



# Lecture 8: Heteroskedasticity

- Causes
- Consequences
- Detection
- Fixes

# Assumption MLR5: Homoskedasticity

$$\text{var}(u \mid x_1, x_2, \dots, x_j) = \sigma^2$$

- In the multivariate case, this means that the variance of the error term does not increase or decrease with any of the explanatory variables  $x_1$  through  $x_j$ .
- If MLR5 is untrue, we have heteroskedasticity.

# Causes of Heteroskedasticity

- Error variance can increase as values of an independent variable increase.
  - Ex: Regress household security expenditures on household income and other characteristics. Variance in household security expenditures will increase as income increases because you can't spend a lot on security unless you have a large income.
- Error variance can increase with extreme values of an independent variable (either positive or negative)
- Measurement error. Extreme values may be wrong, leading to greater error at the extremes.

## Causes of Heteroskedasticity, cont.

- Bounded independent variable. If  $Y$  cannot be above or below certain values, extreme predictions have restricted variance. (See example in 5<sup>th</sup> slide after this one.)
- Subpopulation differences. If you need to run separate regressions, but run a single one, this can lead to two error distributions and heteroskedasticity.
- Model misspecification:
  - form of included variables (square, log, etc.)
  - exclusion of relevant variables

## [ Not Consequences of Heteroskedasticity: ]

- MLR5 is not needed to show unbiasedness or consistency of OLS estimates. So violation of MLR5 **does not** lead to biased estimates.
- Since  $R^2$  is based on overall sums of squares, it is unaffected by heteroskedasticity.
- Likewise, our estimate of root mean squared error is valid in the presence of heteroskedasticity.

# Consequences of heteroskedasticity

- OLS model is no longer B.L.U.E. (best linear unbiased estimator)
  - Other estimators are preferable
- With heteroskedasticity, we no longer have the “best” estimator, because error variance is biased.
  - incorrect standard errors
  - Invalid t-statistics and F statistics
  - LM test no longer valid



# [ Detection of heteroskedasticity: graphs ]

- Conceptually, we know that heteroskedasticity means that our predictions have uneven variance over some combination of  $X$ s.
  - Simple to check in bivariate case, complicated for multivariate models.
- One way to visually check for heteroskedasticity is to plot predicted values against residuals
  - This works for either bivariate or multivariate OLS.
- If heteroskedasticity is suspected to derive from a single variable, plot it against the residuals
- This is an ad hoc method for getting an intuitive feel for the form of heteroskedasticity in your model

# Let's see if the regression from the [2010 midterm has heteroskedasticity (DV is high school g.p.a.)

```
. reg hsgpa male hisp black other agedol dfreq1 schattach msgpa r_mk income1
    antipeer
```

