UNIVERSITY OF CALIFORNIA
College of Engineering
Department of Electrical Engineering
and Computer Sciences

C. SPANOS **Special Issues in Semiconductor Manufacturing** **EECS 290H**
Fall 2005

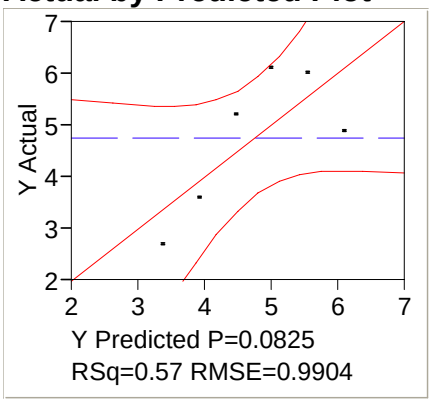
PROBLEM SET No. 6
Official Solutions

1. (a) Using the method of least squares fit a straight line to the following data. What are the least squares estimates of the slope and intercept of the line?

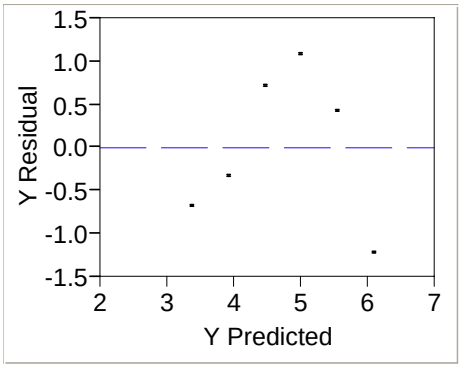
X	10	20	30	40	50	60
Y	2.7	3.6	5.2	6.1	6.0	4.9

I do the classical analysis fitting the model y = a + b(x-xbar)

Actual by Predicted Plot



Residual by Predicted Plot



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	5.2115714	5.21157	5.3133
Error	4	3.9234286	0.98086	Prob > F
C. Total	5	9.1350000		0.0825

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.75	0.404322	11.75	0.0003
X-Xbar	0.0545714	0.023675	2.31	0.0825

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
X-Xbar	1	1	5.2115714	5.3133	0.0825

(b) Calculate 99% confidence intervals for the slope and intercept.

The 99% interval can be found by looking at the t-statistics. For the intercept, the estimate of sigma is 0.922. This means that the estimated Intercept, is distributed according to the student-t distribution (with 4 degrees of freedom, since out of the 6 original degrees of freedom, 2 were consumed to estimate the

intercept and the slope of the model) as follows:

$$\frac{Int}{0.404} \sim t_4 \Rightarrow P\{4.75 - 0.404 \cdot t_{4,0.005} \le Int \le 4.75 + 0.404 \cdot t_{4,0.005}\} =$$

$$P\{4.75 - 0.404 \cdot 4.604 \le Int \le 4.75 + 0.404 \cdot 4.604\} = P\{2.89 \le Int \le 6.61\} = 0.99$$

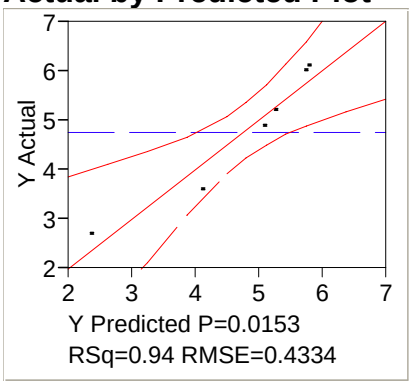
The slope can be analyzed in a very similar fashion:

$$\frac{Slope}{0.02367} \sim t_4 \Rightarrow P\{0.0546 - 0.02367 \cdot t_{4,0.005} \le Slope \le 0.0546 + 0.02367 \cdot t_{4,0.005}\} =$$

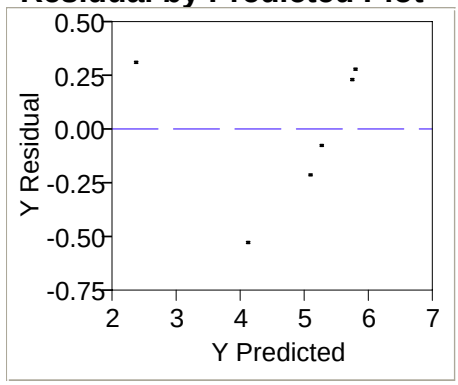
$$P\{0.0546 - 0.02367 \cdot 4.604 \le Slope \le 0.0546 + 0.02367 \cdot 4.604\} = P\{-0.05438 \le Slope \le 0.1636\} = 0.99$$

