STAT 2131:
Applied Statistical Methods I
HW #1
Due Thursday September 5th

For Problems 1 and 2 (the problems from the text): please do the calculations "by hand" (i.e. do not run SAS or R). Assume, when applicable, that $\infty \cdot 0 = 0$.

1. 1.21 from KNNL.

2. 2.6a-d from KNNL. For parts c and d, do not find the p-value but evaluate the hypothesis test at the $\alpha = 0.05$ level.

3. Suppose data are generated as $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $E(\epsilon_i \mid x_1, \ldots, x_n) = 0$ for all $i = 1, \ldots, n$. Show that the ordinary least squares estimates for β_0 and β_1 are unbiased.

4. We have looked at using ordinary least squares/maximum likelihood estimation for the simple linear regression model. This problem considers an alternative estimation procedure. For simplicity, we will assume that the variance σ^2 is known.

   You observe outcomes y_i from $i = 1, \ldots, n$ subjects and assume that the data follow the linear model

   $$y_i = \beta_0 + x_i\beta_1 + \epsilon_i$$

   where x_i are known deterministic covariates, β_0 and β_1 are unknown deterministic parameters, and ϵ_i are independent and identically distributed mean zero Gaussian random variables with *known* variance 1. Given some $\lambda \geq 0$, you decide to estimate β_1 with $\tilde{b}_1$ that minimizes the penalized sum-of-squares

   $$\text{PSS}_\lambda(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 + \lambda\beta_1^2$$

   so that

   $$\left(\tilde{b}_0, \tilde{b}_1\right) = argmin_{\beta_0, \beta_1 \in \mathbb{R}} \text{PSS}_\lambda(\beta_0, \beta_1).$$

   (a) What are $\tilde{b}_0, \tilde{b}_1$ when $\lambda = 0$? When $\lambda = \infty$?

   (b) A collaborator claims that $\tilde{b}_1$ always has smaller variance than the best linear unbiased estimator of β_1. Prove or disprove this claim.

   (c) Show that $\tilde{b}_1$ is a biased estimate for β_1 when $\lambda > 0$.

5. A researcher is interested in evaluating the efficacy of a new treatment over conventional treatment in improving the quality of life among depression patients. Suppose she randomly assigns 50 patients to the new treatment group and the other 50 patients to the conventional treatment group. It is reasonable to assume that the quality of

life for the new treatment group and the conventional treatment gorup both follow a normal distribution, with means μ_1 and μ_2, respectively. Without loss of generality, the standard deviations of the two groups are assumed to be 1. The researcher expects to see a difference in the means $\Delta = \mu_1 - \mu_2 = 0.6$. She plans to employ a simple one-sided test for the two groups comparison at the significance level $\alpha = 0.05$. What is the power of detecting such a difference based on the sample she has?

6. (Use computer) Set $x_i = i/100$ for $i = 1, \ldots, n$, generate $y_i = 1 + 2x_i + \epsilon_i$ with $\epsilon_i \overset{i.i.d}{\sim} N(0, 1)$ for $i = 1, \ldots, n$. Then fit a simple linear regression model and obtain $\hat{\beta}_1$, as well as a 95% confidence interval. Repeat that 100 times.

   (a) What is the empirical bias and variance of $\hat{\beta}_1$?

   (b) What is the empirical coverage of CI for β_1?

   (c) Do (a) and (b) agree with their theoretical values?

1.19. **Grade point average.** The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	118	119	120
X_i:	21	14	28	...	28	16	28
Y_i:	3.897	3.885	3.778	...	3.914	1.860	2.948

a. Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function.

b. Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?

c. Obtain a point estimate of the mean freshman GPA for students with ACT test score $X = 30$.

d. What is the point estimate of the change in the mean response when the entrance test score increases by one point?

*1.20. **Copier maintenance.** The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data below have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers serviced and Y is the total number of minutes spent by the service person. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	43	44	45
X_i:	2	4	3	...	2	4	5
Y_i:	20	60	46	...	27	61	77

a. Obtain the estimated regression function.

b. Plot the estimated regression function and the data. How well does the estimated regression function fit the data?

c. Interpret b_0 in your estimated regression function. Does b_0 provide any relevant information here? Explain.

d. Obtain a point estimate of the mean service time when $X = 5$ copiers are serviced.

*1.21. **Airfreight breakage.** A substance used in biological and medical research is shipped by air-freight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y). Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	4	5	6	7	8	9	10
X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here?

b. Obtain a point estimate of the expected number of broken ampules when $X = 1$ transfer is made.

c. Estimate the increase in the expected number of ampules broken when there are 2 transfers as compared to 1 transfer.

d. Verify that your fitted regression line goes through the point $(\bar{X}, \bar{Y})$.

1.22. **Plastic hardness.** Refer to Problems 1.3 and 1.14. Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours, and Y is hardness in Brinell units. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	14	15	16
X_i:	16	16	16	...	40	40	40
Y_i:	199	205	196	...	248	253	246

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here?

b. Obtain a point estimate of the mean hardness when $X = 40$ hours.

c. Obtain a point estimate of the change in mean hardness when X increases by 1 hour.

1.23. Refer to **Grade point average** Problem 1.19.

a. Obtain the residuals e_i. Do they sum to zero in accord with (1.17)?

b. Estimate σ^2 and σ. In what units is σ expressed?

*1.24. Refer to **Copier maintenance** Problem 1.20.

a. Obtain the residuals e_i and the sum of the squared residuals $\sum e_i^2$. What is the relation between the sum of the squared residuals here and the quantity Q in (1.8)?

b. Obtain point estimates of σ^2 and σ. In what units is σ expressed?

*1.25. Refer to **Airfreight breakage** Problem 1.21.

a. Obtain the residual for the first case. What is its relation to ε_1?

b. Compute $\sum e_i^2$ and MSE. What is estimated by MSE?

1.26. Refer to **Plastic hardness** Problem 1.22.

a. Obtain the residuals e_i. Do they sum to zero in accord with (1.17)?

b. Estimate σ^2 and σ. In what units is σ expressed?

*1.27. **Muscle mass.** A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79. The results follow; X is age, and Y is a measure of muscle mass. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	58	59	60
X_i:	43	41	47	...	76	72	76
Y_i:	106	106	97	...	56	70	74

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here? Does your plot support the anticipation that muscle mass decreases with age?

b. Obtain the follow

*2.6. Refer to **Airfreight breakage** Problem 1.21.

a. Estimate β_1 with a 95 percent confidence interval. Interpret your interval estimate.

b. Conduct a t test to decide whether or not there is a linear association between number of times a carton is transferred (X) and number of broken ampules (Y). Use a level of significance of .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

c. β_0 represents here the mean number of ampules broken when no transfers of the shipment are made—i.e., when $X = 0$. Obtain a 95 percent confidence interval for β_0 and interpret it.

d. A consultant has suggested, on the basis of previous experience, that the mean number of broken ampules should not exceed 9.0 when no transfers are made. Conduct an appropriate test, using $\alpha = .025$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

e. Obtain the power of your test in part (b) if actually $\beta_1 = 2.0$. Assume $\sigma\{b_1\} = .50$. Also obtain the power of your test in part (d) if actually $\beta_0 = 11$. Assume $\sigma\{b_0\} = .75$.