group project 2

Group 05

2022/3/15

Contents


```
film <- read.csv("dataset5.csv")
```

Introduction of dataset

question of dataset

Imagine you have been asked by a film producer to investigate the following question of interest:

- Which properties of films influence whether they are rated by IMDB as greater than 7 or not?

You should conduct an analysis to answer your question using a Generalised Linear Model (GLM). Following your analyses, you should then summarise your results in the form of a presentation.

explain each variables

- film.id – The unique identifier for the film
- year – Year of release of the film in cinemas
- length – Duration (in minutes)
- budget – Budget for the films production (in $1000000s)
- votes – Number of positive votes received by viewers
- genre – Genre of the film
- rating – IMDB rating from 0-10

data processing

create a col seperate the rate: >7(1), <=7(0)

```
film <- film %>%
  mutate(rating.large7 = cut(rating, breaks = c(0,7,Inf), labels=c(0,1))) %>%
  dplyr::select(-film_id, -rating)%>%
  na.omit()
```

Exploratory data analysis

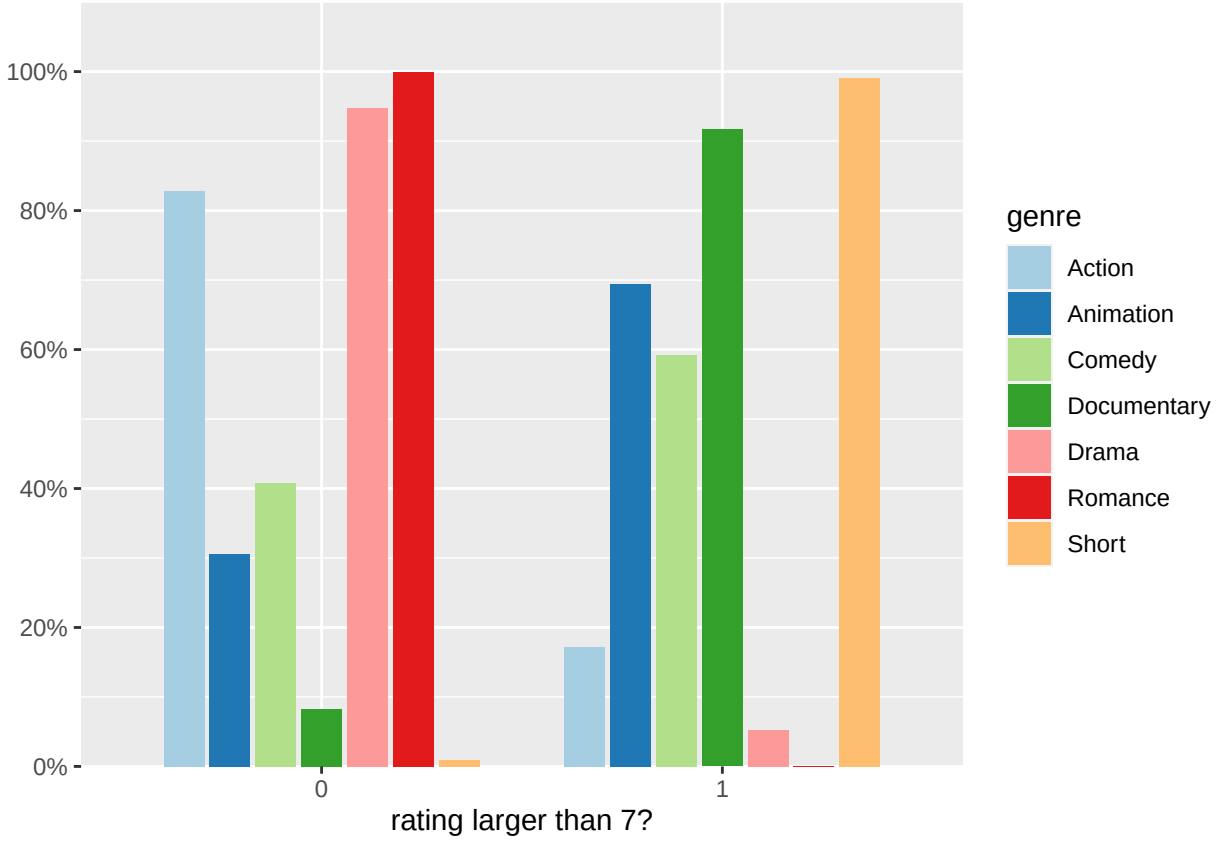
the distribution of rating.large7 by genre

```
film %>%
  group_by(genre, rating.large7)%>%
  summarise(n = n())
```

```
## `summarise()` has grouped output by 'genre'. You can override using the `.groups` argument.
```

```
## # A tibble: 13 x 3
## # Groups:   genre [7]
##    genre       rating.large7     n
##    <chr>       <fct>         <int>
##  1 Action      0               563
##  2 Action      1               117
##  3 Animation   0                49
##  4 Animation   1               111
##  5 Comedy      0               224
##  6 Comedy      1               325
##  7 Documentary 0                11
##  8 Documentary 1               121
##  9 Drama       0               620
## 10 Drama       1                34
## 11 Romance     0                15
## 12 Short       0                 1
## 13 Short       1               104
```

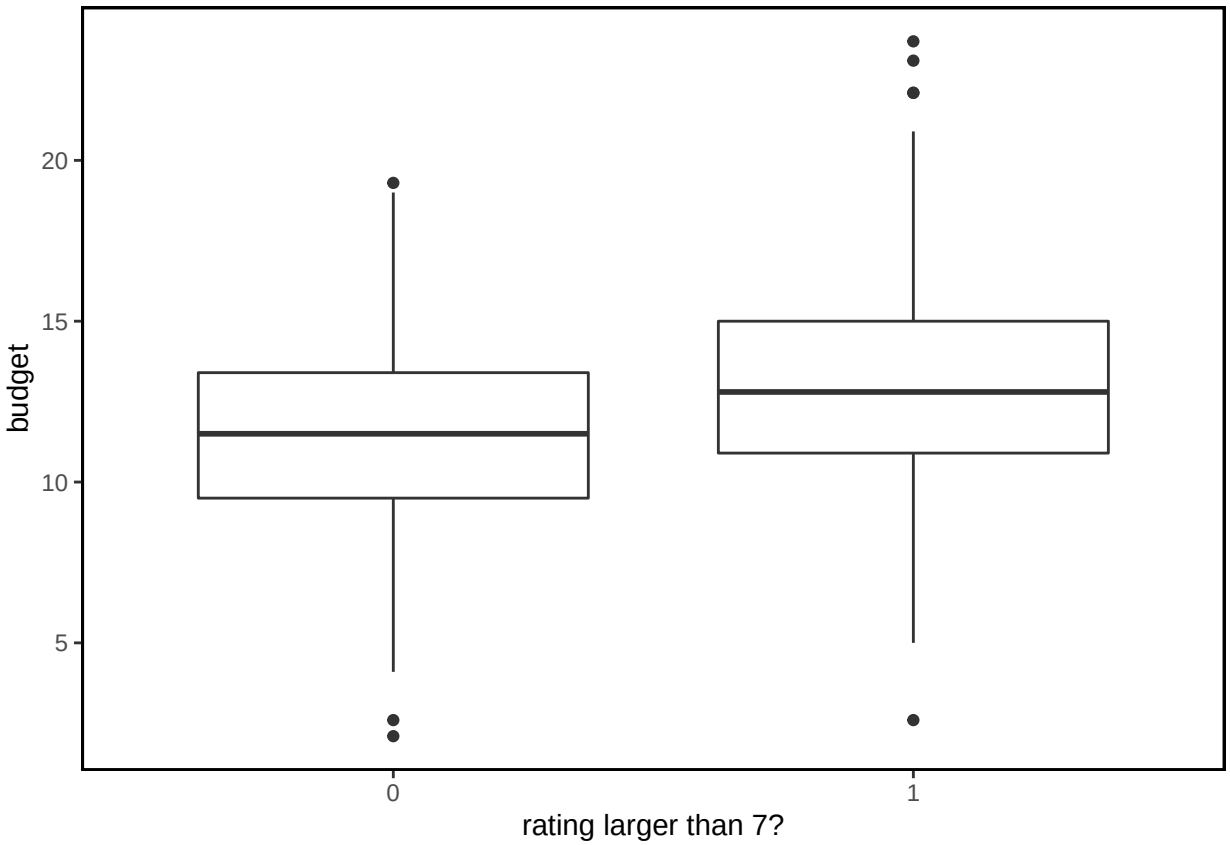
```
plot_xtab(film$rating.large7,film$genre,show.values =FALSE,show.total =FALSE,
axis.labels =c("0","1"),
axis.titles=c("rating larger than 7?"))
```



