**STAT 441/541 Regression Project Guidelines (Individual Project)**

**1. Obtain or Create a Data Set**

**2. Requirements**

**Requirement 1**: At least two quantitative independent variables and a quantitative continuous response variable.

**Requirement 2:** The data must be applicable to a regression model situation. That is, the quantitative values for independent variables must be meaningful and not represent categories.

**3. Report Outline**

* 1. **Introduction**
* Introduction
* Objective(s) of Analysis
* Explanation of the dependent and independent variables

**3.2 Exploratory Data Analysis based on summary statistics for all variables**

The objective is to provide an exploratory data analysis of variables in the data set.

This should include:

* A table of numerical summary statistics, number of observations, average, standard deviation, minimum, maximum, etc. with meaningful comments about each variable
* Exploration of bivariate relationships, i.e. scatterplot matrix; correlations and p-values; with meaningful comments

**3.3 Regression Analysis**

**3.3.A Propose and Fit a Regression Model**

1. State the model using correct notation
2. Give the assumptions for regression analysis
3. Fit the model, show the model results from the summary for regression analysis, and give the estimated model with correct notation
4. Interpret all model coefficient(s),

* For all independent variables, interpret coefficients (for a unit increase in xi the change in the mean of y etc.)
* For the intercept, explain whether or not it is meaningful and interpretable

**3.3.B Use the output from 3.3.A to perform hypothesis tests:**

1. Perform an overall *F* test to see if all ’s for partial slopes are simultaneously equal to zero
2. Perform the *t*-test for ***individual*** coefficients

State the hypotheses, use output with *p-values* to perform the tests, and provide the decision and conclusion in the context of your data

**3.3.C** **Residuals and diagnostics**

* Provide diagnostic plots and check all assumptions
* Perform the Shapiro-Wilks Test for normality of the errors. Give hypotheses, test statistic, p-value, decision and conclusion
* Perform the Breusch-Pagan test for constant error variance. Give hypotheses, test statistic, p-value, decision and conclusion
* Comment if the assumptions have been violated or met.

**3.3.D Check for outliers and influential observations.**

Use influence measures to test for outliers and influential observations.

* Which observations are flagged as influential?
* Which observations are flagged for Cook’s distance values greater than one?

Make an assertion as to whether you should include or exclude any potential outliers (you do not need to perform further analysis).

**3.3.E Check for multicollinearity**

Provide VIF values and comment using a threshold value of 10.

**3.3.F Conclusion**

Provide a conclusion for your analysis. As part of your conclusion, explain how you would proceed with your regression analysis if further modeling is needed.

**4. Report**

Format 12pt, 1 inch margins, 1.5 line spacing.

Upload a single pdf file of your report on D2L in the *Multiple Regression Project* folder.

A picture containing text, receipt

Description automatically generated

Call:

lm(formula = yield\_y ~ may\_ndvi + june\_ndvi + july\_ndvi + august\_ndvi,

data = dataobj)

Residuals:

Min 1Q Median 3Q Max

-90.877 -11.556 2.626 12.807 97.090

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 105.685 3.943 26.804 < 2e-16 \*\*\*

may\_ndvi -31.553 7.484 -4.216 2.74e-05 \*\*\*

june\_ndvi -52.215 8.349 -6.254 6.19e-10 \*\*\*

july\_ndvi -35.516 12.405 -2.863 0.00429 \*\*

august\_ndvi 170.641 11.524 14.807 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.62 on 894 degrees of freedom

Multiple R-squared: 0.4461, Adjusted R-squared: 0.4437

F-statistic: 180 on 4 and 894 DF, p-value: < 2.2e-16