

Stat 8061—First Examination, October 22, 2001

Mandible data

These data relate gestational age in weeks to mandible length in 158 fetuses of gestational age of 42 weeks or less. The computer output has been EDITED, but the information that remains is adequate to fill in the blanks.

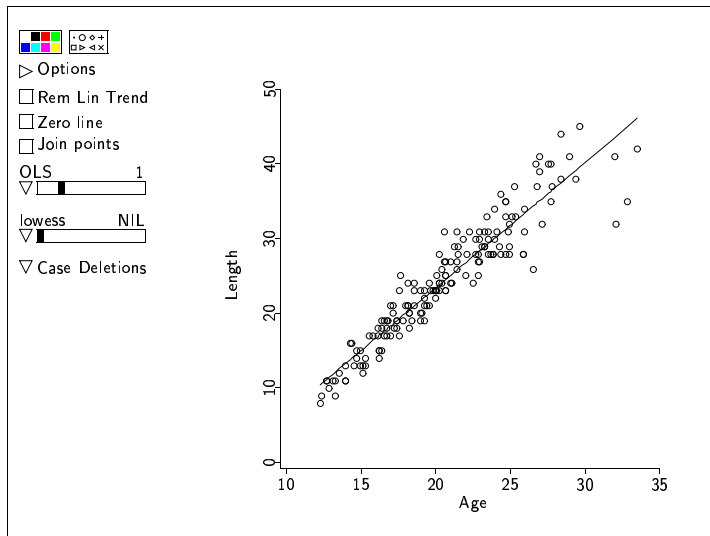


Figure 1: Mandible data.

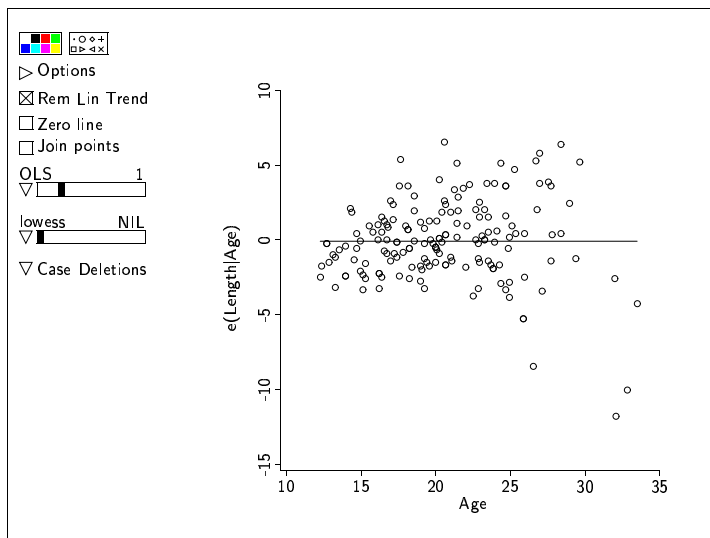


Figure 2: Mandible data.

Data set = Mandible, Summary Statistics

Variable	N	Average	Std. Dev	Minimum	Median	Maximum
Length	167	24.497	8.1225	8.0000	24.000	45.000
Age	167	20.646	4.5584	12.286	20.286	33.571

Data set = Mandible, Sample Correlations

	Length	Age
Length	1.0000	0.9418
Age	0.9418	1.0000

Name of Dataset = Mandible; Name of Model = L1
 Normal Regression Model
 Mean function = Identity
 Response = Length
 Predictors = (Age)

Coefficient Estimates

Label	Estimate	Std. Error	t-value
Constant	-10.1491	0.986127	-10.292
Age	1.67812	0.0466471	35.975

R Squared:
 Sigma hat:
 Number of cases: 167
 Degrees of freedom:

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression		9713.35		1294.18	0.0000
Residual		1238.39			

Data set = Mandible, Name of Fit = L1
 Normal Regression
 Kernel mean function = Identity
 Response = Length
 Terms = (Age)
 Term values = (20)
 Prediction = 23.4132, se(pred) = 2.74796, weight = 1
 Leverage = 0.0061, Max(h_i) = 0.0544
 Estimated population mean value = 23.4132, se = 0.214127

