STAT 401A - Statistical Methods for Research Workers

Multiple regression analysis

Jarad Niemi (Dr. J)

Iowa State University

last updated: November 11, 2014

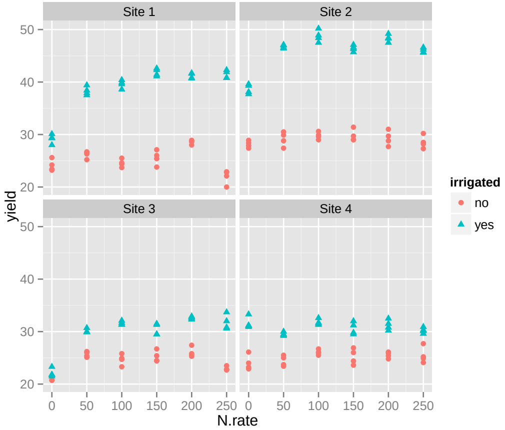
Potato yield

Based on exercise 10.25:

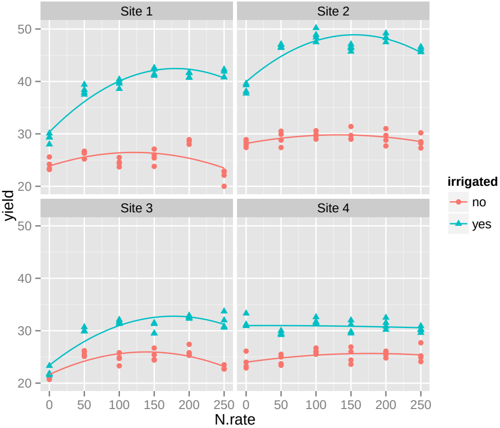
> *Nitrogen and water are important factors influencing potato production. The yield (t / ha) response of Russet Burbank potatoes to six rates of N fertilization (0-250 kg N / ha) with and without supplemental irrigation was studied at four on farm sites in 1995 in the upper St-John River Valley of New Brunswick, Canada.*
>
> *Belanger et al., (2000) "Yield Response of Two Potato Cultivars to Supplemental Irrigation and N fertilization in New Brunswick," American Journal of Potato Research 77:11-21*

Build a model that accurately captures the relationship between nitrogen rate and yield in these data.

```
   yield N.rate irrigated   site
1   23.4      0        no Site 1
2   24.2      0        no Site 1
3   23.2      0        no Site 1
4   25.6      0        no Site 1
5   26.3     50        no Site 1
6   25.2     50        no Site 1
7   26.5     50        no Site 1
8   26.7     50        no Site 1
9   24.6    100        no Site 1
10  23.7    100        no Site 1
11  25.5    100        no Site 1
12  24.4    100        no Site 1
13  25.4    150        no Site 1
14  23.8    150        no Site 1
15  27.1    150        no Site 1
16  26.0    150        no Site 1
17  28.0    200        no Site 1
18  28.9    200        no Site 1
19  28.8    200        no Site 1
20  28.6    200        no Site 1
21  22.9    250        no Site 1
22  22.8    250        no Site 1
23  22.1    250        no Site 1
24  20.0    250        no Site 1
25  30.1      0       yes Site 1
```


Site 1
Site 2
Site 3
Site 4
yield
N.rate
50
40
30
20
0
50
100
150
200
250
irrigated
no
yes

Foreshadowing



Building a model

Section 10.4.7 Informal Tests in Model Fitting:

> *Tests for hypotheses about regression coefficients – t-tests and extra-sum-of-squares F-tests – are valuable for two purposes: for formally providing evidence regarding questions of interest in a final model and for exploring models by testing potential terms at the exploratory stage.*

H_0: the βs are zero

H_1: at least one of the βs are non-zero

Steps to building the model

Add the following variables in order

1. nitrogen rate
2. irrigated
3. site
4. site*irrigated
5. rate*site*irrigated
6. rate^2*site*irrigated

