Biostatistics 140.655, 2017-18
QUIZ 4 Solution

Quiz Guidelines:

Please read the following quiz guidelines carefully:

- For this quiz, you are to work ALONE. You may use your course notes and lab materials to help answer the questions.
- Submit your answers to Courseplus by 5pm Friday March 9th
- DO NOT discuss this quiz or your solution to this quiz with other students from the course starting Wednesday March 7th through Friday March 9th. The solution to the quiz will be available Saturday, March 10th.
- By submitting your answers to Courseplus, you are acknowledging that you have read the guidelines carefully and will adhere to these guidelines.

Scientific Background:

This quiz is an analysis of data from the Supplemental Nutrition Assistance Program (SNAP) which is the new name for the federal Food Stamp Program. Data include county-level population size and SNAP enrollees for 3,205 US counties for years 2000 to 2010. The population size and SNAP enrollees are based on available data from July of the corresponding year. The outcome of interest is the proportion of the county population that is enrolled in the SNAP. In addition, a metro vs. non-metro indicator for each county is provided. There is no missing data.

Notation:

Y_{ij} is the proportion enrolled in SNAP for county i (i in 1, 2, …, 3205) during year j (j = 200, 2001, …, 2010), this value ranges between 0 and 100.

$year_{ij}$ is the year of observation; 2000 to 2010

$metro_i$ is the indicator that county i is a metro county (1) vs. a non-metro county (0)

Objectives:

- **PRIMARY OBJECTIVE:** to estimate the yearly trend in SNAP enrollment separately for metro and non-metro counties and determine if the recent recession (starting 2007) had a different effect on SNAP enrollment for metro vs. non-metro counties.
- **SECONDARY OBJECTIVE:** summarize the variation in the growth rates of SNAP enrollment across the counties.

Figures 1 and 2 below summarize the data across years and metro/non-metro designation.

Figure 1: Distribution of the proportion of county population enrolled in SNAP by year (2000 to 2010) and metro/non-metro designation.

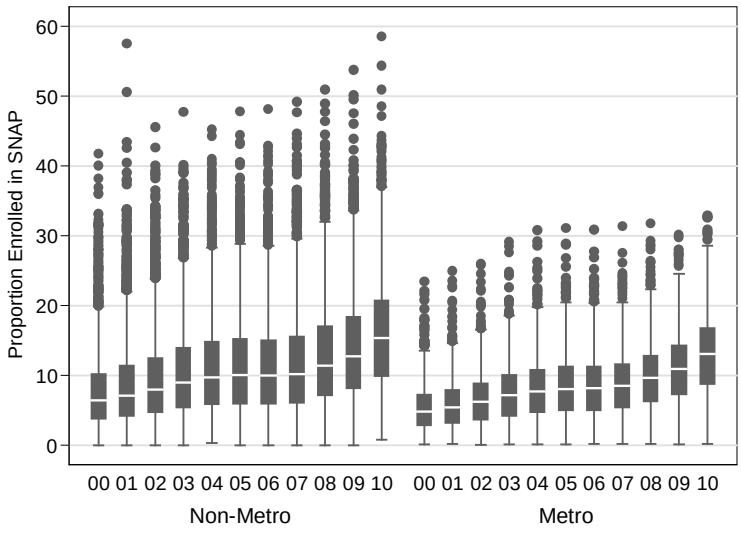
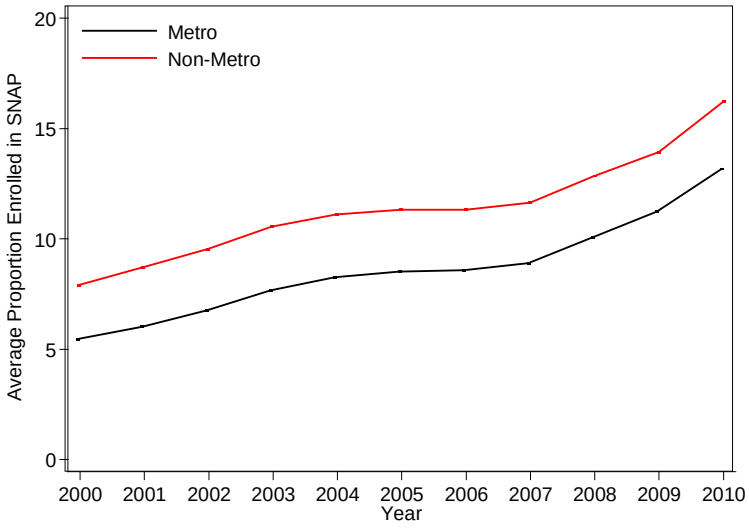