the distribution of rating.large7 by other numerical variables

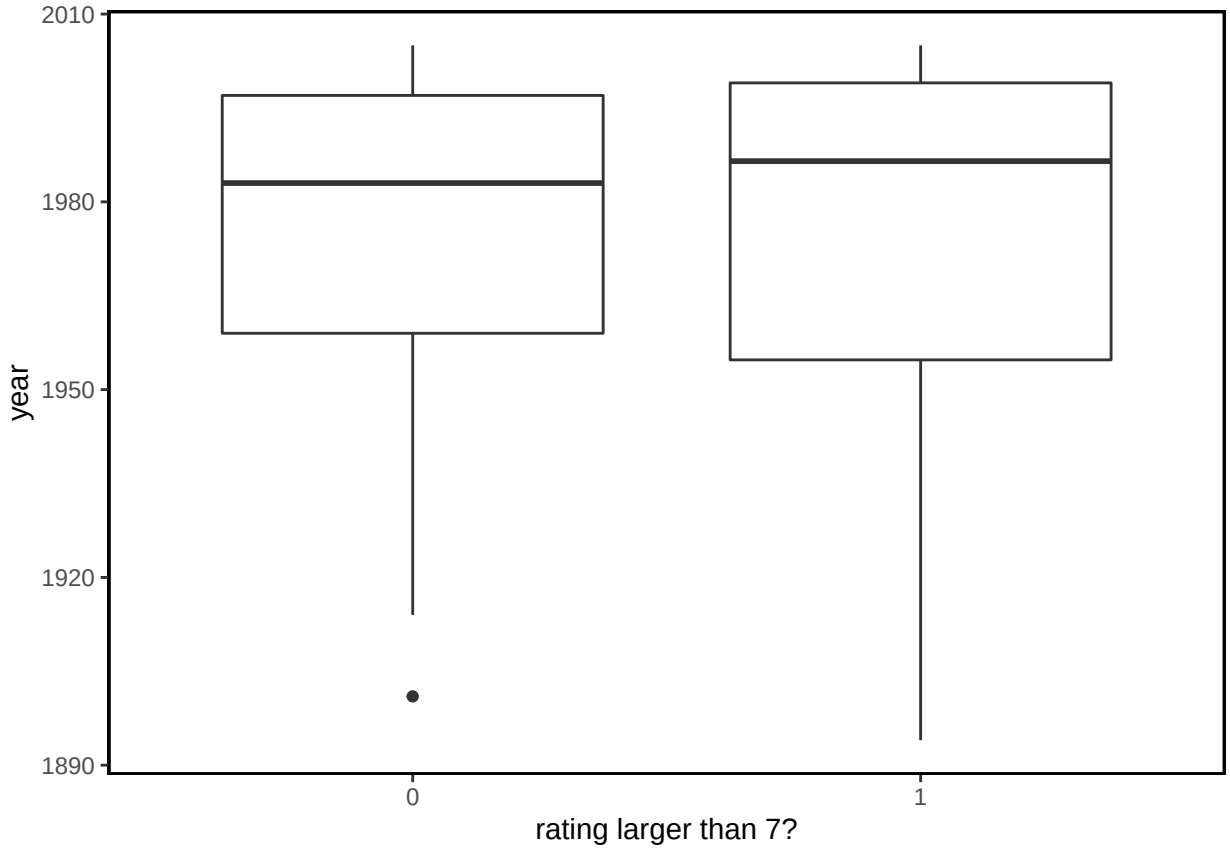
```
## budget
film.plot1<-ggplot(film, aes(y=budget,x=rating.large7))

film.plot1+geom_boxplot()+xlab("rating larger than 7?")+
theme(panel.background =element_rect(fill ="transparent",colour =NA),
plot.background =element_rect(fill ="transparent",colour =NA),
panel.border =element_rect(fill =NA,colour ="black",size =1))
```



