**Assignment 4**

*The objective of this assignment is to assess the email marketing strategy of a certain firm. It also helps you understand some of the metrics that can be used to examine the effectiveness of email marketing.*

*Consumers are sent emails as part of the email (digital) marketing strategy of the firm and consumers can respond to these emails by opening emails. Moreover, if the consumers are sufficiently annoyed with the emails (or for a variety of other reasons) they may decide to unsubscribe (also called opt-out) from the email lists. Also included is information about the number of visits the consumers make to the store (this includes both the combination of online and offline store).*

*The following are the columns with their explanation.*

*Customer: The ID of the consumer to track them.*

*Date Format: This indicates the week the email was sent.*

*Week: This is a numerical representation of the week.*

*Weekend: This takes the value of 1 is the email is sent on a weekend; 0 otherwise.*

*Holiday: This takes the value of 1 if the email is sent on a holiday; 0 otherwise.*

*Subject Length: The length of the subject line of the email header (of the specific campaign).*

*Opt-out: This takes the value of 1 for consumers who have eventually opted out; 0 for those who have not.*

*When Opt-out: This denotes the week when the respective consumer has opted out. This takes the value of 1 for the week when the customer opts out. Note: This should also be 1 for all the weeks after, since once the customer opts out he/she does not opt in.*

*Opened: This denotes if the particular email campaign was opened by the customer.*

*Total\_sent: This denotes the total cumulative emails that have been sent to the customer until that time period.*

*Total\_open: This denotes the total cumulative emails that have been opened by the customer until that time period (out of the total sent).*

*Total\_visits: The total number of store visits made by the customer during that week (a metric to measure the effectiveness that email campaigns have on other activities).*

*The specific questions that need to be answered for the assignment are mentioned in the following page.*

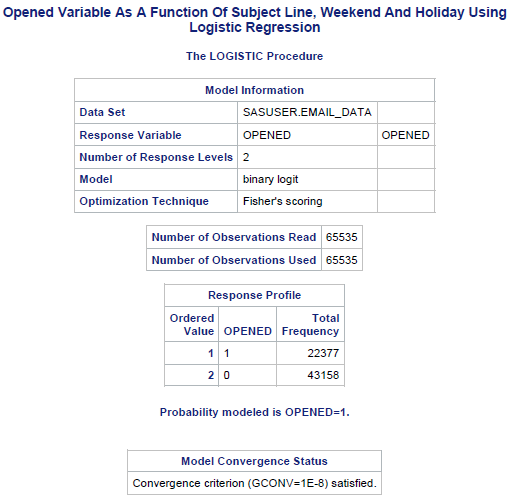
1. *It is instructive to understand the factors that might impact the decision of consumers to open a specific email campaign or not. Let us understand how the factors/variables of subject line length, weekend and holiday affect if a specific email campaign is opened by customers. You should estimate the model (wherein the dependent variable is the probability if a campaign is opened by the customer and the independent variables are subject line length, weekend and holiday). The model will be a logit model as follows:*

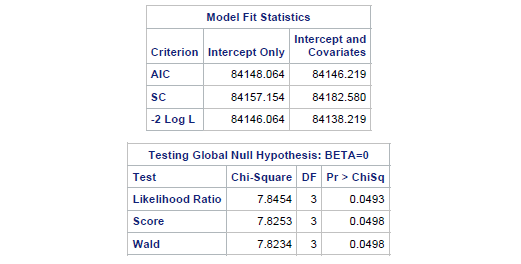
**

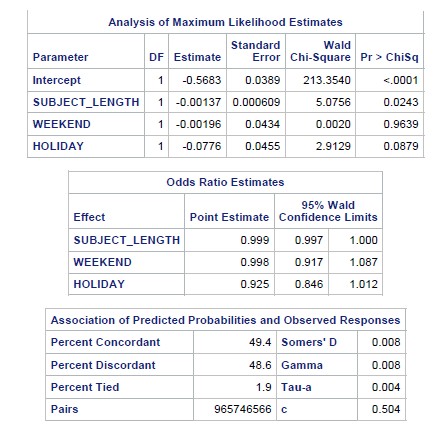
*Report your results and summarize your findings. How do you interpret your parameters/coefficients? What do you find?*

**Regression Results:**

Performing a logistic regression and creating the above model yields the following outputs for the maximum likelihood estimates:







The model can therefore be written as follows:



**Analysis and Conclusion:**

­From the maximum likelihood estimates of the coefficients, the following conclusions can be obtained:

* For every one unit increase in the subject length, the log odds of opening a marketing email (versus not opening the email) decreases by 0.000137.
* For every day that is a weekend, the log odds of opening a marketing email (versus not opening the email) decreases by 0.00196.
* For every day that is a holiday, the log odds of opening a marketing email (versus not opening the email) decreases by 0.0776.

Observing the model fit statistics, a larger value of -2LogL in intercept only model as opposed to intercept with covariates indicates that intercept only model is a better estimator of log odds of the emails opened.

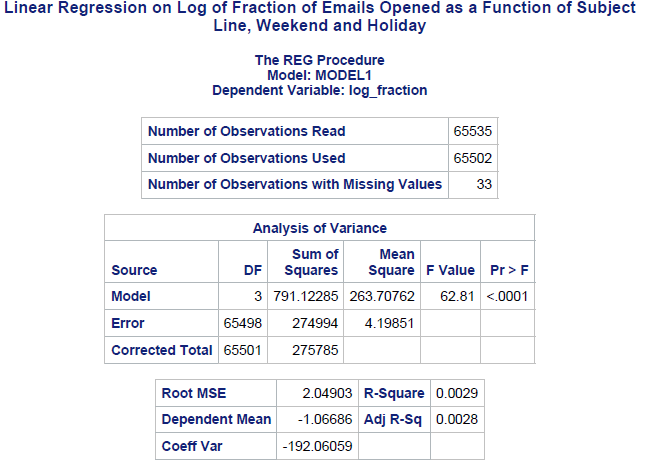
**Findings:**

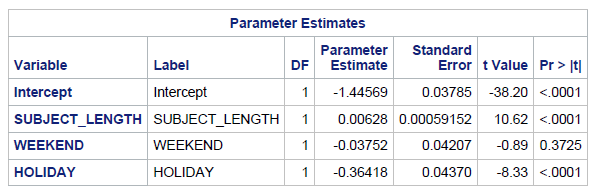
* On contrary to the general belief that more emails would be opened during holidays and weekends, we would be able to conclude from the p-values for these variables that they are statistically insignificant in the prediction of whether or not an email would be opened.
* Only length of the subject line seems to be a good predictor of the log odds of opening an email. It would be better to create a new model omitting weekend and holidays from the set of independent variables.

1. *Another relevant metric that can be used is the fraction of emails opened by the customers to the number of the emails received by him/her. Let . Since this is a fraction, it is better to use this after transforming it appropriately. So consider the following regression with the logistic transformation given below. You need to estimate the model given below.*

*Report your results and summarize your findings. How do you interpret your parameters/coefficients? What do you find? Contrast your results here with those of #1.*

**Regression Results**





The regression model is as follows:

**Analysis and Conclusion:**

From the extremely low R squared value of this model, it can be understood that this model is very bad predictor of the log odds of number of marketing emails opened.

Just going through the coefficients, we would be able to arrive at the following conclusions:

* For every unit increase in the length of subject line, the log odds of opening an email increases by 0.00628.
* For every day that is a weekend, the log odds of opening an email reduces by 0.03752.
* For every day that is a holiday, the log odds of opening an email reduces by 0.36418.

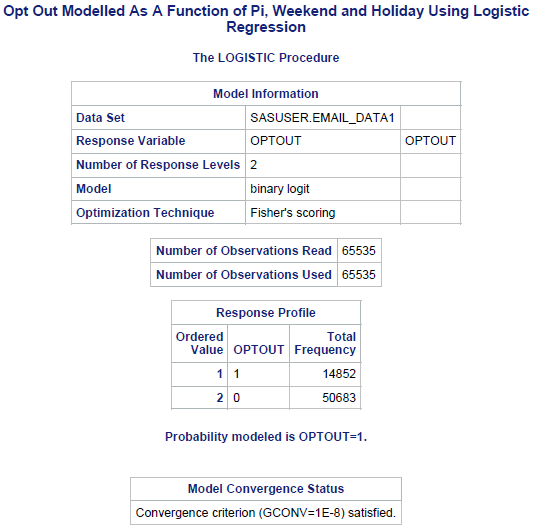
**Findings:**

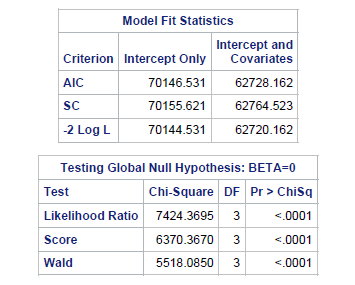
* The p-values of the variable weekend is significantly higher than alpha. Therefore this variable could be omitted to obtain a better model prediction of the log odds of opening an email.
* Though the values of parameters are expected to be the same while running both a logistic regression and a linear regression on the log(odds), it can be observed that parameter estimates from both the models are not consistent with each other and are in fact contradicting each other.
* According to the logistic regression model, the parameter estimates and their corresponding p-values, suggested that the variables Weekend and Holiday are not significant for the prediction of the log odds of an email being opened. Also, it could be understood that the log odds of an email being opened is reduced by a unit increase in the subject line length.
* However, according to the linear regression model, it can be observed that in addition to the Subject Line Length, the variable Holiday also helps in prediction of the values of log odds of an email being opened. Also, as per this model, log odds of an email being opened increases with increase in Subject Line Length.
* This contradiction could be because of the imputation of zero’s at multiple levels with the value 0.01 to avoid any errors due to division by zero or taking a log of zero. The contradiction could also be due to the internal optimization techniques used by the package proc logistic as opposed to the proc reg in SAS.
* The extremely low R squared values and contradicting results obtained from this model, helps us conclude that it is always advisable to use a logistic regression for a binary target variable and modelling it manually as a linear regression is not advisable.

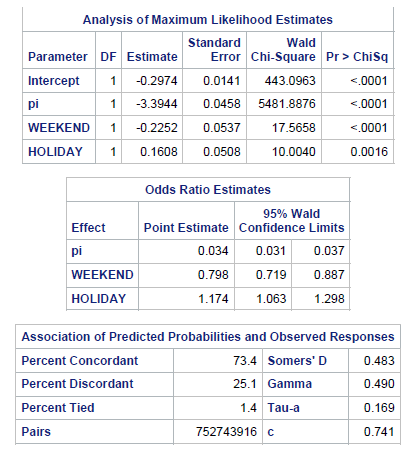
1. *Let us understand the factors that impact consumers opt-out decision (equivalently, consumers decision to unsubscribe) from the email campaign of the company. Moreover, we will do this at the individual consumer level and use the variable you created in #2. Estimate the following model:*

*Report your results and summarize your findings. How do you interpret your parameters/coefficients? What do you find? Contrast your results here with those of #1.*

**Regression Results:**







1. Now, let us see how the metrics at the store are influenced by the email campaigns. We will use store visits as a proxy for this. Estimate the following model:

Note: Since total visits is an integer, to get appropriate results you should either do a Poisson regression or take the logarithm of total visits and use that instead.

Report your results and summarize your findings. How do you interpret your parameters/coefficients? What do you find?

1. Compare and contrast consumers who unsubscribe versus those who do not. Also, compare the behavior of consumers before they opt out and after they opt out (this is relevant for only those consumers who have opted out). You can develop appropriate models with the relevant variables as you wish (e.g., store visits, opt out etc.).