**DSCI 5240: Data Mining**

**Homework 01**

**Group 09:** Anji Lanke, Chris Benton, Cody Holmes, Raheyma Khan, Srinivas Rao Kolla

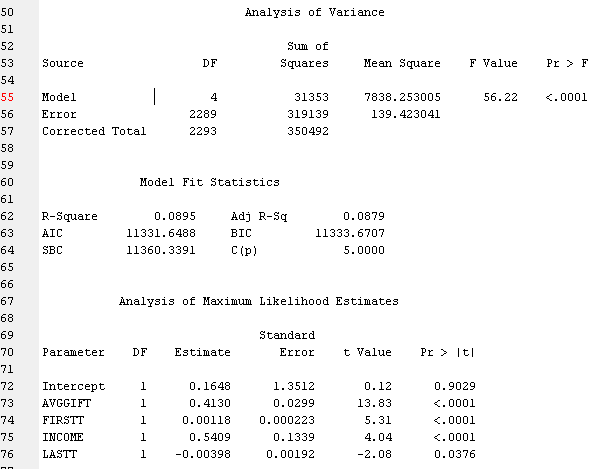
# **HW 1 – 1:**

# Section 1: Brief Description of the Data File

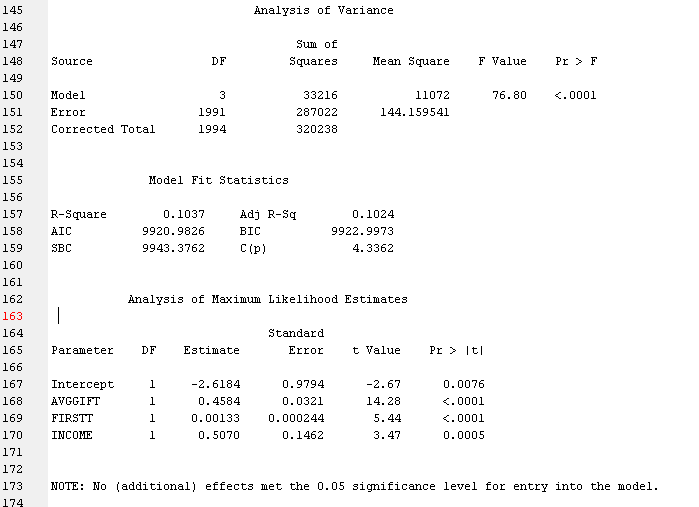
The data consists of 22 variables and 2988 observations; TARGET\_B and TARGET\_D are dependent variables, while the rest are independent variables used for predicting the dependent variables. TARGET\_B is a dummy variable with values 0 or 1, and TARGET\_D is an interval. The dependent variables include five nominal variables and 15 interval variables. In part 1 of the assignment we will use the predictor variables AVGGIFT, FIRSTT, INCOME, LASTT, AGE, MALEMILI and MALEVET to form simple linear regression models that will predict the response variable TARGET\_D.

# Section 2: Linear Regression Output

## Regression Output for AVGGIFT, FIRSTT, INCOME, and LASTT Variables



## Regression Output for Stepwise Model of Variables AVGGIFT, FIRSTT, INCOME, LASTT, AGE, MALEMILI, and MALEVET



# Section 3: Answers to Q1 – Q7

## Q1 Answer

8.95% of the variance in the dependent variable of TARGET\_D is explained by the independent variables of AVGGIFT, FIRSTT, INCOME, and LASTT.

## Q2 Answer

1. For every one unit increase in AVGGIFT, TARGET\_D is expected to increase by 0.1646, holding all other variables constant.
2. For every one unit increase in FIRSTT, TARGET\_D is expected to increase by 0.4130, holding all other variables constant.
3. For every one unit increase in INCOME, TARGET\_D is expected to increase by 0.00118, holding all other variables constant.
4. For every one unit increase in LASTT, TARGET\_D is expected to decrease by 0.00398, holding all other variables constant.

## Q3 Answer

To determine if the variables of AVGGIFT, FIRTT, INCOME, and LASTT are statistically significant, a level of confidence must be established. Common levels of confidence are p-values of 0.1, 0.05, and 0.01. If we assume a level of significance at the p-value of 0.05, all variables are statistically significant predictors of TARGET\_D. Interestingly, if we assumed a level of significance at the p-value of 0.01, all variables would still be significant predictors except the LASTT variable.

## Q4 Answer

The variables now selected in the model are AVGGIFT, FIRSTT, and INCOME.

## Q5 Answer

Given the new stepwise model, three of the original four variables are significant well beyond a p-value of 0.01. The LASTT variable is not.

## Q6 Answer

The paradox is explained by the fact that stepwise regression takes multiple combinations of variables to see which produce the best model. Best model is defined as a model that most increases the R-square value. Within each iteration of the stepwise regression, for instance, the complete output shows the value of R-square increasing at every step, finally stopping an R-square value of 0.1037.

