Homework Assignment 8

Total points:130

Due: Friday Nov 18, 2022 at 5PM

Please email your answer (compiled pdf file from R markdown) and R code to Yen-Yi Ho (hoyen@stat.sc.edu).

Instructions: feel free to discuss the homework with other students. However, each student must conduct their own analyses and write-up their own solutions. Write as if for a scientific journal. Be brief and accurate.

Background

Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.) and employers across the United States. MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance held by and available to U.S. workers. We will use the MEPS data collected during 2009 for the following analysis.

The MEPS data is available at
http://people.stat.sc.edu/hoyen/Stat704/Data/h129.RData

Use the following code to get needed variables for the following analysis,

```
>dat<-data.frame(h129$TOTEXP09, h129$SEX,
 h129$RACEX, h129$ASTHDX, h129$DOBMM,
 h129$DOBYY,h129$ADSMOK42, h129$CALUNG, h129$CHDDX,
 h129$EDUCYR, h129$POVCAT09,  h129$MARRY09X,
 h129$SEATBE53, h129$PERWT09F, h129$AGE09X)
```

and the codebook at
https://meps.ahrq.gov/data_stats/download_data_files_codebook.jsp?PUFId=H129
to look up the meaning of each variable.

I. Advanced Inference for Linear Regression

Use MEPS data set to address the question of whether men and women use roughly the same quantity of medical services at each age. That is, determine whether the age-dependence of average transformed medical expenditures (log10($+100)) is roughly the same for men and for women.

1. Plot total expenditure versus age, put a lowess smooth line (use lowess function in R). To make a **"useful" plot**, plot expenditures above $10,000 near $10,000 or use a logarithmic scale, **jitter** the points and use a plotting symbol with less mass. (6 points)
2. Comment on the linear relationship you observed in the figure above. Fit a regression model with spline transformation (if needed) to describe the effect of age on medical expenditure. (10 points)
3. Use different color and line type, put lowess smooth lines for male and female, respectively, in the figure plotted in (1). (4 points)
4. Based on the figure plotted in (3), comment on whether gender ``modifies" the age-expenditure relationship. Fit a MLR model to describe the dependence of expenditures on just age and gender (include interaction terms if needed.) (10 points)
5. Make residual plots, a Q-Q plot and a boxplot of the residuals obtained from the model described in (4). Comment on whether the assumptions in the model described in (4) seem to be reasonable in light of the residual plots, the Q-Q plot and boxplot. (15 points)
6. The model estimates in (4) is likely to be substantially influenced by one or a few extreme observations. A better alternative is to model lexp=log10(expenditure + 100). Fit a MLR of **lexp** on age (with splines if needed), gender, and interaction between age (with splines if needed) and gender. (5 points)
7. Interpret each of the coefficients in the model. (10 points)
8. Test the null hypothesis that the dependence of mean(Lexpend) on age is the same for men and women. (5 points)
9. Make a plot of the difference between women and men in the expected **Lexpend** as a function of age based on the model in (6). (10 points)
10. Plotted the DFBETAs for age against the predicted Lexpend to identify highly influential observations. How many highly influential observations do you find? Drop them and refit the model. (10 points)

11. Use the **esticon** function in R to calculate the estimated difference between women and men in average(Lexpend) and it standard error at 40, 65, and 80 years of age. Complete the table below. (10 points)

Age	Estimated Diff In **Lexpend**: Women-Men	Std Error	95% CI
40			
65			
80			

We now want to use the fitted model to estimate the difference in expenditures ($), not log10($+100), between a man and woman of the same age. Note that the model prediction of median expenditure is a non-linear function of the model parameters.

12. Complete the table for median expenditures ($) between a man and woman. Use the delta method and bootstrap approach to estimate the standard error of difference in median expenditures. (15 points)

Age	Estimated Diff In Median expend(**$**): Women-Men	Std Error (Delta Method)	Std Error (bootstrap)	95% CI
40				
65				
80				

13. Summarize your regression findings in a brief paragraph as if for a public health journal. Use coefficient estimates and confidence intervals in the text. (15 points)
14. Write your own lm function in R to perform multiple linear regression analysis using **matrix representation** (need to return **intercept, slope and MSE)**. (5 points)