DOCUMENT RESUME

ED 380 490 TM 022 850

AUTHOR	Yoon, Jina
TITLE	Evaluating Replicability of Regression Results Using the Jackknife Statistic.
PUB DATE	Jan 95
NOTE	17p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 26, 1995).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Evaluation Methods; *Predictor Variables; *Regression (Statistics); *Research Methodology; Statistical Analysis; *Statistical Significance
IDENTIFIERS	Confidence Intervals (Statistics); *Jackknifing Technique; *Research Replication; T Test

ABSTRACT

Contrary to popular opinion, significance testing does not inform the researcher of the likelihood of the replication of results from current research findings. Result replicability has been ignored by researchers because of an overreliance on significance testing. Several alternatives have been offered to provide the researcher with more information than the limited contribution of significance testing. One such method employed to determine the stability of results within different subtests of the existing data is the "jackknife." Using a hypothetical data set of 15 cases and 2 predictor variables, the jackknife technique is applied to the interpretation of regression results. The jackknifed coefficients are computed to evaluate the stability of beta weights and the R-squared value. In addition, confidence intervals and t-statistics are calculated to facilitate the interpretation of the jackknifed coefficients. (Contains 17 references and 5 tables.) (Author/SLD)

Evaluating Replicability of Regression Results Using the Jackknife Statistic

JINA YOON
Texas A&M University
Department of Educational Psychology
College Station, TX 77843-4225

(Running Head: Jackknife Technique)

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, January 26, 1995.

Abstract

Contrary to popular opinion, significance testing does not inform the researcher of the likelihood of the replication of results from current research findings. Result replicability has been ignored by researchers because of an over-reliance on significance testing. Several alternatives have been offered to provide the researcher with more information than the limited significance testing's contribution. One such method employed to determine the stability of results within different subsets of the existing data set is the "Jackknife."

Using a hypothetical data set of 15 cases and two predictor variables, the jackknife technique is applied to the interpretation of regression results. The jackknifed coefficients are computed to evaluate the stability of beta weights and the R squared value. In addition, confidence intervals and t-statistics are calculated to facilitate the interpretation of the jackknifed coefficients.

Science has contributed to the accumulation of knowledge, expanding intellectual boundaries. Based on previous studies, a researcher formulates hypotheses and designs a study to support these hypotheses. The findings are then reported in the field and become a foundation for future studies. Scientific method encourages researchers to prove or refute a theory through empirical results. However, research findings have limited value if the results can not be replicated in future research. Carver (1987) emphasizes that "Replication is the cornerstone of science" (p. 392). Consistent results from replication strengthens confidence in the hypothesis and in the theory from which the hypothesis was derived (Borg & Gall, 1989). However, even though the study is carefully designed and carried out, lack of replicability indicates that conclusions are based on sample specific results and are not probably generalizable to future studies, thus making little contribution to the existing knowledge (Thompson, 1994).

Daniel (1989) states that "there is always the possibility that ... results may simply capitalize on artifacts of the sample employed" (p.1). Taylor (1991) elaborates more:

> "Artifacts of the sample" include such features as outliers and the chance selection of an atypical sample which differs substantially from the population. Characteristics such as the one just mentioned lead to biased results and hence to the reporting of inaccurate conclusions. Compounding the problem, the smaller the sample size is, the greater is the risk of sample specific results. (p. 10)

In the literature, "result replicability," "generalizability," "sample specificity," and "invariance testing" are interchangeably used to refer to the likelihood of obtaining the same research results in future research (Taylor, 1991). Unfortunately, far too few researchers have paid attention to replicability issues (Cohen, 1994; Thompson 1993). Carver (1978) explains why replicability has not been seriously considered by researchers:

> Too often statistical significance is substituted for actual replicative evidence; too often statistical significance covers up an inferior research design. Nothing in the

> logic of statistics allows a statistically significant result to be interpreted as directly reflecting the probability that the result can be replicated. It is a fantasy to hold that statistical significance reflects the degree of confidence in the replicability or reliability of results. (p.386)

Thompson (1989) argues that the statistical significance, result importance, and replicability are somewhat distinct subjects that require the researcher's special attention and that these three issues are not answered by only testing for statistical significance. Statistical significance testing yields p(calculated), i.e., the probability of obtaining these particular sample statistics with the given sample size, given the null hypothesis is true. Based on the p(calculated), researchers reject or do not reject the null hypothesis. However, researchers often interpret the p(calculated) as a probability that the results are replicable or reliable. In an effort to inform researchers of limitations of statistical significance testing, many important explanations (Carver, 1978; Huberty, 1987; Thompson, 1989) have been provided with suggestions to overcome the prevalent misuse of statistical significance testing. Using an example data set with varying sample sizes, Thompson (1989) demonstrates that statistical significance testing is primarily a function of sample size. As the sample size increases, statistically nonsignificant results become significant. That is, a small mean difference can be statistically significant with a large enough sample size, leading the researcher to reject the null hypothesis.

