

Appendix C

Datasets

In this appendix, we provide a brief description of each of the SAS datasets used in the book in the order in which they first occur. A partial or full listing of the data is presented either in this appendix or in the designated chapter as indicated. Also, the datasets and programs used in this book are all available through the author's Web page at

<http://support.sas.com/publishing/authors/vonesh.html>.

In some of the examples, the SAS datasets are stand alone datasets that are called from within the SAS program while in other examples, the dataset will be provided within the program itself (such as described for the dental growth data below). The output/data analyses for this book were generated using SAS software (Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA).

C.1 Dental growth data

The dental growth data listed in Output 2.1 of Chapter 2 are from Potthoff and Roy (1964). The data consists of orthodontic measurements (the distance in mm from the center of the pituitary to the pterygomaxillary fissure) of 16 boys and 11 girls measured successively at ages 8, 10, 12 and 14 years. The initial dataset is contained within the SAS program,

Example 2.2.1 - Dental growth data.sas

and is arranged vertically within the program as follows:

```
data dental_data;
  input gender person age y;
  if gender=1 then sex='boys';
    else sex='girls';
  _age_=age;
cards;
1 1 8 26
1 1 10 25
1 1 12 29
1 1 14 31
1 2 8 21.5
1 2 10 22.5
1 2 12 23
1 2 14 26.5
:
```

The data are from Table 1 of Potthoff and Roy (1964) and are made available here with permission from Oxford University Press.

C.2 Bone mineral density data

The bone mineral density data are from Vonesh and Chinchilli (1997; Table 5.4.1, pp. 228-230) and reproduced here with permission from Taylor & Francis Group, LLC - Books. The original data are arranged horizontally and made available as dataset TBBMD within the SAS program

Example 2.2.2 - Bone mineral density data.sas

shown in §2.2.2 and partially reproduced below.

```
data TBBMD;
  input (subject group date1 tbbmd1 date2 tbbmd2 date3 tbbmd3
  date4 tbbmd4 date5 tbbmd5)
(3. @5 $char1. @7 date7. @15 5.3 @21 date7.
@29 5.3 @35 date7. @43 5.3 @49 date7. @57 5.3
@63 date7. @71 5.3);
```

```

if tbbmd1^=. & tbbmd2^=. & tbbmd3^=. &
   tbbmd4^=. & tbbmd5^=. then data='Complete';
else data='Incomplete';
format date1 date2 date3 date4 date5 date7. ;
cards;
101 C 01MAY90 0.815 05NOV90 0.875 24APR91 0.911 30OCT91 0.952 29APR92 0.970
102 P 01MAY90 0.813 05NOV90 0.833 15APR91 0.855 21OCT91 0.881 13APR92 0.901
103 P 02MAY90 0.812 05NOV90 0.812 17APR91 0.843 23OCT91 0.855 15APR92 0.895
104 C 09MAY90 0.804 12NOV90 0.847 29APR91 0.885 11NOV91 0.920 19JUN92 0.948
105 C 14MAY90 0.904 10DEC90 0.927 06MAY91 0.952 04NOV91 0.955 27APR92 1.002
106 P 15MAY90 0.831 26NOV90 0.855 20MAY91 0.890 17DEC91 0.908 15JUL92 0.933
107 P 23MAY90 0.777 05DEC90 0.803 15MAY91 0.817 27NOV91 0.809 04MAY92 0.823
108 C 23MAY90 0.792 03DEC90 0.814
109 C 23MAY90 0.821 03DEC90 0.850 06MAY91 0.865 05NOV91 0.879 29APR92 0.908
110 P 30MAY90 0.823 10DEC90 0.827 21MAY91 0.839 25NOV91 0.885 15MAY92 0.923
111 C 30MAY90 0.828 05DEC90 0.873 07AUG91 0.935 12FEB92 0.952
112 P 04JUN90 0.797 17DEC90 0.818 13MAY91 0.817 16DEC91 0.847 18MAY92 0.862
:

```

A partial listing of the data, shown above, includes the subject ID, treatment group, date of each visit and the TBBMD response at that visit.

C.3 ADEMEX adequacy data

The ADEMEX adequacy data are from Paniagua et. al. (2002) and were made available courtesy of Dr. Ramon Paniagua. The data are available as a SAS dataset, ADEMEX_adequacy_data. The following is a partial listing of the full dataset made available to users through the SAS Press Companion Web site at support.sas.com/companionsites.

Obs	ptid	Visit	Trt	pKtV	pCCr	Month	NumberActive	TriYear
1	01 001	2	1	1.66953	44.9940	-1	481	0
2	01 002	2	0	1.84879	53.2581	-1	484	0
3	01 002	8	0	1.60299	48.4590	4	439	4
4	01 002	10	0	1.89197	54.7183	9	374	12
5	01 002	12	0	1.66400	48.2390	13	332	16
6	01 002	14	0	1.65904	44.6819	18	281	20
7	01 002	16	0	1.20457	35.5130	22	238	24
8	01 002	18	0	1.72249	45.6907	26	165	28
9	01 004	2	0	1.71246	52.1098	-1	484	0
10	01 004	8	0	1.66711	46.4633	4	439	4

The variables `ptid` and `Trt` represent the patient ID and treatment group indicator which is 0 if a patient was randomized to the control (standard dose) group and 1 if to the treated (high dose) group. The variables `pKtV` and `pCCr` are measures of dialysis adequacy and represent, respectively, peritoneal urea Kt/V which is defined as urea clearance, Kt (in L/week), divided by total body water, V (in L), and peritoneal creatinine clearance, or `pCCr` (L/week/1.73m²), which is creatinine clearance normalized to a 1.73 m² body surface area. The variable `Month` is the time elapsed from the date of randomization to the date of the adequacy measurements rounded to the nearest whole month. A value of -1 corresponds to the time at which baseline adequacy measurements were collected just prior to randomization to ensure patients met the inclusion/exclusion criteria of the study. The variables `NumberActive` and `TriYear` are derived SAS variables representing the number of active patients that are in the study at baseline (month -1 set equal to 0 for plotting purposes) and at yearly tertiles of follow-up (i.e., `TriYear` = 4, 8, 12, 16,... months post randomization). These last two variables are used only in conjunction with the ANNOTATE feature of PROC GPLOT to create Figure 2.1.

C.4 MCM2 biomarker data

The MCM2 biomarker data are from Helenowski et. al. (2011) and were made available courtesy of Irene Helenowski. The data are available as a SAS dataset, `MCM2_data`. The following is a partial listing of the data.

Obs	Row	Col	Subject	MCM2Index	Core
1	4	2	16	23.8908	1
2	3	2	16	26.1251	2
3	2	2	16	31.0541	3
4	1	2	16	32.0144	4
5	4	1	16	.	6
6	3	1	16	31.9534	7
7	2	1	16	26.7631	8
8	1	1	16	32.5031	9
9	4	2	17	52.9183	1
10	3	2	17	36.4055	2

The variables `Row` and `Col` represent the spatial coordinates of the prostate gland core sample (`Core`) from which the response variable, `MCM2index`, was obtained.

C.5 Estrogen hormone data

The estrogen hormone data are from Gann et. al. (2006) and were made available courtesy of Dr. Peter Gann. The data are available as a SAS dataset, `Estradiol_data`. A partial listing of the data is given in Output 3.4. The data consist of each subject's ID (`Subject_ID`), the treatment group to which the subject was randomized (`Group`), the visit at which measurements were obtained (`Visit`), the breast on which nipple aspirate fluid (NAF) samples were obtained (`Breast`), and the log concentrations of NAF estradiol (`NAF_Estradiol`), saliva estradiol (`Saliva_Estradiol`) and serum estradiol (`Serum_Estradiol`).

C.6 ADEMEX peritonitis and hospitalization data

The ADEMEX peritonitis and hospitalization data (Examples 4.3.1, 5.2.4, and 7.1.1) are from Paniagua et. al. (2002) and were made available courtesy of Dr. Ramon Paniagua. The data are available as a SAS dataset, `ADEMEX_Peritonitis_data`. A partial listing of the data is given in Output 4.1. The data consist of the patient's ID (`ptid`), treatment group (`Trt`), age (`Age`), gender (`Sex`, 0=male, 1=female), diabetic status (`Diabetic`, 0=non-diabetic, 1=diabetic), baseline serum albumin (`Albumin`, in g/dL), months on dialysis prior to randomization (`PriorMonths`), months at risk for peritonitis (`MonthsAtRisk`), episodes of peritonitis (`Episodes`) and number of hospital admissions (`Hosp`).

C.7 Respiratory disorder data

The respiratory disorder data are from Stokes et. al. (2000) and is available in the SAS Sample Library under Example 5 for PROC GENMOD. A partial listing of the data as analyzed in the book and which is based on the dataset

made available courtesy of SAS Institute Inc., is given in Output 4.5. The data consists of a patient's center id (**Center**), the patient's ID (**ID**), the treatment group to which the patient was randomized (**Treatment**), patient gender (**Sex**), age (**Age**), the baseline binary response (**Baseline**, 0=negative response, 1=positive response), the visit at which follow-up responses were obtained (**Visit**) and the binary response at each follow-up visit (**y**).

