Stats 506, F18, Problem Set 2

Xun Wang, xunwang@umich.edu

October 13,2018

Question 1

Table 1: **Table 1.** *RECS2015 usage*

	Electricity/kwh	Natural gas/100 cubic feet	Propane/gallons	Fuel oil/gallons
estimate	1.267235e+12	39629221888	3951633152	3380928512
standard error	1.372967e+10	1032082304	492229504	288820928
CI_lower bound	1.240326e+12	38597139584	3459403648	3092107584
CI_upper bound	1.294145e+12	40661304192	4443862656	3669749440

Question 2: Including Plots

Part a

Using "fdause" command to import XPT files into Stata and then merging two files by "seqn", there are 8305 observatiosn left after deleting not matched observations.

Part b

First, we recode "permanent root fragments" as "permanent" and drop individuals for whom upper right 2nd bicuspid tooth was not assessed. Then we see "tooth not present" and "permanent" as categorical variable 1, and "primary tooth" as 0. Using logistic regression to estimate, we get the relationship between age (in months) and the probability that an 1 appears.

Table 2: **Table 2.** *Logistic Regression* Relationship between age and probability of tooth=1

variables		value
ridagemn	coefficient	0.069678***
	standard error	2.5658e-03
	z value	27.1564
	p-value	0
	confidence interval	0.064649 - 0.074707
Constant	coefficient	-8.35936***
	standard error	0.32349
	z value	-25.8408
	p-value	0
	confidence interval	-8.99340 - -7.72532

Because we select"tooth not present" and "permanent tooth" as 1, so get the regression equation:

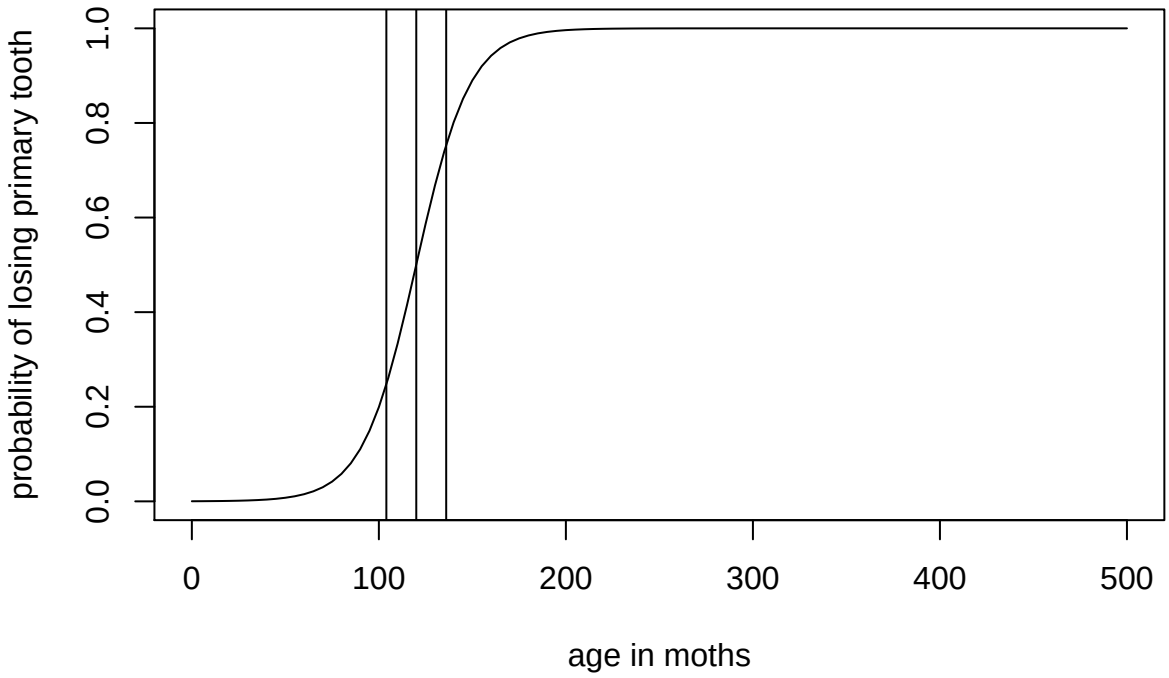
```
## LP= 0.069678 *ridagemn -8.35936

## log(P(lose primary)/P(primary))=LP

## log(P(primary)/P(lose primary))=-LP

## P(primary)=exp(-LP)+/(1+exp(-LP))

## P(primary)=exp(-0.069678*ridagemn+8.35936)/(1+exp(-0.069678*ridagemn+8.35936))
```

In the following, we use the fitted model in stata to estimate the ages at which 25, 50, and 75% of individuals who lose their primary upper right 2nd bicuspid. Then we round these results to the nearest month. the results in months are:

104,120,136 respectively.

We could get a range of representative age values (in year) by taking the floor of the smallest one and the ceiling of the largest one. That are:

8,9,10,11,12 years old.



Part c

First, we add gender variable into the existing model. After fitting, we get BIC of this new model is 1542.055, which is bigger than the the BIC of the original model(1533.407). So, we do not add gender into the model.

Then, we want to improve the model by adding race variable. We manipulate these data by creating indictors for four kind of races: Mexican American, Non-Hispanic White, Non-Hispanic Black and Other Race. We find that White has the largest population, so we see white race as reference. After that, we add other three races into model one by one and test their BIC. The results are as following:

Mexican American: BIC=1542.285

Other Race: BIC=1541.932

Non-Hispanic Black: BIC=1529.281<1533.407

So, we add black as a categorical variable into the model.

At last, we come to test the poverty income ratio in demographic sheet. BIC of the final model equals 1462.895<1529.281. So, the poverty income ratio should be retained in the final model.

The regression table is as following:

Table 3: **Table 3.** *Logistic Regression* Regression table of the final model

variables		value
ridagemn	coefficient	0.071375***
	standard error	2.7063e-03

variables		value
	z value	26.3735
	p-value	0
	confidence interval	0.066070 - 0.076679
1.black	coefficient	0.49498***
	standard error	0.14892
	z value	3.32371
	p-value	8.8830e-04
	confidence interval	0.20309 - 0.78687
indfmpir	coefficient	-0.11907***
	standard error	0.045378
	z value	-2.62402
	p-value	8.6898e-03
	confidence interval	-0.20801 - -0.030134
Constant	coefficient	-8.46029***
	standard error	0.35103
	z value	-24.1013
	p-value	0
	confidence interval	-9.14829 - -7.77228

Part d

In Stata, we could get the adjusted predictions and marginal effects simply by using "margins" command.

First, we should keep in mind that the representative ages determined in part b are 8,9,10,11,12 years old, which equals 96,108,120,132,144 months in our model.

Use margins, we could calculate the Adjusted predctions at the mean: at the mean of each variables and at each of the representative ages and produnce a table by outreg2 command. We import this table as below:

Table 4: **Table 4.** *Logistic Regression* Regression table of Adjusted predctions at the mean. At the mean of black and poverty income ratio, at representative age

age/years		Adjusted predictions
8_at	coefficient	0.14591***
	standard error	0.012762
	z value	11.4332
	p-value	0
	confidence interval	0.12089 - 0.17092
9_at	coefficient	0.28688***
	standard error	0.016653
	z value	17.2269
	p-value	0
	confidence interval	0.25424 - 0.31952
10_at	coefficient	0.48648***
	standard error	0.017442
	z value	27.8915
	p-value	0
	confidence interval	0.45230 - 0.52067
11_at	coefficient	0.69049***
	standard error	0.015459
	z value	44.6655
	p-value	0

age/years		Adjusted predictions
	confidence interval	0.66019 - 0.72079
12_at	coefficient	0.84009***
	standard error	0.011788
	z value	71.2678
	p-value	0
	confidence interval	0.81699 - 0.86319

Use margins,dydx command, we could calculate the marginal effects at the mean of retained black categorical variables and at the same representative ages as part b.The table of marginal effects at the mean:

Table 5: **Table 5.** *Logistic Regression* Regression table of Marginal effects at the mean. At the mean of poverty income ratio, at 0 and 1 of black,at representative age

age/months		margins
8_at	coefficient	0.066838***
	standard error	0.021677
	z value	3.08332
	p-value	2.0471e-03
	confidence interval	0.024351 - 0.10932
9_at	coefficient	0.10567***
	standard error	0.032775
	z value	3.22400
	p-value	1.2641e-03
	confidence interval	0.041429 - 0.16991
10_at	coefficient	0.12301***
	standard error	0.036534
	z value	3.36704
	p-value	7.5979e-04
	confidence interval	0.051407 - 0.19462
11_at	coefficient	0.10083***
	standard error	0.028985
	z value	3.47849
	p-value	5.0425e-04
	confidence interval	0.044015 - 0.15764
12_at	coefficient	0.061634***
	standard error	0.017514
	z value	3.51908
	p-value	4.3305e-04
	confidence interval	0.027307 - 0.095962

Also use margins,dydx command, we could calculate the average marginal effects of black categorical variables and at therepresentative ages 96,108,120,132,144 months.

Table 6: **Table 6.** *Logistic Regression* Regression table of Average marginal effects. At 0 and 1 of black,at representative age

age/months		margins
8_at	coefficient	0.067064***
	standard error	0.021696

age/months		margins
	z value	3.09108
	p-value	1.9943e-03
	confidence interval	0.024541 - 0.10959
9_at	coefficient	0.10515***
	standard error	0.032627
	z value	3.22287
	p-value	1.2691e-03
	confidence interval	0.041205 - 0.16910
10_at	coefficient	0.12193***
	standard error	0.036301
	z value	3.35900
	p-value	7.8224e-04
	confidence interval	0.050786 - 0.19308
11_at	coefficient	0.10039***
	standard error	0.028907
	z value	3.47281
	p-value	5.1504e-04
	confidence interval	0.043732 - 0.15705
12_at	coefficient	0.061892***
	standard error	0.017578
	z value	3.52096
	p-value	4.2999e-04
	confidence interval	0.027439 - 0.096345

Part e

Finally, we will refit the final model from part c using svy command. First, we need to point out how the survey was designed. Using svyset command, we tells stata to treat "wtmec2yr" as weights, "sdmvstra" as strata, "sdmvpsu" as primary sampling units. And the "vce(linearized)" means survey should has Taylor linearized variance estimation.

Then we get the regression table of the complexed model based on survey environment.

Table 7: **Table 7.** *Logistic Regression* Regression table of the final model with survey command

variables		value
ridagemn	coefficient	0.061941***
	standard error	7.2296e-03
	z value	8.56766
	p-value	3.6649e-07
	confidence interval	0.046531 - 0.077351
1.black	coefficient	0.54349***
	standard error	0.14619
	z value	3.71760
	p-value	2.0634e-03
	confidence interval	0.23189 - 0.85510
indfmpir	coefficient	-0.081181
	standard error	0.052192
	z value	-1.55543
	p-value	0.14069
	confidence interval	-0.19243 - 0.030064

variables		value
Constant	coefficient	-7.51602***
	standard error	0.86156
	z value	-8.72373
	p-value	2.9173e-07
	confidence interval	-9.35239 - -5.67964

The coefficients of these variables do change but they do not change too much. While, the p-value of poverty income ratio increases and this variable changes from significant to non-significant. I think the differences may because of the ignorance of survey design at first. The stratified sampling mechanism and the weights of the survey change the z-value and coefficients of these variables.

Question 3

Part a

Using "read.xport" command to import XPT files into R and then merging two files by "SEQN", there are 8305 observatiosn left after deleting not matched observations.

Part b

First, we recode "permanent root fragments" as "permanent" and drop individuals for whom upper right 2nd bicuspid tooth was not assessed. Then we see "tooth not present" and "permanent" as categorical variable 1, and "primary tooth" as 0. Using logistic regression to estimate, we get the relationship between age (in months) and the probability that 1 appears.

Table 8: **Table 8.** *Logistic Regression* Relationship between age and probability of tooth=1

term	estimate	std.error	statistic	p-value
(Intercept)	-8.35936	0.32349	-25.84117	0
RIDAGEMN	0.06968	0.00257	27.15687	0

Because we select"tooth not present" and "permanent tooth" as 1, so get the regression equation as the first equation as below. If we want to know the relationship between age (in months) and the probability of "primary",we need to have an extra step.

```
## LP= 0.06967781 *ridagemn -8.359363
```

```
## log(P(lose primary)/P(primary))=LP
```

```
## log(P(primary)/P(lose primary))=-LP
```

```
## P(primary)=exp(-LP)+/(1+exp(-LP))
```

```
## P(primary)=exp( -0.06968 *ridagemn+ 8.35936 )/(1+exp( -0.06968 *ridagemn+ 8.35936 ))
```

In the following, we use the fitted model in R to estimate the ages at which 25, 50, and 75% of individuals who lose their primary upper right 2nd bicuspid. Then we round these results to the nearest month. The results in months are:

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

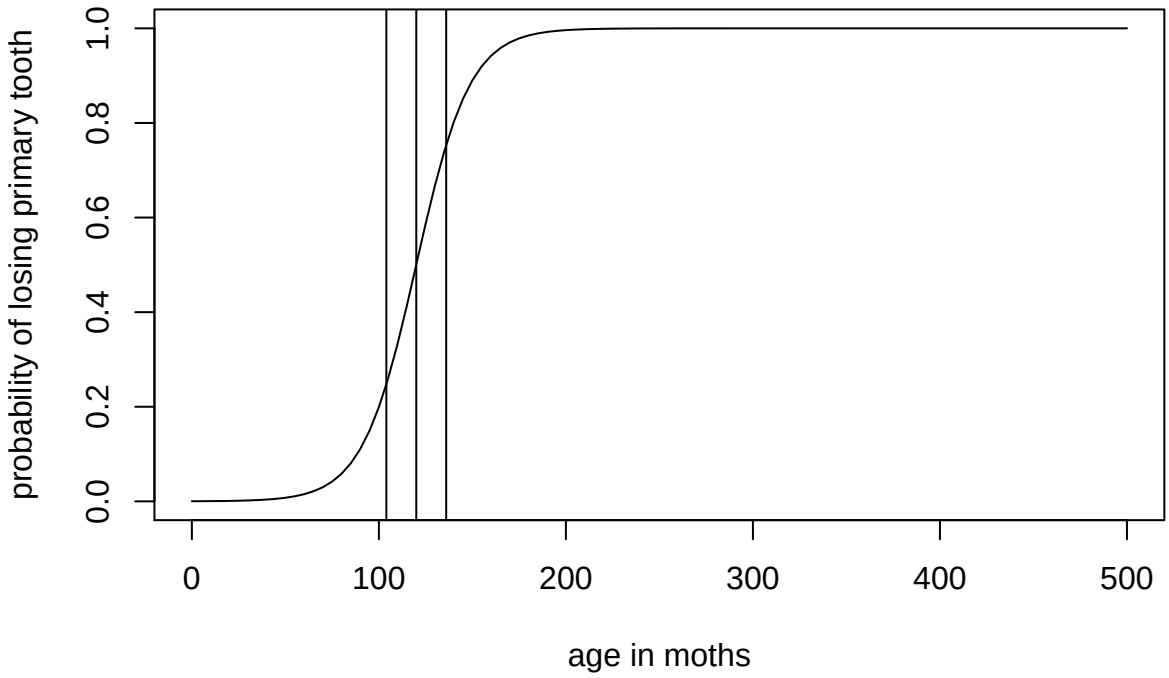
```
## (Intercept)
##         120
```

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

We could get a range of representative age values (in year) by taking the floor of the smallest one and the ceiling of the largest one. That are:

```
## [1]  8  9 10 11 12
```

Then we show the plot of this relationship and 25,50,75% quantile.



Part c

First, we add gender variable into the existing model. After fitting, we get BIC of this new model is:

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

which is bigger than the the BIC of the original model:

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

So, we do not add gender into the model.

Then, we want to improve the model by adding race variable. We manipulate these data by creating indictors for four kind of races: Mexican American, Non-Hispanic White, Non-Hispanic Black and Other Race. We find that White has the largest population, so we see white race as reference. After that, we add other three races into model one by one and test their BIC. The results are as following:

```
## Mexican American: BIC= 1542.285
```

```
## Other Race: BIC= 1541.932
```

```
## Non-Hispanic Black: BIC= 1529.281 <1533.407
```

So, we add black as a categorical variable into the model.

At last, we come to test the poverty income ratio in demographic sheet. BIC of the final model equals:

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

which is smaller than the BIC value of 1529.281.

So, the poverty income ratio should be retained in the final model.

The regression table is as following:

Table 9: **Table 9.** *Logistic Regression* Regression table of the final model

term	estimate	std.error	statistic	p-value
(Intercept)	-8.46029	0.35102	-24.10180	0.00000
RIDAGEMN	0.07137	0.00271	26.37413	0.00000
black	0.49498	0.14892	3.32374	0.00089
INDFMPIR	-0.11907	0.04538	-2.62405	0.00869

Part d

First, we should keep in mind that the representative ages determined in part b is 8,9,10,11,12, which equals 96,108,120,132,144 in our model.

Use the predict function, we could calculate the Adjusted predctions at the mean: at the mean of each variables and at each of the representative ages and the result is:

Table 10: **Table 10.** *Logistic Regression* Regression table of Adjusted predctions at the mean. At the mean of black and poverty income ratio, at representative age

age/years	8_at	9_at	10_at	11_at	12_at
Adajusted Predictions	0.14591	0.28688	0.48648	0.69049	0.84009

Continuing, we could calculate the marginal effects at the mean of retained black categorical variables and at the same representative ages as part b.The table of marginal effects at the mean:

Table 11: **Table 11.** *Logistic Regression* Regression table of Marginal effects at the mean. At the mean of poverty income ratio, at 0 and 1 of black,at representative age

age/years	8_at	9_at	10_at	11_at	12_at
margins	0.06684	0.10567	0.12301	0.10083	0.06163

Also use predict command, we could calculate the average marginal effects of black categorical variables and at the representative ages 8,9,10,11,12 years.

Table 12: **Table 12.** *Logistic Regression* Regression table of Average marginal effects. At 0 and 1 of black,at representative age

age/years	8_at	9_at	10_at	11_at	12_at
margins	0.06706	0.10515	0.12193	0.10039	0.06189