Tutorial 5

keywords: OLS estimator, multiple linear regression, interpretation, ceteris paribus, predict, interpretation, variation, R squared

estimated reading time: 30 minutes

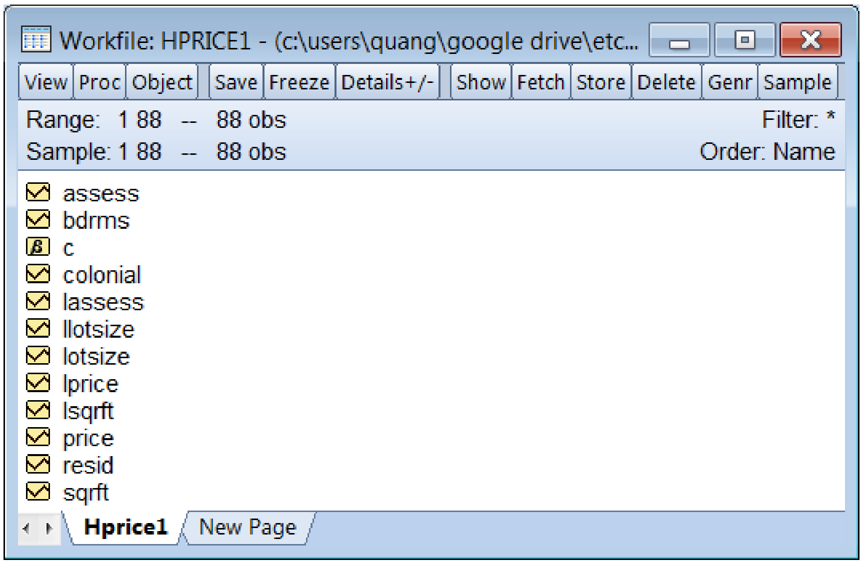
Quang Bui

March 26, 2018

Question 1

Multiple linear regression model and interpreting coefficients

EViews workfile: $hprice.wf1$

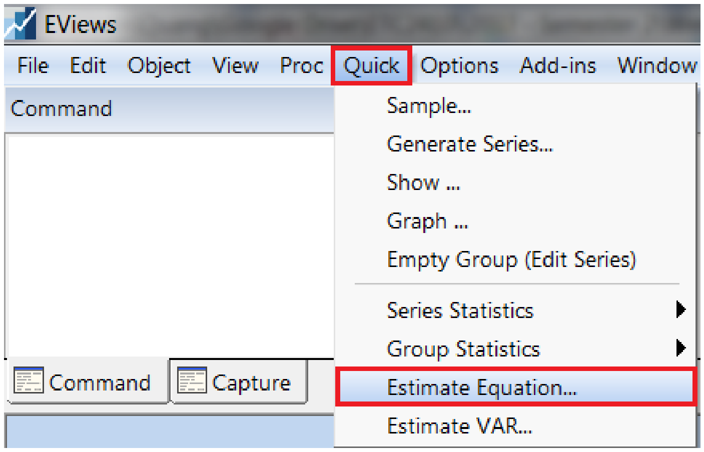


i. Estimate the model of $price$ on a constant, $sqrft$ and $bdrms$ and write out the results in equation form.

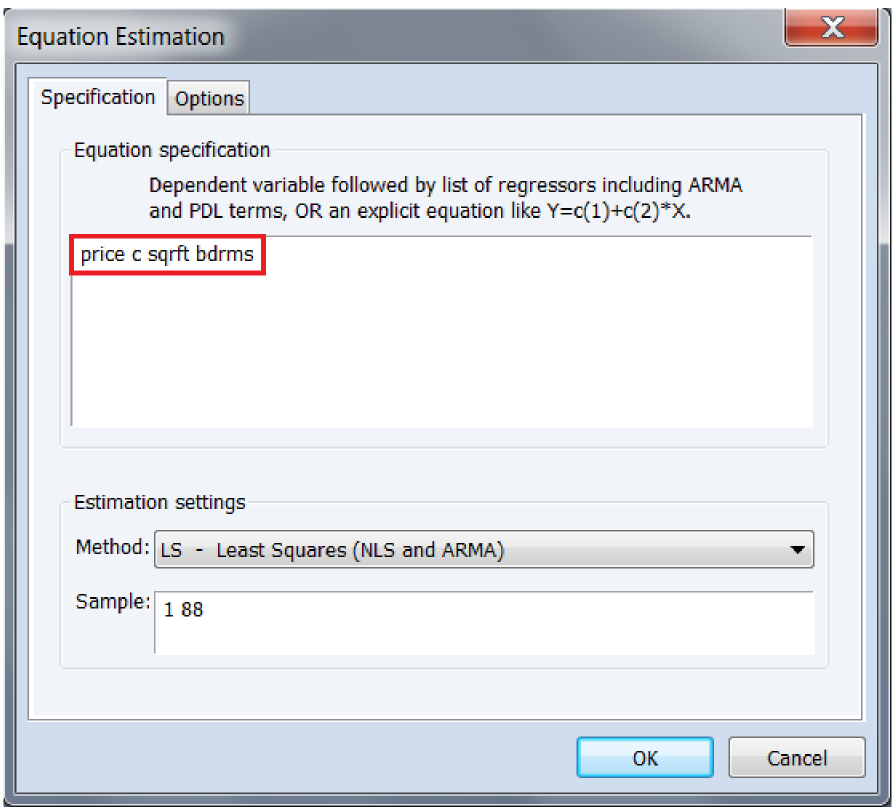
$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$

- $price$ - house price ($'000)
- $sqrft$ - area of the house (square foot)
- $bdrms$ - no. of bedrooms

$$Quick \rightarrow Estimate\ Equation$$



$$Equation\ Estimation : price\ c\ sqrft\ bdrms$$



Dependent Variable: PRICE
Method: Least Squares
Date: 07/16/17 Time: 21:25
Sample: 1 88
Included observations: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−19.31500	31.04662	−0.622129	0.5355
SQRFT	0.128436	0.013824	9.290506	0.0000
BDRMS	15.19819	9.483517	1.602590	0.1127

R-squared	0.631918	Mean dependent var	293.5460
Adjusted R-squared	0.623258	S.D. dependent var	102.7134
S.E. of regression	63.04484	Akaike info criterion	11.15907
Sum squared resid	337845.4	Schwarz criterion	11.24352
Log likelihood	−487.9989	Hannan-Quinn criter.	11.19309
F-statistic	72.96353	Durbin-Watson stat	1.858074
Prob(F-statistic)	0.000000		

Table 1: Regression output of *price* on a constant, *sqrft* and *bdrms*

When reporting the estimated model, we must not forget to include a 'hat' above the dependent variable and $se(\hat{\beta}_j)$ underneath $\hat{\beta}_j$ in parenthesis,

$$\widehat{price} = \underset{(se(\hat{\beta}_0))}{\hat{\beta}_0} + \underset{(se(\hat{\beta}_1))}{\hat{\beta}_1} sqrft + \underset{(se(\hat{\beta}_2))}{\hat{\beta}_2} bdrms$$

$$\widehat{price} = \underset{(31.0466)}{-19.3150} + \underset{(0.0138)}{0.1284} sqrft + \underset{(9.4835)}{15.1982} bdrms$$

ii. What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

Background

Interpretation of estimated coefficients for multiple linear regression models

Suppose we estimate a model of y on a constant, x_1, and x_2,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

if x_1 and x_2 changes by Δx_1 and Δx_2 respectively then,

$$x_1 \textit{ becomes } x_1 + \Delta x_1$$

$$x_2 \textit{ becomes } x_2 + \Delta x_2$$

which will change $\hat{y}$,

$$\hat{y} \textit{ becomes } \hat{y} + \Delta \hat{y}$$

This then gives us the following equation,

$$\begin{aligned} \hat{y} + \Delta \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 (x_1 + \Delta x_1) + \hat{\beta}_2 (x_2 + \Delta x_2) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 \\ &= \hat{y} + \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 \end{aligned}$$

Since $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, it must follow that,

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

$\therefore$ the change in $\hat{y}$ for a 1-unit change in x_1, holding x_2 constant, is $\hat{\beta}_1$,

$$\Delta x_2 = 0$$

$$\Delta x_1 = 1$$

$$\begin{aligned} \Delta \hat{y} &= \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 \\ &= \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times 0 \\ &= \hat{\beta}_1 \end{aligned}$$

and the change in $\hat{y}$ for a 1-unit change in x_2, holding x_1 constant, is $\hat{\beta}_2$,

$$\Delta x_2 = 1$$

$$\Delta x_1 = 0$$

$$\begin{aligned} \Delta \hat{y} &= \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 \\ &= \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 1 \\ &= \hat{\beta}_2 \end{aligned}$$

As we can see, $\hat{\beta}_1$ and $\hat{\beta}_2$ have a partial effect (ceteris paribus) interpretation!

