432 Class 08

https://thomaselove.github.io/432-2024

2024-02-08

## Today’s Agenda

* The Bechdel-Wallace Test and the Favorite Movies Data
* Three Logistic Regressions with glm() & lrm()
  + Using tidy, glance and augment from broom
  + Making Predictions with our models
  + Interpreting exponentiated coefficients as odds ratios
  + Likelihood Ratio and other ANOVA Tests
  + ROC curve and the Area under the Curve
  + Validating Model Summaries

## Today’s R Setup

knitr::opts\_chunk$set(comment = NA)  
  
library(readxl) # import from Excel Sheet  
library(skimr) # can help with exploration/cleaning  
library(broom)  
library(janitor)  
library(gt)  
library(naniar)  
library(pROC) # helps us plot ROC curves  
library(rms) # also loads Hmisc  
library(tidyverse)  
  
theme\_set(theme\_bw())

# “Favorite Movies” Data

## Ingest Data from an Excel Sheet

mov23\_full <- read\_xlsx("c08/data/movies\_2023-10-24.xlsx")  
  
dim(mov23\_full)

[1] 201 90

### Select Today’s Nine Variables

mov23 <- mov23\_full |>  
 select(film\_id, bw\_rating, year, mpa, metascore,   
 gross\_world, comedy, drama, film) |>  
 type.convert(as.is = FALSE) |> # makes all character variables factors  
 mutate(film\_id = as.character(film\_id),  
 film = as.character(film))  
  
dim(mov23)

[1] 201 9

## The Bechdel Test

The Bechdel Test, or Bechdel-Wallace Test was popularized by Alison Bechdel’s comic, in a 1985 strip called [The Rule](https://dykestowatchoutfor.com/the-rule/).

* from <https://bechdeltest.com/>

The Bechdel-Wallace Test is a simple way to gauge the active presence of female characters in Hollywood films and just how well rounded and complete those roles are[[1]](#footnote-26).

## Passing the Bechdel-Wallace Test

To pass the test, a movie must have all three of the following.

* at least two (named) women
* who talk to each other
* about something besides a man

mov23 <- mov23 |>  
 mutate(bechdel = factor(ifelse(bw\_rating == 3, "Pass", "Fail")))  
mov23 |> count(bechdel, bw\_rating)

# A tibble: 5 × 3  
 bechdel bw\_rating n  
 <fct> <dbl> <int>  
1 Fail 0 16  
2 Fail 1 54  
3 Fail 2 11  
4 Pass 3 112  
5 <NA> NA 8

## Some Data Cleanup

1. Drop the films missing the bechdel information.
2. Create an age variable and use it instead of year, and
3. Rescale world-wide gross by dividing by $1,000,000.

mov23 <- mov23 |>  
 filter(complete.cases(bechdel)) |>  
 mutate(age = 2024-year,  
 gross = gross\_world/1000000)  
  
mov23 |> tabyl(bechdel) |> adorn\_pct\_formatting()

bechdel n percent  
 Fail 81 42.0%  
 Pass 112 58.0%

## More Data Cleanup

How about the MPA ratings?

summary(mov23$mpa)

G NR PG PG-13 R TV-G TV-MA TV-PG   
 6 6 50 68 61 1 0 1

Let’s collapse to the two largest categories, plus “Other”

mov23 <- mov23 |> mutate(mpa3 = fct\_lump\_n(mpa, n = 2))  
mov23 |> tabyl(mpa3) |> adorn\_pct\_formatting() |>   
 gt() |> tab\_options(table.font.size = 24)

| mpa3 | n | percent |
| --- | --- | --- |
| PG-13 | 68 | 35.2% |
| R | 61 | 31.6% |
| Other | 64 | 33.2% |

## Any Missing Data?

## select the variables we're actually going to use  
  
mov23a <- mov23 |>  
 select(film\_id, film, bechdel, age, gross, metascore, mpa3, comedy, drama)  
  
miss\_var\_summary(mov23a)

# A tibble: 9 × 3  
 variable n\_miss pct\_miss  
 <chr> <int> <dbl>  
1 metascore 4 2.07  
2 gross 3 1.55  
3 film\_id 0 0   
4 film 0 0   
5 bechdel 0 0   
6 age 0 0   
7 mpa3 0 0   
8 comedy 0 0   
9 drama 0 0

## Which movies are missing data?

mov23a |> filter(!complete.cases(metascore, gross)) |>  
 select(film\_id, metascore, gross, film)

# A tibble: 6 × 4  
 film\_id metascore gross film   
 <chr> <dbl> <dbl> <chr>   
1 M043 NA 0.0750 Dilwale Dulhania Le Jayenge (The Brave Heart Will T…  
2 M048 50 NA Eurovision Song Contest: The Story of Fire Saga   
3 M078 NA NA High School Musical 2   
4 M132 38 NA Murder Mystery   
5 M141 NA 0.135 Pather Panchali   
6 M200 NA 31.0 Yeh Jawaani hai Deewani

## Today, we use complete cases

For today, let’s just drop the films with missing data.

mov23a <- mov23a |> drop\_na()  
  
mov23a

# A tibble: 187 × 9  
 film\_id film bechdel age gross metascore mpa3 comedy drama  
 <chr> <chr> <fct> <dbl> <dbl> <dbl> <fct> <int> <int>  
 1 M001 3 Idiots Fail 15 6.03e+1 67 PG-13 1 1  
 2 M002 8 1/2 Pass 61 1.96e-1 93 Other 0 1  
 3 M003 10 Things I Hate … Pass 25 5.35e+1 70 PG-13 1 1  
 4 M004 2001: A Space Ody… Fail 56 6.64e+1 84 Other 0 0  
 5 M005 About Elly (Darba… Pass 15 8.79e-1 87 Other 0 1  
 6 M006 About Time Pass 11 8.71e+1 55 R 1 1  
 7 M007 Alien Pass 45 1.06e+2 89 R 0 0  
 8 M008 Amadeus Pass 40 5.21e+1 87 Other 0 1  
 9 M009 Avatar Pass 15 2.92e+3 83 PG-13 0 0  
10 M010 Avengers: Infinit… Pass 6 2.80e+3 78 PG-13 0 0  
# ℹ 177 more rows

## Codebook, part 1

| Variable | Description |
| --- | --- |
| film\_id | identifying code (M-001 through M-201) |
| film | title of film |
| bechdel | Bechdel Test Result (Pass or Fail) |
| age | 2024 - Year of release (1942-2023) |
| gross | Worldwide gross income in $millions |

Data Sources: <https://www.imdb.com/> and <https://bechdeltest.com>

## Codebook, part 2

| Variable | Description |
| --- | --- |
| metascore | Metacritic score (from critics: 0-100 scale) |
| mpa3 | MPA rating (now PG-13, R, Other) |
| comedy | Is Comedy one of the movie’s three genres listed at IMDB? (1 = Yes, 0 = No) |
| drama | Is Drama one of the movie’s three genres listed at IMDB? (1 = Yes, 0 = No) |

Data Sources: <https://www.imdb.com/> and <https://bechdeltest.com>

## Skim the mov23a data?

skim\_output <- skim(mov23a)  
summary(skim\_output)

Data summary

|  |  |
| --- | --- |
| Name | mov23a |
| Number of rows | 187 |
| Number of columns | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| factor | 2 |
| numeric | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

## Character Summary

skimr::yank(skim\_output, "character") |>   
 gt() |> tab\_options(table.font.size = 24)

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| film\_id | 0 | 1 | 4 | 4 | 0 | 187 | 0 |
| film | 0 | 1 | 3 | 54 | 0 | 187 | 0 |

### Factor Summary

yank(skim\_output, "factor") |>   
 gt() |> tab\_options(table.font.size = 24)

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| bechdel | 0 | 1 | FALSE | 2 | Pas: 107, Fai: 80 |
| mpa3 | 0 | 1 | FALSE | 3 | PG-: 66, R: 61, Oth: 60 |

## Numeric Summary

skimr::yank(skim\_output, "numeric")

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 0 | 1 | 21.14 | 14.23 | 1 | 11.00 | 18.00 | 28.0 | 82.00 | ▇▆▂▁▁ |
| gross | 0 | 1 | 339.56 | 456.03 | 0 | 55.06 | 180.56 | 469.6 | 2923.71 | ▇▂▁▁▁ |
| metascore | 0 | 1 | 71.61 | 15.67 | 9 | 61.00 | 72.00 | 84.0 | 100.00 | ▁▁▅▇▆ |
| comedy | 0 | 1 | 0.35 | 0.48 | 0 | 0.00 | 0.00 | 1.0 | 1.00 | ▇▁▁▁▅ |
| drama | 0 | 1 | 0.56 | 0.50 | 0 | 0.00 | 1.00 | 1.0 | 1.00 | ▆▁▁▁▇ |

## Splitting the sample?

We have 187 films in our mov23a tibble.

* It turns out that a logistic regression model needs about 96 observations just to fit a reasonable intercept term.
* Each additional coefficient we fit requires another 10-20 observations just so that we *might* validate well.

Here, we want to explore six predictors (age, mpa3, metascore, gross, comedy and drama.)

* Does it make sense to split our data into separate training and testing samples?

# Model 1. Using age to predict Pr(bechdel = Pass)

## The Logistic Regression Model

Here, our *event* will be “movie passes the bechdel-Wallace test” (bechdel = Pass)

## Model mod\_1

mod\_1 <- glm((bechdel == "Pass") ~ age,  
 data = mov23a, family = binomial(link = "logit"))  
mod\_1

Call: glm(formula = (bechdel == "Pass") ~ age, family = binomial(link = "logit"),   
 data = mov23a)  
  
Coefficients:  
(Intercept) age   
 0.77951 -0.02291   
  
Degrees of Freedom: 186 Total (i.e. Null); 185 Residual  
Null Deviance: 255.3   
Residual Deviance: 250.6 AIC: 254.6

## Tidied mod\_1 coefficients

tidy(mod\_1, conf.int = TRUE, conf.level = 0.90) |>   
 gt() |> fmt\_number(decimals = 3) |> tab\_options(table.font.size = 24)

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- | --- |
| (Intercept) | 0.780 | 0.275 | 2.839 | 0.005 | 0.335 | 1.240 |
| age | -0.023 | 0.011 | -2.127 | 0.033 | -0.041 | -0.006 |

Note that I haven’t done any exponentiating here.

