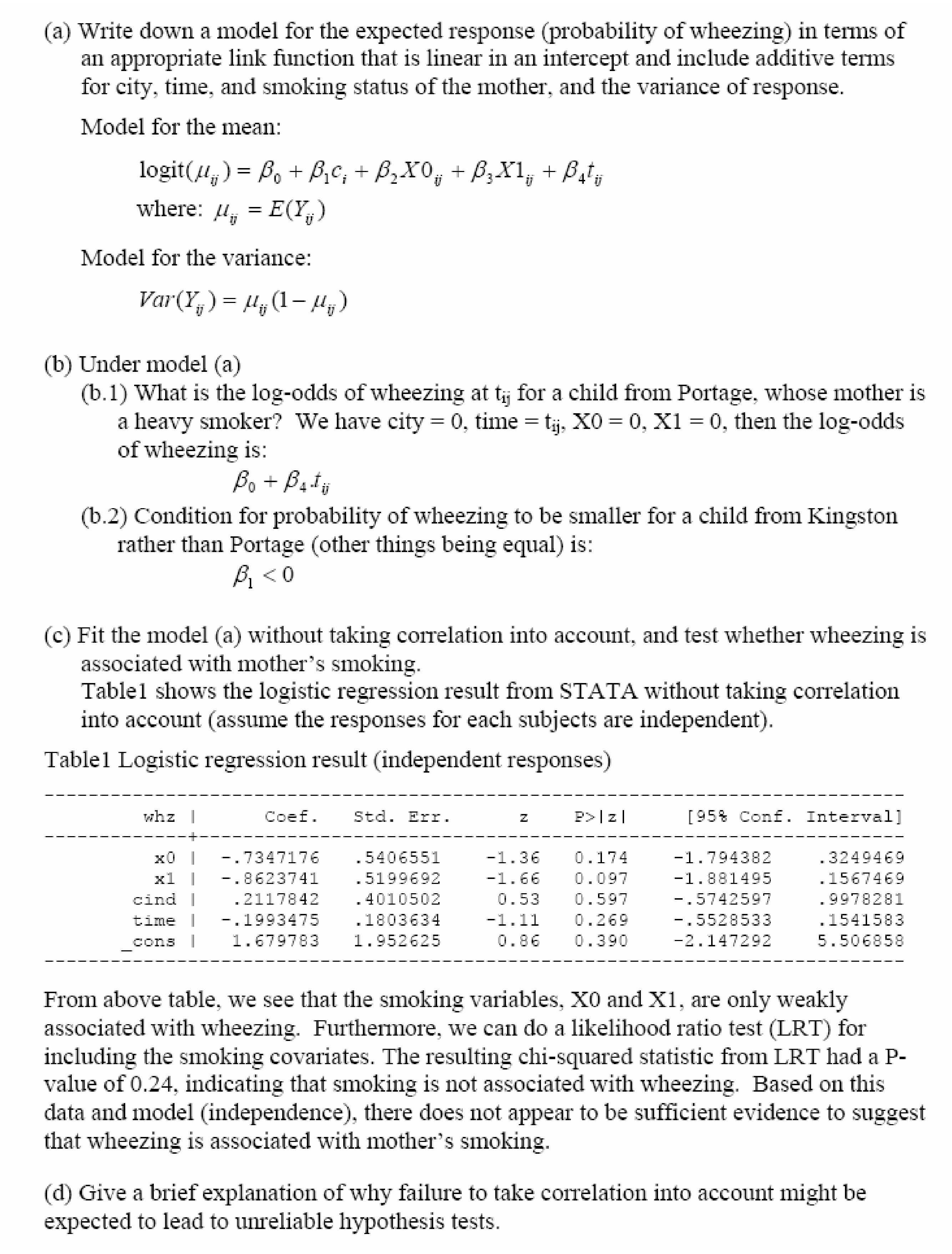


(a) Write down a model for the expected response (probability of wheezing) in terms of an appropriate link function that is linear in an intercept and include additive terms for city, time, and smoking status of the mother, and the variance of response.

