```r
setwd("C:/Users/JPRS1/Desktop/STAT3014/Major project/STAT3014-Major-Project")
proj_dat = read.csv("cleanedData.csv",row.names=1)
dataSteph = proj_dat[which(proj_dat$AGEC >= 18), ]

carb = cut(dataSteph$CHOPER1, breaks=c(-1, 45, 65, 100), labels=c("low", "med", "high"))
protein = cut(dataSteph$PROPER1, breaks=c(-1, 15, 25, 100), labels=c("low", "med", "high"))
fat = cut(dataSteph$FATPER1, breaks=c(-1, 20, 35, 100), labels=c("low", "med", "high"))

library(MASS)
#bmi, age, exercise,
dat1.var = dataSteph[,c(1,2,8, 11, 12, 14, 77, 86, 64, 65, 68)]
names(dat1.var) = c("bmi", "age", "mins.phys", "waist.cm", "bmr", "ses", "mins.sed", "sex", "protein",
```

can we predict who is obese according to BMI?

```r
bmi.class = cut(dat1.var$bmi, breaks=c(18.5, 30, 65), labels=c("norm","obese")) #categorical
bmi.num = ifelse(bmi.class=="norm", 0,1) #binary
bmi1 = dat1.var$bmi #numeric
# lot's of NA's
X = data.matrix(dat1.var[2:8])

#obese as binary for variables
mod = glm(bmi.num~X, family = binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = bmi.num ~ X, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.02058  -0.33978  -0.10160   0.05488   2.91087
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.331e+01  1.222e+00 -35.439  < 2e-16 ***
## Xage         5.237e-02  4.384e-03  11.946  < 2e-16 ***
## Xmins.phys   2.282e-04  1.566e-04   1.458  0.14493
## Xwaist.cm    1.753e-01  6.678e-03  26.252  < 2e-16 ***
## Xbmr         2.152e-03  1.056e-04  20.374  < 2e-16 ***
## Xses        -4.063e-02  2.999e-02  -1.355  0.17546
## Xmins.sed   -9.633e-05  3.435e-05  -2.804  0.00504 **
## Xsex         5.271e+00  1.940e-01  27.171  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8980.5  on 7593  degrees of freedom
## Residual deviance: 3663.1  on 7586  degrees of freedom
##   (1841 observations deleted due to missingness)
## AIC: 3679.1
##
## Number of Fisher Scoring iterations: 7
```

```
#obese as continuous for variables
mod1 = glm(bmi1~X)
summary(mod1)
```

```
##
## Call:
## glm(formula = bmi1 ~ X)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.7045  -1.5527  -0.1067   1.3694  22.1730
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.352e+01  4.082e-01 -57.618  < 2e-16 ***
## Xage         5.637e-02  2.500e-03  22.545  < 2e-16 ***
## Xmins.phys   1.733e-04  9.783e-05   1.772 0.076512 .
## Xwaist.cm    2.313e-01  3.452e-03  66.990  < 2e-16 ***
## Xbmr         2.661e-03  5.757e-05  46.216  < 2e-16 ***
## Xses        -7.011e-02  1.963e-02  -3.572 0.000356 ***
## Xmins.sed   -5.931e-05  2.194e-05  -2.704 0.006869 **
## Xsex         6.070e+00  8.994e-02  67.489  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 5.892893)
##
##     Null deviance: 239060  on 7715  degrees of freedom
## Residual deviance:  45422  on 7708  degrees of freedom
##   (1719 observations deleted due to missingness)
## AIC: 35593
##
## Number of Fisher Scoring iterations: 2
```

Have a smaller AIC when the data is run as continuous instead of binary. Using the binary variable model we have a log linear model of $log(p/1-p) = -43.3 + 0.05 * age + 0.18 * waist.cm + 0.002 * bmr - 0.00 * mins.sed + 5.26 * sex$ This tells us that the best model for describing BMI includes the following relationships: - Higher age have a higher BMI - Larger waist measurements have higher BMI - Higher BMR have higher BMI - More minutes per week sedentary is related to a higher BMI - Females have a higher BMI than males

Part 3: want to see if people of different diets have different behaviours and demographics. Use deviance tests

$H_0 : nullmodel(interceptonly)$ $H_1 : onefactormodel, protein$

In both hypothesis testing using the AIC and the deviance test we failed to reject the null hypothesis and prefer the one factor model for protein. BMI is only significant with respect to protein. To fit the mean percentage of obesity.
Part 3 - Model Selection using AIC Stepwise forward regression was used to select the most informative variables, which were included in a generalised linear model (GLM). GLMs were used because they are able

to handle different types of data including binary, categorical, and numerical. Logistic regression was used for binary data and classical regression was used for continuous data. A 5% significance level was chosen as a threshold for the inclusion of the model variables. In this section the aim was to determine the effectiveness of the three macro- nutrients for predicting gender, socio-economic status, Basal Betabolic Rate (BRM), BMI, age, and mintues spent sedentary for subjects in this study. These variables were chosen because they were found to be significant predictors of BMI. Can we determine the type of diet someone may have according to their SES, age, gender, waist measurement, bmr, time spent sedentary, energy intake, and time spend doing exercise.

BMI - BMISC

```
bmi.class = cut(dat1.var$bmi, breaks=c(18.5, 30, 65), labels=c("norm","obese")) #categorical
bmi.num = ifelse(bmi.class=="norm", 0,1) #binary

bmi.macro = glm(bmi.num~protein+carb+fat, family = binomial)
best.marco = stepAIC(bmi.macro, scope = list(upper = ~protein*carb*fat, lower = ~1))
```

```
## Start:  AIC=9271.44
## bmi.num ~ protein + carb + fat
##
##                 Df Deviance    AIC
## - carb           2   9258.8 9268.8
## - fat            2   9260.0 9270.0
## <none>               9257.4 9271.4
## + protein:carb   3   9255.3 9275.3
## + carb:fat       3   9256.8 9276.8
## + protein:fat    4   9255.9 9277.9
## - protein        2   9296.9 9306.9
##
## Step:  AIC=9268.77
## bmi.num ~ protein + fat
##
##                Df Deviance    AIC
## - fat           2   9260.4 9266.4
## <none>              9258.8 9268.8
## + carb          2   9257.4 9271.4
## + protein:fat   4   9257.7 9275.7
## - protein       2   9301.1 9307.1
##
## Step:  AIC=9266.38
## bmi.num ~ protein
##
##            Df Deviance    AIC
## <none>          9260.4 9266.4
## + fat       2   9258.8 9268.8
## + carb      2   9260.0 9270.0
## - protein   2   9303.3 9305.3
```

```
summary(best.marco)
```

```
##
## Call:
## glm(formula = bmi.num ~ protein, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9411  -0.8141  -0.7520   1.4338   1.6741
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.11858    0.04722 -23.688  < 2e-16 ***
## proteinmed   0.18444    0.05788   3.187  0.00144 **
## proteinhigh  0.53357    0.08089   6.596 4.22e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9303.3  on 7834  degrees of freedom
## Residual deviance: 9260.4  on 7832  degrees of freedom
##   (1600 observations deleted due to missingness)
## AIC: 9266.4
##
## Number of Fisher Scoring iterations: 4
```

```
#deviance test
mod12 = glm(bmi.num~protein, family = binomial)
mod13 = glm(bmi.num~fat, family = binomial)
mod14 = glm(bmi.num~carb, family = binomial)
a = anova(mod12, test = "Chisq")
b = anova(mod13, test = "Chisq")
c = anova(mod14, test = "Chisq")
a #protein is significant. So now look for protein and fat and protein and carb.
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: bmi.num
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     7834     9303.3
## protein  2   42.887      7832     9260.4 4.868e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
b
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: bmi.num
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                  7834     9303.3
## fat   2   2.1258      7832     9301.1   0.3454
```

```r
c
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: bmi.num
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                  7834     9303.3
## carb  2   4.0074      7832     9299.3   0.1348
```

```r
mod15 = glm(bmi.num~protein+fat, family = binomial)
mod16 = glm(bmi.num~protein+carb, family = binomial)
d = anova(mod15, test = "Chisq")
e = anova(mod16, test = "Chisq")
d
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: bmi.num
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     7834     9303.3
## protein  2   42.887      7832     9260.4 4.868e-10 ***
## fat      2    1.604      7830     9258.8    0.4484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
e
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: bmi.num
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     7834     9303.3
## protein  2   42.887      7832     9260.4 4.868e-10 ***
## carb     2    0.416      7830     9260.0     0.812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pars = coef(best.marco)
logistic = function(x){1/(1+exp(-x))}
prop.low = logistic(pars[1])
prop.med = logistic(pars[1]+pars[2])
prop.high = logistic(pars[1]+pars[2]+pars[3])
```

For BMI as a binary varible we have the best model using stepAIC as: $log(p/1-p) = -1.12 + 0.18 * MedProtein + 0.53 * HighProtein$ The estimated proprotion of obese people within the low protein group is 24.6% (since only working with intercept). The estimated proprotion of obese people within the medium protein group is 28.2%. The estimated proprotion of obese people within the high protein group is 40.1%.

