Statistical Analysis of Pima Indians using R

Arpan Niyogi

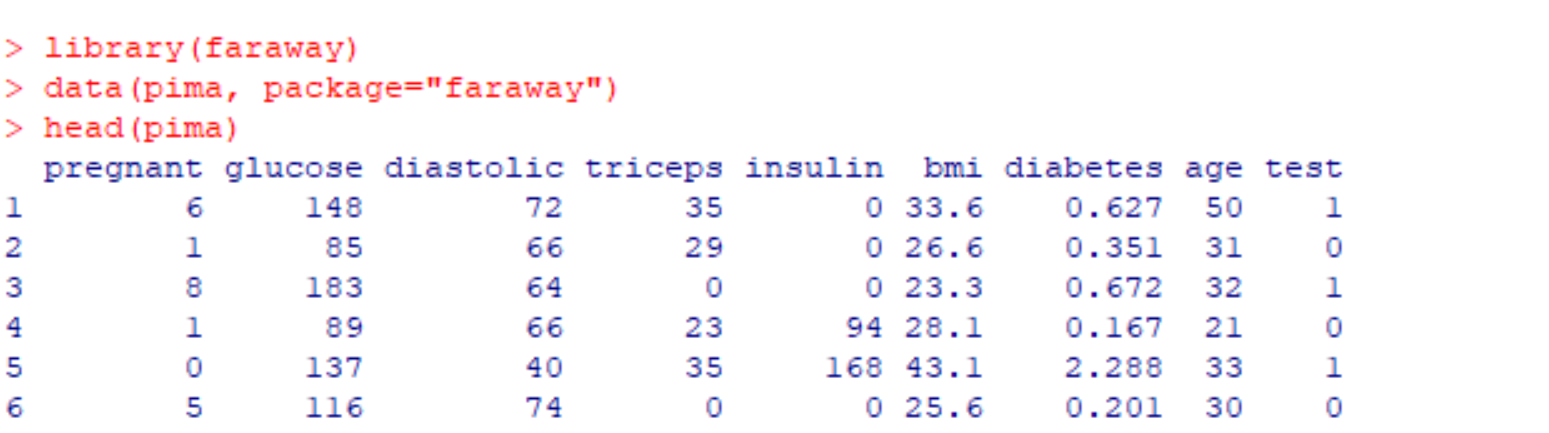
Section: 1

ABSTRACT

Statistical analysis of the Pima Indians diabetes dataset and interpretation of the findings. Figuring out how to find the overall summary of the data, as well as how to do various types of stepwise regression in order to understand the true significance of each variable involved in the dataset. In addition, interpretation of graphs and understanding how they relate to the raw dataset.

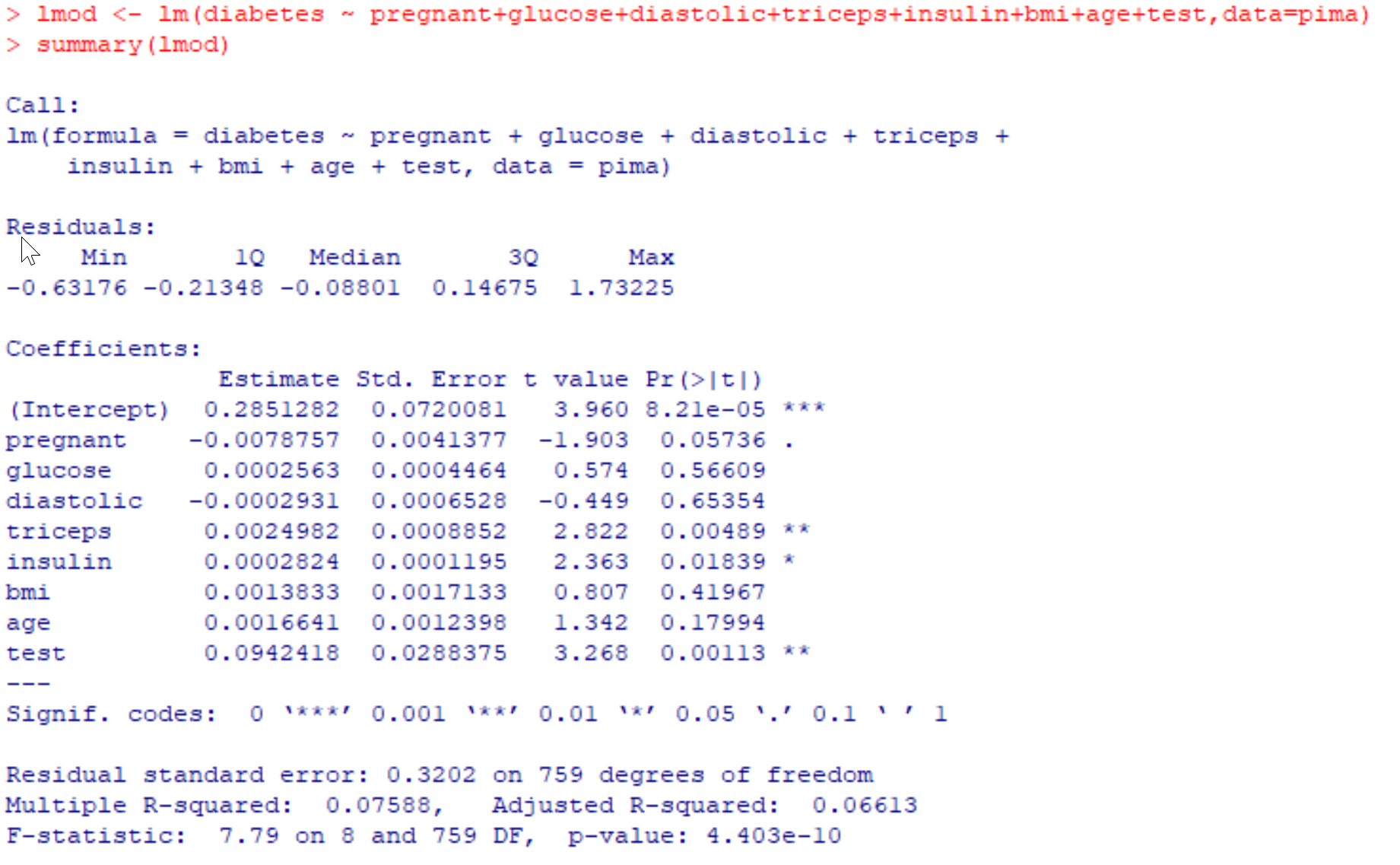
INTRODUCTION

The goal is to analyze the Pima Indians dataset and actually understand what all the numbers indicate in the below snippet of code using R. We want to summarize all the information from the raw dataset as well as perform various kinds of stepwise regression in order to understand the importance of each variable. If a nonsignificant variable is found, it is removed from the model. Therefore, we need to find these variables and remove them if they are nonsignificant.



