Stats 506, F18, Problem Set 2

Zi Wang, tlwangzi@umich.edu

10/19/2018

Question 1

Table 1: Estimates and 95% confidence intervals of the four national means for residential energy consumption

Total	Estimate	lwr of 95% CI	upr of 95% CI
Electricity usage	10720.36100	10492.71473	10948.00727
Natural gas usage	226.53752	213.48458	239.59046
Propane usage	33.42942	25.26796	41.59088
Fuel oil or kerosene usage	28.60146	23.81263	33.39029

Question 2

a.

Use "import" command to read both data sets into Stata Use "merge" command to merge them together by the participant id SEQN.

Stata code:
import sasxport "https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/OHX_D.XPT"
save oral, replace
import sasxport "https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DEMO_D.XPT"
merge 1:1 seqn using oral
save syn, replace

b.

The summary of the model:

	(1)
VARIABLES	Model sign ridagemn
ridagemn	0.0697***
	(0.00257)
Constant	-8.359***
	(0.323)
Observations	7,563

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The BIC of the model: 1533.407

The age at which 25% of individuals lose their primary upper right 2nd bicuspid: 104

The age at which 50% of individuals lose their primary upper right 2nd bicuspid: 120

The age at which 75% of individuals lose their primary upper right 2nd bicuspid: 136

The range of representative age values: 8, 9, 10, 11, 12

c.

Add gender to the model and show the summary:

	(1)
VARIABLES	Model ridagemn riagendr
ridagemn	0.0697***
	(0.00257)
2.riagendr	0.0702
	(0.132)
Constant	-8.397***
	(0.332)
Observations	7,563

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The BIC: 1542.055
It is higher than 1533.407, so do not retain gender.

Add category Mexican American to the model and show the summary:

VARIABLES	(1) Model sign ridagemn+mex
ridagemn	0.0704***
	(0.00265)
1.mex	0.0386
	(0.143)
Constant	-8.459***
	(0.337)
Observations	7,246

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The BIC: 1542.285
It is higher than 1533.407, so do not retain category Mexican American.

Add category Non-Hispanic Black to the model and show the summary:

VARIABLES	(1) Model sign ridagemn+black
ridagemn	0.0701***
	(0.00259)
1.black	0.521***
	(0.145)
Constant	-8.567***
	(0.334)
Observations	7,563

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The BIC: 1529.281
It is lower than 1533.407, so retain category Non-Hispanic Black.

Add category Other to the model and show the summary:

VARIABLES	(1) Model sign ridagemn+black+other
ridagemn	0.0703***
	(0.00260)
1.black	0.567***
	(0.149)
1.other	0.337
	(0.233)
Constant	-8.636***
	(0.339)
Observations	7,563

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The BIC: 1536.103
It is higher than 1529.281, so do not retain category Other.

Add poverty income ratio to the model:

VARIABLES	(1) Model sign ridagemn+black+indfmpir
ridagemn	0.0714***
	(0.00271)
1.black	0.495***
	(0.149)
indfmpir	-0.119***
	(0.0454)
Constant	-8.460***
	(0.351)
Observations	7,246

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The BIC: 1462.895
It is lower than 1529.281, so retain poverty income ratio.

The final model:

VARIABLES	(1) Model sign ridagemn+black+indfmpir
ridagemn	0.0714***
	(0.00271)
1.black	0.495***
	(0.149)
indfmpir	-0.119***
	(0.0454)
Constant	-8.460***
	(0.351)
Observations	7,246

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

d.

1.

Adjusted Predictions at the mean (for other values) at the representative ages (From 8 to 12):

VARIABLES	(1) Adjusted predictions
1bn._at	0.146***
	(0.0128)
2._at	0.287***
	(0.0167)
3._at	0.486***
	(0.0174)
4._at	0.690***
	(0.0155)
5._at	0.840***
	(0.0118)
Observations	7,246

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The plot can be seen from the corresponding "marginsplot" command in ps2_q2.do

2.

The marginal effects at the mean of black at the same representative ages:

VARIABLES	(1) The marginal effects at the mean	(2) The marginal effects at the mean
1bno._at	0	
	(0)	
2o._at	0	
	(0)	
3o._at	0	
	(0)	
4o._at	0	
	(0)	
5o._at	0	
	(0)	
1bn._at		0.0668***
		(0.0217)
2._at		0.106***
		(0.0328)
3._at		0.123***
		(0.0365)
4._at		0.101***
		(0.0290)
5._at		0.0616***
		(0.0175)
Observations	7,246	7,246

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The plot can be seen from the corresponding "marginsplot" command in ps2_q2.do

3.

