

Best Linear Unbiased Prediction-Variance Structure

References

Searle, S.R. 1971 Linear Models, Wiley

Schaefer, L.R., Linear Models and Computer Strategies in Animal Breeding

Lynch and Walsh Chapter 8

Linear vs non-linear

Linear
2nd order Polynomial

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2 + b_3 X_{3i} + b_4 X_{4i}^2 + b_5 (X_{1i} X_{2i}) + \varepsilon_i$$

Non-linear

$$Y_i = b_0 e^{-b_1 X_i} \varepsilon_i$$

log-linear

$$\ln(Y_i) = \ln(b_0) - b_1 X_i + \ln(\varepsilon_i)$$

Why Linear

Taylor Expansion

$$Y = f(X)$$

$$Y \approx f(a) + \frac{f'(a)(x-a)}{1!} + \frac{f''(a)(x-a)^2}{2!} + \dots + \frac{f^n(a)(x-a)^n}{n!}$$

$$Y = e^{-X} \quad Y' = -e^{-X} \quad Y'' = e^{-X}$$

At $a=0$

$$Y \approx 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \dots$$

Lower Order Terms Are more Important than higher

Works for other values of a but not as exact, example a=.1

$$Y = e^{-X}$$

a=.1

$$Y = e^{-.1} = .904837$$

$$Y \approx 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \dots$$

$$Y \approx 1 - x = 1 - .1 = .9$$

$$Y \approx 1 - x + \frac{x^2}{2!} = 1 - .1 + \frac{.1^2}{2!} = .905$$

$$Y \approx 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} = 1 - .1 + \frac{.1^2}{2!} - \frac{.1^3}{3!} = .904833$$

