

Consequences of violating assumptions of OLS
 Testing our assumptions
 Regression diagnostics
 Multicollinearity

The basic assumptions of OLS are:

1. The X values are fixed - i.e. we usually don't treat them as a *sample*.
2. Errors have zero mean - i.e. $E[\hat{\epsilon}_i] = 0$
3. Errors have the same constant variance (homoscedasticity)
4. Errors are uncorrelated one with another
5. Errors are normally distributed
6. The linear model correctly describes the relationship between Y and the predictor variables. This assumption means that we believe that the predictors are not affected by multicollinearity and that the relationship between Y and a particular X is not better expressed by a curvilinear rather than a linear model.

CONSEQUENCES OF VIOLATING THESE ASSUMPTIONS

Problem	Biased b	Biased SE	Invalid t & F tests	High var(b) [inefficient]
Nonlinear relationship	yes	yes	yes	---
Omit relevant X variable	yes	yes	yes	---
Include irrelevant X	no	no	no	yes
X measured with error	yes	yes	yes	---
Heteroscedasticity	no	yes	yes	yes
Autocorrelation	no	yes	yes	yes
X correlated with $\hat{\epsilon}$	yes	yes	yes	---
Nonnormal $\hat{\epsilon}$	no	no	yes	yes
Multicollinearity	no	no	no	yes

HOW DO WE INVESTIGATE WHETHER ASSUMPTIONS ARE VIOLATED?

Assumptions 1 and 2. We usually simply assume these are true. We do so for two quite different reasons: we have no way of testing (2) and it turns out that violating (1) is not important.

The estimation procedures we use always make the estimated errors have mean = 0, so we have no way of testing whether or not the assumption is true. Many investigations of the effects of our X's actually not being fixed have shown that violating this assumption does not seriously affect any of our results.

Assumptions 3,5,6:

1. Look at correlation matrices and scatterplot matrices to detect multicollinearity, non-linearity and heteroscedasticity
2. Look at plots of residuals vs predicted Y values, again to look for non-linearity and heteroscedasticity
3. Use band regression
4. Do the tests for normality

Assumption 4: Look for autocorrelation - correlation among the values of the variables for different cases. An easy example to think about is temperature - in a hot summer, it's likely that successive months will be hot, so that the errors in successive months may well be correlated.

Assumption 6. See which cases are especially influential - which ones are seriously influencing our estimates. Examine

issues of multicollinearity.

CHECKING ASSUMPTIONS 3,5,6:

Check relationships among variables:

```
. use nations.dta

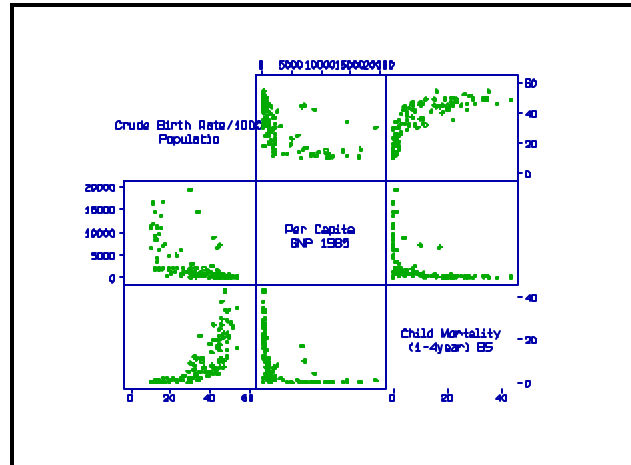
. corr birth gnpcap chldmort
(obs=109)

    Be concerned about the correlation between
    gnpcap and chldmort
    |   birth   gnpcap chldmort
-----+-----
    birth |   1.0000
    gnpcap | -0.6263   1.0000
    chldmort |  0.7773  -0.5047   1.0000
```

```
. graph birth gnpcap chldmort, matrix
label
```

Be concerned about curvilinearity of the relationships - we may not be specifying our model properly

Note that these checks don't help us look for heteroscedasticity



```
. regress birth gnpcap chldmort
```

Source	SS	df	MS
Model	13604.1783	2	6802.08913
Residual	6471.96852	106	61.0563068
Total	20076.1468	108	185.890248

