

# Stata Software for Measurement Error Models

This material was put together by Roberto G. Gutierrez of the Stata Corporation. Users should consult <http://www.stata.com/merror/> as the web site for this software.

## 1 Installing the software

Installing the software of Hardin & Carroll (2003) is very easy, and requires Stata version 8.0 or later (the current version as of the publication of our book is Stata 9.1). From a “net-aware” Stata, simply type

```
. net install http://www.stata-journal/software/sj3-4/st0049
```

which installs the following Stata programs

`qvf` – Generalized linear models with (possibly) instrumental variables and fast bootstrap

`rcal` – Generalized linear models with regression calibration

`simex` – Simulation extrapolation in generalized linear models

`simexplot` – SIMEX plots after estimation with `simex`

If you find the exact website hard to remember, you can use Stata’s net searching tools to find the software. The Stata command, `findit`, is very useful in this regard, e.g.

```
. findit simex
```

or

```
. findit measurement error models
```

will allow you to then point-and-click your way towards installing the appropriate software.

## 2 Getting the Framingham Data

The Framingham data are in Stata format, saved as `framingham.dta`. It can be loaded into Stata, described, and partially listed as follows

```
. use framingham, clear  
(Framingham Heart Study)
```

```
. describe
```

```
Contains data from framingham.dta
```

```
obs:      1,615      Framingham Heart Study  
vars:      14      20 Jan 2006 09:31  
size:      62,985 (99.4% of memory free)  (_dta has notes)
```

variable name	storage type	display format	value label	variable label
age	byte	%9.0g		age at exam 2
sbp21	int	%9.0g		first systolic blood pressure at exam 2
sbp22	int	%9.0g		second systolic blood pressure at exam 2
sbp31	int	%9.0g		first systolic blood pressure at exam 3
sbp32	int	%9.0g		second systolic blood pressure at exam 3

```

smoke      byte   %9.0g      present smoking at exam 1
cholest2   int    %9.0g      serum cholesterol at exam 2
cholest3   int    %9.0g      serum cholesterol at exam 3
firstchd   byte   %9.0g      first evidence of CHD occurring
              at exam 3 through 6 (1 == yes)

sbp2       float  %9.0g      (sbp21 + sbp22) / 2
sbp3       float  %9.0g      (sbp31 + sbp32) / 2
lsbp2      float  %9.0g      log(sbp2 - 50)
lsbp3      float  %9.0g      log(sbp3 - 50)
lcholest3  float  %9.0g      log(cholest3)

```

---

Sorted by:

. notes

.\_dta:

1. Source: Carroll, Ruppert, Stefanski & Crainiceanu (2006)

. list firstchd age smoke cholest3 in 1/10

	firstchd	age	smoke	cholest3
1.	0	56	0	295
2.	0	38	1	255
3.	0	54	1	287
4.	0	42	1	285
5.	0	47	1	240
6.	0	43	1	208
7.	0	58	1	208
8.	0	43	0	334
9.	0	43	1	242
10.	1	36	0	244

The dataset is titled *Framingham Heart Study*. It should be available on the Stata website for direct download within Stata. That is, readers should be able to type

```
. use http://www.stata-press.com/data/nlmem/framingham, clear
```

and they would then have the data ready for analysis in Stata. The directory name `nlmem` stands for Non-linear measurement error models.

### 3 Analysis With One Error-Prone Covariate, Known Error Variance

#### 3.1 Background

- Variables  $Z$  measured without error: `age`, `smoke` (smoking status) and `cholest3` (Cholesterol)
- Variable  $W$  measured with known measurement error: `lsbp3` (Transformed systolic blood pressure)
- Measurement error variance: 0.1260

#### 3.2 Naive Analysis

We begin with a naive analysis, treating `lsbp3` as if it were measured exactly. In Stata, this analysis can be done with the commands `logistic` or `glm`, but here we use `qvf` which, when used without instrumental variables, is analogous to `glm` with option `irls` for iterated re-weighted least squares.

```
. qvf firstchd age smoke cholest3 lsbp3, family(binomial)
Generalized linear models          No. of obs   =   1615
Optimization      : MQL Fisher scoring      Residual df   =   1610
                    (IRLS EIM)              Scale param   =     1
Deviance          =   825.345216             (1/df) Deviance = .5126368
Pearson          =   1462.41452             (1/df) Pearson  = .908332
Variance Function: V(u) = u(1-u)           [Bernoulli]
Link Function    : g(u) = log(u/(1-u))     [Logit]
Standard Errors  : EIM Hessian
```

firstchd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.05683	.0117455	4.84	0.000	.0338092	.0798508
smoke	.5731081	.2497074	2.30	0.022	.0836907	1.062526
cholest3	.0078488	.0021162	3.71	0.000	.003701	.0119965
lsbp3	1.524242	.3887089	3.92	0.000	.7623869	2.286098
_cons	-14.1818	1.786893	-7.94	0.000	-17.68404	-10.67955

noting that we simply need to specify `family(binomial)` since the logit link is canonical (and the default) for that family.