From our estimated model, the changed in estimated house price depends on the change square footage and no. of bedrooms,

$$\Delta\widehat{price} = \hat{\beta}_1 \Delta sqrft + \hat{\beta}_2 \Delta bdrms$$

Note: The estimated intercept coefficient does not change the estimated house price.

If square footage is held constant,

$$\Delta sqrft = 0$$

then the change in the estimated house price depends only on the change in no. of bedrooms,

$$\begin{aligned}\Delta\widehat{price} &= \hat{\beta}_1 \Delta \times 0 + \hat{\beta}_2 \Delta bdrms \\ &= \hat{\beta}_2 \Delta bdrms\end{aligned}$$

Therefore, the estimated increase in house price for for an additional bedroom, <u>holding square footage constant</u>,

$$\begin{aligned}\Delta\widehat{price} &= \hat{\beta}_2 \times 1 \\ &= 15.1982 \times 1 \\ &= 15.1982\end{aligned}$$

$$\$15,198.20$$

iii. What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).

$$\Delta bdrms = 1$$

$$\Delta sqrft = 120$$

$$\begin{aligned}\Delta\widehat{price} &= \hat{\beta}_1 \Delta sqrft + \hat{\beta}_2 \Delta bdrms \\ &= 0.1284 \times 140 + 15.1982 \times 1 \\ &= 33.12\end{aligned}$$

$$\$33,120$$

The change in estimated house price is greater here than in ii) because we are also increasing the size of the house. In ii), we estimated the change in house price for an additional bedroom but kept the size of the house the same.

iv. What percentage of the variation in price is explained by square footage and number of bedrooms?

$$R^2 = 63.2\%$$

63.2% of the variation in house price is explained by square footage and number of bedrooms.

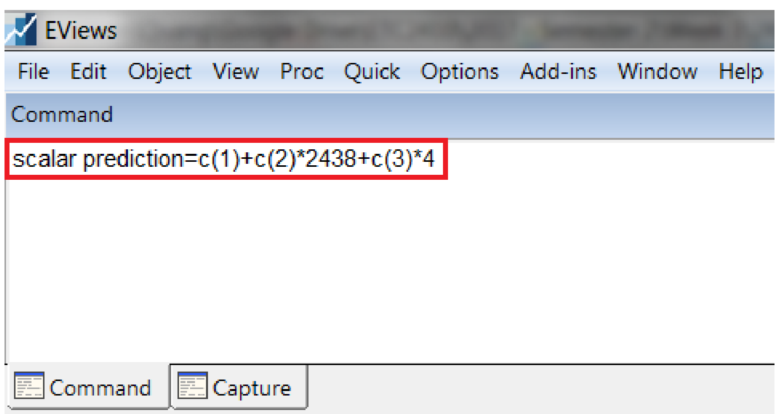
v. The first house in the sample has $sqrft = 2438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

$$\begin{aligned}
\widehat{price} &= -19.3150 + 0.1284 sqrft + 15.1982 bdrms \\
\widehat{price}_1 &= -19.3150 + 0.1284 sqrft_1 + 15.1982 bdrms_1 \\
&= -19.3150 + 0.1284 \times 2438 + 15.1982 \times 4 \\
&= 354.6052
\end{aligned}$$

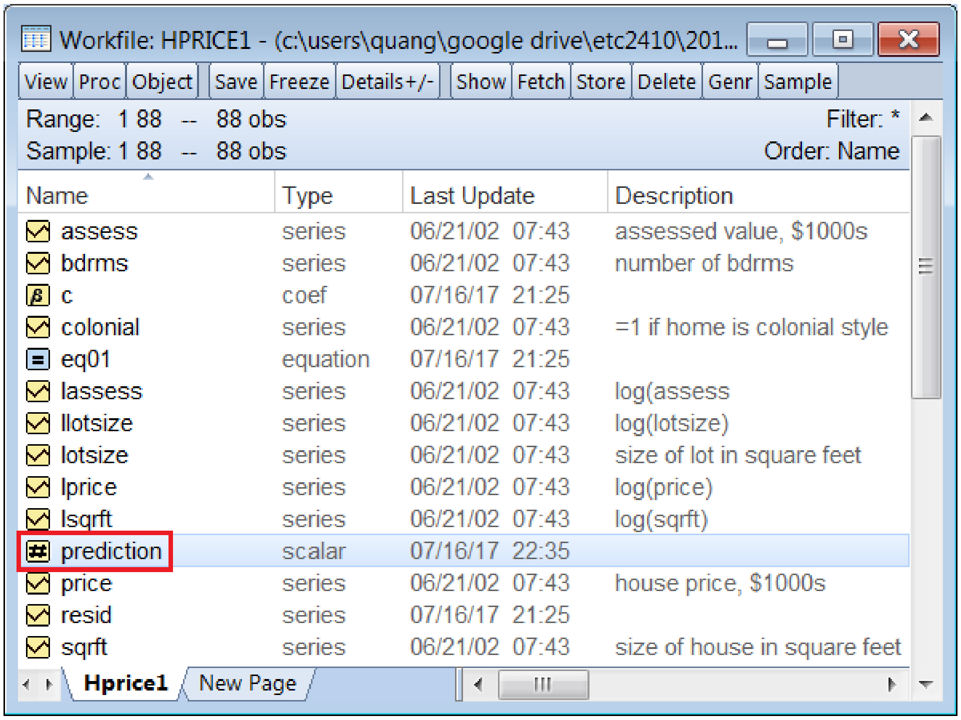
$$\$354,605$$

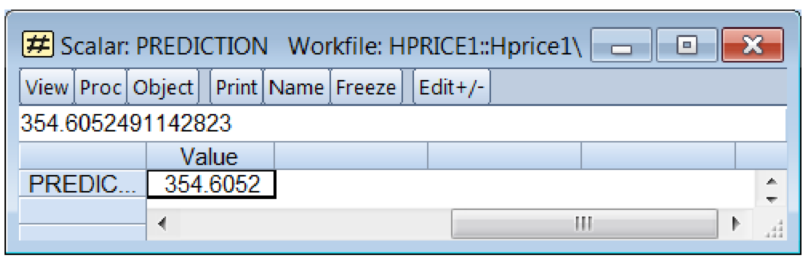
To perform this calculation in EViews,

$$Command\ Window : scalar\ prediction\ \ = c(1) + c(2)^{*}2438 + c(3)^{*}4$$



(*press Enter to execute code*)





vi. The actual selling price of the first house in the sample was \$300,000 (so $price_1$=300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

$$\hat{u}_i = price_i - \widehat{price}_i$$

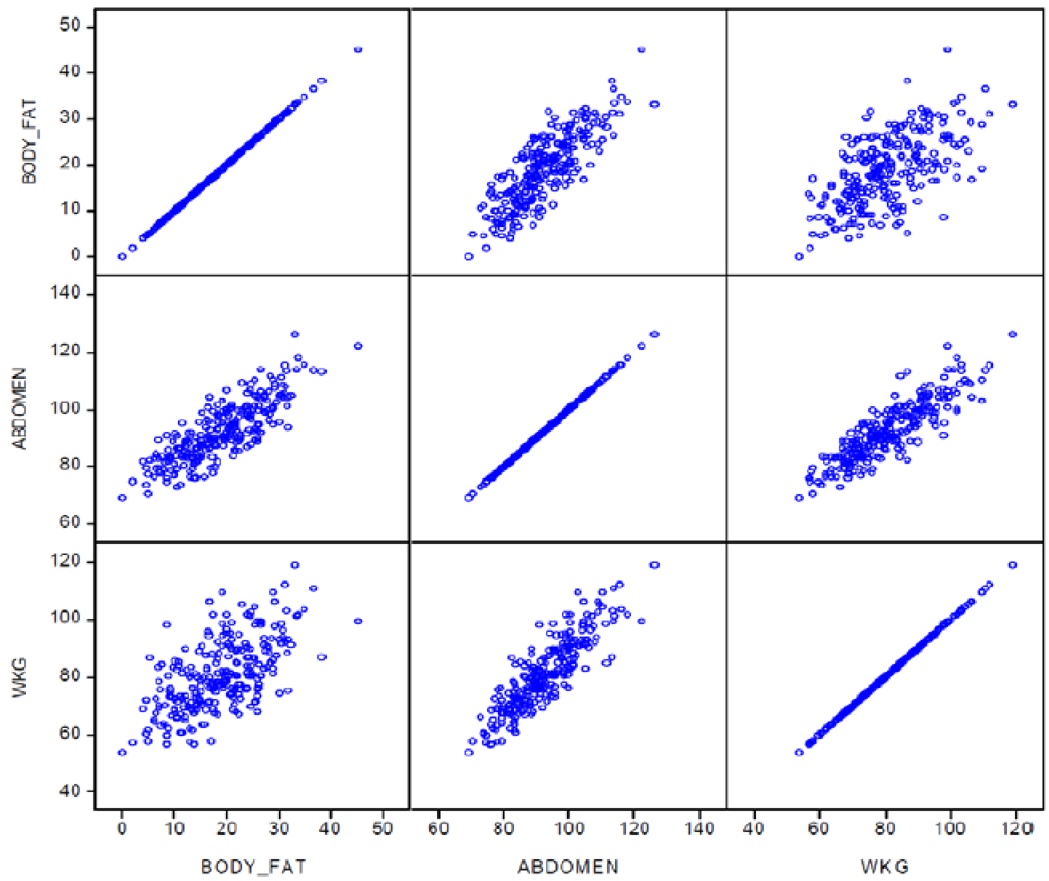
$$\begin{aligned}\hat{u}_1 &= price_1 - \widehat{price}_1 \\ &= 300 - 354.605 \\ &= -54.605\end{aligned}$$

