Data simulation experiments

BM

23/10/2020

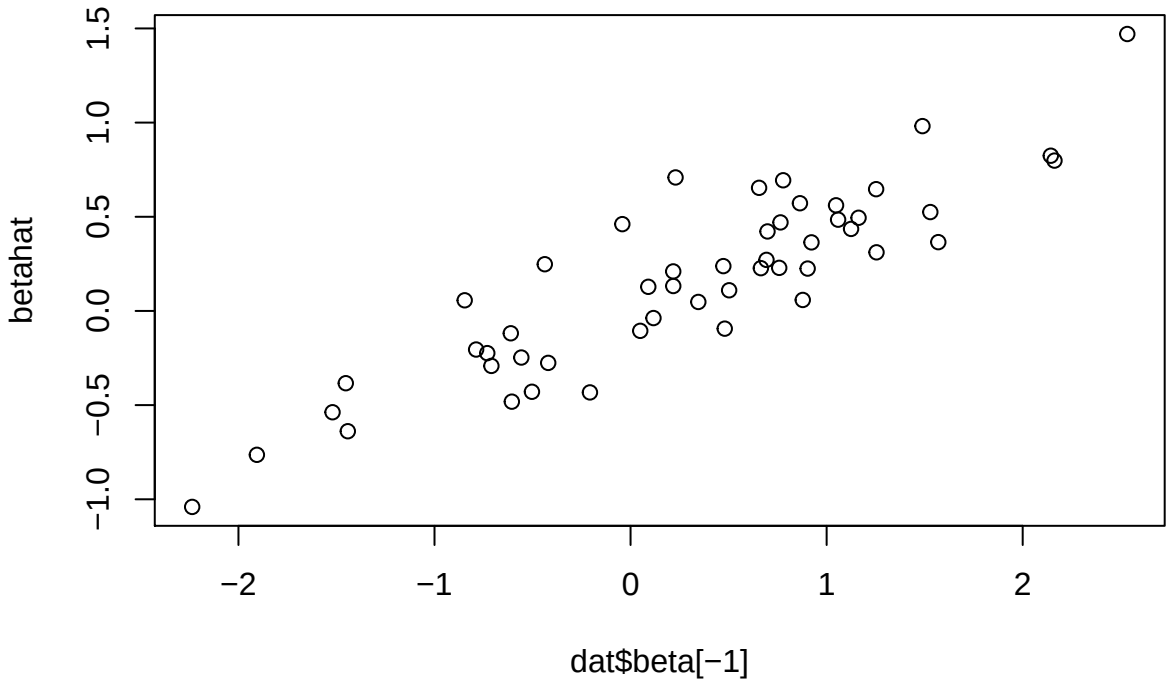
Case control simulation 1

1. Sample covariates on a population,
2. sample covariate effects from log-F prior,
3. calculate linear predictor of the logistic regression using all covariate effects,
4. simulate disease status, and
5. sample cases and controls from population.

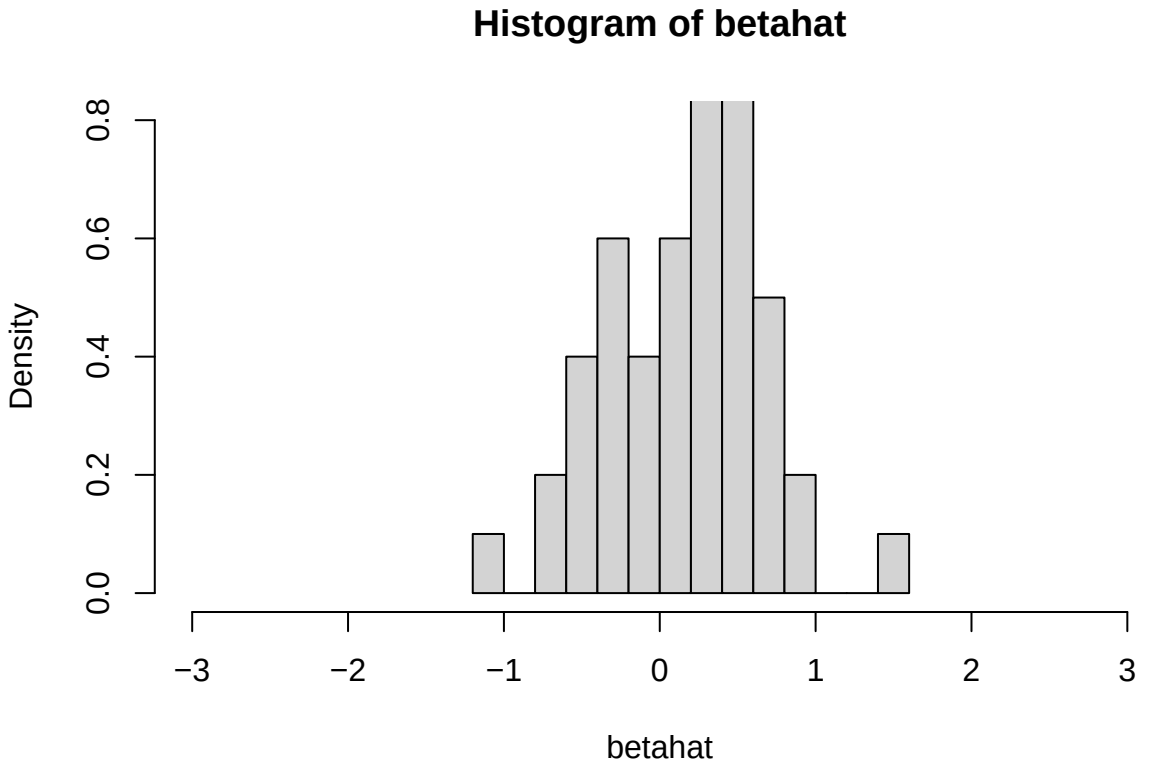
```
ncase <- ncon <- 100
K <- 50 # num SNPs
m <- 4 # log-F parameter
N <- 20000 # popn size
beta0 <- (-4) # intercept

ccsim1 <- function(MAF) {
  X <- cbind(1,matrix(rbinom(N*K,size=2,p=MAF),ncol=K))
  beta <- c(beta0,log(rf(K,m,m)))
  linpred <- X %*% beta
  p <- exp(linpred)/(1+exp(linpred))
  cc <- rbinom(N,size=1,prob=p)
  caseind <- sample((1:N)[cc==1],size=ncase,replace=FALSE)
  conind <- sample((1:N)[cc==0],size=ncon,replace=FALSE)
  list(cc=c(rep(1,ncase),rep(0,ncase)),X = X[c(caseind,conind),],
       beta = beta)
}
```

