Department of Mathematics and Statistics
The University of Toledo

Master's Comprehensive Examination
Applied Statistics

April 20, 2013

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.

Nonparametric Statistics:

1. (20 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 $Z_i = Y_i - X_i$.

   a. (5 points) Give the model assumptions required to be able to perform the 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 $(9, 8), (2, 3), (4, 7)$ and $(12, 15)$, can we reject H_0 in favor of $H_1 : \theta_X < \theta_Y$ at (nominal) level of significance $\alpha = 0.20$? Use the exact distribution from part b.

   d. (5 points) Use the formulas given below to find $var_0(T^+)$ in this case. Do you get the same answers as $var(T^+)$ in part b?

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

Linear Statistical Model:

2. (30 points) Creatinine clearance (y) is an important measure of kidney function, but is difficult to obtain in a clinical office setting because it requires 24-hour urine collection. To determine whether this measure can be predicted from some data that are easily available, a kidney specialist obtained the data that follow for 33 male subjects. The predictor vari11bles are serum creatinine concentration (x_1), age $(x_2$), and weight (x_3). The SAS output is given in the attached paper. Please answer following questions:

   a. (5 points) Use C_p criterion to choose the best subset of variables in the full model

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon.$$

   But theoretical arguments suggest use of the following regression function (use this model for b, c, d, e):

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln\left(140 - x_2\right) + \beta_3 \ln x_3 + \varepsilon.$$

   b. (5 points) Find $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$.

   c. (5 points) Obtain the variance inflation factors. Are there indications that serious multicollinearity problems exist here? Explain.

d. (5 points) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = 0.10$. State the decision rule and conclusion. Obtain the diagonal elements of the hat matrix and identify any outlying X observations.

e. (10 points) Cases 28 and 29 are relatively far outlying with respect to their Y values. Obtain DFFITS, DFBETAS, and Cook's distance values for these cases to assess their influence. What do you conclude?

(Hint: The sample size is 33, which we consider as a large sample size.)

3. (15 points) Suppose $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ where $\epsilon \sim MVN(\mathbf{0}, \mathbf{H})$, where $\mathbf{Y}^T = (\mathbf{Y}_1^T, \cdots, \mathbf{Y}_N^T)$ with $\mathbf{Y}_i^T = (Y_{i1}, \cdots, Y_{in})$.

a. (5 points) Give the formula for the ordinary least squares estimator $\tilde{\beta}$.

b. (5 points) Give the formula for the generalized least squares estimator $\hat{\beta}$.

c. (5 points) Which estimator do you prefer? Briefly explain.

4. (35 points) A study was conducted to test whether a new treatment for muscle strength is effective. The subjects were randomly assigned to the treatment group or placebo group. There are N subjects in each group. Measurements of muscle strength were taken at times 0 (baseline), 2, and 8 months. Assume that there is a correlation between the observations from the same subject and no correlation between the observations from any two different subjects.

a. (10 points) Suppose the observations are $\mathbf{Y}^T = (\mathbf{Y}_1^T, \cdots, \mathbf{Y}_{2N}^T)$, where $\mathbf{Y}_i^T = (Y_{i1}, Y_{i2}, Y_{i3})$ and Y_{ik} being the observation for the i-th subject at the k-th occasion. Let the baseline (month 0) and the placebo group as the reference group, and vector β is the linear model coefficients. Write down the linear model for this study using dummy variables. Label the variables in the model clearly.

b. (5 points) Let $\mathbf{L}$ denote a matrix of known weights. The null hypothesis that the patterns of change over time do not differ in the two treatment groups can be expressed as $H_0 : \mathbf{L}\beta = \mathbf{0}$. Describe an appropriate weight matrix $\mathbf{L}$ for this null hypothesis.

c. (10 points) The Appendix is the SAS code and outputs for the models with three different covariance structures. Choose the most appropriate model and *briefly explain.*

d. (10 points) Using the model you chose above, interpret the estimated regression coefficients in terms of the effect of the treatments on the patterns of change in muscle strength.