MATERIALS AND METHODS

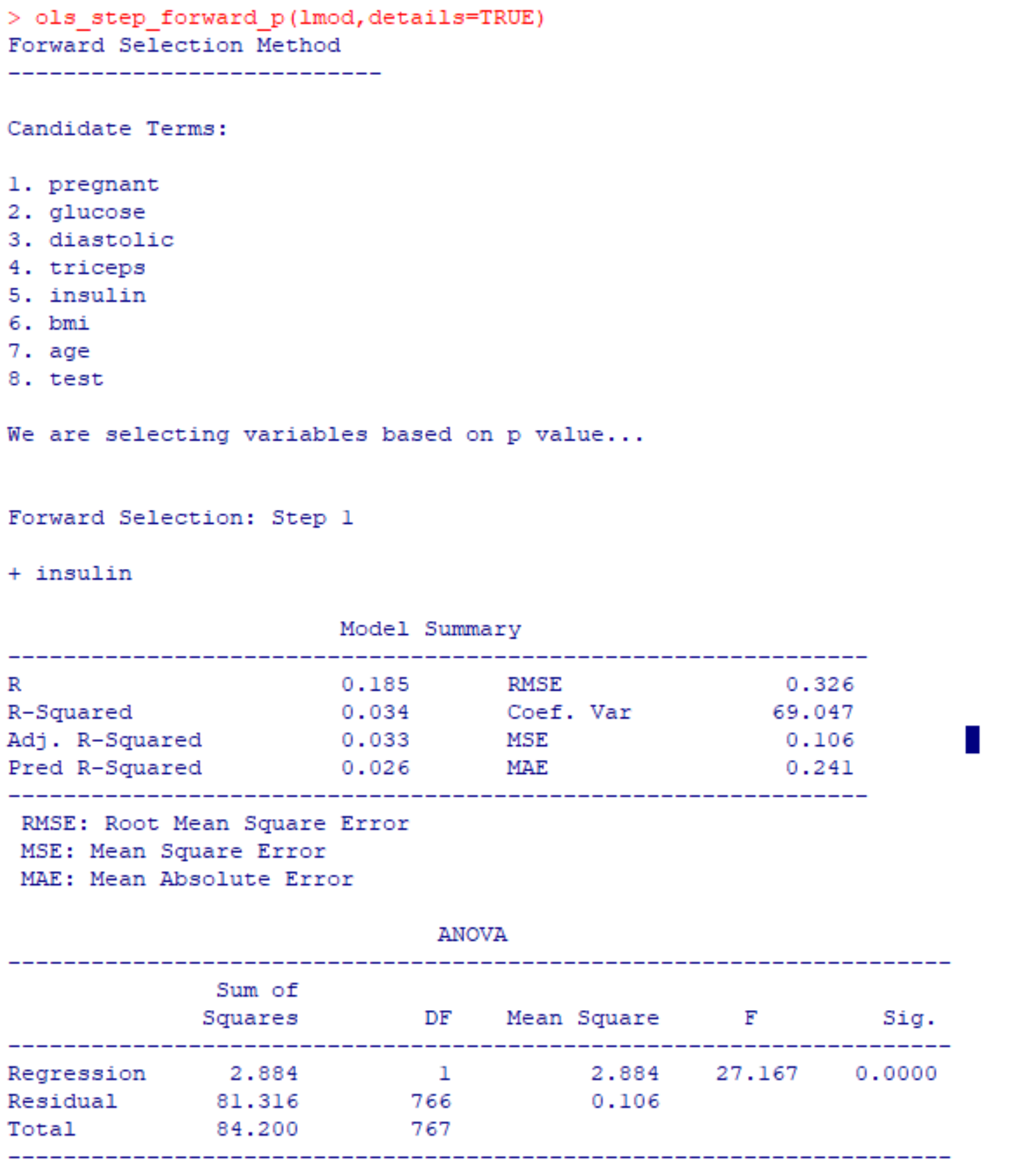
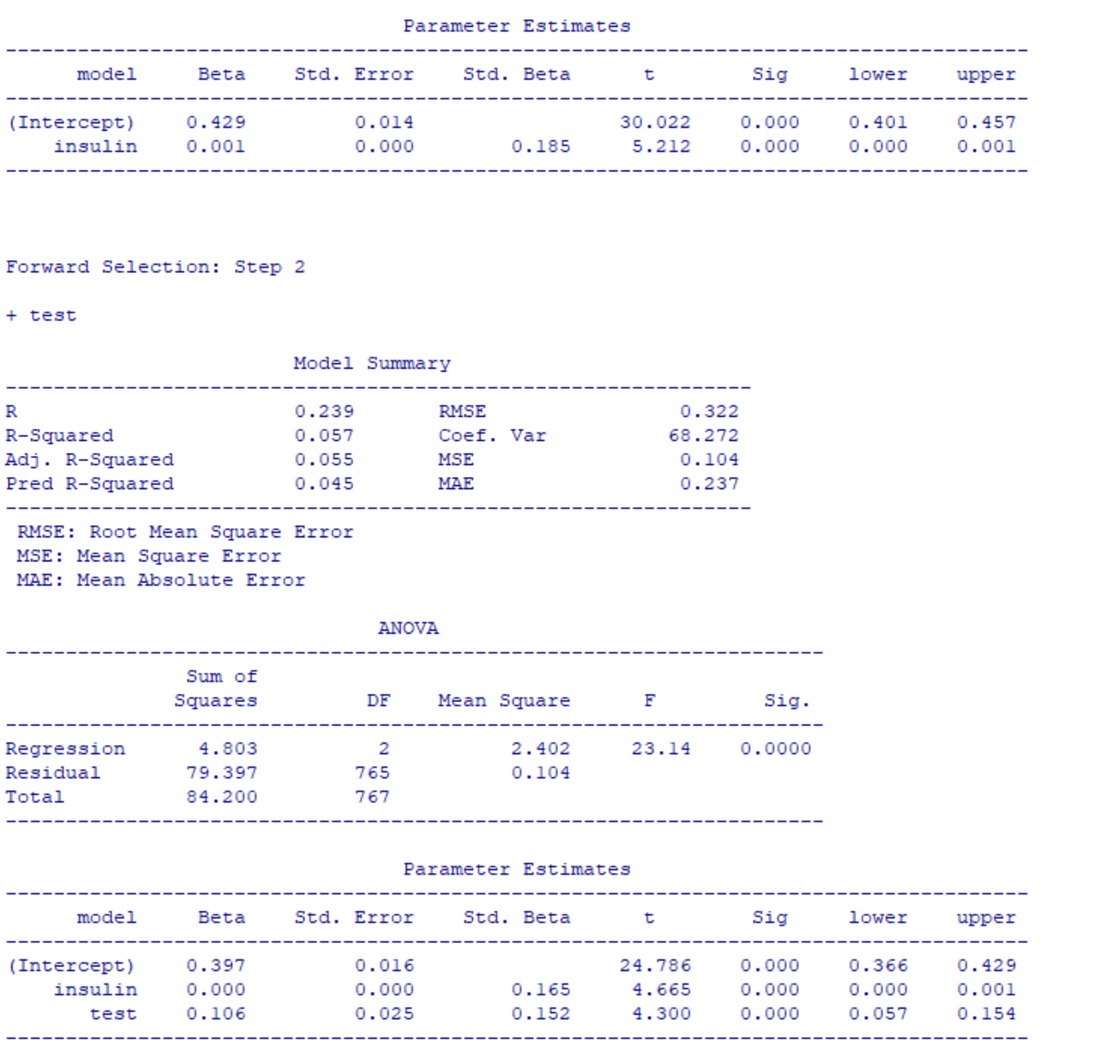
Stepwise Regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. We can use forward and/or backward regression depending on how the variables are added or removed from the model. First, we would have to create a linear model with the diabetes variable being the independent variable. Consequently, all the other variables would be dependent/predictor variables in the model. Below is the linear model.

