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?
  Answer: 200. Because the corrected total is 199,so the there are 199+1 students were scored, that is 200.

- Is the overall model significant?
  Answer: Yes. Because the p-value of F-test is less than 0.05.

- What is the F-value (1-?)?
  Answer: F-value =MSR/MSE = 2385.93019/51.09630 = 46.69477

- What is the R-square for this model (2-?)?
  Answer: R-square = SSR/SST = 9543.72074/19508 =0.4892

- What is the formula for this model?
  Answer: Y=12.325 + 0.389*math -2.009*female + 0.049*socst + 0.335*read

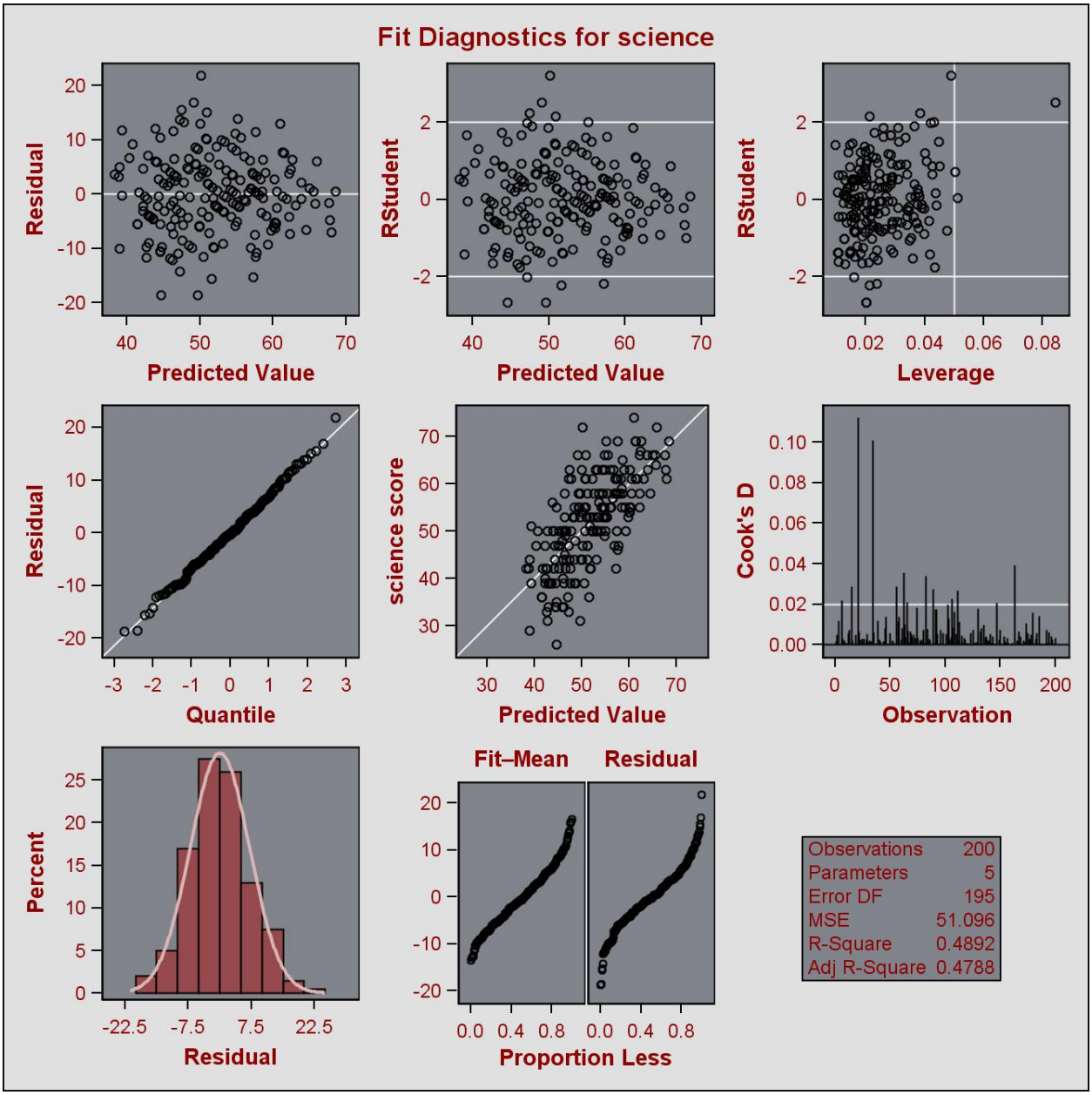
- Is this a good model? Why or why not?
  Answer: No, because p-value for female and socst are over 0.05, they are not significant.

