PREFACE

Most calculations were done using Microsoft Excel XP. For you convenience, all worksheets can be downloaded via the internet: http://www.stephan-brumme.com/studies/statistik.html

PROBLEM 1

? *A study in surgery examined the increase (Y) in pancreatic intraductal pressure (PIP) in response to doses of a potent cholinesterase inhibitor (x). Six different doses were administrated to one dog (see below for results).*

These are the measured values:

x	Y
0	14.6
5	24.5
10	21.8
15	34.5
20	35.1
25	43.0

Table 1 *Recorded PIP increase*

First of all, you should enable Excel's powerful *Data Analysis* add-in. To do so, go to the add-in manager (menu *Tools*) and check the appropriate box. Now you can access the add-in in the *Tools* menu, too. Unfortunately, I do not have access to an English version of Excel and thus some problems arise in giving you the correct English terms. An internet page provided this translation table, I hope it is accurate (see http://www.unifr.ch/stat/alt/Unterlagen/Stat-II/Regression/RegressionUsingExcel.pdf):

English	German / Deutsch
Regression Statistics	Regressions-Statistik
Multiple R	Multipler Korrelationskoeffizient
R Square	Bestimmtheitsmaß
Adjusted R Square	Adjustiertes Bestimmtheitsmaß
Standard Error	Standardfehler
Observations	Beobachtungen
ANOVA	Analyse der Varianzen
Degrees of Freedom (df)	Freiheitsgrade
Summed Squares (SS)	Quadratsummen
Mean Squares (MS)	Mittlere Quadratsummen
F	Prüfgröße
Significance F	F krit.
Regression	Regression
Residual	Residue
Total	Gesamt
Coefficients	Koeffizienten
Standard Error	Standardfehler
t Stat	t-Statistik
P-value	P-Wert
Intercept	Schnittpunkt

Table 2 *English/German Statistics Terms*

When I ran the add-in for the first time, I was deeply impressed by the enormous size of the resulting worksheet. However, I did not grasp the meaning of each and every value and often had to take a look at Excel's manual. These are the formulas behind the most important estimators:

estimator	formula
mean	$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$
standard deviation	$s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$
covariance	$s_{xy}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot (y_i - \overline{y})^2$

Table 3 *Basic estimators*

Excel's regression statistics are based on these formulas (where m denotes the number of influencing factors and b stands for the coefficient of regression):

attribute	formula
multiple R	$r_{xy} = \frac{s_{xy}}{\sqrt{s_x s_y}}$
R square	$\sqrt{r_{xy}}$
adjusted R square	$B_{adj} = 1 - \left(\frac{\sum_{i=1}^{n} (y_i - \overline{y})^2 - b^2 \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}{n-m-1} \right) \cdot \left(\frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2 \right)^{-1}$
standard error	$s_{x,y} = \sqrt{\frac{1}{n \cdot (n-2)} \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 - \left(\frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \right) \right)}$
observations	n

Table 4 *Regression statistics*

The ANOVA (ANalysis Of VAriances) table relies on these formulas:

	df	SS	MS	F	signif. F
regression	1	$q_1 = (n-1) \cdot \frac{s_{xy}^2}{s_x^2}$	$w_1 = q_1$	$v_0 = \frac{w_1}{w_2}$	$F_{1,n-2}(v_0)$
residual	$n-1$	$q_2 = q - q_1$	$w_2 = \frac{q_2}{n-2}$		
total	n	$q = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_i \right)^2$			

Table 5 *ANOVA*

The most important table printed by the Data Analysis gives several values of the intercept and X:

	coefficient	**standard error**	**t stat**	**P-value**
intercept	$k = \bar{y} - b\bar{x}$	$s_{k_{yx}}$	$t_{0,s}$	$t_{n-2}\left(t_{0,s}\right)$
X variable 1	$b = \frac{s_{xy}}{s_x^2}$	$s_{b_{yx}}$	$t_{0,K}$	$t_{n-2}\left(t_{0,K}\right)$

Table 6 *Coefficients, confidence intervals etc., part I*

	lower $\tilde{\gamma}$ %	**upper $\tilde{\gamma}$ %**	**lower γ %**	**upper γ %**
intercept	$k - l_{k,\tilde{\gamma}\%}$	$k + l_{k,\tilde{\gamma}\%}$	$k - l_{k,\gamma\%}$	$k + l_{k,\gamma\%}$
X variable 1	$b - l_{b,\tilde{\gamma}\%}$	$b + l_{b,\tilde{\gamma}\%}$	$b - l_{b,\gamma\%}$	$b + l_{b,\gamma\%}$

Table 7 *Coefficients, confidence intervals etc., part II*

The last two tables contain some variables not mentioned so far. $s_{k_{yx}}$ and $s_{b_{yx}}$ estimate the standard error of k and b, respectively.

$$s_{k_{yx}} = s_{y,x} \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}^2\right)}}$$

$$s_{b_{yx}} = \frac{s_{y,x}}{\sum_{i=1}^{n}\left(x_i - \bar{x}^2\right)}$$

$t_{0,K}$ and $t_{0,s}$ give the probability of the t-distribution under the assumption or hypothesis "is equal to zero".

$$t_{0,K} = s_x \cdot \sqrt{(n-1)\cdot(n-2)} \cdot \frac{b}{\sqrt{a}}$$

$$t_{0,s} = \sqrt{n-2} \cdot \frac{-\bar{x}\cdot b + \bar{y}}{h \cdot \sqrt{a}}$$

Last but not least, the confidence intervals:

$$l_{b,\gamma} = \frac{t_{n-2}^{-1} \cdot \frac{1+\gamma}{2} \cdot \sqrt{a}}{s_x} \cdot \sqrt{(n-1)\cdot(n-2)}$$

$$t_{k,\gamma} = \frac{t_{n-2}^{-1} \cdot \frac{1+\gamma}{2} \cdot h \cdot \sqrt{a}}{\sqrt{n-2}}$$

The formulas are almost identical for γ and $\tilde{\gamma}$, just insert the desired one.

