HW4_SOLUTION

gustavoc@msu.edu

12/06/20201

Single marker regression (aka indpendent screening, aka GWAS)

```
load('~/Dropbox/STAT_COMP/2020/Xy.RData')

## Single-marker regression test
 SMR=matrix(nrow=ncol(X),ncol=4,NA)
 colnames(SMR)=c('Estimate','SE','z-stat','p-value')

 for(i in 1:ncol(X)){
  SMR[i,]=ls.print(lsfit(y=y,x=X[,i]),print.it=F)[[2]][[1]][2,] # can also use summary(lm(y~X[,i]))$coe
 }

 head(SMR)
```

```
##         Estimate         SE     z-stat   p-value
## [1,] 0.066375838 0.11438827 0.58026785 0.5617471
## [2,] 0.071140176 0.12721343 0.55921907 0.5760247
## [3,] 0.019019788 0.05703887 0.33345310 0.7387993
## [4,] 0.007430585 0.05438721 0.13662376 0.8913309
## [5,] 0.002325077 0.17841395 0.01303193 0.9896026
## [6,] 0.021287279 0.05337528 0.39882278 0.6900323
```

Adjusting p-values and determining significance

```
BONF=p.adjust(SMR[,4],method='bonferroni')
BONF_SIG=BONF<0.05

HOLM=p.adjust(SMR[,4],method='holm')
HOLM_SIG=HOLM<0.05

FDR=p.adjust(SMR[,4],method='fdr')
FDR_SIG=FDR<0.05
```

Q1 Which SNPs were significant for each criteria

Bonferroni

```
 colnames(X)[BONF_SIG]
```

```
## [1] "SNP_ 588"   "SNP_ 609"   "SNP_ 611"   "SNP_ 10892" "SNP_ 10898"
## [6] "SNP_ 10905" "SNP_ 10906" "SNP_ 10911" "SNP_ 10922"
```

Holm

```
colnames(X)[HOLM_SIG]
```

```
## [1] "SNP_ 588"   "SNP_ 609"   "SNP_ 611"   "SNP_ 10892" "SNP_ 10898"
## [6] "SNP_ 10905" "SNP_ 10906" "SNP_ 10911" "SNP_ 10922"
```

FDR

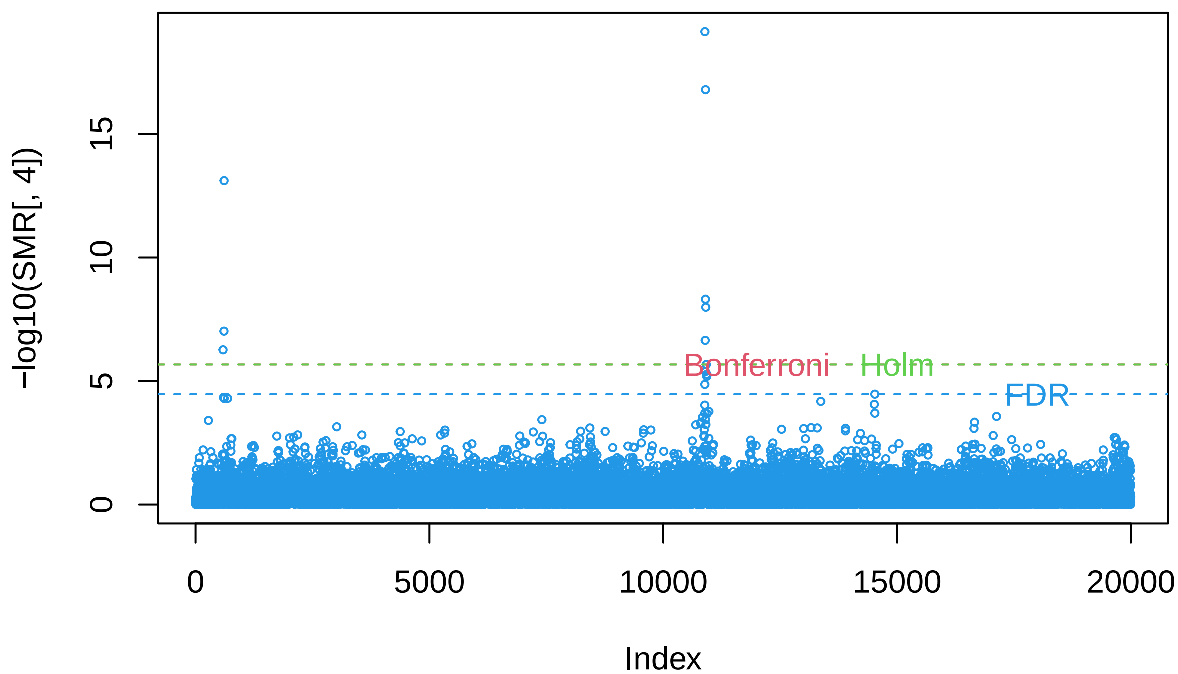
```
colnames(X)[FDR_SIG]
```

```
##  [1] "SNP_ 588"   "SNP_ 609"   "SNP_ 611"   "SNP_ 10891" "SNP_ 10892"
##  [6] "SNP_ 10898" "SNP_ 10902" "SNP_ 10905" "SNP_ 10906" "SNP_ 10911"
## [11] "SNP_ 10922" "SNP_ 10932" "SNP_ 10933" "SNP_ 14527"
```

