Problem Set 2, Solutions

Stats 506, Fall 2018

Due: Wednesday October 17, 5pm

Instructions

- You should submit a single compressed archive (`.zip`) which contains the following files:
  - `ps2.pdf` or `ps2.html`
  - `ps2.Rmd`
  - `ps2_q1.do`
  - `recs2015_usage.csv`
  - `ps2_q2.do`
  - `ps2_q2.log`
  - `ps2_q3.R`
- Each of the files above and an example `.zip` file is available at the Stats506_F18 (https://github.com/jbhender/Stats506_F18/) repo.

Question 1 [25 points]

Use Stata to estimate the following national *totals* for residential energy consumption:

- Electricity usage in kilowatt hours
- Natural gas usage, in hundreds of cubic feet
- Propane usage, in gallons
- Fuel oil or kerosene usage, in gallons

In your analysis, be sure to properly weight the individual observations. Use the replicate weights to compute standard errors. At the end of your `.do` file, write the estimates and standard errors to a delimited file `recs2015_usage.csv`.

In your `.Rmd` read `recs2015_usage.csv` and produce a nicely formatted table with estimates and 95% confidence intervals.

Solution: See `ps2_q1.do` for computations in Stata. It can be run from the command line using `stata -b ps2_q1.do`.

```
# Libraries: ----------------------------------------------------------------
library(tidyverse)

# Data: ---------------------------------------------------------------------
usage = readr::read_delim('./recs2015_usage.csv', delim = ',')

# Rescale to "billions" = 1e9: ----------------------------------------------
m = qnorm(.975)

usage = usage %>%
  mutate(
    total = total / 1e9,
    std_error = std_error / 1e9,
    lwr = total - m*std_error,
    upr = total + m*std_error
  )

# Table: --------------------------------------------------------------------
cap = '**Estimated US residential energy consumption in 2015.**'
usage %>%
  arrange( desc(total) ) %>%
  transmute(
     `Energy Source`=
       c('Electricity (billions of kwh)',
         'Natural Gas (billions of cubic feet)',
         'Propane (billions of gallons)',
         'Kerosene (billions of gallons)'
        ),
    `Residential Consumption` =
      sprintf('%6.1f (%.1f, %.1f)', total, lwr, upr)
  ) %>%
  knitr::kable( caption = cap, align = 'c'  )
```

Estimated US residential energy consumption in 2015.

Energy Source	Residential Consumption
Electricity (billions of kwh)	1267.2 (1240.3, 1294.1)
Natural Gas (billions of cubic feet)	39.6 (37.6, 41.7)
Propane (billions of gallons)	4.0 (3.0, 4.9)
Kerosene (billions of gallons)	3.4 (2.8, 3.9)

Question 2 [35 points]

For this question you should use the 2005-2006 NHANES ORAL Health data available here (https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&CycleBeginYear=2005) and the demographic data available here (https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2005). Your analyses for this question should be done in Stata, though you may create plots and format tables using R within Rmarkdown.

For part (b-d), you can ignore the survey aspect of the data and analyze it as if the data were a simple random sample.

a. [5 points] Determine how to read both data sets into Stata and merge them together by the participant id SEQN.

b. [5 points] 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. You can recode permanent root fragments as permanent and drop individuals for whom this tooth was not assessed. Use the fitted model to estimate the ages at which 25, 50, and 75% of individuals lose their primary upper right 2nd bicuspid. Round these to the nearest month. Choose 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.

c. [10 points] In the regression above, control for demographics in the following way:
   - Add gender to the model and retain it if it improves the BIC.
   - Create indicators for each race/ethnicity category using the largest as the reference and collapsing 'Other Hispanic' and 'Other'. In order of group size in the sample, add each category retaining those that improve BIC.
   - Add poverty income ratio to the model and retain it if it improves BIC.

