Homework #8

Stat 545 (GLM & Categorical Data Analysis)

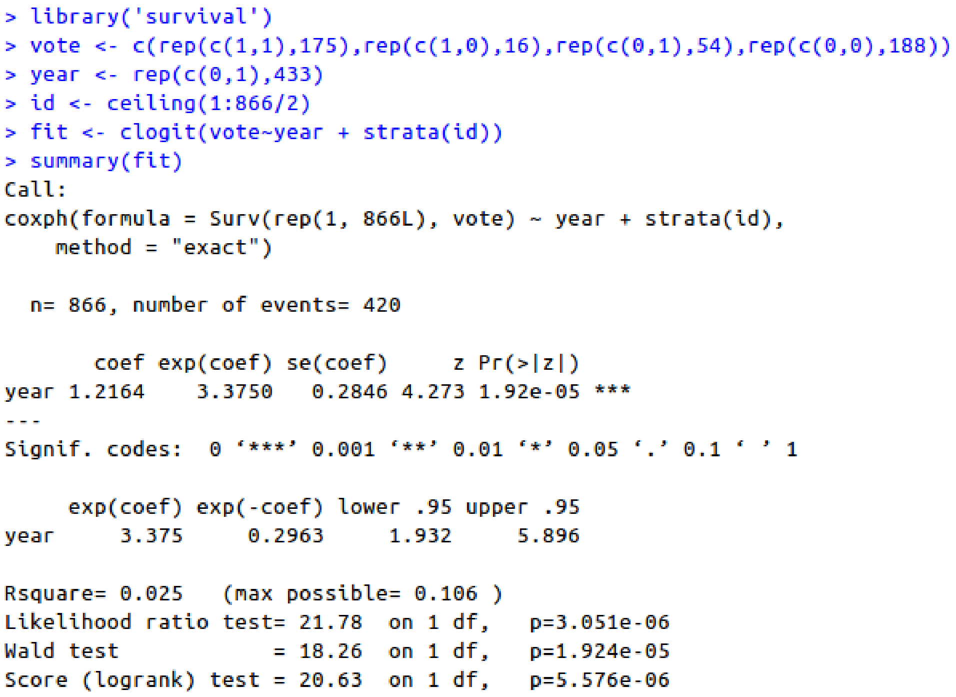
Quan Zhou

November 29, 2015

Email: qz9@rice.edu

1. **Problem 8.1**

   (a) The model is fit by R function *clogit*.



```
> library('survival')
> vote <- c(rep(c(1,1),175),rep(c(1,0),16),rep(c(0,1),54),rep(c(0,0),188))
> year <- rep(c(0,1),433)
> id <- ceiling(1:866/2)
> fit <- clogit(vote~year + strata(id))
> summary(fit)
Call:
coxph(formula = Surv(rep(1, 866L), vote) ~ year + strata(id),
    method = "exact")

  n= 866, number of events= 420

       coef exp(coef) se(coef)     z Pr(>|z|)
year 1.2164    3.3750   0.2846 4.273 1.92e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
year     3.375     0.2963     1.932     5.896

Rsquare= 0.025   (max possible= 0.106 )
Likelihood ratio test= 21.78  on 1 df,   p=3.051e-06
Wald test            = 18.26  on 1 df,   p=1.924e-05
Score (logrank) test = 20.63  on 1 df,   p=5.576e-06
```

   Hence $\hat{\beta} = 1.216$ and $\exp(\hat{\beta}) = 3.375 = 54/16$. $\beta$ describes the conditional association between the 2004 vote and 2008 vote for each fixed individual. The exponential of its MLE is equal to the ratio of the off-diagonal counts in the table. $\exp(\beta)$ may also be called the true odds ratio for each individual.

   Since R reports a p-value $1.9 \times 10^{-5}$ for $\hat{\beta}$, there is strong evidence that $\beta$ is actually bigger than 0. Voter preference for Democrats has increased.

   (b) $\alpha_i$ represents a fixed subject-specific effect for each individual. A larger $\alpha_i$ means the individual has a stronger tendency to vote for Democrats in both 2004 and 2008.