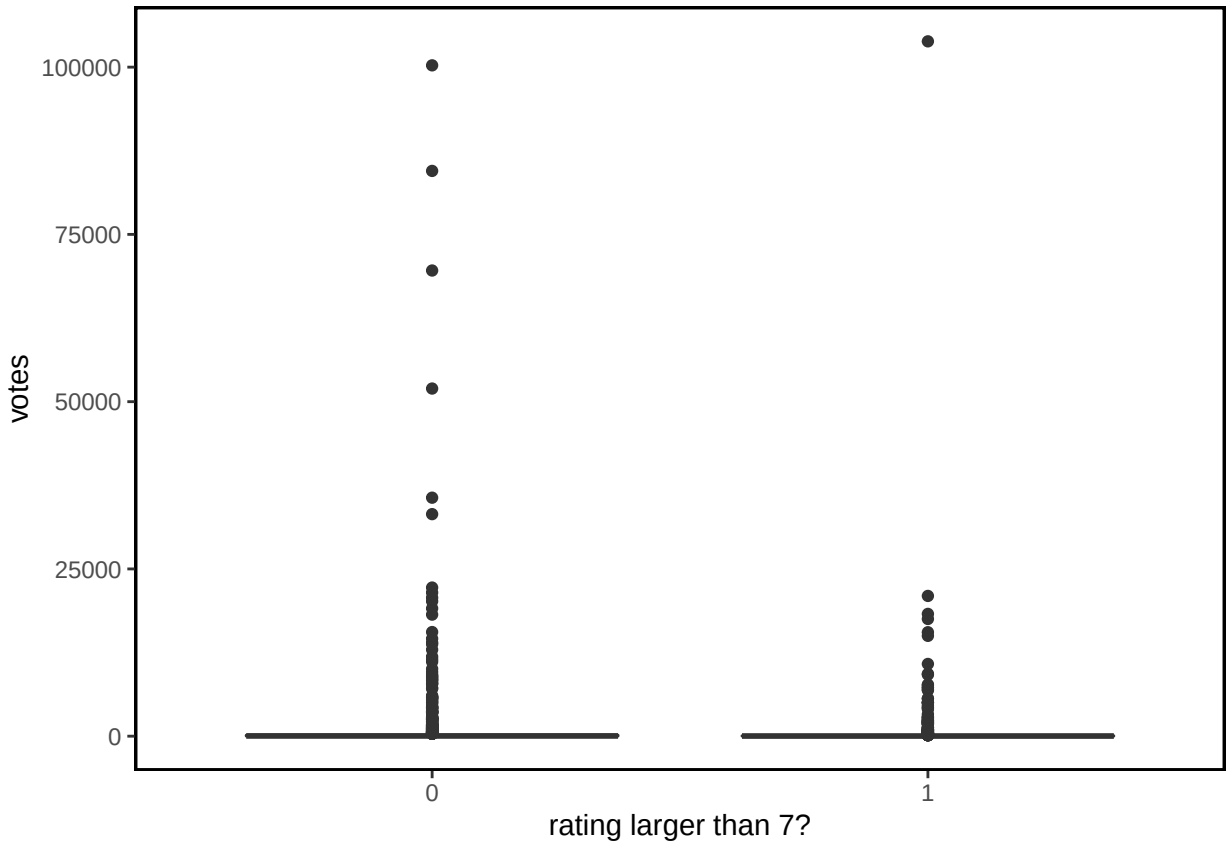

```
## year
film.plot2<-ggplot(film, aes(y=year,x=rating.large7))

film.plot2+geom_boxplot()+xlab("rating larger than 7?")+
theme(panel.background =element_rect(fill ="transparent",colour =NA),
plot.background =element_rect(fill ="transparent",colour =NA),
panel.border =element_rect(fill =NA,colour ="black",size =1))
```



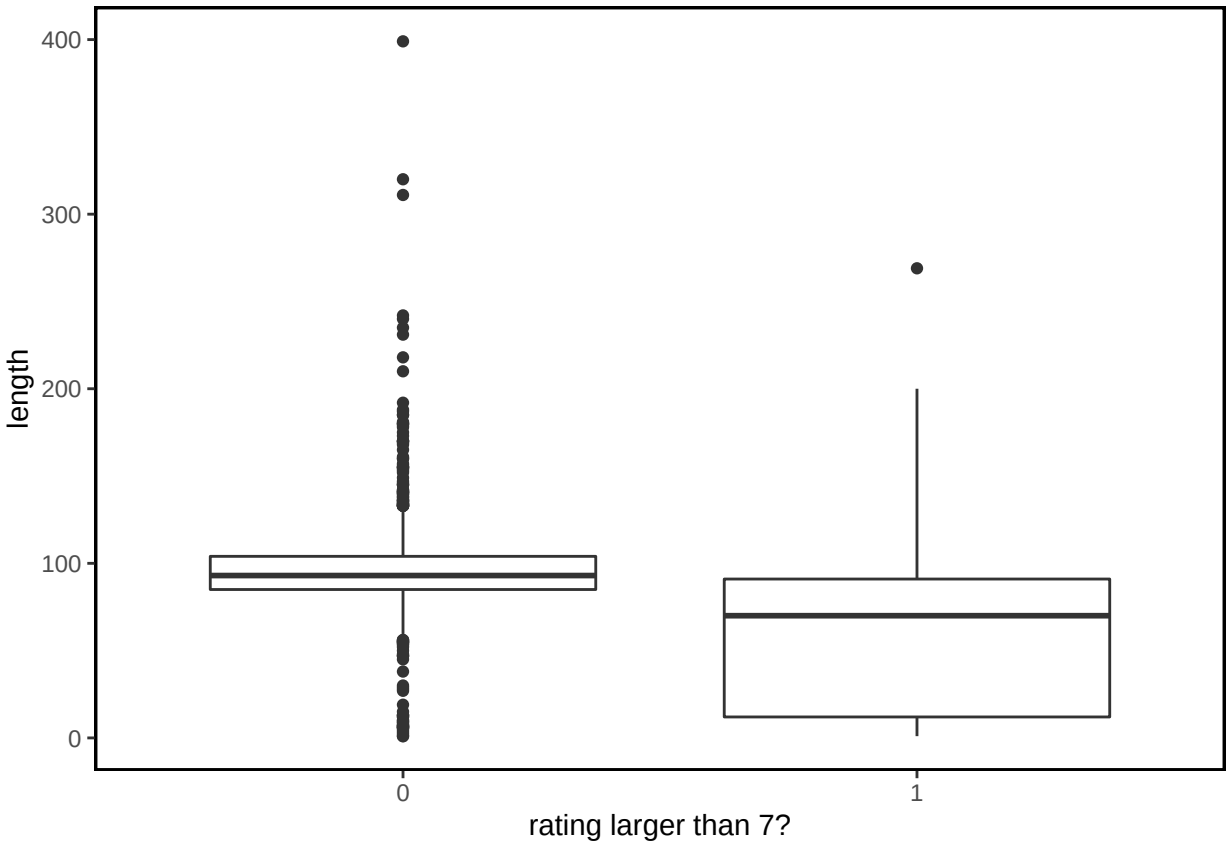

```
## votes
film.plot3<-ggplot(film, aes(y=votes,x=rating.large7))

film.plot3+geom_boxplot()+xlab("rating larger than 7?")+
theme(panel.background =element_rect(fill ="transparent",colour =NA),
plot.background =element_rect(fill ="transparent",colour =NA),
panel.border =element_rect(fill =NA,colour ="black",size =1))
```




```
## length
film.plot4<-ggplot(film, aes(y=length,x=rating.large7))

film.plot4+geom_boxplot()+xlab("rating larger than 7?")+
theme(panel.background =element_rect(fill ="transparent",colour =NA),
plot.background =element_rect(fill ="transparent",colour =NA),
panel.border =element_rect(fill =NA,colour ="black",size =1))
```



formal analysis

stepwise: choosing which variables need to be removed.

Since Model2(removed year) has lowest AIC=1310.1 and deviance D=1290.1 we will go ahead and compared the three link function in our model

```
model <-  glm(rating.large7 ~ year + length + budget + votes + genre, family = binomial(link = "logit")

logit.step <- step(model,direction='both')
```

```
## Start:  AIC=1311.03
## rating.large7 ~ year + length + budget + votes + genre
##
##          Df Deviance    AIC
## - year    1   1290.1 1310.1
## <none>        1289.0 1311.0
## - votes   1   1293.0 1313.0
## - length  1   1599.9 1619.9
## - budget  1   1663.1 1683.1
## - genre   6   2110.1 2120.1
##
## Step:  AIC=1310.07
```

```
## rating.large7 ~ length + budget + votes + genre
##
##          Df Deviance    AIC
## <none>        1290.1 1310.1
## + year    1   1289.0 1311.0
## - votes   1   1294.3 1312.3
## - length  1   1609.3 1627.3
## - budget  1   1664.8 1682.8
## - genre   6   2130.3 2138.3
```

```
summary(logit.step)
```

```
##
## Call:
## glm(formula = rating.large7 ~ length + budget + votes + genre,
##     family = binomial(link = "logit"), data = film)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7603  -0.4107  -0.1148   0.2614   4.1785
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.618e+00  4.428e-01  -8.170 3.09e-16 ***
## length           -5.129e-02  3.566e-03 -14.382  < 2e-16 ***
## budget            4.962e-01  3.147e-02  15.770  < 2e-16 ***
## votes             3.609e-05  1.672e-05   2.159   0.0309 *
## genreAnimation   -1.974e-01  3.378e-01  -0.584   0.5590
## genreComedy       2.749e+00  1.835e-01  14.983  < 2e-16 ***
## genreDocumentary  4.840e+00  4.098e-01  11.809  < 2e-16 ***
## genreDrama       -2.040e+00  2.580e-01  -7.907 2.64e-15 ***
## genreRomance     -1.361e+01  3.138e+02  -0.043   0.9654
## genreShort        4.209e+00  1.050e+00   4.008 6.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2982.5  on 2294  degrees of freedom
## Residual deviance: 1290.1  on 2285  degrees of freedom
## AIC: 1310.1
##
## Number of Fisher Scoring iterations: 14
```

compare different link functions

the AIC and BIC in model1 is the smallest, and the Pseudo-R^2 is the largest. Hence we choose 'logit' link function to fit our model

Model 1: logit link

```
model1 <- glm(rating.large7 ~ length + budget + votes + genre, family = binomial(link = "logit"), data =
summary(model1)
```

```
##
## Call:
## glm(formula = rating.large7 ~ length + budget + votes + genre,
##     family = binomial(link = "logit"), data = film)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7603  -0.4107  -0.1148   0.2614   4.1785
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.618e+00  4.428e-01  -8.170 3.09e-16 ***
## length           -5.129e-02  3.566e-03 -14.382  < 2e-16 ***
## budget            4.962e-01  3.147e-02  15.770  < 2e-16 ***
## votes             3.609e-05  1.672e-05   2.159   0.0309 *
## genreAnimation   -1.974e-01  3.378e-01  -0.584   0.5590
## genreComedy       2.749e+00  1.835e-01  14.983  < 2e-16 ***
## genreDocumentary  4.840e+00  4.098e-01  11.809  < 2e-16 ***
## genreDrama       -2.040e+00  2.580e-01  -7.907 2.64e-15 ***
## genreRomance     -1.361e+01  3.138e+02  -0.043   0.9654
## genreShort        4.209e+00  1.050e+00   4.008 6.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2982.5  on 2294  degrees of freedom
## Residual deviance: 1290.1  on 2285  degrees of freedom
## AIC: 1310.1
##
## Number of Fisher Scoring iterations: 14
```

```
summ(model1)
```

Observations	2295
Dependent variable	rating.large7
Type	Generalized linear model
Family	binomial
Link	logit

Model 2: probit link

```
model2 <- glm(rating.large7 ~ length + budget + votes + genre, family = binomial(link = "probit"), data
```

$\chi^2(9)$	1692.40
Pseudo-R^2 (Cragg-Uhler)	0.72
Pseudo-R^2 (McFadden)	0.57
AIC	1310.07
BIC	1367.45

	Est.	S.E.	z val.	p
(Intercept)	-3.62	0.44	-8.17	0.00
length	-0.05	0.00	-14.38	0.00
budget	0.50	0.03	15.77	0.00
votes	0.00	0.00	2.16	0.03
genreAnimation	-0.20	0.34	-0.58	0.56
genreComedy	2.75	0.18	14.98	0.00
genreDocumentary	4.84	0.41	11.81	0.00
genreDrama	-2.04	0.26	-7.91	0.00
genreRomance	-13.61	313.77	-0.04	0.97
genreShort	4.21	1.05	4.01	0.00

Standard errors: MLE

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model2)
```

```
##
## Call:
## glm(formula = rating.large7 ~ length + budget + votes + genre,
##     family = binomial(link = "probit"), data = film)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1883  -0.4440  -0.0870   0.2847   4.8941
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.083e+00  2.342e-01  -8.895  < 2e-16 ***
## length           -2.565e-02  1.792e-03 -14.312  < 2e-16 ***
## budget            2.637e-01  1.604e-02  16.442  < 2e-16 ***
## votes             1.937e-05  8.570e-06   2.261   0.0238 *
## genreAnimation    6.224e-02  1.818e-01   0.342   0.7321
## genreComedy       1.468e+00  9.656e-02  15.202  < 2e-16 ***
## genreDocumentary  2.639e+00  2.074e-01  12.720  < 2e-16 ***
## genreDrama       -1.117e+00  1.319e-01  -8.468  < 2e-16 ***
## genreRomance     -4.742e+00  7.702e+01  -0.062   0.9509
## genreShort        2.307e+00  4.497e-01   5.131 2.88e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2982.5  on 2294  degrees of freedom
## Residual deviance: 1314.7  on 2285  degrees of freedom
```

```
## AIC: 1334.7
##
## Number of Fisher Scoring iterations: 14
```

```
summ(model2)
```

Observations	2295
Dependent variable	rating.large7
Type	Generalized linear model
Family	binomial
Link	probit

$\chi^2(9)$	1667.81
Pseudo-R^2 (Cragg-Uhler)	0.71
Pseudo-R^2 (McFadden)	0.56
AIC	1334.66
BIC	1392.04

	Est.	S.E.	z val.	p
(Intercept)	-2.08	0.23	-8.89	0.00
length	-0.03	0.00	-14.31	0.00
budget	0.26	0.02	16.44	0.00
votes	0.00	0.00	2.26	0.02
genreAnimation	0.06	0.18	0.34	0.73
genreComedy	1.47	0.10	15.20	0.00
genreDocumentary	2.64	0.21	12.72	0.00
genreDrama	-1.12	0.13	-8.47	0.00
genreRomance	-4.74	77.02	-0.06	0.95
genreShort	2.31	0.45	5.13	0.00

Standard errors: MLE

Model 3: complementary log-log link

```
model3 <- glm(rating.large7 ~ length + budget + votes + genre, family = binomial(link = "cloglog"), data
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model3)
```

