**Instructions:**

**You can use Word, Excel, Power Point and SAS to answer the questions in this exam. There are a total of six (6) multi-part questions, with point values noted for each question.**

**Please show your calculations, or the details of your program(s) for each problem. Include your SAS programs and output in your submission. The SAS programs should be commented so that each step is clearly explained.**

**Combine all your answers/files into a single zipped file and post the zipped file to “Final Submissions” in Moodle.**

**Important: Please fill out the optional survey of the class at bottom of this page, as well as in CANVAS.**

**Problem #1: (15 points)**

X number of high school students are scored on various tests, such as science, math, and social studies (**socst**). The variable **female** is a dichotomous variable, coded 1 if the student was female and 0 if male. Using the multiple regression analysis results below, answer the following questions:

* How many students were scored?

The number of students should be the number of corrected total plus 1.

So there are 200 students scored

* Is the overall model significant?

Pr<0.0001, so the model is significant

* What is the F-value (1-?)?

F-value=MSR/MSE=2385.93019/(7.14817)^2=46.69477

* What is the R-square for this model (2-?)?

R-square=SSR/SST=9543.72074/19508=0.4892

* What is the formula for this model?

Y=12.32529+0.38931\*math-2.00976\*female+0.04984\*socst+0.33530\*read

* Is this a good model? Why or why not?

The model is significant and the residual distributes evenly. However, the coefficients of variables are not that significant, especially the Pr for socst is 0.4241. So it is not a good model

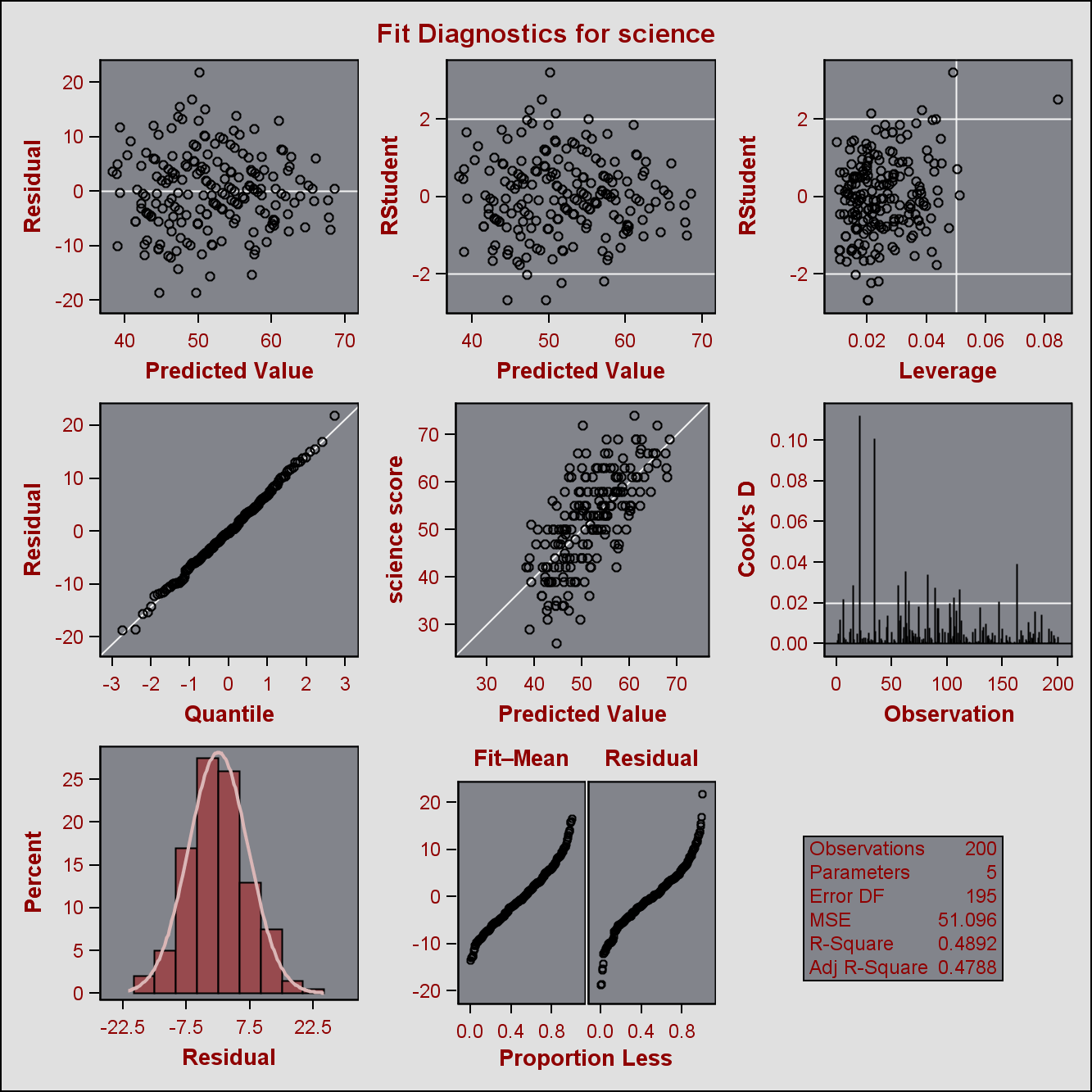
* Would you change the model? If yes, How?

