**Problem #1:**

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? **200**
* Is the overall model significant? **Yes**
* What is the F-value (1-?)? **46.69**
* What is the R-square for this model (2-?)? **0.4892**
* What is the formula for this model?

**12.32529+.3893102\*math+2.009765\*female+0498443\*socst+.3352998\*read**

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

**No, On observing the residuals they do not follow a pattern. The percent Vs residual graph follows normal distribution. The F-value is also significant. But on considering our level of significance to 5 % except socst all other variables are significant.**

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

**Yes, we can perform Max-R and select the most best/significant variables and then run my model**.

**Problem #2:**

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

**Problem #3:**

**Please check attached code file Question\_3.sas**

**Problem #4:**

**Please check attached code file Question\_4.sas**

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 | 1 | 40 |
| B | 0 | 70 |

**log(odds)= -6.3634-1.0241\*Anger\_Treatment+0.1190\*Anxiety\_Treatment**

**Patient A :-**

**Anger\_Treatment=1 and Anger\_Anxiety=40**

**log(odds)= -6.3634-1.0241+0.1190\*40= -7.3875+4.76= -2.6275**

**Odds for patient A= e^(-2.6275)=0.07225**

**Probability=odds/1+odds= 0.067389 =6.7%**

**Patient B:-**

**Anger\_Treatment=0 Anxiety\_Treatment=70**

**log(odds)= -6.3634+0.1190\*70 = 1.9666**

**odds= e^(1.9666) = 7.14633**

**Probability= odds/1+odds=0.8772 = 87.72%**

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

**odds for patient A - 0.07225**

**odds for patient B - 7.14633**

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

**Odd Ratio =0.07225/7.14633**

=**0.0101100844769273**

**Problem #5:**

**Please check attached code file Question\_5.sas**

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

**3 Principal Components.**

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

**5 Principal Components**

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

**Problem #6:**

**Please check attached code file Question\_6.sas**