Homework 5

2/9/2024

Question 1

Part a)

Using the *titanic_dat* data frame, conduct a logistic regression analysis to predict whether passengers' survival (*Survived*) is predicted by the cost of fare paid by passengers (*Fare*). Write out the logistic regression equation on the odds scale. Please write the equation in the most simplified format (i.e., write it in the form of $intercept * slope^{X_{predictor}}$ with each estimate converted to odds ratios; don't write it in the form of $e^{intercept+slope*X_{predictor}}$; remember to simplify as much as possible). Interpret the slope of *Fare*.

```
# Code

mod1 <- glm(Survived ~ Fare, data = titanic_dat, family='binomial')
summary(mod1)
```

```
##
## Call:
## glm(formula = Survived ~ Fare, family = "binomial", data = titanic_dat)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.896828   0.107616  -8.334  < 2e-16 ***
## Fare         0.015997   0.002502   6.394 1.61e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 901.25  on 712  degrees of freedom
## AIC: 905.25
##
## Number of Fisher Scoring iterations: 5
```

```
# Exponetiated coefficients
exp(mod1$coefficients)
```

```
## (Intercept)        Fare
##   0.4078613   1.0161258
```

Regression Equation:

$$\widehat{Odds}_{Survival} = 0.41 + 1.02 \times \text{Fare}_i$$

Interpretation:

- **Slope of Fare (odds)**: For every one-unit increase in the cost of the *Fare* paid by the passenger, the odds of survival ($Survival = 1$) increases by a factor of 1.02.

Part b)

Create a graph that demonstrates the results of the analysis above. The graph should show *Fare* (on the x-axis) plotted against the predicted **probability** of *Survived*. The graph should include the predicted probabilities for each value of fare and a fit line (hint: For an example of how to add non-linear fit lines to a graph, see the example in the lab where we did this using a separate data frame called `line.data)`. The y-scale on the graph should go from 0% to 100%, and the x- and y-axes should be appropriately labeled. You are welcome to use a jitter on the data points and/or to adjust the points' alpha level (transparency), but it is OK if the points clump on the fit line somewhat (as long as the graph looks cleans and legible, that is the most important thing). You do not need any extra panels for predicted odds or predicted logits.

