Homework 3

Alex Ojemann

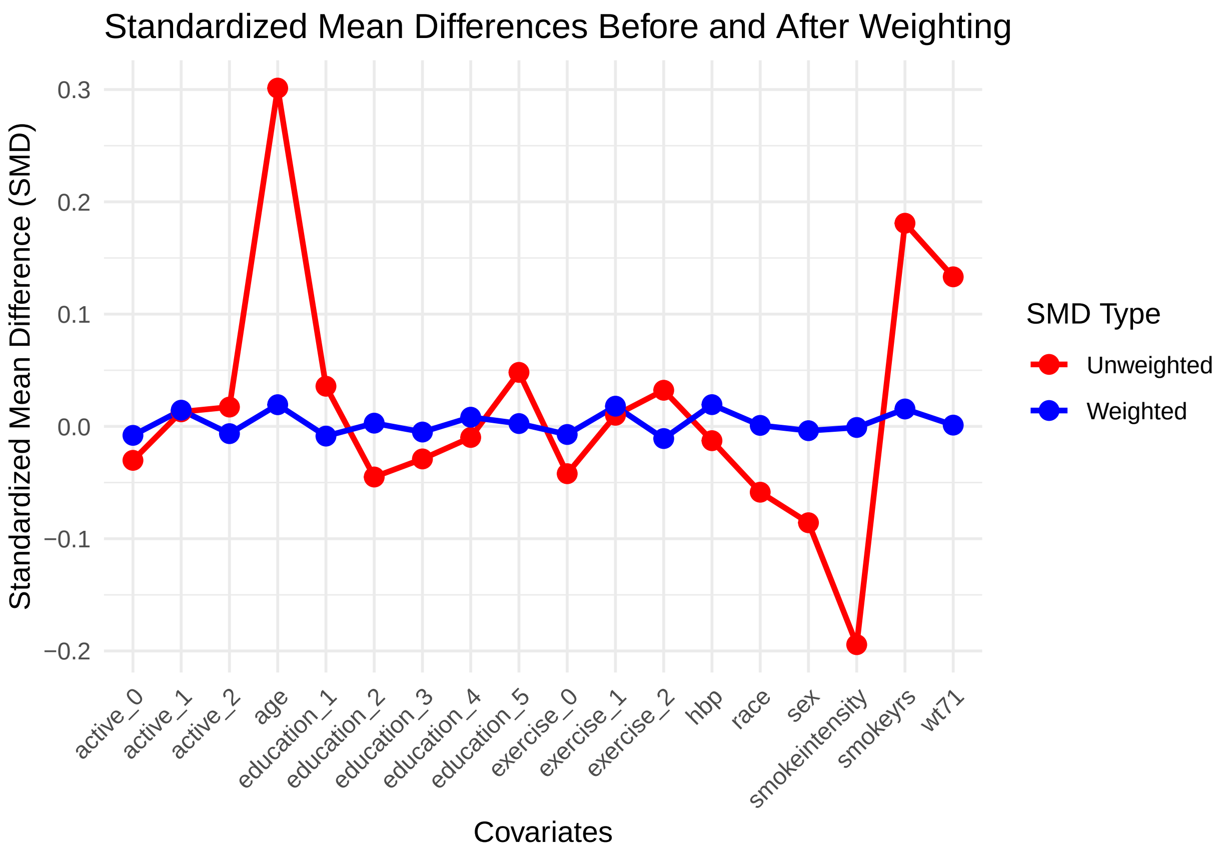
2024-10-03

Problem 1

Part a

The following logistic regression model predicts whether a participant quit smoking using the following covariates: sex, race, age, education, smoking intensity, number of years smoked for, exercise level, activity level, weight in 1971, and high blood pressure.

