D4-61

N87-24899

EARLY ESTIMATION OF RESOURCE EXPENDITURES AND PROGRAM SIZE

Prepared by

COMPUTER SCIENCES CORPORATION

D. Card

For

GODDARD SPACE FLIGHT CENTER

Under

Contract NAS 5-24300

June 1982

1. INTRODUCTION

A substantial amount of software engineering research effort has been focused on the development of software cost estimation models. A concensus (of sorts) has emerged on that topic. The following relationship is widely accepted:

$$H_s = aL^b \qquad (1)$$

where H_s = staff-hours of effort
L = lines of code
a = a constant
b = a constant

The Software Engineering Laboratory (SEL) has devised a measure of lines of code based on the origin of the delivered code that is substituted in the equation above. This is

$$L_{dev} = N + E + 0.2\ (S+O) \qquad (2)$$

where L_{dev} = "developed" lines of code
N = newly implemented lines of code
E = extensively modified lines of code
S = slightly modified lines of code
O = old (unchanged) lines of code

Equation 1 using "developed" lines of code has given good results as an estimator of development effort. (The analyses in this document are based on a sample of 20 ground-based attitude systems). Table 13 shows a regression analysis that produced a correlation of 0.99 and an estimate of b of 1.1 when the value of a was fixed at 1.0 in Equation 1. Despite these encouraging results, this model has two significant limitations. These are the following:

- The substantial amount of development work done in activities other than code implementation may not be adequately considered in the lines of code measure.

- The lines of code, whether "delivered" or "developed", is not known accurately until late in the development cycle when accurate estimates are less useful.

The purpose of this memorandum is to discuss these limitations and to propose some alternative estimation models that can be used earlier in the development process, e.g., during requirements analysis and preliminary design.

2. MODELS OF WORK

The obvious alternative to lines of code as a measure of the work done is pages of documentation. Although only a portion of a software development team is involved in coding, almost everyone produces some documentation. This includes requirements, design, and operations documents. Table 1 compares the components of developed lines of code with pages of documentation as estimators or programmer hours. A regression model based on the two most strongly correlated measures is described in Table 2. This model showed the following relationship:

$$H_p = 0.056\ N + 4.15D \qquad (3)$$

where H_p = programmer hours
N = newly implemented lines of code
D = pages of documentation

A similar comparison is made in Table 3 for these measures as estimators of staff-hours (including programmer, manager, and other hours). A regression model based on the two most strongly correlated measures is described in Table 4. This model showed the following relationship:

$$H_s = 0.051\ N + 7.10D \qquad (4)$$

where H_s = staff-hours
N = newly implemented lines of code
D = pages of documentation

The correlation coefficient (r) associated with each of the relationships expressed in Equations 3 and 4 was 0.97, comparable to that obtained by substituting Equation 2 for L in Equation 1. These results suggest that the best measures of work done are lines of new code and pages of documentation. Reused lines of code do not seem to contribute directly to resource expenditures. However, the requirements analysis and design effort involved in reusing previously developed code may be included in the pages of documentation measure.

Although pages of documentation appears to be an important measure of work, it has the same limitation as lines-of-code measures. Pages of documentation cannot be determined accurately early in the development cycle. The next sections discusses some other measures that can be used to develop models for early estimation of resource expenditures and program size.

3. MODELS FOR EARLY ESTIMATION

Few objective measures are available early in the software development process. The following five measures were considered in this analysis:

- Number of subsystems - requirements analysis
- Number of data sets - preliminary design
- Complexity (PRICE-S) - preliminary design
- Number of new modules - detailed design
- Number of reused modules (extensively modified, slightly modified, and old) - detailed design

The following sections discuss the use of these measures for early estimation of program size and resource expenditures.

