UNIVERSITY OF TORONTO

Faculty of Arts and Science

APRIL 2010 EXAMINATIONS
STA 303 H1S / STA 1002 HS

Duration - 3 hours

Aids Allowed: Calculator

LAST NAME: SOLUTIONS **FIRST NAME:** ________________

STUDENT NUMBER: ________________________________

- There are 27 pages including this page.
- The last page is a table of formulae that may be useful. For all questions you can assume that the results on the formula page are known.
- A table of the chi-square distribution can be found on page 26.
- Total marks: 90

1abcd	1efg	1hi	2ab	2cde	2fghi

3a	3bcdef	4abcd	4efg	5

1. A study was carried out to investigate the effects of heredity and environment on intelligence. From adoption registers, researchers selected samples of adopted children whose biological parents and adoptive parents came from either the very highest or the very lowest socio-economic status (SES) categories. They attempted to obtain samples of size 10 from each combination (1. high adoptive SES and high biological SES, 2. high adoptive SES and low biological SES, 3. low adoptive SES and high biological SES, and 4. low SES for both parents). However, only 8 children belonged to combination 3. The 38 selected children were given intelligence quotient (IQ) tests. Some output from SAS for this analysis is given below and on the next 2 pages. The variables `adoptive` and `biologic` each take on the values `High` and `Low`, indicating the SES of the respective parents.

adoptive	biologic	IQ Mean	IQ Std	IQ N
High	High	119.60	12.25	10.00
	Low	103.60	12.71	10.00
Low	High	107.50	11.94	8.00
	Low	92.40	15.41	10.00

```
                     The GLM Procedure

                  Class Level Information
             Class          Levels    Values
             adoptive            2    High Low
             biologic            2    High Low

          Number of Observations Read          38
          Number of Observations Used          38

Dependent Variable: IQ

                                        Sum of
Source                       DF        Squares    Mean Square   F Value   Pr > F
Model                        (A)         (C)      1257.003509      7.19   0.0007
Error                        (B)    5941.200000       (D)
Corrected Total              37     9712.210526

              R-Square     Coeff Var      Root MSE       IQ Mean
              0.388275      12.50799      13.21897      105.6842
```

Output continues on the next page

(Question 1 continued)

Source	DF	Type I SS	Mean Square	F Value	Pr > F
adoptive	1	1477.632749	1477.632749	8.46	0.0064
biologic	1	2291.471895	2291.471895	13.11	0.0009
adoptive*biologic	1	1.905882	1.905882	0.01	0.9174

Source	DF	Type III SS	Mean Square	F Value	Pr > F
adoptive	1	1277.388235	1277.388235	7.31	0.0106
biologic	1	2275.788235	2275.788235	13.02	0.0010
adoptive*biologic	1	1.905882	1.905882	0.01	0.9174

The GLM Procedure

Class Level Information

Class	Levels	Values
adoptive	2	High Low
biologic	2	High Low

Number of Observations Read	38
Number of Observations Used	38

Dependent Variable: IQ

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3769.104644	1884.552322	11.10	0.0002
Error	35	5943.105882	169.803025		
Corrected Total	37	9712.210526			

R-Square	Coeff Var	Root MSE	IQ Mean
0.388079	12.32999	13.03085	105.6842

Source	DF	Type I SS	Mean Square	F Value	Pr > F
adoptive	1	1477.632749	1477.632749	8.70	0.0056
biologic	1	2291.471895	2291.471895	13.49	0.0008

Source	DF	Type III SS	Mean Square	F Value	Pr > F
adoptive	1	1276.005229	1276.005229	7.51	0.0096
biologic	1	2291.471895	2291.471895	13.49	0.0008

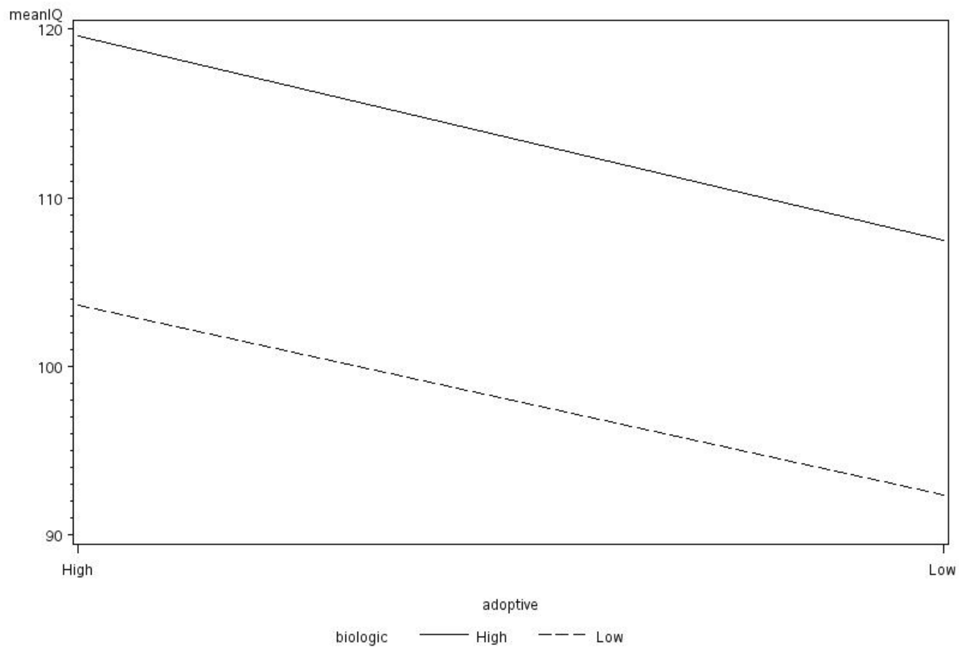
Level of adoptive	N	--------------IQ------------- Mean	Std Dev
High	20	111.600000	14.6625193
Low	18	99.111111	15.6238464

