AARHUS
UNIVERSITET
BUSINESS AND SOCIAL SCIENCES

SAS/STAT

Statistical analysis

Analytics Group

See more at Helpdesk.asb.dk

Analytics Group

Agenda

- Brief refresher on Library references and proc contents
- Simple statistics and confidence intervals
- The help function
- Hypothesis testing
- ANOVA
- Regression analysis
  - Preliminary analysis
  - Test of assumptions
- Enhancing your output
  - HTML and .pdf output and ODS Graphics

Practical information

- Brief introduction to each topic.
- After each introduction – assignments for 10-15 minutes.
- Duration 3 hours – approximately

Why SAS/Stat?

- Why use SAS programming for hypothesis testing, ANOVA, regression analysis and so on?
- You have more options and modifications available to use in your analysis
- Once you have everything set up and working with preliminary analysis, the analysis itself and assumptions testing you can easily apply the same analysis over and over again to other variables and different datasets.
- This is just as fast as the usual pointing and clicking people know from graphical software.

Data library

- SAS library references: Highway to data
- Default library references
  - SAS User
  - Work (temporary)
- User defined library references
  - The physical location on your computer/server where you want SAS to retrieve data sets from and store them in.

library reference

- In general

```
libname name 'path';
run;
```

- Examples:

```
libname asb ' C:\documents\school\class';
run;
```

```
libname mt 'M:\master thesis\datasets';
run;
```

Proc contents

- Proc contents can be used to get general information about the data set at hand

```
proc contents data=libname.dataset;
run;
```

Proc Means

- Produces simple statistics such as:
  - Mean
  - Std. dev
  - variance
  - Confidence intervals
  - Kurtosis
  - Min/max
  - Etc.

Proc Means

```
Proc means data=dataset <options>;
Title 'Free text' ;
Var varx vary;
Run;
```

- You can personalize your output by inserting a title of your choice in the procedure. This can be done in pretty much all SAS procedures.
- Options:
  - Mean
  - Std
  - Clm
  - And many many more…

Proc Means output

Mean	Kurtosis	Skewness	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean	N
10.5861290	0.1360422	0.5146521	1.3874141	10.0772214	11.0950366	31

The Help function

- To see information about the *Means* procedure, type:
  - *Help Proc Means statement* in the Command Bar or
  - Press F1 and type *Proc Means*

- Here you can get:
  - A generel overview of the different procedures
  - An overview of the different *Options* avaiable in the procedure
  - Syntax examples
  - See details on how the different measures are calculated.
  - Concrete examples on how the procedurecould be used (only for some of the procedures)

Proc Ttest

- Proc Ttest can be used for hypothesis testing on:
  - One sample
  - Two independent samples
  - Paired samples

Proc Ttest – One sample

- In general

```
proc ttest data=dataset h0=x <options>;
var variable;
run;
```

- Available options:
  - Alpha=x the significance level for the test where default is 5%
  - Sides = 2 for the two sided test(default), sides=u for the upper one-sided test in which the alternative hypothesis indicates a mean greater than the null value, and sides=l for the lower two sided test in which the alternative hypothesis indicates a mean less than the null value.
  - Plots(showh0) shows a graphic illustration of the test. Only possible using ODS Graphics which will be shown later

Output for H_0: Runtime ≥ 10

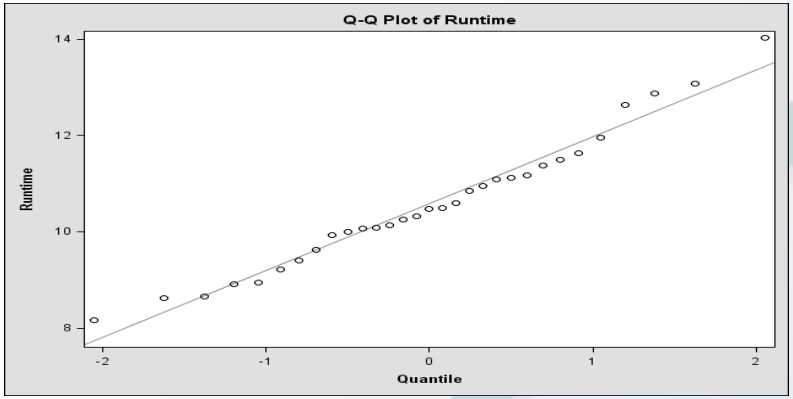
The TTEST Procedure

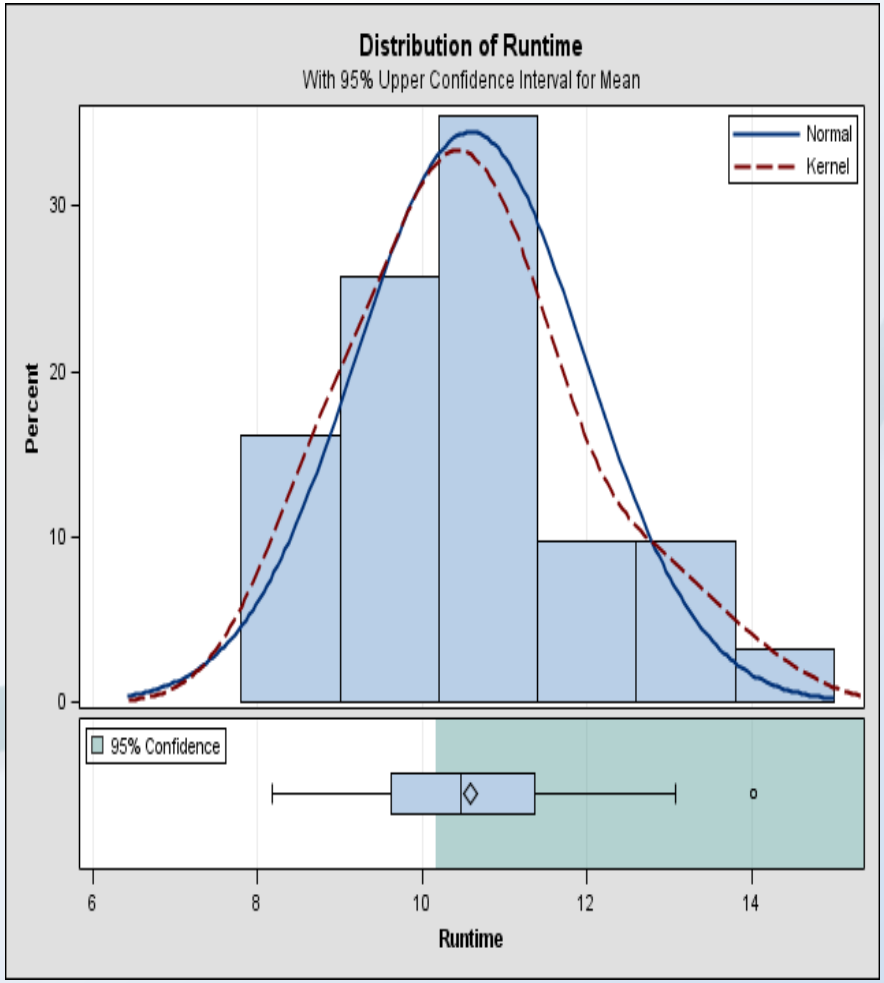
Variable: Runtime

N	Mean	Std Dev	Std Err	Minimum	Maximum
31	10.5861	1.3874	0.2492	8.1700	14.0300

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
10.5861	10.1632	Infty	1.3874	1.1087	1.8545

