

NOTICE. Unless otherwise indicated, all materials on this page and linked pages at the blue.temple.edu, astro.temple.edu, and rbtaylor.net addresses are the sole property of Ralph B. Taylor and © 1999-2001 by Ralph B. Taylor. All users have the right to freely access and copy these pages provided that they: acknowledge the source, do not make changes on any pages, and do not charge more than copying costs for distribution.

More Topics on Multiple Regression:

Assessing Change;  
Different Predictor Sets;  
Dummy Variables;  
Interaction Terms;  
Multicollinearity;  
Curvilinearity

Learning objectives

- Understand, conceptually, what happens when you add outcome scores from an earlier time as a predictor
- How to use dummy variables to represent categorical and continuous variables
- How to construct interaction terms in regression, and how to interpret
- How to test for significant explained variance of additional predictors
- How to decide if multicollinearity exists
- A curvilinear relationship?

CONTROLLING FOR EARLIER LEVELS OF AN OUTCOME

Usually, correlation does not imply causality

In general, with multiple regression it is difficult to conclude that our predictors "caused" scores on our outcome. Multiple regression is built on correlations, and we know that "correlation does not necessarily imply causation." We find a correlation between ice cream sales and monthly temperature, so we try to predict temperature with ice cream sales. If we can do so we do not conclude that ice cream sales therefore cause a rise in temperature.

The above example may seem silly, but only because the laws of nature are well known. We know what does cause temperature variations, and ice cream sales is not one of those factors.

In the realm of the social sciences, however, matters are less clearcut. You are examining the relationship between the violent crime rate and the divorce rate in states, for example. You use the divorce rate to predict the violent crime rate. But you also could use the violent crime rate to predict the divorce rate, arguing that high crime levels in general elevate stress levels in the population (or increase distrust in the population; or make people less committed to certain standards of public behavior) and therefore result in higher divorce rates.

In short, in many instances in the social sciences, other than the

theory we are using, we have little reason for firmly believing that one variable is an outcome, and another a predictor.<sup>1</sup>

#### Time ordering of variables

But there are situations where there are exceptions, where we can make some limited inferences about causality based on the correlations.

**If a variable measures a behavior or attitude that occurred earlier than the behaviors or attitudes measured by a second variable, then we might think that the first variable could cause the second.**

For example, say you are studying neighborhood crime rates. You have reported crime rates for 1980 and 1990. It makes no sense to treat the 1990 crime rates as predictors of the 1980 crime rates. But it does make sense to use the 1980 crime rates as predictors of the 1990 crime rates.

So when you have predictors measuring attributes that occurred before the attributes measured by the outcome, you can make some limited inferences about causality.

#### Spurious correlation

Of course this situation does not rule out spurious correlation.<sup>2</sup> There might be a third variable (Z) that is causing both the predictor (X) and the outcome (Y). For example a high poverty rate in neighborhoods in 1980 (Z) might cause both a high crime rate in 1980 (X) and a high crime rate in 1990 (Y).

To deal with the possibility of spurious correlation: if there is an additional variable that you think might influence both your predictor and your outcome, include it as a predictor in your multiple regression analysis.

#### Controlling for earlier levels of the outcome

You can make even stronger inferences about causality if in your regression model you include as a predictor a measure of your outcome at an earlier point in time.

Hamilton provides such an example (p. 68). His outcome is postshortage (1981) water usage. He includes as a predictor 1980 (preshortage) water usage. He also includes income as a predictor. The inclusion of preshortage water usage changes the conceptual meaning of the outcome variable.

In order to understand what is happening, we need to remind ourselves about the logic of partialling.

Suppose that I regressed postshortage (1981) water usage on preshortage (1980) usage. I have now decomposed postshortage usage into two parts. First there is the familiar predicted portion:

$$\text{Post-shortage-predicted} = A + B * \text{Pre-shortage}$$

This is the portion predictable from pre-shortage usage, which will correlate

---

1 Of course there are exceptions when we have variables that clearly cannot cause another variable. For example age cannot be a result of a person's level of fear of crime.

2 For a detailed discussion of spurious correlation see Blalock Social Statistics pp. 469-475. Hamilton does not appear to cover this.

1.0 with scores on pre-shortage usage.

