ProjetII

Sven Spörri, Thomas Castiglione

07/12/2020

R environment

We use the package ggplot2 for plotting.

```
library(ggplot2)
```

2. Data exploration

We load the data genotypes.txt, variants_info_openSNP.txt, phenotypes.txt and covariates.txt.

```
genotypes<-read.table(’ressources/genotypes.txt’,header=T)
phenotypes<-read.table(’ressources/phenotypes.txt’, header=T)
variants.info<-read.table(’ressources/variants_info_openSNP.txt’,header=T)
covariates<-read.table(’ressources/covariates.txt’,header=T)
```

1. linear regression

We make changes to the data format to make it easier to handle, and perform linear regression.

```
covariates.pheno<-cbind(covariates,phenotypes$HEIGHT)
rownames(covariates.pheno)<-covariates$CHALLENGE_ID
covariates.pheno<-covariates.pheno[,-1]
colnames(covariates.pheno)<-c(’SEX’,’PLATFORM’,’HEIGHT’)

lin.reg.cov<-lm(HEIGHT~., data=covariates.pheno)
summary(lin.reg.cov)
```

```
##
## Call:
## lm(formula = HEIGHT ~ ., data = covariates.pheno)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.8618  -5.0818  -0.0518   5.5237  23.4798
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            165.1018     0.4359 378.794  < 2e-16 ***
## SEXM                    14.6184     0.5577  26.210  < 2e-16 ***
## PLATFORMancestry        -0.3951     0.9121  -0.433  0.66502
## PLATFORMftdna-illumina  -7.2203     1.8862  -3.828  0.00014 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.639 on 780 degrees of freedom
## Multiple R-squared:  0.472,  Adjusted R-squared:  0.4699
## F-statistic: 232.4 on 3 and 780 DF,  p-value: < 2.2e-16
```

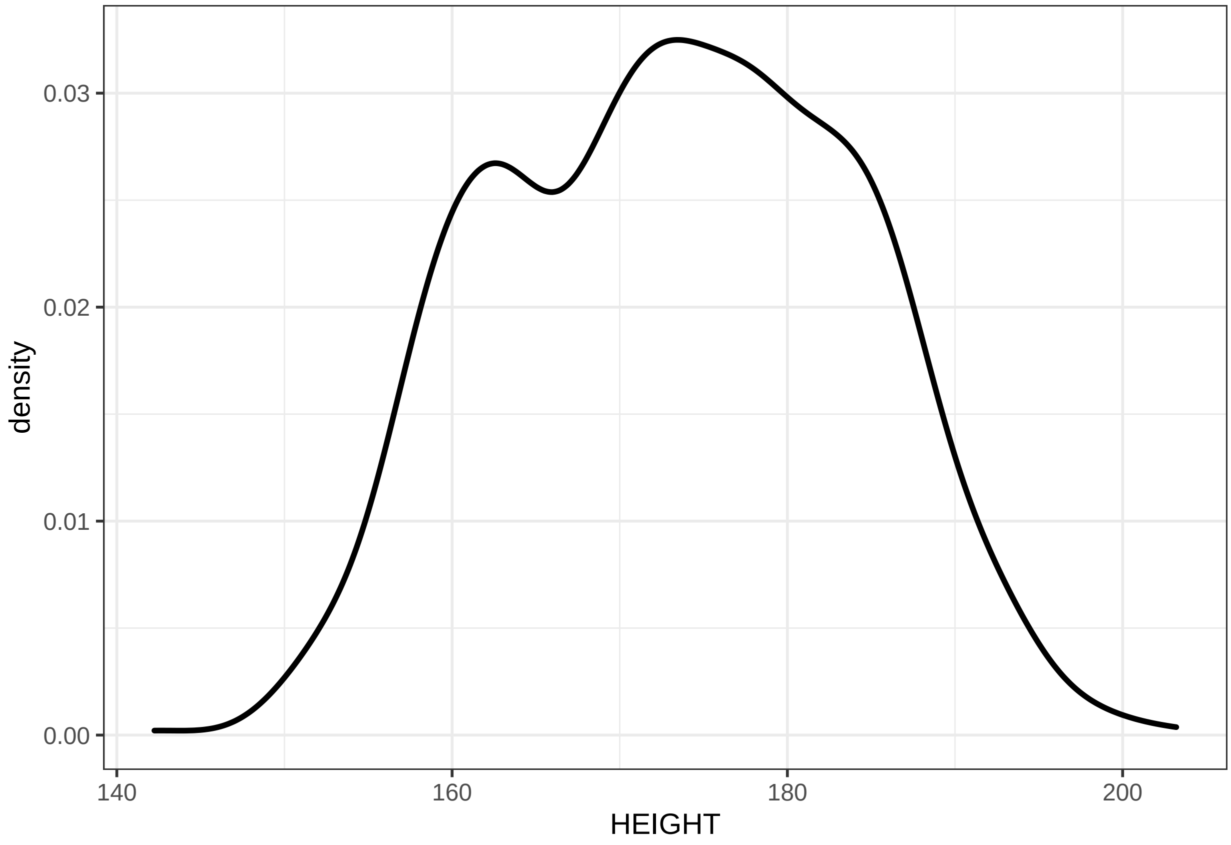
```
cat("number of sample coming from ftdna-illumina = ",
    nrow(covariates.pheno[covariates.pheno$PLATFORM=='ftdna-illumina',]))
```

```
## number of sample coming from ftdna-illumina =  17
```

The multiple R-squared value of 0.472 shows us that the covariates do impact the phenotype, but they do not fully explain its variations. Interestingly, people who sequenced their genome with the platform ftdna-illumina are significantly smaller (by 7cm in the mean). This is very probably an artifact in our data since there are only 17 people who used this platform.

