

Biostatistics

Math 322 — Spring 2009

Homework 12 — Due 04/24

Problem 1

Given the following data:

<i>Y</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
51.4	0.2	17.8	24.6	18.9
72.0	1.9	29.4	20.7	8.0
53.2	0.2	17.0	18.5	22.6
83.2	10.7	30.2	10.6	7.1
57.4	6.8	15.3	8.9	27.3
66.5	10.6	17.6	11.1	20.8
98.3	9.6	35.6	10.6	5.6
74.8	6.3	28.2	8.8	13.1
92.2	10.8	34.7	11.9	5.9
97.9	9.6	35.8	10.8	5.5
88.1	10.5	29.6	11.7	7.8
94.8	20.5	26.3	6.7	10.0
62.8	0.4	22.3	26.5	14.3
81.6	2.3	37.9	20.0	0.5
62.6	16.3	19.3	27.3	17.2

The data are available as a computer readable file on
http://www.math.wustl.edu/~hjelle/m322/m322_090424table1.txt.

- a) Fit the multiple regression model

$$Y = \alpha + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D + e$$

to the data, computing the sample partial-regression coefficients.

What is the predicted value of Y at $A = 5.2$, $B = 22.6$, $C = 7.8$, and $D = 12.2$?

- b) Use an appropriate hypothesis test to test if any of the partial-regression coefficients are non-zero.

- c) For each partial-regression coefficient, carry out a hypothesis test to test if that coefficient is non-zero.
- d) Use a variable selection strategy to select a “best model” based on the data.

Problem 2

The data below provide the unemployment rates during the 10 years from 1950 to 1959 together with an index of industrial production from the Federal Reserve Board.

Year	Unemployment	Index of production
1950	3.1	113
1951	1.9	123
1952	1.7	127
1953	1.6	138
1954	3.2	130
1955	2.7	146
1956	2.6	151
1957	2.9	152
1958	4.7	141
1959	3.8	159

Primarily we want to find out how unemployment is related to industrial production. Of course, other factors will also have played a role, (can you explain the apparant drop in unemployment for the years 1951 – 1953?) and many of those factors will have changed over the course of the decade. As a surrogate for these other factors, the year will be included as an independent variable in the regression model.

- a) Fit a multiple regression model to the data.
- b) Is the model statistically significant? Which of the variables are statistically significant? Report p -values.

Problem 3

Refer to the article “Storks Deliver Babies ($p = 0.008$)”¹. The birthrate data are available in machine readable form on http://www.math.wustl.edu/~hjelle/m322/m322_090424table3.txt.

Can you find a more plausible explanation of the birth rate than “Storks Deliver Babies”?

¹Matthews, R. (2002). Storks Deliver Babies ($p = 0.008$). *Teaching Statistics*, 22(2), 36 – 38. Available through the course web site.

Problem 4 (Based on Problems 11.30 – 11.35 in the book)

The Update to the Task Force Report on Blood Pressure Control in Children reported the observed 90th percentile of systolic blood pressure in single years of age from age 1 to 17 based on prior studies. The data for boys of an average height are given in Table 11.21.

Table 11.21: 90th percentile of systolic blood pressure (SBP) in boys ages 1 – 17

Age	SBP	Age	SBP
1	104	10	118
2	106	11	120
3	107	12	122
4	108	13	124
5	109	14	126
6	111	15	129
7	112	16	131
8	114	17	133
9	116	18	136

Suppose we seek a more efficient way to display the data and choose linear regression to accomplish this task.

- Fit a regression line relating age to SBP.
- Use the estimated regression line to predict the systolic blood pressure for an average 13-year-old boy. What is the standard error of this prediction?
- Plot the residuals from your regression analysis. Do they seem to be independent and normally distributed?

Explain why would it make sense to consider the model

$$SBP = \alpha + \beta_1 \text{Age} + \beta_2 \text{Age}^2 + e?$$

Carry out this as a multiple regression. Plot the residuals from this analysis and compare them with the previous residuals.