Then we have the residual portion:

$$\text{Post-shortage}_{\text{residual}} = \text{Post-shortage}_{\text{actual}} - \text{Post-shortage}_{\text{predicted}}$$

This is post-shortage usage after removing pre-shortage levels. Of course, it will correlate 0 with pre-shortage usage levels.

More specifically, this residualized variable is both:

- \* the portion of post-shortage usage for each household not predictable from that household's pre-shortage usage; and
- \* the portion of post-shortage usage for each household not predictable given, overall, how all the households changed their usage from the pre- to the post-shortage period.

In other words it represents **unexpected change**: that portion of the outcome not predictable from earlier levels, and not predictable from the overall changes occurring during the period.<sup>3</sup>

Hamilton uses the residuals in the partial leverage plot on p. 70. The Y variable ( $e_{Y|X2}$ ) is post-usage after controlling for (partialling out) pre-shortage usage (X2). Scores above 0 on the vertical axis used more water during the post-shortage period than they "should" have given their earlier usage levels, and given overall changes in all households between pre- and post-shortage times. They showed unexpected increases in the sense that their residual is greater than 0, not simply in the sense that their usage went up. This is a residual not a difference score.

**When you include an earlier level of your outcome as a predictor in your regression, the portion of the outcome remaining for the other predictors to predict is unexpected change on the outcome; unexpected change occurring between the period bounded by your predictor and your outcome.**

Say I have the following regression:

$$\begin{aligned} \text{1990 Neighborhood Robbery Rate} = & \text{Constant} + \\ & \text{1980 Neighborhood Robbery Rate} + \\ & \text{1980 Poverty} \end{aligned}$$

If I find a significant coefficient for poverty I can conclude that the 1980 poverty rate led to an unexpected increase in neighborhood robbery rates. (This is EQUIVALENT to saying that 1980 poverty was connected positively to 1990 robbery rates while controlling for 1980 robbery rates.) Assuming I have ruled out spurious correlation, this comes pretty close to a causal statement. This is about as close to causality as we can get with regression.

---

3 Bohrstedt, G (1969). "Observation on the measurement of change." In E. F. Borgatta and G. Bohrstedt (eds.) Sociological methodology 1969 113-133. San Francisco: Jossey Bass.

Note that unexpected change scores are different from using difference scores to measure change (e.g., later level - earlier level). Such scores are suboptimal for a number of reasons (see Bohrnstedt 1979). For example, they do not take into account how all the cases might have changed between an earlier time and a later time.

#### DUMMY VARIABLES

We have been over dummy variables before. But I want to pull together what we already know about dummy variables, and introduce the idea of multiple dummy variables.

##### Single dummy, categorical basis

Dummy variables represent categorical variables and a nominal level of measurement. Cases that have a certain characteristic are coded 1 (repeat offenders; southern states; gentrifying neighborhoods; women); cases that do not have these characteristics are coded 0 (first-time offenders; non-southern states; non-gentrifying neighborhoods; men). Darlington (pp. 66-70) calls these indicator variables.

Hamilton (pp. 85-92) distinguishes between intercept dummy variables and slope dummy variables. His slope dummy variables are really a product of multiplying a dummy variable times another measurement variable. In this handout when I refer to dummy variables I refer only to intercept dummy variables.

The unstandardized slope or the b weight for dummy variables reflects the difference between the two groups on the Y variable.

For example, suppose I regress the state-level rate of imprisonment per 1,000 population against a dummy variable for southern location. The results are as follows:

---

Listwise Deletion of Missing Data

Equation Number 1    Dependent Variable..    PRISRA85    RATE PRISONERS / 100K POP

Block Number 1.    Method:    Enter    SOUTH

Variable(s) Entered on Step Number 1..    SOUTH    SOUTH CENSUS REGION: DUMMY

Multiple R	.50107	Analysis of Variance			
R Square	.25107		DF	Sum of Squares	Mean Square
Adjusted R Square	.23547	Regression	1	71124.71529	71124.71529
Standard Error	66.48250	Residual	48	212156.26471	4419.92218
		F =	16.09185	Signif F =	.0002

----- Variables in the Equation -----

Variable	B	SE B	Beta	T	Sig T
SOUTH	80.852941	20.155467	.501074	4.011	.0002
(Constant)	148.147059	11.401654		12.993	.0000

End Block Number 1 All requested variables entered.