- Start with simulations for MAF = 0.5

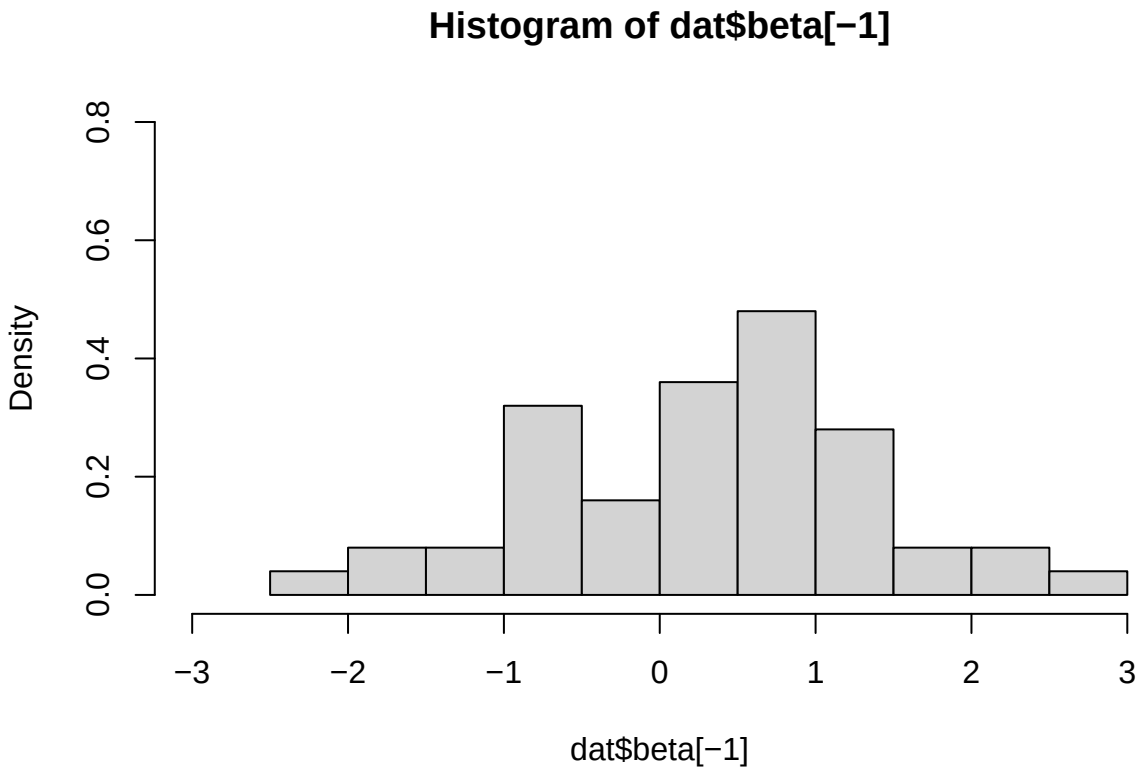
```
set.seed(123)
dat <- ccsim1(MAF=.5)
betahat <- rep(NA,K)
for(i in 1:K) {
  XX <- dat$X[,1+i]
  betahat[i] <- coef(glm(dat$cc~XX,family=binomial()))[[2]]
}
plot(dat$beta[-1],betahat) # correlated, but betahats attenuated
```




```
hist(betahat,freq=FALSE,nclass=10,ylim=c(0,.8),xlim=c(-3,3))
```

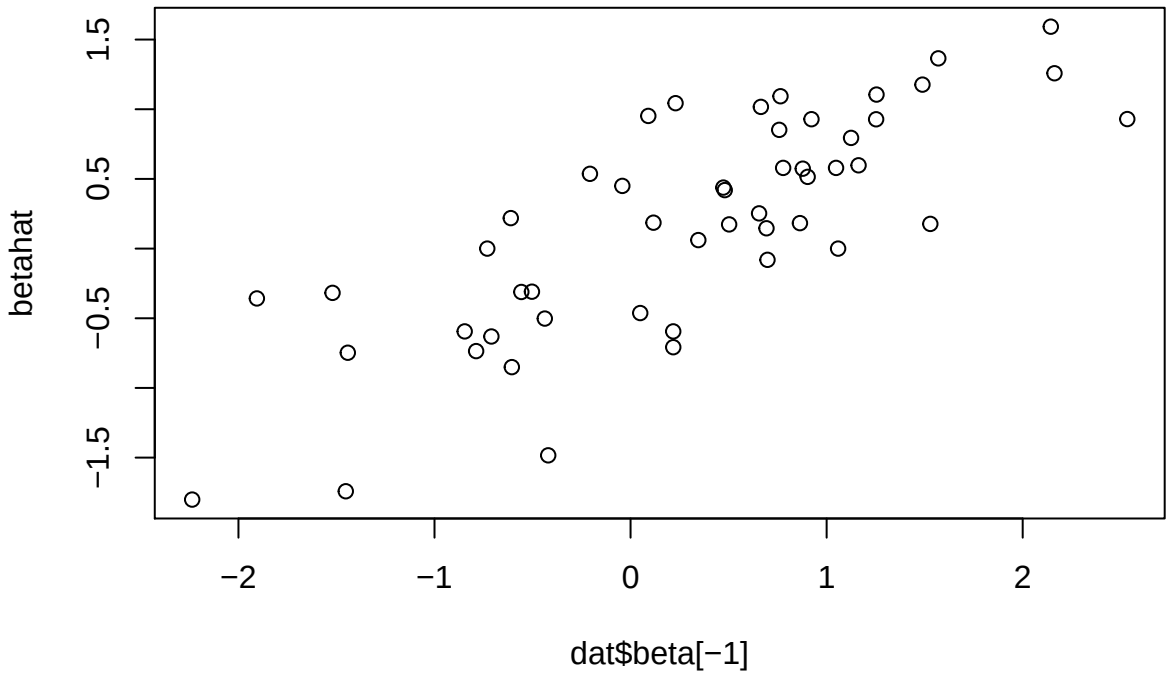



```
hist(dat$beta[-1],freq=FALSE,nclass=10,ylim=c(0,.8),xlim=c(-3,3))
```

