STA 4504/5503
Sample questions for exam 2

1. For the 23 space shuttle flights that occurred before the Challenger mission in 1986, Table 1 shows the temperature ($^\circ F$) at the time of the flight and whether at least one of the six primary O-rings suffered thermal distress (1 = yes, 0 = no). The first attached SAS printout shows the use of various models for analyzing these data.

   (a) For the logistic regression model using temperature as a predictor for the probability of thermal distress, calculate the estimated probability of thermal distress at 31°, the temperature at the time of the Challenger flight.

   (b) At the temperature at which the estimated probability equals 0.5, give a linear approximation for the change in the estimated probability per degree increase in temperature.

   (c) Interpret the estimated effect of temperature on the odds of thermal distress.

   (d) Test the hypothesis that temperature has no effect, using the likelihood-ratio test. Interpret results.

(e) Suppose we treat the (0, 1) response as if it has a normal distribution, and fit a linear model for the probability. Report the prediction equation, and find the estimated probability of thermal distress at 31°. Comment on the suitability of this model.

2. For the horseshoe crab data, in class we modeled the effects of color and weight on the presence of satellites. We now consider the effects of color and the width of the carapace shell. The second attached SAS printout shows the use of various models for analyzing these data, where 'satell' denotes the number of satellites for a female horseshoe crab.

   (a) First, suppose we model the number of satellites directly. Using the prediction equation for the Poisson loglinear model with width as the predictor, find the estimated expected number of satellites for a crab having width 25.0 *cm*.

   (b) Explain how to interpret the estimated effect of a 1-cm increase in width on the expected number of satellites for a female horseshoe crab.

(c) Suppose that instead of having a separate observation for each female horseshoe crab, the data file had summarized the data in the following way:

Width	Number of crabs	Number of satellites
20	11	12
21	14	20
22	21	38
etc.		

Explain how one could then model the expected number of satellites per horseshoe crab as a function of width.

(d) Now consider a logistic regression model for the response, "whether a female crab has at least one satellite," with $Y = 1$ for 'yes' and $Y = 0$ for 'no.' Include both color and width in the model. Carefully show how to set up dummy variables for four colors (i.e., define the dummy variables, as software would set them up for you). Report the prediction equation.

(e) Using the odds ratio, interpret the parameter estimate for the coefficient of the dummy variable for the first color in this prediction equation.

(f) Using the prediction equation from part (b), construct a prediction equation for the probability $\hat{\pi}$ (i.e., not for the logit of $\hat{\pi}$) of a satellite for crabs of the first color, as a function of width.

(g) Construct and interpret a 95% confidence interval for the effect of a 1-cm increase in width on the odds that a crab has a satellite, controlling for color.

3. True-False questions.

(a) _____ An ordinary regression model that treats the response Y as having a normal distribution is a special case of a generalized linear model, with normal random component and identity link function. As a result, one can do ordinary regression and ANOVA using software (such as PROC GENMOD in SAS) that is designed to fit generalized linear models.

(b) _____ One question in the 1990 General Social Survey asked subjects how many times they had had sexual intercourse in the previous month. The sample means were 5.9 for the male respondents and 4.3 for the female respondents; the sample variances were 54.8 and 34.4. The modal response for each gender was 0. For a Poisson GLM using gender as a pre-

dictor, there is severe evidence of overdispersion, suggesting that negative binomial regression would be better.

(c) _____ For a sample of retired subjects in Florida, a contingency table is used to relate X = cholesterol (8 ordered levels) to Y = whether the subject has symptoms of heart disease (yes = 1, no = 0). For the linear logit model $\text{logit}[P(Y = 1)] = \alpha + \beta x$ fitted to the 8 binomials in the 8×2 contingency table by assigning scores to the 8 cholesterol levels, the deviance statistic equals 6.0. Thus, this model provides a poor fit to the data.

(d) _____ In the example mentioned in (c), at the lowest cholesterol level, the observed number of heart disease cases equals 31. The adjusted residual equals 1.35. This means that the model predicted 29.65 cases (i.e., 1.35 = 31 - 29.65).

4. Multiple choice question. Circle the letter(s) for the correct response(s). More than one response may be correct.

   Let π denote the probability that a randomly selected respondent supports current laws legalizing abortion, predicted using gender of respondent ($G = 0$, male; $G = 1$, female), religious affiliation ($R_1 = 1$, Protestant, 0 otherwise; $R_2 = 1$, Catholic, 0 otherwise; $R_1 = R_2 = 0$, Jewish), and political party affiliation ($P_1 = 1$, Democrat, 0 otherwise; $P_2 = 1$, Republican, 0 otherwise, $P_1 = P_2 = 0$, Independent). The logit model with main effects has prediction equation

$$\text{logit}(\hat{\pi}) = .11 + .16G - .57R_1 - .66R_2 + .47P_1 - 1.67P_2$$

   For this prediction equation,

   a. Females are more likely than males to support legalized abortion, controlling for religious affiliation and political party affiliation.

   b. Controlling for gender and religious affiliation, the estimated odds that a Democrat supports legalized abortion equal $e^{.47-(-1.67)}$ times the estimated odds that a Republican supports legalized abortion.

   c. The estimated probability that a male Jewish Independent supports legalized abortion equals $e^{.11}/(1 + e^{.11})$.

   d. The estimated probability of supporting legalized abortion is highest for female Jewish Independents.

