2004

More Estimation of Genetic Parameters

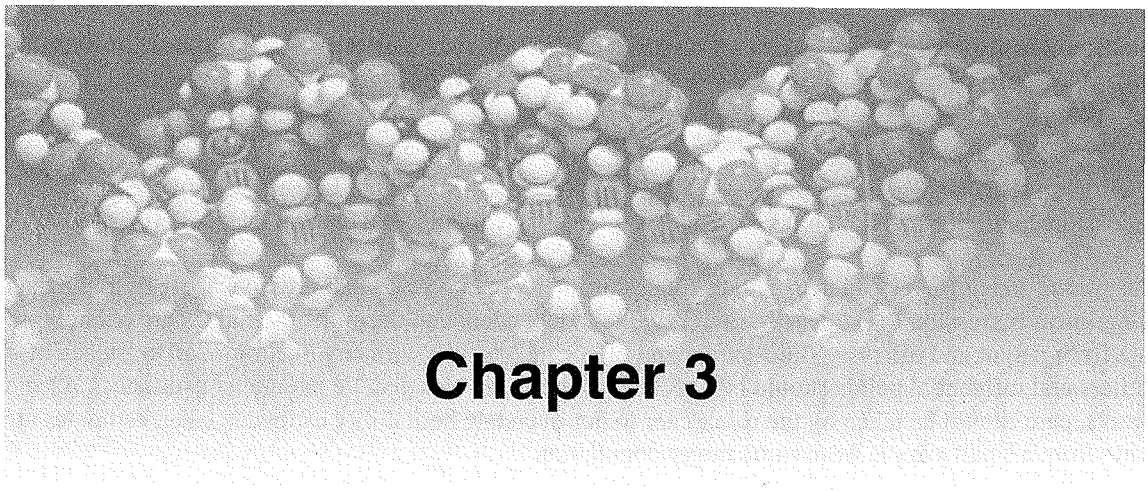
Kenneth J. Stalder
Iowa State University, stalder@iastate.edu

Arnold M. Saxton
University of Tennessee

More Estimation of Genetic Parameters

Abstract
This chapter addresses genetic parameters and estimation methods not covered in other chapters. Topics include regression designs for estimating heritability and genetic correlations, inbreeding, heterosis or crossbreeding, and realized heritability estimated from selection expe1iments. These techniques are from classic quantitative genetics, providing information on the genetics of populations using only pedigree and phenotypic information. The primary objective is to demonstrate how SAS software can be used to obtain this information, with minimal genetic background provided. More details are available in excellent texts such as Falconer and Mackay (1996) and Lynch and Walsh (1998).

Disciplines
Agriculture | Animal Sciences | Genetics and Genomics



Chapter 3

More Estimation of Genetic Parameters

Kenneth J. Stalder and Arnold M. Saxton

3.1 Introduction

This chapter addresses genetic parameters and estimation methods not covered in other chapters. Topics include regression designs for estimating heritability and genetic correlations, inbreeding, heterosis or crossbreeding, and realized heritability estimated from selection experiments. These techniques are from classic quantitative genetics, providing information on the genetics of populations using only pedigree and phenotypic information. The primary objective is to demonstrate how SAS software can be used to obtain this information, with minimal genetic background provided. More details are available in excellent texts such as Falconer and Mackay (1996) and Lynch and Walsh (1998).

3.2 Genetic Parameters Estimated with Regression

Regression was one of the first methods used to estimate heritability and genetic correlations, producing a direct measure of resemblance of relatives. Francis Galton used this technique when he collected statistical information from parents and their offspring (Hartl and Clark, 1989). The true relationship between any two variables may or may not be of a linear nature. Regardless of the "true" association, linear regression

can serve as a method of approximation (Lynch and Walsh, 1998). The general form of a linear regression equation is as follows:

$$y = \alpha + \beta x + e,$$

where y is the dependent or response variable, α is the y intercept value, β is the slope of the line or the regression coefficient, x is the independent or explanatory variable, and e is the residual error.

The goal of linear regression is to find estimates of intercept and slope that provide a best fit to the data. In using linear regression to estimate heritability, the independent and dependent variables are phenotypic data from parent and offspring, mid-parent and offspring, etc. As with any linear model, existence of outliers and influential points should be checked, as these can bias genetic parameter estimates. As always, care should be taken by the researcher to not overstate heritability estimates when estimates are made from a small sample from some larger population.

Regression methods for estimating genetic parameters are often preferred because the association of phenotypic records and genetic relationship among offspring and parents is easily attainable from field data. Additionally, this method of estimating genetic parameters is unbiased by parental selection. Lastly, least squares techniques used to estimate regressions are not as computationally demanding as other estimation procedures. Lynch and Walsh (1998) outline other useful properties of least squares regression analysis.

Biologically, the degree of resemblance of relatives depends on a variety of factors: the rearing environment of individuals, genetic relationships, etc. There are a variety of relationships among members of an extended family. It is reasonable to assume that closer relationship might lead to more phenotypic resemblance among relatives compared to more distant relationships. If there is no strong genetic relationship or no resemblance among relatives, then phenotype of one relative will not help predict the other.

There are a variety of regression designs that could be used to estimate genetic parameters. These include one offspring on dam, one offspring on sire, one offspring on mid-parent, mean offspring on dam, mean offspring on sire, individual offspring records on copied dam records, etc. This section will cover the more common methods of using regression to estimate genetic parameters.

The formula (Falconer and Mackay, 1996) for calculating regression of offspring on parent is

$$b_{OP} = \frac{COV_{OP}}{\sigma_P^2},$$

where COV_{OP} is the covariance of offspring on parents and σ^2_P is the parental variance.

For genetic interpretation of this statistical quantity, theory states (Hartl and Clark, 1989) that the offspring–parent slope is

$$b_{OP} = 1/2\,\frac{V_A}{V_P} = \frac{1}{2}h^2,$$

where b_{OP} is the regression of offspring on parent, ½ is used because the regression involves only a single parent, and V_A/V_P or h^2 is heritability.

