Data Taming Assignment 3 — SOLUTIONS_0

Bill S. Preston Esq.

Trimester 1, 2024

Setup

```
#Load the required packages
library(tidyverse)
library(tidymodels)
library(modelr)
library(car)
```

Q1. Loading the data

```
# Your student number goes here
ysn<-1
# Calculate your (student number + 2) modulo 3
filenum <- (ysn+2) %% 3
filenum
```

```
## [1] 0
```

```
filename<-paste0("./data/merry_",filenum,".csv")
filename
```

```
## [1] "./data/merry_0.csv"
```

```
# Read in the data
merry_raw <- read_csv(filename)
# Display the first 10 lines of the data
merry_raw
```

```
## # A tibble: 30,000 x 6
##    RHBMM Accuracyyy AGE    DRESS Home   Jail
##    <dbl>      <dbl> <chr>  <chr> <chr>  <chr>
##  1     0      0.672 senior black city   no
##  2     0      0.723 middle black city   no
##  3     1      0.639 youth  black city   no
##  4     0      0.872 middle green city   no
```

```
## 5     0       0.670 middle black city   no
## 6     0       0.625 youth  red   city   no
## 7     0       0.673 senior red   city   no
## 8     0       0.748 middle black city   no
## 9     0       0.742 senior red   city   no
## 10    1       0.913 youth  green forest no
## # i 29,990 more rows
```

Q2. Variable types

- `RHBMM`: **Categorical nominal**. This is just the name assigned to the status of being a Merry Man.
- `Accuracy`: **Quantitative continuous**. This is a value between 0 and 1 (so not an integer), and since we don't know how many arrows each archer shot to judge their accuracy, it could in principle, be any number in this range. So it is quantitative continuous.
- `AGE`: **Categorical ordinal**. This is the name of the category of age of the archer, and age is naturally ordered.
- `DRESS`: **Categorical nominal**. This is just the name of the colour of the archer's clothing, and so it doesn't seem there would be any ordering.
- `Home`: **Categorical nominal** This is just the name of the place where the archer live, and so it doesn't seem there would be any ordering.
- `Jail`: **Categorical nominal**. This is the name assigned to the status of the archer having been in jail.

Q3. Taming the data

```
## Convert column names to snakecase
merry <- rename(merry_raw,
                rhbmm=RHBMM,
                accuracy=Accuracyyy,
                age=AGE,
                dress=DRESS,
                home=Home,
                jail=Jail)

## Convert rhbmm to yes and no factors
merry$rhbmm<-fct_recode(as.character(merry$rhbmm),"no"="0","yes"="1")

## Convert age, dress, home to factors. Jail to logical
merry$jail<-fct_recode(as.character(merry$jail),"FALSE"="no","TRUE"="yes")
merry <- mutate(merry,
                age=as.factor(age),
                dress=as.factor(dress),
                home=as.factor(home),
                jail=as.logical(jail)
                )
merry
```

```
## # A tibble: 30,000 x 6
##    rhbmm accuracy age    dress home   jail
```

```
##     <fct>      <dbl> <fct>  <fct> <fct>  <lgl>
##   1 no         0.672 senior black city   FALSE
##   2 no         0.723 middle black city   FALSE
##   3 yes        0.639 youth  black city   FALSE
##   4 no         0.872 middle green city   FALSE
##   5 no         0.670 middle black city   FALSE
##   6 no         0.625 youth  red   city   FALSE
##   7 no         0.673 senior red   city   FALSE
##   8 no         0.748 middle black city   FALSE
##   9 no         0.742 senior red   city   FALSE
##  10 yes        0.913 youth  green forest FALSE
## # i 29,990 more rows
```