As mentioned earlier, I use the German edition of Excel. It computes for the given six observations (not translated, see Table 2, the decimal point is a comma in Germany):

Regressions-Statistik	
Multipler Korrelationskoeffizient	0,9567
Bestimmtheitsmaß	0,9153
Adjustiertes Bestimmtheitsmaß	0,8941
Standardfehler	3,3904
Beobachtungen	6

Table 8 *Regression statistics*

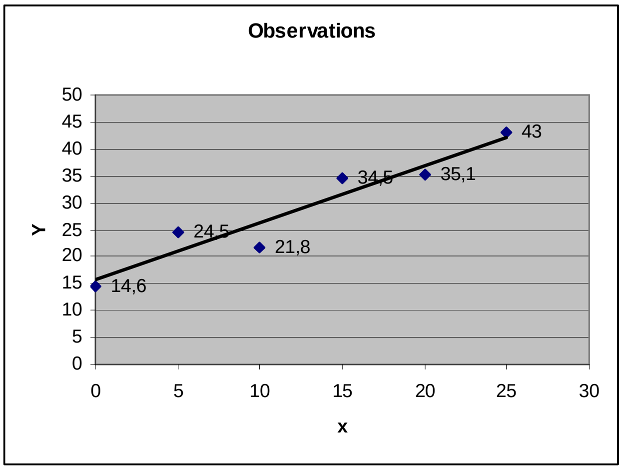

Figure 1 *Observations*

All observations seem to be pretty close to a linear regression (Figure 1). However, a residual plot could make you believe in the opposite:

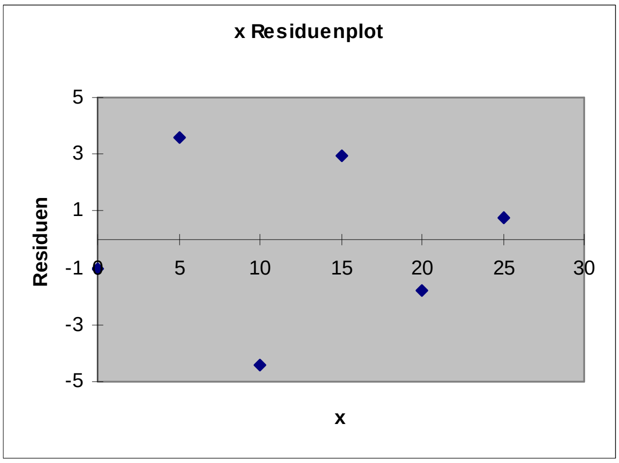

Figure 2 *Residual plot*

Since it seems to be impossible to get a distinct decision based on the visual impression, I have to stick to the numerical facts (all numbers rounded to four digits):

	Freiheitsgrade (df)	Quadrat-summen (SS)	Mittlere Quad-rat-summe (MS)	Prüfgröße (F)	F krit
Regression	1	496,8893	496,8893	43,2275	0,0028
Residue	4	45,9790	11,4948		
Gesamt	5	542,8683			

Table 9 *ANOVA*

The linear regression of Figure 1 follows the equation

$$y = f(x) = 1.0657 \cdot x + 15.5952$$

You will find the numbers in the first column of Table 10:

	Koeffizienten	Standardfehler	t-Statistik	P-Wert
Schnittpunkt	15,5952	2,4538	6,3556	0,0031
x	1,0657	0,1621	6,5748	0,0028

Table 10 *Coefficients, confidence intervals etc., part I*

	Untere 95%	Obere 95%	Untere 95,0%	Obere 95,0%
Schnittpunkt	8,7824	22,4081	8,7824	22,4081
x	0,6157	1,5158	0,6157	1,5158

Table 11 *Coefficients, confidence intervals etc., part II*

These were quite a lot tables and figures generated from only six observations. So what do I conclude ? First of all, Excel did most of the work with a few clicks, it happened much faster and easier than the export to Word. Then, Excel did more than I wanted – so it took me some time to actually find out whether the linear regression can be accepted or must be rejected. The last three tables (Table 9, Table 10 and Table 11) contain that information three times: F krit is below 0.05, i.e. 5%, the P-value is below 0.05, and the confidence intervals contain both the intercept (Schnittpunkt) and x. Therefore, I suppose there is a linear correlation between the pancreatic intraductal pressure (PIP) and doses of a potent cholinesterase inhibitor.

Poor dog that had to suffer. God bless you.

PROBLEM 2

?

Dairy scientists have recently carried out several studies on protein biosynthesis milk and the accompanying decomposition of nucleic acids into various constituents. The paper "Metabolites of Nucleic Acids in Bovine Milk" (J. of Dairy Science (1984):723-728) reported the accompanying data on milk production (x, kg/day) and milk protein (y, kg/day) for Holstein-Friesian cows.

x	y
42.7	1.20
40.2	1.16
38.2	1.07
37.6	1.13
32.2	0.96
32.2	1.07
28.0	0.85
27.2	0.87
26.6	0.77
23.0	0.74
22.7	0.76
27.8	0.69
21.3	0.72
20.2	0.64

Table 12 *Milk production (x) vs. milk protein (y) in kg/day*

The observations are well distributed insofar as about half of them are above and half of them are below the estimated linear regression:

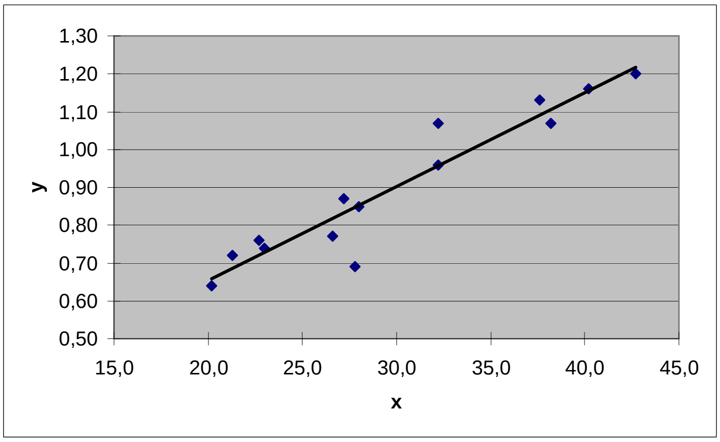

Figure 3 *Linear regression*

Note that the figure's axes were shifted and do not start with zero. The formula of the linear regression:

$$y = f(x) = 0.0249 \cdot x + 0.1567$$

The way I go to examine the data does not differ in any aspect from problem 1, hence I present you a slightly shortened version of the results but this time I translated all terms to English in order to enhance readability:

	df	SS	MS	F	Significance F
regression	1	0.4338	0.4338	109.2981	0.0000
residual	12	0.0476	0.0040		
total	13	0.4814			

Table 13 *ANOVA*

	coefficients	standard error	t-stat	P-value
intercept	0.1567	0.0733	2.1388	0.0537
x	0.0249	0.0024	10.4546	0.0000

Table 14 *Coefficients, confidence intervals etc., part I*

	lower 95%	upper 95%	lower 95.0%	upper 95.0%
intercept	-0.0029	0.3163	-0.0029	0.3163
x	0.0197	0.0300	0.0197	0.0300

Table 15 *Coefficients, confidence intervals etc., part II*

Because of *Significance F* being very low (≈ 0.00000022), I infer that the daily milk protein production of a Holstein-Friesian cow linearly depends on its daily milk production. Without the outlier of Table 12 (marked red) the relationship would be even stronger.

PROBLEM 3

? *A study examines the influence of manganese (Mn) on the growth of wheat. The observations consist of the height in cm (y) and the amount of added manganese (x). It may be necessary to apply a suitable transformation to x.*

x	ln x	y
0.368	-1.00	32
0.670	-0.40	37
1.000	0.00	44
1.221	0.20	45
2.718	1.00	46
7.389	2.00	42
16.445	2.80	42
24.533	3.20	40
29.964	3.40	37
54.598	4.00	30

Table 16 *Milk production (x) vs. milk protein (y) in kg/day*

Mrs. Liero gave a hint to use the natural logarithm of x. The following diagrams show that she was right, indeed:

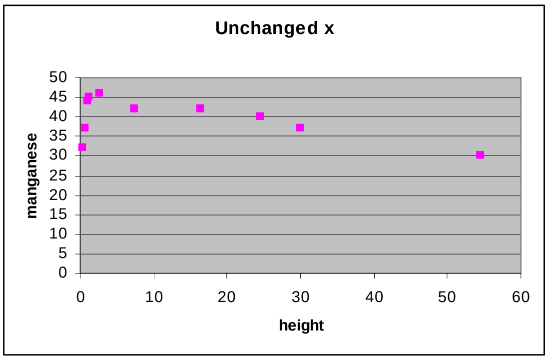

Figure 4 *Initial setting*

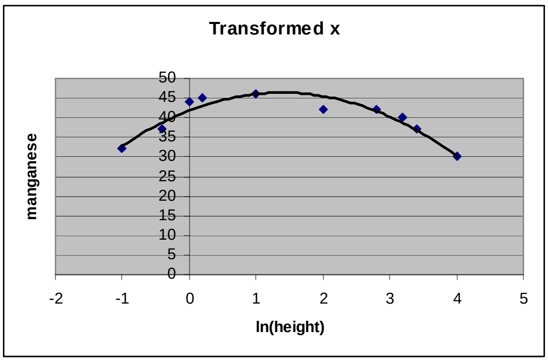

Figure 5 *Applying ln(x)*

Furthermore, she told us to *not* always try to find a linear relationship. This problem seems to contain a polynomial one with a degree of two. I added it to Figure 5:

$$y = f(x) = -2.3624 \cdot x^2 + 6.5817 \cdot x + 41.7427$$

Due to some reasons I do not know, Excel cannot directly compute the formula above. While it is able to add a polynomial trend line, it offers no obvious way to display the coefficients used. Smart students know how to use the internet – I am not smart but awfully lazy and found the according tips and tricks within a few seconds: if you create a new column holding the squares of ln(x) and include these cells in the area of x then Excel's regression analysis tells you the formula like it did for linear regression in former times.

input x		response
ln x	(ln x)^2	y
-1.00	1.00	32
-0.40	0.16	37
0.00	0.00	44
0.20	0.04	45
1.00	1.00	46
2.00	4.00	42
2.80	7.84	42
3.20	10.24	40
3.40	11.56	37
4.00	16.00	30

Table 17 *Excel's modified input*

And now the usual tables, of course they have an additional row since I found a polynomial relationship:

	df	SS	MS	F	Significance F
regression	2	237.5199	118.7600	30.8124	0.0003
residual	7	26.9801	3.8543		
total	9	264.5000			

Table 18 *ANOVA*

	coefficients	standard error	t-stat	P-value
intercept	41.7427	0.8522	48.9800	0.0000
ln x	6.5817	1.0017	6.5703	0.0003
(ln x)^2	-2.3624	0.3074	-7.6861	0.0001

Table 19 *Coefficients, confidence intervals etc., part I*

	lower 95%	upper 95%	lower 95.0%	upper 95.0%
intercept	39.7275	43.7579	39.7275	43.7579
ln x	4.2130	8.9505	4.2130	8.9505
(ln x)^2	-3.0892	-1.6356	-3.0892	-1.6356

Table 20 *Coefficients, confidence intervals etc., part II*

Again, significance F is small enough (definitely below 0.05) to accept the hypothesis that the plant height depends on the concentration of manganese. The maximum height can be achieved by adding about one unit of manganese (do not ask me what "unit" means in that context – 1g ?).

