Introductory Econometrics

Tutorial 4 Solutions

<u>PART A:</u> To be done before you attend the tutorial. The tutors will ask you questions based on this part and that will be the basis for your participation point. The solutions will be made available at the end of the week.

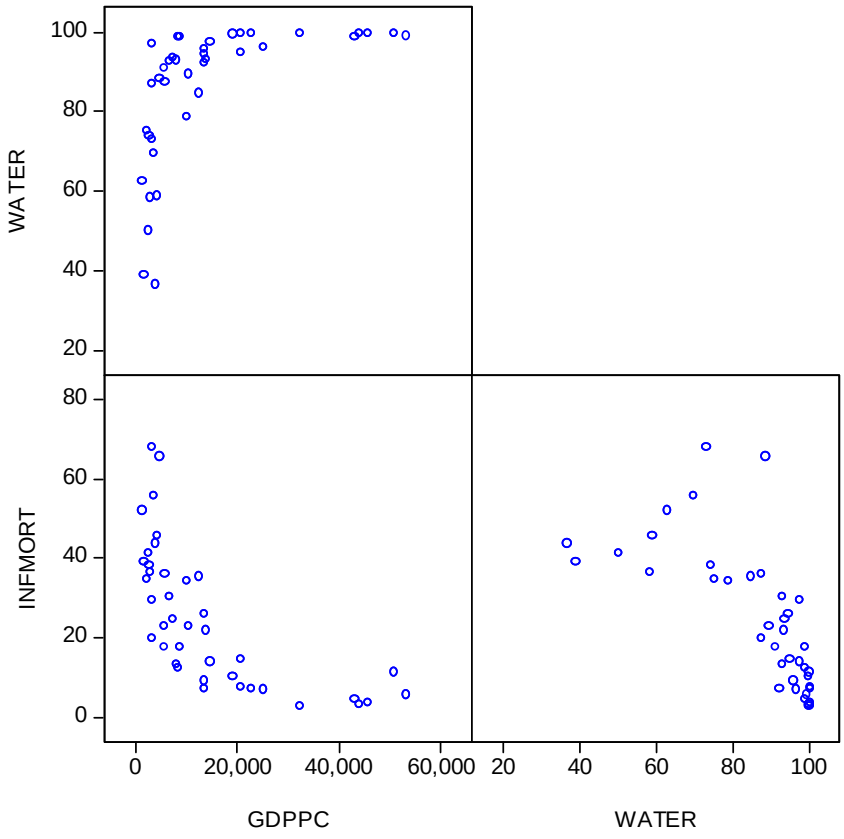
In this exercise, you need to continue working with WDI.xlsx data set that you used in tutorial 2.

1. This part is data cleaning, a very important step in the real world because in the real world data sets are messy. If you have any problems, please go to consultation hours of any member of the teaching team and sort it out.

2. Four countries (Kosovo, St. Kitts and Nevis, St. Martin and Uzbekistan) have missing data in 2015. Countries with very low access to basic drinking water are poor under-developed countries, and on the other extreme, countries with 100% access to basic drinking water are mostly developed countries or rich developing countries, although there are exceptions such as Nauru.

3. More data work.

4. The graphs show that there is a negative relationship between infant mortality and each of GDP per capita and access to basic drinking water services. It also shows a positive relationship between GDPPC and WATER. These all make sense. The scatter plots further show that these relationships are not linear. Perhaps the most linear one is the relationship between INFMORT and WATER. These make sense because given the state of the medical knowledge in 2015, there is a limit on how low infant mortality can get, and when a country achieves that, further changes in GDPPC or access to drinking services (or any other variable) cannot push mortality lower than that. Also, % of population using basic drinking services cannot go above 100. For countries with 100% access to drinking water, changes in other variables cannot be associated with any further increase in access to drinking water.



5. Sample correlation coefficients also show negative correlation of similar magnitude between infant mortality and GDP per capita and infant mortality and basic drinking water services, and positive relationship between drinking water services and GDP per capita.