Number of obs = 109
 F(2, 106) = 111.41
 Prob > F = 0.0000
 R-square = 0.6776
 Adj R-square = 0.6715
 Root MSE = 7.8139

birth	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
gnpcap	-.0009631	.000196	-4.914	0.000	-.0013517 - .0005745
chldmort	.7512455	.0775466	9.688	0.000	.5975018 .9049893
_cons	28.37934	1.42717	19.885	0.000	25.54983 31.20884

```
. predict birthhat

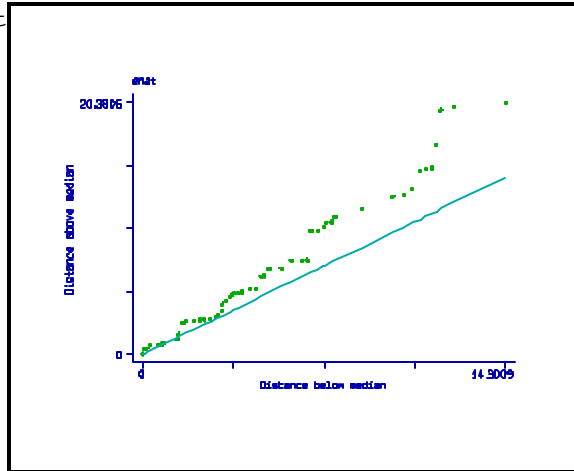
. predict ehat, resid
```

Residual analysis: Look for normality of errors, heteroscedasticity

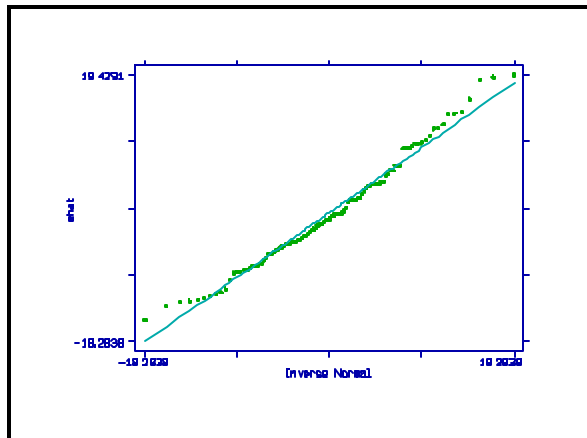
. sum ehat Errors ALWAYS will have mean =0

Variable	Obs	Mean	Std. Dev.	Min	Max
ehat	109	2.73e-09	7.741165	-15.25247	19.42905

. symplot ehat

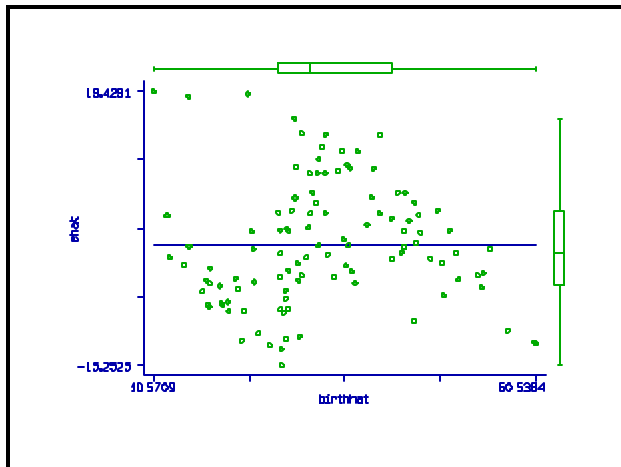
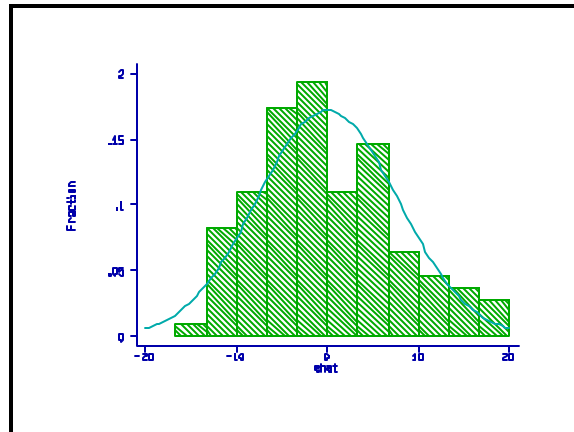