Wasit measurement

```
#AIC model selection
waist = dat1.var$waist.cm
aovwaist = lm(waist~protein+fat+carb)
bestmod.waist = stepAIC(aovwaist , scope = list(upper = ~protein*fat*carb, lower = ~1))
```

```
## Start:  AIC=42823.48
## waist ~ protein + fat + carb
##
##                  Df Sum of Sq     RSS   AIC
## - fat             2    860.88 1749962 42823
## <none>                        1749101 42823
## + protein:carb    3    898.98 1748202 42825
## + protein:fat     4    842.48 1748259 42828
## - protein         2   1925.52 1751027 42828
## + fat:carb        3     89.73 1749011 42829
## - carb            2   2392.59 1751494 42830
##
## Step:  AIC=42823.39
## waist ~ protein + carb
##
##                  Df Sum of Sq     RSS   AIC
## <none>                        1749962 42823
## + fat             2    860.88 1749101 42823
## + protein:carb    3    975.89 1748986 42825
## - carb            2   1683.28 1751645 42827
## - protein         2   2409.35 1752371 42830
```

```
summary(bestmod.waist)
```

```
##
## Call:
## lm(formula = waist ~ protein + carb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.413 -10.627  -0.732   9.672  47.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  93.3089     0.3727 250.330  < 2e-16 ***
## proteinmed    0.3184     0.3831   0.831  0.40603
## proteinhigh   1.8730     0.5826   3.215  0.00131 **
## carbmed      -0.8955     0.3554  -2.520  0.01176 *
## carbhigh     -1.9062     1.2555  -1.518  0.12899
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.86 on 7929 degrees of freedom
##   (1501 observations deleted due to missingness)
## Multiple R-squared:  0.003254,   Adjusted R-squared:  0.002751
## F-statistic: 6.471 on 4 and 7929 DF,  p-value: 3.398e-05
```

```
#deviance test
mod17 = lm(waist~protein)
mod18 = lm(waist~fat)
mod19 = lm(waist~carb)
a = anova(mod17)
b = anova(mod18)
c = anova(mod19)
a #protein is significant and explains the most of the within factor variation out of the three variabl
```

```
## Analysis of Variance Table
##
## Response: waist
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## protein      2    4030 2014.91   9.123 0.0001103 ***
## Residuals 7931 1751645  220.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
b
```

```
## Analysis of Variance Table
##
## Response: waist
##             Df  Sum Sq Mean Sq F value Pr(>F)
## fat          2     192  95.818  0.4329 0.6486
## Residuals 7931 1755483 221.345
```

```
c
```

```
## Analysis of Variance Table
##
## Response: waist
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## carb         2    3304 1651.87  7.4762 0.0005704 ***
## Residuals 7931 1752371  220.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#protein
mod20 = lm(waist~protein+fat)
mod21 = lm(waist~protein+carb) #signif
d = anova(mod20)
e = anova(mod21)
d
```

```
## Analysis of Variance Table
##
## Response: waist
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## protein      2    4030 2014.91  9.1215 0.0001104 ***
## fat          2     152   75.79  0.3431 0.7095905
## Residuals 7929 1751494  220.90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
e
```

```
## Analysis of Variance Table
##
## Response: waist
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## protein      2    4030 2014.91  9.1295 0.0001096 ***
## carb         2    1683  841.64  3.8134 0.0221129 *
## Residuals 7929 1749962  220.70
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod21 = lm(waist~protein+carb+fat) #not signif
f = anova(mod21)
f
```

```
## Analysis of Variance Table
##
## Response: waist
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## protein      2    4030 2014.91  9.1317 0.0001093 ***
## carb         2    1683  841.64  3.8143 0.0220927 *
## fat          2     861  430.44  1.9508 0.1422311
## Residuals 7927 1749101  220.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod21 = lm(waist~protein*carb) #not signif
f = anova(mod21)
f
```

```
## Analysis of Variance Table
##
## Response: waist
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## protein         2    4030 2014.91  9.1311 0.0001094 ***
## carb            2    1683  841.64  3.8141 0.0220978 *
## protein:carb    3     976  325.30  1.4742 0.2193941
## Residuals    7926 1748986  220.66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#mean waist size - still zero?
mn.fat = 93.3089
mn.high.pro = 93.3089 +1.8730
mn.med.carb = 93.3089 +1.8730 - 0.8955
```

In both cases stepAIC and deviance test find that the best model for waist measurements relies on protein and carbs. The best model is: $Y_{ijk} = 93.3089 + 1.87 * HighProtein - 0.90 * MedCarb$. Because were are now doing normal regression we are looking at the mean within each group, not the proportion. The largest mean waist size in this model was found to be 95.12cm for people with a high protein diet.

