UCD School of Mathematics and Statistics

STAT40840: Data programming with SAS

Laura Kirwan

Lecture 8

Lecture 8: Analysing Data

Objectives – 8.1

- Use SAS procedures to summarise data numerically
- Use SAS procedures to summarise data graphically

UNIVARIATE procedure

- descriptive statistics based on moments (including skewness and kurtosis), quantiles or percentiles (such as the median), frequency tables, and extreme values
- histograms that optionally can be fitted with probability density curves for various distributions and with kernel density estimates
- cumulative distribution function plots (cdf plots). Optionally, these can be superimposed with probability distribution curves for various distributions.

UNIVARIATE procedure

- quantile-quantile plots (Q-Q plots), probability plots, and probability-probability plots (P-P plots). These plots facilitate the comparison of a data distribution with various theoretical distributions.
- goodness-of-fit tests for a variety of distributions including the normal
- the ability to inset summary statistics on plots

UNIVARIATE procedure

- the ability to analyse data sets with a frequency variable
- the ability to create output data sets containing summary statistics, histogram intervals, and parameters of fitted curves

Scenario

We will use the bodyweight dataset (from assignment1).

We will use the UNIVARIATE procedure to summarise the dataset and check for outliers.

UNIVARIATE procedure

Numerical descriptive statistics.

```
proc univariate data=work.bodyweight;
     var Bodyweight0 Energy_Intake0
         Bodyweight6 Energy_Intake6 ;
run;
```

L8_D1.sas

Viewing the Output

```
proc univariate data=work.bodyweight;
     var Bodyweight0 Energy_Intake0
         Bodyweight6 Energy_Intake6 ;
run;
```

Moments			
N	176	Sum Weights	176
Mean	2218.53448	Sum Observations	390462.068
Std Deviation	667.066134	Variance	444977.227
Skewness	0.82012438	Kurtosis	0.72050898
Uncorrected SS	944124575	Corrected SS	77871014.8
Coeff Variation	30.0678732	Std Error Mean	50.2820018

L8_D1.sas

Viewing the Output

```
proc univariate data=work.bodyweight;
     var Bodyweight0 Energy_Intake0
         Bodyweight6 Energy_Intake6 ;
run;
```

Basic Statistical Measures			
Location		Variability	
Mean	2218.534	Std Deviation	667.06613
Median	2149.905	Variance	444977
Mode	.	Range	3521
		Interquartile Range	833.84928

L8_D1.sas

Viewing the Output

```
proc univariate data=work.bodyweight;
     var Bodyweight0 Energy_Intake0
         Bodyweight6 Energy_Intake6 ;
run;
```

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	44.12184	Pr > |t|	<.0001
Sign	M	88	Pr >= |M|	<.0001
Signed Rank	S	7788	Pr >= |S|	<.0001

L8_D1.sas

Viewing the Output

```
proc univariate data=work.bodyweight;
     var Bodyweight0 Energy_Intake0
         Bodyweight6 Energy_Intake6 ;
run;
```

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	4580.79
99%	4311.06
95%	3493.39
90%	3175.00
75% Q3	2561.61
50% Median	2149.90
25% Q1	1727.76
10%	1430.50
5%	1326.31
1%	1071.71
0% Min	1059.51

L8_D1.sas

Viewing the Output

```
proc univariate data=work.bodyweight;
     var Bodyweight0 Energy_Intake0
         Bodyweight6 Energy_Intake6 ;
run;
```

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1059.51	23	3709.19	137
1071.71	8	3930.66	129
1087.08	205	3999.04	29
1114.40	217	4311.06	122
1209.17	207	4580.79	132

L8_D1.sas

Viewing the Output

```
proc univariate data=work.bodyweight;
     var Bodyweight0 Energy_Intake0
         Bodyweight6 Energy_Intake6 ;
run;
```

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	44	20.00	100.00

L8_D1.sas

ODS

Use the output delivery system (ODS) to select the output that you wish to print.

```
ods trace on;
proc univariate data=work.bodyweight;
     var Bodyweight0 Energy_Intake0
Bodyweight6 Energy_Intake6 ;
run;
ods trace off;
```

L8_D2.sas

ods trace produces a list of output tables in the log

ODS

```
*Output Added:
-------------
Name:          Moments
Name:          BasicMeasures
Name:          TestsForLocation
Name:          Quantiles
Name:          ExtremeObs;
```

L8_D2.sas

ODS

We use the ods to select the output that we want

```
ods select Moments ExtremeObs;
proc univariate data=work.bodyweight;
     var Bodyweight0 Energy_Intake0
Bodyweight6 Energy_Intake6;
run;
```

L8_D2.sas

Exercise 1

Run the code in L8_E1.sas. Does it run correctly?

Why / why not?

```
ods select Moments ExtremeObs;
proc univariate data=work.bodyweight noprint;
     var Bodyweight0 Energy_Intake0
Bodyweight6 Energy_Intake6;
run;
```

Exercise 1 - solution

Run the code in L8_E1.sas. Does it run correctly?

