FINAL SOLUTION 2009

The association between maternal smoking and respiratory health of children
Outcome variable: wheezing (binary: 0, 1)
C: City (1 = Kingston, 0 = Portage)
Repeated measurements "t": Once a year (age (t) = 9, 10, 11, 12)
Mother's smoking status (categorical: 0, 1, 2, with dummy variables X_1 and X_2)
Scientific question: to assess and compare the effects of smoking patterns on wheezing patterns

---

```
** Read in Dataset (Wide)
infile id str10 city age9 smk9 whz9 age10 smk10 whz10 age11 smk11 whz11 age12
smk12 whz12

using wheeze2.raw, clear

** Convert to long format
reshape long smk whz , i(id) j(age)
drop age9-age12

** Generate the moderate and heavy smoker indicator
gen smk1 = 1 if smk == 1
replace smk1 = 0 if smk1 == .
gen smk2 = 1 if smk == 2
replace smk2 = 0 if smk2 ==.
```

(a) 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 city, for smoking (moderate and heavy), and time. Also, write down var(y_{ij}) given the nature of the response. Interpret the coefficients in your model.

Let y_{ij} be the response at time t_{ij} = 9, 10, 11, and 12 for the ith child.

<u>Link function:</u> $$g(E(y_{ij})) = \log\left(\frac{E(y_{ij})}{1-E(y_{ij})}\right)$$

<u>Systematic part:</u> $$\log\left(\frac{E(y_{ij})}{1-E(y_{ij})}\right) = \beta_0 + \beta_1 c_i + \beta_2 x_{1ij} + \beta_3 x_{2ij} + \beta_4 t_{ij}$$

<u>Random part:</u>

The binary responses are correlated, and the diagonal element of covariance matrix are:

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

<u>Model coefficient interpretation:</u>

On the Population-level:

β_0: log odds of wheezing for children from Portage with non-smoker mothers at birth.

β_1: log odds ratio of wheezing comparing children from Kingston to children from Portage with the same mother smoking status and age.

β_2 : log odds ratio of wheezing comparing same-age children whose mothers are moderate smokers to children whose mothers are non-smokers from the same city.
β_3 : log odds ratio of wheezing comparing same-age children whose mothers are heavy smokers to children whose mothers are non-smokers from the same city.
β_4 : log odds ratio of wheezing due to one year increase in age of children from the same city and same mother's smoking status.

(b) Under your model in (a)

The log odds of wheezing for a child from Kingston whose mother is a heavy smoker at t_{ij} is $\beta_0 + \beta_1 + \beta_3 + \beta_4 t_{ij}$.

```
xi: xtgee whz age smk1 smk2 i.city, nolog f(bin) l(logit) corr(ind)

i.city            _Icity_1-2          (_Icity_1 for city==kingston omitted)

GEE population-averaged model                   Number of obs      =       128
Group variable:                         id      Number of groups   =        32
Link:                                logit      Obs per group: min =         4
Family:                           binomial                     avg =       4.0
Correlation:                   independent                     max =         4
                                                Wald chi2(4)       =      4.10
Scale parameter:                         1      Prob > chi2        =    0.3930

Pearson chi2(128):                  126.26      Deviance           =    147.94
Dispersion (Pearson):              .986411      Dispersion         =  1.155751

------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1993475   .1803634    -1.11   0.269    -.5528533    .1541583
        smk1 |  -.1276565   .4582412    -0.28   0.781    -1.025793    .7704798
        smk2 |   .7347176   .5406551     1.36   0.174    -.3249469    1.794382
    _Icity_2 |  -.2117842   .4010502    -0.53   0.597    -.9978281    .5742597
       _cons |    1.15685   1.899495     0.61   0.543    -2.566093    4.879792
------------------------------------------------------------------------------
```

(c) The investigators 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.

We fit a longitudinal logistic regression model assuming 'independent' correlation structure. When adjusted by age and city, mother's smoking status is 'not' significantly associated with wheezing. P-values for both smk1 (the mother is moderate smoker) and smk2 (the mother is heavy smoker) are larger than alpha-level 0.05 (0.781 and 0.174, respectively). Testing smk1 and smk2 simultaneously, the p-value was 0.235, showing that those two variables together overall was not statistically significant either.

$H_0 : \beta_2 = 0$	-> p = 0.781 > 0.05; therefore, failed to reject the null
$H_0 : \beta_3 = 0$	-> p = 0.174 > 0.05; therefore, failed to reject the null
$H_0 : \beta_2 = \beta_3 = 0$	-> p = 0.2325 > 0.05; therefore, failed to reject the null

(d) Why might the analysis in (c) be unreliable?

Failure to take into account within-subject correlation leads to incorrect estimation of the standard error for the estimated coefficients. Thus, hypothesis tests about those coefficients based on their standard error give incorrect results, from which we may draw incorrect conclusion.

(e) Logistic regression in longitudinal data with taking into account correlation among repeated measurements on the same subject

Link function: $g(E(y_{ij})) = \log\left(\frac{E(y_{ij})}{1-E(y_{ij})}\right)$

Systematic part: $\log\left(\frac{E(y_{ij})}{1-E(y_{ij})}\right) = \beta_0 + \beta_1 c_i + \beta_2 x_{1ij} + \beta_3 x_{2ij} + \beta_4 t_{ij}$,

Where y_{ij} is the response, and t_{ij} is 9, 10, 11, and 12

Random part: the responses are correlated Bernoulli, and need specify the correlation matrix.

$$\mathrm{var}[y_{ij}] = \mu_{ij}(1-\mu_{ij})$$
$$T_i^{1/2} = \sqrt{Var(Y_i)}$$
$$\Gamma_i = corr.matrix.for.subjec.i$$
$$Var(Y_i) = T_i^{1/2}\Gamma_i T_i^{1/2}$$

(f) Fit your model in (e) to the data and conduct a test of the null hypothesis in part (c). State your conclusion as a meaningful sentence. Do the results agree with those in part (c)? Give a possible explanation for this, citing results from your output to support your explanation.

```
. xi: xtgee whz age smk1 smk2 i.city, nolog f(bin) l(logit) corr(unst)
i.city           _Icity_1-2          (_Icity_1 for city==kingston omitted)

GEE population-averaged model                   Number of obs      =       128
Group and time vars:                    id age  Number of groups   =        32
Link:                                    logit  Obs per group: min =         4
Family:                               binomial                 avg =       4.0
Correlation:                      unstructured                 max =         4
                                                Wald chi2(4)       =      4.68
Scale parameter:                             1  Prob > chi2        =    0.3223

------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.2144158   .1746147    -1.23   0.219    -.5566543    .1278228
        smk1 |  -.0223768   .4500519    -0.05   0.960    -.9044624    .8597087
        smk2 |   .8193055   .5183241     1.58   0.114    -.1965911    1.835202
    _Icity_2 |  -.2001139   .4154962    -0.48   0.630    -1.014471    .6142437
       _cons |   1.284055   1.824173     0.70   0.481    -2.291259     4.85937
------------------------------------------------------------------------------


. test  smk1  smk2

 ( 1)  smk1 = 0
 ( 2)  smk2 = 0
```

```
          chi2(  2) =    3.33
        Prob > chi2 =    0.1890
```

From the STATA output

$H_0{:}\beta_2 = 0$ -> p = 0.960 > 0.05; therefore, failed to reject the null

$H_0{:}\beta_3 = 0$ -> p = 0.114 > 0.05; therefore, failed to reject the null

$H_0{:}\beta_2 = \beta_3 = 0$ -> p = 0.1890 > 0.05; therefore, failed to reject the null

Therefore, when adjusted by age and city, mother's smoking status is 'not' significantly associated with wheezing at 0.05 alpha-level. This result agrees with those in part (c). This is because the within-subject correlation is relatively small so that independent assumption for the correlation structure will not affect the model inference very much.

(g) Do you think a simpler model for correlation may be plausible? Select and explain a correlation model you feel is most plausible, and fit this model to the data.

Based on the correlation structure estimated from (f) with an unstructured correlation, either of the exponential or the exchangeable model is suitable for this dataset. However, the correlation between age 10 and 11 (0.27) may not be treated as independent. To be conservative, the unstructured correlation with robust variance is used for the following analyses.

```
. xtcorr
Estimated within-id correlation matrix R:
        c1       c2       c3       c4
r1   1.0000
r2  -0.0932   1.0000
r3   0.0543   0.2669   1.0000
r4   0.0231  -0.0708   0.0768   1.0000

. xi:xtgee whz i.city smk1 smk2 age, nolog f(bin) link(logit) corr(uns) robust
           |               Semi-robust
       whz |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
      city |   .2001139    .411357     0.49   0.627     -.606131    1.006359
      smk1 |  -.0223768   .4658936    -0.05   0.962    -.9355115    .8907578
      smk2 |   .8193055   .4853743     1.69   0.091    -.1320106    1.770622
       age |  -.2144158   .1804719    -1.19   0.235    -.5681342    .1393027
      cons |   1.083942   1.929807     0.56   0.574    -2.698411    4.866294
------------------------------------------------------------------------------

. test smk1 smk2
             ( 1)  smk1 = 0.0
             ( 2)  smk2 = 0.0              chi2(  2) =    3.60
                                              Prob > chi2 =     0.1651
. test city
             ( 1)  city = 0.0                chi2(  1) =    0.24
                                             Prob > chi2 =    0.6266
```

The analysis shows that there is no statistically significant evidence that wheezing is associated with mother's smoking status (p-value .17) at alpha-level 0.05, after adjusting for other confounders. City is also not a statistically significant risk factor of wheezing (p-value .63), after adjusting for other confounder.

(h) From your fit in (g), estimate the probability that child from Kingston whose mother is heavy smoker wheeze at the initial visit. And, estimate of the probability that child from Kingston whose mother does not smoke wheeze at the initial visit. What can you conclude?

The model fit in (g) looks as follows:

$$\log\left(\frac{E(y_{ij})}{1-E(y_{ij})}\right) = 1.08 + 0.20c_i - 0.02x_{1ij} + 0.82x_{2ij} - 0.21t_{ij}$$

For the first child:

```
. lincom _cons+smk2+_Icity_2+9*age

 ( 1)  _Icity_2 + smk2 + 9 age + _cons = 0

------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0264948   .4029385    -0.07   0.948    -.8162398    .7632503
```

→ The probability is **0.49 (95% CI: 0.31 – 0.68)**

For the second child:

```
. lincom _cons+_Icity_2+9*age

 ( 1)  _Icity_2 + 9 age + _cons = 0

------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.8458002   .4879743    -1.73   0.083    -1.802212    .1106118
```

→ The probability is **0.30 (95% CI: 0.14 – 0.52)**

The probability of wheezing for a child with heavy smoker mother is higher than that of a child with non-smoking mother, when other variables are held equal. However, this is not statistically significant, since the p-value for heavy-smoking status greater than alpha-level 0.05. Also the two confidence intervals overlap. Therefore, we cannot draw statistically significant conclusion.

(i.1) One could imagine that wheezing at a particular time might be dependent on past and present maternal smoking behavior. Write down model and fit it, and report finding.

An example model with past maternal smoking behavior (with both previous moderate and heavy smoking status):

$$\log\left(\frac{E(y_{ij})}{1-E(y_{ij})}\right) = \beta_0 + \beta_1 c_i + \beta_2 x_{1ij} + \beta_2^{/} x_{1ij-1} + \beta_3 x_{2ij} + \beta_3^{/} x_{2ij-1} + \beta_4 t_{ij}$$

Based on the STATA output below, the model can be specified as follows:

$$\log\left(\frac{E(y_{ij})}{1-E(y_{ij})}\right) = -0.38 + 0.57c_i + 0.038x_{1ij} - 0.22x_{1ij-1} + 1.27x_{2ij} + 0.0052x_{2ij-1} - 0.107\ t_{ij}$$

```
. xi:xtgee whz i.smk1 i.smk1_lag1 i.smk2 i.smk2_lag1 age city, nolog f(bin) l(logit)
corr(unst) robust

GEE population-averaged model                   Number of obs      =        96
Group and time vars:                 id age     Number of groups   =        32
Link:                                 logit     Obs per group: min =         3
Family:                            binomial                    avg =       3.0
Correlation:                   unstructured                    max =         3
                                                Wald chi2(6)       =      8.05
Scale parameter:                          1     Prob > chi2        =    0.2348

                               (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
             |             Semi-robust
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Ismk1_1 |   .0377312   .5635067     0.07   0.947    -1.066722    1.142184
_Ismk1_lag~1 |  -.2158187    .555965    -0.39   0.698     -1.30549    .8738527
    _Ismk2_1 |   1.273384   .6339722     2.01   0.045     .0308217    2.515947
_Ismk2_lag~1 |   .0052262   .9997849     0.01   0.996    -1.954316    1.964769
         age |  -.1071572   .3212241    -0.33   0.739    -.7367449    .5224306
        city |   .5702549    .524679     1.09   0.277     -.458097    1.598607
       _cons |  -.3758572   3.685178    -0.10   0.919    -7.598673    6.846958
```

(i.2) One could imagine that wheezing at a particular time might be dependent on previous wheezing. Perhaps children who have already exhibited such behavior are more prone to show it again. Write down model and fit it, and report finding.

The model with previous wheezing:

$$\log\left(\frac{E(y_{ij})}{1-E(y_{ij})}\right)=\beta_0+\beta_1 c_i+\beta_2 x_{1ij}+\beta_3 x_{2ij}+\beta_4 t_{ij}+\beta_4 y_{ij-1}$$

Based on the STATA output, the model can be specified as follows (log odds)

$$\log\left(\frac{E(y_{ij})}{1-E(y_{ij})}\right)=0.73+0.55c_i+0.10x_{1ij}+1.21x_{2ij}-0.19t_{ij}-0.80y_{ij-1}$$

```
. xi:xtgee whz i.city i.smk1 i.smk2 age i.whz_lag1, nolog f(bin) l(logit)
corr(unst) robust

GEE population-averaged model                   Number of obs      =        96
Group and time vars:                 id age     Number of groups   =        32
Link:                                 logit     Obs per group: min =         3
Family:                            binomial                    avg =       3.0
Correlation:                   unstructured                    max =         3
                                                Wald chi2(5)       =      7.11
Scale parameter:                          1     Prob > chi2        =    0.2129
------------------------------------------------------------------------------
             |             Semi-robust
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Icity_1 |   .5517445   .5532821     1.00   0.319    -.5326685    1.636157
    _Ismk1_1 |    .103979   .5853943     0.18   0.859    -1.043373    1.251331
    _Ismk2_1 |     1.2126   .5990186     2.02   0.043     .0385451    2.386655
         age |  -.1926388   .3275985    -0.59   0.557      -.83472    .4494424
_Iwhz_lag1_1 |  -.8014573   .6043182    -1.33   0.185    -1.985899    .3829845
       _cons |   .7271864   3.736305     0.19   0.846    -6.595837     8.05021
```

From the STATA output, we conclude that, at an alpha-level of 0.05 and controlling for other covariates, on the population level:

1) past maternal smoking is not significantly associated with child wheezing.
2) past wheezing is not significantly associated with present child wheezing.
3) however, after controlling for previous maternal smoking status or previous wheezing status, current maternal heavy smoking is associated with increased odds of wheezing.

(j) Write down a logistic regression model with random intercept and additive terms for city, for smoking and time.

$$\text{logit}P(y_{ij}=1\,|\,U_i)=(\beta_0+U_i)+\beta_1 c_i+\beta_2 x_{1ij}+\beta_3 x_{2ij}+\beta_4 t_{ij}$$

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

β_0 : average baseline log odds of wheezing at age 0 for a typical (Ui = 0) child from Protage with a non-smoker mother.
β_1 : subject-specific change in log odds of wheezing of a child being from Kingston to Protage, adjusted for the other variables.
β_2 : subject-specific log odds ratio of wheezing comparing moderate to non-smoker mother, adjusted for other variables.
β_3 : subject-specific log odds ratio of wheezing comparing heavy to non-smoker mother, adjusted for other variables.
β_4 : subject-specific change in log odds ratio of wheezing due to one year increase in age, adjusted for other variables.
U_i : random deviation of baseline odds from β_0 for individual i
v^2 : variance of the random deviations U_i .