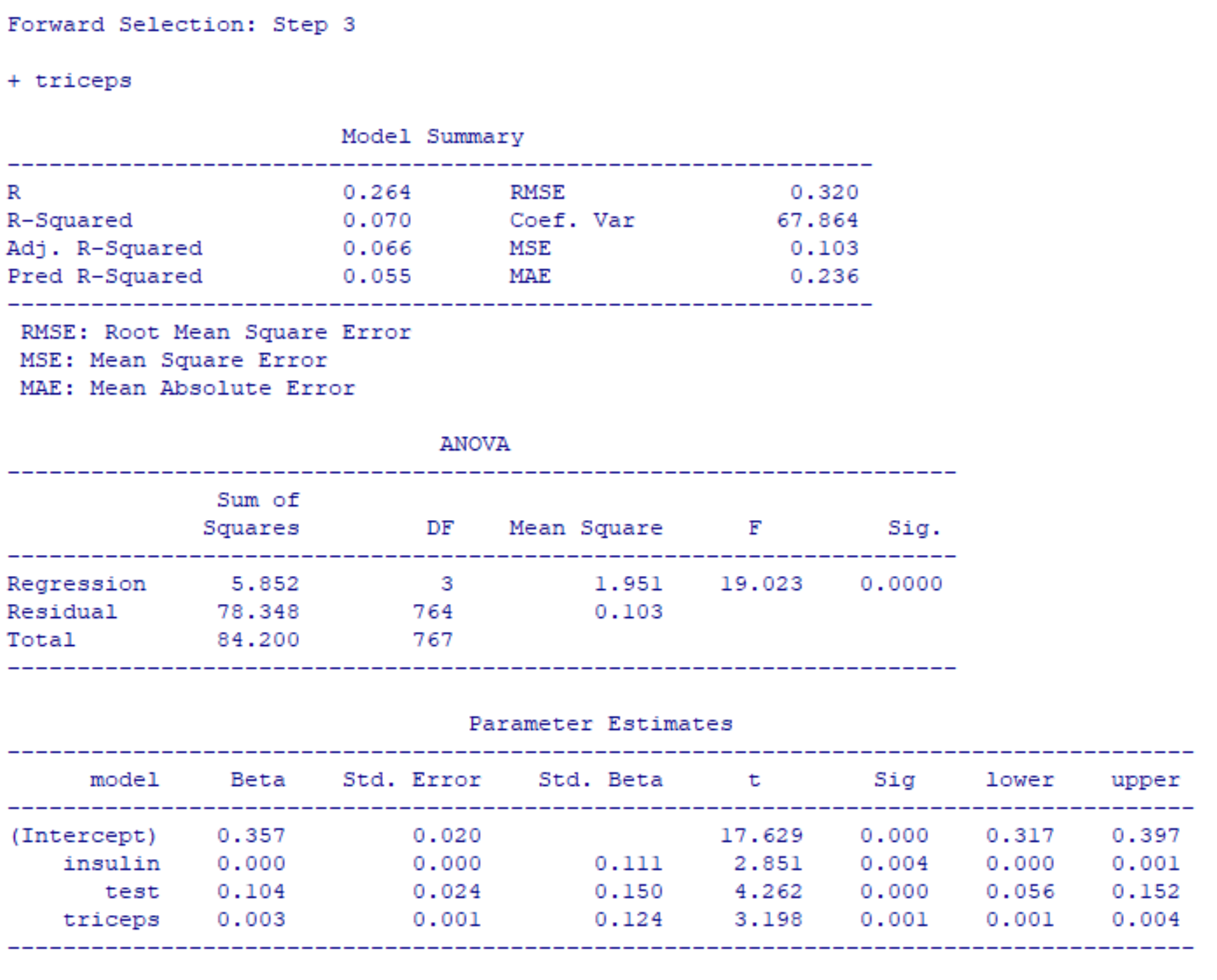
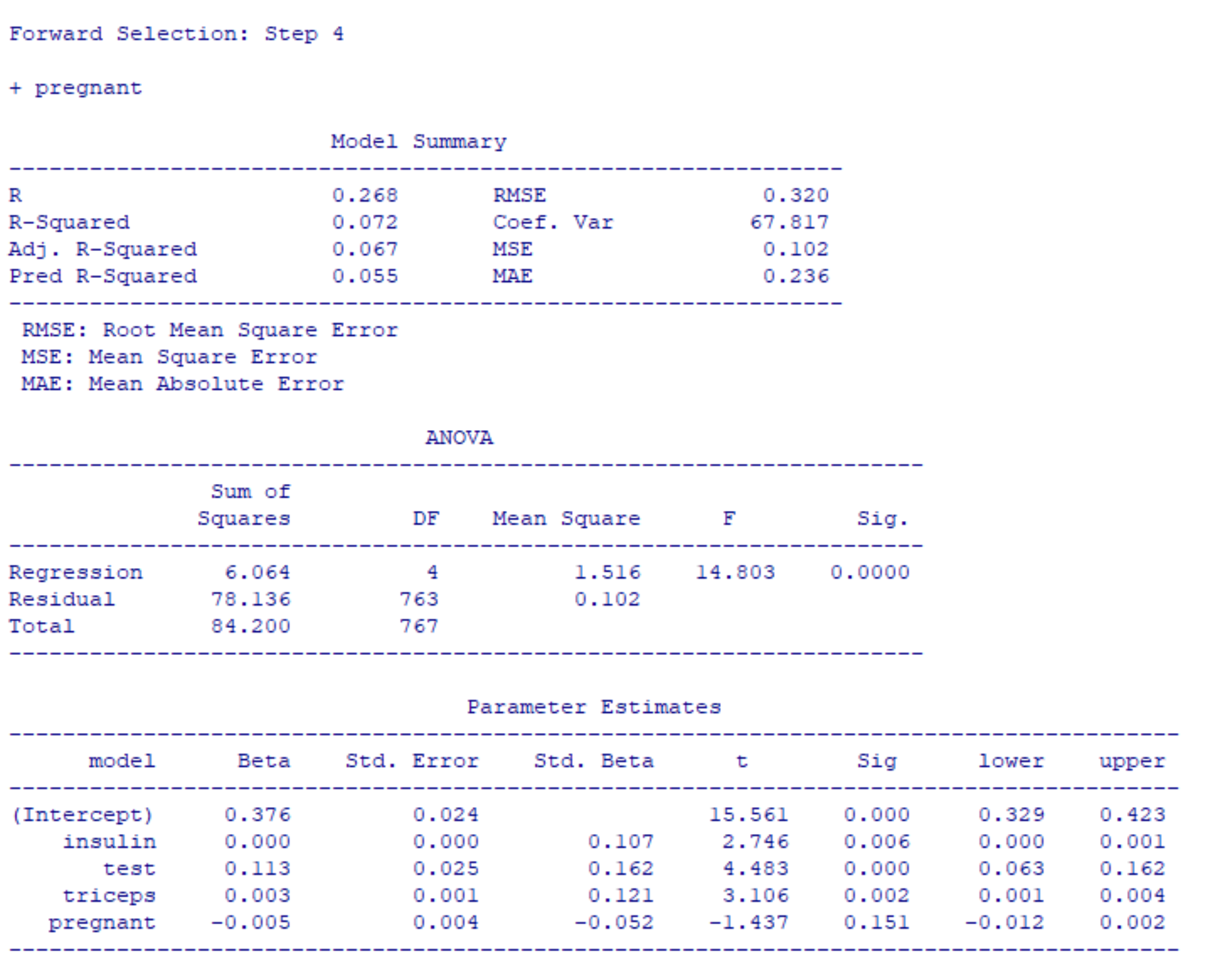


