Relation between Population RMSEA and *p*-Value: 2-Dimensional Approach in Assessing and Reporting Goodness of Fit in Structural Equation Modeling

Xiaoxu Li The Chinese University of Hong Kong
Kit-Tai Hau The Chinese University of Hong Kong
Herbert W. Marsh Oxford University

Abstract

In describing fit of models in structural equation modeling, besides using various goodness of fit indexes, one can choose to report the upper-bound of RMSEA confidence interval (e.g., .043) at a commonly used probability criterion (e.g., $p < .05$). Alternatively one can report the *p*-value using a popular RMSEA close-fit criterion (e.g., .05). In essence both approaches summarize the RMSEA-*p* curves into some quantities, but with reference to different common standards ($p \leq .05$, RMSEA $\leq .05$ respectively). SAS CNONCT/PROBCHI codes are provided to search for boundary conditions in which the two approaches may provide different information and conclusion. We differentiate the implicit and explicit roles of substantial knowledge against statistical conclusion and recommend using RMSEA-*p* curves.

Relation between Population RMSEA and *p*-Value: 2-Dimensional Approach in Assessing and Reporting Goodness of Fit in Structural Equation Modeling

In essence, structural equation modeling (SEM) is statistically testing the closeness of a multivariate population to a certain targeted constrained model subspace. Among various fit indexes, the Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993; Steiger & Lind, 1980) is one of the favorite statistics used by researchers to provide an intuitive summary of the discrepancy between the population and the model subspace by rescaling them to a comparable metric (Li, Hau & Marsh, 2006). The present study compares the use of two statistical reporting approaches related to RMSEA, which are respectively made on the basis of (i) the upper-bound of RMSEA at a fixed significance level (e.g., .05) (reporting RMSEA$_{\text{upper-bound}}$ approach) and (ii) the *p*-value of a fixed RMSEA level (e.g., close fit .05) (reporting *p*-value approach). We also alert researchers the interrelatedness of these two approaches, and recommend better utilization of the RMSEA-*p* curve in model fit assessment with particular attention to the underlying substantial knowledge.

Statistical Conclusion and Inherent Error Probability

The chi-square sample statistics T_{sample} with the non-central parameter λ (Noncentral Chi-square Distribution, 2006) are related in the following equations. At a given sample size N=n+1 and degree of freedom *df,*

$$T_{sample} \equiv n \times \mathrm{F}_{\text{Discrepancy}}(\text{Model Subspace}, S_{\text{sample-cov}}) \xrightarrow{L} \chi^2_{\text{noncentral}}(df, \lambda)$$

$$\text{wherein } \lambda \equiv n \times \mathrm{F}_{\text{Discrepancy}}(\text{Model Subspace}, \Sigma_{\text{population-cov}}) = df \times n \times \text{RMSEA}^2$$

Consider a simple case with just two discrete plausible populations, H_0 ($\lambda = \lambda_0$, a better fit model) vs. H_1 ($\lambda = \lambda_1 > \lambda_0$, a worse fit model) with a criterion chi-square statistics T_C as follows:

$$\mathrm{P}(T_{sample} < T_C \mid H_0) \approx \mathrm{P}(T < T_C \mid T \sim \chi^2_{\text{noncentral}}(df, \lambda_0)) = 1 - \alpha$$
$$\mathrm{P}(T_{sample} \geq T_C \mid H_0) \approx \mathrm{P}(T \geq T_C \mid T \sim \chi^2_{\text{noncentral}}(df, \lambda_0)) = \alpha = \text{Asymptotic Type I Error}$$
$$\mathrm{P}(T_{sample} < T_C \mid H_1) \approx \mathrm{P}(T < T_C \mid T \sim \chi^2_{\text{noncentral}}(df, \lambda_1)) = 1 - \pi = \text{Asymptotic Type II Error}$$
$$\mathrm{P}(T_{sample} \geq T_C \mid H_1) \approx \mathrm{P}(T \geq T_C \mid T \sim \chi^2_{\text{noncentral}}(df, \lambda_1)) = \pi = \text{Asymptotic Power}$$

According to the classic statistical testing theory, it is meaningless to report the rejection of H_0 at $T_{\text{sample}} \geq T_C$ without mentioning the α value. Symmetrically, no statistical conclusion can be drawn for $T_{\text{sample}} < T_C$ and consequentially accepting H_0 if 1-π is unknown. The conclusion "H_0 is retained / not rejected" implies that it is substantial knowledge that helps the decision to accept or retain the model while statistical information contributes little to the decision process.

