Biostatistics 140.656, 2019-20

Final Exam

**Guidelines**:

1. Data Analysis
   1. You are encouraged to work on the data analysis in teams/groups.
   2. You may consult your notes, labs, homework solutions, textbooks and have questions answered via the discussion forum
2. Short Answer Questions
   1. You MUST write and submit your own answers to the short answer questions
   2. Please provide concise answers; some questions ask you to specify/define a model, others ask you answer a question (1 or 2 sentence answers should be enough).
3. Submission
   1. In your submission, please include an acknowledgement of any students that you collaborated with on the completion of the required data analysis.
   2. Be sure to provide your solution in 12-point font.
   3. Your solution should be submitted via Dropbox. Please check courseplus for deadlines.

**Scientific Background:**

You will be analyzing data from a survey conducted in Guatemala in 1987. The survey identified a nationally representative sample of 5160 mothers, between 15 and 44 years of age, with the primary purpose of understanding factors that could affect the immunization status of children who were born in the previous 5 years and alive at the time of the interview.

The data available represent 2,158 children (level 1) aged 1 – 4 years of age from 1595 mothers (level 2) from within 161 communities (level 3). The 2,158 children received at least one immunization. The outcome of interest is whether the child received the full set of immunizations.

In 1986, the government of Guatemala undertook a campaign to immunize the population against major childhood illnesses. The immunization campaign visited most of the country and often located children in their own households. The full set of immunizations at the time of the campaign included three doses of DPT vaccine (against diphtheria, whooping cough, and tetanus), three doses of polio vaccine, one dose of BCG (antituberculosis) if available. The campaign provided the immunizations free of charge on the following schedule in 1986: May 17-18, July 5-6 and August 16-17. As you can imagine, this campaign required a massive effort and staff; an evaluation of the campaign estimated that roughly 32,000 people, including 25,000 volunteers were involved.

Due to local mayoral elections that occurred in 1987, the government did not replicate the campaign; requiring families to seek out and pay for immunizations at local health care centers/providers. The 1986 campaign was largely successful in increasing the immunization coverage; but absent the immunization days, immunization coverage fell sharply in 1987. In subsequent years, immunization days were held regularly.

The survey conducted in 1987 offers an opportunity to evaluate the likelihood of children receiving the full set of immunization during the campaign (children of eligible ages the summer of 1986) and absent the campaign (children of eligible ages in 1987, who were not alive or who were not of eligible age in 1986); thus, creating a natural experiment. The key exposure variable from the 1987 survey is the indicator for whether the child was at least 2 years old at the time of the survey/interview, in which case the child was eligible to receive all immunizations during the campaign. If this variable is associated with immunization status, there is some indication that the campaign worked.

In this final exam, you will be conducting/interpreting a series of analyses to address questions that are relevant to government health officials both from the prospective of improving the immunization coverage rate but also to explore factors related to differences in the odds of full immunization that could provide insight into how to improve the design and implementation of future campaigns.

Objectives:

1. Absent the immunization campaign, estimate and interpret the impact of mother’s ethnicity (indigenous vs. not) on the odds of receiving the full course of immunizations and the contextual effect of mother’s ethnicity. This information may promote the use of different strategies for campaign implementation across communities.
2. Estimate and interpret the population average/marginal effect of the immunization campaign.
3. Estimate the heterogeneity of the effect of the immunization campaign across the Guatemalan communities and determine if community characteristics explain observed heterogeneity. Similar to objective 1, this data may be used to modify future campaigns based on community characteristics to improve impact.

**Data:**

The data set is called **guatemala.csv,** which can be downloaded directly from our website. The dataset contains children *k* nested in mother *j* nested in community *i*. It contains the following subset of variables.

**Level 1 (children)**

* immun: dummy variable for child being immunized, the response variable.
* kid2p: child at least 2 years old at the time of the interview (indicator for exposure to the campaign or not)

**Level 2 (mother)**

* mom: identifier for mother
* Ind: Indigenous Ethnicity (indigenous vs. not)

**Level 3 (community)**

* cluster: identifier for communities
* rural: dummy variable for community being rural
* pcInd: percent of indigenous mothers in the community (calculated variable, ranges from 0 to

**Brief EDA of the hierarchical clustering:**

How many communities are in the study, how many mothers and how many children?