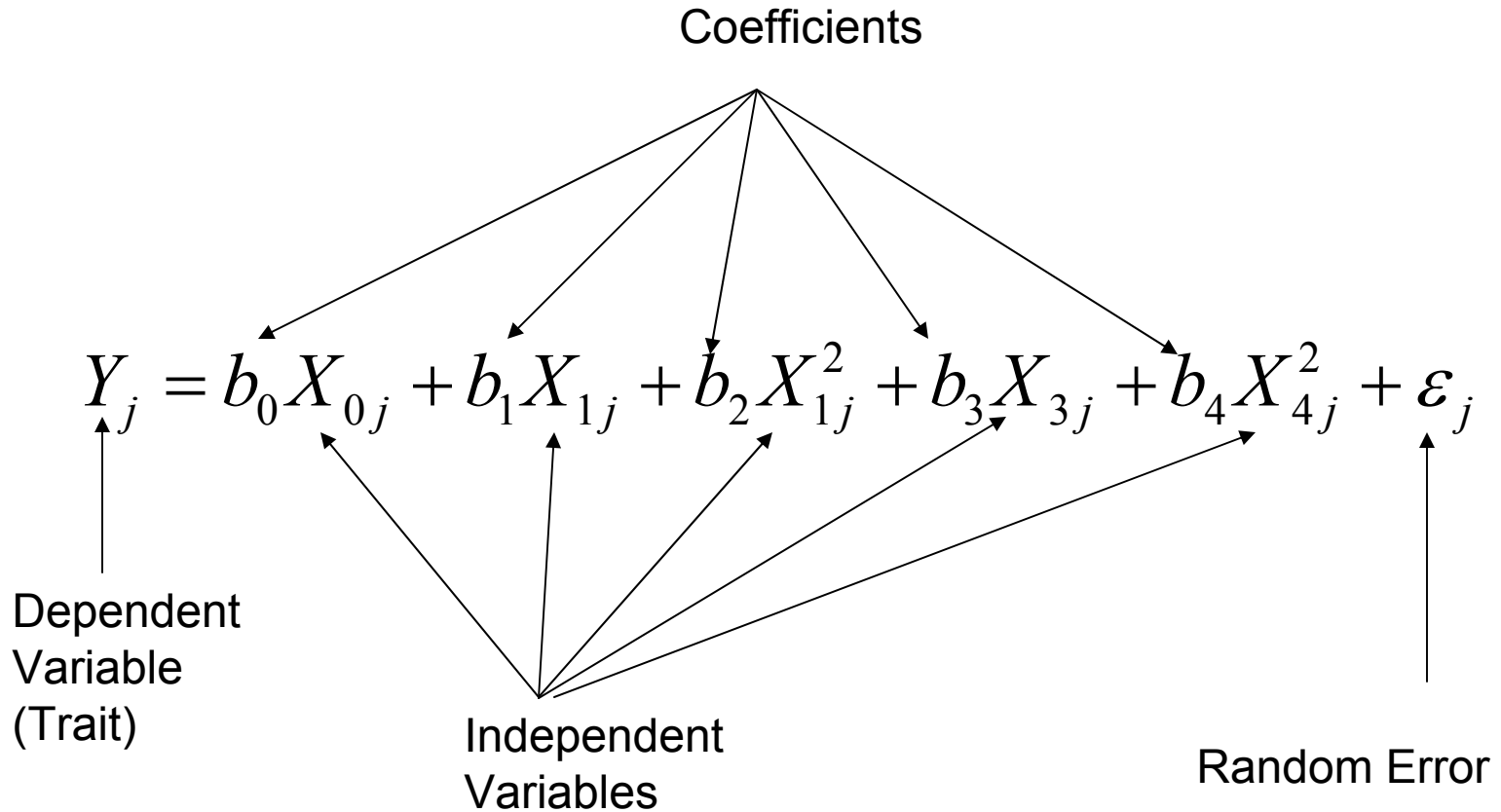
Generality

- Any underlying unknown function can be approximated by a polynomial equation (linear Model)
 - Lower order terms are more important than higher order
 - Model does not have any basis in biological function
 - Even highly non-linear systems can be approximated by a linear model with only lower order terms
 - Purely Descriptive
 - Allows tests of hypothesis related to treatment effects
 - Allows limited prediction (expansion is around a point)

Linear Model

- Can be used to approximate highly non-additive genetic systems, including dominance and epistasis
- Predictive ability is fairly good, even if underlying mode of gene action is non-additive
- Linear Models Extensively Used in Animal Breeding

One Random effect Linear Model



Matrix Notation

$$Y_1 = b_0 X_{01} + b_1 X_{11} + b_2 X_{11}^2 + b_3 X_{31} + b_4 X_{41}^2 + \varepsilon_1$$

$$Y_2 = b_0 X_{02} + b_1 X_{12} + b_2 X_{12}^2 + b_3 X_{32} + b_4 X_{42}^2 + \varepsilon_2$$

⋮

$$Y_n = b_0 X_{0n} + b_1 X_{1n} + b_2 X_{1n}^2 + b_3 X_{3n} + b_4 X_{4n}^2 + \varepsilon_n$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{01} & X_{11} & X_{11}^2 & X_{31} & X_{41}^2 \\ X_{02} & X_{12} & X_{12}^2 & X_{32} & X_{42}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{0n} & X_{1n} & X_{1n}^2 & X_{3n} & X_{4n}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \boldsymbol{\varepsilon}$$

Estimation

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}$$

Ordinary Least Squares

- Independent variables (X)
 - fixed
 - measured without error
- Residuals
 - Random
 - Independently and Identically Distributed (IID) with Mean 0 and variance σ^2

Independently and Identically Distributed with Mean 0 and variance σ^2

$$V(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})]^2$$

$$V(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_e^2 \end{bmatrix}$$

$$V(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma_e^2$$

The error distribution from which each observation is sampled is the same

No Environmental Correlations

When would these assumptions be violated?