(c) Comment on the data and analysis, and carry out any further analysis you think is appropriate.
Clearly, this model is not fitting well, and the slope parameter is not significant event at the 5% level. Looking at the residual plot, I think that the model fit would be enhanced if I fit a quadratic function:

Actual by Predicted Plot



Residual by Predicted Plot



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	8.5715714	4.28579	22.8199
Error	3	0.5634286	0.18781	Prob > F
C. Total	5	9.1350000		0.0153

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.715	0.453391	8.19	0.0038
X	0.0545714	0.01036	5.27	0.0133
(X-35)*(X-35)	-0.003	0.000709	-4.23	0.0242

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
X	1	1	5.2115714	27.7492	0.0133
X*X	1	1	3.3600000	17.8905	0.0242

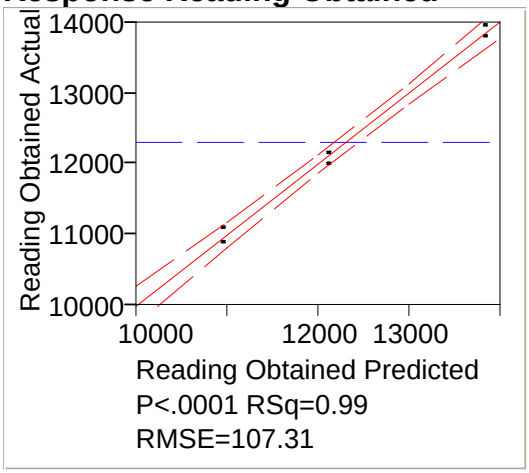
2. You have been asked to calibrate the "Inspector", an instrument used for in-situ measurement of photoresist thickness. You are given three reference wafers known to have exact resist thicknesses of 12,050Å, 13,580Å and 11,030Å. Using a purely randomized sequence, you have collected the following data from the new instrument:

Reference Used	Reading Obtained
12,050	12,000

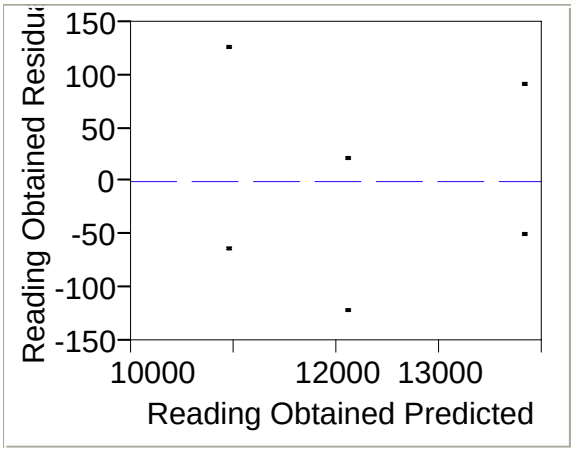
11,030	11,090
12,050	12,145
13,580	13,810
11,030	10,900
13,580	13,950

Do a complete regression analysis to answer the following: What is the best model that describes this instrument? (Consider straight lines only - is there any "lack of fit"?). Is a non-zero intercept necessary? Would it be OK if while using this instrument in the future you simply used the readings without any correction? Explain. Estimate the standard error of the instrument.

Response Reading Obtained



Residual by Predicted Plot



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	8493158.0	8493158	737.5281
Error	4	46062.8	11516	Prob > F
C. Total	5	8539220.8		<.0001

Lack Of Fit

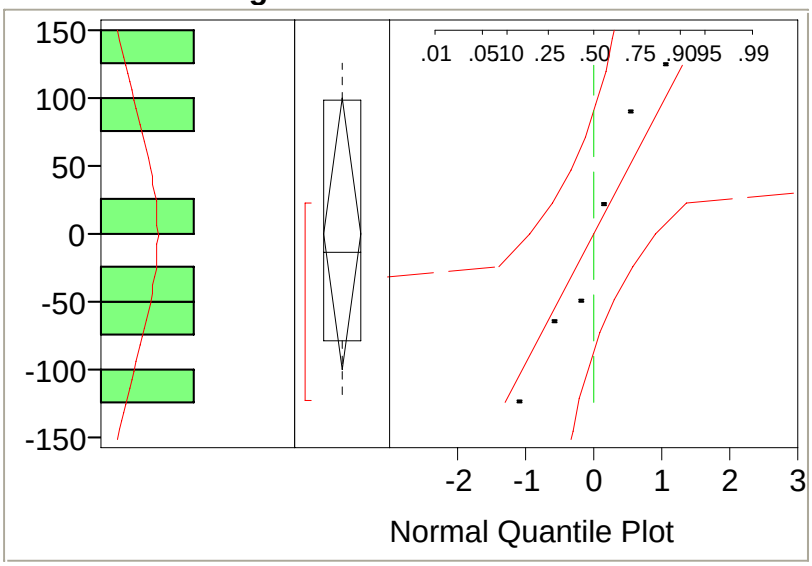
Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	1	7700.329	7700.3	0.6022
Pure Error	3	38362.500	12787.5	Prob > F
Total Error	4	46062.829		0.4943
				Max RSq
				0.9955

