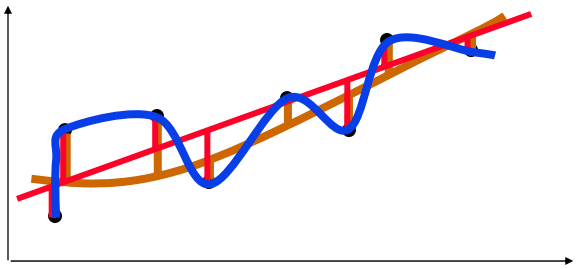
Regression Analysis

Simple Regression
Multivariate Regression
Stepwise Regression
Replication and Prediction Error

Regression Analysis



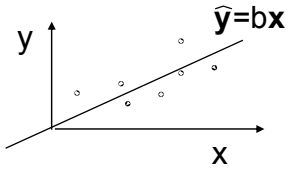
- In general, we "fit" a model by minimizing a metric that represents the error.

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

- The sum of squares gives closed form solutions and minimum variance for linear models.

The Simplest Regression Model

Line through the origin:



$$y_u=\beta x_u+\varepsilon_u \qquad u=1,2,...,n \qquad \varepsilon_u \sim N(0,\sigma_R^2)$$

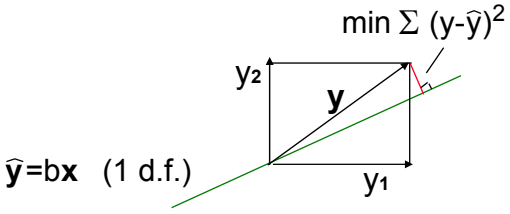
$$\min S = \min \sum_{u=1}^{n} (y_u - \beta x_u)^2: \quad \text{estimate of } \sigma_R^2$$

$$\hat{y}=bx \qquad \eta_u=\beta x_u$$

b: estimate of β

$\hat{y}$: estimate of η_u, the true value of the model.

Using the Normal Equation



Using the Normal Equation (cont)

Choose b so that the residual vector is perpendicular to the model vector...

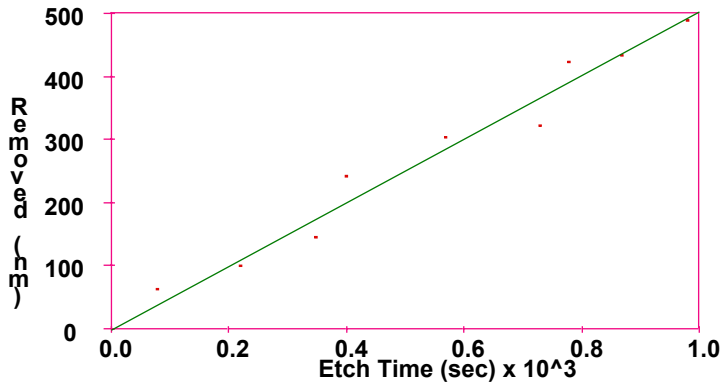
$$\sum (y-\hat{y}) \cdot x = 0 \Rightarrow \sum (y - bx) \, x = 0 \Rightarrow$$

$$b = \frac{\Sigma xy}{\Sigma x^2} \text{ (est. of } \beta) \quad s^2 = \frac{S_R}{n-1} \text{ (est. of } \sigma_R^2)$$

$$V(b) = \frac{s^2}{\Sigma x^2} \quad 67\% \text{ conf: } b \pm \sqrt{\frac{s^2}{\Sigma x^2}}$$

$$\text{Signif. test:} \quad t = \frac{b - \beta^*}{\sqrt{\frac{s^2}{\Sigma x^2}}} \sim t_{n-1}$$

Etch time vs removed material: y = bx



Data File:regression Dependent Variable:Removed (nm)

Variable Name	Coefficient	Std. Err. Estimate	t Statistic	Prob > t
Etch Time (sec)	5.0098e-1	1.6199e-2	3.0927e+1	1.33e-8

Model Validation through ANOVA

The idea is to decompose the sum of squares into orthogonal components.