PROBLEM 4

? *Image quality of monitors is an important characteristic, affecting among other things- extent of eyestrain and work efficiency. The paper "Image Quality Determines Differences in Reading Performance and Perceived Image Quality with CRT and Hard Copy Displays" (Human Factors (1991):459-469) reported on an experiment in which image quality (x) and average time for a group of subjects to read certain passages (y, in seconds) were determined. The accompanying data was read from a graph that appeared in the paper.*

x	y
4.30	8.0
4.55	8.3
5.55	7.8
5.65	7.25
5.95	7.7
6.30	7.5
6.45	7.6
6.45	7.2

Table 21 *Image quality vs. time needed to read certain passages*

Creating a diagram and adding a trend line does not require any extraordinary skills, which are the main reason why I did them first. When I inserted the green linear trend line, I did not feel satisfied because of the big deviations and inserted a yellow polynomial trend line (degree two) as well. They do not differ much so choose the linear one because of its simplicity.

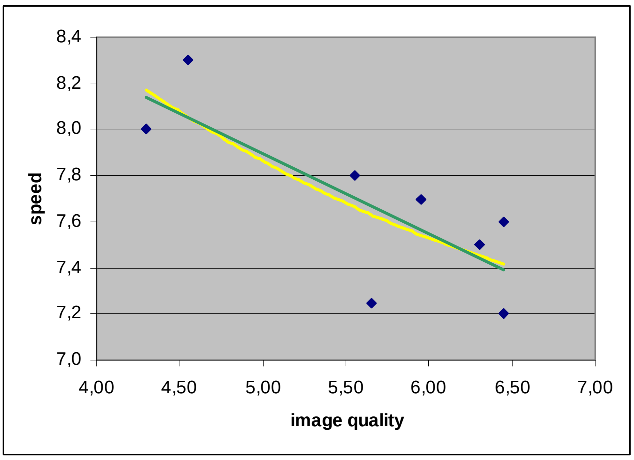

Figure 6 *Observations plotted with two trend lines*

$$y = f(x) = -9.6378 \cdot x - 0.3485$$

	df	SS	MS	F	Significance F
regression	1	0.5872	0.5872	9.5883	0.0212
residual	6	0.3675	0.0612		
total	7	0.9547			

Table 22 *ANOVA*

	coefficients	standard error	t-stat	P-value
intercept	9.6378	0.6419	15.0149	0.0000
x	-0.3485	0.1125	-3.0965	0.0212

Table 23 *Coefficients, confidence intervals etc., part I*

	lower 95%	upper 95%	lower 95.0%	upper 95.0%
intercept	8.0671	11.2084	8.0671	11.2084
x	-0.6239	-0.0731	-0.6239	-0.0731

Table 24 *Coefficients, confidence intervals etc., part II*

Maybe shifting the origin of Figure 6 away from zero was not as good as intended; it falsified the diagram by zooming too close to the observations and loosing the context. I accept the hypothesis at a 0.05 level: the image quality of a CRT monitor does play an important role when reading texts.

PROBLEM 5

? *The accompanying data appeared in the paper "Determination of Biological Maturity and Effect of Harvesting and Drying Conditions on Milling Quality of Paddy" (J. of Ag. Engr. Research (1975):353-361). The dependent variable y is yield (kg/ha) of paddy, a grain farmed in India, and x is the number of days after flowing at which harvesting took place.*

x	y
16	2508
18	2518
20	3304
22	3423
24	3057
26	3190
28	3500
30	3883
32	3823
34	3646
36	3708
38	3333
40	3517
42	3241
44	3103
46	2776

Table 25 *Yield y after x days*

According to the diagram, it may be preferable to choose a polynomial function.

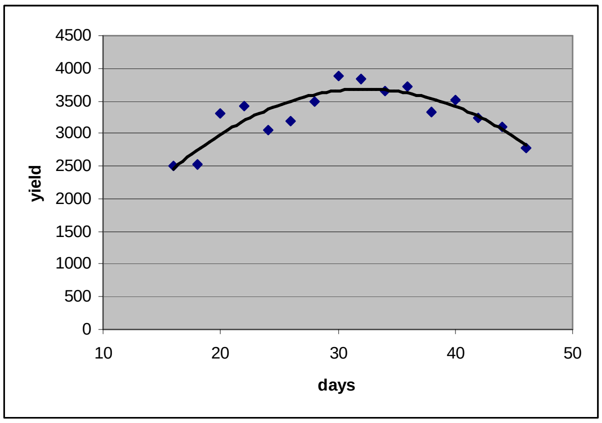

Figure 7 *Paddy growth*

When taking the same algorithm I did in problem 3, Excel gives me these tables that do not surprise me hardly:

	df	SS	MS	F	Significance F
regression	2	2,084,779	1,042,389	25.0765	0.0000
residual	13	540,388	41,568		
total	15	2,625,167			

Table 26 *ANOVA*

	coefficients	standard error	t-stat	P-value
intercept	-1070.3977	617.2527	-1.7341	0.1065
x	293.4829	42.1776	6.9583	0.0000
x^2	-4.5358	0.6744	-6.7255	0.0000

Table 27 *Coefficients, confidence intervals etc., part I*

	lower 95%	upper 95%	lower 95.0%	upper 95.0%
intercept	-2403.8908	263.0954	-2403.8908	263.0954
x	202.3637	384.6022	202.3637	384.6022
x^2	-5.9928	-3.0788	-5.9928	-3.0788

Table 28 *Coefficients, confidence intervals etc., part II*

$$y = f(x) = -4.6358 \cdot x^2 + 293.4829 \cdot x - 1070.3977$$

All observed values spread very well above and below the graph. Therefore, the remarkably low significance F can be trusted.