BMR

```
bmr = dat1.var$bmr
aovbmr = lm(bmr~protein+fat+carb)
bestmod.bmr = stepAIC(lm(bmr~protein*fat*carb) , scope = list(upper = ~protein*fat*carb, lower = ~1))
```

```
## Start:  AIC=113764.2
## bmr ~ protein * fat * carb
##
##                    Df Sum of Sq        RSS    AIC
## - protein:fat:carb  3   6466763 1.1765e+10 113763
## <none>                          1.1759e+10 113764
##
## Step:  AIC=113762.6
## bmr ~ protein + fat + carb + protein:fat + protein:carb + fat:carb
##
##                    Df Sum of Sq        RSS    AIC
## <none>                          1.1765e+10 113763
## - protein:carb      3   9114384 1.1774e+10 113763
## - fat:carb          3   9372626 1.1775e+10 113763
## + protein:fat:carb  3   6466763 1.1759e+10 113764
## - protein:fat       4  17423107 1.1783e+10 113766
```

```
summary(bestmod.bmr)
```

```
##
## Call:
## lm(formula = bmr ~ protein + fat + carb + protein:fat + protein:carb +
##     fat:carb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2511.7  -970.7  -209.2   874.4  5974.0
##
## Coefficients: (2 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6792.647     96.398  70.465   <2e-16 ***
## proteinmed           -112.766    112.298  -1.004   0.3153
## proteinhigh            47.487    158.116   0.300   0.7639
## fatmed                 24.462    104.140   0.235   0.8143
## fathigh              -220.555    109.426  -2.016   0.0439 *
## carbmed                20.846    107.151   0.195   0.8458
## carbhigh             -190.932    151.825  -1.258   0.2086
## proteinmed:fatmed      -1.422    110.280  -0.013   0.9897
## proteinhigh:fatmed    -80.165    160.700  -0.499   0.6179
## proteinmed:fathigh    260.455    125.217   2.080   0.0376 *
## proteinhigh:fathigh   257.490    180.806   1.424   0.1545
## proteinmed:carbmed     84.053     75.538   1.113   0.2659
## proteinhigh:carbmed  -195.072    144.509  -1.350   0.1771
## proteinmed:carbhigh   408.888    284.417   1.438   0.1506
## proteinhigh:carbhigh       NA         NA      NA       NA
## fatmed:carbmed       -148.411    104.695  -1.418   0.1564
## fathigh:carbmed        59.353    130.962   0.453   0.6504
## fatmed:carbhigh       193.588    333.467   0.581   0.5616
## fathigh:carbhigh           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1213 on 7992 degrees of freedom
##   (1426 observations deleted due to missingness)
## Multiple R-squared:  0.003575,   Adjusted R-squared:  0.00158
## F-statistic: 1.792 on 16 and 7992 DF,  p-value: 0.02645
```

```
#deviance test
mod17 = lm(bmr~protein)
mod18 = lm(bmr~fat)
mod19 = lm(bmr~carb)
a = anova(mod17)
b = anova(mod18)
c = anova(mod19)
a
```

```
## Analysis of Variance Table
##
## Response: bmr
##             Df     Sum Sq Mean Sq F value  Pr(>F)
## protein      2 7.0396e+06 3519812   2.388 0.09188 .
## Residuals 8006 1.1801e+10 1473962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
b
```

```
## Analysis of Variance Table
##
## Response: bmr
##             Df     Sum Sq Mean Sq F value Pr(>F)
## fat          2 1.9246e+06  962290  0.6526 0.5207
## Residuals 8006 1.1806e+10 1474600
```

```
c
```

```
## Analysis of Variance Table
##
## Response: bmr
##             Df     Sum Sq Mean Sq F value Pr(>F)
## carb         2 4.0924e+06 2046193  1.3879 0.2497
## Residuals 8006 1.1803e+10 1474330
```

In this case we accept the null model.