Q4. Splitting data in training and testing sets

```r
set.seed(ysn)
merry_split<-initial_split(merry, prop=2/3)
merry_train<-training(merry_split)
merry_test<-testing(merry_split)
merry_train
```

```
## # A tibble: 20,000 x 6
##     rhbmm accuracy age    dress home   jail
##     <fct>      <dbl> <fct>  <fct> <fct>  <lgl>
##   1 no         0.660 youth  black city   FALSE
##   2 yes        0.548 middle black city   FALSE
##   3 no         0.549 middle red   city   FALSE
##   4 yes        0.521 middle black city   TRUE
##   5 yes        0.915 youth  red   city   TRUE
##   6 yes        0.771 middle green forest TRUE
##   7 no         0.719 senior black city   FALSE
##   8 no         0.713 youth  black city   FALSE
##   9 no         0.731 youth  black city   FALSE
##  10 no         0.742 senior red   city   TRUE
## # i 19,990 more rows
```

```r
merry_test
```

```
## # A tibble: 10,000 x 6
##     rhbmm accuracy age    dress home   jail
##     <fct>      <dbl> <fct>  <fct> <fct>  <lgl>
##   1 no         0.670 middle black city   FALSE
##   2 no         0.625 youth  red   city   FALSE
##   3 no         0.742 senior red   city   FALSE
##   4 yes        0.913 youth  green forest FALSE
##   5 yes        0.701 middle black forest FALSE
##   6 yes        0.601 senior green city   FALSE
##   7 yes        0.783 youth  red   city   FALSE
##   8 no         0.684 senior black city   FALSE
##   9 yes        0.766 senior black forest FALSE
##  10 yes        0.734 middle green city   FALSE
## # i 9,990 more rows
```

Q5. Logistic model with no interactions

```
lr_spec = logistic_reg(mode="classification") %>% set_engine("glm")
lr_spec
```

```
## Logistic Regression Model Specification (classification)
##
## Computational engine: glm
```

```
fitindivs <- fit(lr_spec,rhbmm ~ ., data = merry_train)
summary(fitindivs$fit)
```

```
##
## Call:
## stats::glm(formula = rhbmm ~ ., family = stats::binomial, data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.31349    0.09499 -13.828  < 2e-16 ***
## accuracy     1.39992    0.12740  10.989  < 2e-16 ***
## agesenior   -0.03011    0.04409  -0.683    0.495
## ageyouth    -0.03714    0.03538  -1.050    0.294
## dressgreen   1.06839    0.03917  27.279  < 2e-16 ***
## dressred    -0.41640    0.03709 -11.227  < 2e-16 ***
## homeforest   1.70451    0.03863  44.123  < 2e-16 ***
## jailTRUE     0.20421    0.04329   4.717  2.4e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27337  on 19999  degrees of freedom
## Residual deviance: 23623  on 19992  degrees of freedom
## AIC: 23639
##
## Number of Fisher Scoring iterations: 4
```

Q6. Effect on DRESS and AGE variables

```
model_matrix(merry_train,~dress)
```

```
## # A tibble: 20,000 x 3
##    `(Intercept)` dressgreen dressred
##            <dbl>      <dbl>    <dbl>
##  1             1          0        0
##  2             1          0        0
##  3             1          0        1
##  4             1          0        0
```

```
##   5              1           0         1
##   6              1           1         0
##   7              1           0         0
##   8              1           0         0
##   9              1           0         0
## 10              1           0         1
## # i 19,990 more rows
```

```
model_matrix(merry_train,~age)
```

```
## # A tibble: 20,000 x 3
##    `(Intercept)` agesenior ageyouth
##            <dbl>     <dbl>    <dbl>
##   1             1         0        1
##   2             1         0        0
##   3             1         0        0
##   4             1         0        0
##   5             1         0        1
##   6             1         0        0
##   7             1         1        0
##   8             1         0        1
##   9             1         0        1
## 10             1         1        0
## # i 19,990 more rows
```

Q6(a)

- The dress variable has been split into 2 new binary variables: dressgreen and dressred.
- The age variable has also been split into 2 new binary variables: agesenior and ageyouth.

Q6(b)

- We see that the level black is missing from model matrix for dress, and so black is the reference level.
- Similarly, see that middle is missing from the model matrix for age and so it must be the reference level.

Q7 Number of lines in model

There are 3 levels in age, 3 levels in dress, 2 levels in home and 2 levels in jail. So we have $3 \times 3 \times 2 \times 2 = 36$ number of lines.

