Homework 6

Pete Rigas (pbr43@cornell.edu), BTRY 6830

November 14, 2020

---

1 Problem 1

1.1 Part A

There are several reasons as to why it is advantageous for the phenotype in a GWAS to conform to a normal distribution. One of the easier reasons to immediately spot is that, from general facts about genetic inference, we know that for a model focusing on one locus, the linear regression is of the form

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \ ,$$

with $\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon)$. From this equality, we know that it would certainly be imperative to have the phenotype conform to a normal distribution so that as we are testing null hypotheses, in which we are looking to determine whether β_a and β_d vanish or not, the normal distribution assumption gives us the ability to accurately test positions of the genome, including genotype associations, that could be of importance to the causal polymorphism that we want to learn more about. For a GWAS, it is immediately clear that having the assumption that many independent, random effects is approximately normal helps us, as shown in HW 5, allows us to calculate several quantities of interest quickly, given that the distribution is normal.

More generally, it is important for the phenotype to have a normal distribution so that we can we can carry out the linear regression model to estimate positions of interest in the genome with Y, as seen in the expression for Y above. So the assumption of a normal distribution shows that this type of distribution is important for several reasons, from actually being able to carry out computations more easily to also applying different Central Limit Theorems to determine how the data is distributed, relative to the mean.

1.2 Part B

In the case that we reject a null hypothesis for a genotype in a GWAS, we know from Lecture 15 slides that for every marker, the phenotype

$$X = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \ ,$$

for ϵ normally distributed, as already stated in **A**, implies that the hypothesis that we would be interested in testing in our GWAS are of the form

$$H_0 : \text{Cov}(Y, X_a) = 0 \cap \text{Cov}(Y, X_d) = 0 \ ,$$

or,

$$H_A = \text{Cov}(Y, X_a) \neq 0 \cup \text{Cov}(Y, X_d) \neq 0 \ .$$

With this recapitulation, we observe that our rejection of the null hypothesis would imply that our marker is definitely not correlated with the causal polymorphism that we are trying to determine. As mentioned in the most recent lecture before spring break, if we were even in the case that there was no causal polymorphism, we could have more than one population in our sample; the population itself could have more than one subset of measured genotypes; or, finally, the mean value of the phenotype could also differ. In this case, rejecting the null hypothesis could potentially, experimentally, give a false positive because the MAF that we would want to measure between populations would give us a low p value. Altogether, I have provided more than 2 reasons as to why rejecting the null hypothesis for a GWAS analysis could not indicate the position of a causal polymorphism; each of the reasons that have been provided show that variability in either set of the observables that we use for a GWAS, either the genotypes, phenotypes or even the mean or variance of the sampling population, can impact the position at which we think we should test to find a causal polymorphism.

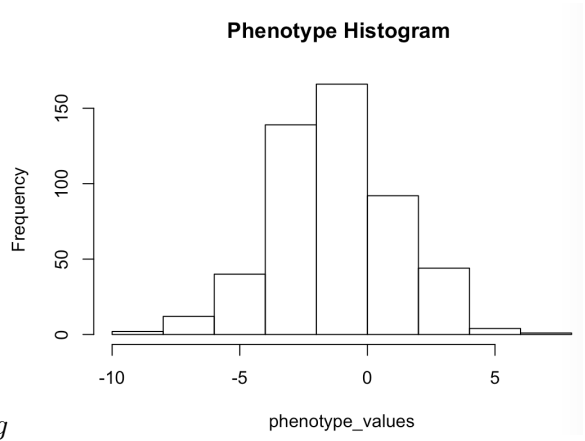
As a sidenote, this false positive could also occur from the fact that we performed genotyping errors, or also that we encountered a case of disequilibrium in which there is no linkage. We can correct for these problems with covariates.

2 Problem 2

2.1 Part A

```
> setwd("/Users/peterigas/Downloads/")
> phenotypes <- read.table("QG19 - hw6_phenotypes (1).txt", stringsAsFactors = FALSE , col.names =1
> length(count.fields("QG19 - hw6_genotypes (2).txt"))
[1] 1000
```

2.2 Part B



(40).*png*

```
phenotypes <- as.matrix(phenotypes)
hist(phenotypes)
```

2.3 Part C

```
> genotypes <- read.delim2("QG19 - hw6_genotypes (2).txt", col.names=1 , stringsAsFactors = F)
> genotypes <- read.delim(file.choose())
> length(count.fields("QG19 - hw6_phenotypes (1).txt"))
[1] 500
```

2.4 Part D

```
sample_number <- nrow(genotypes)/2
genotype_columns <-ncol(genotypes)

library(MASS)

xa_converter <- function(geno){
   geno_count <- table(t(geno))
     minor_allele <- names(geno_count[geno_count == min(geno_count)])
    xa_code <- ifelse(geno == minor_allele, 1,0)
     xa_result <- colSums(xa_code) - 1
     return(xa_result)
 }

xa_matrix <- matrix(NA,
                    nrow = nrow(genotypes)/2,
                     ncol = ncol(genotypes))


for (i in 1:(ncol(genotypes)/2)){
    xa_matrix[i,] <- xa_converter(genotypes[c(2*i -1, 2*i),])
 }
```

2.5 Part E

2.5.1 R Code for MLE

```
    MLE_calculator <- function(phenotype , xa , xd){
    n_samples <- length(xa)

    X_mx <- cbind(1,xa_input,xd)

    MLE_beta <- ginv(t(X_mx) %*% X_mx) %*% t(X_mx) %*% phenotype

    return(MLE_beta)
}
```

2.5.2 R Code for p values

```
    pvalues <- function(phenotype, xa, xd_){
    n_samples <- length(xa)

    X_mx <- cbind(1,xa,xd)
```

```
    MLE_beta <- ginv(t(X_mx) %*% X_mx) %*% t(X_mx) %*% phenotype
    y_hat <- X_mx %*% MLE_beta

    SSM <- sum((y_hat - mean(phenotype))^2)
    SSE <- sum((phenotype - y_hat)^2)

    df_M <- 2
    df_E <- n_samples - 3

    MSM <- SSM / df_M
    MSE <- SSE / df_E

    Fstatistic <- MSM / MSE

    pval <- pf(Fstatistic, df_M, df_E, lower.tail = FALSE)

    return(pval)
}

p_vector <- c()

for (k in 1:ncol(xa_matrix)){
    pval_vec <- c(p_vec, pvalues(phenotypes, xa_matrix[,k], xd_matrix[,k]))
}
```

2.6 Part F

```
    plot(-log10(pval), main = "Manhattan plot")
```

2.7 Part G

```
    number_genotypes <- nrow(pvalues)
    log_phenotypes <- -log(pvalues[,i])

    expected_pvalues <- sort(runif(length(pval)), decreasing = FALSE)
    observed_pvalues <- sort(pval, decreasing = FALSE)
    plot(-log10(expected_pvalues), -log10(observed_pvalues), main = "QQplot")
```

2.8 Part H

Overall, the QQ Plot is not that great. To justify our observations, we observe that the QQ plot, overall, comes off the line $y = x$ too early. So from our discussion in lecture, we furthermore know that this would not make the QQ plot necessarily a very good one. However, on the other hand if we are looking just at the beginning of the QQ plot, where it stays particularly close to the curve, we would then conclude that it would be a good QQ plot, only in that region. So it really depends on which part of the graph we are looking at, but as a whole, this QQ plot is not that great. In a real world example, we would introduce more covariates so that our QQ plot does not have values that depart from the line that quickly. Also, we could introduce genotype filtering to make our QQ plot more reflective of realistic data that we would expect to gather from a GWAS.

2.9 Part I

Even with this Bonferroni correction, we could still encounter a biological false positive.

```
which(pvals < (0.05 / 1194))
```

2.10 Part J

From the Manhattan plot and previous arguments, we observe that the significant 'hits' are locations in the plot in which, after taking the $-\log$ transformation, are accentuated as points that appear very high on the plot. In all reality, every genotype that we measure, and can observe to some degree, depends on several cases of linkage disequilibrium. In the case of humans, a safe answer is that such a peak includes one causal genotype, from our human genome. In other cases, if this peak represents a significant part of the genome, then each peak could represent more than one causal genotype. This would be important for someone to consider because when one is doing a GWAS, we often encounter that we can try to isolate causal genotypes that are significant in the genome, but that we do not exactly know how many genotypes in the genome that each one of the peaks on the Manhattan Plot would have. To be completely safe, we observe that each tower in our Manhattan plot would mean that there is one causal genotype corresponsing to these positions in the genome. Therefore, the answer really depends on the genome that we are talking about, whether human or animal, as well as how sure we can be of the number of genotypes corresponding to each individual peak in the Manhattan plot.

3 Problem 3

In order to show that the given equality holds, first, we observe that

$$\mathcal{L}(\beta, \sigma_\epsilon^2, \mathbf{x}|\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma_\epsilon^2)^n}\exp\Big(\sum(y_i - \hat{y}_i)^2\Big) = \frac{\prod_{i=1}^n \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma_\epsilon^2)^n}\exp(y_i - \hat{y}_i)^2}{\exp(\frac{-n(\bar{x}-\mu)^2}{2\sigma^2})} \ ,$$

from which we know that the Likelihood is maximized over all μ such that the exponent in the exponential term above from the Normal distribution is minimized. Next, we trivially observe that taking the logarithm of the likelihood function above, in terms of β and σ_ϵ^2, gives,

$$\mathcal{L} = -\frac{n}{2}\log(2\pi\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^n (y_i - \hat{y}_i)^2 \ ,$$

from the alternate expression of the likelihood,

$$\mathcal{L} = \frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{n}{2}}}\exp\Big(\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{2\sigma^2}\Big) \ ,$$

