

LDA 140.655

Answer Key to Homework 2 2009

1. A study was conducted in West Java, Indonesia, to determine the effects of vitamin A deficiency in preschool children. The investigators were particularly interested in whether children with vitamin A deficiency were at increased risk of developing respiratory infection, which is one of the leading causes of death in this part of the world. 250 children were recruited in the study, and their age in years, gender (0 =male, 1 =female), and whether they suffered vitamin A deficiency (0 =no, 1 =yes) was recorded at an initial clinic visit (time 0). Also recorded was the response, whether the child was suffering from a respiratory infection (0 =no, 1 =yes). The children then were examined again at 3 month intervals for a year (at 3,6,12, and 15 months after the first visit) and the presence or absence of respiratory infection was recorded at each of these visits. Luckily, all children we seen at all visits, so there were no missing data.

The data file has the following columns:

Column	Description
1	Child id
2	Response (0 or 1)
3	Time (in months)
4	Gender (male = 0, female=1)
5	Vitamin A not deficient =0, deficient =1)
6	age in years)

- (a) Let  $\mathbf{y}_i$  be the vector of responses for the  $i$ th child, consisting of elements  $y_{ij}$ , the observations on whether the child has a respiratory infection at time  $t_{ij}$  (recorded in months). Write down a model for  $E(y_{ij})$  in terms of an appropriate link function that is linear in an intercept and include additive terms for time, age, gender, and vitamin A status. Also, write down  $\text{var}(y_{ij})$  given the nature of the response

$$\text{Let } g[E(y_{ij})] = \log\left(\frac{E(y_{ij})}{1 - E(y_{ij})}\right), \quad \log\left(\frac{E(y_{ij})}{1 - E(y_{ij})}\right) = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{age}_{ij} + \beta_3 \text{gender}_{ij} + \beta_4 \text{vita}_{ij},$$

where  $y_{ij}$  is the response and  $t_{ij}$  is 0, 3, 6, 9, 12 or 15. Also, given that the response is Bernoulli, we assume  $\text{var}(y_{ij}) = E(y_{ij})[1 - E(y_{ij})]$ . We model the log odds of the mean response (probability of respiratory infection, RI) by a linear function.

- (b) The investigators had not taken a course in longitudinal analysis; thus, they were unaware that measurements on the same child might be correlated. They fit the model in (a) without taking correlation into account, treating all the observations from all children as if they were unrelated. Based on this fit, is there sufficient evidence to suggest that the mean pattern of respiratory response is associated with the presence or absence of vitamin A deficiency? State the null hypothesis corresponding to this issue in terms of your model (a), cite the test statistic and p-value on which you base this conclusion, and state your conclusion as a meaningful sentence.

Fit the model in (a) with independent correlation structure:

```
. xi: xtgee infect time gender age vitA, nolog f(bin) l(logit) corr(ind)

GEE population-averaged model
Group variable:          id
Link:                    logit
Family:                  binomial
Correlation:            independent

Number of obs          =      1500
Number of groups       =        250
Obs per group: min     =         6
                   avg     =        6.0
                   max     =         6
Wald chi2(4)           =      38.71
```

```

Scale parameter:                1          Prob > chi2          =      0.0000
Pearson chi2(1500):             1492.29      Deviance              =      1782.54
Dispersion (Pearson):          .9948573  Dispersion             =      1.188361

```

infect	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
time	.0168947	.0111999	1.51	0.131	-.0050567 .0388462
gender	-.5576829	.114908	-4.85	0.000	-.7828984 -.3324674
age	-.0715987	.0297275	-2.41	0.016	-.1298636 -.0133338
vitA	.2869989	.1175161	2.44	0.015	.0566716 .5173263
_cons	-.5575417	.162073	-3.44	0.001	-.8751989 -.2398845

```

. test vitA
( 1) vitA = 0

```

```

          chi2( 1) =      5.96
        Prob > chi2 =      0.0146

```

**We obtained a Z statistics of 2.44 (p-value = 0.015) for the coefficient on vitA. Thus we can reject the null at a 0.05 significance level and conclude that there is evidence that the mean response is associated with the presence or absence of vitA.**

(c) Because you have taken a course in longitudinal data analysis, the investigators called you in for help with an improved analysis. Write down an extended model to (a) that takes into account correlation among repeated measurements on the same subject.

