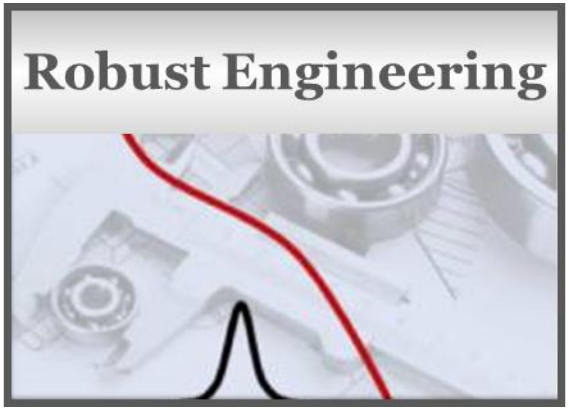
Module: Robust Engineering

Design of Experiments & Response Surface Modelling



Session 3: Model Selection

In This Session We Will …

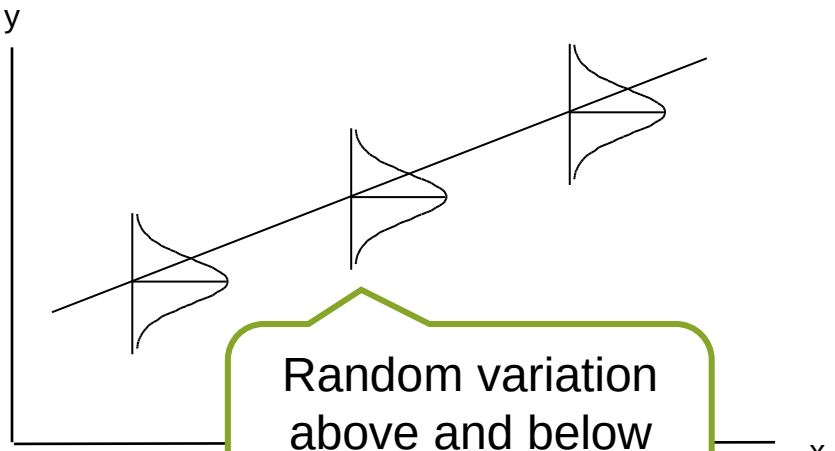
- Describe the statistical model underpinning regression analysis.
- Explain how to calculate the standard error of a regression coefficient.
- Apply t tests to the individual regression coefficients.
- Show that the standard errors may depend on which x-variables are in the fitted equation,
  - and that this may have a serious effect on our conclusions.
- Define a Variance Inflation Factor as a measure of the potential change in the standard error.

The Statistical Model Underpinning Regression Analysis

- A statistical model is a set of assumptions about the way our data set was generated.
- The model ***guides our analysis,***
  - we also use the data to ***check*** that the model is a reasonable one.
- When we use multiple regression to fit a 1st order surface with three x's, the underlying model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

This part defines a 1st order response surface



Random variation above and below the transfer function

ε represents random variation of the observations above and below the response surface

Standard Errors of Regression Coefficients

- In the statistical model;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

the β's are the unknown **parameters.**

Our regression coefficients b_0, b_1, etc are **estimates** of β_0, β_1, etc

- Because the y-values have a random component, so do the estimated regression coefficients.

The **standard error** of the coefficient is the 'sigma' for this random variation

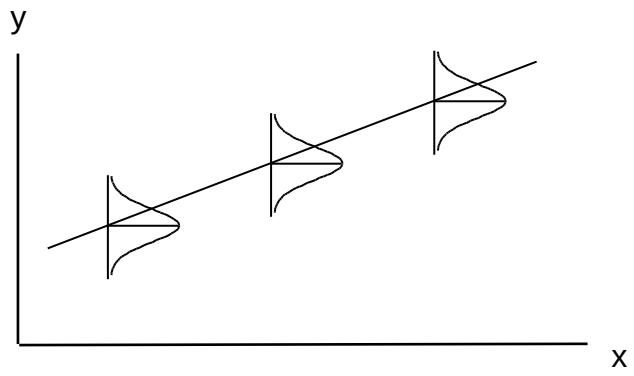
Standard Errors of Regression Coefficients (Cont.)

