Case Study: Blood Pressure

Rahiq Raees

Research Question:

For this case study, a genetic data set is generated based on a complex genetic model we developed at GSK. There are 500 predictors (483 genetic markers and 17 clinical covariates). The goal is to identify the 'true' predictors among the 500 variables and, at the same time, control the false discovery rate. Therefore, the objectives are:

1. Identify 'true' genes and clinical covariates
2. Control False Discovery (number of true X's versus number of false X's identified)

```
data_bp = read.table("bp.txt", header = TRUE)

head(data_bp)
```

```
##   sbp gender married smoke exercise age weight height overwt race alcohol trt
## 1 133      F       N     N        3  60    159     56      3    1       2   0
## 2 115      M       N     Y        1  55    107     65      1    1       2   0
## 3 140      M       N     Y        1  18    130     59      2    1       1   0
## 4 132      M       Y     N        2  19    230     57      3    2       3   1
## 5 133      M       N     N        2  58    201     74      2    1       3   0
## 6 138      F       N     N        3  55    166     67      2    1       1   1
##   bmi stress salt chldbear income educatn g1 g2 g3 g4 g5 g6 g7 g8 g9 g10
## 1  35      2    2        2      2       2  0  1  0  2  2  2  2  2  0   0
## 2  17      2    2        1      3       2  1  2  0  2  1  2  2  2  0   1
## 3  26      3    2        1      1       3  0  1  0  2  1  2  2  1  1   2
## 4  49      3    3        1      1       2  0  2  0  2  0  1  1  1  1   2
## 5  25      2    2        1      2       3  2  2  0  2  0  2  2  1  1   1
## 6  25      2    1        3      2       3  0  2  0  2  1  1  1  0  2   2
```

```
# obtain data from data_bp, names are corresponding to description pdf
sbp = data_bp$sbp

gender = data_bp$gender
married = data_bp$married
smoke = data_bp$smoke
age = data_bp$age
weight = data_bp$weight
height = data_bp$height
bmi = data_bp$bmi
overwt = data_bp$overwt
race = data_bp$race
exercise = data_bp$exercise
alcohol = data_bp$alcohol
stress = data_bp$stress
salt = data_bp$salt
chldbear = data_bp$chldbear
```

```
income = data_bp$income
educatn = data_bp$educatn
trt = data_bp$trt # to coincide with dataset, name trt is used here

g1 = data_bp$g1
g2 = data_bp$g2
g3 = data_bp$g3
g4 = data_bp$g4
g5 = data_bp$g5
g6 = data_bp$g6
g7 = data_bp$g7
g8 = data_bp$g8
g9 = data_bp$g9
g10 = data_bp$g10
```

```
# all the dummy variable names start with "dv_"
# and followed by their variable names with meanings in dataset
dv_gender = relevel(factor(gender), ref = "M")
dv_married = relevel(factor(married), ref = "N")
dv_smoke = relevel(factor(smoke), ref = "N")
dv_overwt = relevel(factor(overwt), ref = "1")
dv_race = relevel(factor(race), ref = "1")
dv_exercise = relevel(factor(exercise), ref = "1")
dv_alcohol = relevel(factor(alcohol), ref = "1")
dv_stress = relevel(factor(stress), ref = "1")
dv_salt = relevel(factor(salt), ref = "1")
dv_chldbear = relevel(factor(chldbear), ref = "1")
dv_income = relevel(factor(income), ref = "1")
dv_educatn = relevel(factor(educatn), ref = "1")
dv_trt = relevel(factor(trt), ref = "1")
```

Identified the categorical variables and created the corresponding dummy variables. For gender, used "M" as the reference group. For all other binary variables, used "N" as the reference group. For categorical variables, used the group with value '1' as the reference group.

Determining which variables should not be included together in the regression model:

Variable "bmi" should not be included in the model, since we can obtain that by "height" and "weight". Variable "gender" should not be included in the model as well, since we already have "chldbear" standing for childbearing potential. Not having a "1" in "chldbear" is equivalent as stating the subject as a female. Variable "overwt" should not be included in the model, since the judgement of being overweight or not is fully based on weight, height and bmi, which we have all the information.

To mathematically check if these predictors are of high correlation to others in the model, we have regressed "bmi" and "overwt" over the other predictors:

```
bmi_check = lm(bmi ~ dv_gender + dv_married + dv_smoke + age + weight + height + overwt +
                 dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
                 dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 + g3 +
                 g4 + g5 + g6 + g7 + g8 + g9 + g10, data = data_bp)
```

```
summary(bmi_check):

Residual standard error: 1.464 on 465 degrees of freedom
Multiple R-squared:  0.9727,    Adjusted R-squared:  0.9707
F-statistic: 487.8 on 34 and 465 DF,  p-value: < 2.2e-16
```

Since the R-squared value is very close to 1, we can safely conclude that "bmi" is highly correlated with at least one predictor already in the model, thus we can exclude it.

```
overwt_check = lm(overwt ~ dv_gender + dv_married + dv_smoke + age + weight + height + bmi +
                  dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
                  dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 + g3 +
                  g4 + g5 + g6 + g7 + g8 + g9 + g10, data = data_bp)
```

`summary(overwt_check)`:

```
Residual standard error: 0.3921 on 465 degrees of freedom
Multiple R-squared:  0.8169,    Adjusted R-squared:  0.8035
F-statistic: 61.01 on 34 and 465 DF,  p-value: < 2.2e-16
```

Now regressing "overwt" over the other predictors, we still get a relatively high r-squared value, again signifying that "overwt" is highly correlated with at least one predictor already in the model, thus we exclude this one as well.

Overall, including such redundant variables in our model might result in multicollinearity and hence our statistical inferences might decrease the accuracy. Besides, redundant variables could result in selecting bad models if the selection criteria RSS is used.

```
fit_full = lm(sbp ~ dv_married + dv_smoke + age + weight + height +
                dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
                dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 + g3 +
                g4 + g5 + g6 + g7 + g8 + g9 + g10, data = data_bp)
summary(fit_full)
```

```
##
## Call:
## lm(formula = sbp ~ dv_married + dv_smoke + age + weight + height +
##     dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
##     dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 +
##     g3 + g4 + g5 + g6 + g7 + g8 + g9 + g10, data = data_bp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.327 -17.921  -0.155  15.769  66.878
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   98.55312   17.38836   5.668 2.54e-08 ***
## dv_marriedY    3.23353    2.34021   1.382  0.16772
## dv_smokeY     10.27299    2.35141   4.369 1.54e-05 ***
## age            0.15225    0.08638   1.763  0.07862 .
## weight         0.19125    0.02961   6.459 2.65e-10 ***
## height        -0.45801    0.19655  -2.330  0.02022 *
## dv_race2       0.19939    2.94925   0.068  0.94613
## dv_race3       1.02319    5.38068   0.190  0.84927
## dv_race4      -5.81281    5.83773  -0.996  0.31990
## dv_exercise2 -11.87925    2.92312  -4.064 5.66e-05 ***
## dv_exercise3 -11.16955    2.71170  -4.119 4.50e-05 ***
## dv_alcohol2    2.71773    2.86977   0.947  0.34412
## dv_alcohol3   12.69439    2.89366   4.387 1.42e-05 ***
## dv_stress2     2.65799    2.89645   0.918  0.35927
## dv_stress3     5.76087    2.89666   1.989  0.04731 *
```

```
## dv_salt2       2.07267    2.88015   0.720  0.47211
## dv_salt3       1.24880    2.80631   0.445  0.65653
## dv_chldbear2  -0.90332    2.90226  -0.311  0.75575
## dv_chldbear3   3.86852    3.04050   1.272  0.20389
## dv_income2     1.89036    2.80035   0.675  0.49998
## dv_income3     5.39092    2.86114   1.884  0.06016 .
## dv_educatn2   -0.10213    2.84554  -0.036  0.97138
## dv_educatn3    0.02383    2.82510   0.008  0.99327
## dv_trt0       14.56146    2.91367   4.998 8.23e-07 ***
## g1            -1.10985    1.88581  -0.589  0.55646
## g2             3.33753    2.28629   1.460  0.14502
## g3             0.14202    2.20481   0.064  0.94867
## g4             1.04341    3.02544   0.345  0.73034
## g5             1.56016    2.12356   0.735  0.46290
## g6            -4.02052    2.72658  -1.475  0.14101
## g7             8.98933    3.09837   2.901  0.00389 **
## g8            -1.38636    2.45609  -0.564  0.57271
## g9             1.95338    2.17376   0.899  0.36932
## g10            0.07273    1.75617   0.041  0.96698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.16 on 466 degrees of freedom
## Multiple R-squared:  0.2456, Adjusted R-squared:  0.1921
## F-statistic: 4.596 on 33 and 466 DF,  p-value: 2.206e-14
```

