Debug SynSurr

Height

```
# read in the data
pheno <- readRDS("Data/Old/height_imputed.rds")

# read in the genetic data
G <- BEDMatrix::BEDMatrix(path = "Data/allchromosome.bed", simple_names = TRUE)

# a random SNP
i <- sample(1:ncol(G), size = 1)
print(i)
```

```
## [1] 125328
```

```
g <- as.numeric(G[as.character(pheno$f.eid), i]) # snp i
length(g)
```

```
## [1] 349474
```

```
g_complete <- g[!is.na(g)]
length(g_complete)
```

```
## [1] 348737
```

```
X.cov <- cbind(
  g_complete,
  (pheno %>%
    select(
      f.21022.0.0, f.22001.0.0,
      starts_with("PC")
    ))[!is.na(g), ]
)
dim(X.cov)
```

```
## [1] 348737     13
```

```
colnames(X.cov)
```

```
##  [1] "g_complete"  "f.21022.0.0" "f.22001.0.0" "PC1"         "PC2"
##  [6] "PC3"         "PC4"         "PC5"         "PC6"         "PC7"
## [11] "PC8"         "PC9"         "PC10"
```

```
X.cov <- scale(X.cov)
X.cov <- cbind(X.cov, rep(1, nrow(X.cov))) # append intercept
X.cov[, 1] <- g_complete # dont scale G
X.cov[, 3] <- pheno$f.22001.0.0[!is.na(g)] # dont scale sex


# Check Imputed Linear
```

```r
summary(lm(pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)]))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6230 -0.4836  0.0052  0.4950  4.2172
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.031513   0.001835   17.17   <2e-16 ***
## pheno$imputed_linear[!is.na(g)] 0.633931   0.001843  343.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7662 on 174317 degrees of freedom
##   (174418 observations deleted due to missingness)
## Multiple R-squared:  0.4042, Adjusted R-squared:  0.4042
## F-statistic: 1.183e+05 on 1 and 174317 DF,  p-value: < 2.2e-16
```

```r
summary(lm(pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)] + X.cov - 1))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)] +
##     X.cov - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6771 -0.4523  0.0090  0.4624  4.2261
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## pheno$imputed_linear[!is.na(g)]  0.023847   0.009119   2.615  0.00892 **
## X.covg_complete                  0.001904   0.002549   0.747  0.45508
## X.covf.21022.0.0                -0.121019   0.005459 -22.169  < 2e-16 ***
## X.covf.22001.0.0                 1.358080   0.015202  89.336  < 2e-16 ***
## X.covPC1                        -0.009723   0.001867  -5.206 1.93e-07 ***
## X.covPC2                         0.040082   0.009518   4.211 2.54e-05 ***
## X.covPC3                         0.039735   0.009566   4.154 3.27e-05 ***
## X.covPC4                        -0.009461   0.001747  -5.416 6.11e-08 ***
## X.covPC5                        -0.033986   0.003639  -9.340  < 2e-16 ***
## X.covPC6                         0.034364   0.003130  10.979  < 2e-16 ***
## X.covPC7                         0.015108   0.002520   5.995 2.04e-09 ***
## X.covPC8                        -0.005082   0.001747  -2.909  0.00363 **
## X.covPC9                        -0.012140   0.001725  -7.038 1.95e-12 ***
## X.covPC10                        0.023877   0.001816  13.151  < 2e-16 ***
## X.cov                           -0.593941   0.007362 -80.682  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6974 on 174304 degrees of freedom
```

```
##   (174418 observations deleted due to missingness)
## Multiple R-squared:  0.5069, Adjusted R-squared:  0.5069
## F-statistic: 1.195e+04 on 15 and 174304 DF,  p-value: < 2.2e-16
```

```
# Check Permuted Outcome
permuted <- sample(pheno$yhat[!is.na(g)])
summary(lm(pheno$int[!is.na(g)] ~ permuted))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ permuted)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2143 -0.6503 -0.0247  0.6300  4.3351
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.031399   0.002377  13.207   <2e-16 ***
## permuted    0.002025   0.002378   0.852    0.394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9926 on 174317 degrees of freedom
##   (174418 observations deleted due to missingness)
## Multiple R-squared:  4.161e-06,  Adjusted R-squared:  -1.575e-06
## F-statistic: 0.7254 on 1 and 174317 DF,  p-value: 0.3944
```

```
summary(lm(pheno$int[!is.na(g)] ~ permuted + X.cov - 1))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ permuted + X.cov - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6766 -0.4524  0.0093  0.4627  4.2314
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## permuted          0.001123   0.001671    0.672  0.50137
