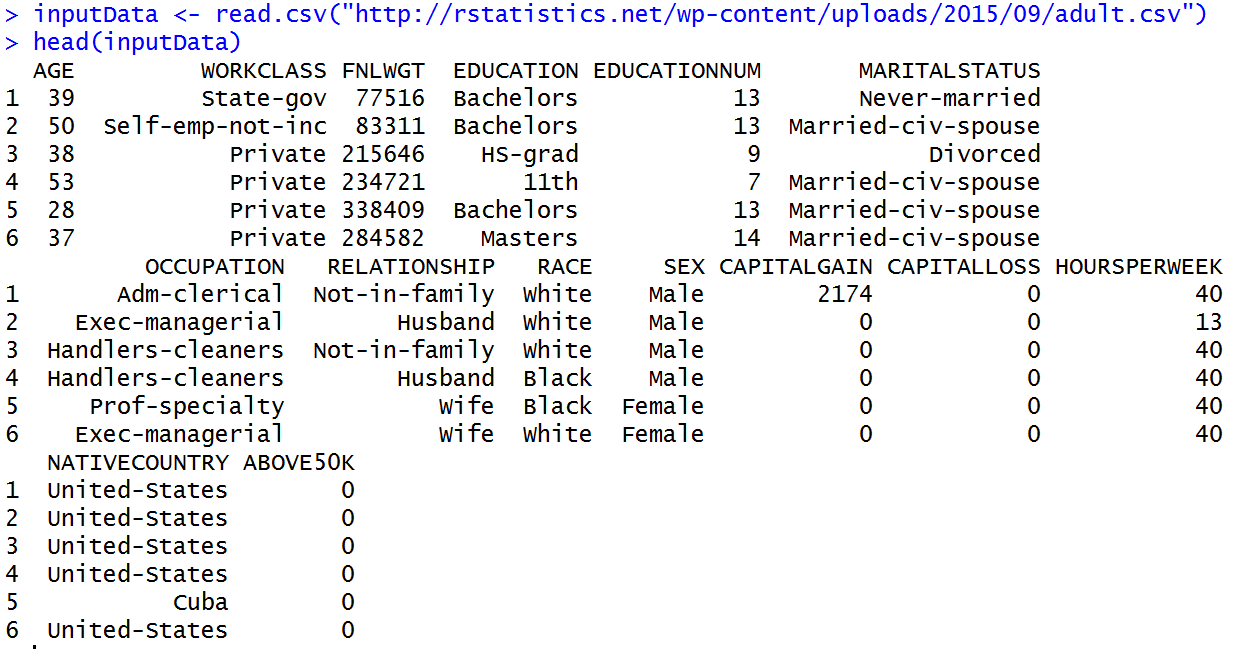
**Module 6: Additional Exercises with Answers**