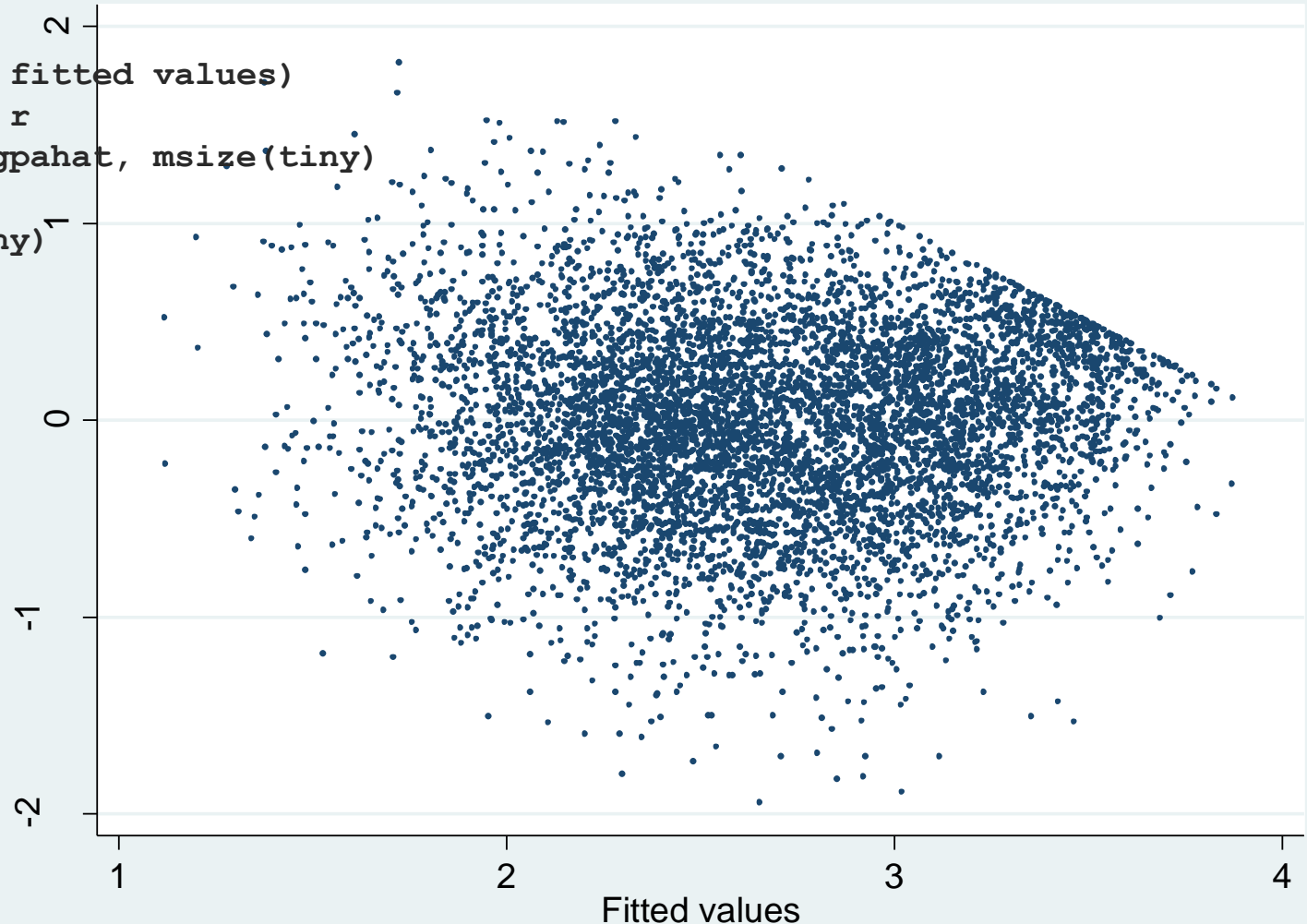
Source	SS	df	MS	Number of obs =	6574
Model	1564.98297	11	142.271179	F( 11, 6562) =	610.44
Residual	1529.3681	6562	.233064325	Prob > F =	0.0000
Total	3094.35107	6573	.470766936	R-squared =	0.5058
				Adj R-squared =	0.5049
				Root MSE =	.48277

hsgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
male	-.1574331	.0122943	-12.81	0.000	-.181534	-.1333322
hisp	-.0600072	.0174325	-3.44	0.001	-.0941806	-.0258337
black	-.1402889	.0152967	-9.17	0.000	-.1702753	-.1103024
other	-.0282229	.0186507	-1.51	0.130	-.0647844	.0083386
agedol	-.0105066	.0048056	-2.19	0.029	-.0199273	-.001086
dfreq1	-.0002774	.0004785	-0.58	0.562	-.0012153	.0006606
schattach	.0216439	.0032003	6.76	0.000	.0153702	.0279176
msgpa	.4091544	.0081747	50.05	0.000	.3931294	.4251795
r_mk	.131964	.0077274	17.08	0.000	.1168156	.1471123
income1	1.21e-06	1.60e-07	7.55	0.000	8.96e-07	1.52e-06
antipeer	-.0167256	.0041675	-4.01	0.000	-.0248953	-.0085559
_cons	1.648401	.0740153	22.27	0.000	1.503307	1.793495