There seems to be some assymetry - with larger positive errors



. qnorm ehat the plot looks pretty normal

```
. graph ehat, bin(12) norm xlabel ylabel
```

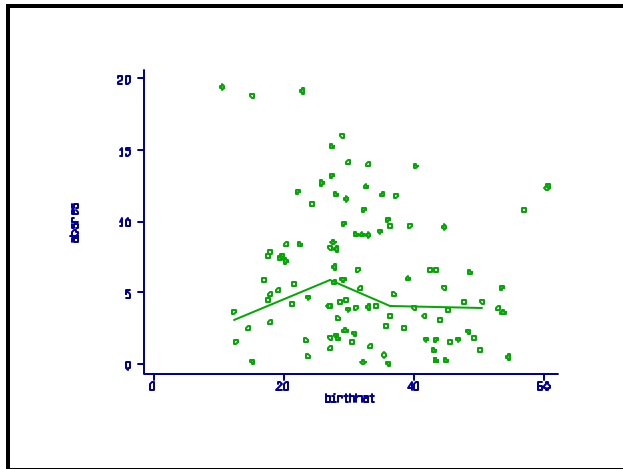


```
. graph ehat birthhat, yline(0) twoway box
```

BUT there is clear evidence of heteroscedasticity - a patterning in the residuals

Band plots: Band plots are another way of looking for heteroscedasticity. They divide the horizontal axis into a number of bands - here 4 bands. Then, within a band, they find the median of the values on the horizontal axis (i.e., the predicted Y's), and the median of the values on the vertical axis (here the absolute values of the residuals). They form these four points, (median of \hat{Y} , median absres) for each band, and connect them. If there were homoscedasticity, these points would lie approximately on a straight line - no difference in the absres no matter where you were on the predicted Y's.

```
. gen absres = abs(ehat)
. graph absres birthhat, connect(m) bands(4) xlabel ylabel
```



Again there is evidence of heteroscedasticity -- in this case because the median absolute residual varies among the bands.

CHECKING ASSUMPTION 6: INFLUENCE STATISTICS

Another set of questions we'd like to ask concerns whether there is one observation that is particularly influential in determining our results, or a few. There are several diagnostic techniques that can be used.

DFbetas: For a particular variable, say *gnpcap*, we can ask, how much would the coefficient of *gnpcap* change if we were to drop the first observation, or if we were to put observation 1 back in the data set and drop observation 2, and so on. More generally, $DFbeta_k$ measures how much the coefficient of X_k would change if we dropped ONE observation. It is *standardized* in the sense that the difference between the coefficient estimated from all cases and the coefficient estimated omitting case *i* is divided by the estimated standard error of the coefficient:

The equation for the *change in \hat{a}_k* if observation *i* is dropped is:

$$DFbeta_{ik} = \frac{b_k - b_{k(i)}}{s_{e(i)} / \sqrt{RSS_k}}$$

STATA will generate DFbetas after the `regress` command is used.

```
. quietly regress birth gnpcap chldmort          quietly: no output is printed
. dfbeta gnpcap chldmort
DFgnpcap:  DFbeta(gnpcap)
DFchldmo:  DFbeta(chldmort)
. summ DFgnpcap DFchldmo
```

Variable	Obs	Mean	Std. Dev.	Min	Max
DFgnpcap	109	.0016865	.1432527	-.1663877	1.072631
DFchldmo	109	-.0004444	.1097267	-.5188906	.3429605

Here we find that when any one observation is omitted, the coefficient of *gnpcap* is changed by anywhere between $-.167sd$ to $1.07sd$ -- not a significant change. Similarly, the changes in the coefficient of *chldmo* vary between $-.52sd$ and $.34sd$ -- even less than for the first variable.

If there were observations for which the value of Dfbeta were large, say 5 or more, we would certainly want to look at those observations -- to see if, for example, they were highly unusual and might have been coded wrong, or they were cases that should be looked at closely to understand why they had such a large effect on the coefficient of a particular variable.

REMEMBER: Dfbeta's are measures of the effect of each observation on a particular coefficient. An observation may have a large effect on one coefficient, but not on any others.