The standard errors above are analytical. If we instead wanted bootstrap standard errors, we add options `bstrap` and `seed()`, the latter necessary only for reproducibility.

```
. qvf firstchd age smoke cholest3 lsbp3, family(binomial) bstrap seed(10001)
Generalized linear models          No. of obs   =   1615
Optimization      : MQL Fisher scoring      Residual df   =   1610
                    (IRLS EIM)              Scale param   =     1
Deviance          =   825.345216             (1/df) Deviance = .5126368
Pearson          =   1462.41452             (1/df) Pearson  = .908332
Variance Function: V(u) = u(1-u)           [Bernoulli]
Link Function    : g(u) = log(u/(1-u))     [Logit]
Standard Errors  : Bootstrap
```

firstchd	Coef.	Bootstrap Std. Err.	z	P> z	[95% Conf. Interval]	
age	.05683	.0111139	5.11	0.000	.0350473	.0786128
smoke	.5731081	.2337721	2.45	0.014	.1149231	1.031293
cholest3	.0078488	.0019552	4.01	0.000	.0040167	.0116808
lsbp3	1.524242	.345349	4.41	0.000	.8473706	2.201114
_cons	-14.1818	1.5101	-9.39	0.000	-17.14154	-11.22205

For all the software covered here, the default number of bootstrap replications is 199, but this can be overruled via option `brep()`, e.g. `brep(500)`.

### 3.3 Regression Calibration

For a regression calibration analysis, we use command `rcal`. In this example, we assume the measurement error variance is known and thus we need to define the known variance as a Stata matrix and pass this matrix to `rcal` by using option `suunit()`.

```
. matrix U = 0.01260 // define the m.e. variance
. rcal (firstchd = age smoke cholest3) (wlsbp:lslbp3), family(binomial) suunit(
> U)

Regression calibration                No. of obs      =      1615
Link          = Logit
Residual df =          1610                Wald F(4,1610)   =      17.58
                                                Prob > F        =      0.0000
                (IRLS EIM)                Scale param     =      .908332
Variance Function: V(u) = u(1-u)        [Bernoulli]
Link Function   : g(u) = log(u/(1-u))    [Logit]
```

firstchd	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0537147	.0114134	4.71	0.000	.031328	.0761014
smoke	.5816822	.2380915	2.44	0.015	.1146804	1.048684
cholest3	.0075968	.0020234	3.75	0.000	.0036281	.0115655
wlsbp	2.047573	.5001969	4.09	0.000	1.066468	3.028678
_cons	-16.26699	1.703024	-9.55	0.000	-19.60736	-12.92661

In the above, note the slightly different syntax for specifying the model: the response and covariates measured without error are in the first *equation*, separated by parentheses, while the single variable measured with error, `lslbp3`, comprises the second equation, labelled as `wlsbp`.

As was the case last time, the above are analytical standard errors. Switching to bootstrap standard errors works just as before

```
. rcal (firstchd = age smoke cholest3) (wlsbp:lslbp3), family(binomial) suunit(
> U) bstrap seed(10002)

Regression calibration                No. of obs      =      1615
                                                Bootstrap reps  =       199
Residual df =          1610                Wald F(4,1610)   =      20.05
                                                Prob > F        =      0.0000
                (IRLS EIM)                Scale param     =      .908332
Variance Function: V(u) = u(1-u)        [Bernoulli]
Link Function   : g(u) = log(u/(1-u))    [Logit]
```

firstchd	Coef.	Bootstrap Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0537147	.0108181	4.97	0.000	.0324956	.0749338
smoke	.5816822	.2412495	2.41	0.016	.1084861	1.054878
cholest3	.0075968	.0020114	3.78	0.000	.0036516	.0115421
wlsbp	2.047573	.5125924	3.99	0.000	1.042154	3.052992
_cons	-16.26699	2.189779	-7.43	0.000	-20.5621	-11.97187

