Heteroscedasticity

[NOTE: These notes draw heavily from Berry and Feldman, and, to a lesser extent, Allison, and Pindyck and Rubinfeld.]

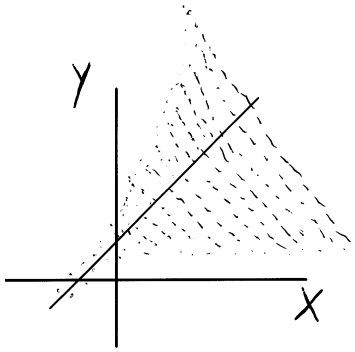
What heteroscedasticity is. Recall that OLS makes the assumption that

$V(\varepsilon_j) = \sigma^2$ for all j. That is, the variance of the error term is constant. (Homoscedasticity). If the error terms do not have constant variance, they are said to be heteroscedastic.

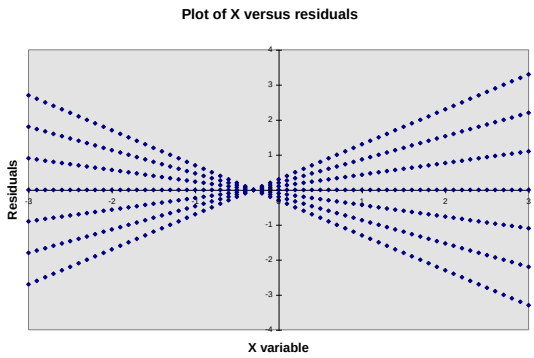
When heteroscedasticity might occur.

- Errors may increase as the value of an IV increases. For example, consider a model in which annual family income is the IV and annual family expenditures on vacations is the DV. Families with low incomes will spend relatively little on vacations, and the variations in expenditures across such families will be small. But for families with large incomes, the amount of discretionary income will be higher. The mean amount spent on vacations will be higher, and there will also be greater variability among such families, resulting in heteroscedasticity. Note that, in this example, a high family income is a necessary but not sufficient condition for large vacation expenditures. Any time a high value for an IV is a necessary but not sufficient condition for an observation to have a high value on a DV, heteroscedasticity is likely.

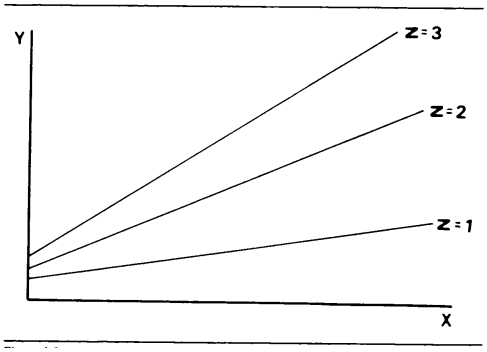
  Similar examples: Error terms associated with very large firms might have larger variances than error terms associated with smaller firms. Sales of larger firms might be more volatile than sales of smaller firms.



- Errors may also increase as the values of an IV become more extreme in either direction, e.g. with attitudes that range from extremely negative to extremely positive. This will produce something that looks like an hourglass shape:



- Measurement error can cause heteroscedasticity. Some respondents might provide more accurate responses than others. (Note that this problem arises from the violation of another assumption, that variables are measured without error.)
- Heteroscedasticity can also occur if there are subpopulation differences or other interaction effects (e.g. the effect of income on expenditures differs for whites and blacks). (Again, the problem arises from violation of the assumption that no such differences exist or have already been incorporated into the model.) For example, in the following diagram suppose that Z stands for three different populations. At low values of X, the regression lines for each population are very close to each other. As X gets bigger, the regression lines get further and further apart. This means that the residual values will also get further and further apart.



- Other model misspecifications can produce heteroskedasticity. For example, it may be that instead of using Y, you should be using the log of Y. Instead of using X, maybe you should be using X^2, or both X and X^2. Important variables may be omitted from the model. If the model were correctly specified, you might find that the patterns of heteroskedasticity disappeared.

Consequences of heteroscedasticity. Note that heteroscedasticity is often a by-product of other violations of assumptions. These violations have their own consequences which we will deal with elsewhere. For now, we'll assume that other assumptions except heteroscedasticity have been met. Then,

- Heteroscedasticity does *not* result in biased parameter estimates.