   (c) The MLE of $\beta$ for population averaged effect model is simply the log odds ratio.

$$\hat{\beta}_2 = \log \frac{229 \times 242}{191 \times 204} = 0.352$$

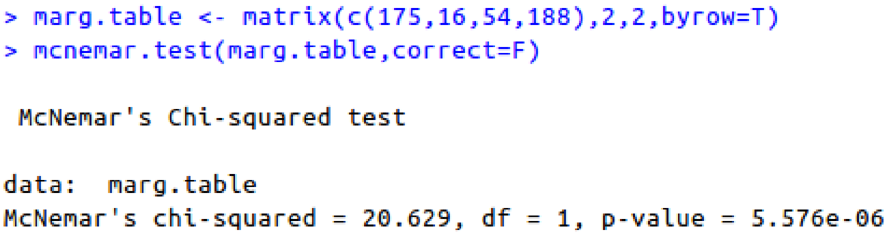
The MLE of $\beta$ for subject-specific model is given above, $\hat{\beta}_1 = 1.216$.

They are not same and $\hat{\beta}_2 < \hat{\beta}_1$. This is expected. As is illustrated in Figure 13.1 of the textbook, when two individuals have very different $\alpha_i$, their probability curves $P(Y_i = 1)$ v.s. $x_i$ (in this problem, $x_i \in [0, 1]$) are spaced far apart. The marginal model tries to fit a curve that is averaged over all the individuals' curves and thus it has a shallower slope. By the approximation formula of Zeger et al. (1988),

$$\hat{\beta}_1 \approx \hat{\beta}_2 (1 + 0.346\sigma^2)^{-1/2} \tag{1}$$

where $\sigma^2$ is the variance of $\alpha_i$ if a random effect model is assumed. In our problem, $\sigma^2$ is big because most people didn't change their side, which implies many people have a very large $\alpha_i$ while many others have a very small $\alpha_i$. Therefore, by (1), $\hat{\beta}_1$ should be much larger.

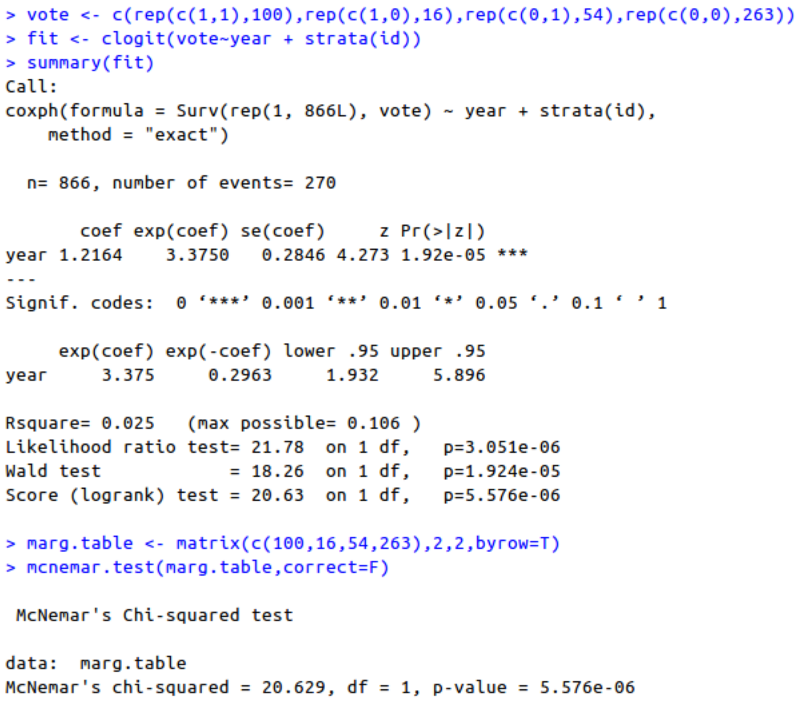
(d) The McNemar's test is done in R with no continuity correction.



```
> marg.table <- matrix(c(175,16,54,188),2,2,byrow=T)
> mcnemar.test(marg.table,correct=F)

 McNemar's Chi-squared test

data:  marg.table
McNemar's chi-squared = 20.629, df = 1, p-value = 5.576e-06
```

The p-value is similar to that of conditional logistic regression. This is expected. McNemar test is actually the score test. Asymptotically, it is equivalent to Wald test and likelihood ratio test. Therefore, the p-value of the conditional logistic regression, which is usually computed by Wald test, is often very close to the p-value of McNemar test, especially when the sample size is large.