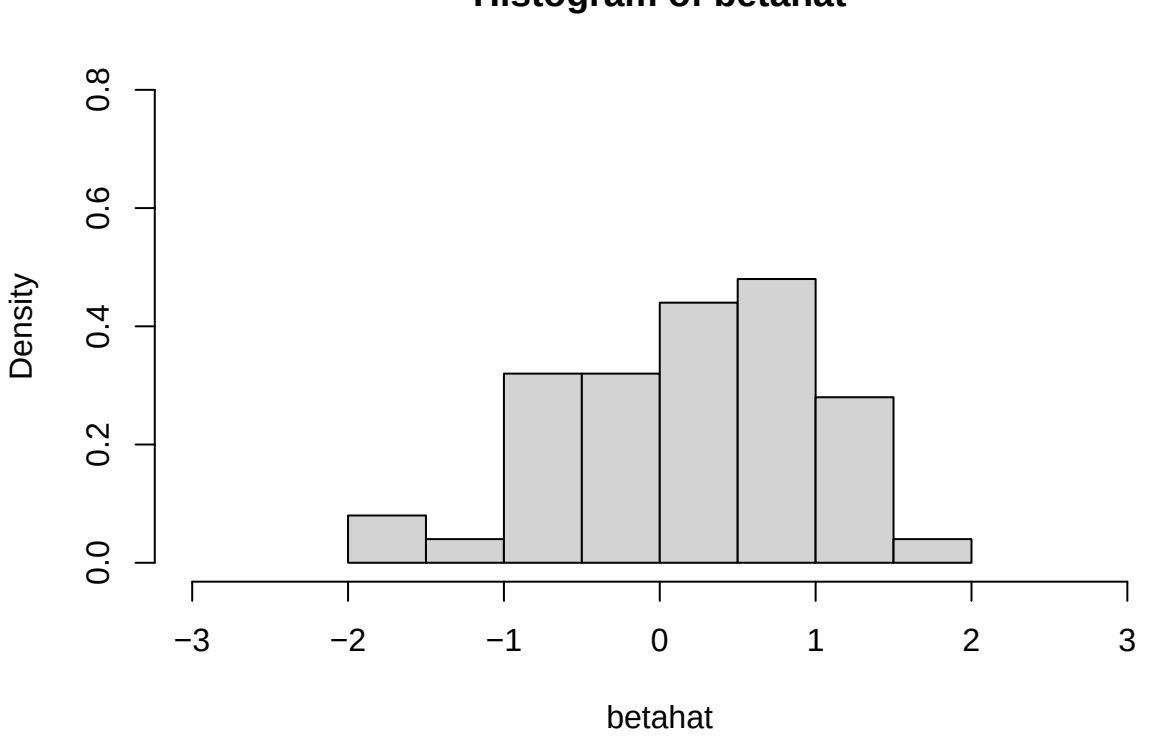


- $\hat{\beta}$'s are correlated with the β's but are attenuated. I guess this is why DY and SC's methods are suggesting that m is very large (variance of log-F small).
- Next try MAF $= 0.5$