3.1 PROGRAM SIZE

The correlations of the measures described here with delivered lines of code are compared in Table 5. Three regression models were developed (Tables 6, 7, and 8). The two most useful of these are the following:

$$L_{del} = 7596\ S \qquad (5)$$

$$L_{del} = 168N + 195R \qquad (6)$$

where L_{del} = delivered lines of code
S = number of subsystems
N = number of new modules
R = number of reused modules

Equation 5 ($r = 0.99$) defines an estimating relationship for program size that can be used during the requirements analysis phase. Equation 6 ($r = 0.98$) defines an estimating relationship of comparable reliability that can be used during the design phase.

3.2 RESOURCE EXPENDITURES

The correlations of the measures described here with staff-hours of effort are compared in Table 9. Three regression models were developed (Tables 10, 11, and 12). The two most useful of these are the following:

$$H_s = 1634\ S \qquad (7)$$

$$H_s = 45\ N + 28\ R \qquad (8)$$

where H_s = staff-hours
S = number of subsystems
N = number of new modules
R = number of reused modules

Equation 7 ($r = 0.93$) defines an estimating relationship for resource expenditures that can be used during the requirements analysis phase. Equation 8 ($r = 0.94$) defines an

estimating relationship of higher reliability that can be used during the design phase.

4. CONCLUSION

The preceding analysis has demonstrated two important points. These are the following:

- New measures of productivity which incorporate other development products besides lines of code must be investigated. Pages of documentation is a good candidate.
- Effective estimates of program size and resource expenditures can be made using measures that are available early in the development cycle.

Table 1. Components of Programmer Effort

N= 20 REGRESSION MODELS FOR DEPENDENT VARIABLE PGMRHRS

NUMBER IN MODEL	R-SQUARE	VARIABLES IN MODEL
1	0.08943231	OLDLINES
1	0.49044658	MODLINES
1	0.80450662	NEWLINES
1	0.85046601	DOCPAGES
2	0.49386580	MODLINES OLDLINES
2	0.80450674	NEWLINES MODLINES
2	0.81581865	NEWLINES OLDLINES
2	0.85129683	OLDLINES DOCPAGES
2	0.85198581	MODLINES DOCPAGES
2	0.85887286	NEWLINES DOCPAGES
3	0.81685888	NEWLINES MODLINES OLDLINES
3	0.85257247	MODLINES OLDLINES DOCPAGES
3	0.85888650	NEWLINES OLDLINES DOCPAGES
3	0.86233681	NEWLINES MODLINES DOCPAGES
4	0.86261078	NEWLINES MODLINES OLDLINES DOCPAGES

Table 2. Model of Programmer Effort

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: PGMRHRS

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	2	106607575032.508650	53303787516.254325	130.92	0.0001	0.935679	36.2358
ERROR	18	7328452413.491348	407136245.193964		STD DEV		PGMRHRS MEAN
UNCORRECTED TOTAL	20	113936027446.000000			20177.617431		55684.20000000

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
NEWLINES	1	103353493284.406580	253.85	0.0001	1	448211536.116894	1.10	0.3080
DOCPAGES	1	3254081748.102075	7.99	0.0112	1	3254081748.102075	7.99	0.0112

PARAMETER	ESTIMATE	T FOR HO: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
NEWLINES	0.55992183	1.05	0.3080	0.53364909
DOCPAGES	41.51867299	2.83	0.0112	14.68585102

Table 3. Components of Total Staff Effort

N= 20 REGRESSION MODELS FOR DEPENDENT VARIABLE MANHRS

NUMBER IN MODEL	R-SQUARE	VARIABLES IN MODEL
1	0.09707264	OLDLINES
1	0.53222266	MODLINES
1	0.81364699	NEWLINES
1	0.88750508	DOCPAGES
2	0.53593783	MODLINES OLDLINES
2	0.81530674	NEWLINES MODLINES
2	0.82758155	NEWLINES OLDLINES
2	0.88779376	MODLINES DOCPAGES
2	0.88803206	OLDLINES DOCPAGES
2	0.89026281	NEWLINES DOCPAGES
3	0.82762271	NEWLINES MODLINES OLDLINES
3	0.88823599	MODLINES OLDLINES DOCPAGES
3	0.89028560	NEWLINES OLDLINES DOCPAGES
3	0.89106130	NEWLINES MODLINES DOCPAGES
4	0.89106286	NEWLINES MODLINES OLDLINES DOCPAGES

