Department of Mathematics and Statistics
The University of Toledo

Master's Comprehensive Examination
Applied Statistics

April 19, 2014

Instructions:

Do all four problems;

Show all of your computations;

Prove all of your assertions or quote appropriate theorems;

This is three-hour open book examination.

1. (25 points) This problem relates to paired data $(X_1, Y_1), \ldots\ldots, (X_n, Y_n)$, with distributions $F(x)$ and $G(y)$ for X and Y. Denote $Z_i = Y_i - X_i$. Answer the following questions:

a. (5 points) Give the model assumptions required to be able to perform the Wilcoxon signed rank test for equality of medians for these distributions.

b. (5 points) If $n = 4$ and given that there are only two distinct values for $|Z|$s with two ties each, find the conditional distribution of T^+, the signed rank test statistic, under the assumptions given in part a and under $H_0 : \theta_X = \theta_Y$,. Use this distribution to find $E(T^+)$ and $var(T^+)$ in this case.

c. (5 points) If the data is $(6, 5), (10, 13), (8, 7)$ and $(15, 18)$, can we reject H_0 in favor of $H_1 : \theta_X < \theta_Y$ at (nominal) level of significance $\alpha = 0.10$? Use the exact distribution from part b.

d. (5 points) Use the formulas given in the text to find $E_0(T^+)$ and $var_0(T^+)$ in this case. Do you get the same answers as you did at the end of part b? Hint: the formula for $var_0(T^+)$ is

$$var_0\left(T^+\right) = (24)^{-1}\left[n\,(n+1)\,(2n+1) - \frac{1}{2}\sum_{j=1}^{g} t_j\,(t_j - 1)\,(t_j + 1)\right]$$

e. (5 points) Using the normal approximation for T^+, find the approximate P-value for the test in part c. Do you obtain the same final result using the normal approximation?

2. (25 points) In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product, the following results were obtained from the experiment based on a completely randomized design. Based on the SAS program and output in the appendix, answer the following questions:

a. (2 points) Give the equation of the fitted regression line using both explanatory variables.

b. (3 points) Use the C_p criterion to select the best subset of variables for this problem. Summarize the results and explain your choice of the best model.

In the following problems, use the model which have both explanatory variables to predict the response brand liking.

c. (5 points) Obtain the studentized deleted residuals for observation 14. Use the Bonferroni outlier test procedure with $\alpha = 0.10$ to identify whether it is an outlying Y observation.

d. (5 points) Use the diagonal elements of the hat matrix to identify whether observation 14 is an outlying X observation. State the decision rule and conclusion.

e. (10 points) Obtain DFFITS, DFBETAS and Cook's distance values for observation 14 to assess its influence. What do you conclude? (Hint: $F_{3,13}(0.5) = 0.8316$ and $n = 16$ is considered as a small sample size)

3. (30 points) y is the type of salmon (1: Alaskan; 2: Canadian); x_1 is the diameters of rings of the first year fresh water growth (hundredths of an inch); x_2 is the diameters of rings of the first year marine water growth (hundredths of an inch), x_3 is the sex (1: female; 2: male).

The first interest is whether there is any association between the probability of Alaskan Salmon and x_1, x_2, x_3. Using the R output in the appendix, answer the following questions:

a. (5 points) Write down the full model.

b. (5 points) Based on the R output, write down the most appropriate model. Provide at least two reasons for your choice.

c. (5 points) Interpret the estimates of the coefficients in the appropriate model you chose above.

In the following, we only use $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. The second interest is to assign $\mathbf{x}_0 = \begin{pmatrix} 108 \\ 368 \end{pmatrix}$ to one of these two populations using the sample information in the appendix.

d. (5 points) Use Fisher's linear classification rule to classify $\mathbf{x}_0$.

e. (5 points) Use the logistic regression model you chose in (b) to classify $\mathbf{x}_0$. (Use $p = 0.5$ as the cutoff point)

f. (5 points) Suppose that $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is bivariate normal with the same covariance matricies for these two kinds of salmon. Which of the methods (e) and (f) do you prefer? Briefly explain why.

4. (20 points) Suppose that two random samples are selected from two populations $N_2(\mu_1, \Sigma)$ and $N_2(\mu_2, \Sigma)$. The sample sizes are n_1 and n_2, respectively.

a. (3 points) What is the distribution of $\bar{\mathbf{X}} = \begin{pmatrix} \bar{\mathbf{X}}_1 \\ \bar{\mathbf{X}}_2 \end{pmatrix}$? Explain clearly why and find the mean vector and covariance matrix. $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ are two sample means.

b. (3 points) What is the distribution of $\mathbf{a}'(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$? Explain clearly why and find the mean and the covariance matrix, where $\mathbf{a}$ is a constant vector.

c. (4 points) Is the following a Hotelling's T^2? Explain clearly why. What is the distribution under $H_0 : \mu_1 = \mu_2$?