### 3.4 SIMEX

Analysis using SIMEX is achieved using the `simex` command. The syntax is identical to that for `rcal`, with the exception that analytical standard errors are not available. By default `simex` will

not calculate any standard errors, but we can request bootstrap standard errors as previously

```
. simex (firstchd = age smoke cholest3) (wlsbp:lspb3), family(binomial) suunit
> (U) bstrap seed(10003)
Estimated time to perform bootstrap: 2.28 minutes.

Simulation extrapolation          No. of obs      =      1615
                                Bootstraps reps =       199
Residual df =      1610          Wald F(4,1610) =      20.66
                                Prob > F          =      0.0000

Variance Function: V(u) = u(1-u)      [Bernoulli]
Link Function      : g(u) = log(u/(1-u)) [Logit]
```

firstchd	Coef.	Bootstrap Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0541254	.0120696	4.48	0.000	.0304516	.0777992
smoke	.5822821	.2290522	2.54	0.011	.1330103	1.031554
cholest3	.007754	.002152	3.60	0.000	.003533	.0119751
wlsbp	1.854374	.4400707	4.21	0.000	.991202	2.717545
_cons	-15.48957	1.834077	-8.45	0.000	-19.087	-11.89214

Once `simex` has been run, we can examine the extrapolant function with command `simexplot`

```
. simexplot wlsbp
```

producing the graph in Figure 1.

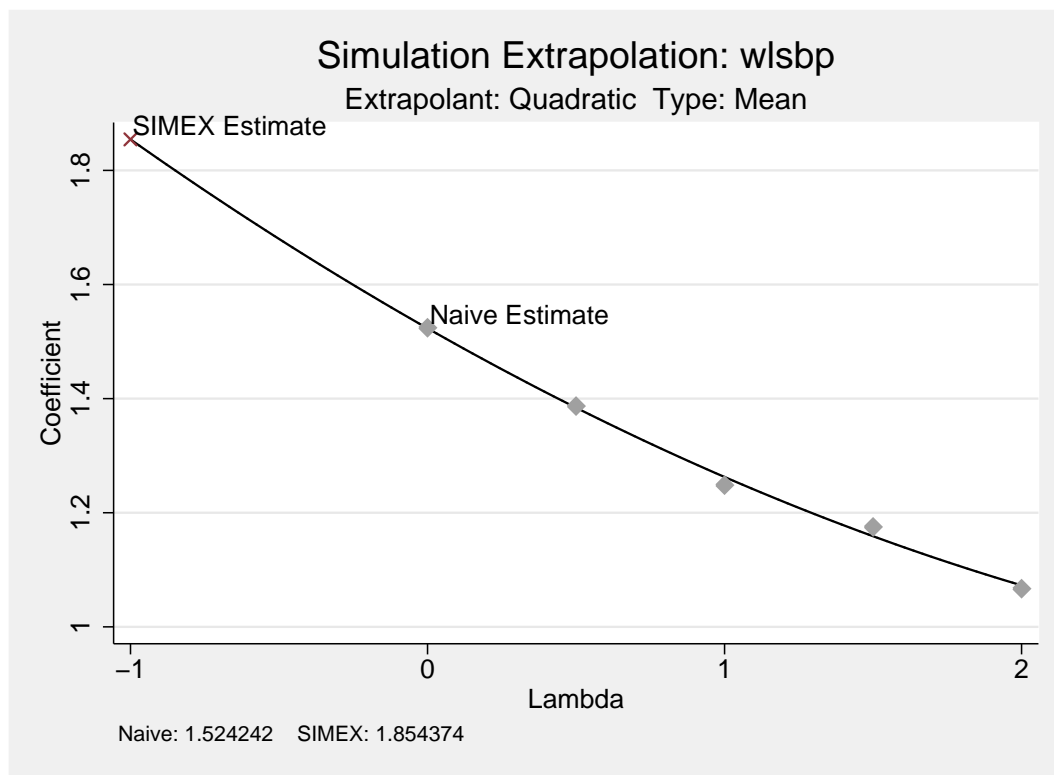


Figure 1: Simex plot for proposed analysis #1

The quadratic extrapolant function is the default for `simex`, with `linear` and `rational` being options. In this example, but the non-linear (rational linear) curve fit refused to converge and estimation reverted to the quadratic extrapolant.