**Extend the model in (a) to account for correlation structure.**

$$E(y_{ij}) = \frac{\exp(\beta_0 + \beta_1 t_{ij} + \beta_2 age_{ij} + \beta_3 gender_{ij} + \beta_4 vitA_{ij})}{1 + \exp(\beta_0 + \beta_1 t_{ij} + \beta_2 age_{ij} + \beta_3 gender_{ij} + \beta_4 vitA_{ij})}$$

$$\text{var}[y_{ij}] = \mu_{ij}(1 - \mu_{ij})$$

$$\text{var}[Y_i] = \text{diagonal matrix with diagonal elements } \text{var}(y_{ij})$$

$$T_i^{1/2} = \sqrt{\text{Var}(Y_i)}$$

So, the covariance matrix taking into account the correlation structure is:  $T_i^{1/2} \Gamma_i T_i^{1/2}$

(d) Fit your model in (d) to the data. Assuming that your stated model for correlation is correct, conduct a test of the null hypothesis in part (b). State your conclusion as a meaningful sentence. Do the results agree with those in part (b)? Give a possible explanation for this, citing results from your output to support your explanation.

**Fit model from (c) using the unstructured correlation assumption**

```

. xtgee infect time gender age vitA, nolog f(bin) l(logit) corr(uns)

GEE population-averaged model
Group and time vars:      id time          Number of obs      =      1500
Link:                      logit          Number of groups   =      250
Family:                    binomial        Obs per group: min =      6
Correlation:              unstructured      avg                =     6.0
                                                max                =      6
                                                Wald chi2(4)       =     14.40
Scale parameter:          1              Prob > chi2        =     0.0061

```

	infect	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
time		.017177	.0081949	2.10	0.036	.0011154 .0332386
gender		-.5339991	.2139473	-2.50	0.013	-.9533282 -.1146701
age		-.0771982	.0553928	-1.39	0.163	-.1857662 .0313698
vitA		.2840136	.2189155	1.30	0.195	-.1450528 .7130801
_cons		-.5469366	.2655082	-2.06	0.039	-1.067323 -.0265501

Or look at the Z statistics 1.30 and the p-value=0.195. Thus we cannot reject the null at a 0.05 significance level. We conclude that there is no evidence that the mean response is associated with the presence absence of vitA.

The results of this test do not agree with the one in part (B). Looking at the working correlation matrix Estimated within-id correlation matrix R:

	c1	c2	c3	c4	c5	c6
r1	1.0000					
r2	0.5623	1.0000				
r3	0.4606	0.5757	1.0000			
r4	0.4240	0.5629	0.5587	1.0000		
r5	0.5035	0.5251	0.4250	0.4342	1.0000	
r6	0.4480	0.5636	0.5148	0.5189	0.5097	1.0000

We notice that there is substantial correlation between obs taken on the same subject. Thus, it is important to incorporate this correlation structure into the model.

We will try two other correlation models: exchangeable and ar1.

. xtgee infect time gender age vitA, nolog f(bin) l(logit) corr(exc)

```
GEE population-averaged model
Group variable:          id          Number of obs      =      1500
Link:                   logit        Number of groups   =       250
Family:                 binomial     Obs per group: min =         6
Correlation:           exchangeable  avg                =        6.0
Scale parameter:      1              max                =         6
Wald chi2(4)          =       14.69
Prob > chi2          =       0.0054
```

	infect	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
time		.0168792	.0078796	2.14	0.032	.0014355 .0323228
gender		-.5550523	.2156628	-2.57	0.010	-.9777436 -.132361
age		-.0744012	.0558017	-1.33	0.182	-.1837704 .0349681
vitA		.2757902	.2206841	1.25	0.211	-.1567427 .7083232
_cons		-.5443887	.2661777	-2.05	0.041	-1.066087 -.0226901

. xtgee infect time gender age vitA, nolog f(bin) l(logit) corr(ar1)

```
GEE population-averaged model
Group and time vars:    id time     Number of obs      =      1500
Link:                   logit        Number of groups   =       250
Family:                 binomial     Obs per group: min =         6
Correlation:           AR(1)         avg                =        6.0
Scale parameter:      1              max                =         6
Wald chi2(4)          =       16.53
Prob > chi2          =       0.0024
```

	infect	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
time		.01778	.0128053	1.39	0.165	-.007318 .0428779

gender		-.5245489	.1763746	-2.97	0.003	-.8702367	-.1788611
age		-.0768369	.0456722	-1.68	0.092	-.1663527	.0126789
vitA		.2861879	.1804226	1.59	0.113	-.067434	.6398097
_cons		-.5617749	.2333908	-2.41	0.016	-1.019212	-.1043374

The coefficients and standard errors of the coefficients for the gee with the exchangeable correlation structure match those for the gee with the unstructured correlation quite well, whereas the ar1 correlation structure does not produce such similar correlations. However all three correlation structure do not change the estimate coefficients qualitatively and we obtain the same conclusions for the effect of vitamin A on RI. We will select the final correlation structure to be the parametric correlation structure that produces estimates of the coefficients and the standard error of the coefficients that most closely match the estimates using the unstructured correlation matrix. Our final model will hence use an exchangeable correlation structure. In this way, we increase the power of our model since we use up fewer degrees of freedom estimating the exchangeable correlation matrix than we would estimating the unstructured correlation matrix. Finally to guard against mis-specifying the correlation structure, we will also use the robust variance estimates.

### Final Model

```
. xtgee infect time gender age vitA, nolog f(bin) l(logit) corr(exc) robust
```

GEE population-averaged model		Number of obs	=	1500
Group variable:	id	Number of groups	=	250
Link:	logit	Obs per group: min	=	6
Family:	binomial	avg	=	6.0
Correlation:	exchangeable	max	=	6
		Wald chi2(4)	=	15.24
Scale parameter:	1	Prob > chi2	=	0.0042

(Std. Err. adjusted for clustering on id)

		Semi-robust				
	infect	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
time		.0168792	.0082807	2.04	0.042	.0006493 .033109
gender		-.5550523	.2168697	-2.56	0.010	-.9801092 -.1299955
age		-.0744012	.0546842	-1.36	0.174	-.1815803 .032778
vitA		.2757902	.2232926	1.24	0.217	-.1618552 .7134357
_cons		-.5443887	.2759358	-1.97	0.049	-1.085213 -.0035644

(e) Based on your final model in (e), conduct a test of the null hypothesis in part (b). State your conclusion in a meaningful sentence. Do the results agree with those in part (b)? Give a possible explanation for this, citing results from your output to support your explanation.

After controlling for time, gender, and age, from the above model, we do not find a statistically significant association between vitamin A deficiency and respiratory infection. This conclusion differs from (b) where correlation between repeated measurements of the same child is not considered. The estimated effect of vitamin A remains similar (log OR of 0.275); however the standard error increases two-fold.

(f) Is there sufficient evidence to suggest that the probability of respiratory infection changed over the 15 month study period? Is there sufficient evidence to suggest that it is worthwhile to take gender into account in understanding the risk of respiratory infection in this population of children?

```
. test time
```

( 1) time = 0	chi2( 1) =	4.15
	Prob > chi2 =	0.0415

```

. test age
( 1) age = 0
      chi2( 1) =    6.55
      Prob > chi2 = 0.0105

      chi2( 1) =    1.85
      Prob > chi2 = 0.1737

. test gender
( 1) gender = 0
      chi2( 2) =    8.97
      Prob > chi2 = 0.0113

. test age gender
( 1) age = 0
( 2) gender = 0
      chi2( 2) =    8.97
      Prob > chi2 = 0.0113

```

The above tests show that there is evidence to suggest that it is worthwhile taking the gender and time into account. Based on the estimates from (f), girls had a lower risk of RI and the risk of RI decreased over the study period. However, age does not associate with respiratory infection AFTER controlling for the other variables.

(g) Specify a logistic regression model with random intercept and additive terms for time, age, gender, and vitamin A status.

$$\text{logit}P(y_{ij} = 1 | U_i) = (\beta_0 + U_i) + \beta_1 t_{ij} + \beta_2 \text{age}_{ij} + \beta_3 \text{gender}_{ij} + \beta_4 \text{vita}_{ij}$$

$$U_i \sim N(0, \sigma^2)$$

(h) Base on model (h), what describes the probability of respiratory infection for a 3-year old boy with random intercept  $U_i = 0$  at the at the second visit who does not have vitamin A deficiency (in terms of model parameters)? What describes a similar child but with vitamin A deficiency?