```
##
## Call:
## glm(formula = rating.large7 ~ length + budget + votes + genre,
##     family = binomial(link = "cloglog"), data = film)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.6733  -0.4994  -0.2382   0.2738   3.1342
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.927e+00  2.712e-01 -10.791  < 2e-16 ***
## length           -2.510e-02  1.919e-03 -13.080  < 2e-16 ***
## budget            2.725e-01  1.771e-02  15.386  < 2e-16 ***
## votes             1.852e-05  9.592e-06   1.931   0.0535 .
## genreAnimation    3.375e-01  1.981e-01   1.704   0.0883 .
## genreComedy       1.664e+00  1.177e-01  14.134  < 2e-16 ***
## genreDocumentary  2.805e+00  1.886e-01  14.873  < 2e-16 ***
## genreDrama       -1.403e+00  1.921e-01  -7.300 2.87e-13 ***
## genreRomance     -2.442e+01  8.045e+04   0.000   0.9998
## genreShort        2.367e+00  3.432e-01   6.898 5.27e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2982.5  on 2294  degrees of freedom
## Residual deviance: 1384.6  on 2285  degrees of freedom
## AIC: 1404.6
##
## Number of Fisher Scoring iterations: 25
```

```
summ(model3)
```

Observations	2295
Dependent variable	rating.large7
Type	Generalized linear model
Family	binomial
Link	cloglog

$\chi^2(9)$	1597.91
Pseudo-R^2 (Cragg-Uhler)	0.69
Pseudo-R^2 (McFadden)	0.54
AIC	1404.55
BIC	1461.94

$$D_0 - D_1 = 2982.5 - 2357.5 = 1692.401 > \chi^2(0.95, 9) = 16.91898$$

So we reject H0, and we can say that the model1 fits the data better than saturated model.

```
model1$null.deviance - model1$deviance
```

```
## [1] 1692.401
```

	Est.	S.E.	z val.	p
(Intercept)	-2.93	0.27	-10.79	0.00
length	-0.03	0.00	-13.08	0.00
budget	0.27	0.02	15.39	0.00
votes	0.00	0.00	1.93	0.05
genreAnimation	0.34	0.20	1.70	0.09
genreComedy	1.66	0.12	14.13	0.00
genreDocumentary	2.80	0.19	14.87	0.00
genreDrama	-1.40	0.19	-7.30	0.00
genreRomance	-24.42	80451.51	-0.00	1.00
genreShort	2.37	0.34	6.90	0.00

Standard errors: MLE

```
df = model1$df.null - model1$df.residual
qchisq(p=0.95, df = df)
```

```
## [1] 16.91898
```

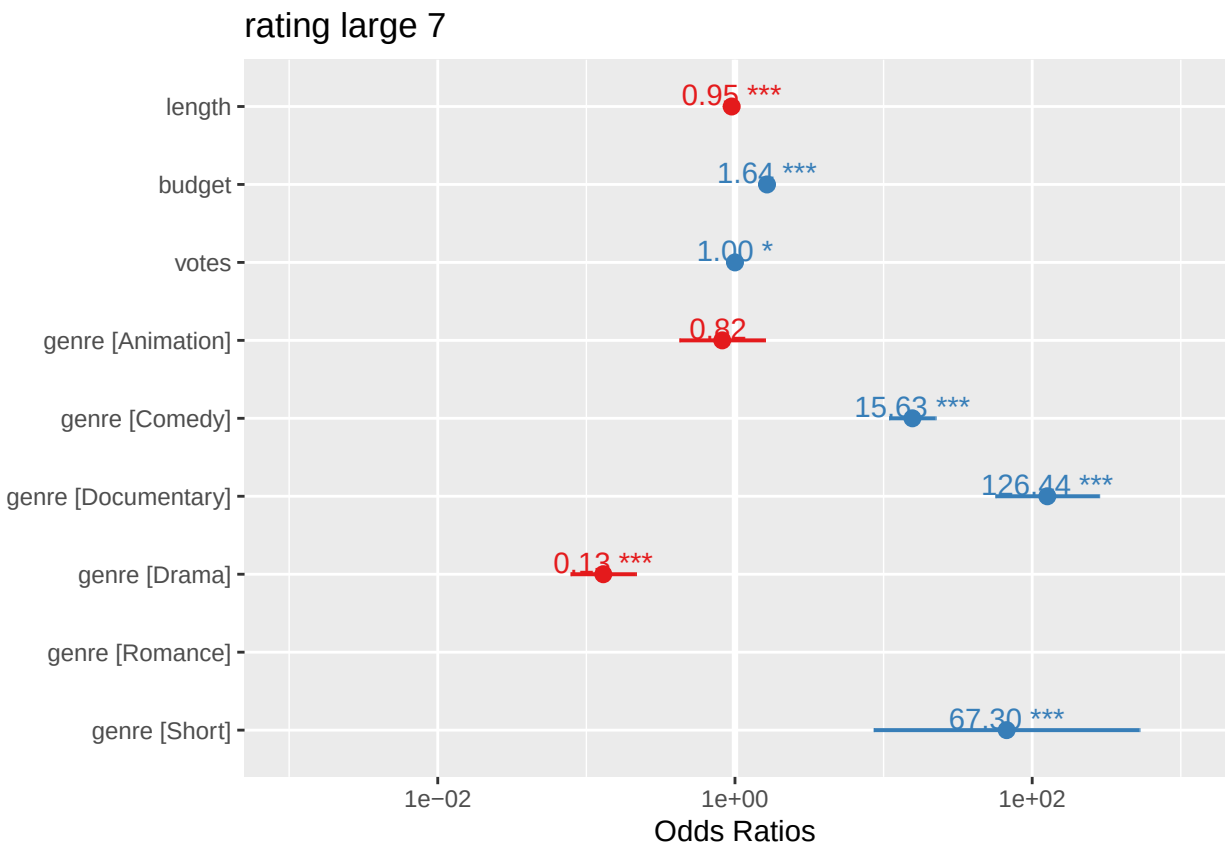
odds ratios of model1

```
plot_model(model1,show.values=TRUE)+
  scale_y_log10(limits = c(0.001, 1000))
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