Table 4. Model of Total Staff Effort

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: MANHRS

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	2	238607979283.655180	119303989641.827590	167.97	0.0001	0.949144	32.3398
ERROR	18	12784812713.344818	710267372.963601		STD DEV		MANHRS MEAN
UNCORRECTED TOTAL	20	251392791997.000000			26650.841881		82408.75000000

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
NEWLINES	1	229088029089.200590	322.54	0.0001	1	370237194.893545	0.52	0.4796
DOCPAGES	1	9519950194.454598	13.40	0.0018	1	9519950194.454598	13.40	0.0018

PARAMETER	ESTIMATE	T FOR HO: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
NEWLINES	0.50889234	0.72	0.4796	0.70485019
DOCPAGES	71.01442466	3.66	0.0018	19.39725019

Table 5. Comparison of Early Size Estimators

N= 20 REGRESSION MODELS FOR DEPENDENT VARIABLE TOTLINES

NUMBER IN MODEL	R-SQUARE	VARIABLES IN MODEL
1	0.05087892	COMPLXTY
1	0.60752653	NODATSET
1	0.68352123	REMODS
1	0.77270837	NEWMOD
1	0.96967766	NOSUBSYS
2	0.60988357	NODATSET COMPLXTY
2	0.71154503	REMODS COMPLXTY
2	0.77923131	NODATSET NEWMOD
2	0.79114322	NEWMOD COMPLXTY
2	0.87902381	NODATSET REMODS
2	0.93248454	NEWMOD REMODS
2	0.97280856	NOSUBSYS NODATSET
2	0.97356977	NOSUBSYS NEWMOD
2	0.97363300	NOSUBSYS COMPLXTY
2	0.97737439	NOSUBSYS REMODS
3	0.79403377	NODATSET NEWMOD COMPLXTY
3	0.88406426	NODATSET REMODS COMPLXTY
3	0.93555143	NEWMOD REMODS COMPLXTY
3	0.93928611	NODATSET NEWMOD REMODS
3	0.97487138	NOSUBSYS NODATSET NEWMOD
3	0.97550227	NOSUBSYS NEWMOD COMPLXTY
3	0.97689986	NOSUBSYS NODATSET COMPLXTY
3	0.97746439	NOSUBSYS NEWMOD REMODS
3	0.97757499	NOSUBSYS NODATSET REMODS
3	0.97907109	NOSUBSYS REMODS COMPLXTY
4	0.94078488	NODATSET NEWMOD REMODS COMPLXTY
4	0.97742335	NOSUBSYS NODATSET NEWMOD COMPLXTY
4	0.97764845	NOSUBSYS NODATSET NEWMOD REMODS
4	0.97907134	NOSUBSYS NEWMOD REMODS COMPLXTY
4	0.97955179	NOSUBSYS NODATSET REMODS COMPLXTY
5	0.97956096	NOSUBSYS NODATSET NEWMOD REMODS COMPLXTY

Table 6. Minimal Size Estimating Model

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: TOTLINES

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	1	47483674672.6913720	47483674672.6913720	1143.24	0.0001	0.983652	17.5551
ERROR	19	789152506.3086271	41534342.4372962		STD DEV		TOTLINES MEAN
UNCORRECTED TOTAL	20	48272827179.0000000			6444.7143022		36711.35000000

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
NOSUBSYS	1	47483674672.6913720	1143.24	0.0001	1	47483674672.6913690	1143.24	0.0001