Model for the mean:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 c_i + \beta_2 X0_{ij} + \beta_3 X1_{ij} + \beta_4 t_{ij}$$

where: $\mu_{ij} = E(Y_{ij})$

Model for the variance:

$$Var(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$$

(b) Under model (a)

(b.1) What is the log-odds of wheezing at t_{ij} for a child from Portage, whose mother is a heavy smoker? We have city = 0, time = t_{ij}, X0 = 0, X1 = 0, then the log-odds of wheezing is:

$$\beta_0 + \beta_4 . t_{ij}$$

(b.2) Condition for probability of wheezing to be smaller for a child from Kingston rather than Portage (other things being equal) is:

$$\beta_1 < 0$$

(c) Fit the model (a) without taking correlation into account, and test whether wheezing is associated with mother's smoking.

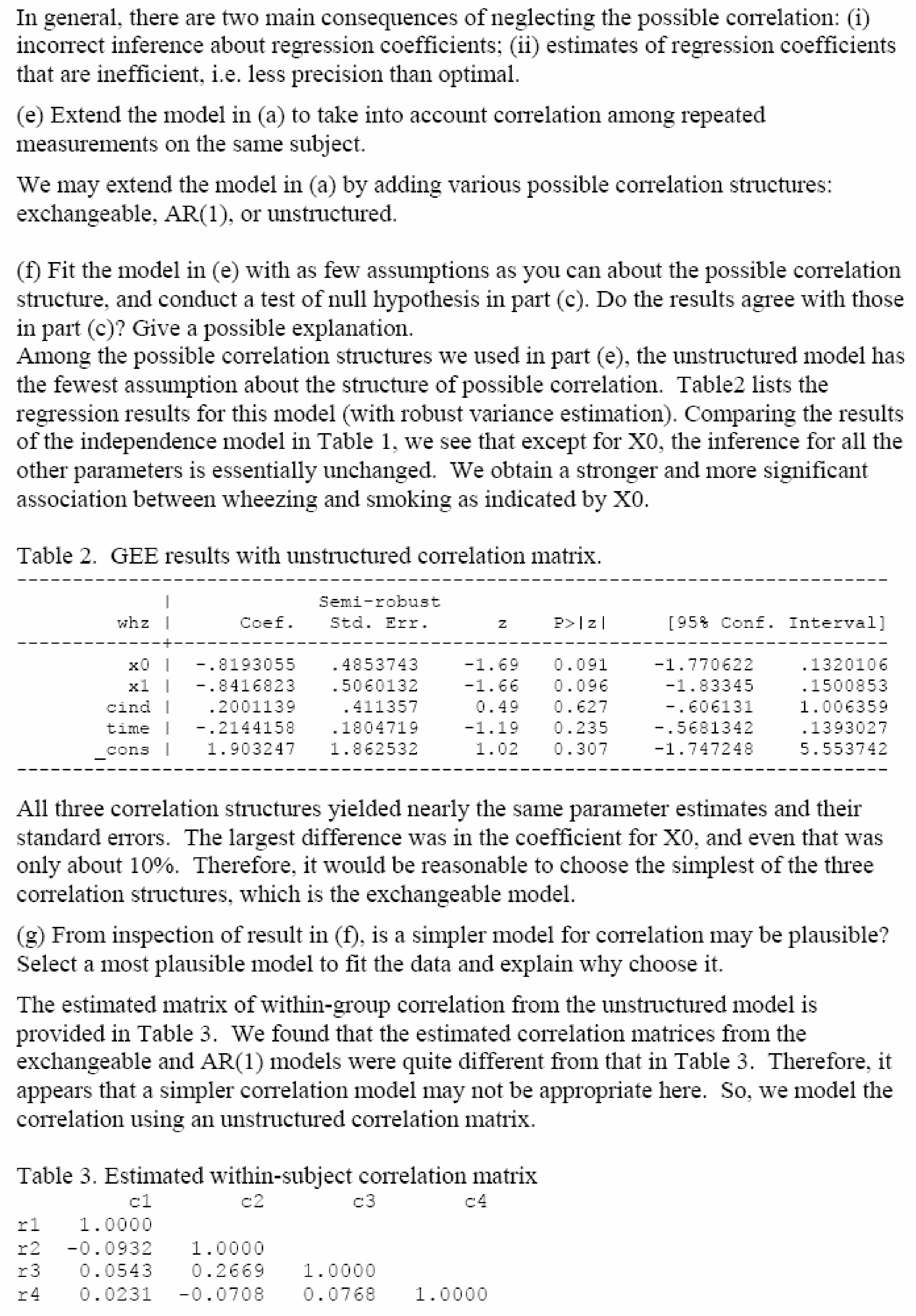
Table1 shows the logistic regression result from STATA without taking correlation into account (assume the responses for each subjects are independent).