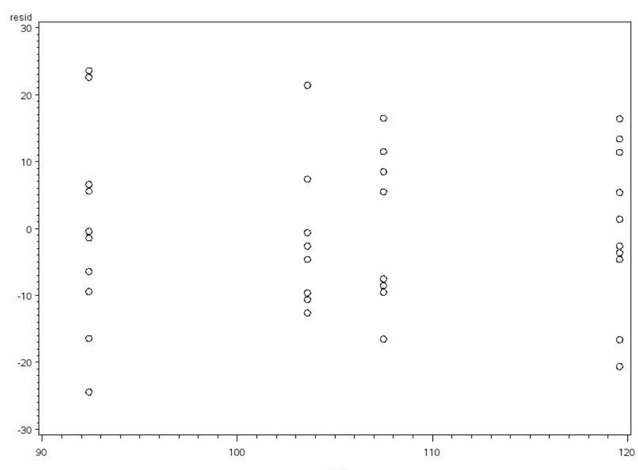
Least Squares Means

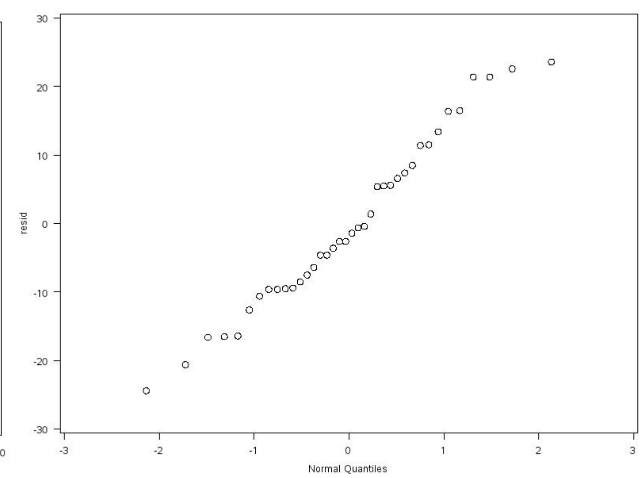
adoptive	IQ LSMEAN
High	111.600000
Low	99.976471

Output continues on the next page

(Question 1 continued)







Questions begin on the next page.

Continued

(a) (4 marks) Some numbers in the SAS output on page 2 have been replaced by letters. What are the missing values?

(A) = 3

(B) = 34

(C) = 3771.011

(D) = 174.7412

(b) (1 mark) Two linear models have been fit in the output above. In the first linear model, how many β's (coefficients of terms in the linear model) must be estimated?

4 (including β_0)

(c) (2 marks) Why can the first model be considered a saturated model? Explain why, in this case, it is possible to carry out inference.

The explanatory variables are categorical and using indicator variables in the model gives estimates of the response for each possible value of the explanatory variables, so the model can be considered saturated. It is possible to carry out inference because there are multiple observations for each combination of values of the explanatory variables.

(d) (2 marks) What is being tested by the test with p-value 0.9174? What do you conclude?

$H_0: \beta_3 = 0$ versus $H_a: \beta_3 \neq 0$, given the other variables are in the model, where β_3 is the coefficient of the adoptive-biological parent interaction term in the model.
Conclude that there is no evidence against H_0, so the way the adoptive parent's SES contributes to IQ doesn't differ with the biological parent's SES.

(Question 1 continued)

(e) (2 marks) For the second linear model, some "Least Squares Means" are given. Explain clearly how they are calculated.

For the high (in adoptive parent's SES) group, the least squares mean is $\hat{\beta}_0 + \hat{\beta}_1 + \frac{1}{2}\hat{\beta}_2$ *(the effect of biological parent SES is averaged out).*
For the low group, the least squares mean is $\hat{\beta}_0 + \frac{1}{2}\hat{\beta}_2$.

(f) (2 marks) Why does one of the "Least Squares Means" differ from the means given in the table above the least squares means?

There are unequal sample sizes. The LSmeans calculation weights each group the same, ignoring that one group has fewer observations.

(g) (3 marks) From the results of this study, what do you conclude about the relationship between parental socio-economic status and IQ? Quote relevant p-values to support your conclusions.

The effect of adoptive parent SES is the same regardless of biological parent SES and vice versa ($p = 0.9174$).
There is strong evidence that the mean IQ differs with biological parent SES ($p = 0.0010$ *or* 0.0008).
There is strong evidence that the mean IQ differs with adoptive parent SES ($p = 0.0106$ *or* 0.0096).

(h) (3 marks) The first graph on page 4 is a plot of the mean IQ of the children, classified by the socio-economic status of their adoptive and biological parents. Explain how it illustrates your conclusions from part (g).