PARAMETER	ESTIMATE	T FOR HO: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
NOSUBSYS	7595.77764277	33.81	0.0001	224.64861839

Table 7. Optimal Size Estimating Model

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: TOTLINES

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	2	46817506873.1303250	23408753436.5651620	289.53	0.0001	0.969852	24.4930
ERROR	18	1455320305.8696746	80851128.1038708		STD DEV		TOTLINES MEAN
UNCORRECTED TOTAL	20	48272827179.0000000			8991.7255354		36711.35000000

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
NEWMOD	1	43332342186.0297470	535.95	0.0001	1	6521075210.1470580	80.66	0.0001
REMODS	1	3485164687.1005773	43.11	0.0001	1	3485164687.1005768	43.11	0.0001

PARAMETER	ESTIMATE	T FOR HO: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
NEWMOD	168.14286708	8.98	0.0001	18.72241579
REMODS	195.10551281	6.57	0.0001	29.71672397

Table 8. Alternative Size Estimating Model

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: TOTLINES

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	3	47824276670.3418870	15941425556.7806290	604.18	0.0001	0.990708	13.9920
ERROR	17	448550508.6581125	26385324.0387125		STD DEV		TOTLINES MEAN
UNCORRECTED TOTAL	20	48272827179.0000000			5136.6646804		36711.35000000

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
NOSUBSYS	1	47483674672.6913720	1799.62	0.0001	1	5853376833.9617080	221.84	0.0001
REMODS	1	177236696.2114050	6.72	0.0190	1	130771745.4506511	4.96	0.0398
COMPLXTY	1	163365301.4391105	6.19	0.0235	1	163365301.4391105	6.19	0.0235

PARAMETER	ESTIMATE	T FOR HO: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
NOSUBSYS	7408.97828817	14.89	0.0001	497.43495025
REMODS	52.08554530	2.23	0.0398	23.39599237
COMPLXTY	-46.45837911	-2.49	0.0235	18.67090462

Table 9. Comparison of Early Resource Estimators

N= 20 REGRESSION MODELS FOR DEPENDENT VARIABLE MANHRS

NUMBER IN MODEL	R-SQUARE	VARIABLES IN MODEL
1	0.01590804	COMPLXTY
1	0.45501952	REMODS
1	0.45541937	NODATSET
1	0.70892650	NEWMOD
1	0.72646057	NOSUBSYS
2	0.45626141	NODATSET COMPLXTY
2	0.46118614	REMODS COMPLXTY
2	0.61892671	NODATSET REMODS
2	0.71178663	NODATSET NEWMOD
2	0.72655608	NOSUBSYS REMODS
2	0.72877159	NOSUBSYS NODATSET
2	0.74250148	NOSUBSYS COMPLXTY
2	0.74906820	NOSUBSYS NEWMOD
2	0.76095094	NEWMOD COMPLXTY
2	0.76680965	NEWMOD REMODS
3	0.61906266	NODATSET REMODS COMPLXTY
3	0.72903802	NOSUBSYS NODATSET REMODS
3	0.74310346	NOSUBSYS REMODS COMPLXTY
3	0.74504992	NOSUBSYS NODATSET COMPLXTY
3	0.76091707	NOSUBSYS NODATSET NEWMOD
3	0.76689486	NOSUBSYS NEWMOD REMODS
3	0.76956103	NODATSET NEWMOD REMODS
3	0.77156956	NODATSET NEWMOD COMPLXTY
3	0.78545331	NOSUBSYS NEWMOD COMPLXTY
3	0.80024180	NEWMOD REMODS COMPLXTY
4	0.74860158	NOSUBSYS NODATSET REMODS COMPLXTY
4	0.77089041	NOSUBSYS NODATSET NEWMOD REMODS
4	0.80025824	NOSUBSYS NEWMOD REMODS COMPLXTY
4	0.80557302	NOSUBSYS NODATSET NEWMOD COMPLXTY
4	0.80898615	NODATSET NEWMOD REMODS COMPLXTY
5	0.81071231	NOSUBSYS NODATSET NEWMOD REMODS COMPLXTY