An example of regression of offspring on a single parent is milk traits from daughters and dams in a dairy herd, since no records exist for the male parent. To conduct this type of study, phenotypic information needs to be collected from the parent and from the offspring. Example data for somatic cell count (SCC) from the University of Tennessee Dairy Experiment Station are read into SAS with this program:

```
data one;
   input CowName$ DamID CurrSCC CurrMilk LactNum DamSCC DamMilk DamLact;
datalines;
  TRST11    3871716   100   47.1      3       650       .       5
  ZUKR02    3878083   152   54.4      2       162      38.1     5
  GENE01    3924135    62   52.6      3        54      34.4     5
  ANCH01    3933356    38   43.5      1       162      34.4     5
  LXUS01    3933356    41   49.0      2       162      34.4     5
  BUCK01    3953108   141   32.6      2       200      54.4     5
  VIEW10    3953973   162      .      2        29      54.4     5
  DAN15     3973832    87   43.5      2        29      58.0     4
  MONT09    3973868    38   38.1      1       100      47.1     3
  HRDL03    3986622   650   36.3      2       214      49.0     4
  DCLO04    4024311   162   61.7      1      1715      23.5     3
  MONT05    4024314   348   50.8      1        31      56.2     3
  FLAG01  110128090    13   54.4      1        13      68.9     3
  JOUR02  110128317   246   38.1      1        62         .     3
  AVRY12  110128438    81      .      1       746      49.0     3
  BRTA77  110128456   152      .      1      1131      41.7     3
  DCLO01  110409807    22   38.1      1        13      47.1     3
;
proc mixed;
   model currscc = damscc /solution outp=rrr influence;
   estimate 'Heritability' damscc 2;
run;
proc univariate plot normal data=rrr;
 var resid;
run;
```

Calculation of the slope of the regression line is easily obtained with any of the linear model procedures in SAS. Here PROC MIXED is used, with the advantage of being able to estimate twice the slope, something PROC REG will not do. Also shown in the MODEL statement is creation of a data set named RRR that contains the residuals, and a request for influential diagnostics. Residuals are processed by PROC UNIVARIATE to check normality and identify outliers.

Thus, heritability can be estimated as twice the regression slope, and the standard error of the estimate is automatically provided by the software. Output 3.1 provides a heritability estimate of 1% with a standard error of 17%. Low heritability is expected for this disease-related trait, and the large standard error reflects the small experiment. Care must be taken in interpretation, as the estimates should be made from a reasonable number of parents. If too few sires or dams are used, the genetic parameter estimates obtained are likely to be biased.

Not shown in Output 3.1 are the influential diagnostics, which did not find any potential problems, and the PROC UNIVARIATE results, which did flag observation 10, with offspring SCC of 650, as a potential outlier. This point is easily seen in Output 3.2. A decision should be made to delete this observation if scientifically justified, because it increases variability and contributes to the standard error of 17%.

Output 3.1 SAS output for heritability estimation using regression of offspring on parent.

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate
Residual	25975

Fit Statistics

-2 Res Log Likelihood	213.0
AIC (smaller is better)	215.0
AICC (smaller is better)	215.3
BIC (smaller is better)	215.7

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	145.21	47.6585	15	3.05	0.0082
DamSCC	0.004835	0.08469	15	0.06	0.9552

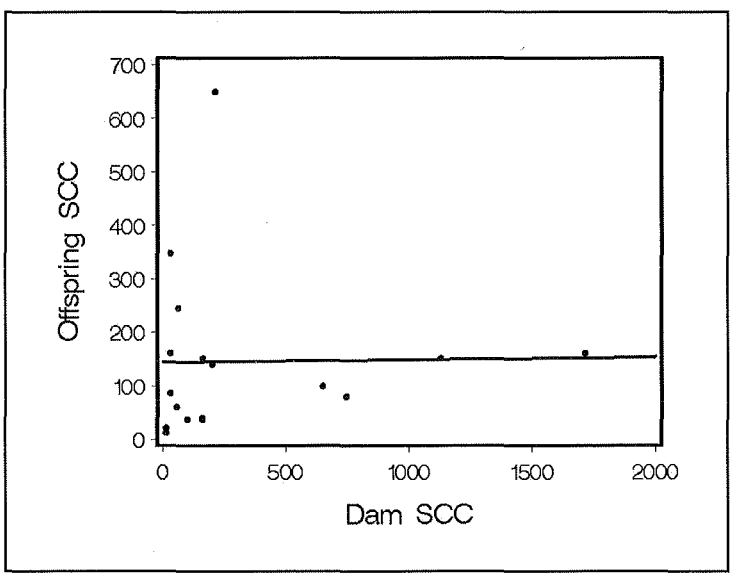
Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
DamSCC	1	15	0.00	0.9552

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Heritability	0.009670	0.1694	15	0.06	0.9552

Output 3.2 Regression of somatic cell count in dairy cows on corresponding values for their dam.



Typical SAS code and options for plotting the offspring on parent regression are shown here; these produce Output 3.2.

```
proc gplot;
 goptions ftext=swissb hpos=30 vpos=30;
 axis1 minor=none w=3 label=(a=90 'Offspring SCC');
 axis2 minor=none w=3 label=( 'Dam SCC');
 symbol1 v=dot i=rl w=3;
 plot currscc*damscc/haxis=axis2 vaxis=axis1;
run; quit;
```

Regression of offspring on mid-parent average is another method that can be used to estimate genetic parameters. This is commonly used when data are available from several offspring from a set of parents, as is the case with litter-bearing species like swine. This method assumes that one of the parents is mated to a single parent of the opposite sex. In most cases this will mean that one female is mated to a single male, but each male may be mated to several females. Assumptions are equal means and variance among the population of males and females, and autosomal inheritance of the trait (so equal resemblance between offspring and sire and dam occurs). If male and female parental means and variances are not equal, then it will be better to calculate a regression of each sex of offspring performance on each individual parent rather than on the midpoint of both parents. Methods that do handle sexes with unequal means or variances are outlined very well in Falconer and Mackay (1996).

Falconer and Mackay (1996) demonstrate that the covariance of offspring and both parents (mid-parent) can be calculated in the following manner:

$$b_{O,MP} = \frac{\text{cov}(O,[P1+P2]/2)}{V_{MP}} = \frac{\text{cov}(O,P1)/2 + \text{cov}(O,P2)/2}{[\sigma^2_{P1} + \sigma^2_{P2}]/4}$$

$$= \frac{V_A/4 + V_A/4}{V_P/2} = h^2$$

These calculations use basic statistical facts for variances and covariances and assume both parents have the same phenotypic variance. Thus the regression of offspring on mid-parent value is two times the regression of offspring on a single parent, and is a direct measure of heritability. Standard error of the heritability estimate will be the standard error of the regression coefficient.

