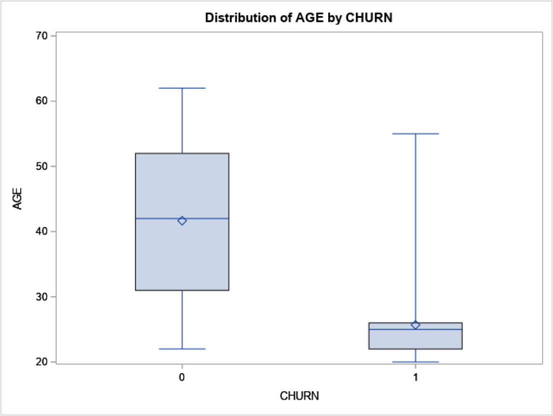
Ramadan Gannud
Assignment-6

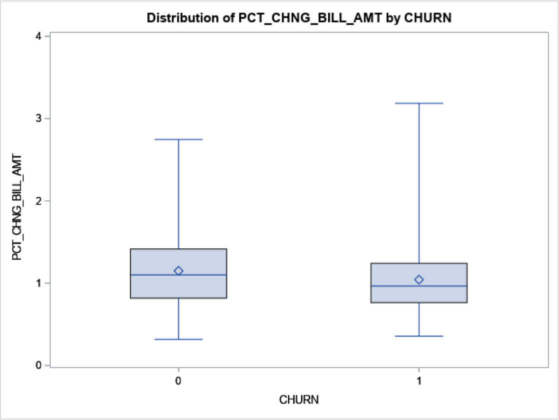
Problem 1 [10 pts] Churn analysis – to be answered by everyone
Given the large number of competitors, cell phone carriers are very interested in analyzing and predicting customer retention and churn. The primary goal of churn analysis is to identify those customers that are most likely to discontinue 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:

a) **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.**

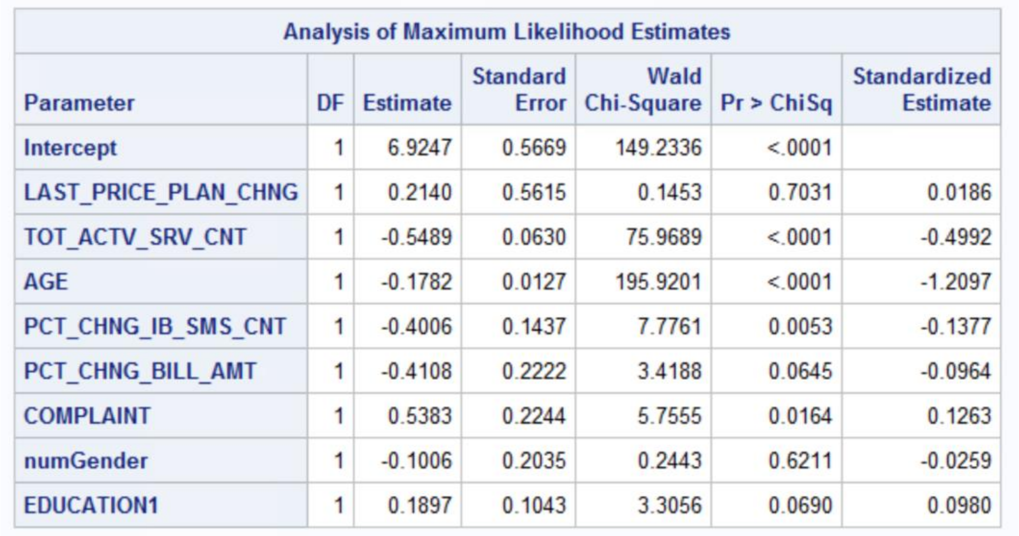




In the first box plot where churn is plotted against age, we notice that people who switched providers are mostly young. The mean and the median are around 25 years old while 75% of the users where churn is 1 are younger than 26 and older than 22. On the other hand, 75% of people who have not switched provider are older than 30. The mean and the median are around 42 years old and the range around the median to the first and third quartile is wider than it is in churn = 1. To illustrate, 50% of people who switched provider are in range of 22 to 26, at the same time, 50% of people who did not switch provider are in range of 31 to 52 approximately.

The change in bill amount does not affect the churn probabilities as much as age does. In the second box plot, we see that the median and mean of the incoming bill where churn = 0 are almost the same to them where churn = 1. They are only slightly higher for people who did not switch provider. The range of the whisker after the third quartile for people who switched is a little higher than it is for people who did not switch provider.

b) 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.



