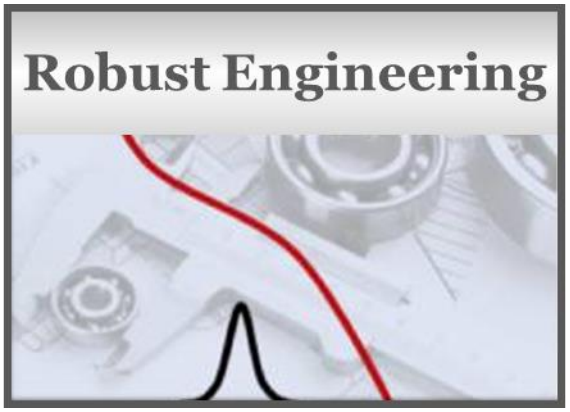


Session 2: Multiple Regression

In This Session We Will…

- Fit a 1st order response surface to a set of process data, using the least squares method.
- Explain the following terms: regression coefficient, fitted response value.
- Define and interpret two success criteria for a response surface equation: R^2 and PRESS RMSE.
- Show how R^2 can be calculated from an ANOVA table.
- Show how PRESS and PRESS RMSE are calculated from the residuals and the leverages.
- Construct 3D plots of the fitted surface.
- Explain why it is often helpful to work with coded x-variables.
- Introduce the model matrix M and give a formula for the regression coefficients.

Sintering of Bearing Liners

- The process is as follows:
  - deposit the powdered coating material onto sheet steel;
  - heat, stretch and re-heat the coated sheets;
  - form the liners.
- In a particular plant the machinery was antiquated, resulting in poor control of the furnace temperature and the stretching percentage
- Production data was collected to investigate whether better control of these parameters would give improved quality (increased bond strength of the coating)

20 sheets were followed through the process; the actual furnace temperatures and the stretching percentage were recorded

The Data

Row	Temp 1 (x_1, C)	Stretch $(x_2, \%)$	Temp 2 (x_3, C)	Bond strength (y, N)
1	860	2.8	869	90.7
2	850	2.9	850	85.8
3	861	3.0	867	97.2
4	861	2.5	865	86.5
5	888	2.1	877	92.5
6	864	2.3	864	88.5
7	875	2.6	876	95.1
8	853	2.2	861	79.0
9	866	2.1	866	84.6
10	884	2.2	890	92.2
11	868	2.9	874	96.3
12	888	2.6	888	104.4
13	887	2.0	886	92.8
14	849	2.7	861	84.1
15	877	2.6	874	93.6
16	855	2.1	858	83.5
17	864	2.7	867	89.1
18	874	2.1	869	88.9
19	853	2.2	857	77.7
20	889	3.0	883	104.1

Fitting A Response Surface

- We will fit a 1st order response surface (transfer function) for three x-variables which can be represented generically by:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

b is a ***regression coefficient***

Intercept

Linear term in each x-variable

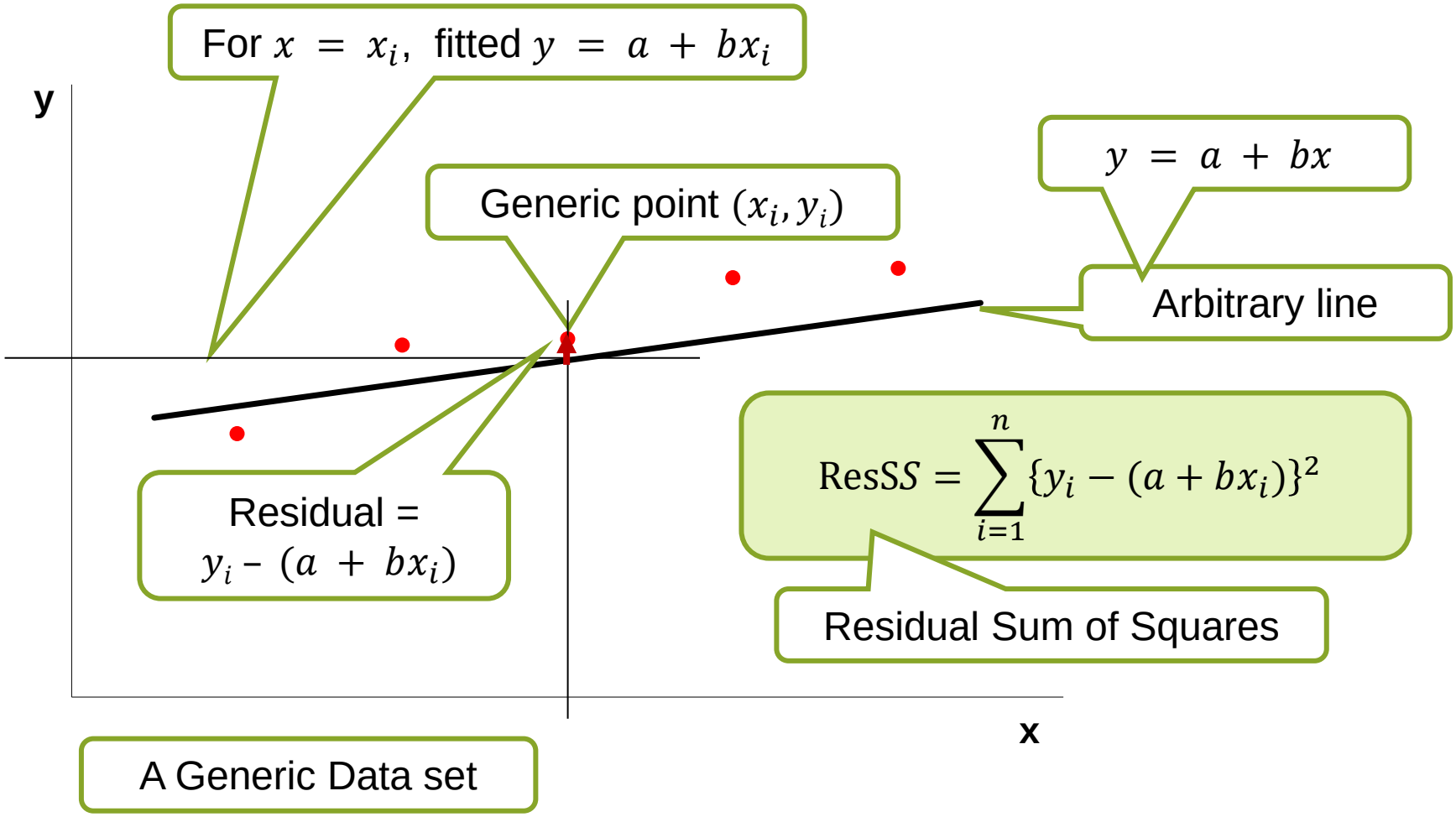
The values of the regression coefficients will define the fitted surface

- The equation will be fitted by the method of least squares.

