A primer on statistics genetics

Seth D. Temple (sdtemple@uw.edu)
4th year PhD Student in Statistics

April 18, 2023

Credit to:

- BIOST 311 course slides
  - Charlie Wolock, Nina Galanter, Taylor Okonek (UW BIOST PhDs)
- BIOST/STAT 550 course slides
  - Sharon Browning (UW BIOST)
  - Timothy Thornton (UW BIOST, Regeneron)
- BIOST 533 course slides
  - Amy Willis (UW BIOST)

Agenda

1. Contingency tables
   - Why? ---> quality control, testing HWE
2. Hypothesis testing
   - Why? ---> controlling false discoveries in GWAS
3. Linear models
   - Why? ---> interpreting GWAS, PRS
4. Principal components analysis (PCA)
   - Why? ---> quality control, population stratification

Contingency tables

With focus on chi-squared tests

Credit: Browning, Temple

Allelic Chi-square test

- General form of chi-square tests is:
  - $\sum(O-E)^2/E$
  - Sum over entries in table
  - O is observed, E is expected (under the null hypothesis)
- The allelic chi-square test assumes each allele is independent

Observed	Allele A	Allele B	Total
Cases	54	146	200
Controls	84	116	200
Totals	138	262	400

Expected	Allele A	Allele B	Total
Cases	138/2=69	131	200
Controls	69	131	200
Totals	138	262	400

$$X^2 = \frac{(54-69)^2}{69} + \frac{(84-69)^2}{69} + \frac{(146-131)^2}{131} + \frac{(116-131)^2}{131} = 9.96$$

df = 1

P-value in R: pchisq(9.96,1,lower.tail=F)

P-value is 0.0016. Conclude there is a difference.

Genotypic tests

- Assume genotype counts are multinomial within each population.
- Do not need HWE. Do need lack of relatedness.
- Can test one genotype against the others: e.g., Does AA genotype frequency differ between cases and controls?
  - Could use z-test or chi-square test as for allelic tests.
  - This is a 1 d.f. test
- Can test the overall hypothesis that one or more genotypes has a frequency difference between cases and controls.
  - Can use chi-square test on the k x 2 contingency table (for k genotype classes).
  - d.f. = k-1

Genotypes/ Alleles	**AA**	**AB**	**BB**	**Genotype Totals**	**A**	**B**	**Allele Totals**
Cases	11	32	57	100	54	146	200
Controls	20	44	36	100	84	116	200

Genotypic test. $H_0 : p_1^{AA} = p_2^{AA}$ and $p_1^{AB} = p_2^{AB}$ and $p_1^{BB} = p_2^{BB}$

```
> x=matrix(c(11,20,32,44,57,36),nrow=2)
> x
     [,1] [,2] [,3]
[1,]   11   32   57
[2,]   20   44   36
> chisq.test(x)

        Pearson's Chi-squared test

data:  x
X-squared = 9.2496, df = 2, p-value = 0.009806
```

Allelic vs genotypic tests

- Allelic tests are great when the effect of the allele is additive
  - Individuals with two copies of the allele are more likely to get the disease than individuals with one copy, who are more likely to get the disease than those with zero copies.
  - Only one degree of freedom (one free parameter) helps power.
- If we know the disease is dominant or recessive or over-dominant we might use a genotypic test.
- In practice, we usually use an allelic test.
- The allelic tests we talked about so far assume HWE.

Chi-square tests

- $\sum(O-E)^2/E$
- The sum is over categories of observed counts, O (here the three genotypes).
- Calculate the expected counts, E, under the null hypothesis.
- The HWE chi-square test is **NOT the same** as the contingency table chi-square test.

HWE chi-square test

- $X^2 = \frac{(n_{AA} - e_{AA})^2}{e_{AA}} + \frac{(n_{AB} - e_{AB})^2}{e_{AB}} + \frac{(n_{BB} - e_{BB})^2}{e_{BB}}$
  - $e_{AA} = n\hat{p}^2, e_{AB} = 2n\hat{p}(1 - \hat{p}), e_{BB} = n(1 - \hat{p})^2$
    - $\hat{p}$ is the sample frequency of the A allele.
- The degrees of freedom is 1

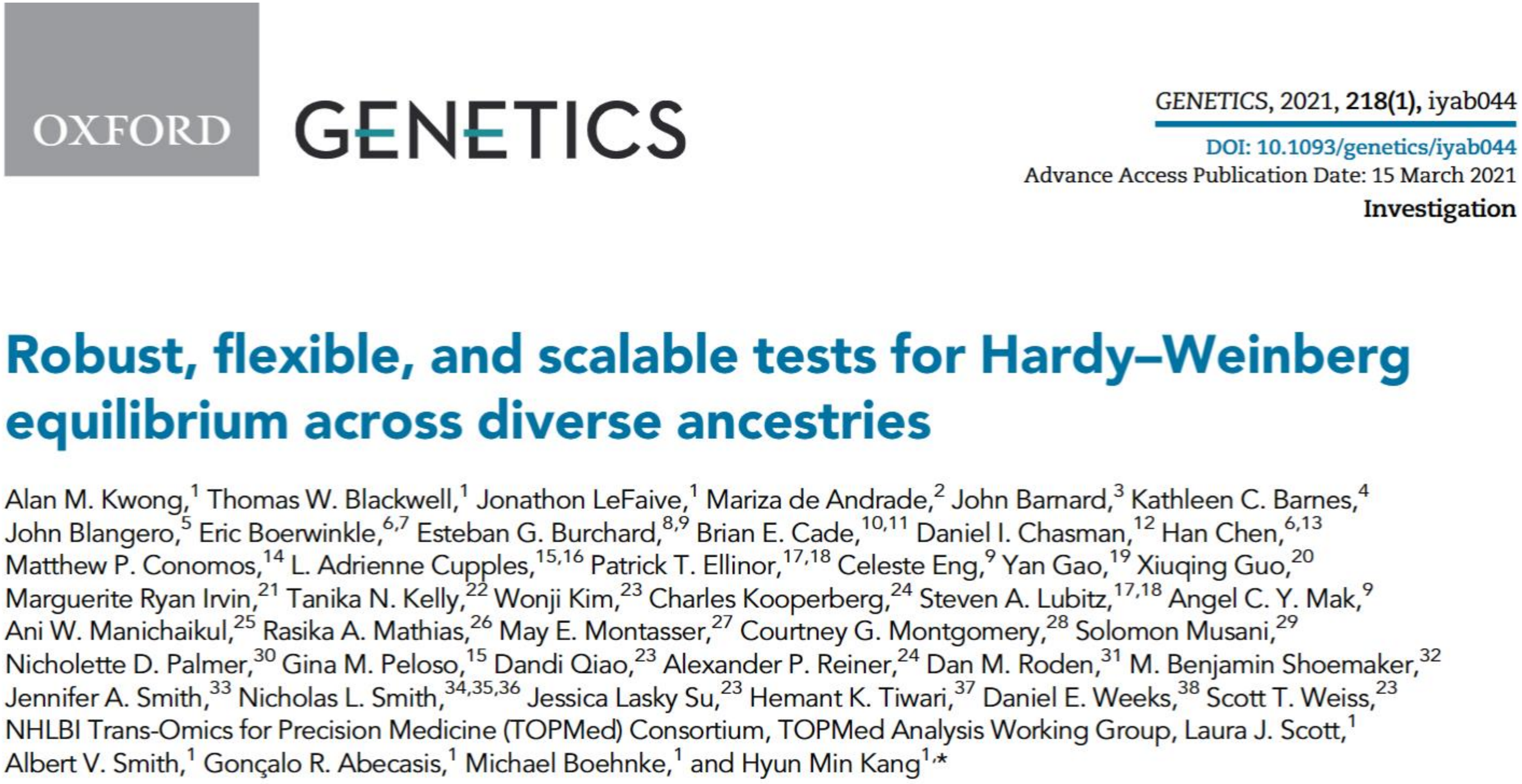
Study	Sample size (n)	AA	AB	BB
1	100	36	48	16
2	100	30	60	10
3	100	45	30	25

- For all these, $\hat{p} = 0.6, n = 100$
- For all studies, $e_{AB} = 100 \times 2 \times 0.6 \times 0.4 = 48,$
- For Study 1, $D = 0, X^2 = 0$
- For Study 2, $D = \frac{60-48}{2} = 6, X^2 = \frac{6^2}{.6^2 \times .4^2 \times 100} = 6.25$
- For Study 3, $D = \frac{30-48}{2} = -9, X^2 = \frac{9^2}{.6^2 \times .4^2 \times 100} = 14.0625$

Summary of HWE tests

- Testing for HWE needs **specialized** tests.
  - DON'T try to use general tests for contingency tables!
- There are multiple available tests. We looked at the LR test, the Chi-square test, and the exact test.
- If the **sample size is low**, the **exact test** will be best.
- We only considered two alleles, but all the tests are extendable to more than two alleles.
  - The **HardyWeinberg package** only offers support for two alleles.
  - The exact test is slow to compute if the sample size is large and there are multiple alleles.