$$-\$54,605$$

Based on our estimated model, the buyer underpaid, however, we have not considered other features that impact house price e.g. number of baths, age of house, whether it has been renovated etc.

Question 2

We would like to make an "app" where users input their easy to measure body characteristics and the app predicts their body fat percentage. We start with making an app for men. We have data on body fat percentage ($BODY\_FAT$), weight in kg (WKG) and abdomen circumference in cm ($ABDOMEN$) for 251 adult men. The matrix of scatter plots of each pair of these three variables in our sample is given below.



Without estimating any regressions, explain what these plots can tell us about each of the following (the correct answer for one of these is "nothing"):

Background

OLS estimator for a simple linear regression model

For the following simple linear regression model,

$$y = \beta_0 + \beta_1 x_1 + u$$

the OLS estimates of β_0 and β_1 can be expressed by the following formulas,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$$

$$\hat{\beta}_1 = \frac{\widehat{Cov(y, x_1)}}{\widehat{Var(x_1)}}$$

or in matrix notation,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \bar{y} - \hat{\beta}_1 \bar{x}_1 \\ \\ \dfrac{\widehat{Cov(y, x_1)}}{\widehat{Var(x_1)}} \end{bmatrix}$$

since $\widehat{Var(x_1)} > 0$, the sign of $\hat{\beta}_1$ depends directly on the sign of $\widehat{Cov(y, x_1)}$.

For the multiple linear regression model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

the OLS estimate of β_1 is not equal to $\dfrac{\widehat{Cov(y, x_1)}}{\widehat{Var(x_1)}}$,

$$\hat{\beta}_1 \neq \frac{\widehat{Cov(y, x_1)}}{\widehat{Var(x_1)}}$$

$\therefore$ the sign of $\hat{\beta}_1$, in the estimated multiple linear regression model, does not depend directly on the sign of $\widehat{Cov(y, x_1)}$.

(a) the sign of the coefficient of $ABDOMEN$ in a regression of $BODY\_FAT$ on a constant and $ABDOMEN$

$$BODY\_FAT = \beta_0 + \beta_1 ABDOMEN + u$$

$$\widehat{BODY\_FAT} = \hat{\beta}_0 + \hat{\beta}_1 ABDOMEN$$

For the simple regression model of $BODY\_FAT$ on a constant and $ABDOMEN$ the OLS estimates of β_0 and β_1 are given by the following formulas,

$$\hat{\beta}_0 = \overline{BODY\_FAT} - \hat{\beta}_1 \overline{ABDOMEN}$$

$$\hat{\beta}_1 = \frac{\widehat{Cov(BODY\_FAT, ABDOMEN)}}{\widehat{Var(ABDOMEN)}}$$

From the scatter plot, we can see that $BODY\_FAT$ and $ABDOMEN$ have a positive linear relationship,

$$\begin{aligned}
&\therefore \widehat{Cov(BODY_FAT, ABDOMEN)} > 0 \\
&\implies \hat{\beta}_1 > 0
\end{aligned}$$

(b) the sign of the coefficient of WKG in a regression of $BODY\_FAT$ on a constant and WKG

$$(different\ model\ so\ I'm\ using\ a\ different\ greek\ letter)$$

$$\begin{aligned}
BODY_FAT &= \alpha_0 + \alpha_1 WKG + u \\
\widehat{BODY_FAT} &= \hat{\alpha}_0 + \hat{\alpha}_1 WKG
\end{aligned}$$

For the simple regression model of $BODY\_FAT$ on a constant and WKG the OLS estimates of α_0 and α_1 are given by the following formulas,

$$\begin{aligned}
\hat{\alpha}_0 &= \overline{BODY_FAT} - \hat{\alpha}_1 \overline{WKG} \\
\hat{\alpha}_1 &= \frac{\widehat{Cov(BODY_FAT, WKG)}}{\widehat{Var(WKG)}}
\end{aligned}$$

From the scatter plot, we can see that $BODY\_FAT$ and WKG have a positive linear relationship,

$$\begin{aligned}
&\therefore \widehat{Cov(BODY_FAT, WKG)} > 0 \\
&\implies \hat{\alpha}_1 > 0
\end{aligned}$$

(c) which of the two regressions explained in parts (a) and (b) is likely to have a better fit?

$$\widehat{BODY\_FAT} = \hat{\beta}_0 + \hat{\beta}_1 ABDOMEN \quad (1)$$

$$\widehat{BODY\_FAT} = \hat{\alpha}_0 + \hat{\alpha}_1 WKG \quad (2)$$

The first estimated model is likely to fit the data better than the second. Why?

An OLS regression line of (1) through the scatter plot of $BODY\_FAT$ against $ABDOMEN$ would have a smaller sum of squared residuals (SSR) than the OLS regression line of (2) through the scatter plot of $BODY\_FAT$ against WKG.

Since,

$$R^2 = 1 - \frac{SSR}{SST}$$

and SST (sum of squared totals) is the same for both estimated models,

$$SST = \sum_{i-1}^{n} (BODY\_FAT_i - \overline{BODY\_FAT})$$

then the R^2 of (1) is likely to be higher than the R^2 of (2).

The scatter plot of $BODY\_FAT$ against $ABDOMEN$, $BODY\_FAT$ is less dispersed around $\overline{BODY\_FAT}$ for each value of $ABDOMEN$ than it is for WKG. (Think about R^2.)

(d) the sign of the coefficient of WKG in a regression of $BODY\_FAT$ on a constant, $ABDOMEN$ and WKG.

Scatter plots cannot tell us anything about the correlation of body fat and weight after the influence of abdomen has been taken out. (Think about 2 people with the same abdomen circumference i.e. controlling for $ABDOMEN$ but one weights more than the other. Since both have the same abdomen circumference, the one that is heavier will have weight distributed elsewhere in his body that the other male does not e.g. broader shoulders, thicker quads, fuller chest etc. If both males have the same abdomen circumference, the one with the bigger shoulders, quads, chest, etc. is likely to have a better physique and also likely to have less body fat.)

