AUTOMATIC EXTRACTION OF CONTENT - SIGNIFICANT SENTENCES
LA PREPARATION AUTOMATIQUE DES PRECIS

John M. Carroll & John Cakarnis
Computer Science Department
University of Western Ontario

ABSTRACT

Content-significant sentences can be extracted automatically in decreasing order of importance from scientific papers available in machine-sensible format. Ordering is accomplished according to multiple regression of a non-linear combination of variables. The dependent variable is a subjective weighting of sentence importance. The independent variables include computational linguistic properties such as each word's relative frequency, the type-to-token ratio, mean-word length, and the predominent parts of speech encountered. (Il est possible de reduire un document a un cinquième de sa longeur originelle si on peut identifier les phrases les plus importantes. Nous avons découvert plusieures charactéristiques de text qui assistent à prendre des decisions. Il y a des repetitions de mots; la moyènne des nombres des lettres dans chaque mot; le nombre et la longeur de constructions grammatiques dans chaque phrase, etc. Nous avons utilisé les moyens statistiques dans notre oeuvre en particulier l'analyse de regression. Avec cette technique nous avons fait les precis sur quelques documents concernant la santé des animaux. Les savants dont nous avons obtenu ces documents nous ont dit que nos precis sont trés acceptables.)

OBJECTIVE

There are basically two ways to analyze documentary text automatically. The most common way is to store some form of lexicon or thesaurus and search for these words or stems in the subject text. The other is to make inferences regarding content from measurable characteristics of the text.

AUTOMATIC EXTRACTION OF SENTENCES

The former approach seems to have attracted most attention, a promising piece of work being the fact-retrieval program recently described by O'Connor (1973). We chose to explore the purely computational approach. Quite possibly an ideal text analysis regieme would incorporate elements from both approaches.

MEASUREMENT OF SIGNIFICANCE

We will now discuss six proposed measures of sentence significance: semantic weight, average word length, length of sub-sentence construction, type-token ratio, verb-ishness, noun-ishness. (See Table II).

We define a sentence as a string ending with a period followed by a space or line feed/carriage return and preceded by three or more non-blank, non-numeric characters.

Sentence Weight (1), i.e. Semantic Weight x_1

The idea of measuring the semantic weight of sentences originated with Luhn (1958). It derives from the premise that sentences pertinent to the central theme of document will be rich in words semantically related to the theme and that these words, with allowance for potential distortions arising from synonyms, homographs, and virtual mentions, can be identified by their local occurrence frequencies.

Edmundson (1961) and others pointed out that a better means for identifying words central to the theme of a document would be the ratio between their local and global occurrence frequencies.

On the surface, this would seem to require storing an extensive lexicon and, indeed, in earlier work we did that very thing (CARROLL & ROELOFFS 1969). However, recently Langland (1973) and others have shown that on the average words retain many of their unique characteristics even when severely truncated. Therefore, we used a table of trigraphic occurrence frequencies from Pratt (1939) modified to focus on initial trigraphs as our global lexicon. This required storing a list of only 419 trigraphs.

eg	aΔΔ	.0108
	abℓ	.0020
	abo	.0011
	acc	.0010
	ach	.0011

etc.

AUTOMATIC EXTRACTION OF SENTENCES

The significance of all independent variables (x_1 to x_6) were determined by regression analysis (see Computational Procedures).

Word Length (2) (x_2)

Inasmuch as we would be dealing with scientific literature, we postulated that in writing a content significant sentence, the author would choose words having a high degree of denotative precision and that such words would tend to be long ones. This premise lead us to identify another potentially determining characteristic: characters per word. (Averaged over a sentence).

Length of Construction (3) (x_3)

Content-significant sentences, likewise, can be observed to be on average more complex than other sentences. A list of some 25 prepositions proved to be reliable delimiters of clauses and phrases. Using these delimiters, we could arrive at a measure of the complexity of a sentence: the number of words per presentential grammatical construction divided by the number of constructions per sentence.

Type-To-Token Ratio (4) (x_4)

This measure is commonly used by computational linguists and we decided to compute it on a sentence by sentence basis and include it as one of our possibly determining characteristics. It is the ratio of the number of distinct stems encountered to the total number of appearances of these stems. It gives vastly different results when computed over discrete sentences than when computed over an entire document.

Verb-ishness (5) and Noun-ishness (6) (x_5) (x_6)

It is a point of common knowledge that nouns and verbs tend to carry the communications load in language. See, for example, Yuen Ren Chao "Language and Symbolic Systems", Cambridge, 1968, p. 91. However, in a post-inflectional language such as English, it is by no means easy to identify nouns and verbs without storing a vast lexicon. We attempted, therefore, to measure the relative noun-ishness and verb-ishness of sentences by counting on a sentence-by-sentence basis on one hand, the occurrences of common determiners, adjectival endings, and nominal endings and, on the other hand, the occurrences of auxiliary verbs, adverbial endings, and verbal endings. These counts were normalized by dividing by the number of words in the sentence. The evaluation of techniques for determining noun-ishness and verb-ishness were necessarily empirical and more work can profitably be done in this area.