Average Marginal Effect of black at the representative ages:

	(1)	(2)
VARIABLES	Average marginal effect	Average marginal effect
1bno._at	0	
	(0)	
2o._at	0	
	(0)	
3o._at	0	
	(0)	
4o._at	0	
	(0)	
5o._at	0	
	(0)	
1bn._at		0.0671***
		(0.0217)
2._at		0.105***
		(0.0326)
3._at		0.122***
		(0.0363)
4._at		0.100***
		(0.0289)
5._at		0.0619***
		(0.0176)
Observations	7,246	7,246

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The plot can be seen from the corresponding "marginsplot" command in ps2_q2.do

e.

Compare the two model:

	(1)	(2)
VARIABLES	Model sign ridagemn+black+indfmpir	Model svy:sign ridagemn+black+indfmpir
ridagemn	0.0714***	0.0619***
	(0.00271)	(0.00723)
1.black	0.495***	
	(0.149)	
indfmpir	-0.119***	-0.0812
	(0.0454)	(0.0522)
black		0.543***
		(0.146)
Constant	-8.460***	-7.516***
	(0.351)	(0.862)
Observations	7,246	7,246

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Comments on the differences:

The coefficient of the variable ridagemn is more signficant from the model using svy than the other one.

The coefficient of the variable black is less signficant from the model using svy than the other one.

The coefficient of the variable indfmpir is more signficant from the model using svy than the other one.

The coefficient of the constant is more signficant from the model using svy than the other one.

Reason: the coefficient with lower p-value is more significant.

Question 3

a.

```
health = read.xport("OHX_D.XPT")
demo = read.xport("DEMO_D.XPT")
syn = demo %>%
  left_join(health,by = 'SEQN')
```

b.

The original model:

```
log_reg = glm(sign ~ RIDAGEMN,family = binomial(link = 'logit'),
              data = syn_select)
```

The summary of the model:

```
summary(log_reg)
```

```
##
## Call:
## glm(formula = sign ~ RIDAGEMN, family = binomial(link = "logit"),
##     data = syn_select)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8910   0.0000   0.0000   0.0498   2.8962
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.359363   0.323490  -25.84   <2e-16 ***
## RIDAGEMN     0.069678   0.002566   27.16   <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: 5763.4  on 7562  degrees of freedom
## Residual deviance: 1515.5  on 7561  degrees of freedom
## AIC: 1519.5
##
## Number of Fisher Scoring iterations: 10
```

The BIC of the model:

```
## [1] 1533.407
```

The age at which 25% of individuals lose their primary upper right 2nd bicuspid:

```
## (Intercept)
##         104
```

The age at which 50% of individuals lose their primary upper right 2nd bicuspid:

```
## (Intercept)
##         120
```

The age at which 75% of individuals lose their primary upper right 2nd bicuspid:

```
## (Intercept)
##         136
```

The lower bound and upper bound of the range of representative age values:

```
## (Intercept) (Intercept)
##           8          12
```

c.

Add gender to the model and show the summary:

```
log_reg_gen = glm(sign ~ RIDAGEMN + RIAGENDR,
                  family = binomial(link = 'logit'),data = syn_select)
```

```
summary(log_reg_gen)
```

```
##
## Call:
## glm(formula = sign ~ RIDAGEMN + RIAGENDR, family = binomial(link = "logit"),
##     data = syn_select)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8977   0.0000   0.0000   0.0507   2.9087
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.397415   0.332086 -25.287   <2e-16 ***
## RIDAGEMN        0.069698   0.002567  27.148   <2e-16 ***
## RIAGENDRFemale  0.070195   0.131971   0.532    0.595
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5763.4  on 7562  degrees of freedom
## Residual deviance: 1515.3  on 7560  degrees of freedom
## AIC: 1521.3
##
## Number of Fisher Scoring iterations: 10
```

the BIC:

```
## [1] 1542.055
```

It is higher than BIC_orig = 1533.407, so do not retain gender.

Add category Mexican American to the model and show the summary:

```
log_reg_W_Mex = glm(sign ~ RIDAGEMN + Mex,
                    family = binomial(link = 'logit'),data = syn_select_1)
```

```
summary(log_reg_W_Mex)
```

```
##
## Call:
## glm(formula = sign ~ RIDAGEMN + Mex, family = binomial(link = "logit"),
##     data = syn_select_1)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8892   0.0000   0.0000   0.0501   2.9000
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.370770   0.327455 -25.563   <2e-16 ***
## RIDAGEMN     0.069681   0.002566  27.157   <2e-16 ***
## Mex          0.032033   0.138808   0.231    0.817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5763.4  on 7562  degrees of freedom
## Residual deviance: 1515.5  on 7560  degrees of freedom
## AIC: 1521.5
##
## Number of Fisher Scoring iterations: 10
```

the BIC:

```
## [1] 1542.285
```

It is higher than BIC_orig = 1533.407, so do not retain category Mexican American.

Add category Non-Hispanic Black to the model and show the summary:

```
log_reg_W_Black = glm(sign ~ RIDAGEMN + Black,
                      family = binomial(link = 'logit'),data = syn_select_1)
```

```
summary(log_reg_W_Black)
```

```
##
## Call:
## glm(formula = sign ~ RIDAGEMN + Black, family = binomial(link = "logit"),
##     data = syn_select_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8805   0.0000   0.0000   0.0486   2.9347
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.567218   0.334198 -25.635  < 2e-16 ***
## RIDAGEMN     0.070075   0.002592  27.031  < 2e-16 ***
## Black        0.520727   0.145267   3.585 0.000338 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5763.4  on 7562  degrees of freedom
## Residual deviance: 1502.5  on 7560  degrees of freedom
## AIC: 1508.5
##
```

```
## Number of Fisher Scoring iterations: 10
```

the BIC:

```
BIC_W_Black
```

```
## [1] 1529.281
```

It is lower than BIC_orig = 1533.407, which means it improves the BIC, so retain category Non-Hispanic Black.

Add category Other to the model and show the summary:

```
log_reg_W_Black_Other = glm(sign ~ RIDAGEMN + Black + Other,
                            family = binomial(link = 'logit'),
                            data = syn_select_1)
```

```
summary(log_reg_W_Black_Other)
```

```
##
## Call:
## glm(formula = sign ~ RIDAGEMN + Black + Other, family = binomial(link = "logit"),
##     data = syn_select_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8805   0.0000   0.0000   0.0476   2.9540
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.636156   0.339150 -25.464  < 2e-16 ***
## RIDAGEMN     0.070262   0.002602  27.003  < 2e-16 ***
## Black        0.567418   0.149025   3.808  0.00014 ***
## Other        0.337474   0.232955   1.449  0.14743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5763.4  on 7562  degrees of freedom
## Residual deviance: 1500.4  on 7559  degrees of freedom
## AIC: 1508.4
##
## Number of Fisher Scoring iterations: 10
```

the BIC:

```
BIC_W_Black_Other
```

```
## [1] 1536.103
```

It is higher than BIC_W_Black = 1529.281, so do not retain category Other.

Add poverty income ratio(PIR) to the model:

```
log_reg_W_Black_PIR = glm(sign ~ RIDAGEMN + Black + INDFMPIR,
                          family = binomial(link = 'logit'),
                          data = syn_select_2)
```

```r
summary(log_reg_W_Black_PIR)
```

```
##
## Call:
## glm(formula = sign ~ RIDAGEMN + Black + INDFMPIR, family = binomial(link = "logit"),
##     data = syn_select_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9353   0.0000   0.0000   0.0458   2.8760
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.460288   0.351023 -24.102  < 2e-16 ***
## RIDAGEMN     0.071375   0.002706  26.374  < 2e-16 ***
## Black        0.494980   0.148923   3.324 0.000888 ***
## INDFMPIR    -0.119073   0.045378  -2.624 0.008689 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5534.6  on 7245  degrees of freedom
## Residual deviance: 1427.3  on 7242  degrees of freedom
## AIC: 1435.3
##
## Number of Fisher Scoring iterations: 10
```

the BIC:

```r
BIC_W_Black_PIR
```

```
## [1] 1462.895
```

It is lower than BIC_W_Black = 1529.281, which means it improves the BIC, so retain category poverty income ratio.