DF	t Value	Pr > t
30	2.35	0.0127





Analytics Group

Proc Ttest – Independent samples

- In general

```
Proc ttest <options>;
Class variable;
Var analysis_variable;
Run;
```

- Where
  - Class is the grouping variable
  - Var specifies the response variable to be used in calculation
- Options:
  - Alpha for significance level
  - Cochran produced p-values for the unequal variances situation

Output

The TTEST Procedure

Variable: V5 (Højde i cm)

V3	N	Mean	Std Dev	Std Err	Minimum	Maximum
Mand	363	182.8	6.4742	0.3398	163.0	198.0
Kvinde	205	169.9	6.5775	0.4594	154.0	192.0
Diff (1-2)		12.9724	6.5116	0.5689		

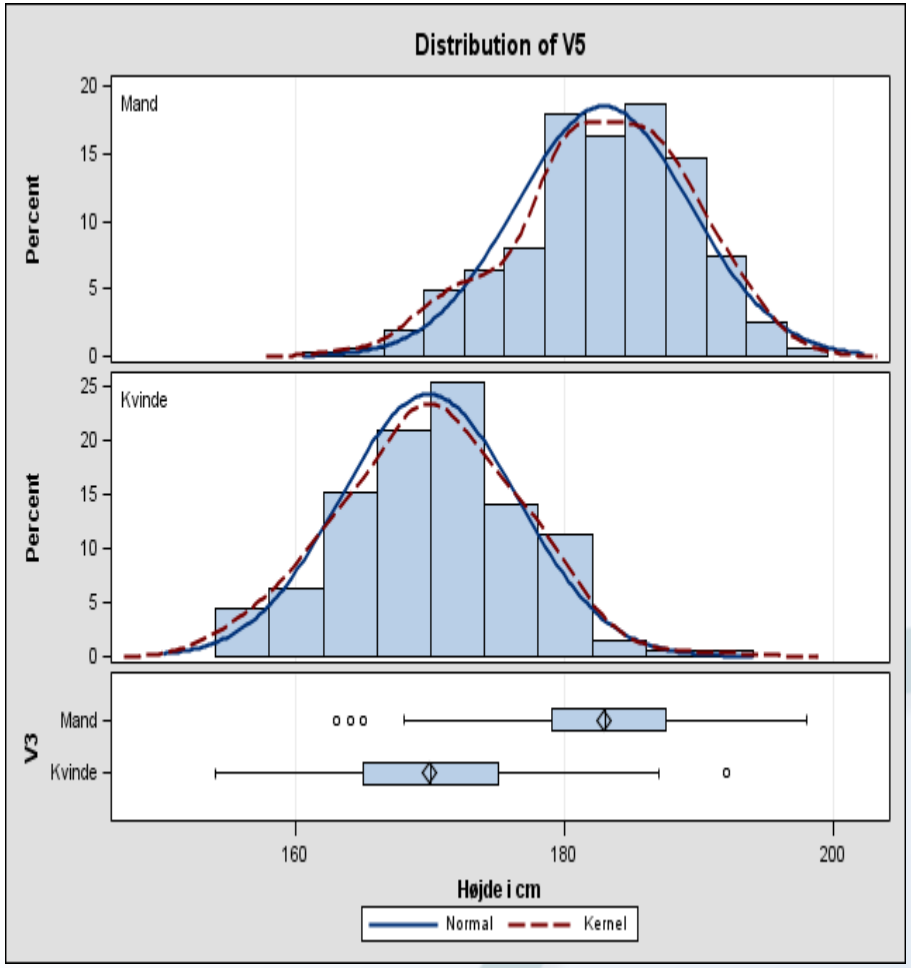
V3	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Mand		182.8	182.2	183.5	6.4742	6.0350	6.9828
Kvinde		169.9	169.0	170.8	6.5775	5.9965	7.2842
Diff (1-2)	Pooled	12.9724	11.8550	14.0898	6.5116	6.1534	6.9144
Diff (1-2)	Satterthwaite	12.9724	11.8492	14.0956			

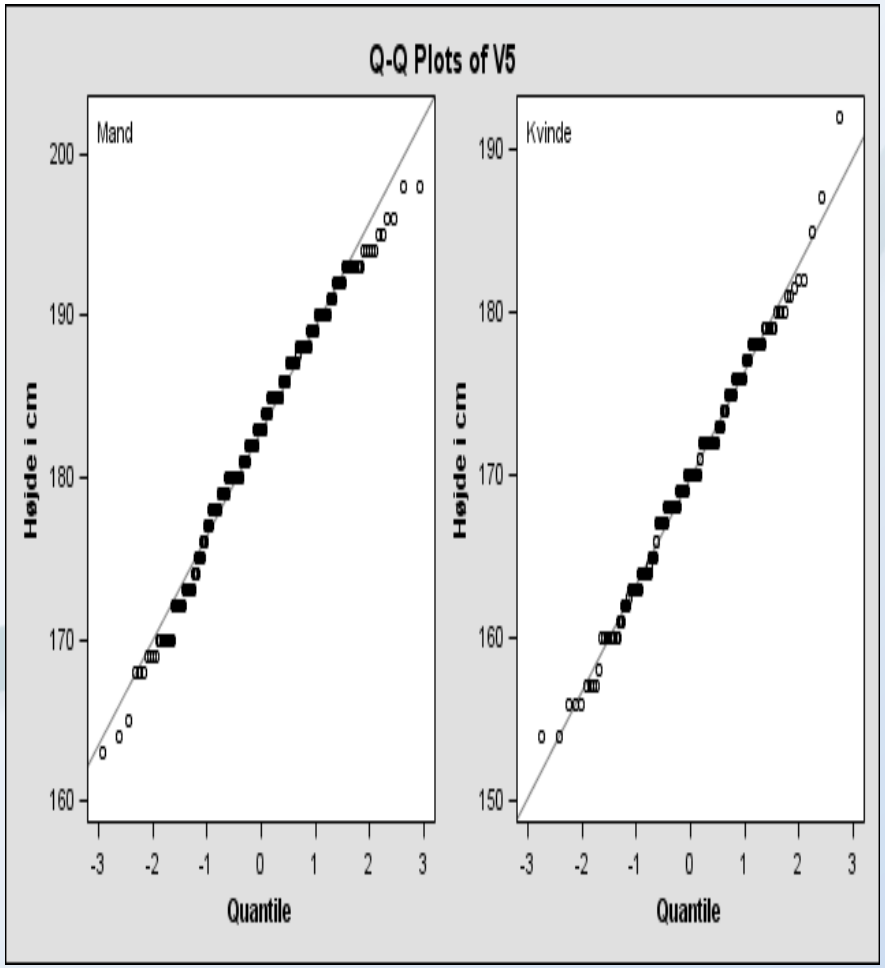
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	566	22.80	<.0001
Satterthwaite	Unequal	417.81	22.70	<.0001

