반복측정 자료의 분석 고찰

김 호

서울대학교 보건대학원

* 반복측정자료(repeated measures data)
하나의 관측단위로부터 두 번 이상의 측정을 통하여 얻어진 자료

ex 1) 신생아들의 체중을 한 달 간격으로 생후 일년간 측정, 기록
ex 2) 두 처방에서의 결과(예: 체중, 혈압 혹은 심장박동수 등)를 석 달마다 일년간 에 걸쳐 관측 기록

* 두 가지의 중요한 요인
1) 시간 : 환자내 효과 (within-subjects factors)
2) 처리 :환자간 효과 (between-subjects factors)

* 두 가지 관심 사항은
1) 각 처리의 평균값이 시간에 따라 어떻게 변하는지⇒시간의 주효과
2)처리의 효과(처리간의 차이)가 시간에 따라 어떻게 변하는지→시간과 처리의 교호작용

1) 반응변수가 연속인 경우 (예: 혈압, 체중, 콜레스테롤 level 등) -> 정규분포 사용 -> SAS Proc Mixed

2) 반응변수가 이산형 경우 (예: 유병여부 등) -> GEE -> SAS Proc Genmod

* 공분산구조 :
반복자료 분석이 일반적인 통계분석과 가장 다른 점
- 일반적인 통계 선형모형에서는 각 관측치의 오차항은 독립적이라는 가정이 필수적
- 반복측정 자료분석에서는
  + 각 각의 환자들은 독립적
  + 한 환자 안에서의 측정치, 즉 같은 환자의 다른 시점들에서 관측된 값들의 오차항 사이에는 상관관계가 존재한다고 가정
  + 한 환자 안의 관찰치들 간의 상관관계는 관측 시점의 간격에 따라 다르게 가정되는 것이 보통

  + 반복측정 자료 분석의 주요 목적은 시간에 따른 처리효과의 비교이지만 모형의 구축 단계에서는 이 상관관계 구조의 설정에 가장 많은 노력을 기울임. 상관관계구조의 올바른 선택은 반복측정 자료 분석에서 가장 중요한 과정.

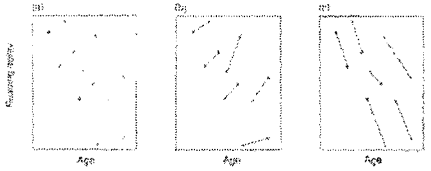
EXAMPLE : BP Data

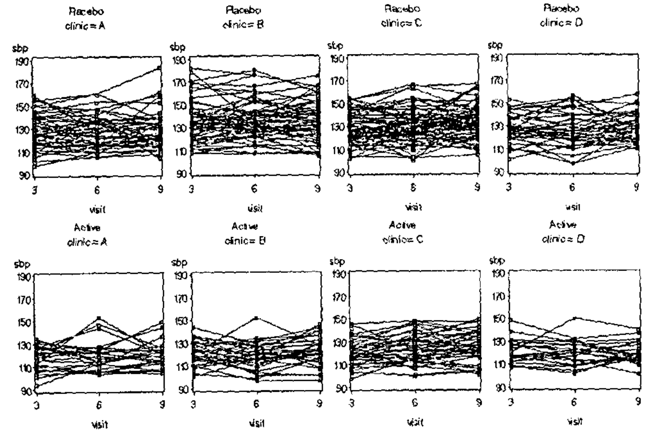
Treatment of Mild Hypertension Trial (TOMHS) by Neaton, et. al. (1993)

경미한 고혈압 환자의 치료를 위한 확률화(randomized), 양쪽-맹검법(double-blind), 비교-처치군(placebo-controlled)을 이용한 임상

연구의 주목 : 비교군에 비하여 처치군에서 약효를 확인할 수 있는가 하는 것이다.

〈그림 2〉 반복자료 분석의 시작 (그림 그리기)





* 결측치에 대한 문제

-연구 초기에는 대부분의 환자에게서 자료,
-연구가 진행될수록 제외(lost to follow-up or droop-out)되는 환자가 증가
-BP 데이터 3번의 follow-up 중 하나 이상에서 결측치 있는 환자는 13%.

PROC GLM	PROC MIXED
	id,y1,x1,x2,x3
id,y1,y2,y3,x1,x2,x3	id,y2,x1,x2,x3
	id,y3,x1,x2,x3
	1,10,1,2,3
	1,12,1,2,3
1,10,12,13,1,2,3	1,131,2,3
2,11,13, . . ,2,1,4	2,11,2,1,4
	2,13,2,1,4
	2, . ,2,1,4

〈표2〉 PROC GLM과 PROC MIXED에서의 입력자료의 비교

데이터를 분석의 기본 목적 :
-다른 모든 설명변수의 효과를 제어한 후의 SBP에 대한 TRT의 효과의 유의성을 보는 것
-동시에 같은 환자들의 관측치가 가지는 가능한 상관관계를 제어

SBP 값들은 다변량 정규분포를 따른다는 가정
-평균은 설명변수들의 선형모형으로 표현
-비교적 간단한 구조의 분산-공분산

* 분산-공분산행렬 (*variance-covariance matrix*)

-단변량 경우 $Var(\varepsilon)=\sigma^2$

-다변량 (삼변량) 경우

$$Var(\underline{\varepsilon})=Var\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\end{pmatrix}=\begin{pmatrix}var(\varepsilon_1) & cov(\varepsilon_1,\varepsilon_2) & cov(\varepsilon_1,\varepsilon_3)\\ cov(\varepsilon_2,\varepsilon_1) & var(\varepsilon_2) & cov(\varepsilon_2,\varepsilon_3)\\ cov(\varepsilon_3,\varepsilon_1) & cov(\varepsilon_3,\varepsilon_2) & var(\varepsilon_3)\end{pmatrix}=\begin{pmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{pmatrix}$$

-혼합대칭(compound symmetry) 공분산 모형

$$Var(\underline{\varepsilon})=Var\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\end{pmatrix}=\begin{pmatrix}\sigma_1+\sigma^2 & \sigma_1 & \sigma_1\\ \sigma_1 & \sigma_1+\sigma^2 & \sigma_1\\ \sigma_1 & \sigma_1 & \sigma_1+\sigma^2\end{pmatrix}$$

-일반선형모형 (독립인 오차항)

$$Var(\underline{\varepsilon})=Var\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\end{pmatrix}=\begin{pmatrix}\sigma^2 & 0 & 0\\ 0 & \sigma^2 & 0\\ 0 & 0 & \sigma^2\end{pmatrix}$$