Predicting Annual Income of Individuals

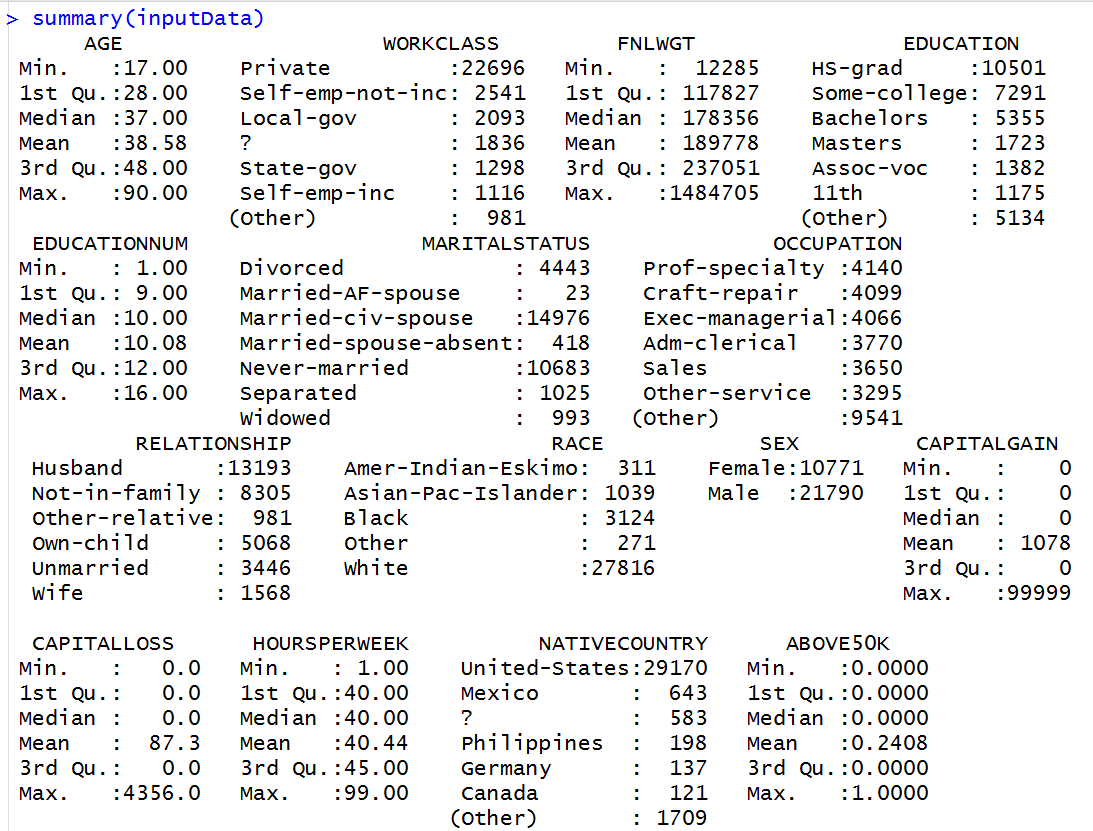
In this example, we will use logistic regression to predict if an individual earns more than $50k in a year or not. First read the ‘Adult’ CSV file using the following command (you need to be connected to the internet and it will take few moments for data to be loaded):

