Paper 435-2013

JOINT MODELING OF MIXED OUTCOMES IN HEALTH SERVICES RESEARCH

Joseph C. Gardiner

Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI

ABSTRACT

Outcomes with different attributes, of continuous, count and categorical types are often encountered jointly in many settings. For example, two widely used measures of healthcare utilization, length of stay (LOS) and cost can be analyzed jointly with LOS as a count and cost as continuous. Occurrence of an adverse event (binary) would impact both outcomes. For fitting marginal distributions and assessing the impact of explanatory variables on outcome SAS® offers a number of procedures. Correlation and clustering are additional features of these outcomes that must be addressed in analyses. We survey some SAS procedures, GLIMMIX, COPULA, PHREG and QLIM that can be applied to modeling multivariate outcomes of mixed types. Examples from the extant literature are used to demonstrate the application of the procedures.

INTRODUCTION

Hospital-acquired infections and injuries lead to increase in utilization of healthcare resources and cost. Preventable adverse events such as sepsis, pressure ulcers and falls are regarded as caused by medical error or poor medical management.[1] Healthcare reform has instituted polices that view preventable adverse events as a defect in care, the treatment which is not reimbursable. Ensuring patient safety and improving health care delivery are a growing concern of all healthcare professionals. The occurrence of an adverse event, for example a pressure ulcer, is a binary outcome Y_1 with consequence for length of stay Y_2 measured as a count or continuous outcome and hospital cost Y_3 as a continuous response. Variables that impact $\mathbf{Y}=(Y_1,Y_2,Y_3)$ are patient characteristics such as age, gender, race, presenting comorbidity and hospital-level factors. The challenge in analyses is to specify a joint model for $\mathbf{Y}$ given these explanatory variables $\mathbf{z}$ regarded as exogenous. One approach is to model each outcome separately using a generalized linear model, appropriate to the response type, by structuring the mean, $E(Y_k\,|\,\mathbf{z}_k)$ and variance, $Var(Y_k\,|\,\mathbf{z}_k)$, $k=1,2,3$. The covariates $\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3$ need not be the same. In applications some overlap is warranted but excluding some variables from a model for one outcome that are included in a model for another outcome is often needed for purposes of interpretation and identification. An approach to link the outcomes is through a shared random effect ζ in $E(\mathbf{Y}\,|\,\mathbf{z},\zeta)$ or by structuring the covariance matrix $Var(\mathbf{Y}\,|\,\mathbf{z})$ to include potential correlations. Copula regression is another approach that has received some attention.[2]

If we are interested in the effect of Y_1 on the utilization measures we might consider a joint distribution $f(Y_1,Y_2,Y_3\,|\,\mathbf{z})=f(Y_1\,|\,\mathbf{z})f(Y_2,Y_3\,|\,Y_1,\mathbf{z})$ where the generic notation $f(\cdot\,|\,\cdot)$ stands for a conditional distribution. This makes Y_1 potentially endogenous in a model of the second term. Many empirical applications in the econometric literature must deal with both endogeneity and sample selection.[3, 4]

EXAMPLE

The data set is drawn from a sample of hospital discharges for the year 2003 for one mid-western state.[5] Patient-level covariates are: age at admission (restricted to 18 to 84 years), gender, race, a measure of overall presenting comorbidity as assessed by the Charlson Comorbidity Index (CCI),[6] the number of procedures undergone (NPR), and an indicator for obesity (OBESE). These were obtained from ICD-9CM diagnosis and procedure codes. The assessment of the presence of a pressure ulcer (PU) on the discharge record is determined from diagnosis codes (up to 15 codes) with ICD-9CM stem code 707.xx. A small number of records with a principal diagnosis of a PU were excluded. We also restrict to discharges with at least one day for length of stay (LOS). The resulting data set has 12,152 discharges.

Some characteristics of the sample are: presence of PU 2.4%, obese 8.5%, female 61.7%, white race 68.8%, black race 14.1%, age ≥65 years, 39.8%, no comorbidity (CCI=0) 41.6%, and no procedures (NPR=0) 39.6%.

The outcomes for our analysis are: Y_1 =PU, Y_2 =LOS (in days), and Y_3 =CHG for total hospital charge (in \$). Consider the generalized linear model $g_k\left(E(Y_{ik}\,|\,\mathbf{z}_i)\right)=\mathbf{z}'_{ik}\beta_k$, $k=1,2,3$ where the subscript i denotes the individual hospital discharge, and g_k a link function for the outcome indexed by k. Specifically, we take $g_1(u)=\Phi^{-1}(u)$, the probit link, and $g_2(u)=g_3(u)=u$, the identity link. Other natural alternatives are the logit link for g_1 and the log link for g_2, g_3 if we assume that Y_2 is Poisson or Negative-Binomial, and Y_3 is Gamma distributed. Different covariates from the constellation $\mathbf{z}_i$ may be used in the three model equations.

The following formats are applied:

```
proc format;
value female 1='female' 0='male';
value race 1='white' 2='black' 6='other';
value PU 0='No' 1-high='Yes';
value cci 3-high='3+';
value affirm 0='No' 1='Yes';
value npr 0='none' other='1+';
run;
```

Each outcome can be analyzed separately. For example, for the probit model for Y_1, the procedures LOGISTIC, GENMOD, GLIMMIX, QLIM would estimate the model parameters by maximum likelihood. For Y_2 and Y_3 we assume a lognormal model. The procedures GLIMMIX, LIFEREG, SEVERITY and QLIM are some choices for estimation. We will use GLIMMIX to estimate the three marginal models jointly. For analysis, the data set must be pivoted to have three records for each discharge, one for each type of outcome, covariates specific to outcome and a single variable RESPONSE that contains the outcome. Additionally, the appropriate distribution (DIST) and link function (LINK) are defined.

The display of the file `trivar_glx` for 2 discharges (SUBJID=5 and 6) is shown next. For example, the covariates AGE and CCI are used in each model, but OBESE is used only in the model for Y_1 and is omitted in the models for Y_2 and Y_3 by defining OBESE as identically zero. The lognormal distribution for Y_3 is the same as using a normal (Gaussian) distribution for $\log Y_3$. It avoids naming conflicts.