* 고정효과(fixed effect)와 임의효과 (random effect)

1) 고정효과

-전통적인 선형모형에서 고려하는 효과

-하나의 고정된 값 (모수)이고 우리는 적절한 통계적 방법을 통하여 이들을 추정

-연구자가 관심이 있는 효과의 수준이 일정하고 모든 수준에서 자료를 수집했다면

2) 임의효과

-고정된 값이 아니고 분포를 가진 효과로서

-전체 분포에서 하나의 실현된 경우를 데이터를 통하여 관측

-효과의 분산을 추정

-효과의 수준이 다양하고 자료에서 관측하는 효과들은 전체 효과의 일부분으로 가정

고정효과의 예) 특정 처방에 대한 효과

- 그 증상에 대한 처방이 다양하게 존재하지 않고
- 연구자가 수집한 방법에만 관심이 있다면 (예를 들어 두 가지 방법을 비교한다고 할 때 이 두 가지 처방이 여러 가지 가능한 처방 방법에서의 일부 샘플이라고 생각하지 않고 연구자가 이 두 가지 처방의 비교에만 관심이 있고 연구의 결과를 이 두 가지 처방에 대한 비교로만 국한 시켜도 무방)

임의효과의 예) 환자의 효과 :

-다양하게 존재하고

-연구자가 모든 환자들의 효과를 수집할 수도 없고

-자료에서의 환자효과는 전체 연구집단에서의 임의 추출(random sampling)된 것이라고 가정해도 무방

* 두 가지 분석 방법

1)최초 관측치(baseline observation)를 설명변수로 하고 나머지 관측치들 종속변수

2)최초관측치와 각 시점에서의 관측치의 차이를 종속변수로

결국 동등한 모형으로

해석하고자 하는 방향에 따라 선택할

여기에서는 전자의 방법을 사용

SBP의 최초방문 시에 비한 SBP의 증가분이 아닌 최초방문 SBP 값으로 보정한 후의 각각 방문 시점에서의 SBP 값에 관심이 있게된다.

<모형 1> 일반 선형 모형 (General Linear Model)

서로 독립이고 동일한 분포를 가지는(iid) 오차항 가정

```
proc mixed data=bp;
   class trt visit complier clinic stratum;
   model sbp = sbpbl trt
      visit trt*visit
      complier trt*complier
      clinic trt*clinic
      stratum trt*stratum;
run;
```

PROC MIXED는 혼합효과 모형을 실행하게 하고 DATA=BP는 사용할 데이터의 이름을 나타낸다. CLASS문에서는 비연속 변수를 설정해 주고 MODEL문에서 고려되는 설명변수들을 설정해주고 있다. # cov parameter = 1

<모형 2> 혼합대칭 공분산 (Compound Symmetry) 모형

```
proc mixed data=bp;
   class trt visit complier clinic stratum person;
   model sbp = sbpbl trt
      visit trt*visit
      complier trt*complier
      clinic trt*clinic
      stratum trt*stratum;
   repeated visit / type=cs sub=person;
run;
```

TYPE=CS 대각선은 공통의 값을 가지고 비대각의 값은 또 다른 값을 가지는 형태, 반복자료분석의 가장 간단한 형태, # parameter = 2

<모형 2>와 동등한 임의 절편모형

```
proc mixed data=bp;
   class trt visit complier clinic stratum person;
   model sbp = sbpbl trt
      visit trt*visit
      complier trt*complier
      clinic trt*clinic
      stratum trt*stratum;
   random intercept /
sub=person;
run;
```

parameter = 2

<모형 3> 임의계수(Random Coefficients) 모형

```
proc mixed data=bp;
   class trt visit complier clinic stratum person;
   model sbp = sbpbl trt
      visit trt*visit
      complier trt*complier
      clinic trt*clinic
      stratum trt*stratum /
ddfm=bw;
   random int visitlin /
type=un sub=person;
run;
```

parameter = 4

<모형 4> 비구조화(Unstructured) 분산 모형

```
proc mixed data=bp;
   class trt visit complier clinic stratum person;
   model sbp = sbpbl trt
      visit trt*visit
      complier trt*complier
      clinic trt*clinic
      stratum trt*stratum;
   repeated visit / type=un sub=person r rcorr;
run;


# parameter = 6 (분산 3 개, 공분산 3 개)
```

<모형 5> 이질성 (heterogeneous) 일반 선형 모형

두 군간에 분산이 다른 것을 모형화

```
proc mixed data=bp;
   class trt visit complier clinic stratum;
   model sbp = sbpbl trt
      visit trt*visit
      complier trt*complier
      clinic trt*clinic
      stratum trt*stratum;
   repeated / sub=person group=trt;
run;

# parameter = 2
```

<모형 6> 이질성 혼합대칭 모형

```
proc mixed data=bp;
   class trt visit complier clinic stratum person;
   model sbp = sbpbl trt
      visit trt*visit
      complier trt*complier
      clinic trt*clinic
      stratum trt*stratum;
   repeated visit / type=cs sub=person group=trt;
run;

# parameters = 4 , 비교군에서 2 개, 처리군에서 2 개,
```

<모형 7> 이질성 임의계수 모형

```
proc mixed data=bp;
   class trt visit complier clinic stratum person;
   model sbp = sbpbl trt
      visit trt*visit
      complier trt*complier
      clinic trt*clinic
      stratum trt*stratum / ddfm=bw;
   random int visitlin / type=un sub=person group=trt;
   repeated / sub=person group=trt;
run;

# parameters = 8  각각의 처치군에서 4 개씩 8 개
```

<모형 8> 이질성 비구조화 분산 모형

```
proc mixed data=bp;
   class trt visit complier clinic stratum person;
   model sbp = sbpbl trt
      visit trt*visit
      complier trt*complier
      clinic trt*clinic
      stratum trt*stratum;
   repeated visit / type=un sub=person group=trt;
run;

# parameters = 12  각각의 처치군에서 6 개씩 12 개
```

<표 3>BP 데이터를 분석하기 위한 모형들

번호	모형	분산-공분산 행렬의 모수 개수	BIC
1	일반 선형화	1	-4033.5
2	혼합대칭	2	-3957.6
3	임의 계수	4	-3962.2
4	비구조화	6	-3968.2
5	이질성 일반 선형화	2	-4027.1
6	이질성 혼합대칭	4	-3957.1
7	이질성 임의계수	8	-3968.3
8	이질성 비구조화	12	-3980.5

모형을 적합시킨 후에는 잔차 검사한다. (잔차분석)

```
proc mixed data=bp;
   class trt visit complier clinic stratum person;
   model sbp = sbpbl trt
      visit trt*visit
      complier trt*complier
      clinic trt*clinic
      stratum trt*stratum / p;
   repeated visit / type=cs sub=person group=trt;
   make 'predicted' out=p noprint;
   id trt visit clinic person;
run;
```

```
     Covariance Parameter Estimates

Cov Parm     Subject    Group    Estimate


Variance     person     trt 1     69.0629   = $\hat{\sigma}^2$
CS           person     trt 1     34.9670   = $\hat{\sigma}_1$
Variance     person     trt 2     87.2927   = $\hat{\sigma}^2$
CS           person     trt 2     71.7782   = $\hat{\sigma}_1$

Fitting information

Res Log Likelihood                   -3943.3
Akaike's Information Criterion        -3947.3
Schwarz's Bayesian Criterion          -3955.0
```

```
proc mixed data=bp;
   class trt visit complier clinic stratum person;
   model sbp = sbpbl trt visit clinic trt*clinic/s;
   repeated visit / type=cs sub=person group=trt;
   lsmeans trt*clinic / cl;
run;
```

```
Type 3 Tests of Fixed Effects

             Num     Den
Effect       DF      DF     F Value    Pr > F

sbpbl         1      345     194.84    <.0001
trt           1      345      21.38    <.0001
visit         2      682       4.90    0.0077
trt*visit     2      682       0.45    0.6393
complier      1      345       0.47    0.4914
trt*complier  1      345       0.05    0.8263
clinic        3      345       6.25    0.0004
trt*clinic    3      345       3.76    0.0112
stratum       1      345       0.06    0.8021
trt*stratum   1      345       1.86    0.1736
```

```
          Least Squares Means
                              Standard
Effect      trt clinic Estimate  Error    DF  t Value  Pr>|t|
trt*clinic   1  A      121.30   1.4817   349   81.87   <.0001
trt*clinic   1  B      123.21   1.3044   349   94.46   <.0001
trt*clinic   1  C      122.46   1.1529   349  106.21   <.0001
trt*clinic   1  D      122.38   1.6317   349   75.00   <.0001
trt*clinic   2  A      126.79   1.4116   349   89.82   <.0001
trt*clinic   2  B      136.02   1.2604   349  107.92   <.0001
trt*clinic   2  C      129.45   1.1777   349  109.92   <.0001
trt*clinic   2  D      127.90   1.5799   349   80.95   <.0001
```

* TRT (CLINIC)

CLINIC=A 126.79-121.30=5.49,
B 136.02-123.21=12.81,
C 129.45-122.46=6.99,
D 127.90-122.38=5.52

```
lsmeans trt / diff cl e;
lsmeans trt / at sbpbl=170 cl;
```

```
   Least Squares Means

                 Standard
Eff trt sbpbl Est   Error  DF   t Value  P> |t|
trt  1   .   122.34 0.7051 349  173.51  <.0001
trt  2   .   130.04 0.6834 349  190.29  <.0001
trt  1  170  138.64 1.4373 349   96.46  <.0001
trt  2  170  146.35 1.3221 349  110.69  <.0001

     Differences of Least Squares Means
                      Standard
Eff trt _trt  sbpbl Est Error    DF    t Val

trt  1   2  .  -7.7021  0.9848   349   -7.82
```

Summary

- * 반복측정 자료 분석은 한 개체 안에서 반복적으로 측정된 자료간에는 상관이 있다는 사실에 부합하는 분석 방법이다.
- * 독립성을 가정한 일반 선형모형(일반 회귀분석이나 분산분석)을 사용하는 것은 가정에 어긋나는 모형을 사용하는 것이다.
- * 연구의 목적이 상관관계의 분석이 아닌 약효의 차이에 대한 것이라도 반복자료 분석 기법을 이용해야만 상관성을 보정한 후의 효과를 얻을 수 있다.
- * 올바른 약효를 추정하기 위해서는 올바른 공분산 구조를 선택해야한다.
- * 올바른 약효를 추정하기 위해서는 교호작용의 선택 등 변수 선택에 유의해야한다.

generalized linear model

1) linear component (predictor) : $\eta_i = x_i'\beta$

2) expected value : link function $g(\mu_i) = g(E(Y)) = x_i'\beta = \eta_i$

3) exponential family : $Var(Y_i) = \frac{\phi V(\mu_i)}{w_i}$

ϕ : dispersion parameter, $Var(\mu_i)$: variance function w_i : weight

Examples of generalized linear models

Traditional linear model	response var : conti.
dist'n : normal	link : identity, $\eta = \mu$
Logistic regression	response var : proportion
dist'n : binomial	link : logit, $\eta = \log\left(\frac{\mu}{1-\mu}\right)$

What is Generalized_Linear Model ?

traditional linear model : $y_i = x_i'\beta + \epsilon_i$

Expected value $\mu_i = x_i'\beta$

Extension

1) dist'n other than normal

2) restriction on the ranges

3) variance could not be a constant (function of mean)

Poisson regression	response var : count
dist'n : Poisson	link : log, $\eta = \log(\mu)$

Gamman model with log link

response var : positive conti var   dist'n : gamma

link : log, $\eta = \log(\mu)$

Popular link functions

identity : $\eta = \mu$

logit : $\eta = \log\left(\frac{\mu}{1-\mu}\right)$

probit : $\eta = \Phi^{-1}(\mu)$, $\Phi^{-1}(\cdot)$:cdf of $N(0,1)$

power : $\eta = \begin{cases} \mu^{\lambda} & \text{if } \lambda \neq 0 \\ \log(\mu) & \text{if } \lambda = 0 \end{cases}$

log : $\eta = \log(\mu)$

complementary log-log : $\eta = \log(-\log(1-\mu))$

Dist'ns and variance functions

normal : $V(\mu) = 1$

binomial : $V(\mu) = \mu(1-\mu)$

Poisson : $V(\mu) = \mu$

gamma : $V(\mu) = \mu^2$

inverse Gaussian : $V(\mu) = \mu^3$

Poisson Regression Example

(Insurance claims data)

```
data insure;
input n c car$ age;
ln = log(n);
datalines;
   500   42  small  1
   1200  37  medium 1
   100    1  large  1
   400  101  small  2
   500   73  medium 2
   300   14  large  2
;
```

$$\log(\mu_i) = \log(N_i) + \beta_0 + CAR_i(1)\beta_1 + CAR_i(2)\beta_2 + CAR_i(3)\beta_3 + AGE_i(1)\beta_4 + AGE_i(2)\beta_5$$

where $CAR_i(j)$ and $AGE_i(j)$ are indicator variables.

