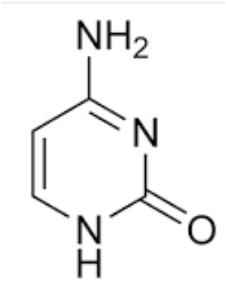
Exploration 4.1: Fatty Acids and DNA

Learning Goals:

- Describe the association between two quantitative variables numerically and graphically
- Interpret least-squares regression models between two quantitative variables
- Compare and contrast separate means vs. linear regression models



More and more evidence exists that fatty acids circulating in the blood are associated with cardiovascular and brain health. Some fatty acids are obtained through diet whereas others are synthesized by the body from dietary fatty acids and other sources. One gene that has recently been identified in fatty acid metabolism is FADS1 (fatty acid desaturase 1, on chromosome 11). Some people have a genetic variant in the FADS1 gene, such that at one place in the gene the person has the base cytosine (C) instead of thymine (T). This "T to C" genetic variant may be inherited from either the mother or the father, both parents, or not at all. Thus, people can have either 0 (neither mother nor father), 1 (from the mother or father) or 2 copies (from both mother and father) of cytosine (C) in the FADS1 gene.

STEP 1: Ask a research question.

In this study, the researcher wished to know whether the cytosine genetic variation in FADS1 is associated with the metabolism of fatty acids in the blood as measured by the Arachidonic Acid to di-homo-gamma-linolenic fatty acid ratio. Lower values of this ratio indicate higher amounts of fatty acids in the blood and thus, more efficient use of dietary fatty acids.

STEP 2: Design a study and collect data.

The Framingham Heart Study is a well-known study of cardiovascular health. In a sample of 100 individuals from the Framingham Heart Study researchers determined the cytosine genetic variation in the FADS1 gene, as well as the Arachidonic Acid to di-homo-gamma-linolenic (AA %:dgLA %) fatty acid ratio in the blood.

STEP 3: Explore the data.

Copy/paste the data **fattyAcid** data into the Multiple Variables applet (or just type fattyAcid.txt in the data window and press Use Data twice). Create a histogram and numerical summaries for the FattyAcidRatio variable.

1. Describe the shape, center, and variability of the distribution of the fatty acid ratio in the blood of these 100 study participants.

These materials were developed by the STUB Network and supported by the National Science Foundation under Grant NSF-DBI 1730668. They are covered under the Creative Commons license BY-NC which allows users to distribute, adapt, and build upon the materials for noncommercial purposes only, and only so long as attribution is given to the STUB Network.

Subset by the categorical variable Cytosine and obtain the summary statistics for the one-variable separate means model, and check the **ANOVA Table** box. Summarize or include a screen capture of your output below.

2. Is there an association between the fatty acid ratio and the genetic variantion in this sample? Describe the nature of this association. Does it appear to be better (in terms of lower ratios) to inherit the cytosine variant from both parents, either (one) parent, or neither parent?(What would the histograms look like if the number of parents was not related to the fatty acid ratios? What if Cytosine explained all of the variation in fatty acid ratios?) Explain.

3. What proportion of the variation in the fatty acid ratio is explained by the cytosine variable? What is the *SSModel* value? Is the association between the fatty acid ratio and the genetic variant statistically significant (cite both the p-value and the *F*-statistic)?

4. Write out the prediction equation for the separate means model. Check Show residual box. What is the value of the residual SE? What is the value of the *SSError*?

5. What does this model predict for the fatty acid ratio when cytosine is inherited from neither parent, 1 parent, or both parents?

Neither parent:

One parent:

Both parents:

6. What do you notice about the change in the predicted ratio as you move from 0 to 1 to 2 parents? In other words, how big is the change from 0 to 1 as compared to the change from 1 to 2?

The Linear Regression Model

Your observation that the change from 0 to 1 is very similar to the change from 1 to 2 parents suggests that another way we can model this relationship is with a line. To explore the possible linear association between these two variables, we will first examine a *scatterplot* of two quantitative variables. So rather than phrasing our observations in terms of inheritance from neither, one, or both parents we will simply refer to 0, 1, or 2 copies.

Return to the applet and remove the *Cytosine* variable from the Subset by box. Drag the *Copies* variable into the Explanatory box.

7. Do you see a general trend between the Fatty Acid Ratio and the number of copies? Increasing or decreasing? Would it be reasonable to call it a linear trend?

You may recall that one way to summarize a linear trend is to fit a *least squares regression line* that minimizes the sum of the squared residuals from the response values to the values predicted by the line.

Check the **Show Equation** box to see the least squares regression line, which passes (as closely as possible) through the means of the responses for each value of the number of copies.

8. Record the equation below, using good statistical notation.

Interpreting the Regression Model

Definition: The statistical equation for a line, called the ***regression line***, is generally written as: $\hat{y} = b_0 + b_1 x$, where b_0 is the *y*-intercept of the line, b_1 is the slope of the line, and $\hat{y}$ is the predicted value of the response for a particular value of *x*.

9. Identify the *y*-intercept and the slope in the fatty acid regression equation provided (include units).

y-intercept:

slope:

Key Idea:

- The intercept is the predicted response when the value of the explanatory variable is equal to 0, but this is not always a useful or meaningful prediction; it is often *extrapolation* (using the model to predict responses for explanatory variable values outside of the range of x-values in the sample)
- The slope is the change in predicted response when the value of the explanatory variable increases by one unit. A key feature of the linear model is that the interpretation of the slope doesn't depend on what two *x* values you are talking about, as long as they are one unit apart.

10. In the context of this study, write an interpretation for the *y*-intercept.

11. Use the regression model to predict the fatty acid ratio when someone has 0, 1, and 2 copies of cytosine.

0 parents: 1 parent: 2 parents:

What is the difference in the predicted values as you move from 0 to 1 and 1 to 2 parents?

12. In the context of this study, write an interpretation for the slope.

Key idea: When the explanatory variable is quantitative, the regression model makes predictions based on the assumption of a linear change in the predicted response that is associated with a one-unit change of the explanatory variable.

Evaluating the Regression Model

13. How do the predictions of the Fatty Acid Ratio made with the regression model compare to what you found earlier when we used the separate means model? What are the implications of this?

The predictions from the separate means model match the mean response for each of the 3 possible number of parents exactly, but the linear model seems to be doing almost as well! More quantitatively, the *SSError* and the *Residual SE* both give summary information about how well the separate means model compares to the regression model. The *SSError* and *Residual SE* are both based on the residuals. We can calculate the *residual* for each person by comparing the observed ratio to the predicted ratio.

14. Determine the residual for the first person in the data set. Remember: the residual is the observed response minus the predicted response. Interpret this residual in the context of this study.

If we determine the residual for each individual in the data set and then sum the squared residuals, we get the *SSError* = $\sum(y_i - \hat{y}_i)^2$. Taking the square root of the *MSError* gives the residual SE, or the size of a typical residual in this data set.

15. Fill in the table below with information from the ANOVA tables for both the separate means and linear models and the residual SE values for each model. (The standard error of the residuals for the regression model can obtained by checking the Regression SE box, by looking at the distribution of the model residuals, or by taking the square root of *MSError*.)

	DF	SSModel	SSError	MSError	Residual SE
Separate Means Model					
Linear Regression Model					

16. Compare the *SSError* from the linear model to the *SSError* of the single means model you computed earlier. How close are they? How do the R^2 values for the two models compare?

17. How does the residual standard error for the linear model compare to the separate means model?

18. What are the implications of having similar residual standard errors and explained variation in both models?

Which Model is Better?

The separate means model using the treatment groups will always predict the observed group means exactly. Because of this, in terms of *SSError*, linear regression can never do better than the separate means model. The linear regression model is a simplification of the separate means model. But when the treatment group means have a linear relationship with the explanatory variable, the predictions from the linear regression model will be very close to the observed group means. In this case, the two models are practically equivalent, with almost identical residual SE and R^2 values.

Key Idea: When two models have similar *SSError* values, choose the model that is simpler (e.g., uses fewer degrees of freedom). You can also favor the model that makes more sense in context (e.g., agrees with past research studies). There is not always "one best" model but you may want to explore a couple in more depth.

The linear regression model has an advantage in terms of interpretation. With the regression model we have a built-in statistic, the slope, which helps us describe *how* the fatty acid ratio is related to number of cytosine copies. Because the slope is negative, we can conclude there is a *decreasing* relationship between the fatty acid ratio and the number of copies of cytosine. As the number of copies of cytosine goes up, the predicted fatty acid ratio goes down (a good thing). In addition, using the value of the slope, we can quantify the rate of change in the predicted fatty acid ratio. We predict a decrease of 2.64 in the predicted fatty acid ratio for each additional copy of the genetic variant.

Assessment of Linearity

To complete our assessment of the appropriateness of the regression model for these data we should examine plots of the residuals. The residuals represent the variation left-over in the response variable after adjusting for the linear trend. If we see a pattern in those residuals vs. our explanatory variable, then this tells us that the linear model is not appropriate (or at least that there could be a better model out there).

Key Idea: Using the linear regression model requires that the association between the explanatory variable and the response variable is linear. This means that after fitting the regression line (removing the linear trend in the data), the residuals should exhibit no patterns when plotted against the explanatory variable or against the predicted (aka fitted) values.

Check the **Show Residuals** box and either sketch or include a screen capture of the residuals vs. predicted values graph.

19. Does this graph exhibit any sort of pattern? That is, do any of the "stacks" sit mostly above or below the horizontal line at a residual of 0, or does each stack of residuals appear to be mostly randomly scattered about the horizontal line at a residual of 0? Do the "width" of the stacks seem approximately the same?

STEP 4: Draw Inferences beyond the data.
This step, which will help assess whether the observed linear association could have happened "by chance alone" as well as how to use confidence intervals for the slope and for the predicted values, will be discussed in the second part of this module. Do keep in mind that even though we may think the association is reasonably linear, it would not make sense in context to use this model to make predictions for values other than 0, 1, or 2 copies as those are the only possible explanatory variable values in this context.

STEP 5: Formulate conclusions.

In this sample of participants, we found that each additional copy of cytosine was associated with a decline in the average fatty acid ratio. Because we could translate the number of copies to an ordinal, numeric variable and that variable had a linear relationship with the response variable (residual plots look relatively random), it makes sense to treat the number of copies as a quantitative variable and use a linear model to predict the fatty acid ratio. We found that in simplifying from the separate means model to the linear model we don't lose much information (predictive ability). Both the R^2 and SE residual for the linear regression model are very similar to those of the separate means model (around 36% of the variation in fatty acid ratios can be explained by cytosine). And by simplifying to the regression model we gain interpretability: each associational copy of cytosine predicts a decrease of 2.64 in the fatty acid ratio. With fatty acid ratios ranging from 5 to 20, a decrease of 2.64 seems meaningful.

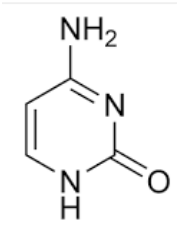
STEP 6: Look back and ahead.

Due to the relatively large R^2 value and somewhat large sample size, you probably predict that this association is statistically significant. Do be cautious that you can't draw any cause and effect conclusions from this observational study.

Exploration 4.2: Fatty Acids and DNA (contd.)

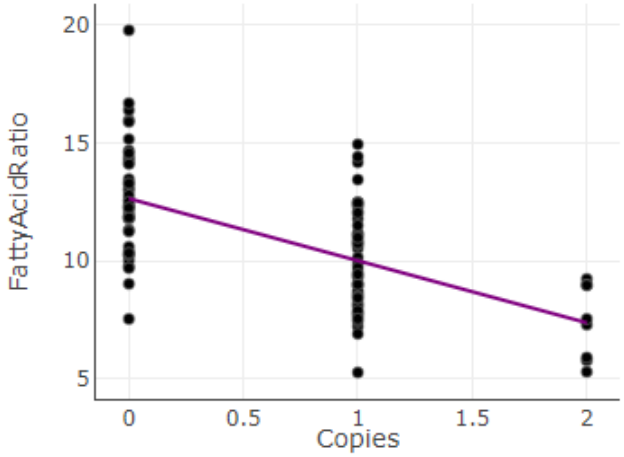
Learning Goals:

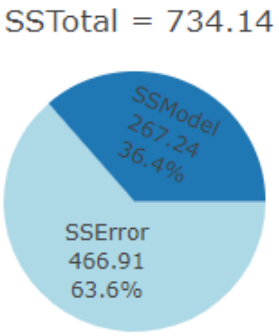
- Carry out simulation-based inference to assess the evidence of a linear association between the quantitative explanatory and response variables
- Use a theory-based approach to assess the strength of evidence of a linear association between the quantitative explanatory and response variables
- Evaluate the validity of the theory-based approach using residual plots



Recall the study on the genetic variation, cytosine instead of thymine in the FADS1 gene, and its association with fatty acid metabolism from Exploration 4.1. The fatty acid ratio and the number of copies of the cytosine variant in the FADS1 gene were determined for a sample of 100 participants in the Framingham Heart Study. A lower fatty acid ratio is better, as lower values indicate greater synthesis of the fatty acids in the blood. The question of interest in this study is whether the cytosine genetic variation in FADS1 is associated with the Arachidonic Acid to di-homo-gamma-linolenic fatty acid ratio in the blood. You found that the fatty acid ratio has a decreasing, linear association with the number of copies (0, 1, or 2) of the cytosine variant inherited from parents.

Figure 1: The association between fatty acid ratio and the number of copies of cytosine inherited





The prediction equation, found through least-squares regression, was:

$$predicted\ fatty\ acid\ ratio = 12.7 - 2.64 \times Number\ of\ copies$$

and explained about 36% of the variation in fatty acid ratios.

1. Make a prediction: Do you think that this association will be statistically significant? Why or why not?

This is a reasonable R^2 value for an observational study and a relatively large sample size, but we still need to officially assess the statistical significance of this linear association – could the sample show this strong of a linear association if the population itself did not have an association?

STEP 4: Draw inferences beyond the data.

Let's now think about how to find a p-value to assess the statistical significance of not just the "association" but more specifically the "linear association" between fatty acid ratio and the number of copies of cytosine.

2. Write out the null and alternative hypotheses we are interested in testing; make sure you are using the phrase "linear association."

Simulation-Based Inference

Let's see how to apply the 3S Strategy (statistic, simulate, strength of evidence) to test for a linear association.

3. Outline the simulation method you could use to determine whether the linear association is statistically significant. In particular, what statistic could you use to measure the strength of the linear association in the observed data? How would you shuffle the data?

Copy/paste the data, **fattyAcid2**, into the Analyzing **Two Quantitative Variable** applet. The response variable (FattyAcidRatio) is listed first so press the **(Response, Explanatory)** button. Press **Use Data** twice, and make sure there are 100 observations. Check the **Show Regression Line** and **R-squared** boxes.

Choice of Statistic

4. One choice of statistic for summarizing the linear association is R^2. Why do you think it's reasonable to use R^2 to measure the linear association between fatty acid ratio and the number of copies of cytosine? What types of R^2 values will you consider evidence against the null hypothesis?

Simulation

Check the **Show Shuffle Options** box. Notice, on the right, this applet gives you a choice of 6 statistics. Select **R-squared** as the statistic. Keep the **Number of Shuffles** set to 1. With the **Plot** radio button selected, press **Shuffle Y-values** a few times.

5. How do the scatterplots of the shuffled data compare to the scatterplot of the original data? Does this comparison make sense? What types of values are you getting for the shuffled R^2? Do these values make sense? Explain briefly.

Strength of Evidence

Now set the applet to a total of 1,000 random shuffles. Copy/paste or sketch the histogram of the shuffled R^2 values below.

6. Where is the observed R^2 value located in the null distribution of shuffled R^2 values?

7. Determine the approximate p-value for the observed R^2. Based on this p-value should we reject or fail to reject H_o?

Alternative Statistics

8. Now let's repeat the simulation process, but this time using the slope as the statistic. What value of the poplation slope, β_1, is assumed by the null hypothesis?

9. Rewrite the null and alternative hypotheses in terms of the population slope, β_1.

Note: The alternative hypothesis is a general statement that there is a linear association between the response and explanatory variable in the population. Because we have not specified a direction to the relationship, we use ≠ in our alternative hypothesis about the population slope.

Reset the shuffling by unchecking and rechecking the Show Shuffle Options box. Set the **Number of Shuffles** back to 1 and use the pull-down menu to change the statistic in the right panel to the Slope. Press **Shuffle** a few times.

10. What is a typical value for the shuffled slope when the fatty acid values get shuffled to different values of the number of copies of cytosine? How do the shuffled slopes compare to the observed slope?

11. Now get up to at least 1,000 shuffles. Based on the *scatterplot* of simulated lines (shown in grey), does it seem plausible that the observed regression slope (shown in red) happened by chance alone when there is no real association between the two variables? Explain how you are deciding.

12. Based on the *histogram* of simulated slopes, does it seem plausible that the observed regression slope happened by chance along when there is no real association between the two variables? Explain how you are deciding.

13. Approximate the two-sided p-value and include a screen capture or sketch of the null distribution showing the p-value.

Standardized Statistics

It's also possible to create a *standardized statistic* for the slope:

$$\textit{Standardized slope} = \frac{\text{observed slope} - \text{mean of slopes}}{\text{SD of slopes}}$$

14. What are some potential advantages to using a standardized statistic, rather than R^2 or the sample slope, to evaluate the strength of evidence against the null hypothesis?

15. In your histogram of 1,000 or more shuffled slopes, what are the mean and standard deviation of the shuffled slopes in the null distribution? Why do you think the mean is close to zero?

16. Use a mean of 0, and the standard deviation of the null distribution to standardize the observed slope. Write a sentence interpreting the standardized slope.

Theory-Based Approach

We can also estimate the standard deviation of the null distribution for the slopes using a mathematical formula (shown below). In this formula, SD(X) is the standard deviation of the *x*-values. For this study, that refers to the standard deviation of the number of copies.

Definition: The theory-based standard error of the slope is calculated as follows:

$$Standard\ error\ of\ slope, SE(b) = \frac{SE\ of\ residuals}{\left(\sqrt{n-1}\right)SD(X)}$$

where *X* represents the explanatory variable.

Return to the left panel of the regression applet. Check the **Show descriptive statistics** box to determine SD(X). Note that *n* is the sample size which is 100 for this study. Recall, the *SE of residuals* for the linear regression model is called "Regression SE" in the applet. Check that box.

17. Now, use the values displayed in the applet to calculate the theory-based SE of the slope, SE(b_1). How does it compare to the standard deviation of the null distribution for the slopes from the simulation you performed earlier?

The theory-based standard error should be similar to, but not necessarily extremely close to, the standard deviation of the simulated null distribution of slopes, even if you had a large number of shuffles. The above mathematical formula is based on the chance variation due to random sampling, rather than the chance variation due to random assignment being modeled by the applet. The key idea to remember is your goal is to estimate the chance variation in the slopes "by chance alone."

18. Based on the mathematical formula, what would happen to the theory-based SE of the slopes if the variation in the explanatory variable increased but everything else stayed the same? That is, if the values of the explanatory variable were more spread out? Explain visually why this relationship makes intuitive sense (in determining how much slopes vary from sample to sample).

19. Based on the mathematical formula, what would happen to the theory-based SE of the slopes if the sample size was smaller? Larger?

20. Use the theory-based standard error of the slope you found in in #17 to standardize the observed slope.

This standardized statistic is referred to as the ***t-statistic***, because across many random samples it tends to follow a *t*-distribution.

In the applet, change the statistic from the slope to the *t*-statistic. Check the box to **Overlay *t*-distribution**.

21. Does the *t*-distribution appear to reasonably approximate the null distribution of the theory-based *t*-statistic?

Enter the theory-based *t*-statistic you calculated earlier and use it to obtain a two-sided p-value. Include a screen capture of your results.

22. Are the simulated p-value and theory-based p-value similar?

Back on the left-side of the applet you can check the **Regression Table** box to verify your SE and *t*-statistic calculations.

Key Idea: When the validity conditions are met, the null distribution of the standardized slope ($t = b_1 / SE(b_1)$) can be approximated by a *t*-distribution with $n - 2$ degrees of freedom.

