COMM 581 - Assignment #8
Logistic Regression

Name: _____Gurpal Bisra_____ **Total: 22 marks**
Due date: Monday Nov. 14, 2015 (11pm)

Background:

A magazine company is trying to figure out who to target with email advertisements for a children's magazine: Kid Creative. When they send an email, they only want to advertise three magazines to each potential customer, so they want to maximize their probability of making a sale, thereby maximizing their overall revenue. If they advertise magazines that someone doesn't buy, they have wasted that opportunity. They send out some experimental email advertisements to people on their mailing list. The sample includes people who have already purchased a magazine from the magazine company. The magazine company purchases additional information on these customers from a third party credit agency to create a profile of these customers. They want to see which variables can be used to predict if someone will buy the magazine Kid Creative.

Please submit your R script file for this assignment as part of your assignment PDF. Clearly label each model that you used in the assignment. (1 mark)

Questions

1. What relationship do you expect between each potential explanatory variable and the response variable? Why? (You can say "no relationship", however, you must explain why.) (**2 marks**)

I would expect a sigmoidal curve between the response variable and the several explanatory variables. The response variable is buy (i.e. whether to purchase the magazine Kid Creative or not). First, I would imagine the variable buy plotted against income or unemployed would exhibit a sigmoidal curve because one can expect someone to spend money on a kid's magazine after a certain annual income given that the magazine is not a living expense. Likewise, the variables dual income and professional job would exhibit a sigmoidal curve for the same reason. Next, I would expect the graphs of buy plotted against any of married, children, gender, previously bought children's magazine, previously bought parenting magazine, or children explanatory variables to also exhibit a sigmoidal curve. One would is married or who already expressed interested in having or raising children may be inclined to purchase the magazine Kid Creative to nurture their own children. Generally speaking, women are more likely to be nurturing to their children than men so one's gender may affect their likelihood of purchasing the magazine Kid Creative. Furthermore, I expect the plot of buy as a function of college to appear sigmoidal since educated people may place a higher emphasis on learning on their own children – especially if they took courses in psychology or sociology. Finally, I would imagine retired to also display a sigmoidal curve when the response variable is buy. For instance, one who is retired may purchase the magazine Kid Creative either for their own grandchildren or for other children as a gift.

2. Which 3-5 potential explanatory variables do you think are most likely to be good predictors of whether or not someone will buy a children's magazine? (Which potential explanatory variables do you think will have the strongest relationship with the response variable?) (**0.5 mark**)

Some potential explanatory variables are: (1) income since those with more disposable income will purchase the magazine Kid Creative; (2) gender (i.e. females) since they are more likely to be caretakers; (3) dual income instead of married because there is more disposable income; (4) college (i.e. education) because people who may have post-secondary education may value reading more; (5) previously purchased a children's or parenting magazine because such individuals would be more likely to make additional purchases having overcome their initial skepticism regarding the value of magazines.

3. Fit a null model with the intercept only. What is the estimate of the intercept for this model? Convert this estimate from logit (log odds) units to probability units? (**0.25 marks**)

```
> z.null <- glm(buy ~ 1, data=mydata, family="binomial"(link="logit"))
> summary(z.null)

Call:
glm(formula = buy ~ 1, family = binomial(link = "logit"), data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6411  -0.6411  -0.6411  -0.6411   1.8349

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.47796    0.09912  -14.91   <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: 646.05  on 672  degrees of freedom
Residual deviance: 646.05  on 672  degrees of freedom
AIC: 648.05

Number of Fisher Scoring iterations: 4
```

Hence the intercept is -1.47796. And the estimate in probability units is 0.1857.

```
> exp(-1.47796)/(1 + exp(-1.47796))
[1] 0.1857357
```

4. Using the original data, calculate the probability of a purchase. What do you notice about this value and the intercept from the null model? (**0.25 marks**)

Using the original data, the probability a purchase is 0.1857, which is the same as the estimate of the intercept from logit (log odds) units in probability units.

```
> mean(mydata$buy)
[1] 0.1857355
```

5. Use R to obtain the log likelihood for the null model.

The log likelihood for the null model is -323.0265.

```
> logLik(z.null)
'log Lik.' -323.0265 (df=1)
```

Single variable models

6. Fit a model with **income** only. Use R to obtain the log likelihood for this model.

```
> summary(z1)

Call:
glm(formula = buy ~ income, family = binomial(link = "logit"),
    data = mydata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.00280  -0.22466  -0.05466  -0.01655   2.87445

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.344e+00  8.315e-01  -11.24   <2e-16 ***
income       1.494e-04  1.336e-05   11.18   <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: 646.05  on 672  degrees of freedom
Residual deviance: 249.32  on 671  degrees of freedom
AIC: 253.32

Number of Fisher Scoring iterations: 7

> logLik(z1)
'log Lik.' -124.6621 (df=2)
```

7. Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R) for the model including **income**. Test the significance of the model (include all four steps of your hypothesis test). (**1 mark**)

```
> anova(z.null, z1, test="Chi")
Analysis of Deviance Table

Model 1: buy ~ 1
Model 2: buy ~ income
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       672     646.05
2       671     249.32  1   396.73 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z.null) - logLik(z1))
'log Lik.' 396.7289 (df=1)
```

Step 3: Compare to Chi-statistic

(p-value = 2.2e-16) < (α = 0.05)

($X^2_{1,1-0.05}$= 3.84) < (G = 396)

Step 4: Therefore, the regression is significant. I reject the null hypothesis and the so go with more complex model which has income in model.

8. Show the full calculation (based on log likelihood values from R) of the AIC value for the model including **income**. (**0.25 marks**)

```
> # Calculate AIC
> -2*(logLik(z1))+2*(1+1)
'log Lik.' 253.3241 (df=2)
> AIC(z1)     # takes into account number of variables (penalized if more)
[1] 253.3241
> AICc(z1)    # takes into account number of variables and sample size
[1] 253.3421
```

9. Fit a model with **married** only. Use R to obtain the log likelihood for this model.

```
> summary(z2)

Call:
glm(formula = buy ~ Married, family = binomial(link = "logit"),
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9687  -0.4201  -0.4201  -0.4201   2.2232

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3830     0.1718 -13.870   <2e-16 ***
Married1      1.8699     0.2184   8.563   <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: 646.05  on 672  degrees of freedom
Residual deviance: 564.47  on 671  degrees of freedom
AIC: 568.47

Number of Fisher Scoring iterations: 5

> logLik(z2)
'log Lik.' -282.2328 (df=2)
```

The log likelihood for the null model is -282.2328.

10. Write the full calculation for the likelihood ratio test statistic (based on log likelihood values from R) for the model including **married**. Test the significance of the model (include all four steps of your hypothesis test). (**1 mark**)

```
> anova(z.null, z2, test="Chi")
Analysis of Deviance Table

Model 1: buy ~ 1
Model 2: buy ~ Married
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       672     646.05
2       671     564.47  1   81.588 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z.null) - logLik(z2))
'log Lik.' 81.58756 (df=1)
```

Step 3: Compare to Chi-statistic

(p-value = 2.2e-16) < (α = 0.05)

($X^2_{1,1-0.05}$= 3.84) < (G = 81.58756)

Step 4: Therefore, regression is significant. I reject the null hypothesis and the so go with more complex model and have married in my model.

11. Show the full calculation (based on log likelihood values from R) of the AIC value for the model including **married**. (**0.25 marks**)

The AIC value for the model including married is -568.4655.

```
> -2*(logLik(z2))+2*(1+1)
'log Lik.' 568.4655 (df=2)
> AIC(z2)     # takes into account number of variables (penalized if more)
[1] 568.4655
> AICc(z2)    # takes into account number of variables and sample size
[1] 568.4834
```

12. Fit 11 models, **each with one of the following explanatory variables**: income, gender, married, education, professional job, retired, unemployed, dual income, children, bought children's magazine previously, and bought parenting magazine previously. Record information about each model in the following table. Organize your models from lowest AIC value to highest AIC value. Example table below. (**4 marks**)

The table below summarizes my models from lowest AIC value to highest AIC value. For information how I obtained my values, please refer to my code which is found at the end of this document.

