# DATA 603 Assignment 3

Michael Ellsworth

November 29th, 2019

# Problem 1

*A study investigated characteristics associated with y = whether a cancer patient achieved remission (1=yes, 0=no). An important explanatory variable was a labeling index (LI=percentage of “labeled” cells) that measures proliferative activity of cells after a patient receives an injection of tritiated thymidine. Fit a logistic regression model in order to answer the following questions. The data is provided in remission.scv file.*

## a

*Estimate the probabilities of a cancer patient achieved remission when LI=15 and LI=37. Comment on your results.*

```
remission_model_a <- glm(data = remission, y ~ LI, family = "binomial")
summary(remission_model_a)
```

```
##
## Call:
## glm(formula = y ~ LI, family = "binomial", data = remission)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.15450  -0.52277   0.09487   0.39040   2.42125
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.01530    0.68349  -10.26   <2e-16 ***
## LI           0.31838    0.03061   10.40   <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: 553.31  on 399  degrees of freedom
## Residual deviance: 266.84  on 398  degrees of freedom
## AIC: 270.84
##
## Number of Fisher Scoring iterations: 6
```

```
predict(remission_model_a, data.frame(LI = 15), type = "response")
```

```
##          1
## 0.09624499
```

```
predict(remission_model_a, data.frame(LI = 37), type = "response")
```

```
##         1
## 0.9915459
```

Based on the P-value of the summary of the logistic regression model where LI is a predictor of remission, we can say with confidence that LI is a predictor of remission as the P-value is less than 0.05.

If the percentage of “labeled” cells is 15%, the probability that the cancer patient has achieved remission is 0.096. If the percentage is 37%, the probability of remission is 0.992.

## b

*Interpret the effect of LI in terms of the odds.*

```
coefficients(remission_model_a)
```

```
## (Intercept)          LI
##  -7.0153000   0.3183759
```

$$e^{\beta_0+\beta_1*LI} = e^{-7.015+0.318*LI}$$

```
exp(0.3183759)
```

```
## [1] 1.374893
```

For every 1% increase in LI, the odds of remission is multiplied by 1.375 or, there is an increase of 37.5% of the odds of remission.

## c

*Compute 95% confidence intervals for the logistic regression coefficient and its associated odds. Give an interpretation.*

```
confint(remission_model_a)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %     97.5 %
## (Intercept) -8.4561952 -5.7656003
## LI           0.2626089  0.3831168
```

The 95% confidence interval for $\beta_1$ is 0.2626 and 0.3831. This suggests that an increase in LI by 1% would increase the log odds of remission between 0.2626 and 0.3831.

```
exp(0.2626089)
```

```
## [1] 1.300318
```

```
exp(0.3831168)
```

```
## [1] 1.466849
```

```
exp(confint(remission_model_a))
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %      97.5 %
## (Intercept) 0.0002125794 0.003133514
## LI          1.3003180838 1.466849322
```

The 95% confidence interval for $e^{\widehat{\beta_1}}$ is 1.3003 and 1.4668. This suggests that the relationship between LI and remission is positive as the confidence interval covers a range greater than 1. An increase in LI by 1% would increase the odds of remission between 30% and 47%.

## d

*Use the Wald Z test and Likelihood ratio test to confirm that the LI predictor is associated with remission of cancer at alpha = 0.05.*

```
wald.test(b = coef(remission_model_a), Sigma = vcov(remission_model_a), Terms = 2)
```

```
## Wald test:
## ----------
## 
## Chi-squared test:
## X2 = 108.2, df = 1, P(> X2) = 0.0
```

```
remission_model_intercept <- glm(data = remission, y ~ 1, family = "binomial")
lrtest(remission_model_intercept, remission_model_a)
```

| | #Df | LogLik | Df | Chisq | Pr(>Chisq) |
|---|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 1 | -276.6536 | NA | NA | NA |
| 2 | 2 | -133.4202 | 1 | 286.4666 | 2.927255e-64 |

2 rows

Based on a P-value less than 0.05 for both the Wald Z test and the Likelihood Ratio test, we can confirm that LI is a predictor of remission.

