Sample Statistical Workflow:
Simple Linear Regression

Adam Davey

February 7, 2015

> Test every assumption of your analysis you can. Know how every assumption you violate affects your results.
>
> Adam

1 Preamble

Simple linear regression models the association between a single predictor of interest (independent variable, predictor) and a response variable (dependent variable, outcome). Some of the most important assumptions include linearity of association between predictor and response, normality of residuals, and independence of predictors and residuals which includes homoscedasticity. Other important assumptions that we will address later include independence of observations, and that predictors are measured free from error.

2 Steps in a Simple Linear Regression Workflow

Below, we present a set of sample steps that should be performed whenever you are performing a simple linear regression analysis. It is not necessary for you to literally perform every one of these steps with every single analyses. However, this workflow should essentially cover the range of requirements and assumptions that you are expected to be responsible for under most circumstances. Know, however, that it is in no way exhaustive and that research conventions in your area of research may differ in important ways from the expectations for biostatistics. Likewise, many areas of biostatistics are developing very quickly and so current conventions are subject to change.

2.1 Frequencies/Summary Statistics

Look at the univariate distribution of both your predictor and your outcome variable using frequencies and/or summary statistics (M, SD, Min, Max, etc.). In Stata, you would use the `tab y x, missing` command and `summ y x`. If you have multiple groups (say treatment and control), then you will need to look at the distributions *within* levels of those grouping variables. You can use the `by` command (e.g., `by tx, sort: tab1 y x`) or in some cases produce the tables directly (e.g., `tab y x, missing` if x is the categorical variable. For continuous variables, the `tabstat` command is very useful (e.g., `tabstat y, by(x) columns(statistics) statistics(n mean sd min max)`). If your predictor is categorical or ordinal, be sure to look for categories with very small numbers of observations. If, for example, your predictor is race with 4 categories (e.g., White, African American, Asian, and American Indian) but you have a very small number of observations in one or more categories you may need to consider revising your coding so as to omit small groups or simplify coding (e.g., White / non-White).

2.2 Univariate plots (Y and X)

Like frequencies, univariate plots help you to understand the nature of your distributions. There are plenty of good ways to plot univariate data. At minimum, you should consider histograms (`hist y` or `hist y, by(x)`), boxplots (`graph box y` or `graph box y, over(x)`), and normal quantile plots (Q-Q plots, `qnorm y` – qnorm cannot be run by groups so you would have to run separately using `qnorm y if x==0` and `qnorm y if x==1`). In histograms, look at the shape of the distribution. Is it symmetric or asymmetric (i.e., skewed)? Is it unimodal or multimodal? Are there any gaps or bumps in the distribution? What is the range of observed values? How much of the possible range of values is actually observed? Check to ensure that the data do not contain impossible values, extreme values, and that missing value codes (e.g., -9, -99) are not being treated as observed values. Boxplots provide a different kind of information about your distributions. Specifically, depending on the software package, they display points such as the median, 25^{th} and 75^{th} %iles (for the box) and other functions of the interquartile range (for the whiskers) and extreme values. Normal quantile plots show you how your distribution aligns with the quantiles expected if the data followed a normal distribution. Deviations away from the 45° line show the location of deviations from normality. From these plots you can determine whether the points tend to pile up in some parts of the distribution, or if the distribution is highly skewed. These plots are also a good way to determine whether transformations to improve the shape of a distribution (such as a square root or log transformation) have been effective.

2.3 Testing Normality of Distributions

A number of formal statistical tests are available to test whether your distribution *deviates* significantly from normality. Stata offers several tests for normality. For sample sizes between $4 \leq N \leq 2000$ Stata suggests using the Shapiro-Wilk W test (`swilk y`). For sample sizes between $5 \leq N \leq 5000$ the Shapiro-Francia $W\prime$ test is available (`sfrancia y`).

There is also a more general Kolmogorov-Smirnov test that can compare the observed distribution to any kind of distribution (e.g., `summarize y; ksmirnov y = normal((y-r(mean))/r(sd))`) including comparing two distributions across a second variable (`ksmirnov y, by(x)`).
Some people consider these tests to be relatively uninformative for a variety of reasons.

1. First, in the situations where it is most likely to matter (i.e., your sample size is insufficient to rely on the central limit theorem), you have very lower power to detect deviations from normality.

2. Second, with reasonable sample sizes, many of these tests will return significant results even for relatively minor violations of normality.

3. Third, most statistical techniques are reasonably robust to violations of normality and so the real question is whether these violations could have affected the results in your analyses.

2.4 Scatterplot

Scatterplots are very useful bivariate plots to examine potential associations between pairs of variables. From these plots, you will get a sense of whether the association between two variables is strong (you can see it) or weak (you can't see it), positive (higher values in x correspond with higher values of y, i.e., positive slope) or negative (higher values in x correspond with lower values of y, i.e., negative slope), and whether points tend to bunch up in some locations, such as in a corner with a lot of values at or near the floor on both variables. This is a situation that can give undue influence to the small number of points that are not at the floor/ceiling. You can also get a sense of how the variance in one variable may vary as a function of the variance in the other variable. The basic syntax to create a scatterplot in Stata looks something like this `twoway (scatter y x, sort)` and is most useful with continuous predictors. If that works, great. In some cases, you will need to change the options such that points are larger or smaller, by modifying the "marker properties" option.

2.5 Lowess

"Linearity" is one of the key assumptions of simple linear regression. But this doesn't mean that regression can only fit straight lines. Rather, the "linear" part simply means "linear in the model parameters." By including both linear and quadratic terms in a model, for example, it is possible to model curvilinear associations (linear in the linear and quadratic terms). But in order to correctly specify the association, you first need to determine whether a linear association is appropriate and, if not, what an appropriate functional form for the association might be. One good solution here is to apply a lowess regression model, which stands for locally weighted smoothed regression (`lowess y x`). In looking at lowess curves, try not to pay too much attention to minor wiggles. That is almost always just normal sampling variation even when the true association follows a straight line. Likewise, in most data sets there are very few points at the extremes of your predictor variable. This

can sometimes lead to erratic behavior in the lines toward the extremes. In these cases, it is often helpful to "Winsorize" extreme values (http://en.wikipedia.org/wiki/Winsorising) which means to systematically recode high/low values to the corresponding value of some specific centile. If we wanted to recode extreme values to the corresponding 5^{th} and 95^{th} %iles, respectively, we could first find the centiles (`centile y, centile(5 95)`) and then recode (`recode y (min/xl=xl) (xu/max=xu) if y!=.` where xl and xu are the 5^{th} and 95^{th} %iles, respectively). Alternatively, you can type `findit winsor` for a program that will perform these transformation for you automatically. Repeat the lowess plot using the Winsorized variables. Any remaining curvature toward the extremes are now more credible as evidence of meaningful nonlinearity. In class, we will consider several ways of estimating models with nonlinear associations. For now, just worry about trying to determine whether a linear model is a reasonable way of modeling the association.

2.6 Estimate Model

If you have spent sufficient time on the preliminaries, getting to know your data and variables, then regression estimation is a straightforward task. We now regress the response variable on the predictor. In general, it's usually a good idea to estimate robust standard errors. In Stata, type `regress y x, vce(robust)` in case assumptions of homoscedasticity are not strictly met. If you have met the assumptions, then standard errors will be only very slightly larger that using the OLS approach; if you have violated the assumption, then these results will protect you against Type I errors in this case, so are to be preferred. (In fact, I almost always estimate the model both ways so that I know whether results differ, but then still present the robust standard errors.) With one predictor, that's almost always all there is to model estimation. In the multivariable approach, framework, we will considerably elaborate the steps and tools available to us at this point. At that point, we will also consider strategies for modeling nonlinear associations.