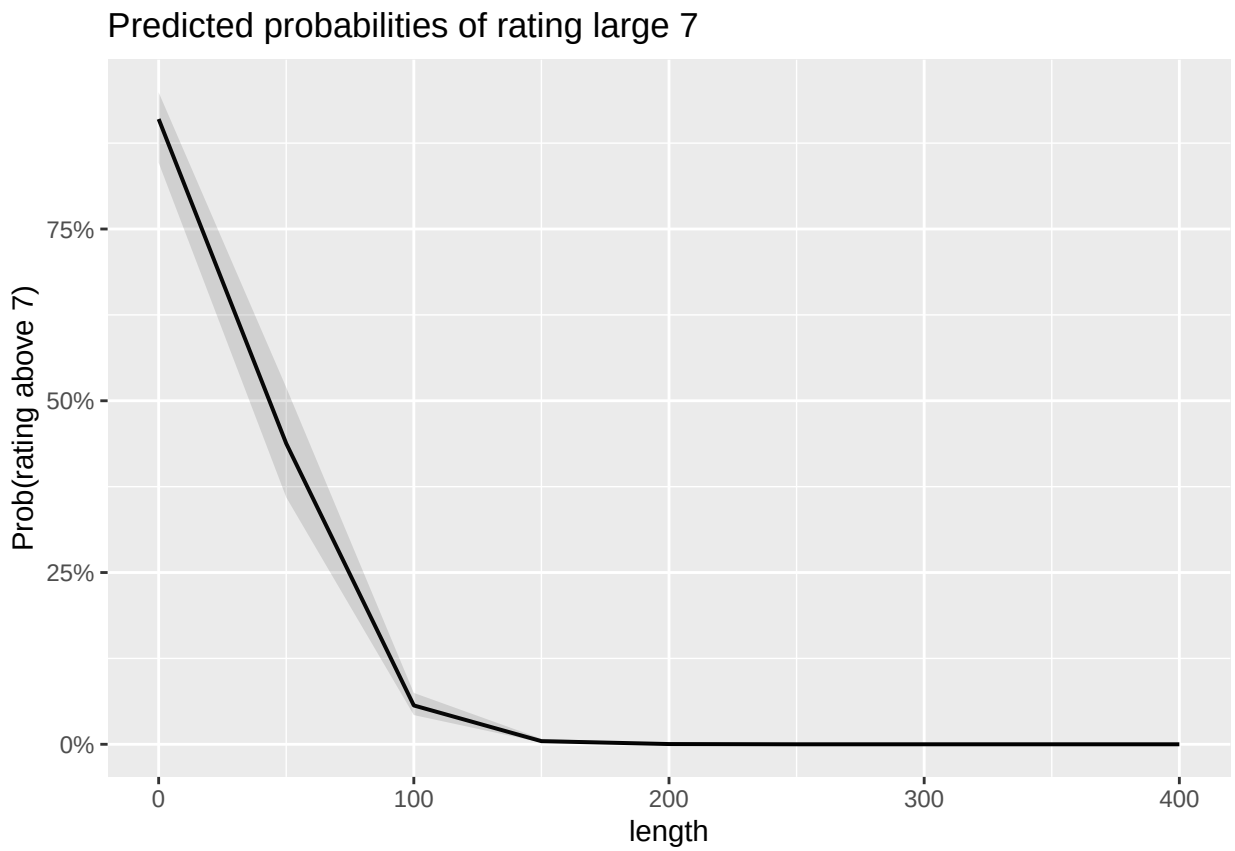
```
## Warning: Removed 1 rows containing missing values (geom_text).
```



prediction

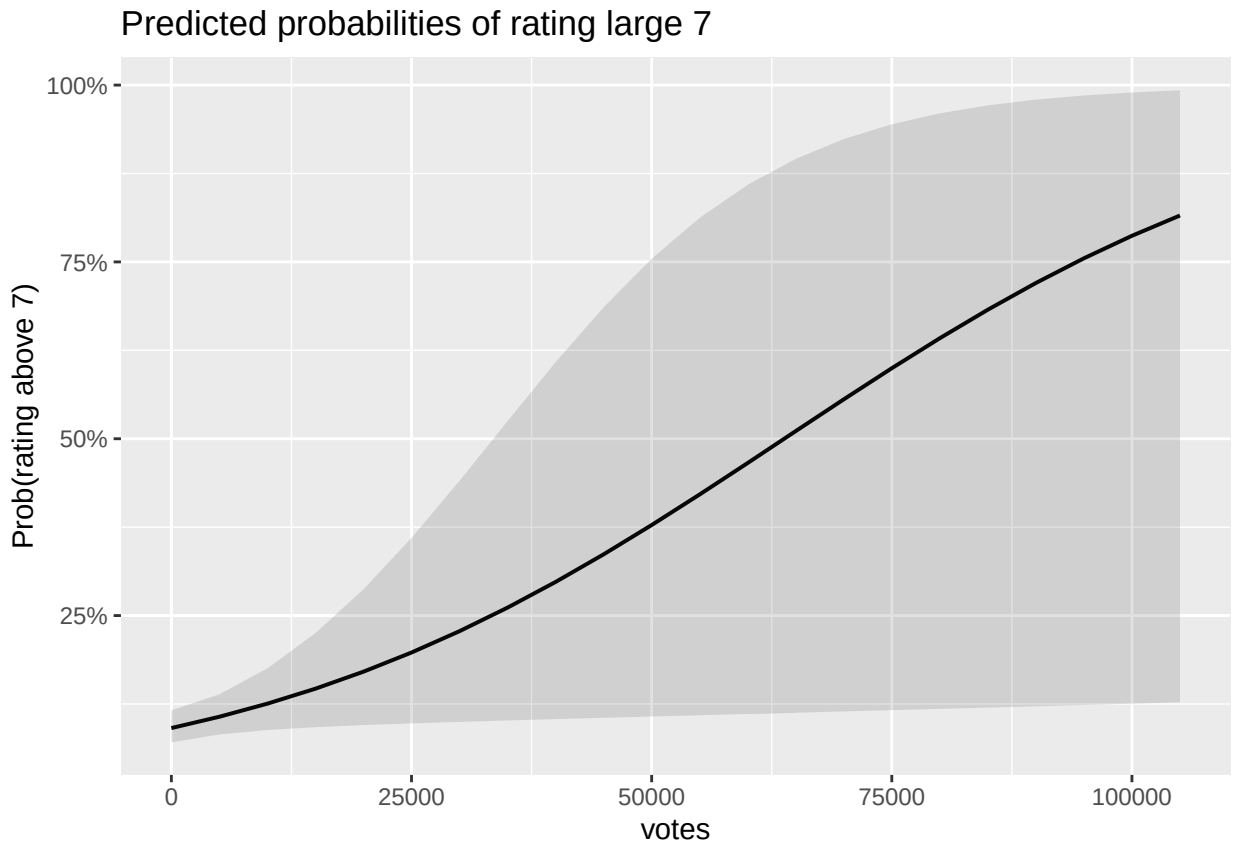
```
plot_model(model1,type="pred",terms=c("length"),axis.title=c("length","Prob(rating above 7)"))
```

```
## Data were 'prettified'. Consider using `terms="length [all]"` to get smooth plots.
```



