STAT 615 (Baron)

Data entry Print, plot, simple stats

Refers to the datafile "pulse.dat" (in Unix). In Windows, write the full path.

```
DATA;
INFILE pulse;
INPUT ROW PULSE1 PULSE2 RAN SMOKES SEX HEIGHT WEIGHT ACTIVITY;

PROC PRINT;

PROC PLOT HPERCENT=33 VPERCENT=33;
PLOT WEIGHT*HEIGHT;
PLOT WEIGHT*HEIGHT=SEX;
PLOT PULSE1*WEIGHT='*' PULSE2*WEIGHT='o' / OVERLAY;

PROC MEANS;
VAR PULSE1 PULSE2 SMOKES SEX HEIGHT WEIGHT;

PROC UNIVARIATE NORMAL PLOT;
VAR HEIGHT;

RUN;
```

Reduce the size of plots to 33% (option)

plots two graphs on the same picture

Options verifying Normality

PROC PRINT

Obs	ROW	PULSE1	PULSE2	RAN	SMOKES	SEX	HEIGHT	WEIGHT	ACTIVITY
1	1	64	88	1	2	1	66.00	140	2
2	2	58	70	1	2	1	72.00	145	2
3	3	62	76	1	1	1	73.50	160	3
...	...	...	...	...	...	...	...	...	...
92	92	76	76	2	2	2	61.75	108	2

PROC PLOT

```
Plot of WEIGHT*HEIGHT.  A=1, B=2, etc.

WEIGHT |
  250 +
      |
      |               A
      |
  200 +               A
      |
      |         A     B B C B
      |
      |         B   BA  BAA A
      |
  150 +       C D FAD B D CAA
      |
      | A A   A A G B D A B B
      |
      |   B D   CAA B B
      |
  100 +   B A A
      -+---------+---------+---------+-
      60        65        70        75

                 HEIGHT

Plot of WEIGHT*HEIGHT=SEX.
 (NOTE: 48 obs hidden.)

WEIGHT |
  250 +
      |
      |                        1
      |
  200 +                        1
      |
      |                  1     1 1 1 1
      |
      |                  1   11  111 1
      |
  150 +              1 1 111 1 1 111
      |
      | 2 2   2 2 1 1 2 1 1 1
      |
      |   2 2   222 1 2
      |
  100 +   2 2 2
      -+---------+---------+---------+-
      60        65        70        75

                 HEIGHT

Plot of PULSE1*WEIGHT='*'.
Plot of PULSE2*WEIGHT='o'.
 (NOTE: 90 obs hidden.)

PULSE1 |
  150 +
      |
      |              o
      |
      |            o
      |
      |         oooooo
  100 +         *** *
      |         oooo oo o o
      |       *  *******  *
      |       o    oooooo o
      |        **************  *
      |        oooooooooooo o  o
      |         *********** *
      |         o oooooo  o o
   50 +            ***
      |           o ooo
      --+------+------+------+------+--
       50    100    150    200    250

                  WEIGHT
```

Plots "1" for men "2" for women

The MEANS Procedure

PROC MEANS

Variable	N	Mean	Std Dev	Minimum	Maximum
PULSE1	92	72.8695652	11.0087052	48.0000000	100.0000000
PULSE2	92	80.0000000	17.0937943	50.0000000	140.0000000
SMOKES	92	1.6956522	0.4626519	1.0000000	2.0000000
SEX	92	1.3804348	0.4881540	1.0000000	2.0000000
HEIGHT	92	68.7173913	3.6592907	61.0000000	75.0000000
WEIGHT	92	145.1521739	23.7393978	95.0000000	215.0000000

The UNIVARIATE Procedure
Variable: HEIGHT

PROC UNIVARIATE

Moments

$\bar{X}$ $\quad S$ $\quad \Sigma(X_i-\bar{X})^2$

N	92	Sum Weights	92
Mean	68.7173913	Sum Observations	6322
Std Deviation	3.65929071	Variance	13.3904085
Skewness	-0.2246482	Kurtosis	-0.8044121
Uncorrected SS	435649.875	Corrected SS	1218.52717
Coeff Variation	5.32513042	Std Error Mean	0.38150743

$$\frac{\Sigma(X-\bar{X})^3/n}{S^3}$$ measures symmetry

ΣX_i^2

$\hat{\gamma} = \frac{S}{\mu}\cdot 100\%$

$S/\sqrt{n}$

$$\frac{\Sigma(X-X)^4/n}{S^4}-3$$ measures deviation from Normality

Basic Statistical Measures

Location		Variability	
Mean	68.71739	Std Deviation	3.65929
Median	69.00000	Variance	13.39041
Mode	68.00000	Range	14.00000
		Interquartile Range	6.00000

(max−min) $\quad$ (Q3−Q1)

Confidence intervals and two-sample t-tests

STAT 615 (Baron)

```
DATA C;
INFILE pulse;
INPUT ROW PULSE1 PULSE2 RAN SMOKES SEX HEIGHT WEIGHT ACTIVITY;
DIFF = PULSE2-PULSE1;

PROC TTEST DATA=C;
CLASS SEX;
VAR DIFF;

PROC TTEST DATA=C;
CLASS SMOKES;
VAR DIFF;

RUN;
```

Does the difference in pulses depend on gender? on smoking?

The TTEST Procedure

Statistics

Variable	SEX	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
DIFF	1	57	2.6454	5.4561	8.2669	8.9434	10.593	12.995	1.4031	-8	42
DIFF	2	35	4.0221	9.8571	15.692	13.74	16.986	22.256	2.8712	-14	48
DIFF	Diff (1-2)		-10.11	-4.401	1.3041	11.672	13.373	15.658	2.8717		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
DIFF	Pooled	Equal	90	-1.53	0.1289
DIFF	Satterthwaite	Unequal	50.4	-1.38	0.1745

Two tests for $H_0: \mu_1 = \mu_2$ vs $H_A: \mu_1 \neq \mu_2$

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
DIFF	Folded F	34	56	2.57	0.0017

Test for $H_0: \sigma_1 = \sigma_2$ vs $H_A: \sigma_1 \neq \sigma_2$

The TTEST Procedure

Statistics

Variable	SMOKES	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
DIFF	1	28	1.817	6.1786	10.54	8.8931	11.248	15.31	2.1257	-14	36
DIFF	2	64	3.9501	7.5469	11.144	12.265	14.399	17.439	1.7999	-6	48
DIFF	Diff (1-2)		-7.459	-1.368	4.7226	11.81	13.531	15.843	3.0659		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
DIFF	Pooled	Equal	90	-0.45	0.6565
DIFF	Satterthwaite	Unequal	65.2	-0.49	0.6249

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
DIFF	Folded F	63	27	1.64	0.1579

Contingency Tables and chi-square tests

```
DATA C;
INFILE pulse;
INPUT ROW PULSE1 PULSE2 RAN SMOKES SEX HEIGHT WEIGHT ACTIVITY;

proc freq data=c;
tables (sex smokes)*activity / chisq;

RUN;
```

TABLE OF SEX BY ACTIVITY

SEX	Frequency / Percent / Row Pct / Col Pct	ACTIVITY 0	1	2	3	Total
1	Frequency	1	5	35	16	57
	Percent	1.09	5.43	38.04	17.39	61.96
	Row Pct	1.75	8.77	61.40	28.07	
	Col Pct	100.00	55.56	57.38	76.19	
2	Frequency	0	4	26	5	35
	Percent	0.00	4.35	28.26	5.43	38.04
	Row Pct	0.00	11.43	74.29	14.29	
	Col Pct	0.00	44.44	42.62	23.81	
Total	Frequency	1	9	61	21	92
	Percent	1.09	9.78	66.30	22.83	100.00

$n_{1\cdot} = \sum_j n_{1j}$

$n_{1\cdot}/n_{\cdot\cdot}$

$n_{2\cdot}$

$n_{2\cdot}/n_{\cdot\cdot}$

$n_{\cdot\cdot}$

STATISTICS FOR TABLE OF SEX BY ACTIVITY

Statistic	DF	Value	Prob
Chi-Square	3	3.118	0.374

p-value = $P_{H_0}\{\chi^2 > 3.118\}$

χ^2 - statistic

Sample Size = 92
WARNING: 38% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

$(r-1)(c-1) = (1)(3)$

TABLE OF SMOKES BY ACTIVITY

SMOKES	Frequency / Percent / Row Pct / Col Pct	ACTIVITY 0	1	2	3	Total
1	Frequency	1	3	19	5	28
	Percent	1.09	3.26	20.65	5.43	30.43
	Row Pct	3.57	10.71	67.86	17.86	
	Col Pct	100.00	33.33	31.15	23.81	
2	Frequency	0	6	42	16	64
	Percent	0.00	6.52	45.65	17.39	69.57
	Row Pct	0.00	9.38	65.63	25.00	
	Col Pct	0.00	66.67	68.85	76.19	
Total	Frequency	1	9	61	21	92
	Percent	1.09	9.78	66.30	22.83	100.00

The SAS System 79

STATISTICS FOR TABLE OF SMOKES BY ACTIVITY

Statistic	DF	Value	Prob
Chi-Square	3	2.771	0.428

Sample Size = 92
WARNING: 38% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

$\}$ χ^2 test may be inaccurate.
Therefore, combine the first two columns (on the back)

Monte Carlo method of estimating p-values

```
/* Computing p-value for the Shapiro-Wilk test of normality
   for small samples, using Monte Carlo simulations           */

DATA A;
INFILE pulse;
INPUT row pulse1 pulse2 ran smokes sex height weight activity;

   PROC PRINT DATA=A; TITLE 'Data set A. Initial sample';

PROC UNIVARIATE DATA=A NOPRINT;      /* supresses the output       */
VAR height;
OUTPUT OUT=B NORMAL=Wobs N=N;        /* save Shapiro-Wilk statistic and sample size */

DATA B; SET B; Nruns = 1000;         /* Nruns = number of MC runs  */

   PROC PRINT DATA=B; TITLE 'Data set B. Observed statistic';

DATA MC; SET B;
RETAIN SEED 1660519;
DO MCrun=1 TO Nruns;
  DO j=1 TO N;                       /* Generate Nruns samples of  */
    CALL RANNOR(SEED, X);            /* size N of normal variables */
    OUTPUT;
  END;
END;

   PROC PRINT DATA=MC; TITLE 'Data set MC. Simulated samples';

PROC UNIVARIATE DATA=MC NOPRINT;
VAR X;
CLASS MCrun;                         /* Compute Shapiro-Wilk statistic W */
OUTPUT OUT=C NORMAL=W;               /* for each sample                  */

DATA B; SET B;                       /* Make K to have the same    */
DO i=1 TO Nruns; OUTPUT; END;        /* dimension as Kgenerated    */

DATA D; MERGE B C;                   /* The indicator is 1 if      */
indicator = (W < Wobs);              /* W < Wobs and 0 if W > Wobs */

   PROC PRINT DATA=D; TITLE 'Data set D. Results of simulations';

PROC MEANS DATA=D NOPRINT;           /* The probability of W<Wobs  */
VAR indicator;                       /* is estimated by the        */
OUTPUT OUT=E MEAN=Pvalue;            /* proportion of W < Wobs     */

   PROC PRINT DATA=E; TITLE 'Estimated p-value';
   VAR Pvalue;                       /* Report the p-value         */

RUN;
```