C.8 Epileptic seizure data

The epileptic seizure data are from Thall and Vail (1990) and is available in the SAS Sample Library under Example 7 for PROC GENMOD. A partial listing of the data, which were made available courtesy of SAS Institute Inc., is given in Output 4.8. The data consists of the patient's ID (**ID**), the number of epileptic seizures in a two-week period (**y**), the visit at which these measurements were obtained (**Visit**), the baseline seizure count over an eight (8) week period (**bline**), and the patient's age (**Age**). Included in that listing is the log baseline count normalized to a two week period (**y0**).

C.9 Schizophrenia data

The schizophrenia data were made available courtesy of Dr. Donald Hedeker with permission from Dr. Nina Schooler. The data are available as a SAS dataset, **Schizophrenia_data**. A partial listing of the data is given in Output 4.14. The data consists of the patient's ID (**ID**), the IMPS79 four-point ordinal score (**IMPS79o**), the treatment group to which the patient was randomized (**Trt**, 0=placebo control, 1=drug), the week of follow-up at which measurements were obtained (**Week**) and the square root of week (**SWeek**).

C.10 LDH enzyme leakage data

The LDH enzyme leakage data are from Gennings, Chinchilli and Carter (1989) and are reproduced here with permission from *The Journal of the American Statistical Association* (Copyright 1989 by the American Statistical Association. All rights reserved). The data are available as a SAS dataset, **LDH_Ezyme_data**. A partial listing of the data is given in Output 4.20. The data as listed consists of the flask ID containing the isolated

hepatocyte suspensions (Flask) as well as the SAS variables: LDH0 LDH001 LDH025 LDH05 LDH1 LDH2 LDH3 which represent the percentage LDH leakage (i.e., $100 \times$ proportion) at times 0, 0.01, 0.25, 0.5, 1.0, 2.0 and 3.0 hours, respectively.

C.11 Orange tree data

The orange tree data are from Draper and Smith (1981) and are reproduced here with permission from John Wiley and Sons, Inc. A complete listing of the data is given in Output 4.23 and consists of the tree ID (**Tree**), the days on which trunk circumference was measured (**Days**) and the trunk circumference of the tree measured in mm (**y**).

C.12 Soybean growth data

The soybean growth data are from Davidian and Giltinan (1993a) and are reproduced here with permission from the publisher (Taylor and Francis Group, <http://www.informaworld.com>). The data are available as a SAS dataset, **Soybean_data**. The following is a partial listing (first 20 plots) from the full dataset.

Year	Plot	Genotype	D1	W1	D2	W2	D3	W3	D4	W4	D5	W5
1988	1	F	14	0.106	21	0.261	28	0.666	35	2.110	42	3.560
1988	2	F	14	0.104	21	0.269	28	0.778	35	2.120	42	2.930
1988	3	F	14	0.108	21	0.291	28	0.667	35	2.050	42	3.810
1988	4	F	14	0.105	21	0.299	28	0.844	35	1.320	42	2.240
1988	5	F	14	0.101	21	0.273	28	0.848	35	1.950	42	4.770
1988	6	F	14	0.106	21	0.337	28	0.699	35	1.530	42	3.870
1988	7	F	14	0.102	21	0.275	28	0.767	35	1.450	42	3.950
1988	8	F	14	0.103	21	0.273	28	0.742	35	1.410	42	3.010
1988	9	P	14	0.131	21	0.338	28	0.701	35	1.660	42	4.250
1988	10	P	14	0.128	21	0.404	28	0.897	35	1.780	42	3.910
1988	11	P	14	0.131	21	0.379	28	1.126	35	2.440	42	3.890
1988	12	P	14	0.154	21	0.357	28	1.181	35	1.830	42	4.710
1988	13	P	14	0.139	21	0.328	28	0.932	35	1.990	42	3.460
1988	14	P	14	0.139	21	0.389	28	1.094	35	2.130	42	4.040
1988	15	P	14	0.145	21	0.366	28	0.799	35	1.610	42	3.510
1988	16	P	14	0.130	21	0.355	28	1.090	35	2.280	42	3.940
1989	17	F	14	0.047	20	0.136	27	0.272	34	0.983	41	1.796
1989	18	F	14	0.035	20	0.112	27	0.268	34	1.053	41	1.715
1989	19	F	14	0.044	20	0.146	27	0.360	34	0.775	41	1.641
1989	20	F	14	0.049	20	0.097	27	0.227	34	0.833	41	1.354

Year	Plot	Genotype	D6	W6	D7	W7	D8	W8	D9	W9	D10	W10
1988	1	F	49	6.230	56	8.710	63	13.350	70	16.341	77	17.750
1988	2	F	49	5.290	56	9.500	63	.	70	16.966	77	17.746
1988	3	F	49	6.130	56	10.280	63	18.080	70	20.182	77	21.810
1988	4	F	49	4.680	56	8.820	63	15.090	70	14.660	77	14.004
1988	5	F	49	6.010	56	9.910	63	.	70	19.262	77	.
1988	6	F	49	5.600	56	9.430	63	13.730	70	17.380	77	19.932
1988	7	F	49	4.940	56	9.640	63	.	70	17.880	77	17.724
1988	8	F	49	5.260	56	9.810	63	12.850	70	18.214	77	19.680
1988	9	P	49	9.240	56	12.150	63	16.780	70	15.925	77	17.272
1988	10	P	49	7.400	56	10.070	63	18.860	70	17.011	77	27.370
1988	11	P	49	6.910	56	12.490	63	15.670	70	23.763	77	21.490
1988	12	P	49	10.710	56	9.910	63	15.510	70	14.957	77	21.800
1988	13	P	49	7.020	56	11.790	63	15.830	70	15.920	77	17.441
1988	14	P	49	7.620	56	12.480	63	17.930	70	14.421	77	30.271
1988	15	P	49	6.790	56	9.950	63	14.540	70	19.280	77	22.572
1988	16	P	49	4.960	56	10.920	63	14.020	70	17.994	77	22.370
1989	17	F	55	4.117	69	11.037	84	10.403
1989	18	F	55	6.728	69	10.313	84	11.088
1989	19	F	55	3.748	69	10.875	84	16.027
1989	20	F	55	5.988	69	10.585	84	10.911

The variables `Year`, `Plot` and `Genotype` are the calendar year in which measurements were taken, an ID variable identifying the plot of land from which plants were sampled, and the genotype of soybean planted, respectively. The variables `D1-D10` represent the days after planting on which plants were sampled and `W1-W10` are the corresponding average leaf weights from the sampled plants.

C.13 High flux hemodialyzer data

The high flux hemodialyzer data are from Vonesh and Carter (1992) and the dataset, which is available with the software program MIXNLIN 3.1 (Copyright 1995 by Edward F. Vonesh), was made available here with permission from the author. The data are available as a SAS dataset, `Hemodialyzer_data`. A partial listing of the data (data from the first two dialyzers) is presented below.

Obs	Dialyzer	Qb	TMP	UFR
1	K80790	200	24.0	64.5
2	K80790	200	50.5	2011.5
3	K80790	200	99.5	3846.0
4	K80790	200	148.5	4498.5

5	K80790	200	202.0	5176.5
6	K80790	200	249.5	4657.5
7	K80790	200	297.0	4081.5
8	K80789	200	24.0	372.0
9	K80789	200	54.0	1888.5
10	K80789	200	99.5	3469.5
11	K80789	200	147.5	4030.5
12	K80789	200	200.0	4447.5
13	K80789	200	250.0	4243.5
14	K80789	200	301.0	4465.5

The data consists of the dialyzer ID (Dialyzer), the blood flow rate at which the dialyzer was run (Q_b , in ml/min), the transmembrane pressures at which ultrafiltration was measured (TMP, in mmHg) and the ultrafiltration rate (UFR, in ml/hour).

C.14 Cefamandole pharmacokinetic data

The cefamandole PK data are from Davidian and Giltinan (1993b) and are reproduced here with permission from Elsevier. The data are available as a SAS dataset, Cefamandole_data. The following is a partial listing (first two subjects) from the full dataset.

Obs	Subject	time	Conc
1	1	10	127.00
2	1	15	87.00
3	1	20	47.40
4	1	30	39.90
5	1	45	24.80
6	1	60	17.90
7	1	75	11.70
8	1	90	10.90
9	1	120	5.70
10	1	150	2.55
11	1	180	1.84
12	1	240	1.50
13	1	300	0.70
14	1	360	0.34
15	2	10	120.00

16	2	15	90.10
17	2	20	70.00
18	2	30	40.10
19	2	45	24.00
20	2	60	16.10
21	2	75	11.60
22	2	90	9.20
23	2	120	5.20
24	2	150	3.00
25	2	180	1.54
26	2	240	0.73
27	2	300	0.37
28	2	360	0.19

The data consists of the subject ID (`Subject`), time at which concentrations were obtained (`time`) and the concentration of cefamandole in $\mu\text{g}/\text{mL}$ (`Conc`).

C.15 MDRD data

The MDRD data are provided courtesy of Dr. Gerald Beck from the Cleveland Clinic Foundation on behalf of the MDRD Study Group. The data are available as a SAS dataset, `MDRD_data`. The following is a partial listing (first two subjects) of the data.