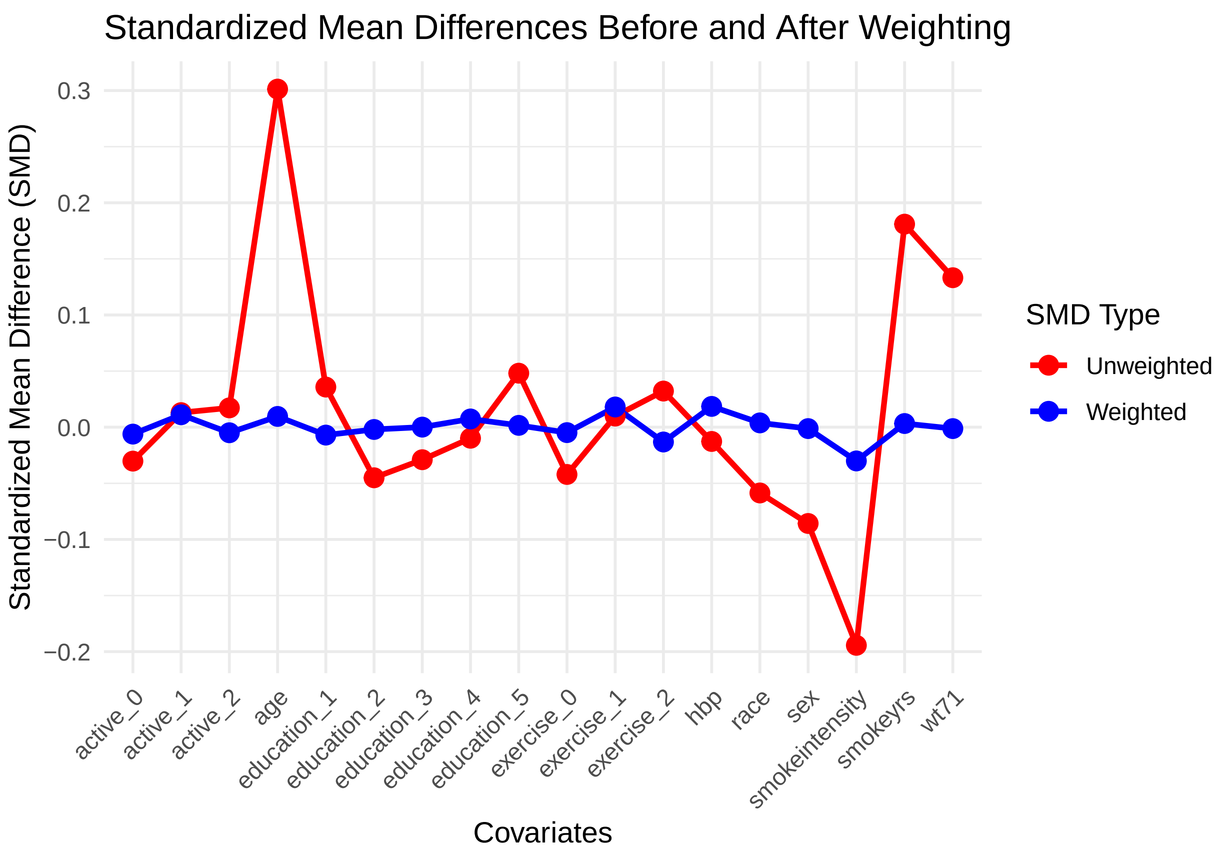
```
##
## Call:
## glm(formula = qsmk ~ sex + race + age + education + smokeintensity +
##     smokeyrs + exercise + active + wt71 + hbp, family = binomial(),
##     data = nhefs)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.404687   0.469988  -5.116 3.11e-07 ***
## sex1           -0.491585   0.141912  -3.464 0.000532 ***
## race1          -0.786979   0.202444  -3.887 0.000101 ***
## age             0.047381   0.009649   4.910 9.08e-07 ***
## education2     -0.136395   0.188772  -0.723 0.469962
## education3     -0.018052   0.168621  -0.107 0.914744
## education4     -0.013412   0.261659  -0.051 0.959119
## education5      0.368661   0.218824   1.685 0.092039 .
## smokeintensity -0.024236   0.005468  -4.432 9.32e-06 ***
## smokeyrs       -0.027809   0.009740  -2.855 0.004301 **
## exercise1       0.292360   0.172537   1.694 0.090175 .
## exercise2       0.380230   0.179278   2.121 0.033931 *
## active1         0.016889   0.129334   0.131 0.896102
## active2         0.063898   0.208581   0.306 0.759343
## wt71            0.006341   0.004133   1.534 0.124944
## hbp             0.022646   0.062330   0.363 0.716356
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1876.3  on 1628  degrees of freedom
## Residual deviance: 1780.5  on 1613  degrees of freedom
## AIC: 1812.5
##
## Number of Fisher Scoring iterations: 4
```



Part b

The following logistic regression model includes all of the covariates from part a and squared and logarithmic terms of each of the continuous covariates.