Why / why not?

```
ods select Moments ExtremeObs;
proc univariate data=work.bodyweight noprint;
     var Bodyweight0 Energy_Intake0
Bodyweight6 Energy_Intake6;
run;
```

Output is not created with the `noprint` option.

```
WARNING: Output 'ExtremeObs' was not created.  Make sure that the output
object name, label, or path is spelled correctly.  Also, verify that the
appropriate procedure options are used to produce the requested output
object.  For example, verify that the NOPRINT option is not used.
```

UNIVARIATE procedure

Frequency data

```
ods select Frequencies;
proc univariate data=work.bodyweight1 freq;
    var gender;
run;
```

L8_D3.sas

UNIVARIATE procedure

Frequency data

```
ods select Frequencies;
proc univariate data=work.bodyweight1 freq;
    var gender;
run;
```

The UNIVARIATE Procedure
Variable: Gender

Frequency Counts			
Value	Count	Percents	
		Cell	Cum
0	101	45.9	45.9
1	119	54.1	100.0

L8_D3.sas

UNIVARIATE procedure

Grouping descriptive statistics

```
proc sort data=work.bodyweight1;
     by gender;
run;

proc univariate data=work.bodyweight1;
     by gender;
     var bodyweight0;
run;
```

L8_D4.sas

UNIVARIATE procedure

Grouping descriptive statistics

```
proc univariate data=work.bodyweight1;
     by gender;
     var bodyweight0;
run;
```

We get separate tables for males and females

```
---------------------------------------- Gender=1 ----------------------------------------

                          The UNIVARIATE Procedure
                           Variable:  Bodyweight0

                                 Moments

N                         119    Sum Weights               119
Mean                68.4033613    Sum Observations         8140
Std Deviation       12.5152919    Variance            156.632531
Skewness            1.14582501    Kurtosis            1.87374681
Uncorrected SS          575286    Corrected SS        18482.6387
Coeff Variation     18.2963112    Std Error Mean      1.14727493
```

L8_D4.sas

UNIVARIATE procedure

Saving summary statistics using OUT= output dataset

```
proc univariate data=work.bodyweight1;
     var bodyweight0;
     output out=work.means mean=m_weight0 ;
run;
```

L8_D5.sas

UNIVARIATE procedure

Saving summary statistics using OUT= output dataset

```
proc univariate data=work.bodyweight1;
     var bodyweight0;
     output out=work.means mean=m_weight0 ;
run;
```

A dataset called work.means is created with one variable and one observation

Obs	m_weight0
1	75.4218

L8_D5.sas

UNIVARIATE procedure

Saving summary statistics using ods output

```
ods output Moments=work.Moments;
proc univariate data=work.bodyweight1;
     var bodyweight0;
run;
```

L8_D5.sas

UNIVARIATE procedure

Saving summary statistics using ods output

```
ods output Moments=work.Moments;
proc univariate data=work.bodyweight1;
     var bodyweight0;
run;
```

A dataset called work.moments is created

Obs	VarName	Label1	cValue1	nValue1	Label2	cValue2	nValue2
1	Bodyweight0	N	220	220.000000	Sum Weights	220	220.000000
2	Bodyweight0	Mean	75.4218182	75.421818	Sum Observations	16592.8	16593
3	Bodyweight0	Std Deviation	14.7023987	14.702399	Variance	216.160526	216.160526
4	Bodyweight0	Skewness	0.62738205	0.627382	Kurtosis	0.24931139	0.249311
5	Bodyweight0	Uncorrected SS	1298798.3	1298798	Corrected SS	47339.1553	47339
6	Bodyweight0	Coeff Variation	19.4935617	19.493562	Std Error Mean	0.99123552	0.991236

L8_D5.sas

UNIVARIATE procedure

Computing Confidence Limits for the Mean, Standard Deviation, and Variance

```
proc univariate data=work.bodyweight1
cibasic;
      var bodyweight0;
run;
```

L8_D6.sas

UNIVARIATE procedure

Computing Confidence Limits for the Mean, Standard Deviation, and Variance

```
proc univariate data=work.bodyweight1
cibasic;
      var bodyweight0;
run;
```

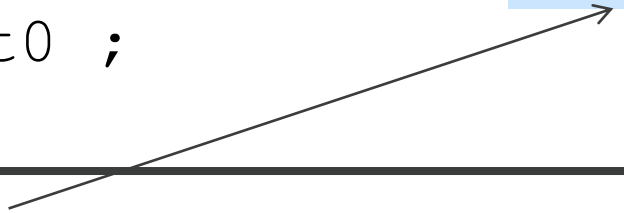
Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	75.42182	73.46824	77.37540
Std Deviation	14.70240	13.44497	16.22133
Variance	216.16053	180.76710	263.13147