Obs	ptid	GFR	Months	DietL_NormBP	DietL_LowBP	log_protein
1	1	24.33	0	1	0	-1.47
2	1	23.75	1.74	1	0	-1.47
3	1	21.84	4.5	1	0	-1.47
4	1	18.72	7.95	1	0	-1.47
5	1	15.51	11.89	1	0	-1.47
6	1	14.97	16.26	1	0	-1.47
7	1	12.67	20.14	1	0	-1.47
8	1	12.08	24.34	1	0	-1.47
9	1	14.15	28.42	1	0	-1.47
10	1	10.57	32.16	1	0	-1.47
11	1	10.27	35.84	1	0	-1.47
12	2	18.82	0	1	0	-2.3
13	2	22.61	2.04	1	0	-2.3
14	2	26.12	4.4	1	0	-2.3
15	2	14.05	8.25	1	0	-2.3

16	2	20.19	12.22	1	0	-2.3
17	2	22.61	16.13	1	0	-2.3
18	2	22.63	24.11	1	0	-2.3
19	2	25.09	28.29	1	0	-2.3
20	2	22.63	32.43	1	0	-2.3
21	2	23.3	36.07	1	0	-2.3
22	2	26.71	39.79	1	0	-2.3

Obs	ptid	DietK_NormBP	DietK_LowBP	FollowupMonths	Dropout
1	1	0	0	38.37	0
2	1	0	0	38.37	0
3	1	0	0	38.37	0
4	1	0	0	38.37	0
5	1	0	0	38.37	0
6	1	0	0	38.37	0
7	1	0	0	38.37	0
8	1	0	0	38.37	0
9	1	0	0	38.37	0
10	1	0	0	38.37	0
11	1	0	0	38.37	0
12	2	0	0	39.79	0
13	2	0	0	39.79	0
14	2	0	0	39.79	0
15	2	0	0	39.79	0
16	2	0	0	39.79	0
17	2	0	0	39.79	0
18	2	0	0	39.79	0
19	2	0	0	39.79	0
20	2	0	0	39.79	0
21	2	0	0	39.79	0
22	2	0	0	39.79	0

The data consists of the subject ID (ptid), GFR in ml/min/1.73 m² (GFR), time at which GFR was measured (Months), four 0-1 indicator variables indicating which treatment group patients were randomized to (DietK_LowBP, DietK_NormBP, DietL_LowBP, DietL_NormBP), the log baseline urine protein concentration (log_protein), the time to dropout (FollowupMonths), and a binary indicator variable of dropout (Dropout, 0=administratively censored, 1=dropout). While we did not include the variable log_protein in any of the analyses presented in Chapter 6, §6.6.2, it was included with the dataset as

a possible baseline predictor of a subject's GFR slope.

C.16 Theophylline data

The theophylline data are from Boeckmann et. al. (1992) and subsequently Pinheiro and Bates (1995) and is available in the SAS Sample Library under Documentation Example 1 for PROC NLMIXED. A partial listing of the data, as made available courtesy of SAS Institute Inc., is given in Output 7.3. The data consists of the subject ID (**subject**), time in hours at which concentrations were obtained (**time**), the concentration of theophylline in mg/L (**conc**), the dose per kg body weight (**dose**) and the subject's baseline weight in kg (**wt**).

C.17 Phenobarbital data

The phenobarbital data are from Grasela and Donn (1985) and are made available courtesy of the Resource Facility for Population Kinetics (RFPK) from their website at <http://www.rfpk.washington.edu> with funding support from NIH/NIBIB grant P41-EB01975. The data were downloaded from the RFPK website and a SAS dataset created using the following programming statements.

```
data pheno;
  input ID time amt wt vapgar conc EVID EVID1;
  datalines;
    1   0.    25.0   1.4    7     .      1 1
    1   2.0    .     1.4    7    17.3    0 0
    1   12.5   3.5   1.4    7     .      1 1
    1   24.5   3.5   1.4    7     .      1 1
    1   37.0   3.5   1.4    7     .      1 1
    1   48.0   3.5   1.4    7     .      1 1
    1   60.5   3.5   1.4    7     .      1 1
    1   72.5   3.5   1.4    7     .      1 1
    1   85.3   3.5   1.4    7     .      1 1
    1   96.5   3.5   1.4    7     .      1 1
    1  108.5   3.5   1.4    7     .      1 1
    1  112.5    .     1.4    7    31.0    0 0
```

```

2    0.    15.0    1.5    9    .    1 1
2    2.0    .    1.5    9    9.7    0 0
2    4.0    3.8    1.5    9    .    1 1
2   16.0    3.8    1.5    9    .    1 1
:
data PHENO.Phenobarbital_data(rename=(ID=ID
amt=dose wt=weight vapgar=apgar EVID=eventid));
set pheno;
drop EVID1;
run;

```

The data consists of the subject ID (ID), the time in hours following the administration of the initial dose (time), the dose per kg body weight in $\mu\text{g}/\text{kg}$ (dose), the weight at birth in kg (weight), the five minute Apgar score (apgar), the serum concentration of phenobarbital in $\mu\text{g}/\text{L}$ (conc), and a binary event indicator variable (eventid) that defines whether an initial or intermittent dose was delivered at a given time point (eventid=1) or whether a serum concentration was measured (eventid=0) at that time point.

C.18 ADEMEX GFR and survival data

The ADEMEX GFR and survival data (Examples 7.3.1) are from Paniagua et. al. (2002) and were made available courtesy of Dr. Ramon Paniagua. The data are available as a SAS dataset, ADEMEX_GFR_data, for which a partial listing of the data is given below.

Obs	ptid	Trt	PriorMonths	Age	Sex	Diabetic	GFR0	nPNAO
1	01 001	1		2	34	1	0	3.3198 0.85084
2	01 028	0		1	54	0	0	5.1160 .
3	01 028	0		1	54	0	0	5.1160 .
4	01 028	0		1	54	0	0	5.1160 .
5	01 028	0		1	54	0	0	5.1160 .
6	01 028	0		1	54	0	0	5.1160 .
7	01 028	0		1	54	0	0	5.1160 .
8	01 028	0		1	54	0	0	5.1160 .
9	01 034	0		3	69	0	0	1.0083 0.47856
10	01 034	0		3	69	0	0	1.0083 0.47856
11	01 034	0		3	69	0	0	1.0083 0.47856
12	01 035	1		2	37	1	0	1.1291 0.82493

13	01	035	1	2	37	1	0	1.1291	0.82493
14	01	035	1	2	37	1	0	1.1291	0.82493
15	01	035	1	2	37	1	0	1.1291	0.82493
16	01	035	1	2	37	1	0	1.1291	0.82493

Obs	ptid	Albumin0	Months	GFR_ml_min	ITTdeath	ITTtime
1	01 001	2.96	0.0000	3.3198	0	28.5197
2	01 028	3.48	0.0000	5.1160	0	32.3026
3	01 028	3.48	5.0000	10.8301	0	32.3026
4	01 028	3.48	8.9145	4.8665	0	32.3026
5	01 028	3.48	13.3224	6.3700	0	32.3026
6	01 028	3.48	16.9737	2.8615	0	32.3026
7	01 028	3.48	20.6908	1.0970	0	32.3026
8	01 028	3.48	25.7895	0.8632	0	32.3026
9	01 034	2.37	0.0000	1.0083	0	31.4474
10	01 034	2.37	5.8224	0.0533	0	31.4474
11	01 034	2.37	9.6711	0.0292	0	31.4474
12	01 035	2.29	0.0000	1.1291	0	29.3750
13	01 035	2.29	5.4934	0.4041	0	29.3750
14	01 035	2.29	11.0526	1.4380	0	29.3750
15	01 035	2.29	19.1447	1.0409	0	29.3750
16	01 035	2.29	23.2566	0.8830	0	29.3750

The data consists of the subject ID (ptid), treatment group (Trt, 0=control or standard dose, 1=treated or high dose), prior months on dialysis prior to randomization (PriorMonths), patient age (Age), gender (Sex, 0=male, 1=female), diabetic status (Diabetic, 0=non-diabetic, 1=diabetic), baseline GFR in ml/min (GFR0), baseline value of normalized protein nitrogen appearance or nPNA in g/kg/day (nPNA0), baseline serum albumin (Albumin0), times at which GFR were measured (Months), GFR in ml/min (GFR_ml_min), a binary indicator of whether a patient died (ITTdeath, 0=censored, 1=died) and the corresponding patient survival time in months (ITTtime).

Appendix D

Select SAS macros

In this appendix we present select SAS macros that are used in several examples throughout the book. Each macro is available through the author's Web page at

<http://support.sas.com/publishing/authors/vonesh.html>.

The macros are provided as is without any expressed or implied warranty.

D.1 The GOF Macro

The %GOF macro computes summary goodness-of-fit measures based on methods described by Vonesh et. al. (1996) and Vonesh and Chinchilli (1997). Specifically, it takes output generated from MIXED, GENMOD, GLIMMIX, or NLMIXED and summarizes various goodness-of-fit measures using R-square type criteria as well as concordance correlation coefficients described by Vonesh et. al. (1996) and Vonesh and Chinchilli (1997). Both the R-square and concordance correlation coefficients provide goodness-of-fit measures that can prove useful when assessing model fit and when performing variable selection in the LME, GLME and GNLME regression setting (see Vonesh et. al. 1996; Vonesh and Chinchilli, 1997). In addition, one can use the %GOF macro to provide an additional test for determining whether the "working" covariance structure may be reasonable. This test has been shown to provide reasonable type I and type II errors for models with Gaussian errors but it can be sensitive to models where the distribution of the residual errors is highly skewed (Vonesh et. al., 1996).