When looking at the calculation of heritability by regression of offspring on single parent or offspring on mid-parent average, biological meaning can be slightly different. If the means and variances are equal between males and females, regression of offspring on mid-parent is likely the best estimate of "effective" heritability. It is best because it factors in both parents when estimating heritability while the heritability estimate obtained from regression of offspring on a single parent is made from either parent, but not both. There are cases such as sex-limited traits where an estimate can be obtained only from one parent and one sex of offspring.

It should be noted that when estimating heritability for a given trait from a data set, several methods can be used. These methods include the regression methods described in this chapter or other methods outlined in this book. If more than one method is used to estimate heritability from the same data set, it is likely that the estimates will differ. Falconer and Mackay (1996) estimated the heritability of abdominal bristle number in *Drosophila melanogaster* by three different methods and arrived at heritability estimates differing by .05. However, all were within range of the standard errors of the estimates. Additionally, heritability estimates are not static and as more records are added to a data set, heritability estimates can change. Similarly, heritability estimates can and often do differ depending on the population being evaluated.

Offspring–parent relationships can be used to estimate genetic correlations. To do this using the regression formula previously outlined in this section, one phenotypic character of interest must be measured in the offspring and the other phenotypic trait of interest must be measured in the parents. For example, protein content might be estimated in corn from the offspring while starch content is measured in the parental lines. The opposite could also be done (starch content measured in the offspring while protein content is measured in the parental line). If both measures are available the arithmetic mean should be used (Falconer and Mackay, 1996). The covariances for offspring and parents are needed for both traits, in this case protein and starch content. The genetic correlation then can be given as

$$r_A = \frac{COV_{XY}}{\sqrt{(COV_{XX}COV_{YY})}}.$$

Users should be aware that calculation of genetic correlations often has some undesirable characteristics. The genetic correlations frequently have large sampling errors. Because of their large sampling errors, the precision of genetic correlations is often less than desired. Additionally, genetic correlations are often population dependent because of differing gene frequencies in various populations (Falconer and Mackay, 1996) and should not be compared across different populations.

Using the somatic cell count example above, all possible regressions of offspring traits on parent traits are done with the following code, producing Output 3.3:

```
proc glm;
 model currscc currmilk=dammilk;
 estimate '2*DamMilk slope' dammilk 2;
run; quit;
proc glm;
 model currscc currmilk=damscc;
 estimate '2*DamSCC slope' damscc 2;
run; quit;
```

Slopes are multiplied by two to estimate genetic correlation for single offspring regressed on single parent, with different multipliers needed for other types of data (Falconer and Mackay, 1996). Estimates of additive genetic correlation are 2.57 and .018, clearly unstable for this small amount of data. Standard errors are produced automatically. The GLM procedure is used here, as it allows multiple dependent variables to be analyzed, whereas PROC MIXED does not. For regression models with no random effects as in these examples, PROC GLM and PROC MIXED will produce identical results.

Output 3.3 All possible regressions of two offspring traits on parent traits.

```
                          The GLM Procedure
                       Number of observations    17
                         Dependent Variables With
                    Equivalent Missing Value Patterns
                                              Dependent
                     Pattern          Obs     Variables
                           1           15     CurrSCC
                           2           12     CurrMilk
NOTE: Variables in each group are consistent with respect to the presence or
      absence of missing values.

Dependent Variable: CurrSCC
                                              Standard
  Parameter                    Estimate           Error     t Value    Pr > |t|
  2*DamMilk slope           2.57122313      7.73543873        0.33      0.7449

Dependent Variable: CurrMilk
                                              Standard
  Parameter                    Estimate           Error     t Value    Pr > |t|
  2*DamMilk slope          -0.49150166      0.40922263       -1.20      0.2574

Dependent Variable: CurrSCC
                                              Standard
  Parameter                    Estimate           Error     t Value    Pr > |t|
  2*DamSCC slope            0.00966995      0.16937735        0.06      0.9552

Dependent Variable: CurrMilk
                                              Standard
  Parameter                    Estimate           Error     t Value    Pr > |t|
  2*DamSCC slope            0.01853753      0.00947190        1.96      0.0740
```

3.3 Genetic Gain and Realized Heritability

Prediction of response to selection is described in Chapter 4. In this section, observed response to selection is used to estimate genetic gain and realized heritability. This type of information is useful to assess the effectiveness of genetic selection. If progress is too slow, changes in the selection program must be considered.

Selection experiments for a variety of traits have been and continue to be conducted. These generally have an unselected control line, used to monitor environmental changes. Selected lines may be selected in one direction or selected divergently. Falconer and Mackay (1996) provide an introduction to advantages of various designs. Provided that the selected and unselected individuals were derived from the same original population and that performance for a given trait diverges over time, realized heritability can be calculated. Realized heritability is the ratio of change in population mean per unit selection differential and can be calculated by (Van Vleck et al., 1987)

$$h^2_{realized} = \frac{\overline{P}_{PSP} - \overline{P}_{PRP}}{\overline{P}_S - \overline{P}},$$

where $\overline{P}_{PSP}$ is the performance of the progeny from selected parents, $\overline{P}_{PRP}$ is the performance of the progeny from random parents (if progeny from random parents does not exist, the population mean, P, can be used), and $\overline{P}_S - \overline{P}$ is the selection differential, or selected parent average minus parental population average.