AUTOMATIC EXTRACTION OF SENTENCES

COMPUTATIONAL PROCEDURES

Linear Model

We decided to try the technique of multiple linear regression to find out how much to weigh each of these sentence characteristics, that is, determination of the b's. We experimented with a data bank consisting of 14 scientific papers ranging in length from 1,000 to 3,000 words each and dealing with a variety of subjects from astronomy to zoology.

We determined the value of the dependent variable for each sentence by applying manually the decision rules set forth in Table I.

(1) Our initial regression equation was of the form:

$$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 + b_6 X_6$$

The meaning of the independent variables are summarized in Table II.

(2) We carried out an error-sum-of-squares analysis and determined that independent variables X_1, X_2, X_3, and X_5 were sufficiently significant to be retained in our model. Figure 1A shows the standard error of the fit plotted against the number of independent variables.

Nonlinear Combination

An examination of the covariance matrix cells in which X_1, X_2, X_3 and X_5 were involved persuaded us that there were perhaps some important interactions between variables. This examination led to the reintroduction of X_4 in combined form but not X_6 and resulted in a nonlinear combination of variables of the form

(3) $$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_5 X_5 + b_{14} X_1 X_4 + b_{23} X_2 X_3 + b_{24} X_2 X_4 + b_{25} X_2 X_5$$

which was investigated.

We carried out an error-sum-of-squares analysis and determined that the regression model should be:

(4) $$y = b_0 + b_2 X_2 + b_3 X_3 + b_{14} X_1 X_4 + b_{24} X_2 X_4 + b_{25} X_2 X$$

with the values:

$$
\begin{aligned}
y = & -34.6 \\
& + 12.3 \text{ X (characters/word)} \\
& + 10.9 \text{ X (words/construction)} \\
& + \ 0.2 \text{ X (semantic weight) X (type-token ratio)} \\
& - 10.5 \text{ X (characters/word) X (type-token ratio)} \\
& - \ 6.7 \text{ X (characters/word) X (verb-ishness)}
\end{aligned}
$$

Figure 2A shows the standard error of fit plotted against the number of independent variables.

The negative coefficients in the last three terms suggested that there might be an inverse relationship involving the type-token ratio and the word-length measure. We, therefore, tried a 10-variable regression with the model:

$$
\begin{aligned}
(5) \quad y = & \ b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_5 X_5 \\
& + b_{14} X_1 X_4 + b_{23} X_2 X_3 + b_{24} X_2 X_4 + B_{25} X_2 X_5 \\
& + b_{21} (X_2)^{-1} + b_{41} (X_4)^{-1}
\end{aligned}
$$

We carried out an error-sum-of-squares analysis and determined that the regression model should be:

$$
\begin{aligned}
(6) \quad y = & \ b_0 + b_3 X_3 + 0.15 b_{14} X_1 X_4 + 15.9 X_2 X_4 \\
& + b_{21} (X_2)^{-1} + B_{41} (X_4)^{-1}
\end{aligned}
$$

or

$$
\begin{aligned}
y = & - 189.0 \\
& + 15.5 \text{ X (words/construction)} \\
& + \ 0.15 \text{ X (semantic weight) X (type-token ratio)} \\
& + 15.9 \text{ X (characters/word) X (type-token ratio)} \\
& +110.5 \text{ X (words/character)} \\
& + 58.4 \text{ X (token-type ratio).}
\end{aligned}
$$

Figure 3A shows the standard error of fit plotted against the number of independent variables.

Table III lists the beginning and final values of mean sum of squares for error, regression coefficient and statistical significance (F) for each of the three analysis.

RESULTS

We tested our model against a collection of papers on veterinary medicine supplied by the State University of Iowa, Ames, Iowa. The collection consisted of three independent papers and a long report in five distinct parts, a total of some 13 thousand words.

We picked of the highest weighted sentences from each paper until we had assembled about 20 percent of the original length of each paper (or two sentences in the case of one short paper) and arrayed the sentences in occurrence order to produce condensed papers. Table IV gives the gross results of automatically condensing these papers.

Minimal post editing (a dozen manual interventions) sufficed to make the condensed papers read smoothly but would have been unnecessary were scientific sense the sole criterion.

