Applied Statistics

MS Comprehensive Examination

April 22, 2006

Instructions:

Please answer all six questions.

Record your answers in your blue books.

Show all of your computations.
Prove all of your assertions or quote the appropriate theorems.
Books, notes, and calculators *may be used.*

You have three hours.

Comprehensive Examination April 22, 2006

1. (15 points) A factorial experiment is conducted to study the influence of two operating temperatures and two types of face-plate glass in the light output of an oscilloscope tube. Two replicates are obtained at each treatment. Refer to the SAS output in **Appendix One** for the questions.

(a). Write the model and indicate the meaning of model parameters.

(b). Use $\alpha = 0.05$ in the analysis. Is there a significant interaction effect? Briefly explain whether there is agreement between the profile plot and your conclusion.

(c). Write the contrast for testing the main effect of the fact-plate glass.

2. (35 points) The research goal is to fit a regression model for the content genre diversity (y). The independent variables include the year (x_1), the audience diversity (x_2), and the number of cable networks (x_3). Refer to Appendix Two and Three. You may need to do some calculation to answer all the questions.

The SAS output for the model that includes x_1 and x_2 is in **Appendix Two**. Refer to Appendix Two for the questions (a)-(f).

(a). If the audience diversity remains constant, what is the annual change in the content genre diversity?

(b). Write the numerical equation for prediction.

(c). Predict the content genre diversity $\hat{y}_0$ at $x_1 = 2004$ and $x_2 = 0.79$ by the above equation and construct a 95% prediction interval. Is this interval wider than a 95% interval for E(y_0)? Briefly explain why.

(d). Calculate the residual, the studentized residual, and the studenttized deleted residual for year 2004.

(e). What are R^2 and adjusted R_a^2? Which one is a better measure for model selection? Briefly explain why.

(f). Find the hat matrix H.

The SAS output for the backward model selection is in **Appendix Three**.

(g). According to Appendix Three, what is the most appropriate model? Briefly explain the backward procedure.

Here is a dataset, with statistics, that we will work with in Problem 3. These are rates of healing, as measured by the number of new cells formed at the site of a cut, in one hour.

```
11    12    14    18    22    22    23    23    26    27    28    29    30    33    34    35    35    40
Variable    N   Mean    StDev
Rate       18  25.67   8.32
```

3. (25 points) Give details in every part below.

a. Find the 5-number summary (median, 1st & 3rd quartiles, and minimum and maximum) and display a box plot of this data. From this information, does the normal distribution seem to be a plausible model for this data? Explain.

b. Give a 90% confidence interval for the mean of this distribution.

c. Give a 90% confidence interval for the standard deviation of this distribution.

d. Use parametric methods to test whether the mean is less than or equal to 22 versus the alternative that it is greater. Use level of significance $\alpha = 0.05$.

e. Use the sign test to test whether the median is less than or equal to 22 versus the alternative that it is greater. Use level of significance $\alpha = 0.05$.

f. Use the signed rank test to test whether the median is less than or equal to 22 versus the alternative that it is greater. Use level of significance $\alpha = 0.05$.

g. Comment on the agreement between the results in d,e, and f. Do they agree? Why or why not? Which result would you use (and why)?

4. (5 points) If I'm going to do a simulation study on the power of a test, and I want to guarantee that, with 90% certainty, no matter what the true value of the power, my simulation-based power estimates will be within .01 of the actual value, what number of simulation repetitions should I use? Give your reasoning.

5. (10 points) The following data comes from a gamma distribution with shape parameter $\alpha = 1.5$. If α were 1, this would be exponentially distributed data. The task here is to perform a chi-square goodness of fit test (use level of significance = .05) to see whether or not we can reject the null hypothesis that this data does come from an exponential distribution. Statistics are provided that might help. The specifics of the test are up to you, but you need to justify the choices you make.

```
0.3    0.7    1.0    1.0    1.1    1.5    1.7    2.4    2.6    2.8    2.9    3.0    3.8    3.8    4.1
4.1    4.2    4.3    5.0    5.0    5.2    5.4    5.9    6.3    6.4    6.5    6.6    7.2    7.7    9.2
Variable    N    Mean   StDev   Minimum       Q1   Median       Q3   Maximum
x          30   4.052   2.309     0.330   2.208    4.074    6.029     9.195
```

6. (10 points) In this problem we explore whether or not pre-school (school for 3 & 4 year olds, before kindergarten) helps poor children later in life. Two groups of children were followed from early childhood until they were adults. One group of 62 attended pre-school; out of these, 38 used welfare as an adult. The other (control) group of 61 did not attend pre-school; out of these, 49 used welfare as an adult.

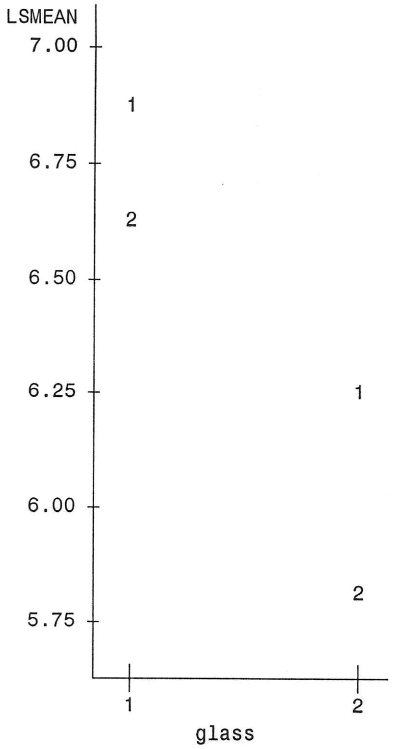
a. Find a 92% confidence interval for the difference between the two proportions who used welfare as an adult. Comment on the applicability of the method you choose to use.

b. Test the null hypothesis that the probability of requiring welfare as an adult is the same for the two groups versus the alternative that the probability is greater for the control group. Use level of significance $\alpha = .05$. Comment on any assumptions required for the validity of this test.

c. Use a 2x2 table approach to do a chi-square test of homogeneity between these two populations. Use level of significance $\alpha = .05$. Comment on any assumptions required for the validity of this test. Compare and contrast this test and the one in part b. What is the relationship between the two statistics?

