Simulated GenProbe Data

Marshall Brown

October 26, 2015

Simulating data with known correlation structure.

I use the method described here:

http://www.r-bloggers.com/easily-generate-correlated-variables-from-any-distribution-without-copulas/

Assuming that the (scaled) continuous variables are approximately normal, we:

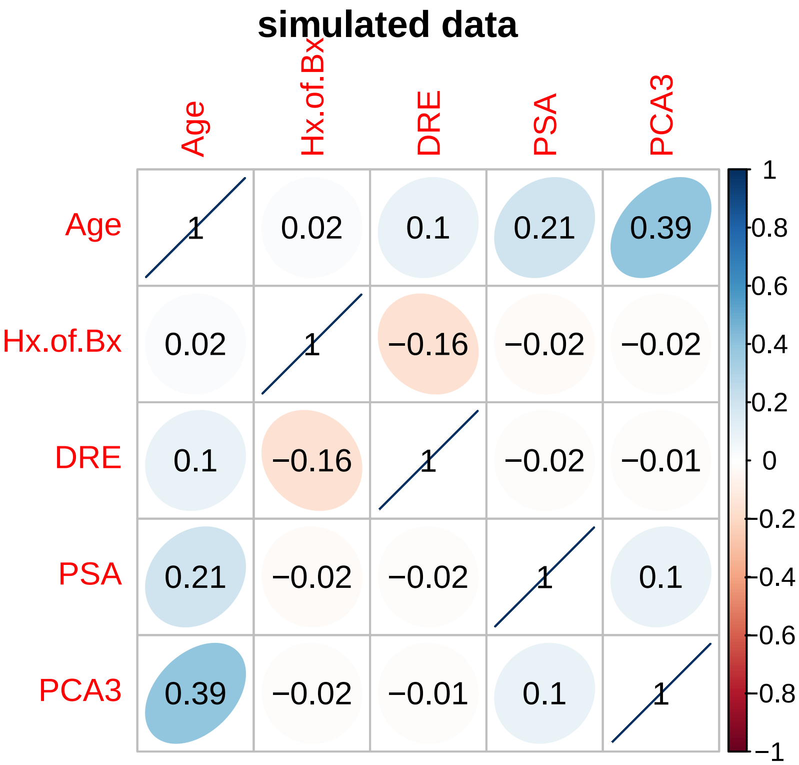
1. Draw variables from a joint normal distribution using the observed mean/covariance matrix estimated from the GenProbe data.

- Simulated Age, log(PSA), and log(PCA3) values are directly simulated from this multivariate normal distribution.

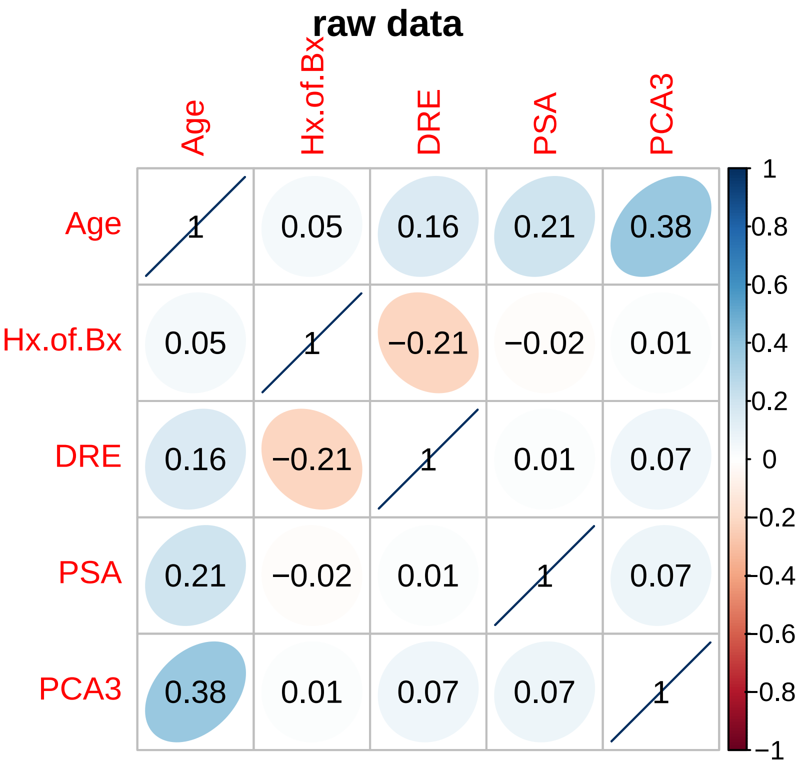
2. Apply the univariate normal CDF of variables to derive probabilities for binary variables 'suspicious DRE' and 'history of biopsy.'
3. Apply the inverse binomial CDF with $p = \hat{p}$ estimated from the raw data to simulate binary variables.

