EXPERIMENT 6: HERITABILITY AND REGRESSION

DAY ONE: INTRODUCTION TO HERITABILITY AND REGRESSION

OBJECTIVES:

Today you will be learning about some of the basic ideas and tools used in quantitative genetics. By the end of today's lab you should

- Understand basic regression analysis and statistical hypothesis testing
- Understand the theory behind estimating heritability of height from a data set

INTRODUCTION:

We often want to know if, and to what extent, two variables are related, e.g., height and weight. Do taller people tend to weigh more than shorter people? Of course, because those two variables are auto correlated, that is, because height is one of the three dimensions that comprise volume, it will necessarily contribute to greater weight - but to what extent? We can use regression analysis to measure this relationship. Figure 1 shows a scatter plot of height and weight for a sample of men. Each point represents one person's height and weight. Note that x is, by convention, the independent, or causative, variable, and y is the dependent, or response, variable. For example, as you gain height your weight will increase, but as you gain weight you don't grow taller; therefore, height is the independent variable in this example.

![height-weight scatter plot](https://via.placeholder.com/150)

From this scatter plot you can see that while the shortest person is not the lightest, and the tallest is not the heaviest, there does appear to be a positive relationship. We can quantify this relationship by compressing the many data point into a single regression, or prediction, line whose slope indicates the overall relationship.
The line has the form you learned in algebra:

\[ y = ax + b \]

- \( a \) = the slope of the line (change in \( x \) divided by change in \( y \))
- \( b \) = a constant that is specific to each data set

How is the line itself determined relative to data? The line represents the overall minimized distances from the \( y \) values. Specifically, for every data point the distance to a potential line (\( \Delta y \)) is measured. This distance is squared, and the sum of all such squared distances is calculated. The line that minimizes the sum of the squared distances for all of the data is the “least squares” or “regression” line. This is an excellent job for a computer.

The equation for the regression line is shown in the upper-right corner of the chart. It indicates that the slope is 2. What does this mean? For every 1 inch change in height there is, on average, a 2 pound change in weight. How much will a 60 inch man tend to weigh? \((60 \times 2) + 40 = 160\) pounds. The prediction equation cannot perfectly predict \( y \) for each value of \( x \) because of error (scatter) around the line. However, it does predict the average relationship between \( x \) and \( y \).

(An aside: Regression is an odd term, in that it doesn’t actually describe what we are doing: relating two variables. The term regression was coined in the early 1900’s by the geneticist Francis Galton. He noticed that tall parents tended to produce children that were not quite as tall and short parents tended to produce children that were not quite as short. He developed the “regression analysis” described above to understand this phenomenon, which he called: “regression toward mediocrity.”)

**R-SQUARED = \( R^2 \) (THE SQUARED CORRELATION)**

Now that we understand the average relationship between two variables we might ask: How well does variable \( x \) explain variable \( y \)? In the example above, if you know height, how well can you predict weight? This is determined by the squared correlation or \( R \)-squared \((R^2)\).
To understand the $R^2$, we have to look more closely at regression.

There are 3 sources of variability shown in the above figure:

**Regression variability**: $y_i$-hat; the minimum distance between a point on the least squares line and the mean value of $y$ ($y$-bar).

**Error variability**: $y_i$; the difference between each $y$ value and $y$-hat.

**Total variability**: $y_i$ - $y$-bar; the difference between a value of $y$ and the mean value of $y$.

For each source of variability the sum of the squares is calculated:

<table>
<thead>
<tr>
<th>SOURCE OF VARIABILITY</th>
<th>SUM OF SQUARES</th>
</tr>
</thead>
<tbody>
<tr>
<td>REGRESSION</td>
<td>$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$</td>
</tr>
<tr>
<td>ERROR</td>
<td>$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$</td>
</tr>
<tr>
<td>TOTAL</td>
<td>$SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$</td>
</tr>
</tbody>
</table>

Note that the SSE is exactly what was used to create the least squares regression line earlier. Thus, $R^2 = \frac{SSR}{SS_y}$. It is the proportion of the total variability in the data that is removed (accounted for) by $x$.

Notice also what factors affect the $R^2$ value: the slope of the regression (which is measured by the magnitude of SSR) and tightness of points to the regression line (which is measured by SSE). A slope of zero implies that $y$-hat is zero; hence, the regression variability is zero. In other words, as $x$ increases $y$ remains constant. The equation for the line becomes $y=a$. The more the data hugs the regression line, the smaller is the SSE and therefore, the bigger is the SSR relative to the total variability.
Returning to our height-weight regression chart (p. 75), we see that the R-squared = 0.2. That is, 20% of the variation in weight can be explained by variation in height. The other 80% is unexplained. What might be some of the factors that contribute to that 80% of unexplained variation in weight? Most generally, the remaining variation in weight would be due to thickness and body composition (e.g., percent lean body mass). You might then ask what affects those variables? Health, nutrition, exercise, etc.

**STATISTICAL SIGNIFICANCE AND SAMPLING:**

$R^2$ does not tell us if the effect we are seeing is produced by chance. In the example of height-weight, this is somewhat trivial since logic dictates a relationship. But, in general, we will want to know how confident we can be that the $R^2$ is truly different from zero and not the result of noise due to: (1) **sampling error** (the chance sampling of values that differ from the true population), (2) **measurement error**, which has two components (a) **random** (chance fluctuations in measurement accuracy; e.g., some people read the scale a little from the left and over-estimate their weight, while others read from the right and under-estimate their weight) and (b) **systematic** (consistently incorrect measures in one direction; e.g., if the scale used to measure weight was under-calibrated by 2 pounds).

We are generally not so interested in the **sample** of data we collect as we are in the **population** from which it came. For example, most people would rather know heritability of height in humans, than the heritability of height of students taking Bio 184 this semester. However, practical restraints dictate that we can rarely measure everyone. Usually, we must **sample** the population which we are interested in understanding. A **sample** is a subset of individuals observed from a population. **Sampling in way that minimizes bias is a critical aspect of experimental design.** To minimize bias is an art that requires long thought and intimate familiarity with the system in question.

To answer the question of statistical significance requires an analysis of variance (ANOVA). This statistic goes beyond the scope of this class. In brief: The ANOVA uses the SSE, SSR and SSyy to estimate the probability that error present relative to the total variability would occur by chance. These values are standardized (converted to an *F-statistic*) which is then used to generate a p-value (as we saw with the Chi-square test).
BROAD SENSE HERITABILITY = $H^2$: THE FRACTION OF VARIATION IN A TRAIT THAT IS DUE TO GENETIC VARIATION

Recall the following properties of heritability:

• It is a property of populations, not individuals
• Characters are heritable only if similarity arises from shared alleles
• It is specific to a given environment

NARROW SENSE HERITABILITY = $h^2$: THE FRACTION OF VARIATION IN A TRAIT THAT IS DUE TO ADDITIVE GENETIC VARIATION

Measurement:

$h^2$ can be statistically estimated from similarities between relatives. Most commonly, F1 values are regressed against mid-parent values (mean of parent values) for a population or a sample of a population. The slope of the regression line estimates heritability. This makes intuitive sense: if an increase in the value of a trait among parents corresponds to an increase in value among offspring, it is implied that there is an inherited factor at work. Within a standard environment, this can inferred to be due to additive genetic variation. Often environments are not standardized, but instead assumed to be variable, but not different between parent and offspring. This assumption is often made when working with human data. Note, $h^2$ is not the fraction of offspring variation that is explained by parental variation (that would be the r-squared); it is the fraction of phenotypic variation that is due to additive genetic variation.

$h^2$ range: 0 - 1
$h^2 > 0$ means that additive genetic variance is present
$h^2 = 0$ means that there is no additive genetic variation. (Genes are still relevant to the trait; there are no traits without genes)
$h^2 = 1$ means that additive genetic variation perfectly predicts trait variation (but it doesn’t mean that environment is unimportant, or even less important than genetic variation). Environmental variation is being removed by rearing under a standard environment.

Before the next lab period, COLLECT AND CONVERT TO CM heights of the persons listed below. If you do not have information for one or more of the individuals that is OK. Only include what you know to be correct.

• you
• up to two of your full-siblings
• your biological parents
• 1 male and 1 female friend

Fill in the table below before the next meeting
Our next meeting will be in the computer lab announced in class. **WE WILL MEET AT THE COMPUTER LAB, NOT IN THE GENETICS LAB.** It is critical that you be at the computer lab on time to share your data at the start of the lab. If you think you will be late or absent, email your data ahead of time to a lab mate.

**DAY TWO: REGRESSION ANALYSIS OF HEIGHT DATA**

**OBJECTIVE:**

- Use regression to estimate heritability of height
- Use regression to test the hypothesis that friends assort by height

Your instructor will guide you through the computer activities. Go to:

http://www.csus.edu/indiv/h/hollandb/genetics/index.htm

and download the “height data” template file (this may be possible by simply clicking on the above URL; if not then type the URL into your browser.)

**Preparing a spread sheet:**

1. Your instructor will call on each person to call-out their data, in the order shown below. Enter the data into your spread sheet beneath each category (sample data is shown below for student #1).

```
<table>
<thead>
<tr>
<th></th>
<th>Yourself</th>
<th>Mom</th>
<th>Dad</th>
<th>Sib 1</th>
<th>Sib 2</th>
<th>Male friend</th>
<th>Fem. friend</th>
<th>Averages</th>
</tr>
</thead>
<tbody>
<tr>
<td>feet/inches</td>
<td>cm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

2. After entering the raw data: Copy the first row of data under **Averages**.

   for example: 99.0 107.5 83.5 (These cells contain formulas for manipulating the raw data.)

3. Select all of the cells below **F1** for which there is corresponding raw data. Paste the copied values. You should now see values in all three columns beneath **Averages** for every cell that you have raw data.

**Make the following regression plots using directions below:**

1. Parent-offspring (parent in x-axis)
   
   Student-friend (friend in x-axis)
Making a regression plot with Excel:
The goal is to produce charts look like the sample charts shown above (bogus data is included so that you can see a trend line, equation for the line and r-squared values).

1. Select a cell from the left-most column, immediately below your last row of raw data.
2. Click the chart icon
3. Select scatter plot
4. Next
5. Series tab
6. Add
7. Series name: (e.g., parent-offspring).
8. Select the x-values box
9. Select exactly the data that you want to see plotted on the x-axis (e.g., Parent values).
10. Repeat the last two steps for the Y-axis data.
11. Next
12. Label the axes.
13. Select the Legend tab.
14. Deselect the show legend box.
15. Next
16. Finish
17. Pull down the Chart menu
18. Select Add trend line. Linear will be selected by default.
19. Select Options tab. Select the last two boxes (display equation and R-squared). While holding down the Ctrl button, use your cursor and left mouse button to select the equations on the chart. Drag it to a white area where it is legible. If needed, you can increase the font size of the box with the format menu.
20. Decrease the range for the y and x-axes by double clicking on each axis. A format box will appear. Select the scale tab. Set the minimum to zero and the maximum to 200 (or greater if needed).
21. To change the chart size, select it (black squares will appear around it). Select a corner square while depressing the left mouse button and drag it to a new size.
22. After the first chart is complete, copy and paste it to make a 2nd chart (this way you don't have to re-format the new chart). Select the new chart and pull down the Chart menu. Select: source data. For the x and y-axes, select the data you want to plot (student and mid-friend, respectively).

23. If time permits, your instructor will lead you through an analysis of the plots you have generated. If not, the analysis will be performed during the next laboratory period. If this is the case, be sure to bring all of your data with you next time and think about your results prior to class.

24. Print, email or save to a personal disk, your spread-sheet with charts. You can continue to work on the charts. Include a copy of your spread-sheet data and charts (all can fit on 1 page) with your lab report.