Explanatory variable	AIC value	G statistic	p value	Include variable in model?
Income	253.3241	396	2.2e-16	Yes
Married	586.4655	81.58756	2.2e-16	Yes
Dual Income	593.6976	56.35545	6.049e-14	Yes
Professional Job	616.7322	33.32085	7.814e-09	Yes
Education (i.e. College)	624.8341	25.21897	5.118e-07	Yes
Previously bought Children's Magazine	637.5077	12.54536	0.0003972	Yes
Gender (i.e. Female)	645.957	4.096057	0.04298	Yes
Unemployed	**646.4233**	**3.629809**	**0.05675**	**No**
Previously bought Parenting Magazine	**646.6884**	**3.364706**	**0.06661**	**No**
Retired	**649.762**	**0.291094**	**0.5895**	**No**
Children	**649.7904**	**0.2626163**	**0.6083**	**No**

13. What do you notice about the AIC values and the results of the likelihood ratio tests? Which AIC values indicate a better fit? (**0.5 marks**)

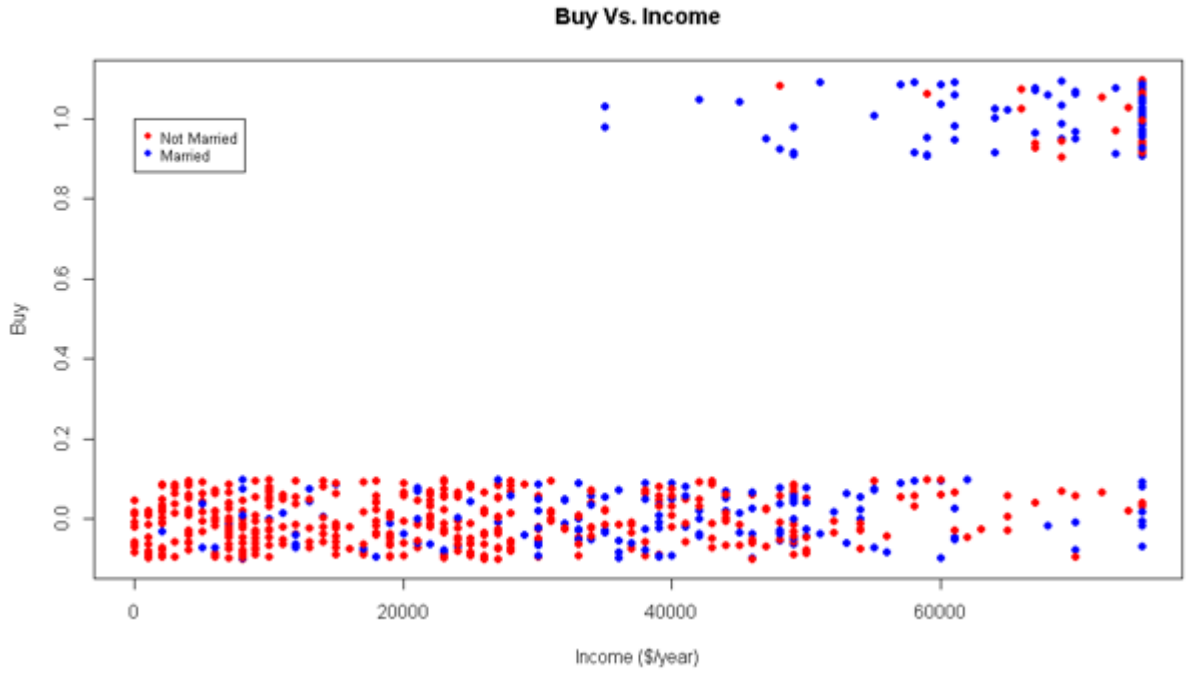
Higher AIC values indicate a worse regression fit (i.e. one fails to reject the null hypothesis as the regression is not significant) when compared to lower AIC values which indicate a better fit (i.e. one rejects the null hypothesis as the regression is significant). This means that a good regression model can predict the response variable better if it has lower AIC values.

14. Out of all the possible explanatory variables, are there any that you think are redundant or correlated? Out of redundant variables, which one would you prefer to use? Why? **(1 mark)**

Some explanatory variables which are redundant/correlated are: (1) married and dual income; and (2) income and unemployed; (3) income and professional; (4) income and dual income. Purchasing children's magazines are not an essential purchase required for one to live. That said, it is logical to assume that when someone is married, both partners have an income in contemporary society. Likewise, one can assume someone has an income if they are not unemployed. Moreoever, one can assume a professional would have a high income as professional jobs include roles like doctor, engineer, lawyer etc. Furthermore, there are people in the dataset who have an income but are themselves unemployed; they likely receive money from their spouse.

I would prefer using dual income instead of married because dual income would imply a couple has more disposable income to spend on a child's magazine. Similarly, I would prefer to use income over the factor unemployed to fit the data because one who has a higher annual income would have more disposable income to spend on children's magazines.

15. Using R, create a graph of purchase vs. income with a method to visualize married. What are some overall patterns that you see? How can these help you answer your research question? **(1 mark)**




```
plot(jitter(buy, f = 0.5)~ income, data=mydata, pch=16, col = c("red","blue")[as.factor
(mydata$Married)], main = "Buy Vs. Income", xlab = "Income ($/year)", ylab = "Buy")
```

Based on the graph of purchases vs. income, where the differences between married and not married are colored, I see a pattern that married people appear more likely to purchase magazines. Based on this observation, I can improve my regression model by including married in addition to income, as my explanatory variables.

Model A: Model with income and married:

16. Create a model with the following explanatory variables: income, married, the interaction between income and married. (buy ~ income*married)

```
> z12 <- glm(buy ~ income + Married + income*Married, data=mydata, family="binomial"(link="logit"))
> AIC(z12)                              # AIC value
[1] 233.6734
> -2*(logLik(z.null) - logLik(z12))     # G value
'log Lik.' 420.3797 (df=1)
> anova(z.null, z12, test="Chi")        # p value
Analysis of Deviance Table

Model 1: buy ~ 1
Model 2: buy ~ income + Married + income * Married
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       672     646.05
2       669     225.67  3   420.38 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

17. Does this model meet the assumptions of generalized linear models? (**1 mark**)

1. Statistical independence of observations:

The assumption of independence of observations cannot be verified because I am not provided any information on when or where the data was collected. For instance, if I wanted to prove my data depended on another explanatory variable, I would need to look at my data at 2 different time points.

First, it might be possible that the same individual was sampled more than once. For instance, one may have not bought the magazine, had a high income, been married etc. one year but then have bought the magazine, had a low income, and divorced years later. Counting the same individual across different time points may produce different buy values. Furthermore, since the form of communication is email, I would imagine that space is not affecting the assumption of independence of observations.

2. Correct specification of link function:

Since the response variable is binary, the correct linking function “Logit” was used. This link function is an appropriate match for the errors too.

3. **Variance correspond to what is expected from the link function:**

First, one can obtain the mean of proportions of successes (i.e. total successes/total) by the following code:

```
> mean(mydata$buy)
[1] 0.1857355
```

In fact, p = the proportion of successes = 0.1857355. Given that the response variable can take on values of 0 or 1, one can expect the distribution of successes (i.e. buy = 1) and failures (i.e. buy = 0) to be modeled as a binomially distributed function. Therefore, the expected variance can be computed as follows:

expected variance = p*(1-p)

```
> expected_variance = 0.1857355*(1-0.1857355)
> expected_variance
[1] 0.1512378
```

Finally, this value can be compared to what is calculated by R:

```
> var(mydata$buy)
[1] 0.1514629
```

Both variances are more or less equal values.

Alternatively, I could compare the residual deviance from what is obtained in the summary command.

18. What does the interaction between income and married allow in this model allow? (**0.25 marks**)

The interaction between income and married allows one to test the situation where the simultaneous influence of both variables on the response variable buy is not additive. If it is the case that the co-efficient corresponding to the interaction is not zero (i.e. the interaction is significant), then one requires more information other than a customer's income and marital status to predict whether they would be likely to purchase the magazine Kid Creative. Consequently, the observed slopes in the graph of buy in the log odds ratios units plotted against income may be modified.

19. Test the likelihood of the whole model compared to the likelihood of the null model using a likelihood ratio test. Show your calculation of the test statistic using the log-likelihoods from R. Confirm your results using R. (**1 mark**)

