Developments In Business Simulation & Experiential Exercises, Volume 23, 1996

AN ECONOMETRIC MULTIPLE REGRESSION CASE IN EXPERIENTIAL LEARNING

Craig G. Harms, University of North Florida, 4567 St. Johns Bluff Rd., S., Jacksonville, FL 32224-2645 (904) 646-2780

ABSTRACT

In the rush to create a more realistic learning environment, creative faculty are developing extensive case studies for classroom analysis. This paper presents the development and analysis of a multiple regression model using econometric data to help predict sales of a sporting good consumer product within a given geographic region or metroplex. The students are fully involved in the process from model conception to data collection to quantitative analysis. The case requires three weeks of time, utilizing 90 minutes of class time during each of the three weeks.

INTRODUCTION

A company is contemplating expanding its distribution network by opening new warehouses in one and possibly two additional cities in the U.S.A. Currently the company operates 15 facilities in every corner of the country. Obviously a location for new operations is a multi-faceted decision. One of the critical pieces that will go into the final decision is a quantitative model that will help predict expected sales penetration in the new region where the warehouse is opened.

The following question was posed to a group of “junior executives” in the company: “What demographic and econometric variables will help predict sales of our product?” Rather than merely discussing the model, the class proceeded through development, data collection, data base creation, and finally quantitative analysis and final model development.

DEVELOPING A CAUSAL MODEL

Historic sales data existed for the previous calendar year from each of the current 15 warehouses, which became the values of the dependent variable. Students were immediately worried about the small sample size. Although a small sample size, no additional data was available. Because of the small sample size, the validity of the model is more difficult to defend. Quite possibly a longitudinal study could be performed where sales data from each warehouse for a number of years would increase the sample size. But, as previously stated, only one year of data was available.

The case requires each student to present one independent variable, quantify the independent variable, hypothesize how its relationship would affect the dependent variable, and explain a plan of attack to secure the actual values of the independent variable.

Students are notorious for coming up with a causal variable that makes perfect sense, such as “competition” and then not be able to quantify it or have any notion of how to acquire the information. Needless to say most competitors are not going to give you their private sales and research data! An example of the various independent variables that were suggested are listed in Table One.

TABLE ONE

Causal Variables Used to Determine Sales in Existing Markets

1. College enrollment in the city
2. Days of precipitation in the city
3. Average relative humidity
4. Athletic oriented clubs
5. Average gross income per person
6. Population between 15 and 40 years of age
7. Number of sporting goods shoe stores
8. Mean temperature
9. Expected change in population in five years
10. Expected percentage change in population in five years
11. Percent of workforce in service industries
12. Average personal disposable income
13. Population in SMSA
14. Average age of population
15. Number of sunny days
16. Number of high schools
17. Level of unemployment
18. Per capita personal income
19. Students in K-12
20. Percent of owner occupied homes
21. Number of births in city last year

DATA COLLECTION

Each student was assigned to research and obtain the values for his or her particular independent variable at each of the 15 cities in the study and note very carefully where the data is obtained since we will go back to collect data for possible future city locations for the various independent variables that became part of the final model(s).

Some of the data was easy to obtain. The school library has an extensive government section, which includes volumes of statistical data from the U.S. Labor Department (at least some of our tax money is well spent). Students also made trips to the weather bureau and called the Chamber of Commerce in each of the cities.

The most unique search was to find the number of high schools in the city. Our library also has a very good telephone section with up-to-date phone books from several hundred cities. One student looked in the government section of the phone book and counted the number of high schools in each city. (What a project!)

Within one week each student turned in a sheet of paper with the values for his/her independent variable at each of the 15 cities as well as their hypothesis as to the relationship with sales. A discussion in class quickly brought about several arguments about the hypotheses. In particular was the discussion about the number of days of precipitation (variable #2). Some students believed that as the number of days of rain goes up, sales would decrease because the activity that uses the product would occur less-rain "dampens" the activity. But others thought rain would bring people to the mall and buy more product for lack of anything else to do.

The students were engaged and the beauty of this case came forth. The "rain" variable would be included in the database and evaluated by the students. At that point we would find whether the relationship was direct (positive B-coefficient) or inverse (negative B-coefficient) or not significant in the first place.

CREATION OF THE DATABASE

As previously stated, model inception and data collection required one week. After one-week information for variables #1 through #18 in Table One was collected from the students.