Example of records for two discharges **(N=36,456 records)**

dist	link	PU	LOS	CHG	L_CHG	rtype	response	SUBJID	Age	CCI	obese
Binary	probit	0	30	44649	10.7066	1	0.0000	5	60	6	0
Lognormal	identity	0	30	44649	10.7066	2	30.0000	5	60	6	0
Normal	identity	0	30	44649	10.7066	3	10.7066	5	60	6	0
Binary	probit	0	9	18542	9.8278	1	0.0000	6	50	1	1
Lognormal	identity	0	9	18542	9.8278	2	9.0000	6	50	1	0
Normal	identity	0	9	18542	9.8278	3	9.8278	6	50	1	0

The following syntax will fit the three marginal models with a single innovation of proc GLIMMIX. A similar example is described in the GLIMMIX documentation for two outcomes, one binary with the default logit link and the other Poisson with the default log link.[7] Then the link=byobs(link) is redundant. The ddfm=none option is used to obtain p-values based on standard normal distribution. GLIMMIX defaults to ddfm=residual =N – #parms. Because N is very large and the total number of parameters is 28, there is no practical difference.

```
proc glimmix data=trivar_glx noclprint;
class SUBJID rtype female race cci obese npr PU;
model response(event='1')=rtype female*rtype race*rtype age*rtype cci*rtype
                        obese*rtype npr*rtype/ noint solution
                        link=byobs(link) dist=byobs(dist) ddfm=none;
format female female. race race. PU PU. ;
format obese affirm. cci cci. npr npr.;
run;
```

Table 1: Marginal models for outcomes PU, log(LOS) and log(CHG)

		BINARY (PU)			LOGNORMAL (LOS)			LOGNORMAL (CHG)		
Effect	**Class**	**Estimate**	**StdErr**	**Probt**	**Estimate**	**StdErr**	**Probt**	**Estimate**	**StdErr**	**Probt**
Intercept		–1.8048	0.1548	<.0001	0.9267	0.0336	<.0001	8.3493	0.0330	<.0001
Female	female	–0.1220	0.0536	0.0228	0.0098	0.0136	0.4743	–0.1736	0.0134	<.0001
	male	ref			ref			ref		
Race	black	0.2281	0.0680	0.0008	0.1398	0.0191	<.0001	0.0712	0.0188	0.0002
	other	–0.0470	0.0817	0.5646	0.0380	0.0176	0.0313	0.0446	0.0173	0.0100
	white	ref			ref			ref		
Age		0.0058	0.0019	0.0019	0.0078	0.0004	<.0001	0.0115	0.0004	<.0001
CCI	0	–0.9889	0.0929	<.0001	–0.5029	0.0199	<.0001	–0.3904	0.0195	<.0001
	1	–0.5560	0.0728	<.0001	–0.3482	0.0201	<.0001	–0.1397	0.0197	<.0001
	2	–0.2398	0.0674	0.0004	–0.1843	0.0219	<.0001	–0.0621	0.0215	0.0039
	3+	ref			ref			ref		
Obese	No	–0.2430	0.0794	0.0022	na			na		
	Yes	ref								
NPR	1+	0.2061	0.0563	0.0003	0.2851	0.0135	<.0001	0.8896	0.0133	<.0001
	none	ref			ref			ref		
Scale					0.5075	0.0065		0.4901	0.0063	
–2 LogL		2447.78			26243.07			25819.55		
Pearson χ^2/DF		0.9082			0.5079			0.4905		

Scale: variances σ_2^2 for log(LOS) and σ_3^2 for log(CHG); na: not applicable, covariate omitted; ref: reference category.

Table 1 is assembled from the `solution` request. With the `noint` option the class variable rtype plays the role of an explicit intercept term and crossing all effects with rtype in the model statement ensures estimates specific to each response type. For rtype=1 the option `event='1'` models $P[Y_{i1}=1\,|\,\mathbf{z}_i]$. For rtype=2 and rtype=3, normal distributions are fitted after log transformation, that is Y_{i2} =log(LOS) and Y_{i3} =log(CHG) are normally distributed. A structural formulation of the model is given by

$$Y_{i1}^{*}=\mathbf{z}_{i1}'\beta_1+\varepsilon_{i1}\,,\ Y_{i2}=\mathbf{z}_{i2}'\beta_2+\varepsilon_{i2},Y_{i3}=\mathbf{z}_{i3}'\beta_3+\varepsilon_{i3}.$$

The observables are the indicator $Y_{i1}=[Y_{i1}^{*}>0]$, Y_{i2} and Y_{i3}. The model error is $\varepsilon_i=(\varepsilon_{i1},\varepsilon_{i2},\varepsilon_{i3})\sim N(\mathbf{0},\Sigma)$, where $\Sigma=\begin{bmatrix}1 & \rho_{12}\sigma_2 & \rho_{13}\sigma_3\\ \rho_{12}\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3\\ \rho_{13}\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2\end{bmatrix}$.

The model in table 1 assumes $\Sigma=\text{diag}(1,\sigma_2^2,\sigma_3^2)$ which is the same as estimating the three responses individually, except for the issue of the degrees of freedom. The estimates of σ_2^2 and σ_3^2 are in the row labeled 'scale' in the table. In general $Cov(Y_{i1},Y_{ik}\,|\,\mathbf{z}_i)=\rho_{1k}\sigma_k\phi(\mathbf{z}_{i1}'\beta_1)$, $k=2,3$, $E(Y_{i1}\,|\,\mathbf{z}_i)=\Phi(\mathbf{z}_{i1}'\beta_1)$ and $Var(Y_{i1}\,|\,\mathbf{z}_i)=\Phi(\mathbf{z}_{i1}'\beta_1)\left(1-\Phi(\mathbf{z}_{i1}'\beta_1)\right)$ where ϕ and Φ denote the density and cumulative distribution of the standard normal distribution. GLIMMIX will structure the variance matrix of $\mathbf{Y}_i=(Y_{i1},Y_{i2},Y_{i3})$ as $Var(\mathbf{Y}_i\,|\,\mathbf{z}_i)=\mathbf{A}_i^{1/2}\mathbf{R}_i\mathbf{A}_i^{1/2}$ where $\mathbf{R}_i$ is a user-specified 3×3 covariance structure and $\mathbf{A}_i$ is the diagonal matrix of the variances of (Y_{i1},Y_{i2},Y_{i3}). It is not possible to choose a structure to match the structural formulation.

The model for the two continuous responses (Y_{i2},Y_{i3}) = (log(LOS), log(CHG)) can be estimated by restricting the data set to

```
data=trivar_glx(where=(rtype in(2,3)))
```

and supplying the statement:

```
random residual/subject=subjid type=un v=1 vcorr=1;
```

Using residual pseudo-likelihood we get the same estimates of σ_2 and σ_3, and additionally $\rho_{23}=0.6193$. This is precisely the partial correlation of (Y_{i2},Y_{i3}) after controlling for covariates female, race, age, cci, npr. It could be verified using proc CORR.

Although a more careful evaluation of the potential correlates of the three outcomes is necessary, including an assessment of plausible interactions, we see that the estimates in table 1 are generally in the expected direction. Higher comorbidity, older age, and undergoing one or more procedures are associated with longer LOS and hospital charge. Gender has a significant effect on hospital charge, lower for females compared to males. For LOS the gender effect is in the opposite direction, but not significant. Among correlates of the likelihood of acquiring a pressure ulcer during the hospital stay we see that higher comorbidity, older age, obesity and male gender are associated with higher probability of a pressure ulcer. Other studies have reported higher incidence among patients who are older, thinner (based on body mass index), incontinent, immobile, and with poor nutritional intake.[8, 9]

STRUCTURAL MODEL

To estimate a joint model using the aforementioned structural form, we use the data set `trivar` of 12,152 hospital discharges. For each discharge there is one record (line) for the covariates and the outcomes Y_1 =PU, Y_2=L_LOS, Y_3= L_CHG, the latter two for the logged response. Proc QLIM is harnessed to estimate the parameters. Three model statements are required together with the `endogenous` statement to declare PU as a binary indicator and specify the probit model, $E(Y_{i1} \mid \mathbf{z}_i) = \Phi(\mathbf{z}'_{i1}\beta_1)$. Both L_LOS and L_CHG have the default normal marginals. The following syntax will invoke maximum likelihood estimation of the parameters of the 3-equation system.

```
proc qlim data= trivar method=newrap;
class female race cci obese npr PU;
endogenous PU~discrete(order=formatted dist=normal);
model PU= female race age cci obese npr;
model L_LOS = female race age cci npr;
model L_CHG = female race age cci npr;
format female female. race race. PU PU.;
format obese affirm. cci cci. npr npr.;
run;
```

Table 2: Joint estimation of outcomes PU, log(LOS) and log(CHG)

		BINARY (PU)			**LOGNORMAL (LOS)**			**LOGNORMAL (CHG)**		
Effect	**Class**	**Estimate**	**StdErr**	**Probt**	**Estimate**	**StdErr**	**Probt**	**Estimate**	**StdErr**	**Probt**
Intercept		–1.7127	0.1551	<.0001	0.9267	0.0336	<.0001	8.3493	0.0330	<.0001
FEMALE	female	–0.0941	0.0535	0.0787	0.0098	0.0136	0.4743	–0.1736	0.0134	<.0001
	male	ref			ref			ref		
RACE	black	0.2100	0.0676	0.0019	0.1398	0.0191	<.0001	0.0712	0.0188	0.0002
	other	–0.0461	0.0808	0.5682	0.0380	0.0176	0.0313	0.0446	0.0173	0.0100
	white	ref			ref			ref		
AGE		0.0049	0.0019	0.0084	0.0078	0.0004	<.0001	0.0115	0.0004	<.0001
CCI	0	–1.0074	0.0930	<.0001	–0.5029	0.0199	<.0001	–0.3904	0.0195	<.0001
	1	–0.5841	0.0730	<.0001	–0.3482	0.0201	<.0001	–0.1397	0.0197	<.0001
	2	–0.2606	0.0672	0.0001	–0.1843	0.0219	<.0001	–0.0621	0.0215	0.0039
	3+	ref			ref			ref		
Obese	No	–0.2358	0.0785	0.0027	na			na		
	Yes	ref								
NPR	1+	0.1366	0.0563	0.0153	0.2851	0.0135	<.0001	0.8896	0.0133	<.0001
	none	ref			ref			ref		
ρ_{23}		0.6194	0.0056	<.0001						
ρ_{12}		0.2665	0.0247	<.0001						
ρ_{13}		0.0816	0.0242	0.0007						
Scale					σ_2 0.7124	0.0046		σ_3 0.7001	0.0045	

na: not applicable, covariate omitted; ref: reference category.

Comparing the results in tables 1 and 2, the properties of maximum likelihood estimators of the mean and variance in the normal distribution lead to the same estimates for the models for log(LOS) and log(CHG). The results for the probit model are only slightly different. The likelihood is constructed from $f(Y_1 = y_1, Y_2, Y_3 \mid \mathbf{z}) = P[Y_1 = y_1 \mid Y_2, Y_3, \mathbf{z}] f(Y_2, Y_3 \mid \mathbf{z})$. The second term is the bivariate normal density

$$\phi_2(u_2,u_3)=\frac{1}{2\pi\sigma_2\sigma_3(1-\rho_{23}^2)^{1/2}}\exp\left(-\tfrac{1}{2}(u_2^2+u_3^2-2\rho_{23}u_2u_3)/(1-\rho_{23}^2)\right)$$

where $u_k=(Y_k-\mathbf{z}_k'\beta_k)/\sigma_k$, k=2,3. The term $P[Y_1=1\,|\,Y_2,Y_3,\mathbf{z}]$ is evaluated from the conditional distribution of ε_1 given $(\varepsilon_2,\varepsilon_3)$ which is normally distributed with mean $E(\varepsilon_1\,|\,\varepsilon_2,\varepsilon_3)=(1-\rho_{23}^2)^{-1}\left\{(\rho_{12}-\rho_{13}\rho_{23})\varepsilon_2/\sigma_2+(\rho_{13}-\rho_{12}\rho_{23})\varepsilon_3/\sigma_3\right\}$ and variance $Var(\varepsilon_1\,|\,\varepsilon_2,\varepsilon_3)=1-(1-\rho_{23}^2)^{-1}\left\{\rho_{12}(\rho_{12}-\rho_{13}\rho_{23})+\rho_{13}(\rho_{13}-\rho_{12}\rho_{23})\right\}$.[3] It involves all three correlation parameters.

Proc QLIM also permits testing of linear hypotheses on parameters. The following `test` statement gives the Wald and likelihood ratio test for testing $H_0:\rho_{12}=\rho_{13}=\rho_{23}=0$. Note that parameter names created by the QLIM procedure are used to specify the hypothesis.

```
test "nocorr" _Rho.L_LOS.L_CHG=0,
              _Rho.L_LOS.PU=0,
              _Rho.L_CHG.PU/wald lr;
```

Test Results				
Test	**Type**	**Statistic**	**Pr > ChiSq**	**Label**
"nocorr"	Wald	110.54	<.0001	_Rho.L_LOS.L_CHG = 0 , _Rho.L_LOS.PU = 0 , _Rho.L_CHG.PU = 0
"nocorr"	L.R.	5991.0	<.0001	_Rho.L_LOS.L_CHG = 0 , _Rho.L_LOS.PU = 0 , _Rho.L_CHG.PU = 0

ENDOGENEITY

Consider the system $Y_{i1}^*=\mathbf{z}_{i1}'\beta_1+\varepsilon_{i1}$, $Y_{i2}=\mathbf{z}_{i2}'\beta_2+\alpha Y_{i1}+\varepsilon_{i2}$, where $Y_{i1}=[Y_{i1}^*>0]$. The covariates $\mathbf{z}_i=(\mathbf{z}_{i1},\mathbf{z}_{i2})$ are exogenous by which is meant that $(\varepsilon_{i1},\varepsilon_{i2})$ is independent of $\mathbf{z}_i$. Since Y_{i1} appears on the right-hand side of the second equation it is potentially correlated with the error ε_{i2}, that is Y_{i1} is endogenous. We can estimate this model by maximum likelihood in proc QLIM, although in general it regards all right-hand side variables as exogenous. The reason why the estimation works is because the likelihood is constructed in two parts: on $Y_1=1$ using $f(Y_1=1,Y_2\,|\,\mathbf{z})=P[\varepsilon_1>-\mathbf{z}_1'\beta_1\,|\,Y_2,\mathbf{z}]f(Y_2\,|\,\mathbf{z})$ and similarly on $Y_1=0$. With Y_1=PU and Y_2=log(LOS) the syntax to estimate the model is

```
proc qlim data= trivar method=newrap;
class female race cci obese npr;
endogenous PU~discrete(dist=normal);
model PU= female race age cci obese npr;
model L_LOS = PU female race age cci npr;
format female female. race race.;
format obese affirm. cci cci. npr npr.;
run;
```

