Homework 6

Due Monday 11/22

Turn in R code and output – try to use `R markdown`.

Qinyuan Xing, Min Yu, and Wenmin Zhang will lead HW discussion.

1. (60 points) *Modeling 7-Day COVID-19 Case Count*

   The dataset for this question uses the data from HW4: *covid_data_pdb_2.csv*. You will carry out data analysis modeling the 7-day COVID-19 case count in a county defined as:

   $$y = covid\_count\_sep17 - covid\_count\_sep10.$$

   Set $y = 0$ for the two counties with $y < 0$. Use the following variables as predictors:

   - $LAND\_AREA$: land area in square miles;
   - $pct\_URBANIZED\_AREA\_POP\_CEN\_2010$: percent of urbanized land;
   - $pct\_Males\_ACS\_14\_18$: percent males;
   - $pct\_Pop\_**\_ACS\_14\_18$: percent ** years old where ** = 5–17, 18–24, 25–44, 45–64, 65 plus;
   - $pct\_Inst\_GQ\_CEN\_2010$: percent in group quarters (e.g. dorms, prisons);
   - $pct\_Hispanic\_ACS\_14\_18$: percent Hispanic origin;
   - $pct\_NH\_Blk\_alone\_ACS\_14\_18$: percent non-Hispanic Black;
   - $pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18$: percent in poverty;
   - $pct\_Not\_HS\_Grad\_ACS\_14\_18$: percent not high school graduates;
   - $pct\_College\_ACS\_14\_18$: percent college graduates;
   - $pct\_No\_Health\_Ins\_ACS\_14\_18$: percent without health insurance;
   - $pct\_Civ\_unemp\_16p\_ACS\_14\_18$: percent of 16 and older population unemployed;
   - $pct\_Diff\_HU\_1yr\_Ago\_ACS\_14\_18$ percent that have moved in the past year;
   - $pct\_NON\_US\_Cit\_ACS\_14\_18$: percent that are not U.S. citizens;
   - $avg\_Tot\_Prns\_in\_HHD\_ACS\_14\_18$: average number of people living in the household;
   - $avg\_Agg\_HH\_INC\_ACS\_14\_18$: average household income;
   - $pct\_Vacant\_Units\_ACS\_14\_18$: percent of housing units that are vacant (not lived in);
   - $pct\_Renter\_Occp\_HU\_ACS\_14\_18$: percent of housing units that are rented (vs. occupied by owner);
   - $avg\_Agg\_House\_Value\_ACS\_14\_18$: average home value;
   - $pct\_HHD\_No\_Internet\_ACS\_14\_18$: percent of households that do not have internet;
   - $pct\_Single\_Unit\_ACS\_14\_18$: percent of housing units that are single family (vs. multi-family – e.g. apartment buildings);
   - *division*: geographic groupings of states – https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf.

This list excludes some predictors from the full list based on EDA detailed in HW4 solution code $HW4\_Q1.R$ posted on Blackboard.

(a) Find summary statistics of the COVID-19 case count. Do you see evidence of overdispersion?

(b) Fit a Poisson regression with all of the above predictors. Test its goodness-of-fit via the appropriate deviance test. Make sure to report the null and alternative hypothesis, test statistic, critical value and conclusion.

(c) State and interpret $exp\left(\hat{\beta}_{pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18}\right)$. Is $pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18$ statistically significant? Carry out the appropriate Wald or deviance test to answer this question.

(d) Does it make sense to include an *offset* in this analysis? What variable would be a sensical *offset* index associated with the COVID-19 case count? Explain why.

(e) Fit a Poisson regression with all of the above predictors *and include your chosen offset from part (d)*. Test its goodness-of-fit via the appropriate deviance test. Make sure to report the null and alternative hypothesis, test statistic, critical value and conclusion.

(f) State and interpret $exp\left(\hat{\beta}_{pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18}\right)$ from the Poisson regression with the offset. Is $pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18$ statistically significant in the Poisson regression with the offset? Carry out the appropriate Wald or deviance test to answer this question.

(g) Compare your conclusions from part (c) and part (f) regarding $pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18$. Why do you think including an offset does or does not affect the analysis. Use summary statistics of the offset variable to support your answer.

(h) Fit a negative binomial regression with all of the above predictors *and include your chosen offset from part (d)*. Test goodness-of-fit via the appropriate deviance test. Make sure to report the null and alternative hypothesis, test statistic, critical value and conclusion.

(i) State and interpret the estimated dispersion parameter from the negative binomial model with the offset. Is there evidence of overdispersion?