The faculty member is allocated one week to enter the data for each of the 18 independent variables into a database, prepare disks for each student, and pass them out. Realism is not for the lazy faculty member.

WHAT MAKES A GOOD MODEL'

This case culminates a six-week study (and five other cases) involving simple and multiple regression, ANOVA, dummy tables, transformations, and residual analysis. From the beginning of this case, the ground rules included no residuals or transformations. Developing the multiple regression model that "passed mustard" was enough of a task.

A good model included the following criterion:

a. Correlation coefficients between independent variables must be smaller than 0.6 (or -0.6)—try to minimize the multi-colinearity!
b. A model that passes a .05 F-test (F of about 4.0).
c. A model with three or more independent variables.
d. At least one and hopefully more than one of the independent variables must pass a .05 t-test. (Trying to find only models with all variables passing the .05 t-test was impossible.)
e. None of the independent variables with calculated t-statistics of less than 1.2 (or -1.2). (There was much room for "marginal" t's)
f. An adjusted R-square of 0.7 or larger.
g. The percent deviation (e-i/Y-hat: error term divided by the model forecast) less than 20% for at least one-half of the observations.
h. The plot of the multiple regression Y-hats versus the driver variable (high t-independent variable) visually appear as a "skinny banana."

Students were told to work alone and that grades would not be awarded solely on the value of adjusted R-square--which was the "bottom-line" measure of performance. Defending their models and explaining why they accepted certain independent variables and measures of performance was much more important than just good numbers.

DATA ANALYSIS

Armed with a disk containing the database, students went about developing *two* multiple regression models-with completely *unique* variables. Why two? If information for one of the independent variables did not arrive in time or was lost, only one of the two models would be inoperable.

THE DAY OF MODEL PRESENTATION

An entire class period was allocated for model presentation. The exercise consisted of two unique parts. First, students were given a set of independent variable values for each of the 18 variables. They applied both models to these values, picking their particular group of significant variables. The list of test values is presented in Table Two.

TABLE TWO
Test Values of All Independent Variables

	Variable	Test Value in New City
1.	College enrollment in the city	73,000
2.	Days of Precipitation in the city	78.0
3.	Average relative humidity	55.0
4.	Athletic oriented clubs	16.7
5.	Average gross income per person	16,000
6.	Population between 15 and 40 years of age	301,667
7.	Number of sporting goods shoe stores	518
8.	Mean temperature	61.0
9.	Expected change in population in five years	40,000
10.	Expected % of change in population in five years	+6.0
11.	Percent of workforce in service industries	35.0
12.	Average personal disposable income	8,900
13.	Population in SMSA	500,156
14.	Average age of population	31.9
15.	Number of sunny days	120
16.	Number of high schools	32
17.	Level of unemployment	8.3
18.	Per capita personal income	13,678

Each student went to the board and wrote the values of expected sales (Y-hat) for both models. Of the 4Q sales values (20 students
* two models per) there was only *one* duplication. The students had worked alone and found a large variety of good models, and some poor ones too. The range of sales values (once the math errors were removed) was 750,000 to 1,245,000. Over 70 percent were between 950,000 and 1,100,000. The students were very surprised that with so many combinations, so much latitude as to model acceptance, and the small number of observations, the range of the Y-hats is very tight. The students did perform expertly!

The second part of the class was spent by the presentation of specific models with statistical details. Two models are presented below. They are not the best or the worst, merely two good models from a sample of 39 unique models.

A FIRST MODEL

The first model includes three independent variables: number of high schools (#16), average humidity (#3), and average personal disposable income (#12). Table Three presents the correlation matrix. The multi-colinearity test passes, as the largest independent variable correlation is 0.29. Table Four presents the ANOVA table. Although two of the independent variables display rather weak t-statistics, both contribute to the adjusted R-square measure. Table Five presents the percent deviation table (right column). Nine of 15 deviations are within plus or minus 20 percent. Figure One is a plot of the Y-i's (symbol: x) and Y-hats (symbol: o) versus the number of high schools.

TABLE THREE
Correlation Matrix

	COL. 0	1	2	3
Row 0	1.00	.94	.32	9
Row 1	.94	1.00	.22	2
Row 2	.32	.22	1.00	6
Row 3	.39	.29	.16	10

TABLE FOUR
ANOVA Table

Source	Sum of Square	df	Mean Square	F
Model	4571644753326.000	3.	1523881584442.000	36.219
Error	462820999168.000	11.		
Total	5034465752494.000	14.		

