

EXPERIMENT 6: HERITABILITY AND REGRESSION

DAY ONE: INTRODUCTION TO HERITABILITY AND GROUP PROBLEM SOLVING

OBJECTIVES:

- Understand the partitioning of continuous phenotypic variation into genotypic and environmental components.
- Understand broad and narrow sense heritability.
- Understand methods for estimating heritability.
- Understand basic regression analysis and statistical hypothesis testing.
- Understand sampling and sampling errors.

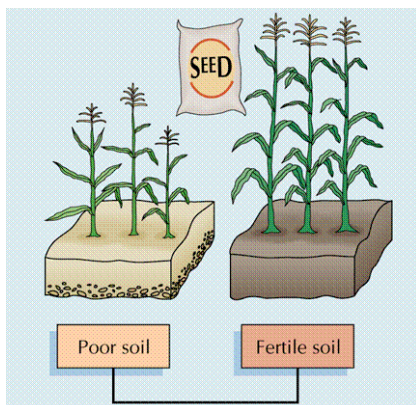
INTRODUCTION:

Before Proceeding: picture your closest male and female friends that you have met since adulthood (one each). The reason for doing this will be explained later. Write down their first names below:

Best male friend (met since adulthood): _____

Best female friend (met since adulthood): _____

If there is heritable variation among individuals within a population, and if there are differences in survival and/or reproductive success among the individuals, then the population will evolve. **Heritability** is a *prerequisite* for evolution.



Genotype: The genetic constitution of an individual organism.

Phenotype: The form of the trait shown by an individual.

Phenotype is usually the outcome of both genotype and environment.

The phenotype of the plants in **figure 1*** is a simplified example of environmental variation (soil quality) interacting with genotype, to produce a variety of phenotypes (e.g., height, weight, leaf width, etc.). Differences within each soil group (poor or fertile) are due to genetics (the seeds came from different parents). Differences between groups are due to the environment (soil quality). We intuitively recognize that the phenotypic expression of many traits can depend on environmental factors. Other examples: A starved animal will have a 'thin' phenotype, regardless of whether it carries genetic variation that might be called

'thin' or 'obese' alleles; wealthy individuals may be healthier than relatively poorer individuals because of access to health care, regardless of any differences in their genes.

Francis Galton (1822-1911) asked to what extent is a given trait in an individual determined by genotype and by environment? In other words, to what extent is it **nature and nurture**? However, it is not possible

to tease apart the influences of genotype and environment **within one individual**. For example, it makes no sense to say that an individual gets X% of her intelligence from her genes, and Y% from her environment. The only valid way to contrast 'nature' and 'nurture' is a comparative approach **between** individuals. Our question now becomes: To what extent is the **variation among individuals** for a trait due to *genetic* variation (among individuals), and to what extent is it due to *environmental* variation (among individuals)? This brings us to a modern definition of heritability.

HERITABILITY: The fraction of the phenotypic variation in a trait that is due to genetic differences. (Jay L. Lush - 1945).

The simplest model for variation in a quantitative trait partitions phenotypic variation into variation due to genetic differences between individuals, and variation due to environmental differences between individuals.

Notations:

V_P = the total *phenotypic variation* observed for a trait.

V_G = the fraction of the phenotypic variation that is due to *genetic differences* among individuals.

V_E = the fraction of the phenotypic variation that is due to *differences in the environmental* conditions to which the individuals were exposed.

Therefore, $V_P = V_G + V_E$

Heritability = genetic variance/phenotypic variance, therefore:

$$\text{Heritability} = \frac{V_G}{V_P}$$

But, because we recognize that V_P is affected by V_G and V_E , we can rewrite Heritability:

$$\text{Heritability} = \frac{V_G}{V_G + V_E}$$

Heritability is often discussed in two different ways: **broad sense** and **narrow sense**. The **BROAD SENSE**, heritability (H^2) is that fraction of the total phenotypic variation in a population that is caused by *any* genetic differences among individuals (i.e., we have been discussing broad sense heritability above).

$$H^2 = \frac{V_G}{V_G + V_E}$$

However, V_G can be broken down into additive effects and dominance effects.

V_A = the fraction of the phenotypic variance that is due to *additive genetic* differences.

V_D = the fraction of the phenotypic variance that is due to *dominant genetic* differences.

Therefore $V_G = V_A + V_D$.

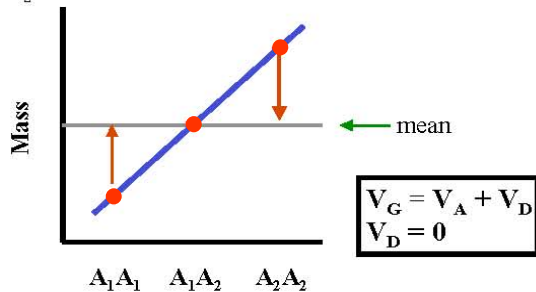
HERITABILITY = h^2 (NARROW SENSE): FOR ANY TRAIT, THE FRACTION OF VARIATION IN A POPULATION THAT IS DUE TO **ADDITIVE** VARIATION IN THE GENES

$$h^2 = \frac{V_A}{V_A + V_D + V_E} = \frac{V_A}{V_P}$$

Additive vs. Dominance

Scenario #1: Codominance

– An A_2 allele confers one unit of mass increase



Scenario #2: Dominance

– A_1A_2 genotype has A_2A_2 phenotype

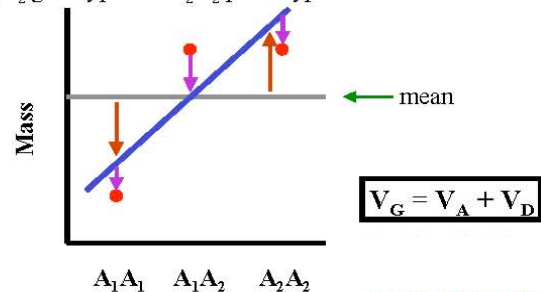


Figure 2*. Two hypothetical relationships between a trait (mass) and genotype. The slopes in both scenarios are the same, but the mean and variance differ. The red arrows indicate the difference between the best fit line and the mean, and are due to *additive* effects. The purple arrows indicate the difference between the observed values and best fit line, and indicate the effect of *dominance*. Notice that dominance increases the noise around the linear relationship, but not necessarily the slope.

Measuring Broad Sense Heritability with Clonal Transplant Experiments

Measuring V_P for a trait is straight forward: measure the trait (e.g. flower mass) for all individuals and calculate the variance: **Variance = $\sum (x_i - \text{mean})^2 / N$**

Measuring V_G or V_E is more difficult, but if you could measure V_E then you could estimate V_G by subtracting V_E from V_P (because $V_P = V_G + V_E$). Fortunately, we can accomplish this through transplant experiments. Using clonal replicates, we can plant identical 'individuals' in different environments. Variation among the replicates (i.e., within each clone) *must* be due to environmental differences since they all share an identical genotype. If you perform the same experiment using several independent clones (to create a reciprocal transplant experiment) you can get an overall estimate of V_E by taking the mean variance across all replicates. You could now estimate $V_G (= V_P - V_E)$, as well as estimate heritability $H^2 = V_G / V_P$.

Measuring Narrow Sense Heritability with Parent-Offspring Regression

Heritability can also be estimated by comparing relatives through the use of a scatter plot. By plotting 'mid-parent' value (x-axis) by mean progeny value (y-axis) for a particular trait across several families we can observe to what extent progeny resemble their parents. Often environments are not standardized, but instead, assumed to vary, but not differ on average between parent and offspring. If progeny resemble

their parents on average then the slope will be one (e.g. $h^2 = 1$). If progeny do not resemble their parents any more than they do parents of an unrelated individual, the slope of the best-fit line will be zero (e.g. $h^2 = 0$; figure 3)

