Psychology Research and Applications, Vol. 1, No. 2, June 2019
https://dx.doi.org/10.22606/pra.2019.12001

Can the Null Hypothesis Be "Proven" by Using Large Numbers of Predictor Variables?

Michelle Perez, Walter R. Schumm, Abdullah AlRashed, Duane W. Crawford

School of Family Studies and Human Services, College of Health and Human Sciences, Kansas State University, USA
Email: schumm@ksu.edu

Abstract. Multivariate techniques have become commonplace in the social sciences. Some scholars have attempted to "prove" the null hypothesis by predicting one variable from a host of others. At least three risks or dangers accompany such attempts. First, the addition of multiple predictors may eventually render the apparent significance of any one independent variable to non-significance even if the underlying effect size remains medium to large. Second, the use of multiple predictors implies the potential for hundreds to millions of alternative models that could have been tested along with the models reported in an article. Third, ruling out the direct effects of a variable does not necessarily prove that the variable has no indirect effects on the outcome(s). Using a small data set of survey research in which truck drivers participated as respondents, we show how a theoretically and empirically strong relationship between income satisfaction and job satisfaction could be rendered non-significant (despite a medium or larger effect size) by adding enough other predictor variables, even if those additions did not make much theoretical sense. Thus, it is possible that other misinterpretations of the null hypothesis have occurred in the scientific literature when models with dozens of predictor variables have been used by authors who have had a goal of "proving" the null hypothesis.

Keywords: Truck drivers, income satisfaction, job satisfaction, methodology, indirect effects

1 Introduction

Scholars have known for a long time that multivariate analyses, including ordinary least squares regression analysis, have numerous limitations (Schumm, 1982; Schumm, Southerly, & Figley, 1980). However, in recent years, one limitation has continued to appear through attempts by both progressive and conservative scholars to "prove the null hypothesis" (Schumm, 2010). Technically, the null hypothesis cannot be proven; it can only be rejected or accepted, because further research might yield results in which the null was rejected. However true that statement might be (Herek, 2006, p. 610), many scholars have essentially disregarded it (Schumm, 2018). A statistically non-significant result may reflect the truth of a null hypothesis but it may also reflect a small sample size, low statistical power, poor measurement, or other methodological limitations (Aczel et al., 2018).

1.1 How the Null Hypothesis Is Often "Proven"

The usual process is that someone states a hypothesis that variable X is not related to variable Y. Upon initial multivariate analysis, it appears that variable X is significantly related to variable Y. Since the researcher's goal is to find a non-significant relationship between X and Y, a tempting option is to continue to add other independent variables to the regression or other multivariate models until X is no longer significant statistically and then declare "victory" or "proof" that indeed X is not related to Y, as you had expected.

1.2 Examples of Attempts to "Prove" the Null Hypothesis

For example, Calzo et al. (2017) started with a model in which one outcome variable was predicted by parental sexual orientation (lesbian or gay, bisexual, and heterosexual, with the latter as the reference)

and one other independent variable. Having a bisexual parent was related to a higher level of a child's emotional and mental health difficulties. After they tested three more models and added another ten independent variables, neither of the parental sexual orientation variables were significant, even though the regression coefficients remained at moderate effect sizes (b = .21 and .26).

Longbein and Yost (2009) with a sample of N = 141 to 153, used as many as 63 or more independent variables from which they concluded that state level approval of same-sex marriage had no effects on their outcome measures. Rosenfeld (2010) predicted educational progress as a function of more than 70 independent variables, including same-sex cohabiting parental status, finding non-significant results, even though the regression result suggested at least a small size negative association.

1.3 Researcher Bias

Bakker and Wicherts (2011) found that when errors were made in statistical reporting, the errors usually favored the researcher's expectations or biases; thus, if the initial goal is to "prove" the null, one might expect that such would eventually be found if the researchers were to torture their data long enough. Often, data are not available for other scholars to check and see if a different combination of predictor variables might have left X as significantly (or not) related to Y. Therefore, we wanted to demonstrate with a small data set (available to others on request) how one could start with a regression equation in which a predictor variable X was significantly related to Y, with sound theoretical support, and if one were to add enough other predictor variables to the model, one could "prove" that X was not related to Y, at least not significantly, even if the effect size was in the medium to large range.

1.4 Hypotheses

Our theory was that (1) truck driver income should predict truck driver satisfaction with income and that (2) truck driver satisfaction with income should predict truck driver satisfaction with their quality of life as a trucker (essentially their job satisfaction). We also expected to find a significant indirect effect from income to satisfaction with quality of life with truck driver satisfaction with income as a mediating variable. We expected these relationships to be not only statistically significant but to be substantial in magnitude with medium to large effect sizes.