Table 10. Minimal Resource Estimating Model

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: MANHRS

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	1	219779472655.727810	219779472655.727810	132.09	0.0001	0.874247	49.4977
ERROR	19	31613319341.272186	1663858912.698536		STD DEV		MANHRS MEAN
UNCORRECTED TOTAL	20	251392791997.000000			40790.426729		82408.75000000

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
NOSUBSYS	1	219779472655.727810	132.09	0.0001	1	219779472655.727790	132.09	0.0001

PARAMETER	ESTIMATE	T FOR HO: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
NOSUBSYS	16341.56500608	11.49	0.0001	1421.86489246

Table 11. Optimal Resource Estimating Model

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: MANHRS

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	2	224429018553.875590	112214509276.937780	74.91	0.0001	0.892742	46.9657
ERROR	18	26963773443.124404	1497987413.506911		STD DEV		MANHRS MEAN
UNCORRECTED TOTAL	20	251392791997.000000			38703.842361		82408.75000000

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
NEWMOD	1	217210830613.300650	145.00	0.0001	1	47581329259.352713	31.76	0.0001
REMODS	1	7218187940.574946	4.82	0.0415	1	7218187940.574943	4.82	0.0415

PARAMETER	ESTIMATE	T FOR HO: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
NEWMOD	454.18954107	5.64	0.0001	80.58847287
REMODS	280.78375723	2.20	0.0415	127.91220057

Table 12. Alternative Resource Estimating Model

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: MANHRS

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	3	224740240779.482830	74913413593.160930	47.78	0.0001	0.893980	48.0476
ERROR	17	26652551217.517166	1567797130.442186		STD DEV		MANHRS MEAN
UNCORRECTED TOTAL	20	251392791997.000000			39595.418049		82408.75000000

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
NEWMOD	1	217210830613.300650	138.55	0.0001	1	35681483401.478013	22.76	0.0002
REMODS	1	7218187940.574946	4.60	0.0466	1	7506110162.185972	4.79	0.0429
COMPLXTY	1	311222225.607251	0.20	0.6615	1	311222225.607251	0.20	0.6615

PARAMETER	ESTIMATE	T FOR HO: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
NEWMOD	479.86369347	4.77	0.0002	100.58688857
REMODS	289.54047766	2.19	0.0429	132.32647759
COMPLXTY	-62.79840646	-0.45	0.6615	140.94778457

Table 13. Log Model of Resource-Size Relationship

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: MANHRS

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	1	2348.55070494	2348.55070494	19095.41	0.0001	0.999006	3.2512
ERROR	19	2.33681641	0.12299034		STD DEV		MANHRS MEAN
UNCORRECTED TOTAL	20	2350.88752135			0.35069978		10.78669231

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
DEVLINES	1	2348.55070494	19095.41	0.0001	1	2348.55070494	19095.41	0.0001

PARAMETER	ESTIMATE	T FOR HO: PARAMETER=0	PR > |T|	STD ERROR OF ESTIMATE
DEVLINES	1.10558402	138.19	0.0001	0.00800069

Table 14. SUMMARY STATISTICS FOR MEASURES

MEASURE	MEAN	MEDIAN
DOCPAGES	995	762
NEWLINES	25,928	13,550
TOTLINES	36,711	16,265
PGMRHRS	5,568.4	3,051.1
MANHRS	8,240.9	4,541.8
NEWMOD	133	88
REMODS	76	32

