***Investigating genetic profiles of fatty acids and cardiometabolic traits***

***Learning goals:***

* Contrasting linear and non-linear genetic models and connecting to disease modes of inheritance

***Assumes***

* Basic understanding of ANOVA

**STEP 1:** **Ask a research question.**

More and more evidence exists that circulating fatty acids (in the blood) are associated with cardiovascular and brain health. Some fatty acids are obtained through diet while others are synthesized by the body from dietary fatty acids and other sources. Understanding dietary, genetic and other factors that explain variation in fatty acid levels is a current research question of interest.

One gene that has been recently identified in fatty acid metabolism is FADS1 (fatty acid desaturase 1, on chromosome 11). Some people have a particular genetic variant in the FADS1 gene, having cytosine (C) instead of thymine (T) (at polymorphism rs174547). This genetic variant may be inherited from either your mother or your father, both parents, or not at all. Thus, we can summarize this genetic variant as having either 0 (neither mother nor father), 1 (from the mother or father) or 2 copies (from both mother and father) of cytosine (C).

Researchers want to know: Is this genetic variation in FADS1 associated with the Arachidonic Acid to di-homo-gamma-linolenic fatty acid ratio in the blood?

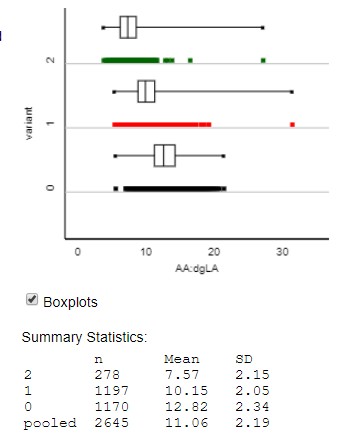
**STEP 2: Design a study and collect data.**

The Framingham Heart Study is a well-known study of cardiovascular health. In a sample of 2,645 individuals from the Framingham Heart Study we have measured genetic variations of the FADS1 gene, as well as the Arachidonic Acid to di-homo-gamma-linolenic (AA %:dgLA %) fatty acid ratio. Lower values of this ratio of fatty acids are better as lower values indicate greater synthesis of the fatty acids by the body.

**STEP 3: Explore the data.**

We used the ***Multiple Means applet*** (available at [www.isi-stats.com](http://www.isi-stats.com)) to investigate the relationship between the number of copies of the genetic variant (0, 1 or 2) with the AA to dgLA ratio.

**Figure 1. Distribution of the ratio AA:dgLA by number of variants**



As we see in Figure 1, there is a strong relationship between the number of variants (0,1 or 2) and the AA:dgLA ratio, as well as a few extreme outliers (values of the AA:dgLA ratio above 25).

|  |
| --- |
| **Think about it:** What statistical method could we use to explore how much of the variation in AA:dgLA ratios are explained by the rs174547 variant within FADS1? |

Analysis of variance is one approach to explore how much of the variation in the AA:dgLA ratio is explained by the rs174547 variant in the FADS1 gene. Table 1 illustrates the sum of squares and degrees of freedom for this model.

**Table 1. Analysis of variance table**

|  |  |  |
| --- | --- | --- |
| **Source** | **df** | **SS** |
| **FADS1** | 2 | 8028.94 |
| **Error** | 2642 | 12683.01 |
| **Total** | 2644 | 20711.95 |

Table 1 illustrates that a substantial amount of the variation in the AA:dgLA ratio is explained by variation within FADS1 (the explained sum of squares is 8028 and the unexplained is only 12683). This is further visualized in Figure 2.

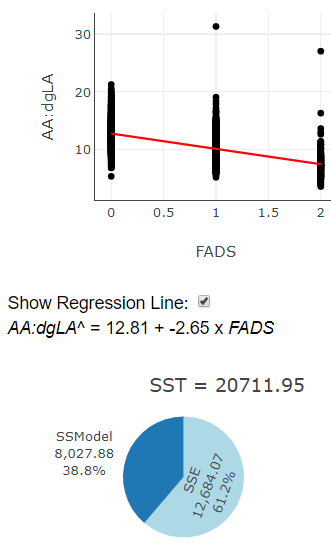
**Figure 2. Pie chart of variation explained**

**Contrasting this approach with the linear regression model**

Not only does the fatty acid ratio differ based on the genetic variant, but the association appears to be *linear* – the means of the three distributions could be modelled as following on a line. One downside of the ANOVA approach shown above is that it does not provide numeric values that directly connect the statistical model to the genetic model.

