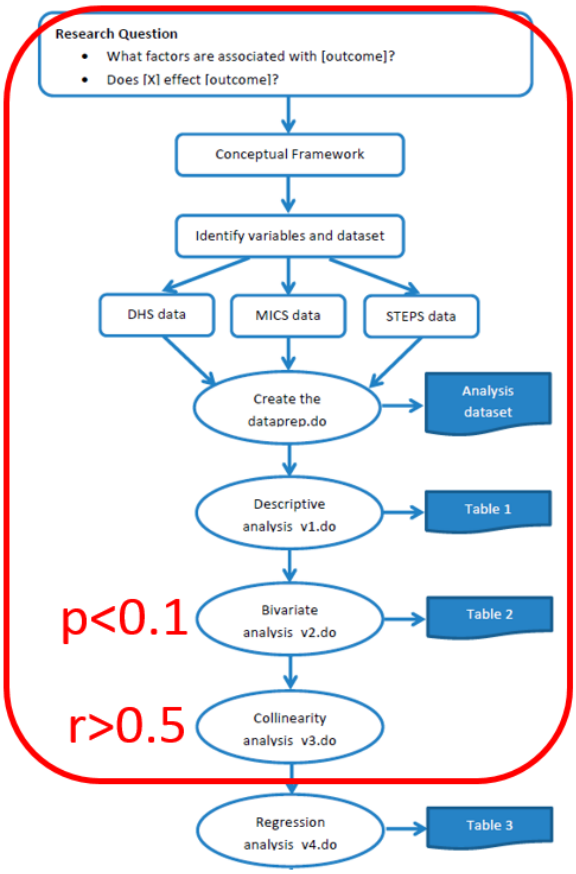
Stata: Multivariate Statistics – General Explanatory Modeling

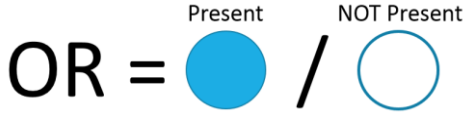
Topics: Manual backward stepwise logistic regression



1. Review of steps before logistic regression

Building a regression model is fast and easy once other foundational analysis steps have been performed. Before we build a multivariate model, we should (1) test the independent associations between each covariate and the outcome in bivariate analysis, and only advanced those variables which are statically significant at p<0.1, or which we decide *a priori* need to be in the model for conceptual reasons. (2) We test the correlations among all covariates with the Pearson's R statistic to identify collinear pairs associated at r>0.5, and retained only one of the variables (usually the one most strongly associated with the outcome) to ensure a stable model. Watch the bivariate statistics video at www.populationsurveyanalysis.com if you have questions about that.

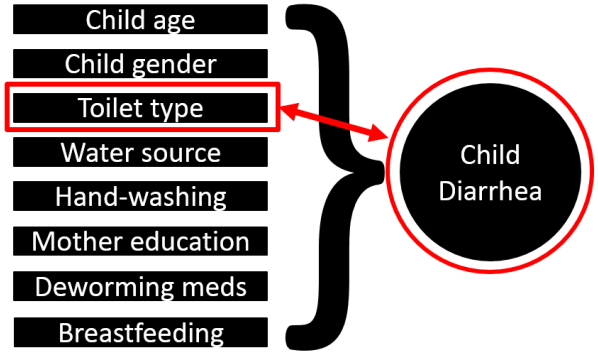
2. Logistic regression



Logistic regression is used to model the odds of a binary outcome. Results are reported as odds ratios – a ratio of the odds that the outcome occurs over the odds the outcome does not occur.

Regression analysis allows us to consider the effects/associations of multiple variables at once. Multivariate modeling has an advantage over bivariate modeling by identifying the additional explanatory power of a given variable, accounting for any overlap with other explanatory variables.

For example, we can identify the magnitude and direction of association between household toilet type and child diarrhea accounting for the related



influence of drinking water source and other household social-economic (SES) factors. Likewise, we can identify the additional, unique contribution of water source toward "explaining" childhood diarrhea beyond other SES factors.

In regression analysis we often use terms like "predictor" or "explanatory variable" which imply causation. With cross-sectional survey data, we cannot determine which factors caused the outcome, we can only identify factors that were associated with the outcome. I use causal terms like predictor occasionally to clarify how regression modeling works, but please remind yourself, these are associations only (just two things happening at the same time).

3. General explanatory modeling

In general explanatory modeling, we interpret all variables that remained statically significant in the model. A general explanatory model is used to answer the question "What factors are associated with the outcome?" This is different from other modeling approaches, for example hypothesis test modeling which we review in a separate lecture.

Variable	OR (95% CI)	p-value
Child age		
0-11 months	**1.00**	
12-23 months	**1.64 (1.25,1.91)**	**0.002**
24-59 months	**1.12 (0.90,1.29)**	**0.342**
Child female	**0.93 (0.89,0.99)**	**0.048**
Toilet type unimproved	**1.32 (1.39,1.85)**	**0.030**
Water source unimproved	**1.16 (1.07,1.32)**	**0.042**
Poor hand-washing	1.12 (0.95,1.98)	0.143
Mother's has secondary+	0.90 (0.85,1.11)	0.263
Deworming meds last 6mo.	**0.54 (0.42,0.63)**	**<0.001**
Exclusive breastfed	1.13 (0.90,1.23)	0.732

Statistically significant at p<0.05

We start the multivariate modeling process by fitting a model with all potential (non-collinear) covariates. We could stop here, and only interpret those variables which are statistically significant. In fact, this approach is taken in some social sciences where the conceptual framework drives the analysis.