**inputData <- read.csv("http://rstatistics.net/wp-content/uploads/2015/09/adult.csv")**

Let’s have a look at the first 6 records and also a summary of the dataset



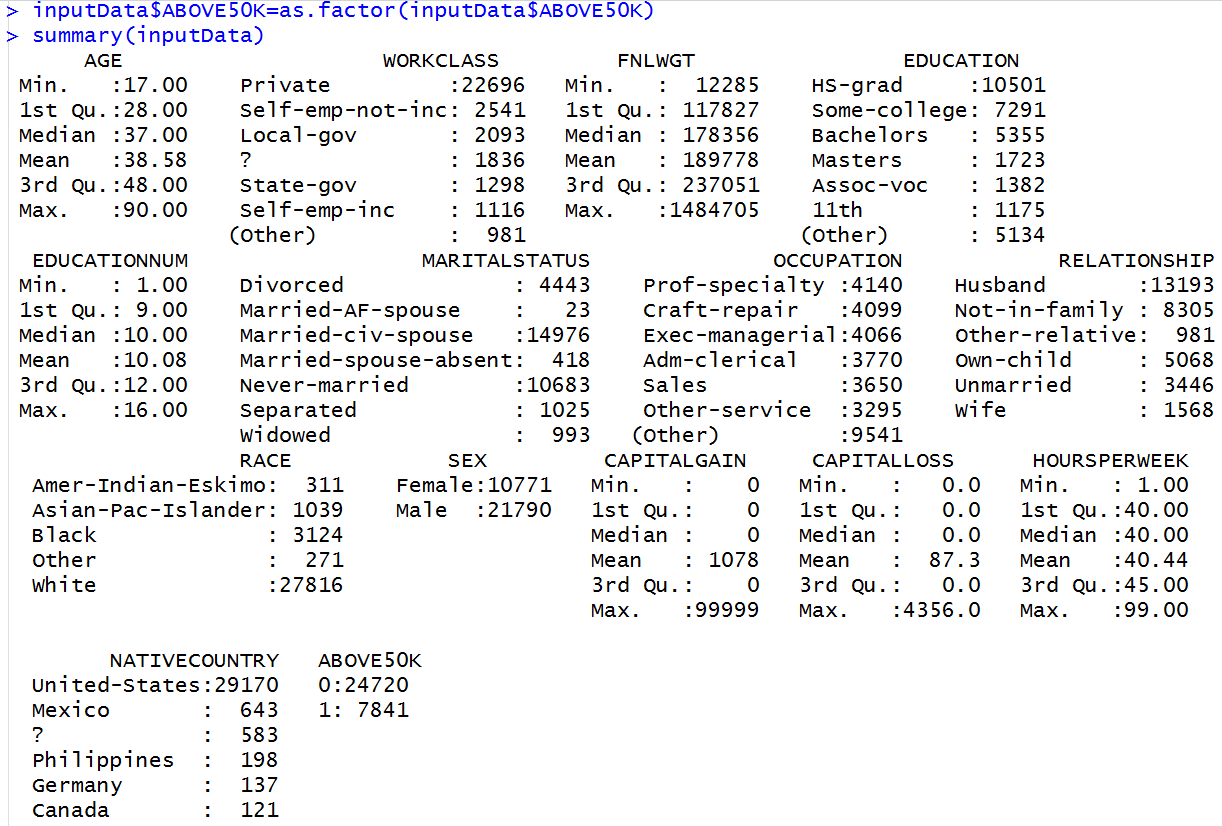
And



The variable ‘ABOVE50K’ is the variable that we are trying to predict. Currently, the variable is coded as a numeric variable that takes 0 and 1s. To use logistic regression, that is a classification method, we need to convert this variable to a factor (i.e. categorical variable):

inputData$ABOVE50K=as.factor(inputData$ABOVE50K)

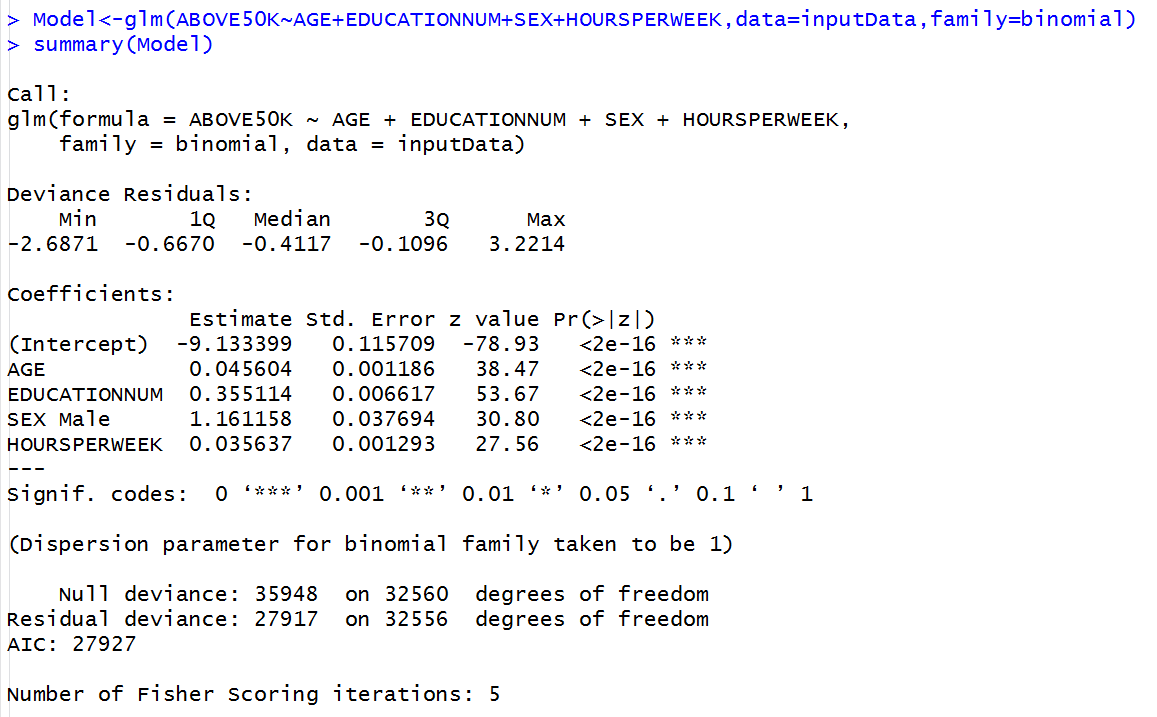
let’s look at the summary again:



**Questions**

**Q1. Build a logistic regression model based on AGE, EDUCATIONNUM (Number of years of Education), SEX and HOURSPERWEEK (number of hours worked per week) to predict ABOVE50K.**

Model<-glm(ABOVE50K~AGE+EDUCATIONNUM+SEX+HOURSPERWEEK,data=inputData,family = binomial)



**Q2. Which of these variable are statistically significant?**

The z-value is very high resulting in very small p-values for coefficients of variables, implying that they are all statistically significant.