However, when $T_{\text{sample}} < T_C$, a known and reported 1-π value and no matter how imperfect it seems (but less than .5, e.g., .2) (Hancock, 2006), it provides an explicit statistic conclusion to reject H_1 at an error probability not greater than 1-π. Noteworthily, how small the α or 1-π should be set depends on substantial knowledge. We will discuss its role in various kinds of decision-making process in the discussion section below.

Reporting RMSEA$_{upper\text{-}bound}$ vs. Reporting p-Value

In reality, the plausible population RMSEA is constituted of a range in a continuum interval rather than taking two alternative discrete values. There is a monotonous relation between p and RMSEA at an observed T_{sample}, a known df, and a specific n as follows:

$$p \equiv p_{left} = \mathrm{P}(T < T_{sample} \mid T \sim \chi^2_{\text{noncentral}}(df, df \times n \times \text{RMSEA}^2))$$

The smaller the RMSEA value is, the larger p will become. MacCallum, Browne and Sugawara (1996) have highlighted how the RMSEA confidence interval (CI) can be utilized in hypothesis testing.

Browne and Cudeck (1993) labeled RMSEA ≤ .05 as "close fit", .05 < RMSEA ≤ .08 as "fair fit", RMSEA > .1 as "poor fit" while MacCallum, Browne and Sugawara (1996) suggested .08 < RMSEA ≤ .1 as "mediocre fit". Thus, when researchers would like to report RMSEA$_{\text{upper-bound}}$ = .105, they might have difficulty because this specific value has not been a criterion reported in the literature. However, it is totally legitimate to take an alternative approach and compromise by using a slightly different error probability criterion (e.g., $p \leq .055$). For a certain RMSEA criterion (e.g., .105), it is possible and not difficult to calculate the corresponding p-value for empirical results obtained in a study.

In this study, we propose the use of RMSEA-p curve and identify cases in which this method provides more information than either the "Reporting RMSEA$_{\text{upper-bound}}$" and "Reporting p-value" approaches. We will also examine how the valuable information is potentially lost in the summarization process of these two approaches.

Method

The position and shape of the RMSEA-p curve depend on only three parameters—df, n, and noncentral chi-square sample T. It is easy to draw a RMSEA-p curve through any given point (RMSEA$_2$, p_2) by fitting an appropriate T parameter. Statistically it means a researcher could conclude on the fitness of a certain model with RMSEA$_{\text{upper-bound}}$ = RMSEA$_2$ value at the explicit error probability p_2.

Even if both RMSEA$_2$ and p_2 are a bit larger than the recommended criterion RMSEA$_1$, p_1 (p_1=.05 and RMSEA$_1$=.10; while p_2=.10, RMSEA$_2$=.105), the statistical conclusion (rejection of RMSEA>.105 at error probability as low as .10) could still be of great value to researchers. The unfortunate situation is that researchers only report a rather big RMSEA$_{\text{upper-bound}}$ on p_1=.05 or a rather high p-value on RMSEA$_1$=.10.

On the SAS platform, we vary *df* and *n* in incremental small steps to search for the cases that our proposed RMSEA-*p* curve method contributes more than the reporting “$\text{RMSEA}_{\text{upper-bound}}$“ or “*p*-value” approaches (see SAS codes in Appendix). Specifically, we increase *df* from 2 to 250 by step of 1, and *n* from 30 to 300 by step of 1, 305 to 1305 by step of 5, 1310 to 10000 by step of 20.

Result

With the Reporting $\text{RMSEA}_{\text{upper-bound}}$ approach, the problem is sometimes serious when both *df* and *n* are small. For example, when *n*=35 and *df*=2, the corresponding $\text{RMSEA}_{\text{upper-bound}}$ is .18 at $p = .05$. That means, with the $\text{RMSEA}_{\text{upper-bound}}$ approach, little information on fit can be drawn with respect to the model when it is marginally “not-poor” at an error probability as low as 1/10. In Figure 1, the 90% RMSEA CI of the curve is relatively large and the curve appears to be rather flat. So a small change in *p*-value may result in a dramatic increase in RMSEA, thus easily rendering the $\text{RMSEA}_{\text{upper-bound}}$ approach useless.

