MATH 6364 Introduction to Biostatistics

Midterm Exam

Instructor: Dr. George Yanev

1. Suppose that based on data $(x_1, y_1), \ldots, (x_n, y_n)$ we wish to fit the linear regression model $y = \beta_0 + \beta_1 x$. Write down the matrices X, $X^T X$ and $(X^T X)^{-1}$ for this model. Find the estimates of β_0 and β_1, and calculate the corresponding variances. Show that the entries of the H matrix are:

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

2. Often the conditions of the problem dictate that the intercept β_0 must be zero, e.g. the sales revenue as a function of the number of units sold or the gas mileage of a car as a function of the weight of the car. This is called regression through the origin. Show that the estimate of the slope β_1 when fitting the line $y = \beta_1 x$ based on the data (x_1, y_1), (x_2, y_2), ... , (x_n, y_n) is

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

#3.

Consider the punting data in Exercise 2.11. Fit three separate models

$$E(y_i) = \beta_0 + \beta_1 x_{1i} \qquad \text{(model 1)}$$
$$E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \qquad \text{(model 2)}$$
$$E(y_i) = \beta_0 + \beta_2 x_{2i} \qquad \text{(model 3)}$$

Let x_{1i} be right-leg strength; x_{2i} is left-leg strength.

(a) Plot the residuals for these three models. Plot against $\hat{y}$. Comment on which model seems to be most appropriate.

(b) Test

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

Make appropriate conclusion.

Data for Q3.

2.11 In an experiment to determine the influence of certain physical measures on the performance of punters in American football, 13 punters were used as subjects in an experiment in which the average distance on 10 punts was measured. In addition, measures of left leg and right leg strength (lb lifted) were taken via a weight lifting test. The following data were taken. All subjects use their right legs for punting.

Subject	Left Leg (lb)	Right Leg (lb)	Average Punting Distance
1	170	170	162 ft 6 in.
2	130	140	144 ft 0 in.
3	170	180	147 ft 6 in.
4	160	160	163 ft 6 in.
5	150	170	192 ft 0 in.
6	150	150	171 ft 9 in.
7	180	170	162 ft 0 in.
8	110	110	104 ft 10 in.
9	110	120	105 ft 8 in.
10	120	130	117 ft 7 in.
11	140	120	140 ft 3 in.
12	130	140	150 ft 2 in.
13	150	160	165 ft 2 in.

Data analyzed for the Department of Health, Physical Education and Recreation by the Statistical Consulting Center, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1983.

#4.

3.5 Consider the squid data in Example 3.2.

(a) Generate the residuals for the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

(b) Compute 95% confidence intervals on the mean response and 95% prediction intervals on a new observation at the conditions of the 22 specimen.

(c) Compute a new multiple regression using regressors x_2, x_4, and x_5. Compute s^2, the standard errors of prediction, and 95% confidence intervals on the mean response at the regressor locations for the 22 specimen.

(d) Use the information from (a), (b), and (c) to make a choice between the full model and the reduced model in part (c).

(e) For the squid data, test

$$H_0: \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0 \qquad H_1: \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \neq 0$$

Draw conclusions. Comment on the results.

Data for Q4.

TABLE 3.2 Squid weight and beak measurements

x_1	x_2	x_3	x_4	x_5	y
1.31	1.07	0.44	0.75	0.35	1.95
1.55	1.49	0.53	0.90	0.47	2.90
0.99	0.84	0.34	0.57	0.32	0.72
0.99	0.83	0.34	0.54	0.27	0.81
1.05	0.90	0.36	0.64	0.30	1.09
1.09	0.93	0.42	0.61	0.31	1.22
1.08	0.90	0.40	0.51	0.31	1.02
1.27	1.08	0.44	0.77	0.34	1.93
0.99	0.85	0.36	0.56	0.29	0.64

