How do we choose our model?

How do you decide which independent variables?

If you want to read more about this, try Studenmund, A.H. *Using Econometrics* Chapter 7. (either 3rd or 4th Editions)

1. Theory

Is the variable's place in the equation certain and theoretically sound? Most important!

2. T-test

Is the variable's estimated coefficient significant in the expected direction (one-sided test)?

3. $\overline{R}^2$

Does the overall fit of the equation improve when then variable is added to the equation?

4. Bias

Do other variables' coefficients change significantly when the variable is added to the equation?

If all of these are true, then the variable belongs in the equation.

Dummy Variables

1. Intercept Dummy Variables

   - A dummy variable that changes the constant or intercept term
   - $Y_i = \beta_0 + \beta_1 X + \beta_2 D_i + \varepsilon$

2. Seasonal Dummies (Multi-alternative Dummies)
   - Dummy variables used to represent qualitative variables that take on more than two alternatives

   Use one less dummy variable than there are alternatives. Each dummy will represent one condition.

3. Slope Dummies
    - A dummy variable that changes the slope of the relationship between x and y

4. Dummy Dependent Variables
    - Dummy variable is used as the dependent variable

Example #1: More than 2 categories (more than one dummy variable)

Education can be thought of as:
(1) not having earned a high school diploma,
(2) having earned only a high school diploma, and
(3) having more education than a high school diploma.

Then we use two dummies. We'll call them D_1 and D_2.

D_1=1 if you have only a high school diploma
0 otherwise

D_2=1 if you have more education than a high school diploma
0 otherwise.

What are all the possibilities?

you have more than a high school degree	D_1= _____ and D_2= ____
you have a high school diploma and nothing beyond that	D_1= ____ and D_2= ____
you have not earned any diploma	D_1= ____ and D_2= ____

CAUTION: Don't include too many dummies or you'll have to explain each data point!

CAUTION: Don't include a dummy that only takes a value of 1 for one data point and zero for all other observations. This 'one-time' dummy acts to eliminate that observation from the data set, improving the fit artificially.

Some ideas for using dummy variables:

- Could use dummy for seasonal changes if you have data where each case is at a different time point.

Illustration #1: If the data has been recorded quarterly, you will need 3 dummy variables.

D_1 = { 1 in Quarter 1
0 otherwise

D_2 = { 1 in Quarter 2
0 otherwise

D_3 = { 1 in Quarter 3
0 otherwise

	D_1	D_2	D_3
Quarter 1			
Quarter 2			
Quarter 3			
Quarter 4			

Illustration #2: Dummy for Time Series Data
If you were interested to study the impact of a particular event on a given variable, a dummy variable could be used for this. For example, the impact of Sept 11 on airline travel could be modeled with a dummy variable.

D = 0 for 1980 -2001
D = 1 for 2002 to present

The sign of the coefficient of D will give the direction of any shift in 2001.

Interaction Terms

⇒ Interaction terms are products of two or more independent variables.

⇒ Allow for differences in effect of an explanatory factor across categories or levels of another factor

⇒ Used when the change in Y with respect to one independent variable depends on the level of another independent variable.

⇒ Can interact with dummies or continuous variables.

$\Rightarrow$ For a continuous(independent) and a dummy:

- Allows the slope between the dependent and independent variable to be different depending on whether the condition specified by the dummy is met.

- Used whenever the impact of an independent variable on the dependent variable is hypothesized to change if some qualitative condition is met.

1. What does the regression equation look like with an interaction in it?

This is called a slope dummy variable

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + \varepsilon_I$$

This one has an interaction between an independent variable, X_i , and a dummy variable, D_i.

What effect does the interaction have on the slope (that is, the change in Y brought about by a change in X)?

When D=0, $\Delta Y/\Delta X=\beta_1$ This is the slope for the reference group

When D=1, $\Delta Y/\Delta X=\beta_1+\beta_3$ This is the slope for the indicated group

The slope (or the coefficient of X) changes when the condition specified by D is met.

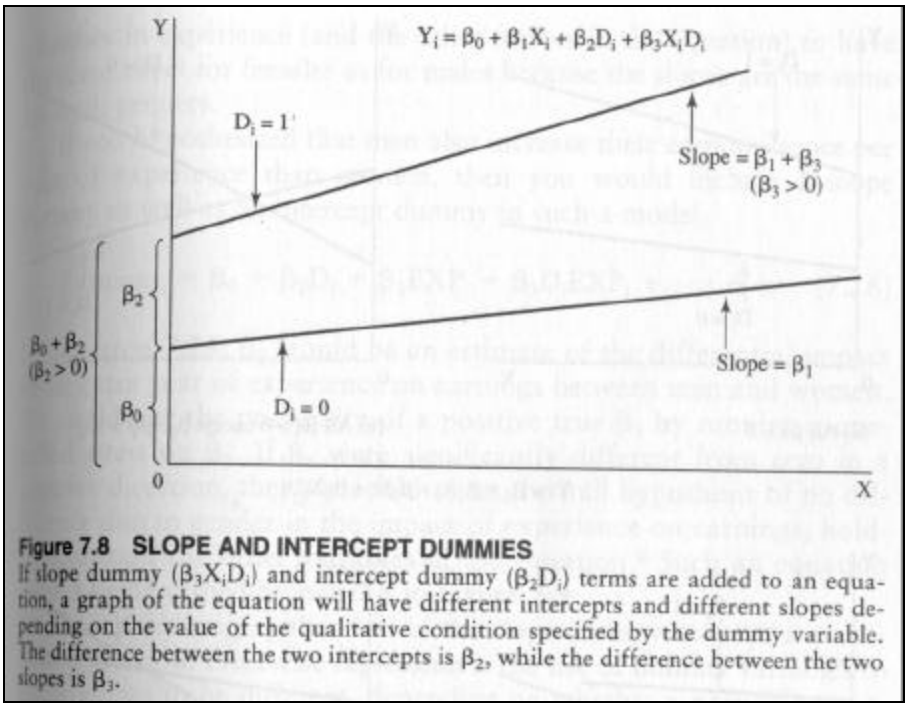

Figure 7.8 SLOPE AND INTERCEPT DUMMIES
If slope dummy ($\beta_3 X_i D_i$) and intercept dummy ($\beta_2 D_i$) terms are added to an equation, a graph of the equation will have different intercepts and different slopes depending on the value of the qualitative condition specified by the dummy variable. The difference between the two intercepts is β_2, while the difference between the two slopes is β_3.

