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
```

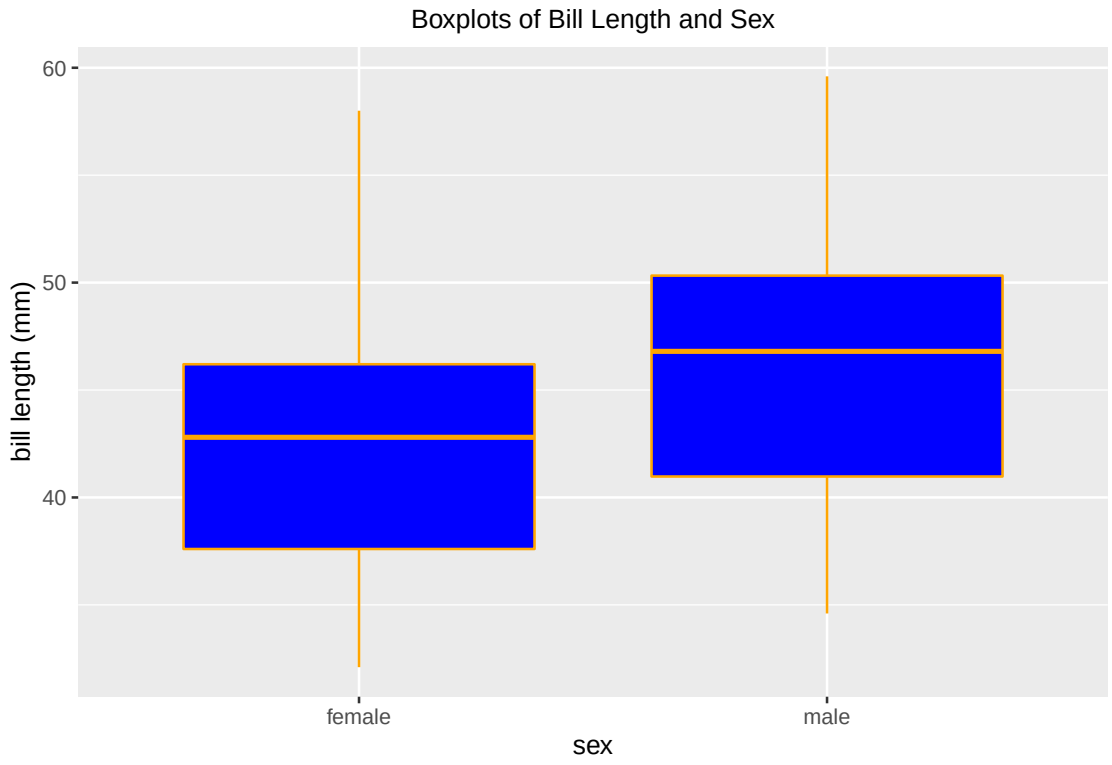
```
indices_of_training_observations <- sample.int(number_of_observations, floor(0.8 * number_of_observ
training_data_set <- data_set[indices_of_training_observations, ]
testing_data_set <- data_set[-indices_of_training_observations, ]
```

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

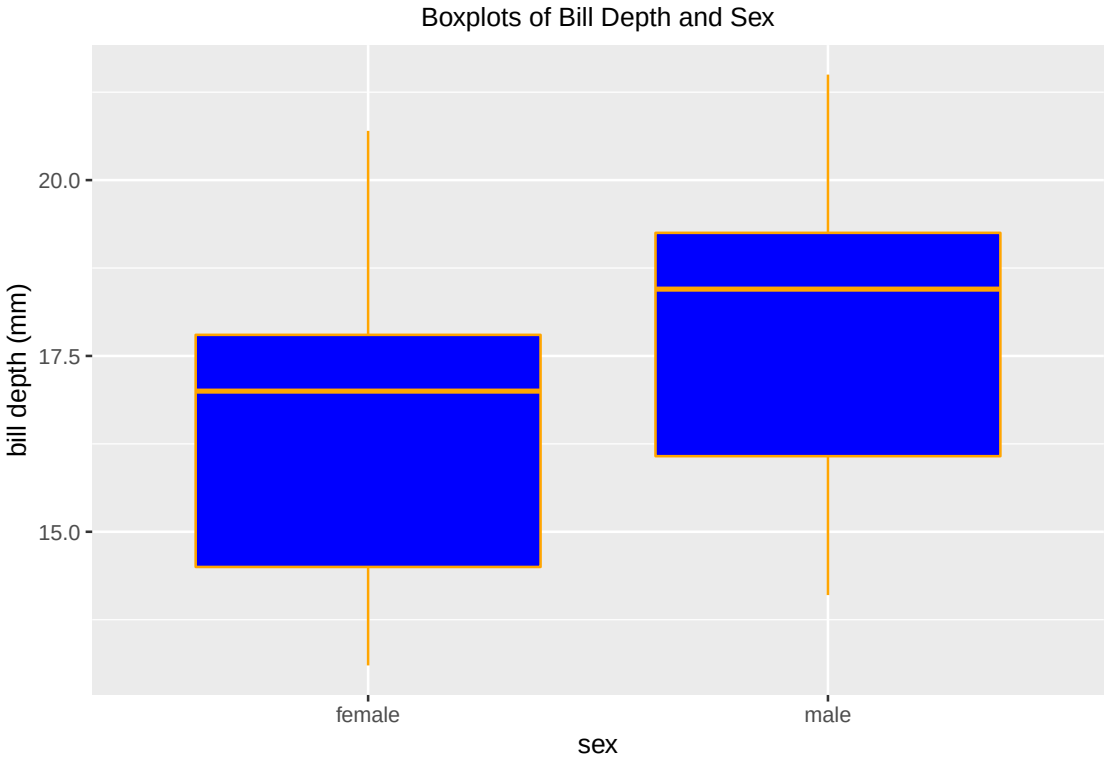
   a) Create some data visualizations to explore the relationship between the various body measurements and the gender of penguins. Be sure to briefly comment on your data visualizations.

      Boxplots of bill_length_mm, bill_depth_mm, flipper_length_mm, and body_mass_g are presented below.

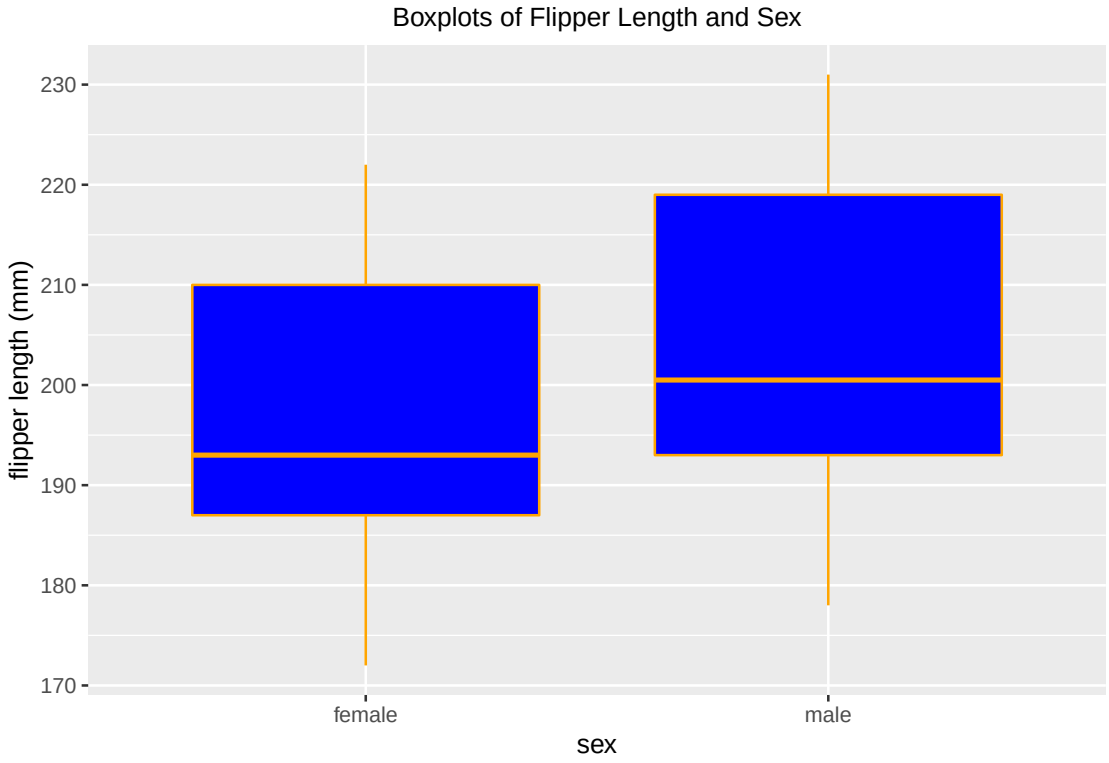
```
library(ggplot2)
ggplot(data_set, aes(x = sex, y = bill_length_mm)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
    labs(
        y = "bill length (mm)",
        title = "Boxplots of Bill Length and Sex"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
    )
```



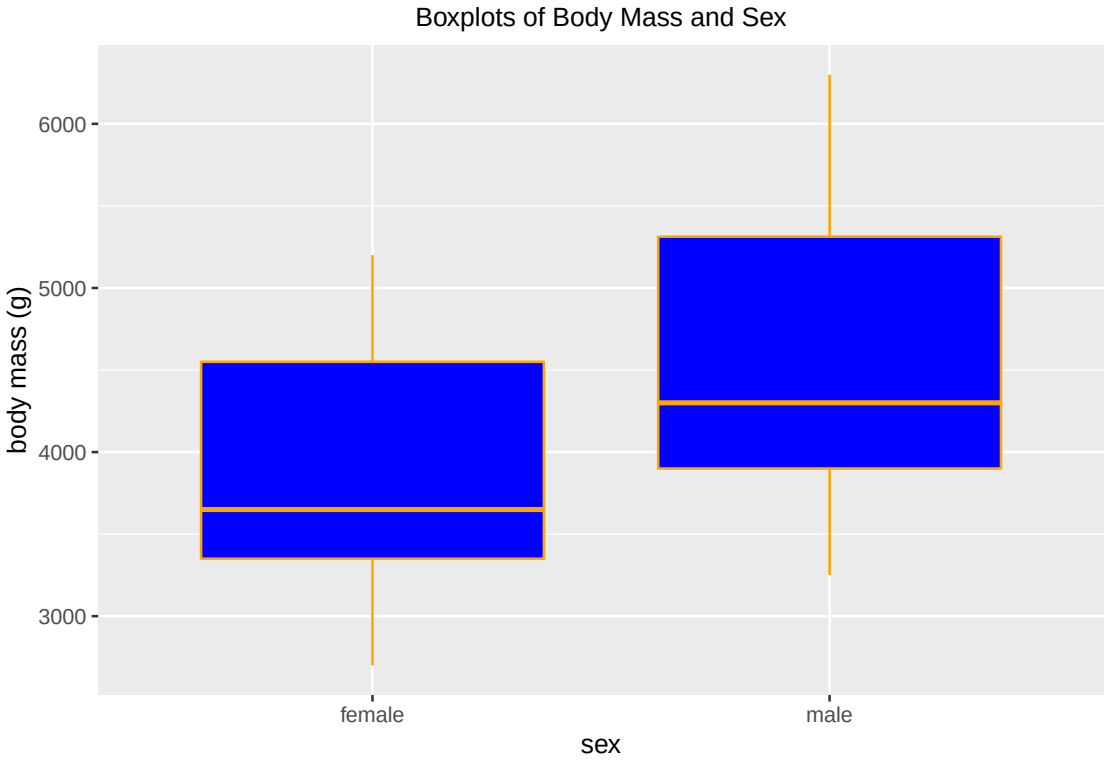