L8_D6.sas

UNIVARIATE procedure

Graphical descriptive statistics: creating a histogram

```
proc univariate data=work.bodyweight1 noprint;
     histogram bodyweight0 ;
run;
```



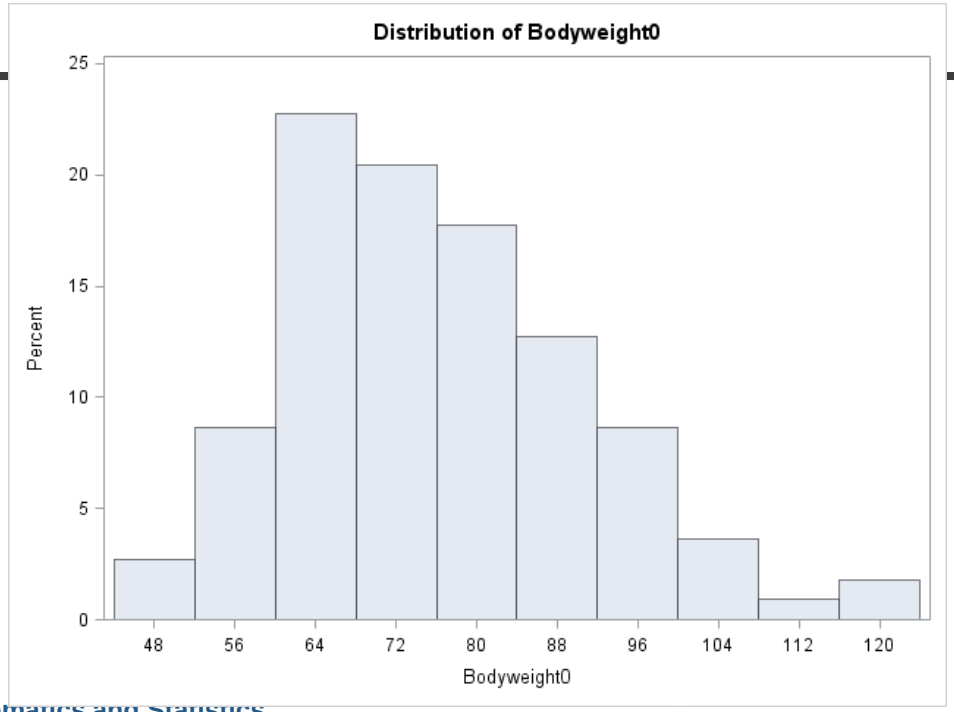
Suppresses automatic printing of tables

L8_D7.sas

UNIVARIATE procedure

Graphical descriptive statistics: creating a histogram

```
proc univariate data=work.bodyweight1 noprint;
     histogram bodyweight0 ;
run;
```



L8_D7.sas

UNIVARIATE procedure

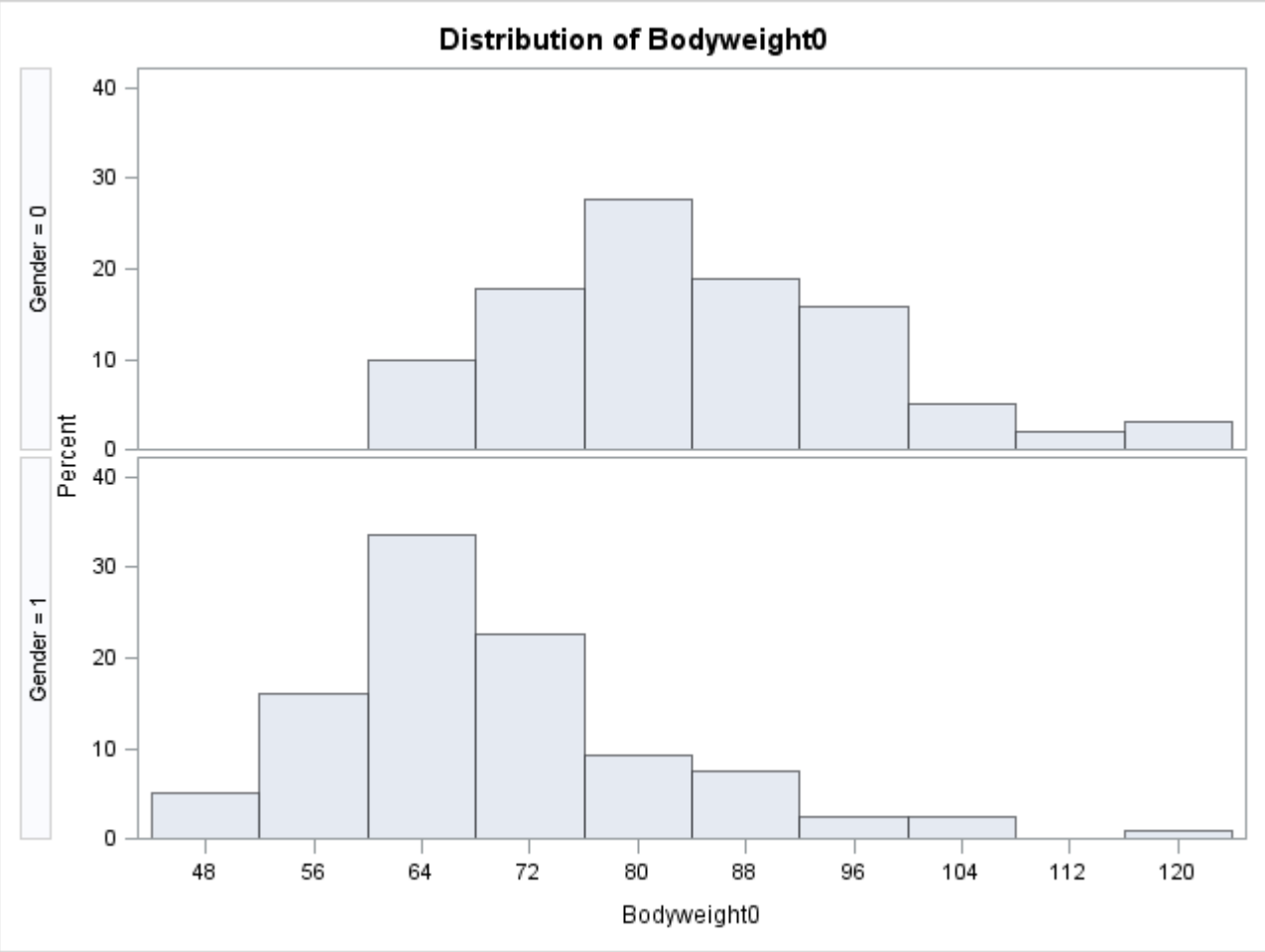
Comparing groups: creating a comparative histogram

```
proc univariate data=work.bodyweight1 noprint;
     class gender;
     histogram bodyweight0 ;
run;
```

