If you have difficulty with any part, or if you need help, you can talk with me or with the course assistants, Han Han or Joe Guinness. 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.dat` 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.** A covariate $x$ is a function that associates with each statistical unit $u$ a value $x(u) \equiv x_u$. If the values are in the real line or in a vector space (where they may be added or subtracted), $x$ is said to be a *quantitative* covariate. Otherwise, if $x$ takes values in a finite set of levels, $x$ is said to be a factor, or a *qualitative factor* usually denoted by $A, B, \ldots$. The levels may be ordered or unordered. A *treatment factor* has levels that can, in principle, be assigned to the units by the experimenter; A *classification factor* is an intrinsic property of the units over which the experimenter has no control.

A relation $R$ is a function on pairs of units, usually symmetric, i.e, $R(u, u') = R(u', u)$ for every ordered pair. The identity $\delta_{u,u'}$ (Kronecker's delta) is a relation; a metric is an example of a real-valued relation on the points in a space. An equivalence relation is a binary or Boolean function that is (i) symmetric, (ii) reflexive $R(u, u) = 1$ for every $u$, and (iii) transitive meaning that $R(u, u') = R(u', u'') = 1$ implies $R(u, u'') = 1$. In factorial models, a block factor is an equivalence relation, usually coded as a list of block labels rather than a matrix.

(a) In a randomized experiment, one usually insists on covariates being measured prior to treatment assignment. Explain briefly, i.e. in 1-2 sentences, the reasons for this.

(b) A randomized pharmaceutical trial with human subjects is conducted on patients recruited at nine participating hospitals (sites) across the U.S. Numerous baseline variables are measured at recruitment, including age, race, sex and marital status. Patients are assigned at random to one of three treatment groups, control, half dose, full dose. Patients are volunteers who have given informed consent to participate, but are not told of their treatment status. The response is 5-year survival. How would you classify the variables *site*, *age*, *race*, *sex*, *marital status*. One or two sentences of explanation is sufficient for each.

(c) 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 |

in which $A, B$ are qualitative factors. The symbol $\mathbf{1}$ denotes the vector whose components are all one, i.e. $\mathbf{1}_u = 1$ for each unit, or the associated one-dimensional subspace of constant functions. 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 matrices for the linear models $y \sim A+1$ and $y \sim A-1$, indicating which vectors or columns are omitted in the conventional parameterization in $R$ or $S$. Explain why, or in what sense, the two models are equivalent. 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.)

(ii) Each symbol appearing in a model formula refers to a vector subspace of $\mathcal{R}^n$ with a particular basis implied. Repeat part (i) with $A = x$ treated as a quantitative covariate and $B$ as a qualitative factor.

(iii) 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.

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

(iv) 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. 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.

**2.** Eight isolates of rose blackspot fungus (from eight different areas) were grown for 20 days at seven different temperatures ranging from 55°F to 85°F. The log weights in mg are reported in the file `fungus.txt`.

(i) Complete the decomposition of the total sum of squares:

| Source | SS | df | MS | F | $p$-value(%) |
|---|---|---|---|---|---|
| Isolate | | 7 | | | 0.19% |
| Temp | | 6 | | | |
| Resid | 0.4934 | 42 | | | |

MS denotes the mean square, and F is the relevant $F$-ratio.

(ii) Complete the decomposition of the sum of squares due to temperature

| Source | SS | df | MS | F | $p$-value(%) |
|---|---|---|---|---|---|
| P1/1 | 0.0117 | 1 | 0.0117 | | |
| P2/P1 | | | | | |
| P3/P2 | | | | | |
| P4/P3 | | | | | |
| Temp/P4 | | | | | |
| Temp/P3 | 0.0439 | 4 | | | |
| Total Temp | | 6 | | | |
| Resid | 0.4934 | 42 | | | |

Here $P_r$ denotes the subspace of polynomials of degree $r$ or less in temperature, $P_0 \equiv 1$ is the one-dimensional subspace of constant functions, $1 \subset P_1 \subset \cdots \subset P_6 \equiv \mathit{Temp}$, and $P_4/P_3$ is the quotient space. If $\mathcal{X} \subset \mathcal{X}' \subset \mathcal{R}^n$ are two subspaces, the sum of squares associated with $\mathcal{X}$ is the squared length of the projection, $\|P_{\mathcal{X}} y\|^2$, and the sum of squares associated with $\mathcal{X}'/\mathcal{X}$ is the additional squared length $\|P_{\mathcal{X}'} y\|^2 - \|P_{\mathcal{X}} y\|^2$.

(iii) What does the preceding table tell you about the effect of temperature on growth? Estimate the temperature at which the growth rate is a maximum.

**3.** 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 recorded twice.

(i) A circuit is a closed loop, $i_0, \ldots, i_{n-1}, i_n = i_0$ 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

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

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 $3 \times 3$ 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 $6 \times 6$ 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 $\alpha$. Use this formula to compute $\hat{\alpha}$ for each of the three electrolytes. Explain why $(\alpha_1, \ldots, \alpha_5)$ and $(\alpha_1, \ldots, \alpha_5) + (c, c, c, c, c)$ 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 $\alpha$ varies from one electrolyte to another. Obtain the relevant sums of squares, their degrees of freedom, and compute the appropriate $F$-statistic.

| | Mg | Zn | Fe | Pb | Cu |
|---|---|---|---|---|---|
| Electrolyte O | | | | | |
| 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.dat`.

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

**4.** GLM Exercise 1.6.

**5.** GLM Exercise 3.11.

**6.** 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 $(x_1/x_2)^5 x_3^{-7/2}$?