2 Methods

2.1 Sample

In the late 1990's, the senior author, after obtaining approval from the university IRB board (project number 1651, 1998), used non-random sampling at truckstops on major highways in the Midwestern United States to collect surveys from over-the-road truck drivers. Of the 82 surveys handed out, 58 (70.7%) were returned. Thus, the sample included 58 participants, of whom 15.5% were women, 84.5% were long-haul drivers, 72.4% were single drivers (versus team drivers), 86.2% were company drivers, spending up to 29 days a month away from home (mean = 20.5, SD = 7.3), 15.5% drove with a pet cat or dog, 48.3% had a high school education (6.9% less, 24.1% more), with ages from 25 to 64 (mean = 43.1, SD = 9.4), 86.2% White, of whom 25.9% were in their first marriage, 25.9% were divorced, and 32.7% were remarried after the death of their spouse or a divorce. Their number of children ranged between none and 8, with mean = 2.14 (SD = 1.4). Most had been married only once (51.5%) or twice (24.1%) or three or more times (17.2%)(1.7% never married). Most of them were driving 2,000 to 2,999 miles a week (31.0%) or 3,000 to 3,999 miles a week (44.8%) with a few 4,000 or more miles (10.3%) and a few under 2,000 miles a week (8.6%). The number of days off driving ranged up to 20 a month (mean = 6.2, SD = 3.5). Total driving income ranged between less than $10,000 (1.7%) to $50,000 or greater (8.6%), with most earning between $20,000 and 39,999 (58.6%). Of those 37 drivers currently married, duration of marriage ranged from one to 46 years (mean = 16.6, SD = 11.4).

2.2 Measures

In addition to demographic measures discussed above, the survey included several Likert-type measures of satisfaction, including sexual satisfaction, income satisfaction, quality of life satisfaction, driver quality of life, marital satisfaction, satisfaction with relationship with spouse, satisfaction with spouse as a partner, satisfaction with children, satisfaction with relationship with children, satisfaction as a parent, satisfaction with parenting difficulties while on the road, satisfaction with parenting difficulties while at home, satisfaction with quality of communication while on the road, and satisfaction with quality of communication while at home with responses including very satisfied, moderately satisfied, satisfied, moderately unsatisfied, and very unsatisfied. Others items included how often the driver felt lonely on the road, with responses of extremely often, very often, often, not very often, not extremely often, and "I do not feel lonely"; driver level of stress, with responses of very high level of stress, high level of stress, moderate level of stress, low level of stress, and very low level of stress; driver coping with stress, with responses of very effective, moderately effective, effective, moderately ineffective, and very ineffective. Cronbach's alpha for the three marital satisfaction items, comprising the Kansas Marital Satisfaction Scale (Perrone, Webb, Wright, Jackson, & Ksiazak, 2006) was only 0.70, with n = 38; we did not use that scale because of the substantial loss of cases. Cronbach's alpha for the three items of the Kansas Parental Satisfaction Scale (Perrone et al., 2006) was 0.84.

2.3 Analyses

We used ordinary least squares multiple regression to predict satisfaction with driver quality of life from other predictor variables, including several demographic variables and other measures of satisfaction, loneliness, stress, or coping with stress. At first our model included only three variables – satisfaction with income, driver's satisfaction with his/her quality of life, and total income. Our final model in which we predicted driver satisfaction included gender, age, attendance at religious services, number of times married, number of children, type of driver, miles driven per week, days away from home, days off, income satisfaction, driver level of stress, coping with stress, taking a pet along, feeling lonely, total income, quality of life satisfaction, sexual satisfaction, parenting difficulty on the road, parenting difficulty at home, calling home while on the road, quality of family communication on the road, quality of family communication while at home, the Kansas Parental Satisfaction Scale, and satisfaction with relationship with spouse. Using so many variables with so few subjects does, of course, violate traditional rules of thumb that there should be several cases per variable in multivariate analyses (Schumm et al., 1980).

3 Results

3.1 Bivariate Results

Income predicted income satisfaction (b = .31, p < .03, Cohen's d = .65; Cohen, 1988, p. 22) while income satisfaction predicted driver satisfaction (b = .63, p < .001, d = 1.62). Income did not predict driver satisfaction after the inclusion of income satisfaction in the regression equation (b = .01, p = .932). A Sobel test [quantpsy.org/sobel/sobel.htm] for the indirect effect of income was significant (2.19, p < .03). In the univariate model, clearly satisfaction with income as a trucker strongly and significantly predicted satisfaction with driver quality of life. It is important to note that the effect size found for income satisfaction was over twice as large as that indicated by Cohen (1992) to be deemed a "large" effect.