Table 3: Joint model for log (LOS) and PU

		LOGNORMAL (LOS)			BINARY (PU)		
Parameter	**Class**	**Estimate**	**StdErr**	**Probt**	**Estimate**	**StdErr**	**Probt**
Intercept		0.9042	0.0338	<.0001	–1.8040	0.1548	<.0001
PU	Yes	0.4904	0.1164	<.0001			
	No	ref					
FEMALE	female	0.0126	0.0136	0.3547	–0.1222	0.0536	0.0227
	male	ref			ref		
RACE	black	0.1328	0.0191	<.0001	0.2290	0.0681	0.0008
	other	0.0388	0.0176	0.0270	–0.0470	0.0817	0.5650
	white	ref			ref		
AGE		0.0078	0.0004	<.0001	0.0057	0.0019	0.0021
CCI	0	–0.4761	0.0208	<.0001	–0.9886	0.0929	<.0001
	1	–0.3263	0.0206	<.0001	–0.5554	0.0729	<.0001
	2	–0.1718	0.0220	<.0001	–0.2394	0.0674	0.0004
	3+	ref			ref		
NPR	1+	0.2793	0.0135	<.0001	0.2092	0.0573	0.0003
	none	ref			ref		
Obese	No	na			–0.2441	0.0795	0.0021
	Yes				ref		
Scale σ_2		0.7090	0.0045	<.0001			
ρ_{12}		–0.0188	0.0682	0.7824			

na: not applicable, covariate omitted; ref: reference category.

From table 3 we see that the directions of the estimated effects remain the same as in previous analyses. For pressure ulcer (PU) incidence the coefficient is positive indicating an impact of lengthening LOS. The Wald test for endogeneity $H_0:\rho_{12}=0$ cannot be rejected.

A quantity of interest in the above model is the expected LOS. Because Y_2 =log(LOS) we calculate $E\left(\exp(Y_2)\,|\,Y_1,\mathbf{z}\right)=\exp(\mathbf{z}_2'\beta_2+\alpha Y_1+½\sigma_2^2)\dfrac{\Phi(\mathbf{z}_1'\beta_1+\rho_{12}\sigma_2)}{\Phi(\mathbf{z}_1'\beta_1)}$. The effect of the correlation is in the second term, called the smear.[10] We compute this expression for the full dataset twice: first assuming counterfactually that Y_1 =1 in all records, and second also counterfactually that Y_1 =0 in all records. Then we compute the sample averages.

Add the option `outest=est` to the proc QLIM statement to save the parameter estimates and add the statement `output out=stats_q xbeta errstd;` Then `xbeta_PU` is the estimated $\mathbf{z}_1'\beta_1$ in the PU model and `xbeta_L_LOS` is the estimated $\mathbf{z}_2'\beta_2+\alpha Y_1$ in the log(LOS) model. Estimates of α and σ_2 are named `L_LOS_PU` and `errstd_L_LOS` respectively.

```
data expted;
if _n_=1 then set est(obs=1 keep=_rho L_LOS_PU);
set stats_q(keep=PU xbeta_PU xbeta_L_LOS errstd_L_LOS);
smear=CDF("normal", xbeta_PU + errstd_L_LOS *_rho)/CDF("normal", xbeta_PU);
Mean_0= exp(xbeta_L_LOS-L_LOS_PU*(PU=1)+.5* errstd_L_LOS**2)*smear;
Mean_1= exp(xbeta_L_LOS+L_LOS_PU*(PU=0)+.5* errstd_L_LOS**2)*smear;
run;
```

The sample means of mean_1 and mean_0 are respectively, 7.33 days and 4.49 days. Their ratio should be exp(α)−1 =0.633. Our simple analysis should not be interpreted inferentially because many other important factors that might influence both LOS and CHG have not been evaluated. We have also assumed that $\sigma_2^2 = Var(\varepsilon_{i2})$ is constant. Homoscedasticity is untenable for LOS and CHG even when modeled on the log transformed scale. Proc QLIM offers some options to model heteroscedasticity in σ_2^2 through the HETERO statement. For example if $\sigma_2^2(\mathbf{z}_i) = \sigma^2 \exp(\mathbf{z}_i'\gamma)$ with $\mathbf{z}_i$ =(female cci0 cci1 cci2), we would use

```
hetero L_LOS~ female cci0 cci1 cci2 npr1/link=exp noconst;
```

For numerical stability we created dummy variables cci0, cci1, cci2 corresponding to the levels of the comorbidity index CCI in table 3. The same syntax can be used for calculation of the smear, mean_1 and mean_0.

COPULAS

As mentioned previously among the challenges in the analysis of multivariate outcomes of mixed types is the specification of a joint distribution that accommodates the different measurement scales and dependencies among the outcomes. It might be relatively easy to specifiy a marginal model for each outcome and then link them together through a random effect. This is what we attempted to achieve in our first example using Proc GLIMMIX. Copulas provide a general approach to link the specified marginal distributions to get a joint distribution for the outcomes. For example, focussing on (Y_2, Y_3) for log(LOS) and log(CHG), a joint distribution function $F(y_2, y_3) = C\left(F_2(y_2), F_3(y_3)\right)$ is constructed from their marginal distributions F_2, F_3 using a copula C. The copula C as applied here is a continuous joint distribution function on the unit square $(u_1, u_2) \in [0,1]^2$ for dependent random variables (U_1, U_2) whose marginal distributions are uniform on [0, 1]. For a thorough discussion of copulas see Nelson (2006).[11]

For outcomes (Y_2, Y_3) we previously fitted a bivariate normal model. Distibutions that might give better fit to LOS and cost, are the log-logistic for LOS and log-normal or Gamma for cost. The practical use of a copula is to infuse dependence in (Y_2, Y_3). This dependence is a property of the copula and not of the marginals. Proc COPULA offers five copula functions for fitting and simulating of a joint distribution.[12] They are the normal (Gaussian), Student's t, Clayton, Frank and Gumbel-Hougard copulas. The lattter three are members of Archimedian families that can be constructed as $C(u_1, u_2) = \varphi^{-1}\left(\varphi(u_1) + \varphi(u_2)\right)$ where the generator $\varphi : [0,1] \to [0,\infty]$ is a continuous, convex, stricting decreasing function, with $\varphi(0) = \infty, \varphi(1) = 0$.

A convex combination of copulas is a copula, and so are continous mixtures of a familiy of copulas.[11] A generator for an Archimedean copula is easily obtained from the Laplace transform of a non-negative random variable X. If $\psi(t) = E\left(\exp(-tX)\right), t \ge 0$ then $\varphi(t) = \psi^{-1}(t)$ is a generator for an Archimedean copula. Three simple copulas are the *independence* copula Π, the *Fréchet lower bound W* and *Fréchet upper bound M* defined by $\Pi(u_1, u_2) = u_1 u_2$, $W(u_1, u_2) = \max\{0, u_1 + u_2 - 1\}$, $M(u_1, u_2) = \min\{u_1, u_2\}$. All copulas C are captured by the Fréchet bounds in the sense that $W \le C \le M$.