2.7 Evaluate Model

A standard regression model provides a variety of information that can be used to evaluate the results. For example, the formal test of association between the model predictor(s) and the response variable (i.e., that the regression coefficients are all 0) corresponds with the F test. Is the omnibus F test significant at your selected α? If not, you're done. There is no association between predictor(s) and criterion. You cannot reject the null hypothesis. If the F statistic is significant, then you can look at how much variance in the response variable your model predicts. This is estimated in the multiple R^2, which ranges from 0 (absolutely no association) to 1.0 (perfect association). What kind of R^2 value is "good" depends very strongly on the discipline and domain. In some areas, anything less than 0.7 might be considered a poor model fit, but in many areas of public health, anything greater than 0.1 might be considered a success. Anything less than 0.01 would generally be considered to lack practical importance for most applications. Finally, the model will provide you with estimated regression coefficients, standard errors, and the corresponding t-statistics and their p-values. In a simple linear regression model with 1 predictor, the p-value for the t-statistic for your model coefficient will be the same as for your omnibus F

statistic. Likewise, the value of F will be equal to y^2. When looking at your regression coefficients, start by looking at two things.

1. First, are the effects in the right (expected) direction? If your predictor is participation in the treatment group, does the direction of the effect go in the right direction? Is participation in the treatment associated with better outcomes? If the effect is not in the expected direction, there may be something wrong, such as with the coding of your variables. This is the time to check.

2. Second, do the standard errors look reasonable? If you have a large sample size, you should expect very precise estimates on the sampling distribution of the regression coefficients. If the standard errors are very large, there may be something wrong with the model and estimation. This is something we will revisit again when considering multivariable models. On the other hand, very small standard errors can also be a sign of problems with variable coding and estimation. Beware of small regression coefficients associated with large t-values. Although most statistics packages pay very careful attention to numerical accuracy, computers use something called floating point arithmetic which can sometimes introduce a great deal of imprecision.

2.8 Regression Diagnostics

Many books have been written on regression diagnostics, so we will only begin to scratch the surface in terms of what can be done to evaluate regression models. First, we need to calculate predicted values ($\hat{y}_i$s) and residuals (ϵ_is). Stata makes this easy. To obtain fitted values, type `predict yhat` (or whatever you want the variable containing predicted values to be called. To calculate the residuals, type `generate resid = y - yhat` (better yet, type `predict resid, resid`). One assumption of the regression model is that residuals are normally distributed. We can evaluate this with a histogram (`hist resid, normal`) and formal test of normality (`swilk resid`) and a QQ plot. We can test whether the dependent variable is correctly specified using a "hat test", which regresses the response variable on the estimated residuals and squared residuals (`linktest`). Look for a nonsignificant coefficient for the squared residuals (`_hatsq`). Heteroscedasticity can be tested statistically using `estat hettest` if the model is not estimated using robust standard errors and by `estat szroeter x` with robust standard errors. Influential points are those that have a high influence on the estimated regression coefficient values. They can be identified using `dfbeta x` if the model is not estimated using robust standard errors. Next, we need to consider the distribution of residuals in relation to other aspects of the regression model. Stata includes a variety of diagnostic plots for regression models. Good choices for all regression models include:

- A plot of residuals against fitted values. One assumption of the regression model is that the predicted values are independent of the residuals. So we expect to see no discernable pattern of associations. To obtain this plot in Stata, type `rvfplot, yline(0)`

- A plot of residuals against predicted values. The assumption of homoscedasticity suggests that the variance of the residuals is the same at all levels of the predictor. To obtain this plot in Stata, type `rvpplot x, yline(0)`.
- Leverage points are those far from other values in a model and they warrant additional attention once identified. A plot of leverage values by squared residuals can be obtained by typing `lvr2plot`.

2.9 Tabling Results

Different disciplines have different expectations for how results from regression models should be presented. In Public Health, critical information typically includes estimated regression coefficients, upper and lower confidence limits, and a p-value to 4 decimals. Many areas of the social sciences prefer presenting estimated regression coefficients, standard errors, t-values, and associated p-values. Stata contains a few features that can help for tabling regression output so that you don't have to type the values in (which is tedious, time-consuming, and error-prone).
Every time you estimate a regression model in Stata, you have the option to temporarily save the estimation results. After typing `regress y x`, for example, you can type `estimates store model1`. You can look at the list of saved estimates by typing `estimates dir`.
Save model estimates can be "replayed" (`estimates replay model1`) or tabled, albeit with somewhat limited options. For example, a reasonable table can be displayed using `estimates table model1, star(.05 .01 .001) stats(N r2_a)`. Much better, but somewhat more complicated to get the hang of is a routine called outreg2 (`findit outreg2`). This will export tables in a variety of formats including Word (RTF), Excel, CSV, LaTeX, text, and as Stata data files. Here is some sample syntax to export a serviceable regression table in Excel format for an epidemiological audience. `outreg2 [model1] using "RegTable", replace excel stats(coef ci pval) bdec(2) cdec(2) pdec(3) bracket(ci) noaster sideway`. It produces output similar to what is below. Note that p-values that display as 0.0000 have been changed to read 0.0001. Along with titles and headings, this should be all you need to change on the table generated from Stata.

