GWAS 6: Confounders and covariates

Matti Pirinen, University of Helsinki

Updated: 21-Mar-2023.

The slide set referred to in this document is "GWAS 6".

When we do a GWAS, our hope is to identify variants that point to biological mechanisms behind the phenotypes, since by knowing those mechanisms we could understand better how each phenotype is generated and, importantly, how harmful phenotypes could be influenced.

Often this agenda is summarized by saying that we look for **causal variants**, i.e., variants whose biological functions contribute to the system that determines the phenotype of interest.

Next we look at extensions of our basic regression models that allow us to decrease the amount of both false positives (i.e. non-causal variants labelled as interesting) and false negatives (i.e. causal variants not labelled as interesting), by including some additional variables into the regression models. Those additional variables, (such as sex, age or population structure), whose effects on the phenotype are not of our primary interest, but that we may include in the analysis for other reasons, are called **covariates**.

To conceptually separate different settings with respect to how and why we should make use of covariates, we define two concepts about relationships between variables:

Independence.

- **Intuition**: Variables Y and X are independent (in a population S) if knowing the value of one of them does not tell anything more about the value of the other than what we can already learn by the population distribution of the other.
- **Definition**: Y and X are independet (denoted by $Y \perp X$) if $P(X, Y) = P(X)P(Y)$, or, equivalently, $P(Y \,|\, X = x) = P(Y)$ for all values of x.
- **Example**: Consider a SNP and denote by A_1 and A_2 the 0-1 indicator of the minor allele at the SNP in the two genomes of an individual. Assume we know that MAF is 0.3 in the population, i.e., $P(A_i = 1) = 0.3$ and $P(A_i = 0) = 0.7$ for $i = 1, 2$. The alleles in the two genomes of an individual from the population are independent if and only if the joint probability distribution of A_1 and A_2 in the population is

A_1	A_2	$P(A_1, A_2) = P(A_1)P(A_2)$	$X = A_1 + A_2$
0	0	$0.7 \cdot 0.7 = 0.49$	0
0	1	$0.7 \cdot 0.3 = 0.21$	1
1	0	$0.3 \cdot 0.7 = 0.21$	1
1	1	$0.3 \cdot 0.3 = 0.09$	2

That is, the genotype X follows the Hardy-Weinberg equilibrium proportions $P(X) = (0.49, 0.42, 0.09)$. The independence means that even if we learned that individual i had inherited allele 1 from the father, our estimate of the probability that the maternal allele of i is also 1 remains 0.3, i.e., the same as it was before we knew about the state of i's paternal allele.

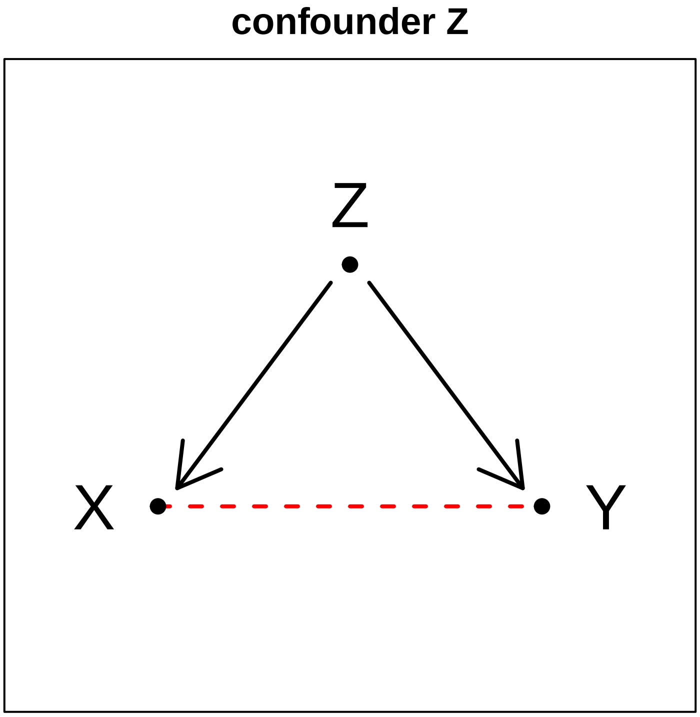
In short: Independent variables do not carry additional information about each other on top of what can be learned from their marginal distribution in the population.

Causal relationship.

- **Intuition**: Variable X causally affects variable Y if a hypothetical intervention that directly changes only the value of X but not the value of any other variable than X, will also (indirectly) affect Y.
- **Example**. You come home every day at 8pm and you (A) press the light switch and (B) take off your shoes. Observation: (C) Your dark home becomes illuminated. Out of A and B, A is a causal effect of C, since if you do *everything else as usual* except not hit the light switch, there will be no light. However, B is not a causal effect of C since if you do *everything else as usual* except that you leave your shoes on, there will still be light.
- **Warning:** Causality is a thorny concept that we can't quite ever define precisely at the same level of mathematical rigor as we can do with other concepts in statistics. However, "causality" is what we would like to have in the end, for example, when we aim to build an effective medical intervention, and therefore we need to educate our intuition and understanding about what that causality means.
  Good news is that it is easy to understand how genetic variants can causally affect traits, which is our main business here; Things get much more complicated when we wonder what should we do with other covariates available that can have all kinds of causal and non-causal relationships between both the genetic variants and the traits of interest.

6.1 Confounding ("correlation does not imply causation")

Confounder (intuitive definition). When we study the relationship between a predictor X and an outcome Y, a confounder Z is a variable that is independently associated with X and Y and can therefore cause a spurious association between X and Y in an analysis that ignores Z.



The mantra “correlation does not imply causation” means that the observed association / correlation between X and Y could be due to a third variable Z that is associated with both X and Y, and hence an X-Y association can appear even without any causal relationship between X and Y. And indeed, this is how most correlations are: Not causal.

Confounding in the Catalogue of Bias.

Example 6.1. If in a heart disease case-control GWAS, the cases are heart disease patients from Helsinki and controls are population samples from the Estonian biobank, then a standard GWAS analysis comparing allele frequencies between cases and controls would indicate statistical differences throughout the genome. However, these associations would not be only heart disease related biologically causal signals, but also spurious associations caused by genetic population structure. Indeed, all variants that have different allele frequencies between Finland and Estonia would show an association with case-control status in this sample, independent of whether the variant has anything to do with the biology of heart disease. Here population structure (Z) between Finland and Estonia is a confounder that is associated with both the SNP frequencies (X) and the outcome (Y). Z is associated with X because population membership affects allele frequencies and Z is associated with Y because our case-control sample was collected in such a way that there is an association with case status and population label (cases were from Helsinki, controls were from Estonia). This exampe is a caricature of a badly designed case-control GWAS that suffers from the population stratification problem (Slides 2-4).

Example 6.2. Since there is a difference (of about 1.5 cm) in average (sex and age adjusted) adult height between East and West Finland, and there is a relatively strong allele frequency gradient between EF and WF, any GWAS on adult height with Finnish samples that does not account for the confounding effect of population structure, risks producing false positives. Note that since variation in height is to a large part genetic, it my well be that some of the E-W allele frequency differences are actually causing some of the observed height differences, but the interpretation of the results would be difficult. This is because the known population structure would confound the results for all those variants that are not biologically linked to height and result in false positive association, and the population structure would also bias the effect estimates at the true height variants.

What can we do with such confounders? The idea is that the analyses should be **stratified** according to the possible levels of the confounder. For example, if we have samples from Finland and from Estonia, we should do separate analyses in both countries and combine the results. The confounder cannot confound the results when all allele frequency comparisons are done among samples that share the level of the confounder, e.g., the two levels of confounder could be having a Finnish background or having an Estonian background. (But note that in Example 6.1, the study design is such that we would have no power left after adjusting for population label because we had no controls from Finland and no cases from Estonia! See slides 6-9 for a more realistic example.) More generally, this idea of stratifying the analysis by the levels of the confounder is implemented by including the confounders as **covariates** in the GWAS regression model. Then we talk about **adjusting the analysis for the covariate**. Technically, multi-level discrete confounders are expanded as indicator variables, one per each level of the confounder, and continuous confounders are simply included as continuous covariates in the regression model. To understand the confounding from regression model’s point of view, let’s define it more precisely.

Confounder (mathematical definition for GWAS setting). Suppose that for fixed values of μ, β and γ and a joint distribution of the random variables X, Y and Z in a population, the phenotype Y truly follows the regression model M:

$$\text{Model M:} \qquad Y \sim \mu + Z\gamma + X\beta.$$

We call the covariate Z a **confounder** of the association between X and Y, if $\beta = 0$ in model M, but, when the covariate Z is omitted and the association between X and Y is described by a simpler model M’:

$$\text{Model M'}: \qquad Y \sim \mu' + X\beta',$$

then $\beta' \neq 0$.

Essentially, this definition says that a confounder is a variable whose omission from a GWAS regression model will cause a spurious association between the genotype and the phenotype. For example, when the population structure (Z) is included as a covariate in a height GWAS in Finland (Example 6.2), it explains away some of the east-west difference in height in Finland. Consequently, those genetic variants (X), whose apparent association with height (Y) is **only** through their geographic allele frequency differences, will not show association in this model (of type M) where population structure is already explicitly included as a covariate. However, when the population structure variable is omitted from the model (of type M'), then all variants that have east-west difference in allele frequency, will also show a statistical association with height, even if their height-association is only through the link between population structure and height and not through a direct biological effect. Thus, population structure is a confounder, and we must include it as a covariate or otherwise we get false positives.

Note that a more general definition of confounding could say that a confounder of X-Y association is any variable that when omitted from the model will *change* the effect estimate of X on Y (i.e. $\beta' \neq \beta$). Our GWAS-motivated definition is narrower: We are only worried about cases where there is no true effect ($\beta = 0$), but where a regression model without appropriate covariates will mislead us by producing an effect ($\beta' \neq 0$).

(A word of warning: Even though we always want to avoid confounding, we still shouldn't always include covariate Z in the model simply because Z is associated with both X and Y! There is a possible problem of collider bias there, that we will study later.)