PROBLEM 6

? *The metabolic rate and the weight of 14 ordinary cows have been tracked for a time. A renowned scientist proposes a correlation (polynomial, degree two) between both values. Is he right ? Estimate all coefficients of the regression function and its error deviation.*

x (weight)	y (metabolic rate)
110	235
110	198
110	173
230	174
230	149
230	124
360	115
360	130
360	102
360	95
505	122
505	112
505	98
505	96

Table 29 *Metabolic rates of cows*

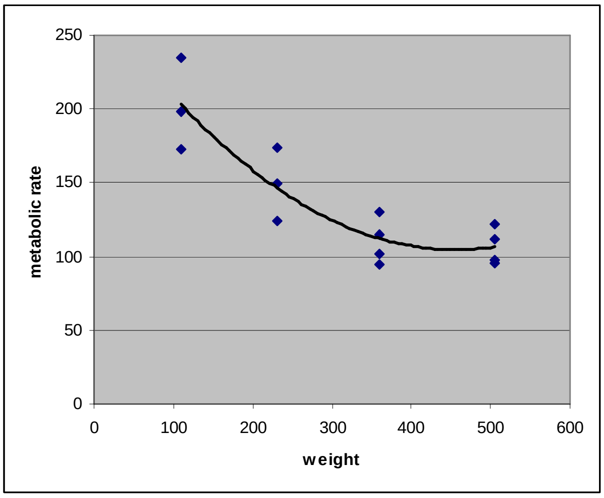

Figure 8 *Metabolism of cows*

The overall visual impression seems to prove the proposal but reveals a lack of data, too. Only four distinct weights were actually recorded – not a lot if you consider the total number of cows on earth ☺

Regressions statistics	
Multiple R	0.9033
R Square	0.8160
Adjusted R Square	0.7825
Standard Error	19.9851
Observations	14

Table 30 *Regression statistics*

	df	SS	MS	F	Significance F
regression	2	19,481	9,740	24.3886	0.0001
residual	11	4,393	399		
total	13	23,875			

Table 31 *ANOVA*

	coefficients	standard error	t-stat	P-value
intercept	275.2588	26.7107	10.3052	0.0000
x	-0.7481	0.1954	-3.8284	0.0028
x^2	0.00082	0.00031	2.6701	0.0218

Table 32 *Coefficients, confidence intervals etc., part I*

	lower 95%	upper 95%	lower 95.0%	upper 95.0%
intercept	216.4690	334.0487	216.4690	334.0487
x	-1.1782	-0.3180	-1.1782	-0.3180
x^2	0.00014	0.00149	0.00014	0.00149

Table 33 *Coefficients, confidence intervals etc., part II*

$$y = f(x) = -0.0008 \cdot x^2 + 0.7481 \cdot x + 275.2588$$

The p-value of x^2 comes close to the 0.05 barrier.

PROBLEM 7

? *The data given were taken from the paper "Applying stepwise multiple Regression Analysis to the reaction of formaldehyde with cotton cellulose". The dependent variable y (durable press rating) is a quantitative measure of wrinkle resistance. The four independent variables used in the model building process are HCHO, i.e. formaldehyde, concentration (x1), catalyst ratio (x2), curing temperature (x3), and curing time(x4).*

HCHO	catalyst ratio	temperature	time	durable press
8	4	100	1	1.4
2	4	180	7	2.2
7	4	180	1	4.6
10	7	120	5	4.9
7	4	180	5	4.6
7	7	180	1	4.7
7	13	140	1	4.6
5	4	160	7	4.5
4	7	140	3	4.8
5	1	100	7	1.4
8	10	140	3	4.7
2	4	100	3	1.6
4	10	180	3	4.5
6	7	120	7	4.7
10	13	180	3	4.8
4	10	160	5	4.6
4	13	100	7	4.3
10	10	120	7	4.9
5	4	100	1	1.7
8	13	140	1	4.6
10	1	180	1	2.6
2	13	140	1	3.1
6	13	180	7	4.7
7	1	120	7	2.5
5	13	140	1	4.5
8	1	160	7	2.1
4	1	180	7	1.8
6	1	160	1	1.5

Table 34 *Four independent inputs vs. durable press rating*

The diagram was composed from the estimated y and the measured y.

$$y = f(x) = 0.1607 \cdot x_1 + 0.2198 \cdot x_2 + 0.0112 \cdot x_3 + 0.1020 \cdot x_4 - 0.9122$$

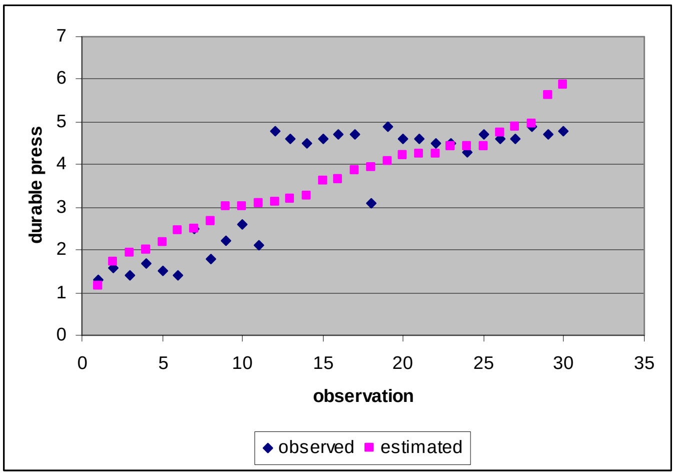

Figure 9 *Estimated and observed durable press rating*

The remarkable break between observation 11 and 12 caused me to establish a theory: do *all* input parameters significantly influence the durable press rating ?

