Methods in Epidemiologic Research

Sample Problems

Chapter 14 - Linear Regression

Preparation

We will use the -brazil- dataset and evaluate how various factors influence how many days a person had diarrhea assuming that they had at least one bout of diarrhea during the month.

If you are using Stata, some code to select the people with >1 day with diarrhea is

```
keep if di_d_ln!=.
```

We will also generate a new variable that represented the proportion of the municipality that was rural

If you are using Stata, some code to generate the variable is

```
gen m_prp_rural= m_pop_rural/ m_pop_total
```

We will then save a version of the dataset with just the variables of interest for this exercise.

If you are using Stata, some code to generate the variable is

```
keep mun comm fam id di_d di_d_ln fam_n age sex cistern income_social ///
     c_rec m_prp_rural m_hdi
save brazil_lin_reg_exc.dta, replace
```

Note: The sample size in this reduced dataset is only 496 individuals. Consequently, we will often retain variables in regression models even if their significance (P-value) falls between 0.05 and 0.1.

Caution

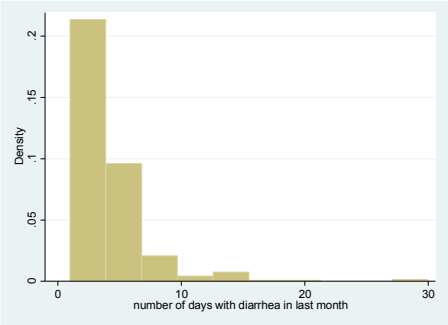
The data are clustered (people within households within communities within municipalities) so the assumption of independence of observations that is required for linear regression is clearly violated. The biggest effect of this clustering is likely to be an overestimation of the statistical significance of some predictors (particularly at the community and municipality level). Never-the-less, we will ignore clustering for these exercises (methods for dealing with clustering are covered in Chapters 20-24 of the text)]. The results from these analyses can then be compared with those we obtain when we account for clustering in the exercises for Chapter 21.

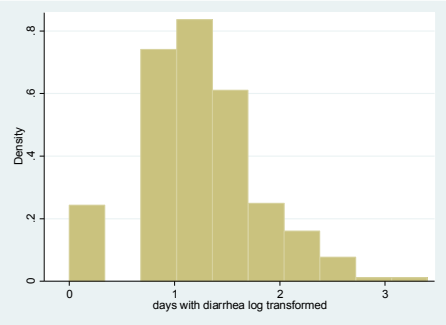
Questions

1. Simple and multiple linear regression

   (a) The first thing we should do is look at the distribution of the outcome of interest (days with diarrhea -di_d-). Generate histograms of the original and a log-transformed version of the variable.

```
sum di_d di_d_ln
     histogram di_d, name(raw, replace)
     histogram di_d_ln, name(ln, replace)
```





The raw data (left) is highly skewed. The log-transformed data (right) are closer to a normal distribution and we will consider this acceptable for the purpose of these exercises. The adequacy of the assumption of normality will be evaluated later. The gap in the graph on the right is a function of the raw data not being truly continuous (ie measured in 1 day increments). An alternative way to analyse such data would be to use some form of model for count data (eg Poisson).

(b) For each of the following variables (predictors) compute basic descriptive statistics (to determine what scale they are) and then fit simple linear regressions to determine if there is any evidence that they are associated with -di_d_ln-: -fam_n-, -age-, -sex-, -cistern-, -income_social-, -c_rec-, -m_hdi-, -m_prp_rural-. Interpret the coefficients for -cistern- and -m_hdi-

-age-, -sex, and -cistern-, are dichotomous variables and all the rest are treated as continuous. The statistical significance (P-value) of each predictor was: -fam_n- (P=0.361), -age- P=0.639), -sex- (P=0.162), -cistern- (P=0.022), -income_social- (P<0.001), -c_rec- (P=0.002), -m_hdi- (P=0.003), -m_prp_rural- (P=0.661). Only the models for -cistern- and -m_hdi- are shown below.

```
. reg di_d_ln cistern
      Source |       SS       df       MS              Number of obs =     496
-------------+------------------------------           F(  1,   494) =    5.31
       Model |   1.9414216     1   1.9414216           Prob > F      =  0.0216
    Residual |  180.613849   494  .365615079           R-squared     =  0.0106
-------------+------------------------------           Adj R-squared =  0.0086
       Total |  182.555271   495  .368798527           Root MSE      =  .60466

------------------------------------------------------------------------------
     di_d_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cistern |  -.1282025   .0556351    -2.30   0.022     -.237513    -.018892
       _cons |   1.205359   .0347943    34.64   0.000     1.136996    1.273722
------------------------------------------------------------------------------
```