## mod\_1 predicts a movie with age = 50

Estimated Probability of Passing Bechdel-Wallace Test: 41%.

## Three extra movies (not in mov23a)

new3\_a <- tibble(age = c(50, 50, 20),   
 film = c("Godfather II", "Chinatown", "Incredibles"))  
  
augment(mod\_1, newdata = new3\_a, type.predict = "response") |>   
 gt() |> tab\_options(table.font.size = 24)

| age | film | .fitted |
| --- | --- | --- |
| 50 | Godfather II | 0.4094498 |
| 50 | Chinatown | 0.4094498 |
| 20 | Incredibles | 0.5796185 |

## Exponentiating mod\_1 Coefficients

tidy(mod\_1, exponentiate = TRUE, conf.int = TRUE, conf.level = 0.90) |>   
 gt() |> fmt\_number(decimals = 3) |> tab\_options(table.font.size = 24)

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- | --- |
| (Intercept) | 2.180 | 0.275 | 2.839 | 0.005 | 1.397 | 3.456 |
| age | 0.977 | 0.011 | -2.127 | 0.033 | 0.960 | 0.995 |

The exponentiated slope coefficient (for age) is very useful. Suppose we compare two films. The older movie was made 1 year earlier than the newer movie. What does our mod\_1 say about the effect of the movie’s age?

* The exponentiated coefficient for age, 0.977, describes the **relative odds** of passing the Bechdel-Wallace test.

## Interpreting the Relative Odds

The movie whose age is one year older has 0.977 times the odds (97.7% of the odds) of the younger movie of passing the Bechdel-Wallace test, according to mod\_1.

* Movie A: age = 10, has log odds(pass) = 0.780 - 0.023 (10) = 0.55, so odds(pass) = exp(0.55) = 1.733
* Movie B: age = 9, has log odds(pass) = 0.780 - 0.023 (9) = 0.573, so odds(pass) = exp(0.573) = 1.774
* Relative odds (A vs. B) are thus 1.733 / 1.774 = 0.977

## Relative Odds with 2-year gap

Exponentiated age coefficient is 0.977, according to mod\_1. What does this imply about the impact on the odds of passing when we have a 2-year difference in age?

* Movie C: age = 12, has log odds(pass) = 0.780 - 0.023 (12) = 0.504, so odds(pass) = exp(0.504) = 1.655329
* Movie A: age = 10, has log odds(pass) = 0.780 - 0.023 (10) = 0.55, so odds(pass) = exp(0.55) = 1.733253
* Relative odds (C vs. A) are 1.655329 / 1.733253 = 0.955

Note that , as well.

## Relative Odds with 10-year gap

Exponentiated age coefficient is 0.977, according to mod\_1. What does this imply about a 10-year difference in age?

* Movie D: age = 20, has log odds(pass) = 0.780 - 0.023 (20) = 0.32, so odds(pass) = exp(0.32) = 1.377128
* Movie A: age = 10, has log odds(pass) = 0.780 - 0.023 (10) = 0.55, so odds(pass) = exp(0.55) = 1.733253
* Relative odds (D vs. A) are 1.377128 / 1.733253 = 0.79

Note that , as well.

## summary of mod\_1

summary(mod\_1)

Call:  
glm(formula = (bechdel == "Pass") ~ age, family = binomial(link = "logit"),   
 data = mov23a)  
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 0.77951 0.27455 2.839 0.00452 \*\*  
age -0.02291 0.01077 -2.127 0.03338 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 255.32 on 186 degrees of freedom  
Residual deviance: 250.60 on 185 degrees of freedom  
AIC: 254.6  
  
Number of Fisher Scoring iterations: 4

## Likelihood Ratio Test: Model 1

* compares model mod\_1 to a null model (with only an intercept term)

anova(mod\_1, test = "LRT")

Analysis of Deviance Table  
  
Model: binomial, link: logit  
  
Response: (bechdel == "Pass")  
  
Terms added sequentially (first to last)  
  
 Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
NULL 186 255.32   
age 1 4.7223 185 250.60 0.02977 \*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Other ANOVA options for glm()

* We can also get Rao’s efficient score test (test = "Rao") or Pearson’s chi-square test (test = "Chisq")

anova(mod\_1, test = "Rao")

Analysis of Deviance Table  
  
Model: binomial, link: logit  
  
Response: (bechdel == "Pass")  
  
Terms added sequentially (first to last)  
  
 Df Deviance Resid. Df Resid. Dev Rao Pr(>Chi)   
NULL 186 255.32   
age 1 4.7223 185 250.60 4.7132 0.02993 \*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## What’s in glance(mod\_1)?

glance(mod\_1) |>  
 gt() |> tab\_options(table.font.size = 24)

| null.deviance | df.null | logLik | AIC | BIC | deviance | df.residual | nobs |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 255.325 | 186 | -125.3014 | 254.6027 | 261.0649 | 250.6027 | 185 | 187 |

* nobs = we fit null model and mod\_1 using 187 observations
* null model (intercept) has 186 residual df (df.null) with null.deviance of 255.3
* mod\_1 (includes age) has 185 residual df (df.residual) with deviance of 250.6
* Think of deviance quantifying what has not yet been explained by model
  + Our mod\_1 has deviance= -2\*log likelihood (logLik`)
  + These are the elements of the ANOVA tests we discussed
* AIC and BIC for comparing models for the same outcome, as in linear regression