L8_D7.sas

UNIVARIATE procedure

Comparing groups: creating a comparative histogram



L8_D7.sas

UNIVARIATE procedure

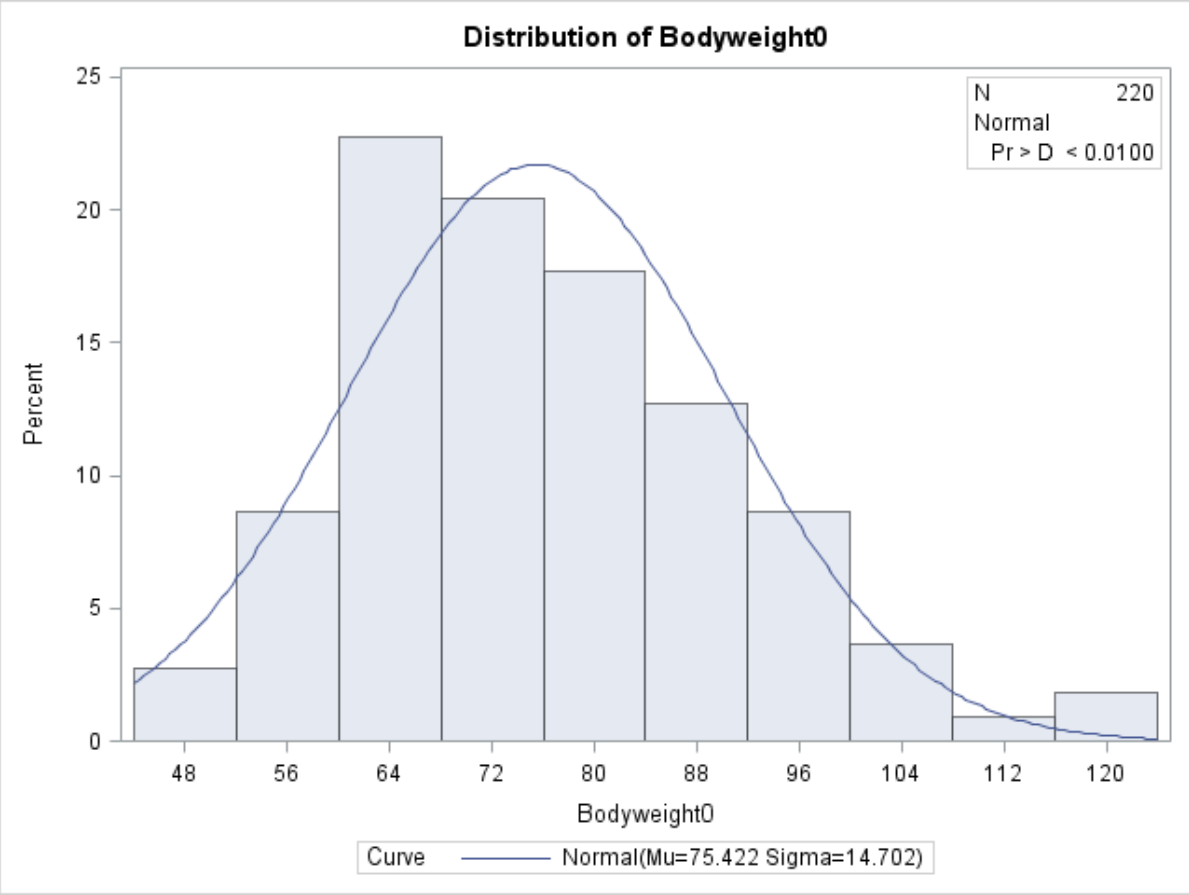
Adding a normal curve to a histogram

```
proc univariate data=work.bodyweight1 noprint;
    histogram bodyweight0/ normal;
    inset n normal(ksdpval) / pos = ne ;
run;
```

