**CSC423: Data Analysis And Regression / CSC 324: Data Analysis & Statistical Software II**

**Assignment-6** | **Total Points: 10 pts for CSC 423/ CSC 324**

**Due Date: 03/06/2018 by 11:59 pm**

***Note: For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.***

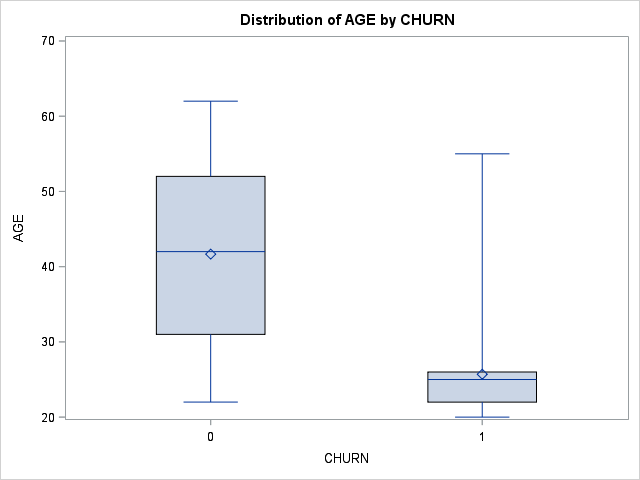
**Problem 1 [10 pts] Churn analysis – to be answered by everyone**

Given the large number of competitors, cell phone carriers are very interested in analyzing and predicting customer retention and churn. The primary goal of churn analysis is to identify those customers that are most likely to discontinue using your service or product. The dataset churn\_train.csv contains information about a random sample of customers of a cell phone company. For each customer, company recorded the following variables:

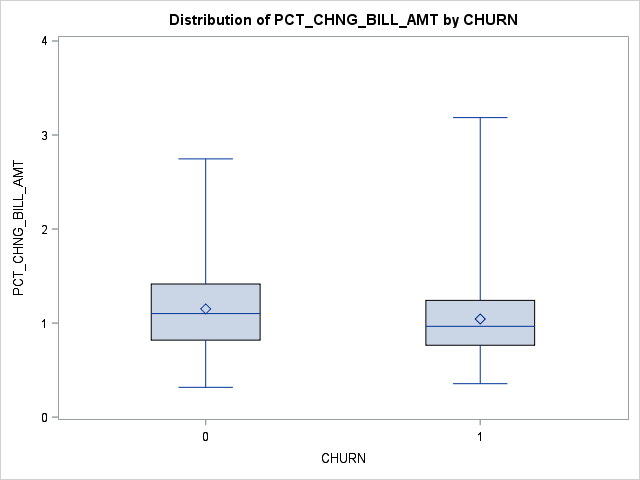
1. CHURN: 1 if customer switched provider, 0 if customer did not switch
2. GENDER: M, F
3. EDUCATION (categorical): code 1 to 6 depending on education levels
4. LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT: No. of days since last price plan change
5. TOT\_ACTV\_SRV\_CNT: Total no. of active services
6. AGE: customer age
7. PCT\_CHNG\_IB\_SMS\_CNT: Percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS
8. PCT\_CHNG\_BILL\_AMT: Percent change of latest 2 months bill amount wrt previous 4 months bill amount
9. COMPLAINT: 1 if there was at least a customer’s complaint in the two months, 0 no complaints

The company is interested in a churn predictive model that identifies the most important predictors affecting probability of switching to a different mobile phone company (churn = 1). Answer the following questions:

1. Create two boxplots to analyze the observed values of age and PCT\_CHNG\_BILL\_AMT by churn value. Analyze the boxplots and discuss how customer age and changes in bill amount affect churn probabilities. Include the boxplots.



The older the person, the less likely they will switch providers

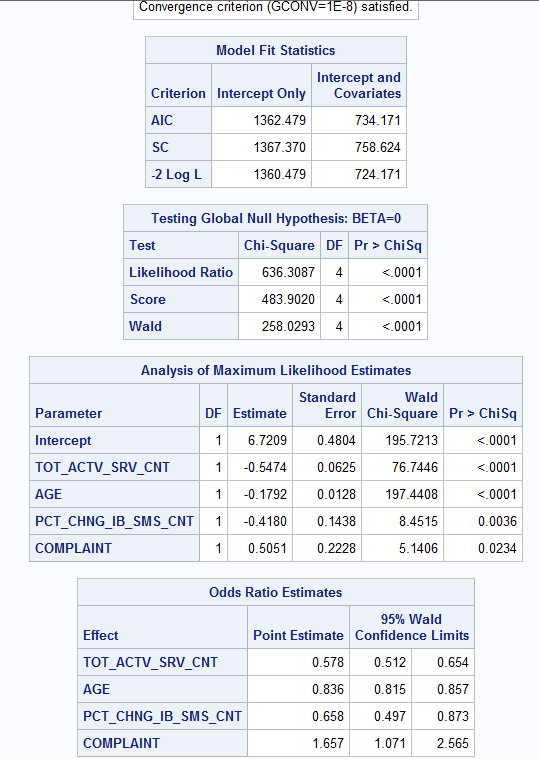
. 

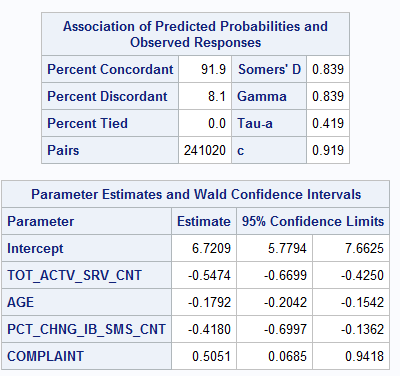
There is not much difference in the percent change bill amount between the two churn values.

1. Fit a logistic regression model to predict the churn probability using the data in the dataset (Churn is the response variable and the remaining variables are the independent x-variables). Remove x-variables that are not significant using alpha=0.05. Include the SAS output. Write down the expression of the fitted model. (HINT: probability of interest is p = pr(churn = 1)

I removed gender, last\_price\_place\_chng, pct\_chng\_bill\_amnt, and all education dummy variables because they were insignificant.

Log(churn =1/churn=0) = 6.72 – 0.55Tot\_ACTV\_SRV\_CNT -0.18Age – 0.42Pct\_chng\_ib\_sms\_cnt +0.51Complaint





1. Analyze the final logistic regression model and discuss the effect of each variable on the churn probability. Discuss results in terms of odds ratios.

With each addition of active services the probability of churn decreases by 57.8%.

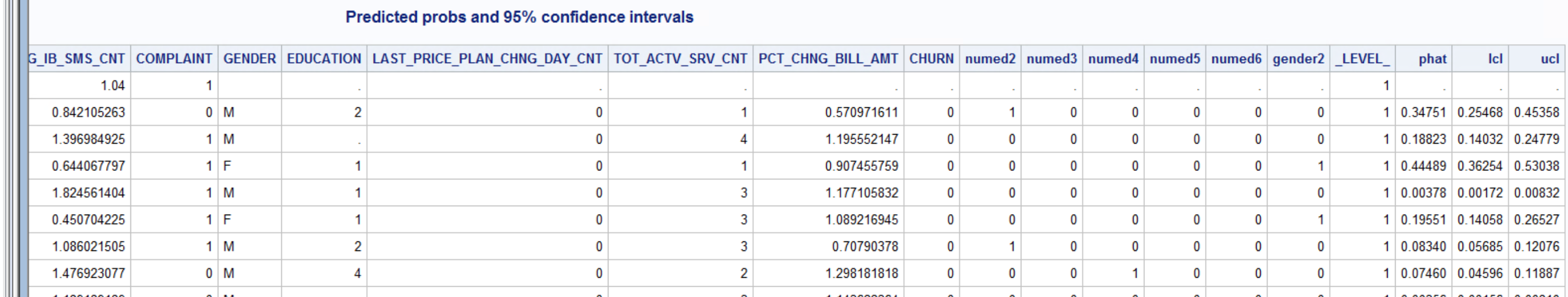
With each year of age, the probability of churn decreases by 83.6%.

With each percentage increase of the latest 2 months incoming SMS wrt previous 4 months incoming SMS, the probability of churn decreases by 65.8%.

With each additional complaint, the probability of churn increases by 165%.

1. Using SAS, compute the predicted churn probability and the confidence interval for a male customer who is 43 years old, and has the following information LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT=0, TOT\_ACTV\_SRV\_CN=4, PCT\_CHNG\_IB\_SMS\_CNT= 1.04, PCT\_CHNG\_BILL\_AMT= 1.19, and COMPLAINT =1. Include the output, interpret and explain the 3 values you obtained.

I was not able to obtain a predicted value. I am sure my code is correct, but phat, lcl, and ucl all came up blank.



1. Copy and paste your FULL SAS code into the word document along with your answers.

Title "Analysis of Churn data";

**PROC** **IMPORT** datafile = 'S:\Homeworks\churn\_train.csv' out=churn replace;

delimiter = ',';

getnames=YES;

**proc** **print**;

**run**;

**data** churn;

set churn;

numed2 = (education = **2**);

numed3 = (education = **3**);

numed4 = (education = **4**);

numed5 = (education = **5**);

numed6 = (education = **6**);

gender2 = (gender = 'F');

**run**;

**proc** **print**;

**run**;

\*boxplot;

**proc** **sort** data = churn;

by churn;

**proc** **boxplot** data = churn;

plot age\*churn;

**run**;

**proc** **sort** data = churn;

by churn;

**proc** **boxplot** data = churn;

plot PCT\_CHNG\_BILL\_AMT \*churn ;

**run**;

\*fit model;

**proc** **logistic** data=churn;

model churn(event='1') = gender2 last\_price\_plan\_chng\_day\_cnt tot\_actv\_srv\_cnt age pct\_chng\_ib\_sms\_cnt pct\_chng\_bill\_amt complaint numed2 numed3 numed4 numed5 numed6;

**run**;

\*fit model using setpwise;

\*fit model;

**proc** **logistic** data=churn;

model churn(event='1') = gender2 last\_price\_plan\_chng\_day\_cnt tot\_actv\_srv\_cnt age pct\_chng\_ib\_sms\_cnt pct\_chng\_bill\_amt complaint numed2 numed3 numed4 numed5 numed6/selection = stepwise;

**run**;

\*re-run model with significant variables;

**proc** **logistic** data=churn;

model churn(event='1') = tot\_actv\_srv\_cnt age pct\_chng\_ib\_sms\_cnt complaint/cl;

**run**;

\*part d;

**data** new;

input age TOT\_ACTV\_SRV\_CN PCT\_CHNG\_IB\_SMS\_CNT COMPLAINT ;

datalines;

43 4 1.04 1 0

;

**proc** **print**;

**run**;

**data** pred;

set new churn;

**run**;

**proc** **print**;

**run**;

\*run prediction, p is the predicted probabilty, lower is the lower CI, upper is the upper CI;

**proc** **logistic** data=pred;

model churn(event='1') = tot\_actv\_srv\_cnt age pct\_chng\_ib\_sms\_cnt complaint;

output out=pred2 p=phat lower=lcl upper=ucl;

**run**;

**proc** **print** data = pred2;

title 'Predicted probs and 95% confidence intervals';

**run**;