Stata Exercise (1):

In this example we are going to replicating some of the results in the paper: Godlonton and Thornton: “Learning from Others’ HIV testing: Updating beliefs and responding to risk” published in American Economic Review Papers and Proceedings, 2013.

As background:

This selected dataset comes from the larger Malawi Longitudinal and Family Health Survey, a longitudinal dataset that researchers at the University of Pennsylvania manage. This dataset is a longitudinal dataset spanning 1998 to 2010 covering 3 districts in Malawi and approximately one in four households in 1998. It is available for public use: <http://malawi.pop.upenn.edu/malawi-data-mlsfh>

In 2004, a randomized controlled trial was implemented led by Thornton (2008) whereby individuals were randomly assigned to receive a small monetary incentive (between $0 and $3) to learn their HIV results. VCT sites were randomly determined ensuring exogenous variation in distance to clinic as well.

Key variable definitions:

**got** is a binary indicator for those individual who learnt their HIV results

**any** is a binary indicator for whether an individual received any incentive

**tinc** the incentive amount the individual received, in usd

**villnum** is a village indicator

**distvct** is the distance from the individual to the vct site

**\_diedevr** is the number of people they know has ever died of HIV

**\_diedyr** is the number of people they know has died of HIV in past year

**\_diedrel** is the number of relative they know has ever died of HIV

**\_ccuse3, \_ccuse4** is currently using condoms with spouse in wave 3 and 4 respectively

**\_rcuse3ALL, \_rcuse4ALL** recently used condom with any partner in wave 4 and 4 respectively

**\_yrcheat3, \_yrcheat4** engaged in multiple concurrent partnerships in last year in wave 3 and 4 respectively

**villwght** is the village weight

**sample** is the sample we will use for table 2

1. Create the following variables needed for table 2:
   1. The fraction of the village that learnt their results (Vgot)
   2. The average incentive in the village (Vtinc)
   3. The fraction of the village that received any incentive (Vany)
   4. The average distance to the vct center in the village (Vdistvct)
   5. The number of individuals in the village (Vcount)
2. Summarise these new measures. Compare them to Panel A, columns 3 and 4.

Restrict the sample to sample = 1. See the footnote to Table 1 for the sample restrictions imposed.

1. Replicate table 2, shown below. Cluster the standard errors by village. Use the village weights using pweights (sampling weights). Include the following controls: age, age squared, male indicator, married indicator, educational attainment in 2004, dummy indicator for missing educational attainment indicator in 2004, dummy indicator for missing age information, dummy indicator for missing marital status, size of the village. Include lagged dependent variables.
   1. Why do you think the dummy indicators for the missing variables are used as controls?
   2. Refer to table 1 panel B, why is it important to include the lagged dependent variables?
   3. Interpret the coefficient on the proportion of village that got HIV results in each of the columns:
      1. Col 1:
      2. Col 2:
      3. Col 3:
      4. Col 4:
      5. Col 5:
      6. Col 6: