values = TRUE, transform = NULL,
           title = "Length of Film- Log-Odds (Excellent films)", show.p = FALSE, vline.color = "cyan")
```



This tells us $\alpha = 2.61$ and that $\beta = -0.04$. We can see that the p-values of the coefficients are significant even at the highest level. Both 95% confidence intervals do not contain zero. We can conclude that this is a good performing model.



Model 3 The third model investigates the relationship between the budget of a film and whether or not the film received a rating over 7. The equation for this model is:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \text{budget} \tag{3}$$

where,

- p is the probability that the film is ranked over 7,
- budget is the budget of a film in \$1000000s
- α is the intercept value
- β is the regression coefficient.

```r
# create logistic regression model between over7 and the budget of a film and show summary
model3 = glm(over7 ~ budget, data = films,
             family = binomial(link = "logit"))
model3 %>%
  summary()
```

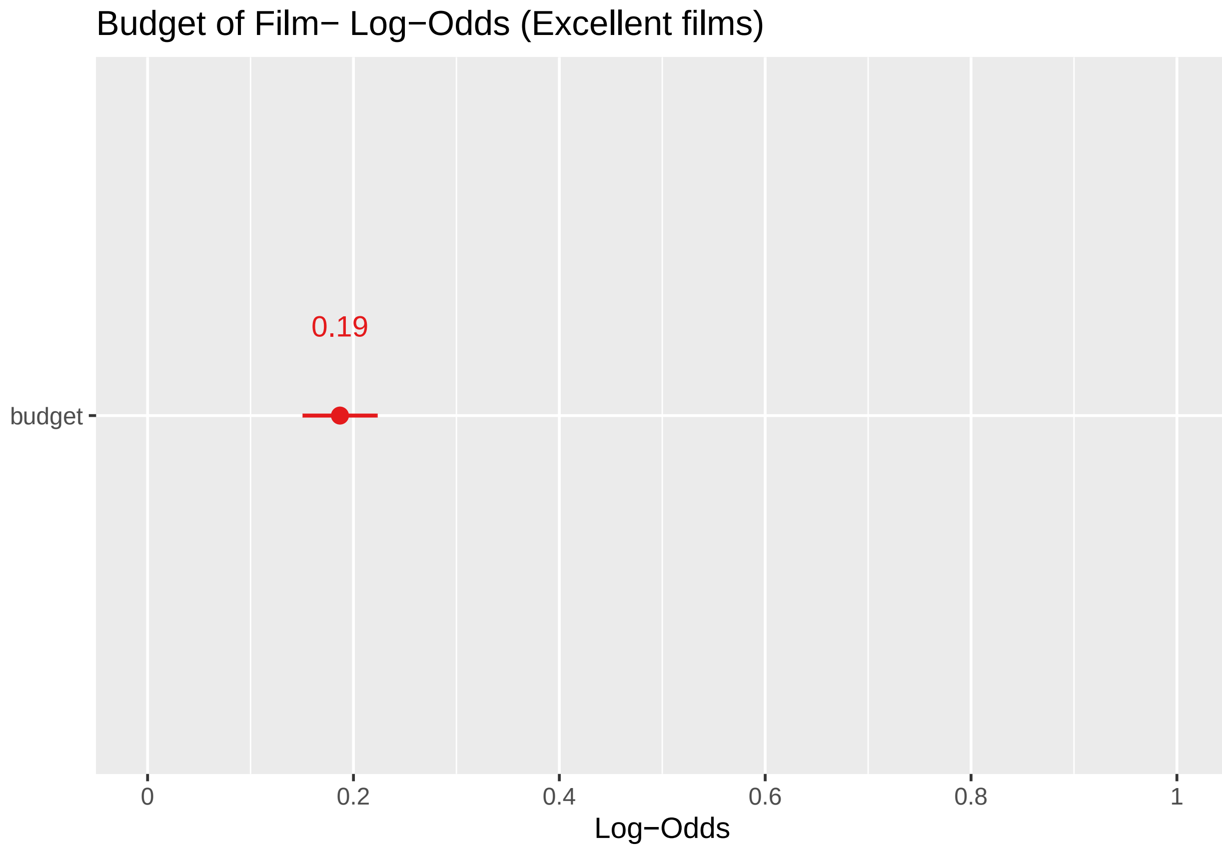
```
##
## Call:
## glm(formula = over7 ~ budget, family = binomial(link = "logit"),
##     data = films)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6079  -0.9151  -0.7210   1.2415   2.1882
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.98983    0.23639  -12.65   <2e-16 ***
## budget       0.18686    0.01849   10.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.5  on 1833  degrees of freedom
## Residual deviance: 2220.0  on 1832  degrees of freedom
## AIC: 2224
##
## Number of Fisher Scoring iterations: 4
```

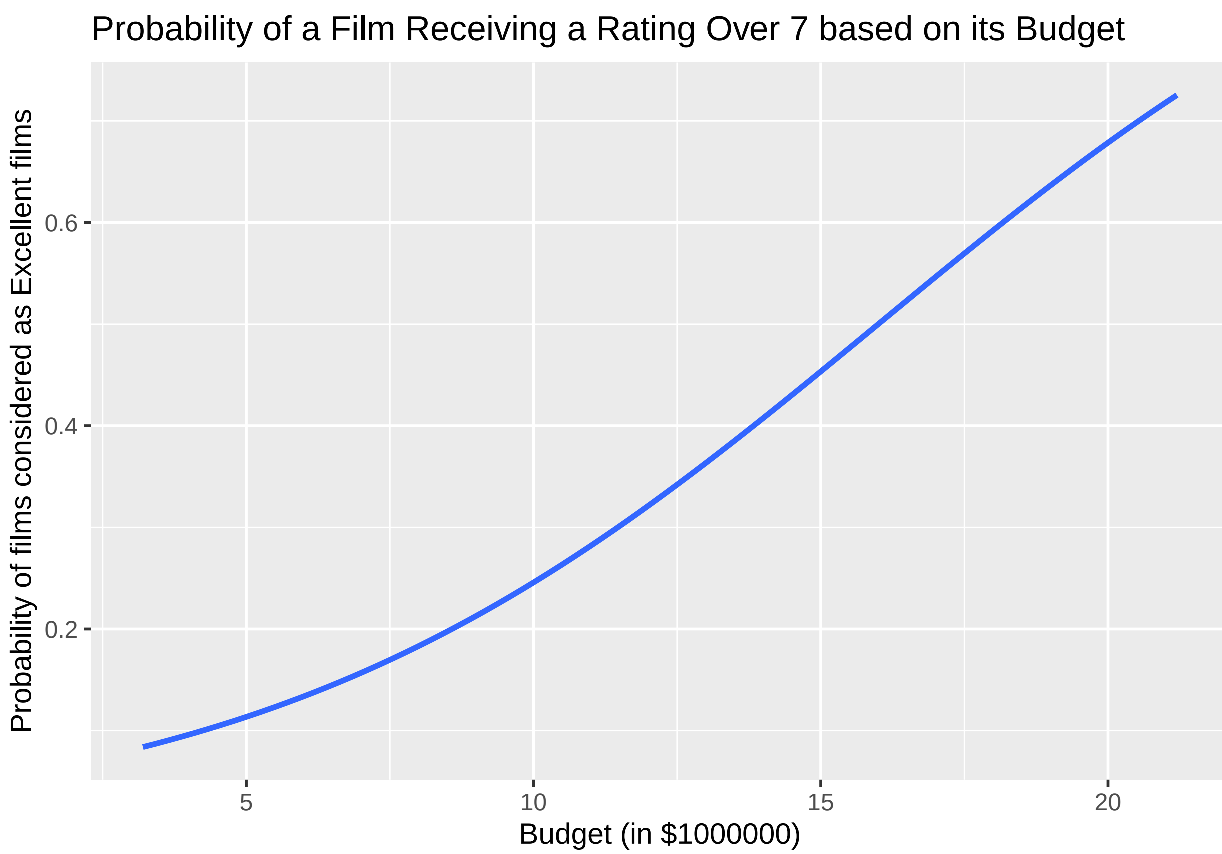
```
# create logistic regression model between over7 and the budget of a film and show summary
confint(model3) %>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	-3.4595018	-2.532449
budget	0.1509963	0.223504

```
plot_model(model3, show.values = TRUE, transform = NULL,
           title = "Budget of Film- Log-Odds (Excellent films)", show.p = FALSE, vline.color = "cyan")
```



This tells us α = -2.99 and that β = 0.19. We can see that the p-values of the coefficients are significant even at the highest level. Both 95% confidence intervals do not contain zero. We can conclude that this is a good performing model.



