Assessing Normality – The Microwave Radiation Dataset

- SAS code is posted as microwave.sas

/* Display a box plot */

data set1; set set1;
  w=1; run;

PROC BOXPLOT DATA=set1;
  PLOT x*w / BOXSTYLE=SCHEMATICID;
  Title "Microwave Radiation";
run;
/* Check data quality: compute summary statistics and tests of normality */

PROC UNIVARIATE DATA=set1 NORMAL;
   VAR x;
RUN;
......

Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W</td>
<td>0.85793</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D</td>
<td>0.230299</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq</td>
<td>0.366602</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq</td>
<td>2.101398</td>
</tr>
</tbody>
</table>
Extreme Observations

<table>
<thead>
<tr>
<th>Value</th>
<th>Obs</th>
<th>Value</th>
<th>Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>19</td>
<td>0.3</td>
<td>38</td>
</tr>
<tr>
<td>0.01</td>
<td>13</td>
<td>0.3</td>
<td>39</td>
</tr>
<tr>
<td>0.02</td>
<td>35</td>
<td>0.3</td>
<td>41</td>
</tr>
<tr>
<td>0.02</td>
<td>17</td>
<td>0.4</td>
<td>20</td>
</tr>
<tr>
<td>0.02</td>
<td>12</td>
<td>0.4</td>
<td>40</td>
</tr>
</tbody>
</table>
/* Compute a Q-Q plot */

PROC RANK DATA=set1 NORMAL=blom OUT=set1;
    VAR x; RANKS q;
    RUN;

PROC PRINT DATA=set1; run;

....... 

/* Create high quality graphs with SASGRAPH */

goptions targetdevice=ps300 rotate=portrait;

/* Windows users can use the following options 
goptions device=WIN target=ps rotate=portrait; */

/* Specify features of the plot */
PROC GPLOT DATA=set1;
   PLOT x*q / vaxis=axis2 haxis=axis1;
TITLE1 ls=1.5in H=3.0 F=swiss "NORMAL PROBABILITY PLOT";
TITLE2 H=2.3 F=swiss "Microwave Radiation";
footnote ls=1.0in;
   RUN;
NORMAL PROBABILITY PLOT

Microwave Radiation

Ordered Data vs. Standard Normal Quantiles
/* Compute natural logarithm of radiation values */

data set1; set set1;
   logx = log(x);
run;

PROC BOXPLOT DATA=set1;
   PLOT logx*w / BOXSTYLE=SCHEMATICID;
   Title "Log-Radiation";
   Title2 " ";
run;
/* Compute tests of normality */

PROC UNIVARIATE DATA=set1 NORMAL;
   VAR logx;
RUN;

/* Compute a Q-Q plot */

PROC RANK DATA=set1 NORMAL=blom OUT=set1;
   VAR logx; RANKS q;
RUN;
The UNIVARIATE Procedure
Variable: logx

Moments

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>42</td>
<td>Sum Weights</td>
<td>42</td>
</tr>
<tr>
<td>Mean</td>
<td>-2.3841075</td>
<td>Sum Observations</td>
<td>-100.13251</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>0.90507309</td>
<td>Variance</td>
<td>0.81915731</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.6885442</td>
<td>Kurtosis</td>
<td>0.38678213</td>
</tr>
<tr>
<td>Uncorrected SS</td>
<td>272.312122</td>
<td>Corrected SS</td>
<td>33.5854495</td>
</tr>
<tr>
<td>Coeff Variation</td>
<td>-37.962764</td>
<td>Std Error Mean</td>
<td>0.13965581</td>
</tr>
</tbody>
</table>

Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W 0.938782</td>
<td>Pr &lt; W 0.0259</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D 0.152115</td>
<td>Pr &gt; D 0.0158</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq 0.154638</td>
<td>Pr &gt; W-Sq 0.0208</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq 0.890959</td>
<td>Pr &gt; A-Sq 0.0218</td>
</tr>
</tbody>
</table>
PROC GPLOT DATA=set1;
   PLOT logx*q / vaxis=axis2 haxis=axis1;
TITLE1 ls=1.5in H=3.0 F=swiss "NORMAL PROBABILITY PLOT";
TITLE2 H=2.3 F=swiss "Logarithm of Radiation";
footnote ls=1.0in;
RUN;
Assessing bivariate and multivariate normality

• If sample observations $X_1, ..., X_n$ come from a $N_p(\mu, \Sigma)$, we know that

$$\delta^2_i = (x_i - \mu)' \Sigma^{-1} (x_i - \mu) \sim \chi^2_p.$$ 

• By substituting $\bar{x}, S$ for the population mean vector, we can estimate the sample squared distances $d^2_j$. For large $n - p$ (at least 25), the $d^2_j$ should behave approximately like independent $\chi^2_p$ random variables.