Question 4

EViews workfile: *tute5discrim.wf1*

tute5discrim.wf1 contains zip code level data i.e. each observation is an area/location/zip code district in the US. Information about each district is held in the following variables:

$$income - median\ family\ income\ in\ a\ zip\ code\ district$$
$$prpblck - proportion\ of\ the\ population\ that\ is\ black\ in\ a\ zip\ code\ district$$
$$prppov - proportion\ of\ the\ population\ that\ is\ in\ poverty\ in\ a\ zip\ code\ district$$
$$psoda - price\ of\ medium\ soda\ in\ a\ zip\ code\ district$$

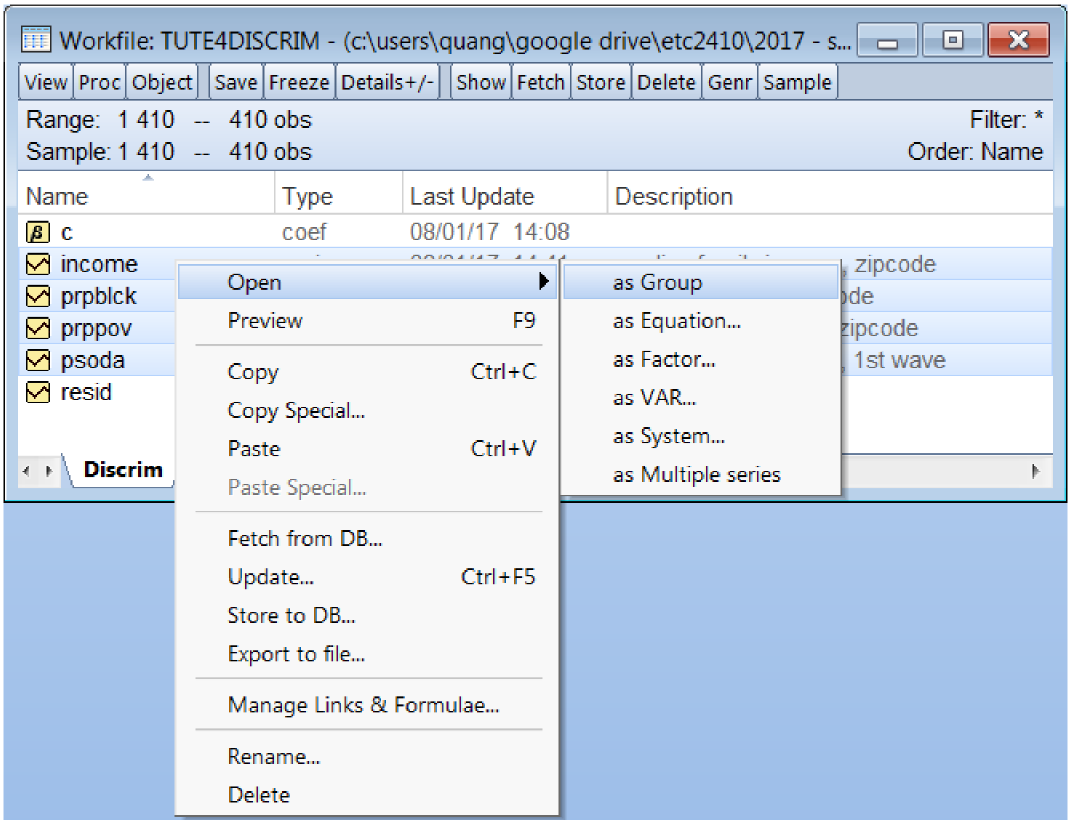
	INCOME	PRPBLCK	PRPPOV	PSODA
1	44534	0.171154	0.036579	1.12
2	44534	0.171154	0.036579	1.06
3	41164	0.047360	0.087907	1.06
4	50366	0.052839	0.059123	1.12
5	72287	0.034480	0.025415	1.12
6	44515	0.059133	0.083500	1.06
7	62056	0.018677	0.029235	1.17
8	53655	0.004906	0.033760	1.17
9	31314	0.921056	0.203682	1.18
10	31314	0.921056	0.203682	1.17
11	31314	0.921056	0.203682	1.06
12	31314	0.921056	0.203682	1.06
13	31314	0.921056	0.203682	1.05
14	38569	0.013911	0.084540	1.17
15	60657	0.010212	0.059816	1.15
16	60657	0.010212	0.059816	1.27
17	47891	0.006090	0.038651	1.06
18	36705	0.003541	0.111877	1.06
19	43022	0.010452	0.060849	1.06
20	79025	0.007387	0.013591	1.20

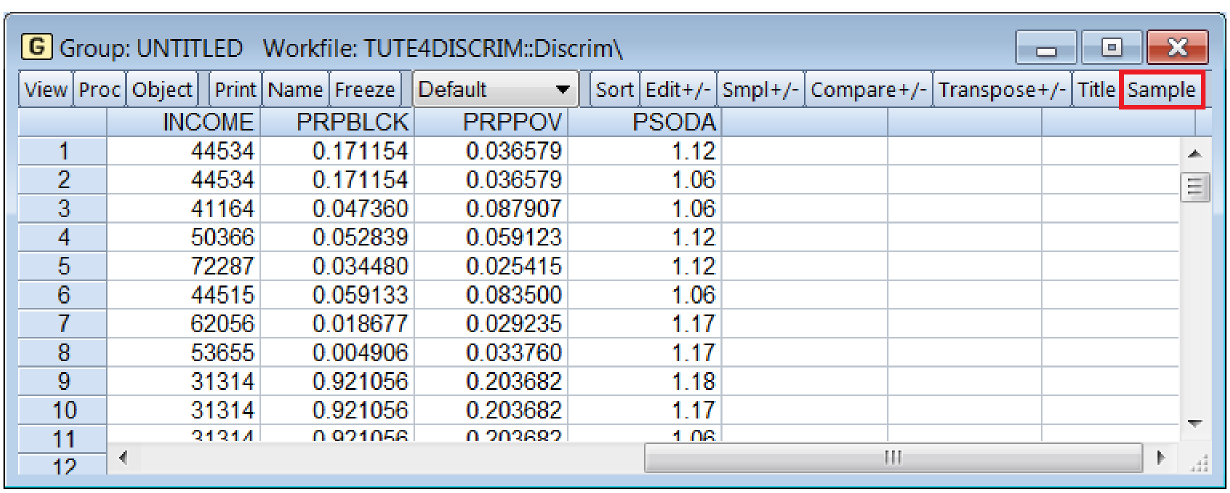
Table 2: Data on *income*, *prpblck*, *prppov* and *psoda* for the first 20 observations in our sample of 410 districts (there are some missing values in our sample).

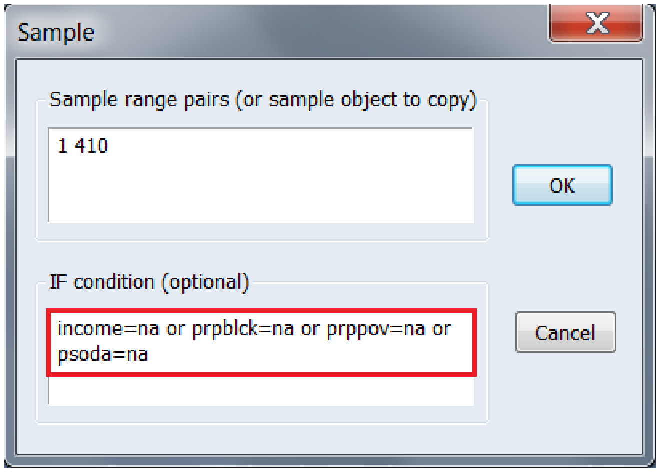
Use the data to see if fast-food restaurants charge higher prices in areas with a large concentration of blacks.

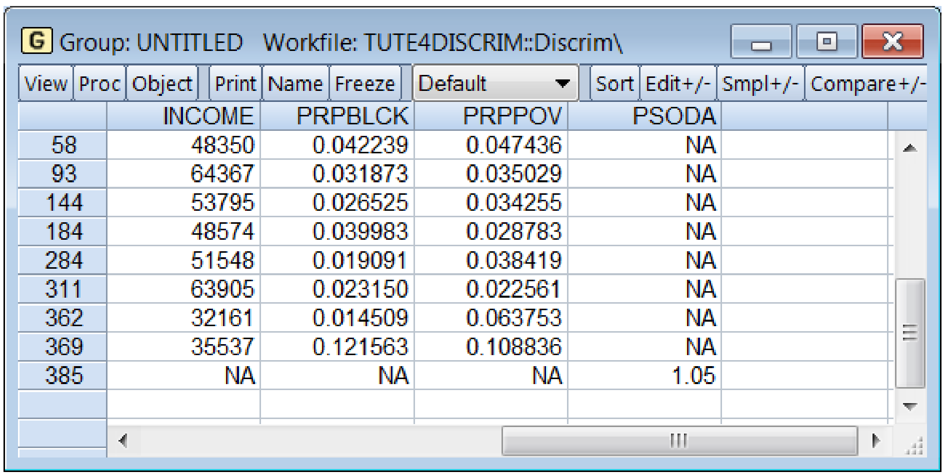
(i) Find the sample mean and sample standard deviation of *prpblck* and *income*. What are the units of measurement of *prpblck* and *income*?