Model 4 The next model is investigating the relationship between the length of a film and whether or not the film received a rating over 7. The equation for this model is:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \cdot \log(\text{votes}) \tag{4}$$

where,

- p is the probability that the film is ranked over 7,
- votes is the number of positive votes the film received by viewers,
- α is the intercept value
- β is the regression coefficient.

```r
# create logistic regression model between over7 and the log of the number of positive votes
# the film received from viewers and show summary
model4 = glm(over7 ~ log(votes), data = films,
             family = binomial(link = "logit"))
model4 %>%
  summary()
```

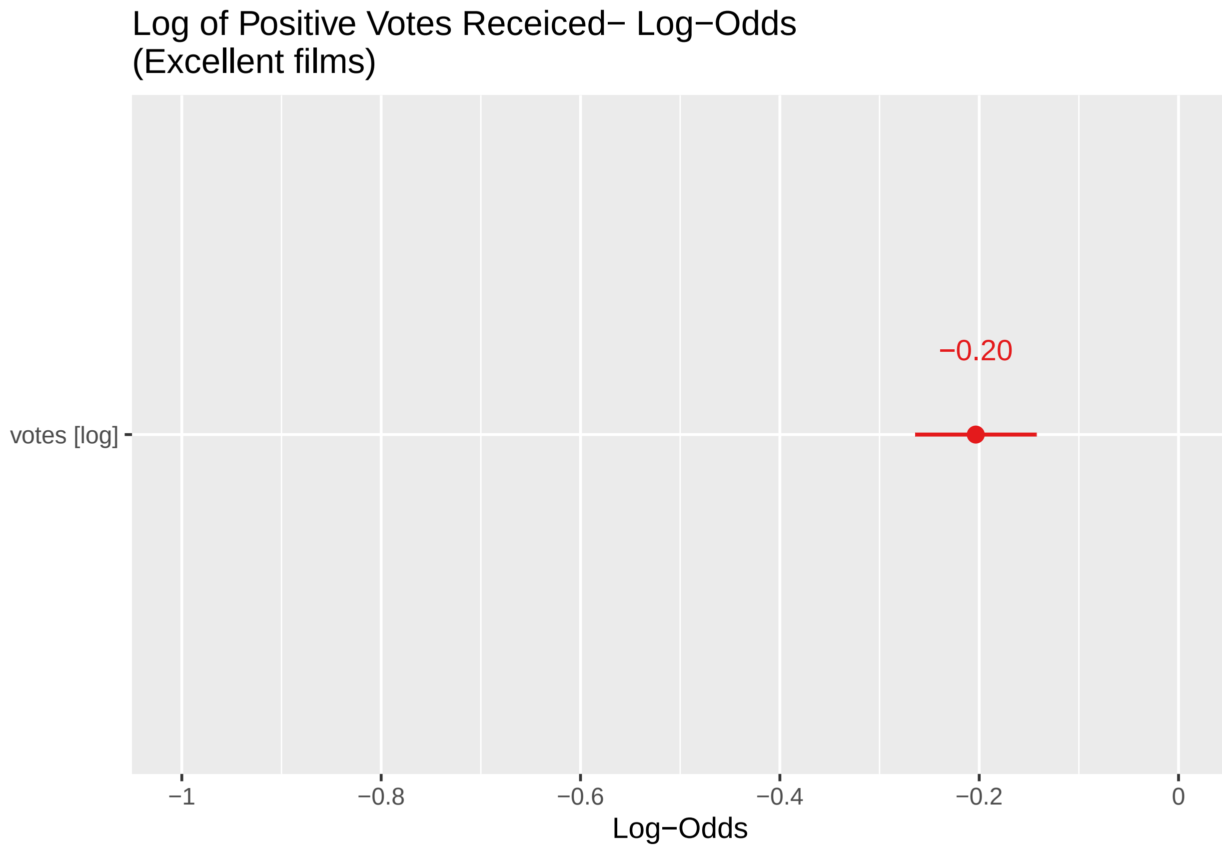
```
##
## Call:
## glm(formula = over7 ~ log(votes), family = binomial(link = "logit"),
##     data = films)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0675  -0.9450  -0.7992   1.3536   2.1834
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.06327    0.12310   0.514    0.607
## log(votes)  -0.20346    0.03103  -6.558 5.47e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.5  on 1833  degrees of freedom
## Residual deviance: 2284.2  on 1832  degrees of freedom
## AIC: 2288.2
##
## Number of Fisher Scoring iterations: 4
```

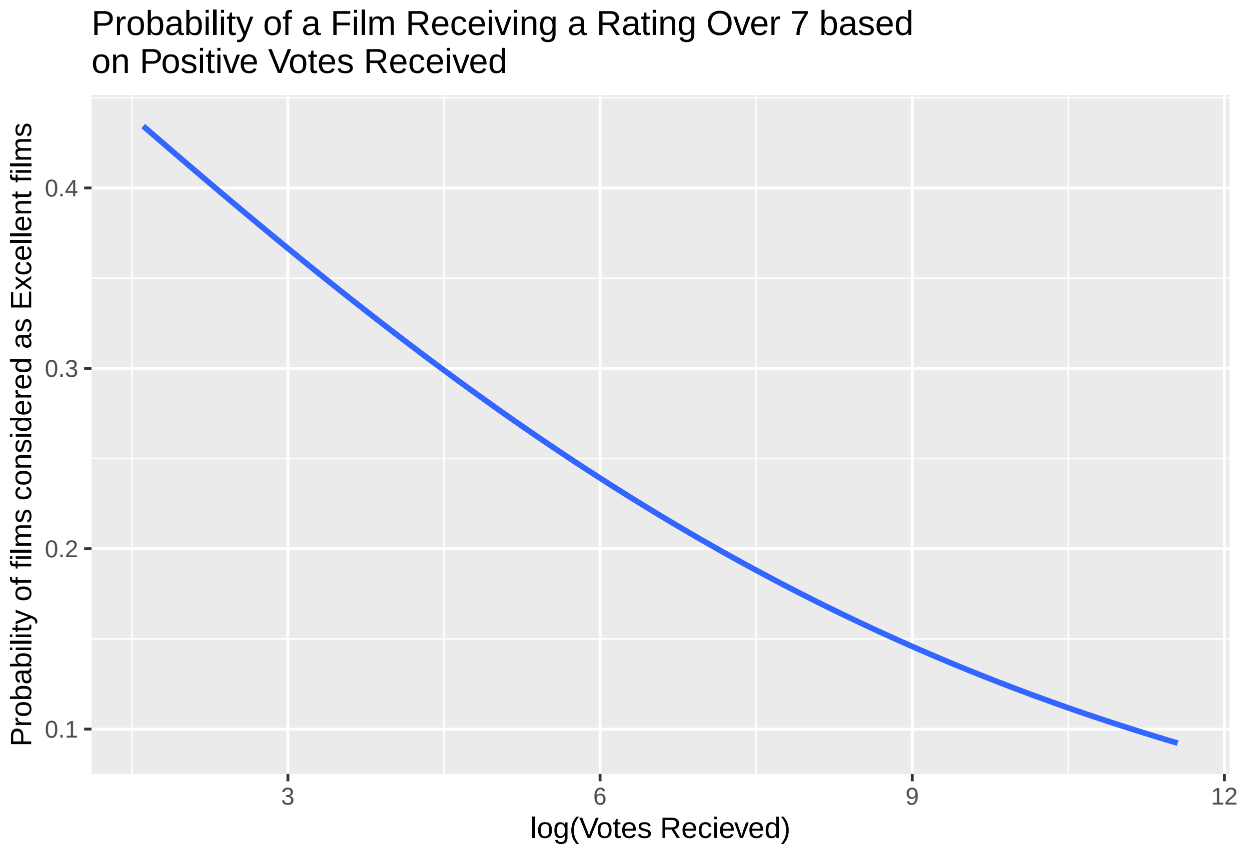
```
# find coeff confidence intervals and plot the model
confint(model4) %>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	-0.1769765	0.3057451
log(votes)	-0.2652308	-0.1435425

```
plot_model(model4, show.values = TRUE, transform = NULL,
           title = "Log of Positive Votes Receiced- Log-Odds (Excellent films)", show.p = FALSE, vline.c
```



This tells us $\alpha = 0.06$ and that $\beta = -0.2$. We can see that the p-values of the β coefficient is significant even at the highest level but the intercept is not. We can conclude that this is a model that performs okay but not as well as others.



We can compare the objective criteria of these models using the AIC and BIC. It appears that the second model (rating modeled with the movie length) it the best as it has the lowest AIC and BIC values.

```
# compare AIC and BIC of model1- model4
AIC(model1, model2, model3, model4)
```

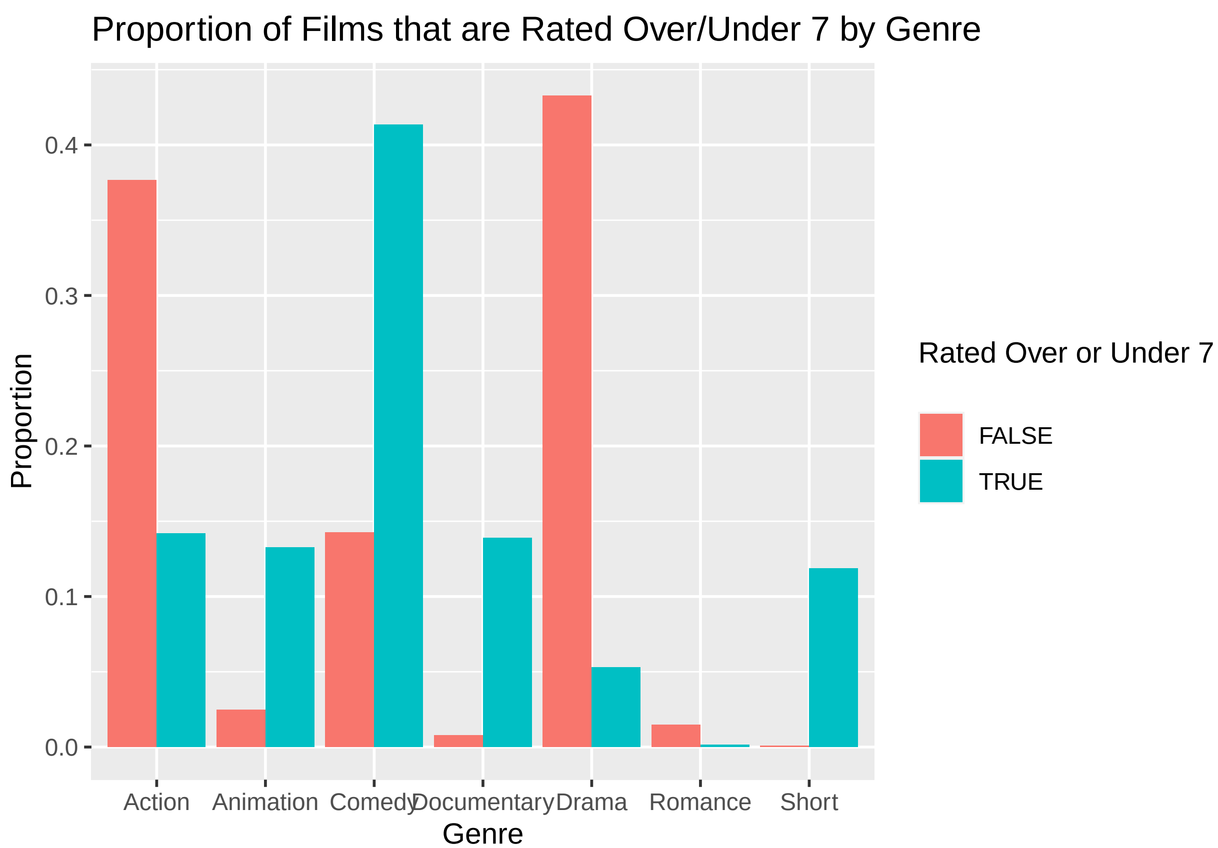
```
##        df      AIC
## model1  2 2332.500
## model2  2 1764.540
## model3  2 2224.033
## model4  2 2288.244
```