Testing for HWE is typically applied as a quality control. We expect most SNPs to be in HWE.



OXFORD GENETICS

GENETICS, 2021, **218(1)**, iyab044
DOI: 10.1093/genetics/iyab044
Advance Access Publication Date: 15 March 2021
Investigation

Robust, flexible, and scalable tests for Hardy–Weinberg equilibrium across diverse ancestries

Alan M. Kwong,[1] Thomas W. Blackwell,[1] Jonathon LeFaive,[1] Mariza de Andrade,[2] John Barnard,[3] Kathleen C. Barnes,[4] John Blangero,[5] Eric Boerwinkle,[6,7] Esteban G. Burchard,[8,9] Brian E. Cade,[10,11] Daniel I. Chasman,[12] Han Chen,[6,13] Matthew P. Conomos,[14] L. Adrienne Cupples,[15,16] Patrick T. Ellinor,[17,18] Celeste Eng,[9] Yan Gao,[19] Xiuqing Guo,[20] Marguerite Ryan Irvin,[21] Tanika N. Kelly,[22] Wonji Kim,[23] Charles Kooperberg,[24] Steven A. Lubitz,[17,18] Angel C. Y. Mak,[9] Ani W. Manichaikul,[25] Rasika A. Mathias,[26] May E. Montasser,[27] Courtney G. Montgomery,[28] Solomon Musani,[29] Nicholette D. Palmer,[30] Gina M. Peloso,[15] Dandi Qiao,[23] Alexander P. Reiner,[24] Dan M. Roden,[31] M. Benjamin Shoemaker,[32] Jennifer A. Smith,[33] Nicholas L. Smith,[34,35,36] Jessica Lasky Su,[23] Hemant K. Tiwari,[37] Daniel E. Weeks,[38] Scott T. Weiss,[23] NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Analysis Working Group, Laura J. Scott,[1] Albert V. Smith,[1] Gonçalo R. Abecasis,[1] Michael Boehnke,[1] and Hyun Min Kang[1,*]

Multiple testing

With focus on Bonferroni adjustment

Credit: Browning, Temple

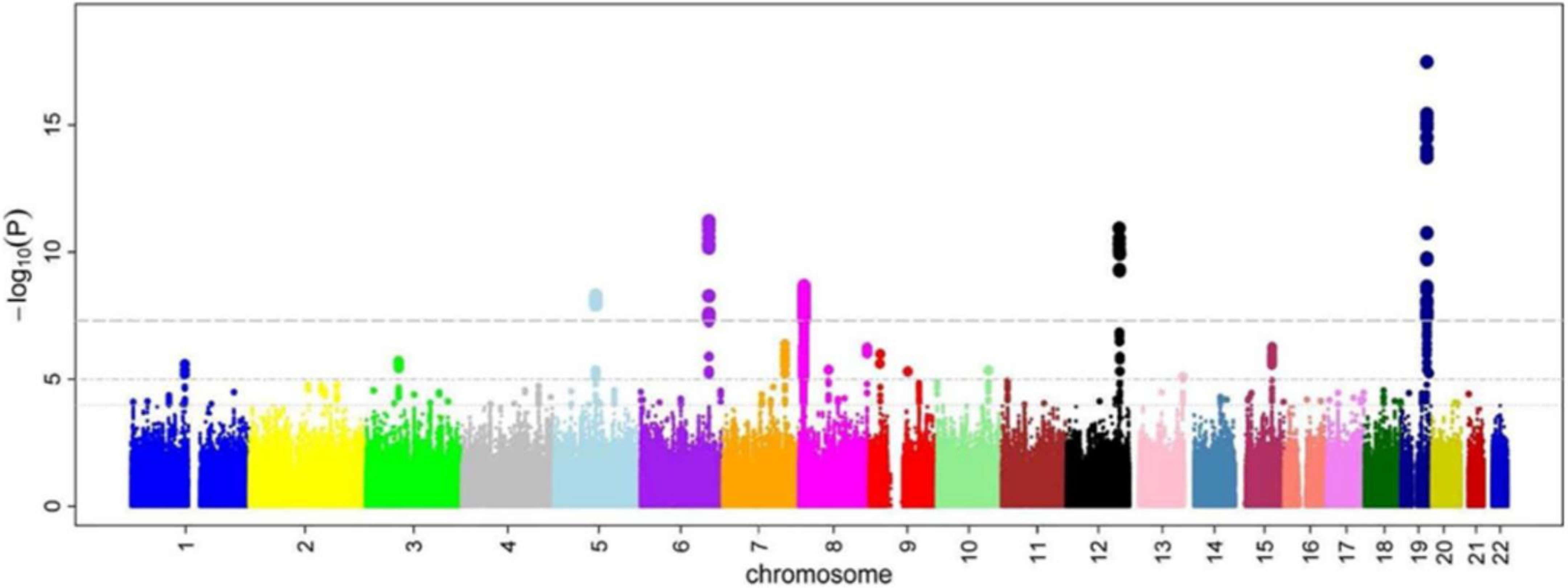
Multiple testing

- If the null hypothesis (e.g., no association between allele and disease) is true, the p-value for one test should be approx. uniformly distributed between 0 and 1.
- So, if testing many SNPs, and the null hypothesis is true for all of them
  - Around 5% will give p-values less than 0.05
  - Around 1% will give p-values less than 0.01
- So, using a standard p-value threshold will result in **a lot of false positive results** (which may or may not be a problem).
- The concept of **multiple testing is REALLY important** in many genetics, because **the genome is BIG**, and we are often scanning the genome rather than testing a small number of plausible loci.

Bonferroni correction

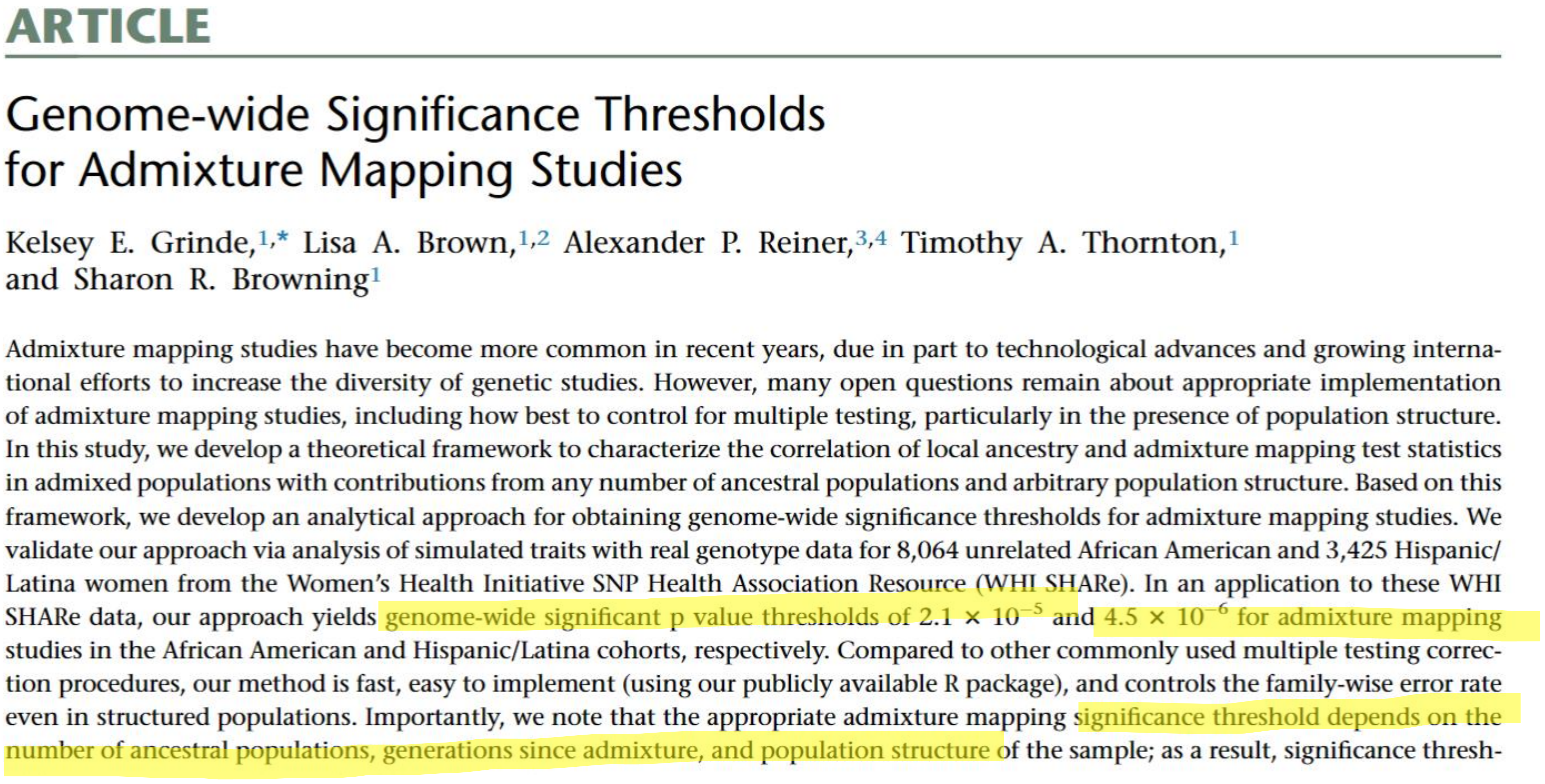
- Suppose $X_i \sim \mathrm{Unif}(0,1)$ independently, $i = 1, \dots, n$
- $P(\min X_i \leq x) = 1 - P(\min X_i > x) = 1 - \prod_{i=1}^{n} P(X_i > x)$
  $= 1 - (1 - x)^n \text{ for } 0 \leq x \leq 1$
  $\approx xn$ for small x.
- To control family-wise error rate (FWER), want $P(\min p_i \leq x) = \alpha$, which implies $x \approx \alpha/n$ is the appropriate threshold.
- Assume independent tests under null hypothesis.