```
ggplot(data_set, aes(x = sex, y = bill_depth_mm)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
    labs(
        y = "bill depth (mm)",
        title = "Boxplots of Bill Depth and Sex"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
    )
```




```
ggplot(data_set, aes(x = sex, y = flipper_length_mm)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
    labs(
        y = "flipper length (mm)",
        title = "Boxplots of Flipper Length and Sex"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
    )
```




```
ggplot(data_set, aes(x = sex, y = body_mass_g)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
    labs(
        y = "body mass (g)",
        title = "Boxplots of Body Mass and Sex"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
    )
```



Male penguins tend to have longer, deeper bills, longer flippers, and greater body masses. The interquartile ranges of bill length, bill depth, flipper length, and body mass are approximately 10 mm, 3 mm, 25 mm, and 1300 g.

All of these body measurements differ similarly between male and female penguins. About three quarters of penguins have bill lengths and bill depths less than the appropriate medians; about three quarters of penguins have flipper lengths and body masses greater than the appropriate medians. Bill length and bill depth are right skewed; flipper length and body mass are left skewed.

There is significant overlap in the boxplots for the body measurements; many female and male penguins have similar body measurements.

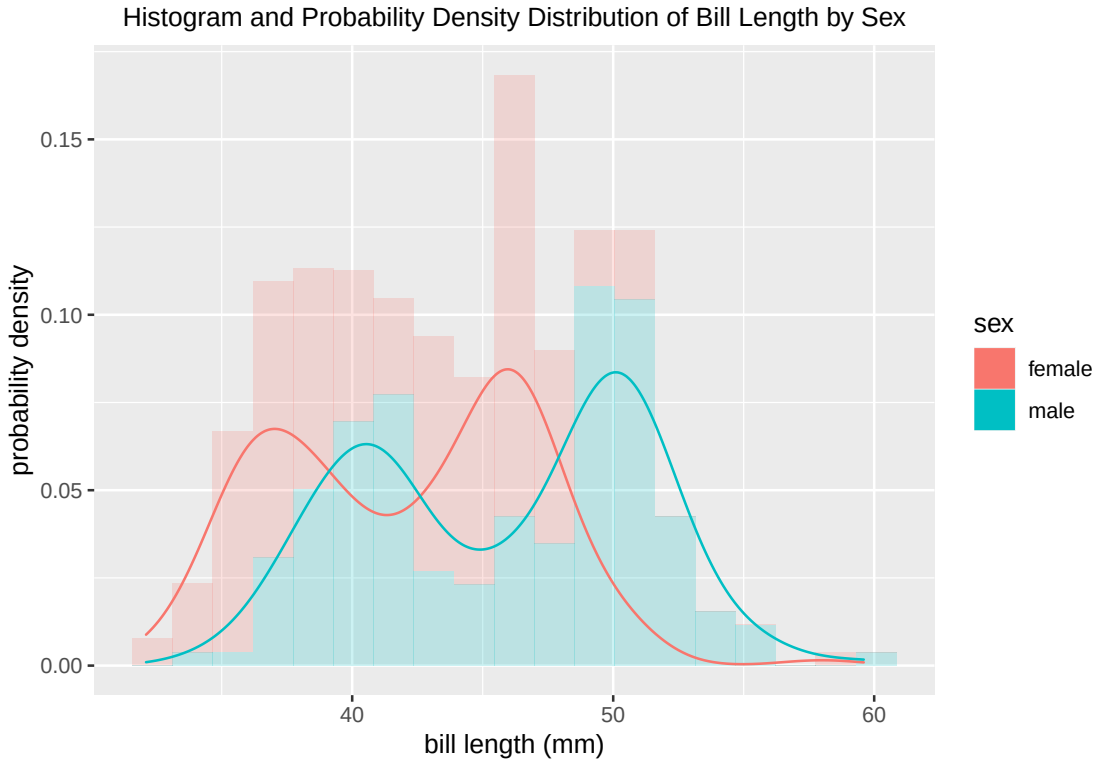
Density plots for bill_length_mm, bill_depth_mm, flipper_length_mm, and body_mass_g are presented below. The density plots for bill length and bill depth are roughly bimodal and left skewed; higher proportions of female and male penguins have lower bill length and bill depth. The density plots for males is shifted right relative to the density plot for females; males generally have greater bill lengths and bill depths than females. The density plots are more or less the same shape, and there is significant overlap. Male bill lengths and depths are more spread out.

Density plots for flipper_length_mm and body_mass_g are presented below. The density plots for flipper_length_mm and body_mass_g are roughly bimodal and right skewed; higher proportions of female and male penguins have higher bill length and bill depth. The density plots for males is shifted right relative to the density plot for females; males generally have greater flipper lengths and body masses than females. The density plots are more or less the same shape, and there is significant overlap. Male flipper lengths and body masses are more spread out.