Linear Regression Model:

$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \beta_5 X_{5,i} + \beta_6 X_{6,i} + \beta_7 X_{7,i} + \beta_8 X_{8,i} + \beta_9 X_{9,i} + \beta_{10} X_{10,i} + \beta_{11} X_{11,i} + \beta_{12} X_{12,i}$
$+ \beta_{13} X_{13,i} + \beta_{14} X_{14,i} + \beta_{15} X_{15,i} + \beta_{16} X_{16,i} + \beta_{17} X_{17,i} + \beta_{18} X_{18,i} + \beta_{19} X_{19,i} + \beta_{20} X_{20,i} + \beta_{21} X_{21,i} + \beta_{22} X_{22,i} + \beta_{23} X_{23,i}$
$+ \beta_{24} X_{24,i} + \beta_{25} X_{25,i} + \beta_{26} X_{26,i} + \beta_{27} X_{27,i} + \beta_{28} X_{28,i} + \beta_{29} X_{29,i} + \beta_{30} X_{30,i} + \beta_{31} X_{31,i} + \beta_{32} X_{32,i} + \beta_{33} X_{33,i}$
$+ \epsilon_i$ for i = 1,2,...n **The estimated regression parameters that I will interpret are:**
$\hat{\beta}_0$
$\hat{\beta}_1$ which corresponds to the binary variable $X_{1,i}$
$\hat{\beta}_9$ and $\hat{\beta}_{10}$ which corresponds to the Exercise Level categorical variable with dummy variables $X_{9,i}$ and $X_{10,i}$
$\hat{\beta}_3$ which corresponds to the continuous variable $X_{3,i}$
$\hat{\beta}_{24}$ which corresponds to the continuous variable $X_{24,i}$

Descriptions of the corresponding variables:

$$X_{1,i} = \begin{cases} 1 & \text{if } i^{th} \text{ individual is married} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{9,i} = \begin{cases} 1 & \text{if } i^{th} \text{ individual has a medium exercise level} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{10,i} = \begin{cases} 1 & \text{if } i^{th} \text{ individual has a high exercise level} \\ 0 & \text{otherwise} \end{cases}$$

$X_{3,i}$ = age in years of i^{th} individual
$X_{24,i}$ = type of minor allele(s) (0 = 0 minor allele, 1 = 1 minor allele, 2 = minor alleles) that the i^{th} individual has

Interpretation:
$\hat{\beta}_0 = 98.553$ is the mean systolic blood pressure when all predictors are equal to zero.

$\hat{\beta}_1 = 3.23353104$ is the mean difference in systolic blood pressure between an individual that is married and an individual that is not married that have the same age, weight,height,genes and share the same categories for all categorical variables.

$\hat{\beta}_3 = -11.87924618$ is the change in systolic blood pressure with one additional year of age when fixing all other variables.

$\hat{\beta}_9 = -0.45800894$ is the mean difference in systolic blood pressure between an individual that has a medium level of exercise and an individual that has a low level of exercise that have the same age, weight,height,genes and share the same categories for all categorical variables.

$\hat{\beta}_{10} = 0.19938550$ is the mean difference in systolic blood pressure between an individual that has a high level of exercise and an individual that has a low level of exercise that have the same age, weight,height,genes and share the same categories for all categorical variables.

$\hat{\beta}_{24} = 2.07266627$ is the change in systolic blood pressure when the value of gene 1 is increased by 1 (values of g1 are 0,1,2) when fixing all other variables.

```
# Create reduced models for each of the ten genes involved
# testing if B15=0
fit_rg1<- lm(sbp ~ dv_married + dv_smoke + age + weight + height +
               dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
               dv_chldbear + dv_income + dv_educatn + dv_trt + g2 + g3 + g4 +
              g5 + g6 + g7 + g8 + g9 + g10, data = data_bp)

anova(fit_rg1,fit_full )
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g2 + g3 + g4 + g5 + g6 +
##     g7 + g8 + g9 + g10
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    467 295265
## 2    466 295045  1     219.3 0.3464 0.5565
```

The corresponding P-value is 0.5565 which is greater than 0.05. As a result, we do not reject the null hypothesis and conclude that gene 1 has no effect on systolic blood pressure levels.

```
# testing if B16= 0
fit_rg2<-lm(sbp ~ dv_married + dv_smoke + age + weight + height +
```

```
                dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
                dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g3 + g4 +
              g5 + g6 + g7 + g8 + g9 + g10, data = data_bp)

anova(fit_rg2,fit_full )
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g3 + g4 + g5 + g6 +
##     g7 + g8 + g9 + g10
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    467 296394
## 2    466 295045  1    1349.2 2.131  0.145
```

The corresponding P-value is 0.145 which is greater than 0.05. As a result, we do not reject the null hypothesis and conclude that gene 2 has no effect on systolic blood pressure levels.

```
# testing if B17 = 0
fit_rg3<- lm(sbp ~ dv_married + dv_smoke + age + weight + height +
                dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
                dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 + g4 + g5
             + g6 + g7 + g8 + g9 + g10, data = data_bp)

anova(fit_rg3,fit_full )
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g4 + g5 + g6 +
##     g7 + g8 + g9 + g10
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    467 295048
## 2    466 295045  1    2.6269 0.0041 0.9487
```

The corresponding P-value is 0.9487 which is greater than 0.05. As a result, we do not reject the null hypothesis and conclude that gene 3 has no effect on systolic blood pressure levels.

```
# testing if B18 = 0
fit_rg4<- lm(sbp ~ dv_married + dv_smoke + age + weight + height +
                dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
                dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 + g3  +
               g5 + g6 + g7 + g8 + g9 + g10, data = data_bp)

anova(fit_rg4,fit_full )
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g5 + g6 +
##     g7 + g8 + g9 + g10
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    467 295121
## 2    466 295045  1    75.306 0.1189 0.7303
```

The corresponding P-value is 0.7303 which is greater than 0.05. As a result, we do not reject the null hypothesis and conclude that gene 4 has no effect on systolic blood pressure levels.

```
# testing if B19 = 0
fit_rg5<- lm(sbp ~ dv_married + dv_smoke + age + weight + height +
               dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
               dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4
             +  g6 + g7 + g8 + g9 + g10, data = data_bp)

anova(fit_rg5,fit_full )
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g6 +
##     g7 + g8 + g9 + g10
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    467 295387
## 2    466 295045  1    341.75 0.5398 0.4629
```

The corresponding P-value is 0.4629 which is greater than 0.05. As a result, we do not reject the null hypothesis and conclude that gene 5 has no effect on systolic blood pressure levels.

```
# testing if B20= 0
fit_rg6<- lm(sbp ~ dv_married + dv_smoke + age + weight + height +
               dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
               dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 + g3 +
              g4 +  g5 + g7 + g8 + g9 + g10, data = data_bp)

anova(fit_rg6,fit_full )
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g7 + g8 + g9 + g10
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
```

```
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    467 296422
## 2    466 295045  1    1376.7 2.1743  0.141
```

The corresponding P-value is 0.141 which is greater than 0.05. As a result, we do not reject the null hypothesis and conclude that gene 6 has no effect on systolic blood pressure levels.

```
#testing if B21 = 0
fit_rg7<- lm(sbp ~ dv_married + dv_smoke + age + weight + height +
               dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
               dv_chldbear + dv_income + dv_educatn + dv_trt +g1+ g2 + g3 +g4 +
              g5 + g6  + g8 + g9 + g10, data = data_bp)

anova(fit_rg7,fit_full )
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g8 + g9 + g10
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    467 300375
## 2    466 295045  1    5329.5 8.4176 0.003892 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The corresponding P-value is 0.003892 which is less than 0.05. As a result, we reject the null hypothesis and conclude that gene 7 has a significant effect on systolic blood pressure levels.

```
# testing if B22 = 0
fit_rg8<- lm(sbp ~ dv_married + dv_smoke + age + weight + height +
               dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
               dv_chldbear + dv_income + dv_educatn + dv_trt +g1+ g2 + g3 +g4 +
              g5 + g6 + g7 + g9 + g10, data = data_bp)

anova(fit_rg8,fit_full )
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g9 + g10
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    467 295247
```

```
## 2    466 295045  1    201.73 0.3186 0.5727
```

The corresponding P-value is 0.5727 which is greater than 0.05. As a result, we do not reject the null hypothesis and conclude that gene 8 has no effect on systolic blood pressure levels.

```
# testing if B23 = 0
fit_rg9<- lm(sbp ~ dv_married + dv_smoke + age + weight + height +
               dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
               dv_chldbear + dv_income + dv_educatn + dv_trt +g1+ g2 + g3 +g4 +
              g5 + g6 + g7 + g8 + g10, data = data_bp)

anova(fit_rg9,fit_full )
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g10
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    467 295556
## 2    466 295045  1    511.27 0.8075 0.3693
```

The corresponding P-value is 0.3693 which is greater than 0.05. As a result, we do not reject the null hypothesis and conclude that gene 9 has no effect on systolic blood pressure levels.

```
# testing if B24 = 0
fit_rg10<- lm(sbp ~ dv_married + dv_smoke + age + weight + height +
               dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
               dv_chldbear + dv_income + dv_educatn + dv_trt + g1+g2 + g3 +g4 +
               g5 + g6 + g7 + g8 + g9, data = data_bp)

anova(fit_rg10,fit_full )
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    467 295046
## 2    466 295045  1     1.086 0.0017  0.967
```

The corresponding P-value is 0.967 which is greater than 0.05. As a result, we do not reject the null hypothesis and conclude that gene 10 has no effect on systolic blood pressure levels.

Of the 10 individual tests we have done on g1 - g10, we can conclude that only g7 has a significant effect on systolic blood pressure levels. Further, we cannot make a conclusion about whether these 10 genes have a

joint effect on SBP because the tests we have run above are testing the individual effects of each of the 10 genes.

```
# create a linear model including all the possible interactions between the 10 genes
fit_int<- lm(sbp ~ dv_married + dv_smoke + age + weight + height +
               dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
              dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4
            + g5 + g6 + g7 + g8 + g9 + g10 +  g1:g2 + g1:g3+ g1:g4+ g1:g5 + g1:g6
            + g1:g7+ g1:g8+ g1:g9+ g1:g10 + g2:g3+ g2:g4+ g2:g5 + g2:g6+ g2:g7 +
              g2:g8+ g2:g9+ g2:g10 +g3:g4+ g3:g5 + g3:g6+ g3:g7+ g3:g8+ g3:g9+
              g3:g10+ g4:g5 + g4:g6+ g4:g7+ g4:g8+ g4:g9+ g4:g10+g5:g6+ g5:g7+
              g5:g8+ g5:g9+ g5:g10+g6:g7+ g6:g8+ g6:g9+ g6:g10+g7:g8+ g7:g9+
              g7:g10+g8:g9+ g8:g10 + g9:g10, data = data_bp)


summary(fit_int)
```

```
##
## Call:
## lm(formula = sbp ~ dv_married + dv_smoke + age + weight + height +
##     dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
##     dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 +
##     g3 + g4 + g5 + g6 + g7 + g8 + g9 + g10 + g1:g2 + g1:g3 +
##     g1:g4 + g1:g5 + g1:g6 + g1:g7 + g1:g8 + g1:g9 + g1:g10 +
##     g2:g3 + g2:g4 + g2:g5 + g2:g6 + g2:g7 + g2:g8 + g2:g9 + g2:g10 +
##     g3:g4 + g3:g5 + g3:g6 + g3:g7 + g3:g8 + g3:g9 + g3:g10 +
##     g4:g5 + g4:g6 + g4:g7 + g4:g8 + g4:g9 + g4:g10 + g5:g6 +
##     g5:g7 + g5:g8 + g5:g9 + g5:g10 + g6:g7 + g6:g8 + g6:g9 +
##     g6:g10 + g7:g8 + g7:g9 + g7:g10 + g8:g9 + g8:g10 + g9:g10,
##     data = data_bp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.968 -16.167  -0.039  15.119  62.907
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.18786   38.32454   1.414 0.158124
## dv_marriedY    3.59594    2.41257   1.491 0.136841
## dv_smokeY     10.93569    2.45075   4.462 1.04e-05 ***
## age            0.15836    0.09148   1.731 0.084169 .
## weight         0.19386    0.03104   6.246 1.03e-09 ***
## height        -0.45948    0.20401  -2.252 0.024822 *
## dv_race2       1.84231    3.12264   0.590 0.555517
## dv_race3       3.01558    5.70108   0.529 0.597120
## dv_race4      -4.54375    6.06563  -0.749 0.454217
## dv_exercise2 -11.52220    3.06874  -3.755 0.000198 ***
## dv_exercise3 -10.61630    2.87389  -3.694 0.000250 ***
## dv_alcohol2    4.47354    3.00468   1.489 0.137273
## dv_alcohol3   12.84673    3.00396   4.277 2.35e-05 ***
## dv_stress2     2.36530    3.01051   0.786 0.432497
## dv_stress3     6.82944    3.00680   2.271 0.023631 *
## dv_salt2       1.74455    3.00004   0.582 0.561208
## dv_salt3       0.64414    2.93299   0.220 0.826273
## dv_chldbear2  -1.50535    2.99501  -0.503 0.615495
```

```
## dv_chldbear3   4.56001    3.18412   1.432 0.152854
## dv_income2     0.84444    2.98183   0.283 0.777166
## dv_income3     4.37181    3.02051   1.447 0.148536
## dv_educatn2    1.21477    2.94234   0.413 0.679920
## dv_educatn3    0.27029    2.88851   0.094 0.925492
## dv_trt0       14.36812    3.02000   4.758 2.70e-06 ***
## g1           -18.05551   14.64100  -1.233 0.218183
## g2            29.70345   17.85486   1.664 0.096935 .
## g3            35.45293   16.59922   2.136 0.033271 *
## g4            24.23153   17.18880   1.410 0.159358
## g5           -10.74625   18.15023  -0.592 0.554120
## g6           -11.72987   19.33462  -0.607 0.544392
## g7            17.57637   20.57426   0.854 0.393431
## g8            -0.57504   19.06561  -0.030 0.975953
## g9            32.68211   17.70876   1.846 0.065662 .
## g10          -33.90156   13.26602  -2.556 0.010953 *
## g1:g2          1.34860    4.41797   0.305 0.760324
## g1:g3          3.47869    3.93076   0.885 0.376667
## g1:g4          8.84429    5.50415   1.607 0.108839
## g1:g5          0.50784    3.57047   0.142 0.886964
## g1:g6         -5.82910    4.69221  -1.242 0.214821
## g1:g7          5.84852    5.37158   1.089 0.276869
## g1:g8         -0.92503    4.06861  -0.227 0.820256
## g1:g9         -6.81107    3.97789  -1.712 0.087590 .
## g1:g10         0.72681    3.00070   0.242 0.808732
## g2:g3        -12.66827    6.00568  -2.109 0.035501 *
## g2:g4        -14.33535    8.57436  -1.672 0.095290 .
## g2:g5         -0.98021    4.58325  -0.214 0.830754
## g2:g6          3.08031    5.64235   0.546 0.585405
## g2:g7         -2.19237    6.44030  -0.340 0.733714
## g2:g8         -3.04819    5.54293  -0.550 0.582663
## g2:g9          4.98597    4.96999   1.003 0.316334
## g2:g10         2.71505    3.71153   0.732 0.464870
## g3:g4         -6.31981    5.51074  -1.147 0.252108
## g3:g5         12.88253    4.03465   3.193 0.001514 **
## g3:g6         -1.45915    5.67737  -0.257 0.797295
## g3:g7         -4.88561    6.64521  -0.735 0.462622
## g3:g8         -2.09984    5.14648  -0.408 0.683471
## g3:g9         -6.04570    4.72554  -1.279 0.201472
## g3:g10         3.90026    3.54910   1.099 0.272420
## g4:g5         -0.78454    7.05227  -0.111 0.911474
## g4:g6          1.75663    7.32061   0.240 0.810480
## g4:g7         -2.12440    8.47407  -0.251 0.802173
## g4:g8          4.08964    7.09952   0.576 0.564893
## g4:g9        -14.62363    6.87779  -2.126 0.034067 *
## g4:g10        10.79461    5.38395   2.005 0.045607 *
## g5:g6          3.51752    5.04998   0.697 0.486474
## g5:g7          9.81982    6.20829   1.582 0.114463
## g5:g8         -9.25042    4.83114  -1.915 0.056202 .
## g5:g9         -1.17515    4.45581  -0.264 0.792114
## g5:g10         0.40151    3.37282   0.119 0.905297
## g6:g7         -3.15695    4.06141  -0.777 0.437416
## g6:g8         -1.62774    5.34010  -0.305 0.760659
## g6:g9         11.12825    5.48959   2.027 0.043276 *
```

```
## g6:g10        -1.45926    4.34482  -0.336 0.737143
## g7:g8          1.44448    5.79410   0.249 0.803249
## g7:g9        -13.45216    6.56163  -2.050 0.040971 *
## g7:g10         5.08191    4.81037   1.056 0.291369
## g8:g9          1.39445    4.14066   0.337 0.736459
## g8:g10         4.51099    3.85023   1.172 0.242014
## g9:g10        -1.91350    3.10673  -0.616 0.538281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.97 on 421 degrees of freedom
## Multiple R-squared:  0.3288, Adjusted R-squared:  0.2044
## F-statistic: 2.644 on 78 and 421 DF,  p-value: 2.509e-10
```

```
# compare our linear model without interactions of the genes to the model with all of the gene interact

anova(fit_full,fit_int)
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
## Model 2: sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10 + g1:g2 + g1:g3 + g1:g4 + g1:g5 +
##     g1:g6 + g1:g7 + g1:g8 + g1:g9 + g1:g10 + g2:g3 + g2:g4 +
##     g2:g5 + g2:g6 + g2:g7 + g2:g8 + g2:g9 + g2:g10 + g3:g4 +
##     g3:g5 + g3:g6 + g3:g7 + g3:g8 + g3:g9 + g3:g10 + g4:g5 +
##     g4:g6 + g4:g7 + g4:g8 + g4:g9 + g4:g10 + g5:g6 + g5:g7 +
##     g5:g8 + g5:g9 + g5:g10 + g6:g7 + g6:g8 + g6:g9 + g6:g10 +
##     g7:g8 + g7:g9 + g7:g10 + g8:g9 + g8:g10 + g9:g10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    466 295045
## 2    421 262495 45     32550 1.1601 0.2292
```

The corresponding P-value is 0.2292 which is greater than 0.05. As a result, we do not reject the null hypothesis and conclude that the genes do not have a joint effect on SBP.

Ultimately, using predictors that have a high correlation with each other will have little to no effect on the prediction model. The predictor "gender" is directly correlated with "chldbear", as 1 in "chldbear" implies male and 2 and 3 imply female. We also identified "overweight" and "bmi" as redundant variables highly correlated with other predictors such as "height" and "weight".

For example, using linear regression while omitting "gender", "overweight", and "bmi" yields the following. We created a testing data set to use in each of the following examples:

```
defaultW <- getOption("warn")
options(warn = -1)

train1 <- data_bp[1:250,]

test1 <- data_bp[251:500,]

drop <- c("gender","overwt","bmi")
```

```
train1 = train1[,!(names(train1) %in% drop)]
test1 = test1[,!(names(test1) %in% drop)]

example_1 = lm(sbp ~ relevel(factor(married), ref = "N") + relevel(factor(smoke), ref = "N")
               + age + weight + height + relevel(factor(race), ref = "1") +
                  relevel(factor(exercise), ref = "1") + relevel(factor(alcohol), ref = "1")
               + relevel(factor(stress), ref = "1") + relevel(factor(salt), ref = "1") +
                 relevel(factor(chldbear), ref = "1") + relevel(factor(income), ref = "1")
               + relevel(factor(educatn), ref = "1")+ relevel(factor(trt), ref = "1")
               + g1 + g2 + g3 + g4 + g5 + g6 + g7 + g8 + g9 + g10, data = train1)

Y1 = predict(example_1, newdata = test1, interval="prediction", level=0.95)
options(warn = defaultW)
```

The predicted range for response Y given x = test is from 96.3352273 to 151.3582108. Now, let's add in "gender", which we've already set in the test data set to match "chldbear":

```
defaultW <- getOption("warn")
options(warn = -1)

train2 <- data_bp[1:250,]

test2 <- data_bp[251:500,]

drop <- c("overwt","bmi")

train2 = train2[,!(names(train2) %in% drop)]
test2 = test2[,!(names(test2) %in% drop)]

example_2 = lm(sbp ~ relevel(factor(married), ref = "N") + relevel(factor(smoke), ref = "N")
               + relevel(factor(gender), ref = "M") + age + weight + height +
                 relevel(factor(race), ref = "1") + relevel(factor(exercise), ref = "1")
               + relevel(factor(alcohol), ref = "1")
               + relevel(factor(stress), ref = "1") + relevel(factor(salt), ref = "1") +
                 relevel(factor(chldbear), ref = "1") + relevel(factor(income), ref = "1") +
                  relevel(factor(educatn), ref = "1")+ relevel(factor(trt), ref = "1")
               + g1 + g2 + g3 + g4 + g5 + g6 + g7 + g8 + g9 + g10, data = train2)

Y2 = predict(example_2, newdata = test2, interval="prediction", level=0.95)
options(warn = defaultW)
```

The resultant range for Y given "test" is 96.3352273 to 151.3582108, which is the exact same as the range given above. Even though we've added another predictor to the model, since it is directly correlated with another predictor in the model, it has no effect on the prediction.

To explain why this happens, let's look at the summary of the regression models "example_1" and "example_2", where "bmi" and "overweight" are omitted from both models and "gender" is omitted from "example_1"

Running summary(example_1) and summary(example_2), we get a wide array of numbers. To save space, I'll only include the relevant values here:

```
(1): summary(example_1):
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   98.55312   17.38836   5.668 2.54e-08
```

```
dv_chldbear2  -0.90332    2.90226  -0.311  0.75575
dv_chldbear3   3.86852    3.04050   1.272  0.20389

(2): summary(example_2):
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   98.55312   17.38836   5.668 2.54e-08
dv_genderF     3.86852    3.04050   1.272  0.20389
dv_chldbear2  -4.77184    3.19486  -1.494  0.13596
dv_chldbear3        NA         NA      NA       NA
```

First, notice how the intercept does not change. Even though we added another predictor, the intercept remained the same. Second, notice how the regression parameters for dv_genderF + dv_chldbear2 in (2) = dv_chldbear2 in (1). Last, notice that the regression parameter dv_genderF in (2) = dv_chldbear3 in (1). This is occuring because the magnitude of the regression parameters are being recalculated as we add another predictor, but the overall effect of the variable on the prediction remains the same.

Let's say, using (1), that we have a childbearing female, this will have an effect of -0.90332 on our prediction. In (2), since we have added gender into the equation, and this is a child-bearing female, dv_gender = F and dv_chldbear = 2. The total effect on the predictor is 3.86852 - 4.77184 = -0.90332, which is the same effect as (1).

Now let's say we have a non-childbearing woman. In (1), dv_chldbear = 3 which will have an effect of 3.86852 on our prediction. Using (2), we have dv_gender = F and dv_chldbear = 3, and this will have an effect of 3.86852 + 0 on our prediction, the same as (1).

Ultimately, the individual regression parameters change, but the overall prediction does not, when a predictor that is directly correlated with another predictor present in the model is added. Predictors added that have a high correlation to others present will also not have much of an effect on the prediction.

Forward selection

```r
library('MASS')

fit_full2 = lm(sbp ~ dv_married + dv_smoke + age + weight + height +
                dv_race + dv_exercise + dv_alcohol + dv_stress + dv_salt +
                dv_chldbear + dv_income + dv_educatn + dv_trt + g1 + g2 + g3 +
                g4 + g5 + g6 + g7 + g8 + g9 + g10, data = data_bp)

fit0 <- lm(sbp~1, data=data_bp)
stepAIC(fit0, direction="forward", scope=list(upper=fit_full2, lower=fit0))
```

```
## Start:  AIC=3333.02
## sbp ~ 1
##
##               Df Sum of Sq    RSS    AIC
## + weight       1   20737.8 370337 3307.8
## + dv_smoke     1   14634.9 376440 3316.0
## + g7           1    7912.3 383163 3324.8
## + dv_exercise  2    8895.5 382179 3325.5
## + dv_alcohol   2    8137.8 382937 3326.5
## + dv_trt       1    6186.9 384888 3327.0
## + height       1    5345.9 385729 3328.1
## + g8           1    4347.8 386727 3329.4
## + g10          1    3124.7 387950 3331.0
## + g9           1    2690.4 388384 3331.6
## <none>                     391075 3333.0
```

```
## + g6           1    1555.3 389520 3333.0
## + dv_married   1    1437.4 389637 3333.2
## + g2           1     631.7 390443 3334.2
## + age          1     548.9 390526 3334.3
## + g3           1     482.3 390593 3334.4
## + g1           1     435.4 390639 3334.5
## + dv_stress    2    1770.1 389305 3334.8
## + g5           1     105.2 390970 3334.9
## + g4           1      33.0 391042 3335.0
## + dv_chldbear  2    1461.7 389613 3335.2
## + dv_income    2     858.0 390217 3335.9
## + dv_salt      2     364.5 390710 3336.6
## + dv_educatn   2     197.5 390877 3336.8
## + dv_race      3     986.8 390088 3337.8
##
## Step:  AIC=3307.78
## sbp ~ weight
##
##               Df Sum of Sq    RSS    AIC
## + dv_smoke     1   12083.0 358254 3293.2
## + dv_alcohol   2   11025.3 359312 3296.7
## + dv_trt       1    9373.3 360964 3297.0
## + dv_exercise  2   10811.0 359526 3297.0
## + g7           1    7477.8 362859 3299.6
## + height       1    5963.3 364374 3301.7
## + g8           1    3992.6 366344 3304.4
## + g9           1    3198.5 367138 3305.4
## + dv_married   1    2468.9 367868 3306.4
## + g10          1    2236.0 368101 3306.8
## + g6           1    1917.1 368420 3307.2
## <none>                     370337 3307.8
## + dv_chldbear  2    2458.2 367879 3308.5
## + age          1     565.4 369772 3309.0
## + g2           1     444.1 369893 3309.2
## + g3           1     355.3 369982 3309.3
## + g1           1     340.9 369996 3309.3
## + g5           1     328.2 370009 3309.3
## + g4           1       5.6 370331 3309.8
## + dv_stress    2    1134.3 369203 3310.2
## + dv_income    2     756.5 369581 3310.8
## + dv_salt      2     238.9 370098 3311.5
## + dv_educatn   2      80.1 370257 3311.7
## + dv_race      3    1085.0 369252 3312.3
##
## Step:  AIC=3293.19
## sbp ~ weight + dv_smoke
##
##               Df Sum of Sq    RSS    AIC
## + dv_alcohol   2   12090.9 346163 3280.0
## + dv_trt       1   10578.1 347676 3280.2
## + dv_exercise  2   11874.0 346380 3280.3
## + g7           1    6567.8 351686 3285.9
## + height       1    4842.1 353412 3288.4
## + g8           1    2784.4 355470 3291.3
```

```
## + g9           1    2383.9 355870 3291.9
## + dv_married   1    2078.9 356175 3292.3
## + g10          1    1900.9 356353 3292.5
## + g6           1    1882.7 356371 3292.6
## <none>                     358254 3293.2
## + dv_chldbear  2    2032.8 356221 3294.3
## + age          1     583.8 357670 3294.4
## + g2           1     513.1 357741 3294.5
## + g5           1     366.6 357887 3294.7
## + g3           1     254.4 358000 3294.8
## + g1           1     110.4 358144 3295.0
## + g4           1      21.3 358233 3295.2
## + dv_income    2    1388.9 356865 3295.3
## + dv_stress    2    1090.1 357164 3295.7
## + dv_salt      2     158.7 358095 3297.0
## + dv_educatn   2     142.1 358112 3297.0
## + dv_race      3     597.3 357657 3298.4
##
## Step:  AIC=3280.03
## sbp ~ weight + dv_smoke + dv_alcohol
##
##               Df Sum of Sq    RSS    AIC
## + dv_trt       1   13209.4 332954 3262.6
## + dv_exercise  2   12515.8 333647 3265.6
## + g7           1    6097.7 340065 3273.1
## + height       1    4201.3 341962 3275.9
## + g8           1    2863.9 343299 3277.9
## + g9           1    2110.1 344053 3279.0
## + g6           1    1740.3 344423 3279.5
## <none>                     346163 3280.0
## + dv_married   1    1371.9 344791 3280.0
## + g10          1    1330.7 344832 3280.1
## + age          1    1148.1 345015 3280.4
## + g2           1     526.2 345637 3281.3
## + g5           1     482.2 345681 3281.3
## + g3           1     263.9 345899 3281.6
## + dv_chldbear  2    1571.1 344592 3281.8
## + g1           1     182.6 345981 3281.8
## + g4           1       1.2 346162 3282.0
## + dv_income    2    1159.9 345003 3282.4
## + dv_stress    2    1064.5 345099 3282.5
## + dv_educatn   2       8.2 346155 3284.0
## + dv_salt      2       1.3 346162 3284.0
## + dv_race      3     704.1 345459 3285.0
##
## Step:  AIC=3262.58
## sbp ~ weight + dv_smoke + dv_alcohol + dv_trt
##
##               Df Sum of Sq    RSS    AIC
## + dv_exercise  2   14255.1 318699 3244.7
## + g7           1    5928.5 327025 3255.6
## + height       1    3962.6 328991 3258.6
## + g8           1    2706.5 330247 3260.5
## + g9           1    2328.7 330625 3261.1
```

```
## + g6           1    1656.7 331297 3262.1
## + g10          1    1519.2 331435 3262.3
## + age          1    1512.5 331441 3262.3
## <none>                     332954 3262.6
## + dv_married   1    1017.3 331936 3263.0
## + g5           1     588.3 332365 3263.7
## + g2           1     573.3 332381 3263.7
## + dv_chldbear  2    1729.4 331224 3264.0
## + dv_income    2    1501.4 331452 3264.3
## + dv_stress    2    1472.6 331481 3264.4
## + g3           1      73.4 332880 3264.5
## + g1           1      61.2 332893 3264.5
## + g4           1       0.8 332953 3264.6
## + dv_salt      2       8.1 332946 3266.6
## + dv_educatn   2       3.5 332950 3266.6
## + dv_race      3     836.5 332117 3267.3
##
## Step:  AIC=3244.7
## sbp ~ weight + dv_smoke + dv_alcohol + dv_trt + dv_exercise
##
##               Df Sum of Sq    RSS    AIC
## + g7           1    5981.8 312717 3237.2
## + height       1    3328.2 315370 3241.4
## + g8           1    2388.9 316310 3242.9
## + g9           1    2209.2 316489 3243.2
## + age          1    1792.4 316906 3243.9
## + g6           1    1771.5 316927 3243.9
## <none>                     318699 3244.7
## + g10          1    1125.7 317573 3244.9
## + g5           1    1092.3 317606 3245.0
## + dv_chldbear  2    2264.5 316434 3245.1
## + g2           1     795.5 317903 3245.4
## + dv_married   1     679.3 318019 3245.6
## + dv_income    2    1935.5 316763 3245.7
## + dv_stress    2    1604.4 317094 3246.2
## + g3           1      96.5 318602 3246.5
## + g4           1      25.3 318673 3246.7
## + g1           1       6.8 318692 3246.7
## + dv_salt      2     105.2 318593 3248.5
## + dv_educatn   2     105.2 318593 3248.5
## + dv_race      3     832.3 317866 3249.4
##
## Step:  AIC=3237.22
## sbp ~ weight + dv_smoke + dv_alcohol + dv_trt + dv_exercise +
##     g7
##
##               Df Sum of Sq    RSS    AIC
## + height       1    3812.5 308904 3233.1
## + age          1    2023.2 310694 3236.0
## <none>                     312717 3237.2
## + dv_chldbear  2    2485.7 310231 3237.2
## + dv_income    2    2399.4 310317 3237.4
## + g2           1     998.3 311719 3237.6
## + dv_married   1     945.1 311772 3237.7
```

```
## + g6           1     848.9 311868 3237.9
## + g9           1     597.1 312120 3238.3
## + g10          1     499.9 312217 3238.4
## + dv_stress    2    1659.6 311057 3238.6
## + g3           1     289.2 312428 3238.8
## + g5           1     129.6 312587 3239.0
## + g4           1      60.7 312656 3239.1
## + g8           1      11.3 312706 3239.2
## + g1           1       6.8 312710 3239.2
## + dv_salt      2      64.9 312652 3241.1
## + dv_educatn   2      19.4 312698 3241.2
## + dv_race      3     687.7 312029 3242.1
##
## Step:  AIC=3233.09
## sbp ~ weight + dv_smoke + dv_alcohol + dv_trt + dv_exercise +
##     g7 + height
##
##               Df Sum of Sq    RSS    AIC
## + age          1   1976.14 306928 3231.9
## + dv_income    2   2610.09 306294 3232.8
## <none>                     308904 3233.1
## + dv_married   1   1065.09 307839 3233.4
## + g2           1    927.24 307977 3233.6
## + g6           1    740.56 308164 3233.9
## + dv_stress    2   1945.05 306959 3233.9
## + g10          1    631.55 308273 3234.1
## + g9           1    561.87 308343 3234.2
## + dv_chldbear  2   1756.30 307148 3234.2
## + g3           1    142.09 308762 3234.9
## + g1           1     70.29 308834 3235.0
## + g5           1     58.46 308846 3235.0
## + g4           1     35.20 308869 3235.0
## + g8           1      6.28 308898 3235.1
## + dv_salt      2    146.50 308758 3236.9
## + dv_educatn   2      5.69 308899 3237.1
## + dv_race      3    789.65 308115 3237.8
##
## Step:  AIC=3231.88
## sbp ~ weight + dv_smoke + dv_alcohol + dv_trt + dv_exercise +
##     g7 + height + age
##
##               Df Sum of Sq    RSS    AIC
## + dv_income    2   2488.11 304440 3231.8
## <none>                     306928 3231.9
## + dv_married   1   1094.83 305833 3232.1
## + g2           1    989.70 305939 3232.3
## + g6           1    824.27 306104 3232.5
## + dv_chldbear  2   1900.49 305028 3232.8
## + g9           1    650.18 306278 3232.8
## + g10          1    598.80 306329 3232.9
## + dv_stress    2   1804.91 305123 3232.9
## + g3           1    175.14 306753 3233.6
## + g5           1     48.85 306879 3233.8
## + g1           1     31.78 306896 3233.8
```

