Conference on Applied Statistics in Agriculture 1992 - 4th Annual Conference Proceedings

AN EXAMPLE OF PATH ANALYSIS APPLIED TO CLASSIFICATION VARIABLES APPLIED TO CLASSIFICATION VARIABLES

Richard E. Lund

Albert L. Scharen

Recommended Citation

Lund, Richard E. and Scharen, Albert L. (1992). "AN EXAMPLE OF PATH ANALYSIS APPLIED TO CLASSIFICATION VARIABLES APPLIED TO CLASSIFICATION VARIABLES," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1407

An Example of Path Analysis Applied to Classification Variables

Richard E. Lund and Albert L. Scharen
Montana State University
Bozeman, MT 59717-0002

Abstract

Path analysis was originally proposed to decompose and interpret causal linear relationships among a set of continuous stochastic variables. Research designs necessarily employed the natural variation in the system rather than the technique of controlling independent variables by selection of levels and categories which is emphasized in many experimental designs. Path coefficients are closely related to correlation coefficients, the size of which will be controlled when variation in the system is controlled. We examine a data set produced by research related to world-wide occurrence of a wheat pathogen and describe techniques for applying path analysis to its variables, some of which were merely categorical. Limitations in interpretation are noted.

1. Introduction

Path analysis, as originally conceived by Sewall Wright (1921 and 1960), was applicable primarily to decomposing and interpreting causal linear relationships among a set of continuous stochastic variables. Research designs necessarily utilize the natural variation in the system. In simple cases path coefficients are estimated by standardized regression coefficients obtained from one or more multiple regressions. Li (1977) offers a simple introduction to path analysis while Kempthorne (1973) offers a more advanced treatment. Applications in agriculture apparently are rare but examining those by Karlsson, et. al. (1988) and Dewey and Lu (1959) may be valuable.

Modern experimental designs do not rely on natural variation, but instead generally control the values for independent variables by selecting specific levels for study. Subsets of potential categories for categorical variables frequently are chosen by criteria. A major goal for these and other decisions about an experimental design is to improve efficiency in determining relationships among variables. But standardized regression coefficients are closely related to correlation coefficients for which it is known that experimental control of variation and selection of experimental material for study will influence the size of the correlations.

Categorical variables are generally accommodated in regression analysis by creating a set of indicator variables, one for each category. An indicator variable

---

Contribution J-2822 from Montana Agricultural Experiment Station.

generally is assigned the value one to denote an observation of that category and the value zero otherwise. (The actual regression may involve only a subset or a linear transform of these variables in order to handle their linear dependence). Path diagrams can become exceedingly complex when defining individual path coefficients for each variable. Moreover, standardized regression coefficients become inappropriate estimators for path coefficients when replication is unequal across categories due to its influence upon the size of coefficients.

Our objective was to 1) examine a problem associated with studying world-wide occurrence of a wheat pathogen and 2) show results when applying a technique proposed by Henry (1983) to develop a path analysis diagram displaying the relation among pertinent variables, some of which are categorical. Limitations in the interpretation resulting from researcher selection of categories for study are noted.

2. Research Data Base

Septoria nodorum blotch (SNB) (caused by a fungus) is a serious disease of wheat in many parts of the world (Eyal, et. al. 1987). Damage to crops can be considerable where environmental conditions are favorable for disease. Yield losses can be explained largely by reductions in kernel weight (KW). Much research effort has been oriented toward relating yield losses and reduced KW to disease symptoms appearing on plant leaves. Genetic resistance to SNB is known and efforts to produce resistant cultivars have been in progress for several years.

The U.S. Department of Agriculture (USDA) Septoria nursery was established in 1973 to support such effort. Approximately 30 winter wheat and 30 spring wheat cultivars have been grown continuously since then in many of the Septoria-prone countries around the world. Data referenced herein come from a uniform group of ten representative wheat cultivars that had been grown in the six years 1982 through 1987 at seven North American and European locations (See Table 1).

The ten cultivars were selected on the basis of 1) a range of known resistance and susceptibility; 2) genotypes having adaptability to a wide range of growing conditions; and 3) inclusion in the nursery at all locations during at least three of the six years under study. Data were deleted for certain wheat cultivar, location and year combinations affected by winter kill, extreme drought or the occurrence of other diseases to a degree that would have masked the effects of SNB. Additional details regarding the data base and recent research on SNB appear in Scharen, et. al. (1991).

Nursery plots at Bozeman, Montana consisted of four rows/cultivar, each 1.5 m in length and sown 30 cm apart. Since natural infection of SNB does not occur at Bozeman, two rows of each cultivar were inoculated, while the other two remained untreated and served as controls. Disease symptoms (in particular the relative area covered by lesions on the leaves) were assessed visually and recorded on a 0 to 9

scale approximately 2 weeks after the last inoculation. This measurement is called leaf-symptom hereafter.

Yields were determined by harvesting 25 heads at random from each inoculated and control plot. After threshing individually, kernels from each head were counted, weighed, and KW calculated. The response due to infection was quantified by dividing mean KW for inoculated plots by the mean KW for control plots (with rescaling by multipying by 100), where means were taken over all rows and replications within the experiment. Hereafter, this ratio is called relative kernel weight (RKW).

Since SNB occurs naturally at all locations other than Montana, data for the other locations relate to comparing artifically inoculated plots with control plots that were protected from infection by applications of a foliar fungicide. Complete details of experimental methods used at the six locations over a period of six years were not always furnished as a part of the nursery reporting procedure.

3. Creation of Explanatory Model and Path Diagram

A general linear model (GLM) defined RKW to be a function of the categorical factors and measured variables. These were location, year-within-location, cultivar and leaf-symptom.

The 29 categories for location and year-within-location were aggregated to form a supplementary categorical factor called infection-level. Assignment of categories low, medium and high for infection-level was determined by separating all data into three approximately equally sized groups on the basis of a low, medium or high mean value for RKW taken across all cultivars for each year at each location. Terms were included in the fitted models which expressed the interactive effect of leaf-sympton with the major components of location and year-within-location by the surrogate infection-level by leaf-sympton interaction.

If ten cultivars had been grown at seven locations over six years in a full factorial arrangement, the data set would contain $10 \times 7 \times 6 = 420$ observations. While the data set was nearly complete with respect to cultivar by location combinations (See Table 1), failure to grow all cultivars at all sites every year and a few missing measurements limited the data set to 170 useable records. The large reduction in data set size was regrettable and reduced precision in estimation. However, the authors feel that data omission did not bias the results, since data omission was largely related to classification by year (a nested rather than a crossed factor).

Preliminary analyses were performed by the GLM and REG procedures in the SAS System statistical package. Statistical tests for GLM were based upon type III

sums of squares due to the non-balance resulting from missing data. Interpretation of test results followed procedures proposed by Searle (1987) for unbalanced designs.

Path diagram construction becomes complex, is subject to mis-interpretation and seldom is applied to categorical data. Our viewpoint, while investigating the relevance of the measurable symptoms of infection, is that the categorical factors location, year-within-location and cultivar represent other unknown and unmeasurable environmental and genotypic variation which can be considered continuous variables for which path analysis is meaningful. Qualifications related to this viewpoint appear later.

We followed a procedure proposed by Henry to create a new continuous variable corresponding to each categorical factor. A linear transform was applied to the indicator variables defined during processing by procedure GLM in SAS, using the GLM regression coefficients as the transformation coefficients. While the indicator variables are not unique, the summations of the product of regression coefficients times values for the indicator variables for each classifying factor produces a unique value (ignoring scaling) for each data point (record). Use of procedure REG in SAS for subsequent regressions of RKW upon the newly created variables produced the standardized regression coefficients displayed in the path diagram. The sums-of-squares associated with the three categorical factors (locations, years within locations and cultivars) in the analysis of the first phase equals that associated with the three new continuous variables in the second phase.

4. Results and Discussion

Means across years for all data are presented in Table 1 by location and cultivar. Variation in the level of symptoms and the relative influence of infection on grain yield components were clearly related to location. Reduction in yield as measured by RKW reached 40 to 60% in Austria.

Analysis of RKW by GLM provided an R^2 of 0.75 with $p < 0.001$ for the effects of location, $p < 0.003$ for year-within-location and $p < 0.006$ for cultivar. Leaf-symptom, combining both its linear effect and its interaction with the factor called infection-level (to represent its interaction with location and year-within-location), was significant at $p < 0.058$. More specifically, the linear effect of leaf-symptom upon RKW and the non-linearity expressed by its interaction was examined in the GLM by relating RKW to leaf-symptom linearly across all cultivars by three equally spaced slopes for the three categories of infection-level. A test for variation in slope across the infection-levels was found to be significant at $p < 0.02$ while the particular slope for the low level of infection did not differ significantly from zero. A simple curve, such as a quadratic polynomial, did not adequately express the non-linear relation found between leaf-symptom and RKW. Other forms of interaction between the categorical factors and the factors with the measured variables could not be substantiated.