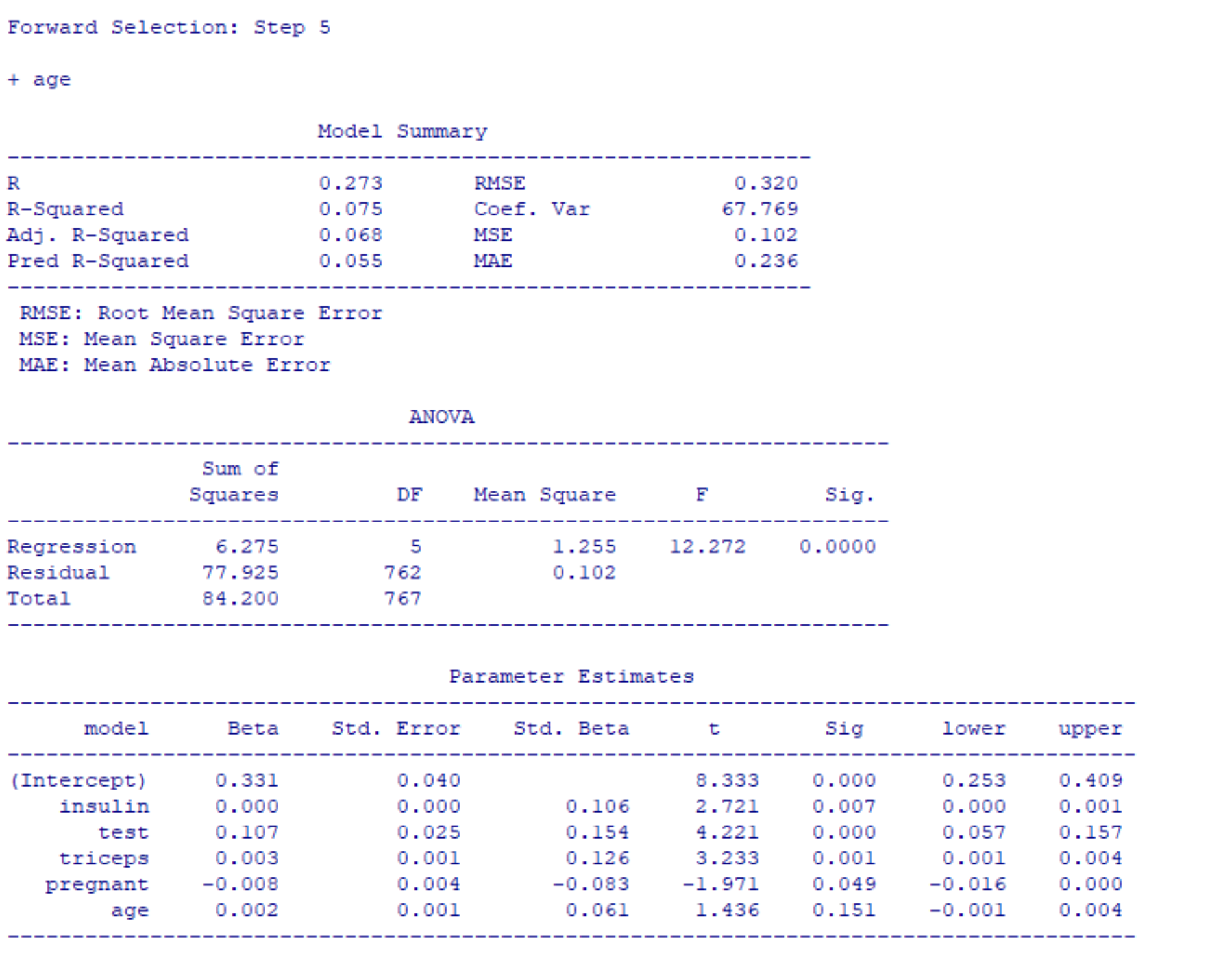
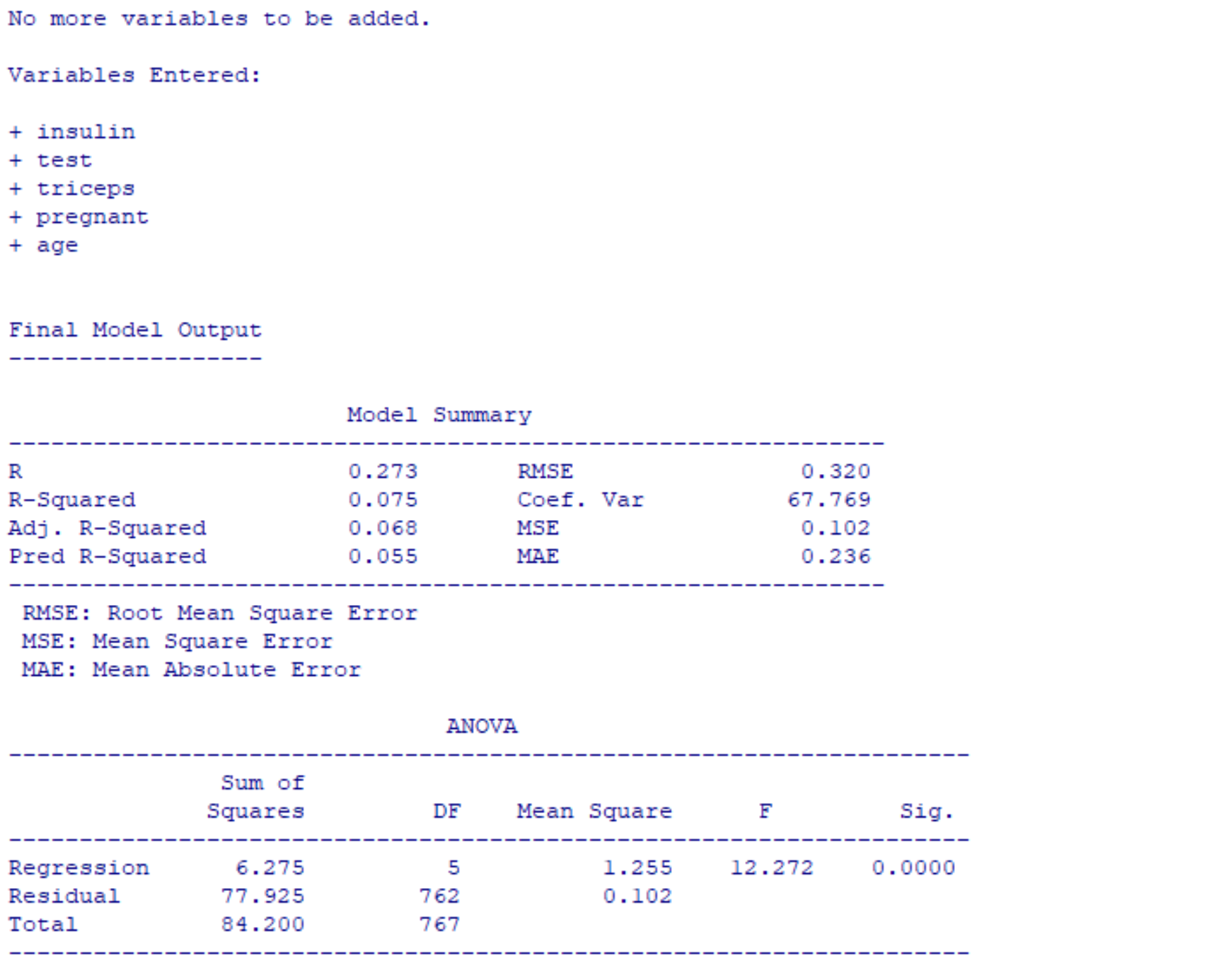
MATERIALS AND METHODS (Cont.)