## X.covg_complete   0.001898   0.002549    0.745  0.45640
## X.covf.21022.0.0 -0.134607   0.001670  -80.608  < 2e-16 ***
## X.covf.22001.0.0  1.396852   0.003353  416.591  < 2e-16 ***
## X.covPC1         -0.009729   0.001867   -5.210 1.90e-07 ***
## X.covPC2          0.040082   0.009518    4.211 2.54e-05 ***
## X.covPC3          0.039740   0.009566    4.154 3.27e-05 ***
## X.covPC4         -0.009449   0.001747   -5.409 6.36e-08 ***
## X.covPC5         -0.033973   0.003639   -9.336  < 2e-16 ***
## X.covPC6          0.034356   0.003130   10.976  < 2e-16 ***
## X.covPC7          0.015096   0.002520    5.990 2.10e-09 ***
## X.covPC8         -0.005063   0.001747   -2.898  0.00375 **
## X.covPC9         -0.012144   0.001725   -7.040 1.92e-12 ***
## X.covPC10         0.023879   0.001816   13.152  < 2e-16 ***
## X.cov            -0.611773   0.002775 -220.446  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6974 on 174304 degrees of freedom
##   (174418 observations deleted due to missingness)
## Multiple R-squared:  0.5069, Adjusted R-squared:  0.5069
## F-statistic: 1.195e+04 on 15 and 174304 DF,  p-value: < 2.2e-16
```

```
# SynSurr with linear regression
SurrogateRegression::FitBNR(
  t = pheno$int[!is.na(g)],
  s = pheno$imputed_linear[!is.na(g)],
  X = X.cov
)@Regression.tab %>%
  filter(Outcome == "Target" & Coefficient == "g_complete")
```

```
##   Outcome Coefficient       Point          SE           L           U
## 1  Target  g_complete 0.001901229 0.002548793 -0.003094313 0.006896771
##           p
## 1 0.4557078
```

```
# SynSurr with permuted outcome
SurrogateRegression::FitBNR(
  t = pheno$int[!is.na(g)],
  s = permuted,
  X = X.cov
)@Regression.tab %>%
  filter(Outcome == "Target" & Coefficient == "g_complete")
```

```
##   Outcome Coefficient       Point          SE           L           U
## 1  Target  g_complete 0.001898784 0.002548816 -0.003096804 0.006894372
##           p
## 1 0.4562916
```

```
# Check Random Forest
summary(lm(pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)]))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7687 -0.3826  0.0042  0.3866  3.8866
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.030319   0.001449   20.92   <2e-16 ***
## pheno$yhat[!is.na(g)] 0.786823   0.001449  542.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6051 on 174317 degrees of freedom
##   (174418 observations deleted due to missingness)
## Multiple R-squared:  0.6284, Adjusted R-squared:  0.6284
## F-statistic: 2.947e+05 on 1 and 174317 DF,  p-value: < 2.2e-16
```

```r
summary(lm(pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)] + X.cov - 1))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)] + X.cov -
##     1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0372 -0.3810  0.0017  0.3856  3.8466
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## pheno$yhat[!is.na(g)]  0.6246973  0.0024023 260.043  < 2e-16 ***
## X.covg_complete        0.0025614  0.0021635   1.184  0.23645
## X.covf.21022.0.0      -0.0163938  0.0014885 -11.013  < 2e-16 ***
## X.covf.22001.0.0       0.4071072  0.0047525  85.661  < 2e-16 ***
## X.covPC1              -0.0075965  0.0015852  -4.792 1.65e-06 ***
## X.covPC2               0.0395728  0.0080791   4.898 9.68e-07 ***
## X.covPC3               0.0423110  0.0081200   5.211 1.88e-07 ***
## X.covPC4              -0.0044891  0.0014830  -3.027  0.00247 **
## X.covPC5              -0.0333378  0.0030888 -10.793  < 2e-16 ***
## X.covPC6               0.0284226  0.0026570  10.697  < 2e-16 ***
## X.covPC7               0.0141705  0.0021391   6.625 3.49e-11 ***
## X.covPC8               0.0005616  0.0014831   0.379  0.70496
## X.covPC9              -0.0102359  0.0014641  -6.991 2.74e-12 ***
## X.covPC10              0.0212589  0.0015412  13.794  < 2e-16 ***
## X.cov                 -0.1581420  0.0029312 -53.951  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.592 on 174304 degrees of freedom
##   (174418 observations deleted due to missingness)
## Multiple R-squared:  0.6447, Adjusted R-squared:  0.6447
## F-statistic: 2.109e+04 on 15 and 174304 DF,  p-value: < 2.2e-16
```

```r
# SynSurr with permuted outcome
SurrogateRegression::FitBNR(
  t = pheno$int[!is.na(g)],
  s = pheno$yhat[!is.na(g)],
  X = X.cov
)@Regression.tab %>%
  filter(Outcome == "Target" & Coefficient == "g_complete")
```

```