A small example was used by Muir (1986) to illustrate statistical issues, in which *Tribolium* was selected for low body weight. This code shows the data and SAS analysis, and produces Output 3.4:

```
data one;
 input generation bw1 bw2 control;
 bw=bw1; rep=1; diff=bw-control; output;
 bw=bw2; rep=2; diff=bw-control; output;
datalines;
1  216.9  212.1  207.1
2  215.9  212.0  214.4
3  198.0  201.7  215.9
4  193.4  167.8  223.1
5  177.1  161.0  224.3
6  190.2  177.5  213.0
7  171.4  168.6  215.2
8  150.5  131.8  230.4
9  136.7  126.0  233.2
;

proc reg;    ❶
 model bw diff=generation;
 model bw = generation control;
run;

data one; set one;
 classgen=generation;
run;
proc mixed;    ❷
 class classgen rep ;
 model bw = control generation classgen /htype=1;
 random rep rep*generation;
run;

proc mixed;    ❸
 class classgen rep ;
 model bw = control generation  /htype=1 solution;
 random rep rep*generation;
run;
```

❶ A simple linear regression of response, or deviation from the control line, over generation number will estimate selection response per generation if the selection differential is constant. However, to get realized heritability, the slope must be divided by the selection differential value. Alternatively, the regression can be done using cumulative selection differential as the X variable. Results from these analyses in Output 3.4 show a selection response of –10 mg per generation if the control information is not used. Selection response is –12 mg when deviated from the control, or –6.8 mg when the control is used as a covariate, as suggested by Muir (1986). If a constant selection differential of 4 mg is assumed (i.e., the parent's body weight is 4 mg lower than the population average), then realized heritability would be 6.8/4 = 1.7, with standard error similarly calculated from the output. The 170% heritability simply reflects that for each mg that the parent body weight is lower than the average, the progeny body weight is 1.7 mg lower, something that genetically is theoretically impossible.

❷ The experiment has two replicate selected lines, and Muir (1986) suggests a more appropriate framework for testing if the selection response is different from zero. In particular, REP variation must be controlled, and a "pure error" term based on reps should be used for testing. SAS code for implementing this in PROC MIXED is given, with REPs declared as random. REP*GENERATION creates the correct error term for testing the linear regression over generations. CLASSGEN is used to create dummy variables that address all other differences among generation means, ensuring this variation does not affect statistical tests (it did affect the regression testing above). Results suggest weak evidence for a non-zero selection response (P=.11). The slope of –8.87 has been affected by the presence of CLASSGEN in the model, and this has also made the standard error unusable.

❸ By dropping CLASSGEN from the model, the selection response and standard error now match the correct regression results. Note that CLASSGEN did not greatly affect the test (P=.11), but this in general may not be true.

Output 3.4 Modeling results for genetic gain in *Tribolium* body weight.

```
                          The REG Procedure
                        Dependent Variable: bw
                          Parameter Estimates
                        Parameter       Standard
    Variable      DF     Estimate          Error    t Value    Pr > |t|
    generation     1    -10.12417        1.05611      -9.59      <.0001

                          The REG Procedure
                        Dependent Variable: diff
                          Parameter Estimates
                        Parameter       Standard
    Variable      DF     Estimate          Error    t Value    Pr > |t|
    generation     1    -12.47250        1.46788      -8.50      <.0001

                          The REG Procedure
                        Dependent Variable: bw
                          Parameter Estimates
                        Parameter       Standard
    Variable      DF     Estimate          Error    t Value    Pr > |t|
    generation     1     -6.87684        1.18723      -5.79      <.0001
    control        1     -1.38282        0.37674      -3.67      0.0023

                          The MIXED Procedure
                       Solution for Fixed Effects
                                   Standard
Effect        classgen    Estimate       Error      DF    t Value    Pr > |t|
generation                -8.8719      916359       1      -0.00      1.0000

                       Type 1 Tests of Fixed Effects
                            Num     Den
              Effect         DF      DF    F Value    Pr > F
              control         1       6     130.36    <.0001
              generation      1       1      33.61    0.1087
              classgen        6       6       1.23    0.4023

                          The MIXED Procedure
                       Solution for Fixed Effects
                                Standard
    Effect        Estimate       Error      DF    t Value    Pr > |t|
    generation     -6.8768      1.1984       1      -5.74      0.1098

                       Type 1 Tests of Fixed Effects
                            Num     Den
              Effect         DF      DF    F Value    Pr > F
              control         1      13     127.92    <.0001
              generation      1       1      32.93    0.1098
```

3.4 Inbreeding and Relationship

Inbreeding is the mating of individuals that are related by having common ancestry. Inbreeding coefficients are represented by the symbol F as defined by Sewall Wright (1922). The inbreeding coefficient represents the probability of alleles being identical by descent. F represents "fixation" of an allele, where one allele for a gene has a frequency of 100%, or complete loss of genetic variation. For an individual to be inbred, its parents must be related. Without using molecular techniques it is not possible to actually measure homozygosity changes; one can only estimate probabilities. It is important to note that individuals may have the same inbreeding coefficient, but may not be homozygous at the same loci.

Inbreeding generally has adverse effects on lowly heritable traits or those associated with fitness, known as inbreeding depression. Such traits include reproductive and survivability traits. Inbreeding effects are observed in both plants and animals, though animals generally show more inbreeding depression. The general formula for calculating inbreeding is

$$F_X = \sum\left[(1/2)^n (1+F_A)\right],$$

where F_X is the inbreeding coefficient of individual X, Σ means that summation occurs across all common ancestors, n is the number of individuals in the path connecting the sire and dam of X through the common ancestor A (including sire and dam), and F_A is the inbreeding coefficient of the common ancestor.

The inbreeding coefficient is measured relative to a particular breed or generation at a specified time. It is common to trace a pedigree back six generations or more. Hence, F represents the increase in homozygosity as a result of mating related individuals since the reference date six generations ago. However, if a pedigree can be traced back only three generations, then F represents the increase in homozygosity as a result of mating related individuals since the reference date three generations ago. It is important to note that F represents only the relationship of an individual back to some point where the parentage is known.

There are several forms of inbreeding that will result in variation in the accumulation of inbreeding in a population. Selfing, or cloning, will result in the most intense form of inbreeding. Selfing is common among some plant species; however, it is not naturally possible with animals. Full-sib or parent–offspring types of matings are the most intense form of inbreeding possible with animals. Half-sib, grandparent–grandoffspring, uncle–niece, and aunt–nephew types are equal in inbreeding and would result in less inbred individuals. Lastly, cousins could be mated together and result in even lower inbreeding coefficients than those previously described.

Relationships between any pair of individuals within a pedigree can be computed easily. Inbreeding of an individual is equal to the relationship of its parents. When inbreeding calculations are made, relationships must be available. The general form of the relationship equation, which is very similar to that seen in the calculation of inbreeding, is as follows:

$$R_{XY} = \frac{\sum\left[\left(1/2^n\right)\left(1+F_A\right)\right]}{\sqrt{\left(1+F_X\right)\left(1+F_Y\right)}},$$

where R_{XY} is the coancestry coefficient between individual X and individual Y, Σ means that summation occurs across all common ancestors, n is the number of individuals in the path connecting X and Y (inclusive), F_A is the inbreeding coefficient of the path's common ancestor, F_X is the inbreeding coefficient of individual X, and F_Y is the inbreeding coefficient of individual Y.

