**BIOST 2050: Longitudinal and Clustered Data Analysis**

**Homework Assignment #1**

**Due: Wednesday, September 20, 2023 by 11:59pm**

Answer the questions and justify your answers. Note that you will receive a major deduction if you answer a question by giving only the software output without justifying your answer.

Until now, most pediatric studies examining the relationship between asthma and COVID-19 have been ecological and provided limited insights. To enrich this understanding, an observational study was conducted to evaluate the association at an individual level. The study population consists of children and adolescents aged 2 to 21 years who were diagnosed and hospitalized with asthma.

In this **cross-sectional** dataset, the primary interest of the investigators was to determine whether asthma patients diagnosed with COVID-19 experienced more severe asthma outcomes compared to their counterparts without a COVID-19 diagnosis. In this context, a “more severe asthma outcome” is defined as longer hospital stay.

The dataset asthma\_cross (available as asthma\_cross.csv for free format, and asthma\_cross.dta for Stata 17) is a cross-sectional dataset containing variables specifically relevant to the cross-sectional research aim. Descriptions of these variables can be found in the table provided below.

A separate cohort of asthma-diagnosed patients with multiple hospital/clinic visits was obtained. In this **longitudinal** data set, the primary research questions were: (1) whether receiving an influenza vaccination is associated with improved asthma outcomes, specifically in the context of emergency department (ED) visit due to asthma exacerbation, and (2) whether racial disparities impact this association.

The asthma\_long dataset (available as asthma\_long.csv for free format, and asthma\_long.dta for Stata 17) is a longitudinal dataset containing variables specifically tailored to address longitudinal research questions. Descriptions of these variables can be found in the table provided below.

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| ***asthma\_cross (cross-sectional data set)*** | |
| personid | patient’s unique identification number |
| patientage | patient’s age (in years) |
| female | 0=male, 1=female |
| black | 0=white, 1=black |
| currentcovid | the patient tested positive for COVID-19 |
| coexisting | the patient has other respiratory disorders coexisting (0=no, 1=yes) |
| influenzavaccine | whether the patient had a flu shot for that season (0=no, 1=yes) |
| los | length of hospital stay (in hours) |
| ***asthma\_long (longitudinal data set)*** | |
| personid | patient’s unique identification number |
| encounterid | hospital/clinic visit (1, 2, 3, 4, 5, 6, 7) |
| patientage | patient’s age (in years) |
| eosinophils | lab result for eosinophil counts in blood. eosinophile is a type of white blood cell |
| eosin03 | eosinophils >=0.3/L (0=no, 1=yes) |
| fev1fvc | ratio of FEV1/FVC, which is a pulmonary function test |
| fev1fvc75 | fev1fvc >=75% (0=no, 1=yes) |
| black | patient’s race/ethnicity (0=white, 1=black) |
| influenzavaccine | whether the patient had a flu shot for that season (0=no, 1=yes) |
| coexisting | whether the patient has other respiratory disorders coexisting (0=no, 1=yes) |
| exacerbationed | Emergency Department (ED) visit due to asthma exacerbation (0=no, 1=yes) |

1. Descriptive statistical analysis for **cross-sectional** data. (Total 14 pts)

Complete Table 1 below, describing the distributions of all relevant variables (including the outcome variable) for asthma patients with and without COVID-19. For each entry, provide descriptions of the values, such as the minimum or maximum of variable X. Please also provide inference test results and justifications of the test chosen for comparing between asthma patients with and without COVID-19 diagnosis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **All**  **(N= 178)** | **Without COVID-19**  **(n= 120)** | **With COVID-19**  **(n=58 )** | **P-value\*** |
| Patient age (in years) | 7.43 (4.31)  Mean (sd) | 7.28 (4.23) | 7.72 (4.47) | t = -0.63  p = 0.53 |
| Female sex | 41.57% | 40.83% Female | 43.10% Female | X2 = 0.016  p = 0.90 |
| Black race | 47.19% | 45% Black | 51.72% Black | X2 = 0.47  p = 0.50 |
| Other respiratory disorders coexisting | 8.99& | 6.67% with Comorbidity | 13.79% with Comorbidity | X2 = 1.63  p = 0.20 |
| Influenza vaccine | 50% | 43.33% vaccinated | 63.79% vaccinated | X2 = 5.7543  p = 0.016\* |
| Length of hospital stay (in hours) | 37.38 (24.69)  Mean (sd) | 32.98 (22.194) | 46.49 (27.20) | t = -3.29  p = 0.001\* |
| \* = <0.05 | | | | |

