Homework 2

Due Monday 9/27

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

Rui Chen, Xuanyu Chen, Armando Garia and Yuanzhe He will lead HW discussion.

1. (20 points) *Linear Regression MLEs Using Numerical Optimization*

   This question uses the wage data from HW1. We are again interested in a linear regression of `wage` as a function of the continuous covariates `year` and `age`. Recall that then this linear model is defined as

   $$\mathbf{y} = \beta_0 + \beta_1 * \texttt{year} + \beta_2 * \texttt{age} + \epsilon$$

   where ϵ is a n-dimensional vector of errors distributed $\mathcal{N}(0, \sigma^2)$. You will be finding the MLEs $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ using numerical optimization to maximize the loglikelihood for the linear model which is

   $$\begin{aligned} \ell(\beta_0, \beta_1, \beta_2) &= \log\left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - \beta_0 - \beta_1 * \texttt{year}_i - \beta_2 * \texttt{age}_i)^2 / 2\sigma^2}\right] \\ &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 * \texttt{year}_i - \beta_2 * \texttt{age}_i)^2, \end{aligned}$$

   where for simplicity we will assume that $\sigma^2 = 1669.5$, i.e. it is fixed.

   (a) Use the optim function* to find the values of β_0, β_1 and β_2 that maximize the loglikelihood for the linear model with fixed $\sigma^2 = 1669.5$. Use the MLEs calculated in Q2 of HW1 as the initial (a.k.a. starting) values.

   (b) How does the estimate from part (a) compare to the results from Q2 of HW1? Note that $\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \mathbf{y}$ from Q2 of HW1 is the closed form solution for the MLE. Provide comment about the similarity/difference.

   * To get started understanding the `optim` function for a linear model, see https://www.r-bloggers.com/2013/03/how-to-use-optim-in-r/ and https://www.r-bloggers.com/2020/04/optimisation-of-a-linear-regression-model-in-r/.

2. (50 points) *EDA for COVID-19 Prevalence Data*

   In HW 3 you will be modeling COVID-19 case rates in the U.S. as a function of characteristics of the geographic area. 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_Age*5*p_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_*16*p_ACS_*14_18: percent of 16 and older population unemployed;
- *pct_Diff_HU_*1*yr_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.

(a) Suggest a few questions that would be interesting to study with this data.

(b) Desribe a hypothesis test you might consider carrying out with this data.

(c) Would you be more interested in inference or prediction with this data? Explain.

(d) Use the `summary` command to obtain descriptive statistics of all variables. Make sure all variables are assigned the correct type (e.g. numeric, factor, etc.). Comment on any interesting or unusual findings.

(e) Use the `hist` command to describe the variable of interest: covid case rate. Overlay the empirical distribution using the `density` command. Comment on any interesting or unusual findings.

(f) Use the `pairs` and `cor` commands to depict two-way relationships between all continuous variables including covid case rate. Comment on any interesting or unusual findings.

(g) Use the `boxplot` command to depict two-way relationships between any factor variable and covid case rate. Comment on any interesting or unusual findings.

(h) Carry out further EDA using descriptive statistics and/or graphical displays for any interesting relationships you discovered.

(i) Based on your EDA, which variables do you think are strongly related to covid case rate? Why?

3. (30 points) *Fitting Linear Regression and Extracting Information*

Fit the linear regression model from Q1 using the `lm` command in R. Note that the `lm` function uses the model formulae discussed in Lecture 2.

(a) State the coefficient estimates. How do these compare to HW1/Q2, HW1/Q3(c), and HW2/Q1? Comment on why the similarity and/or differences.

(b) Provide an interpretation for the regression coefficient associated with `age`.

(c) Find the residuals three ways and show that they give the same result (e.g. using `summary` or a logical vector):

- by direct calculation of $y_i - \hat{y}_i$ (you do not have to calculate the fitted values by hand, you can extract them from the model object),
- by name from the model object, and
- using $ to get the residual component of the model object.

(d) Find the AIC two ways and show that they give the same result:

- by direct calculation using the `logLik` function to extract the loglikelihood, and
- using the `AIC` function.

(e) Find the MSE two ways and show that they give the same result:

- by direct calculation based on residuals calculated (in any of the three ways) in part (c), and
- by extracting the correct element from the `summary` of a linear model object output. *Hint: MSE is equivalent to $\hat{\sigma}^2$ for a linear model.*

(f) Carry out your own forward selection algorithm using AIC as the evaluation criteria, where at each step the model with the smaller AIC is better. Based on your procedure, state whether to include `year`, `age` or both in the linear model.