```
> anova(z.null, z12, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ 1
Model 2: buy ~ income + Married + income * Married
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       672     646.05
2       669     225.67  3   420.38 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z.null) - logLik(z12))     # G value
'log Lik.' 420.3797 (df=1)
```

Step 3: Compare to Chi-statistic

(p-value = 2.2e-16) < (α = 0.05)

($X^2_{3,1-0.05}$= 7.81) < (G = 420.3997)

Step 4: Therefore, the regression is significant. I reject the null hypothesis and the so go with more complex model which has income, married and the interaction in model for now.

20. Test each variable using a likelihood ratio test. This will require you to fit models that eliminate one variable. Show your calculation of the test statistics using the log-likelihoods from R. Confirm your results using R. (**1 mark**)

So far, I have compared the following models:

(1) Buy ~ 1 AND Buy ~ income

(2) Buy ~ 1 AND Buy ~ Marriage

I determined the variables income and married are significant for determining whether someone will purchase the magazine Kid Creative (i.e. Buy = 1).

There are a total of 4 models which could be used in such analyses which I will denote as:

(1) z1: Buy ~ income

(2) z2: Buy ~ Married

(3) z13 Buy ~ income + Married

(4) z12: Buy ~ income + Married + income*Married

Testing the Significance of Married [i.e. Compare (1) and (4)]

```
> anova(z12, z1, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ income + Married + income * Married
Model 2: buy ~ income
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       669     225.67
2       671     249.32 -2  -23.651 7.317e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z1) - logLik(z12))    # G value
'log Lik.' 23.65074 (df=2)
```

Step 3: Compare to Chi-statistic

(p-value = 7.317e-06) < (α = 0.05)

($X^2_{2,1-0.05}$= 5.99) < (G = 23.65074)

Step 4: Therefore, the more complex model is significantly better. I reject the null hypothesis and the so go with more complex model which includes the variable married in model.

Testing the Significance of income [i.e. Compare (2) and (4)]

```
> anova(z12, z2, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ income + Married + income * Married
Model 2: buy ~ Married
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       669     225.67
2       671     564.47 -2  -338.79 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z2) - logLik(z12))    # G value
'log Lik.' 338.7921 (df=2)
```

Step 3: Compare to Chi-statistic

(p-value = 2.2e-16) < (α = 0.05)

($X^2_{2,1-0.05}$= 5.99) < (G = 338.7921)

Step 4: Therefore, the more complex model is significantly better. I reject the null hypothesis and the so go with more complex model which includes the variable income in the model.

21. If both variables should remain in the model, test the interaction term only using a likelihood ratio test (you will have to fit a reduced Model A that excludes the interaction). Show your calculation of the test statistics using the log-likelihoods from R. Confirm your results using R. (**1 mark**)

Testing the Significance of the Interaction [i.e. Compare (3) and (4)]

```
> anova(z12, z3, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ income + Married + income * Married
Model 2: buy ~ Female
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       669     225.67
2       671     641.96 -2  -416.28 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z3) - logLik(z12))     # G value
'log Lik.' 416.2836 (df=2)
```

Step 3: Compare to Chi-statistic

(p-value = 2.2e-16) < (α = 0.05)

($X^2_{2,1-0.05}$= 5.99) < (G = 338.7921)

Step 4: Therefore, the more complex model is significantly better. I reject the null hypothesis and the so go with more complex model which includes married, income, and the interaction in the model.

Model B: Model with income, married and professional job:

22. Create a model with the following explanatory variables: income, married, the interaction between income and married, job, and the interaction between job and income. (buy ~ income*married + income*job)

```
> z14 <- glm(buy ~ income + Married + Professional + income*Married + income*Professional, dat
a=mydata, family="binomial"(link="logit"))
```

23. What does the interaction between job and income allow in this model? (**0.25 marks**)

The interaction between income and professional job allows one to test the situation where the simultaneous influence of both variables on the response variable "buy" is not additive. If it is the case that the co-efficient corresponding to the interaction is not zero (i.e. the interaction is significant), then one requires more information other than a customer's income and professional employment status to predict whether they would be likely to purchase the magazine Kid Creative. Consequently, the observed slopes in the graph of buy in the log odds ratios units plotted against income may be modified.

24. Compare this model (Model B) to Model A, which did not have job, using a likelihood ratio test. Does including professional job as a variable improve the model? (**1 mark**)

    You could continue trying different variables to add to your model. Choose the variables that explain the most variation on their own and add those next. For this assignment, we will stop developing models at this point.

Now I will compare a total of 2 models which I will denote as:

(Model A) z12: Buy ~ income + Married + income*Married

(Model B) z14: Buy ~ income + Married + Professional + income*Married + income*Professional

Testing the Significance of the variable Unemployment [i.e. Compare (Model A) and (Model B)]

```
> anova(z14, z12, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ income + Married + Professional + income * Married + income *
    Professional
Model 2: buy ~ income + Married + income * Married
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       667     224.93
2       669     225.67 -2 -0.74237   0.6899
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z12) - logLik(z14))    # G value
'log Lik.' 0.742373 (df=4)
```

Step 3: Compare to Chi-statistic

(p-value = 0.6899) > (α = 0.05)

($X^2_{2,1-0.05}$= 5.99) > (G = 0.742373)

Step 4: Therefore, the more complex model is not significantly better. I fail to reject the null hypothesis and will go with less complex model. This means I will keep the explanatory variables marriage, income, and the interaction in the model but not professional or the interaction between income and professional. Hence, the explanatory variable professional job does not improve the model.

Testing the Significance of the variable Dual Income [i.e. Compare (Model A) and (Model C)]

Model C: z15 <- glm(buy ~ income + Married + Dual.Income + income*Married + income*Dual.Income, data=mydata, family="binomial"(link="logit"))

```
> anova(z15, z12, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ income + Married + Dual.Income + income * Married + income *
    Dual.Income
Model 2: buy ~ income + Married + income * Married
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       667     222.77
2       669     225.67 -2  -2.9078   0.2337
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z12) - logLik(z15))    # G value
'log Lik.' 2.907851 (df=4)
```

Step 3: Compare to Chi-statistic

(p-value = 0.2337) > (α = 0.05)

($X^2_{2,1-0.05}$= 5.99) > (G = 2.907851)

Step 4: Therefore, the more complex model is not significantly better. I fail to reject the null hypothesis and will go with less complex model. This means I will keep the explanatory variables married, income, and the interaction in the model but not dual income or the interaction between income and dual income.

Testing the Significance of the variable Unemployment [i.e. Compare (Model A) and (Model D)]

Model D: z16 <- glm(buy ~ income + Married + Unemployed + income*Married + income*Unemployed, data=mydata, family="binomial"(link="logit"))

```
> anova(z16, z12, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ income + Married + Unemployed + income * Married + income *
    Unemployed
Model 2: buy ~ income + Married + income * Married
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       667     225.27
2       669     225.67 -2 -0.40279   0.8176
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z12) - logLik(z16))     # G value
'log Lik.' 0.4027884 (df=4)
```

Step 3: Compare to Chi-statistic

(p-value = 0.8176) > (α = 0.05)

($X^2_{2,1-0.05}$= 5.99) > (G = 0.4027884)

Step 4: Therefore, the more complex model is not significantly better. I fail to reject the null hypothesis and will go with less complex model. This means I will keep the explanatory variables married, income, and the interaction in the model but not unemployed or the interaction between income and unemployed.

Testing the Significance of the variable College [i.e. Compare (Model A) and (Model E)]

Model E: z17 <- glm(buy ~ income + Married + College + income*Married + income*College, data=mydata, family="binomial"(link="logit"))

```
> anova(z17, z12, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ income + Married + College + income * Married + income
    College
Model 2: buy ~ income + Married + income * Married
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       667     225.24
2       669     225.67 -2 -0.43036   0.8064
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z12) - logLik(z17))     # G value
'log Lik.' 0.4303565 (df=4)
```

Step 3: Compare to Chi-statistic

(p-value = 0.8064) > (α = 0.05)

($X^2_{2,1-0.05}$= 5.99) > (G = 0.4303565)

Step 4: Therefore, the more complex model is not significantly better. I fail to reject the null hypothesis and the so go with less complex model (i.e. keep married, income, and the interaction in the model but not college or the interaction between income and college).

Testing the Significance of the variable Prev.Child.Mag [i.e. Compare (Model A) and (Model F)]

Model F: z18 <- glm(buy ~ income + Married + Prev.Child.Mag + income*Married + income*Prev.Child.Mag, data=mydata, family="binomial"(link="logit"))

```
> anova(z18, z12, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ income + Married + Prev.Child.Mag + income * Married +
    income * Prev.Child.Mag
Model 2: buy ~ income + Married + income * Married
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       667     220.18
2       669     225.67 -2  -5.4943  0.06411 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z12) - logLik(z18))    # G value
'log Lik.' 5.494329 (df=4)
```

Step 3: Compare to Chi-statistic

(p-value = 0.06411) > (α = 0.05)

($X^2_{2,1-0.05}$= 5.99) > (G = 5.494329)

Step 4: Therefore, the more complex model is not significantly better. I fail to reject the null hypothesis and will go with less complex model. This means I will keep the explanatory variables

married, income, and the interaction in the model but not Prev.Child.Mag or the interaction between income and Prev.Child.Mag.

Testing the Significance of the variable Female [i.e. Compare (Model A) and (Model G)]

Model G: z19 <- glm(buy ~ income + Married + Female + income*Married + income*Female, data=mydata, family="binomial"(link="logit"))

```
> anova(z19, z12, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ income + Married + Female + income * Married + income *
    Female
Model 2: buy ~ income + Married + income * Married
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       667     211.82
2       669     225.67 -2  -13.857 0.0009796 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z12) - logLik(z19))    # G value
'log Lik.' 13.85679 (df=4)
```

Step 3: Compare to Chi-statistic

(p-value = 0.0009796) < (α = 0.05)

($X^2_{2,1-0.05}$= 5.99) < (G = 13.85679)

Step 4: Therefore, the more complex model is significantly better. I reject the null hypothesis and the so go with more complex model which includes the explanatory variables married, income, the interaction between married and income, gender, and the interaction between gender and income for now.

Now, I must test whether keeping the interaction term between gender and income is necessary:

Testing the Significance of the variable Female [i.e. Compare (Model G) and (Model H)]

Model H: z20 <- glm(buy ~ income + Married + Female + income*Married, data=mydata, family="binomial"(link="logit"))

```
> anova(z19, z20, test="Chi")       # p value
Analysis of Deviance Table

Model 1: buy ~ income + Married + Female + income * Married + income *
    Female
Model 2: buy ~ income + Married + Female + income * Married
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       667     211.82
2       668     212.43 -1 -0.61632   0.4324
```

Step 1: Hypothesis

H0: No difference between the two models (restriction is justified → use simpler model)

H1: The two models have significantly different likelihoods (restriction is not justified → use more complex model)

Step 2: Calculate G-statistic

$G = lambda_{LR} = 2\ln(\frac{L_R}{L_R})$

```
> -2*(logLik(z20) - logLik(z19))     # G value
'log Lik.' 0.616321 (df=5)
```

Step 3: Compare to Chi-statistic

(p-value = 0.4324) > (α = 0.05)

($X^2_{2,1-0.05}$= 5.99) > (G = 0.616321)

Step 4: Therefore, the more complex model is not significantly better. I fail to reject the null hypothesis and the so go with less complex model which includes the explanatory variables married, income, the interaction between married and income, and gender.

To summarize, the model which best improves the model is Model z20:

z20 <- glm(buy ~ income + Married + Female + income*Married, data=mydata, family="binomial"(link="logit"))

That said, for the purposes of this homework, I will use Model z12 for questions 25-29:

z12 <- glm(buy ~ income + Married + income*Married, data=mydata, family="binomial"(link="logit"))

Final model: With the variables that you have decided should be included

25. For your final model, calculate the Pseudo-R^2 and the scaled Pseudo-R^2. (**1 mark**)

I calculated my Pseudo-R^2 value to be -0.003130574 and Scaled Pseudo-R^2 value to be 0.1807.

Pseudo-R^2

```
> # Pseudo R2
> 1 - (logLik(z.null)/logLik(z12))^(2/nrow(mydata))
'log Lik.' -0.003130574 (df=1)
```

Scaled Pseudo-R^2

= (1 - (logLik(z.null)/logLik(z12))^(2/nrow(mydata)))/(1-logLik(z.null)^(2/nrow(mydata)))
= (1 - ((-323.0265))/(-112.8367)^(2/(673)))/(1-(-323.0265)^(0.002971768))
= 0.1807

- note: I could only compute this value using a calculator

26. Calculate the AIC value for your final model. Compare the AIC value of your final model to the AIC values for the models that had only one of the variables that are included in your final model (Single variable models). Put all of these models and AIC values in a table to make them easy to compare. How much does the AIC value improve from the single variable models of income, married and professional job to your final model? (**2 marks**)

```
> AIC(z12)
[1] 233.6734
> AICc(z12)
[1] 233.7333
```

Explanatory variable(s)	AIC value	Include variable in model?	Difference between Final Model and Single Explanatory Variable
Final Model: income + Married + income*Married	**233.6734**	**Yes**	**0**
Income	253.3241	Yes	**-19.6507**
Married	586.4655	Yes	**-352.792**
Dual Income	593.6976	**No**	-360.024
Professional Job	616.7322	**No**	**-383.059**
Education (i.e. College)	624.8341	**No**	-391.161
Previously bought Children's Magazine	637.5077	**No**	-403.834
Gender (i.e. Female)	645.957	**No**	-412.284
Unemployed	**646.4233**	**No**	-412.75
Previously bought Parenting Magazine	**646.6884**	**No**	-413.015
Retired	**649.762**	**No**	-416.089
Children	**649.7904**	**No**	-416.117

When using my final model, the AIC value improve from the single variable models of income, married, and professional job by 19.6507, 352.792, and 383.059, respectively.

27. Create a classification table for this data based on the final model. Include the full classification table in your output. Remember that you can output dataframes to a .csv file using write.csv().

The code I used to create my classification table, using my final model z12, is shown below:

```
# Create an empty dataframe that you will fill with
df <- data.frame(matrix(ncol = 9, nrow = 51))
colnames(df) <- c("correct.event", "correct.non.event", "incorrect.event", "incorrect
.non.event", "correct.percent", "sensitivity", "specificity", "false.pos", "false.neg")
df

prob.level <- seq(0, 1, length.out=51) # create a vector with different possible
probabilities
prob.level
class.table.data <- cbind(prob.level, df) # combine your vector of probabilities and
your empty dataframe
class.table.data # Your dataframe has one row for each probability cut-off

# fill empty cells in your dataframe with 0
class.table.data$correct.non.event <- rep(c(0), c(51))
class.table.data$correct.event <- rep(c(0), c(51))
class.table.data$incorrect.non.event <- rep(c(0), c(51))
class.table.data$incorrect.event <- rep(c(0), c(51))
class.table.data

# This loop will try out the different probability cut-off values and fill in how many
correct and incorrect events and non-events you have based on your data.
for (i in 1:51) {
    class.table <- table(mydata$buy, fitted(z12) > class.table.data$prob.level[i])

    col.true.num <- grep("TRUE", colnames(class.table))
    col.false.num <- grep("FALSE", colnames(class.table))

    if (length(col.true.num) > 0) {
        class.table.data$incorrect.non.event [i] <- class.table[1, col.true.num]
        class.table.data$correct.event [i] <- class.table[2, col.true.num] }

    if (length(col.false.num) > 0) {
        class.table.data$correct.non.event [i] <- class.table[1, col.false.num]
        class.table.data$incorrect.event [i] <- class.table[2, col.false.num] }  }

class.table.data

# You will use this information to fill in the rest of your classification table.
class.table.data$correct.percent <- (class.table.data$correct.event + class.table
.data$correct.non.event)/nrow(mydata)
class.table.data$sensitivity <- (class.table.data$correct.event)/nrow(mydata)
class.table.data$specificity <- (class.table.data$correct.non.event)/nrow(mydata)
class.table.data$false.neg <- (class.table.data$incorrect.non.event)/nrow(mydata)
class.table.data$false.pos <- (class.table.data$incorrect.event)/nrow(mydata)
class.table.data