L8_D7.sas

UNIVARIATE procedure

Adding a normal curve to a histogram



L8_D7.sas

UNIVARIATE procedure

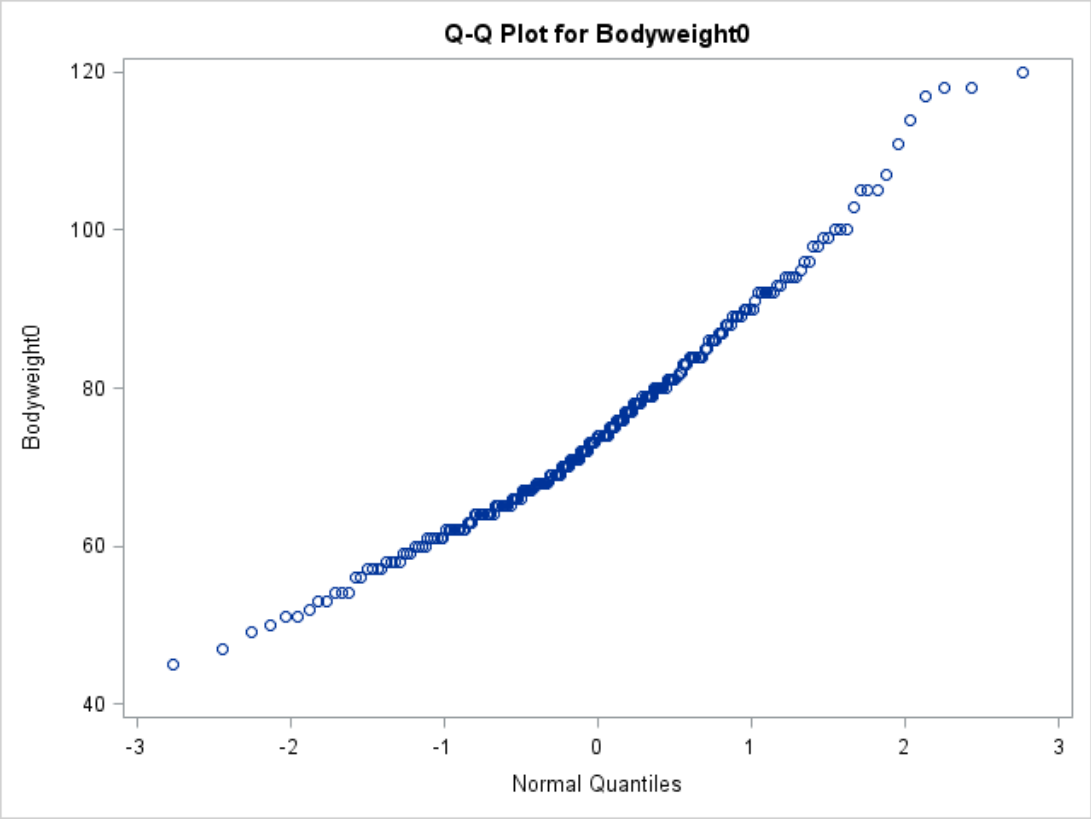
Producing a Q-Q plot

```
proc univariate data=work.bodyweight1 noprint
normal;
   qqplot bodyweight0 / normal;
run;
```

L8_D8.sas

UNIVARIATE procedure

Producing a Q-Q plot

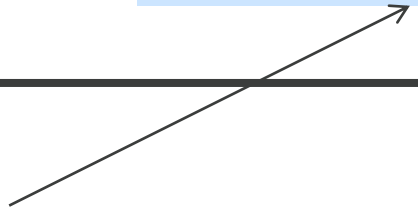


L8_D8.sas

UNIVARIATE procedure

Adding a distribution reference line to a Q-Q plot

```
proc univariate data=work.bodyweight1 normal noprint;
    qqplot bodyweight0 / normal (mu=est sigma=est);
run;
```