Some of the observations in our data set contains missing values. Can see this when sorting our data,



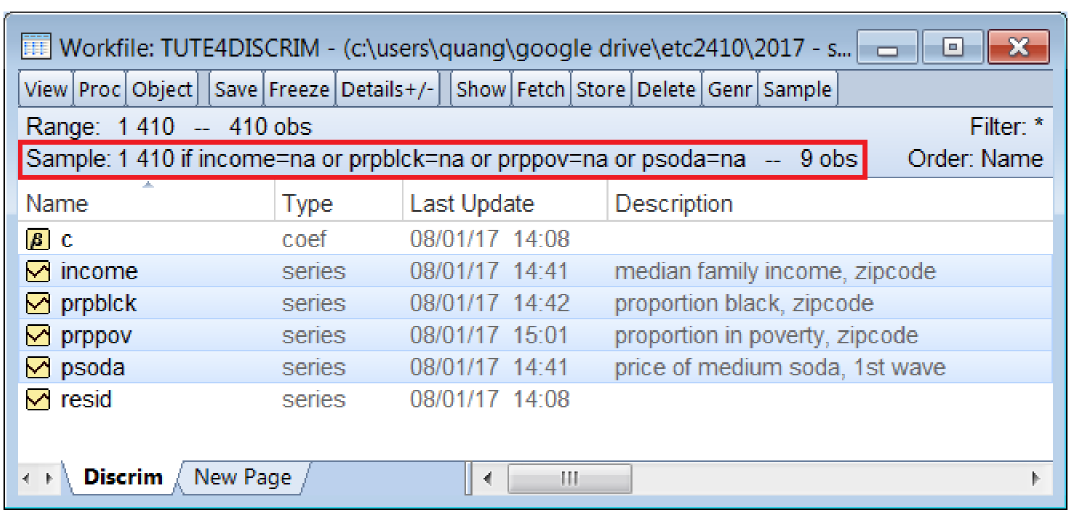




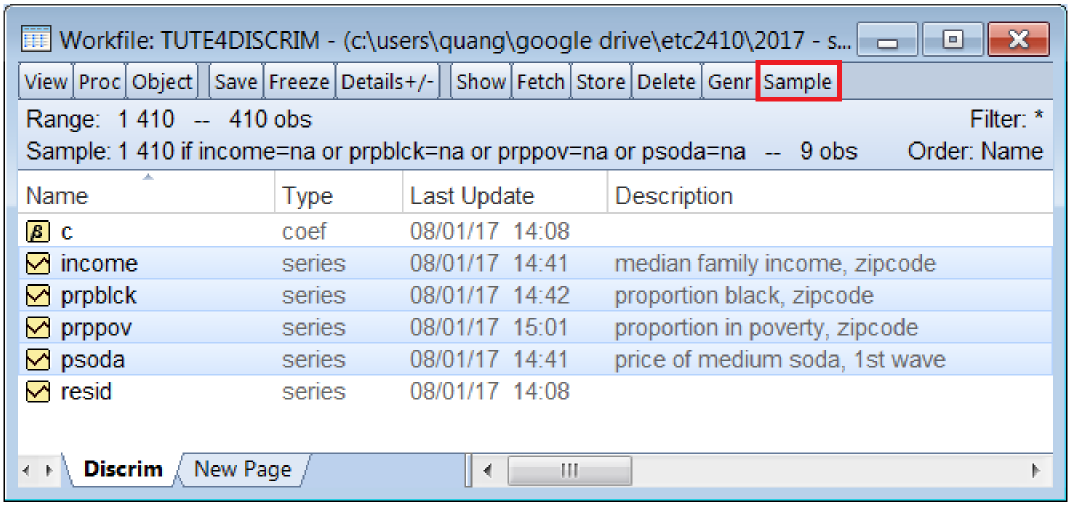


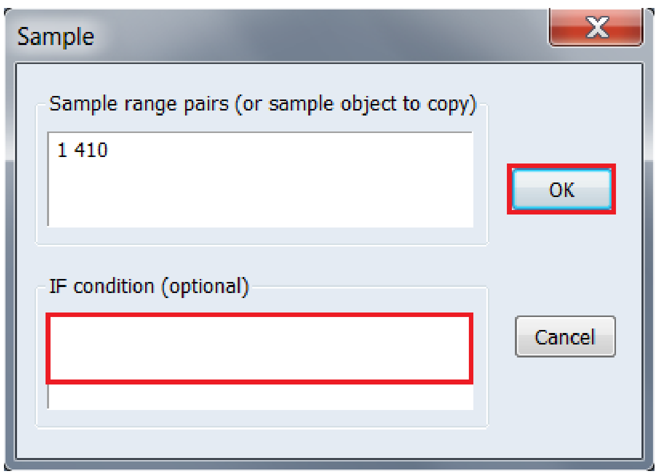
	INCOME	PRPBLCK	PRPPOV	PSODA
58	48350	0.042239	0.047436	NA
93	64367	0.031873	0.035029	NA
144	53795	0.026525	0.034255	NA
184	48574	0.039983	0.028783	NA
284	51548	0.019091	0.038419	NA
311	63905	0.023150	0.022561	NA
362	32161	0.014509	0.063753	NA
369	35537	0.121563	0.108836	NA
385	NA	NA	NA	1.05

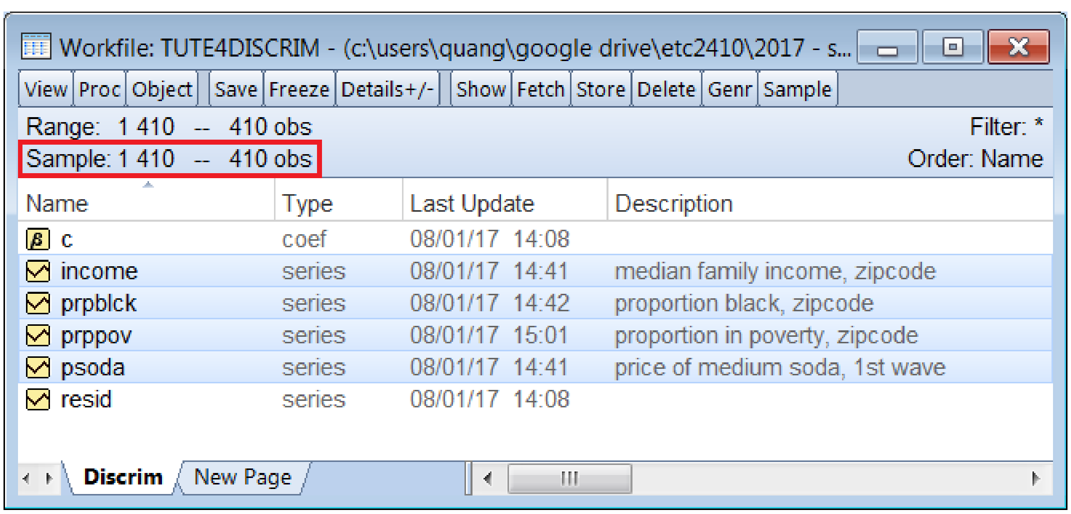
We can see that the 385^{th} district in our sample data set has a missing value for *income*, *prpblck* and *prppov*. The $58^{th}, 93^{rd}, 144^{th}, 284^{th}, 311^{th}, 362^{nd}$ & 369^{th} district in our sample data set as a missing value for *psoda*.



(Ensure that sample is set back to the original data set)