In your pdf document, include a nicely formatted regression table for the final model and an explanation of the model fitting process.

d. [10 points] Use the `margins` command to compute:

   1. Adjusted predctions at the mean (for other values) at each of the representative ages determined in part b.
   2. The marginal effects at the mean of any retained categorical variables at the same representative ages.
   3. The average marginal effect of any retained categorical varialbes at the representative ages.

e. [5 points] Refit your final model from part c using `svy` and comment on the differences. Include a nicely formatted regression table and cite evidence to justify your comments.

You should use the following command to set up the survey weights (ftp://ftp.cdc.gov/pub/health_statistics/nchs/tutorial/nhanes/Continuous/descriptive_mean.do):

```
svyset sdmvpsu [pweight=wtmec2yr], strata(sdmvstra) vce(linearized)
```

Solution: The Stata code for this question is available as `ps2_q2.do` and its output as `ps2_q2.log`. Additional files output are documented below: - -

a. See section "a" of `ps2_q2.do`. After merging we have 8,305 matches. We then drop cases with missing data:

- 623 with `tooth4` (OHX04HTC) missing or not not assessed,
- 119 with missing age in months (RIDAGEMN)

This leaves 8,305 - 742 = 7,563 cases with complete data on key variables. Additional cases have missing poverty income ratio (`pir` or INDFMPIR). We will make a decision about these cases in part c.

b. The nearest whole ages that span the unadjusted 25 and 75 percentiles of the estimated model for the absence of a primary upper right 2nd bicuspid are ages 8 and 12. We will use as representative ages (8, 9, 10, 11, 12)*12 months. See the script for computation details.

c. In this section we first create indicators for racial/ethinic groups based on RIDRETH1. Specifically, we always include 'Non-Hispanic White' in the reference group and create indicators for (listed in order of group size):
   - 'Non-Hispanic Black'

- ‘Mexican American’
- ‘Other race’ collapsing ‘Other Race - Including Multi-Racial’ and ‘Other Hispanic’.

Starting from a base model with age as the only independent variable, we use forward selection to build the model using the BIC criterion to make decisions at each step. We consider individual terms in the following, fixed, order: - Gender - ‘Non-Hispanic Black’ - ‘Mexican American’ - ‘Other race’ - PIR.

After deciding on main effects we consider interactions for retained variables with age. The reults are

```
estat_out = readxl::read_excel('./ps2_q2_c.xls',
  col_names = c('Model', 'Observations', 'loglik(null)', 'loglik(model)', 'df',
                'aic', 'bic') )

# Function to clean up the ouput for printing
clean_model = function(x){
  x = stringr::str_replace(x, "age", "Age (months),")
  x = stringr::str_replace(x, "gender", "Gender")
  x = stringr::str_replace(x, "i.black|black", "Non-Hispanic Black,")
  x = stringr::str_replace(x, "i.mexamer", "Mexican American")
  x = stringr::str_replace(x, "i.other", "Other race")
  x = stringr::str_replace(x, "pir", "Poverty Income Ratio")
  x = stringr::str_replace(x, ",$", "")
  x = stringr::str_replace(x, ",#", "#")
  x = stringr::str_replace(x, "c\\.", "")
  x = stringr::str_replace(x, "_cons", "Intercept")
  x
}

cap = '**Model Selection.** The sequence of independent variables considered
and associated BIC values are shown here. Observe that we use only the cases
without missing Poverty Income Ratio when deciding whether to include it in the
model.'
cap = stringr::str_replace(cap, '\n', ' ')

estat_out %>%
  transmute( Model = clean_model(Model), Observations, BIC = round(bic, 1)) %>%
  mutate( Decision = c('-', ifelse(diff(BIC)<0, 'Keep', 'Drop')),
          Decision = ifelse(grepl('reduced', Model), '-', Decision),
          Observations = format(Observations, big.mark = ',')
  ) %>%
  knitr::kable( caption = cap )
```

Model Selection. The sequence of independent variables considered and associated BIC values are shown here. Observe that we use only the cases without missing Poverty Income Ratio when deciding whether to include it in the model.

Model	Observations	BIC	Decision
Age (months)	7,563	1533.4	-
Age (months), Gender	7,563	1542.1	Drop
Age (months), Non-Hispanic Black	7,563	1529.3	Keep
Age (months), Non-Hispanic Black, Mexican American	7,563	1533.1	Drop
Age (months), Non-Hispanic Black, Other race	7,563	1536.1	Drop

Model	Observations	BIC	Decision
Age (months), Non-Hispanic Black, (reduced)	7,246	1460.9	-
Age (months), Non-Hispanic Black, Poverty Income Ratio	7,246	1462.9	Drop
Age (months)##Non-Hispanic Black	7,563	1538.1	Drop

Below is a regression table for the final model.

```
q2_c_reg = readxl::read_excel('./ps2_q2_c_coef.xls',
                              skip = 1)
terms = names(q2_c_reg[,-1])
est = as.numeric(q2_c_reg[1,-1])
se = as.numeric(q2_c_reg[2,-1])
terms = clean_model(terms)
ind = which(!grepl('^0', terms))

or = sprintf('%4.2f (%4.2f, %4.2f)', exp(est), exp(est - m*se), exp(est + m*se))

terms = stringr::str_replace(terms, "^1\\.", "")

srs_reg = tibble(Variable = terms[ind], Coefficient = est[ind], SE = se[ind]) %>%
  mutate(
    OR = or[ind], OR = ifelse(grepl('Intercept', Variable), '-', OR)
  ) %>%
  rename( `OR (95% CI)` = OR)

cap = '**Coefficient table.** The table shows the fitted
coefficients for the final logistic regression model for the probability
of an absent primary tooth. In conrast to the earlier table, this verion comes from
using svy: logit to account for the complex survey weights.'
cap = stringr::str_replace(cap, '\n', ' ')

srs_reg %>% knitr::kable( digits = 2, caption = cap)
```

Coefficient table. The table shows the fitted coefficients for the final logistic regression model for the probability of an absent primary tooth. In conrast to the earlier table, this verion comes from using svy: logit to account for the complex survey weights.

Variable	Coefficient	SE	OR (95% CI)
Age (months)	0.07	0.00	1.07 (1.07, 1.08)
Non-Hispanic Black	0.52	0.15	1.68 (1.27, 2.24)
Intercept	-8.57	0.33	-

d. See the Stata script for the specific calls to margins. Here we just provide nicely formatted results.

1. Adjusted predictions

```
## Read margins output: ------------------------------------------------------
adj_pred_est = readxl::read_excel('./ps2_q2_d.xls', sheet = 'adj_pred_est',
                                  col_names = FALSE)
adj_pre_est = as.numeric(adj_pred_est)

adj_pred_v = readxl::read_excel('./ps2_q2_d.xls', sheet = 'adj_pred_v',
                                col_names = FALSE)
adj_pred_se = sqrt(diag(as.matrix(adj_pred_v)))

## Format a nice table: ------------------------------------------------------
cap = 'Probability Primary Tooth is Absent at Select Ages.'
tibble(
  `Age (years)` = 8:12,
  `Probability Primary Tooth is Absent at the Mean` =
    sprintf('%4.2f (%4.2f, %4.2f)', adj_pred_est, adj_pred_est - m*adj_pred_se,
            adj_pred_est + m*adj_pred_se)
) %>%
  knitr::kable()
```

Age (years)	Probability Primary Tooth is Absent at the Mean
8	0.15 (0.13, 0.18)
9	0.30 (0.27, 0.33)
10	0.50 (0.46, 0.53)
11	0.69 (0.67, 0.72)
12	0.84 (0.82, 0.86)

