

Mathematics 231

Lecture 32
Liam O'Brien

1

Announcements

- Today
 - Multiple Regression
 - Effect Modification

2

Modifying Variables

- A **modifying variable** changes the effect of a predictor on the outcome.
- Modifying variables are included in the model by adding an **interaction term**.
- These are generated by multiplying two covariates (predictors) together.
- When an interaction term is present, you generally want to include its **main effects** regardless of the p-values associated with them.

3

Modifying Variables

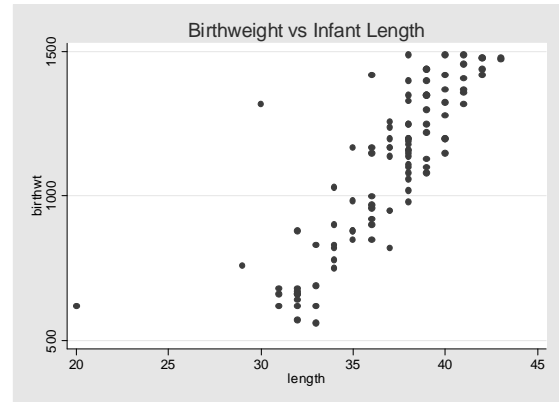
- For example, if you have two predictors, x_1 and x_2 , the interaction term would be $x_1 * x_2$.
- If $x_1 * x_2$ is a statistically significant predictor, then x_1 and x_2 must also remain individually in the model.
- The effect of x_1 on the outcome is dependent upon the value of x_2 .
- Interactions can include more than two main effects, but their interpretations become difficult.

4

Example: Birthweights

- Let's consider the birthweight data.
- We already know that toxemia, by itself, is not a significant predictor, but that with gestational age it is.
- Now consider infant length as a predictor of birthweight.

5



6

Example: Birthweights

```
. gen lentox=length*toxemia
. regress birthwt length toxemia lentox
```

Source	SS	df	MS	Number of obs = 100		
Model	4997922.76	3	1665974.25	F(3, 96)	=	72.08
Residual	2218819.99	96	23112.7083	Prob > F	=	0.0000
Total	7216742.75	99	72896.3914	R-squared	=	0.6925
				Adj R-squared	=	0.6829
				Root MSE	=	152.03

birthwt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
length	57.47765	4.690349	12.25	0.000	48.16738	66.78792
toxemia	-1192.974	443.1596	-2.69	0.008	-2072.639	-313.3094
lentox	30.50423	11.7999	2.59	0.011	7.081609	53.92686
_cons	-1007.631	172.6113	-5.84	0.000	-1350.262	-665.0001

7

Example: Birthweights

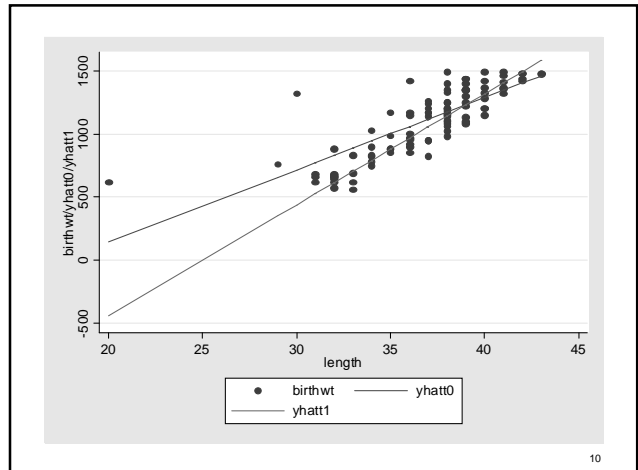
- The effect of infant length on birthweight depends on whether toxemia is present or not.
- $E(Y) = -1008 + 57.5(\text{length}) - 1193(\text{toxemia}) + 30.5(\text{length} * \text{toxemia})$
- For those with toxemia:
 $E(Y) = -2201 + 88(\text{length})$
- For those without toxemia:
 $E(Y) = -1008 + 57.5(\text{length})$

8

Plotting Both Regression Lines

- Knowing the two regression equations (from setting toxemia to 0 and 1), use Stata to generate two sets of predicted values.
- `. gen yhatt0=-1008+57.5*length`
- `. gen yhatt1=-2201+88*length`
- Then use the "Overlaid twoway graphs" to place these on the same plot as the observed data.

9



The Effect of a Modifying Variable

- The interaction term has changed the slope.
- The interaction term is of primary interest when its coefficient is significantly different from zero.
- The terms that make up the interaction must be singly included in the model (the main effects).
- Often, these main effects may not individually have coefficients that are significantly different from zero.
- They also often do not have meaningful interpretations if an interaction term is present.

11

Collinearity

- Collinearity is always present to some degree when you include an interaction term.
- **Colliearity** occurs when two or more explanatory variables are correlated to the extent that they convey essentially the same information about the variation in the response.
- One symptom of colliearity is the instability of the estimated coefficients and their standard errors (i.e., the standard errors become large).

12

Okay, So I'm Confused

- This is the point... there is no set-in-stone method for finding the most appropriate model.
- The goal is to find a model that balances prediction (e.g., a large R^2) with parsimony (not too many predictors).
- There are several semi-standard methods for doing this, and may result in different final models.

13

Choosing Predictors

- Ideally, we should have some prior knowledge as to which variables might be relevant.
- To study fully all of the predictors, it would be necessary to run a separate regression analysis for each possible combination of variables.
- While such a procedure would be thorough, it would be terribly time consuming.
- More frequently we use a stepwise approach to choose the best-fitting model.

14

Stepwise Regression

- Most statistical packages have automated routines built in that will perform a systematic method for obtaining the best model.
- The simplest of these are the **forward selection** and **backward selection** methods.
- The automated routines are not recommended for use by any self-respecting statistician, but they can be done "manually." We won't do them here.
- The best model results from careful statistical thought, combined with subject matter considerations.

15

Forward Selection

- **Forward selection** begins with nothing in the model and introduces one variable at a time.
- The variable that has the smallest p-value for its coefficient is added (or largest test statistic).
- This variable is kept in the model, then from the remaining explanatory variables, we repeat the procedure.
- We do this until no variables, when introduced to the model, have an adequately small p-value (or cause the estimates of the other coefficients to become unstable).

16

Backward Selection

- **Backward selection** begins with all the variables in the model and drops the one with the largest p-value (smallest test statistic).
- From the remaining variables the one with the next largest p-value is dropped.
- We repeat this procedure until all of the remaining variables have small p-values.
- Note that it is entirely possible that the final models will differ depending on whether forward, or backward, selection is used.

17

Stepwise Regression

- More complicated procedures exist in which variables may be added after they have been dropped (or dropped after they've been added), according to certain rules.
- These procedures are called **stepwise regression** procedures, and are generally preferred.
- However, they are difficult to do without using an automated computer routine.
- What about interaction terms?

18