STAT 212-501 (Fall 2022)

Homework 1

A. Do the following two exercises from your textbook *(Akritas, First Edition)*:

(i) Problem 5 (pg. 426 of textbook).

(ii) Problem 6 (pg. 426 of textbook).

(**Note:** for answering part (b) of Problem 5 above, you need to report *both* the least squares regression lines, one fitted to the data on (y, t) and the other on $(\log(y), t)$, and in each case, report the equation of those lines as the 'predictive equation for the bacteria count Y at time t' asked for in the question.)

B. In addition, for Problem 6 in Part **A** (ii) above, do the following exercises:

(i) Fit a least squares regression line to the data with $y = \textit{output}$ and $x = 1/(\textit{wind speed})$.

(ii) Produce *all* relevant plots of the residuals (as discussed in slides 19-20 of Chapter 1A lecture notes) from your fitted linear model in part (i) above and use them to comment on whether the model assumptions (i.e. linearity of the regression curve $E(Y|X)$, equal error variances and normality of the errors ϵ_i's) seem reasonable. (**Note:** you must present *all five plots* discussed in the notes.)

(iii) Find the coefficient of determination R^2 and interpret it.

(iv) Find a 99% confidence interval (CI) for the slope (β_1) of your linear model fitted in part (i) above.

(v) Find a 95% confidence interval for the *average* output when the wind speed is 3.2.

(vi) Suppose the wind speed at a particular windmill is 9.05. Find a 95% prediction interval for the output of this mill, i.e. an interval in which you are 95% sure the output of this windmill will be.

C. Recall the derivation of the least squares regression line shown in slides 12-14 of Chapter 1A notes.

(i) Complete the steps outlined in those slides and show that the slope ($\hat{\beta}_1$) and intercept ($\hat{\beta}_0$) estimates for the least squares regression line are given by:

$$\hat{\beta}_1 \;=\; \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 \;=\; \bar{y} - \hat{\beta}_1 \bar{x},$$

where all notations have their standard meanings as used in the notes. (**Hint:** in the final step, to arrive at the desired expression of $\hat{\beta}_1$ as above, you may need to use the fact that $\sum_{i=1}^n (x_i - \bar{x}) = 0$ and $\sum_{i=1}^n (y_i - \bar{y}) = 0$, both of which follow quite easily from the definitions of $\bar{x}$ and $\bar{y}$.)

(ii) Further, let $\hat{\rho}$ denote the sample correlation coefficient between Y and X (defined formally below), and let $\hat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$ and $\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$. (These notations are also defined in slides 41-43 of Chapter 1A notes.) Then, prove that $\hat{\beta}_1$ from above also satisfies the following form:

$$\hat{\beta}_1 \;=\; \hat{\rho}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}, \quad \text{where} \quad \hat{\rho} \;=\; \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \;\; \hat{\sigma}_Y = \sqrt{\hat{\sigma}_Y^2} \;\text{ and }\; \hat{\sigma}_X = \sqrt{\hat{\sigma}_X^2},$$

with $\hat{\sigma}_Y^2$ and $\hat{\sigma}_X^2$ as above. What does this result tell us about the nature of the relationship between $\hat{\beta}_1$ and $\hat{\rho}$? And also between $\hat{\beta}_1^2$ and $\hat{\rho}^2$ (which is the same as R^2, the coefficient of determination)?

(**Note:** For all proofs above, you *must show* in detail *all your steps* in arriving at the final conclusions.)

General Instructions and Things to Keep in Mind:

1. All homework **submissions must be made online via Canvas.**

2. Your solutions must be uploaded **as a single pdf file** with your **name, course-section and email id clearly printed** on the first page. Your solutions may include a combination of typed pages and/or hand-written documents (properly scanned) and/or `R` codes with outputs (embedding screenshots of these is acceptable). But they **must** be all combined into a single pdf file and submitted in `Canvas`.

3. It is **your responsibility to ensure** that your **uploaded homework solution is complete, clear and fully legible,** especially if there are scans of hand-written documents involved. If not, the TA may be forced to ignore the affected questions and deduct all allotted points!

4. The **deadline is strict.** (No unwarranted exceptions and/or extension requests will be entertained.)

5. For all exercises, you may use a standard scientific calculator or `R/RStudio` for any numerical calculations required. In either case, you must show all relevant intermediate steps to get to the result.

6. For all software implementations via `R/RStudio`, you **must include all the relevant `R` code along with the outputs, and a clear statement of your final answer(s) to the question(s) asked.**

7. For some problems, you may need to use one or more of the tables of critical values for the Normal, t, χ^2, F and Q distributions. These are available in your textbook (Tables A.3-A.7, pg. 495-499).

8. For all exercises, you should **show all your work,** including intermediate calculations and all relevant `R` codes/outputs, as applicable. Otherwise the TA may choose **not** to give you any partial credit.

9. In some cases, the TA may choose to grade a selected set of problems only. Please keep this policy in mind and make every attempt to **answer all questions** and their subparts.

10. Review the syllabus very carefully for **all the guidelines and policies** regarding homework assignments. You are **required to abide by them** strictly. Finally, all homework **grading related questions/concerns must be directly addressed to the TA.**