	GDPPC	WATER	INFMORT
GDPPC	1.000000	0.554186	-0.681636
WATER	0.554186	1.000000	-0.699874
INFMORT	-0.681636	-0.699874	1.000000

6. The $\mathbf{y}$ vector will be 39×1 (39 because Uzbekistan has missing data on water, therefore that observation is dropped automatically by the software), and the $\mathbf{X}$ matrix will be 39×3. The $\boldsymbol{\beta}$ vector will be 3×1. The first 3 rows of $\mathbf{y}$ and $\mathbf{X}$ are:

$$\mathbf{y} = \begin{bmatrix} 21.9 \\ 10.3 \\ 12.5 \\ \vdots \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 13724.72 & 93.46643 \\ 1 & 19101.30 & 99.62730 \\ 1 & 8195.934 & 98.92365 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Do not forget to bring your answers to PART A and a copy of the tutorial questions to your tutorial.

PART B: You do not need to hand this part in. It will be covered in the tutorial. It is still a good idea to attempt these questions before the tutorial.

1. (*Post-multiplying a matrix by a vector produces a linear combination of the columns of the matrix):* Let

$$\mathbf{X} = \begin{bmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}$$

and

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 0.7 \\ 0.2 \end{bmatrix}.$$

Compute $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, and show that the result is 0.7 times the first column of $\mathbf{X}$ plus 0.2 times the second column of $\mathbf{X}$.

- $\widehat{\mathbf{y}} = \begin{bmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0.7 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \times 0.7 + \begin{bmatrix} 3 \\ 2 \\ 2 \\ 1 \end{bmatrix} \times 0.2 = \begin{bmatrix} 1.3 \\ 1.1 \\ 1.1 \\ 0.9 \end{bmatrix}$

2. Let's generalise the result in question 1. Suppose

$$\underset{n\times 3}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$$

and

$$\underset{3\times 1}{\widehat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

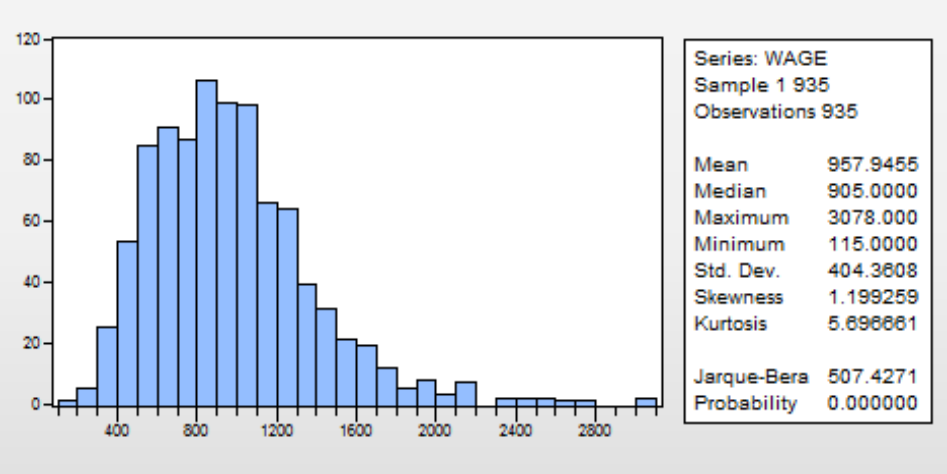
Show that $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ is an $n \times 1$ vector which is a linear combination (a weighted sum) of columns of $\mathbf{X}$ with weights given by the elements of $\widehat{\boldsymbol{\beta}}$. That is:

$$\widehat{\mathbf{y}} = \text{first column of } \mathbf{X}\times\hat{\beta}_1 + \text{second column of } \mathbf{X}\times\hat{\beta}_2 + \text{third column of } \mathbf{X}\times\hat{\beta}_3$$

In fact this is not specific to $\mathbf{X}$ having 3 columns. It is true for any $n \times k$ matrix $\mathbf{X}$ and $k \times 1$ vector $\widehat{\boldsymbol{\beta}}$.

- $\widehat{\mathbf{y}} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} x_{11}\hat{\beta}_1 + x_{12}\hat{\beta}_2 + x_{13}\hat{\beta}_3 \\ x_{21}\hat{\beta}_1 + x_{22}\hat{\beta}_2 + x_{23}\hat{\beta}_3 \\ \vdots \\ x_{n1}\hat{\beta}_1 + x_{n2}\hat{\beta}_2 + x_{n3}\hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} \times\hat{\beta}_1 + \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} \times\hat{\beta}_2 + \begin{bmatrix} x_{13} \\ x_{23} \\ \vdots \\ x_{n3} \end{bmatrix} \times\hat{\beta}_3$

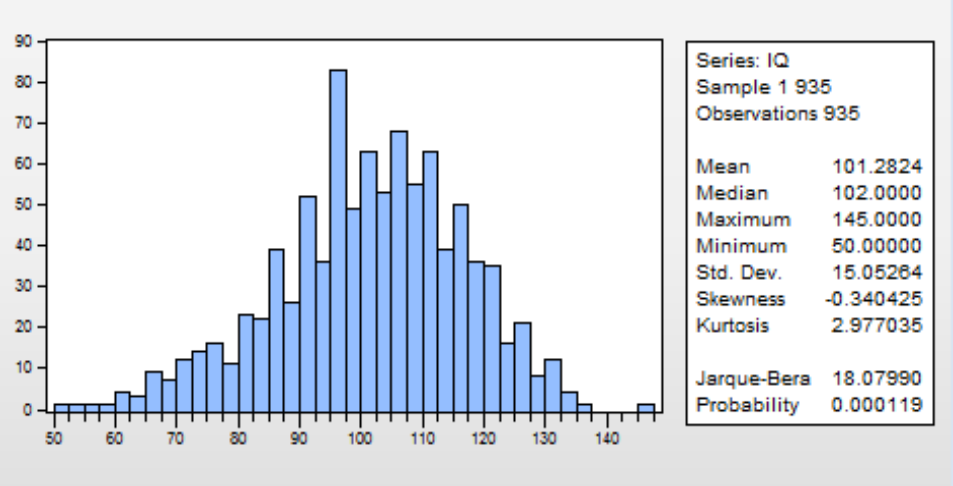
3. This question is based on question C4 in Chapter 2 of the textbook. The dependent variable is *wage* and the independent variable is IQ.
Preliminary analysis of the data (not asked in the question, but a step that we usually take before estimating a regression equation: "Look" at these variables (meaning that examine their histograms, summary statistics, scatter plot of wage against IQ, sample correlation coefficient between wage and IQ, and summarise your insights from just looking at data through these views).



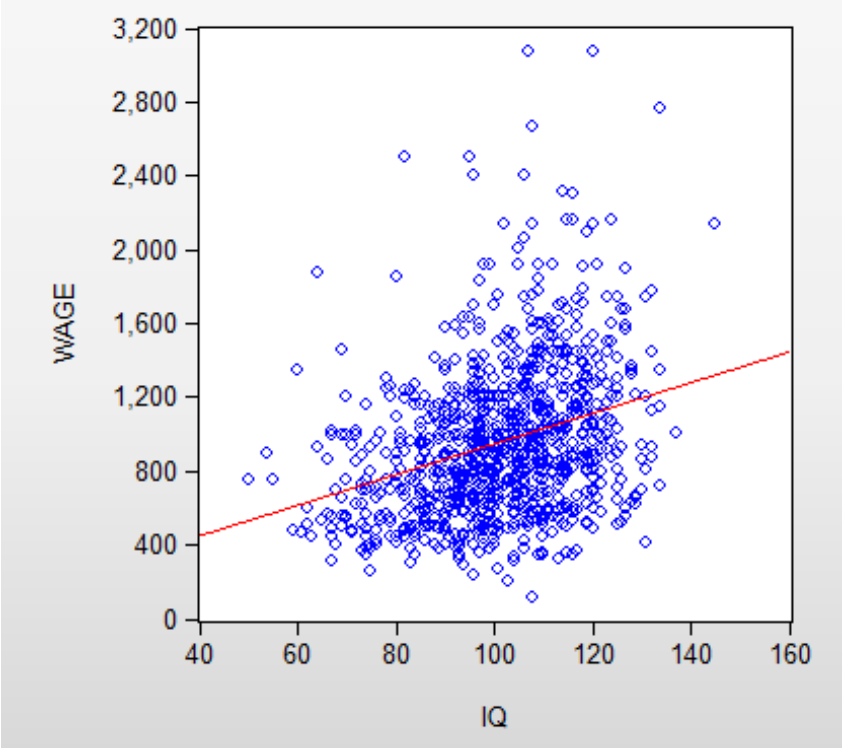
Wage is positively skewed. The sample mean is \$958, but because of this skewness the median \$905 is a better indicator of the central tendency than the mean. Need to make sure that the

one outlying observation with \$3078 wage does not affect the analysis.