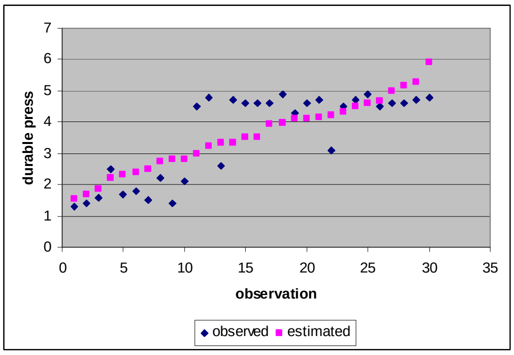

Figure 10 *Without curing time*

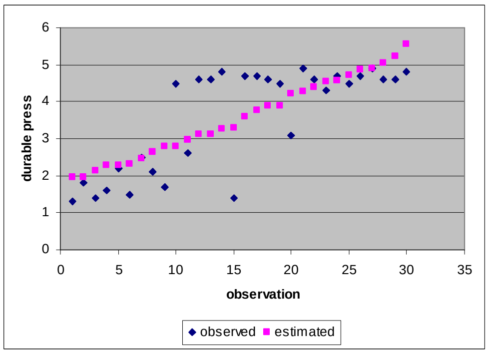

Figure 11 *Without curing time and curing temperature*

Comparing Excel's regression analysis yields:

	F significance	multiple R	R square	adjusted R square
Figure 9	3.8455E-06	0.8321	0.6924	0.6432
Figure 10	3.3477E-06	0.8095	0.6553	0.6155
Figure 11	4.7872E-06	0.7723	0.5964	0.5665

Table 35 *Comparing my three models*

In my eyes, the second model fits best the data. Its equation is:

$$y = f(x) = 0.1555 \cdot x_1 + 0.2139 \cdot x_2 + 0.0109 \cdot x_3 - 0.3883$$

PROBLEM 8

? *Milk samples were obtained from 14 Holstein-Friesian cows, and each was analyzed to determine uric acid concentration (µ mol/L). In addition to acid concentration, the total milk production (kg/day) was recorded for each cow.*

milk production	acid concentration
42.7	92
40.2	120
38.2	128
37.6	110
32.2	153
32.2	162
28.0	202
27.2	140
26.6	218
23.0	195
22.7	180
21.8	193
21.3	238
20.2	213

Table 36 *Acid concentration found in milk*

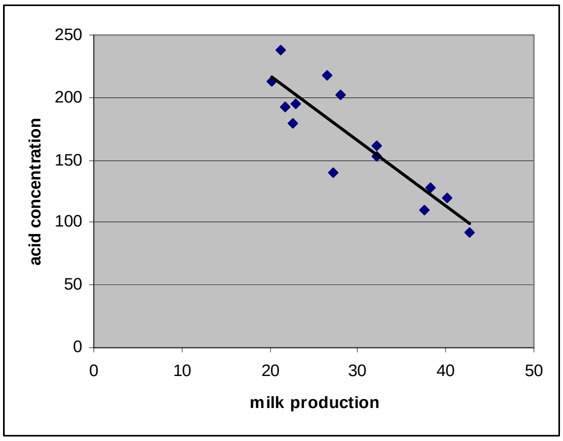

Table 37 *Linear regression analysis*

Repeating the same steps a hundred times bores even a simple-minded Bachelor of Science in Software Engineering like me …

$$y = f(x) = -5.2027 \cdot x + 321.2413$$

	df	SS	MS	F	Significance F
regression	1	20,625	20,625	44.6965	0.0000
residual	12	5,537	461		
total	13	26,163			

Table 38 *ANOVA*

	coefficients	standard error	t-stat	P-value
intercept	321.2413	23.7123	13.5474	0.0000
x	-5.2027	0.7782	-6.6855	0.0000

Table 39 *Coefficients, confidence intervals etc., part I*

	lower 95%	upper 95%	lower 95.0%	upper 95.0%
intercept	269.5766	372.9060	269.5766	372.9060
x	-6.8982	-3.5071	-6.8982	-3.5071

Table 40 *Coefficients, confidence intervals etc., part II*

Again, there is a relationship between the acid concentration and the total milk production of these cows. Maybe the whole regression analysis technique is biased since all the test generate “positive” results. Just my guess.

PROBLEM 9

? *25 observations on y = catch intake (number of fish), x1 = water temperature , x2 minimum tide height (m), x3 = number of pumps running, x4 = speed (knots), x5 = wind-range of direction (degrees) constitute a subset of the data that appeared in the paper "Multiple Regression Analysis for Forecasting Critical Fish Influxes at Power Station Intakes" (J. Applied Ecol. (1983)).*

temperature	tide height	pumps	speed	wind-range	catch intake
17	6.7	0.5	4	10	50
42	7.8	1	4	24	30
1	9.9	1.2	4	17	120
11	10.1	0.5	4	23	30
8	10	0.9	4	18	20
30	8.7	0.8	4	9	160
2	10.3	1.5	4	13	40
6	10.5	0.3	4	10	150
11	11	1.2	3	9	50
14	11.2	0.6	3	7	100
53	12.9	1.8	3	10	90
9	13.2	0.2	3	12	50
4	16.2	0.7	3	6	80
3	15.8	1.6	3	7	120
7	16.2	0.4	3	10	50
9	15.8	1.2	3	9	60
10	16	0.8	3	12	90
7	16.2	1.2	3	5	160
12	17.1	0.7	3	10	90
12	17.5	0.8	3	12	110
26	17.5	1.2	3	18	130
14	17.4	0.8	3	9	60
18	17.4	1.1	3	13	30
14	17.8	0.5	3	8	160
5	18	1.6	3	10	40

Table 41 *Various parameters influencing fish intake*

PROBLEM 10

? *The data for this example come from a study by Stamey et al (1989) that examined the correlation between the level of prostate specific antigen (PSA) an a number of clinical measures, in 97 men who where about to receive a radical prostatectomy. The goal is to predict the log of PSA (lpsa) from a number of measurements including log-cancer-volume (lcavol), log prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).*