```
BIC(model1, model2, model3, model4)
```

```
##        df      BIC
## model1  2 2343.528
## model2  2 1775.569
## model3  2 2235.062
## model4  2 2299.272
```

Model 5 The next model is investigating the relationship between the genre of a film and whether or not the film received a rating over 7. We will look at the counts of films in each category an how many received an excellent score.

```
##   over7       Action   Animation       Comedy Documentary       Drama    Romance
##       0 37.7% (462)   2.3% (28) 14.4% (177)    0.7%   (9)  43.3% (531) 1.4% (17)
##       1 13.8%   (84) 13.3% (81) 41.9% (255)   13.5% (82)   5.1%   (31) 0.2%   (1)
##        Short
##    0.1%   (1)
##   12.3% (75)
```



The equation for this model is:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre} \tag{5}$$

where,

- p is the probability that the film is ranked over 7,
- votes is the number of positve votes the film recieved by viewers,
- α is the intercept value
- β_{genre} is the regression value for the i^th genre.

```
# plot proportions of films over and under 7 that are from each genre
model5 = glm(over7 ~ genre, data = films,
             family = binomial(link = "logit"))
model5 %>%
  summary()
```

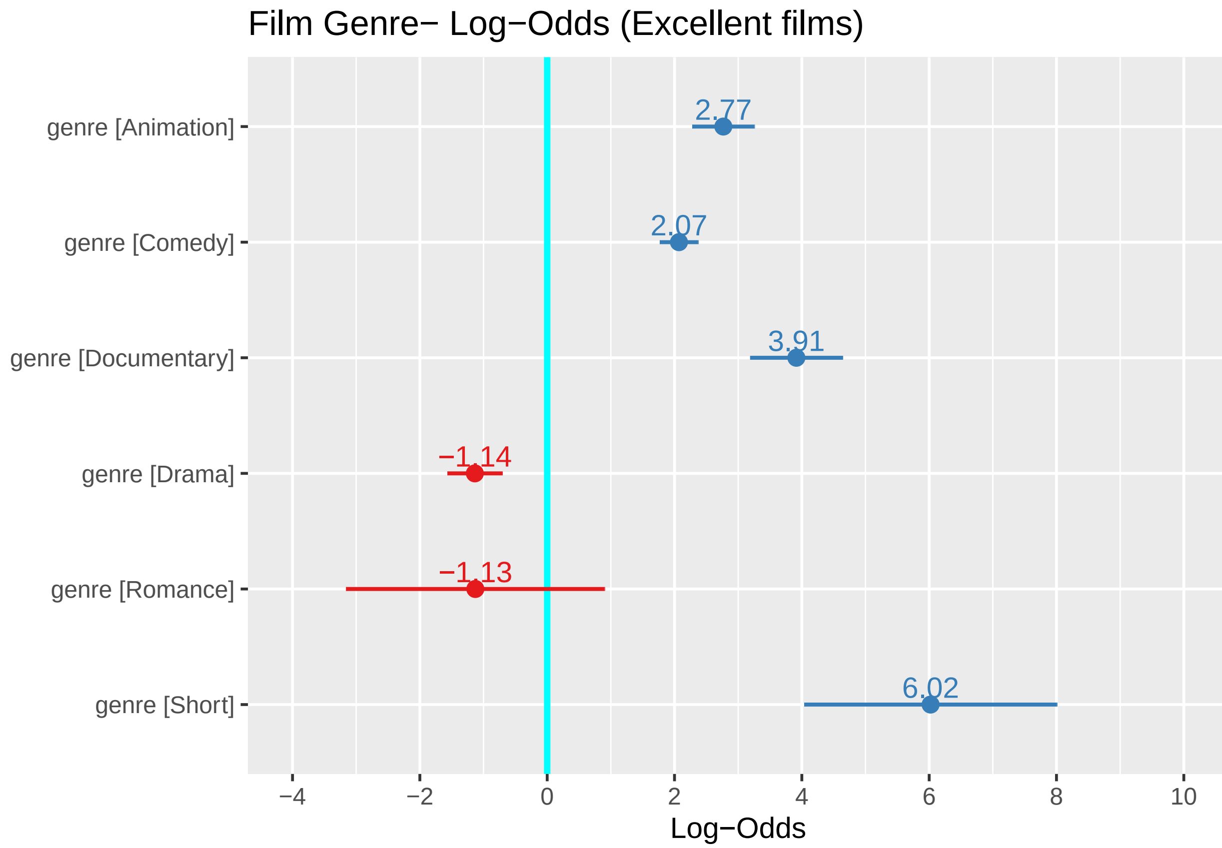
```
##
```

```
## Call:
## glm(formula = over7 ~ genre, family = binomial(link = "logit"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9430  -0.5780  -0.3369   0.4564   2.4073
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.7047     0.1186 -14.372  < 2e-16 ***
## genreAnimation     2.7670     0.2493  11.101  < 2e-16 ***
## genreComedy        2.0699     0.1538  13.462  < 2e-16 ***
## genreDocumentary   3.9142     0.3706  10.561  < 2e-16 ***
## genreDrama        -1.1360     0.2196  -5.174 2.29e-07 ***
## genreRomance      -1.1285     1.0358  -1.089    0.276
## genreShort         6.0222     1.0128   5.946 2.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.5  on 1833  degrees of freedom
## Residual deviance: 1494.7  on 1827  degrees of freedom
## AIC: 1508.7
##
## Number of Fisher Scoring iterations: 6
```

```
# find coeff confidence intervals and plot the model
confint(model5) %>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	-1.943783	-1.4782167
genreAnimation	2.290217	3.2699036
genreComedy	1.772459	2.3756085
genreDocumentary	3.238244	4.7069454
genreDrama	-1.579834	-0.7162757
genreRomance	-4.026532	0.4748886
genreShort	4.494912	8.8993975

```
plot_model(model5, show.values = TRUE, transform = NULL,
           title = "Film Genre- Log-Odds (Excellent films)", show.p = FALSE, vline.color = "cyan")
```



This tells us α = -1.70 and that β_i values are 2.77, 2.07, 3.91, -1.14, -1.13 and 6 for animation, comedy, documentary, drama, romance and short respectively. We can see that the p-values of every coefficient except the romance genre is significant even at the highest level. We can see that if a film is in the animation, comedy, documentary or short film categories it will improve the chances of the film getting a high score. We can conclude that this is a model that performs well and we can experiment with removing the romance genre films to see how it affects the model performance.

Full Model We are now going to look at the full model with every explanatory variable in the model. This model has the following equation:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre} + \beta_2 \cdot \log(\text{votes}) + \beta_3 \cdot \text{length} + \beta_4 \cdot \text{budget} + \beta_5 \cdot \text{year} \tag{6}$$

where,

- p is the probability that the film is ranked over 7,
- votes is the number of positve votes the film recieved by viewers,
- genre is the genre of the film,
- length is the length of the film in minutes,
- budget is the budget of the film in $1000000,
- α is the intercept value,
- β_{genre} is the regression value for the i^th genre,
- β_i is the regression value for the i^th variable.