$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{\text{pooled}} \right]^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$$

Use the information about these two random samples in the appendix and the quantiles of the relevant distributions

d. (5 points) Test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$ at the significance level $\alpha = 0.05$.

e. (5 points) Construct a 95% simultaneous confidence intervals and Bonferroni simultaneous confidence intervals for $\mu_{11} - \mu_{21}$ and $\mu_{12} - \mu_{22}$, and compare them.

Appendix for Problem 2

SAS Output

```
data Brand;
input liking moisture sweet;
cards;
   64.0    4.0    2.0
   73.0    4.0    4.0
   61.0    4.0    2.0
   76.0    4.0    4.0
   72.0    6.0    2.0
   80.0    6.0    4.0
   71.0    6.0    2.0
   83.0    6.0    4.0
   83.0    8.0    2.0
   89.0    8.0    4.0
   86.0    8.0    2.0
   93.0    8.0    4.0
   88.0   10.0    2.0
   95.0   10.0    4.0
   94.0   10.0    2.0
  100.0   10.0    4.0
;
run;
proc reg data=Brand;
model liking = moisture sweet/r ;
run;
proc reg data=Brand;
model liking = moisture sweet/selection = cp b;
run;
proc reg data=Brand;
model liking = moisture sweet / r vif influence tol;
plot r.*(moisture sweet);
run;
quit;
```

```
                                     The SAS System

                                   The REG Procedure
                                     Model: MODEL1
                               Dependent Variable: liking

                          Number of Observations Read          16
                          Number of Observations Used          16

                                  Analysis of Variance

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

      Model                     2    1872.70000      936.35000     129.08    <.0001
      Error                    13      94.30000        7.25385
      Corrected Total          15    1967.00000

                Root MSE              2.69330    R-Square     0.9521
                Dependent Mean       81.75000    Adj R-Sq     0.9447
                Coeff Var             3.29455

                                  Parameter Estimates

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

           Intercept     1       37.65000        2.99610      12.57      <.0001
           moisture      1        4.42500        0.30112      14.70      <.0001
           sweet         1        4.37500        0.67332       6.50      <.0001

                                     The SAS System

                                   The REG Procedure
                                     Model: MODEL1
                               Dependent Variable: liking

                                   Output Statistics

      Dependent Predicted    Std Error            Std Error  Student                    Cook's
 Obs   Variable     Value Mean Predict  Residual   Residual Residual    -2-1 0 1 2          D

   1    64.0000   64.1000       1.3126   -0.1000      2.352  -0.0425 |       |       |   0.000
   2    73.0000   72.8500       1.3126    0.1500      2.352   0.0638 |       |       |   0.000
   3    61.0000   64.1000       1.3126   -3.1000      2.352   -1.318 |     **|       |   0.180
   4    76.0000   72.8500       1.3126    3.1500      2.352    1.339 |       |**     |   0.186
   5    72.0000   72.9500       0.9987   -0.9500      2.501   -0.380 |       |       |   0.008
   6    80.0000   81.7000       0.9987   -1.7000      2.501   -0.680 |      *|       |   0.025
   7    71.0000   72.9500       0.9987   -1.9500      2.501   -0.780 |      *|       |   0.032
   8    83.0000   81.7000       0.9987    1.3000      2.501    0.520 |       |*      |   0.014
   9    83.0000   81.8000       0.9987    1.2000      2.501    0.480 |       |       |   0.012
  10    89.0000   90.5500       0.9987   -1.5500      2.501   -0.620 |      *|       |   0.020
  11    86.0000   81.8000       0.9987    4.2000      2.501    1.679 |       |***    |   0.150
  12    93.0000   90.5500       0.9987    2.4500      2.501    0.979 |       |*      |   0.051
  13    88.0000   90.6500       1.3126   -2.6500      2.352   -1.127 |     **|       |   0.132
  14    95.0000   99.4000       1.3126   -4.4000      2.352   -1.871 |    ***|       |   0.363
  15    94.0000   90.6500       1.3126    3.3500      2.352    1.424 |       |**     |   0.211
  16   100.0000   99.4000       1.3126    0.6000      2.352    0.255 |       |       |   0.007

                         Sum of Residuals                           0
                         Sum of Squared Residuals            94.30000
```

```
                   Predicted Residual SS (PRESS)      148.37802
                                  The SAS System      10:12 Thursday, April 17, 2014   3

                                The REG Procedure
                                  Model: MODEL1
                           Dependent Variable: liking

                               C(p) Selection Method

                      Number of Observations Read          16
                      Number of Observations Used          16


      Number in                            -----------Parameter Estimates-----------
        Model         C(p)    R-Square      Intercept    moisture          sweet