$$\log\left(\frac{\mu_i}{N_i}\right) = \beta_0 + CAR_i(1)\beta_1 + CAR_i(2)\beta_2 + CAR_i(3)\beta_3 + AGE_i(1)\beta_4 + AGE_i(2)\beta_5$$

```
proc genmod data=insure;
     class car age;
     model c = car age / dist   = poisson
                         link   = log
                         offset = ln;   run;

                 The GENMOD Procedure

                   Model Information

          Data Set                WORK.INSURE
          Distribution                Poisson
          Link Function                   Log
          Dependent Variable                c
          Offset Variable                  ln
          Observations Used                 6

                Class Level Information

        Class      Levels    Values

        car             3    large medium small
        age                2    1 2

            Criteria For Assessing Goodness Of Fit

    Criterion              DF          Value       Value/DF

    Deviance                2          2.8207       1.4103
    Scaled Deviance         2          2.8207       1.4103
    Pearson Chi-Square      2          2.8416       1.4208
    Scaled Pearson X2       2          2.8416       1.4208
    Log Likelihood                   837.4533
```

P-VALUE of 2.8207 = .24 (from chi-sq df=2)

```
Algorithm converged.

                    The SAS System                      4


                   The GENMOD Procedure

              Analysis Of Parameter Estimates

                                        Wald 95%
                              Standard   Confidence Limits      Chi-
```

```
Parameter          DF  Estimate     Error     Lower     Upper   Square

Intercept           1  -1.3168    0.0903   -1.4937   -1.1398   212.73
car       large     1  -1.7643    0.2724   -2.2981   -1.2304    41.96
car       medium    1  -0.6928    0.1282   -0.9441   -0.4414    29.18
car       small     0   0.0000    0.0000    0.0000    0.0000      .
age        1        1  -1.3199    0.1359   -1.5863   -1.0536    94.34
age        2        0   0.0000    0.0000    0.0000    0.0000      .
Scale               0   1.0000    0.0000    1.0000    1.0000

                  Analysis Of Parameter
                        Estimates

               Parameter            Pr > ChiSq

               Intercept              <.0001
               car        large       <.0001
               car        medium      <.0001
               car        small         .
               age         1          <.0001
               age         2            .
               Scale

NOTE: The scale parameter was held fixed.
```

expected values

```
   n  car     age              expected value                       obs
 500  small    1   exp(log(500) - 1.3168             - 1.3199) = 36  42
1200  medium   1   exp(log(1200) - 1.3168 - 0.6928 - 1.3199) = 43  37
 100  large    1   exp(log(100) - 1.3168 - 1.7643 - 1.3199) = 1    1
 400  small    2   exp(log(400) - 1.3168                   ) = 107 101
 500  medium   2   exp(log(500) - 1.3168 - 0.6928          ) = 67  73
 300  large    2   exp(log(300) - 1.3168 - 1.7643          ) = 14  14
```

Logistic Regression Approach

```
proc genmod data=insure;
     class car age;
     model c/n = car age / dist  = binomial    ;   run;

                    Model Information

          Data Set                      WORK.INSURE
          Distribution                   Binomial
          Link Function                     Logit
          Response Variable (Events)           c
```

```
        Response Variable (Trials)              n
        Observations Used                       6
        Number Of Events                      266
        Number Of Trials                     3000


              Criteria For Assessing Goodness Of Fit

    Criterion              DF         Value       Value/DF

    Deviance                2         1.7221       0.8611
    Scaled Deviance         2         1.7221       0.8611
    Pearson Chi-Square      2         1.7283       0.8641
    Scaled Pearson X2       2         1.7283       0.8641
    Log Likelihood                 -606.2836

Algorithm converged.

                   The GENMOD Procedure

              Analysis Of Parameter Estimates

                                          Wald 95%
```

Parameter		DF	Estimate	Standard Error	Confidence Limits Lower	Upper	Chi-Square
Intercept		1	−1.0256	0.1042	−1.2297	−0.8214	96.96
car	large	1	−1.9978	0.2824	−2.5512	−1.4443	50.05
car	medium	1	−0.8148	0.1385	−1.0862	−0.5434	34.63
car	small	0	0.0000	0.0000	0.0000	0.0000	.
age	1	1	−1.4890	0.1431	−1.7695	−1.2085	108.25
age	2	0	0.0000	0.0000	0.0000	0.0000	.
Scale		0	1.0000	0.0000	1.0000	1.0000	

Generalized Estimating Equations

The non-independence of observations for a given subject can be characterized in terms of a correlation matrix for each subject

For the response vector for the i-th subject

$Y_i' = (Y_{i1}, Y_{i2}, \cdots, Y_{in_i})$

the correlation between the j-th and j'-th response is given by

$Corr(Y_{ij}, Y_{ij'}) = \rho_{ijj'}$

and the correlation matrix for the i-th subject is a $(n_i \times n_i)$ matrix, e.g. for $n_i = 3$, we have

$$R_i = \begin{vmatrix} 1 & \rho_{i12} & \rho_{i13} \\ \rho_{i12} & 1 & \rho_{i23} \\ \rho_{i13} & \rho_{i23} & 1 \end{vmatrix}$$

Working correlation structures

For the i-th subject, let $R_i(\alpha)$ be a $(n_i \times n_i)$ matrix.

We allow the (size of) correlation matrices for different subjects to vary, but the general structure, defined by α is the same for all subjects.

Correlation structures

Independence, Exchangeable, Stationary,
Autoregressive, Arbitrary

Exchangeable : all correlations are the same

$Corr(Y_{ij}, Y_{ij'}) = \rho$ for all j, j' where $i \neq j$.