From the above model, for a 3-year old boy without vitamin A deficiency on the 2<sup>nd</sup> visit ( $t = 3$ ), the probability of respiratory infection is

$$\frac{\exp(\beta_0 + 3\beta_1 + 3\beta_2)}{1 + \exp(\beta_0 + 3\beta_1 + 3\beta_2)}$$

A similar boy with vitamin A deficiency has probability

$$\frac{\exp(\beta_0 + 3\beta_1 + 3\beta_2 + \beta_4)}{1 + \exp(\beta_0 + 3\beta_1 + 3\beta_2 + \beta_4)}$$

(i) Fit the above logistic regression model with random intercept. Compare the estimated coefficients and their standard errors with those obtained from model (e). Explain any differences you observe. Are the two models equivalent?

```

. xtlogit infect time gender age vita
Random-effects logistic regression
Group variable: id
Number of obs      =    1500
Number of groups   =     250

Random effects u_i ~ Gaussian
Obs per group: min =     6
                avg =    6.0
                max =     6

Log likelihood     = -666.24996
Wald chi2(4)      =    15.26
Prob > chi2       =    0.0042

```

```

-----+-----
infect |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----

```

time		.0341493	.0160027	2.13	0.033	.0027846	.065514
gender		-1.09796	.4154391	-2.64	0.008	-1.912206	-.2837144
age		-.140227	.1071796	-1.31	0.191	-.3502951	.0698411
vitA		.6023991	.4273532	1.41	0.159	-.2351978	1.439996
_cons		-1.169446	.5275679	-2.22	0.027	-2.20346	-.1354319
-----							
/lnsig2u		2.084724	.1801202			1.731695	2.437753
-----							
sigma_u		2.835907	.2554021			2.377019	3.383384
rho		.7096895	.0371102			.6320095	.7767637
-----							

Likelihood-ratio test of rho=0: chibar2(01) = 450.04 Prob >= chibar2 = 0.000

The coefficient estimates obtained from the random effect model are similar to that from the marginal model. More importantly, the inferences drawn from these estimates do not differ qualitatively. The effects of time and gender remain statistically significant, while the effect of vitamin A remains non-significant.

(j) Estimate the probability in (i) with the random effect model in (j) and the GEE model in (f).

After fitting the random effect and GEE model, we can use the `lincom` post-estimation command to obtain the predicted value and confidence interval on the log-odds scale.

```
. lincom 3*time+3*age+_cons
. iincom 3*time+3*age+vitA+_cons
```

For the random effect model,

$$\frac{\exp(\hat{\beta}_0 + 3\hat{\beta}_1 + 3\hat{\beta}_2)}{1 + \exp(\hat{\beta}_0 + 3\hat{\beta}_1 + 3\hat{\beta}_2)} = 0.18 \text{ (0.10-0.31)}; \frac{\exp(\hat{\beta}_0 + 3\hat{\beta}_1 + 3\hat{\beta}_2 + \hat{\beta}_4)}{1 + \exp(\hat{\beta}_0 + 3\hat{\beta}_1 + 3\hat{\beta}_2 + \hat{\beta}_4)} = 0.29 \text{ (0.16-0.48)}$$

For the GEE model,

$$\frac{\exp(\hat{\beta}_0 + 3\hat{\beta}_1 + 3\hat{\beta}_2)}{1 + \exp(\hat{\beta}_0 + 3\hat{\beta}_1 + 3\hat{\beta}_2)} = 0.33 \text{ (0.25-0.41)}; \frac{\exp(\hat{\beta}_0 + 3\hat{\beta}_1 + 3\hat{\beta}_2 + \hat{\beta}_4)}{1 + \exp(\hat{\beta}_0 + 3\hat{\beta}_1 + 3\hat{\beta}_2 + \hat{\beta}_4)} = 0.39 \text{ (0.30-0.49)}$$

And from both models, the two estimated probabilities overlap.

(k) Report and interpret the estimated variance of the random effect and the intra-class correlation.

The estimated degree of heterogeneity in log-odds is  $2.83^2 = 8.00$ , indicating significant difference in baseline propensity of respiratory infection among the children. Also, there exists high intra-class correlation (0.71) on the log-odds of infection within a child.

