11.220 Quantitative Reasoning & Statistical Methods for Planners I
Spring 2009

Computer Lab #4 Apr 24th, 2009

Regression: Bivariate/Multivariate Model, Log Transformation and Categorical Variable

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 “nbawage.dta”

This dataset contains NBA players’ wages and their personal characteristics.

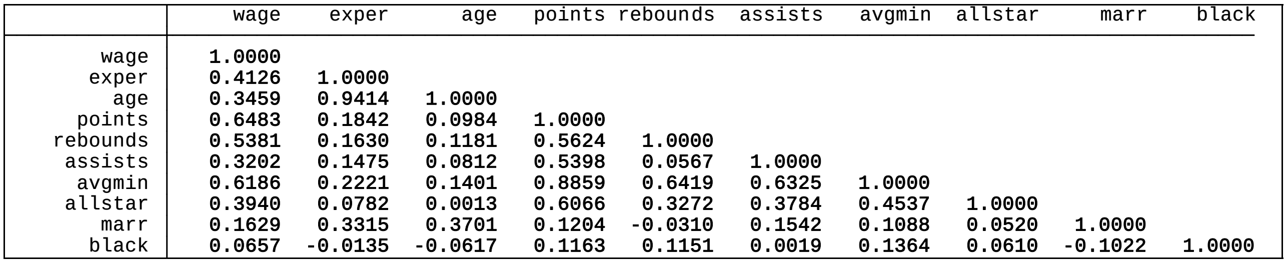
wage	annual salary (million $)
exper	years as a professional player
age	age (in years)
point	points per game
rebounds	rebounds per game
assists	assists per game
avgmin	minutes per game
allstar	=1 if allstar player
marr	=1 if married
black	=1 if black

Scripts in the Command Window

```
///change this part to your own local directory
      cd E:\MIT\09Spring\STATALAB\DATA
      use nbawage, clear
      log using log, text replace
      sum
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |       270    1.428924    1.001422        .15       5.74
       exper |       270    5.133333    3.401946          1         18
         age |       270    27.40741    3.392089         21         41
      points |       270    10.18815    5.901037        1.2       29.8
    rebounds |       270    4.401481    2.887197         .5       17.3
-------------+--------------------------------------------------------
     assists |       270    2.404444    2.090388          0       12.6
      avgmin |       270    23.97278    9.713654   2.888889   43.08537
     allstar |       270    .1148148    .3193903          0          1
        marr |       270    .4444444    .4978268          0          1
       black |       270    .8037037    .3979328          0          1
```

```
        corr
```



```
             |     wage    exper      age   points rebounds  assists   avgmin  allstar     marr    black
-------------+------------------------------------------------------------------------------------------
        wage |   1.0000
       exper |   0.4126   1.0000
         age |   0.3459   0.9414   1.0000
      points |   0.6483   0.1842   0.0984   1.0000
    rebounds |   0.5381   0.1630   0.1181   0.5624   1.0000
     assists |   0.3202   0.1475   0.0812   0.5398   0.0567   1.0000
      avgmin |   0.6186   0.2221   0.1401   0.8859   0.6419   0.6325   1.0000
     allstar |   0.3940   0.0782   0.0013   0.6066   0.3272   0.3784   0.4537   1.0000
        marr |   0.1629   0.3315   0.3701   0.1204  -0.0310   0.1542   0.1088   0.0520   1.0000
       black |   0.0657  -0.0135  -0.0617   0.1163   0.1151   0.0019   0.1364   0.0610  -0.1022   1.0000
```

1) Bivariate regression model (Uncontrolled regression)

```
/// Run regression of wage on exper,points respectively
      reg wage exper
      reg wage points
```

*Note: If we do not control for other variables, all slope coefficients appear to be statistically significant.

2) Multivariate regression model (Controlled ~)

```
/// Run regression of wage on points control for exper, or even more
predictors
      reg wage points exper
      reg wage points exper rebounds
```

*Note: Since “exper” and “age” have very high correlation (0.94), we do not want to include both in one model. Likewise, “points” and “avgmin” are also highly correlated (0.88), we can include either but not both.

3) Ln-linear and Ln-ln model

```
/// Plot the distribution of wage, points and exper
      histogram wage, normal          /*wage is positively skewed*/
      histogram points, normal        /*points is positively skewed*/
      histogram exper, normal        /*exper is positively skewed*/
/// Create new variable using log transformed data
      gen lwage = ln(wage)
      gen lpoints = ln(points)
      gen lexper = ln(exper)
```

*Note: Sometimes we do log transform when the variable is positively skewed, even if we do not detect obvious non-linear relationship. The function of log transform is to compress the high end values and stretch the low end values.

```
/// Add label to the new variables
      label variable lwage "natural log of wage"
      label variable lpoints "natural log of points"
      label variable lexper "natural log of exper"
```

```
/// Run Ln-linear regression use the new variables
      Reg lwage points exper
```

$$\ln wage = -1.207 + 0.083 po\operatorname{int} s + 0.079 \exp er$$

*Note: Remember when the slope coefficient is very small, $e^{\hat{\beta}} - 1 \approx \hat{\beta}$.This means, every 1 unit difference in per game points is associated with 8.3% difference in wages, controlled for years as professional player. Or, you can say, every 1unit difference in years as professional player is associated with 7.9% difference in wages, controlled for points per game.

```
/// Run Ln-ln regression use the new variables
      reg lwage lpoints lexper
```

$$\ln wage = -1.95 + 0.70 \ln point s + 0.37 \ln \exp er$$

*Note: This means, every 1% difference in points is associated with 0.7% difference in wages, controlled for years as professional player. Or, you can say, every 1% difference in exper is associated with 0.3% difference in wages, controlled for points per game.

4) Including categorical variable in the model

```
/// Simple regression on allstar
      reg wage allstar
```

$$wage = 1.28 + 1.23 allstar$$

*Note: This is interpreted as, the mean wage for non allstar player is $1.28million, and being allstar player means a $1.23 million more, if we do not control for other variables.

```
/// Run regression on allstar, control for exper
      reg wage exper allstar
```

$$wage = 0.717 + 0.113 \exp er + 1.141 allstar$$

*Note: allstar is still statistically significant. We need some more calculation to explain the slope coefficients. If we plug in the mean of exper(5.13), we get the adjusted mean wages of non-allstar players, which is $1.30 million, and being allstar player means 1.14 millions more in wage.

```
/// Run regression on allstar, control for points
      reg wage points allstar
```

*Note: Now allstar becomes insignificant. Intuitively we know that points and allstar should be highly correlated, we do not need to include both in one model.

Exercises

1: Run a regression of wage on “age” and “rebounds” respectively and together, what do you find?
2: Plot the distribution of “age” and “rebounds”, do you worry about asymmetric distribution? If yes, take log transformation and rerun the regression.
3: Is there any difference between married and unmarried, black or nonblack players? Develop a model to capture the difference, if any.