**HW1: Practice working with data in R**

**Background**

***Note: Scenario and all data were created for the purpose of this assignment; the data file does not include any real monitoring data nor does the narrative reflect true program history.***

***The objective of this homework assignment is to provide an opportunity for us to assess your R programming skills, with a focus on data prep, visualization and a simple model.***

Onchocerciasis is transmitted among humans by the black fly. Left untreated, it can result in debilitating skin or eye disease. Additional background information on onchocerciasis can be found here: <http://www.who.int/mediacentre/factsheets/fs374/en/>

The national onchocerciasis control program in Burkina Faso was implemented from 1995-2005. From 1995-1998, a series of community-level surveys were conducted to identify areas that might require intervention. These surveys were implemented in areas known or suspected to have high levels of onchocerciasis.

Two types of interventions were implemented at the community-level: 1) annual MDA with ivermectin targeted to the whole population aged 5 years and older; and 2) vector control. MDA treats the infection, under the assumption that if a large enough proportion of a community is treated every year, transmission would be reduced over time. MDA was only implemented in communities where the prevalence of onchocerciasis infection ≥20% among adults (infection was defined by presence of skin nodules). Vector control (insecticide, larvacide, etc.) included interventions designed to reduce the population of black flies, with the aim of preventing new transmission events. Not all communities received both interventions.

From 1999-2003, periodic monitoring surveys were conducted in these communities to determine if the interventions reduced nodule prevalence of onchocerciasis.

In 2004-2005, in order to qualify for additional funds for future MDA and vector control, surveys were repeated using a rapid diagnostic test to detect onchocerciasis antibodies among children less than 10 years of age among all communities in areas known or suspected to have onchocerciasis. Nodule prevalence was also measured as in earlier surveys. Due to problems with survey implementation, some study sites are missing at least one of the two diagnostic measures.

You are tasked with using programmatic data to conduct analysis. These datasets include:

1. Baseline\_eval.csv : a data file of geo-referenced prevalence data from community-based surveys
2. MDA\_coverage.csv : district level summary data of MDA coverage (% of district population treated)
3. Monitoring.csv: mid-program community level prevalence data
4. vector.csv : community level data on vector control implementation for 2001 (yes/no)
5. Data dictionary (excel)

**Questions:**

1. Consider the data from the evaluation surveys (2004-2005). You will notice that onchocerciasis prevalence is measured using two different diagnostics, with some communities only having one diagnostic reported.
   1. Present descriptive statistics of the study data.
   2. Make a scatter plot with nodule prevalence on the Y-axis and ov16 diagnostic on the X-axis.
   3. Do you notice any missing or implausible data in the baseline\_eval file?
2. Merge the vector control, MDA and baseline/evaluation datasets together:
   1. What variables do you think would be unique identifiers (UIDs) that you would use to merge these datasets?
   2. Merge the vector control data with the baseline and evaluation data.
   3. Now merge that table with the MDA data (you will notice this data is recorded at the district level, so will require you to match on a different UID).
   4. Finally, merge the community monitoring data with the rest of the study by village ID. Create a line chart that plots a line for each community as follows: Year on the x-axis, community level prevalence on the Y-axis, a group variable for whether or not vector control was implemented in 2001. Use nodule prevalence for the data from baseline and evaluation and skin snip from the community monitoring file. Only plot villages that have data reported in the community monitoring data file in this figure (each village should contribute three prevalence observations).
3. Create a simple linear model of the interventions: Treat all observed prevalence values as equivalent (make no correction for diagnostic type) and generate an estimate of the effect of MDA and vector control on the prevalence of onchocerciasis. Does MDA and/or vector control appear to reduce the prevalence of onchocerciasis? Follow these steps:
   1. Create an outcome variable for the difference in prevalence from baseline to evaluation [use difference in nodule prevalence if reported at both time points, or substitute Ov16 if nodule data missing at evaluation].
   2. Create a variable for MDA coverage that is the total number of rounds of MDA implemented.
   3. Create a binary variable for vector control (better to recode any vector control=1; no vector control=0; note the variable is currently coded as 2=no; 1=yes. Make sure this coding doesn’t alter your regression results).
   4. Run a linear model using the glm function, with the difference in prevalence as your outcome and MDA rounds and vector control as independent variables. Paste the output of your model here:
   5. Based on the results, which intervention had a stronger effect on the prevalence of onchocerciasis at the community level?
4. Think about the narrative and review the datasets. List three sources of bias that may exist in the data and explain how you would attempt to account for it in the analysis (you do not need to implement these approaches, just describe what you think should be done). If you need additional data sources not provided here, feel free to list those in your response:

Fill in your responses: