Assignment #11

Troy Edwards

2023-04-11

Question 1

Calculate the conditional probability that a person survives given their sex and passenger-class: (You will have to find a creative way to summarize the data using either Rstudio or Excel. Google search 'pivot table', this is how I would summarize my data using Excel data sheets. Answer in fractions AND decimals.)

```
pt_data <- raw_data[order(raw_data$Sex, decreasing = TRUE),]
pt_data$Survived <- ifelse(pt_data$Survived == 1, "Survived", "Died")
pt_data$Sex <- lapply(pt_data$Sex, str_to_title)
pt_data[pt_data$Pclass == 1,]$Pclass <- "1st class"
pt_data[pt_data$Pclass == 2,]$Pclass <- "2nd class"
pt_data[pt_data$Pclass == 3,]$Pclass <- "3rd class"

pt <- PivotTable$new()
pt$addData(pt_data)
pt$addColumnDataGroups("Survived", caption = "{value}", dataSortOrder = "desc")
pt$addRowDataGroups("Pclass", caption = "{value}")
pt$addRowDataGroups("Sex")
pt$defineCalculation(calculationName = "Total", summariseExpression = "n()")
cat(pt$getLatex())
```

		Survived	Died	Total
1st class	Male	45	77	122
	Female	91	3	94
	Total	136	80	216
2nd class	Male	17	91	108
	Female	70	6	76
	Total	87	97	184
3rd class	Male	47	296	343
	Female	72	72	144
	Total	119	368	487
Total		342	545	887

$$
\begin{aligned}
P(\text{Survived} \mid \text{Female}) &= \frac{233}{314} = 0.7420482 \\
P(\text{Survived} \mid \text{Male}) &= \frac{109}{573} = 0.1902269 \\
P(\text{Survived} \mid \text{1st Class}) &= \frac{136}{216} = 0.6296296 \\
P(\text{Survived} \mid \text{2nd Class}) &= \frac{87}{184} = 0.4728261 \\
P(\text{Survived} \mid \text{3rd Class}) &= \frac{119}{487} = 0.2443532 \\
P(\text{Survived} \mid \text{Female} \wedge \text{1st Class}) &= \frac{91}{94} = 0.9680851 \\
P(\text{Survived} \mid \text{Female} \wedge \text{2nd Class}) &= \frac{70}{76} = 0.9210526 \\
P(\text{Survived} \mid \text{Female} \wedge \text{3rd Class}) &= \frac{72}{144} = 0.5000000 \\
P(\text{Survived} \mid \text{Male} \wedge \text{1st Class}) &= \frac{45}{122} = 0.3688525 \\
P(\text{Survived} \mid \text{Male} \wedge \text{2nd Class}) &= \frac{17}{108} = 0.1574074 \\
P(\text{Survived} \mid \text{Male} \wedge \text{3rd Class}) &= \frac{47}{343} = 0.1370262
\end{aligned}
$$

Question 2

How much did people pay to be on the ship? Let X = the random variable of cost for ticket, calculate the expectation of fare conditioned on passenger-class.

