Sophisticated, flexible routines for statistical model building.

Applications of the Hewlett-Packard System 45

Over the last several years, HP has worked with a local university to develop a comprehensive library of statistics programs. These programs, together with HP's experience developing desktop computing systems, give you the most advanced solutions available on the market today to your statistical problems.

Now, Hewlett-Packard offers you a Regression Analysis software package with unprecedented flexibility for handling your data. It is designed to run on HP's System 45 — a powerful, friendly desktop computer which features a typewriter-like keyboard, a CRT screen, a built-in, high-speed magnetic tape drive and an enhanced BASIC language processor — all integrated into a single, easy-to-use package.

The Regression Analysis programs give you a choice of three regression techniques and a collection of routines for entering, editing and transforming your data for later analysis. They provide, as well, an extensive set of summary statistics routines. Besides all these, there is a program for analyzing residuals from any of the regression programs.

A variety of routines for extensive data manipulation.

An important part of this software package is the collection of routines for entering, editing, transforming and otherwise manipulating your data to prepare it for later analysis. Or, if preliminary analysis indicates the data should be modified in some way, these routines can be used, followed by a second analysis.

Uniform data structure. The data array in this program is compatible with all other statistics packages available from Hewlett-Packard. This means you need not reenter your data when you want to analyze it with this program or any other HP program. In fact, your data can be collected independent of any program, stored on tape or disk and then analyzed using any of the regression programs described above.

Summary statistics. The following statistical routines for up to 1500 data points are included:

Basic Statistics:
- Mean
- Variance
- Standard Deviation
- Second Moment
- Skewness
- Kurtosis
- t
- Standard Error
- Confidence Interval for Mean
- Coefficient of Variation
- Correlation Coefficient for each pair of variables

Order Statistics:
- Range
- Midrange
- Median
- Percentiles
- Tukey's Middlemeans

Transformation routines. These useful routines allow you to choose more sophisticated models than simple linear regression. You create new variables by using any of these 16 transformations:

1. a X^b + c
2. a log(bX) + c
3. a ln(bX) + c
4. a exp(bX) + c
5. a (b^{cx})
6. a cos(bX) + c
7. a sin(bX) + c
8. a + arcsin(bX) + c
9. aX + bY + c
10. $aX^b Y^c$
11. a log (bX + cY)
12. a ln (bX + cY)
13. a cos (bX + cY)
14. a sin (bX + cY)
15. PROUND (X,a)
16. DROUND (X,a)

You may also enter a transformation routine of your own definition.

A choice of regression methods.

The program package gives you a choice of three regression techniques to analyze your data: polynomial regression (up to degree nine); multiple linear regression (up to 20 variables); and multiple regression with four variable selection procedures.

Polynomial regression. This program fits a polynomial of the form $Y=a_0 + a_1x + a_2x^2 +...+ a_nx^n$ to your data, using the least squares method to determine the coefficients. Program output includes:

- mean, variance, standard deviation and coefficient of variation of both variables,
- the correlation coefficient and standard error of estimate,
- a preliminary analysis of variance table to assist in selecting the degree of polynomial desired,
- a final analysis of variance table which includes the F-ratio associated with each degree,
- the computed regression coefficients together with their associated standard errors and t-values,
- the confidence interval for each coefficient.

You can also obtain a scatter plot of your data, together with plots of any of the polynomials you have chosen to examine. Figure 1 draws a sample data set with the results from polynomial regression.

Multiple linear regression. This program fits a model of the form $Y=a_0+a_1x_1+a_2x_2+...+a_mx_m$ to your data, using the least squares method. Output includes a matrix which gives the correlation between each pair of variables; the coefficient of determination and standard error of estimate; an analysis of variance table including the F-ratio associated with each variable; and the regression coefficients with their associated standard errors, t-values and confidence intervals.

Bus Passenger Service Time

OBS#	Variable # 1	Variable # 2
1	1.00000	1.40000
2	1.00000	2.80000
3	1.00000	3.00000
4	1.00000	1.80000
5	1.00000	2.60000
6	2.00000	4.70000
7	2.00000	8.00000
8	2.00000	3.00000
27	13.00000	25.20000
28	17.00000	33.50000
29	19.00000	33.70000

****
POLYNOMIAL REGRESSION ON DATA SET:
Bus Passenger Service Time
****

Where: Dependent variable = TIME
Independent variable = NUMBER

VARIABLE	N	MEAN	VARIANCE	STANDARD DEVIATION	COEFFICIENT OF VARIATION
NUMBER	29	6.00000	23.07143	4.80327	80.05454
TIME	29	12.10690	82.79495	9.09917	75.15694

CORRELATION = .980337144461

Selected degree of regression = 1
R-SQUARED = .961060916526
STANDARD ERROR OF ESTIMATE = 1.82848718332