## Q7 Answer

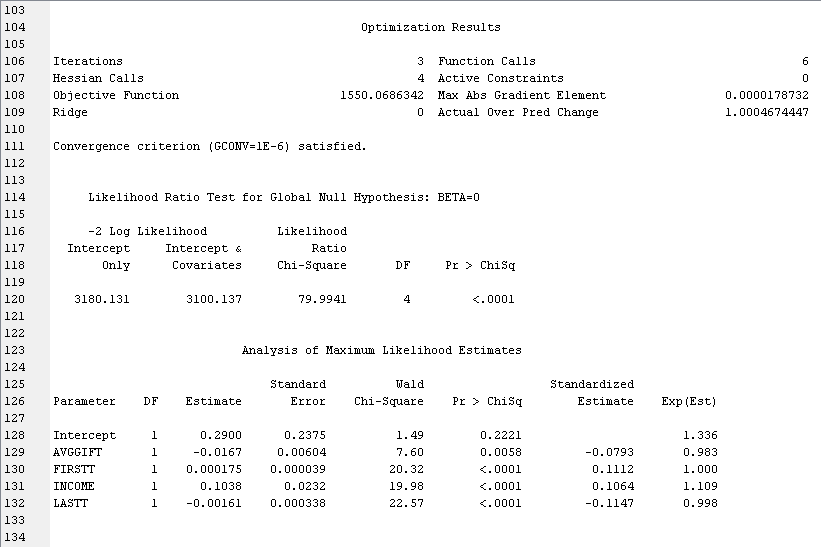
Adding the three extra variables did not have any effect on the model. This is because our final model only contains the three variables AVGVGIFT, FIRSTT and INCOME which were part of the first model. Moreover, the final model has a higher R-square value because it does not include the variable LASTT. LASTT was becoming insignificant at p-value = 0.01 in the first model, and therefore has been removed by the final model.

**HW 1 - 2:**

# Section 1: Brief Description of the Data File

The data consists of 22 variables and 2988 observations; TARGET\_B and TARGET\_D are dependent variables, while the rest are independent variables used for predicting the dependent variables. TARGET\_B is a dummy variable with values 0 or 1, and TARGET\_D is an interval. The dependent variables include five nominal variables and 15 interval variables. In part 2 of the assignment we will use the predictor variables AVGGIFT, FIRSTT, INCOME, and LASTT to form a logistic regression model that will predict the response variable TARGET\_B.

# Section 2: Logistic Regression Output





# Section 3: Answers to Q1 – Q4

**Q1 Answer**

The model is valid. The estimated probability of the predicted values for TARGET\_B falls between 0 and 1. Moreover, -2xlog likelihood has a smaller value for model with intercept and covariates. The chi square test also shows the model is valid at all levels of significance.

**Q2 Answer**

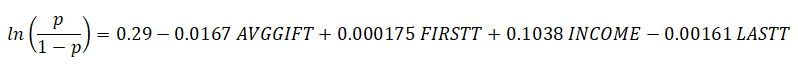
The variables AVGGIFT and LASTT have odds ratios less than 1, which implies that the probability of them not occurring is greater than the probability of them occurring. FIRSTT has an odds ratio equal to 1 which means that it is equally likely for the event to occur or not. INCOME has odds ratio greater than 1 so the probability of it occurring is greater.

**Q3 Answer**

A level of confidence such as .1, .05, or .01 has to be set in order to determine significance. Based on any of the common levels of confidence, all variables (FIRSTT, INCOME, LASTT, and AVGGIFT) are significant predictors for TARGET\_B. While AVGGIFT has a Chi-Square of .0058, it is still below the lowest common level of confidence of .01.

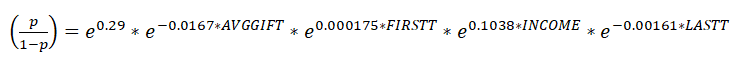
**Q4 Answer**

The equation of the model is



The odds ratio is equal to *(p/1-p).* The coefficients of independent variables in the above equation give the rate of change in the log of odds ratio as the independent variables change. A one unit increase in AVGGIFT reduces the log of odds ratio by 0.0167 units. A one unit increase in FIRSTT increases the log of odds ratio by 0.000175 units. A one unit increase in INCOME increases the log of odds ratio by 0.1038 units. A one unit increase in LASTT decreases the log of odds ratio by 0.00161 units.

Changing our logistic regression equation to the one below makes it easier to understand the coefficients.



Now, a one unit increase in AVGGIFT causes the odds ratio to increase by e^-0.0167 = 0.983 units. A one unit increase in FIRSTT causes the odds ratio to increase by e^0.000175 = 1.000 units. A one unit increase in INCOME causes the odds ratio to increase by e^0.1038 = 1.109 units. A one unit increase in LASTT causes the odds ratio to increase by e^-0.00161 = 0.998 units.