In the M1 module, statistics for engineering, least squares was used to fit a straight line to data.

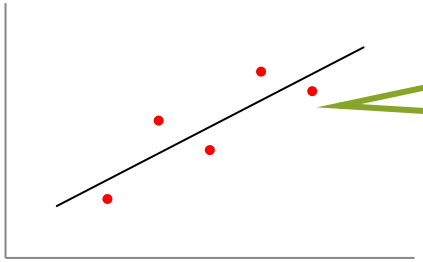
In here it is briefly described as background information.

Background: Least Square



Fitting A Response Surface (Cont.)

- Using the **Least squares** method the regression coefficients are selected such that the sum of squared **residuals** is as small as possible.



A residual is the distance from the data to the surface, measured in the y-direction

- For the sintering data the best-fitting 1st order surface is:

$$y = -340.2 + 0.3763\,x_1 + 12.28\,x_2 + 0.0843x_3$$

The formula that generates the coefficients will be introduced later

The main focus of this study was to evaluate the effects of the individual x's (represented by the regression coefficients)

…. but we will begin by looking at properties of the equation as a whole

Success Criteria For A Response Surface Equation

There are two types of numerical success criterion:

I. How well does the equation fit the data?

- we will define a measure known as R^2.

We should also check that the equation makes sense in engineering terms

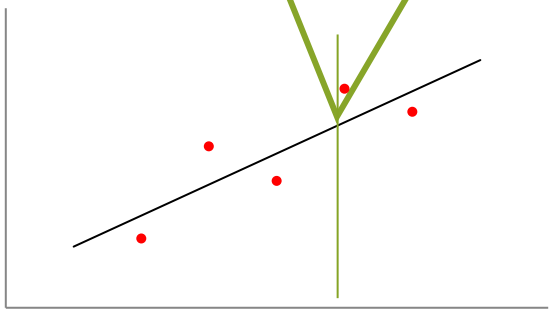
II. Is the model useful?

- for most response surface models, usefulness means **predicting** well.
- the ideal measure is prediction performance on new data ('evaluation data' or 'validation data') but ….
- in many studies evaluation data is not available, so we describe a method called **cross-validation** and a measure called **PRESS RMSE**.

A Numerical Measure of Fit: R^2

- R^2 is the squared of Pearson correlation coefficient between the **actual** y-values and the **fitted** y-values.

The fitted y-value is read off the response surface at the given x-values



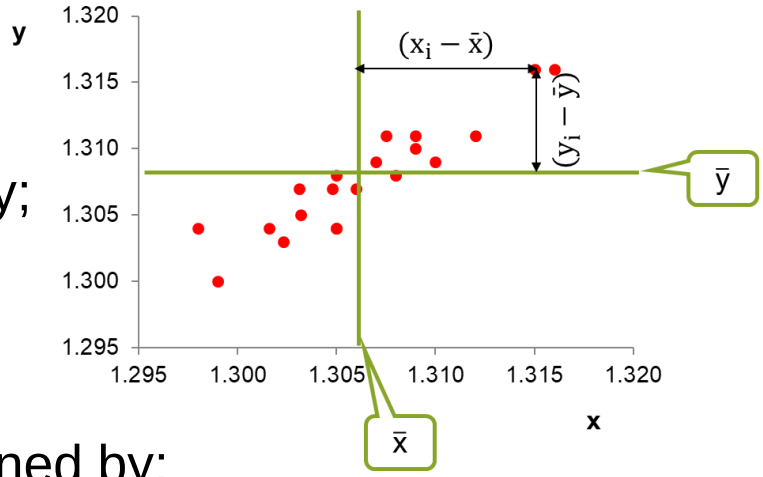
R^2 is a **multiple** correlation coefficient

Row	x1	x2	x3	Actual y	Fitted y
1	860	2.8	869	90.7	91.04
2	850	2.9	850	85.8	86.91
3	861	3.0	867	97.2	93.71
4	861	2.5	865	86.5	87.40
5	888	2.1	877	92.5	93.66
6	864	2.3	864	88.5	85.99
7	875	2.6	876	95.1	94.82
8	853	2.2	861	79.0	80.37
9	866	2.1	866	84.6	84.45
10	884	2.2	890	92.2	94.48
11	868	2.9	874	96.3	95.70
12	888	2.6	888	104.4	100.72
13	887	2.0	886	92.8	92.81
14	849	2.7	861	84.1	85.00
15	877	2.6	874	93.6	95.41
16	855	2.1	858	83.5	79.64
17	864	2.7	867	89.1	91.15
18	874	2.1	869	88.9	87.72
19	853	2.2	857	77.7	80.03
20	889	3.0	883	104.1	105.59

Background: The Pearson correlation coefficient

For a set of bivariate data,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$$



- The Pearson correlation coefficient is defined by;

$$r = \frac{SPxy}{\sqrt{SSx}\sqrt{SSy}}$$

- **SP** stands for sum of products and SPxy is defined by;

$$SPxy = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- **SS** stands for sum of squares; SSx and SSy are defined by;

$$SSx = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad SSy = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Plotting Actual Against Fitted y-values

