Data Mining Student Assignments

Assignment #1: (20 points)   Due Date: _____
Based on Lesson 1

Examine the four remaining variables with Risk to determine if any one of them has a stronger linear relationship with Risk. You are to submit a copy of the scatterplots and the four resulting linear regression models, including the corresponding correlation coefficient. Choose the prediction model that you believe gives the strongest linear relationship with Risk and explain why. Support your choice with an analysis of the scatterplot, correlation coefficient, and the residual plot.

The entire assignment should be typed. Hand-written assignments will not be accepted.

Rubric

Model	Points Possible	Points Earned
Risk vs Age	4	
Risk vs Lab	4	
Risk vs Xray	4	
Risk vs Beds	4	
Explanation of best choice of model	4	
Total	20	

Assignment#2 (30 points)   Due Date: _____
Based on Lesson 2

Load the *consumer.dat* file from the AP Statistics folder on the Shared Directory.

The data in the table are the average annual prices of various market basket goods that are used in the determination of the Consumer Price Index. Create an appropriate model to predict the price of beef.

You have been assigned to one of 3 teams: Team Forward, Team Backward and Team Step-wise. Each team will complete their corresponding variable selection process and post the results to the class Wiki page in Blackboard. The team posting should include the output from the selection process in R and a residual plot of the dominant explanatory variable to the residuals. (10 points for completion)

Each member of each team is required to post and contribute to the discussion about the two remaining prediction models. Compare the goodness of fit of the three models by examining R^2 and the residual plot of the dominant explanatory variable to the residuals. Determine which model is the most appropriate in your discussions. (5 points for completion for each discussion = total is 10 points)