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	204	362	1.03	0.7891

Output





Proc Ttest – Paired observations

- In general

```
Proc ttest <options>;
Paired var1*var2
 Run;
```

Output

The TTEST Procedure

Difference: V5 - V6

N	Mean	Std Dev	Std Err	Minimum	Maximum
557	10.6302	9.4895	0.4021	-25.0000	33.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
10.6302	9.8404	11.4199	9.4895	8.9630	10.0821

DF	t Value	Pr > |t|
556	26.44	<.0001

The dataset *b_cars*

- Contains the following information on 92 cars:
  - Price
  - Miles per gallon
  - Engine size
  - Horsepower
  - RPM
  - Fuel tank capacity
  - Weight.
  - And some other variables we will not use

Assignment 1

A. Produce the following statistics for the variables Midprice and horsepower from the dataset b_cars: Average, standard deviation, number of observations, kurtosis and skewness.

B. Make a 90% confidence interval for the variable midprice.

C. Test the hypothesis H_0: midprice = 21using a 1% significance level.

Hint: you might need to use the help function to see the options for the different procedures.

ANOVA

- Proc ANOVA
  - Only works on balanced samples! Rarely the case in practice, but the procedure is quite good at it.
- Proc GLM
  - Handles any ANOVA design and some regressions and ANCOVA designs.
  - Proc Reg is better for regressions and gives more options!

Proc GLM

- In general

```
Proc glm data=dataset;
Class independent_var;
Model dependent = independent;
Run;
```

- Where
  - Dependent is the continuous dependent variable
  - Independent is the categorical independent variable

Proc GLM

- The previous code only produces the ANOVA table.
- This will only tell you whether to reject the null or not.
- In case we reject the null hypothesis we would like to know where the difference is.
- We also need to assess the assumptions underlying ANOVA

Proc GLM – further options

- Under the model statement you can insert a number of additional lines and options
- Levenes test:
  - Means ID_var /hovtest=levene(type=abs)
  - instead of levenes you could write Bartlett for the Bartlett test. If you leave out the type GLM will produce squared differences as default.
- Multiple comparison adjustements
  - lsmeans ID_var /adjust=bon;
  - Requests the Bonferroni CI. Default is T. You can also write Tukey and many others
  - The Alpha level can also be adjusted in the usual manner in this line.

Partial output

The GLM Procedure

Dependent Variable: weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	9038.86667	4519.43333	17.95	<.0001
Error	27	6798.50000	251.79630		
Corrected Total	29	15837.36667			

R-Square	Coeff Var	Root MSE	weight Mean
0.570730	10.54825	15.86809	150.4333

Partial output

Levene's Test for Homogeneity of weight Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
trt	2	123.1	61.5253	0.97	0.3933
Error	27	1719.2	63.6742		

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

trt	weight LSMEAN	LSMEAN Number
Control	160.600000	1
Thiouracil	126.000000	2
Thyroxin	164.700000	3

Least Squares Means for effect trt
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: weight

i/j	1	2	3
1		0.0001	1.0000
2	0.0001		<.0001
3	1.0000	<.0001	

ANOVA assumptions

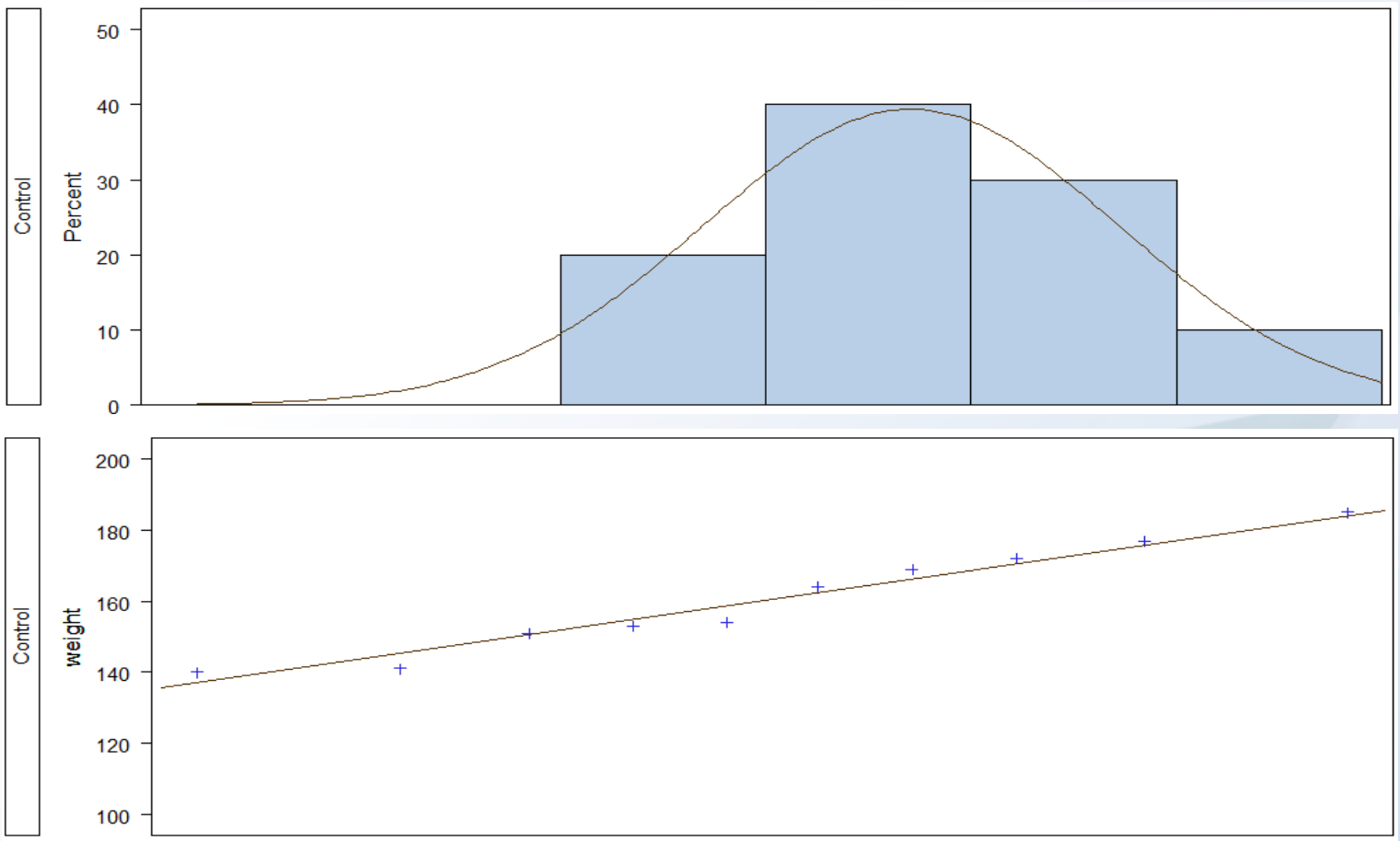
- We already tested for equal variances using Levenes test
- We need to see if our variable is normally distributed using histograms and a PP-plot.