write.csv(class.table.data, file = "ClassTable.csv")
```

The output table I created is shown below:

	A	B	C	D	E	F	G	H	I	J	K
1		prob.level	correct.event	correct.non.event	incorrect.event	incorrect.non.event	correct.percent	sensitivity	specificity	false.pos	false.neg
2	1	0	125	0	0	548	0.185735513	0.185735513	0	0	0.814264487
3	2	0.02	124	410	1	138	0.79346211	0.184249629	0.609212481	0.001485884	0.205052006
4	3	0.04	124	433	1	115	0.827637444	0.184249629	0.643387816	0.001485884	0.170876672
5	4	0.06	122	446	3	102	0.843982169	0.18127786	0.662704309	0.004457652	0.151560178
6	5	0.08	122	462	3	86	0.867756315	0.18127786	0.686478455	0.004457652	0.127786033
7	6	0.1	121	472	4	76	0.881129272	0.179791976	0.701337296	0.005943536	0.112927192
8	7	0.12	120	473	5	75	0.881129272	0.178306092	0.70282318	0.007429421	0.111441308
9	8	0.14	120	476	5	72	0.885586924	0.178306092	0.707280832	0.007429421	0.106983655
10	9	0.16	119	484	6	64	0.895988113	0.176820208	0.719167905	0.008915305	0.095096582
11	10	0.18	118	486	7	62	0.897473997	0.175334324	0.722139673	0.010401189	0.092124814
12	11	0.2	117	491	8	57	0.903417533	0.17384844	0.729569094	0.011887073	0.084695394
13	12	0.22	113	502	12	46	0.913818722	0.167904903	0.745913819	0.017830609	0.068350669
14	13	0.24	113	506	12	42	0.919762259	0.167904903	0.751857355	0.017830609	0.062407132
15	14	0.26	113	507	12	41	0.921248143	0.167904903	0.753343239	0.017830609	0.060921248
16	15	0.28	112	508	13	40	0.921248143	0.166419019	0.754829123	0.019316493	0.059435364
17	16	0.3	112	509	13	39	0.922734027	0.166419019	0.756315007	0.019316493	0.05794948
18	17	0.32	112	511	13	37	0.925705795	0.166419019	0.759286776	0.019316493	0.054977712
19	18	0.34	112	511	13	37	0.925705795	0.166419019	0.759286776	0.019316493	0.054977712
20	19	0.36	112	517	13	31	0.9346211	0.166419019	0.76820208	0.019316493	0.046062407
21	20	0.38	111	520	14	28	0.937592868	0.164933135	0.772659733	0.020802377	0.041604755
22	21	0.4	111	520	14	28	0.937592868	0.164933135	0.772659733	0.020802377	0.041604755
23	22	0.42	109	521	16	27	0.936106984	0.161961367	0.774145617	0.023774146	0.040118871
24	23	0.44	108	522	17	26	0.936106984	0.160475483	0.775631501	0.02526003	0.038632987
25	24	0.46	108	522	17	26	0.936106984	0.160475483	0.775631501	0.02526003	0.038632987
26	25	0.48	104	524	21	24	0.933135215	0.154531947	0.778603269	0.031203566	0.035661218
27	26	0.5	104	524	21	24	0.933135215	0.154531947	0.778603269	0.031203566	0.035661218
28	27	0.52	101	524	24	24	0.928677563	0.150074294	0.778603269	0.035661218	0.035661218
29	28	0.54	99	526	26	22	0.928677563	0.147102526	0.781575037	0.038632987	0.03268945
30	29	0.56	99	526	26	22	0.928677563	0.147102526	0.781575037	0.038632987	0.03268945
31	30	0.58	93	530	32	18	0.925705795	0.138187221	0.787518574	0.047548291	0.026745914
32	31	0.6	93	530	32	18	0.925705795	0.138187221	0.787518574	0.047548291	0.026745914
33	32	0.62	93	531	32	17	0.927191679	0.138187221	0.789004458	0.047548291	0.02526003
34	33	0.64	93	533	32	15	0.930163447	0.138187221	0.791976226	0.047548291	0.022288262
35	34	0.66	93	533	32	15	0.930163447	0.138187221	0.791976226	0.047548291	0.022288262
36	35	0.68	90	533	35	15	0.925705795	0.133729569	0.791976226	0.052005944	0.022288262
37	36	0.7	89	533	36	15	0.924219911	0.132243685	0.791976226	0.053491828	0.022288262
38	37	0.72	89	533	36	15	0.924219911	0.132243685	0.791976226	0.053491828	0.022288262
39	38	0.74	88	534	37	14	0.924219911	0.130757801	0.79346211	0.054977712	0.020802377
40	39	0.76	85	534	40	14	0.919762259	0.126300149	0.79346211	0.059435364	0.020802377
41	40	0.78	83	535	42	13	0.918276374	0.12332838	0.794947994	0.062407132	0.019316493
42	41	0.8	78	535	47	13	0.910846954	0.11589896	0.794947994	0.069836553	0.019316493
43	42	0.82	72	538	53	10	0.906389302	0.106983655	0.799405646	0.078751857	0.014858841
44	43	0.84	72	538	53	10	0.906389302	0.106983655	0.799405646	0.078751857	0.014858841
45	44	0.86	46	541	79	7	0.872213967	0.068350669	0.803863299	0.117384844	0.010401189
46	45	0.88	44	541	81	7	0.869242199	0.0653789	0.803863299	0.120356612	0.010401189
47	46	0.9	0	548	125	0	0.814264487	0	0.814264487	0.185735513	0
48	47	0.92	0	548	125	0	0.814264487	0	0.814264487	0.185735513	0
49	48	0.94	0	548	125	0	0.814264487	0	0.814264487	0.185735513	0
50	49	0.96	0	548	125	0	0.814264487	0	0.814264487	0.185735513	0
51	50	0.98	0	548	125	0	0.814264487	0	0.814264487	0.185735513	0
52	51	1	0	548	125	0	0.814264487	0	0.814264487	0.185735513	0

28. For sensitivity, specificity, false negatives, false positives, which do you want to maximize or minimize, and which do you not care about? Why? (**2 marks**)

Sensitivity refers to predicting an event will happen when it actually does happen (i.e. this is one way of being correct), while specificity refers to predicting an event will not happen when it actually does not happen (i.e. the other way of being correct). Therefore, one would want to maximize both sensitivity and specificity. In R, we code these values as follows:

Sensitivity <- (class.table.data$correct.event)/nrow(mydata)

Specificity <- (class.table.data$correct.non.event)/nrow(mydata)

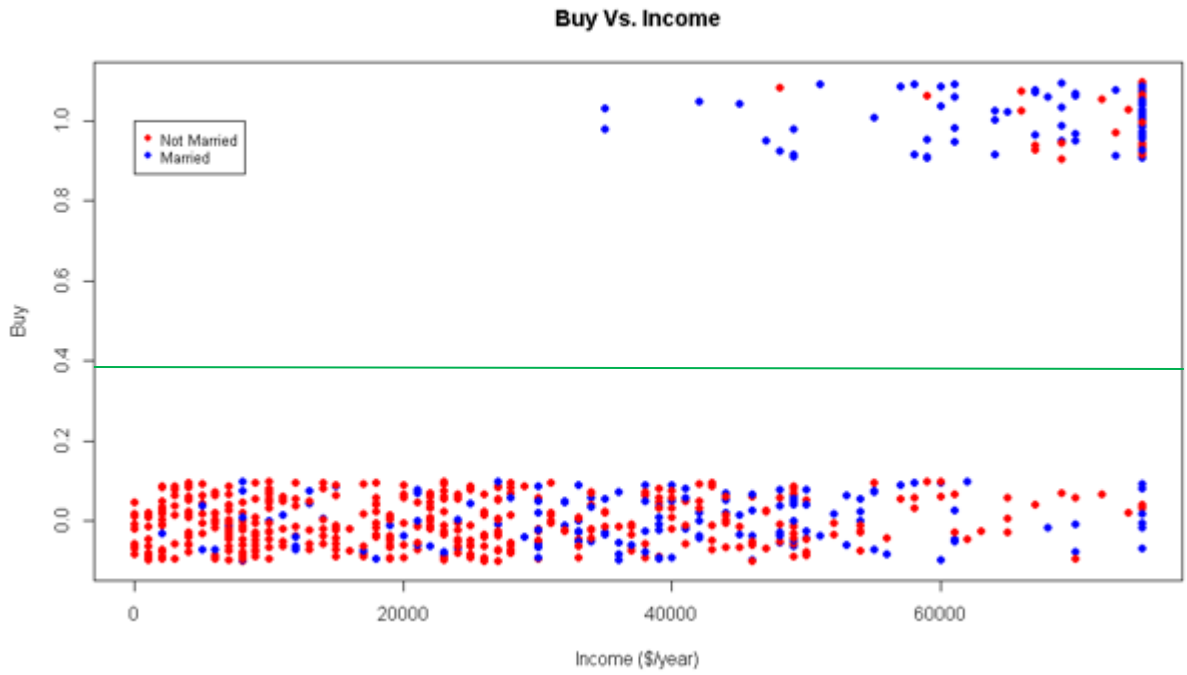
False Negatives <- (class.table.data$incorrect.non.event)/nrow(mydata)

False Positives <- (class.table.data$incorrect.event)/nrow(mydata)

One would want to minimize the number of false negatives and false positives. Such events means our models do not predict the data well or for certain cases. When these events occur for our dataset, it means the company fails to accurately predict which customers would purchase the magazine Kid Creative. Since their goal is to maximize profits, any such errors would mean they failed to capture the market. In particular, a false negative would be worse than a false positive because a potential purchase was never made (i.e. lost opportunity).

29. You decide to maximize the percentage of correct of predictions. Why? What probability cut-off should you use to maximize the percent of correct predictions? (**0.5 marks**)

The maximum number of "correct.percent' occurs (i.e. 0.937592868) when the probability is 0.38 or 0.40. Therefore, will select my cutoff to be 0.39 because it makes sense to pick 1 value. I would like to maximize the percentage of correct predictions because that would ensure the company supplies as many Kid Creative magazines as people demand; excess printed copies will not lose the company money overall. An example of how this cutoff would be used in practice is shown below:


Buy Vs. Income
Not Married
Married
Buy
0.0
0.2
0.4
0.6
0.8
1.0
0
20000
40000
60000
Income ($/year)

R Code:

```
1  ##################################
2  # ASSIGNMENT 8 - LOGISTIC REGRESSION
3  # Instructor: Martha Essak
4  # Gurpal Bisra
5  # Student Number: 69295061
6  # Nov. 8, 2016
7  ##################################
8
9  # Who more likey to purchase subscription for company
10 # 1 = Yes, 0 = No to questions
11 # Income = $/year
12 # females are more caretakers
13 # married = more likely to have children, but not say more likely to buy kid's magazine
   (bad variable to use)
14 # College = yes because would affect culturally
15 # professional = may not be meaningful
16 # children = maybe
17 # unemployed repeating income, so decide which variable better for us to use*
18 # dual income = similar to married, but better than married
19 # previously purchased children's magazine = yes affects
20 # previously purchase parenting magazine = yes affects
21
22 ####################
23 # CREATE THE DATASET
24 ####################
25
26 # read in the csv file
27 mydata <- read.csv("KidCreative dataset.csv", header=TRUE)
28
29 str(mydata)
30
31 # 'data.frame': 673 obs. of  12 variables:
32 # $ buy            : int  0 1 0 1 0 0 0 0 0 0 ...
33 # $ income         : int  24000 75000 46000 70000 43000 24000 26000 38000 39000 49000
   ...
34 # $ Female       : int  1 1 1 0 1 1 1 1 1 0 ...
35 # $ Married      : int  0 1 1 1 0 1 1 1 0 1 ...
36 # $ College    : int  1 1 0 0 0 0 1 0 1 0 ...
37 # $ Professional: int  1 1 0 1 0 0 0 0 1 0 ...
38 # $ Retired      : int  0 0 0 0 0 0 1 1 0 1 ...
39 # $ Unemployed     : int  0 0 0 0 0 0 0 0 0 0 ...
40 #$ Dual.Income    : int  0 1 1 0 0 0 0 0 0 0 ...
41 # $ Children       : int  0 0 1 0 0 0 1 0 0 0 ...
42 # $ Prev.Child.Mag : int  0 1 0 1 0 0 0 0 0 0 ...
43 # $ Prev.Parent.Mag: int  0 0 0 0 1 0 0 0 0 0 ...
44
45 # IF CHANGE 0 AND 1 TO NAMES
46 # mydata$Is.Female <- factor(mydata$Is.Female, labels = c("male", "female"))
47     # write the value for 0 first, then the value for 1
48
49 ### 1.  What relationship do you expect between each potential explanatory variable and
   the response variable? Why? (You can say "no relationship", however, you must explain
   why.) (2 marks)
50
```

```r
### 2.  Which 3-5 potential explanatory variables do you think are most likely to be good predictors of whether or not someone will buy a children's magazine? (Which potential explanatory variables do you think will have the strongest relationship with the response variable?)  (0.5 mark)

#################
# PLOT THE DATA
#################

plot(buy ~ income, data=mydata, pch=16, col = "blue", main = "Buy Vs. Income", xlab = "Income ($/year)", ylab = "Buy")

# What pattern do you see?
# looks like class notes, sigmoidal curve

# Lowess curve
lines(lowess(mydata$income, mydata$buy, delta=0.1), col="red")
plot(jitter(buy, f = 0.5)~ income, data=mydata, pch=16, col = c("red","blue")[as.factor(mydata$Married)])

# What shape do you expect the model to take?
# sigmoidal, but get flat curve


### 3.  Fit a null model with the intercept only. What is the estimate of the intercept for this model? Convert this estimate from logit (log odds) units to probability units?

####################################
# NULL MODEL / INTERCEPT ONLY MODEL
####################################

z.null <- glm(buy ~ 1, data=mydata, family="binomial"(link="logit"))
summary(z.null)
  # can also do partial F-test which is equivalent here to a global F-test

# The intercept is -1.47796 in log odds (the estimate of the mean on the logit scale).
# What is the estimate of the population proportion?
# You need to use the inverse equation to obtain this

exp(-1.47796)/(1 + exp(-1.47796))
  # = 0.5, so means half the sample did purchase and other half didn't purchase
  # only pick 1 number as cutoff point

# This might be a natural cut-off point, where any predicted values below the population mean would get a predicted value of 0, and any above would get a predicted value of 1

### 4.  Using the original data, calculate the probability of a purchase. What do you notice about this value and the intercept from the null model? (0.25 marks)

mean(mydata$buy)
# Note that your model has the same mean for predicted values

### 5.  Use R to obtain the log likelihood for the null model.
logLik(z.null)

### 6.  Fit a model with income only. Use R to obtain the log likelihood for this model.

```

```
##############################
# INTERPRETING CO-EFFICIENTS
##############################
# Fit the Continuous-only model
# glm = generalized linear model
# link says values can only take on 0 or 1
z1 <- glm(buy ~ income, data=mydata, family="binomial"(link="logit"))

summary(z1)

# Coefficients:
#     Estimate Std. Error z value Pr(>|z|)
# (Intercept) -9.344e+00  8.315e-01  -11.24   <2e-16 ***
#    income       1.494e-04  1.336e-05   11.18   <2e-16 ***

logLik(z1)

### 7.  Write the full calculation for the likelihood ratio test statistic (based on
log likelihood values from R) for the model including income. Test the significance of
the model (include all four steps of your hypothesis test). (1 mark)

################################
# LIKELIHOOD RATIO TEST
################################

# You can compare the null model and your model using a likelihood ratio test
anova(z.null, z1, test="Chi")

# Analysis of Deviance Table
#
# Model 1: buy ~ 1
# Model 2: buy ~ income
# Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
# 1       672     646.05
# 2       671     249.32  1   396.73 < 2.2e-16 ***
#    ---
#    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Calculate G value
-2*(logLik(z.null) - logLik(z1))
  # log Lik.' 396.7289 (df=1)

# Look on chi-squared distribution table for critical value
# critica value = 3.84
  # rejecting null hypothesis, so go with more complex model (have age in model)

# 8.    Show the full calculation (based on log likelihood values from R) of the AIC
value for the model including income. (0.25 marks)
# AIC = -2ln(L) + 2*(k+s)
    # k = levels of y -1
    # k = 2 - 1 = 1 for binary y(0,1)
    # s = number of predictor variables, including any indicator variables = 1

# Calculate AIC
-2*(logLik(z1))+2*(1+1)

# Double check with codes
library(AICcmodavg)
# AIC, smaller (more negative) is better
AIC(z1)     # takes into account number of variables (penalized if more)
AICc(z1)    # takes into account number of variables and sample size

### 9.  Fit a model with married only. Use R to obtain the log likelihood for this
model.
```

```
##############################
# Model with Married Only
##############################
# Fit the categorical-only model
# glm = generalized linear model
# link says values can only take on 0 or 1
mydata$Married <- factor(mydata$Married)

z2 <- glm(buy ~ Married, data=mydata, family="binomial"(link="logit"))

summary(z2)
# Coefficients:
#    Estimate Std. Error z value Pr(>|z|)
# (Intercept)  -2.3830     0.1718 -13.870   <2e-16 ***
#    Married1      1.8699     0.2184   8.563   <2e-16 ***

logLik(z2)