Variance-covariance matrix of the coefficient estimates

	Constant	Age
Constant	0.97245	-0.044925
Age	-0.044925	0.0021760

Correlation matrix of the coefficient estimates

	Constant	Age
Constant	1.0000	-0.9766
Age	-0.9766	1.0000

Corn data

These data relate average corn *Yield*, in bushels per acre, to summer *Rainfall* in inches and mean high temperature *Temp* for the growing season in degrees F for the Years 1890 to 1956. The variable *Year* refers to the year of measurement. These data represent an average over several Midwestern states. Here is some computer output for these data, and a few graphs.

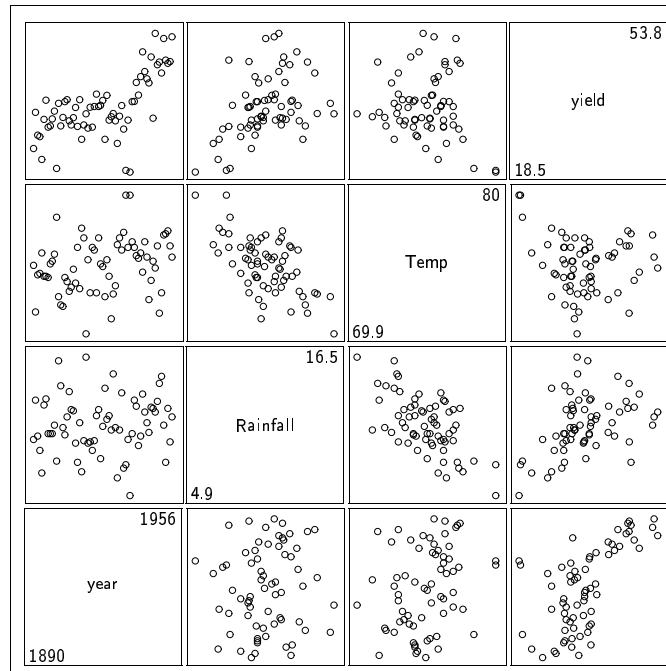


Figure 3: Scatterplot matrix for the corn data.

Data set = Corn, Summary Statistics

Variable	N	Average	Std Dev	Minimum	Median	Maximum
Rainfall	67	10.834	2.359	4.9	10.7	16.5
Temp	67	74.896	2.0082	69.9	75.	80.
Year	67	1923.	19.485	1890.	1923.	1956.
Yield	67	35.176	7.8905	18.5	34.	53.8

Data set = Corn, Sample Correlations

	Rainfal	Temp	Year	Yield
Rainfall	1.0000	-0.6206	0.0761	0.4169
Temp	-0.6206	1.0000	0.2776	-0.1458
Year	0.0761	0.2776	1.0000	0.6472
Yield	0.4169	-0.1458	0.6472	1.0000