For each new model:

1. Show the mathematical model
2. Show SAS code to fit the model
3. Show a plot to help interpret the model

Simple linear regression model

Consider the simple linear regression model with

- Y_i: yield for observation i
- N_i: nitrogen rate for observation i

$$Y_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2)$$

with

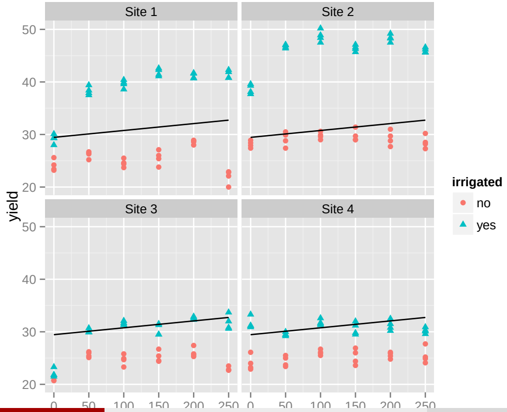
$$\mu_i = \beta_0 + \beta_1 N_i$$

```
DATA potato;
  INFILE 'potato.csv' DSD FIRSTOBS=2;
  INPUT site $ year irrigated $ Nrate meanyield yield;

PROC GLM;
  MODEL yield = Nrate / SOLUTION;
  RUN;
```

```
                                        The GLM Procedure

Dependent Variable: yield

                                              Sum of
       Source                      DF        Squares      Mean Square    F Value    Pr > F
       Model                        1      240.25400        240.25400       4.29    0.0397
       Error                      190    10639.18913         55.99573
       Corrected Total            191    10879.44313

                       R-Square     Coeff Var      Root MSE    yield Mean
                       0.022083      24.06844      7.483030      31.09062

                                                 Standard
                Parameter         Estimate          Error    t Value    Pr > |t|
                Intercept      29.45312500     0.95739062      30.76      <.0001
                Nrate           0.01310000     0.00632431       2.07      0.0397
```

Simple linear regression

```
p <- ggplot(d, aes(x=N.rate, y=yield))+
     geom_point(aes(shape=irrigated, color=irrigated))+
     facet_wrap(~site)
m = lm(yield~N.rate, d)
d2 = data.frame(N.rate = 0:250)
d2$yield = predict(m,d2)
p+geom_line(aes(x=N.rate, y=yield), d2)
```



Parallel lines model for irrigation

Add irrigation into the model

- I_i: indicator that the observation was irrigated

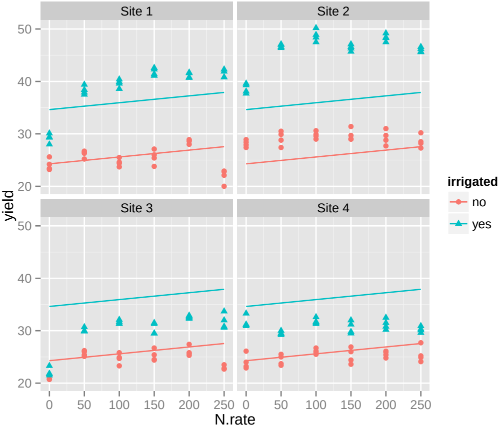
$$I_i = \begin{cases} 1 & \text{if observation } i \text{ was irrigated} \\ 0 & \text{if observation } i \text{ was not irrigated} \end{cases}$$

