Quiz 1. Thursday. in class

Ch 1 - Ch 3.

Describing Relationships *Among* Variables (Ch. 4)

y, $x_1, \ldots, x_p$.

Yifan Zhu

Iowa State University

Outline

Polynomial Regression

Multiple Regression

Polynomial Regression

- Simple linear regression: fit a line:

$$y_i \approx b_0 + b_1 x_i$$

p-1 predictors.

- Polynomial regression: fit a polynomial:

$$y_i \approx b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3 + \cdots + b_{p-1} x_i^{p-1}$$

  - The p coefficients $b_0, b_1, \ldots, b_{p-1}$ are estimated by minimizing the loss function below using the least squares principle:

$$S(b_0, \ldots, b_{p-1}) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i + \cdots + b_{p-1} x_i^{p-1}))^2$$

  - In practice, we make a computer find the coefficients for us. This class uses JMP. See https://www.stat.iastate.edu/statistical-software-jmp for JMP installation and JMP Help and Resource.

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \qquad X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{bmatrix} \quad n \times p.$$

$$b = (X^T X)^{-1} X^T y = \begin{bmatrix} b_0 \\ \vdots \\ b_{p-1} \end{bmatrix}.$$

Example: fly ash cylinders

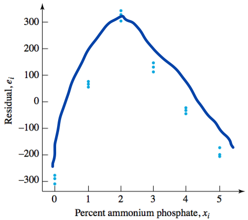
- A researcher studied the compressive strength of concrete-like fly ash cylinders. The cylinders were made with varying amounts of ammonium phosphate as an additive.
- We want to investigate the relationship between the amount ammonium phosphate added and compressive strength.

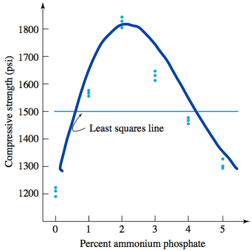
Additive Concentrations and Compressive Strengths for Fly Ash Cylinders

x, Ammonium Phosphate (%)	y, Compressive Strength (psi)	x, Ammonium Phosphate (%)	y, Compressive Strength (psi)
0	1221	3	1609
0	1207	3	1627
0	1187	3	1642
1	1555	4	1451
1	1562	4	1472
1	1575	4	1465
2	1827	5	1321
2	1839	5	1289
2	1802	5	1292

Simple linear regression fit: $\widehat{y}_i = 1498.4 - .6381x_i$

x	y	$\hat{y}$	$e = y - \hat{y}$	x	y	$\hat{y}$	$e = y - \hat{y}$
0	1221	1498.4	−277.4	3	1609	1496.5	112.5
0	1207	1498.4	−291.4	3	1627	1496.5	130.5
0	1187	1498.4	−311.4	3	1642	1496.5	145.5
1	1555	1497.8	57.2	4	1451	1495.8	−44.8
1	1562	1497.8	64.2	4	1472	1495.8	−23.8
1	1575	1497.8	77.2	4	1465	1495.8	−30.8
2	1827	1497.2	329.8	5	1321	1495.2	−174.2
2	1839	1497.2	341.8	5	1289	1495.2	−206.2
2	1802	1497.2	304.8	5	1292	1495.2	−203.2