From these rearrangements, we will
from the fact that from this particular Likelihood, we are obtaining an estimate $\hat{\mu} = \bar{x}$, which gives the ratio

$$\Delta = \frac{\mathcal{L}_{H_0}}{\mathcal{L}_{H_A}} \ ,$$

which in the fraction above represents the likelihoods from the null and alternative hypotheses, respectively. With this approach, we will now analyze both the left and side sides of the equality

$$\left(\left(\frac{2}{n-3}\right)F_{2,n-3}+1\right)^{-\frac{n}{2}}=\Delta\ ,$$

starting with the ratio of likelihoods that we first gave, in which we directly substitute the 2 exponential terms,

$$\frac{\frac{1}{(2\pi)^{\frac{n}{2}}(\sigma_\epsilon)^n}\exp(-\frac{1}{2\sigma_\epsilon^2}\sum(y_i-\hat{y})^2)}{\frac{1}{(2\pi)^{\frac{n}{2}}(\sigma_\epsilon)^n}\exp(-\frac{1}{2\sigma_\epsilon}\sum(y_i-\hat{y})^2)}\ ,$$

and because the 2π terms cancel, with the σ_ϵ^2 term remaining in each fraction because we are looking at the null hypothesis in the numerator, versus the alternative hypothesis in the denominator, looking up the MLE for the variance, which is equivalent to the MLE of the variance $\hat{\sigma}^2$ given either the alternative or null hypotheses, gives a similar expression for the MLE of the variance is useful for us. Also, from previous expression of the given MLE in **3**, we also know that

$$\frac{\partial l}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\ ,$$

which vanishes when the following condition is satisfied,

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\ .$$

Before proceeding, observe that there was some abuse of notation, because I intentionally did not cancel the σ_ϵ^2 factors in the numerator and denominator of the expression above. It is important to observe that we keep these terms and do not cancel them out because under the separate null and alternative hypotheses, the variances that appear to be the same, from the mere notation, actually are different.

From these arguments, we recover the maximun likeilihood estimators of the $sigma_\epsilon^2$ under the null and alternative hypotheses, precisely from the exponential terms in the ratio for Δ above. Now that it is clear that this ratio is actually a ratio of Maximum Likelihood estimators, we can analyze the equality to make sure that it holds by determining whether any similar expression can be expressed in terms of the F statistic with degrees of freedom 2 and $n-3$.

To do this, we will introduce rearrangementsm, from the definitions, on the other side of the equality that was previously stated. Specifically, we will make the association that the F statistic, by definition is a ratio of SSEs. As a result, we will rearrange terms to then show that the equality, after rearrangements on the left and right sides together, relies on expressions between SSEs and MLEs.

From the F-Statistic side of the equality,

$$\left(\frac{2}{n-3}F_{2,n-3}+1)\right)^{\frac{-n}{2}} = \left(\frac{2}{n-3}\frac{\frac{\text{SSE}(\hat{\theta}_0)-\text{SSE}(\hat{\theta}_1)}{2}}{\frac{\text{SSE}(\hat{\theta}_1)}{n-3}}\right)^{\frac{-n}{2}}$$
$$= \left(\frac{\text{SSE}(\hat{\theta}_0)-\text{SSE}(\hat{\theta}_1)}{\text{SSE}(\hat{\theta}_1} + \frac{\text{SSE}(\hat{\theta}_1}{\text{SSE}(\hat{\theta}_1)}\right)^{\frac{-n}{2}}$$
$$= \left(\frac{\text{SSE}(\hat{\theta}_0)}{\text{SSE}(\hat{\theta}_1)}\right)^{\frac{-n}{2}}.$$

From these rearrangements from the other side of the equality, we now observe that these terms are, by definition, equal to the terms that we recovered from the previous side of the equality in earlier steps. To see that this is in fact the case, we know that $y_{\hat{H}_0} = \hat{\beta}_u$, and also that $y_{\hat{H}_A} = \hat{\beta}_u + X_a\hat{\beta}_a + X_d\hat{\beta}_d$. So from basic properties of the linear regression and making use of these equalities, we are then able to clearly see that the terms in the numerator and denominator that we have given are clearly equivalent to terms that we have recovered from the MLE ratio, in previous steps. Immediately,

$$\left(\frac{\frac{\sum_i(y_i-\hat{\beta}_u)^2}{n}}{\frac{\sum_i(y_i-(\hat{\beta}_a+X_a\hat{\beta}_a+X_d\hat{\beta}_d)^2}{n}}\right)^{\frac{-n}{2}},$$

clearly demonstrating that these terms are equivalent to what we got for the MLEs previously, because

$$= \left(\frac{\text{MLE}(\hat{\sigma_\epsilon^2}}{\text{MLE}(\hat{\sigma_\epsilon^2}}\right)^{\frac{-n}{2}},$$

and also that,

$$= \left(\frac{\hat{\sigma_\epsilon^2}}{\hat{\sigma_\epsilon^2}}\right)^{\frac{-n}{2}}.$$

To conclude, we have shown that the equality holds by manipulating both sides of the equality. With the given expression for the likelihood, we then rearranged terms to discover that the terms actually depend on rearrangements of SSEs and MLEs.

4 References

- Stack Exchange and Overflow for ideas about the simulations, and for the proof in **3**.