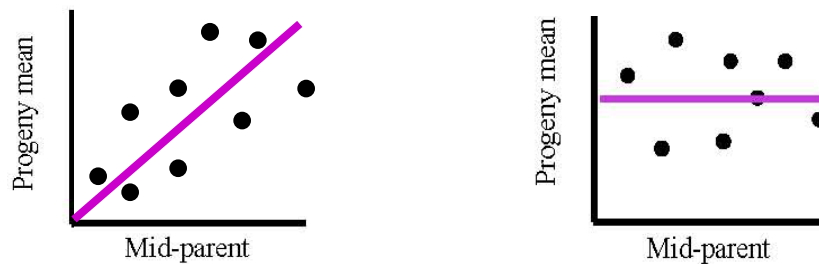


Figure 3*. The slope of the parent-offspring regression line estimates h^2 . If heritable, a change in the value of a trait among parents corresponds to a change in value among offspring. Shown above are the theoretical maximum ($h^2 = 1$, left figure) and minimum ($h^2 = 0$, right figure) heritability.

h^2 range: 0 - 1

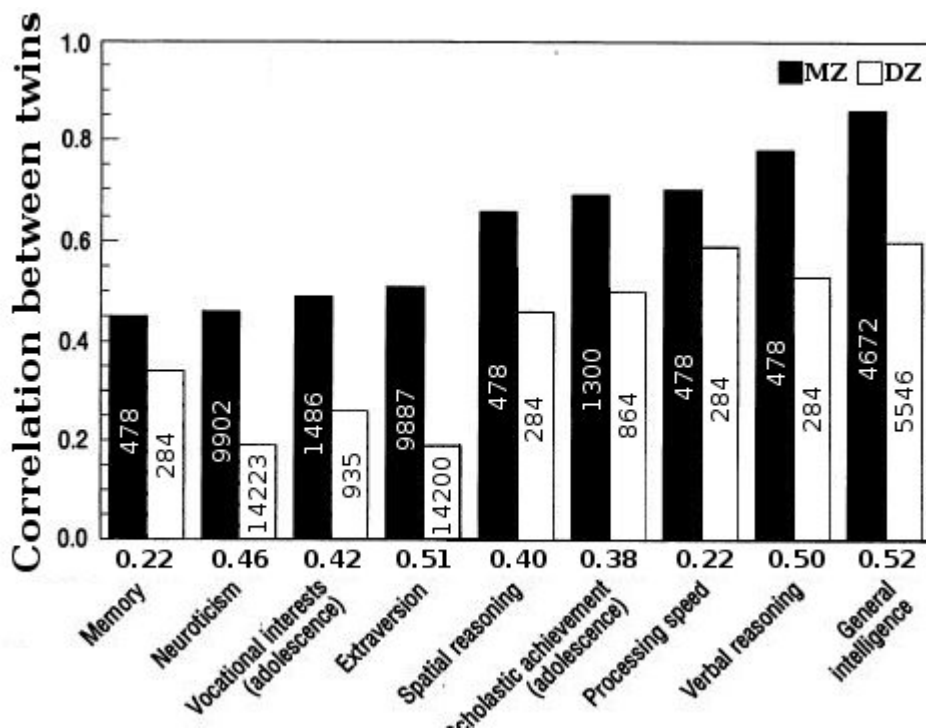
$h^2 > 0$ means that additive genetic variance is present

$h^2 = 0$ 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*. (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.)

Most h^2 estimates fall within 0.2-0.6, although higher and lower estimates are not uncommon.

Measuring Heritability of traits in humans



This frequently is estimated by comparing resemblances between twins. Identical, monozygotic (MZ), twins are twice as genetically related as fraternal, dizygotic (DZ), twins; so heritability is approximately twice the difference in correlation (r) between MZ and DZ twins, $h^2 = 2(r(MZ) - r(DZ))$.

Figure 4. Estimated heritability for nine psychological traits as estimated from [twin](#)

[studies](http://en.wikipedia.org/wiki/Heritability). In all families the twins were raised together (sample size shown inside bars). Modified from <http://en.wikipedia.org/wiki/Heritability>.

SELECTION

For a trait to evolve it must: 1) vary across individuals, 2) some variants of the trait must be more fit than others, and 3) trait variation must be heritable. Therefore, h^2 can also be estimated by trying to make the trait evolve and measuring the response to selection (Figure 5).

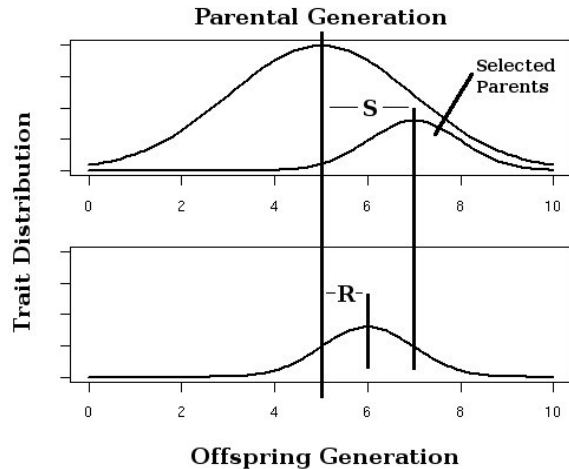


Figure 5. Trait distributions for parents (top) and offspring (bottom) within a hypothetical population. The response to selection (**R**) (mean value of the offspring minus the mean value of the parental population) divided by the strength of selection (**S**) (the mean value of the selected parents (those allowed to mate) minus the mean of the entire parental population) is an estimate of h^2 . In summary: $h^2 = R/S$. Modified from <http://en.wikipedia.org/wiki/Image:Resp-to-sel.jpg>

Important properties of h^2

- Traits are heritable only if similarity arises from shared alleles.
- It is specific to a given environment.
- It is a property of a population, not an individual.
- It does not imply that the trait is fixed (e.g., hair color, weight, height, extraversion, and intelligence are all heritable and can be modified).
- h^2 is NOT the fraction of variation among offspring that is explained by variation in the parents; that is the R-squared (R^2), or, coefficient of determination.
- h^2 IS, instead, an estimate of the fraction of the phenotypic variation within the population that is due to additive variation in their genes.

REGRESSION

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 6 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.

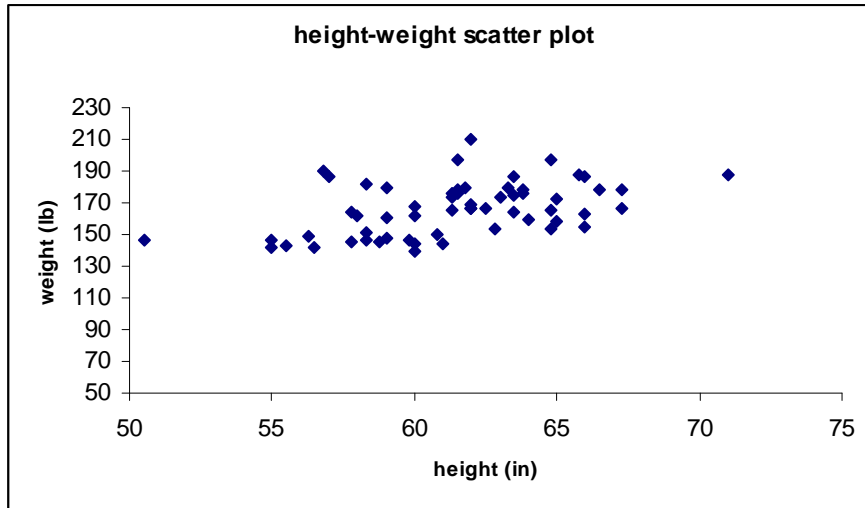


Figure 6. Height-weight for a group of men. From this scatter plot you can see that while the shortest person is not the lightest, and the tallest is not the heaviest, there does appear to be a positive relationship (correlation).

Regression takes this one step further by estimating a specific value of Y , for every X value. For example, if someone is 68 inches tall, what is your best estimate for their weight? 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 (Δy) 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.

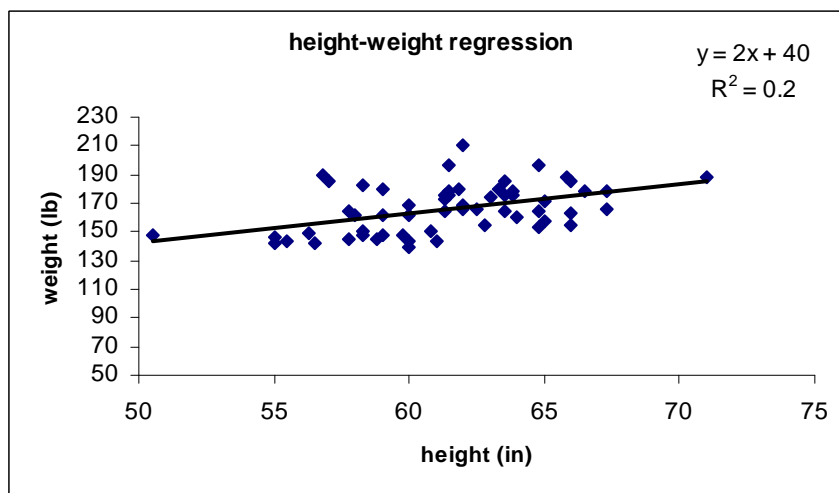


Figure 7.

Height-weight regression for a group of men. 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 .

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

Now that we understand the *average* relationship between two variables let's return to the correlation. We might ask: **How well does variable x explain the variation in y ?** In the example above, if you know height, how well can you predict weight? This is measured by the *squared correlation* or *R-squared* (r^2).

To understand the R^2 , we have to look more closely at regression.

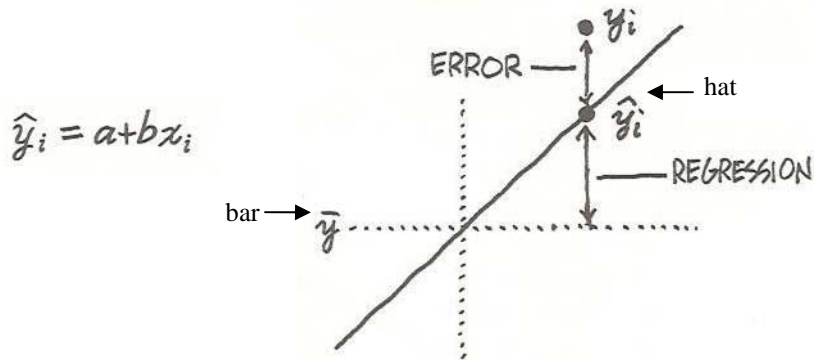


Figure 8. Shows three sources of variability:

Error variability: $y_i - \hat{y}_i$; the vertical distance between each y value and the regression line (\hat{y}_i is corresponding point on the regression line to each y_i value).

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

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

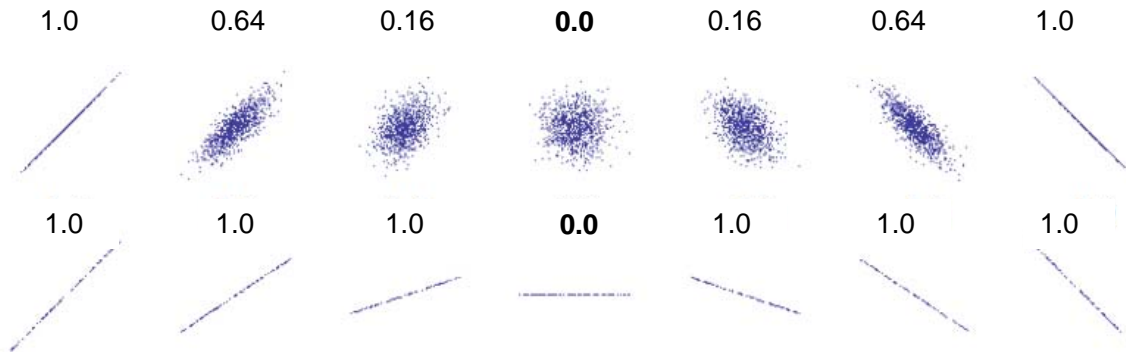
For each source of variability the sum of the squares is calculated:

SOURCE OF VARIABILITY	SUM OF SQUARES
REGRESSION (signal)	$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
ERROR (noise)	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
TOTAL (signal+noise)	$SS_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$

The SSE is exactly what was used to create the least squares regression line earlier. Thus, $R^2 = SSR/SS_{yy}$. It is the proportion of the total variability in the data that is removed (accounted for) by x . The remainder is unexplained variation (called error, or noise).

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 \hat{y} 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. You should be able to understand each source of variability verbally. The math above is provided as an aide because it unambiguously defines each concept. To help calibrate your intuition, figure 9 shows R^2 values for 14 sets of x, y data.



Adapted from <http://en.wikipedia.org/wiki/Correlation>

Figure 9. R-squared values for 14 data sets.

Returning to our height-weight regression chart (Fig. 7), we see that the $R^2 = 0.2$. That is, 20% of the variation in weight can be explained by variation in height. The other 80% is unexplained (noise, or, error). 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....

By analogy: **the slope of a regression is like an average** (measure of central tendency) and the R^2 is **an index of the variance** (as R^2 increases the variance decreases).

SAMPLING

The squared correlation 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 constraints dictate that we can rarely measure an entire in a population. 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 a 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 SS_{yy} 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).

THINGS TO DO:

Work in small groups to answer the questions below. All of the questions can be answered through studying the lab manual.

1. The table below shows the flower mass for a hypothetical clonal plant. Estimate H^2 of flower mass in this population. Hint: V_P and V_E have been calculated for you. What does "clone variance" represent (what is the variable within each clone)?

environment	Flower mass (g)								
	Clones A			Clones B			Clones C		
1	8	9	8	7	7	7	5	5	4
2	9	8	7	5	6	6	4	4	3
3	7	5	8	3	3	5	4	4	4
4	9	5	4	6	5	4	3	4	3
5	6	8	8	5	4	3	2	3	2
6	6	9	9	4	4	4	2	2	2
Mean	7.5	7.3	7.3	5.0	4.8	4.8	3.3	3.7	3.0
Clone variance	1.2	1.4	1.3	1.2	1.2	1.2	1.1	1.0	0.9
Mean clone variance	1.17								
Variance for all plants	1.46								
V_g									
H^2									

2. In fig. 4, why is it important that the twins be raised together?
3. Give two reasons why is it difficult to estimate V_E in the study shown in fig. 4?
4. The estimated heritability for general intelligence is 0.52. Does this mean that 52% of your intelligence is determined by your genes? Provide your rationale.
5. What is the estimated heritability of the data from fig. 5?
6. The data sets from figure 9 indicate that the R^2 is proportional to the _____.

7. The data sets from figure 9 indicate that the R^2 is not sensitive to _____?

Before the next lab period:

1. **COLLECT 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.
 - a. You
 - b. up to two of your full-siblings who are adults (finished growing)
 - c. your biological parents
 - d. 1 male and 1 female friend (who you identified at the beginning of the lab). The friends should be people you met after you and they were done growing.
2. **CONVERT TO CM** Download from 184 Home: "Heritability.height.Excel.2007" or (if your computer is not Office 2007 compatible) "Heritability.height.Excel.2003". You will find four sheets (see tabs in lower left area). They are self explanatory. **Fill in the table below and bring the information to out next meeting.**

	Yourself	Mom	Dad	Sib 1	Sib 2	Male friend	Fem. friend
cm							

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 CLASS HEIGHT DATA

OBJECTIVE:

- Use regression to estimate heritability of height.
- Use regression to test the hypothesis that friends assort by height.
- Use your lab manual to understand the difference between heritability and R^2 .

