

**Homework Assignment 2**  
 (Due on 5/22 9:45am in class)

**Problem 1. Regression through the origin.** We will consider a special case of the simple linear regression model, where the intercept term is assumed to be zero from the outset (this is often assumed in the calibration of certain measuring devices). Let

$$Y_i = x_i\beta + \epsilon_i,$$

where  $E(\epsilon_i) = 0$ ,  $V(\epsilon_i) = \sigma^2$ , and the  $\epsilon_i$ s are uncorrelated. In what follows we will treat the  $x_i$ s as known constants.

- (a) Find the least squares estimate of  $\beta$ , call it  $b$ ; i.e., find  $\beta$  that minimizes  $Q(\beta) = \sum_{i=1}^n (Y_i - x_i\beta)^2$ .
- (b) Show that  $E(b) = \beta$ .
- (c) Show the  $V(b) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$ .
- (d) Investigate whether  $b$  is a BLUEs estimator of  $\beta$ .

**Problem 2.** Suppose a sample of 10 types of compact cars reveals the following one-day rental prices (in dollars) for Hertz and Thrifty, respectively:

Renter	Car Type									
	A	B	C	D	E	F	G	H	I	J
Hertz	37.16	14.36	17.59	19.73	30.77	26.29	30.03	29.02	22.63	39.21
Thrifty	29.49	12.19	15.07	15.17	24.52	22.32	25.30	22.74	19.35	34.44

- (a) Explain why this is a paired-sample problem.
- (b) Use a graph to determine whether the assumption of normality is reasonable.
- (c) Using a p-value, test at  $\alpha = 0.05$  whether Thrifty has a lower true mean rental rate than Hertz via a t-test.

**Problem 3.** Supposed our observed data follow a normal error simple linear regression model,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i^{iid} \sim N(0, \sigma^2).$$

Now suppose we have a sample of size  $n$ , where  $n$  is an even number, and we divide the sample up into non-overlapping pairs,  $(i_1, j_1), (i_2, j_2), \dots, (i_{n/2}, j_{n/2})$ ; i.e., we arrange our sample as  $\{(x_{i_1}, y_{i_1}), (x_{j_1}, y_{j_1})\}, \dots, \{(x_{i_{n/2}}, y_{i_{n/2}}), (x_{j_{n/2}}, y_{j_{n/2}})\}$ . Consider the following estimator of  $\beta_1$ :

$$\tilde{\beta}_1 = \frac{2}{n} \sum_{k=1}^{n/2} \frac{Y_{i_k} - Y_{j_k}}{x_{i_k} - x_{j_k}}.$$

For instance, if we had  $n = 6$  observations and paired the indices as  $(1, 6), (2, 5), (3, 4)$ , the estimator would be

$$\tilde{\beta}_1 = \frac{1}{3} \left[ \frac{Y_1 - Y_6}{x_1 - x_6} + \frac{Y_2 - Y_5}{x_2 - x_5} + \frac{Y_3 - Y_4}{x_3 - x_4} \right].$$

- (a) Show that  $\tilde{\beta}_1$  is an unbiased estimator of  $\beta_1$ .
- (b) Without any additional calculations, what can you say about the variance of  $\tilde{\beta}_1$  as it compares to the variance of the usual least squares estimator  $\hat{\beta}_1$ ?

**Problem 4.** An economist compiled data on the productivity improvements last year for a sample of firms producing electronic computing equipment. The firms were classified according to the level of their average expenditures for research and development in the past three years (Low and Moderate). The results of the study follows:

Rating	1	2	3	4	5	6	7	8	9	10	11	12
Low	7.6	8.2	6.8	5.8	6.9	6.6	6.3	7.7	6.0			
Moderate	6.7	8.1	9.4	8.6	7.8	7.7	8.9	7.9	8.3	8.7	7.1	8.4

- (a) Use R to prepare side-by-side box plots for the two samples. Do the spreads seem to differ across samples?
- (b) Use a graph to determine whether the assumption of normality is reasonable.
- (c) Using R, obtain the p-value from the test of  $H_0 : \sigma_1^2 = \sigma_2^2$ . What do you conclude here? **Note that you can only use a cumulative distribution function in R.**
- (d) Using a p-value from a t-test, test at  $\alpha = 0.05$  whether the firms rated Moderate have a significantly higher mean productivity improvement than those rated Low. **Note that you can only use a cumulative distribution function in R. Also note that you may want to consider what you have discovered in parts (a)-(c))**

- (e) Obtain and interpret a 95% CI for the difference in mean productivity improvement between firms rated Moderate and those rated Low.

**Problem 5.** Refer again to the salt concentration data. Assume a model of the form

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

is appropriate, where  $Y$  is the salt concentration and  $x$  is the roadway area. Do the following **by hand**. (You can use the following results:  $\bar{x} = 0.824$ ,  $\bar{y} = 17.135$ ,  $\sum_i x_i^2 = 17.2502$ ,  $\sum_i x_i y_i = 346.793$ .)

- (a) Find  $\hat{\beta}_0$  and  $\hat{\beta}_1$  and 95% confidence intervals about  $\beta_0$  and  $\beta_1$ . Give the estimated regression equation. You may use the fact that, for this model fit,  $MSE = 3.21$ .
- (b) Estimate the mean salt concentration when 1.5 percent of the watershed area consists of paved roads. Provide the 95% confidence interval. You may use the fact that, for this model fit,  $MSE = 3.21$ .
- (c) Predict what salt concentration we would actually observe when 1.5 percent of the watershed area consists of paved roads. Give the 95% prediction interval. You may use the fact that, for this model fit,  $MSE = 3.21$ .

**Problem 6.** Discussion questions: Provide a brief response summarizing your thoughts.

- (a) The regression function relating the production output ( $Y$ ) by an employee taking a training program for ( $X$ ) hours, where  $X$  ranges from 40 to 100. An observer concludes that the training program does not increase productivity since  $\beta_1$  is less than 1. Comment.
- (b) Evaluate the statement "For the least squares method to be fully valid, it is required that the distribution of  $Y$  be normal." Comment.
- (c) Recall, we saw that  $\sum_{i=1}^n e_i = 0$  and we have used  $e_i$  as a surrogate for the unobserved  $\epsilon_i$ , thus does this imply that  $\sum_{i=1}^n \epsilon_i = 0$ . Comment.