Sedentary

```
sed = dat1.var$mins.sed
aovsed = lm(sed~protein+fat+carb)
bestmod.sed = stepAIC(aovsed , scope = list(upper = ~protein*fat*carb, lower = ~1))
```

```
## Start:  AIC=134997.6
## sed ~ protein + fat + carb
##
##                 Df Sum of Sq        RSS    AIC
## <none>                       1.6133e+10 134998
## - fat            2  11085121 1.6145e+10 135000
## - carb           2  12191586 1.6146e+10 135001
## + protein:carb  3   3319015 1.6130e+10 135002
## + fat:carb      3   1012037 1.6132e+10 135003
## + protein:fat   4   2185902 1.6131e+10 135004
## - protein       2  32139733 1.6166e+10 135012
```

```
summary(bestmod.sed)
```

```
##
## Call:
## lm(formula = sed ~ protein + fat + carb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2481.3  -959.4  -175.3   801.1  6920.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2350.15      54.91  42.798  < 2e-16 ***
## proteinmed    -92.92      31.40  -2.959   0.0031 **
## proteinhigh  -198.91      47.42  -4.194 2.76e-05 ***
## fatmed        108.18      47.25   2.289   0.0221 *
## fathigh       131.18      52.92   2.479   0.0132 *
## carbmed       -73.04      32.34  -2.258   0.0239 *
## carbhigh     -203.00     106.66  -1.903   0.0570 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1310 on 9396 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.003764,   Adjusted R-squared:  0.003128
## F-statistic: 5.917 on 6 and 9396 DF,  p-value: 3.533e-06
```

```
#deviance test
mod17 = lm(sed~protein)
mod18 = lm(sed~fat)
mod19 = lm(sed~carb)
a = anova(mod17)
b = anova(mod18)
c = anova(mod19)
a
```

```
## Analysis of Variance Table
##
## Response: sed
##             Df     Sum Sq  Mean Sq F value   Pr(>F)
## protein      2 2.0753e+07 10376530  6.0308 0.002413 **
## Residuals 9400 1.6174e+10  1720602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
b #most signif
```

```
## Analysis of Variance Table
##
## Response: sed
##             Df     Sum Sq  Mean Sq F value    Pr(>F)
## fat          2 2.6136e+07 13067843  7.5975 0.0005048 ***
## Residuals 9400 1.6168e+10  1720029
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
c
```

```
## Analysis of Variance Table
##
## Response: sed
##              Df     Sum Sq Mean Sq F value  Pr(>F)
## carb         2 1.3670e+07 6834957  3.9707 0.01889 *
## Residuals 9400 1.6181e+10 1721355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#fat
mod20 = lm(sed~fat+protein) #signif
mod21 = lm(sed~fat+carb)
d = anova(mod20)
e = anova(mod21)
d
```

```
## Analysis of Variance Table
##
## Response: sed
##              Df     Sum Sq  Mean Sq F value    Pr(>F)
## fat          2 2.6136e+07 13067843  7.6065 0.0005003 ***
## protein      2 2.2635e+07 11317678  6.5878 0.0013835 **
## Residuals 9398 1.6146e+10  1717987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
e
```

```
## Analysis of Variance Table
##
## Response: sed
##              Df     Sum Sq  Mean Sq F value   Pr(>F)
## fat          2 2.6136e+07 13067843  7.5971 0.000505 ***
## carb         2 2.6872e+06  1343604  0.7811 0.457925
## Residuals 9398 1.6166e+10  1720109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#fat+protein
mod20 = lm(sed~fat+protein+carb)
d = anova(mod20)
d #signif
```

```
## Analysis of Variance Table
##
## Response: sed
##              Df     Sum Sq  Mean Sq F value    Pr(>F)
## fat          2 2.6136e+07 13067843  7.6106 0.0004982 ***
## protein      2 2.2635e+07 11317678  6.5913 0.0013786 **
## carb         2 1.2192e+07  6095793  3.5501 0.0287590 *
## Residuals 9396 1.6133e+10  1717055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#interaction
mod20 = lm(sed~fat*protein*carb)
d = anova(mod20)
d
```

```
## Analysis of Variance Table
##
## Response: sed
##                    Df     Sum Sq  Mean Sq F value    Pr(>F)    
## fat                 2 2.6136e+07 13067843  7.6089 0.0004991 ***
## protein             2 2.2635e+07 11317678  6.5898 0.0013807 ** 
## carb                2 1.2192e+07  6095793  3.5493 0.0287826 *  
## fat:protein         4 2.1859e+06   546475  0.3182 0.8659717    
## fat:carb            3 1.7567e+06   585564  0.3409 0.7957239    
## protein:carb        3 7.6641e+06  2554710  1.4875 0.2157338    
## fat:protein:carb    3 7.0028e+06  2334263  1.3591 0.2532933    
## Residuals        9383 1.6115e+10  1717450                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Best model is the full additive model.

Sex

```r
#AIC
sex = ifelse(dat1.var$sex==1,1,0) #1 = male
glm.sex = glm(sex~protein+fat+carb, family= binomial)
bestmod.sex = stepAIC(glm.sex, scope = list(upper = ~protein*fat*carb, lower = ~1))
```

```
## Start:  AIC=12974.55
## sex ~ protein + fat + carb
##
##                Df Deviance   AIC
## + protein:fat   4    12952 12974
## <none>               12961 12975
## + fat:carb      3    12960 12980
## + protein:carb  3    12960 12980
## - protein       2    12975 12985
## - fat           2    12998 13008
## - carb          2    12998 13008
##
## Step:  AIC=12973.7
## sex ~ protein + fat + carb + protein:fat
##
##                Df Deviance   AIC
## <none>               12952 12974
## - protein:fat   4    12961 12975
## + fat:carb      3    12948 12976
## + protein:carb  3    12950 12978
## - carb          2    12992 13010
```