The condensed papers were judged by the scientists who supplied them to represent adequately the sense and principal content of the originals. Evaluation of techniques for extracting meaning from text can probably never satisfy all critics. The author has tried use of "juries" [5], but in scientific work the audience is typically limited and the judgement of actual users appears to be more a meaningful criteria than compounding of ignorance.

The author thanks Dr. Norman E. Hutton of the College of Veterinary Medicine of the State University of Iowa and his colleagues for serving as adjudicators.

This work was financed in part by the National Research Council under grant number A-7132.

REFERENCES

CARROLL, J.M. & ROELOFFS, R. Computer Selection of Keywords Using Word-Frequency Analysis, JASIS, Vol. 20, 1969. 227p.

EDMUNDSON, H.P.& WYLLYS, R.E. Automatic Abstracting and Indexing - Survey and Recommendations. CACM, Vol. 4, 1961. 226p.

LANGLAND, K.G. A Study of Key Abbreviation Schemes, M.Sc. Thesis, UWO, April 1973.

AUTOMATIC EXTRACTION OF SENTENCES

REFERENCES (cont'd)

LUHN, H.P. The Automatic Creation of Literature Abstracts. IBM Jour. R & D, Vol. 2, 1958. 154p.

LUHN, H.P. An Experiment in Auto-Abstracting. Intern. Conf. on Scientific Information, Washington, D.C., November 16-21 1958.

O'CONNOR, J. Text Searching Retrieval of Answer-Sentences and other Answer Packages. JASIS, Vol. 24. 1973. 445p.

PRATT, F. Secret and Urgent, The Story of Codes and Cyphers. Blue Ribbon Books, Garden City, N.Y. 1939.

TABLE I - Decision rules for manually assigning experimental sentence significance values (Y)

Subjective Classification of Sentence	Base Value	Differential Value
Title	100	---
Subject	50	50 (ie 100-50)
Purpose	33	17
Major Definition	25	8
Minor Definition	20	5
Major Fact	17	3
Minor Fact	14	3
Reference to Prior Work	12	2
Laboratory Procedure	11	1
Subheading	10	1
Nil	1	9

Number of Concepts in Sentence	Multiplier of Differential Value
0	9
9	0.95
8	.91
7	.85
6	.78
5	.70
4	.6
3	.5
2	.3
1	.1

Y = Sentence Value = Base Value + Differential Value X Multiplier

TABLE II - Meanings of independent variables

X_1 Semantic weight/number of words in sentence

X_2 Number of characters in sentence / number of words in sentence

X_3 Number of words in sentence / number of sub-sentence grammatical constructions (ie number of delimiters + 1)

X_4 Type-token-ratio (number of unique words in sentence/ total number of words in sentence)

X_5 Verb-ishness: number of verb-adverb indicators present/ number of words in sentence (is, am, are, was, were, have, has, had, isn't, well, to, may, can could, might, be, should, would, ing, ed, es, ly)

X_6 Noun-ishness: number of noun-adjective indicators present/number of words in sentence (the, a, an, and, this, ty, ion, ness, gt, nt, ry, ic, ian, ism, ce, "comma")

TABLE III - Summary results of regression analysis

Model	Type	Mean error sum of squares	Regression Coefficiant	Statistical Significance
(1)	Linear Comb. 6 variables	755.42	.3928	3.1627
(2)	Linear Comb. 4 variables	748.55	.3819	4.526
(3)	Nonlinear Comb. 8 variables	751.67	.4179	2.699
(4)	Nonlinear Comb. 5 variables	738.41	.4095	4.230
(5)	Nonlinear & inverse Comb. 10 variables	724.94	.4687	2.815
(6)	Nonlinear & inverse Comb. 5 variables	698.39	.4589	5.603

TABLE IV - Gross results of condensing eight papers on veterinary medicine

Paper number	Original length (words)	Length of condensate (words)
1	1,800	250
2	987	250
3	3,567	400
4	220	60
5	1,473	270
6	907	260
7	897	290
8	3,307	562
	13,158	2,342 (18 percent overall)

LEGENDS

Figure 1A - Standard error of fit versus number of independent variables (linear combination)

Figure 1B - Snedecor's F statistic versus number of independent variables

Figure 2A - Standard error of fit versus number of independent variables (nonlinear combination)

Figure 2B - F statistic versus number of independent variables

Figure 3A - Standard error of fit versus number of independent variables (non-linear combination with inverse relationships)