Assuming that there is no dependence:

$$H_0: \beta^* = 0$$

$$\underset{\text{n}}{\underset{\text{total}}{\sum y_u^2}} = \underset{\text{p}}{\underset{\text{model}}{\sum \hat{y}_u^2}} + \underset{\text{n-p}}{\underset{\text{residual}}{\sum (y_u - \hat{y}_u)^2}}$$

Model Validation through ANOVA (cont)

Assuming a specific model:

$$H_0: \beta^* = b$$

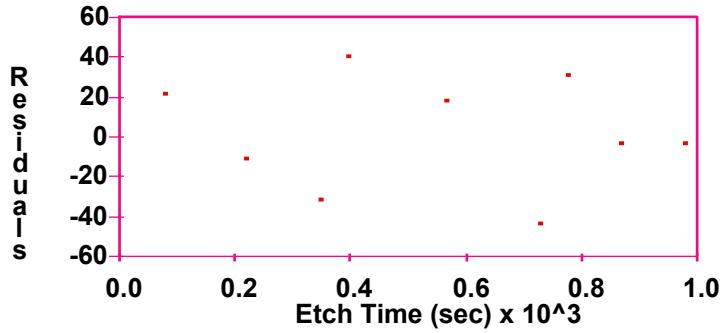
$$\underset{\text{n}}{\underset{\text{total}}{\sum (y_u - \beta^* x_u)^2}} = \underset{\text{p}}{\underset{\text{model}}{\sum (\hat{y}_u - \beta^* x_u)^2}} + \underset{\text{n-p}}{\underset{\text{residual}}{\sum (y_u - \hat{y}_u)^2}}$$

The ANOVA table will answer the question:

Is there a relationship between x *and* y*?*

ANOVA table and Residual Plot

Source	Sum of Squares	Deg. of Freedom	Mean Squares	F-Ratio	Prob>F
Model	1.8293e+5	1	1.8293e+5	1.9801e+2	2.17e-6
Error	6.4669e+3	7	9.2385e+2		
Total	1.8939e+5	8			



A More Complex Regression Equation

actual estimated

$$\eta = \alpha + \beta (x - \bar{x}) \qquad \hat{y} = a + b (x - \bar{x})$$

$$y_i \sim N(\eta_i, \sigma^2)$$

Minimize $R = \Sigma(y_i - \hat{y}_i)^2$ to estimate α and β

$$a = \bar{y} \qquad b = \frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

Are a and b good estimators of α and β?

$$E[a] = \alpha \qquad E[b] = \frac{\Sigma(x_i - \bar{x})E[y_i]}{\Sigma(x_i - \bar{x})^2} = \beta$$

Variance Estimation:

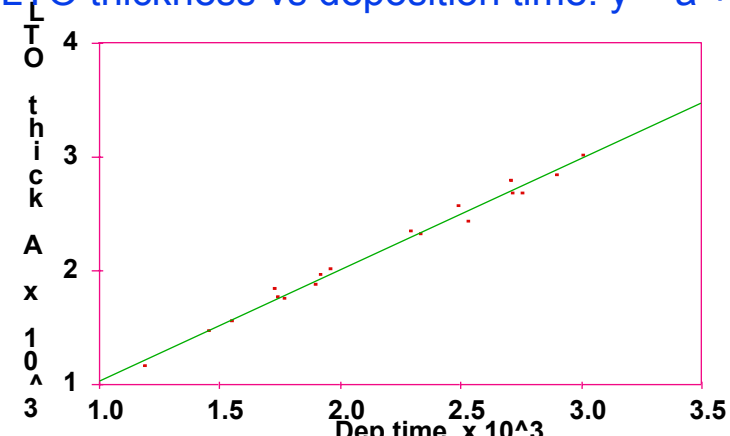
Note that all variability comes from y_i!

$$V[a] = V\left[\frac{\Sigma y_i}{k}\right] = \frac{1}{k^2}\Sigma V[y_i] = \frac{\sigma^2}{k}$$

$$V[b] = V\left[\frac{\Sigma (x_i-\bar{x})y_i}{\Sigma (x_i-\bar{x})^2}\right] = \frac{\sigma^2}{\Sigma (x_i-\bar{x})^2}$$

min var. thanks to least squares!