In contrast, the Reporting *p*-value approach is inferior to our proposed RMSEA-*p* curve approach when *n* is large, or when the *df* is large and *n* is reasonably not too small. For example, when n = 9370 and df = 16, or when df = 240 and n = 325, the respectively *p*-values are larger than .50 when the criterion is chosen to be RMSEA=.10. That means, the Reporting *p*-value approach provides no conclusion on the popular criterion RMSEA=.10 whereas the *p*-value of the model is actually quite low at .10 when RMSEA is .105.

In Figure 1, the 90% CI of the two curves is small and the curves appear steep between $.05 < p <.95$. Such curves may cross a popular RMSEA criterion (e.g., .10) and makes the *p*-value report approach inappropriate on this criterion.

In our specific design, our explored result proves that when the RMSEA at $p = .05$ is larger than .12, the RMSEA-*p* curve will cross at a point with RMSEA=.10 whose *p*-value is smaller than .13; and vise versa. That means, researcher find that $\text{RMSEA}_{\text{upper-bound}}$ is .105 with p<.10. They may hope to try how about the case p<.05. Surely changing *p* criterion from .10 to .05 will make $\text{RMSEA}_{\text{upper-bound}}$ report increase. If the increased change is too big, for example, from .105 to more than .12, researchers may think a change in RMSEA criterion may perform better. Then after changing RMSEA criterion from .105 to .1, researchers will see *p* value does not increase too much, i.e., lower than .13. The point (RMSEA =.105, *p*=.10) acts as a pivot around which curves like poles.

Conclusion & Discussion

Two-dimensional RMSEA-p curve Reporting Approach

In this study, we show that both the traditional reporting of RMSEA-*p* relation either along the commonly used .05 significance probability (horizontal line in Figure 1) or adopting a frequently used RMSEA critical value (e.g., not-poor fit RMSEA= .10; vertical line in Figure 1) may lose valuable statistical information. One may speculate whether we can always simplistically combine the results from both reporting approaches. As demonstrated in this study, there is no such a case that both the $\text{RMSEA}_{\text{upper-bound}}$ and *p*-value approaches significantly perform worse than the RMSEA-*p* curve approach at the same time. This does not rule out, however, the existence of exceptional cases in other situations. As always in any simulation studies, the range of differences shown in this study is specific to our particular choice of (RMSEA_2, p_2) or (RMSEA_1, p_1). Geometrically, a case will lose most information when the curve concaves downward and leftward to the anchor point (RMSEA_1, p_1). Thus, under these circumstances, both the fixing RMSEA_1 and the fixing p_1 approaches will lead to poor results. In contrast, the optimal points lie rather low under the straight line joining the *p*-value reporting point fixed at RMSEA_1 and the $\text{RMSEA}_{\text{upper-bound}}$ reporting point fixed at p_1.

Actually we recommend the reporting of the two-dimensional RMSEA-*p* curves itself, with the *p* axis rescaled to $\log_{10}(p)$, the RMSEA axis rescaled to RMSEA^2, and the commonly chosen criteria with summarizing points being highlighted. Although the Chi-square value, degree of freedom, and sample size together have sufficient information to reproduce the RMSEA-*p* curve, it will be much easier if readers are provided with these graphs to help their decision-making and statistical conclusion. With these curves, researchers could weigh and see the compromise between the error probability and goodness of fit, and then judge themselves where their optimal tradeoff point should be.

Although there may be suggestions among academic on some other criteria, these critical values are still arbitrary. Imagine that human had evolved with eight fingers and had chosen an octal rather than a decimal system. Likely we would have chosen 1/16 (.0625), rather than 1/20 (.05) as the commonly used probability criterion. The RMSEA criterion for not-poor-fit might also be 1/8 rather than 1/10. Thus, it becomes obvious that it is reasonable to report a conclusion based on $\text{RMSEA}_{\text{upper-bound}}$ at any other personal favor (e.g., RMSEA = .105) to optimize the tradeoff with the corresponding error probability.

Roles of Substantial Knowledge

We will discuss the different roles that substantial knowledge plays in modeling. Firstly, substantial knowledge imposes concrete interpretations on the targeted geometric model subspace. In general, SEM as the statistical tool just provides conclusion on the model subspace without further information on the preference of one specification of the structural equations to another which occupies an identical subspace in geometry but with a different interpretation in substantial meaning (Hershberger, 2006).

Secondly, just what our RMSEA-*p* reporting purports, substantial knowledge helps in resolving the compromise between the explicit error probability level and the goodness of model fit through scales specific to a particular application.