Q8 Interacting model

```
fitall<-fit(lr_spec,rhbmm ~ (.)^2, data = merry_train)
Anova(fitall$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##                 LR Chisq Df Pr(>Chisq)
## accuracy          126.08  1  < 2.2e-16 ***
## age                 0.99  2    0.61089
## dress            1422.39  2  < 2.2e-16 ***
## home             2302.49  1  < 2.2e-16 ***
## jail               22.68  1  1.910e-06 ***
## accuracy:age        3.37  2    0.18524
## accuracy:dress      2.86  2    0.23892
## accuracy:home      15.41  1  8.658e-05 ***
## accuracy:jail       0.15  1    0.70207
## age:dress           5.84  4    0.21119
## age:home            4.10  2    0.12850
## age:jail            5.35  2    0.06905 .
## dress:home        363.28  2  < 2.2e-16 ***
## dress:jail          1.39  2    0.49909
## home:jail           0.86  1    0.35466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction terms significant at the 99% significance level are:

- `accuracy:home`
- `dress:home`

Q9 Backwards stepwise regression

Q9(a)

```r
m1<-fit(lr_spec,rhbmm ~ (.)^2-age:jail-home:jail-dress:jail-accuracy:jail, data = merry_train)
Anova(m1$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##                 LR Chisq Df Pr(>Chisq)
## accuracy          126.15  1  < 2.2e-16 ***
## age                 1.01  2     0.6032
## dress            1421.35  2  < 2.2e-16 ***
## home             2306.79  1  < 2.2e-16 ***
## jail               22.68  1  1.910e-06 ***
## accuracy:age        3.40  2     0.1824
## accuracy:dress      2.80  2     0.2461
## accuracy:home      15.56  1  7.996e-05 ***
## age:dress           6.14  4     0.1887
## age:home            4.17  2     0.1242
## dress:home        363.90  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q9(b)

THESE SIGNIFICANCE LEVELS ARE PROBABLY WRONG IN EACH STUDENT'S VERSION

- `accuracy:dress` is least significant

```
m2<-fit(lr_spec,rhbmm ~ (.)^2-age:jail-home:jail-dress:jail-accuracy:jail-accuracy:dress, data = merry_
Anova(m2$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##                LR Chisq Df Pr(>Chisq)
## accuracy         126.15  1  < 2.2e-16 ***
## age                1.06  2  0.5891292
## dress           1421.35  2  < 2.2e-16 ***
## home            2307.51  1  < 2.2e-16 ***
## jail              22.58  1   2.02e-06 ***
## accuracy:age       3.35  2  0.1875577
## accuracy:home     13.81  1  0.0002027 ***
## age:dress          6.24  4  0.1816249
## age:home           4.16  2  0.1246289
## dress:home       362.69  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- `accuracy:age` is least significant

```
m3<-fit(lr_spec,rhbmm ~ (.)^2-age:jail-home:jail-dress:jail-accuracy:jail-accuracy:dress-accuracy:age, d
Anova(m3$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##                LR Chisq Df Pr(>Chisq)
## accuracy         126.15  1  < 2.2e-16 ***
## age                1.06  2  0.5891292
## dress           1423.72  2  < 2.2e-16 ***
## home            2308.12  1  < 2.2e-16 ***
## jail              22.61  1  1.981e-06 ***
## accuracy:home     13.55  1  0.0002326 ***
## age:dress          6.21  4  0.1838951
## age:home           3.76  2  0.1528440
## dress:home       362.63  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- `age:dress` is least significant (interaction) term

```
m4<-fit(lr_spec,rhbmm ~ (.)^2-age:jail-home:jail-dress:jail-accuracy:jail-accuracy:dress-accuracy:age-ag
Anova(m4$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##                LR Chisq Df Pr(>Chisq)
## accuracy         126.50  1  < 2.2e-16 ***
## age                1.06  2  0.5891292
## dress           1423.72  2  < 2.2e-16 ***
## home            2307.52  1  < 2.2e-16 ***
## jail              22.80  1  1.795e-06 ***
## accuracy:home     13.58  1  0.0002285 ***
## age:home           3.33  2  0.1894673
## dress:home       363.40  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- age:home is least significant (interaction) term