==

Data set A. Initial sample

Obs	row	pulse1	pulse2	ran	smokes	sex	height	weight	activity
1	1	64	88	1	2	1	66.00	140	2
.	.	.	.	.	.	.	.	.	.
92	92	76	76	2	2	2	61.75	108	2

Data set B. Observed statistic

Obs	N	Wobs	Nruns
1	92	0.96656	10000

Data set MC. Simulated samples

Obs	N	Wobs	Nruns	SEED	MCrun	j	X
1	92	0.96656	10000	1597779957	1	1	-0.02050
.	.	.	.	.	.	.	.
920000	92	0.96656	10000	404972177	10000	92	0.33889

Data set D. Results of simulations

Obs	N	Wobs	Nruns	i	MCrun	W	indicator
1	92	0.96656	10000	1	1	0.99086	0
.	.	.	.	.	.	.	.
10000	92	0.96656	10000	10000	10000	0.99215	0

Estimated p-value

Obs	Pvalue
1	0.0181

LINEAR REGRESSION

STAT 615 (Baron)

```
DATA pulses;
INFILE pulse;
INPUT row pulse1 pulse2 ran smokes gender height weight activity;

PROC REG;                          /* Basic form of the procedure          */
MODEL weight=height;  run;         /* The model is:  response = predictor(s) */

PROC REG DATA=pulses  ALPHA=0.10  LINEPRINTER;
                    /* Lineprinter puts graphs into the same output file */
MODEL weight=height / CLB CLI CLM;
/* (1-α)-confidence intervals for parameters and mean values and prediction intervals */
TEST  height=3;                                   /* Test Ho: slope = 3 */
PLOT  weight*height='*'  PREDICTED.*height = 'o' / OVERLAY;
      /* Plot observed and predicted values by different symbols on the same picture. */
PLOT  RESIDUAL.*height;                           /*  Plot residuals.  */
RUN;                                              /* End of Proc Reg.  */
```

The REG Procedure
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	31592	31592	144.38	<.0001
Error	90	19692	218.80247		
Corrected Total	91	51284			

Root MSE	14.79197	R-Square	0.6160
Dependent Mean	145.15217	Adj R-Sq	0.6117

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
Intercept	1	-204.74078	29.15972	-7.02	<.0001	-253.20310	-156.27845
height	1	5.09177	0.42375	12.02	<.0001	4.38751	5.79602

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	90% CL Mean		90% CL Predict		Residual
1	140.0000	131.3159	1.9246	128.1172	134.5145	106.5249	156.1068	8.6841
2	145.0000	161.8665	2.0768	158.4149	165.3180	137.0416	186.6913	-16.8665
3	160.0000	169.5041	2.5467	165.2716	173.7366	144.5587	194.4495	-9.5041
. . .								
91	150.0000	136.4076	1.7053	133.5735	139.2417	111.6611	161.1541	13.5924
92	108.0000	109.6758	3.3309	104.1400	115.2117	84.4766	134.8751	-1.6758

Sum of Residuals	0
Sum of Squared Residuals	19692
Predicted Residual SS (PRESS)	20624

Simple Linear Regression

STAT 615 (Baron)

```
DATA C;
INFILE pulse;
INPUT row pulse1 pulse2 ran smokes gender height weight activity;

PROC REG LINEPRINTER;
MODEL weight=height / CLI CLM;
PLOT weight*height='*' PREDICTED.*height = 'O' / OVERLAY;
PLOT RESIDUAL.*height;

RUN;
```

asks for plots in the same file as all outputs

CLI – prediction intervals for individual responses

CLM – confidence intervals for mean responses for each X.

Plot $\{ y_i$ vs X_i, $\hat{y}_i$ vs $X_i \}$ together, then e_i vs X_i

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	31592	31592	144.38	<.0001
Error	90	19692	218.80247		
Corrected Total	91	51284			

ANOVA components and F-test

$S \longrightarrow$ Root MSE	14.79197	R-Square	0.6160
Dependent Mean	145.15217	Adj R-Sq	0.6117
Coeff Var	10.19067		

Parameter Estimates

	Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
$b_0 \longrightarrow$	Intercept	1	-204.74078	29.15972	-7.02	<.0001
$b_1 \longrightarrow$	height	1	5.09177	0.42375	12.02	<.0001

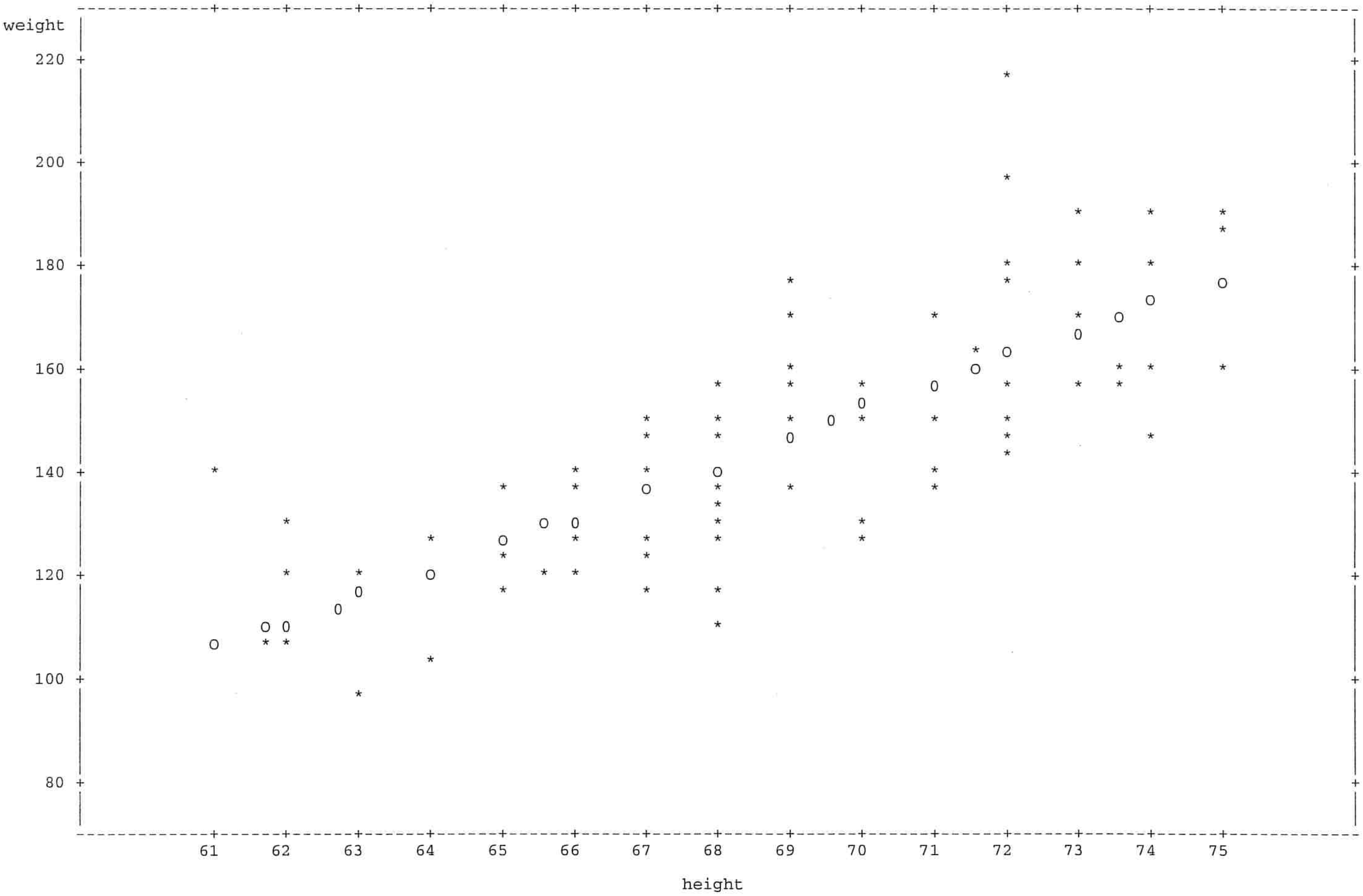
p-values for $H_0: \beta_0 = 0$, $H_0: \beta_1 = 0$