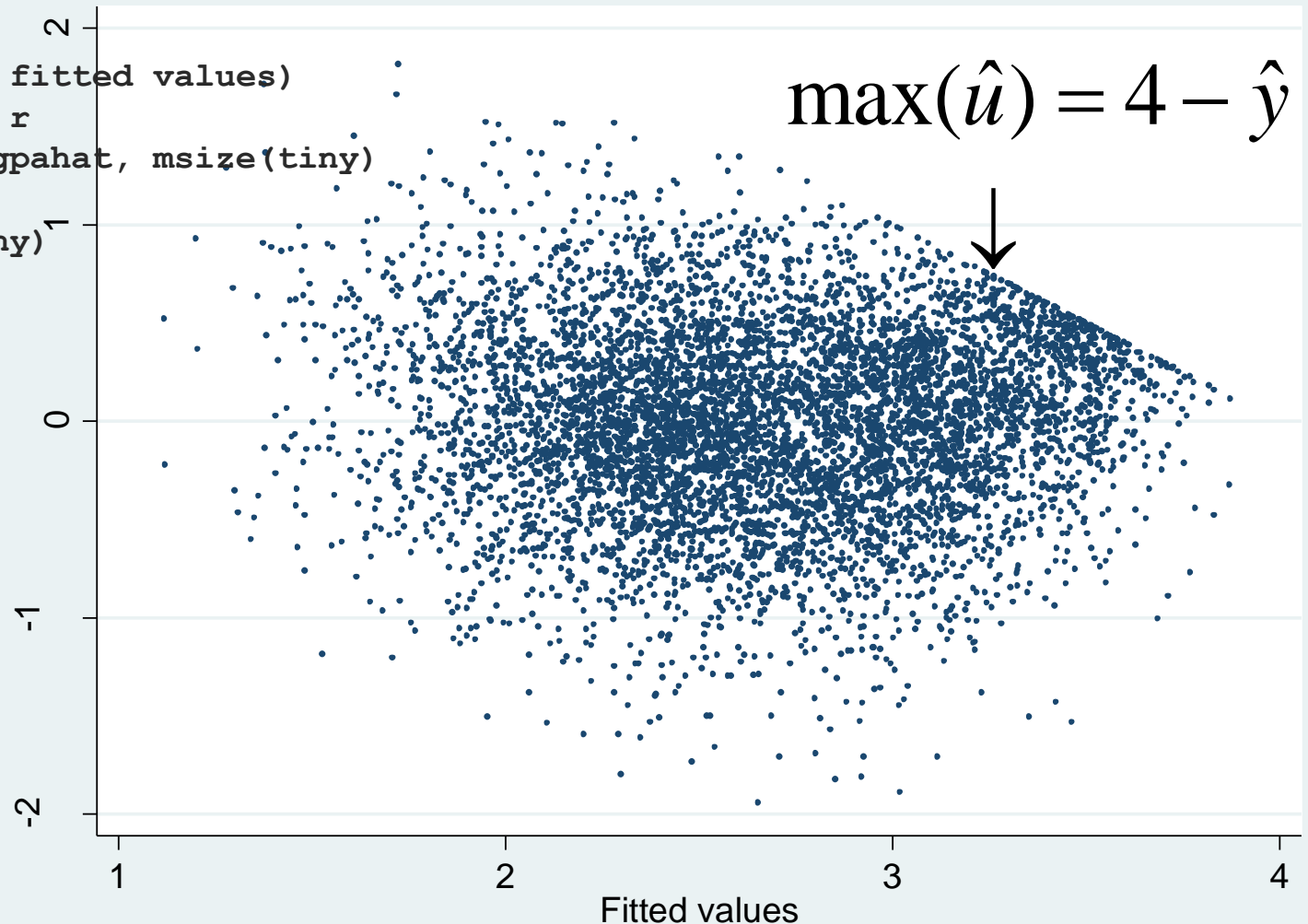
Let's see if the regression from the midterm has heteroskedasticity . . .

```
. predict gpahat  
(option xb assumed; fitted values)  
. predict residual, r  
. scatter residual gpahat, msize(tiny)  
or . . .  
. rvfplot, msize(tiny)
```



[ Let's see if the regression from the  
midterm has heteroskedasticity . . . ]

```
. predict gpahat  
(option xb assumed; fitted values)  
. predict residual, r  
. scatter residual gpahat, msize(tiny)  
or . . .  
. rvfplot, msize(tiny)
```

