

Reading assignment: Neter, Kutner, Nachtsheim & Wasserman, Chapter 13

Written assignment: Solutions will be posted on Friday, May 3.

Final Exam: On-campus students: Tuesday, May 7, 9:45 am-11:45 am., Physics Room 5.
 Distance students: as arranged.

1. For each of the following models, state whether it is a linear model, an intrinsically linear model, or a non linear model. In the case of an intrinsically linear model, indicate how it can be transformed to a linear model by describing a suitable transformation. In each case, ϵ_i , $i=1,2,\dots,n$, denote independent random errors.

(a) $Y_i = \exp(\beta_0 + \beta_1 X_i) \epsilon_i$

(b) $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$

(c) $Y_i = [1 + \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i)]^{-1}$

(d) $Y_i = \exp(\beta_0 + \beta_1 \log(X_{1i})) + \epsilon_i$

(e) $Y_i = \beta_0 X_{1i}^{b_1} X_{2i}^{b_2} e_i^{b_3}$

(f) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}^{\beta_3} + \epsilon_i$

2. The yield of a chemical process depends on temperature (X_1) and pressure (X_2). In one study, process yield (Y) was measured for $n=18$ runs of the process under various temperature and pressure conditions. The data are posted in the file `pnonlin.dat` with one line of data for each run. The variables are in the following order on each line.

Y = yield of the process

X_1 = temperature ($^{\circ}\text{C}$)

X_2 = pressure (psi)

Use least squares analysis to fit the following model

$$Y_i = \beta_0 X_{1i}^{b_1} X_{2i}^{b_2} + e_i$$

where $\epsilon_i \sim \text{NID}(0, \sigma^2)$ $i=1, \dots, 18$. Some code is posted as `pnonlin.ssc` and `pnonlin.sas` on the course web page.

- (a) To obtain starting values, ignore the additive random error in the nonlinear model shown above and use least squares to fit the model

$$\log(Y_i) = \log(\beta_0) + \beta_1 \log(X_{1i}) + \beta_2 \log(X_{2i}) + \text{error}$$

List your starting values for β_0 , β_1 , and β_2 .

- (b) Using the starting values from part (a), find the least squares estimates of the parameters in the non-linear model.
- (c) Make a plot showing the observed data and the estimated curve for the mean yield. Does this model appear to be adequate?
- (d) Plot the residuals against the values of X_1 and X_2 , on separate plots. What do these plots indicate?
- (e) Create a normal probability plot for the residuals. What does this plot suggest?
- (f) Regardless of your results in parts (c), (d) and (e), assume that the proposed non linear model is appropriate and use the large sample distributional properties of the least squares estimators to
- (i) Construct 95% confidence intervals for β_0 , β_1 and β_2 .
 - (ii) Construct a 95% confidence interval for the mean yield when $X_1 = 50$ and $X_2 = 60$.
 - (iii) Test the null hypothesis $H_0: \beta_1 = \beta_2$ versus the alternative $H_A: \beta_1 \neq \beta_2$. Indicate how your test was performed. State your conclusion.

3. In an enzyme kinetics study, the velocity of a reaction (Y) is expected to be related to the concentration (X) of a substance as follows:

$$Y_i = \frac{\mathbf{b}_0 X_i}{\mathbf{b}_1 + X_i} + \varepsilon_i$$

Observations were made at $n=18$ concentrations and the observed velocity of the reaction was recorded. That data are stored in the file posted as kinetics.dat on the course web page. There is one line for each run with Y_i preceding X_i on the i -th line. Some code is posted as kinetics.ssc and kinetics.sas on the course web page.

- (a) To obtain starting values, ignore the additive error and note that

$$Y_i^{-1} \approx \frac{1}{\mathbf{b}_0} + \frac{\mathbf{b}_1}{\mathbf{b}_0} X_i^{-1}.$$

Use least squares regression analysis to obtain estimates of $(\beta_0)^{-1}$ and (β_1/β_0) , and list your starting values for β_0 and β_1 .

- (b) Using the starting values from Part (a), find the least squares estimates of β_0 and β_1 in the original non-linear model.
 - (c) Make a plot showing the estimated mean velocity and the observed data. Does the model appear to be adequate?
 - (d) Plot the residuals against the concentration (X). What does this plot suggest?
 - (e) Construct a normal probability plot for the residuals. What does this plot suggest?
 - (f) Construct 95% confidence intervals for β_0 and β_1 .
 - (g) Estimate the concentration (X) at which the expected velocity of the reaction is 25. Report a 95% confidence interval for the concentration.
4. Suppose X_1, X_2, \dots, X_n is a random sample from a population with c.d.f. $F(x)$ and density function $f(x)$. Let θ denote the median of the population. Suppose $n = 2m$ is an even integer. Then, a consistent estimator for the median is obtained by ordering the observations from smallest to largest, $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, and computing the sample median

$$\hat{q}_n = (X_{(m)} + X_{(m+1)})/2$$

When the form of the c.d.f. is known, the exact finite sample distribution of \hat{q}_n can be obtained from the theory of order statistics. Unless n is very small, however, the evaluation of moments, such as the variance of \hat{q}_n , can be quite tedious. Standard asymptotic theory reveals that

$$n \text{Var}(\hat{q}_n) \approx \frac{1}{4[f(\theta)]^2} \quad \text{as } n \rightarrow \infty,$$

where $f(\theta)$ is the density function evaluated at θ , the true median for the population. You could use $(4n[f(\hat{q}_n)]^2)^{-1}$ as an estimate of $\text{Var}(\hat{q}_n)$, if you knew the form of the density function. The bootstrap provides a reasonably good estimator for the standard error of the sample median, as well as a confidence interval for the median, without any knowledge of the c.d.f. $F(x)$ or density $f(x)$. Basically, the bootstrap replaces the derivation of asymptotic theoretical results with simulation.

- (a) The file posted as `heart1.dat` on the course web page contains survival times (in days) for 69 patients who received heart transplants. There is one line of data for each patient and each line contains a survival time. Compute the sample median of the survival times. Use bootstrap methods to obtain a standard error for the sample median and 95% percentile confidence limits

for the population median. Use 5000 bootstrap samples. S-PLUS code for doing this is posted in the file heart.ssc.

- (b) Repeat part (a) for the survival times for a similar set of 34 patients that did not receive heart transplants these data are posted as heart2.dat. (S-PLUS code is available in the file posted as heart.ssc.)
 - (c) Use bootstrap methods to test the null hypothesis that the median survival time for patients who receive heart transplants is equal to the median survival time for similar patients who do not receive heart transplants. State your conclusion. (S-PLUS code is available in the file posted as heart.ssc.)
5. Consider the model you fit to the chemical process data in Problem 2. Use bootstrap methods to complete the following. Some S-PLUS code is posted on the course web page as pnonlin.ssc.
- (a) Construct individual 95% confidence intervals for β_0 , β_1 and β_2 . Use empirical bootstrap methods for each of the following resampling schemes:
 - i. Resample cases with replacement.
 - ii. Keep the model matrix fixed and resample centered residuals
 - iii. Keep the model matrix fixed and resample centered residuals that have also been re-scaled to approximately have the same variance as the actual random errors.
 - (b) Construct a 95% confidence interval for the expected yield of the process when it is run at a temperature of 100 °C and at a pressure of 50 psi.
 - (c) Construct a 95% prediction interval for the yield that will be observe when the process is at a temperature of 100 °C and at a pressure of 50 psi.