            2       3.0000      0.9521       37.65000     4.42500        4.37500
            1      43.2190      0.7964       50.77500     4.42500              .
            1     216.9475      0.1557       68.62500           .        4.37500
                                  The SAS System

                                The REG Procedure
                                  Model: MODEL1
                           Dependent Variable: liking

                      Number of Observations Read          16
                      Number of Observations Used          16

                               Analysis of Variance

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

      Model                       2   1872.70000      936.35000     129.08    <.0001
      Error                      13     94.30000        7.25385
      Corrected Total            15   1967.00000

                  Root MSE              2.69330    R-Square     0.9521
                  Dependent Mean       81.75000    Adj R-Sq     0.9447
                  Coeff Var             3.29455

                               Parameter Estimates

                  Parameter      Standard                                       Variance
  Variable   DF    Estimate         Error   t Value   Pr > |t|     Tolerance    Inflation

  Intercept   1    37.65000       2.99610     12.57     <.0001             .            0
  moisture    1     4.42500       0.30112     14.70     <.0001       1.00000      1.00000
  sweet       1     4.37500       0.67332      6.50     <.0001       1.00000      1.00000

                                  Model: MODEL1
                           Dependent Variable: liking
                                Output Statistics

      Dependent Predicted    Std Error            Std Error  Student                     Cook's
  Obs  Variable     Value Mean Predict  Residual   Residual Residual    -2-1 0 1 2           D

    1   64.0000   64.1000       1.3126   -0.1000      2.352  -0.0425 |      |      |   0.000
    2   73.0000   72.8500       1.3126    0.1500      2.352   0.0638 |      |      |   0.000
    3   61.0000   64.1000       1.3126   -3.1000      2.352   -1.318 |    **|      |   0.180
    4   76.0000   72.8500       1.3126    3.1500      2.352    1.339 |      |**    |   0.186
    5   72.0000   72.9500       0.9987   -0.9500      2.501   -0.380 |      |      |   0.008
    6   80.0000   81.7000       0.9987   -1.7000      2.501   -0.680 |     *|      |   0.025
    7   71.0000   72.9500       0.9987   -1.9500      2.501   -0.780 |     *|      |   0.032
    8   83.0000   81.7000       0.9987    1.3000      2.501    0.520 |      |*     |   0.014
    9   83.0000   81.8000       0.9987    1.2000      2.501    0.480 |      |      |   0.012
   10   89.0000   90.5500       0.9987   -1.5500      2.501   -0.620 |     *|      |   0.020
   11   86.0000   81.8000       0.9987    4.2000      2.501    1.679 |      |***   |   0.150
   12   93.0000   90.5500       0.9987    2.4500      2.501    0.979 |      |*     |   0.051
   13   88.0000   90.6500       1.3126   -2.6500      2.352   -1.127 |    **|      |   0.132
   14   95.0000   99.4000       1.3126   -4.4000      2.352   -1.871 |   ***|      |   0.363
   15   94.0000   90.6500       1.3126    3.3500      2.352    1.424 |      |**    |   0.211
   16  100.0000   99.4000       1.3126    0.6000      2.352    0.255 |      |      |   0.007
                                Output Statistics

                     Hat Diag       Cov                 -------------DFBETAS-------------
     Obs  RStudent          H     Ratio     DFFITS    Intercept    moisture         sweet

       1   -0.0409     0.2375    1.6667    -0.0228      -0.0216      0.0157        0.0117
       2    0.0613     0.2375    1.6659     0.0342       0.0087     -0.0235        0.0175
       3   -1.3606     0.2375    1.0842    -0.7593      -0.7178      0.5226        0.3895
       4    1.3860     0.2375    1.0680     0.7735       0.1962     -0.5324        0.3968
       5   -0.3669     0.1375    1.4256    -0.1465      -0.1199      0.0442        0.0988
       6   -0.6649     0.1375    1.3225    -0.2655       0.0241      0.0800       -0.1790
       7   -0.7672     0.1375    1.2769    -0.3063      -0.2506      0.0924        0.2065
       8    0.5046     0.1375    1.3841     0.2015      -0.0183     -0.0607        0.1358
       9    0.4651     0.1375    1.3972     0.1857       0.0732      0.0560       -0.1252
      10   -0.6044     0.1375    1.3473    -0.2413       0.1243     -0.0728       -0.1627
      11    1.8230     0.1375    0.7080     0.7279       0.2867      0.2195       -0.4907
      12    0.9778     0.1375    1.1712     0.3904      -0.2011      0.1177        0.2632
      13   -1.1397     0.2375    1.2250    -0.6360       0.0147     -0.4378        0.3263
      14   -2.1027     0.2375    0.6507    -1.1735       0.8388     -0.8077       -0.6020
      15    1.4897     0.2375    1.0022     0.8314      -0.0192      0.5722       -0.4265
      16    0.2457     0.2375    1.6425     0.1371      -0.0980      0.0944        0.0704

                   Sum of Residuals                           0
                   Sum of Squared Residuals            94.30000
                   Predicted Residual SS (PRESS)      148.37802
```