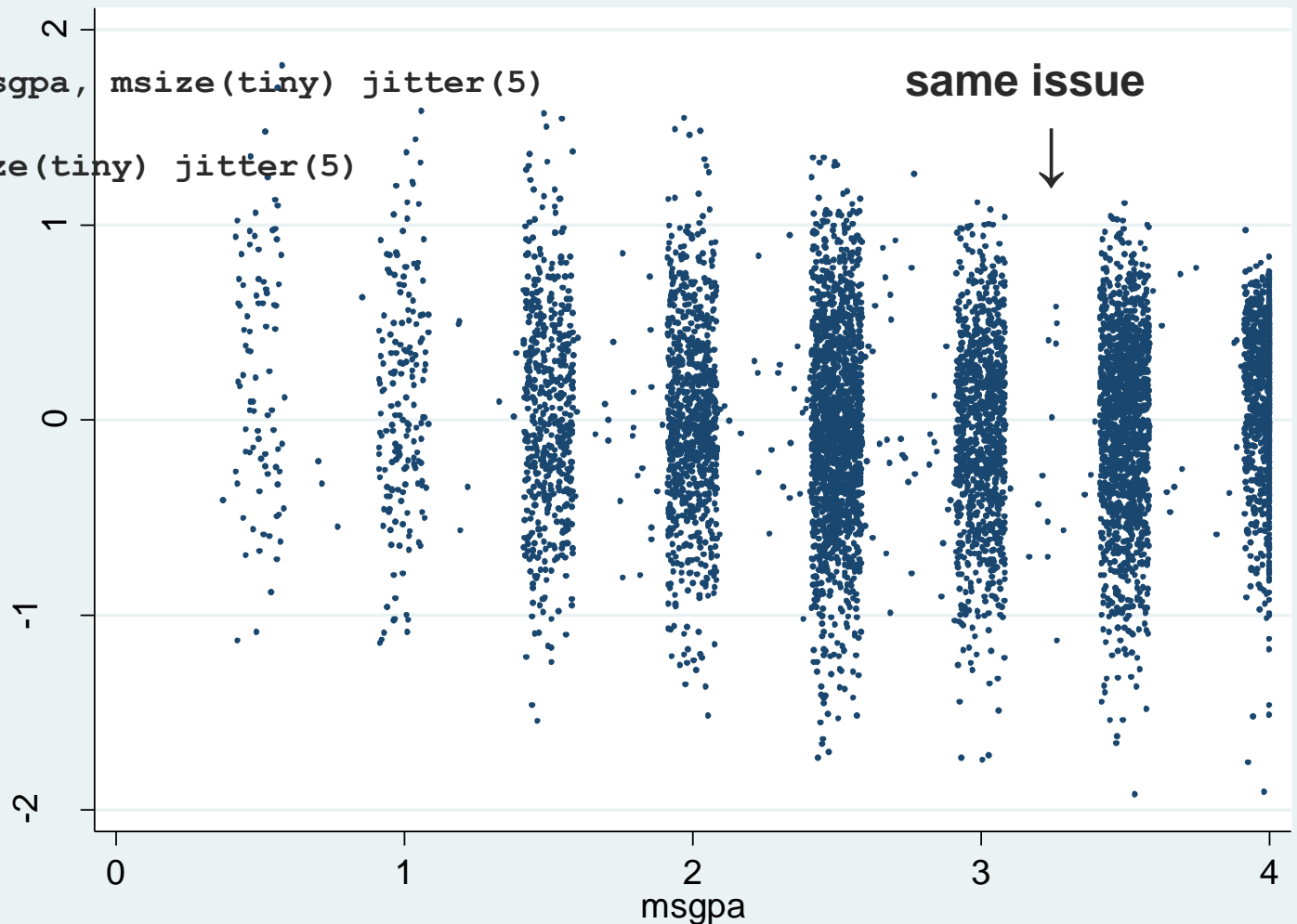


# [ Let's see if the regression from the 2010 midterm has heteroskedasticity ]

- This is not a rigorous test for heteroskedasticity, but it has revealed an important fact:
  - Since the upper limit of high school gpa is 4.0, the maximum residual, and error variance, is artificially limited for good students.
- With just this ad-hoc method, we strongly suspect heteroskedasticity in this model.
- We can also check the residuals against individual variables:

Let's see if the regression from the 2010 midterm has heteroskedasticity

```
. scatter residual msgpa, msize(tiny) jitter(5)  
or . . .  
. rvpplot msgpa, msize(tiny) jitter(5)
```



# Other useful plots for detecting heteroskedasticity

- **twoway (scatter resid fitted) (lowess resid fitted)**
  - Same as rvfplot, with an added smoothed line for residuals – should be around zero.
  - You have to create the “fitted” and “resid” variables
- **twoway (scatter resid var1) (lowess resid var1)**
  - Same as rvpplot var1, with smoothed line added.

# [ Formal tests for heteroskedasticity ]

- There are many tests for heteroskedasticity.
- Deriving them and knowing the strengths/weaknesses of each is beyond the scope of this course.
- In each case, the null hypothesis is homoskedasticity:

$$H_0 : E(u^2 \mid x_1, x_2, \dots, x_k) = E(u^2) = \sigma^2$$

- The alternative is heteroskedasticity.



# [ Formal test for heteroskedasticity: “Breusch-Pagan” test ]

- 1) Regress Y on Xs and generate squared residuals
- 2) Regress squared residuals on Xs (or a subset of Xs)
- 3) Calculate  $LM = n \cdot R_{\hat{u}^2}^2$ , ( $N \cdot R^2$ ) from regression in step 2.
- 4) LM is distributed chi-square with  $k$  degrees of freedom.
- 5) Reject homoskedasticity assumption if  $p$ -value is below chosen alpha level.



# [ Formal test for heteroskedasticity: “Breusch-Pagan” test, example

- After high school gpa regression (not shown):

```
. predict resid, r  
. gen resid2=resid*resid
```

```
. reg resid2 male hisp black other agedol dfreq1 schattach msgpa r_mk income1 antipeer
```

Source	SS	df	MS	Number of obs =	6574
Model	12.5590862	11	1.14173511	F( 11, 6562) =	9.31
Residual	804.880421	6562	.12265779	Prob > F =	0.0000
				R-squared =	0.0154
				Adj R-squared =	0.0137
Total	817.439507	6573	.124363229	Root MSE =	.35023

resid2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
male	-.0017499	.008919	-0.20	0.844	-.019234 .0157342
hisp	-.0086275	.0126465	-0.68	0.495	-.0334188 .0161637
black	-.0201997	.011097	-1.82	0.069	-.0419535 .0015541
other	.0011108	.0135302	0.08	0.935	-.0254129 .0276344
agedol	-.0063838	.0034863	-1.83	0.067	-.013218 .0004504
dfreq1	.000406	.0003471	1.17	0.242	-.0002745 .0010864
schattach	-.0018126	.0023217	-0.78	0.435	-.0063638 .0027387
msgpa	-.0294402	.0059304	-4.96	0.000	-.0410656 -.0178147
r_mk	-.0224189	.0056059	-4.00	0.000	-.0334083 -.0114295
income1	-1.60e-07	1.16e-07	-1.38	0.169	-3.88e-07 6.78e-08
antipeer	.0050848	.0030233	1.68	0.093	-.0008419 .0110116
_cons	.4204352	.0536947	7.83	0.000	.3151762 .5256943





# [ Formal test for heteroskedasticity: Breusch-Pagan test, example ]

```
. di "LM=", e(N) * e(r2)  
LM= 101.0025
```

```
. di chi2tail(11, 101.0025)  
1.130e-16
```

- We emphatically reject the null of homoskedasticity.
- We can also use the global F test reported in the regression output to reject the null ( $F(11, 6562) = 9.31$ ,  $p < .00005$ )
- In addition, this regression shows that middle school gpa and math scores are the strongest sources of heteroskedasticity. This is simply because these are the two strongest predictors and hsgpa is bounded.



# [ Formal test for heteroskedasticity: Breusch-Pagan test, example

- We can also just type “`ivhetttest, nr2`” after the initial regression to run the LM version of the Breusch-Pagan test identified by Wooldredge.

```
. ivhetttest, nr2
OLS heteroskedasticity test(s) using levels of IVs only
Ho: Disturbance is homoskedastic
White/Koenker nR2 test statistic      : 101.002  Chi-
sq(11) P-value = 0.0000
```

- Stata documentation calls this the “White/Koenker” heteroskedasticity test, based on Koenker, 1981.
- This adaptation of the Breusch-Pagan test is less vulnerable to violations of the normality assumption.

# Other versions of the Breusch-Pagan test

- **Note**, “`estat hetttest`” and “`estat hetttest, rhs`” also produce commonly-used Breusch-Pagan tests of the null of homoskedasticity, they’re older versions, and are biased if the residuals are not normally distributed.

# Other versions of the Breusch-Pagan test

- `estat hetttest, rhs`
  - From Breusch & Pagan (1979)
  - Square residuals and divide by mean so that new variable mean is 1
  - Regress this variable on Xs
  - Model sum of squares / 2  $\sim \chi_k^2$
- `estat hetttest`
  - Square residuals and divide by mean so that new variable mean is 1
  - Regress this variable on yhat
  - Model sum of squares / 2  $\sim \chi_1^2$

# [ Other versions of the Breusch-Pagan test ]

```
. estat hettest, rhs
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
```

```
Ho: Constant variance
```

```
Variables: male hisp black other agedol dfreq1 schattach msgpa r_mk income1  
antipeer
```

```
chi2(11)      =    116.03
```

```
Prob > chi2   =    0.0000
```

```
. estat hettest
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
```

```
Ho: Constant variance
```

```
Variables: fitted values of hsgpa
```

```
chi2(1)      =     93.56
```

```
Prob > chi2   =    0.0000
```

- In this case, because heteroskedasticity is easily detected, our conclusions from these alternate BP tests are the same, but this is not always the case.

# Other versions of the Breusch-Pagan test

- We can also use these commands to test whether homoskedasticity can be rejected with respect to a subset of the predictors:

```
. ivhettest hisp black other, nr2
OLS heteroskedasticity test(s) using user-supplied indicator variables
Ho: Disturbance is homoskedastic
    White/Koenker nR2 test statistic      :    2.838  Chi-sq(3) P-value =
    0.4173
```

```
. estat hettest hisp black other
```

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: hisp black other

```
chi2(3)          =    3.26
Prob > chi2       =    0.3532
```

## Tests for heteroskedasticity: White's test, complicated version

- 1) Regress  $Y$  on  $X$ s and generate residuals, square residuals
- 2) Regress squared residuals on  $X$ s, squared  $X$ s, and cross-products of  $X$ s (there will be  $p = k(k+3)/2$  parameters in this auxiliary regression, e.g. 11  $X$ s, 77 parameters!)
- 3) Reject homoskedasticity if test statistic (LM or  $F$  for all parameters but intercept) is statistically significant.
  - With small datasets, the number of parameters required for this test is too many.

