**CSC 435 LAB 04 Points: 20**

**Correlation, Simple & Multiple Regression**

Your Team Member names:

1. Padraic Darby

2. Stryder Coleman

Dr. John, recently elected as the chair of Rochester Hospital, has hired your consulting firm to help him develop some statistical models that can be used for monitoring hospital performance. He would like to know what variables are most relevant in predicting hospital expenses, so that hospitals may be monitored on an ongoing basis.

You have been asked to develop a model to predict total expenditures for hospitals and to identify the significant variables. You have decided to use regression model to accomplish this task. You have access to a hospital database that consists of several variables that are listed in Table 1.

**Purpose:**

This lab introduces students to Correlation, Simple Regression & Multiple Regression concepts. After completing tasks in this lab, you should be able to:

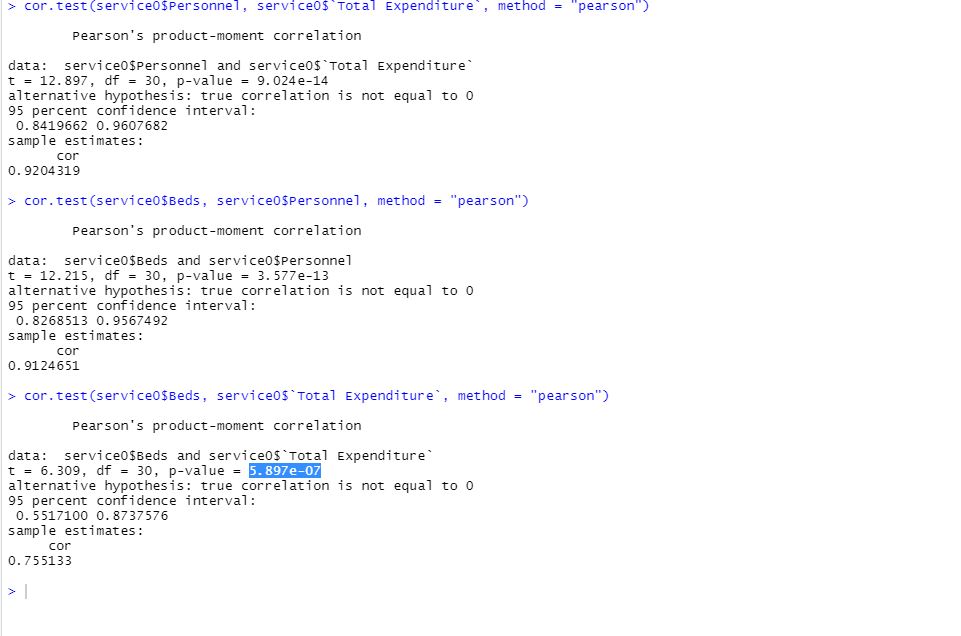
* Perform correlation analysis & analyze results to discover how the variables are related.
* Perform a simple linear regression analysis & interpret the results such as such as p-value, R2, residuals, and regression equation.
* Perform a Multiple Regression Analysis and interpret the results such as p-value, adjusted R2,VIF, and regression equation.
* Predict dependent variable based on given values for independent variables by using the regression equation.
* Table 1 - Details of the variables in the data file (Meta Data).

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Details** | **Value/(Units)** | **Data Type** |
| A | Service | 0 - psychiatric hospital 1 - general medical hospital | Ordinal |
| B | Beds | Number of beds | Continuous |
| C | Admissions | Number of admissions per year | Continuous |
| D | Outpatient Visits | Number of outpatient visits per year | Continuous |
| E | Births | Number of births per year | Continuous |
| F | Personnel | Total number of personnel | Continuous |
| G | Total Expenditure | Annual expenditures in $1,000 | Continuous |

**Lab Instructions:**

1. Fill in the Data Types (Nominal, Ordinal, Continuous) in Table1
2. Verify the data types of the variables in Table1. Make changes to the R variables if necessary.
3. Investigate whether correlation exist between all continuous variables by **Service Level**. **Consider all possible pairwise correlations by service level**. Identify the correlation coefficients, and check whether the correlation coefficients are significant (Alpha = 0.05) or not. Use the Table below to list the three **highest** correlation coefficients for the two service levels that are significant. Are there any differences between the two? Report a screenshot of your test result below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Service Level- 0(psychiatric hospital)** | | | |  |
| **Variable 1** | **Variable 2** | **Correlation Coefficient** | **Interpretation in non-statistical term** | **Significance** |
| Personnel | Total Expenditure | 0.92 | High Positive Correlation | 9.024e-14 |
| Beds | Personnel | 0.91 | High Positive Correlation | 3.577e-13 |
| Beds | Total Expenditure | 0.76 | High Positive Correlation | 5.897e-07 |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Service Level- 1(general medical hospital)** | | | |  |
| **Variable 1** | **Variable 2** | **Correlation Coefficient** | **Interpretation in non statistical term** | **Significance** |
| Personnel | Total Expenditure | 0.97 | High Positive Correlation | 2.2e-16 |
| Admissions | Total Expenditure | 0.91 | High Positive Correlation | 2.2e-16 |
| Admissions | Personnel | 0.90 | High Positive Correlation | 2.2e-16 |



1. Pick the independent variable (**Personnel)** and perform a Simple Linear Regression analysis with the dependent variable (**Total Expenditure**). State the Null and Alternate Hypotheses; interpret the Output, and the regression equation.