LTO thickness vs deposition time: y = a + bx

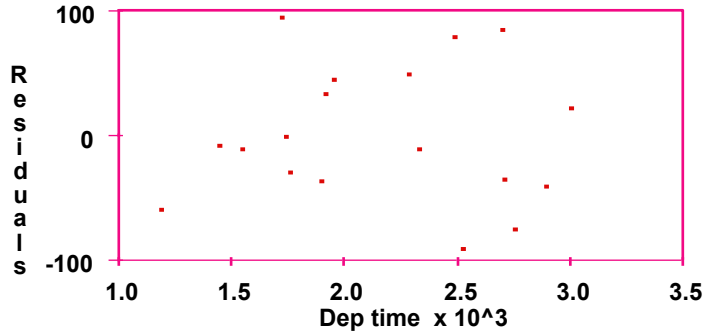


Data File: regression Dependent VariableLTO thick A

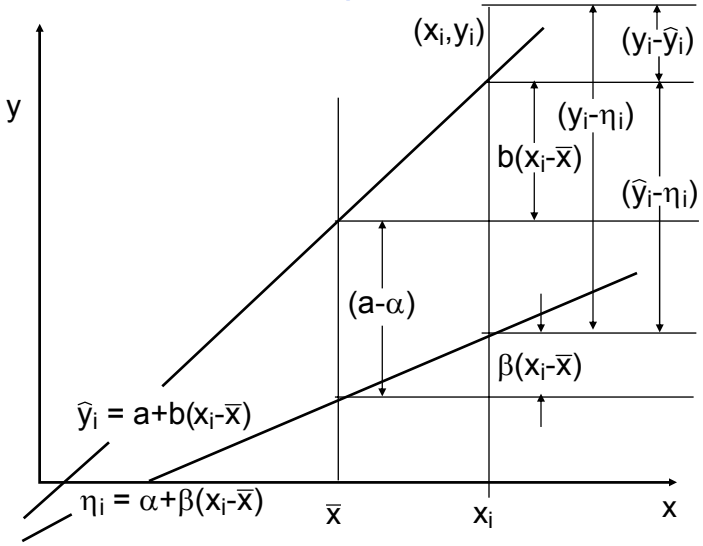
Variable Name	Coefficient	Std. Err. Estimate	t Statistic	Prob > t
Constant	6.0352e+1	5.6058e+1	1.0766e+0	2.98e-1
Dep time	9.7456e-1	2.5155e-2	3.8743e+1	3.02e-17

Anova table and Residual Plot

Source	Sum of Squares	Deg. of Freedom	Mean Squares	F-Ratio	Prob>F
Model	4.7725e+6	1	4.7725e+6	1.5010e+3	3.02e-17
Error	5.0872e+4	16	3.1795e+3		
Total	4.8233e+6	17			



ANOVA Representation



Note differences between "true" and "estimated" model.

ANOVA Representation (cont)

$$(y_i-\eta_i) = (a-\alpha) + (b-\beta)(x_i-\bar{x}) + (y_i-\hat{y}_i)$$

$$\Sigma(y_i-\eta_i)^2 = k(a-\alpha)^2 + (b-\beta)^2\Sigma(x_i-\bar{x}) + \Sigma(y_i-\hat{y}_i)^2$$

$$(k) \qquad (1) \qquad (1) \qquad (k-2)$$

$$\sim\sigma^2\chi^2(k) \qquad \sim\sigma^2\chi^2(1) \qquad \sim\sigma^2\chi^2(1) \qquad \sim\sigma^2\chi^2(k-2)$$

In this way, the significance of the model can be analyzed in detail.

