



*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

*No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.*

## VALIDATING PREDICTING EQUATIONS: THE SUPPLY OF FEEDER CALVES IN WEST VIRGINIA

John P. Kuehn

The objectives of this article were first to develop a viable predicting model of the West Virginia feeder calf supply using calves marketed as the dependent variable. The second objective was to validate the predicting model using the "leave-out-one-year" procedure and to derive an alternative predicting equation using the jackknife technique. The purpose of the emphasis on the second objective was to provide a simple and direct demonstration, of a useful and necessary technique, for the large group of applied economists, who often use econometric methods, but who do not consider themselves to be econometric specialists.

### INTRODUCTION

The use of regression analysis in the Agricultural Economics literature has been widespread. It is an extremely useful tool and many researchers are taking advantage of its usefulness. Unfortunately, however, the expanded number of applications of the technique has resulted in some abuses. One such problem is the validation of predicting equations.

Results are often presented with little or no validation; and predictions based on these models can be questioned. The best means of validating a prediction model is to compare the model's prediction to the actual phenomenon being predicted. This is not usually practical, however, since if this information was known, there would be no need for prediction. When validation does take place, it usually involves prediction of a subset of observations (the last five years of a twenty year period, for example). In some cases, however, the data for this subset are already incorporated in the predicting model. A more "honest" appraisal of the regression coefficients according to Mosteller and Tukey (1977) is to predict one or more subsets of observations that are not already incorporated in this model. One method of achieving this is called the Jackknife.

The name "jackknife" is intended to suggest the broad usefulness of a technique as a substitute for specialized tools that may not be available, just as the Boy Scout's trusty tool serves so variedly . . . the basic idea is to assess the effect of each of the groups (observations) into which the data have been divided, not by the results for that group (observation) alone . . . but rather through the effect upon the body of data (predicting equation) that results from *omitting* that group (observation), (Mosteller and Tukey, 1977).

The use of this technique is not new. In addition to Tukey's (1958) and Miller's (1964) work, Hartley and Hartley (1968) used it for estimating variances in simultaneous equations. Miller (1974 b) reviewed the technique and provided a list of 45 references. In agricultural economics research, the jackknife has been used to validate simultaneous equation models such as by Thompson, Sprott and Callen (1972). However, the method is not usually clearly explained or simply demonstrated. The objective of this paper is to provide an explanation of the method and to demonstrate it using a single equation predicting model.

### Objectives

The specific objectives of this paper are:

- 1) To predict the number of calves marketed in West Virginia for one year by a single equation lagged regression model;

and

- 2) To validate that equation by means of the "leave-out-one-year" procedure and
- 3) To compare the regression model to a jackknife model.

### THE MODEL

The first step of the analysis was to hypothesize a set of variables which influenced the number of calves marketed yearly in West Virginia. Emphasis was on variables which had a prior effect on the dependent variable, so all proposed independent variables were lagged one and two years. The following variables were considered: deflated feeder calf prices (deflated by CPI), calves born, calf deaths, feeder calf imports, the number of beef cattle on farms (Jan. 1), the number of beef cows on farms (Jan. 1), beef cattle prices deflated, cattle on feed in the United States (all other variables were for West Virginia), average feeder calf weight, deflated land values, the number of beef farms, the number of milkcows on farms (Jan. 1) and deflated milk prices.<sup>1</sup>

A correlation matrix was run for the set of independent variables to determine the nature and degree of interrelationships. A number of changes in the list of independent variables to be included in the model was made based on the correlation results. Two new variables were formed by combining inter-related variables: 1) FC = calves born—calf deaths + feeder calf imports and 2) BC = the number of beef cattle on farms in West Virginia—the number of beef cows on farms in West Virginia (Jan. 1). Also, the number of milk cows on farms in West Virginia was eliminated from the analysis due to its associative correlation with many of the variables which were considered to be more important influences on the dependent variable.