- This transformation reduces the amount of correlation among variables, but for this example the simulated variables follow pretty close to the observed correlation structure (see below).

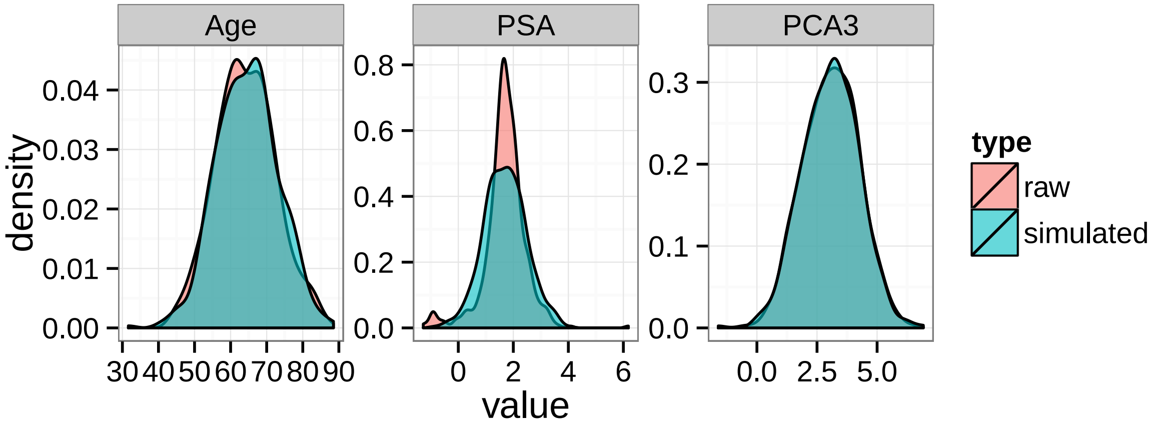
Nex I show the correlation and marginal distributions in the raw vs. simulated data.