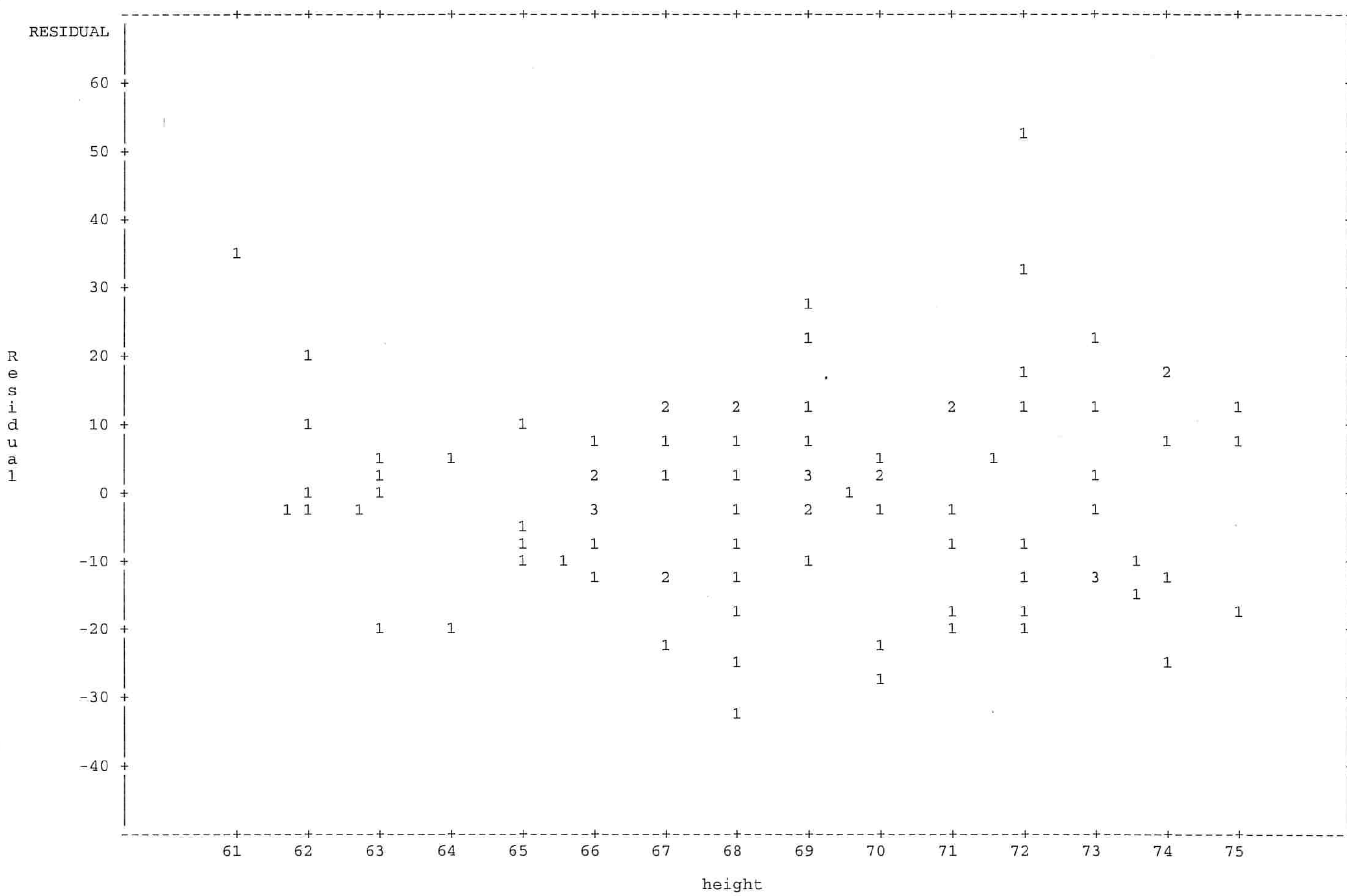
The REG Procedure
Model: MODEL1
Dependent Variable: weight

Output Statistics

Obs	Dep Var weight (y_i)	Predicted Value ($\hat{y}_i$)	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	140.0000	131.3159	1.9246	127.4922	135.1395	101.6813	160.9504	8.6841
2	145.0000	161.8665	2.0768	157.7405	165.9924	132.1914	191.5415	-16.8665
3	160.0000	169.5041	2.5467	164.4447	174.5635	139.6849	199.3233	-9.5041
4	190.0000	166.9582	2.3815	162.2269	171.6895	137.1930	196.7235	23.0418
5	155.0000	146.5912	1.5468	143.5181	149.6642	117.0441	176.1382	8.4088
6	165.0000	166.9582	2.3815	162.2269	171.6895	137.1930	196.7235	-1.9582
7	150.0000	161.8665	2.0768	157.7405	165.9924	132.1914	191.5415	-11.8665
8	190.0000	172.0500	2.7183	166.6496	177.4504	142.1711	201.9289	17.9500
9	195.0000	161.8665	2.0768	157.7405	165.9924	132.1914	191.5415	33.1335
10	138.0000	156.7747	1.8204	153.1581	160.3912	127.1661	186.3832	-18.7747
11	160.0000	172.0500	2.7183	166.6496	177.4504	142.1711	201.9289	-12.0500
12	155.0000	161.8665	2.0768	157.7405	165.9924	132.1914	191.5415	-6.8665
13	153.0000	151.6829	1.6351	148.4344	154.9314	122.1171	181.2488	1.3171
14	145.0000	136.4076	1.7053	133.0198	139.7954	106.8261	165.9891	8.5924
15	170.0000	156.7747	1.8204	153.1581	160.3912	127.1661	186.3832	13.2253
16	175.0000	161.8665	2.0768	157.7405	165.9924	132.1914	191.5415	13.1335
17	175.0000	146.5912	1.5468	143.5181	149.6642	117.0441	176.1382	28.4088
18	170.0000	166.9582	2.3815	162.2269	171.6895	137.1930	196.7235	3.0418
19	180.0000	172.0500	2.7183	166.6496	177.4504	142.1711	201.9289	7.9500
20	135.0000	131.3159	1.9246	127.4922	135.1395	101.6813	160.9504	3.6841
21	170.0000	156.7747	1.8204	153.1581	160.3912	127.1661	186.3832	13.2253
22	157.0000	151.6829	1.6351	148.4344	154.9314	122.1171	181.2488	5.3171
23	130.0000	151.6829	1.6351	148.4344	154.9314	122.1171	181.2488	-21.6829
24	185.0000	177.1418	3.0767	171.0294	183.2541	147.1260	207.1575	7.8582
25	140.0000	105.8570	3.6156	98.6739	113.0401	75.6050	136.1090	34.1430
26	120.0000	131.3159	1.9246	127.4922	135.1395	101.6813	160.9504	-11.3159
27	130.0000	141.4994	1.5718	138.3766	144.6221	111.9471	171.0517	-11.4994
28	138.0000	141.4994	1.5718	138.3766	144.6221	111.9471	171.0517	-3.4994
29	121.0000	116.0405	2.8719	110.3350	121.7461	86.1050	145.9761	4.9595
30	125.0000	151.6829	1.6351	148.4344	154.9314	122.1171	181.2488	-26.6829
31	116.0000	141.4994	1.5718	138.3766	144.6221	111.9471	171.0517	-25.4994
32	145.0000	146.5912	1.5468	143.5181	149.6642	117.0441	176.1382	-1.5912
33	150.0000	146.5912	1.5468	143.5181	149.6642	117.0441	176.1382	3.4088
34	112.0000	114.7676	2.9618	108.8834	120.6518	84.7975	144.7378	-2.7676
35	125.0000	141.4994	1.5718	138.3766	144.6221	111.9471	171.0517	-16.4994
36	190.0000	172.0500	2.7183	166.6496	177.4504	142.1711	201.9289	17.9500
37	155.0000	156.7747	1.8204	153.1581	160.3912	127.1661	186.3832	-1.7747
38	170.0000	146.5912	1.5468	143.5181	149.6642	117.0441	176.1382	23.4088
39	155.0000	151.6829	1.6351	148.4344	154.9314	122.1171	181.2488	3.3171
40	215.0000	161.8665	2.0768	157.7405	165.9924	132.1914	191.5415	53.1335
.....								
90	110.0000	110.9488	3.2374	104.5171	117.3804	80.8664	141.0312	-0.9488
91	150.0000	136.4076	1.7053	133.0198	139.7954	106.8261	165.9891	13.5924
92	108.0000	109.6758	3.3309	103.0584	116.2933	79.5531	139.7985	-1.6758
				$\underbrace{\qquad}_{\text{conf. intervals for } E(y/X_i)}$		$\underbrace{\qquad}_{\text{prediction intervals for } y_i}$		$\underbrace{\qquad}_{e_i = y_i - \hat{y}_i}$

Sum of Residuals	0
Sum of Squared Residuals	19692
Predicted Residual SS (PRESS)	20624


weight
220
200
180
160
140
120
100
80
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
height


RESIDUAL
60
50
40
30
20
10
0
-10
-20
-30
-40
Residual
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
height

STAT 615 (Baron)

Data of type "Character" and PROC IMPORT

```
DATA presidents;                                /*   COMMENTS:                                        */
Length name $ 40;                               /* Introduce a string variable of at most 40 characters */
Infile presid firstobs=6 PAD;                   /* Tells SAS to read data starting from line 6          */

     /*  Option PAD tells SAS to fill the rest of each line with spaces, if variable "name" appears shorter than 40 characters  */

Input expected 1-4 actual 7-10 name $ 13-40;    /* Tells SAS that the first 4 columns contain variable  */
                                                /* "expected", columns 7-10 contain "actual", columns   */
                                                /* 13-33 contain "name"                                 */
                                                /* The dollar "$" says that "name" is a string variable */

PROC PRINT data=presidents;
run;
```

Obs	name	expected	actual
1	ANDREW JOHNSON	17.2	10.3
2	ULYSSES S. GRANT	22.8	16.4
3	RUTHERFORD B. HAYES	18.0	15.9
4	JAMES A. GARFIELD	21.2	0.5
5	CHESTER A. ARTHUR	20.1	5.2
6	GROVER CLEVELAND	22.1	23.3
7	BENJAMIN HARRISON	17.2	12.0
8	WILLIAM MCKINLEY	18.2	4.5
[illegible]	THEODORE ROOSEVELT	26.1	17.3
[illegible]	WILLIAM H. TAFT	20.3	21.2
11	WOODROW WILSON	17.1	10.9
12	WARREN G. HARDING	18.1	2.4
13	CALVIN COOLIDGE	21.4	9.4
14	HERBERT C. HOOVER	19.0	35.6
[illegible]	FRANKLIN D. ROOSEVELT	21.7	12.1
[illegible]	HARRY S. TRUMAN	15.3	27.7
17	DWIGHT D. EISENHOWER	14.7	16.2
18	JOHN F. KENNEDY	28.5	2.8
19	LYNDON B. JOHNSON	19.3	9.2

===

/* Reading from Excel or other formats */

```
PROC IMPORT DATAFILE = "C:\...\...\presid.xls"    /* Full path                           */
     OUT = presidents                             /* Name the new dataset                */
     DBMS = EXCEL                                 /* Type of external file               */
     REPLACE;
    GETNAMES=YES;                                 /* Read variable names from the 1st row  */
run;


/* or, in SAS, use File -> Import data… (ask to see the generated code)                   */
```

STAT 615
(Baron)

```
/* Monte Carlo study for the estimation of the actual coverage probability
   of a standard confidence interval about the population proportion        */

DATA A;
SEED   = 64378492;
N      = 15;
p      = 0.2;
alpha  = 0.05;
Nruns  = 100000;

DO MCrun = 1 to Nruns;
  CALL RANBIN(SEED, N, p, X);
  Phat    = X/N;
  Margin = PROBIT(1-alpha/2) * sqrt( Phat*(1-Phat)/N );
  Left_Bound  = Phat - Margin;
  Right_Bound = Phat + Margin;
  Coverage    = ( p GE Left_Bound  AND  p LE Right_Bound );
OUTPUT;
END;

PROC MEANS DATA=A N MEAN STDERR; VAR N p alpha Phat Coverage;
run;
```