Confidence Limits of an Estimate

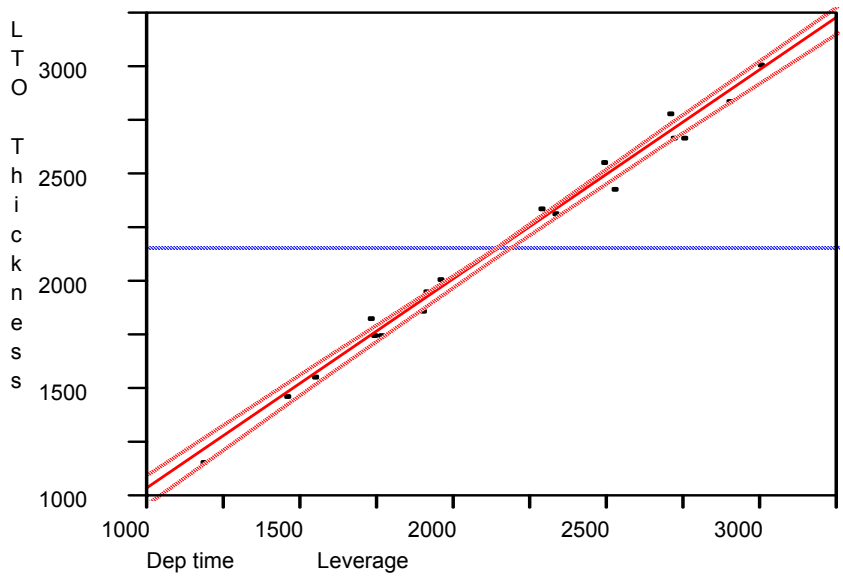
$$y_0 = \bar{y} + b(x_0 - \bar{x})$$

$$V(\hat{y}_0) = V(\bar{y}) + (x_0 - \bar{x})^2 V(b)$$

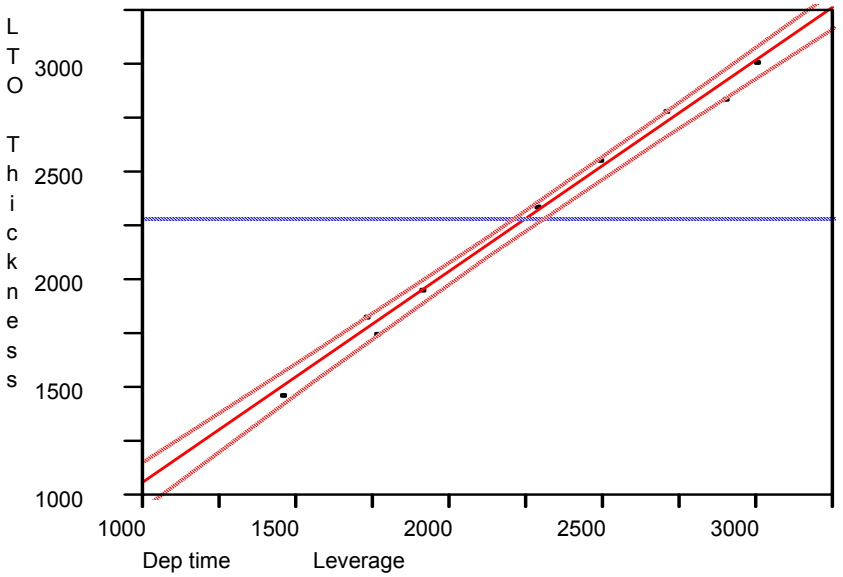
$$\hat{V}(\hat{y}_0) = \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}\right] s^2$$

prediction interval: $\hat{y}_0 +/- t_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{y}_0)}$

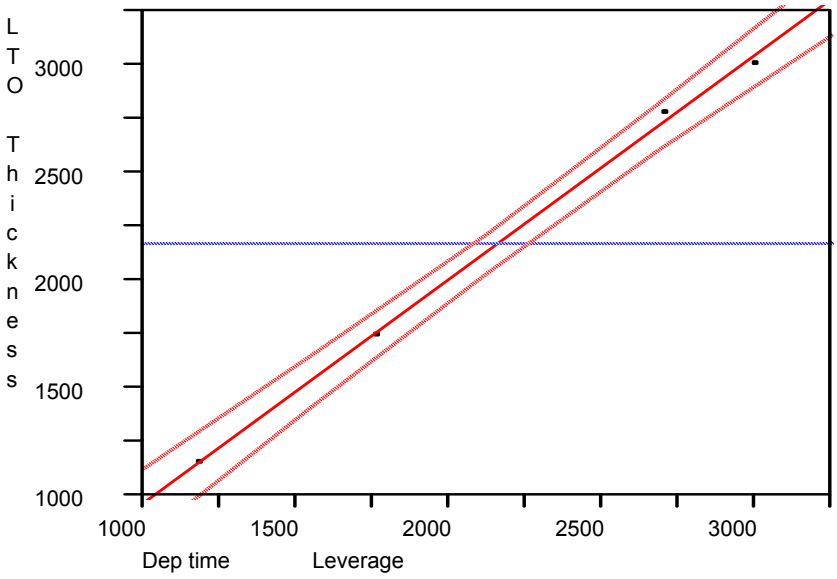
Confidence Interval of Prediction (all points)



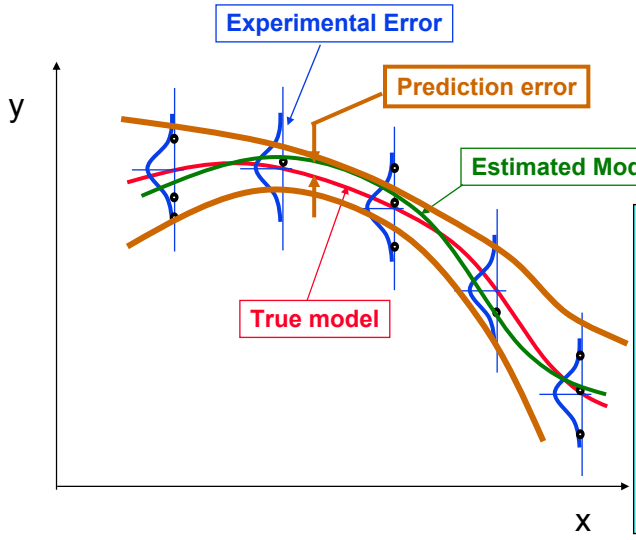
Confidence Interval of Prediction (half the points)



Confidence Interval of Prediction (1/4 of points)

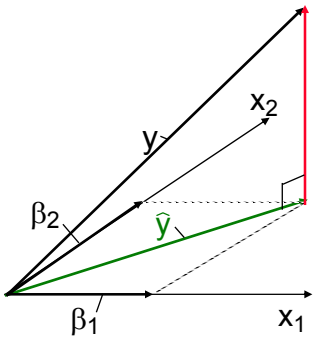


Prediction Error vs Experimental Error



- Experimental Error
  - Does not depend on location or sample size.
- Prediction Error
  - depends on location
  - gets smaller as sample size increases.

