

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] 80090

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] 348952

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] 348952      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
```

```
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
## -4.6233 -0.4839  0.0051  0.4950  4.2172
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.031622   0.001834   17.24  <2e-16 ***
## pheno$imputed_linear[!is.na(g)] 0.634012   0.001843  344.09  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7661 on 174432 degrees of freedom
## (174518 observations deleted due to missingness)
## Multiple R-squared:  0.4043, Adjusted R-squared:  0.4043
## F-statistic: 1.184e+05 on 1 and 174432 DF,  p-value: < 2.2e-16
summary(lm(pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)] + X.cov))

##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)] +
##      X.cov)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.6764 -0.4523  0.0090  0.4622  4.2250
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -0.592928   0.007191 -82.454  < 2e-16 ***
## pheno$imputed_linear[!is.na(g)] 0.023674   0.009117   2.597  0.00942 **
## X.covg_complete                 0.006088   0.011429   0.533  0.59425
## X.covf.21022.0.0               -0.121237   0.005458 -22.212  < 2e-16 ***
## X.covf.22001.0.0               1.358391   0.015199  89.377  < 2e-16 ***
## X.covPC1                       -0.009708   0.001867  -5.200 2.00e-07 ***
## X.covPC2                       0.040440   0.009516   4.250 2.14e-05 ***
## X.covPC3                       0.040122   0.009564   4.195 2.73e-05 ***
## X.covPC4                       -0.009431   0.001746  -5.400 6.68e-08 ***
## X.covPC5                       -0.034020   0.003638  -9.351  < 2e-16 ***
## X.covPC6                       0.034168   0.003129  10.918  < 2e-16 ***
## X.covPC7                       0.015272   0.002519   6.062 1.35e-09 ***
## X.covPC8                       -0.004934   0.001746  -2.825  0.00472 **
## X.covPC9                       -0.012215   0.001724  -7.084 1.41e-12 ***
## X.covPC10                      0.023987   0.001815  13.216  < 2e-16 ***
## X.cov                          NA          NA      NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.6974 on 174419 degrees of freedom
## (174518 observations deleted due to missingness)
## Multiple R-squared: 0.5065, Adjusted R-squared: 0.5064
## F-statistic: 1.278e+04 on 14 and 174419 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
## -4.2170 -0.6495 -0.0258  0.6265  4.3369
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0316020  0.0023768  13.296  <2e-16 ***
## permuted    -0.0004946  0.0023797  -0.208    0.835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9927 on 174432 degrees of freedom
## (174518 observations deleted due to missingness)
## Multiple R-squared: 2.477e-07, Adjusted R-squared: -5.485e-06
## F-statistic: 0.0432 on 1 and 174432 DF, p-value: 0.8353

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

##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ permuted + X.cov)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.6746 -0.4523  0.0092  0.4628  4.2276
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.610629   0.002285 -267.230 < 2e-16 ***
## permuted      -0.000728   0.001672  -0.435  0.66326
## X.covg_complete  0.006094   0.011429  0.533  0.59388
## X.covf.21022.0.0 -0.134733   0.001669 -80.706 < 2e-16 ***
## X.covf.22001.0.0  1.396883   0.003352 416.744 < 2e-16 ***
## X.covPC1        -0.009712   0.001867  -5.202 1.98e-07 ***
## X.covPC2         0.040445   0.009516  4.250 2.14e-05 ***
## X.covPC3         0.040132   0.009565  4.196 2.72e-05 ***
## X.covPC4        -0.009420   0.001747  -5.394 6.91e-08 ***
## X.covPC5        -0.034007   0.003638  -9.347 < 2e-16 ***
## X.covPC6         0.034155   0.003129 10.914 < 2e-16 ***
## X.covPC7         0.015262   0.002519  6.058 1.38e-09 ***
## X.covPC8        -0.004918   0.001746  -2.816 0.00486 **
## X.covPC9        -0.012215   0.001724  -7.084 1.40e-12 ***
```

```
## X.covPC10          0.023993   0.001815   13.219   < 2e-16 ***
## X.cov              NA          NA          NA          NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6974 on 174419 degrees of freedom
## (174518 observations deleted due to missingness)
## Multiple R-squared:  0.5064, Adjusted R-squared:  0.5064
## F-statistic: 1.278e+04 on 14 and 174419 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      p
## 1 Target g_complete 0.006084183 0.01142932 -0.01631687 0.02848524 0.5944967

# 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      p
## 1 Target g_complete 0.006092206 0.01142943 -0.01630906 0.02849347 0.5940142