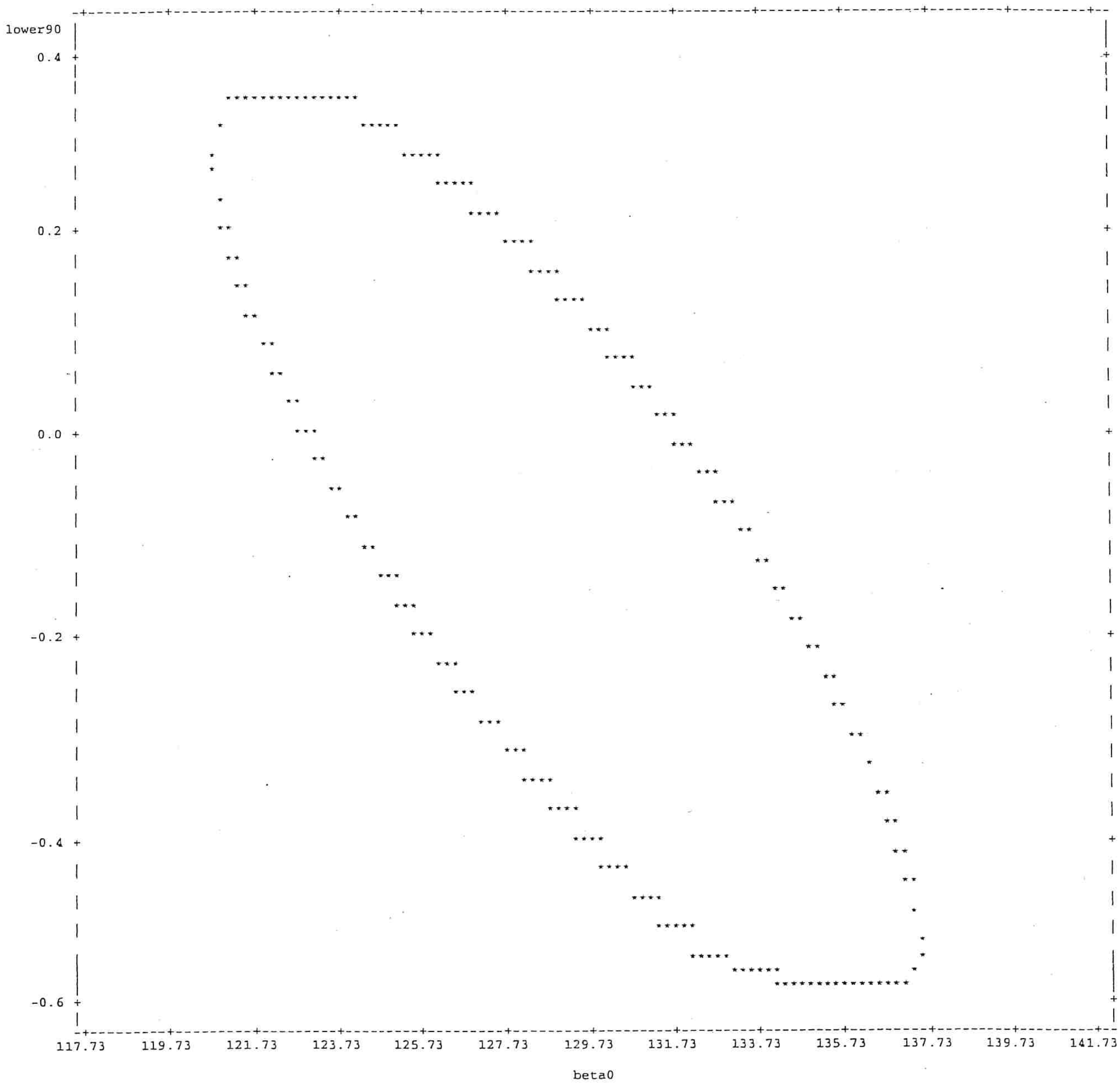
The MEANS Procedure

Variable	N	Mean	Std Error
N	100000	15.0000000	0
p	100000	0.2000000	0
alpha	100000	0.0500000	0
Phat	100000	0.2000280	0.000326764
Coverage	100000	0.8149200	0.0012281

STAT 615 (Baron)

Confidence ellipse for (β_0, β_1)

```
DATA c;   INFILE peru;
INPUT number age years weight height chin calf forearm pulse systol;

PROC MEANS NOPRINT;
VAR years;      OUTPUT OUT=x n=n              /* sample size   */
                         mean=xbar            /* mean of X_i   */
                          css=Sxx;            /* sum of X_i**2 corrected for the mean (Sxx) */

PROC REG DATA=c OUTEST=est NOPRINT;
MODEL systol = years;

data est; set est;
   s = _RMSE_;           /* root MSE = estimated standard deviation */
   b0 = INTERCEPT;       /* estimated intercept b0 */
   b1 = years;           /* estimated slope b1 */
KEEP s b0 b1;

DATA ellipse; MERGE x est;
    sb1 = s/sqrt(Sxx);               /* standard deviation of b1 */
    sb0 = s*sqrt(1/n+xbar**2/Sxx);  /* standard deviation of b0 */
    F90 = 2*finv(0.90,2,n)*s**2;    /* 90% upper bounds for the quadratic form */

do beta0=b0-3*sb0 by 0.1 to b0+3*sb0;
   D = (n*xbar*(beta0-b0))**2 - (n*xbar**2+Sxx)*(n*(beta0-b0)**2-F90);  /* discriminant */
   if D < 0 then do;          upper90 = .; lower90 = .;          end;
   else do;        upper90 = b1+(-n*xbar*(beta0-b0)+sqrt(D))/(n*xbar**2+Sxx);
                   lower90 = b1+(-n*xbar*(beta0-b0)-sqrt(D))/(n*xbar**2+Sxx); end;
OUTPUT; END;

PROC PLOT DATA=ellipse;
PLOT (lower90 upper90)*beta0 = '*' / overlay box;
RUN;
```



STAT 615 (Baron)

Regression Diagnostics – I

```
PROC REG LINEPRINTER;
MODEL weight=height / R ;
PAINT RSTUDENT. >2 OR RSTUDENT.<-2 / SYMBOL='*';
PLOT weight*height RSTUDENT.*PREDICTED. ;
REWEIGHT RSTUDENT.>2 OR RSTUDENT.<-2 / WEIGHT = 0.25;
REFIT;
PRINT;
OUTPUT OUT=D RSTUDENT=R;

PROC ANOVA DATA=C;
CLASS height;
MODEL weight=height;

PROC MODEL DATA=C;
PARMS b0 b1;
weight=b0+b1*height;
fit weight / WHITE BREUSCH=(height);
fit weight / BREUSCH=(gender smokes);

PROC UNIVARIATE DATA=D NORMAL PLOT;
VAR R;
RUN;
```

R – analysis of residuals

identify potential outliers and paint them on the plot; reweight

To obtain SS_{PE} for Lack-of-fit test

Testing homoscedasticity

Testing Normality

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	31592	31592	144.38	<.0001
Error	90	19692	218.80247		
Corrected Total	91	51284			

$SSE = 19692$

Root MSE	14.79197	R-Square	0.6160
Dependent Mean	145.15217	Adj R-Sq	0.6117
Coeff Var	10.19067		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-204.74078	29.15972	-7.02	<.0001
height	1	5.09177	0.42375	12.02	<.0001

Output Statistics

Obs	Dep Var weight	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	140.0000	131.3159	1.9246	8.6841	14.666	0.592	| |* |	0.003
2	145.0000	161.8665	2.0768	-16.8665	14.645	-1.152	| **| |	0.013
3	160.0000	169.5041	2.5467	-9.5041	14.571	-0.652	| *| |	0.006
4	190.0000	166.9582	2.3815	23.0418	14.599	1.578	| |*** |	0.033
. . .								
91	150.0000	136.4076	1.7053	13.5924	14.693	0.925	| |* |	0.006
92	108.0000	109.6758	3.3309	-1.6758	14.412	-0.116	| | |	0.000
i	y_i	$\hat{y}_i$	$s_{\hat{y}_i}$	e_i	s_{e_i}	r_i		

One * for each 0.5 deviation of r_i from 0.

Weighted Least Squares (reweighted outliers)

Weight: REWEIGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	29692	29692	177.18	<.0001
Error	90	15082	167.58030		
Corrected Total	91	44775			

Root MSE	12.94528	R-Square	0.6632
Dependent Mean	144.48315	Adj R-Sq	0.6594
Coeff Var	8.95972		

!

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-203.66894	26.19112	-7.78	<.0001
height	1	5.06527	0.38053	13.31	<.0001

Lack-of-fit test:

The ANOVA Procedure

Class Level Information

Partition 92 obs. into 21 groups

Class	Levels	Values
height	21	61 61.75 62 62.75 63 64 65 65.5 66 67 68 69 69.5 70 71 71.5 72 73 73.5 74 75

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	35192.59814	1759.62991	7.76	<.0001
Error	71	16091.27143	226.63763		
Corrected Total	91	51283.86957			

$SS_{PE} = 16091$

Then $F_{LOF} = \dfrac{(19692-16091)/19}{16091/71}$

The MODEL Procedure

No evidence against H_0: $Var\ \varepsilon_i = const$

Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
weight	White's Test	2.05	2	0.3597	Cross of all vars
	Breusch-Pagan	1.60	1	0.2062	height, 1

Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
weight	Breusch-Pagan	2.08	2	0.3530	gender, smokes, 1