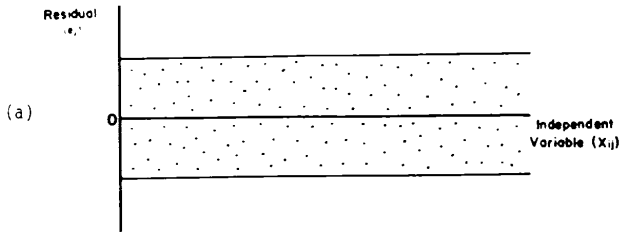
- However, OLS estimates are no longer BLUE. That is, among all the unbiased estimators, OLS does not provide the estimate with the smallest variance. Depending on the nature of the heteroscedasticity, significance tests can be too high or too low.
  - ✓ Why does this occur? Recall that OLS is designed to minimize Σe_i^2. It therefore gives more weight to the people with potentially the largest error terms. For example, suppose you have two cases, where Y = 100 for the first case and Y = 10,000 for the second. For the first case, it won't matter that much whether the predicted Y is off by 10, 20, or 30; either way, the squared residual will be fairly small. It will make much more difference if the predicted value for the second case is off by 100, 200, or 300. Ergo, the calculations wind up placing greater emphasis on the more extreme cases. This is undesirable, because the extreme cases are the ones furthest from the true regression line, hence they give you the least information about where the true regression line is.
  - ✓ Or, as Allison puts it: "The reason OLS is not optimal when heteroskedasticity is present is that it gives equal weight to all observations when, in fact, observations with larger disturbance variance contain less information than observations with smaller disturbance variance."
- In addition, the standard errors are biased when heteroskedasticity is present. This in turn leads to bias in test statistics and confidence intervals.
- Fortunately, unless heteroscedasticity is "marked," significance tests are virtually unaffected, and thus OLS estimation can be used without concern of serious distortion. But, severe heteroscedasticity can sometimes be a problem.

Warning: Heteroskedasticity can be very problematic with methods besides OLS. For example, in logistic regression heteroskedasticity can produce biased and misleading parameter estimates. I talk about such concerns in my categorical data analysis class.

Detecting Heteroscedasticity

Visual Inspection. Do a visual inspection of residuals plotted against fitted values; or, plot the IV suspected to be correlated with the variance of the error term. In Stata, after running a regression, you could use the `rvfplot` (residuals versus fitted values) or `rvpplot` command (residual versus predictor plot, e.g. plot the residuals versus one of the X variables included in the equation). In SPSS, plots could be specified as part of the Regression command.

- ✓ In a large sample, you'll ideally see an "envelope" of even width when residuals are plotted against the IV. In a small sample, residuals will be somewhat larger near the mean of the distribution than at the extremes. Thus, if it appears that residuals are roughly the same size for all values of X (or, with a small sample, slightly larger near the mean of X) it is generally safe to assume that heteroscedasticity is not severe enough to warrant concern.



- ✓ If the plot of residuals shows some uneven envelope of residuals, so that the width of the envelope is considerably larger for some values of X than for others, a more formal test for heteroscedasticity should be conducted.

Breusch-Pagan / Cook-Weisberg Test for Heteroskedasticity. The Breusch-Pagan test is designed to detect any linear form of heteroscedasticity. This test is an option built in to Stata; you could do it in SPSS but it would take more work. You run a regression, and then give the `estat hettest` command (or, `hettest` alone will work). Using the reg01 data,

```
. use http://www.nd.edu/~rwilliam/stats2/statafiles/reg01.dta, clear
. reg  income educ jobexp

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   46.33
       Model |  1538.22521     2  769.112605            Prob > F      =  0.0000
    Residual |  282.200265    17  16.6000156            R-squared     =  0.8450
-------------+------------------------------           Adj R-squared =  0.8267
       Total |  1820.42548    19  95.8118671            Root MSE      =  4.0743

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.933393   .2099494     9.21   0.000     1.490438    2.376347
      jobexp |   .6493654   .1721589     3.77   0.002     .2861417    1.012589
       _cons |  -7.096855   3.626412    -1.96   0.067    -14.74791    .5542051
------------------------------------------------------------------------------

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of income

         chi2(1)      =     0.12
         Prob > chi2  =   0.7238
```

Breusch-Pagan / Cook-Weisberg tests the null hypothesis that the error variances are all equal versus the alternative that the error variances are a multiplicative function of one or more variables. For example, in the default form of the `hettest` command shown above, the alternative hypothesis states that the error variances increase (or decrease) as the predicted values of Y increase, e.g. the bigger the predicted value of Y, the bigger the error variance is. A large chi-square would indicate that heteroscedasticity was present. In this example, the chi-square value was small, indicating heteroskedasticity was probably not a problem (or at least that if it was a problem, it wasn't a multiplicative function of the predicted values).

Besides being relatively simple, `hettest` offers several additional ways of testing for heteroskedasticity; e.g. you could test for heteroskedasticity involving one variable in the model, several or all the variables, or even variables that are not in the current model. Type `help hettest` or see the Stata reference manual for details.