### 10. Write the full calculation for the likelihood ratio test statistic (based on
log likelihood values from R) for the model including married. Test the significance of
the model (include all four steps of your hypothesis test). (1 mark)

# You can compare the null model and your model using a likelihood ratio test
anova(z.null, z2, test="Chi")

# Analysis of Deviance Table
#
# Model 1: buy ~ 1
# Model 2: buy ~ income
# Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
# 1       672     646.05
# 2       671     249.32  1   396.73 < 2.2e-16 ***
#    ---
#    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Calculate G value
-2*(logLik(z.null) - logLik(z2))
# log Lik.' 81.58756 (df=1)

# Look on chi-squared distribution table for critical value
# critica value = 3.84
# rejecting null hypothesis, so go with more complex model (have age in model)

# 8.    Show the full calculation (based on log likelihood values from R) of the AIC
value for the model including income. (0.25 marks)
# AIC = -2ln(L) + 2*(k+s)
# k = levels of y -1
# k = 2 - 1 = 1 for binary y(0,1)
# s = number of predictor variables, including any indicator variables = 1

### 11. Show the full calculation (based on log likelihood values from R) of the AIC
value for the model including married. (0.25 marks)

# Calculate AIC
-2*(logLik(z2))+2*(1+1)

# Double check with codes
library(AICcmodavg)
# AIC, smaller (more negative) is better
AIC(z2)     # takes into account number of variables (penalized if more)
AICc(z2)    # takes into account number of variables and sample size

```

```
218 ################################################################################
219
220 ### 12. Fit 11 models, each with one of the following explanatory variables: income,
    gender, married, education, professional job, retired, unemployed, dual income,
    children, bought children's magazine previously, and bought parenting magazine
    previously. Record information about each model in the following table. Organize your
    models from lowest AIC value to highest AIC value. Example table below. (4 marks)
221
222 # already did income
223 # already did Married
224 mydata$Female <- factor(mydata$Female)              # Female
225 mydata$College <- factor(mydata$College)            # College
226 mydata$Professional <- factor(mydata$Professional)  # Professional
227 mydata$Retired <- factor(mydata$Retired)            # Retired
228 mydata$Unemployed <- factor(mydata$Unemployed)      # Unemployed
229 mydata$Dual.Income <- factor(mydata$Dual.Income)    # Dual.Income
230 mydata$Children <- factor(mydata$Children)          # Children
231 mydata$Prev.Child.Mag <- factor(mydata$Prev.Child.Mag)      # Prev.Child.Mag
232 mydata$Prev.Parent.Mag <- factor(mydata$Prev.Parent.Mag)    # Prev.Parent.Mag
233
234 str(mydata)
235
236 # 'data.frame': 673 obs. of  12 variables:
237 #    $ buy            : int  0 1 0 1 0 0 0 0 0 0 ...
238 # $ income         : int  24000 75000 46000 70000 43000 24000 26000 38000 39000 49000
    ...
239 # $ Female         : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 2 2 2 1 ...
240 # $ Married        : Factor w/ 2 levels "0","1": 1 2 2 2 1 2 2 2 1 2 ...
241 # $ College        : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 2 1 2 1 ...
242 # $ Professional   : Factor w/ 2 levels "0","1": 2 2 1 2 1 1 1 1 2 1 ...
243 # $ Retired        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 2 1 2 ...
244 # $ Unemployed     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
245 # $ Dual.Income    : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 1 1 1 ...
246 # $ Children       : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 1 1 1 ...
247 # $ Prev.Child.Mag : Factor w/ 2 levels "0","1": 1 2 1 2 1 1 1 1 1 1 ...
248 # $ Prev.Parent.Mag: Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
```

```
249
250 # $ Female          : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 2 2 2 1 ...
251 z3 <- glm(buy ~ Female, data=mydata, family="binomial"(link="logit"))
252 AIC(z3)                                # AIC value
253 -2*(logLik(z.null) - logLik(z3))       # G value
254 anova(z.null, z3, test="Chi")          # p value
255
256 # $ College         : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 2 1 2 1 ...
257 z4 <- glm(buy ~ College, data=mydata, family="binomial"(link="logit"))
258 AIC(z4)                                # AIC value
259 -2*(logLik(z.null) - logLik(z4))       # G value
260 anova(z.null, z4, test="Chi")          # p value
261
262 # $ Professional    : Factor w/ 2 levels "0","1": 2 2 1 2 1 1 1 1 2 1 ...
263 z5 <- glm(buy ~ Professional, data=mydata, family="binomial"(link="logit"))
264 AIC(z5)                                # AIC value
265 -2*(logLik(z.null) - logLik(z5))       # G value
266 anova(z.null, z5, test="Chi")          # p value
267
268 # $ Retired         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 2 1 2 ...
269 z6 <- glm(buy ~ Retired, data=mydata, family="binomial"(link="logit"))
270 AIC(z6)                                # AIC value
271 -2*(logLik(z.null) - logLik(z6))       # G value
272 anova(z.null, z6, test="Chi")          # p value
273
274 # $ Unemployed      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
275 z7 <- glm(buy ~ Unemployed, data=mydata, family="binomial"(link="logit"))
276 AIC(z7)                                # AIC value
277 -2*(logLik(z.null) - logLik(z7))       # G value
278 anova(z.null, z7, test="Chi")          # p value
279
280 # $ Dual.Income     : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 1 1 1 ...
281 z8 <- glm(buy ~ Dual.Income, data=mydata, family="binomial"(link="logit"))
282 AIC(z8)                                # AIC value
283 -2*(logLik(z.null) - logLik(z8))       # G value
284 anova(z.null, z8, test="Chi")          # p value
285
286 # $ Children        : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 1 1 1 ...
287 z9 <- glm(buy ~ Children, data=mydata, family="binomial"(link="logit"))
288 AIC(z9)                               # AIC value
289 -2*(logLik(z.null) - logLik(z9))      # G value
290 anova(z.null, z9, test="Chi")         # p value
291
292 # $ Prev.Child.Mag : Factor w/ 2 levels "0","1": 1 2 1 2 1 1 1 1 1 1 ...
293 z10 <- glm(buy ~ Prev.Child.Mag, data=mydata, family="binomial"(link="logit"))
294 AIC(z10)                              # AIC value
295 -2*(logLik(z.null) - logLik(z10))     # G value
296 anova(z.null, z10, test="Chi")        # p value
297
298 # $ Prev.Parent.Mag: Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
299 z11 <- glm(buy ~ Prev.Parent.Mag, data=mydata, family="binomial"(link="logit"))
300 AIC(z11)                              # AIC value
301 -2*(logLik(z.null) - logLik(z11))     # G value
302 anova(z.null, z11, test="Chi")        # p value
303
304 ##############################################################################
305
306 ### 15. Using R, create a graph of purchase vs. income with a method to visualize
    married. What are some overall patterns that you see? How can these help you answer
    your research question? (1 mark)
307
308 plot(jitter(buy, f = 0.5)~ income, data=mydata, pch=16, col = c("red","blue")[as.factor
    (mydata$Married)], main = "Buy Vs. Income", xlab = "Income ($/year)", ylab = "Buy")
309
310 legend(1, 1, legend=c("Not Married", "Married"),
311        col=c("red", "blue"), pch=16:16, cex=0.8)
312
```

```r
################################################################################

### 16. Create a model with the following explanatory variables: income, married, the interaction between income and married. (buy ~ income*married)
z12 <- glm(buy ~ income + Married + income*Married, data=mydata, family="binomial"(link="logit"))
AIC(z12)                                # AIC value
-2*(logLik(z.null) - logLik(z12))       # G value
anova(z.null, z12, test="Chi")          # p value

### 17. Does this model meet the assumptions of generalized linear models? (1 mark)

# . Statistical independence of observations
# . Correct specification of link function
# . Variance correspond to what is expected from the link function

### 18. What does the interaction between income and married allow in this model allow? (0.25 marks)


### 19. Test the likelihood of the whole model compared to the likelihood of the null model using a likelihood ratio test. Show your calculation of the test statistic using the log-likelihoods from R. Confirm your results using R. (1 mark)