6.1.1 Detecting confounding The large scale of data in GWAS provides us with an opportunity to assess whether a genome-wide confounding effect seems to be present. We should be worried, if there is a too large a number of apparent associations throughout the genome. A traditional way to assess this has been to look at the QQ-plot at the middle point of the distribution (at the median of the test statistic values) and compare it to the corresponding value under the null hypothesis that there are no true associations at all. The ratio of the median of the observed distribution to the theoretical median of the null distribution is called the **genomic control parameter** λ. The test statistic used in this QQ-plot is often the chi-square statistic of association, but also -log10(P-value) is used. If λ is much > 1, that could indicate that we have a problem with some omitted confounder (such as population structure) that causes elevated signals across the genome. Unfortunately, this approach is not that useful anymore with very large and statistically powerful GWAS, since there λ can elevate substantially from 1 already because of a true polygenic signal distributed across the genome. Instead, LD-score regression (that we will discuss later) is a better way to indicate whether we seem to have missed some important confounders.

Genomic control (GC). Many GWAS have used *genomic control* to adjust for observed inflation (i.e. too many too large values) of the test statistic distribution. The procedure is to divide each chi-square test statistic value by the GC parameter λ (defined above) and this way force the observed test statistic distribution to match the null expectation at the median value. Such method can reasonably adjust for a confounder only if that confounder affects all variants in a similar way. However, considering population structure as an example confounder, some variants are much more geographically stratified than others and hence an adjustment of all variants by the same constant λ is unlikely to be an appropriate solution. GC also keeps the order of the variants the same before and after its application, which seems a too inflexible property to properly model that confounders can work differently with different variants. Furthermore, GC can give a misleading impression that there is no confounding left because the "corrected"" distribution has $\lambda = 1$, whereas in reality GC may have severely undercorrected the most biased variants and overcorrected the less biased variants. Sometimes GC has also been applied to the standard error estimates by multiplying them with $\sqrt{\lambda}$, which has the same effect on the test statistic as dividing the chi-square statistic by λ. This approach again reveals the inflexibility of GC: The effect size estimates are not adjusted at all even though it feels that they should be adjusted according to each variant's correlation with the confounders.

As opposed to GC, regression models that include a confounder as covariate allow each variant to be adjusted according to how correlated the variant is with the covariate.

6.1.2 Avoiding confounding How can we know that we have adjusted for all the relevant confounders? Unfortunately, we can't ever be sure about this in an observational study. But we can at least avoid the common pitfalls causing typical confounding effects in GWAS:

- **Population structure** that is associated with the phenotype will cause inflated P-values across the genome and this can be dealt with by including population structure as a covariate or by using *mixed models* (that we will talk about later).
- **Sample ascertainment** that has been done based on phenotype causes a risk that different phenotypic groups may not match well each other with respect to their genetic backgrounds, because different phenotypic groups may have been collected from different places/times/circumstances. Case-control sampling is the most prominent example of this potential problem. This does not mean that case-control sampling is not a good strategy. It just means that it opens up possible problems that must be taken care of. In GWAS setting, confounding due to sample ascertainment can be seen as a subcategory of general confounding by population structure.
- **Genotyping** of the samples has been done in batches that are not independent of the phenotype, which causes a risk that errors and biases in sample handling and genotyping process lead to spurious genotype-phenotype associations, because such biases may affect differently different batches, and different batches have different phenotype distributions. For example, if cases have been genotyped in Lab 1 in 2009 and controls in Lab 2 in 2015 then technical differences in genotyping process could lead to spurious genotype-disease associations. Often genotyping batches are used as covariates, which adjusts the analysis for the differences in the phenotypic means of the batches, but this strategy cannot be followed if the case-control status is to a large part associated with batches. In those cases, one simply needs to do particularly careful quality control between the genotyping batches. To avoid the batch problem, ideally, one should randomise the samples to genotyping batches, because then the batch could not be associated with the phenotype. Unfortunately, this is rarely possible in practice, since studies tend to combine many sources of existing genotyping data rather than genotyping all samples anew.

Example 6.3. Confounding by population structure. Let's assume that our individuals are geographically spread around the line from Turku (in Southwest Finland) to Kajaani (in Eastern Finland) that is the gradient of the major population structure in Finland. Each individual i has a coordinate $u_i \in [0,1]$, where 0=Turku, 1=Kajaani. For each variant k, we determine its allele 1 frequency on the line by first sampling a Turku-Kajaani difference $d_k \sim \text{Uniform}(-0.5, 0.5)$, and then sampling Turku frequency $t_k \sim \text{Uniform}(0.5, 1)$, if $d_k < 0$; and $t_k \sim \text{Uniform}(0, 0.5)$, if $d_k \geq 0$. Now allele 1 frequency at position $u \in [0,1]$ is $t_k + u \cdot d_k$. So allele frequency changes smoothly from t_k in Turku to $t_k + d_k$ in Kajaani.

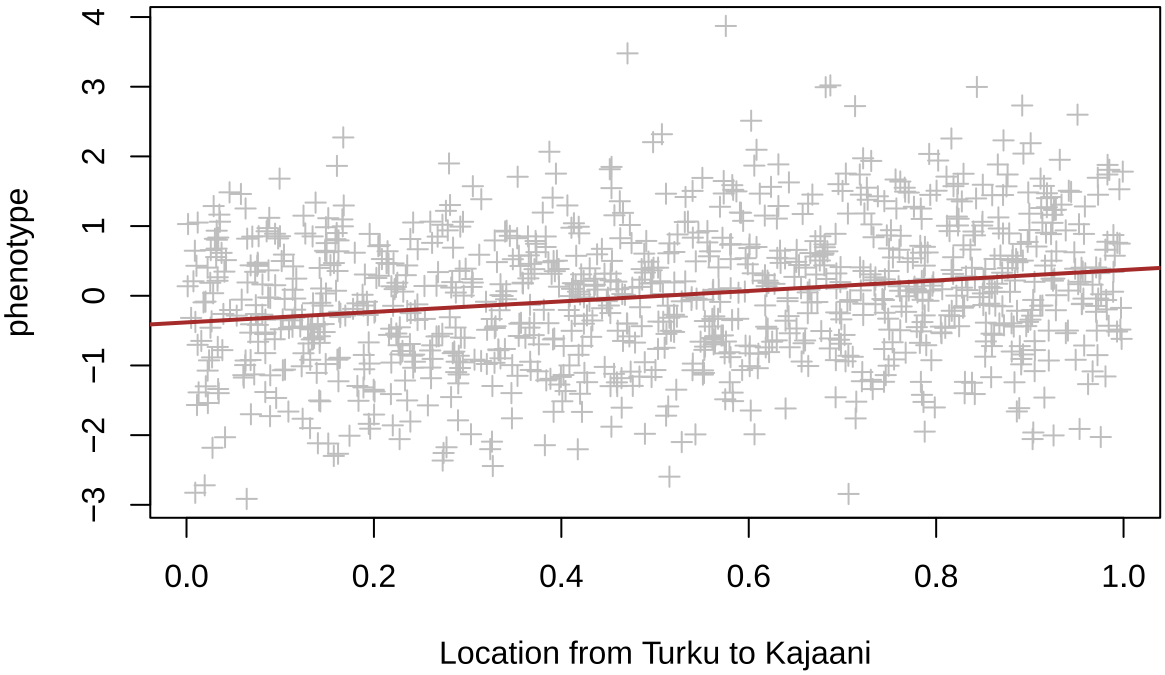
Let's simulate $p = 1000$ such variants for $n = 1000$ individuals randomly picked along the line from Turku to Kajaani.

```r
n = 1000 #individuals
p = 1000 #SNPs
u = runif(n, 0, 1) #coordinates of inds
d = runif(p, -0.5, 0.5) #allele freq difference btw T and K
tu = runif(p, 0, 0.5) #when d>=0 then tu is in (0,0.5)
tu[ d < 0 ] = 1 - tu[ d < 0 ] #when d<0 then tu is in (0.5,1)
X = matrix(NA, nrow = n, ncol = p) #SNP data matrix, rows individuals, columns SNPs
for(k in 1:p){
  f = tu[k] + u*d[k] #allele 1 frequency for each individual
  X[,k] = rbinom(n, size = 2, prob = f) #genotypes from varying frequencies
}
```

Here we have a set of genotypes that carry a population structure effect.

Let's assume that we study a phenotype Y that depends on the environment described by u as $Y = u + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$, but there is no direct effect of any of our p variants on the phenotype.

```
y = u + rnorm(n) #phenotype has a geographical cline but no genetic effect from X
y = scale(y) #standardize trait to have mean = 0 and var = 1 for convenience
#Check how it looks like
plot(u, y, xlab = "Location from Turku to Kajaani", ylab = "phenotype", pch = 3, col = "gray")
abline( lm(y ~ u), col = "brown", lwd = 2 ) #add the regression line to plot
```



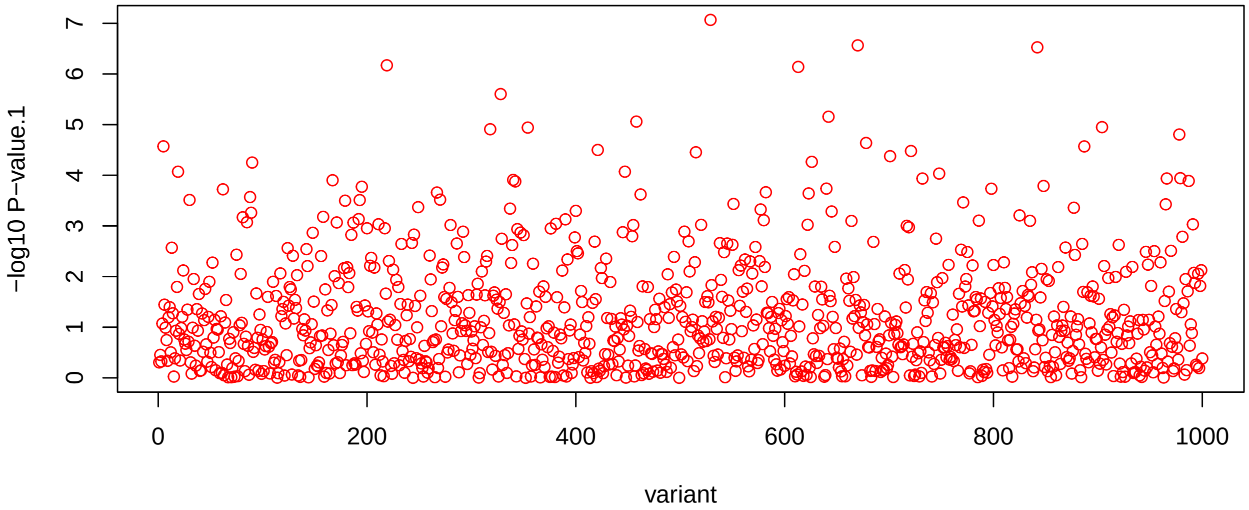
So we have a cline in the phenotype where the mean increases as we move from Turku towards Kajaani.