- Would you change the model? If yes, How?
  Answer: Yes. Remove variable socst and run the model again.

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

I. The model is NOT a good model because variable shelf2 and "Intercept" are not significant at 5%
II. The model is NOT a good model because variable shelf1 is not significant at 5%
III. 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.
IV. Both a and c

Answer: B

Problem #3: (20 points)

I. 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).

Answer:

Variable Weight_father Entered: R-Square = 0.9207 and C(p) = 7.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	6583.83851	1097.30642	276.82	<.0001
Error	143	566.85482	3.96402		
Corrected Total	149	7150.69333			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.10417	5.37537	0.00149	0.00	0.9846
Age_oldest_child	1.19844	0.09786	594.50331	149.97	<.0001
Weight_oldest_child	0.07914	0.00836	355.57696	89.70	<.0001
Height_mother	0.30516	0.07431	66.85085	16.86	<.0001
Weight_mother	-0.01133	0.00582	14.99668	3.78	0.0537
Height_father	0.29605	0.07366	64.02408	16.15	<.0001
Weight_father	-0.00917	0.00845	4.66901	1.18	0.2796

During forward selection procedure,p-value of variables Age_oldest_child, Weight_oldest_child,height_father and Height_mother meet the 0.05000 significance level, so they were included in the model.

So Model **A** is: *y* = 0.104 +*1.198**(Age_oldest_child) +0.079*(*Weight_oldest_child*) +0.305*(height_mother)+0.296*(height_father)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6579.16950	1315.83390	331.53	<.0001
Error	144	571.52383	3.96892		
Corrected Total	149	7150.69333			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.52776	5.36450	0.03841	0.01	0.9218
Age_oldest_child	1.21129	0.09720	616.35434	155.30	<.0001
Weight_oldest_child	0.07769	0.00825	351.65930	88.60	<.0001
Height_mother	0.32174	0.07276	77.59926	19.55	<.0001
Weight_mother	-0.01282	0.00566	20.36365	5.13	0.0250
Height_father	0.25345	0.06237	65.53135	16.51	<.0001

During backward procedure, p-value of variable Weight_father is over 0.05000 significance level, so it was dropped from the model.

So Model **B** is: *y* = 0.528 +*1.211**(Age_oldest_child) +0.078*(*Weight_oldest_child*) +0.322*(height_mother) -0.013*(Weight_mother)+0.253*(height_father)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6579.16950	1315.83390	331.53	<.0001
Error	144	571.52383	3.96892		
Corrected Total	149	7150.69333			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.52776	5.36450	0.03841	0.01	0.9218
Age_oldest_child	1.21129	0.09720	616.35434	155.30	<.0001
Weight_oldest_child	0.07769	0.00825	351.65930	88.60	<.0001
Height_mother	0.32174	0.07276	77.59926	19.55	<.0001
Weight_mother	-0.01282	0.00566	20.36365	5.13	0.0250
Height_father	0.25345	0.06237	65.53135	16.51	<.0001

Stepwise procedure has the same result with backward procedure, so the model is same is model **B,** that is *y* = 0.528 +*1.211**(Age_oldest_child) +0.078*(*Weight_oldest_child*) +0.322*(height_mother) -0.013*(Weight_mother)+0.253*(height_father)

II. Find the best subset of the three variables

Variable Height_father Entered: R-Square = 0.9088 and C(p) = 22.5495

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6498.41596	2166.13865	484.85	<.0001
Error	146	652.27738	4.46765		
Corrected Total	149	7150.69333			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	14.03591	4.50096	43.44601	9.72	0.0022
Age_oldest_child	1.21966	0.10259	631.52269	141.35	<.0001
Weight_oldest_child	0.07545	0.00864	340.98721	76.32	<.0001
Height_father	0.33073	0.06348	121.28277	27.15	<.0001

Bounds on condition number: 3.6082, 24.657

The above model is the best 3-variable model found.