SAS OUTPUT

Problem 2.

```
data final2;
input y x1 x2 x3;
datalines;
  132.0   0.71   38.0   71.0
   53.0   1.48   78.0   69.0
   ......
   57.0   1.37   68.0   52.0
   ;
run;
proc reg data=final2;
model y = x1 x2 x3/selection = cp b;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y

C(p) Selection Method

Number of Observations Read	33
Number of Observations Used	33

Number in Model	C(p)	R-Square	Parameter Estimates: Intercept	x1	x2	x3
3	4.0000	0.8548	120.04728	-39.93933	-0.73677	0.77642
2	22.4041	0.7527	176.24154	-43.41076	-0.65689	.
2	29.1518	0.7189	104.32904	-53.85323	.	0.66441
1	42.3306	0.6429	154.66173	-55.55969	.	.
2	52.8666	0.6002	84.83155	.	-1.21762	0.94466
1	81.6508	0.4461	150.71892	.	-1.17040	.
1	146.8485	0.1197	24.96849	.	.	0.83043

```
data final2;
   set final2;
   newy = log(y);
   newx1 = log(x1);
   newx2 = log(140-x2);
   newx3 = log(x3);
run;

proc reg data=final2;
model newy = newx1 newx2 newx3 / r vif influence tol;
run;
quit;

data final2;
   set final2;
   newy = log(y);
   newx1 = log(x1);
   newx2 = log(140-x2);
```