Multivariate Regression



$\eta = \beta_1 x_1 + \beta_2 x_2$

The Residual is $\perp$ to $\hat{y}, x_1, x_2$.

Coefficient Estimation: $\Sigma(y-\hat{y})x_1=0 \quad \Sigma(y-\hat{y})x_2=0$

$$\Sigma y x_1 - b_1 \Sigma x_1^2 - b_2 \Sigma x_1 x_2 = 0$$
$$\Sigma y x_2 - b_2 \Sigma x_2^2 - b_1 \Sigma x_1 x_2 = 0$$

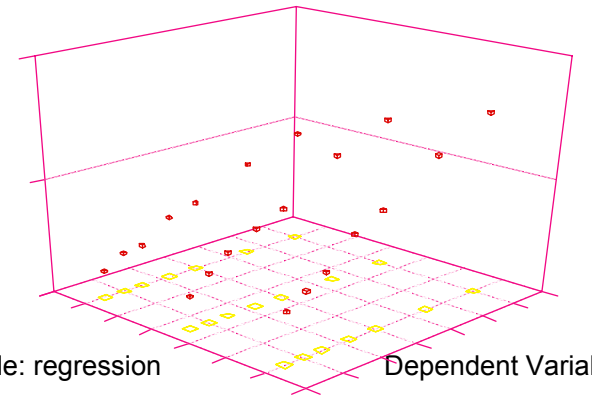
Variance Estimation:

$$s^2 = \frac{S_R}{n-p}$$

$$\hat{V}(b_1) = \frac{1}{1-\rho^2}\frac{s^2}{\Sigma x_1^2}$$

$$\hat{V}(b_2) = \frac{1}{1-\rho^2}\frac{s^2}{\Sigma x_2^2} \qquad \rho = \frac{-\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \Sigma x_2^2}}$$

Thickness vs time, temp: y = a + b1 x1 + b2 x2



Data File: regression Dependent Variable: tox nm

Variable Name	Coefficient	Std. Err. Estimate	t Statistic	Prob > t
Constant	-7.0363e+2	7.1769e+1	-9.8041e+0	1.10e-8
temp	7.1429e-1	6.9976e-2	1.0208e+1	7.49e-9
time min	8.6874e-1	3.8905e-2	2.2330e+1	3.72e-9

Anova table and Correlation of Estimates

Source	Sum of Squares	Deg. of Freedom	Mean Squares	F-Ratio	Prob>F
Model	2.5828e+4	2	1.2914e+4	3.0141e+2	1.45e-14
Error	7.7121e+2	18	4.2845e+1		
Total	2.6599e+4	20			

	Tox	Temp	Time
tox nm	1.000	0.410	0.896
temp	0.410	1.000	0.000
time min	0.896	0.000	1.000

Multiple Regression in General

$$[x_1\ x_2\ \cdots\ x_n][b] = [y] + [e]$$

$$\text{minimize} \left| Xb - y \right|^2 = |e|^2 = (y - Xb)^T (y - Xb)$$

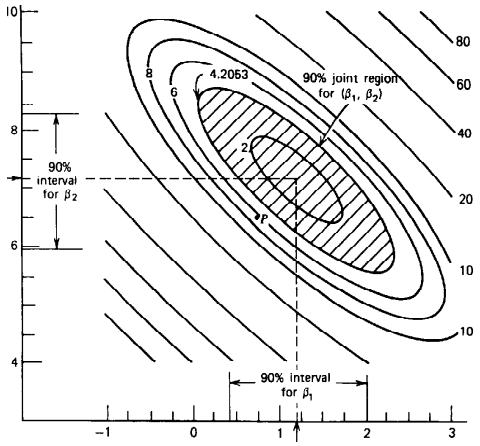
$$\text{or, min } -e^T Xb + e^T y \quad \text{which is equiv. to: } (y - Xb)^T Xb = 0$$

