

# Linear Modelling in Stata

## Session 6: Further Topics in Linear Modelling

Stephen Pye

Arthritis Research UK Epidemiology Unit  
University of Manchester



01/11/2011

# This Week

- Categorical Variables
  - Comparing outcome between groups
  - Comparing slopes between groups (Interactions)
- Confounding
- Variable Selection
- Other considerations
  - Polynomial Regression
  - Transformation
  - Regression through the origin

# Categorical Variables

- None of the linear model assumptions mention the distribution of  $\mathbf{x}$ .
- Can use  $\mathbf{x}$ -variables with any distribution
- This enables us to compare different groups

## Dichotomous Variable

- Let  $\mathbf{x} = \mathbf{0}$  in group A and  $\mathbf{x} = \mathbf{1}$  in group B.
- Linear model equation is  $\hat{Y} = \beta_0 + \beta_1 \mathbf{x}$
- In group A,  $\mathbf{x} = \mathbf{0}$  so  $\hat{Y} = \beta_0$
- In group B,  $\mathbf{x} = \mathbf{1}$  so  $\hat{Y} = \beta_0 + \beta_1$
- Hence the coefficient of  $\mathbf{x}$  gives the mean difference between the two groups.

## Dichotomous Variable Example

- $x$  takes values 0 or 1
- $Y$  is normally distributed with variance 1, and mean 3 if  $x = 0$  and 4 if  $x = 1$ .
- We wish to test if there difference in the mean value of  $Y$  between the groups with  $x = 0$  and  $x = 1$

# Dichotomous Variable: Stata output

```
. regress Y x
```

Source	SS	df	MS	
Model	5.91402766	1	5.91402766	Number of obs = 40
Residual	38.5735226	38	1.0150927	F( 1, 38) = 5.83
Total	44.4875503	39	1.14070642	Prob > F = 0.0207
				R-squared = 0.1329
				Adj R-squared = 0.1101
				Root MSE = 1.0075

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.7690272	.3186052	2.41	0.021	.1240447	1.41401
_cons	3.149821	.2252879	13.98	0.000	2.69375	3.605892

## Dichotomous Variables and the T-Test

- Differences in mean between two groups usually tested for with t-test.
- Linear model results are *exactly* the same.
- Linear model assumptions are *exactly* the same.
  - Normal distribution in each group
  - Same variance in each group
- A t-test is a special case of a linear model.
- Linear model is far more versatile (can adjust for other variables).

## Categorical Variable with Several Categories

- What can we do if there are more than two categories ?
- Cannot use  $\mathbf{x} = \mathbf{0}, \mathbf{1}, \mathbf{2}, \dots$
- Instead we use “dummy” or “indicator” variables.
- If there are  $\mathbf{k}$  categories, we need  $\mathbf{k} - \mathbf{1}$  indicators.

## Three Groups: Example

Group	$x_1$	$x_2$	$\hat{Y}$	$\sigma^2$	
A	0	0	3	1	Baseline Group
B	1	0	5	1	
C	0	1	4	1	

- $\beta_0 = \hat{Y}$  in group A
- $\beta_1 =$  difference between  $\hat{Y}$  in group A and  $\hat{Y}$  in group B
- $\beta_2 =$  difference between  $\hat{Y}$  in group A and  $\hat{Y}$  in group C

# Three Groups: Stata Output

```
. regress Y x1 x2
```

Source	SS	df	MS	
Model	43.9178814	2	21.9589407	Number of obs = 60
Residual	55.4225097	57	.972324732	F( 2, 57) = 22.58
Total	99.3403911	59	1.68373544	Prob > F = 0.0000
				R-squared = 0.4421
				Adj R-squared = 0.4225
				Root MSE = .98607

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x1	2.071885	.3118212	6.64	0.000	1.447474 2.696296
x2	.7633442	.3118212	2.45	0.017	.138933 1.387755
_cons	2.95576	.2204909	13.41	0.000	2.514234 3.397285

## Comparing Groups

- In the previous example, groups B and C both compared to group A.
- Can we compare groups B and C as well ?
- In group B,  $\hat{Y} = \beta_0 + \beta_1$
- In group C,  $\hat{Y} = \beta_0 + \beta_2$
- Hence difference between groups is  $\beta_1 - \beta_2$
- Can use `lincom` to obtain this difference, and test its significance.