from AH Studenmund. 1997 *Using Econometrics: A Practical Guide* p. 237

2. You need both the slope dummy and the intercept dummy in the equation.

The above regression line has both a **slope dummy** and an **intercept dummy** (a dummy that does not get multiplied by anything). This is necessary in most cases since just including a slope dummy would bias the slope by forcing it to explain more than it should, for example, changes in the mean between two groups. An intercept dummy best explains this sort of change. So, the model should include an intercept dummy (plain dummy term) where there is a slope dummy (a dummy multiplied by a predictor).

Think carefully about your hypotheses about the direction of the relationship between the dummies and the outcomes since these terms make the model very flexible.

3. How do you test for significance of interaction terms?

⇒ To test for differences in slopes between the categories, use the t-test on the interaction term.

⇒ To test overall differences in the regression relationship with and without the inclusion of an interaction use an F-test (#2).

Example #2: Dummy Variable Interacting with a Continuous (Independent) Variable

Does extensive media coverage of a military crisis influence public opinion on how to respond to the crisis?

Political scientists at UCLA came up with a model concerning the 1990 Persian Gulf War, precipitated by Iraq leader Saddam Hussein's invasion of Kuwait. They developed a model to analyze the level of support Americans had for military (rather than diplomatic) response to the crisis. The dependent variable ranges from 0 (preference for a diplomatic response) to 4 (preference for military response.

Here is the model they developed based on data from 1,763 U.S. Citizens.

$$E(y)=\beta_0 + \beta_1x_1 + \beta_2x_2 +\beta_3x_3 +\beta_4x_4 +\beta_5x_5 +\beta_6x_6 +\beta_7x_7 +\beta_8x_2x_3 +\beta_9x_2x_4$$

where:

x_1 = Level of TV news exposure in a selected week (number of days)
x_2 = Knowledge of seven political figures (1 point for each correct answer)
x_3 = Dummy variable for Gender (1 if male, 0 if female)
x_4 = Dummy variable for Race (1 if nonwhite, 0 if white)
x_5 = Partisanship (0-6 scale, where 0 = strong Democrat and 6 = strong Republican)
x_6 = Defense spending attitude (1-7 scale, where 1 = greatly decrease spending and 7 = greatly increased spending)
x_7 = Education

The regression results:

Variable	β estimate	Standard Error	Two-Tailed p-Value
TV news exposure (x_1)	.02	.01	.03
Political knowledge (x_2)	.07	.03	.03
Gender (x_3)	.67	.11	<.001
Race (x_4)	-.76	.13	<.001
Partisanship (x_5)	.07	.01	<.001
Defense spending (x_6)	.20	.02	<.001
Education (x_7)	.07	.02	<.001
Knowledge X Gender (x_2x_3)	-.09	.04	.02
Knowledge X Race (x_2x_4)	.10	.06	.08
R^2=.194, F=46.88 (p<.001)			

Source: Iyengar, S. and Simon, A. 1993. News coverage of the Gulf Crisis and public opinion. *Communication Research* 20,: 380 (Table 2)

1) Interpret the β estimates for TV news exposure.

2) Is there enough support to say that TV news exposure is associated with support for a military resolution of the crisis?

3) Is there sufficient evidence to say that the relationship between support for a military resolution and gender depends on political knowledge?

4) What is the effect of knowledge on support for military resolution for men?

$E(y)=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(1) + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_2(1) + \beta_9 x_2 x_4$

5) What is the effect of knowledge on support for military resolution for women?

$E(y)=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(0) + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_2(0) + \beta_9 x_2 x_4$

6) What test would you use to answer the following question: Overall, does gender affect support for a military resolution?

Example #3: Interacting 2 continuous variables

Fowles and Loeb hypothesized that drunk driving fatalities are more likely at high altitude because higher elevations diminish the oxygen intake of the brain, which increases the impact of a given amount of alcohol.

F_i= Traffic fatalities per vehicle mile (by state)
B_i= per capita consumption of beer
S_i= average highway driving speed
D_i= dummy (1=state has a vehicle inspection program, 0=no inspection program)
A_i= average altitude of metro areas (1000s of feet)

$$\hat{F}_i = -2.33 - 0.024B_i + 0.17S_i - 0.24D_i - 0.35A_i + 0.023B_i \cdot A$$

(t-statistics)	(-0.8)	(1.53)	(-0.96)	(-1.07)	(1.97)
Hypothesized relationship	+	+	-	+	+?

n=48
adjusted R^2=.499

The interaction in this model is between two continuous variables, consumption rate of beer and altitude. The effect on the outcome of each the two variables involved in the interaction depends on the interaction coefficient and the coefficient on the original variable.

1) Does the average altitude of metropolitan areas in the state affect the relationship between per capita beer consumption and the rate of traffic fatalities?

2) How much higher a fatality rate do we expect for an average altitude of metro area increase by 1000 feet?

3) Does the altitude affect the overall regression relationship explaining fatality rate? (How would you approach this?)