# ECON2228 Notes 7

Christopher F Baum

Boston College Economics

2014–2015

# Chapter 8: Heteroskedasticity

In laying out the standard regression model, we made the assumption of *homoskedasticity* of the regression error term: that its variance is assumed to be constant in the population, conditional on the explanatory variables.

The assumption of homoskedasticity fails when the variance changes in different segments of the population: for instance, if the variance of the unobserved factors influencing individuals' saving increases with their level of income. In such a case, we say that the error process is *heteroskedastic*.

This does not affect the optimality of ordinary least squares for the computation of point estimates–and the assumption of homoskedasticity did not underly our derivation of the OLS formulas. But if this assumption is not tenable, we may not be able to rely on the interval estimates of the parameters: on their confidence intervals, and $t-$statistics derived from their estimated standard errors.

The Gauss–Markov theorem, proving the optimality of least squares among linear unbiased estimators of the regression equation, does not hold in the presence of heteroskedasticity. If the error variance is not constant, then OLS estimators are no longer BLUE.

How, then, should we proceed? The classical approach is to test for heteroskedasticity, and if it is evident, try to model it. We can derive modified least squares estimators (known as *weighted least squares*) which will regain some of the desirable properties enjoyed by OLS in a homoskedastic setting.

But this approach is sometimes problematic, since there are many plausible ways in which the error variance may differ in segments of the population: depending on some of the explanatory variables in our model, or perhaps on some variables that are not even in the model. We can use weighted least squares effectively if we can derive the correct weights, but may not be much better off if we cannot convince ourselves that our application of weighted least squares is valid.

# Robust standard errors

Fortunately, developments in econometric theory have made it possible to avoid these quandaries. Methods have been developed to adjust the estimated standard errors in an OLS context for *heteroskedasticity of unknown form*: to develop what are known as *robust* standard errors. Most statistical packages now support the calculation of these robust standard errors when a regression is estimated.

If heteroskedasticity is a problem, the robust standard errors will differ from those calculated by OLS, and we should take the former as more appropriate. How can you compute these robust standard errors? In Stata, one merely adds the option `, robust` to the `regress` command.

The ANOVA F-table will be suppressed (as will the adjusted $R^2$ measure), since neither is valid when robust standard errors are being computed, and the term “robust” will be displayed above the standard errors of the coefficients to remind you that robust errors are in use.

How are robust standard errors calculated? Consider a model with a single explanatory variable. The OLS estimator can be written as:

$$b_1 = \beta_1 + \frac{\sum (x_i - \bar{x})\, u_i}{\sum (x_i - \bar{x})^2}$$

This gives rise to an estimated variance of the slope parameter:

$$Var\left(b_1\right) = \frac{\sum\left(x_i - \bar{x}\right)^2 \sigma_i^2}{\left(\sum\left(x_i - \bar{x}\right)^2\right)^2} \tag{1}$$

This expression reduces to the standard expression from Chapter 2 if $\sigma_i^2 = \sigma^2$ for all observations:

$$Var\left(b_1\right) = \frac{\sigma^2}{\sum\left(x_i - \bar{x}\right)^2}$$

But if $\sigma_i^2 \neq \sigma^2$ this simplification cannot be performed on (1). How can we proceed? Halbert White showed (in a famous article in *Econometrica*, 1980) that the unknown error variance of the $i^{th}$ observation, $\sigma_i^2$, can be consistently estimated by $e_i^2 -$that is, by the square of the OLS residual from the original equation.

This enables us to compute robust variances of the parameters. For instance, (1) can now be computed from the OLS residuals, and its square root will be the robust standard error of $b_1$.

This carries over to multiple regression; in the general case of $k$ explanatory variables,

$$Var\left(b_j\right) = \frac{\sum r_{ij}^2 e_i^2}{\left(\sum \left(x_{ij} - \bar{x}_j\right)^2\right)^2} \tag{2}$$

where $e_i^2$ is the square of the $i^{th}$ OLS residual, and $r_{ij}$ is the $i^{th}$ residual from regressing variable $j$ on all other explanatory variables.