The final model: sign ~ RIDAGEMN + Black + INDFMPIR

```
##
## Call:
## glm(formula = sign ~ RIDAGEMN + Black + INDFMPIR, family = binomial(link = "logit"),
##     data = syn_select_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9353   0.0000   0.0000   0.0458   2.8760
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.460288   0.351023 -24.102  < 2e-16 ***
## RIDAGEMN     0.071375   0.002706  26.374  < 2e-16 ***
## Black        0.494980   0.148923   3.324 0.000888 ***
## INDFMPIR    -0.119073   0.045378  -2.624 0.008689 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5534.6  on 7245  degrees of freedom
## Residual deviance: 1427.3  on 7242  degrees of freedom
## AIC: 1435.3
##
## Number of Fisher Scoring iterations: 10
```

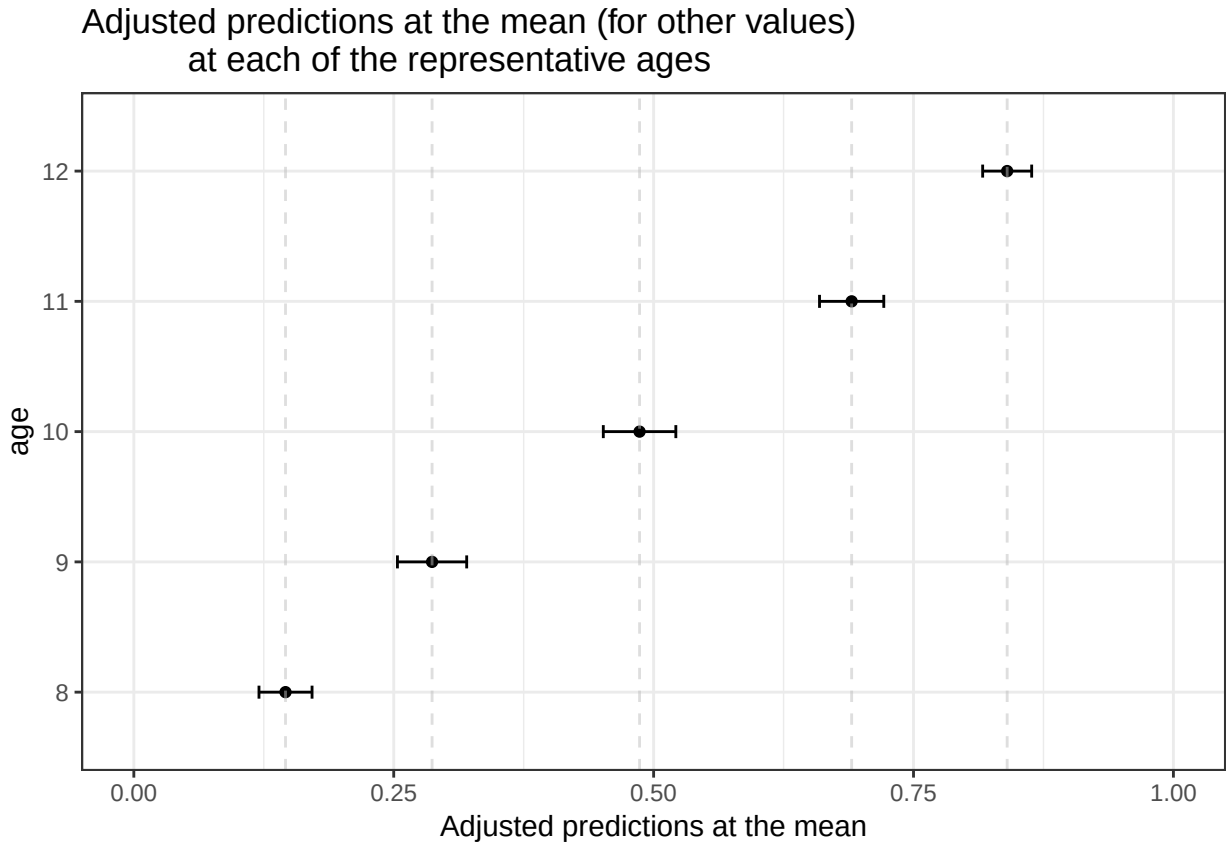
d.

1.