Leverage: Another concept is the *leverage* of a particular case - how much the *combination* of X values in a particular case may influence results. It looks basically at how unusual, in the sense of being far from the average, a particular set of X's is. It uses the *hat statistics*.

You may recall that the matrix equation for the predicted values of Y, \hat{Y} , is given by

$$\hat{Y} = X \hat{\beta}$$

and

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Combining these two equations,

$$\hat{Y} = X(X'X)^{-1}X'Y$$

The *hat matrix*, H, is the matrix that multiplies Y to produce the expected values of Y. It is so-named because it *puts the hat* on Y:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$$

The *hat statistic for case i* is the value on the diagonal of this hat matrix, which we call h_i . It is what multiplies Y_i itself. If that element is unusually large, it says that case i is being given a very large weight in determining the estimated Y_i . This is what is meant by *leverage*.

The leverage of a case enters into estimating the variance of the residuals for that case. Specifically,

$$\text{Var}[e_i] = s_e^2(1-h_i)$$

so that, the greater the leverage of case i, the smaller the estimated variance of the estimated error for that case -- we do better in estimating the residuals for cases that have high leverage, i.e. these cases tend to put our regression line close to themselves.

We can find the values of h_i in STATA by, again after the `regress` command, issuing the command:

```
. predict h, hat
```

We try **not** to have cases with high leverage. One rule of thumb used when looking at leverage:

$\max(h_i)$	$\#0.2$	safe
$0.2 < \max(h_i)$	$\#0.5$	risky
$\max(h_i)$	>0.5	avoid if possible

Standardized residual: The *standardized residual* estimates how far our observed value is from predicted, *adjusting* for the leverage of that particular point. That is, if the point is highly influential, it has pulled the predicted values toward itself. The z residual adjusts for this: the larger h_i , the smaller the denominator of

$$z_i' = \frac{e_i}{s_e \sqrt{1-h_i}}$$

Cook's distance: The statistic referred to as *Cook's distance* or *Cook's D* measures the influence of case i on the model as a whole - not just on a single coefficient. It is related both to the standardized error and the leverage. *Cook's D* adjusts not only for the leverage but also for the number of predictor variables. It is this statistic, which takes into account the estimated error, the leverage, and the number of predictors, that is used to judge how influential a particular observation is overall.

$$D_i' = \frac{z_i^2 h_i}{K(1-h_i)}$$

Two rules of thumb:	D_i is large if it is greater than 1	absolute cutoff
	D_i is larger if it is greater than $4/n$	size-adjusted cutoff

Checking assumption 6: Multicollinearity

Multicollinearity occurs when the X values, the predictors, are themselves intercorrelated. One measure of that intercorrelation is the multiple correlation coefficient that results from the regression of each X variable on all others in the model, e.g. if there are 8 predictors, we can estimate the model:

$$X_3 = \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2 + \hat{a}_4 X_4 \dots + \hat{a}_8 X_8 + \hat{a}$$

Then R_3^2 gives the proportion of variation in X_3 that is explained by its relation to the other predictor variables. It is sometimes referred to as the proportion of the variance of X_3 that is *shared* with the other predictors.

Tolerance: The *tolerance* of X_3 is the proportion that is *not* explained by the other variables, i.e.

$$\text{tolerance of } X_3 = 1 - R_3^2$$

It is the proportion of the variance of X_3 that is not shared with the other variables in our analysis. If the tolerances

are low (say .1 or .2) there are multicollinearity problems.

A guide to dealing with multicollinearity problems:

1. look at the correlation matrix
2. experiment with adding and dropping the suspect variables - do se's and coefficients change much?
3. Cope:
 - a. keep variables in equation, but understand interpretation
 - b. drop one or more of the variables, but understand interpretation.
 - c. combine variables (develop index through principal component analysis or other means)
 - d. ridge regression
 - e. collect more data

AN EXAMPLE

Using the dataset **nations.dta**, I regressed birth rates on population characteristics.

