HW 11 & 12

Matt Scheffel

2022-12-05

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ROCR)
library(faraway)
library(palmerpenguins)
#install.packages("gridExtra")
library(gridExtra)
```

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

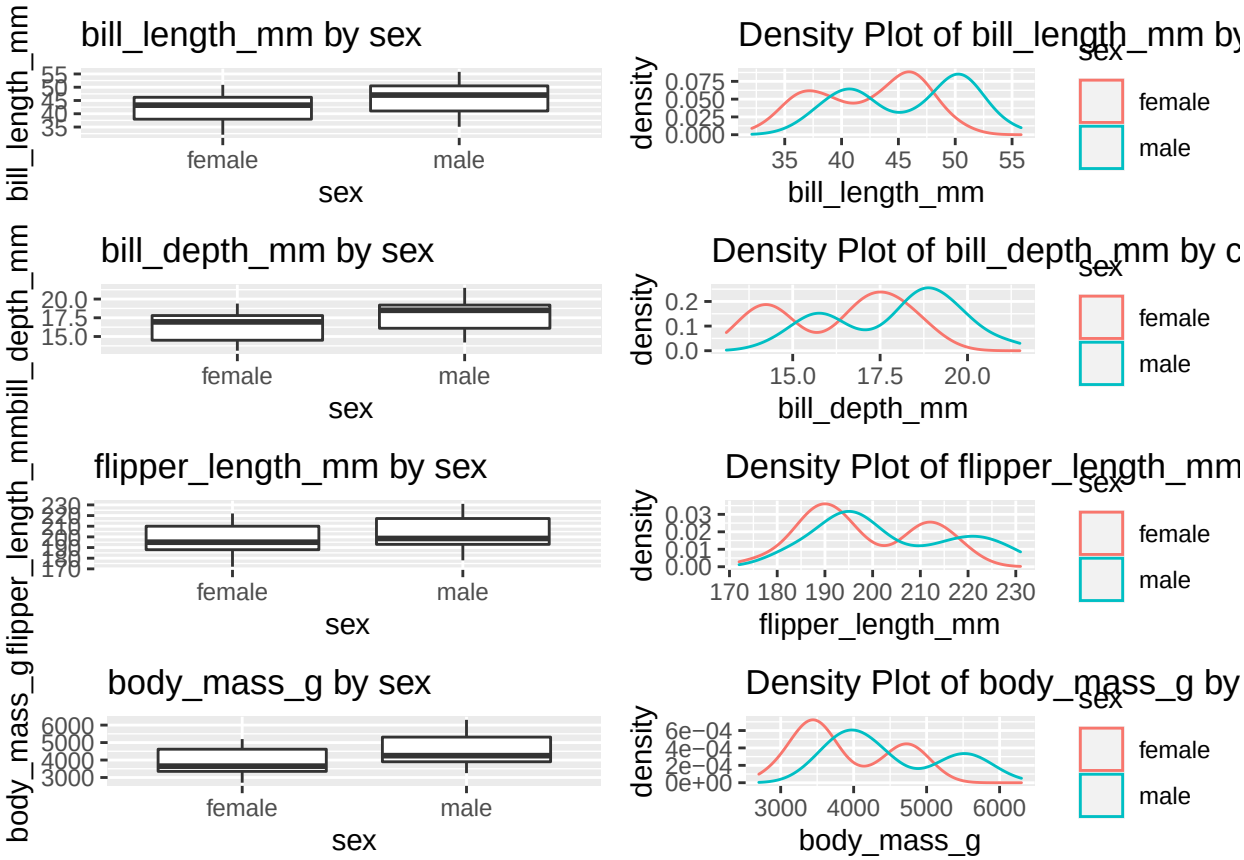
```
library(ggplot2)