Conclusion	*Graph*
There is no interaction.	*The lines are parallel.*
There is a biological parent effect.	*The lines differ vertically.*
There is an adoptive parent effect.	*The lines are not horizontal.*

(i) (4 marks) Do you trust your conclusions from part (g)? Why or why not?

For valid inferences we need:

- *Independent observations – Assume that there is no relationship among any of the children or parents.*
- *Same variance in all groups – The standard deviations of IQ are close to equal for each adoptive-biological parent combination.*
- *Normally distributed errors – There are no outliers. The normal quantile plot doesn't indicate any serious departures from normality.*

Since these conditions appear to be met, we can trust the inferences.
(Note that the p-values are either very large or very small, so even if they are only approximately correct the conclusions would not differ.)

2. Some of the debate about capital punishment in the U.S. has revolved around the rôle race plays in the decision to use it. The 674 subjects considered in this question were the defendants in murder cases in Florida between 1976 and 1987. SAS output for 4 models is given below and on the next 3 pages. The variables are:
`V` - the race of the victim (either black (`B`) or white (`W`))
`D` - the race of the defendant (either black (`B`) or white (`W`))
`C` - verdict for capital punishment (yes (`Y`) or no (`N`))

MODEL 1

```
                        The GENMOD Procedure

                          Model Information
                    Distribution           Poisson
                    Link Function              Log
                    Dependent Variable       count

               Number of Observations Read           8
               Number of Observations Used           8

                       Class Level Information
                     Class      Levels    Values
                     V               2    B W
                     D               2    B W
                     C               2    N Y

                  Criteria For Assessing Goodness Of Fit
      Criterion                     DF           Value        Value/DF
      Deviance                       4        402.8353        100.7088
      Scaled Deviance                4        402.8353        100.7088
      Pearson Chi-Square             4        419.5584        104.8896
      Scaled Pearson X2              4        419.5584        104.8896
      Log Likelihood                         2725.4956
      Full Log Likelihood                    -220.4376
      AIC (smaller is better)                 448.8752
      AICC (smaller is better)                462.2085
      BIC (smaller is better)                 449.1930

Algorithm converged.

                  Analysis Of Maximum Likelihood Parameter Estimates
                                 Standard   Wald 95% Confidence        Wald
Parameter       DF   Estimate       Error          Limits        Chi-Square   Pr > ChiSq
Intercept        1     3.6172      0.1255     3.3713      3.8632      830.72       <.0001
V         B      1    -1.1753      0.0907    -1.3531     -0.9974      167.81       <.0001
V         W      0     0.0000      0.0000     0.0000      0.0000        .            .
D         B      1    -0.9277      0.0855    -1.0953     -0.7602      117.81       <.0001
D         W      0     0.0000      0.0000     0.0000      0.0000        .            .
C         N      1     2.1874      0.1279     1.9367      2.4380      292.53       <.0001
C         Y      0     0.0000      0.0000     0.0000      0.0000        .            .
Scale            0     1.0000      0.0000     1.0000      1.0000
```


Continued

MODEL 2

The GENMOD Procedure

Model Information

Distribution	Poisson
Link Function	Log
Dependent Variable	count

Number of Observations Read	8
Number of Observations Used	8

Class Level Information

Class	Levels	Values
V	2	B W
D	2	B W
C	2	N Y

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	3	22.2659	7.4220
Scaled Deviance	3	22.2659	7.4220
Pearson Chi-Square	3	19.7018	6.5673
Scaled Pearson X2	3	19.7018	6.5673
Log Likelihood		2915.7803	
Full Log Likelihood		-30.1529	
AIC (smaller is better)		70.3058	
AICC (smaller is better)		100.3058	
BIC (smaller is better)		70.7030	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept			1	3.8526	0.1239	3.6097	4.0955	966.09	<.0001
V	B		1	-3.3737	0.2542	-3.8721	-2.8754	176.08	<.0001
V	W		0	0.0000	0.0000	0.0000	0.0000	.	.
D	B		1	-2.2751	0.1516	-2.5722	-1.9780	225.30	<.0001
D	W		0	0.0000	0.0000	0.0000	0.0000	.	.
C	N		1	2.1874	0.1279	1.9367	2.4380	292.53	<.0001
C	Y		0	0.0000	0.0000	0.0000	0.0000	.	.
V*D	B	B	1	4.4654	0.3041	3.8694	5.0614	215.64	<.0001
V*D	B	W	0	0.0000	0.0000	0.0000	0.0000	.	.
V*D	W	B	0	0.0000	0.0000	0.0000	0.0000	.	.
V*D	W	W	0	0.0000	0.0000	0.0000	0.0000	.	.

