

4

DTIC FILE COPY

CRM 89-292 / December 1989

AD-A217 642

# Computing Mean Responses in Nonlinear Models

Edward S. Cavin

DTIC  
ELECTE  
FEB 05 1990  
S B D



*A Division of Hudson Institute*

**CENTER FOR NAVAL ANALYSES**

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

**DISTRIBUTION STATEMENT A**  
Approved for public release;  
Distribution Unlimited

90 02 05 0 37

**APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED**

Work conducted under contract N00014-87-C-0001.

This Research Memorandum represents the best opinion of CNA at the time of issue. It does not necessarily represent the opinion of the Department of the Navy.

REPORT DOCUMENTATION PAGE				
1a. REPORT SECURITY CLASSIFICATION <b>UNCLASSIFIED</b>		1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY		3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for Public Release; Distribution Unlimited		
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE				
4. PERFORMING ORGANIZATION REPORT NUMBER(S)  CRM 89-292		5. MONITORING ORGANIZATION REPORT NUMBER(S)		
6a. NAME OF PERFORMING ORGANIZATION  Center for Naval Analyses		6b. OFFICE SYMBOL <i>(if applicable)</i>  CNA	7a. NAME OF MONITORING ORGANIZATION	
6c. ADDRESS <i>(City, State, and ZIP Code)</i>  4401 Ford Avenue Alexandria, Virginia 22302-0268		7b. ADDRESS <i>(City, State, and ZIP Code)</i>		
8a. NAME OF FUNDING ORGANIZATION  Office of Naval Research		8b. OFFICE SYMBOL <i>(if applicable)</i>  ONR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER  N00014-87-C-0001	
8c. ADDRESS <i>(City, State, and ZIP Code)</i>  800 North Quincy Street Arlington, Virginia 22217		10. SOURCE OF FUNDING NUMBERS		
		PROGRAM ELEMENT NO. 65153M	PROJECT NO. C0031	TASK NO.
				WORK UNIT ACCESSION NO.
11. TITLE <i>(Include Security Classification)</i> <b>Computing Mean Responses in Nonlinear Models</b>				
12. PERSONAL AUTHOR(S) <b>Edward S. Cavin</b>				
13a. TYPE OF REPORT  Final		13b. TIME COVERED  FROM TO		14. DATE OF REPORT <i>(Year, Month, Day)</i>  December 1989
15. PAGE COUNT  9				
16. SUPPLEMENTARY NOTATION				
17. COSATI CODES			18. SUBJECT TERMS <i>(Continue on reverse if necessary and identify by block number)</i>	
FIELD	GROUP	SUB-GROUP	Approximation (mathematics), Computations, Econometrics, Estimates, Mathematical models, Nonlinear algebraic equations, Nonlinear analysis, Response	
12	02			
19. ABSTRACT <i>(Continue on reverse if necessary and identify by block number)</i>  It is common in empirical studies using nonlinear models to estimate the mean response by evaluating the nonlinear response function at the mean value of its argument(s). However, this procedure conceptually is flawed if the response function has significant curvature in the neighborhood of the mean. Ideally, one should evaluate the estimated response function for each of the estimated responses. In general, there will be some nonzero approximation error if one instead simply evaluates the response function at the mean of the independent variables(s). Furthermore, if the variability is significant in the independent variable(s) of interest, the approximation error of using the evaluate at the mean procedure increases. This paper examines the magnitude of the approximation error, and attempts to identify situations in which somewhat more computationally intensive procedures should be used.				
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input type="checkbox"/> UNCLASSIFIED / UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION <b>UNCLASSIFIED</b>	
22a. NAME OF RESPONSIBLE INDIVIDUAL			22b. TELEPHONE <i>(Include Area Code)</i>	22c. OFFICE SYMBOL



# CENTER FOR NAVAL ANALYSES

A Division of Hudson Institute

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

17 January 1990

## MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 89-292

Encl: (1) CNA Research Memorandum 89-292, *Computing Mean Responses in Nonlinear Models*, by Edward S. Gavin, Dec 1989

1. Enclosure (1) is forwarded as a matter of possible interest.
2. It is common in empirical studies using nonlinear models to estimate the mean response by evaluating the nonlinear response function at the mean value of its argument(s). However, this procedure conceptually is flawed if the response function has significant curvature in the neighborhood of the mean. Ideally, one should evaluate the estimated response function for each of the estimated responses. In general, there will be some nonzero approximation error if one instead simply evaluates the response function at the mean of the independent variable(s). Furthermore, if the variability is significant in the independent variable(s) of interest, the approximation error of using the "evaluate at the mean" procedure increases. This paper examines the magnitude of the approximation error, and attempts to identify situations in which somewhat more computationally intensive procedures should be used.