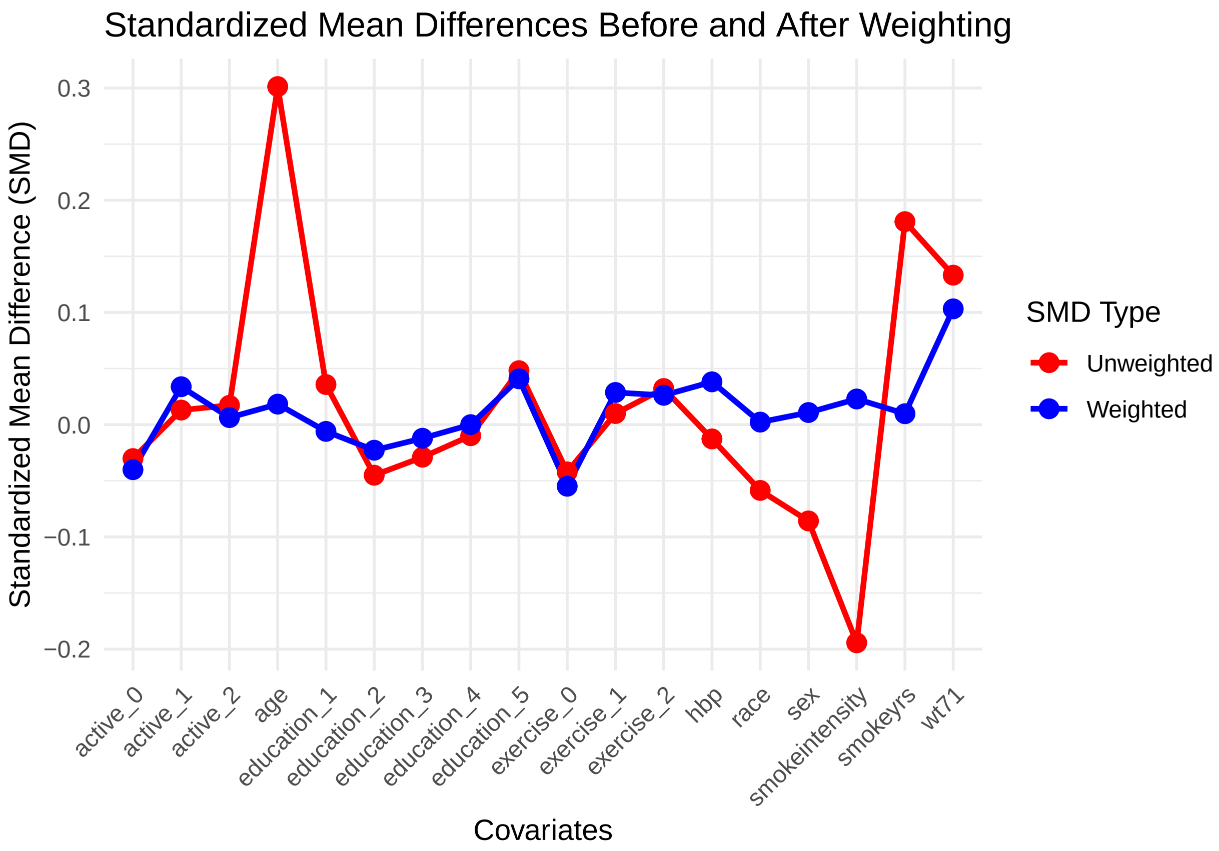
```
##
## Call:
## glm(formula = formula, family = binomial(), data = nhefs)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            19.3425788 27.2791477   0.709 0.478287
## sex1                   -0.5247104  0.1495193  -3.509 0.000449 ***
## race1                  -0.8449665  0.2068143  -4.086  4.4e-05 ***
## age                     0.0140384  0.3952226   0.036 0.971665
## age_squared            -0.0001913  0.0021961  -0.087 0.930571
## log_age                 2.3379913  8.5016565   0.275 0.783313
## education2             -0.0765825  0.1919260  -0.399 0.689878
## education3              0.0238010  0.1717569   0.139 0.889787
## education4             -0.0407497  0.2659140  -0.153 0.878206
## education5              0.3763507  0.2217205   1.697 0.089619 .
## smokeintensity         -0.0284073  0.0451235  -0.630 0.528993
## smokeintensity_squared  0.0004460  0.0005293   0.843 0.399465
## log_smokeintensity     -0.3294653  0.3723330  -0.885 0.376228
## smokeyrs               -0.1617775  0.0777325  -2.081 0.037415 *
## smokeyrs_squared        0.0017412  0.0008508   2.047 0.040694 *
## log_smokeyrs            0.8581547  0.7164581   1.198 0.231005
## exercise1               0.3087091  0.1741477   1.773 0.076281 .
## exercise2               0.3762316  0.1809236   2.080 0.037571 *
## active1                -0.0113069  0.1311830  -0.086 0.931314
## active2                 0.0563246  0.2108349   0.267 0.789353
## wt71                    0.2271321  0.1725681   1.316 0.188111
## wt71_squared           -0.0005836  0.0005298  -1.102 0.270666
## log_wt71               -9.7007508  6.7872876  -1.429 0.152932
## hbp                     0.0228222  0.0628670   0.363 0.716587
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1876.3  on 1628  degrees of freedom
## Residual deviance: 1761.7  on 1605  degrees of freedom
## AIC: 1809.7
##
## Number of Fisher Scoring iterations: 4
```



