Group 9: Anji Lanke, Chris Benton, Raheyma Khan, Sriniva Rao Kolla, and Cody Holmes
HW1-1

DSCI 5240: Data Mining
Homework 01

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

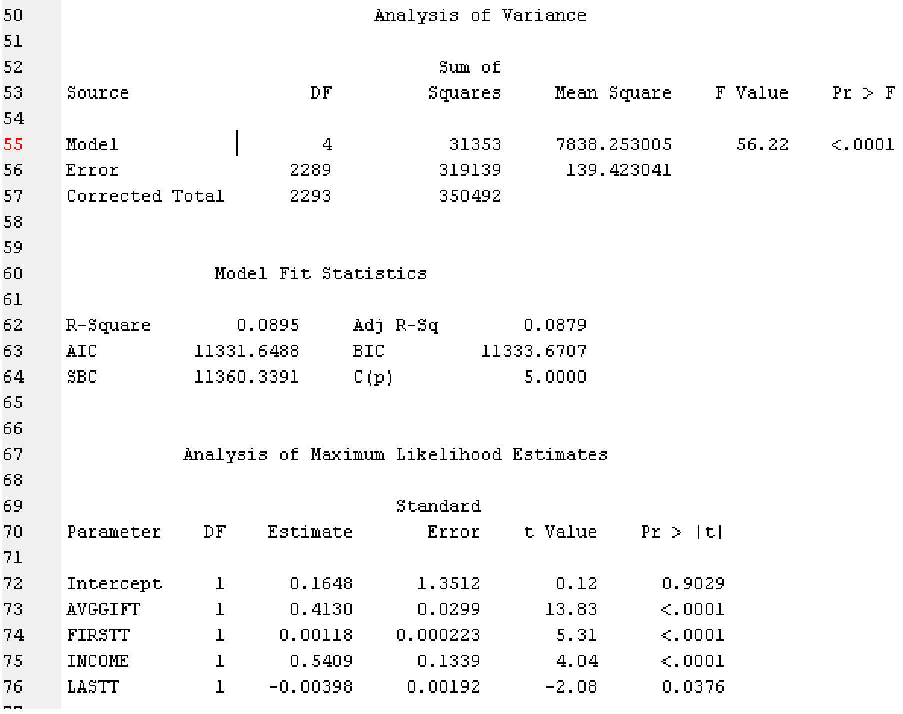
HW 1 – 1:

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

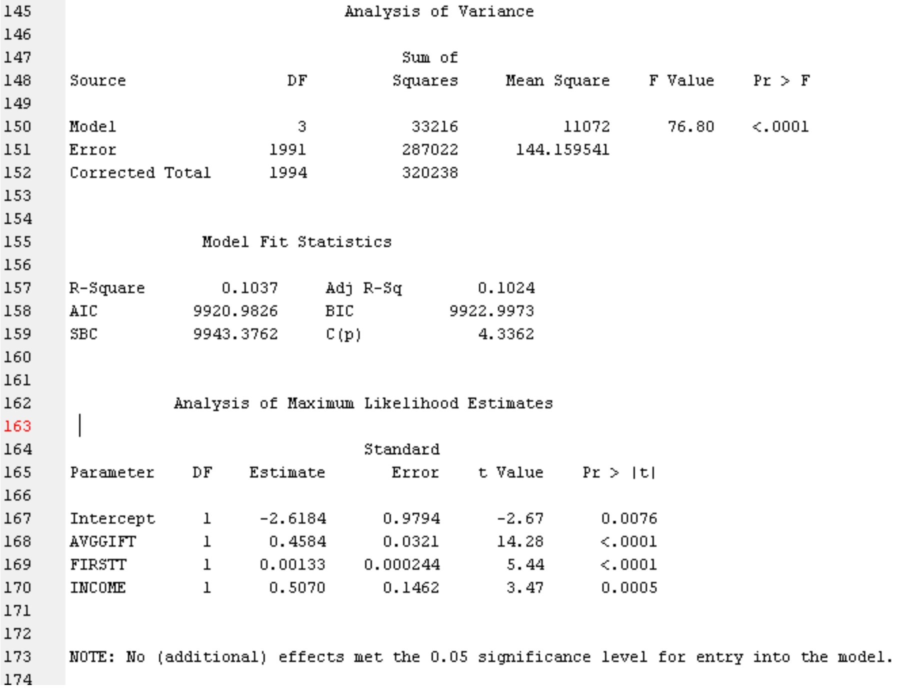
Regression Output for AVGGIFT, FIRSTT, INCOME, and LASTT Variables



```
                              Analysis of Variance

                                     Sum of
Source                 DF           Squares      Mean Square    F Value    Pr > F

Model                   4             31353      7838.253005      56.22    <.0001
Error                2289            319139       139.423041
Corrected Total      2293            350492


              Model Fit Statistics

R-Square        0.0895     Adj R-Sq        0.0879
AIC         11331.6488     BIC         11333.6707
SBC         11360.3391     C(p)            5.0000


           Analysis of Maximum Likelihood Estimates

                                 Standard
Parameter    DF    Estimate         Error    t Value    Pr > |t|

Intercept     1      0.1648        1.3512       0.12      0.9029
AVGGIFT       1      0.4130        0.0299      13.83      <.0001
FIRSTT        1     0.00118      0.000223       5.31      <.0001
INCOME        1      0.5409        0.1339       4.04      <.0001
LASTT         1    -0.00398       0.00192      -2.08      0.0376
```

Group 9: Anji Lanke, Chris Benton, Raheyma Khan, Sriniva Rao Kolla, and Cody Holmes
HW1-1

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



```
                               Analysis of Variance


                                     Sum of
Source                  DF          Squares     Mean Square    F Value    Pr > F

Model                    3            33216           11072      76.80    <.0001
Error                 1991           287022      144.159541
Corrected Total       1994           320238


              Model Fit Statistics

R-Square        0.1037     Adj R-Sq        0.1024
AIC          9920.9826     BIC          9922.9973
SBC          9943.3762     C(p)            4.3362


           Analysis of Maximum Likelihood Estimates

                                Standard
Parameter    DF    Estimate       Error    t Value    Pr > |t|

Intercept     1     -2.6184      0.9794      -2.67      0.0076
AVGGIFT       1      0.4584      0.0321      14.28      <.0001
FIRSTT        1     0.00133    0.000244       5.44      <.0001
INCOME        1      0.5070      0.1462       3.47      0.0005


NOTE: No (additional) effects met the 0.05 significance level for entry into the model.
```

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.

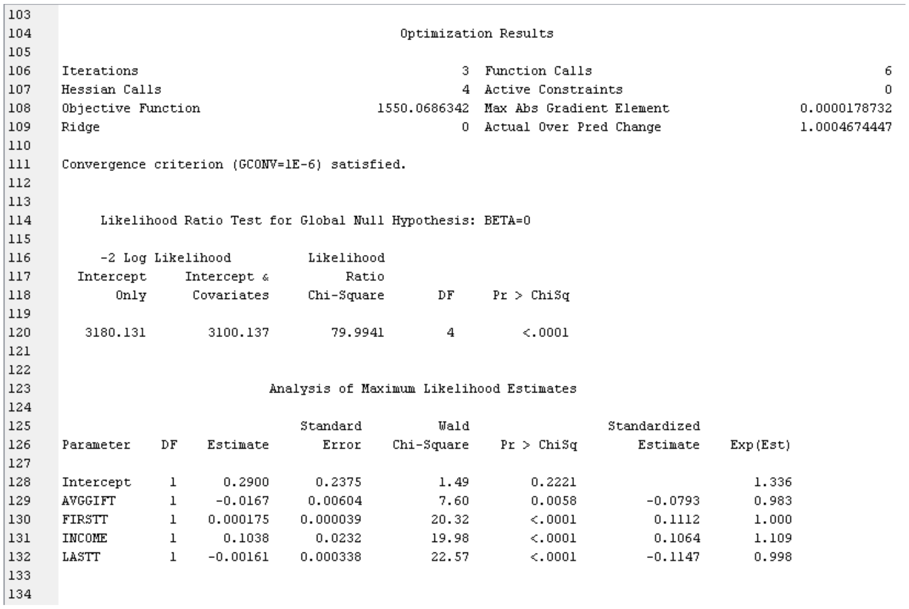
HW 1 - 2:

Section 1: Brief Description of the Data File

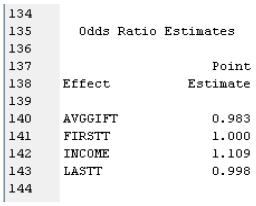
The data consists of 22 variables and 2988 observations; TARGET_B and TARGET_D are dependent variables, while the rest are independent variables used for predicting the dependent variables. TARGET_B is a dummy variable with values 0 or 1, and TARGET_D is an interval. The dependent variables include five 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

Group 9: Anji Lanke, Chris Benton, Raheyma Khan, Sriniva Rao Kolla, and Cody Holmes
HW1-1



```
                                          Optimization Results

Iterations                                        3  Function Calls                                  6
Hessian Calls                                     4  Active Constraints                              0
Objective Function                     1550.0686342  Max Abs Gradient Element             0.0000178732
Ridge                                             0  Actual Over Pred Change              1.0004674447

Convergence criterion (GCONV=1E-6) satisfied.


     Likelihood Ratio Test for Global Null Hypothesis: BETA=0

     -2 Log Likelihood          Likelihood
  Intercept     Intercept &          Ratio
       Only      Covariates     Chi-Square       DF     Pr > ChiSq

   3180.131        3100.137        79.9941        4         <.0001


                           Analysis of Maximum Likelihood Estimates

                                  Standard          Wald                  Standardized
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq        Estimate    Exp(Est)

Intercept     1      0.2900      0.2375          1.49        0.2221                       1.336
AVGGIFT       1     -0.0167     0.00604          7.60        0.0058         -0.0793       0.983
FIRSTT        1    0.000175    0.000039         20.32        <.0001          0.1112       1.000
INCOME        1      0.1038      0.0232         19.98        <.0001          0.1064       1.109
LASTT         1    -0.00161    0.000338         22.57        <.0001         -0.1147       0.998
```



```
   Odds Ratio Estimates

                   Point
Effect          Estimate

AVGGIFT            0.983
FIRSTT             1.000
INCOME             1.109
LASTT              0.998
```

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

Commented [KR1]: @Lanke, Anji I have copied your answers here.
Can you please explain your first three answers. I have different answers for them actually.

Commented [LA2R1]: For Q1, I have considered probability instead of odds ratio. Could you please correct me if i am wrong.

Commented [KR3R1]: I'm not sure about this answer myself. I was thinking we will prove validity by likelihood ratio chi square or -2xlog likelihood. But your answer is also making sense.

Commented [BC4R1]: Can we include a bit about the Likelihood Ratio as well? I think both would work well for this question and it might be good to inlcude both.

Commented [KR5]: My answer for Q2 is also different. Please verify.

Commented [LA6R5]: Both of our answers are same. I did not make it clear interms of event occur or not. Replaced my answer with yours.

Commented [KR7R5]: Okay. Thank you for the clarification!

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

Commented [KR8]: My answer for Q4 is different from Anji's. Please let me know if its correct.

Commented [LA9R8]: My bad, i missed the exponential part in the formula. We will go with your answer.

$$ln\left(\frac{p}{1-p}\right) = 0.29 - 0.0167\ AVGGIFT + 0.000175\ FIRSTT + 0.1038\ INCOME - 0.00161\ LASTT$$

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.

$$\left(\frac{p}{1-p}\right) = e^{0.29} * e^{-0.0167*AVGGIFT} * e^{0.000175*FIRSTT} * e^{0.1038*INCOME} * e^{-0.00161*LASTT}$$

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.