# Problem 2

*The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. The RMS Titanic was the largest ship a float at the time it entered service and was the second of three Olympic-class ocean liners operated by the White Star Line. The Titanic was built by the Harland and Wolff shipyard in Belfast. Thomas Andrews, her architect, died in the disaster. The training-dataset with 891 examples is provided in titanic.csv file and the list of Variables with a short description is provided:*

- *Survived: Survival*
- *PassengerId: Unique Id of a passenger. pclass: Ticket class*
- *Pclass: Ticket class*
- *sex: Sex*
- *Age: Age in years*

## a

*Test if the chances of survival of passengers in Titanic depends on those variables at alpha = 0.05*

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

```
## 
## Call:
## glm(formula = Survived ~ PassengerId + Pclass + Sex + Age, family = "binomial", 
##     data = titanic)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2259  -0.7183  -0.4399   0.6442   2.2287  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  2.129e+00  3.151e-01   6.757 1.41e-11 ***
## PassengerId  8.937e-05  3.383e-04   0.264 0.791652    
## Pclass2     -8.443e-01  2.449e-01  -3.448 0.000566 ***
## Pclass3     -1.913e+00  2.149e-01  -8.904  < 2e-16 ***
## Sexmale     -2.649e+00  1.848e-01 -14.339  < 2e-16 ***
## Age          4.523e-03  6.097e-03   0.742 0.458166    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  826.27  on 885  degrees of freedom
## AIC: 838.27
## 
## Number of Fisher Scoring iterations: 4
```

Age and PassengerId are not predictors of survival as they are insignificant as shown by the summary above.

To confirm that survival is dependent on passenger class and sex, we can use the wald z test.

```
titanic_model_reduced <- glm(data = titanic, Survived ~ Pclass + Sex, family = "binomial")
wald.test(b = coef(titanic_model_reduced), Sigma = vcov(titanic_model_reduced), Terms = 2:3)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 82.1, df = 2, P(> X2) = 0.0
```

Since the P-value is less than 0.05, we can confirm that survival is dependent on passenger class and sex.

# b

*From part a), use the likelihood ratio test to check whether the variable age should be in the full model.*

```
lrtest(titanic_model_reduced, titanic_model)
```

| | #Df | LogLik | Df | Chisq | Pr(>Chisq) |
|---|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 4 | -413.4442 | NA | NA | NA |
| 2 | 6 | -413.1365 | 2 | 0.6153969 | 0.735137 |

2 rows

Since the P-value calculated from the likelihood ratio test is above 0.05, we can say that the larger model does not perform better than the reduced model, or we failed to reject $H_0$. In other words, the model without Age performs better than the model with Age. Age should not be in the full model.

# c

*Write the logit and logistic regression model for predicting the chances of survival of passengers in Titanic.*

```
titanic_model_reduced_c <- glm(data = titanic, Survived ~ Pclass + Sex, family = "binomial")
coefficients(titanic_model_reduced_c)
```

```
## (Intercept)     Pclass2     Pclass3     Sexmale
##   2.2971232  -0.8379523  -1.9054951  -2.6418754
```

```
summary(titanic_model_reduced_c)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex, family = "binomial", data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1877  -0.7312  -0.4476   0.6465   2.1681
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.2971     0.2190  10.490  < 2e-16 ***
## Pclass2      -0.8380     0.2447  -3.424 0.000618 ***
## Pclass3      -1.9055     0.2141  -8.898  < 2e-16 ***
## Sexmale      -2.6419     0.1841 -14.351  < 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: 1186.66  on 890  degrees of freedom
## Residual deviance:  826.89  on 887  degrees of freedom
## AIC: 834.89
##
## Number of Fisher Scoring iterations: 4
```