## 4 Analysis With One Error-Prone Covariates, Replicates

### 4.1 Background

- Variables  $Z$  measured without error: age, smoke, cholest3
- Variable  $W$  measured with unknown measurement error: wlsbp with two replicate measurements lsbp2 and lsbp3
- Measurement error variance: unknown, estimated from the replicates

### 4.2 Naive Analysis

For the naive analysis, we want the average of the replicates, lsbp2 and lsbp3, to be the variable measured without error, in which case we need to generate this new variable first

```
. gen mlsbp = (lsbp2 + lsbp3) / 2
. qvf firstchd age smoke cholest3 mlsbp, family(binomial)

Generalized linear models          No. of obs   =       1615
Optimization      : MQL Fisher scoring      Residual df   =       1610
                  (IRLS EIM)              Scale param   =         1
Deviance          =    824.099326           (1/df) Deviance = .5118629
Pearson          =    1471.51585           (1/df) Pearson = .913985
Variance Function: V(u) = u(1-u)          [Bernoulli]
Link Function     : g(u) = log(u/(1-u))    [Logit]
Standard Errors  : EIM Hessian
```

firstchd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.0554057	.0118037	4.69	0.000	.0322709	.0785405
smoke	.5930756	.2499898	2.37	0.018	.1031045	1.083047
cholest3	.0078725	.0021084	3.73	0.000	.0037402	.0120048
mlsbp	1.706105	.4174073	4.09	0.000	.8880022	2.524209
_cons	-14.94906	1.899273	-7.87	0.000	-18.67156	-11.22655

for analytical standard errors. Switching to bootstrap

*(Continued on next page)*

```

. qvf firstchd age smoke cholest3 mlsbp, family(binomial) bstrap seed(10004)
Generalized linear models                No. of obs    =    1615
Optimization      : MQL Fisher scoring   Residual df   =    1610
                  (IRLS EIM)           Scale param   =     1
Deviance          =    824.099326        (1/df) Deviance = .5118629
Pearson          =    1471.51585         (1/df) Pearson  = .913985
Variance Function: V(u) = u(1-u)        [Bernoulli]
Link Function    : g(u) = log(u/(1-u))   [Logit]
Standard Errors  : Bootstrap

```

firstchd	Coef.	Bootstrap Std. Err.	z	P> z	[95% Conf. Interval]	
age	.0554057	.0103033	5.38	0.000	.0352116	.0755998
smoke	.5930756	.2389307	2.48	0.013	.1247801	1.061371
cholest3	.0078725	.0018464	4.26	0.000	.0042536	.0114913
mlsbp	1.706105	.4232597	4.03	0.000	.8765317	2.535679
_cons	-14.94906	1.797272	-8.32	0.000	-18.47165	-11.42647

### 4.3 Regression Calibration

For regression calibration, note that the equation labelled as `wlsbp` now has two variables in it, representing the replicate measurements. Since we estimate the measurement error by using the replicates, the option `suunit()` is no longer necessary

```

. rcal (firstchd = age smoke cholest3) (wlsbp:lsbp2 lsbp3), family(binomial)
Regression calibration                No. of obs    =    1615
Link      = Logit
Residual df =    1610                Wald F(4,1610) =    17.85
                                           Prob > F       =    0.0000
                                           Scale param   =    .913985
Variance Function: V(u) = u(1-u)        [Bernoulli]
Link Function    : g(u) = log(u/(1-u))   [Logit]

```

firstchd	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0535814	.0114046	4.70	0.000	.031212	.0759508
smoke	.6012368	.2391313	2.51	0.012	.1321953	1.070278
cholest3	.0077435	.0020176	3.84	0.000	.0037861	.0117009
wlsbp	2.010981	.4718496	4.26	0.000	1.085477	2.936485
_cons	-16.1729	1.815754	-8.91	0.000	-19.73439	-12.61141

At this point we mention that, whenever appropriate, you can add option `robust` to get Sandwich estimates of standard errors. This is fairly standard throughout all of Stata.