Let's do a "GWAS" of phenotype Y on our $p = 1000$ variants. We collect only P-values from this GWAS.

```
#collect only P-values, i.e., element [2,4] of lm's summary()
pval.1 = apply(X, 2, function(x){summary(lm(y ~ x))$coeff[2,4]})
summary(pval.1) #median of P-values should be 0.5 under the null
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000001 0.0232523 0.1213125 0.2608591 0.4376539 0.9980830
```

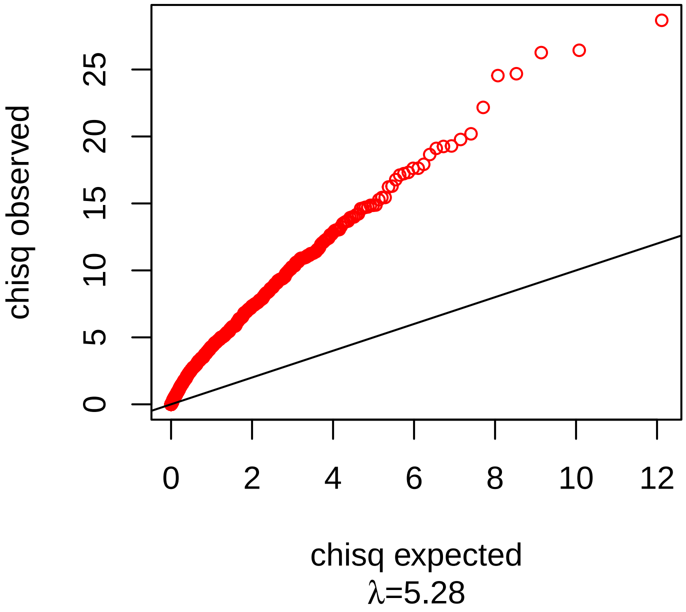
```
plot(1:p, -log10(pval.1), xlab = "variant", ylab = "-log10 P-value.1", col = "red")
```



We have clearly inflated P-values compared to the null hypothesis as our median P-value is 0.12 while under the null the median should be 0.5. Additionally, with 1000 null variants, we would not expect many P-values < 0.001, but here we have a lot of those (in the picture, seen as rising above -log10 = 3).

Let's next draw a QQ-plot to compare the observed association statistics with those expected under the null hypothesis. We could use -log10 P-values to make the QQ-plot, but since it is traditionally done using chisq-statistics, let's also do it like that here. We can turn the P-values to the chi-square test statistics by `chisq = qchisq(pval, df = 1, lower = F)`.

```
#Under NULL p-values are Uniformly distributed between 0 and 1,
#hence chisq-stats are expected to be:
expect.stats = qchisq(ppoints(p), df = 1, lower = F)
obs.stats = qchisq(pval.1, df = 1, lower = F)
lambda = median(obs.stats) / median(expect.stats) #GC lambda = ratio at medians
qqplot(expect.stats, obs.stats, xlab = "chisq expected", ylab = "chisq observed",
       sub = substitute(paste(lambda, "=", lam), list(lam = signif(lambda,3))),
       cex = 0.8, col = "red")
abline(0,1)
```



If observed P-values were an independent sample from the null distribution, then the QQ-plot would follow the diagonal line y=x. Here it deviates strongly from the diagonal *throughout* the distribution, as also measured by $\lambda > 5$. This means that there are hugely more association in the genome than we would expected under the null. This type of an inflated QQ-plot points to confounding by some variable omitted from the model. (Here the confounder is the geographical location).

Before we adjust the analysis for the location, let's update our phenotype slightly so that it is also affected by one of our SNPs.

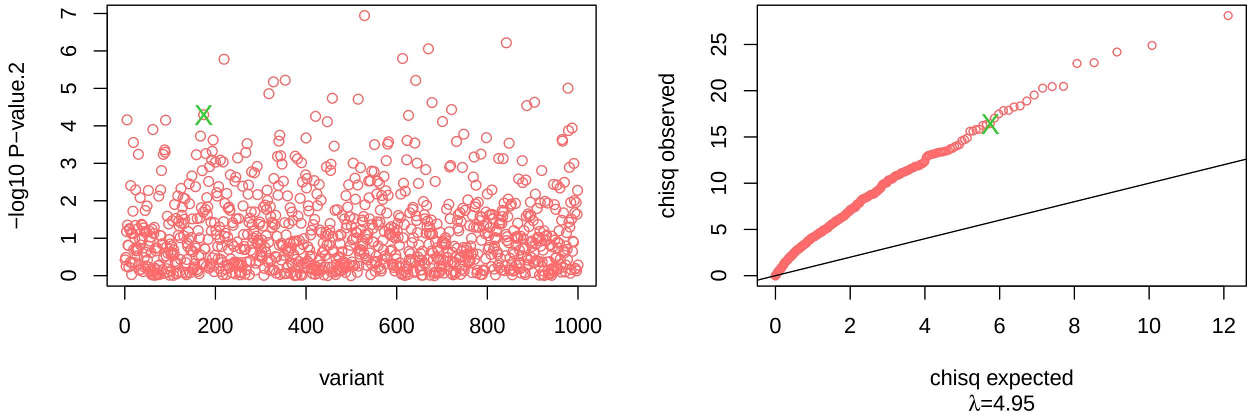
```
snp.id = sample(1:p, size = 1) #Randomly choose one SNP to have an effect

#We'll make this SNP to explain about p.eff  of the trait variance
p.eff = 0.015 #target variance explained
f.mean = tu[snp.id] + 0.5*d[snp.id] #allele frequency at mid point
b = sqrt(p.eff/(2*f.mean*(1 - f.mean))) #effect size corresp. to var explained = p.eff
y = scale(y + X[,snp.id]*b) #new trait is the old one added by the single SNP effect
```

Let's do the association test for the updated trait and make a Manhattan plot and QQ-plot. We mark by green 'X' the SNP that has the effect.

```
pval.2 = apply(X, 2, function(x){summary(lm(y ~ x))$coeff[2,4]})
summary(pval.2) #median would be 0.5 under the null
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000001 0.0258836 0.1335911 0.2653399 0.4517021 0.9997591
```

```
par(mfrow = c(1,2))
plot(1:p,-log10(pval.2), xlab = "variant",
     ylab = "-log10 P-value.2", col = "indianred1")
#Mark the effect SNP by a green cross
points(snp.id, -log10(pval.2[snp.id]), pch = "X", col = "limegreen", cex = 1.3)
#Make a QQ-plot
obs.stats = qchisq(pval.2, df = 1, lower = F)
lambda = median(obs.stats) / median(expect.stats) #GC lambda = ratio at medians
qqplot(expect.stats, obs.stats, xlab = "chisq expected", ylab = "chisq observed",
       sub = substitute(paste(lambda, "=", lam), list(lam = signif(lambda,3))),
       cex = 0.8, col = "indianred1")
points(expect.stats[p + 1 - rank(obs.stats)[snp.id]],
       obs.stats[snp.id], pch = "X", col = "limegreen", cex = 1.3)
abline(0,1)
```



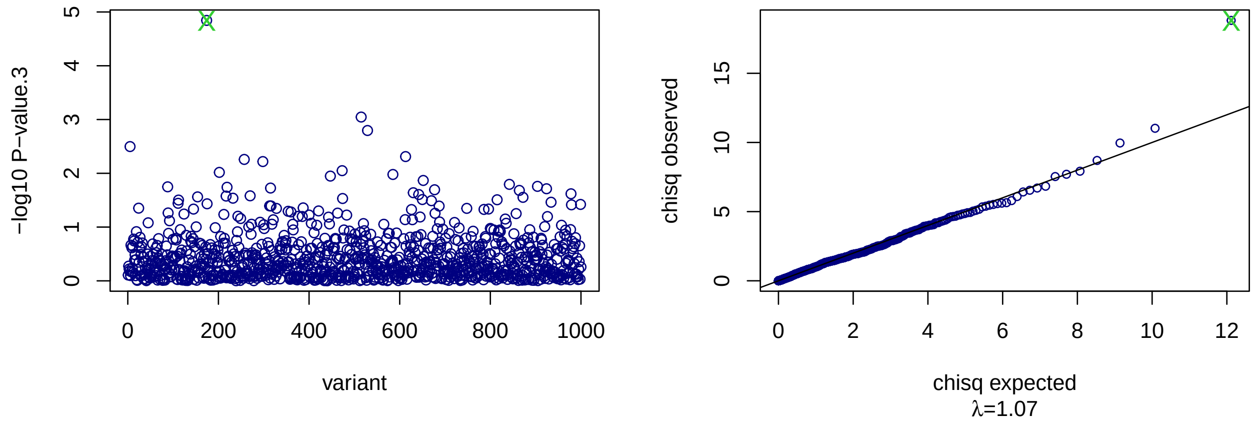
Not only do we have a lot of false positives but also the true association, having a P-value ~1e-4, is masked by the false associations due to the genetic population structure.