$$\mu_i = \beta_0 + \beta_1 N_i + \beta_2 I_i$$

- If $I_i = 0$, then $\mu_i = \beta_1 N_i + \beta_0$
- If $I_i = 1$, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_2$

```
PROC GLM;
  CLASS irrigated(ref='no');
  MODEL yield = Nrate irrigated / SOLUTION;
  RUN;

                                  The GLM Procedure

Dependent Variable: yield

                                           Sum of
      Source                      DF      Squares     Mean Square    F Value    Pr > F
      Model                        2   5377.99483      2688.99742      92.38    <.0001
      Error                      189   5501.44829        29.10819
      Corrected Total            191  10879.44313


                 R-Square     Coeff Var      Root MSE    yield Mean
                 0.494326      17.35314      5.395201      31.09062


                                                Standard
           Parameter            Estimate             Error    t Value    Pr > |t|
           Intercept         24.28020833 B      0.79251407      30.64      <.0001
           Nrate              0.01310000        0.00455978       2.87      0.0045
           irrigated yes     10.34583333 B      0.77873016      13.29      <.0001
           irrigated no       0.00000000 B       .                .           .
```

Parallel lines model for irrigation



Additive model for irrigation and site

Add site into the model:

- $S1_i$ indicator that the observation was from Site 1
- $S2_i$ indicator that the observation was from Site 2
- $S3_i$ indicator that the observation was from Site 3

$$Sj_i = \begin{cases} 1 & \text{if observation } i \text{ was from Site } j \\ 0 & \text{if observation } i \text{ was not from Site } j \end{cases}$$

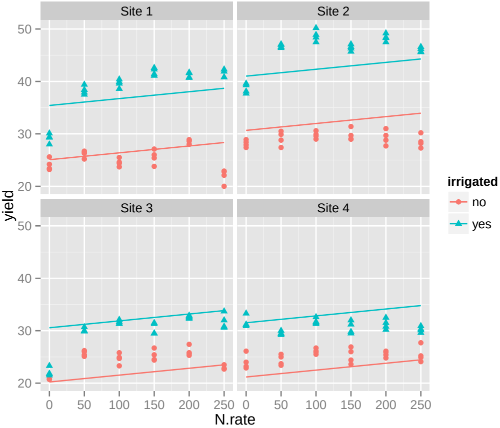
$$\mu_i = \beta_0 + \beta_1 N_i + \beta_2 I_i + \beta_3 S1_i + \beta_4 S2_i + \beta_5 S3_i$$

- If $I_i = 0$ and Site 4, then $\mu_i = \beta_1 N_i + \beta_0$
- If $I_i = 0$ and Site 1, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_3$
- If $I_i = 1$ and Site 4, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_2$
- If $I_i = 1$ and Site 1, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_2 + \beta_3$

```
PROC GLM;
  CLASS site irrigated(ref='no');
  MODEL yield = Nrate irrigated site / SOLUTION;
  RUN;
```

```
                                The GLM Procedure
Dependent Variable: yield

                                       Sum of
      Source                  DF      Squares     Mean Square    F Value    Pr > F
      Model                    5   8621.00088      1724.20018     142.00    <.0001
      Error                  186   2258.44225        12.14216
      Corrected Total        191  10879.44313

                  R-Square     Coeff Var      Root MSE    yield Mean
                  0.792412      11.20775      3.484561      31.09062


      Source                  DF  Type III SS     Mean Square    F Value    Pr > F
      Nrate                    1   240.254000      240.254000      19.79    <.0001
      irrigated                1  5137.740833     5137.740833     423.13    <.0001
      site                     3  3243.006042     1081.002014      89.03    <.0001


                                                Standard
         Parameter              Estimate           Error    t Value    Pr > |t|
         Intercept           21.17083333 B    0.67209922      31.50      <.0001
         Nrate                0.01310000      0.00294499       4.45      <.0001
         irrigated yes       10.34583333 B    0.50295300      20.57      <.0001
         irrigated no         0.00000000 B     .                .         .
         site      Site 1     3.89791667 B    0.71128296       5.48      <.0001
         site      Site 2     9.49375000 B    0.71128296      13.35      <.0001
         site      Site 3    -0.95416667 B    0.71128296      -1.34      0.1814
         site      Site 4     0.00000000 B     .                .         .
```