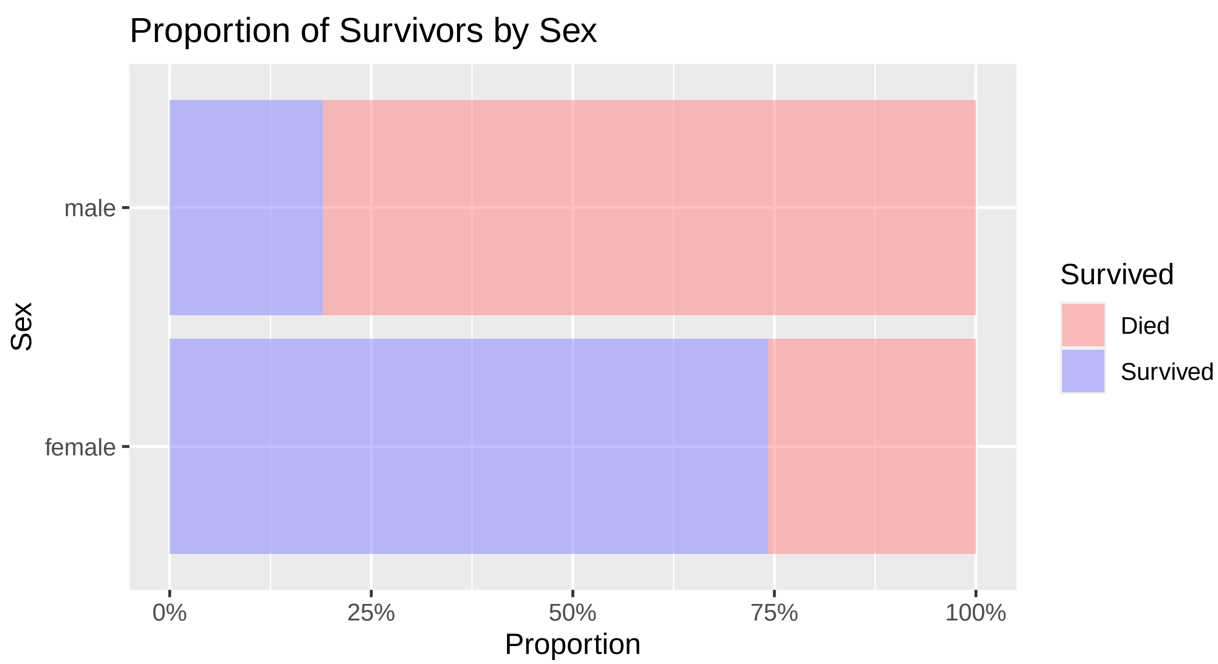
$$
\begin{aligned}
E(X \mid \text{1st Class}) &= \$84.15 \\
E(X \mid \text{2nd Class}) &= \$20.66 \\
E(X \mid \text{3rd Class}) &= \$13.71
\end{aligned}
$$

Question 4

Create a plot (a stacked bar chart for example, but you can choose anything) to compare passenger survival numbers/percentages per gender. Write a few sentences summarizing what you can conclude from your display only. You might want to have your graphs displaying relative frequencies instead of raw counts.

```
ggplot(titanic, aes(Sex)) +
    geom_bar(
        aes(fill = ifelse(Survived == 1, "Survived", "Died")),
```

```
        position = "fill"
    ) +
    coord_flip() +
    scale_y_continuous(labels = scales::percent) +
    scale_fill_manual(
        name = "Survived",
        values = setNames(
            c("#ff888888", "#8888ff88"),
            c("Died", "Survived")
        )
    ) +
    labs(
        x = "Sex",
        y = "Proportion",
        title = "Proportion of Survivors by Sex"
    )
```