Appendix for Problem 3

The probability of Canadian salmon is modelled as "success". The following are the output of two logistic models.

R Output

```
glm(formula = group ~ x1 + x2 + x3, family = binomial)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.50501    6.39367   0.548 0.583555
x1           0.12642    0.03570   3.541 0.000398 ***
x2          -0.04865    0.01457  -3.339 0.000842 ***
x3           0.28156    0.83383   0.338 0.735614
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 138.629  on 99  degrees of freedom
Residual deviance:  38.674  on 96  degrees of freedom
AIC: 46.674
Number of Fisher Scoring iterations: 7
######################################################
glm(formula=group~ x1 + x2, family = binomial)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.92484    6.31518   0.621 0.534275
x1           0.12605    0.03586   3.515 0.000439 ***
x2          -0.04854    0.01452  -3.342 0.000831 ***
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 138.629  on 99  degrees of freedom
Residual deviance:  38.788  on 97  degrees of freedom
AIC: 44.788
Number of Fisher Scoring iterations: 7
######################################################
Analysis of Deviance Table
Model 1: group ~ x1 + x2
Model 2: group ~ x1 + x2 + x3
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        97     38.788
2        96     38.674  1  0.11461    0.735
######################################################
```

Sample Statistics

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} 98 \\ 430 \end{pmatrix} \quad \bar{\mathbf{x}}_2 = \begin{pmatrix} 137 \\ 367 \end{pmatrix} \quad \mathbf{S}_{\text{pooled}} = \begin{pmatrix} 676 & -649 \\ -649 & 2138 \end{pmatrix} \quad \mathbf{S}_{\text{pooled}}^{-1} = \begin{pmatrix} 0.002 & 0.001 \\ 0.001 & 0.001 \end{pmatrix}$$

Appendix for Problem 4

Descriptive Statistics of Two Random Samples

Population 1	$n_1 = 10$	$\bar{\mathbf{x}}_1 = \begin{pmatrix} 1.54 \\ 10.23 \end{pmatrix}$	$\mathbf{S}_1 = \begin{pmatrix} 11.1 & 4.8 \\ 4.8 & 2.7 \end{pmatrix}$	$\mathbf{S}_1^{-1} = \begin{pmatrix} 0.4 & -0.7 \\ -0.7 & 1.6 \end{pmatrix}$
Population 2	$n_2 = 10$	$\bar{\mathbf{x}}_2 = \begin{pmatrix} 1.71 \\ 8 \end{pmatrix}$	$\mathbf{S}_2 = \begin{pmatrix} 14.8 & 5.1 \\ 5.1 & 2.9 \end{pmatrix}$	$\mathbf{S}_2^{-1} = \begin{pmatrix} 0.2 & -0.3 \\ -0.3 & 0.9 \end{pmatrix}$
$\mathbf{S}_{\text{pooled}} = \begin{pmatrix} 13 & 5 \\ 5 & 2.8 \end{pmatrix}$ $\mathbf{S}_{\text{pooled}}^{-1} = \begin{pmatrix} 0.2 & -0.4 \\ -0.4 & 1.1 \end{pmatrix}$				

The 0.95th Quantiles of Some F_{df_1,df_2} Distributions

		df_2			
		16	17	18	19
df_1	1	4.494	4.451	4.414	4.381
	2	3.634	3.592	3.555	3.522
	3	3.239	3.197	3.160	3.127
	4	3.007	2.965	2.928	2.895

Quantiles of Some t_{df} Distributions

df	0.95th quantile	0.975th quantile	0.9875th quantile
18	1.734	1.729	1.725
19	2.101	2.093	2.086
20	2.445	2.433	2.423