- No difference between Holm and Bonferroni (both yielded 9 discoveries)
- FDR gave 14 discoveries.

Manhattan plot

```
## Manhattan plot
plot(-log10(SMR[,4]),cex=.5,col=4 )
abline(h= -log10(max(SMR[BONF_SIG,4])),lty=2,col=2)
abline(h= -log10(max(SMR[HOLM_SIG,4])),lty=2,col=3)
abline(h= -log10(max(SMR[FDR_SIG,4])),lty=2,col=4)
text(x=12000,y=-log10(max(SMR[BONF_SIG,4])), label='Bonferroni',col=2)
text(x=15000,y=-log10(max(SMR[HOLM_SIG,4])), label='Holm',col=3)
text(x=18000,y=-log10(max(SMR[FDR_SIG,4])), label='FDR',col=4)
```

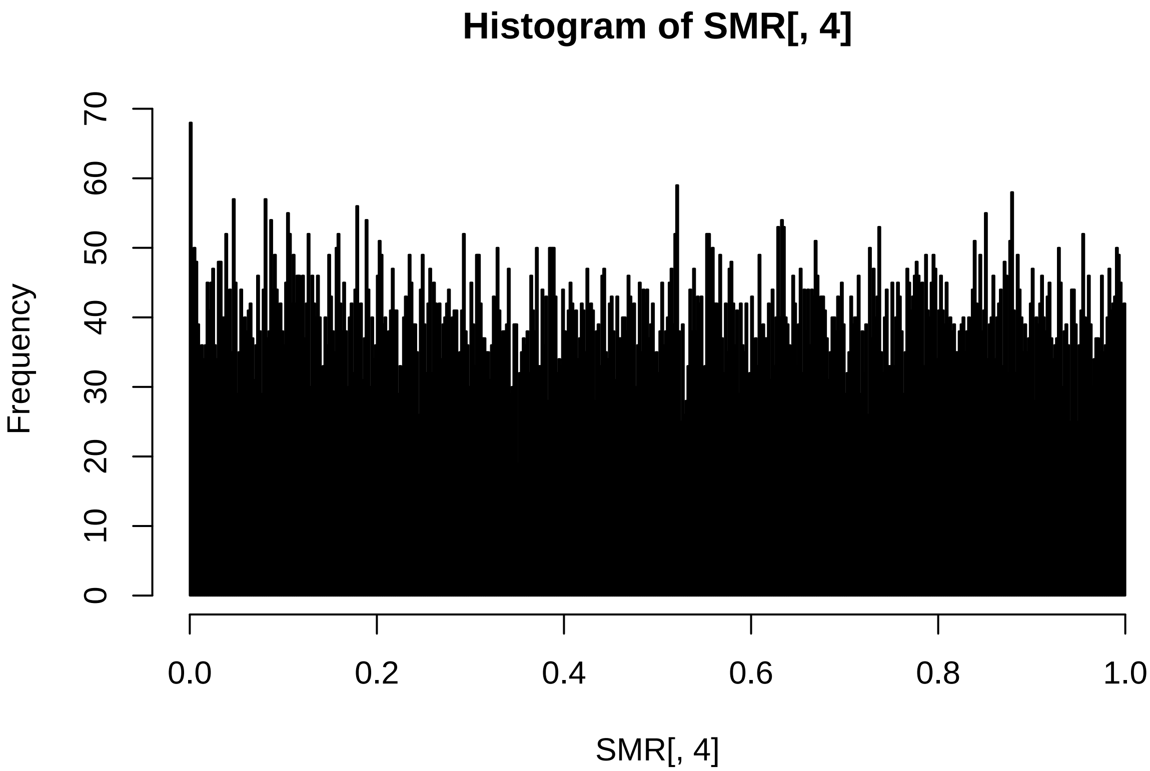