Figure 3B - F statistic versus number of independent variables

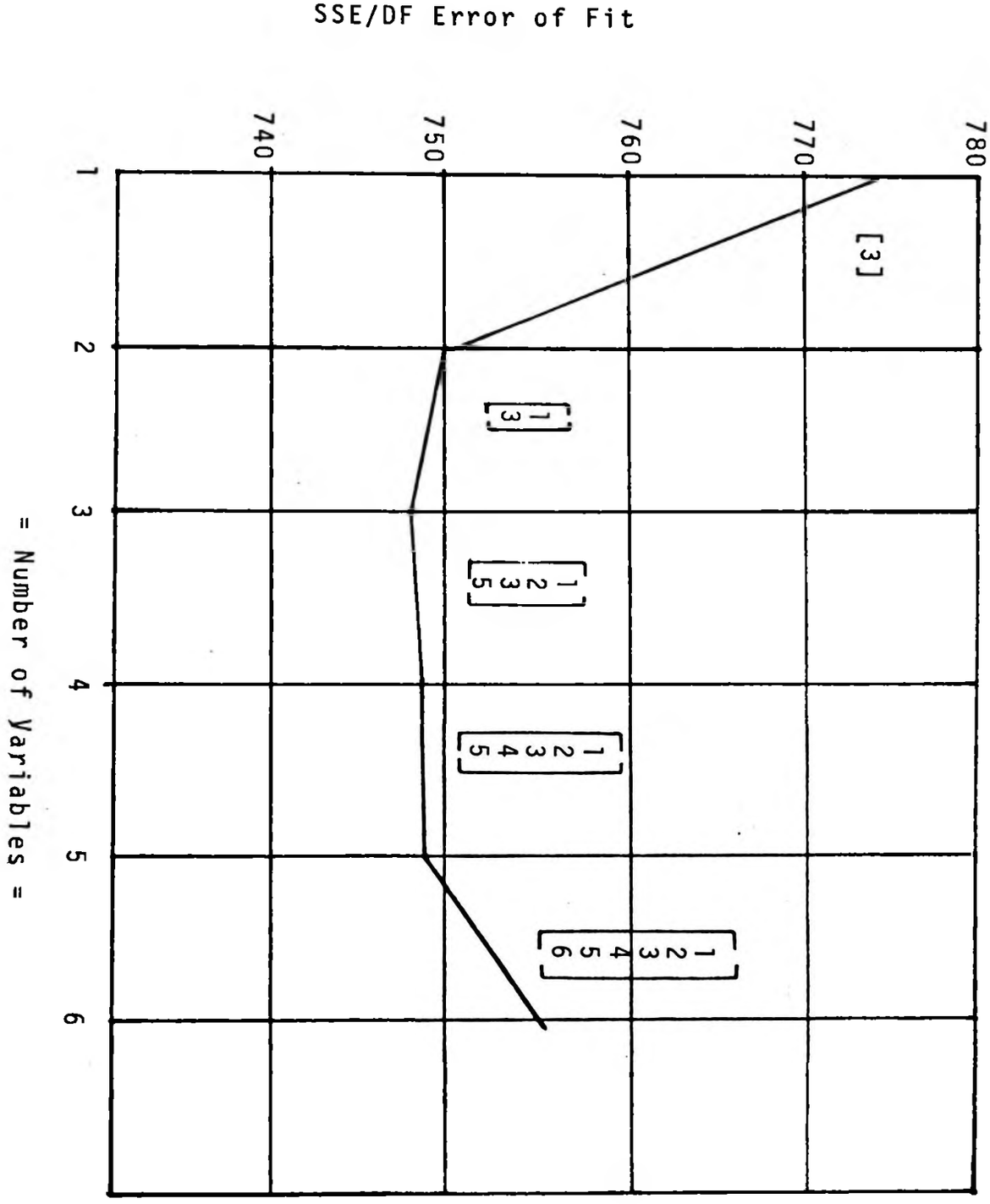
SSE/DF Error of Fit
780
770
760
750
740
1
2
3
4
5
6
= Number of Variables =
[3]
[1 3]
[1 2 3 5]
[1 2 3 4 5]
[1 2 3 4 5 6]

Fig. 1A

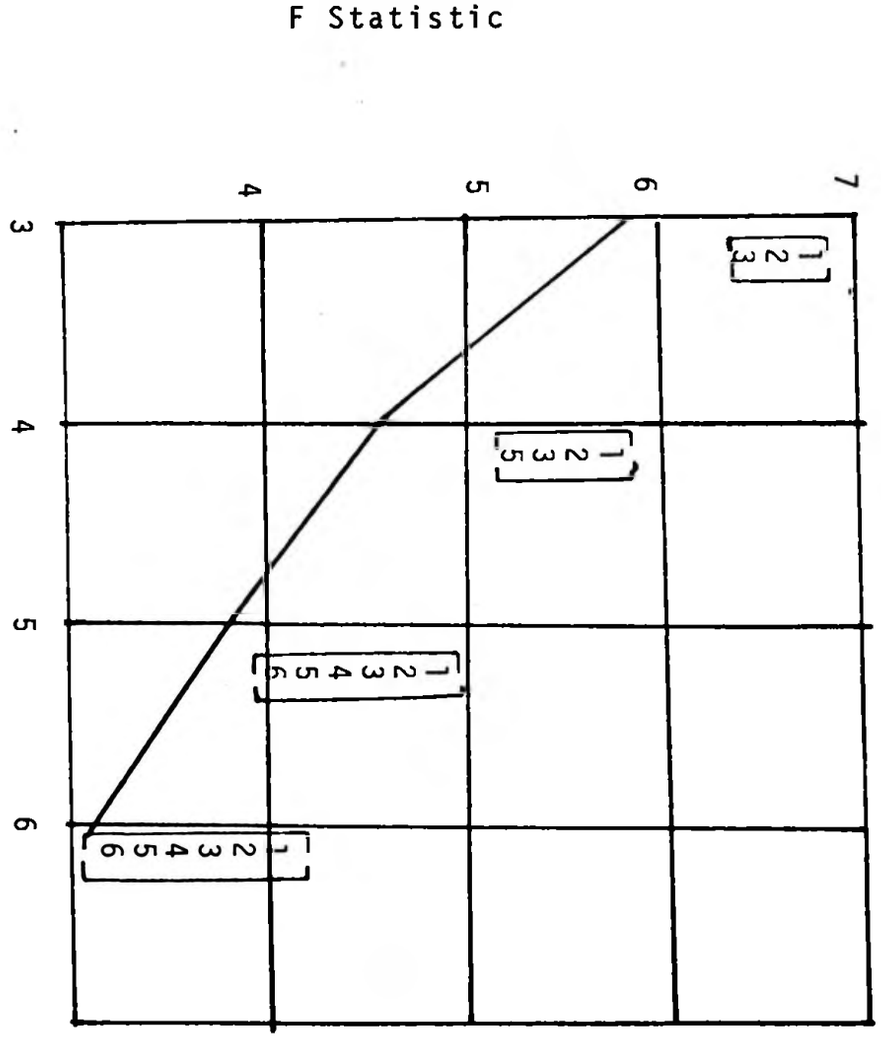
F Statistic
7
6
5
4
3
4
5
6
1 2 3
1 2 3 5
1 2 3 4 5 6
2 3 4 5 6 1
= Number of Variables =

Fig. 1B

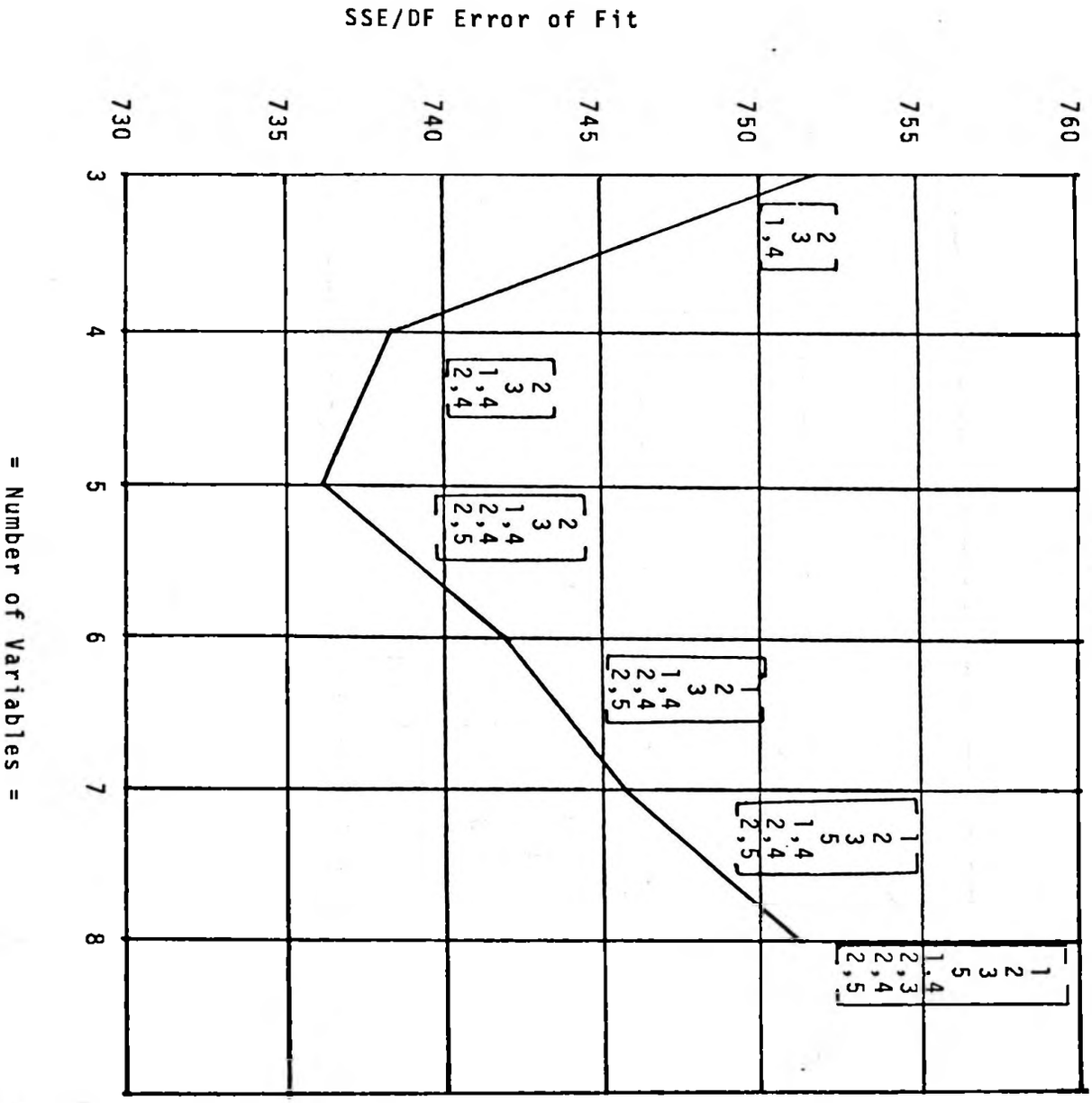
SSE/DF Error of Fit
760
755
750
745
740
735
730
3
4
5
6
7
8
= Number of Variables =
1 2 3 5 1,4 2,3 2,4 2,5
1 2 3 5 1,4 2,4 2,5
1 2 3 1,4 2,4 2,5
2 3 1,4 2,4 2,5
2 3 1,4 2,4
2 3 1,4

Fig. 2A

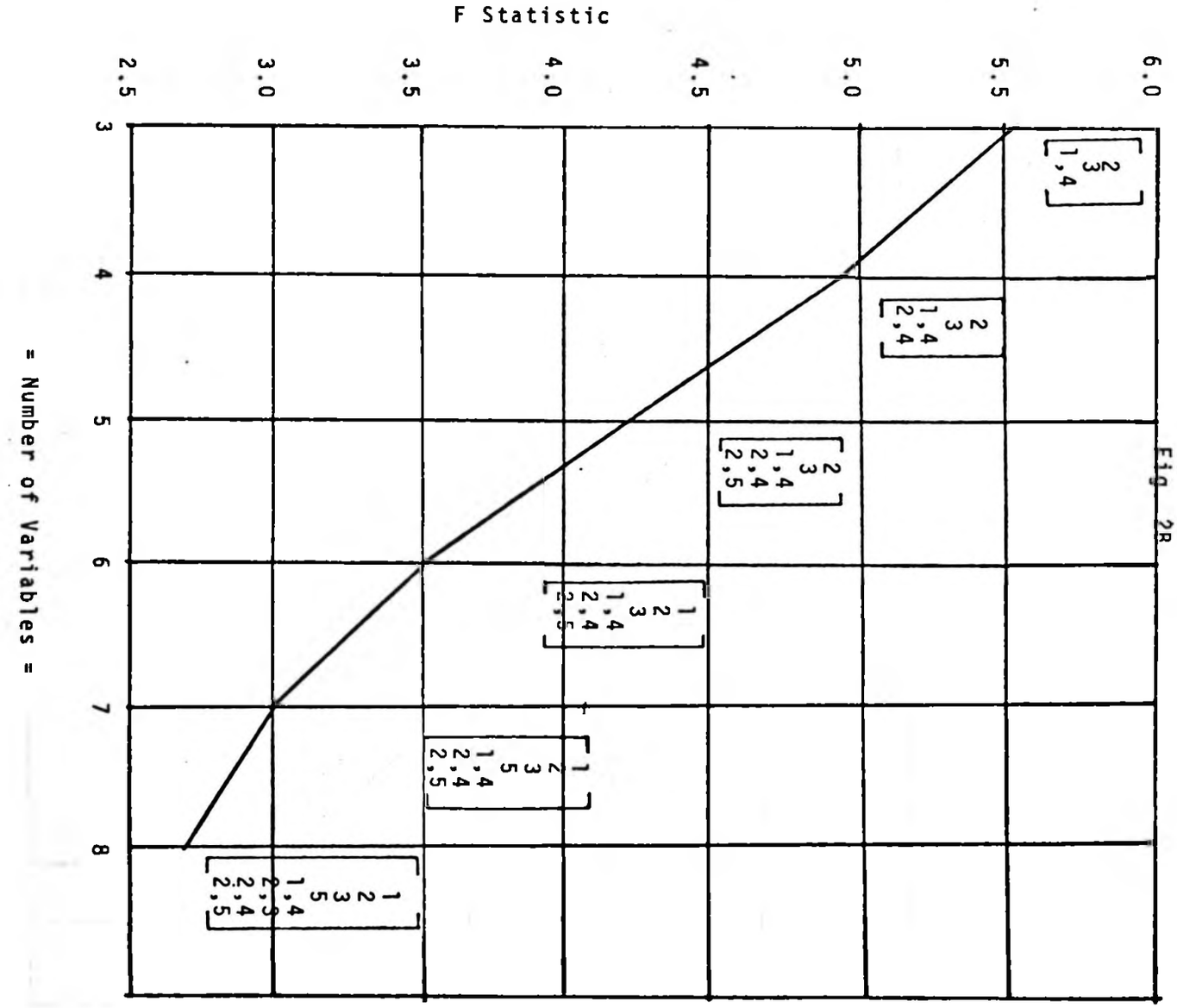
F Statistic
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
3
4
5
6
7
8
= Number of Variables =
2 3 1,4
2 3 1,4 2,4
2 3 1,4 2,4 2,5
1 2 3 1,4 2,4 2,5
2 3 5 1,4 2,4 2,5
1 2 3 5 1,4 2,3 2,4 2,5