If Wright's relationship coefficient is needed, it is simply twice the coancestry R given here.

SAS PROC INBREED allows users to calculate the inbreeding and relationship coefficients from a defined pedigree. Inbreeding coefficients can be calculated for very large pedigree files. PROC INBREED can conduct an inbreeding analysis assuming that individuals belong either to the same generation or to non-overlapping generations. This example shows input of a simple pedigree, with codes identifying the individual and both parents, if known.

```
options ls=78;
data one;
 input  indiv mom dad;
datalines;
 5 1 .
 6 1 .
 8 5 6
 9 8 .
 10 8 .
 11 9 10
;
proc inbreed matrix ;
run;
```

A period is used to indicate missing data, as with any SAS data. Here individual 8 is from a mating of half sibs, which then produces half sibs that are mated to give individual 11. Note that individuals should be ordered from oldest to youngest, so that an individual has defined parents before it is used as a parent. Otherwise, undefined parents are automatically assigned as unknown and unrelated. PROC INBREED assumes the first three unused variables in the data set are codes for individual and parents, unless specified otherwise with the VAR statement. Results are shown in Output 3.5; note the inbreeding coefficients on the diagonal of the matrix. Individual 8 has "1" as a common ancestor, with three individuals in the chain from mother to father of "8," giving F=1/8. Individual 11 also has three individuals in the chain running through the common ancestor "8," but since "8" is inbred, the inbreeding of "11" is F=(1/8)*(1+1/8) = 9/64.

Off-diagonal elements in Output 3.5 are relationships between individuals represented by the row and column labels. The relationship values of 0.25 for full sibs and 0.125 for half sibs can be recognized. The relationship of individuals 9 and 10 is 0.140625, illustrating the fact that the inbreeding of an individual, in this case "11," equals the coancestry relationship of its parents.

Output 3.5 Genetic relationships from a sequence of half-sib matings.

The INBREED Procedure

Inbreeding Coefficients

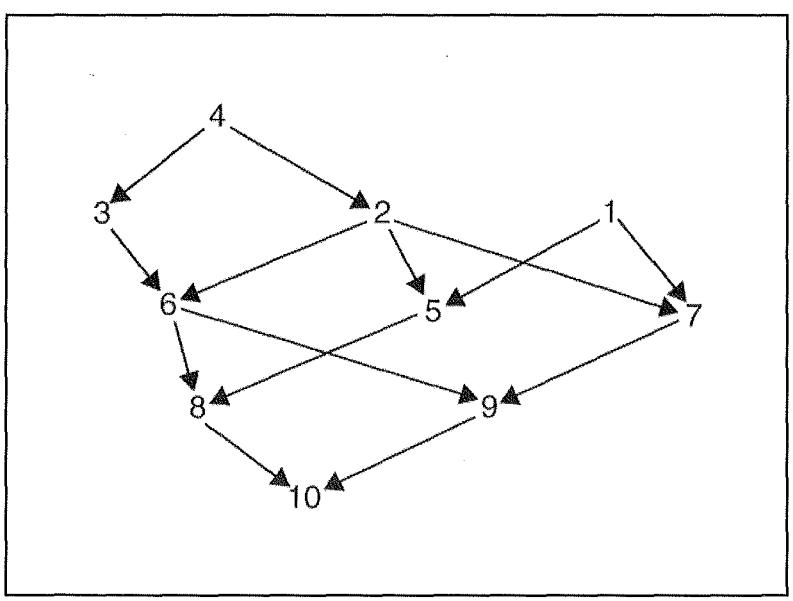
indiv	mom	dad	1	5	6	8	9	10
1			.	0.2500	0.2500	0.2500	0.1250	0.1250
5	1		0.2500	.	0.1250	0.3125	0.1563	0.1563
6	1		0.2500	0.1250	.	0.3125	0.1563	0.1563
8	5	6	0.2500	0.3125	0.3125	0.1250	0.2813	0.2813
9	8		0.1250	0.1563	0.1563	0.2813	.	0.1406
10	8		0.1250	0.1563	0.1563	0.2813	0.1406	.
11	9	10	0.1250	0.1563	0.1563	0.2813	0.3203	0.3203

Inbreeding Coefficients

indiv	mom	dad	11
1			0.1250
5	1		0.1563
6	1		0.1563
8	5	6	0.2813
9	8		0.3203
10	8		0.3203
11	9	10	0.1406

Number of Individuals 7

Figure 3.1 A complex pedigree.



As a second example, a slightly more complex pedigree, shown in Figure 3.1, is analyzed by the following SAS program.

```
data one;
 input  gen sex$ indiv mom dad cov;
datalines;
 0 F 3 4 .   .05
 0 M 2 4 .   .05
 0 F 1 . .   .05
 1 F 5 1 2   .
 1 M 6 3 2   .
 1 F 7 1 2   .
 2 F 8 5 6   .
 2 M 9 7 6   .
 3 F 10 8 9  .
;
proc inbreed ;
 var indiv dad mom cov;
 matings 3/4 , 8/9 , 2/3;
run;
proc inbreed average init=.05 matrix;
 class gen;
 gender sex;
 var indiv dad mom;
run;
```

This program includes information on generation (GEN), gender (SEX), and initial relationships (COV) existing in the base generation. The first PROC INBREED uses a VAR statement to identify the variables containing pedigree information. VAR is required here because pedigree codes are not the first variables in the data set. Additionally, the COV variable is listed, containing covariances (equals two times coancestry relationship) between parents for that individual. This is particularly useful when parents are unknown. For larger pedigrees, calculation of all relationships may not be of interest, so the MATRIX option is not used. Instead, the MATING statement is used to choose specific pairs for relationship calculation. Results from the first PROC INBREED are in Output 3.6. The relationship of parent–offspring for "3" and "4" is increased from the usual 0.25 due to the COV variable specifying that the parents of "3" are related. For the relationship of "8" and "9," the reader is challenged to identify the seven paths through the three common ancestors.