---

The b weight tells us that in southern as compared to non-southern states, the rate of imprisonment is 80.8 prisoners per 100,000 population higher. You can interpret the t-test in the usual way.

The beta weight for a dummy variable, however, is not especially informative. The standard error for a beta weight depends in part on the standard deviation of the predictor variable. The standard deviation of a dummy variable may be driven in part by researcher choice. I explain below.

### Single Dummy, Interval basis

#### Creating

You can create dummy variables where the basis is numerical rather than categorical. Some examples: the upper quartile of a distribution; all cases below the lower hinge of a distribution; the top 10% of arrestees on previous arrests, and so on. The final dummy variable you create is the same as a dummy variable based on a categorical variable, except that the researcher has explicitly chosen how many cases will have the value 1 on the dummy, and how many will have the value 0. Therefore she has "chosen" the standard deviation for the dummy variable. And, since we know that the formula for beta weights uses standard deviations on X, this will influence the beta weight.

For example, I might be interested in isolating the states that are in the top 25% on ALLBOOZ and SPIRITS. I sort the file on each variable, and note the value for the highest 25%. I know from the stem and leaf plot that the 75% percentile (upper hinge) is 42.710 for ALLBOOZ.

I create a dummy variable for ALLBOOZ as follows.

---

```
COMPUTE BOOZDUM=0 .  
IF (ALLBOOZ>=42.710) BOOZDUM=1 .  
EXECUTE .  
FREQUENCIES=BOOZDUM .  
EXECUTE .
```

---

[Note use of >= "greater than or equal to"]

This sets the value of the variable to 1 for all cases where the average amount of alcohol purchased equals or exceeds 42.710 gallons per year.

If you wanted to focus on SPIRITS you could proceed as follows, after doing identifying the 75th percentile.

---

```
COMPUTE SP75DUM=0 .
IF SPIRITS>=2.90 THEN LET SP75DUM=1 .
EXECUTE .
FREQUENCIES VARIABLES=SP75DUM .
EXECUTE .
```

---

### Cutpoints

Where on your continuous variable you choose to draw the line between 0 and 1 with your dummy is up to you. It may reflect half, or scores above average, or the top 25 percent. It's up to you. Just be sure you have a clear theoretical rationale.

### Multiple dummies

You can include more than one dummy variable in a regression. When you do so, however, your interpretation of the b weight changes dramatically for the categorical variables. IF YOU INCLUDE TWO (OR MORE) DUMMY VARIABLES IN A MULTIPLE REGRESSION, THE B WEIGHT FOR EACH DUMMY CONTRASTS THE CASES SCORING 1 ON THE DUMMY WITH CASES SCORING 0 ON ALL THE DUMMIES. Cases scoring 0 on all the dummies are in the base category or the reference category. Hamilton calls this the omitted category. (Darlington, pp. 226-227, calls it the base cell. Cohen and Cohen call it the reference string.)

Say for example I wanted to look at the effects of the different regions of the country. I could create a regional dummy for each region. But I would have to leave ONE REGION with no dummy variable.

For example in the census data file - USDANU12 -- I have four regions

REGION	REGION\$
1	Northeast
2	Midwest
3	South
4	West

I have already created my dummy for SOUTH. Suppose I want to create a dummy for the northeast (NEDUM) and a dummy for the west (WESTDUM).

What statements would I use for this?

When I am done here is how the different regions score on the original variable and on these dummy variables.

REGION	REGION\$	NEAST	SOUTH	WEST
1	Northeast	1	0	0
2	Midwest	0	0	0
3	South	0	1	0
4	West	0	0	1

Note that each region except for the Midwest scores 1 on one and only one of the dummies. The Midwest scores 0 on all three dummies. It is the reference category. This is appropriate, we could argue conceptually, since the midwest is one of the safest regions of the country, and it also represents the "heartland." The terms omitted category, base cell, and reference category or reference string are equivalent.