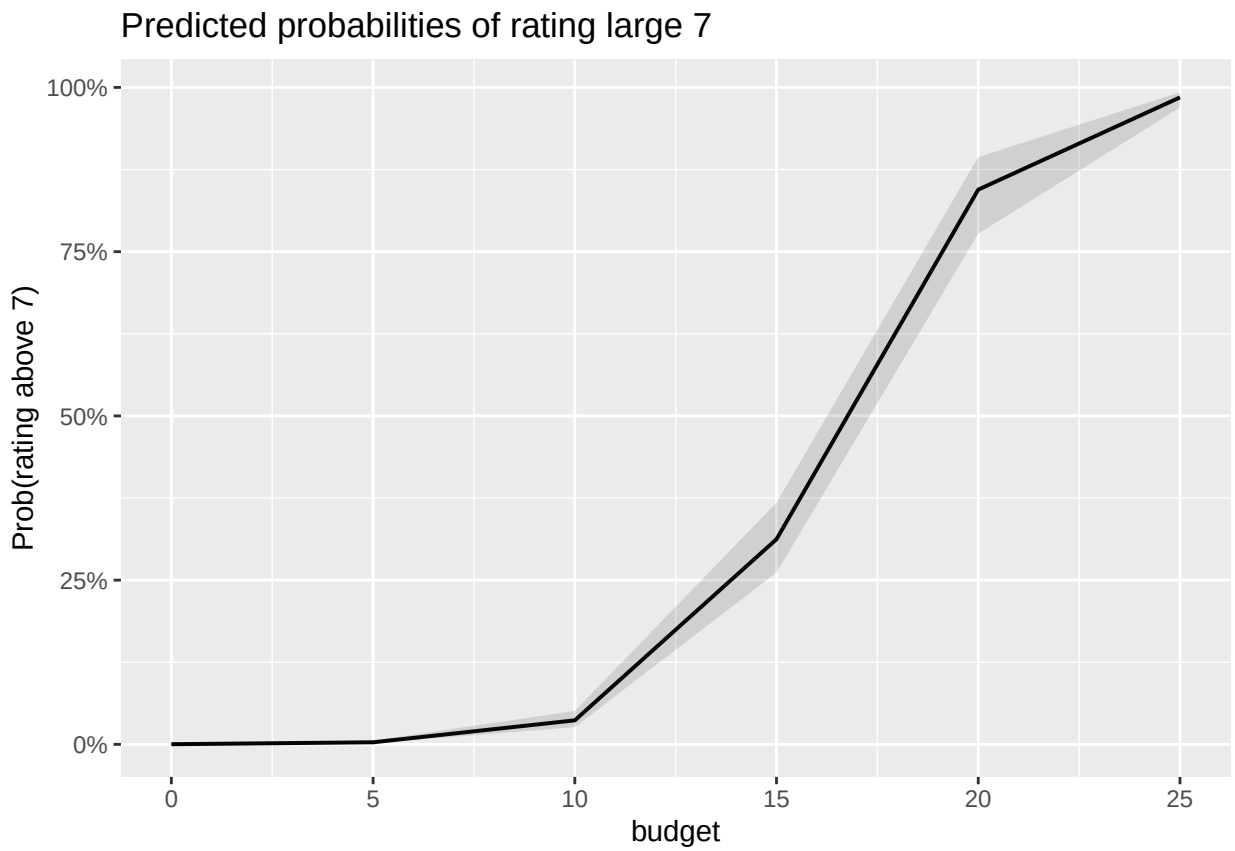

```
plot_model(model1,type="pred",terms=c("votes"),axis.title=c("votes","Prob(rating above 7)"))
```

```
## Data were 'prettified'. Consider using 'terms="votes [all]"' to get smooth plots.
```

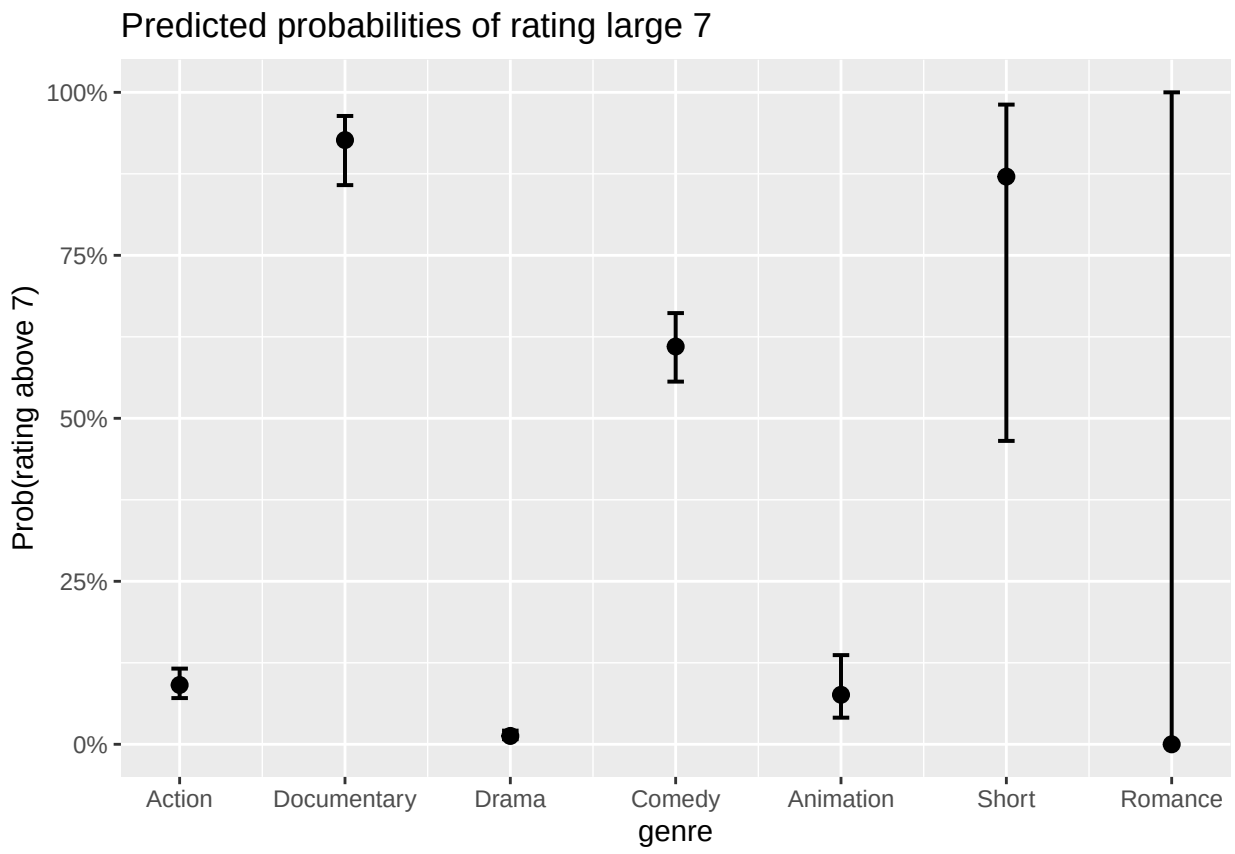



```
plot_model(model1,type="pred",terms=c("budget"),axis.title=c("budget","Prob(rating above 7)"))
```

```
## Data were 'prettified'. Consider using 'terms="budget [all]"' to get smooth plots.
```