- In general

```
Proc univariate data=dataset name <options>;
Class indep_var
Histogram var /normal;
Probplot var /normal <options>;
Run; Quit;
```

- A 45-degree reference line can be added to the pp-plot using the option: (mu=est sigma=est)
- The Univariate also produces a great deal of statistics so the normallity assumption can be tested numerically.

Partial output



Assignment 2

The dataset *ratgrowth* contains data on the effects of three different drug treatments on rat growth.

A. Test whether the mean weight of rats is equal whatever the treatment they receive, against the alternative that at least two means differ. i.e.:

H_0: $\mu_{Control} = \mu_{Thyroxin} = \mu_{Thiouracil}$

H_A: At least two μ differ

B. Test the assumptions underlying ANOVA

SAS regression methods

- SAS /STAT & /ETS (Econometric Time Series)
  - Reg — OLS
  - Logistic — Logistic regression
  - Probit — Probit regression
  - Autoreg — ARCH and GARCH models
  - Genmod — Poisson regression

  - And about 25 other methods.

Proc Reg shotcommings

- No dummy variables
  - As dependent variable
- Interaction effects and quadratic terms are only possible through a data step.

Preliminary analysis - correlations

- The correlations can tell us if it is possible to exclude some variables from the model estimation. If the correlation between two or more variables are high they essentially describe the same thing.
- High correlations are unwanted in the regression since it will infect the model with multicollinearity.

Correlations

- In general

```
Proc corr data=dataset <options>;
Var variable-list;
Run;
```

- Proc Corr computes (as default) the Pearson correlations between a number of variables.

- Options:
  - *Nosimple* supresses simple statistics in output
  - *Rank* ranks the correlations from highest to lowest in absolute value
  - Noprob: excludes the hypothesis test on the correlation being equal to zero.

Output

The CORR Procedure

7 Variables: Performance Rest_Pulse Run_Pulse Age Weight Runtime Maximum_Pulse

Pearson Correlation Coefficients, N = 31

	Performance	Rest_ Pulse	Run_Pulse	Age	Weight	Runtime	Maximum_ Pulse
Performance	1.00000	-0.47957	-0.31369	-0.22943	-0.10544	-0.98841	-0.22035
Rest_Pulse	-0.47957	1.00000	0.35246	-0.15087	0.04397	0.45038	0.30512
Run_Pulse	-0.31369	0.35246	1.00000	-0.31607	0.18152	0.31365	0.92975
Age	-0.22943	-0.15087	-0.31607	1.00000	-0.24050	0.19523	-0.41490
Weight	-0.10544	0.04397	0.18152	-0.24050	1.00000	0.14351	0.24938
Runtime	-0.98841	0.45038	0.31365	0.19523	0.14351	1.00000	0.22610
Maximum_Pulse	-0.22035	0.30512	0.92975	-0.41490	0.24938	0.22610	1.00000

Scatter- and box plots

- Proc (G)(Box)Plot outputs a simple plot of two variables. It is often easier to visually check if any of the independent variables have any explanatory power in relation to the dependent variable.