\*\*\* NOTE \*\*\* In the current data file the dummies we have are for: MWEST WEST and SOUTH; if you were to use all of those NORTHEAST would be the reference string

As usual, if I run a simple correlation, or a simple regression, the b weight for the dummy will tell me the difference between states scoring 1 and states scoring 0 on that dummy. BUT if I do a multiple regression putting all of these dummies in together, then the b weight is contrasting states scoring 1 on the variable with the states in the reference category.

Each  $b_j$  equals the difference between the mean of the corresponding cell and the mean of the base cell. (Darlington p. 235; see also Hamilton p. 97.)

What we are doing here is using dummy variables to imitate one-way analysis of variance. (See Hamilton 95-101 for more detail).

Here are some examples for you to interpret.

DEP VAR:PRISRA85	N:	50	MULTIPLE R:	.581	SQUARED MULTIPLE R:	.338
VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	1.271	0.192	0.000	.	6.625	0.000
SOUTH	1.022	0.254	0.608	0.6302521	4.025	0.000
NEDUM	-0.100	0.293	-0.049	0.6968641	-0.340	0.735
WESTDUM	0.478	0.266	0.268	0.6486486	1.797	0.079

DEP VAR:VIOLRA85	N:	50	MULTIPLE R:	.289	SQUARED MULTIPLE R:	.084
VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	309.250	58.975	0.000	.	5.244	0.000
SOUTH	148.875	78.017	0.339	0.6302521	1.908	0.063
NEDUM	83.528	90.086	0.157	0.6968641	0.927	0.359
WESTDUM	133.519	81.784	0.286	0.6486486	1.633	0.109

DEP VAR:PROPRA85	N:	50	MULTIPLE R:	.564	SQUARED MULTIPLE R:	.318
VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	3907.500	250.932	0.000	.	15.572	0.000
SOUTH	107.250	331.952	0.050	0.6302521	0.323	0.748
NEDUM	191.611	383.305	0.073	0.6968641	0.500	0.620
WESTDUM	1383.885	347.980	0.601	0.6486486	3.977	0.000

The t tests for each dummy reflect the significance of the mean difference between states coded 1 on that dummy and states coded 0 on all of the

dummies.

[YOU HAVE JUST READ AN EXTREMELY IMPORTANT POINT: re-read it.]

Furthermore, a (the constant) gives you the average score on the outcome for cases in the reference category; in this example, the midwest. SO if you add each dummy's b weight to the constant, you obtain the average score on the outcome for cases coded 1 on that dummy.

The Y predicted for each dummy = the mean for the cases scoring 1 on that dummy. So for the mean property crime rate in the South, since

$$Y_{\text{predicted}} = b_1 * \text{SOUTH} + b_2 * \text{NEDUM} + b_3 * \text{WESTDUM} + a$$

$$Y_{\text{predicted in South}} = 107 * 1 + 192 * 0 + 1384 * 0 + 3908 = 4015$$

You can check to see if this is true by running statistics by each regional group.

REGION		PRISRA85	VIOLRA85	PROPRA85
1 NE	(9)	1.171	392.778	4099.111
2 MW	(12)	1.271	309.250	3907.500
3 SO	(16)	2.293	458.125	4014.750
4 WE	(13)	1.749	442.769	5291.385

#### INTERACTION TERMS

In regression we usually assume that the effects of each of our predictor variables on the outcome is independent of the effects of the other predictors on the outcome. But this might not always be true. A certain predictor might have a more or less powerful impact on the outcome at certain levels of the other predictor(s).

#### The Usual Multiple Regression Model of Independent Impacts

For example, you know that imprisonment rates are higher in southern states. You also know that they are higher in states where more hard liquor is purchased per person (SPIRITS). You of course have a state-level theory about why this is the case.

For example, you might argue that states with harder drinking populations have populations that are harder to control, and less amenable to non-incarcerative sanctions. SO they have to throw more of them in prison.

You run the regression to see if each has an independent impact on the outcome, while controlling for the other predictor. Here is what you find:

---

\* \* \* \* M U L T I P L E R E G R E S S I O N \* \* \* \*

Listwise Deletion of Missing Data



R Square	.58871		DF	Sum of Squares	Mean Square
Adjusted R Square	.55216	Regression	4	166771.58950	41692.89737
Standard Error	50.88318	Residual	45	116509.39050	2589.09757

F = 16.10325      Signif F = .0000

----- Variables in the Equation -----

Variable	B	SE B	Beta	T	Sig T
SOUTH	82.360253	17.794367	.510415	4.628	.0000
SPIRITS	16.740931	7.990948	.212823	2.095	.0418
VIOLRA85	.140849	.052488	.383019	2.683	.0102
PROPRA85	.010900	.010848	.146227	1.005	.3204
(Constant)	.581343	39.359565		.015	.9883

End Block Number 1 All requested variables entered.