Let's then use the geographical position as a covariate in the analysis.

```
pval.3 = apply(X, 2, function(x){summary(lm(y ~ x + u))$coeff[2,4]})
summary(pval.3) #median would be 0.5 under the null
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000143 0.2431451 0.4843497 0.5031942 0.7666718 0.9979641
```

```
par(mfrow = c(1,2))
plot(1:p,-log10(pval.3), xlab = "variant",
     ylab = "-log10 P-value.3", col = "navy")
#Mark the effect SNP by a green cross
points(snp.id, -log10(pval.3[snp.id]), pch = "X", col = "limegreen", cex = 1.3)
#Make a QQ-plot
obs.stats = qchisq(pval.3, df = 1, lower = F)
lambda = median(obs.stats) / median(expect.stats) #GC lambda = ratio at medians
qqplot(expect.stats, obs.stats, xlab = "chisq expected", ylab = "chisq observed",
       sub = substitute(paste(lambda, "=", lam), list(lam = signif(lambda,3))),
       cex = 0.8, col = "navy")
points(expect.stats[p + 1 - rank(obs.stats)[snp.id]],
       obs.stats[snp.id], pch = "X", col = "limegreen", cex = 1.3)
abline(0,1)
```



What a beautiful class room example we have here! The confounder correction in the regression model gave us:

- P-value distribution that follows the null, except for the true effect.
- Ability to see the true effect more clearly (here leading to lower P-value) after the confounding effect has been explained away from the phenotype.

And what if we did not know each individual's geographic location u in the first place? Then we would adjust the analysis for the principal components (PCs) of the population structure! In practice, first 10 PCs are often included in GWAS as the first one alone might not yet capture the relevant structure. In general, one should check how many PCs explain the phenotype and consider including them into the model, as we do below.

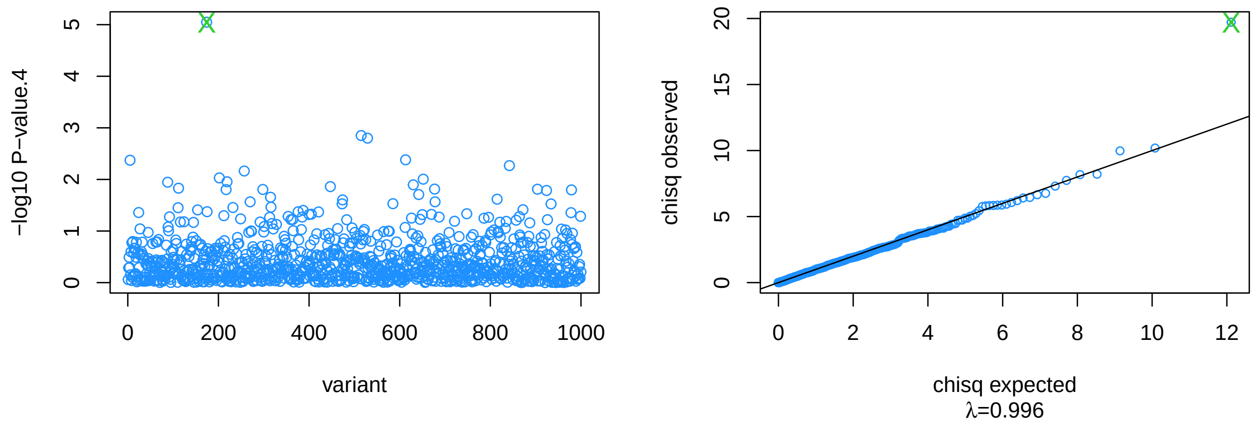
```
pca = prcomp(X, scale. = T) #Make PCA
summary(lm( y ~ pca$x[,1:10] ) )$coeff
```

```
##                       Estimate  Std. Error       t value     Pr(>|t|)
## (Intercept)       -2.861268e-17 0.030824224 -9.282531e-16 1.000000e+00
## pca$x[, 1:10]PC1   2.659184e-02 0.003919903  6.783800e+00 2.012343e-11
## pca$x[, 1:10]PC2  -1.043886e-02 0.015930062 -6.552928e-01 5.124316e-01
## pca$x[, 1:10]PC3  -4.017938e-02 0.016021508 -2.507840e+00 1.230640e-02
## pca$x[, 1:10]PC4   7.878589e-04 0.016046023  4.909995e-02 9.608496e-01
## pca$x[, 1:10]PC5   2.200017e-02 0.016098531  1.366595e+00 1.720629e-01
## pca$x[, 1:10]PC6   1.604440e-02 0.016211901  9.896681e-01 3.225784e-01
## pca$x[, 1:10]PC7  -8.147639e-03 0.016241908 -5.016430e-01 6.160303e-01
## pca$x[, 1:10]PC8  -1.452559e-02 0.016300806 -8.910967e-01 3.730940e-01
## pca$x[, 1:10]PC9   2.004513e-02 0.016319913  1.228262e+00 2.196409e-01
## pca$x[, 1:10]PC10 -3.392253e-02 0.016378707 -2.071136e+00 3.860498e-02
```

Let's use the first 3 since also PC3 seems to have some predictive power in addition to PC1.

[What exactly was the logic of choosing 3 PCs? As PCs are orthogonal to each other, their individual P-values from the joint model with 10 PCs indicate whether they seem better predictors than just random noise. As it does not harm to include some more PCs even if they are not predictive, we don't need to be particularly strict here with the inclusion criteria. Since someone might say that P-value of 0.01 for the 3rd PC suggests that it is important, we just include it here. And when we include PC3, then we also include PC2 because PC2 is expected to capture more of the genetic structure than PC3. What about PC10? Well, we can run the analysis also with PCs 1-10 and see that the results don't change.]

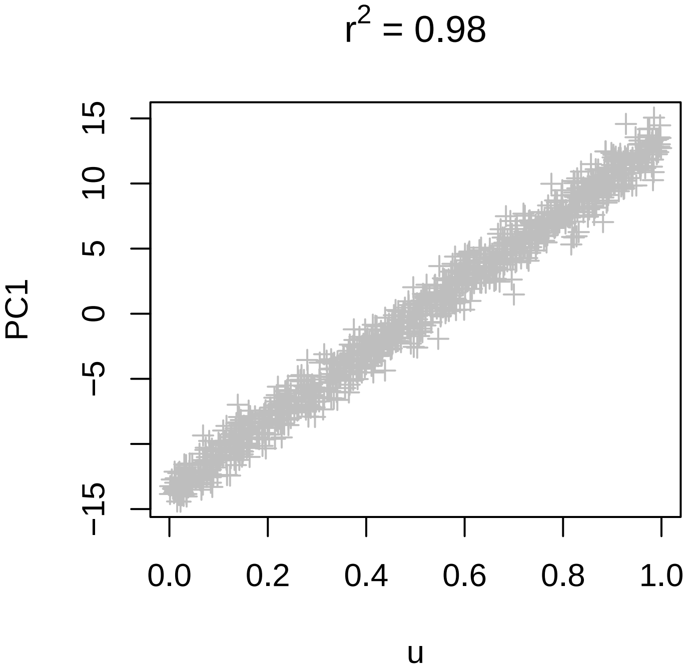
Let's do the GWAS using PCs 1-3 as covariates and make a Manhattan plot and QQ-plot as above (suppressing code from the document).



Results are almost identical to those of the perfect location information.

Let's check how well PC1 captures the location information u.

```
plot(u, pca$x[,1], xlab = "u", ylab = "PC1", col = "gray", pch = 3,
     main = substitute(paste(r^2," = ",r.val),
                       list(r.val = signif(cor(u,pca$x[,1])^2,2))))
```



PC1 captures u very accurately, and hence here we can control for confounding by population structure simply by adding PC1 as a covariate in the regression model.

We have seen above that we should always include potential confounders in the GWAS regression model or otherwise we risk getting false positive genotype-phenotype associations. But what if we have a covariate that is not a confounder? Should we include that in the model? Unfortunately, there is no simple answer to this question and we will next look at some prominent cases.

6.2 Collider bias

Wouldn't it be interesting if we had an autosomal variant that were associated with biological sex? In other words, a variant in autosomal genome would have different allele frequencies in males than in females. We don't know of any such variant but here we keep on trying to find it with some help from a "collider" variable.

Consider a variant with MAF 0.5 in population and whose distribution is independent of biological sex. Let's test its effect on sex.

```
f = 0.5 #MAF
N = 50000 #sample N males and N females from population
y = rep(c(0,1), each = N) #"phenotype" = biological sex (0 = female, 1 = male)
x = rbinom(2*N, size = 2, prob = f) #genotypes in males and females
summary( glm(y ~ x, family = "binomial") )$coeff[2,]
```

```
##     Estimate   Std. Error      z value     Pr(>|z|)
## -0.013417702  0.008963701 -1.496893124  0.134421070
```

No association, as expected.