The revised list of variables was incorporated into a stepwise regression model for the years 1950-1976. The following equation resulted:

$$\begin{aligned} \text{CVS} = & 179.0725 - 0.3452 \text{FC}_2 - 2.0766 \text{BCPRD}_1 \\ & (13.8575) (0.1058) \quad (0.4317) \\ & + 5.3594 \text{ONFEED}_2 - 0.4284 \text{LANDVD}_1 \\ & (1.9241) \quad (0.1087) \\ & + 0.0245 \text{BFARMS}_1 \quad (1) \\ & (0.0041) \end{aligned}$$

where CVS = the number of calves marketed in West Virginia

FC<sub>2</sub> = (Calves born—calf deaths + feeder calf imports) lagged two years

BCPRD<sub>1</sub> = deflated beef cattle prices lagged one year

ONFEED<sub>2</sub> = the number of cattle on feed in the U.S. lagged two years

LANDVD<sub>1</sub> = the average sale value of land in West Virginia lagged one year

BFARMS<sub>1</sub> = The number of beef farms in West Virginia lagged one year

All variables were statistically significant at better than the .01 level. The overall F value was 28.4 and the R<sup>2</sup> was 0.8820. The Durbin-Watson statistic was 1.8697 which was in the indeter-

John P. Kuehn is Associate Professor of Agricultural Economics, West Virginia University.

minant range. To assess for multicollinearity, each independent variable ( $x$ ) was regressed on the remaining independent variables to determine if one of the  $X$ 's could be predicted by the remaining  $X$ 's. The low  $R^2$  values obtained, indicated no significant relationships existed. Then, a principal components test (eigenvectors) was run on the final stepwise model. The proportion of variation was examined and found to exist in all five dimensions, indicating the problem of multicollinearity was not serious. If most of the variation occurred in only three dimensions, it could be determined that two or more  $X$ 's were colinear or redundant. After the determination that the model was sound in terms of multicollinearity, the final step was validation.

### VALIDATION

The first step of the validation procedure was to run 25 separate regression equations.<sup>2</sup> Each equation predicted a given year's number of calves without the use of data from that year. For example, the number of calves marketed in 1960 was predicted by an equation incorporating the variables from (1) for the years 1950 through 1959 and 1961 through 1976. The objective of this procedure was to compare these predictions to those of equation (1). It was not expected that the results of the separate runs would be better than those of equation (1) but if (1) was a valid predictor the results would be similar. Any extreme variation between the two predictions would indicate

problems in the predictive ability of the original model, especially with regard to extreme variation in the data for one or more years.

Table 1 shows the predictions for the two procedures and compares them to the actual observations. The absolute differences between the predicted and actual observations were summed and averaged. The absolute deviations between equation (1) and the actual observations averaged 5.08 indicating that equation (1)'s future prediction should be correct within 5080 animals. The absolute deviations between the separate run predictions and the actual observations was 7.16<sup>3</sup>.

The similarity of predictions between the two equations along with the apparent stability of the separate runs coefficients could lead to the conclusion that the stepwise model was a valid predictor. However, a careful comparison of the predictions of the two equations reveals a potential problem. The variation in absolute deviation was fairly similar for the first 21 years, however, during the last three years, the deviation in the separate runs predictions increased substantially compared to those of the stepwise model. This deviation could be due to extreme or outlying values in the input data and raises questions as to the predictive reliability of the stepwise equation.

### The Jackknife

Extreme values in the data can have a significant effect on a least square regression. The regression coefficients as well as the

Table 1  
Predictions of Calf Marketings in West Virginia (1000) By the Stepwise Model, the Validation Equations (Separate Runs) and the Jackknife Model.

YEAR	Stepwise Model Predictions	Absolute Deviations (a)	Separate Runs Predictions	Absolute Deviations (a)	Jackknife Model Predictions	Absolute Deviations (a)	Actual Observations
1952	139	2	142	5	117	20	137
3	147	12	142	17	126	33	159
4	174	3	179	8	155	16	171
5	155	—	154	1	133	22	155
6	150	3	151	4	131	16	147
7	142	1	142	1	125	18	143
8	138	7	139	8	122	9	131
9	125	3	125	3	110	12	122
1960	120	3	121	4	107	10	117
1	125	2	126	3	114	9	123
2	127	—	127	—	116	11	127
3	129	1	129	1	117	13	130
4	130	1	130	1	119	12	131
5	129	11	127	13	118	22	140
6	125	14	124	15	115	24	139
7	119	4	120	4	110	6	116
8	124	6	125	7	115	3	118
9	126	7	127	8	117	2	119
1970	124	6	125	7	115	3	118
1	127	2	126	3	117	12	129
2	126	1	127	2	116	9	125
3	103	14	101	16	95	22	117
4	89	7	95	13	81	1	82
5	101	9	105	13	96	4	92
6	98	8	84	22	95	11	106
Average Deviation (a)		5.08		7.16		12.80	