## Evaluating prediction quality?

The Receiver Operating Characteristic (ROC) curve is one approach. We can calculate the Area under this curve (sometimes labeled AUC or just C). AUC falls between 0 and 1.

| AUC | Interpretation |
| --- | --- |
| 0.5 | A coin-flip. Model is no better than flipping a coin. |
| 0.6 | Still a fairly weak model. |
| 0.7 | Low end of an “OK” model fit. |
| 0.8 | Pretty good predictive performance. |
| 0.9 | Outstanding predictive performance. |
| 1.0 | Perfect predictive performance. |

## How well does mod\_1 predict?

1. Collected predicted probabilities for our mov23a data:

predict.prob1 <- predict(mod\_1, type = "response")

1. Calculate the ROC curve (roc() from pROC package)

roc1 <- roc(mod\_1$data$bechdel, predict.prob1)

Setting levels: control = Fail, case = Pass

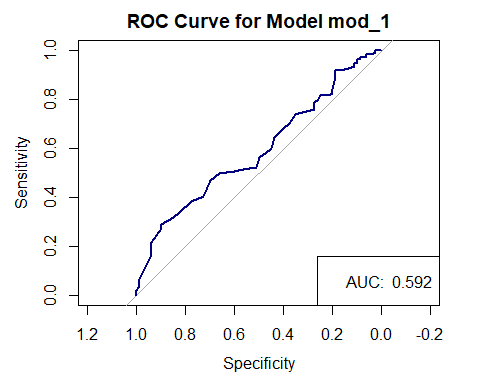
Setting direction: controls < cases

roc1

Call:  
roc.default(response = mod\_1$data$bechdel, predictor = predict.prob1)  
  
Data: predict.prob1 in 80 controls (mod\_1$data$bechdel Fail) < 107 cases (mod\_1$data$bechdel Pass).  
Area under the curve: 0.5918

## Plotting the ROC Curve for mod\_1

plot(roc1, main = "ROC Curve for Model mod\_1", lwd = 2, col = "navy")  
legend('bottomright', legend = paste("AUC: ", round\_half\_up(auc(roc1),3)))



## mod\_1 summaries after lrm fit

d <- datadist(mov23a); options(datadist = "d")  
mod1\_lrm <- lrm((bechdel == "Pass") ~ age, data = mov23a,   
 x = TRUE, y = TRUE)  
mod1\_lrm

Logistic Regression Model  
  
lrm(formula = (bechdel == "Pass") ~ age, data = mov23a, x = TRUE,   
 y = TRUE)  
  
 Model Likelihood Discrimination Rank Discrim.   
 Ratio Test Indexes Indexes   
Obs 187 LR chi2 4.72 R2 0.033 C 0.592   
 FALSE 80 d.f. 1 R2(1,187)0.020 Dxy 0.184   
 TRUE 107 Pr(> chi2) 0.0298 R2(1,137.3)0.027 gamma 0.188   
max |deriv| 2e-09 Brier 0.239 tau-a 0.090   
  
 Coef S.E. Wald Z Pr(>|Z|)  
Intercept 0.7795 0.2746 2.84 0.0045   
age -0.0229 0.0108 -2.13 0.0334

* Here, C = C statistic (AUC), Somers’ d = Dxy.
* Note that C = 0.5 + Dxy/2, by definition.

## Bootstrap validate mod\_1 summaries

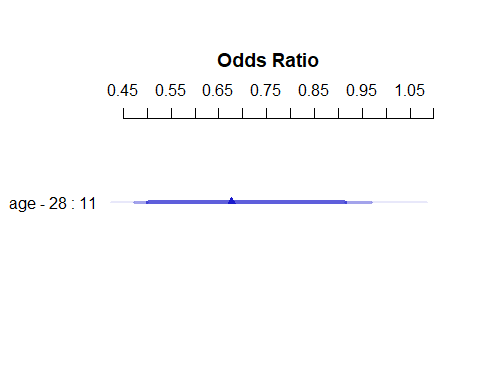
set.seed(20240208); validate(mod1\_lrm, B = 50)

index.orig training test optimism index.corrected n  
Dxy 0.1836 0.2113 0.1763 0.0350 0.1487 50  
R2 0.0335 0.0519 0.0335 0.0185 0.0150 50  
Intercept 0.0000 0.0000 0.1943 -0.1943 0.1943 50  
Slope 1.0000 1.0000 0.5401 0.4599 0.5401 50  
Emax 0.0000 0.0000 0.1731 0.1731 0.1731 50  
D 0.0199 0.0343 0.0199 0.0144 0.0055 50  
U -0.0107 -0.0107 0.0010 -0.0117 0.0010 50  
Q 0.0306 0.0450 0.0189 0.0261 0.0045 50  
B 0.2388 0.2332 0.2416 -0.0085 0.2473 50  
g 0.3508 0.4197 0.3508 0.0689 0.2819 50  
gp 0.0848 0.0984 0.0848 0.0136 0.0711 50

Since C = 0.5 + Dxy / 2, our index-corrected (e.g., bootstrap-validated) C statistic = 0.5 + (0.1487/2) = 0.57435