Lewis R. Cabe  
Director

Manpower and Training Program

Distribution List:  
Reverse page

Subj: Center for Naval Analyses Research Memorandum 89-292

Distribution List

**SNDL**

A1 ASSTSECNAV MRA  
A1 DASN MANPOWER  
A1 DASN PERSONNEL AND FAMILY MATTER  
A6 HQMC MPR & RA  
    Attn: Code M  
    Attn: Code MP  
    Attn: Code MR  
    Attn: Code MA  
    Attn: Code MH  
  
A6 HQMC TRAINING  
FF38 USNA  
FF42 NAVPGSCOL  
FF44 NAVWARCOL  
FKQ6D NAVPERSRANDCEN  
V12 CGMCCDC  
V12 CG MCRDAC - QUANTICO

**OPNAV**

OP-01  
OP-11  
OP-13  
OP-15  
OP-81  
OP-813C

# Computing Mean Responses in Nonlinear Models

Edward S. Cavin



*A Division of Hudson Institute*

---

**CENTER FOR NAVAL ANALYSES**

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

---

**ABSTRACT**

It is common in empirical studies using nonlinear models to estimate the mean response by evaluating the nonlinear response function at the mean value of its argument(s). However, this procedure conceptually is flawed if the response function has significant curvature in the neighborhood of the mean. Ideally, one should evaluate the estimated response function for each of the estimated responses. In general, there will be some nonzero approximation error if one instead simply evaluates the response function at the mean of the independent variable(s). Furthermore, if the variability is significant in the independent variable(s) of interest, the approximation error of using the "evaluate at the mean" procedure increases. This paper examines the magnitude of the approximation error, and attempts to identify situations in which somewhat more computationally intensive procedures should be used.



<b>Accession For</b>	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification _____	
By _____	
Distribution/ _____	
<b>Availability Codes</b>	
<b>Dist</b>	<b>Avail and/or Special</b>
A-1	

## CONTENTS

	Page
Introduction .....	1
Algebraic Representation of Mean Response .....	1
Empirical Examples .....	5
Mean Response in Region of Significant Curvature .....	5
Mean Response in Region of Little Curvature .....	6
Conclusion .....	7

## INTRODUCTION

Once the parameters of a nonlinear econometric model have been estimated, it usually is of interest to use these estimates to generate an estimate of mean response. The most common practice is to compute the value of the function at the mean values of the independent variables. Although this procedure works for linear models, it is, strictly speaking, inappropriate for nonlinear models. Rather, one should compute the value of the function for each observation and then estimate the mean response by the mean predicted value of the function.

However, adherents of the "evaluate at the means" procedure may fairly ask whether it makes any practical difference to compute the mean response in this somewhat more complicated fashion.<sup>1</sup> This paper describes some conditions under which it matters how the mean response is calculated and provides some examples for these cases.

## ALGEBRAIC REPRESENTATION OF MEAN RESPONSE

The choice of how to evaluate the mean response of a nonlinear function depends on how nonlinear the function is local to the mean of the independent variable and on the variability of the independent variables. These considerations can be made more explicit with the aid of the following second-order Taylor expansion of some arbitrary function about the mean of its argument.<sup>2</sup> Let

$$F(X_i) \approx F(\bar{X}) + F^{(1)}(X_i - \bar{X}) + \frac{1}{2} F^{(2)}(\bar{X})(X_i - \bar{X})^2 . \quad (1)$$

Then

$$\frac{1}{N} \sum F(X_i) \approx F(\bar{X}) + \frac{1}{2} F^{(2)} \frac{1}{N} \sum (X_i - \bar{X})^2 , \quad (2)$$

and

$$\left| \frac{1}{N} \sum F(X_i) - F(\bar{X}) \right| \approx \frac{1}{2} F^{(2)} \text{Var}(X) . \quad (3)$$

---

1. Obviously, if the function in question is easy to evaluate, there is not much excuse for using the mean values of the independent variables. However, if function evaluation is cumbersome, this is a more legitimate concern.

