Quantitative Genomics HW 4

Noelle Wheeler

2023-03-15

Problem 1

Consider the (slightly idealized) genetics behind behind one of Mendel's famous experiments with pea plants (look Mendel up on wikipedia if you are new to genetics), where for two alleles A_1 and A_2 the phenotype of a pea is guaranteed to be 'yellow' if the genotype is either A_1A_1 or A_1A_2 and guaranteed to be 'green' if the genotype is A_2A_2. If you were to code a random variable for this system Y(yellow)$= 1$ and Y(green)$= 0$ and used a (genetic) linear regression to model this case, what would be the true values of the parameters β_μ, β_a, β_d, and σ_ϵ^2?

For this system we can use the parameters $\beta_\mu = \frac{3}{4}$, $\beta_a = -\frac{1}{2}$, and $\beta_d = \frac{1}{4}$. We would get σ_ϵ^2 from our actual data points but since this is a perfect system where no points fall outside of the line we can set $\sigma_\epsilon^2 = 0$.

Problem 2

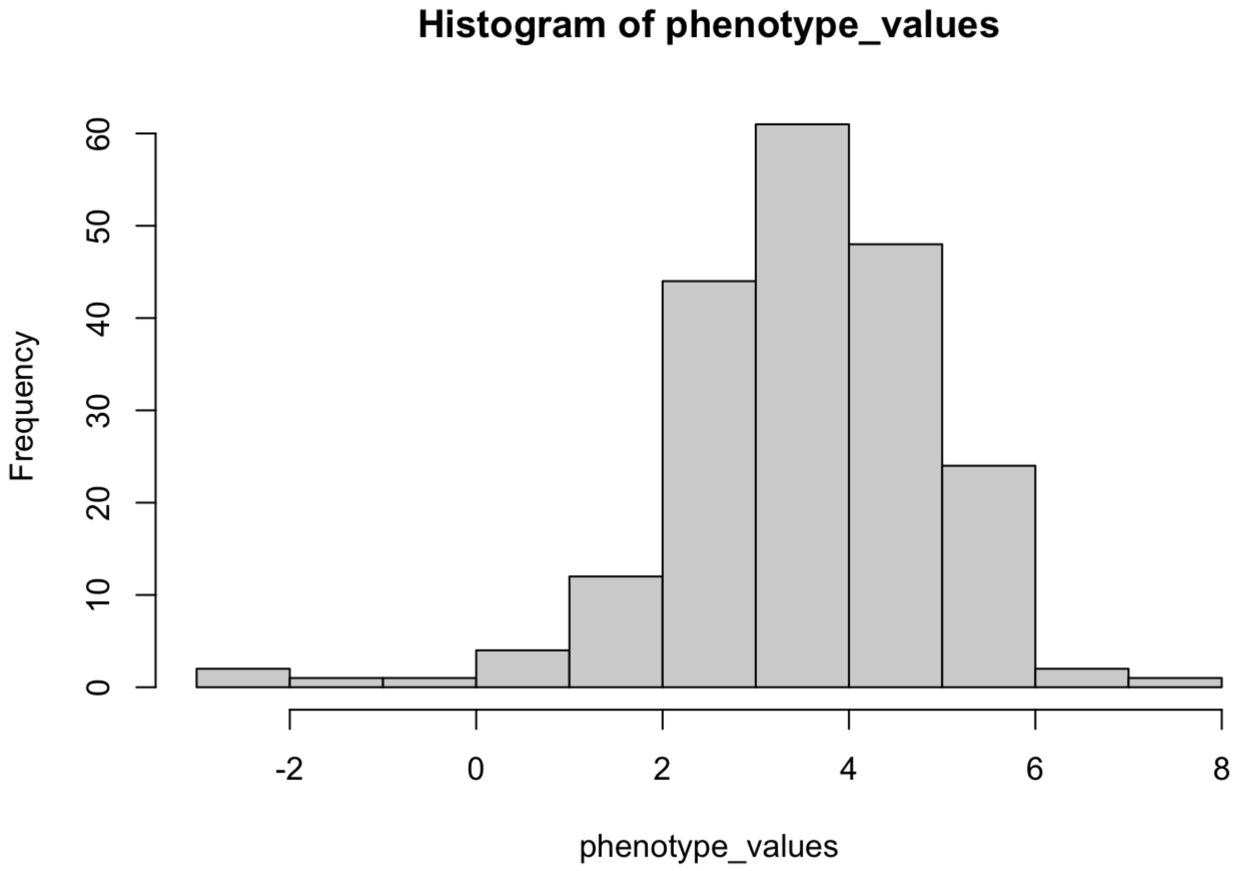
a. Write code that inputs the phenotype, plus an (additional) line of code that calculates the number of samples n and report the number n.

```
phenotypes <- read.table("/Users/noelawheeler/Downloads/QG23 - hw4_phenotypes.txt")
print(paste0("There are ", nrow(phenotypes), " samples."))
```

```
## [1] "There are 200 samples."
```

b. histogram of phenotypes.

```
phenotype_values <- unlist(phenotypes, use.names = FALSE)
hist(phenotype_values)
```



c. Write code that inputs the genotype data plus a line of code that outputs the number of genotypes N and sample size n and report these number.

```
genotypes <- read.table("/Users/noelawheeler/Downloads/QG23 - hw4_genotypes.txt")
print(paste0("There are ", ncol(genotypes), " genotypes."))
```

```
## [1] "There are 1000 genotypes."
```

```
print(paste0("The sample size is  ", nrow(genotypes)))
```

```
## [1] "The sample size is  400"
```

d. Write code that converts your genotype data input in part [c] into two new matrices, the first a matrix where each genotype is converted to the appropriate Xa value and the second where each genotype is converted to the appropriate Xd value