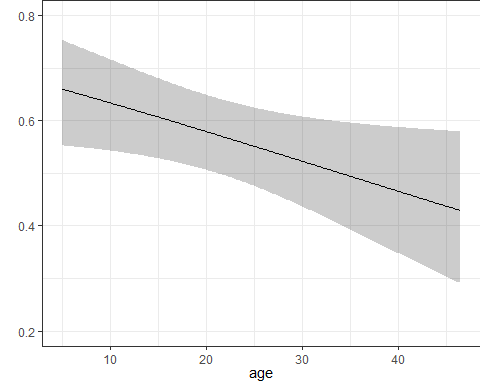
## Effects Plot for mod\_1

plot(summary(mod1\_lrm))



## mod\_1 on Probability Scale

ggplot(Predict(mod1\_lrm, fun = plogis))



# Predicting Pr(bechdel = Pass) using three predictors

## Model mod\_2

mod\_2 <- glm((bechdel == "Pass") ~ age + metascore + mpa3, data = mov23a,   
 family = binomial(link = logit))  
  
mod\_2

Call: glm(formula = (bechdel == "Pass") ~ age + metascore + mpa3, family = binomial(link = logit),   
 data = mov23a)  
  
Coefficients:  
(Intercept) age metascore mpa3R mpa3Other   
 1.453139 -0.027700 -0.009322 -0.162648 0.482545   
  
Degrees of Freedom: 186 Total (i.e. Null); 182 Residual  
Null Deviance: 255.3   
Residual Deviance: 246.5 AIC: 256.5

## Our mod\_2 Equation

## Tidied mod\_2 coefficients

tidy(mod\_2, conf.int = TRUE, conf.level = 0.90) |>   
 gt() |> fmt\_number(decimals = 3) |> tab\_options(table.font.size = 24)

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- | --- |
| (Intercept) | 1.453 | 0.743 | 1.957 | 0.050 | 0.251 | 2.705 |
| age | -0.028 | 0.012 | -2.347 | 0.019 | -0.048 | -0.009 |
| metascore | -0.009 | 0.010 | -0.920 | 0.358 | -0.026 | 0.007 |
| mpa3R | -0.163 | 0.375 | -0.434 | 0.664 | -0.781 | 0.454 |
| mpa3Other | 0.483 | 0.402 | 1.200 | 0.230 | -0.172 | 1.154 |

## Predictions for our three extra movies

new3\_b <- tibble(  
 film = c("Godfather II", "Chinatown", "Incredibles"),  
 age = c(50, 50, 20),   
 metascore = c(90, 92, 90),   
 mpa3 = c("R", "R", "Other") )  
  
augment(mod\_2, newdata = new3\_b, type.predict = "response") |>   
 gt() |> tab\_options(table.font.size = 24)

| film | age | metascore | mpa3 | .fitted |
| --- | --- | --- | --- | --- |
| Godfather II | 50 | 90 | R | 0.2822175 |
| Chinatown | 50 | 92 | R | 0.2784562 |
| Incredibles | 20 | 90 | Other | 0.6324414 |

## Tidied mod\_2 Odds Ratios

After exponentiating…

tidy(mod\_2, exponentiate = TRUE,   
 conf.int = TRUE, conf.level = 0.90) |>   
 gt() |> fmt\_number(decimals = 3) |> tab\_options(table.font.size = 24)

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- | --- |
| (Intercept) | 4.277 | 0.743 | 1.957 | 0.050 | 1.285 | 14.949 |
| age | 0.973 | 0.012 | -2.347 | 0.019 | 0.953 | 0.991 |
| metascore | 0.991 | 0.010 | -0.920 | 0.358 | 0.974 | 1.007 |
| mpa3R | 0.850 | 0.375 | -0.434 | 0.664 | 0.458 | 1.575 |
| mpa3Other | 1.620 | 0.402 | 1.200 | 0.230 | 0.842 | 3.171 |

## glance for mod\_1 and mod\_2

bind\_rows(glance(mod\_1), glance(mod\_2)) |>  
 mutate(model = c("mod\_1", "mod\_2")) |>  
 gt() |>   
 fmt\_number(columns = logLik:BIC, decimals = 1) |>  
 tab\_options(table.font.size = 24)

| null.deviance | df.null | logLik | AIC | BIC | deviance | df.residual | nobs | model |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 255.325 | 186 | -125.3 | 254.6 | 261.1 | 250.6027 | 185 | 187 | mod\_1 |
| 255.325 | 186 | -123.3 | 256.5 | 272.7 | 246.5368 | 182 | 187 | mod\_2 |

* What conclusions does this output suggest?

## Compare “Nested” Models

* This is OK since mod\_1 is a subset of mod\_2

anova(mod\_1, mod\_2, test = "LRT")

Analysis of Deviance Table  
  
Model 1: (bechdel == "Pass") ~ age  
Model 2: (bechdel == "Pass") ~ age + metascore + mpa3  
 Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
1 185 250.60   
2 182 246.54 3 4.0659 0.2544

* Rao’s efficient score test (test = "Rao") yields p = 0.2622
* Pearson’s test (test = "Chisq") also yields p = 0.2544
* Conclusions?