```
   newx3 = log(x3);
run;
proc reg data=final2;
model newy = newx1 newx2 newx3 / r vif influence tol;
run;
quit;
```

The REG Procedure
Model: MODEL1
Dependent Variable: newy

Number of Observations Read	33
Number of Observations Used	33

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4.62887	1.54296	62.54	<.0001
Error	29	0.71551	0.02467		
Corrected Total	32	5.34438			

Root MSE	0.15708	R-Square	0.8661
Dependent Mean	4.37072	Adj R-Sq	0.8523
Coeff Var	3.59381		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	-2.04269	1.01919	-2.00	0.0545	.	0
newx1	1	-0.71195	0.09203	-7.74	<.0001	0.74665	1.33932
newx2	1	0.74736	0.15696	4.76	<.0001	0.75182	1.33011
newx3	1	0.75745	0.15923	4.76	<.0001	0.98422	1.01603

The REG Procedure
Model: MODEL1
Dependent Variable: newy

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	4.8828	4.8864	0.0500	-0.003632	0.149	-0.0244	| | |	0.000
2	3.9703	3.9698	0.0478	0.000518	0.150	0.00346	| | |	0.000
3	3.9120	3.9436	0.0659	-0.0316	0.143	-0.221	| | |	0.003
4	4.4067	4.3660	0.0641	0.0407	0.143	0.284	| | |	0.004
5	4.7005	4.7238	0.0621	-0.0233	0.144	-0.162	| | |	0.001
6	4.6052	4.6292	0.0561	-0.0241	0.147	-0.164	| | |	0.001
7	4.2195	4.1230	0.0555	0.0965	0.147	0.657	| |* |	0.015
8	4.5218	4.6108	0.0414	-0.0890	0.152	-0.587	| *| |	0.006
9	4.0943	4.1016	0.0367	-0.007273	0.153	-0.0476	| | |	0.000
10	4.5433	4.6207	0.0485	-0.0774	0.149	-0.518	| *| |	0.007
11	4.6540	4.4836	0.0407	0.1703	0.152	1.123	| |** |	0.023
12	4.5850	4.7136	0.0522	-0.1286	0.148	-0.868	| *| |	0.023
13	4.7185	4.7315	0.0587	-0.0130	0.146	-0.0890	| | |	0.000

14	4.8283	4.8458	0.0493	-0.0175	0.149	-0.117			0.000
15	4.6821	4.5210	0.0438	0.1611	0.151	1.068		|** |	0.024
16	3.4012	3.6018	0.0672	-0.2006	0.142	-1.412	**|		0.112
17	4.7095	4.5983	0.0459	0.1112	0.150	0.741		|* |	0.013
18	4.8675	4.7270	0.0556	0.1406	0.147	0.957		|* |	0.033
19	4.5433	4.4023	0.0544	0.1410	0.147	0.957		|* |	0.031
20	4.8675	4.6410	0.0684	0.2265	0.141	1.602		|*** |	0.150
21	4.0775	4.3153	0.0583	-0.2378	0.146	-1.630	***|		0.106
22	3.6376	3.7238	0.0697	-0.0863	0.141	-0.613	*|		0.023
23	4.1744	4.1084	0.0369	0.0659	0.153	0.432			0.003
24	4.4427	4.4087	0.0602	0.0339	0.145	0.234			0.002
25	4.9416	5.0503	0.0567	-0.1086	0.146	-0.742	*|		0.021
26	4.3820	4.5841	0.0647	-0.2021	0.143	-1.412	**|		0.102
27	3.7612	3.8109	0.0541	-0.0497	0.147	-0.337			0.004
28	4.3175	4.0077	0.0497	0.3098	0.149	2.079		|**** |	0.120
29	3.7136	4.0633	0.0431	-0.3497	0.151	-2.315	****|		0.109
30	4.7875	4.8941	0.0757	-0.1066	0.138	-0.774	*|		0.045
31	3.9512	4.1000	0.0379	-0.1487	0.152	-0.976	*|		0.015
32	4.2905	4.0041	0.0414	0.2863	0.152	1.890		|*** |	0.067
33	4.0431	3.9223	0.0606	0.1208	0.145	0.834		|* |	0.030

Output Statistics

Obs	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	newx1	newx2	newx3
1	-0.0240	0.1012	1.2802	-0.0080	0.0006	0.0048	-0.0017	0.0006
2	0.003400	0.0926	1.2681	0.0011	0.0007	-0.0000	-0.0008	-0.0002
3	-0.2177	0.1760	1.3870	-0.1006	0.0415	-0.0759	-0.0123	-0.0497
4	0.2794	0.1668	1.3657	0.1250	-0.0527	0.0228	-0.0230	0.1041
5	-0.1589	0.1565	1.3593	-0.0684	-0.0310	0.0474	0.0059	0.0382
6	-0.1611	0.1273	1.3137	-0.0616	-0.0292	0.0518	0.0377	0.0031
7	0.6500	0.1247	1.2382	0.2453	0.2071	-0.1181	-0.1928	0.1067
8	-0.5807	0.0696	1.1789	0.1568	0.0033	0.0902	0.0613	-0.0637
9	-0.0468	0.0545	1.2166	-0.0112	-0.0004	-0.0047	0.0024	-0.0021
10	-0.5113	0.0952	1.2253	-0.1658	0.0129	0.0793	0.0687	-0.0948
11	1.1279	0.0670	1.0325	0.3023	0.0601	-0.1594	-0.1694	0.0941
12	-0.8644	0.1104	1.1641	-0.3044	0.2203	-0.0238	-0.0880	-0.2467
13	-0.0875	0.1397	1.3362	-0.0352	-0.0135	0.0231	0.0003	0.0188
14	-0.1152	0.0985	1.2740	-0.0381	-0.0022	0.0258	-0.0030	0.0050
15	1.0705	0.0776	1.0626	0.3105	0.0166	-0.1453	-0.1567	0.1457
16	-1.4382	0.1828	1.0588	-0.6802	-0.0604	-0.4809	0.1009	-0.0127
17	0.7346	0.0855	1.1657	0.2246	-0.1398	0.0763	0.1791	0.0309
18	0.9553	0.1252	1.1570	0.3614	-0.3057	0.1200	0.2591	0.1989
19	0.9556	0.1198	1.1498	0.3525	-0.1836	0.2116	0.2902	-0.0178
20	1.6490	0.1897	0.9802	0.7980	-0.3991	0.2977	0.7076	-0.1159
21	-1.6808	0.1380	0.9088	-0.6725	-0.4385	0.2330	0.0718	0.5664
22	-0.6060	0.1969	1.3600	-0.3001	-0.2317	-0.0250	0.0992	0.2421
23	0.4257	0.0551	1.1868	0.1028	-0.0072	0.0541	-0.0077	0.0205
24	0.2300	0.1470	1.3388	0.0955	-0.0500	0.0580	0.0810	-0.0070
25	-0.7358	0.1303	1.2256	-0.2849	0.1118	0.1365	-0.0959	-0.0785
26	-1.4376	0.1699	1.0426	-0.6504	0.3584	-0.2938	-0.5806	0.0494
27	-0.3316	0.1185	1.2851	-0.1216	-0.0020	-0.0839	0.0129	-0.0108
28	2.2146	0.1001	0.6705	0.7385	0.5299	-0.1506	-0.5768	-0.1874
29	-2.5200	0.0753	0.5494	-0.7190	-0.1973	-0.3100	-0.1334	0.4202
30	-0.7687	0.2321	1.3783	-0.4226	0.1418	0.1660	0.1079	-0.3322
31	-0.9748	0.0581	1.0690	-0.2422	-0.0234	-0.0790	0.0817	-0.0535
32	1.9828	0.0694	0.7310	0.5414	0.1742	0.2497	0.0126	-0.2652
33	0.8290	0.1488	1.2269	0.3467	0.2705	-0.0153	-0.1063	-0.2906

Sum of Residuals	0
Sum of Squared Residuals	0.71551
Predicted Residual SS (PRESS)	0.91672

Appendix for Problem 4

SAS Program

```
*******************************************************************************
data exercise;
     infile 'F:\exercise.dat';
     input id group y0 y2 y8;
     y=y0; day=0; output;
     y=y2; day=2; output;
     y=y8; day=8; output;
     drop y0 y2 y8;
run;
data exercise;
     set exercise;
time=day;
run;
title1 Unstructured covariance;
proc mixed method=reml noclprint=10;
     class id group time;
     model y = group time group*time / s chisq;
     repeated time / type=un subject=id r rcorr;
run;
title1 Exponential covariance;
proc mixed noclprint=10;
     class id group time;
     model y = group time group*time / s chisq;
     repeated time / type=sp(exp)(day) subject=id r rcorr;
run;
title1 Autoregressive covariance;
proc mixed noclprint=10;
     class id group time;
     model y = group time group*time / s chisq;
     repeated time / type=ar(1) subject=id r rcorr;
run;
*******************************************************************************
```

SAS Output

```
*******************************************************************************
                Unstructured covariance for strength data                12
                          The Mixed Procedure
                           Model Information

Data Set                          WORK.EXERCISE
Dependent Variable                y
Covariance Structure              Unstructured
Subject Effect                    id
Estimation Method                 REML
Residual Variance Method          None
Fixed Effects SE Method           Model-Based
Degrees of Freedom Method         Between-Within

                    Class Level Information

  Class     Levels    Values

  id            37    not printed
  group          2    1 2
  time           3    0 4 8