-2*(logLik(z.null) - logLik(z12))       # G value
anova(z.null, z12, test="Chi")          # p value

### 20. Test each variable using a likelihood ratio test. This will require you to fit models that eliminate one variable. Show your calculation of the test statistics using the log-likelihoods from R. Confirm your results using R. (1 mark)

#There are a total of 4 models which could be used in such analyses which I will denote as:
    # (1) z1: Buy ~ income
    # (2) z2: Buy ~ Married
    # (3) z13  Buy ~ income + Married
    # (4) z12: Buy ~ income + Married + income*Married

# Testing the Significance of Married [i.e. Compare (1) and (4)]
anova(z12, z1, test="Chi")          # p value
-2*(logLik(z1) - logLik(z12))       # G value

# Testing the Significance of income [i.e. Compare (2) and (4)]
anova(z12, z2, test="Chi")          # p value
-2*(logLik(z2) - logLik(z12))       # G value

### 21. If both variables should remain in the model, test the interaction term only using a likelihood ratio test (you will have to fit a reduced Model A that excludes the interaction). Show your calculation of the test statistics using the log-likelihoods from R. Confirm your results using R. (1 mark)

# Model 3
z13 <- glm(buy ~ income + Married, data=mydata, family="binomial"(link="logit"))

# Testing the Significance of income [i.e. Compare (3) and (4)]
anova(z12, z3, test="Chi")          # p value
-2*(logLik(z3) - logLik(z12))       # G value

```

```
################################
# MODEL B
################################

### Model B: Model with income, married and professional job:
### 22. Create a model with the following explanatory variables: income, married, the interaction between income and married, job, and the interaction between job and income. (buy ~ income*married + income*job)

z14 <- glm(buy ~ income + Married + Professional + income*Married + income*Professional, data=mydata, family="binomial"(link="logit"))

# Testing the Significance of Unemployed [i.e. Compare (3) and (4)]
anova(z14, z12, test="Chi")       # p value
-2*(logLik(z12) - logLik(z14))    # G value

################################
# MODELS C THROUGH G
################################

#-------------------------------------------------------------------------------
# Testing the Significance of the variable Dual Income [i.e. Compare (Model A) and (Model C)]

# Model C
z15 <- glm(buy ~ income + Married + Dual.Income + income*Married + income*Dual.Income, data=mydata, family="binomial"(link="logit"))

anova(z15, z12, test="Chi")       # p value
-2*(logLik(z12) - logLik(z15))    # G value

#-------------------------------------------------------------------------------
# Testing the Significance of the variable Unemployed [i.e. Compare (Model A) and (Model D)]

# Model D
z16 <- glm(buy ~ income + Married + Unemployed + income*Married + income*Unemployed, data=mydata, family="binomial"(link="logit"))

anova(z16, z12, test="Chi")       # p value
-2*(logLik(z12) - logLik(z16))    # G value

#------------------------------------------------------------------------------- # Testing the Significance of the variable College [i.e. Compare (Model A) and (Model E)]

# Model E
z17 <- glm(buy ~ income + Married + College + income*Married + income*College, data=mydata, family="binomial"(link="logit"))

anova(z17, z12, test="Chi")       # p value
-2*(logLik(z12) - logLik(z17))    # G value
```

```r

#------------------------------------------------------------------------------------- #
Testing the Significance of the variable Prev.Child.Mag [i.e. Compare (Model A) and
(Model F)]

# Model F
z18 <- glm(buy ~ income + Married + Prev.Child.Mag + income*Married + income*Prev.Child
.Mag, data=mydata, family="binomial"(link="logit"))

anova(z18, z12, test="Chi")       # p value
-2*(logLik(z12) - logLik(z18))    # G value

#-------------------------------------------------------------------------------------
# Testing the Significance of the variable Female [i.e. Compare (Model A) and (Model G)]

# Model G
z19 <- glm(buy ~ income + Married + Female + income*Married + income*Female, data=mydata,
family="binomial"(link="logit"))

anova(z19, z12, test="Chi")       # p value
-2*(logLik(z12) - logLik(z19))    # G value


# Model H
z20 <- glm(buy ~ income + Married + Female + income*Married, data=mydata, family="binomia
l"(link="logit"))

anova(z19, z20, test="Chi")       # p value
-2*(logLik(z20) - logLik(z19))    # G value

################################
# FINAL MODEL
################################

### 25. For your final model, calculate the Pseudo-R2 and the scaled Pseudo-R2. (1 mark)

# Pseudo R2
1 - (logLik(z.null)/logLik(z12))^(2/nrow(mydata))

logLik(z.null)
# 'log Lik.' -323.0265 (df=1)

#Scaled R2
(1 - (logLik(z.null)/logLik(z12))^(2/nrow(mydata)))/(1-logLik(z.null)^(2/nrow(mydata)))
(1 - (logLik(z.null)/logLik(z12))^(2/nrow(mydata)))/(1 - (-323.0265^(2/nrow(mydata))))

### 26. Calculate the AIC value for your final model. Compare the AIC value of your final
model to the AIC values for the models that had only one of the variables that are
included in your final model (Single variable models). Put all of these models and AIC
values in a table to make them easy to compare. How much does the AIC value improve from
the single variable models of income, married and professional job to your final model?
(2 marks)

library(AICcmodavg)
# AIC, smaller (more negative) is better
AIC(z12)     # takes into account number of variables (penalized if more)
# [1] 233.6734
AICc(z12)    # takes into account number of variables and sample size
# [1] 233.7333

################################
# CLASSIFICATION TABLE
################################

### 27. Create a classification table for this data based on the final model. Include the
full classification table in your output. Remember that you can output dataframes to a
.csv file using write.csv().

# This will try out many different cut-off points to give you an idea of how to maximize
or minimize different values.
# For example, you might want to maximize percentage of correct predictions.
# For example, you might want to minimize false negatives.

```

```r
# Create an empty dataframe that you will fill with
df <- data.frame(matrix(ncol = 9, nrow = 51))
colnames(df) <- c("correct.event", "correct.non.event", "incorrect.event", "incorrect.non.event", "correct.percent", "sensitivity", "specificity", "false.pos", "false.neg")
df

prob.level <- seq(0, 1, length.out=51) # create a vector with different possible probabilities
prob.level
class.table.data <- cbind(prob.level, df) # combine your vector of probabilities and your empty dataframe
class.table.data # Your dataframe has one row for each probability cut-off

# fill empty cells in your dataframe with 0
class.table.data$correct.non.event <- rep(c(0), c(51))
class.table.data$correct.event <- rep(c(0), c(51))
class.table.data$incorrect.non.event <- rep(c(0), c(51))
class.table.data$incorrect.event <- rep(c(0), c(51))
class.table.data

# This loop will try out the different probability cut-off values and fill in how many correct and incorrect events and non-events you have based on your data.
for (i in 1:51) {
    class.table <- table(mydata$buy, fitted(z12) > class.table.data$prob.level[i])

    col.true.num <- grep("TRUE", colnames(class.table))
    col.false.num <- grep("FALSE", colnames(class.table))

    if (length(col.true.num) > 0) {
        class.table.data$incorrect.non.event [i] <- class.table[1, col.true.num]
        class.table.data$correct.event [i] <- class.table[2, col.true.num] }

    if (length(col.false.num) > 0) {
        class.table.data$correct.non.event [i] <- class.table[1, col.false.num]
        class.table.data$incorrect.event [i] <- class.table[2, col.false.num] }  }

class.table.data

# You will use this information to fill in the rest of your classification table.
class.table.data$correct.percent <- (class.table.data$correct.event + class.table.data$correct.non.event)/nrow(mydata)
class.table.data$sensitivity <- (class.table.data$correct.event)/nrow(mydata)
class.table.data$specificity <- (class.table.data$correct.non.event)/nrow(mydata)
class.table.data$false.neg <- (class.table.data$incorrect.non.event)/nrow(mydata)
class.table.data$false.pos <- (class.table.data$incorrect.event)/nrow(mydata)
class.table.data

write.csv(class.table.data, file = "ClassTable.csv")

### 28. For sensitivity, specificity, false negatives, false positives, which do you want to maximize or minimize, and which do you not care about? Why?     (2 marks)

### 29. You decide to maximize the percentage of correct of predictions. Why? What probability cut-off should you use to maximize the percent of correct predictions? (0.5 marks)
```