(a) Between predicted and actual observations.



intercept can be altered and the predictive ability of the model could be lessened due to one or two outlying years of data. If these outlying values are not common occurrences, the jackknife model can improve future predictions for years when these extremes do not occur.

In effect, the jackknife equation is a weighted average of the separate runs equations discussed earlier. Extremes in a particular year do not significantly affect predictions since the equation is based on the series of each year without that year's data.

In the case of the feeder calf model, the following steps are necessary:

1. Each of the separate runs (leaving out that year's data) equations are arrayed in a matrix, variable by variable including the intercept. There were 25 separate equations, one for each year.
2. The original equation (1) is placed at the top of the matrix variable by variable (Table 2).
3. A new matrix is then formed by multiplying each variable in equation (1) by 25. Then, for each year of separate runs, each independent variable is multiplied by 24 and subtracted from its counterpart (already multiplied by 25) in equation (1). The new matrix will then have 25 equations of pseudo-coefficients (Table 3).
4. The jackknife equation is then formed by averaging the coefficients of each variable in the new matrix.
5. The standard errors are calculated by first calculating the standard deviations (SD) for each variable. The standard error then equals  $\sqrt{SD^2 \div 25}$ .

6. The  $t'$  values are determined by dividing each of the new coefficients by its standard error. These values approximate the student's  $t$  values. As sample size increases, these values asymptotically approach the actual student's  $t$ . It was judged that with 27 observations of data  $t'$  adequately approximated  $t$ .

The jackknife equation from Table 3 then is as follows:

$$\begin{aligned} \text{CVS} = & 192.6619 - 0.3641 \text{FC}_2 - 2.2716 \text{BCPRD}_1 \\ & (42.0109) \quad (0.1203) \quad (0.6353) \\ & + 4.132 + \text{ONFEED}_2 - 0.3303 \text{LANDVD}_1 \\ & (2.2737) \quad (0.2183) \\ & + 0.0187 \text{BFARMS}_1 \\ & (0.0037) \end{aligned} \quad (2)$$

### CONCLUSIONS

Two of the variables in the jackknife equation were not statistically significant ( $\text{ONFEED}_2$  and  $\text{LANDVD}_1$ ). The fact that this occurred raises questions as to the stability of the stepwise prediction model . . . more instability than a noraml evaluation of the statistical results would indicate.

The pseudocoeficients in Table 3 are indicative of internal stability. A close examination of these coefficients shows an apparently greater than normal amount of fluctuation in the last three years. An examination of the actual input data revealed a large increase in West Virginia land values in the last few years of the study period. Between 1974 and 1976, land values increased 24 percent *after* being deflated by the consumer price

Table 2  
Regression Coefficients of Separate Runs Leaving Out the Data From  
the Year Being Predicted—Comparison to Stepwise Model Coefficients.