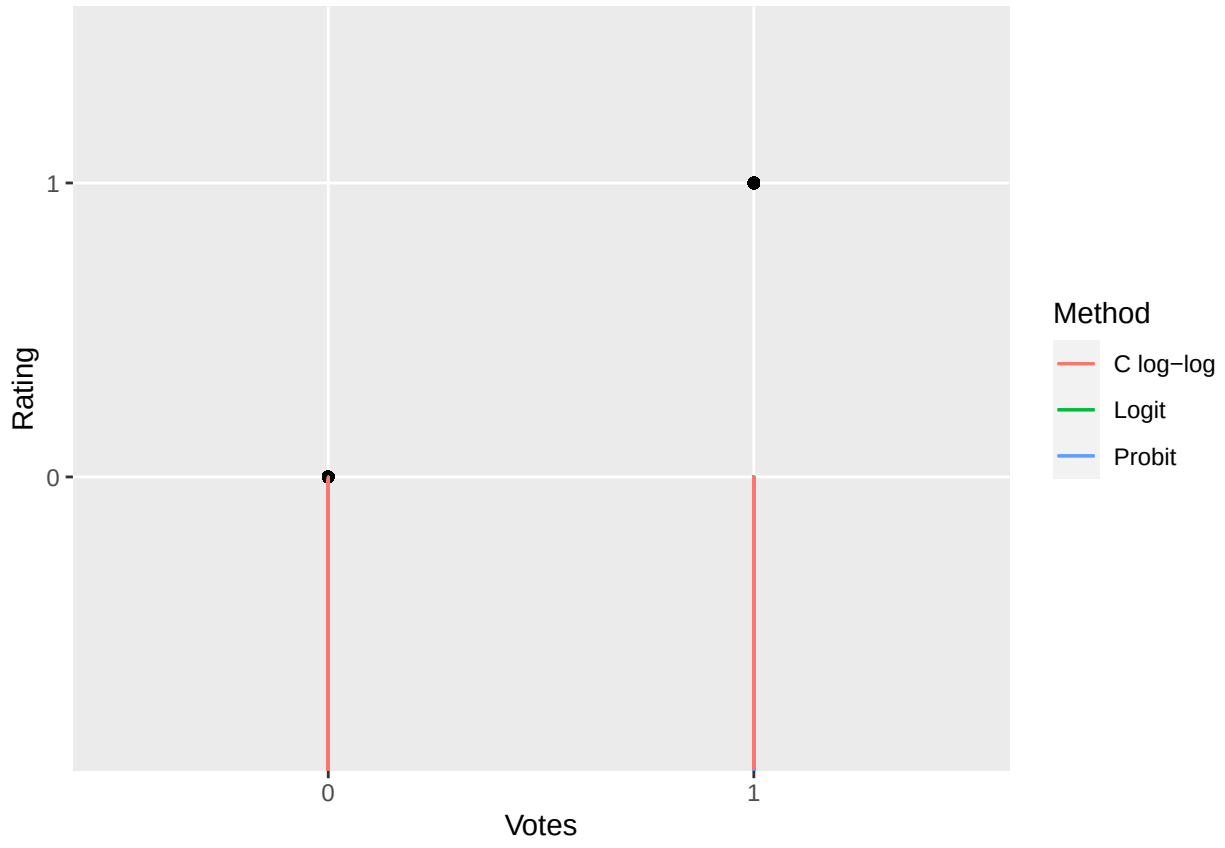



```
plot_model(model1,type="pred",terms=c("genre"),axis.title=c("genre","Prob(rating above 7)"))
```



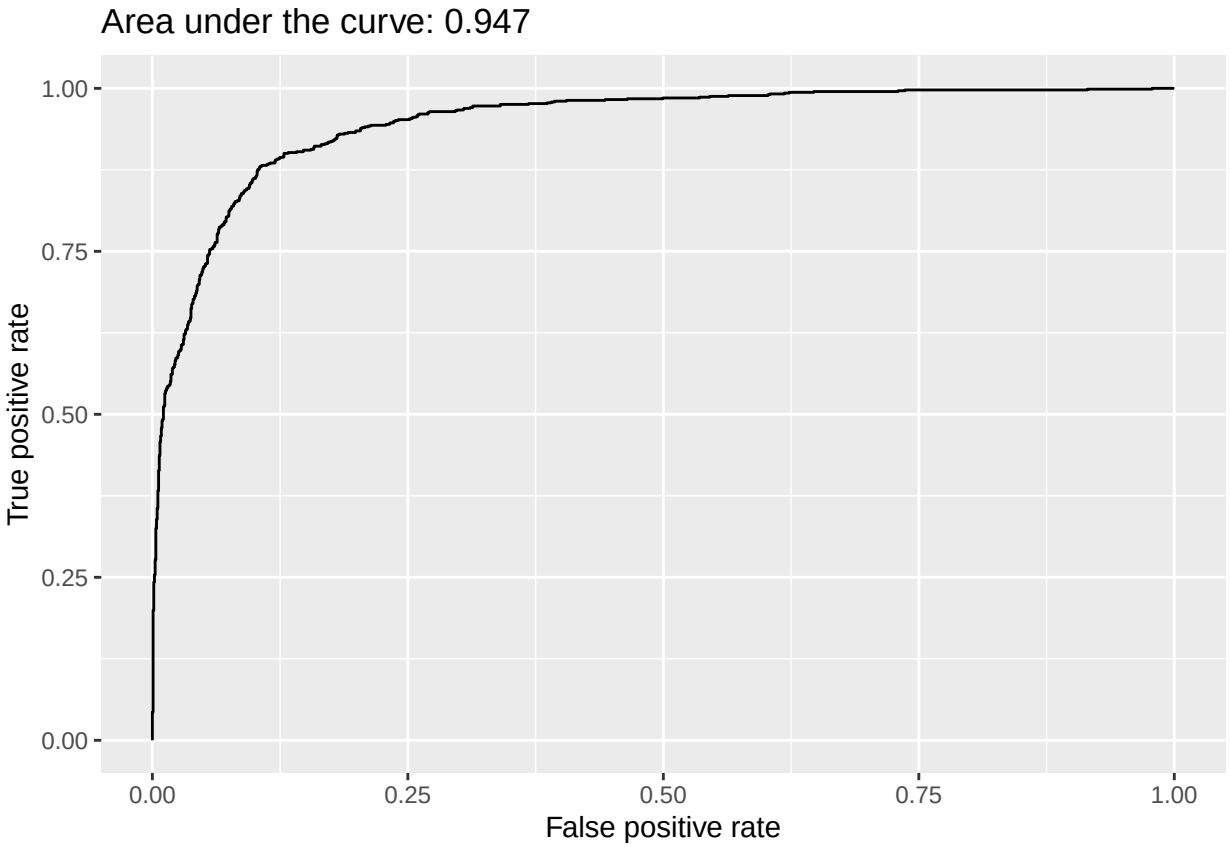
Model Comparison

```
film_p <- data.frame(film = film,
                     logit = fitted(model1),
                     probit = fitted(model2),
                     cloglog = fitted(model3))
ggplot(film_p, aes(y = film$rating.large7, x = film$rating.large7)) +
geom_point() + xlab("Votes") + ylab("Rating") +
geom_line(aes(x = film$rating, y = logit, colour = "Logit")) +
geom_line(aes(x = film$rating, y = probit, colour = "Probit")) +
geom_line(aes(x = film$rating, y = cloglog, colour = "C log-log")) +
guides(colour = guide_legend("Method"))
```



ROC

```
film$Prid <- predict(model1, film, type="response")
score <- prediction(film$Prid,film$rating.large7)
perf <- performance(score,"tpr","fpr")
auc <- performance(score,"auc")
perfd <- data.frame(x= perf@x.values[1][[1]], y=perf@y.values[1][[1]])
p4<- ggplot(perfd, aes(x= x, y=y)) + geom_line() +
xlab("False positive rate") + ylab("True positive rate") +
ggtitle(paste("Area under the curve:", round(auc@y.values[[1]], 3)))
p4
```



AUC = 0.947 indicated that model 1 is very good at predicting the films rating greater than 7 given all predictor variables.