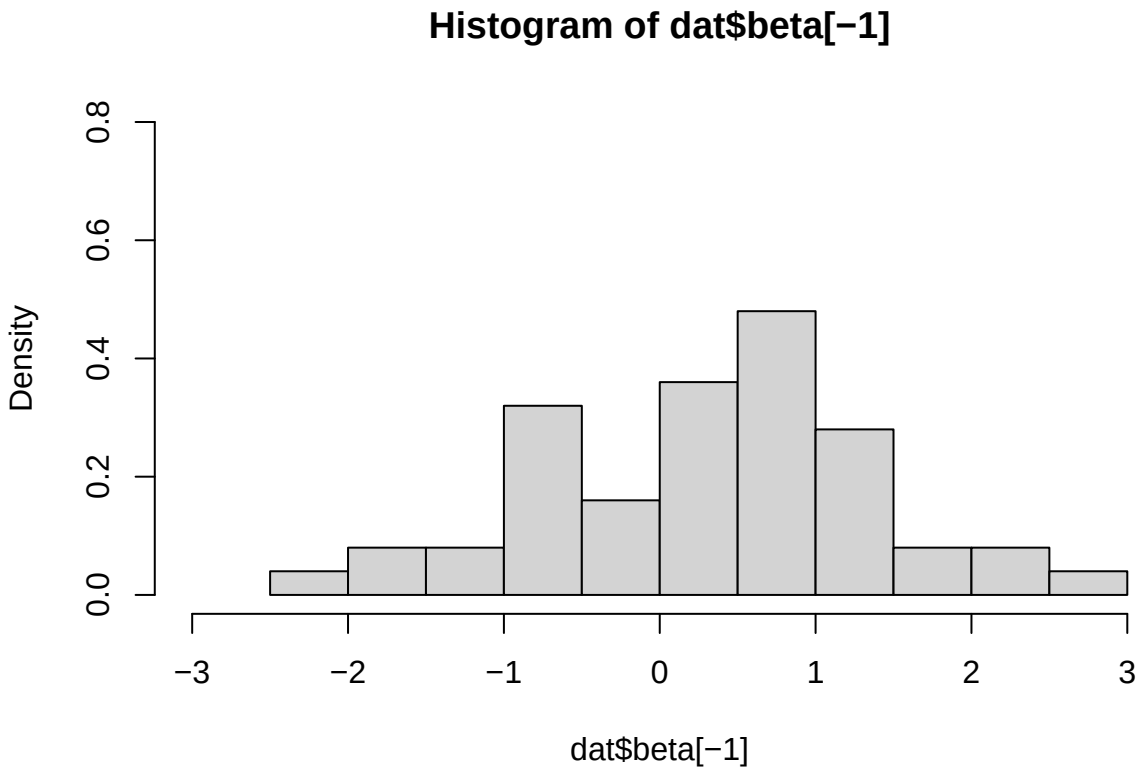
```
set.seed(123)
dat <- ccsim1(MAF=.05)
betahat <- rep(NA,K)
for(i in 1:K) {
  XX <- dat$X[,1+i]
  betahat[i] <- coef(glm(dat$cc~XX,family=binomial()))[[2]]
}
plot(dat$beta[-1],betahat) # betahats still attenuated, but less so
```




```
hist(betahat,freq=FALSE,nclass=10,ylim=c(0,.8),xlim=c(-3,3))
```




```
hist(dat$beta[-1],freq=FALSE,nclass=10,ylim=c(0,.8),xlim=c(-3,3))
```

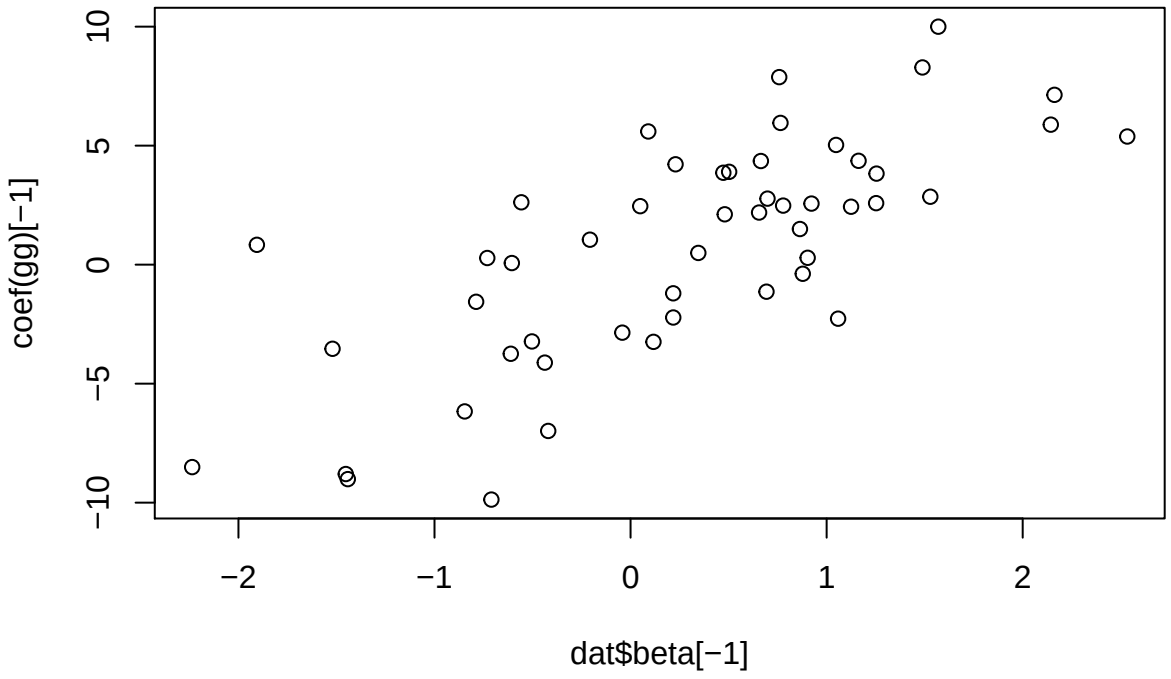


- $\hat{\beta}$'s are less attenuated, but still too small in magnitude.
- Consistent with the fact that DY and SC's likelihoods for m sometimes have a max.
- Can we fit a joint model in X's, like the one used to simulate the data? No, it fails with current sample size.

```
XX <- dat$X[,-1]
colnames(XX)<- paste0("X",1:K)
dd <- data.frame(cc=dat$cc,XX)
gg <- glm(cc ~ ., data=dd, family=binomial())
head(round(cbind(dat$beta,coef(gg),c(1,betahat)),3)) # too unstable
```

```
##                [,1]   [,2]   [,3]
## (Intercept) -4.000 -7.541  1.000
## X1           2.143  5.882  1.592
## X2           1.163  4.362  0.598
## X3          -0.557  2.620 -0.311
## X4           1.529  2.853  0.177
## X5           0.693 -1.136  0.146
```

```
plot(dat$beta[-1],coef(gg)[-1])
```



