

Stat 8061—Solutions to First Examination, Fall 2001

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.

The point cloud is not elliptical; the mean function is not linear (this is evident in the residual plot for larger values of Age), and the variance function is not constant. Normality requires an elliptical point cloud with a straight mean function and constant variance function.

2. Consider two regression mean functions:

$$\begin{aligned}E(\text{MandibleLength}|\text{age}) &= \eta_0 + \eta_1 \text{age} \\E(\text{age}|\text{MandibleLength}) &= \gamma_0 + \gamma_1 \text{MandibleLength}\end{aligned}$$

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 $\hat{\gamma}_1 = 1/\hat{\eta}_1$?

If r is the sample correlation $\hat{\eta}_1 = r \times (SD_y/SD_x)$, while $\hat{\gamma}_1 = r \times (SD_x/SD_y)$, and we have $\hat{\gamma}_1 = 1/\hat{\eta}_1$ only if $r = 0, +1$ or -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?

The prediction at 25 weeks is $-10.1491 + 1.67812 \times 25$. Its standard error is larger than at age=20 weeks because the mean of age in the data is 20.646, and the standard error of prediction is smallest at the mean. Prediction intervals depend on the normality of the errors, not the asymptotic normality of the estimates. If the assumption of normality of errors is not acceptable, then the assumption of normality for prediction intervals is not acceptable.

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.

Look at the t -value, state the hypothesis it tests, and look at the added-variable plot. Unfortunately, the labels under the added-variable plots are in the wrong order.

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.

The verbal description of the test was wrong (but has been fixed as shown above). The null hypothesis is always the smaller model, and the alternative is always the larger one. The test is therefore of $NH: E(y|x) = \eta_0 + \eta_1 \text{Year}$ versus $AH: E(y|x) = \eta_0 + \eta_1 \text{Year} + \eta_2 \text{Temp} + \eta_3 \text{Rainfall}$, and

$$F = [(2387.87 - 1766.44)/2]/28.0387 \sim F(2, 63)$$

The residual sum of squares under the null model must be obtained from the backward elimination output. The usual assumption is normality of the response given the predictors, Asymptotic justification for the test is available under the even weaker assumption that at worst the alternative hypothesis holds, and that residual variance is constant.

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.

The goal of subset selection as stated in class is to find coefficients that are effectively equal to zero (this can be framed differently). Backward elimination is a numerical algorithm for examining a few of the 2^k possible subset models that is likely to examine the “best” models. Starting with the full k -term mean function, terms are deleted one at a time so that the term removed causes the smallest increase in C_I , and estimate of the scaled mean square error of prediction. Better models have $C_I \approx k(I)$, the number of elements in subset I . Subset selection may be hard because of collinearity, computational issues, difficulty in making inferences, or other problems.

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.

Fit the usual regression with response y and x converted to a factor.

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.

Fit with weighted least squares, response is \bar{y}_i , term is x_i , and weights are m_i .

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.

The lowess smooth locally fits by estimating straight lines. If the smoothing parameter is too large, the overall fit tends to look like a straight line even if the true mean function is curved.

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.

Compute the cubic regression of $y - 4$ on $x - 10$ with no intercept.

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,

$$E(y|x) = \eta_0 + \eta_1 x \tag{1}$$

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.

$$E(y|x, z) = \eta_0 + \eta_1 x + \eta_2 z$$

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?

We have that

$$\begin{aligned} E(y|x) &= E(E(y|x, z)) \\ &= \eta_0 + \eta_1 x + \eta_2 E(z|x) \end{aligned}$$

This will give (1) if $E(z|x)$ is independent of x , or if $E(z|x)$ is linear in x .

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?

$$\begin{aligned} \text{var}(y|x) &= E(\text{var}(y|x, z)) + \text{var}(E(y|x, z)) \\ &= \sigma^2 + \eta_2^2 \text{var}(z|x) \end{aligned}$$

Since z is binary, the term $\text{var}(z|x)$ will be a constant only if $E(z|x)$ is constant.