BankHWfile-Summary

RyanYip

2018骞11鏈<88>15鏃

## There is :%d NA value 0

## age education contact month   
## Min. :20.00 primary : 63 cellular :536 may :185   
## 1st Qu.:33.00 secondary:303 telephone: 28 apr : 84   
## Median :38.00 tertiary :198 0v : 75   
## Mean :41.05 feb : 56   
## 3rd Qu.:48.00 jan : 36   
## Max. :80.00 aug : 34   
## (Other): 94   
## campaign poutcome y   
## Min. : 1.000 failure:452 Min. :0.000   
## 1st Qu.: 1.000 success:112 1st Qu.:0.000   
## Median : 1.000 Median :0.000   
## Mean : 1.927 Mean :0.234   
## 3rd Qu.: 2.000 3rd Qu.:0.000   
## Max. :11.000 Max. :1.000   
##

### Question 1

**Hypotheses 1: variable education has significant effect on output variable y,that is**

* H0:the coefficient of education is equal to 0 in logistic model
* H1: the coefficient of education is not equal to 0 in logistic model

**Hypotheses 2: variable age has significant effect on output variable y,that is**

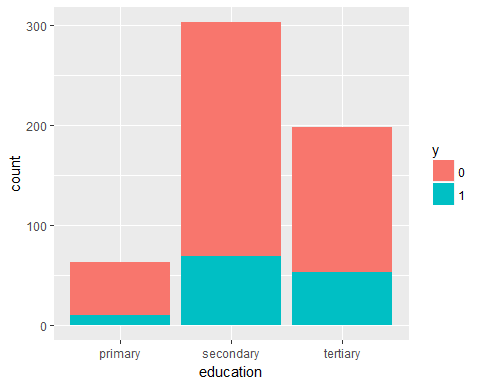
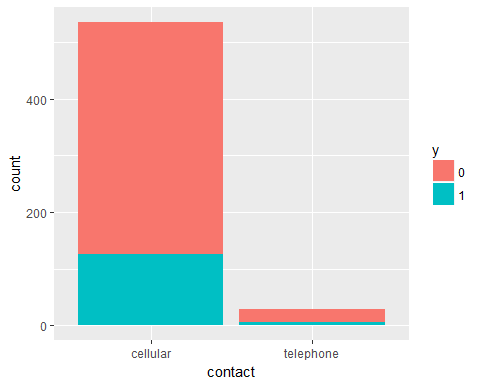
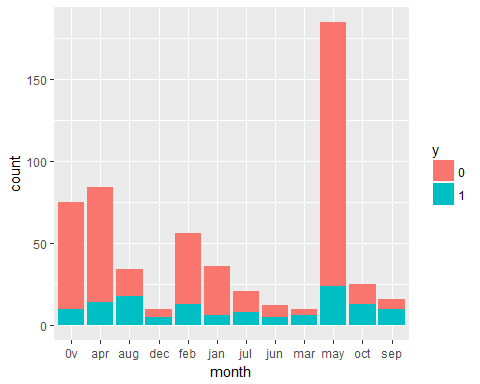
* H0:the coefficient of age is equal to 0 in logistic model
* H1: the coefficient of age is not equal to 0 in logistic model

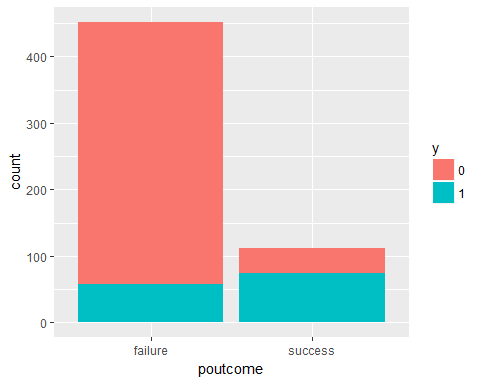
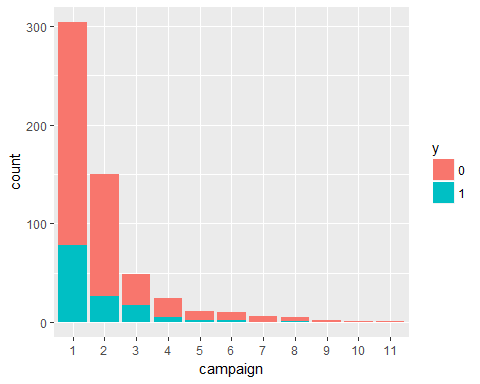
I conducted the hypotheses test of age and education because, according to common sense, these two variables have a relatively large relationship with the dependent variable, so I am interested and want to see if there is any influence.And in reality,It is stange that *education* and *age* should have effect on whether has the client subscribed a term deposit.o I suggest that the company can pay more attention to age and education level when considering customers,they mat be important message and need to be consider further.

### Question 2

#### preliminary analysis

## Loading required package: ggplot2





We make preliminary analysis on data by simple graph analysis,due to these variables are most categorical,so only do barplot.According to these plot,in *contact* variable,the numeber in celluar type is the most,but number of not having the client subscribed a term deposit also has big ratio;in *education*,the count in secondary is most. ;and in telephone has the same result;In *month*,I think responce(y) has a differnce across month,and last contact in **May** is the highest;And in *campain*, the most is in low level number,but in high number,it is almost that client has not subscribed a term deposit.In *poutcome*,it shows that when poutcome is fail,the more number of client has not subscribed a term deposit.

### Question 3

Corresponding to question 1,*age* and *education* are the two variables I want to test,So I control respectively.And I use all variables as predictor in full model.

The follow model I do not consider the relationship across predictors.

model.full<-glm(y~age+education+contact+month+campaign+poutcome,data=BankHWWfile,family = "binomial") # with age,edu  
  
model.reduce.edu<-glm(y~age+contact+month+campaign+poutcome,data=BankHWWfile,family = "binomial") # without edu  
model.reduce.age<-glm(y~education+contact+month+campaign+poutcome,data = BankHWWfile,family = "binomial") #without age  
#summary(model.full)

### Question 4

Interpret the output of model in Question 3

## Analysis of Deviance Table  
##   
## Model 1: y ~ age + education + contact + month + campaign + poutcome  
## Model 2: y ~ age + contact + month + campaign + poutcome  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 537 448.88   
## 2 539 453.76 -2 -4.8782 0.08724 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Accoring to the anova tables,we know that the p-value is larger than 0.05,that is,the **Hypotheses 1** is not significant,so we can not reject the H0,that is,variable *education* should not be in model,which means *education* has few effect on response variable *y*.

## Analysis of Deviance Table  
##   
## Model 1: y ~ age + education + contact + month + campaign + poutcome  
## Model 2: y ~ education + contact + month + campaign + poutcome  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 537 448.88   
## 2 538 449.21 -1 -0.33103 0.5651

The result of age is similiar with education,the p-value is also larger than 0.05,so it is not significant and variable *age* should not be in model,which means *age* has few effect on response variable *y*.

According to the above result,they do not support my Hypotheses in statistics,which is not consistent with common sense?,may be I should take them into further consideration and consider other factors.