Stat 104
Regression Project
Due 8am EST December 7, 2017

Collaboration.
You must work by yourself on this project. No collaboration with anyone else (in this class or not) is permitted. You may consult with the teaching staff as the need arises.

Comment on the role of your TF
Please only consult with your TF when you run into a problem, not when you are just looking to have someone check your work. Otherwise, contacts with your TF may become excessive and this project tends to become a "joint project with your TF" rather than "your own project."

Overview
An individual "regression report" is required of all students. Submit the report as you would a homework assignment. Each student will analyze the same data set, described below. The report is expected to be 5-10 pages in length.

Project Background
We are interested in the general question of what factors impact health care utilization spending among the elderly. Medicare, the federal health insurance program for the elderly, is the fastest growing expense in the federal budget. Knowledge of what factors contribute to health care expenditures will possibly help identify what sort of programs to implement to reduce future expenditures. Our data comes from the 2005 Medical Expenditures Panel Survey. A description of the variables is at the end of this project document. The explanatory variable is `totalexp`.

The data may be loaded into R using the command
```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/hospvisits.csv")
```

Assignment

Imagine that you work for a statistical consulting firm that has been asked to determine factors that contribute to (in a positive or negative way) health care costs among the elderly. Given the data, you will build a regression equation with `totalexp` (or possibly some transformation of this variable) as the dependent variable, and characteristics of the population as regressors.

Here is a simple example, taken directly from the data set that you will be using (you can do better than this).

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/hospvisits.csv"
)
> names(mydata)
 [1] "totalexp"  "age"       "marital"   "educ"      "income"    "srhealth"
 [7] "mntl_hlth" "phy_lim"   "bmi"       "chd"       "high_chol" "diabetes"
[13] "dr_visits" "msa"       "race_grp"  "smoker"    "male"      "high_bp"
[19] "hosp_vis"

> fit=lm(totalexp~age+marital+educ+income+bmi,data=mydata)
> summary(fit)

Call:
lm(formula = totalexp ~ age + marital + educ + income + bmi,
    data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-10426  -6048  -4148    360 100249

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.039e+03  6.350e+03  -1.108   0.2680
age          1.584e+02  6.899e+01   2.297   0.0219 *
marital      6.278e+01  5.223e+02   0.120   0.9043
educ         1.289e+00  1.217e+02   0.011   0.9916
income      -2.778e-02  1.946e-02  -1.428   0.1538
bmi          1.467e+02  8.035e+01   1.825   0.0683 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11710 on 793 degrees of freedom
Multiple R-squared:  0.01367,   Adjusted R-squared:  0.007448
F-statistic: 2.198 on 5 and 793 DF,  p-value: 0.05273
```

There are many models that could be set up.

You will need to think about all of the technical problems that can arise in regression. For this data set multicollinearity, heteroscedasticity, and outlying observations may be problems. You will probably wish to carry out some hypothesis tests as part of your work; these will be of dubious value if the assumptions of the normal multiple linear regression model are badly violated.

The deliverable for this project is a written report submitted in canvas. The report itself must be no more than 10 pages, including graphs and any R output you wish to include. The report must present, at the end, a single model that is the final product of the work. We want your model building process explained in an easy to follow format (this includes why variables were excluded, transformed or modified, and any diagnostics that were performed).

Evaluation

The grade on the assignment will depend on two things.

1. *Clari*ty (50%). The report must clearly indicate how you went about your work, including what models were considered.

2. *Substance* (50%). The project should use the tools developed in our class in an appropriate and correct manner. The report should anticipate questions that a technically critical reader might ask. For example, if the model can predict negative health care expenditures for certain reasonable values of the regressors, and there is no discussion of this fact, there are problems. Similarly, if heteroscedasticity is likely to be a problem with a particular function form, then the report must indicate how this was handled.

Notes:
Note that we enjoy graphs like regression diagnostic plots and scatter plots of response versus explanatory variables (at least one).

Clearly walk the reader through the analysis performed, but in a readable way without using a lot of computer output or jargon.

Your report should have a conclusion section written in "non-statistical" style that clearly summarizes the major conclusions from the final analysis and discuss these results. This section should also identify any major weaknesses of this analysis and discuss the likely impact of such weaknesses on your conclusions. The final piece of this section should discuss any policy implications of the study and possible future studies that may be needed to follow-up or confirm the current findings. Try not to get carried away with the implications.

Data Code Book

Variable name	Definition
age	Age in years
male	Dummy variable, =1 if male, 0 otherwise
race_grp	Race categorical variable, =1 if the respondent is white non-Hispanic, =2 if black non-Hispanic, =3 if other non-Hispanic, =4 if Hispanic
marital	Marital status categorical variable, =1 if married, =2 if widowed, =3 if divorced or separated, =4 if never married
income	Annual family income
educ	Years of education
msa	Dummy variable, =1 if live in MSA, =0 otherwise
bmi	Body mass index, weight in kilograms/(height in cm-squared)
smoker	Dummy variable, =1 if a current smoker, =0 otherwise
high_bp	Dummy variable, =1 if respondent has high blood pressure, =0 otherwise
high_chol	Dummy variable, =1 if respondent has high cholesterol, =0 otherwise
phy_lim	Dummy variable, =1 if respondent has a physical limitation, =0 otherwise
diabetes	Dummy variable, =1 if respondent has diabetes, =0 otherwise
chd	Dummy variable, =1 if respondent has chronic heart disease, =0 otherwise
srhealth	Self reported health status, =1 if excellent, =2 if very good, =3 if good, =4 if fair, =5 if poor
mntl_hlth	Self reported mental health, =1 if excellent, =2 if very good, =3 if good, =4 if fair, =5 if poor
dr_visits	Doctor visits in 2005. This variable is an integer value (0,1,2…)
hosp_vis	Hospital visits in 2005. This variable is an integer value (0,1,2…)
totalexp	Total expenditures on medical care in 2005. This variable measures the value of medical care consumed by the respondent, not what they paid out of pocket.