• A $\chi^2$ plot is similar to a Q-Q plot and can be computed for the sample squared distances.
Chi Square Plot

• First order the squared distances from smallest to largest to get the $d^2_{(i)}$.

• Next compute the probability levels as before: $\hat{p} = (i-0.5)/n$.

• Then compute the $n$ chi-square quantiles $q_{c,p}(\hat{p}_i)$ for the $\chi^2_p$ distribution.

• Finally plot the pairs $(d^2_{(i)}, q_{c,p}(\hat{p}_i))$ and check if they approximately fall on a straight line.

• For a $\chi^2$ distribution with $\nu$ degrees of freedom, the function to find the chi-square quantiles in SAS is $q_{i,\nu} = \text{cinv}(\hat{p}_i, \nu)$.
Stiffness of boards: Example 4.14

- Four measures of stiffness on each of \( n = 30 \) boards were obtained.

- Data are shown on Table 4.3 in book, including the 30 sample squared distances.

- The 30 probability levels are computed as
  \[
  \frac{1 - 0.5}{30} = 0.017, \quad \frac{2 - 0.5}{30} = 0.05, \ldots, \quad \frac{30 - 0.5}{30} = 0.983.
  \]

- Quantiles can be calculated using the cinv function in SAS, for \( p = 4 \) degrees of freedom and the 30 probability levels computed above.
Example (cont’d)
Board Stiffness Data

Posted in the data folder on the course web page as board.stiffness.dat

```
1  1889  1651  1561  1778
2  2493  2048  2087  2197
3  2119  1700  1815  2222
4  1645  1627  1110  1533
 .  .  .  .  .
 .  .  .  .  .
 .  .  .  .  .
 .  .  .  .  .
28  1655  1675  1414  1597
29  2326  2301  2065  2234
30  1490  1382  1214  1284
```
/* This program computes summary statistics and tests of univariate normality for the board stiffness data in example 4.14. It also creates a chi-square probability plot to test for multivariate normality. PROC IML in SAS is used to generate points for the chi-square probability plot, to compute a sample covariance matrix, the sample mean vector, sample correlation matrix, partial correlation matrix, and related t-tests. This program is posted in the SAS folder of the course web page as board.stiffness.sas */
data set1;
    INFILE 'c:\stat501\board.stiffness.dat';
    INPUT Board X1 X2 X3 X4;
run;

/* First print the data file */

PROC PRINT DATA=set1;
    Title "Board Stiffness Data";
    RUN;
### Board Stiffness Data

| B | o | a | b r X X X X Q Q Q Q | s d 1 2 3 4 Z 1 2 3 4 |
|---|---|---|-------------------|----------------|---|---|---|---|
| 1 | 1 | 1889 | 1651 | 1561 | 1778 | 1 | 0.12462 | -0.12462 | 0.29421 | 0.38198 |
| 2 | 2 | 2493 | 2048 | 2087 | 2197 | 1 | 1.60982 | 1.02411 | 1.60982 | 1.17581 |
| 3 | 3 | 2119 | 1700 | 1815 | 2222 | 1 | 0.89292 | 0.20866 | 1.17581 | 1.36087 |
| . | . | . | . | . | . | . | . | . | . |
| 27 | 27 | 2168 | 1896 | 1701 | 1834 | 1 | 1.02411 | 0.66800 | 0.89292 | 0.47279 |
| 28 | 28 | 1655 | 1675 | 1414 | 1597 | 1 | -0.89292 | -0.04144 | -0.20866 | -0.29421 |
| 29 | 29 | 2326 | 2301 | 2065 | 2234 | 1 | 1.36087 | 1.60982 | 1.36087 | 1.60982 |
| 30 | 30 | 1490 | 1382 | 1214 | 1284 | 1 | -1.36087 | -1.36087 | -1.17581 | -1.36087 |
/* You can make a scatterplot matrix and other data displays using the interactive data analysis option in SAS. Click on "Solutions" at the top of the SAS editing window.
Select "Analysis". Then click on "Interactive Data Analysis". Select the "Work" library and select the set1 data set. A spreadsheet containing the data will appear. Click on the "analyze" button that now appears at the top of the editing window and select scatter plot. Select the variables to appear on the plot. */
/* Compute correlations */