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	6.9247	0.5669	149.2336	<.0001	
LAST_PRICE_PLAN_CHNG	1	0.2140	0.5615	0.1453	0.7031	0.0186
TOT_ACTV_SRV_CNT	1	-0.5489	0.0630	75.9689	<.0001	-0.4992
AGE	1	-0.1782	0.0127	195.9201	<.0001	-1.2097
PCT_CHNG_IB_SMS_CNT	1	-0.4006	0.1437	7.7761	0.0053	-0.1377
PCT_CHNG_BILL_AMT	1	-0.4108	0.2222	3.4188	0.0645	-0.0964
COMPLAINT	1	0.5383	0.2244	5.7555	0.0164	0.1263
numGender	1	-0.1006	0.2035	0.2443	0.6211	-0.0259
EDUCATION1	1	0.1897	0.1043	3.3056	0.0690	0.0980

Since the Gender variable has qualitative values, we should define a dummy variable.

numGender = 1 if Gender = M

numGender = 0 if Gender = F

	Z1 numGender
M	1
F	0

Also, change education missing values to zeros in EDUCATION1 new variable.
Variables (EDUCATION1, LAST_PRICE_PLAN_CHNG_DAY_CNT, PCT_CHNG_BILL_AMT, numGender) have p-values that are higher than 0.05 which make them insignificant X-variables. Thus, we should exclude these previous variables and keep the ones that have p-values less than 0.05 which are (TOT_ACTV_SRV_CNT, AGE, PCT_CHNG_IB_SMS_CNT, COMPLAINT).

The expression of the fitted model:
Log (churn=1 / churn=0) = 6.9247 – 0.5489 * TOT_ACTV_SRV_CNT – 0.1782 * AGE – 0.4006 PCT_CHNG_IB_SMS_CNT + 0.5383 * COMPLAINT
COMPLAINT: 1 if there was at least a customer's complaint in the two months, 0 no complaints

c) Analyze the final logistic regression model and discuss the effect of each variable on the churn probability. Discuss results in terms of odds ratios.
Log (p/(1-p)) = 6.9247 – 0.5489 * TOT_ACTV_SRV_CNT – 0.1782 * AGE – 0.4006 PCT_CHNG_IB_SMS_CNT + 0.5383 * COMPLAINT

Log odds log(p/(1-p)) of churn probabilities decreases by 0.5489 when the total number of active services increases by 1, and decreases by 0.1782 when age increases by 1, and decreases by 0.4006 when the Percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS increases by 1%, and increases by 0.5383 when there is at least a customer's complaint in the last two months and COMPLAINT = 1.

By using the anti-log function, we can determine the change in the Y-variable as follows:

$e^{(-0.5489)}$ = 0.57758, [(0.57758-1)*100] = - 42.24%. The odds of churn p/(1-p) decreases by 42.24% when the total number of active service increases by 1.

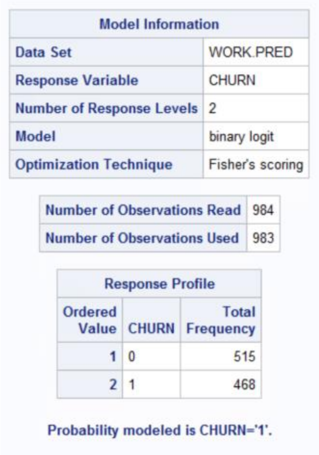
$e^{(-0.1782)}$ = 0.8367, [(0.8367-1)*100] = - 16.32%. The odds of churn p/(1-p) decreases by 16.32% when age increases by 1.

$e^{(-0.4006)}$ = 0.6699, [(0.6699-1)*100] = - 33.008%. The odds of churn p/(1-p) decreases by 33.008% when Percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS increases by 1%.

$e^{(0.5383)}$ = 1.7131, [(1.7131-1)*100] = 71.31%. The odds of churn p/(1-p) increases by 71.31% when there is a complaint in the past two months i.e COMPLAINT = 1.

d) **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.**

Assuming that EDUCATION1 = 0 for missing EDUCATIONS.