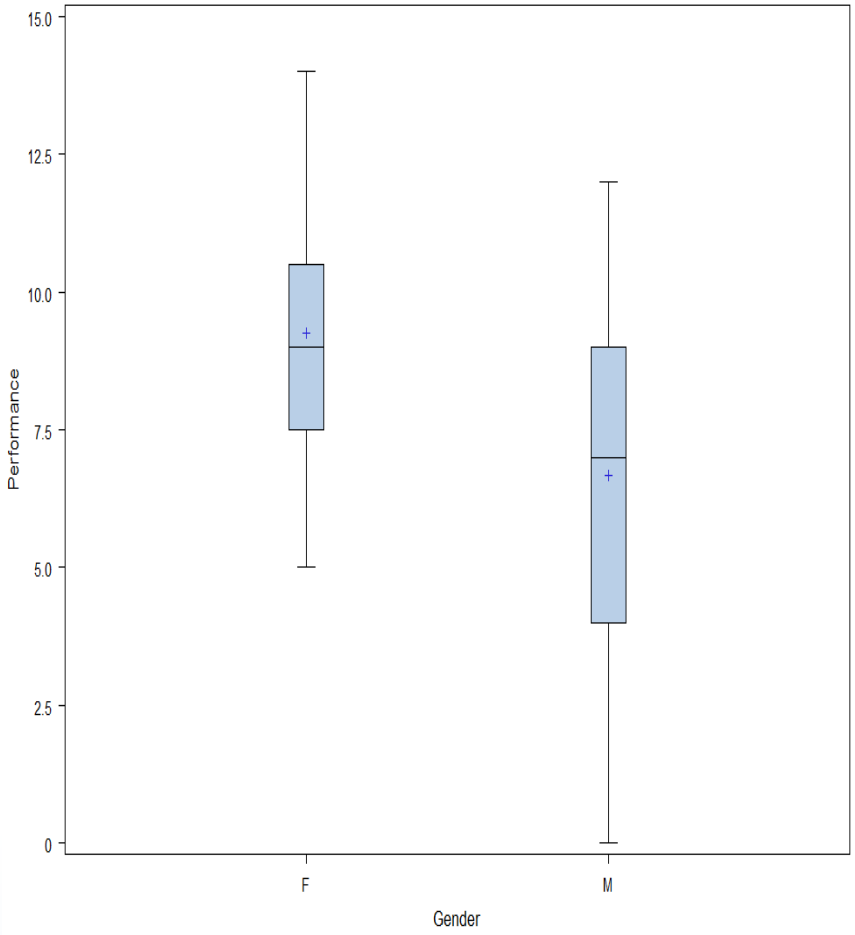
- In general

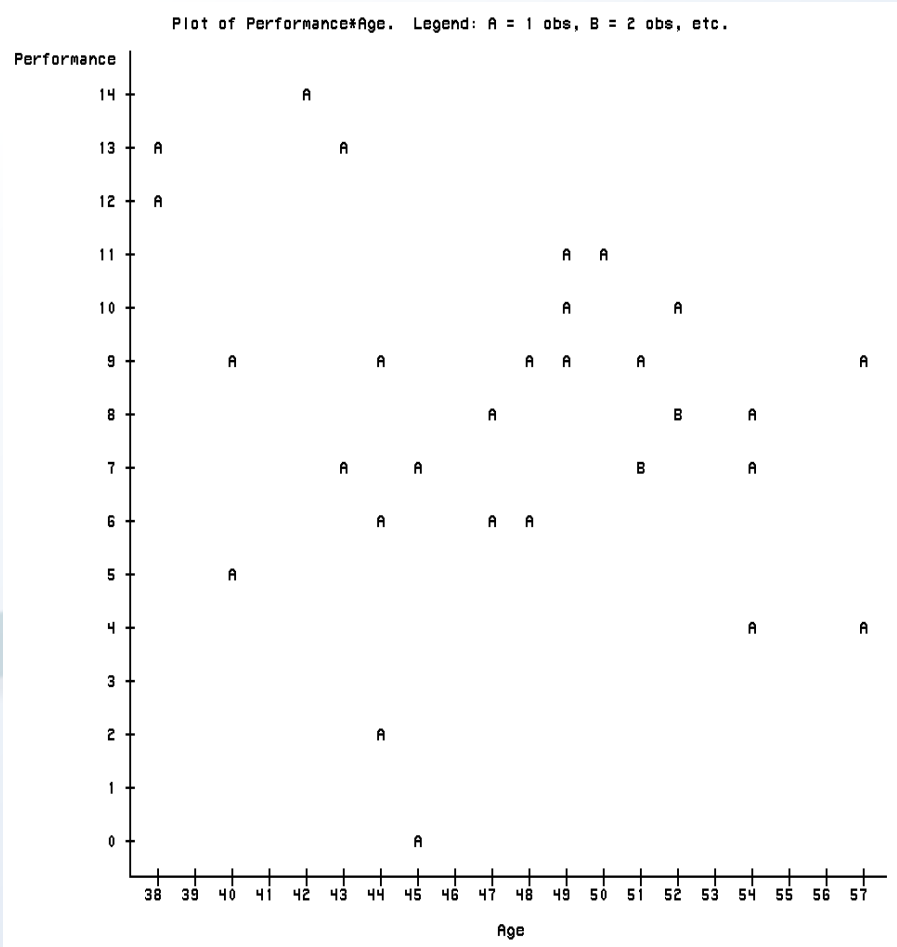
```
Proc (g)(box)plot data=dataset;
Plot Y*X <options>;
Plot Y*Z <options>;
Plot Y* (X Z V) <options>;
Run; Quit;
```

- You can insert linear reference lines on the plot using: /vref= -x 0 x in the option
- Note: When producing boxplots you might need to sort the dataset by the classification variable first.

AARHUS UNIVERSITET
BUSINESS AND SOCIAL SCIENCES

Output





Analytics Group

Regression analysis

- We are interested in determining whether the variables: Runtime, Age, Weight, Run_pulse, Rest_pulse, maximum_pulse and Performance, can help predict the Oxygen consumption of the 92 high school students.

- We estimate the initial model using OLS:
  - Oxygen_consumption = $\beta_0 + \beta_1$*Performance + β_2*Runtime + β_3*Age + β_4*Weight + β_5*Run_pulse + β_6*Rest_pulse + β_7*Maximum_pulse + ε

Regression analysis

- In general

```
Proc reg data=dataset name <options>;
Model DV = IV1 IV2 IV3 </options>;
<Options>;
Run; Quit;
```

- It is also possible to make plots in the regression procedure.
- Model options:
  - Noint: fits a model without the intercept term
  - Aic: computes Akaike's information criterion and stores it in the outest = dataset
  - VIF: computes the Variance Inflation Factor
  - HCC: Outputs the heteroscedasticity-consistent standard errors of the parameter estimates
  - White, DW and many many more…

Output

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen_Consumption

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.03251	103.14750	18.32	<.0001
Error	23	129.52204	5.63139		
Corrected Total	30	851.55455			

Root MSE	2.37306	R-Square	0.8479
Dependent Mean	47.37581	Adj R-Sq	0.8016
Coeff Var	5.00900		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	93.33753	36.49782	2.56	0.0176
Runtime	1	-2.08804	2.22856	-0.94	0.3585
Age	1	-0.21066	0.10519	-2.00	0.0571
Performance	1	0.25756	1.02373	0.25	0.8036
Maximum_Pulse	1	0.30490	0.13990	2.18	0.0398
Rest_Pulse	1	-0.01389	0.07114	-0.20	0.8469
Run_Pulse	1	-0.36618	0.12299	-2.98	0.0067
Weight	1	-0.07741	0.05681	-1.36	0.1862

Assignment 3

A. Check the correlations for the b_cars dataset to check for any potential multicolliniarity between the variables: CityMPG, EngineSize and HorsePower. Supress the simple statistics.

B. Make a plot of the variables RPM, FuelTankSize and Weight to visually assess the correlations.

C. Use the b_cars dataset to regress the Price of cars on the variables: CityMPG, EngineSize, HorsePower, RPM, FuelTankSize and Weight.

D. Write down the final model.

Testing the assumptions

- Constant variance – Homoskedasticity
  - Compute the heteroscedasticity-consistent standard errors
    - Compare OLSstd. errors with HRSE errors. If the difference is large it could be a sign of heteroscedasticy.->Only possible in SAS 9.2 = option /hcc
  - Plot the standardized residuals against the unstandardized predicted values of Y.
    - Ideally there should not be any clustering or decreasing/increasing effect.

- Errors are uncorrelated with explanatory variables
  - Plot the standardized residuals against unstandardized predicted value
    - Again no clustering or the like.
  - Plot the standardized residuals against all the explanatory variables

- Normally distributed residuals
  - PP-plot of the residuals
  - Histogram of the residuals

- No perfect multicollinearity
  - Compute VIF
    - If the VIF estimates are relatively high it could be a sign of multicollinearity

Constant variance?

- We have already seen how to plot two variables against each other.
- However, to do that we need to save the residuals and the predicted values.