The *Gaussian* copula is defined by $C_\theta(u_1,u_2)=\Phi_2\left(\Phi^{-1}(u_1),\Phi^{-1}(u_2)\right)$, $\theta\in[-1,1]$ where Φ_2 is the bivariate normal distribution function with correlation θ, unit variances, and zero means. Association measures are Spearman's rho $\rho_\theta=6\pi^{-1}\arcsin(\tfrac{1}{2}\theta)$ and Kendall's tau $\tau_\theta=2\pi^{-1}\arcsin(\theta)$.[13]

The *Student's t* copula is defined by $C_\theta(u_1,u_2)=\mathrm{T}_2\left(\mathrm{T}^{-1}(u_1),\mathrm{T}^{-1}(u_2)\right),\theta=(\nu,\phi),\nu\in(1,\infty),\phi\in[-1,1]$ where T is the univariate central *t*-distribution function with ν degrees of freedom , T_2 is the bivariate t-distribution function with correlation ϕ and ν degrees of freedom. Kendall's tau $\tau_\theta=2\pi^{-1}\arcsin(\phi)$, but there is no closed expression for Spearman's rho.

The *Gumbel-Hougaard* copula is defined by $C_\theta(u_1,u_2)=\exp\left(-\left[\{-\log u_1\}^\theta+\{-\log u_2\}^\theta\right]^{1/\theta}\right)$, $\theta\in[1,\infty)$. Kendall's tau $\tau_\theta=1-\theta^{-1}$ but there is no simple form for Spearman's ρ.

EXAMPLE

Using the data set of 12,152 hospitals discharges we explore fitting a copula to LOS and CHG. As before the measures are log-transformed and the model is $Y_{i2}=\mathbf{z}'_{i2}\beta_2+\sigma_2\varepsilon_{i2},Y_{i3}=\mathbf{z}'_{i3}\beta_3+\sigma_3\varepsilon_{i3}$ where the covariates $\mathbf{z}_{i2},\mathbf{z}_{i3}$ are female gender, race, age, the comorbidity index (CCI) and number of procedures (NPR). We first fit log-logistic distributions to LOS and CHG which means $\varepsilon_{i2},\varepsilon_{i3}$ have the logistic (survival) distribution $S(u)=(1+e^u)^{-1}$, $-\infty<u<\infty$. The following syntax fits the regression model for LOS with the same syntax for CHG instead of LOS.

```
ods output parameterestimates=parms_L;
proc lifereg data= trivar;
class female race cci npr;
model LOS=female race age cci npr/dist=llogistic;
format female female. race race. cci cci. npr npr.;
output out=stats_LOS cres=cres_LOS sres=sres_LOS;
run;
```

Standardized residuals (SRES) and Cox-Snell residuals (CRES) are computed as: $s_{ik}=(Y_{ik}-\mathbf{z}'_{ik}\hat{\beta}_k)/\hat{\sigma}_k$ and $c_{ik}=-\log S(s_{ik}),k=2,3$ respectively. Under the assumed model $\{c_{ik}:1\le i\le n\}$ should behave like a sample from the exponential distribution with mean=1.[14] Use proc LIFETEST to estimate the cumulative hazard function *H* regarding CRES as "time". Overall fit can be gauged visually to see if there is gross departure from the exponential cumulative hazard $H_e(t)=t$.[15]

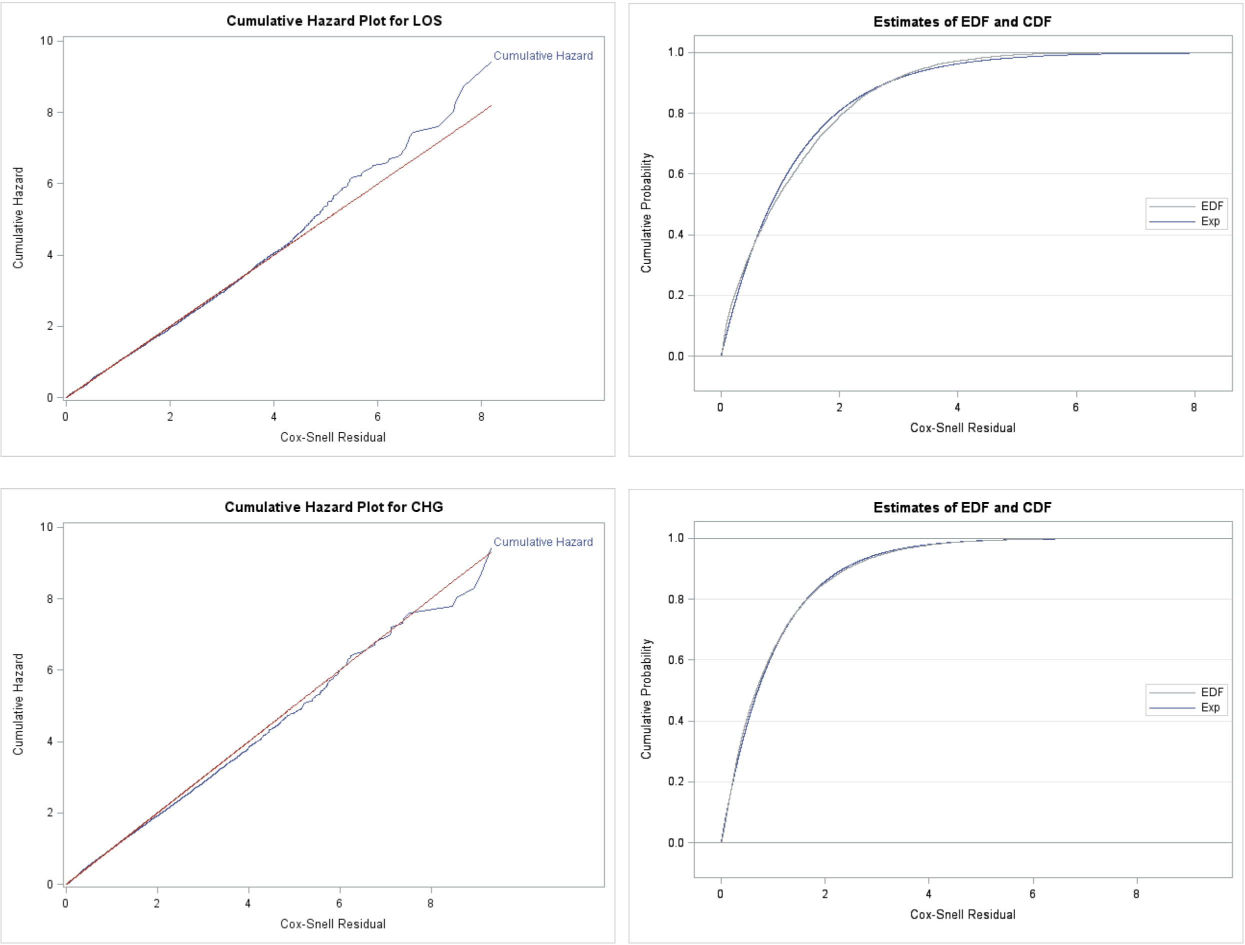
```
proc lifetest data=stats_LOS notable outsurv=surv_L;
time cres_LOS; run;

data surv_L;
set surv_L(where=(survival>0));
logsurv=-log(survival); run;
```

