432 Class 08

https://thomaselove.github.io/432-2023/

2023-02-09

Today's Agenda

- The Favorite Movies Data
- The Bechdel Test
- Fitting Three Logistic Regression Models with `glm()` and `lrm()`
  - Using tidy, glance and augment from `broom`
  - Making Predictions with the model
  - Interpreting exponentiated coefficients as odds ratios
  - Likelihood Ratio Tests
  - ROC curve and the Area under the Curve
  - Summaries from `lrm`
  - Validating Model Summaries

See Chapters 19-21 in our Course Notes for more on these models.

Today's R Setup

```
knitr::opts_chunk$set(comment = NA)
options(width = 55) # for slides

library(googlesheets4) # import from Google Sheet
library(broom)
library(janitor)
library(knitr)
library(naniar)
library(pROC)
library(rms)
library(tidyverse)

theme_set(theme_bw())
```

Section 1

Our "Favorite Movies" Data

Get The Movies Data from a Google Sheet

```
gs4_deauth()
mov23_full <- read_sheet("https://docs.google.com/spreadsheets

dim(mov23_full)
```

```
[1] 156  33
```

Select Today's Variables

```
mov23 <- mov23_full |>
  select(film_id, bechdel, year, mpa, meta_score,
         gross_ww_2023, comedy, drama, country, film) |>
  type.convert(as.is = FALSE) |>
  mutate(film_id = as.character(film_id),
         film = as.character(film))

dim(mov23)
```

```
[1] 156  10
```

The Bechdel Test

The Bechdel 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]. To pass the test, a movie has to have:

1. at least two (named) women in it
2. who talk to each other
3. about something besides a man

   The Bechdel Test, or Bechdel-Wallace Test was popularized by Alison Bechdel's comic, in a 1985 strip called The Rule.

- from https://bechdeltest.com/

[1] See https://feministfrequency.com/video/the-bechdel-test-for-women-in-movies/

How Many of Our Favorites Pass the Bechdel Test?

```
mov23 |> tabyl(bechdel) |> adorn_pct_formatting()
```

```
 bechdel  n percent valid_percent
    Fail 63   40.4%         41.4%
    Pass 89   57.1%         58.6%
    <NA>  4    2.6%             -
```

Some Cleaning Up and Rescaling of Variables

Since `bechdel` will be our outcome today, we'll drop those films who are missing this information.

```
mov23 <- mov23 |>
  filter(complete.cases(bechdel))
```

We'll also create an age variable and use it instead of `year`, and we'll make sure that `bech` is 1 when the film passes the test, and 0 when the film fails.

```
mov23 <- mov23 |>
  mutate(age = 2023-year,
         bech = ifelse(bechdel == "Pass", 1, 0))
```

Codebook

Variable	Description
`film_id`	identifying code (M-001 through M-156)
`bech`	0 = Failed Bechdel Test or 1 = Passed Test
`age`	2003 - Year of release (1942-2022), so age in years
`mpa`	MPA rating (G, PG, PG-13, R or NR)
`meta_score`	Metacritic score (from critics: 0-100 scale)
`gross_ww_23`	Worldwide gross income in millions of 2023 US dollars
`comedy`	Is comedy one of the three genres listed at IMDB?
`drama`	Is drama one of the three genres listed at IMDB?
`country`	country of origin (first listed at IMDB)
`film`	title of film

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

How Much Missing Data Are We To Deal With?

```
miss_var_summary(mov23) |> filter(n_miss > 0)
```

```
# A tibble: 1 x 3
  variable   n_miss pct_miss
  <chr>       <int>    <dbl>
1 meta_score      3     1.97
```

Which films are missing meta_score?

```
miss_case_summary(mov23) |> filter(n_miss > 0)
```

```
# A tibble: 3 x 3
   case n_miss pct_miss
  <int>  <int>    <dbl>
1    29      1     8.33
2   101      1     8.33
3   151      1     8.33
```

Identifying the films with missing data

```
mov23 |> select(film_id, film, meta_score, country) |>
  slice(c(29, 101, 151))
```

```
# A tibble: 3 x 4
  film_id film                              meta_sc~1 country
  <chr>   <chr>                                 <int> <fct>
1 M-029   Dilwale Dulhania Le Jayenge              NA India
2 M-105   Pather Panchali                          NA India
3 M-155   Yeh Jawaani hai Deewani                  NA India
# ... with abbreviated variable name 1: meta_score
```

How Many of Our Favorites are U.S. Movies?

```
mov23 <- mov23 |>
  mutate(usa = ifelse(country == "USA", 1, 0))
mov23 |> tabyl(usa, country)
```

```
 usa Australia Canada France Germany India Ireland
   0         1      1      3       3     3       1
   1         0      0      0       0     0       0
 Italy Japan Lebanon New Zealand Norway Spain
     3     5       1           4      1     2
     0     0       0           0      0     0
 United Kingdom USA
             14   0
              0 110
```

We'll drop the three films from India (no `meta_score`)

```
mov23 <- mov23 |> filter(complete.cases(meta_score))
```

How About the MPA Ratings?