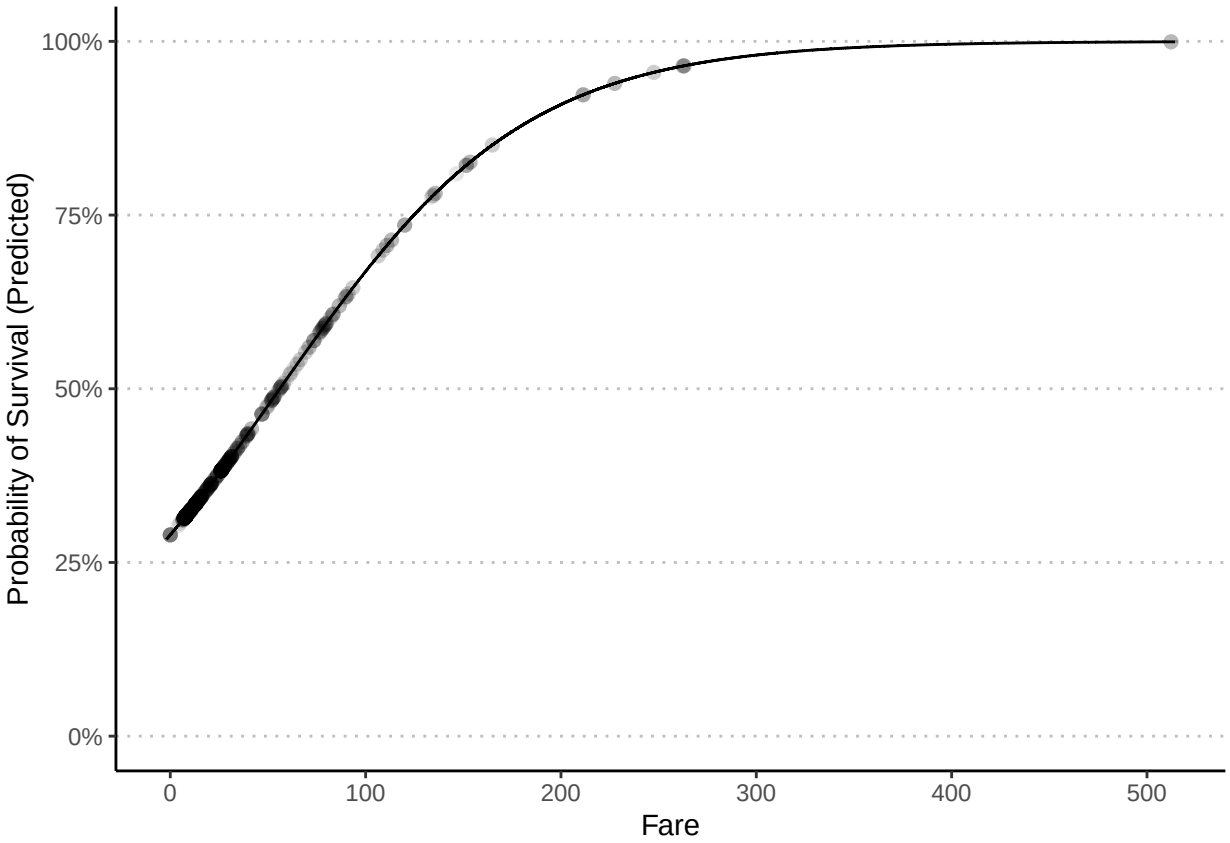
```
titanic_dat <- titanic_dat |>
  dplyr::mutate(
    prob = exp(predict(mod1)) / (1 + exp(predict(mod1)))
  )


line_data <- data.frame(
  Fare = seq(from = min(titanic_dat$Fare, na.rm = T) - 2,
to = max(titanic_dat$Fare, na.rm = T) + 2,
by = .01))

line_data$prob <- exp(predict(mod1, line_data)) / (1 + exp(predict(mod1, line_data)))

titanic_dat |>
  ggplot(aes(x = Fare, y = prob)) +
  geom_line(data = line_data, aes(x = Fare, y = prob)) +
  geom_point(alpha = .1, size = 2) +
  labs(y = "Probability of Survival (Predicted)") +
  scale_y_continuous(breaks = c(0,.25, .5, .75, 1), limits = c(0, 1),
                     label = percent) +
  theme_classic()  +
  theme(panel.grid.major.y =
          element_line(size = .5, color = "grey", linetype = "dotted"))
```

```
## Warning: The ‘size‘ argument of ‘element_line()‘ is deprecated as of ggplot2 3.4.0.
## i Please use the ‘linewidth‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Part c)

Create a Null Model for *Survived* (i.e., an intercept-only logistic regression model). Then, calculate the accuracy rate (the percent of correct predictions, or in other words, $1 - error.rate$) for the Null Model and for the model created in Part a (i.e., *Survived* predicted by *Fare*), and report these percentages below. Please report these values as percentages and not decimals.

Note: To calculate the accuracy rate for these models, assume that a predicted probability of Survival greater than or equal to 50% means that a person was predicted to Survive ($Survived = 1$), and that a predicted probability below 50% means that a person was predicted to Not Survive ($Survived = 0$).

```
# Code
null_model <- glm(Survived ~ 1, data = titanic_dat, family = "binomial")

titanic_dat <- titanic_dat |>
  dplyr::mutate(
    prob_null = exp(predict(null_model)) / (1 + exp(predict(null_model))),
    accuracy_mod1 = dplyr::case_when(
      (Survived == 1 & prob > .5 ) | (Survived == 0 & prob < .5 ) ~ 1,
      TRUE ~ 0
    ),
    accuracy_null = dplyr::case_when(
      (Survived == 1 & prob_null > .5 ) | (Survived == 0 & prob_null < .5 ) ~ 1,
      TRUE ~ 0
    )
  )
```

Answers

- **Accuracy Rate of Null Model**: In the null model, the correct prediction frequency was 59.38%
- **Accuracy Rate of Model 1**: In the model of `Survival` predicted by `Fare`, 66.95% of the cases were correctly classified.

Question 2

Part a)

In the table below, report the Deviance and AIC for the following regression models (some of these models you already created, some you have not). Round all decimals to 2 places.

- Model 0: The Null model
- Model 1: *Survived* predicted by *Fare*
- Model 2: *Survived* predicted by *Age*
- Model 3: *Survived* predicted by *Sex*

```
# Code
fit_null <- summary(null_model)
fit_fare <- summary(mod1)
fit_age <- summary(glm(Survived ~ Age, data = titanic_dat, family = "binomial"))
fit_sex <- summary(glm(Survived ~ Sex, data = titanic_dat, family = "binomial"))
```

Model	Deviance	AIC
Model 0 (Null Model)	964.52	966.52
Model 1 (Fare)	901.25	905.25
Model 2 (Age)	960.23	964.23
Model 3 (Sex)	750.7	754.7

Part b)

Based on the AIC from the table in *Part a*, answer the following questions in the space below:

1. Which variables (*Fare*, *Age*, and/or *Sex*) were significant predictors of Surviving?
2. Which variable (*Fare*, *Age*, or *Sex*) was the **best** predictor of Surviving? How can you tell?

Answers:

1. *Fare*, *Age*, and *Sex*.
2. *Sex* is the best predictor because it gives the greatest decrease in AIC relative to the null model.

Question 3

Create a regression model with *Survived* predicted by *Age* + *Fare* + *Sex*. Then, use this model to predict the probability of the following two people surviving the titanic (given the information provided about them below). Report the predicted probabilities, and report these values as percentages below.

1. Rose, a 17-year old female who paid a Fare of $84.
2. Jack, a 23-year old male who snuck aboard without paying any Fare ($0).

```
# Code

mod3 <- lm(Survived ~ Age + Fare + Sex, data=titanic_dat)

rose <- data.frame(Age = 17, Sex = "female", Fare = 84)
jack <- data.frame(Age = 23, Sex = "male", Fare = 0)

# Rose probability
exp(predict(mod3, rose)) / (1 + exp(predict(mod3, rose)))
```

```
##         1
## 0.6970624
```

```
# Jack probability
exp(predict(mod3, jack)) / (1 + exp(predict(mod3, jack)))
```

```
##         1
## 0.5429404
```

Answers

- **Probability of Rose Surviving**: 69.71%
- **Probability of Jack Surviving**: 54.29%

Question 4

The code for a complex regression model is provided below, in which *Survived* is predicted by a three way interaction between *Age*, *Fare*, and Sex. Run the code and look at the summary output. Then, read the statements below. Based on the analysis output and your understanding of logistic regression, indicate which of the statements are correct and which ones are incorrect (it is possible for more than one statement to be correct); focus on the underlined portion of each statement. For each statement, provide a brief explanation of why the statement was correct or not (one to two sentences will do). You may write additional code if you would like (however, you technically have all the information that you need to answer the question from the output below).

1. There was a significant three-way interaction between *Age*, *Fare*, and *Sex* predicting survival status.
2. Each slope indicates how the predicted score of *Survived* will increase or decrease for every one unit increase in each respective predictor.
3. The model intercept reflects the estimated logit of *Survived* for females of an average age and who paid an average fair.
4. This model was significantly improved relative to the null model.