. codebook cluster mom kid

----------------------------------------------------------------------------------------------------

cluster (unlabeled)

----------------------------------------------------------------------------------------------------

type: numeric (int)

range: [1,240] units: 1

unique values: 161 missing .: 0/2,158

mean: 145.858

std. dev: 59.3406

percentiles: 10% 25% 50% 75% 90%

63 94 148 202 226

----------------------------------------------------------------------------------------------------

mom (unlabeled)

----------------------------------------------------------------------------------------------------

type: numeric (int)

range: [2,2782] units: 1

unique values: 1,595 missing .: 0/2,158

mean: 1502.63

std. dev: 751.435

percentiles: 10% 25% 50% 75% 90%

498 859 1471.5 2208 2571

----------------------------------------------------------------------------------------------------

kid (unlabeled)

----------------------------------------------------------------------------------------------------

type: numeric (int)

range: [2,4627] units: 1

unique values: 2,158 missing .: 0/2,158

mean: 2445.98

std. dev: 1267.71

percentiles: 10% 25% 50% 75% 90%

775 1346 2373.5 3636 4255

We have 161 communities; 1595 mothers and 2158 children.

How many mothers in each community?

bys cluster mom: gen kid\_counter = \_n

bys cluster: egen junk = count(mom) if kid\_counter==1

bys cluster: egen num\_moms = min(junk)

drop junk

bys cluster: gen cluster\_counter = \_n

summ num\_moms if cluster\_counter==1

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

num\_moms | 161 9.906832 6.022667 1 37

The average number of mothers per community was roughly 10 with a range from 1 to 37.

How many children in each community?

. bys cluster: egen cluster\_kids = count(cluster)

. summ cluster\_kids if cluster\_counter==1

Variable | Obs Mean Std. Dev. Min Max

-------------+---------------------------------------------------------

cluster\_kids | 161 13.40373 8.750271 1 55

The number of children per community ranges from 1 to 55. Average number of children per community is 13.

How many children per mother?

. tab num\_kids if kid\_counter==1

num\_kids | Freq. Percent Cum.

------------+-----------------------------------

1 | 1,063 66.65 66.65

2 | 501 31.41 98.06

3 | 31 1.94 100.00

------------+-----------------------------------

Total | 1,595 100.00

The number of children per mother ranges from 1 to 3 with 67% of the mothers contributing data from a single child.

**Brief EDA of the primary outcome**

What is the sample prevalence of immunization?

. summ immun

Variable | Obs Mean Std. Dev. Min Max

-------------+---------------------------------------------------------

immun | 2,158 **.4467099** .4972673 0 1

What is the sample prevalence of immunization by study period (during or post campaign)?

-----------------------------------------------------------------------

-> kid2p = 0

Variable | Obs Mean Std. Dev. Min Max

-------------+---------------------------------------------------------

immun | 492 **.2845528** .4516604 0 1

-----------------------------------------------------------------------

-> kid2p = 1

Variable | Obs Mean Std. Dev. Min Max

-------------+---------------------------------------------------------

immun | 1,666 **.4945978** .5001209 0 1

**Objective 1:** Use the data on immunization of eligible children absent the campaign in 1987 (i.e. *kid2p* = 0) to determine if:

1. within a community, there is a difference in the odds of a child receiving the full course of immunizations comparing mothers of indigenous vs. non-indigenous ethnicity
2. mother’s ethnicity acts as a contextual effect on the odds of completing the full course of immunizations

NOTE: The data from 1987, absent the campaign, represents 2-level data since each mother only contributes one child in the pre-campaign time period. Therefore, the communities are the level-2 units and the mother/child pairs are the level-1 units.

1. Write out a single marginal model that addresses the two goals of this objective. Be sure to include an assumption about the correlation between the binary indicator of completing the full course of immunization for two moms from the same community. For each parameter (i.e. ) in the model, provide a written interpretation. Be sure to indicate which parameter or linear combination of parameters will address objective a) and b).
2. Fit the model and present your answers for objectives a) and b). Be sure to include the estimates and statistical evidence (i.e. confidence intervals; p-values optional).

**Objective 2:** Define the marginal effect of the immunization campaign as the ratio of the odds of completing the full course of immunizations for children who were exposed to the campaign and the odds of completing the full course of immunizations for children who were not exposed to the campaign,

.

where and are the binary indicators for completing the full course of immunizations and exposure to the campaign, respectively, for child *k* from mother *j* nested within community *i.*

Your objective is to estimate this marginal effect and perform a hypothesis test of the null hypothesis that the marginal effect is 1.