After we summarize the model, we need to perform forward stepwise regression on the model to check the significance of each variable. Forward selection is a type of stepwise regression which begins with an empty model and adds in variables one by one. In each forward step, we add the one variable that gives the single best improvement to your model. Similarly, we also perform backward elimination/selection on the model for similar reasons. All the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation. The combination of these two methods of stepwise regression will help us understand the true significance of each variable efficiently.

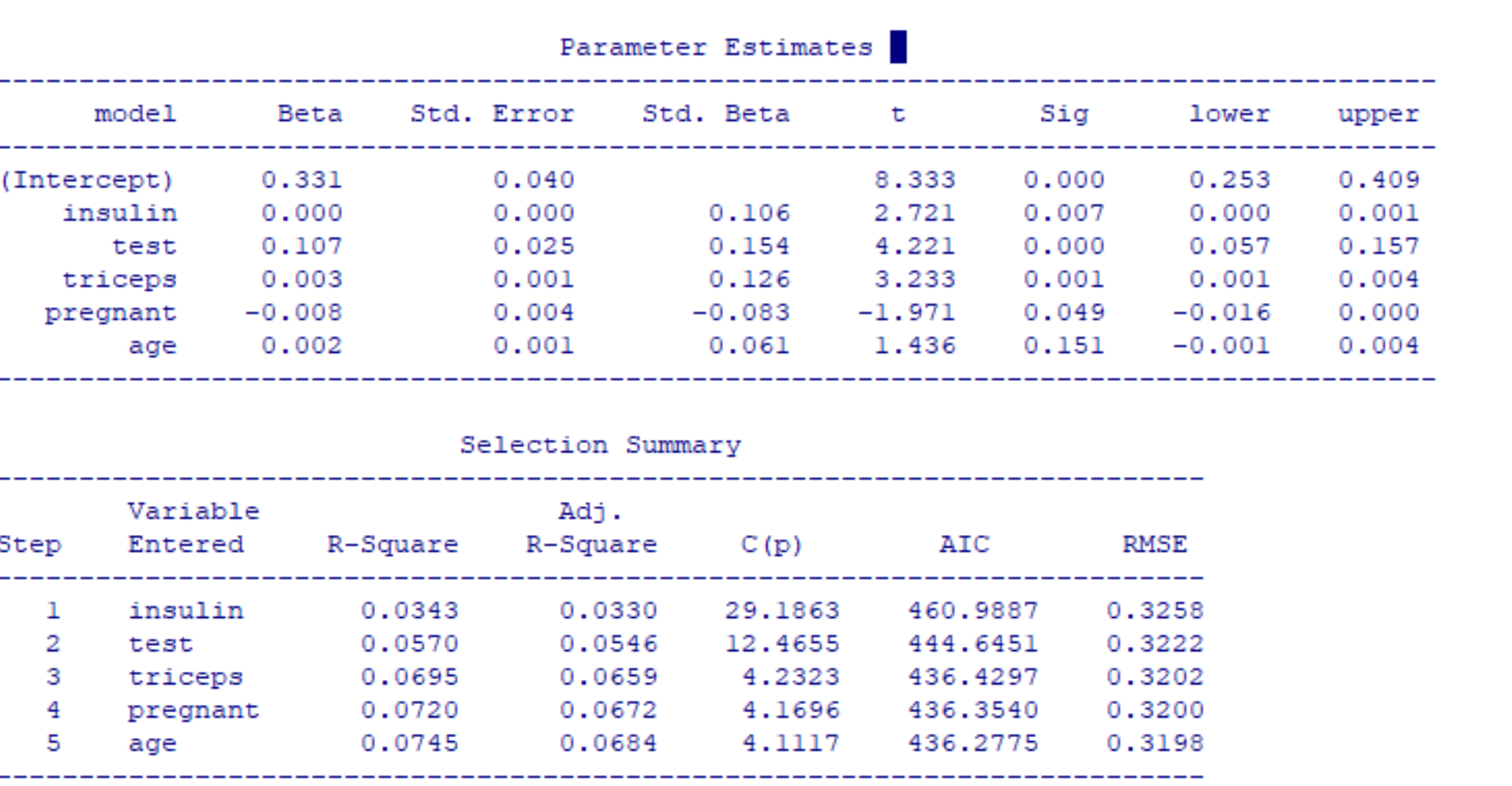
RESULTS

After performing forward stepwise regression, we can see that the variables that are kept in the model are: Insulin, test, triceps, pregnant, and age. Below are the steps shown using R. 