Case control simulation 2

- Use the Qin and Zhang method.
- If $g(x)$ is the covariate distribution in controls, the distribution in cases is proportional to $g(x)\exp(x\beta)$.
- If $g(x)$ is $binomial(2,p)$, where p is the MAF (under a rare disease, the distribution in controls is about that in the population), then

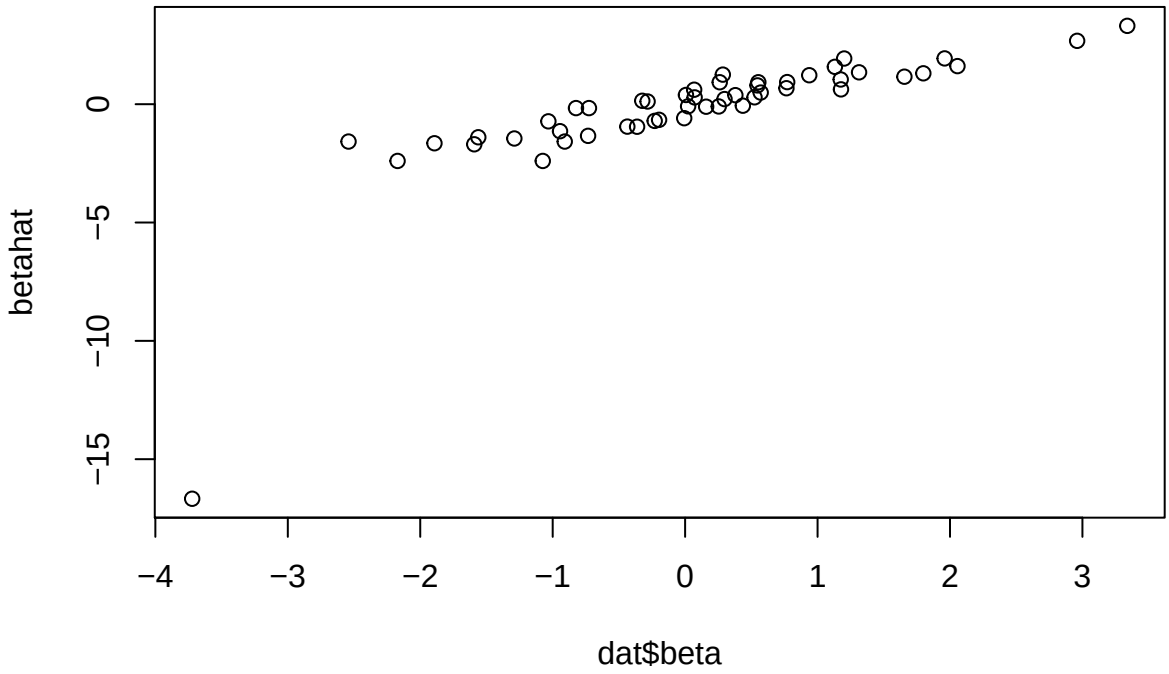
$$g(x)\exp(x\beta) = \begin{cases} (1-p)^2 & x=0 \\ 2p(1-p)\exp(\beta) & x=1 \\ p^2\exp(2\beta) & x=2 \end{cases},$$

which has normalizing constant $(1-p)^2 + 2p(1-p)\exp(\beta) + p^2\exp(2\beta)$.

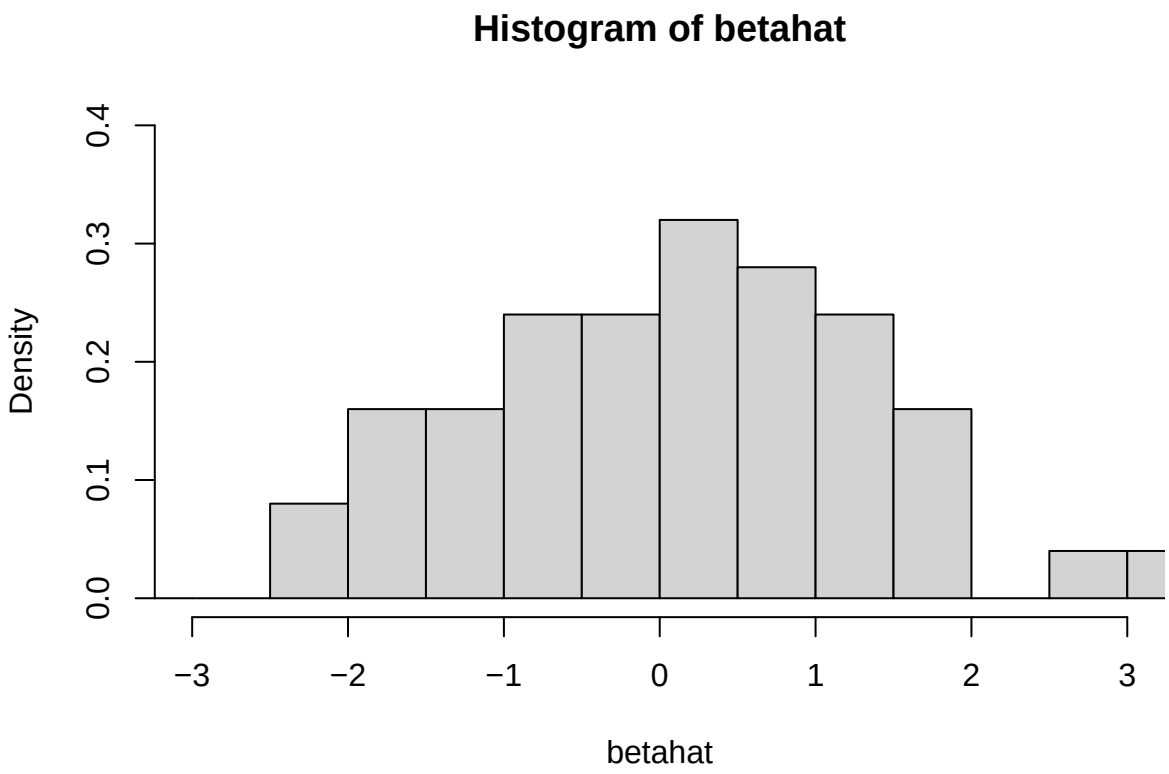
```
rcase <- function(n,beta,MAF) {
  p0 <- (1-MAF)^2
  p1 <- 2*MAF*(1-MAF)*exp(beta)
  p2 <- MAF^2*exp(2*beta)
  pp <- c(p0,p1,p2)
  sample(0:2,size=n,replace=TRUE,prob=pp)
}

ccsim2 <- function(MAF) {
  Xcon <- matrix(rbinom(ncon*K,size=2,prob=MAF),ncol=K)
  beta <- log(rf(K,m,m))
  Xcase <- matrix(NA,ncol=K,nrow=ncase)
  for(i in 1:K) {
    Xcase[,i] <- rcase(ncase,beta[i],MAF)
  }
  X <- rbind(Xcase,Xcon)
  list(cc=c(rep(1,ncase),rep(0,ncase)),X = X,beta = beta)
}
```