Continued

(Question 2 continued)

MODEL 3

The GENMOD Procedure

Model Information

Distribution	Poisson
Link Function	Log
Dependent Variable	count

Number of Observations Read	8
Number of Observations Used	8

Class Level Information

Class	Levels	Values
V	2	B W
D	2	B W
C	2	N Y

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	5.3940	2.6970
Scaled Deviance	2	5.3940	2.6970
Pearson Chi-Square	2	5.8109	2.9054
Scaled Pearson X2	2	5.8109	2.9054
Log Likelihood		2924.2162	
Full Log Likelihood		-21.7170	
AIC (smaller is better)		55.4339	
AICC (smaller is better)		139.4339	
BIC (smaller is better)		55.9106	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept			1	4.0610	0.1258	3.8145	4.3076	1042.18	<.0001
V	B		1	-4.9710	0.5675	-6.0833	-3.8588	76.74	<.0001
V	W		0	0.0000	0.0000	0.0000	0.0000	.	.
D	B		1	-2.2751	0.1516	-2.5722	-1.9780	225.30	<.0001
D	W		0	0.0000	0.0000	0.0000	0.0000	.	.
C	N		1	1.9526	0.1336	1.6908	2.2144	213.68	<.0001
C	Y		0	0.0000	0.0000	0.0000	0.0000	.	.
V*C	B	N	1	1.7045	0.5237	0.6780	2.7310	10.59	0.0011
V*C	B	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
V*C	W	N	0	0.0000	0.0000	0.0000	0.0000	.	.
V*C	W	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
V*D	B	B	1	4.4654	0.3041	3.8694	5.0614	215.64	<.0001
V*D	B	W	0	0.0000	0.0000	0.0000	0.0000	.	.
V*D	W	B	0	0.0000	0.0000	0.0000	0.0000	.	.
V*D	W	W	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale			0	1.0000	0.0000	1.0000	1.0000		

MODEL 4

The GENMOD Procedure

Model Information

Distribution	Poisson
Link Function	Log
Dependent Variable	count
Number of Observations Read	8
Number of Observations Used	8

Class Level Information

Class	Levels	Values
V	2	B W
D	2	B W
C	2	N Y

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	0.3798	0.3798
Scaled Deviance	1	0.3798	0.3798
Pearson Chi-Square	1	0.1978	0.1978
Scaled Pearson X2	1	0.1978	0.1978
Log Likelihood		2926.7234	
Full Log Likelihood		-19.2098	
AIC (smaller is better)		52.4197	
AICC (smaller is better)		.	
BIC (smaller is better)		52.9758	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept			1	3.9668	0.1374	3.6976	4.2361	833.78	<.0001
V	B		1	-5.6696	0.6459	-6.9355	-4.4037	77.06	<.0001
V	W		0	0.0000	0.0000	0.0000	0.0000	.	.
D	B		1	-1.5525	0.3262	-2.1918	-0.9132	22.66	<.0001
D	W		0	0.0000	0.0000	0.0000	0.0000	.	.
C	N		1	2.0595	0.1458	1.7736	2.3453	199.40	<.0001
C	Y		0	0.0000	0.0000	0.0000	0.0000	.	.
V*D	B	B	1	4.5950	0.3135	3.9805	5.2095	214.78	<.0001
V*D	B	W	0	0.0000	0.0000	0.0000	0.0000	.	.
V*D	W	B	0	0.0000	0.0000	0.0000	0.0000	.	.
V*D	W	W	0	0.0000	0.0000	0.0000	0.0000	.	.
D*C	B	N	1	-0.8678	0.3671	-1.5872	-0.1483	5.59	0.0181
D*C	B	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
D*C	W	N	0	0.0000	0.0000	0.0000	0.0000	.	.
D*C	W	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
V*C	B	N	1	2.4044	0.6006	1.2273	3.5816	16.03	<.0001
V*C	B	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
V*C	W	N	0	0.0000	0.0000	0.0000	0.0000	.	.
V*C	W	Y	0	0.0000	0.0000	0.0000	0.0000	.	.

Continued

(Question 2 continued)

(a) (4 marks) For each of the 4 models for which output is given, give a practical interpretation of the relationships among the variables (assuming that the model is appropriate).

MODEL 1: `V`, `D`, `C` *are independent*
MODEL 2: `V`, `D` *are not independent; they are jointly independent of* `C`
MODEL 3: `C`, `D` *are conditionally independent, conditional on* `V`
MODEL 4: the effect of each variable depends on the value of each other variable, but each of these interactions is the same for the values of the third variables

(b) (4 marks) Show how the value for the "Full Log Likelihood" is calculated for model 1. Give your answer in terms of the observed counts y_{ijk}.

Likelihood:

$$\prod_{i,j,k} \frac{e^{-\mu_{ijk}} \mu_{ijk}^{y_{ijk}}}{y_{ijk}!}$$

Log-likelihood:

$$-\sum_i \sum_j \sum_k \mu_{ijk} + \sum_i \sum_j \sum_k y_{ijk} \log(\mu_{ijk}) - \sum_i \sum_j \sum_k \log(y_{ijk}!)$$

To get the value in the table, plug

$$\hat{\mu}_{ijk} = 674 \frac{y_{i\cdot\cdot}}{674} \frac{y_{\cdot j \cdot}}{674} \frac{y_{\cdot\cdot k}}{674}$$

into the above for μ_{ijk}*, where* $y_{i\cdot\cdot} = \sum_j \sum_k y_{ijk}$*, etc.*

(Question 2 continued)

(c) (1 mark) For model 1, explain why the degrees of freedom for the "Criteria For Assessing Goodness Of Fit" is 4.

The number of observed counts is 8. There are 4 β's in the model (β_0 plus the coefficient of one dummy variable for each classification). So the required degrees of freedom is $8 - 4$.

(d) (5 marks) Use a likelihood ratio test to compare the fits of models 1 and 3. State the null and alternative hypotheses, the test statistic, the distribution of the test statistic under the null hypothesis, the p-value, and your conclusion.