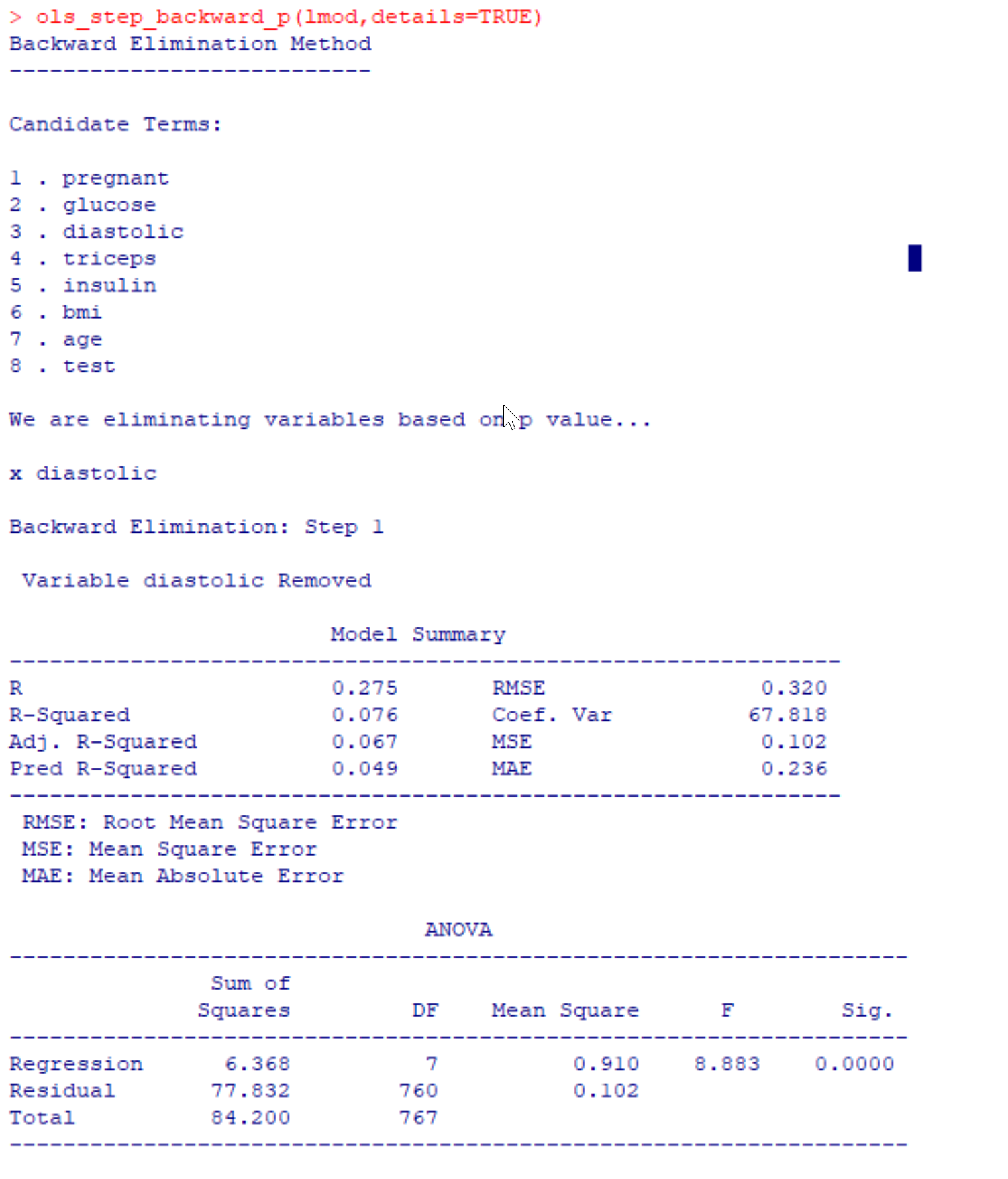
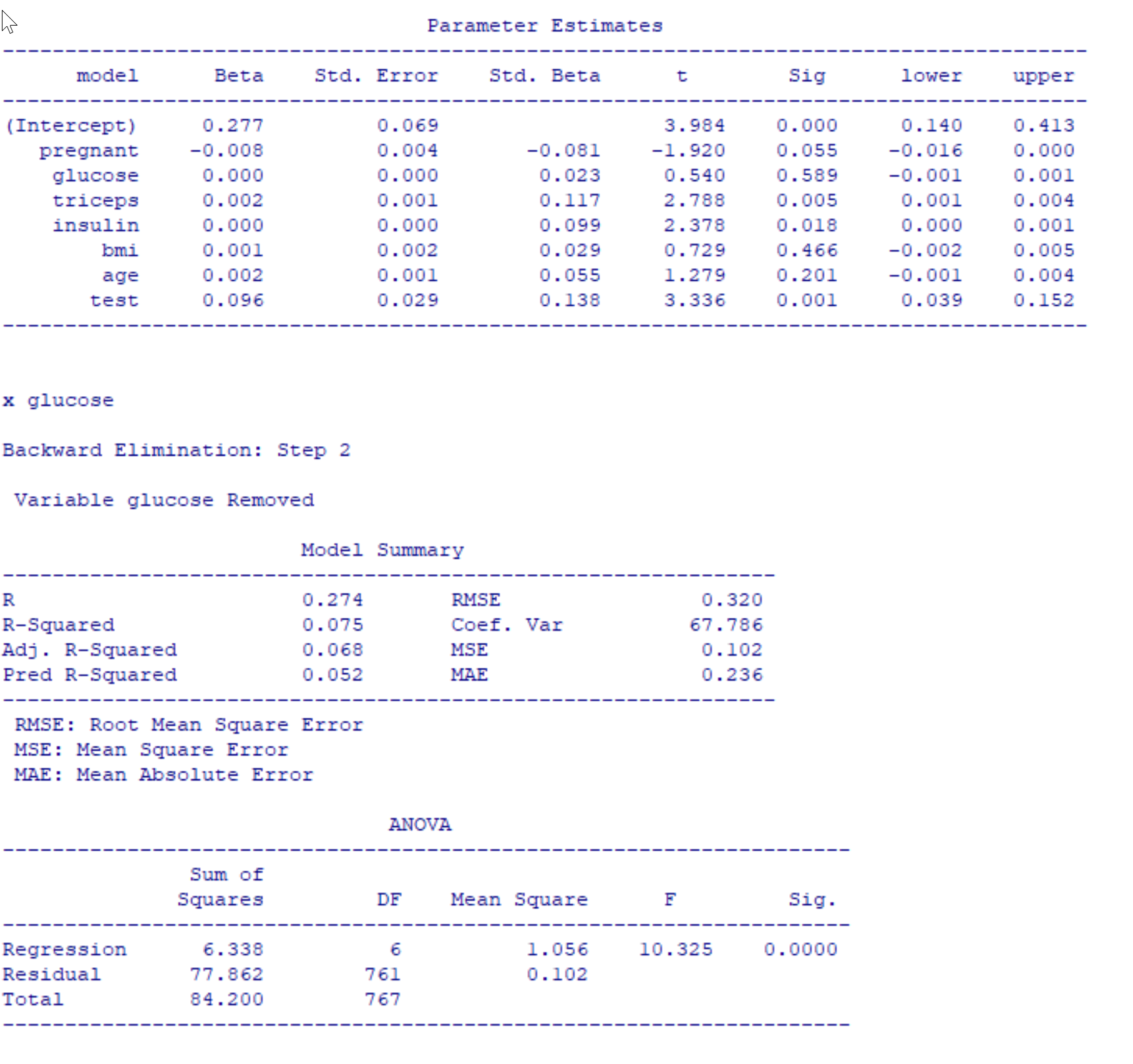
 

