

■ RICK HESSE, Pepperdine University

Truly Independent Variables

by Rick Hesse, Feature Editor

The danger of having a package with a multiple regression tool is that it is so easy to use, but there are hidden dangers aplenty. If the package used by the students doesn't automatically check for multicollinearity, the results may be misleading. My first rule of data that I give my MBA students is "Look at the data!" and that includes making sure that independent variables for a multiple regression are truly independent. To this end I developed a simple template that uses the **RSQ** function in Excel to test all the combinations of a simple linear regression for each pair of independent variables and the dependent variable.

The Initial Analysis

A national company has data for the sales (\$000) for 14 cities across the country, with the number of sales agents, marketing budget, and number of stores in their area given in Figure 1.

Inputs for the template are the labels for the dependent variable (**B17**) and up to 10 independent variables (**C17:L17**), the values of up to 100 dependent variables (**B18:B117**), and the independent variables (**C18:L117**). The equations for the white cells in the table are illustrated by the formula in **B6**.

B6: =IF(\$D\$18="", "", RSQ(B18:B117, \$D\$18:\$D\$117)) copy to C7:D7

This formula checks to see if there is any data for that variable. If not, the cell remains empty. Otherwise the r^2 is computed using the **RSQ**(variable1, variable2) function to determine the biased r^2 for a simple linear regression (straight line fit minimizing squared error) using the two variables at the intersection. For cell **B6**, this would be Sales(k) versus Stores(#). Note that the shaded (blue) cell, **B4**, computes the

same value, because the table is symmetrical. The r^2 value is the same regardless of whether Sales(k) is defined as the dependent or independent variable. The major diagonal is 100%, which is the r^2 for a variable predicting itself—which it does it perfectly every time. Rows 8→14 have been hidden in this case.

Figure 1 shows that the conditional formatting rules for cell **C7** turns the value bold, red and with a strike-through. This indicates that the row/column variables are possibly not independent. The rule for the conditional formatting is shown in Figure 2 and is for all the non-colored cells below the major diagonal. Note that cell **E5** has the same value as cell **C7** but no extra conditional formatting.

This rule turns on the formatting if either r^2 for the dependent variable is less than the r^2 of the two independent variables. In this case, the r^2 for Sales vs. Sales Agents is only 80.19% while the r^2 for Sales Agents and Budget is higher (85.25%). This means that Sales Agents explains Budget better than Sales (the dependent variable) and since it is the weaker variable (80.19% versus 94.05%), it should be eliminated from any multiple regression analysis.

Eliminating Colinearity

We must now eliminate any independent variable that shows colinearity, or dependence upon another independent variable. For this example, we need to delete the label and values for Sales Agents (**E17:E31**), and our resulting table has no more colinearity, as shown in Figure 3.

These two variables are independent in terms of each other and also due to the fact that the number of stores and the marketing budget are known or under



Rick Hesse

is professor of quantitative methods at Pepperdine University in the Graziadio Graduate School of Business and Management. He received his BS, MS, and DSc at Washington University School of Engineering in

applied math and computer science. Dr. Hesse is the author of *Managerial Spreadsheet Modeling & Analysis and Applied Management Science: A Quick & Dirty Approach* (with Gene Woolsey), articles in numerous journals, and software for personal computers. Rick was the first professor to be awarded the Outstanding Civilian Service Medal by the Department of the Army at West Point in 1982, and was the winner of the Decision Sciences Institute's Innovative Instructional Award in 1981.

Rick Hesse

Pepperdine University
Graziadio Graduate School of Business
and Management
Malibu, CA 90265
rickhesse@aol.com

	A	B	C	D	E
1	r² Table (Coefficient of Determination)				
2	Sales Forecasting				
3	r ²	Sales(k)	Budget (\$k)	# Stores	Sales Agents
4	Sales(k)	100.00%	94.05%	18.93%	80.19%
5	Budget (\$k)	94.05%	100.00%	4.27%	85.25%
6	# Stores	18.93%	4.27%	100.00%	2.34%
7	Sales Agents	80.19%	85.25%	2.34%	100.00%
15					
16		Y	X1	X2	X3
17		Sales(k)	Budget (\$k)	# Stores	Sales Agents
18		\$2,811	\$218	2	4
19		\$4,047	\$252	6	5
20		\$1,659	\$55	4	2
21		\$1,433	\$106	1	3
22		\$5,480	\$404	5	7
23		\$1,593	\$99	2	2
24		\$5,965	\$507	4	6
25		\$2,174	\$128	3	2
26		\$1,842	\$47	5	1
27		\$1,580	\$75	3	1
28		\$2,312	\$120	4	2
29		\$2,414	\$150	3	3
30		\$2,916	\$225	2	4
31		\$2,264	\$90	5	1

Figure 1: Initial r² analysis for sales(k).