2. Functions with vector arguments present no special difficulties to the following discussion; a scalar argument is used purely for expository convenience.

What determines the value of equation 3, which represents the difference in the two methods for computing mean response? First, there is the curvature of the function  $F$  local to the mean of its argument. If  $F$  is linear, then  $F^{(2)}$  is zero and the two methods trivially are identical. If, on the other hand,  $F$  is strongly curved in the neighborhood of  $\bar{X}$ , then there can be serious error in using  $F(\bar{X})$ .

Second, even if  $F$  is strongly curved, if  $\text{Var}(X)$  is small the error in approximating  $\frac{1}{N} \sum F(X_i)$  with  $F(\bar{X})$  will be also. With a large  $\text{Var}(X)$ , even small amounts of curvature in the function can create significant error. The intuition behind this observation is rather obvious: if the range of variation in  $X$  is large, the chord connecting extreme values of  $X$  can lie farther away from the function  $F$ .<sup>1</sup>

These conditions can be made more explicit with the aid of a common nonlinear statistical model for discrete outcomes--the binary logit model. The binary logit model may be specified as follows:

$$y = X\beta + \epsilon \quad (4)$$

$$q = \begin{cases} 1 & \text{if } y \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$P(q = k) = F(y), \quad K = \{0,1\},$$

where  $F(y)$  is the logistic distribution function,  $y$  is a continuous variable, and  $q$  is a binary indicator variable.

Figure 1 shows the cumulative distribution and density functions ( $F(X)$  and  $f(x)$ , respectively) for the logistic distribution. From figure 1, it is obvious that the curvature of the function  $F(X)$  is smallest when  $X$  is near zero and for extreme values of  $X$ . Thus, the curvature of  $F(X)$  is smallest when  $F(X)$  is relatively small, close to 0.5, or relatively large.

---

1. A third factor, which is more subtle than the first two mentioned, is that the curvature of  $F$  is estimated with an error that depends on, among other things, the degree of "smoothness" in the data. If the  $X$  variables are highly clustered, the estimate of the parameters of  $F$ , and therefore of  $F^{(2)}$ , may be sensitive to slight perturbations in  $X$ . If the population is substantially homogeneous in its measured characteristics (i.e., the independent variables), errors from using the mean value of independent variables should be minimized.

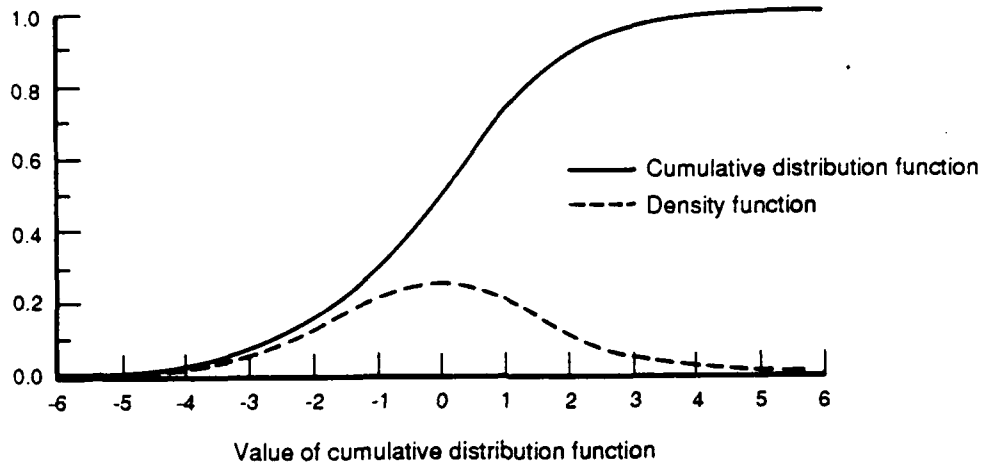


Figure 1. Logistic probability density and distribution

Figure 2 graphs  $F^{(2)}(X)$  as a function of  $F(X)$ . When  $F(X)$  lies in the ranges 0.07 to 0.35 and 0.65 to 0.93, the distribution function has significant curvature. Therefore, except for empirical situations in which the mean response is less than about 0.07, between 0.35 and 0.65, or greater than 0.93, curvature may create difficulties in approximating  $\frac{1}{N} \sum F(X)$  by  $F(\bar{X})$ .