Additive model for irrigation



Parallel lines model for each irrigation-site combination

Add the irrigation-site interaction:

$$\begin{aligned}\mu_i = \quad & \beta_0 + \beta_1 N_i + \beta_2 I_i + \beta_3 S1_i + \beta_4 S2_i + \beta_5 S3_i \\ & + \beta_6 S1_i I_i + \beta_7 S2_i I_i + \beta_8 S3_i I_i\end{aligned}$$

- If $I_i = 0$ and Site 4, then $\mu_i = \beta_1 N_i + \beta_0$
- If $I_i = 0$ and Site 1, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_3$
- If $I_i = 0$ and Site 2, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_4$
- If $I_i = 0$ and Site 3, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_5$
- If $I_i = 1$ and Site 4, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_2$
- If $I_i = 1$ and Site 1, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_3 + \beta_2 + \beta_6$
- If $I_i = 1$ and Site 2, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_4 + \beta_2 + \beta_7$
- If $I_i = 1$ and Site 3, then $\mu_i = \beta_1 N_i + \beta_0 + \beta_5 + \beta_2 + \beta_8$

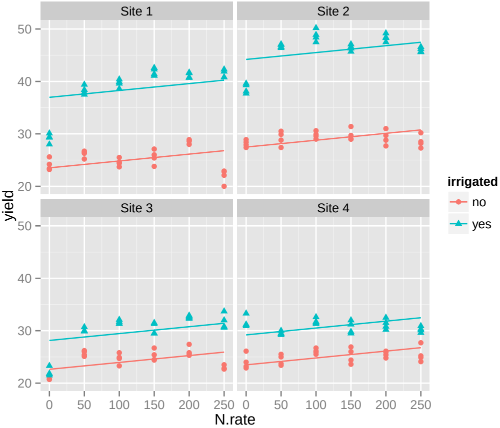
```
PROC GLM;
  CLASS site irrigated(ref='no');
  MODEL yield = Nrate irrigated|site / SOLUTION;
  RUN;
```

```
                                    The GLM Procedure
Dependent Variable: yield

                                              Sum of
      Source                      DF         Squares     Mean Square    F Value    Pr > F
      Model                        8      9762.59546      1220.32443     199.96    <.0001
      Error                      183      1116.84767         6.10299
      Corrected Total            191     10879.44313

                    R-Square     Coeff Var      Root MSE    yield Mean
                    0.897343      7.945879      2.470424      31.09062


      Source                      DF     Type III SS     Mean Square    F Value    Pr > F
      Nrate                        1      240.254000      240.254000      39.37    <.0001
      site                         3     3243.006042     1081.002014     177.13    <.0001
      irrigated                    1     5137.740833     5137.740833     841.84    <.0001
      site*irrigated               3     1141.594583      380.531528      62.35    <.0001
```

```
PROC GLM;
  CLASS site irrigated(ref='no');
  MODEL yield = Nrate irrigated|site / SOLUTION;
  RUN;
```

```
                                 The GLM Procedure
Dependent Variable: yield

                                          Sum of
      Source                     DF      Squares     Mean Square    F Value    Pr > F
      Model                       8   9762.59546      1220.32443     199.96    <.0001
      Error                     183   1116.84767         6.10299
      Corrected Total           191  10879.44313

                                                      Standard
      Parameter                         Estimate         Error    t Value    Pr > |t|

      Intercept                      23.48750000 B    0.56780729      41.37      <.0001
      Nrate                           0.01310000      0.00208789       6.27      <.0001
      irrigated       yes             5.71250000 B    0.71314986       8.01      <.0001
      irrigated       no              0.00000000 B     .                .         .
      site            Site 1          0.02916667 B    0.71314986       0.04      0.9674
      site            Site 2          3.99166667 B    0.71314986       5.60      <.0001
      site            Site 3         -0.85000000 B    0.71314986      -1.19      0.2348
      site            Site 4          0.00000000 B     .                .         .

      site*irrigated Site 1 yes       7.73750000 B    1.00854621       7.67      <.0001
      site*irrigated Site 1 no        0.00000000 B     .                .         .
      site*irrigated Site 2 yes      11.00416667 B    1.00854621      10.91      <.0001
      site*irrigated Site 2 no        0.00000000 B     .                .         .
      site*irrigated Site 3 yes      -0.20833333 B    1.00854621      -0.21      0.8366
      site*irrigated Site 3 no        0.00000000 B     .                .         .
      site*irrigated Site 4 yes       0.00000000 B     .                .         .
```