IQ seems to be representative, with mean close to 100 and standard deviation close to 15. There is a very smart person with IQ of 145, so need to ensure that it does not influence the analysis too much.



The scatter plot of wage against IQ, in particular with the regression line included, shows that the relationship between wage and IQ, although seems positive (sample correlation coefficient is 0.309), is not linear. The variation of wage around the regression line seems to be higher at higher IQs.



Sample: 1 935
Included observations: 935

Correlation	IQ	WAGE
IQ	1.000000	
WAGE	0.309088	1.000000

(a) *Estimation, interpretation of the slope coefficient and R^2 of the regression*: Estimate a simple regression model where a one-point increase in *IQ* changes *wage* by a constant dollar amount. Use this model to find the predicted increase in *wage* for an increase in *IQ* of 15 points. Does *IQ* explain most of the variation in *wage*? What is the relationship between the R^2 of this regression and the sample correlation coefficient between wage and IQ? Name your estimated equation **eq01.** Save the residuals of this regression in a variable called **uhat01.** points. Does *IQ* explain most of the variation in *wage*? Name your estimated equation **eq01.** Save the residuals of this regression in a variable called **uhat01.**

- Dependent Variable: WAGE
  Method: Least Squares
  Date: 07/28/16 Time: 12:46
  Sample: 1 935
  Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	116.9916	85.64153	1.366061	0.1722
IQ	8.303064	0.836395	9.927203	0.0000

R-squared	0.095535	Mean dependent var	957.9455
Adjusted R-squared	0.094566	S.D. dependent var	404.3608
S.E. of regression	384.7667	Akaike info criterion	14.74529
Sum squared resid	1.38E+08	Schwarz criterion	14.75564
Log likelihood	-6891.422	Hannan-Quinn criter.	14.74924
F-statistic	98.54936	Durbin-Watson stat	1.802114
Prob(F-statistic)	0.000000		

  If *IQ* increases by 15 points (*i.e.* one standard deviation), the predicted increase in wage is $8.303 \times 15 = 124.545$ (get students to use their calculators).
  The R^2 is very low, so there is a lot of unexplained variation. Note that in a regression with only one explanatory variable, R^2 is the square of sample correlation coefficient between the dependent variable and the explanatory variable. Ask them to verify that on their calculators $0.309088^2 = 0.095535$.

(b) Interpretation of the intercept: What does the intercept in eq01 mean? Now, run a regression of wage on a constant and (IQ-100), and name it eq02. Compare the results with your results in eq01 and note all similarities and differences. Save the residuals of this regression in a variable called uhat02. Open uhat01 and uhat02 side by side and see if they are different. What is the interpretation of the intercept in eq02?

- Intercept in eq01 does not have any meaningful interpretation because there is no

individual with IQ of zero.

Dependent Variable: WAGE
Method: Least Squares
Date: 07/28/16 Time: 12:48
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	947.2980	12.62885	75.01066	0.0000
IQ-100	8.303064	0.836395	9.927203	0.0000

R-squared	0.095535	Mean dependent var	957.9455
Adjusted R-squared	0.094566	S.D. dependent var	404.3608
S.E. of regression	384.7667	Akaike info criterion	14.74529
Sum squared resid	1.38E+08	Schwarz criterion	14.75564
Log likelihood	-6891.422	Hannan-Quinn criter.	14.74924
F-statistic	98.54936	Durbin-Watson stat	1.802114
Prob(F-statistic)	0.000000		

Only the intercept has changed. The estimate of the slope, its standard error, R-squared and the standard error of regression are all exactly the same as in eq01. The intercept, however, is now meaningful. It shows the predicted wage for a person with IQ of 100, i.e. average IQ.