Formulas for the R-square and concordance correlation coefficients used in the macro are briefly described in Example 5.2.1. For evaluating the goodness-of-fit of fixed effects, the average model R^2 computed in %GOF (as well as %GLIMMIX_GOF in the following section) is

$$R_{Avg}^2 = 1 - \frac{\sum_{i=1}^n (y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^n (y_{ij} - \bar{y})^2} \quad (\text{D.1})$$

where $\hat{y}_{ij} = E(y_{ij} | \mathbf{b}_i = \mathbf{0}) = \mu_{ij}(\mathbf{x}'_{ij}, \hat{\boldsymbol{\beta}}, \mathbf{z}_{ij}, \mathbf{0}) = \mu_{ij}(\hat{\boldsymbol{\beta}}, \mathbf{0})$ is the i^{th} subject's predicted value under the mixed-effects model (LME, GLME or GNLME) but evaluated at the average random effect, $\mathbf{b}_i = \mathbf{0}$, and $\bar{y} = \sum_{i=1}^n y_{ij}/N$ is the overall mean of the response variable. The conditional model R^2 , denoted by R_{Cond}^2 , is defined in the same way as in (D.1) but with $\hat{y}_{ij} = E(y_{ij} | \hat{\mathbf{b}}_i) = \mu_{ij}(\mathbf{x}'_{ij}, \hat{\boldsymbol{\beta}}, \mathbf{z}_{ij}, \hat{\mathbf{b}}_i) = \mu_{ij}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}_i)$ representing the i^{th} individual's predicted value evaluated at the individual's estimated random effect, $\mathbf{b}_i = \hat{\mathbf{b}}_i$.

The average model R^2 in (D.1) measures a reduction in residual variation explained by the fixed-effects alone while the corresponding conditional model R^2 measures a reduction in explained variation due to fitting both fixed and random effects versus explained variation due to fitting an overall mean. Both of these R^2 measures are a special case of a more general formulation described by Vonesh and Chinchilli (1997) (see also Edwards et. al., 2008) in which one can evaluate goodness-of-fit based on an R^2 statistic that may interpreted as "explained residual variation" of \mathbf{y}_i relative to some specified null model and some specified covariance structure.

The macro also computes a goodness-of-fit measure based on a concordance correlation coefficient developed by Lin (1989) and adapted to goodness-of-fit by Vonesh et. al. (1996) and Vonesh and Chinchilli (1997). The average model concordance correlation is defined as

$$\hat{\rho}_{c,Avg} = 1 - \frac{\sum_{i=1}^n (y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^n (y_{ij} - \bar{y})^2 + \sum_{i=1}^n (\hat{y}_{ij} - \bar{y})^2 + N(\bar{y} - \hat{\bar{y}})^2} \quad (\text{D.2})$$

where $\hat{\bar{y}} = \sum_{i=1}^n \hat{y}_{ij}/N$ is the overall mean of the individual predicted values evaluated at the average random effect, i.e., $\hat{y}_{ij} = \mu_{ij}(\hat{\boldsymbol{\beta}}, \mathbf{0})$. The conditional concordance correlation, $\hat{\rho}_{c,Cond}$, is also given by (D.2) but with the predicted values evaluated at the subject-specific random effects, i.e., $\hat{y}_{ij} = \mu_{ij}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}_i)$. The concordance correlation is a measure of how well the predicted values agree with the observed values. As a measure of agreement, it reflects how

well a scatter plot of observed versus predicted values falls about the line of identity. Consequently, it does not require specification of a null model (e.g., an overall intercept model with no fixed effects) that is otherwise required for R^2 type goodness-of-fit measures (Vonesh and Chinchilli, 1997).

A comparative evaluation of how well these goodness-of-fit statistics along with several other proposed goodness-of-fit statistics perform for LME models was investigated by Orelien and Edwards (2008). Based on simulations, Orelien and Edwards (2008) demonstrate that when the goal is to identify an appropriate model based on the fixed effects alone (i.e., fixed-effects variable selection), one should avoid the use of the conditional model R_{Cond}^2 and $\hat{\rho}_{c,Cond}$, in favor of the average model R_{Avg}^2 and $\hat{\rho}_{c,Avg}$. An alternative R-square statistic suitable for evaluating goodness-of-fit of the fixed effects components of a LME model was proposed by Edwards et. al. (2008). Their proposed R^2 statistic measures multivariate association between repeated measurements and the fixed effects of a LME model fit to longitudinal data. It offers a number of advantages including 1) the ability to compare a full model with a null model obtained by excluding all fixed effects (except typically the intercept) while retaining the same covariance structure for both models and 2) the ability to extend this measure to a partial R^2 statistic. It is not entirely clear whether their proposed R^2 statistic can be applied in a straightforward manner to GLME and GNLME models but this is certainly an area worth pursuing.

Presented below is a detailed description of the macro arguments one must supply whenever one makes a call to the macro as well as the basic syntax used to call the macro. In terms of the macro arguments one must provide, the PROC= argument is required. It informs the macro which procedure was used to obtain parameter estimates. If one uses the macro %NLINMIX for a particular analysis, one can still use the %GOF macro but when specifying which procedure was used, one would specify PROC=MIXED since the MIXED procedure is used by the macro %NLINMIX to iteratively compute parameter estimates for the specified model. The %GOF macro will then use the final parameter estimates from the last call to MIXED from within the %NLINMIX macro. The PARMS= and COVPARMS= arguments instruct the %GOF macro as to what datasets contain the model parameter estimates including the covariance parameter estimates (except in those cases when one uses GENMOD - see the detailed syntax notes below for details). The DATA= argument identifies the SAS dataset containing the observed and predicted response from the model and is a required argu-

ment. The FITSTATS= argument specifies the SAS dataset containing the final fit statistics as generated under the SAS procedure. This argument is strictly optional and one may choose to ignore it. Finally, the required RESPONSE= argument specifies the name of the response variable while the PRED_IND= argument and PRED_AVG= argument specify, respectively, each individual's predicted mean response and the average individual's predicted mean response. The former is a function of both the fixed-effects and random-effects parameters (as determined when one has a random effects model) as they relate to the subject-specific mean response. The latter is a function of only the fixed-effects and represents what the mean response is when the random effects are set equal to 0 (if random effects are in fact present). For marginal models such as the GNLM, the value of PRED_IND= and PRED_AVG= should be one in the same.

Finally, the OMEGA= and OMEGA_R= arguments are optional arguments that specify datasets containing, respectively, the model-based and empirical-based asymptotic covariance matrices of the regression parameter estimates. When specified, these two arguments instruct the %GOF macro to compute the pseudo-likelihood ratio test of Vonesh et. al. (1996) in order that one might test the adequacy of the assumed model-based covariance structure. For analyses that require one to use the MIXED or NLMIXED procedures, one will need to run the analyses twice in order to output these two datasets. Also, when running NLMIXED, one will need to specify starting values for *all* model parameters starting with the regression parameters first followed by the variance-covariance parameters of the assumed covariance structure. This is necessary in order for the macro to compute the pseudo-likelihood ratio test based strictly on the specifications of the assumed mean structure (see the detailed syntax comments for the GOF macro for a further explanation).

Below is the basic syntax of the program as made available through the SAS Press Companion Web site at support.sas.com/companionsites. Included is a description of other optional arguments like the TITLE= and REF= arguments not discussed above but which are described in detail below. Applications illustrating its use may be found in Examples 4.5.1 and 4.5.2.

Program

```
/*===== MACRO GOF ======
```

Author: Edward F. Vonesh, PhD
Written: May 14, 2004
Modified: Sep 18, 2008
Modified: May 18, 2010
Program: This macro program performs the following:

- 1) This macro calculates R-square (R) and concordance correlation (Rc) Goodness-of-Fit (GOF) statistics for the SAS procedures: MIXED, GLIMMIX, NLMIXED and GENMOD and is based on the paper by Vonesh, Chinchilli and Pu (Biometrics, 52:572-587, 1996).
- 2) This macro was originally written on May 14, 2004 to include unadjusted and adjusted R-square and concordance correlations (Rc).
- 3) The program was modified on Sep 18, 2008 to allow R-square and concordance correlation coefficients from NLMIXED and GENMOD output. It also contains model fit statistics from the procedures.
- 4) The program was modified on May 18, 2010 to allow the use of the pseudo-likelihood ratio test of Vonesh et. al. (1996) to test whether the model based covariance matrix of the parameter estimates is equal to the robust covariance matrix of the parameter estimates. The modifications in this version of macro GOF yield essentially the same results as the macro GLIMMIX_GOF but with the following EXCEPTIONS:
 - GLIMMIX_GOF bases the variance-covariance concordance correlation coefficient (CCC) on a slight modification of the variance-covariance CCC by Vonesh (see the GLIMMIX documentation for details) whereas macro GOF is based on the original publication of Vonesh, Chinchilli and Pu (Biometrics 52:572-587, 1996).
 - GLIMMIX_GOF bases adjusted R-Square and adjusted Rc (CCC) values on the number of regression

parameters whereas GOF bases the adjusted R-Square and Rc values on the total number of parameters.