(e) No. Neither of them would change. This is because both McNemar's test and the conditional likelihood only depend on the off-diagonal counts. For McNemar's test, recall that the p-value is evaluated under the null distribution and the diagonal counts only influence the inferences under the alternative, e.g., how much heterogeneity there exists. For conditional logistic regression, observe that the MLE of $\beta$ given the whole table is the same as the MLE given only the off-diagonal cells. This is because, for any $\beta \in \mathbb{R}$, the likelihood of the diagonal cells can always take any value in $\mathbb{R}$ by choosing the appropriate values of $\alpha_i$.

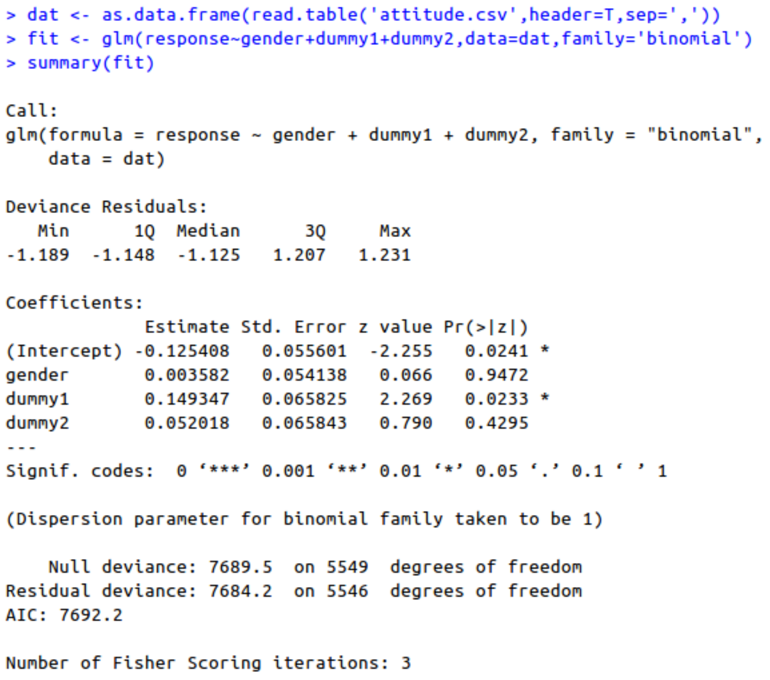


```
> vote <- c(rep(c(1,1),100),rep(c(1,0),16),rep(c(0,1),54),rep(c(0,0),263))
> fit <- clogit(vote~year + strata(id))
> summary(fit)
Call:
coxph(formula = Surv(rep(1, 866L), vote) ~ year + strata(id),
    method = "exact")

  n= 866, number of events= 270 

       coef exp(coef) se(coef)     z Pr(>|z|)    
year 1.2164    3.3750   0.2846 4.273 1.92e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
year     3.375     0.2963     1.932     5.896

Rsquare= 0.025   (max possible= 0.106 )
Likelihood ratio test= 21.78  on 1 df,   p=3.051e-06
Wald test            = 18.26  on 1 df,   p=1.924e-05
Score (logrank) test = 20.63  on 1 df,   p=5.576e-06

> marg.table <- matrix(c(100,16,54,263),2,2,byrow=T)
> mcnemar.test(marg.table,correct=F)

	McNemar's Chi-squared test

data:  marg.table
McNemar's chi-squared = 20.629, df = 1, p-value = 5.576e-06
```

2. **Problem 8.2**

(a) The ‘response’ column in the data provided is treated as $Y$. Logistic regression is fitted in R. The fitted model is

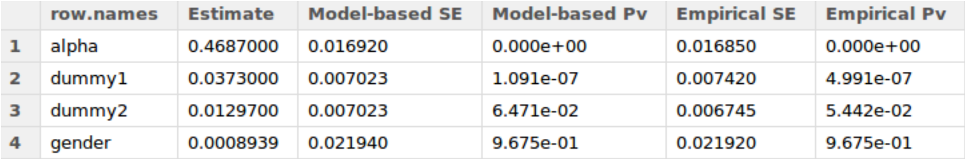
$$\text{logit}(P(Y_t = 1)) = -0.125 + 0.149I(t = 1) + 0.0520I(t = 2) + 0.00358X \qquad (2)$$