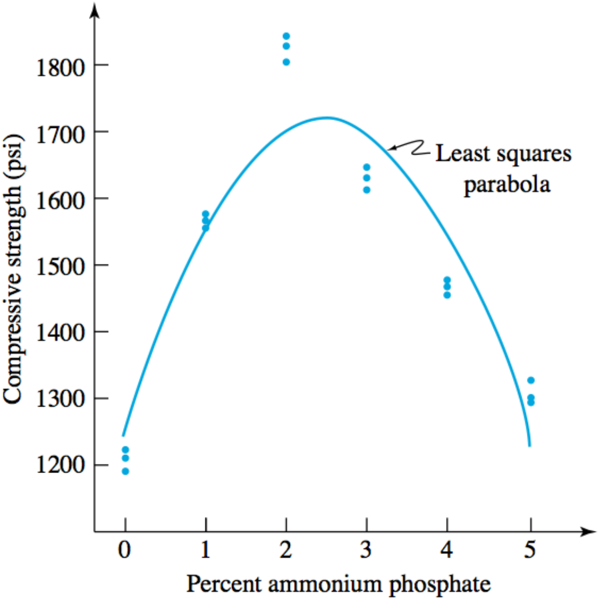




Quadratic fit: $\hat{y}_i = 1242.9 + 382.7x - 76.7x_i^2$

```
Regression Analysis

The regression equation is
y = 1243 + 383 x - 76.7 x**2

   Predictor          Coef       StDev          T        P
bo Constant        1242.89       42.98      28.92    0.000
b1 x                382.67       40.43       9.46    0.000
b2 x**2            -76.661       7.762      -9.88    0.000

S = 82.14         R-Sq = 86.7%       R-Sq(adj) = 84.9%

Analysis of Variance

Source            DF         SS          MS         F        P
Regression         2     658230      329115     48.78    0.000
Residual Error    15     101206        6747
Total             17     759437

Source        DF       Seq SS
x              1           21
x**2           1       658209
```

Quadratic fit: $\hat{y}_i = 1242.9 + 382.7x - 76.7x_i^2$



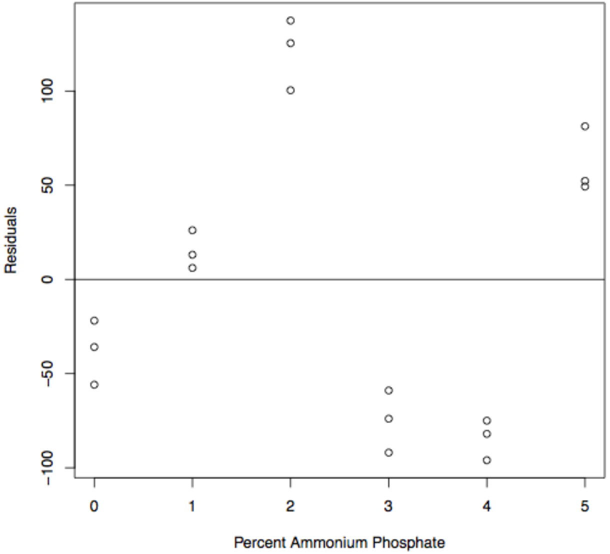
$R^2 = 86.7\%$

- The parabolic fit explained 86.7% of the variation in compressive strength.
- Note: for polynomial regression (and later, multiple regression) R^2 does not equal the squared correlation r_{xy}^2 between x and y. last time: linear regression $r_{xy}^2 = r_{y\hat{y}}^2$
- Instead, $\boxed{R^2 = r_{y\hat{y}}^2}$:

$$r_{y\hat{y}} = \frac{\sum(y_i - \overline{y})(\widehat{y}_i - \overline{\widehat{y}}_i)}{\sqrt{\sum(y_i - \overline{y})^2}\sqrt{\sum(\widehat{y}_i - \overline{\widehat{y}}_i)^2}}$$

$$R^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

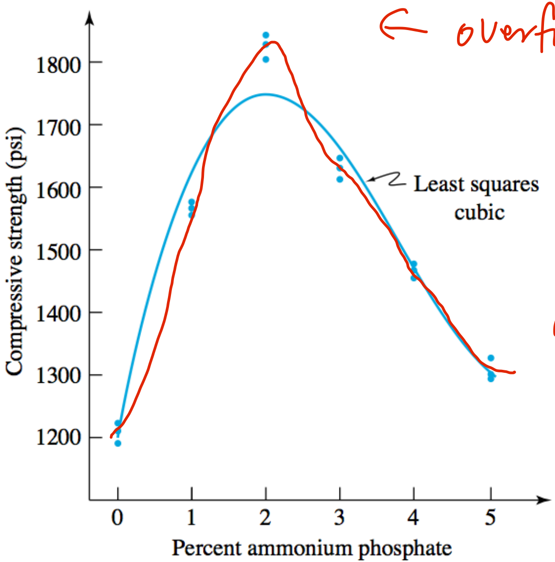
Residuals for the quadratic fit have less of a pattern than those of the linear fit.