TABLE 3.2 Continued

TABLE 3.2
Continued

Squid weight and beak measurements

x_1	x_2	x_3	x_4	x_5	y
1.34	1.13	0.45	0.77	0.37	2.08
1.30	1.10	0.45	0.76	0.38	1.98
1.33	1.10	0.48	0.77	0.38	1.90
1.86	1.47	0.60	1.01	0.65	8.56
1.58	1.34	0.52	0.95	0.50	4.49
1.97	1.59	0.67	1.20	0.59	8.49
1.80	1.56	0.66	1.02	0.59	6.17
1.75	1.58	0.63	1.09	0.59	7.54
1.72	1.43	0.64	1.02	0.63	6.36
1.68	1.57	0.72	0.96	0.68	7.63
1.75	1.59	0.68	1.08	0.62	7.78
2.19	1.86	0.75	1.24	0.72	10.15
1.73	1.67	0.64	1.14	0.55	6.88

Rudolf J. Freund, SAS Tutorial, "Regression with SAS with Emphasis on PROC REG" (Paper presented at Eighth Annual SAS Users Group International Conference, New Orleans, Louisiana, January 16–19, 1983).

a. Determine an approximate 95% prediction interval for the fuel efficiency of an automobile weighing 2000 pounds. The computer output does not give you the information to construct exact prediction intervals. Approximate the prediction intervals, assuming that the sample size n is large enough to allow you to ignore the parameter estimation uncertainty.

b. Determine an approximate 95% prediction interval for the fuel efficiency of an automobile weighing 1500 pounds.

2.11. Discuss the functional relationship between the coefficient of determination R^2 and the F ratio.

2.12. Occasionally, a model is considered in which the intercept is known to be zero a priori. Such a model is given by

$$y_i = \beta_1 x_i + \epsilon_i, i = 1, 2, \ldots, n$$

where the errors ϵ_i follow the usual assumptions.

a. Obtain the LSEs $(\hat{\beta}_1, s^2)$ of (β_1, σ^2).

b. Define $e_i = y_i - \hat{\beta}_1 x_i$. Is it still true that $\sum_{i=1}^n e_i = 0$? Why or why not?

c. Show that $V(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^n x_i^2$.

2.13. The data listed in the file **sriver** include the water content of snow on April 1 (x) and the water yield from April to July (y) in the Snake River watershed in Wyoming. Information on $n = 17$ years (from 1919 to 1935) is listed (see Weisberg, 1980).

a. Fit a regression through the origin ($y = \beta_1 x + \epsilon$), and find $\hat{\beta}_1$ and s^2. Obtain a 95% confidence interval for β_1.

b. A more general model for the data includes an intercept,

$$y = \beta_0 + \beta_1 x + \epsilon.$$

Is there convincing evidence that suggests that the simpler model in (a) is an appropriate representation?

2.14. Often, researchers need to calibrate measurement processes. For that they use a set of known x's to obtain observed y's, then fit a model called the calibration model and use this model to convert future measured y's back into the corresponding x's.

The following is an example taken from analytical chemistry where the process is the assay of the element calcium. Determining calcium in the presence of other elements is quite tricky. The following table records the quantities of calcium in carefully prepared solutions (x) and the corresponding analytical results (y):

x	4	8	12.5	16	20	25	31	36	40	40
y	3.7	7.8	12.1	15.6	19.8	24.5	31.1	35.5	39.4	39.5

a. Fit a simple linear regression of y as a function of x. List the assumptions that you make.

b. Calculate a 95% confidence interval for the intercept of your model.

c. Calculate a 95% confidence interval for the slope of your model.

d. In this context two properties may be expected:

i. When $x = 0$, then $y = 0$; if there is no calcium present, your technique should not find any.

ii. If the empirical technique is any good at all, then the slope in the simple linear regression should be 1.

Is there evidence for (i)? For (ii)?

e. If you accept (i) as a condition to be imposed on the model a priori, then the model reduces to

$$y = \beta x + \epsilon$$

Redo part (c) and reexamine property (ii) for your new model.

f. Explain why the results in (d) and (e) are different.

2.15. The following data give the monthly machine maintenance cost (y) in hundreds of dollars