```
complex.model <-
  glm(Survived ~ Age * Fare * Sex,
    family = "binomial",
    data = titanic_dat) %>%
  summary()

complex.model
```

```
##
## Call:
## glm(formula = Survived ~ Age * Fare * Sex, family = "binomial",
##     data = titanic_dat)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.1231467  0.4671529   2.404 0.016206 *
## Age              -0.0321516  0.0180072  -1.785 0.074183 .
## Fare             -0.0152546  0.0129267  -1.180 0.237968
## Sexmale          -2.0077288  0.5835864  -3.440 0.000581 ***
## Age:Fare          0.0015767  0.0005771   2.732 0.006295 **
## Age:Sexmale       0.0067298  0.0212314   0.317 0.751263
## Fare:Sexmale      0.0234538  0.0145952   1.607 0.108065
## Age:Fare:Sexmale -0.0015390  0.0006070  -2.536 0.011224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 695.32  on 706  degrees of freedom
## AIC: 711.32
##
## Number of Fisher Scoring iterations: 7
```

Statement 1

- **Is the Statement Correct:** YES
- **Why:** The coefficient for the three way interaction (`Age:Fare:Sexmale`) is statistically significant with $p = 0.011$.

Statement 2

- **Is the Statement Correct:** NO
- **Why:** Each slope indicates how the predicted logit of *Survived* will increase or decrease for every one unit increase in each respective predictor, keeping all the other predictors constant.

Statement 3:

- **Is the Statement Correct:** NO
- **Why:** Because the data is not centered, the model intercept reflects the estimated logit of *Survived* for females of age 0 and who paid 0 in fare.

Statement 4

- **Is the Statement Correct**: YES
- **Why**: The improvement of the model with predictors over a null model is reflected by the decrease of 269.2 units in the deviance.

Extra Credit

Part a)

Based on your answers to Question 3, how much **more** likely was Rose to survive than Jack? Write your answer in the space below, and include any code used to calculate your answer in the chunk below. Round to 2 decimal places. This question is worth an extra 0.5 points.

Answer: The odds for Rose to survive are 2.30, whereas the odds for Jack to survive are 1.19. We can find how likely Rose was to survive than Jack by calculating the odds ratio: $\frac{2.30}{1.89} = 1.93$. Therefore, Rose was 1.94 times more likely to have survived than Jack.

```
odds_rose <- exp(predict(mod3, rose))
odds_jack <- exp(predict(mod3, jack))

odds_ratio <- odds_rose/odds_jack
round(odds_ratio, 2)
```

```
##    1
## 1.94
```

Part b)

Based on the analysis output in Question 4, in the space below, fill in the blank for the following statement:

"For a female who paid $20 in fair, the logit of survival increased by [BLANK] for every one unit increase in Age".

Enter the number in space below, rounded to 3 decimal places. Write any code that you use for calculations in the space below. This question is worth an extra 0.5 points.

Answer: For a female who paid $20 in fair, the logit of survival increased by -0.0006 for every one unit increase in Age

```
titanic_dat$Fare20 <- titanic_dat$Fare - 20

mod4 <- glm(Survived ~ Age * Fare20 * Sex, family = "binomial", data = titanic_dat)
summary(mod4)
```

```
##
## Call:
## glm(formula = Survived ~ Age * Fare20 * Sex, family = "binomial",
##     data = titanic_dat)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.8180554  0.3343023   2.447 0.014403 *
## Age                -0.0006185  0.0117704  -0.053 0.958092
## Fare20             -0.0152546  0.0129267  -1.180 0.237968
## Sexmale            -1.5386528  0.4381000  -3.512 0.000445 ***
## Age:Fare20          0.0015767  0.0005771   2.732 0.006295 **
## Age:Sexmale        -0.0240507  0.0149703  -1.607 0.108150
## Fare20:Sexmale      0.0234538  0.0145952   1.607 0.108065
```

```
## Age:Fare20:Sexmale -0.0015390  0.0006070  -2.536 0.011224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 695.32  on 706  degrees of freedom
## AIC: 711.32
##
## Number of Fisher Scoring iterations: 7
```