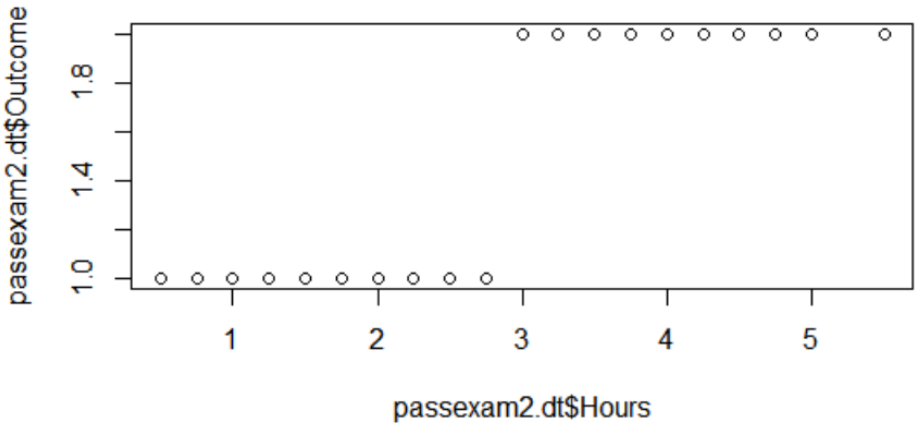
Solution to Exercise 7.1 Logistic Regression

Logistic Regression on Binary Y:

1. Re-run the Rscript passexam.R on passexam2.csv dataset. What is the cause of the error? Explain.

   Solution in Rscript: passexam2.R. Cause of Error: Perfect Separation of Data.



2. Execute logistic regression on default.csv dataset to predict default:

   a. Verify the baseline reference level for default.

      levels(default.dt$Default) # Baseline is Default = "No"

   b. Which variables are statistically insignificant?

```
21  m1 <- glm(Default ~ . , family = binomial, data = default.dt)
22
23  summary(m1)
```

```
Call:
glm(formula = Default ~ ., family = binomial, data = default.dt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1419  -0.0557  -0.0203   3.7390

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.152e+01  4.379e-01 -26.300  < 2e-16 ***
GenderM      6.471e-01  2.363e-01   2.739  0.00616 **
AvgBal       5.737e-03  2.319e-04  24.737  < 2e-16 ***
Income       3.022e-06  8.203e-06   0.368  0.71260
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

   Gender and avg balance are sig. Income is not sig in the presence of avg bal and gender.

c. Keeping only statistically significant variables, show the confusion matrix.

```
> table1
        m2.predict
Actual    No   Yes
   No   9628    39
   Yes   228   105
```

```
> round(prop.table(table1),3)
        m2.predict
Actual    No   Yes
   No  0.963 0.004
   Yes 0.023 0.011
```

d. Using set.seed(2) with 70-30 train-test splt, and keeping only statistically significant variables, show the trainset confusion matrix and testset confusion matrix.

```
> table3
                 m4.predict.train
Trainset.Actual    No  Yes
            No   6741   26
            Yes   156   77
```

```
> table4
                m4.predict.test
Testset.Actual    No  Yes
            No  2889   11
            Yes   72   28
```

e. An analyst commented that AvgBal is a weak predictor of Default. Do you agree? Explain.

Disagree. The model coefficient and hence OR depends on the unit. $1 change is not realistic. $500 change (almost 1 standard deviation) would be more realistic and becomes strong predictor of Default.

See Rscript solution: default.R

Questions for Research Paper [Freitas et. al. (2012)] Reading:

1. How are outliers determined? Why is this impt?

   LOS > Geometric avg + 2SD. In order to get categorical Y for logistic reg.

2. What is the difference between adjusted Odds Ratio and unadjusted Odds Ratio?

   Adjusted Odds Ratio uses model coefficient from the model that includes all input variables. Unadjusted Odds Ratio uses model coefficient from the model that has only one input variable.

3. How did Freitas et. al. (2012)] identify high risk factors? *Hint: See their Table 1.*

   OR Confidence Interval that does not contain 1 and OR bigger than 1. The bigger the OR, the higher the risk.

Logistic Regression on Multi-category Y:

1. Set Service Rating = Neutral as the baseline reference level for Rating, in rating.csv dataset.

```
16  ratings.df$Rating <- relevel(ratings.df$Rating, ref = "Neutral")
17  levels(ratings.df$Rating)  # Baseline is now changed to "Neutral"
```

2. Develop Logistic regression to predict Rating using the multinom() function from Rpackage nnet. Which variables are statistically significant.

```
> summary(ratings.fit)
Call:
multinom(formula = Rating ~ . - Cust, data = ratings.df)

Coefficients:
     (Intercept)        WTQ       WTP LocationB   LocationC
Bad    -3.911087  0.3353235 0.0714015 0.1522743 -0.03359073
Good    1.195214 -0.2388393 0.1420355 0.1017307  0.39957803
```

Note the two linear equations for Z. One for Rating = Bad, another for Rating = Good.

```
> OR.CI
, , Bad

                2.5 %    97.5 %
(Intercept) 0.00340963 0.1175346
WTQ         1.24051365 1.5763647
WTP         0.68601901 1.6814440
LocationB   0.38765887 3.4979535
LocationC   0.31630077 2.9561279

, , Good

                2.5 %     97.5 %
(Intercept) 0.9354157 11.6719979
WTQ         0.7089447  0.8748518
WTP         0.7707552  1.7236695
LocationB   0.3867445  3.1691152
LocationC   0.5372330  4.1391041
```

The OR CI shows that only WTQ is statistically significant for Bad service rating and Good service rating as the confidence interval excludes 1. [Recall that OR is just a fraction.]

Alternatively, we can calculate the p-values to get the same conclusion. Only WTQ's p-value is less than 0.05.

```
> pvalue
     (Intercept)          WTQ       WTP LocationB LocationC
Bad  1.486471e-05 4.109339e-08 0.7548881 0.7861266 0.9530192
Good 6.341463e-02 8.492339e-06 0.4890781 0.8496359 0.4430074
```

3. What is the model predicted service rating for each of the case in the dataset?

   See the predicted_class vector in Rscript solution.

4. Show the confusion matrix.

	Model.Predict		
Trainset.Actuals	Neutral	Bad	Good
Neutral	17	15	18
Bad	9	56	3
Good	13	3	44

   See Rscript solution: rating.R