```
m5<-fit(lr_spec,rhbmm ~ (.)^2-age:jail-home:jail-dress:jail-accuracy:jail-accuracy:dress-accuracy:age-ag
Anova(m5$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##                LR Chisq Df Pr(>Chisq)
## accuracy         126.44  1  < 2.2e-16 ***
## age                1.06  2  0.5891292
## dress           1424.29  2  < 2.2e-16 ***
## home            2307.52  1  < 2.2e-16 ***
## jail              22.72  1  1.871e-06 ***
## accuracy:home     13.51  1  0.0002367 ***
## dress:home       363.84  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q9(c)

- age is least significant

```
m6<-fit(lr_spec,rhbmm ~ (.)^2-age:jail-home:jail-dress:jail-accuracy:jail-accuracy:dress-accuracy:age-ag
Anova(m6$fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: rhbmm
##                LR Chisq Df Pr(>Chisq)
## accuracy         126.44  1  < 2.2e-16 ***
## dress           1425.38  2  < 2.2e-16 ***
## home            2307.61  1  < 2.2e-16 ***
## jail              22.73  1  1.864e-06 ***
## accuracy:home     13.51  1  0.0002371 ***
## dress:home       363.93  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q10

Q10(a)

`accuracy:home` and `dress:home` are significant at the 95% significance level

Q10(b)

- If you live in the forest, you might be more likely to hunt for food, so you would be better at archery.
- Also, if you live in the forest, you might be less likely to wear red, so that you are camouflaged.

Q11 General form of our estimated log-odds function

$$\hat{r}_i = \hat{\beta}_0 + \hat{\beta}_1 a_i + \hat{\beta}_2 d_i^{(g)} + \hat{\beta}_3 d_i^{(r)} + \hat{\beta}_4 h_i + \hat{\beta}_5 j_i + \hat{\beta}_6\,(a_i \times h_i) + \hat{\beta}_7\,(d_i^{(g)} \times h_i) + \hat{\beta}_8\,(d_i^{(r)} \times h_i)$$

where

- a is the `acc` variable, which is a real number between 0 and 1.
- $d^{(g)}$ is the `dressgreen` class, a binary integer equal to
  - "1" for green clothes
  - "0" for non-green clothes
- $d^{(r)}$ the `dressred` class, a binary integer equal to
  - "1" for red clothes
  - "0" for non-red clothes
- h the `home` class, a binary integer equal to
  - "0" for a city home
  - "1" for a forest home
- j the `jail`, a binary integer equal to
  - "1" for having been to jail
  - "0" for never having been to jail

Q12

Q12(a)

- There are $3 \times (2^2) = 12$ lines. The variables `home` and `jail` each have two options, and `dress` has 3 options. (We saw that the `dress` variable was split into 2 new binary variables, `dressred` and `dressgreen`, but they can't both be 1 at the same time.)

Q12(b)

- No, we have an interaction term $a \times h$, and so the coefficient of `accuracy` (the gradient) will change for the two values of `home`, 0 or 1.

Q13

```r
summary(m6$fit)
```

```
##
## Call:
## stats::glm(formula = rhbmm ~ (.)^2 - age:jail - home:jail - dress:jail -
##     accuracy:jail - accuracy:dress - accuracy:age - age:dress -
##     age:home - age, family = stats::binomial, data = data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.26557    0.10277 -12.315  < 2e-16 ***
## accuracy               1.20195    0.14195   8.467  < 2e-16 ***
## dressgreen             1.00770    0.04226  23.843  < 2e-16 ***
## dressred              -0.12597    0.04215  -2.988 0.002804 **
## homeforest             1.21854    0.22721   5.363 8.18e-08 ***
## jailTRUE               0.20580    0.04325   4.759 1.95e-06 ***
## accuracy:homeforest    1.19227    0.32494   3.669 0.000243 ***
## dressgreen:homeforest  1.49250    0.19701   7.576 3.57e-14 ***
## dressred:homeforest   -1.10104    0.08550 -12.877  < 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: 27337  on 19999  degrees of freedom
## Residual deviance: 23254  on 19991  degrees of freedom
## AIC: 23272
##
## Number of Fisher Scoring iterations: 6
```