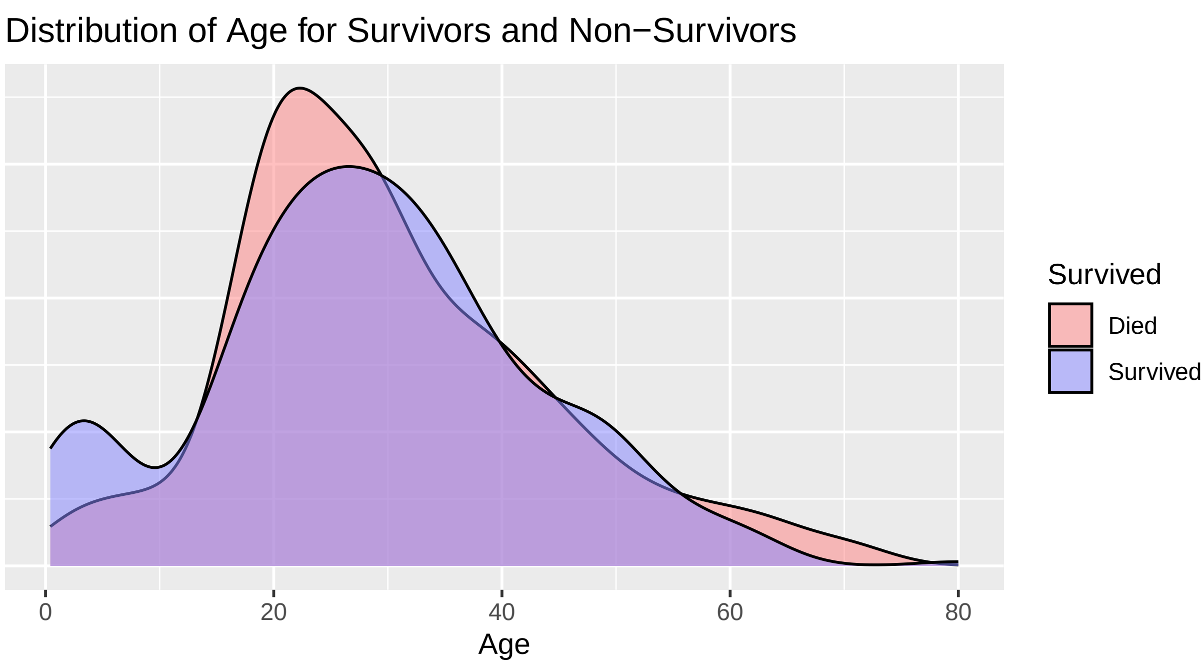


Question 5

Create a plot (a comparative boxplot for example, but feel free to be creative) to compare age distributions of those who survived and those who did not. Write a few sentences summarizing what you can conclude from your display only.

```
ggplot(titanic, aes(Age)) +
    geom_density(aes(fill = ifelse(Survived == 1, "Survived", "Died"))) +
    scale_fill_manual(
```

```
        name = "Survived",
        values = setNames(
            c("#ff888888", "#8888ff88"),
            c("Died", "Survived")
        )
    ) +
    labs(
        x = "Age",
        title = "Distribution of Age for Survivors and Non-Survivors"
    ) +
    theme(
        axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        axis.title.y = element_blank()
    )
```

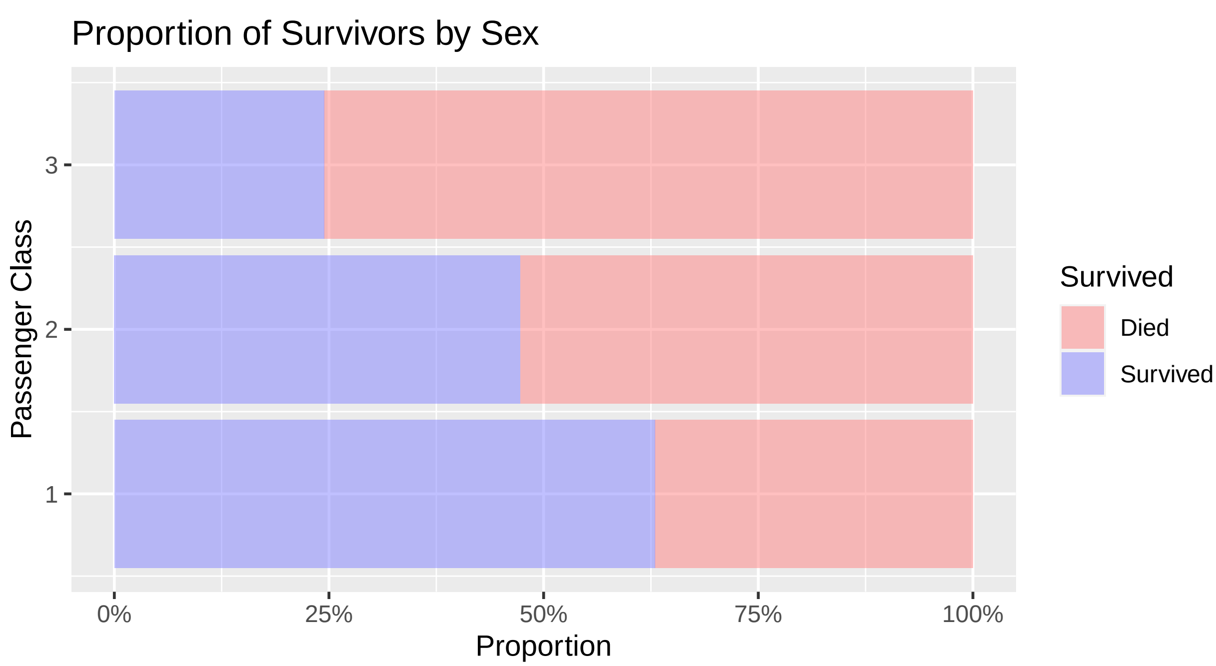


The distributions are somewhat similar in shape, but there are a lot more old non-survivors and a lot more young survivors.