H_0*: the coefficients of the* `V-C` *and* `V-D` *interaction terms are 0*
H_a*: at least one of these coefficients is not 0*
Test statistic: $402.8353 - 5.3940 = 397.4413$
Under H_0*, this is an observation from a chi-square distribution with 2 degrees of freedom*
From tables, $p < 0.005$
There is strong evidence that at least one of these coefficients is not 0 so the smaller model is not appropriate.

(e) (4 marks) Carry out the Deviance Goodness-of-Fit test for model 3. State the null and alternative hypotheses, the test statistic, the distribution of the test statistic under the null hypothesis, the p-value, and your conclusion.

H_0*: the coefficents of the* `C-D` *and* `V-C-D` *interaction terms are 0 (since these terms would be in the saturated model)*
H_a*: at least one of these coefficients is not 0*
Test statistic: 5.3940
Under H_0*, this is an observation from a chi-square distribution with 2 degrees of freedom*
From tables, $0.05 < p < 0.1$
There is weak evidence that at least one of the interaction terms has a non-zero coefficient, so the saturated model fits the data better.

(Question 2 continued)

(f) (2 marks) Using model 4, what is the estimated count of the number of cases with a verdict of capital punishment for which the defendant and victim were both white?

$\exp\{3.9668\} = 52.8$

(g) (3 marks) Using model 4, estimate the odds of receiving a verdict in favour of capital punishment if the defendant was black.

If `V` *is* `W`*, the odds are* $1/\exp(2.0595 - 0.8676) = 0.3037$
If `V` *is* `B`*, the odds are* $1/\exp(2.0595 - 0.8676 + 2.4044) = 0.0274$

(h) (4 marks) For model 4, what evidence is available from the SAS output that the model is adequate? What else would you like to see to ensure that the Wald tests are appropriate?

The deviance is small so the model fits the data as well as the saturated model with no extra-Poisson variation.
Would like to see the residuals to check for outliers and the expected counts in each cell to ensure that they are large enough for the large-sample tests to be (at least approximately) correct.

(i) (2 marks) Which of the 4 models would you choose for these data? Why?

MODEL 4
The deviance is smallest and the saturated model does not fit significantly better. Moreover, in part (e) it was shown that MODEL 3 doesn't fit as well as the saturated model.

3. Below is some additional output analysing the data from question 2. `nCapital` is the number of cases for which the verdict was for capital punishment.

MODEL A

```
                    The LOGISTIC Procedure
                     Model Information
         Response Variable (Events)     nCapital
         Response Variable (Trials)     m
         Model                          binary logit
         Optimization Technique         Fisher's scoring

             Number of Observations Read           4
             Number of Observations Used           4
             Sum of Frequencies Read             674
             Sum of Frequencies Used             674

                      Response Profile
                Ordered     Binary           Total
                  Value     Outcome      Frequency
                      1     Event               68
                      2     Nonevent           606

                  Class Level Information
                  Class     Value     Design Variables
                  V         B         1
                            W         0
                  D         B         1
                            W         0

                  Model Convergence Status
        Quasi-complete separation of data points detected.
WARNING: The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown
         are based on the last maximum likelihood iteration. Validity of the model fit is
         questionable.

                    Model Fit Statistics
                                           Intercept
                             Intercept           and
               Criterion          Only    Covariates
               AIC             442.843       426.577
               SC              447.356       444.630
               -2 Log L        440.843       418.577

             Testing Global Null Hypothesis: BETA=0
     Test                 Chi-Square       DF     Pr > ChiSq
     Likelihood Ratio        22.2659        3         <.0001
     Score                   19.7018        3         0.0002
     Wald                    14.6545        3         0.0021
```

Output for MODEL A continues on the next page.

Output for MODEL A continued

```
                    Type 3 Analysis of Effects
                                  Wald
            Effect      DF    Chi-Square    Pr > ChiSq
            V            1        0.0032        0.9547
            D            1        5.0991        0.0239
            V*D          1        0.0020        0.9640

            Analysis of Maximum Likelihood Estimates
                                   Standard          Wald
Parameter         DF    Estimate      Error    Chi-Square    Pr > ChiSq
Intercept          1     -2.0556     0.1459      198.5297        <.0001
V         B        1    -11.3015      198.8        0.0032        0.9547
D         B        1      0.8426     0.3731        5.0991        0.0239
V*D       B B      1      8.9663      198.8        0.0020        0.9640

     Association of Predicted Probabilities and Observed Responses
           Percent Concordant     35.3    Somers' D    0.261
           Percent Discordant      9.1    Gamma        0.589
           Percent Tied           55.6    Tau-a        0.047
           Pairs                 41208    c            0.631
```

MODEL B

```
                        The LOGISTIC Procedure

                          Model Information

        (SOME OUTPUT OMITTED HERE THAT IS THE SAME AS FOR MODEL A)

                      Model Convergence Status
           Convergence criterion (GCONV=1E-8) satisfied.

                        Model Fit Statistics
                          Intercept   Intercept and
             Criterion         Only      Covariates
             AIC            442.843         424.957
             SC             447.356         438.496
             -2 Log L       440.843         418.957

              Testing Global Null Hypothesis: BETA=0
        Test                 Chi-Square       DF     Pr > ChiSq
        Likelihood Ratio        21.8861        2         <.0001
        Score                   18.7847        2         <.0001
        Wald                    16.2460        2         0.0003
```

Output for MODEL B continues on the next page.