(j) Recall that Poisson regression is a special case of negative binomial regression occurring when the dispersion parameter is zero. Because we have nested models, we can use a LRT (deviance) test to assess the statistical significance of the dispersion parameter. Test for overdispersion by carrying out a deviance test comparing the Poisson regression model with an offset to the negative binomial model with an offset. Make sure to report the null and alternative hypothesis, test statistic, critical value and conclusion.

(k) State and interpret $exp\left(\hat{\beta}_{pct\_No\_Health\_Ins\_ACS\_14\_18}\right)$ from the negative binomial regression with the offset. Is $pct\_No\_Health\_Ins\_ACS\_14\_18$ statistically significant in the negative binomial regression with the offset? Carry out the appropriate Wald or deviance test.

(l) Compare $exp\left(\hat{\beta}_{pct\_No\_Health\_Ins\_ACS\_14\_18}\right)$ and its statistical significance for the negative binomial regression with an offset versus the Poisson regression with an offset. Explain why your conclusions about the effect of $pct\_No\_Health\_Ins\_ACS\_14\_18$ are the same or different between these two models.

(m) Using AIC as the selection criteria, which of the following models would you use to model COVID-19 case count? Explain your choice.
  - Poisson regression, no offset;
  - Poisson regression with an offset;
  - negative binomial regression, no offset;
  - negative binomial regression with an offset.

2. (40 points) *Modeling 7-Day COVID-19 Case Count Over Time*

This question uses the data *covid_data_pdb_3.csv* posted on Blackboard under HW6. You will model the 7-day COVID-19 case count in a county (where county is defined uniquely by the variable `GIDSTCO`) for *three* 7-day periods:

- September 11 – September 17 ($week = 1$),
- September 18 - September 24 ($week = 2$), and
- September 25 – October 1 ($week = 3$)

as a function of the following predictors:

- $pct\_URBANIZED\_AREA\_POP\_CEN\_2010$: percent of urbanized land;
- $pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18$: percent in poverty;
- $pct\_No\_Health\_Ins\_ACS\_14\_18$: percent without health insurance;

These variables are standardized in the HW6 dataset to avoid computation issues.

(a) For each county, do you expect the three 7-day counts to have positive, negative or no correlation? Explain.

(b) Fit a Poisson regression model [with offset $Tot\_Population\_ACS\_14\_18$] via GEE with an unstructured correlation. State and interpret the estimated correlation matrix. What does this suggest about a reasonable working correlation structure? Why?

(c) Fit the same model as in part (b) but with an exchangeable correlation structure. State and interpret the estimated correlation matrix. How does this compare to the correlation matrix estimated in part (b)? Which correlation structure would you use?

(d) Recall that the `Robust S.E.` and `Robust z` protect against incorrect assumptions about the variance structure, where incorrect assumptions may lead to invalid inference. Compare the `Robust z` to the `Naive z` for each of the predictors from the GEE with your chosen correlation structure. Explain why you think they are similar or different.

(e) State and interpret $exp\left(\hat{\beta}_{pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18}\right)$ for the GEE fit with your chosen correlation structure. Is $pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18$ statistically significant? Carry out the appropriate Wald test to answer this question – make sure to choose and use the appropriate S.E. and z.

(f) Fit a Poisson regression [with offset $Tot\_Population\_ACS\_14\_18$] via GLM ignoring the clustered nature of the data. Compare the coefficient estimates, standard errors, and conclusions from Wald inference on each predictor from this independence model to the GEE model with your chosen correlation structure. Would you choose the GLM or GEE?

(g) Fit a Poisson regression random intercept model [with offset $Tot\_Population\_ACS\_14\_18$] using `glmer`. State and interpret the estimated variance of the random effect $\hat{\sigma}^2$. What does this suggest about the effect of the clustering structure? Why?

(h) State and interpret $exp\left(\hat{\beta}_{pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18}\right)$ for the GLMM fit in part (g). Is $pct\_Prs\_Blw\_Pov\_Lev\_ACS\_14\_18$ statistically significant? Carry out the appropriate Wald test to answer this question. How does this compare to your conclusions from the GEE analysis in part (e)?

(i) Compare the coefficient estimates, standard errors, and conclusions from Wald inference on each predictor from the GLM in part (f) to the GLMM model from part (g). Describe and explain any similarities and differences.

(j) Would you choose the GLM or GLMM? Support your choice with the appropriate deviance test. Make sure to report the null and alternative hypothesis, test statistic, critical value and conclusion.