NOTES 1. For the SAS procedure GENMOD, adjusted R-square and adjusted concordance correlation coefficients (CCC) are computed based ONLY on the number of fixed-effects regression parameters in the model as the estimated variance-covariance matrix for the repeated measures setting are method-of-moment estimates that are not tied to a likelihood-based objective function. For the SAS procedures MIXED, GLIMMIX and NLMIXED, it is based on all of the estimated parameters, both the fixed-effects regression parameters and the variance-covariance parameters provided the parameter that is estimated has an associated degrees of freedom. For example, if one specifies the following GLIMMIX statement

```
RANDOM _Residual_;
```

then SAS will output all of the fixed-effects regression parameters plus an overdispersion parameter that has 0 degrees of freedom. This macro GOF will NOT include this overdispersion parameter when it computes the adjusted R-square or adjusted concordance correlation coefficient based on the number of parameters used to compute the predicted values.

2. It should also be noted that the dataset containing predicted values must be created from the appropriate SAS procedure in accordance with the goals for running macro GOF. For example, if one is running a model with random effects using MIXED, GLIMMIX or NLMIXED and one would like to see both population-average R-square and subject-specific

R-square values (Vonesh and Chinchilli, 1997), then one should create an appropriate dataset containing both the predicted values for the typical subject (random effects set to 0) and the predicted values for each subject (random effects set to the BLUP values). For example, in MIXED one would specify both the OUTP= and OUTPM= options of the MODEL statement specifying PA and SS predicted values. In GLIMMIX, one would use the OUTPUT OUT statement with specifications pred(noblup ilink)=’name’ where ‘name’ is the name of the predicted value when the random effects are set to 0 and pred(blup ilink)=’name’ where ‘name’ is the name of the predicted value when random effects are set to their EBLUP values. Likewise one will need to specify both kinds of predicted values in NLMIXED using the SAS statement PREDICT expression OUT=’name of dataset’.

3. Note that the predicted values correspond the conditional means of the response variable. For a mixed-effects model, let the conditional mean given a random effect, b_i , be denoted by $E(Y_{ij}|b_i)=f(\beta, b_i)$ where $f(\beta, b_i)$ is a possibly nonlinear conditional mean of response Y_{ij} for i th subject at j th occasion. The typical subject mean response is $E(Y_{ij}|b_i=0)=f(\beta, 0)$ and the subject-specific (SS) mean response is $E(Y_{ij}|b_i)=f(\beta, b_i)$ where β is the vector of fixed-effects regression parameters. If the model is already a marginal model such that $E(Y_{ij})=f(\beta)$, then the macro GOF will only report goodness-of-fit on marginal or PA-based predicted values, $E(Y_{ij})=f(\hat{\beta})$ whereas if the model is a mixed-effects model, it will report goodness-of-fit on both the typical mean response, $E(Y_{ij}|b_i=0)=f(\hat{\beta}, 0)$, denoted PRED_AVG within macro GOF and the SS mean response, $E(Y_{ij}|b_i)=f(\hat{\beta}, b_i)$ denoted by PRED_IND within GOF PROVIDED one specifies values of both PRED_AVG and PRED_IND. If one only specifies PRED_AVG (the minimum default), then macro GOF will

report the same PA and SS R-square and Rc values based on PRED_AVG.

Macro Key:

- WHERE = Where clause defining subgroups within which to perform GOF
- PROC = Defines which procedure was used to create predicted and observed response values.
Valid values are:
 - PROC = MIXED
 - PROC = GLIMMIX
 - PROC = NLMIXED
 - PROC = GENMOD
- PARMS = SAS dataset 'name' containing fixed-effects regression parameters. For the different SAS procedures, these correspond to the following ODS OUTPUT statements:
 - MIXED: ODS OUTPUT SolutionF='name'
 - GLIMMIX: ODS OUTPUT ParameterEstimates='name'
 - NLMIXED: ODS OUTPUT ParameterEstimates='name'
 - GENMOD: ODS OUTPUT ParameterEstimates='name' if no REPEATED statement is used.
This is useful for univariate applications.
 - GENMOD: ODS OUTPUT GEEEmpPEst='name' if a REPEATED statement is used
- IMPORTANT NOTE:**
For NLMIXED, one should create two datasets based on the dataset created by the ODS OUTPUT ParameterEstimates='name'. The first dataset should contain ONLY the fixed-effects regression parameters and this is the dataset one should specify with the PARMS= argument.
- COVPARMS = SAS dataset containing unique variance-covariance parameters from MIXED or GLIMMIX and NLMIXED (see below for NLMIXED explanation).
The dataset corresponds to:
 - MIXED: ODS OUTPUT CovParms='name'

is established by default regardless of whether one specified a RANDOM or REPEATED statement

- GLIMMIX: ODS OUTPUT CovParms='name' if a RANDOM statement is used
- NLMIXED: ODS OUTPUT ParameterEstimates='name'
(See IMPORTANT NOTE below)

IMPORTANT NOTE:
For NLMIXED, one should create two datasets based on the dataset created by the ODS OUTPUT ParameterEstimates='name'. The second dataset should contain ONLY the variance-covariance parameters and this is the dataset one should specify with the COVPARMS= argument.

DATA = SAS dataset containing predicted values. For the different SAS procedures, this dataset would be constructed based on the following statements:
- MIXED: MODEL /OUTP='name1' OUTPM='name2'
- GLIMMIX: OUTPUT OUT='name' pred(options)=
- NLMIXED: PREDICT expression OUT='name'

RESPONSE = Primary response variable

SUBJECT = Identifies the SAS variable representing the unique SUBJECT ID. This must be specified if one wishes to test whether the robust covariance matrix and the model-based covariance matrix are equal. This variable must be included in the DATA= dataset above.

PRED_AVG = Predicted response for the average individual based on the PA predicted response, $f(\beta)$

PRED_IND = Predicted response for each individual based on the SS predicted response, $f(\beta, \beta_i)$

FITSTATS = SAS dataset containing fit statistics from the given SAS procedure. This is OPTIONAL within the macro GOF and is designed to let one print the fit statistics that SAS outputs by default. In MIXED, GLIMMIX and NLMIXED, this dataset corresponds to the statement:
- ODS OUTPUT FitStatistics='name'.

Note that for GENMOD, if one runs a model without a REPEATED statement, the dataset corresponds to

- ODS OUTPUT ModelFit = 'name'

but if one runs GENMOD with a REPEATED statement, then such a dataset should be given by the following

- ODS OUTPUT GEEFitCriteria = 'name'

OMEGA = SAS dataset containing the model-based covariance matrix of the model parameter estimates. This is used for the pseudo-likelihood ratio test of Vonesh et. al. (Biometrics, 1996). In MIXED, GLIMMIX, NLMIXED and GENMOD, the dataset corresponds to the statement:

- MIXED: ODS OUTPUT CovB='name'
- must specify CovB option in MODEL
- GLIMMIX: ODS OUTPUT CovB='name'
- must specify CovB option in MODEL
- NLMIXED: ODS OUTPUT CovMatParmEst='name'
- must specify COV option in NLMIXED
- GENMOD: ODS OUTPUT GEENcov='name'
- must specifiy MCOVB and ECOVB
options in the REPEATED statement

IMPORTANT NOTE:

For NLMIXED, one should make sure that the regression parameters are listed first within the PARMS= statement (one should ALWAYS use a PARMS= statement if one wishes to perform the pseudo-likelihood ratio test) so that the macro will correctly identify the correct elements of the overall covariance matrix associated with ONLY the regression parameter estimates and NOT any of the model variance-covariance parameter estimates.

OMEGA_R = SAS dataset containing the robust covariance matrix of the model parameter estimates. This is used for the pseudo-likelihood ratio test of Vonesh et. al. (Biometrics, 1996). In MIXED, GLIMMIX, NLMIXED and GENMOD, the dataset

corresponds to the statement:

- MIXED: ODS OUTPUT CovB='name'
 - must specify CovB option in MODEL and the EMPIRICAL option of the MIXED statement
- GLIMMIX: ODS OUTPUT CovB='name'
 - must specify CovB option in MODEL and the EMPIRICAL option of the GLIMMIX statement
- NLMIXED: ODS OUTPUT CovMatParmEst='name'
 - must specify COV option in NLMIXED and the EMPIRICAL option of the NLMIXED statement
- GENMOD: ODS OUTPUT GEERcov='name'
 - must specifiy MCovB and ECovB options in the REPEATED statement

IMPORTANT NOTE:

For NLMIXED, one should make sure that the regression parameters are listed first within the PARMS= statement (one should ALWAYS use a PARMS= statement if one wishes to perform the pseudo-likelihood ratio test) so that the macro will correctly identify the correct elements of the overall covariance matrix associated with ONLY the regression parameter estimates and NOT any of the model variance-covariance parameter estimates.

