**Assignment Part I**

You will complete this part of the assignment using SPSS Statistics.

You have been provided a data set, HELP, in Excel format (posted on BB under HW3) that includes scores on factors pertaining to mental health of a person. The descriptions for the variables in the dataset are as following:

|  |  |
| --- | --- |
| **Variable** | **Label** |
| Age | Age at baseline (in years) |
| Female | Gender of the respondent, 0= Male, 1=Female |
| PSS\_FR | Perceived Social Support from Friends |
| Homeless | One or more nights on the street or shelter in past 6 months, 0= Not Homeless, 1= Homeless |
| PCS | Physical Health Composite Score-Baseline for a person |
| MCS | Mental Health Composite Score – Baseline for a person |
| CESD | Total Score, Baseline for a person. |

For Part I of the homework, you will be looking at mental health for the subjects for whom details are provided in the dataset, HELP. First, you will be running a model to look at the continuous measure “MCS” which is the mental component score of the SF-36 quality of life instrument/questionnaire. The MCS values range from 0 to 100 where the population norm for “normal mental health quality of life” is considered to be a 50. If you score higher than 50 on the MCS you have mental health better than the population norm and vice versa - if your MCS scores are less than 50 then your mental health is considered to be worse than the population norm.

Learn more about MCS at [MCS](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5029523/)

Here is a list of tasks that you need to perform using SPSS Statistics for HELP dataset.

1. Run a simple linear regression for MCS variable (using it as a dependent variable) using the CESD variable as predictor variable (which is a more specific measure for depression). **Write the equation of the final fitted model (i.e. what is the intercept and the slope)? Write a sentence describing the model results (interpret the intercept and slope).** Copy/paste Coefficient Table from SPSS Statistics output below.

[INSERT RELEVANT OUTPUT HERE]

2. How much variability in MCS variable does the CESD variable explain? (What is the value of R2?) Write a sentence describing how well the CESD variable does in predicting the variable MCS?

3. Run a second linear regression model for the MCS variable putting in all of the other variables in the data subset as predictor variables:

* Age
* Female
* PSS\_FR
* Homeless
* PCS
* CESD

Copy/Paste the model results with the coefficients and tests and model fit statistics below.

[INSERT RELEVANT OUTPUT HERE]

4. Which variables are significant in the model? Write a sentence or two describing the impact of these variables for predicting mental component scores. Run the VIFs to check for multi-collinearity issues. Identify the variables with multi-collinearity issues. Copy/Paste the output with VIF values for each variable.

[INSERT RELEVANT OUTPUT HERE]

5. Remove the variables which are not significant in the model in part 4 above and those which VIF more than 4 and run it again. Copy/Paste the model results with the coefficients and tests and model fit statistics below. Which variables are significant now?

[INSERT RELEVANT OUTPUT HERE]

6. Which model (from part 1 and 2 or the one from part 3 and 4 or part 5) is better to explain the dependent variable, MCS? Write your answer and explain why you think this model is better than others?

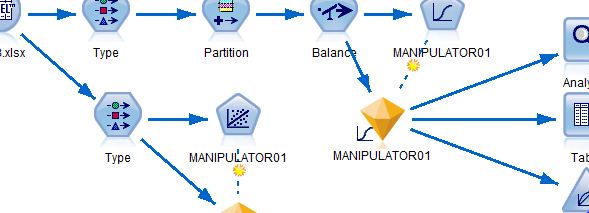
**Assignment Part II**

***Predicting Earnings Manipulation by Firms***

Earning manipulations involve deliberate steps by companies to bring reported earnings to a desired level. Some of the banks extending loans to companies suspect that these companies are manipulating their earnings to secure the loans. Your task is to use the eight financial indices described below to predict the earning manipulators using multiple regression and logistic regression based on SPSS Modeler 18.2. The Excel provided for this assignment, Manipulator\_Firms\_Data.xls, contains data on 1,239 firms where the outcome variable of interest, *C\_Manipulator* (1/0) is known.

**After completing all your analyses**, provide a picture of your stream showing all the nodes by replacing the image in Figure 1 below with a screenshot of your stream (NOTE: Image below may not show all the nodes you need to include in your stream).

**Figure 1.**



**Data Preparation**

1. **Select and prepare Excel data (same as for Assignment 2)**

Change the data format from “General” to “Number” for the variables *DSRI, GMI, AQI, SGI, DEPI, SGAI, ACCR,* and *LEVI.* These variables should display at 4 - 5 decimals. Use this newly created file as your source file. **DO NOT SORT THIS DATASET DIFFERENTLY.**

1. **Properly configure the Type node**.
   1. Make sure to “Read Values” first.
   2. Select the appropriate measurement for the 8 input variables listed under 1 above.
   3. The dependent variable is *C\_MANIPULATOR (*with values 1/0*)*.
2. **Partition the data set (same as for HW2)**

Attach a **Partition node** to the Type node using 50% of the data for training and 50% for testing.

