Exploratory Data Analasis with R

The data file `bea-2006.csv` contains information about the economies of the 366 metropolitan statistical areas" (cities) of the US in 2006. In particular, it lists, for each city, the population, the total value of all goods and services produced for sale in the city that year per person (per capita gross metropolitan product", pcgmp), and the share of economic output coming from four selected industries.

Formulate in clear and concise sentences your hypothesis, reasoning and motivation for any statistical analysis you will implement in R and interpretations of the R outputs. You will present your analysis in a Rmarkdown report that will contain reproducible codes, all necessary graphs and outputs and your complete reasoning. All figures should be clearly labeled and readable.

1. Load the data file and verify that it has 366 rows and 7 columns. Why should it have seven columns, when the paragraph above described only six variables?

2. Calculate summary statistics for the six numerical-valued columns.

3. Make univariate EDA plots for population and for per-capita GMP, and de- scribe their distributions in words. (Use the commands `hist` and `boxplot`.)

4. Make a bivariate EDA plot for per-capita GMP as a function of population. Describe the relationship in words.

5. Using only the functions `mean, var, cov, sum` and `arithmetic`, calculate the slope and intercept of the least-squares regression line.

6. What are the slope and intercept returned by the function lm? Does it agree with your answer in the previous part? Should it?

7. Add both lines to the bivariate EDA plot. (Add only one line, of course, if you think they are the same.) Comment on the fit. Do the assumptions of the simple linear regression model appear to hold? Are they any places where the fit seems better than others?

8. Find Pittsburgh in the data set. What is the population? The per-capita GMP? The per-capita GMP predicted by the model? The residual for Pittsburgh?

9. What is the mean squared error of the regression? That is, what is $n^{-1}\sum_{i=1}^{n} e_i^2$ where $e_i = Y_i - \hat{Y}_i$ is the residual.

10. Is the residual for Pittsburgh large, small, or typical compared to the mean squared error?

11. Make a plot of residuals (vertical axis) against population (horizontal axis). What should this look like if the assumptions of the simple linear regression model hold? Is the actual plot compatible with those assumptions? Explain.

continued on page 2

12. Make a plot of squared residuals (vertical axis) against population (horizontal axis). What should this look like if the assumptions of the simple linear regression model hold? Is the actual plot compatible with those assumptions? Explain.

13. State, carefully, the interpretation of the estimated slope; be sure to refer to the actual variables of the problem, not abstract ones like "the predictor variable" or X.

14. What per-capita GMP does the model predict for a city with 10^5 more people than Pittsburgh?

15. What does the model predict would happen to Pittsburgh's per- capita GMP if, by a policy intervention, we added 10^5 people to the population?