STA613/CBB540: Statistical methods in computational biology

Linear Regression (1/1/17)

Lecturer: Barbara Engelhardt *Scribe: Ethan Hada*

1. Linear regression

1.1. **Linear regression basics.** Linear regression is a technique used for modeling data Y conditioned on random variables X.
It is often used for data analysis or for prediction problems.
It can be very useful in understanding the relationship between a quantitative variable Y and variable X.

We will look at data that are IID - Independent and Identically Distributed

Our data today will consist of:

$$x : \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$
$$y : scalar\ in\ \Re$$

- X is a vector of *predictors* for linear regression
- Y is a scalar *response* variable
- β is a scalar *coefficient.*

Our goal in linear regression is to estimate the coefficients, including a slope and an intercept, describing the relationship between X and Y:

- β_0 : intercept
- β_1 : slope.

Augment the vector X with a 1 to include an intercept term β_0 in the model.

Then we can model the linear regression as:

$$\begin{aligned} Y &= \beta_0 + \beta_1 x \\ &= \beta^T \hat{X}, \end{aligned}$$

where the second equation assumes the augmented X. An example system we might model with linear regression is shown in Figure 1.

We project x^* onto our coefficient line, and, in this framework, this yields the best guess for what y^* will be.
We're going to assume, for the time being, that our Y variables are *quantitative*, which means that they are in $\Re$, the reals. Y can represent a quantitative trait, including many different clinical

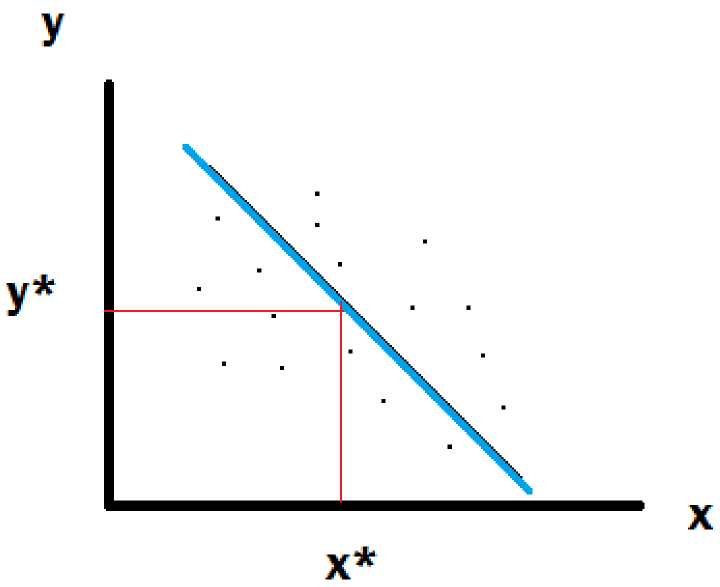

FIGURE 1. ***E***xample linear regression. x* is our predictor for y*. We see larger values of x correlated to smaller values of y in this linear regression.

traits, such as gene expressions, heights, levels of iron in the blood, etc.
What does the linear regression look like with n samples of paired data $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, where each x_i is now a vector of length p?

$$y_i = \hat{\beta}_0 + \sum_{i=0}^{p} x_{i,j}\hat{\beta}_j$$

What this means is that y is a function of the intercept, plus a scaled Y, or an affine transformation of X. Y can also be represented as:

$$y = \tilde{x}^T \beta$$

where

$$\tilde{x} : \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{bmatrix}$$

How can we judge the fit of a line? What makes two coefficients β different?

We must first talk about what a *residual* is.

1.2. Residual Sum of Squares.

A *residual* is:

$$y_i - x_i^T \hat{\beta}$$

We can write $\hat{y}_i = x_i^T \hat{\beta}$, where $\hat{y}_i$ is the estimate of y_i from our model.

The residual quantifies the distance between the prediction $\hat{y}_i$ and the actual y_i.