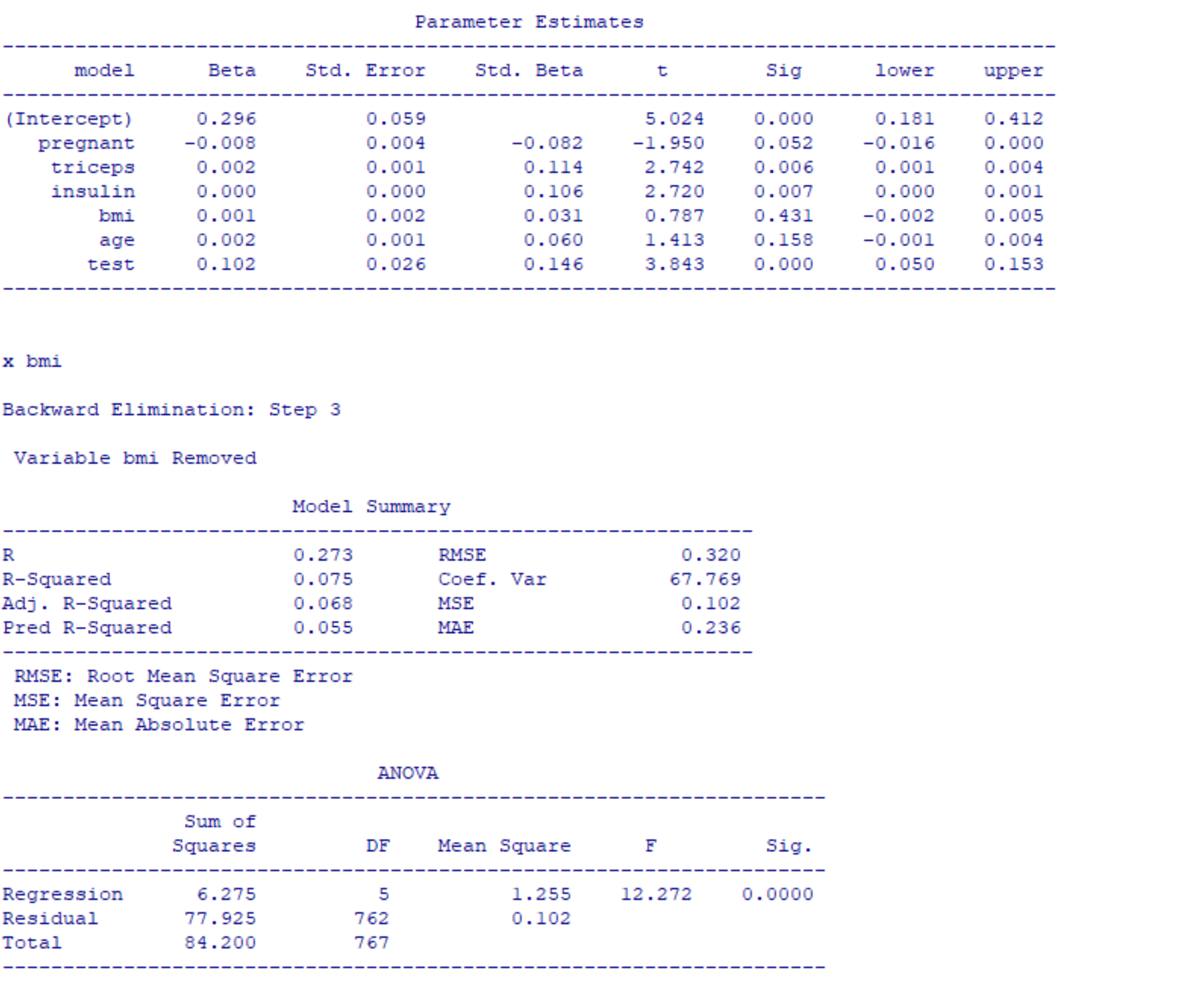
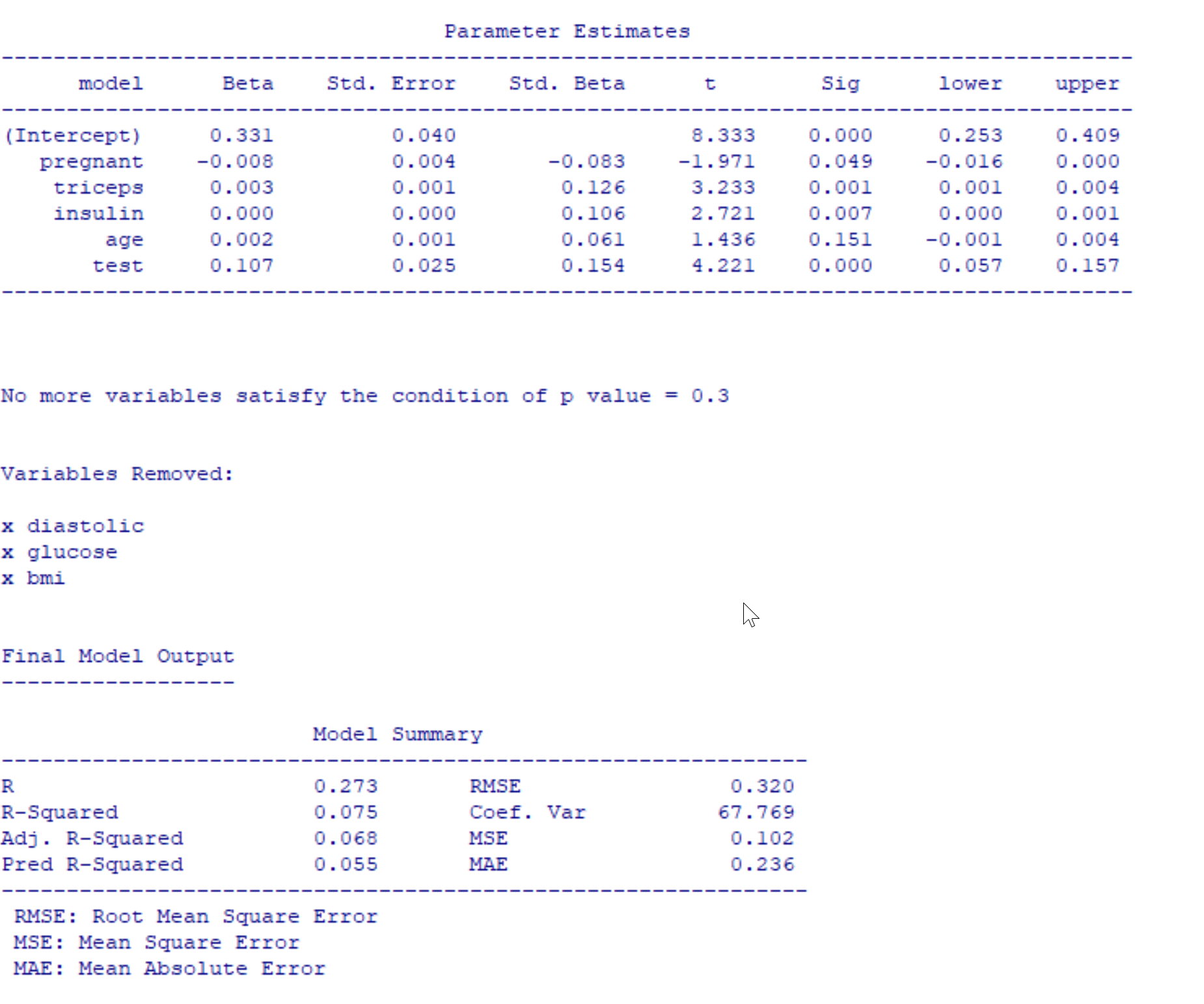
 