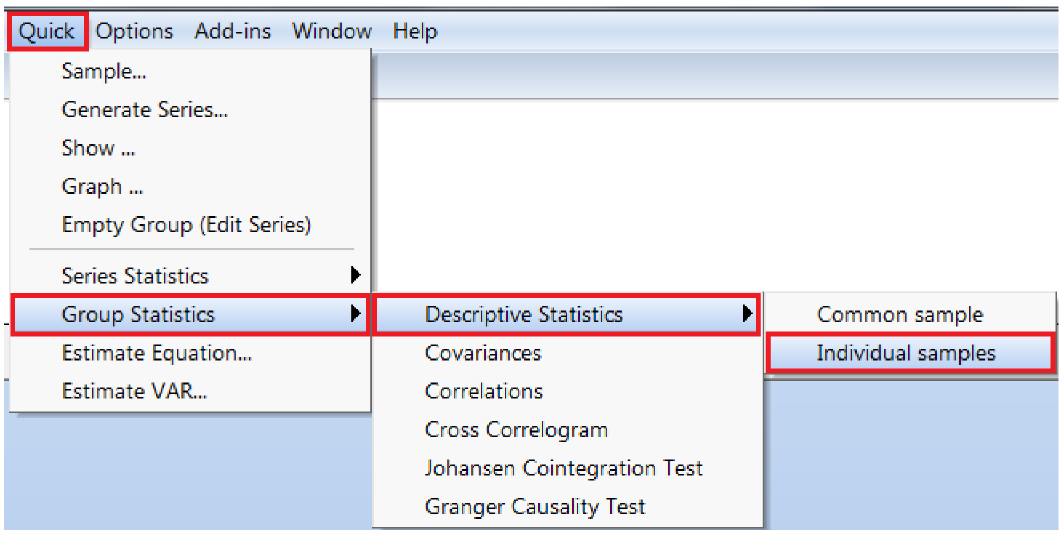




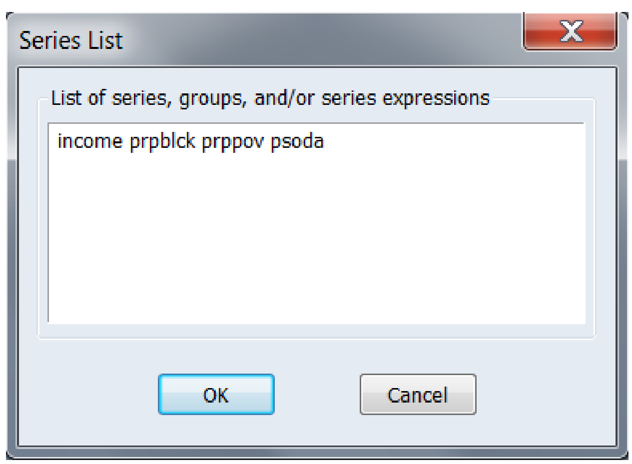


Because of the missing values in our data set, we should be careful when obtaining summary statistics for a group of variables. To obtain summary statistics for the variables the *prpblck*, *income*, *prppov* and *psoda* with each variable's individual sample in EViews,

Quick → *Group Statistics* → *Descriptive Statistics* → *Individual Sample*



then type in the variables of interest in the *Series List* dialog box,

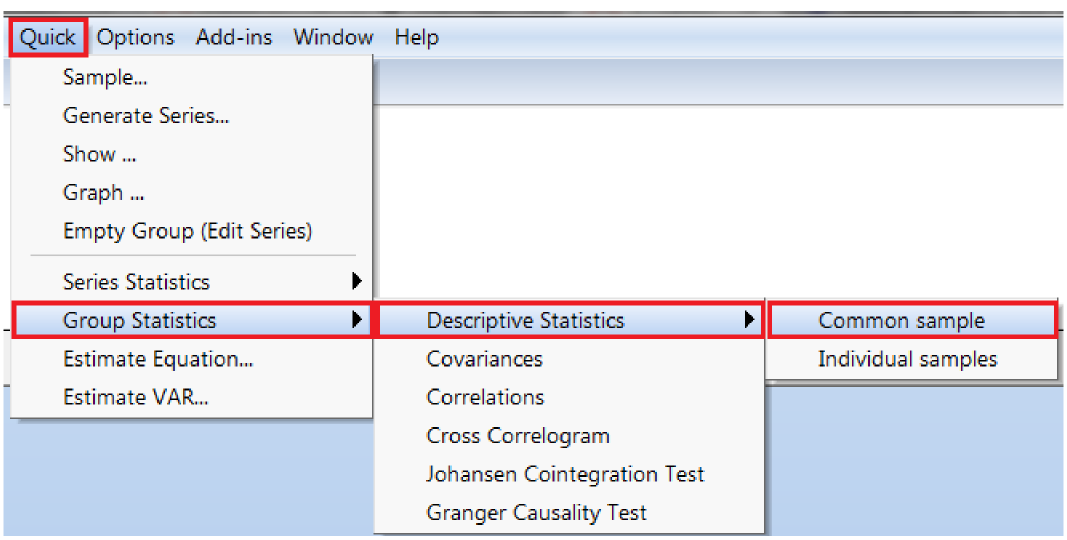


	INCOME	PRPBLCK	PRPPOV	PSODA
Mean	47053.78	0.113486	0.071297	1.044876
Median	46272.00	0.041444	0.044441	1.060000
Maximum	136529.0	0.981658	0.418480	1.490000
Minimum	15919.00	0.000000	0.004298	0.730000
Std. Dev.	13179.29	0.182416	0.067439	0.088687
Skewness	0.962831	2.700012	2.222999	0.348905
Kurtosis	7.551386	10.56841	8.212019	4.582298
Jarque-Bera	416.2135	1473.100	799.8001	50.09267
Probability	0.000000	0.000000	0.000000	0.000000
Sum	19244998	46.41594	29.16060	420.0400
Sum Sq. Dev.	$7.09E + 10$	13.57651	1.855573	3.154044
Observations	409	409	409	402

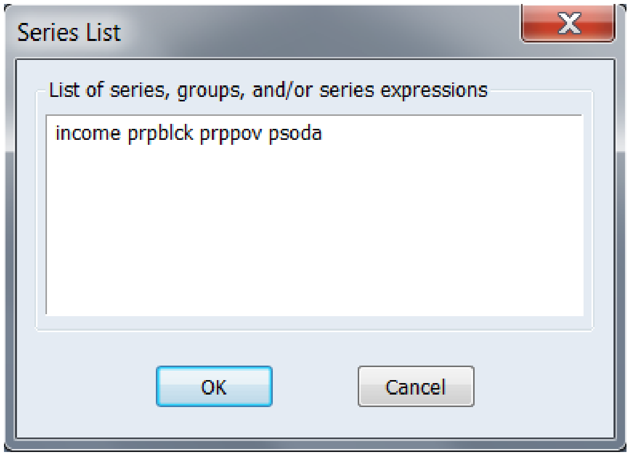
Table 3: Descriptives statistics of *median family income*, *proportion of the population that is black*, *proportion of the population in poverty* and *price of medium soda* for each variable's individual sample of districts.

To obtain summary statistics for the variables the *prpblck*, *income*, *prppov* and *psoda* using the *common sample* in EViews,

Quick → *Group Statistics* → *Descriptive Statistics* → *Common Sample*



then type in the variables of interest in the *Series List* dialog box,



	INCOME	PRPBLCK	PRPPOV	PSODA
Mean	46999.40	0.114955	0.071774	1.044863
Median	46255.00	0.042239	0.044441	1.060000
Maximum	136529.0	0.981658	0.418480	1.490000
Minimum	15919.00	0.000000	0.004298	0.730000
Std. Dev.	13215.33	0.183875	0.067924	0.088798
Skewness	0.980441	2.666880	2.200406	0.348907
Kurtosis	7.615445	10.35573	8.075490	4.571177
Jarque-Bera	420.1710	1379.368	754.0096	49.38217
Probability	0.000000	0.000000	0.000000	0.000000
Sum	18846761	46.09700	28.78153	418.9900
Sum Sq. Dev.	$6.99E+10$	13.52401	1.845495	3.154017
Observations	401	401	401	401

Table 4: Descriptives statistics of *median family income*, *proportion of the population that is black*, *proportion of the population in poverty* and *price of medium soda* using a common sample.

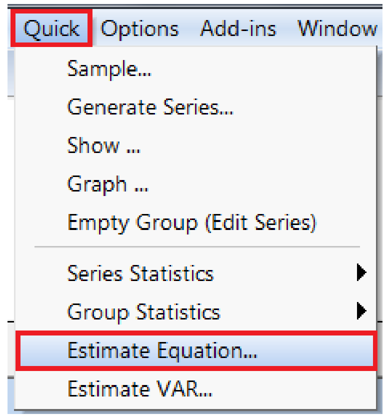
(ii) Consider a model to explain the price of soda in a district, *psoda* in terms of the proportion of the population that is black in a district (*prpblck*) and the median income in a district (*income*)

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + u$$

Estimate this model by OLS and report the results in equation form, including the sample size and R^2. (Do not use scientific notation when reporting the estimates.)