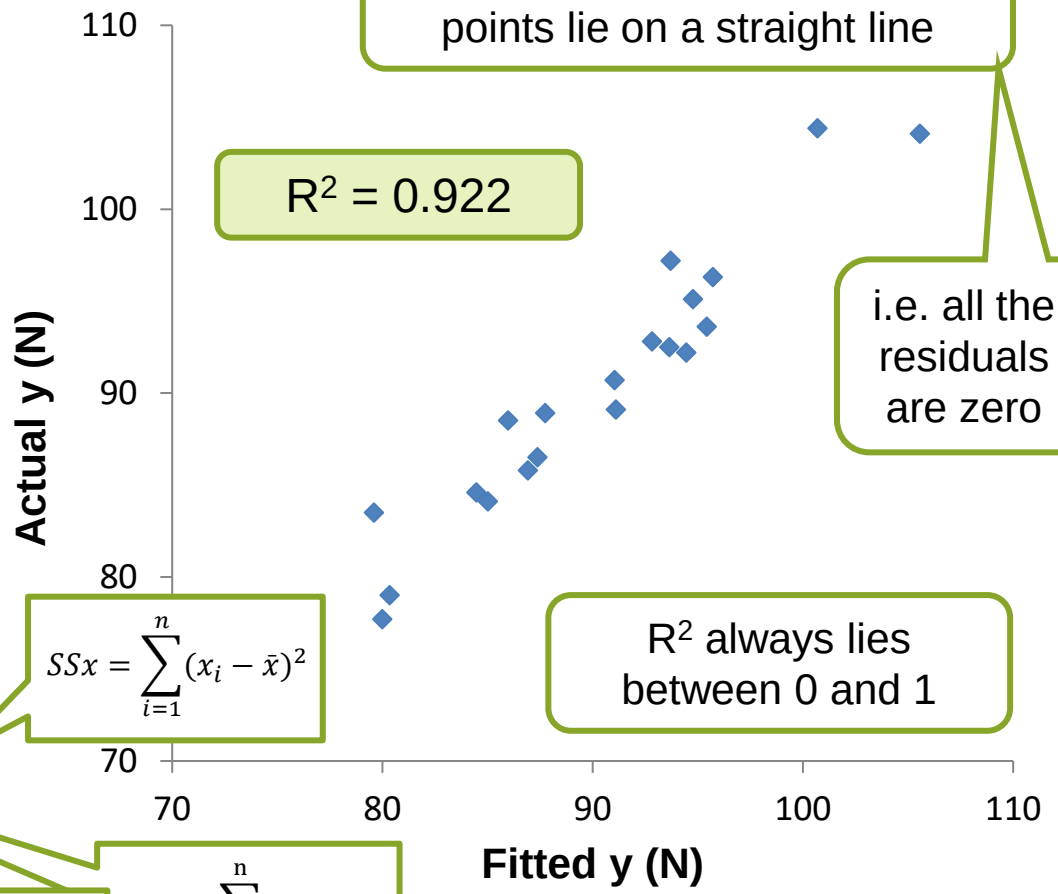
y (Actual y)	x (Fitted y)	y - $\bar{y}$	x - $\bar{x}$	SPxy	SSy	SSx
90.7	91.04	0.37	0.71	0.26	0.14	0.51
85.8	86.91	-4.53	-3.42	15.50	20.52	11.71
97.2	93.71	6.87	3.38	23.20	47.20	11.41
86.5	87.40	-3.83	-2.93	11.23	14.67	8.59
92.5	93.66	2.17	3.33	7.22	4.71	11.08
88.5	85.99	-1.83	-4.34	7.95	3.35	18.86
95.1	94.82	4.77	4.49	21.43	22.75	20.18
79.0	80.37	-11.33	-9.96	112.87	128.37	99.25
84.6	84.45	-5.73	-5.88	33.68	32.83	34.54
92.2	94.48	1.87	4.15	7.75	3.50	17.19
96.3	95.70	5.97	5.37	32.08	35.64	28.87
104.4	100.72	14.07	10.39	146.25	197.96	108.05
92.8	92.81	2.47	2.48	6.13	6.10	6.16
84.1	85.00	-6.23	-5.33	33.19	38.81	28.38
93.6	95.41	3.27	5.08	16.60	10.69	25.76
83.5	79.64	-6.83	-10.69	73.02	46.65	114.28
89.1	91.15	-1.23	0.82	-1.01	1.51	0.68
88.9	87.72	-1.43	-2.61	3.74	2.04	6.83
77.7	80.03	-12.63	-10.30	130.08	159.52	106.07
104.1	105.59	13.77	15.26	210.15	189.61	232.91
Average			Sum	891.31	966.58	891.31
90.3	90.33		r	0.960		
			R^2	0.922		

$R^2 = 1$ if and only if all the points lie on a straight line

$R^2 = 0.922$

i.e. all the residuals are zero

R^2 always lies between 0 and 1



$$SSx = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SSy = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SPxy = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$r = \frac{SPxy}{\sqrt{SSx}\sqrt{SSy}}$$

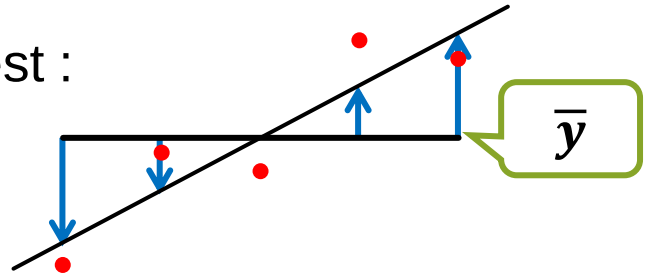
$\bar{y}$

$\bar{x}$

Alternative Method Of Calculating R^2: ANOVA Table

- As always in ANOVA, the bottom line is the 'Total sum of squares' for the y values: we call this SSy.
- In multiple regression there are several ways of splitting this into components; the split shown below is the simplest :

Residual sum of squares (ReSS) = sum of squares of difference between $actual\ y$ and $fitted\ y$.



Source of variation	Sum of squares	Degrees of freedom
Regression	RegSS	RegDF
Residual	ResSS	ResDF
Total (around the mean)	SSy	Total DF (for variation)

Regression sum of squares (RegSS) = sum of squares of blue arrows

ANOVA Details

Source of variation	Sum of squares	Degrees of freedom (DoF)
Regression	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	p
Residual	$\sum_{i=1}^{n}\{y_i - (a + bx_i)\}^2$	n – p – 1
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	n – 1

p = number of x-variables

The ‘Total sum of squares’ is a measure of the overall amount of variation in the response values

n = number of rows of data

ANOVA For The Sintering Data

Actual y	Fitted y	Fit y - $\bar{y}$	Act y - Fit y	Act y - $\bar{y}$	RegSS	ResSS	SSy
90.7	91.04	0.71	-0.34	0.37	0.51	0.12	0.14
85.8	86.91	-3.42	-1.11	-4.53	11.71	1.23	20.52
97.2	93.71	3.38	3.49	6.87	11.41	12.20	47.20
86.5	87.40	-2.93	-0.90	-3.83	8.59	0.81	14.67
92.5	93.66	3.33	-1.16	2.17	11.08	1.34	4.71
88.5	85.99	-4.34	2.51	-1.83	18.86	6.31	3.35
95.1	94.82	4.49	0.28	4.77	20.18	0.08	22.75
79.0	80.37	-9.96	-1.37	-11.33	99.25	1.87	128.37
84.6	84.45	-5.88	0.15	-5.73	34.54	0.02	32.83
92.2	94.48	4.15	-2.28	1.87	17.19	5.18	3.50
96.3	95.70	5.37	0.60	5.97	28.87	0.36	35.64
104.4	100.72	10.39	3.68	14.07	108.05	13.51	197.96
92.8	92.81	2.48	-0.01	2.47	6.16	0.00	6.10
84.1	85.00	-5.33	-0.90	-6.23	28.38	0.81	38.81
93.6	95.41	5.08	-1.81	3.27	25.76	3.26	10.69
83.5	79.64	-10.69	3.86	-6.83	114.28	14.90	46.65
89.1	91.15	0.82	-2.05	-1.23	0.68	4.21	1.51
88.9	87.72	-2.61	1.18	-1.43	6.83	1.40	2.04
77.7	80.03	-10.30	-2.33	-12.63	106.07	5.43	159.52
104.1	105.59	15.26	-1.49	13.77	232.91	2.22	189.61
Average	90.3	90.33			Sum	891.31	75.27	966.58
					R^2	0.922		

$\bar{y}$

Source of variation	Sum of squares	DoF
Regression (RegSS)	891.27	3
Residual (ResSS)	75.33	16
Total (SSy)	966.58	19

If the 1st order model fits the data well, RegSS will be much bigger than ResSS

R^2 is defined as $\frac{\text{RegSS}}{\text{SSy}} = \frac{891.27}{966.58} = 0.922$

$$SSy = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Equivalently, $R^2 = 1 - \frac{\text{ResSS}}{\text{SSy}}$

$$ResSS = \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2$$

$$RegSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

We will discuss the interpretation of R^2 after introducing cross-validation

Cross-Validation

- This a technique for model evaluation. In this method we collect only one set of data but let each run take a turn at acting as a validation run.
- If there are n runs in the training data, the equation is fitted to n–1 runs.
  - a prediction is made for the remaining one.
  - the prediction error is measured.

Called the **leave one out** prediction error

- This idea is used successfully in many areas of statistics.
- In multiple regression the calculations are very easy, because the leave one out prediction error is:

$$PE = \frac{\text{residual}}{(1 - \text{leverage})}$$

M matrix is introduced next

The leverage value measures the 'pull' of the point on the fitted surface.

leverage values are the diagonal elements of the matrix $M(M'M)^{-1}M'$

The Model Matrix M

- M has a row for each line of data and a column for each term in the fitted equation:

Model Matrix M is often called '**X**'

The intercept column consists entirely of 1's

The other columns contain the values of the x-variables

Intercept	x_1	x_2	x_3
1	860	2.8	869
1	850	2.9	850
1	861	3.0	867
1	861	2.5	865
1	888	2.1	877
1	864	2.3	864
1	875	2.6	876
1	853	2.2	861
1	866	2.1	866
1	884	2.2	890
1	868	2.9	874
1	888	2.6	888
1	887	2.0	886
1	849	2.7	861
1	877	2.6	874
1	855	2.1	858
1	864	2.7	867
1	874	2.1	869
1	853	2.2	857
1	889	3.0	883

Matrix M plays a substantial role in relation to modelling and it will become clear as we progress through this course.

Calculating Leverage Values Using $\mathbf{M(M'M)^{-1}M'}$

M'																			
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
860	850	861	861	888	864	875	853	866	884	868	888	887	849	877	855	864	874	853	889
2.8	2.9	3.0	2.5	2.1	2.3	2.6	2.2	2.1	2.2	2.9	2.6	2.0	2.7	2.6	2.1	2.7	2.1	2.2	3.0
869	850	867	865	877	864	876	861	866	890	874	888	886	861	874	858	867	869	857	883

M			
1	860	2.8	869
1	850	2.9	850
1	861	3.0	867
1	861	2.5	865
1	888	2.1	877
1	864	2.3	864
1	875	2.6	876
1	853	2.2	861
1	866	2.1	866
1	884	2.2	890
1	868	2.9	874
1	888	2.6	888
1	887	2.0	886
1	849	2.7	861
1	877	2.6	874
1	855	2.1	858
1	864	2.7	867
1	874	2.1	869
1	853	2.2	857
1	889	3.0	883

1. Transpose

2. Multiply

M'M			
2.000E+01	1.737E+04	4.960E+01	1.740E+04
1.737E+04	1.508E+07	4.306E+04	1.511E+07
4.960E+01	4.306E+04	1.252E+02	4.315E+04
1.740E+04	1.511E+07	4.315E+04	1.514E+07

3. Inverse

$(M'M)^{-1}$			
3.774E+02	2.578E-01	-7.555E-01	-6.888E-01
2.578E-01	2.054E-03	7.790E-03	-2.368E-03
-7.555E-01	7.790E-03	4.827E-01	-8.281E-03
-6.888E-01	-2.368E-03	-8.281E-03	3.178E-03

4. Multiply

$M(M'M)^{-1}$			
-1.574E+00	-1.195E-02	9.893E-02	1.351E-02
8.859E+00	1.329E-02	2.266E-01	-2.403E-02
-8.982E-02	-3.600E-03	2.198E-01	3.127E-03
1.666E+00	-2.759E-03	-4.976E-03	9.111E-04
6.638E-01	2.116E-02	-8.712E-02	-2.157E-02
3.279E+00	4.212E-03	-6.987E-02	-7.715E-03
-2.377E+00	7.234E-04	6.126E-02	1.893E-03
2.585E+00	-1.205E-02	-1.790E-01	9.626E-03
2.568E+00	2.025E-03	-1.674E-01	-4.438E-03
-9.397E+00	-1.706E-02	-1.777E-01	2.839E-02
-3.031E+00	-6.579E-03	1.681E-01	9.627E-03
-7.291E+00	-9.945E-04	6.315E-02	9.249E-03
-5.718E+00	-2.986E-03	-2.177E-01	1.023E-02
1.176E+00	-1.637E-02	3.122E-02	1.496E-02
-4.838E-01	9.566E-03	9.340E-02	-9.199E-03
5.242E+00	-1.622E-03	-1.868E-01	-3.817E-03
9.104E-01	2.237E-04	9.837E-02	-1.492E-03
2.564E+00	1.135E-02	-1.299E-01	-1.385E-02
5.340E+00	-2.582E-03	-1.459E-01	-3.087E-03
-3.891E+00	1.601E-02	3.054E-01	-1.232E-02