The square root of this quantity is the *heteroskedasticity-robust standard error*, or the "White" standard error, of the $j^{th}$ estimated coefficient. It may be used to compute the *heteroskedasticity-robust* $t-$*statistic*, which then will be valid for tests of the coefficient even in the presence of heteroskedasticity of unknown form.

Likewise, $F$-statistics, which would also be biased in the presence of heteroskedasticity, may be consistently computed from the regression in which the robust standard errors of the coefficients are available.

If we have this better mousetrap, why would we want to report OLS standard errors–which would be subject to bias, and thus unreliable, if there is a problem of heteroskedasticity? If (and only if) the assumption of homoskedasticity is valid, the OLS standard errors are preferred, since they will have an exact $t-$distribution at any sample size.

The application of robust standard errors is justified as the sample size becomes large. If we are working with a sample of modest size, and the assumption of homoskedasticity is tenable, we should rely on OLS standard errors.

As robust standard errors are very easily calculated in most statistical packages, it is a simple task to estimate both sets of standard errors for a particular equation, and consider whether inference based on the OLS standard errors is fragile. In large data sets, it has become increasingly common practice to report the robust standard errors.

# Testing for heteroskedasticity

We may want to demonstrate that the model we have estimated does not suffer from heteroskedasticity, and justify reliance on OLS and OLS standard errors in this context. How might we evaluate whether homoskedasticity is a reasonable assumption? If we estimate the model via standard OLS, we may then base a test for heteroskedasticity on the OLS residuals.

If the assumption of homoskedasticity, conditional on the explanatory variables, holds, it may be written as:

$$H_0 : Var\left(u|x_1, x_2, ..., x_k\right) = \sigma^2$$

And a test of this null hypothesis can evaluate whether the variance of the error process appears to be independent of the explanatory variables. We cannot observe the variances of each observation, of course, but as above we can rely on the squared OLS residual, $e_i^2$, to be a consistent estimator of $\sigma_i^2$.

One of the most common tests for heteroskedasticity is derived from this line of reasoning: the *Breusch–Pagan* test. The BP test involves regressing the squares of the OLS residuals on a set of variables, such as the original explanatory variables, in an auxiliary regression:

$$e_i^2 = d_0 + d_1 x_1 + d_2 x_2 + ... d_k x_k + v \tag{3}$$

If the magnitude of the squared residual, which is a consistent estimator of the error variance of that observation, is not related to any of the explanatory variables, then this regression will have no explanatory power: its $R^2$ will be small, and its ANOVA $F-$statistic will indicate that it does not explain any meaningful fraction of the variation of $e_i^2$ around its own mean. Note that although the OLS residuals have mean zero, and are in fact uncorrelated by construction with each of the explanatory variables, that does not apply to their squares.

The Breusch–Pagan test can be conducted by either the ANOVA $F-$statistic from (3), or by a large-sample form known as the Lagrange multiplier statistic: $LM = n \times R^2$ from the auxiliary regression. Under $H_0$ of homoskedasticity, $LM \sim \chi^2_k$.

The Breusch–Pagan test can be computed with the `estat hettest` command after `regress`.

```
regress price mpg weight length
estat hettest
```

which would evaluate the residuals from the regression for heteroskedasticity, with respect to a linear combination of the original explanatory variables: the $\hat{y}$ values from the regression. The null hypothesis is that of homoskedasticity; if a small $p-$value is received, the null is rejected in favor of heteroskedasticity. That is, the auxiliary regression (which is not shown) had a meaningful amount of explanatory power.