```
## + g4           1     29.18 306899 3233.8
## + g8           1      0.22 306928 3233.9
## + dv_salt      2    133.86 306794 3235.7
## + dv_educatn   2      3.44 306925 3235.9
## + dv_race      3    877.29 306051 3236.4
##
## Step:  AIC=3231.81
## sbp ~ weight + dv_smoke + dv_alcohol + dv_trt + dv_exercise +
##     g7 + height + age + dv_income
##
##               Df Sum of Sq    RSS    AIC
## <none>                     304440 3231.8
## + dv_married   1   1169.09 303271 3231.9
## + g6           1   1105.86 303334 3232.0
## + g2           1    983.45 303457 3232.2
## + g9           1    676.76 303763 3232.7
## + dv_stress    2   1730.46 302710 3233.0
## + g10          1    497.47 303943 3233.0
## + dv_chldbear  2   1558.05 302882 3233.2
## + g3           1    175.65 304265 3233.5
## + g1           1     41.61 304399 3233.7
## + g4           1     17.95 304422 3233.8
## + g5           1     17.82 304422 3233.8
## + g8           1      0.48 304440 3233.8
## + dv_salt      2    117.95 304322 3235.6
## + dv_educatn   2      6.92 304433 3235.8
## + dv_race      3    855.27 303585 3236.4
##
## Call:
## lm(formula = sbp ~ weight + dv_smoke + dv_alcohol + dv_trt +
##     dv_exercise + g7 + height + age + dv_income, data = data_bp)
##
## Coefficients:
##  (Intercept)        weight     dv_smokeY   dv_alcohol2   dv_alcohol3
##     112.7370        0.1870       11.0571        1.9125       12.5150
##      dv_trt0  dv_exercise2  dv_exercise3            g7        height
##      14.0765      -10.8987      -11.2405        5.6124       -0.4598
##          age    dv_income2    dv_income3
##       0.1462        2.6649        5.5651
```

Backward selection

```
stepAIC(fit_full2, direction="backward")
```

```
## Start:  AIC=3258.14
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_race +
##     dv_exercise + dv_alcohol + dv_stress + dv_salt + dv_chldbear +
##     dv_income + dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 +
##     g6 + g7 + g8 + g9 + g10
##
##               Df Sum of Sq    RSS    AIC
## - dv_race      3     683.1 295728 3253.3
## - dv_educatn   2       1.4 295047 3254.1
```

```
## - dv_salt      2     332.9 295378 3254.7
## - g10          1       1.1 295046 3256.1
## - g3           1       2.6 295048 3256.1
## - g4           1      75.3 295121 3256.3
## - g8           1     201.7 295247 3256.5
## - g1           1     219.3 295265 3256.5
## - g5           1     341.8 295387 3256.7
## - dv_chldbear  2    1580.7 296626 3256.8
## - g9           1     511.3 295556 3257.0
## - dv_income    2    2299.8 297345 3258.0
## <none>                     295045 3258.1
## - dv_married   1    1208.8 296254 3258.2
## - dv_stress    2    2536.3 297581 3258.4
## - g2           1    1349.2 296394 3258.4
## - g6           1    1376.7 296422 3258.5
## - age          1    1967.0 297012 3259.5
## - height       1    3438.0 298483 3261.9
## - g7           1    5329.5 300375 3265.1
## - dv_smoke     1   12084.8 307130 3276.2
## - dv_alcohol   2   13701.7 308747 3276.8
## - dv_exercise  2   14730.7 309776 3278.5
## - dv_trt       1   15813.6 310859 3282.2
## - weight       1   26417.9 321463 3299.0
##
## Step:  AIC=3253.29
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_stress + dv_salt + dv_chldbear + dv_income +
##     dv_educatn + dv_trt + g1 + g2 + g3 + g4 + g5 + g6 + g7 +
##     g8 + g9 + g10
##
##               Df Sum of Sq    RSS    AIC
## - dv_educatn   2       7.0 295735 3249.3
## - dv_salt      2     265.5 295994 3249.7
## - g10          1       4.0 295732 3251.3
## - g3           1       4.2 295733 3251.3
## - g4           1      76.6 295805 3251.4
## - g8           1     234.0 295962 3251.7
## - g1           1     266.1 295994 3251.7
## - g5           1     344.4 296073 3251.9
## - dv_chldbear  2    1680.0 297408 3252.1
## - g9           1     505.0 296233 3252.1
## - dv_income    2    2335.8 298064 3253.2
## <none>                     295728 3253.3
## - dv_stress    2    2421.8 298150 3253.4
## - g2           1    1332.0 297060 3253.5
## - dv_married   1    1361.3 297090 3253.6
## - g6           1    1449.1 297177 3253.7
## - age          1    1903.4 297632 3254.5
## - height       1    3292.8 299021 3256.8
## - g7           1    5621.4 301350 3260.7
## - dv_alcohol   2   13396.1 309124 3271.4
## - dv_smoke     1   12744.3 308473 3272.4
## - dv_exercise  2   14689.1 310417 3273.5
## - dv_trt       1   15798.8 311527 3277.3
```

```
## - weight       1   26175.2 321904 3293.7
##
## Step:  AIC=3249.31
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_stress + dv_salt + dv_chldbear + dv_income +
##     dv_trt + g1 + g2 + g3 + g4 + g5 + g6 + g7 + g8 + g9 + g10
##
##               Df Sum of Sq    RSS    AIC
## - dv_salt      2     271.3 296007 3245.8
## - g10          1       3.8 295739 3247.3
## - g3           1       3.9 295739 3247.3
## - g4           1      78.1 295813 3247.4
## - g8           1     247.5 295983 3247.7
## - g1           1     264.1 295999 3247.8
## - g5           1     341.3 296077 3247.9
## - dv_chldbear  2    1693.8 297429 3248.2
## - g9           1     517.3 296253 3248.2
## - dv_income    2    2342.0 298077 3249.3
## <none>                     295735 3249.3
## - dv_stress    2    2425.7 298161 3249.4
## - g2           1    1338.4 297074 3249.6
## - dv_married   1    1356.5 297092 3249.6
## - g6           1    1449.8 297185 3249.8
## - age          1    1906.0 297641 3250.5
## - height       1    3338.5 299074 3252.9
## - g7           1    5650.7 301386 3256.8
## - dv_alcohol   2   13565.0 309300 3267.7
## - dv_smoke     1   12761.5 308497 3268.4
## - dv_exercise  2   14825.0 310560 3269.8
## - dv_trt       1   15835.6 311571 3273.4
## - weight       1   26339.5 322075 3290.0
##
## Step:  AIC=3245.76
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_stress + dv_chldbear + dv_income + dv_trt +
##     g1 + g2 + g3 + g4 + g5 + g6 + g7 + g8 + g9 + g10
##
##               Df Sum of Sq    RSS    AIC
## - g3           1       5.0 296012 3243.8
## - g10          1       7.4 296014 3243.8
## - g4           1      77.9 296084 3243.9
## - g8           1     211.9 296219 3244.1
## - g1           1     257.3 296264 3244.2
## - g5           1     338.2 296345 3244.3
## - dv_chldbear  2    1662.7 297669 3244.6
## - g9           1     489.1 296496 3244.6
## - dv_income    2    2372.4 298379 3245.8
## <none>                     296007 3245.8
## - dv_stress    2    2403.8 298410 3245.8
## - g2           1    1287.6 297294 3245.9
## - dv_married   1    1304.3 297311 3246.0
## - g6           1    1415.2 297422 3246.1
## - age          1    1933.8 297940 3247.0
## - height       1    3222.1 299229 3249.2
```