3.2 Multivariate Results

Would controlling for multiple predictor variables potentially render the apparent effect of income satisfaction to the null in terms of statistical significance? Adding variables in this data set also reduces the available sample size due to missing data on several of the variables. Predicting satisfaction with driver quality of life from the 24 variables of gender, age, attendance at religious services, number of times married, number of children, type of driver, miles driven per week, days away from home, days off,

income satisfaction, driver level of stress, coping with stress, taking a pet along, feeling lonely, total income, quality of life satisfaction, sexual satisfaction, parenting difficulty on the road, parenting difficulty at home, calling home while on the road, quality of family communication on the road, quality of family communication while at home, the Kansas Parental Satisfaction Scale, and satisfaction with relationship with spouse yielded an adjusted R-squared of 0.834 (n = 34) with beta = .28 ($p < .10$) for income satisfaction. Drivers who were more significantly ($p < .05$) more satisfied with their quality of life included males ($b = .37$), younger drivers ($b = .46$), those married more times ($b = .36$), who didn't take pets along ($b = .34$), those who had fewer difficulties on the road with parenting ($b = .45$), who called home more often ($b = .49$), and who had greater satisfaction with their relationship with their spouse ($b = .43$).

3.3 Limitations/Objections

One limitation of the research is the smaller sample size. It could be argued that the main reason the results for income satisfaction became non-significant was more the smaller sample size than the number of variables entered. To test that objection, we re-ran the same model with 24 predictor variables using the SPSS mean substitution option, yielding n = 58. The overall adjusted R-squared was 0.482 but the regression coefficient for income satisfaction ($b = .325$, $d = .69$) still remained non-significant ($p < .07$). None of the other predictors were statistically significant in the mean substitution model. Therefore, we would argue that the large number of predictor variables was playing at least part of a role in reducing the apparent covariance between income satisfaction and job satisfaction. It could be argued that our ratio of cases to variables was small (below 5), which is correct; however, that wasn't a barrier in the Langbein and Yost (2009) study in which the number of predictors variables was at least 65 in some analyses for samples as small as n = 141. It could also be argued that our adjusted R-squared was unacceptably large, at between 0.482 and 0.834; however, Langbein and Yost (2009, p. 301) reported R-squared values between .85 and 0.94, higher than those we obtained.

4 Discussion

Our simple model featured medium to large effect sizes with income predicting income satisfaction and income satisfaction predicting driver job satisfaction, with a significant indirect effect of income on job satisfaction, a result that common sense would anticipate. However, through a combination of reducing sample size and adding many variables to a multivariate regression model, it was possible to make an initially strong empirical relationship and a strong theoretical relationship (income satisfaction predicting job satisfaction) disappear in terms of statistical significance, even though the effect size remained medium or greater, with Cohen's $d = 0.59$. This suggests that in many cases, it may be possible to "prove" the null hypothesis by overwhelming any effect of one variable with the effects of multiple other variables. We are also concerned with approaches to multivariate analysis that don't use much theory but just predict one variable from a set of other variables whether those other variables make any theoretical sense or not. Furthermore, as we used so many variables but reported results for only one result, it should be noted that the use of any of the 24 variables represented a choice (use or not use). We had to start with at least one variable, so our choices concerned using or not using the other 23 variables. As we added variables, it was our impression that most models featured a significant result for satisfaction with income out of the $(23)^2$ or 8, 388, 608 models.

We have seen models featuring as many as 65 (Langbein and Yost, 2009) or even 77 independent variables (Rosenfeld, 2010), which would represent far more possibilities, yet those authors presented results for only a handful of models among the tens of millions possible). Rosenfeld (2010) in a footnote (#10, p. 768) noted that if you restricted the model to highly educated parents, then children of heterosexual couples made significantly more progress through school than the children of same-sex couples, indicating there was at least one model that contradicted his primary reported models. How many others might there have been that were not reported? In the same way, we presented only one complete model out of more than eight million possibilities, reaching a conclusion that would disagree with most of the other millions of models. While our analysis can be criticized for a low ratio of cases to variables (34-58/24), Langbein and Yost (2009) had at least one situation of 141 cases for 65 variables, a