```r
m6$fit$coefficients
```

```
##           (Intercept)              accuracy            dressgreen
##            -1.2655714             1.2019451             1.0077020
##              dressred            homeforest              jailTRUE
##            -0.1259669             1.2185360             0.2058022
##   accuracy:homeforest dressgreen:homeforest   dressred:homeforest
##             1.1922699             1.4924954            -1.1010408
```

```r
m6coeffs<-as.numeric(m6$fit$coefficients)
beta0=m6coeffs[1]
beta1=m6coeffs[2]
beta2=m6coeffs[3]
beta3=m6coeffs[4]
beta4=m6coeffs[5]
beta5=m6coeffs[6]
beta6=m6coeffs[7]
beta7=m6coeffs[8]
beta8=m6coeffs[9]
```

This output gives us the equation

$\hat{r}_i = -1.27 + 1.2a_i + 1.01d_i^{(g)} - 0.126d_i^{(r)} + 1.22h_i + 0.206j_i + 1.19\,(a_i \times h_i) + 1.49\,(d_i^{(g)} \times h_i) - 1.1\,(d_i^{(r)} \times h_i)$

Q14

Q14(a)

```
dg<-0
dr<-1
h<-1
j<-1
int1<- beta0+beta2*dg+beta3*dr+beta4*h+beta5*j+beta7*dg*h+beta8*dr*h
int1
```

```
## [1] -1.068241
```

```
slope1<-beta1+beta6*h
slope1
```

```
## [1] 2.394215
```

This gives us an estimated line:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -1.07 + 2.39\ a_i$$

Q14(b)

```
dg<-0
dr<-0
h<-0
j<-0
int2<- beta0+beta2*dg+beta3*dr+beta4*h+beta5*j+beta7*dg*h+beta8*dr*h
int2
```

```
## [1] -1.265571
```

```
slope2<-beta1+beta6*h
slope2
```

```
## [1] 1.201945
```

This gives us an estimated line:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -1.27 + 1.2\ a_i$$

Q15

```r
merry_pred =
  bind_cols(merry_test[,"rhbmm"],
            predict(m6, merry_test, type = "class"),
            predict(m6, merry_test, type = "prob")
  )
merry_pred
```

```
## # A tibble: 10,000 x 4
##    rhbmm .pred_class .pred_no .pred_yes
##    <fct> <fct>          <dbl>     <dbl>
##  1 no    no           0.613       0.387
##  2 no    no           0.655       0.345
##  3 no    no           0.623       0.377
##  4 yes   yes          0.00958     0.990
##  5 yes   yes          0.164       0.836
##  6 yes   yes          0.386       0.614
##  7 yes   no           0.611       0.389
##  8 no    no           0.609       0.391
##  9 yes   yes          0.144       0.856
## 10 yes   yes          0.349       0.651
## # i 9,990 more rows
```

Q16

Q16(a)

```r
cm1 <- merry_pred %>%
  conf_mat(
    .pred_class,
    truth = rhbmm
  )
cm1
```

```
##           Truth
## Prediction   no  yes
##        no  2981 1973
##        yes 1344 3702
```

```r
merry_pred %>% accuracy(
  .pred_class,
  truth = rhbmm
)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.668
```

Q16(b)

```
sens1 <- tidy(cm1)[4,2] / (tidy(cm1)[4,2] + tidy(cm1)[3,2])
sens1
```

```
##       value
## 1 0.6523348
```

```
merry_pred %>% sens(
  .pred_class,
  truth = rhbmm,
  event_level="second"
)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 sens    binary         0.652
```