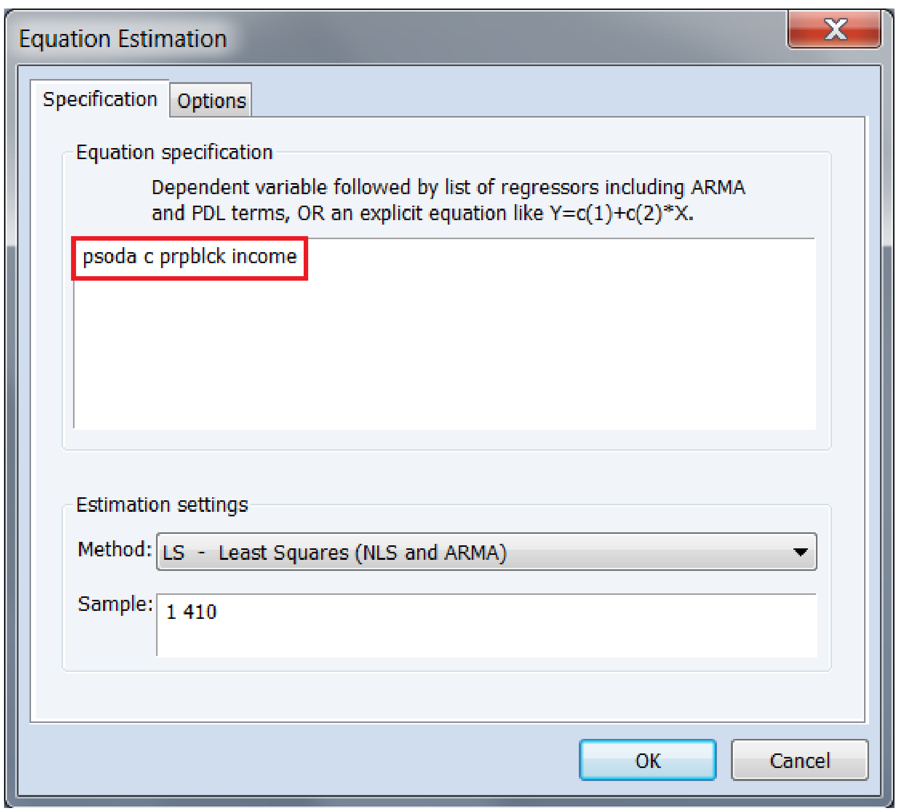
To estimate this model in EViews,

$$Quick \to Estimate\ Equation$$



then in the *Equation Estimation* dialog box type in,

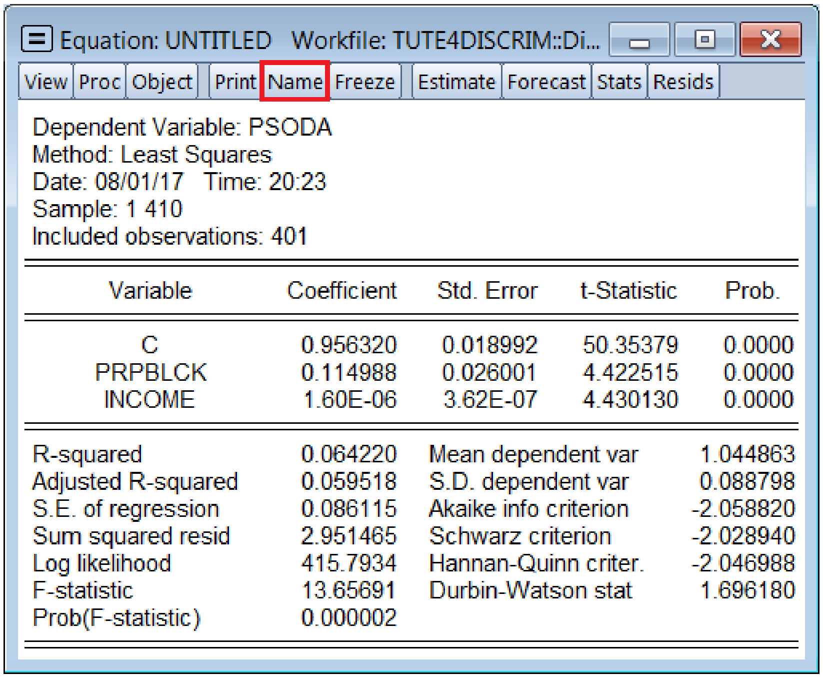
$$psoda\ c\ prpblck\ income$$



To name (save) the estimated equation,

$$Name \rightarrow Name\ to\ identify\ object : eq01$$

$$(This\ names\ the\ equation\ \boldsymbol{eq01})$$



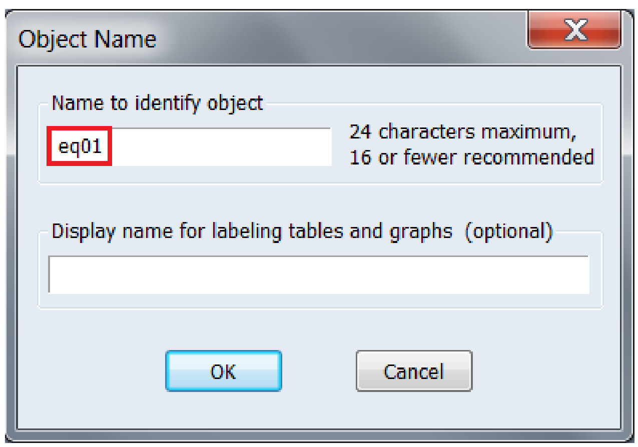
Equation: UNTITLED Workfile: TUTE4DISCRIM::Di...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: PSODA
Method: Least Squares
Date: 08/01/17 Time: 20:23
Sample: 1 410
Included observations: 401

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.956320	0.018992	50.35379	0.0000
PRPBLCK	0.114988	0.026001	4.422515	0.0000
INCOME	1.60E-06	3.62E-07	4.430130	0.0000

R-squared	0.064220	Mean dependent var	1.044863
Adjusted R-squared	0.059518	S.D. dependent var	0.088798
S.E. of regression	0.086115	Akaike info criterion	-2.058820
Sum squared resid	2.951465	Schwarz criterion	-2.028940
Log likelihood	415.7934	Hannan-Quinn criter.	-2.046988
F-statistic	13.65691	Durbin-Watson stat	1.696180
Prob(F-statistic)	0.000002		



The estimated model,

$$\widehat{psoda} = \hat{\beta}_0 + \hat{\beta}_1 prpblck + \hat{\beta}_2 income$$

Interpret the coefficient on *prpblck* (with a 0.1 increase in *prpblck*)

$\hat{\beta}_1 = 0.1150$

The model estimates that for a 0.1 increase in the proportion of blacks in a district i.e. a 10 percentage-point increase (not a 10 percent increase), the price of soda in that district will increase by $0.1150 \times 0.1 = 0.0115$ i.e. \$0.0115 or about 1.2 cents, on average, holding median family income constant.

Is $\hat{\beta}_1 = 0.1150$ economically large?

Although a 1.2 cent increase in soda price for a 0.1 increase in the proportion of the population in a district that is black, holding median family income constant, does not seem large, if we compare the soda price between districts with and without a black population, holding median family income constant, we find that the difference is 11.50 cent,

$$\Delta income = 0$$

$$\Delta prpblck = 1$$

$$\begin{aligned}\Delta\widehat{psoda} &= \hat{\beta}_1 \Delta prpblck + \hat{\beta}_2 \Delta income \\ &= \hat{\beta}_1 \times 1 \\ &= 0.1150\end{aligned}$$

Whether this is large depends on the average soda price. From our sample of districts, the average soda price was \$1.04 so an \$11.5 cent difference would seem large.

(iv) Reporting estimated models and rescaling

Data initially obtained may not be in a convenient scale for regression analysis. In our example, *psoda* and *income* are both is measured in dollars and we obtained the following estimated model,

$$\widehat{psoda} =$$

The interpretation of the estimated coefficient of income,

When a district's median family income increases by \$1, we estimate that the price of soda in that district will increase by \$0.00000016, holding district's proportion of the population that is black constant.

Although nothing is mathematically wrong with this, it leads to a discussion of changes that are so small as to seem irrelevant. By changing the unit of measurement of *psoda* to cent, we obtain the following estimated model,

Dependent Variable: PSODA_CENT
Method: Least Squares
Date: 08/12/17 Time: 16:08
Sample: 1 410
Included observations: 401

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	95.63197	1.899201	50.35379	0.0000
PRPBLCK	11.49882	2.600064	4.422515	0.0000
INCOME	0.000160	$3.62E-05$	4.430130	0.0000

R-squared	0.064220	Mean dependent var	104.4863
Adjusted R-squared	0.059518	S.D. dependent var	8.879777
S.E. of regression	8.611470	Akaike info criterion	7.151520
Sum squared resid	29514.65	Schwarz criterion	7.181400
Log likelihood	−1430.880	Hannan-Quinn criter.	7.163352
F-statistic	13.65691	Durbin-Watson stat	1.696180
Prob(F-statistic)	0.000002		

Table 5: Regression output of the *price of soda in cents* on a constant, the *proportion of the population that is black* and *median family income in $.*

$$\widehat{psoda\_cent} =$$

and the interpretation of the estimated coefficient of *income* becomes,

When a district's median family income increases by $1, we estimate that the price of soda in that district increases by 0.00016 cent, holding the proportion of the population that is black in a district constant.