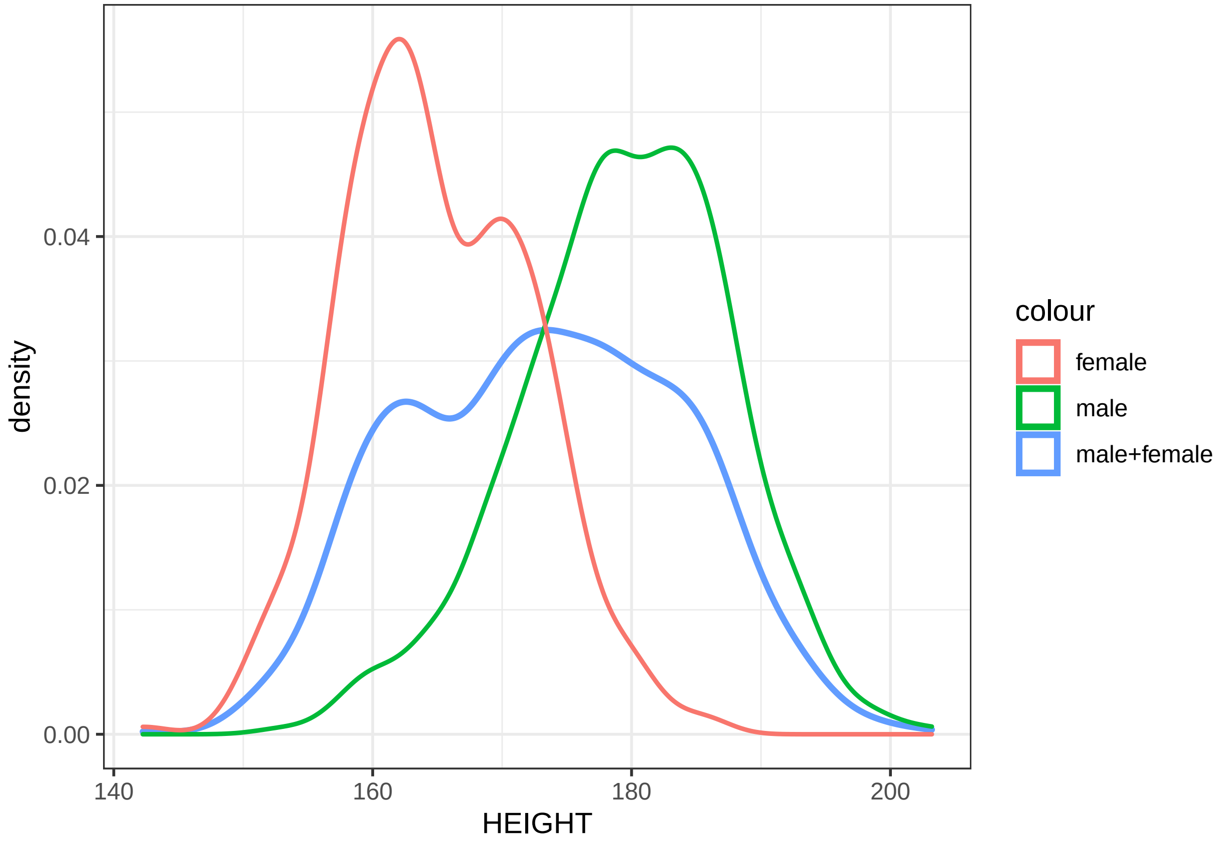
2. Plotting the phenotype distribution

```
ggplot(phenotypes)+geom_density(aes(HEIGHT),size=1)+theme_bw()
```



```
covariates.pheno.M<-covariates.pheno[apply(covariates.pheno,1,function(x) x[1]==’M’),]
covariates.pheno.F<-covariates.pheno[apply(covariates.pheno,1,function(x) x[1]==’F’),]

ggplot()+geom_density(data=covariates.pheno,aes(HEIGHT,color=’male+female’),size=1.1)+
  geom_density(data=covariates.pheno.M,aes(HEIGHT,color=’male’),size=0.8)+
  geom_density(data=covariates.pheno.F,aes(HEIGHT,color=’female’),size=0.8)+ theme_bw()
```



The mean height of men is higher than the mean height of women. The bell curve for men is a little bit skew towards the right.

Genome-Wide Association Study

1.

running GWAS without covariates.

```
gwas<-function(SNP,phenos=phenotypes$HEIGHT){
  df<-data.frame(cbind(SNP,phenos))
  colnames(df)<-c(’SNP’,’HEIGHT’)
  fit<-lm(HEIGHT~.,data=df)
  summary.coefs<-summary(fit)$coefficient
  data.frame(coef=summary.coefs[2,1],intercept=summary.coefs[1,1],pval=summary.coefs[2,4])
}
gwas.with.cov<-function(SNP,cov=covariates.pheno){
  df<-data.frame(cbind(SNP,cov))
  colnames(df)<-c(’SNP’,colnames(covariates.pheno))
  fit<-lm(HEIGHT~.,data=df)
  summary.coefs<-summary(fit)$coefficients
  data.frame(coef=summary.coefs[2,1],intercept=summary.coefs[1,1],pval=summary.coefs[2,4])
}
```

```
gwas.res<-data.frame(coef=NULL,pval=NULL)
for(i in 1:ncol(genotypes)){
  gwas.res<-rbind(gwas.res,gwas(genotypes[,i]))
}
rownames(gwas.res)<-colnames(genotypes)
gwas.res<-cbind(gwas.res, variants.info$CHR)
```