Figure 1. Relationship of Modules to Size

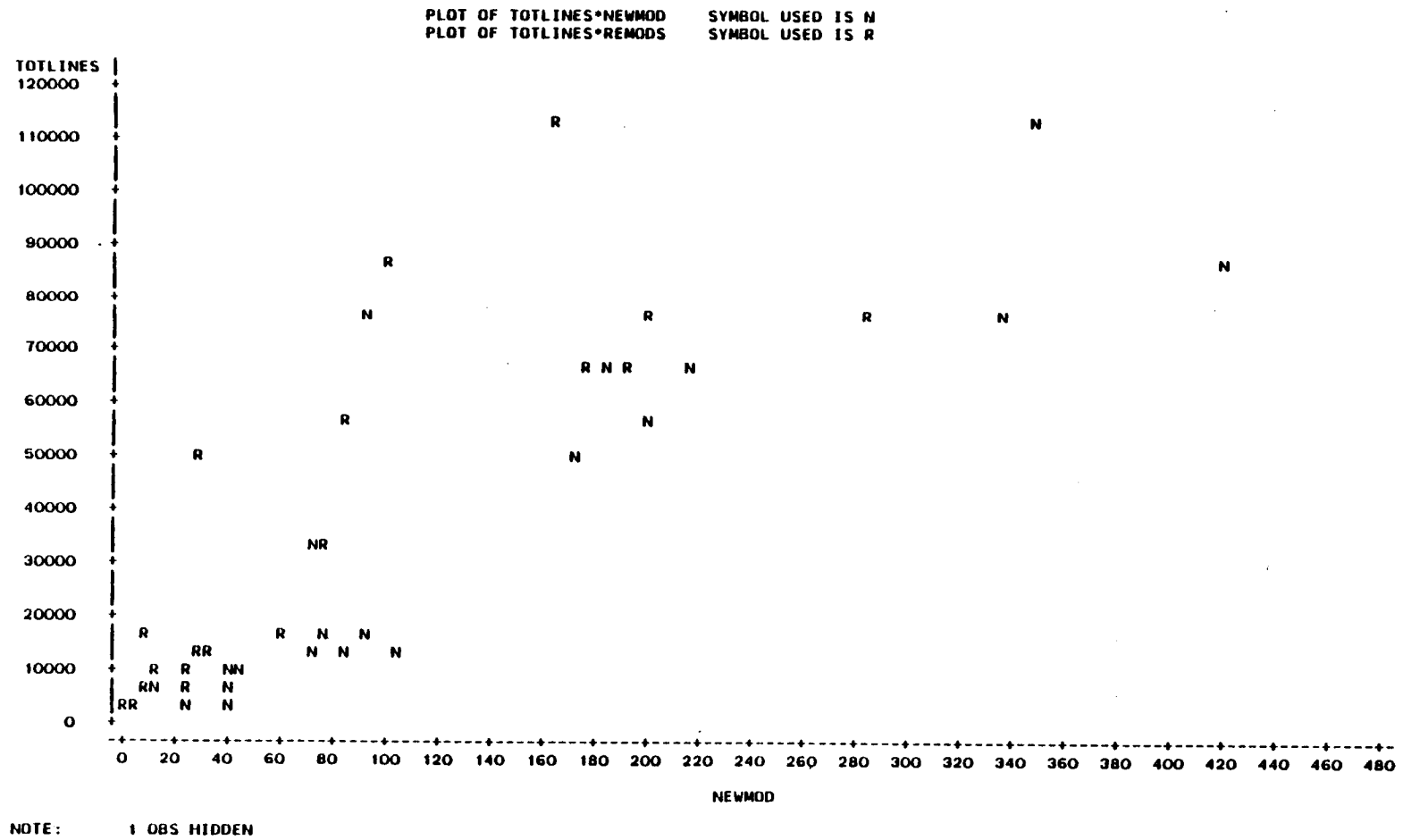

Figure 2. Relationship