So α is ρ

Stationary m-dependent

$$\begin{aligned} &corr(Y_{ij}, Y_{ij+1}) = \rho_1 \\ &corr(Y_{ij}, Y_{ij+2}) = \rho_2 \\ &\cdots \qquad \cdots \\ &corr(Y_{ij}, Y_{ij+m}) = \rho_m \\ &corr(Y_{ij}, Y_{ij+m'}) = 0, \quad \text{if } m' > m \end{aligned}$$

Example 1. Stationary 1-dependent

$$n_i = 4,\quad R_i(\rho_1) = \begin{bmatrix} 1 & \rho_1 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & 1 \\ 0 & \rho_1 & 1 & \rho_1 \\ 0 & 0 & \rho_1 & 1 \end{bmatrix}$$

Example 2. Stationary 2-dependent

$$n_i = 5,\quad R_i(\rho_1, \rho_2) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & \rho_2 & 0 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ 0 & \rho_2 & \rho_1 & 1 & \rho_1 \\ 0 & 0 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

Autoregressive (AR-1)

$$Corr(Y_{ij}, Y_{ij'}) = \rho^{|t_{ij} - t_{ij'}|}$$

Example.

$$n_i = 5,\quad R_i(\rho) = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Arbitrary correlation

no restriction on $R(a)$ so n(n+1)/2 elements

$$R_i(a) = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$$

SAS options

AR|AR(1) : Autoregressive(1)
EXCH|CS : exchangeable
IND : independent
MDEP : m-dependent
UNSTR|UN : unrestricted (arbitrary)
USER|FIXED : fixed, user-specified correlation matrix

TYPE=user(1.0 .9 .8 .6

.9 1.0 .9 .8
.8 .9 1.0 .9
.6 .8 .9 1.0)

The Variance formula

$$Var(Y_i) = \Sigma_i = V_i^{1/2} R_i V_i^{1/2}$$

where

$$\Sigma_i = \begin{vmatrix} Var(Y_{i1}) & Cov(Y_{i1}, Y_{i2}) & \cdots & Cov(Y_{i1}, Y_{in_i}) \\ Cov(Y_{i1}, Y_{i2}) & Var(Y_{i2}) & \cdots & Cov(Y_{i2}, Y_{in_i}) \\ \cdots & \cdots & \cdots & \cdots \\ Cov(Y_{i1}, Y_{in_i}) & Cov(Y_{i2}, Y_{in_i}) & \cdots & Var(Y_{in_i}) \end{vmatrix}$$

and

$$R_i = \begin{vmatrix} 1 & Corr(Y_{i1}, Y_{i2}) & \cdots & Corr(Y_{i1}, Y_{in_i}) \\ Corr(Y_{i1}, Y_{i2}) & 1 & \cdots & Corr(Y_{i2}, Y_{in_i}) \\ \cdots & \cdots & \cdots & \cdots \\ Corr(Y_{i1}, Y_{in_i}) & CorrY_{i2}, Y_{in}) & \cdots & 1 \end{vmatrix}$$

$$V_i^{1/2} = \begin{vmatrix} \sqrt{\phi V(\mu_{i1})} & & 0 \\ & \sqrt{\phi V(\mu_{i2})} & \\ & 0 & \sqrt{\phi V(Y_{in_i})} \end{vmatrix}$$

The Generalized Estimating Equations (GEE's)

$$\sum_{i=1}^{K} [\frac{-a\mu_i'}{d\beta}] \{\Sigma_i(\alpha)\}^{-1} [Y_i - \mu_i] = 0$$
$$(p \times n_i) \quad (n_i \times n_i) \quad (n_i \times 1) \quad (p \times 1)$$

Note that the above GEE's can be expressed as a function of β alone by first replacing α by a consistent estimator $\hat{\alpha}(Y, \beta, \phi)$ and then replacing ϕ in $\hat{\alpha}$ by a consistent estimator $\hat{\phi}(Y, \beta)$. Note in particular, that μ_i can be expressed in terms of β using the inverse link function and that $Var(Y_{ij}) = \phi V(\mu_{ij})$.

Variance Estimation in the GEE Approach

Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2 \cdots, \hat{\beta}_p)'$ denote the estimated regression coefficients obtained by solving the GEE's for β. Also $\hat{\alpha}$ and $\hat{\phi}$ denote the estimates of α and ϕ.

Then a **robust estimator** of the variance- covariance matrix for $\hat{\beta}$ is given by the following matrix formula

$$\widehat{Var}(\hat{\beta}) = \hat{M}_0^{-1} \hat{M}_1 \hat{M}_0^{-1}$$

: sandwich estimator

where $\hat{M}_0 = \sum_{i=1}^{K} [\frac{-d\widehat{\mu}}{d\beta}] [\Sigma_i(\alpha)]^{-1} [\frac{-d\widehat{\mu}}{d\beta}]$

$$\hat{M}_1 = \sum_{i=1}^{K} [\frac{-d\widehat{\mu}}{d\beta}] [\Sigma_i(\alpha)]^{-1} (y_i - \hat{\mu}_i)$$
$$(y_i - \hat{\mu}_i)' [\Sigma_i(\alpha)]^{-1} [\frac{-d\widehat{\mu}}{d\beta}]$$

Issues regarding GEE-Based analyses of longitudinal data

1. Why dose a GEE analysis for a univariate GLM model give different variance estimates for the estimated regression coefficients than those obtained using a program that fits a GLM model directly? And if the variance estimates are different, which method of variance estimation is better ?

2. If no matter which working correlation matrix structure is used, the estimated regression coefficients are roughly the same, why not use an independent working correlation structure all the same ?

3. If, for a given set of data, the use of different working correlation structure yields (typically slightly) different estimated regression coefficients and/or different variances and

covariances corresponding to theses coef., how do we decide which working corr str is *most appropreate* for the data being analyzed ?

Answers to Issues

1. The estimated var-cov matrix of $\widehat{\beta}_1, \cdots \widehat{\beta}_b$ obtained using a GEE is a **robust sandwich** estimator involving an estimated covariance matrix of residuals $(y_i - \widehat{\mu}_i)(y_i - \widehat{\mu}_i)'$.

This var-cov matrix estimator is robust because it is a consistent estimate of the appropriate var-cov matrix regardless of whether the working correlation matrix used is correct or not.

If an univariate GLM is fit using a GEE analysis, the GEE var-cov matrix estimator will still be a robust estimator, *since it uses a scalar* version of the estimated covariance matrix of residuals.

If a program that fits GLM's directly is used (e.g. GLIM), the var cov matrix is not robust.

Which var-cov estimator is better ?
The **robust estimator is better** (for validity reasons), although the non-robust estimator may give better statistical efficiency if both the working corr. structure and the mean-variance relationship specified for the GEE analysis are correct (which you never *really know*).
The chief advantage of the robust variance estimator is that it provides the correct (valid) estimate of the appropriate population var-cov matrix regardless whether the working correlation structure used is correct or whether the mean-variance relationship specified is correct.

Note that both the robust and non-robust retimators can be in error if the model being fitted is incorrectly specified (due to mispecifying the link function and/or the set of predictors). Moreover, the estimated regression coef. can be in error if the model is incorrectly specified.

2. *Even if the GEE-based* estimated regression coef. are roughly the same regardless of which working correlation structure is used, it is not appropriate to use the independence working corr. structure exclusively.