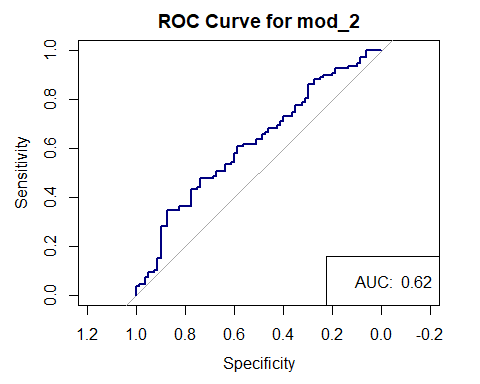
## Plotting the ROC curve for mod\_2

predict.prob2 <- predict(mod\_2, type = "response")  
roc2 <- roc(mod\_2$data$bechdel, predict.prob2)

Setting levels: control = Fail, case = Pass

Setting direction: controls < cases

plot(roc2, main = "ROC Curve for mod\_2", lwd = 2, col = "navy")  
legend('bottomright', legend = paste("AUC: ",round\_half\_up(auc(roc2),3)))



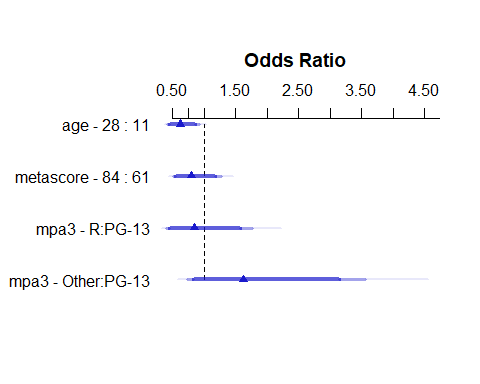
## mod\_2 via lrm fit

d <- datadist(mov23a); options(datadist = "d")  
mod2\_lrm <- lrm((bechdel == "Pass") ~ age + metascore + mpa3,   
 data = mov23a, x = TRUE, y = TRUE)  
  
mod2\_lrm

Logistic Regression Model  
  
lrm(formula = (bechdel == "Pass") ~ age + metascore + mpa3, data = mov23a,   
 x = TRUE, y = TRUE)  
  
 Model Likelihood Discrimination Rank Discrim.   
 Ratio Test Indexes Indexes   
Obs 187 LR chi2 8.79 R2 0.062 C 0.620   
 FALSE 80 d.f. 4 R2(4,187)0.025 Dxy 0.240   
 TRUE 107 Pr(> chi2) 0.0666 R2(4,137.3)0.034 gamma 0.240   
max |deriv| 1e-05 Brier 0.234 tau-a 0.118   
  
 Coef S.E. Wald Z Pr(>|Z|)  
Intercept 1.4531 0.7426 1.96 0.0504   
age -0.0277 0.0118 -2.35 0.0189   
metascore -0.0093 0.0101 -0.92 0.3578   
mpa3=R -0.1626 0.3747 -0.43 0.6642   
mpa3=Other 0.4825 0.4020 1.20 0.2300

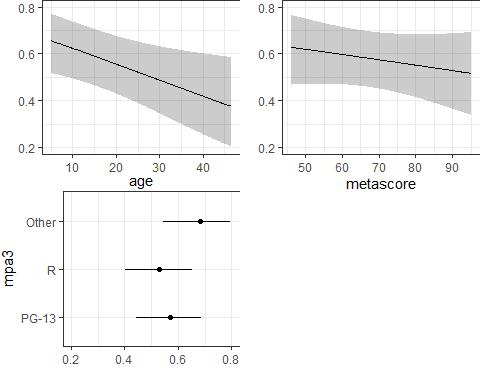
## Effects Plot for mod\_2

plot(summary(mod2\_lrm))



## mod\_2 on Probability Scale

ggplot(Predict(mod2\_lrm, fun = plogis))



## Bootstrap validate mod\_2 summaries

set.seed(202402082); validate(mod2\_lrm, B = 75)

index.orig training test optimism index.corrected n  
Dxy 0.2397 0.2795 0.2003 0.0792 0.1605 75  
R2 0.0616 0.0894 0.0443 0.0451 0.0165 75  
Intercept 0.0000 0.0000 0.0841 -0.0841 0.0841 75  
Slope 1.0000 1.0000 0.7122 0.2878 0.7122 75  
Emax 0.0000 0.0000 0.0903 0.0903 0.0903 75  
D 0.0416 0.0639 0.0282 0.0357 0.0060 75  
U -0.0107 -0.0107 0.0054 -0.0161 0.0054 75  
Q 0.0523 0.0746 0.0228 0.0518 0.0005 75  
B 0.2336 0.2277 0.2403 -0.0126 0.2463 75  
g 0.5025 0.6138 0.4184 0.1955 0.3070 75  
gp 0.1181 0.1398 0.0996 0.0402 0.0779 75

Since C = 0.5 + Dxy / 2, our index-corrected (e.g., bootstrap-validated) C statistic = 0.5 + (0.1605/2) = 0.58025

* For mod\_1, our validated C was 0.57435 and was 0.0150

# Predicting Pr(bechdel = Pass) using five predictors (leaving out mpa3)

## Model mod\_3

mod\_3 <- glm((bechdel == "Pass") ~ age + metascore +   
 gross + comedy + drama,  
 data = mov23a, family = binomial(link = logit))  
  
mod\_3

Call: glm(formula = (bechdel == "Pass") ~ age + metascore + gross +   
 comedy + drama, family = binomial(link = logit), data = mov23a)  
  