Some important features to note are as follows. There are no missing data (all 496 observations are used in the analysis). Although the model is statistically significant, the R^2 is very low (1%). Having a water cistern in the house appears to reduce the number of days with diarrhea by 0.13 units (on the log scale).

```
. reg di_d_ln m_hdi
      Source |       SS       df       MS              Number of obs =     496
-------------+------------------------------           F(  1,   494) =    8.64
       Model |  3.13926445     1  3.13926445           Prob > F      =  0.0034
    Residual |  179.416006   494  .363190296           R-squared     =  0.0172
-------------+------------------------------           Adj R-squared =  0.0152
       Total |  182.555271   495  .368798527           Root MSE      =  .60265

------------------------------------------------------------------------------
     di_d_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       m_hdi |  -1.777397   .6045574    -2.94   0.003    -2.965218   -.5895758
       _cons |   2.236346   .3687261     6.07   0.000     1.511881     2.96081
------------------------------------------------------------------------------
```

*This model explains slightly nore of the variability in di_d_ln (R^2=1.7%), but that is still quite low. It appears that for each unit increase in the human development index for the municipality, there is a reduction in # of days with diarrhea of 1.8 log-units. This looks like a very big effect, but the 25th and 75th percentiles of -m_hdi- are 0.58 and 0.62 so going from a relatively low m_hdi municipality to a relatively high m_hdi entails an increase of 0.04 untis (which would equate to a reduction in -di_d_ln- of 0.04*1.8=0.07 log-units (less than the effect of having a water cistern).*

(c) Interpret the root MSE from the regression model for -cistern-

The standard deviation (a measure of the variability) of the original outcome variable (-di_d_ln) is 0.607. After accounting for whether or not household had a water cistern, the SD of the residuals is 0.605. The minimal reduction in the SD reflects the fact that water cistern only explains about 1% of the variability in the outcome.

(d) Compare the estimate for the effect of -cistern- from a simple linear regression to that obtained from a model with all the predictors listed in 1(b) included. Repeat the process for -m_hdi-.

The simple regression model for cistern is shown above. The multiple regression model is:

```
. reg di_d_ln fam_n age sex cistern income_social c_rec m_hdi m_prp_rural

      Source |       SS       df       MS              Number of obs =     496
-------------+------------------------------           F(  8,   487) =    7.20
       Model |  19.3181566     8  2.41476957           Prob > F      =  0.0000
    Residual |  163.237114   487  .335189146           R-squared     =  0.1058
-------------+------------------------------           Adj R-squared =  0.0911
       Total |  182.555271   495  .368798527           Root MSE      =  .57896
------------------------------------------------------------------------------
     di_d_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       fam_n |  -.0038309   .0146794    -0.26   0.794    -.0326736    .0250118
         age |   .0005477   .0015907     0.34   0.731    -.0025779    .0036732
         sex |   .0371437   .0526638     0.71   0.481    -.0663325    .1406199
     cistern |  -.0747975   .0552226    -1.35   0.176    -.1833015    .0337065
income_social |   .2879666   .0595239     4.84   0.000     .1710113     .404922
       c_rec |  -.0034193   .0011423    -2.99   0.003    -.0056638   -.0011748
       m_hdi |  -3.363706   .7779517    -4.32   0.000    -4.892262    -1.83515
  m_prp_rural |  -.7113484   .2125473    -3.35   0.001    -1.128971   -.2937255
       _cons |   3.426497   .5372685     6.38   0.000     2.370846    4.482147
------------------------------------------------------------------------------
```