Figure 2: Average proportion enrolled in SNAP among metro/non-metro counties for years 2000 to 2010



Linear Mixed Model specification: To address the primary and secondary objectives of the study, we will fit a linear mixed model that allows for a county-specific estimate for the proportion enrolled in SNAP based on the county type (metro/non-metro county) and time (linear spline with knot at 2007). The county-specific estimates will include a random intercept and random slopes for both the linear and spline terms in the model. The model is given below:

$$Y_{ij} = \beta_{0i} + \beta_{1i}(year_{ij} - 2007) + \beta_{2i}(year_{ij} - 2007)^{+} + \beta_3 metro_i + \beta_4 metro_i \times (year_{ij} - 2007) + \beta_5 metro_i \times (year_{ij} - 2007)^{+} + \varepsilon_{ij}$$

where $(year_{ij} - 2007)^{+} = 0$ if $year_{ij} - 2007 < 0$ and $(year_{ij} - 2007)$ if $year_{ij} - 2007 \geq 0$, $metro_i = 1$ if county *i* is a metro county and 0 if county *i* is a non-metro county.

The random effects are defined as follows:

$$\beta_{0i} = \beta_0 + b_{0i}$$
$$\beta_{1i} = \beta_1 + b_{1i}$$
$$\beta_{2i} = \beta_2 + b_{2i}$$

$$\begin{pmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{21} & \sigma_2^2 \end{bmatrix} \right)$$

And we make the following independence assumptions for the within subject residuals and random effects.

$$\varepsilon_{ij} \sim Independent\ N(0, \sigma^2), Cov(b_{0i}, \varepsilon_{ij}) = 0, Cov(b_{1i}, \varepsilon_{ij}) = 0, Cov(b_{2i}, \varepsilon_{ij}) = 0$$

This model was fit using mixed in STATA and the following output was obtained:

```
Mixed-effects ML regression                     Number of obs      =     33774
Group variable: id                              Number of groups   =      3137

                                                Obs per group: min =         4
                                                               avg =      10.8
                                                               max =        11


                                                Wald chi2(5)       =   9607.74
Log likelihood = -63548.475                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       y_100 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      time_c |   .5678269   .0081568    69.61   0.000     .5518398     .583814
     time_sp |   .6246083   .0406965    15.35   0.000     .5448445     .704372
       metro |  -2.812391   .2764718   -10.17   0.000    -3.354266   -2.270516
             |
 metro#c.time_c |
          1  |  -.0201355   .0157611    -1.28   0.201    -.0510267    .0107556
             |
metro#c.time_sp |
          1  |   .1938918   .0785994     2.47   0.014     .0398397    .3479438
             |
       _cons |   12.24922   .1430656    85.62   0.000     11.96882    12.52963
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                  sd(time_c) |   .3583502   .0053946      .3479315    .3690809
                 sd(time_sp) |   1.601304   .0303115      1.542984     1.66183
                   sd(_cons) |   6.825124   .0869663      6.656784    6.997721
        corr(time_c,time_sp) |  -.2788328   .0213522     -.3201314   -.2364764
          corr(time_c,_cons) |   .6696656   .0111615      .6472047    .6909649
         corr(time_sp,_cons) |  -.2080311   .0208766     -.2485633   -.1667717
-----------------------------+------------------------------------------------
                sd(Residual) |   1.031977   .0046761      1.022853    1.041183
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(6) = 94931.65   Prob > chi2 = 0.0000


. lincom time_c + 1.metro#c.time_c

 ( 1)  [y_100]time_c + [y_100]1.metro#c.time_c = 0

------------------------------------------------------------------------------
       y_100 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .5480702   .0134278    40.82   0.000     .5217521    .5743883
------------------------------------------------------------------------------
```

1. The model we described and fit above assumes that
    a. $Var(Y_{ij})$ is constant with time
    b. $Var(Y_{ij})$ is constant with county type (metro vs. non-metro)
    c. $Var(Y_{ij})$ differs with time
    d. $Var(Y_{ij})$ differs by county type (metro vs. non-metro)