Output for MODEL B continued

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
V	1	16.0262	<.0001
D	1	5.5889	0.0181

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.0595	0.1458	199.3973	<.0001
V	B	1	-2.4044	0.6006	16.0262	<.0001
D	B	1	0.8678	0.3671	5.5889	0.0181

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
V B vs W	0.090	0.028	0.293
D B vs W	2.382	1.160	4.890

Association of Predicted Probabilities and Observed Responses

Percent Concordant	35.3	Somers' D	0.261
Percent Discordant	9.1	Gamma	0.589
Percent Tied	55.6	Tau-a	0.047
Pairs	41208	c	0.631

(a) (4 marks) Give test statistics and p-values for **two** tests comparing models A and B. What do you conclude? (As part of your conclusion, you should be choosing one of model A or B.)

i. *Wald test that coefficient of* `V-D` *interaction term is 0 has test statistic 0.0020 and p-value 0.9640*

ii. *Likelihood Ratio Test has test statistic 418.957-418.577=0.38 and p-value from chi-square table with 1 degree of freedom* $0.1 < p < 0.9$

Conclusion: There is no evidence against the hypothesis that the coefficient of the `V-D` *interaction term is 0. So choose MODEL B.*

(Question 3 continued)

(b) (2 marks) For the model you chose in part (a), describe the relationship among the 3 variables.

`V` *affects the probability of receiving capital punishment* ($p < 0.0001$)
`D` *affects the probability of receiving capital punishment* ($p < 0.0181$)
How each of `V` *and* `D` *affects the probability of receiving capital punishment doesn't vary with the value of the other.*

(c) (2 marks) Using model B, estimate the odds of receiving a verdict in favour of capital punishment if the defendant and victim were both black.

$\exp(-2.0595 - 2.4044 + 0.8678) = 0.0274$

(d) (2 marks) The SAS output for model A includes the message below. Explain what the message indicates.

```
                    Quasi-complete separation of data points detected.
WARNING: The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown
         are based on the last maximum likelihood iteration. Validity of the model fit is
         questionable.
```

Values of the explanatory variables `V` *and* `D` *almost perfectly divide the observations into capital punishment / not capital punishment groups.*

(e) (2 marks) For model A, what are the hypotheses for the likelihood ratio test under the heading "Testing Global Null Hypothesis: BETA=0" in the SAS output? What do you conclude?

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ *where* β_1, β_2, β_3 *are the coefficients of the indicator variables for* `V`, `D` *and their interaction*
H_a: *at least one of these coefficients is not zero*
$p < 0.0001$ *So there is strong evidence that at least one of the coefficients is not 0.*

(f) (2 marks) Do you prefer the analysis carried out on these data in question 2 or question 3? Why?

I prefer the logistic regression analysis of question 3. The interpretation is simpler. And for these data, of the 3 variables there is one clear reponse variable.

4. A study followed the orthodontic growth of 27 children (16 males and 11 females). At ages 8, 10, 12, and 14, the distance (in millimeters) from the center of the pituitary to pterygomaxillary fissure was measured. The investigators were interested in how the growth of this distance varied as the boys and girls grew. In the analysis below, age was treated as a categorical variable.

   Some SAS output is given below for 3 models that were fit to the resulting data.

MODEL I

```
                   The Mixed Procedure

                   Model Information
Dependent Variable            distance
Covariance Structure          Compound Symmetry
Subject Effect                subject(sex)
Estimation Method             REML
Residual Variance Method      Profile
Fixed Effects SE Method       Model-Based
Degrees of Freedom Method     Between-Within

                 Class Level Information
  Class      Levels    Values
  sex             2    Female Male
  subject        27    F01 F02 F03 F04 F05 F06 F07
                       F08 F09 F10 F11 M01 M02 M03
                       M04 M05 M06 M07 M08 M09 M10
                       M11 M12 M13 M14 M15 M16
  age             4    8 10 12 14


                       Dimensions
          Covariance Parameters             2
          Columns in X                     15
          Columns in Z                      0
          Subjects                         27
          Max Obs Per Subject               4

                 Number of Observations
      Number of Observations Read            108
      Number of Observations Used            108
      Number of Observations Not Used          0

                    Iteration History
Iteration    Evaluations    -2 Res Log Like       Criterion
        0              1       470.49084642
        1              1       423.40853283      0.00000000
                  Convergence criteria met.
```

Output for MODEL I continues on the next page.

Output for MODEL I continued

Estimated R Correlation Matrix for subject(sex) F01 Female

Row	Col1	Col2	Col3	Col4
1	1.0000	0.6245	0.6245	0.6245
2	0.6245	1.0000	0.6245	0.6245
3	0.6245	0.6245	1.0000	0.6245
4	0.6245	0.6245	0.6245	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	subject(sex)	3.2854
Residual		1.9750

Fit Statistics

-2 Res Log Likelihood	423.4
AIC (smaller is better)	427.4
AICC (smaller is better)	427.5
BIC (smaller is better)	430.0

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	47.08	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
age	3	75	35.35	<.0001
sex	1	25	9.29	0.0054
age*sex	3	75	2.36	0.0781