Let's collapse to the three largest categories.

```
mov23 <- mov23 |> mutate(mpa3 = fct_lump_n(mpa, n = 2))

mov23 |> tabyl(mpa3, mpa) |>
  adorn_totals(where = c("row", "col"))
```

```
  mpa3 G NR PG PG-13  R Total
 PG-13 0  0  0    52  0    52
     R 0  0  0     0 52    52
 Other 4  2 39     0  0    45
 Total 4  2 39    52 52   149
```

Splitting the sample?

We have 149 films in our `mov23` tibble.

- It turns out that a logistic regression model needs about 96 observations just to fit a reasonable intercept term.
- Each additional coefficient we need to fit requires another 10-20 observations for us to get results that will validate well.

Here, we have seven predictors (age, mpa3, meta_score, gross_ww_23, comedy, drama and usa) we want to explore.

Does it make sense to split the sample into separate training and testing samples?

Section 2

Model 1. Using `year` to predict Pr(`bechdel` = Pass)

The Logistic Regression Model

$$logit(event) = log\left(\frac{Pr(event)}{1 - Pr(event)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

$$odds(event) = \frac{Pr(event)}{1 - Pr(event)}$$

$$Pr(event) = \frac{odds(event)}{odds(event) + 1}$$

$$Pr(event) = \frac{exp(logit(event))}{1 + exp(logit(event))}$$

Model 1

```
mod_1 <- glm(bech ~ age,
             data = mov23, family = binomial(link = "logit"))

mod_1$coefficients
```

```
(Intercept)         age
 0.96239924 -0.03146865
```

$$\log\left[\frac{\widehat{P(\text{bech} = 1)}}{1 - \widehat{P(\text{bech} = 1)}}\right] = 0.962 - 0.031(\text{age})$$

Tidied Model 1 coefficients

```
tidy(mod_1, conf.int = TRUE, conf.level = 0.90) |>
  kable(digits = c(0, 3, 3, 2, 3, 2, 2))
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.962	0.309	3.11	0.002	0.47	1.49
age	-0.031	0.013	-2.51	0.012	-0.05	-0.01

Predicting Pr(pass Bechdel) for a 50 year old movie

$$\log\left[\frac{P(\widehat{\text{bech}=1})}{1-P(\widehat{\text{bech}=1})}\right] = 0.962 - 0.031(\text{age})$$

$$logit(bechdel = Pass) = 0.962 - .031(50) = -0.588$$

$$odds(bechdel = Pass) = exp(-0.588) = 0.5554$$

$$Pr(bechdel = Pass) = 0.5554/(1 + 0.5554) = 0.357$$

Estimated Percentage Chance of Passing Bechdel is 35.7%.

Predictions for three movies (not in mov23 data)

Movie	Year	Age
The Godfather, Part II	1974	49
Chinatown	1974	49
The Incredibles	2004	19

```
new3_a <- tibble(age = c(49, 49, 19),
                 film = c("Godfather II", "Chinatown", "Incredib

augment(mod_1, newdata = new3_a, type.predict = "response")
```

```
# A tibble: 3 x 3
    age film          .fitted
  <dbl> <chr>           <dbl>
1    49 Godfather II    0.359
2    49 Chinatown       0.359
3    19 Incredibles     0.590
```

Tidied Model 1 coefficients (after exponentiating)

```
tidy(mod_1, exponentiate = TRUE, conf.int = TRUE, conf.level =
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.618	0.309	3.114	0.002	1.592	4.415
age	0.969	0.013	-2.515	0.012	0.948	0.989

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 can we conclude about the effect of the movie's age based on mod_1? The exponentiated coefficient for age, 0.969, describes the relative odds of passing the Bechdel test.

- Specifically, the movie whose age is one year older has 0.969 times the odds (96.9% of the odds) of the younger movie of passing the Bechdel test, according to our model mod_1.