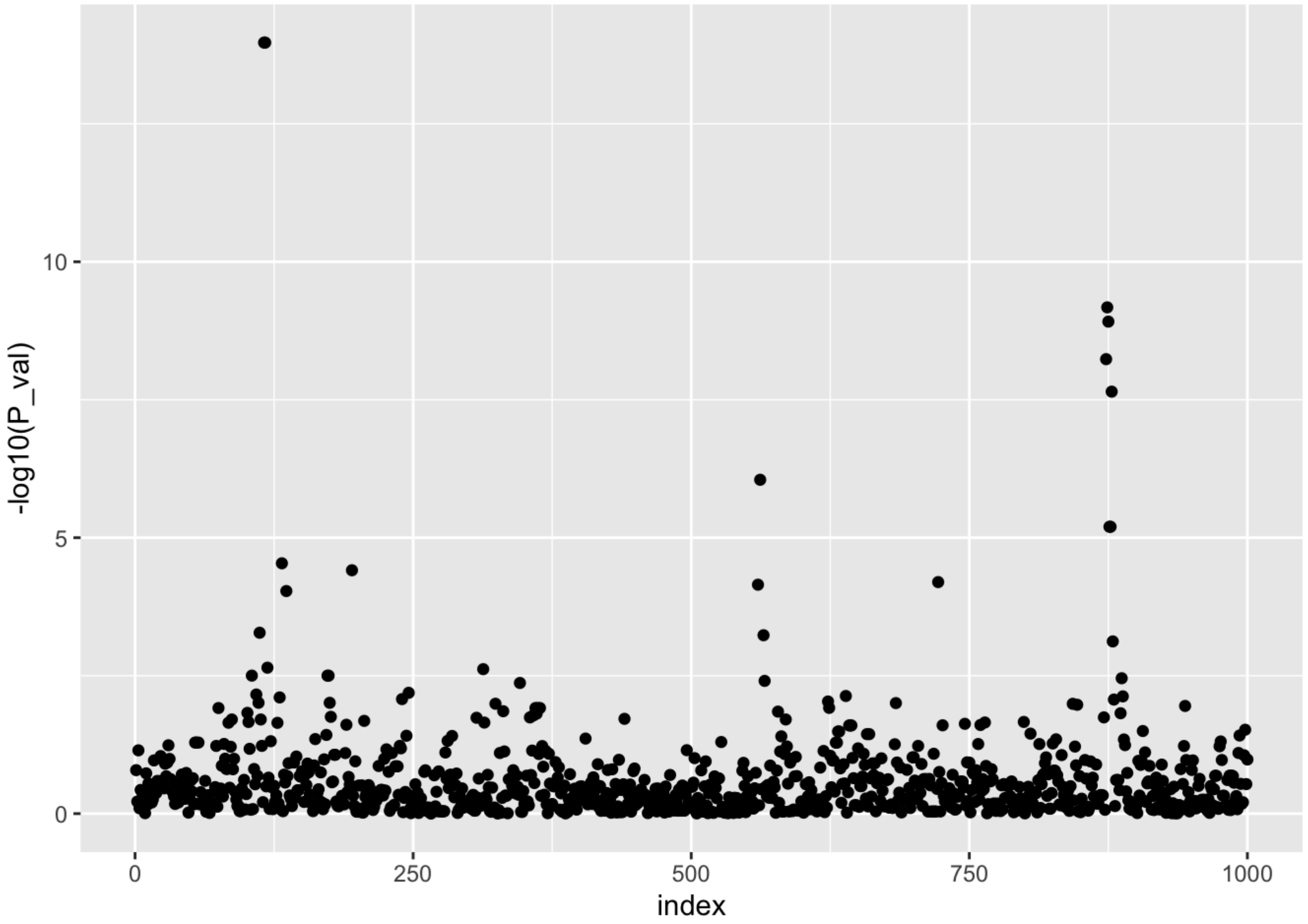
```
# separate matrices into first allele and second allele
genotypes_odd <- genotypes[as.logical(seq(nrow(genotypes)) %% 2),]
genotypes_even <- genotypes[!as.logical(seq(nrow(genotypes)) %% 2),]

# create x_d
x_d <- genotypes_odd[,] == genotypes_even[,]
x_d <- ifelse(x_d, -1, 1)

# create x_a
# find most common allele for each column
major_allele <- c()
for (i in seq(1:1000)){
  freq_allele <- tail(names(sort(table(genotypes_even[[i]]))), 1)
  major_allele <- append(major_allele, freq_allele)
}

# convert into matrix
mat <- matrix(replicate(200,major_allele),nrow=1000)
mat <- t(mat)

# switch 1 in x_d to 0 in x_a
x_a <- x_d
x_a[x_a == 1] <- 0

# if both alleles equal the minor allele, change -1 to 1
major_allele_even <-  genotypes_even[,] == mat[,]
major_allele_odd <-  genotypes_odd[,] == mat[,]

indices <- which(major_allele_even == FALSE & major_allele_odd == FALSE, arr.ind = TRUE)
x_a[indices] <- 1
```

e. Write code to calculate $MLE(\hat{\beta}) = [\hat{\beta}_{\mu}, \hat{\beta}_a, \hat{\beta}_d]$ for each genotype in the dataset, an F-statistic for each genotype, and a p-value for each genotype using the R function pf(F-statistic, df1, df2, lower.tail = FALSE). PLEASE NOTE (!!): that you may NOT use an existing R function for ANY of these calculations other than the calculation of the p-value

```r
library(MASS)
# create matrix for B-hat values
MLE_beta_all <- matrix(0, nrow = 1000, ncol = 3)
y <- as.matrix(phenotypes)
# create vector for F-statistic/ P-values
F_stat <- c()
P_val <- c()
# set F-statistic parameters
n_samples <- length(x_a[,1])
df_M <- 3 - 1
df_E <- n_samples - 3
# loop through each genotype
for (i in seq(1:1000)){
  xa_input <- x_a[,i]
  xd_input <- x_d[,i]
  # bind columns with Bu = 1
  x <- cbind(rep(1,length(xa_input)), xa_input, xd_input)
  # calculate b-hat for each genotype
  MLE_beta <- ginv(t(x) %*% x) %*% t(x) %*% y
  MLE_beta_all[i,] <- MLE_beta
  # calulate y-hat
  y_hat <- x %*% MLE_beta
  # calculate F-statistic
  SSM <- sum((y_hat - mean(y))^2)
  SSE <- sum((y - y_hat)^2)
  MSM <- SSM / df_M
  MSE <- SSE / df_E
  Fstatistic <- MSM / MSE
  F_stat <- append(F_stat, Fstatistic)
  # calculate p-value
  pval <- pf(Fstatistic, df_M, df_E,lower.tail = FALSE)
  P_val <- append(P_val, pval)
}
```

f. Write code to produce a Quantile-Quantile (QQ) plot for your p-values PLEASE NOTE (!!): do NOT use an R function (=write your own code to produce the Manhattan plot) but DO provide your code AND your QQ plot.

```r
library(ggplot2)
plot_df <- data.frame(index = 1:length(P_val), pval = P_val)
ggplot(plot_df, aes(index, -log10(P_val))) + geom_point()
```



h. Do you consider the QQ plot to indicate that you have ‘good’ model fit in this case?
Yes, this appears to have a good model fit since most loci aren’t highly correlated with the phenotype. However, there are some loci and some SNPs that do seem to have a larger effect on the phenotype.

i. Write code that uses a Bonferroni correction to produce an overall study controlled Type I error of 0.05 to assess whether to reject the null hypothesis for each genotype, where your code also outputs the number of each genotype for which you rejected the null (remember: the genotypes are provided in order along the genome!). Report the numbers of all genotypes for which you rejected the null

```
p_adj <- p.adjust(P_val, method = "bonferroni")
num_reject <- sum(p_adj <= .05)
print(paste0("There are ",num_reject, " genotypes that reject the null hypothesis"))
```

```
## [1] "There are 11 genotypes that reject the null hypothesis"
```

```
location_reject <- which(p_adj <= .05)
print(location_reject)
```

```
##  [1] 116 117 132 195 562 873 874 875 876 877 878
```

j. Assuming the set of genotypes for which you rejected the null hypothesis in part [i] do indeed indicate the positions of causal genotypes in the genome, how many causal genotypes do you think these significant genotypes are indicating overall? Explain your reasoning using no more than two sentences.

I think these genotypes are indicating about 5 casual genotypes. Since many of these genotypes are close in location, it would make sense that one of those is a casual genotype and the others are in linkage disequilibrium with that casusal genotype. There seems to be five distinct genotype location clusters.

Problem 3

$$\hat{\beta}_1 = \frac{\mathrm{Cov}(y,x)}{\mathrm{Var}(x)}$$

$$MLE(\hat{\beta}) = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y}$$

For the simple regression model with parameters $\beta = [\beta_0, \beta_1]$, show that equation (2) can be re-written such that the element β_1 of the output vector has the form of equation (1).

Since there is only one single predictor, $\mathbf{x}^{\mathrm{T}}\mathbf{x}$ equals the variance of X and $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ equals the covariance between $\mathbf{x}^{\mathrm{T}}$ and Y. This means that $(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y}$ equals $\mathrm{Cov}(y,x) * \mathrm{Var}(x)^{-1}$, which can be rewritten as $\frac{\mathrm{Cov}(y,x)}{\mathrm{Var}(x)}$.