Continued

MODEL II

(The output was edited to remove Class Level Information and Number of Observations (both same as model I) and Iteration History (convergence criteria were met).)

```
                 The Mixed Procedure

                   Model Information
Dependent Variable            distance
Covariance Structures         Variance Components,
                              Autoregressive
Subject Effect                subject(sex)
Estimation Method             REML
Residual Variance Method      Profile
Fixed Effects SE Method       Model-Based
Degrees of Freedom Method     Containment

                            Dimensions
          Covariance Parameters            3
          Columns in X                    15
          Columns in Z                    27
          Subjects                         1
          Max Obs Per Subject            108

                 Estimated R Correlation Matrix
               for subject(sex) F01 Female
  Row        Col1         Col2         Col3         Col4
    1      1.0000     -0.05822     0.003390     -0.00020
    2    -0.05822       1.0000     -0.05822     0.003390
    3    0.003390     -0.05822       1.0000     -0.05822
    4    -0.00020     0.003390     -0.05822       1.0000

             Covariance Parameter Estimates
       Cov Parm         Subject         Estimate
       subject(sex)                       3.3423
       AR(1)            subject(sex)    -0.05822
       Residual                           1.9206

                     Fit Statistics
         -2 Res Log Likelihood          423.3
         AIC (smaller is better)        429.3
         AICC (smaller is better)       429.5
         BIC (smaller is better)        433.2

             Type 3 Tests of Fixed Effects
                   Num      Den
    Effect          DF       DF     F Value    Pr > F
    age              3       75       37.60    <.0001
    sex              1       25        9.28    0.0054
    age*sex          3       75        2.51    0.0649
```

(Question 4 continued)

MODEL III

(The output was edited to remove Class Level Information and Number of Observations (both same as models I and II) and Iteration History (convergence criteria were met).)

```
                     The Mixed Procedure

                  Model Information
Dependent Variable          distance
Covariance Structure        Unstructured
Subject Effect              subject(sex)
Estimation Method           REML
Residual Variance Method    None
Fixed Effects SE Method     Model-Based
Degrees of Freedom Method   Between-Within

                          Dimensions
          Covariance Parameters            10
          Columns in X                     15
          Columns in Z                      0
          Subjects                         27
          Max Obs Per Subject               4

                Estimated R Correlation Matrix
              for subject(sex) F01 Female
  Row        Col1        Col2        Col3        Col4
    1      1.0000      0.5707      0.6613      0.5216
    2      0.5707      1.0000      0.5632      0.7262
    3      0.6613      0.5632      1.0000      0.7281
    4      0.5216      0.7262      0.7281      1.0000

                    Fit Statistics
         -2 Res Log Likelihood          414.0
         AIC (smaller is better)        xxxxx
         AICC (smaller is better)       436.5
         BIC (smaller is better)        447.0

           Null Model Likelihood Ratio Test
             DF    Chi-Square      Pr > ChiSq
              9         56.46          <.0001

            Type 3 Tests of Fixed Effects
                  Num     Den
    Effect         DF      DF    F Value    Pr > F
    age             3      25      34.45    <.0001
    sex             1      25       9.29    0.0054
    age*sex         3      25       2.93    0.0532
```

(Question 4 continued)

(a) (1 mark) The models include the interaction of sex and age. Explain in practical terms why this was included in the models.

We are interested in how the change in distance with age differs between sexes.

(b) (2 marks) The model was fit using the mixed models procedure in SAS. Explain why the model is "mixed".

The mixed models procedure can be used to fit models with fixed effects (age, sex) and random effects (subject).

(c) (4 marks) Write the model that was fit in model I, carefully defining all terms. (Do not write the fitted equation; write the model in terms of its parameters.)

$$\begin{aligned} Y_{ijk} &= \beta_0 + \beta_1 I_{[sex=F],ijk} + \beta_2 I_{[age=8],ijk} + \beta_3 I_{[age=10],ijk} + \beta_4 I_{[age=12],ijk} \\ &\quad + \beta_5 I_{[sex=F],ijk} * I_{[age=8],ijk} + \beta_6 I_{[sex=F],ijk} * I_{[age=10],ijk} \\ &\quad + \beta_7 I_{[sex=F],ijk} * I_{[age=12],ijk} + \epsilon_{ijk} \end{aligned}$$

for the ith subject in the jth sex ($j = 1, 2$) *at age* k ($k = 1, 2, 3, 4$)
where $I_{[condition],ijk}$ *is 1 if the condition is met for observation ijk and is 0 otherwise*
Y_{ijk} *is the distance*
ϵ_{ijk} *are random errors with* $Var(\epsilon_{ijk}) = \sigma^2_\epsilon$ *and* $Cov(\epsilon_{ijk}, \epsilon_{ijn}) = \sigma^2_{subj}$ ($k \neq n$)

(The solution for this question could also be given in matrix terms. Also, the model fit in SAS did not include a random effect for subject so this is not included in the model here and the covariance parameter for the random effect is not mentioned in part (d).)

(d) (2 marks) For model I, why is the number of covariance parameters equal to 2?

The 2 covariance parameters are the variance of the error term and the covariance between observations on the same subject.

(Question 4 continued)

(e) (1 mark) What is the value of AIC for model III?

$$414.0 + 2(10) = 434.4$$

(f) (2 marks) AR(1) is a commonly used covariance structure in situations such as this, where observations are taken over time. Explain why it is not an appropriate covariance structure for these data by comparing at least 2 different kinds of information given in the SAS output.