Fig 2R

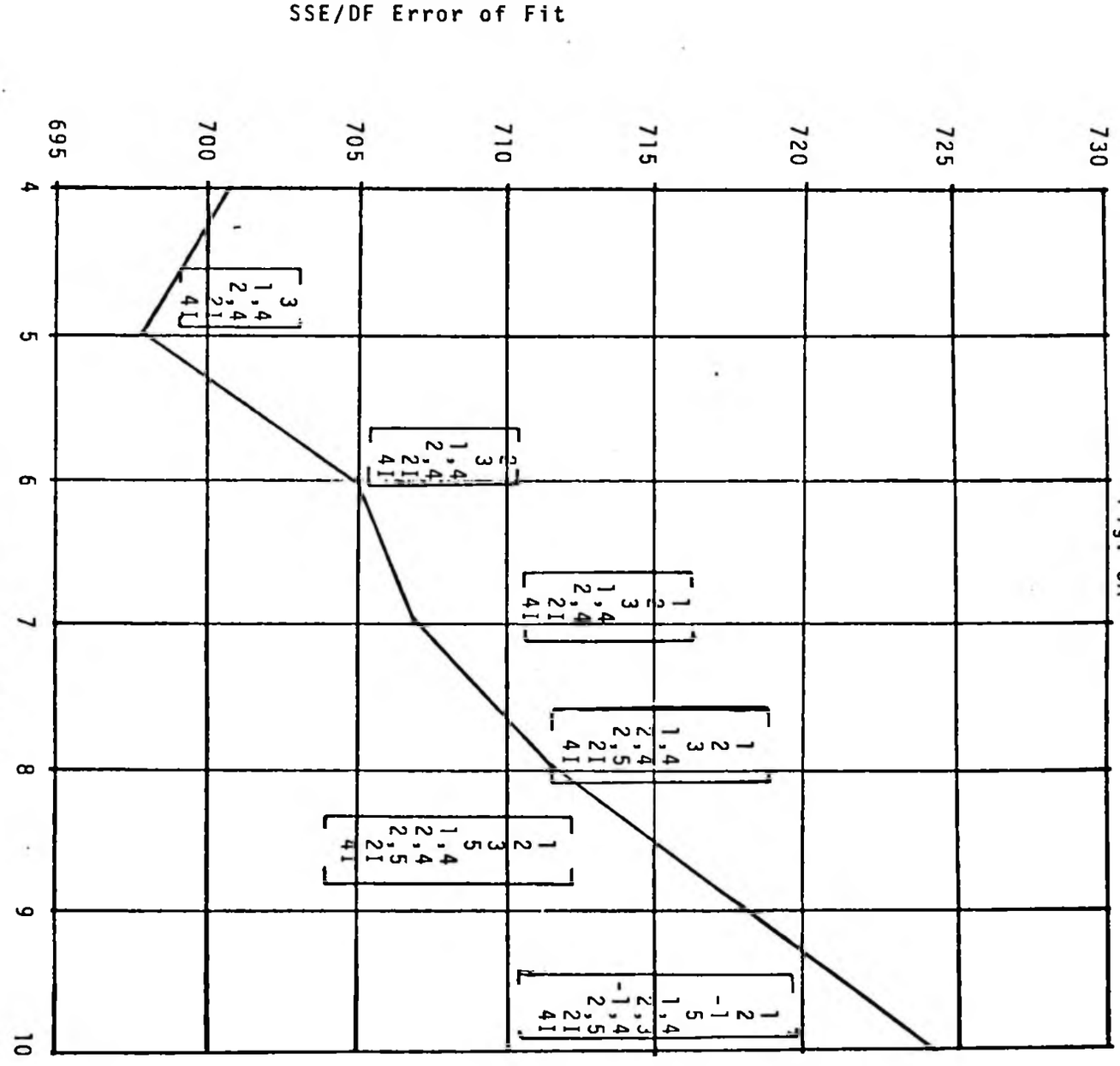
SSE/DF Error of Fit
695
700
705
710
715
720
725
730
4
5
6
7
8
9
10
= Number of Variables =