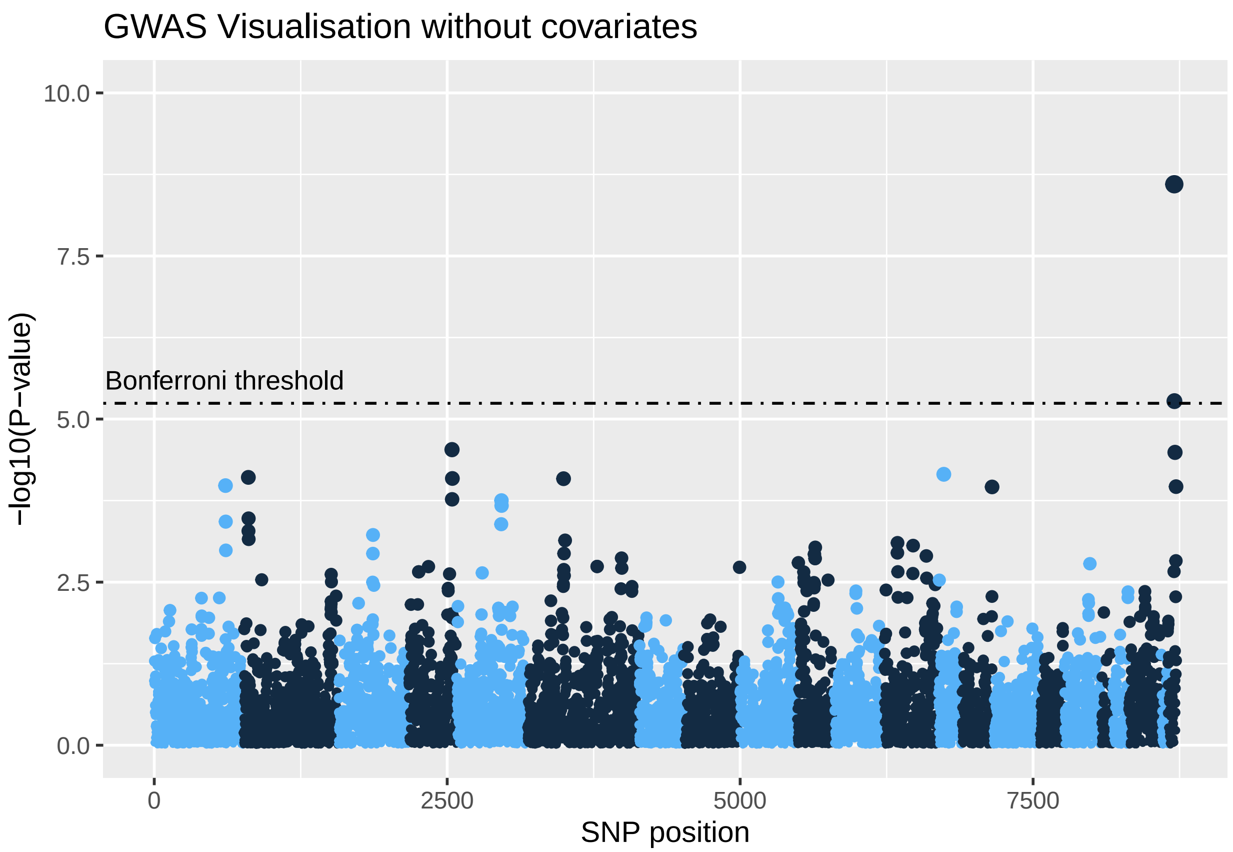
running GWAS with covariates.

```
gwas.res.with.cov<-data.frame(coef=NULL,pval=NULL)
for(i in 1:ncol(genotypes)){
  gwas.res.with.cov<-rbind(gwas.res.with.cov,gwas.with.cov(genotypes[,i]))
}
rownames(gwas.res.with.cov)<-colnames(genotypes)
gwas.res.with.cov<-cbind(gwas.res.with.cov,variants.info$CHR)
```