# The `lincom` Command

- `lincom` is short for linear combination.
- It can be used to calculate linear combinations of the parameters of a linear model.
- Linear combination =  $\mathbf{a}_j\beta_j + \mathbf{a}_k\beta_k + \dots$
- Can be used to find differences between groups  
(Difference between Group B and Group C =  $\beta_1 - \beta_2$ )
- Can be used to find mean values in groups  
(Mean value in group B =  $\beta_0 + \beta_1$ ).

# Stata Output from `lincom`

```
. lincom x1-x2
```

```
( 1)  x1 - x2 = 0.0
```

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	1.308541	.3118212	4.20	0.000	.6841298 1.932952

```
. lincom _cons + x1
```

```
( 1)  x1 + _cons = 0.0
```

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	5.027645	.2204909	22.80	0.000	4.586119 5.46917

## The `xi` Command

- Generating dummy variables can be tedious and error-prone
- Stata can do it for you with the command `xi`
- Add "`xi:` " to the start of the command
- Identify categorical variables by adding "`i.`" to the start of their name.
- `xi` generates new variables beginning with "`_I`".
- For example, suppose that the variable `group` contains the values "A", "B" and "C" for the three groups in the previous example.

# Stata Output from the `xi` Command

```
. xi: regress Y i.group
i.group          _Igroup_1-3          (_Igroup_1 for group==A omitted)
```

Source	SS	df	MS	Number of obs = 60		
Model	43.9178814	2	21.9589407	F( 2, 57)	=	22.58
Residual	55.4225097	57	.972324732	Prob > F	=	0.0000
				R-squared	=	0.4421
				Adj R-squared	=	0.4225
				Root MSE	=	.98607
Total	99.3403911	59	1.68373544			

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_Igroup_2	2.071885	.3118212	6.64	0.000	1.447474	2.696296
_Igroup_3	.7633442	.3118212	2.45	0.017	.138933	1.387755
_cons	2.95576	.2204909	13.41	0.000	2.514234	3.397285

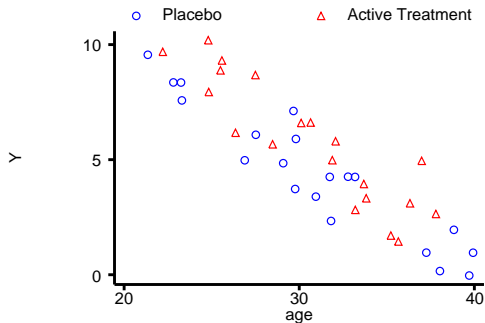
## Linear Models and ANOVA

- Differences in mean between more than two groups usually tested for with ANOVA.
- Linear model results are *exactly* the same.
- Linear model assumptions are *exactly* the same.
- ANOVA is a special case of a linear model.
- Linear model is far more versatile (can adjust for other variables).

## Mixing Categorical & Continuous Variables

- So far, we have only seen either continuous or categorical predictors in a linear model.
- No problem to mix both.
- E.g. Consider a clinical trial in which the outcome is strongly associated with age.
- To test the effect of treatment, need to include both age and treatment in linear model.
- Once upon a time, this was called Analysis of Covariance (ANCOVA)

# Example Clinical Trial: simulated data



# Stata Output Ignoring the Effect of Age

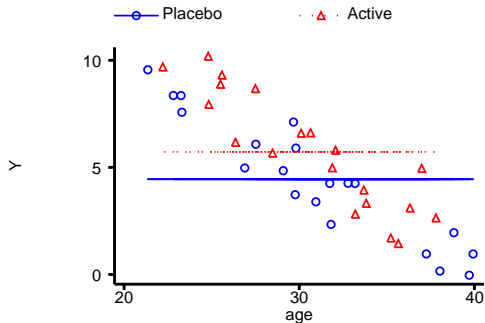
```
. regress Y treat
```

Source	SS	df	MS			
Model	16.2113323	1	16.2113323	Number of obs =	40	
Residual	297.306333	38	7.82385087	F( 1, 38) =	2.07	
Total	313.517665	39	8.03891449	Prob > F =	0.1582	
				R-squared =	0.0517	
				Adj R-squared =	0.0268	
				Root MSE =	2.7971	

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
treat	1.273237	.8845253	1.44	0.158	-.5173906	3.063865
_cons	4.448076	.6254539	7.11	0.000	3.181911	5.714241

# Observed and predicted values from linear model ignoring age



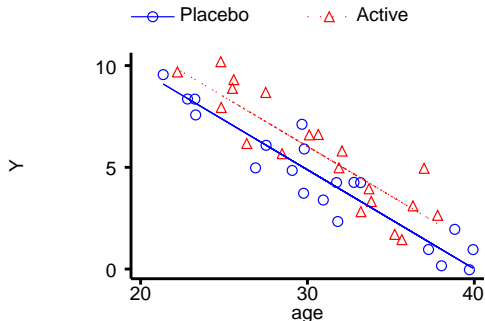
# Stata Output Including the Effect of Age

```
. regress Y age treat
```