- Simulate covariates with MAF = 0.05 and fit single-SNP models to see what the coefficients look like.
  - We see one estimated coefficient of about −15, which I guess means non-convergence or monotone likelihood for that covariate.

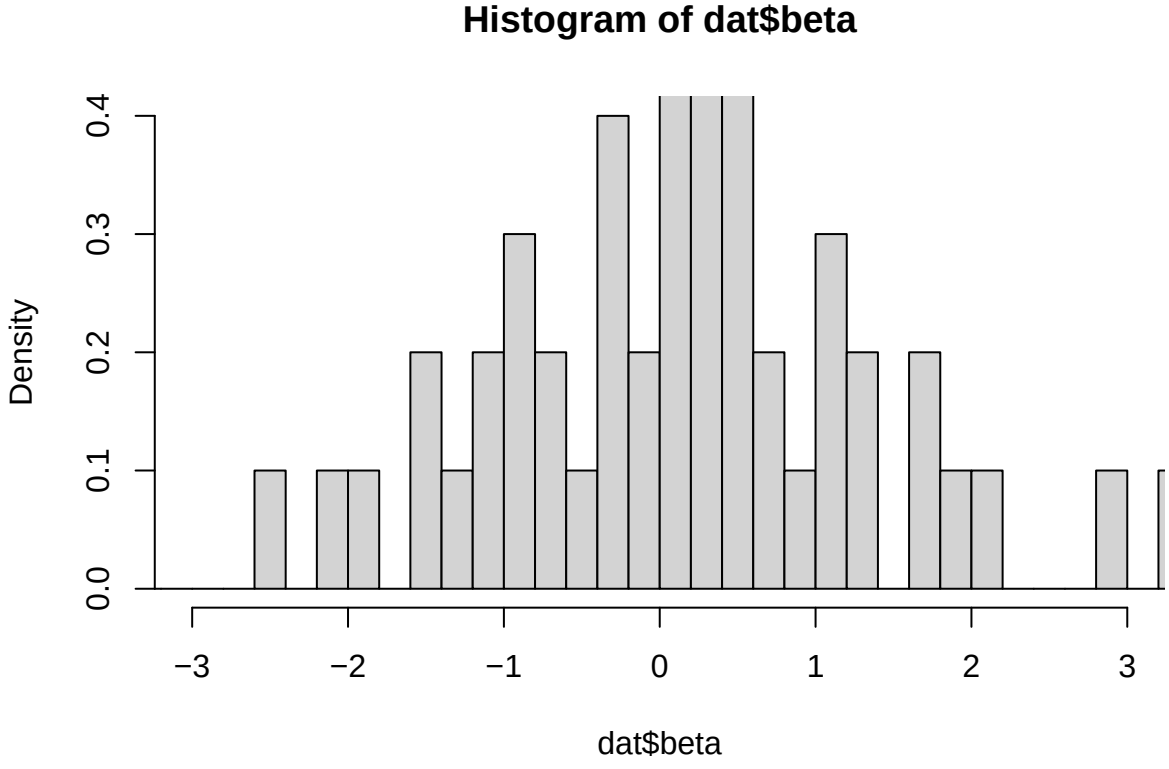
```
set.seed(123)
dat <- ccsim2(MAF=.05)
betahat <- rep(NA,K)
for(i in 1:K) {
  XX <- dat$X[,i]
  betahat[i] <- coef(glm(dat$cc~XX,family=binomial()))[[2]]
}
plot(dat$beta,betahat)
```




```
hist(betahat,freq=FALSE,nclass=30,ylim=c(0,.4),xlim=c(-3,3))
```



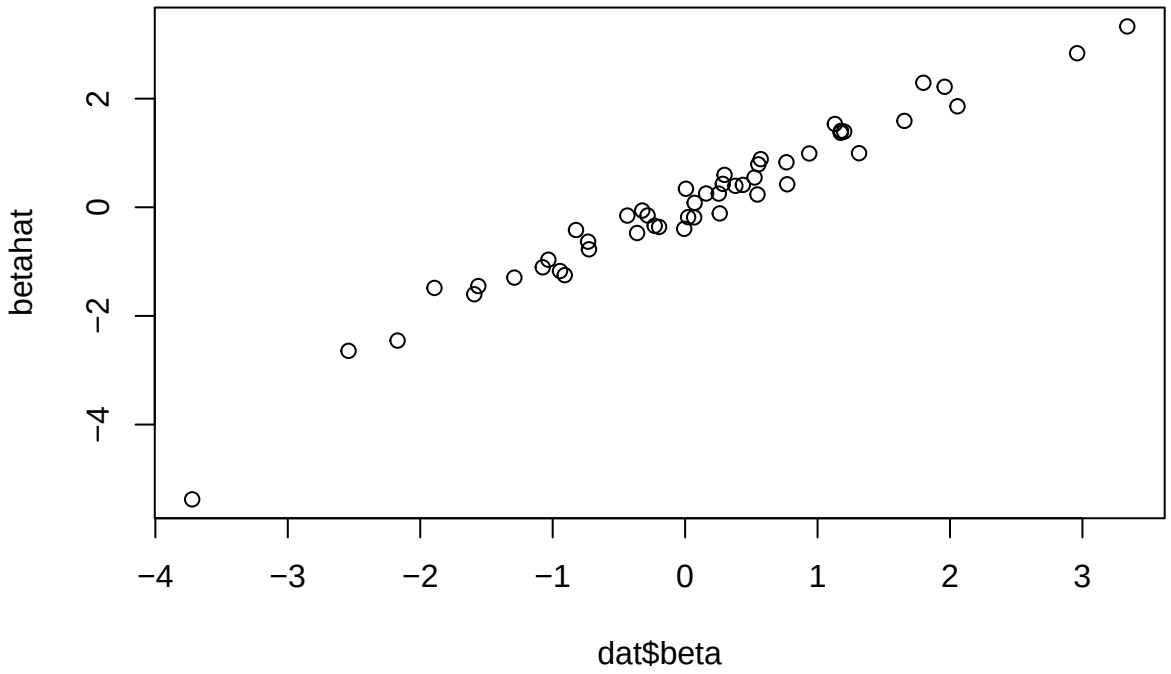

```
hist(dat$beta,freq=FALSE,nclass=30,ylim=c(0,.4),xlim=c(-3,3))
```



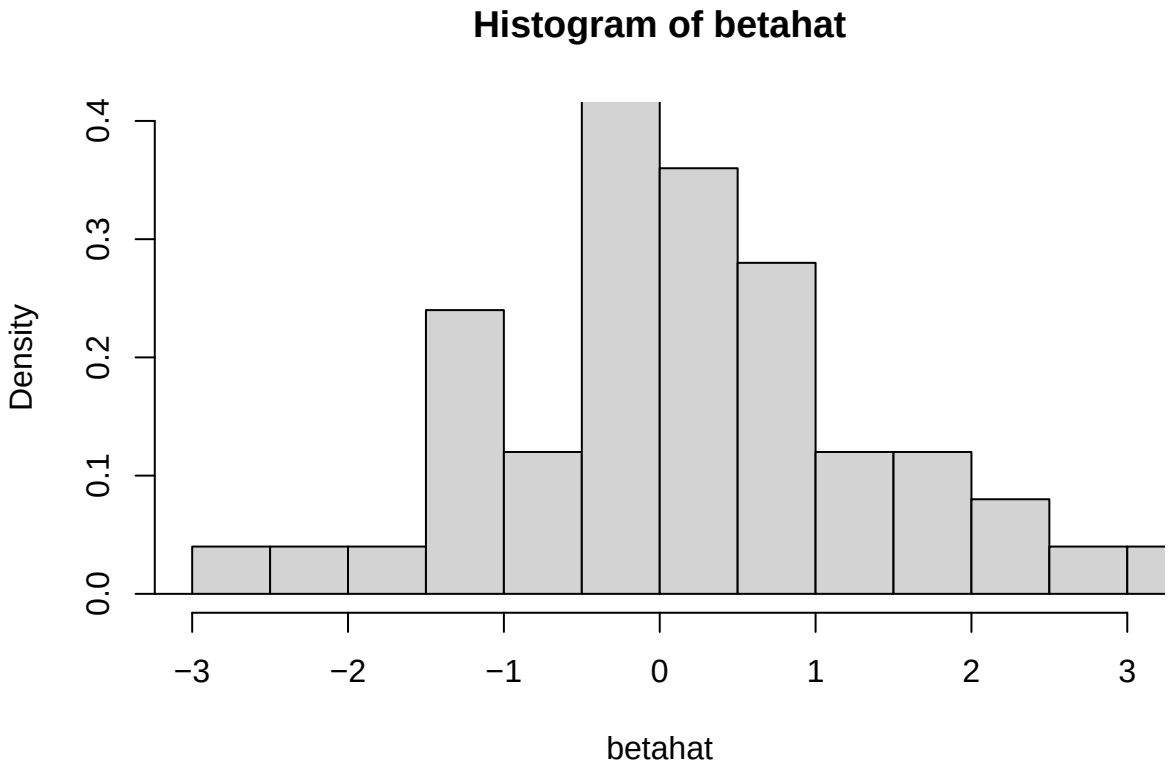
- Repeat with MAF $= 0.5$

```
set.seed(123)
dat <- ccsim2(MAF=.5)
```

