Homework11

Dima Mikhaylov

11/22/2021

Question 1: penguins data set, the questions below should be answered using the training set.

```
library(palmerpenguins)
Data<-penguins
##remove penguins with gender missing
Data<-Data[complete.cases(Data[ , 7]),-c(2,8)]
##80-20 split
set.seed(1)
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
train<-Data[sample, ]
test<-Data[-sample, ]
head(train)
```

```
## # A tibble: 6 × 6
##   species   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex
##   <fct>              <dbl>         <dbl>             <int>       <int> <fct>
## 1 Chinstrap           50.2          18.8               202        3800 male
## 2 Gentoo              50.2          14.3               218        5700 male
## 3 Adelie              38.1          17.6               187        3425 female
## 4 Chinstrap           51            18.8               203        4100 male
## 5 Chinstrap           52.7          19.8               197        3725 male
## 6 Gentoo              49.6          16                 225        5700 male
```

Check if target is binary, i.e factor

```
class(train$sex)
```

```
## [1] "factor"
```

Convert binary sex to 1 for ‘male’ and 0 for ‘female’

```
train$sex <- ifelse(train$sex=="male",1,0)
train$sex <- factor(train$sex)
head(train)
```

```
## # A tibble: 6 × 6
##   species   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex
##   <fct>              <dbl>         <dbl>             <int>       <int> <fct>
## 1 Chinstrap           50.2          18.8               202        3800 1
## 2 Gentoo              50.2          14.3               218        5700 1
## 3 Adelie              38.1          17.6               187        3425 0
## 4 Chinstrap           51            18.8               203        4100 1
## 5 Chinstrap           52.7          19.8               197        3725 1
## 6 Gentoo              49.6          16                 225        5700 1
```

Check if target is still coded as factor:

```
class(train$sex)
```

```
## [1] "factor"
```

a. Create some data visualizations to explore the relationship between the various body measurements and the gender of penguins. Be sure to briefly comment on your data visualizations.

Boxplots of numeric predictors by binary sex: `bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, and `body_mass_g`.