$$X^T Xb = X^T y$$

$$b = (X^T X)^{-1} X^T y \qquad V(b) = (X^T X)^{-1} \sigma^2$$

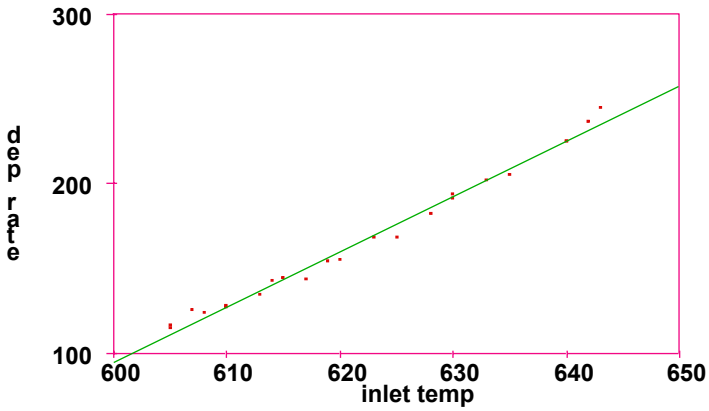
Joint Confidence Region for x_1 x_2

$$S = S_R\left[1 + \frac{p}{n-p} F_\alpha(p, n-p)\right]$$



$$\sum (\beta_1 - b_1)^2 \sum x_1^2 + 2(\beta_1 - b_1)(\beta_2 - b_2) \sum x_1 x_2 + (\beta_2 - b_2)^2 \sum x_2^2 = S - S_R$$

What if a "linear" model is not enough?

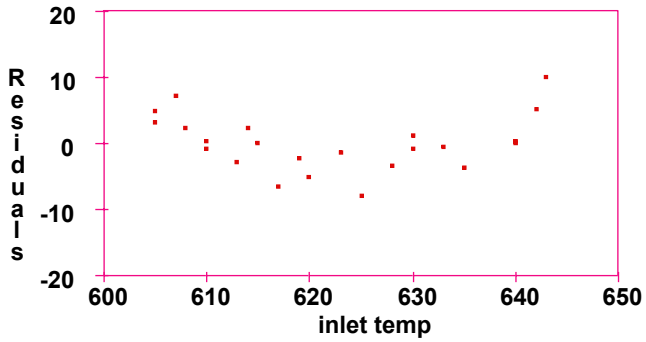


Data File: regression Dependent Variable: dep rate

Variable Name	Coefficient	Std. Err. Estimate	t Statistic	Prob > t
Constant	-1.8502e+3	4.6425e+1	-3.9853e+1	3.72e-9
inlet temp	3.2426e+0	7.4592e-2	4.3471e+1	3.72e-9

ANOVA table and Residual Plot

Source	Sum of Squares	Deg. of Freedom	Mean Squares	F-Ratio	Prob>F
Model	3.6490e+4	1	3.6490e+4	1.8897e+3	0.00e+0
Error	4.0550e+2	21	1.9309e+1		
Total	3.6895e+4	22			



Multiple Regression with Replication

$$S_E=\frac{1}{2}\sum (y_{i1}-y_{i2})^2 \qquad S_{LF}=S_R-S_E$$

$$\sum_i^k \sum_v^{n_i} (y_{iv}-\eta_i)^2 =$$

$$(a-\alpha)^2 \sum_i^k \eta_i + (b-\beta)^2 \sum_i^k \eta_i (x_i-\bar{x})^2 + \sum_i^k \eta_i (\bar{y}_{i.}-\hat{y}_i)^2 + \sum_i^k \sum_v^{n_i} (y_{iv}-\bar{y}_{i.})^2$$

$$1 \qquad\qquad 1 \qquad\qquad k-2 \qquad\qquad \sum_i^k \eta_i - k$$

$$\sum_i^k \sum_v^{n_i} (y_{iv}-\bar{y})^2 = \sum_i^k \sum_v^{n_i} (y_{iv}-\bar{y}_{i.})^2 + \sum_i^k \eta_i (\bar{y}_{i.}-\hat{y}_i)^2 + \sum_i^k \eta_i (\bar{y}-\hat{y}_i)^2$$

Pure Error vs. Lack of Fit Example

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	17	401.01171	23.5889	21.0360
Pure Error	4	4.48543	1.1214	**Prob > F**
Total Error	21	405.49714		0.0047

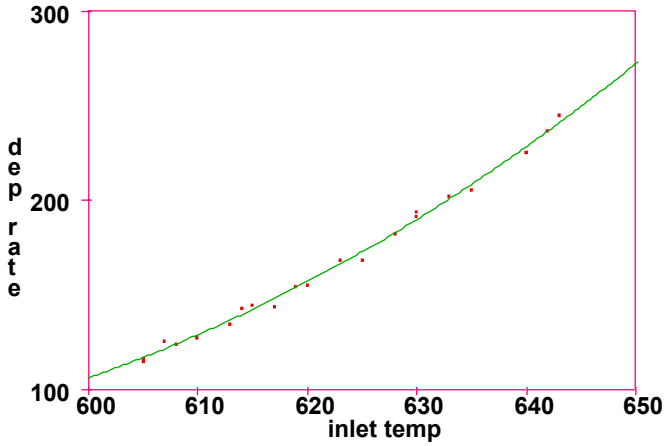
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1850.159	46.4247	-39.85	0.0000
inlet temp	3.242592	0.07459	43.47	0.0000