Question 6

Create a plot (a comparative bar chart for example, once again there are many options) to compare survival numbers/percentages versus non-survival numbers per ticket class. Write a few sentences summarizing what you can conclude from your display only.

```r
ggplot(titanic, aes(Pclass)) +
    geom_bar(
        aes(fill = ifelse(Survived == 1, "Survived", "Died")),
        position = "fill"
    ) +
    coord_flip() +
    scale_y_continuous(labels = scales::percent) +
    scale_fill_manual(
        name = "Survived",
        values = setNames(
            c("#ff888888", "#8888ff88"),
            c("Died", "Survived")
        )
    ) +
    labs(
        x = "Passenger Class",
        y = "Proportion",
        title = "Proportion of Survivors by Sex"
    )
```



The chart shows that the passengers in more prestigious passenger classes had a higher probability of surviving than those in less prestigious passenger classes.

Question 7

Fit a logistic model using age to predict survival. Is there a statistically significant relationship between these two variables? If so, in what direction is the relationship? Interpret the magnitude of the relationship in context.

```r
regres01 <- glm(
    Survived ~ Age,
    data = titanic,
    family = binomial("logit")
)
summary(regres01)
```

```
##
## Call:
## glm(formula = Survived ~ Age, family = binomial("logit"), data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0864  -1.0017  -0.9439   1.3562   1.5806
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.209189   0.159494  -1.312   0.1897
## Age         -0.008774   0.004947  -1.774   0.0761 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.8  on 886  degrees of freedom
## Residual deviance: 1179.6  on 885  degrees of freedom
## AIC: 1183.6
##
## Number of Fisher Scoring iterations: 4
```

$$P(\text{Survived}) = \frac{e^u}{1+e^u} \text{ where } u = -0.209189 - 0.008774(\text{Age})$$

The relationship is not statistically significant at the 0.05 level since $p = 0.0761 > \alpha = 0.05$. Since $\beta_1 = -0.008774$, we can conclude that for every 1 year increase in age, the *odds* of survival are multiplied by $e^{-0.008774}$.

Question 8

Using your model, what is the probability that a passenger aged 48 survived the Titanic? What about a 14 year old? 78 year old? 5 year old?

```
create_predict_logreg_function <- function(model) {
    function(x) {
        (function(u) {
            exp(u) / (1 + exp(u))
        })(
            summary(model)$coefficients[1, 1] +
            summary(model)$coefficients[2, 1] * x
        )
    }
}

predict_regres01 <- create_predict_logreg_function(regres01)

printf("48: %.7f", predict_regres01(48))
```

```
## [1] "48: 0.3474294"
```

```
printf("14: %.7f", predict_regres01(14))
```

```
## [1] "14: 0.4177468"
```

```
printf("78: %.7f", predict_regres01(78))
```

```
## [1] "78: 0.2903699"
```

```
printf("5: %.7f", predict_regres01(5))
```

```
## [1] "5: 0.4370703"
```

$$
\begin{aligned}
P(\text{Survived} \mid \text{Age} = 48) &= 0.3474294 \\
P(\text{Survived} \mid \text{Age} = 14) &= 0.4177468 \\
P(\text{Survived} \mid \text{Age} = 78) &= 0.2903699 \\
P(\text{Survived} \mid \text{Age} = 5) &= 0.4370703
\end{aligned}
$$