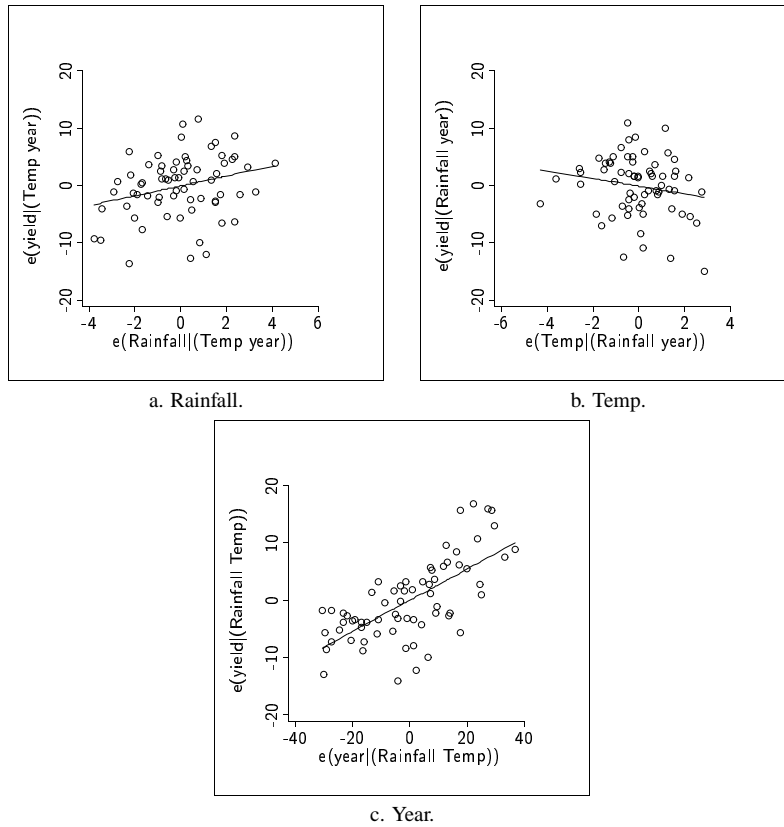


Figure 4: Added-variables plots for the corn data.

```

Data set = Corn, Name of Fit = L1
Normal Regression, Kernel mean function = Identity
Response      = Yield
Terms         = (Rainfall Temp Year)
Coefficient Estimates
Label      Estimate      Std. Error      t-value      p-value
Constant  -449.353             64.7731        -6.937       0.0000
Rainfall   0.863139            0.373209         2.313       0.0240
Temp      -0.680397            0.455015        -1.495       0.1398
Year       0.273602            0.0368804       7.419       0.0000

R Squared:          -----
Sigma hat:          5.29516
Number of cases:    67
Degrees of freedom: 63

Summary Analysis of Variance Table
Source      df      SS      MS      F      p-value
Regression  3      2342.72  780.908  27.85  0.0000
Residual    63     1766.44  28.0387

Variance-covariance matrix of the coefficient estimates
Constant  4195.6    -1.3032   -3.3341   -2.0444
Rainfall  -1.3032    0.13929  0.11375   -0.0045375
Temp      -3.3341    0.11375  0.20704   -0.0069707
Year      -2.0444   -0.0045375 -0.0069707  0.0013602
          Constant  Rainfall  Temp      Year

```

Data set = Corn, Name of Fit = L2

Normal Regression

Kernel mean function = Identity

Response = Yield

Terms = (Temp)

Coefficient Estimates

Label	Estimate	Std. Error	t-value	p-value
Constant	78.0751	36.1232	2.161	0.0344
Temp	-0.572784	0.482144	-1.188	0.2392

R Squared: 0.0212513

Sigma hat: 7.86603

Number of cases: 67

Degrees of freedom: 65

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	1	87.325	87.325	1.41	0.2392
Residual	65	4021.84	61.8744		

Data set = Corn, Name of Fit = L3

Normal Regression

Kernel mean function = Identity

Response = Yield

Terms = (Rainfall Year)

Coefficient Estimates

Label	Estimate	Std. Error	t-value	p-value
Constant	-460.310	64.9758	-7.084	0.0000
Rainfall	1.23697	0.279762	4.422	0.0000
Year	0.250694	0.0338705	7.402	0.0000

R Squared: 0.554864

Sigma hat: 5.34605

Number of cases: 67

Degrees of freedom: 64

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	2	2280.03	1140.01	39.89	0.0000
Residual	64	1829.13	28.5802		

Backward Elimination: Sequentially remove terms

that give the smallest change in C_I.

All fits include an intercept.

Current terms: (Rainfall Temp Year)

	df	RSS	k	C _I
Delete: Temp	64	1829.13	3	4.236
Delete: Rainfall	64	1916.41	3	7.349
Delete: Year	64	3309.57	3	57.036

Current terms: (Rainfall Year)

	df	RSS	k	C _I
Delete: Rainfall	65	2387.87	2	22.163
Delete: Year	65	3394.84	2	58.077

Name: _____

Stat 8061—First Examination Problems

You may use up to two pages of notes. Write all your answers on the exam. Partial credit will be given where appropriate, so attempt to answer all questions. An answer of $(77 - 13)/4$ will get the same credit as the answer 8, so there is no need to simplify computations. There are 10 questions, each worth 10 points. Good luck!