Variable	Estimated Coefficient	Estimated Std. Dev.	T-stat
Intercept	-706019.900		
Number of High Schools	17205.2462	1891.9858	9.0938
Average Relative Humidity	6907.1309	5885.1872	1.1736
Personal Disp. Income P	91.5453	75.1986	1.2174

R-squared = .908 R = .953 Adjusted R-squared = .871
Std. Error of Est. = 205121.000
Continue? 1+Yes, and do NOT print Y(i)s, Y(hats), and c(i)s;
2 = print Y(i)s, Y(hats), and c(i)s;
3 = go to Main Menu; 9 = End of program;
4 = print X(1)s, Y(i)s, Y(hats), and e(i)s;
5 = go back to top of ANOVA table.
Or, 6 = **store the residuals just computed for overlay graphics**
Enter 7 for Help screen on overlay residuals

TABLE FIVE
Percent Deviations

Other # I	X(1) Value X(i)	Historic Y(i)	Forecasted Y-hat	Deviation E(i)	%Dev
1	25.0000	495122.0000	879323.3355	-384201.335	-77.60
2	5.0000	696590.0000	535218.4122	161371.588	23.17
3	107.0000	2423567.0000	2401129.4696	22437.530	.93
4	17.0000	423429.0000	420204.9343	322437.530	.76
5	27.0000	1134567.0000	859873.0500	3224.066	24.21
6	10.0000	689556.0000	619912.7002	69643.300	10.10
7	31.0000	613217.0000	911241.0528	-298024.053	-48.60
8	25.0000	1016115.0000	901303.2714	114811.729	11.60
9	16.0000	748455.0000	671530.7949	76924.205	10.28
10	23.0000	765099.0000	7735993.0835	-8500.083	-1.11
11	21.0000	1023448.0000	781308.4849	242139.515	23.66
12	103.0000	2249475.0000	2205787.9873	43687.013	1.94
13	18.0000	707650.0000	871340.4865	-163690.486	-23.13
14	22.0000	579441.0000	680012.9249	-100571.925	-17.36
15	16.0000	588484.0000	642428.4742	-53944.474	-9.17

FIGURE ONE
Plot of Historical Observations and Forecasts

```
2323400.00:-----------------------------------------------------------o
2223232.00:---------------------------------------------------------x--
2123064.00:---------------------------------------------------------o--
2022895.00:------------------------------------------------------------
1922727.00:------------------------------------------------------------
1822559.00:------------------------------------------------------------
1722391.00:------------------------------------------------------------
1622223.00:------------------------------------------------------------
1522055.00:------------------------------------------------------------
1421886.00:------------------------------------------------------------
1321718.00:------------------------------------------------------------
1221550.00:------------------------------------------------------------
1121382.00:-------------x----------------------------------------------
1021214.00:---------x--------------------------------------------------
 921045.71:------------x-----------------------------------------------
 820877.50:--------o---oo-o--------------------------------------------
 720709.40:------x--o-o------------------------------------------------
 620541.30:x--x--o-x-o-------------------------------------------------
 520373.10:o--o--x---x-----x-------------------------------------------
 420204.90:-------o----x-----------------------------------------------

          4   11   20   29   37   46   54   63   71   80   89   97  106
      1 = Main Menu;   2 = Graph other option;   9 = Graph Menu
                                        Number of High Schools
```

A SECOND MODEL

This model contains four independent variables: number of sporting goods shoe stores (#7), expected change in population (#9), percent of workforce in service industries (#1 1), and per capital personal income (#18). The various measures in the second model are very similar to the first model. A winner is not chosen. Both student models "pass mustard." Table Six presents the correlation matrix. The multi-colinearity test was passed, as the largest independent variable correlation is 0.46. Table Seven presents the ANOVA table. Although two of the independent variables display rather weak t-statistics, both contribute to the adjusted R-square measure. Table Eight presents the percent deviation table (right column). Eleven of 15 deviations are within plus or minus 20 percent. Figure Two is a plot of the Y-i's (symbol: x) and Y-hats (symbol: o) versus the number of shoe stores

TABLE SIX
Correlation Matrix

	Col. 0	1	2	3	4
Row 0	1.00	.92	.47	.37	.30
Row 1	.92	1.00	.32	.40	.45
Row 2	.47	.32	1.00	-.21	.13
Row 3	.37	.40	-.21	1.00	.46
Row 4	.30	.45	.13	.46	1.00

TABLE SEVEN
ANOVA Table

Source	Sum of Square	df	Mean Square	F
Model	4628833885587.094	4.	1157208471396.773	28.529
Error	405631918080.000	10.	40563191808.000	
Total	5034465803667.094	14.		