of System to Size

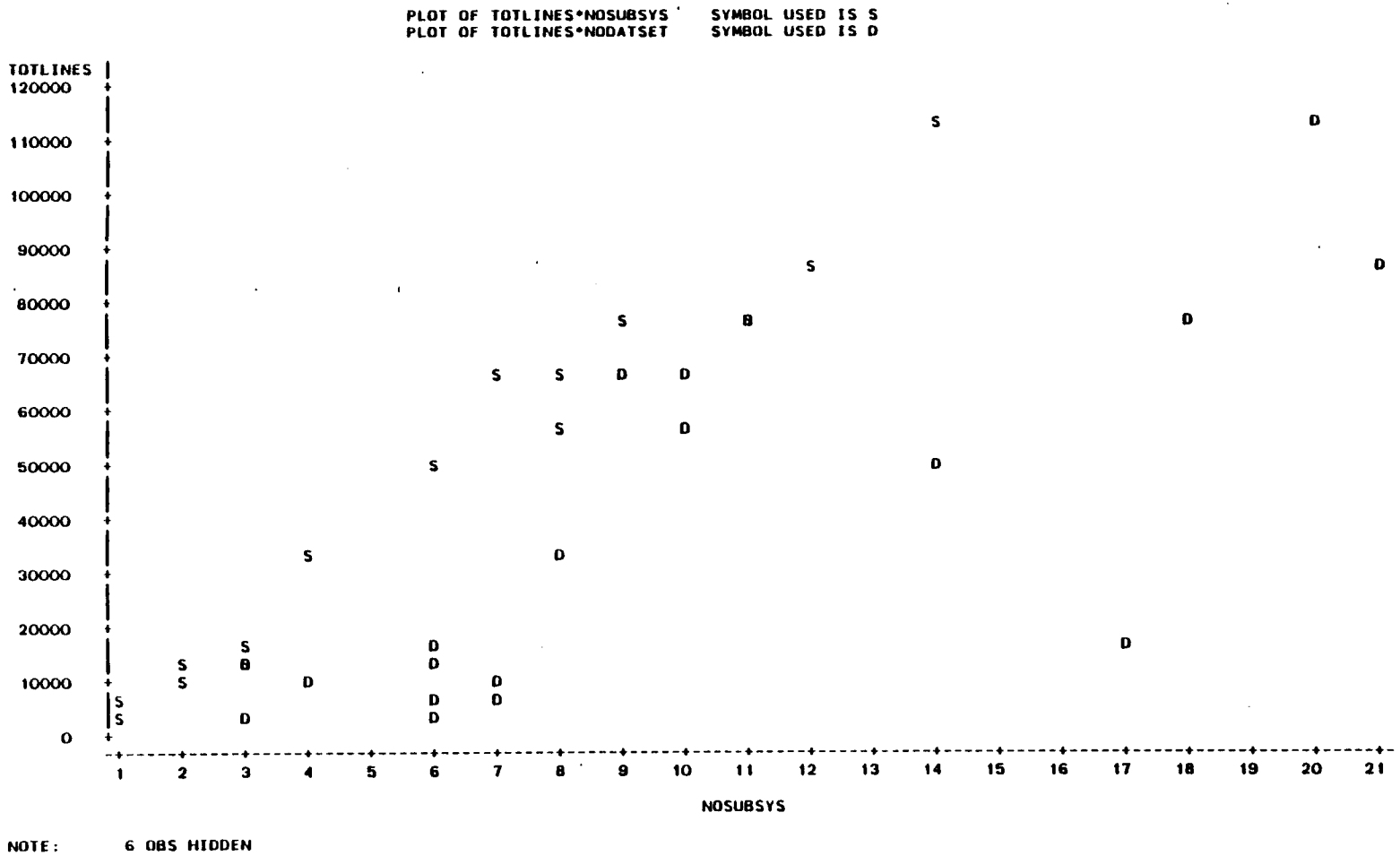

Figure 3. Relationship of Modules to Total Staff Effort