When the price of soda is rescaled from dollars to cents we do so by multiplying the original variable *psoda* by 100,

$$psoda\_cent = 100psoda$$

and the estimated coefficients will rescale according,

$$\hat{\beta}_0^* = 100\hat{\beta}_0$$
$$\hat{\beta}_1^* = 100\hat{\beta}_1$$
$$\hat{\beta}_2^* = 100\hat{\beta}_2$$

where,

$$\widehat{psoda} = \hat{\beta}_0 + \hat{\beta}_1 prpblck + \hat{\beta}_2 income$$
$$\widehat{psoda\_cent} = 100\hat{\beta}_0 + 100\hat{\beta}_1 prpblck + 100\hat{\beta}_2 income$$
$$\widehat{psoda\_cent} = \hat{\beta}_0^* + \hat{\beta}_1^* prpblck + \hat{\beta}_2^* income$$

If we also rescale median family income from dollars to \$'000, we obtain the following estimated model,

$$\widehat{psoda\_cent} = \hat{\beta}_0^* + \hat{\beta}_1^* prpblck + 1000\hat{\beta}_2^* income\_thousand$$

Dependent Variable: PSODA_CENT
Method: Least Squares
Date: 08/12/17 Time: 16:28
Sample: 1 410
Included observations: 401

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	95.63197	1.899201	50.35379	0.0000
PRPBLCK	11.49882	2.600064	4.422515	0.0000
INCOME_THOUSAND	0.160267	0.036177	4.430130	0.0000

R-squared	0.064220	Mean dependent var	104.4863
Adjusted R-squared	0.059518	S.D. dependent var	8.879777
S.E. of regression	8.611470	Akaike info criterion	7.151520
Sum squared resid	29514.65	Schwarz criterion	7.181400
Log likelihood	−1430.880	Hannan-Quinn criter.	7.163352
F-statistic	13.65691	Durbin-Watson stat	1.696180
Prob(F-statistic)	0.000002		

Table 6: Regression output of the *price of soda in cents* on a constant, the *proportion of the population that is black* and *median family income in \$'000.*

$$\widehat{psoda\_cent} =$$

which provides a more meaningful interpretation,

When a district's median family income increases by \$1,000, the model estimates that the price of soda in that district will increase by 0.16 cents, holding the proportion of the population that is black in a district constant.

And if we also express *prpblck* in percentage points,

$$prpblck\% = 100 \times prpblck$$

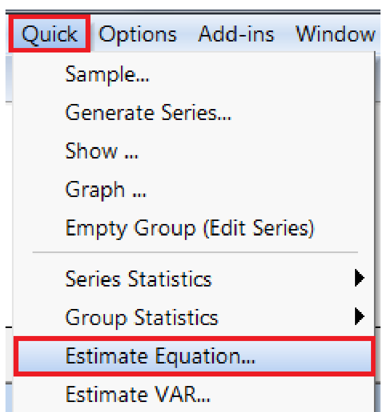
$$\widehat{psoda\_cent} = \hat{\beta}_0^* + \frac{1}{1000}\hat{\beta}_1^* prpblck\% + 1000\hat{\beta}_2^* income\_thousand$$

As we can see, the data has been rescaled without changing the real underlying relationship between the price of soda and median family income. The interpretation remains mathematically correct and the magnitudes are more relevant and easy for discussion.

(iii) Compare the estimate from part (ii) with the simple regression estimate from *psoda* on *prpblck*. Is the discrimination effect larger or smaller when you control for *income*?

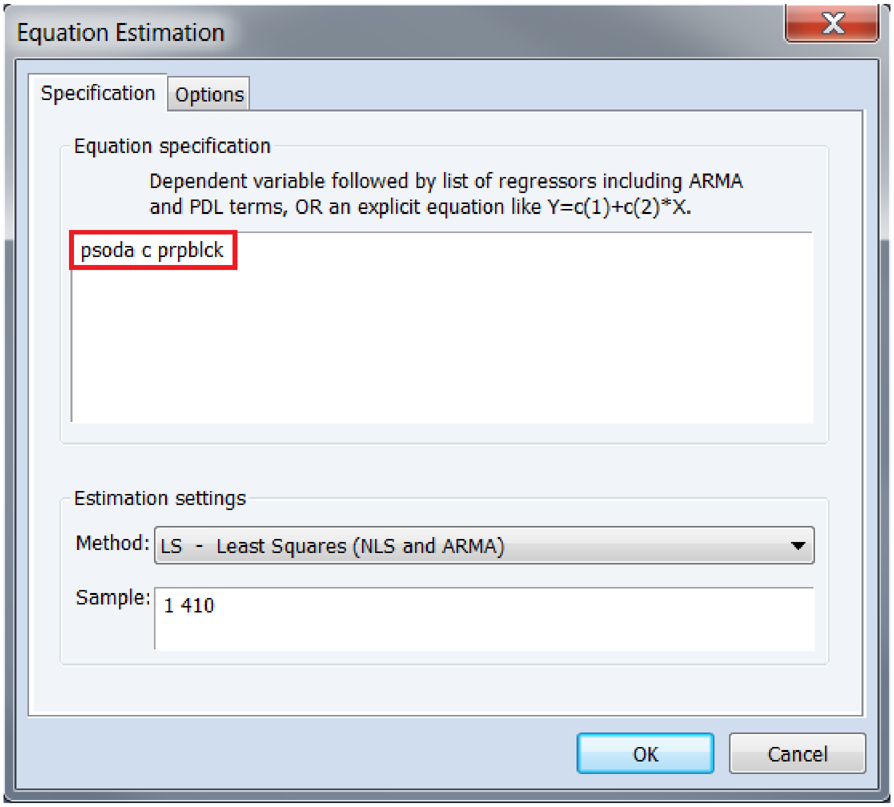
To estimate this model in EViews,

$$Quick \rightarrow Estimate\ Equation$$



then in the *Equation Estimation* dialog box type in,

$$psoda\ c\ prpblck$$



Dependent Variable: PSODA
Method: Least Squares
Date: 08/12/17 Time: 17:14
Sample: 1 410
Included observations: 401

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.037399	0.005190	199.8668	0.0000
PRPBLCK	0.064927	0.023957	2.710146	0.0070

R-squared	0.018076	Mean dependent var	1.044863
Adjusted R-squared	0.015615	S.D. dependent var	0.088798
S.E. of regression	0.088102	Akaike info criterion	−2.015673
Sum squared resid	3.097007	Schwarz criterion	−1.995753
Log likelihood	406.1425	Hannan-Quinn criter.	−2.007785
F-statistic	7.344894	Durbin-Watson stat	1.611081
Prob(F-statistic)	0.007015		

Table 7: Regression output of *psoda* on a constant and *prpblck*.

$$\widehat{psoda} =$$

The discrimination effect is larger then we control for median family income. For our estimated simple and multiple regression model (not controlling then controlling median family income),

$$\widehat{psoda} = \hat{\alpha}_0 + \hat{\alpha}_1 prpblck$$

$$\widehat{psoda} = \hat{\beta}_0 + \hat{\beta}_1 prpblck + \hat{\beta}_2 income$$

$\hat{\alpha}_1$ and $\hat{\beta}_1$ have the following algebraic relationship,

$$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1$$

where $\hat{\delta}_1$ is the estimated slope of coefficient of *income* regressed on a constant and *prpblck*,

$$\widehat{income} = \hat{\delta}_0 + \hat{\delta}_1 prpblck$$

$$\widehat{income} = \underset{(692.7061)}{50608.36} - \underset{(3227.179)}{31321.63} prpblck$$

Dependent Variable: INCOME
Method: Least Squares
Date: 08/13/17 Time: 05:02
Sample: 1 410
Included observations: 409

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	50608.36	692.7061	73.05892	0.0000
PRPBLCK	−31321.63	3227.179	−9.705576	0.0000

R-squared	0.187946	Mean dependent var	47053.78
Adjusted R-squared	0.185951	S.D. dependent var	13179.29
S.E. of regression	11890.97	Akaike info criterion	21.60982
Sum squared resid	$5.75E + 10$	Schwarz criterion	21.62945
Log likelihood	−4417.209	Hannan-Quinn criter.	21.61759
F-statistic	94.19820	Durbin-Watson stat	1.035961
Prob(F-statistic)	0.000000		

Table 8: Regression output of *income* on a constant and *prpblck*.