Yes. Just remove the variable socst because the variable is not significant

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 4 | 9543.72074 | 2385.93019 | **1-?** | <.0001 |
| **Error** | 195 | 9963.77926 | 51.09630 |  |  |
| **Corrected Total** | 199 | 19508 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 7.14817 | **R-Square** | **2-?** |
| **Dependent Mean** | 51.85000 | **Adj R-Sq** | 0.4788 |
| **Coeff Var** | 13.78624 |  |  |

| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | |
| **Intercept** | Intercept | 1 | 12.32529 | 3.19356 | 3.86 | 0.0002 | 6.02694 | 18.62364 |
| **math** | math score | 1 | 0.38931 | 0.07412 | 5.25 | <.0001 | 0.24312 | 0.53550 |
| **female** |  | 1 | -2.00976 | 1.02272 | -1.97 | 0.0508 | -4.02677 | 0.00724 |
| **socst** | social studies score | 1 | 0.04984 | 0.06223 | 0.80 | 0.4241 | -0.07289 | 0.17258 |
| **read** | reading score | 1 | 0.33530 | 0.07278 | 4.61 | <.0001 | 0.19177 | 0.47883 |



**Problem #2: select one (5 points)**

A software package has produced the following output for a regression model estimating the nutritional ratings of cereals, based on the location of the cereal on a super market shelf (shelf1, shelf2). Is this model a good regression model?

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** | 45.22003 | 2.23245 | 20.26 | <.0001 |
| **shelf1** | **1** | 0.92541 | 3.73561 | 0.25 | 0.8050 |
| **shelf2** | **1** | -10.24721 | 3.67798 | -2.79 | 0.0068 |

1. The model is NOT a good model because variable shelf2 and “Intercept” are not significant at 5%
2. **The model is NOT a good model because variable shelf1 is not significant at 5%**
3. The model is NOT a good model because the location of cereal (“shelf1 vs. shelf2) has nothing to do with ratings and cannot cause a change in cereal ratings.
4. Both a and c

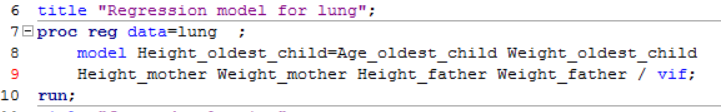
**Answer: B**

**Problem #3: (20 points)**

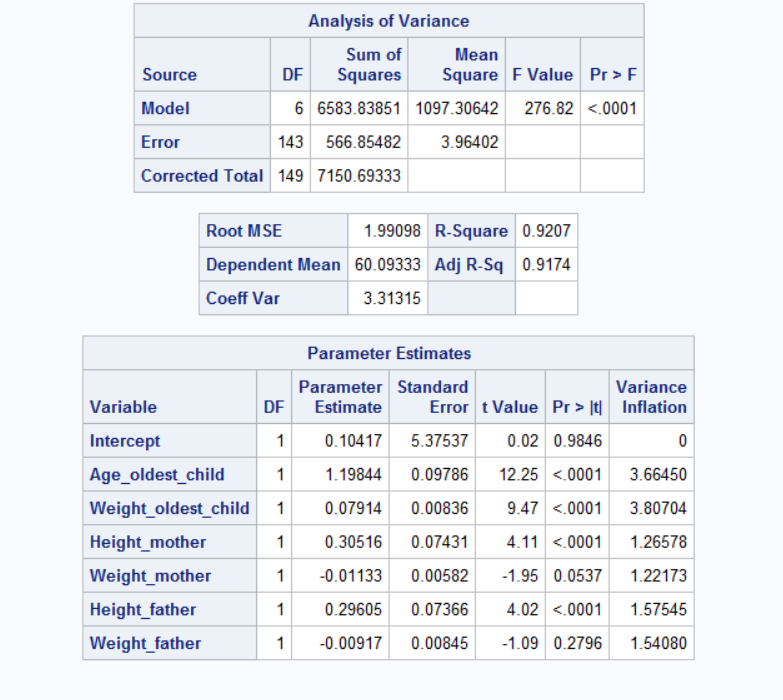
1. Use the Lung dataset in CANVAS, and forward, backward, and stepwise selection methodologies to develop multiple regression models for “HEIGHT of Oldest Child” as dependent variable and “AGE of Oldest Child”, “WEIGHT of Oldest Child”, “HEIGHT of Mother”, “WEIGHT of Mother, “HEIGHT of Father” and “WEIGHT of Father” as independent variables. (Do not perform any data transformation).

