CS 194-172 HW 2

Collaborators: Zaid Ahmad

Sat Oct 4th 2021

Stefan Biclmeier

a) Prove (1) $\frac{e^{\beta_0+\beta_1 \cdot x}}{1+e^{\beta_0+\beta_1 \cdot x}} = (2)\ \frac{1}{1+e^{-\beta_0-\beta_1 \cdot x}}$

$\Rightarrow$ take (1): $\frac{e^{\beta_0+\beta_1 \cdot x} \cdot e^{-\beta_0-\beta_1 \cdot x}}{\left(1+e^{\beta_0+\beta_1 \cdot x}\right)\left(e^{-\beta_0-\beta_1 \cdot x}\right)} = \frac{e^{\beta_0+\beta_1 \cdot x-\beta_0-\beta_1 \cdot x}}{e^{-\beta_0-\beta_1 \cdot x}+e^{\beta_0+\beta_1 \cdot x-\beta_0-\beta_1 \cdot x}}$

$= \frac{e^0}{e^{-\beta_0-\beta_1 x}+e^0} = \frac{1}{1+e^{-\beta_0-\beta_1 \cdot x}} = (2)$ ✓✓

$1 \doteq 1 \doteq 1$

number one, sorry for any confusion.

b) $g(z) = \frac{1}{1+e^{-z}} \Rightarrow g'(z) = \frac{0 \cdot \left(1+e^{-z}\right) - \left(1+e^{-z}\right)' \cdot 1}{\left(1+e^{-z}\right)^2}$

$= \frac{-\left(0-e^{-z}\right)}{\left(1+e^{-z}\right)^2} = \frac{e^{-z}}{\left(1+e^{-z}\right)^2} = \underbrace{\frac{1}{\left(1+e^{-z}\right)}}_{g(z)\ \checkmark} \cdot \underbrace{\left(\frac{e^{-z}}{\left(1+e^{-z}\right)}\right)}_{\text{get it to } 1-g(z)!}$

$= \frac{1}{\left(1+e^{-z}\right)} \cdot \left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) = \frac{1}{1+e^{-z}}\left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}}\right) = \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right)$

$= g(z) \cdot \left(1-g(z)\right)$ ✓✓ $= g'(z)$

SB

1c) $f(x; \beta_0, \beta_1) = \frac{1}{1+e^{-\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x)}}$

$\Rightarrow$ with $z = \beta_0 + \beta_1 \cdot x$

$\Rightarrow f(x; \beta_0, \beta_1) = g(\beta_0 + \beta_1 \cdot x)$ with $g(z) = \frac{1}{1+e^{-z}}$

1d) We will go from general to specific to answer this question.

Probability of observing a 1 (data) given a specific datapoint x & β_0, β_1 (model)

$\Rightarrow P(Y=1 | x; \beta_0, \beta_1) = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x)}}$, using rule of total probability & knowing only 2 outcomes, 0 & 1

$P(Y=0 | x; \beta_0, \beta_1) = 1 - \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x)}}$

$\Rightarrow$ for any given data point (y, x): $P(Y=y | X; \beta_0, \beta_1) = \left(\frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x)}}\right)^y \cdot \left(\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}\right)^{(1-y)}$

(Bernoulli probability mass function)

$\Rightarrow$ Now, for the equivalent of a sequence in Markov model, a dataset with n observations, the likelihood of a logistic regression is:

$\Rightarrow L(\beta_0, \beta_1) = \prod_{i=1}^{n} P(Y=y_i | X=x_i; \beta_0, \beta_1)$, using $\frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = g(\beta_0+\beta_1 \cdot x)$

$= \prod_{i=1}^{n} \left(g(\beta_0+\beta_1 \cdot x_i)\right)^{y_i} \cdot \left(1 - g(\beta_0+\beta_1 x_i)\right)^{1-y_i}$

$\Rightarrow \log L(\beta_0, \beta_1) = \sum_{i=1}^{n} \log\left(g(\beta_0+\beta_1 x_i)^{y_i} \cdot \left(1 - g(\beta_0+\beta_1 \cdot x_i)\right)^{1-y_i} \right)$