Answer: according to the best subset procedure, the best subset of the three variables is Y=14.036 + 1.219*(Age_oldest_child) + 0.075*(Weight_oldest_child) + 0.331*(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.

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

Answer:

```
proc logistic data=heart_attack descending;
class anger_treatment(ref='0') / param=ref;
model heart_attack_2 = anger_treatment anxiety_treatment;
quit;
```

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-6.3634	3.2139	3.9203	0.0477
Anger_Treatment	1	1	-1.0241	1.1711	0.7647	0.3818
Anxiety_Treatment		1	0.1190	0.0550	4.6884	0.0304

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Anger_Treatment 1 vs 0	0.359	0.036	3.565
Anxiety_Treatment	1.126	1.011	1.255

According to the result, the p-value of anger_treatment is not significant, so omit this variable .

```
proc logistic data=heart_attack descending;
class anger_treatment(ref='0') / param=ref;
model heart_attack_2 = anxiety_treatment;
quit;
```

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.0925	3.1710	5.0027	0.0253
Anxiety_Treatment	1	0.1246	0.0553	5.0791	0.0242

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Anxiety_Treatment	1.133	1.016	1.262

The remaining variable is considered significant and retained in the model.'

So logit = -7.0925+ 0.1246* anxiety_treatment.

We may estimate the probability that a particular person will have a second heart_attack given values for the predictor variables.

II. Using your model:

a. 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

b. What are the odds for patient A and patient B?

c. What is the odds ratio of A over B?

Answer:

From the model:

logit = -7.0925+ 0.1246* anxiety_treatment

a. For A: g(x)= -7.0925 + 0.1246*40 = -2.1085

P(heart_attack_2|A)=$e^{(-2.1085)}$/ (1+$e^{(-2.1085)}$) = 0.108

For B: g(x)= -7.0925 + 0.1246*70 = 1.6295

P(heart_attack_2|A)=$e^{(1.6295)}$ / (1+$e^{(1.6295)}$) = 0.836

b. For A: odds=$e^{(-2.1085)}$ =0.121

For B: odds=$e^{(1.6295)}$ =5.101

c. The odds ratio of A over B : 0.121/5.101 =0.024

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

i. How many principal components should be used to explain at least 85 percent of the variability in data?

Answer: According to the development result, we can easily find there are 3 principal components should be used to explain at least 85 percent of the variability in data.