What does `glance(mod_1)` tell us?

```
glance(mod_1) |> kable(digits = 1)
```

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
203	148	-98.1	200.1	206.1	196.1	147	149

Likelihood Ratio Test: Model 1

- compares model mod_1 to a null model
- can also get Rao's efficient score test (test = "Rao")
- or Pearson's chi-square test (test = "Chisq")

```
anova(mod_1, test = "LRT")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: bech

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                   148     202.99
age   1   6.8917       147     196.10  0.00866 **
---
```

How do we evaluate 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 mov23 data:

```
predict.prob1 <- predict(mod_1, type = "response")
```

2. Calculate the ROC curve

```
roc1 <- roc(mod_1$data$bech, predict.prob1)
roc1
```

```
Call:
roc.default(response = mod_1$data$bech, predictor = predict.pr

Data: predict.prob1 in 63 controls (mod_1$data$bech 0) < 86 ca
Area under the curve: 0.612
```

Plotting the ROC Curve for mod_1

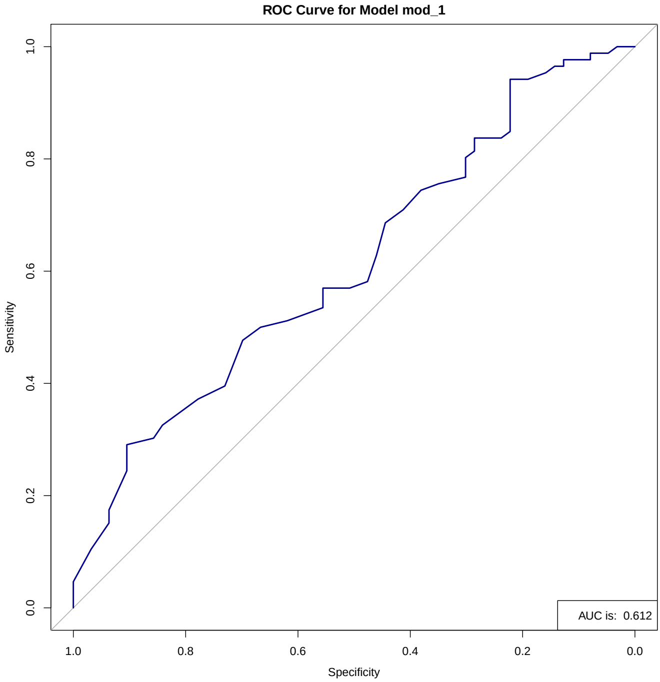
The complete output from the call to roc1 was

```
Call:
roc.default(response = mod_1$data$bech,
                    predictor = predict.prob1)
Data: predict.prob1 in 63 controls (mod_1$data$bechdel 0)
                        < 86 cases (mod_1$data$bechdel 1).
Area under the curve: 0.612
```

The actual plot will be on the next slide.

```
plot(roc1, main = "ROC Curve for Model mod_1",
     lwd = 2, col = "blue4")
legend('bottomright',
       legend = paste("AUC is: ",round_half_up(auc(roc1),3)))
```

Note that I used `#| fig-asp: 1` to obtain a square plot.


ROC Curve for Model mod_1
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC is: 0.612

Model Summaries via lrm fit

```
d <- datadist(mov23)
options(datadist = "d")

mod1_lrm <- lrm(bech ~ age, data = mov23,
                x = TRUE, y = TRUE)
```

What's in mod1_lrm?

```
> mod1_lrm
Logistic Regression Model

 lrm(formula = bech ~ age, data = mov23, x = TRUE, y = TRUE)

                      Model Likelihood      Discrimination    Rank Discrim.
                         Ratio Test            Indexes           Indexes
 Obs           149    LR chi2      6.89     R2       0.061    C       0.612
  0             63    d.f.            1     R2(1,149)0.039    Dxy     0.224
  1             86    Pr(> chi2) 0.0087   R2(1,109.1)0.053    gamma   0.229
 max |deriv| 1e-06                          Brier    0.233    tau-a   0.110

          Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept  0.9624 0.3091  3.11  0.0018
 age       -0.0315 0.0125 -2.51  0.0119
```