Calculating Leverage Values Using $\mathbf{M(M'M)^{-1}M'}$

M'

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
860	850	861	861	888	864	875	853	866	884	868	888	887	849	877	855	864	874	853	889
2.8	2.9	3.0	2.5	2.1	2.3	2.6	2.2	2.1	2.2	2.9	2.6	2.0	2.7	2.6	2.1	2.7	2.1	2.2	3.0
869	850	867	865	877	864	876	861	866	890	874	888	886	861	874	858	867	869	857	883

$M(M'M)^{-1}$

-1.574E+00	-1.195E-02	9.893E-02	1.351E-02
8.859E+00	1.329E-02	2.266E-01	-2.403E-02

5. Multiply

$M(M'M)^{-1}M'$

0.166	0.039	0.147	0.070	-0.130	0.001	0.062	0.082	-0.015	0.104	0.148	0.068	-0.006	0.179	0.011	0.008	0.081	-0.071	0.028	0.028
0.039	0.385	0.145	0.080	0.060	0.099	0.024	0.002	0.032	-0.283	0.047	-0.091	-0.192	0.062	0.099	0.078	0.117	0.066	0.098	0.133
0.147	0.145	0.181	0.065	-0.083	0.007	0.071	0.015	-0.038	-0.006	0.156	0.061	-0.073	0.139	0.057	-0.023	0.104	-0.057	0.003	0.130
0.070	0.080	0.065	0.065	0.004	0.057	0.036	0.085	0.055	0.026	0.052	0.011	0.015	0.094	0.029	0.078	0.058	0.035	0.082	0.002
-0.130	0.060	-0.083	0.004	0.351	0.106	0.054	-0.053	0.123	-0.023	-0.077	0.070	0.144	-0.181	0.139	0.063	0.007	0.227	0.033	0.164
0.001	0.099	0.007	0.057	0.106	0.092	0.024	0.075	0.098	-0.018	-0.011	-0.014	0.040	0.024	0.048	0.114	0.040	0.109	0.106	0.001
0.062	0.024	0.071	0.036	0.054	0.024	0.073	0.005	0.017	0.082	0.083	0.105	0.064	0.032	0.071	-0.006	0.054	0.029	-0.003	0.121
0.082	0.002	0.015	0.085	-0.053	0.075	0.005	0.197	0.106	0.102	0.016	-0.037	0.064	0.156	-0.039	0.162	0.033	0.039	0.158	-0.168
-0.015	0.032	-0.038	0.055	0.123	0.098	0.017	0.106	0.127	0.040	-0.038	-0.010	0.098	0.014	0.030	0.140	0.018	0.130	0.124	-0.052
0.104	-0.283	-0.006	0.026	-0.023	-0.018	0.082	0.102	0.040	0.397	0.091	0.201	0.268	0.082	-0.009	0.001	-0.004	-0.011	-0.011	-0.029
0.148	0.047	0.156	0.052	-0.077	-0.011	0.083	0.016	-0.038	0.091	0.160	0.113	-0.001	0.126	0.050	-0.043	0.085	-0.062	-0.023	0.125
0.068	-0.091	0.061	0.011	0.070	-0.014	0.105	-0.037	-0.010	0.201	0.113	0.204	0.148	-0.001	0.085	-0.073	0.039	0.010	-0.074	0.182
-0.006	-0.192	-0.073	0.015	0.144	0.040	0.064	0.064	0.098	0.268	-0.001	0.148	0.261	-0.033	0.038	0.049	-0.017	0.104	0.023	0.007
0.179	0.062	0.139	0.094	-0.181	0.024	0.032	0.156	0.014	0.082	0.126	-0.001	-0.033	0.237	-0.030	0.075	0.081	-0.072	0.096	-0.080
0.011	0.099	0.057	0.029	0.139	0.048	0.071	-0.039	0.030	-0.009	0.050	0.085	0.038	-0.030	0.109	-0.001	0.058	0.079	-0.002	0.178
0.008	0.078	-0.023	0.078	0.063	0.114	-0.006	0.162	0.140	0.001	-0.043	-0.073	0.049	0.075	-0.001	0.189	0.028	0.116	0.177	-0.130
0.081	0.117	0.104	0.058	0.007	0.040	0.054	0.033	0.018	-0.004	0.085	0.039	-0.017	0.081	0.058	0.028	0.075	0.016	0.039	0.087
-0.071	0.066	-0.057	0.035	0.227	0.109	0.029	0.039	0.130	-0.011	-0.062	0.010	0.104	-0.072	0.079	0.116	0.016	0.179	0.094	0.039
0.028	0.098	0.003	0.082	0.033	0.106	-0.003	0.158	0.124	-0.011	-0.023	-0.074	0.023	0.096	-0.002	0.177	0.039	0.094	0.171	-0.119
0.028	0.133	0.130	0.002	0.164	0.001	0.121	-0.168	-0.052	-0.029	0.125	0.182	0.007	-0.080	0.178	-0.130	0.087	0.039	-0.119	0.381

Calculating Leverage Values Using $\mathbf{M(M'M)^{-1}M'}$

M'																			
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
860	850	861	861	888	864	875	853	866	884	868	888	887	849	877	855	864	874	853	889
2.8	2.9	3.0	2.5	2.1	2.3	2.6	2.2	2.1	2.2	2.9	2.6	2.0	2.7	2.6	2.1	2.7	2.1	2.2	3.0
869	850	867	865	877	864	876	861	866	890	874	888	886	861	874	858	867	869	857	883

$\mathbf{M(M'M)^{-1}}$			
-1.574E+00	-1.195E-02	9.893E-02	1.351E-02
8.859E+00	1.329E-02	2.266E-01	-2.403E-02