Thompson (1989) also argues that statistically significant results do not necessarily indicate the importance of the results. He emphasizes that investigating effect sizes allows researchers to examine the results' importance. However, as Thompson (1989) points out, neither statistical significance testing nor effect sizes inform researchers of the replicability of results. One way to examine the likelihood that results will be replicated in future research is to repeat the study with a new sample using the same or similar methods. Due to time constraints and limited energy, this approach has not been favored by researchers.

There are three methods available to researchers to examine replicability without implementing the same study with a new sample. These are a jackknife method developed by Tukey (1958), a bootstrap method developed by Efron (1983), and a cross-validation method, illustrated by Thompson (1989). These methods are internal replicability techniques, using the existing sample data to estimate result replicability. The present paper applies the jackknife technique, examining replicability of multiple regression results. A brief explanation of the jackknife technique will be offered. In addition, a jackknife computation will be demonstrated with a small hypothetical data set, followed by regression analysis.

Jackknife Statistic

The jackknife statistic was developed by Tukey based on research by Quenouille and Jones as a measure of replicability (Fenwick, 1979). According to Miller (1964), Tukey named this method "jackknife" because of its versatile usage, like a scout's jackknife. Crask and Perreault (1977) describe the jackknife as "partitioning out the impact of effect of a particular subset of data on an estimate derived from the total sample" (p. 61).

The jackknife procedure involves omitting one observation (or a subset of observations of a fixed size) from the original data set and recalculating the original statistical estimator (e.g., beta weights and multiple R squared). Each observation (or a subset of observations) is omitted in turn and the statistical estimator is calculated with the truncated data set. The next step involves the calculation of "pseudovalues" (Quenouille, 1956) and the jackknifed estimator, which is the average of the pseudovalues. Daniel (1987) provides the following procedure for computing the jackknife estimator:

> A given sample of size N is partitioned into k subsets of size M (kM=N). All subsets must be of the same size (M) and may be as small as one case or as large as the largest multiplicative factor of N. A predictive estimator (e.g., a discriminant function [or canonical function] coefficient), designated as theta-prime (0′) is then computed using all k of the subsamples from the original sample of size N. The same

estimator is also computed with the i subset (i=1 to k) omitted from the sample. This estimator is designated as 0i´. This procedure is repeated k times with a different subset omitted each time. Before computing the jackknifed estimator, weighted combinations of the 0´ and 0i´ values are computed. These weighted values are called pseudovalues and are designated by the letter J. The pseudovalues are computed using the equation:

(1) $Ji\ (0') = \underline{k}\ 0' - (\underline{k}-1)\ 0i'$

where $i=1, 2, 3, \ldots, \underline{k}$.

The average of the pseudovalues is the jackknifed estimator:

(2) $J\ (0') = [\text{Sum } Ji\ (0')] / \underline{k}$

where $i= 1, 2, 3, \ldots, \underline{k}$. (p.10)

Next, the jackknifed estimator is interpreted. According to Tukey (1958) and Crask and Perreault (1977), the jackknifed estimator is normally distributed. Based on this postulation, the stability of the jackknifed estimator is evaluated as the confidence interval about the estimator. When the jackknifed estimator lies within the confidence interval, it is considered stable. Alternatively, a t-statistic can be calculated by dividing the jackknifed estimator by the standard error of the mean for the pseudovalues and determine whether the calculated t-value is greater than the critical t-value with degrees of freedom.

The jackknife technique has been reported to have advantages over other internal replicability methods. First, Taylor (1991) states that the jackknife is especially appropriate when the sample size is small. Cross-validation technique splits the sample into two groups and then the prediction equations for the one group is used for the other group's prediction. This technique reduces sample size by arbitrarily dividing the data. Daniel (1989) argues that this is especially problematic if the original sample size is small. The second advantage is illustrated by Crask and Perreault (1977):

The jackknife statistic is a general method for reducing the bias in an estimator while providing a measure of the variance of the resulting estimator by sample reuse. The result of the procedure is an unbiased, or nearly unbiased, estimator and its associated approximate confidence interval. (p. 61)

Tucker and Daniel (1992) describe a third advantage of the jackknife technique. It allows researchers to estimate changes in sampling error that may result from a single observation's uniqueness. In addition, by omitting one observation at a time, one can see the impact of any outliers on analysis.