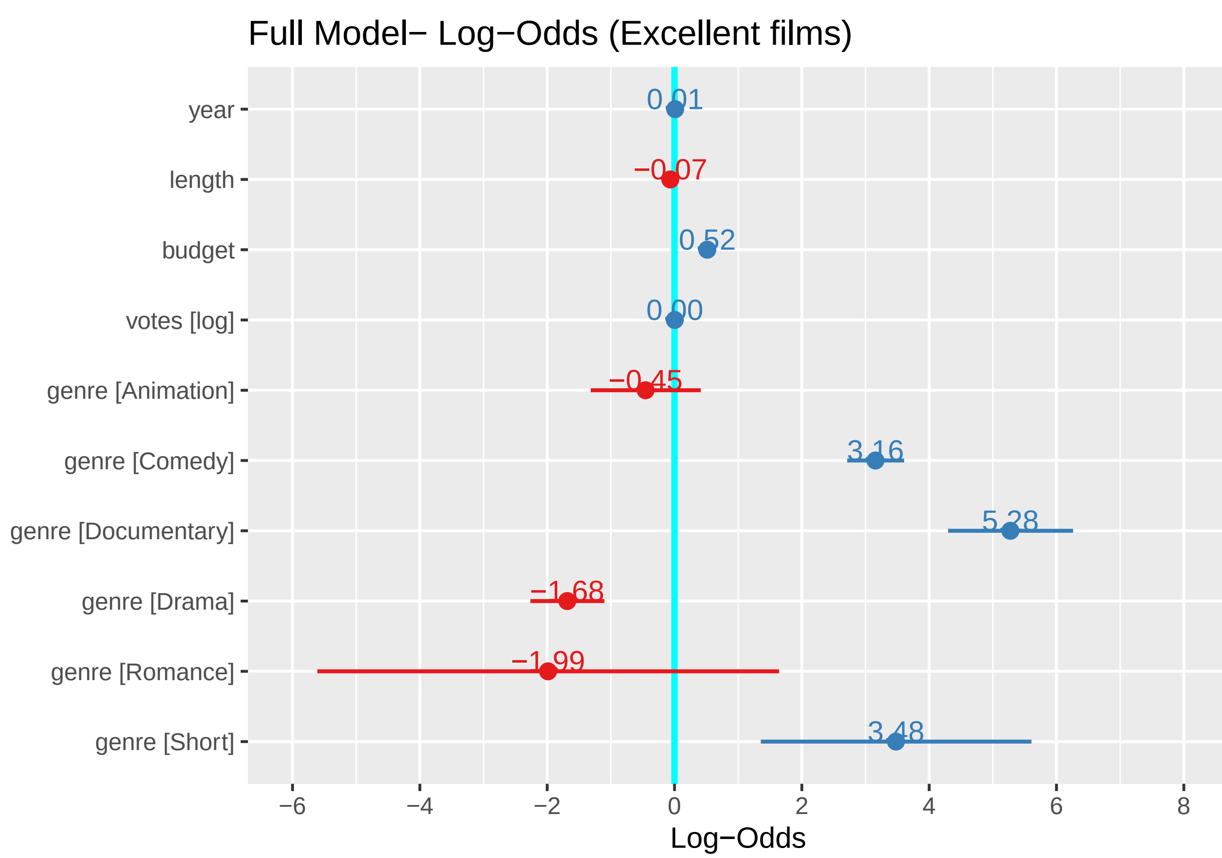
```r
# create logistic regression model between over7 and all of the explanatory variables
model6 = glm(over7 ~  year + length + budget + log(votes) + genre, data = films,
             family = binomial(link = "logit"))
model6 %>%
  summary()
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + log(votes) + genre,
##     family = binomial(link = "logit"), data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8467  -0.3347  -0.1085   0.1569   3.9304
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -23.543163   7.603127  -3.097  0.00196 **
## year               0.010425   0.003868   2.695  0.00704 **
## length            -0.066094   0.004881 -13.541  < 2e-16 ***
## budget             0.515096   0.037389  13.777  < 2e-16 ***
## log(votes)         0.003991   0.052281   0.076  0.93915
## genreAnimation    -0.454016   0.439159  -1.034  0.30122
## genreComedy        3.156805   0.225752  13.983  < 2e-16 ***
## genreDocumentary   5.275585   0.498513  10.583  < 2e-16 ***
## genreDrama        -1.684525   0.294512  -5.720 1.07e-08 ***
## genreRomance      -1.986251   1.847771  -1.075  0.28240
## genreShort         3.478327   1.082201   3.214  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  936.87  on 1823  degrees of freedom
## AIC: 958.87
##
## Number of Fisher Scoring iterations: 7
```

```r
# find coeff confidence intervals and plot the model
confint(model6) %>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	-38.5748578	-8.7408209
year	0.0028891	0.0180676
length	-0.0760037	-0.0568476
budget	0.4439093	0.5906270
log(votes)	-0.0991066	0.1060937
genreAnimation	-1.3237972	0.4008454
genreComedy	2.7251537	3.6110335
genreDocumentary	4.3524877	6.3194921
genreDrama	-2.2836194	-1.1257113
genreRomance	-5.8143422	0.7705535
genreShort	1.7470723	6.4232203

```
plot_model(model6, show.values = TRUE, transform = NULL,
           title = "Full Model- Log-Odds (Excellent films)", show.p = FALSE, vline.color = "cyan")
```



We can see that many of the variables are significant but there are also a substantial number that are not even at the 10% level. Using stepwise regression we can look to find an optimal model. As we found in the table earlier the "best model" includes length, budget, genre and year.

```
# find optimal model using stepwise regression- try forward, backwards and both directions
logit.step.forward = step(model6,direction="forward")
```

```
## Start:  AIC=958.87
## over7 ~ year + length + budget + log(votes) + genre
```

```r
summary(logit.step.forward)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + log(votes) + genre,
##     family = binomial(link = "logit"), data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8467  -0.3347  -0.1085   0.1569   3.9304
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -23.543163   7.603127  -3.097  0.00196 **
## year              0.010425   0.003868   2.695  0.00704 **
## length           -0.066094   0.004881 -13.541  < 2e-16 ***
## budget            0.515096   0.037389  13.777  < 2e-16 ***
## log(votes)        0.003991   0.052281   0.076  0.93915
## genreAnimation   -0.454016   0.439159  -1.034  0.30122
## genreComedy       3.156805   0.225752  13.983  < 2e-16 ***
## genreDocumentary  5.275585   0.498513  10.583  < 2e-16 ***
## genreDrama       -1.684525   0.294512  -5.720 1.07e-08 ***
## genreRomance     -1.986251   1.847771  -1.075  0.28240
## genreShort        3.478327   1.082201   3.214  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  936.87  on 1823  degrees of freedom
## AIC: 958.87
##
## Number of Fisher Scoring iterations: 7
```

```r
logit.step.backward = step(model6,direction="backward")
```

```
## Start:  AIC=958.87
## over7 ~ year + length + budget + log(votes) + genre
##
##              Df Deviance     AIC
## - log(votes)  1   936.88  956.88
## <none>            936.87  958.87
## - year        1   944.25  964.25
## - budget      1  1217.76 1237.76
## - length      1  1236.52 1256.52
## - genre       6  1578.07 1588.07
##
## Step:  AIC=956.88
## over7 ~ year + length + budget + genre
##
##          Df Deviance     AIC
```

```
## <none>         936.88  956.88
## - year     1   944.64  962.64
## - budget   1  1217.83 1235.83
## - length   1  1263.40 1281.40
## - genre    6  1581.53 1589.53
```

```r
summary(logit.step.backward)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + genre, family = binomial(link = "logit"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8460  -0.3352  -0.1081   0.1569   3.9271
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -23.647995   7.477584  -3.163  0.00156 **
## year               0.010482   0.003795   2.762  0.00574 **
## length            -0.066013   0.004761 -13.864  < 2e-16 ***
## budget             0.515046   0.037383  13.778  < 2e-16 ***
## genreAnimation    -0.449210   0.434572  -1.034  0.30128
## genreComedy        3.159643   0.222730  14.186  < 2e-16 ***
## genreDocumentary   5.272927   0.497292  10.603  < 2e-16 ***
## genreDrama        -1.684283   0.294467  -5.720 1.07e-08 ***
## genreRomance      -1.981099   1.845435  -1.074  0.28304
## genreShort         3.478463   1.082170   3.214  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  936.88  on 1824  degrees of freedom
## AIC: 956.88
##
## Number of Fisher Scoring iterations: 7
```

```r
logit.stepwise = step(model6,direction="both")
```

```
## Start:  AIC=958.87
## over7 ~ year + length + budget + log(votes) + genre
##
##              Df Deviance     AIC
## - log(votes)  1   936.88  956.88
## <none>            936.87  958.87
## - year        1   944.25  964.25
## - budget      1  1217.76 1237.76
## - length      1  1236.52 1256.52
## - genre       6  1578.07 1588.07
##
```

```
## Step:  AIC=956.88
## over7 ~ year + length + budget + genre
##
##               Df Deviance     AIC
## <none>              936.88  956.88
## + log(votes)  1   936.87  958.87
## - year        1   944.64  962.64
## - budget      1  1217.83 1235.83
## - length      1  1263.40 1281.40
## - genre       6  1581.53 1589.53
```

```
summary(logit.stepwise)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + genre, family = binomial(link = "logit"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8460  -0.3352  -0.1081   0.1569   3.9271
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -23.647995   7.477584  -3.163  0.00156 **
## year               0.010482   0.003795   2.762  0.00574 **
## length            -0.066013   0.004761 -13.864  < 2e-16 ***
## budget             0.515046   0.037383  13.778  < 2e-16 ***
## genreAnimation    -0.449210   0.434572  -1.034  0.30128
## genreComedy        3.159643   0.222730  14.186  < 2e-16 ***
## genreDocumentary   5.272927   0.497292  10.603  < 2e-16 ***
## genreDrama        -1.684283   0.294467  -5.720 1.07e-08 ***
## genreRomance      -1.981099   1.845435  -1.074  0.28304
## genreShort         3.478463   1.082170   3.214  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  936.88  on 1824  degrees of freedom
## AIC: 956.88
##
## Number of Fisher Scoring iterations: 7
```

```
summary(stepAIC(model6))
```

```
## Start:  AIC=958.87
## over7 ~ year + length + budget + log(votes) + genre
##
##               Df Deviance     AIC
## - log(votes)  1   936.88  956.88
## <none>              936.87  958.87
```