One alternative is to use a linear model (simple linear regression). We used the Multiple Regression applet (available at [www.isi-stats.com](http://www.isi-stats.com)) to calculate the linear regression model for this data using the number of copies of the variant (0, 1 or 2) as the predictor and the AA:dgLA ratio as the response. The output is seen in Figure 3 below.

**Figure 3. Regression line predicting the AA:dgLA ratio by FADS1 genetic variant**



A few important things are worth noting about this model. First, we see that the variance explained by the model is (SS=8027.88 with linear model vs. 8028.94 from the ANOVA model; the SSModel (or R2) is 38.8)%) is nearly identical. Why is the ANOVA model slightly better? The ANOVA model will always predict each group mean exactly (see Table 2 below) because it is fitting a separate means model. The regression model has an advantage though, in terms of interpretation. With the regression model we have built in language, the slope, which helps us describe how the fatty acid ratio is related to the genetic variant.

**Table 2. Actual and predicted values of the AA:dgLA ratio**

|  |  |  |  |
| --- | --- | --- | --- |
| **FADS variant** | **Actual means** | **Predicted by ANOVA** | **Predicted by linear regression** |
| **0** | 12.82 | 12.82 | 12.81 |
| **1** | 10.15 | 10.15 | 10.16 |
| **2** | 7.57 | 7.57 | 7.51 |

Arguably, there is more interpretability from the linear model because the slope of the model (-2.65) means something.

|  |
| --- |
| Think about it: What does the slope of the linear model (-2.65) mean genetically? |

Because the slope is negative, we understand there is a decreasing relationship between fatty acid ratio and the number of copies of the genetic variant. As the number of copies of the genetic variant goes up, the fatty acid ratio goes down. In addition, using the value of the slope, we can quantify the rate of change in the fatty acid ratio. We predict a decrease of 2.65 in the fatty acid ratio for each additional copy of the genetic variant. This suggests an ‘additive mode of inheritance’ such that the protein made by FADS1 is more efficient with the variant and the more copies of the variant you have, you are able to convert more and more AA to dgLA.

It makes sense in this study to think of the explanatory variable, genetic variant, as quantitative since not inheriting, inheriting from your mother or father, and inheriting from both has a natural quantitative meaning for the number of copies of the variant. In addition, because the response variable, fatty acid ratio, has a linear relationship with the number of copies of the genetic variant, in simplifying from the separate means model to the linear model we don’t lose any information. Both the R2 and SE residual for the linear model are very similar to those of the separate means model. And by simplifying to the regression model we gain interpretability.

**Step 4: Draw Inferences beyond the data.**

But there is another key aspect (beyond interpretability) as to why the linear model may be better in this case. The full ANOVA table is shown below. In this case, the explained variation is highly statistically significant (p<0.0001), and the F-statistic is 836.255.

**Table 3. Analysis of variance table for separate means model**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **df** | **SS** | **MS** | **F** | **p-value** |
| **FADS1** | 2 | 8028.94 | 4014.47 | 836.255 | <0.0001 |
| **Error** | 2642 | 12683.01 | 4.80 |  |  |
| **Total** | 2644 | 20711.95 |  |  |  |

However, when we contrast this with the linear model (see Table 4), the F statistic is higher. In this case, it doesn’t matter much in terms of statistical significance, however in some cases it could.

**Table 4. Analysis of variance table for linear regression model**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **df** | **SS** | **MS** | **F** | **p-value** |
| **FADS1** | 1 | 8027.88 | 8027.88 | 1672.82 | <0.0001 |
| **Error** | 2643 | 12684,07 | 4.799 |  |  |
| **Total** | 2644 | 20711.95 |  |  |  |

|  |
| --- |
| **Think about it:** Why is the F-statistic larger for the linear model? |

The F-statistic is larger for the linear because the model is simpler. Statistically, the more complex your model is the you more you ‘pay’ ---sometimes the extra complexity is worth it because you explain a lot more variation. However, sometimes the extra complexity isn’t worth it because you don’t explain a lot more variation. This is the case here.

**Extensions we anticipate developing**

1. Testing hardy-weinberg equilibrium with this data using a chi-square GOF
2. Using simulation to model null distributions/having students brainstorm the statistics to evaluate different hypotheses and understand the more ‘restrictive’ null for the linear model
3. Evaluating a gene-diet interaction (see below for quick synopsis) using a two-variable, interaction model.

**Potential extension**

Fractional shortening is associated with reduced heart function. We found a significant statistical interaction between genetic variation in the FFAR4 gene and the fatty acid EPA on fractional shortening values within the Framingham Heart Study.