PROC CORR DATA=set1;
   VAR X1-X4;
RUN;
The CORR Procedure

4 Variables:   X1     X2     X3     X4

Simple Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>30</td>
<td>1909</td>
<td>330.10681</td>
<td>57273</td>
</tr>
<tr>
<td>X2</td>
<td>30</td>
<td>1750</td>
<td>318.60653</td>
<td>52486</td>
</tr>
<tr>
<td>X3</td>
<td>30</td>
<td>1509</td>
<td>303.17831</td>
<td>45274</td>
</tr>
<tr>
<td>X4</td>
<td>30</td>
<td>1725</td>
<td>322.84356</td>
<td>51749</td>
</tr>
</tbody>
</table>
Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1.00000</td>
<td>0.90840</td>
<td>0.89011</td>
<td>0.89794</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>0.90840</td>
<td>1.00000</td>
<td>0.78821</td>
<td>0.78810</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>0.89011</td>
<td>0.78821</td>
<td>1.00000</td>
<td>0.92310</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td>0.89794</td>
<td>0.78810</td>
<td>0.92310</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>
/* Display box plots */

DATA set1; SET set1; Z=1; run;

PROC BOXPLOT DATA=set1;
   PLOT (X1-X4)*Z / BOXSTYLE=SCHEMATICID;
run;

/* Check the quality of the data. Compute
   summary statistics and tests of normality
   for each variable */

PROC UNIVARIATE DATA=set1 NORMAL;
   VAR X1-X4;
RUN;
### Variable: X1

#### Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W 0.931138</td>
<td>Pr &lt; W 0.0526</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D 0.1197</td>
<td>Pr &gt; D &gt;0.1500</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq 0.093607</td>
<td>Pr &gt; W-Sq 0.1340</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq 0.579696</td>
<td>Pr &gt; A-Sq 0.1253</td>
</tr>
</tbody>
</table>

### Variable: X2

#### Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W 0.912744</td>
<td>Pr &lt; W 0.0175</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D 0.180222</td>
<td>Pr &gt; D 0.0141</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq 0.160401</td>
<td>Pr &gt; W-Sq 0.0175</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq 0.874727</td>
<td>Pr &gt; A-Sq 0.0226</td>
</tr>
</tbody>
</table>

206
### Variable: X3

Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>---p Value---</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W</td>
<td>0.932577</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D</td>
<td>0.125836</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq</td>
<td>0.086811</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq</td>
<td>0.609371</td>
</tr>
</tbody>
</table>

### Variable: X4

Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>---p Value---</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W</td>
<td>0.961272</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D</td>
<td>0.101426</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq</td>
<td>0.06725</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq</td>
<td>0.411431</td>
</tr>
</tbody>
</table>
PROC RANK DATA=set1 NORMAL=BLOM OUT=set1;
   VAR X1-X4; RANKS Q1-Q4;
   RUN;

PROC PRINT DATA=set1; run;

     B
     o
     a
     b r X X X X Q Q Q Q s d 1 2 3 4 Z 1 2 3 4

     1  1 1889 1651 1561 1778 1  0.12462 -0.12462  0.29421  0.38198
     2  2 2493 2048 2087 2197 1  1.60982  1.02411  1.60982  1.17581
     3  3 2119 1700 1815 2222 1  0.89292  0.20866  1.17581  1.36087
     4  4 1645 1627 1110 1533 1 -1.02411 -0.38198 -1.60982 -0.52024
goptions device=WIN target=ps rotate=portrait;

/* Specify features of the plot */

axis1 label=(f=swiss h=2.5
    "Standard Normal Quantiles")
  ORDER = -2.5 to 2.5 by .5
  value=(f=triplex h=1.6)
  length= 5.5in;

axis2 label=(f=swiss h=2.5 a=90 r=0
    "Ordered Data ")
  value=(f=triplex h=2.0)
  length = 5.0in;

SYMBOL1 V=CIRCLE H=2.0 ;
PROC GPLOT DATA=set1;
    PLOT X1*Q1 / vaxis=axis2 haxis=axis1;
TITLE1 ls=1.5in H=3.0 F=swiss "NORMAL PROBABILITY PLOT";
TITLE2 H=2.3 F=swiss "X1";
footnote ls=1.0in;
    RUN;