Final Step:

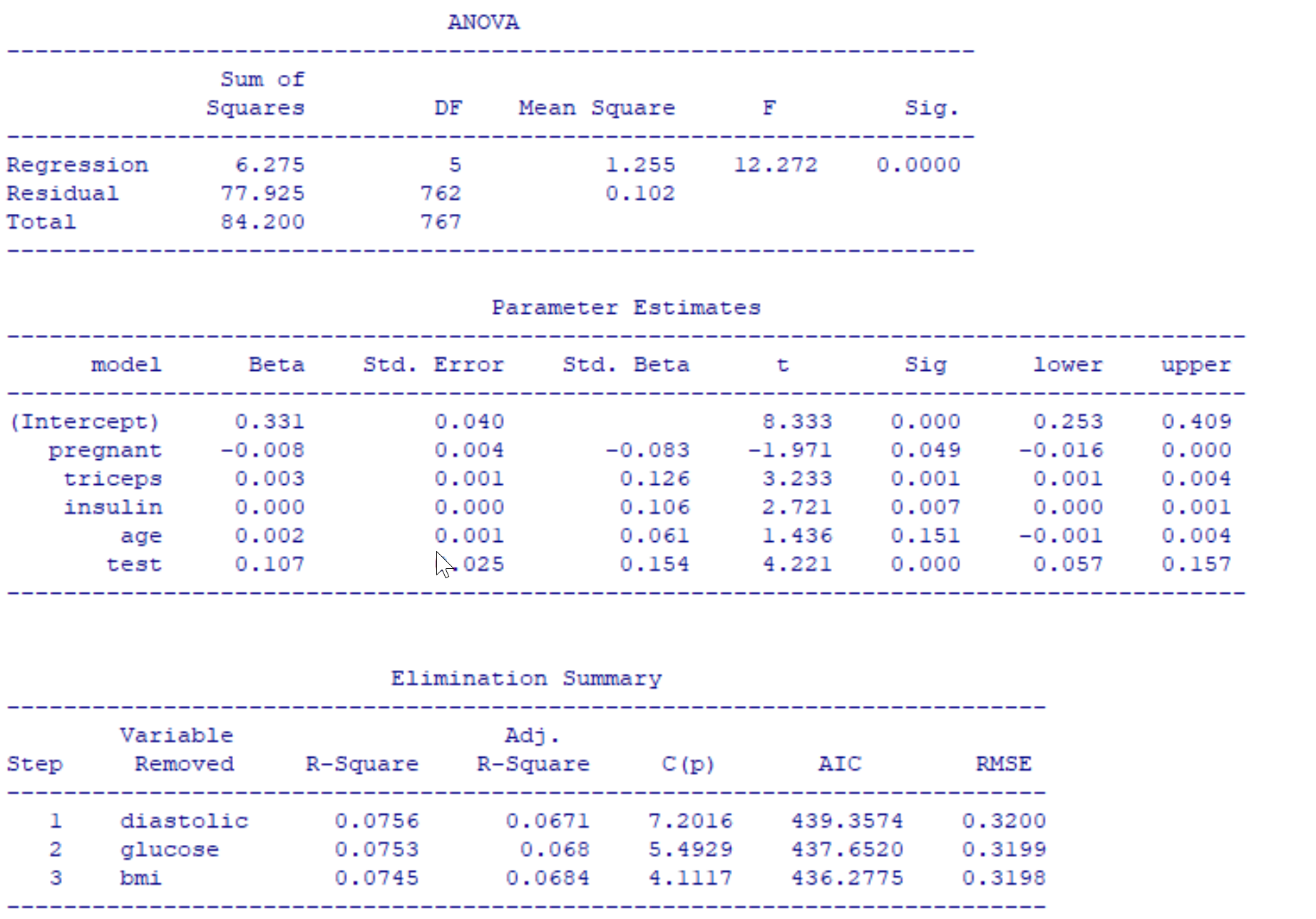


After performing backward stepwise regression, we can see that the variables that are removed from the model are: diastolic, glucose, bmi. Below are the steps shown using R.

Final Step:



DISCUSSION

After we created our linear model for the Pima Indians dataset, we summarized it and we have efficiently gotten information such as the p-value, adjusted R-squared value, F-statistic, Residuals, etc. This snippet of information could be found in the introduction section where the snippet of code displays the summary of the linear model.

Then, we performed forward stepwise regression. We did this in order to weed out any nonsignificant variables so as to make out model more efficient. Similarly, we performed backward elimination. This is another method in order to remove relatively inconsequential variables from the regression model. The findings have been shown in the results section.

Overall, summarizing linear regression models and performing stepwise regressions are incredibly important in analysis of data.

LITERATURE CITED

1. NCSS, LLC. All Rights Reserved. Stepwise Regression. PDF. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise\_Regression.pdf

2. PIMA Indian Diabetes - Logistic Regression with R. (2017, April) Retrieved from <https://www.kaggle.com/ksp585/pima-indian-diabetes-logistic-regression-with-r>

# 3. Using Predictive Models to Classify Pima Indians Diabetes Database. (2017, December 27). Retrieved from <https://rstudio-pubs-static.s3.amazonaws.com/346228_a62c6c91d5cf40869cd5aef7206826ae.html>