**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.**

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

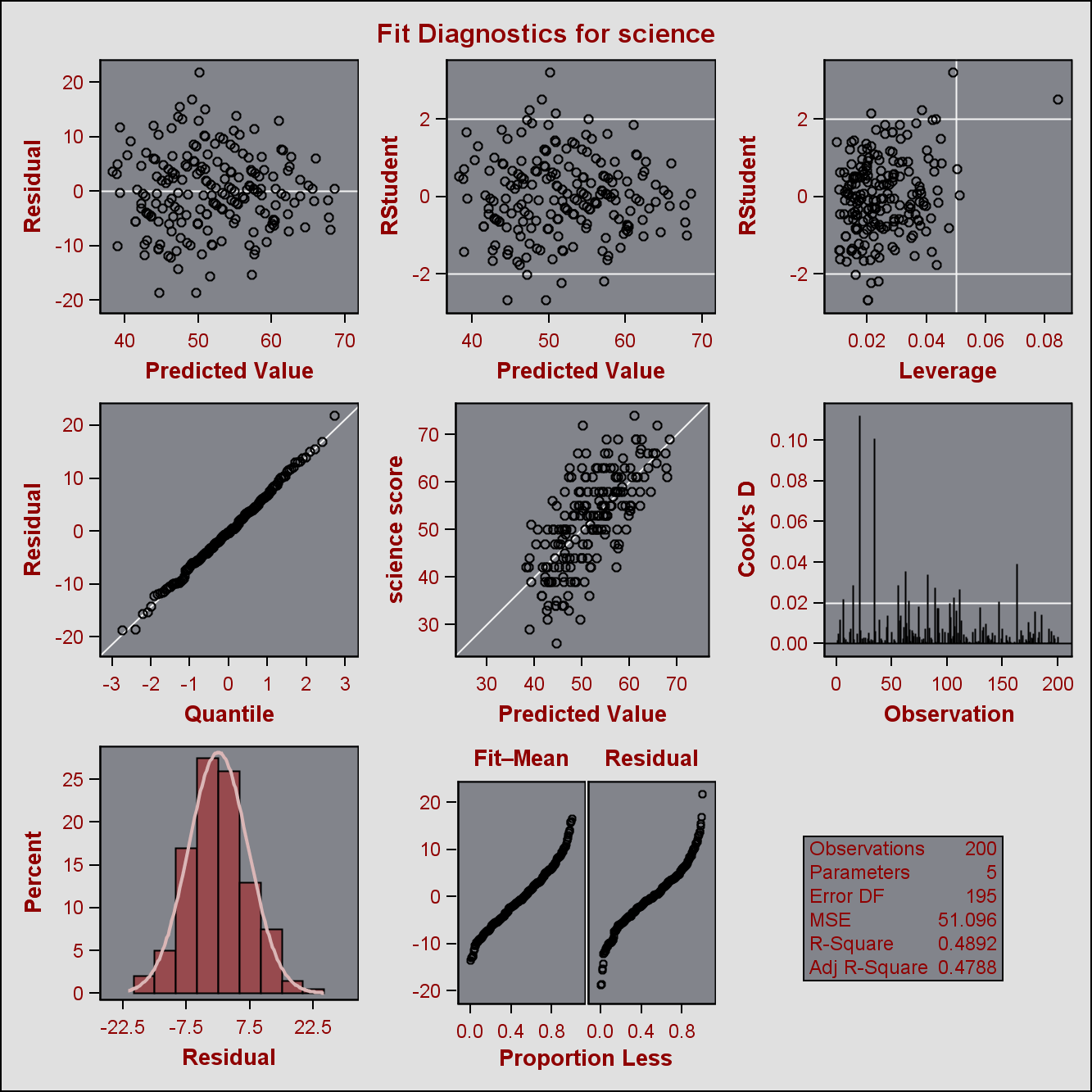
X number of high school students are scored on various tests, such as science, math, and social studies (**socst**). The variable **female** is a dichotomous variable, coded 1 if the student was female and 0 if male. Using the multiple regression analysis results below, answer the following questions:

* How many students were scored?
* Is the overall model significant?
* What is the F-value (1-?)?
* What is the R-square for this model (2-?)?
* What is the formula for this model?
* Is this a good model? Why or why not?
* Would you change the model? If yes, How?

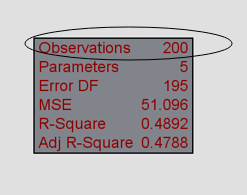
| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 4 | 9543.72074 | 2385.93019 | **1-?** | <.0001 |
| **Error** | 195 | 9963.77926 | 51.09630 |  |  |
| **Corrected Total**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

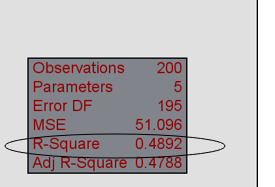


**Solution:**

1.200 

2.Yes

3.46.695=2385.93019/51.09630

4.0.4892

5.Science Score=Math Female Socst Read

6.Not good. Because the P value of socst and female are over 0.05, so they are not significant.