```
## - g7           1    5567.3 301574 3253.1
## - dv_alcohol   2   13341.2 309348 3263.8
## - dv_smoke     1   12869.3 308876 3265.0
## - dv_exercise  2   14567.8 310574 3265.8
## - dv_trt       1   15878.1 311885 3269.9
## - weight       1   26415.5 322422 3286.5
##
## Step:  AIC=3243.77
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_stress + dv_chldbear + dv_income + dv_trt +
##     g1 + g2 + g4 + g5 + g6 + g7 + g8 + g9 + g10
##
##               Df Sum of Sq    RSS    AIC
## - g10          1       7.4 296019 3241.8
## - g4           1      80.6 296092 3241.9
## - g8           1     209.9 296221 3242.1
## - g1           1     260.2 296272 3242.2
## - g5           1     339.0 296351 3242.3
## - dv_chldbear  2    1677.7 297689 3242.6
## - g9           1     494.8 296506 3242.6
## - dv_income    2    2368.8 298380 3243.8
## <none>                     296012 3243.8
## - dv_stress    2    2398.9 298410 3243.8
## - dv_married   1    1301.8 297313 3244.0
## - g6           1    1418.1 297430 3244.2
## - g2           1    1446.6 297458 3244.2
## - age          1    1929.7 297941 3245.0
## - height       1    3252.5 299264 3247.2
## - g7           1    5568.2 301580 3251.1
## - dv_alcohol   2   13371.9 309383 3261.9
## - dv_smoke     1   12868.4 308880 3263.1
## - dv_exercise  2   14588.6 310600 3263.8
## - dv_trt       1   15998.3 312010 3268.1
## - weight       1   26455.9 322467 3284.6
##
## Step:  AIC=3241.79
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_stress + dv_chldbear + dv_income + dv_trt +
##     g1 + g2 + g4 + g5 + g6 + g7 + g8 + g9
##
##               Df Sum of Sq    RSS    AIC
## - g4           1      85.4 296104 3239.9
## - g8           1     210.2 296229 3240.1
## - g1           1     266.0 296285 3240.2
## - g5           1     344.1 296363 3240.4
## - dv_chldbear  2    1676.4 297695 3240.6
## - g9           1     631.2 296650 3240.9
## <none>                     296019 3241.8
## - dv_income    2    2393.3 298412 3241.8
## - dv_stress    2    2395.5 298414 3241.8
## - dv_married   1    1333.1 297352 3242.0
## - g6           1    1451.1 297470 3242.2
## - g2           1    1502.5 297522 3242.3
## - age          1    1941.7 297961 3243.1
```

```
## - height       1    3245.5 299265 3245.2
## - g7           1    5642.3 301661 3249.2
## - dv_alcohol   2   13454.5 309474 3260.0
## - dv_smoke     1   12875.3 308894 3261.1
## - dv_exercise  2   14657.6 310677 3261.9
## - dv_trt       1   15991.6 312011 3266.1
## - weight       1   26827.0 322846 3283.2
##
## Step:  AIC=3239.93
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_stress + dv_chldbear + dv_income + dv_trt +
##     g1 + g2 + g5 + g6 + g7 + g8 + g9
##
##               Df Sum of Sq    RSS    AIC
## - g8           1     209.5 296314 3238.3
## - g1           1     292.5 296397 3238.4
## - g5           1     416.2 296521 3238.6
## - dv_chldbear  2    1645.8 297750 3238.7
## - g9           1     635.8 296740 3239.0
## - dv_stress    2    2348.8 298453 3239.9
## - dv_income    2    2359.0 298463 3239.9
## <none>                     296104 3239.9
## - dv_married   1    1296.0 297400 3240.1
## - g6           1    1400.7 297505 3240.3
## - g2           1    1425.0 297529 3240.3
## - age          1    1921.9 298026 3241.2
## - height       1    3212.4 299317 3243.3
## - g7           1    5567.9 301672 3247.2
## - dv_alcohol   2   13661.5 309766 3258.5
## - dv_smoke     1   12794.6 308899 3259.1
## - dv_exercise  2   14829.7 310934 3260.4
## - dv_trt       1   15940.1 312045 3264.1
## - weight       1   26848.0 322952 3281.3
##
## Step:  AIC=3238.28
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_stress + dv_chldbear + dv_income + dv_trt +
##     g1 + g2 + g5 + g6 + g7 + g9
##
##               Df Sum of Sq    RSS    AIC
## - g1           1     279.3 296593 3236.8
## - g5           1     393.6 296707 3236.9
## - dv_chldbear  2    1627.8 297942 3237.0
## - g9           1     451.4 296765 3237.0
## - dv_stress    2    2235.8 298550 3238.0
## <none>                     296314 3238.3
## - dv_income    2    2393.3 298707 3238.3
## - dv_married   1    1345.2 297659 3238.5
## - g2           1    1372.2 297686 3238.6
## - g6           1    1438.0 297752 3238.7
## - age          1    1980.1 298294 3239.6
## - height       1    3246.5 299560 3241.7
## - g7           1    5730.5 302044 3245.9
## - dv_alcohol   2   13783.6 310097 3257.0
```

```
## - dv_smoke     1   12651.3 308965 3257.2
## - dv_exercise  2   14700.6 311014 3258.5
## - dv_trt       1   15840.7 312155 3262.3
## - weight       1   26746.7 323061 3279.5
##
## Step:  AIC=3236.75
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_stress + dv_chldbear + dv_income + dv_trt +
##     g2 + g5 + g6 + g7 + g9
##
##               Df Sum of Sq    RSS    AIC
## - g5           1     364.2 296957 3235.4
## - dv_chldbear  2    1634.5 298228 3235.5
## - g9           1     456.7 297050 3235.5
## - dv_stress    2    2224.2 298817 3236.5
## - g2           1    1141.9 297735 3236.7
## - dv_income    2    2366.1 298959 3236.7
## <none>                     296593 3236.8
## - dv_married   1    1349.8 297943 3237.0
## - g6           1    1395.2 297988 3237.1
## - age          1    2076.0 298669 3238.2
## - height       1    3136.9 299730 3240.0
## - g7           1    5650.3 302243 3244.2
## - dv_alcohol   2   13737.2 310330 3255.4
## - dv_smoke     1   13057.1 309650 3256.3
## - dv_exercise  2   14862.4 311456 3257.2
## - dv_trt       1   16159.5 312753 3261.3
## - weight       1   26790.6 323384 3278.0
##
## Step:  AIC=3235.37
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_stress + dv_chldbear + dv_income + dv_trt +
##     g2 + g6 + g7 + g9
##
##               Df Sum of Sq    RSS    AIC
## - dv_chldbear  2    1538.9 298496 3234.0
## - g9           1     490.9 297448 3234.2
## - dv_stress    2    2102.5 299060 3234.9
## - g6           1    1134.7 298092 3235.3
## - g2           1    1148.4 298106 3235.3
## <none>                     296957 3235.4
## - dv_income    2    2449.1 299406 3235.5
## - dv_married   1    1289.8 298247 3235.5
## - age          1    2087.7 299045 3236.9
## - height       1    3280.2 300238 3238.9
## - g7           1    5623.6 302581 3242.7
## - dv_alcohol   2   13593.1 310550 3253.7
## - dv_smoke     1   12974.4 309932 3254.7
## - dv_exercise  2   14594.3 311552 3255.4
## - dv_trt       1   16013.6 312971 3259.6
## - weight       1   26547.0 323504 3276.2
##
## Step:  AIC=3233.95
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
```