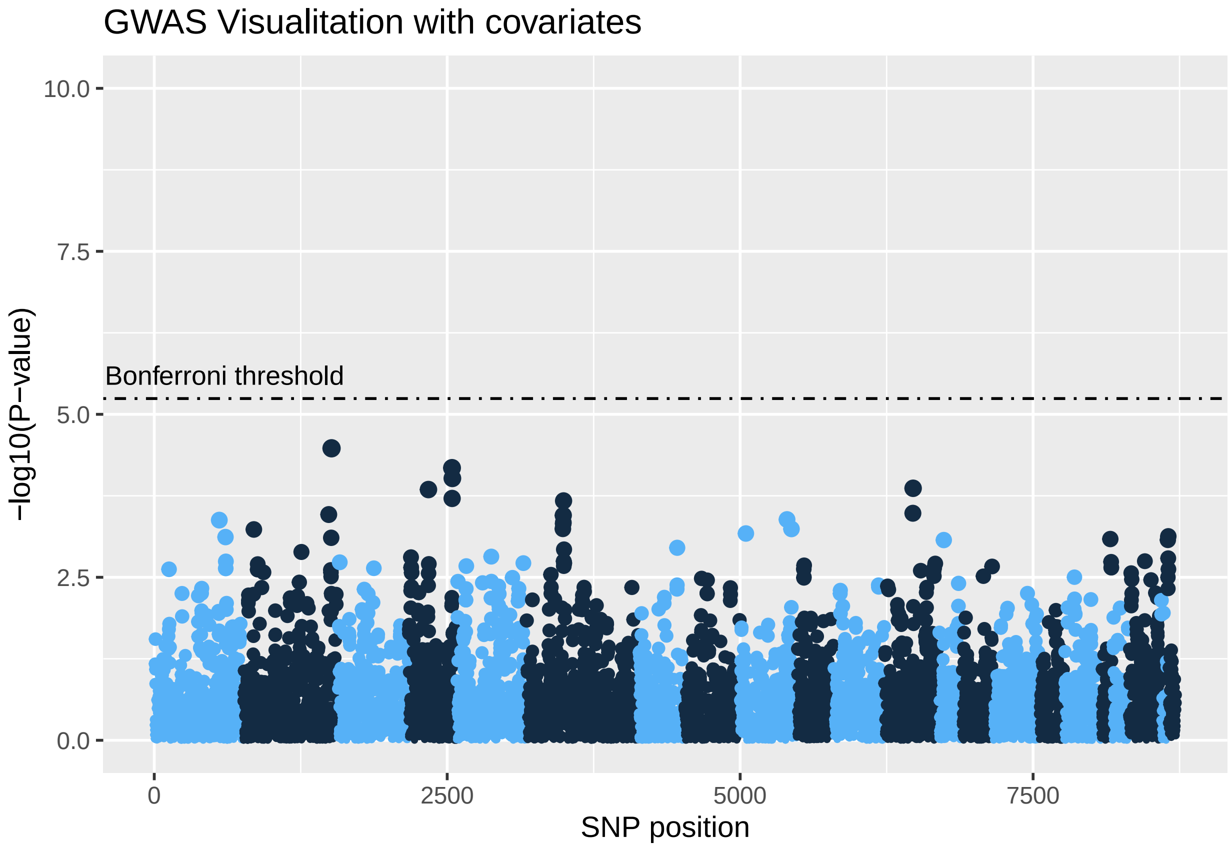
2. Manhattan plots.

```
manhattan.plot<-function(gwas.out,title){
colnames(gwas.out)<-c(’coef’,’intercept’,’pval’,’CHR’)
bonferroni_thresh<-0.05/nrow(gwas.out)
ggplot(gwas.out,aes(x=1:nrow(gwas.out),y=-log10(pval),color=CHR%%2,size=-log10(pval)))+
  geom_point(shape=20)+
  geom_hline(yintercept = -log10(bonferroni_thresh),linetype=’dotdash’)+
  scale_size_continuous(range = c(1,4))+
  scale_y_continuous(limits=c(0,10))+
  theme(legend.position = ’none’)+
  annotate(geom=’text’, x=600,y=5.6, label=’Bonferroni threshold’,size=3.5)+
  labs(x=’SNP position’,y=’-log10(P-value)’)+
  ggtitle(title)
}
```

```
manhattan.plot(gwas.res, ’GWAS Visualisation without covariates’)
```



```
manhattan.plot(gwas.res.with.cov, 'GWAS Visualitation with covariates')
```



3. comparing results

We can see from the GWAS results without covariates that one of the very last SNPs is has a high significance, but not on the GWAS with the covariates. This my come from the fact that making GWAS without covariates does not take in account the difference between samples with chromosomes XX or XY.

Also, the GWAS with covariates does not shows any SNP having significance higher than the Bonferroni threshold. But some of them get close to it, meaning they may impact the height, but without high significance.

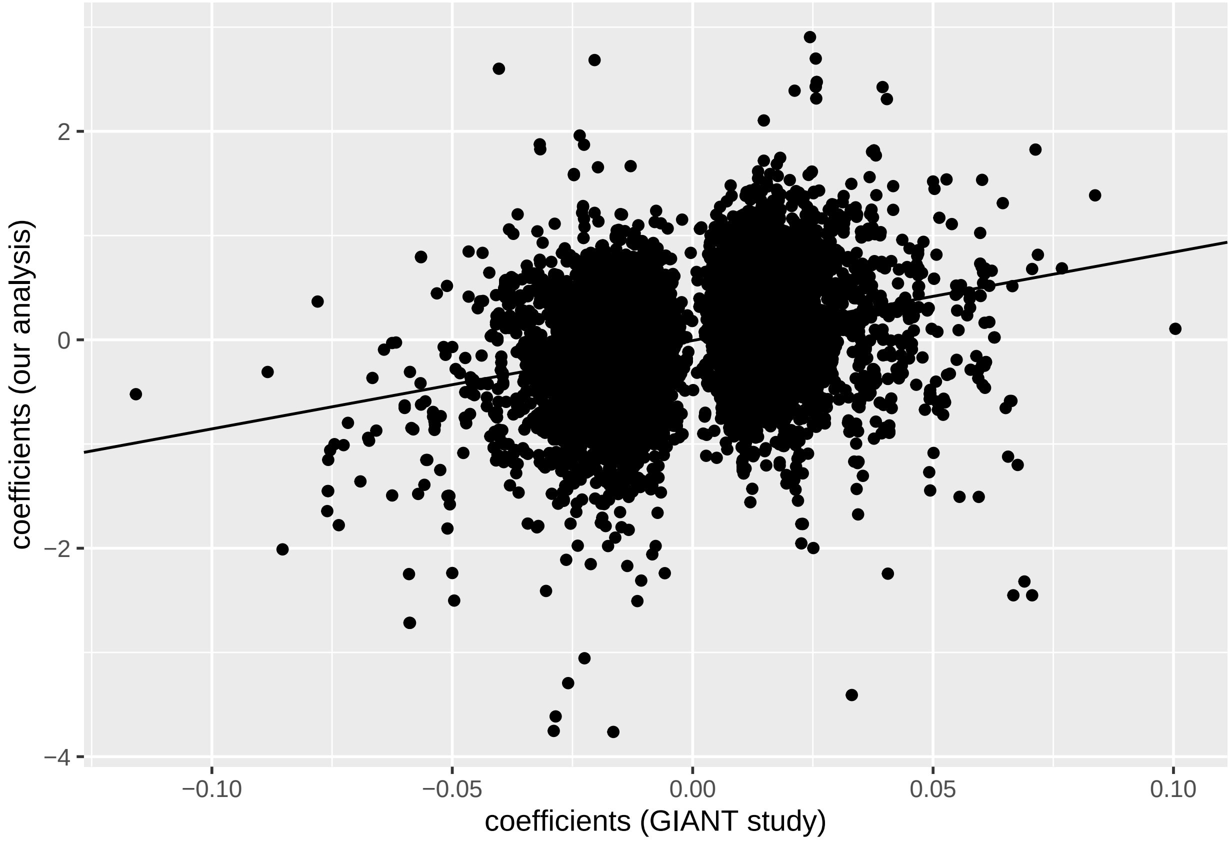
Meta-analysis with the GIANT Study

```
# read data from giant study
giant.summary <- read.table('ressources/summary_GWAS_giant.txt',header=T)
print(paste0('# of SNPs not available in GIANT dataframe: ',
             nrow(gwas.res.with.cov) - nrow(giant.summary)))
```

```
## [1] "# of SNPs not available in GIANT dataframe: 20"
```

```
# add rownames as column to allow merge on column SNP
gwas.res.with.cov.new <- cbind(data.frame(SNP=rownames(gwas.res.with.cov)),gwas.res.with.cov)
# merge giant study data with our data
```

```
merged <- merge(giant.summary,gwas.res.with.cov.new,by='SNP')
# linear fit between coefficients from giant study and our analysis
coefs <- summary(lm(merged$coef~merged$BETA))$coefficients
#plot
ggplot(merged,aes(x=BETA,y=coef)) +
  geom_point() +
  geom_abline(intercept=coefs[1,1],slope=coefs[2,1]) +
  xlab(paste0('coefficients (GIANT study)')) + ylab(paste0('coefficients (our analysis)'))
```



```
print('Correlation coefficient: ')
```

```
## [1] "Correlation coefficient: "
```

```
cor.test(merged$coef,merged$BETA)
```

