Midterm

Stat 623

Due on Nov 05, 2020

- **Answer all questions. 3(e) is for extra credit.**
- This is a take home examination. Be sure to make a photocopy in case the original gets lost. This is not a collaborative exam, so please do not discuss any part of this exam or your answer with another person.
- In order to receive full credit for a problem, you should show all of your work and explain your reasoning. Good work can receive substantial partial credit even if the final answer is incorrect.
- Attach codes at the end.

1. This question is about the analysis of the data *birth.txt*. The response is the mean interval in months between successive births in the same family. Thus a family of size k gives rise to $k-1$ intervals, $1-2, 2-3, \ldots$, each of which corresponds to a particular sex sequence MM, MF, FM or FF (M=Male, F=Female). Any systematic pattern is of potential interest.

   (a) Calculate the residual sum of squares for the linear regressions (additive models)

      i. response $\sim$ family size + birth order + sex sequence

      ii. response $\sim$ family size + reverse birth order + sex sequence

      iii. response $\sim$ family size * birth order + sex sequence

      iv. response $\sim$ family size * reverse birth order + sex sequence

   (b) Comment on the difference between the first two models, and how this difference affects the interpretation.

   (c) Comment also on any unexpected or unusual effects. As part of this analysis, you should first check to see whether transformation might be helpful.

(d) Give a one-paragraph summary in your own words of the main patterns of variation in these data.

2. The following Table provides counts for a truncated sample of 486 galaxies, binned by redshift (x_1) and brightness (x_2). Distance from earth is an increasing function of redshift, while apparent brightness of a galaxy is a decreasing function of magnitude. In this survey, counts were limited to galaxies having

$$1.22 \leq x_1 \leq 3.22, \qquad 17.2 \leq x_2 \leq 21.5.$$

The upper limit of x_1 and the lower limit of x_2 reflect the difficulty of measuring very dim and distant galaxies. The range of redshift has been divided into 15 equal bins and likewise 18 equal intervals for brightness

		redshift(further) $\longrightarrow$														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	18	1	6	6	3	1	4	6	8	8	20	10	7	16	9	4
	17	3	2	3	4	0	5	7	6	6	7	5	7	6	8	5
	16	3	2	3	3	3	2	9	9	6	3	5	4	5	2	1
	15	1	1	4	3	4	3	2	3	8	9	4	3	4	1	1
	14	1	3	2	3	3	4	5	7	6	7	3	4	0	0	1
	13	3	2	4	5	3	6	4	3	2	2	5	1	0	0	0
	12	2	0	2	4	5	4	2	3	3	0	1	2	0	0	1
	11	4	1	1	4	7	3	3	1	2	0	1	1	0	0	0
Brightness	10	1	0	0	2	2	2	1	2	0	0	0	1	2	0	0
$\uparrow$	9	1	1	0	2	2	2	0	0	0	0	1	0	0	0	0
	8	1	0	0	0	1	1	0	0	0	0	1	1	0	0	0
	7	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
	6	0	0	3	1	1	0	0	0	0	0	0	0	0	0	0
	5	0	3	1	1	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

(a) Use Poisson regression with different functions of redshift (x_1) and brightness (x_2) and evaluate their fit to the data using plots of the fitted mean, and appropriate residual plots. In particular evaluate the following mean models on the log scale.

model 1	linear in x_1 and x_2
model 2	quadratic in x_1 and x_2
model 3	cubic in x_1 and x_2

(b) Based on the plots in 2(a), estimated regression coefficients and AIC values, which model(s) appear appropriate in order to characterize the rate (or density) of galaxies as a function of redshift and brightness? Also provide a summary of the fit of the model(s) that appear most appropriate.

3. Byssinosis is an occupational lung disease caused by exposure to cotton dust in inadequately ventilated working environments. The file *cotton.txt* contains information, obtained from a survey conducted by a textile company, on the prevalence of byssinosis. The file lists the observed prevalence of byssinosis (affected, not affected), by race (white = 1, non white = 2), sex (male = 1, female = 2), smoking habits (two levels), length of employment (three levels), and dustiness of the work environment (three levels). In the last three cases, higher-numbered categories denote larger values (more smoking, longer employment and increased dustiness). Parts (a) and (b) are based on the assumption that the main-effects linear logistic model is substantially correct.

(a) Fit the main-effects linear logistic model and summarize the results. Explain how the residual degrees of freedom is calculated for the deviance.

(b) Interpret the regression coefficient of sex. Construct an approximate 90% confidence interval for the odds ratio (males vs females) of contracting byssinosis.

(c) Estimate the excess risk (i.e. the difference between the risk of an outcome in two groups) associated with cotton dust and provide approximate confidence intervals. How fast does the risk increase with dust level? If necessary, give separate figures for males and females or for smokers and non-smokers.

(d) What conclusions can you draw from the analysis?

(e) Beginning with the complete main-effects model, look for significant interactions. For example you can try fitting a model with main effects + all pairwise interactions. After detecting the significant main effects and interactions, interpret the model thus obtained.

4. Suppose x follows a Beta distribution Beta$(r\nu, r(1-\nu))$, r known,

$$f(x) = \frac{\Gamma(r)}{\Gamma(r\nu)\Gamma(r(1-\nu))} x^{r\nu-1}(1-x)^{r(1-\nu)-1}, \quad 0 < x < 1.$$

Write the above density in the exponential family form

$$f(y) = \exp\Big\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big\}.$$

Furthermore, consider generalized regression (GLM) associated with

$$\mu = b'(\theta), \quad g(\mu) = \eta = z^T\beta.$$

Identify the relevant components necessary for use in a GLM: (1) the quantity y, (2) the canonical parameter θ, (3) the dispersion parameter ϕ, (4) the functions $a(.), b(.), c(.)$, (5) the mean μ, and (6) the canonical link function $g(.)$. It is OK if you can not solve all these functions explicitly. Providing some implicit formulas and relationships will suffice.