Variable	Estimated Coefficient	Estimated Std. Dev.	T-stat
Intercept	410364.800		
Number of High Schools	1181.8948	155.0078	7.6247
Average Relative Humidity	2.9732	1.1985	2.4809
Personal Disp. Income P	38629.4217	25037.1731	1.5249
	-80.5839	43.1616	-1.8670

R-squared = .919 R = .959 Adjusted R-squared = .875
Std. Error of Est. = 201403.100

TABLE EIGHT
Percent Deviations

Other # I	X(1) Value X(i)	Historic Y(i)	Forecasted Y-hat	Deviation E(i)	%Dev
1	157.0000	495122.0000	522925.3678	-27803.368	-5.62
2	31.0000	696590.0000	475299.5838	221290.416	31.77
3	1230.0000	2423567.0000	2306154.1836	117412.816	4.84
4	277.0000	423429.0000	799468.3723	-376039.372	-88.81
5	639.0000	1134567.0000	1037135.0099	97431.990	8.59
6	85.0000	689556.0000	387403.1386	302152.861	43.82
7	313.0000	613217.0000	590477.4854	22739.515	3.71
8	653.0000	1016115.0000	1073124.4089	-57009.409	-5.61
9	396.0000	748455.0000	804689.0568	-56234.057	-7.51
10	371.0000	765099.0000	724430.0308	40668.969	5.32
11	470.0000	1023448.0000	865754.9998	157693.000	15.41
12	1646.0000	2249475.0000	2255510.6843	-6035.684	-.27
13	527.0000	70765.0000	848464.8886	-140814.889	-19.90
14	381.0000	579441.0000	765550.9482	-189109.948	-32.12
15	114.0000	.5884840000	697826.4187	-109342.419	-18.58

1 = Continue, 8 = Go To Top of List, 2 = Go To End of List
3 = Go to Main Menu; 9 = End of program

FIGURE TWO
Plot of Historical Observations and Forecasts

```
2321760.00:--------------------------------------------x---------------
2219952.00:--------------------------------------------o--------------o
2118143.00:------------------------------------------------------------
2016335.00:------------------------------------------------------------
1914527.00:------------------------------------------------------------
1812719.00:------------------------------------------------------------
1710910.00:------------------------------------------------------------
1609102.00:------------------------------------------------------------
1507294.00:------------------------------------------------------------
1405486.00:------------------------------------------------------------
1303677.00:------------------------------------------------------------
1201869.00:------------------------------------------------------------
1100061.00:----------------------x-------------------------------------
 998252.60:----------------x-----oo------------------------------------
 896444.30:------------------------------------------------------------
 794636.10:---------o---o--o-o-----------------------------------------
 692827.90:x--o--------oo----x-----------------------------------------
 591019.60:--x-------x-------------------------------------------------
 489211.40:---xo-----o--x----------------------------------------------
 387403.10:o-o------x--------------------------------------------------

           30 138  273  408  542  677  812  946 1081 1216 1350 1485 1620
                     Number of Sporting Goods Shoe Stores
```

CONCLUSION

My first comment to the students is that this case is very similar to a project that their boss may assign as soon as they are hired. The project can be accomplished in a short amount of time and without handholding from other people in the firm. And the findings are meaningful!

Experiential Topics, whether in-class simulations or realistic, multi-faceted projects are so critical to the ability of students to hit the ground running when they are hired by a large corporation. It amazes me how weak students are when presented with a complicated project. They are very good at rote learning, yet are weak at creative thought and project completion.

This case has been a big success for mc and a good conclusion to our six weeks of regression analysis. I can assure possible employers that my students have dealt with a complicated and realistic quantitative case.

If class time is short, the data collection portion of the case can be eliminated by giving students a disk with a database already developed. This cuts down on their legwork, yet does allow students to deal with all of the statistical measures that they have previously studied. Whichever method is employed, the results of model prcscntation gratifies mc cvcry time.

REFERENCES

1. Coleman, B. Jay, *The Analysis of Statistical Relationships* (93 page booklet), 1993.

2. Neter, John & Wasserman, William, *Applied Linear Statistical Models,* Richard D. Irwin, 1974.