Fig. 3A

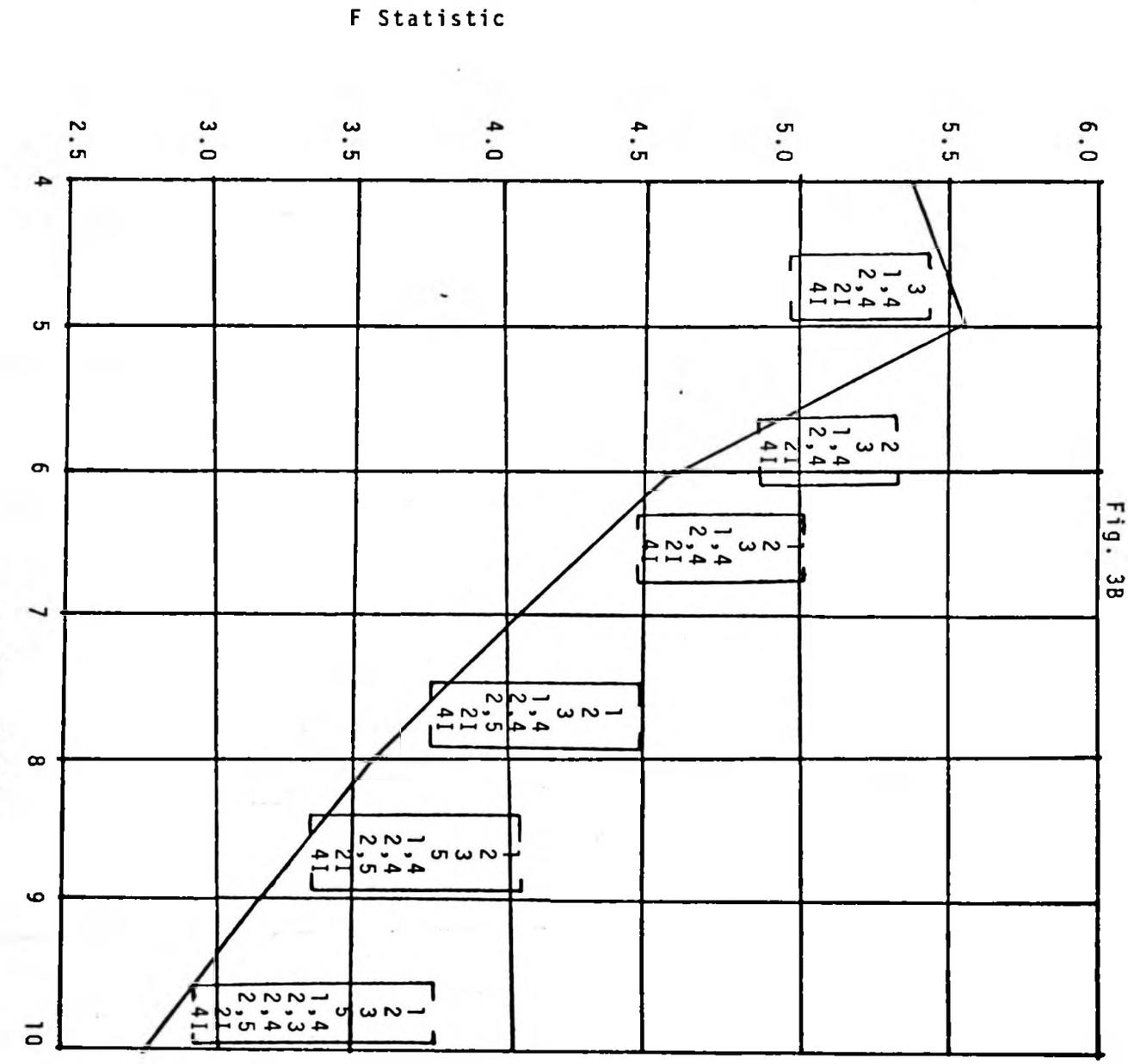
F Statistic
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
4
5
6
7
8
9
10
= Number of Variables =

Fig. 3B