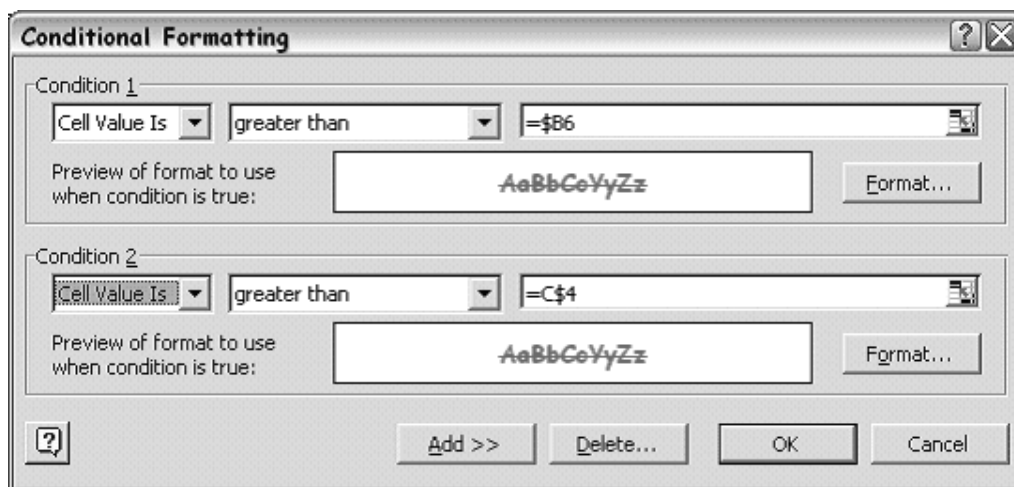


Figure 2: Conditional formatting for lower part of r² table.

control of the company. A point of discussion might be whether the weaker variable, number of stores, should be included in a multiple regression analysis.

Second Example: Median House Prices

I wanted personally to do some analysis of the median house price in my area, Ventura, California. I used data provided from UCSB's Economic Forecast Project 2005 Report (thanks to Dr. Bill Watkins, Director). I wanted to see if I could forecast the Median House price in Ventura County, which is situated between Pepperdine (Malibu) and UCSB (Santa Barbara). I chose possible independent variables as Year, Retail sales (\$M) in Ventura County, Tourism income (\$M) from hotels, and GCP/capita (\$k). The data and results are shown in Figure 3.

Analysis using the R² template indicates that all four independent variables that I chose are strongly related to each other. In fact, these variables predict each other better than median house price. If we use only the strongest (Retail sales) variable, then we run into another problem because only Year is truly independent (can be known with certainty or set to a value). The retail sales would have to be forecast for 2005 (and all the other variables if they were in the model) to be able to forecast the median house price in Ventura. Thus the search needs to continue for other variables which are truly independent if I want a multiple regression. If I want a single variable regression model using years, then I could fit a quadratic, cubic or even an exponential curve, using the coded year (1 = 1993) as the independent variable.

Conclusion

The fact that all of these economic factors are related to each other highlights the difficulty of finding truly independent variables that are independent of each other and that are known or can be con-

r^2	Sales(k)	Budget(\$k)
Sales(k)	100.00%	94.05%
Budget(\$k)	94.05%	100.00%
# Stores	18.93%	4.27%

Figure 3: Colinearity removed from data.

trolled (like a budget, or dosage of a chemical, etc.). This template provides a quick look at the data and I will leave the choice of the forecasting model up to each professor. I thought that showing this simple template (available as always through the DSI website) might help your students to investigate independence before they just shove all the data into a computerized forecasting model. ■

Excel file discussed in article is available from the Decision Line Web site at www.decisionsciences.org/decisionline.

	A	B	C	D	E	F
1	r^2 Table (Coefficient of Determination)					
2	Sales(k) Independent Variables					
3	r^2	MedianPrice	Year	Retail (\$M)	Tourism (\$M)	GCP/Cap (\$k)
4	Median Price	100.00%	77.08%	85.29%	71.30%	79.37%
5	Year	77.08%	100.00%	97.52%	94.65%	96.81%
6	Retail (\$M)	85.29%	97.52%	100.00%	95.54%	98.85%
7	Tourism (\$M)	71.30%	94.65%	95.54%	100.00%	96.49%
8	GCP/Cap (\$k)	79.37%	96.81%	98.85%	96.49%	100.00%
15						
16		Y	X1	X2	X3	X4
17		Median Price	Year	Retail (\$M)	Tourism (\$M)	GCP/Cap (\$k)
18		\$211.0	1993	\$3.9	\$77.3	\$28.1
19		\$206.6	1994	\$4.3	\$89.7	\$28.8
20		\$199.9	1995	\$4.4	\$83.9	\$29.9
21		\$205.7	1996	\$4.6	\$85.6	\$31.1
22		\$219.3	1997	\$4.9	\$99.5	\$33.9
23		\$233.8	1998	\$5.2	\$101.2	\$36.6
24		\$255.0	1999	\$5.9	\$114.7	\$42.8
25		\$295.1	2000	\$6.5	\$126.4	\$47.0
26		\$322.6	2001	\$6.8	\$125.5	\$50.2
27		\$372.4	2002	\$7.2	\$128.1	\$52.6
28		\$462.5	2003	\$7.7	\$132.1	\$55.0
29		\$587.8	2004	\$8.3	\$138.9	\$57.4

Figure 4: Median house price data for Ventura County.

INNOVATION, from page 7

(in states as diverse as New York, Texas, and Georgia) at which successful implementation has already occurred, have included vastly differing characteristics: a private university in a rural setting with a residential, traditional-aged student body; a public university in an urban setting with a largely commuter, non-traditional student body in the business school; and a public university in a rural setting with a majority (~80 percent) of traditional students in the undergraduate body, but with a substantial number of non-traditional students at both levels.

Furthermore, the sizes of these institutions varied from under 3,000 to over 10,000 students, while the character of these institutions included those with heavy engineering roots to one with very traditional liberal-arts and teacher education roots. Among the universities were a national doctoral/research university and two Masters I (comprehensive) universities. This approach has been used successfully in honors courses and non-honors courses. So it is fair to say that significant transferability has been demonstrated across institutions, and indeed significantly

different types of institutions and settings.

In each of these situations, course objectives have been demonstrably met for students and clients, non-majors and undecided students have gained a working knowledge of the "ways of knowing" and the "ways of doing," and majors have experienced exciting applications of the knowledge in their field. ■