Question 9

Fit a logistic model using gender to predict survival. Is there a statistically significant relationship between these two variables? If so, in what direction is the relationship? Interpret the magnitude of the relationship in context.

```r
regres02 <- glm(
    Survived ~ Sex,
    data = titanic,
    family = binomial("logit")
)
summary(regres02)
```

```
##
## Call:
## glm(formula = Survived ~ Sex, family = binomial("logit"), data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6462  -0.6496  -0.6496   0.7725   1.8218
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0566     0.1290   8.191 2.58e-16 ***
## Sexmale      -2.5051     0.1672 -14.980  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.77  on 886  degrees of freedom
## Residual deviance:  916.12  on 885  degrees of freedom
## AIC: 920.12
##
## Number of Fisher Scoring iterations: 4
```

$$P(\text{Survived}) = \frac{e^u}{1+e^u} \text{ where } u = 1.0566 - 2.5051(\text{Male})$$

The relationship between these variables is significant at the 0.05 level because $p = 9.9476 \times 10^{-51} < \alpha = 0.05$. The direction of the relationship is negative. Since $\beta_1 = -2.5051$, we can conclude that the *odds* of a male surviving are $e^{-2.5051}$ times the odds of a female surviving.

Question 10

Using your model, what is the probability that a female survived the titanic? How does this compare to your answer in question 1)?

```r
predict_regres02 <- create_predict_logreg_function(regres02)
```

```r
printf("female: %.7f", predict_regres02(0))
```

```
## [1] "female: 0.7420382"
```

$$P(\text{Survived} \mid \text{Female}) = 0.7420382$$

This answer is almost exactly equal to my answer from question 1). The error is only 0.00001.

Question 11

Fit a logistic model using class to predict survival. Is there a statistically significant relationship between these two variables? If so, in what direction is the relationship? Interpret the magnitude of the relationship in context.

```r
regres03 <- glm(
    Survived ~ Pclass,
    data = titanic,
    family = binomial("logit")
)
summary(regres03)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass, family = binomial("logit"),
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4382  -0.7602  -0.7602   0.9374   1.6629
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.43907    0.20742   6.938 3.98e-12 ***
## Pclass      -0.84423    0.08719  -9.683  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.8  on 886  degrees of freedom
## Residual deviance: 1082.1  on 885  degrees of freedom
## AIC: 1086.1
##
## Number of Fisher Scoring iterations: 4
```

$$P(\text{Survived}) = \frac{e^u}{1+e^u} \text{ where } u = 1.43907 - 0.84423(\text{Class})$$

The relationship beteen these two variables is significant at the 0.05 level because $p = 3.57423 \times 10^{-22} < \alpha = 0.05$. The direction of the relationship is negative. Since $\beta_1 = -0.84423$, we can conclude that the odds of survival are multiplied by $e^{-0.84423}$ for every numeric increase in class (i.e. 1st $\rightarrow$ 2nd, 2nd $\rightarrow$ 3rd).

Question 12

```
predict_regres03 <- create_predict_logreg_function(regres03)

printf("1st: %.7f", predict_regres03(1))
```

```
## [1] "1st: 0.6444743"
```

```
printf("3rd: %.7f", predict_regres03(3))
```

```
## [1] "3rd: 0.2509373"
```

$$P(\text{Survived} \mid \text{1st Class}) = 0.6444743$$
$$P(\text{Survived} \mid \text{3rd Class}) = 0.2509373$$

These answers were pretty close to what I got in question 1).

Question 13

Using the variables gender, class and age, explore a logistic regression model that applies all three variables. You may also look at interactive terms. After you find the best model, use it to predict the survival probabilities of:

a) *A female, 1st class passenger, 58 years old.*
b) *A male, 3rd class passenger, 34 years old.*
c) *A male, 2nd class passenger, 44 years old.*
d) *A female, 3rd class passenger, 12 years old.*