(l) One could imagine that respiratory infection at a particular time might be dependent on previous infection. Perhaps children exhibited such behavior are more prone to show it again. Fit a logistic model that examines this potential phenomenon. Report and interpret the odds ratio estimates of the effect of previous respiratory infection.

```
* shift the infection indicator down by 1 row
. gen infect_prev = infect[_n-1]
* drop the first visit of all subjects
. by id: drop if _n==1
```

**GEE model with exchangeable correlation:**

```
. xtgee infect infect_prev time gender age vitA, nolog f(bin) l(logit) corr(exc)
robust
```

```
GEE population-averaged model          Number of obs      =      1250
Group variable:                        id                 Number of groups   =       250
Link:                                  logit              Obs per group: min =         5
Family:                                binomial            avg                  =       5.0
Correlation:                           exchangeable        max                  =         5
Scale parameter:                        1                  Wald chi2(5)       =      30.33
                                          Prob > chi2         =      0.0000
```

(Std. Err. adjusted for clustering on id)

	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]	
infect_prev	.6266044	.1776886	3.53	0.000	.2783411	.9748676
time	.0139394	.0101137	1.38	0.168	-.005883	.0337618
gender	-.5243144	.1995905	-2.63	0.009	-.9155046	-.1331241
age	-.0751472	.0504864	-1.49	0.137	-.1740988	.0238044
vitA	.2393817	.203166	1.18	0.239	-.1588162	.6375797
_cons	-.7559597	.2708649	-2.79	0.005	-1.286845	-.2250742

### Random intercept model:

```
. xtlogit infect infect_prev time gender age vitA
```

```
Random-effects logistic regression          Number of obs      =      1250
Group variable: id                         Number of groups   =       250
Random effects u_i ~ Gaussian              Obs per group: min =         5
                                          avg                  =       5.0
                                          max                  =         5
Wald chi2(5)                               =      23.48
Prob > chi2                                 =      0.0003
Log likelihood = -564.86277
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
infect_prev	.7714571	.3158682	2.44	0.015	.1523667	1.390547
time	.0262321	.0208573	1.26	0.209	-.0146475	.0671116
gender	-.9711682	.3811278	-2.55	0.011	-1.718165	-.2241715
age	-.1325172	.0965802	-1.37	0.170	-.321811	.0567765
vitA	.437492	.3823023	1.14	0.252	-.3118068	1.186791
_cons	-1.219728	.4984402	-2.45	0.014	-2.196653	-.2428029
/lnsig2u	1.712958	.3069945			1.11126	2.314656
sigma_u	2.354854	.3614637			1.743038	3.181421
rho	.6276411	.071747			.4801135	.7546945

```
Likelihood-ratio test of rho=0: chibar2(01) = 33.70 Prob >= chibar2 = 0.000
```

**We found evidence suggesting that previous respiratory infection (a lag of 3 months) was significantly associated with a recurrence from both the population-averaged and the subject-specific model. The odds ratios are 1.87 from the GEE with exchangeable correlation and 2.16 from the random effect model.**

(m) Given all the data analyses you have conducted so far, write a brief summary discussing:

1. The statistical model you assumed, and why you choose it
2. The analyses you conducted, the assumptions you made and why you made them
3. The results, addressing the interests of the investigators as described above.

**We investigated effects of vitamin A deficiency on the risk of developing respiratory infection in 250 preschool children. We highlighted the importance of using modeling approaches that account for correlations between repeated measurements from the same child. Based on the assumption that these correlations are lag-independent and identical for each child, we used the GEE method to estimate a population-averaged odds ratio for vitamin A deficiency. After controlling for gender, age, time since first visit, no statistically significant effect was found. We also explored the sensitivity of our estimates due to different working correlation matrices and found that our exchangeable assumption produced very similar standard errors compared to that using the saturated (unstructured) working correlation matrix. A logistic regression model with random intercepts was also examined and, similarly, we found no statistically significant effect for vitamin A deficiency on respiratory infection. Furthermore, based on the random effect model, we found that girls and older children were less likely to develop respiratory infection, after controlling for the other variables. The OR associated with girls compared to boys is 0.37 ( $CI_{95} = 0.20-0.72$ ) and the OR associated with one year increase in age is 0.27 ( $CI_{95} = 0.19-1.05$ ). Our analyses also found that the children were slightly more likely to develop respiratory infection further into the study with an OR of 1.03 ( $CI_{95} = 1-1.06$ ) per month. Finally, we found evidence that a previous respiratory infection (lag 3 month) was associated with reoccurrence: population-averaged OR = 1.87 ( $CI_{95} = 1.32-2.64$ ) and subject-specific OR = 2.16 ( $CI_{95} = 1.16-4.01$ ).**