Table 1: Sample Regression Table Output Produced by `outreg2` in Stata

	Model 11		
Variables	y	CI	p
x	1.44	[0.123 - 2.748]	0.032
Constant	7.74	[6.828 - 8.660]	0.001
Observations	160		
R-squared	0.029		

2.10 Writing Up Results

Prior to analysis, distributions for all variables were examined. The distribution of the response variable, y, was tested for normality and no substantial deviations were observed (Shapiro-Wilk W = 0.9835, p=0.0543). Preliminary evidence for linearity of association between x and y was examined by way of a scatterplot with lowess curve overlaid. (Actually, because x is dichotomous, the association can only be linear so this part should be omitted.) Y was regressed on x in simple linear regression with robust standard errors. The overall model suggested a significant association between x and y ($F(1, 158) = 4.67, p = 0.0323$) although x accounted for only a small proportion of the variance in y ($R^2 = 0.02$). The regression equation was $y = 7.74 + 1.44 \times x$, suggesting that participants with x scored 1.44 units higher (95%CI: 0.123-2.748) on y than participants without x. [Alternatively, for a continuous predictor, you could say something like "Each additional point higher on x was associated with 1.44 unit higher mean values of y."] Model regression coefficients, confidence intervals, and p-values are shown in Table 1. A variety of regression diagnostic plots were evaluated for violations of assumptions. Some slight violation of normality of residuals was observed (Shapiro-Wilk W = 0.9819, p=0.0342), but there was no evidence of heteroscedasticity ($\chi^2(1) = 0.06, p = 0.8084$). Inspection of leverage and influence plots did not identify any outliers or points with high influence on model parameters.
In Table 2 below, the same results are shown in a different format. Choose the one that best suits your needs and your disciplinary conventions.

Table 2: Another Sample Regression Table Output Produced by `outreg2` in Stata

	Model 2			
Variables	b	SE(b)	t	pl
x	1.44	0.66	2.16	0.0323
Constant	7.74	0.46	16.69	0.0000
Observations	160			
R-squared	0.029			

Checklist: Simple Linear Regression

☐ Frequencies and summary statistics for response and predictor.

☐ Test normality of response variable.

☐ Univariate plots (histograms, box plots, Q-Q plots) for response and predictor.

☐ Scatterplot of response by predictor with lowess plot.

☐ Estimate model with OLS and robust standard errors.

☐ Evaluate model. Proceed if omnibus (F) test significant; otherwise stop. Verify that coefficient is in the expected direction and that estimates and standard errors appear to have been estimated appropriately.

☐ Evaluate regression diagnostics. Calculate fitted values and residuals; test normality of residuals; evaluate variety of regression diagnostic plots (residuals by fitted, residuals by predictor; leverage by squared residuals); evaluate heteroscedasticity and influence statistics (dfbeta).

☐ Table results. Results should include regression coefficients, standard errors, t-values, and p-values, or else regression coefficients, 95% CI, and p-values, plus number of observations used in analysis and model R^2. Format your results in accordance with the style and conventions for your area of research.

☐ Write up results; note any violations of assumptions and their effects, if any, on results. Refer to model table.