Output 3.6 Selected relationships for the Figure 3.1 pedigree.

```
                    The INBREED Procedure

          Inbreeding Coefficients of Matings
              dad     mom      Coefficient
              3       4            0.2687
              8       9            0.2991
              2       3            0.1468

            Number of Individuals     10
```

The second PROC INBREED in the program above illustrates using generation and gender codes to obtain summaries. Results for the last two generations are shown in Output 3.7. If parents in one generation cannot be used in subsequent generations, relationships among all individuals in the pedigree are not useful. Instead, PROC INBREED reports relationship matrices within generation. Summaries of relationships and inbreeding are printed for the gender combinations by specifying the AVERAGE option.

Output 3.7 Inbreeding and relationship summaries for a complex pedigree.

```
                    The INBREED Procedure

                              gen = 2
                       Inbreeding Coefficients
          indiv     dad     mom            8            9
          8         6       5         0.1772       0.2991
          9         6       7         0.2991       0.1772

      Averages of Inbreeding Coefficient Matrix in Generation 2
                                Inbreeding       Coancestry
         Male X Male                0.1772           0.0000
         Male X Female                   .           0.2991
         Female X Female            0.1772           0.0000
         Over Sex                   0.1772           0.2991

                     Number of Males          1
                     Number of Females        1
                     Number of Individuals    2

                              gen = 3
                       Inbreeding Coefficients
                  indiv     dad     mom           10
                  10        9       8        0.2991

      Averages of Inbreeding Coefficient Matrix in Generation 3
                                       Inbreeding
                  Male X Male              0.0000
                  Male X Female                 .
                  Female X Female          0.2991
                  Over Sex                 0.2991

                      Number of Males          0
                      Number of Females        1
                      Number of Individuals    1
```

Other uses for information provided by PROC INBREED may be of interest. The inbreeding coefficients can be used in other SAS procedures to account for differences in degree of inbreeding among individuals. This is particularly important when analyzing data that might be influenced by the degree of inbreeding an individual has, such as traits with low heritability. The inbreeding coefficient or the deviation from population mean inbreeding might be used as a covariate in the analysis using PROC GLM or PROC MIXED.

PROC INBREED can be used as a selection tool to examine the inbreeding that exists in a current breeding population. Additionally, PROC INBREED can be used to determine the inbreeding coefficient of any particular mating. The user could create dummy individuals, and PROC INBREED would determine inbreeding coefficients of these matings with the MATINGS statement as above. Matings that minimize the accumulation of inbreeding in the population can be selected for implementation.

Another important consideration may be to determine effective population size. The reason effective population size is discussed in this chapter is because of its relationship to the increase or buildup of inbreeding in a population. This shows the extreme effect on the increase of inbreeding that occurs in a population that will occur when working with populations that are small. With the use of artificial insemination, embryo transfer, and other reproductive technologies, the number of breeding animals needed becomes greatly reduced and affects population size. Additionally, when effective population size becomes small, there is a greater chance that genes may be lost because of random genetic drift.

Effective population size can be reduced substantially when related animals are extensively used to produce the next generation of individuals, effectively reducing the average inbreeding coefficient. However, it does not substantially affect the rate at which inbreeding accumulates (Falconer and Mackay, 1996). Effective population size, usually represented by N_e, denotes the number of individuals that would give rise to the calculated sampling variance or rate of inbreeding if the animals bred in the manner of the idealized population (Falconer and Mackay, 1996). In many mammalian breeding populations, males are allowed to mate with more than one female. This gives rise to the case where family size differences exist between males and females. When this is the case, the general form of calculating effective population size is denoted (Hill, 1979) by

$$N_e = \frac{8N}{V_{km} + V_{kf} + 4},$$

where N_e is the effective population size, N is the number of breeding individuals, V_{km} is the variance of male family size, and V_{kf} is the variance of female family size.

If the variance of the family size does not differ between males and females in the population, this equation reduces substantially. See Falconer and Mackay (1996) for deviations from this equation. The effect of effective population size on the accumulation of inbreeding in a given population is given by the approximation

$$\Delta F = \frac{1}{8N_m} + \frac{1}{8N_f},$$

where ΔF is the change in average inbreeding in a population, N_m is the number of breeding males, and N_f is the number of breeding females. In many laboratory and animal breeding experiments, it is desirable to minimize inbreeding. This can be done by appropriate choice of individuals to become parents of the next generation. This will reduce the variation in family size (V_k) in the formula to calculate effective population size. It should be noted that avoiding close matings in any one generation will reduce the accumulation of inbreeding in that generation, but it does not reduce the overall rate of inbreeding accumulation.

3.5 Heterosis, or Hybrid Vigor

The use of hybrid seed corn was popularized by former United States Secretary of Agriculture Henry Wallace. Wallace founded what is now Pioneer Hybrid International, one of the largest seed suppliers in the world. Today, many animal and plant breeders take advantage of the heterosis first described in the seed corn industry. In fact, most animals used for commercial production are the result of breed or line crosses designed to take advantage of heterosis.

Why is it important to maximize heterosis? Because it is a free source of improved performance and profits. Producers need only to develop a planned mating system in which breeds or lines of plants are chosen appropriately to capture the heterosis from crossing lines. It is important to remember that the expression of heterosis occurs only with continual crossing of pure lines or breeds. Heterosis can be maximized only when highly inbred lines or divergent breeds are crossed.

Offspring produced by crossing of inbred lines will increase the productivity of traits that were shown to suffer from inbreeding depression in the inbred lines. Heterosis, or hybrid vigor, can be described as the increased performance of crossbred offspring over the average performance of pure parents. This

phenomenon is the result of increased heterozygosity of the offspring that results from the crossing of inbred strains. The frequency of unfavorable homozygous genotypes is reduced when crossing occurs that makes the animals or plants more vigorous and adaptable to a wider range of environments. This adaptability and increased vigor results in increased performance.