Try different models for non-constant variance

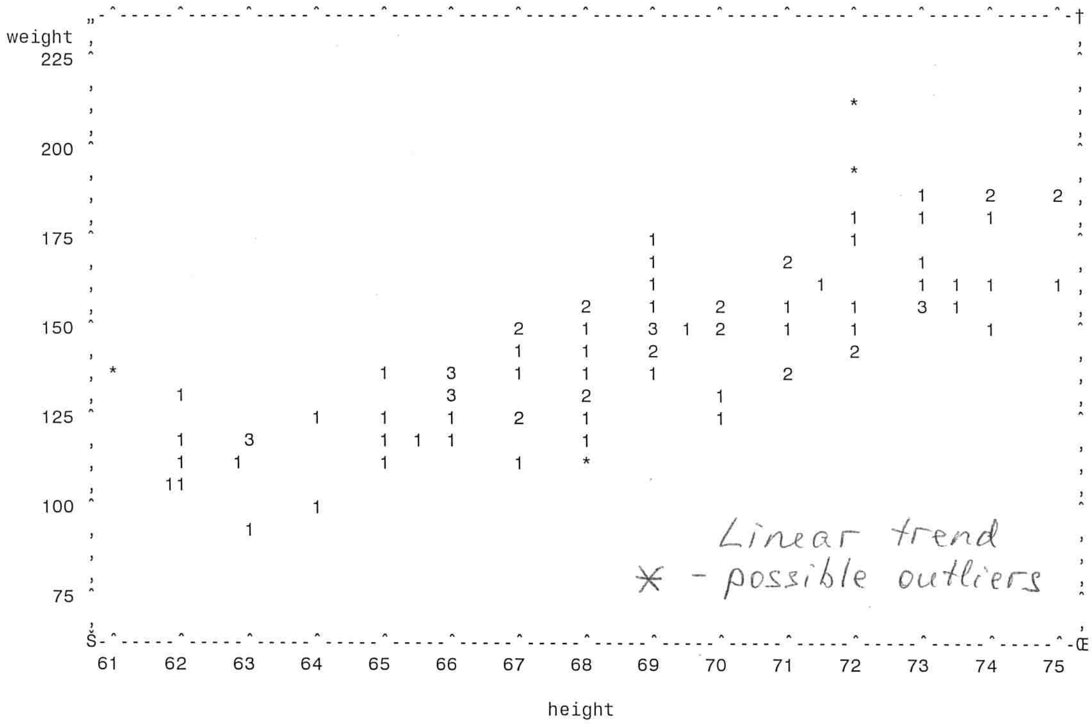
The UNIVARIATE Procedure

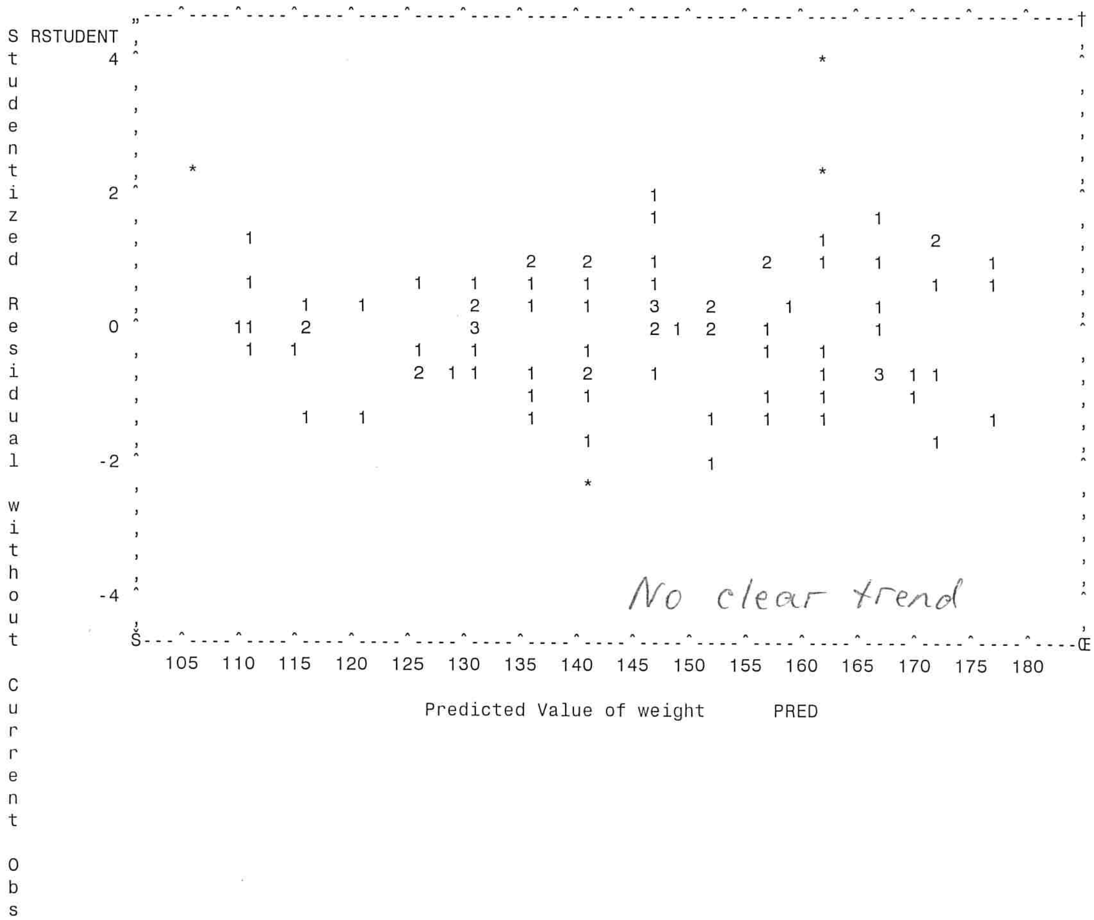
Variable: R (Studentized Residual without Current Obs)

Tests for Normality

No evidence against H_0: $\varepsilon_i \sim Normal$

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.984703	Pr < W	0.3585
Kolmogorov-Smirnov	D	0.072229	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.067134	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.414496	Pr > A-Sq	>0.2500


weight
225
200
175
150
125
100
75
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
height
Linear trend
* - possible outliers


Studentized Residual without Current Obs
RSTUDENT
4
2
0
-2
-4
105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180
Predicted Value of weight PRED
No clear trend

Multivariate Regression

DATA C; INFILE homes FIRSTOBS=3;
INPUT home price footage land rooms baths;
PROC PRINT;
PROC REG;
MODEL price = footage land rooms baths / XPX I;
RUN;

asks for the matrix (X^TX)

asks for $(X^TX)^{-1}$

$p-1=4$ predictors

Response

Data

Obs	home	price	footage	land	rooms	baths
1	1	179000	3060	0.7500	8	2.0
2	2	126500	1600	0.2600	8	1.5
3	3	134500	2000	0.7000	8	1.0
4	4	125000	1300	0.6500	5	1.0

. . . .

The REG Procedure

X^TX

n

Model Crossproducts X'X X'Y Y'Y

Variable	Intercept	footage	land	rooms	baths	price
Intercept	149	271830	147.1521	1080	285	22843256
footage	271830	547281406	291118.2872	2079488	567776	44221525040
land	147.1521	291118.2872	708.21780849	1081.7646	311.847	28341164.25
rooms	1080	2079488	1081.7646	8164	2170	170893748
baths	285	567776	311.847	2170	620	46590190
price	22843256	44221525040	28341164.25	170893748	46590190	3.755274E12

$\sum_i X_{ij}$

X^Ty

y^Ty

Number of Observations Read	150
Number of Observations Used	149
Number of Observations with Missing Values	1

$(X^T X)^{-1}$

X'X Inverse, Parameter Estimates, and SSE

Variable	Intercept	footage	land	rooms	baths	price
Intercept	0.1842932787	0.0000391793	-0.002247295	-0.033112733	-0.003569677	53664.375468
footage	0.0000391793	8.9792995E-8	-1.528842E-6	-0.000020268	-0.000028532	29.432598777
land	-0.002247295	-1.528842E-6	0.0018551891	0.0006053889	-0.000618887	8258.0550804
rooms	-0.033112733	-0.000020268	0.0006053889	0.0098556114	-0.001017169	1344.5862531
baths	-0.003569677	-0.000028532	-0.000618887	-0.001017169	0.0332541747	14664.061156
price	53664.375468	29.432598777	8258.0550804	1344.5862531	14664.061156	80824871495
	b_0	b_1	b_2	b_3	b_4	SS_{Err}

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.723394E11	43084862498	76.76	<.0001
Error	144	80824871495	561283830		
Corrected Total	148	2.531643E11			

Root MSE	23691	R-Square	0.6807
Dependent Mean	153310	Adj R-Sq	0.6719
Coeff Var	15.45324		

Parameter Estimates

	Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
b_0 →	Intercept	1	53664	10171	5.28	<.0001
b_j	footage	1	29.43260	7.09925	4.15	<.0001
	land	1	8258.05508	1020.43502	8.09	<.0001
	rooms	1	1344.58625	2351.97689	0.57	0.5684
	baths	1	14664	4320.30445	3.39	0.0009

STAT 615 (Baron)

Regression of $(y_i - \bar{y})$ on $(X_i - \bar{X})$

```
DATA c;
INFILE pulse;
INPUT row pulse1 pulse2 ran smokes sex height weight activity;

PROC REG;
MODEL height=weight;
title 'Regression HEIGHT on (I) WEIGHT';

PROC REG NOPRINT;
MODEL height = ;
OUTPUT OUT=c RESIDUAL=e;

PROC REG NOPRINT;
MODEL weight = ;
OUTPUT OUT=c RESIDUAL=e2;

PROC REG;
MODEL e=e2;
TITLE 'Regression RESID (height on (I)) on RESID (weight on (I))';

RUN;
=============================================================================
                    Regression HEIGHT on (I) WEIGHT
Model: MODEL1
Dependent Variable: HEIGHT
                           Sum of         Mean
  Source          DF      Squares       Square      F Value       Prob>F

  Model            1    750.63136    750.63136      144.384       0.0001
  Error           90    467.89581      5.19884
  C Total         91   1218.52717

      Root MSE       2.28010     R-square       0.6160
      Dep Mean      68.71739     Adj R-sq       0.6117
      C.V.           3.31808


                    Parameter      Standard    T for H0:
   Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

   INTERCEP   1     51.156502    1.48066539        34.550        0.0001
   WEIGHT     1      0.120983    0.01006846        12.016        0.0001

         Regression RESID (height on (I)) on RESID (weight on (I))

Model: MODEL1
Dependent Variable: E          Residual

                           Sum of         Mean
  Source          DF      Squares       Square      F Value       Prob>F

  Model            1    750.63136    750.63136      144.384       0.0001
  Error           90    467.89581      5.19884
  C Total         91   1218.52717

      Root MSE       2.28010     R-square       0.6160
      Dep Mean      -0.00000     Adj R-sq       0.6117
      C.V.      -1.230098E17


                    Parameter      Standard    T for H0:
   Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

   INTERCEP   1  -6.20203E-16    0.23771655        -0.000        1.0000
   E2         1      0.120983    0.01006846        12.016        0.0001
```

```
DATA C;
INFILE squid;
INPUT x1 x2 x3 x4 x5 y;

proc reg;
model y=x1 x2 x3 x4 x5;
title 'Regression Y on (I) x1 x2 x3 x4 x5';

proc reg noprint;
model y=x1 x2;
output out=c residual=e1;

proc reg noprint;
model x3 x4 x5=x1 x2;
output out=c residual=e3 e4 e5;

proc reg;
model e1=e3 e4 e5 / noint;
title 'Regression RESIDUALS(y on (i) x1 x2) on RESIDUALS(x3 x4 x5 on (i) x1 x2)';
run;
```

STAT 615 (Baron)

Extra SS principle

===

Regression Y on (I) x1 x2 x3 x4 x5

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	208.00723	41.60145	84.070	0.0001
Error	16	7.91752	0.49485		
C Total	21	215.92475			

SSErr (Full)

Root MSE	0.70345	R-square	0.9633
Dep Mean	4.19500	Adj R-sq	0.9519

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-6.512215	0.93356074	-6.976	0.0001
X1	1	1.999413	2.57333840	0.777	0.4485
X2	1	-3.675096	2.77365962	-1.325	0.2038
X3	1	2.524486	6.34749500	0.398	0.6961
X4	1	5.158082	3.66028320	1.409	0.1779
X5	1	14.401162	4.85599380	2.966	0.0091

Estimated Model 1 (Full)

Regression RESIDUALS(y on (i) x1 x2) on RESIDUALS(x3 x4 x5 on (i) x1 x2)

NOTE: No intercept in model. R-square is redefined.