We can estimate this marginal effect by fitting the following logistic regression model:

An unbiased estimate of can be obtained by fitting this logistic regression model assuming the observations in the data are independent (i.e. working independence assumption). However, to obtain a valid hypothesis test, we must account for the correlation in the data. Given that the data represents 3 levels of nesting, we will not be able to apply the generalized estimating equations approach. Hence, we will utilize a bootstrap procedure.

Run the Stata or R code that is provided (see final.do and final.R).

1. Interpret the estimate of the marginal effect and provide statistical evidence for or against the null hypothesis.
2. If we ignore the correlation in the data and fit a logistic regression model assuming independence of the kids within mom and moms within communities, we estimate the marginal effect to be 2.46 with 95% CI for the marginal effect (1.98, 3.06). Compared to this interval (assuming independence), the bootstrap BCa confidence interval for the marginal effect is wider. Is this what you expect and why?
3. Suppose that we wanted to define the marginal effect of the campaign as with corresponding statistical test of the null hypothesis that the marginal effect is 0. **In words** (you do NOT have to implement this), describe how you would obtain an estimate of this marginal effect and how you could conduct the hypothesis test using a bootstrap procedure.

**Objective 3:** The goal of Objective 3 is to estimate the heterogeneity in the effect of the campaign across the communities in Guatemala and determine if two community level characteristics explain the observed heterogeneity.

You consider the following model:

The fit of this model is presented below:

meqrlogit immun kid2p || cluster: kid2p, cov(uns) || mom: , intp(12)

Mixed-effects logistic regression Number of obs = 2,158

----------------------------------------------------------------------------

| No. of Observations per Group Integration

Group Variable | Groups Minimum Average Maximum Points

----------------+-----------------------------------------------------------

cluster | 161 1 13.4 55 12

mom | 1,595 1 1.4 3 12

----------------------------------------------------------------------------

Wald chi2(1) = 42.50

Log likelihood = -1348.8518 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

immun | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

kid2p | 1.933637 .2966054 6.52 0.000 1.352302 2.514973

\_cons | -1.986354 .3108885 -6.39 0.000 -2.595684 -1.377024

------------------------------------------------------------------------------

------------------------------------------------------------------------------

Random-effects Parameters | Estimate Std. Err. [95% Conf. Interval]

-----------------------------+------------------------------------------------

cluster: Unstructured |

var(kid2p) | 1.978795 1.064116 .6896979 5.677311

var(\_cons) | 3.671978 1.461849 1.682786 8.012563

cov(kid2p,\_cons) | -1.981638 1.14151 -4.218957 .2556801

-----------------------------+------------------------------------------------

mom: Identity |

var(\_cons) | 5.889154 1.3969 3.699551 9.37469

------------------------------------------------------------------------------

LR test vs. logistic model: chi2(4) = 199.32 Prob > chi2 = 0.0000

. meqrlogit, or

Mixed-effects logistic regression Number of obs = 2,158

----------------------------------------------------------------------------

| No. of Observations per Group Integration

Group Variable | Groups Minimum Average Maximum Points

----------------+-----------------------------------------------------------

cluster | 161 1 13.4 55 12

mom | 1,595 1 1.4 3 12

----------------------------------------------------------------------------

Wald chi2(1) = 42.50

Log likelihood = -1348.8518 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

immun | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

kid2p | 6.914616 2.050912 6.52 0.000 3.866314 12.36628

\_cons | .1371947 .0426523 -6.39 0.000 .0745948 .2523284

------------------------------------------------------------------------------

Note: \_cons estimates baseline odds (conditional on zero random effects).

