# Homework 3

36-401, Fall 2015

Due at the start of class, 24 September 2015

1. (50) *Urban economies* The data file `http://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/03/bea-2006.csv` contains information about the economies of the 366 "metropolitan statistical areas" ($\approx$ 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.

   (a) (2) 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?

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

   (c) (6) Make univariate EDA plots for population and for per-capita GMP, and describe their distributions in words.

   (d) (6) Make a bivariate EDA plot for per-capita GMP as a function of population. Describe the relationship in words.

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

   (f) (3) What are the slope and intercept returned by the function `lm`? Does it agree with your answer in the previous part? Should it?

   (g) (4) Add both lines to the bivariate EDA plot. (Add only one line, of course, if you think they are the same.) Comment in the fit. Do the assumptions of the simple linear regression model appear to hold? Are there any places where the fit seems better than others?

   (h) (4) 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?

   (i) (2) What is the mean squared error of the regression?

   (j) (1) Is the residual for Pittsburgh large, small, or typical compared to the mean squared error? (Be careful!)

   (k) (3) 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.

(l) (3) 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.

(m) (3) 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 "$X$".

(n) (3) What per-capita GMP does the model predict for a city with $10^5$ more people than Pittsburgh?

(o) (3) What, if anything, 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?

2. *Fun with residuals* (10) Prove that the following are all true for any least-squares regression line. Throughout, $e_i$ refers to the residuals, $Y_i$ to the observed responses, $\hat{Y}_i$ to the fitted values, and bars indicate sample averages. *General hint:* this problem involves no probability, expectation values or variances.

   (a) (2) The residuals average to zero: $\overline{e} = 0$

   (b) (4) The sample covariance of the residuals with $X$ is zero: $\overline{eX} - \bar{e}\bar{X} = 0$.

   (c) (4) The mean of the fitted values equals the mean of the observed response values: $\overline{\hat{Y}} = \overline{Y}$.

3. *Regression through the origin* (30) In many situations, we know that $Y$ should be exactly proportional to $X$, but don't know the proportionality factor. This is the special case of the simple linear regression model where $\beta_0 = 0$, called "regression through the origin".

   (a) (10) Find the least squares estimate of the slope for regression through the origin. Begin by writing out the function to be minimized in terms of the $X_i$, the $Y_i$, $n$, and the model parameter (or parameters — which?); then differentiate it to get a "normal" or "estimating" equation (or equations), and finally solve the equation(s) for the slope. Express your final answer as a formula involving sample averages of $X$, $Y$, and their products and powers. Explain how (if at all) your answer differs from the usual least-squares estimate.

   (b) (5) Show that this estimator of the slope is unbiased, *assuming* the regression-through-the-origin model. Is this estimator of the slope unbiased for simple linear regression in general? (We know the usual least-squares estimator of the slope is unbiased.)

(c) (5) Show that the fitted value at $x$ can be written as $\beta_1 x + \sum_{i=1}^{n} c_i \epsilon_i$; find an explicit formula for the weights $c_i$.

(d) (5) Use (3c) to show that the fitted values are unbiased, and find a formula for the variance of the fitted values as a function of $\sigma^2$, $n$, the $X_i$ and $X$.

(e) (5) Suppose that the noise around the line through the origin is Gaussian, independent of $X$ and independent across observations. Find maximum likelihood estimate of the parameter(s). Does it agree or disagree with the least-squares estimate?

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is either properly integrated with a tool like R Markdown or knitr, or included as a separate R file. In the former case, both the knitted and the source file are included. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output. For full credit, all code runs, and the Markdown file knits (if applicable).

Extra Credit (10) *Trigonometric regression* If the predictor variable $X$ is an angle, it is often related to the response $Y$ by $Y_i = \alpha_0 + \alpha_1 \cos X_i + \alpha_2 \sin X_i + \epsilon_i$. Assume that the $\epsilon_i \sim N(0, \sigma^2)$, and are independent of $X_i$.

1. (2) Write down the log-likelihood function for this model.
2. (6) Find formulas for the maximum likelihood estimates of $\alpha_0$, $\alpha_1$, $\alpha_2$ and $\sigma^2$. Simplify as much as possible.
3. (1) Explain how this model is related to the model $Y_i = \gamma_0 + \gamma_1 \sin(X_i - \gamma_2) + \epsilon_i$.
4. (2) Can you find formulas for the maximum likelihood estimators of the $\gamma$ parameters in terms of the MLEs for the $\alpha$ parameters?