Ex SS

SSErr (Full)

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	8.73523	2.91174	6.620	0.0033
Error	18	7.91752	0.43986		
C Total	21	16.65275			

SSErr (Red.)

Root MSE	0.64553	R-square	0.5246
Dep Mean	0.00000	Adj R-sq	0.4495

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
E3	1	2.524486	5.98447568	0.422	0.6781
E4	1	5.158082	3.45094810	1.495	0.1523
E5	1	14.401162	4.57827486	3.146	0.0056

Estimated Model 4 (Regression of residuals)

Model Selection

```
DATA;  INFILE squid;
INPUT x1 x2 x3 x4 x5 y;

PROC RSQUARE;                                        /* R-square etc. for all models  */
MODEL y=x1 x2 x3 x4 x5 / ADJRSQ CP SSE MSE;

PROC REG;                                            /* Stepwise election of the best */
MODEL y=x1 x2 x3 x4 x5 / SELECTION=STEPWISE;         /*       set of predictors       */

PROC REG;                                            /* Backward selection with       */
model y=x1 x2 x3 x4 x5 / SELECTION=BACKWARD          /* specified significance levels */
                     SLENTRY=0.05 SLSTAY=0.05;
RUN;
```

The RSQUARE Procedure
R-Square Selection Method

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	SSE	Variables in Model
1	0.9455	0.9428	5.7794	0.58836	11.76713	x5
1	0.9223	0.9184	15.9084	0.83897	16.77941	x1
1	0.9140	0.9097	19.5335	0.92866	18.57326	x3
1	0.8857	0.8800	31.8621	1.23370	24.67401	x4
1	0.8812	0.8753	33.8244	1.28225	25.64506	x2
2	0.9584	0.9540	2.1456	0.47259	8.97925	x4 x5
2	0.9575	0.9530	2.5565	0.48329	9.18259	x1 x5
2	0.9486	0.9432	6.4172	0.58384	11.09305	x2 x5
2	0.9477	0.9422	6.8317	0.59464	11.29813	x3 x5
2	0.9386	0.9322	10.7846	0.69759	13.25422	x1 x3
2	0.9288	0.9213	15.0819	0.80951	15.38074	x3 x4
2	0.9229	0.9148	17.6524	0.87646	16.65275	x1 x2
2	0.9226	0.9145	17.7619	0.87931	16.70693	x1 x4
2	0.9167	0.9079	20.3553	0.94686	17.99026	x2 x3
2	0.8960	0.8850	29.3891	1.18213	22.46056	x2 x4
3	0.9616	0.9553	2.7354	0.46008	8.28144	x2 x4 x5
3	0.9592	0.9524	3.8138	0.48973	8.81505	x1 x4 x5
3	0.9585	0.9516	4.0951	0.49746	8.95427	x3 x4 x5
3	0.9584	0.9515	4.1525	0.49904	8.98266	x1 x2 x5
3	0.9575	0.9504	4.5563	0.51014	9.18251	x1 x3 x5
3	0.9488	0.9403	8.3265	0.61379	11.04816	x2 x3 x5
3	0.9420	0.9323	11.3275	0.69629	12.53321	x1 x2 x3
3	0.9386	0.9284	12.7821	0.73628	13.25299	x1 x3 x4
3	0.9314	0.9200	15.9325	0.82289	14.81195	x2 x3 x4
3	0.9229	0.9101	19.6254	0.92441	16.63935	x1 x2 x4
4	0.9630	0.9543	4.1582	0.47034	7.99580	x1 x2 x4 x5
4	0.9619	0.9530	4.6037	0.48331	8.21625	x2 x3 x4 x5
4	0.9593	0.9497	5.7556	0.51684	8.78628	x1 x3 x4 x5
4	0.9588	0.9491	5.9859	0.52354	8.90021	x1 x2 x3 x5
4	0.9432	0.9298	12.7951	0.72175	12.26972	x1 x2 x3 x4
5	0.9633	0.9519	6.0000	0.49485	7.91752	x1 x2 x3 x4 x5

The REG Procedure

Stepwise Selection: Step 1
Variable x5 Entered: R-Square = 0.9455 and C(p) = 5.7794

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	204.15762	204.15762	347.00	<.0001
Error	20	11.76713	0.58836		
Corrected Total	21	215.92475			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-5.48434	0.54474	59.63564	101.36	<.0001
x5	20.85657	1.11964	204.15762	347.00	<.0001

Stepwise Selection: Step 2

Variable x4 Entered: R-Square = 0.9584 and C(p) = 2.1456

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	206.94550	103.47275	218.95	<.0001
Error	19	8.97925	0.47259		
Corrected Total	21	215.92475			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-6.33511	0.60088	52.53154	111.16	<.0001
x4	4.15425	1.71041	2.78787	5.90	0.0252
x5	15.01600	2.60568	15.69476	33.21	<.0001

All variables left in the model are significant at the 0.1500 level.
No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x5		1	0.9455	0.9455	5.7794	347.00	<.0001
2	x4		2	0.0129	0.9584	2.1456	5.90	0.0252

The REG Procedure

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.9633 and C(p) = 6.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	208.00723	41.60145	84.07	<.0001
Error	16	7.91752	0.49485		
Corrected Total	21	215.92475			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-6.51221	0.93356	24.07918	48.66	<.0001
x1	1.99941	2.57334	0.29873	0.60	0.4485
x2	-3.67510	2.77366	0.86876	1.76	0.2038
x3	2.52449	6.34750	0.07827	0.16	0.6961
x4	5.15808	3.66028	0.98269	1.99	0.1779
x5	14.40116	4.85599	4.35219	8.80	0.0091

Backward Elimination: Step 1

Variable x3 Removed: R-Square = 0.9630 and C(p) = 4.1582

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	207.92895	51.98224	110.52	<.0001
Error	17	7.99580	0.47034		

Corrected Total	21	215.92475

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-6.34844	0.81683	28.41077	60.40	<.0001
x1	1.95314	2.50625	0.28565	0.61	0.4465
x2	-3.16590	2.39880	0.81926	1.74	0.2044
x4	5.16889	3.56841	0.98687	2.10	0.1657
x5	15.63277	3.64667	8.64356	18.38	0.0005

Backward Elimination: Step 2

Variable x1 Removed: R-Square = 0.9616 and C(p) = 2.7354

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	207.64331	69.21444	150.44	<.0001
Error	18	8.28144	0.46008		
Corrected Total	21	215.92475			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-5.98038	0.65914	37.87384	82.32	<.0001
x2	-2.88979	2.34647	0.69781	1.52	0.2340
x4	6.87926	2.78280	2.81161	6.11	0.0236
x5	17.10900	3.08189	14.17911	30.82	<.0001

Backward Elimination: Step 3

Variable x2 Removed: R-Square = 0.9584 and C(p) = 2.1456

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	206.94550	103.47275	218.95	<.0001
Error	19	8.97925	0.47259		
Corrected Total	21	215.92475			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-6.33511	0.60088	52.53154	111.16	<.0001
x4	4.15425	1.71041	2.78787	5.90	0.0252
x5	15.01600	2.60568	15.69476	33.21	<.0001

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x3	4	0.0004	0.9630	4.1582	0.16	0.6961
2	x1	3	0.0013	0.9616	2.7354	0.61	0.4465
3	x2	2	0.0032	0.9584	2.1456	1.52	0.2340

Dummy variables

STAT 615 (Baron)

```
DATA C;
INFILE pulse;
INPUT row pulse1 pulse2 ran smokes gender height weight activity;
if activity=0 then delete;
if activity=1 then z1=1; else z1=0;          /* Defining  dummy variables manually */
if activity=2 then z2=1; else z2=0;
```

(2 dummy variables for 3 levels of activity)

same result

```
PROC REG;
MODEL pulse1=height weight z1 z2;

PROC GLM;
CLASS activity;                              /* Automatic definition of dummy variables */
MODEL pulse1 = height weight activity;
```

Treating activity as a categorical variable

```
PROC GLM;                                    /* A model with interactions */
CLASS smokes gender activity;
MODEL pulse1 = weight activity weight*smokes smokes|gender|height;

run;
```

weight*smokes: interaction (different slopes for smokers and non-smokers)

smokes|gender|height: including all interactions

The REG Procedure

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	996.36360	249.09090	2.28	0.0675
Error	86	9406.77926	109.38115		
Corrected Total	90	10403			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	109.52800	26.36944	4.15	<.0001
height	1	-0.45978	0.48930	-0.94	0.3500
weight	1	-0.04053	0.07495	-0.54	0.5901
z1	1	7.84646	4.18090	1.88	0.0639
z2	1	0.47400	2.67773	0.18	0.8599