YEAR	Intercept	FC <sub>2</sub>	BCPRD <sub>1</sub>	ONFEED <sub>2</sub>	LANDVD <sub>1</sub>	BFARMS <sub>1</sub>
All Years						
Stepwise Model	179.0725	-0.3452	-2.0766	5.3594	-0.4284	0.0245
1952	178.3849	-0.3522	-1.9740	5.3457	-0.4290	0.0248
3	172.3609	-0.2949	-2.1912	5.5073	-0.4421	0.0233
4	194.8279	-0.4119	-2.3266	5.4376	-0.4221	0.0260
5	180.8639	-0.3490	-2.0779	5.2883	-0.4259	0.0244
6	174.9588	-0.3343	-2.0891	5.4631	-0.4299	0.0249
7	180.5031	-0.3505	-2.0743	5.3482	-0.4284	0.0245
8	177.4499	-0.3291	-2.0934	5.1747	-0.4209	0.0243
9	177.8127	-0.3349	-2.0363	5.2247	-0.4252	0.0242
60	182.3965	-0.3415	-2.0439	4.9469	-0.4139	0.0267
1	183.4739	-0.3491	-2.0849	5.0903	-0.4201	0.0239
2	179.0949	-0.3452	-2.0766	5.3582	-0.4283	0.0245
3	178.6492	-0.3449	-2.0761	5.3739	-0.4280	0.0246
4	178.8107	-0.3450	-2.0747	5.3632	-0.4278	0.0246
5	178.7317	-0.3537	-2.0305	5.0929	-0.3995	0.0246
6	173.5301	-0.3443	-2.0366	5.2959	-0.4039	0.0252
7	181.2343	-0.3463	-2.0727	5.3026	-0.4316	0.0242
8	183.0112	-0.3543	-2.0991	5.4454	-0.4413	0.0244
9	185.8028	-0.3661	-2.1354	5.4869	-0.4416	0.0246
70	182.2943	-0.3593	-2.1106	5.5572	-0.4404	0.0248
1	180.2768	-0.3431	-2.0723	5.1216	-0.4162	0.0242
2	178.3511	-0.3480	-2.0848	5.5401	-0.4365	0.0249
3	194.4351	-0.3827	-2.2919	4.9804	-0.4227	0.0246
4	153.0499	-0.2858	-1.6831	5.8839	-0.4356	0.0247
5	174.5773	-0.3442	-2.1431	5.2001	-0.3649	0.0248
6	157.7749	-0.3025	-1.7328	7.4672	-0.6363	0.0268

**Table 3**  
**Pseudocoefficients Derived by Multiplying the All Years Stepwise**  
**Model Variables by 25 and Subtracting the Counterpart Variables**  
**(Multiplied by 24) for Each Year.**

YEAR	Intercept	FC <sub>2</sub>	BCPRD <sub>1</sub>	ONFEED <sub>2</sub>	LANDVD <sub>1</sub>	BFARMS <sub>1</sub>
1952	195.5749	-0.1772	- 4.5390	5.6882	-0.4140	0.0173
3	340.1509	-1.5524	+ 0.6738	1.8098	-0.0995	0.0533
4	-199.0571	+1.2556	+ 3.9234	3.4826	-0.5796	-0.0115
5	136.0789	-0.2540	- 2.0454	7.0658	-0.4884	0.0269
6	277.8013	-0.6068	- 1.7766	3.6674	-0.3924	0.0149
7	144.7381	-0.2180	- 2.1318	5.6282	-0.4284	0.0245
8	218.0149	-0.7316	- 1.6734	9.7922	-0.6084	0.0293
9	209.3077	-0.5924	- 3.0438	8.5922	-0.5052	0.0317
60	99.2965	-0.4340	- 2.8614	15.2594	-0.7764	-0.0283
1	73.4389	-0.2516	- 1.8774	11.8178	-0.6276	0.0389
2	178.5349	-0.3452	- 2.0766	5.3882	-0.4308	0.0245
3	189.2317	-0.3524	- 2.0886	5.0114	-0.4380	0.0221
4	185.3557	-0.3500	- 2.1222	5.2682	-0.4428	0.0221
5	187.2517	-0.1412	- 3.1830	11.7554	-1.1220	0.0221
6	312.0901	-0.3668	- 3.0366	6.8834	-1.0164	0.0077
7	127.1893	-0.3188	- 2.1702	6.7226	-0.3516	0.0317
8	84.5437	-0.1268	- 1.5366	3.2954	-0.1188	0.0269
9	17.5453	+0.1564	- 0.6654	2.2994	-0.1116	0.0221
70	101.7493	-0.0668	- 1.2606	0.6122	-0.1404	0.0173
1	150.1693	-0.3956	- 2.1798	11.0666	-0.7212	0.0317
2	196.3861	-0.2780	- 1.8798	1.0226	-0.2340	0.0149
3	-189.6299	+0.5548	+ 3.0906	14.4554	-0.5634	0.0221
4	803.6149	-1.7708	-11.5206	- 7.2286	-0.2556	0.0197
5	286.9573	-0.3692	- 0.4806	9.1826	-1.9524	0.0173
6	90.2149	-1.3700	-10.3278	-45.2278	+4.5612	-0.0307
<hr/>						
F <sub>2</sub> (mean)	192.6619	-0.3641	- 2.2716	4.1324	-0.3303	0.0187
J (Standard Deviation)	210.0543	0.6013	3.1767	11.3686	1.0916	0.0185
J (Standard Error)	42.0109	0.1203	0.6353	2.2737	0.2183	0.0037