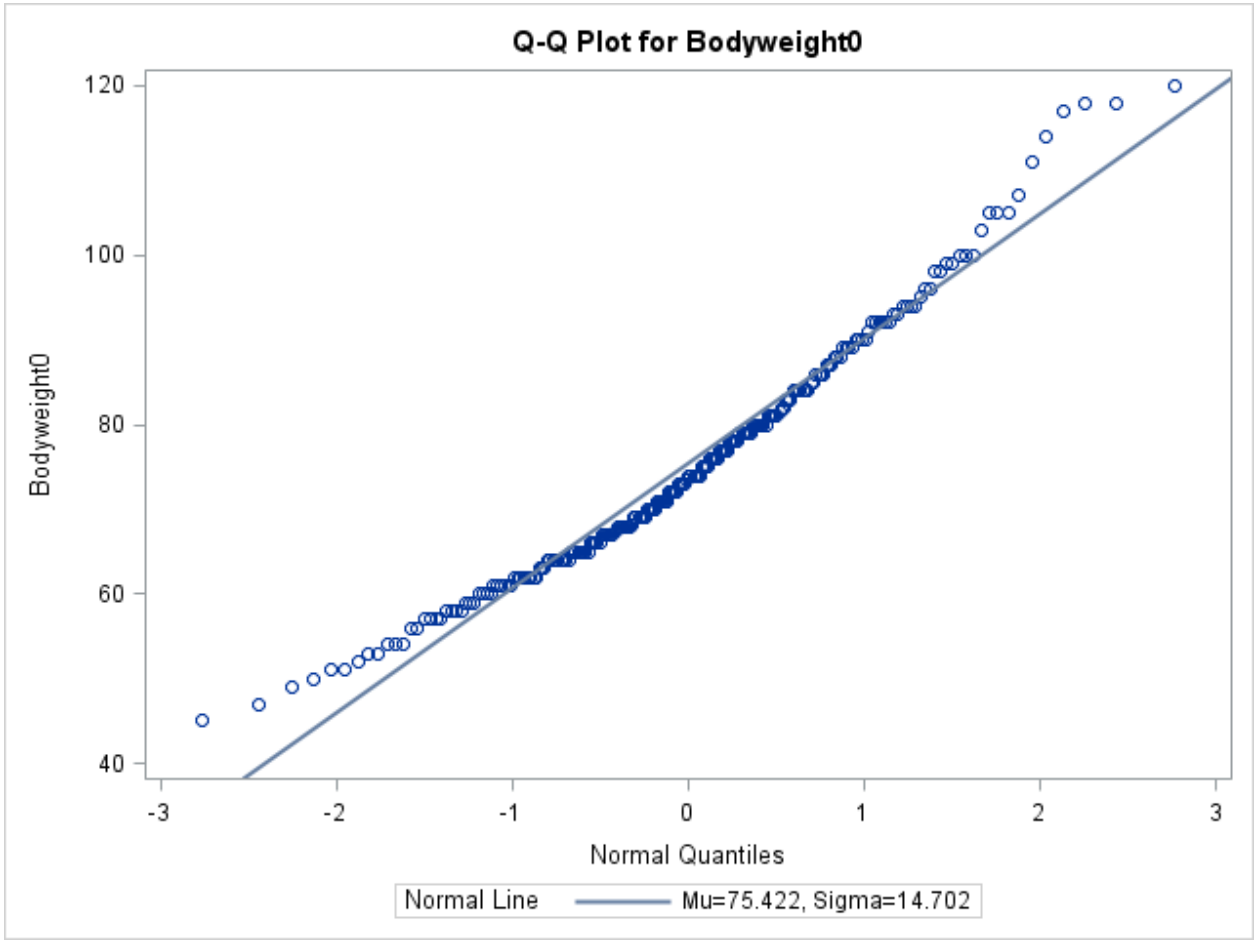


`est` option means SAS will estimate the mean and variance from the data. You can specify particular values to test

L8_D8.sas

UNIVARIATE procedure

Adding a distribution reference line to a Q-Q plot



L8_D8.sas

Lecture 8: Analysing Data

Objectives – 8.2

- Use a SAS procedure to assess correlations
- Use a SAS procedure to conduct a regression analysis

CORR procedure

Calculating a correlation coefficient

```
proc corr data=work.bodyweight1;
     var  Bodyweight0 Age;
run;
```

L8_D9.sas

CORR procedure

Calculating a correlation coefficient

```
proc corr data=work.bodyweight1;
     var  Bodyweight0 Age;
run;
```

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Bodyweight0	220	75.42182	14.70240	16593	45.00000	120.00000
Age	220	43.34091	13.79025	9535	18.00000	72.00000