GWAS significance threshold: 5e-8



Multiple testing correction in genome-wide association studies

- 5×10^{-8} has been a standard p-value threshold.
  - **Where does this come from?**
- May need different thresholds for
  - Different populations
  - Array vs. WGS vs. WES
  - Sex chromosomes, mtDNA
  - Different analyses (admixture mapping, IBD mapping, etc.)



ARTICLE

Genome-wide Significance Thresholds for Admixture Mapping Studies

Kelsey E. Grinde,[1,*] Lisa A. Brown,[1,2] Alexander P. Reiner,[3,4] Timothy A. Thornton,[1] and Sharon R. Browning[1]

Admixture mapping studies have become more common in recent years, due in part to technological advances and growing international efforts to increase the diversity of genetic studies. However, many open questions remain about appropriate implementation of admixture mapping studies, including how best to control for multiple testing, particularly in the presence of population structure. In this study, we develop a theoretical framework to characterize the correlation of local ancestry and admixture mapping test statistics in admixed populations with contributions from any number of ancestral populations and arbitrary population structure. Based on this framework, we develop an analytical approach for obtaining genome-wide significance thresholds for admixture mapping studies. We validate our approach via analysis of simulated traits with real genotype data for 8,064 unrelated African American and 3,425 Hispanic/Latina women from the Women's Health Initiative SNP Health Association Resource (WHI SHARe). In an application to these WHI SHARe data, our approach yields genome-wide significant p value thresholds of 2.1×10^{-5} and 4.5×10^{-6} for admixture mapping studies in the African American and Hispanic/Latina cohorts, respectively. Compared to other commonly used multiple testing correction procedures, our method is fast, easy to implement (using our publicly available R package), and controls the family-wise error rate even in structured populations. Importantly, we note that the appropriate admixture mapping significance threshold depends on the number of ancestral populations, generations since admixture, and population structure of the sample; as a result, significance thresh-

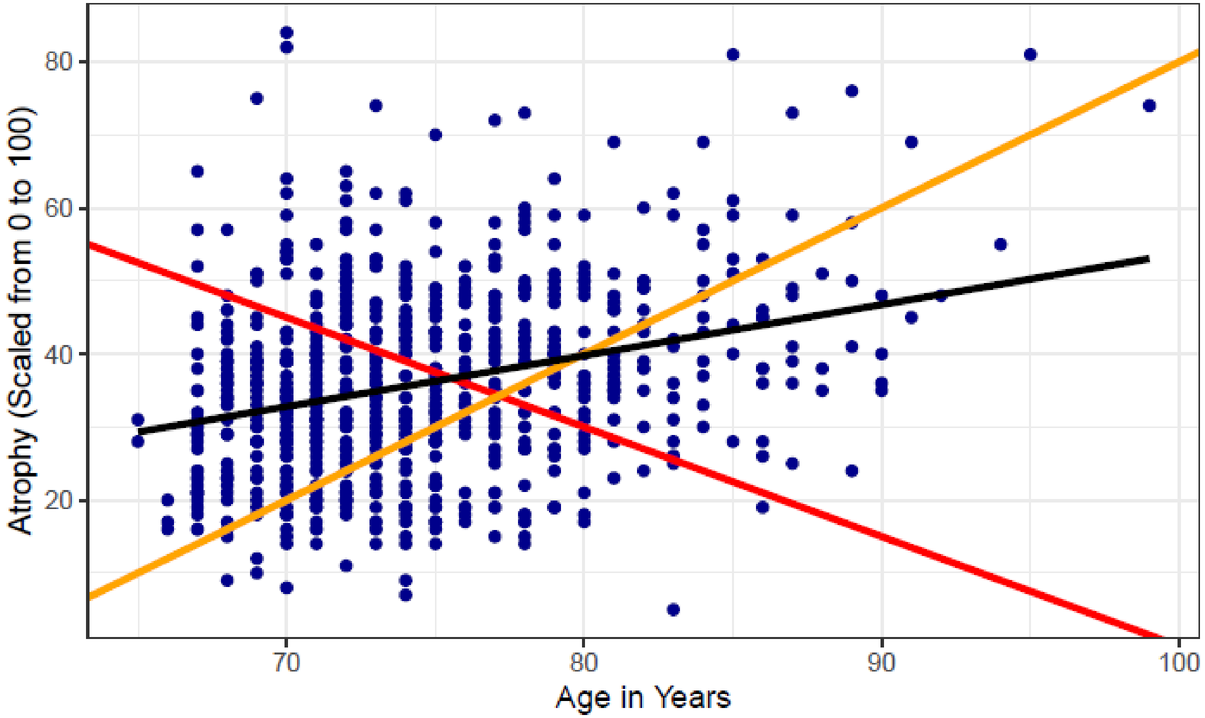
Linear models

With focus on interpretation

Credit: Galanter, Okonek, Wolock

Simple linear regression

Simple linear regression is a statistical tool that allows us to estimate the “best” fitting line through two variables. In general, we use linear regression with **quantitative** outcomes. Below we show our guesses for the best fitting line and the line estimated from simple linear regression in **black**:



Simple linear regression

Lines take the form $y = a + bx$. When writing regression models we'll use the notation $E[Y \mid X] = \beta_0 + \beta_1 X$. Translating this, we have

- β_0: The intercept of the linear regression line
- β_1: The slope of the linear regression line
- $E[Y]$: The average value of Y in the population
- $E[Y \mid X]$: The average value of Y in the population, given the predictor X

Simple linear regression

Lines take the form $y = a + bx$. When writing regression models we'll use the notation $E[Y \mid X] = \beta_0 + \beta_1 X$. Translating this, we have

- β_0: The intercept of the linear regression line
- β_1: The slope of the linear regression line
- $E[Y]$: The average value of Y in the population
- $E[Y \mid X]$: The average value of Y in the population, given the predictor X
- Examples: Y is atrophy score, X is whether someone has been diagnosed with congestive heart failure. The following might be true for the population of US adults over 65
  - $E[Y] = 35$ points (average score in the population)
  - $E[Y \mid X = 0] = 34$ points (average score for those in the population not diagnosed with CHF)
  - $E[Y \mid X = 1] = 40$ points (average score for those in the population diagnosed with CHF)

Coefficient interpretation

The coefficients (β_0, β_1) in our simple linear regression model $E[Y \mid X] = \beta_0 + \beta_1 X$ often have useful interpretations.

How do you interpret β_0 and β_1?

(Hint: think about how we interpret a and b in $y = a + bx$)

- β_0 is the mean value of Y among subjects with $X = 0$
- β_1 is the difference in mean value of Y comparing two groups that differ by one unit in X

Interpreting slopes: mathematical explanation

For a regression model $E[Y \mid X] = \beta_0 + \beta_1 X$, we noted that the interpretation of the slope β_1 is the difference in mean value of Y comparing two groups that differ by one unit in X.

Why is this the correct interpretation? Let's do some algebra...

- $E[Y \mid X = x] = \beta_0 + \beta_1 x$
- $E[Y \mid X = (x + 1)] = \beta_0 + \beta_1(x + 1) = \beta_0 + \beta_1 x + \beta_1$
- $E[Y \mid X = (x + 1)] - E[Y \mid X = x] = \beta_1$

Practice: interpreting intercepts in context

Suppose we fit a linear regression model with atrophy score (0 to 100) as our outcome, and age in years as our predictor:

$$E[\text{atrophy} \mid \text{age}] = -16.06 + 0.70 \times \text{age}$$

Pollev: Which of the following is the correct interpretation?

1. A person of age 0 will have atrophy score equal to -16.06.
2. Among all people of age 0, the average atrophy score is -16.06.

Practice: interpreting intercepts in context

Suppose we fit a linear regression model with atrophy score (0 to 100) as our outcome, and age in years as our predictor:

$$E[\text{atrophy} \mid \text{age}] = -16.06 + 0.70 \times \text{age}$$

Pollev: Which of the following is the correct interpretation?

1. ~~A person of age 0 will have atrophy score equal to -16.06.~~
2. Among all people of age 0, the average atrophy score is -16.06.