index. This abrupt change, in effect, diagnoses land values as a statistically outlying variable in the last year or so of the study. The fact that this fluctuation occurred in the last part of the study period casts strong doubts on the usefulness of the stepwise predictor. When the last year's prediction of the stepwise equation is compared to the actual number of calves marketed (see Figure 1), it can be seen that the direction of the predicting line changes in relation to the actual trend.

It is possible that equation (2) could be a better predictor despite the lack of significance of two variables.<sup>4</sup> The yearly predictions from this equation are already presented in Table 1. It can be seen that the predictions do not fit the actual observations as well as equation (1), however, the absolute deviations in the last few years are lower than those of the stepwise model.

The implications arising from the validation procedure are that the stepwise equation could be the better predictor if the increase in land values slows; and the jackknife equation may be the better predictor if outlying values recur. Both equations will be used to predict calf marketings for 1977 and later years and results will be compared to the actual observations when they become known.

The results of this analysis should provide a scientific warning signal to researchers predicting economic phenomena. What originally appeared to be statistically sound predicting model was found to be of questionable value after being subjected to a vigorous validation procedure.

#### FOOTNOTES

<sup>1</sup>Sources of data included *Agricultural Statistics*, U.S.D.A., Washington D.C., U.S. Government Printing Office, 1950-1976; *Agricultural Prices, Annual Summary*, U.S.D.A. S.R.S., Crop Reporting Board, Washington, D.C., U.S. Government Printing Office, 1950-1976; and West Virginia Department of Agriculture, Charleston, WV.

<sup>2</sup>Actually, there were 27 observations (years of data) in the model, but since lags of two years were involved, the first two years' predictions were omitted due to missing values.

<sup>3</sup>An alternative method of comparing predicted to actual observations involves the use of indices of dispersion. A discussion of the technique can be found in Hee.

<sup>4</sup>It should be noted that if the jackknife model is used for predictive purposes, it should be cross-validated by using the separate runs (leave-out-one-year) procedure.



## REFERENCES

- Barr, A. J., J. H. Goodnight, J. P. Sail and J. T. Helwig. *A User's Guide to SAS 76*. SAS Institute, Raleigh, North Carolina, 1967.
- Hartley, H. O. and M. J. Hartley. "Jack-knife Variance Estimation for Simultaneous Equations," Paper presented at the 1968 Winter Meeting, Econometrica Society, Evanston, Illinois.
- Hee, Olman. "Tests for Predictability of Statistical Models," *AJAE*, 48 (1966):1479-1484.
- Miller, R. G. Jr. "A Trustworthy Jackknife," *Ann. Math. Stat.*, 35(1964): 1594-1605.
- \_\_\_\_\_. "An Unbalanced Jackknife." *Ann. of Stat.*, 2(1974a):880-891.
- \_\_\_\_\_. "The Jackknife—A Review." *Biometrika*, 61(1974b):1-15.
- Mosteller, Frederick and John W. Tukey. *Data Analysis and Regression*. Addison-Wesley Pub., Reading, Mass., 1977.
- Thompson, R. G., J. M. Sprott and R. W. Callen. "Demand, Supply and Price Relationships for the Broiler Sector, with Emphasis on the Jackknife Method." *AJAE*, 54(1972):245-248.
- Tukey, J. W. "Bias and Confidence in Not-Quite Large Samples." Abstract in *Ann. Math. Stat.*, 1958, p. 614.
- Wallace, T. Dudley. "Pretest Estimation in Regression: A Survey." *AJAE*, 59(1977):431-443.