The reason for considering corr. structures other than for independence is that even though the GEE estimated *regression coeff. will usually* not differ much regardless of the working corr. structure used, the estimated var-cov matrix of the GEE estimates may *differ substantially* for different working corr. structures. Such differences can lead to different statistical inference results and may therefore. necessitate choosing among different corr. str.
3. If the use of different working corr. str. gives *different estimated regression coeff.* and their estimated var and cov, one reasonable way to choose the appropriate structure requires the investigators to compare the biological or clinical relevance of alternative choices, and to choose the structure considered most relevant. Also it may be appropriate to compute simple estimates of correlation matrices and to let such estimates suggest an appropriate working corr. str. for the GEE analysis.

An alternative apprpach to this problem (of choosing an appropriate corr. str.) is to consider modeling the corr. str., as well as modeling the responses. (likelihood ratio test, AIC, or BIC)

Example : Six Cities Study of health effects on pollution.

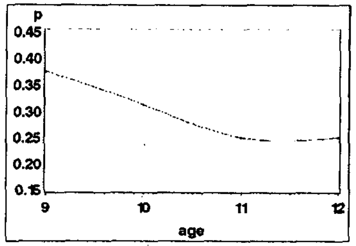
Response : Wheezing status of sixteen children at ages 9, 10, 11, and 12.

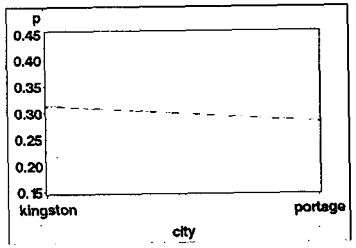
Explanatory vars : city of residence, age maternal smoking status at the particular age (time varying covariate).

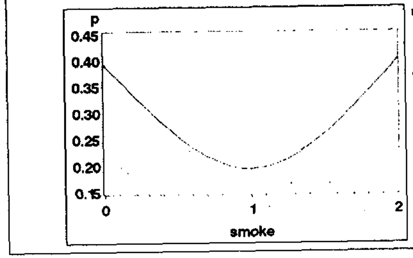
```
data six;
     input case city$ @;
     do i=1 to 4;
        input age smoke wheeze @@;
        output;
     end;
     datalines;
   1 portage   9 0 1  10 0 1  11 0 1  12 0 0
   2 kingston  9 1 1  10 2 1  11 2 0  12 2 0
   3 kingston  9 0 1  10 0 0  11 1 0  12 1 0
   4 portage   9 0 0  10 0 1  11 0 1  12 1 0
   5 kingston  9 0 0  10 1 0  11 1 0  12 1 0
   6 portage   9 0 0  10 1 0  11 1 0  12 1 0
   7 kingston  9 1 0  10 1 0  11 0 0  12 0 0
   8 portage   9 1 0  10 1 0  11 1 0  12 2 0
   9 portage   9 2 1  10 2 0  11 1 0  12 1 0
  10 kingston  9 0 0  10 0 0  11 0 0  12 1 0
```

```
   11 kingston  9 1 1  10 0 0  11 0 1  12 0 1
   12 portage   9 1 0  10 0 0  11 0 0  12 0 0
   13 kingston  9 1 0  10 0 1  11 1 1  12 1 1
   14 portage   9 1 0  10 2 0  11 1 0  12 2 1
   15 kingston  9 1 0  10 1 0  11 1 0  12 2 1
   16 portage   9 1 1  10 1 1  11 2 0  12 1 0
   ;
goptions gunit=pct border ftext=swissb htext=6 ;

%macro p(var) ;

proc sort data=six out=se;
by &var ;

proc means data=se mean;
var wheeze;
by &var;
output out=s mean=p;

proc gplot;
symbol1 i=spline v=dot;
plot p*&var/vaxis=0.15 to 0.45 by 0.05;
run;

 %mend p;
```

```
%p(age);
%p(city);
%p(smoke);
```



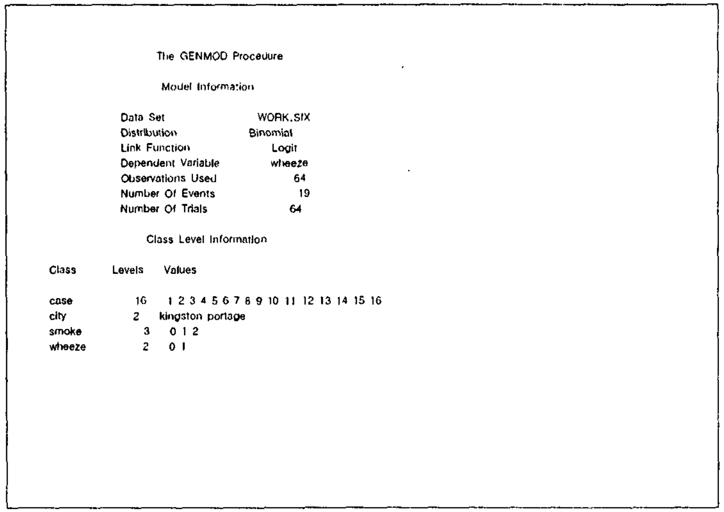




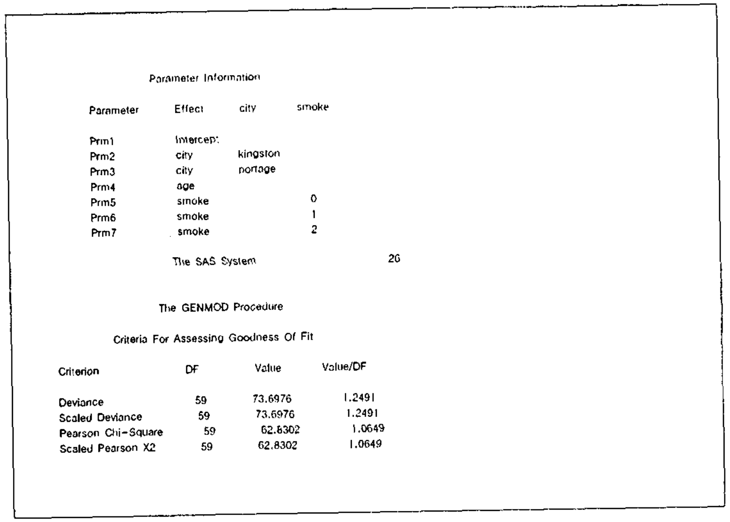
```
proc genmod data=six ;
     class case city smoke;
     model  wheeze = city age smoke  /  dist=bin;
     repeated  subject=case / type=exch covb
corrw;
   run;
```

