Stats 506, F18, Problem Set 2

Xinye Jiang (xinyej@umich.edu)

October 16, 2018

Question 1

The table and figure below show estimates and their 95% confidence intervals of national totals for for residential energy consumption about electricity usage in kilowatt hours, natural gas usage in hundreds of cubic feet, propane usage in gallons and kerosene usage in gallons. We can see that the nation consumes more propane than kerosene.

Table 1: **Table 1.** *National totals for residential energy consumption.* Each row shows estimate and 95% confidence interval.

	Estimate	lwr	upr
Electricity usage in kilowatt hours	1.267235e+12	1.240325e+12	1.294145e+12
Natural gas usage in hundreds of cubic feet	3.962922e+10	3.760638e+10	4.165207e+10
Propane usage in gallons	3.951633e+09	2.986881e+09	4.916385e+09
Kerosene usage in gallons	3.380928e+09	2.814850e+09	3.947007e+09

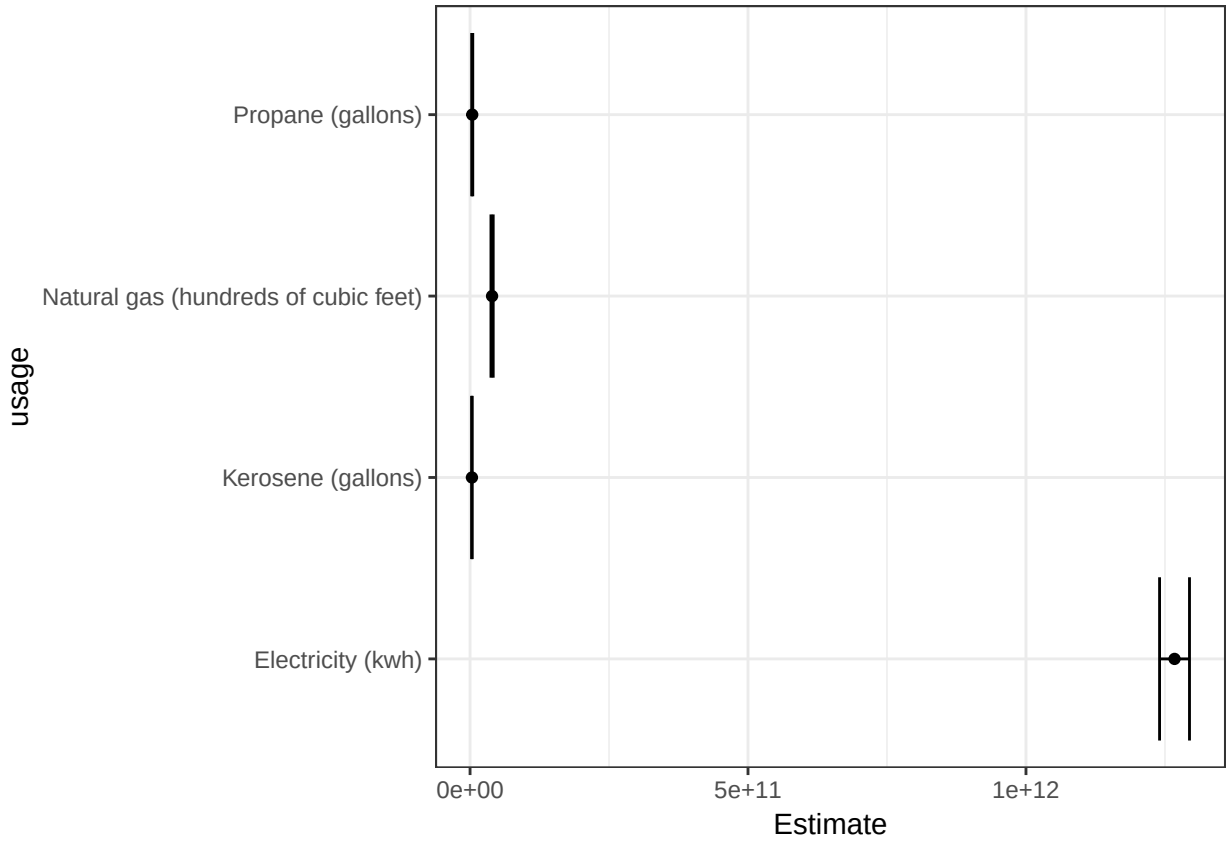

Figure 1: **Figure 1.** *National totals for residential energy consumption.*

Question 2

b.

Use logistic regression to estimate the relationship between age (in months) and the probability that an individual has a primary rather than a missing or permanent upper right 2nd bicuspid. The probability that an individual has a primary upper right 2nd bicuspid equals to 1 - the probability that an individual loses one. Fit a logisic regression over age and the probability that an individual loses a primary upper right 2nd bicuspid (denote it as ur2b) and I get a model: ur2b = -8.359362 + .0696778 * age . And we can see from the model that when age increases, individuals are more likely to lose their primary upper right 2nd bicuspid.

The estimated ages at which 25, 50 and 75% of individuals lose their primary upper right 2nd bicuspid by the fitted model (rounded to the nearest month) are:

```
## [1] 104 120 136
```

A range of representative age values with one year increments by taking the floor (in years) of the 25%-ile and the ceiling (in years) of the 75%-ile are:

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

c.

The final logistic regression model is 'ur2b ~ ridagemn(age) + i.race_black + indfmpir(pir)'. ('ur2b' means the probability that an individual loses a primary upper right 2nd bicuspid)

Table 2: **Table 2.** *Regression table for the final model.* Each row shows coefficient estimate, its 95% confidence interval and p-value.

UR2B	Coefficient	95% CI	pvalue
Intercept	-8.46029	(-9.14829, -7.77228)	0.000
age	0.07137	(0.06607, 0.07668)	0.000
race black	0.49498	(0.20309, 0.78687)	0.001
poverty income ratio	-0.11907	(-0.20801, -0.03013)	0.009

Model fitting process:

Table 3: **Table 3.** *Model fitting process.* Each row shows one step in the logistic regression model fitting process.

Step	Model	BIC	Improve or Not	Decision
1	ur2b ~ age	1533.407	/	Retain age
2	ur2b ~ age + gender	1542.055	No	Not retain gender
3	ur2b ~ age + race_mexican	1542.285	No	Not retain race_mexican
4	ur2b ~ age + race_black	1529.281	Yes	Retain race_black
5	ur2b ~ age + race_black + race_other	1536.103	No	Not retain race_other
6	ur2b ~ age + race_black + pir	1462.895	Yes	Retain pir

By the above steps described in Table 3, I get the final model. The variables 'race_black' and 'pir' improve

BIC, so I retain them in the final model. The variables 'gender', 'race_mexican' and 'race_other' do not improve BIC, so I do not retain them.

d.

d.1 Adjusted predictions at the mean (for other values) at each representative age:

Table 4: **Table 4.** *Adjusted predictions at the mean at each representative age.* Each row shows an adjusted prediction at the mean with its 95% CI at a representative age. Rows are sorted by age.

Age	Adjusted Prediction	95% CI
8	0.14591	(0.12089, 0.17092)
9	0.28688	(0.25424, 0.31952)
10	0.48648	(0.45230, 0.52067)
11	0.69049	(0.66019, 0.72079)
12	0.84009	(0.81699, 0.86319)

d.2 The marginal effects at the mean of any retained categorical variables at the representative ages:

Table 5: **Table 5.** *The marginal effect at the mean of race_black at each representative age.* Each row shows the marginal effect at the mean with its 95% CI at a representative age. Rows are sorted by age.

Age	Marginal effect at the mean	95% CI
8	0.06684	(0.02435, 0.10932)
9	0.10567	(0.04143, 0.16991)
10	0.12301	(0.05141, 0.19462)
11	0.10083	(0.04402, 0.15764)
12	0.06163	(0.02731, 0.09596)

d.3 The average marginal effects of any retained categorical variables at the representative ages:

Table 6: **Table 6.** *The average marginal effect of race_black at each representative age.* Each row shows the average marginal effect with its 95% CI at a representative age. Rows are sorted by age.

Age	Average marginal effect	95% CI
8	0.06706	(0.02454, 0.10959)
9	0.10515	(0.04120, 0.16910)
10	0.12193	(0.05079, 0.19308)
11	0.10039	(0.04373, 0.15704)
12	0.06189	(0.02744, 0.09634)

We can see that the marginal effects at the mean and the average marginal effects of any retained categorical

variables at each representative age are quite close.

e. Refit the final model from part c using 'svy'.

Table 7: **Table 7.** *Regression table for the final model refitted using svy.* Each row shows coefficient estimate, its 95% confidence interval and p-value.

UR2B	Coefficient	95% CI	pvalue
Intercept	-7.51602	(-9.35239, -5.67964)	0.000
age	0.06194	(0.04653, 0.07735)	0.000
race black	0.54349	(0.23189, 0.85510)	0.002
poverty income ratio	-0.08118	(-0.19243, 0.03006)	0.141

After refitting the final model from part c using "svy", the variable 'pir' becomes not statistically significant any more as compared to before, which can be seen from the change of its corresponding p-value. Before refitting, p-value of 'pir' equals to 0.009 and is smaller than 0.05, while after refitting, p-value of 'pir' equals to 0.141 and is bigger than 0.05. And almost all coefficients' 95% CIs have bigger range of values than before, which can been seen from the above table.

The differences are due to in part b-d we ignore the survey aspect of the data and analyze it as if the data are from a simple random sample, while in part e we consider the survey setting.

Question 3

b.

Use logistic regression to estimate the relationship between age (in months) and the probability that an individual has a primary rather than a missing or permanent upper right 2nd bicuspid. The probability that an individual has a primary upper right 2nd bicuspid equals to 1 - the probability that an individual loses one. Fit a logisic regression over age and the probability that an individual loses a primary upper right 2nd bicuspid (denote it as ur2b) and I get a model: ur2b = -8.359363 + .069678 * age . And we can see from the model that when age increases, individuals are more likely to lose their primary upper right 2nd bicuspid.

The estimated ages at which 25, 50 and 75% of individuals lose their primary upper right 2nd bicuspid by the fitted model (rounded to the nearest month) are listed below:

```
## [1] 104 120 136
```

A range of representative age values with one year increments by taking the floor (in years) of the 25%-ile and the ceiling (in years) of the 75%-ile are listed below:

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

c.

The final logistic regression model: ur2b ~ age + race_black + pir ('ur2b' means the probability that an individual loses a primary upper right 2nd bicuspid)

```
##
## Call:
## glm(formula = ur2b == "lost" ~ age + race_black + pir, family = binomial(link = "logit"),
##     data = oral_demo_c)
##
## 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 ***
## age             0.071375   0.002706  26.374  < 2e-16 ***
## race_blackTRUE  0.494980   0.148923   3.324 0.000888 ***
## pir            -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
```

Table 8: **Table 8.** *Regression table for the final model.* Each row shows coefficient estimate, its 95% confidence interval and p-value.

Coefficients	Estimate	95% CI	pvalue
Intercept	-8.46029	(-9.14828, -7.77230)	0.00000
age	0.07137	(0.06607, 0.07668)	0.00000
race black	0.49498	(0.20310, 0.78686)	0.00089
poverty income ratio	-0.11907	(-0.20801, -0.03013)	0.00869

Model fitting process:

Table 9: **Table 9.** *Model fitting process.* Each row shows one step in the logistic regression model fitting process.

Step	Model	BIC	Improve or Not	Decision
1	ur2b ~ age	1533.407	/	Retain age
2	ur2b ~ age + gender	1542.055	No	Not retain gender
3	ur2b ~ age + race_mexican	1542.285	No	Not retain race_mexican
4	ur2b ~ age + race_black	1529.281	Yes	Retain race_black
5	ur2b ~ age + race_black + race_other	1536.103	No	Not retain race_other
6	ur2b ~ age + race_black + pir	1462.895	Yes	Retain pir

By the steps described in the above table, I get the final model. The variables 'race_black' and 'pir' improve BIC, so I retain them in the final model. The variables 'gender', 'race_mexican' and 'race_other' do not improve BIC, so I do not retain them.

d.

d.1 Adjusted predictions at the mean (for other values) at each representative age:

Table 10: **Table 10.** *Adjusted predictions at the mean at each representative age.* Each row shows an adjusted prediction at the mean at a representative age. Rows are sorted by age.

Age	Adjusted Prediction
8	0.14591
9	0.28688
10	0.48648
11	0.69049
12	0.84009

d.2 The marginal effects at the mean of any retained categorical variables at the representative ages:

Table 11: **Table 11.** *The marginal effect at the mean of race_black at each representative age.* Each row shows the marginal effect at the mean at a representative age. Rows are sorted by age.

Age	Marginal effect at the mean of race_black
8	0.06684
9	0.10567
10	0.12301
11	0.10083
12	0.06163

d.3 The average marginal effects of any retained categorical variables at the representative ages:

Table 12: **Table 12.** *The average marginal effect of race_black at each representative age.* Each row shows the average marginal effect at a representative age. Rows are sorted by age.

Age	Average marginal effect of race_black
8	0.06706
9	0.10515
10	0.12193
11	0.10039
12	0.06189

We can see that the marginal effects at the mean and the average marginal effects of any retained categorical variables at each representative age are quite close.

Compare the outputs from Stata in question 2 and the outputs from R in question 3, we can see that the results are the same.