Section 3

Model 2. Predicting Pr(`bechdel` = Pass) using four predictors

Model 2

```r
mod_2 <- glm(bech ~ age + meta_score +
               mpa3 + usa, data = mov23,
             family = binomial(link = logit))
```

$$
\log\left[\frac{P(\widehat{\text{bech} = 1})}{1 - P(\widehat{\text{bech} = 1})}\right] = 2.288 - 0.035(\text{age}) \\
- 0.018(\text{meta\_score}) - 0.172(\text{mpa3}_{\text{R}}) \\
+ 0.378(\text{mpa3}_{\text{Other}}) - 0.052(\text{usa})
$$

Predictions for three movies (not in mov23 data)

```
new3_b <- tibble(meta_score = c(90, 92, 90), mpa3 = c("R", "R'
                 usa = c(1, 1, 1), age = c(49, 49, 19),
                 film = c("Godfather II", "Chinatown", "Incredil

augment(mod_2, newdata = new3_b, type.predict = "response")
```

```
# A tibble: 3 x 6
  meta_score mpa3    usa   age film          .fitted
       <dbl> <chr> <dbl> <dbl> <chr>           <dbl>
1         90 R         1    49 Godfather II    0.222
2         92 R         1    49 Chinatown       0.216
3         90 Other     1    19 Incredibles     0.588
```

Tidied Model 2 coefficients

```
tidy(mod_2, exponentiate = TRUE,
     conf.int = TRUE, conf.level = 0.90) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	9.860	0.943	2.427	0.015	2.186	49.402
age	0.965	0.014	-2.556	0.011	0.942	0.987
meta_score	0.983	0.011	-1.537	0.124	0.964	1.001
mpa3R	0.842	0.415	-0.413	0.679	0.425	1.669
mpa3Other	1.459	0.466	0.811	0.417	0.683	3.181
usa	0.950	0.395	-0.131	0.896	0.492	1.810

Compare mod_1 to mod_2 with glance()

```
bind_rows(glance(mod_1), glance(mod_2)) |>
  mutate(model = c("1", "2")) |>
  kable(digits = 1)
```

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs	model
203	148	-98.1	200.1	206.1	196.1	147	149	1
203	148	-95.7	203.4	221.5	191.4	143	149	2

- What conclusions does this output suggest?

Compare Models 1 and 2 with ANOVA

- compares model mod_1 to a null model

```
anova(mod_1, mod_2, test = "LRT")
```

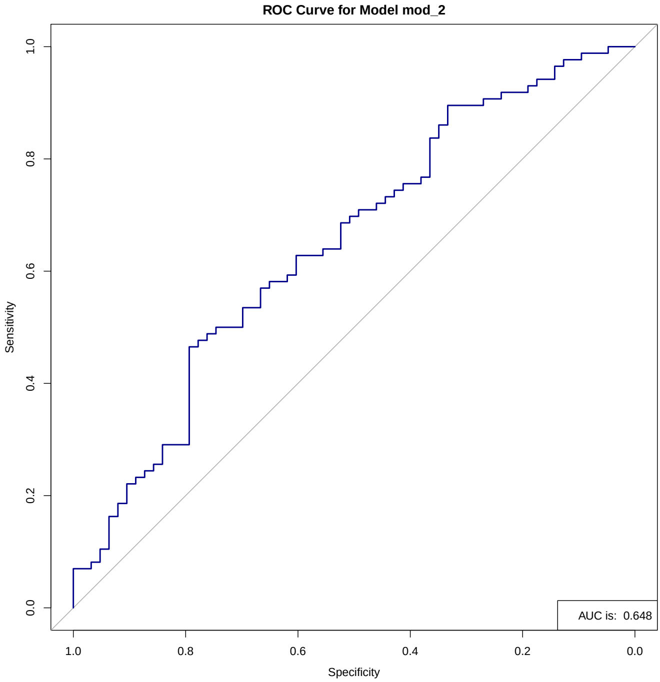
```
Analysis of Deviance Table

Model 1: bech ~ age
Model 2: bech ~ age + meta_score + mpa3 + usa
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       147     196.10
2       143     191.44  4   4.6591   0.3241
```

- Rao's efficient score test (test = "Rao") yields p = 0.3359
- Pearson's chi-square test (test = "Chisq") also yields p = 0.3241
- Conclusions?

Plotting the ROC curve for Model mod_2

```
predict.prob2 <- predict(mod_2, type = "response")
roc2 <- roc(mod_2$data$bech, predict.prob2)
plot(roc2, main = "ROC Curve for Model mod_2",
     lwd = 2, col = "blue4")
legend('bottomright',
       legend = paste("AUC is: ",round_half_up(auc(roc2),3)))
```

Result on Next Slide


ROC Curve for Model mod_2
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC is: 0.648

Model Summaries via lrm fit

```
d <- datadist(mov23)
options(datadist = "d")

mod2_lrm <- lrm(bech ~ age + meta_score + mpa3 + usa,
                data = mov23, x = TRUE, y = TRUE)
```

What's in mod2_lrm?