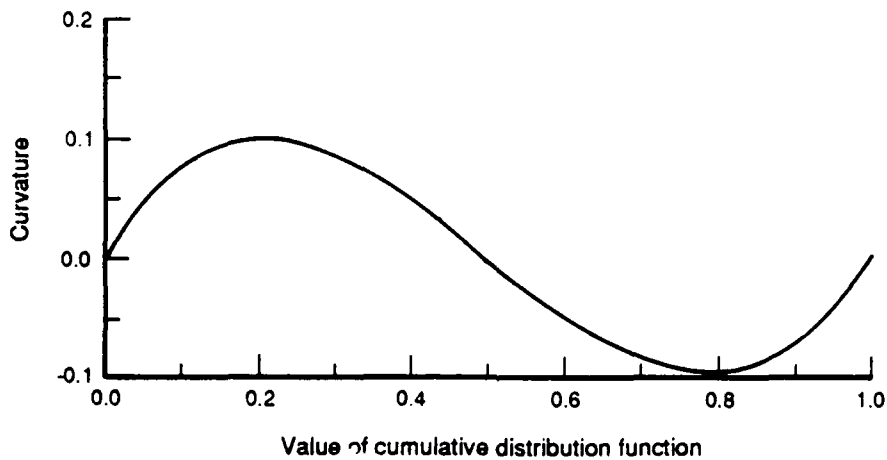


Figure 2. Curvature of logistic probability distribution

As shown in figure 3, the actual error in approximation depends on the variance of the right-hand-side variables as well as the curvature of the distribution function local to the mean response. Figure 3 shows how the approximation error is related to the variance (the variance of the argument of the distribution in logit models generally is less than unity). The effect of smaller variance essentially is to compress the curvature function presented in figure 2.

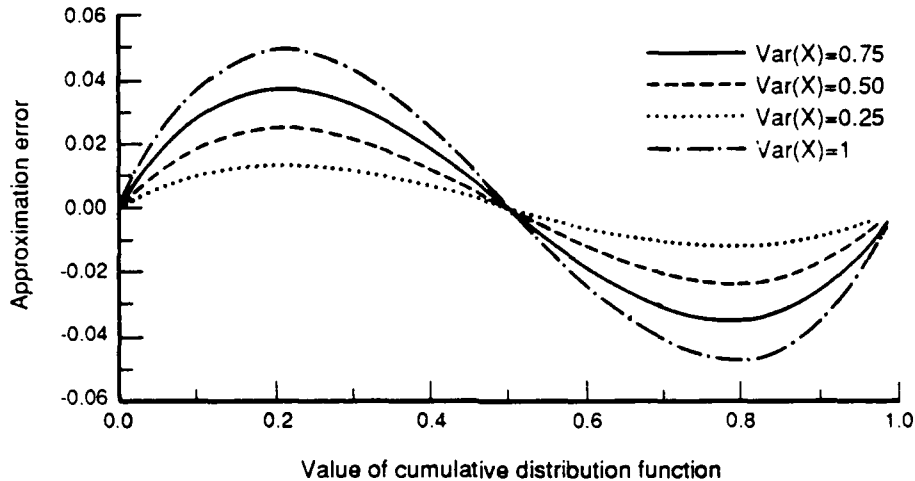


Figure 3. Approximation error (logistic distribution)

The following general observations may be made regarding the error associated with estimating the mean response using the mean value of the independent variables.

- In most standard probability models, if the mean of the dependent variable lies near 0.5, the cumulative distribution function is essentially linear, and the error is close to zero.
- Similarly, in these probability models one need be less cautious when dealing with very likely or very unlikely events.
- However, in many empirical situations, the mean response may be 0.1 to 0.3 or 0.7 to 0.9, which are ranges in which the approximation error is greatest.
- Variation in the magnitude of the variance of the independent variables does not change the dependence of the approximation error on the curvature of  $F(X)$ , but does affect its amplitude.

