CSC423 Data Analysis and Regression
SAS Procedures for Linear Regression

Suppose that the goal of the statistical analysis is to predict future values of a response variable Y from one or more independent variables X's that affect the changes in Y. We'll now consider only those cases where the association between the Y variable and the X variables is <u>linear.</u>

PROC GPLOT
The PROC GPLOT creates a scatter plot of point for two variables.

PROC GPLOT DATA=*dataset-name;*
PLOT*y-variable***x-variable<=third-variable>;* ← *y-variable* is plotted on the y-axis
RUN; & *x-variable* is plotted on the x-axis

If the plot statement is y-*variable***x-variable=third-variable*;

then the values of the two variables are plotted against a third classification variable, that should be a character variable. Points on the graph will be displayed with different colors according to the *third-variable.*

The **symbol** option can be used to define alternative characters (instead of a dot) to be plotted on the graph. More than one symbol can appear.

```
SYMBOL1 VALUE=plus;
PROC GPLOT;
PLOT yvar*xvar;
RUN;
```

PROC CORR
The CORR procedure is a statistical procedure for numeric random variables that computes the Pearson correlation coefficient. The default correlation analysis includes descriptive statistics, Pearson correlation statistics, and probabilities for each analysis variable. For each pair of variables, the computed probability is the p-values for the test on the correlation coefficient being equal to zero.

PROC CORR;
BY *variable*; ←optional: *variable* is a group variable that classifies the observations.
VAR *variable(s)*; ← *variables* to correlate – produces correlation matrix for the listed variables.

PROC REG

It is a general-purpose procedure for linear regression. PROC REG provides the following capabilities:

- multiple MODEL statements;
- nine model-selection methods;
- tests of linear hypotheses on model parameters;
- model diagnostics; predicted values, residuals, studentized residuals, confidence limits, influence statistics and correlation;
- plots
    - plot model fit summary statistics and diagnostic statistics
    - produce normal probability-probability (P-P) plots for statistics such as residuals
    - specify special options to plot confidence intervals, and prediction intervals
    - display the fitted model equation, summary statistics, and reference lines on the plot
    - control the graphics appearance with PLOT statement options and with global graphics statements including the TITLE, FOOTNOTE, NOTE, SYMBOL, and LEGEND statements.

The PROC REG statement is required. To fit a model to the data, you must specify the MODEL statement.

PROC REG <DATA=*dataset-name*> < /*options*>;
MODEL *y-variable=x-variable(s);*
PLOT *yvariable*xvariable* <=*symbol*>
        < ...*yvariable*xvariable*> <=*symbol*> < / options > **;**
OUTPUT < OUT**=***SAS-data-set* > *keyword=names*
        < ... *keyword=names* > **;**
RUN;

1. The CORR option in PROC REG computes the correlation matrix for all the variables listed in MODEL.

    PROC REG <DATA = dataset-name> / CORR;
    MODEL y-variable=x-variables;
    RUN:

2. The PLOT statement

    More than one *yvariable*xvariable* pair can be specified to request multiple plots. The *yvariables* and *xvariables* can be
    - any variables specified in the VAR or MODEL statement before the first RUN statement
    - *keyword***.**, where *keyword* is a regression diagnostic statistic available in the OUTPUT statement – note the period after the keyword.
      For example,
      plot predicted.*residual.;

generates one plot of the predicted values versus the residuals for each dependent variable in the MODEL statement. These statistics can also be plotted against any of the variables in the VAR or MODEL statements. Possible keywords are:
Predicted. (or pred. or p.) = predicted values;
Residual. (or r.) = residuals;
Student. = studentized residuals;
Npp. = normal probability plot;

Specialized plots are requested with special options. The CONF option plots the 95% confidence intervals for the new predicted values Y, while the PRED option plots the 95% prediction intervals.