similarly low ratio of cases to variables. Allen and Price (2015) critiqued Langbein and Yost (2009, 2015) for this in terms of low power, but our comments aim more at the effect of the addition of multiple variables for diminishing the apparent substantive importance of independent variables, even if the effect sizes remain of small to medium nature. If one expands the concept of models beyond one dependent variable and multiple independent variables, one might consider the possibilities for intervening variables. Along those lines, if variable X does not significantly predict variable Y after controlling for dozens of other variables, that may show that X does not have a direct effect on Y under specific conditions, but it does not prove that X has no indirect effects through other variables on Y. In our case, income might not have a direct effect on job satisfaction, but it might well have an indirect effect on job satisfaction operating through the intervening or mediating variable of income satisfaction. For example, Calzo, Mays, Bjorkenstam, Bjorkenstam, Kosidou, and Cochran (2017) found that children with bisexual parents reported significantly elevated rates of emotional and mental health difficulties until their final model, with the most (k = 13) independent variables, was displayed. Adding parental distress to the model seemed to be the variable that "did in" the apparent effect of parental sexual orientation. What may have been overlooked is that parental sexual orientation might have had an indirect effect on the child outcome through parental distress as a mediating variable. Furthermore, the effect sizes reported for lesbian or gay ($d = 0.43$) or bisexual ($d = 0.54$) parent sexual orientation were in the small to medium range, regardless of their statistical non-significance. Thus, these issues appear to be involved in older and more recently published articles – in other words, they have not gone away no matter how often they may have been criticized.

Aczel et al. (2018) surveyed null findings in three psychology journals for 2015 and found that 72% of the cases involving findings in favor of the null hypothesis did not in fact permit inference that the effect was absent; fewer than 5% of the null findings provided strong evidence that the null hypothesis was correct. In other words, our example may not be an atypical case in which a null finding seems statistically correct but is probably not accurate. When researchers use large numbers of independent variables (see Schumm, 2018, for other examples) how are other scholars to know that the final model presented actually represents the most typical result of the many possible other models? If the researchers have not tested each of their millions of possibilities (unlikely), then they should share the process by which they initiated and expanded their model to include as many variables as they eventually used and for which they reported final results. Researchers should also report the effect sizes for their results, even when the results do not appear to be significant statistically. Making data available for other researchers to allow for independent replications would also be helpful.

References

1. Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B.,,,,,,Wagenmakers, E-J. (2018). Quantifying support for the null hypothesis in psychology: an empirical investigation. *Advances in Methods and Practices in Psychological Science*, 1(3), 357-366.
2. Allen, D. W., & Price, J. (2015). Same-sex marriage and negative externalities: A critique, replication, correction of Langbein and Yost. *Econ Journal Watch*, *12*, 142-160.
3. Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavioral Research*, *43*, 666-678.
4. Calzo, J. P., Mays, V. M., Bjorkenstam, C., Bjorkenstam, E., Kosidou, K., & Cochran, S. D. (2017). Parental sexual orientation and children's psychological well-being: 2013-2015 National Health Interview Survey. *Child Development*, online advance.
5. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.
6. Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.
7. Herek, G. M. (2006). Legal recognition of same-sex relationships in the United States: a social science perspective. *American Psychologist*, *61*, 607-621.
8. Langbein, L., & Yost, M. A. (2009). Same-sex marriage and negative externalities. *Social Science Quarterly*, *90*, 292-308.
9. Langbein, L., & Yost, M. A., Jr. (2015). Still no evidence of negative outcomes from same-sex marriage. *Econ Journal Watch*, *12*, 161-163.

10. Perrone, K. M., Webb, L. K., Wright, S. L., Jackson, Z. V., & Ksiazak, T. M. (2006). Relationship of spirituality to work and family roles and life satisfaction among gifted adults. *Journal of Mental Health Counseling*, *28*, 253-268.
11. Rosenfeld, M. J. (2010). Nontraditional families and childhood progress through school. *Demography*, *47*, 755-775.
12. Schumm, W. R. (1982). Integrating theory, measurement, and data analysis in family studies survey research. *Journal of Marriage and the Family*, *44*, 983-998.
13. Schumm, W. R. (2010). Statistical requirements for properly investigating a null hypothesis. *Psychological Reports*, *107*, 953-971.
14. Schumm, W. R. (2018). *Same-sex parenting research: a critical assessment*. London: Wilberforce Press.
15. Schumm, W. R., Southerly, W. T., & Figley, C. R. (1980). Stumbling block or stepping stone: path analysis in family studies. *Journal of Marriage and the Family*, *42*, 251-262.