An alternative is to plot the empirical distribution function (EDF) and a fitted exponential distribution for $\{c_{ik} : 1 \le i \le n\}$ using proc SEVERITY.

```
proc severity data=surv_L plots=(cdf);
dist exp;
loss CRES_LOS;
run;
```

Figure 1: Cumulative Hazard and Empirical Distribution of Cox-Snell Residuals



In Figure 1 the right hand side EDF plots are the default output from proc SEVERITY. The left-hand side CRES plots are generated using, for example

```
proc sgplot data=surv_L;
series x=cres_LOS y=logsurv/curvelabel='Cumulative Hazard';
series x=cres_LOS y=cres_LOS;
label logsurv='Cumulative Hazard';
title "Cumulative Hazard Plot for LOS";
run;
```

We might be inclined to accept the log-logistic model for CHG, but for LOS it is rather tenuous. A quantitative assessment of the goodness-of-fit with Kolmogorov-Smirov, Anderson-Darling or Cramer-von Mises statistics is outside the scope of the present article. Perhaps another distribution for LOS such as the Pareto or a Coxian phase-type might be appropriate.[16, 17]

ESTIMATING A COPULA MODEL

We begin by assessing which of the five copulas available in proc COPULA would be a viable option for fitting a joint distribution to log-transformed (LOS, CHG). To this objective save the standardized residuals (SRES) in a data set `residuals_all`.

Let $\{(s_{i2}, s_{i3}) : 1 \le i \le n\}$ be a SRES sample. Using the EDFs, $F_{2n}(y_2) = n^{-1}\sum_{i=1}^{n}[s_{i2} \le y_2]$, $F_{3n}(y_3) = n^{-1}\sum_{i=1}^{n}[s_{i3} \le y_3]$ the sample is transformed to pseudo data $\{(U_{i2}, U_{i3}) : 1 \le i \le n\}$ where the components have uniform marginals: $U_{i2} = F_{2n}(s_{i2}),\ \ U_{i3} = F_{3n}(s_{i3})$. For each copula C the likelihood is constructed for the pseudo data. Maximum likelihood estimation (MLE) gives estimates of the association parameters of the copula. For the Gumbel-Hougaard copula use the following syntax, saving the estimated association parameter θ in the data set OUTCOPULA=`assoc`. The results of the five estimations are assembled in Table 4.

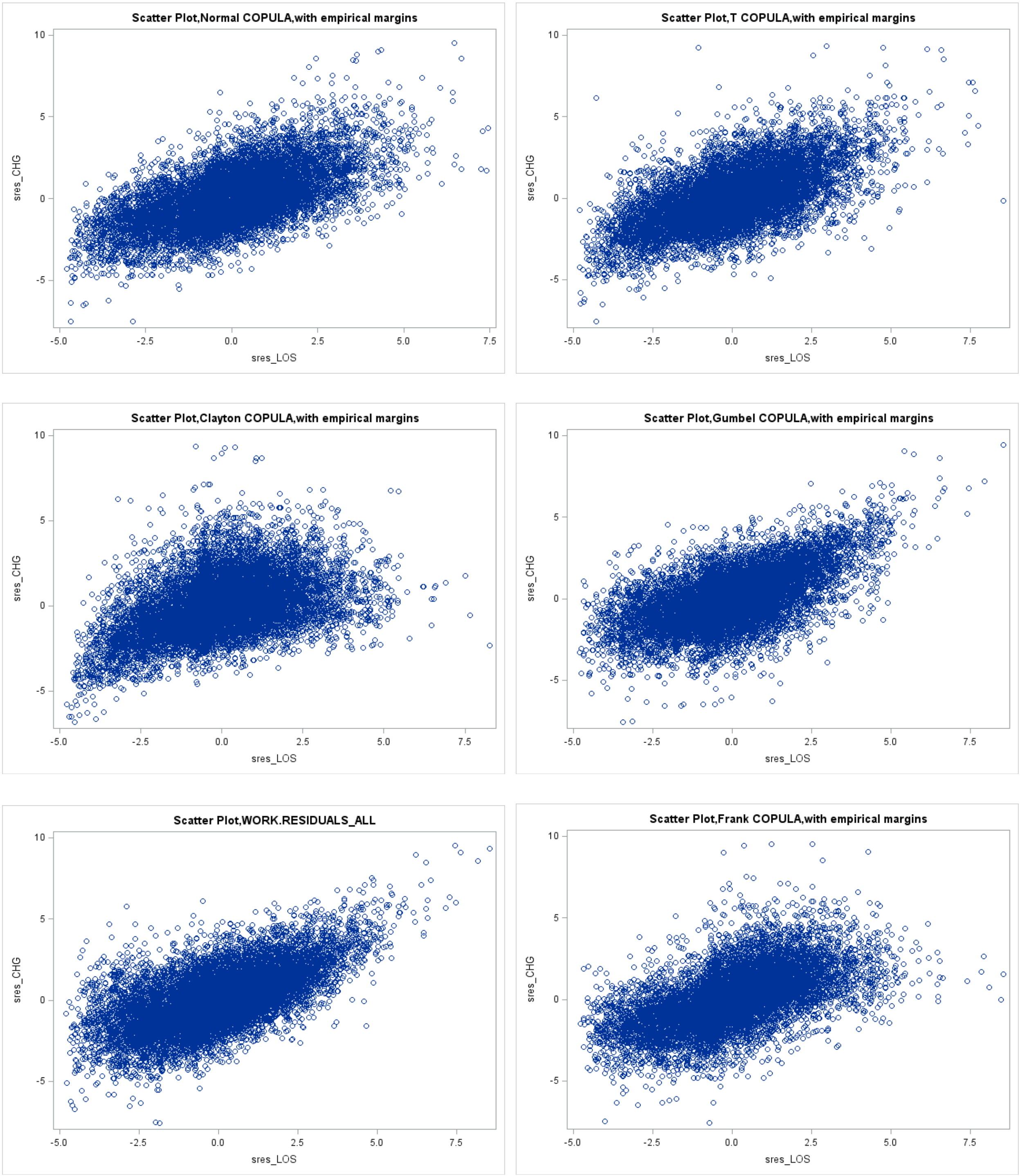
```
proc copula data=residuals_all;
var  sres_LOS sres_CHG;
fit gumbel/marginals    =empirical
            outcopula   =assoc
            method      =mle;
simulate/marginals      =empirical
            ndraws     =10000
            seed       =30213
            plots      =(data=original scatter) ;
run;
```

Table 4: Copula Association Parameter Estimates by MLE, N=12,152

Distribution	Parameter	Estimate	Standard Error	t Value
t	DF, ν	8.133635	0.665042	12.23
	Correlation, ϕ	0.614905		
Gaussian	Correlation, θ	0.610811		
Clayton	Association, θ	0.824992	0.016562	49.81
Gumbel	Association, θ	1.707093	0.012343	138.31
Frank	Association, θ	4.535135	0.064000	70.86

Having estimated the association parameter we now simulate a sample of NDRAWS from the copula. The SIMULATE statement is added to the above syntax after the FIT statement. To obtain the simulated sample $\{(\tilde{s}_{b2}, \tilde{s}_{b3}) : 1 \le b \le B\}$ with the same marginal distributions as the EDFs (F_{2n}, F_{3n}) of the original data, use the MARGINALS=EMPIRICAL option. The PLOTS= option also requests a scatter plot of the simulated sample with the same marginals as the original data.

