

Lab 5

Multiple Linear Regression

In this lab we will discuss examples of model selection in multiple linear regression.

We will use two datasets. The first is the Peru bloodpressure data from lecture 4, and the second is the births data from the Stata book (in hs8).

We will look at four of the six automated model selection methods Stata offers: Forward, Backward, Forward Stepwise, Backward Stepwise (and not look at Forward Hierarchical or Backward Hierarchical which are hypothesis driven).

To reiterate Ron's last paragraph about stepwise methods receiving criticism for exaggerating relationships between variables, some of that criticism is provided here: <http://www.stata.com/support/faqs/stat/stepwise.html>.

Forward selection:

The simplest data-driven model building approach is called forward selection. In this approach, one adds variables to the model one at a time. At each step, each variable that is not already in the model is tested for inclusion in the model. The most significant of these variables is added to the model, so long as its p-value is below some pre-set level. It is customary to set this value above the conventional .05 level at say .10 or .15, because of the exploratory nature of this method.

```
. stepwise, pe(0.1): regress response predictors...
```

Backward selection:

Forward selection has drawbacks, including the fact that each addition of a new variable may render one or more of the already included variables non-significant. An alternate approach which avoids this is backward selection. Under this approach, one starts with fitting a model with all the variables of interest (following the initial screen). Then the least significant variable is dropped, so long as it is not significant at our chosen critical level. We continue by successively re-fitting reduced models and applying the same rule until all remaining variables are statistically significant.

```
. stepwise, pr(0.1): regress response predictors...
```

Stepwise selection:

Stepwise selection is a method that allows moves in either direction, dropping or adding variables at the various steps. **Backward stepwise** selection involves starting off in a backward approach and then potentially adding back variables if they later appear to be significant. The process is one of alternation between choosing the least significant variable to drop and then re-considering all dropped variables (except the most recently dropped) for re-introduction into the model. This means that two separate significance levels must be chosen for deletion from the model and for adding to the model. The adding significance must be more stringent (smaller p-value) than the deletion significance (larger p-value).

Forward stepwise selection is also a possibility, though not as common. In the forward approach, variables once entered may be dropped if they are no longer significant as other variables are added.

```
. stepwise, pe(0.1) pr(0.15): regress response predictors...  
. stepwise, pe(0.1) pr(0.15) forward: regress response predictors...
```

Use the Peru dataset and generate the variable fraction (as in lecture 4).

```
clear
use peru
generate fraction=years/age
list in 1/5
```

```
-----+-----
| age  years  weight  height  chin  forearm  calf  pulse  systol  diastol  fraction |
|-----+-----|
1. | 21      1      71     1629    8      7     12.7   88     170     76     .047619 |
2. | 22      6     56.5   1569   3.3     5      8      64     120     60     .2727273 |
3. | 24      5      56     1561   3.3     1.3    4.3    68     125     75     .2083333 |
4. | 24      1      61     1619   3.7     3      4.3    52     148     120    .0416667 |
5. | 25      1      65     1566    9     12.7   20.7   72     140     78      .04 |
|-----+-----
```

A big scatterplot matrix with the response variable systol at top.

```
graph matrix systol fraction weight height chin forearm calf pulse
```

Get help on the stepwise procedure so that the syntax makes sense.

```
help stepwise
```

Backward selection. The predictor variable with the largest p-value larger than 0.10 is removed from the model.

```
stepwise, pr(0.1): regress systol fraction weight height chin forearm calf
pulse
```

```
begin with full model
p = 0.8637 >= 0.1000  removing calf
p = 0.6953 >= 0.1000  removing pulse
p = 0.6670 >= 0.1000  removing forearm
p = 0.2745 >= 0.1000  removing height
p = 0.1534 >= 0.1000  removing chin
```

Source	SS	df	MS	Number of obs =	39
Model	3090.07324	2	1545.03662	F(2, 36) =	16.16
Residual	3441.36266	36	95.5934072	Prob > F =	0.0000
-----+-----				R-squared =	0.4731
-----+-----				Adj R-squared =	0.4438
Total	6531.4359	38	171.879892	Root MSE =	9.7772

systol	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
fraction	-26.76722	7.217801	-3.71	0.001	-41.40559 -12.12884
weight	1.216857	.2336873	5.21	0.000	.7429168 1.690796
_cons	60.89592	14.28088	4.26	0.000	31.93295 89.85889

Forward selection. The predictor variable with the smallest p-value smaller than 0.10 is added from the existing model.

```
. stepwise, pe(0.1): regress systol fraction weight height chin forearm calf
pulse
```

```
begin with empty model
p = 0.0007 < 0.1000  adding weight
p = 0.0007 < 0.1000  adding fraction
```

Source	SS	df	MS	Number of obs =	39
Model	3090.07324	2	1545.03662	F(2, 36) =	16.16
Residual	3441.36266	36	95.5934072	Prob > F =	0.0000
-----+-----				R-squared =	0.4731

-----+-----					Adj R-squared =	0.4438
Total		6531.4359	38	171.879892	Root MSE	= 9.7772
-----+-----						
systol		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----						
weight		1.216857	.2336873	5.21	0.000	.7429168 1.690796
fraction		-26.76722	7.217801	-3.71	0.001	-41.40559 -12.12884
_cons		60.89592	14.28088	4.26	0.000	31.93295 89.85889
-----+-----						