(Show your chosen equation, justification on why you chose it, and your work answering parts a-d above. You can justify it using a table showing your progression on finding the best model.)

```
regres04 <- glm(
    Survived ~
```

```r
        Age +
        Sex +
        Pclass,
    data = titanic,
    family = binomial("logit")
)
summary(regres04)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass, family = binomial("logit"),
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6858  -0.6588  -0.4102   0.6386   2.4493
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.878511   0.463474  10.526  < 2e-16 ***
## Age         -0.034361   0.007134  -4.816 1.46e-06 ***
## Sexmale     -2.589163   0.186933 -13.851  < 2e-16 ***
## Pclass      -1.230538   0.124957  -9.848  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.77  on 886  degrees of freedom
## Residual deviance:  801.61  on 883  degrees of freedom
## AIC: 809.61
##
## Number of Fisher Scoring iterations: 5
```

```r
regres05 <- glm(
    Survived ~
        Age *
        Sex *
        Pclass,
    data = titanic,
    family = binomial("logit")
)
summary(regres05)
```

```
##
```

```
## Call:
## glm(formula = Survived ~ Age * Sex * Pclass, family = binomial("logit"),
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9201  -0.6417  -0.4397   0.5002   2.5186
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          5.592312   2.162247   2.586   0.0097 **
## Age                  0.020447   0.064164   0.319   0.7500
## Sexmale             -3.984800   2.315690  -1.721   0.0853 .
## Pclass              -1.711391   0.754931  -2.267   0.0234 *
## Age:Sexmale         -0.050170   0.067746  -0.741   0.4590
## Age:Pclass          -0.013077   0.023043  -0.567   0.5704
## Sexmale:Pclass       0.973958   0.827834   1.177   0.2394
## Age:Sexmale:Pclass   0.004287   0.025314   0.169   0.8655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.77  on 886  degrees of freedom
## Residual deviance:  772.02  on 879  degrees of freedom
## AIC: 788.02
##
## Number of Fisher Scoring iterations: 6
```

```r
regres06 <- glm(
    Survived ~
        Age +
        Sex +
        Pclass +
        Sex : Pclass,
    data = titanic,
    family = binomial("logit")
)
summary(regres06)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + Sex:Pclass, family = binomial("logit"),
##     data = titanic)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2772  -0.6577  -0.4728   0.4647   2.3389
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     7.741802   0.925642   8.364  < 2e-16 ***
## Age            -0.035537   0.007378  -4.817 1.46e-06 ***
## Sexmale        -6.087233   0.889970  -6.840 7.93e-12 ***
## Pclass         -2.305385   0.308926  -7.463 8.48e-14 ***
## Sexmale:Pclass  1.397528   0.325970   4.287 1.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.77  on 886  degrees of freedom
## Residual deviance:  777.08  on 882  degrees of freedom
## AIC: 787.08
##
## Number of Fisher Scoring iterations: 6
```

```
regres07 <- glm(
    Survived ~
        Age +
        Sex +
        Pclass +
        Age : Sex +
        Sex : Pclass,
    data = titanic,
    family = binomial("logit")
)
summary(regres07)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + Age:Sex + Sex:Pclass,
##     family = binomial("logit"), data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0295  -0.6356  -0.4534   0.4852   2.4379
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)     6.72546    1.01047   6.656 2.82e-11 ***
## Age            -0.01531    0.01225  -1.250  0.21146
## Sexmale        -4.57263    1.13801  -4.018 5.87e-05 ***
## Pclass         -2.11598    0.31292  -6.762 1.36e-11 ***
## Age:Sexmale    -0.03109    0.01544  -2.014  0.04400 *
## Sexmale:Pclass  1.12125    0.34714   3.230  0.00124 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.77  on 886  degrees of freedom
## Residual deviance:  773.05  on 881  degrees of freedom
## AIC: 785.05
##
## Number of Fisher Scoring iterations: 6
```

```r
regres08 <- glm(
    Survived ~
        Age +
        Sex +
        Pclass +
        Age : Sex +
        Sex : Pclass +
        Age : Pclass,
    data = titanic,
    family = binomial("logit")
)
summary(regres08)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + Age:Sex + Sex:Pclass +
##     Age:Pclass, family = binomial("logit"), data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8810  -0.6427  -0.4385   0.4967   2.5256
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     5.891199   1.283274   4.591 4.42e-06 ***
## Age             0.010754   0.028768   0.374 0.708533
## Sexmale        -4.328793   1.150043  -3.764 0.000167 ***
## Pclass         -1.818275   0.425335  -4.275 1.91e-05 ***
```

```
## Age:Sexmale    -0.039092   0.017488  -2.235 0.025396 *
## Sexmale:Pclass  1.102295   0.346900   3.178 0.001485 **
## Age:Pclass     -0.009530   0.009519  -1.001 0.316733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.77  on 886  degrees of freedom
## Residual deviance:  772.05  on 880  degrees of freedom
## AIC: 786.05
##
## Number of Fisher Scoring iterations: 6
```