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, ]
```

1A

```
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")))
}
box1 <- boxplot_pen(train, "sex", "bill_length_mm")
dens1 <- density_pen(train, "sex", "bill_length_mm")
box2 <- boxplot_pen(train, "sex", "bill_depth_mm")
dens2 <- density_pen(train, "sex", "bill_depth_mm")
box3 <- boxplot_pen(train, "sex", "flipper_length_mm")
dens3 <- density_pen(train, "sex", "flipper_length_mm")
box4 <- boxplot_pen(train, "sex", "body_mass_g")
dens4 <- density_pen(train, "sex", "body_mass_g")
grid.arrange(box1,dens1,box2,dens2,box3,dens3,box4,dens4, ncol = 2, nrow = 4)
```



Comments:

From our visualizations, we can see that males have higher body measurements values and medians for each variable. The density plots for males are also shifted further right, indicating their higher values.

1B

```
regression1<-glm(sex ~ ., family="binomial", data=train)
summary(regression1)
```

```
##
## 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- and p-values, we can see that Flipper Length is not significant and can be dropped. ($P > 0.05$)

1C

```
regression2<-glm(sex ~ . - flipper_length_mm, family="binomial", data=train)
summary(regression2)
```

```
##
## 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
```

Regression equation:

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

$I_1 = 1$ for Chinstrap penguins and $I_2 = 1$ for Gentoo penguins. Both of these values are zero for Adelie penguins.

1D

Holding the penguin species constant, the data shows all body measurement coefficients have a positive value. This demonstrates that with these body measurements, the (log) odds of a penguin being a male will increase as the body measurements of the penguin increase.

1E

The coefficient for bill length is 0.09513. Contextually, this means that on average for a bill length increase, the estimated (log) odds of a penguin being male increases by 0.09513, while other variables (bill depth, flipper length, body mass) are held constant.

1F

```
data2 <- data.frame(bill_length_mm=49, bill_depth_mm=15, flipper_length_mm=220, body_mass_g=5700, speci

print(predict(regression2, data2))
```

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

```
odds2<-exp(predict(regression2,data2))
print(odds2)
```

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

```
maleprob<-odds2/(1+odds2)
print(maleprob)
```

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

Log Odds:

1 6.462668

Odds:

1 640.7683

Probability:

1 0.9984418

1G

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

H_A : one or more of the cofficients in H_0 is a non-zero

ΔG^2:

```
changeg2 <- regression2$null.deviance - regression2$deviance
changeg2
```

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

```
1 - pchisq(changeg2, 5)
```

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

Test-statistic = 298.4472. P-value = 0.

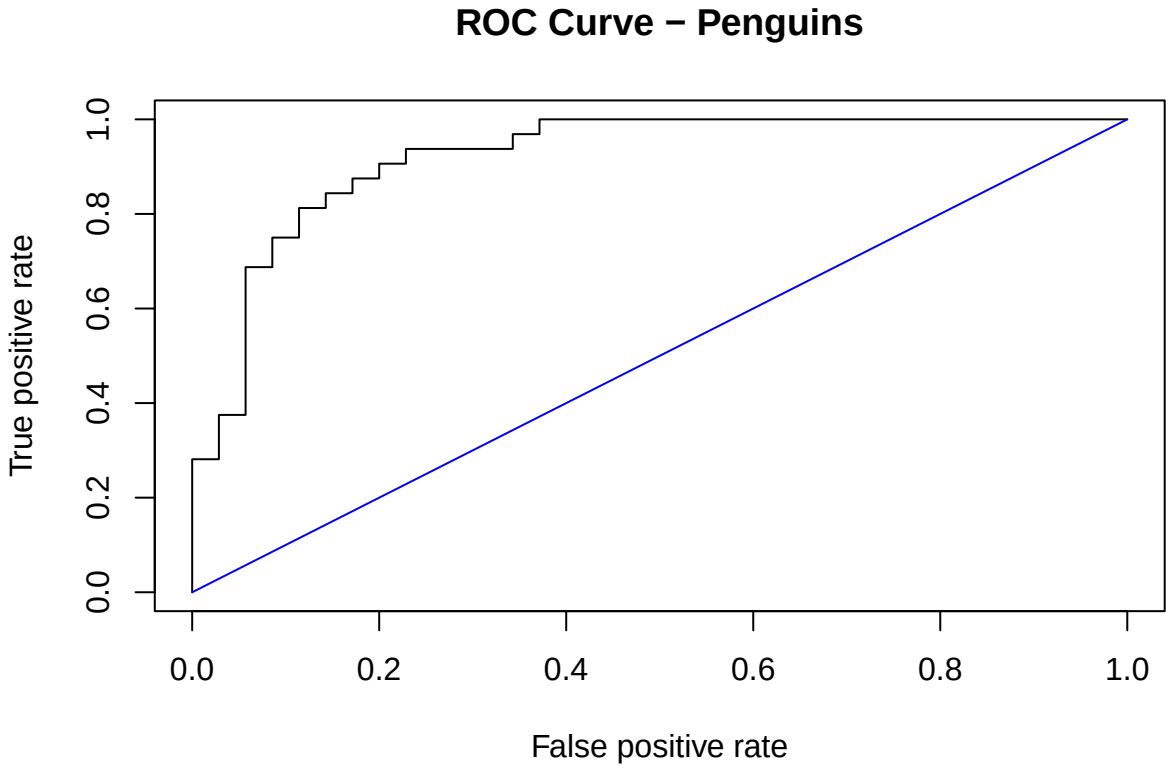
Conclusion: Thus, we reject the null hypothesis. This logistic regression in part 1C proves to be useful in comparison to the intercept-only model.