$= \sum_{i=1}^{n} \log\left(g(\beta_0+\beta_1 \cdot x_i)^{y_i}\right) + \log\left(\left(1 - g(\beta_0+\beta_1 \cdot x_i)\right)^{1-y_i}\right)$

$= \sum_{i=1}^{n} y_i \log\left[g(\beta_0+\beta_1 \cdot x_i)\right] + (1-y_i) \cdot \log\left[1 - g(\beta_0+\beta_1 \cdot x_i)\right]$

✓✓

SB
Oct 4th
2021
CS194-172

(e) The log function is a monotonically increasing function, as $f(x) = \log x$, and $f'(x) = \frac{1}{x}$ if log is to the base e ~~because~~ With $x > 0$ as per definition of the log, $\frac{1}{x} > 0$! That means if we maximize $\log_e x$ will also maximize. $\Rightarrow$ OK to use, and more convenient to compute.

* which we do, $\max L(\beta_0, \beta_1)$

(f) There are 2 ways to calculate the gradients – one extensive, one simpler by substitution. I'll do one each. $\log = \ln$

$$\frac{\partial \log L(\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} y_i \cdot \underbrace{\frac{1}{g(\beta_0+\beta_1 x_i)}}_{\frac{\partial \log(g)}{\partial g}} \cdot \underbrace{g(\beta_0+\beta_1 x_i)\cdot(1-g(\beta_0+\beta_1 x_i))}_{\frac{\partial g(\beta_0+\beta_1 x_i)}{\partial(\beta_0+\beta_1 x_i)}} \cdot \underbrace{1}_{\frac{\partial \beta_0+\beta_1 x_i}{\partial \beta_0}}$$

$$+ (1-y_i) \cdot \frac{1}{1-g(\beta_0+\beta_1 x_i)} \cdot (-1) \cdot g(\beta_0+\beta_1 x_i) \cdot (1-g(\beta_0+\beta_1 x_i)) \cdot 1$$

$$= \sum_{i=1}^{n} y_i - y_i\, g(\beta_0+\beta_1 x_i) - g(\beta_0+\beta_1 x_i) + y_i\, g(\beta_0+\beta_1 x_i)$$

$$= \sum_{i=1}^{n} y_i - g(\beta_0+\beta_1 x_i) = \sum_{i=1}^{n} y_i - \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$$

for $\frac{\partial \log L(\beta_0,\beta_1)}{\partial \beta_1}$ we use: $f = \log(\quad)$ & $h = \beta_0 + \beta_1 x_i$ with $f' = \frac{\partial f}{\partial g(h)}$ & $h' = \frac{\partial h}{\partial \beta_1}$

$$= \sum_{i=1}^{n} y_i \cdot f'(g(h)) \cdot g'(h) \cdot h' + (1-y_i) \cdot f'(1-g(h)) \cdot (1-g(h))' \cdot h'$$

$$= \sum_{i=1}^{n} y_i \cdot \frac{1}{g(h)} \cdot g(h)(1-g(h)) \cdot \underbrace{\frac{\partial h}{\partial \beta_1}}_{=x_i} + (1-y_i)\frac{1}{1-g(h)} \cdot (-1) \cdot g(h)(1-g(h)) \cdot x_i$$

$$= \sum_{i=1}^{n} \left(y_i - y_i\, g(h) - g(h) + y_i\, g(h)\right) \cdot x_i = \sum_{i=1}^{n} \left(y_i - g(h)\right) \cdot x_i$$

$$= \sum_{i=1}^{n} \left(y_i - \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}\right) \cdot x_i \quad \text{(v)}$$

1g) Initalize $\beta_0 = 0$, $\beta_1 = 0$;

Do:

$$\beta_0^{new} = \beta_0^{old} + \alpha \cdot \sum_{i=1}^{n} y_i - g\left(\beta_0^{old} + \beta_1^{old} \cdot x_i\right)$$

$$\beta_1^{new} = \beta_1^{old} + \alpha \cdot \sum_{i=1}^{n} \left(y_i - g\left(\beta_0^{old} + \beta_1^{old} \cdot x_i\right)\right) \cdot x_i$$

Until: $\| \beta_0^{new} - \beta_0^{old} \| < \text{threshold}$ and $\| \beta_1^{new} - \beta_1^{old} \| < \text{threshold}$

α being step size of gradient ascent.

Homework 2 – ct'd

SB
CompGen
Oct 8th 2021

2a) Prove that FWER $\leq \alpha$ for rejecting all H_0^j at $p_j \leq \frac{\alpha}{m}$ (after Bonferri correction)

FWER = P(rejecting ≥ 1 H_0^j falsely) for all $H_1, \dots, H_m$

$= P(\bigcup_{j=1}^{m}$ rejecting H_0^j falsely$)$

knowing that $P(A \cup B) \leq P(A) + P(B)$, no matter if A and B are independent

$\Rightarrow$ FWER $= P(\bigcup_{j=1}^{m}$ rejecting H_0^j falsely$) \leq \sum_{j=1}^{m} P($rejecting H_0^j falsely$)$

$\Rightarrow$ FWER$(\alpha) \leq \sum_{j=1}^{m} \alpha$, with Bonferri correction follows

$\Rightarrow$ FWER$(\alpha) \leq \sum_{j=1}^{m} \frac{\alpha}{m} = m \cdot \frac{\alpha}{m} = \alpha$

$\Rightarrow$ FWER$(\alpha) \leq \alpha$

Bonferri correction controls FWER(α) at $\leq \alpha$! for a family of hypothesis. The p-values need not be independent because of clear case of inequality – we do not require to know if events / hypothesis are independent. (which is probably helpful in variant analysis because of LD)

Oct 15th 2021

CS 194-172 – Computational Genomics – HW 2

Stefan Bielmeier

Problem 1h)
With perfectly linearly separable data, why might gradient ascent still not converge to stable values of β0 and β1?
=> one reason could be that the maximum of the log likelihood function does not exist. That can be the case with perfectly linearly separable data, where the data points can be completely separated by a linear function into two same-size categories,
where X*β1 < (threshold of separation) corresponds to Y = 0, and X*β1 > (threshold of separation) correspond to Y = 1.
In that case, if we initialize β0– the intercept – with 0, it would converge quickly to an appropriate value: the threshold. β1 however could be made artificially large, yielding the same, linear separation at that certain threshold X-value, and it still would not converge to a global maximum of the log likelihood function. It could be increased infinitely, with which it is not converging.

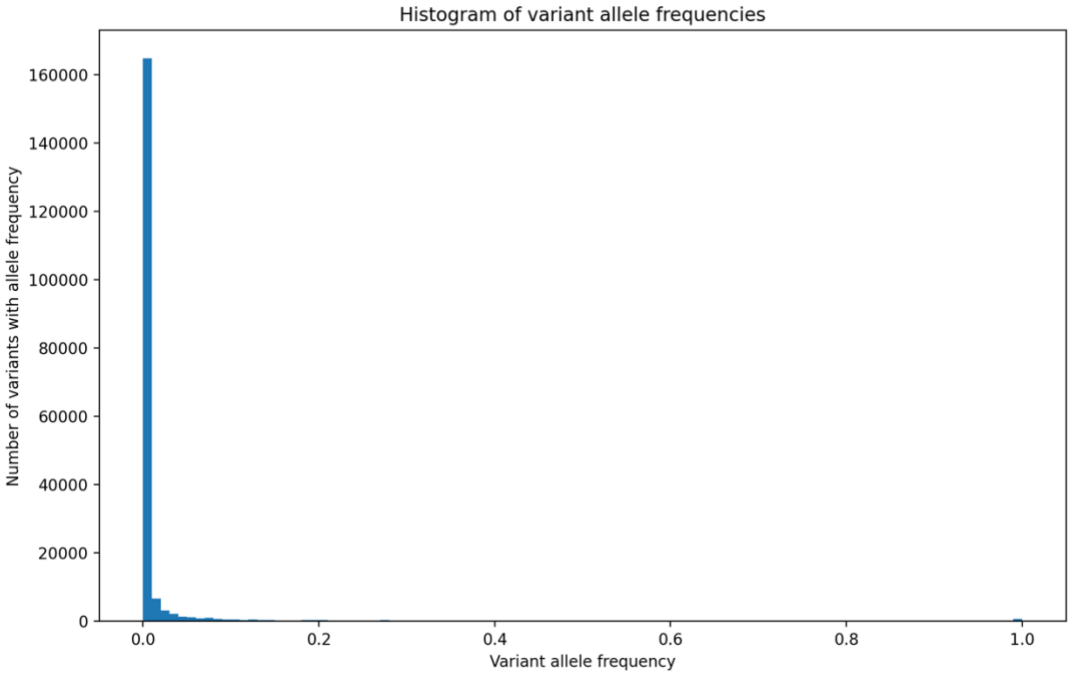
Problem 1i)
This is hard to say, multiple options are mentioned in literature: Bayesian inference (estimate β1 via a probability prior), or exact logistic regression.