Welch’s Two Sample t-tests were performed for Patient Age by COVID-19 Status and Length of Hospital Stay (in hours) by COVID-19 Status, since this is an appropriate test for comparing two independent sample means. Chi-squared tests of homogeneity were conducted to test if the distributions of female/male sex, Black/non-Black race, Comorbid/Non-comorbid, Vaccinated/Unvaccinated, were the same when sampling from Asthma patients with or without COVID-19.

1. Answer the following questions pertaining to the primary objectives of the **cross-sectional** phase of the study. (Total 26 pts)
   1. **Fit a linear regression model for the length of hospital stay, with the primary predictor being whether the patient had COVID-19. Potential adjusting covariates include the patient’s age, sex, race, presence of other coexisting respiratory disorders, having an influenza vaccination, and their interactions with the primary predictor. Note: To decide the optimal model, you will need to (1) find an appropriate transformation of the outcome variable, and (2) determine the best set of covariates by performing variable selection. (12 pts)**

I used a natural log transformation of the length of stay, which resolved the right-skew in the length of stay in the dataset. I used backward selection to identify the best set of main effect covariates:

sw regress log\_los patientage female black currentcovid coexisting influenzavaccine, pr(

> 0.05)

begin with full model

p = 0.8243 >= 0.0500 removing influenzavaccine

p = 0.1733 >= 0.0500 removing female

p = 0.1519 >= 0.0500 removing black

Source | SS df MS Number of obs = 178

-------------+---------------------------------- F(3, 174) = 12.13

Model | 9.03759098 3 3.01253033 Prob > F = 0.0000

Residual | 43.2055408 174 .248307706 R-squared = 0.1730

-------------+---------------------------------- Adj R-squared = 0.1587

Total | 52.2431318 177 .295158937 Root MSE = .4983

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

log\_los | Coefficient Std. err. t P>|t| [95% conf. interval]

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

patientage | .0273181 .0087153 3.13 0.002 .0101167 .0445194

coexisting | .2800072 .1316212 2.13 0.035 .0202277 .5397868

currentcovid | .3366767 .0803109 4.19 0.000 .1781678 .4951856

\_cons | 3.124417 .0782588 39.92 0.000 2.969959 3.278876

I then used backward selection with pairwise interactions of covariates that had significant main effects

sw regress log\_los (patientage currentcovid coexisting) agecoexisting agecovid coexistin

> gcovid, lockterm1 pr(0.05)

begin with full model

p = 0.9850 >= 0.0500 removing agecoexisting

p = 0.1420 >= 0.0500 removing agecovid

p = 0.1632 >= 0.0500 removing coexistingcovid

Source | SS df MS Number of obs = 178

-------------+---------------------------------- F(3, 174) = 12.13

Model | 9.03759098 3 3.01253033 Prob > F = 0.0000

Residual | 43.2055408 174 .248307706 R-squared = 0.1730

-------------+---------------------------------- Adj R-squared = 0.1587

Total | 52.2431318 177 .295158937 Root MSE = .4983

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

log\_los | Coefficient Std. err. t P>|t| [95% conf. interval]

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

patientage | .0273181 .0087153 3.13 0.002 .0101167 .0445194

currentcovid | .3366767 .0803109 4.19 0.000 .1781678 .4951856

coexisting | .2800072 .1316212 2.13 0.035 .0202277 .5397868

\_cons | 3.124417 .0782588 39.92 0.000 2.969959 3.278876

The final model formula is log(los) = beta0 + 0.027\*age + .337\*currentcovid + .28\*coexisting + error

* 1. For your model in 2(a), write down the formulas for the predicted outcome for patients who test either positive or negative for COVID-19. Note: plug in appropriate values for other adjusting covariates in this formula. (8 pts)

For simplicity, assume coexisting == 0 and age == 7

Exp(Log(los)\_estimate) = exp(3.124 + 0.027\*(7 years old) + 0.337) = exp(3.65) = 38.47 hours for COVID-19 positive

Exp(Log(los)\_estimate) = exp(3.124 + 0.027 \* (7 years old) = exp(3.313) = 27.47 hours for COVID-19 negative

* 1. Based on your model in 2(a), do older children tend to have longer lengths of hospital stay compared with younger children? Does that depend on whether they have COVID-19 or not? (6 pts)

Based on a significance value of p < .05, the main effect of age on length of hospital stay is significantly different from zero, so I reject the null hypothesis that there is no effect of age on length of hospital stay. This main effect is positive, indicating that older children have longer hospital stays. There is no statistical reason to reject the null hypothesis that there is no interaction between age and COVID status because the interaction between age and COVID status was non-significant when the model was fit with the interaction term. However, COVID status does increase length of hospital stay, when holding age and comorbidity-presence constant.

**BIOST 2050: Longitudinal and Clustered Data Analysis**

**Homework Assignment #1 (part 2)**

**Due: Friday, September 22, 2023**

Answer the questions and justify your answers. Note that you will receive a major deduction if you answer a question by giving only the software output without justifying your answer.

Enterprise zones (EZs) are geographic areas in which companies can qualify for a variety of incentives, subsidies, or tax credits. EZ programs are established to encourage businesses to remain, locate, or expand in areas with high unemployment and significant poverty, thereby aiding in the revitalization of these depressed regions.

A study was conducted to analyze panel data concerning an enterprise zone program in the state of Indiana. Data were gathered from 22 area unemployment claims offices from 1980 to 1988. Of these, six were designated as EZs starting in 1984 and an additional four began their EZ designation in 1985. The remaining 12 unemployment claims offices did not have an EZ in their area during the study's timeframe.

A region was selected as an EZ based on several factors: the population and geographic size of the area, the economic indicators (such as unemployment and poverty rates), and the decision of the state government. An EZ designation is typically intended to last 10 years.

Analytic dataset enterprise.dta contains the following variables:

**Variable Label**

area: unemployment claims area identifier ()

year: calendar year ()

uclms: number of unemployment claims ()

uclms1: baseline number of unemployment claims ()

ez: designation as an enterprise zones (0=no, 1=yes) ()

time:time 1, …, 9 () corresponding to calendar years 1980-1988

1. Let be the number of unemployment claims for area () at calendar time (). Fit a model to investigate the first study aim: **relationship between the number of unemployment claims and calendar year; and how the relationship varies when comparing areas included in the enterprise zones with those that are not included in an enterprise zone.**
   1. Write down the formula of the model assuming random area-specific intercepts. Explicitly specify the fixed effects and the random effects, if any, and justify your choices. (12 pts)

uclmsij – uclms1j = beta\_0 + zeta\_j + beta\_1\*time + beta\_2\*ez + beta\_3\*ez\*time + episilonij

where zeta\_j = random-intercept, and betas are fixed intercepts/effects

* 1. **Fit the model in 3(a) using a statistical package. (8 pts)**

> ri\_fit\_mle <- lmerTest::lmer(uclms\_diff ~ 1 + time\*ez + (1 | area), data = enterprise, REML = FALSE)

> summary(ri\_fit\_mle)

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: uclms\_diff ~ 1 + time \* ez + (1 | area)

Data: enterprise

AIC BIC logLik deviance df.resid

4803.4 4823.2 -2395.7 4791.4 192

Scaled residuals:

Min 1Q Median 3Q Max

-4.0392 -0.5510 -0.0138 0.4257 5.8937

Random effects:

Groups Name Variance Std.Dev.

area (Intercept) 1.114e+09 33381

Residual 1.510e+09 38856

Number of obs: 198, groups: area, 22

Fixed effects:

Estimate Std. Error df t value Pr(>|t|)

(Intercept) 17394.31 9568.08 51.64 1.818 0.0749 .

time -12025.06 1354.70 187.13 -8.877 5.56e-16 \*\*\*

ez -32478.06 32214.91 177.59 -1.008 0.3147

time:ez 3235.56 4507.24 177.48 0.718 0.4738

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) time ez

time -0.571

ez -0.120 0.123

time:ez 0.172 -0.297 -0.955

* 1. **Based on the fitted model in 3(b), is there a time effect? Describe it and specify the null and alternative hypotheses associated with the test. Note: please describe the time effect thoroughly (that is, interpret your finding and not just simply say “there is a (or there is no) significant time effect based on the p value <0.05 (or ≥0.05).” (6 pts)**

Yes, there is a significant time effect. Assuming panel data, each increment in time (1 year) isis associated with a 12,025 person decrease in the expected difference of unemployment claims relative to baseline unemployment claims.

* 1. **Does the relationship between the number of unemployment claims and time vary when comparing areas included in the enterprise zones with those that are not included in an enterprise zone? Answer this question by specifying the null and alternative hypotheses and the associated test results. (6 pts)**

The null hypothesis is that the fixed effect (beta\_3) coefficient of the time by enterprise zone interaction is zero. The alternative hypothesis is that this coefficient is not equal to zero. This can be tested with a t-test, where t = beta\_3\_estimate / standard\_error(beta\_3\_estimate) given in the output. This statistical test does not reject the null hypothesis in my model.

* 1. **Based on the fitted model in 3(b), how much of the total variation in the number of unemployment claims is attributable to area variability? (6 pts)**

ri\_fit\_reml <- lmerTest::lmer(uclms\_diff ~ 1 + time\*ez + (1 | area), data = enterprise)

summary(ri\_fit\_reml)

performance::icc(ri\_fit\_reml)

# Intraclass Correlation Coefficient

Adjusted ICC: 0.434

Unadjusted ICC: 0.312

The proportion of total variance in change in number of unemployment claims due to inter-area variation is 43.4%

1. Fit a model to investigate the second study aim: **whether the inclusion in the enterprise zones would reduce the number of unemployment claims.**
   1. Write down the formula of the model assuming random area-specific intercepts, random area-specific time effects, and possible intercept-slope covariance. (12 pts)

uclmsij – uclms1j = beta\_0 + zeta\_0j + beta\_1\*timeij + zeta\_1j\*time\_ij + beta\_2\*ezij + beta\_3\*ezij\*time + episilonij

where zeta\_0j = random-intercept for area

where zeta\_1j = random area-specific time effect

* 1. **Fit the model using a statistical package. (8 pts)**

> rc\_fit\_mle <- lmerTest::lmer(uclms\_diff ~ 1 + time\*ez + (1 + time || area), data = enterprise, REML = FALSE)

> summary(rc\_fit\_mle)

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: uclms\_diff ~ 1 + time \* ez + (1 + time || area)

Data: enterprise

AIC BIC logLik deviance df.resid

4761.5 4784.5 -2373.7 4747.5 191

Scaled residuals:

Min 1Q Median 3Q Max

-1.7673 -0.3899 -0.0312 0.2883 6.2220

Random effects:

Groups Name Variance Std.Dev.

area (Intercept) 1.892e+08 13754

area.1 time 4.636e+07 6809

Residual 1.096e+09 33103

Number of obs: 198, groups: area, 22

Fixed effects:

Estimate Std. Error df t value Pr(>|t|)

(Intercept) 15239.35 6509.20 44.31 2.341 0.0238 \*

time -11119.33 2043.41 45.81 -5.442 1.99e-06 \*\*\*

ez -27597.75 27564.70 165.32 -1.001 0.3182

time:ez 1131.08 4267.55 191.16 0.265 0.7913

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) time ez

time -0.535

ez -0.192 0.128

time:ez 0.331 -0.337 -0.908

* 1. **For a specific calendar year, estimate the conditional covariance of the number of unemployment claims between two different areas and . (6 pts)**

The unemployment claims in two different areas, given their covariates, are independent for any given calendar year. Therefore, their conditional covariance is zero.

* 1. **Based on the fitted model in 4(b), would inclusion in the enterprise zones reduce the number of unemployment claims? Answer this question by specifying the null and alternative hypotheses associated and the associated test results. (6 pts)**

The null hypothesis is that the fixed effect (beta\_2) coefficient of enterprise zones is zero (i.e., there is no effect of enterprise zones on expected difference in unemployment claims). The alternative hypothesis is that this coefficient is not equal to zero. This can be tested with a t-test, where t = beta\_2\_estimate / standard\_error(beta\_2\_estimate) given in the model summary in b). This statistical test does not reject the null hypothesis in my model. This result holds regardless of whether the interaction term between time and ez is included or excluded in the model.