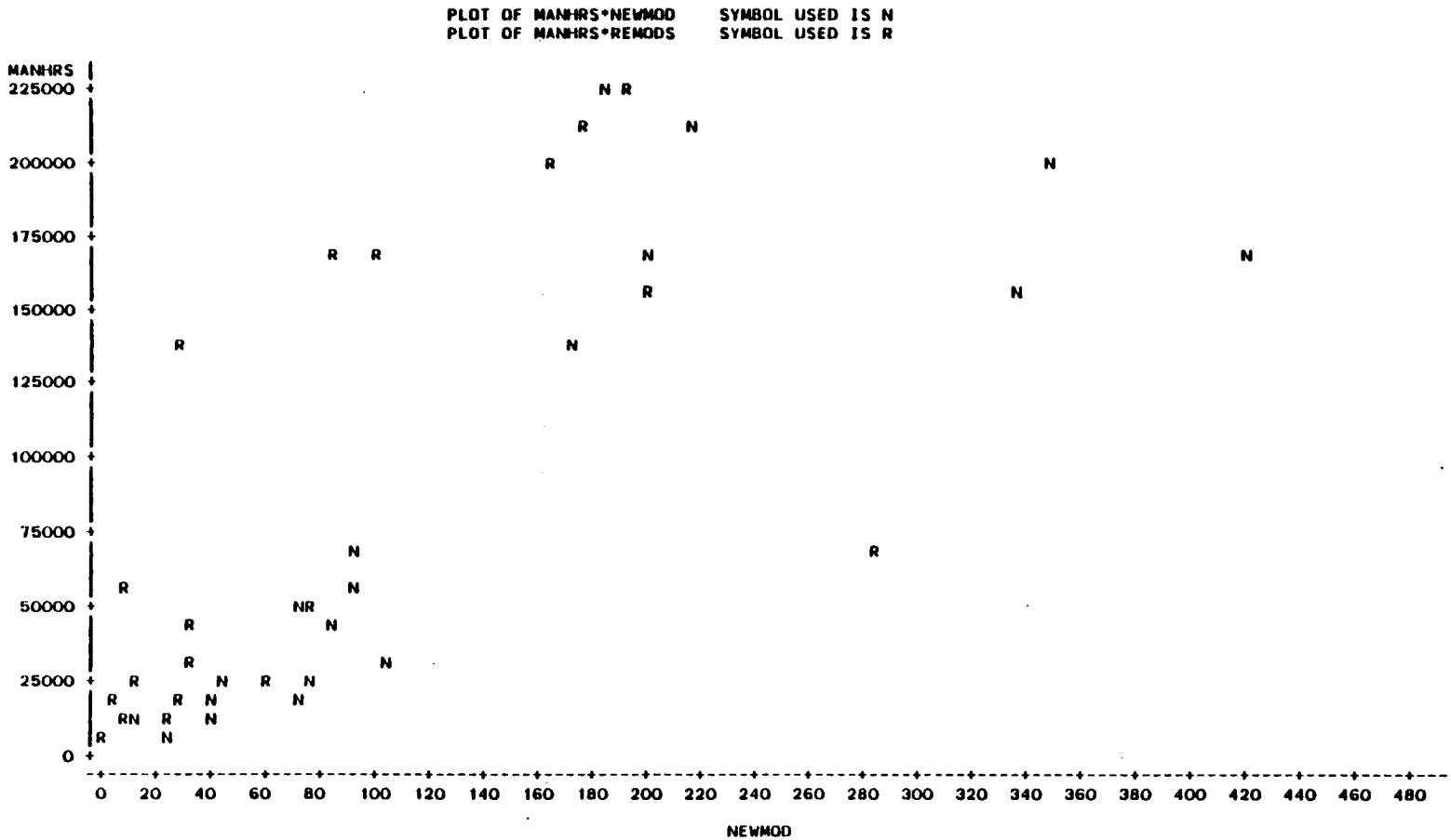

Figure 4. Relationship of System to Total Staff Effort