## Tests for heteroskedasticity: White's test, simple version

- 1) Regress  $Y$  on  $X$ s and generate residuals, square residuals, fitted values, squared fitted values
- 2) Regress squared residuals on fitted values and squared fitted values:
$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + v$$
- 3) Reject homoskedasticity if test statistic (LM or F) is statistically significant.



# Tests for heteroskedasticity: White's test, example

```
. reg r2 gpahat gpahat2
```

Source	SS	df	MS	Number of obs =	6574
Model	10.4222828	2	5.2111414	<b>F( 2, 6571) =</b>	<b>42.43</b>
Residual	807.017224	6571	.122814979	Prob > F =	0.0000
Total	817.439507	6573	.124363229	R-squared =	0.0127
				Adj R-squared =	0.0124
				Root MSE =	.35045

r2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gpahat	.0454353	.0816119	0.56	0.578	-.1145505	.2054211
gpahat2	-.023728	.0152931	-1.55	0.121	-.0537075	.0062515
_cons	.2866681	.1067058	2.69	0.007	.0774901	.4958461

```
. di "LM=",e(r2)*e(N)
```

```
LM= 83.81793
```

```
. di chi2tail(2,83.81893)
```

```
6.294e-19
```

- Again, reject the null hypothesis.

## Tests for heteroskedasticity: White's test

- This test is not sensitive to normality violations
- The complicated version of the White test can be found using the “`whitetst`” command after running a regression.

```
. whitetst
```

```
White's general test statistic : 223.1636  
Chi-sq(72) P-value = 2.3e-17
```

- Note: the degrees of freedom is less than 77 because some auxiliary variables are redundant and dropped (e.g. the square of any dummy variable is itself).

## In-class exercise

- Work on questions 1 through 7 on the heteroskedasticity worksheet.

# [ Fixes for heteroskedasticity ]

- Heteroskedasticity messes up our variances (and standard errors) for parameter estimates
- Some methods tackle this problem by trying to model the exact form of heteroskedasticity: weighted least squares
  - Requires some model for heteroskedasticity.
  - Re-estimates coefficients and standard errors
- Other methods do not deal with the form of the heteroskedasticity, but try to estimate correct variances: robust inference, bootstrapping
  - Useful for heteroskedasticity of unknown form
  - Adjusts standard errors only

# Fixes for heteroskedasticity: heteroskedasticity-robust inference

$$\text{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2} = \frac{\sigma^2}{SST_x}, \text{ if } \sigma_i^2 = \sigma^2 \forall i \quad \leftarrow \text{the ideal}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2} \quad \leftarrow \text{robust variance estimator}$$

- The robust variance estimator is easy to calculate post-estimation. It reduces to the standard variance estimate under homoskedasticity.
- In Stata, obtaining this version of the variance is very easy: `reg y x, robust`

# Heteroskedasticity-robust inference, example

```
. quietly reg hsgpa male hisp black other agedol dfreq1 schattach msgpa r_mk income1  
    antipeer  
. estimates store ols  
. quietly reg hsgpa male hisp black other agedol dfreq1 schattach msgpa r_mk income1  
    antipeer, robust  
. estimates store robust  
. estimates table ols robust, stat(r2 rmse) title("High school GPA models") b(%7.3g)  
    se(%6.3g) t(%7.3g)
```

High school GPA models

-----			
Variable	ols	robust	
-----+-----			
male	-.157	-.157	← parameter estimates, unchanged
	.0123	.0124	← standard errors
	-12.8	-12.7	← T-statistics
hisp	-.06	-.06	
	.0174	.0173	
	-3.44	-3.46	
black	-.14	-.14	
	.0153	.0157	
	-9.17	-8.91	
other	-.0282	-.0282	
	.0187	.0186	
	-1.51	-1.52	
agedol	-.0105	-.0105	
	.0048	.0048	
	-2.19	-2.19	

# Heteroskedasticity-robust inference, example cont.

High school GPA models, cont.

Variable	ols	robust
dfreq1	-.00028 4.8e-04	-.00028 5.4e-04
	-.58	-.509
schattach	.0216 .0032	.0216 .0034
	6.76	6.4
msgpa	.409 .0082	.409 .0088
	50.1	46.3
r_mk	.132 .0077	.132 .0079
	17.1	16.6
income1	1.2e-06 1.6e-07	1.2e-06 1.5e-07
	7.55	7.87

High school GPA models, cont.

Variable	ols	robust
antipeer	-.0167 .0042	-.0167 .0043
	-4.01	-3.9
_cons	1.65 .074	1.65 .0752
	22.3	21.9
r2	.506	.506
rmse	.483	.483

legend: b/se/t

- Despite solid evidence for heteroskedasticity in this model, very little changes when heteroskedasticity-robust standard errors are calculated.
- Why did the estimates change so little?

