MATHS 7107
Data Taming
Practical 10

Classification models

Preliminaries

- Set up a project in RStudio
- Now load the packages
  - tidyverse
  - tidymodels
  - palmerpenguins
  - harrypotter (which we will use for colouring our graphs)
- Then load the dataset penguins.

Source: easy-peasy.ai

1 Part 1

First we will start by looking at how to measure a model using yardstick. We will fit a regression model, and also a classification model to the penguins dataset and then have a look at assessing them.

1.1 What are we modelling?

First let's look at the data that we are going to model. First a linear model.

Question:

1. Make a scatterplot of body_mass_g against bill_length_mm. Does it look like there is a linear relationship?

Second, we'll fit a logistic model for the categorical response variable sex against body_mass_g. But in order to make a scatterplot, we'll need to recode the sex variable to integers. We expect males to be heavier, so we'll make them a 1 and the females a 0. The code to do this is a bit annoying, but let's get on with it.

```r
p1<-mutate(penguins,
           sex01=as.integer(as.character((
             fct_recode(penguins$sex,`0`="female", `1`="male"))
             )),
           .after=bill_length_mm)
p1
```

```
## # A tibble: 344 x 9
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  1 Adelie  Torgersen           39.1          18.7               181        3750
##  2 Adelie  Torgersen           39.5          17.4               186        3800
##  3 Adelie  Torgersen           40.3          18                 195        3250
##  4 Adelie  Torgersen           NA            NA                  NA          NA
##  5 Adelie  Torgersen           36.7          19.3               193        3450
##  6 Adelie  Torgersen           39.3          20.6               190        3650
##  7 Adelie  Torgersen           38.9          17.8               181        3625
##  8 Adelie  Torgersen           39.2          19.6               195        4675
##  9 Adelie  Torgersen           34.1          18.1               193        3475
## 10 Adelie  Torgersen           42            20.2               190        4250
## # i 334 more rows
## # i 3 more variables: sex <fct>, year <int>, sex01 <int>
```

Question:

2. Make a scatterplot of `body_mass_g` against `sex01`. Does it look like we may be able to predict sex with body mass?

1.2 Create the models

```r
lm1<-lm(flipper_length_mm ~ body_mass_g, penguins)
summary(lm1)
```

```
##
## Call:
## lm(formula = flipper_length_mm ~ body_mass_g, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.7626  -4.9138   0.9891   5.1166  16.6392
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.367e+02  1.997e+00   68.47   <2e-16 ***
## body_mass_g 1.528e-02  4.668e-04   32.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.913 on 340 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.7583
## F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

```r
logreg_spec <- logistic_reg( mode = "classification" )

logreg1 <- logreg_spec %>%
  set_engine( "glm" ) %>%
  fit( sex ~ body_mass_g, data = penguins )

summary(logreg1$fit)
```

```
##
## Call:
## stats::glm(formula = sex ~ body_mass_g, family = stats::binomial,
##     data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.1625416  0.7243906  -7.127 1.03e-12 ***
## body_mass_g  0.0012398  0.0001727   7.177 7.10e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 461.61  on 332  degrees of freedom
## Residual deviance: 396.64  on 331  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 400.64
##
## Number of Fisher Scoring iterations: 4
```

Questions:

3. For model 1, what is the response variable and what are the predictors?

4. For model 2, what is the response variable and what are the predictors?

1.3 Getting predictions

For yardstick, we will need predicted values, we obtain that using the `predict()` function. Here I will add a variety of predictions to the original dataset.

First we run this bit of code to make sure that we have the right tibble to put into our `bin_cols()` command:

```r
rename(as_tibble(predict(lm1, penguins)),.pred_reg=value)
```

```
## # A tibble: 344 x 1
##    .pred_reg
##        <dbl>
##  1      194.
##  2      195.
##  3      186.
##  4       NA
##  5      189.
##  6      192.
##  7      192.
```

```
##   8       208.
##   9       190.
## 10       202.
## # i 334 more rows
```

Yes, this looks good, and so we're ready to put all our predictions into a single tibble. (The `predict()` function automatically produces a tibble for the classification model, so we don't have to do that ourselves.)

