Assessment #6
Beiming Zhang

Conceptual Part

Question 1

a)

- The probability that Shengqi will default is:

$$\hat{y} = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2+\widehat{\beta_3}x_3}}{1+e^{\widehat{\beta_0}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2+\widehat{\beta_3}x_3}} = \frac{e^{-1+0.05x_1-0.01x_2-0.5x_3}}{1+e^{-1+0.05x_1-0.01x_2-0.5x_3}}$$

$$\hat{y} = \frac{e^{-1+0.05\times5-0.01\times25-0.5\times1}}{1+e^{-1+0.05\times5-0.01\times25-0.5\times1}} = \frac{e^{-1.5}}{1+e^{-1.5}} = 0.1824$$

b)

- The odds that Shengqi will default is:

$$\frac{\hat{y}}{1-\hat{y}} = e^{\widehat{\beta_0}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2+\widehat{\beta_3}x_3} = e^{-1.5} = 0.2231$$

Programming Part

Question 1

a)

Output:

Train dataset:

0	1
0.6135	0.3865

Test dataset:

0	1
0.6188	0.3812

- The proportion of passengers who survived in the training data is 38.65%, and the proportion of passengers who survived in the test data is 38.12%.

b)

Output:

```
Call:
glm(formula = Survived ~ Gender + Child + Fare + Class, family = binomial,
    data = titanic_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1755  -0.6942  -0.4369   0.6527   2.4596

Coefficients:
```

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.8123     0.5608    5.015 5.32e-07 ***
Gendermale     -2.7384     0.2775   -9.866  < 2e-16 ***
Child           1.0362     0.3754    2.760  0.00577 **
Fare10--20     -0.4844     0.3973   -1.219  0.22275
Fare20--30     -0.6740     0.4351   -1.549  0.12136
Fare30+        -0.5443     0.4506   -1.208  0.22707
ClassThree     -2.3751     0.4573   -5.193 2.06e-07 ***
ClassTwo       -0.8897     0.4405   -2.020  0.04342 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 593.78  on 444  degrees of freedom
Residual deviance: 415.32  on 437  degrees of freedom
AIC: 431.32

Number of Fisher Scoring iterations: 4
```

- The lowest p-value corresponding to Gender is less than 0.001, the lowest p-value corresponding to Child is less than 0.01, and the lowest p-value corresponding to Class is less than 0.05. Therefore, we can consider the variables Gender, Child, and Class to be statistically significant.
- Conversely, the p-values corresponding to Fare are all greater than 0.1, hence we can consider the Fare variable to be not statistically significant.

c)

Output:

```
Call:
glm(formula = Survived ~ Gender + Child + Class, family = binomial,
    data = titanic_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1338  -0.7034  -0.4080   0.6807   2.2483

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.1681     0.3111    6.969 3.20e-12 ***
Gendermale     -2.6148     0.2623   -9.970  < 2e-16 ***
Child           0.8769     0.3557    2.466   0.0137 *
ClassThree     -1.9976     0.3153   -6.335 2.38e-10 ***
```

ClassTwo -0.8238 0.3442 -2.393 0.0167 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 593.78 on 444 degrees of freedom
Residual deviance: 418.06 on 440 degrees of freedom
AIC: 428.06

Number of Fisher Scoring iterations: 4

- All the p-values of the remaining variables are less than 0.05, therefore the remaining variables are statistically significant.

d)

Output (first of 15):

4	6	9	10	11
0.89735099	0.07985625	0.54253048	0.90214255	0.74028489
12	13	14	17	18
0.89735099	0.07985625	0.07985625	0.17258936	0.21917324
21	23	24	26	30
0.21917324	0.74028489	0.39014467	0.54253048	0.07985625

[1] 1 0 1 1 1 1 0 0 0 0 0 1 0 1 0

- The survival rates of each passenger and the final prediction can be seen in the running results in R.

e)

Output:

	0	1
0	0.5381	0.0807
1	0.1166	0.2646

[1] 0.8027

- The percentage of passengers for whom my prediction was accurate is 80.27%.