Source	SS	df	MS	
Model	266.979443	2	133.489722	Number of obs = 40
Residual	46.5382219	37	1.25778978	F( 2, 37) = 106.13
Total	313.517665	39	8.03891449	Prob > F = 0.0000
				R-squared = 0.8516
				Adj R-squared = 0.8435
				Root MSE = 1.1215

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.487518	.034527	-14.12	0.000	-.5574763	-.4175597
treat	1.151167	.3547587	3.24	0.002	.4323578	1.869977
_cons	19.50491	1.095446	17.81	0.000	17.28533	21.7245

# Observed and predicted values from linear model including age



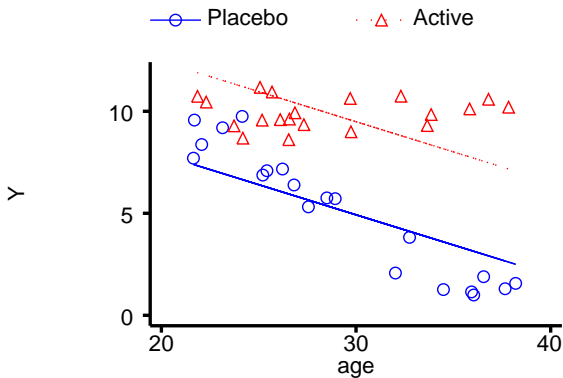
# Interactions

- In previous example, assumed that the effect of age was the same in treated and untreated groups.
- I.e. regression lines were parallel.
- This may not be the case.
- If the effect of one variable varies accord to the value of another variable, this is called “interaction” between the variables.

## Interaction Example

- Consider the clinical trial in the previous example
- Suppose treatment reverses the effect of aging, so that  $\hat{Y}$  is constant in the treated group.
- Thus the difference between the treated and untreated groups will increase with increasing age.
- Need to fit different intercepts and different slopes in the two groups.

# Interaction Data



# Regression Equations

- Need to fit the two equations

$$Y = \begin{cases} \beta_{00} + \beta_{10} \times \text{age} + \varepsilon & \text{if } \text{treat} = 0 \\ \beta_{01} + \beta_{11} \times \text{age} + \varepsilon & \text{if } \text{treat} = 1 \end{cases}$$

- These are equivalent to the equation

$$Y = \beta_{00} + (\beta_{01} - \beta_{00}) \times \text{treat} + \beta_{10} \times \text{age} + (\beta_{11} - \beta_{10}) \times \text{age} \times \text{treat} + \varepsilon.$$

- I.e. the output from stata can be interpreted as

**\_cons** The intercept in the untreated group  
**treat** The difference in intercept between the treated and untreated groups  
**age** The slope with age in the untreated group  
**treat\*age** The difference in slope between the treated and untreated groups

# Interactions: Stata Output

```
. xi:regress Y i.treat*age
i.treat      _Itreat_0-1      (naturally coded; _Itreat_0 omitted)
i.treat*age  _ItreXage_#      (coded as above)
```

Source	SS	df	MS			
Model	106.885564	3	35.6285212	Number of obs =	40	
Residual	25.4601434	36	.707226205	F( 3, 36) =	50.38	
				Prob > F =	0.0000	
				R-squared =	0.8076	
				Adj R-squared =	0.7916	
				Root MSE =	.84097	
Total	132.345707	39	3.39347967			

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_Itreat_1	-5.98171	1.52061	-3.93	0.000	-9.065651	-2.897769
age	-.2732318	.0337249	-8.10	0.000	-.3416291	-.2048345
_ItreXage_1	.2888497	.0518905	5.57	0.000	.183611	.3940885
_cons	15.45171	1.004261	15.39	0.000	13.41497	17.48844

# Interactions: Using `lincom`

- `lincom` can be used to calculate the slope in the treated group:

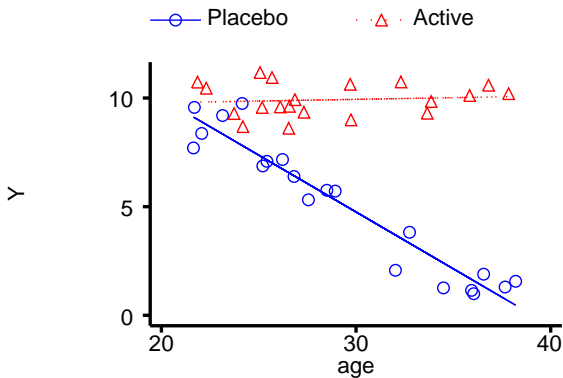
```
. lincom age+_ItrExAge_1
```

```
( 1)  age + _ItrExAge_1 = 0.0
```

	Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
(1)		.015618	.0394367	0.40	0.694	-.0643634	.0955993

- Can also be used to calculate intercept in treated group. However, this is not interesting since
  - We are unlikely to be interested in subjects of age 0
  - The youngest subjects in our sample were 20, so we are extrapolating a long way from the data.

# Interactions: Predictions from Linear Model



# The `testparm` Command

- Used to test a number of parameters simultaneously
- Syntax: `testparm varlist`
- Test  $\beta = \mathbf{0}$  for all variables in *varlist*
- Produces a  $\chi^2$  test on  $k$  degrees of freedom, where there are  $k$  variables in *varlist*.

# Confounding

- A linear model shows association.
- It does not show *causation*.
- Apparent association may be due to a third variable which we haven't measured.
- Confounding is about causality, and knowledge of the mechanisms are required to decide if a variable is a confounder.

# Confounding Example: Fuel Consumption

```
. regress mpg foreign
```

Source	SS	df	MS	
Model	378.153515	1	378.153515	Number of obs = 74
Residual	2065.30594	72	28.6848048	F( 1, 72) = 13.18
Total	2443.45946	73	33.4720474	Prob > F = 0.0005

				R-squared = 0.1548
				Adj R-squared = 0.1430
				Root MSE = 5.3558

mpg	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
foreign	4.945804	1.362162	3.631	0.001	2.230384	7.661225
_cons	19.82692	.7427186	26.695	0.000	18.34634	21.30751

```
. regress mpg foreign weight
```

Source	SS	df	MS	
Model	1619.2877	2	809.643849	Number of obs = 74
Residual	824.171761	71	11.608053	F( 2, 71) = 69.75
Total	2443.45946	73	33.4720474	Prob > F = 0.0000

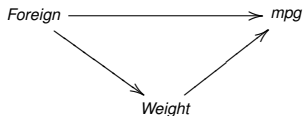
				R-squared = 0.6627
				Adj R-squared = 0.6532
				Root MSE = 3.4071

mpg	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
foreign	-1.650029	1.075994	-1.533	0.130	-3.7955	.4954421
weight	-.0065879	.0006371	-10.340	0.000	-.0078583	-.0053175
_cons	41.6797	2.165547	19.247	0.000	37.36172	45.99768

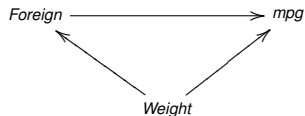
# What is Confounding ?

- What you see is not what you get
- $\hat{Y} = \beta_0 + \beta_1 x$
- Two groups differing in  $x$  by  $\Delta x$  will differ in  $Y$  by  $\beta_1 \Delta x$
- If we change  $x$  by  $\Delta x$ , what happens to  $\hat{Y}$  ?
- If it changes by  $\beta_1 \Delta x$ , no confounding
- If it changes by anything else, there is confounding

# Path Variables vs. Confounders



*Weight is a path variable*



*Weight is a confounder*

## Identifying a Confounder

- Is a cause of the outcome irrespective of other predictors
- Is associated with the predictor
- Is not a consequence of the predictor

## Allowing for Confounding

- In theory, adding a confounder to a regression model is sufficient to adjust for confounding.
- Then parameters for other variables measure the effects of those variables when confounder does not change.
- This assumes
  - Confounder measured perfectly
  - Linear association between confounder and outcome
- If either of the above are not true, there will be *residual confounding*

# Variable Selection

- May wish to reduce the number of predictors used in a linear model.
  - Efficiency
  - Clearer understanding
- Several suggested methods
  - Forward selection
  - Backward Elimination
  - Stepwise
  - All subsets

## Forward Selection

- Choose a significance level  $p_e$  at which variables will enter the model.
- Fit each predictor in turn.
- Choose the most significant predictor.
- If its significance level is less than  $p_e$ , it is selected.
- Now add each remaining variable to this model in turn, and test the most significant.
- Continue until no further variables are added.

## Backward Elimination

- Starts with all predictors in model.
- Removes the least significant.
- Repeat until all remaining predictors significant at chosen level  $p_r$ .
- Has the advantage that all parameters are adjusted for the effect of all other variables from the start.
- Can give unusual results if there are a large number of correlated variables.