Part c

Here, all of the covariates from part a and each of the pairwise interaction terms between covariates are candidates for the final logistic regression model which is selected using stepwise feature selection with BIC as the information criterion.

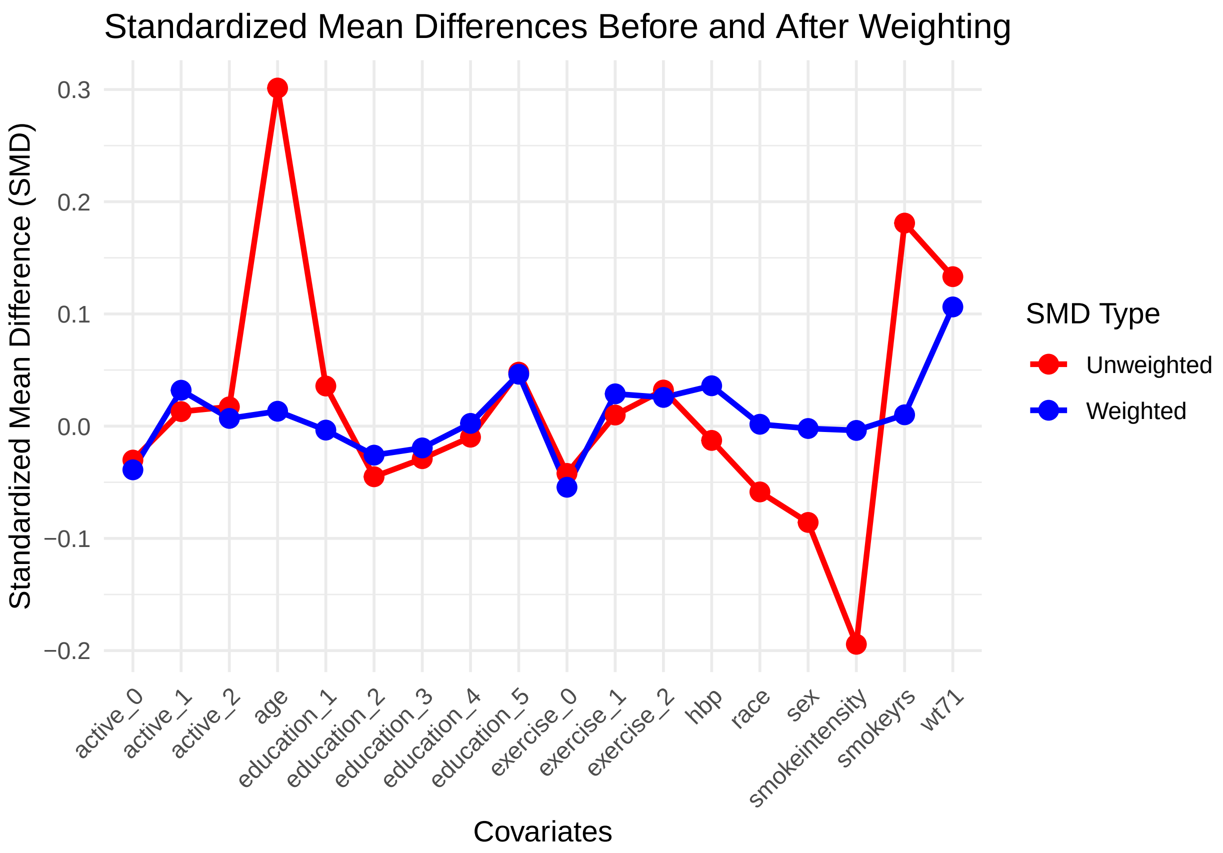
```
##
## Call:
## glm(formula = qsmk ~ sex + race + age + smokeintensity + smokeyrs +
##     sex:smokeintensity, family = binomial(), data = nhefs)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.886799   0.301662  -6.255 3.98e-10 ***
## sex1                0.050593   0.241092   0.210 0.833784
## race1              -0.734872   0.195525  -3.758 0.000171 ***
## age                 0.047945   0.009569   5.010 5.43e-07 ***
## smokeintensity     -0.012080   0.006563  -1.841 0.065658 .
## smokeyrs           -0.028403   0.009643  -2.945 0.003224 **
## sex1:smokeintensity -0.032961   0.011300  -2.917 0.003537 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1876.3  on 1628  degrees of freedom
## Residual deviance: 1786.4  on 1622  degrees of freedom
## AIC: 1800.4
##
## Number of Fisher Scoring iterations: 4
```