Your instructor will guide you through the computer activities. Download:

<http://www.csus.edu/indiv/h/hollandb/184-genetics/lab/Heritability/heritability.student.download.xls>

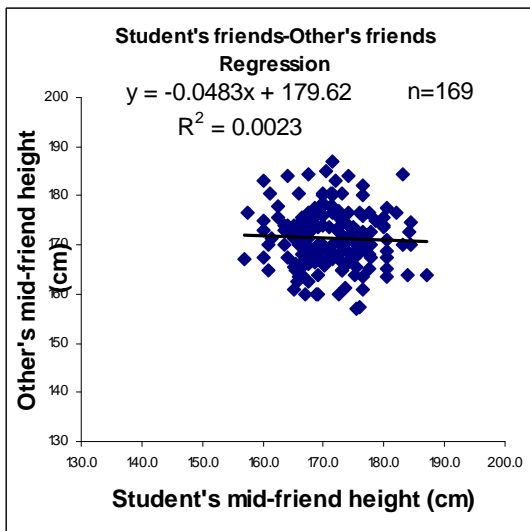
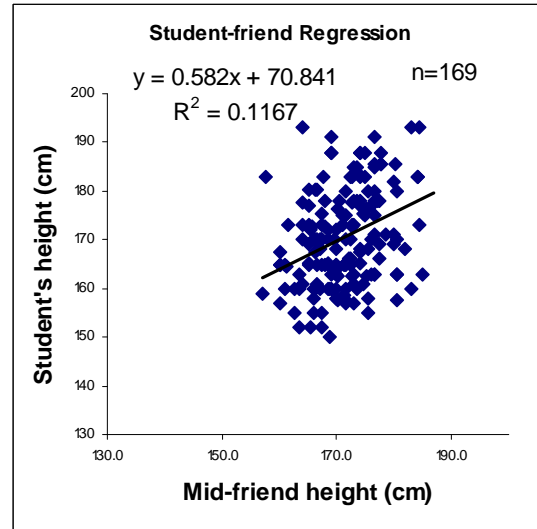
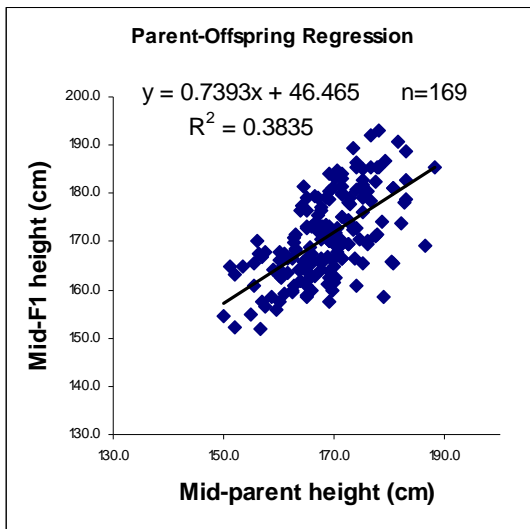
Preparing a spread sheet:

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).

	Bio 184	All heights in cm (2.54 cm/inch)						Averages		
#	Student	Mom	Dad	Sib 1	Sib 2	Male friend	Fem. friend	mid F1	mid parent	mid-friend
1	100	105	110	98	99	88	79	99.0	107.5	83.5

2. Make the following regression plots using directions below:
 1. Parent - offspring (parent in x-axis; causative variable)
 2. Student - friend (friend in x-axis; this is arbitrary,)

3. Other's mid friend - student's mid friend regression (mid friend in x-axis; arbitrary)



Making regression plots with Excel: The goal is to produce charts look like the sample charts shown above (the data was collected over several years of BIO 184 labs, and is included so that you can see examples of how your figures should look (formatting, labels, axes ranges, etc.).