Regression Analysis Results

Thompson (1992) reports that there are two basic applications for regression analysis. One focuses on obtaining accurate mathematical formula for prediction of the dependent variable while the other focuses on explaining the way that prediction works. The regression analysis yields various coefficients. Beta weight and multiple R squared are usually interpreted. Beta weights inform researchers of how much credit is given to a particular variable for predicting the dependent variable values while multiple R squared informs the researcher of what percentage of the variance in the dependent variables is explained by the variance of predictor variables.

However, beta weights are influenced by the collinearity among predictor variables. Using a heuristic data set, Thompson and Borrello (1985) demonstrated that when predictor variables are correlated, only interpreting beta weights leads to an inaccurate estimation of a variable's predictive power. They suggest that structure coefficients do not fluctuate with the correlation between predictor variables and are a more accurate indicator of the predictive power of a predictor variable.

For the present study, a hypothetical data set with 15 cases and two predictor variables, Y1 and Y2, was analyzed using the SPSS regression procedure. The data set and the correlation matrix are provided in Tables 1 and 2. The regression results from the SPSS program are reported in Table 3.

Insert Tables 1, 2 and 3 about here

The results indicates that about 96% of the variance of dependent variable is accounted for by the variance of the two predictor variables, X1 and X2. Beta weights for variable X1 and X2 are -1.0877 and -.32686. Squared Structure Coefficients indicate that the proportion of the YHat explained by the predictor X1 and X2 is about 96% and 4%, suggesting variable X1's strong predictive power.

An Application of the Jackknife Technique

A jackknife statistic was computed to evaluate the replicability of the multiple regression analysis described above. After the data set (n=15) was analyzed with regression procedure in the SPSS, each truncated data set was analyzed repeatedly with a sample size of 14, yielding R squared and beta weights for X1 and X2. Then, the pseudovalues and jackknifed coefficient were computed with a spreadsheet program and reported in Table 4. In order to interpret the jackknifed coefficients, a t-statistic was calculated by dividing the jackknifed coefficient by the standard error. As Table 4 indicates, the jackknifed coefficients (beta) for variable X1 and X2 are -1.0835 and -0.3841. A t calculated=-7.8062 for X1 was obtained with its absolute value exceeding the t critical value of 2.145. A t calculated= -1.6255 for X2 was also obtained with its absolute value failing to exceed 2.145. These results indicate that variable X1's beta weight (-1.0877) is stable and other researchers are likely to obtain a similar beta weight in future studies. Meanwhile, variable X2's beta weight (-0.3269) is sample specific and will not generalizable to the population. Considering that variable X2 has an extreme outlier (case 15), the beta weight for variable X2 with different samples will not be consistent.

Insert Table 4 about here

Table 4 provides the jackknifed coefficient for R squared=0.9674 with standard error of 0.0233. A t calculated =41.5193 exceeds the t critical value of 2.145, indicating its generalizability. In these sample data, variable X1 has most of the predictive power with an effect size of about 0.88 while variable X2 has an effect size of about 0.8. It appears that the replicability of R squared was primarily influenced by the stability of the beta weight on variable X1. In other words, despite of lack of replicability of variable X2's beta weight, the predictive strength of X1 and X2 is still quite strong in the jackknifed analyses because the one variable, X1, explains almost all the variance in the dependent variable without much assistance from variable X2's predictive ability.

Additionally, Table 5 provides 95% confidence intervals for the jackknifed coefficients(J0′). For each jackknifed coefficient, a margin of error was computed by its standard error times Z critical value of 1.96. Then, the confidence interval is J0′ ± margin of error. In all cases, the original coefficient lies between the confidence interval constructed.

Insert Table 5 about here

Summary

Despite its importance, the generalizability of research results has been ignored by many researchers. Carver (1978) argues that the misunderstanding of significance testing primarily has led to this problem. Evidence for replicability strengthens confidence in research results. The jackknife statistic is a technique to evaluate the replicability of a study without repeating the same study with a new sample.

The jackknife technique has advantages over other internal replicability techniques such as bootstrap and cross-validation: its appropriateness with small sample, its strength as an unbiased estimator, and its ability to estimate changes in sampling error. However, Taylor (1991) warns that parameters for interpreting invariance coefficients such as jackknifed coefficients have not been established and as Fish (1986) suggests, the interpretation of invariance results calls for the researcher's judgment.