Questions about the Mandible Data

1. Based on the computer output provided, summarize the evidence for or against the possibility that $(Length, Age)$ follows a bivariate normal distribution.

2. Consider two regression mean functions:

$$E(MandibleLength|age) = \eta_0 + \eta_1 age$$

$$E(age|MandibleLength) = \gamma_0 + \gamma_1 MandibleLength$$

Why is it not true that $\hat{\gamma}_1 = 1/\hat{\eta}_1$? Are there conditions (which are not satisfied in this problem) such that $\hat{\gamma}_1 = 1/\hat{\eta}_1$?

3. Assume the model fit on the handout is appropriate. (a) Give the value of a prediction of mandible length at gestation age 25 weeks. (b) Is the standard error of prediction at 25 weeks larger or smaller than the standard error of prediction at 20 weeks (you don't need to compute it)? How do you know? (c) Do you think it is reasonable to use the t -distribution to get a prediction interval for mandible length at age 25 weeks? Why or why not?

Questions about the Corn data

4. Summarize the evidence in the computer output about the importance of *Temp* after adjusting for *Rainfall* and *Year*; it is up to you to decide what the evidence is, and tell me why it is useful.

5. Test the hypothesis that the mean function for *Yield* depends only on *Year* against the alternative that it depends on all three predictors. State the hypotheses, the test statistic, including the numerical values of quantities from the computer output supplied, and the distribution of the test. State the fewest assumptions you need to make for the test to be used.

6. Using the computer output labeled “Backward Elimination” for reference, explain how backward elimination works, and the goals behind its use in subset selection. Also, give ONE reason why subset selection is a hard problem.

Other Questions

7. An industrial experiment is carried out to study $E(y|x)$, where x is temperature and y the hardness of the resulting product. In the experiment, (1) x is observed at each of five values, x_1, \dots, x_5 ; and (2) At the i -th value of x , a random sample of m_i parts are measured. Use this information to answer the next two questions.

7.1. Describe a model you can fit for which the residual sum of squares will be exactly the same as the sum of squares for pure error.

7.2. If the data you received consisted only the mean of y at each value of x and the sample sizes, describe how you would compute estimates of the slope and intercept for simple linear regression, assuming again that the variance function for a single observation is constant.

8. Explain **briefly** why a lowess smooth of a plot of y versus x with smoothing parameter that is too large will generally give a biased estimate of $E(y|x)$. Sketching a graph may help.

9. Suppose it is known that

$$E(y|x) = \eta_0 + \eta_1 x + \eta_2 x^2 + \eta_3 x^3$$

with $\text{var}(y|x) = \sigma^2$. Further, suppose that it is *known for certain* that $E(y|x = 10) = 4$. Given a sample $(x_i, y_i), i = 1, 2, \dots, n$, describe how you would estimate the parameters in the cubic polynomial.

10. Suppose you have a regression problem $F(y|x)$ with one predictor x , and you have decided that the simple linear regression mean function,

$$E(y|x) = \eta_0 + \eta_1 x \quad (1)$$

is appropriate for your data. Now suppose a new variable z becomes available, and you now want to look at $F(y|x, z)$. Suppose further that z is binary, either 0 or 1, representing the presence or absence of some characteristic.

10.1. Write down a mean function $E(y|x, z)$ that has a common slope for $z = 0$ and $z = 1$, but different intercepts for each of the two categories.

10.2. Assuming all the parameters in the mean function you wrote down in the last subproblem are non-zero, and that z is not constant, under what conditions can *both* the mean function you derived in the last subproblem and the mean function (1) be true?

10.3. To the mean function in the first subproblem we add the assumption that $\text{var}(y|x, z) = \sigma^2$, a constant. What is the variance function $\text{var}(y|x)$, and under what conditions is it constant?