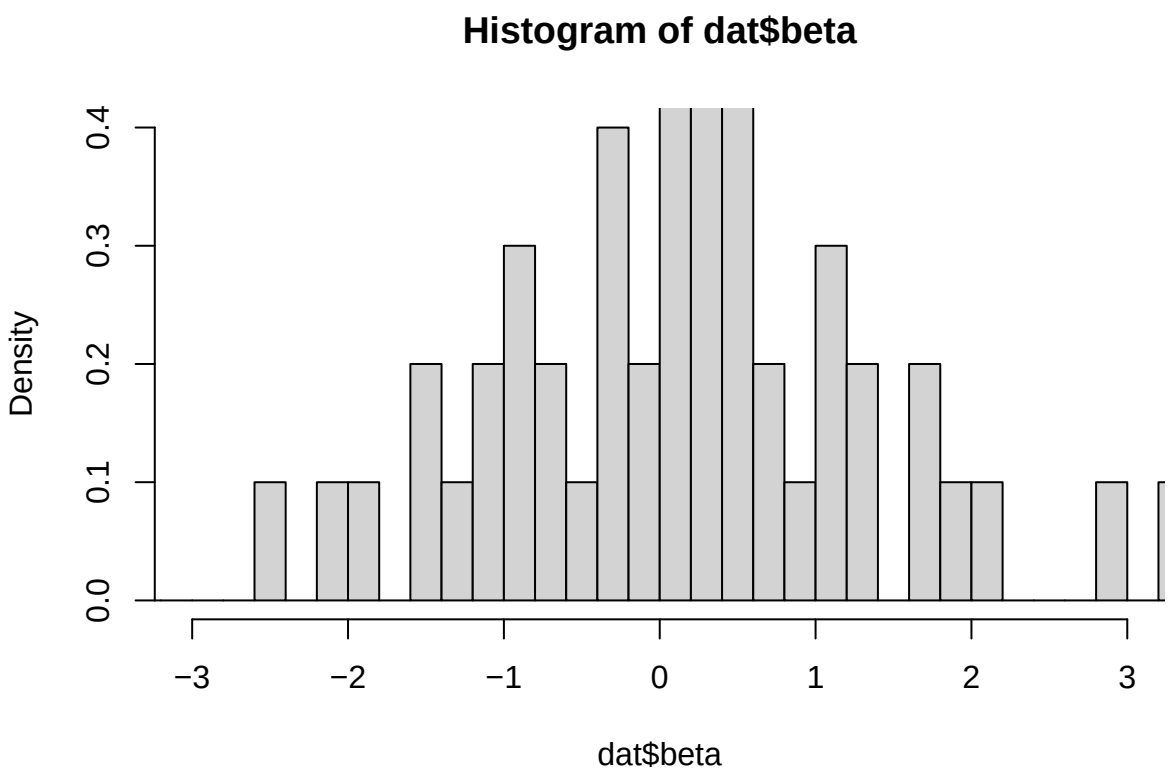
```
betahat <- rep(NA,K)
for(i in 1:K) {
  XX <- dat$X[,i]
  betahat[i] <- coef(glm(dat$cc~XX,family=binomial()))[[2]]
}
plot(dat$beta,betahat)
```




```
hist(betahat,freq=FALSE,nclass=30,ylim=c(0,.4),xlim=c(-3,3))
```



```r
hist(dat$beta,freq=FALSE,nclass=30,ylim=c(0,.4),xlim=c(-3,3))
```



Q&Z with population as a confounder

- Simulate under population stratification with
  1. population-SNV correlation from different SNV MAFs in different populations
  2. population-disease correlation by a population main effect on disease risk
- Simulate population status first, and then SNVs conditional on population status.

Simulate population status

- Let D=0 or 1 denote disease status, with 0 for controls and 1 for cases.
- Let sub-population S take values 0 or 1 and say their frequency is f_0 and f_1, respectively in the general population. We take these frequencies to be the same among controls (rare disease assumption).
- Start by simulating population with log-OR β_s using the Q&Z method.
- The distributon of S in controls, $P(S = s|D = 0)$, is $bernoulli(f_1)$, and the distribution in cases, $P(S = s|D = 1)$ is proportional to

$$f_s \exp(s\beta_s) = \begin{cases} f_0 & s = 0 \\ f_1 \exp(\beta_p) & s = 1 \end{cases}$$

- Use the above distributions to simulate S in controls and cases.

Simulate SNVs conditional on population status

- We specify the distribution of the SNV in controls and then work out its distribution in cases.

  - We need to start with the joint distribution of the SNV and sub-population, and then find the relevant conditional distributions.
- For a given SNV X, its distribution in controls differs by sub-population, with MAF p_0 in sub-population 0 and MAF p_1 in sub-population 1.
- Let $g_s(x)$ denote distribution in population s, i.e., $P(X = x|S = s, D = 0) = g_s(x)$.
- The joint distribution of X and S in controls is $P(X = x, S = s|D = 0) = f_s g_s(x)$.
- Let β_x denote the log-OR for the SNV. I'm going to assume a joint disease risk model with log-OR $= s\beta_s + x\beta_x$.
- Then the joint distribution of X and S in cases is proportional to

$$f_s g_s(x) \exp(s\beta_s + x\beta) = \begin{cases} f_s \exp(s\beta_s)(1-p_s)^2 & x = 0 \\ f_s \exp(s\beta_s) 2p_s(1-p_s)\exp(\beta) & x = 1 \\ f_s \exp(s\beta_s) p_s^2 \exp(2\beta) & x = 2 \end{cases},$$

- From here you can show (but check me on this) that

$$P(S = s|D = 1) \propto f_s \exp(s\beta_s) \sum_{x'} g_s(x') \exp(x'\beta_x).$$

- Notice that this is **different** from the distribution used to simulate S in the previous section, which was proportional to $f_s \exp(s\beta_s)$.
  - The two distributions are the same only when $\beta_x = 0$.
  - A large β_x would tilt the distribution of S in favour of the sub-population with the larger MAF, I think.
  - But, to maintain the idea of a dataset with multiple markers on a sample of cases and controls, I propose that we stick with $P(S = s|D = 1) \propto f_s \exp(s\beta_s)$. Let me know what you think.
- We then have

$$P(X = x|S = s, D = 1) = \frac{P(X = x, S = s|D = 1)}{P(S = s|D = 1)} = \frac{g_s(x)\exp(x\beta_x)}{\sum_{x'} g_s(x')\exp(x'\beta_x)} \propto g_s(x)\exp(x\beta_x).$$