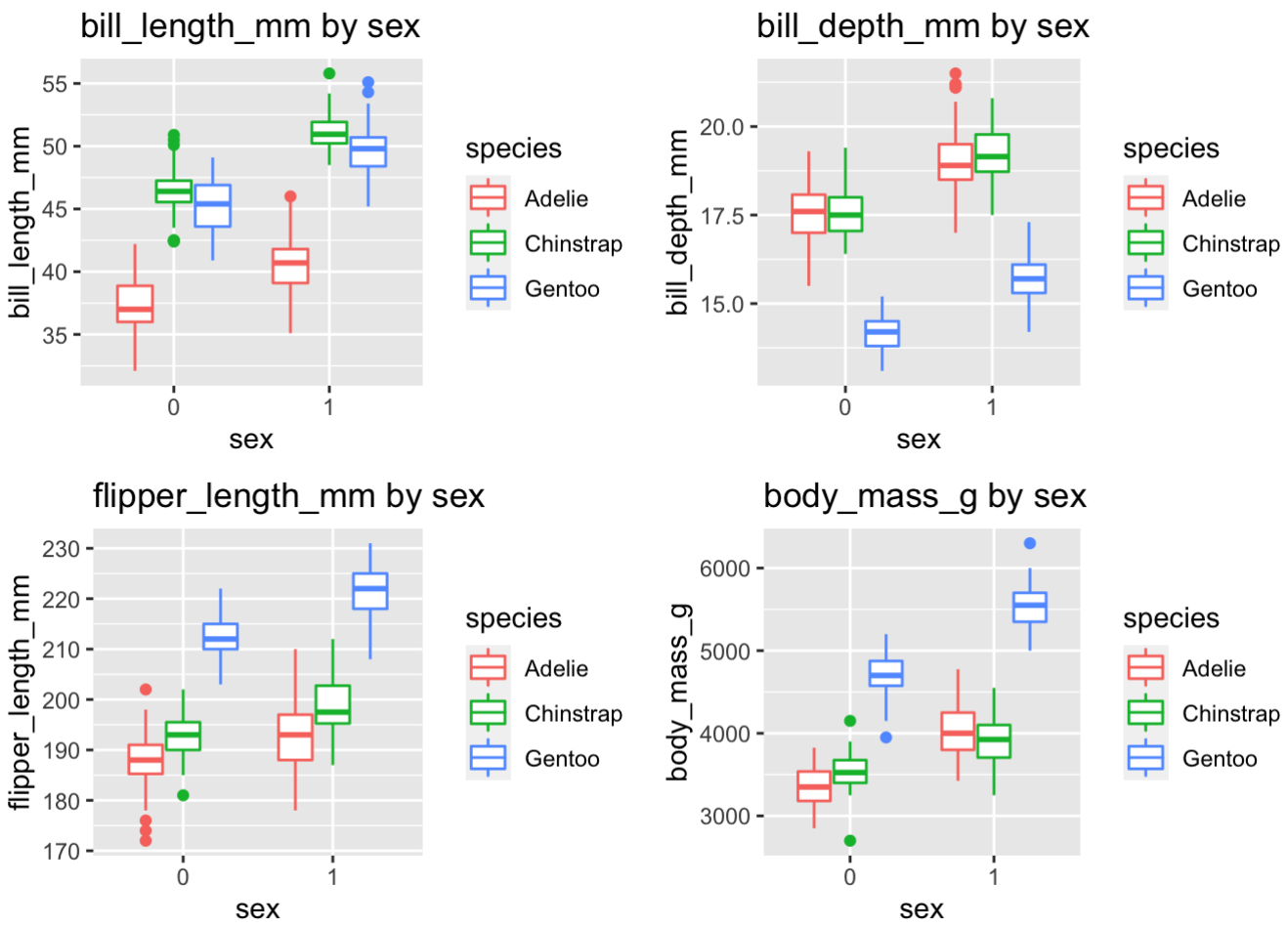
```r
# Create 4 boxplot objects for respective numeric predictors:
library(ggplot2)
bp1<-ggplot(train, aes(x=sex, y=bill_length_mm, color=species))+
geom_boxplot()+
labs(x="sex", y="bill_length_mm", title="bill_length_mm by sex")

bp2<-ggplot(train, aes(x=sex, y=bill_depth_mm, color=species))+
geom_boxplot()+
labs(x="sex", y="bill_depth_mm", title="bill_depth_mm by sex")

bp3<-ggplot(train, aes(x=sex, y=flipper_length_mm, color=species))+
geom_boxplot()+
labs(x="sex", y="flipper_length_mm", title="flipper_length_mm by sex")

bp4<-ggplot(train, aes(x=sex, y=body_mass_g, color=species))+
geom_boxplot()+
labs(x="sex", y="body_mass_g", title="body_mass_g by sex")

## Produce the 4 boxplots in a 2 by 2 matrix
library(gridExtra)
library(grid)
grid.arrange(bp1, bp2, bp3, bp4, ncol = 2, nrow = 2)
```



Conclusion: from EDA boxplots above it seems that males (coded as 1) tend to be larger and heavier than females (coded as 0). Therefore, it seems warranted that sex can be used as a target variable for binary classification. Noteworthy is that there is a significant intergroup variability depending on the `species` type. For example, Gentoo females tend to have larger `body_mass_g` and `flipper_length_mm` than Chinstrap and Adelie males.

b. Use R to fit the logistic regression model. Based on the results of the Wald tests for the individual coefficients, which predictor(s) appears to be insignificant in the model?

```
full_model<-glm(sex ~ ., family="binomial", data=train)
summary(full_model)
```

```
##
## Call:
## glm(formula = sex ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.85959  -0.10720   0.00061   0.06817   3.02072
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -94.355394  17.638204  -5.349 8.82e-08 ***
## speciesChinstrap  -10.608813   2.634752  -4.026 5.66e-05 ***
## speciesGentoo     -10.384568   3.565641  -2.912  0.00359 **
## bill_length_mm      1.025200   0.238593   4.297 1.73e-05 ***
## bill_depth_mm       2.287977   0.516595   4.429 9.47e-06 ***
## flipper_length_mm  -0.088318   0.065040  -1.358  0.17450
## body_mass_g         0.008094   0.001662   4.871 1.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 368.619  on 265  degrees of freedom
## Residual deviance:  68.297  on 259  degrees of freedom
## AIC: 82.297
##
## Number of Fisher Scoring iterations: 8
```

Based on the summary results, `flipper_length_mm` has large p-value associated with Wald test and thus could be seen as insignificant.

c. Based on your answer in part 1b, drop the predictor(s) and refit the logistic regression. Write out the estimated logistic regression equation. If you did not drop any predictor, write out the logistic

regression equation from part 1b.

```
reduced_model<-glm(sex ~ species + bill_length_mm + bill_depth_mm + body_mass_g, family="binomial", data=train)
summary(reduced_model)
```

```
##
## Call:
## glm(formula = sex ~ species + bill_length_mm + bill_depth_mm +
##     body_mass_g, family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.52269  -0.11388   0.00063   0.06524   3.01858
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.032e+02  1.706e+01  -6.051 1.44e-09 ***
## speciesChinstrap -1.042e+01  2.544e+00  -4.096 4.20e-05 ***
## speciesGentoo    -1.238e+01  3.383e+00  -3.661 0.000251 ***
## bill_length_mm    9.513e-01  2.210e-01   4.303 1.68e-05 ***
## bill_depth_mm     2.099e+00  4.684e-01   4.481 7.41e-06 ***
## body_mass_g       7.714e-03  1.625e-03   4.746 2.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 368.619  on 265  degrees of freedom
## Residual deviance:  70.172  on 260  degrees of freedom
## AIC: 82.172
##
## Number of Fisher Scoring iterations: 8
```

Once `flipper_length_mm` was dropped all other coefficients appear statistically significant, below is the logistic regression equation:

log(p/(1-p)) = -103.2 - 10.42 * I1 - 12.38 * I2 + 0.9513 * bill_length_mm + 2.099 * bill_depth_mm + 0.007714 * body_mass_g

where: I1 = 1 for Chinstrap and I2 = 1 for Gentoo species.

d. Based on your estimated logistic regression equation in part 1c, how would you generalize the relationship between some of the body measurement predictors and the (log) odds of a penguin being male?

Estimated log odds of being a male increase with all body measurements differently, i.e the log odds increase on average by 0.9513 with increase in `bill_length_mm`, by 2.099 with increase in `bill_depth_mm`, and by 0.007714 with increase in `body_mass_g` while controlling for all other predictors.

e. Interpret the estimated coefficient for bill length contextually.

Coefficient for bill_length_mm is 0.9513 and, therefore, for every millimeter of bill length, the estimated log odds of a penguin being a male increases by 0.9513, while controlling for all other predictors. In other words, for an additional millimeter of observed bill length, the estimated odds of being a male penguin should be multiplied by a factor of exp(0.9513)=2.589073.

f. Consider a Gentoo penguin with bill length of 49 mm, bill depth of 15 mm, flipper length of 220 mm, and body mass of 5700 g. What are the log odds, odds, and probability that this penguin is male?

```
# Create a vector with new data using species + bill_length_mm + bill_depth_mm + body_mass_g,
newdata <- data.frame(species="Gentoo", bill_length_mm=49, bill_depth_mm=15, flipper_length_mm=220, body_mass_g=5
700)

# Make prediction for log odds
predict(full_model, newdata)
```

```
##        1
## 6.519736
```

```
## Convert to odds
odds<-exp(predict(full_model,newdata))
odds
```

```
##        1
## 678.3991
```

```
## Convert odds to probability
prob<-odds/(1+odds)
prob
```

```
##         1
## 0.9985281
```

g. Conduct a relevant hypothesis test to assess if the logistic regression in part 1c is a useful model. Be sure to write out the null and alternative hypotheses, report the value of the test statistic, and write a relevant conclusion.

Testing if the full model (orginal predictors) is a usefull model. Hypothesis:

* H0: B1=B2=B3=B4=B5=0

* Ha: at least one of the coefficients in H0 is not zero

To check this, compute delta G2:

```
deltaG2<-full_model$null.deviance - full_model$deviance
cat("Delta G2 statistic is", deltaG2)
```

```
## Delta G2 statistic is 300.3223
```

Test the result for significance:

```
cat("Corresponding p-value is", 1-pchisq(deltaG2,5))
```

```
## Corresponding p-value is 0
```

Conclusion: reject H0 and accept Ha because p-value associated with computed dG2 statistics is very small. In other words, the data supports the claim that full model is useful when compared to naive intercept only model.

Question 2: follow-up flu stot study of 159 elderly clients.

a. Interpret the estimated coefficient for x3, gender, in context.

B3 coefficient associated with x3, `gender`, is equal to 0.43397. Thus the estimated log odds of taking a flu shot for males (coded as 1) is 0.43397 higher than for females (coded as 0), when controlling for all other predictors.

b. Conduct the Wald test for B3. State the null and alternative hypotheses, calculate the test statistic, and make a conclusion in context.

Testing if B3 is (statistically) significantly different from zero. Hypothesis:

*** H0: B3=0**

*** Ha: B3 is NOT 0, two-tailed test**

```
B3_hat <- 0.43397
B3_se <- 0.52179
W <- B3_hat / B3_se
cat("Wald test statistic is", W)
```

```
## Wald test statistic is 0.8316947
```

```
2 * pnorm(W, lower.tail = FALSE)
```

```
## [1] 0.4055813
```

Conclusion: W stats is 0.8316947 and is associated with large p-value of 0.4055813. Therefore, fail to reject H0. In other words, we can drop x3, `gender` , while leaving the other predictors in the model.

c. Calculate a 95% confidence interval for B3, and interpret the interval in context.

```
z <- 1.96 # Z value for a/2 of 0.25
multiplier <- z*B3_se
cat("CI for B3 [", B3_hat-multiplier, ",",  B3_hat+multiplier, "]")
```

```
## CI for B3 [ -0.5887384 , 1.456678 ]
```

Conclusion: the interval includes 0 and thus one should fail to reject H0 that states that B3=0. In other words, the data does not support claim that B3 is different from zero becuase even 95% CI includes 0.

d. Comment on whether your conclusions from parts 2b and 2c are consistent.

Yes, I think the conclusions are consistent, because in both 2b and 2c one failes to reject H0. In other words, the slope B3 is likely 0.

e. Suppose you want to drop the coefficients for age and gender, B1 and B3. A logistic regression model for just awareness was fitted, and the output is shown below. Carry out the appropriate hypothesis test to see if the coefficients for age and gender can be dropped.

* H0: B1 = B3 = 0

* Ha: at least one of the slopes, either B1 or B3, is not 0

```
full_deviance <- 105.09
reduced_deviance <- 113.20
deltaG2_reduced<-reduced_deviance-full_deviance
cat("Reduced delta G2 statistic is", deltaG2_reduced)
```

```
## Reduced delta G2 statistic is 8.11
```

```
cat("Corresponding p-value is", 1-pchisq(deltaG2_reduced,2)) # df = number of vars dropped
```

```
## Corresponding p-value is 0.01733548
```

Conclusion: reject H0 and accept Ha because p-value associated with computed reduced dG2 statistics is very small. In other words, the data supports the claim that full model is useful when compared to reduced model, thus we should have not dropped `age` and `gender`.

f. Based on your conclusion in question 2e, what are the estimated odds of a client receiving the flu shot if the client is 70 years old, has a health awareness rating of 65, and is male? What is the estimated probability of this client receiving the flu shot?

Estimated odds of (Y=1) can be computed as (p/(1-p)) = exp(Intercept + B1 * `age` + B2 * `aware` + B3 * `gender`)

```
shot_odds <- exp(-1.17716 + 0.07279 * 70 - 0.09899 * 65 + 0.43397 * 1)
shot_odds
```

```
## [1] 0.1246507
```

```
# check propability:
shot_odds/(1+shot_odds)
```

```
## [1] 0.110835
```