Subject: k fold cross validation for logistic regression in R
Posted by dhivvyajp@am.amrita.edu on Thu, 16 Feb 2023 08:53:34 GMT

Dear Experts,

I am working on DHS7 dataset. I was able to do logistic regression for 70% training and 30% testing data. But When I am trying to do k fold cross validation instead of 70: 30 split up, came across surveyCV package. I am getting the following error. Kindly let me know how can I fix this issue.

```
> set.seed(2023)
> svylogistic <- svyglm(formula =
InternetUsage~RuralOrUrban+AgeGroup+WealthIndex+SchoolingCom
pleted+Religion+Caste+MaritalStatus+Occupation+Gender+Litera cy+OwnsMobile,
design=my_design, family=quasibinomial())
> cv.svyglm(svylogistic, nfolds=3, na.rm = FALSE)
Error in if (clusterID %in% c("0", "1")) { : the condition has length > 1
```

Subject: Re: k fold cross validation for logistic regression in R
Posted by Bridgette-DHS on Thu, 16 Feb 2023 13:25:07 GMT

Following is a response from Senior DHS staff member, Tom Pullum:

If you were using Stata, I would want to be sure that your svyset command includes "singleunit(centered)" (or one of the other options in the parentheses). The India surveys are huge, but it is still possible that singleunit is needed when you are dividing the sample up. I hope R has something equivalent to singleunit (I do not use R myself). Beyond that, I have no suggestions. Other users may be able to help.

Subject: Re: k fold cross validation for logistic regression in R
Posted by dhivvyajp@am.amrita.edu on Fri, 17 Mar 2023 07:45:48 GMT

I tried the following also. But not able to fix the error.

```
> cv.svy(train, formulae = "
InternetUsage~RuralOrUrban+AgeGroup+WealthIndex+SchoolingCom
pleted+Religion+Caste+MaritalStatus+Occupation+Gender+Litera cy+OwnsMobile ", method =
"logistic", nfolds=3, strataID = train$strata, clusterID = train$Cluster, nest = T, weightsID =
```

train$samplewt)

.....................................Error in .subset2(x, i, exact = exact) : no such index at level 1[/color]

> cv.svy(train, formulae = " InternetUsage~RuralOrUrban+AgeGroup+WealthIndex+SchoolingCom pleted+Religion+Caste+MaritalStatus+Occupation+Gender+Litera cy+OwnsMobile ", method = "logistic", nfolds=3, strataID = train$strata, clusterID = train$Cluster, nest = F, weightsID = train$samplewt)

.....................................Error in .subset2(x, i, exact = exact) : no such index at level 1

> cv.svyglm(svylogistic, nfolds=3,na.rm=FALSE)

.....................................Error in if (clusterID %in% c("0", "1")) { : the condition has length > 1

Can anyone help me in fixing this error?

Subject: Re: k fold cross validation for logistic regression in R
Posted by Bridgette-DHS on Mon, 27 Mar 2023 12:05:00 GMT
View Forum Message <> Reply to Message

Following is a response from DHS Senior Analysis & Research Manager, Shireen Assaf:


```
# install and load the packages you need
install.packages("survey")
library(survey)

# setting your survey design
# To identify the survey design, you need three variables: weight, psu, and strata

# creating the sampling weight variable.
IRdata$wt <- IRdata$v005/1000000

mysurvey<-svydesign(id=IRdata$v021, data=IRdata, strata=IRdata$v022,  weight=IRdata$wt,
nest=T)
options(survey.lonely.psu="adjust")
```

#now you can use the svy commands in the survey package and use the "mysurvey" sample design object. Check the commands you can use in the survey package.

for example

table of variable v313 (FP use) this is after you attach your data. You can also use svyby, svyglm, etc.

```
svytable(~v313, mysurvey)
```