PROC GPLOT DATA=set1;
    PLOT X2*Q2 / vaxis=axis2 haxis=axis1;
TITLE1 ls=1.5in H=3.0 F=swiss "NORMAL PROBABILITY PLOT";
TITLE2 H=2.3 F=swiss "X2";
footnote ls=1.0in;
    RUN;
PROC GPLOT DATA=set1;
   PLOT X3*Q3 / vaxis=axis2 haxis=axis1;
TITLE1 ls=1.5in H=3.0 F=swiss "NORMAL PROBABILITY PLOT";
TITLE2 H=2.3 F=swiss "X3";
footnote ls=1.0in;
RUN;

PROC GPLOT DATA=set1;
   PLOT X4*Q4 / vaxis=axis2 haxis=axis1;
TITLE1 ls=1.5in H=3.0 F=swiss "NORMAL PROBABILITY PLOT";
TITLE2 H=2.3 F=swiss "X4";
footnote ls=1.0in;
RUN;
NORMAL PROBABILITY PLOT

X1

Ordered Data vs. Standard Normal Quantiles
NORMAL PROBABILITY PLOT

X3

Ordered Data

Standard Normal Quantiles
/* Use PROC IML to compute summary statistics partial correlations, related t-tests, a chi-square Q-Q plot for assessing multivariate normality, and a goodness-of-fit test */

DATA set2; SET set1;
  KEEP X1-X4;
RUN;

PROC IML;
START NORMAL;
  USE set2;           /* Enter the data */
  READ ALL INTO X;

N=NROW(X); /* Number of observations is N */
P=NCOL(X); /* Number of traits is p */
SUM=X[+, ]; /* Total for each trait */
A=X'X-SUM'SUM/N; /* Corrected crossproducts matrix */
S=A/(N-1); /* Sample covariance matrix */
XBAR=SUM/N; /* Sample mean vector */
SCALE=INV(SQRT(DIAG(A)));
R=SCALE*A*SCALE; /* Sample correlation matrix */
PTR=J(P,P);
TR=J(P,P);
DF=N-2;
DO I1=1 TO P; /* T-tests for correlations */
    IP1=I1+1;
    DO I2=IP1 TO P;
        TR[I1,I2]=SQRT(N-2)#R[I1,I2]/SQRT(1-R[I1,I2]##2);
        TR[I2,I1]=TR[I1,I2];
        PTR[I1,I2]=(1-PROBT(ABS(TR[I1,I2]),DF))*2;
        PTR[I2,I1]=PTR[I1,I2];
    END;
END;
RINV=INV(R); /* Partial Correlations */
SCALER=INV(SQRT(DIAG(RINV)));
PCORR=-SCALER*RINV*SCALER;
DO I = 1 TO P;
   PCORR[I,I]=1.0;
END;
PTPCORR=J(P,P);
TPCORR=J(P,P);
DF=N-P;
DO I1=1 TO P; /* T-tests for partial correlations */
   IP1=I1+1;
   DO I2=IP1 TO P;
      TPCORR[I1,I2]=SQRT(N-P)#PCORR[I1,I2]/
      SQRT(1-PCORR[I1,I2]##2);
      TPCORR[I2,I1]=TPCORR[I1,I2];
      PTPCORR[I1,I2]=
      (1-PROBT(ABS(TPCORR[I1,I2]),DF))*2;
      PTPCORR[I2,I1]=PTPCORR[I1,I2];
   END;
END;
PRINT,,,," THE SAMPLE MEAN VECTOR";
PRINT XBAR;
PRINT,,,," THE SAMPLE CORRELATION MATRIX";
PRINT R;
PRINT,,,," VALUES OF T-TESTS FOR ZERO CORRELATIONS";
PRINT TR;
PRINT,,,," P-VALUES FOR THE T-TESTS FOR ZERO CORRELATIONS";
PRINT PTR;
PRINT,,,," THE MATRIX OF PARTIAL CORRELATIONS CONTROLLING FOR";
PRINT " ALL VARIABLES NOT IN THE PAIR";
PRINT PCORR;
PRINT,,,," VALUES OF T-TESTS FOR PARTIAL CORRELATIONS";
PRINT TPCORR;
PRINT,,,," P-VALUES FOR T-TESTS FOR PARTIAL CORRELATIONS";
PRINT PTPCORR;
/* Compute plotting positions
   for a chi-square probability plot */

