

Homework 3

36-401/607 Section B (Wasserman) Due Friday, September 22 at 3:00

1. (10 points) Suppose we have data $(X_1, Y_1), \dots, (X_n, Y_n)$. We fit a linear model using least squares to get $\hat{\beta}_0$ and $\hat{\beta}_1$. Let $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. (We do not assume that the linear model is true.) Prove that

$$\frac{1}{n} \sum_{i=1}^n \hat{Y}_i = \frac{1}{n} \sum_{i=1}^n Y_i.$$

2. (40 points) Assume the linear model with Normal errors. In other words, assume that

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon_1, \dots, \epsilon_n \sim N(0, \sigma^2)$.

- (a) (20) Let $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$ where $e_i = Y_i - \hat{Y}_i$. Show that this estimator is unbiased. In other words, show that $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.

Hint: You may use this fact that we stated in class: $\frac{\sum_{i=1}^n e_i^2}{\sigma^2} \sim \chi_{n-2}^2$.

- (b) (20) We saw that the maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2.$$

What is the amount of bias in this estimator? What happens to the bias as the sample size n becomes large?

Note: Recall (from 36-226) that “amount of bias” is defined as the difference between the expected value of an estimator and the true parameter value. Of course, for unbiased estimators, this will be zero.

3. (50 points) *Urban economies.*

This question will help you prepared for writing full reports on the results of statistical analyses. Hence, when asked to explain, comment, or discuss, use complete sentences and answer thoroughly **in the context of the problem** for full credit. Use R when appropriate for the homework problems. You do not need to hand in code, but you do need to hand in graphs and a writeup of the R output. **Figures should be clearly labeled and readable.**

The data file <http://www.stat.cmu.edu/~larry/=stat401/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. (Use the commands `hist` and `boxplot`.)
- (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 on 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? That is, what is $n^{-1} \sum_{i=1}^n e_i^2$ where $e_i = Y_i - \hat{Y}_i$ is the residual.
- (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?