Appendix One

```
******************************************************************************
                               The SAS System                                1

                             The GLM Procedure

                          Class Level Information

                       Class         Levels    Values

                       temp               2    1 2

                       glass              2    1 2

                  Number of Observations Read           8
                  Number of Observations Used           8
******************************************************************************
                               The SAS System                                2

                             The GLM Procedure
Dependent Variable: rep

                                          Sum of
 Source                      DF          Squares     Mean Square    F Value    Pr > F

 Model                        3       1.17625000      0.39208333       0.11    0.9503

 Error                        4      14.36430000      3.59107500

 Corrected Total              7      15.54055000

             R-Square     Coeff Var      Root MSE      rep Mean

             0.075689      29.66753      1.895013      6.387500

 Source                      DF      Type III SS     Mean Square    F Value    Pr > F

 temp                         1       0.24500000      0.24500000       0.07    0.8068
 glass                        1       0.91125000      0.91125000       0.25    0.6409
 temp*glass                   1       0.02000000      0.02000000       0.01    0.9441
******************************************************************************
                               The SAS System                                3

                             The GLM Procedure
                            Least Squares Means

                       temp    glass       rep LSMEAN

                       1       1           6.85000000
                       1       2           6.27500000
                       2       1           6.60000000
                       2       2           5.82500000
     Plot of LSMEAN*glass=temp.
```



Appendix Two

```
data one;
  input x1 x2 y x3 @@;
  cards;
    1984     0.61     0.84      32
    1985     0.61     0.85      39
    1986     0.55     0.87      44
    1987      0.6     0.87      47
    1988      0.6     0.87      49
    1989     0.62     0.87      52
    1990     0.63     0.87      56
    1991     0.62     0.87      65
    1992     0.63     0.87      70
    1993     0.67     0.88      76
    1994     0.72     0.86     110
    1995     0.74     0.87     123
    1996     0.74     0.88     151
    1997     0.74     0.89     161
    1998     0.74     0.89     196
    1999     0.76     0.89     224
    2000     0.76     0.89     234
    2001     0.77     0.88     244
    2002     0.77     0.88     272
    2003     0.78     0.89     320
    2004     0.79     0.89     353
;
run;
proc reg;
```

```
model y=x1 x2/covb;
run;
proc reg;
model y=x1 x2 x3/selection=backward;
run;
```

```
***************************************************************************
                          Analysis of Variance

                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F

Model                     2         0.00275        0.00137      25.41    <.0001
Error                    18      0.00097400     0.00005411
Corrected Total          20         0.00372

          Root MSE                  0.00736     R-Square     0.7384
          Dependent Mean            0.87476     Adj R-Sq     0.7094
          Coeff Var                 0.84092

                           Parameter Estimates

                           Parameter       Standard
   Variable     DF          Estimate          Error    t Value    Pr > |t|

   Intercept     1          -6.19932        1.63610      -3.79      0.0013
   x1            1           0.00360     0.00084282       4.27      0.0005
   x2            1          -0.15423        0.06822      -2.26      0.0364

                         Covariance of Estimates

     Variable           Intercept                 x1                  x2

     Intercept       2.6768119831       -0.001378883        0.1056303555
     x1              -0.001378883       7.1035084E-7         -0.00005458
     x2              0.1056303555        -0.00005458         0.004654122
```

Appendix Three

```
***************************************************************************

                      Backward Elimination: Step 0

        All Variables Entered: R-Square = 0.7544 and C(p) = 4.0000

                          Analysis of Variance

                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F

Model                     3         0.00281     0.00093636      17.40    <.0001
Error                    17      0.00091472     0.00005381
Corrected Total          20         0.00372
```

```
                 Parameter     Standard
    Variable      Estimate        Error   Type II SS  F Value  Pr > F

    Intercept     -7.99285      2.36244   0.00061591    11.45  0.0035
    x1             0.00450      0.00120   0.00075416    14.02  0.0016
    x2            -0.15554      0.06804   0.00028117     5.23  0.0354
    x3         -0.00005717   0.00005446   0.00005929     1.10  0.3085

                 Bounds on condition number: 20.719, 126.26
-----------------------------------------------------------------------------
                       Backward Elimination: Step 1
          Variable x3 Removed: R-Square = 0.7384 and C(p) = 3.1019


                          Analysis of Variance

                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F

Model                     2         0.00275        0.00137      25.41    <.0001
Error                    18      0.00097400     0.00005411
Corrected Total          20         0.00372


                 Parameter     Standard
    Variable      Estimate        Error   Type II SS  F Value  Pr > F

    Intercept     -6.19932      1.63610   0.00077689    14.36  0.0013
    x1             0.00360   0.00084282   0.00098773    18.25  0.0005
    x2            -0.15423      0.06822   0.00027656     5.11  0.0364

                 Bounds on condition number: 10.108, 40.433
-----------------------------------------------------------------------------
    All variables left in the model are significant at the 0.1000 level.
*****************************************************************************
```

```
                           The REG Procedure
                             Model: MODEL1
                         Dependent Variable: y

                     Summary of Backward Election

        Variable   Number  Partial    Model
  Step  Removed    Vars In R-Square  R-Square     C(p)   F Value  Pr > F

     1  x3               2   0.0159    0.7384   3.1019      1.10  0.3085
```