ANOV

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F-VALUE
TOTAL	28	2318.25862		
REGRESSION	1	2227.98776	2227.98776	666.39
X^1	1	2227.98776	2227.98776	666.39
RESIDUAL	27	90.27087	3.34337	

VARIABLE	REGRESSION COEFFICIENTS STD. FORMAT	E-FORMAT	STANDARD ERROR REG. COEFFICIENT	T-VALUE
'CONSTANT'	.96417	.964172393400E+00	.54919	1.76
X^1	1.85712	.185712074305E+01	.07194	25.81

	COEFFICIENT	95 % CONFIDENCE INTERVAL LOWER LIMIT	UPPER LIMIT
'CONSTANT'	.96417	-.16293	2.09128
X^1	1.85712	1.70948	2.00477

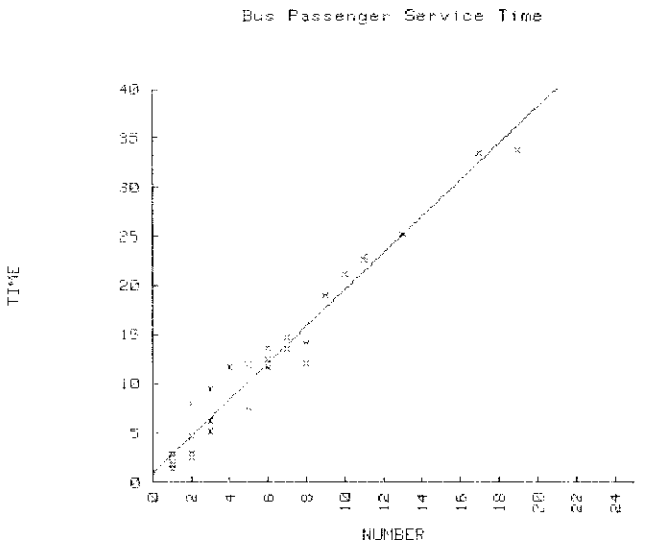

Figure 1. Sample polynomial regression showing the relationship between the number of passengers and service time.

```
**********************************************************************
                    STEPWISE REGRESSION ON
                          SAMPLE
**********************************************************************

Dependent variable:Water Use
Independent variable(s) =   Temp(C)
                            Production
                            Days
                            Payroll

Tolerance = .01
F-value for inclusion = 4
F-value for deletion = 4

                        CORRELATION MATRIX

                Temp(C) Production       Days    Payroll  Water Use
Temp(C)       1.0000000  -.0923502   .2685939  -.1073726   .2503744
Production                1.0000000   .1057436   .9184717   .6308869
Days                                  1.0000000   .0318312  -.0887635
Payroll                                           1.0000000   .4134486
Water Use                                                     1.0000000

**********************************************************************

                F TO  PART        F TO        REGRESSION COEFFICIENTS       STD
#--VARIABLE     ENTER  CORR   TOL DELETE  STD.FORMAT            E-FORMAT     ERROR
 1.Temp(C)       1.00  .250 1.000
 2.Production    9.92  .631 1.000
 3.Days           .12  .089 1.000
 4.Payroll       3.09  .415 1.000

**********************************************************************
STEP NUMBER 1
VARIABLE'Production' ADDED
R-SQUARED =  .39801825079

                                AOV

SOURCE              DF      SUM OF SQUARES         MEAN SQUARE         F-VALUE
TOTAL               16       3861967.52900
REGRESSION           1       1537133.56850       1537133.56850            9.92
RESIDUAL            15       2324833.96050        154988.93123

STANDARD ERROR =  393.686836102

                F TO  PART        F TO        REGRESSION COEFFICIENTS       STD
#--VARIABLE     ENTER  CORR   TOL DELETE  STD.FORMAT            E-FORMAT     ERROR
 1.Temp(C)       2.66  .399  .991
 2.Production                          9.92       .03765  .976476057819E-01    .0310
 3.Days           .59  .202  .989
 4.Payroll       5.79  .541  .156

Constant =  2499.97697139

**********************************************************************
STEP NUMBER 2
VARIABLE'Payroll' ADDED
R-SQUARED =  .57420302157

                                AOV

SOURCE              DF      SUM OF SQUARES         MEAN SQUARE         F-VALUE
TOTAL               16       3861967.52900
REGRESSION           2       2217553.42436       1108776.71218            9.44
RESIDUAL            14       1644414.10465        117458.15033

STANDARD ERROR =  342.721680568

                F TO  PART        F TO        REGRESSION COEFFICIENTS       STD
#--VARIABLE     ENTER  CORR   TOL DELETE  STD.FORMAT            E-FORMAT     ERROR
 1.Temp(C)       3.10  .439  .988
 2.Production                         13.26       .24853  .248502359013E+00    .0683
 3.Days          1.83  .351  .962
 4.Payroll                             5.79    -23.70679 -.237067948772E+02   9.8498

Constant =  5177.19960983

Tolerance value too small and/or F-values insufficient to proceed.
```

Figure 2. Sample stepwise regression showing the effect of average external temperature, production yield, number of working days in the month, and number of employees on water use.

