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

(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-.

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

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

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.)

(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?

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

(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?

(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.

(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?

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

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.

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

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

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

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

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

(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.