Seedbox Data Science Application Test

David Gutkovsky

2018-08-27

For every question, I will give a clear explanation of the code and results. RStudio and Rmarkdown are used to complete this test.

Importing data

The first step is to import the data from the CSV files to Rstudio. Also, I will merge both datasets using the foreign key *sample_id* in **transData**.

```
testSamples=read.csv("testSamples.csv")
transData=read.csv("transData.csv")
mydata=merge(testSamples,transData,by.x="sample_id",by.y="sample_id")
attach(mydata)
mydata<-mydata[order(transaction_id),]
```

Question 1

What is the aproximate probability distribution between the test group and the control group

Given that users will randomly assigned to a group, an adequat approximation of the probability distribution would be the binomial distribution. Let TG_i be defined as the variable indicating if the user i is placed into the control group ($TG_i = 0$) or in the test group ($TG_i = 1$). Therefore $TG \sim \text{Binomial}(n, p)$ where n is the number of observations and p is the rate of success.

Let $\hat{p}$ be an estimator of the rate of success p. The maximum likelihood estimator is $\hat{p} = \frac{\sum_{i=1}^{n} TG_i}{n}$. The following code performs this task:

```
nobs=nrow(testSamples)
ntest=sum(test_group)
freq=ntest/nobs
```

We have that $\hat{p} = 0.0565965$. In sum, users had approximately a 5% of being assigned to the test group.

Question 2

Is a user that must call-in to cancel more likely to generate at least 1 addition REBILL?

Two techniques are used to answer this question. First, I calculate the frequency of rebilling of each group and see if the difference is statistical significant. Second, a logit model is estimated to verify is the coefficient related to group assignation is statistical significant.

Difference between two proportions

Let f_c be the frequency of transactions of users in the control group that were rebilling and f_t be the same for the test group. Our statistical test is the following:

$$H_0 : f_c - f_t = 0$$
$$H_1 : f_c - f_t > 0$$

In other words, the null hypothesis is the proportion of transactions that were rebilling made by users in the control group is equal to the proportion in the test group.

In this part of the code, I calculate $\hat{f}_c$ and $\hat{f}_t$ which are estimators of the population value. Here, a dummy variable is created if the transaction type is a rebill (dum_rebill=1)

```
library(plyr)
count_test<-count(mydata,"mydata$test_group")
nc=count_test[1,2]
nt=count_test[2,2]

mydata$dum_rebill<-0
mydata$dum_rebill[mydata$transaction_type=="REBILL"]<-1

mydata_control<-subset(mydata,test_group==0)
Rc=sum(mydata_control$dum_rebill)
fc=Rc/nc

mydata_test<-subset(mydata,test_group==1)
Rt=sum(mydata_test$dum_rebill)
ft=Rt/nt
```

The frequency of rebilling in the control group is *0.93* and *0.95* in the test group. The next question we must ask is if these proportions are statistically different. To do so, we must calculate a variance to create a statistic. Under the null hypothesis, $f_t = f_c = f$. Therefore, it is possible to show that the variance of the difference is given by

$$V(\hat{f}_t - \hat{f}_c) = f(1-f)\left(\frac{1}{n_t} + \frac{1}{n_c}\right)$$

where n_j is the number of observations in group j. The following code chunk calculates the variance.

```
#variance under the null hypothesis
f=(Rt+Rc)/(nc+nt)
var_f=f*(1-f)*(1/nt+1/nc)
```

The last step is to calculate the statistic $\frac{\hat{f}_t - \hat{f}_c}{\sqrt{\hat{f}(1-\hat{f})\left(\frac{1}{n_t}+\frac{1}{n_c}\right)}}$ and then compared to critical value on a standard normal distribution. I chose a significance level of 5% and the critical value is 1.645 (one-sided test).

```
#porprotion difference testing
z=(ft-fc)/sqrt(var_f)
```

Here, our statistic z=*3.7*>1.645. Therefore, we conclude that a user that must call-in to cancel is more likely to generate at least 1 addition rebill.

A logit model can be estimated. Let *Rbill* be a dummy variable equal to 1 if the transaction type is *Rebill* or 0 otherwise. Then, using the logistic function,

$$p(Rbill = 1) = \frac{e^{\beta_0 + \beta_1 \text{Group}}}{1 + e^{\beta_0 + \beta_1 \text{Group}}}$$

where Group is also a dummy variable equal to 1 if the user is assigned to the test group or 0 if assigned in the control group. By using maximum likelihood, we can estimate β_0 and β_1

```
#Logit
library(ISLR)
glm.fit=glm(dum_rebill~test_group,data=mydata,family=binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = dum_rebill ~ test_group, family = binomial, data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4335   0.3261   0.3882   0.3882   0.3882
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.54753    0.06056  42.066  < 2e-16 ***
## test_group   0.36015    0.09846   3.658 0.000254 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3499.1  on 7429  degrees of freedom
## Residual deviance: 3485.5  on 7428  degrees of freedom
## AIC: 3489.5
##
## Number of Fisher Scoring iterations: 5
```

According to our results, β_1 is positif and statistically different from zero (p-value under 5%). This concords with the results obtained with the previous test.

Question 3

Is a user that must call-in to cancel more likely to generate more revenues?

Just like for previous question, I perform two tests. First, I create a statistic to see if the revenue per transaction in the test group is larger than in the control group. Second, I estimate a linear regression.

Since the number of transactions are not the same in each group, revenue per transaction will be used. Here, I consider every type of transaction, even the ones with a negative value.

The first test can be summarized by

$$H_0 : Rev_t - Rev_c = 0$$
$$H_1 : Rev_t - Rev_c > 0$$

where Rev_j is population mean of the revenue per transaction in group j = Control, Test.

Let $\bar{Rev}_j$ be the sample average for group j. It is possible to show that these sample averages are unbiased estimator the population mean i.e. $\mathbb{E}(\bar{Rev}_j) = Rev_j$. Therefore, $\mathbb{E}(\bar{Rev}_t - \bar{Rev}_c) = Rev_t - Rev_c$. The following code calculates the mean.

```
revc=mean(mydata_control$transaction_amount)
revt=mean(mydata_test$transaction_amount)
```

The variance of the difference is given by $V(Rev_t - Rev_c) = \frac{\sigma_t^2}{n_t} + \frac{\sigma_c^2}{n_c}$, where σ_j^2 is the population variance for group j. The variance of the sample for each group j, S_j^2 is an unbiased estimator for the population variance. The following code calculated the standard deviation for each group.

```
#standard deviation
sigmac=sd(mydata_control$transaction_amount)
sigmat=sd(mydata_control$transaction_amount)
```

The last step is to construct the statistic and compare it to a critical value from a standard normal distribution. The statistic is given by

$$z_{rev} = \frac{(\bar{Rev}_t - \bar{Rev}_c)}{\sqrt{\frac{S_t^2}{n_t} + \frac{S_c^2}{n_c}}}$$

```
sd_diff=sqrt(sigmac^2/nc+sigmat^2/nt)
z_rev=(revt-revc)/sd_diff
```

We obtain that *18*>1.645, suggesting that users that must call-in to cancel are more likely to generate more revenues.

Our second test is to estimate a linear regression (OLS). Let Rev_i be the transaction amount for user i. Then, we can estimate the following equation

$$Rev_i = \beta_0 + \beta_1 \text{Group} + e_i$$

where e_i are the residuals. The following code estimates the last equation.

```
lm.fit=lm(transaction_amount~test_group,data=mydata)
summary(lm.fit)
```

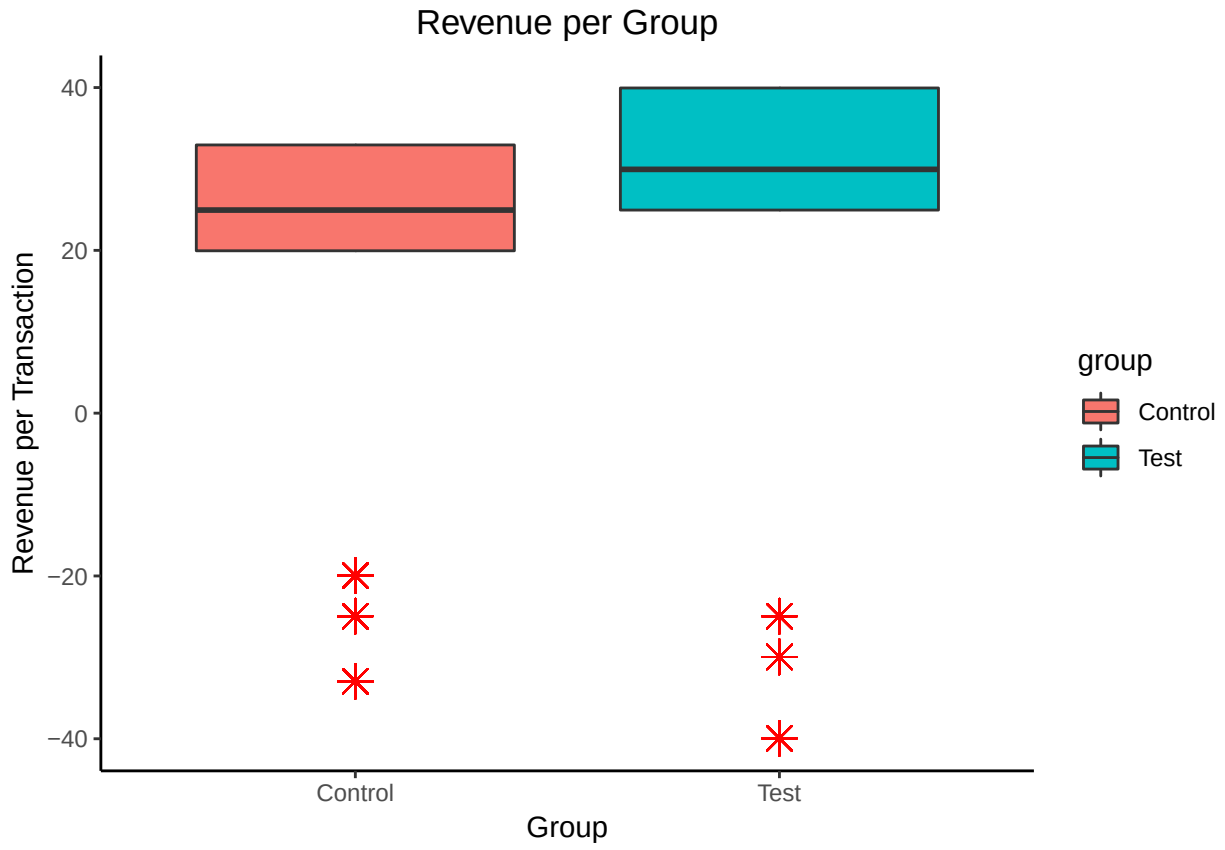
```
##
## Call:
## lm(formula = transaction_amount ~ test_group, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.185  -2.232   1.715  10.768  11.715
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.1824     0.2317   95.73   <2e-16 ***
## test_group    6.0523     0.3436   17.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.75 on 7428 degrees of freedom
## Multiple R-squared:  0.0401, Adjusted R-squared:  0.03997
```

```
## F-statistic: 310.3 on 1 and 7428 DF,  p-value: < 2.2e-16
```

According to our results, β_1 is positive and statistical significant (p-value<0.05). This concords with the result of the previous test.

Using a boxplot (ggplot2 package), we can have a clear visualization of our results.

```
#Visualization
library(ggplot2)
mydata$group<-""
mydata$group[test_group==0]="Control"
mydata$group[test_group==1]="Test"
ggplot(mydata,aes(x=mydata$group,y=mydata$transaction_amount,fill=group))+
  geom_boxplot(outlier.colour = "red",outlier.shape=8,outlier.size=4)+
  labs(x="Group",y="Revenue per Transaction",title="Revenue per Group")+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```



The red stars are outiliers. We can see that the test group revenue average per transaction is higher than the one from the control group.

Question 4

Is a user that must call-in more likely to produce a higher chargeback rate(CHARGEBACKs/REBILLs)?

I will use a different definition of a chargeback rate. Let Charge_i be a dummy variable that is equal to 1 if transaction i is a chargeback and 0 otherwise. Then for each group j, I calculate the following chargeback rate:

$$\text{chargerate}_j = \frac{\sum_{i=1}^{n_j} \text{Charge}_i}{n_j}$$

This chargeback rate has the same significance as the other mentioned in the question and easier to build statistical tests around it.

The methodology to answer the question will be same as the first test in question 2:

$$H_0 : \text{chargerate}_t - \text{chargerate}_c = 0$$
$$H_1 : \text{chargerate}_t - \text{chargerate}_c > 0$$

In the following code chunk, I calculate the chargeback rate for each group, the variance of the unbiased estimator under the null hypothesis and the statistic.

```
mydata_control$dum_chargeback<-0
mydata_control$dum_chargeback[mydata_control$transaction_type=="CHARGEBACK"]<-1
sum_chargeback_c=sum(mydata_control$dum_chargeback)
chargeback_rate_control=sum_chargeback_c/nc

mydata_test$dum_chargeback<-0
mydata_test$dum_chargeback[mydata_test$transaction_type=="CHARGEBACK"]<-1
sum_chargeback_t=sum(mydata_test$dum_chargeback)
chargeback_rate_test=sum_chargeback_t/nt

p=(sum_chargeback_t+sum_chargeback_c)/(nc+nt)
var_p=p*(1-p)*(1/nc+1/nt)
z_charge=(chargeback_rate_control+chargeback_rate_test)/sqrt(var_p)
```

The statistic $z_{\text{charge}} = 13 > 1.645$ which indicates that the test group (user must call-in to cancel) has statistically higher chargeback rate.