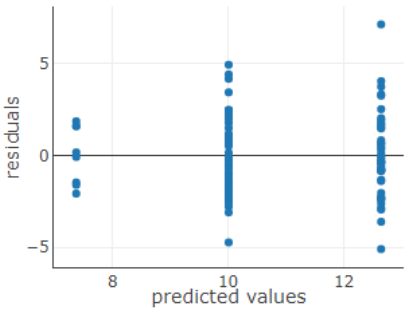
The theory-based SE of the slopes and p-value from the *t*-statistic are based on the following ***validity conditions***.

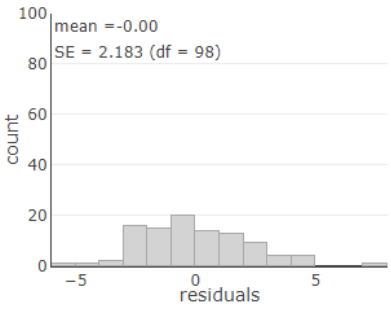
Key Idea: The validity conditions for the theory-based *t*-test for the slope are considered met if

- The residuals vs. predicted values graph does not show any strong evidence of curvature or other patterns (Linearity)
- The responses can be considered independent of each other (Independence)
- The histogram of the residuals is approximately symmetric with no large outliers (Normality)
- The residuals vs. predicted values graph shows a constant width (Equal variance)

The linearity assumption was checked in Exploration 4.1. The remaining assumptions are the same as what you have seen previously. As before, the independence assumption is checked using the study design: If the data are a random sample from some larger population, or the experiment used random assignment, the observations are considered independent. To check the other assumptions, we examine a histogram of the residuals and the residuals vs. predicted values plot shown in Figure 2.

Figure 2: Residual by predicted plot and histogram of the residuals from the linear regression model of fatty acid ratio versus the number of copies of cytosine





23. Do you consider the independence, normality, and equal standard deviation assumptions met for these data? Cite explicit evidence which supports your conclusions.

In Exploration 4.1, you saw how the various sums of squares (*SSTotal, SSModel,* and *SSError*) are calculated in the context of the linear model. Recall the *mean square* values are computed by dividing the *sum of squares* values by the associated degrees of freedom and that the *F-statistic = MSModel/MSError.* We now interpret this *F*-statistic as the ratio of the variability explained by the regression line to the unexplained variation about the regression line (rather than differences in groups means to within group variation as before). The df total is still the number of observations minus one (for estimating the overall mean). The df for model with one quantitative explanatory variable is 1, corresponding to the slope.

Check the box to display the **ANOVA Table**.

24. Report the value of the *F*-statistic and the corresponding p-value. Do they also provide strong evidence of a genuine linear association for this process? Explain.

Key idea: For theory-based inference in the context of a linear regression model, the *F*-statistic (and p-value) from the ANOVA table are equivalent to the *t*-statistic (and p-value) pertaining to the slope in the Regression table. In particular, $F = t^2$.

With strong evidence of a genuine linear association, we next want to estimate the rate of change in the response (slope), with a measure of precision.

Check the **95% Confidence interval for slope** box (Make sure the **Regression Table** box is still checked).

25. Confirm that this interval is roughly equal to $b_1 \pm 2\ \mathrm{SE}(b_1)$.

26. Write a sentence interpreting this interval in the context of this study. (You should be interpreting both the slope in context and defining the population/larger process to which you are generalizing.)

Definition: A t-confidence interval for the population slope, β_1, is given by:

$$b_1 \pm t^* \times (SE\ of\ b_1),$$

where b_1 represents the sample slope, and the multiplier $t^*$is the critical value corresponding to a t distribution with degrees of freedom being the same as the error df. For a 95% confidence interval, t^* will be approximately 2.

STEP 5: Formulate conclusions.

27. Write a paragraph summarizing the results of your analysis. Be sure to address the following:

- Is there a statistically significant linear association between the fatty acid ratio and the number of copies of the genetic variant? What is your evidence?
- What is the nature of that association? (Use the confidence interval for the slope.)
- To what population would you generalize these conclusions?
- Is a cause-and-effect conclusion valid?

STEP 6: Look back and ahead.

Future studies could look to see how common the **variant** is in individuals of different ethnicity (the Framingham sample is Caucasian) and whether the resulting association with the fatty acid ratio can be linked to medical outcomes (e.g., heart disease).