1.

a) Let $\beta_0 + \beta_1 x = \alpha$

$$(1): \frac{e^{\alpha} \cdot \frac{1}{e^{\alpha}}}{(1+e^{\alpha})\frac{1}{e^{\alpha}}}$$

$$\Rightarrow \frac{1}{\frac{1}{e^{\alpha}} + 1} = \frac{1}{1+e^{-\alpha}} = \frac{1}{1+e^{(-\beta_0 - \beta_1 x)}} \quad [2]$$

b) Quotient rule:

$$\frac{dy}{dx} = \frac{v\frac{du}{dx} - u\frac{dv}{dx}}{v^2}$$

$$g'(z) = [0 - 1(-e^{-x})] \cdot [g(z)]^2$$

$$= (e^{-x})(g(z))^2$$

$$= \frac{e^{-x}}{1+e^{-x}} g(z)$$

$$= [1 - g(z)]\, g(z)$$

c)

$$f(x; \beta_0, \beta_1) = g(\beta_0 + \beta_1 x)$$

d) Using the PMF of a Bernoulli Distribution:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} P(x_i | z_i; \beta_0, \beta_1)$$

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

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

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

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

e) Because logarithms are monotonically increasing functions, so maximizing $\log(x)$ maximizes x.

4) Let $\alpha_i = \beta_0 + \beta_1 x_i$

$$\frac{d}{d\alpha} \log(g(\alpha)) = \frac{1}{g(\alpha)} g'(\alpha) \alpha'$$

$$= \frac{1}{g(\alpha)} g(\alpha)(1-g(\alpha)) \alpha'$$

$$= \alpha'(1-g(\alpha))$$

$$\frac{d}{d\alpha} \log(1-g(\alpha)) = \frac{1}{1-g(\alpha)} -g'(\alpha)\alpha'$$

$$= \frac{-1}{1-g(\alpha)} g(\alpha)(1-g(\alpha))\alpha' = -g(\alpha)\alpha'$$

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

$$= \sum_{i=1}^{n} y_i - y_i g(\alpha) - g(\alpha) + y_i g(\alpha)$$

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

$$\frac{\partial \log L}{\partial \beta_1} = \sum_{i=1}^{n} y_i x_i (1 - g(\alpha)) + (1-y_i) x_i g(\alpha)$$

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

g) $\beta_0 \leftarrow \beta_0 + \eta \frac{\partial \log(L)}{\partial \beta_0}$

$\beta_1 \leftarrow \beta_1 + \eta \frac{\partial \log L}{\partial \beta_1}$

} where η is the step size

$\frac{\partial \log(L)}{\partial \beta_0}$, $\frac{\partial \log(L)}{\partial \beta_1}$ from part (f)

h) If the separation is "too good", the logit could explode to $-\infty$, i.e. $g(\beta_0 + \beta_1 x_0) \to 1$, $\log(1 - g(x)) \to -\infty$ and $g(x) \to 0$, $\log(g(x)) \to -\infty$ respectively

i) This can be fixed by adding or subtracting a small value ε. For example, $\log(1 - g(x) - \varepsilon)$ and $\log(g(x) + \varepsilon)$.

2.

a) $P\left(\bigcup_{i=1}^{m_0} p_i \leq \frac{\alpha}{m}\right) \leq \sum_{i=1}^{m_0} P\left(p_i \leq \frac{\alpha}{m}\right) = m_0 \frac{\alpha}{m} \leq \alpha$

This is from Bonferroni's inequality.

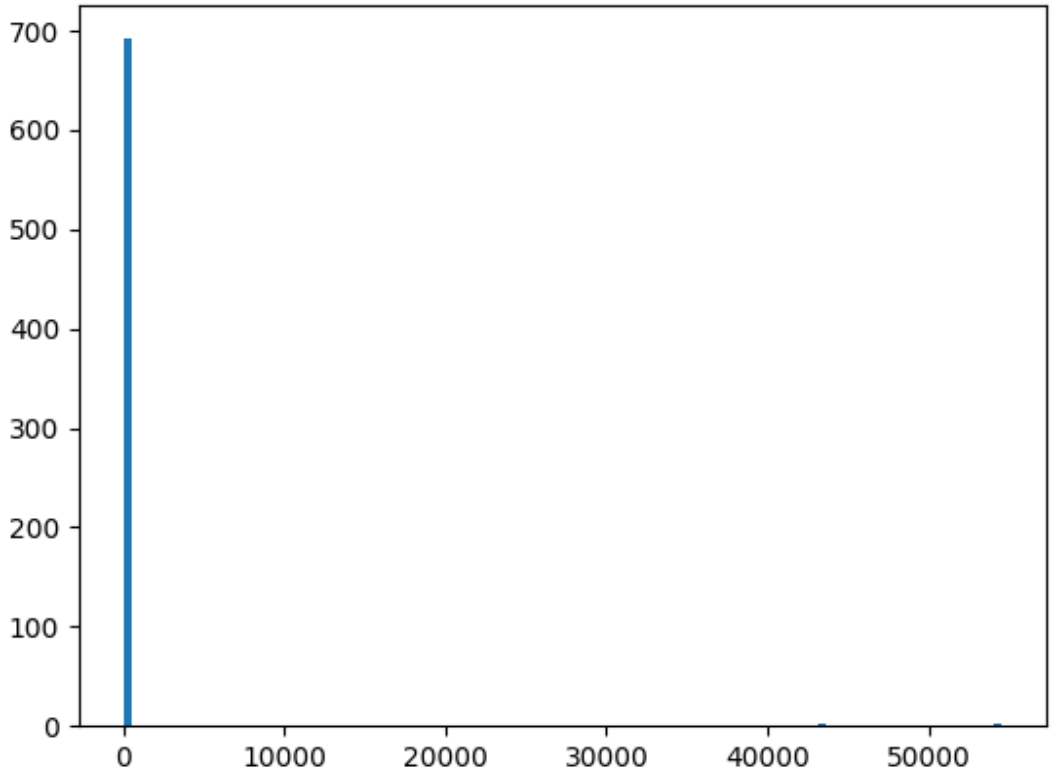
It doesn't require independence of p-values.

b) See code.

c) See code.

3a)
TOTAL SNPS: 191079

3b)
Less than 1%: 692



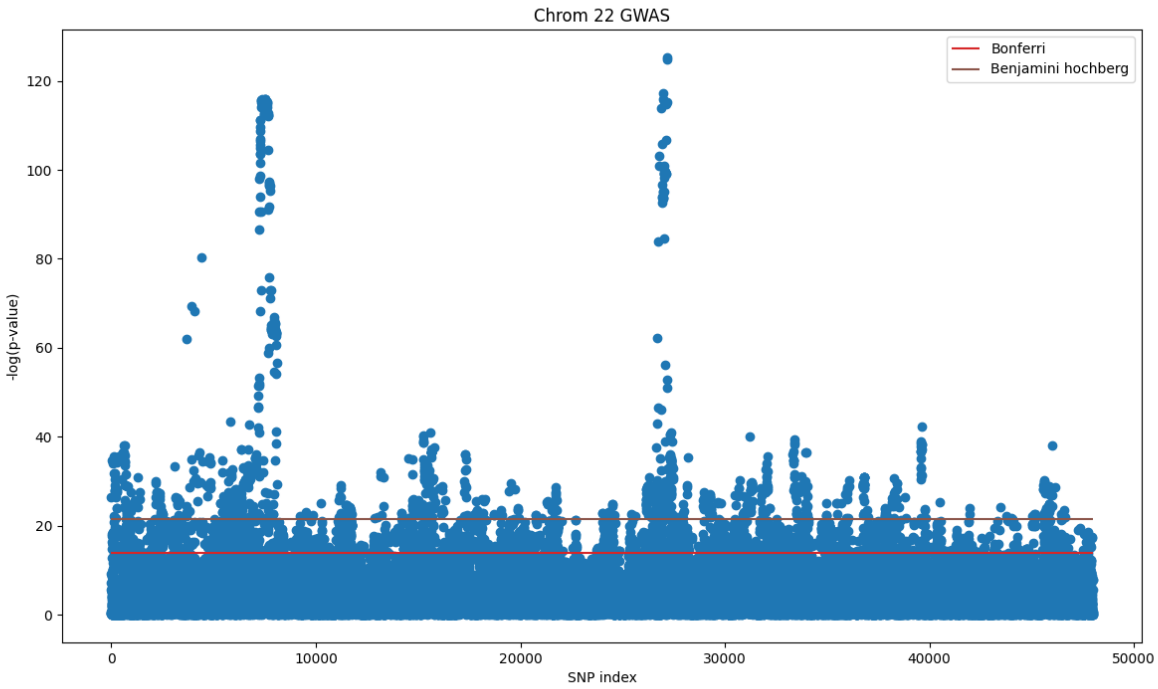
3c)
NUM SAMPLES: 2504
AVERAGE VARIANTS PER INDIVIDUAL: 16042.29

3d)
A dbSNP ID is an ID that links a particular variant to a known database of variants.
NUM dbSNP: 0

3e)
A Phred quality score is a measure of the quality of the bases observed from automatic DNA sequencing.
A score of 45 means that the probability of this being an incorrect base call is 1 in 31,622.78.

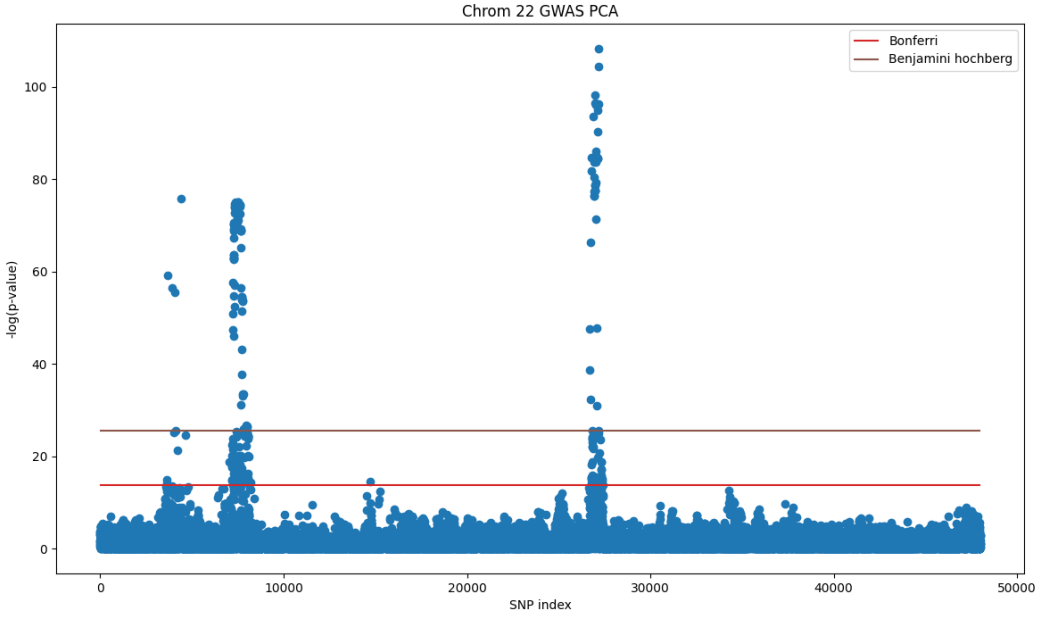
4a)
See code

4b)



4c)
REJECTED BONFERRI: 4007
REJECTED BENJAMINI-HOCHBERG: 304

4d)
This is because the corrections only help reduce the false positive rate (with certain tradeoffs) and doesn't necessarily eliminate all false positives. In the end, it is is only a heuristic that depends on an alpha that is also a heuristic. It is also difficult to separate raw input signals from the noise



4e)

REJECTED PCA BONFERRI: 258
REJECTED PCA BENJAMINI-HOCHBERG 4

4f)
This is because PCA removes some multicollinearity by picking out vectors that contribute to the dosage matrix most strongly but are also orthogonal. Hence it gives logistic regression stronger signals to learn a regressive model on.

4g)
This is because individuals of African descent could have very different SNPs or other variations that lead to certain phenotypes compared to individuals of European descent. Doing a PCA would also likely lead to very different principal components as well.