##   Outcome Coefficient       Point          SE           L           U
## 1  Target  g_complete 0.002477691 0.002364012 -0.002155688 0.007111069
##           p
## 1 0.2945985
```

FEV1

```r
# read in the data
pheno <- readRDS("Data/Old/fev1_imputed.rds")
```

```r
# a random SNP
i <- sample(1:ncol(G), size = 1)
print(i)
```

```
## [1] 249873
```

```r
g <- as.numeric(G[as.character(pheno$f.eid), i]) # snp i
length(g)
```

```
## [1] 260878
```

```r
g_complete <- g[!is.na(g)]
length(g_complete)
```

```
## [1] 259712
```

```r
X.cov <- cbind(
  g_complete,
  (pheno %>%
    select(
      f.21022.0.0, f.22001.0.0,
      starts_with("PC")
    ))[!is.na(g), ]
)
dim(X.cov)
```

```
## [1] 259712     13
```

```r
colnames(X.cov)
```

```
##  [1] "g_complete"  "f.21022.0.0" "f.22001.0.0" "PC1"         "PC2"
##  [6] "PC3"         "PC4"         "PC5"         "PC6"         "PC7"
## [11] "PC8"         "PC9"         "PC10"
```

```r
X.cov <- scale(X.cov)
X.cov <- cbind(X.cov, rep(1, nrow(X.cov))) # append intercept
X.cov[, 1] <- g_complete # dont scale G
X.cov[, 3] <- pheno$f.22001.0.0[!is.na(g)] # dont scale sex


# Check Imputed Linear
summary(lm(pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)]))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5508 -0.4233  0.0488  0.4734  4.0110
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.006556   0.002047   3.202  0.00136 **
## pheno$imputed_linear[!is.na(g)] 0.678349   0.002057 329.809  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.7373 on 129697 degrees of freedom
##   (130013 observations deleted due to missingness)
## Multiple R-squared:  0.4561, Adjusted R-squared:  0.4561
## F-statistic: 1.088e+05 on 1 and 129697 DF,  p-value: < 2.2e-16
```

```
summary(lm(pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)] + X.cov - 1))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)] +
##     X.cov - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6167 -0.4130  0.0523  0.4701  4.1437
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## pheno$imputed_linear[!is.na(g)]  0.045217   0.011652   3.881 0.000104 ***
## X.covg_complete                 -0.009020   0.002861  -3.153 0.001616 **
## X.covf.21022.0.0                -0.345031   0.007091 -48.654  < 2e-16 ***
## X.covf.22001.0.0                 1.098534   0.019109  57.488  < 2e-16 ***
## X.covPC1                        -0.008449   0.002257  -3.743 0.000182 ***
## X.covPC2                         0.008248   0.011514   0.716 0.473768
## X.covPC3                         0.024875   0.011566   2.151 0.031504 *
## X.covPC4                        -0.004371   0.002117  -2.065 0.038955 *
## X.covPC5                        -0.001957   0.004399  -0.445 0.656423
## X.covPC6                         0.011770   0.003788   3.107 0.001890 **
## X.covPC7                        -0.002494   0.003045  -0.819 0.412784
## X.covPC8                         0.003932   0.002109   1.865 0.062253 .
## X.covPC9                        -0.004619   0.002090  -2.210 0.027092 *
## X.covPC10                        0.006010   0.002191   2.744 0.006079 **
## X.cov                           -0.496105   0.009573 -51.826  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7271 on 129684 degrees of freedom
##   (130013 observations deleted due to missingness)
## Multiple R-squared:  0.4711, Adjusted R-squared:  0.471
## F-statistic:  7700 on 15 and 129684 DF,  p-value: < 2.2e-16
```

```
# Check Permuted Outcome
permuted <- sample(pheno$yhat[!is.na(g)])
summary(lm(pheno$int[!is.na(g)] ~ permuted))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ permuted)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5174 -0.6787  0.0075  0.6753  4.2003
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.005336   0.002777   1.921   0.0547 .
## permuted    -0.001157   0.002790  -0.415   0.6782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9997 on 129697 degrees of freedom
##   (130013 observations deleted due to missingness)
## Multiple R-squared:  1.327e-06,  Adjusted R-squared:  -6.383e-06
## F-statistic: 0.1721 on 1 and 129697 DF,  p-value: 0.6782
```

```
summary(lm(pheno$int[!is.na(g)] ~ permuted + X.cov - 1))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ permuted + X.cov - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6235 -0.4129  0.0523  0.4707  4.1494
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## permuted          0.001510   0.002029    0.744 0.456913