A strict definition of heterosis is the difference of offspring performance from average performance of parents. There are really three types of heterosis. The first type is individual heterosis, which is described as the performance advantage of a crossbred offspring over purebred parents. The second type is maternal heterosis, which is described as the advantage of a crossbred mother over a purebred mother. The last type is paternal heterosis, described as the advantage of a crossbred father over a purebred father. This type is not as important as maternal heterosis, particularly in commercial animal production. Individual heterosis can be calculated by

$$H = \frac{\frac{(A\times B)+(B\times A)}{2} - \frac{(A\times A)+(B\times B)}{2}}{\frac{(A\times A)+(B\times B)}{2}} * 100 ,$$

where $A\times B$ and $B\times A$ represent the performance for a given trait from an individual or the mean of a group of individuals produced from the reciprocal cross of pure lines A and B, and $A\times A$ and $B\times B$ represent the performance for a given trait from an individual or the mean of a group of individuals produced from the matings of pure lines.

Maternal heterosis is the advantage of having a crossbred dam compared to a purebred dam and is usually if not exclusively observed in animal species. Maternal heterosis is exhibited one generation after the cross is made to produce the crossbred female. This is the result of better performance of traits like milking ability, number of individuals born in litter-bearing species, etc. Table 3.1 shows how maternal heterosis can be captured as compared to matings where no maternal heterosis exists. Notice that the offspring in every case captures 100 percent of the individual heterosis. Paternal heterosis can be estimated in a similar manner. However, paternal heterosis is not as important in the commercial livestock industries as a whole as it once was, because of the widespread use of artificial insemination.

Table 3.1 Example matings illustrating occurrence of maternal heterosis.

Maternal Line	Sire Line	Offspring	Amount of Maternal Heterosis (%)
$B\times B$	A	$A\times(B\times B)$	0
$C\times C$	A	$A\times(C\times C)$	0
$B\times C$	A	$A\times(B\times C)$	100
$C\times B$	A	$A\times(C\times B)$	100

The analysis of this type of data can easily be done using PROC GLM or PROC MIXED and the ESTIMATE statement. As the formulas above suggest, heterosis calculations are simply a series of comparisons among means. As an example, an experiment studying five corn lines and their first generation crosses was conducted by Dr. Dennis West at the University of Tennessee. The study did not include reciprocal crosses, meaning if the cross Male 1 by Female 2 was made, then the cross Female 1 by Male 2 was not. Parts of the program are shown here, with only estimates involving the first three lines shown to save space:

```
data one;
 input plot entry rep  year loc$ par1 par2 lodge height earht standpcnt buyield
kgyield;
datalines;
108 1      1      1998   KnoxTN 1  2   15.6   1.98   1.12   94     121.2  7604
207 2      1      1998   KnoxTN 1  3   16.4   2.37   1.31   99     132.4  8303
106 3      1      1998   KnoxTN 1  4   6      2.22   1.31   99     141.5  8871
109 4      1      1998   KnoxTN 1  5   7.5    2.16   1.28   99     139.1  8722
...more datalines...
;
proc mixed data=one;
 class par1 par2 year loc rep;
 model kgyield = par1*par2;    ❶
 random loc year(loc) rep*year(loc);    ❷
 estimate 'pure line mean' intercept 5 par1*par2 1 0 0 0 0   1 0 0 0   1 0 0  1
0   1/divisor=5 ;    ❸
 estimate 'avg heterosis' par1*par2  -4 2 2 2 2   -4 2 2 2   -4 2 2  -4 2   -
4/divisor=20;
 ** avg of 4 heterosis values per line ;
 estimate 'heterosis 1' par1*par2  -4 2 2 2 2  -1 0 0 0   -1 0 0  -1 0   -
1/divisor=8;
 estimate 'heterosis 2' par1*par2  -1 2 0 0 0  -4 2 2 2   -1 0 0  -1 0   -
1/divisor=8;
 estimate 'heterosis 3' par1*par2  -1 0 2 0 0  -1 2 0 0   -4 2 2  -1 0   -
1/divisor=8;
** deviation of line heterosis from avg heterosis;
estimate 'dev heterosis 1' par1*par2  -12 6 6 6 6   3 -4 -4 -4   3 -4 -4  3 -4
3/divisor=40;
estimate 'dev heterosis 2' par1*par2  3 6 -4 -4 -4  -12 6 6 6   3 -4 -4  3 -4
3/divisor=40;
estimate 'dev heterosis 3' par1*par2  3 -4 6 -4 -4   3 6 -4 -4  -12 6 6  3 -4
3/divisor=40;
***** cross heterosis;
 estimate 'hij 1-2'  par1*par2 -1 2 0 0 0  -1 /divisor=2;
 estimate 'hij 1-3'  par1*par2 -1 0 2 0 0   0 0 0 0  -1/divisor=2;
estimate 'hij 2-3'  par1*par2  0 0 0 0 0  -1 2 0 0  -1/divisor=2;
***** specific heterosis;
 estimate 'Sij 1-2' par1*par2 0   12   -4   -4   -4  0   -4   -4   -4  4  0  0
4  0  4   /divisor=20;
 estimate 'Sij 1-3' par1*par2  0   -4   12   -4   -4  4   -4  0  0  0   -4   -4
4  0  4   /divisor=20;
estimate 'Sij 2-3' par1*par2  4   -4   -4  0  0  0   12   -4   -4  0   -4   -4
4  0  4   /divisor=20;

 lsmeans par1*par2;
run;
```

❶ After creating the working SAS data set ONE, a mixed model analysis is used to analyze the data. In order to work with each of the cross means, the model has only the interaction term of the two parent lines. Main effects of parents could be included in the model, but these would make the ESTIMATE statements that follow much more complex.

❷ As dictated by the experimental design, any random effects must be addressed. Here YEAR, LOCATION, and REP blocking terms are used to remove those sources of variation. The presence of random effects makes PROC MIXED the best choice for statistical analysis.

❸ ESTIMATE statements are used to produce comparisons of interest. After a label in quotes, coefficients are assigned to each cross mean (PAR1*PAR2) as needed to produce the desired information. Note that when a mean is being estimated, a coefficient for the intercept is needed in addition to the cross means. The DIVISOR option requests all coefficients be divided by the given number, allowing awkward fractions to be easily entered. As with any ESTIMATE or CONTRAST, order of the coefficients is critical, as the correct coefficient must be matched with the corresponding cross. An easy way to verify cross order is to request least squares means, the order there being identical. Besides estimated values of heterosis, variability explained by comparisons may be of interest. These can be investigated using CONTRAST statements, using a similar set of coefficients.