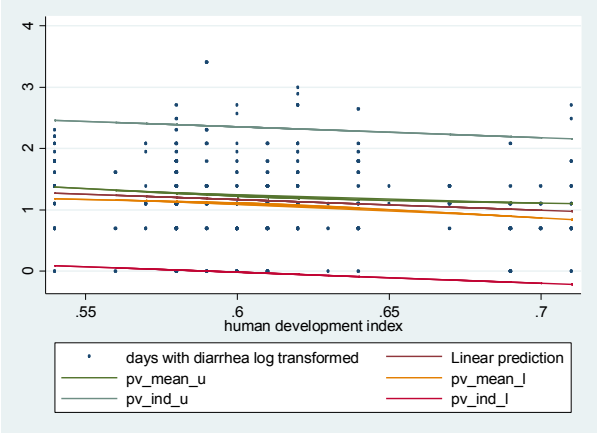
The apparent effect of water cistern has dropped from -0.13 to -0.07. Given that none of the added variables are likely to intervening variables, this suggest that one or more of them have some confounding effect on the impact of water cistern. (More on confounding later).

The simple regression model for -m_hdi- is shown above. The multiple regression model is:

```
.          reg di_d_ln fam_n age sex cistern income_social ///
>                 c_rec m_hdi m_prp_rural

      Source |       SS       df       MS              Number of obs =     496
-------------+------------------------------           F(  8,   487) =    7.20
       Model |  19.3181566     8  2.41476957            Prob > F      =  0.0000
    Residual |  163.237114   487  .335189146           R-squared     =  0.1058
-------------+------------------------------           Adj R-squared =  0.0911
       Total |  182.555271   495  .368798527           Root MSE      =  .57896
------------------------------------------------------------------------------
      di_d_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
        fam_n |  -.0038309   .0146794    -0.26   0.794    -.0326736    .0250118
          age |   .0005477   .0015907     0.34   0.731    -.0025779    .0036732
          sex |   .0371437   .0526638     0.71   0.481    -.0663325    .1406199
      cistern |  -.0747975   .0552226    -1.35   0.176    -.1833015    .0337065
income_social |   .2879666   .0595239     4.84   0.000     .1710113     .404922
        c_rec |  -.0034193   .0011423    -2.99   0.003    -.0056638   -.0011748
        m_hdi |  -3.363706   .7779517    -4.32   0.000    -4.892262    -1.83515
  m_prp_rural |  -.7113484   .2125473    -3.35   0.001    -1.128971   -.2937255
        _cons |   3.426497   .5372685     6.38   0.000     2.370846    4.482147
------------------------------------------------------------------------------
```

The magnitude of the coefficient for -m_hdi- has increased from -1.78 to -3.36. This is either due to removal of confounding effects or inclusion of intervening variables in the model. It turns out that -m_prp_rural- is highly (negatively) correlated with -m_hdi- (ρ=-0.66). As the proportion of a municipality that is rural goes up, the average human development index goes down. When you compare 2 municipalities with the same proportion rural (ie the multivariable model), the effect of -m_hdi- is much greater than when you compare 2 municipalities randomly chosen from the population.

(e) Graph the relationship between -di_d_ln- and -m_hdi- and include confidence intervals for the mean and for individual observations.



The confidence interval around the mean prediction is quite narrow (partly reflecting the fact that the mean number of days with diarrhea doesn't change much with -m_hdi-). However, the prediction interval for individual observations is very wide, reflecting the fact that the model has low predictive ability so it is hard to predict the number of days with diarrhea for an individual.

(f) How well do models with just -cistern- and with all predictors do in predicting the number of days with diarrhea?

As noted above, the model with just -cistern- only explains about 1.1% of the variability in -di_d_ln while the full multivariable model explains about 10.6%. There is still a lot of variability between individuals in how long they have diarrhea that is not explained by the predictors in this model.

2. Coding predictors, confounding and interaction.

(a) The size of the community (-c_rec-) is a continuous variable (range in values from 4 to 104). Compare models with c_rec as the only predictor in which c_rec is included as a continuous variable or as a 4-level categorical variable (0-16, 17-31, 32-46 and 47-104). (These cutpoints were chosen because they are the quartiles of the variable.)