LOG OF BIRTH RATE ANALYSIS: This is a copy of my log. You can replicate the analysis on your own computers.

```
. * I am interested in investigating the relationship between the birth rate and other
measures in society. One hypothesis - if a country has low mortality, it will have low
fertility because more children survive to be adults. Therefore want to look at the
relationship to the death rate - or to infant or child mortality.
```

```
. * second hypothesis: wealthier countries will have lower birth rates because of the
availability of family planning services and because they want to have more educated
children -- therefore want to include gnpcap and a measure of education
```

```
. * third hypothesis: there is greater access to health care in urban areas of most
countries. Therefore want to look at %urban
```

```
. * countries with better food supply are also better off - in ways that may more
directly relate to survival and fertility than gnpcap
```

```
. * will want to worry about multicollinearity, unusual observations, transformations,
etc.
```

```
. * step 1: correlations
```

```
. corr birth gnpcap chldmort infmort death energy food urban school1 school2 school3
(obs=94)
```

	birth	gnpcap	chldmort	infmort	death	energy	food
birth	1.0000						
gnpcap	-0.6191	1.0000					
chldmort	0.7730	-0.5027	1.0000				
infmort	0.8772	-0.6274	0.9454	1.0000			
death	0.6108	-0.3989	0.8874	0.8162	1.0000		
energy	-0.6713	0.9183	-0.5218	-0.6524	-0.3653	1.0000	
food	-0.7862	0.6854	-0.7017	-0.7834	-0.5567	0.7172	1.0000
urban	-0.6922	0.6868	-0.6535	-0.7299	-0.6082	0.6672	0.6909
school1	-0.5259	0.2390	-0.7214	-0.6625	-0.7359	0.2638	0.4703
school2	-0.8406	0.7106	-0.7391	-0.8503	-0.6217	0.7588	0.8017

```

school3|  -0.6991   0.6365  -0.6188  -0.6941  -0.5041   0.7001   0.6512
      |      urban school1  school2  school3
-----+-----
urban|      1.0000
school1|  0.4813   1.0000
school2|  0.7832   0.5646   1.0000
school3|  0.6850   0.4183   0.8112   1.0000
    
```

. * good candidates: food urban school2 death

```

. graph food, bin(11) norm xlabel
. qnorm food
. symplot food
. ladder food          none helped
    
```

```

. graph urban, bin(11) norm xlabel
. qnorm urban
. symplot urban
. ladder urban          none helped
    
```

```

. graph school2, bin(11) norm xlabel
. qnorm school2
. symplot school2
. ladder school2
    
```

. ladder death

Transformation	formula	Chi-sq(2)	P(Chi-sq)
cube	death^3	40.39	0.000
square	death^2	21.73	0.000
raw	death	7.07	0.029
square-root	sqrt(death)	4.28	0.118
log	log(death)	1.43	0.490
reciprocal root	1/sqrt(death)	11.82	0.003
reciprocal	1/death	33.39	0.000
reciprocal square	1/(death^2)	.	0.000
reciprocal cube	1/(death^3)	.	0.000

```

. gen deathlog = log(death)
. graph death1, bin(11) norm xlabel
. qnorm death1
. symplot death1
    
```

```

. * graph relationship with birth
. graph birth food urban school2 death1, matrix
    
```

. * calculate tolerances

```

. regress food urban school2 death1          R-squared = 0.6599
. regress urban school2 death1 food          R-squared = 0.6886
. regress school2 death1 food urban          R-squared = 0.7750
. regress death1 food urban school2          R-squared = 0.4458
    
```

FIRST MODEL

```
. regress birth deathl food urban school2
```

Source	SS	df	MS			
Model	14123.3368	4	3530.83421	Number of obs =	102	
Residual	4326.0357	97	44.5983062	F(4, 97) =	79.17	
Total	18449.3725	101	182.667055	Prob > F =	0.0000	
				R-squared =	0.7655	
				Adj R-squared =	0.7558	
				Root MSE =	6.6782	

birth	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
deathlog	2.331468	2.032175	1.147	0.254	-1.701836	6.364772
food	-.0069966	.0020166	-3.470	0.001	-.0109989	-.0029943
urban	-.0032375	.0464446	-0.070	0.945	-.0954172	.0889422
school2	-.2591044	.047461	-5.459	0.000	-.3533014	-.1649074
_cons	57.65646	7.211043	7.996	0.000	43.34453	71.96838

```
. test deathlog urban
```

```
( 1) deathlog = 0.0
( 2) urban = 0.0
      F( 2, 97) = 0.74
      Prob > F = 0.4797
```

The problem with accepting these results is that schooling is a result of being in urban areas and of a lower death rate. I prefer to start over.