In epidemiology, we tend to perform the additional process of stepwise regression to arrive at a reduced model with only the most important "explanatory" variables. The goal of backward stepwise regression is to identify key factors that are associated with the outcome. The stepwise process provides a systematic way to arrive at the simplest model with the most explanatory power.

4. How to perform manual backward stepwise logistic regression in Stata

The command for logistic regression with survey data is straight forward. Take account of the survey design with an `svy:` statement, then specify the `logistic` command. Next list the outcome

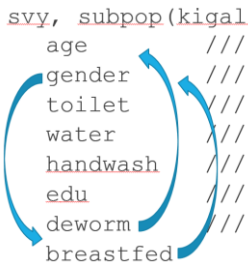
```
svy: logistic diarrhea ///
   age         ///
   gender      ///
   toilet      ///
   water       ///
   handwash    ///
   edu         ///
   deworm      ///
   breastfed
```

```
svy, subpop(kigali): logistic diarrhea ///
   age         ///
   gender      ///
   toilet      ///
   water       ///
   handwash    ///
   edu         ///
   deworm      ///
   breastfed
```

variable (note, we can only have one outcome variable per model), and then list all covariates.

We can optionally incorporate a subpopulation statement in the `svy:` statement to analyze a subset of the sample.

5. Manual backward stepwise regression is a formula.



1. Arrange all covariates from most to least important based on the conceptual framework. If there are covariates that must remain in the model regardless of their statistical significance, for example, age or urban/rural residence, then put them at the top of your list.

2. Run the full model (with all covariates).

```
svy, subpop(kigali): logistic diarrhea ///
   age         ///
   deworm      ///
   breastfed   ///
   toilet      ///
   water       ///
   handwash    ///
   gender      ///
   edu
test edu cat1 cat2
   *p-value = 0.367
```

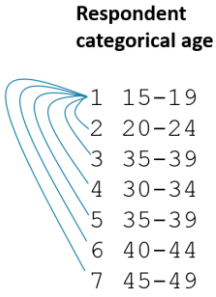
3. Test all variables for statistical significance at p<0.05 starting with the bottom variable.
    a. If the bottom variable is statistically significant, retain it in the model and test the next most important variable, and so on.
    b. If the bottom variable is not statistically significant, create a new model removing the non-significant variable, run the new model, and test significance of all variables starting again from the bottom variable.

4. Repeat step 3 until you are left with only covariates that are significant at p<0.05 or covariates that you decided were important to keep in the model regardless of statistical significance.

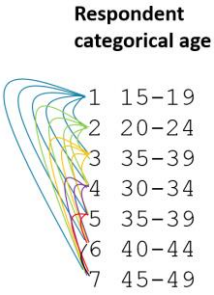
6. Dummy variables

Categorical covariates are far easier to interpret in logistic regression than continuous covariates, which is why we have been generating and analyzing categorical variables all along. Stata has

```
xi:svy, subpop(kigali): logistic diarrhea ///
    i.age           ///
    i.gender        ///
    i.toilet        ///
    i.water         ///
    i.handwash      ///
    i.edu           ///
    i.deworm        ///
    i.breastfed
```



diarrhea	Odds Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age_categorical						
20-24	1.167399	.2808356	0.64	0.520	.7276322	1.872952
25-29	.7836149	.1888855	-1.01	0.312	.4879631	1.258399
30-34	.6749488	.165116	-1.61	0.109	.4173404	1.091569
35-39	.6729353	.167726	-1.59	0.113	.4123425	1.098218
40-44	.5896339	.1553977	-2.00	0.046	.351285	.9897041
45-49	.3661241	.1198072	-3.07	0.002	.1924674	.6964653
_cons	.1967158	.0456354	-7.01	0.000	.1246963	.310331




```
xi:svy: logistic diarrhea i.age_categorical

test _Iage_categ_2 _Iage_categ_3 _Iage_categ_4 ///
     _Iage_categ_5 _Iage_categ_6 _Iage_categ_7
```

no way to differentiate numerical values that form categories versus numerical values in a continuous variable, so we have to specify this.

We tell Stata that a variable is categorical by placing an `i.` in front of it. We only *need* to include the `i.` for variables with three or more categories, though I include it with binary variables, too.

For variables with three or more categories, Stata will compare all groups. The output will display the p-values for a single set of comparisons to the reference category, however, if we want to generate an overall p-value for all comparisons, then we will need to create a dummy variable for each category to be included in a statistical test. By adding `xi:` to the beginning of the model statement, and including `i.` in front of categorical variables, Stata will generate temporary dummy variables for all categories and add them to the bottom of our dataset.

7. Global test for variable significance

As I just mentioned, Stata will display the p-values for a single set of comparisons to the reference category, however, we can generate an overall p-value that summarizes statistical significance of all comparison across all categories in the variable. This global p-value tells us whether the categorical variable is contributing the model; if any of its odds ratios are different.

To generate the global p-value, run the model, then run a test statement by typing `test` and then listing all of the dummy variables created for that variable. Note that the reference dummy variable is omitted. This is the global p-value for a multi-category variable.

8. Learning logistic regression

This course does not explain logistic regression. For an introduction to regression modelling in general, check out the Udacity Intro to Statistics Course which is available for free at www.udacity.com. For a more complete introduction to Logistic regression, watch select lectures from the edX Health in Numbers course PH207x available at www.edx.org.