**Q3. How does the probability of earning above $50k changes with these variables?**

The coefficient is positive for all variables implying that the probability of earnings being above the $50k increases with all variables. As for SEX, the default value is Female (alphabetically before Male) which is used for the base model. If the SEX is male there will be an additional 1.16 added to the output (that is the logarithm of the odds of earning more than $50k)

**Q4. James and Hannah are two individuals. Given the information below, what is the probability that each of them is earning more than $50k a year?**

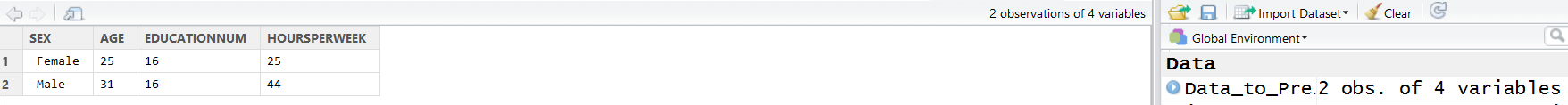
**Hannah: Female, 25 Year, 16 Years of Education, Working part time 25 hours a week   
James: Male, 31 Year, 16 Years of Education, Working 44 hours a week**

We create a new dataset, first record for Hannah and second record for James:

Data\_to\_Predict=data.frame(SEX=c(' Female' ,' Male'), AGE=c(25,31),

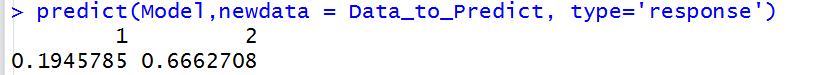
EDUCATIONNUM=c(16,16), HOURSPERWEEK=c(25,44))

Now you should have a new dataframe called Data\_to\_Predict in your global environment area. Double click on it to see the content of it.



Let’s try to predict now:

predict(Model,newdata = Data\_to\_Predict, type='response')



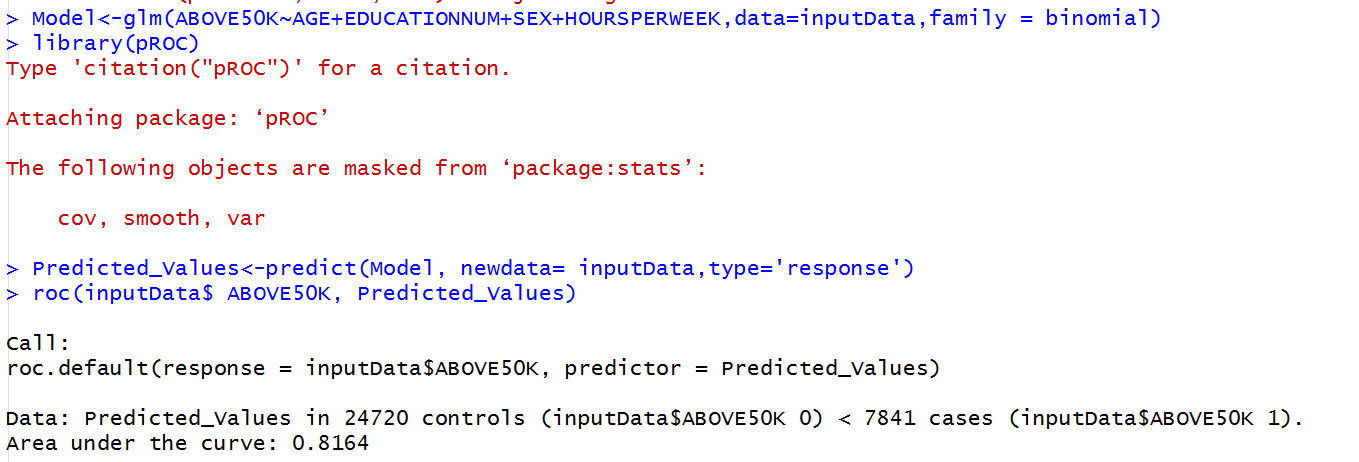
So the probability of Hannah (the first record) to earn more than $50k is 19% (perhaps her mother, Elizabeth, is now worried again!) and the probability that James earns more than $50k is 66%.

**Q5. What is the accuracy of this model in terms of Area Under Curve (AUC) of ROC ?**

library(pROC)  
Predicted\_Values<-predict(Model, newdata= inputData,type='response')   
roc(inputData$ ABOVE50K, Predicted\_Values)

If you are getting error by calling the pROC library, you need to install the package first that is :

install.packages('pROC')



AUC is 0.81. The model is pretty accurate!