*(Continued on next page)*

```

. rcal (firstchd = age smoke cholest3) (wlsbp:lspb2 lspb3), family(binomial) ro
> bust
Regression calibration                               No. of obs      =      1615
Link          = Logit
Residual df   =      1610                          Wald F(4,1610)   =      21.65
                                                    Prob > F         =      0.0000
                                                    Scale param     =      .913985
                (IRLS EIM)
Variance Function: V(u) = u(1-u)                  [Bernoulli]
Link Function   : g(u) = log(u/(1-u))              [Logit]

```

firstchd	Semi-Robust		t	P> t	[95% Conf. Interval]	
	Coef.	Std. Err.				
age	.0535814	.0107486	4.98	0.000	.0324986	.0746642
smoke	.6012368	.2442884	2.46	0.014	.12208	1.080394
cholest3	.0077435	.0019718	3.93	0.000	.0038759	.0116112
wlsbp	2.010981	.4636976	4.34	0.000	1.101467	2.920496
_cons	-16.1729	1.735387	-9.32	0.000	-19.57676	-12.76905

And, finally, bootstrap standard errors

```

. rcal (firstchd = age smoke cholest3) (wlsbp:lspb2 lspb3), family(binomial) bs
> trap seed(10005)
Regression calibration                               No. of obs      =      1615
                                                    Bootstrap reps  =      199
Residual df   =      1610                          Wald F(4,1610)   =      23.15
                                                    Prob > F         =      0.0000
                                                    Scale param     =      .913985
                (IRLS EIM)
Variance Function: V(u) = u(1-u)                  [Bernoulli]
Link Function   : g(u) = log(u/(1-u))              [Logit]

```

firstchd	Bootstrap		t	P> t	[95% Conf. Interval]	
	Coef.	Std. Err.				
age	.0535814	.0106838	5.02	0.000	.0326257	.0745371
smoke	.6012368	.2218764	2.71	0.007	.1660399	1.036434
cholest3	.0077435	.001876	4.13	0.000	.0040639	.0114231
wlsbp	2.010981	.4641468	4.33	0.000	1.100586	2.921377
_cons	-16.1729	1.982635	-8.16	0.000	-20.06172	-12.28409

## 4.4 SIMEX

For SIMEX we have

```

. simex (firstchd = age smoke cholest3) (wlsbp:lspb2 lspb3), family(binomial) b
> strap seed(10006)
Estimated time to perform bootstrap: 2.25 minutes.
Simulation extrapolation                               No. of obs      =      1615
                                                    Bootstraps reps =      199
Residual df   =      1610                          Wald F(4,1610)   =      21.61
                                                    Prob > F         =      0.0000
Variance Function: V(u) = u(1-u)                  [Bernoulli]
Link Function   : g(u) = log(u/(1-u))              [Logit]

```

firstchd	Bootstrap		t	P> t	[95% Conf. Interval]	
	Coef.	Std. Err.				
age	.0537021	.0112988	4.75	0.000	.0315403	.075864
smoke	.5991448	.2520173	2.38	0.018	.1048283	1.093461
cholest3	.0077896	.0020684	3.77	0.000	.0037324	.0118467
wlsbp	1.905424	.4326691	4.40	0.000	1.05677	2.754078

_cons	-15.73008	1.878278	-8.37	0.000	-19.41421	-12.04596
-------	-----------	----------	-------	-------	-----------	-----------

and plotted with

```
. simexplot wlsbp
```

producing the graph in Figure 2.