Ordinary Least Squares Estimator

$$\sum_{j=1}^n \varepsilon_j^2 = \boldsymbol{\varepsilon}' \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \varepsilon_j = Y_j - E(Y_j)$$

$$E(Y_j) = \sum_{i=0}^k b_i X_{ij} \quad \varepsilon_j = Y_j - \sum_{i=0}^k b_i X_{ij}$$

$$\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon} = \sum_{j=1}^n \varepsilon_j^2 = \sum_{j=1}^n \left(Y_j - \sum_{i=0}^k b_i X_{ij} \right)^2$$

Find Solutions such that the sum of the residuals squared is minimum

Least Square Estimators

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \sum_{j=1}^n \varepsilon_j^2 = \sum_{j=1}^n \left(Y_j - \sum_{i=0}^m b_i X_{ij} \right)^2$$

$$\frac{\partial(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})}{\partial b_i} = 2 \sum_{j=1}^n \left(Y_j - \sum_{i=0}^m b_i X_{ij} \right) [-b_i X_{ij}]$$

Set=0 for each i and solve system

Normal Equations

$$\begin{bmatrix} \sum x_{0j}^2 & \sum x_{0j}x_{1j} & \cdots & \sum x_{0j}x_{kj} \\ \sum x_{0j}x_{1j} & \sum x_{1j}^2 & \cdots & \sum x_{1j}x_{kj} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{0j}x_{kj} & \sum x_{1j}x_{kj} & \cdots & \sum x_{kj}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} \sum x_{0j}y_j \\ \sum x_{1j}y_j \\ \vdots \\ \sum x_{kj}y_j \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X}\mathbf{B} = \mathbf{X}'\mathbf{Y}$$

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

$$V(\hat{\mathbf{B}}) = \sigma_e^2 (\mathbf{X}'\mathbf{X})^{-1}$$

Prediction

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$$

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

$$V(\hat{\mathbf{Y}}) = V(\mathbf{X}\hat{\mathbf{B}})$$

$$V(\hat{\mathbf{B}}) = \sigma_e^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$V(\hat{\mathbf{Y}}) = \mathbf{X}V(\hat{\mathbf{B}})\mathbf{X}'$$

$$V(\hat{\mathbf{Y}}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma_e^2$$

Example Factor Affecting Fatty Acid

From Gill, J. Design and Analysis of experiments

Fatty Acid	Amount over Weight (Kg)	Age
10	6	28
20	12	40
17	10	32
12	8	36
11	9	34

OLS by IML

- *data from Gill, Design and Analysis of experiments;
- *Demonstrates OLS estimation via matrix methods;

```
proc iml;
start main;
y={ 10,
    20,
    17,
    12,
    11 };
x={ 1 6 28,
    1 12 40,
    1 10 32,
    1 8 36,
    1 9 34 };
```

```
b=inv(x`*x)*x`*y;
Yhat=x*b;
e=y-x*b;
sse=e`*e;
df=1/2;
rms=sse#df;
vb=inv(x`*x)#rms;
Vyhat=x*inv(x`*x)*x`#rms;
print b vb y yhat e sse rms vyhat;
finish main;
run;quit;
```

BY GLM

- **data** one;
- input fatty_acid over_wt age;
- cards;
- 10 6 28
- 20 12 40
- 17 10 32
- 12 8 36
- 11 9 34
- ;
- **proc glm;**
- model fatty_acid=over_wt age
/ solution;
- **run;**
- quit;
- Compare results from IML to GLM

Lab Problem 5.1

Inbreeding data (Data set 1) Averaged Over 20 Reps (see program next slide)

	Gen	Y1	Y2	Y3	F
• Fit the Generation Means for each trait to a second order model against F	1	110.4	137.5	2.28	0.000
– Use IML	2	147.4	136.8	2.06	0.157
– Use GLM	3	122.0	130.6	1.70	0.281
• Plot each trait against F	4	102.0	129.9	2.02	0.333
• Why do you get a different relationships for each trait?	5	64.2	122.9	1.40	0.396
	6	38.5	121.0	1.20	0.459
	7	12.8	120.0	1.22	0.491
	8	-21.8	118.4	0.66	0.584
	9	-20.0	115.2	1.03	0.624
	10	-56.5	115.1	1.06	0.657

SAS program to calculate Generation means for each replicate

```
data a1;  
input Rep Gen sire dam animal y1 y2 y3;  
cards;
```

1	1	2	10001	20001	341.561	140.218	-1.32
1	1	3	10002	20002	316.23	143.202	-0.598

```
Proc sort;by gen;  
proc means noprint;by gen;var y1 y2 y3;  
output out=m1 mean=y1 y2 y3;  
proc print; run;
```