---

You can see that the effects of SOUTH and SPIRITS are still each significant. The model tells you about the main effect of each predictor; this is what non-contingent impacts means.

#### Idea of Contingent Impacts: Interaction Effects

##### Example

Sometimes you might be interested in the pattern of how a case scores on two variables. Here is an example that was an issue a few years ago, in November 1992, in New Jersey. It has come up again recently in Tinicum Township, PA, in a court case there.

Civil rights advocates argued that state police in the area were using arrest profiles up and down Route 95 in order to decide who to pull over and search for drugs. The advocates said that numerous motorists were detained simply because they were African American or Hispanic, AND because they had out of state plates on their cars.

What they were saying was that Hispanics (H) and African Americans (AA) did not have a much higher rate of detention than Whites (W), and those with out of state plates did not have a much higher detention rate than those with in state plates. BUT: those who are African American or Hispanic AND who ALSO have out of state plates did have a high rate of detention.

Here are the two main effects:

Race:            Probability of Detention:

W                Low

AA or H        Low to moderate

Plates:        Probability of Detention:

In S            Low  
 Out of S      Low to Moderate

Here is the interaction effect.

		Race	
		White	African American or Hispanic
Plates	In State	Low	Low to Moderate
	Out of State	Low to Moderate	HIGH

In other words there is an effect on the outcome that depends upon a particular combination of scores on the predictor: being both out of state and being African American or Hispanic.

Going back to our example with states, southern location, and spirits. You might argue, for example, that the effects of higher alcohol consumption on imprisonment rates are higher in southern as compared to non-southern states. Your theoretical logic might rely on interactions between temperature and alcohol as they influence violence, or some other mechanism.

The Mechanics of Creating the Interaction Effect

The best way to test for an interaction is to make dichotomous dummy variables out of your constituent variables, coding them so that the cell of most interest scores 1 on both dummy variables.<sup>4</sup>

		DUMMY 2	
		0	1
DUMMY 1	0		
	1		GROUP OF MOST INTEREST

Can you see how you would do this in the highway example above?

In our states example with spirits and SOUTH, we already have the dummy for southern location. We need to create a dummy variable for spirits. The simplest approach is to dichotomize the variable at the median (2.48 gallons/person/year). Your dummy will be coded 0 for states below that value,

---

4 Allison, P. D. (1977). "Testing for interaction in multiple regression." American Journal of Sociology 83.

and 1 for states at or above that value. You can do it with an IF statement, either in the compute dialog box, or in a syntax box. Here is what the syntax box commands might look like.

---

```
* COMMENT: Going to create a dummy for spirits variable
* States above median will be 1
* .
COMPUTE SPIRDUM=0 .
IF (SPIRITS>=2.48) SPIRDUM=1 .
EXECUTE .
* RUN FREQUENCIES TO BE SURE VARIABLE CREATED PROPERLY
* .
FREQUENCIES VARIABLE=SPIRDUM .
EXECUTE .
```

---

HINT: Check file info from the main menu to be sure you have less than 50 variables in your file. Otherwise, if you are running the student version, you will find it hard to save the file.

To see how SOUTH and SPIRDUM relate to each other, and how many states are in each cell, run Crosstabs

Your results should look like this:

		SPIRDUM	
		0	1
South	0	15	19
	1	10	6

You have six states scoring at or above the median on SPIRITS that are in the southern region. You can look at your original data to see what they are.

When we create interaction terms in multiple regression we multiply main effects, creating an interaction term that is a product of dummy variables.

The advantage of this approach, suggested by Allison, of creating dummies and then products of dummies, is that you can more easily interpret your results. You will see this below. If you create interaction terms based on variables that are not dummies, then it is harder to figure out what the term is telling you.