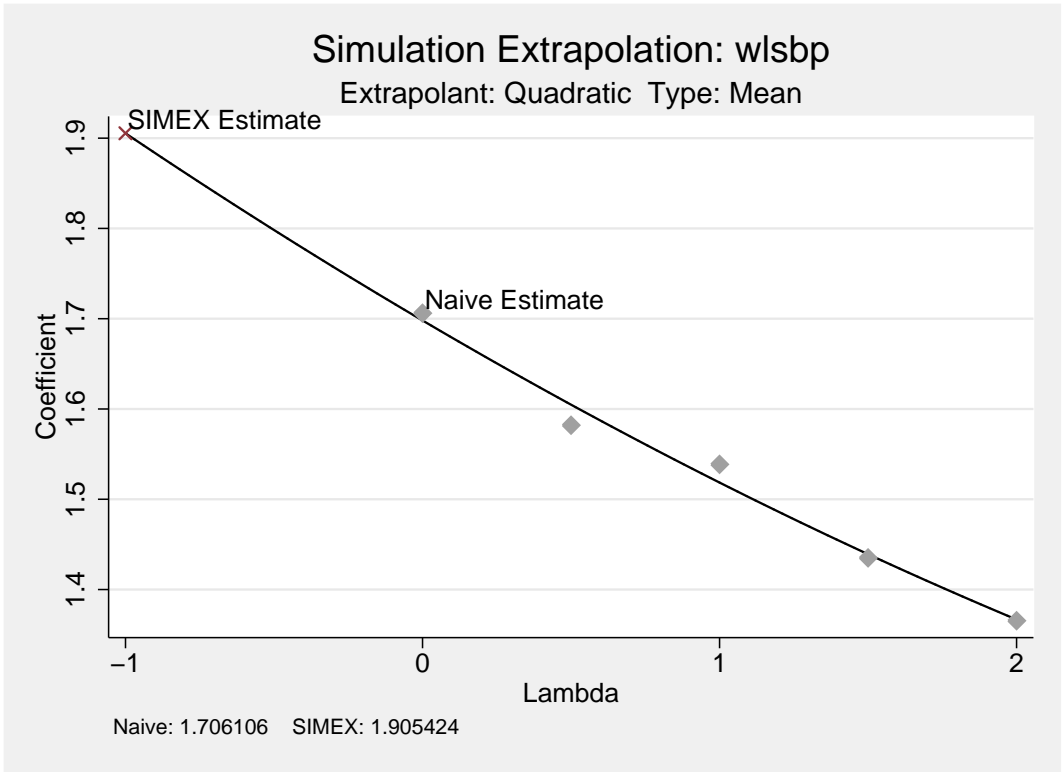


Figure 2: Simex plot for proposed analysis #2

## 5 Analysis With Two Error-Prone Covariates, Known Error Covariance Matrix

### 5.1 Background and Important Calling Convention

In this analysis, we have two variables measured with error, namely transformed systolic blood pressure and log-cholesterol. In the example, the measurement error covariance matrix is assumed known.

There is an **important convention** here: we input the measurement error covariance matrix so that the diagonals correspond to the measurement error variances for transformed systolic blood pressure first, and then log-cholesterol. When calling regression calibration and SIMEX, the variables must be entered exactly in the same order, or strange results will occur.

- Variables  $Z$  measured without error: `age`, `smoke`
- Variables  $W$  measured with known measurement error: `lsbp3` and `cholest3`
- Measurement error variance known:

$$\mathbf{V} = \begin{bmatrix} 0.01260208144393 & 0.00067287539939 \\ 0.00067287539939 & 0.00845894714763 \end{bmatrix}$$

### 5.2 Naive Analysis

```
. qvf firstchd age smoke lcholest3 lsbp3, family(binomial)
Generalized linear models          No. of obs   =    1615
Optimization      : MQL Fisher scoring      Residual df   =    1610
                  (IRLS EIM)              Scale param   =     1
Deviance          =    824.240423           (1/df) Deviance = .5119506
Pearson          =    1458.82744           (1/df) Pearson  = .906104
Variance Function: V(u) = u(1-u)          [Bernoulli]
Link Function    : g(u) = log(u/(1-u))    [Logit]
Standard Errors  : EIM Hessian
```

firstchd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.056446	.0117413	4.81	0.000	.0334334	.0794585
smoke	.572659	.2498046	2.29	0.022	.0830509	1.062267
lcholest3	2.039176	.5435454	3.75	0.000	.9738468	3.104506
lsbp3	1.518676	.3889605	3.90	0.000	.7563275	2.281025
_cons	-23.39799	3.413942	-6.85	0.000	-30.0892	-16.70679

With bootstrap standard errors:

```
. qvf firstchd age smoke lcholest3 lsbp3, family(binomial) bstrap seed(10001)
Generalized linear models          No. of obs   =    1615
Optimization      : MQL Fisher scoring      Residual df   =    1610
                  (IRLS EIM)              Scale param   =     1
Deviance          =    824.240423           (1/df) Deviance = .5119506
Pearson          =    1458.82744           (1/df) Pearson  = .906104
Variance Function: V(u) = u(1-u)          [Bernoulli]
Link Function    : g(u) = log(u/(1-u))    [Logit]
Standard Errors  : Bootstrap
```

firstchd	Coef.	Bootstrap Std. Err.	z	P> z	[95% Conf. Interval]	
age	.056446	.0111898	5.04	0.000	.0345143	.0783776

smoke	.572659	.2337119	2.45	0.014	.1145921	1.030726
lcholest3	2.039176	.4900323	4.16	0.000	1.078731	2.999622
lsbp3	1.518676	.3510057	4.33	0.000	.8307176	2.206635
_cons	-23.39799	2.926921	-7.99	0.000	-29.13465	-17.66133