Pearson Correlation Coefficients, N = 220
Prob > |r| under H0: Rho=0

	Bodyweight0	Age
Bodyweight0	1.00000	0.27696 <.0001
Age	0.27696 <.0001	1.00000

L8_D9.sas

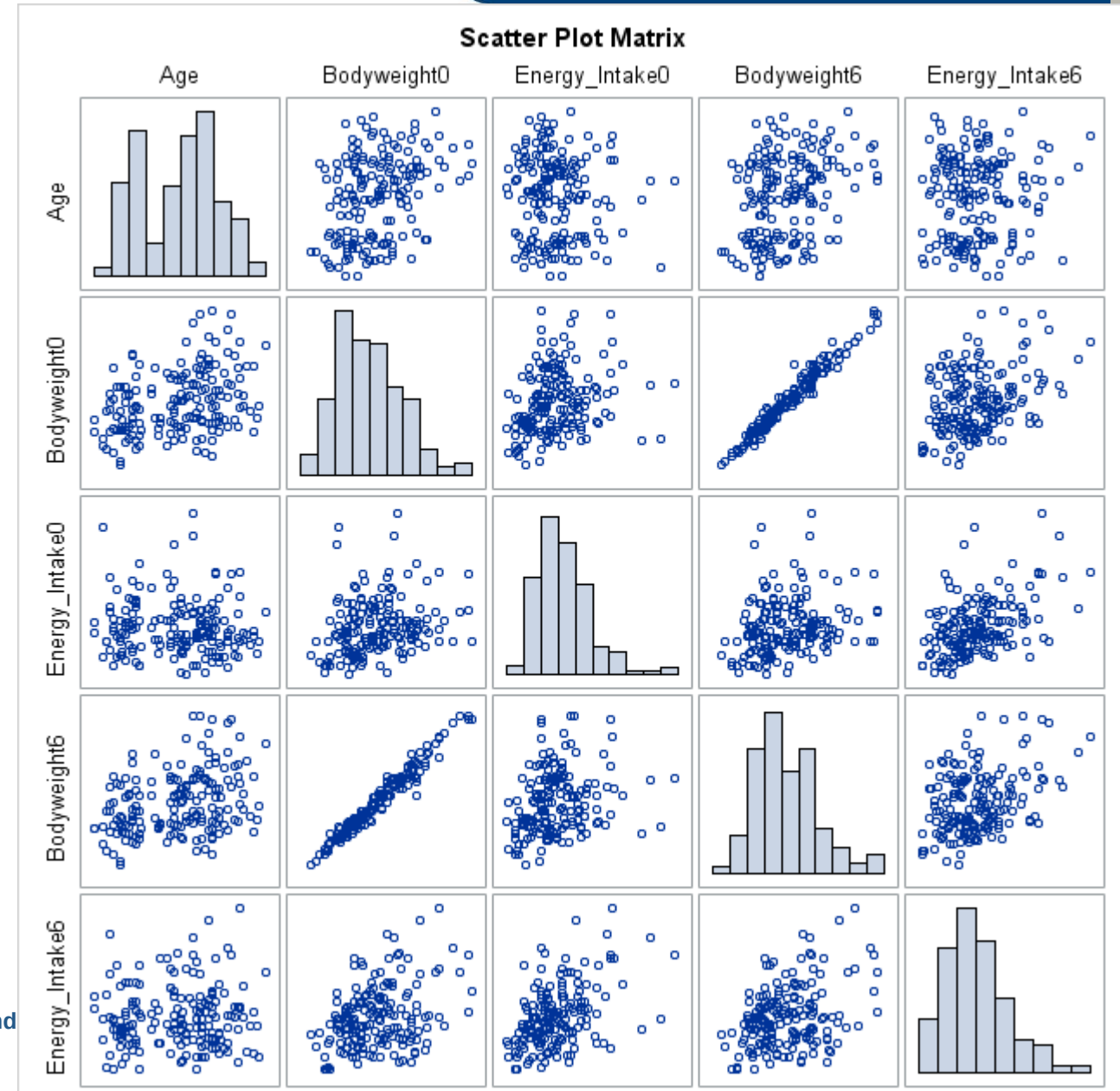
CORR procedure

Producing a scatterplot matrix

```
ods graphics on;
proc corr data=work.bodyweight1 nomiss
plots=matrix(histogram);
   var Age  Bodyweight0 Energy_Intake0
Bodyweight6 Energy_Intake6;
run;
ods graphics off;
```

L8_D10.sas

CORR procedure



L8_D10.sas

CORR procedure

Computing partial correlations

```
proc corr data=work.bodyweight1;
     var  Bodyweight0 Energy_Intake0;
     Partial age;
run;
```

L8_D11.sas

CORR procedure

Computing partial correlations

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Age	219	43.37443	13.81285	9499	18.00000	72.00000
Bodyweight0	219	75.35525	14.70281	16503	45.00000	120.00000
Energy_Intake0	219	2631	758.64564	576145	1338	5900

Pearson Partial Correlation Coefficients, N = 219
Prob > |r| under H0: Partial Rho=0

	Bodyweight0	Energy_Intake0
Bodyweight0	1.00000	0.29422 <.0001
Energy_Intake0	0.29422 <.0001	1.00000