The GLM Procedure

Multicollinearity

STAT 615 (Baron)

```
DATA C;
INFILE forest;
INPUT AGE HD N MDBH;

PROC REG DATA=C;
MODEL MDBH = HD AGE N / COLLIN TOL VIF;

RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: MDBH

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8.60560	2.86853	25.79	<.0001
Error	16	1.77990	0.11124		
Corrected Total	19	10.38550			

Root MSE	0.33353	R-Square	0.8286
Dependent Mean	6.26500	Adj R-Sq	0.7965
Coeff Var	5.32374		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	5.07300	0.52784	9.61	<.0001	.	0
HD	1	0.09968	0.01873	5.32	<.0001	0.18638	5.36536
AGE	1	-0.09741	0.05244	-1.86	0.0817	0.19665	5.08522
N	1	-0.00290	0.00062451	-4.64	0.0003	0.80108	1.24831

Tolerance $= 1 - R_j^2$

Variance Inflation $\to VIF_j = \dfrac{1}{1-R_j^2}$

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	HD	AGE	N
1	3.93496	1.00000	0.00127	0.00044404	0.00049800	0.00191
2	0.04558	9.29148	0.01797	0.02137	0.05106	0.34816
3	0.01575	15.80776	0.87202	0.04876	0.00145	0.38388
4	0.00372	32.53216	0.10874	0.92942	0.94700	0.26604

Eigenvalue: Eigenvalues of $(X'X)$, ordered

Condition Index: $\sqrt{\dfrac{\lambda_{max}}{\lambda_j}}$

Condition number: $\varkappa = \sqrt{\dfrac{\lambda_{max}}{\lambda_{min}}}$

Proportion of Variation: $p_{je} = \dfrac{A_{je}^2/\lambda_e}{VIF_j}$

where

A = matrix that orthogonalizes $X'X$

A_{je} = principal components

Scaled $X'X$ is used: all diagonals = 1

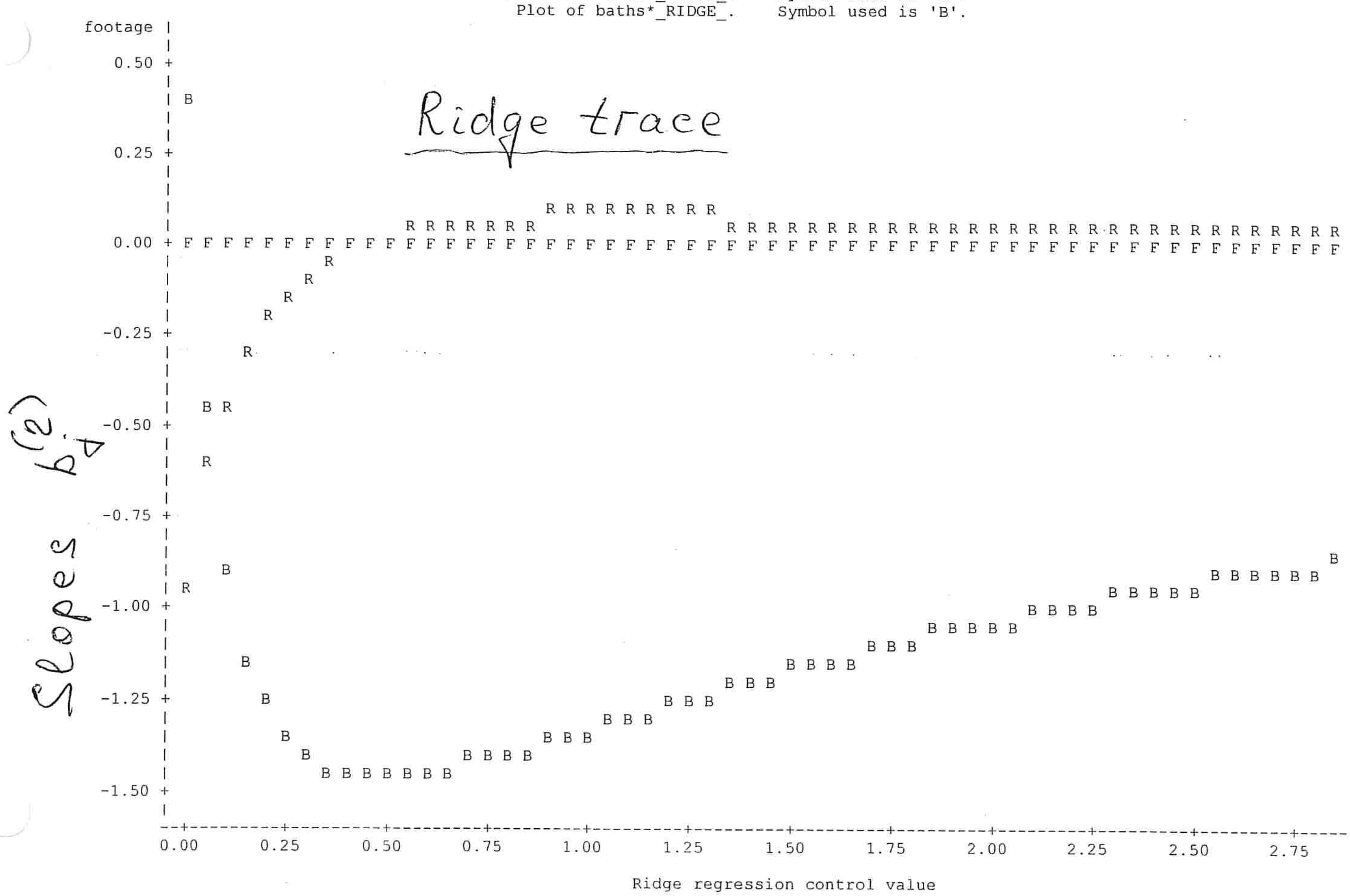
Ridge regression
Treating multicollinearity
STAT 615 (Baron)

```
DATA C;
INFILE homes firstobs=3;
INPUT price footage land rooms baths;

PROC REG OUTEST=RIDGE OUTVIF RIDGE=0 to 3.0 by 0.05;
MODEL price = footage land rooms baths;

DATA slopes; SET ridge(where=(_type_='RIDGE'));
DATA vif;    SET ridge(where=(_type_='RIDGEVIF'));

PROC PLOT data=slopes;
PLOT footage*_RIDGE_='F' land *_RIDGE_='L'
     rooms  *_RIDGE_='R' baths*_RIDGE_='B'  / overlay;

PROC PLOT data=vif;
PLOT footage*_RIDGE_='F' land *_RIDGE_='L'
     rooms  *_RIDGE_='R' baths*_RIDGE_='B'  / overlay;

proc print data=slopes;
proc print data=vif;

RUN;
```

type : 'parms' = $\underset{\sim}{b}^{(0)}$
'ridge' = $\underset{\sim}{b}^{(2)}$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	89.18575	20.17285	4.42	<.0001
footage	1	0.00017763	0.00014595	1.22	0.2255
land	1	-0.02337	0.01261	-1.85	0.0659
rooms	1	-0.93825	2.23118	-0.42	0.6747
baths	1	0.37931	4.28200	0.09	0.9295

```
Plot of footage*_RIDGE_.   Symbol used is 'F'.
Plot of land*_RIDGE_.      Symbol used is 'L'.
Plot of rooms*_RIDGE_.     Symbol used is 'R'.
Plot of baths*_RIDGE_.     Symbol used is 'B'.
```



STAT 615 (Baron)

Influential
Data analysis
and
Outliers

```
DATA C;
INFILE forest;
INPUT AGE HD N MDBH;

PROC REG DATA=C LINEPRINTER;
MODEL MDBH = HD AGE N / INFLUENCE R;

PAINT COOKD.>0.15;
PLOT   RSTUDENT.*PREDICTED.   COOKD.*MDBH;

REWEIGHT RSTUDENT.>5 OR RSTUDENT.<-5 / WEIGHT=0.25;
REWEIGHT COOKD. > 0.05 / WEIGHT=0.5;
REWEIGHT COOKD. > 0.10 / WEIGHT=0;

REFIT;  PRINT;      run;
```

The REG Procedure
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.07300	0.52784	9.61	<.0001
HD	1	0.09968	0.01873	5.32	<.0001
AGE	1	-0.09741	0.05244	-1.86	0.0817
N	1	-0.00290	0.00062451	-4.64	0.0003

Output Statistics

				e_i		r_i			$r_{i(-i)}$	h_{ii}		
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	7.0000	6.9058	0.1451	0.0942	0.300	0.314	|	0.006	0.3048	0.1892	1.5576	0.1472
2	5.0000	5.2162	0.1763	-0.2162	0.283	-0.764	*|	0.057	-0.7531	0.2795	1.5488	0.4691
3	6.2000	5.7748	0.1391	0.4252	0.303	1.403	|**	0.104	1.4501	0.1739	0.9272	0.6653
4	5.2000	5.6244	0.1716	-0.4244	0.286	-1.484	**|	0.198	-1.5473	0.2647	0.9737	0.9283
5	6.2000	6.1862	0.1312	0.0138	0.307	0.0449	|	0.000	0.0435	0.1547	1.5308	0.0186
6	5.2000	5.1056	0.1740	0.0944	0.285	0.332	|	0.010	0.3223	0.2723	1.7304	0.1971
7	6.2000	6.3933	0.0980	-0.1933	0.319	-0.606	*|	0.009	-0.5939	0.0864	1.2911	0.1826
8	6.4000	6.1704	0.1134	0.2296	0.314	0.732	|*	0.018	0.7211	0.1157	1.2773	0.2608
9	6.4000	6.5881	0.1892	-0.1881	0.275	-0.685	*|	0.056	-0.6731	0.3219	1.6949	0.4638
10	6.4000	6.9782	0.1309	-0.5782	0.307	-1.885	***|	0.162	-2.0690	0.1540	0.5605	0.8827
11	5.4000	5.2048	0.1495	0.1952	0.298	0.655	|*	0.027	0.6425	0.2010	1.4536	0.3223
12	6.4000	6.5451	0.1773	-0.1451	0.283	-0.514	*|	0.026	-0.5014	0.2825	1.6881	0.3146
13	5.4000	5.7041	0.1542	-0.3041	0.296	-1.028	**|	0.072	-1.0304	0.2138	1.2525	0.5374
14	6.7000	6.8622	0.1083	-0.1622	0.315	-0.514	*|	0.008	-0.5022	0.1055	1.3539	0.1725
15	6.7000	6.8822	0.1080	-0.1822	0.316	-0.577	*|	0.010	-0.5649	0.1049	1.3294	0.1934
16	5.9000	5.6869	0.1515	0.2131	0.297	0.717	|*	0.033	0.7057	0.2062	1.4311	0.3597
17	6.9000	6.9995	0.1419	-0.0995	0.302	-0.330	|	0.006	-0.3203	0.1811	1.5383	0.1506
18	6.9000	6.4114	0.0892	0.4886	0.321	1.520	|***	0.044	1.5914	0.0715	0.7470	0.4415
19	6.9000	6.7711	0.1704	0.1289	0.287	0.449	|	0.018	0.4379	0.2609	1.6648	0.2602
20	7.9000	7.2895	0.2002	0.6105	0.267	2.288	|****	0.737	2.7016	0.3603	0.4144	2.0276

Output Statistics

	DFBETAS			
Obs	Intercept	HD	AGE	N
1	0.0247	-0.0188	0.0516	-0.0798
2	0.1572	0.2287	-0.1545	-0.4077
3	0.4684	0.1840	-0.3988	-0.1259
4	-0.6521	0.5003	-0.2991	0.2904
5	0.0164	0.0034	-0.0070	-0.0121
6	0.0933	-0.1434	0.0850	0.0305
7	0.0734	0.0612	-0.0979	-0.0525
8	0.0204	0.1288	-0.1577	0.0458
9	0.3857	-0.0187	-0.0881	-0.3052
10	0.0470	-0.6952	0.5366	0.2503
11	-0.0088	-0.1855	0.1040	0.2232
12	0.1313	0.1841	-0.2613	-0.0599
13	-0.4901	0.0237	0.1475	0.2306
14	0.0675	-0.0338	-0.0230	0.0100
15	-0.0305	-0.1188	0.0830	0.0998
16	-0.1649	-0.0623	0.0353	0.2999
17	-0.0399	-0.1243	0.1084	0.0828
18	-0.0643	0.1779	-0.1478	0.0652
19	0.0744	0.2254	-0.2296	-0.0931
20	-0.1745	-0.3834	1.0042	-0.8208

STAT 615 (Baron)

NONLINEAR REGRESSION