**Hypothesis:**

Ho: Personnel is not significant

Ha: Personnel is significant

**Output Interpretation:**

1. P-value and your conclusion based on this: (report the F-statistic result below)

2.2e-16

We reject the null hypotheses

1. Find the confidence interval of the Personnel coefficient. Report your result below.

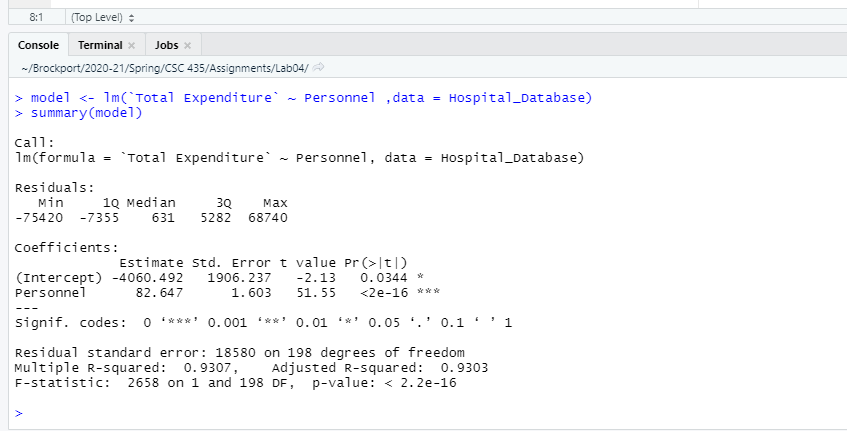
B1 = 82.647

SE = 1.603

2SE = 3.206

79.441 < B1 < 85.853

1. Calculate the percentage error of the model based on the model sigma (RSE). Report a screenshot of your code and the result below.



1. R2 value and your interpretation:

R2 = 0.9307

93.07% of the expenditure data can be explained by the model

1. Write and interpret the regression equation:

Total Expenditure = -4060.492 + 86.647x

x-personnel number

The line has a positive slope of 92.647 and a y-intercept of -4060.492

1. Now run a multiple regression analysis with **Total Expenditure** as dependent variable and all other variables as independent variables.The final model should have only significant independent variables. State the statistical hypotheses; interpret adjusted R-Square, multicollinearity (VIF), andinterpret regression equation (Prediction formula).

**Hypothesis for the overall model:**

Ho: Every is coefficient is equal to zero

Ha: At least one of the coefficients is not equal to zero

**Hypothesis for the independent variables:**

Ho:

Personnel = 0

Beds = 0

Admissions = 0

Outpatient Visits = 0

Births = 0

Ha:

Personnel≠ 0

Beds ≠ 0

Admissions ≠ 0

Outpatient Visits ≠ 0

Births ≠ 0

**Output Interpretations:**

1. P-value for the overall model and its interpretation

P-value = 2.2e-16 indicating at least one of the predictor values is significantly related to the outcome variable.

1. List all significant independent variables and their p-values:

Personnel 63.2398 3.0998 20.401 < 2e-16 \*\*\*

Admissions 3.3035 0.5306 6.226 2.85e-09 \*\*\*

Births -4.5343 2.2019 -2.059 0.040785

1. List and interpret R-Square adjusted:

R-Square adjusted= 0.9436.

There are too many variables being used causing the R-Square and R-Square adjustedto differ slightly. Therefore, some of the variable(s) should be removed.

1. Test Multi-Collinearity interpret the result and report the outcome here.

vif(model2)

Personnel Admissions Births

4.645637 8.905122 3.928589

There is some collinearity between admissions and another variable; therefore, we should remove admissions.

1. **List and interpret the Prediction Formula**:

Text

Description automatically generated

Total Expenditure = -4560.637 + 77.95x +5.202y

x-number of personnel

y-number of births

The line as positive slopes of 77.95 and 5.202 and an intercept of -4560.637.2

1. With the regression equation you derived in previous step, predict the Total Expenses for a hospital where: Service = 1, Beds = 500, Admissions = 10,000, Outpatient visits = 100,000, Births = 1,000, and personnel = 2500?

Total Expenditure = -4560.637 + 77.95(2500) +5.202(1000) = $195,516.363