# 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
## -3.7688 -0.3827  0.0041  0.3867  3.8867
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.030378   0.001449   20.97   <2e-16 ***
## pheno$yhat[!is.na(g)] 0.786886   0.001449  543.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6051 on 174432 degrees of freedom
## (174518 observations deleted due to missingness)
## Multiple R-squared:  0.6284, Adjusted R-squared:  0.6284
## F-statistic: 2.95e+05 on 1 and 174432 DF,  p-value: < 2.2e-16
summary(lm(pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)] + X.cov))
```

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)] + X.cov)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.0360 -0.3810  0.0018  0.3858  3.8452
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.1567443   0.0026089  -60.081 < 2e-16 ***
## pheno$yhat[!is.na(g)]  0.6247245   0.0024016  260.133 < 2e-16 ***
## X.covg_complete    0.0117659   0.0097014    1.213  0.22521
## X.covf.21022.0.0   -0.0165224   0.0014881  -11.103 < 2e-16 ***
## X.covf.22001.0.0    0.4071163   0.0047510   85.691 < 2e-16 ***
## X.covPC1          -0.0076005   0.0015848   -4.796 1.62e-06 ***
## X.covPC2           0.0400338   0.0080775    4.956 7.20e-07 ***
## X.covPC3           0.0427403   0.0081185    5.265 1.41e-07 ***
## X.covPC4          -0.0044529   0.0014826   -3.003  0.00267 **
## X.covPC5          -0.0332783   0.0030883  -10.776 < 2e-16 ***
## X.covPC6           0.0282792   0.0026564   10.646 < 2e-16 ***
## X.covPC7           0.0142599   0.0021383    6.669 2.59e-11 ***
## X.covPC8           0.0006603   0.0014825    0.445  0.65603
## X.covPC9          -0.0102524   0.0014636   -7.005 2.48e-12 ***
## X.covPC10          0.0212806   0.0015406   13.813 < 2e-16 ***
## X.cov              NA          NA          NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.592 on 174419 degrees of freedom
## (174518 observations deleted due to missingness)
## Multiple R-squared:  0.6444, Adjusted R-squared:  0.6444
## F-statistic: 2.258e+04 on 14 and 174419 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      p
## 1 Target   g_complete 0.009045909 0.01060138 -0.01173241 0.02982423 0.393506
```

FEV1

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

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

```
## [1] 237238
g <- as.numeric(G[as.character(pheno$f.eid), i]) # snp i
length(g)

## [1] 260878
g_complete <- g[!is.na(g)]
length(g_complete)

## [1] 260520
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] 260520      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
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
## -4.5507 -0.4229  0.0487  0.4734  4.0110
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.006575   0.002044   3.217  0.00129 **
## pheno$imputed_linear[!is.na(g)] 0.678302   0.002053 330.452 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7371 on 130101 degrees of freedom
## (130417 observations deleted due to missingness)
## Multiple R-squared:  0.4563, Adjusted R-squared:  0.4563
## F-statistic: 1.092e+05 on 1 and 130101 DF, p-value: < 2.2e-16
```

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

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$imputed_linear[!is.na(g)] +
##     X.cov)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.6246 -0.4125  0.0523  0.4698  4.1348
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.505444   0.009125 -55.388 < 2e-16 ***
## pheno$imputed_linear[!is.na(g)]  0.045399   0.011629   3.904 9.47e-05 ***
## X.covg_complete  0.006072   0.007292   0.833 0.405041
## X.covf.21022.0.0 -0.344822   0.007079 -48.712 < 2e-16 ***
## X.covf.22001.0.0  1.098329   0.019072  57.587 < 2e-16 ***
## X.covPC1         -0.008445   0.002253  -3.748 0.000178 ***
## X.covPC2          0.009106   0.011496   0.792 0.428298
## X.covPC3          0.025713   0.011547   2.227 0.025957 *
## X.covPC4         -0.004263   0.002113  -2.017 0.043658 *
## X.covPC5         -0.002496   0.004392  -0.568 0.569823
## X.covPC6          0.012292   0.003782   3.250 0.001152 **
## X.covPC7         -0.002115   0.003040  -0.696 0.486559
## X.covPC8          0.003843   0.002105   1.826 0.067869 .
## X.covPC9         -0.004517   0.002086  -2.165 0.030370 *
## X.covPC10         0.006275   0.002187   2.869 0.004116 **
## X.cov              NA          NA          NA          NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.727 on 130088 degrees of freedom
## (130417 observations deleted due to missingness)
## Multiple R-squared:  0.4712, Adjusted R-squared:  0.4712
## F-statistic: 8281 on 14 and 130088 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
## -4.5176 -0.6786  0.0068  0.6752  4.2015
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.005339   0.002773   1.925  0.0542 .
## permuted     0.001505   0.002782   0.541  0.5885
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9997 on 130101 degrees of freedom
## (130417 observations deleted due to missingness)
## Multiple R-squared:  2.25e-06, Adjusted R-squared:  -5.437e-06
## F-statistic: 0.2927 on 1 and 130101 DF, p-value: 0.5885
```

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

