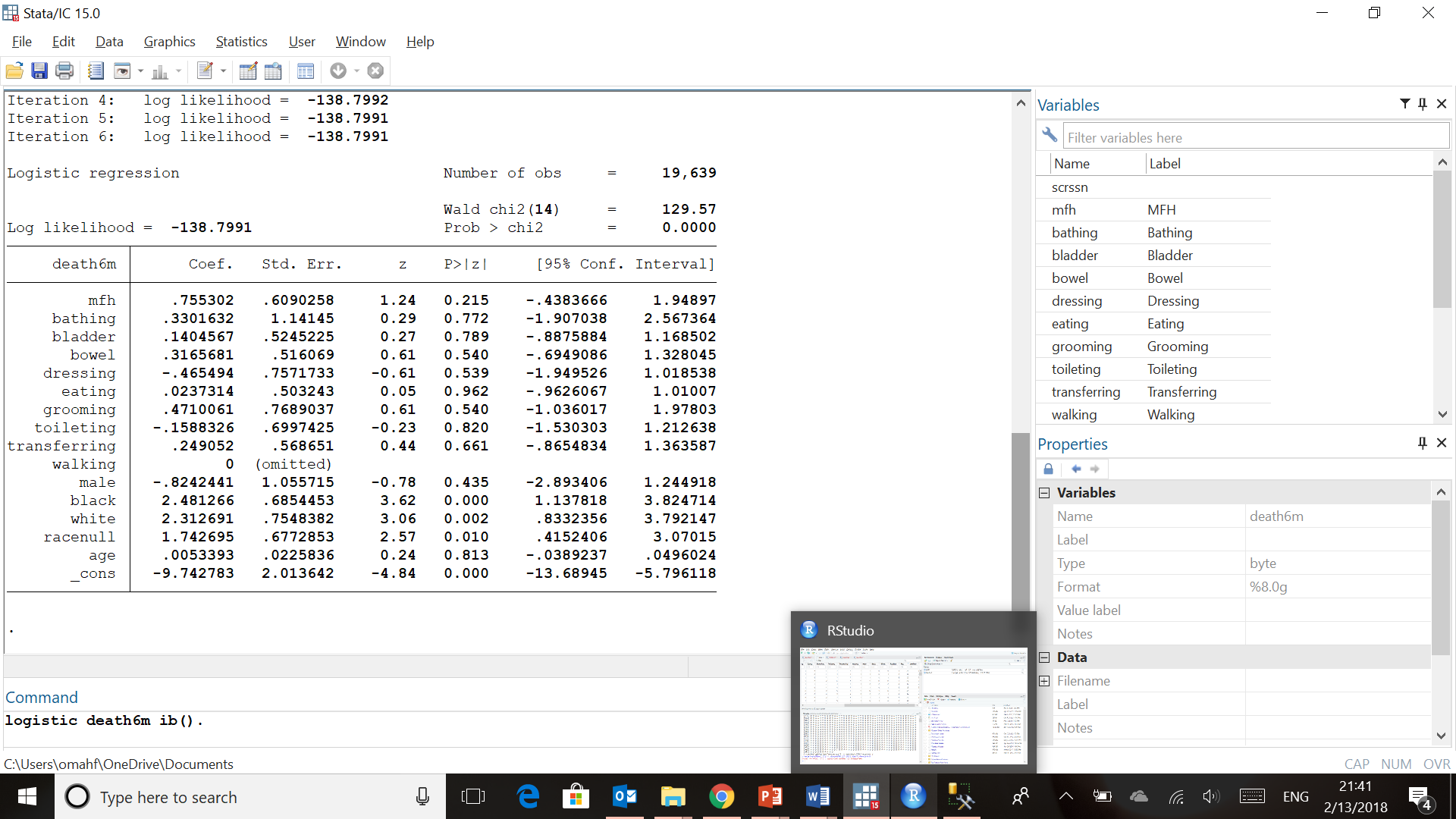
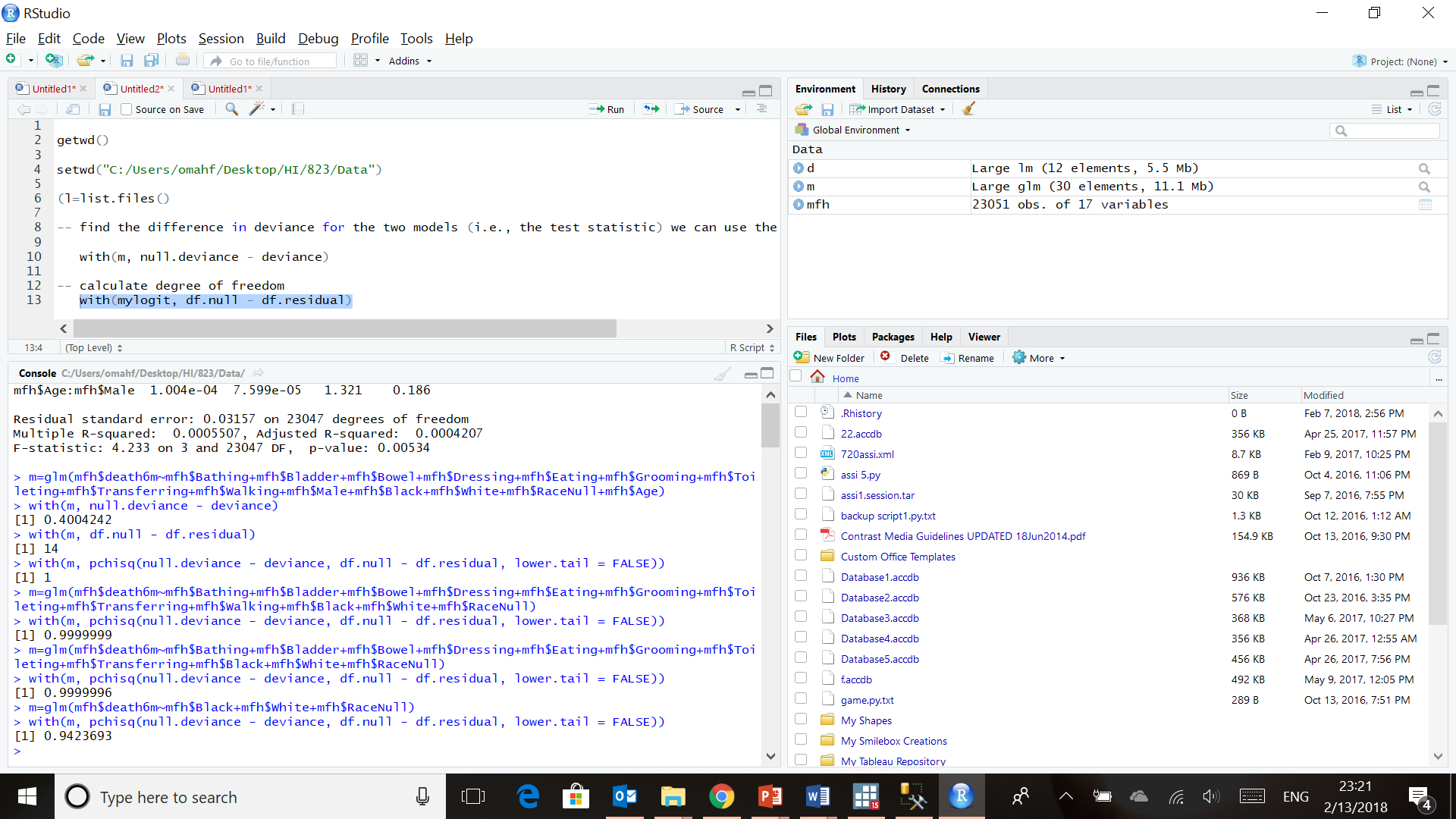
**Question 3:** Create a regression model to explain the relationship among the variables and survival.  

From our logistic regression result we can see that all race variables appear to statistically significant, which means that race is the only predictor. However, our model is not significant .

1. Use plots of residuals to test regression assumptions

Since we are predicting binary variable we used logistic regression which does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables.

1. Explain the fit of the model to the data.



From the result above we can see that the chi-square of 0.4with 14 degrees of freedom and an associated p-value of 1 tells us that our model does not fits significantly better than an empty model.

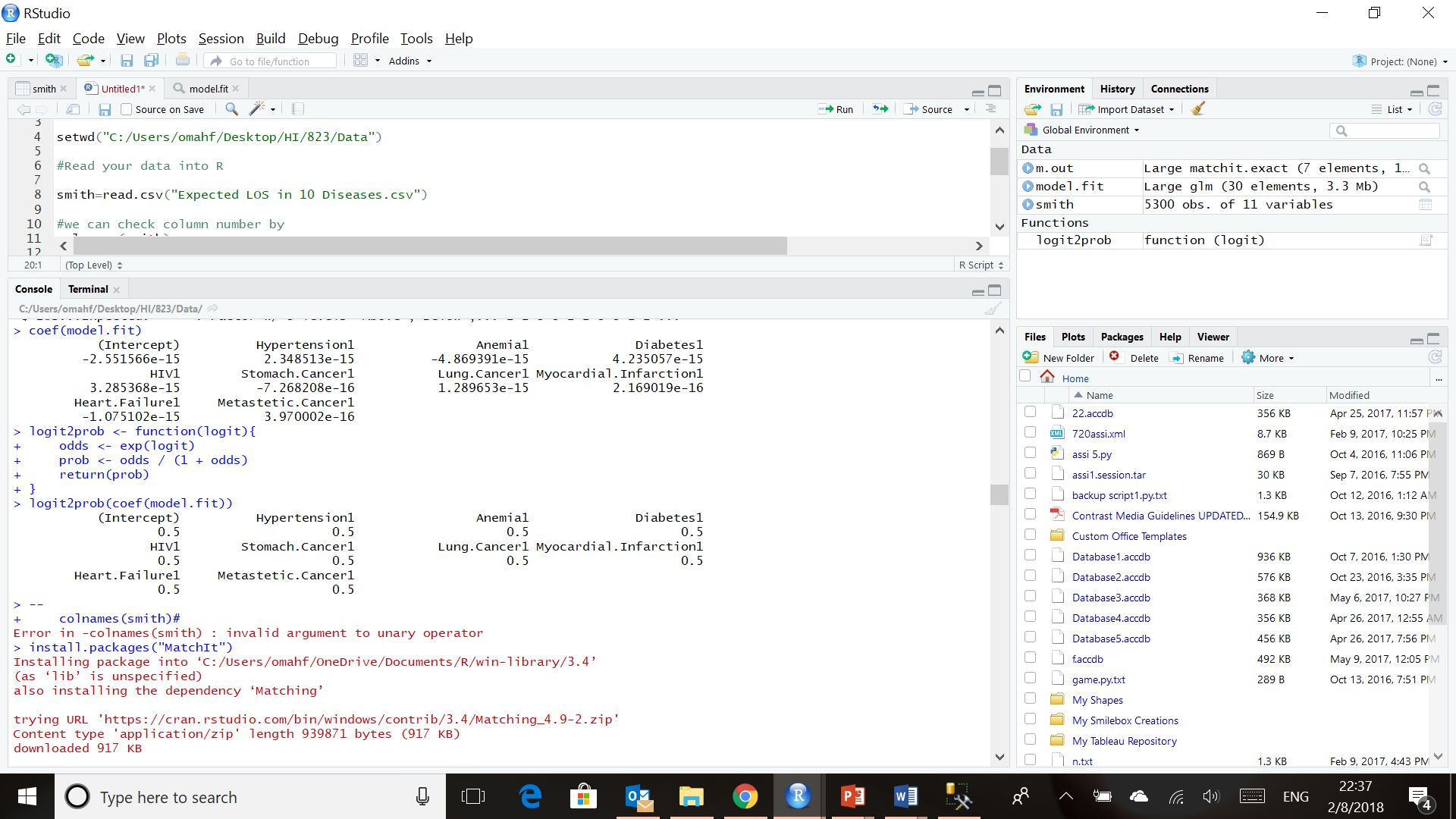
1. List the top 4 predictors of survival (list these predictors using English language and not coded data).

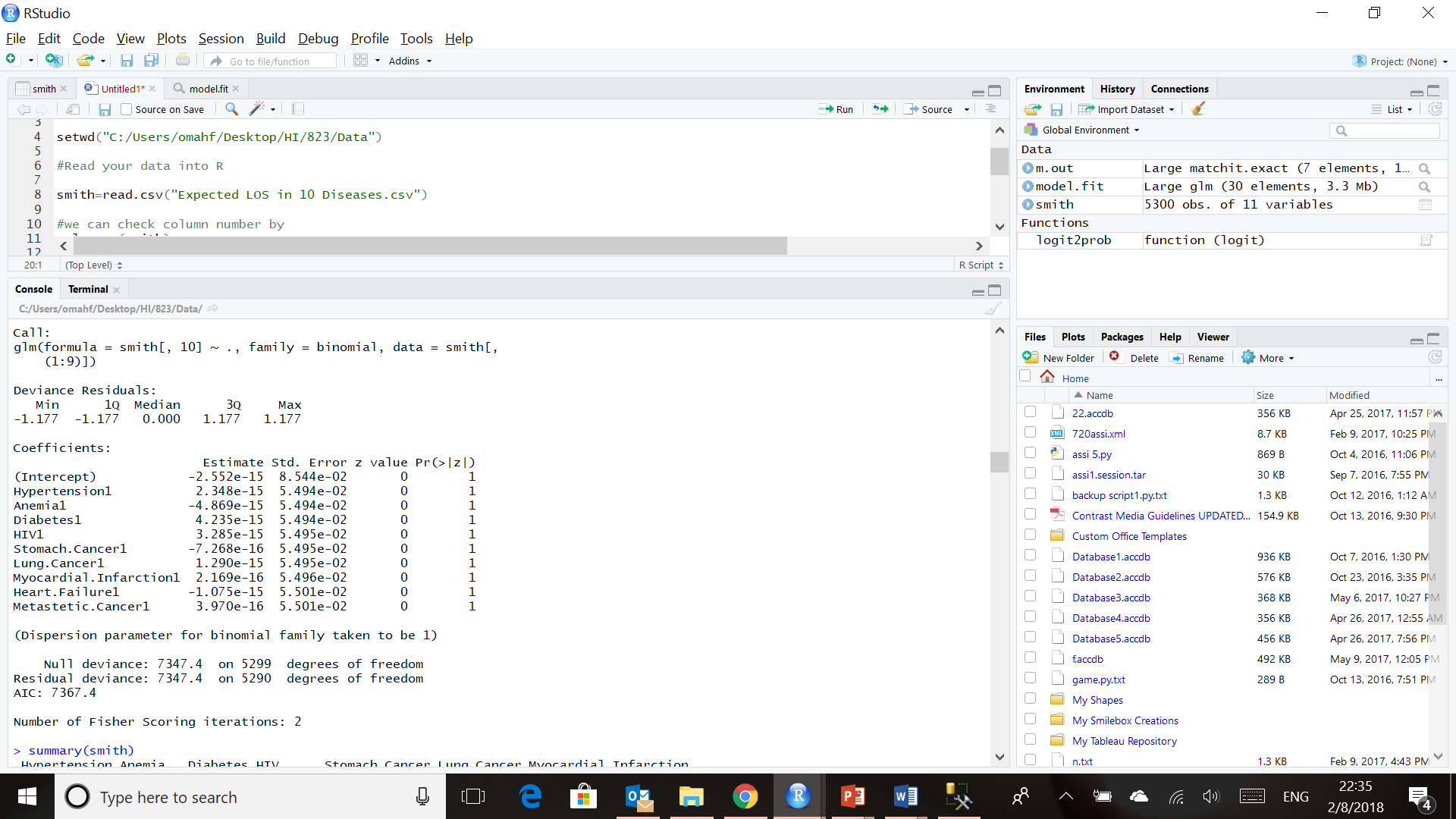
The only significant predictor to survival was race in general. All the three race groups were statistically significant predictors to survival.

1. Describe, in English, if the MFH program contributes to survival.  Provide the evidence for your claim.

from our logistic regression result , we can conclude that MFH does not contribute in prediction wither the patient will survive or not.