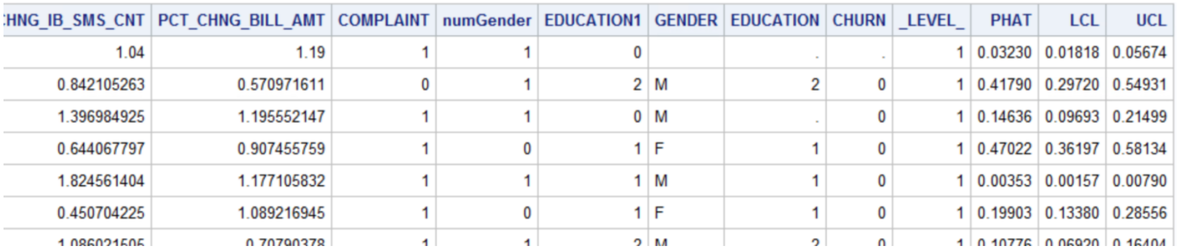
Model Information	
Data Set	WORK.PRED
Response Variable	CHURN
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	984
Number of Observations Used	983

Response Profile

Ordered Value	CHURN	Total Frequency
1	0	515
2	1	468

Probability modeled is CHURN='1'.



HNG_IB_SMS_CNT	PCT_CHNG_BILL_AMT	COMPLAINT	numGender	EDUCATION1	GENDER	EDUCATION	CHURN	_LEVEL_	PHAT	LCL	UCL
1.04	1.19	1	1	0		.	.	1	0.03230	0.01818	0.05674
0.842105263	0.570971611	0	1	2	M	2	0	1	0.41790	0.29720	0.54931
1.396984925	1.195552147	1	1	0	M	.	0	1	0.14636	0.09693	0.21499
0.644067797	0.907455759	1	0	1	F	1	0	1	0.47022	0.36197	0.58134
1.824561404	1.177105832	1	1	1	M	1	0	1	0.00353	0.00157	0.00790
0.450704225	1.089216945	1	0	1	F	1	0	1	0.19903	0.13380	0.28556
1.086021505	0.70790378	1	1	2	M	2	0	1	0.10776	0.06920	0.16404

Predicted churn probability = 0.0323

95% confidential interval is (0.01818, 0.05674) which means that 95% of the time, the predicted probability will fall within 0.01818 and 0.05674.

The corresponding 95% confidence limits for the odd ratio are $e^{(0.01818)}$-1 = 1.83%, $e^{(0.05674)}$-1 = 5.838%. That means the odds of having customer switch provider when the customer is a 43 years old male with the above description will increase between 1.83% to 5.838%.

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

```
*a;
proc import datafile="S:\HW6\churn_train.csv" out=myd replace;
delimiter=',';
getnames=yes;
run;

Proc print data=myd (obs= 20);
run;

*Boxplot - by AGE;
proc sort;
by CHURN;
RUN;

PROC BOXPLOT;
PLOT AGE*CHURN;
RUN;

*Boxplot - by PCT_CHNG_BILL_AMT*CHURN;
PROC BOXPLOT;
PLOT PCT_CHNG_BILL_AMT*CHURN;
RUN;

*Create dummy variable for Gender;
data churn;
set myd;
numGender = 1;
if GENDER = 'F' then numGender = 0;
EDUCATION1 = EDUCATION;
If EDUCATION = '.' then EDUCATION1 = 0;
run;

Proc print data=churn (obs= 20);
run;

*b;
*fit full logistic model;
PROC LOGISTIC data=churn;
MODEL CHURN (EVENT = '1') = EDUCATION1 LAST_PRICE_PLAN_CHNG_DAY_CNT
TOT_ACTV_SRV_CNT AGE PCT_CHNG_IB_SMS_CNT PCT_CHNG_BILL_AMT COMPLAINT
numGender/STB;
RUN;

*d;
*(a) Create prediction dataset;
DATA NEW;
INPUT LAST_PRICE_PLAN_CHNG_DAY_CNT TOT_ACTV_SRV_CNT AGE
PCT_CHNG_IB_SMS_CNT PCT_CHNG_BILL_AMT COMPLAINT numGender EDUCATION1;
DATALINES;
0 4 43 1.04 1.19 1 1 0
PROC PRINT;
RUN;

*(b) Merge prediction dataset with original dataset;
DATA PRED;
SET NEW churn;
```

```
RUN;
PROC PRINT;
RUN;

*(c) Run prediction;
PROC LOGISTIC data=PRED;
MODEL CHURN (EVENT = '1') = LAST_PRICE_PLAN_CHNG_DAY_CNT TOT_ACTV_SRV_CNT
AGE PCT_CHNG_IB_SMS_CNT PCT_CHNG_BILL_AMT COMPLAINT numGender EDUCATION1;
OUTPUT OUT=PRED P=PHAT LOWER=LCL UPPER=UCL;
RUN;

*(d) Print predicted probabilities and confidence intervals;

PROC PRINT data=PRED;
RUN;
```