E=X-(J(N,1)*XBAR);
D=VECDIAG(E*INV(S)*E'); /* Squared Mah. distances */
RD = RANK(D); /* Compute ranks */
   RD=(RD-.5)/N;
PD2=P/2;
   Q=2*GAMINV(RD,PD2); /* Plotting positions */
   DQ=D||Q;
CREATE CHISQ FROM DQ ; /* Open a file to store results */
APPEND FROM DQ;
/* Compute test statistic */

\[
 rpn = t(D) * Q - \frac{(\text{sum}(D) * \text{sum}(Q))}{N};
 rpn = \frac{rpn}{\sqrt{\left(\text{ssq}(D) - \frac{(\text{sum}(D)^2)}{N}\right) \times \left(\text{ssq}(Q) - \frac{(\text{sum}(Q)^2)}{N}\right)}};
\]

/* Simulate a p-value for the correlation test */

ns=10000;
pvalue=0;
do i=1 to ns;
    do j1=1 to n;
        do j2=1 to p;
            x[j1,j2] = rannor(-100);
        end;
    end;
end;
SUMX=X[+, ];
A=t(X)*X-t(SUMX)*SUMX/N;
S=A/(N-1);
XBAR=SUMX/N;
E=X-(J(N,1)*XBAR);
DN=VECDIAG(E*INV(S)*t(E));
RN = RANK(DN);
RN =(RN-.5)/N;
PD2=P/2;
QN=2*GAMINV(RN,PD2);
rpnn = t(DN)*QN-(sum(DN)*sum(QN))/N;
rpnn = rpnn/sqrt((ssq(DN)-(sum(DN)**2)/N)
    *(ssq(QN)-(sum(QN)**2)/N));
if(rpn>rpnn) then pvalue=pvalue+1;
end;
pvalue=pvalue/ns;

print,,,,,"Correlation Test of Normality";
print N P rpn pvalue;

FINISH;

RUN NORMAL;
### The Sample Mean Vector

\( \text{XBAR} \)

\[
1909.1 \quad 1749.5333 \quad 1509.1333 \quad 1724.9667
\]

### The Sample Correlation Matrix

\( R \)

\[
\begin{array}{cccc}
1 & 0.908396 & 0.8901079 & 0.8979365 \\
0.908396 & 1 & 0.7882129 & 0.7881034 \\
0.8901079 & 0.7882129 & 1 & 0.9231013 \\
0.8979365 & 0.7881034 & 0.9231013 & 1 \\
\end{array}
\]
VALUES OF T-TESTS FOR ZERO CORRELATIONS

<table>
<thead>
<tr>
<th></th>
<th>TR</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.496422</td>
<td>10.334636</td>
<td>10.795644</td>
<td></td>
</tr>
<tr>
<td>11.496422</td>
<td>1</td>
<td>6.7774009</td>
<td>6.7749149</td>
<td></td>
</tr>
<tr>
<td>10.334636</td>
<td>6.7774009</td>
<td>1</td>
<td>12.701875</td>
<td></td>
</tr>
<tr>
<td>10.795644</td>
<td>6.7749149</td>
<td>12.701875</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

P-VALUES FOR THE T-TESTS FOR ZERO CORRELATIONS

<table>
<thead>
<tr>
<th></th>
<th>PTR</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.068E-12</td>
<td>4.631E-11</td>
<td>1.73E-11</td>
<td></td>
</tr>
<tr>
<td>4.068E-12</td>
<td>1</td>
<td>2.322E-7</td>
<td>2.3371E-7</td>
<td></td>
</tr>
<tr>
<td>4.631E-11</td>
<td>2.322E-7</td>
<td>1</td>
<td>3.852E-13</td>
<td></td>
</tr>
<tr>
<td>1.73E-11</td>
<td>2.3371E-7</td>
<td>3.852E-13</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
THE MATRIX OF PARTIAL CORRELATIONS CONTROLLING FOR ALL VARIABLES NOT IN THE PAIR

PCORR

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0.7192895</th>
<th>0.2644769</th>
<th>0.3777683</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-0.018395</td>
<td>-0.107278</td>
<td></td>
</tr>
<tr>
<td>0.7192895</td>
<td>1</td>
<td>0.6117527</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2644769</td>
<td>-0.018395</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3777683</td>
<td>-0.107278</td>
<td>0.6117527</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