- Our statistical model has two parts, a part that is random and a part that depends on the x-variables.
- The formula for a standard error (SE) also has two parts,
  - but they are multiplied together, not added.
- For a typical regression coefficient b, we can write the SE of b as:

A number that depends only on the levels of the x's

$$k_b \cdot \sigma$$

The SD of the random part of the response



Calculating k_b

- k_b is calculated from the model matrix M and is the **square root** of the diagonal element of $(M'M)^{-1}$.

Matrix M once again

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

	b_0	b_1	b_2	b_3
b_0	**0.052**	0.033	0.007	-0.038
b_1	0.033	**0.805**	0.071	-0.927
b_2	0.007	0.071	**0.120**	-0.074
b_3	-0.038	-0.927	-0.074	**1.25**

Coefficient	k_b
b_1	0.897
b_2	0.346
b_3	1.118

Intercept	x_1	x_2	x_3
1	-0.45	0.60	-0.05
1	-0.95	0.80	-1.00
1	-0.40	1.00	-0.16
1	-0.40	0.00	-0.26
1	0.95	-0.80	0.34
1	-0.25	-0.40	-0.31
1	0.30	0.20	0.28
1	-0.80	-0.60	-0.45
1	-0.15	-0.80	-0.18
1	0.75	-0.60	1.00
1	-0.05	0.80	0.21
1	0.95	0.20	0.89
1	0.90	-1.00	0.81
1	-1.00	0.40	-0.44
1	0.40	0.20	0.21
1	-0.70	-0.80	-0.61
1	-0.25	0.40	-0.18
1	0.25	-0.80	-0.03
1	-0.80	-0.60	-0.66
1	1.00	1.00	0.63
1	-0.45	0.60	-0.05

$(M'M)^{-1}$ is a symmetric matrix with one row and one column for each regression coefficient

Matrix M for the sintering data based on coded values of x -variables

Estimating σ And Calculating The Standard Errors

- We **estimate σ** by s, the RMS residual, which is found from the ANOVA table.

Reminder: $Residual\ DF = n - p - 1$, where p is the no. of x-variables

Source of variation	Sum of squares	Degrees of freedom
Regression (RegSS)	891.30	3
Residual (ResSS)	75.28	16
Total (SSy)	966.58	19

From TS2

$$\text{RMS residual} = \sqrt{\frac{\text{ResSS}}{\text{ResDF}}} = \sqrt{\frac{75.28}{16}} = 2.17 \text{ N}$$

We can now test the significance of individual regression coefficients

- Multiply k_b by our estimate of σ to calculate **standard errors.**

Coefficient	k_b	SE
b_1	0.897	1.95
b_2	0.346	0.75
b_3	1.118	2.43

Significance Tests On The Regression Coefficients

- Why should we test the coefficients?
- From other areas of statistics we know that **apparent** effects may be purely the result of random variation.

> For example, two sets of data will usually have different **sample means**, but the **distribution means** may be the same

- In the sintering example, the apparent effect of changing Temp1 from 849 to 889 (–1 to +1 on the coded scale) is:

$$2 \times \text{coded regression coefficient for } x_1$$

$$= \ 2 \times 7.54 = 15.1 \text{ N}$$

  - but we don't want to spend money on a new temperature control device unless we have evidence that x_1 has a real effect on bond strength.

Significance Tests On The Regression Coefficients (Cont.)

- For a general term in the equation, the formal procedure is as follows:
  - let β denote the unknown regression coefficient.
  - set up a null hypothesis H_0: $\beta = 0$.
  - let b denote our estimate of β i.e. our fitted coefficient .
  - calculate the test statistic.

$$T = \frac{b}{\text{SE of } b}$$

This is a ‘t test’

- Calculate a p-value, using a central t distribution with degrees of freedom $\nu = ResDF$ (Residual Degrees of Freedom).
- If the p-value is small, reject H_0, i.e. assume that this variable has a real effect on the response.

Example: Calculating The p-value For x_3

- In the sintering data, Residual degrees of freedom (ResDF) = 16 so we use the t_{16} distribution.