Table 1. Space shuttle data

Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD
1	66	0	2	70	1	3	69	0	4	68	0
5	67	0	6	72	0	7	73	0	8	70	0
9	57	1	10	63	1	11	70	1	12	78	0
13	67	0	14	53	1	15	67	0	16	75	0
17	70	0	18	81	0	19	76	0	20	79	0
21	75	1	22	76	0	23	58	1			

Note: Ft = flight no., Temp = temperature, TD = thermal distress (1 = yes, 0 = no). Data based on Table 1 in *J. Amer. Statist. Assoc*, 84: 945-957, (1989), by S. R. Dalal, E. B. Fowlkes, and B. Hoadley.

```
data shuttle;
input temp distress  @@;  n=1;
cards;
66 0   70 1  69 0   68 0   67 0   72 0   73 0   70 0
57 1   63 1   70 1   78 0   67 0   53 1   67 0  75 0
70 0   81 0   76 0   79 0   75 1   76 0   58 1
proc genmod;
  model distress/n = temp / dist=bin  link=logit ; * model 1;
proc genmod;
  model distress/n =   / dist=bin   link=logit; * model 2;
proc genmod;
  model distress = temp  /  dist=nor  link=identity; * model 3;
```

Results of model fitting for space shuttle data:

Model 1

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	21	20.3152	0.9674
Pearson Chi-Square	21	23.1691	1.1033
Log Likelihood	.	-10.1576	.

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	15.0429	7.3786	4.1563	0.0415
TEMP	1	-0.2322	0.1082	4.6008	0.0320

Model 2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	22	28.2672	1.2849
Pearson Chi-Square	22	23.0000	1.0455
Log Likelihood	.	-14.1336	.

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-0.8267	0.4532	3.3278	0.0681

Model 3

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	21	3.3386	0.1590
Pearson Chi-Square	21	3.3386	0.1590
Log Likelihood	.	-10.4411	.

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	2.9048	0.8046	13.0323	0.0003
TEMP	1	-0.0374	0.0115	10.5473	0.0012

Results of model fitting for horseshoe crab data:

```
data crab;
input color spine width satell weight;
if satell>0 then y=1; if satell=0 then y=0; n=1;
cards;
 2  3  28.3  8  3050
 3  3  22.5  0  1550
 1  1  26.0  9  2300
 .... (etc. for 173 lines)
    ;
proc genmod;
  model satell =  / dist=poi  link=log; * model 1;
proc genmod;
  model satell = width / dist=poi link=log; * model 2;
proc genmod;
  model y/n =  / dist=bin link=logit; * model 3;
proc genmod;
  model y/n = width / dist=bin  link=logit obstats; * model 4;
proc genmod;  class color;
   model y/n = width color / dist=bin link=logit; * model 5;
run;
```

```
Model 1
                 Criteria For Assessing Goodness Of Fit

          Criterion             DF         Value      Value/DF

          Deviance             172      632.7917        3.6790
          Pearson Chi-Square   172      584.0436        3.3956
          Log Likelihood         .       35.9898             .

                    Analysis Of Parameter Estimates

      Parameter    DF    Estimate     Std Err   ChiSquare  Pr>Chi

      INTERCEPT     1      1.0713      0.0445    579.5444  0.0000

-------------------------------------------------------------------
```

Model 2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	171	567.8786	3.3209
Pearson Chi-Square	171	544.1570	3.1822
Log Likelihood	.	68.4463	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-3.3048	0.5422	37.1444	0.0000
WIDTH	1	0.1640	0.0200	67.5107	0.0000

Model 3

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	172	225.7585	1.3125
Pearson Chi-Square	172	173.0000	1.0058
Log Likelihood	.	-112.8793	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	0.5824	0.1585	13.4929	0.0002

Model 4

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	171	194.4527	1.1372
Pearson Chi-Square	171	165.1434	0.9658
Log Likelihood	.	-97.2263	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-12.3508	2.6287	22.0749	0.0000
WIDTH	1	0.4972	0.1017	23.8872	0.0000

Observation Statistics

Y	N	Pred	Lower	Upper	Resraw	Reschi	Resdev
1	1	0.84823287	0.7528514	0.91114836	0.15176713	0.42299119	0.57375966
0	1	0.2380991	0.12998546	0.39527988	-0.2380991	-0.559023	-0.7374806
1	1	0.6404177	0.55808923	0.71523445	0.3595823	0.74932029	0.94407061

Model 5

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	168	187.4570	1.1158
Pearson Chi-Square	168	168.6590	1.0039
Log Likelihood	.	-93.7285	.

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-12.7151	2.7618	21.1965	0.0000
WIDTH		1	0.4680	0.1055	19.6573	0.0000
COLOR	1	1	1.3299	0.8525	2.4335	0.1188
COLOR	2	1	1.4023	0.5484	6.5380	0.0106
COLOR	3	1	1.1061	0.5921	3.4901	0.0617
COLOR	4	0	0.0000	0.0000	.	.

<u>Formulas</u>

$$-2(L_0 - L_1)$$

$$z = \hat{\beta}/\text{ASE}$$

$$\log(\mu) = \alpha + \beta x$$

$$\pi(x) = \alpha + \beta x \qquad \pi = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

$$\log(\tfrac{\pi}{1-\pi}) = \alpha + \beta x$$

$$\hat{\pi} = .5 \text{ at } x = -\hat{\alpha}/\hat{\beta}$$

$$\text{incremented rate of change} = \hat{\beta}\hat{\pi}(1-\hat{\pi})$$

$$\text{logit}(\pi) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \qquad \pi = \frac{\exp(\alpha+\beta_1 x_1+\cdots+\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\cdots+\beta_k x_k)}$$