Adjusted Predictions at the mean (for other values) at the representative ages (From 8 to 12):

Table 2: Adjusted predictions at the mean (for other values) at each of the representative age

Age	Adjusted predictions at the mean
8	0.1459060
9	0.2868807
10	0.4864818
11	0.6904898
12	0.8400912

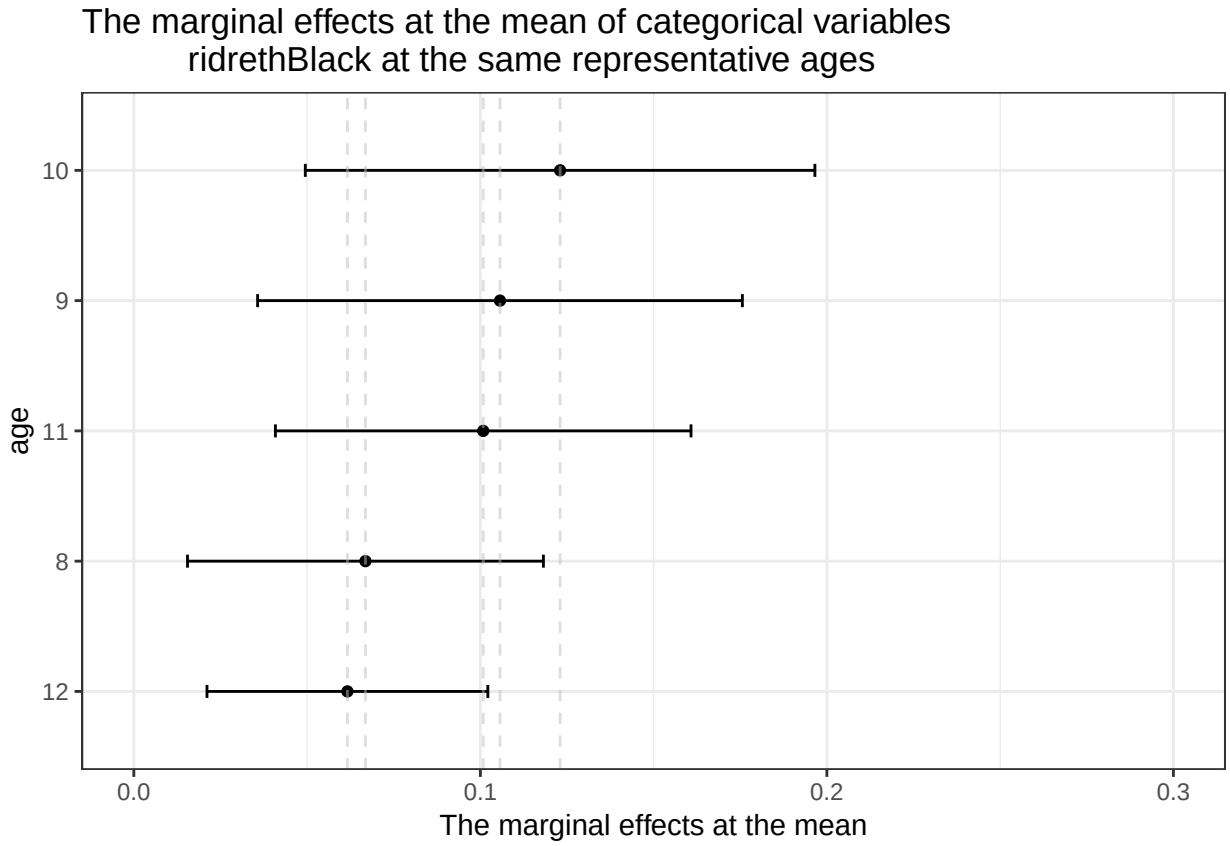


2.

The marginal effects at the mean of Black at the same representative ages:

Table 3: The marginal effects at the mean of categorical variables ridrethBlack at the same representative age

Age	The marginal effects at the mean
8	0.0668380
9	0.1056674
10	0.1230125
11	0.1008256
12	0.0616343



3.

Use "margins" command in the "margins" package with the "at" option:

```
q_tot = c(96,108,120,132,144)
mem0 = margins(log_reg_W_Black_PIR_1, at = list(RIDAGEMN = q_tot))
mem0
```

```
## Average marginal effects at specified values
## glm(formula = sign ~ RIDAGEMN + factor(Black) + INDFMPIR, family = binomial(link = "logit"),     data
##  at(RIDAGEMN) RIDAGEMN INDFMPIR  Black1
```

```
##            96 0.008987 -0.01499 0.06706
##           108 0.014430 -0.02407 0.10515
##           120 0.017427 -0.02907 0.12193
##           132 0.015068 -0.02514 0.10039
##           144 0.009676 -0.01614 0.06189
```

So, Average Marginal Effect of Black at the representative ages:

Table 4: The average marginal effect of Black at the representative ages

Age	The average marginal effect
8	0.06706
9	0.10515
10	0.12193
11	0.10039
12	0.06189