```
ggplot() +
    geom_histogram(data = data_set, aes(x = data_set$bill_length_mm, y = ..density.., fill = se
    geom_density(data = data_set, aes(x = data_set$bill_length_mm, color = sex)) +
    labs(
        x = "bill length (mm)",
```

```
        y = "probability density",
        title = "Histogram and Probability Density Distribution of Bill Length by Sex"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(angle = 0)
    )
```

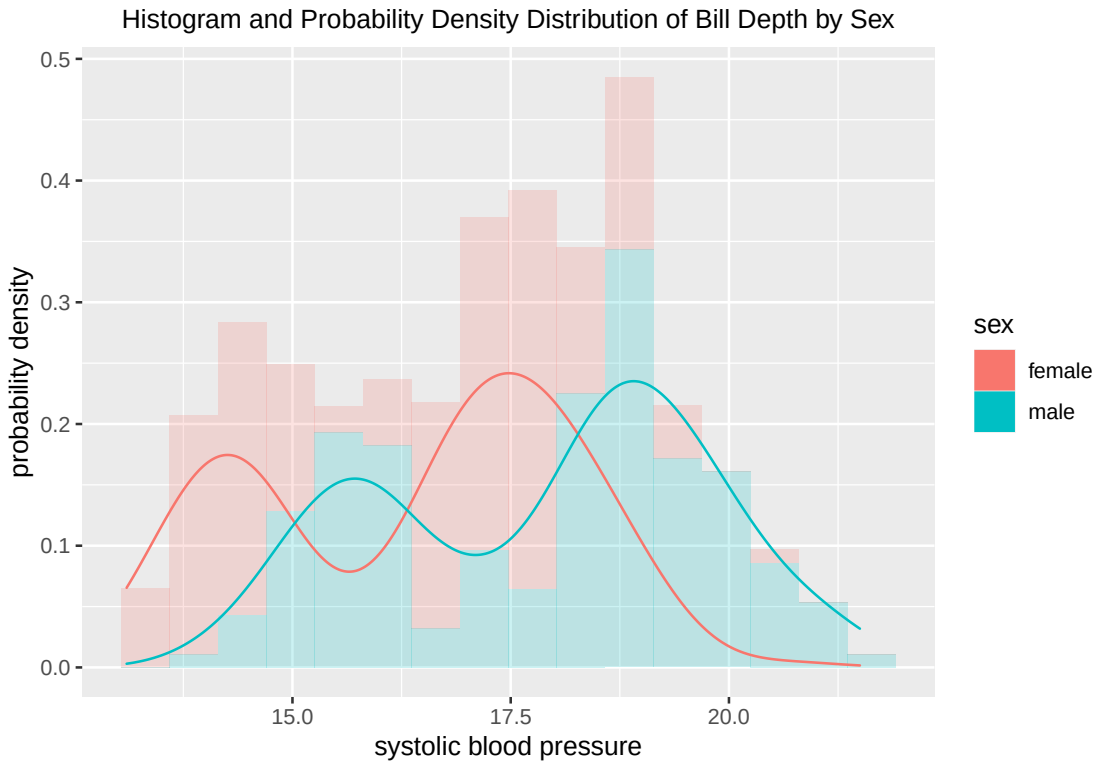
```
## Warning: Use of `data_set$bill_length_mm` is discouraged. Use `bill_length_mm` instead.
## Use of `data_set$bill_length_mm` is discouraged. Use `bill_length_mm` instead.
```




```
ggplot() +
    geom_histogram(data = data_set, aes(x = data_set$bill_depth_mm, y = ..density.., fill = sex
    geom_density(data = data_set, aes(x = data_set$bill_depth_mm, color = sex)) +
    labs(
        x = "systolic blood pressure",
        y = "probability density",
        title = "Histogram and Probability Density Distribution of Bill Depth by Sex"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(angle = 0)
    )
```