#### What It Means

Notice what will happen when we multiply two dummy variables together. The only cases that will score "1" on the interaction term will be those scoring "1" on both dummy variables. In other words the interaction term that we create contrasts the states that are both southern and above the median on



```

SPIRITS      26.902635    9.354005    .342007    2.876    .0060
SOUTH       92.134245    19.188079    .570988    4.802    .0000
(Constant)   76.166685    27.189501    2.801    .0074

```

End Block Number 1 All requested variables entered.

```

-> REGRESSION
-> /MISSING LISTWISE
-> /STATISTICS COEFF OUTS R ANOVA
-> /CRITERIA=PIN(.05) POUT(.10)
-> /NOORIGIN
-> /DEPENDENT prisra85
-> /METHOD=ENTER spirits south spirsout .

```

\* \* \* \* MULTIPLE REGRESSION \* \* \* \*

Listwise Deletion of Missing Data

Equation Number 1 Dependent Variable.. PRISRA85 RATE PRISONERS / 100K POP

Block Number 1. Method: Enter SPIRITS SOUTH SPIRSOUT

Variable(s) Entered on Step Number

```

1.. SPIRSOUT
2.. SPIRITS  AVG N GALLONS HARD LIQUOR PURCHASED PER
3.. SOUTH    SOUTH CENSUS REGION: DUMMY

```

Multiple R	.62993	Analysis of Variance			
R Square	.39681		DF	Sum of Squares	Mean Square
Adjusted R Square	.35748	Regression	3	112409.75797	37469.91932
Standard Error	60.94745	Residual	46	170871.22203	3714.59178
		F =	10.08722	Signif F =	.0000

----- Variables in the Equation -----

Variable	B	SE B	Beta	T	Sig T
SPIRITS	22.140465	9.670042	.281466	2.290	.0267
SOUTH	70.266163	23.293956	.435464	3.016	.0042
SPIRSOUT	52.989658	33.074375	.228770	1.602	.1160
(Constant)	88.908292	27.904612		3.186	.0026

You can follow these steps for interpreting.<sup>5</sup>

First, is the sign for the coefficient correct? You had predicted that spirits purchase in southern states would have more impact on imprisonment than spirits purchase in non southern states. Did it come out as predicted?

Second, construct the cell means for each cell in the interaction term. You can do this by means of the MEANS command, from the statistics menu. Here is what the syntax command would look like:

```

MEANS
  TABLES=prisra85 BY spirdum BY south
  /CELLS MEAN STDDEV COUNT

```

<sup>5</sup> These steps are appropriate only if you have created the interaction term out of two dummy variables and coded the dummy so that 1 on the interaction term is the case of interest.

/FORMAT= LABELS .

You must enter one of the predictors as a SECOND LAYER predictor; if you do this properly you will have TWO "BY"s in the syntax. If you have only one "BY" you have not done it properly.

You should get means that look like this:

Summaries of            PRISRA85    RATE PRISONERS / 100K POP 1985 - ST & FE  
By levels of            SPIRDUM  
                          SOUTH            SOUTH CENSUS REGION: DUMMY

Variable	Value	Label	Mean	Std Dev	Cases
For Entire Population			174.0200	76.0345	50
SPIRDUM	.00		167.8400	51.7580	25
SOUTH	0	NON-SOUTHERN	146.1333	32.9759	15
SOUTH	1	SOUTHERN	200.4000	59.0672	10
SPIRDUM	1.00		180.2000	95.1048	25
SOUTH	0	NON-SOUTHERN	149.7368	88.4325	19
SOUTH	1	SOUTHERN	276.6667	23.8551	6

+

Total Cases = 50

Third you want to learn if the additional explained variance in the outcome variable, provided by adding in the interaction term, is statistically significant. More formally, you want to test the null hypothesis that on the population of states the  $R^2_{\text{added}}$  by the interaction term is 0. You do this with an F test for the  $R^2_{\text{increment}}$ .

In the sample the  $R^2$  has gone from .363 to .398. The  $R^2_{\text{increment}}$  is .035. You have added 3.5% additional explained variance with this interaction term.

#### F test for $R^2$ Increment

An F test for  $R^2_{\text{increment}}$  exists, and it is better to use that F test than it is to rely on the t value associated with the b weight for the interaction term. Why?