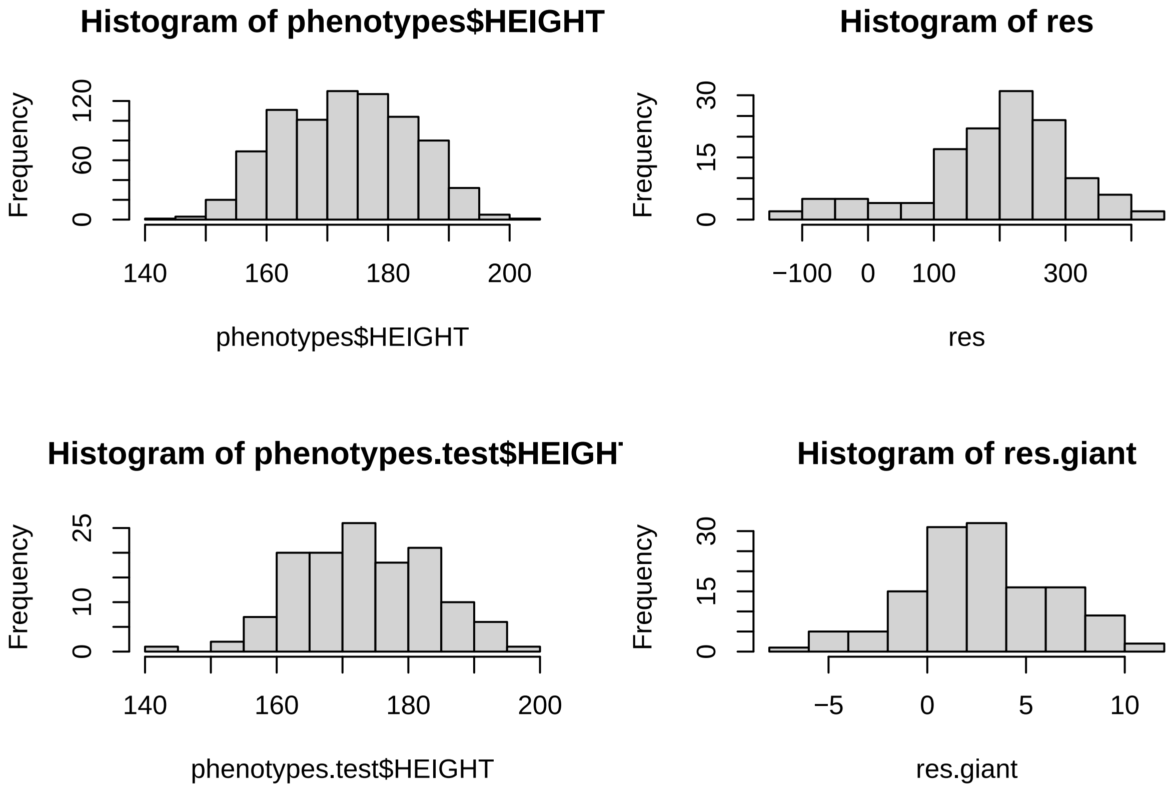
```
##
##  Pearson's product-moment correlation
##
## data:  merged$coef and merged$BETA
## t = 28.82, df = 8700, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2759081 0.3142697
## sample estimates:
```

```
##       cor
## 0.2952079
```

Interestingly, the two sets of coefficients do not seem to have the same scale. The coefficients in the GIANT study are roughly 10 times smaller than the coefficients we found. This could be due to scaling of the response variable in the GIANT study before fitting the individual linear models.

5. Genomic predictions

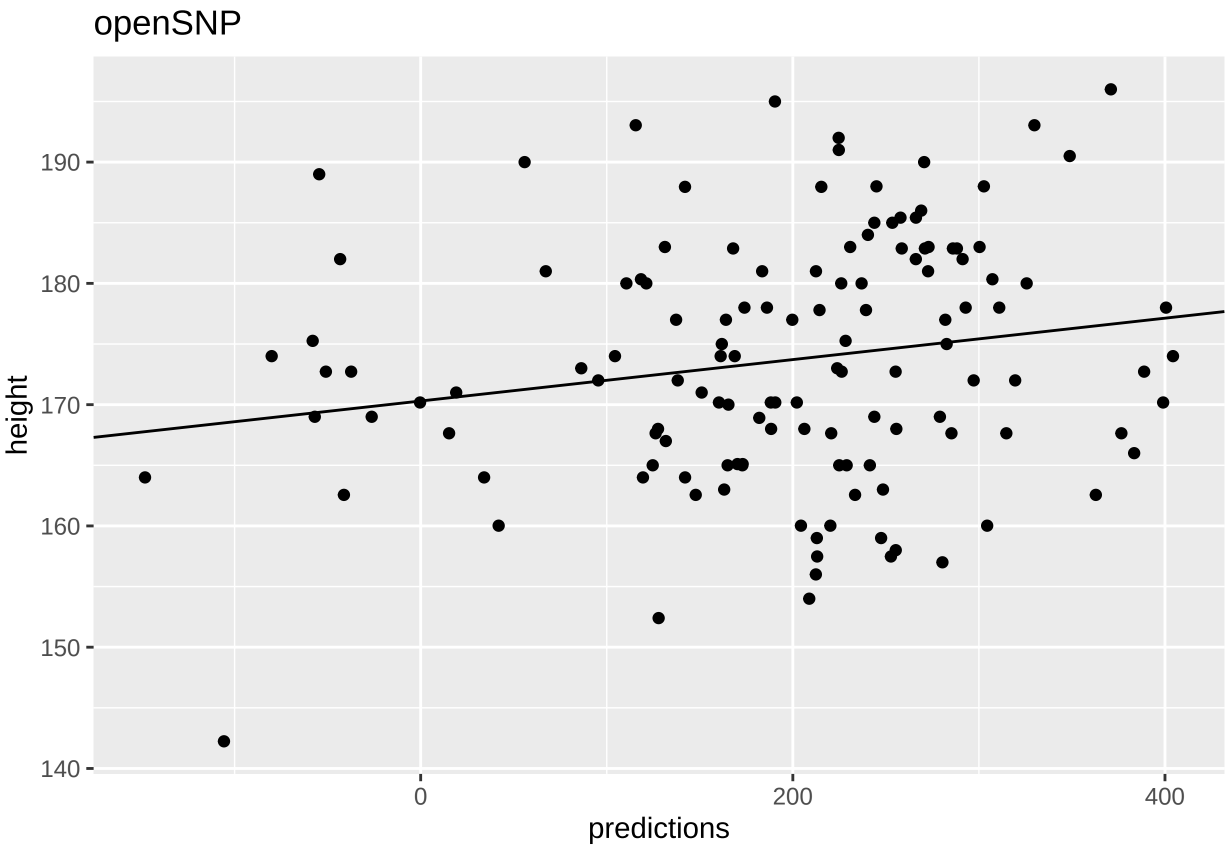
```
genotypes.test <- read.table('ressources/genotypes_test.txt',header=T)
phenotypes.test <- read.table('ressources/phenotypes_test.txt',header=T)
res <- as.matrix(genotypes.test[,giant.summary$SNP]) %*%
  as.matrix(gwas.res.with.cov[giant.summary$SNP,]$coef)
res.giant <- as.matrix(genotypes.test[giant.summary$SNP]) %*%
  as.matrix(giant.summary$BETA)
par(mfrow=c(2,2))
hist(phenotypes$HEIGHT)
hist(res)
hist(phenotypes.test$HEIGHT)
hist(res.giant)
```