The code to generate the coefficients from the two models was:

```
reg di_d_ln c_rec           /* continuous predictor */
     estimates store cont
egen c_rec_c4=cut(c_rec), at(0 17 32 47 999) icodes
reg di_d_ln i.c_rec_c4  /* categorical predictor */
     estimates store categ
estimates table cont categ
```

and the estimates were:

```
-------------------------------------
    Variable |    cont        categ
-------------+-----------------------
       c_rec | -.00352462
             |
    c_rec_c4 |
           1 |              -.12408338
           2 |               .02770988
           3 |              -.21971667
             |
       _cons |  1.2819739    1.2341668
-------------------------------------
```

When treated as a continuous predictor, the model assumes a straight line relationship (in this case a slight downward (negative) slope as community size increases. This assumption of a linear effect may not legitimate.Compared to the smallest communities, the next largest (c_rec_c4=1) and the largest (c_rec_c4=3) have shorter duration of illness, but average size communities (c_rec_c4=2) have longer illness. The model with the continuous predictor has an R^2 of only 1.8% while that with the categorical predictor has an R^2 of 2.8% ... but this improved predictability is acheived at the cost of having 3 predictors instead of one. Never-the-less, there is an increase in the adjusted R^2 from 1.6% to 2.2%

(b) Fit a model with all of the predictors (listed in 1(b)). Interpret the intercept. How could the "interpretability" of this value be improved?

The model computer in 1(d) has an intercept of 3.43. This indicates that an individual with all

predictors set to "0" would have an expected log-days of illness of 3.43. This individual would be 0 yrs old, be from a family of size=0, live in a community with no households, etc. Clearly such an individual can not exist. To improve the interpretability of the intercept we will either centre all continuous predictors by subtracting the mean value (eg. -age-), or set their "0' value to either a reasonable value (eg -c_rec-=50) or the smallest legitimate value (eg -fam_n-=2). The code to do this is:

```
gen fam_n_min=fam_n-2
gen age_ct=age-16.6
gen c_rec_50=c_rec-50
gen m_hdi_ct=m_hdi-.608
gen m_prp_rural_25=m_prp_rural-.25
reg di_d_ln fam_n_min age_ct sex cistern income_social ///
     c_rec_50 m_prp_rural_25 m_hdi_ct
```

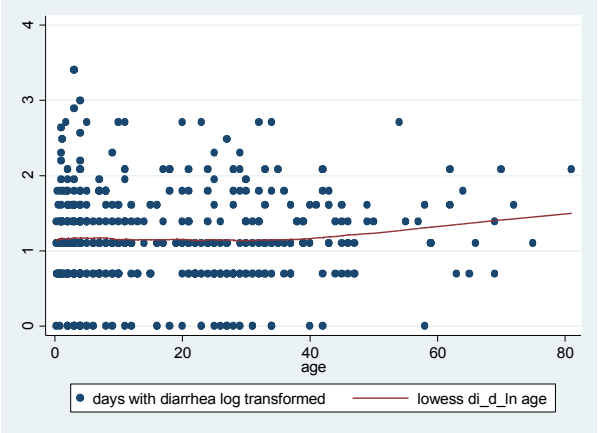
The resulting model is:

```
.reg di_d_ln fam_n_min age_ct sex cistern income_social c_rec_50 m_prp_rural_25
m_hdi_ct

...output omitted
------------------------------------------------------------------------------
       di_d_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
     fam_n_min |  -.0038309   .0146794    -0.26   0.794    -.0326736    .0250118
        age_ct |   .0005477   .0015907     0.34   0.731    -.0025779    .0036732
           sex |   .0371437   .0526638     0.71   0.481    -.0663325    .1406199
       cistern |  -.0747975   .0552226    -1.35   0.176    -.1833015    .0337065
 income_social |   .2879666   .0595239     4.84   0.000     .1710113     .404922
      c_rec_50 |  -.0034193   .0011423    -2.99   0.003    -.0056638   -.0011748
m_prp_rural_25 |  -.7113484   .2125473    -3.35   0.001    -1.128971   -.2937255
      m_hdi_ct |  -3.363706   .7779517    -4.32   0.000    -4.892262    -1.83515
         _cons |   1.033991   .0721742    14.33   0.000       .89218    1.175803
------------------------------------------------------------------------------
```

Now the constant (1.03) represents someone from a family of size=2, of average age, of sex=0 (male), from a household that does not have a water cistern or receive income from social assistance, is from a 50 household community and lives in a municipality that is 25% rural and has an average human development index.