Depending upon a number of factors, your interaction term may be more or less correlated with your main effect predictors. (See Hamilton pp. 91-92.) If the interaction term correlates highly with the main effects it inflates the standard error of your b weight, biasing your t test downward. The F test for  $R^2_{\text{increment}}$ , however, is not so influenced.

Furthermore, the F test for  $R^2_{\text{increment}}$  is actually a very general test, and can be used to find out, whenever you add additional predictors, if the change in  $R^2$  is significant.

Hamilton calls this the F-test for sets of coefficients (pp. 80-81).<sup>6</sup> The example Hamilton uses is concerned with the additional variance explained by a set of predictors (K) , once a first set of predictors (H) has already been entered. You are testing whether a "complex model, with K parameters, significantly improves upon a simpler model with H [i.e., fewer] parameters (0<H<K)" (Hamilton p. 80).

The logic and test used is the same when you are adding just one predictor, as you do when you are adding a test for curvilinearity, or a test for an interaction effect, or several predictors.

In short, you are testing for the unique contribution of a set of variables, given that an initial set of variables has already entered. The set may have one or more variables in it.

If set H is the set already entered, and set K is the H set already entered plus the additional predictors, the F test to determine if the change in R squared is significant is as follows (Darlington p. 124; Hamilton's formula is in terms of residual sums of squares, p. 80):

$$F_{\text{increment}} = \frac{[ [R(K)^2 - R(H)^2 ] / Q ]}{[ [ 1 - R(K)^2 ] / df_{\text{res}} ]}$$

where Q = number of ADDITIONAL regressors (predictors); I.E., Q = [K - H]

R<sup>2</sup>(K) = R<sup>2</sup> when all predictors entered  
 R<sup>2</sup>(H) = R<sup>2</sup> when just predictors from set H entered

dfres = residual degrees of freedom = n-K-1

So in our example here:

Q = 1 [we are adding just one new predictor]

R<sup>2</sup>(K) = .397

R<sup>2</sup>(H) = .363

dfres = 50-3-1=46

$$F = \frac{.034/1}{(1-.397)/46} = \frac{.034}{.0131} = 2.595$$

We then go to our lookup table, find F<sub>critical</sub> for 1 and 46 df, and see if our F<sub>obtained</sub> exceeds our F<sub>critical</sub> and allows us to reject the null hypothesis stated above.

---

6

Darlington (pp. 122-123) calls this three way decomposition.

This is a statistic that you can calculate by hand. Alternatively, you can ask SPSS to calculate it for you. You do this by telling SPSS to enter two sets of predictors. In the first set you have the main effect. In the second set you have the interaction effect. You want it to tell you the R squared change after entering the second set. The you want to be sure to ask it for the results for each block of predictors. Tthe statistic you must request is CHA for CHANGE. Here is the syntax:

---

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHA END
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT prisra85
  /METHOD=ENTER spirits south / METHOD=ENTER spirsout .
```

---

The output of interest is the **F change** under the **R square change**.

#### MULTICOLLINEARITY

With regression it is always easy to add one more predictor. But sometimes adding more predictors leaves you with "less" because it creates problems of collinearity (two predictors highly correlated with one another) or multicollinearity (several predictors highly correlated with one another).

Hamilton defines multicollinearity as "too-high intercorrelations among X variables" (p. 133). Darlington (sec. 3.4.3 and 5.3.2) defines collinear sets of regressors "as one whose member variables are highly correlated and largely interchangeable." (p. 125) The correlation between the variables diminishes the unique contribution of each regressor to the regression. (Darlington p. 82) It is harder to disentangle unique effects of two variables, or more than two variables, when they are highly correlated.

If you have a collinear set of predictors, deleting one of the members will not lower the R squared substantially.

Tolerance also tells us about collinearity of a regressor with other regressors; low tolerance indicates high collinearity.

Tolerance is telling you about the percentage of variance in that predictor that is independent of all the other predictors in the model.

**ARBITRARY RULE OF THUMB:** When tolerance goes below .10 for one or more predictors in a small data set, you may have problems with multicollinearity. A small data set would be less than 50-80 cases. When tolerance goes below .05 or .01 with larger data sets you also may have problems.

What are the effects of collinearity?

1. It can reduce the b weight for a regressor, because that regressor and the other regressors with which it correlates highly "fight" with each other to explain the same portion of the Y variable. It can even cause a b weight to change sign.

2. It can affect the significance test of a b weight, because standard errors of b weights among collinear regressors are inflated (Hamilton p. 134). It makes it harder to find statistically significant b weights.

How do you deal with multicollinearity?

Hamilton (p. 136) provides a specific list of things you can do.<sup>7</sup> Some of the things include:

1. Look at matrix of predictors.
2. Regress predictors on one another.
3. Delete suspect variables and see if standard errors change.

For your purposes, your best bet is to

- a. closely examine the matrix of correlations between predictors, looking out for correlations above .50 in small samples (less than 100 cases). In larger samples it is harder to decide when multicollinearity is a problem.
- b. closely examine tolerances, and be cautious when you have predictors with tolerance levels below .10 or .05.

But when you are doing this, be aware that a large number of moderately large correlations (.45) may present more of a problem than one fairly large correlation (.6). This is because collinearity is a property of sets of predictors, not just pairs of predictors.

Most important of all, look carefully at your matrix of predictors at the front end, and try and eliminate predictors that contribute substantially to collinearity.

For example, in the regression we are doing here - but WITHOUT the interaction term, if we had asked for multicollinearity diagnostics, we would have the following syntax:

---

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA END COLLIN TOL
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT prisra85
/METHOD=ENTER spirits south .
```