Suppose then that this genetic variant had an effect on height. Let's say that the allele 1 increases height by 1cm and the SD around the genotype means is 5cm. The mean height is 175cm in males and 165cm in females. Let's repeat the test for the genotype-sex association but let's include height as covariate.

```
z = 164 + x*1 + 10*y #mean is 164,165,166 for females with genotype 0,1,2; and +10 for males
z = z + rnorm(2*N, 0, 5) #variation with SD=5cm around the sex + genotype means for each individual
summary( glm(y ~ x + z, family = "binomial") )$coeff[2:3,]
```

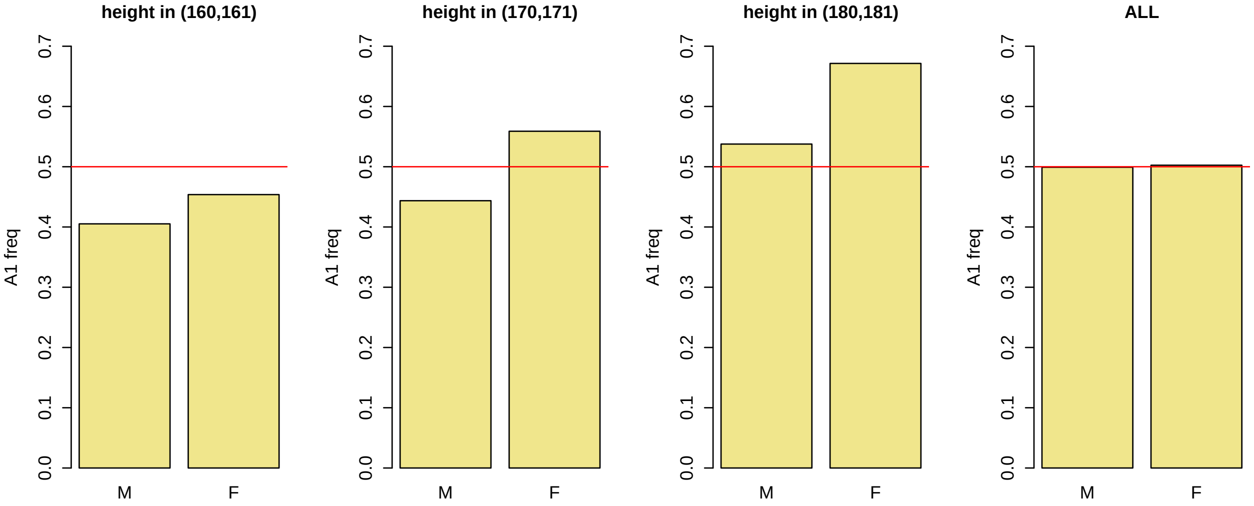
```
##     Estimate  Std. Error   z value      Pr(>|z|)
## x -0.4139522 0.013596669 -30.44512 1.390148e-203
## z  0.4010095 0.002454253 163.39373  0.000000e+00
```

This test shows that, unsurprisingly, height is highly predictive of sex, but also, surprisingly(?), the genotype x has now become strongly associated with sex. This raises two questions: What has just happened? and Is that a problem?

What? A way to think about GWAS for a binary phenotype (here sex), that has been adjusted for a covariate, is that the test is asking whether the phenotype groups (here males and females) have different genotype distributions *within a fixed level of the covariate value* (here among individuals who have the same height), and then takes a weighted average of those results over all possible levels of the covariate, where the "weights" represent the accuracy of the results reflecting the effective sample size at each value of the covariate.

Let's demonstrate how the sex-specific allele 1 frequency varies by the height of the individuals, and, as a comparison, what is the allele 1 frequency in ALL individuals. Let's use three bins for height values, each of length 1 cm.

```
cut.points = matrix(c(160,161,
                      170,171,
                      180,181),
                    byrow = T, ncol = 2) #bins for height
par(mfrow = c(1,4))
for(ii in 1:(1 + nrow(cut.points))){
  if(ii <= nrow(cut.points)){ #this is a height stratified plot
    ind.m = (y == 1 & z > cut.points[ii,1] & z <= cut.points[ii,2]) #males
    ind.f = (y == 0 & z > cut.points[ii,1] & z <= cut.points[ii,2]) #females
    title.txt = paste0("height in (",cut.points[ii,1],",",cut.points[ii,2],")")
  }else{ #this last plot is for ALL individuals, independent of height
    ind.m = (y == 1) #males
    ind.f = (y == 0) #females
    title.txt = "ALL"
  }
  barplot(matrix(
    c(sum(x[ind.m])/2/sum(ind.m), # male allele 1 frequency
      sum(x[ind.f])/2/sum(ind.f)), ncol = 2), # female allele 1 frequency
    ylim = c(0,0.7), col = "khaki", ylab = "A1 freq",
    main = title.txt, cex.main = 1.3, cex.lab = 1.3, cex.axis = 1.3, cex.names = 1.3,
    names.arg = c("M","F"))
  abline(h = 0.5, col="red")
}
```



We see that even though the allele 1 frequency is 0.5 for ALL males and ALL females (4th panel), in every height-stratified comparison, the allele 1 frequency is higher in females than in males.

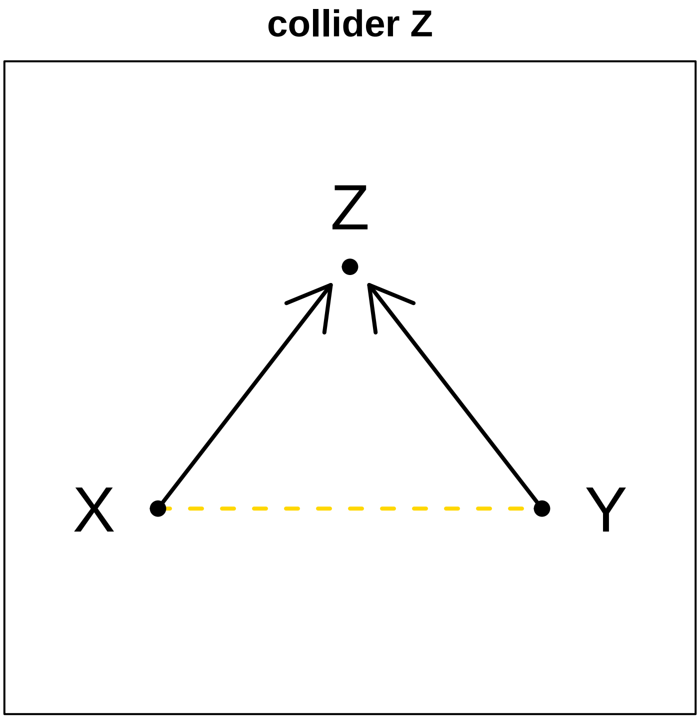
Consider the middle bin (170,171). Males in that bin are a bit shorter than average male, so their allele 1 frequency is a bit less than the average of 0.5. On the other hand, the females in this bin are a bit taller than average female and hence their allele 1 frequency is a bit higher than the average of 0.5.

Consider then the first bin (160,161). Not many males belong to this bin, but those who do are clearly shorter than the males on average and hence they tend not to have height increasing alleles. Consequently, the allele 1 frequency is much lower for these males than its population frequency. Even though also the females in this bin are shorter than average, for them the deviation from the mean height is not that large as for the males in this bin, and hence the drop from the population allele frequency is also less for these females than it is for the males in this bin.

And similar argument can be made for the 3rd bin that only contains very tall women, who therefore have a high frequency of allele 1, whereas males are only a bit above the male average.

All in all, even though there is no association between allele 1 and sex in the population, still, for any fixed level of height, females have higher relative frequency of allele 1 than males. This creates a positive statistical association between the height increasing allele and being female in an analysis that adjusts for height. Given that there is no association with the variant and sex in the general population, we consider this association misleading. This phenomenon, called collider bias, was also demonstrated in practice with the UK Biobank data by Day et al. 2016 (slide 11).

Collider bias can create an association between uncorrelated variables X and Y if both X and Y are causally affecting Z, and Z is used as a covariate in the regression model $Y \sim X + Z$. (slide 10)



Collider bias in the Catalogue of Bias.

This shows that when studying the association between X and Y, we shouldn't blindly adjust for all covariates that are associated both with X and Y, but think also what the model means in terms of what could be causing what. We must adjust for confounders to avoid false positives but we (most likely) do not want to adjust for colliders because such associations are not causal for the phenotype Y we are studying.

6.3. Mediation

Suppose that we study Type 2 Diabetes (Y) and we have BMI information (Z) available. We know that high BMI is a risk factor for T2D. For now, let's suppose that high BMI is a causal risk factor for T2D.

Let's generate data for $n = 500,000$ individuals where each unit of BMI is increasing logOR of T2D by 0.15 and where we have two variants X_1 and X_2 with MAF 0.5 of which X_1 has a causal effect on T2D of logOR of 0.15 and X_2 has an effect on BMI of one unit per allele 1, but X_2 has no direct effect on T2D.

```
n = 5e5 #large n to get accurate frequencies in BMI bins later
f = 0.5 #MAF
preval = 0.15 #population prevalence of T2D
mean.bmi = 27
x = replicate(2, rbinom(n, size = 2, prob = f)) #genotypes at two variants
z = rnorm(n, mean.bmi, 3) + x[,2]*1 #bmi values in the sample affected by x2
log.odds = log(preval/(1-preval)) + (z - mean.bmi)*0.15 + (x[,1] - 2*f)*0.15 #logodds of T2D depend on
p.t2d = exp(log.odds)/(1 + exp(log.odds)) #probability of T2D for each individual
y = rbinom(n, size = 1, prob = p.t2d) #simulate binary T2D status
table(y)/n #check that sample prevalence is reasonably close to given 'preval'
```

```
## y
##        0        1
## 0.819632 0.180368
```

```
#Let's do association test for the variants
summary( glm(y ~ x, family = "binomial") )$coeff[2:3,]
```

```
##     Estimate  Std. Error  z value      Pr(>|z|)
## x1 0.1422925 0.005217518 27.27207 9.097332e-164
## x2 0.1461203 0.005213999 28.02461 8.147087e-173
```

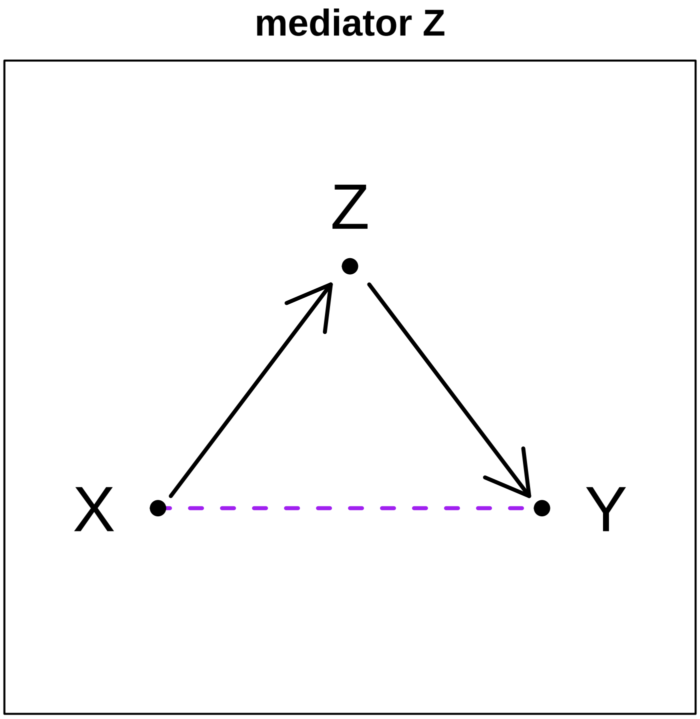
```
#Let's do association test for the variants by ADJUSTING for BMI
summary( glm(y ~ x + z, family = "binomial") )$coeff[2:4,]
```

```
##        Estimate  Std. Error     z value      Pr(>|z|)
## x1 0.1467049640 0.005297922  27.6910378 8.952148e-169
## x2 0.0006037135 0.005425930   0.1112645  9.114066e-01
## z  0.1507729575 0.001284674 117.3628021  0.000000e+00
```

In the first model $Y \sim X_1 + X_2$, that does not adjust for BMI, we see associations for all risk SNPs of T2D, even for those that affect T2D risk only through their effects on BMI.