Parallel lines model for irrigation*site



Independent lines model for each irrigation-site combination

Add the site-irrigation combination interacted with nitrogen rate

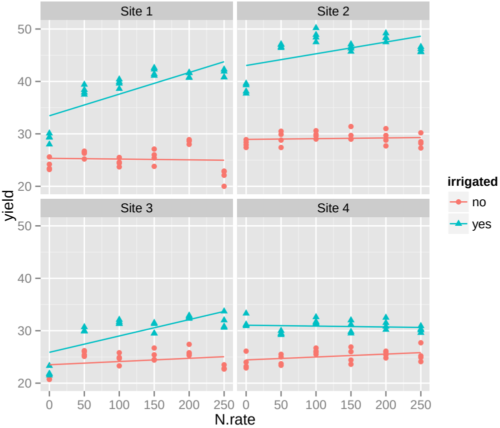
$$
\begin{aligned}
\mu_i = \quad & \beta_0 + \beta_1 I_i \\
& + \beta_2 S1_i + \beta_3 S2_i + \beta_4 S3_i \\
& + \beta_5 S1_i I_i + \beta_6 S2_i I_i + \beta_7 S3_i I_i \\
& + \beta_8 N_i \\
\\
& + \beta_9 N_i I_i \\
& + \beta_{10} N_i S1_i + \beta_{11} N_i S2_i + \beta_{12} N_i S3_i \\
& + \beta_{13} N_i S1_i I_i + \beta_{14} N_i S2_i I_i + \beta_{15} N_i S3_i I_i
\end{aligned}
$$

- If $I_i = 0$ and Site 4, then $\mu_i = \beta_0 + \beta_8 N_i$
- If $I_i = 1$ and Site 3, then $\mu_i = \beta_0 + \beta_1 + \beta_4 + \beta_7 + (\beta_8 + \beta_9 + \beta_{12} + \beta_{15}) N_i$

```
PROC GLM;
  CLASS site irrigated(ref='no');
  MODEL yield = irrigated|site|Nrate / SOLUTION;
  RUN;

                                        The GLM Procedure
Dependent Variable: yield

                                                Sum of
       Source                      DF          Squares     Mean Square    F Value    Pr > F
       Model                       15      10091.35274       672.75685     150.24    <.0001
       Error                      176        788.09038         4.47779
       Corrected Total            191      10879.44313

                       R-Square     Coeff Var      Root MSE    yield Mean
                       0.927562      6.806161      2.116078      31.09062


       Source                      DF      Type III SS     Mean Square    F Value    Pr > F
       irrigated                    1       929.263638      929.263638     207.53    <.0001
       site                         3      1043.558521      347.852840      77.68    <.0001
       site*irrigated               3       269.564182       89.854727      20.07    <.0001
       Nrate                        1       240.254000      240.254000      53.65    <.0001
       Nrate*irrigated              1       145.146446      145.146446      32.41    <.0001
       Nrate*site                   3        71.216679       23.738893       5.30    0.0016
       Nrate*site*irrigated         3       112.394161       37.464720       8.37    <.0001
```

```
                                The GLM Procedure
Dependent Variable: yield

                                                          Standard
    Parameter                         Estimate               Error    t Value    Pr > |t|

    Intercept                      24.42857143 B        0.76575242      31.90      <.0001

    irrigated            yes        6.62142857 B        1.08293746       6.11      <.0001

    site                 Site 1     0.90595238 B        1.08293746       0.84      0.4040
    site                 Site 2     4.51309524 B        1.08293746       4.17      <.0001
    site                 Site 3    -0.92142857 B        1.08293746      -0.85      0.3960

    site*irrigated       Site 1 yes 1.49285714 B        1.53150484       0.97      0.3310
    site*irrigated       Site 2 yes 7.47380952 B        1.53150484       4.88      <.0001
    site*irrigated       Site 3 yes -4.25119048 B       1.53150484      -2.78      0.0061

    Nrate                           0.00557143 B        0.00505839       1.10      0.2722


    Nrate*irrigated      yes       -0.00727143 B        0.00715365      -1.02      0.3108

    Nrate*site           Site 1    -0.00701429 B        0.00715365      -0.98      0.3282
    Nrate*site           Site 2    -0.00417143 B        0.00715365      -0.58      0.5606
    Nrate*site           Site 3     0.00057143 B        0.00715365       0.08      0.9364

    Nrate*site*irrigated Site 1 yes 0.04995714 B        0.01011679       4.94      <.0001
    Nrate*site*irrigated Site 2 yes 0.02824286 B        0.01011679       2.79      0.0058
    Nrate*site*irrigated Site 3 yes 0.03234286 B        0.01011679       3.20      0.0016
```