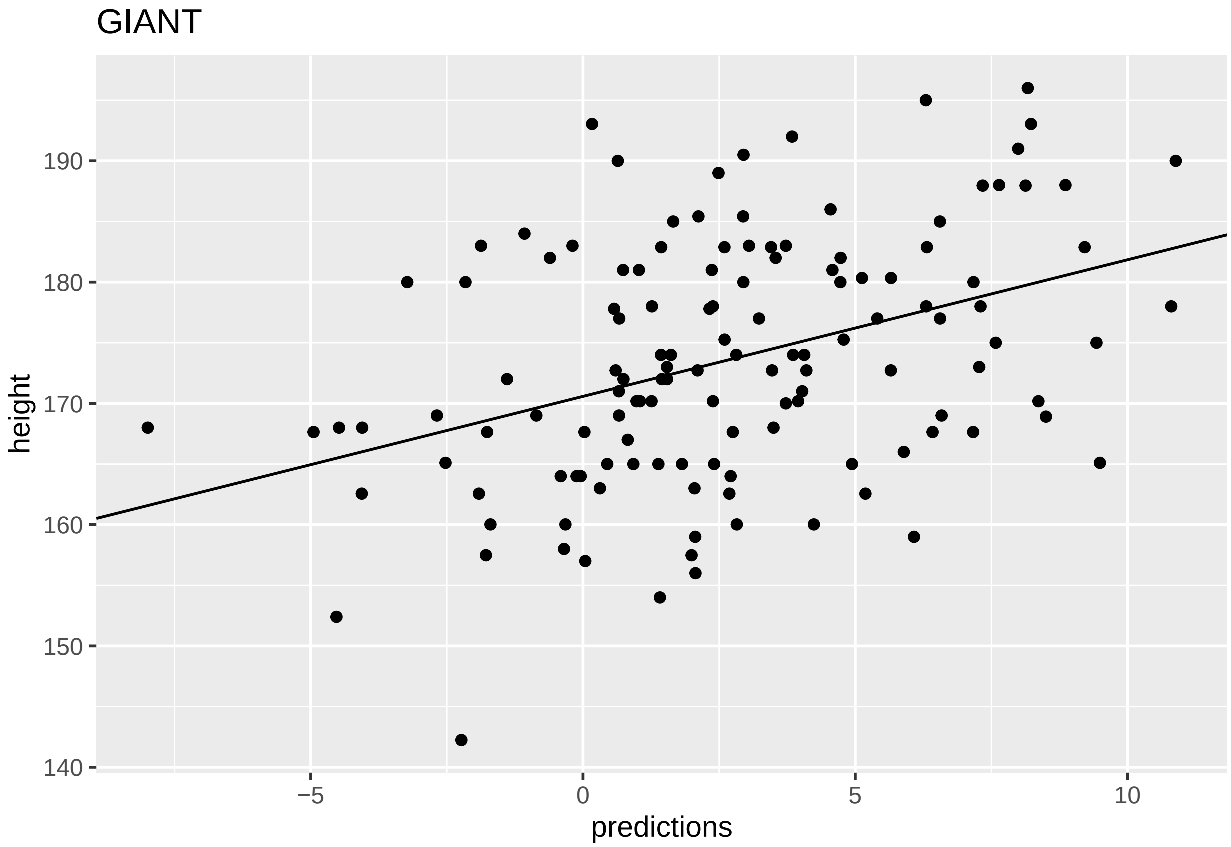
Multiplying the coefficients we found in our analiysis and the coefficients from the GIANT study yields predictions that are not yet on the right scale. It can be observed on the plots above that the predictions made by the GIANT study are closer to a normal distribution than the predictions we got from our analysis. This can be explained by the number of participants in the two studies (n bigger for GIANT).

To get an estimate of the accuracy of our predictions we find the Rˆ2 value between our predictions and the real values in the test set.

```
fit <- lm(phenotypes.test$HEIGHT~res)
ggplot(data.frame(height=phenotypes.test$HEIGHT,predictions=res),aes(y=height,x=predictions)) +
  geom_point() +
  geom_abline(intercept=fit$coefficients[1],slope=fit$coefficients[2]) +
  ggtitle(’openSNP’)
```



```
fit.giant <- lm(phenotypes.test$HEIGHT~res.giant)
ggplot(data.frame(height=phenotypes.test$HEIGHT,predictions=res.giant),aes(y=height,x=predictions)) +
  geom_point() +
  geom_abline(intercept=fit.giant$coefficients[1],slope=fit.giant$coefficients[2]) +
  ggtitle(’GIANT’)
```

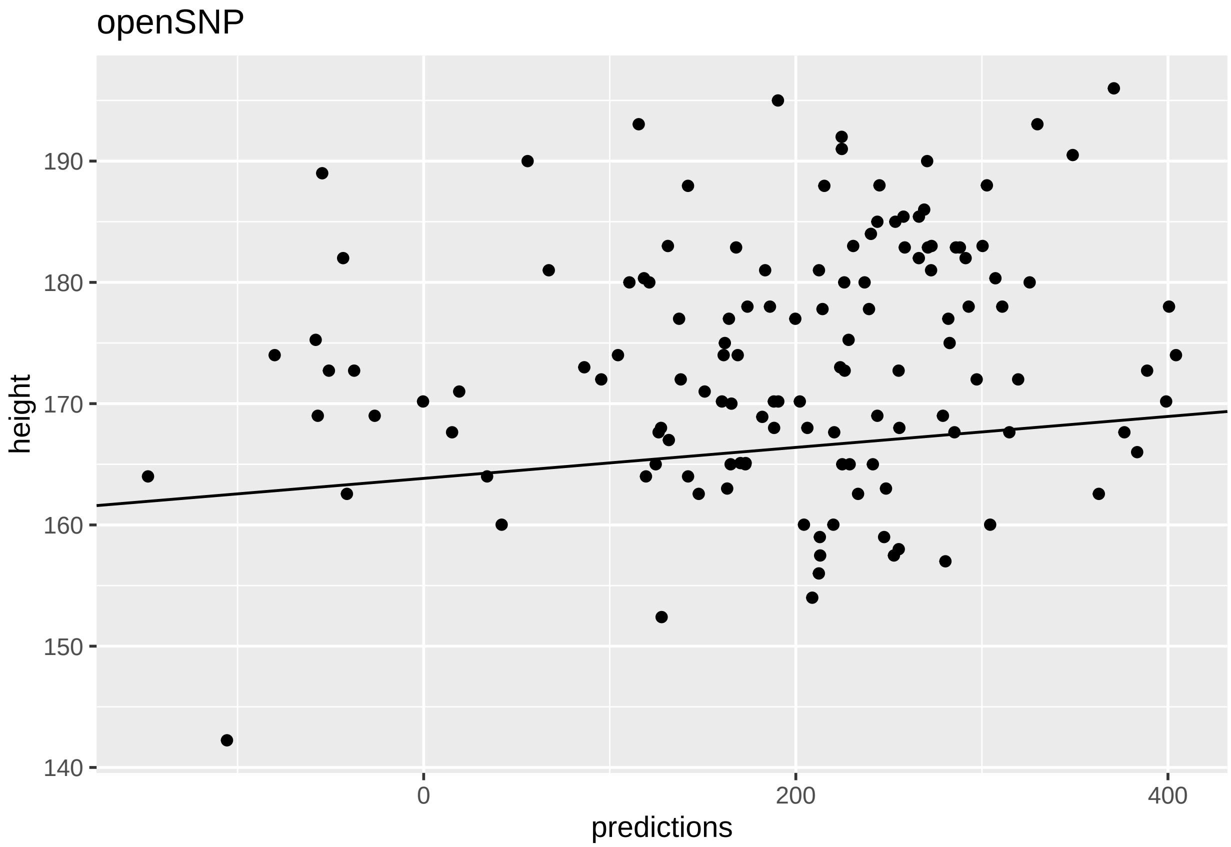


```
cat('R^2:', '\n',
    'openSNP:','\t', as.numeric(summary(fit)['r.squared']), '\n',
    'GIANT: ','\t', as.numeric(summary(fit.giant)['r.squared']))
```