The test displays the *LM* statistic and its $p-$value versus the $\chi_k^2$ distribution. If a rejection is received, one should rely on robust standard errors for the original regression. Although we have demonstrated the Breusch–Pagan test by employing a combination of the original explanatory variables, the test may be used with any set of variables: including those not in the regression, but suspected of being systematically related to the error variance, such as the size of a firm, or the wealth of an individual.

```
. eststo, ti("iid"):reg price mpg weight length
      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  3,    70) =   12.98
       Model |  226957412     3  75652470.6           Prob > F      =  0.0000
    Residual |  408107984    70  5830114.06           R-squared     =  0.3574
-------------+------------------------------           Adj R-squared =  0.3298
       Total |  635065396    73  8699525.97           Root MSE      =  2414.6

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -86.78928   83.94335    -1.03   0.305     -254.209    80.63046
      weight |   4.364798   1.167455     3.74   0.000     2.036383    6.693213
      length |  -104.8682   39.72154    -2.64   0.010    -184.0903   -25.64607
       _cons |   14542.43   5890.632     2.47   0.016      2793.94    26290.93
------------------------------------------------------------------------------
(est1 stored)
. estat hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of price
         chi2(1)      =    16.21
         Prob > chi2  =   0.0001
. eststo, ti("robust"): qui reg price mpg weight length, robust
(est2 stored)
```

```
. esttab,star(* 0.1 ** 0.05 *** 0.01) mti nonum
```

| | iid | robust |
|---|---|---|
| mpg | -86.79 | -86.79 |
| | (-1.03) | (-0.95) |
| weight | 4.365*** | 4.365** |
| | (3.74) | (2.36) |
| length | -104.9** | -104.9* |
| | (-2.64) | (-1.86) |
| _cons | 14542.4** | 14542.4** |
| | (2.47) | (2.18) |
| N | 74 | 74 |

t statistics in parentheses
* p<0.1, ** p<0.05, *** p<0.01

The Breusch-Pagan test is a special case of *White's general test for heteroskedasticity*. The sort of heteroskedasticity that will damage OLS standard errors is that which involves correlations between squared errors and explanatory variables. White's test takes the list of explanatory variables $\{x_1, x_2, ..., x_k\}$ and augments it with squares and cross products of each of these variables.

The White test then runs an auxiliary regression of $e_i^2$ on the explanatory variables, their squares, and their cross products. Under the null hypothesis, none of these variables should have any explanatory power, if the error variances are not systematically varying.

The White test is another $LM$ test, of the $n \times R^2$ form, but involves a much larger number of regressors in the auxiliary regression. In the example above, rather than just including `mpg weight length`, we would also include $\texttt{mpg}^2$, $\texttt{weight}^2$, $\texttt{length}^2$, $\texttt{mpg}\times\texttt{weight}$, $\texttt{mpg}\times\texttt{length}$, and $\texttt{weight}\times\texttt{length}$: 9 regressors in all, giving rise to a test statistic with a $\chi^2_{(9)}$ distribution.

How can you perform White's test? Give the command `estat imtest, white` after your regression. The command will automatically generate these additional variables and perform the test after a `regress` command. Since Stata knows what explanatory variables were used in the regression, you need not specify them.

```
. qui reg price mpg weight length

. imtest, white

White´s test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =     39.59
         Prob > chi2  =    0.0000

Cameron & Trivedi´s decomposition of IM-test
```

| Source | chi2 | df | p |
|---|---|---|---|
| Heteroskedasticity | 39.59 | 9 | 0.0000 |
| Skewness | 16.16 | 3 | 0.0011 |
| Kurtosis | 0.13 | 1 | 0.7136 |
| Total | 55.89 | 13 | 0.0000 |

The null of homoskedasticity is overwhelmingly rejected, and *i.i.d.* standard errors should not be used.

# Weighted least squares estimation

As an alternative to using heteroskedasticity-robust standard errors, we could transform the regression equation if we had knowledge of the form taken by heteroskedasticity. For instance, if we had reason to believe that:

$$Var(u|x) = \sigma^2 h(x)$$

where $h(x)$ is some function of the explanatory variables that could be made explicit (e.g. $h(x) = income$), we could use that information to properly specify the correction for heteroskedasticity.

What would this entail? Since in this case we are saying that $Var(u|x) \propto income$, then the standard deviation of $u_i$, conditional on $income_i$, is $\sqrt{income_i}$. Thus could be used to perform *weighted least squares*: a technique in which we transform the variables in the regression, and then run OLS on the transformed equation.

