11.220 Quantitative Reasoning & Statistical Methods for Planners I

Spring 2009

Computer Lab #3 — Apr 3rd, 2009

Analyze data: T-test, ANOVA and Correlation

Tips to get the software and data work:

To use STATA on Linux system

```
type "add stata" in the terminal
type "xstata" in the terminal
```

To use flash drive on Linux system

```
type "add consult" in the terminal
type "tellme root" and pay attention to the password it gives you
type "attach-usb" and then enter that password
The path will be "/mnt/usb/foldername"
type "detach-usb", and give the same password to detach f-drive
```

Metadata of "Hedonic.dta"

This data set contains observations on house prices and attributes in the city of Newton.

id	house code
price	sale price
lot	lot size
style	building style
year_b	year when the house was built
size	total areas of living space
room	number of rooms
bed	number of bedrooms
bath	number of bathrooms
q1	interior condition of the house: "above","average","bellow"
q2	bathroom condition: "above","average","bellow"
year_s	year of sale
old	dummy variable = 1 if the house was built before 1930

STATA commands used in today's class

ttest	compare the sample means or other descriptive statistics values
oneway	one-way analysis of variance
anova	analysis of variance
corr	simple correlation among variables
twoway scatter	produce scatter plot of outcome vs. predictor
graph matrix	produce multiple twoway scatter plot at a time

Scripts in the real Command Window

```
cd E:\MIT\09Spring\STATALAB\DATA (change this part to your own local directory)
use hedonic, clear
log using log1, text
summarize
```

1) **T-test (One sample and two independence samples)**

```
/// Compare the mean of one variable to some constant value
      ttest size = 1770      * Can we reject H0 μ = 1770? Why?
      ttest size = 2000      * Can we reject H0 μ = 2000? Why?
      ttest size = 1770, level(99)    * Can we reject H0 μ = 1770 now?
```

*Note: This is to infer whether the mean of the population equals 1700 or 2000, given the sample mean we already know.

```
/// Compare the mean of two different variables
      ttest bed = bath, unpaired
```

*Note: Here I use the option "unpaired" since the means are from different variables. "Paired" ttest is by default, which is designed to compare the means of the same variable from different samples. Think of a "pre-post" situation.

```
/// Compare the mean price of old houses vs. new houses
      tab old, summarize(price)
      ttest price, by(old)    * Can we reject H0: μp_old = μp_new?
      ttest price, by(old) unequal
```

*Note: If we concern that the samples may have different variances, we need to include "unequal" option.

2) Analysis of Variance

```
/// See whether the old houses and new houses have equal variance?
      oneway price, old       * Check the chi2 value, what do you find?
      anova price, old
```

3) Things to do before run into "regression"

```
/// See the simple correlation among variables before we do regression,
and this help us to roughly determine which predictors to be included.
      corr price lot year_b size room bed bath year_s

/// Plot the outcome against some predictors
      graph matrix price lot size room, half

/// Plot "price" against "lot" with fitted linear regression line
      twoway scatter price lot || lfit price lot

/// Plot "price" against "lot" with 95% confidence interval
      twoway scatter price lot || lfitci price lot

/// Do simple linear regression!
      regress price lot
```

Exercises

1: Test whether the μ of lot size = 8600? 8900? On a 95% confidence level.

2: Test whether the μ of lot size is statistically different between new and old house.

3. Test whether the variances of price are different for houses with different interior quality? Hint: use "q1" to divide the data into 3 groups.

4. Plot "price" against "room" with fitted regression line and confidence interval.