```
spec1 <- tidy(cm1)[1,2] / (tidy(cm1)[1,2] + tidy(cm1)[2,2])
spec1
```

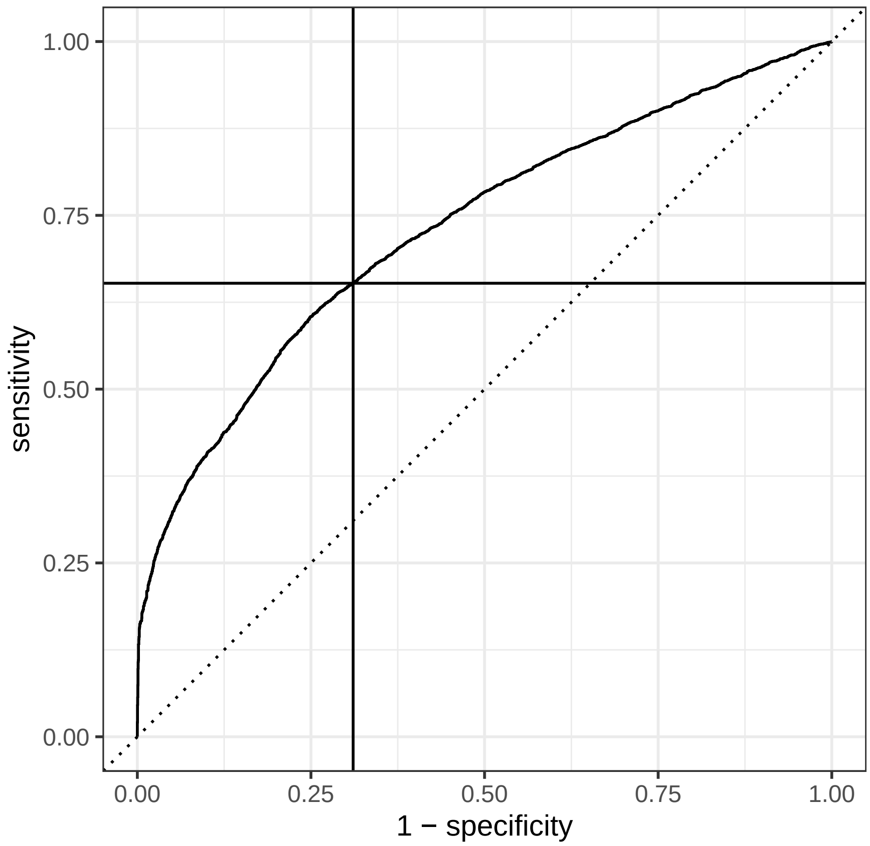
```
##       value
## 1 0.6892486
```

```
merry_pred %>% spec(
  .pred_class,
  truth = rhbmm,
  event_level="second"
)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 spec    binary         0.689
```

Q16(c)

```
merry_pred %>%
  roc_curve(
    .pred_yes,
    truth = rhbmm,
    event_level = "second"
  ) %>%
  autoplot()+
  geom_vline(xintercept=as.numeric(1-spec1))+
  geom_hline(yintercept=as.numeric(sens1))
```



Q16(d)

```
merry_pred %>%
  roc_auc(
    .pred_yes,
    truth = rhbmm,
    event_level = "second"
  )
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.730
```

Q17

```
new_archer <- tibble(
  accuracy=112/116,
  age="youth",
  jail=as.logical("FALSE"),
  home="forest",
```

```r
  dress="green"
)
new_archer
```

```
## # A tibble: 1 x 5
##   accuracy age   jail  home   dress
##      <dbl> <chr> <lgl> <chr>  <chr>
## 1    0.966 youth FALSE forest green
```

```r
pprob<-predict(m6, new_archer, type="prob")
pprob
```

```
## # A tibble: 1 x 2
##   .pred_no .pred_yes
##      <dbl>     <dbl>
## 1  0.00845     0.992
```

```r
predict(m6, new_archer, type="class")
```

```
## # A tibble: 1 x 1
##   .pred_class
##   <fct>
## 1 yes
```

We predict that the new archer is Merry Man, with probability 99.2%.