Stata Syntax: Simple Linear Regression

```
/*****************************************************************
* Sample Simple Linear Regression Syntax
* PH 8012, Spring 2014
* Adam Davey
* Requires outreg2
* Type: findit outreg2 to install
*****************************************************************/

#delimit;
clear all;
capture log close;
log using "mylog.log", replace;

* Below simulates some data for the workflow;
set seed 12345;
set obs 1000;
generate age = int(20 + 60*runiform());
generate iq = int(100 + 16*rnormal() - 0.15*(age-50));

* Step 2.1 Frequencies / Summary Statistics;
tab1 iq age, missing;
summarize iq age;
* Alternatively, use syntax below for considerably more information;
*summarize iq age, detail;

* Step 2.2 Univariate Plots;
hist iq;
graph export "hist_iq.pdf", replace;
graph box iq;
graph export "box_iq.pdf", replace;
qnorm iq;
graph export "qnorm_iq.pdf", replace;

hist age;
graph export "hist_age.pdf", replace;
graph box age;
graph export "box_age.pdf", replace;
qnorm age;
graph export "qnorm_age.pdf", replace;

* Step 2.3 Testing Normality of Distributions;
swilk iq;
swilk age;
```

```
* Step 2.4 Scatterplot;
scatter iq age, sort;
graph export "scatter_iq_age.pdf", replace;

* Step 2.5 Lowess;
lowess iq age, sort;
graph export "lowess_iq_age.pdf", replace;

* Step 2.6 Estimate Model;
* First, estimate and store OLS model;
regress iq age;
estimates store ols;

* Then estimate and store model with robust SEs;
regress iq age, vce(robust);
estimates store robust;

* Step 2.7 Evaluate Model;
* No additional analyses or syntax for this section, so here's a handy figure;
graph twoway (lfitci iq age) (scatter iq age);
graph export "twoway_fitted_cis.pdf", replace;

* Step 2.8 Regression Diagnostics;
* Things we can do with robust standard errors;
* First ensure robust standard error model is active;
estimates restore robust;
* Predicted values;
predict iqhat, xb;
* Predicted residuals;
predict iqres, resid;
hist iqres, normal;
graph export "hist_iqres.pdf", replace;
swilk iqres;
* Hat test;
linktest;
qnorm iqres;
graph export "qnorm_iqres.pdf", replace;
szroeter age;
rvfplot, yline(0);
graph export "rvfplot.pdf", replace;
rvpplot age, yline(0);
graph export "rvpplot_age.pdf", replace;

* Things we can do with OLS standard errors;
```

```
* First restore OLS estimates;
estimates restore ols;
estat hettest;
dfbeta age;
lvr2plot;
graph export "lvr2plot.pdf", replace;

* Step 2.9 Tabling Results;
* NEVER type numbers you don't need to;
* Let the computer do the heavy lifting for you;
* Example 1 -- direct to Excel;
* Crap table, but all the information is where you put it;
quietly: estimates replay robust;
putexcel set "regression1.xls", sheet("Robust SEs") replace;
putexcel F1=("Number of obs") G1=(e(N));
putexcel F2=("F")             G2=(e(F));
putexcel F3=("Prob > F")      G3=(Ftail(e(df_m), e(df_r), e(F)));
putexcel F4=("R-squared")     G4=(e(r2));
putexcel F5=("Adj R-squared") G5=(e(r2_a));
putexcel F6=("Root MSE")      G6=(e(rmse));
matrix a = r(table)';
matrix a = a[.,1..6];
putexcel A8=matrix(a, names);
quietly: estimates replay ols;
putexcel set "regression1.xls", sheet("OLS SEs") modify;
putexcel F1=("Number of obs") G1=(e(N));
putexcel F2=("F")             G2=(e(F));
putexcel F3=("Prob > F")      G3=(Ftail(e(df_m), e(df_r), e(F)));
putexcel F4=("R-squared")     G4=(e(r2));
putexcel F5=("Adj R-squared") G5=(e(r2_a));
putexcel F6=("Root MSE")      G6=(e(rmse));
matrix a = r(table)';
matrix a = a[.,1..6];
putexcel A8=matrix(a, names);
* Example 2 -- also direct to Excel;
* Nicer table, but labels need to be changed;
outreg2 [robust ols] using "regression2",
 replace excel stats(coef ci pval)
 bdec(2) cdec(2) pdec(3) bracket(ci) noaster sideway;

log close;

* Convert text output to PDF;
translate mylog.log mylog.pdf, replace;
```