Figure 2: Scatter Plots of 10,000 simulated samples from five copulas (Original data at bottom left)



The original scatter plot of the residuals (N=12,152) is in the bottom left hand corner. Other scatter plots are from the simulated data (N=10,000) of their respective copulas. Visual examination of these scatter plots suggest that the Gumbel copula is closer to the original data than any of the others. Comparisons based on Kolmogorov-Smirov, Anderson-Darling or Cramer-von Mises statistics could be made.[18]

ESTIMATION OF THE GUMBEL-HOUGAARD COPULA

Proc COPULA does not currently support copula regression models. Our objective is to estimate the parameters of a bivariate Gumbel-Hougaard regression model for log-transformed (LOS, CHG), $Y_{i2} = \mathbf{z}'_{i2}\beta_2 + \sigma_2\varepsilon_{i2}, Y_{i3} = \mathbf{z}'_{i3}\beta_2 + \sigma_3\varepsilon_{i3}$ where $(\varepsilon_{i2}, \varepsilon_{i3})$ have marginal logisic distributions. Proc NLMIXED is harnessed to perform the optimation of the likelihood constructed from the density function $c_\theta(u_1,u_2)$ of the copula which is given by

$$c_\theta(u_1,u_2) = C_\theta(u_1,u_2)(u_1u_2)^{-1}\left(\tilde{u}_1\tilde{u}_2\right)^{1-1/\theta}\left((\tilde{u}_1+\tilde{u}_2)^{1/\theta}+\theta-1\right)/\left((\tilde{u}_1+\tilde{u}_2)^{2-1/\theta}\right)$$

where $\tilde{u}_1 = (-\log u_1)^\theta, \tilde{u}_2 = (-\log u_2)^\theta$. Expressed in terms of $e_{i2} = (Y_{i2} - \mathbf{z}'_{i2}\beta_2)/\sigma_2$ and $e_{i3} = (Y_{i3} - \mathbf{z}'_{i3}\beta_3)/\sigma_3$ the joint density is $f(e_2,e_3) = c_\theta\left(F(e_2),F(e_3)\right)f(e_2)f(e_3)/\sigma_2\sigma_3$ where F and f are respectively, the standard logistic cumulative distribution and density functions.

Initial values for the parameters $(\beta_2,\beta_3,\sigma_2,\sigma_3,\theta)$ are obtained from the previously fitted marginal distributions with proc LIFEREG and from proc COPULA for the association parameter θ. We have assembled them into a single data set `parms_init` combining the three data sets `parms_L, parms_C` and `assoc.`

```
proc nlmixed data=trivar gconv=0;
dummy=1;
parms/data=parms_init;

race_b=(race=2); race_o=(race not in (1 2));
cci0=(cci=0); cci1=(cci=1); cci2=(cci=2);
npr1=(npr>=1);

xb=b0+b1*female+b2*race_b+b3*race_o+b4*age+b5*cci0+b6*cci1+b7*cci2+ b8*npr1;
xc=c0+c1*female+c2*race_b+c3*race_o+c4*age+c5*cci0+c6*cci1+c7*cci2+ c8*npr1;

e1=(log(los)-xb)/b9;
e2=(log(chg)-xc)/c9;

u1=CDF("LOGISTIC", e1); u1t=(-log(u1))**theta;
u2=CDF("LOGISTIC", e2); u2t=(-log(u2))**theta;

JLIK1=LOGPDF("LOGISTIC",e1)+LOGPDF("LOGISTIC", e2)-log(b9)-log(c9);

JLIK2=-(u1t+u2t)**(1/theta);
JLIK3=-log(u1)-log(u2)+(theta-1)*(log(-log(u1))+log(-log(u2)));

JLIK4=log(theta-1-JLIK2);
JLIK5=(-2+(1/theta))*log(u1t+u2t);

JLIK=JLIK1+JLIK2+JLIK3+JLIK4+JLIK5;

model dummy~general(JLIK);
run;
```

Table 5: Gumbel-Hougaard Copula for log(LOS) and log(CHG) with logistic marginals

		LOG-LOGISTIC (LOS)				LOG-LOGISTIC (CHG)			
Parameter	**Class**	**Estimate**	**STDERR**	**tValue**	**Probt**	**Estimate**	**STDERR**	**tValue**	**Probt**
Intercept		0.7814	0.0324	24.12	<.0001	8.2306	0.0317	259.26	<.0001
FEMALE	female	–0.0066	0.0132	–0.50	0.6180	–0.1832	0.0130	–14.09	<.0001
RACE	black	0.1280	0.0181	7.06	<.0001	0.0749	0.0181	4.15	<.0001
RACE	other	0.0369	0.0166	2.23	0.0260	0.0376	0.0167	2.25	0.0246
AGE		0.0094	0.0004	25.33	<.0001	0.0125	0.0004	33.67	<.0001
CCI	0	–0.4309	0.0194	–22.20	<.0001	–0.3423	0.0190	–18.03	<.0001
CCI	1	–0.2691	0.0193	–13.95	<.0001	–0.0989	0.0191	–5.19	<.0001
CCI	2	–0.1461	0.0211	–6.94	<.0001	–0.0530	0.0208	–2.55	0.0109
NPR	1+	0.3099	0.0130	23.88	<.0001	0.8992	0.0128	70.20	<.0001
Scale		0.3995	0.0029	136.40	<.0001	0.3950	0.0029	135.62	<.0001
Theta		1.7006	0.0145	117.39	<.0001				
–2 Log L		45868							

The results from the maximum likelihood estimation are shown in Table 5. The estimates and their standard errors differ from their naïve counterparts from fitting marginal models, ignoring the association. If θ=1 the Gumbel-Hougaard copula reduces to the independence copula. A formal test of $H_0 : \theta = 1$ would be rejected based on the Wald test, which is not surprising from the association seen in figure 2. Because testing H_0 places the parameter value on boundary of the parameter space, the asymptotic distribution of the likelihood ratio test statistic is generally non-standard. A comparison of above model with a bivariate Gaussian copula model (table 2, middle and right panels) by a formal likelihood ratio test for two non-nested models[19] will support the Gumbel-Hougaard copula.

If estimates of mean LOS and mean CHG are desired for a specified covariate profile, it can be requested from an ESTIMATE statement. Because both $\sigma_2 < 1$ and $\sigma_3 < 1$, $E(LOS \mid \mathbf{z}) = \exp(\mathbf{z}'\beta_2)\Gamma(1+\sigma_2)\Gamma(1-\sigma_2)$ and $E(CHG \mid \mathbf{z}) = \exp(\mathbf{z}'\beta_3)\Gamma(1+\sigma_3)\Gamma(1-\sigma_3)$ are finite. It might be desirable to use the logged version in estimation, although results still depend on the asymptotic distribution of the MLE and accuracy of the delta method approximation. Consider the profile, male, age=58, race=white, CCI≥3 and NPR≥1. The following statements are added to the previous proc NLMIXED syntax:

```
estimate 'LOG LOS'  b0+b4*58+b8+LGamma(1+b9)+LGamma(1-b9);
estimate 'LOG CHG'  c0+c4*58+c8+LGamma(1+c9)+LGamma(1-c9);
```

The mean LOS is 6.8 days (95% CI: 6.6, 7.0), mean hospital charge $25,053 (95% CI: 24,191, 25,946).