```
> mod2_lrm
Logistic Regression Model

 lrm(formula = bech ~ age + meta_score + mpa3 + usa, data = mov23,
     x = TRUE, y = TRUE)

                       Model Likelihood      Discrimination    Rank Discrim.
                             Ratio Test             Indexes          Indexes
 Obs           149    LR chi2      11.55      R2          0.100    C        0.648
  0             63    d.f.             5      R2(5,149)0.043    Dxy      0.295
  1             86    Pr(> chi2) 0.0415    R2(5,109.1)0.058    gamma    0.296
 max |deriv| 2e-11                          Brier       0.226    tau-a    0.145

            Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept   2.2885 0.9430  2.43  0.0152
 age        -0.0355 0.0139 -2.56  0.0106
 meta_score -0.0176 0.0114 -1.54  0.1243
 mpa3=R     -0.1715 0.4148 -0.41  0.6792
 mpa3=Other  0.3778 0.4657  0.81  0.4172
 usa        -0.0518 0.3946 -0.13  0.8956
```

Section 4

Model 3. Predicting $\Pr(\texttt{bechdel} = \text{Pass})$ using five predictors

Model 3

```
mod_3 <- glm(bech ~ age + meta_score +
               gross_ww_2023 + comedy + drama,
             data = mov23, family = binomial(link = logit))
```

$$
\begin{aligned}
\log\left[\frac{P(\widehat{\text{bech} = 1})}{1 - P(\widehat{\text{bech} = 1})}\right] = {} & 1.37 - 0.033(\text{age}) \\
& - 0.023(\text{meta\_score}) + 0.001(\text{gross\_ww\_2023}) \\
& + 0.931(\text{comedy}) + 0.842(\text{drama})
\end{aligned}
$$

Tidied Model 3 coefficients (exponentiated)

```
tidy(mod_3, exponentiate = TRUE,
     conf.int = TRUE, conf.level = 0.90) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.935	0.948	1.445	0.148	0.850	19.534
age	0.968	0.014	-2.381	0.017	0.945	0.989
meta_score	0.978	0.012	-1.922	0.055	0.958	0.996
gross_ww_2023	1.001	0.000	2.396	0.017	1.000	1.001
comedy	2.537	0.423	2.201	0.028	1.284	5.187
drama	2.321	0.401	2.101	0.036	1.215	4.563

Compare models with glance()

```
bind_rows(glance(mod_1), glance(mod_2), glance(mod_3)) |>
  mutate(model = c("1", "2", "3")) |>
  kable(digits = 1)
```

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs	model
203	148	-98.1	200.1	206.1	196.1	147	149	1
203	148	-95.7	203.4	221.5	191.4	143	149	2
203	148	-90.3	192.7	210.7	180.7	143	149	3

ANOVA comparison of mod_1 to mod_3

```
anova(mod_1, mod_3, test = "LRT")
```

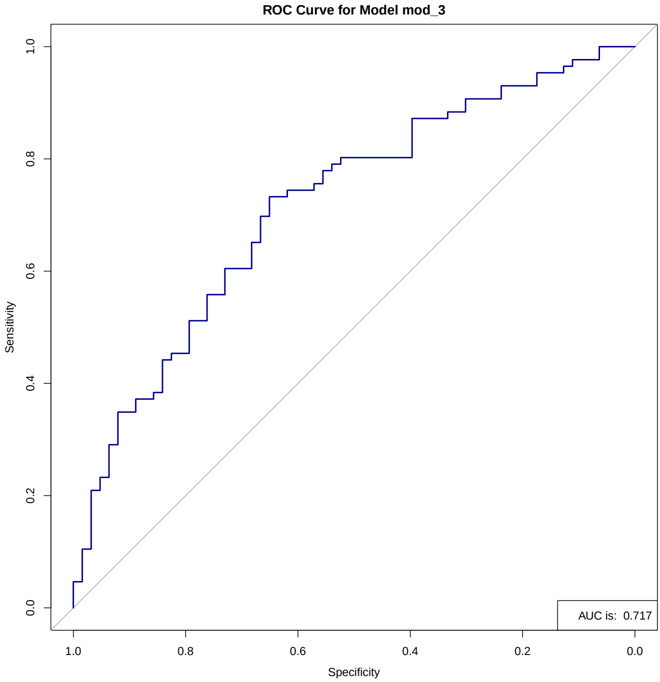
```
Analysis of Deviance Table

Model 1: bech ~ age
Model 2: bech ~ age + meta_score + gross_ww_2023 + comedy + dr
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       147     196.10
2       143     180.65  4   15.448 0.003857 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Rao test: $p = 0.01201$

Plotting the ROC curve for Model mod_3