Possible answers:

- *the AIC for the AR(1) model is greater than the AIC for the CS model*
- *ρ is estimated as negative for the AR(1) model which does not seem appropriate in practical terms*
- *the estimated correlation matrix for the UN model does not show correlation decreasing as time between observations increases which you would expect in an AR(1) model*
- *it is also possible to carry out likelihood ratio tests comparing models (which should be done if this is to count as one of the kinds of information*

(g) (2 marks) How do the conditions for valid inference for this model differ from the conditions needed for a multiple linear regression model?

Not all observations are independent; observations on the same subject are modeled as correlated.
Don't need constant variance; it can be modeled to differ among ages or between genders.

5. (a) (6 marks) In order for inferences to be valid, conditions must be met. Assume standard analyses that were taught in this course are being carried out.

   i. Give two examples of conditions that must be met for both analysis of variance and binomial logistic regression models in order for the inferences to be valid.

   - independent observations
   - correct form of model
   (- no outliers)

   ii. Give two examples of conditions that must be met for the inferences to be valid for an analysis of variance model but which are not necessary for a binomial logistic regression model.

   - normally distributed error terms
   - constant variance (equal for all groups)

   iii. Give two examples of conditions that must be met for the inferences to be valid for a binomial logistic regression model but which are not necessary for an analysis of variance model.

   - Binomial distribution appropriate (no extra-Binomial variation)
   - large sample size for Wald and likelihood ratio tests

(b) (4 marks) Here are two recent quotes from lecture.

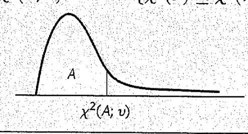
   "What does it mean if you make predictions from a fitted model that does not adequately describe the data?"

   "Only do inference on valid models."

   Imagine it is sometime in the future and you have been hired to do the statistical analysis on the data collected from a scientific study. How will the ideas behind these quotes affect the work you will do? And why is this important?

TABLE B.3 Percentiles of the χ^2 Distribution.

Entry is $\chi^2(A;\nu)$ where $P\{\chi^2(\nu) \leq \chi^2(A;\nu)\} = A$



	A									
ν	.005	.010	.025	.050	.100	.900	.950	.975	.990	.995
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

Continued

Some formulae:

Pooled t-test

$$t_{obs} = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Test for two proportions

$$z_{obs} = (\hat{\pi}_1 - \hat{\pi}_2) \Big/ \sqrt{\hat{\pi}_p(1-\hat{\pi}_p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Linear Regression

$$b_1 = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2} = \frac{\sum X_iY_i - n\overline{XY}}{\sum X_i^2 - n\overline{X}^2} \qquad b_0 = \overline{Y} - b_1\overline{X}$$

One-way analysis of variance

$$\text{SSTO} = \sum_{i=1}^{N}(Y_i-\overline{Y})^2 \qquad \text{SSE} = \sum_{g=1}^{G}\sum_{(g)}(Y_i-\overline{Y}_g)^2 \qquad \text{SSR} = \sum_{g=1}^{G} n_g(\overline{Y}_g-\overline{Y})^2$$

Bernoulli and Binomial distributions

If $Y \sim \text{Bernoulli}(\pi)$

$\text{E}(Y) = \pi$, $\text{Var}(Y) = \pi(1-\pi)$

If $Y \sim \text{Binomial}(m, \pi)$

$\text{E}(Y) = m\pi$, $\text{Var}(Y) = m\pi(1-\pi)$

Logistic Regression with Binomial Response formulae

$$\text{Deviance} = 2\sum_{i=1}^{n}\left\{y_i\log(y_i) + (m_i-y_i)\log(m_i-y_1) - y_i\log(\hat{y}_i) + (m_i-y_i)\log(m_i-\hat{y}_1)\right\}$$

$$D_{res,i} = \text{sign}(y_i - m_i\hat{\pi}_i)\sqrt{2\left\{y_i\log\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i-y_i)\log\left(\frac{m_i-y_i}{m_i-m_i\hat{\pi}_i}\right)\right\}}$$

$$P_{res,i} = \frac{y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

Multinomial distribution for 2×2 table

$$\Pr(\mathbf{Y}=\mathbf{y}) = \frac{n!}{y_{11}!y_{12}!y_{21}!y_{22}!}\pi_{11}^{y_{11}}\pi_{12}^{y_{12}}\pi_{21}^{y_{21}}\pi_{22}^{y_{22}}$$

Poisson distribution

$$\Pr(Y=y) = \frac{\mu^y e^{-\mu}}{y!},\ y = 0, 1, 2, \ldots$$

$$\text{E}(Y) = \mu,\ \text{Var}(Y) = \mu$$

Two-way contingency tables (easily generalizable to three-way tables)

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(y_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \qquad G^2 = 2\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ij}\log\left(\frac{y_{ij}}{\hat{\mu}_{ij}}\right)$$

$$D_{res,ij} = \text{sign}(y_{ij}-\hat{\mu}_{ij})\sqrt{2\left\{y_{ij}\log\left(\frac{y_{ij}}{\hat{\mu}_{ij}}\right) - y_{ij} + \hat{\mu}_{ij}\right\}}$$

$$P_{res,ij} = \frac{y_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

Model Fitting Criteria

$$\text{AIC} = -2\log(L) + 2p \qquad \text{SC} = -2\log(L) + p\log(N)$$