VALUES OF T-TESTS FOR PARTIAL CORRELATIONS

TPCORR

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>5.2794243</th>
<th>1.3983658</th>
<th>2.0804059</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.2794243</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2794243</td>
<td>1</td>
<td>-0.093813</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3983658</td>
<td>-0.093813</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.0804059</td>
<td>-0.550189</td>
<td>3.9432934</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

226
P-VALUES FOR T-TESTS FOR PARTIAL CORRELATIONS

PTPCORR

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0.0000161</th>
<th>0.1738181</th>
<th>0.0474835</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.9259766</td>
<td>0.5868862</td>
<td></td>
</tr>
<tr>
<td>0.0000161</td>
<td>1</td>
<td>0.9259766</td>
<td>0.5868862</td>
<td></td>
</tr>
<tr>
<td>0.1738181</td>
<td>0.9259766</td>
<td>1 0.0005419</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0474835</td>
<td>0.5868862</td>
<td>0.0005419</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Correlation Test of Normality

<table>
<thead>
<tr>
<th>N</th>
<th>P</th>
<th>RPN</th>
<th>PVALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>4</td>
<td>0.9817564</td>
<td>0.4678</td>
</tr>
</tbody>
</table>

227
/* Display the chi-square probability plot
   for assessing multivariate normality */

axis2 label=(f=swiss h=2.5 a=90 r=0
   "Ordered Distances ")
  value=(f=triplex h=2.0)
  length = 5.0in;

axis1 label=(f=swiss h=2.5
   "Chi-Square Quantiles"
  value=(f=triplex h=1.6)
  length= 5.5in;
SYMBOL1 V=CIRCLE H=2.0 ;

PROC GPLOT DATA=CHISQ;
   PLOT Col1*COL2 / vaxis=axis2 haxis=axis1;
TITLE1 ls=1.5in H=3.0 F=swiss "CHI-SQUARE PLOT";
Title2 H=2.3 F=swiss "Board Stiffness Data";
footnote ls=1.0in;
RUN;
CHI-SQUARE PLOT
Board Stiffness Data

Ordered Distances

Chi-Square Quantiles

230
A practical approach to assessing multivariate normality

• For any \( X \), we have seen that \( X \sim MVN \) if and only if \( a'X \) is univariate normally distributed.

• How about taking a large number of projections and evaluating each projection for univariate normality?

• Let us try with \( N \) (large) independent random projections on the unit vector. We will test each projection for univariate normality using the Shapiro-Wilks’ test.
A practical approach to assessing multivariate normality – contd.

- Note that we will have a large number of tests to evaluate, and this means that we will have to account for multiple hypothesis tests that have to be carried out. So we convert all the $p$-values into so-called $q$-values and if all the $q$-values are greater than the desired False Discovery Rate (FDR), then we accept the null hypothesis that $X$ is multivariate normally distributed.

- Note the large number of calculations needed: imperative to write R code efficiently
testnormality <- function(X, numproj = 100000) {
  # note that the value returned is the q-value of the test
  p <- ncol(X)
  n <- nrow(X)
  x <- matrix(rnorm(numproj * p), nrow = p)
    # generate 100,000, standard p-variate
    # normal random variables.
  y <- matrix(sqrt(apply(x^2, 2, sum)), nrow = p, ncol = numproj,
              by = T)
  z <- x / y
tempdat <- as.matrix(X) %*% z

    # this gives rise to a p x numproj matrix called tempdat

    # perform Shapiro-Wilks’ test and calculate individual p-values on
    # each of the numproj observation sets.

pvals <- as.numeric(matrix(unlist(apply(tempdat, 2,
shapiro.test)), ncol=4, by = T)[,2])

return(min(sort(pvals) * numproj / (1:numproj)))
}

    # usage demonstrated in steel.R
The E-statistic (energy) test of Multivariate Normality

Let $X_1, X_2, \ldots, X_n$ be a sample from some $p$-variate distribution. Then, consider the following test statistic:

$$E = n \left[ \frac{2}{n} \sum_{i=1}^{n} E||X_i^* - Z|| - E||Z - Z'|| - \frac{1}{n^2} \sum_{i,j=1}^{n} ||X_i^* - X_j^*|| \right],$$

where $X_i^*, i = 1, 2, \ldots, n$ is the standardized sample and $Z$, and $Z'$ are independent identically distributed standard $p$-variate normal random vectors, and $||| \cdot |||$ denotes Euclidean norm.

- The critical region is obtain by parametric bootstrap.
- implemented in R by the `mvnorm.etest()` function in the `energy` package.