(j.1) The probability of the child with random intercept Ui = 0, from Portage whose mother is heavy smoker at t_{ij}?

$$\exp(\beta_0+\beta_3+\beta_4 t_{ij})/(1+\exp(\beta_0+\beta_3+\beta_4 t_{ij}))$$

(j.2) What describes the probability of wheezing for a child with random intercept $U_i = 2$ from Portage whose mother is moderate smoker at t_{ij}?

$$\exp(\beta_0+2+\beta_2+\beta_4 t_{ij})/(1+\exp(\beta_0+2+\beta_2+\beta_4 t_{ij}))$$

(j.3) What describes the odds ratio of wheezing comparing a child from Kingston whose mother is heavy smoker to if the child is from Protage with a mother who doesn't smoke at t_{ij}? (G*ive answers in terms of model parameters*)

$$\exp(\beta_1+\beta_3)=\exp(\beta_1)\exp(\beta_3)$$

(k) Fit the logistic regression model with random intercept and compare the estimated coefficients and their standard errors with those obtained from model (g). Are the two models equivalent? Also estimate (j.1) and compare these estimates with the population average estimates obtained from model (g). Report and interpret the estimated degree of heterogeneity across children in the log-odds of wheezing not attributable to the covariates.

```
xi:xtlogit whz age i.smk1 i.smk2 i.city, nolog i(id) re

Random-effects logit                           Number of obs      =       128
Group variable (i) : id                        Number of groups   =        32

Random effects u_i ~ Gaussian                  Obs per group: min =         4
                                               Wald chi2(4)       =      3.99
Log likelihood  = -73.927332                   Prob > chi2        =    0.4076
------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.2041793    .183191    -1.11   0.265     -.563227    .1548685
    _Ismk1_1 |  -.1215249   .4714092    -0.26   0.797     -1.04547    .8024202
    _Ismk2_1 |   .7577636   .5637224     1.34   0.179     -.347112    1.862639
    _Icity_1 |   .2168998   .4234064     0.51   0.608    -.6129616    1.046761
       _cons |   .9613227    1.92156     0.50   0.617    -2.804865    4.727511
-------------+----------------------------------------------------------------
    /lnsig2u |  -2.168139   3.734647                     -9.487913    5.151635
-------------+----------------------------------------------------------------
     sigma_u |   .3382163   .6315593                      .0087041    13.14206
         rho |   .0336021   .0368633                       .000023    .9813079
------------------------------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01) =     0.08 Prob >= chibar2 = 0.388
```

From the random-intercept model, the estimated log odds of wheezing for a child with random intercept deviation 0, from Portage whose mother is heavy smoker at time t_{ij} is $1.72-.204^* t_{ij}$ (estimated probability is $\exp(1.72-.204^* t_{ij})/(1+\exp(1.72-.204^* t_{ij}))$). Assuming the child is at age 9,

```
lincom _cons + 9*age + _Ismk2_1

 ( 1)  9 [whz]age + [whz]_Ismk2_1 + [whz]_cons = 0

------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0983727   .5705546     0.17   0.863    -1.019894    1.216639
```

→ Probability is 0.52 (95% CI: 0.26 - 0.77)

The same probability from the GEE model,

```
. xi:xtgee whz i.city smk1 smk2 age, nolog f(bin) link(logit) corr(uns) robust
. lincom _cons + 9*age + smk2

 ( 1)  smk2 + 9 age + _cons = 0

------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .1736191   .3593006     0.48   0.629    -.5305971    .8778353
------------------------------------------------------------------------------
```

→ Probability is 0.54 (95% CI: 0.37 – 0.70)

Numerical differences in point estimates reflect the difference between a population-averaged effect and its individual-level counterpart for models where the link function is not linear. Also as expected, the standard errors estimated from the random effect model are larger than those from the GEE model and the point estimates from the GEE model are small in absolute magnitude than those from the random effect model. The GEE and the random effect model are **not** equivalent.

Rho=0.034 describes the estimated degree of heterogeneity across children in the propensity of wheezing, not due to covariates. This number is relatively small (<5%), which means it may not be necessary to include random effects in the model. Although including random effects will enhance model predictability.

$$rho \approx \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} = \frac{.34^2}{.34^2 + \pi^2 / 3} = 0.033$$