```
penguins_pred <-
  penguins %>%
  bind_cols(
    rename(as_tibble(predict(lm1, penguins)),.pred_reg=value),
    predict(logreg1, penguins),
    predict(logreg1, penguins,
            type = "prob"),
  )

select(penguins_pred,sex, flipper_length_mm, .pred_reg, .pred_class, .pred_female, .pred_male)
```

```
## # A tibble: 344 x 6
##    sex    flipper_length_mm .pred_reg .pred_class .pred_female .pred_male
##    <fct>              <int>     <dbl> <fct>              <dbl>      <dbl>
##  1 male                 181      194. female             0.626      0.374
##  2 female               186      195. female             0.611      0.389
##  3 female               195      186. female             0.756      0.244
##  4 <NA>                  NA       NA  <NA>              NA         NA
##  5 female               193      189. female             0.708      0.292
##  6 male                 190      192. female             0.654      0.346
##  7 female               181      192. female             0.661      0.339
##  8 male                 195      208. male               0.347      0.653
##  9 <NA>                 193      190. female             0.701      0.299
## 10 <NA>                 190      202. male               0.473      0.527
## # i 334 more rows
```

Questions:

5. What is the predicted flipper length for the first penguin?
6. What is the predicted probability of being male for the first penguin?

1.4 Quantitative metrics

```
penguins_pred %>% metrics(.pred_reg, truth=flipper_length_mm)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       6.89
## 2 rsq     standard       0.759
## 3 mae     standard       5.61
```

Question:

7. What is the Root Mean Squared Error for our predictions?

1.5 Categorical metrics

For most of the metrics, we will use the hard classification for the categorical variable as given by `.pred_class`.

We can get the confusion matrix:

```
penguins_pred %>%
  conf_mat(
    .pred_class,
    truth = sex
  )
```

```
##             Truth
## Prediction female male
##     female    109   74
##     male       56   94
```

Question:

8. How many of the females were incorrectly predicted as male?

We can get the sensitivity as follows:

```
penguins_pred %>%
  sens(
    .pred_class,
    truth = sex
  )
```

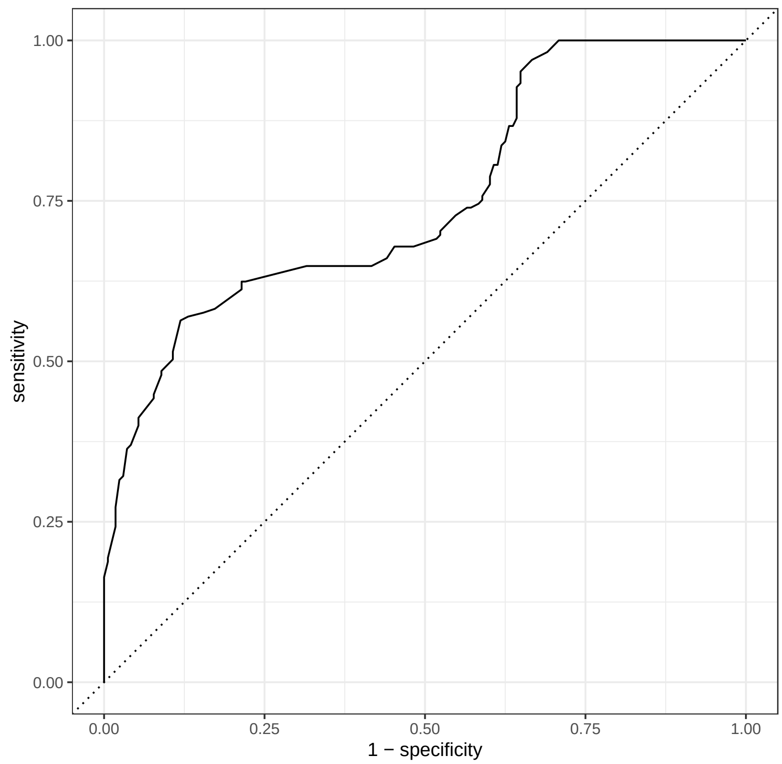
```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 sens    binary         0.661
```

Question:

9. What is the specificity and accuracy?

We can plot the ROC curve using `autoplot()`