Histogram of p-values

Recall that under H0 p-values follow uniform distributions. If your data includes a fraction of alternative hypothesis you should see enrichment of low p-values.

Note that the fraction of Ha's is often small, so you need to use many bins in the histogram, in this case I specified 1,000,

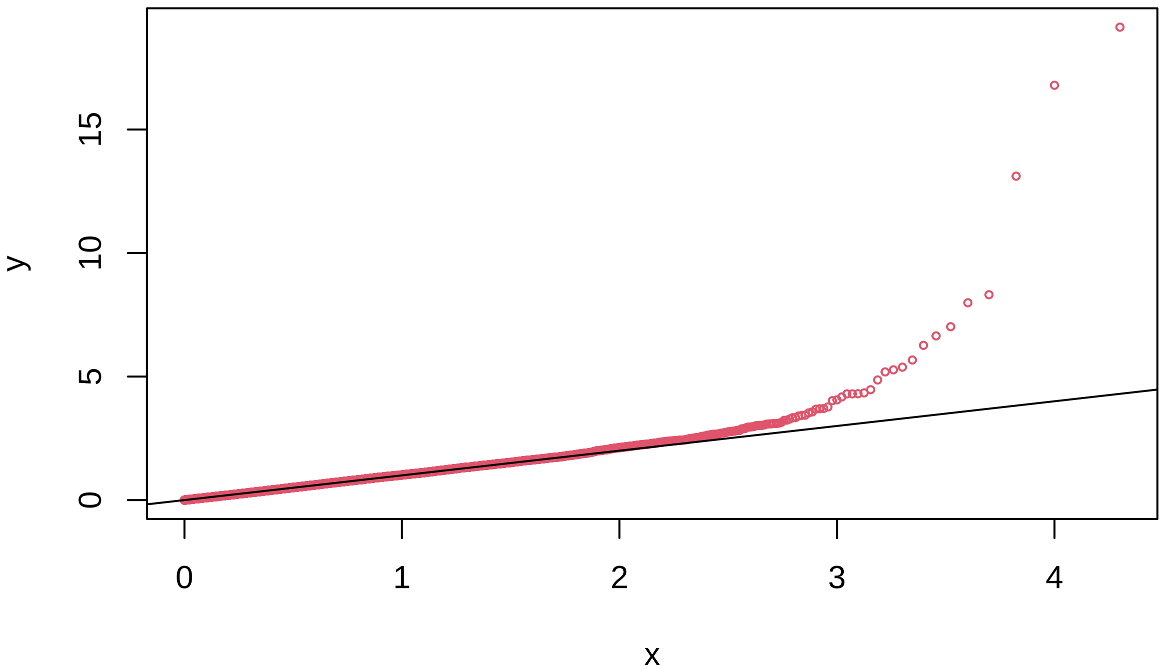
```r
hist(SMR[,4],breaks = 500,col=8)
```



qqplot

Another, better way, to detect enrichment of low p-values is to compare the empirical quantiles of the p-values with theoretical quantiles for the uniform distribution.