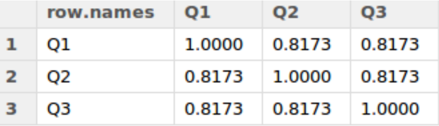
```
> dat <- as.data.frame(read.table('attitude.csv',header=T,sep=','))
> fit <- glm(response~gender+dummy1+dummy2,data=dat,family='binomial')
> summary(fit)

Call:
glm(formula = response ~ gender + dummy1 + dummy2, family = "binomial", 
    data = dat)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
-1.189  -1.148  -1.125   1.207   1.231  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)  
(Intercept) -0.125408   0.055601  -2.255   0.0241 *
gender       0.003582   0.054138   0.066   0.9472  
dummy1       0.149347   0.065825   2.269   0.0233 *
dummy2       0.052018   0.065843   0.790   0.4295  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7689.5  on 5549  degrees of freedom
Residual deviance: 7684.2  on 5546  degrees of freedom
AIC: 7692.2

Number of Fisher Scoring iterations: 3
```

(b) I fit the GEE logistic regression in R by using 'gee' library. The code and output are attached. In the R output, 'naive SE' means the model-based standard error. 'robust SE' means the empirical (sandwich) standard error. The results are summarized in the following tables.

Exchangeable correlation structure:

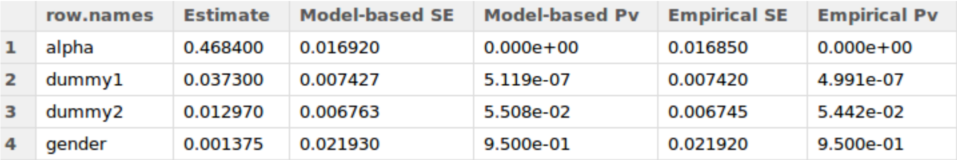


| | row.names | Estimate | Model-based SE | Model-based Pv | Empirical SE | Empirical Pv |
|---|---|---|---|---|---|---|
| 1 | alpha | 0.4687000 | 0.016920 | 0.000e+00 | 0.016850 | 0.000e+00 |
| 2 | dummy1 | 0.0373000 | 0.007023 | 1.091e-07 | 0.007420 | 4.991e-07 |
| 3 | dummy2 | 0.0129700 | 0.007023 | 6.471e-02 | 0.006745 | 5.442e-02 |
| 4 | gender | 0.0008939 | 0.021940 | 9.675e-01 | 0.021920 | 9.675e-01 |

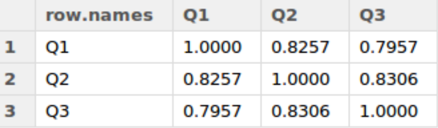


| | row.names | Q1 | Q2 | Q3 |
|---|---|---|---|---|
| 1 | Q1 | 1.0000 | 0.8173 | 0.8173 |
| 2 | Q2 | 0.8173 | 1.0000 | 0.8173 |
| 3 | Q3 | 0.8173 | 0.8173 | 1.0000 |

Unstructured correlation:



| | row.names | Estimate | Model-based SE | Model-based Pv | Empirical SE | Empirical Pv |
|---|---|---|---|---|---|---|
| 1 | alpha | 0.468400 | 0.016920 | 0.000e+00 | 0.016850 | 0.000e+00 |
| 2 | dummy1 | 0.037300 | 0.007427 | 5.119e-07 | 0.007420 | 4.991e-07 |
| 3 | dummy2 | 0.012970 | 0.006763 | 5.508e-02 | 0.006745 | 5.442e-02 |
| 4 | gender | 0.001375 | 0.021930 | 9.500e-01 | 0.021920 | 9.500e-01 |