MODEL 1

```
. regress birth deathl food urban
```

Source	SS	df	MS			
Model	12875.6085	3	4291.8695	Number of obs =	107	
Residual	6615.12046	103	64.2244705	F(3, 103) =	66.83	
Total	19490.729	106	183.874802	Prob > F =	0.0000	
				R-squared =	0.6606	
				Adj R-squared =	0.6507	
				Root MSE =	8.014	

birth	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
deathlog	5.047348	2.371793	2.128	0.036	.3434563	9.75124
food	-.011916	.0019446	-6.128	0.000	-.0157727	-.0080593
urban	-.1351902	.0483365	-2.797	0.006	-.2310543	-.0393262
_cons	58.7373	8.437338	6.962	0.000	42.00383	75.47077

```
. * if we don't include schooling, deathlog is significant
```

```
. * Model 1: birth on deathl food urban
. predict bhatml
(2 missing values generated)
```

```

. predict resm1, resid
(2 missing values generated)

. graph resm1 bhatm1
. graph resm1 bhatm1, yline(0)

. * now look for unusual values
. * dfbetas
. quietly regress birth deathl food urban

. dfbeta deathl food urban
DFdeathl:  Dfbeta(deathlog)
DFfood:    Dfbeta(food)
DFurban:   Dfbeta(urban)

. sum DF*

```

Variable	Obs	Mean	Std. Dev.	Min	Max
DFdeathl	107	-.0005152	.1189294	-.6556519	.4316371
DFfood	107	.0001972	.1122533	-.3251025	.6838872
DFurban	107	.0000239	.1060673	-.263509	.4617972

```

. * nothing unusual

. predict hmod1, h
(2 missing values generated)

. sum h

```

Variable	Obs	Mean	Std. Dev.	Min	Max
hmod1	107	.0373832	.0191376	.0101783	.0958491

```

. * again, nothing unusual

. predict cooksml, cooks
(2 missing values generated)

. sum cook

```

Variable	Obs	Mean	Std. Dev.	Min	Max
cooksml	107	.0111989	.0250435	1.62e-06	.1743918

Here, using the rule that an unusual value is any value greater than $1/\#obs$, then we need to look at countries where the value of cooks is greater than about .04.

```
. list country if cooks>.04
      country
101.   Syria
102. Portugal
103. SriLanka
104.   China
105. UnArEmir
106.   Libya
107.   Kuwait
108.   Malawi
109.   Oman
```

```
. gen cookres =0
```

```
. replace cookres=1 if cooks>.05
(7 real changes made)
```

```
. graph resm1 bhatm1, symbol([cookre])
```

MODEL 2

```
. * mod2 drops the influential observations
```

```
. regress birth death1 food urban if cooks<=.04
```

Source	SS	df	MS	Number of obs =	100
Model	14464.9814	3	4821.66047	F(3, 96) =	112.42
Residual	4117.45859	96	42.8901937	Prob > F =	0.0000
				R-squared =	0.7784
				Adj R-squared =	0.7715
Total	18582.44	99	187.701414	Root MSE =	6.5491

	birth	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
deathlog	5.416917	2.173879	2.492	0.014	1.101801	9.732032
food	-.0129166	.0016776	-7.700	0.000	-.0162466	-.0095866
urban	-.1582633	.0429293	-3.687	0.000	-.2434772	-.0730494
_cons	61.18224	7.494107	8.164	0.000	46.30656	76.05793

```
. predict bhatm2
(2 missing values generated)
```

```
. predict resm2,resid
(2 missing values generated)
```

```
. graph resm2 bhatm2, yline(0) ylabel
```

```
. list country resm2 if resm2>15
```

```
      country      resm2
93.   Jordan    16.26232
98.  SauArabi    21.35169
```

101.	Syria	20.22829
105.	UnArEmir	20.6338
106.	Libya	27.49541
107.	Kuwait	21.95919

I might want to investigate further - because of my hypothesis that countries that are both Arab and Islamic have retained very high fertility rates even when their socioeconomic status has improved.