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
inlet temp	1	1	36489.550	999.9999	0.0000

Dep. rate vs temperature: $y = a + bx + cx^2$



Data File:

Variable Name	Coefficient	Std. Err. Estimate	t Statistic	Prob > t
Constant	8.3391e+3	1.7899e+3	4.6589e+0	1.35e-4
inlet temp	-2.9445e+1	5.7415e+0	-5.1284e+0	4.43e-5
inlet temp ^2	2.6205e-2	4.6028e-3	5.6933e+0	1.19e-5

Pure Error vs. Lack of Fit Example (cont)

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	16	150.24382	9.39024	8.3740
Pure Error	4	4.48543	1.12136	**Prob > F**
Total Error	20	154.72925		0.0264

Parameter Estimates

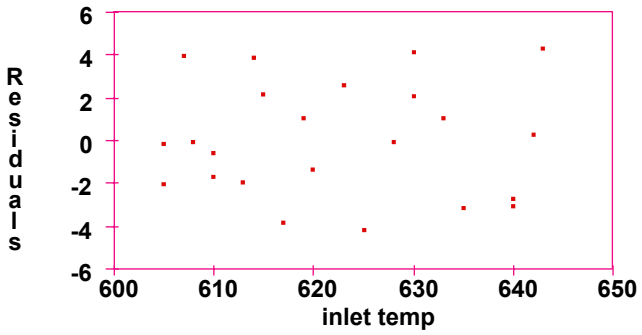
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8339.0507	1789.92	4.66	0.0002
inlet temp^1	-29.44466	5.74154	-5.13	0.0001
inlet temp^2	0.0262051	0.0046	5.69	0.0000

Effect Test

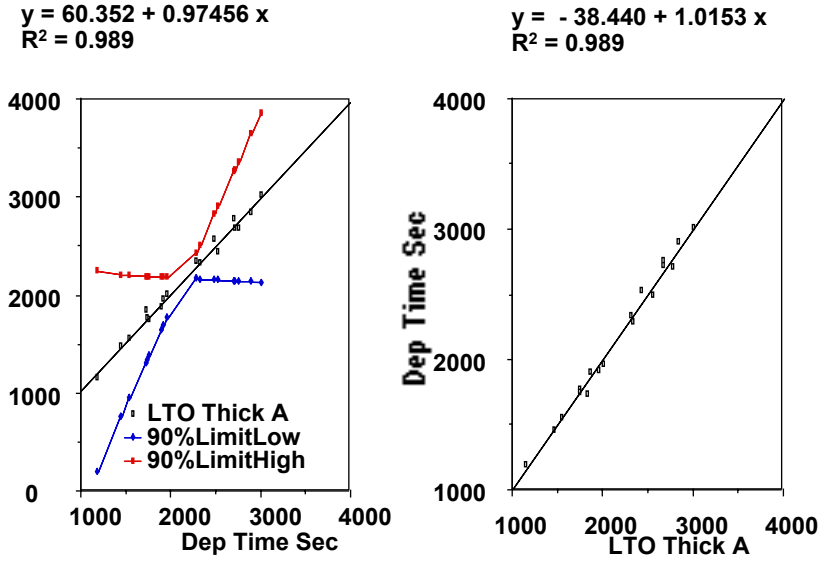
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Poly(inlet temp,2)	2	2	36740.318	999.9999	0.0000

ANOVA table and Residual Plot

Source	Sum of Squares	Deg. of Freedom	Mean Squares	F-Ratio	Prob>F
Model	3.6740e+4	2	1.8370e+4	2.3745e+3	0.00e+0
Error	1.5473e+2	20	7.7365e+0		
Total	3.6895e+4	22			

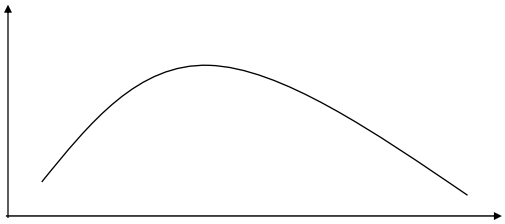


Use regression line to predict LTO thickness...