The SAS System

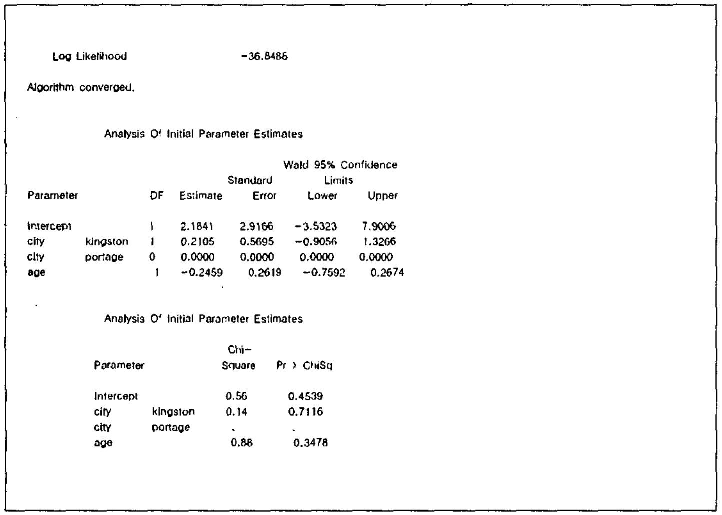
25



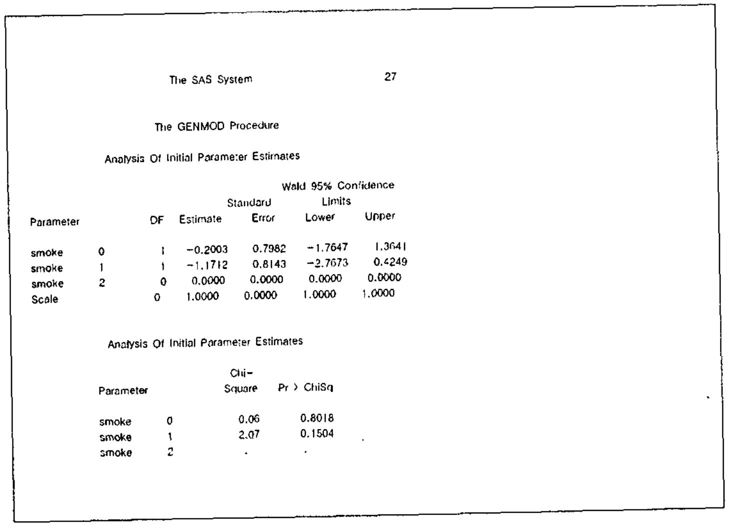
```
                   The GENMOD Procedure

                    Model Information

              Data Set                  WORK.SIX
              Distribution              Binomial
              Link Function                Logit
              Dependent Variable          wheeze
              Observations Used               64
              Number Of Events                19
              Number Of Trials                64

                 Class Level Information

Class       Levels    Values

case          16    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
city           2    kingston portage
smoke          3    0 1 2
wheeze         2    0 1
```



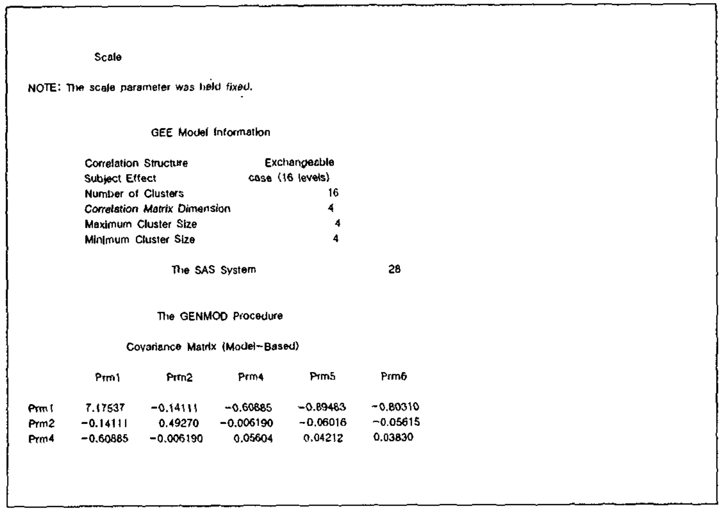
```
                  Parameter Information

      Parameter      Effect       city        smoke

      Prm1           Intercept
      Prm2           city         kingston
      Prm3           city         portage
      Prm4           age
      Prm5           smoke                      0
      Prm6           smoke                      1
      Prm7           smoke                      2

                    The SAS System                      26

                  The GENMOD Procedure

           Criteria For Assessing Goodness Of Fit

Criterion                 DF          Value        Value/DF

Deviance                  59        73.6976          1.2491
Scaled Deviance           59        73.6976          1.2491
Pearson Chi-Square        59        62.8302          1.0649
Scaled Pearson X2         59        62.8302          1.0649
```



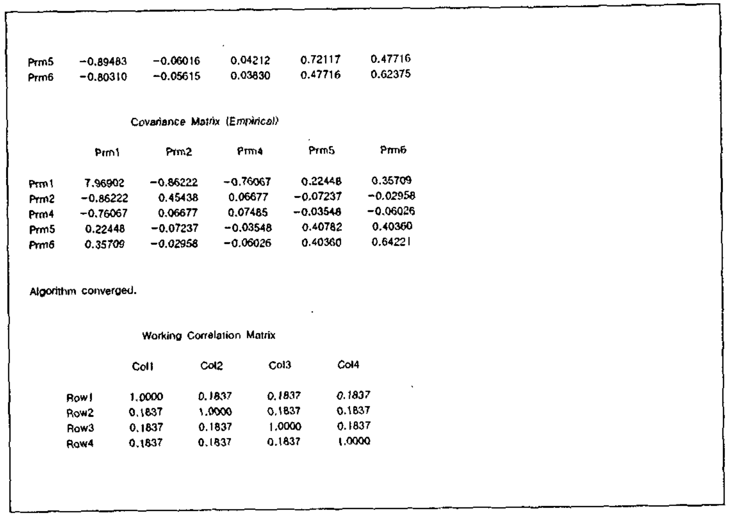
```
  Log Likelihood                      -36.8486

Algorithm converged.

           Analysis Of Initial Parameter Estimates

                                              Wald 95% Confidence
                                  Standard          Limits
Parameter            DF  Estimate    Error     Lower       Upper

Intercept             1    2.1841    2.9166   -3.5323      7.9006
city      kingston    1    0.2105    0.5695   -0.9056      1.3266
city      portage     0    0.0000    0.0000    0.0000      0.0000
age                   1   -0.2459    0.2619   -0.7592      0.2674

           Analysis Of Initial Parameter Estimates

                                Chi-
          Parameter            Square    Pr > ChiSq

          Intercept              0.56       0.4539
          city      kingston     0.14       0.7116
          city      portage       .           .
          age                    0.88       0.3478
```