**Solution:**

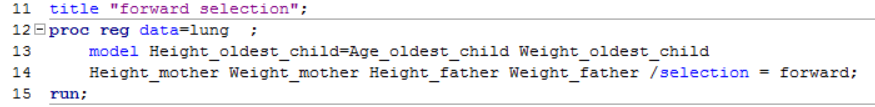
1. Firstly, I did a regression model for the lung dataset using all the variables

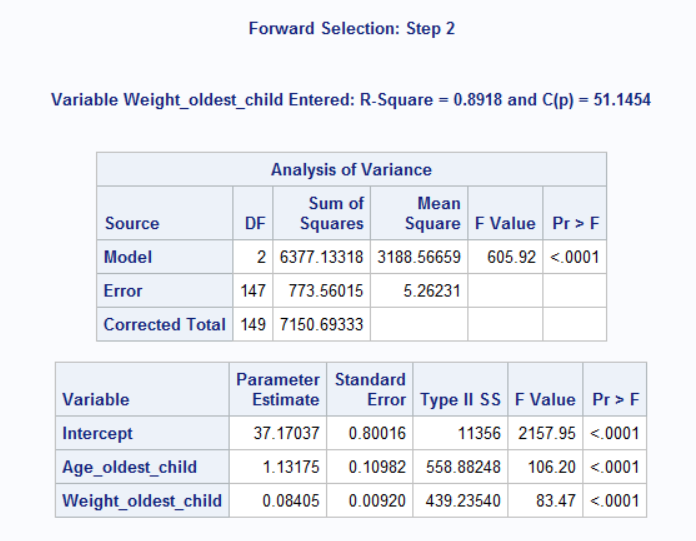


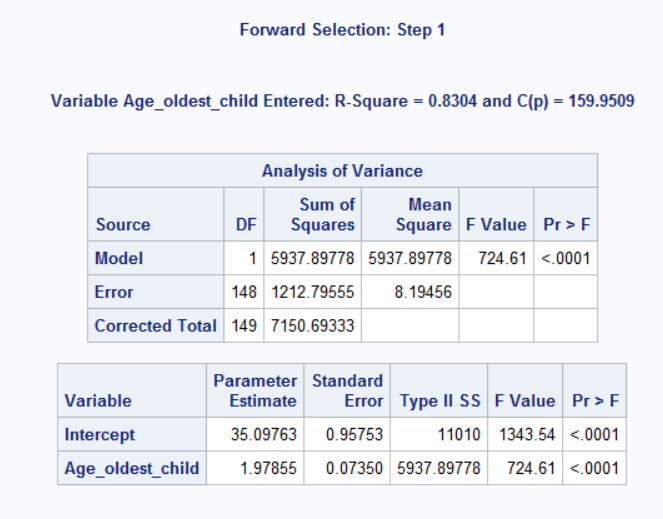
The result is as following

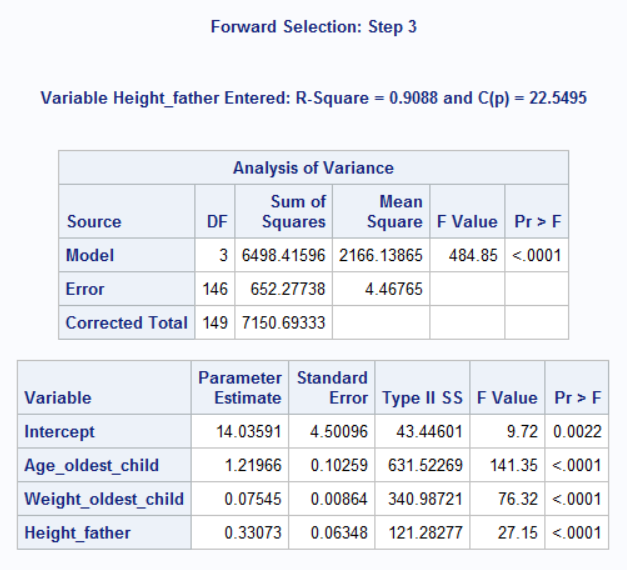
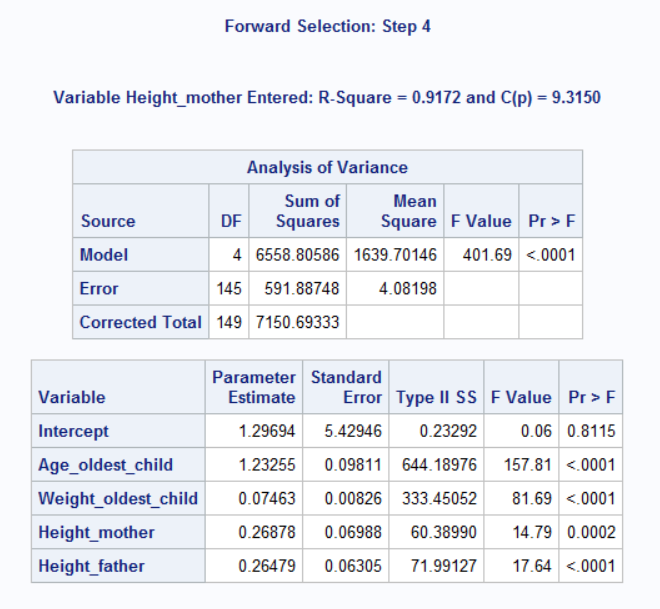


