Principal Components Analysis

Collin Nolte

April 19, 2021

Review – Population Stratification

A few weeks ago, we had considered the the issue of *population stratification*, in which systematic differences in allele frequencies occur as a result of non-random mating

One consequence of this is that, when considering disease association in alleles, we cannot simply consider allele frequency – population stratification results in different populations having a different overall frequency

We were able to account for this in the transmission disequilibrium test by only performing inference of subjects with heterozygous parents – that is, we took additional steps to control for population *confounding*

Review – Population Stratification

Of course, it isn't just differences in allele frequency that can differentiate a population. For example, suppose we were interested in conducting a study on heart disease in Europe

It's certainly true that genetic information is not uniformly mixed: Greek people are more likely to have children with other Greek people, and similarly for the French

However, there are also significant cultural differences between these two populations; smoking is far more common in France than in Greece, while a Mediterranean diet tends to be much higher polyunsaturated fats

Review – Population Stratification

As a consequence of this, we may find an association between heart disease and a particular allele in one of these populations, when in reality this allele has nothing to do with heart disease

On top of allele frequency, this represents another source of *confounding* which can lead to significant bias in population-based studies

Ideally, we would have a way to indicate and control for membership in a population to isolate the genetic effects above those that may be cultural or geographic

Review - Regression

In the last two weeks, we have considered the relationship between a response variable, Y and a collection of explanatory variables, X in the context of a linear model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

Recall that interpretation of β_i is the change in Y given a change in X_i, *all else being equal*

From that discussion, there are two topics that are relevant today: the use of indicator variables to control for different groups (vitamin C in guinea pigs) and the concept of multicollinearity

Review – Indicator variables

With our opening discussion in mind, we might note that it would be convenient if we could also have indicators for our observation indicating to which population they were a member

There are a few difficulties with this, the largest of which is the fact that, in actuality, there are no single, clear-cut populations: even for the French and Greek example, these countries may themselves have any number of subpopulations

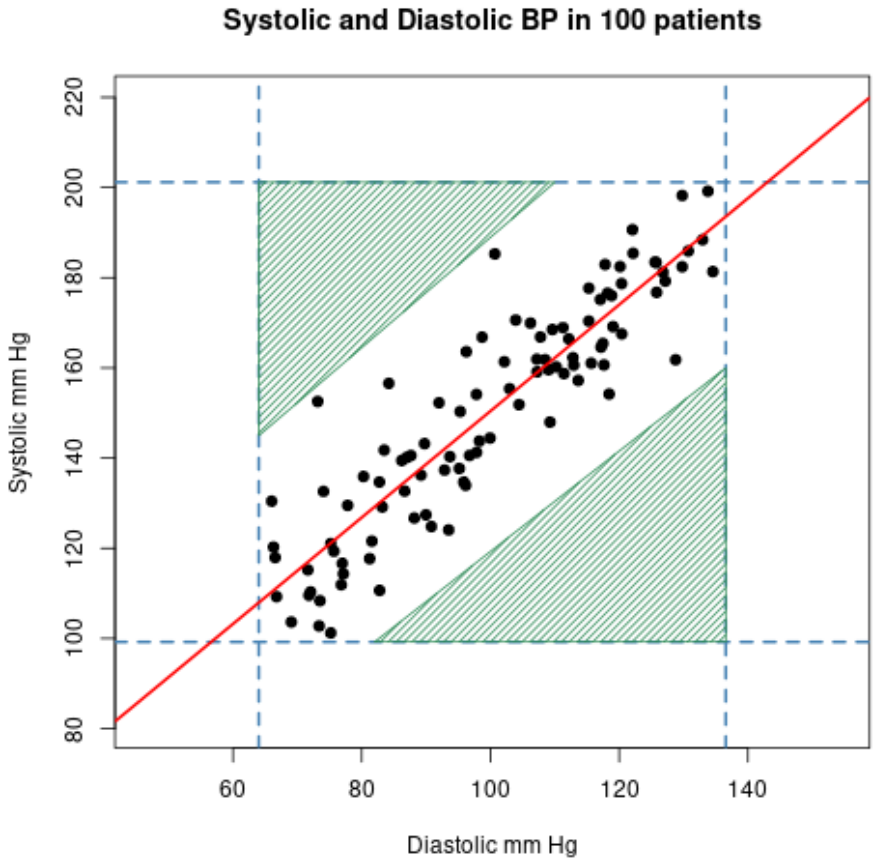
Our goal today will be to determine if we can create something that *resembles* an indicator in the sense that it may function to indicate some sort of membership in different populations