```r
summary(bestmod.sex)
```

```
##
```

```
## Call:
## glm(formula = sex ~ protein + fat + carb + protein:fat, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.379  -1.087  -1.055   1.250   1.427
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.46303    0.11584   3.997 6.41e-05 ***
## proteinmed           -0.25332    0.14222  -1.781   0.0749 .
## proteinhigh          -0.48824    0.19939  -2.449   0.0143 *
## fatmed               -0.31307    0.12212  -2.564   0.0104 *
## fathigh              -0.68138    0.13294  -5.126 2.97e-07 ***
## carbmed              -0.27346    0.05098  -5.364 8.15e-08 ***
## carbhigh             -0.71966    0.17156  -4.195 2.73e-05 ***
## proteinmed:fatmed     0.08241    0.15510   0.531   0.5952
## proteinhigh:fatmed    0.12278    0.21998   0.558   0.5767
## proteinmed:fathigh    0.24136    0.16754   1.441   0.1497
## proteinhigh:fathigh   0.53765    0.23734   2.265   0.0235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13016  on 9434  degrees of freedom
## Residual deviance: 12952  on 9424  degrees of freedom
## AIC: 12974
##
## Number of Fisher Scoring iterations: 4
```

```
#deviance test
#deviance test
mod17 = lm(sex~protein, family = "binomial")
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'family' will be disregarded
```

```
mod18 = lm(sex~fat,  family = "binomial")
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'family' will be disregarded
```

```
mod19 = lm(sex~carb,  family = "binomial")
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'family' will be disregarded
```

```
a = anova(mod17, test = "Chisq")
b = anova(mod18, test = "Chisq")
c = anova(mod19, test = "Chisq")
a
```

```
## Analysis of Variance Table
##
## Response: sex
##             Df  Sum Sq Mean Sq F value Pr(>F)
## protein      2    0.92 0.46090  1.8563 0.1563
## Residuals 9432 2341.83 0.24829
```

```
b #signif: explains more within SS variation
```

```
## Analysis of Variance Table
##
## Response: sex
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## fat          2    3.56 1.77978  7.1763 0.0007686 ***
## Residuals 9432 2339.19 0.24801
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
c  #signif
```

```
## Analysis of Variance Table
##
## Response: sex
##             Df  Sum Sq Mean Sq F value  Pr(>F)
## carb         2    2.18 1.08775  4.3834 0.01251 *
## Residuals 9432 2340.58 0.24815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#fat
mod20 = glm(sex~fat+protein, family = "binomial") #not signif
mod21 = glm(sex~fat+carb, family = "binomial") #signif
d = anova(mod20, test = "Chisq")
e = anova(mod21, test = "Chisq")
d
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sex
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     9434      13016
## fat      2  14.3188      9432      13001 0.0007775 ***
## protein  2   2.9418      9430      12998 0.2297186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
e
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sex
##
## Terms added sequentially (first to last)
##
```

```
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  9434      13016
## fat   2   14.319      9432      13001 0.0007775 ***
## carb  2   26.604      9430      12975 1.671e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#fat+carb
mod22 = glm(sex~fat+carb+protein, family = "binomial") #signif
f = anova(mod22, test = "Chisq")
f
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sex
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     9434      13016
## fat      2   14.319      9432      13001 0.0007775 ***
## carb     2   26.604      9430      12975 1.671e-06 ***
## protein  2   14.150      9428      12961 0.0008458 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
logistic = function(x){1/(1+exp(-x))}
p.add = logistic(0.46 - 0.48- 0.31 - 0.68- 0.27 - 0.72 + 0.54) #probabiliy of being male and being on t
p.add
```

```
## [1] 0.1884673
```

The suboptimal model is : sex~protein+fat+carb+protein:fat when using a significanc level of 0.05. This has the lowest AIC out of all one step models beginning with the full additive model. Our model becomes: $log(p/(1-p)) = 0.46 - 0.48 * HighProtein - 0.31 * MedFat - 0.68 * HighFat - 0.27 * MedCarb - 0.72 * HighCarb + 0.54 * HighProtein : HighFat$ This tells us that: - Less high protein diets

SES

```
dataSteph$SF2SA1QN = as.numeric(as.factor(dataSteph$SF2SA1QN))
ses = ifelse(dataSteph$SF2SA1QN>4,1,0) #decile [1:4] = 0
glmadd.ses = glm(ses~protein+fat+carb, family= binomial)
bestmod.ses = stepAIC(glmadd.ses, scope = list(upper = ~protein*fat*carb, lower = ~1))
```

```
## Start:  AIC=10141.72
## ses ~ protein + fat + carb
##
##                 Df Deviance   AIC
## + protein:fat    4    10118 10140
## - protein        2    10130 10140
## - fat            2    10130 10140
## + protein:carb   3    10121 10141
## <none>                10128 10142
## + fat:carb      3    10124 10144
## - carb          2    10142 10152
##
## Step:  AIC=10139.84
## ses ~ protein + fat + carb + protein:fat
##
##                 Df Deviance   AIC
## <none>                10118 10140
## + protein:carb  3    10113 10141
## - protein:fat   4    10128 10142
## + fat:carb      3    10115 10143
## - carb          2    10134 10152
```