```
## - year        1   944.25  964.25
## - budget      1  1217.76 1237.76
## - length      1  1236.52 1256.52
## - genre       6  1578.07 1588.07
##
## Step:  AIC=956.88
## over7 ~ year + length + budget + genre
##
##          Df Deviance     AIC
## <none>        936.88  956.88
## - year    1   944.64  962.64
## - budget  1  1217.83 1235.83
## - length  1  1263.40 1281.40
## - genre   6  1581.53 1589.53
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + genre, family = binomial(link = "logit"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8460  -0.3352  -0.1081   0.1569   3.9271
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -23.647995   7.477584  -3.163  0.00156 **
## year               0.010482   0.003795   2.762  0.00574 **
## length            -0.066013   0.004761 -13.864  < 2e-16 ***
## budget             0.515046   0.037383  13.778  < 2e-16 ***
## genreAnimation    -0.449210   0.434572  -1.034  0.30128
## genreComedy        3.159643   0.222730  14.186  < 2e-16 ***
## genreDocumentary   5.272927   0.497292  10.603  < 2e-16 ***
## genreDrama        -1.684283   0.294467  -5.720 1.07e-08 ***
## genreRomance      -1.981099   1.845435  -1.074  0.28304
## genreShort         3.478463   1.082170   3.214  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  936.88  on 1824  degrees of freedom
## AIC: 956.88
##
## Number of Fisher Scoring iterations: 7
```

Best AIC Model We are now going to look at the model that has the lowest AIC value of all the possible models. This model has the following equation:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre} + \beta_2 \cdot \text{length} + \beta_3 \cdot \text{budget} + \beta_4 \cdot \text{year} \tag{7}$$

where,

- p is the probability that the film is ranked over 7,
- genre is the genre of the film,
- length is the length of the film in minutes,
- budget is the budget of the film in $1000000,
- year in the year the film was released,
- α is the intercept value,
- β_{genre} is the regression value for the i^th genre,
- β_i is the regression value for the i^th variable.

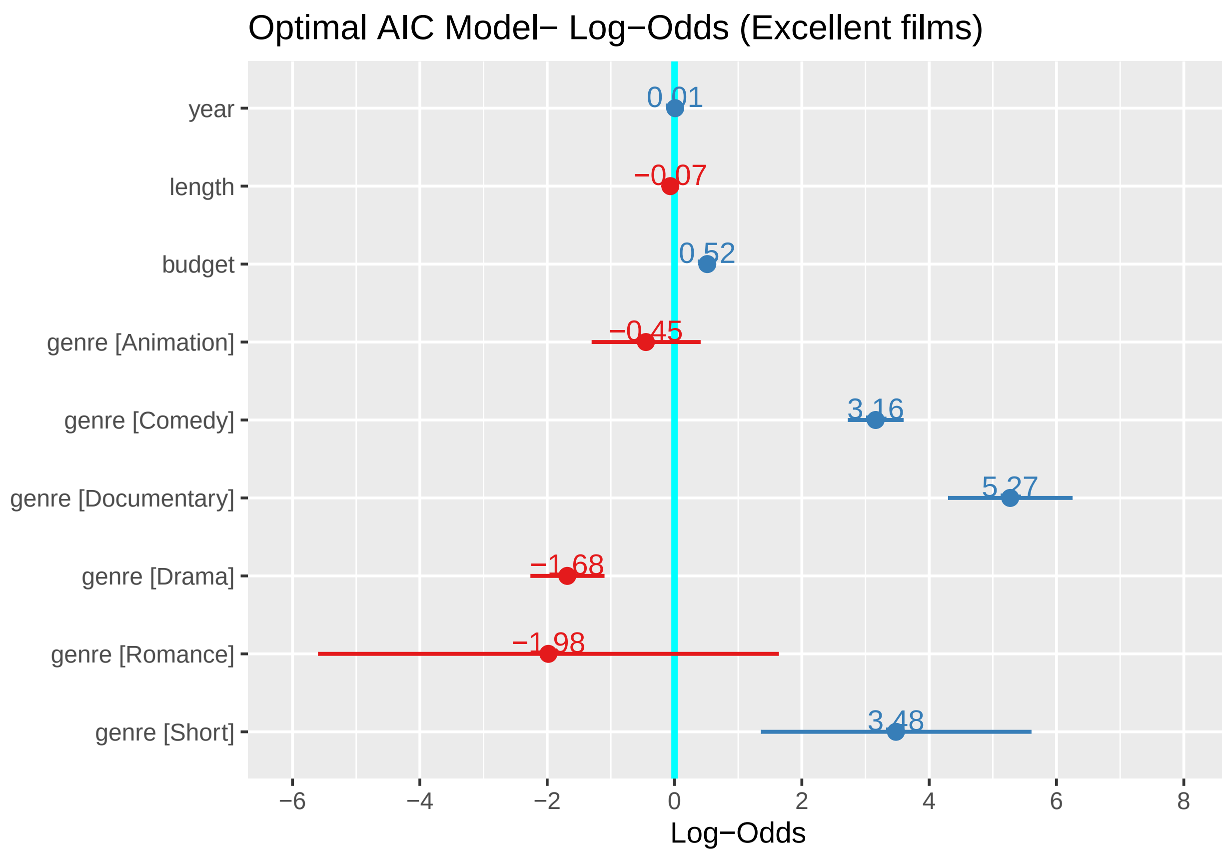
```
# create logistic regression model that was found to have the lowest AIC
model7 = glm(over7 ~  year + length + budget + genre, data = films,
             family = binomial(link = "logit"))
model7 %>%
  summary()
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + genre, family = binomial(link = "logit"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8460  -0.3352  -0.1081   0.1569   3.9271
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -23.647995   7.477584  -3.163  0.00156 **
## year               0.010482   0.003795   2.762  0.00574 **
## length            -0.066013   0.004761 -13.864  < 2e-16 ***
## budget             0.515046   0.037383  13.778  < 2e-16 ***
## genreAnimation    -0.449210   0.434572  -1.034  0.30128
## genreComedy        3.159643   0.222730  14.186  < 2e-16 ***
## genreDocumentary   5.272927   0.497292  10.603  < 2e-16 ***
## genreDrama        -1.684283   0.294467  -5.720 1.07e-08 ***
## genreRomance      -1.981099   1.845435  -1.074  0.28304
## genreShort         3.478463   1.082170   3.214  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  936.88  on 1824  degrees of freedom
## AIC: 956.88
##
## Number of Fisher Scoring iterations: 7
```

```
# find coeff confidence intervals and plot the model
confint(model7) %>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	-38.4417259	-9.1001175
year	0.0030943	0.0179854
length	-0.0756854	-0.0570002
budget	0.4438714	0.5905650
genreAnimation	-1.3095905	0.3967681
genreComedy	2.7337317	3.6077344
genreDocumentary	4.3524542	6.3148446
genreDrama	-2.2832623	-1.1255376
genreRomance	-5.8051736	0.7719487
genreShort	1.7473103	6.4233255

```
plot_model(model7, show.values = TRUE, transform = NULL,
           title = "Optimal AIC Model- Log-Odds (Excellent films)", show.p = FALSE, vline.color = "cyan
```



We can see that all of the variables are significant to a high level except the animation and romance genres.

Optimal Model with only Significant Variables We are going to look at the exact same model as before but we will remove the categories from the data set that are not significant in the model (the animation and romance genre).

```
# filter data to remove insignificant genres
genre.noRomanceandAnimation = films %>%
  filter(genre != "Romance" ) %>%
  filter(genre != "Animation") %>%
  drop_na
```