```
## Warning: Use of `data_set$bill_depth_mm` is discouraged. Use `bill_depth_mm`
## instead.
```

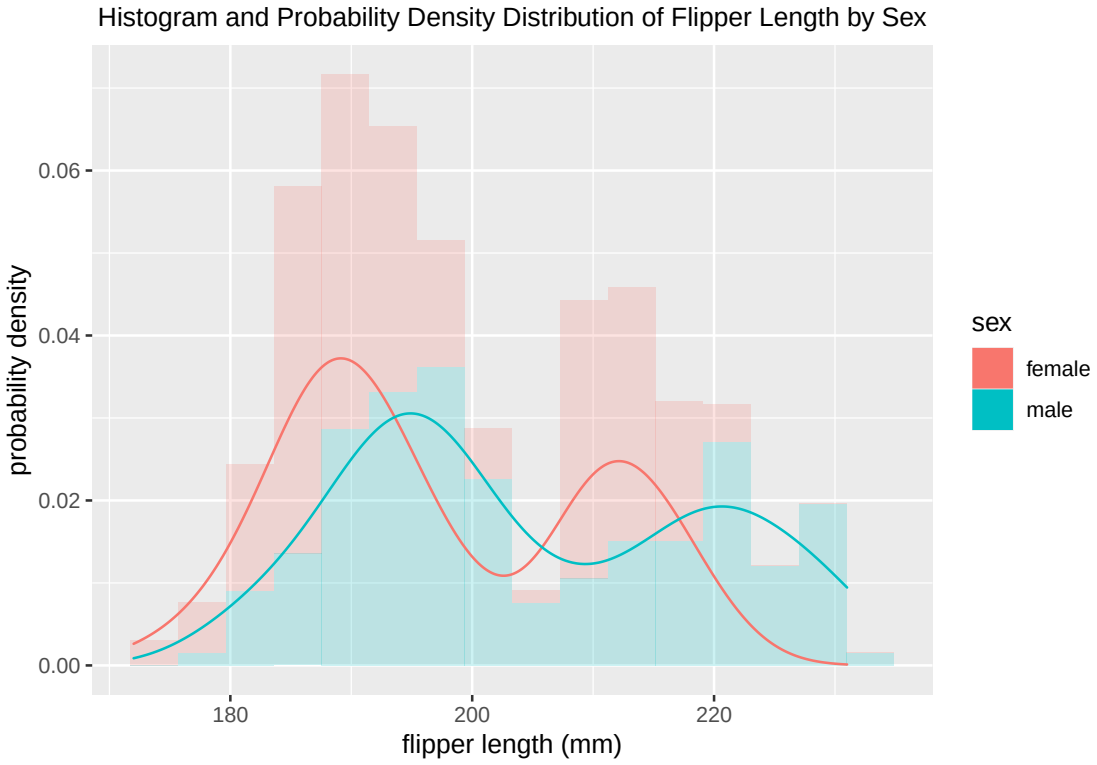
```
## Warning: Use of `data_set$bill_depth_mm` is discouraged. Use `bill_depth_mm`
## instead.
```




```
ggplot() +
    geom_histogram(data = data_set, aes(x = data_set$flipper_length_mm, y = ..density.., fill =
    geom_density(data = data_set, aes(x = data_set$flipper_length_mm, color = sex)) +
    labs(
        x = "flipper length (mm)",
        y = "probability density",
        title = "Histogram and Probability Density Distribution of Flipper Length by Sex"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(angle = 0)
    )
```

```
## Warning: Use of `data_set$flipper_length_mm` is discouraged. Use
## `flipper_length_mm` instead.
```

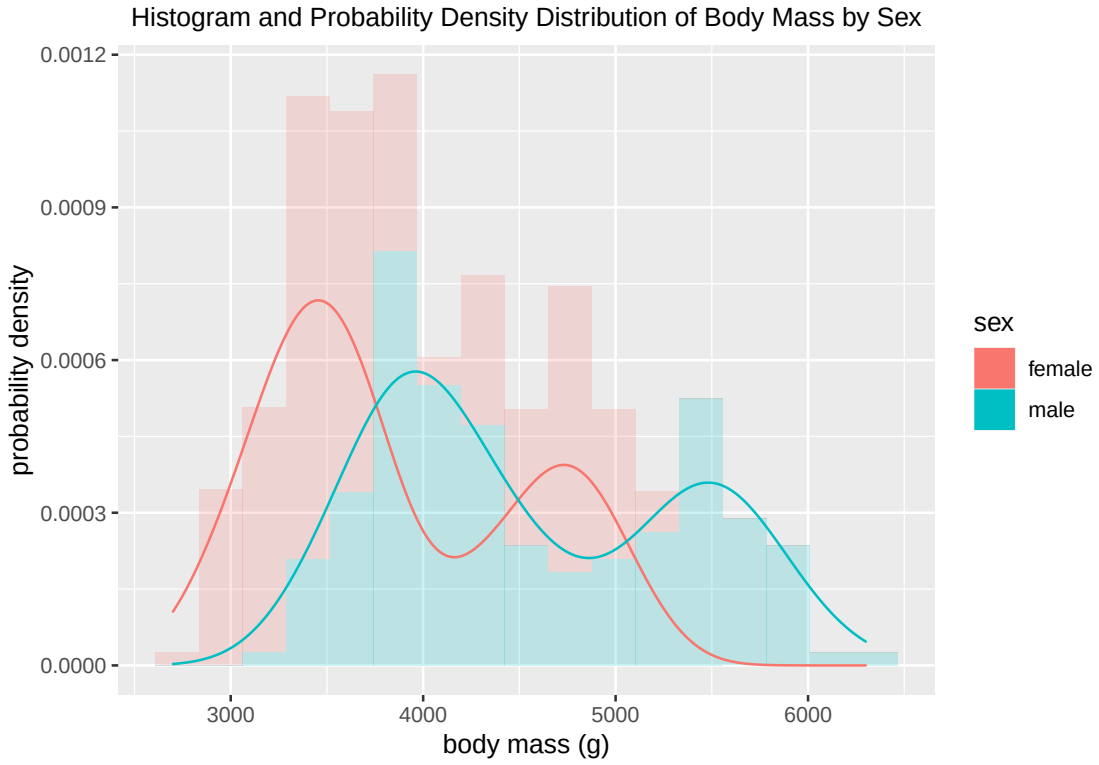
```
## Warning: Use of `data_set$flipper_length_mm` is discouraged. Use
## `flipper_length_mm` instead.
```