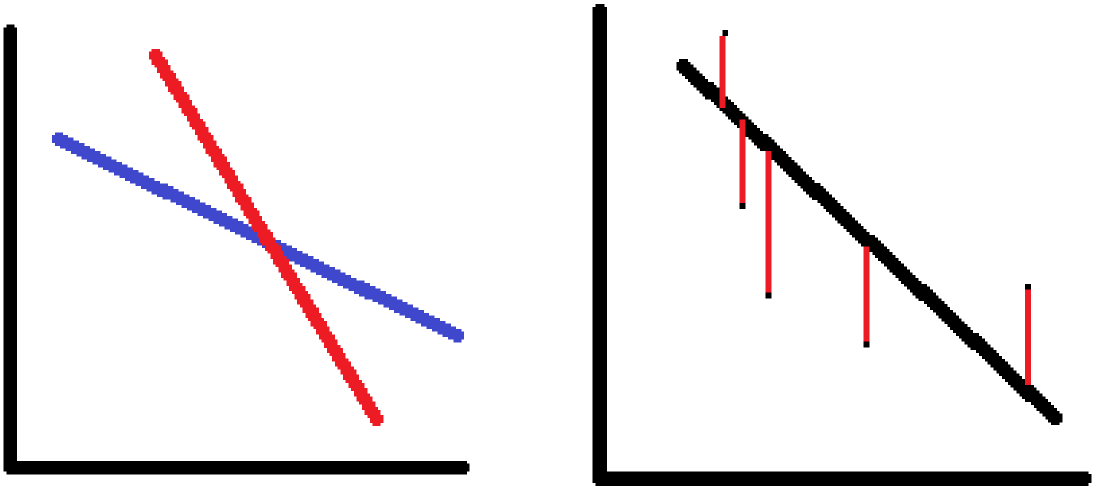

FIGURE 2. How can we tell the difference between our red and blue linear regression estimates (left panel)? (right panel) shows the residuals (in red) between the points y_i and the predictions $\hat{y}_i$.

Residual Sums of Squares (RSS) computes the the residuals across the n data points (squared):

$$RSS(\beta) = \sum_{i=0}^{n} (y_i - x_i^T \hat{\beta})^2$$

We can look at two possible coefficients β now, and compare them to find which one has a smaller $RSS(\beta)$. But this would not be very efficient, because we would have to search through a large number of β terms. Instead, let's optimize this function directly. In order to do this easily, we can put a distribution on the residual term and use the maximum likelihood framework.

We can describe the distribution of the residuals in terms of a normal distribution:

$$y_i = x_i^T \beta + \epsilon_i$$

$$\epsilon_i = N(o, \sigma^2)$$

We assume that the error term, ϵ, is distributed as a Gaussian with mean 0, variance σ^2. At each point on the line, there is a Gaussian distribution describing how far a point y_i is from our residual line, representing our best estimate of y_i. The conditional expectation of Y, then, is

$$E[Y|X, \beta] = \beta^T X,$$

and the expected residual, under this statistical model, is zero:

$$E[Y - \hat{Y}|X, \beta] = 0.$$

In our prediction framework, then, given an estimate of $\hat{\beta}$ and a new x^*, our best guess for y^* would be $\hat{\beta}^T x^*$.

Now we have some terms we can use to explain why we call this a *linear* regression.
The expected value $E[Y|X, \beta]$ of Y conditioned on X and β is a linear function of the parameters

of the system. The function must be linear in the parameters. Some examples of linear functions are:

$$\beta_0 + \beta x^2$$

or

$$\beta_0 + \beta \frac{1}{x}$$

but not

$$\beta_0 + \beta^2 x$$

What about the term *regression*? Initially, Dalton measured the heights of fathers and sons. He found that tall fathers tend to have tall sons, and short fathers tend to have short sons; however, he also found that if a father was very tall, his son would be shorter, and a short father would have a taller son. He called this phenomenon 'regression to the mean.'

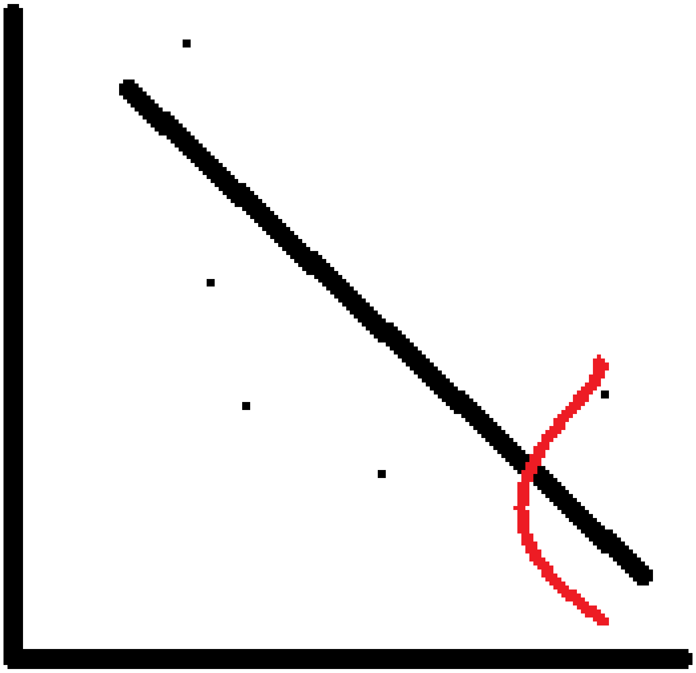

FIGURE 3. The Gaussian distribution of X and Y pairs around our residual line represented in red.

1.3. **Estimating coefficients β.** We can estimate β in two different ways, and both give the same answer. The first way we can estimate is using linear algebra without consideration of the statistical model. The system we solve to obtain a solution for β is:

$$\arg\min_{\beta}(RSS(\beta)) = \arg\min_{\beta}((Y - X^T\beta)^T(Y - X^T\beta))$$

where

$$Y: \text{ vector of length } n$$

$$X: \ n \times p \text{ matrix}$$

$$\beta: \text{vector of length } p$$

We are using matrix notation for Y and X to include all of the data samples n, combining the individual samples into a vector Y and a matrix X.

We can write out our function explicitly, and solve for β without much difficulty:

$$\begin{aligned}\arg\min_{\beta}(RSS(\beta))) &= Y^TY - (X\beta)^TY - Y^TX\beta + (X\beta)^T(X\beta) \\ &= Y^TY - 2(X\beta)^TY + (X\beta)^T(X\beta) \\ &= Y^TY - 2Y^TX\beta + (X\beta)^T(X\beta)\end{aligned}$$

Now, we can take the derivative of the last term with respect to β, and set it equal to zero, and then solve the system. This means we are solving for β by optimizing the residual sum of squares for this system:

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2Y^TX + 2X^TX\beta = 0$$

And solving this system rewards us with our much desired β:

$$\beta = (X^TX)^{-1}X^TY : \text{ Normal Equations}$$

This tells us that β is a function of x^Ty and accounts for x^Tx, an estimate of the covariance of x. This can also be called 'ordinary least squares,' and is the easiest way to solve the system when the matrix X^TX is invertable or *nonsingular*.
The other way to think about this problem is by working in the probabilistic setting where the conditional distribution is normal. We can write out the data likelihood, then take the log of the likelihood. Then we can take the derivative of the log likelihood with respect to β, set this to zero, and then solve. We are afforded some advantages for working with the MLE system, one of which is that we obtain an explicit definition for the variance of the system.

$$\text{MLE} : L(y|x,\beta) = -\frac{n}{2}\log 2\pi - n\log\sigma - \sum_{i=1}^{n}\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}$$

We should recognize this as the likelihood of a normal y with a mean $x^T\beta$ and a variance σ^2.
Now we can take the derivative with respect to β. But only one term will survive because only one term has a β to derivate.

$$\frac{\partial L(y|x,\beta)}{\partial\beta} \longrightarrow -\sum_{i=1}^{n}\frac{(y_i - x_i^T\beta)}{2\sigma^2}$$

This is just a scaled version of the residual sum of squares, $RSS(\beta)$. This illustrates an implicit assumption of a Gaussian residual when we minimize the RSS for the system.

We can also write the normal equations in vector notation:

$$\beta = \left(\frac{1}{n}\sum_{i=1}^{n}X_iX^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_iY_i\right)$$

As mentioned above, we can get a definition for the variance of the system using the MLE method. We assume that the data are homoscedastic, which means that they all have the same variance σ^2. Then we can write out the variance estimate based on the maximum likelihood estimates of the variance for the normal distribution:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_t^T\beta\right).$$

1.4. **Coefficient of Determination.** The RSS is one way of determining how well the linear regression model fits our data, but there are other metrics we can use to quantify this relationship too. For example, we can use the *coefficient of determination*, generally denoted by r^2, and bounded between $0 \leq r^2 \leq 1$.

$$r^2 = \frac{\sum_{i=1}^n \left(X_i^T \beta - \bar{y}\right)^2}{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2},$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the mean of y. Here, the numerator quantifies the variance of our predictions with respect to Y, and the denominator is the variation in Y. This statistic quantifies how much variation in Y is accounted for in the estimation of $\hat{Y}$, or the proportion of the total variation in the response that is explained by the fitted line. If r^2 is near 1, that indicates that our line explains the data well. An r^2 value near 0 indicates that the model does not fit the data well, and it is not likely that there is a linear relationship between X and Y.

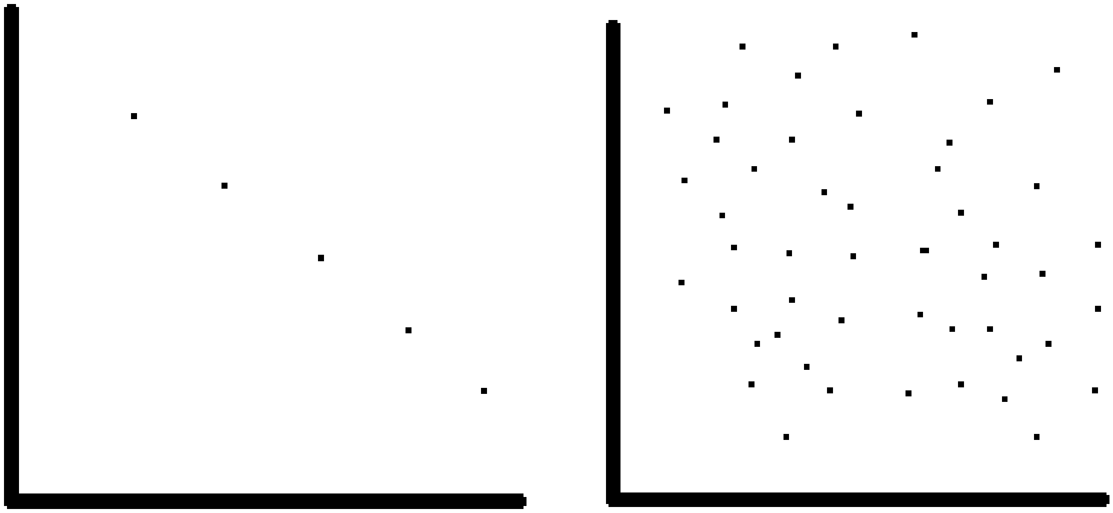

FIGURE 4. In the data on the left, we should expect a fitted line to have an r^2 value near 1 because the data is linear in nature. On the right hand side, we wouldn't expect a fitted line to explain the data very well. In both figures, the x-axis represents X and the y-axis represents the corresponding Y.

What are the assumptions implied of this method of regression investigation?

- We assume that the n samples (x_i, y_i) are IID
- the p dimensions of X are independent
- The data are homoscedastic, which means that all points share the same σ^2 value.

If X or Y are not independent, we can use Bayesian regression for linked parameters (coming up in a few lectures). If the X dimensions are correlated, it is called *coordinated* or collinear. This is a violation of the assumptions we need for ordinary least squares, but we can use one of the models mentioned above for modeling this kind of data.

1.5. **Generalized Least Squares.** This is a form of ordinary least squares in which the variance term is modeled as σ_i^2. We obtain the matrix Σ.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}$$

This changes our estimate for β to:

$$\hat{\beta} = \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1} y$$

We assume that, in this case, we know our $\sigma^2 : \sigma_1^2, \ldots, \sigma_n^2$. If we don't know that, we can use iterative updates to find β and the 'weights' Σ.

1.6. **Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes [Stranger et al., 2007].** Up until this point, we have been thinking about linear regression in the context of prediction of our responses y. This is a common usage of linear regressions, but it isn't the only one. In the paper, linear regression is used to establish a relationship between two variables. First, let's just have a simple review of how genes lead to observable traits.

$$\text{Chromosomes} \longrightarrow \text{DNA} \longrightarrow \text{mRNA} \longrightarrow \text{Proteins}$$

Genes are transcribed into mRNA, and then translated into proteins. Most of the functions in the cell are performed in the proteins. Proteins can be used to identify traits in a subject, but this is a difficult and expensive task at the moment. It is cheaper and easier to measure mRNA levels. So, the hypothesis is that we could use mRNA to measure protein level differences in individuals. In other words, we can look at mRNA to identfy gene expression.
V. E. Cheung, in the mid 2000's, published an influential paper in *Nature* that posed the question:

Are there genetic variants in humans that are associated with the levels of mRNA in an individual?

eQTLs, or expression quantitative trait loci, are genetic variants that associated with levels of mRNA. eQTLs influence a continuous observable trait; they are thought to be involve in regulation of the translation of specific mRNA. One way of quantifying the association between a locus and a gene expression trait is by quantifying the *effect size*, often denoted by β, the coefficient in linear regression models. In the current study, large β values correlate with a large impact of a locus on gene expressions levels, small β values indicate weaker impact of a locus on gene expression levels. The linear regression model discussed in this paper is one that can be used to identify eQTLs.

The relationship investigated in this case was between gene expression values and SNPs (singular nucleotide polymorphisms), or CNVs (copy number variants). SNPs refer to a difference in a single nucleotide in the genome. CNVs refer to a natural variation in the number of copies of a genomic segment. Insertions and deletions (often referred to as *indels*) refer to single or groups of nucleotides being added into or deleted from the genetic code.

$$\begin{array}{ccccc} & A & C & A & T \\ SNP & \longrightarrow & \downarrow & & \\ & A & A & A & T \end{array}$$

SNPs tag for other polymorphisms. This means that there is a correlation between the causal SNP and the identified, *tag* SNP. It has been estimated that on the order of 80% of common CNVs are tagged by common SNPs, so SNPs can be used to identify CNVs with associations to the phenotype of interest. SNPs only consider a single nucleotide, and genome-wide arrays can assay millions of SNPs easily, which makes them easier to measure than CNVs directly.

When investigating genotype-phenotype associations, large sample sizes improve your statistical power. In this study, 210 people provided genetic and gene expression information: 45 Han Chinese, 45 Japanese, 60 Nigerians, and 60 Europeans. Given these population groups, what might happen if a single gene is expressed differently in each group? In sorting the gene expression values from these data, we see gene expression levels often differ by population ancestry.

In this paper, the authors modeled each data set separately to avoid this problem. They could have projected the gene expression data for each gene and each population to the quantiles of a normal distribution, which would eliminate the differences in population-specific expression, but might be too large a correction and eliminate some of the eQTL signals as well.

In this paper, they used r^2 values to quantify possible associations. They identified how much of the total variance of each gene was explained by a SNP. They chose the r^2 threshold for gene-SNP pairs that were called significant using permutations of their data.

The *minor allele frequency* (MAF) is the proportion of the minor allele copies in the population. *Allele* is short for allelomorphs, which variants of a particular genetic locus, and in this case, there are two possible nucleotide alleles for each SNP. MAF quantifies how rare the less frequent allele is in the population, and it is quantified by $MAF(B) = \frac{1}{2n}\sum_{i=1}^{n} 01(g_i, AA) + 11(g_i, AB) + 21(g_i, BB)$, where $1(g_i, XX)$ is the indicator function that takes value 1 when the genotype for individual i is XX, and zero otherwise.

Linkage disequilibrium (LD) is a way to consider the relationship between linked (or co-evolving) SNPs. LD can be best explained using an example (Figure 7). Your grandparents DNA were paired to form your parents' DNA, and your parents DNA created you. The two strands of DNA recombine in the parent to form half of the genome of the offspring. In this model, we can see that segments of the genome that are close together are more likely to be from the same ancestor than what a truly random distribution would predict (meaning: the two loci are independent). This means that segments of the genome close together have a higher probability of coming from the same ancestor than those that are far apart; correspondingly, their probability of occurring together in a genome is not the product of their marginal frequencies in the population.

The *additive assumption* makes the claim that the distance between the means between $\{0, 1\}$ are the same as the distance between the means between $\{1, 2\}$. This is an important assumption to take note of in this system, as it appears to be a commonly observed relationship between SNPs and quantitative phenotypes.

There are other types of models of associations. In recessive associations, we see that the levels of gene expression expected for genotypes with zero or one copies of the minor allele are expected to be the same. We can encode the data with similar response as the same genotype to test for an association in this framework. Similarly, we can consider the over-dominant model, where the homozygotes (AA or BB) at a particular SNP have the same expected gene expression values, but the heterozygotes (AB) have a different expected level of gene expression. It appears in the data

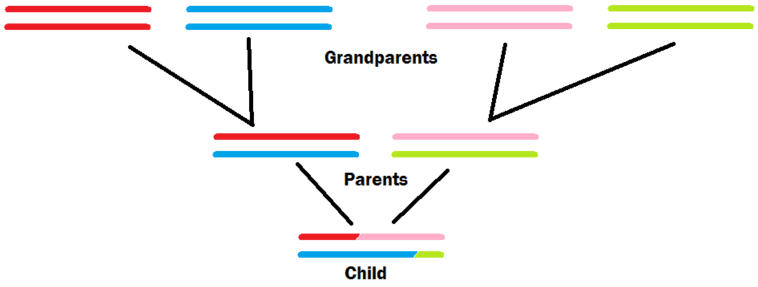

Figure 5. Grandparents donate genetic information to parents, who in turn donate genetic information to offspring. Genetic information is combined between parents in offspring, with one half of each (autosomal) chromosome inhereted from each grandparent. Areas in the chromosomes close to each other are more likely to come from the same ancestor. Linkage disequilibrium is a measure of correlations in nearby sites in a genome.

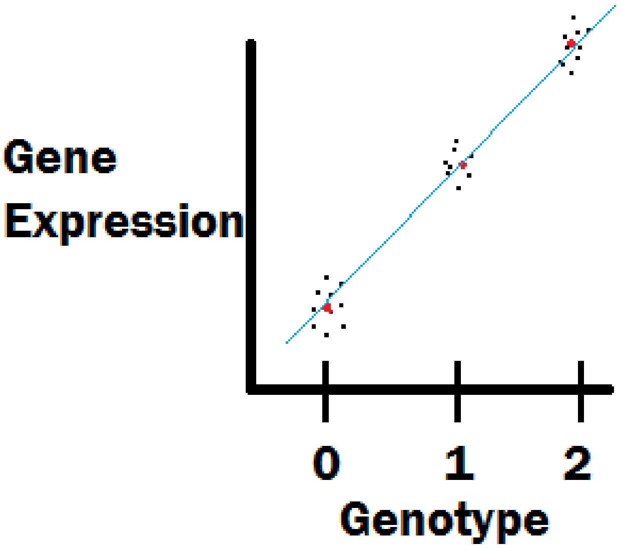

Figure 6. Genotypes are coded as SNP's. The number on the X axis refers to the number of copies of the minor allele. The height on the Y axis refers to the response variable, gene expression. In this graph, we have a lot of collinearity because individuals have 0, 1 or 2 copies of the minor allele (scattered a bit along the x-axis in this plot to show the relationship). This graph would suggest a strong correlation between number of copied minor alleles and gene expression. The β in this case would be greater than 0, because there is a positive slope in the fitted line.

that people have looked at, that the additive association is much more prevalent in eQTLs than any other model.

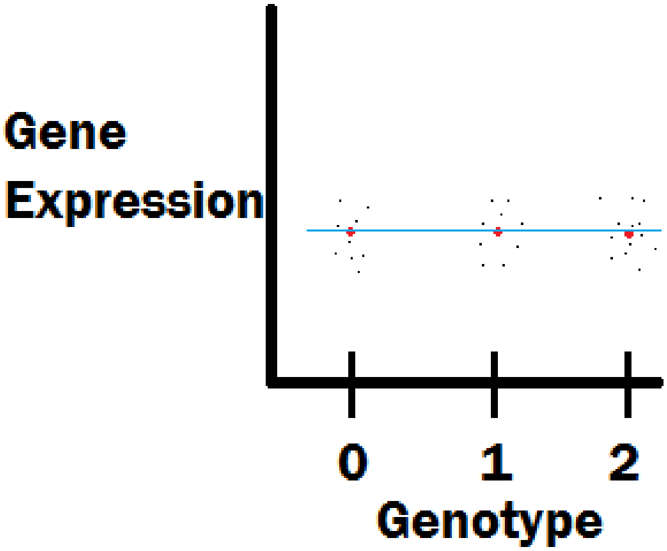

Figure 7. This graph shares axes with Figure 5. In this graph, there is no association between the minor allele copies and the gene expression. Increasing the independent variable does not affect the dependent variable. The β here would be approximately 0.

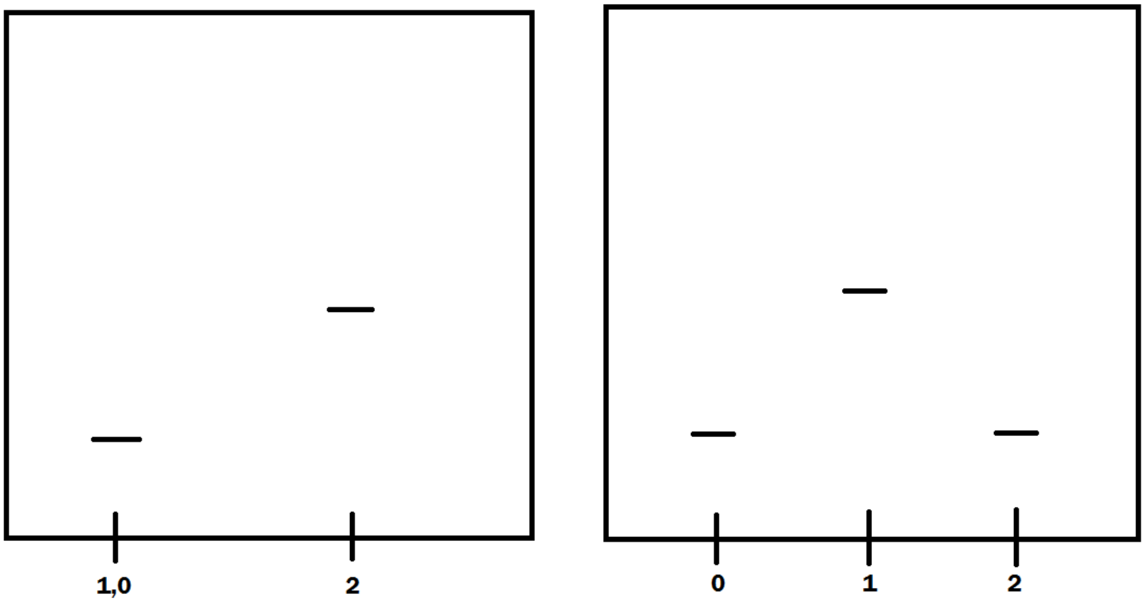

Figure 8. Here, we see an example of a recessive model, and an example of an over-dominant model. On the left, entries 0 and 1 have been grouped together. On the right, we could group entries 0 and 2 in the way 0 and 1 were on the left, which would change the over-dominant model into a recessive one.