------------------------------------------------------------------------------

Random-effects Parameters | Estimate Std. Err. [95% Conf. Interval]

-----------------------------+------------------------------------------------

cluster: Unstructured |

var(kid2p) | 1.978795 1.064116 .6896979 5.677311

var(\_cons) | 3.671978 1.461849 1.682786 8.012563

cov(kid2p,\_cons) | -1.981638 1.14151 -4.218957 .2556801

-----------------------------+------------------------------------------------

mom: Identity |

var(\_cons) | 5.889154 1.3969 3.699551 9.37469

------------------------------------------------------------------------------

LR test vs. logistic model: chi2(4) = 199.32 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

1. Interpret the estimated value of .
2. Provide an interval that contains the effect of the campaign for 95% of the communities in Guatemala.
3. Estimate the and interpret this value within the context of the problem

You then extend this model to include information about the proportion of indigenous mothers in the community and the rural vs. urban status. Specifically, you consider the following model:

The fit of this model is displayed below:

. meqrlogit immun kid2p rural pcind kid2p\_rural kid2p\_pcind ///

> || cluster: kid2p , cov(uns) || mom: , intp(10)

Mixed-effects logistic regression Number of obs = 2,158

----------------------------------------------------------------------------

| No. of Observations per Group Integration

Group Variable | Groups Minimum Average Maximum Points

----------------+-----------------------------------------------------------

cluster | 161 1 13.4 55 10

mom | 1,595 1 1.4 3 10

----------------------------------------------------------------------------

Wald chi2(5) = 55.07

Log likelihood = -1333.9669 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

immun | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

kid2p | 1.58227 .5150844 3.07 0.002 .5727228 2.591816

rural | -1.330542 .5638842 -2.36 0.018 -2.435735 -.2253497

pcind | -1.526359 .6188272 -2.47 0.014 -2.739238 -.3134794

kid2p\_rural | .2823882 .5507157 0.51 0.608 -.7969947 1.361771

kid2p\_pcind | .4056526 .5930128 0.68 0.494 -.756631 1.567936

\_cons | -.4667831 .5012935 -0.93 0.352 -1.4493 .515734

------------------------------------------------------------------------------

------------------------------------------------------------------------------

Random-effects Parameters | Estimate Std. Err. [95% Conf. Interval]

-----------------------------+------------------------------------------------

cluster: Unstructured |

var(kid2p) | 1.927311 1.045977 .6652638 5.583543

var(\_cons) | 3.051023 1.287865 1.333968 6.978235

cov(kid2p,\_cons) | -1.837483 1.067291 -3.929334 .254368

-----------------------------+------------------------------------------------

mom: Identity |

var(\_cons) | 5.870561 1.396145 3.683373 9.356502

------------------------------------------------------------------------------

LR test vs. logistic model: chi2(4) = 175.08 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

. meqrlogit, or

Mixed-effects logistic regression Number of obs = 2,158

----------------------------------------------------------------------------

| No. of Observations per Group Integration

Group Variable | Groups Minimum Average Maximum Points

----------------+-----------------------------------------------------------

cluster | 161 1 13.4 55 10

mom | 1,595 1 1.4 3 10

----------------------------------------------------------------------------

Wald chi2(5) = 55.07

Log likelihood = -1333.9669 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

immun | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

kid2p | 4.865987 2.506394 3.07 0.002 1.773088 13.35401

rural | .2643339 .1490537 -2.36 0.018 .0875334 .798237

pcind | .2173256 .134487 -2.47 0.014 .0646196 .7308994

kid2p\_rural | 1.326293 .7304106 0.51 0.608 .4506813 3.9031

kid2p\_pcind | 1.500281 .889686 0.68 0.494 .4692446 4.796739

\_cons | .627016 .314319 -0.93 0.352 .2347345 1.674867

------------------------------------------------------------------------------

Note: \_cons estimates baseline odds (conditional on zero random effects).

------------------------------------------------------------------------------

Random-effects Parameters | Estimate Std. Err. [95% Conf. Interval]

-----------------------------+------------------------------------------------

cluster: Unstructured |

var(kid2p) | 1.927311 1.045977 .6652638 5.583543

var(\_cons) | 3.051023 1.287865 1.333968 6.978235

cov(kid2p,\_cons) | -1.837483 1.067291 -3.929334 .254368

-----------------------------+------------------------------------------------

mom: Identity |

var(\_cons) | 5.870561 1.396145 3.683373 9.356502

------------------------------------------------------------------------------

LR test vs. logistic model: chi2(4) = 175.08 Prob > chi2 = 0.0000

1. Re-write the model using the sequential model notation; i.e. specify a model for the level-1 outcome, specify a model for parameters defined at level 2, and finally specify a model for parameters defined at level 3.
2. Interpret the estimated value of
3. Interpret the estimated value of
4. Argue for or against the following: Rural vs. urban status and the proportion of indigenous mothers within a community are factors that should be considered in design and implementation of future immunization campaigns.