PROC REG;
MODEL yvar=xvar1 xvar2 xvar3;
PLOTyvar*xvar/nostat; ← draw scatter plot and regression line
PLOT student.*xvar1 student.*predicted.; ← residual plots
PLOT npp.*residual.; ← probability plot for the residuals
PLOT yvar*xvar/CONF; ← draw scatter plot & upper and lower confidence bounds
PLOT yvar*xvar/PRED; ← draw scatter plot & upper and lower prediction bounds.
RUN;

Example – CPU usage.
A study was conducted to examine what factors affect the CPU usage. A set of 38 processes written in a programming language was considered. For each program, data were collected on the CPU usage (`time`) in seconds of time, and the number of lines (`line`) in thousands generated by the program execution. The data file contains data on several variables. We'll restrict our attention to the analysis of the relationship between the two variables above.
This is the SAS code to analyze the data:

```
data cpu;
infile "cpudat.txt"; input time line step device;
linet=line/1000;
label time="CPU time in seconds" line="lines in program execution"
step="number of computer programs" device="mounted devices"
linet="lines in program (in thousands)";

/* scatter plot of time vs line number; */
symbol1 value=dot;
proc gplot;
plot time*linet;
run;

/* produce correlation matrix*/
proc corr;
run;
```

```
/* produce regression analysis: fit regression model,
compute model diagnostics & draw residual plots */

proc reg;
model time=linet;
plot time*linet/nostat pred;
plot (residual. student.)*predicted./nostat;
plot student.*linet/nostat;
plot npp.*student./nostat ;
run;
```

Output

The CORR Procedure

5 Variables: time line step device linet

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
time	38	0.15710	0.13129	5.96980	0.01960	0.46780	CPU time in seconds
line	38	3162	3961	120154	102.00000	14872	lines in program execution
step	38	5.94737	5.13571	226.00000	1.00000	15.00000	number of computer programs
device	38	2.78947	2.24401	106.00000	0	8.00000	mounted devices
linet	38	3.16195	3.96094	120.15400	0.10200	14.87200	lines in program (thousand)

Pearson Correlation Coefficients, N = 38
Prob > |r| under H0: Rho=0

	time	line	step	device	linet
time	1.00000	0.89802	0.90632	0.17841	0.89802
CPU time in seconds		<.0001	<.0001	0.2839	<.0001
line	0.89802	1.00000	0.80355	-0.13825	1.00000
lines in program execution	<.0001		<.0001	0.4078	<.0001
step	0.90632	0.80355	1.00000	0.16083	0.80355
number of computer programs	<.0001	<.0001		0.3347	<.0001
device	0.17841	-0.13825	0.16083	1.00000	-0.13825
mounted devices	0.2839	0.4078	0.3347		0.4078
linet	0.89802	1.00000	0.80355	-0.13825	1.00000
lines in program (thousand)	<.0001	<.0001	<.0001	0.4078	

The SAS System

The REG Procedure

Model: MODEL1

Dependent Variable: time CPU time in seconds

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.51429	0.51429	149.99	<.0001
Error	36	0.12343	0.00343		
Corrected Total	37	0.63772			

Root MSE	0.05856	R-Square	0.8064
Dependent Mean	0.15710	Adj R-Sq	0.8011
Coeff Var	37.27272		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.06298	0.01222	5.16	<.0001
linet	lines in program	1	0.02976	0.00243	12.25	<.0001

The UNIVARIATE Procedure

Variable: resid (Residual)

Moments

N	38	Sum Weights	38
Mean	0	Sum Observations	0
Std Deviation	0.05775874	Variance	0.00333607
Skewness	0.54076603	Kurtosis	-0.7539768
Uncorrected SS	0.12343465	Corrected SS	0.12343465
Coeff Variation	.	Std Error Mean	0.0093697

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	0.05776
Median	-0.02608	Variance	0.00334
Mode	.	Range	0.21021
		Interquartile Range	0.07808

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.932438	Pr < W	0.0240
Kolmogorov-Smirnov	D	0.197791	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.176192	Pr > W-Sq	0.0099
Anderson-Darling	A-Sq	0.999382	Pr > A-Sq	0.0113