# Heteroskedasticity-robust inference of Lagrange multiplier

- The book outlines a very involved set of steps to obtain a Lagrange Multiplier test that is robust to heteroskedasticity.
  - We'll go through these steps, testing whether hisp black and other are jointly significant
- 1) Obtain residuals from restricted model

```
. quietly reg hsgpa male agedol dfreq1 schattach msgpa r_mk income1 antipeer
. predict residuals
```
  - 2) Regress each excluded independent variable on the included independent variables, generate residuals

```
. quietly reg hisp male agedol dfreq1 schattach msgpa r_mk income1 antipeer
. predict rhisp, r
. quietly reg black male agedol dfreq1 schattach msgpa r_mk income1 antipeer
. predict rblack, r
. quietly reg other male agedol dfreq1 schattach msgpa r_mk income1 antipeer
. predict rother, r
```
  - 3) Generate products of residuals from restricted model and residuals from each auxiliary regression

```
. gen phisp=residuals*rhisp
. gen pblack=residuals*rblack
. gen pother=residuals*rother
```



# Heteroskedasticity-robust inference of Lagrange multiplier

- 4) Regress 1 on these three products without a constant,  $N\text{-SSR} \sim \chi^2$  with  $q$  degrees of freedom
- ```
. gen one=1
```
  - ```
. reg one phisp pblack pother, noc
```
  - ```
. di e(N)-e(rss)
```

79.289801
  - ```
. di chi2tail(3,79.289801)
```

4.359e-17
- Based on this test, we'd reject the null that hisp black and other are jointly equal to zero.
  - Another much easier option for heteroskedasticity-robust tests of joint restrictions is to run F-tests *after* a regression model with robust standard errors
- ```
. quietly reg hsgpa male hisp black other agedol dfreq1  
      schattach msgpa r_mk incomel antipeer, robust
```
- ```
. test hisp black other
```
- ```
( 1)  hisp = 0
```
- ```
( 2)  black = 0
```
- ```
( 3)  other = 0
```
- ```
      F(   3,   6562) =    27.01
```
- ```
          Prob > F =    0.0000
```

# [ Obtaining standard errors with bootstrapping ]

Bootstrapping (Wooldredge, pp. 223-4)

- In general, if the distribution of some statistic is unknown, bootstrapping can yield confidence intervals free of distributional assumptions.
- It resamples the dataset with replacement and re-estimates the statistic of interest many times (~1000 is good).
- Conceptually equivalent to drawing many random samples from the population.
- The standard deviation of the statistic of interest from the replications is the standard error of the statistic in the original model.
- This is incorporated into the regress function in Stata
  - `. reg y x, vce(bs, r(N))`
  - N is the number of replications

# [ Obtaining standard errors with bootstrapping ]

Bootstrapping (Wooldredge, pp. 223-4)

- If you are using bootstrapping for a paper, before the bootstrap, use the “`set seed N`” command where N is any particular number. Otherwise, you’ll get different results every time.
- You can also bootstrap other statistics with no obvious distribution, just in case you wanted a confidence interval for them
  - `. bs e(r2), r(1000): reg Y X`
  - `. bs e(rmse), r(1000): reg Y X`
  - `. bs r(p50), r(1000): summarize hsgpa, detail`

# [ Obtaining standard errors with bootstrapping ]

## Bootstrapping (Wooldredge, pp. 223-4)

- After bootstrapping, we can get more information using the command `estat bootstrap, all`
- For each statistic, this reports the following:
  - “bias” : the mean of the bootstrapped estimates minus the estimate from our original model.
  - Normal confidence interval, as reported before
  - Percentile confidence interval: limits defined by 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles of the bootstrapped estimates
  - Bias-corrected confidence interval: normal confidence interval minus bias

# Modeling heteroskedasticity, weighted least squares

- When heteroskedasticity is present, we know that the variance of our error term depends on some function of our  $X$ s

$$\text{Var}(u \mid x) = \sigma^2 h(x)$$

- Usually,  $h(x)$  is unknown, but if it were known, we could undo it by multiplying the regression equation by the inverse of square root  $h(x)$
- This strategy tries to re-weight each observation to “undo” heteroskedasticity.

# Modeling heteroskedasticity, weighted least squares

- Suppose, in the high school gpa regression, we believe that heteroskedasticity is a function of middle school gpa.
- In OLS we minimize the squared error, in WLS we minimize the weighted squared error
- We try to choose the weight such that variance is constant
- So, if middle school gpa is causing heteroskedasticity in our regression model, we can adjust it as follows:

# Modeling heteroskedasticity, weighted least squares

- Transform each variable by dividing by the square root of middle school gpa
- Also, create a new variable that is 1 divided by the square root of middle school gpa
- Run a new regression with all the transformed variables, and the new one, without a constant term.