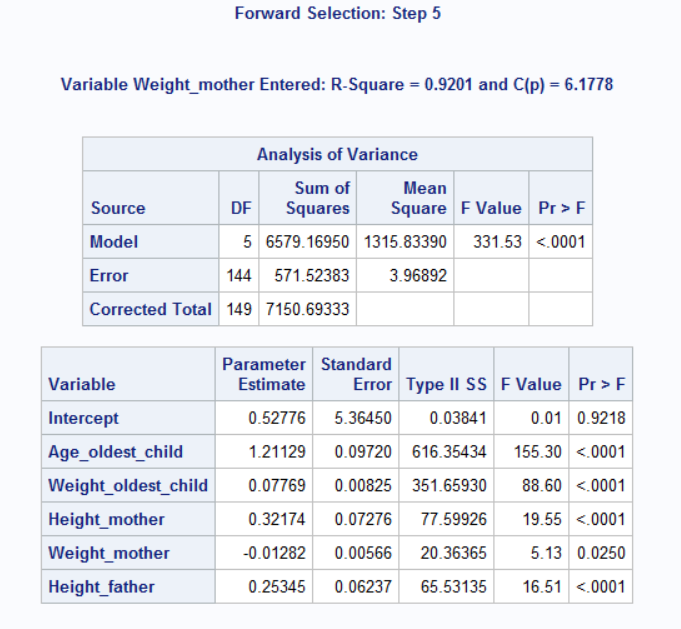
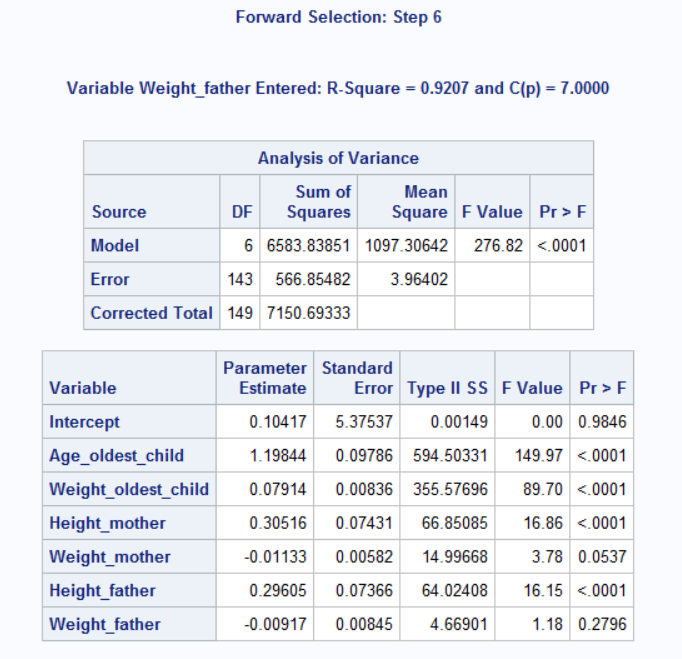
We can see that overall model is good. However, the variable Weight\_mother and Weight\_father are not significant based on the 5% assumption. What’s more, the vif is low for each variable, which means they are not dependent on each other.