Model	AIC
$P(\text{Survived}) = 4.8785 - 0.0344(\text{Age}) - 2.5892(\text{Male}) - 1.2305(\text{Class})$	809.6127
$P(\text{Survived}) = 5.5923 - 0.0204(\text{Age}) - 3.9848(\text{Male}) - 1.7114(\text{Class}) - 0.0502(\text{Age}\times\text{Male}) - 0.0131(\text{Age}\times\text{Class}) + 0.9740(\text{Male}\times\text{Class}) + 0.0043(\text{Age}\times\text{Male}\times\text{Class})$	788.0208
$P(\text{Survived}) = 7.7418 - 0.0355(\text{Age}) - 6.0872(\text{Male}) - 2.3054(\text{Class}) + 1.3972(\text{Male}\times\text{Class})$	787.0783
$P(\text{Survived}) = 6.7255 - 0.0153(\text{Age}) - 4.5726(\text{Male}) - 2.1160(\text{Class}) - 0.0311(\text{Age}\times\text{Male}) + 1.1213(\text{Male}\times\text{Class})$	785.0504
$P(\text{Survived}) = 5.8912 + 0.0107(\text{Age}) - 4.3288(\text{Male}) - 1.8183(\text{Class}) - 0.0391(\text{Age}\times\text{Male}) + 1.1023(\text{Male}\times\text{Class}) - 0.0095(\text{Age}\times\text{Class})$	786.0496

The model I chose was:

$$P(\text{Survived}) = 6.7255 - 0.0153(\text{Age}) - 4.5726(\text{Male}) - 2.1160(\text{Class}) - 0.0311(\text{Age}\times\text{Male}) + 1.1213(\text{Male}\times\text{Class})$$

I chose this model because it had the highest AIC of all the models I tested.

```r
create_predict_multiple_logreg_function <- function(model) {
    function(df) {
        predict.glm(model, newdata = df, type = "response")
    }
}

predict_regres07 <- create_predict_multiple_logreg_function(regres07)

printf(
    "(F, 1, 58): %.7f",
    predict_regres07(data.frame(Sex = "female", Pclass = 1, Age = 58))
)
```

```
## [1] "(F, 1, 58): 0.9763794"
```

```r
printf(
    "(M, 3, 34): %.7f",
    predict_regres07(data.frame(Sex = "male", Pclass = 3, Age = 34))
)
```

```
## [1] "(M, 3, 34): 0.0824994"
```

```r
printf(
    "(M, 2, 44): %.7f",
    predict_regres07(data.frame(Sex = "male", Pclass = 2, Age = 44))
)
```

```
## [1] "(M, 2, 44): 0.1326055"
```

```r
printf(
    "(F, 3, 12): %.7f",
    predict_regres07(data.frame(Sex = "female", Pclass = 3, Age = 12))
)
```

```
## [1] "(F, 3, 12): 0.5483137"
```

$$
\begin{aligned}
P(\text{Survived} \mid \text{Female} \wedge \text{Class} = 1 \wedge \text{Age} = 58) &= 0.9763794 \\
P(\text{Survived} \mid \text{Male} \wedge \text{Class} = 3 \wedge \text{Age} = 34) &= 0.0824994 \\
P(\text{Survived} \mid \text{Male} \wedge \text{Class} = 2 \wedge \text{Age} = 44) &= 0.1326055 \\
P(\text{Survived} \mid \text{Female} \wedge \text{Class} = 3 \wedge \text{Age} = 12) &= 0.5483137
\end{aligned}
$$