SUMMARY

In this article we demonstrated the use of SAS procedures for analyzing multivariate outcomes of dissimilar types. The workhorse for correlated data analysis, proc GLIMMIX can be adapted to the setting discussed in this paper, if an explicit joint distribution is not posited, but dependencies between outcomes need to be acknowledged. The generalized linear (mixed) model is an excellent framework for this type of analysis.

We also discussed a structural model for binary and continuous outcomes where explicit error terms that have a multivariate normal distribution can be exploited to construct a joint likelihood.[4, 20] Here covariates need not be exogenous and indeed interesting applications in econometrics address both endogeneity and sample selection issues. Proc QLIM can be applied in this context, but some attention must be given to the structural implications because currently QLIM does not support models with right-hand side endogenous variables.[12] The basic idea is to parse the joint distribution of say three outcomes (Y_1, Y_2, Y_3) into conditional components suggested by the structure of the model. See Wooldridge (2010) for several applications including some non-likelihood based two-stage methods of estimation.

Although the theory of copulas has been in the literature for many decades, copula regression models, especially in the breadth of empirical applications, have seen some interesting recent developments. This growing field of research is gaining popularity in several areas, in economics, finance, insurance, and health services where correlated binary, count and continuous outcomes are dominant. [21-24] We did not discuss applications with time-to-event outcomes where censoring must be addressed. For example, in survival studies a biomarker (eg, CD4 counts) is assessed at different times during follow-up. Our interest is the impact of the biomarker measurements $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_K)$ on survival time T which might be right censored. Modeling $(T, \mathbf{Y})$ could be approached as a pattern-mixture or as a selection model[25-27] depending on how the joint distribution is constructed. The suite of SAS procedures LIFEREG, PHREG, QLIM, QUANTLIFE and SEVERITY could be used to inform more complex joint models involving copulas.[2, 28-30] It is likely that future enhancements to SAS software will have capabilities for analysis of these models.

REFERENCES

1. Russo CA, Steiner C, Spector W. *Hospitalizations Related to Pressure Ulcers, 2006. HCUP Statistical Brief #64*. Rockville, MD: Agency for Healthcare Research and Quality; 2008.

2. Kolev N, Paiva D. Copula-based regression models: A survey. *Journal of Statistical Planning and Inference.* 2009;139(11):3847-3856.

3. Luo Z, Gardiner JC, Yang N. Estimation of mean response in selected samples with endogenous variables. *Journal of Statistics & Applications.* 2008;3:217-238.

4. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data, 2nd Edition*. Cambridge, MA: Massachusetts Institute of Technology, MIT Press; 2010.

5. HCUP Overview: Healthcare Cost and Utilization Project (HCUP). Rockville, MD: Agency for Healthcare Research and Quality; 2009.

6. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis.* 1987;40(5):373-383.

7. SAS/STAT User's Guide, ver 9.3. Cary, NC: SAS Institute Inc; 2011.

8. Baumgarten M, Margolis DJ, Localio AR, et al. Extrinsic risk factors for pressure ulcers early in the hospital stay: a nested case-control study. *J Gerontol A Biol Sci Med Sci.* 2008;63(4):408-413.

9. Cox J. Predictors of pressure ulcers in adult critical care patients. *Am J Crit Care.* 2011;20(5):364-375.

10. Welsh AH, Zhou XH. Estimating the retransformed mean in a heteroscedastic two-part model. *Journal of Statistical Planning and Inference.* 2006;136(3):860-881.

11. Nelson R. *An Introduction to Copulas, 2nd Edition*. New York, NY: Springer-Verlag; 2006.

12. SAS/ETS User's Guide, ver 9.3. Cary, NC: SAS Institute Inc; 2011.

13. Hougaard P. *Analysis of Multivariate Survival Data*. New York: Springer-Verlag; 2000.

14. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data, 2nd Edition*. New York: Springer-Verlag; 2003.

15. Allison PD. *Survival Analysis using the SAS System--A Practical Guide.Second Edition*. Cary, NC: SAS Institute, Inc; 2010.

16. Tang XQ, Luo ZH, Gardiner JC. Modeling hospital length of stay by Coxian phase-type regression with heterogeneity. *Statistics in Medicine.* 2012;31(14):1502-1516.

17. Gardiner JC. Modeling heavy-tailed distributions in healthcare utilization by parametric and Bayesian methods. Paper 418-2012. Paper presented at: SAS Global Forum, 2012; Orlando, FL.

18. Kole E, Koedijk K, Verbeek M. Selecting copulas for risk management. *Journal of Banking & Finance.* 2007;31(8):2405-2423.

19. Vuong QH. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica.* 1989;57(2):307-333.

20. Maddala GS. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press; 1983.

21. Trivedi PK, Zimmer DM. Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics, Vol 1*. Hanover, MA: NOW Publishers Inc; 2007.

22. de Leon AR, Wu B. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine.* 2011;30(2):175-185.

23. Nikoloulopoulos AK, Karlis D. Multivariate logit copula model with an application to dental data. *Statistics in Medicine.* 2008;27(30):6393-6406.

24. Zhao XB, Zhou X. Applying copula models to individual claim loss reserving methods. *Insurance Mathematics & Economics.* 2010 2010;46(2):290-299.

25. Diggle PJ, Sousa I, Chetwynd AG. Joint modelling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture. *Statistics in Medicine.* 2008;27(16):2981-2998.

26. Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine.* 1997;16(1-3):239-257.

27. Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O. *SAS for Mixed Models Analyses, 2nd Edition.* Cary, NC.: SAS Institute Inc; 2006.

28. Lakhal-Chaieb ML. Copula inference under censoring. *Biometrika.* 2010;97(2):505-512.

29. Cook RJ, Lawless JF, Lee KA. A copula-based mixed Poisson model for bivariate recurrent events under event-dependent censoring. *Statistics in Medicine.* 2010;29(6):694-707.

30. Chen XH, Koenker R, Xiao ZJ. Copula-based nonlinear quantile autoregression. *Econometrics Journal.* 2009;12:S50-S67.

ACKNOWLEDGEMENTS

This research was supported in part by the Agency for Healthcare Research & Quality under Grant 1R01 HS14206. I wish to thank Drs Zhehui Luo at Michigan State University, Jan Chvosta and Hao Chen at the SAS Institute for helpful comments in the preparation of this article.

CONTACT INFORMATION

Joseph C. Gardiner
Division of Biostatistics
Department of Epidemiology and Biostatistics
Michigan State University
East Lansing, MI 48824
jgardiner@epi.msu.edu