References

Borg, W. K. & Gall, M. D. (1989). Educational research: An introduction (5th ed.). New York: Longman.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Crask, M. R., & Perreault, W. D., Jr. (1977). Validation of discriminant analysis in marketing research. Journal of Marketing Research, 14, 60-68.

Daniel, L. G. (1989, January). Use of the jackknife statistic to establish the external validity of discriminant analysis results. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. 305 382)

Efron, B. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 37, 36-48.

Frenwick, I. (1979). Techniques in market measurement: The jackknife. Journal of Marketing Research, 16, 410-414.

Miller, R. T., Jr. (1964). A trustworthy jackknife. The Annals of Mathematical Statistics, 35, 1594-1605.

Quenouille, M. H. (1956). Notes on bias in estimation. Biometrika, 43, 353-360.

Taylor, D. L. (1991, January). Evaluating the sample specificity of discriminant analysis results using the Jackknife statistic. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. 328 574)

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Thompson, B. (1992). Interpreting regression results: beta weights and structure coefficients are both important. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 344 897)

Thompson, B. (Ed.). (1993). Special issue on statistical significance testing, with comments from various journal editors, Journal of Experimental Education, 61(4).

Thompson, B. (1994). The concept of statistical significance testing (An ERIC/AE Clearinghouse Digest #EDO-TM-94-1). Measurement Update, 4(1), 5-6. (ERIC Document Reproduction Service No. ED 366 654)

Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. Educational and psychological Measurement, 45, 203-209.

Tucker, M. L., & Daniel, L. G., Jr. (1992, January). Investigating result stability of canonical function equations with the jackknife technique. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. 305 382)

Tukey, J. W. (1958). Bias and confidence in not-quite large sample. Annals of Mathematical Statistics, 29, 614.

Table 1
Hypothetical Data Set

Case	ID	Y	X1	X2
1	1	78	12	10
2	2	56	30	5
3	3	55	30	7
4	4	49	35	3
5	5	54	30	7
6	6	60	25	1
7	7	65	23	5
8	8	55	28	7
9	9	75	20	1
10	10	80	13	2
11	11	63	24	5
12	12	50	33	10
13	13	55	29	3
14	14	62	25	6
15	15	70	11	45

Table 2
Correlation Coefficients

	Y	X1	X2
Y	1.000		
X1	-.9396**	1.0000	
X2	.1657	-.4529	1.0000

** Statistically significant at level of .01

Table 3

Mulitple Regression Results

Multiple R	.98378	Analysis of Variance			
R Square	.96782		df	Sum of Squares	Mean Square
Adjusted R Square	.96246	Regression	2	1341.78439	670.89219
Standard Error	1.92820	Residual	12	44.61561	3.71797

Variables in the equation

Variable	B	SE B	Beta	T	Sig.T
X1	-1.432065	.076476	-1.087654	-18.726	.0000
X2	-.304790	.054163	-.326855	-5.627	.0001
(Constant)	99.310684	2.159739		45.983	.0000

Table 4
Pseudovalues, Jackknifed coefficient, and t-values

Case Omitted	Beta weights Pseudovalues for X1	Pseudovalues for X2	Pseudovalues for R squared
None	-1.0877	-0.3269	0.9678
1	-0.8619	+0.3308	1.0617
2	-1.0982	-0.2674	0.9639
3	-1.0450	-0.2151	0.9773
4	-1.1293	-0.0226	1.0255
5	-1.0531	-0.2242	0.9895
6	-0.9685	-0.1225	0.8534
7	-1.0578	-0.3281	0.9711
8	-1.0564	-0.2825	0.9352
9	-0.8348	-0.7369	0.7371
10	-0.2800	+0.3658	1.1226
11	-1.0712	-0.3196	0.9664
12	-0.7966	+0.0522	1.0081
13	-1.1184	-0.2310	0.9443
14	-1.0902	-0.3378	0.9667
15	-2.7907	-3.4226	0.9882
Jackknifed Coefficient	-1.0835	-0.3841	0.9674
Standard error of mean	0.1388	0.2363	0.0233
t calculated (df=14)	-7.8062*	-1.6255	41.5193*
t critical (p=.05)	2.145	2.145	2.145

* indicates coefficient stability

Table 5
95% Confidence Intervals for
the Jackknifed Coefficients

	X1	X2	R squared
Original Coefficient	-1.0877*	-0.3269*	0.9678*
Jackknifed Coefficient	-1.0835	-0.3841	0.9674
Lower	-1.3555	-0.8472	0.9217
Upper	-0.8115	0.079	1.0131

* indicates that a coefficient is within the 95% confidence interval.