```
ggplot() +
    geom_histogram(data = data_set, aes(x = data_set$body_mass_g, y = ..density.., fill = sex)
    geom_density(data = data_set, aes(x = data_set$body_mass_g, color = sex)) +
    labs(
        x = "body mass (g)",
        y = "probability density",
        title = "Histogram and Probability Density Distribution of Body Mass by Sex"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(angle = 0)
    )
```

```
## Warning: Use of `data_set$body_mass_g` is discouraged. Use `body_mass_g`
## instead.
```

```
## Warning: Use of `data_set$body_mass_g` is discouraged. Use `body_mass_g`
## instead.
```



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 $flipper_length_mm$ $Z_0 = -1.358$ with a corresponding p value $p = 0.175$. $flipper_length_mm$ appears to be insignificant in the logistic regression model in the context of all the predictors in the model. We can drop $flipper_length_mm$ 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, f
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
```

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

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?

π represents the expected value of our logistic response function given predictor values, the probability that the response is 1, and the probability that a penguin is male. $\frac{\pi}{1-\pi}$ represents the

odds that the response is 1 and the odds that a penguin is male. The log odds of a penguin being male increases with bill length, bill depth, and body mass. The log odds of a penguin being male increases fastest with bill depth.

e) Interpret the estimated coefficient for bill length contextually.

The regression coefficient for *bill length* is 0.951.

For an additional millimeter in bill length on average, the estimated log odds of a penguin being male increases by 0.951, while controlling for the other predictors (*bill depth*, *body mass*, an indicator of whether a penguin is a Chinstrap penguin, and an indicator of whether a penguin is a Gentoo penguin).

For an additional millimeter in bill length on average, the estimated odds of a penguin being male increases by a factor of $exp(0.951)$, while controlling for the other predictors.

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

```
predictor_values <- data.frame(
    bill_length_mm = 49,
    bill_depth_mm = 15,
    flipper_length_mm = 220,
    body_mass_g = 5700,
    species = "Gentoo"
)
log_odds <- predict(generalized_linear_model, predictor_values)
log_odds
```

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

```
odds <- exp(log_odds)
odds
```

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

```
probability <- odds / (1 + odds)
probability
```

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

g) Conduct a relevant hypothesis test to assess if the logistic regression model 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.

We test the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ that the logistic regression coefficients are all 0. The alternate hypothesis is $H_1 : \beta_k \neq 0$ for at least one index of a regression coefficient k.

We consider dropping all 5 predictors of $\{$*bill length*, *bill depth*, *body mass*, $I_{Chinstrap}$, $I_{Gentoo}\}$. The test statistic is $D_{dropped} = D_0 - D_{full}$. The deviance of the dropped predictors follows a χ^2 distribution with degrees of freedom $df_{dropped} = 5$ equal to the number of predictors dropped.

```
D_0 <- generalized_linear_model$null.deviance
D_full <- generalized_linear_model$deviance
D_dropped <- D_0 - D_full
D_dropped
```