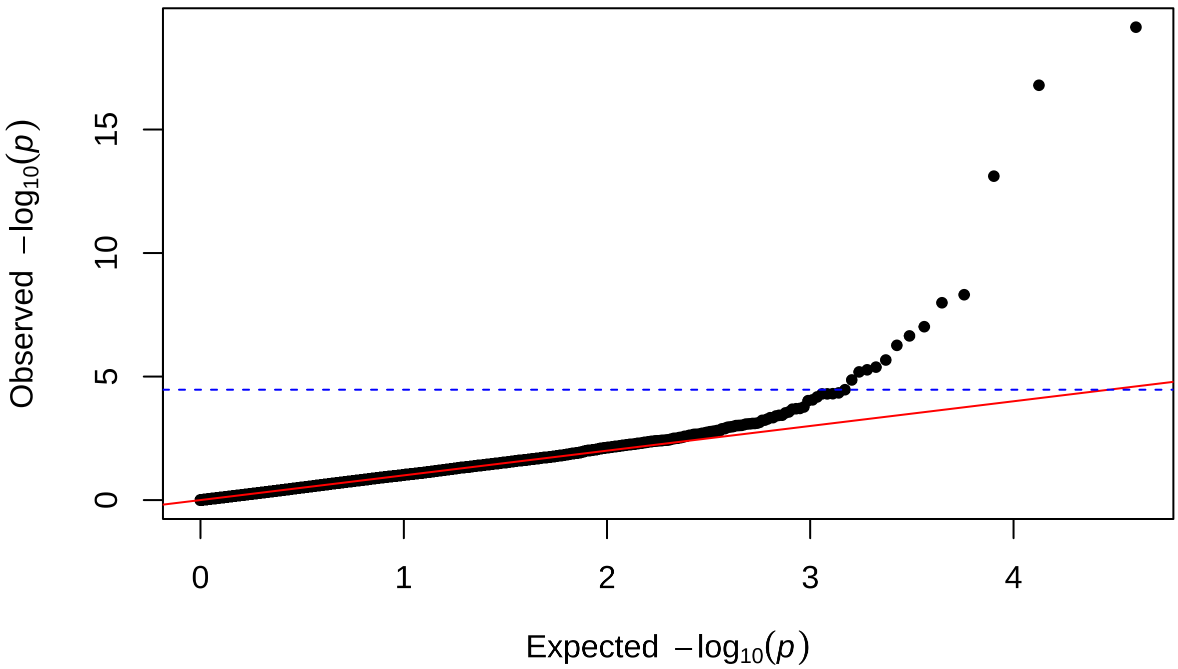
```r
pVal=sort(SMR[,4],decreasing=FALSE)
expectedUnderH0=(1:length(pVal))/length(pVal)

y= -log10(pVal)
x= -log10(expectedUnderH0)

plot(y~x,cex=.5,col=2);abline(a=0,b=1)
```




```
## Using the qqman package
#install.packages(pkg='qqman',repos='https://cran.r-project.org/')
library(qqman)
```

```
##
## For example usage please run: vignette('qqman')
##
## Citation appreciated but not required:
## Turner, (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. Jour
##
```

```
qq(SMR[,4])
abline(h=-log10(max(SMR[FDR_SIG,4])),lty=2,col='blue')
```



Departures from the null?

There seems to be an initial departure from the 45-dgree line at a -log10(pvalue)>2.5 (i.e., p-value < 0.003), but at the begning the departure is linear. Then, starting at the implied FDR-cutoff (`-log10(max(SMR[FDR_SIG,4]))`=4.468, the dashed blue line, we start to see an exponential growth of the quantiles. Personally I think starting at the FDR-implied threshold, there is clear evidence of departure from the 45 degree line, but suggesting a smaller cutoff would also be reasonable, everything depends on what FDR are we willing to tolerate.