- In general

```
Proc reg data=dataset;
Model Y = X </options>;
Output out=datasetname p=varname r=varname student=varname;
Run; Quit;
```

- Option: /Hcc calculates HRSE (In SAS 9.2)
- Where:
  - P / predicted: Unstandardized predicted values
  - Student= Studentized/standardized residuals
  - R / residual: Residuals

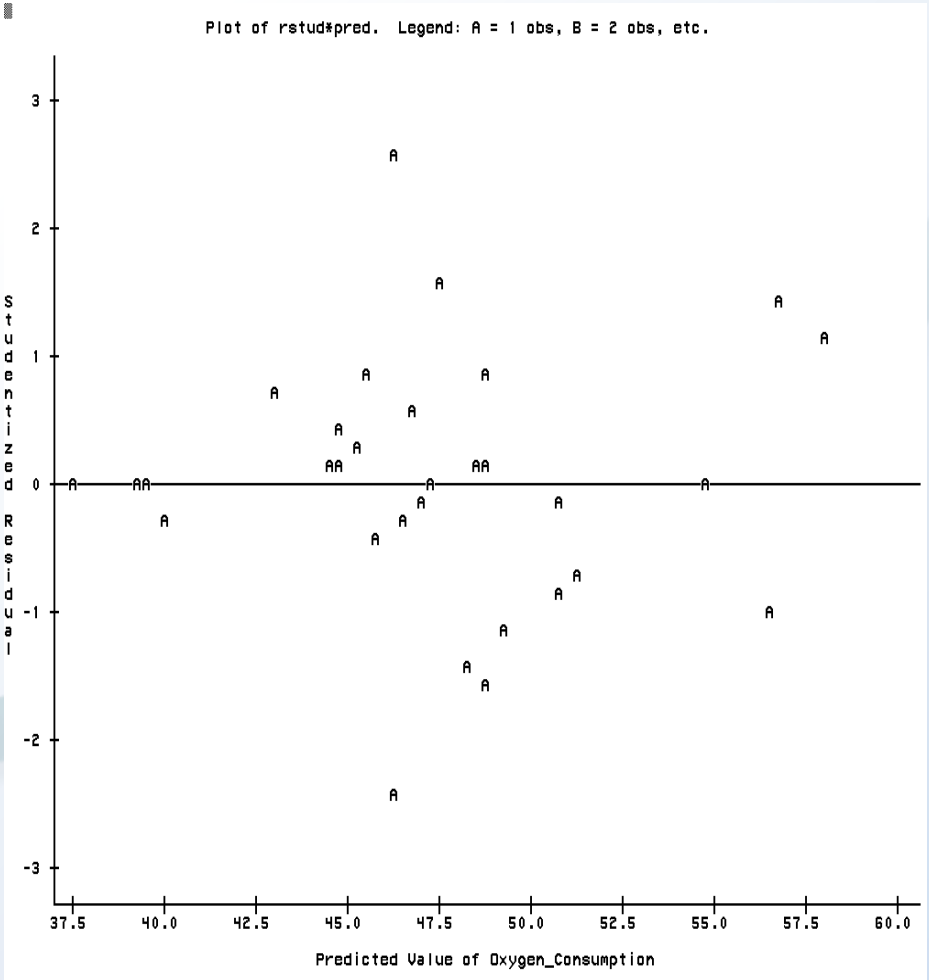
Output

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	93.33753	36.49782	2.56	0.0176
Runtime	1	-2.08804	2.22856	-0.94	0.3585
Age	1	-0.21066	0.10519	-2.00	0.0571
Performance	1	0.25756	1.02373	0.25	0.8036
Maximum_Pulse	1	0.30490	0.13990	2.18	0.0398
Rest_Pulse	1	-0.01389	0.07114	-0.20	0.8469
Run_Pulse	1	-0.36618	0.12299	-2.98	0.0067
Weight	1	-0.07741	0.05681	-1.36	0.1862

Parameter Estimates

Variable	DF	---Heteroscedasticity Consistent-- Standard Error	t Value	Pr > |t|
Intercept	1	19.89129	4.69	0.0001
Runtime	1	1.57811	-1.32	0.1988
Age	1	0.06971	-3.02	0.0061
Performance	1	0.70928	0.36	0.7198
Maximum_Pulse	1	0.10070	3.03	0.0060
Rest_Pulse	1	0.05401	-0.26	0.7994
Run_Pulse	1	0.08787	-4.17	0.0004
Weight	1	0.04601	-1.68	0.1060



Errors are uncorrelated with explanatory variables

- Either:
  - Plot the standardized residuals against unstandardized predicted values
  - Plot the standardized residuals against all the explanatory variables

- This should be fairly easy for you, since you already know how to use Proc Plot and you know how to save the necessary output variables.

Normally distributed residuals

- You already know how to save the standardized residuals.
- You also know how to produce histograms and pp-plots using the Univariate procedure.

Partial output - descriptives

The UNIVARIATE Procedure
Variable: r (Residual)

Moments

N	31	Sum Weights	31
Mean	0	Sum Observations	0
Std Deviation	2.32175508	Variance	5.39054664
Skewness	-0.2006683	Kurtosis	-0.1178117
Uncorrected SS	161.716399	Corrected SS	161.716399
Coeff Variation	.	Std Error Mean	0.41699952

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	2.32176
Median	-0.07841	Variance	5.39055
Mode	.	Range	9.72200
		Interquartile Range	3.48294

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-0.5	Pr >= |M|	1.0000
Signed Rank	S	-4	Pr >= |S|	0.9390

Quantiles (Definition 5)

Quantile	Estimate
100% Max	4.2782917
99%	4.2782917
95%	3.6829807
90%	2.9406003
75% Q3	2.1316197
50% Median	-0.0784141
25% Q1	-1.3513175
10%	-3.1676534
5%	-4.1387784
1%	-5.4437049
0% Min	-5.4437049