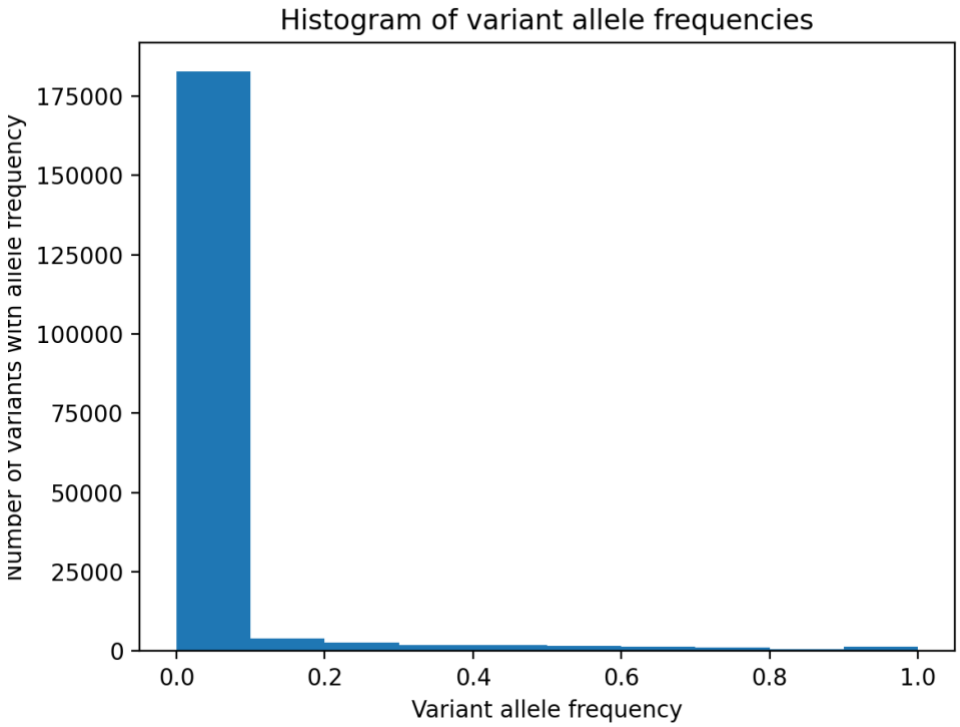
Intuitively I would say add one noise data point to the perfectly linearly separable data so that it's not perfectly linearly separable anymore – basically some value X < X-threshold with Y = 1, and calculate a β1 to see if we can reject it or not. It works because we can no longer separate our data by a linear function with an intercept, but need to actually run the regression.
This may work for large datasets where noise is tolerable.

Problem 3a)
There are 190896 SNPs listed in the file, out of a total of 198102 variants.

Problem 3b)
Provided histogram in 2 different x-axis scales…

Stefan Bielmeier HW2





The number of variants with allele frequencies < 1% is 164.785, which is most of them.

Problem 3c)

There are 2504 samples (individuals).
The average individual contains approx. 11663.579872204473 many distinct variants.
To check if this is correct, sum of all alternative alleles divided by the number of samples should be bigger than the average distinct variant frequency (above): 16042.289137380192
True!

