Hw11

Hyun Suk (Max) Ryoo (hr2ee)

11/22/2021

Set up

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(palmerpenguins)
```

```
## Warning: package 'palmerpenguins' was built under R version 4.0.2
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.0.2
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

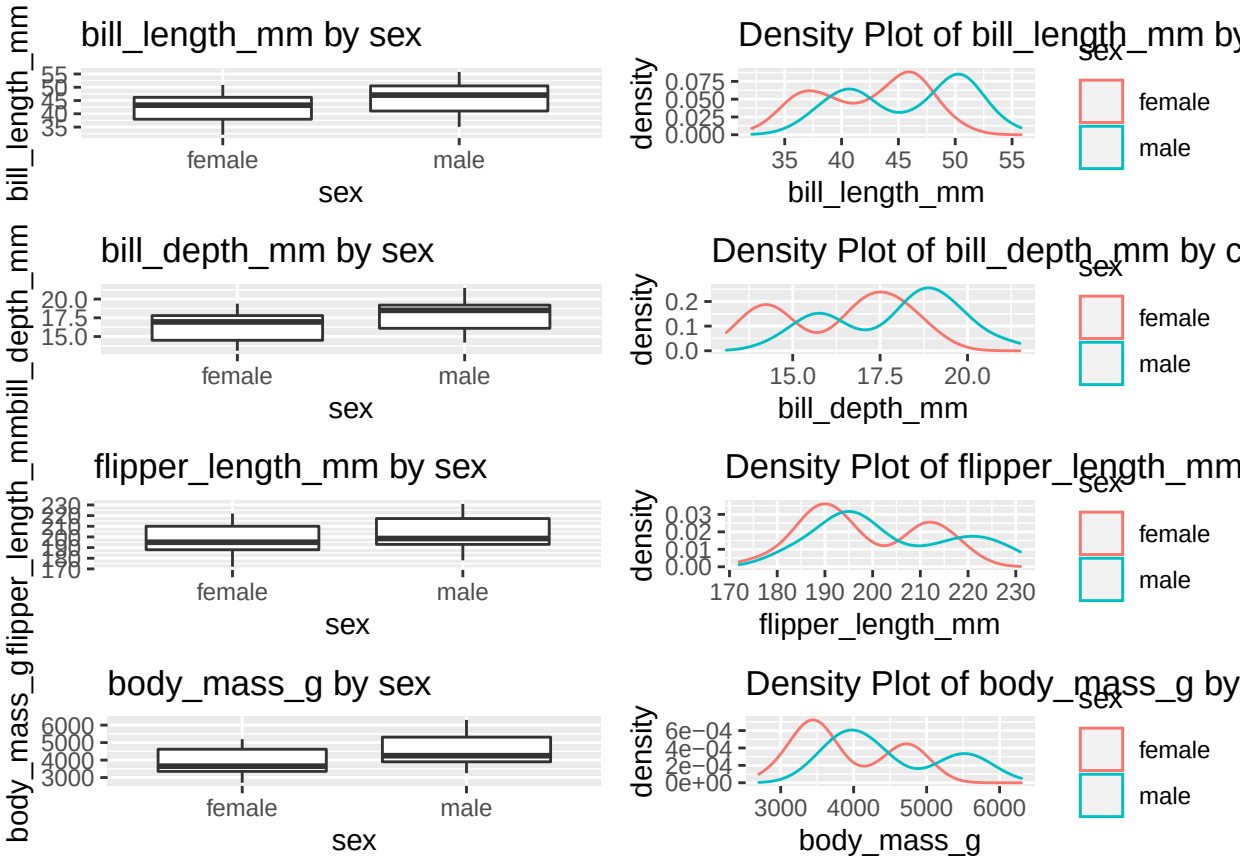
```
Data<-penguins
##remove penguins with gender missing
Data<-Data[complete.cases(Data[ , 7]),-c(2,8)]
##80-20 split
set.seed(1)
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
```

```
train<-Data[sample, ]
test<-Data[-sample, ]
head(train)
```

```
## # A tibble: 6 x 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
```

