Stat 6021: Homework Set 11 and 12

Tom Lever

11/28/22

Note that this is a homework that combines Modules 11 and 12, and is due Dec 5. You can work on questions $1a$ to $1g$ and 2 after Module 11.

1. For this question, we will revisit the `penguins` data set from the `palmerpenguins` package. The data set contains information regarding measurements of adult penguins near Palmer Station, Antarctica. We will focus on using the four measurement variables (bill length, bill depth, flipper length, and body mass) to model the gender of the penguins. Since there are three species involved, we also want to control for species in the logistic regression. We will not consider the island and year in this logistic regression.

   When you read the data in, notice that there are a number of penguins with missing values for gender. Remove these observations from the data frame. Also remove columns 2 and 8 since we are not considering island and year in this logistic regression. Then, randomly split your data into a training and test set ($80 - 20$ split respectively). For reproducibility, use `set.seed(1)` while performing the split. You can run the following block of code to carry out the needed steps.

```
library(palmerpenguins)
data_set <- penguins
data_set <- data_set[complete.cases(data_set[ , 7]), -c(2, 8)]
set.seed(1)
number_of_observations <- nrow(data_set)
number_of_observations
```

```
## [1] 333
```

```
head(data_set, n = 3)
```

```
## # A tibble: 3 x 6
##   species bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex
##   <fct>            <dbl>         <dbl>             <int>       <int> <fct>
## 1 Adelie            39.1          18.7               181        3750 male
## 2 Adelie            39.5          17.4               186        3800 female
## 3 Adelie            40.3          18                 195        3250 female
```

```
indices_of_training_observations <- sample.int(number_of_observations, floor(0.8 * number_of_observ
training_data_set <- data_set[indices_of_training_observations, ]
testing_data_set <- data_set[-indices_of_training_observations, ]
```

   The questions below should be answered using the training data set.

   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.

   b) Use R to fit the logistic regression model (first order only involving the 4 measurement variables and species). Based on the results of the Wald tests for the individual coefficients, which predictor(s) appear to be insignificant in the model?

```
generalized_linear_model <- glm(sex ~ bill_length_mm + bill_depth_mm + flipper_length_mm + body
summary(generalized_linear_model)
```

```
##
## Call:
## glm(formula = sex ~ bill_length_mm + bill_depth_mm + flipper_length_mm +
##     body_mass_g + species, family = "binomial", data = training_data_set)
##
## 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 ***
## 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 ***
## speciesChinstrap  -10.608813   2.634752  -4.026 5.66e-05 ***
## speciesGentoo     -10.384568   3.565641  -2.912  0.00359 **
## ---
## 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
```

The Wald statistic for testing the significance of $flipper_length_m m$ $Z_0 = -1.358$ with a corresponding p value $p = 0.175$. $flipper_length_m m$ appears to be insignificant in the logistic regression model in the context of all the predictors in the model. We can drop $flipper_length_m m$ from the logistic regression model while leaving the other predictors in the model.

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

```
generalized_linear_model <- glm(sex ~ bill_length_mm + bill_depth_mm + body_mass_g + species,
generalized_linear_model
```

```
##
## Call:  glm(formula = sex ~ bill_length_mm + bill_depth_mm + body_mass_g +
##     species, family = "binomial", data = training_data_set)
##
## Coefficients:
##      (Intercept)    bill_length_mm     bill_depth_mm       body_mass_g
##       -1.032e+02         9.513e-01         2.099e+00         7.714e-03
## speciesChinstrap     speciesGentoo
##       -1.042e+01        -1.238e+01
##
## Degrees of Freedom: 265 Total (i.e. Null);  260 Residual
```

```
## Null Deviance:       368.6
## Residual Deviance: 70.17     AIC: 82.17
```

$$
\begin{aligned}
ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) =& \beta_0 \\
&+\beta_{bill\ length}(bill\ length) \\
&+\beta_{bill\ depth}(bill\ depth) \\
&+\beta_{body\ mass}(body\ mass) \\
&+\beta_{Chinstrap} I_{Chinstrap} \\
&+\beta_{Gentoo} I_{Gentoo} \\
ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) =& (-103.2) \\
&+(0.951)(bill\ length) \\
&+(2.099)(bill\ depth) \\
&+(0.00714)(body\ mass) \\
&+(-10.42) I_{Chinstrap} \\
&+(-12.38) I_{Gentoo}
\end{aligned}
$$