Stepwise. The p-value for removal has to be larger than the p-value for adding.

```
. stepwise, pe(0.15) pr(0.1): regress systol fraction weight height chin
forearm calf pulse
pr(.1) <= pe(.15) invalid
r(198);
```

Backward stepwise. The p-value for adding has to be smaller than the p-value for removal. Though it does not happen in this example, predictor variables that were removed early could be added back in at a later stage.

```
. stepwise, pe(0.1) pr(0.15): regress systol fraction weight height chin
forearm calf pulse
begin with full model
p = 0.8637 >= 0.1500 removing calf
p = 0.6953 >= 0.1500 removing pulse
p = 0.6670 >= 0.1500 removing forearm
p = 0.2745 >= 0.1500 removing height
p = 0.1534 >= 0.1500 removing chin
```

-----+-----					Number of obs =	39
Model		3090.07324	2	1545.03662	F(2, 36) =	16.16
Residual		3441.36266	36	95.5934072	Prob > F	= 0.0000
-----+-----						
Total		6531.4359	38	171.879892	R-squared	= 0.4731
-----+-----						
					Adj R-squared	= 0.4438
					Root MSE	= 9.7772
-----+-----						
systol		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----						
fraction		-26.76722	7.217801	-3.71	0.001	-41.40559 -12.12884
weight		1.216857	.2336873	5.21	0.000	.7429168 1.690796
_cons		60.89592	14.28088	4.26	0.000	31.93295 89.85889
-----+-----						

Forward stepwise. The p-value for adding has to be smaller than the p-value for removal. Though it does not happen in this example, predictor variables that were added early could be removed again at a later stage.

```
. stepwise, pe(0.1) pr(0.15) forward: regress systol fraction weight height
chin forearm calf pulse
begin with empty model
p = 0.0007 < 0.1000 adding weight
p = 0.0007 < 0.1000 adding fraction
```

-----+-----					Number of obs =	39
Model		3090.07324	2	1545.03662	F(2, 36) =	16.16
Residual		3441.36266	36	95.5934072	Prob > F	= 0.0000
-----+-----						
Total		6531.4359	38	171.879892	R-squared	= 0.4731
-----+-----						
					Adj R-squared	= 0.4438
					Root MSE	= 9.7772
-----+-----						

systol	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
weight	1.216857	.2336873	5.21	0.000	.7429168	1.690796
fraction	-26.76722	7.217801	-3.71	0.001	-41.40559	-12.12884
_cons	60.89592	14.28088	4.26	0.000	31.93295	89.85889

The births dataset behaves the similarly, with stepwise methods not reincluding or rereoving any predictor variables.

use births

```
stepwise, pe(0.1) pr(0.15): regress bweight lowbw gestwks preterm matage hyp sex
```

```
begin with full model
p = 0.7607 >= 0.1500 removing preterm
p = 0.6119 >= 0.1500 removing matage
```

Source	SS	df	MS	Number of obs =	490
Model	131968728	4	32992182.1	F(4, 485) =	234.16
Residual	68333814.7	485	140894.463	Prob > F =	0.0000
				R-squared =	0.6588
				Adj R-squared =	0.6560
				Root MSE =	375.36
Total	200302543	489	409616.652		

bweight	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lowbw	-874.7961	66.39853	-13.17	0.000	-1005.26	-744.3318
gestwks	116.9463	9.385536	12.46	0.000	98.50497	135.3876
sex	-165.5769	34.09937	-4.86	0.000	-232.5776	-98.57619
hyp	-113.482	49.5704	-2.29	0.022	-210.8813	-16.08277
_cons	-1023.948	372.1152	-2.75	0.006	-1755.104	-292.7906

```
stepwise, pe(0.1) pr(0.15) forward: regress bweight lowbw gestwks preterm matage hyp sex
```

```
begin with empty model
p = 0.0000 < 0.1000 adding lowbw
p = 0.0000 < 0.1000 adding gestwks
p = 0.0000 < 0.1000 adding sex
p = 0.0225 < 0.1000 adding hyp
```

Source	SS	df	MS	Number of obs =	490
Model	131968728	4	32992182.1	F(4, 485) =	234.16
Residual	68333814.7	485	140894.463	Prob > F =	0.0000
				R-squared =	0.6588
				Adj R-squared =	0.6560
				Root MSE =	375.36
Total	200302543	489	409616.652		

bweight	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lowbw	-874.7961	66.39853	-13.17	0.000	-1005.26	-744.3318
gestwks	116.9463	9.385536	12.46	0.000	98.50497	135.3876
sex	-165.5769	34.09937	-4.86	0.000	-232.5776	-98.57619
hyp	-113.482	49.5704	-2.29	0.022	-210.8813	-16.08277
_cons	-1023.948	372.1152	-2.75	0.006	-1755.104	-292.7906

At this point, you will have to take our word for it that variables can be readded or rereoved from models since the datasets we used did not exhibit this feature.

In your homework, you will find that backward stepwise and forward stepwise methods can yield different models even with the same p-value criteria.