Problem 3d)

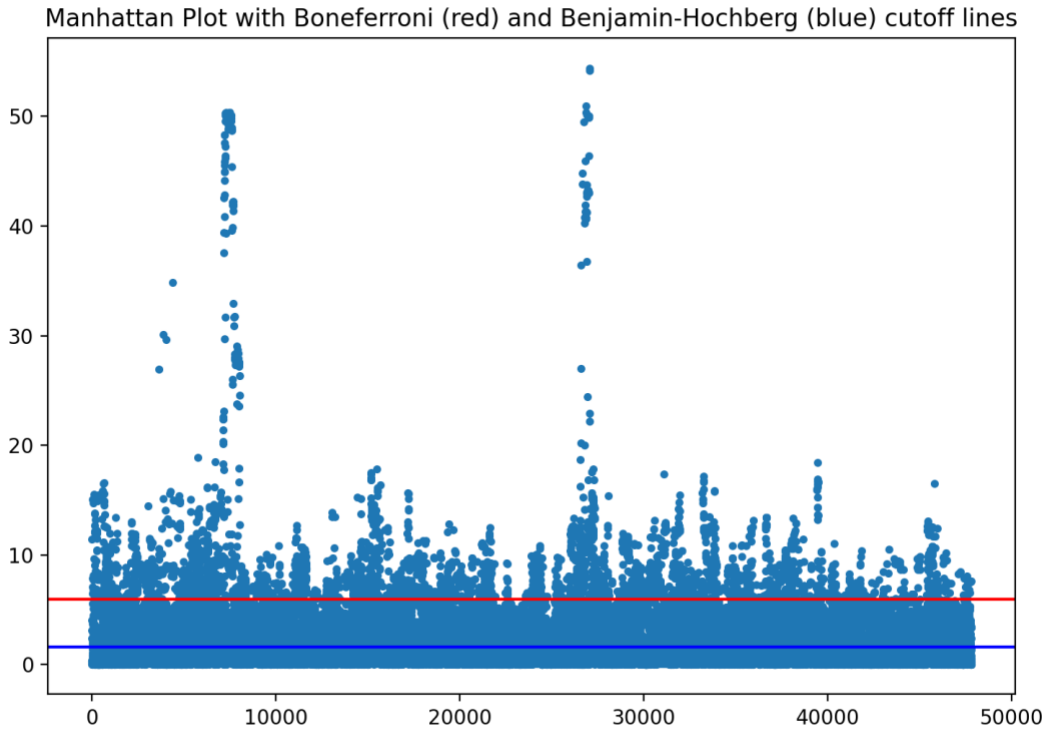
A dbSNP is a known SNP / genetic variation in the free public archive for known genetic variations for different species by National Center for Biotechnology Information (NCBI).
It has an rs number, e.g. rs328 which is without clinical significance (C => (A, G))
Going through all SNPs (rows) in the file, not a single one has an rs number in their ID column.
That means that there are 0 dbSNPs in the VCF file.

Problem 3e)

A Phred quality score is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing.
With a Phred score of 45, what is the accuracy of the variant call?

1) what is the error probability? $P = 10^{(-45/10)} = 0.0000316227766$.
2) The accuracy is thus 1 – P, which is 99.99683772

Problem 4b)



Problem 4c)

Out of 47808 SNPs present in the dataset for chromosome 22,

4013 SNPs are reported as significant after the Bonferroni correction

and **24390** SNPs are reported as significant after the Benjamin-Hochberg correction.

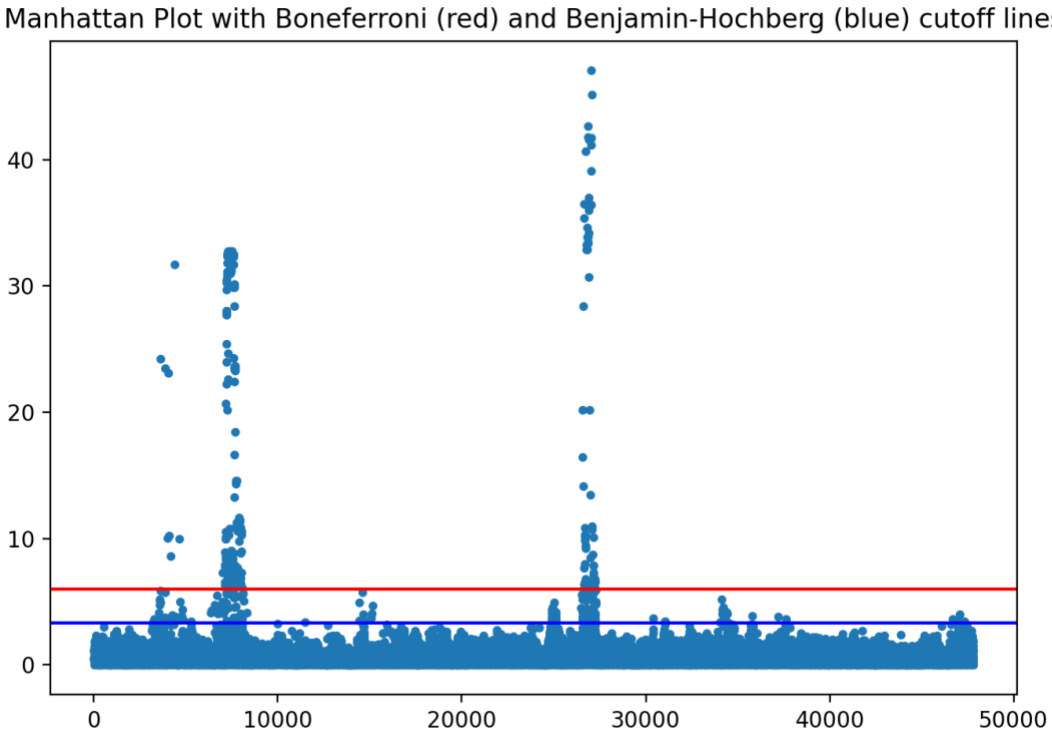
Problem 4d)