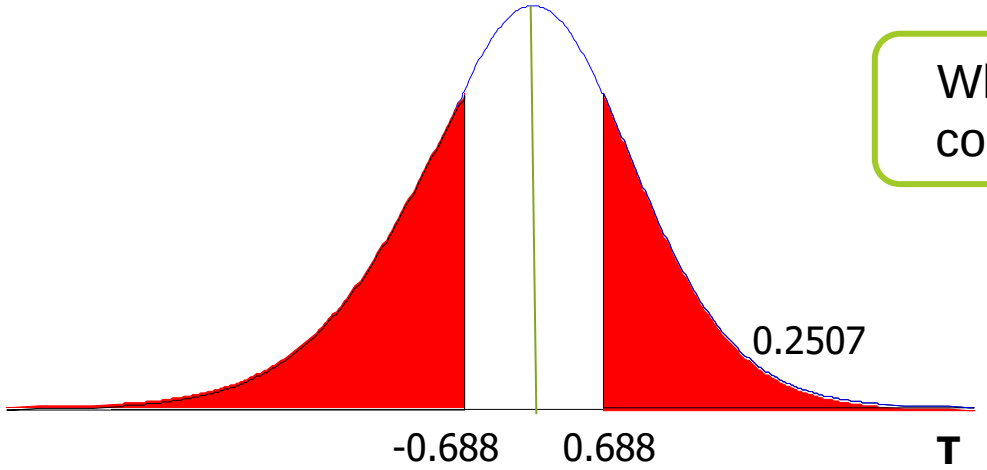
For x_3, T = $\frac{1.67}{2.43}$ = 0.688

Remember (session TS2)

$y = 90.84 + 7.54x_1 + 6.15x_2 + 1.67x_3$

A 'large' T value suggests a real effect

… but how large is large?



When testing regression coefficients the convention is to use a two-sided p-value

... so p = 0.2507×2 = 0.501

A 'small' p-value suggests a real effect

The p-values For Our Data

	Coefficient	SE	t-ratio	p-value
Intercept	90.84			
x_1	7.54	1.95	3.87	0.001
x_2	6.15	0.75	8.20	0.000
x_3	1.67	2.43	0.69	0.501

We aren't usually interested in testing the intercept

Not actually 0, but very small

Choosing The Threshold For Significance

- The conventional threshold for a ‘small’ p-value is 0.05 (5%).
- Experience with industrial experiments, backed up by some theory, suggests that the 5% criterion is much too strict.
  - some authors suggest a threshold of 15% or even higher.

We will use 15%

Even with this higher threshold, b_3 is ‘not significant’

- By this test b_3 is not significantly different from 0.
  - we have an ***inconclusive*** result.
  - we have not shown that $\beta_3 = 0$ or even that β_3 is small in absolute terms.
  - we have been ***unable to detect an effect from x_3***, but this may be because our test has low power (low sensitivity).

Re-fitting The Regression Equation

- Since we have been unable to detect any effect from x_3, we can try ***re-fitting*** the regression equation after omitting this term
  - i.e. use least squares to choose a new equation of the form

$$y = b_0 + b_1 x_1 + b_2 x_2$$

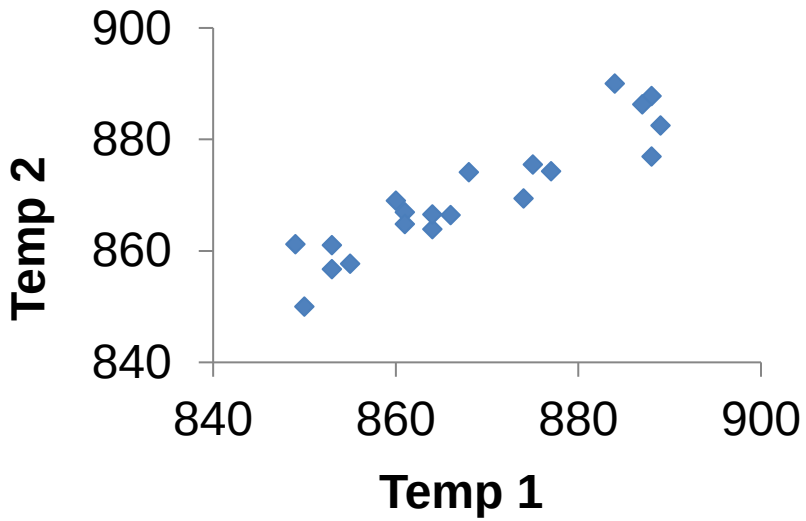
Original equation	Coefficient	SE	t-ratio	p-value
Intercept	90.84			
x_1	7.54	1.95	3.87	0.001
x_2	6.15	0.75	8.20	0.000
x_3	1.67	2.43	0.69	0.502
New equation				
Intercept	90.89			
x_1	8.78	0.73	12.09	0.000
x_2	6.25	0.72	8.62	0.000

The SE for b_1 has changed a lot

… and the t-ratio is much bigger

Why Such A Big Change In The Standard Error of b_2?