Coefficients:  
(Intercept) age metascore gross comedy drama   
 0.8651503 -0.0161145 -0.0113486 0.0009088 0.3614753 0.2969940   
  
Degrees of Freedom: 186 Total (i.e. Null); 181 Residual  
Null Deviance: 255.3   
Residual Deviance: 244.5 AIC: 256.5

## mod\_3 Prediction Equation

## Tidied mod\_3 odds ratios

Coefficients have been exponentiated here…

tidy(mod\_3, exponentiate = TRUE, conf.int = TRUE, conf.level = 0.90) |>   
 gt() |> fmt\_number(decimals = 3) |> tab\_options(table.font.size = 24)

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
| --- | --- | --- | --- | --- | --- | --- |
| (Intercept) | 2.375 | 0.836 | 1.035 | 0.300 | 0.607 | 9.607 |
| age | 0.984 | 0.011 | -1.426 | 0.154 | 0.965 | 1.002 |
| metascore | 0.989 | 0.010 | -1.128 | 0.259 | 0.972 | 1.005 |
| gross | 1.001 | 0.000 | 1.963 | 0.050 | 1.000 | 1.002 |
| comedy | 1.435 | 0.349 | 1.036 | 0.300 | 0.811 | 2.564 |
| drama | 1.346 | 0.341 | 0.872 | 0.383 | 0.771 | 2.372 |

## Compare models with glance()

bind\_rows(glance(mod\_1), glance(mod\_2), glance(mod\_3)) |>  
 mutate(model = c("mod\_1", "mod\_2", "mod\_3")) |>  
 gt() |>   
 fmt\_number(columns = logLik:BIC, decimals = 1) |>  
 tab\_options(table.font.size = 24)

| null.deviance | df.null | logLik | AIC | BIC | deviance | df.residual | nobs | model |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 255.325 | 186 | -125.3 | 254.6 | 261.1 | 250.6027 | 185 | 187 | mod\_1 |
| 255.325 | 186 | -123.3 | 256.5 | 272.7 | 246.5368 | 182 | 187 | mod\_2 |
| 255.325 | 186 | -122.3 | 256.5 | 275.9 | 244.5172 | 181 | 187 | mod\_3 |

## ANOVA comparison of mod\_1 to mod\_3

anova(mod\_1, mod\_3, test = "LRT")

Analysis of Deviance Table  
  
Model 1: (bechdel == "Pass") ~ age  
Model 2: (bechdel == "Pass") ~ age + metascore + gross + comedy + drama  
 Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
1 185 250.60   
2 181 244.52 4 6.0855 0.1928

* Rao test: p = 0.2239
* Note that mod\_1 is nested in mod\_3 but mod\_2 isn’t.

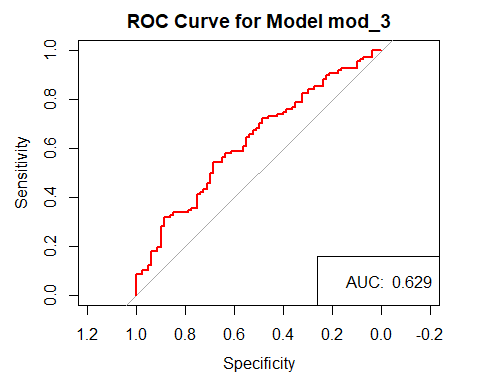
## ROC curve for mod\_3

predict.prob3 <- predict(mod\_3, type = "response")  
roc3 <- roc(mod\_3$data$bechdel, predict.prob3)

Setting levels: control = Fail, case = Pass

Setting direction: controls < cases

plot(roc3, main = "ROC Curve for Model mod\_3", lwd = 2, col = "red")  
legend('bottomright', legend = paste("AUC: ",round\_half\_up(auc(roc3),3)))



## Fit mod\_3 via lrm()

d <- datadist(mov23a); options(datadist = "d")  
  
mod3\_lrm <- lrm((bechdel == "Pass") ~ age +   
 metascore + gross + comedy + drama,  
 data = mov23a, x = TRUE, y = TRUE)

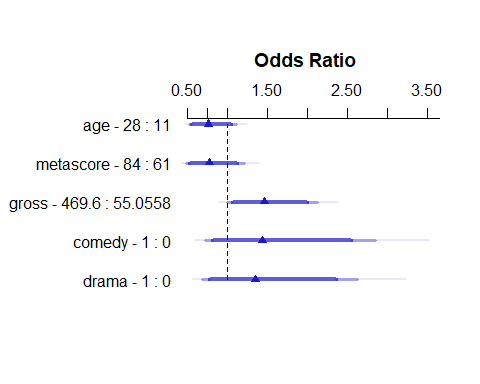
## mod\_3 via lrm() summaries

mod3\_lrm

Logistic Regression Model  
  
lrm(formula = (bechdel == "Pass") ~ age + metascore + gross +   
 comedy + drama, data = mov23a, x = TRUE, y = TRUE)  
  
 Model Likelihood Discrimination Rank Discrim.   
 Ratio Test Indexes Indexes   
Obs 187 LR chi2 10.81 R2 0.075 C 0.629   
 FALSE 80 d.f. 5 R2(5,187)0.031 Dxy 0.259   
 TRUE 107 Pr(> chi2) 0.0553 R2(5,137.3)0.041 gamma 0.259   
max |deriv| 5e-06 Brier 0.232 tau-a 0.127   
  