Logit:

$$\widehat{logit} = \begin{cases} 2.297 - 2.642 & \text{if Passenger Class is 1 and Sex is Male} \\ 2.297 - 0.838 - 2.642 & \text{if Passenger Class is 2 and Sex is Male} \\ 2.297 - 1.905 - 2.642 & \text{if Passenger Class is 3 and Sex is Male} \\ 2.297 & \text{if Passenger Class is 1 and Sex is Female} \\ 2.297 - 0.838 & \text{if Passenger Class is 2 and Sex is Female} \\ 2.297 - 1.905 & \text{if Passenger Class is 3 and Sex is Female} \end{cases}$$

Logistic:

$$\widehat{\pi} = \begin{cases} \frac{e^{2.297-2.642}}{1+e^{2.297-2.642}} & \text{if Passenger Class is 1 and Sex is Male} \\ \frac{e^{2.297-0.838-2.642}}{1+e^{2.297-0.838-2.642}} & \text{if Passenger Class is 2 and Sex is Male} \\ \frac{e^{2.297-1.905-2.642}}{1+e^{2.297-1.905-2.642}} & \text{if Passenger Class is 3 and Sex is Male} \\ \frac{e^{2.297}}{1+e^{2.297}} & \text{if Passenger Class is 1 and Sex is Female} \\ \frac{e^{2.297-0.838}}{1+e^{2.297-0.838}} & \text{if Passenger Class is 2 and Sex is Female} \\ \frac{e^{2.297-1.905}}{1+e^{2.297-1.905}} & \text{if Passenger Class is 3 and Sex is Female} \end{cases}$$

## d

*Interpret the effect of Class and Sex in terms of the odds ratio from the logistic regression model in part c)*

```
exp(coefficients(titanic_model_reduced_c))
```

```
## (Intercept)     Pclass2     Pclass3     Sexmale 
##  9.94552998  0.43259543  0.14874898  0.07122757
```

If all other variables are held constant:

- The odds of survival of passengers in class 2 are 0.432 times the odds for passengers in class 1
- The odds of survival of passengers in class 3 are 0.149 times the odds for passengers in class 1
- The odds of survival of passengers in class 3 are 0.344 times the odds for passengers in class 2
- The odds of survival of male passengers are 0.071 times the odds for female passengers