```
predict.prob3 <- predict(mod_3, type = "response")
roc3 <- roc(mod_3$data$bech, predict.prob3)
plot(roc3, main = "ROC Curve for Model mod_3",
     lwd = 2, col = "blue4")
legend('bottomright',
       legend = paste("AUC is: ",round_half_up(auc(roc3),3)))
```

Result on Next Slide


ROC Curve for Model mod_3
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC is: 0.717

Model Summaries via lrm fit

```
d <- datadist(mov23)
options(datadist = "d")

mod3_lrm <- lrm(bech ~ age + meta_score +
                gross_ww_2023 + comedy + drama,
                 data = mov23, x = TRUE, y = TRUE)
```

What's in `mod3_lrm`?

```
> mod3_lrm
Logistic Regression Model

 lrm(formula = bech ~ age + meta_score + gross_ww_2023 + comedy +
     drama, data = mov23, x = TRUE, y = TRUE)

                       Model Likelihood      Discrimination    Rank Discrim.
                          Ratio Test            Indexes           Indexes
 Obs            149    LR chi2      22.34     R2          0.187    C       0.717
  0              63    d.f.             5     R2(5,149)0.110    Dxy     0.434
  1              86    Pr(> chi2) 0.0005    R2(5,109.1)0.147    gamma   0.434
 max |deriv| 1e-06                            Brier       0.210    tau-a   0.213

               Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept      1.3698 0.9479  1.45  0.1484
 age           -0.0326 0.0137 -2.38  0.0173
 meta_score    -0.0226 0.0118 -1.92  0.0546
 gross_ww_2023  0.0006 0.0003  2.40  0.0166
 comedy         0.9309 0.4229  2.20  0.0277
 drama          0.8421 0.4009  2.10  0.0357
```

Store Validated mod1_lrm and mod3_lrm summaries

```
set.seed(4321)
v1 <- validate(mod1_lrm)

set.seed(4322)
v3 <- validate(mod3_lrm)
```

Now, let's look at the validated Somers' d statistics:

- **AUC = 0.5 + (Somer's d)/2**

```r
v1["Dxy",]
```

```
     index.orig        training            test 
     0.22406792      0.20010649      0.22406792 
       optimism index.corrected               n 
    -0.02396143      0.24802935     40.00000000 
```

```r
v3["Dxy",]
```

```
     index.orig        training            test 
     0.43373939      0.46513170      0.39150978 
       optimism index.corrected               n 
     0.07362192      0.36011747     40.00000000 
```

How about the Nagelkerke R^2 after validation?

```
v1["R2",]
```

```
  index.orig       training            test
 0.060756199     0.057704674     0.060756199
    optimism index.corrected               n
-0.003051525     0.063807724    40.000000000
```

```
v3["R2",]
```

```
  index.orig       training            test
  0.18715264      0.21708060      0.15967581
    optimism index.corrected               n
  0.05740479      0.12974785     40.00000000
```

- Conclusions?

Predictions for three movies (not in mov23 data)

```
new3_c <- tibble(meta_score = c(90, 92, 90), comedy = c(0, 0,
                 gross_ww_2023 = c(288.741, 175.946, 992.372),
               film = c("Godfather II", "Chinatown", "Incredib

augment(mod_3, newdata = new3_c, type.predict = "response")
```

```
# A tibble: 3 x 7
  meta_score comedy drama gross_w~1   age film  .fitted
       <dbl>  <dbl> <dbl>     <dbl> <dbl> <chr>   <dbl>
1         90      0     1      289.    49 Godf~   0.224
2         92      0     1      176.    49 Chin~   0.204
3         90      0     0      992.    19 Incr~   0.339
# ... with abbreviated variable name 1: gross_ww_2023
```

Actual Bechdel Test Results

Film	Bechdel Rating	Result
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

Next Time

1. Walking through necessary analyses for Project A's logistic regression model
2. Plotting and Interpreting Effect Sizes from Logistic Regression Models (see Chapters 21-22)