Here comes a very, very, very, very, very, very, long table:

lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa	train
-0.580	2.769	50	-1.386	0	-1.386	6	0	-0.431	T
-0.994	3.320	58	-1.386	0	-1.386	6	0	-0.163	T
-0.511	2.691	74	-1.386	0	-1.386	7	20	-0.163	T
-1.204	3.283	58	-1.386	0	-1.386	6	0	-0.163	T
0.751	3.432	62	-1.386	0	-1.386	6	0	0.372	T
-1.050	3.229	50	-1.386	0	-1.386	6	0	0.765	T
0.737	3.474	64	0.615	0	-1.386	6	0	0.765	F
0.693	3.540	58	1.537	0	-1.386	6	0	0.854	T
-0.777	3.540	47	-1.386	0	-1.386	6	0	1.047	F
0.223	3.245	63	-1.386	0	-1.386	6	0	1.047	F
0.255	3.604	65	-1.386	0	-1.386	6	0	1.267	T
-1.347	3.599	63	1.267	0	-1.386	6	0	1.267	T
1.613	3.023	63	-1.386	0	-0.598	7	30	1.267	T
1.477	2.998	67	-1.386	0	-1.386	7	5	1.348	T
1.206	3.442	57	-1.386	0	-0.431	7	5	1.399	F
1.541	3.061	66	-1.386	0	-1.386	6	0	1.447	T
-0.416	3.516	70	1.244	0	-0.598	7	30	1.470	T
2.288	3.649	66	-1.386	0	0.372	6	0	1.493	T
-0.562	3.268	41	-1.386	0	-1.386	6	0	1.558	T
0.182	3.825	70	1.658	0	-1.386	6	0	1.599	T
1.147	3.419	59	-1.386	0	-1.386	6	0	1.639	T
2.059	3.501	60	1.475	0	1.348	7	20	1.658	F
-0.545	3.376	59	-0.799	0	-1.386	6	0	1.696	T
1.782	3.452	63	0.438	0	1.179	7	60	1.714	T
0.385	3.667	69	1.599	0	-1.386	6	0	1.732	F
1.447	3.125	68	0.300	0	-1.386	6	0	1.766	F
0.513	3.720	65	-1.386	0	-0.799	7	70	1.800	T
-0.400	3.866	67	1.816	0	-1.386	7	20	1.816	F
1.040	3.129	67	0.223	0	0.049	7	80	1.848	T
2.410	3.376	65	-1.386	0	1.619	6	0	1.895	T
0.285	4.090	65	1.963	0	-0.799	6	0	1.924	T
0.182	6.108	65	1.705	0	-1.386	6	0	2.008	F
1.275	3.037	71	1.267	0	-1.386	6	0	2.008	T
0.010	3.268	54	-1.386	0	-1.386	6	0	2.022	F
-0.010	3.217	63	-1.386	0	-0.799	6	0	2.048	T
1.308	4.120	64	2.171	0	-1.386	7	5	2.086	F
1.423	3.657	73	-0.580	0	1.658	8	15	2.158	T
0.457	2.375	64	-1.386	0	-1.386	7	15	2.192	T
2.661	4.085	68	1.374	1	1.833	7	35	2.214	T
0.798	3.013	56	0.936	0	-0.163	7	5	2.277	T
0.621	3.142	60	-1.386	0	-1.386	9	80	2.298	T
1.442	3.683	68	-1.386	0	-1.386	7	10	2.308	F
0.582	3.866	62	1.714	0	-0.431	6	0	2.327	T
1.772	3.897	61	-1.386	0	0.811	7	6	2.375	F
1.486	3.409	66	1.749	0	-0.431	7	20	2.522	T
1.664	3.393	61	0.615	0	-1.386	7	15	2.553	T
2.728	3.995	79	1.879	1	2.657	9	100	2.569	T
1.163	4.035	68	1.714	0	-0.431	7	40	2.569	F
1.746	3.498	43	-1.386	0	-1.386	6	0	2.592	F
1.221	3.568	70	1.374	0	-0.799	6	0	2.592	F
1.092	3.994	68	-1.386	0	-1.386	7	50	2.657	T
1.660	4.235	64	2.073	0	-1.386	6	0	2.678	T

0.513	3.634	64	1.493	0	0.049	7	70	2.684	F
2.127	4.121	68	1.766	0	1.447	7	40	2.691	F
3.154	3.516	59	-1.386	0	-1.386	7	5	2.705	F
1.267	4.280	66	2.122	0	-1.386	7	15	2.718	T
0.975	2.865	47	-1.386	0	0.501	7	4	2.788	F
0.464	3.765	49	1.423	0	-1.386	6	0	2.794	T
0.542	4.178	70	0.438	0	-1.386	7	20	2.806	T
1.061	3.851	61	1.295	0	-1.386	7	40	2.812	T
0.457	4.525	73	2.326	0	-1.386	6	0	2.842	T
1.997	3.720	63	1.619	1	1.910	7	40	2.854	F
2.776	3.525	72	-1.386	0	1.558	9	95	2.854	T
2.035	3.917	66	2.008	1	2.110	7	60	2.882	F
2.073	3.623	64	-1.386	0	-1.386	6	0	2.882	F
1.459	3.836	61	1.322	0	-0.431	7	20	2.888	F
2.023	3.878	68	1.783	0	1.322	7	70	2.920	T
2.198	4.051	72	2.308	0	-0.431	7	10	2.963	T
-0.446	4.409	69	-1.386	0	-1.386	6	0	2.963	T
1.194	4.780	72	2.326	0	-0.799	7	5	2.973	T
1.864	3.593	60	-1.386	1	1.322	7	60	3.013	T
1.160	3.341	77	1.749	0	-1.386	7	25	3.037	T
1.215	3.825	69	-1.386	1	0.223	7	20	3.056	F
1.839	3.237	60	0.438	1	1.179	9	90	3.075	F
2.999	3.849	69	-1.386	1	1.910	7	20	3.275	T
3.141	3.264	68	-0.051	1	2.420	7	50	3.338	T
2.011	4.434	72	2.122	0	0.501	7	60	3.393	T
2.538	4.355	78	2.326	0	-1.386	7	10	3.436	T
2.648	3.582	69	-1.386	1	2.584	7	70	3.458	T
2.779	3.823	63	-1.386	0	0.372	7	50	3.513	F
1.468	3.070	66	0.560	0	0.223	7	40	3.516	T
2.514	3.474	57	0.438	0	2.327	7	60	3.531	T
2.613	3.889	77	-0.528	1	0.560	7	30	3.565	T
2.678	3.838	65	1.115	0	1.749	9	70	3.571	F
1.562	3.710	60	1.696	0	0.811	7	30	3.588	T
3.303	3.519	64	-1.386	1	2.327	7	60	3.631	T
2.024	3.732	58	1.639	0	-1.386	6	0	3.680	T
1.732	3.369	62	-1.386	1	0.300	7	30	3.712	T
2.808	4.718	65	-1.386	1	2.464	7	60	3.984	T
1.562	3.695	76	0.936	1	0.811	7	75	3.994	T
3.246	4.102	68	-1.386	0	-1.386	6	0	4.030	T
2.533	3.678	61	1.348	1	-1.386	7	15	4.130	T
2.830	3.876	68	-1.386	1	1.322	7	60	4.385	T
3.821	3.897	44	-1.386	1	2.169	7	40	4.684	T
2.907	3.396	52	-1.386	1	2.464	7	10	5.143	F
2.883	3.774	68	1.558	1	1.558	7	80	5.478	T
3.472	3.975	68	0.438	1	2.904	7	20	5.583	F