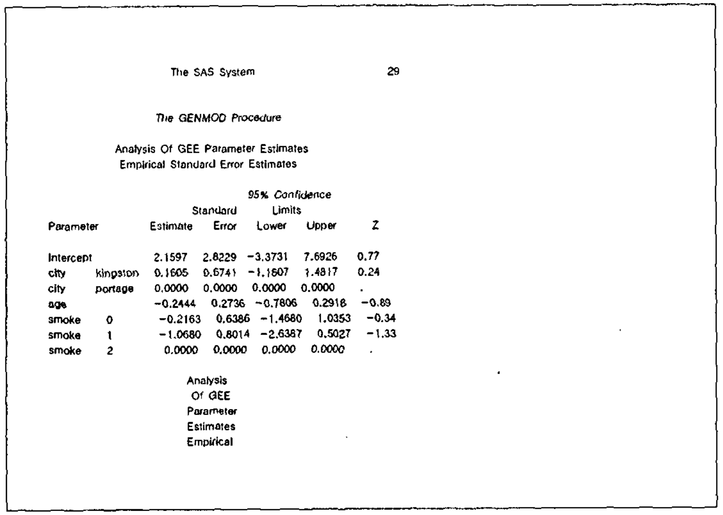
```
                  The SAS System                        27

                The GENMOD Procedure

        Analysis Of Initial Parameter Estimates

                                            Wald 95% Confidence
                                Standard          Limits
Parameter          DF  Estimate    Error     Lower       Upper

smoke      0        1   -0.2003    0.7982   -1.7647      1.3641
smoke      1        1   -1.1712    0.8143   -2.7673      0.4249
smoke      2        0    0.0000    0.0000    0.0000      0.0000
Scale               0    1.0000    0.0000    1.0000      1.0000

        Analysis Of Initial Parameter Estimates

                                Chi-
        Parameter              Square    Pr > ChiSq

        smoke      0             0.06       0.8018
        smoke      1             2.07       0.1504
        smoke      2              .           .
```



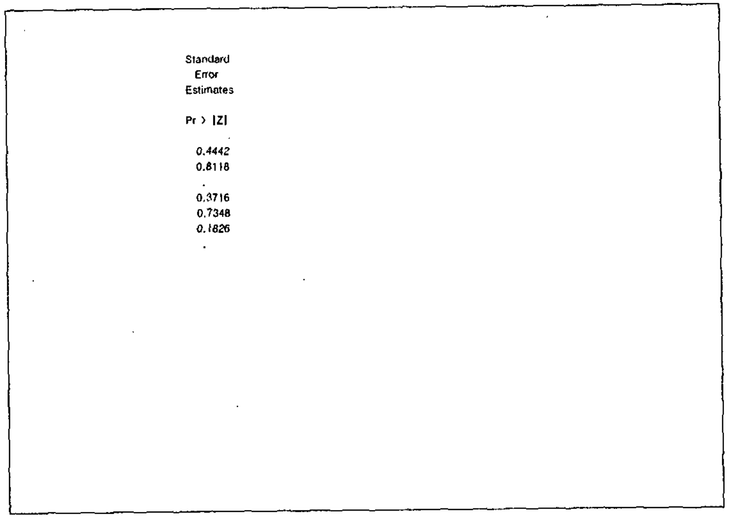
```
          Scale

NOTE: The scale parameter was held fixed.

                      GEE Model Information

          Correlation Structure              Exchangeable
          Subject Effect                  case (16 levels)
          Number of Clusters                           16
          Correlation Matrix Dimension                 4
          Maximum Cluster Size                          4
          Minimum Cluster Size                          4

                       The SAS System                        28

                     The GENMOD Procedure

                 Covariance Matrix (Model-Based)

          Prm1        Prm2        Prm4        Prm5        Prm6

Prm1     7.17537    -0.14111    -0.60885    -0.89483    -0.80310
Prm2    -0.14111     0.49270   -0.006190    -0.06016    -0.05615
Prm4    -0.60885   -0.006190     0.05604     0.04212     0.03830
```



```
Prm5    -0.89483    -0.06016     0.04212     0.72117     0.47716
Prm6    -0.80310    -0.05615     0.03830     0.47716     0.62375

                  Covariance Matrix (Empirical)

          Prm1        Prm2        Prm4        Prm5        Prm6

Prm1     7.96902    -0.86222    -0.76067     0.22448     0.35709
Prm2    -0.86222     0.45438     0.06677    -0.07237    -0.02958
Prm4    -0.76067     0.06677     0.07485    -0.03548    -0.06026
Prm5     0.22448    -0.07237    -0.03548     0.40782     0.40360
Prm6     0.35709    -0.02958    -0.06026     0.40360     0.64221

Algorithm converged.

                   Working Correlation Matrix

              Col1        Col2        Col3        Col4

    Row1     1.0000      0.1837      0.1837      0.1837
    Row2     0.1837      1.0000      0.1837      0.1837
    Row3     0.1837      0.1837      1.0000      0.1837
    Row4     0.1837      0.1837      0.1837      1.0000
```



```
                       The SAS System                        29

                     The GENMOD Procedure

                Analysis Of GEE Parameter Estimates
                 Empirical Standard Error Estimates

                                          95% Confidence
                              Standard        Limits
Parameter            Estimate    Error    Lower    Upper       Z

Intercept             2.1597    2.8229  -3.3731   7.6926    0.77
city      kingston    0.1605    0.6741  -1.1607   1.4817    0.24
city      portage     0.0000    0.0000   0.0000   0.0000     .
age                  -0.2444    0.2736  -0.7806   0.2918   -0.89
smoke        0        -0.2163   0.6386  -1.4680   1.0353   -0.34
smoke        1        -1.0680   0.8014  -2.6387   0.5027   -1.33
smoke        2         0.0000   0.0000   0.0000   0.0000     .

                        Analysis
                         Of GEE
                        Parameter
                        Estimates
                        Empirical
```



```
        Standard
         Error
        Estimates

        Pr > |Z|

         0.4442
         0.8118
           .
         0.3716
         0.7348
         0.1826
           .
```