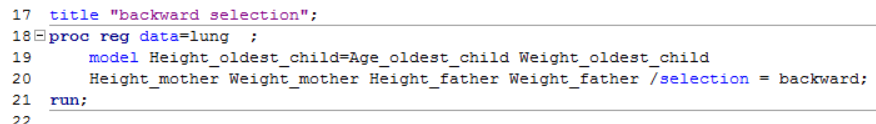
1. **Forward Selection**

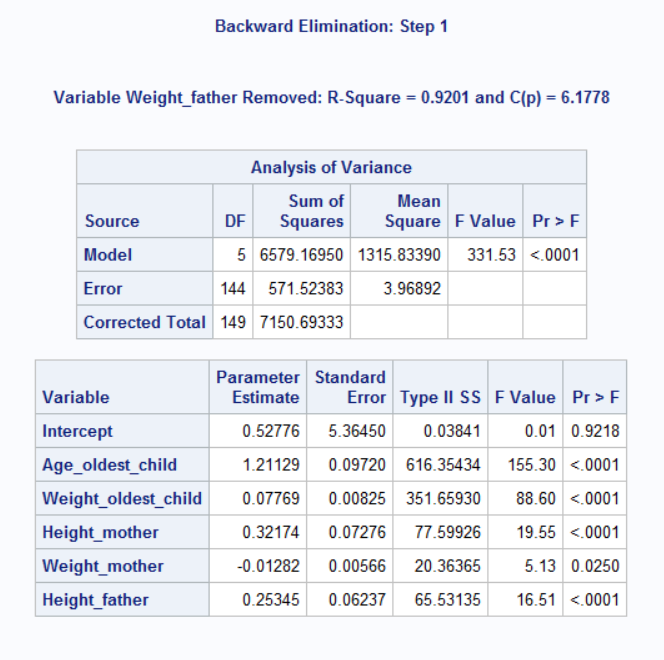
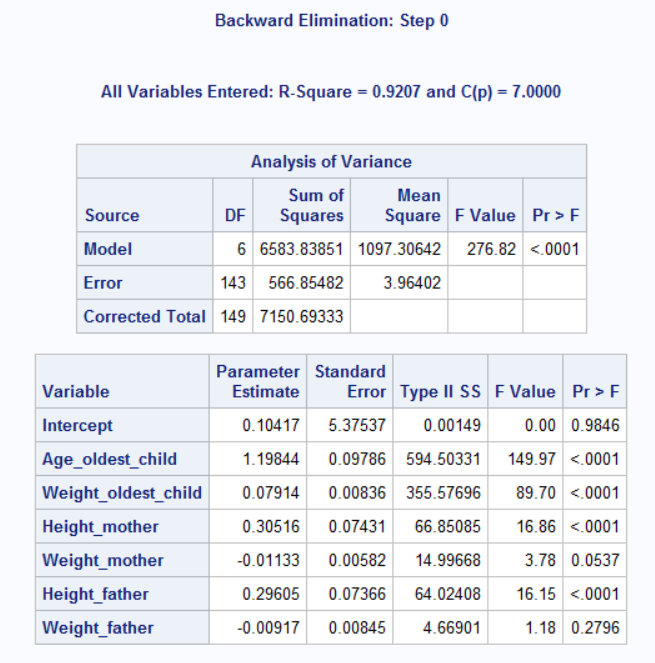
****

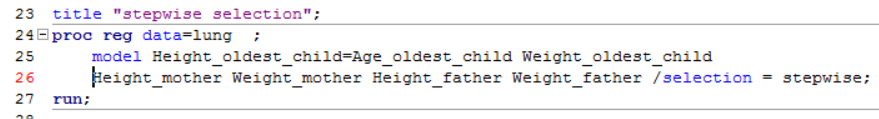


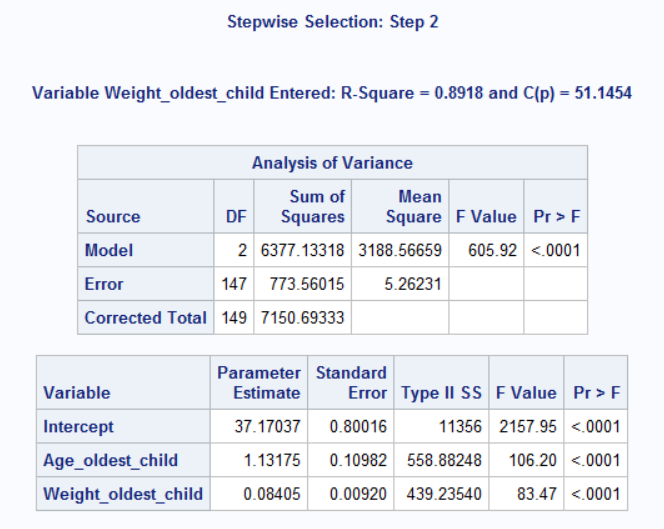
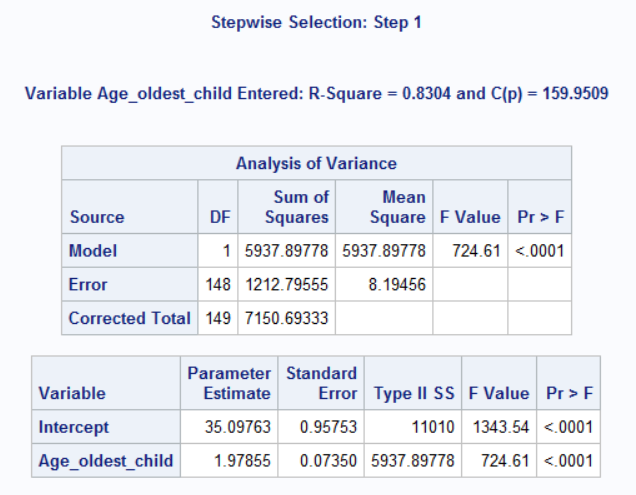