1. **Balance the data set (same as for HW2)**

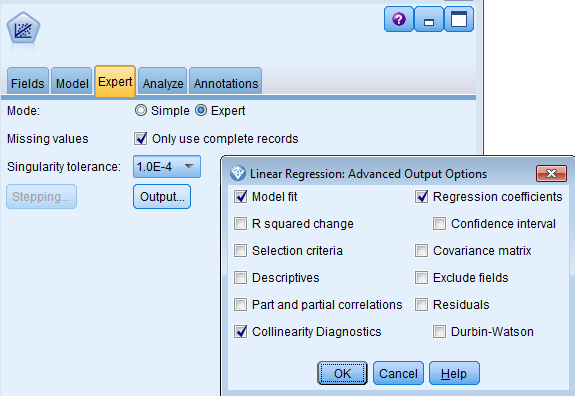
Attach a **Balance node** to the right of the Partition node to allow for oversampling, i.e., duplication of companies in the minority class (the manipulators). For purposes of your analysis, choose a factor of 6 for the condition *Manipulator* = Yes. Make sure to check the “Only balance training data” box.

**Assess Multicollinearity**

1. Perform linear regression to determine if multicollinearity exists among the independent variables.
   1. Attach **another Type node** to the source node. Set the measurement of the *C\_MANIPULATOR* target variable to Continuous. Attach a **Regression node** to this type node. Under the *Expert* tab, choose *Output* to launch the *Advanced Output Options*. Make sure to check *Collinearity Diagnostics* as shown below in Figure 2, then run the node.
   2. Inspect the VIF for each variable. **Is multicollinearity a problem? Why or why not?**

[INSERT RELEVANT OUTPUT HERE - VIF AND TOLERANCE FOR EACH INDEPENDENT VARIABLE]

**Figure 2. Regression Node Configuration**



**Logistic Regression Modeling**

1. To run a **binomial logistic** regression, add a **Logistic node** to your stream. Under the Model tab, make sure you choose the binomial procedure, and use partitioned data.
   1. Provide the **Model Summary** table with goodness of fit indices (Cox & Snell R2, Nagelkerke R2).

[INSERT RELEVANT OUTPUT HERE]

* 1. Provide the final **Variables in the Equation** table with the regression coefficients that is shown at the very bottom of the output.

[INSERT RELEVANT OUTPUT HERE]

1. Attach an **Analysis node** to the data mining nugget, and provide the **coincidence matrices** for Training and Testing.

[INSERT RELEVANT OUTPUT HERE]

1. Attach an **Evaluation node** to the data mining nugget and to create a **Gains chart**, which should be configured with the **Include best line** and **Split by partition** options. Provide the gains chart for Training and Testing partitions. Provide your interpretation of Gains chart i.e. what do you learn from the Gains chart?

[INSERT RELEVANT OUTPUT HERE]

1. Attach another **Evaluation node** to create a **Lift chart**, which should be configured with the **Include best line** and **Split by partition** options. Provide the lift chart for the Training and Testing partitions. Provide your interpretation of Lift chart i.e. what do you learn from the Lift chart?

[INSERT RELEVANT OUTPUT HERE]

1. Attach a **Table node** to the resulting data mining nugget and run it. Using the output from the Table node, provide the information requested below for the company from the Testing Partition that is most likely to be an earnings manipulator. HINT: copy data from the Table node into Excel, filter on Partition, and sort appropriately on *$LP-1*. There could be multiple companies with same probability.

|  |  |
| --- | --- |
| * 1. Company ID |  |
| * 1. Probability of being an earnings manipulator |  |
| * 1. Did the company actually manipulate earnings? |  |

**Model Evaluation**

1. Using the output from 6. – 9., provide the following metrics and details. Note that accuracy metrics and gains/lift should be based on the Testing partition.

|  |  |
| --- | --- |
| * 1. Overall Accuracy |  |
| * 1. Sensitivity (show calculations) |  |
| * 1. False positive rate (show calculations) |  |
| * 1. Gains at the 20th the percentile |  |
| * 1. Lift at the 20th percentile |  |
| * 1. List variables that are NOT significant predictors of earnings manipulations |  |
| * 1. List statistically *significant risk variables* that *increase the odds of being classified as being an earnings manipulator* |  |

1. Based on the output from 6. – 9. and details you provided in 11., evaluate the logistic regression model. How well does the model fit the data? Does the model perform well? Is it an improvement over random guessing? Make sure you support your answer with specifics from the output you generated and your calculations.