Generalized Least Squares (GLS)

- Ordinary Least Squares

- Independent variables
 - fixed
 - measured without error
- Residuals
 - Random
 - Independently and Identically Distributed (IID) with Mean 0 and variance σ^2

- Generalized Least Squares

- Independent variables
 - fixed
 - measured without error
- Residuals
 - Random

$$V(\boldsymbol{\varepsilon}) = \mathbf{V}$$

GLS

Minimize $(\mathbf{y} - \mathbf{X}\mathbf{b})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\mathbf{b})$

Weighting by the inverse of the variance

$$\hat{\mathbf{b}} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} (\mathbf{X}' \mathbf{V}^{-1} \mathbf{y})$$

If

$$\mathbf{V} = \mathbf{I} \sigma_e^2$$

$$\hat{\mathbf{b}} = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{y})$$

Maximum Likelihood (ML)

- Generalized Least Squares
 - Independent variables
 - fixed
 - measured without error
 - Residuals
 - Random

$$V(\boldsymbol{\varepsilon}) = \mathbf{V}$$

- Maximum Likelihood
 - Independent variables
 - fixed
 - measured without error
 - Residuals
 - Random

$$V(\boldsymbol{\varepsilon}) = \mathbf{V}$$

$$\boldsymbol{\varepsilon} \approx N(\mathbf{0}, \mathbf{V})$$

ML

$$L = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{V}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{b})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\mathbf{b})}$$

Maximize w.r.t \mathbf{b}

$$\frac{\partial(\ln L)}{\partial \mathbf{b}} = 0$$

$$\ln L = \ln(C) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{b})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\mathbf{b})$$

$$\frac{\partial(\ln L)}{\partial \mathbf{b}} = -\frac{1}{2}(\mathbf{y} - \mathbf{Xb})' \mathbf{V}^{-1}(-\mathbf{X}) - \frac{1}{2}(-\mathbf{X})' \mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb})$$

$$\frac{\partial(\ln L)}{\partial \mathbf{b}} = (\mathbf{y} - \mathbf{Xb})' \mathbf{V}^{-1} \mathbf{X}$$

$$(\mathbf{y} - \mathbf{Xb})' \mathbf{V}^{-1} \mathbf{X} = 0$$

$$(\mathbf{y}' - (\mathbf{Xb})') \mathbf{V}^{-1} \mathbf{X} = 0'$$

$$(\mathbf{y}' - \mathbf{b}' \mathbf{X}') \mathbf{V}^{-1} \mathbf{X} = 0$$

$$\mathbf{b}' \mathbf{X}' \mathbf{V}^{-1} \mathbf{X} = \mathbf{y}' \mathbf{V}^{-1} \mathbf{X}$$

$$\hat{\mathbf{b}}' = (\mathbf{y}' \mathbf{V}^{-1} \mathbf{X})(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1}$$

$$\hat{\mathbf{b}} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} (\mathbf{X}' \mathbf{V}^{-1} \mathbf{y})$$

Same as GLS

Variance of \mathbf{b}

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

Note if $\mathbf{V} = \mathbf{I}\sigma_e^2$

$$V(\mathbf{b}) = \sigma_e^2 (\mathbf{X}'\mathbf{X})^{-1}$$

Lab Problem: 5.2

Estimate Between Replicate Variance at Each Generation

data a1;

input Rep Gen sire dam animal y1 y2 y3;

cards;

1	1	2	10001	20001	341.561	140.218	-1.32
1	1	3	10002	20002	316.23	143.202	-0.598

proc means noprint;by rep gen;var y1 y2 y3;

output out=m1 mean=y1 y2 y3;

proc sort;by gen;

proc means noprint;by gen;var y1 y2 y3;

output out=m2 mean=my1 my2 my3 std=sb1 sb2 sb3;

proc print; run;

Lab Problem 5.3: Estimate Within Replicate Variance for Each Generation

```
data a1;  
input Rep Gen sire dam animal y1 y2 y3;  
cards;
```