Independent lines model for each site-irrigated combination



Independent curves model for each irrigation-site combination

Add the site-irrigation combination interacted with nitrogen rate squared:

$$
\begin{aligned}
\mu_i = \quad & \beta_0 + \beta_1 I_i \\
& + \beta_2 S1_i + \beta_3 S2_i + \beta_4 S3_i \\
& + \beta_5 S1_i I_i + \beta_6 S2_i I_i + \beta_7 S3_i I_i \\
& + \beta_8 N_i \\
& + \beta_9 N_i I_i \\
& + \beta_{10} N_i S1_i + \beta_{11} N_i S2_i + \beta_{12} N_i S3_i \\
& + \beta_{13} N_i S1_i I_i + \beta_{14} N_i S2_i I_i + \beta_{15} N_i S3_i I_i \\
\\
& + \beta_{16} N_i^2 \\
& + \beta_{17} N_i^2 I_i \\
& + \beta_{18} N_i^2 S1_i + \beta_{19} N_i^2 S2_i + \beta_{20} N_i^2 S3_i \\
& + \beta_{21} N_i^2 S1_i I_i + \beta_{22} N_i^2 S2_i I_i + \beta_{23} N_i^2 S3_i I_i
\end{aligned}
$$

- If $I_i = 0$ and Site 4, then $\mu_i = \beta_0 + \beta_8 N_i + \beta_{16} N_i^2$
- If $I_i = 1$ and Site 3, then $\mu_i = \beta_0 + \beta_1 + \beta_4 + \beta_{12} + (\beta_8 + \beta_9 + \beta_{12} + \beta_{15}) N_i + (\beta_{16} + \beta_{17} + \beta_{20} + \beta_{23}) N_i^2$

```
PROC GLM;
  CLASS site irrigated(ref='no');
  MODEL yield = irrigated|site|Nrate irrigated|site|Nrate*Nrate / SOLUTION;
  RUN;
```

The GLM Procedure

Dependent Variable: yield

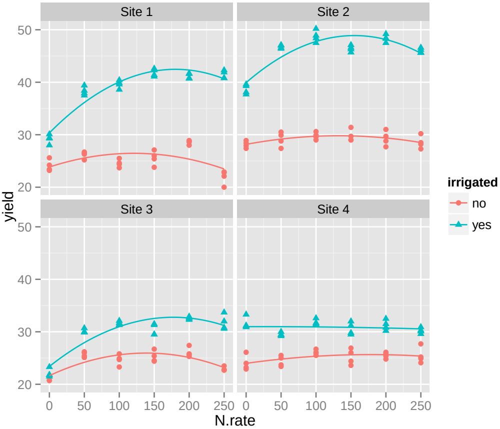
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	10521.10013	457.43914	214.46	<.0001
Error	168	358.34300	2.13299		
Corrected Total	191	10879.44313			