The UNIVARIATE Procedure
Fitted Distribution for r

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	2.321755

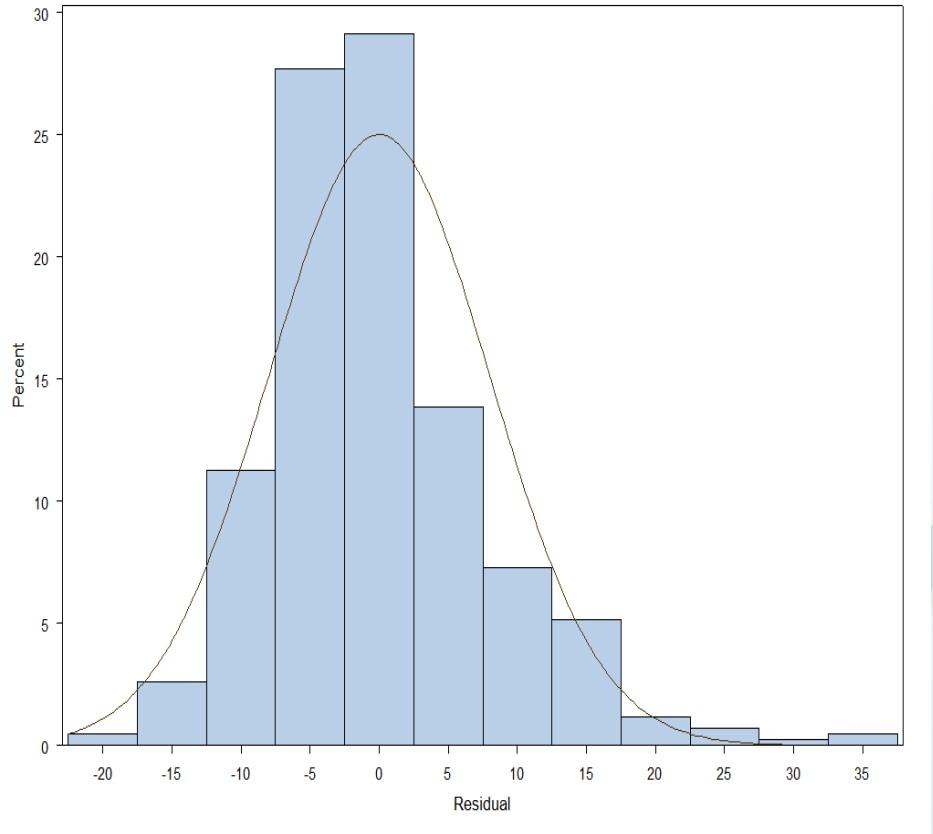
Goodness-of-Fit Tests for Normal Distribution

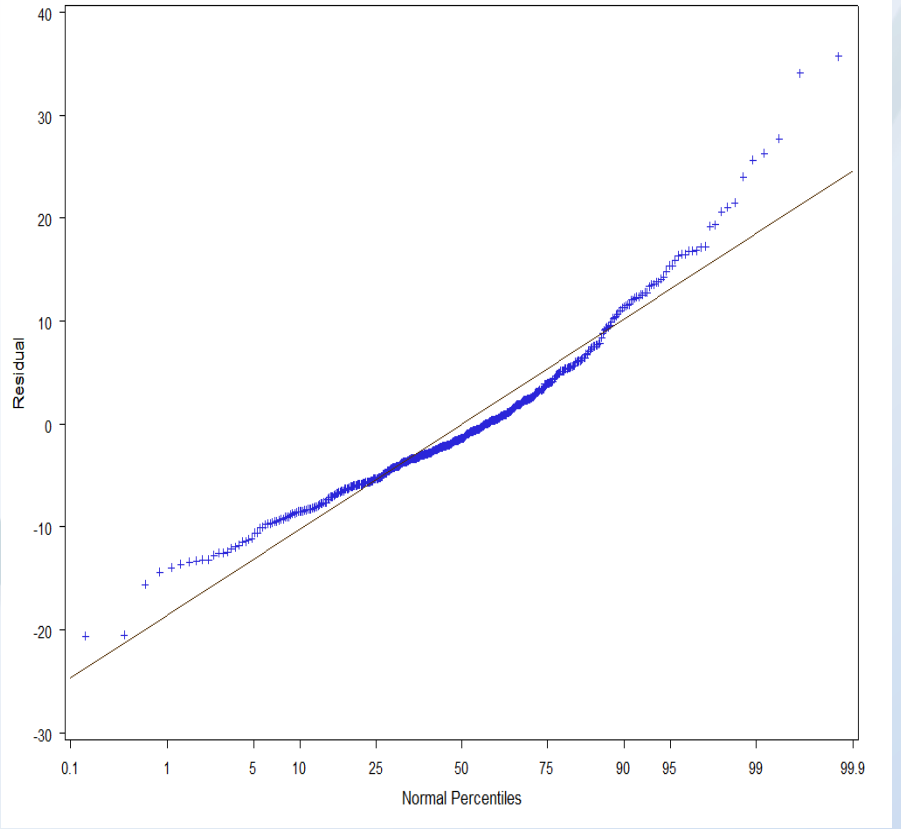
Test	---Statistic----		-----p Value-----	
Kolmogorov-Smirnov	D	0.08823613	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.04125473	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.26992703	Pr > A-Sq	>0.250

Quantiles for Normal Distribution

	------Quantile------	
Percent	Observed	Estimated
1.0	-5.44370	-5.401210
5.0	-4.13878	-3.818947
10.0	-3.16765	-2.975449
25.0	-1.35132	-1.566000
50.0	-0.07841	-0.000000
75.0	2.13162	1.566000
90.0	2.94060	2.975449
95.0	3.68298	3.818947
99.0	4.27829	5.401210

Partial output - Graphs





Assignment 4

- Chech whether the assumptions hold on the final model estimated on the car prices earlier:

  A. Constant variance – Homoskedasticity. Add reference lines at -2 0 2.
  B. Errors are uncorrelated with explanatory variables
  C. Normally distributed residuals and add a 45 degree line to the pp-plot
  D. No perfect multicollinearity

Enhancing your output

- The Output Delivery System (ODS) can be used to enhance your output, and to output your result in various formats
  - HTML
  - PDF
  - Word
  - Excel and many other formats

- In general

```
Ods html/pdf body/file = 'path_filename.fileextension' <options>;
...
Sas code
...
Ods html/pdf close;
```

ODS

- HTML output

```
Ods html body = 'c:\output.html';
Proc univariate data=resids;
var resid;
Histogram resid /normal;
Probplot resid /normal (mu=est sigma=est);
Run; Quit;
Ods html close;
```

- PDF output

```
ods pdf file = 'c:\output.pdf';
Proc univariate data=resids;
var resid;
Histogram resid /normal;
Probplot resid /normal (mu=est sigma=est);
Run;
Quit;
ods pdf close;
```

Output example

- HTML descriptives

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen_Consumption

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance

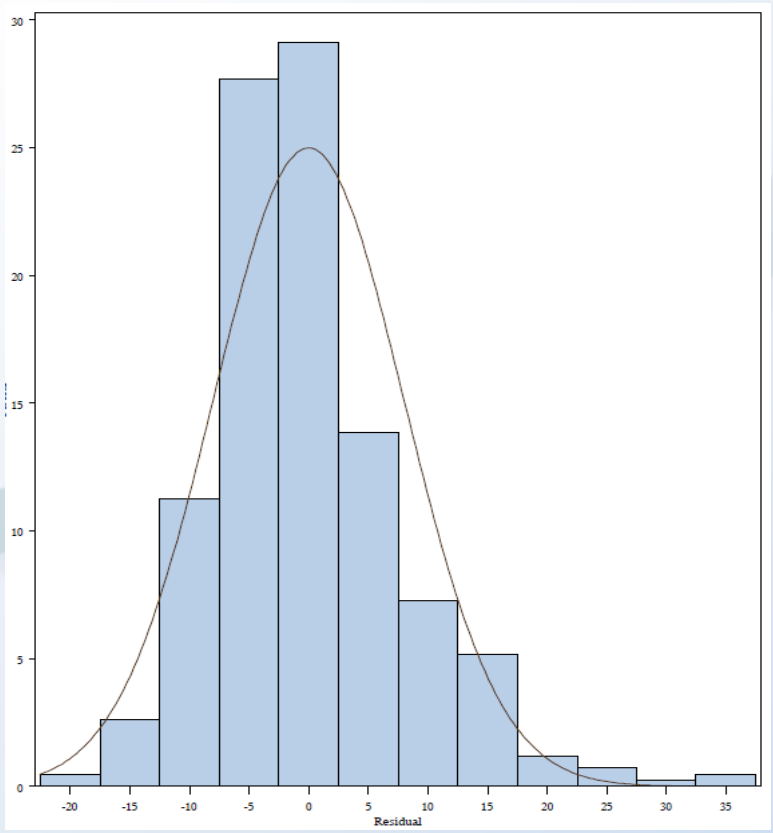
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	689.83816	229.94605	38.39	<.0001
Error	27	161.71640	5.98950		
Corrected Total	30	851.55455			