TITLE	= Additional title information to be included in the output may be specified here Example: TITLE = A Poisson mixed-effects model;
REF	= Specifies that appropriate references to methods be printed. The default (REF=,) does not print references while specifying REF=yes will print references.
OPT	= option to print or not print the results of the SUMMARY statement used in IML. This option will produce summary statistics on the response variable, the PA and SS predicted responses and

```

the corresponding PA and SS residuals (Y-YHAT)
OPT=NOPRINT (no output is printed - default)
OPT=PRINT (output is printed)
PRINTOPT = option to print or not print the results of the
macro. The available options are
PRINTOPT=NOPRINT (no output is printed)
PRINTOPT=PRINT (output is printed - default)

Regardless of which option one selects, the
results are stored in the dataset, _fitting,
which contains a description of the statistic
and its value which one can also print out.
=====
%macro GOF(where,
            proc=,
            parms=,
            covparms=,
            data=,
            subject=,
            fitstats=,
            omega=,
            omega_r=,
            response=,
            pred_ind=,
            pred_avg=,
            title=,
            ref=,
            opt=noprint,
            printopt=print);

```

D.2 The GLIMMIX_GOF Macro

The %GLIMMIX_GOF macro computes the same summary goodness-of-fit measures and pseudo-likelihood ratio test as the %GOF macro but the %GLIMMIX_GOF macro was written for exclusive use with the GLIMMIX procedure. This was done, in part, to take advantage of some of the features of GLIMMIX not available with any of the other procedures. For example,

by specifying the COVB(DETAILS) option of MODEL statement, one can compute the pseudo-likelihood ratio test of the assumed covariance structure without having to output the model-based and empirical-based estimates of the asymptotic covariance matrix of the regression parameter estimates as the COVB(DETAILS) option automatically computes the discrepancy function on which this test is based.

The results one obtains with the %GLIMMIX_GOF macro will differ slightly from results one would obtain using the %GOF macro. For example, the %GLIMMIX_GOF macro does not require one to input the parameter estimates corresponding to the assumed covariance structure. Consequently, the adjusted R-square values and adjusted concordance correlation coefficients (CCC) will differ from that of %GOF in that they are adjusted only for the number of fixed-effects regression parameters in the model whereas the %GOF macro adjusts for all the parameters (except when using the GENMOD procedure to fit the model). Likewise the variance-covariance concordance correlation used to assess how close the model-based asymptotic covariance matrix is to the empirical or robust asymptotic covariance matrix will differ between the two macros for reasons described in the on-line SAS Help and Documentation for GLIMMIX (see the section entitled “Exploring and Comparing Covariance Matrices”).

The macro arguments that one must supply to the macro as well as the basic syntax used to call the macro are summarized below. The macro arguments are similar to those used in the %GOF macro except for the argument COVB_GOF. This argument is needed to test the model-based covariance structure using the pseudo-likelihood ratio test described by Vonesh et. al. (1996) and Vonesh and Chinchilli (1997). The COVB_GOF= argument specifies the dataset containing the discrepancy function and other GOF measures generated under the COVB(DETAILS) option. The dataset is created by the ODS OUTPUT CovBDetails='name' statement. The following is the macro syntax and detailed explanation of the macro arguments. Applications illustrating its use may be found in Examples 2.2.2 and 5.2.1.

Program

```
/*===== MACRO GLIMMIX_GOF ======
 Author:      Edward F. Vonesh, PhD
 Written:     August 11, 2008
 Modified:    August 10, 2009
```

Program: This macro program performs the following:

- 1) This macro calculates R-square (R) and concordance correlation (Rc) Goodness-of-Fit (GOF) statistics for the SAS procedure GLIMMIX and is based on the paper by Vonesh, Chinchilli and Pu (Biometrics, 52:572-587, 1996) as well as the book by Vonesh and Chinchilli (1997).
- 2) This macro also compares the robust sandwich estimator of Var(beta) to a model-based estimator of Var(beta) using the pseudo-likelihood ratio test of Vonesh et al (1996). It should be noted that this macro is intended ONLY for applications that involve a single set of random effects parameters and has not been tested for applications involving nested random effects such as the multi-level random-effects models of Goldstein et al.
- 3) The program was modified on Aug 10, 2009 to allow a modified form of the pseudo-likelihood ratio test of Vonesh et al (1996) to test whether the model-based covariance matrix of the regression parameter estimates is equal to the empirical or robust covariance matrix of those estimates.
- 4) The modifications in this version of the macro yield essentially the same results as the macro GOF but with the following EXCEPTIONS:
 - GLIMMIX_GOF bases the variance-covariance concordance correlation coefficient (CCC) on a slight modification of the variance-covariance CCC by Vonesh (see the GLIMMIX documentation for details) whereas macro GOF is based on the original publication of Vonesh, Chinchilli and Pu (Biometrics 52:572-587, 1996).
 - GLIMMIX_GOF bases adjusted R-Square and adjusted Rc (CCC) values on the number of regression parameters whereas GOF bases the adjusted R-Square and Rc values on the total number of parameters.

=====

=====

NOTES. When running %GLIMMIX_GOF one must use ODS OUTPUT statements to create the necessary datasets which this macro uses to construct concordance correlations and to carry out the necessary pseudo-likelihood ratio test of the assumed variance-covariance structure based on the discrepancy function output by GLIMMIX and described by Vonesh, Chinchilli and Pu (1996) and Vonesh and Chinchilli (1997). The macro arguments used in this macro are defined as follows:

=====

Macro Key:

WHERE = Where clause defining subgroups within which one wishes to perform a GOF. Typically, these subgroups are defined by a BY statement used with the GLIMMIX procedure.

DIMENSION= SAS dataset containing the dimensions of the model run using GLIMMIX. This is the dataset created with ODS OUTPUT Dimensions='name'

PARMS = SAS dataset containing fixed-effects parameters. This is the dataset created with ODS OUTPUT ParameterEstimates='name' statement.

COVB_GOF = SAS dataset containing GOF output when one specifies the COVB(DETAILS) option of MODEL statement. This is the dataset created with ODS OUTPUT CovBDetails='name'

OUTPUT = SAS dataset containing data from OUTPUT OUT= statement from GLIMMIX. When running GLIMMIX with assumed random effects specified by the RANDOM statement, the OUTPUT OUT= statement should contain two sets of predicted values, one for the population-average mean using
PRED(NOBLUP ILINK)='name1'
and one for the subject-specific mean using
PRED(ILINK)='name2'.

One can easily accomplish these tasks using the option ALLSTATS with the OUTPUT OUT= statement as this will produce all the required statistics

RESPONSE = Primary response variable (this is the name of the response variable used in the GLIMMIX MODEL statement).

PRED_AVG = Predicted mean response for typical or average individual based on the inverse link function and the fixed-effects parameters only. If one simply specifies

```
OUTPUT OUT=name / ALLSTATS;
then PRED_AVG=PredMuPA.
```

PRED_IND = Predicted mean response for each individual based on the inverse link function and the fixed-effects and random-effects estimates for each individual. If one simply specifies

```
OUTPUT OUT=name / ALLSTATS;
then PRED_IND=PredMu.
```

OPT = option to print or not print the results of the SUMMARY statement used in IML. This option will produce summary statistics on the response variable, the PA and SS predicted responses and the corresponding PA and SS residuals (Y-YHAT)

- OPT=NOPRINT (no output is printed - default)
- OPT=PRINT (output is printed)

PRINTOPT = option to print or not print the results of the macro. The available options are

- PRINTOPT=NOPRINT (no output is printed)
- PRINTOPT=PRINT (output is printed - default)

Regardless of which option one selects, the results are stored in the dataset, _fitting, which contains a description of the statistic and its value which one can also print out.

```
%macro GLIMMIX_GOF(where,
dimension=,
parms=,
covb_gof=,
```

```

output=,
response=,
pred_ind=,
pred_avg=,
opt=noprint,
printopt=print);

```

D.3 The CCC Macro

The %CCC macro was used in the analysis of the MCM2 biomarker data (Example 2.2.4). The macro was written specifically for applications in which the goal is to assess the level of agreement between two or more assays, two or more raters or, in general, two or more measurement sources for a particular response variable taken on the same subject or experimental unit. The basic setup involves p sets of measurements of the same response variable taken from the same subject or experimental unit from p different sources. The goal is to estimate the degree of concordance between measurements taken from the different sources. As described in §2.2.4, this can be done either on a pairwise basis or on an overall basis.

The macro %CCC computes a concordance correlation matrix consisting of pairwise CCC's between each of the $p(p - 1)/2$ pairs of measurements assuming some specified variance-covariance structure. It also computes an overall concordance correlation coefficient (OCCC) based on the work of Lin (1989, 2000), Barnhart et. al. (2002) and Carrasco and Jover (2003). The overall CCC or OCCC is used to assess the overall agreement between all p sets of measurements ($p \geq 2$). Following Helenowski et. al. (2011), estimation of a CCC matrix and the OCCC is based on fitting the marginal linear repeated measures ANOVA model

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

to a set of p measurements from p different sources taken from the same individual. Under this model, y_{ij} is a measurement taken from the i^{th} individual ($i = 1, \dots, n$) corresponding to the j^{th} measurement source ($j = 1, \dots, p$), μ is an overall mean, τ_j is the effect due to the j^{th} source (e.g., different raters or instruments or, in the MCM2 biomarker example, different prostate gland sections), and the ϵ_{ij} are random error terms with $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{ip})'$ assumed to be normally distributed as $N_p(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$ for some specified variance-

covariance structure $\Sigma(\boldsymbol{\theta})$. One can include covariate adjustments by simply extending the repeated measures ANOVA model to the repeated measures ANCOVA model

$$y_{ij} = \mu + \tau_j + \sum_{k=1}^s \beta_k x_{ik} + \epsilon_{ij}$$

where $\{x_{ik}; k = 1, \dots, s\}$ are a set of select covariates one wishes to adjust for. One can then estimate the unadjusted (the ANOVA model) or adjusted (the ANCOVA model) means, $\mu_j = \mu + \tau_j$, the variances, σ_j^2 , and correlations, $\rho_{jj'}$, between any two sources j and j' using PROC MIXED. Using ODS OUTPUT statements to output key parameter estimates, one can then estimate a CCC matrix and OCCC under an assumed covariance structure using the SAS macro %CCC in conjunction with the formulas of Carrasco and Jover (2003) and Helenowski et. al. (2011). The macro also includes an estimate of the assumed correlation matrix based on one's specification of $\Sigma(\boldsymbol{\theta})$. Below is a description of the macro programming syntax and detailed explanation of the macro arguments required to run the %CCC macro.

Program

```
/*===== MACRO CCC =====
Author:      Edward F. Vonesh, PhD
Written:     August 22, 2008
Modified:
Program:    This macro program performs the following:
1) Macro CCC calculates an overall correlation matrix
and a concordance correlation matrix of pairwise
concordance correlation coefficients (CCC) based
on a specified marginal linear model with a
specified marginal variance-covariance structure.
The macro also computes an overall concordance
correlation coefficient (OCCC). These summary
measures of agreement are based on papers by
Carrasco and Jover (Biometrics 59:849-858, 2003),
Barnhart, Haber and Song (Biometrics, 58:1020-
1027, 2002), Barnhart and Williamson (Biometrics
57:931-940, 2001) and Helenowski et. al. (The
Int. Journal of Biostatistics, 7:1-21, 2011).
```

NOTES 1. Notation: CCC = Concordance Correlation Coefficient
OCCC = Overall CCC

2. This macro assumes one has run PROC MIXED using a marginal linear model with an intercept, a fixed-effects variable representing say p sources over which one wishes to compute pairwise CCC's and the overall CCC (OCCC). The marginal linear model can also include additional covariates one wishes to adjust for (e.g., Carrasco and Jover, 2003; Helenowski et. al., 2011).

Macro Key:

WHERE -a where clause that defines a where statement for exclusion purposes

COV -defines the dataset from PROC MIXED that contains the marginal covariance matrix. When using the REPEATED statement, this data set is created by specifying the ods statement:
ODS OUTPUT R='name';
where 'name' is the dataset. For example, if one specifies
ODS OUTPUT R=Rvar;
then one would specify
COV=Rvar

MEAN -defines the dataset from PROC MIXED that contains the fixed-effects regression parameter estimates which one obtains using the ods statement:
ODS OUTPUT ParameterEstimates='name'

EFFECT -defines the effect variable from the dataset defined by the above macro argument MEAN. The EFFECT variable should be that variable in the MODEL statement of PROC MIXED that defines what the adjusted or unadjusted effects of the p sources of interest have on the response variable of interest. These p sources are the levels that

one wishes to evaluate how reproducible each of the sources are relative to one another (i.e., what are the pairwise CCC's between the different sources and what is the OCCC across sources).

For example, suppose the MODEL statement of PROC MIXED is the following:

```
MODEL Y = GROUP X /SOLUTION;
```

where X is some continuous covariate one wishes to adjust for and GROUP is a variable that represents the p different sources of interest.

Then one would specify

```
EFFECT=GROUP
```

as the effect variable of interest.

N -defines the dataset from PROC MIXED that contains the number of subjects (experimental units) in the data. This dataset corresponds to the dataset created by the ods statement

```
ODS OUTPUT Dimensions='name'
```

in which case one would specify

```
N='name'
```

LABEL -defines a label for the single response variable which is being measured across the p sources (i.e., p different assay methods, p different raters, etc.).

STRUCTURE -defines a label for the type of variance-covariance structure specified in MIXED.

AIC -defines the dataset from PROC MIXED that summarizes the model goodness-of-fit based on AIC, etc.

TABLE -defines a Table name or number or description of the output from this macro (default is TABLE=1).

STATFMT -define the numeric format for summary measures like N, mean, and standard deviation
The default is STATFMT=6.2

CORRFMT -define the numeric format for all measures of correlation (pairwise CCC's and OCCC).
The default is CORRFMT=6.4

```
=====
%MACRO CCC(where,
            cov=,
            mean=,
            effect=,
            n=,
            label=,
            structure=,
            AIC=,
            table=1,
            statfmt=6.2,
            corrfmt=6.4);
```

D.4 The CONCORR Macro

The %CONCORR macro was also used in the analysis of the MCM2 biomarker data (Example 2.2.4) and like the %CCC macro, the %CONCORR macro was written specifically for applications in which the goal is to assess the level of agreement between two or more assays, two or more raters or, in general, two or more measurement sources for a particular response variable taken on the same subject or experimental unit. However, unlike the %CCC macro which is model-based, the %CONCORR macro simply computes estimates of the CCC for each specified pair of measurements based on method-of-moment estimates of the means, standard deviations and correlation restricted to complete data on each pair of measurements. This is useful when assessing the level of agreement between several different measurement sources in the presence of missing data or when the number of sources exceeds the number of individuals making it impossible to use the %CCC macro with an unstructured variance-covariance structure as illustrated in Example 2.2.4. Below is a description of the macro programming syntax and detailed explanation of the macro arguments required to run the %CONCORR macro.

Program

```
===== MACRO CONCORR =====
Author:      Edward F. Vonesh, PhD
```

Written: July 13, 1992

Modified: April 26, 2010

Program: This macro program performs the following:

- 1) Macro CONCORR computes the concordance correlation coefficients (CCC) for a group of variables. It is based on the CCC as defined by Lin in a series of papers (Biometrics, 1989, 1992).
- 2) The macro CONCORR was modified on April 26, 2010 to allow one to compute pairwise estimates of the CCC when faced with unbalanced data. The macro computes summary statistics (N, MEAN, STD) and both the Pearson correlation (R) and the Lin concordance correlation (Rc) for pairwise variables defined by a VAR statement. To use this newer version of CONCORR, we assume that the concordance correlations are needed on pairwise variables which are listed side by side in the VAR statement. Thus specifying the following VAR statement

VAR Y1 X1 Y2 X2 Y3 X3 ...

will cause CONCORR to print the summary statistics N MEAN STD and correlation coefficients R and Rc for Y1 with X1 and Y2 with X2 and Y3 with X3 etc.

NOTES 1. As a simple example, one can run the following code to illustrate how the CONCORR macro works.

```
DATA TEST;
  DO I=1 TO 20;
    X=NORMAL(945723848);
    Y=.15+1.05*X + NORMAL(45678123);
    OUTPUT;
  END;
%CONCORR(BY=,LASTBY=, DATA=TEST, VAR=Y X, FORMAT=12.2,);
```

Macro Key:

WHERE = a where clause that defines a where statement for

exclusion purposes

BY = Defines a SAS BY statement that will compute the pairwise CCC's by the levels of the different groups specified in the BY argument

LASTBY = Defines the value of the last variable listed in the above BY statement. This is needed in order for the macro to sort and print results from IML

NOTE: If one specifies one or more BY variables then one MUST specify the LASTBY variable or risk running into errors

BYLABEL = Lists labels for the BY variables. The following example illustrates its use:

```
BY = ModelType Typical,
BYLABEL={'Model', 'Typical'},
```

In this example we label ModelType by 'Model' and Typical by 'Typical'. Note that the value of BYLABEL should be in the matrix character format in order to be directly implemented within the program

BYFORMAT = Lists formats for the BY variables
The default is BYFORMAT=\$char8.

DATA = Defines the SAS dataset containing the variables

VAR = Lists the variables one wishes to compute the concordance correlations for. It assumes this list results in an even number of variables with concordance correlations desired on each pair of variables listed in the VAR. Thus if we specify the VAR statement as

```
VAR=Y1 X1 Y2 X2 Y3 X3,
```

then we will get a column-wise printing of correlations between Y1 and X1 and between Y2 and X2 and between Y3 and X3, etc. This simplifies the output and presents everything in a one table format.

VARLABEL = Lists labels for the various variables listed in the VAR statement using same structure as used for the BYLABEL argument. For example,

```
VAR = y_obs y_pred,
```

```

        VARLABEL = {'Observed Y' 'Predicted Y'}
could be used to label the variables y_obs and
y_pred
STATFMT = Define the numeric format for summary measures
          like N, mean, and standard deviation
          The default is STATFMT=6.2
FORMAT   = Define the numeric format for the pairwise
          correlations (CCC's and Pearson correlation)
          The default is FORMAT=4.2
LABEL1   = Defines a label for the odd numbered variables
          in the VAR statement
          The default is LABEL1=Variables
LABEL2   = Defines a label for the even numbered variables
          in the VAR statement
          The default is LABEL2=With
=====
Important Note: All BY variables must be character so one MUST
                convert these to character variables within
                the dataset prior to running CONCORR.
=====
=====*/
%MACRO CONCORR(where,
               by=dum82152,
               lastby=dum82152,
               bylabel=,
               byformat=$char8.,
               data=_LAST_,
               var=,
               varlabel=,
               label1=Variables,
               label2=With,
               statfmt=6.2,
               format=4.2);

```

D.5 The COVPARMS Macro

The %COVPARMS macro was written exclusively for use with the NLMIXED procedure and is designed to provide users with reasonable starting values for both the fixed-effects parameters and the variance-covariance parameters of a nonlinear or generalized nonlinear mixed-effects model. Briefly, the macro provides users with a SAS dataset containing OLS estimates of the regression parameters and method-of-moment type estimates for the random-effect variance-covariance parameters obtained by fitting a pseudo linear mixed-effects model to the residuals of an OLS fit to the data. For a discussion on how one can use this macro to help decide on an appropriate random-effects structure for a given GNLME model, the reader is referred to §5.3.6 as well as Examples 5.4.1, 5.4.2 and 7.2.1. Here we present the macro syntax required to run the macro as well as a detailed explanation of the macro arguments and required input needed to run the macro.