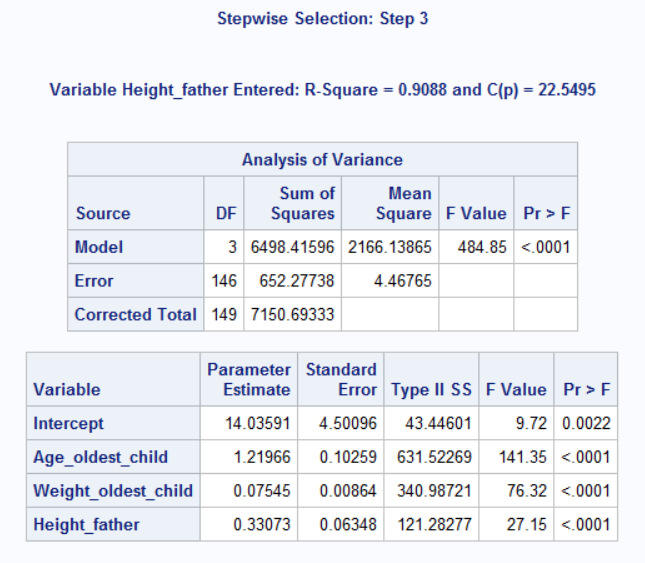
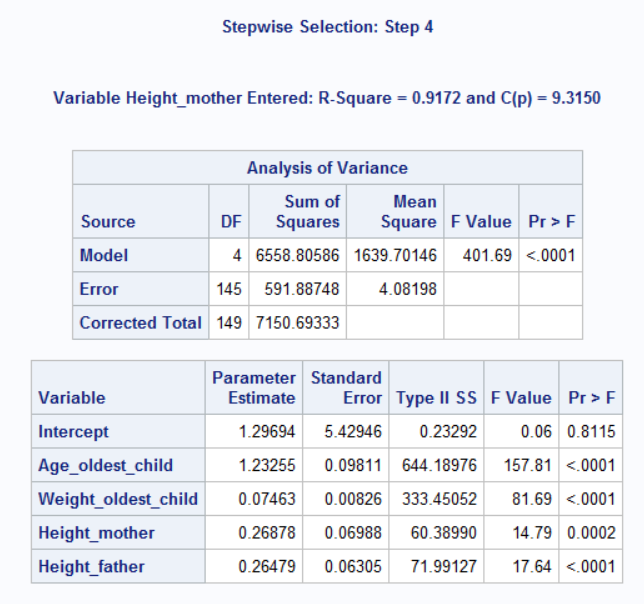