Table 42 *Various parameters influencing fish intake*

Mrs. Liero generated a nice SPSS plot:

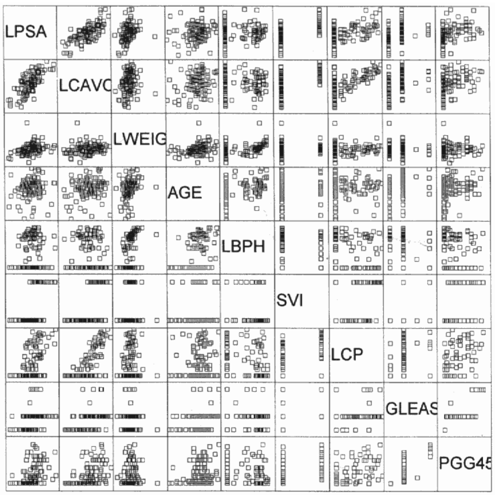

Figure 12 *Correlations*

In the end, 67 out of 97 records are valid (attribute *train* is *true*).

Regressions statistics	
Multiple R	0.8333
R Square	0.6944
Adjusted R Square	0.6522
Standard Error	0.7123
Observations	67

Table 43 *Regression statistics*

	df	SS	MS	F	Significance F
regression	8	66.8551	8.3569	16.4716	0.0000
residual	58	29.4264	0.5074		
total	66	96.2814			

Table 44 *ANOVA*

	coefficients	standard error	t-stat	P-value
intercept	0.4292	1.5536	0.2762	0.7833
x_1 (lcavol)	0.5765	0.1074	5.3663	0.0000
x_2 (lweight)	0.6140	0.2232	2.7508	0.0079
x_3 (age)	-0.0190	0.0136	-1.3959	0.1681
x_4 (lbph)	0.1448	0.0705	2.0558	0.0443
x_5 (svi)	0.7372	0.2986	2.4693	0.0165
x_6 (lcp)	-0.2063	0.1105	-1.8669	0.0670
x_7 (gleason)	-0.0295	0.2011	-0.1467	**0.8839**
x_8 (pgg45)	0.0095	0.0054	1.7378	0.0875

Table 45 *Coefficients, confidence intervals etc., part I*

	lower 95%	upper 95%	lower 95.0%	upper 95.0%
intercept	-2.6807	3.5390	-2.6807	3.5390
x_1 (lcavol)	0.3615	0.7916	0.3615	0.7916
x_2 (lweight)	0.1672	1.0608	0.1672	1.0608
x_3 (age)	-0.0462	0.0082	-0.0462	0.0082
x_4 (lbph)	0.0038	0.2859	0.0038	0.2859
x_5 (svi)	0.1396	1.3348	0.1396	1.3348
x_6 (lcp)	-0.4275	0.0149	-0.4275	0.0149
x_7 (gleason)	-0.4321	0.3731	-0.4321	0.3731
x_8 (pgg45)	-0.0014	0.0204	-0.0014	0.0204

Table 46 *Coefficients, confidence intervals etc., part II*

The p-value of x_7 is the largest one and far than the 0.05 threshold. In consequence, I suppose it does not influence significantly the level of PSA and remove it. After that, the whole regression analysis should be repeated but one will find out that the p-value of x_3 is too high. Throughout the next iteration, only x_1 (lcavol), x_2 (lweight), x_4 (lbph), and x_5 (svi) remain.

Regressions statistics	
Multiple R	0.8119
R Square	0.6592
Adjusted R Square	0.6372
Standard Error	0.7275
Observations	67

Table 47 *Regression statistics*

	df	SS	MS	F	Significance F
regression	4	63.4665	15.8666	29.9781	0.0000
residual	62	32.8150	0.5293		
total	66	96.2814			

Table 48 *ANOVA*

	coefficients	standard error	t-stat	P-value
intercept	-0.3259	0.7800	-0.4179	0.6775
x_1 (lcavol)	0.5055	0.0926	5.4614	0.0000
x_2 (lweight)	0.5388	0.2207	2.4413	0.0175
x_4 (lbph)	0.1400	0.0704	1.9885	**0.0512**
x_5 (svi)	0.6718	0.2732	2.4589	0.0167

Table 49 *Coefficients, confidence intervals etc., part I*

	lower 95%	upper 95%	lower 95.0%	upper 95.0%
intercept	-1.8851	1.2332	-1.8851	1.2332
x_1 (lcavol)	0.3205	0.6906	0.3205	0.6906
x_2 (lweight)	0.0976	0.9800	0.0976	0.9800
x_4 (lbph)	-0.0007	0.2808	-0.0007	0.2808
x_5 (svi)	0.1257	1.2180	0.1257	1.2180

Table 50 *Coefficients, confidence intervals etc., part II*

x_4's p-value is still above 0.05 but removing it would increase … (todo !)