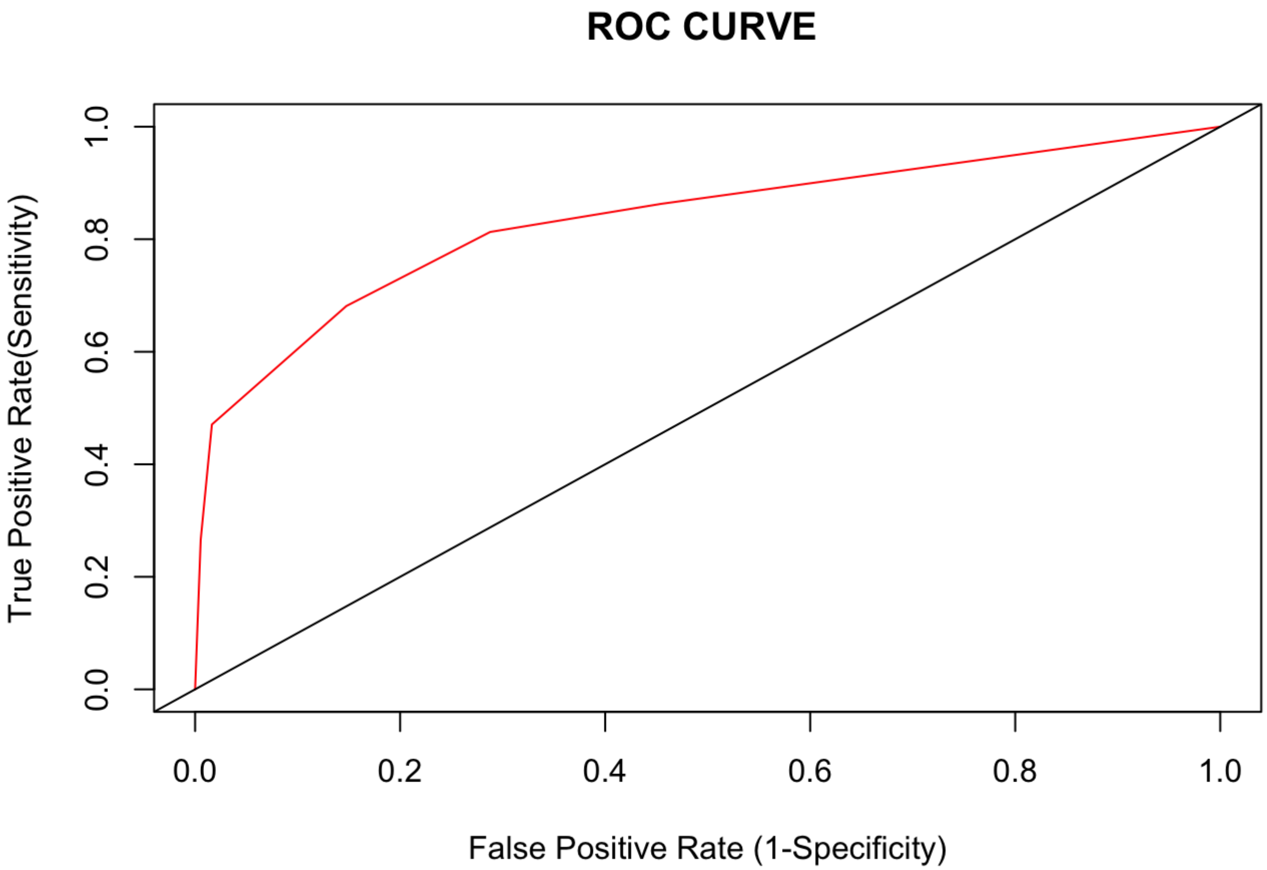
## e

*Report Deviance, AIC, ROC with AUC (the model fit) from the model in part c)*

The deviance of the model from part c is 826.89.

The AIC of the model from part c is 834.89.

```
prob_e <- predict(titanic_model_reduced_c, type = c("response"))
pred_e <- prediction(prob_e, titanic$Survived)
perf_e <- performance(pred_e, measure = "tpr", x.measure = "fpr")
plot(perf_e, col = 2, main = "ROC CURVE ", xlab = "False Positive Rate (1-Specificity)", ylab = "True Positive Ra
te(Sensitivity)")
abline(0,1)
```



```
auc(roc(titanic$Survived, prob_e))
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
## Area under the curve: 0.8328
```

AUC is 0.8328

## f

*Build the logistic regression with interation terms and write the logit model. Use the likelihood ratio test to confirm your result.*

```
titanic_model_interact <- glm(data = titanic, Survived ~ (Pclass + Sex)**2, family = "binomial")
summary(titanic_model_interact)
```

```
##
## Call:
## glm(formula = Survived ~ (Pclass + Sex)^2, family = "binomial",
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6248  -0.5853  -0.5395   0.4056   1.9996
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.4122     0.5868   5.815 6.06e-09 ***
## Pclass2          -0.9555     0.7248  -1.318  0.18737
## Pclass3          -3.4122     0.6100  -5.594 2.22e-08 ***
## Sexmale          -3.9494     0.6161  -6.411 1.45e-10 ***
## Pclass2:Sexmale  -0.1850     0.7939  -0.233  0.81575
## Pclass3:Sexmale   2.0958     0.6572   3.189  0.00143 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance:  798.1  on 885  degrees of freedom
## AIC: 810.1
##
## Number of Fisher Scoring iterations: 6
```