5. Multiply

Leading diagonals = Leverage values

$\mathbf{M(M'M)^{-1}M'}$																			
0.166	0.039	0.147	0.070	-0.130	0.001	0.062	0.082	-0.015	0.104	0.148	0.068	-0.006	0.179	0.011	0.008	0.081	-0.071	0.028	0.028
0.039	0.385	0.145	0.080	0.060	0.099	0.024	0.002	0.032	-0.283	0.047	-0.091	-0.192	0.062	0.099	0.078	0.117	0.066	0.098	0.133
0.147	0.145	0.181	0.065	-0.083	0.007	0.071	0.015	-0.038	-0.006	0.156	0.061	-0.073	0.139	0.057	-0.023	0.104	-0.057	0.003	0.130
0.070	0.080	0.065	0.065	0.004	0.057	0.036	0.085	0.055	0.026	0.052	0.011	0.015	0.094	0.029	0.078	0.058	0.035	0.082	0.002
-0.130	0.060	-0.083	0.004	0.351	0.106	0.054	-0.053	0.123	-0.023	-0.077	0.070	0.144	-0.181	0.139	0.063	0.007	0.227	0.033	0.164
0.001	0.099	0.007	0.057	0.106	0.092	0.024	0.075	0.098	-0.018	-0.011	-0.014	0.040	0.024	0.048	0.114	0.040	0.109	0.106	0.001
0.062	0.024	0.071	0.036	0.054	0.024	0.073	0.005	0.017	0.082	0.083	0.105	0.064	0.032	0.071	-0.006	0.054	0.029	-0.003	0.121
0.082	0.002	0.015	0.085	-0.053	0.075	0.005	0.197	0.106	0.102	0.016	-0.037	0.064	0.156	-0.039	0.162	0.033	0.039	0.158	-0.168
-0.015	0.032	-0.038	0.055	0.123	0.098	0.017	0.106	0.127	0.040	-0.038	-0.010	0.098	0.014	0.030	0.140	0.018	0.130	0.124	-0.052
0.104	-0.283	-0.006	0.026	-0.023	-0.018	0.082	0.102	0.040	0.397	0.091	0.201	0.268	0.082	-0.009	0.001	-0.004	-0.011	-0.011	-0.029
0.148	0.047	0.156	0.052	-0.077	-0.011	0.083	0.016	-0.038	0.091	0.160	0.113	-0.001	0.126	0.050	-0.043	0.085	-0.062	-0.023	0.125
0.068	-0.091	0.061	0.011	0.070	-0.014	0.105	-0.037	-0.010	0.201	0.113	0.204	0.148	-0.001	0.085	-0.073	0.039	0.010	-0.074	0.182
-0.006	-0.192	-0.073	0.015	0.144	0.040	0.064	0.064	0.098	0.268	-0.001	0.148	0.261	-0.033	0.038	0.049	-0.017	0.104	0.023	0.007
0.179	0.062	0.139	0.094	-0.181	0.024	0.032	0.156	0.014	0.082	0.126	-0.001	-0.033	0.237	-0.030	0.075	0.081	-0.072	0.096	-0.080
0.011	0.099	0.057	0.029	0.139	0.048	0.071	-0.039	0.030	-0.009	0.050	0.085	0.038	-0.030	0.109	-0.001	0.058	0.079	-0.002	0.178
0.008	0.078	-0.023	0.078	0.063	0.114	-0.006	0.162	0.140	0.001	-0.043	-0.073	0.049	0.075	-0.001	0.189	0.028	0.116	0.177	-0.130
0.081	0.117	0.104	0.058	0.007	0.040	0.054	0.033	0.018	-0.004	0.085	0.039	-0.017	0.081	0.058	0.028	0.075	0.016	0.039	0.087
-0.071	0.066	-0.057	0.035	0.227	0.109	0.029	0.039	0.130	-0.011	-0.062	0.010	0.104	-0.072	0.079	0.116	0.016	0.179	0.094	0.039
0.028	0.098	0.003	0.082	0.033	0.106	-0.003	0.158	0.124	-0.011	-0.023	-0.074	0.023	0.096	-0.002	0.177	0.039	0.094	0.171	-0.119
0.028	0.133	0.130	0.002	0.164	0.001	0.121	-0.168	-0.052	-0.029	0.125	0.182	0.007	-0.080	0.178	-0.130	0.087	0.039	-0.119	0.381

Cross-Validation For The Sintering Data

Leading diagonal elements of matrix $M(M'M)^{-1}M'$

$$PE = \frac{\text{residual}}{(1 - \text{leverage})}$$

Actual y	Fitted y	Residual	Leverage	PE	PE^2
90.7	91.05	-0.35	0.166	-0.42	0.176
85.8	86.93	-1.13	0.385	-1.84	3.376
97.2	93.71	3.49	0.181	4.26	18.154
86.5	87.39	-0.89	0.065	-0.95	0.907
92.5	93.65	-1.15	0.351	-1.77	3.141
88.5	85.98	2.52	0.092	2.77	7.694
95.1	94.79	0.31	0.073	0.33	0.112
79.0	80.36	-1.36	0.197	-1.69	2.868
84.6	84.48	0.12	0.127	0.14	0.019
92.2	94.47	-2.27	0.397	-3.76	14.164
96.3	95.72	0.58	0.160	0.69	0.477
104.4	100.71	3.69	0.204	4.63	21.465
92.8	92.83	-0.03	0.261	-0.04	0.002
84.1	85.02	-0.92	0.237	-1.21	1.453
93.6	95.44	-1.84	0.109	-2.06	4.260
83.5	79.61	3.89	0.189	4.80	22.993
89.1	91.12	-2.02	0.075	-2.18	4.772
88.9	87.75	1.15	0.179	1.40	1.963
77.7	80.00	-2.30	0.171	-2.77	7.694
104.1	105.57	-1.47	0.381	-2.38	5.646
Average	90.3	90.33		PRESS	121.3
				PRESS RMSE	2.46

The PRESS RMSE is defined as

$$\sqrt{\frac{\text{PRESS}}{n}}$$

$$\text{PRESS RMSE} = \sqrt{\frac{121.3}{20}} = 2.46 \text{ N}$$

= Prediction Residual Sum of Squares

The sum of squares of these errors is called PRESS

Interpreting R^2 And PRESS RMSE

- Note that R^2 is dimensionless whereas PRESS RMSE is measured in the units of the response variable.
- The fact that R^2 is dimensionless has advantages and disadvantages;
  - advantage: we can compare the fit of surfaces in widely different contexts.
  - disadvantage: although bigger R^2 is better, there is no rule to tell us how big is big enough.