Practice: interpreting slopes in context

Suppose we fit a linear regression model with atrophy score (0 to 100) as our outcome, and age in years as our predictor:

$$E[\text{atrophy} \mid \text{age}] = -16.06 + 0.70 \times \text{age}$$

Pollev: Which of the following is the correct interpretation?

1. For every one year increase in age, average atrophy score increases by 0.70 points.
2. Comparing two groups of people who differ by one year in age, the difference in average atrophy score will be 0.70 points, with the higher average score in the older of the two groups.

Practice: interpreting slopes in context

Suppose we fit a linear regression model with atrophy score (0 to 100) as our outcome, and age in years as our predictor:

$$E[\text{atrophy} \mid \text{age}] = -16.06 + 0.70 \times \text{age}$$

Pollev: Which of the following is the correct interpretation?

1. ~~For every one year increase in age, average atrophy score increases by 0.70 points.~~
2. Comparing two groups of people who differ by one year in age, the difference in average atrophy score will be 0.70 points, with the higher average score in the older of the two groups.

Why?

This was an observational study! The word “increases” in the first interpretation implies causality, which we cannot assume in an observational study.

Why do we care about the slope?

When

- $\beta_1 = 0$: there is no linear association between X and Y
- $\beta_1 > 0$: there is a positive linear association between X and Y
- $\beta_1 < 0$: there is a negative linear association between X and Y

Consider the **scientific question:** *is there an association between atrophy and age?*

We can use linear regression to answer this question:

1. Fit the model $E[\text{atrophy} \mid \text{age}] = \beta_0 + \beta_1 \times \text{age}$
2. Check whether or not $\beta_1 = 0$ (estimate, CI, p-value)*

* Note: our **null hypothesis** here (that there is *no* association between atrophy and age) can be written as $H_0 : \beta_1 = 0$

Hypothesis tests

Consider the regression model:

$$E[Y \mid X] = \beta_0 + \beta_1 X$$

We know that β_1 quantifies the association between X and Y.

To test whether there is a *statistically significant* association between X and Y, we need to test whether $\beta_1 = 0$:

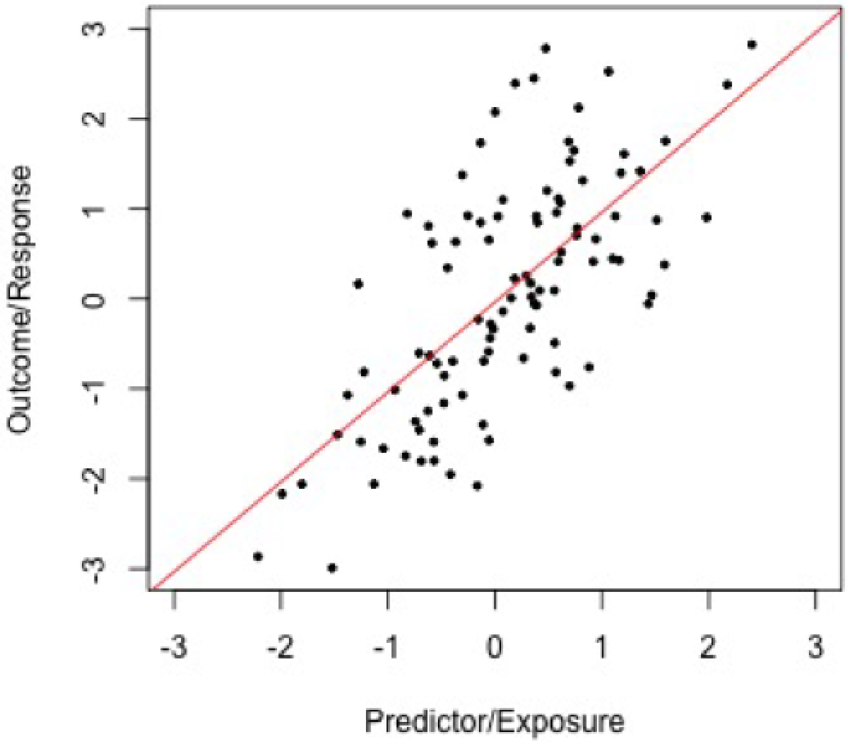
- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

Interpret the p-value as we have before in Chapter 0:

- $p < \alpha$: **reject** H_0, "we have evidence to suggest that X is associated with Y"
- $p > \alpha$: **fail to reject** H_0, "we do not have enough evidence to support the hypothesis that X is associated with Y"

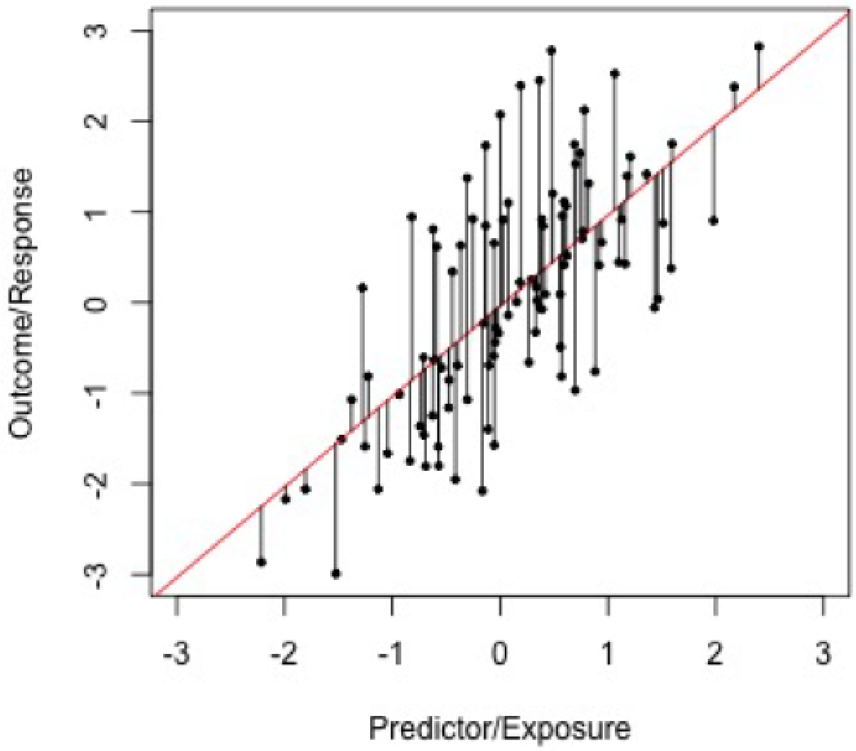
Least squares estimation

What is R doing *under the hood* to get these regression coefficient estimates?



Least squares estimation

Least squares: minimize the sum of the squared distances from the observed points to the fitted line



Least squares estimation: technical details

Least squares: minimize the sum of the squared distances from the observed points to the fitted line

Breaking this down...

- The distance from an observed point for an individual i to the fitted line is $Y_i - [\beta_0 + \beta_1 X_i]$
- Squaring the distance we get $(Y_i - [\beta_0 + \beta_1 X_i])^2$
- Summing the squared distances over all observations we get $\sum_{i=1}^{n}(Y_i - [\beta_0 + \beta_1 X_i])^2$
- Minimizing we get

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n}(Y_i - [\beta_0 + \beta_1 X_i])^2$$

This has a closed form solution given by $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, $\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$

Statisticians love to tell you about **assumptions** they take about data

Linear regression assumptions

We can write the *classical* linear regression assumptions under the following model as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \text{ for } i = 1, \ldots n$$

- Linearity: There is a linear relationship between X_i and the true population conditional mean $Y_i \mid X_i$
  - i.e. $E[Y_i \mid X_i] = \beta_0 + \beta_1 X_i$
- Independence: The errors ϵ_i are independent of each other
- Normality: ϵ_i are normally distributed
- Equal variance (homoscedasticity): ϵ_i have the same variance
  - $\epsilon_i \sim N(0, \sigma^2)$, where σ^2 does not depend on i

These are commonly called the LINE assumptions, which should hopefully make them easier to remember!

Independence assumption

Independence assumption: The errors ϵ_i, $i = 1, \ldots, n$ are independent of each other.

What does it *mean* for this assumption to be violated?

> Violating the independence assumption means that our observations are sampled in a way such that their responses are *dependent*. Here are a few examples of when this might happen:
>
> - We observe multiple individuals over time, and collect data on their outcomes at multiple time points. We expect the responses for a given individual to be *dependent* over time (e.g. my weight tomorrow *depends* on my weight today)
> - We collect standardized test scores from individuals in multiple schools. We expect students within the same to school (perhaps) to score similarly, as they've had similar educational experiences

Normality assumption

Main concern is small sample size !!!

Normality assumption: The errors ϵ_i, $i = 1, \ldots, n$ are normally distributed.

What does it *mean* for this assumption to be violated?

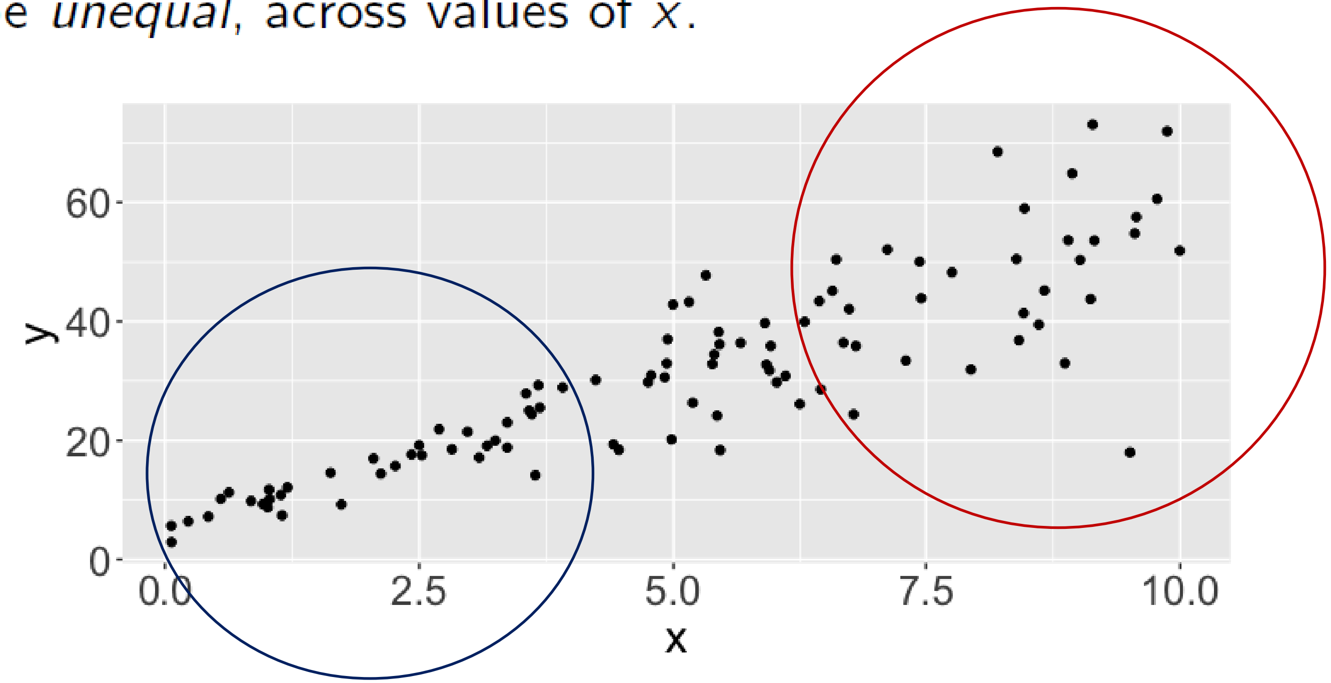
> Violating the normality assumption means that our errors are not normally distributed. We like our errors to be normally distributed because our confidence intervals and hypothesis tests are based on the Normal or t distribution, which assume underlying Normal data.

Question: But what about the Central Limit Theorem? We know from the CLT that the sampling distribution of the sample mean is approximately normal in large samples. . . does the CLT also apply to regression coefficients?

Equal variance assumption

How can we tell by looking at a scatterplot of our outcome (y) vs. our predictor (x)?

If our errors ϵ_i have *unequal* variance, the spread of the y values should be *unequal*, across values of x.



Linear regression assumptions: Recap

The *classical* linear regression (LINE) assumptions are:

- Linearity: There is a linear relationship between X_i and the true population conditional mean $Y_i \mid X_i$
- Independence: The errors are independent of each other
- Normality: The errors are normally distributed
- Equal variance: The errors have the same variance

Diagnostics for each of these assumptions are:

- Linearity: Scatterplot of fitted values vs. residuals
- Independence: Think about the data collection!
- Normality: Histogram of residuals
- Equal variance: Scatterplot of fitted values vs. residuals

Why transform X and/or Y?

Why might we want to transform our variables?

- Transforming can *make our interpretation of* β_0 and β_1 *more scientifically meaningful* (e.g., so we're not talking about the brain atrophy of someone who is 0 years old)

- Transforming can *change the units of the outcome and/or predictor* (e.g., difference in atrophy, comparing two groups that differ by one year in age or two groups that differ by one month in age)

- The relationship between the outcome and the predictor may be nonlinear, and transforming the variables can more accurately match the true relationship.

Simple logistic regression: interpretation

Suppose we fit the simple logistic regression model

$$\log\left(\text{Odds}[Y = 1 \mid X]\right) = \beta_0 + \beta_1 X,$$

where our outcome Y is binary. What is the interpretation of the following quantities?

1. β_0: log odds of $Y = 1$ when $X = 0$
2. β_1: difference in log odds of $Y = 1$ comparing $X = 1$ and $X = 0$ groups
3. $\exp(\beta_0)$: odds of $Y = 1$ when $X = 0$
4. $\exp(\beta_1)$: ratio of odds of $Y = 1$ comparing $X = 1$ and $X = 0$ groups (in other words, the odds ratio!)

Multiple logistic regression

$$\log\left(\text{Odds}[Y=1|X_1,\cdots,X_p]\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p$$

Interpretation:

- β_1: difference in log odds (of $Y=1$) comparing groups that differ by 1 unit in X_1 but are the same with respect to $X_2,\ldots,X_p$

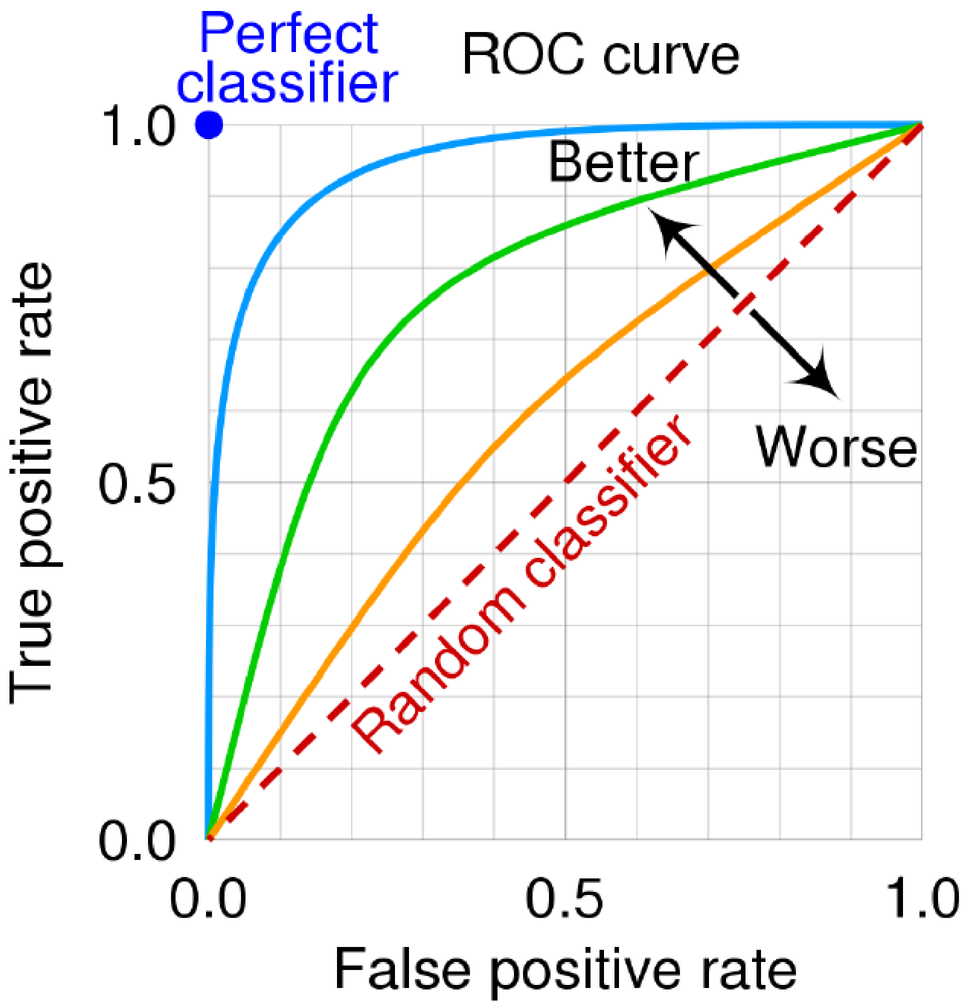
$$\begin{aligned}&\log\left(\text{Odds}[Y=1|X_1=x_1+1, X_2=x_2\cdots, X_p=x_p]\right)\\ &\qquad - \log\left(\text{Odds}[Y=1|X_1=x_1, X_2=x_2\cdots, X_p=x_p]\right)\\ &= (\beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + \cdots \beta_p x_p) - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p)\\ &= \beta_1\end{aligned}$$

- e_1^{β}: odds ratio (of $Y=1$) comparing groups that differ by 1 unit in X_1 but are the same with respect to $X_2,\ldots,X_p$

$$e^{\beta_1} = \exp\Big(\log\left(\text{Odds}[Y=1|X_1=x_1+1, X_2=x_2\cdots, X_p=x_p]\right)$$

Hard to interpret

ROC curve


Perfect classifier
ROC curve
1.0
0.5
0.0
Better
Worse
Random classifier
True positive rate
0.0
0.5
1.0
False positive rate

Generalized linear models: error distribution

Besides the link function and mean model, GLMs also require a distribution on the outcome in the model:

- Linear regression: Y has a Normal distribution with mean $E[Y \mid X] = \beta_0 + \beta_1 X$ and variance σ^2.
- Logistic regression: Y has a Binomial distribution with probability (mean) $\text{logit}(E[Y \mid X]) = \beta_0 + \beta_1 X$
- 🌶️ Poisson regression: Y has a Poisson distribution with probability (mean) $\log(E[Y \mid X]) = \beta_0 + \beta_1 X$

If you're unfamiliar with the Poisson distribution, that's okay! Just know that there are three essential pieces of a GLM:

1. link function
2. mean model
3. outcome distribution

How linear models fit into StatGen

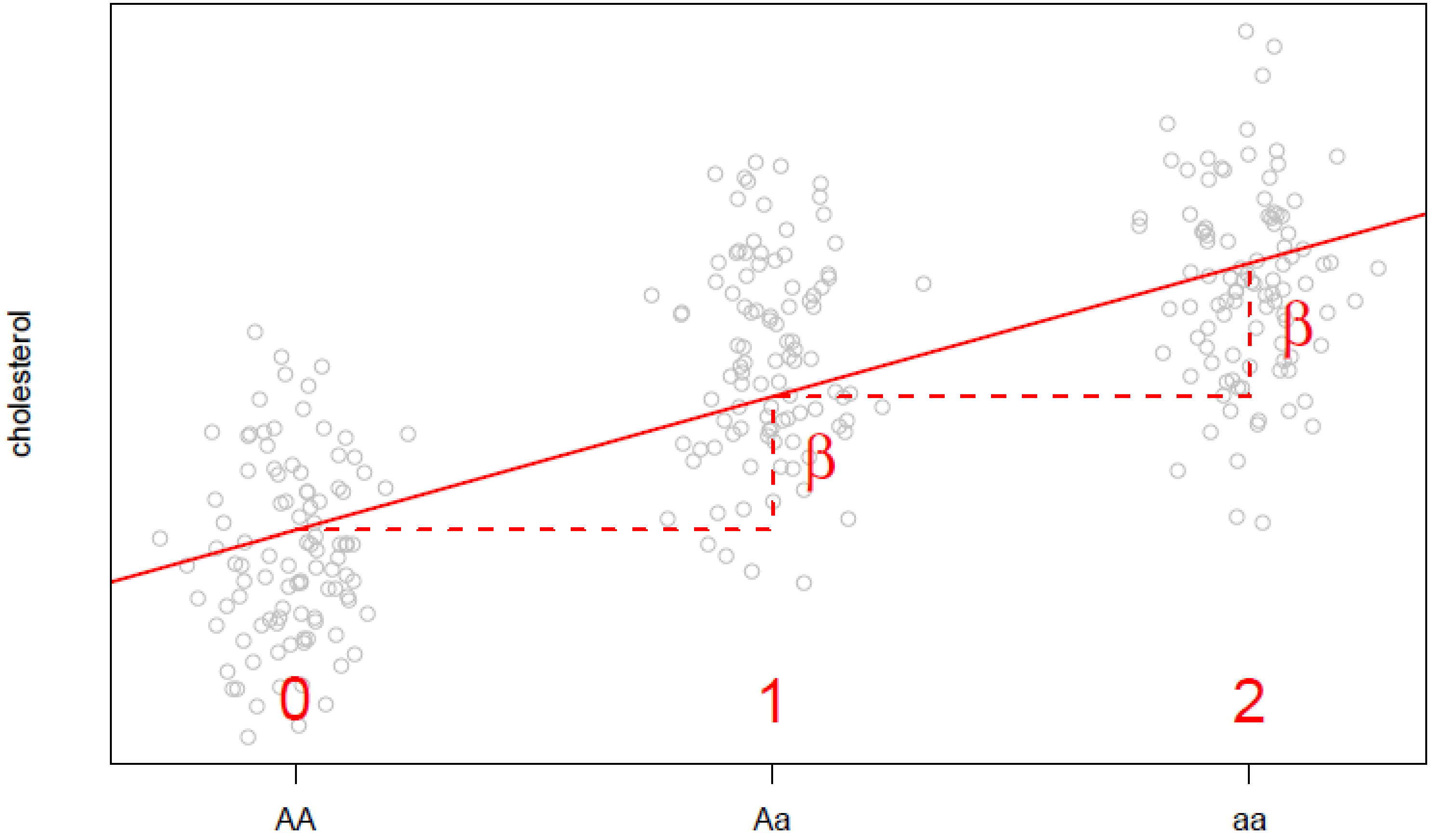
- Some genome-wide hypothesis testing is based on linear models
  - Know how to interpret the estimate
- Polygenic risk scores will be formed by fitting a linear model on summary statistics from prior GWAS
  - Know how to interpret risk stratification
  - Know how to communicate risk stratification to others
- In the generalized linear model framework, we can study data beyond case/control !!!

Credit: Thornton

Linear regression with SNPs

Many analyses fit the ‘additive model’

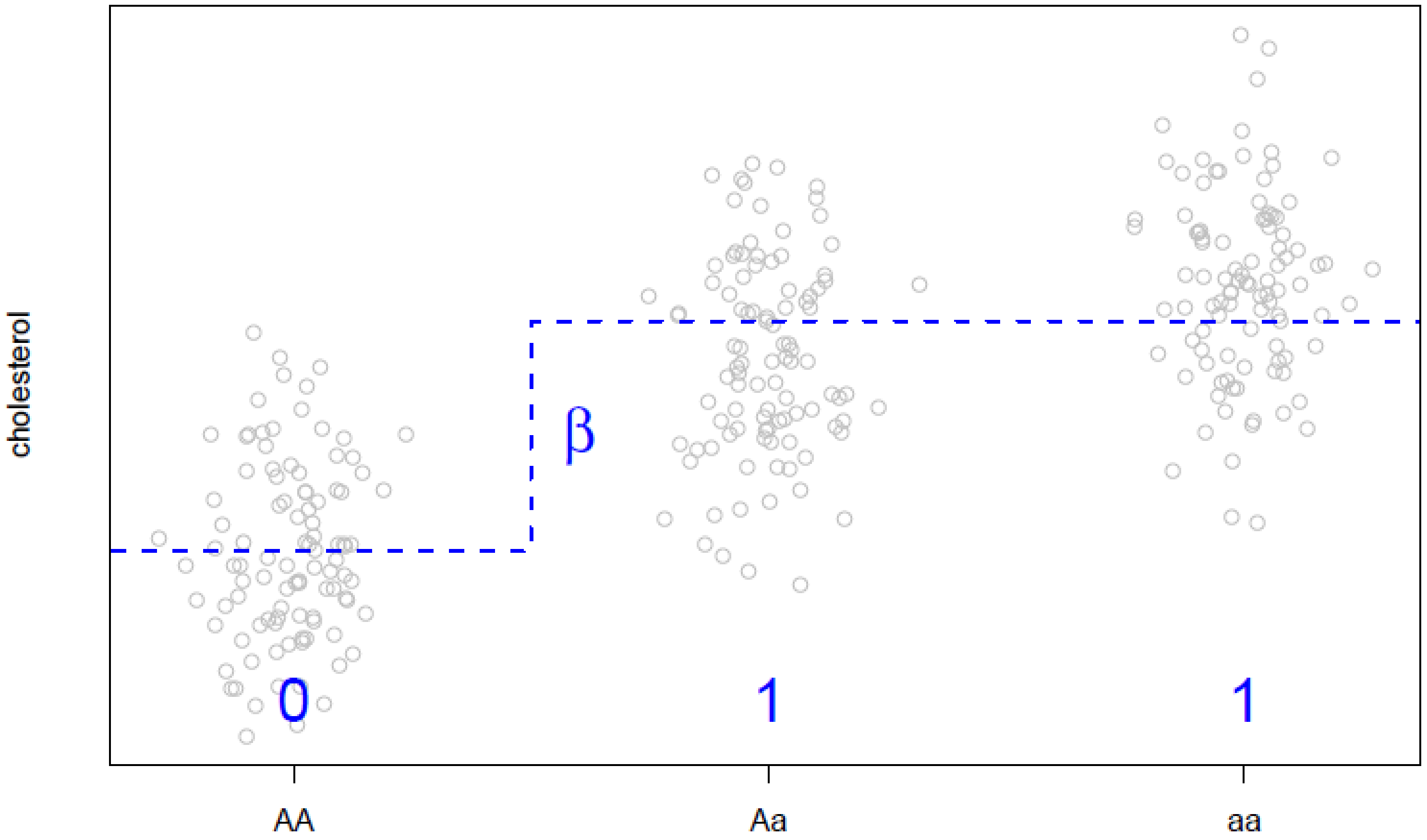
$$y = \beta_0 + \beta \times \#\text{minor alleles}$$



Linear regression, with SNPs

An alternative is the 'dominant model';

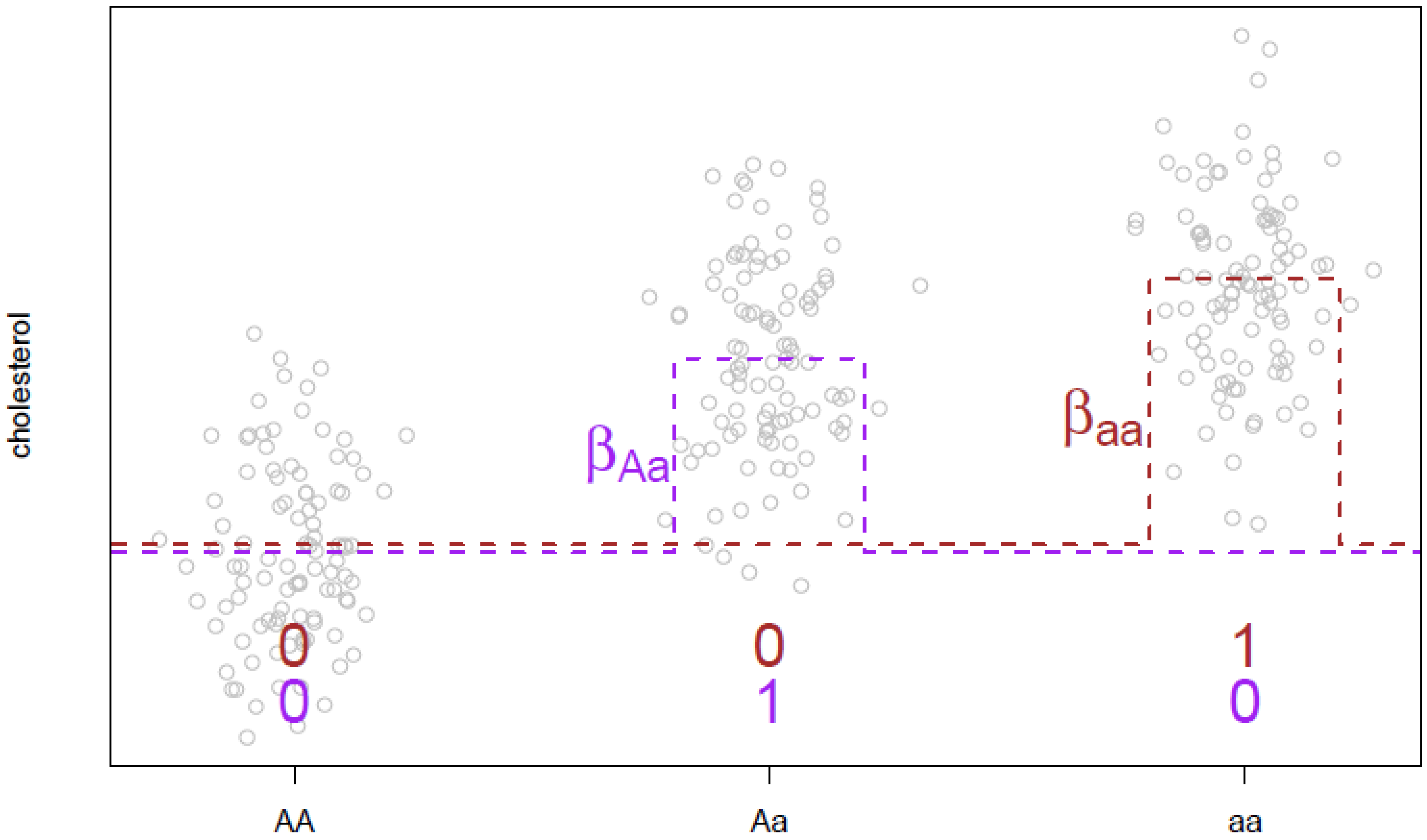
$$y = \beta_0 + \beta \times (G \neq AA)$$



Linear regression, with SNPs

Finally, the 'two degrees of freedom model';

$$y = \beta_0 + \beta_{Aa} \times (G == Aa) + \beta_{aa} \times (G == aa)$$



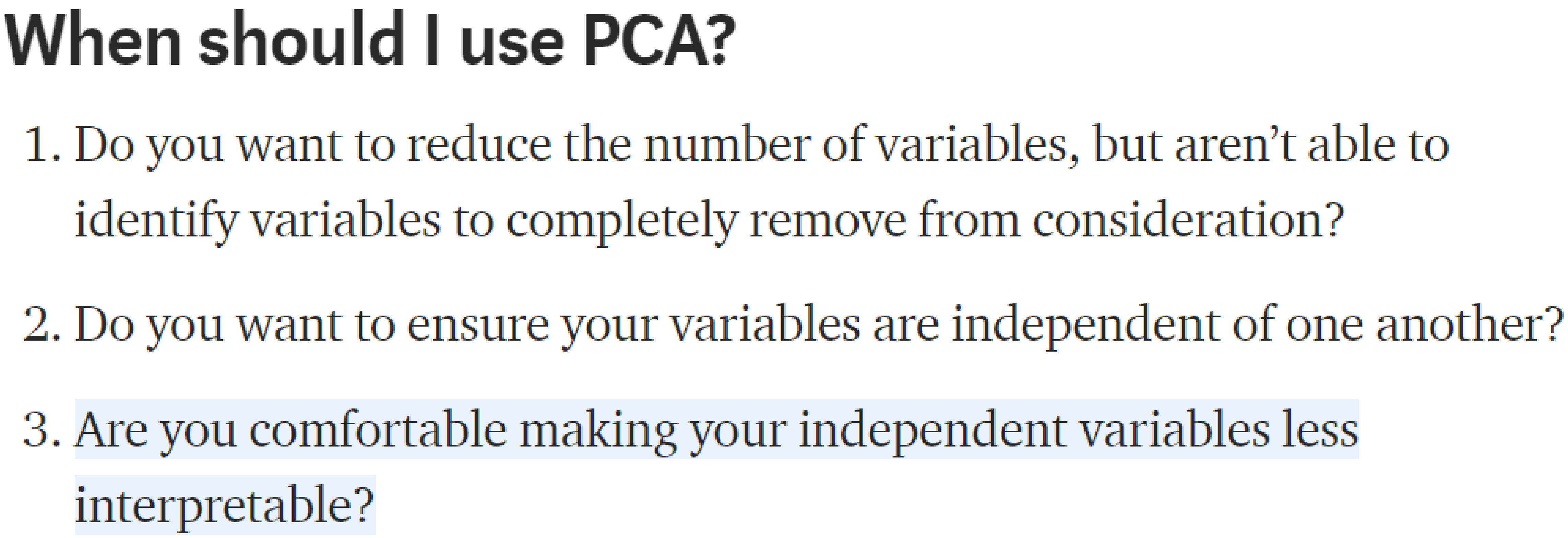
Principal components analysis

With focus on population stratification

Credit: Temple

Dimensionality Reduction

- Address overfitting
- Either eliminate or extract features
- Simpler model is easier to interpret
  - Use PCs (lose interpretability)
- Rank deficiency (e.g., linear regression)
  - Use PCs (lose interpretability)



When should I use PCA?

1. Do you want to reduce the number of variables, but aren’t able to identify variables to completely remove from consideration?
2. Do you want to ensure your variables are independent of one another?
3. Are you comfortable making your independent variables less interpretable?

Singular value decomposition

Theorem (Singular value decomposition): Consider a matrix $\mathbf{X} \in \mathbb{R}^{\mathbf{n}\times\mathbf{p}}$. We can write $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where

- $\mathbf{U}$ is an orthogonal $n \times n$ matrix ($\mathbf{U}^T\mathbf{U} = \mathbf{I}$).
- $\mathbf{V}$ is an orthogonal $p \times p$ matrix ($\mathbf{V}^T\mathbf{V} = \mathbf{I}$).
- $\mathbf{D}$ is "diagonal" ($D_{ij} = 0$ for all $i \neq j$) with nonnegative diagonals ($D_{ii} \geq 0$ for all $i \leq \min\{n, p\}$)

By convention, we write order the diagonals of $\mathbf{D}$ to be non-decreasing: $D_{11} \geq D_{22} \geq \ldots$.