Based on the summary of the interaction model, we can say that the interaction term of passenger class and sex is significant.

```
lrtest(titanic_model_reduced_c, titanic_model_interact)
```

| | #Df | LogLik | Df | Chisq | Pr(>Chisq) |
|---|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 4 | -413.4442 | NA | NA | NA |
| 2 | 6 | -399.0484 | 2 | 28.79147 | 5.597724e-07 |

2 rows

Based on the P-value of the likelihood ratio, our assumption that the full model with the interaction term performs better than the reduced model without the interaction term is confirmed.

```
coefficients(titanic_model_interact)
```

```
##     (Intercept)         Pclass2         Pclass3         Sexmale 
##       3.4122472      -0.9555114      -3.4122472      -3.9493901 
## Pclass2:Sexmale Pclass3:Sexmale 
##      -0.1849918       2.0957553
```

Logit:

$$\widehat{logit} = \begin{cases} 3.412 - 3.949 & \text{if Passenger Class is 1 and Sex is Male} \\ 3.412 - 0.956 - 3.949 - 0.185 & \text{if Passenger Class is 2 and Sex is Male} \\ 3.412 - 3.412 - 3.949 + 2.096 & \text{if Passenger Class is 3 and Sex is Male} \\ 3.412 & \text{if Passenger Class is 1 and Sex is Female} \\ 3.412 - 0.956 & \text{if Passenger Class is 2 and Sex is Female} \\ 3.412 - 3.412 & \text{if Passenger Class is 2 and Sex is Female} \end{cases}$$

