**Stat 402 – Applied Regression Homework one** – **Professor Esfandiari**

**Due Monday October 19th by eleven PM**

Needs to be posted on CCLE before the deadline. Pdf files only. You can do this individually or in a group. If you do it as a group, submit one file with the name of all group members

**Problem one.** Using the following campus climate data set (data folder week one), answer the following questions…

1. Create a scatterplot for showing UCLA students’ perception of friendliness on our campus (friendlyenvp), predictor, as a function of their perception of academic satisfaction on our campus (academicenvp), outcome.
2. Now create a scatterplot like the one shown on (same plot as page 21 of lecture one). Remember to install the “car” package. Explain what the lines in this plot show? What do you conclude from the two boxplots?
3. As you see from the two plots you created in parts a and b, there is very little data below friendlyenvp = 40 and below academicenvp = 10. Create a new subset by deleting the above range of data, attach this new data set and draw the scatter plot you drew in part b. Use the following reference as guideline for sub-setting your data. I have also included the hsb2 data in the data folder of week one so you can recreate what you find in the following reference. Comment how everything changed. [https://stats.idre.ucla.edu/r/modules/subsetting-data/](about:blank)
4. Now create two linear models for the prediction of UCLA students’ perception of academics from the friendliness by using: 1) The original campus climate data, and 2) the subset you created. Compare the two models in terms of the slope and R-squared and comment on any changes that happened.
5. Interpret the slope, intercepts, and R-squared for model resulting from the subset of the data you created within context
6. Conduct exploratory data analysis by creating the histograms, qqplot, plot of residuals vs. predictor. One quick way to perform exploratory data analysis is to use the common
7. plot(name of the model) function. This will provide us with the majority of the plots we need.
8. Draw the plot of friendlyenvironmentp^ (Y^) vs residual and vs. academicenvp. Are the different or the same? Explain why they are the same or different?
9. Conduct the ncv test to show that the principle of equality of error variance holds.

**Problem two**.

1. In regression, conceptually speaking what do we mean by the principle of equality of error variance.
2. In regression, mathematically what do we mean by the principle of least squares?
3. In regression plot of residuals vs. X or residuals vs. Y^ serve equally well for checking the principle of equality of error variance. Why is this the case? Explain conceptually and mathematically.
4. Prove that sum of square of total = Sum of square of regression + sum of square of residual. (see answers to review exercise one)
5. Prove that slope and intercept result from placing the derivative of the sum of square of residuals equal to zero.

**Problem three** Two researchers are studying the relationship between math and physics scores. Researcher A finds the covariance to be 700, Researchers B finds covariance to be 900. They both use a sample size of 100.

1. Can we say that researcher B showed a stronger relationship between math and physics scores, yes and no and why? Make your point mathematically and conceptually.
2. Why do they call correlation standardized covariance?

**Problem four**.

1. Using R, calculate the following, using the subset that you created. Variable Mean SD Variance friendlyenvp academicenvp

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Mean | SD | Variance |
| friendlyenvp |  |  |  |
| academicenvp |  |  |  |

1. Using R, calculate the coefficient of correlation and covariance between UCLA students’ perception of academics and friendliness of our environment at UCLA.

Using what you reported in parts a and b, calculate…

1. TSS (total sum of square)
2. SSX 3. RSS (residual sum of square)
3. The standard error of the slope.
4. The t-test of the slope.
5. F test of R-squared.
6. Show that F = t squared – Remember that this holds true only in linear regression