Thirdly, we have some concern over the general ignorance of the error probability associated with the report of any point estimate. Specifically, some researchers tend to report *df*, *n*, and the point estimates of a number of goodness-of-fit indexes and then assert the fit of the model when these point estimates are within some commonly chosen criteria. The role of substantial knowledge is ambiguous here. With a big enough sample size, researchers can easily be contented to expect a nearby *p*-value low enough for their substantial applications. This is similar to what we have discussed in the second point above. However, if the sample size is really small, a model with acceptable point estimates of goodness-of-fit indexes will still have large *p*-values on targeted criteria of respective goodness-of-fit indexes, which in general is not explicitly made known to readers (this point to be further elaborated below).

Fourthly, with a satisfactorily chosen *p,* sometimes the SEM analysis cannot explicitly reject the model to be bad (or conversely good) using the two bounds of RMSEA CI. In these cases, the substantial knowledge may play an even more important role to select and retain one of the two susceptible null hypotheses -- the not-good-fit, or the positive and often favorable not-bad-fit hypotheses. Whether the above asymmetric choice is more appropriately described as "a substantial conclusion without statistical significance" or "a statistical decision based on substantial knowledge" reflects the profound distinction in thought between Fisher school and Neyman-Pearson school (Hubbard & Armstrong, 2005). Considering most SEM researches are inference type in nature, we prefer the former.

Further Implications

Some researchers might have doubt whether the RMSEA-*p* curve approach involves a potential trap of capitalization on chance. A capitalization on chance artifact arises through an exploratory selection from a pool of estimated confirmatory models. A typical case is the unconscious reusing of one single set of data in both model respecification and final confirmation (Bollen, 1989; Ting, 1998). Specifically the exploration using the RMSEA-*p* curve in a series of confirmatory models with respective hypotheses in the population RMSEA and then selecting the most appealing one to report makes some researchers to suspect a similar capitalization on chance problem.

Even if that is the case, both model search and data reuses are not academic ethic problems if they are honestly and fully noted by the researchers. A compromise in the methods being chosen may also provide meaningful results, as long as both the researchers and readers are aware of and do not over-interpret the results. At this juncture, one may feel safer to report the full RMSEA-*p* than a single pair of *p* value and RMSEA bound alone. However, the RMSEA-*p* approach is not a model search in nature, even only one favorable point chosen from the RMSEA-*p* is reported with *df* and *n*. As all information of the series of models is available in any one in them, it is just one complete model rather than a series of selectable models that is to be explored.

Every classic statistical test takes a similar form that it is an assessment of the probability of observing within a sample domain, with some given parameters. A true capitalization on chance of model search makes the selected conclusion false because the assumption other than the involved parameters in the selected model is contingent upon the observed data, which subsequently bias the resulted probability. The point selection on the RMSEA-*p* curve does not change any assumption other than the asserted parameter RMSEA, so it is justified to say that the final *p* derived is as safe as those from the conventional approaches.

The comparison of our suggested approach to general classic statistical tests may lead to the query why such an approach has not been generally used in other statistical models involving parameter range test. One possible reason is the difference between quantitative criteria of “social science” and of “natural science”. The .05 *p* criterion originates from a historical period when statistics developed with intensive industrial applications with high precision (Dallal, 2007). While in social research with greater complexity, many researchers may look for results with more tolerance, and may wish to have criterion of .08 or even much larger. Another reason that two-dimensional approaches have not been more widely used is that in contrast to many other researches, in SEM, the RMSEA provides a scale-free measurement for model fitness across studies (Browne & Cudeck, 1993; Steiger & Lind, 1980). While in general, statistical test like *t*-test, the tested parameter is too problem-specific to be used as a quantitative criterion. So, standard *t*-test, *F*-test and χ^2-test only consider a qualitative conclusion. A third related reason is that SEM is extensively applied in researches with a wide spectrum in falsifiability. In a less falsifiable research field, the research question is more likely to be “to what extent can the population be sketched in the way”, while in a high falsifiable research field, the research question asks, “can the population be replicated in the way”. In the former situation, a tolerant RMSEA criterion could serve much better than in the latter.

We would like to make a comparison of our approach to the Nobel-winning Markowitz model (Markowitz, 1987) in finance. In that model, profit and risk are a pair of trading-off traits that always counteract each other. Any decision that only considers one aspect and not the other would not be useful. This rationale is rather universal and is not limited to the finance domain. Many quantitative researchers are used to the conventional single dimension kind of summarizing a model, just like investors had got used to a unitary assessment on asset. However, a model could be good or bad in more than one dimension, just as an investment could be desirable or not in another dimension, i. e. risk. The traditional *t*-test is fixed in the dimension that is called goodness of fit in SEM. When researchers say a *t*-test model is good or bad, they just mean the significance while the goodness of fit is always fixed with extreme precision by default. The controversy on goodness (or sometimes effect) and statistical significance generally emphasize the goodness or effect, while our two-dimensional curve approach means to a more comprehensive summary of both.