In some applications we may have R^2 very close to 1, but the prediction errors are still too large for the model to be useful

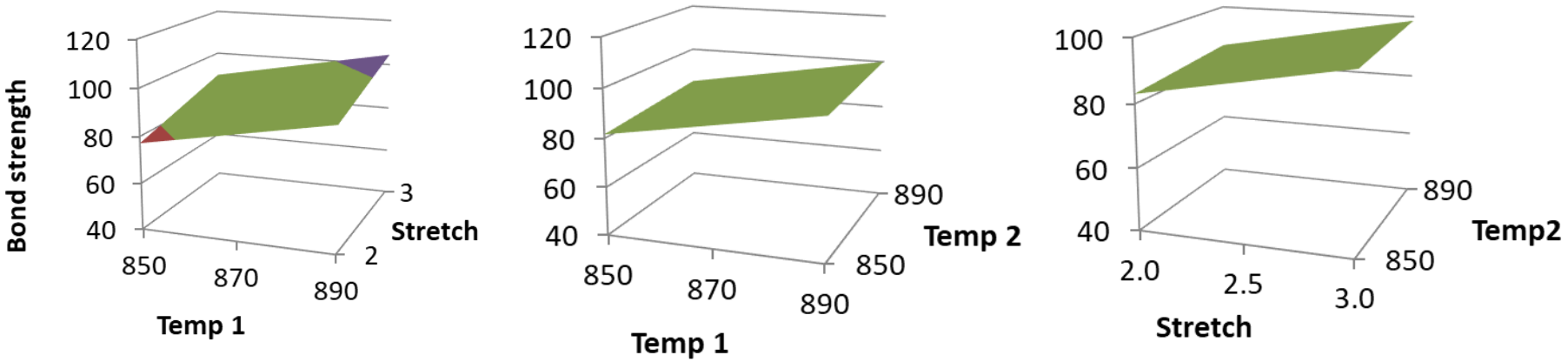
- In the sintering example, PRESS RMSE tells us that we are able to predict bond strength within 2.5 N (on average).
  - this was regarded as very good for the sintering process.

Interpreting The Regression Coefficients

- The response surface equation is:

$$y = -340.2 + 0.3763\,x_1 + 12.28\,x_2 + 0.0843x_3$$

- As a first step in interpretation we plot the surface against two x's at a time, keeping the third one fixed.



Over the observed ranges of the x-variables, the predicted effect on the response is is much the same for x_1 (Temp 1) and x_2 (stretch).

…. although the regression coefficients have very different magnitudes

Coding The x-Variables

- We explore the alternate way to represent the x-variables.
- Interpretation of regression output is often easier if the x-variables are coded (standardized):
  - use –1 to represent the lowest observed value.
  - use +1 to represent the highest.
  - code the intermediate values by linear interpolation.
  - e.g. for x_1:

Temp C	Code
849	–1
889	+1
869	0
860	–0.45

$$= \frac{860 - 869}{889 - 849} \times 2$$

An alternative coding method is to subtract the average of each x, then divide by their SD's

We use the ±1 coding because it fits in better with designed experiments

The Coded x-Variables

x_1 coded	x_2 coded	x_3 coded
-0.45	0.60	-0.05
-0.95	0.80	-1.00
-0.40	1.00	-0.15
-0.40	0.00	-0.25
0.95	-0.80	0.35
-0.25	-0.40	-0.30
0.30	0.20	0.30
-0.80	-0.60	-0.45
-0.15	-0.80	-0.20
0.75	-0.60	1.00
-0.05	0.80	0.20
0.95	0.20	0.90
0.90	-1.00	0.80
-1.00	0.40	-0.45
0.40	0.20	0.20
-0.70	-0.80	-0.60
-0.25	0.40	-0.15
0.25	-0.80	-0.05
-0.80	-0.60	-0.65
1.00	1.00	0.65

The Regression Equation With Coded x-Variables

$$y = 90.83 + 7.53x_1 + 6.14x_2 + 1.69x_3$$

We should really set up a new notation for the coded x-variables, but we will ignore this!

- The magnitudes of the regression coefficients b_1 and b_2 are now roughly equal.
- Also, the intercept is now a meaningful number;
  - 90.83 is the predicted bond strength when all x's are at coded 0.
  - i.e. at the 'centre' of the data.

Formula To Generate Coefficients

- Let $\underline{y}$ be a column vector containing the response values.
- Let M be the model matrix.
- Let $\underline{b}$ be a column vector containing the regression coefficients; in our example:

$$\underline{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

- Then;

$$\underline{b} = (M'M)^{-1} M'\underline{y}$$

Matrix M once again.

You won't need to work directly with this formula.

Calculating (Coded) Coefficients Using $(M'M)^{-1}\ M'\underline{y}$

$$y = 90.83 + 7.53\ x_1 + 6.14\ x_2 + 1.69 x_3$$

M' (Coded)

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
-0.45	-0.95	-0.40	-0.40	0.95	-0.25	0.30	-0.80	-0.15	0.75	-0.05	0.95	0.90	-1.00	0.40	-0.70	-0.25	0.25	-0.80	1.00
0.60	0.80	1.00	0.00	-0.80	-0.40	0.20	-0.60	-0.80	-0.60	0.80	0.20	-1.00	0.40	0.20	-0.80	0.40	-0.80	-0.60	1.00
-0.05	-1.00	-0.15	-0.25	0.35	-0.30	0.30	-0.45	-0.20	1.00	0.20	0.90	0.80	-0.45	0.20	-0.60	-0.15	-0.05	-0.65	0.65

1. Transpose

M (Coded)

1	-0.45	0.60	-0.05
1	-0.95	0.80	-1.00
1	-0.40	1.00	-0.15
1	-0.40	0.00	-0.25
1	0.95	-0.80	0.35
1	-0.25	-0.40	-0.30
1	0.30	0.20	0.30
1	-0.80	-0.60	-0.45
1	-0.15	-0.80	-0.20
1	0.75	-0.60	1.00
1	-0.05	0.80	0.20
1	0.95	0.20	0.90
1	0.90	-1.00	0.80
1	-1.00	0.40	-0.45
1	0.40	0.20	0.20
1	-0.70	-0.80	-0.60
1	-0.25	0.40	-0.15
1	0.25	-0.80	-0.05
1	-0.80	-0.60	-0.65
1	1.00	1.00	0.65