- In multiple regression, the standard error for one x-variable ***depends on which other variables are in the equation***, unless the variables are orthogonal (uncorrelated).
- In this case x_1 (Temp 1) and x_3 (Temp 2) are highly correlated because heating and re-heating occurred on the same day.



Temp 1 and Temp 2 are both affected by ambient conditions in the plant

The Effect of 'Unstable' Standard Errors

- Changes in the SE may have a serious effect on our conclusions.
- To see this, suppose that we had omitted Temp 1 from our first analysis:

	Coefficient	SE	t-ratio	p-value
Intercept	90.53			
x_2	5.49	0.99	5.57	0.000
x_3	10.35	1.24	8.36	0.000

x_3 now looks very significant

- If the correlations between predictors (x-variables) are small, the standard errors will not change much if other variables are included or omitted.

In regression software, this idea is operationalized by calculating Variance Inflation Factors

Variance Inflation Factors (VIFs)

- For variable x_j the **variance** of the regression coefficient b_j is the ***square of the SE***, and the VIF is defined as:

$$\left\{\frac{\text{SE of } b_j \text{ if the full model is fitted}}{\text{SE of } b_j \text{ if all other x's are removed}}\right\}^2$$

i.e. we just fit the intercept and x_j

Variable	VIF
x_1	7.1
x_2	1.1
x_3	7.0

All VIFs = 1 if the variables are uncorrelated

As a rough guide, we would like to have all VIFs < 2, so this is not a good set of data for regression analysis

Conclusions From The Sintering Example

- The equation as a whole gave acceptably small prediction errors.
  - this is promising – if they can control the x-variables better they will also be able to control bond strength.
- Even using a ‘relaxed’ 15% criterion for statistical significance, the analysis failed to detect the effect of ‘Temp 2’, the temperature at the re-heating stage of the process.
  - this might be due to low power caused by correlation between the x’s,
  - in fact a designed experiment run later showed that all 3 variables do affect the bond strength.

In This Session We Have …

- Described the statistical model underpinning regression analysis.
- Explained how to calculate the standard error of a regression coefficient.
- Applied t tests to the individual regression coefficients.
- Shown that the standard errors depend on which x-variables are in the fitted equation, unless they are uncorrelated.
  - and that this may have a serious effect on our conclusions.
- Defined a Variance Inflation Factor as a measure of the potential change in the standard error.

Session 3: Model Selection

Tutorial and Exercise

Tutorial

- **Session TS03: Model Selection**
- **Objectives**

  Carry out residual analysis and check if the regression coefficients are significant and have a real effect.
- **Engineering Scenario**

  Tutorial 03 is a continuation of Tutorial 02.

  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 TS03: Model Selection**
- **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**

  1. Fit a regression model and carry out residual analysis.
  2. Read the tutorial data into a Pandas dataframe.
  3. Code the factors and fit a 1st order response surface using the coded factors.
  4. Generate and analyse residual plots.
  5. Evaluate model coefficients.

Exercise

- **Session TS03: Model Selection**
- **Objectives**

  Fit an improved 1st order response surface for predicting overall efficency of a chemical plant.
- **Engineering Scenario**

  Exercise 03 is a continuation of Exercise 02.

  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 TS03: Model Selection**
- **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.
- **Tutorial Task**
  1. Read the tutorial data into a Pandas dataframe and remove the two observations (4 and 21).
  2. Transform the input variables to coded units [-1, 1].
  3. Re-fit the multiple regression to the new data to derive a response surface equation with stack loss as the y-variable and coded x-variables as predictors. Discuss the improvement in the model quality when compared to the previous exercise by checking the diagnostics (PRESS RMSE).
  4. Improve the model by removing insignificant terms. NB. Check VIFs.
  5. Produce a residual plot (use deleted residuals) and a Normal plot of the residuals. Discuss the interpretation of these plots.