```
summary(bestmod.ses)
```

```
##
## Call:
## glm(formula = ses ~ protein + fat + carb + protein:fat, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8356  -0.7277  -0.6987  -0.6297   1.9422
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.18602    0.13920  -8.520   <2e-16 ***
## proteinmed            0.31338    0.16860   1.859   0.0631 .
## proteinhigh          -0.12254    0.24752  -0.495   0.6205
## fatmed                0.13784    0.14745   0.935   0.3499
## fathigh              -0.04306    0.16033  -0.269   0.7882
## carbmed              -0.20883    0.05999  -3.481   0.0005 ***
## carbhigh             -0.53563    0.22105  -2.423   0.0154 *
## proteinmed:fatmed    -0.34454    0.18402  -1.872   0.0612 .
## proteinhigh:fatmed   -0.11499    0.27164  -0.423   0.6721
## proteinmed:fathigh   -0.27794    0.19898  -1.397   0.1625
## proteinhigh:fathigh   0.31207    0.28881   1.081   0.2799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10145  on 9434  degrees of freedom
## Residual deviance: 10118  on 9424  degrees of freedom
## AIC: 10140
##
## Number of Fisher Scoring iterations: 4
```

```
mod23 = lm(ses~protein, family = "binomial")
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'family' will be disregarded
```

```
mod24 = lm(ses~fat,  family = "binomial")
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'family' will be disregarded
```

```r
mod25 = lm(ses~carb,  family = "binomial")
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'family' will be disregarded
```

```r
a = anova(mod23, test = "Chisq")
b = anova(mod24, test = "Chisq")
c = anova(mod25, test = "Chisq")
a
```

```
## Analysis of Variance Table
##
## Response: ses
##             Df Sum Sq Mean Sq F value Pr(>F)
## protein      2    0.4 0.19856  1.1259 0.3244
## Residuals 9432 1663.5 0.17636
```

```r
b
```

```
## Analysis of Variance Table
##
## Response: ses
##             Df  Sum Sq  Mean Sq F value Pr(>F)
## fat          2    0.03 0.015159  0.0859 0.9177
## Residuals 9432 1663.84 0.176404
```

```r
c #signif
```

```
## Analysis of Variance Table
##
## Response: ses
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## carb         2    2.16 1.08100  6.1358 0.002173 **
## Residuals 9432 1661.71 0.17618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#carbs
mod20 = glm(sex~carb+protein, family = "binomial") #signif
mod21 = glm(sex~carb+fat, family = "binomial") #signif
d = anova(mod20, test = "Chisq")
e = anova(mod21, test = "Chisq")
d
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sex
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                     9434      13016
## carb     2   8.7716      9432      13007  0.01245 *
## protein  2   9.1697      9430      12998  0.01021 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
e
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sex
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  9434      13016
## carb  2    8.772      9432      13007   0.01245 *
## fat   2   32.151      9430      12975 1.043e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#fat+carb
mod22 = glm(sex~fat+carb+protein, family = "binomial") #signif
f = anova(mod22, test = "Chisq")
f
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sex
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     9434      13016
## fat      2   14.319      9432      13001 0.0007775 ***
## carb     2   26.604      9430      12975 1.671e-06 ***
## protein  2   14.150      9428      12961 0.0008458 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Beginning with the full additive model out best model according to one step AIC with a significance level of 0.05 is ses ~ protein + fat + carb + protein:fat. Summary statistics give the model as: $Y_{ijk} = -1.19 - 0.21 * MedCarb - 0.54 * HighCarb$ This indicates that if you are in a lo

ENERGY (BMR) - EIBMR1

```
bmr = dataSteph$EIBMR1
aovadd.bmr = lm(bmr~protein+fat+carb)
bestmod.bmr = stepAIC(aovadd.bmr , scope = list(upper = ~protein*fat*carb, lower = ~1))
```

```
## Start:  AIC=-11046.75
## bmr ~ protein + fat + carb
##
##                 Df Sum of Sq    RSS    AIC
## + protein:carb  3     2.499 2010.3 -11051
## <none>                      2012.8 -11047
## + fat:carb      3     0.532 2012.3 -11043
## + protein:fat   4     0.732 2012.1 -11042
## - carb          2    18.215 2031.0 -10979
## - fat           2    20.505 2033.3 -10970
## - protein       2   121.492 2134.3 -10581
##
## Step:  AIC=-11050.7
## bmr ~ protein + fat + carb + protein:carb
##
##                Df Sum of Sq    RSS    AIC
## <none>                      2010.3 -11051
## + fat:carb      3    0.5166 2009.8 -11047
## - protein:carb  3    2.4991 2012.8 -11047
## + protein:fat   4    0.1869 2010.1 -11043
## - fat           2   22.3289 2032.6 -10966
```

```r
summary(bestmod.bmr)
```

```
##
## Call:
## lm(formula = bmr ~ protein + fat + carb + protein:carb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4237 -0.3482 -0.0594  0.2814  4.8070
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.37294    0.02427  56.575  < 2e-16 ***
## proteinmed           -0.20207    0.01917 -10.542  < 2e-16 ***
## proteinhigh          -0.47155    0.02362 -19.961  < 2e-16 ***
## fatmed                0.15775    0.02018   7.819 6.02e-15 ***
## fathigh               0.21107    0.02243   9.408  < 2e-16 ***
## carbmed              -0.15260    0.02160  -7.064 1.75e-12 ***
## carbhigh             -0.26327    0.05032  -5.232 1.72e-07 ***
## proteinmed:carbmed    0.07214    0.02631   2.742  0.00612 **
## proteinhigh:carbmed   0.12378    0.05177   2.391  0.01684 *
## proteinmed:carbhigh   0.01659    0.10825   0.153  0.87823
## proteinhigh:carbhigh       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5013 on 7999 degrees of freedom
##   (1426 observations deleted due to missingness)
## Multiple R-squared:  0.07867,    Adjusted R-squared:  0.07763
## F-statistic: 75.89 on 9 and 7999 DF,  p-value: < 2.2e-16
```