## Stepwise Selection

- Combination of preceding methods.
- Variables are added one at a time.
- Each time a variable is added, all the other variables are tested to see if they should be removed.
- Must have  $p_r > p_e$ , or a variable could be entered and removed on the same step.

## All Subsets

- Can try every possible subset of variables.
- Can be hard work: 10 predictors = 1023 subsets.
- Need a criterion to choose best model.
- Adjusted  $R^2$  is possible, there are others.
- Not implemented in stata.

# Problems with Variable Selection

- Significance Levels
  - Hypotheses tested are not independent.
  - Variables chosen for testing not randomly selected.
  - Hence significance levels not equal to nominal levels.
  - Less of a problem in large samples.
- Differences in Models Selected
  - Models chosen by different methods may differ.
  - Choice of significance level will affect models.
  - Need common sense.

## Variable Selection in Stata

- Command `sw regress` is used for forwards, backwards and stepwise selection.
- Option `pe` is used to set significance level for inclusion
- Option `pr` is used to set significance level for exclusion
- Set `pe` for forwards, `pr` for backwards and both for stepwise regression.

# Variable Selection in Stata: Example

```
. sw regress weight price hdroom trunk length turn displ gratio, pe(0.05)
```

```
p = 0.0000 < 0.0500 adding length
p = 0.0000 < 0.0500 adding displ
p = 0.0015 < 0.0500 adding price
p = 0.0288 < 0.0500 adding turn
```

Source	SS	df	MS	Number of obs	-	74
Model	41648450.8	4	10412112.7	F( 4, 69)	-	293.75
Residual	2445727.56	69	35445.3269	Prob > F	-	0.0000
				R-squared	-	0.9445
				Adj R-squared	-	0.9413
				Root MSE	-	188.27
Total	44094178.4	73	604029.841			

weight	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
length	19.38601	2.328203	8.327	0.000	14.74137 24.03064
displ	2.257083	.467792	4.825	0.000	1.323863 3.190302
price	.0332386	.0087921	3.781	0.000	.0156989 .0507783
turn	23.17863	10.38128	2.233	0.029	2.468546 43.88872
_cons	-2193.042	298.0756	-7.357	0.000	-2787.687 -1598.398

```
. sw regress weight price hdroom trunk length turn displ gratio, pr(0.05)
```

```
p = 0.6348 >= 0.0500 removing hdroom
p = 0.5218 >= 0.0500 removing trunk
p = 0.1371 >= 0.0500 removing gratio
```

Source	SS	df	MS	Number of obs	-	74
Model	41648450.8	4	10412112.7	F( 4, 69)	-	293.75
Residual	2445727.56	69	35445.3269	Prob > F	-	0.0000
				R-squared	-	0.9445
				Adj R-squared	-	0.9413
				Root MSE	-	188.27
Total	44094178.4	73	604029.841			

weight	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
price	.0332386	.0087921	3.781	0.000	.0156989 .0507783
turn	23.17863	10.38128	2.233	0.029	2.468546 43.88872
displ	2.257083	.467792	4.825	0.000	1.323863 3.190302
length	19.38601	2.328203	8.327	0.000	14.74137 24.03064

# Polynomial Regression

- If association between  $x$  and  $Y$  is non-linear, can fit polynomial terms in  $x$ .
- Keep adding terms until the highest order term is not significant.
- Parameters are meaningless: only entire function has meaning.

# Transformations

- If  $\mathbf{Y}$  is not normal or has non-constant variance, it may be possible to fit a linear model to a transformation of  $\mathbf{Y}$ .
- Interpretation becomes more difficult after transformation.
- Log transformation has a simple interpretation.
  - $\log(\mathbf{Y}) = \beta_0 + \beta_1 \mathbf{x}$
  - when  $\mathbf{x}$  increases by 1,  $\log(\mathbf{Y})$  increases by  $\beta_1$ ,
  - $\mathbf{Y}$  is multiplied by  $e^{\beta_1}$
- Transforming  $\mathbf{x}$  is not normally necessary unless the problem suggests it.

## Regression through the origin

- You may know that if  $\mathbf{x} = \mathbf{0}$ ,  $\mathbf{y} = \mathbf{0}$ .
- Stata can force the regression line through the origin with the option `nocons`.
- However
  - $R^2$  is calculated differently and cannot be compared to conventional  $R^2$ .
  - If we have no data near the origin, should not force line through the origin.
  - May obtain a better fit with a non-zero intercept if there is measurement error.