If we were estimating a simple savings function from the dataset `saving.dta`, in which `sav` is regressed on `inc`, and believed that there might be heteroskedasticity of the form above, we would perform the following transformations:

```
gen sd=sqrt(inc)
gen wsav=sav/sd
gen kon=1/sd
gen winc=inc/sd
regress wsav winc kon, noc
```

# Original regression:

```
. eststo, ti("OLS"):regress sav inc

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    6.49
       Model |    66368437     1    66368437           Prob > F      =  0.0124
    Residual |  1.0019e+09    98  10223460.8           R-squared     =  0.0621
-------------+------------------------------           Adj R-squared =  0.0526
       Total |  1.0683e+09    99  10790581.8           Root MSE      =  3197.4

------------------------------------------------------------------------------
         sav |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .1466283   .0575488     2.55   0.012     .0324247     .260832
       _cons |   124.8424   655.3931     0.19   0.849    -1175.764    1425.449
------------------------------------------------------------------------------
(est1 stored)
```

```
. bcuse saving, clear nodesc
. gen sd=sqrt(inc)
. gen wsav=sav/sd
. gen kon=1/sd
. gen winc=inc/sd
```

## WLS regression:

```
. regress wsav winc kon, noc
      Source |       SS           df       MS            Number of obs =       100
-------------+----------------------------------         F(   2,     98) =     14.30
       Model |  25251.0121         2  12625.506          Prob > F      =    0.0000
    Residual |  86513.4811        98  882.790623         R-squared     =    0.2259
-------------+----------------------------------         Adj R-squared =    0.2101
       Total |  111764.493       100  1117.64493         Root MSE      =    29.712

------------------------------------------------------------------------------
        wsav |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        winc |   .1717555   .0568128     3.02   0.003     .0590124    .2844986
         kon |  -124.9528   480.8606    -0.26   0.796    -1079.205    829.2995
------------------------------------------------------------------------------
```

Note that there is no constant term in the weighted least squares (WLS) equation, and that the coefficient on `winc` still has the same connotation: that of the marginal propensity to save. In this case, though, we might be thankful that Stata (and most modern packages) have a method for estimating WLS models by merely specifying the form of the weights:

```
regress sav inc [aw=1/inc]
```

In this case, the “aw” indicates that we are using “analytical weights”, Stata’s term for this sort of weighting, and the analytical weight is specified to be the inverse of the observation variance (not its standard error).

If you run this regression, you will find that its coefficient estimates and their standard errors are identical to those of the transformed equation. with less hassle than the latter, in which the summary statistics (F-statistic, $R^2$, predicted values, residuals, etc.) pertain to the transformed dependent variable (`wsav`) rather than the original variable.

# WLS with analytical weights:

```
. eststo, ti("WLS"):regress sav inc [aw=1/inc]
(sum of wgt is   1.3877e-02)
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    9.14
       Model |  58142339.8     1  58142339.8           Prob > F      =  0.0032
    Residual |   623432468    98   6361555.8           R-squared     =  0.0853
-------------+------------------------------           Adj R-squared =  0.0760
       Total |   681574808    99  6884594.02           Root MSE      =  2522.2

------------------------------------------------------------------------------
         sav |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .1717555   .0568128     3.02   0.003     .0590124    .2844986
       _cons |  -124.9528   480.8606    -0.26   0.796    -1079.205    829.2994
------------------------------------------------------------------------------
(est2 stored)
```

```
. esttab,star(* 0.1 ** 0.05 *** 0.01) mti nonum
-------------------------------------------------
                      OLS                WLS
-------------------------------------------------
inc                 0.147**            0.172***
                   (2.55)             (3.02)

_cons               124.8             -125.0
                   (0.19)            (-0.26)
-------------------------------------------------
N                     100                100
-------------------------------------------------
t statistics in parentheses
* p<0.1, ** p<0.05, *** p<0.01
```

The use of this sort of WLS estimation is less popular than it was before the invention of “White” standard errors; in theory, the transformation to homoskedastic errors will yield more attractive properties than even the use of “White” standard errors, conditional on our proper specification of the form of the heteroskedasticity. But of course we are not sure about that, and imprecise treatment of the errors may not be as attractive as the less informed technique of using the robust estimates.