1. Open an Excel 2007 file (if using Excel 2003 see instructions at [Regression with Excel 2003](#), supplemental lab material, lab 6)
2. Select a cell from the left-most column, immediately below your last row of raw data
3. Click Insert tab on Menu bar
4. Select *scatter* plot with only markers
5. Under Design, select the grey (left most) data point choice graphic
6. Right click on Chart; choose Select Data
7. Click Add; The cursor should be inside the Series Name box; click the cell above the data column you wish to add (e.g., mid parent)
8. Move cursor to the Series X values; select the data within that same column

9. Repeat above for Series y values (e.g., F1)
10. Click OK, (window closes); Click OK (window closes); data should be visible in chart
11. Right click on axis; select format axis
12. On new window: Axis Options: select fixed for Minimum (enter 130) and Maximum (enter 210); click Close
13. If the data points are connected by a line: right click on data point, select Format Data Series > Line Color > No line > Close
14. Change Menu tab from Design to Layout
15. Click twice on Axis Titles icon; on pull down menu, for one of the two options, follow the menu and choose the second option from the top; repeat with second of the two options; you should now see Axis Title next to each axis
16. Click on Axis Title and change to appropriate title shown in lab manual; repeat for other axis and Chart Title
17. Select Trendline icon from Menu; choose Linear Trendline;
18. Right click on the trendline; choose Format Trendline; Select Display Equation on Chart; and Display R-Squared on Chart
19. To change the chart size, select a corner while and drag it to a new size
20. After the first chart is complete, copy and paste it to make a 2nd chart (this way you do not have to re-format the new chart).
21. Right click on the new chart and choose Select Data. Return to directions above as needed to insert appropriate data and re-label this chart.
22. 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.
23. Email, or save to a personal storage device, 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.

DAY THREE: DISCUSSION OF RESULTS AND GROUP PROBLEM SOLVING

OBJECTIVES:

Today you will be integrating your results with the concepts introduced during day 1.

- Discuss basic regression analysis and statistical hypothesis testing with respect to your results.
- Answer questions about your results using concepts drawn from your lab manual.

THINGS TO DO:

Work in small groups to answer the questions below. All of the questions can be answered through studying the lab manual.

Parent-Offspring Regression

1. Within your sample, is there evidence that height is heritable? What is the heritability?

2. How well does parental height predict offspring height? (How much of the variation in F1 height is explained by parental height)?
3. What is the remainder of the variation in F1 height attributable to?
4. What is the average F1 height?
5. What is the average parent height?
6. Why might the average student height be different than the average parent height?
7. Published estimates of human height heritability estimate it at 0.65. Is our measure of heritability close or far from the mark? What do you attribute the difference to?

Student-Friend Regression

8. Why did you use student height in this regression instead of mid-F1 height?
9. What would each of the following results between Student and mid friend height, imply?
 - a. positive slope
 - b. negative slope
 - c. no slope
10. What are some of the factors that might have contributed to the observed positive relationship between Student and mid-friend height?
11. How much of the variation among student height is explained by variation in their friends height?
 - a. Is height a big factor in predicting who will form friendships?
 - b. In this regression, is the X-variable causing the Y-variable, or are they both affecting the relationship?
12. Suppose that in this regression, all of the data fell perfectly along the best fit line, what would that imply?
13. Do you think people intentionally (consciously) form friendships based on height?
 - a. Might height be an unconscious factor in forming friendships?

14. When a participant does not know the nature of their role in an experiment (e.g., whether they are in the 'experimental' or 'control' treatment) this is called '**single blind**'. When the experimenter also does not know a participants role in the experiment, it is called '**double blind**'.
- a. Why did we ask you to choose your friends before reading the lab manual?
 - b. What was the value of performing this experiment 'blind'?

Other's mid-friend - student's mid-friend regression

15. What is the purpose of regressing "others mid-friend - your mid-friend height"?
- a. Is there any reason to expect a relationship between these variables?
 - b. Is this a kind of control?
16. Do these results imply that the results seen in the previous two figures are **not** due to chance?
17. Do these results give you more confidence that the relationships seen in the other figures are real?

*Some components of this document were modified from a lab by J. Brown.