2. The estimated mean proportion enrolled in SNAP in 2007 in non-metro counties is 12.25 (SE: 0.14). The estimated standard deviation of the random intercept is 6.83 (SE: 0.09). Which of the following is the most accurate interpretation of these two values?
    a. Roughly 95% of non-metro counties in 2007 have expected mean proportion enrolled in SNAP ranging from 12.25 +/- 2 (0.14)
    b. Roughly 95% of non-metro counties in 2007 have expected mean proportion enrolled in SNAP ranging from 12.25 +/- 2 (6.83)
    c. Roughly 95% of non-metro counties in 2007 have expected mean proportion enrolled in SNAP ranging from 12.25 +/- 2 (0.09)
    d. Roughly 95% of non-metro counties in 2007 have expected mean proportion enrolled in SNAP ranging from 6.83 +/- 2 (0.09)

3. Prior to 2007, the proportion enrolled in SNAP among metro counties increases by 0.55% per year (95% CI: 0.52 to 0.57). There is considerable variation in the growth rates in SNAP enrollment across counties. The estimate of the standard deviation in the linear growth rate prior to 2007 in metro counties is:
    a. 0.358 (SE: 0.005)
    b. 1.601 (SE: 0.030)
    c. 6.825 (SE: 0.087)
    d. 0.670 (SE: 0.011)

4. We estimate that the correlation between the random intercept (b_{0i}) and the linear random slope (b_{1i}) is 0.67. This correlation can be interpreted as:
    a. Counties with higher than average growth in enrollment in SNAP prior to 2007 will tend to have higher than average enrollment in SNAP in 2007.
    b. Counties with higher than average growth in enrollment in SNAP prior to 2007 will tend to have lower than average enrollment in SNAP in 2007.
    c. Counties with lower than average growth in enrollment in SNAP prior to 2007 will tend to have higher than average enrollment in SNAP in 2007.
    d. This value is not important as it is not statistically significantly different from 0 and therefore, these two random effects are independent.

5. Suppose now that counties receive additional federal revenue to support social worker and child healthcare outreach if the proportion of the county population enrolled in SNAP exceeds 15 percent. NOTE: I'm making this up!

   You specified the random intercept logistic model below:

   $$log\left[\frac{P(Y_{ij} > 15)}{P(Y_{ij} \leq 15)}\right] = \beta_{0i} + \beta_1(year_{ij} - 2007) + \beta_2(year_{ij} - 2007)^+ + \beta_3 metro_i + \beta_4 metro_i \times (year_{ij} - 2007) + \beta_5 metro_i \times (year_{ij} - 2007)^+$$

   where $(year_{ij} - 2007)^+ = 0$ if $year_{ij} - 2007 < 0$ and $(year_{ij} - 2007)$ if $year_{ij} - 2007 \geq 0$, $metro_i = 1$ if county *i* is a metro county and 0 if county *i* is a non-metro county.

   The random effects are defined as follows: $\beta_{0i} = \beta_0 + b_{0i}, b_{0i} \sim N(0, \sigma_0^2)$

   The interpretation for β_1 is:

   a. The yearly change in the log odds of greater than 15 percent of the population enrolled in SNAP during 2000 to 2007 among all counties included in the study.
   b. The yearly change in the log odds of greater than 15 percent of the population enrolled in SNAP during 2000 to 2007 for a given/specific non-metro county.
   c. The odds of greater than 15 percent of the population enrolled in SNAP in 2007 among all counties included in the study.
   d. The odds of greater than 15 percent of the population enrolled in SNAP in 2007 for the "average" non-metro county.

6. If you were to add a random slope for $(year_{ij} - 2007)$ to the random intercept logistic regression model above; i.e. define $\beta_{1i} = \beta_1 + b_{1i}, b_{1i} \sim N(0, \sigma_1^2)$. The interpretation for β_1:

   a. cannot be determined by the information provided.
   b. remains the same; adding a random slope to the random intercept logistic regression model does not change any of the interpretations.
   c. changes; β_1 now represents the yearly change in the log odds of greater than 15 percent of the population enrolled in SNAP during 2000 to 2007 for the non-metro county with "average" yearly change.