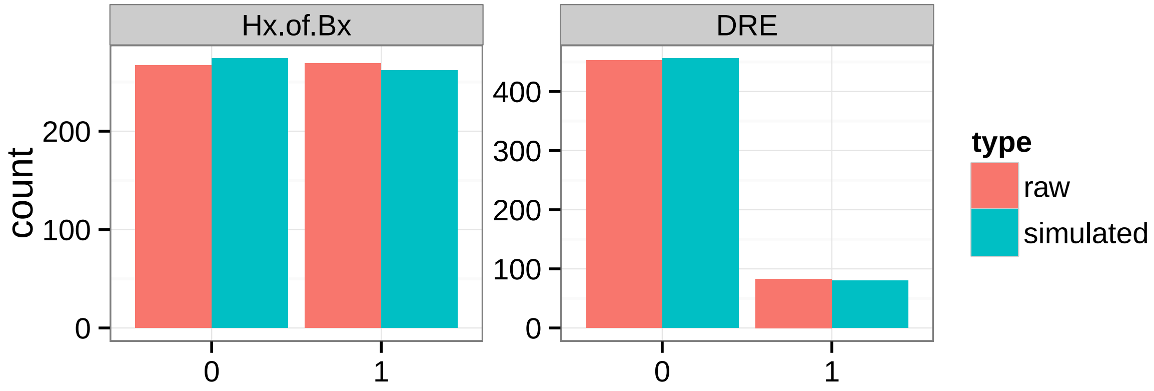


simulated data



raw data



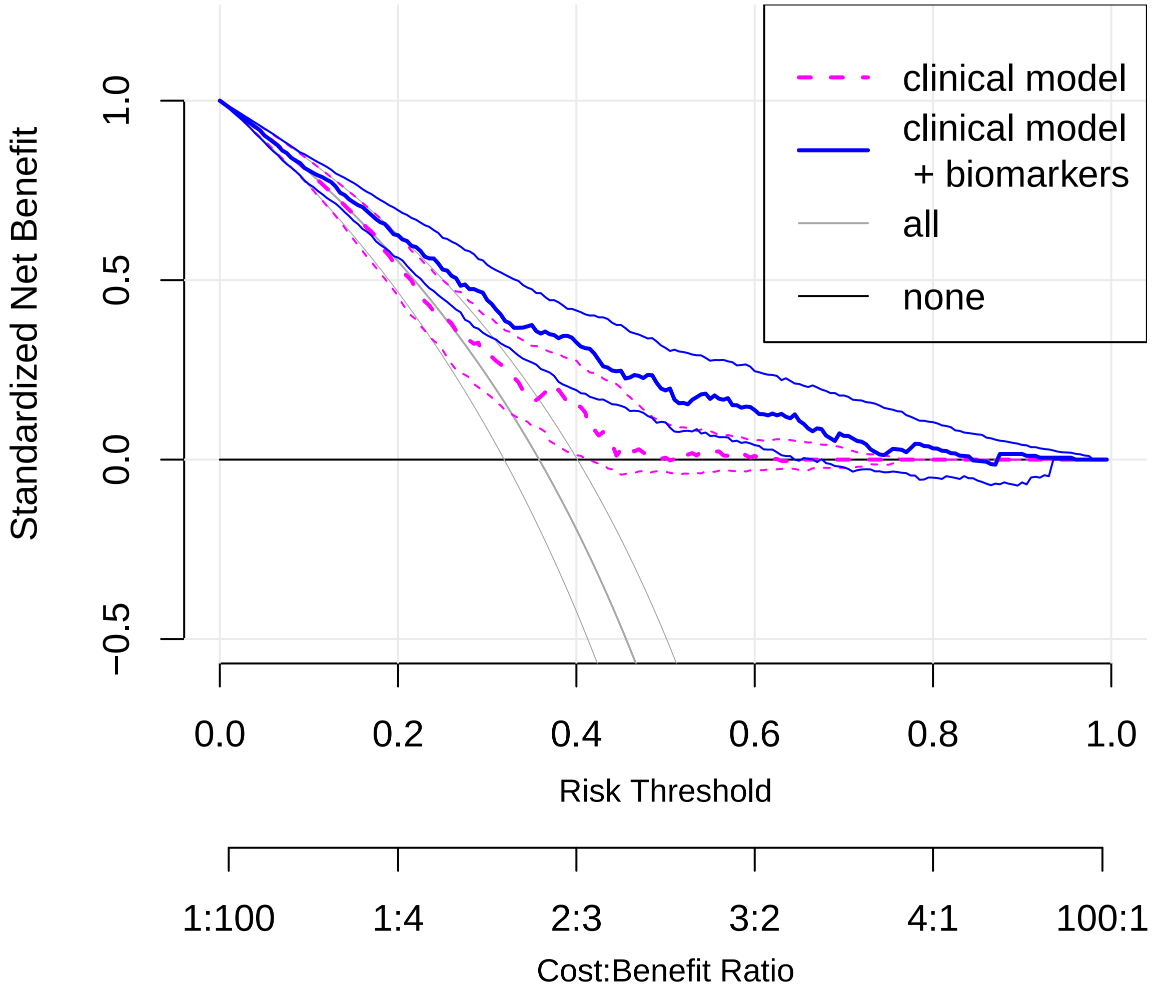


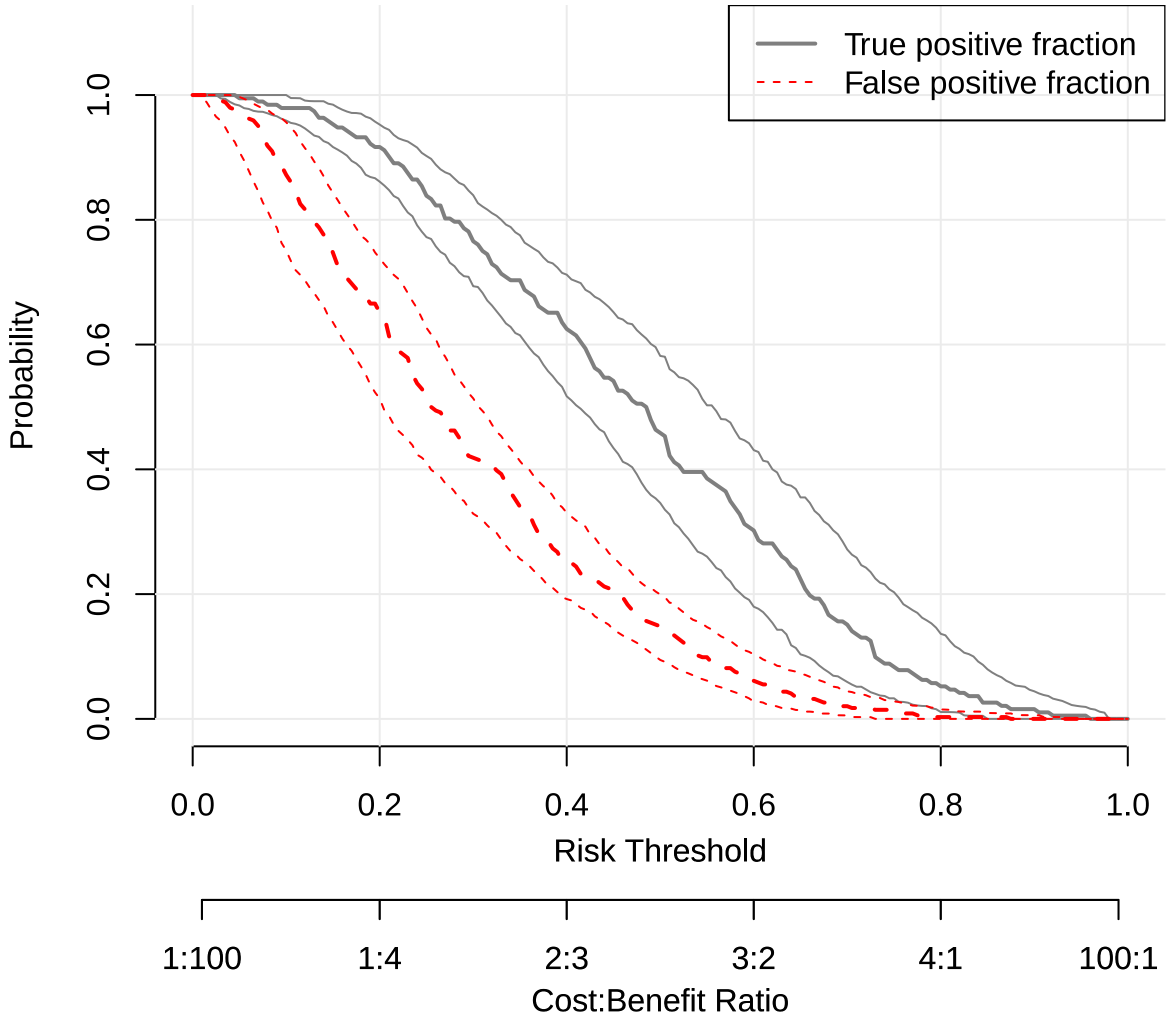
Simulating outcome

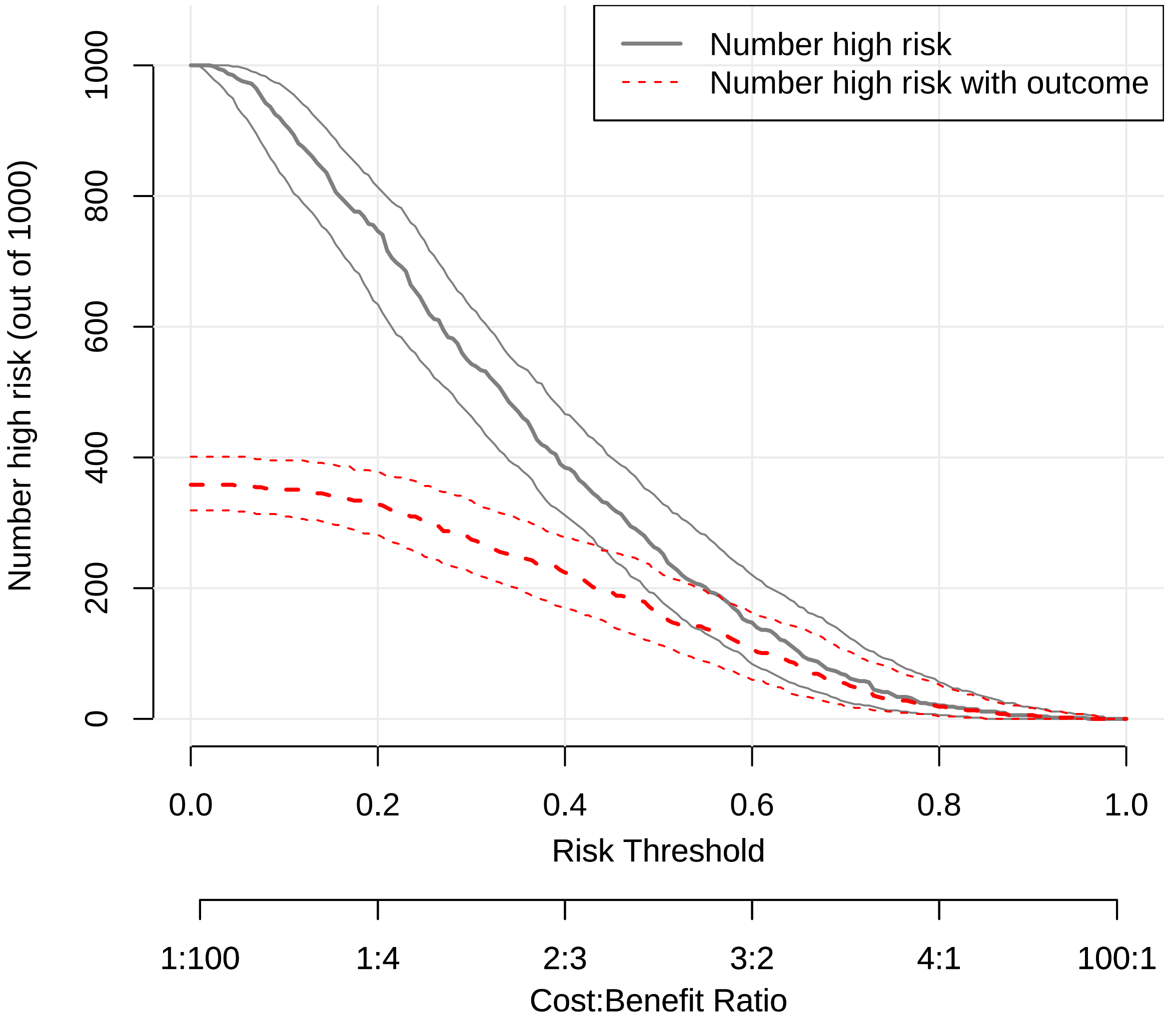
First I fit a logistic regression model using the raw data with age, DRE, history of biopsy, PSA and PCA3. I then simulate cancer outcome from a logistic model using the same coefficients from the model fit on the raw data.

```
##           (Intercept)          Age   Hx.of.Bx       DRE       PSA      PCA3
## obs.coef    -2.748036 -0.005166188 -0.7954109 0.5796367 0.4158020 0.6464313
## new.coef    -2.691541 -0.020114025 -0.4787605 0.7706856 0.5347999 0.8101323
```

Create plots for manuscript




True positive fraction
False positive fraction
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Risk Threshold
1:100
1:4
2:3
3:2
4:1
100:1
Cost:Benefit Ratio


Number high risk
Number high risk with outcome
Number high risk (out of 1000)
0
200
400
600
800
1000
0.0
0.2
0.4
0.6
0.8
1.0
Risk Threshold
1:100
1:4
2:3
3:2
4:1
100:1
Cost:Benefit Ratio