```
##     dv_alcohol + dv_stress + dv_income + dv_trt + g2 + g6 + g7 +
##     g9
##
##               Df Sum of Sq    RSS    AIC
## - g9           1     536.8 299033 3232.9
## - dv_stress    2    2006.0 300502 3233.3
## - g2           1    1124.0 299620 3233.8
## - g6           1    1169.4 299666 3233.9
## <none>                     298496 3234.0
## - dv_married   1    1351.7 299848 3234.2
## - dv_income    2    2818.5 301315 3234.7
## - age          1    1964.5 300461 3235.2
## - height       1    4138.1 302634 3238.8
## - g7           1    5628.8 304125 3241.3
## - dv_alcohol   2   13832.0 312328 3252.6
## - dv_exercise  2   14145.1 312641 3253.1
## - dv_smoke     1   13195.4 311692 3253.6
## - dv_trt       1   15722.5 314219 3257.6
## - weight       1   26965.0 325461 3275.2
##
## Step:  AIC=3232.85
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_stress + dv_income + dv_trt + g2 + g6 + g7
##
##               Df Sum of Sq    RSS    AIC
## - dv_stress    2    2181.2 301214 3232.5
## - g2           1    1044.6 300078 3232.6
## <none>                     299033 3232.9
## - g6           1    1290.6 300324 3233.0
## - dv_married   1    1313.1 300346 3233.0
## - dv_income    2    2794.6 301828 3233.5
## - age          1    1881.1 300914 3234.0
## - height       1    4191.2 303224 3237.8
## - g7           1    6963.1 305996 3242.4
## - dv_alcohol   2   13910.7 312944 3251.6
## - dv_exercise  2   14198.5 313232 3252.0
## - dv_smoke     1   13484.0 312517 3252.9
## - dv_trt       1   15585.4 314618 3256.3
## - weight       1   26619.1 325652 3273.5
##
## Step:  AIC=3232.48
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_income + dv_trt + g2 + g6 + g7
##
##               Df Sum of Sq    RSS    AIC
## - g2           1     934.5 302149 3232.0
## - dv_married   1    1077.5 302292 3232.3
## - g6           1    1177.9 302392 3232.4
## <none>                     301214 3232.5
## - dv_income    2    2854.3 304069 3233.2
## - age          1    2025.7 303240 3233.8
## - height       1    3891.3 305106 3236.9
## - g7           1    6681.5 307896 3241.5
## - dv_alcohol   2   13902.2 315116 3251.0
```

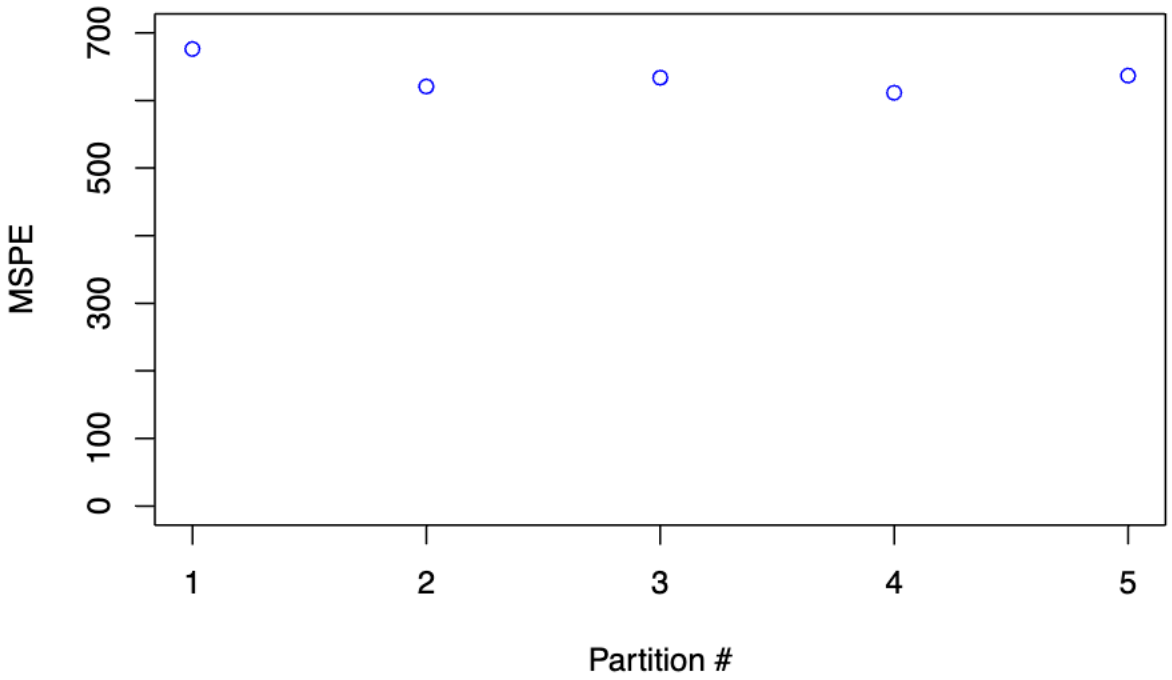
```
## - dv_exercise  2   14132.3 315347 3251.4
## - dv_smoke     1   13933.5 315148 3253.1
## - dv_trt       1   15058.1 316272 3254.9
## - weight       1   27166.6 328381 3273.7
##
## Step:  AIC=3232.03
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_income + dv_trt + g6 + g7
##
##               Df Sum of Sq    RSS    AIC
## - g6           1    1122.3 303271 3231.9
## - dv_married   1    1185.5 303334 3232.0
## <none>                     302149 3232.0
## - dv_income    2    2852.2 305001 3232.7
## - age          1    1966.4 304115 3233.3
## - height       1    3976.2 306125 3236.6
## - g7           1    6471.3 308620 3240.6
## - dv_alcohol   2   13816.6 315965 3250.4
## - dv_exercise  2   13873.5 316022 3250.5
## - dv_smoke     1   13830.1 315979 3252.4
## - dv_trt       1   14930.0 317079 3254.1
## - weight       1   27482.6 329631 3273.6
##
## Step:  AIC=3231.89
## sbp ~ dv_married + dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_income + dv_trt + g7
##
##               Df Sum of Sq    RSS    AIC
## - dv_married   1    1169.1 304440 3231.8
## <none>                     303271 3231.9
## - dv_income    2    2562.4 305833 3232.1
## - age          1    1878.5 305150 3233.0
## - height       1    4101.8 307373 3236.6
## - g7           1    7546.3 310817 3242.2
## - dv_alcohol   2   13897.2 317168 3250.3
## - dv_exercise  2   13956.6 317228 3250.4
## - dv_smoke     1   14258.4 317529 3252.9
## - dv_trt       1   14858.3 318129 3253.8
## - weight       1   28285.7 331557 3274.5
##
## Step:  AIC=3231.81
## sbp ~ dv_smoke + age + weight + height + dv_exercise + dv_alcohol +
##     dv_income + dv_trt + g7
##
##               Df Sum of Sq    RSS    AIC
## <none>                     304440 3231.8
## - dv_income    2    2488.1 306928 3231.9
## - age          1    1854.2 306294 3232.8
## - height       1    3972.8 308413 3236.3
## - g7           1    7216.4 311657 3241.5
## - dv_exercise  2   14383.0 318823 3250.9
## - dv_alcohol   2   14744.6 319185 3251.5
## - dv_smoke     1   14709.9 319150 3253.4
## - dv_trt       1   15270.6 319711 3254.3
```

```
## - weight       1   27669.0 332109 3273.3

##
## Call:
## lm(formula = sbp ~ dv_smoke + age + weight + height + dv_exercise +
##     dv_alcohol + dv_income + dv_trt + g7, data = data_bp)
##
## Coefficients:
##  (Intercept)     dv_smokeY           age        weight        height
##     112.7370       11.0571        0.1462        0.1870       -0.4598
## dv_exercise2  dv_exercise3   dv_alcohol2   dv_alcohol3    dv_income2
##     -10.8987      -11.2405        1.9125       12.5150        2.6649
##   dv_income3       dv_trt0            g7
##       5.5651       14.0765        5.6124
```

5-fold cross validation

```
n <- nrow(data_bp)

k <- 5

g <- cut(1:n, breaks=k, labels=FALSE)
set.seed(123)
g <- sample(g, n, replace=TRUE)

mspe <- data.frame(matrix(ncol = 2, nrow = k))
colnames(mspe) <- c("mod1", "mod2")

for(i in 1:k)
{
  train_i <- data_bp[which(!(g == i)), ]
  test_i <- data_bp[which(g == i), ]


  # FITS MODEL 1
   fit_i1 <- lm(sbp ~ weight + relevel(factor(smoke), ref = "N") +
                  relevel(factor(alcohol), ref = "1") +
                  relevel(factor(trt), ref = "1") +
   relevel(factor(exercise), ref = "1") + g7 + height + age +
     relevel(factor(income), ref = "1"),data = train_i)
  # COMPUTES PREDICTED VALUES ON TEST DATASET

  yhat_i1 <- predict(fit_i1, test_i)
  # COMPUTES MSPE FOR MODEL 1
  mspe_i1 <- mean((test_i$sbp - yhat_i1)^2)

  # FITS MODEL 2
   fit_i2 <- lm(sbp ~ relevel(factor(smoke), ref = "N") + age + weight + height +
                  relevel(factor(exercise), ref = "1") +
    relevel(factor(alcohol), ref = "1") + relevel(factor(income), ref = "1") +
      relevel(factor(trt), ref = "1") + g7, data=train_i)
  # COMPUTES PREDICTED VALUES ON TEST DATASET
  yhat_i2 <- predict(fit_i2, test_i)
  # COMPUTES MSPE FOR MODEL 1
```

```
  mspe_i2 <- mean((test_i$sbp - yhat_i2)^2)
  # SAVES MSPE FOR BOTH MODELS
  mspe[i, ] <- c(mspe_i1, mspe_i2)
}

# PLOTS MSPE BY PARTITION
plot(1:k, mspe$mod1, ylim=c(0,700), col="red", xlab="Partition #", ylab="MSPE")
points(1:k, mspe$mod2, col="blue")
```




```
dev.off()
```

```
## null device
##           1
```

```
# AVERAGES MSPE'S OVER PARTITIONS
apply(mspe, 2, mean)
```

```
##     mod1     mod2
## 635.8417 635.8417
```

Based on our results, the two models produce the same averaged MSPE. This happens is because forward selection and backward selection produce the same result, the only difference is the order of the variables. Therefore, we can choose either one as our final prediction model.