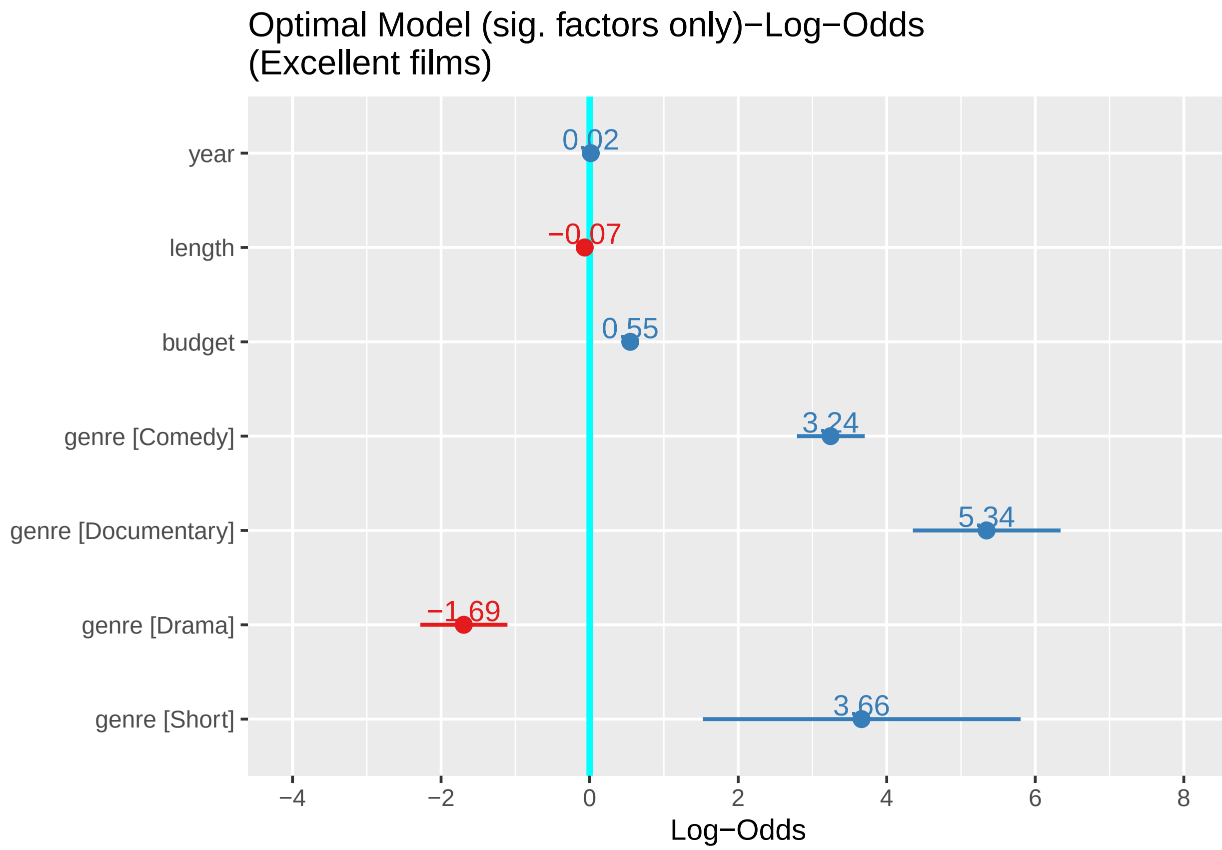
```r
# create logistic regression model that was found to have lowest AIC but with non significant variables

model8 = glm(over7 ~  year + length + budget + genre, data = genre.noRomanceandAnimation,
             family = binomial(link = "logit"))
model8 %>%
  summary()
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + genre, family = binomial(link = "logit"),
##     data = genre.noRomanceandAnimation)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8475  -0.3409  -0.1115   0.0969   3.9369
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -33.480306   7.955409  -4.208 2.57e-05 ***
## year              0.015229   0.004030   3.779 0.000157 ***
## length           -0.066091   0.005086 -12.995  < 2e-16 ***
## budget            0.547144   0.040452  13.526  < 2e-16 ***
## genreComedy       3.244307   0.230305  14.087  < 2e-16 ***
## genreDocumentary  5.344683   0.505907  10.565  < 2e-16 ***
## genreDrama       -1.694471   0.297013  -5.705 1.16e-08 ***
## genreShort        3.661236   1.090358   3.358 0.000786 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2110.12  on 1706  degrees of freedom
## Residual deviance:  844.14  on 1699  degrees of freedom
## AIC: 860.14
##
## Number of Fisher Scoring iterations: 7
```

```r
# find coeff confidence intervals and plot the model
confint(model8) %>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	-49.2573755	-18.0373440
year	0.0073999	0.0232139
length	-0.0764343	-0.0564756
budget	0.4703544	0.6291131
genreComedy	2.8049669	3.7088028
genreDocumentary	4.4087112	6.4046152
genreDrama	-2.2993607	-1.1314241
genreShort	1.9075070	6.6147107

```
plot_model(model8, show.values = TRUE, transform = NULL,
           title = "Optimal Model (sig. factors only)-Log-Odds (Excellent films)", show.p = FALSE, vlin
```



Now every single variable in this model is significant to the highest level and the AIC is significantly less.

Simplified Optimal Model When iterating through all of the possible models we found that the model that includes just the length, budget and genre of a film performs very similarly to the optimal model but it has the added benefit of being simpler. This model has the following equation:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_{genre} + \beta_2 \cdot \text{length} + \beta_3 \cdot \text{budget} \tag{8}$$

where,

- p is the probability that the film is ranked over 7,
- genre is the genre of the film,
- length is the length of the film in minutes,
- budget is the budget of the film in \$1000000,
- α is the intercept value,
- β_{genre} is the regression value for the i^th genre,
- β_i is the regression value for the i^th variable.

```
# create logistic regression model that is close to having best AIC but is simpler
model9 = glm(over7 ~ length + budget + genre, data = films,
             family = binomial(link = "logit"))
```

```r
model9 %>%
  summary()
```

```
##
## Call:
## glm(formula = over7 ~ length + budget + genre, family = binomial(link = "logit"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9333  -0.3446  -0.1115   0.1661   3.7196
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.066328   0.529523  -5.791 7.01e-09 ***
## length           -0.063254   0.004569 -13.844  < 2e-16 ***
## budget            0.508356   0.036839  13.800  < 2e-16 ***
## genreAnimation   -0.446196   0.433302  -1.030 0.303123
## genreComedy       3.115703   0.220929  14.103  < 2e-16 ***
## genreDocumentary  5.372847   0.494067  10.875  < 2e-16 ***
## genreDrama       -1.637473   0.288105  -5.684 1.32e-08 ***
## genreRomance     -1.979755   1.676791  -1.181 0.237729
## genreShort        3.705407   1.081679   3.426 0.000613 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  944.64  on 1825  degrees of freedom
## AIC: 962.64
##
## Number of Fisher Scoring iterations: 7
```

```r
# find coeff confidence intervals and plot the model
confint(model9) %>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	-4.1166932	-2.0384687
length	-0.0725270	-0.0545981
budget	0.4381783	0.5827352
genreAnimation	-1.3041964	0.3970217
genreComedy	2.6930694	3.5599906
genreDocumentary	4.4586093	6.4084369
genreDrama	-2.2220598	-1.0896041
genreRomance	-5.6272827	0.6036283
genreShort	1.9773583	6.6499762

```r
plot_model(model9, show.values = TRUE, transform = NULL,
           title = "Simplified Optimal Model- Log-Odds (Excellent films)", show.p = FALSE, vline.color =
```

This gives very similar results to the optimal model but it is simpler.

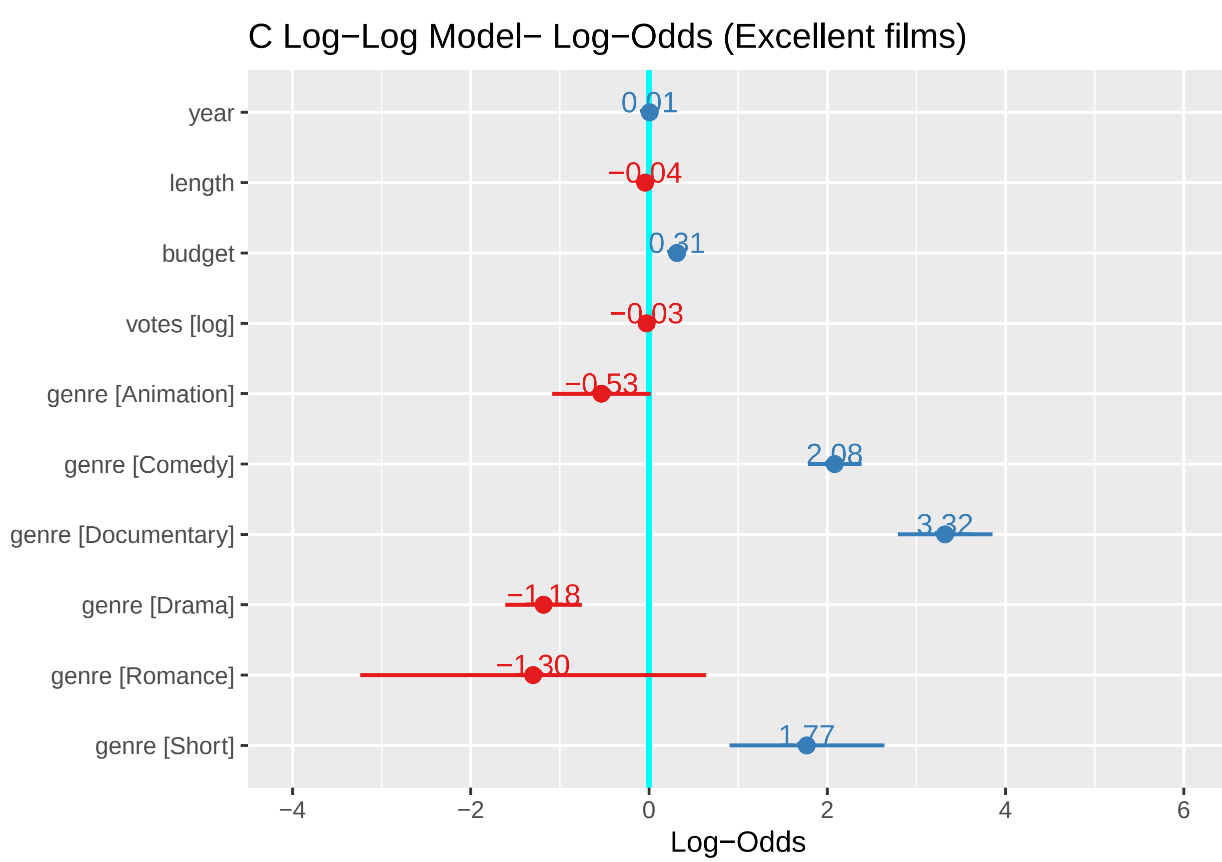
C log-log and Probit Models We will now try using c log-log and a probit models to investigate the relationship between the explanatory variables and a film getting a score above 7. We will use all the explanatory variables in this model. We will also use stepwise regression to find the most optimal models for each method.

```
# create logistic regression model with all of the explanatory variables but use c log-log regression
model10 <- glm(over7~ year + length + budget + log(votes) + genre, data = films, family = binomial(link

summary(model10)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + log(votes) + genre,
##     family = binomial(link = "cloglog"), data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8003  -0.4149  -0.1935   0.0197   3.4125
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -16.224919   4.828015  -3.361 0.000778 ***
## year                 0.007169   0.002458   2.916 0.003541 **
```

```
## length              -0.043637   0.003125 -13.962  < 2e-16 ***
## budget               0.313719   0.023026  13.624  < 2e-16 ***
## log(votes)          -0.026946   0.035538  -0.758 0.448309    
## genreAnimation      -0.534119   0.280560  -1.904 0.056941 .  
## genreComedy          2.082239   0.151294  13.763  < 2e-16 ***
## genreDocumentary     3.320750   0.268756  12.356  < 2e-16 ***
## genreDrama          -1.183308   0.217999  -5.428 5.70e-08 ***
## genreRomance        -1.300908   0.988529  -1.316 0.188173    
## genreShort           1.769844   0.441868   4.005 6.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  982.98  on 1823  degrees of freedom
## AIC: 1005
##
## Number of Fisher Scoring iterations: 9
```