# Modeling heteroskedasticity, weighted least squares

```
. gen con_ms=1/sqrt(msgpa)
. gen hsgpa_ms=hsgpa/sqrt(msgpa)
. gen male_ms=male/sqrt(msgpa)
. . . . etc
. reg hsgpa_ms con_ms male_ms hisp_ms black_ms other_ms agedol_ms dfreq1_ms schattach_ms msgpa_ms r_mk_ms i
> ncomel_ms antipeer_ms, noc
```

| Source   | SS         | df   | MS         | Number of obs = | 6574     |
|----------|------------|------|------------|-----------------|----------|
| Model    | 17706.3813 | 12   | 1475.53178 | F( 12, 6562) =  | 13952.58 |
| Residual | 693.95355  | 6562 | .10575336  | Prob > F =      | 0.0000   |
|          |            |      |            | R-squared =     | 0.9623   |
|          |            |      |            | Adj R-squared = | 0.9622   |
| Total    | 18400.3349 | 6574 | 2.79895572 | Root MSE =      | .3252    |

| hsgpa_ms     | Coef.     | Std. Err. | t      | P> t  | [95% Conf. Interval] |           |
|--------------|-----------|-----------|--------|-------|----------------------|-----------|
| con_ms       | 1.751627  | .0751105  | 23.32  | 0.000 | 1.604386             | 1.898868  |
| male_ms      | -.1602267 | .0129001  | -12.42 | 0.000 | -.1855151            | -.1349384 |
| hisp_ms      | -.0377276 | .0182012  | -2.07  | 0.038 | -.0734079            | -.0020472 |
| black_ms     | -.1319019 | .0157097  | -8.40  | 0.000 | -.1626981            | -.1011057 |
| other_ms     | -.0305844 | .0195973  | -1.56  | 0.119 | -.0690015            | .0078327  |
| agedol_ms    | -.0121919 | .0050095  | -2.43  | 0.015 | -.0220121            | -.0023717 |
| dfreq1_ms    | -2.45e-07 | .0004347  | -0.00  | 1.000 | -.0008525            | .000852   |
| schattach_ms | .022701   | .0032899  | 6.90   | 0.000 | .0162516             | .0291503  |
| msgpa_ms     | .377467   | .0075196  | 50.20  | 0.000 | .362726              | .3922079  |
| r_mk_ms      | .1167528  | .0079359  | 14.71  | 0.000 | .1011959             | .1323097  |
| incomel_ms   | 1.14e-06  | 1.75e-07  | 6.50   | 0.000 | 7.96e-07             | 1.48e-06  |
| antipeer_ms  | -.0195269 | .0042784  | -4.56  | 0.000 | -.027914             | -.0111397 |



# Modeling heteroskedasticity, weighted least squares

- Equivalently (and with much less room for mistakes):

```
. gen weight=1/msgpa  
. reg hsgpa male hisp black other agedol dfreq1  
    schattach msgpa r_mk incomel antipeer  
    [aweight=weight]
```

- The chances that we actually correctly modeled the form of heteroskedasticity are pretty low, but there's no reason we can't estimate weighted least squares with standard errors robust to unknown forms of heteroskedasticity

```
. reg hsgpa male hisp black other agedol dfreq1  
    schattach msgpa r_mk incomel antipeer  
    [aweight=weight], robust
```

# Modeling heteroskedasticity, feasible general least squares (FGLS)

- In practice, exactly modeling  $h(x)$  is infeasible.
- FGLS is a feasible alternative to exactly modeling  $h(x)$
- It assumes that  $h(x)$  is always positive, and of some unknown function of  $X$ s

$$h(X) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \dots + \delta_k x_k)$$

- Resulting estimates are biased but efficient, and have correct t- and F-statistics.



# [ Modeling heteroskedasticity, feasible general least squares (FGLS)

- 1) Regress  $y$  on  $X$ s, obtain residuals.
- 2) Create  $\log(\hat{u}^2)$  by logging squared residuals.
- 3) Regress logged squared residuals on  $X$ s, obtain fitted values
- 4) Exponentiate fitted values
- 5) Re-estimate original equation with  $1/\text{exponentiated fitted values}$  as analytic weight

## Caveats

- All of the preceding assumes that our initial model meets the regression assumptions MLR1 through MLR4.
- If this is not the case, we can't fix the heteroskedasticity problem, we have other issues to deal with.
- Power: if you have little power in your regression (small sample size), you have little power to uncover heteroskedasticity
  - Conversely, much power = easy to discover heteroskedasticity, but might not matter



## In-class exercise, continued

- Questions 8 through 10

[Next time:

---

Homework 9 Problems C8.2, C8.4, C8.6 due 10/21

Read: Wooldridge Chapter 9