2. *Marginal Effect at the Mean for Non-Hispanic Black at representative ages.*

```
## Read margins output: ------------------------------------------------------
me_est = readxl::read_excel('./ps2_q2_d.xls', sheet = 'me_black_est',
                            col_names = FALSE) %>% as.numeric()
me_est = me_est[me_est > 0]

me_v = readxl::read_excel('./ps2_q2_d.xls', sheet = 'me_black_v',
                                col_names = FALSE)
me_se = sqrt(diag(as.matrix(me_v)))
me_se = me_se[me_se > 0]

## Format a nice table: ------------------------------------------------------
cap = '**Difference in Probability of Absent Primary Tooth at Select Ages.**'
tibble(
  `Age (years)` = 8:12,
  `Non-Hispanic Black` =
    sprintf('%4.2f (%4.2f, %4.2f)', me_est, me_est - m*me_se,
            me_est + m*me_se)
) %>%
  knitr::kable(caption = cap)
```

Difference in Probability of Absent Primary Tooth at Select Ages.

Age (years)	Non-Hispanic Black

Age (years)	Non-Hispanic Black
8	0.07 (0.03, 0.12)
9	0.11 (0.05, 0.18)
10	0.13 (0.06, 0.20)
11	0.10 (0.05, 0.16)
12	0.06 (0.03, 0.10)

3. *Average Marginal Effects for Non-Hispanic Black at representative ages.* The result, in this case, is the same as in part 2 as we have only two variables in the model. In general, these would not be same. For instance, if you selected PIR you will have something different here.

e.

```
q2_e_reg = readxl::read_excel('./ps2_q2_e_coef.xls',
                              skip = 1)
terms = names(q2_e_reg[,-1])
est = as.numeric(q2_e_reg[1,-1])
se = as.numeric(q2_e_reg[2,-1])
terms = clean_model(terms)
ind = which(!grepl('^0', terms))

or = sprintf('%4.2f (%4.2f, %4.2f)', exp(est), exp(est - m*se), exp(est + m*se))

terms = stringr::str_replace(terms, "^1\\.", "")

cap = '**Coefficient table using survey weights.** The table shows the fitted
coefficients for the final logistic regression model for the probability
of an absent primary tooth. In conrast to the earlier table, this verion comes from
using svy: logit to account for the complex survey weights.'

cap = stringr::str_replace(cap, '\n', ' ')

svy_reg =
  tibble(Variable = terms[ind], Coefficient = est[ind], SE = se[ind]) %>%
  mutate(
    OR = or[ind], OR = ifelse(grepl('Intercept', Variable), '-', OR)
  ) %>%
  rename( `OR (95% CI)` = OR)
svy_reg %>% knitr::kable( digits = 2 , caption = cap)
```

Coefficient table using survey weights. The table shows the fitted coefficients for the final logistic regression model for the probability of an absent primary tooth. In conrast to the earlier table, this verion comes from using svy: logit to account for the complex survey weights.

Variable	Coefficient	SE	OR (95% CI)
Age (months)	0.06	0.01	1.06 (1.05, 1.08)
Non-Hispanic Black	0.61	0.13	1.84 (1.42, 2.39)
Intercept	-7.65	0.85	-

In the survey setting, the odds ratio between successive ages (in months) is similar to that obtained under the simple random sample assumption. The odds ratio comparing people of race Non-Hispanic Black to all other is a little bit higher than when assuming a simple random sample, 1.84 vs 1.68, but there is substantial overlap between their confidence intervals. Finally, the intercept obtained under the simple random sample assumption is roughly one standard error lower than when using the survey weights leading to a 2-3% difference in the adjusted predictinons at our selected ages.

Question 3 [30 points]

Repeat part a-d of question 2 using **R**. For part d, you may either use the “margins” package or code the computations yourself.

Solution: See the R script `ps2_q3.R` for a solution. It is sufficient to include nicely formatted output from either the R or Stata versions provided they agree. The solution includes both direct computations of the quantities in part 3 and those done using margins. You needed to provide only one version.