 Coef S.E. Wald Z Pr(>|Z|)  
Intercept 0.8652 0.8356 1.04 0.3005   
age -0.0161 0.0113 -1.43 0.1540   
metascore -0.0113 0.0101 -1.13 0.2593   
gross 0.0009 0.0005 1.96 0.0496   
comedy 0.3615 0.3490 1.04 0.3003   
drama 0.2970 0.3408 0.87 0.3834

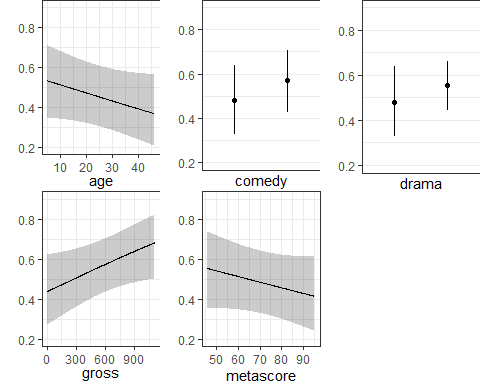
## Effects Plot for mod\_3

plot(summary(mod3\_lrm))



## mod\_3 on Probability Scale

ggplot(Predict(mod3\_lrm, fun = plogis))



## Bootstrap validate mod\_3 summaries

set.seed(202402083); validate(mod3\_lrm, B = 60)

index.orig training test optimism index.corrected n  
Dxy 0.2589 0.2995 0.2095 0.0900 0.1689 60  
R2 0.0754 0.1039 0.0554 0.0485 0.0269 60  
Intercept 0.0000 0.0000 0.0392 -0.0392 0.0392 60  
Slope 1.0000 1.0000 0.7226 0.2774 0.7226 60  
Emax 0.0000 0.0000 0.0790 0.0790 0.0790 60  
D 0.0524 0.0754 0.0368 0.0386 0.0139 60  
U -0.0107 -0.0107 0.0037 -0.0144 0.0037 60  
Q 0.0631 0.0861 0.0331 0.0530 0.0102 60  
B 0.2320 0.2244 0.2387 -0.0143 0.2463 60  
g 0.5663 0.6881 0.4752 0.2129 0.3534 60  
gp 0.1285 0.1494 0.1100 0.0394 0.0891 60

Bootstrap-validated C statistic = 0.5 + (0.1689/2) = 0.58445

* For mod\_1, our validated C was 0.57435 and was 0.0150
* For mod\_2, our validated C = 0.58025, with = 0.0165

## Our Three Extra Movies

new3\_c <- tibble(  
 film = c("Godfather II", "Chinatown", "Incredibles"),  
 age = c(50, 50, 20), metascore = c(90, 92, 90),   
 comedy = c(0, 0, 0), drama = c(1, 1, 0),   
 gross = c(288.741, 175.946, 992.372) )  
  
augment(mod\_3, newdata = new3\_c, type.predict = "response") |>  
 gt() |> tab\_options(table.font.size = 24)

| film | age | metascore | comedy | drama | gross | .fitted |
| --- | --- | --- | --- | --- | --- | --- |
| Godfather II | 50 | 90 | 0 | 1 | 288.741 | 0.4006969 |
| Chinatown | 50 | 92 | 0 | 1 | 175.946 | 0.3710390 |
| Incredibles | 20 | 90 | 0 | 0 | 992.372 | 0.6042738 |

## Actual Bechdel-Wallace Test Results

| Film | Bechdel-Wallace Rating | Bechdel Test |
| --- | --- | --- |
| The Godfather, Part II | 2 | Fail |
| Chinatown | 2 | Fail |
| The Incredibles | 3 | Pass |

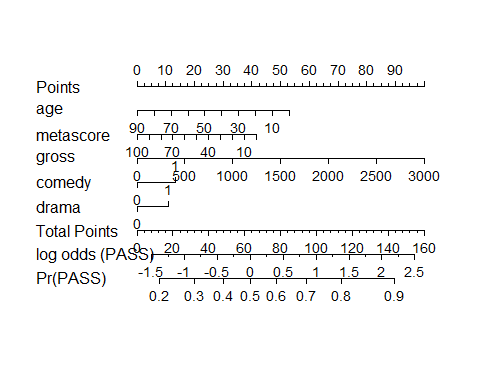
Ratings obtained through API at bechdeltest.com

* 0 means “no two named women”
* 1 means “no talking between the women”
* 2 means “talking only about a man”
* 3 means “passes the test”

Example: <https://bechdeltest.com/api/v1/getMovieByImdbId?imdbid=0071315>

## Nomogram for mod\_3

plot(nomogram(mod3\_lrm, fun = plogis, funlabel = "Pr(PASS)"),  
 lplabel = "log odds (PASS)")



## Next Week (back in person!)

1. Walking through necessary analyses for Project A’s logistic regression model
2. Bringing non-linear terms into our logistic regression models
3. Dealing with missing data more carefully

### Project A Plan due MONDAY 2024-02-12

Get everything into Canvas by Noon Monday, **please**!

1. See <https://feministfrequency.com/video/the-bechdel-test-for-women-in-movies/> [↑](#footnote-ref-26)