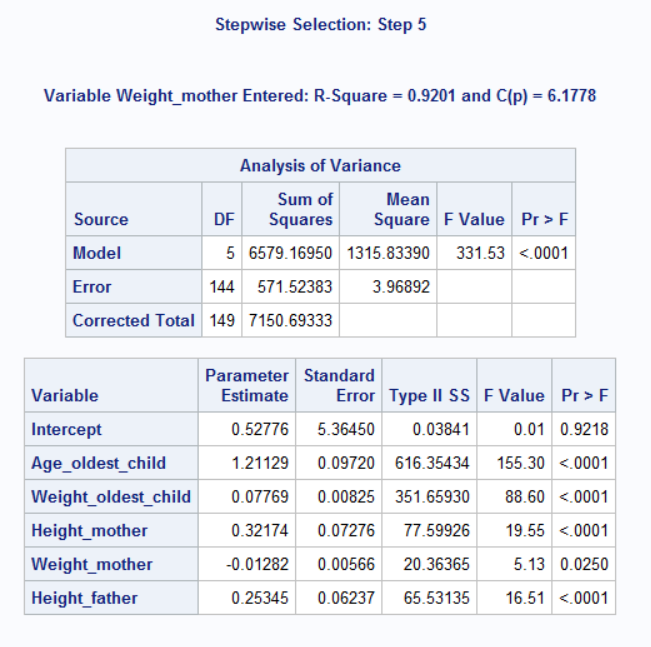
1. **Backward Selection**



1. **Stepwise Selection**



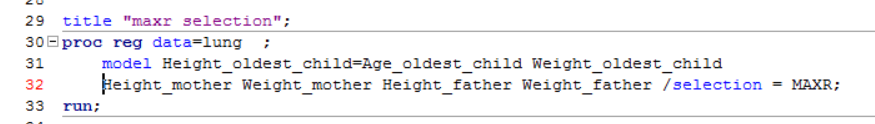


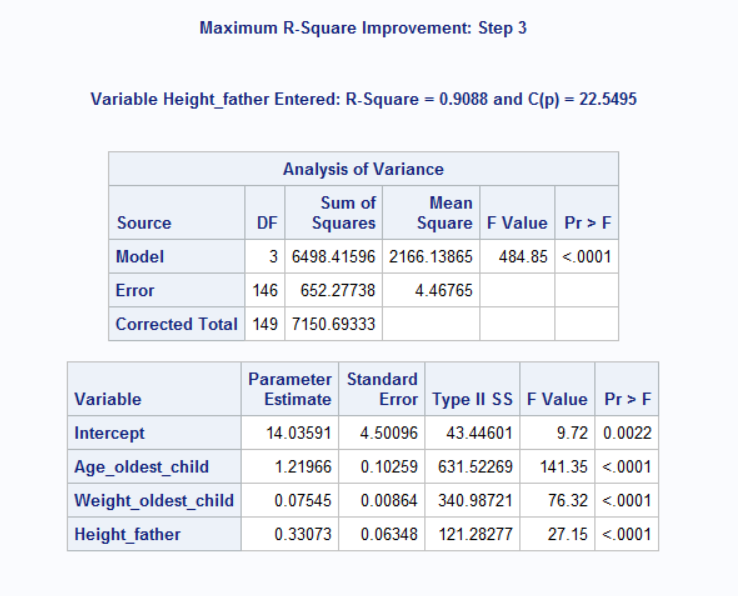


1. Find the best subset of the three variables

**Solution:**

Using the MAXR selection, the result for three variable is





So the best subset of three variables is (Age\_oldest\_child, Weight\_oldest\_child, Height\_father)

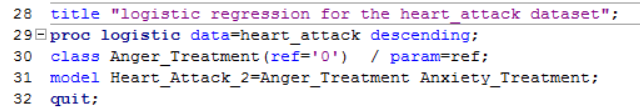
**Problem #4: (20 points)**

The “heart attack” dataset in CANVAS contain the records for twenty heat attack patients. The dependent variable (Heart\_Attack\_2) is an indicator showing whether the patient has had a second heart attack within 1 year (yes=1). The first independent variable “Anger Treatment”, indicates whether the patient completed an anger management treatment or not. The second independent variable (“Anxiety Treatment) shows the level of anxiety treatment of the patient.

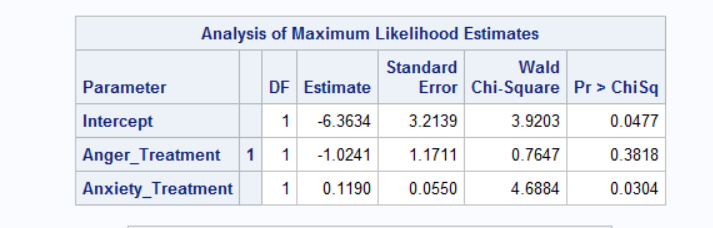
1. Develop a logistic regression model for predicting the probability of the patient having s second heart attack (show your development steps)

**Solution:**

* 1. First, develop a logistic regression model using all the variables

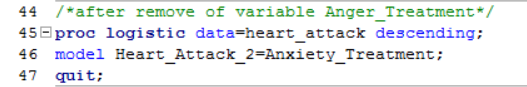


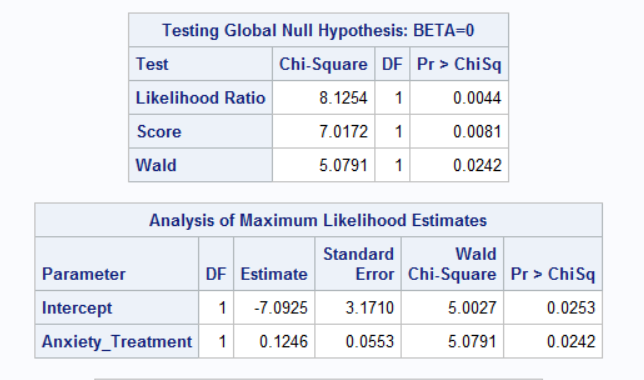
The result is following



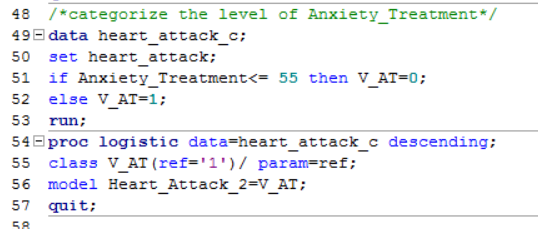
We could see that variable Anger\_Treatment is not significant, so I remove it from the model