Imagine a situation where RMSEA of a model acts as an inverse measure of efficiency of some fund (or profit), *p* value would naturally reflects the corresponding risk. But, we should also be reminded of the conventional precaution on Prosecutor's fallacy (Thompson & Schumann, 1987) that *p* is not the Bayesian probability of erroneous rejection conditional upon the observed data. To control the risk through *p* depends greatly on our substantial knowledge. So, expectations parallel to Markowiz' portfolio strategies to hedge risks are probably too optimistic for the currently proposed RMSEA-*p* curve approach.

References

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley and Sons.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Dallal, G. E. (2007), Why p=0.05? in *The Little Handbook of Statistical Practice*, Retrieved March 16, 2007, from http://www.tufts.edu/gdallal/LHSP.HTM

Hancock, G. R. (2006). Power analysis in covariance structure modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course*, (pp. 69-115). Greenwich, CT: Information Age Publishing

Hershberger, S. L. (2006). The problem of equivalent structural models. In G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling: A second course*, (pp. 13-41). Greenwich, CT: Information Age Publishing.

Jöreskog, K. G., & Sörbom, D. (2003). *LISREL 8.72.* Chicago, IL: Scientific Software International.

Li, X., Hau, K. T., & Marsh, H. W. (2006). Potential uses of RMSEA "scree-plots" in structural equation modeling: Balancing parsimony and approximation. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, 7-11 April.

MacCallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130-149.

Markowitz, H. M. (1987). *Mean-variance Analysis in Portfolio Choice and Capital Markets*. New York: Basil Blackwell.

Noncentral Chi-square Distribution (2006). In *Wikipedia, the Free Encyclopedia*. Wikimedia Foundation, Inc. Retrieved July 20, 2006, from http://en.wikipedia.org/wiki/Noncentral_chi-square_distribution

Hubbard, R., & Armstrong, J. S. (2005). Why we don’t really know what “statistical significance” means: a major educational failure. Retrieved August 01, 2006, from

http://hops.wharton.upenn.edu/ideas/pdf/Armstrong/StatisticalSignificance.pdf

SAS Institute Inc. (2003). *SAS 9.1.3.* Cary, NC: SAS Institute Inc.

Steiger, J. H., & Lind, J. C. (1980). Statistically-based tests for the number of common factors. Paper presented at the May annual meeting of the Psychometric Society, Iowa City, IA, June.

Thompson, W. C. & Schumann, E. L. (1987). Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy. *Law and Human Behavior*, 11(3), 167-187.

Ting, K-f. (1998). The TETRAD approach to model respecification. *Multivariate Behavioral Research*, 33(1), 157-164.

