MET MA 603:

SAS Programming and Applications

The Sort Procedure

Sorting Datasets

When a SAS dataset is created, the order of the observations is determined by when they are written to the output dataset (or if coming from an external file, the order follows that of the external file).

Sorting, or collating, refers to changing the order of the observations in a dataset.

Sorting is based on one or more variables in the dataset.

Unsorted		Sorted by Number		Sorted by Letter		Sorted by Number then Letter	
1	c	1	c	2	a	1	c
2	c	1	d	3	a	1	d
2	a	2	c	1	c	2	a
3	a	2	a	2	c	2	c
3	d	3	a	3	d	3	a
1	d	3	d	1	d	3	d

The Sort Procedure

The **Sort Procedure** orders a dataset according to one or more variables. The example below uses the input dataset "info". The out= option is used to create an output dataset containing the sorted data (if out= is left out, the sorted dataset will overwrite the input dataset).

The BY statement is required. It indicates which variables to sort by. The default sort order is ascending (low to high). Use the key word descending to sort from high to low.

Basic Examples of Proc Sort:

Proc Sort data=info out=info_sorted ;
by num ;
run ;

Proc Sort data=info out=info_sorted ;
by descending num ;
run ;

Practice

The Accounts.sas7bdat dataset has bank account information. Sort the dataset by account number and LastName. Do not overwrite the original dataset!

Do the results look as you would expect?

Rules for Character and Numeric Data

The sort order for numeric data is intuitive, since numbers already have an established order on the real number line. Since Dates are represented by numbers, they follow the rules of numeric sorting.

Character data is more complicated, since it has to account for letters, digits, and symbols.

Digits come before letters, and capital letters come before lower-case letters.

By convention, missing data always comes first in an ascending sort (and last in a descending sort).

Sort Sequence

Sort Sequence is the set of rules used for sorting character data is. Different languages and computer systems follow different rules.

When SAS is run on Windows, the default sorting sequence for character data is based on the ASCII code.

ASCII	Hex	Symbol	ASCII	Hex	Symbol	ASCII	Hex	Symbol	ASCII	Hex	Symbol
0	0	NUL	16	10	DLE	32	20	(space)	48	30	0
1	1	SOH	17	11	DC1	33	21	!	49	31	1
2	2	STX	18	12	DC2	34	22	"	50	32	2
3	3	ETX	19	13	DC3	35	23	#	51	33	3
4	4	EOT	20	14	DC4	36	24	$	52	34	4
5	5	ENQ	21	15	NAK	37	25	%	53	35	5
6	6	ACK	22	16	SYN	38	26	&	54	36	6
7	7	BEL	23	17	ETB	39	27	'	55	37	7
8	8	BS	24	18	CAN	40	28	(	56	38	8
9	9	TAB	25	19	EM	41	29	)	57	39	9
10	A	LF	26	1A	SUB	42	2A	*	58	3A	:
11	B	VT	27	1B	ESC	43	2B	+	59	3B	;
12	C	FF	28	1C	FS	44	2C	,	60	3C	<
13	D	CR	29	1D	GS	45	2D	-	61	3D	=
14	E	SO	30	1E	RS	46	2E	.	62	3E	>
15	F	SI	31	1F	US	47	2F	/	63	3F	?

ASCII	Hex	Symbol	ASCII	Hex	Symbol	ASCII	Hex	Symbol	ASCII	Hex	Symbol
64	40	@	80	50	P	96	60	`	112	70	p
65	41	A	81	51	Q	97	61	a	113	71	q
66	42	B	82	52	R	98	62	b	114	72	r
67	43	C	83	53	S	99	63	c	115	73	s

Sort Sequence Options

The option **SORTSEQ=LINGUISTIC** changes the collation rules to be based on the language that was selected during SAS installation, rather than on ASCII code.

With SORTSEQ=LINGUISTIC, SAS will not differentiate between capital and lowercase letters when sorting.

Sometimes, we store digits as character data, but we want to sort them in numeric order (so that 11 comes before 100, for example). This is done using the additional modification numeric_collation=on.

```
Proc Sort data=txt_info out=txt_info_sorted
     SORTSEQ=LINGUISTIC (numeric_collation=on);
by txt ;
run ;
```

Eliminating Duplicate Observations

Sometimes observations have the same values for one or more variables. The sort option **NODUPKEY** eliminates observations that have the same values for all of the variables listed in the BY statement.

Use the option DUPOUT=dataset to create a dataset which contains all observations that were eliminated due to being duplicates.

Practice

The Accounts.sas7bdat dataset has bank account information. Sort the dataset by account number, with the earliest account (the one with the smallest account number) coming first, and the newest account (the one with the biggest account number) coming last. Some accounts have more than one name on the account – eliminate one of the duplicate records and create a file that contains the duplicate records.

Next, sort the duplicate record dataset by LastName in alphabetical order (that is, instruct SAS to ignore case).

Readings

- Textbook section 4.3, 4.4