Program

```
/*===== MACRO COVPARMS ======
Author:      Edward F. Vonesh, PhD
Written:    September 20, 2009
Modified:
Program:   This macro program performs the following:
1) Macro COVPARMS creates a SAS dataset that contains
parameter starting values for nonlinear mixed-
effects models with random effects. This macro
appends OLS-based parameter estimates from NLMIXED
when one has random effects with all variance-
covariance parameters set to 0 (see example below)
and one creates a ParameterEstimates dataset
(from ODS Output) that contains all fixed effects
parameters and within-subject variance parameters.
2) Specifically, it takes an output dataset from the
PREDICT statement in NLMIXED that contains the
derivatives of the random effects evaluated at
the final OLS estimates from NLMIXED and obtains
estimated variances and covariances of the random
effects and amends these estimates to those of
```

the fixed-effects in a final dataset defined by OUTPUT=.

NOTES 1.To successfully run this macro one MUST

1. Specify a PREDICT statement in an initial call to NLMIXED that will contain the estimated residuals from a least squares (LS) fit to the data in which random effects, although specified initially, are ignored because their variances and covariances are all set equal to 0. This PREDICT statement must include the option DER when specified.
2. Specify an ODS OUTPUT statement containing the initial NLMIXED results one obtains when essentially running a model with no random-effects despite the requirement that one actually does specify the random effects that one wishes to incorporate into the model (see example below). This statement is of the form
ODS OUTPUT ParameterEstimates='name'.
- 2.OUTPUT: The program will print out the initial estimated random-effects variance-covariance matrix as well as initial estimates of the correlations of the random effects (useful for deciding whether to include or exclude one or more random effects in the final model). It will also print the dataset containing all initial starting values of the parameters one is attempting to fit with NLMIXED. One can edit and delete parameters from the dataset to increase the speed at which NLMIXED converges based on reasonable starting values.
- 3.DETAILS: This macro computes approximate starting values for the variance-covariance matrix of the random effects for nonlinear mixed-effects models in NLMIXED. Using input from an initial call to NLMIXED

the macro calls GLIMMIX to compute the variance-covariance parameters of a linear random effects model using a G-side variance-covariance structure specified by the METHOD= option of the macro. One must run NLMIXED with the desired random effects specified in the model but with all values within the RANDOM statement set to 0 so that NLMIXED is performing an OLS analysis. One must also specify the PREDICT= statement in NLMIXED such that the output dataset contains the residual values from the nonlinear function (i.e., resid=response-pred where pred is the predicted response from each observation and response is the response variable specified in the MODEL statement of NLMIXED).

Macro Key:

- PARMS = Defines a SAS dataset defined by
 ODS OUTPUT ParameterEstimates='name'
 that contains OLS parameter estimates from
 NLMIXED when all random effect variance-
 covariance parameters are set equal to 0
- PREDOUT = Defines a SAS dataset containing residual values
 from an initial call to NLMIXED that are defined
 by an ID statement in combination with the
 following general PREDICT statement
 PREDICT 'expression' OUT='dataset name'
 from the initial call to NLMIXED.
- RESID = Defines the name of the variable that is the
 residual ($y - \hat{y}$) from NLMIXED. It must be
 defined within the initial call to NLMIXED and
 placed in an ID statement so that it will be
 included in the dataset
- METHOD = Defines what method one wishes to use to estimate
 the random effect variance-covariance parameters
 to be estimated in NLMIXED. The default is
 METHOD=MSPL which is ML for the pseudo linear
 random effects model.

BOUND = Defines whether the NOBOUND option of GLIMMIX is to be invoked or not. If one leaves BOUND= unspecified, then variance components have a default lower boundary constraint of 0. Else if BOUND=NOBOUND, then this boundary constraint is lifted. The only valid value is BOUND=NOBOUND and the default is BOUND=

NOTE: For small datasets, it is advised to use the default BOUND= option as otherwise the macro may issue the following warning and not produce the desired output.

WARNING: Optimization stopped because of infinite objective function.

WARNING: Output 'g' was not created. Make sure that the output object name, label or path is spelled correctly. Also, verify that the appropriate procedure options are used to produce the requested output object. For example, verify that the NOPRINT option is not used.

RANDOM = Defines the variables having prefix 'Der_' created by NLMIXED corresponding to the dataset specified by the DATA= statement above. These are the random effect derivatives to be used in PROC MIXED as a means to compute the covariance parameters from a first-order approximation.

SUBJECT = Defines a SUBJECT variable needed in MIXED. This should be the same SUBJECT variable used in the initial call to NLMIXED

TYPE = Defines the type of random-effects variance-covariance matrix (typically UN or VC) assumed

COVNAME = Defines a prefix name to be used for the character variable PARAMETER that exists within the dataset defined by PARMs above. This prefix is then used in the creation of values for PARAMETER that specify starting values for the random-effects variance-covariance parameters.

```

OUTPUT    = Defines a SAS dataset, say 'outname' that is to
            contain all of the parameters to be used in a
            call to NLMIXED using the PARMs / data='outname'
            statement. This dataset will include the OLS
            regression parameter estimates in the dataset
            specified by PARMs= statement above as well as
            the variance-covariance parameter estimates from
            this macro having the prefix defined by the
            COVNAME= option above.

=====
%macro covparms(parms=,
                predout=,
                resid=resid,
                method=mspl,
                bound=,
                random=der_u1 der_u2 der_u3,
                subject=subject,
                type=un,
                covname=psi,
                output=parms);

```

D.6 The VECH Macro

The %VECH macro was written exclusively for use with the NLMIXED procedure. It was developed for use in applications where one wishes to fit a GNLME model to a set of data assuming an intrasubject covariance structure other than the default independence structure that NLMIXED assumes (see §5.3.3 for an approach one can take to overcome this limitation of NLMIXED). The macro is designed to provide users with an easy to use naming convention for the variance-covariance parameters of an assumed intra-subject covariance matrix. Examples illustrating its use can be found in Examples 5.4.1 and 5.4.3. Below is a brief description of the macro syntax required to run the macro as well as an explanation of the macro arguments and a complete listing of the macro program.

Program

```
/*===== MACRO VECH =====
Author:      Edward F. Vonesh, PhD
Written:     March 4, 2011
Modified:
Program:    This macro program performs the following:
1) Macro VECH creates a vech representation of the
   intra-cluster covariance matrix defined in the
   NLMIXED code for any particular application that
   requires an intra-cluster or within-subject
   covariance structure (see Ch. 5, Example 5.4.1)
=====
=====

NOTES 1. This macro was written for applications involving
nonlinear mixed-effects models in which one wishes
to include an intra-subject covariance matrix using
the NLMIXED procedure. The sole purpose of this
macro is to provide a naming convention to the
variance-covariance parameters of the intra-subject
covariance matrix so that one can easily specify
this covariance structure using the RANDOM statement
in NLMIXED. Examples of its use are included at the
end of the macro code in this file as well as in
Chapter 5 (Examples 5.4.1, and 5.4.3).
=====

Macro Key:
DIM      = Defines the dimension of the within-subject
            variance-covariance matrix that one wishes
            to include within a RANDOM statement of NLMIXED
COV     = Defines the array name within the NLMIXED program
            that is used to define the within-subject
            variance-covariance matrix. For example, if
            one were to define a 7 x 7 covariance matrix
            using the ARRAY statement
```

```

        ARRAY VARCOV[7,7];
which one then assigns values to using DO
statements from within a call to NLMIXED, then
one would specify
        COV=VARCOV
The default value is
        COV=cov
and there is very little reason to ever change
this default name.

VECHCOV = Defines the name of the VECH vector created by
          the macro containing the unique lower diagonal
          elements of the within-subject covariance matrix
          that is defined by the COV array name above.
The default value is
        VECHCOV=vechcov
and there is very little reason to ever change
this default name.

NAME     = Defines a prefix name for each of the lower
          diagonal components of the variance-covariance
          matrix, or equivalently, the vector components
          of the VECH vector defined by VECHCOV. For
          example, if one wishes all of the lower diagonal
          elements to start with the letters COV and the
          dimension of the original covariance matrix is
          say 5 x 5 then one would specify
        NAME=COV
and the VECHCOV vector would be a 15x1 vector of
dimension 15 = 5*(5+1)/2 with components
VEC1, VEC2,..., VEC15.
The default value is
        NAME=c
which makes inputting the covariance structure
in NLMIXED reasonably simple (see the orange
tree example, Example 5.4.1 for an example).
=====
%macro vech(dim=4, cov=cov, vechcov=vechcov, name=c);
  %global n vech;
  %let n=%eval(&dim*&dim+1)/2;

```

```
array &vechcov. [&n] &name.1-&name.&n;
%let k=0;
%do i=1 %to &dim;
%do j=1 %to &i;
%let k=%eval(&k+1);
&vechcov. [&k] = &cov.[&i, &j];
%end;
%end;
%let temp = &name.1;
%do i=2 %to &n;
%let temp=&temp,&name.&i;
%end;
%let vech = &temp;
%mend vech;
```