### 5.3 Regression calibration

```
. mat V = (0.01260208144393, 0.00067287539939 \ 0.00067287539939, 0.00845894714
> 763)
. rcal (firstchd = age smoke) (wlsbp:lsbp3) (wcholest:lcholest3), family(binomi
> al) suunit(V)
Regression calibration                               No. of obs      =      1615
Link          = Logit
Residual df   =          1610                        Wald F(4,1610)   =      17.63
                                                    Prob > F         =      0.0000
                                                    Scale param     =      .906104
(IRLS EIM)
Variance Function: V(u) = u(1-u)                    [Bernoulli]
Link Function   : g(u) = log(u/(1-u))                [Logit]
```

firstchd	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0526627	.0113888	4.62	0.000	.0303243 .0750012
smoke	.5774215	.2379919	2.43	0.015	.1106151 1.044228
wlsbp	2.014178	.4997267	4.03	0.000	1.033995 2.994361
wcholest	2.768454	.7114252	3.89	0.000	1.373038 4.163871
_cons	-29.33517	3.249715	-9.03	0.000	-35.70928 -22.96105

With bootstrap standard errors:

```
. rcal (firstchd = age smoke) (wlsbp:lsbp3) (wcholest:lcholest3), family(binomi
> al) suunit(V) bstrap seed(10007)
Regression calibration                               No. of obs      =      1615
                                                    Bootstrap reps  =       199
Residual df   =          1610                        Wald F(4,1610)   =      21.74
                                                    Prob > F         =      0.0000
                                                    Scale param     =      .906104
(IRLS EIM)
Variance Function: V(u) = u(1-u)                    [Bernoulli]
Link Function   : g(u) = log(u/(1-u))                [Logit]
```

firstchd	Coef.	Bootstrap Std. Err.	t	P> t	[95% Conf. Interval]
age	.0526627	.01113	4.73	0.000	.0308319 .0744936
smoke	.5774215	.2429321	2.38	0.018	.1009251 1.053918
wlsbp	2.014178	.5024773	4.01	0.000	1.0286 2.999756
wcholest	2.768454	.6975423	3.97	0.000	1.400268 4.136641
_cons	-29.33517	3.858431	-7.60	0.000	-36.90324 -21.76709

### 5.4 SIMEX

```
. simex (firstchd = age smoke) (wlsbp:lsbp3) (wcholest:lcholest3), family(binom
> ial) suunit(V) bstrap seed(10008)
Estimated time to perform bootstrap: 2.40 minutes.
Simulation extrapolation                               No. of obs      =      1615
                                                    Bootstraps reps =       199
Residual df   =          1610                        Wald F(4,1610)   =      24.10
                                                    Prob > F         =      0.0000
```

Variance Function:  $V(u) = u(1-u)$  [Bernoulli]  
 Link Function :  $g(u) = \log(u/(1-u))$  [Logit]

firstchd	Coef.	Bootstrap Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0545443	.0099631	5.47	0.000	.0350023	.0740863
smoke	.5803764	.2638591	2.20	0.028	.0628329	1.09792
wlsbp	1.84699	.4529421	4.08	0.000	.9585718	2.735408
wcholest	2.5346	.7278619	3.48	0.001	1.106944	3.962256
_cons	-27.44831	4.231603	-6.49	0.000	-35.74834	-19.14828