(c) Is the relationship between -age- and -di_d_ln- linear?. If not, how might this situation be dealt with?

A simple scatterplot of -di_d_ln- against -age- does not reveal an obvious pattern, but adding a lowess smoothed curve (graph below - see Chapter 15 of text for explanation of lowess curves) is more informative.



The graph suggest that there might be some curvature to the line. Adding a quadratic term (age^2) to a model with age produces the following:

```
. reg di_d_ln age age_sq
...output omitted ...
------------------------------------------------------------------------------
     di_d_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0074775   .0045515    -1.64   0.101    -.0164203    .0014653
      age_sq |   .0001562   .0000804     1.94   0.052    -1.68e-06    .0003141
       _cons |   1.193361   .0465712    25.62   0.000     1.101858    1.284864
------------------------------------------------------------------------------
```

The quadratic term is borderline significant providing some additional (limited) information that the relationship is not linear.

(d) If you fit a model with -age- and -age^2-, do the coefficients makes sense? If not, what is wrong and how can it be fixed?

It is very hard to interpret coefficients from a queadratic model, but the estimates seem reasonable. This is not always the case. In some situations, the linear effect (-age-) may have a very strange coefficient or very large SE. The VIF of age^2 is 7.7 which is getting close to the point where me might expect serious problems of collinearity. In general, it is a good practice to centre variables before making a quadratic term.

```
    replace age_sq=age_ct^2
         reg di_d_ln age_ct age_sq
         vif
... output omitted ...
------------------------------------------------------------------------------
     di_d_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_ct |  -.0022908   .0022769    -1.01   0.315    -.0067644    .0021829
      age_sq |   .0001562   .0000804     1.94   0.052    -1.68e-06    .0003141
       _cons |   1.112284   .0350556    31.73   0.000     1.043408    1.181161
------------------------------------------------------------------------------
```

The VIF for this model as been reduced to 1.92 ... a much more stable model.

(e) Next we will evaluate possible confounding between -c_rec- and -cistern-. Is it possible that -c_rec- is a confounder for -cistern-? Is it possible that -cistern- is a confounder for -c_rec-? Evaluate which makes sense.

The size of a community might influence how likely it is that an individual household within the community might have a water cistern, so -c_rec- might be a confounder for -cistern-. The opposite can not be true (having a water cistern can no influence the size of the community) so -cistern- can not be a confounder for -c_rec- (it would be considered an intervening variable, not a confounder). Controlling for -c_rec- when evaluating the effect of cistern produces:

```
. reg di_d_ln cistern     c_rec   /* model with c_rec */
... output omitted
------------------------------------------------------------------------------
     di_d_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     cistern |  -.0961306   .0565952    -1.70   0.090    -.2073282     .015067
       c_rec |  -.0031026   .0011664    -2.66   0.008    -.0053942   -.0008109
       _cons |   1.304395    .050814    25.67   0.000     1.204556    1.404233
------------------------------------------------------------------------------
```

Controlling for -c_rec- reduced the beneficial effect of -cistern- from -0.128 to -0.096-. Some of the apparent beneficial effect of -cistern- was attributed to the fact that they were more common in larger communities and illness duration was shorter in large communities

(f) Conceptually, it might make sense that the effect of being in a low-income household (income_social=1) might depend on whether or not the family has a water cistern. Does it?

To evaluate this, we need to create an interaction term between -income_social- and -cistern-. If we do this, the interaction term has a P-value of 0.098 suggesting borderline significance (data not shown). If we retain the interaction term and express the model so the effects of the various combinations of -cistern- and -income_social- are presented, we get the following estimates:

```
cistern#income_social |
                0 1   |   .3632629   .0701224     5.18   0.000     .2254831    .5010427
                1 0   |   .0506168    .097275     0.52   0.603    -.1405137    .2417474
                1 1   |   .2182466   .0772035     2.83   0.005     .0665535    .3699398
```

Being in a household with income from social assistance results in longer duration of illness, but the effect is 0.363 log-units when there is no water cistern in the house, but 0.218 when there is a cistern present.

(g) Build a model in which you retain any predictors (including new predictors created earlier in question 2) that are significant at P<0.1.

