

1. (20 pts.) Let  $Y$  = political ideology (on an ordinal scale from 1 = very liberal to 5 = very conservative),  $x_1$  = gender (1 = female, 0 = male),  $x_2$  = political party (1 = Democrat, 0 = Republican).
- (a) A main effects model with a cumulative logit link gives the output shown. Explain why the output reports four intercepts.

Parameter		DF	Estimate	Standard Error	Wald	95% Confidence Limits
Intercept1		1	-2.5322	0.1489	-2.8242	-2.2403
Intercept2		1	-1.5388	0.1297	-1.7931	-1.2845
Intercept3		1	0.1745	0.1162	-0.0533	0.4023
Intercept4		1	1.0086	0.1232	0.7672	1.2499
gender	female	1	0.1169	0.1273	-0.1327	0.3664
gender	male	0	0.0000	0.0000	0.0000	0.0000
party	democ	1	0.9636	0.1297	0.7095	1.2178
party	repub	0	0.0000	0.0000	0.0000	0.0000

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
gender	1	0.84	0.3586
party	1	56.85	<.0001

- (b) Explain how to describe gender effect on political ideology with an odds ratio.

- (c) Give the hypotheses to which the LR statistic for gender refers, and explain

how to interpret the result of the test.

- (d) When we add an interaction term to the model, we get the output shown. Explain how to find the estimated odds ratio for the gender effect on political ideology for Republicans.

Parameter		DF	Estimate	Standard Error
Intercept1		1	-2.6743	0.1655
Intercept2		1	-1.6772	0.1476
Intercept3		1	0.0424	0.1338
Intercept4		1	0.8790	0.1389
gender	female	1	0.3661	0.1784
gender	male	0	0.0000	0.0000
party	democ	1	1.2653	0.1995
party	repub	0	0.0000	0.0000
gender*party	female democ	1	-0.5091	0.2550
gender*party	female repub	0	0.0000	0.0000
gender*party	male democ	0	0.0000	0.0000
gender*party	male repub	0	0.0000	0.0000

2. (10 pts) You decide to use GEE methods to handle dependent observations because of repeated measurement or clustering of some type.

a. Explain what is meant by a “working correlation matrix.”

b. If you ignore the dependence, will there be bias in your (i) parameter estimates,

(ii) standard error estimates?

3. (30 pts.) The attached printouts at the end of the exam show results of fitting loglinear models and logit models to a data set on  $D$  = defendant's race,  $V$  = victim's race, and  $P$  = death penalty verdict. The data refer to cases in Florida between 1976 and 1987 in which the defendant was convicted of murder.

(a) First consider the loglinear model, denoted by  $(DP, DV, PV)$ , having all the two-factor associations. Describe the association structure implied by this model, in terms of conditional odds ratios relating the three pairs of variables. (i.e., describe characteristics of the model itself, in terms of how it smooths the data, not the particular fit obtained for this data set.)

(b) Show how to test the goodness of fit of model  $(DP, DV, PV)$ . State the hypotheses, test statistic and df value, indicate whether the P-value for the test statistic would be small enough to reject the null hypothesis, and interpret the result.

(c) For the fit of this model, report a summary measure of the association between

the death penalty verdict and victim's race, controlling for defendant's race. Explain how to construct a 95% confidence interval for its population value.

(d) Show how to test the hypothesis of conditional independence between the death penalty verdict and victim's race, controlling for defendant's race, by comparing the fit of the model in (a) to another loglinear model. Interpret.

(e) Now suppose we prefer to treat the death penalty verdict  $P$  as a response variable and  $D$  and  $V$  as predictors. Let  $\pi = P(P = 1)$  denote the probability that  $P$  takes value "yes." Set up dummy variables as SAS does in the printout, and use the results in the printout to write the prediction equation (using the ML estimates) for the logit model with those dummies that is equivalent to

loglinear model (DP, DV, PV).

- (f) Explain why it is more natural to analyze these data using logit models than using loglinear models.
4. (6 pts.) Consider the loglinear model of independence for a two-way contingency table. This has equation for expected frequencies  $\{\mu_{ij}\}$  in an  $I \times J$  contingency table,

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y.$$

Motivate this model, by showing how the definition of statistical independence of two categorical variables implies that a loglinear model of this form holds.

5. (5 pts.) Consider the baseline-category logit model, for a multinomial response variable having  $J$  categories,

$$\log[P(Y = j)/P(Y = J)] = \alpha_j + \beta_j x, \quad j = 1, \dots, J - 1.$$

Show how to use this model to generate a related logit model for  $\log[P(Y = a)/P(Y = b)]$  using an arbitrary pair  $a$  and  $b$  of the response categories.