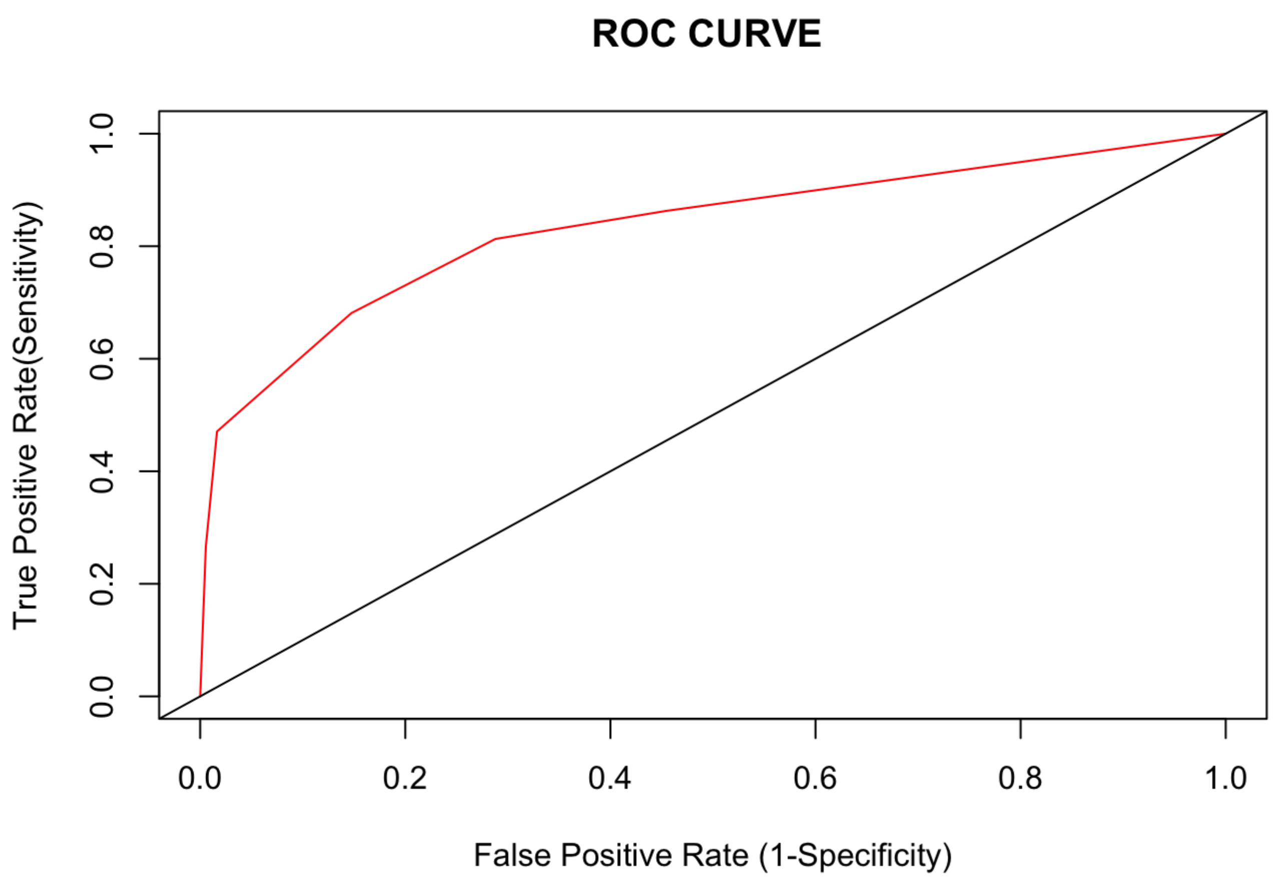
# g

*Report Deviance, AIC, ROC with AUC (the model fit) from the model in part e) and compare the result with the model in part c).*

Deviance: 798.1

AIC: 810.1

```
prob_g <- predict(titanic_model_interact, type = c("response"))
pred_g <- prediction(prob_g, titanic$Survived)
perf_g <- performance(pred_g, measure = "tpr", x.measure = "fpr")
plot(perf_g, col = 2, main = "ROC CURVE ", xlab = "False Positive Rate (1-Specificity)", ylab = "True Positive Ra
te(Sensitivity)")
abline(0,1)
```



```
auc(roc(titanic$Survived, prob_g))
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.8328
```

AUC is 0.8328.

Since the AIC is lower in the interaction model when compared to the model from part c, we can say that the interaction model is a better fit.

# h

*From the model in part e), predict the probability of survival for a 35 years old man who got the third class ticket. Show your work by substituting the effect values and use R command to confirm your result. Comment on your result.*

```
coefficients(titanic_model_interact)
```

```
##     (Intercept)         Pclass2         Pclass3         Sexmale
##       3.4122472      -0.9555114      -3.4122472      -3.9493901
## Pclass2:Sexmale Pclass3:Sexmale
##      -0.1849918       2.0957553
```

Using the coefficients for Male and Passenger Class 3:

```
(exp(3.4122472 - 3.4122472 - 3.9493901 + 2.0957553)) / (1 + exp(3.4122472 - 3.4122472 - 3.9493901 + 2.0957553))
```

```
## [1] 0.1354467
```

Using the predict function to confirm the result:

```
newdata_h = data.frame(Sex = "male", Pclass = "3")
predict(titanic_model_interact, newdata_h, type = "response")
```

```
##         1
## 0.1354467
```

From this result, we can say the probability of survivial for a 35 year old man with a third class ticket is 0.135. This model does not include age and hence, it does not change the probability of survival.

# i

*From the model in part e), predict the probability of survival for a 20 years old man who got the first class ticket. Show your work by substituting the effect values and use R command to confirm your result. Comment on your result.*

Using the coefficients for Male and Passenger Class 1:

```
(exp(3.4122472 - 3.9493901)) / (1 + exp(3.4122472 - 3.9493901))
```

```
## [1] 0.3688525
```

Using the predict function to confirm the result:

```
newdata_i = data.frame(Sex = "male", Pclass = "1")
predict(titanic_model_interact, newdata_i, type = "response")
```

```
##         1
## 0.3688525
```

From this result, we can say the probability of survivial for a 20 year old man with a first class ticket is 0.369. This model does not include age and hence, it does not change the probability of survival.