Homework 3

Due Monday 10/11

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

Mu Hu, Yingfeng Hu and Baldur Hua will lead HW discussion.

1. (75 points) *Modeling COVID-19 Prevalence Data*

   Carry out a linear regression analysis of COVID-19 case rate in the U.S. as a function of the predictors noted below. This question uses the same data you already "got to know" in HW2 Q2. Your analysis should include all steps of statistical modeling presented in lecture (except EDA because you already did that in HW 2). A suggested rough outline of data analysis steps:

   - fit linear regression model with all predictors noted below;
   - check model assumptions, data anomalies (e.g. influential points, outliers, influence, collinearity), model fit statistics;
   - make necessary changes based on data/model checks (e.g. transformations) and compare to previous version of the model;
   - iterate on data/model checking and associated corrections;
   - implement a variable selection procedure;
   - check model assumptions, data anomalies, model fit statistics;
   - make necessary changes based on data/model checks;
   - implement model fit tests and assessments – are you satisfied with your model?;
   - carry out inference for relationships that are interesting to you and interpret findings (you do not have to hypothesis test and interpret all coefficient estimates, but select a few interesting ones – perhaps those you were curious about from HW2).

   Your final write-up should be an "analysis report" that appropriately describes, in words, each step in your analysis (integrated with your code and output). Your HW write-up should "tell the story" of your analysis, that is, do not include plots/statistics/output just for the sake of providing output. Part of the challenge of communicating data analysis is finding the appropriate balance between providing enough information to tell the story accurately and completely, but not providing too much information such that the important parts get lost.

   Below is the description of the dataset from HW 2.

   County-level COVID-19 case counts (as of 9/10/21) are obtained from https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/. The dependent variable is the covid case rate for the county

   $$100*(covid\_count \ / \ Tot\_Population\_ACS\_14\_18)$$

where the denominator is the total population of the county estimated from the 5-year (2014-2018) American Community Survey from the U.S. Census Bureau.

Geographic and demographic characteristics of each county are obtained from the Census Bureau's 2020 Planning Database (PDB). For now consider the following 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 ** = under 5, 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_Othr_Lang_ACS_14_18*: percent primary language other than English;
- *pct_Age5p_Only_Eng_ACS_14_18*: percent >5 years old speaking only English;
- *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_Born_foreign_ACS_14_18*: percent born outside the U.S.;
- *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_w_Computer_ACS_14_18*: percent of households that have a computer;
- *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.

You can read about the data at https://www.census.gov/topics/research/guidance/planning-databases.html and get the detailed description of the predictors at https://www.census.gov/content/dam/Census/topics/research/2020StateandCountyPDBDocumentation_V2.pdf.

The dataset combining the two sources, *covid_data_pdb.csv*, is posted on Blackboard.

2. (25 points) *Comparing Regression Models Assuming Normality*

   This question uses the wage data from HW1 and HW2. We are again interested in a linear regression of `wage` as a function of the continuous covariates `year` and `age`.

   (a) Fit a regression model using `lm`. State the functional form and distributional assumptions of this model.

   (b) Fit the regression model using `glm` with the `family=gaussian(link=identity)` option. State the functional form and distributional assumptions of this model.

   (c) Are the models in part (a) and part (b) the same or different? Show by comparing results (e.g. coefficient estimates, residuals, fitted values) and explain in words why they are the same or different.

   (d) Fit the regression model using `lm`, but with the response log(`wage`) instead of `wage`. State the functional form and distributional assumptions of this model.

   (e) Fit the regression model (with `wage` as the response) using `glm` with the `family=gaussian(link=log)` option. State the functional form and distributional assumptions of this model.

   (f) Are the models in part (d) and part (e) the same or different? Show by comparing results (e.g. coefficient estimates, residuals, fitted values) and explain in words why they are the same or different.