In the second model $Y \sim X_1 + X_2 + Z$, that adjusts for BMI, we see association only for those SNPs whose effect on T2D does not (completely) go through BMI.

Here BMI is **mediating** the effect of SNP2 on T2D while the T2D effect of SNP1 is independent of BMI.

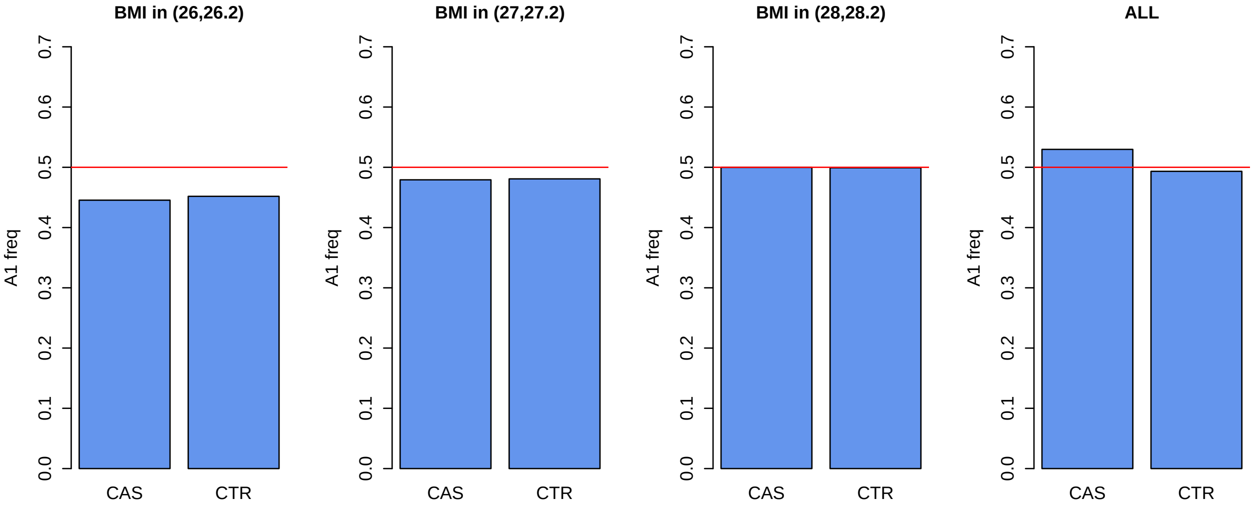


Which one is the correct model to use? Both models are correct for answering their respective questions, which are different from each other. It is for the analyst to decide which question is more relevant: SNPs that affect T2D, no matter which way, or SNPs that affect T2D through something else than BMI? In practice, if we apply both models, we can learn which variants seem to mainly affect BMI and which affect T2D risk also independent of BMI.

Importantly, if two GWAS on T2D (or on any other trait) are compared or combined, one must pay attention to whether same/similar covariates have been used in them. For some SNPs (like SNP2 above), the effect on T2D is very different when it has been adjusted for BMI from when it has not been so adjusted, and combining $\widehat{\beta}$s from such analyses risks comparing apples to oranges.

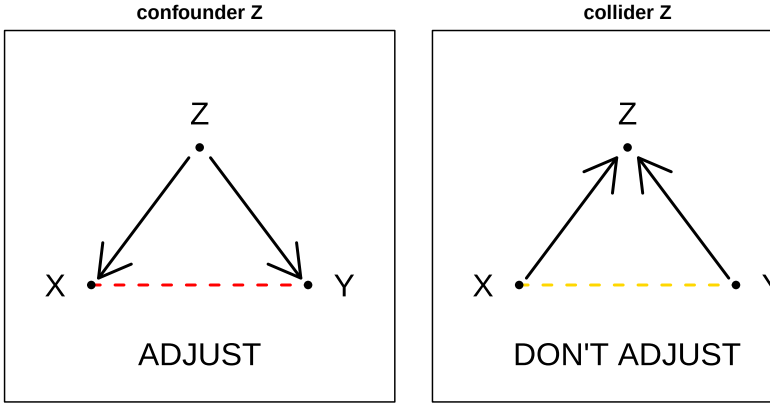
It is instructive to think how the distribution of allele 1 of SNP2, that is completely mediated through BMI, behaves as a function of BMI, as compared to its marginal distribution. Let's repeat the barplot that we used above to demonstrate the collider effect, but now we use T2D case-control label as the binary phenotype and BMI stratified by the cutpoints as the covariate. (We used large n in this example so that the allele frequencies are accurate even within narrow BMI bins.)

```
cut.points = matrix(c(26,26.2,
                      27,27.2,
                      28,28.2),
                    byrow = T, ncol = 2) #cutpoints
par(mfrow = c(1,4))
for(ii in 1:(1 + nrow(cut.points))){
  if(ii <= nrow(cut.points)){ #this is a BMI stratified plot
    ind.s = (y == 1 & z > cut.points[ii,1] & z <= cut.points[ii,2]) #cases
    ind.r = (y == 0 & z > cut.points[ii,1] & z <= cut.points[ii,2]) #controls
    title.txt = paste0("BMI in (",cut.points[ii,1],",",cut.points[ii,2],")")
  }else{ #this last plot is for ALL individuals, independent of BMI
    ind.s = (y == 1) #cases
    ind.r = (y == 0) #controls
    title.txt = "ALL"
  }
  barplot(matrix(
      c(sum(x[ind.s,2])/2/sum(ind.s), # case allele 1 frequency
        sum(x[ind.r,2])/2/sum(ind.r)), ncol = 2), # control allele 1 frequency
      ylim = c(0,0.7), col = "cornflowerblue", ylab = "A1 freq",
      main = title.txt, cex.main = 1.3, cex.lab = 1.3, cex.axis = 1.3, cex.names = 1.3,
      names.arg = c("CAS","CTR"))
  abline(h = 0.5, col = "red")
}
```



We see that, for a fixed value of BMI, there is no difference in allele 1 frequencies in cases and controls, but the allele 1 frequency increases with increasing BMI. However, when we ignore BMI information (4th panel), then allele 1 is more frequent in cases than in controls because it is more frequent in high BMI individuals and high BMI individuals are relatively more frequent among cases than among controls. This is exactly how effects that are mediated through a covariate behave: the effect is observed only when the analysis has not been adjusted for the mediator.

Summary of covariates Z associated with X and Y



6.4 Covariates independent of genetic variants in population samples

First, if a variable is not associated with the phenotype, then we have no reason to include it in the model. So in what follows, we consider covariate W, that is associated with phenotype Y, but is not associated with SNP X at the population level, and hence is not a confounder, collider or mediator at the population level. Omission of such a covariate W from a linear or logistic regression model cannot create a false positive association between X and Y (although it can change the effect size being estimated in logistic regression). Thus we will have a choice to make between two valid models M and M', where the first adjust for W and the latter does not adjust for W

$$\begin{aligned} \text{Model M} &: \quad Y \sim \mu + W\gamma + X\beta, \\ \text{Model M'} &: \quad Y \sim \mu' + X\beta'. \end{aligned}$$

We want to know how models M and M' differ

- in effect size estimates β and β'
- in SEs of β and β'
- in statistical power to detect non-zero genetic effects.

Note on notation: we used Z for covariates associated with X and now we use W for a covariate independent of X.

Ascertained case-control studies will have their own section below and this section 6.4. is about **random population samples**.

Explaining away noise in linear model The main intuition about X-independent covariate W is that it explains away some of that variation of the (quantitative) phenotype Y that is not associated with the SNP X. Thus, from X's point of view, the phenotypic variance explained by W is random noise: variation without any pattern that could be explained by X. When we get rid of this noise, the effect of X on the remaining, noise-reduced, variation of Y, can be more clearly seen.