## EMPIRICAL EXAMPLES

### Mean Response in Region of Significant Curvature

In a recent study of retention among Marines,<sup>1</sup> the intention to reenlist was modeled using a binary logit specification in which a continuous variable measured propensity to remain in the military, and a binary variable indicated intention to quit. Table 1 presents the relevant results.

Table 1. Estimates of logit model of USMC retention

Variable	Parameter estimate	Standard error	Mean of variable
Overall satisfaction with military life (scaled)	-0.539	0.066	1.4
Satisfaction with family services (scaled)	-0.014	0.014	3.3
Perceived chances of good civilian job (scaled)	0.065	0.022	8.6
Length of service (yr)	0.061	0.021	9.0
Whether married	-0.323	0.134	0.7
Whether any minor children	-0.699	0.135	0.5
Age	-0.018	0.020	28.1
Intercept	1.874	0.478	
$Var(X\beta)$	0.536		
$\frac{1}{N} \sum F(X_i\beta)$	0.335		

1. CNA Report 139, *A Study of Marine Corps Family Programs*, by Edward S. Cavin, September 1987.

Using these results, one can calculate the error in approximating the mean response by the distribution function evaluated at the mean of the independent variables:

$$\left| \frac{1}{N} \sum F(X) - F(\bar{X}) \right| = | 0.335 - 0.318 | = 0.017 .$$

Thus, the error is approximately 2 percentage points. Since the logistic distribution function is significantly curved in the neighborhood of the mean of the independent variables (i.e., -0.7627), the error caused by using the approximation of evaluating the distribution function at the mean is relatively large.<sup>1</sup> (It would be somewhat larger, but for the relatively small variance of the  $X_i\beta$ .)

#### Mean Response in Region of Little Curvature

A very different issue recently was addressed using a similar kind of binary logit model, namely the failure rate of certain avionics on U.S. Navy aircraft.<sup>2</sup> These avionics monitor the status of the weapon systems onboard the aircraft, and can be tested using alternative automated testing procedures. For the purposes of this paper, however, the important point is that the failure rate per flight is very low (less than 1 percent).

Table 2 presents the relevant results from this analysis.

As with the previous example, one can calculate the error in approximating the mean response by the distribution function evaluated at the mean of the independent variables:

$$\left| \frac{1}{N} \sum F(X) - F(\bar{X}) \right| = | 0.0070 - 0.0045 | = 0.0025 .$$

In this case, the error is less than 0.3 percent. This is because the logistic distribution has very little curvature in the neighborhood of the mean of the independent variables (i.e., -5.392).<sup>3</sup> (The error would be a little larger if the variance were somewhat larger in this example.)

---

1. The curvature of the distribution function at  $X = -0.7627$  is:  
 $F^{(2)} = 0.078$ .

2. Example is from CNA Research Memorandum 88-175, *A Comparison of the Repair Times and Effectiveness of ATS and IAFTA*, by Robert A. Levy and Beverly Spejewski, January 1989.

3. The curvature of the distribution function at  $X = -5.392$  is:  
 $F^{(2)} = 0.004$ .

Table 2. Estimates of logit model of avionics failure

Variable	Parameter estimate	Standard error	Mean of variable
Whether new test procedure used	0.396	0.510	0.199
Whether new procedure indicated failure	0.852	0.461	0.069
Whether aircraft based at airfield 1	0.119	0.439	0.541
Whether aircraft based at airfield 2	-0.627	0.734	0.095
Whether aircraft based at airfield 3	-0.832	0.710	0.245
Number of flights since last failure	-0.028	0.006	90.2
(Number of flights since last failure) 2	$0.8 \times 10^{-4}$	$0.2 \times 10^{-4}$	14,114
Days since last flight	0.016	0.006	1.63
Intercept	-4.022	0.408	
Var ( $X\beta$ )	0.313		
$\frac{1}{N} \sum F(X_i\beta)$	0.007		

## CONCLUSION

These examples illustrate that the error in approximating the mean response in a nonlinear model by evaluating the nonlinear function at the mean of the independent variables is positively related to the degree of nonlinearity in the function. Given the ease with which nonlinear functions can be evaluated for each observation, and then averaged, the shortcut of function "at the means" should not be used unless one is certain that the function is locally linear.