## X.covg_complete  -0.009018   0.002861   -3.152 0.001620 **
## X.covf.21022.0.0 -0.371414   0.002022 -183.707  < 2e-16 ***
## X.covf.22001.0.0  1.171013   0.004049  289.211  < 2e-16 ***
## X.covPC1         -0.008472   0.002257   -3.753 0.000175 ***
## X.covPC2          0.008215   0.011515    0.713 0.475576
## X.covPC3          0.024897   0.011567    2.152 0.031366 *
## X.covPC4         -0.004342   0.002117   -2.051 0.040270 *
## X.covPC5         -0.001967   0.004399   -0.447 0.654751
## X.covPC6          0.011757   0.003788    3.103 0.001913 **
## X.covPC7         -0.002494   0.003046   -0.819 0.412919
## X.covPC8          0.003953   0.002109    1.874 0.060882 .
## X.covPC9         -0.004631   0.002090   -2.216 0.026703 *
## X.covPC10         0.006028   0.002191    2.751 0.005935 **
## X.cov            -0.529918   0.003991 -132.783  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7272 on 129684 degrees of freedom
##   (130013 observations deleted due to missingness)
## Multiple R-squared:  0.471,  Adjusted R-squared:  0.471
## F-statistic:  7698 on 15 and 129684 DF,  p-value: < 2.2e-16
```

```
# SynSurr with linear regression
SurrogateRegression::FitBNR(
  t = pheno$int[!is.na(g)],
  s = pheno$imputed_linear[!is.na(g)],
  X = X.cov
)@Regression.tab %>%
  filter(Outcome == "Target" & Coefficient == "g_complete")
```

```
##   Outcome Coefficient        Point           SE           L           U
```

```
## 1  Target  g_complete -0.009018047 0.002860713 -0.01462494 -0.003411151
##             p
## 1 0.001619471
```

```
# SynSurr with permuted outcome
SurrogateRegression::FitBNR(
  t = pheno$int[!is.na(g)],
  s = permuted,
  X = X.cov
)@Regression.tab %>%
  filter(Outcome == "Target" & Coefficient == "g_complete")
```

```
##   Outcome Coefficient        Point          SE           L            U
## 1  Target  g_complete -0.009018396 0.002860794 -0.01462545 -0.003411343
##             p
## 1 0.001619286
```

```
# Check Random Forest
summary(lm(pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)]))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5836 -0.3711  0.0519  0.4286  3.9050
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.01944    0.00186  -10.45   <2e-16 ***
## pheno$yhat[!is.na(g)]  0.74700    0.00187  399.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6695 on 129697 degrees of freedom
##   (130013 observations deleted due to missingness)
## Multiple R-squared:  0.5515, Adjusted R-squared:  0.5515
## F-statistic: 1.595e+05 on 1 and 129697 DF,  p-value: < 2.2e-16
```

```
summary(lm(pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)] + X.cov - 1))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)] + X.cov -
##     1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6413 -0.3656  0.0539  0.4272  3.9775
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## pheno$yhat[!is.na(g)]  0.6088629  0.0038181 159.466  < 2e-16 ***
## X.covg_complete       -0.0064144  0.0026159  -2.452 0.014203 *
## X.covf.21022.0.0      -0.0862336  0.0025721 -33.527  < 2e-16 ***
```

```
## X.covf.22001.0.0       0.2671408  0.0067701  39.459  < 2e-16 ***
## X.covPC1               -0.0040872  0.0020642  -1.980 0.047695 *
## X.covPC2               -0.0014426  0.0105289  -0.137 0.891022
## X.covPC3                0.0138460  0.0105764   1.309 0.190488
## X.covPC4               -0.0017038  0.0019357  -0.880 0.378761
## X.covPC5                0.0106826  0.0040234   2.655 0.007929 **
## X.covPC6               -0.0015681  0.0034650  -0.453 0.650864
## X.covPC7               -0.0092371  0.0027851  -3.317 0.000911 ***
## X.covPC8                0.0050401  0.0019282   2.614 0.008952 **
## X.covPC9               -0.0002844  0.0019112  -0.149 0.881705
## X.covPC10              -0.0037443  0.0020041  -1.868 0.061727 .
## X.cov                  -0.1325449  0.0044182 -29.999  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6649 on 129684 degrees of freedom
##   (130013 observations deleted due to missingness)
## Multiple R-squared:  0.5577, Adjusted R-squared:  0.5577
## F-statistic: 1.09e+04 on 15 and 129684 DF,  p-value: < 2.2e-16
```

```
# SynSurr with permuted outcome
SurrogateRegression::FitBNR(
  t = pheno$int[!is.na(g)],
  s = pheno$yhat[!is.na(g)],
  X = X.cov
)@Regression.tab %>%
  filter(Outcome == "Target" & Coefficient == "g_complete")
```

```
##   Outcome Coefficient        Point          SE          L            U
## 1  Target  g_complete -0.008532704 0.002740763 -0.0139045 -0.003160907
##             p
## 1 0.001850339
```