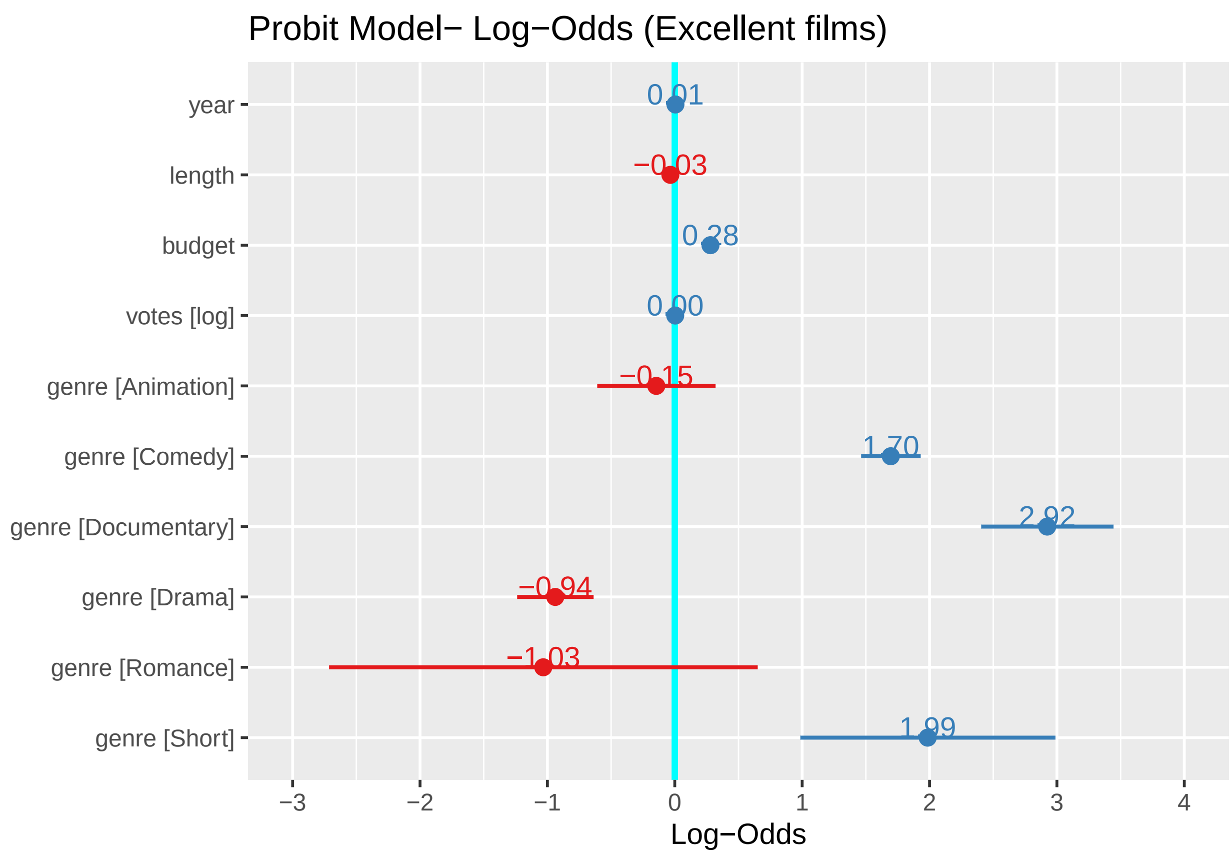
```
# plot the model
plot_model(model10, show.values = TRUE, transform = NULL,
           title = "C Log-Log Model- Log-Odds (Excellent films)", show.p = FALSE, vline.color = "cyan")
```



```r
# create logistic regression model with all of the explanatory variables but use probut regression
model11 <- glm(over7~ year + length + budget + log(votes) + genre, data = films, family = binomial(link

summary(model11)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + log(votes) + genre,
##     family = binomial(link = "probit"), data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8916  -0.3592  -0.0725   0.1351   4.5417
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -11.807117   4.149191  -2.846  0.00443 **
## year                0.005098   0.002113   2.413  0.01581 *
## length             -0.034507   0.002500 -13.803  < 2e-16 ***
## budget              0.280107   0.019439  14.409  < 2e-16 ***
## log(votes)          0.003435   0.028752   0.119  0.90490
## genreAnimation     -0.145286   0.236044  -0.616  0.53822
## genreComedy         1.695746   0.118291  14.335  < 2e-16 ***
## genreDocumentary    2.923809   0.263985  11.076  < 2e-16 ***
## genreDrama         -0.938371   0.152085  -6.170 6.83e-10 ***
## genreRomance       -1.032460   0.857481  -1.204  0.22857
## genreShort          1.985399   0.509786   3.895 9.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.5  on 1833  degrees of freedom
## Residual deviance:  945.9  on 1823  degrees of freedom
## AIC: 967.9
##
## Number of Fisher Scoring iterations: 8
```

```r
# plot the model
plot_model(model11, show.values = TRUE, transform = NULL,
           title = "Probit Model- Log-Odds (Excellent films)", show.p = FALSE, vline.color = "cyan")
```




```r
# find optimal models using c log-log and probit regression
cloglog.step.forward = step(model10,direction="forward")
```

```
## Start:  AIC=1004.98
## over7 ~ year + length + budget + log(votes) + genre
```

```r
summary(cloglog.step.forward)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + log(votes) + genre,
##     family = binomial(link = "cloglog"), data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8003  -0.4149  -0.1935   0.0197   3.4125
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -16.224919   4.828015  -3.361 0.000778 ***
## year               0.007169   0.002458   2.916 0.003541 **
## length            -0.043637   0.003125 -13.962  < 2e-16 ***
## budget             0.313719   0.023026  13.624  < 2e-16 ***
## log(votes)        -0.026946   0.035538  -0.758 0.448309
## genreAnimation    -0.534119   0.280560  -1.904 0.056941 .
```

```
## genreComedy        2.082239   0.151294  13.763  < 2e-16 ***
## genreDocumentary   3.320750   0.268756  12.356  < 2e-16 ***
## genreDrama        -1.183308   0.217999  -5.428 5.70e-08 ***
## genreRomance      -1.300908   0.988529  -1.316 0.188173
## genreShort         1.769844   0.441868   4.005 6.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  982.98  on 1823  degrees of freedom
## AIC: 1005
##
## Number of Fisher Scoring iterations: 9
```

```
cloglog.step.backward = step(model10,direction="backward")
```

```
## Start:  AIC=1004.98
## over7 ~ year + length + budget + log(votes) + genre
##
##              Df Deviance    AIC
## - log(votes)  1   983.56 1003.6
## <none>            982.98 1005.0
## - year        1   992.64 1012.6
## - budget      1  1241.15 1261.2
## - length      1  1246.67 1266.7
## - genre       6  1586.76 1596.8
##
## Step:  AIC=1003.56
## over7 ~ year + length + budget + genre
##
##          Df Deviance    AIC
## <none>        983.56 1003.6
## - year    1   992.66 1010.7
## - budget  1  1241.19 1259.2
## - length  1  1282.16 1300.2
## - genre   6  1589.87 1597.9
```

```
summary(cloglog.step.backward)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + genre, family = binomial(link = "cloglog"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7986  -0.4179  -0.1945   0.0198   3.4423
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -15.590168   4.766679  -3.271  0.00107 **
```

```
## year               0.006833   0.002421   2.822  0.00477 **
## length            -0.044227   0.003076 -14.379  < 2e-16 ***
## budget             0.312552   0.023008  13.584  < 2e-16 ***
## genreAnimation    -0.568963   0.278537  -2.043  0.04108 *
## genreComedy        2.053269   0.148328  13.843  < 2e-16 ***
## genreDocumentary   3.334064   0.268766  12.405  < 2e-16 ***
## genreDrama        -1.185894   0.218220  -5.434 5.50e-08 ***
## genreRomance      -1.335993   0.990063  -1.349  0.17721
## genreShort         1.763519   0.444025   3.972 7.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  983.56  on 1824  degrees of freedom
## AIC: 1003.6
##
## Number of Fisher Scoring iterations: 9
```

```
cloglog.stepwise = step(model10,direction="both")
```

```
## Start:  AIC=1004.98
## over7 ~ year + length + budget + log(votes) + genre
##
##              Df Deviance    AIC
## - log(votes)  1   983.56 1003.6
## <none>            982.98 1005.0
## - year        1   992.64 1012.6
## - budget      1  1241.15 1261.2
## - length      1  1246.67 1266.7
## - genre       6  1586.76 1596.8
##
## Step:  AIC=1003.56
## over7 ~ year + length + budget + genre
##
##              Df Deviance    AIC
## <none>            983.56 1003.6
## + log(votes)  1   982.98 1005.0
## - year        1   992.66 1010.7
## - budget      1  1241.19 1259.2
## - length      1  1282.16 1300.2
## - genre       6  1589.87 1597.9
```

```
summary(cloglog.stepwise)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + genre, family = binomial(link = "cloglog"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -3.7986  -0.4179  -0.1945   0.0198   3.4423
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -15.590168   4.766679  -3.271  0.00107 **
## year               0.006833   0.002421   2.822  0.00477 **
## length            -0.044227   0.003076 -14.379  < 2e-16 ***
## budget             0.312552   0.023008  13.584  < 2e-16 ***
## genreAnimation    -0.568963   0.278537  -2.043  0.04108 *
## genreComedy        2.053269   0.148328  13.843  < 2e-16 ***
## genreDocumentary   3.334064   0.268766  12.405  < 2e-16 ***
## genreDrama        -1.185894   0.218220  -5.434 5.50e-08 ***
## genreRomance      -1.335993   0.990063  -1.349  0.17721
## genreShort         1.763519   0.444025   3.972 7.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  983.56  on 1824  degrees of freedom
## AIC: 1003.6
##
## Number of Fisher Scoring iterations: 9
```

```r
probit.step.forward = step(model11,direction="forward")
```

```
## Start:  AIC=967.9
## over7 ~ year + length + budget + log(votes) + genre
```

```r
summary(probit.step.forward)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + log(votes) + genre,
##     family = binomial(link = "probit"), data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8916  -0.3592  -0.0725   0.1351   4.5417
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -11.807117   4.149191  -2.846  0.00443 **
## year               0.005098   0.002113   2.413  0.01581 *
## length            -0.034507   0.002500 -13.803  < 2e-16 ***
## budget             0.280107   0.019439  14.409  < 2e-16 ***
## log(votes)         0.003435   0.028752   0.119  0.90490
## genreAnimation    -0.145286   0.236044  -0.616  0.53822
## genreComedy        1.695746   0.118291  14.335  < 2e-16 ***
## genreDocumentary   2.923809   0.263985  11.076  < 2e-16 ***
## genreDrama        -0.938371   0.152085  -6.170 6.83e-10 ***
## genreRomance      -1.032460   0.857481  -1.204  0.22857
```

```
## genreShort         1.985399   0.509786   3.895 9.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.5  on 1833  degrees of freedom
## Residual deviance:  945.9  on 1823  degrees of freedom
## AIC: 967.9
##
## Number of Fisher Scoring iterations: 8
```

```
probit.step.backward = step(model11,direction="backward")
```

```
## Start:  AIC=967.9
## over7 ~ year + length + budget + log(votes) + genre
##
##               Df Deviance     AIC
## - log(votes)  1   945.91  965.91
## <none>            945.90  967.90
## - year        1   951.91  971.91
## - budget      1  1225.95 1245.95
## - length      1  1254.18 1274.18
## - genre       6  1582.45 1592.45
##
## Step:  AIC=965.91
## over7 ~ year + length + budget + genre
##
##           Df Deviance     AIC
## <none>         945.91  965.91
## - year     1   952.32  970.32
## - budget   1  1225.97 1243.97
## - length   1  1280.96 1298.96
## - genre    6  1584.57 1592.57
```

```
summary(probit.step.backward)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + genre, family = binomial(link = "probit"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8903  -0.3582  -0.0723   0.1357   4.5314
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -11.903295   4.086943  -2.913  0.00359 **
## year               0.005151   0.002076   2.481  0.01310 *
## length            -0.034440   0.002428 -14.187  < 2e-16 ***
## budget             0.280085   0.019435  14.411  < 2e-16 ***
## genreAnimation    -0.141775   0.233301  -0.608  0.54339
```

```
## genreComedy        1.697973   0.116717  14.548  < 2e-16 ***
## genreDocumentary   2.920990   0.263190  11.098  < 2e-16 ***
## genreDrama        -0.938714   0.152085  -6.172 6.73e-10 ***
## genreRomance      -1.028786   0.856039  -1.202  0.22944
## genreShort         1.984572   0.509619   3.894 9.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  945.91  on 1824  degrees of freedom
## AIC: 965.91
##
## Number of Fisher Scoring iterations: 8
```

```
probit.stepwise = step(model11,direction="both")
```

```
## Start:  AIC=967.9
## over7 ~ year + length + budget + log(votes) + genre
##
##               Df Deviance     AIC
## - log(votes)  1   945.91  965.91
## <none>             945.90  967.90
## - year        1   951.91  971.91
## - budget      1  1225.95 1245.95
## - length      1  1254.18 1274.18
## - genre       6  1582.45 1592.45
##
## Step:  AIC=965.91
## over7 ~ year + length + budget + genre
##
##               Df Deviance     AIC
## <none>             945.91  965.91
## + log(votes)  1   945.90  967.90
## - year        1   952.32  970.32
## - budget      1  1225.97 1243.97
## - length      1  1280.96 1298.96
## - genre       6  1584.57 1592.57
```

```
summary(probit.stepwise)
```

```
##
## Call:
## glm(formula = over7 ~ year + length + budget + genre, family = binomial(link = "probit"),
##     data = films)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8903  -0.3582  -0.0723   0.1357   4.5314
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)      -11.903295   4.086943  -2.913  0.00359 **
## year               0.005151   0.002076   2.481  0.01310 *
## length            -0.034440   0.002428 -14.187  < 2e-16 ***
## budget             0.280085   0.019435  14.411  < 2e-16 ***
## genreAnimation    -0.141775   0.233301  -0.608  0.54339
## genreComedy        1.697973   0.116717  14.548  < 2e-16 ***
## genreDocumentary   2.920990   0.263190  11.098  < 2e-16 ***
## genreDrama        -0.938714   0.152085  -6.172 6.73e-10 ***
## genreRomance      -1.028786   0.856039  -1.202  0.22944
## genreShort         1.984572   0.509619   3.894 9.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2331.49  on 1833  degrees of freedom
## Residual deviance:  945.91  on 1824  degrees of freedom
## AIC: 965.91
##
## Number of Fisher Scoring iterations: 8
```

```r
# compare AIC and BIC of full models using each type of logit, cloglog and probit
AIC(model6,model10,model11)
```

```
##         df       AIC
## model6  11  958.8695
## model10 11 1004.9769
## model11 11  967.8979
```

```r
BIC(model6,model10,model11)
```

```
##         df      BIC
## model6  11 1019.526
## model10 11 1065.634
## model11 11 1028.555
```