```
penguins_pred %>%
  roc_curve(
    .pred_female,
    truth = sex
  ) %>%
  autoplot()
```



Question:

10. What is the AUC for our classification model?

2 Part 2

Now we are going to split the data into a training set and a testing set.

- We will "train" the model on the training set
- And then we will "test" the model on the testing set.

Before you go on, answer this question:

Do you expect the metrics will be better or worse than in Part 1?

2.1 Load and split the data

Back to the penguins — why would you not? First we are going to split our dataset into test data (to save for the very end) and training data.

```
set.seed(2021)
penguin_split <- initial_split(penguins)
penguin_split
```

```
## <Training/Testing/Total>
## <258/86/344>
```

```r
penguins_train <- training(penguin_split)
penguins_test <- testing(penguin_split)
```

Question:

11. How many penguins are in the test dataset?

2.2 Fit the model to the training set

Now we go through the same procedure of fitting models, only this time to the training set:

```r
# training regression model

lm_train<-lm(flipper_length_mm ~ body_mass_g, penguins_train)
summary(lm_train)
```

```
##
## Call:
## lm(formula = flipper_length_mm ~ body_mass_g, data = penguins_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.5884  -4.8446   0.9238   5.3456  14.2796
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.364e+02  2.234e+00   61.03   <2e-16 ***
## body_mass_g 1.532e-02  5.194e-04   29.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.738 on 255 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7733, Adjusted R-squared:  0.7724
## F-statistic: 869.7 on 1 and 255 DF,  p-value: < 2.2e-16
```

```r
# training classification model

logreg_train <- logreg_spec %>%
  set_engine( "glm" ) %>%
  fit( sex ~ body_mass_g, data = penguins_train )

summary(logreg_train$fit)
```

```
##
## Call:
## stats::glm(formula = sex ~ body_mass_g, family = stats::binomial,
##     data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.2344569  0.8284602  -6.318 2.64e-10 ***
## body_mass_g  0.0012289  0.0001952   6.295 3.08e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 350.54  on 252  degrees of freedom
## Residual deviance: 300.90  on 251  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 304.9
##
## Number of Fisher Scoring iterations: 4
```

2.3 Predict on the test set

```
penguins_predt <-
  penguins_test %>%
  bind_cols(
    rename(as_tibble(predict(lm_train, penguins_test)), .pred_reg=value),
    predict(logreg_train, penguins_test),
    predict(logreg_train, penguins_test,
            type = "prob"),
  )
penguins_predt
```

```
## # A tibble: 86 x 12
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  1 Adelie  Torgersen           39.5          17.4               186        3800
##  2 Adelie  Torgersen           40.3          18                 195        3250
##  3 Adelie  Torgersen           NA            NA                  NA          NA
##  4 Adelie  Torgersen           42            20.2               190        4250
##  5 Adelie  Torgersen           37.8          17.1               186        3300
##  6 Adelie  Torgersen           37.8          17.3               180        3700
##  7 Adelie  Torgersen           42.5          20.7               197        4500
##  8 Adelie  Biscoe              37.8          18.3               174        3400
##  9 Adelie  Biscoe              38.8          17.2               180        3800
## 10 Adelie  Dream               37.2          18.1               178        3900
## # i 76 more rows
## # i 6 more variables: sex <fct>, year <int>, .pred_reg <dbl>,
## #   .pred_class <fct>, .pred_female <dbl>, .pred_male <dbl>
```

2.4 Evaluate the models

Questions:

12. Now we'll let you repeat the calculations in Part 1 for the metrics for our predictions on the testing set.
13. So which predictions were better: in Part 1 or Part 2?
14. Why?