Best model is given by $Y_{ijk} = 1.37 - 0.20 * MedProtein - 0.47 * HighProtein + 0.16 * MedFat + 0.21 HighFat - 0.15 * MedCarb - 0.26 * HighCarb + 0.07 MedProtein : MedCarb + 0.12 * HighProtein : MedCarb$

Exercise - ADTOTSE

```r
sedent = dataSteph$ADTOTSE
aovadd.sed = lm(sedent~protein+fat+carb)
bestmod.sed = stepAIC(aovadd.sed, scope = list(upper = ~protein*fat*carb, lower = ~1))
```

```
## Start:  AIC=134997.6
## sedent ~ protein + fat + carb
##
##                  Df Sum of Sq        RSS    AIC
## <none>                        1.6133e+10 134998
## - fat             2  11085121 1.6145e+10 135000
## - carb            2  12191586 1.6146e+10 135001
## + protein:carb    3   3319015 1.6130e+10 135002
## + fat:carb        3   1012037 1.6132e+10 135003
## + protein:fat     4   2185902 1.6131e+10 135004
## - protein         2  32139733 1.6166e+10 135012
```

```
summary(bestmod.sed)
```

```
##
## Call:
## lm(formula = sedent ~ protein + fat + carb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2481.3  -959.4  -175.3   801.1  6920.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2350.15      54.91  42.798  < 2e-16 ***
## proteinmed    -92.92      31.40  -2.959   0.0031 **
## proteinhigh  -198.91      47.42  -4.194 2.76e-05 ***
## fatmed        108.18      47.25   2.289   0.0221 *
## fathigh       131.18      52.92   2.479   0.0132 *
## carbmed       -73.04      32.34  -2.258   0.0239 *
## carbhigh     -203.00     106.66  -1.903   0.0570 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1310 on 9396 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.003764,   Adjusted R-squared:  0.003128
## F-statistic: 5.917 on 6 and 9396 DF,  p-value: 3.533e-06
```

Best model is the full additive model with significance of 0.05. We get the model $Y_{ijk} = 2362.68 - 91.42 * MedProtein - 200.97 * HighProtein + 100.89 MedFat + 123.35 HighFat - 75.94 * MedCarb$

Age - using ANOVA since age is not a factor but a numercial variable.

```
#anova - check if model good or not
#lm - estimate coefficients use summary(lm()) for model
age = dataSteph$AGEC
aovadd.age = lm(age~protein+fat+carb)
bestmod.age = stepAIC(aovadd.age, scope = list(upper = ~protein*fat*carb, lower = ~1))
```

```
## Start:  AIC=54059.1
## age ~ protein + fat + carb
##
##                  Df Sum of Sq     RSS   AIC
## <none>                        2900320 54059
## + protein:fat     4    2245.3 2898074 54060
## - protein       2    1866.7 2902186 54061
## + protein:carb  3     349.1 2899971 54064
## + fat:carb      3     325.9 2899994 54064
## - fat           2   10177.5 2910497 54088
## - carb          2   19303.4 2919623 54118
```

```
summary(bestmod.age)
```

```
##
## Call:
## lm(formula = age ~ protein + fat + carb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.976 -14.180  -1.251  13.356  40.127
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.2510     0.7343  69.795  < 2e-16 ***
## proteinmed    0.7247     0.4195   1.727   0.0841 .
## proteinhigh  -0.4439     0.6337  -0.700   0.4836
## fatmed       -0.7956     0.6320  -1.259   0.2081
## fathigh      -3.1630     0.7077  -4.470 7.93e-06 ***
## carbmed      -3.2154     0.4323  -7.437 1.12e-13 ***
## carbhigh     -6.2463     1.4215  -4.394 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.54 on 9428 degrees of freedom
## Multiple R-squared:  0.008546,   Adjusted R-squared:  0.007915
## F-statistic: 13.54 on 6 and 9428 DF,  p-value: 2.29e-15
```