Credit: Willis

Principal component analysis is **SVD** with preprocessing so that features are sorted by variance explained

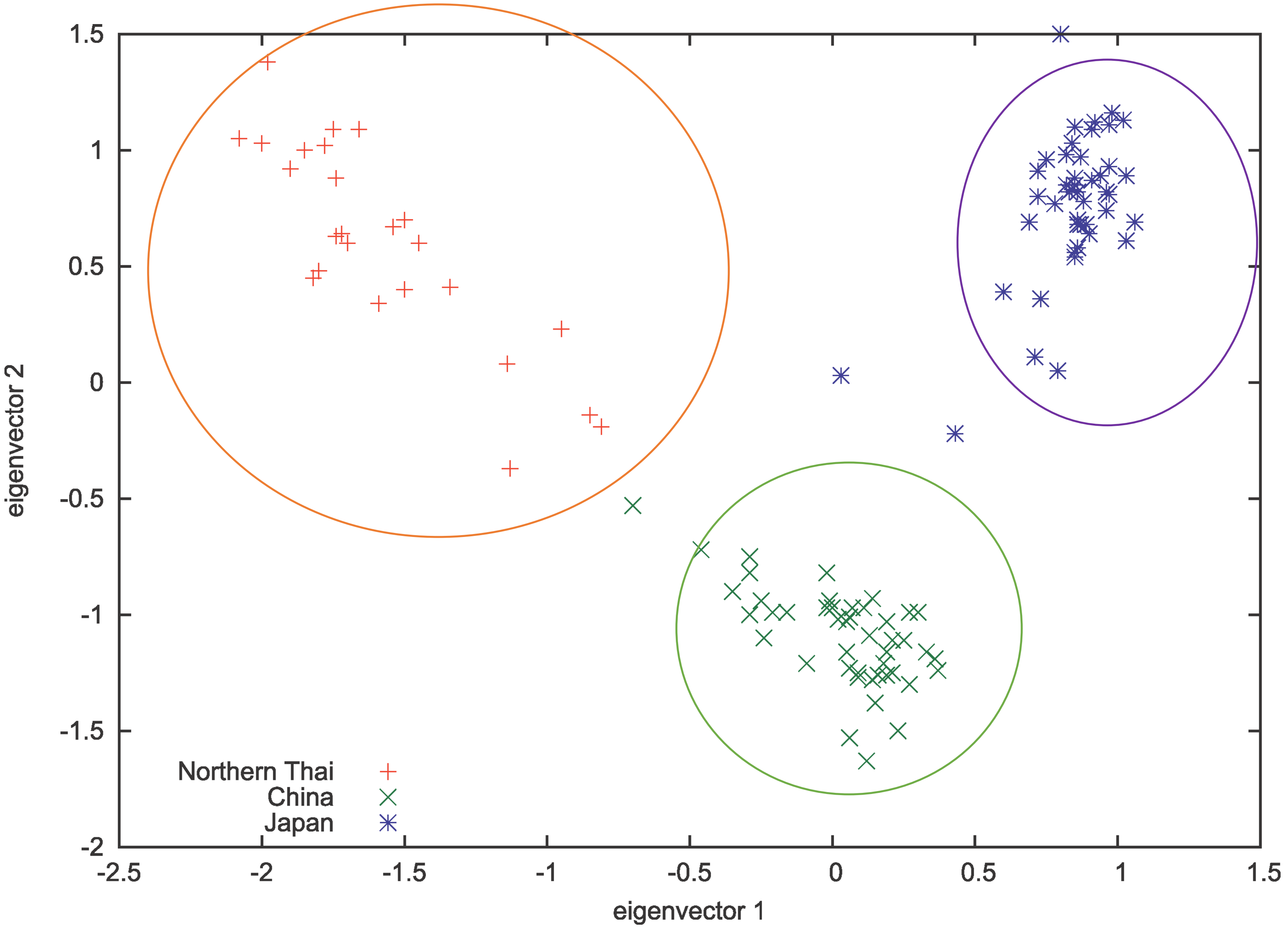
Comments on PCA

- The goal of PCA is to find an orthogonal (perpendicular) matrix U that determines a change of variable such that $Z = XU$
  - The columns of Z are ordered according to increasing variability
  - The columns of Z are uncorrelated
  - Each column of Z is a *linear combination* of the columns of X
  - *Assumes* more variability in a direction better explains the response
- We may use some of columns of Y, called the principal components
  - Use features that explain the most variance

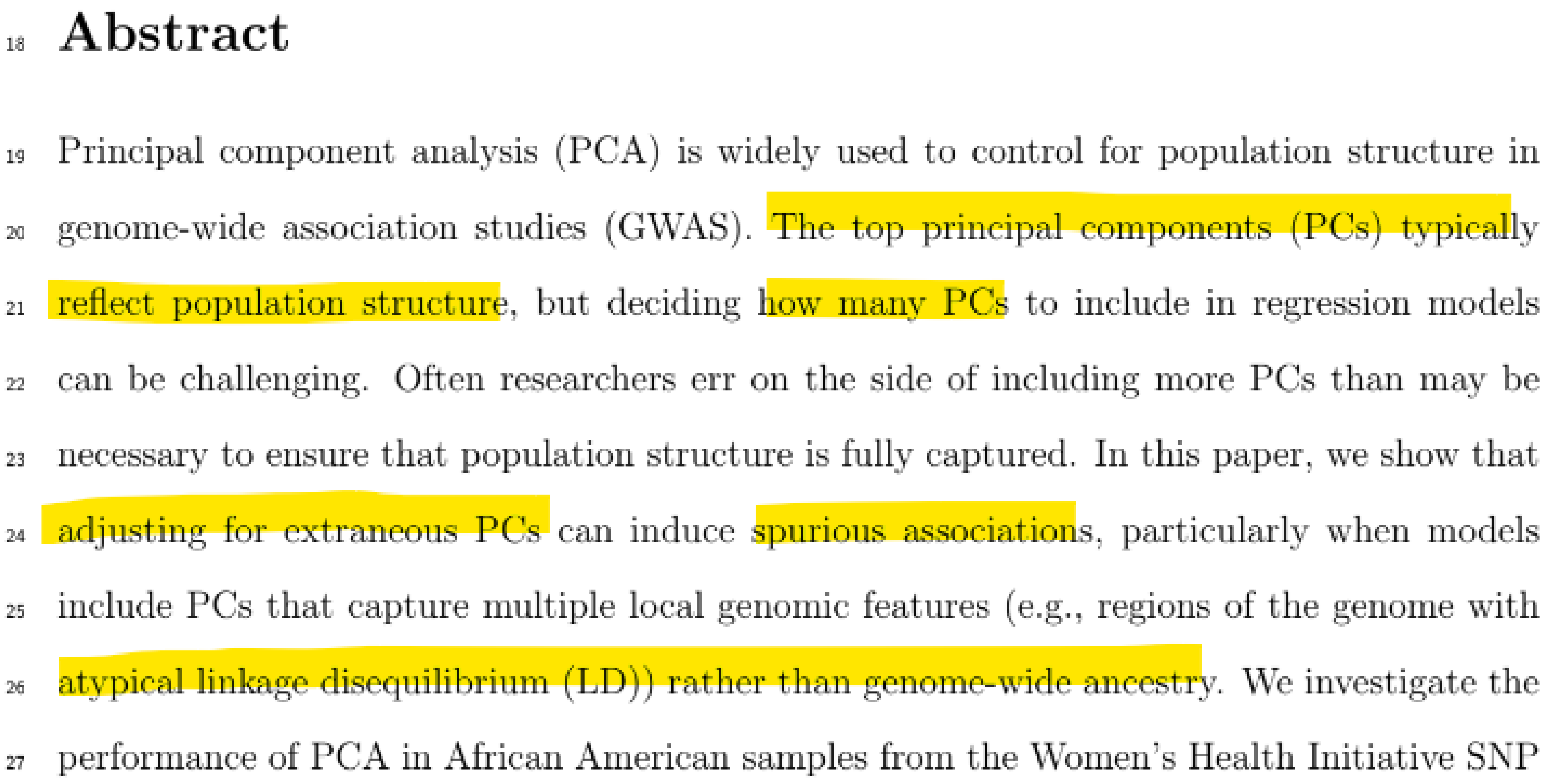
How PCs are used in StatGen

- (Continental) ancestry inference
  - 1st PC separates Out of Africa bottleneck
  - 2nd PC separates Europe and Asia
  - Higher order PCs: different studies?, chromosomal inversions?
  - This sublist assumes global sample, say 1000 Genomes
- Quality control
  - Some clustering in higher order PCs may indicate experiment issues
- Linear models conditional on PCs
  - Hoping to control for population stratification

PCA to infer ancestral populations from dense SNP-array data



Grinde et al. (2023+)



Abstract

Principal component analysis (PCA) is widely used to control for population structure in genome-wide association studies (GWAS). The top principal components (PCs) typically reflect population structure, but deciding how many PCs to include in regression models can be challenging. Often researchers err on the side of including more PCs than may be necessary to ensure that population structure is fully captured. In this paper, we show that adjusting for extraneous PCs can induce spurious associations, particularly when models include PCs that capture multiple local genomic features (e.g., regions of the genome with atypical linkage disequilibrium (LD)) rather than genome-wide ancestry. We investigate the performance of PCA in African American samples from the Women's Health Initiative SNP

Thank you for having me!

More technical materials here:
[Genetic data analyses for admixed and multiethnic samples]
[Genome-wide association study]