R-Square	Coeff Var	Root MSE	yield Mean
0.967062	4.697485	1.460477	31.09062

Source	DF	Type III SS	Mean Square	F Value	Pr > F
irrigated	1	5137.740833	5137.740833	2408.70	<.0001
site	3	3243.006042	1081.002014	506.80	<.0001
site*irrigated	3	1141.594583	380.531528	178.40	<.0001
Nrate	1	240.254000	240.254000	112.64	<.0001
Nrate*irrigated	1	145.146446	145.146446	68.05	<.0001
Nrate*site	3	71.216679	23.738893	11.13	<.0001
Nrate*site*irrigated	3	112.394161	37.464720	17.56	<.0001
Nrate*Nrate	1	298.933393	298.933393	140.15	<.0001
Nrate*Nrate*irrigate	1	29.500952	29.500952	13.83	0.0003
Nrate*Nrate*site	3	73.215565	24.405188	11.44	<.0001
Nrat*Nrat*site*irrig	3	28.097470	9.365823	4.39	0.0053

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		23.98660714	B	0.66183500	36.24	<.0001
irrigated	yes	6.99642857	B	0.93597603	7.48	<.0001
site	Site 1	-0.14017857	B	0.93597603	-0.15	0.8811
site	Site 2	4.17232143	B	0.93597603	4.46	<.0001
site	Site 3	-2.34107143	B	0.93597603	-2.50	0.0133
site*irrigated	Site 1 yes	-0.49821429	B	1.32366999	-0.38	0.7071
site*irrigated	Site 2 yes	4.76696429	B	1.32366999	3.60	0.0004
site*irrigated	Site 3 yes	-5.24375000	B	1.32366999	-3.96	0.0001
Nrate		0.01883036	B	0.01245082	1.51	0.1323
Nrate*irrigated	yes	-0.01852143	B	0.01760812	-1.05	0.2944
Nrate*site	Site 1	0.02436964	B	0.01760812	1.38	0.1682
Nrate*site	Site 2	0.00605179	B	0.01760812	0.34	0.7315
Nrate*site	Site 3	0.04316071	B	0.01760812	2.45	0.0153
Nrate*site*irrigated	Site 1 yes	0.10968929	B	0.02490164	4.40	<.0001
Nrate*site*irrigated	Site 2 yes	0.10944821	B	0.02490164	4.40	<.0001
Nrate*site*irrigated	Site 3 yes	0.06211964	B	0.02490164	2.49	0.0136
Nrate*Nrate		-0.00005304	B	0.00004781	-1.11	0.2688
Nrate*Nrate*irrigate	yes	0.00004500	B	0.00006761	0.67	0.5066
Nrate*Nrate*site	Site 1	-0.00012554	B	0.00006761	-1.86	0.0651
Nrate*Nrate*site	Site 2	-0.00004089	B	0.00006761	-0.60	0.5461
Nrate*Nrate*site	Site 3	-0.00017036	B	0.00006761	-2.52	0.0127
Nrat*Nrat*site*irrig	Site 1 yes	-0.00023893	B	0.00009561	-2.50	0.0134
Nrat*Nrat*site*irrig	Site 2 yes	-0.00032482	B	0.00009561	-3.40	0.0008
Nrat*Nrat*site*irrig	Site 3 yes	-0.00011911	B	0.00009561	-1.25	0.2146

Independent curves model for each site-irrigated combination



Summary

Demonstrated model construction at the exploratory stage using informal tests to help construct the model.