**EPI 510 R assignment 2**

For this assignment, please submit two files: (1) **a file containing text, tables and figures** (.doc, .docx, .pdf or .txt) responding to questions posed in the assignment (questions requiring a written answer or figures are underlined), and (2) **an R script** that performs the requested operations (.R).

1. For R assignment 1, you wrote a script that (a) brought in data from “vipcls.csv”, (b) set to missing all -1’s and implausible values for momage, delges and bw. For this assignment, create a new script, starting with your data management code from assignment 1. We’ll use this as a starting point, make a few modifications, and add additional code to expand our analysis **(1 point)**.
2. Define new categorical variables based upon existing variables. I’ll give the original variable names, and category cutoff values on the left, and the new variable name and the values corresponding to the category cutoffs on the right. **(32 points each, for 16 points total)**
   1. momage (<19, 20-29, 30+) -------------------> momageCat (1, 2, 3)
   2. grade (<12, 13+), ----------------------------> college (0, 1)
   3. marstat (1, 2-4, 5) --------------------------> marCat (1, 2, 3)
   4. partyr (<1, 2+) -----------------------------> monog (1, 0)
   5. pregnum (1 vs 2+) ---------------------------> multipar (0, 1)
   6. delges (<37, 37-42, 43+) -------------------> delgesCat (1, 2, 3)
   7. bw (<1500, 1500-2499, 2500-3999, 4000+) -> bwCat (1, 2, 3, 4).
   8. Finally, create a new variable for Study Clinic (clinic), which is based upon the patid (Clinic 1 has patid 1,000,000 to 1,999,999, Clinic 3 has patid from 3,000,000 to 3,999,999, etc). Clinics (which are fabricated) are as follows: 1=Olympia, 3=Everett, 5=Seattle, 6=Bellingham, 7=Spokane, 8=Bellevue, 9=Tacoma
3. We have variables indicating smoking and alcohol use in the first and second trimester. We want to define 2 new variables categorizing these behaviors across both trimesters:
   1. Create a new variable, called smoke, for smoking during pregnancy with three categories: (1) never smoked, (2) 1st trimester only, and (3) 2nd trimester with or without smoking in the 1st trimester. **(3 points)**
   2. Create a new variable, called drink, for alcohol use during pregnancy with three categories: (1) non-drinker, (2) use in 1st trimester only, and (3) use in 2nd trimester with or without use in the 1st trimester. **(3 points)**
4. Now we’re going to work with date variables:
   1. Create a date variable containing the enrollment date. **(2 points)**
   2. Create a date variable containing the delivery date. **(2 points)**
   3. Create a new variable containing the number of days between enrollment and delivery. **(2 points)**
   4. Determine if there are any delivery dates which precede enrollment. If so, at which clinic was the subject who had date problems enrolled? **(2 points)**

**Seattle (pat id 5248901)**

* 1. Set any implausible number of days between enrollment and delivery to missing in your dataset. **(2 points)**

1. Let’s add labels to produce a neat dataset. Convert all categorical variables to labeled factor variables (raceth, grade, marstat, etoh1, etoh2, deltype, induclab, auglab, intrapih, and all variables created in questions 2 and 3 of this assignment). **(15 points)**
2. Save this dataset as vipclsClean.csv (we’ll be working from this clean dataset in future homework assignments.) **(2 points)**