Random effects model (One factor with random effects)
Example of use of intraclass correlation.

For each gene in a particular genetic line of a plant, we have 3 measurements of gene expression (from three genetically identical organisms). There are 10 distinct genetic lines, and there are 12,488 genes altogether on which we have information.

We want to use the intraclass correlation value to filter the 12,488 genes down to those that have a ‘high’ intraclass correlation (these genes have much of their variability described by genetics, rather than environmental variability).

SAS statements and output:

```sas
proc import datafile="intraclass_corr.csv" out=icc dbms=csv replace;
run;

proc print data=icc (obs=10);
run;
```

```
Obs  Gene  Line  Rep  Expression
1    1     1     1    10.807961556
2    1     1     2    14.622119964
3    1     1     3    6.7486266186
4    1     2     1    -1.977835625
5    1     2     2    -0.713574748
6    1     2     3    -2.645149254
7    1     3     1    14.009476457
8    1     3     2    10.169057997
9    1     3     3    16.387167435
10   1     4     1    0.3552423764
```

Because I’m going to do the same analysis at each gene, I will make sure my data is sorted by gene first.

```sas
proc sort data=icc;
  by Gene;
run;
```

When working with random effects, we will usually use PROC MIXED, though occasionally we will have a reason to use PROC GLM. Here, besides using PROC MIXED, I have saved some of the PROC MIXED output by utilizing an Output Delivery System (ODS) statement. After running the procedure, my ODS statement will create a new data set called ‘covparm’ which holds the estimates for \( \sigma^2 \) and \( \tau^2 \) for each gene. These are the variance components used in the random effects model (there’s a separate such model fit for each gene here). I’m going to use these estimates to estimate the intraclass correlation.
ods output "Covariance Parameter Estimates"=covparm;
proc mixed data=icc;
   class Line;
   by Gene;
   model Expression=;
   random Line;
run;
quit;
ods output close;

Gene=1
The Mixed Procedure
Model Information

Data Set WORK.icc
Dependent Variable Expression
Covariance Structure Variance Components
Estimation Method REML
Residual Variance Method Profile
Fixed Effects SE Method Model-Based
Degrees of Freedom Method Containment

Class Level Information
Class Levels Values
Line 10 1 2 3 4 5 6 7 8 9 10

Dimensions
Covariance Parameters 2
Columns in X 1
Columns in Z 10
Subjects 1
Max Obs Per Subject 30

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line</td>
<td>30.4319</td>
</tr>
<tr>
<td>Residual</td>
<td>9.3871</td>
</tr>
</tbody>
</table>
proc print data=covparm (obs=10);
run;

<table>
<thead>
<tr>
<th>Obs</th>
<th>Gene</th>
<th>CovParm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Line</td>
<td>30.4319</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Residual</td>
<td>9.3871</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Line</td>
<td>39.7134</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>Residual</td>
<td>6.7780</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>Line</td>
<td>87.8875</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>Residual</td>
<td>4.3855</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>Line</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>Residual</td>
<td>9.1654</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>Line</td>
<td>4.0199</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>Residual</td>
<td>6.6496</td>
</tr>
</tbody>
</table>

Now, I have an estimate for $\sigma^2_\tau$ and $\sigma^2$ at each gene.

One thing that we can appropriately use PROC GLM for when utilizing random effects is to the get the expected mean squares of a random effects factor. **BUT PROC GLM CAN DO THE WRONG THING WITH RANDOM EFFECTS**, so if we have random effects, in general, we will use PROC MIXED.

proc glm data=icc;
  class Line;
  by Gene;
  model Expression=Line;
  random Line/q;
run;

The GLM Procedure
Class Level Information
Class Levels Values
Line 10 1 2 3 4 5 6 7 8 9 10

Gene=1
The GLM Procedure
Dependent Variable: Expression

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9</td>
<td>906.14444444</td>
<td>100.682716</td>
<td>10.73</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>187.741545</td>
<td>9.387077</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>29</td>
<td>1093.885989</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source    Type III Expected Mean Square
Line      $\text{Var(\text{Error})} + 3 \text{Var(\text{Line})}$
The histogram of the 12,488 estimated intraclass correlations:

There were about 4500 genes with highly heritable expression ($\rho_I \geq 0.80$).