2. Multiply

M'M (Coded)

2.000E+01	-7.000E-01	-8.000E-01	1.000E-01
-7.000E-01	8.835E+00	-1.210E+00	6.480E+00
-8.000E-01	-1.210E+00	8.880E+00	-3.500E-01
1.000E-01	6.480E+00	-3.500E-01	5.595E+00

3. Inverse

$(M'M)^{-1}$ (Coded)

5.181E-02	3.660E-02	7.968E-03	-4.282E-02
3.660E-02	8.215E-01	7.790E-02	-9.472E-01
7.968E-03	7.790E-02	1.207E-01	-8.281E-02
-4.282E-02	-9.472E-01	-8.281E-02	1.271E+00

4. Multiply

$\underline{y}$

90.7
85.8
97.2
86.5
92.5
88.5
95.1
79.0
84.6
92.2
96.3
104.4
92.8
84.1
93.6
83.5
89.1
88.9
77.7
104.1

Actual y values

$(M'M)^{-1}M'\underline{y}$ (Coded)

90.83
7.53
6.14
1.69

5. Multiply

$(M'M)^{-1}M'$ (Coded)

0.04226	0.0662	0.0516	0.0479	0.0652	0.0523	0.0515	0.037	0.0485	0.0317	0.0478	0.0496	0.0425	0.0377	0.0595	0.0455	0.0523	0.0567	0.0456	0.0686
-0.239	0.2657	-0.072	-0.055	0.4232	0.0842	0.0145	-0.241	0.0405	-0.341	-0.132	-0.02	-0.06	-0.327	0.1913	-0.032	0.0045	0.227	-0.052	0.3203
0.04946	0.1133	0.1099	-0.002	-0.044	-0.035	0.0306	-0.089	-0.084	-0.089	0.0841	0.0316	-0.109	0.0156	0.0467	-0.093	0.0492	-0.065	-0.073	0.1527
0.27015	-0.481	0.0625	0.0182	-0.431	-0.154	0.0379	0.1925	-0.089	0.5678	0.1925	0.185	0.2046	0.2991	-0.184	-0.076	-0.03	-0.277	-0.062	-0.246

ANOVA: Sum of Squares In Multiple Regression

- Let $\underline{y}$ be the vector of n response values, let $\underline{1}$ be a vector of 1's (length n); let M be the model matrix and let $\mathbf{I}$ be an $n \times n$ unit matrix.
- The ANOVA table can be found from the following quadratic forms:

Source of variation	Sum of squares	Degrees of freedom
Regression	$\underline{y}\left(M(M'M)^{-1}M' - \underline{1}(\underline{1}'\underline{1})^{-1}\underline{1}'\right)\underline{y}$	p
Residual	$\underline{y}(\boldsymbol{I} - M(M'M)^{-1}M')\underline{y}$	n – p – 1
Total	$\underline{y}\left(\boldsymbol{I} - \underline{1}(\underline{1}'\underline{1})^{-1}\underline{1}'\right)\underline{y}$	n – 1

In This Session We Have ….

- Fitted a 1st order response surface to a set of process data, using the least squares method.
- Explained the following terms: regression coefficient, residual, fitted response value.
- Defined and interpreted two success criteria for a response surface equation: R^2 and PRESS RMSE.
- Shown how R^2 can be calculated from an ANOVA table.
- Shown how PRESS and PRESS RMSE are calculated from the residuals and the leverages.
- Constructed 3D plots of the fitted surface.
- Explained why it is often helpful to work with coded x-variables.
- Introduced the model matrix M and given a formula for the regression coefficients.

Session 2: Multiple Regression

Tutorial and Exercise

Tutorial

- **Session TS02: Multiple Regression**
- **Objectives**

  Use multiple regression to fit a 1st order response surface in engineering units and in coded units.
- **Engineering Scenario**

  The sintering of bearing liners process is as follows:
  - deposit the powdered coating material onto sheet steel;
  - heat,stretch and re-heat the coated sheets;
  - form the liners

  In a particular plant, the machinery was antiquated resulting in poor control of the furnace temperature and the stretching percentage.

  Production data was collected to investigate whether better control of these parameters would give improved bond strength of the coating).

Tutorial

- **Session TS02: Multiple Regression**
- **Python Environment**

  A self-guided tutorial has been created as a Colab notebook with pre-designed Python code and notes. For this tutorial, follow the instructions in the notes, upload data files and run the code. No modification of code is required. Interpret the results in accordance with the Technical session.

- **Tutorial Task**

  Fit a regression model in engineering units and coded factors.

  1. Read the tutorial data into a Pandas dataframe.
  2. Use multiple regression to fit a 1st order response surface in engineering units.**(for this task, use notebook labelled Tutorial02_part1)**
  3. Code the factors and fit a 1st order response surface using the coded factors.**(for this task, use notebook labelled Tutorial02_part2)**

Exercise

- **Session TS02: Multiple Regression**
- **Objectives**

  Fit a 1st order response surface for predicting overall efficency of a chemical plant.

- **Engineering Scenario**

  The data set stack loss, known as Brownlee's Stack Loss Plant Data is available in the public domain.

  The data-set is obtained from 21 days of operation of a plant for the oxidation of ammonia (NH_3) to nitric acid (HNO_3). The nitric oxides produced are absorbed in a counter current absorption tower".

  (Brownlee, cited by Dodge, slightly reformatted by MM.)

  Stack loss (the dependent variable) is 10 times the percentage of the ingoing ammonia to the plant that escapes from the absorption column unabsorbed; that is, an (inverse) measure of the overall efficiency of the plant.

Exercise

- **Session TS02: Multiple Regression**
- **Python Environment**

  The exercise has been created as a Colab notebook with notes. For this exercise, follow the instructions in the notes, and create your own code using the previous tutorial as a guide. Interpret the results in accordance with the Technical session.
- **Exercise Task**
  1. Read the exercise data into a Pandas dataframe.
  2. Transform the input variables to coded units [-1, 1].
  3. Use multiple regression to derive a response surface equation with stack loss as the y-variable and coded x-variables as predictors and examine the individual regression coefficients.
  4. Calculate the value for PRESS RMSE and compare it with the mean value of 'Stack loss' in the data set. Does the equation appear to be a useful predictor of stack loss?
  5. Analyse the model by looking at regression diagnostics and identify if there any outliers present.