---

<sup>7</sup> Darlington (p. 192) acts like collinearity is not such a big deal. I disagree.

The output gives you a lot of information:

---

Variable	B	SE B	Beta	Tolerance	VIF	T	Sig T
SPIRITS	26.902635	9.354005	.342007	.958211	1.044	2.876	.0060
SOUTH	92.134245	19.188079	.570988	.958211	1.044	4.802	.0000
(Constant)	76.166685	27.189501				2.801	.0074

Collinearity Diagnostics

Number	Eigenval	Cond Index	Variance Proportions		
			Constant	SPIRITS	SOUTH
1	2.33747	1.000	.01764	.01893	.06443
2	.60536	1.965	.01177	.03400	.81629
3	.05718	6.394	.97059	.94707	.11927

---

Regression diagnostics is a complicated business. In addition to the diagnostics you see here you also can get variance inflation factors (VIF), and eigenvalue decomposition of the predictor matrix. For lots of details on this see:

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: Wiley.

For a worked through example see the appendix in:

Covington, J., & Taylor, R. B. (1991). Fear of crime in urban residential neighborhoods: Implications of between and within-neighborhood sources for current models. The Sociological Quarterly, 32, 231-249.

DO WE HAVE A CURVILINEAR RELATIONSHIP?

Your question here is: does the relationship between X and Y have a significant non-linear component?

For example, in the work on fear of crime people argue that increasing physical deterioration results in increased fear among residents. But if deterioration is high, residents may ignore further increases in it. They may "tune out" extremely high levels of the predictor. In other words, the impact of X on Y "flattens out" at high impacts of X.

For some real data on this see:

Taylor, R. B., & Shumaker, S. A. (1990). Local crime as a natural hazard: Implications for understanding the relationship between disorder and fear of crime. American Journal of Community Psychology, 18, 619-642.

In a regression framework, what you are doing is adding an additional predictor,  $X^2$ , and seeing if the squared term adds additional explained variance.

In SPSS I think the simplest way to look at this is through the scatterplots.

1. Run scatterplot with the predictor.
2. Modify chart, asking for the linear regression line, and for it to show you the R squared in the legend. Print out the chart.
3. Go back to the chart, and change the options of how it can show the regression line. Let it show a curvilinear regression line. This is equivalent, in a regression, to adding a square term for a predictor.
4. Note the new R squared value
5. Compute the F test of R squared change, and see if it is significant. You are comparing the difference between the new and the old R squared values.

### Glossary

base cell  
collinearity  
contingent effect  
curvilinearity  
dummy variable  
F test for  $R^2$  increment ( $R^2$  increment)  
independent effect  
interaction effect  
main effect  
multicollinearity  
omitted category  
 $R^2$  increment ( $R^2$  change)  
reference string  
reference cell  
spurious correlation  
tolerance  
unexpected change