And the Adj R-Square is not high.

7.Yes. Remove the socst from this model and do regression again.

And then adjust the newer model.

**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 I and III

**Solution:**

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).
2. Find the best subset of the three variables

**Solution:**

1. libname sasdata "E:\SASDATA\";

**proc** **copy** in=Sasdata out=work;

select Lung;

**run**;

title "Multiple Regression for the cereal dataset";

**proc** **reg** data=Lung;

model Height\_oldest\_child=Age\_oldest\_child Weight\_oldest\_child Height\_mother Weight\_mother Height\_father Weight\_father/dwProb stb selection=forward ;

**quit**;

model Height\_oldest\_child=Age\_oldest\_child Weight\_oldest\_child Height\_mother Weight\_mother Height\_father Weight\_father/dwProb stb selection=backward ;

**quit**;

model Height\_oldest\_child=Age\_oldest\_child Weight\_oldest\_child Height\_mother Weight\_mother Height\_father Weight\_father/dwProb stb selection=stepwise ;

**quit**;

2. **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)
2. 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 |

* 1. What are the odds for patient A and patient B?
  2. What is the odds ratio of A over B?

**Solution:**

**1.**

libname sasdata "E:\SASDATA\";

**data** heart\_attack2;

set sasdata.heart\_attack;

if Anxiety\_Treatment>**60** and anger\_treatment=**1** then v\_Anxiety\_Treatment=**1**;

else v\_Anxiety\_Treatment=**0**;

**run**;

**proc** **logistic** data=heart\_attack2 descending outest=betas covout;

class Anger\_Treatment(ref='1') v\_Anxiety\_Treatment(ref='1')/param=ref;

model Heart\_Attack\_2=Anger\_Treatment v\_Anxiety\_Treatment / selection=backward

slentry=**0.3**

slstay=**0.35**

details

expb

lackfit;

output out=pred p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

**run**;

**quit**;

**proc** **print** data=betas;

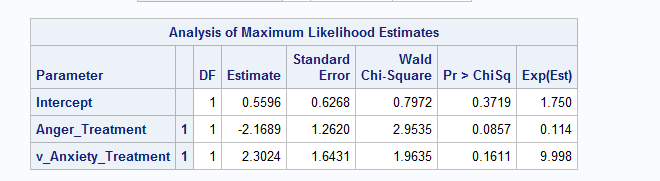
title2 'Parameter Estimates and Covariance Matrix';

**run**;

**proc** **print** data=pred;

title2 'Predicted Probabilities and 95% Confidence Limits';

**run**;



**2.1**

**odds=y/(1-y)=e^(b0+b1x1+b2x2+....)**

For A: the v\_Anxiety\_Treatment=0 Anger\_Treatment=0

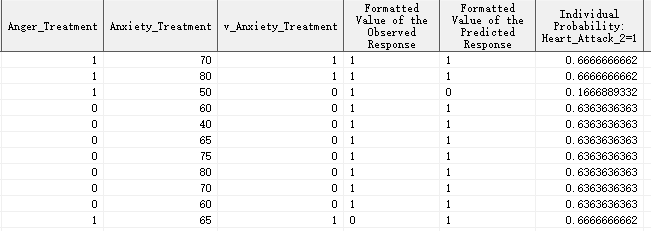
For B: the v\_Anxiety\_Treatment=1 Anger\_Treatment=1

So y=odds/(1+odds)

P(A)= 0.6363599829490800384578858691773

P(B)= 0.66665618201578221572411915614327

But there’s error between the calculation. So I would rather use the result from the SAS



So P(A)= 0.6363636363

P(B)= 0.6666666662

**2.2**

Odds=e^(0.5596-2.1689\*Anger\_Treatment+2.3024\*v\_Anxiety\_Treatment)

Odds(A)= 1.74997237133

Odds(B)= 1.99990564111

**2.3**

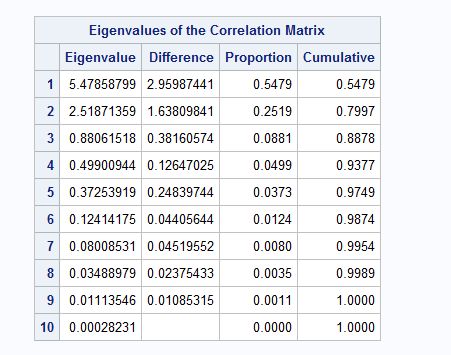
Odds=0.24993326978

**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. How many principal components should be used to explain at least 85 percent of the variability in data?
2. What if the study requires more than 95 percent of variability to be explained, how many variables do you use?
3. radius\_mean
4. texture\_mean
5. perimeter\_mean
6. area\_mean
7. smoothness\_mean
8. compactness\_mean
9. concavity\_mean
10. concave\_points\_mean
11. symmetry\_mean
12. fractal\_dimension\_mean

**Solution:**



1. **3**
2. **5**

**Problem #6: (20 points)**

Assuming the following web structure, calculate the page rank of nodes A through F.

**Solution:**