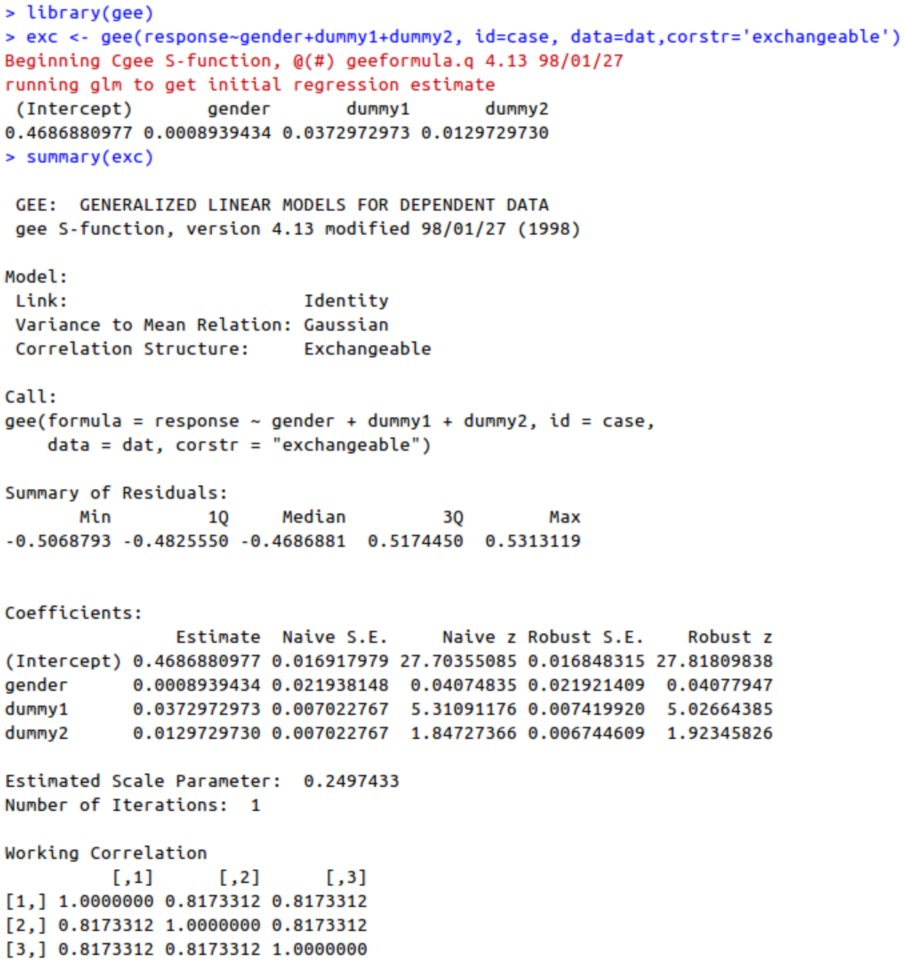


| | row.names | Q1 | Q2 | Q3 |
|---|---|---|---|---|
| 1 | Q1 | 1.0000 | 0.8257 | 0.7957 |
| 2 | Q2 | 0.8257 | 1.0000 | 0.8306 |
| 3 | Q3 | 0.7957 | 0.8306 | 1.0000 |

In either assumption of correlation structure, we have observed a very high estimated correlation between the three questions (around 0.8). This high correlation is indeed expected from Table 11.13. Due to the high correlation, the estimates and standard errors differ significantly from part (a) where we treat the three answers of the same individual as independent. Notice that by taking into consideration the correlation between 3 questions, the evidence for $\beta_1$ become much stronger, which implies that the attitudes toward different questions do differ.