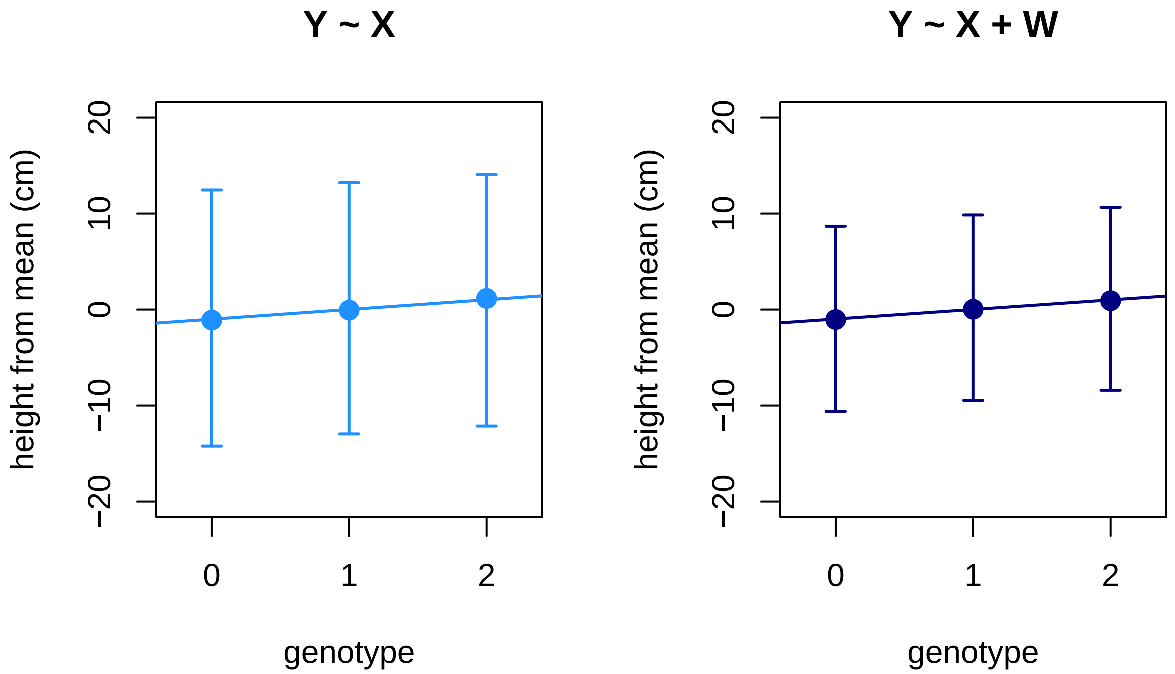
Let's demonstrate this with an example of a GWAS on height (Y) where sex (W) is a highly predictive covariate which is not associated with any autosomal SNP (X).

Example 6.4. We study genetics of height by a sample of 10,000 males and 10,000 females. Distribution of height (in cm) in males is $\mathcal{N}(175, 5^2)$ and in females $\mathcal{N}(165, 5^2)$. Let's assume that there is a genetic variant (with MAF $f = 0.5$) that increases height by 1cm in both sexes.

In this case, adjusting the linear regression for sex is the same as to subtract the sex specific mean height (either 165cm or 175 cm for females and males, respectively) from each individual's height value before the regression. It has an effect of shifting the male and female heights around a common mean value of 0 and shrinking the variance compared to the original population height distribution that had two separate modes (for two sexes) located 10 cm apart. The unadjusted model is equivalent to fitting a linear regression to the height values that have been mean centered by population mean height (170 cm) and that have the same variance as the original height values.

Let's generate the data and make two plots of the genotype phenotype relationship where the first one does not account for sex and the second adjusts for sex.

```r
f = 0.5
b = 1
N = 1e4 #N males and N females
var.expl = 2*f*(1-f)*b^2 #variance explained by the SNP.
w = rep(c(0,1), each = N)
x = rbinom(2*N, size = 2, p = f) #genotypes for females and males
y = 164 + x*b + w*10 #mean height 164,165,166 in geno 0,1,2 in females; +10 in males
y = y + rnorm(2*N, 0, sqrt(5^2 - var.expl)) #variation around mean height for each ind.
y = y - mean(y) #mean center in population by shift of about -170

cols = c("dodgerblue", "navyblue")
par(mfrow = c(1,2))
y.stat = matrix(NA, ncol = 3, nrow = 3)
for(ii in 0:2){y.stat[ii+1,] = as.numeric(quantile(y[x == ii],c(0.025,0.5,0.975)))}
plot(0:2, y.stat[,2], col = cols[1], pch = 19, cex = 1.3,
     ylim = c(-20, 20), xlim = c(-0.3, 2.3), xaxt = "n",
     xlab = "genotype", ylab = "height from mean (cm)", main = "Y ~ X")
axis(1, at = 0:2, labels = 0:2)
arrows(0:2, y.stat[,1], 0:2,y.stat[,3], code = 3, angle = 90,
       lwd = 1.5, length = 0.05, col = cols[1])
lm.all = lm(y ~ x)
abline(lm.all, col = cols[1], lwd = 1.5)

#Let's adjust the phenotype for sex
y.adj = residuals( lm(y ~ w) ) #this subtracts the sex-specific mean from height
y.stat = matrix(NA, ncol = 3, nrow = 3)
for(ii in 0:2){y.stat[ii+1,] = as.numeric(quantile(y.adj[x == ii],c(0.025,0.5,0.975)))}
plot(0:2, y.stat[,2], col = cols[2], pch = 19, cex = 1.3,
     ylim = c(-20, 20), xlim = c(-0.3, 2.3), xaxt = "n", main = "Y ~ X + W",
     xlab = "genotype", ylab = "height from mean (cm)")
axis(1, at = 0:2, labels = 0:2)
arrows(0:2,y.stat[,1],0:2,y.stat[,3], code=3, angle = 90,
       lwd = 1.5, length = 0.05, col=cols[2])
lm.adj = lm(y.adj ~ x)
abline(lm.adj, col = cols[2], lwd = 1.5)
```



We see that the variation in height within the genotype groups is larger in the unadjusted analysis (left panel) than in the adjusted analysis (right panel). Let's see the effect estimates and P-values.

```
summary(lm.all)$coeff[2,] #unadjusted for W
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.010223e+00 7.038656e-02 1.435250e+01 1.749226e-46
```

```
summary(lm.adj)$coeff[2,] #adjusted for W
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 9.936359e-01 4.918877e-02 2.020046e+01 7.638074e-90
```

Both models are estimating the correct effect $\beta = \beta' = 1$, i.e., models are unbiased. However, the accuracy measured by SE is different in the two cases. For linear model, SE$= \sigma_\varepsilon/\sqrt{2nf(1-f)}$, where σ_ε is the error standard deviation, that is, SD of the residuals after the covariates and the SNP effect have been accounted for. By using a covariate W that explains some of the error variance, we can make the SE of the SNP effect smaller, and hence increase the precision of the SNP effect's estimator, and also increase power to detect a non-zero SNP effect.

Let's confirm we have all the pieces to understand the difference between the SEs of the two models. We known that after sex is accounted for, SD in the data should be 5 cm (since it is 5 cm in each of the sexes separately). Thus, we expect that the ratio of the error SDs of the adjusted model/unadjusted model is about

```
5/sd(y)
```

```
## [1] 0.705417
```

This should correspond to the ratio of observed SEs between the models:

```
summary(lm.adj)$coeff[2,2]/summary(lm.all)$coeff[2,2]
```

```
## [1] 0.6988375
```

as seems to be the case.

We conclude that the covariate improves precision of the estimate of the SNP effect by explaining away some of the variance in phenotype. This increase in precision will also lead to higher statistical power. Thus, one should always adjust for W in linear model.

Binary phenotype in population samples A way to think about a GWAS of a binary phenotype with covariate W, such as sex, is that we first split the data into males and females, then we do the logistic regression in each sex separately, and finally we combine the results by weighting each of them with the effective sample size of the corresponding sex.

By splitting the data into groups we will increase SE of the final combined estimate compared to the SE of a single combined analysis that ignores the covariate except when the proportion of cases is exactly the same among the groups, in which case SE does not change between the analyses. Thus, in practice, SE of genetic effect will increase when we include the covariate in the model.

The effect β estimated by the covariate-adjusted model is larger in magnitude than β' of the unadjusted model, but if the prevalence of disease in population is low enough (say, a few percent or less), then this change in effect size is tiny. So in typical disease studies, the estimated genetic effect does not change much between the two possible regression models.

It can be shown that the combined effect of these changes in βs and SEs is always such that the power of the covariate-adjusted model is higher than that of the covariate-ignoring model in logistic regression analysis of population samples. However, for a low-prevalence disease, the difference in power between the two models is very small.

As a conclusion, inclusion of such a covariate W, that is independent of X and does not interact with the effect of X,

- increases the statistical power to detect a non-zero effect of X both in linear and logistic regression model of population samples,
- does not change the effect of X that is being estimated in linear model, but does increase the absolute value of the effect in logistic model,
- increases the precision of effect estimator in linear model, but decreases the precision in logistic model.

Thus, one should always adjust for W in linear model and typically also in logistic model; the exception is if the effect size for the unadjusted logistic regression model is specifically needed for some downstream applications. (Such as, e.g., to combine results with other studies that have also used the unadjusted model.) For GWAS discovery, where we want to maximize power, one should always adjust for W also in logistic regression of population samples.

The above holds for the analysis of a **random sample from the population** where X and W are independent and do not have interaction effects. But typically in GWAS we use highly **ascertained case-control sampling** from a population and then, in our ascertained sample, X and W are not anymore independent and some less intuitive behavior occur.

6.5. Covariates in ascertained case-control GWAS

We want to study whether a SNP X is associated with a disease (or other binary trait) Y and we have available a covariate W, such as age or sex or carrier status of some well-known large effect mutation, which is independent of X **in the general population**. We collect S cases and R controls from the population. The proportion of cases in our sample is $\phi = S/(S+R)$ while the prevalence of the disease in general population is K. Typically, K is 1% or less whereas ϕ is in range of 20% to 50%. Thus, $K \ll \phi$ and the relationship between X and W will be different in our sample than it is in the population.

We will apply logistic regression and the question is whether we should use

$$\text{Model M:} \qquad Y \sim \mu + Z\gamma + X\beta + W\alpha,$$

or the simpler, unadjusted

$$\text{Model M':} \qquad Y \sim \mu' + Z\gamma' + X\beta',$$

where Z represents possible confounders, such as population structure, that we must always adjust for.

This issue has been studied by Pirinen, Donnelly, Spencer (2012), and here is a summary of the main points:

- When W is independent of X in the general population, then W is not a confounder oy X-Y association and therefore also model M’ is a valid model to test for X-Y association, and this is true also in an ascertained case-control study. This means that if the effect β of X is 0 in model M, then also the effect β' of X is 0 in model M’, and no spurious association will be created by ignoring W from the logistic regression model. Both models are valid and we have a choice to make.
- The effect sizes estimated by the two models are different: $|\beta'| \leq |\beta|$.
- The precisions of the two models are also different: $\text{SE}(\beta') \leq \text{SE}(\beta)$.
- Therefore it is not directly evident which model has more power since the above two properties favor the opposite models in terms of maximizing the non-centrality parameter $(\beta/\text{SE})^2$.
- It can be shown that which model is more powerful to detect nonzero genetic effect of X depends (mainly) on the prevalence K of the disease in the population. In typical cases, where $K < 1\%$, the model M’ is often more powerful than M. The opposite is true only for prevalent diseases, say, with $K > 5\%$. When $\phi = K$, i.e., when we are considering a population sample, then model M is always more powerful than model M’, but for low prevalence diseases this difference in favor of M is small.
- There are real examples (e.g. a GWAS on Ankylosing Spondylitis) where an application of model M would have led to over 50% reduction in the non-centrality parameter compared to model M’. (See example below and slides 17) Such situations can come up with relatively rare risk factors with very strong effects on the disease (such as an HLA-B27 allele in Ank. Spond.).