Question 3: The following data provide the length of stay of patients seen by Dr. Smith (Variable Dr Smith=1) and his peer group (variable Dr. Smith = 0).  Answer following questions:

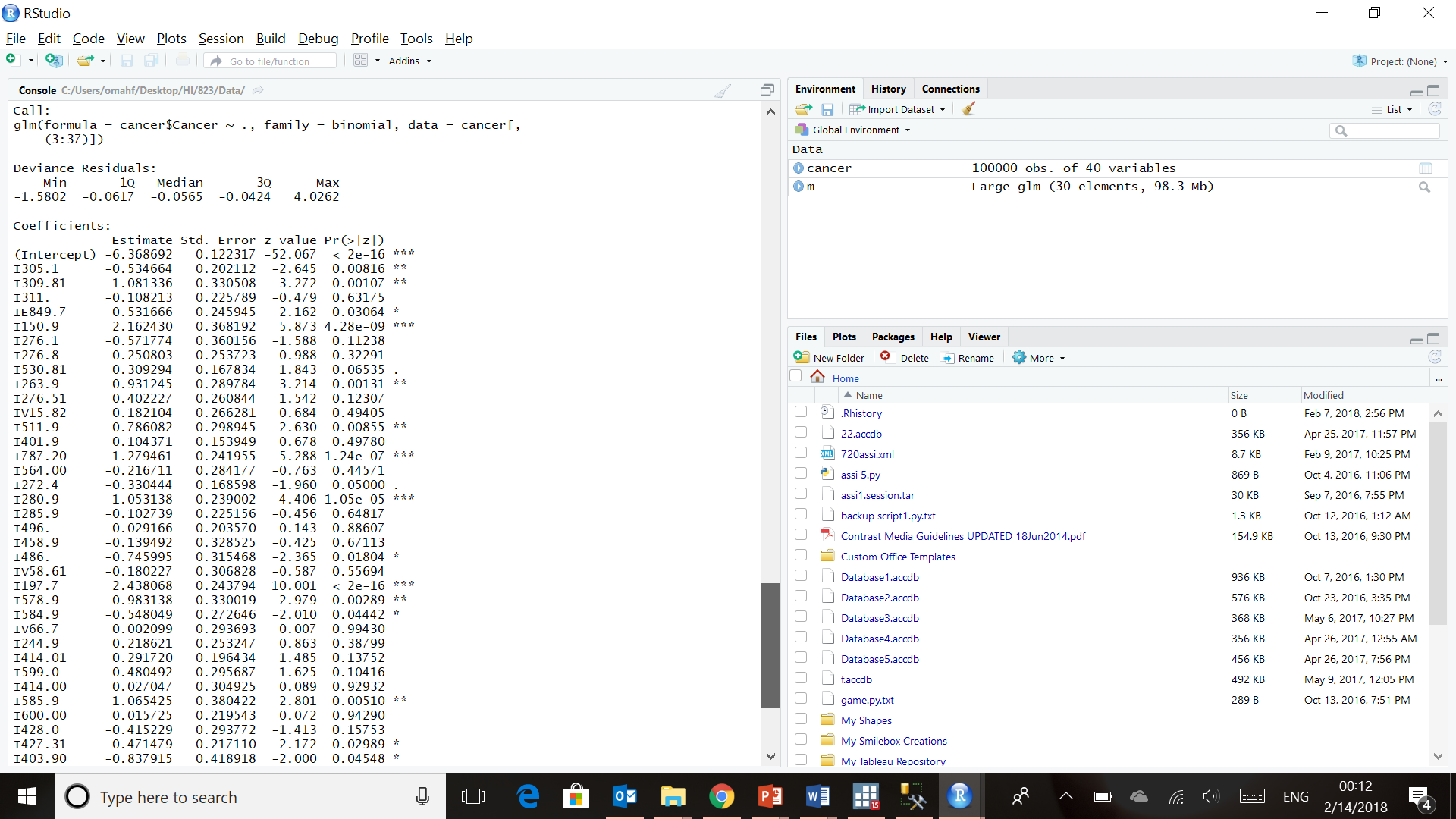
1. Does Dr. Smith see a different set of patients than his peer group?  In particular, what is the probability of patients being seen by Dr. Smith.  Regress the choice of provider on the 9 diagnoses provided.



To answer this question, we run a logistic regression model that shows that there is no significant difference between Dr.smith and his peer group(P-value=1), and the probability 50% that a patients being seen by Dr. Smith.