One reason why I see so many significant SNPs even after controlling for FWER and FDR at alpha = 0.05 is population stratification. Due to linkage disequilibrium, many non-causal SNPs in the same chromosome are inherited together with the causal SNP. That is especially the case when the different subpopulations are observed. That would mean that the proband who has the causal mutation (SNP) would show allele mutations for the other, non-causal SNPs, as well, simply because they are part of a certain subpopulation. When our model encounters such a non-causal SNP, it will still come to the conclusion that that non-causal SNP is significantly associated with the disease (phenotype == 1), even though the non-causal SNP is only significantly associated with the causal SNP through ancestry.

As I study only chromosome 22, it makes sense that my analysis reports high significance levels for many non-causal SNPs, because these non-causal SNPs show similar significant association to the phenotype (according to our model) but are in fact only associated with the 5 functional SNPs due to inheritance. Correcting for FWER and FDR doesn't make a difference, as the significance

levels for the non-causal, LD SNPs are just as high as the causal-SNPs because of very similar alternative allele counts for SNPs that are inherited together.

Problem 4e)



After including the Top 3 principal Components as covariates in the logistic regression analysis of 47808 SNPs present in the dataset for chromosome 22,

253 SNPs are now reported as significant after the Bonferroni correction,
and **481** SNPs are now reported as significant after the Benjamin-Hochberg correction.

Problem 4f)

Including the top (3) principal components for each individual in the logistic regression models reduces the number of significant SNPs because they reduce the multi-collinearity (mutual dependence) between our "features" for each individual. For example, the presence of alternative alleles in non-causal SNPs and causal SNPs for a certain individual caused by ancestry and joint inheritance in a subgroup of a population (population stratification).

PCA can do that by representing all 2504 SNPs for each individual in only 3 top principal components. The 3 principle components represent variance of alternative alleles that occur in one

individual. For example, when we use the principle components as covariates in one logistic regression model to estimate the effect of one single SNP on the phenotype, the principle components account for most of the variance (per definition) for each sample/individual in the model that is caused by population stratification. If the principle components capture most of the variance caused by alternative alleles, the SNP under consideration must itself show greater variance over all samples so that the model determines the its (**$Beta_1$)** to be significantly associated with the phenotype.
Therefore, less SNPs reach **Beta1**s with p-values low enough to pass the threshold of, e.g. 0.05 / number of observations after a Bonferroni correction, leading to less rejected H_0 Hypothesis ($beta_1$). This makes it easier to detect causal SNPs.

Problem 4e)
The model may not perform as well on individuals of African ancestry because they show generally different alternative allele frequencies across their genome, compared to individuals of European ancestry: population stratification. The model may interpret the presence of knowingly non-phenotype causing alternative alleles in individuals of African ancestry as phenotype-causing alternative alleles, simply because it doesn't recognize these alternative alleles to be commonly present, as the model was trained on genomes of people of European ancestry, for whom the absence of such alternative alleles is common.