```
##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ permuted + X.cov)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.6312 -0.4128  0.0523  0.4705  4.1400
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.539369   0.002815  -191.638 < 2e-16 ***
## permuted         0.001288   0.002023    0.637 0.524285
## X.covg_complete  0.006058   0.007293    0.831 0.406133
## X.covf.21022.0.0 -0.371312   0.002018 -183.995 < 2e-16 ***
## X.covf.22001.0.0  1.171091   0.004042  289.725 < 2e-16 ***
## X.covPC1        -0.008469   0.002253   -3.759 0.000171 ***
## X.covPC2         0.009041   0.011496    0.786 0.431625
## X.covPC3         0.025695   0.011547    2.225 0.026069 *
## X.covPC4        -0.004230   0.002113   -2.002 0.045318 *
## X.covPC5        -0.002497   0.004392   -0.568 0.569713
## X.covPC6         0.012280   0.003782    3.247 0.001166 **
## X.covPC7        -0.002120   0.003040   -0.697 0.485569
## X.covPC8         0.003864   0.002105    1.836 0.066390 .
## X.covPC9        -0.004521   0.002086   -2.167 0.030232 *
## X.covPC10        0.006311   0.002187    2.885 0.003911 **
## X.cov              NA          NA          NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.727 on 130088 degrees of freedom
## (130417 observations deleted due to missingness)
## Multiple R-squared:  0.4712, Adjusted R-squared:  0.4711
## F-statistic: 8279 on 14 and 130088 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
)%>%
  filter(Outcome == "Target" & Coefficient == "g_complete")
```

```
## Outcome Coefficient Point SE L U p
## 1 Target g_complete 0.006090789 0.007292399 -0.008202051 0.02038363 0.4035915
```

```
# 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 p
## 1 Target g_complete 0.006062352 0.007292605 -0.008230891 0.02035559 0.4058034
# 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
## -4.5834 -0.3711  0.0520  0.4286  3.9049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.019363   0.001857  -10.43  <2e-16 ***
## pheno$yhat[!is.na(g)]  0.746796   0.001867  399.99  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6695 on 130101 degrees of freedom
## (130417 observations deleted due to missingness)
## Multiple R-squared:  0.5515, Adjusted R-squared:  0.5515
## F-statistic: 1.6e+05 on 1 and 130101 DF, p-value: < 2.2e-16
summary(lm(pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)] + X.cov))

```

```

##
## Call:
## lm(formula = pheno$int[!is.na(g)] ~ pheno$yhat[!is.na(g)] + X.cov)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.6466 -0.3655  0.0539  0.4269  3.9712
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.1398885   0.0035896  -38.971  < 2e-16 ***
## pheno$yhat[!is.na(g)]  0.6081498   0.0038103  159.608  < 2e-16 ***
## X.covg_complete      0.0055445   0.0066688   0.831  0.40575
## X.covf.21022.0.0    -0.0864192   0.0025674  -33.660  < 2e-16 ***
## X.covf.22001.0.0     0.2683327   0.0067568   39.713  < 2e-16 ***
## X.covPC1           -0.0040390   0.0020606  -1.960  0.04999 *
## X.covPC2           -0.0005140   0.0105131  -0.049  0.96101
## X.covPC3            0.0148520   0.0105599   1.406  0.15959
## X.covPC4           -0.0016075   0.0019324  -0.832  0.40549
## X.covPC5            0.0101239   0.0040174   2.520  0.01174 *
## X.covPC6           -0.0011034   0.0034595  -0.319  0.74976

```

```
## X.covPC7          -0.0088513  0.0027806  -3.183  0.00146 **
## X.covPC8          0.0049813  0.0019246   2.588  0.00965 **
## X.covPC9         -0.0002562  0.0019080  -0.134  0.89319
## X.covPC10        -0.0034675  0.0020010  -1.733  0.08312 .
## X.cov              NA          NA          NA          NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6648 on 130088 degrees of freedom
## (130417 observations deleted due to missingness)
## Multiple R-squared:  0.5578, Adjusted R-squared:  0.5577
## F-statistic: 1.172e+04 on 14 and 130088 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      p
## 1 Target g_complete 0.00406326 0.00699338 -0.009643513 0.01777003 0.5612302
```