Hallauer and Miranda (1988) is a good general reference for diallel experiments, but Gardner and Eberhart (1966) should be consulted for statistical details. For more information on variance component estimation, see Chapter 2, "Estimation of Genetic Variances and Covariances by Restricted Maximum Likelihood Using PROC MIXED." In general terms, cross heterosis is as defined above: the deviation of individual cross mean from parental line means. All cross heterosis values for a line can be averaged to give line heterosis, the benefit of using that line in crosses. Then line heterosis values can be averaged to give the overall average heterosis. If the two parental line and average heterosis values are subtracted from cross heterosis, what remains is specific heterosis, heterosis that is specific to this cross above that expected from the two lines involved.

Output 3.8 contains the results from the diallel program, with estimates of the various types of heterosis. Again to save space only the first three line values are reported. On average, crosses gave 1749 kg/ha more yield than parents, and crosses involving line 2 averaged 2021 kg/ha more yield. However, the cross of line 2 with line 1 gave a 2317 kg/ha increase, with only 671 kg/ha of that being specific to the line 1 and 2 cross.

Diallel experiments such as this example are usually designed to include either (1) crosses, (2) crosses and parents, or (3) crosses, parents, and reciprocals. Increased information is available as more relatives are included. Naturally ESTIMATE statements will differ across these designs, and even within a design different ESTIMATE coefficients are needed depending on the number of lines involved. A crude beginning of a PROC IML program is included in the example file that will automatically generate ESTIMATE coefficients for "crosses and parents" experiments. It would be a welcome contribution for a reader to develop an easy-to-use macro for all experimental situations.

This type of analysis can also be used for a variety of crossing systems in animals. Commonly used crossing systems include terminal crosses, where the offspring are destined for market and not retained for further breeding purposes. There are several types of rotational systems that are commonly used, particularly in the commercial livestock industry. These systems include two-breed, three-breed, and four-breed crosses. These rotational systems do not obtain maximum heterosis, but they do have the advantage that replacement females are raised within the system.

Output 3.8 Partial results from the diallel analysis of heterosis.

The Mixed Procedure

Class Level Information

Class	Levels	Values
par1	5	1 2 3 4 5
par2	5	1 2 3 4 5
year	2	1998 1999
loc	5	ColumMO KnoxTN LexingKY MilanTN QuickKY
rep	3	1 2 3

Covariance Parameter Estimates

Cov Parm	Estimate
loc	0
year(loc)	922161
year*rep(loc)	207492
Residual	1109238

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
par1*par2	14	238	16.12	<.0001

Label	Estimate	Estimates Standard Error	DF	t Value	Pr > |t|
pure line mean	4695.53	421.36	238	11.14	<.0001
avg heterosis	1749.13	135.97	238	12.86	<.0001
heterosis 1	1715.08	186.18	238	9.21	<.0001
heterosis 2	2020.97	186.18	238	10.85	<.0001
heterosis 3	1508.86	186.18	238	8.10	<.0001
dev heterosis 1	-34.0514	127.19	238	-0.27	0.7891
dev heterosis 2	271.84	127.19	238	2.14	0.0336
dev heterosis 3	-240.27	127.19	238	-1.89	0.0601
hij 1-2	2317.53	304.03	238	7.62	<.0001
hij 1-3	1270.47	304.03	238	4.18	<.0001
hij 2-3	1408.78	304.03	238	4.63	<.0001
Sij 1-2	-671.31	210.64	238	-3.19	0.0016
Sij 1-3	-1308.68	210.64	238	-6.21	<.0001
Sij 2-3	-1415.09	210.64	238	-6.72	<.0001

The amount of hybrid vigor obtained in a rotational crossbreeding system at equilibrium using purebred or pure-line sires can be predicted by (Bourdon, 2000):

$$\%HybridVigor = \left(\frac{2^n - 2}{2^n - 1}\right) \times 100 \ ,$$

where n is the number of pure breeds or lines involved in the rotational system. When crossbred sires are used in the rotational system, the equation to predict equilibrium hybrid vigor differs (Bourdon, 2000) and is

$$\%HybridVigor = \left(\frac{m\left(2^n - 1\right) - 1}{m\left(2^n - 1\right)}\right) \times 100 \ ,$$

where n is the number of sire types involved in the system and m is the number of breeds present in each sire type. This formula assumes that no breed is present in more than one sire type.

It should be noted that the phenomenon of heterosis is generally lost if two F_1 individuals derived from the crossing of two pure line parents are crossed. The offspring produced from the mating of two F_1 individuals will often exhibit decreased performance values for the traits that exhibited the heterotic effect in the F_1 individuals. In other words, the superior performance observed in crossbred individuals is not transmitted upon mating. This is because the gene combinations are not transmitted to progeny; only individual genes are transmitted to progeny. The gene combinations are rearranged or lost when crossbred animals are mated together, because of random segregation of alleles during meiosis. Additionally, the crossing of different species can often result in reduced reproductive performance as exhibited by the sterility of offspring produced by crossing a horse and an ass.

3.6 References

Bourdon, R. M. 2000. *Understanding Animal Breeding*. 2d ed. Upper Saddle River, NJ: Prentice-Hall.

Falconer, D. S., and T. F. C. Mackay. 1996. *Introduction to Quantitative Genetics*. 4th ed. New York: John Wiley & Sons.

Gardner, C. O., and S. A. Eberhart. 1966. Analysis and interpretation of the variety cross diallel and related populations. *Biometrics* 22:439-452.

Hallauer, A. R., and J. B. Miranda Filho. 1988. *Quantitative Genetics in Maize Breeding*. 2d ed. Ames: Iowa State University Press.

Hartl, D. L., and A. G. Clark. 1989. *Principles of Population Genetics*. 2d ed. Sunderland, MA: Sinauer Associates.

Hill, W. G. 1979. A note on effective population size with overlapping generations. *Genetics* 92:317-322.

Lynch, M., and B. Walsh. 1998. *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

Muir, W. M. 1986. Estimation of response to selection and utilization of control populations for additional information and accuracy. *Biometrics* 42:381-391.

Van Vleck, L. D., E. J. Pollak, and E. A. Branford Oltenacu. 1987. *Genetics for the Animal Sciences*. New York: W. H. Freeman.

Wright, S. 1922. Coefficients of inbreeding and relationship. *Amer. Naturalist* 56:330-339.