Root MSE	2.44734	R-Square	0.8101
Dependent Mean	47.37581	Adj R-Sq	0.7890
Coeff Var	5.16581		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	80.90156	8.81466	9.18	<.0001
Runtime	1	-2.97015	0.34528	-8.60	<.0001
Maximum_Pulse	1	0.35458	0.13482	2.63	0.0139
Run_Pulse	1	-0.37549	0.12363	-3.04	0.0052

PDF histogram



Assignment 5

A. Create a HTML-formatted output file for the estimated final regression model.

B. Create a .pdf output file for one of the tests done in assignment 4.

ODS Graphics

- ODS Graphs is an alternative so SAS/GRAPH
- With SAS/GRAPH you can make just about any graph you desire but it often requires lots of lots of code and the language structure is difficult to master.
- It automatically produces some really cool looking statistical graphics in many of the SAS/STAT and SAS/ETS procedures

ODS Graphics

Component	Procedure	Graphs
Base SAS	CORR	3
SAS/STAT	ANOVA CORRSP GAM GENMOD GLM KDE LIFETEST LOESS LOGISTIC MI MIXED PHREG PRINCOMP PRINQUAL REG	1 1 2 2 2 6 9 10 20 3 12 3 6 1 13
SAS/ETS	ARIMA AUTOREG ENTROPY EXPAND MODEL SYSLIN ROBUSTREG TIMESERIES UCM VARIMAX X12	7 10 3 9 8 4 4 24 27 8 3
SAS High-Performance Forecasting	HPF	25

ODS Graphics

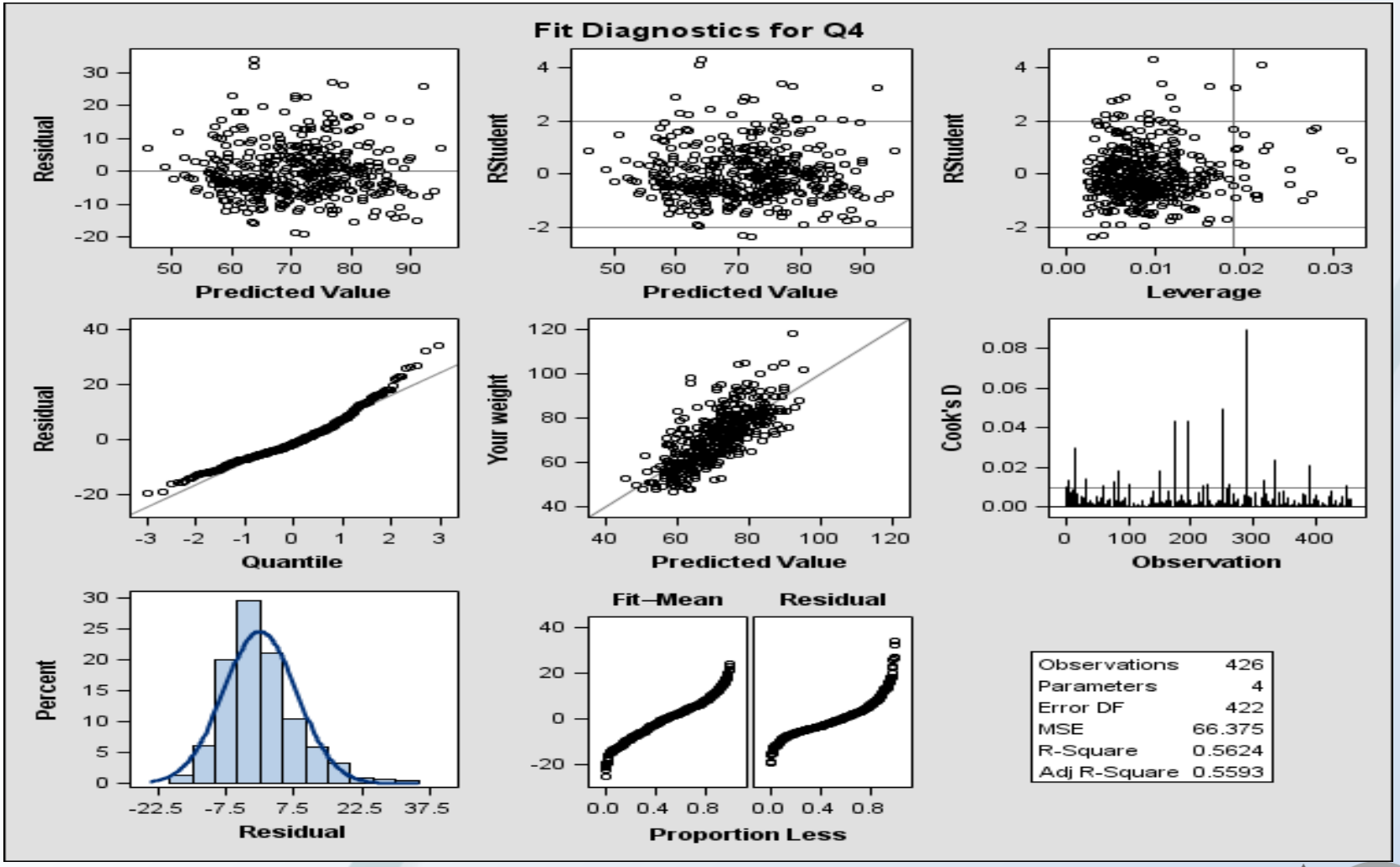
- In general

```
Ods graphics on;
Ods html;
…
SAS code
…
Ods html close;
Ods graphics off;
```

- An example

```
ods graphics on;
ods html;
Proc reg data=stat.dataset;
model q4 = q5 q8 q9;
output out=resdat p=pred student=stdresid;
quit;
ods html close;
ods graphics off;
```

Partial output



Partial output