To evaluate more specifically the parameters and power of each model, we can use R-functions from: https://www.mv.helsinki.fi/home/mjxpirin/log_regression_covariate_functions.R. For binary covariates, such as sex or carrier status of an HLA-allele, we use the function

```
binary.covariate(K, freq.G, or.G, freq.X, or.X, ncases, ncontrols, population.controls = FALSE)
#INPUT
#K, the (target) prevalence of the disease
#freq.G, frequency of risk allele in general population
#or.G, odds-ratio for each copy of the risk allele
#freq.X, frequency of the risk factor of binary exposure
#or.X, odds-ratio of the risk factor
#ncases, number of cases in the case-control sample
#ncontrols, number of controls in the case-control sample
#population.controls, if TRUE then controls have general population frequencies,
                        otherwise controls have proper control frequencies.
```

In the parameters of this function G refers to the tested genetic variant (that we have called X in this document) and X refers to the covariate (that we have called W).

```
source("https://www.mv.helsinki.fi/home/mjxpirin/log_regression_covariate_functions.R")
#Ankylosing Spondylitis with HLA-B27 as binary covariate X
K = 0.0025
freq.X = 0.08 #carriers of HLA-B27 in population
or.X = 49 #OR of B27 on AS risk
binary.covariate(K, freq.G = 0.3, or.G = 1.2, freq.X, or.X,
                 ncases = 2000, ncontrols = 2000, population.controls = FALSE)
```

```
## $K
## [1] 0.002499752
##
## $ncases
## [1] 2000
##
## $ncontrols
## [1] 2000
##
## $pop.a
## [1] -7.664
##
## $case.freq
##            [,1]       [,2]       [,3]
## [1,] 0.08461555 0.08702497 0.02237533
## [2,] 0.35259474 0.36104528 0.09234413
##
## $control.freq
##            [,1]       [,2]        [,3]
## [1,] 0.45171766 0.38715024 0.082951425
## [2,] 0.03841463 0.03277942 0.006986627
##
## $population.controls
## [1] FALSE
##
## $marg.or.G
## [1] 1.195934
##
## $joint.or.G
## [1] 1.2
##
## $marg.or.X
## [1] 48.98191
##
## $joint.or.X
## [1] 49
##
## $marg.se.G
## [1] 0.04792256
##
## $joint.se.G
## [1] 0.0707024
##
```

```
## $NCP.GX.G
## [1] 0.4770184
```

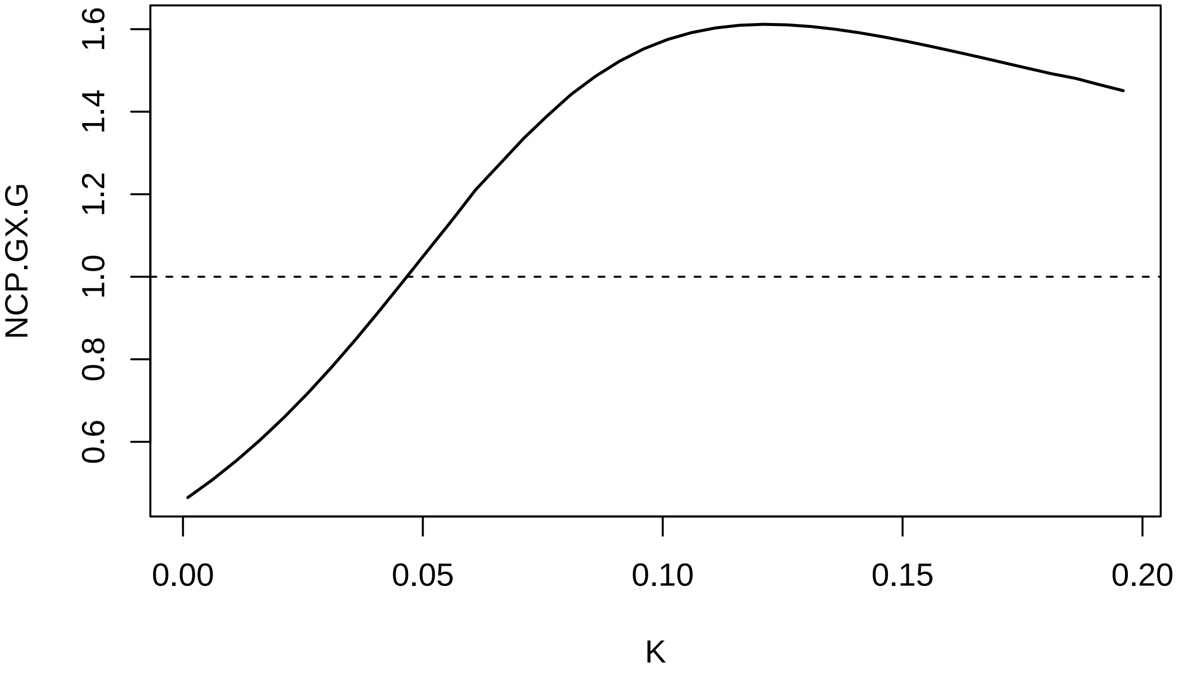
The output tells what are the case or control distribution of genotypes (3 cols) and risk factor (1st row is for non-risk group and 2nd row is for risk group). `marg.or.G` is the marginal odds-ratio for each copy of the risk allele (i.e. effect of G omitting X from the model) and `joint.or.G` is the odds-ratio for each copy of risk allele when adjusted for covariate X.

`marg.or.X` is the marginal odds-ratio for risk factor X = 1 (i.e. effect of risk factor omitting genotype from the model). `joint.or.X` is the marginal odds-ratio for risk factor X = 1 when model has adjusted for genotype G. `marg.se.G` is SE of `marg.or.G` and `joint.se.G` is SE of `joint.or.G`. `NCP.GX.G` is the ratio of the non-centrality parameters between models M (both G and X) and M' (only G not X).

We see that, with these parameters, NCP drops over 50% if GWAS of Ankylosing Spondylitis is adjusted for the HLA-B27 carrier status. Powerwise it corresponds to discarding over half of the valuable samples! More generally, HLA-alleles can have large effects on autoimmune diseases, and adjusting these analyses for HLA-allele should not be the main GWAS strategy because of power loss. On the other hand, these large HLA-effects have a potential to lead to identification of interaction effects between HLA and other genetic variants and such interaction analyses should be pursued after the basic GWAS. For example, both in Ankylosing Spondylitis and in Psoriasis, there is an **interaction effect** between HLA alleles on chr 6 and *ERAP1* locus on chr 5 (slide 18), where *ERAP1* variants only affect disease risk for individuals carrying certain HLA-alleles but not at all for individuals without these HLA-alleles.

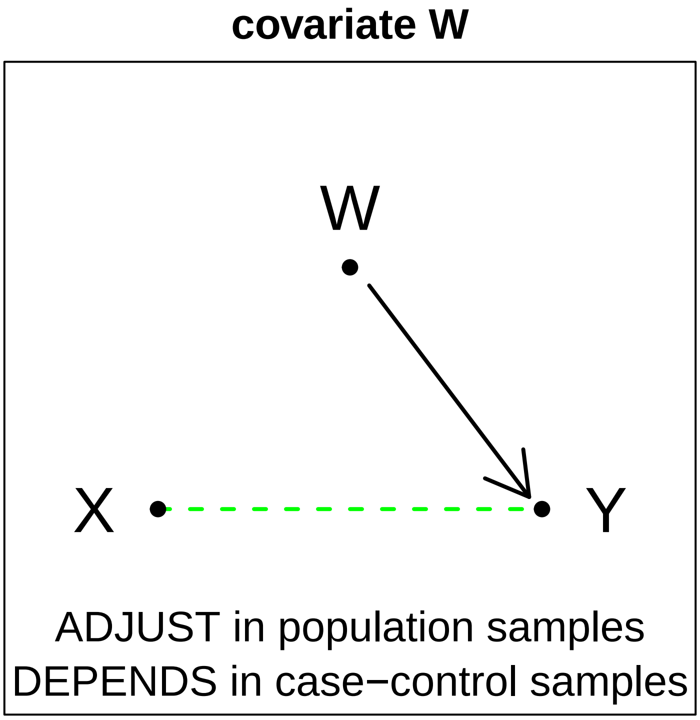
Let's see how the NCP ratio between models M and M' depends on the prevalence of the disease while other disease parameters than prevalence are kept fixed to their values in the Ankylosing Spondylitis example. Let's apply the function to a grid of prevalence values from 0.1% to 20% and plot the results:

```
Ks = seq(0.001, 0.2, 0.005)
res = sapply(Ks,function(K){binary.covariate(K, freq.G = 0.3, or.G = 1.2, freq.X = 0.08,
                                             or.X = 49, ncases = 2000, ncontrols = 2000,
                                             population.controls = FALSE)$NCP.GX.G})
plot(Ks, res, type = "l", xlab = "K", ylab = "NCP.GX.G", lwd = 1.5)
abline(h = 1, lty = 2)
```



This shows that model M' has more power than model M for prevalences below 5%, after which model M starts to have more power. This is a general pattern of covariate adjustment in ascertained case-control studies, but the exact changepoint depends on the parameters of the disease model.

Summary of covariate W independent of genetic variant X in population


covariate W
W
X
Y
ADJUST in population samples
DEPENDS in case−control samples