# One rationale for WLS

One case in which we do know the form of the heteroskedasticity is that of *grouped data*, in which the data we are using has been aggregated from microdata into groups of different sizes. For instance, a dataset with 50 states' average values of income, family size, etc. calculated from a random sample of the U.S. population will have widely varying precision in those average values. The mean values for a small state will be computed from relatively few observations, whereas the counterpart values for a large state will be more precisely estimated.

As we know that the standard error of the mean is $\sigma/\sqrt{n}$, we recognize how this effect will influence the precision of the estimates. How, then, can we use this dataset of 50 observations while dealing with the known heteroskedasticity of the states' errors? This too is weighted least squares, where the weight on the individual state should be its population.

This can be achieved in Stata by specifying "frequency weights", a variable containing the number of observations from which each sample observation represents. If we had state-level data on saving, income and population, we might `regress saving income [fw=pop]` to achieve this weighting.

# OLS:

```
. sysuse census, clear
(1980 Census data by state)
. g pcturban = 100 * popurban / pop
. eststo, ti("OLS"): reg medage pcturban
      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    2.33
       Model |  6.50713318     1  6.50713318           Prob > F      =  0.1334
    Residual |  134.012852    48  2.79193441           R-squared     =  0.0463
-------------+------------------------------           Adj R-squared =  0.0264
       Total |  140.519985    49   2.8677548           Root MSE      =  1.6709

------------------------------------------------------------------------------
      medage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pcturban |   .0252898   .0165655     1.53   0.133    -.0080173     .058597
       _cons |   27.84687   1.133939    24.56   0.000     25.56693     30.1268
------------------------------------------------------------------------------
(est1 stored)
```

# WLS with frequency weights:

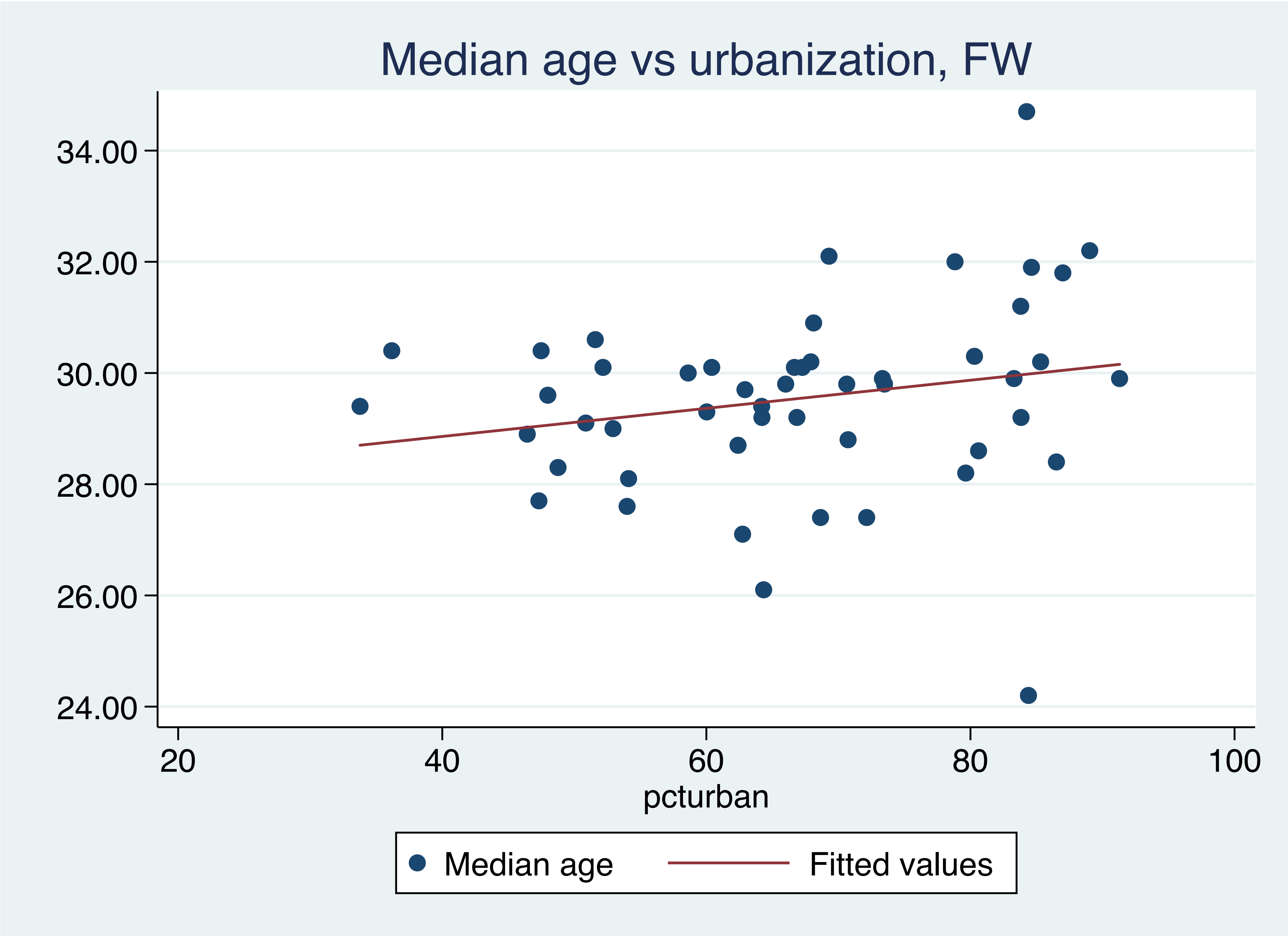
```
. eststo, ti("WLS FW"): reg medage pcturban [fw=pop]
      Source |       SS         df       MS              Number of obs =225907472
-------------+------------------------------             F(  1,225907470) =
>  .
       Model |  61570814.8       1  61570814.8           Prob > F      =  0.0000
    Residual |   555366235225907470  2.45837924             R-squared     =  0.09
> 98
-------------+------------------------------             Adj R-squared =  0.0998
       Total |   616937050225907471  2.73092805             Root MSE      =  1.56
> 79

------------------------------------------------------------------------------
      medage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pcturban |   .0405598   8.10e-06   5004.53   0.000     .0405439    .0405757
       _cons |   27.12268   .0006061   4.5e+04   0.000     27.12149    27.12386
------------------------------------------------------------------------------
(est2 stored)
. predict double wtmedage, xb
```

```
. esttab,star(* 0.1 ** 0.05 *** 0.01) mti nonum
------------------------------------------------
                         OLS           WLS FW
------------------------------------------------
pcturban              0.0253         0.0406***
                      (1.53)       (5004.53)

_cons                  27.85***        27.12***
                     (24.56)      (44752.12)
------------------------------------------------
N                         50       225907472
------------------------------------------------
t statistics in parentheses
* p<0.1, ** p<0.05, *** p<0.01
. tw (scatter medage pcturban, ylab(,angle(0))) ///
>  (lfit medage pcturban, ti("Median age vs urbanization, FW"))
```

When frequency weights are used, the effect of urbanization on median age in a state is precisely estimated. For each additional percent of urban population, the median age increases by 0.04 years, or about two weeks.


Median age vs urbanization, FW
34.00
32.00
30.00
28.00
26.00
24.00
20
40
60
80
100
pcturban
Median age
Fitted values

# A rationale for ratio transformation

One additional observation regarding heteroskedasticity. We often see, in empirical studies, that an equation has been specified in some ratio form—for instance, with per capita dependent and independent variables for data on states or countries, or in terms of financial ratios for firm- or industry-level data.

Although there may be no mention of heteroskedasticity in the study, it is very likely that these ratio forms have been chosen to limit the potential damage of heteroskedasticity in the estimated model. There can certainly be heteroskedasticity in a per-capita form regression on country-level data, but it is much less likely to be a problem than it would be if, say, the levels of GDP were used in that model.

Similarly, scaling firms' values by total assets, or total revenues, or the number of employees will tend to mitigate the difficulties caused by extremes in scale between large corporations and corner stores. Such models should still be examined for their errors' behavior, but the popularity of the ratio form in these instances is an implicit consideration of potential heteroskedasticity.