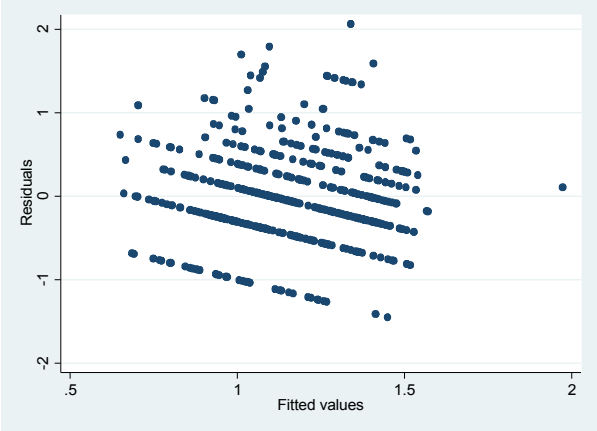
This produces the following model (which will be used for all diagnostic evaluations (question #3):

```
.         reg di_d_ln age_ct age_sq cistern#income_social ///
>                 c_rec m_prp_rural m_hdi

      Source |       SS       df       MS              Number of obs =     496
-------------+------------------------------           F(  8,   487) =    8.34
       Model |    21.98845     8  2.74855625           Prob > F      =  0.0000
    Residual |  160.566821   487  .329705997           R-squared     =  0.1204
-------------+------------------------------           Adj R-squared =  0.1060
       Total |  182.555271   495  .368798527           Root MSE      =   .5742

------------------------------------------------------------------------------
             di_d_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------+--------------------------------------------------------
```

```
                  age_ct |   -.0029284   .0021647    -1.35   0.177    -.0071817    .0013249
                  age_sq |    .0001708   .0000775     2.20   0.028     .0000186     .000323
                         |
  cistern#income_social |
                   0 1   |    .3632629   .0701224     5.18   0.000     .2254831    .5010427
                   1 0   |    .0506168    .097275     0.52   0.603    -.1405137    .2417474
                   1 1   |    .2182466   .0772035     2.83   0.005     .0665535    .3699398
                         |
                  c_rec  |   -.0031899   .0011355    -2.81   0.005     -.005421   -.0009588
            m_prp_rural  |   -.7139844   .2135281    -3.34   0.001    -1.133535   -.2944343
                  m_hdi  |    -3.33322   .7721171    -4.32   0.000    -4.850312   -1.816128
                  _cons  |    3.316236   .5314084     6.24   0.000       2.2721    4.360373
------------------------------------------------------------------------------
```

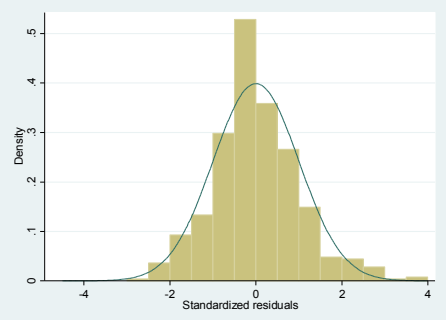
3. We will base our model diagnostics on the model arrived at in 2(g). (Hopefully this includes: -age- (with a quadratic term), -cistern- and -income_social- (and their interaction), -c_rec-, -m_prp_rural-, and -m_hdi-.

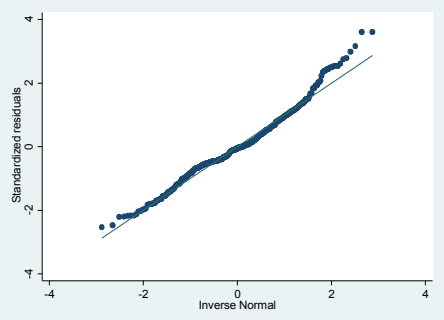
   (a) Evaluate the assumption of homoscedasticity.



There is no obvious pattern to the residuals and a statistical test for heteroscedasticity is completely non-significant (P=0.85) indicating no evidence of violation of the assumption of homoscedasticity.

   (b) Evaluate the normality of the residuals. Do this graphically and using a statistical test.





The histogram shows residuals which are normally distribvuted (except perhaps for a peak of values just below 0). The normal probability plot highlights the fact that there are some positive residuals that are larger than expected. Never-the-less, neither graph reflects a serious departure from normality. The statistical test of normality has a P-value of <0.001 indicating a highly significant departure from normality. This is a commonly observed situation: residuals which

look reasonable when viewed graphically, but a significant statistical test. In these situations, we tend to be guided more by the visual assessment.