Q1 - A

```
boxplot_pen <- function(data, x, y) {
  return(ggplot(data, aes_string(x=x, y=y)) +
    geom_boxplot() +
    labs(x=x, y=y, title=paste(y, "by", x)))
}
density_pen <- function(data, class, field) {
  return(ggplot(data, aes_string(x=field, color=class)) +
    geom_density() +
    labs(title=paste("Density Plot of", field, "by", "class")))
}
bp1 <- boxplot_pen(train, "sex", "bill_length_mm")
dp1 <- density_pen(train, "sex", "bill_length_mm")
bp2 <- boxplot_pen(train, "sex", "bill_depth_mm")
dp2 <- density_pen(train, "sex", "bill_depth_mm")
bp3 <- boxplot_pen(train, "sex", "flipper_length_mm")
dp3 <- density_pen(train, "sex", "flipper_length_mm")
bp4 <- boxplot_pen(train, "sex", "body_mass_g")
dp4 <- density_pen(train, "sex", "body_mass_g")
grid.arrange(bp1,dp1,bp2,dp2,bp3,dp3,bp4,dp4, ncol = 2, nrow = 4)
```



When looking at the box plots for all body measurements the male gender had higher values and medians. Since all body measurements are of different mesaurements and scales, it is hard to distinguished which had the most significant difference, we can observe that for all four body measurments (bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g) were higher for males.

From the density plots we can see that the distribution of all measurments are slightly shifted. All the density plots show a similar shape for females and males, but males' values are shifted/transformed to the right a little, which indiciates that the males had a higher measurments for all the body measurments.

Q1 - B

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

```
##
## 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
```

From the Z values given and p-value, we can see that the predictor flipper_length_mm is not significant. To be exact the p-value of 0.17450 is above the threshold for significance.

Q1 - C

Refitting the logisitc regression will have a summary output like such.

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

```
##
## Call:
## glm(formula = sex ~ . - flipper_length_mm, 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
```

The logistic regression equation is

$$log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -103.2 - 10.42I_1 - 12.38I_2 + 0.09513\text{bill\_length\_mm} + 2.099\text{bill\_depth\_mm} + 0.007714\text{body\_mass\_g}$$

Where $I_1 = 1$ for Chinstrap penguins while $I_2 = 1$ for Gentoo, and both values will be 0 for Adelie species.

Q1 - D

Given that the species is held constant, we can see that all body measurement coefficients have a positive value. Given the body measurments odds of a penguin being male will increase for unit increases of the body measurments.

Q1 - E

The coefficient for bill length is 0.151168. For an additional bill length incraese (on average), the estimated log odds of a penguin being male increases by 0.151168, while controlling the other variables of bill_depth, flipper_length, and body_mass_g.

Q1 - F

```
## make prediction for log odds
newdata <- data.frame(bill_length_mm=49, bill_depth_mm=15, flipper_length_mm=220, body_mass_g=5700, spe
print(predict(result, newdata))
```

```
##        1
## 6.462668
```

```
## Convert to ods
odds<-exp(predict(result,newdata))
print(odds)
```

```
##        1
## 640.7683
```

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

```
##         1
## 0.9984418
```

The estimated log odds for this penguin being male is 6.462668 . The corresponding odds is 640.7683, and the corresponding probability is 0.9984418 .

Q1 - G

$H_0 : \beta_1 = ... = \beta_5 = 0$

H_A : at least one of the cofficients in H_0 is not zero

We can first find ΔG^2.

```
deltag2 <- result$null.deviance - result$deviance
deltag2
```

```
## [1] 298.4472
```

```r
1 - pchisq(deltag2, 5)
```

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

The test statistic for $\Delta G^2 = 298.4472$ with a p-value of 0. So we reject the null hyptohesis. The data supports the claim that our model is useful, compared to the intercept only model.

Q2 - A

The estimated coefficient for x_3 is $\beta_3 = 0.43397$.

The estimated log odds of a client recieves a flu shot for males is 0.43397 higher than than for females when controlling the other variables of age and aware.

Q2 - B

The hypothesis for the Walds tests are as follows.

$$H_0 : \beta_3 = 0 H_A : \beta_3 \neq 0$$

The test-statsitc can be calculated with the following.

$$\begin{aligned} Z &= \frac{\hat{\beta}_3}{SE(\hat{\beta}_3)} \\ &= \frac{0.43397}{0.52179} \\ &= 0.8316947 \end{aligned}$$

With this test statistic we can find the corresponding p-value in R like such.

```r
2*(1-pnorm(0.8316947))
```

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

The corresponding p-value equates to a value of 0.4055813. Since this is a greater value than our threshold, we fail to reject the null hypothesis. Our data supports that gender is not a significant predictor in evaluating the probability of getting a flu shot when given a controlled age and awareness.

Q2 - C

We can utilize the confidence interval for logistic regression from the textbook (Equation 13.25), which is the following

$$\hat{\beta}_j - Z_{\frac{\alpha}{2}} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + Z_{\frac{\alpha}{2}} se(\hat{\beta}_j)$$

The confidence interval computation is shown below.

$$\begin{aligned} \hat{\beta}_3 - Z_{\frac{\alpha}{2}} se(\hat{\beta}_3) \leq \beta_3 \leq \hat{\beta}_3 + Z_{\frac{\alpha}{2}} se(\hat{\beta}_3) \\ 0.43397 - Z_{\frac{0.05}{2}} 0.52179 \leq \beta_j \leq 0.43397 + Z_{\frac{0.05}{2}} 0.52179 \\ 0.43397 - 1.959964 * 0.52179 \leq \beta_j \leq 0.43397 + 1.959964 * 0.52179 \\ -0.5887196 \leq \beta_j \leq 1.456678 \\ (-0.5887196, 1.45666) \end{aligned}$$

The qnorm function was used to calculate the $Z_{\frac{\alpha}{2}}$.

```r
qnorm(0.05/2) * -1
```

```
## [1] 1.959964
```

Thus the 95% confidence interval was (-0.5887196,1.45666). We are 95% confident that the true odds of a client recieving a flu shot for males is between $(e^{-0.5887196}, e^{1.45666}) \rightarrow (0.5550375, 4.291602)$ times the odds of a client recieving a flu shot for females.

Q2 - D

The conclusions of Q2-B and Q2-C are consistent since the interval contains 0 in the 95% confidence interval.

Q2 - E

For this hypothesis test the null and alternate hypothesis are as follows.

$$H_0 : \beta_1 = \beta_3 = 0 H_A : \text{at least one of the cofficients in } H_0 \text{ is not zero}$$

The test statistic is as follows.

$$\begin{aligned} \Delta G^2 &= \text{Res Dev}(Full) - \text{Res Dev}(Reduced) \\ &= 113.20 - 105.09 \\ &= 8.11 \end{aligned}$$

From this statsitc we can compute the p-value to be $1 - pchisq(8.11, 2) = 0.01733548$. Since this p-value is lower than our threshold value, we reject the null hypothesis meaning that we can not drop the predictors of age and gener.

Q2 - F

The full logistic model is as follows.

$$log\left(\frac{\pi}{1-\pi}\right) = -1.17716 + 0.07279age - 0.09899aware + 0.43397gender$$

Given that the client is 70 years old, with a health awareness rating of 65, and is male we can calculate the estimated probability.

$$\begin{aligned} log\left(\frac{\pi}{1-\pi}\right) &= -1.17716 + 0.07279age - 0.09899aware + 0.43397gender \\ &= -1.17716 + 0.07279 * (70) - 0.09899 * (65) + 0.43397 * (1) \\ &= -2.08224 \end{aligned}$$

Therefore, the odds will be computed to $e^{-2.08224} = 0.1246507$ and the corresponding probability is $\frac{odds}{1+odds} = \frac{0.1246507}{1+0.1246507} = 0.110835$