Cubic fit: $\widehat{y}_i = 1188 + 633x - 214x^2 + 18.3x^3$

```
Regression Analysis


The regression equation is
y = 1188 + 633 x - 214 x**2 + 18.3 x**3

Predictor          Coef        StDev         T         P
Constant        1188.05        28.79     41.27     0.000
x                633.11        55.91     11.32     0.000
x**2            -213.77        27.79     -7.69     0.000
x**3             18.281        3.649      5.01     0.000

S = 50.88         R-Sq = 95.2%      R-Sq(adj) = 94.2%

Analysis of Variance

Source            DF          SS          MS          F        P
Regression         3      723197      241066      93.13    0.000
Residual Error    14       36240        2589
Total             17      759437
```

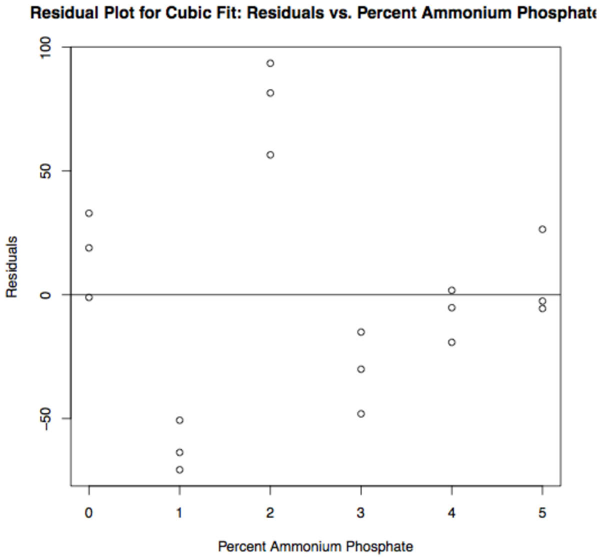
Cubic fit: $\widehat{y}_i = 1188 + 633x - 214x^2 2 + 18.3x^3$



← overfitting

6-degree polymial.

R^2 rose to 95.2%, and the residual plot improved.



Outline

Polynomial Regression

Multiple Regression

Multiple Regression

- **Multiple Regression**: regression on multiple variables:

$$y_i \approx b_0 + b_1 x_{i,1} + b_2 x_{i,2} + b_3 x_{i,3} + \cdots + b_{p-1} x_{i,p-1}$$

  - The p coefficients $b_0, b_1, \ldots, b_{p-1}$ are estimated by minimizing the loss function below using the least squares principle:

$$S(b_0, \ldots, b_p) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i,1} + \cdots + b_{p-1} x_{i,p-1}))^2$$

  - In practice, we make a computer find the coefficients for us. This class uses JMP.

Example: New York rivers data

- Nitrogen content is a measure of river pollution.

Variable	Definition
Y	Mean nitrogen concentration (mg/liter) based on samples taken at regular intervals during the spring, summer, and fall months
X_1	Agriculture: percentage of land area currently in agricultural use
X_2	Forest: percentage of forest land
X_3	Residential: percentage of land area in residential use
X_4	Commercial/Industrial: percentage of land area in either commercial or industrial use

- I will fit each of:

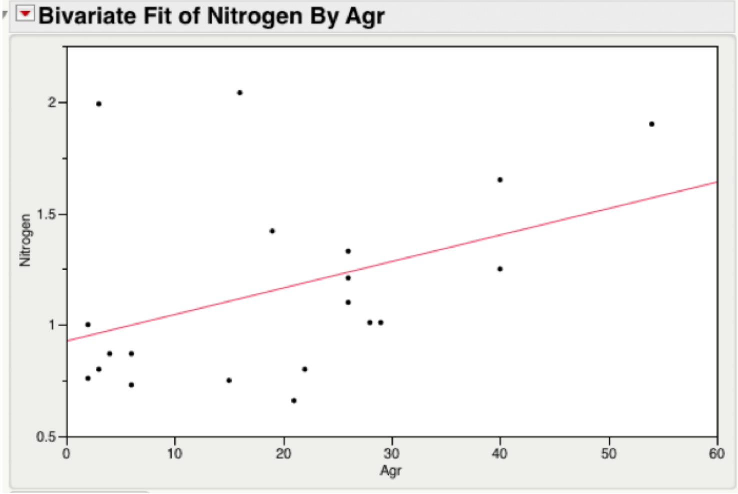
$$\left\{\begin{array}{l}\widehat{y}_i = b_0 + b_1 x_{i,1} \\ \widehat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + b_3 x_{i,3} + b_4 x_{i,4}\end{array}\right.$$

and evaluate fit quality.

Example: New York rivers data

Row	River	Y	X_1	X_2	X_3	X_4
1	Olean	1.10	26	63	1.2	0.29
2	Cassadaga	1.01	29	57	0.7	0.09
3	Oatka	1.90	54	26	1.8	0.58
4	Neversink	1.00	2	84	1.9	1.98
5	Hackensack	1.99	3	27	29.4	3.11
6	Wappinger	1.42	19	61	3.4	0.56
7	Fishkill	2.04	16	60	5.6	1.11
8	Honeoye	1.65	40	43	1.3	0.24
9	Susquehanna	1.01	28	62	1.1	0.15
10	Chenango	1.21	26	60	0.9	0.23
11	Tioughnioga	1.33	26	53	0.9	0.18
12	West Canada	0.75	15	75	0.7	0.16
13	East Canada	0.73	6	84	0.5	0.12
14	Saranac	0.80	3	81	0.8	0.35
15	Ausable	0.76	2	89	0.7	0.35
16	Black	0.87	6	82	0.5	0.15
17	Schoharie	0.80	22	70	0.9	0.22
18	Raquette	0.87	4	75	0.4	0.18
19	Oswegatchie	0.66	21	56	0.5	0.13
20	Cohocton	1.25	40	49	1.1	0.13

$\widehat{y}_i = b_0 + b_1 x_{i,1}$: pollution vs. agricultural land.



- It looks like the data could be roughly linear, although there are too few points to be sure.

$\widehat{y}_i = b_0 + b_1 x_{i,1}$: pollution vs. agricultural land.

Linear Fit

Linear Fit

Nitrogen = 0.9269285 + 0.0118851*Agr

Summary of Fit

RSquare	0.160762
RSquare Adj	0.114137
Root Mean Square Error	0.410975
Mean of Response	1.1575
Observations (or Sum Wgts)	20

Lack Of Fit

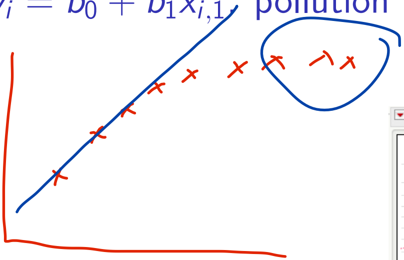
Analysis of Variance

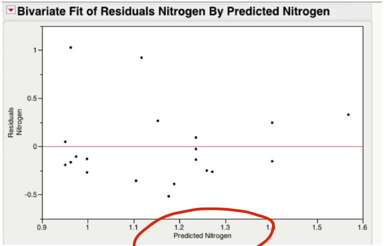
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.5823712	0.582371	3.4480
Error	18	3.0402038	0.168900	**Prob > F**
C. Total	19	3.6225750		0.0798

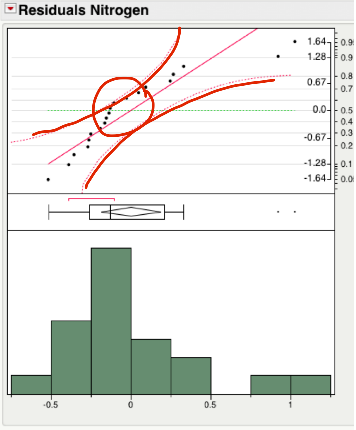
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.9269285	0.154478	6.00	<.0001*
Agr	0.0118851	0.006401	1.86	0.0798

$\widehat{y}_i = b_0 + b_1 x_{i,1}$: pollution vs. agricultural land.







$\hat{y}$

Conclusions: $\widehat{y}_i = b_0 + b_1 x_{i,1}$

- A low R^2 means the model isn't very useful for predicting the pollution of other New York rivers outside our dataset.
- However, the lack of a pattern in the residual plot shows that the model is valid.
- The residuals depart from a bell shape slightly, but not enough to interfere with statistical inference.

① confidence band

② tail violation.

$$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + b_3 x_{i,3} + b_4 x_{i,4}$$

Response Nitrogen

Summary of Fit

RSquare	0.709398
RSquare Adj	0.631904
Root Mean Square Error	0.264919
Mean of Response	1.1575
Observations (or Sum Wgts)	20

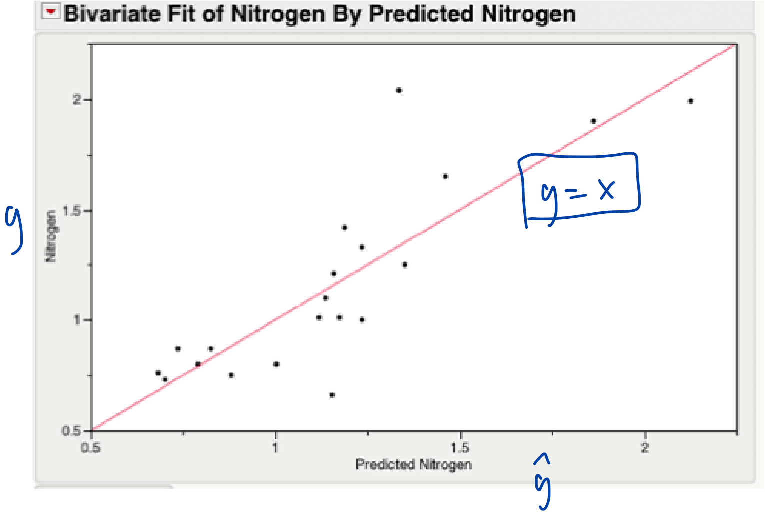
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	2.5698462	0.642462	9.1542
Error	15	1.0527288	0.070182	Prob > F
C. Total	19	3.6225750		0.0006*

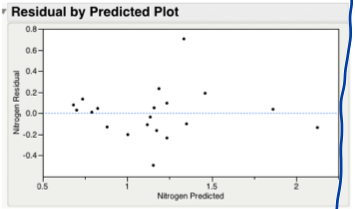
Parameter Estimates

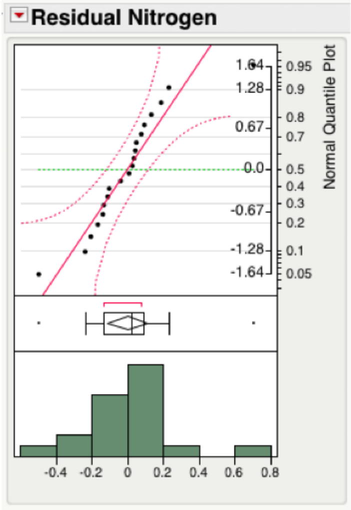
	Term	Estimate	Std Error	t Ratio	Prob>|t|
b_0	Intercept	1.7222135	1.234082	1.40	0.1832
b_1	Agr	0.0058091	0.015034	0.39	0.7046
b_2	Forest	-0.012968	0.013931	-0.93	0.3667
b_3	Rsdntial	-0.007227	0.03383	-0.21	0.8337
b_4	ComIndl	0.3050278	0.163817	1.86	0.0823

Full model: observed pollution values vs fitted values



Full model: residual plots



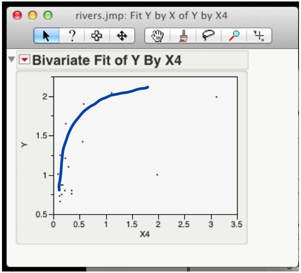


Conclusions: full model

- A higher R^2 indicates that the full model is more useful for predicting river pollution than the agriculture-only model.
- The residual plots show that the full model is valid too.

An even bigger model

- From the scatterplot of y on x_4, it looks like x_4 needs at least a quadratic term.



- I can fit the model:

$\log x_4 +$ $\quad$ x_4^2

$$\widehat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + b_3 x_{i,3} + b_4 x_{i,4} + c x_{i,4}^2$$

which is a combination of polynomial regression and multiple regression.

$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + c f(x_{i1}, x_{i2}, x_{i3}, x_{i4})$

The JMP Spreadsheet

rivers.jmp

rivers.jmp
Source

Columns (7/1)
River
X1
X2
X3
X4
Y
X4^2

Rows

All rows	20
Selected	0
Excluded	0
Hidden	0
Labelled	0

	River	X1	X2	X3	X4	Y	X4^2
1	Olean	26	63	1.2	0.29	1.1	0.0841
2	Cassadaga	29	57	0.7	0.09	1.01	0.0081
3	Oatka	54	26	1.8	0.58	1.9	0.3364
4	Neversink	2	84	1.9	1.98	1	3.9204
5	Hackensack	3	27	29.4	3.11	1.99	9.6721
6	Wappinger	19	61	3.4	0.56	1.42	0.3136
7	Fishkill	16	60	5.6	1.11	2.04	1.2321
8	Honeoye	40	43	1.3	0.24	1.65	0.0576
9	Susquehanna	28	62	1.1	0.15	1.01	0.0225
10	Chenango	26	60	0.9	0.23	1.21	0.0529
11	Tioughnioga	26	53	0.9	0.18	1.33	0.0324
12	West_Canada	15	75	0.7	0.16	0.75	0.0256
13	East_Canada	6	84	0.5	0.12	0.73	0.0144
14	Saranac	3	81	0.8	0.35	0.8	0.1225
15	Ausable	2	89	0.7	0.35	0.76	0.1225
16	Black	6	82	0.5	0.15	0.87	0.0225
17	Schoharie	22	70	0.9	0.22	0.8	0.0484
18	Raquette	4	75	0.4	0.18	0.87	0.0324
19	Oswegatchie	21	56	0.5	0.13	0.66	0.0169
20	Cohocton	40	49	1.1	0.13	1.25	0.0169

R^2 improves

Summary of Fit

RSquare	0.897008
RSquare Adj	0.860226
Root Mean Square Error	0.163247
Mean of Response	1.1575
Observations (or Sum Wgts)	20

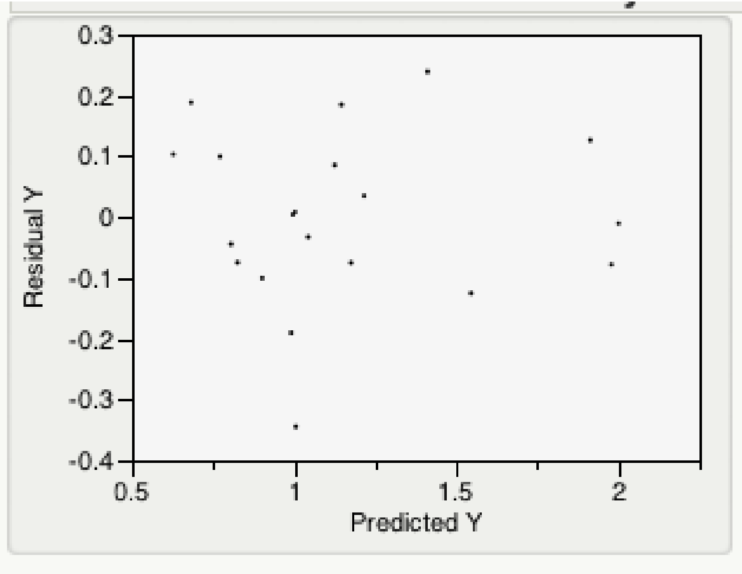
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	3.2494798	0.649896	24.3867
Error	14	0.3730952	0.026650	**Prob > F**
C. Total	19	3.6225750		<.0001*

Parameter Estimates

	Term	Estimate	Std Error	t Ratio	Prob>|t|
b_0	Intercept	1.2942455	0.765169	1.69	0.1129
⋮	X1	0.0049001	0.009266	0.53	0.6052
	X2	-0.010462	0.008599	-1.22	0.2438
b_4	X3	0.0737788	0.026304	2.80	0.0140*
	X4	1.2715886	0.216387	5.88	<.0001*
c	X4^2	-0.532452	0.105436	-5.05	0.0002*

The model looks valid: no pattern in the residuals



The model can be used for statistical inference: the residuals look normally distributed.