Table1 Logistic regression result (independent responses)

```
------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x0 |  -.7347176   .5406551    -1.36   0.174    -1.794382    .3249469
          x1 |  -.8623741   .5199692    -1.66   0.097    -1.881495    .1567469
        cind |   .2117842   .4010502     0.53   0.597    -.5742597    .9978281
        time |  -.1993475   .1803634    -1.11   0.269    -.5528533    .1541583
       _cons |   1.679783   1.952625     0.86   0.390    -2.147292    5.506858
------------------------------------------------------------------------------
```

From above table, we see that the smoking variables, X0 and X1, are only weakly associated with wheezing. Furthermore, we can do a likelihood ratio test (LRT) for including the smoking covariates. The resulting chi-squared statistic from LRT had a P-value of 0.24, indicating that smoking is not associated with wheezing. Based on this data and model (independence), there does not appear to be sufficient evidence to suggest that wheezing is associated with mother's smoking.

(d) Give a brief explanation of why failure to take correlation into account might be expected to lead to unreliable hypothesis tests.



In general, there are two main consequences of neglecting the possible correlation: (i) incorrect inference about regression coefficients; (ii) estimates of regression coefficients that are inefficient, i.e. less precision than optimal.

(e) Extend the model in (a) to take into account correlation among repeated measurements on the same subject.

We may extend the model in (a) by adding various possible correlation structures: exchangeable, AR(1), or unstructured.

(f) Fit the model in (e) with as few assumptions as you can about the possible correlation structure, and conduct a test of null hypothesis in part (c). Do the results agree with those in part (c)? Give a possible explanation.
Among the possible correlation structures we used in part (e), the unstructured model has the fewest assumption about the structure of possible correlation. Table2 lists the regression results for this model (with robust variance estimation). Comparing the results of the independence model in Table 1, we see that except for X0, the inference for all the other parameters is essentially unchanged. We obtain a stronger and more significant association between wheezing and smoking as indicated by X0.

Table 2. GEE results with unstructured correlation matrix.

```
------------------------------------------------------------------------------
             |             Semi-robust
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x0 |  -.8193055   .4853743    -1.69   0.091    -1.770622    .1320106
          x1 |  -.8416823   .5060132    -1.66   0.096     -1.83345    .1500853
        cind |   .2001139    .411357     0.49   0.627     -.606131    1.006359
        time |  -.2144158   .1804719    -1.19   0.235    -.5681342    .1393027
       _cons |   1.903247   1.862532     1.02   0.307    -1.747248    5.553742
------------------------------------------------------------------------------
```

All three correlation structures yielded nearly the same parameter estimates and their standard errors. The largest difference was in the coefficient for X0, and even that was only about 10%. Therefore, it would be reasonable to choose the simplest of the three correlation structures, which is the exchangeable model.

(g) From inspection of result in (f), is a simpler model for correlation may be plausible? Select a most plausible model to fit the data and explain why choose it.

The estimated matrix of within-group correlation from the unstructured model is provided in Table 3. We found that the estimated correlation matrices from the exchangeable and AR(1) models were quite different from that in Table 3. Therefore, it appears that a simpler correlation model may not be appropriate here. So, we model the correlation using an unstructured correlation matrix.

Table 3. Estimated within-subject correlation matrix

```
          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
```

(g.1) As we already found in (f) there is insufficient evidence for smoking to be associated with wheezing.

(g.2) There is also insufficient evidence to suggest that it is worthwhile to take city into account, estimate = 0.20 (95% CI = -0.61, 1.01).

(h) From your fit in (g), provide an estimation of the probability that a child from Kingston whose mother is a heavy smoker, wheezes at the initial visit:

Note that even though there was insufficient evidence for including City and smoking variables, they are still kept in the model. Unless there are good scientific reasons for eliminating covariates, one should keep them in the model. Covariate selection should not be done only on the basis of statistical tests. Hence, our final model is:

$$\text{logit } \Pr(Y=1) = 1.90 + 0.20 * \text{city} - 0.21 * \text{time} - 0.82 * X0 - 0.84 * X1$$

We have time = 9, city = 1, X0 = 0, X1 = 0. Therefore, the estimated probability of this child with a respiratory infection is 0.55.

Provide an estimation of the probability that a child from Kingston whose mother does not smoke, wheezes at the initial visit:
We have time = 9, city = 1, X0 = 1, X1 = 0. Therefore, the estimated probability of this child with a respiratory infection is 0.35. It can be seen that maternal smoking increases the probability of wheezing for the child at baseline.

(i) First we fit a logistic regression model, where the wheezing at any time depends on past and present maternal smoking behavior, in addition to the city and time variables. We use an unstructured correlation model. Specifically, we created smoking variables with lag 1 to indicate past behavior. The results are given below in Table 4.

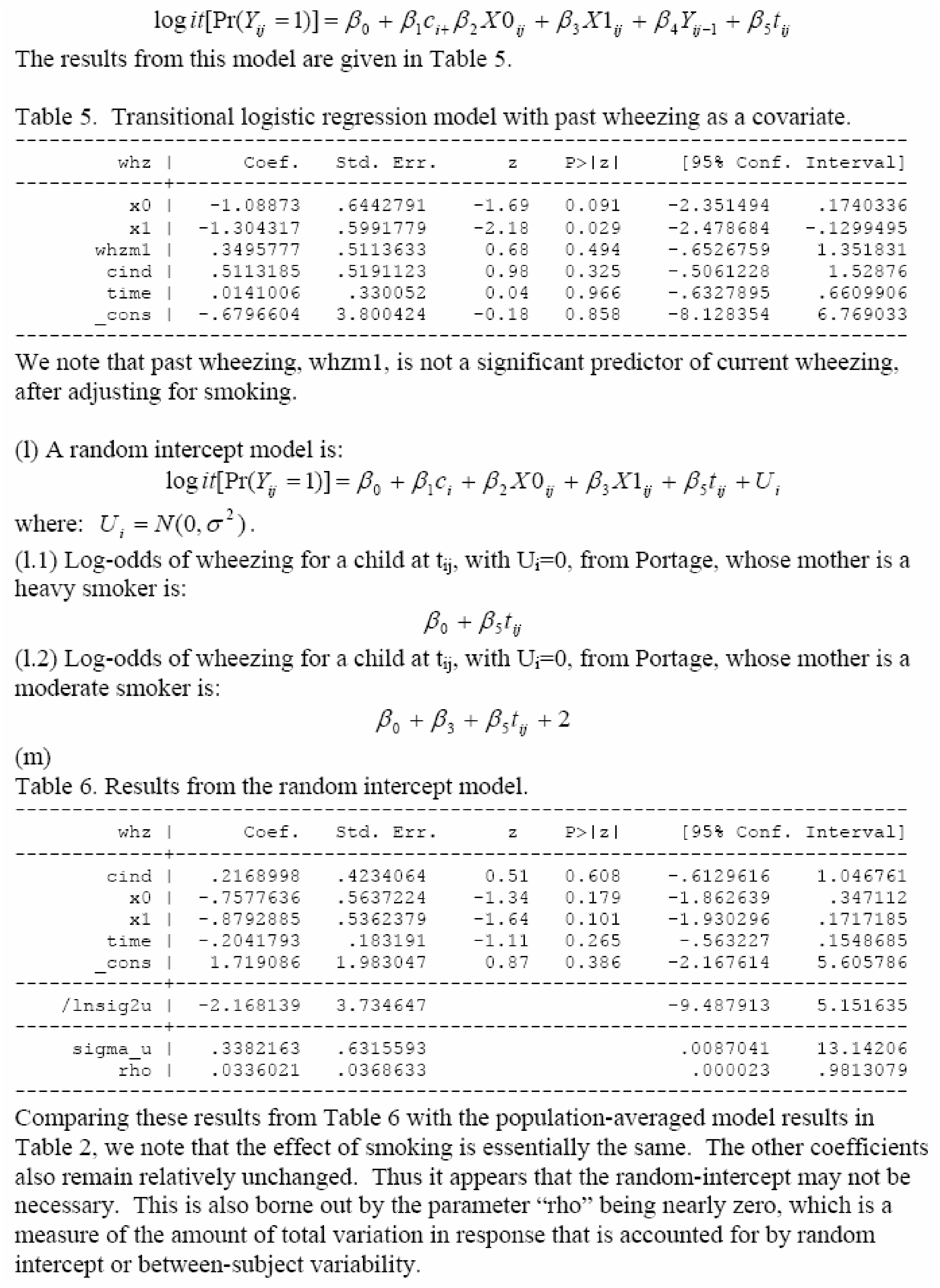
$$\text{logit } [\Pr(Y_{ij} = 1)] = \beta_0 + \beta_1 c_{i+} \beta_2 X0_{ij} + \beta_3 X1_{ij} + \beta_4 X0_{ij-1} + \beta_5 X1_{ij-1} + \beta_6 t_{ij}$$