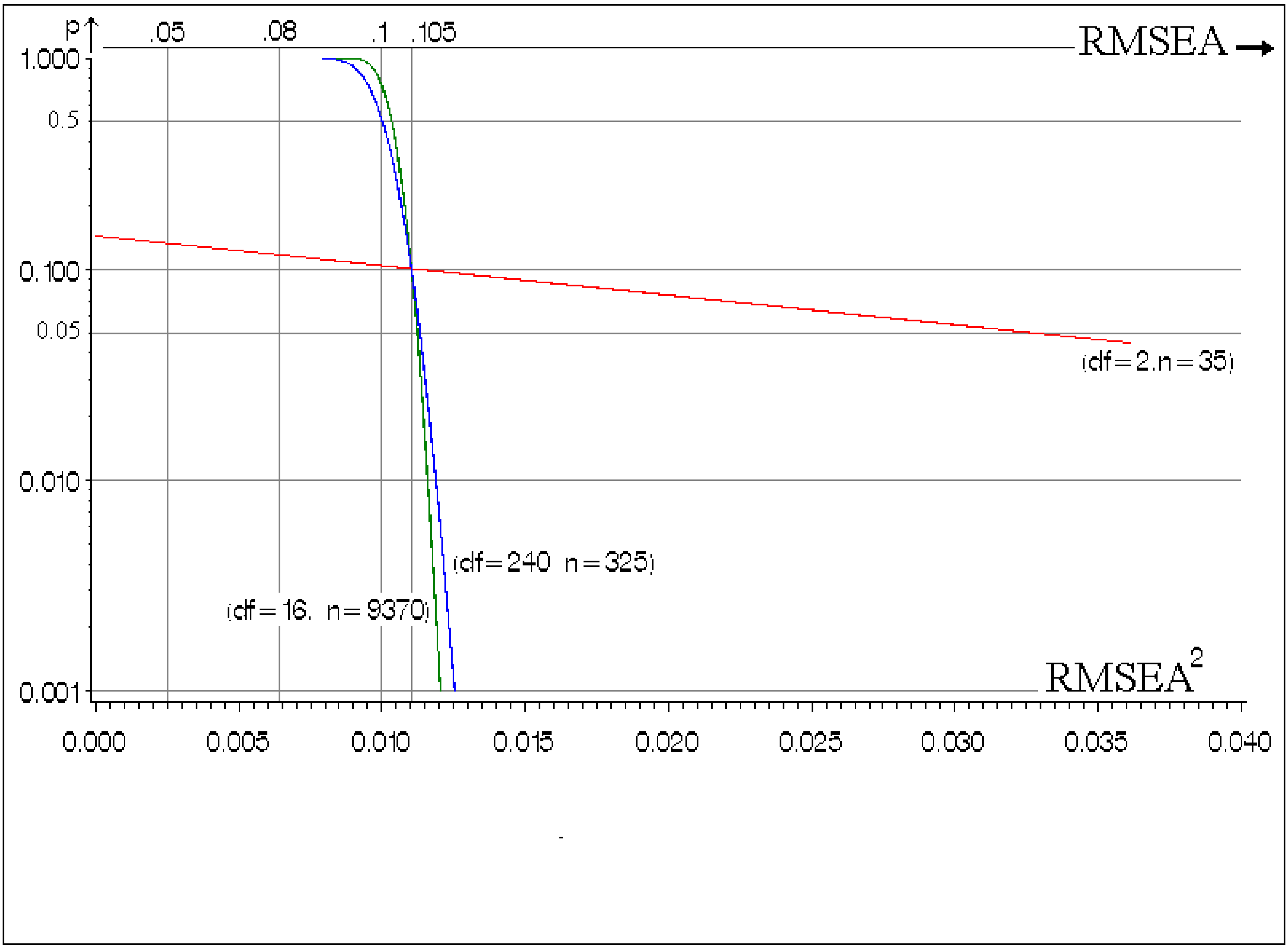

Figure1: Plot of log(p-value) against $\text{RMSEA}^2_{\text{upper-bound}}$

Note. Three curves cross the RMSEA=.105, p=.10 point. The red curve has *df*=2 and n=35, the green one has *df*=16 and n=9370, while the blue one has *df*=240 and n=325.

Appendix SAS Code

```
Data Sasuser.p_RMSEA;
        input p1 p2 RMSEA1 RMSEA2;
        * ( RMSEA1,p1) is the traditional reference point;
        * we investigate the curves through the point (RMSEA2,p2);
        * (RMSEA2,p2) is right-up near to (RMSEA1,p1) ;
        Do df=2 to 250;
                Do n=30 to 300 by 1, 305 to 1305 by 5, 1310 to 10000 by 20;
                        Lambda=n*df*RMSEA2*RMSEA2;
                        T=CInv(p2,df,Lambda);
                        if T~=. Then
                        Do;
                                W_p=Probchi(T,df,n*df*RMSEA1*RMSEA1);
                                W_RMSEA=sqrt(CNONCT(T,df,p1)/df/n);
                        End;
                        output;
                        if T=. then leave;
                End;
        End;
Cards/*p1 p2 RMSEA1 RMSEA2*/;
.05   .1    .1    .105
;

data p_RMSEA_Out;
        set Sasuser.p_RMSEA;
        where W_p<.5 and T~=.;
run;
proc sort data=p_RMSEA_Out;
        by descending w_RMSEA;
run;
proc print data=p_RMSEA_Out (obs=10);
run;
proc sort data=p_RMSEA_Out;
        by descending w_P;
run;
proc print data=p_RMSEA_Out (obs=10);
run;
data p_RMSEA_Out;
        set Sasuser.p_RMSEA;
        S=(W_p-p1)**1.2*(W_RMSEA-RMSEA1);
        where W_p<.6 and T~=.;
run;
proc report data=sasuser.p_RMSEA;
        column w_p,max;
        where w_RMSEA>.12;
run;
```