Multiple linear regression with four variable selection procedures. A limitation of multiple linear regression is that one or more of the independent variables may fail to explain a sufficient amount of the variation of the dependent variable. This program overcomes this limitation by allowing you to examine an analysis of variance table at each step of the routine and select the next variable to be deleted from those previously added. This is called a manual selection procedure. The program also includes three automatic procedures. Forward selection begins with no variables and adds variables until no more are significant. Backward selection begins with all variables and eliminates insignificant ones. Stepwise selection combines these by starting with no variables and then adding or deleting until all variables in the model are significant and the variables omitted are insignificant. Program output includes the correlation matrix, an analysis of variance table, and the regression coefficients with their associated standard errors. Figure 2 shows a sample data set using stepwise selection.

Residual analysis. The residual analysis program may be used after any regression program. It will automatically determine the differences between actual Y values and predicted Y values. This program prints a table of residuals and standardized residuals and plots the standardized residuals either against the observation number, or against any variable. Figure 3 shows the residual analysis of the data in Figure 2.

Here's how to get started with the HP Regression Analysis pack.

To order this software package, specify Part No. 09845-15010. The package includes a user instruction manual, Special Function key overlays, and a program cartridge. An additional Source Documentation manual, containing annotated program listings, tape file descriptions and definitions of variables may be ordered to accompany the software pack by specifying Part No. 09845-15012.

To run this software, you'll need the HP 9845A Desktop Computer with 29k bytes of user memory (Opt. 201). The following hardware may be ordered to enhance your system:

- The internal Thermal Line Printer (Opt. 500)
- The Graphics package (Opt. 370 and 700) or the HP 9872A Plotter for plotting
- The I/O ROM (Opt. 320) for entering data from external devices
- The HP 9885M/S Flexible Disk Drive or the HP 7900 Series Disc Drives for storing larger amounts of data.

OBS#	OBSERVED Y	PREDICTED Y	RESIDUAL	STAND.RES.	SIGNIF.
1	3373.00000	3590.06369	-217.06369	-.63335	
2	3110.00000	3141.55475	-31.55475	-.09207	
3	3180.00000	2951.51323	228.48677	.66668	
4	3293.00000	3182.87562	110.12438	.32132	
5	3390.00000	3791.88564	-401.88564	-1.17263	
6	4287.00000	3835.76492	451.23508	1.31662	
7	3852.00000	3586.10837	265.89163	.77582	
8	3966.00000	3674.51316	291.48684	.85044	
9	3532.00000	3385.08984	146.91016	.42866	
10	3614.00000	3783.51944	-169.51944	-.49468	
11	3896.00000	4169.20360	-273.20360	-.79716	
12	3437.00000	3434.24056	2.75944	.00805	
13	3324.00000	3410.90783	-86.90783	-.25358	
14	3214.00000	3767.12980	-553.12980	-1.61393	
15	4345.00000	3786.36312	558.63688	1.63000	
16	4936.00000	4471.67328	464.32672	1.35482	
17	3624.00000	3810.59316	-186.59316	-.54445	

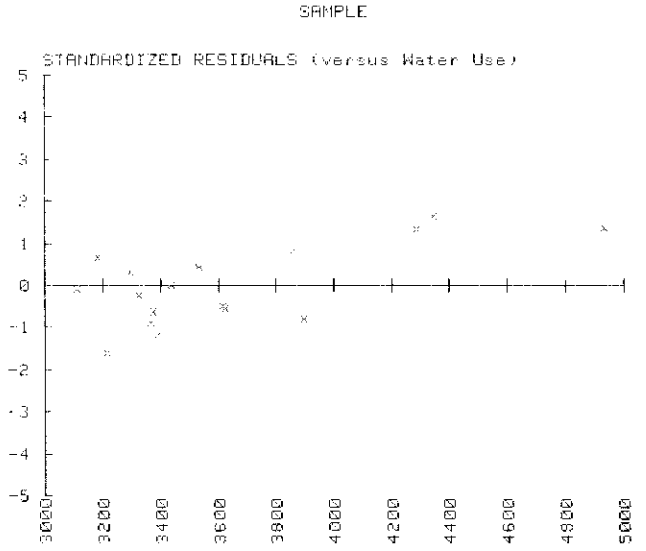

Figure 3. Residual table and plot using the data from Figure 2.

HEWLETT hp PACKARD

Sales and service from 172 offices in 65 countries.
Loveland, Colorado 80537.

Printed in U.S.A.
(11/77)10M

5953-0942(40)