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

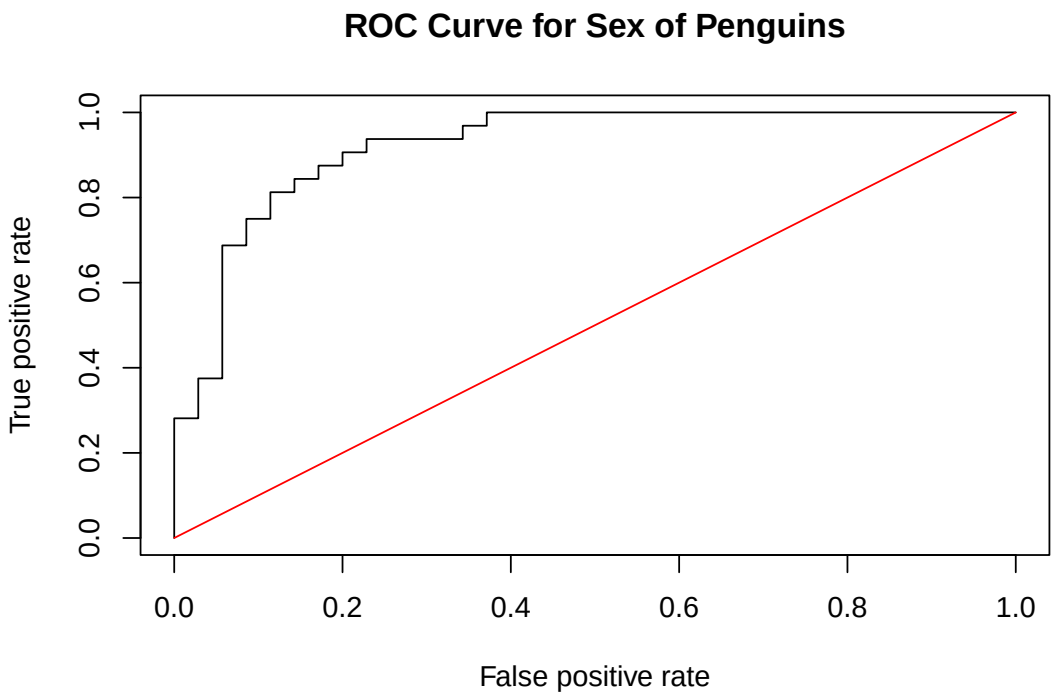
```
number_of_predictors_dropped <- 5
pchisq(D_dropped, number_of_predictors_dropped, lower.tail = FALSE)
```

```
## [1] 2.160184e-62
```

The associated p value is above. We reject the null hypothesis. The data support the claim that our model is useful compared to the intercept-only model.

h) Validate your model on the test data by creating an ROC curve. What does your ROC curve tell you?

```
library(ROCR)
predicted_sex <- predict(generalized_linear_model, newdata = testing_data_set, type = "response
prediction_instance <- prediction(predicted_sex, testing_data_set$sex)
roc_curve <- performance(prediction_instance, measure = "tpr", x.measure = "fpr")
plot(roc_curve, main = "ROC Curve for Sex of Penguins")
lines(x = c(0, 1), y = c(0, 1), col = "red")
```



Our ROC curve tells us that our logistic regression model is abot 90 percent on its way from a logistic regression model that predicts penguin sex randomly and without relying on any predictors (symbolized by the diagonal line) to a logistic regression model that predicts penguin sex perfectly [symbolized by the point (0, 1)].

i) Find the AUC associated with your ROC curve. What does your AUC tell you?

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

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

The Area Under the Curve of the ROC curve tells us that our logistic regression model is about

90 percent on its way from a logistic regression model that predicts penguin sex randomly and without relying on any predictors (symbolized by the area under the diagonal line) to a logistic regression model that predicts indicators of penguin sex perfectly (symbolized by the entire plot area).

j) Create a confusion matrix using a threshold of 0.5. What is the false positive rate? What is the false negative rate? What is the error rate?

```
table(testing_data_set$sex, predicted_sex > 0.5)
```

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

$$FPR = \frac{FP}{TN + FP} = \frac{7}{28 + 7} = \frac{7}{35} = \frac{1}{5} = 0.2 = P\left(\hat{y} = 1 | y = 0\right)$$

$$FNR = \frac{FN}{FN + TP} = \frac{4}{4 + 28} = \frac{4}{32} = \frac{1}{8} = 0.125 = P\left(\hat{y} = 0 | y = 1\right)$$

$$ER = \frac{FP + FN}{n} = \frac{FP + FN}{TN + FP + FN + TP} = \frac{7 + 4}{28 + 7 + 4 + 28} = \frac{11}{67} = 0.164 = P\left(\hat{y} \neq y\right)$$

k) Discuss if the threshold should be changed. If it should be changed, explain why, and create another confusion matrix with a different threshold.

A cutoff value of 0.5 gives the highest accuracy / lowest overall error rate on average. At this point, we have a fairly close-to-ideal ROC curve, a fairly high AUC, a fairly low error rate, a fairly low FPR, and a fairly low FNR. The cutoff value of 0.5 seems satisfactory. We could search for the threshold corresponding to the lowest error rate. Depending on context, we may be more interested in ensuring a high sensitivity / TPR / probability of predicted sex being male when actual sex is male or ensuring high specificity / TNR / probability of predicted sex being female when actual sex is female.

2. A health clinic sent fliers to its clients to encourage everyone to get a flu shot. In a follow-up study, 159 elderly clients were randomly selected and asked if they received a flu shot. A client who received a flu shot was coded $y = 1$, and a client who did not receive a flu shot was coded $y = 0$. Data were also collected on their age, x_1, health awareness rating on a $0 - 100$ scale (higher values indicate greater awareness), x_2, and gender, x_3, where males were coded $x_3 = 1$ and females were coded $x_3 = 0$. A first order logistic model was fitted. The output is displayed in the prompt for this homework.

a) Interpret the estimated coefficient for x_3, gender, in context.

The regression coefficient for `gender` is 0.434.