6. For the following questions, answer true (T) or false (F). (2 pts. each)

- (a) \_\_\_\_\_ For General Social Survey data on  $Y =$  political ideology (categories liberal, moderate, conservative),  $X_1 =$  gender (1 = female, 0 = male), and  $X_2 =$  political party (1 = Democrat, 0 = Republican), the ML fit of the cumulative logit model is  $\text{logit}[\hat{P}(Y \leq j)] = \hat{\alpha}_j + .12x_1 + .96x_2$ . Hence, for each gender, according to this model fit the estimated odds that a Democrat's response is liberal rather than moderate or conservative, and the estimated odds that a Democrat's response is liberal or moderate rather than conservative, is  $e^{.96} = 2.6$  times the corresponding estimated odds for a Republican's response. This odds ratio estimate indicates that in this sample Democrats tended to be more liberal than Republicans.
- (b) \_\_\_\_\_ Subjects suffering from mental depression are measured after 1 week of treatment, 2 weeks of treatment, and 4 weeks of treatment in terms of a (normal, abnormal) response outcome. Covariates are severity of condition at original diagnosis (1 = severe, 0 = mild) and treatment used (1 = new, 0 = standard). Since each subject contributes three observations to the analysis, we can use the GEE (generalized estimating equations) method to fit the model. To use this method, we must choose a "working" correlation matrix for the form of the dependence among the three responses, but the method is robust in the sense that it still gives appropriate estimates and standard errors for large  $n$  even if the actual correlation structure is somewhat different from the one we assumed.
- (c) \_\_\_\_\_ A difference between logit and loglinear models is that the logit model is a generalized linear model assuming a binomial random component whereas the loglinear model is a generalized linear model assuming a Poisson random component. Hence, when both are fitted to a contingency table having 50 cells, the logit model treats the cell counts as 25 binomial observations whereas the loglinear model treats the cell counts as 50 Poisson observations.

- (d) \_\_\_\_\_ When you have an ordinal response, one reason it is usually better to use an model designed for ordinal variables such as the cumulative logit model rather than a model designed for nominal variables such as the baseline-category logit model is that significance tests focus the effect on a smaller number of degrees of freedom, which tends to give greater power (e.g., making it easier to get smaller P-values).

7. Homework (20 pts.)

### OUTPUT FOR PROBLEM 3

```
data loglin;
input v $ d $ p $ count @@;
datalines;
white white yes 53   white white no 414
white black yes 11   white black no 37
black white yes 0    black white no 16
black black yes 4    black black no 139
;
proc genmod; class v d p;                                * Model 1
  model count = v d p v*d d*p
    / dist=poi link=log;
proc genmod; class v d p;                                * Model 2
  model count = v d p v*d d*p v*p
    / dist=poi link=log obstats;

data logistic;
input v $ d $ pen_yes total @@;
datalines;
w w 53 467
w b 11 48
b w 0 16
b b 4 143
;
proc genmod order=data data=logistic; class d v;
  model pen_yes/total = d v / dist=bin link=logit ; * Model 3
run;
```

Model 1:

#### Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	20.7298	10.3649
Pearson Chi-Square	2	22.1413	11.0707

Model 2:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	0.3798	0.3798
Pearson Chi-Square	1	0.1978	0.1978

Parameter		DF	Estimate	Std Err	ChiSquare
INTERCEPT		1	3.9668	0.1374	833.7843
V	black	1	-5.6696	0.6459	77.0565
V	white	0	0.0000	0.0000	.
D	black	1	-1.5525	0.3262	22.6556
D	white	0	0.0000	0.0000	.
P	no	1	2.0595	0.1458	199.3972
P	yes	0	0.0000	0.0000	.
V*D	black black	1	4.5950	0.3135	214.7836
V*D	black white	0	0.0000	0.0000	.
V*D	white black	0	0.0000	0.0000	.
V*D	white white	0	0.0000	0.0000	.
D*P	black no	1	-0.8678	0.3671	5.5889
D*P	black yes	0	0.0000	0.0000	.
D*P	white no	0	0.0000	0.0000	.
D*P	white yes	0	0.0000	0.0000	.
V*P	black no	1	2.4044	0.6006	16.0264
V*P	black yes	0	0.0000	0.0000	.
V*P	white no	0	0.0000	0.0000	.
V*P	white yes	0	0.0000	0.0000	.

Model 3:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	0.3798	0.3798
Pearson Chi-Square	1	0.1978	0.1978

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-3.5961	0.5069	50.3264	0.0000
D	w	1	-0.8678	0.3671	5.5889	0.0181
D	b	0	0.0000	0.0000	.	.
V	w	1	2.4044	0.6006	16.0264	0.0001
V	b	0	0.0000	0.0000	.	.

### Formulas for exam 3:

Baseline-category logit:  $\log[P(Y = j)/P(Y = J)] = \alpha_j + \beta_j x$

$$P(Y = j) = \frac{e^{\alpha_j + \beta_j x}}{1 + e^{\alpha_1 + \beta_1 x} + \dots + e^{\alpha_{J-1} + \beta_{J-1} x}}, \quad j = 1, 2, \dots, J - 1.$$

Cumulative logit:  $\text{logit} [P(Y \leq j)] = \alpha_j + \beta x$

$$P(Y \leq j) = \exp(\alpha_j + \beta x) / [1 + \exp(\alpha_j + \beta x)], \quad j = 1, 2, \dots, J - 1.$$

$$z = (n_{12} - n_{21}) / \sqrt{n_{12} + n_{21}} \quad (\text{McNemar})$$

$$(XY, XZ, YZ) : \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

$$\text{Independence} : \log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

$$\kappa = \frac{\sum_i \pi_{ii} - \sum_i \pi_{i+} \pi_{+i}}{1 - \sum_i \pi_{i+} \pi_{+i}}$$

For comparing proportions with  $n$  matched pairs and counts  $b$  and  $c$  for numbers of different outcomes for the two observations, difference of sample proportions has estimated standard error

$$\frac{\sqrt{(b+c) - (b-c)^2/n}}{n}$$