Review – Multicollinearity

Considered case with systolic and diastolic blood pressure

Impossible to describe change in one variable without change in another

Even though two varibles present, perhaps could be described with one as a combination of the two



Review – Multicollinearity

This idea of creating a new variable out of a combination of the others ones is particularly useful

On one hand, it captures most of the information contained in multiple variables while including fewer in our model; this is great for increasing the value of $n - p$

On the other, it may include information that is not obvious from the other two alone. For example, consider the interpretation of bmi rather than handling height and weight separately

Principal Components

This brings us to the idea of principal components, which is a slightly more general case of what we have just discussed

Given a collection of p covariates, the principal components are p *new* covarites that are linear combinations of the original

Unlike the original covarites, the new principal components are completely independent, with no multicollinearity at all

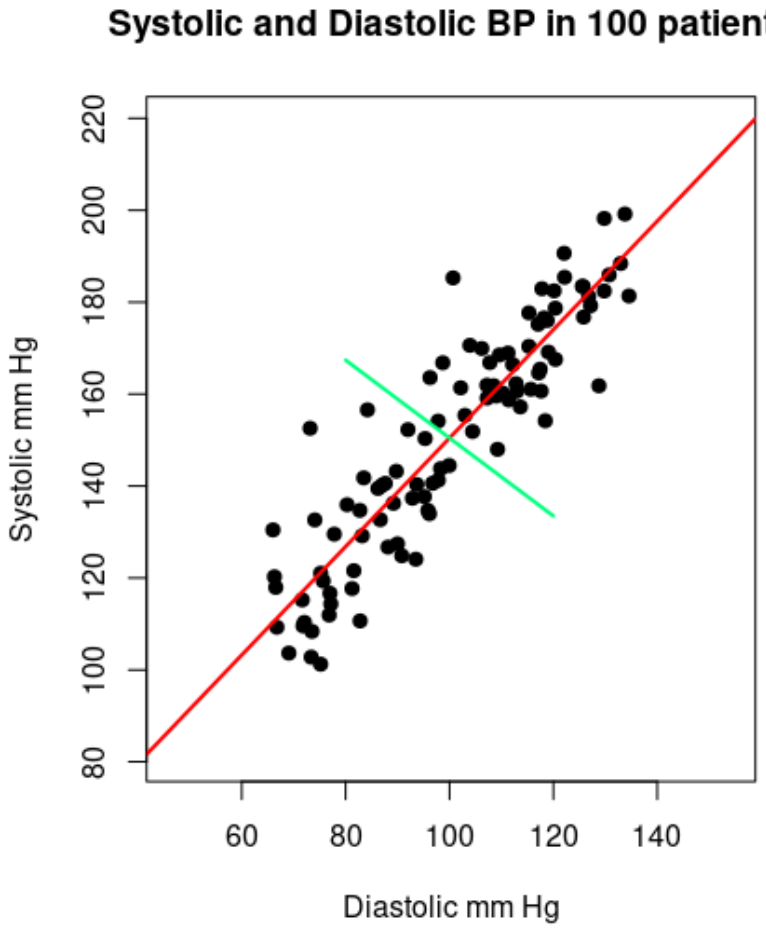
Principal Components

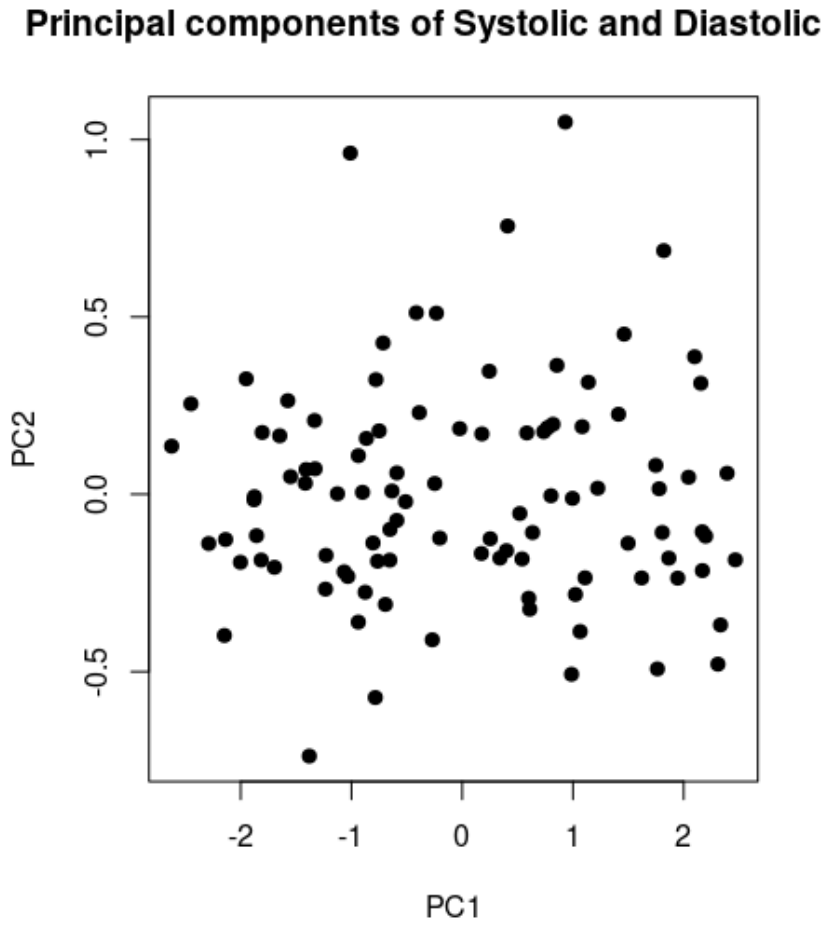
The construction of the new principal components is not arbitrary, however

Recall for blood pressue, we saw that we could capture most of the information with a straight line drawn through the data. We might rephrase this concept of capturing the most information as "capturing the most variance"

As such, the principal components serve to create a collection of new covariates in a descending order, depending on how much covariate variability it is able to "capture"

Principal Components





PC Interpretation

As we have two covariates (systolic and diastolic bp), we also had two principal components

The first, we saw, could be recognized as some combination of systolic and diastolic bp, while the second was ... what was left

This is just to say, while we may be able to create some interpretation of the principal components, they do not necessarily have to *mean* anything

As far as the PCs are concerned, they are nothing more than the reorganization of our covariates into independent pieces. Specifically, independent pieces that are *linear combinations* of the original covariates

Principal Components – Loadings

The amount of each covariate that makes up a particular principal component in its linear combination is know as its *loading*

The loadings themselves may or may not be useful; in a situation where has 10 covariates, it may be the case that the first principal component has three covariates with large loadings, with the remaining seven being small

We might be able to say, then, that these three covariates account for most of the variance in the first principal component. Our present example is unfortunately uninteresting. As they more or less fall on a line, we find their contributions or loadings are about the same

PC recap

Regarding PCs, here is what we have determined so far:

1. Given p covariates, we are able to construct p principal components
2. The principal components are independent of each other, and ordered by amount of information contained
3. As consequence of (2), we may be able to describe more with less, giving us a form of *data/dimension reduction*. This follows since the first few PCs usually contain most of the variance
4. The individual PCs may or may not be directly interpretable, but may be able to tell us things that may not be immediately obvious from the covariates
5. Loadings in a PC may give us an indication of which covariates explain most variance

Example data set

One thing we noted at the beginning of this lecture is that it would be nice to have some covariate that *resembles* and indicator variable for population

We will consider a genetic example here in which, given a collection of SNPs for individuals collected from multiple populations, we will try to determine if any one of them is responsible for some phenotypic outcome

Our goal will be to determine if we can use the information from our covariates (the SNPs) to see if we can construct any principal components that can effectively be serve as our population indicator

SNPs

Quick review, a SNP is a single-nucleotide polymorphism at a specific position in the genome

The dataset we will be using itself is coded as values $\{0, 1, 2\}$, indicating the number of minor alleles at a particular position on the chromosome

For example, if the dominant allele is A, then a value of 0 indicates an A in both chromosomes, a value of 1 indicates a B (minor allele) at the same point in one chromosome, and a value of 2 indicates that both positions contain the minor allele

Example

Going back to what we said about populations and multicollinearity, this should make some sense

We might assume that the SNPs within a given population are highly correlated and, if we can reduce these down to one or two "indicator" covariates, we may be able to effectively control for population

Keep in mind, though, that these aren't *real* indicators: they are not coded as $\{0, 1\}$ and may contain information not strictly related to population membership

Example

For this example, we are going to consider a dataset from 98 population-informative SNPs for 197 individuals

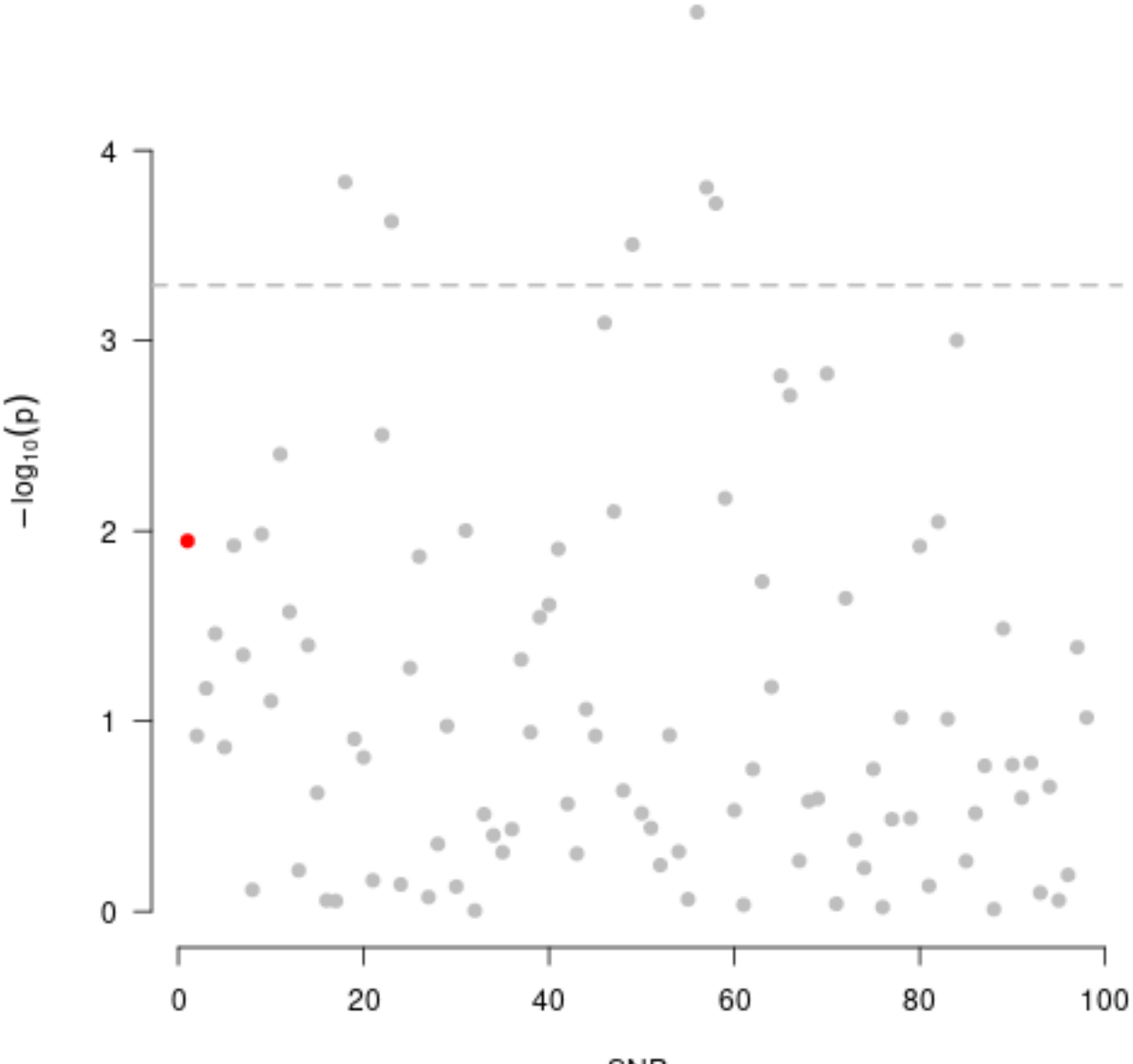
These individuals are made up of four groups: 50 Africans, 47 African Americans, 50 Europeans, and 50 Japanese

While the data is real, we will simulate the phenotype:

$$Y = 0.5\text{SNP}_1 + 1(\text{if J}) - 1(\text{if AA}) + \epsilon$$

That is, only the first SNP will be associated with our phenotype, with higher incidence of occurence in Japanese populations and lower in African Americans

log p-values



Considering principal components

While only considering the relation of the SNPs themselves to the outcome, we were unable to detect the single SNP that was responsible

We might consider using the full collection of SNPs to generate PCs and see if doing so is able to account for the genetic makeup of the subjects

Let's start by considering the SNPs with the largest positive and negative loadings (contributions) in the first principal component

Loadings for first PC

SNP 23 had largest positive loading in first PC

	0	1	2
African	6	19	25
African American	7	29	11
European	46	4	0
Japanese	50	0	0

SNP 31 had largest negative loading in first PC

	0	1	2
African	50	0	0
African American	39	8	0
European	18	28	4
Japanese	20	22	8

Loadings for second PC

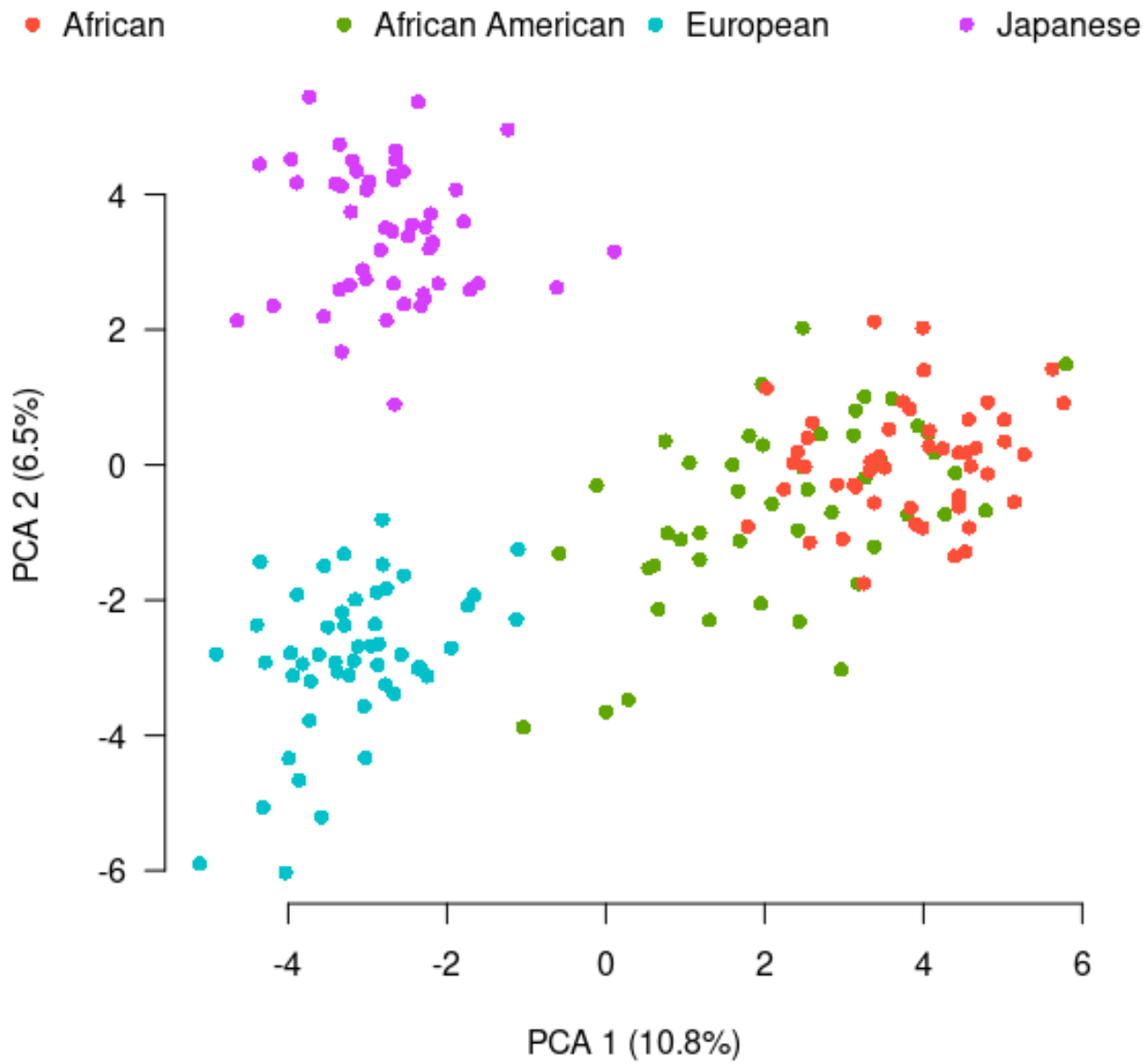
SNP 22 had largest positive loading for second PC

	0	1	2
African	50	0	0
African American	47	0	0
European	50	0	0
Japanese	28	18	4

SNP 63 had largest negative loading for second PC

	0	1	2
African	13	27	10
African American	9	23	15
European	5	25	20
Japanese	49	1	0

Observations by first two PCs



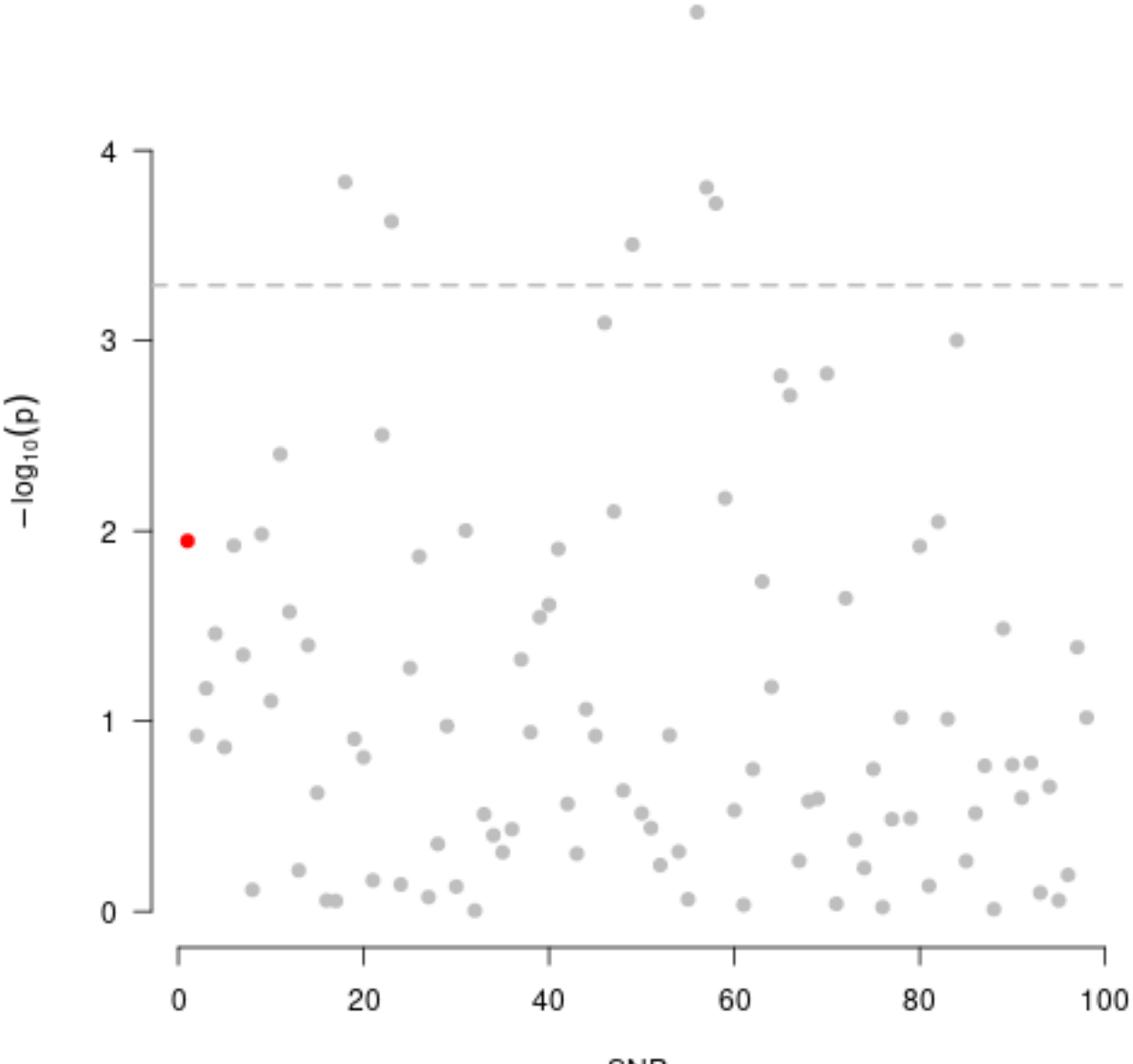
Adjust for population

Here, we see that just plotting our observations with the first two principal components starts to create a collections of groups that match populations

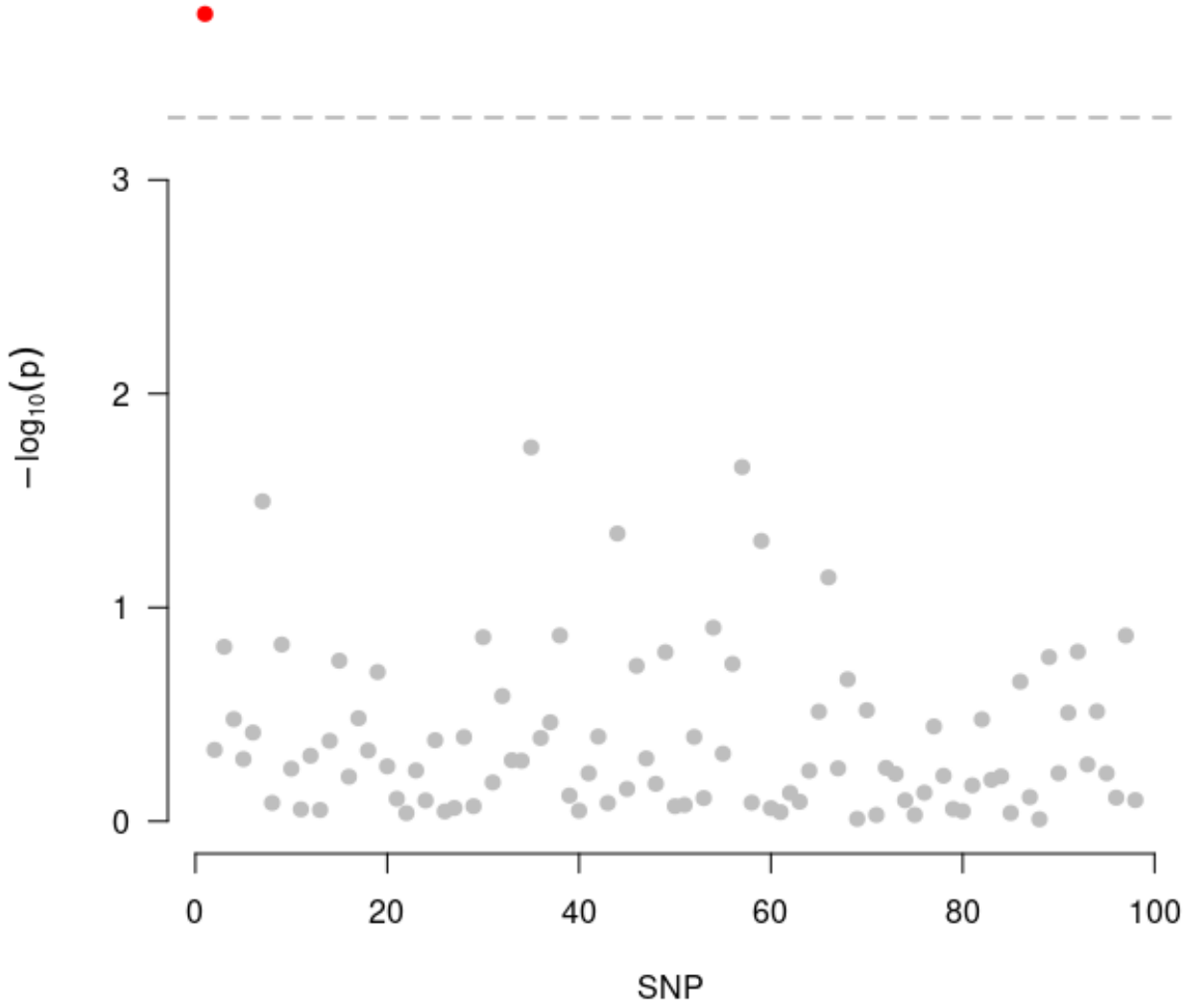
Once these PCs have been constructed, we can include them in our model just as we would any other covariates

Also note that even though we are now including them in our model, we are not really interested in inference on the PCs themselves – here, they only serve to control for the population stratification that already exists

Old log p-values



New log p-values



Review

The principal components of a dataset are a linear combination of the original covariates, given different weights or loadings

The PCs are subsequently organized by the amount of variability explained, resulting in the first few capturing most of the variability

PCs may be interpretable, as in the BP example, or function as control variables, as in the genetic example

References

- Patrick Breheny's notes on statistical genetics (June 29, 2020)
- Wikipedia