(c) We will now turn to evaluating residuals and other diagnostic parameters for individual observations.

The following code generated all of the required residuals and diagnostic parameters.

```
*generate necessary residuals and fit statistics
     capture drop yhat rsta rstu cooksd lev dfit
     predict yhat, xb
     predict rsta, rstandard
     predict rstu, rstudent
     predict cooksd, cooksd
     predict lev, leverage
     predict dfit, dfit
     predict db_hdi, dfbeta(m_hdi)
```

The threshold values for some of the diagnostic parameters were:

```
. scalar th_lev=2*(8+1)/496
. scalar th_cook=4/496
. scalar th_dfit=2*((8+1)/496)^0.5
. scalar th_db=2/(496^0.5)
.
. di "leverage threshold   " th_lev    leverage threshold   .03629032
. di "CooksD threshold   " th_cook     CooksD threshold   .00806452
. di "dfits threshold   " th_dfit      dfits threshold   .26940795
. di "dfbeta threshold   " th_db       dfbeta threshold   .08980265
```

i. Evaluate the standardized residuals. Are there any observations that are clearly outliers? What are the characteristics of the most extreme residuals.

```
. list di_d_ln age cistern income_social c_rec m_prp_rural m_hdi yhat rsta if
abs(rsta)>3, clean noobs

     di_d_ln   age   cistern   income~l   c_rec   m_prp_~l   m_hdi       yhat       rsta
    2.890372     3       Yes         No      36   .2231964     .62   1.097483   3.157362
    3.401197     3        No        Yes      22   .5236546     .59   1.340261   3.605666
    3.401197     3        No        Yes      22   .5236546     .59   1.340261   3.605666
```

There were 3 observations with residuals <-3 or >+3. These were three individuals with 18 or more diarrhea days during the month.

ii. Determine which individuals were potentially influential (high leverage) and actually influential (large Cooks D or DFITS). What were the characteristics of these observations?

The data are not shown here, but the individuals with the highest leverage were all older people (age>=66). There were quite few individuals (n=27) who appeared to have substantial influence on the model (ie Cook'd D or dfits outside the threshold). The most influential results were generally individuals with many days affected (di_d_ln>2) which resulted in large positive residuals and/or were either quite young (<=3) or older (>=54).

iii. Compute the delta-betas for m__hdi. Which individuals had the largest values?

There was an approximately equal mixture of negative (n=12) and positive (n=13) values (data not shown). Most of the influential observations had values of m_hdi that fell in the top 10% of values (ie >=0.69). No big surprise with this. However, care must be taken because this is a

municipal level variable so the influence of a municipality will depend on both its value of hdi and the number of observations from the municipality (range of # of observations from a municipality was 2 to 69).

(d) Refit the original model 3 different ways by leaving out: largest standardized residuals, largest Cook's D and largest delta-betas for m_hdi. What impact did each of these have on the original model in terms of coefficients or their significance.

The coefficients and statistical significance of the predictors in the original and the three revised models are shown below.

Variable	orig	omit_rsta	omit_cooksd	omit_db_hdi
age_ct	-.00292838	-.00158976	-.00179587	-.002239
age_sq	.00017082*	.00014803*	.00017042*	.00015244*
cistern#				
income_soc~l				
0 1	.36326287***	.34113603***	.34819302***	.36131042***
1 0	.05061682	.01543082	-.03090931	.05392547
1 1	.21824662**	.20374347**	.19043324*	.19102622*
c_rec	-.00318988**	-.00296533**	-.00295753**	-.002967**
m_prp_rural	-.71398441***	-.7340622***	-.71882892***	-.69887163***
m_hdi	-3.3332198***	-3.5003274***	-3.5522781***	-3.6427882***
_cons	3.3162364***	3.4252457***	3.4581712***	3.4968252***
N	496	492	492	492

legend: * p<0.05; ** p<0.01; *** p<0.001

In general there was very little change in the magnitude or the statistical significance of any of the coefficients in any of the revised models. This suggest that the model appears to be quite robust.