

If you have difficulty with any part, or if you need help, you can talk with me or with the course assistant, Harry Crane. The work handed in for these assignments should be your own. Collaboration on homeworks is not permitted.

Note: All GLM datasets are available on the web at www.stat.uchicago.edu/~pmcc/glm where GLMxxx.txt refers to the data on page xxx of the GLM book.

Read Applied Statistics, pages 1–50.

Questions 1, 2 and 5 are to be handed in for credit. The remainder are exercises that you should do, some of which may subsequently be used as examples in class.

1. Every household consists of a number of resident individuals, called wage-earners, plus other individuals (dependent children,...) who are not wage-earners. Each wage earner has an income, which might be zero. In a survey with households as sampling units, the response from each household is the list of wage earners together with their annual incomes. The household income is the sum of the incomes of the resident wage-earners.

In the following probability model, the set of households is infinite and the responses are independent and identically distributed for each household. With probability 1/3, the household contains one wage earner whose income is distributed as $N(100, 10)$; with probability 2/3, the household contains two wage earners (labelled in random order) whose incomes are bivariate normal with mean (75, 75), variances 10 and correlation 1/2.

- (i) The mean household income is _____.
- (ii) The variance of the household income is _____.
- (iii) The probability that the household income is more than 125 is _____.
- (iv) An individual i^* is selected uniformly at random from the set of wage earners occupying a large fixed finite population of households, and the individual income $Y(i^*)$ is recorded. Find the mean and variance of $Y(i^*)$.
- (v) An individual i^* is selected at random from a household chosen uniformly at random from a large fixed finite population of households, and the individual income $Y(i^*)$ is recorded. Find the mean and variance of $Y(i^*)$.

(vi) Denote by (m_i, Y_i) the number of wage earners and the household income for household i . The expected value of Y_i/m_i is _____.

Explain briefly the relation between your answers to (iv), (v) and (vi).

2. A factor is a function that associates with each statistical unit (patient, plot, subject,...) a value called the level of the factor. The definition does not exclude a quantitative covariate, but the terminology is most commonly used when the number of levels is finite, and usually moderately small. Association does not imply deliberative or random assignment, so age and sex are instances of factors, sometimes called classification factors by contrast with treatment factors. Consider the following design

unit u	1	2	3	4	5	6	7	8	9
A-level	3	2	2	3	1	2	3	1	1
B-level	1	1	1	2	2	2	3	3	3
Response y	6.74	6.34	5.57	10.86	0.23	4.73	8.57	3.05	2.18

The symbol $\mathbf{1}$ denotes the vector whose components are all one, i.e. $\mathbf{1}_u = 1$ for each unit. The indicator matrix $X = X(A)$ for a factor A is such that $X_{ul} = 1$ if the factor has level l on unit u , i.e. $A(u) = l$.

(i) Write out the model matrix for the linear model $y \sim A+1$, indicating which vectors or columns are omitted in the conventional parameterization in R or S . Write out the model matrix for the linear model $y \sim A+B$, indicating which columns are omitted in the conventional parameterization in R or S . (Omission of a column is equivalent to setting the corresponding coefficient to zero.)

When a factor has ordered levels, it is sometimes helpful to use a polynomial basis rather than the indicator basis. In the contrast matrices displayed below, the rows are the factor levels in increasing order, and the columns indicate polynomial degree, from zero up to the number of levels minus one. For factors having 2–4 levels, the traditional polynomial contrasts are as follows:

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & -3 & 1 & -1 \\ 1 & -1 & -1 & 3 \\ 1 & 1 & -1 & -3 \\ 1 & 3 & 1 & 1 \end{pmatrix}$$

For example, if $k = 4$ and the levels (rows) are denoted by $x = 1, 2, 3, 4$, the linear polynomial is $2x - 5$, the quadratic polynomial is $x^2 - 5x + 5$. These functions give the coefficients in the second and third columns of the 4×4 matrix shown above. Find the cubic polynomial for the last column. Comment briefly on the difference between these and the contrasts used by R or S .

(ii) Repeat part (i) using the polynomial basis instead of the indicator basis.

(iii) Explain how the model matrices in part (ii) are related to those produced by R when the levels are declared to be ordered as in `A <- ordered(A)`.

(iv) For each of the vector subspaces listed below, find the dimension of the space, calculate the squared length of the orthogonal projection of the response y onto each space, and the squared length of the orthogonal complement or residual.

$$\mathbf{1}, \quad A, \quad B, \quad A+B, \quad A.B.$$

(v) For the linear model $y \sim A$ (including $\mathbf{1}$ by default), explain how the coefficients for the conventional indicator basis with intercept are related to those for the polynomial basis.

(vi) Suppose that A is a treatment factor with unordered levels, such as patient medication in a pharmaceutical trial, and that B is a classification factor such as age or sex, or a block factor such as hospital or state or other geographic unit. What sum of squares would you use to test for the presence of a treatment effect, and how would you use it? Explain your reasoning, and compute the relevant statistic or p -value for the data shown above.

3. The file `corn.txt` shows the effect of nitrogen fertilizer on the yield of corn. The experiment was arranged in three blocks or contiguous areas, most likely three parts of the same field, but possibly three different fields. The first column gives the amount of fertilizer used in lbs/acre, and the three columns give the observed yield in lbs per plot for the three blocks.

(i) Calculate the sum of squares associated with variation between blocks, the sum of squares associated with nitrogen, and the residual sum of squares associated with departures from additivity.

(ii) The purpose of blocking is to eliminate recognizable variation between plots, due for example to different soil composition or differences of exposure to sun or rain. Has blocking been effective in this experiment?

(iii) The response to nitrogen appears to be smooth but non-linear. Illustrate this by a suitable plot. Fit the linear model `block + beta log(gamma + x)` for various values of γ in the range 14–20. Plot the residual sum of squares against γ , find the least squares estimate, and compute a 95% confidence interval.

(iv) Test whether the logarithmic function in (iii) is adequate for the relation between nitrogen level and response. What are the appropriate sums of squares to be compared for this purpose?

(v) Can the model in (iii) be substantially improved by response transformation? Check this in two ways, first using Tukey's test and second using the Box-Cox method.

(vi) Give a brief summary of your conclusions.

4. The file `oats.txt` contains data on the Nickel-iron content y of 30 oat plants at various ages x from 4 to 73 days. The plants were grown in sand cultures.

(1) Plot the data and comment on the relationship, if any.

(2) Fit a quadratic model and superimpose the fitted line on the scatterplot.

(3) Use the fitted model to estimate the age at which the nickel-iron content is maximum.

(4) Examine the residuals and comment briefly.

5. The following data were collected as part of a high-school electro-chemical experiment by P. Ohtani. To obtain an observation, two metals i, j , were inserted into an electrolytic solution, and the voltage difference Y_{ij} between i and j recorded by a digital voltmeter. The voltage difference between i and j is, by definition, the negative of the difference between j and i , so each observation is measured once and recorded twice.

(i) A circuit is a closed loop, $i_0, \dots, i_{n-1}, i_n = i_0$ of length $n \geq 0$. Conservation of energy is a condition on the $k \times k$ matrix V to the effect that, on each circuit the sum is zero

$$V(i_0, i_1) + V(i_1, i_2) + \dots + V(i_{n-1}, i_0) = 0.$$

A matrix satisfying this condition is called conservative. Show that each conservative matrix is skew-symmetric. Deduce that the set of conservative matrices is a vector space, closed under vector-space operations. Exhibit a 3×3 skew-symmetric matrix that is not conservative. A skew-symmetric matrix of the form $V(i, j) = \alpha_i - \alpha_j$ is called additive. Prove that every additive matrix is conservative. Prove that every conservative matrix is additive. What is the dimension of the vector space of conservative 6×6 matrices?

The following exercises refer to the linear model for the voltages in which $E(Y_{ij}) = \alpha_i - \alpha_j$ is conservative.

(ii) For a single $k \times k$ table, obtain an expression for the least-squares estimate of α . Use this formula to compute $\hat{\alpha}$ for each of the three electrolytes. Explain why $(\alpha_1, \dots, \alpha_5)$ and $(\alpha_1, \dots, \alpha_5) + (c, c, c, c, c)$ are equivalent as parameter points in the model.

(iii) Assess the evidence for and against the hypothesis that the vector of potentials is constant across electrolytes. That is to say, fit the linear model in which the potentials are constant across electrolytes, and compare the fit with the model in which α varies from one electrolyte to another. Obtain the relevant sums of squares, their degrees of freedom, and compute the appropriate F -statistic.

Electrolyte O

	Mg	Zn	Fe	Pb	Cu
Mg	0.0	0.414	0.807	0.876	1.291
Zn	-0.414	0.0	0.429	0.533	0.886
Fe	-0.807	-0.429	0.0	0.043	0.377
Pb	-0.876	-0.533	-0.043	0.0	0.271
Cu	-1.291	-0.886	-0.377	-0.271	0.0

Electrolyte A

Mg	0.0	0.247	0.856	1.051	1.402
Zn	-0.247	0.0	0.434	0.521	0.867
Fe	-0.856	-0.434	0.0	0.058	0.443
Pb	-1.051	-0.521	-0.058	0.0	0.374
Cu	-1.402	-0.867	-0.443	-0.374	0.0

Electrolyte K

Mg	0.0	0.443	0.895	0.973	1.281
Zn	-0.443	0.0	0.477	0.503	0.856
Fe	-0.895	-0.477	0.0	0.107	0.432
Pb	-0.973	-0.503	-0.107	0.0	0.392
Cu	-1.281	-0.856	-0.432	-0.392	0.0

Data in `glm/electro_chem.txt`.

(iv) Discuss briefly the arguments for and against analysis of these data by linear models after transformation.

6. GLM Exercise 3.11.

7. GLM Exercise 12.3.

8. Study Examples J, P, Q in Applied Statistics. This example may be used in class discussion: be prepared to contribute to the discussion in class. For example J, compute the Yates decomposition by orthogonal polynomial contrasts for both the original and the transformed scale. Compare the half-normal plots and comment on any differences.

Are the data consistent with the model $(x_1/x_2)^5 x_3^{-7/2}$?