We yield very similar results to our binomial regression model, to see whether these models are better suited for our data we can use the AIC values. In each of the stepwise regression methods to optimal model was found to have budget, genre, length and year as the explanatory variables. Our original model has the lowest AIC and BIC values so we will keep using the binomial regression method.

Conclusion

In conclusion, we have investigated which properties influence whether a film receives a rating greater than 7 on the IMDB database. Out of all combinations of the 5 explanatory variables we have found that the best model for predicting whether a film will receives a rating greater than 7 includes the length of the film, the budget of the film, the genre of the film and the year the film was released. We settled on this model by iterating through all the possible combinations of models and choosing the model with the lowest AIC and BIC. We also used stepwise regression to corroborate this choice.

In the optimal model year and length of film have a significant influence on the probability that a film will receive a rating greater than 7 but the relative influence of these variables is small. The log odds of a film being rated over 7 will increase by 0.01 for every unit increase in the year of release of the film. Similarly, the log odds of a film being rated over 7 will decrease by 0.07 for every minute increase in the film length. For every \$1000000 increase of a films budget the log odds that the film will receive a score larger than 7 will increase by 0.52. The biggest influence on the log odds of the film receiving an excellent score is the film genre. In the optimal AIC model we found that only two of the genres are insignificant when predicting if the score of a film will be greater than 7 and these are animation and romance. The categories comedy, documentary and short film all have a positive influence on the log odds of getting a rating larger than 7 with increases of 3.16, 5.27 and 3.48 respectively. If the genre of the film is drama then the log odds of it receiving a score greater than 7 is reduced by 1.68. Therefore in this optimal AIC model we can say that the genre of the film has the largest influence on whether a film will receive a score greater than 7 followed by the budget of the film. Although they are significant in the model the length of the film and the year of the film have a less significant impact on the outcome of the films rating. It is somewhat surprising the the number of positive votes that a film receives from viewers does not have a significant relationship with a film receiving a rating over 7.

Alongside finding the optimal model we also investigated other combinations of explanatory variables that could be used to model film rating. We found that in models with a single explanatory variable that numerical explanatory variable that the length of a film and the budget of a film each had significant impact on the log odds that a film receives a rating over 7. An increase of \$1000000 increases the log odds by 0.18 and a minute increase in the length of a film decreases the log odds by 0.04. When modelling the rating and the film categories we can see that the category on the film has a large influence of the log odds of being greater than 7. The log odds difference in this model for the genres animation, comedy, documentary, drama, romance and short are 2.77, 2.07, 3.91, -1.14, -1.13 and 6 respectively. This is a big range of values, especially compared to the other factors, which means that depending on the category of film the probability of getting an excellent score is very different.

We also looked at altering the dataset by removing the insignificant variables from the optimal AIC model, the animation and romance genres, and fitting the optimal model again. We discovered that the renaming variables stayed significant at the highest level and that their coefficient values stayed very similar at the same time.

We looked at fitting a model which only includes length budget and genre as we found that this model had AIC and BIC values that were very similar to the optimal model but this model has the benefit of being simpler. The coefficients calculated were very similar to the optimal model and the explanatory variables had the same levels of significance. Therefore, unless the data for the year of release was unavailable, there are no significant benefits to using the model with these three variables and we can keep the same optimal model.

Other models we have investigated includes the full model which involves every single explanatory variable. The full model has similar AIC and BIC values to the optimal model we found but the inclusion of the log of the amount of positive votes a film receives is detrimental to the model. Furthermore we assessed whether using probit or a c log-log with the full model would improve model performance. We found that although the results were very similar the original logit regression had smaller AIC values.