1H

```
pred1<-predict(regression2,newdata=test, type="response")

rates<-prediction(pred1, test$sex)

roc_result<-performance(rates,measure="tpr", x.measure="fpr")
plot(roc_result, main="ROC Curve - Penguins")
lines(x = c(0,1), y = c(0,1), col="blue")
```



The ROC Curve is above the line and shows us that the regression model is better than a random guess.

1I

```
AUC<-performance(rates, measure = "auc")
AUC@y.values
```

```
## [[1]]
## [1] 0.9214286
```

The AUC of the ROC curve is 0.9214286. This also indicates that the regression model performs better than a random guess.

1J

```
table(test$sex, pred1>0.5)
```

```
##
##           FALSE TRUE
##   female    28    7
##   male       4   28
```

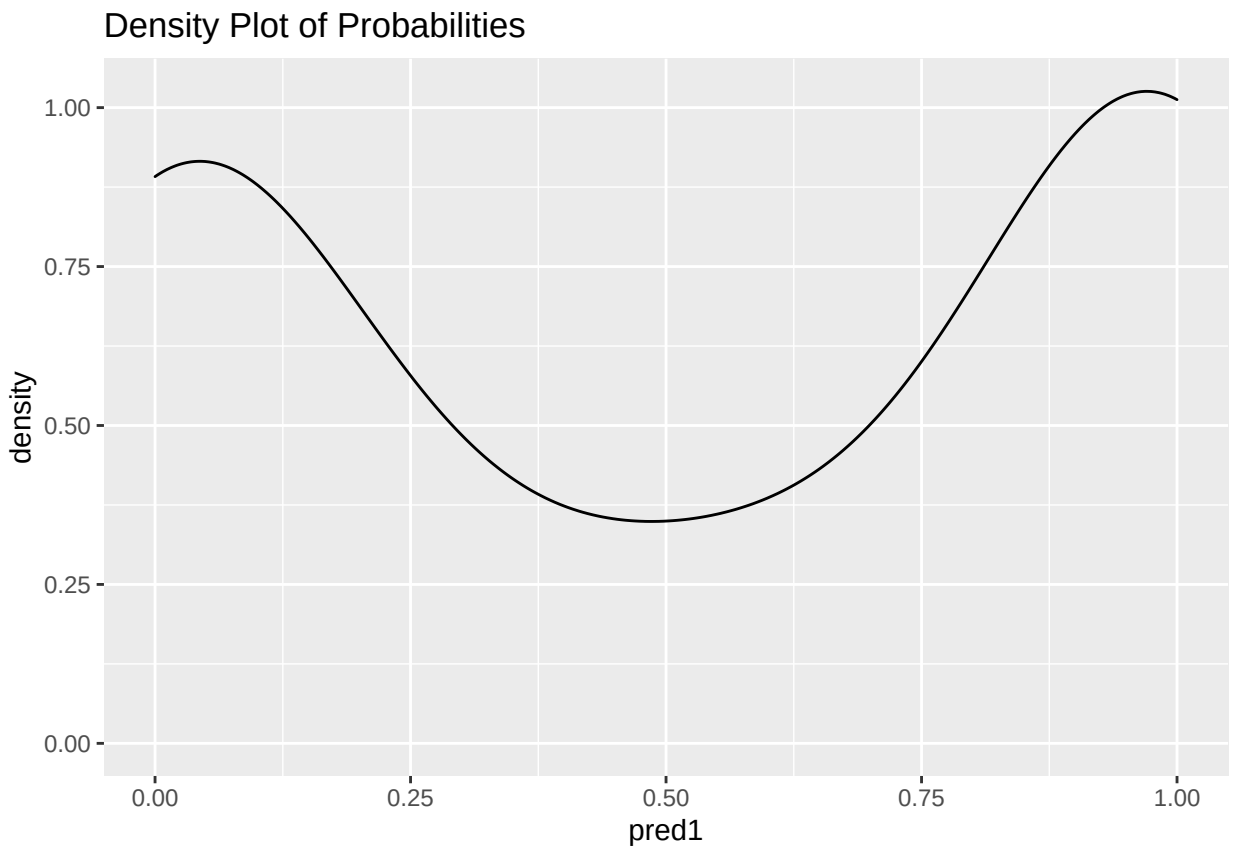
False positive rate = 7/35 = 0.2

False negative rate = 4/32 = 0.125

Error rate = 1 - accuracy = 1 - 56/67 = 0.1641791

1K

```
test<-data.frame(test,pred1)
ggplot(test,aes(x=pred1))+
  geom_density()+
  labs(title="Density Plot of Probabilities")
```



Based upon the results of the density plot, there does not seem to be a significant difference in the prediction probabilities. Thus, the current threshold appears to be adequate and does not need to be changed.

11/29/22

Module 11 & 12 HW

2) a) Coefficient for x_3 (gender): (=0.43 397)

The exponential function ($\exp(0.434) = 1.5434$) indicates that for a member of the female gender, their odds of receiving the flu shot increase by an average of 54.34% compared to members of the male gender.

b) Wald test for β_3:

$H_0: \beta_3 = 0$, $H_a: \beta_3 \neq 0$

w/ significance level = 0.05 ; critical Z-value = $\pm$ 1.96

(test) $Z = \frac{0.434}{0.523} = \boxed{0.829}$

Z-stat (0.829) < Z-critical (1.96)

Thus, we fail to reject the null hypothesis. This indicates that gender is not a significant predictor for which client receives the flu shot.

c) 95% CI for β_3:

Sig. level = 0.05, $Z_{crit} = 1.96$, MoE = (SE) x (Z_{crit})

MoE = (0.523)(1.96) = 1.025

Lower Bound = Estimate − MoE

LB = 0.434 − 1.025 = **−0.591**

UB = Estimate ± M.E = 0.434 + 1.025 = **1.459**

95% CI: (−0.591, 1.459)

Context: In this example, we are 95% confident that the chances of a male client receiving a flu shot is between (−0.591, 1.459) times the odds of a female client receiving the shot.

d) Yes, the conclusions from 2b & 2c are consistent. We know this because the 95% CI contains 0.

e) $H_0: \beta_1 = \beta_3 = 0$, H_a: one of the β coefficients in $H_0 \neq 0$

ΔG^2 = residual deviance (full) − residual deviance (reduced)

= 113.20 − 105.09 = **8.11**

p-value: 1 − pchisq(8.11, 2) = 0.0173

$p <$ threshold

Thus, we reject the null hypothesis and do not drop age and gender as predictors.

→

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

70 years old, awareness = 65, male:

$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.17716 + 0.07279(70) - 0.09899(65) + 0.43397(1)$

$= \boxed{-2.08224}$

Estimated probability of flu shot:

$e^{-2.08224} = 0.1246507$

$\rightarrow \frac{0.1246507}{(1 + 0.1246507} = \boxed{0.110835 \rightarrow \sim 11\%}$