Table 4. Logistic regression of wheezing on past and present smoking.

```
------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x0 |  -1.273384   .6339722    -2.01   0.045    -2.515947   -.0308217
          x1 |  -1.235653   .6262809    -1.97   0.048    -2.463141   -.0081652
        x0m1 |  -.0052262   .9997849    -0.01   0.996    -1.964769    1.954316
        x1m1 |  -.2210449    .810579    -0.27   0.785    -1.809751    1.367661
        cind |   .5702549    .524679     1.09   0.277     -.458097    1.598607
        time |  -.1071572   .3212241    -0.33   0.739    -.7367449    .5224306
       _cons |   .9027534   3.613741     0.25   0.803    -6.180049    7.985556
------------------------------------------------------------------------------
```

We see that the coefficients for past smoking history, x0m1 and x1m1, are not significantly different from zero, after adjusting for the other variables.

Next we fit a transitional logistic regression model, with wheezing status at the previous visit as a covariate. Here we fit an independent logistic regression model because the correlation between repeated measurements is implicitly taken into account by letting current wheezing status be dependent upon previous wheezing status. The model can be written as:



$$\log it[\Pr(Y_{ij}=1)] = \beta_0 + \beta_1 c_{i+} \beta_2 X0_{ij} + \beta_3 X1_{ij} + \beta_4 Y_{ij-1} + \beta_5 t_{ij}$$

The results from this model are given in Table 5.

Table 5. Transitional logistic regression model with past wheezing as a covariate.

```
------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x0 |   -1.08873   .6442791    -1.69   0.091    -2.351494    .1740336
          x1 |  -1.304317   .5991779    -2.18   0.029    -2.478684   -.1299495
       whzm1 |   .3495777   .5113633     0.68   0.494    -.6526759    1.351831
        cind |   .5113185   .5191123     0.98   0.325    -.5061228     1.52876
        time |   .0141006    .330052     0.04   0.966    -.6327895    .6609906
       _cons |  -.6796604   3.800424    -0.18   0.858    -8.128354    6.769033
------------------------------------------------------------------------------
```

We note that past wheezing, whzm1, is not a significant predictor of current wheezing, after adjusting for smoking.

(l) A random intercept model is:

$$\log it[\Pr(Y_{ij}=1)] = \beta_0 + \beta_1 c_i + \beta_2 X0_{ij} + \beta_3 X1_{ij} + \beta_5 t_{ij} + U_i$$

where: $U_i = N(0,\sigma^2)$.

(l.1) Log-odds of wheezing for a child at t_{ij}, with U_i=0, from Portage, whose mother is a heavy smoker is:

$$\beta_0 + \beta_5 t_{ij}$$

(l.2) Log-odds of wheezing for a child at t_{ij}, with U_i=0, from Portage, whose mother is a moderate smoker is:

$$\beta_0 + \beta_3 + \beta_5 t_{ij} + 2$$

(m)

Table 6. Results from the random intercept model.

```
------------------------------------------------------------------------------
         whz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cind |   .2168998   .4234064     0.51   0.608    -.6129616    1.046761
          x0 |  -.7577636   .5637224    -1.34   0.179    -1.862639     .347112
          x1 |  -.8792885   .5362379    -1.64   0.101    -1.930296    .1717185
        time |  -.2041793    .183191    -1.11   0.265     -.563227    .1548685
       _cons |   1.719086   1.983047     0.87   0.386    -2.167614    5.605786
-------------+----------------------------------------------------------------
    /lnsig2u |  -2.168139   3.734647                     -9.487913    5.151635
-------------+----------------------------------------------------------------
     sigma_u |   .3382163   .6315593                      .0087041    13.14206
         rho |   .0336021   .0368633                       .000023    .9813079
------------------------------------------------------------------------------
```

Comparing these results from Table 6 with the population-averaged model results in Table 2, we note that the effect of smoking is essentially the same. The other coefficients also remain relatively unchanged. Thus it appears that the random-intercept may not be necessary. This is also borne out by the parameter "rho" being nearly zero, which is a measure of the amount of total variation in response that is accounted for by random intercept or between-subject variability.