The path diagram in Figure 1 and path coefficients in Table 2 display estimates of the influence of the categories location, year-within-location and cultivar upon RKW both directly and indirectly through the pathway of influencing the biological processes measured by leaf-symptoms which thereby influence RKW. While a path diagram has admitted limitations when applied to categorical variables, its use here concentrates the significant information contained in about eighty categorical means for RKW and leaf-symptom (likewise the information in 80 GLM coefficients) into only seven interpretable path coefficients. These path coefficients were estimated by the standardized coefficients in two regressions (i) RKW over location, year-within-location, cultivar and leaf-symptom and (ii) leaf-symptom over location, year-within-location and cultivar, after transforming the coding for the categorical factors into linear variables as explained earlier. The R^2 values for the two regressions were 0.75 and 0.92 respectively. All regression coefficients were significant at $p < 0.001$.

The path coefficients herein have standardized scaling. Thus, the value 0.45 for the direct path between location and RKW implies that a 1.00 displacement in location leads to a 0.45 alteration in RKW, with all other variables remaining constant and with both RKW and location being measured in standard units (variance of unity). The indirect effects given in Table 2 are the product of the two path coefficients through the leaf-symptom pathway. Thus, the total effect of a 1.00 displacement of location is estimated to produce a 0.45 + (0.37 x 0.81) = 0.75 displacement in RKW, when holding the variables year-within-location and cultivar selection constant.

The standardization of all coefficients in a path diagram simplifies interpretation by enabling comparisons among all variables. However, such standardization can have little meaning and can lead to misinterpretation when applied to categorical variables. As noted earlier, geographic locations, cultivar selections and years studied herein were all determined subjectively when setting up this research. They do not represent a random selection of categories as would generally be implied by their inclusion in a path diagram. A different subjective criterion for selection would lead to an alteration in variation among categories and to an alteration of standardized path coefficients. Consequently, some attention to the GLM coefficients (and least squares means) for particular categories produced during the first phase of the analysis can provide additional clarification.

For example, infection in cultivar JCR11 produced a reduction of 7.1 in RKW as measured by least squares means, which was the largest reduction among all cultivars. Applying the path coefficients in Table 2 suggests that (0.22 / 0.25) x 7.1 = 6.2 of this reduction was related directly and (0.03 / 0.25) = 1.1 was related indirectly through the leaf-symptom pathway. Similarly, the location in Austria was associated with a reduction of 19.1 (differing by more than two standard deviations from the mean of all locations when measured upon the variable scale used in constructing the path diagram) which can be divided into a direct effect of (0.45/0.75) x 19.1 = 11.5 and an indirect effect of (0.30/0.75) x 19.1 = 7.6.

A comparison of the overall reduction of 7.1 in RKW for JCR11 to the reduction of 19.1 for Austria has meaning only when one has an interest in these particular categories and may have no meaning otherwise. Generally, vertical comparisons in Table 2 (and associated coefficients in Figure 1), such as noting that the total effect for location (0.75) is three times that for cultivar selection (0.25), have similar limitations. The relative size of the effects could be altered drastically by using a different set of locations and cultivars, for example by just dropping out either JCR11 or Austria. To the contrary, we believe that comparisons horizontally with in Table 2 and how these comparisons may change from one variable to another do have more general interpretable value.

Path diagrams are designed for linear effects and the effect of the presence of non-linearity in the response to leaf-symptom noted previously need be qualified. The path coefficient with a value of 0.37 relating leaf-symptom to RKW in Figure 1 constitutes an average across the three infection levels. The non-linearity found suggests that if the level of infection is generally high across all cultivars at a particular location for a particular year then the value 0.37 under estimates the correct path coefficient, at which time the value 0.37 should be higher and all entries in the column labeled indirect effects are too small. Conversely, the indirect effects measured by leaf-symptom are over estimated and actually become negligible when an especially low general level of infection occurs at a specific location-year.

In summary, the results obtained from the analysis of the particular locations, years and cultivars we selected show that location and annual local environment are the most important factors contributing to yield reductions as measured by RKW. The direct effect of infection, relative to the indirect effect measured through the leaf-symptom pathway, is marginally stronger with respect to variation in location and annual environment. Conversely, the direct effect is substantially stronger than the indirect effect with respect to variation in cultivar selection.

The selection of genotypes based upon grain yield is a uniformly effective method of obtaining selections that resist Septoria nodorum blotch. Indirect selection through the pathway of disease symptoms is much less effective when data from many locations are considered together.