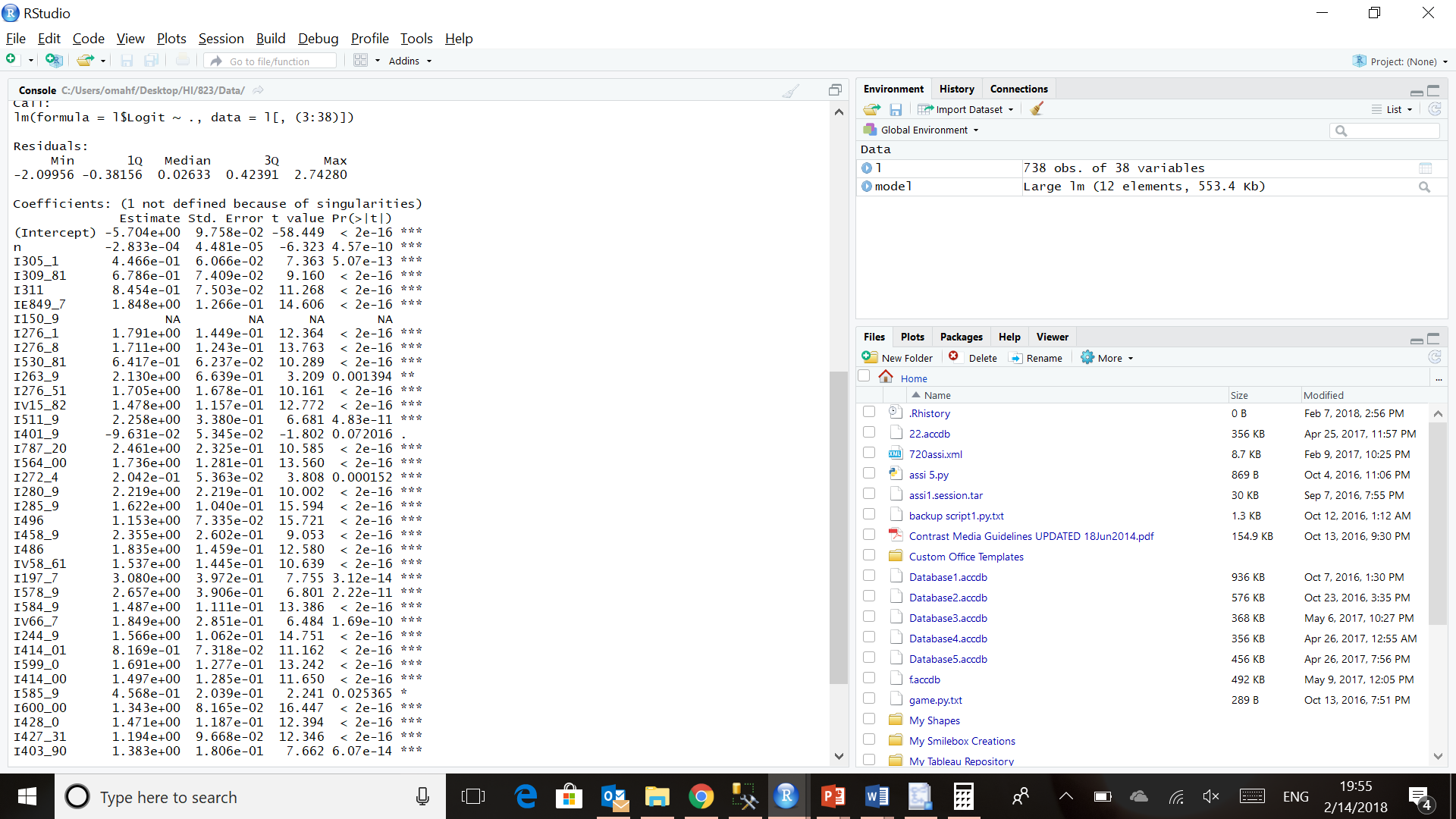
Question 4: The following data provide the survival among cancer patients.  The data provides 35 common comorbidities for patients who have or don't have stomach cancer. Use both logistic and ordinary regression to analyze these data and report the difference of the findings, in particular:

1. Using logistic regression, calculate the propensity to have cancer.



1. Group the diagnoses using SQL.  Transfer the grouped data into Excel. Within the naturally occurring groups of diagnoses, calculate probability of cancer.   Calculate the logit of the probability.  Regress the logit function on the diagnoses.

The image below shows linear regression result .we regress the logit over all the diagnosis.



We can see that all of the variables became significant except for 1401\_9 which is not significant .also we can see that there is differences in the coefficients between the two models