The estimated log odds of a client having received a flu shot is 0.434 higher for males than for females, while controlling for the other predictors.

The estimated odds of a client having received a flu shot is a factor of $exp(0.434)$ higher for males than for females, while controlling for the other predictors.

b) Conduct the Wald test for β_3. State the null and alternative hypotheses, calculate the test statistic, and make a conclusion in context.

We conduct a Wald test a null hypothesis $H_0 : \beta_3 = 0$ that the regression coefficient $\beta_3 = 0$ for individual predictor gender, x_3. Our alternate hypothesis $H_1 : \beta_3 \neq 0$. The test statistic for our Wald test is

$$Z_0 = \frac{\hat{\beta}_3}{SE\left(\hat{\beta}_3\right)} = \frac{0.43397}{0.52179} = 0.832$$

```
test_statistic <- 0.832
significance_level <- 0.05
qnorm(significance_level, lower.tail = FALSE)
```

```
## [1] 1.644854
```

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

```
## [1] 0.4054089
```

Because the test statistic Z_0 is less than a critical value $Z_c = 1.645$ and the p value corresponding to the test statistic Z_0 is greater than a significance level $\alpha = 0.05$, we fail to reject the null hypothesis. We drop the predictor gender, x_3, from our logistic regression model.

c) Calculate a 95-percent confidence interval for β_3, and interpret the interval in context.

$$\left[\hat{\beta}_3 - Z_{\alpha/2} SE\left(\hat{\beta}_3\right), \hat{\beta}_3 + Z_{\alpha/2} SE\left(\hat{\beta}_3\right)\right]$$

$$L = 0.95$$

$$alpha = 0.05$$

```
qnorm(significance_level / 2, lower.tail = FALSE)
```

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

$$Z_{alpha/2} = 1.960$$

$$[0.434 - (1.960)(0.522), 0.434 + (1.960)(0.522)]$$

$$[-0.589, 1.457]$$

Because a 95-percent confidence interval for β_3 contains 0, the population regression coefficient of the logistic regression model for predictor gender, x_3 is not significantly different from 0. We drop predctor gender, x_3, from our logistic regression model.

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

Our conclusions from parts 2b and 2c are consistent. We drop the predictor x_3, gender, in both cases.

e) Suppose you want to drop the coefficients for age and gender, β_1 and β_3. A logistic regression model for just awareness was fitted, and the output is shown in the prompt for this homework. Carry out the appropriate hypothesis test to see if the coefficients for age and gender can be dropped.

Our null hypothesis $H_0 : \boldsymbol{\beta}_{dropped} = \mathbf{0}$. Our alternate hypothesis $H_1 : \boldsymbol{\beta}_{d}ropped \neq \mathbf{0}$.

The deviance of the full model is the residual deviance in the summary for the full model, $D_{full} = 105.09$. The deviance of the reduced model is the residual deviance in the summary for the reduced model, $D_{reduced} = 113.20$. The deviance of the dropped predictors is the difference between the deviance of the reduced model and the deviance of the full model, 8.11. If the deviance of the dropped predictors $D_{dropped} = 8.11$ is greater than a test statistic ${\chi^2}_{\alpha=0.05,dropped=2} = 5.99$, at least one of the regression coefficients in $\beta_{dropped}$ is likely not 0, and we reject the null hypothesis. If the deviance of the dropped predictors $D_{dropped} = 8.11$ is less than a test statistic ${\chi^2}_{\alpha=0.05,dropped=2} = 5.99$, the reduced model is about as good a fit to the data as the full model

and we support the null hypothesis; it is likely the regression coefficients of the dropped predictors are 0. Because the deviance of the dropped predictors $D_{dropped} = 8.11$ is less than a test statistic ${\chi^2}_{\alpha=0.05,dropped=2} = 5.99$, the reduced model is about as good a fit to the data as the full model and we support the null hypothesis; it is likely the regression coefficients of the dropped predictors are 0. We drop the predictors from our logistic regression model.

```
D_full <- 105.09
D_reduced <- 113.20
D_dropped <- D_reduced - D_full
D_dropped
```

```
## [1] 8.11
```

```
significance_level <- 0.05
number_of_dropped_predictors <- 2
qchisq(significance_level, number_of_dropped_predictors, lower.tail = FALSE)
```

```
## [1] 5.991465
```

```
pchisq(D_dropped, number_of_predictors_dropped, lower.tail = FALSE)
```

```
## [1] 0.1502766
```

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?

If we were to keep age, aware, and gender,

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.177 + (0.07279)age + (-0.09899)aware + (0.43397)gender$$

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.082$$

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.082$$

```
log_odds <- -2.082
odds <- exp(log_odds)
odds
```

```
## [1] 0.1246806
```

```
probability <- odds / (1 + odds)
probability
```

```
## [1] 0.1108587
```

If we drop age and gender and go with aware as in part 2e,

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 4.91133 + (-0.11931)aware$$

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 4.91133 + (-0.11931)(65)$$

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.84382$$

```r
log_odds <- -2.84382
odds <- exp(log_odds)
odds
```

```
## [1] 0.05820291
```

```r
probability <- odds / (1 + odds)
probability
```

```
## [1] 0.05500165
```