SIMEX plots:

```
. simexplot wcholest
```

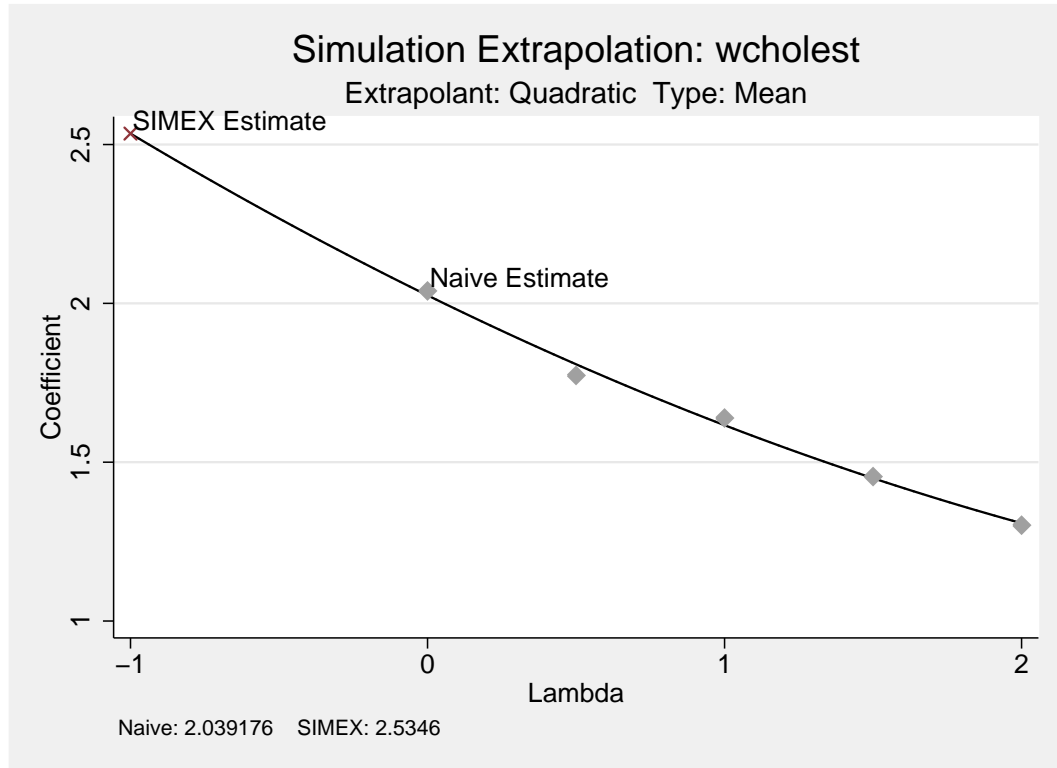


Figure 3: Simex plot for proposed analysis #3, wcholest

```
. simexplot wlsbp
```

and the graphs produced are included at the end of this document. Finally, note that the two final graphs may be combined into one.

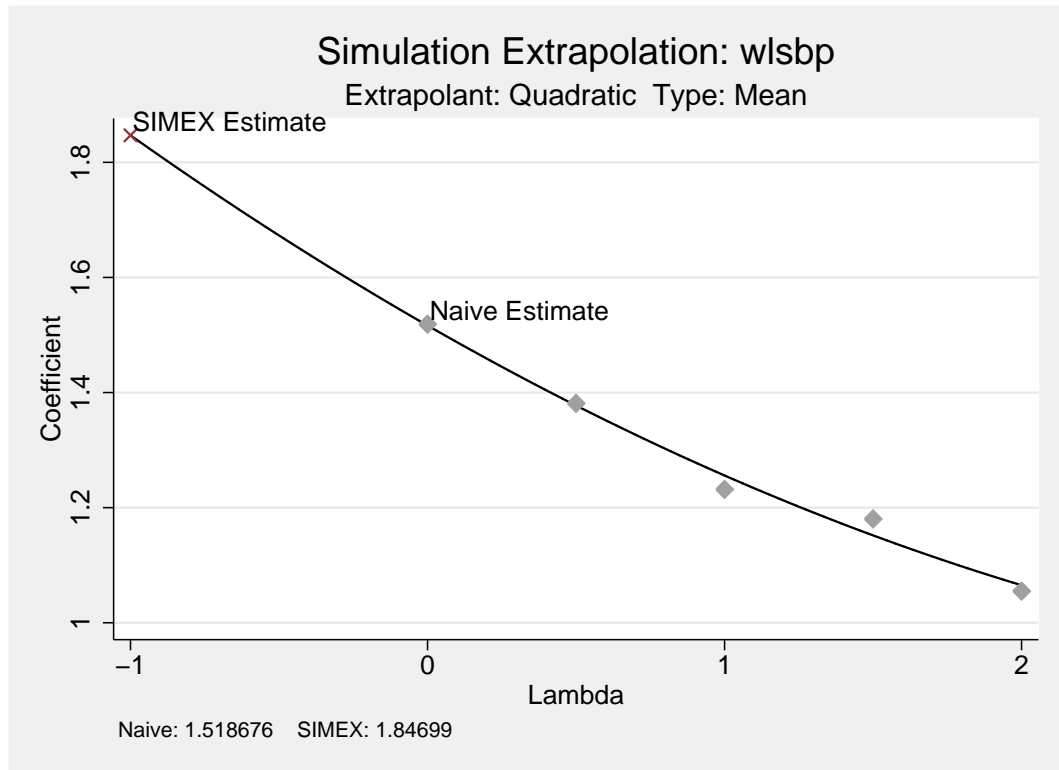


Figure 4: Simex plot for proposed analysis #3, wlsbp