`gretl` Notes

Week 4

Module 2 Week 4

Intro

- Dataset used: `National Election Study 2008.gdt`
  - Open with `File` -> `Open data` -> `User file...`
- Open up a new script editor: `File` -> `Script files` -> `New script` -> `gretl script`
- Checkout summary statistics:
  - ```
    summary --simple
    ```
  - Most Variables are "thermometer" values: higher => more positive feelings of a respondent
- How to get summary statistics across categories:
  - ```
    summary obama_therm --by=race_ethnicity
    ```
- Give `race_ethnicity` variable a label (from data documentation):
  - ```
    setinfo race_ethnicity --description="1=White, 2=Black, 3=Hispanic, 4=Other"
    ```

Two sample hypothesis test for means, with independent sample

- Does the mean for the big business thermometer differ across whether or not someone has the Internet?
  - H_0: $\mu_{bigbus\_withinternet} - \mu_{bigbus\_withoutinternet} = 0$
    H_a: $\mu_{bigbus\_withinternet} - \mu_{bigbus\_withoutinternet} \neq 0$
  - 5 - no internet, 1 - internet:
    - ```
      freq internet
      ```

```
Frequency distribution for internet, obs 1-2323

          frequency    rel.     cum.

   1        1615     69.58%   69.58% ************************
   5         706     30.42%  100.00% **********

Missing observations = 2 ( 0.09%)
```

- Fix coding:

```
series netaccess = (internet==1)
```

- Let `gretl` know it's discrete:

```
discrete netaccess
```

- Fixed version:

```
freq netaccess
```

```
Frequency distribution for netaccess, obs 1-2323

          frequency    rel.     cum.

   0         706     30.42%   30.42% **********
   1        1615     69.58%  100.00% ************************

Missing observations = 2 ( 0.09%)
```

- Summary stats on `bigbusiness` by `netaccess` to see the means we'll compare:

```
summary bigbus_therm --by=netaccess
```

```
Summary statistics for bigbus_therm, by value of netaccess

  netaccess = 0 (n = 706):
    Mean                          59.994
    Median                        60.000
    Minimum                       0.0000
    Maximum                       100.00
    Standard deviation            22.078
    C.V.                         0.36800
    Skewness                    -0.28223
    Ex. kurtosis                 0.42036
    5% percentile                 15.000
    95% percentile                100.00
```

```
    Interquartile range          20.000
    Missing obs.                      0

  netaccess = 1 (n = 1615):
    Mean                         53.182
    Median                       50.000
    Minimum                      0.0000
    Maximum                      100.00
    Standard deviation           20.595
    C.V.                         0.38725
    Skewness                    -0.15362
    Ex. kurtosis                 0.41778
    5% percentile                15.000
    95% percentile               85.000
    Interquartile range          30.000
    Missing obs.                      0
```

- Is this a real difference in means or just by chance?
  - ```
    anova bigbus_therm netaccess
    ```
  - ```
    Analysis of Variance, response = bigbus_therm, treatment =
    netaccess:

                         Sum of squares       df      Mean square

      Treatment                22797.6        1          22797.6
      Residual             1.02821e+06     2319          443.384
      Total                1.05101e+06     2320           453.02

      F(1, 2319) = 22797.6 / 443.384 = 51.4172 [p-value 1e-12]

      Level         n       mean     std. dev

      0           706    59.9943       22.078
      1          1615     53.182       20.595

      Grand mean = 55.2542
    ```
- Conclusion:
  - `p-value 1e-12`
  - p-value is approximately 0. Therefore, we reject Ho, we believe there is a statistically sign difference in the mean of the big business therm between those with and without Internet access
- $\text{T-Stat} = \sqrt{\text{F-Stat}}$:

```
? eval sqrt(51.4172)
7.1705788
```

- Use the T-value and the T-Distribution instead of the F-Distribution:

```
? pvalue t 2319 7.1705788
t(2319): area to the right of 7.17058 = 4.99898e-13
(two-tailed value = 9.99795e-13; complement = 1)
```

- Conclusion:
  - $\text{p-value} = 9.99795e-13$ => Almost the same as the earlier F-Stat-based $\text{p-value} = 1e-12$, as expected!
  - Can use `anova` output to perform t-tests!

Chi-squared test for independence between two nominal variables

- Are the variables internet access and gun ownership related or independent?
  - H_0: `netaccess` and `gunown` are independent
    H_a: `netaccess` and `gunown` are not independent

```
? freq gunown

Frequency distribution for gunown, obs 1-2323

          frequency    rel.     cum.

  1         651     28.63%   28.63% **********
  5        1623     71.37%  100.00% ************************

Missing observations = 49 ( 2.11%)
```

- The coding is messed up again, need to re-code (5 = no, 1 = yes):

```
? series gunown = (gunown==1)
Replaced series gunown (ID 56)
? freq gunown

Frequency distribution for gunown, obs 1-2323

          frequency    rel.     cum.

  0        1623     71.37%   71.37% ************************
  1         651     28.63%  100.00% **********

Missing observations = 49 ( 2.11%)
```

- Perform the Chi-Squared Test for Independence:
  - ```
    ? xtab netaccess gunown

    Cross-tabulation of netaccess (rows) against gunown (columns)

           [   0][   1]  TOT.

    [   0]   547   149    696
    [   1]  1074   502   1576

    TOTAL   1621   651   2272

    51 missing values

    Pearson chi-square test = 25.7635 (1 df, p-value = 3.8591e-07)
    ```
  - $df = (cols - 1) * (rows - 1)$
  - $\text{p-value} = 3.8591e - 07$ => Reject the null hypothesis, the variables `netaccess` and `gunown` are related to one another
- Look at the marginals by `gunown` to investigate the relationship:
  - ```
    ? xtab netaccess gunown --column

    Cross-tabulation of netaccess (rows) against gunown (columns)

           [   0][   1]  TOT.

    [   0]  33.7% 22.9% 30.6%
    [   1]  66.3% 77.1% 69.4%

    TOTAL   1621   651   2272

    51 missing values

    Pearson chi-square test = 25.7635 (1 df, p-value = 3.8591e-07)
    ```
  - Conclusion:
    - Those who own a gun seem to be more likely to have internet access.
- GUI: `View` -> `Cross Tabulation` -> Move `netaccess` and `gunown` to the right -> `OK`

One-way ANOVA

- Is there a difference in the mean of the federal government therm across race/ethnicity?
  - ```
    ? anova fedgov_therm race_ethnicity

    Analysis of Variance, response = fedgov_therm, treatment =
    race_ethnicity:

                         Sum of squares       df      Mean square

      Treatment                 81808.3        3          27269.4
      Residual              1.03246e+06     2311          446.759
      Total                 1.11427e+06     2314          481.533

      F(3, 2311) = 27269.4 / 446.759 = 61.0384 [p-value 5.66e-38]

      Level         n       mean     std. dev

      1          1159    46.3658       19.925
      2           569    58.8067       22.351
      3           509    57.8625       22.260
      4            78     48.141       21.987

      Grand mean = 52.0112
    ```
  - Conclusion:
    - $\text{p-value} = 5.66e-38$ =>
    - Reject H_0, there is a statistically significant difference in mean of `fedgov therm` across race/ethnicity categories
- GUI Anova: `Model` -> `Other linear models` -> `ANOVA` -> Choose Response (`fedgov_therm`) and Treatment (`race_ethinicity`) Variables (No block variable, that's for 2-way ANOVA) -> `OK`
    ```