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1557.776	512.733	-3.04	0.0385
Reference Used	1.1353199	0.041805	27.16	<.0001

From this analysis I conclude that a straight line model is fine. There is no significant lack of fit component. Also, based on this model, the standard error of the instrument is 107.3A. I am not justified to ignore the intercept. If I used a straight line (x=y) to fit these data, (i.e. if I used the readings directly, without any correction factor), then the residuals would not be centered on zero, and the experimental error would be higher (about 180A as opposed to 107A):

Residual Reading Obtained



Normal(1.8e-12,95.9821)

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	Mu	0.00000	-100.727	100.7270
Dispersion	Sigma	95.98211	59.913	235.4071

3. An experimenter intended to perform a full 2^3 factorial design, but two values were lost as follows:

x_1	x_2	x_3	y
-1	-1	-1	3
+1	-1	-1	5
-1	+1	-1	4
+1	+1	-1	6
-1	-1	+1	3
+1	-1	+1	data missing
-1	+1	+1	data missing
+1	+1	+1	6

We now want to estimate the parameters in the model

$$\eta = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

With the data above, is this possible? Explain your answer. If your answer is "yes", set up the equation(s) that you would want to solve. If your answer is "no", what is the minimum amount of additional information that would be required?

It is possible to do the analysis, even with the missing data. Of course, we can no longer do a simple factorial analysis, because the symmetry (and the orthogonality) of the experimental design is destroyed. We can still do regression analysis, however. Note that the degrees of freedom are marginal (we are trying to extract 4 parameters out of 6 experiments).

Summary of Fit

RSquare	1
RSquare Adj	1
Root Mean Square Error	0
Mean of Response	4.5
Observations (or Sum Wgts)	6

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	9.5000000	3.16667	.
Error	2	0.0000000	0.00000	Prob > F
C. Total	5	9.5000000		.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.5	0	.	.
x1	1	0	.	.
x2	0.5	0	.	.
x3	-9.99e-16	0	.	.

Although there are two degrees of freedom for the error, its value is zero. This is a suspicious artifact and it might indicate that something is wrong with the data. The experiment (and the analysis), is fine though, but it does indicate that x3 has a negligible effect. Analysis can be redone without x3:

Summary of Fit

RSquare	1
RSquare Adj	1
Root Mean Square Error	0
Mean of Response	4.5
Observations (or Sum Wgts)	6

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9.5000000	4.75000	.
Error	3	0.0000000	0.00000	Prob > F
C. Total	5	9.5000000		.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.5	0	.	.
x1	1	0	.	.
x2	0.5	0	.	.

4. In regression modeling problems, the residual mean square is often used to estimate the natural variance of the experimental error σ^2. There are circumstances, however, in which it can produce (a) a serious underestimate of σ^2 or (b) a serious overestimate of σ^2. Can you give examples of such circumstances?

Serious underestimate: non-randomized sequence, non-true replications, model is overfitted.
Serious overestimate: when the model is inadequate, either because it is missing terms, or because the overall "shape" of the model is wrong.

5. Suppose data is collected on deposition rate (x_1) and uniformity (x_2) for a CVD process. The covariance matrix is:

$$\mathbf{\Sigma} = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix}$$

(a) What is the first principal component, and what percentage of the total variation does it account for?

For a 2×2 matrix, the eigenvalues are

$$\lambda_{\pm} = \frac{1}{2}\left[(a_{11} + a_{22}) \pm \sqrt{4\, a_{12}\, a_{21} + (a_{11} - a_{22})^2}\,\right],$$

which arises as the solutions of the characteristic equation

$$x^2 - x\,(a_{11} + a_{22}) + (a_{11}\, a_{22} - a_{12}\, a_{21}) = 0.$$

λ_1=124.0 the first eigenvector is [0.707, 0.707], and the total variance explained is 77.5%

λ_2 =36.0 the second eigenvector is [0.707, -0.707], and the total variance explained is 22.5%

(b) Repeat (a) if $\mathbf{\Sigma} = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}$.

λ_1=8024.37 the first eigenvector is [0.99847, 0.055303], and the variance explained is 99.31%

λ_2 =55.63 the second eigenvector is [0.055303, -0.99847], and the variance explained is 0.69%