(c) Discuss what you learned from this exercise.

- Among other things, we learn that if we subtract a constant from one of the explanatory variables, only the OLS estimator of the intercept will change. Everything else, in particular the estimate of the slope, its standard error, the residuals and the predicted wage for all observations will be exactly the same as before. Geometrically, this is because when we add or subtract a multiple of one column to another column of a matrix $\mathbf{X}$ its column space does not change (we are sliding one vector up or down another vector). So, the orthogonal projection of wage on this space will stay the same as before.We can also predict how the intercept will change because:

$$\begin{bmatrix} 1 & IQ_1 \\ 1 & IQ_2 \\ \vdots & \vdots \\ 1 & IQ_n \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1 & IQ_1 - 100 \\ 1 & IQ_2 - 100 \\ \vdots & \vdots \\ 1 & IQ_n - 100 \end{bmatrix} \begin{bmatrix} b_1 + 100b_2 \\ b_2 \end{bmatrix}$$

for any $\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$. So, by observing eq01, we could have guessed that the estimate of intercept in eq02 was going to be $116.9916 + 100 \times 8.303064 = 947.298$ exactly!

4. *(Regression on dummy variables):* Consider a data set in which each observation must belong to only one of two categories. For example, a data set on wage of random sample of n observations from the population of employed people. Each person in the data set is either male or female. Obviously, $n_1 + n_2 = n$. The dummy variable female is a binary variable that is equal to 1 if the individual is female and 0 otherwise. The dummy variable male is a binary variable that is equal to 1 if the individual is male and zero otherwise. Consider regression of wage on these two dummy variables (no constant).

(a) Sketch the $\mathbf{X}$ matrix for this regression (for ease of notation, you can assume that the first n_1 observations are female and the last n_2 observations are male. Obviously, $n_1 + n_2 = n$.). Are the columns of $\mathbf{X}$ linearly independent?

Yes the columns of $\mathbf{X}$ are linearly independent because one is not a multiple of the other.

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} n_1 \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} n_2 \end{matrix}$$

(b) Use the OLS formula $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and derive the OLS estimator in this case. Comment on the result. Verify your result by creating the dummy variable male in wage1tute4 data set and running a regression of wage on female and male (no constant).

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}$$

$$\implies (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{pmatrix} \sum_{i \text{ is a female}} y_i \\ \sum_{i \text{ is a male}} y_i \end{pmatrix}$$

$$\implies \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} \begin{pmatrix} \sum_{i \text{ is a female}} y_i \\ \sum_{i \text{ is a male}} y_i \end{pmatrix} = \begin{pmatrix} \bar{y}_{\text{female}} \\ \bar{y}_{\text{male}} \end{pmatrix}$$

Dependent Variable: WAGE
Method: Least Squares
Date: 02/26/16 Time: 16:31
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
FEMALE	4.587659	0.218983	20.94980	0.0000
MALE	7.099489	0.210008	33.80578	0.0000

R-squared	0.115667	Mean dependent var	5.896103
Adjusted R-squared	0.113979	S.D. dependent var	3.693086

Descriptive Statistics for WAGE
Categorized by values of FEMALE
Date: 02/26/16 Time: 16:33
Sample: 1 526
Included observations: 526

FEMALE	Mean	Std. Dev.	Obs.
0	7.099489	4.160858	274
1	4.587659	2.529363	252
All	5.896103	3.693086	526

(c) Consider now that we had a constant in addition to these two dummy variables. Write down the $\mathbf{X}$ matrix for this case. Are the columns of $\mathbf{X}$ linearly independent? What is the dimension of the column space of $\mathbf{X}$? Remember from first year ETC1000 that you were told that when a regression has a constant, you should only add one dummy variable for an attribute that has two categories (such as male, female). Explain where that rule came from.

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix}$$

No. The columns of $\mathbf{X}$ is now linearly dependent because the first column is the sum of the other two columns. The dimension of the column space of $\mathbf{X}$ is still 2, even though it has 3 columns. One of these columns are redundant, given the other two. Hence, with this $\mathbf{X}$, the matrix $\mathbf{X}'\mathbf{X}$ will not be invertible and OLS estimator cannot be calculated.