1	1	2	10001	20001	341.561	140.218	-1.32
1	1	3	10002	20002	316.23	143.202	-0.598

```
proc means noprint;by rep gen;var y1 y2 y3;  
output out=m1 std=swy1 swy2 swy3;
```

```
proc sort data=m1;by gen;  
proc means noprint;by gen;var swy1 swy2 swy3;  
output out=m2 mean= mswy1 mswy2 mswy3;  
proc print; run;
```

Clearly Between Replicate Variances Increase With Generations

Find GLS Estimates of
Regression Coefficients

Calculate OLS and GLS Estimates of Linear Regression Coefficients

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \quad \hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$$

OLS

```
proc iml;
Start main;
b=inv(x`*x)*x`*y;
Yhat=x*b;
e=y-x*b;
sse=e`*e;
df=1/8;
rms=sse#df;
vb=inv(x`*x)#rms;
print b vb y yhat e sse rms;
finish main;
run;
quit;
```

GLS

```
proc iml;
Start main;
b=inv(x`*inv(v)*x)*x`*inv(v)*y;
Yhat=x*b;
e=y-x*b;
sse=e`*e;
df=1/8;
rms=sse#df;
vb=inv(x`*inv(v)*x)#rms;
print b vb y yhat e sse rms;
finish main;
run;
quit;
```

Where for Trait 3 Y, V and X are

$$y = \{ 2.28460, 2.06908, 1.70690, 2.02030, 1.40335, 1.20441, 1.22212, 0.66843, 1.03762, 1.06027 \};$$
$$V = \{ 0.72 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0, \\ 0 \ 6.00 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0, \\ 0 \ 0 \ 12.39 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0, \\ 0 \ 0 \ 0 \ 11.30 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0, \\ 0 \ 0 \ 0 \ 0 \ 16.97 \ 0 \ 0 \ 0 \ 0 \ 0, \\ 0 \ 0 \ 0 \ 0 \ 0 \ 19.45 \ 0 \ 0 \ 0 \ 0, \\ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 23.45 \ 0 \ 0 \ 0, \\ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 33.52 \ 0 \ 0, \\ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 40.45 \ 0, \\ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 34.46 \};$$
$$X = \{ 1 \ 0.000, \\ 1 \ 0.157, \\ 1 \ 0.281, \\ 1 \ 0.333, \\ 1 \ 0.396, \\ 1 \ 0.459, \\ 1 \ 0.491, \\ 1 \ 0.584, \\ 1 \ 0.624, \\ 1 \ 0.657 \};$$

Compare Estimates and SE

	B		VB	
	2.39		0.023	-0.047
OLS	-2.32		-0.047	0.118

$$t = \frac{-2.32}{.118} = 19.46^{**}$$

	B		VB	
	2.30		0.035	-0.083
GLS	-2.03		-0.083	0.823

$$t = \frac{-2.03}{.832} = 2.43^{NS}$$

LP 5.4 Do same thing for other 2 traits

Conclusions

- Clearly if Proper Error Structure Not Used
 - Wrong Conclusions Occur
 - Replicate Lines Drift Apart
- Solution
 - Replicate Lines
 - Incorporate Drift Variance Into Model
 - Drift is Proportional to F and additive genetic variance
 - See Next Example
 - BLUP will be a partial solution to this problem.

LP 5.5: For Each Trait Find Empirical Relationship Between the within and between population variance

$$V(\textit{Within}) = F + e \quad \text{and } F \quad V(\textit{Between}) = F + e$$

Data a1;

input Gen Mean SW SB F;

VW=SW**2;

VB=SB**2;

CARDS;

1	2.28460	4.71875	0.84600	0.000
2	2.06908	4.34407	2.45756	0.157
3	1.70690	4.07810	3.52271	0.281
4	2.02030	3.72556	3.37658	0.333
5	1.40335	3.73361	4.11996	0.396
6	1.20441	3.86685	4.40742	0.459
7	1.22212	3.59145	4.84650	0.491
8	0.66843	3.47721	5.79003	0.584
9	1.03762	3.11475	6.35566	0.624
10	1.06027	3.06278	5.87509	0.657

PROC GLM;

MODEL VW VB=F;

RUN;

QUIT;

- From the linear regression of within or between variance on F is it possible to estimate the additive and environmental variance?
- What are these?