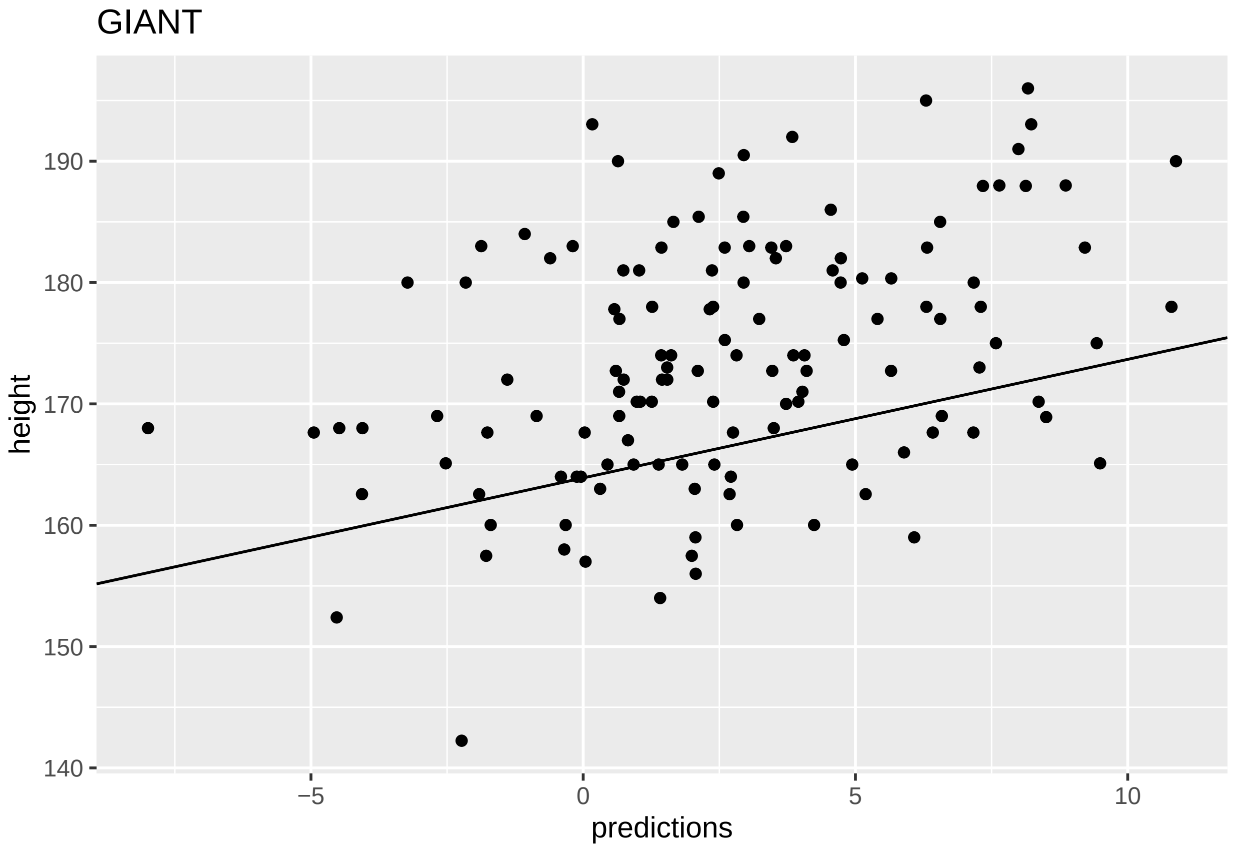
```
## R^2:
##  openSNP:     0.03647657
##  GIANT:       0.1565918
```

We repeat the same analysis but this time we add the covariate SEX to the linear model. We observe that the R^2 value greatly increases through this.

```
covariates.test <- read.table('ressources/covariates_test.txt',header=T)
fit.cov <- lm(phenotypes.test$HEIGHT~res + covariates.test$SEX)
ggplot(data.frame(height=phenotypes.test$HEIGHT,predictions=res),aes(y=height,x=predictions)) +
  geom_point() +
  geom_abline(intercept=fit.cov$coefficients[1],slope=fit.cov$coefficients[2]) +
  ggtitle('openSNP')
```



```
fit.cov.giant <- lm(phenotypes.test$HEIGHT~res.giant+ covariates.test$SEX)
ggplot(data.frame(height=phenotypes.test$HEIGHT,predictions=res.giant),aes(y=height,x=predictions)) +
  geom_point() +
  geom_abline(intercept=fit.cov.giant$coefficients[1],slope=fit.cov.giant$coefficients[2]) +
  ggtitle('GIANT')
```



```
cat(’R^2:’, ’\n’,
    ’openSNP:’,’\t’, as.numeric(summary(fit.cov)[’r.squared’]), ’\n’,
    ’GIANT: ’,’\t’, as.numeric(summary(fit.cov.giant)[’r.squared’]))
```

```
## R^2:
##  openSNP:     0.4325608
##  GIANT:       0.5292932
```