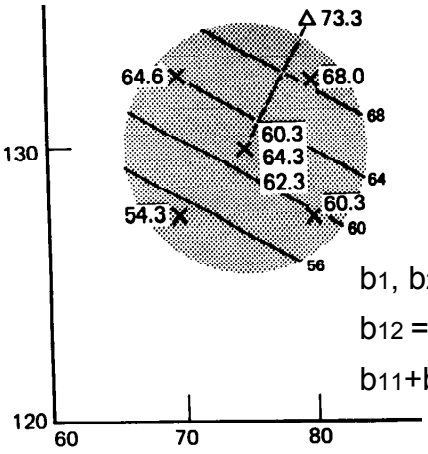
Response Surface Methodology

- Objectives:
- get a feel of I/O relationships
- find setting(s) that satisfy multiple constraints
- find settings that lead to optimum performance
- Observations:
- Function is nearly linear away from the peak
- Function is nearly quadratic at the peak



Building the planar model

A Factorial experiment with center points is enough to build and confirm a planar model.



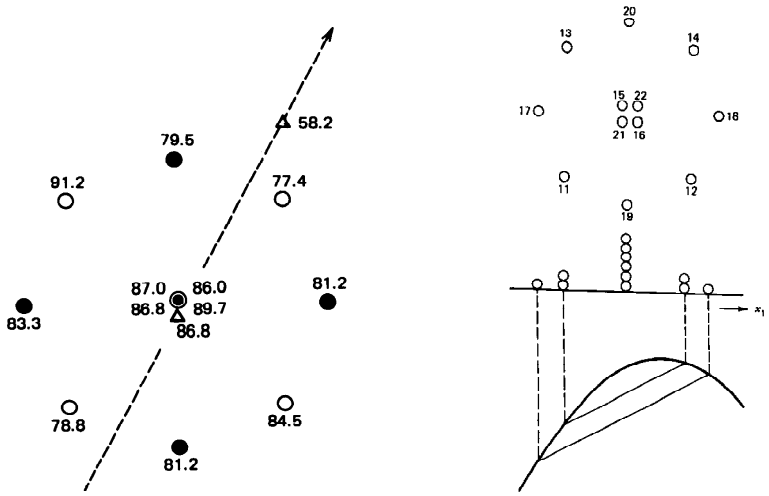
b_1, b_2,

b_{12} = -0.65 +/-0.75

$b_{11}+b_{22}=1/4\Sigma p+1/3\Sigma c$= -0.50 +/-1.15

Quadratic Model and Confirmation Run

Close to the peak, a quadratic model can be built and confirmed by an expanded two-phase experiment.



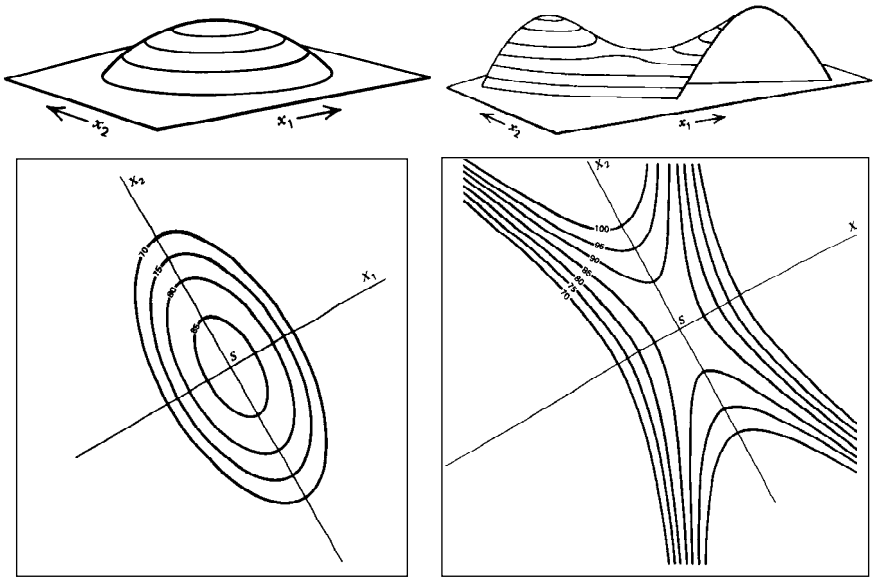
Response Surface Methodology

- RSM consists of creating models that lead to visual images of a response. The models are usually linear or quadratic in nature.
- Either expanded factorial experiments, or regression analysis can be used.
- All empirical models have a random prediction error. In RSM, the average variance of the model is:

$$\overline{V}(\hat{y}) = \frac{1}{n}\sum_{i=1}^{n} V(\hat{y}_i) = \frac{p\sigma^2}{n}$$

- where “p” is the number of model parameters and “n” is the number of experiments.

Response Surface Exploration



"Popular" RSM

- Use singe-stage Box-B or Box-W designs
- Use computer (simulated) experiments
- Rely on "goodness of fit" measures
- Automate model structure generation
- Problems?