                      Dimensions
```

Covariance Parameters	6
Columns in X	12
Columns in Z	0
Subjects	37
Max Obs Per Subject	3

Number of Observations

Number of Observations Read	111
Number of Observations Used	107
Number of Observations Not Used	4

Estimated R Matrix for id 1

Row	Col1	Col2	Col3
1	9.6683	10.1859	9.7866
2	10.1859	12.6032	12.5043
3	9.7866	12.5043	13.7122

Estimated R Correlation Matrix for id 1

Row	Col1	Col2	Col3
1	1.0000	0.9227	0.8500
2	0.9227	1.0000	0.9512
3	0.8500	0.9512	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	id	9.6683
UN(2,1)	id	10.1859
UN(2,2)	id	12.6032
UN(3,1)	id	9.7866
UN(3,2)	id	12.5043
UN(3,3)	id	13.7122

Fit Statistics

-2 Res Log Likelihood	412.5
AIC (smaller is better)	424.5
AICC (smaller is better)	425.4
BIC (smaller is better)	434.1

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
5	140.92	<.0001

Solution for Fixed Effects

Effect	group	time	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			82.5385	0.8119	35	101.66	<.0001
group	1		-1.3190	1.2334	35	-1.07	0.2922
group	2		0	.	.	.	.
time		0	-1.4909	0.4330	35	-3.44	0.0015
time		4	-0.5163	0.2651	35	-1.95	0.0595
time		8	0	.	.	.	.
group*time	1	0	-0.04108	0.6562	35	-0.06	0.9504
group*time	1	4	0.1094	0.3963	35	0.28	0.7842
group*time	1	8	0	.	.	.	.
group*time	2	0	0	.	.	.	.
group*time	2	4	0	.	.	.	.
group*time	2	8	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F

group	1	35	1.36	1.36	0.2437	0.2516
time	2	35	23.37	11.68	<.0001	0.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
group*time	2	35	0.22	0.11	0.8951	0.8955

****

Exponential covariance for strength data 15

The Mixed Procedure

Model Information

Data Set	WORK.EXERCISE
Dependent Variable	y
Covariance Structure	Spatial Exponential
Subject Effect	id
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
id	37	not printed
group	2	1 2
time	3	0 4 8

Dimensions

Covariance Parameters	2
Columns in X	12
Columns in Z	0
Subjects	37
Max Obs Per Subject	3

Number of Observations

Number of Observations Read	111
Number of Observations Used	107
Number of Observations Not Used	4

Estimated R Matrix for id 1

Row	Col1	Col2	Col3
1	11.6177	10.8084	10.0555
2	10.8084	11.6177	10.8084
3	10.0555	10.8084	11.6177

Estimated R Correlation Matrix for id 1

Row	Col1	Col2	Col3
1	1.0000	0.9303	0.8655
2	0.9303	1.0000	0.9303
3	0.8655	0.9303	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
SP(EXP)	id	55.3954
Residual		11.6177

Fit Statistics

-2 Res Log Likelihood	419.8
AIC (smaller is better)	423.8
AICC (smaller is better)	423.9
BIC (smaller is better)	427.0

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	133.62	<.0001

Solution for Fixed Effects

Effect	group	time	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			82.5891	0.7491	35	110.26	<.0001
group	1		-1.4053	1.1374	35	-1.24	0.2249
group	2		0	.	.	.	.
time		0	-1.5415	0.3958	66	-3.89	0.0002
time		4	-0.5803	0.2955	66	-1.96	0.0538
time		8	0	.	.	.	.
group*time	1	0	0.04520	0.5987	66	0.08	0.9400
group*time	1	4	0.2090	0.4416	66	0.47	0.6375
group*time	1	8	0	.	.	.	.
group*time	2	0	0	.	.	.	.
group*time	2	4	0	.	.	.	.
group*time	2	8	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
group	1	35	1.45	1.45	0.2286	0.2367
time	2	66	29.80	14.90	<.0001	<.0001
group*time	2	66	0.36	0.18	0.8364	0.8368

****

Autoregressive covariance for strength data 18

The Mixed Procedure

Model Information

Data Set	WORK.EXERCISE
Dependent Variable	y
Covariance Structure	Autoregressive
Subject Effect	id
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
id	37	not printed
group	2	1 2
time	3	0 4 8

Dimensions

Covariance Parameters	2
Columns in X	12
Columns in Z	0
Subjects	37
Max Obs Per Subject	3

Number of Observations

Number of Observations Read	111
Number of Observations Used	107
Number of Observations Not Used	4

Estimated R Matrix for id 1

Row	Col1	Col2	Col3
1	11.6191	10.8099	10.0570
2	10.8099	11.6191	10.8099
3	10.0570	10.8099	11.6191

Estimated R Correlation Matrix for id 1

Row	Col1	Col2	Col3
1	1.0000	0.9304	0.8656
2	0.9304	1.0000	0.9304
3	0.8656	0.9304	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
AR(1)	id	0.9304
Residual		11.6191

Fit Statistics

-2 Res Log Likelihood	419.8
AIC (smaller is better)	423.8
AICC (smaller is better)	423.9
BIC (smaller is better)	427.0

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	133.62	<.0001

Solution for Fixed Effects

Effect	group	time	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			82.5891	0.7491	35	110.25	<.0001
group	1		-1.4053	1.1375	35	-1.24	0.2249
group	2		0	.	.	.	.
time		0	-1.5415	0.3958	66	-3.89	0.0002
time		4	-0.5803	0.2955	66	-1.96	0.0538
time		8	0	.	.	.	.
group*time	1	0	0.04519	0.5987	66	0.08	0.9401
group*time	1	4	0.2090	0.4416	66	0.47	0.6375
group*time	1	8	0	.	.	.	.
group*time	2	0	0	.	.	.	.
group*time	2	4	0	.	.	.	.
group*time	2	8	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
group	1	35	1.45	1.45	0.2287	0.2367
time	2	66	29.80	14.90	<.0001	<.0001
group*time	2	66	0.36	0.18	0.8364	0.8368