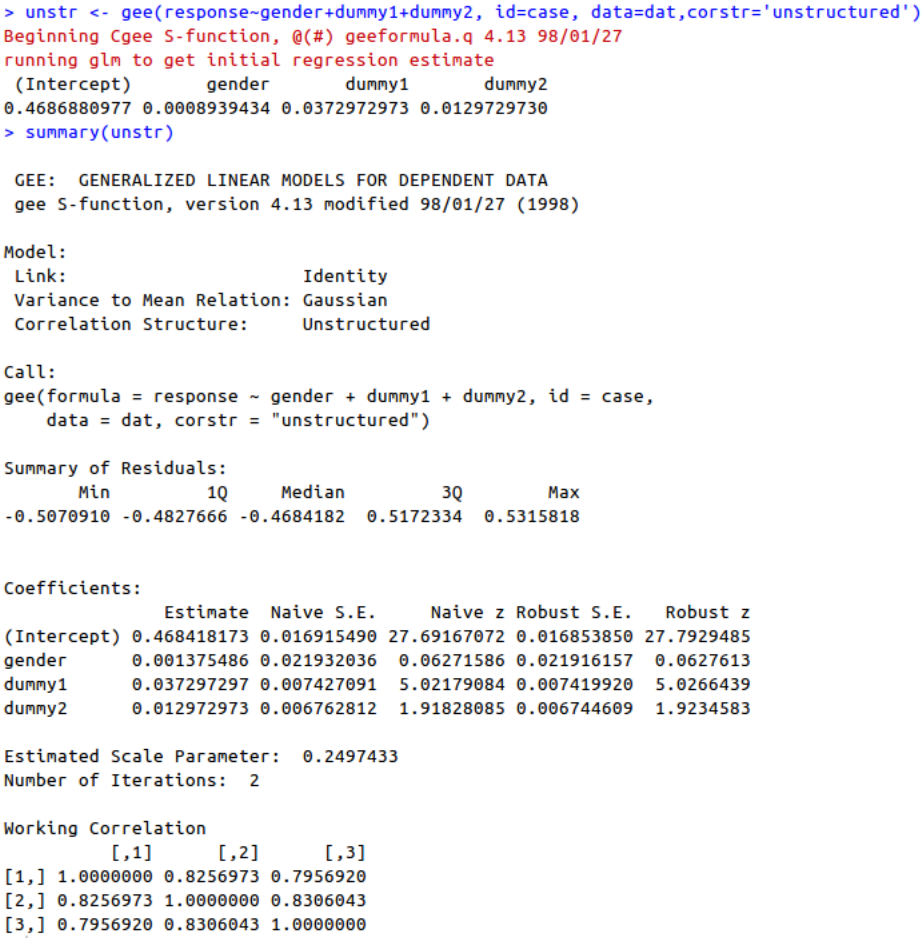
## R code:

Exchangeable:



```
> library(gee)
> exc <- gee(response~gender+dummy1+dummy2, id=case, data=dat,corstr='exchangeable')
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
 (Intercept)       gender       dummy1       dummy2 
0.4686880977 0.0008939434 0.0372972973 0.0129729730 
> summary(exc)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998) 

Model:
 Link:                      Identity 
 Variance to Mean Relation: Gaussian 
 Correlation Structure:     Exchangeable 

Call:
gee(formula = response ~ gender + dummy1 + dummy2, id = case, 
    data = dat, corstr = "exchangeable")

Summary of Residuals:
       Min         1Q     Median         3Q        Max 
-0.5068793 -0.4825550 -0.4686881  0.5174450  0.5313119 


Coefficients:
                Estimate  Naive S.E.     Naive z Robust S.E.    Robust z
(Intercept) 0.4686880977 0.016917979 27.70355085 0.016848315 27.81809838
gender      0.0008939434 0.021938148  0.04074835 0.021921409  0.04077947
dummy1      0.0372972973 0.007022767  5.31091176 0.007419920  5.02664385
dummy2      0.0129729730 0.007022767  1.84727366 0.006744609  1.92345826

Estimated Scale Parameter:  0.2497433
Number of Iterations:  1

Working Correlation
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.8173312 0.8173312
[2,] 0.8173312 1.0000000 0.8173312
[3,] 0.8173312 0.8173312 1.0000000
```

Unstructured:



```
> unstr <- gee(response~gender+dummy1+dummy2, id=case, data=dat,corstr='unstructured')
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
 (Intercept)       gender       dummy1       dummy2
0.4686880977 0.0008939434 0.0372972973 0.0129729730
> summary(unstr)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Unstructured

Call:
gee(formula = response ~ gender + dummy1 + dummy2, id = case,
    data = dat, corstr = "unstructured")

Summary of Residuals:
       Min         1Q     Median         3Q        Max
-0.5070910 -0.4827666 -0.4684182  0.5172334  0.5315818


Coefficients:
               Estimate  Naive S.E.     Naive z Robust S.E.   Robust z
(Intercept) 0.468418173 0.016915490 27.69167072 0.016853850 27.7929485
gender      0.001375486 0.021932036  0.06271586 0.021916157  0.0627613
dummy1      0.037297297 0.007427091  5.02179084 0.007419920  5.0266439
dummy2      0.012972973 0.006762812  1.91828085 0.006744609  1.9234583

Estimated Scale Parameter:  0.2497433
Number of Iterations:  2

Working Correlation
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.8256973 0.7956920
[2,] 0.8256973 1.0000000 0.8306043
[3,] 0.7956920 0.8306043 1.0000000
```