* 1. After the remove of Anger\_Treatment

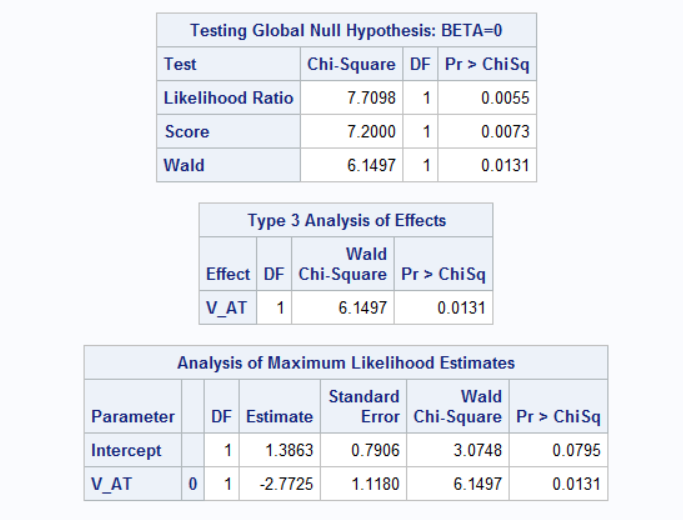




* 1. I also tried to categorize Anxiety\_Treatment into V\_AT that when Anxiety\_Treatment is below 60, V\_AT =0 and when Anxiety\_Treatment is above 60, V\_AT=1



The result is as following



**Conclusion:** we could see that the model is better to remove the variable Anger\_Treatment and there is no need to categorize the variable Anxiety\_Treatment

1. Using your model:
   1. Predict the probabilities of the following two patients (A and B) having a heart attack within the next year?

|  |  |  |
| --- | --- | --- |
| Patient | Anger Treatment | Anxiety Treatment |
| A | 0 | 40 |
| B | 1 | 70 |
|  |  |  |
|  |  |  |

**Solution:**

1. After the remove of Anger\_Treatment, the formula is that

P(heart\_attack)=

For Patient A, the probability to get a heart attack is

=0.1082

For Patient B, the probability to get a heart attack is

=0.8356

* 1. What are the odds for patient A and patient B?

Odds(patient A)=0.1082/（1-0.1082）=0.1213

Odds(patient B)=0.7310/(1-0.7310)=5.0829

* 1. What is the odds ratio of A over B?

Odds ratio=Odds(A)/Odds(B)=0.0446

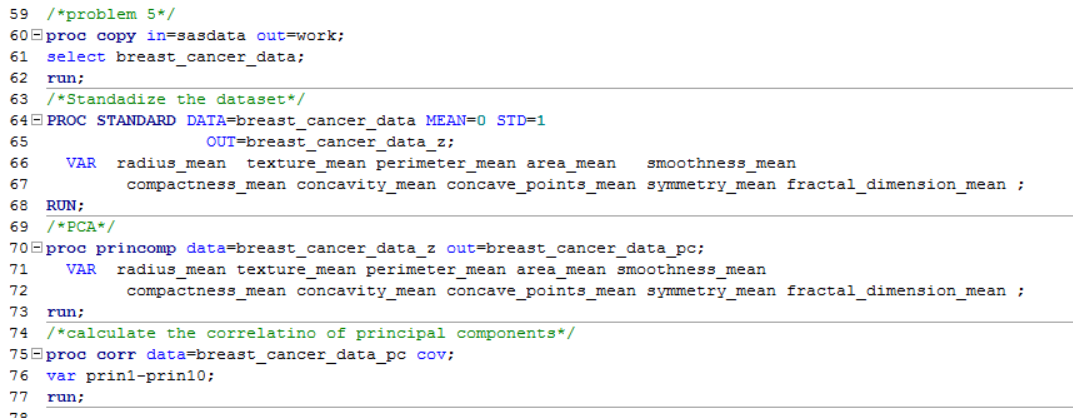
**Problem #5: (20 points)**

The Breast Cancer dataset in CANVAS includes some of the features that are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei in the image. (Source: UCI). Perform PCA analysis on the following 10 variables.

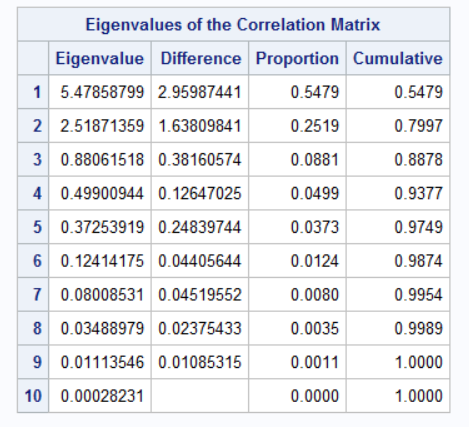
1. radius\_mean
2. texture\_mean
3. perimeter\_mean
4. area\_mean
5. smoothness\_mean
6. compactness\_mean
7. concavity\_mean
8. concave\_points\_mean
9. symmetry\_mean
10. fractal\_dimension\_mean
11. How many principal components should be used to explain at least 85 percent of the variability in data?

**Solution:**

Implement the sas code



we could see from the result that



So we only need ***3*** principal components to explain at least 85 percent of the variability in the data

1. What if the study requires more than 95 percent of variability to be explained, how many variables do you use?

Solution: From the plot above, we find that we only need ***5*** principal components to explain at least 95 percent variability in the data

**Problem #6:**

**See the SAS program**