```
DATA C;
INFILE squid;
INPUT x1 x2 x3 x4 x5 y;

proc NLIN method=GAUSS;
model y=a*x1+exp(b*x2)+c;
parameters a=1, b=1, c=0;

der.a = x1;
der.b = exp(b*x2)*x2;
der.c = 1;

run;
```

} Partial derivatives

```
                 Non-Linear Least Squares Iterative Phase
                Dependent Variable Y   Method: Gauss-Newton
     Iter             A              B                C  Sum of Squares
        0      1.000000       1.000000                0       89.993753
        1      6.310562       0.877580        -8.184051       16.163552
        2      6.397452       0.845963        -8.228067       16.067504
        3      6.418003       0.841275        -8.242082       16.067027
        4      6.421016       0.840711        -8.244281       16.067024
        5      6.421379       0.840645        -8.244547       16.067024
NOTE: Convergence criterion met.


   Non-Linear Least Squares Summary Statistics     Dependent Variable Y

     Source                  DF  Sum of Squares      Mean Square

     Regression               3    587.01427612    195.67142537
     Residual                19     16.06702388      0.84563284
     Uncorrected Total       22    603.08130000

     (Corrected Total)       21    215.92475000


     Parameter      Estimate     Asymptotic            Asymptotic 95 %
                                 Std. Error         Confidence Interval
                                                    Lower         Upper
            A    6.421378626  1.9352691460    2.370842695  10.471914557
            B    0.840644982  0.3336459992    0.142320870   1.538969094
            C   -8.244547413  1.6081107793  -11.610337906  -4.878756920


                       Asymptotic Correlation Matrix

     Corr                    A                  B                  C
     --------------------------------------------------------------
            A                1       -0.956849766       -0.961878292
            B     -0.956849766                  1       0.8492386362
            C     -0.961878292       0.8492386362                  1
```

```
PROC NLIN;
model systol=a+exp(b*weight);
parms a=0, b=1;   title 'UNLUCKY INITIAL PARAMETERS';
der.a = x1; der.b = exp(b*x2)*x2; der.c = 1;

PROC NLIN;
model systol=a+exp(b*weight);
parms a=0, b=0;   title 'LUCKY INITIAL PARAMETERS';
der.a = x1; der.b = exp(b*x2)*x2; der.c = 1;
```

Results may depend on initial values

```
                 UNLUCKY INITIAL PARAMETERS
      DUD                 A               B  Sum of Squares
       -3                 0        1.000000  3.69181433E75
       -2          0.100000        1.000000  3.69181433E75
       -1                 0        1.100000  1.33034364E83
        0                 0        1.000000  3.69181433E75
        1                 0        0.999983  3.68112698E75
                    ...........................

       50                 0        0.686092  7.01623013E51
       51                 0        0.679778  2.33874338E51
WARNING: PROC NLIN failed to converge.


                  LUCKY INITIAL PARAMETERS
      DUD                 A               B  Sum of Squares
       -3                 0               0         629734
       -2          0.100000               0         628748
       -1                 0        0.100000       48891662
        0          0.100000               0         628748
        1        121.649234        0.000628    7398.070407
                       ......................
       15        110.052589        0.044337    4812.211786
       16        110.074308        0.044319    4812.207328
       17        110.074121        0.044320    4812.207328
NOTE: Convergence criterion met.
```

```
proc NLIN method=GAUSS;
model y=a*x1+exp(b*x2)+c;
parameters a=0 to 10, b=0.5 to 1 by 0.1, c=0,3,-5;
der.a = x1;  der.b = exp(b*x2)*x2;  der.c = 1; run;
```

Solution: Grid method

```
      Non-Linear Least Squares Grid Search     Dependent Variable Y
                         A               B               C  Sum of Squares
                         0        0.500000               0      293.741222
                  1.000000        0.500000               0      154.413829
                  2.000000        0.500000               0      115.396435
                            ......................
                  1.000000        0.700000               0      116.979572
                  2.000000        0.700000               0      119.137212

               Dependent Variable Y    Method: Gauss-Newton
     Iter                A               B               C  Sum of Squares
        0         4.000000        1.000000       -5.000000       25.208158
        1         6.310562        0.877580       -8.184051       16.163552
        2         6.397452        0.845963       -8.228067       16.067504
        3         6.418003        0.841275       -8.242082       16.067027
        4         6.421016        0.840711       -8.244281       16.067024
        5         6.421379        0.840645       -8.244547       16.067024
NOTE: Convergence criterion met.
```

Mann-Whitney-Wilcoxon test

```
PROC NPAR1WAY DATA=C WILCOXON;
CLASS GENDER;
VAR PULSE1; run;
```

$X_1 \ldots X_n \sim F(x - \theta_x)$

$y_1 \ldots y_m \sim F(y - \theta_y)$

Test $H_0: \theta_x = \theta_y$

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable PULSE1
Classified by Variable GENDER

GENDER	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	57	2346.50	2650.50	124.079967	41.166667
2	35	1931.50	1627.50	124.079967	55.185714

Average scores were used for ties.

sum of X-ranks among $X_1 .. X_n, y_1 .. y_m$

$m \cdot$ average rank $= \dfrac{m(m+n+1)}{2}$

Wilcoxon Two-Sample Test

Statistic	1931.5000
Normal Approximation	
Z	2.4460
One-Sided Pr > Z	0.0072
Two-Sided Pr > |Z|	0.0144
t Approximation	
One-Sided Pr > Z	0.0082
Two-Sided Pr > |Z|	0.0164

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test

Chi-Square	6.0027
DF	1
Pr > Chi-Square	0.0143

$$K = (m+n-1)\frac{n(\bar{r}_{x} - \bar{r})^2 + m(\bar{r}_{y\cdot} - \bar{r})^2}{\sum_i (r_{ix} - \bar{r})^2 + \sum_j (r_{jy} - \bar{r})^2}$$

STAT 615 (Baron)

Logistic Regression

```
DATA C; INFILE pulse;
INPUT row pulse1 pulse2 ran smokes sex height weight activity;

PROC LOGISTIC;
CLASS sex activity;
MODEL smokes = height weight activity|sex / INFLUENCE COVB SELECTION=BACKWARD;
TEST height=weight;

RUN;
```

Categorical predictors and interactions, similar to Proc GLM.

Various options, similar to Proc REG

```
                         The LOGISTIC Procedure
                           Model Information

             Data Set                        WORK.C
             Response Variable               smokes
             Number of Response Levels       2
             Number of Observations          92
             Link Function                   Logit
             Optimization Technique          Fisher's scoring


                            Response Profile

                   Ordered                        Total
                     Value        smokes      Frequency

                         1             1             28
                         2             2             64


                        Class Level Information

                                        Design Variables

                Class        Value        1        2        3

                sex          1            1
                             2           -1

                activity     0            1        0        0
                             1            0        1        0
                             2            0        0        1
                             3           -1       -1       -1
```

analogue of indicator (dummy) variables

```
Step  0. The following effects were entered: Intercept  height  weight  activity  sex  sex*activity
Step  1. Effect sex*activity is removed:
Step  2. Effect activity is removed:
Step  3. Effect sex is removed:
Step  4. Effect height is removed:
Step  5. Effect weight is removed:
NOTE: All effects have been removed from the model.

                        Summary of Backward Elimination

                  Effect                    Number          Wald
         Step     Removed          DF           In    Chi-Square    Pr > ChiSq

            1     sex*activity      2            4        0.0144        0.9928
            2     activity          3            3        0.3616        0.9481
            3     sex               1            2        0.2546        0.6139
            4     height            1            1        2.3251        0.1273
            5     weight            1            0        3.5211        0.0606


                   Analysis of Maximum Likelihood Estimates

                                           Standard
      Parameter           DF    Estimate      Error    Chi-Square    Pr > ChiSq
      Intercept            1     -0.8267     0.2266       13.3114        0.0003


                        Estimated Covariance Matrix

                         Variable       Intercept
                         Intercept       0.051339

WARNING: The validity of the model fit is questionable.
```

a model

$$\ln \frac{\pi_i}{1-\pi_i} = \alpha + \sum \beta_j X_{ji}$$

is reduced to

$$\ln \frac{\pi_i}{1-\pi_i} = \alpha$$

$$\Downarrow$$

$$\ln \frac{\pi_i}{1-\pi_i} = -.8267$$

$$\Downarrow$$

$$y_i = \frac{e^{-.8267}}{1+e^{-.8267}} + \varepsilon_i$$

```
                               Regression Diagnostics

              Pearson Residual                  Deviance Residual                 Hat Matrix Diagonal

   Case                  (1 unit = 0.19)                   (1 unit = 0.19)                   (1 unit = 7.E-04)
 Number      Value      -8  -4  0 2 4 6 8      Value      -8  -4  0 2 4 6 8      Value      0 2 4 6 8  12  16

      1    -0.6614     |    *    |        |   -0.8519    |    *    |        |    0.0109    |                *|
      2    -0.6614     |    *    |        |   -0.8519    |    *    |        |    0.0109    |                *|
      3     1.5119     |         |       *|    1.5425    |         |      * |    0.0109    |                *|
      4     1.5119     |         |       *|    1.5425    |         |      * |    0.0109    |                *|
      5    -0.6614     |    *    |        |   -0.8519    |    *    |        |    0.0109    |                *|
      6    -0.6614     |    *    |        |   -0.8519    |    *    |        |    0.0109    |                *|
      7    -0.6614     |    *    |        |   -0.8519    |    *    |        |    0.0109    |                *|
etc.
```

$\underbrace{\hspace{6em}}_{\text{Outliers and influential diagnostic}}$