Part d

Here, all of the covariates from part a are candidates for the final logistic regression model which is selected using stepwise feature selection with BIC as the information criterion.

```
##
## Call:
## glm(formula = qsmk ~ sex + race + age + smokeintensity + smokeyrs,
##     family = binomial(), data = nhefs)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.651897   0.289709  -5.702 1.18e-08 ***
## sex1           -0.558627   0.125483  -4.452 8.51e-06 ***
## race1          -0.739328   0.195716  -3.778 0.000158 ***
## age             0.049489   0.009494   5.213 1.86e-07 ***
## smokeintensity -0.023712   0.005404  -4.388 1.15e-05 ***
## smokeyrs       -0.029979   0.009580  -3.129 0.001752 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1876.3  on 1628  degrees of freedom
## Residual deviance: 1795.2  on 1623  degrees of freedom
## AIC: 1807.2
##
## Number of Fisher Scoring iterations: 4
```



Part e

I would use the model from part d. The models from part a and b both have several features that have several covariates that are not significant at the p<0.05 level because there was no feature selection method used to create these models. The model from part c also has two covariates that are not significant at the p<0.05 level despite the covariates being selected using stepwise feature selection and it still has a lower AIC than that of part d.

The weighted SMDs of the model from part d are slightly higher than those of the models from part a and b, but I believe it is more important not to include several statistically insignificant covariates so the model generalizes better to new data than to have weighted SMDs closer to 0 given that none of the weighted SMDs are concerningly large. Only one of the weighted SMDs of the model from part d is above 0.1 (for the wt_71 covariate), and that weighted SMD is less than 0.15, so it is not an enormous concern.

Problem 2

Part i

```
## Estimated ATE, Standard Error, and 95% confidence interval of quitting smoking on change in weight fr
##  - Estimated ATE:  3.16
##  - Standard Error:  1.05
##  - 95% Confidence Interval: [ 1.11 ,  5.22 ]
##
## Estimated ATE, Standard Error, and 95% confidence interval of quitting smoking on systolic blood pres
##  - Estimated ATE:  1.61
##  - Standard Error:  2.52
##  - 95% Confidence Interval: [ -3.32 ,  6.55 ]
##
## Estimated ATE, Standard Error, and 95% confidence interval of quitting smoking on diastolic blood pr
##  - Estimated ATE:  1.3
##  - Standard Error:  1.47
##  - 95% Confidence Interval: [ -1.59 ,  4.18 ]
```

Part ii

```
## IPW1 Results:
```

```
## Estimated ATE, Standard Error, and 95% confidence interval of quitting smoking on change in weight f
##  - Estimated ATE:  3.23
##  - Standard Error:  0.41
##  - 95% Confidence Interval: [ 2.42 ,  4.03 ]
##
## Estimated ATE, Standard Error, and 95% confidence interval of quitting smoking on systolic blood pres
##  - Estimated ATE:  1.46
##  - Standard Error:  0.97
##  - 95% Confidence Interval: [ -0.44 ,  3.36 ]
##
## Estimated ATE, Standard Error, and 95% confidence interval of quitting smoking on diastolic blood pr
##  - Estimated ATE:  1.4
##  - Standard Error:  0.54
##  - 95% Confidence Interval: [ 0.35 ,  2.46 ]
```

```
## IPW2 Results:
```

```
## Estimated ATE, Standard Error, and 95% confidence interval of quitting smoking on change in weight f
##  - Estimated ATE:  3.23
##  - Standard Error:  0.45
##  - 95% Confidence Interval: [ 2.34 ,  4.11 ]
##
## Estimated ATE, Standard Error, and 95% confidence interval of quitting smoking on systolic blood pres
##  - Estimated ATE:  1.46
##  - Standard Error:  1.11
##  - 95% Confidence Interval: [ -0.71 ,  3.63 ]
##
## Estimated ATE, Standard Error, and 95% confidence interval of quitting smoking on diastolic blood pr
##  - Estimated ATE:  1.4
##  - Standard Error:  0.62
##  - 95% Confidence Interval: [ 0.19 ,  2.61 ]
```