Eigenvalues of the Correlation Matrix

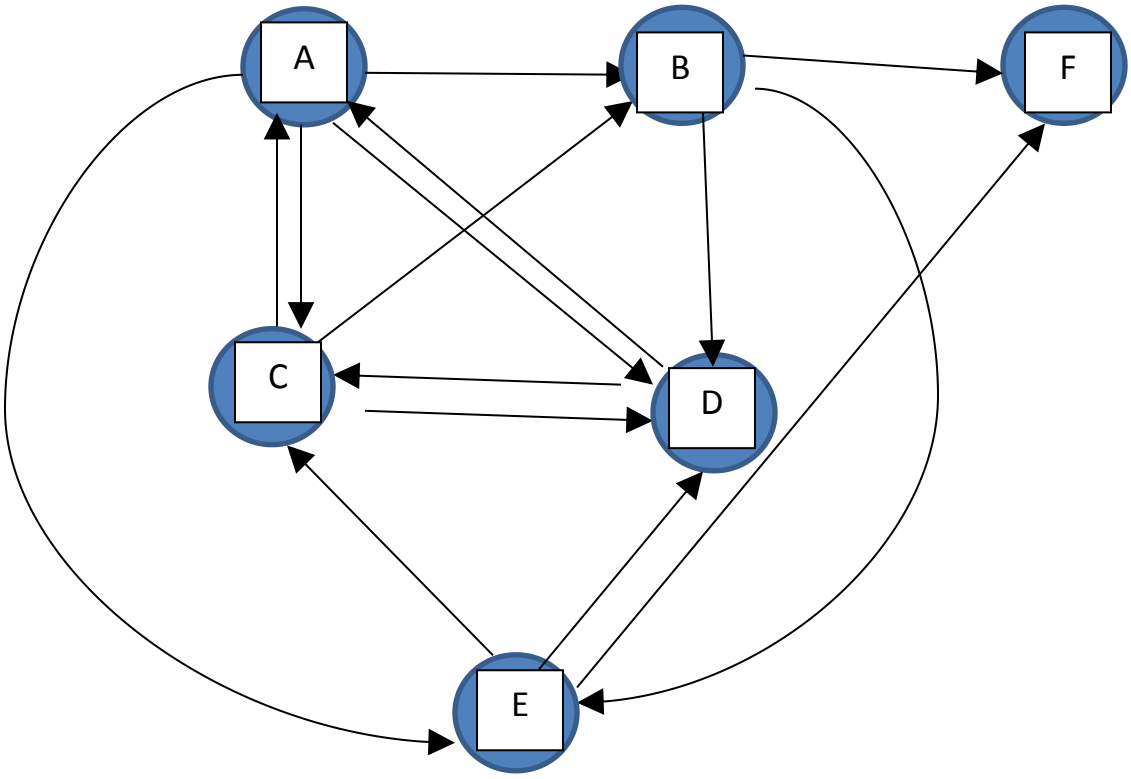
	Eigenvalue	Difference	Proportion	Cumulative
1	5.47858799	2.95987441	0.5479	0.5479
2	2.51871359	1.63809841	0.2519	0.7997
3	0.88061518	0.38160574	0.0881	0.8878
4	0.49900944	0.12647025	0.0499	0.9377
5	0.37253919	0.24839744	0.0373	0.9749
6	0.12414175	0.04405644	0.0124	0.9874
7	0.08008531	0.04519552	0.0080	0.9954
8	0.03488979	0.02375433	0.0035	0.9989
9	0.01113546	0.01085315	0.0011	1.0000
10	0.00028231		0.0000	1.0000

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

Answer: According to the development result, I will use 5 variables.

Problem #6: (20 points)

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



Answer: Because node F is a dead node, so I change it so that it can randomly point to the other nodes.

```
data Arcs;
    infile datalines;
    input Node $ A B C D E F;
    datalines;
A   0   0   1   1   0   1
B   1   0   1   0   0   1
C   1   0   0   1   1   1
D   1   1   1   0   1   1
E   1   1   0   0   0   1
F   0   1   0   0   1   1
;
run;
```

After 50 times iteration, the vector shows little changes at each round.

So the page rank is D > C > A > B > E > F

rank_p50
0.2018666
0.1399931
0.217767
0.2246803
0.1140684
0.1016246

Optional:

1. I am able to define, describe, and clearly state the objectives of data mining specially regression, logistic regression, big data page rank	Strongly Agree √	Agree	Neutral	Disagree	Strongly Disagree
2. I am able to identify relevant data and corresponding databases and data warehouses in SAS	Strongly Agree √	Agree	Neutral	Disagree	Strongly Disagree
3. I am able to describe how to access relevant data.	Strongly Agree	Agree √	Neutral	Disagree	Strongly Disagree
4. I am able to preprocess the data (clean, integrate, transform) in SAS.	Strongly Agree √	Agree	Neutral	Disagree	Strongly Disagree
5. I am able to specify the proper algorithm(s) and data mining technique(s).	Strongly Agree √	Agree	Neutral	Disagree	Strongly Disagree
6. I am able to identify and or develop SAS code to execute the specified algorithm(s)/data mining technique(s).	Strongly Agree √	Agree	Neutral	Disagree	Strongly Disagree
7. I am able to mine and discover models, patterns, dependencies that will enable predictions, make intelligent business and operation decisions, learn and extract nuggets of knowledge.	Strongly Agree √	Agree	Neutral	Disagree	Strongly Disagree
8. I am able to present and document results.	Strongly Agree √	Agree	Neutral	Disagree	Strongly Disagree
9. I am able to input the extracted knowledge to the next iterative steps.	Strongly Agree	Agree √	Neutral	Disagree	Strongly Disagree