L8_D11.sas

REG procedure

Fitting a linear regression

```
ods graphics on;
 proc reg data=work.bodyweight1;
    model bodyweight0 = age energy_intake0;
 run;
```

L8_D12.sas

REG procedure

Fitting a linear regression - estimates

L8_D12.sas

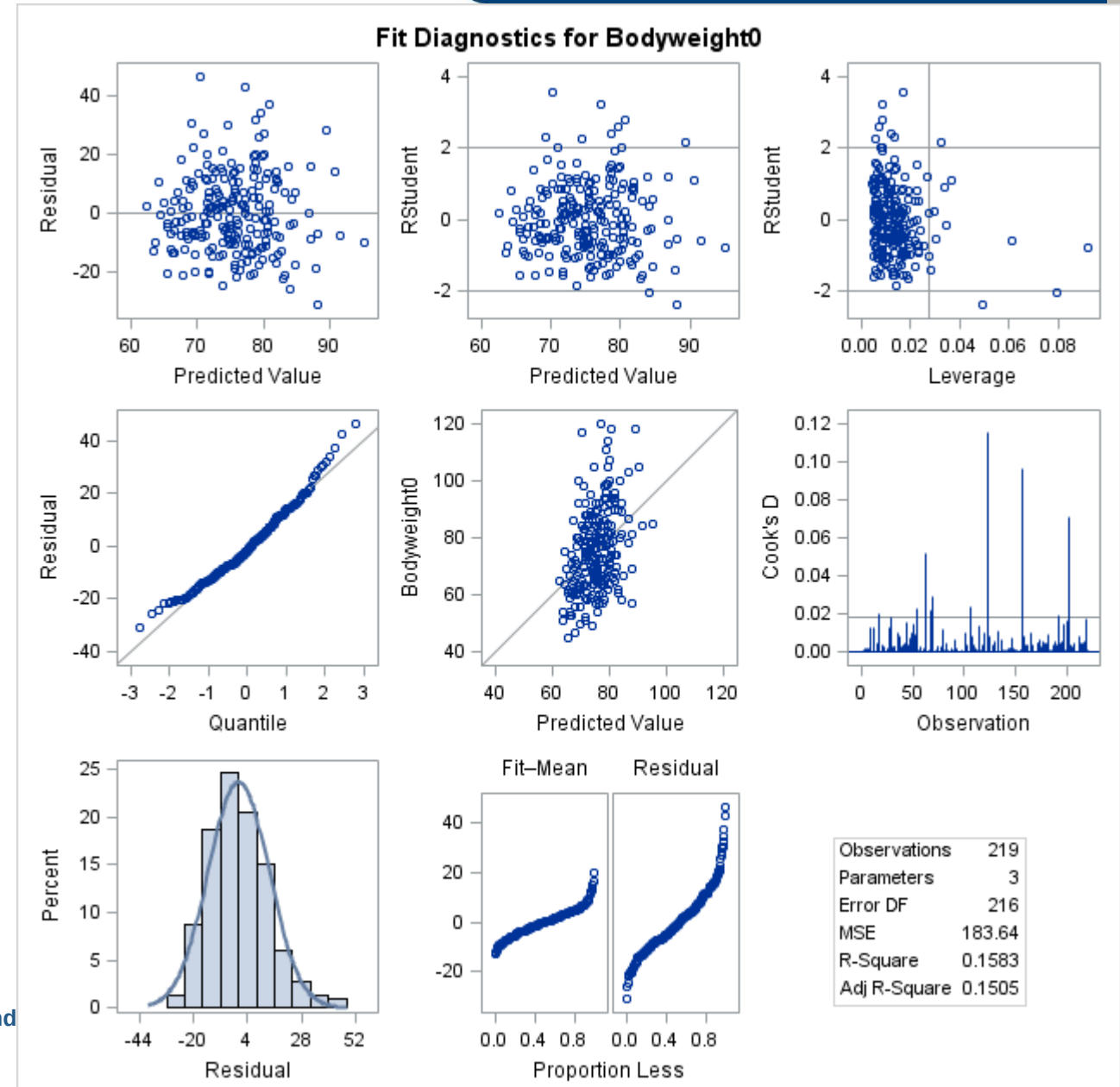
```
ods graphics on;
 proc reg data=work.bodyweight1;
    model bodyweight0 = age energy_intake0;
 run;
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	47.15293	4.53098	10.41	<.0001
Age	1	0.31753	0.06658	4.77	<.0001
Energy_Intake0	1	0.00548	0.00121	4.52	<.0001

REG procedure

Fitting a linear regression - diagnostics

L8_D12.sas



REG procedure

Saving residuals and predicted values in output dataset

```
ods graphics on;
 proc reg data=work.bodyweight1;
    model bodyweight0 = age energy_intake0;
     output out=pred r=r p=p;
 run;
```

L8_D12.sas