5. References

Dewey, D. R. and K. H. Lu (1959), "A correlation and Path-coefficient Analysis of Components of Crested Wheatgrass Seed Production," Agronomy J., 51:515-518.

Eyal, Z., A. L. Scharen, J. M. Prescott and M. Van Ginkel (1987), The Septoria Diseases of Wheat: Concepts and Methods of Disease Management. CIMMYT, Mexico, D.F.

Henry, F. (1983), "Dummy Variable Path Coefficients," Political Methodology, 9:239-248.

Karlsson, M. G., M. P. Pritts and R. D. Heins (1988), "Path Analysis of Chrysanthemum Growth & Development," Hort Science, 23:372-375.

Kempthorne, O. (1973), An Introduction to Genetic Statistics, Iowa State Press.

Li, C. C. (1977), Path Analysis -- a Primer, Boxwood Press.

Scharen, A. L., R. E. Lund and M. E. Dietz-Holmes (1991), "Analysis of Factors that Influence Kernel Weight of Wheat Infected by Septoria nodorum," Plant Breeding, 106:242-249.

Searle, S. R. (1987), Linear Models for Unbalanced Data. John Wiley & Sons, New York.

Table 1. Relative Kernel Weight (RKW) and Symptoms[1] for Selected Entries in the USDA Septoria Nursery for the Years 1982-1987[2]

	No. Years[3]	Bozeman MT, USA Rel RKW	Symp	Griffin GA, USA Rel RKW[4]	Symp	Grunbach Germany Rel RKW	Symp	Kitzeberg Germany Rel RKW	Symp	Vienna Austria Rel RKW	Symp	Rennes France Rel RKW	Symp	Zagreb Yugoslavia Rel RKW	Symp
Winter Wheat															
Yamhill	6	97	65	86	53	75	93	94	97	42	98	81	85	78	85
JCR11	4	84	54	89	65	73	95	88	98	37	87	70	85	84	84
JCR493	4	98	55	82	63	85	94	94	95	56	86	73	93	73	89
JCR655	3	103	64	107	64	80	93	97	96	54	87	90	93	95	81
Hill Sn	3	97	76	90	64	79	94	92	96	55	87	84	93	--	--
DX 365	4	87	76	100	88	88	98	86	98	58	98	84	96	74	87
Spring Wheat															
Fortuna	6	98	75	55	77	79	96	95	75	56	88	80	88	72	99
Espige Gr. I	3	103	75	85	89	73	86	79	54	63	92	78	84	80	97
IsIx-150	3	103	86	--	--	83	84	90	54	42	87	86	67	78	99
IsIx-193	3	100	84	--	--	89	84	94	65	52	76	87	65	83	99

[1]Symptoms are presented as the 00-99 modification of the Saari-Prescott scale(1), where the first digit represents the vertical disease progress, and the second digit is the percentage of total leaf area infected by Septoria nodorum.
[2]Numbers given are derived from means of the actual readings for each entry, at each location, each year.
[3]Number of years cultivar was included in the nursery.
[4]Standard error for RKW varies from 4.1 to 5.8 depending upon number of years cultivar was included in nursery.

Table 2. Direct and Indirect Effects of Environmental and Genetic Factors on Relative Kernel Weight of Wheat Infected with Septoria nodorum.

	Direct	Indirect	Total
Location	0.45	0.30	0.75
Year in Location	0.33	0.20	0.53
Cultivar	0.22	0.03	0.25

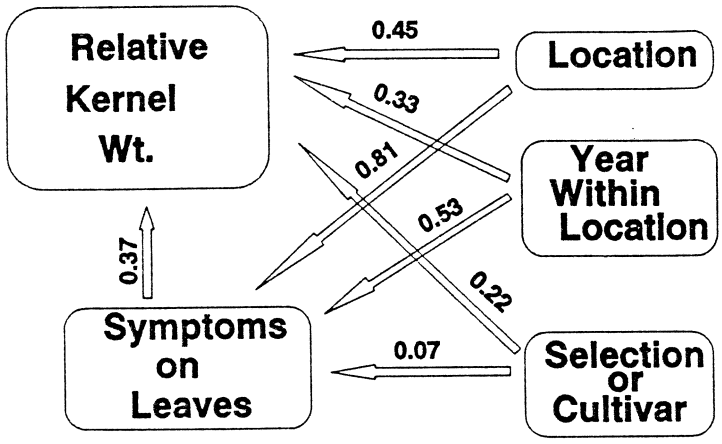

Figure 1. Path analysis of factors influencing kernal weight of wheat infected by *Septoria nodorum*.