Read on your own: Here (roughly) is how `hettest` works.

- First, you run a regression.
- Then, `hettest` uses the `predict` command to compute the predicted values ($\hat{y}$) and the residuals (e).
- The residuals are then squared and also rescaled so that their mean is 1. (This is accomplished by dividing each residual by SS Residual/ N, i.e. each squared residual is divided by the average of the squared residuals). The rescaling is necessary for computing the eventual test statistic.
- The squared residuals are then regressed on $\hat{y}$ and the test statistic is computed. The test statistic is the model (i.e. explained) sums of squares from this regression divided by two (take this part on faith!). If the null is true, i.e. there is no multiplicative heteroskedasticity, the test statistic has a chi-square distribution with 1 degree of freedom.
- If there is no heteroskedasticity, then the squared residuals should neither increase nor decrease in magnitude as $\hat{y}$ increases, and the test statistic should be insignificant.
- Conversely, if the error variances are a multiplicative function of one or more variables (e.g. as X increases, the residuals fall farther and farther from the regression line) then the test statistic will be significant.

Here is how you could interactively do the same thing that `hettest` is doing.

```
. quietly reg  income educ jobexp

. * compute yhat
. predict yhat if e(sample)
(option xb assumed; fitted values)

. * Compute the residual
. predict e if e(sample), resid

. * Square the residual, and rescale it so that the squared values
. * have a mean of 1.  This is needed for the eventual test statistic.
. gen esquare = e^2 / (e(rss)/e(N))

. * Regress squared residuals on yhat
. reg  esquare yhat

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    18) =    0.18
       Model |  .249695098     1  .249695098           Prob > F      =  0.6758
    Residual |  24.8679862    18  1.38155479           R-squared     =  0.0099
-------------+------------------------------           Adj R-squared = -0.0451
       Total |  25.1176813    19  1.32198323           Root MSE      =  1.1754

------------------------------------------------------------------------------
```

```
     esquare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yhat |   .0127408   .0299691     0.43   0.676     -.050222    .0757036
       _cons |   .6889345   .7774684     0.89   0.387     -.944466    2.322335
------------------------------------------------------------------------------

. * Compute test statistic.
. display "Chi Square (1) = " e(mss) / 2
Chi Square (1) = .12484755

. * Display the p value for the chi-square statistic
. display "Prob > chi2 = " chi2tail(1, e(mss)/ 2)
Prob > chi2 = .72383527
```

White's General Test for Heteroskedasticity. The default Breusch-Pagan test specified by `hettest` is a test for linear forms of heteroskedasticity, e.g. as $\hat{y}$ goes up, the error variances go up. In this default form, the test does not work well for non-linear forms of heteroskedasticity, such as the hourglass shape we saw before (where error variances got larger as X got more extreme in either direction). The default test also has problems when the errors are not normally distributed.

White's general test for heteroskedasticity (which is actually a special case of Breusch-Pagan) can be used for such cases. This can be estimated via the command `estat imtest, white` or just `imtest, white`. (Actually, the `white` option seems to matter rarely if ever in my experience; the Stata help says "White's test is usually very similar to the first term of the Cameron-Trivedi decomposition" normally reported by `imtest`.) You can also use Mark Schaffer's `ivhettest` (which offers several additional capabilities) or Baum & Cox's `whitetst`, both available from SSC. As the help for `whitetst` states,

> `whitetst` computes the White (1980) general test for heteroskedasticity in the error distribution by regressing the squared residuals on all distinct regressors, cross-products, and squares of regressors. The test statistic, a Lagrange multiplier measure, is distributed Chi-squared(p) under the null hypothesis of homoskedasticity. See Greene (2000), pp. 507-511. It is a special case of the Breusch-Pagan test for heteroskedasticity, which requires specification of an auxiliary variable list.

NOTE: Part of the reason the test is more general is because it adds a lot of terms to test for more types of heteroskedasticity. For example, adding the squares of regressors helps to detect nonlinearities such as the hourglass shape. In a large data set with many explanatory variables, this may make the test difficult to calculate. Also, the addition of all these terms may make the test less powerful in those situations when a simpler test like the default Breusch-Pagan would be appropriate, i.e. adding a bunch of extraneous terms may make the test less likely to produce a significant result than a less general test would.

Here is an example using `estat imtest, white`:

```
. use http://www.nd.edu/~rwilliam/stats2/statafiles/reg01.dta, clear
. quietly reg income educ jobexp
. estat imtest, white
```

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =      8.98
         Prob > chi2  =    0.1100

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |       8.98      5    0.1100
            Skewness |       2.39      2    0.3022
            Kurtosis |       0.98      1    0.3226
---------------------+-----------------------------
               Total |      12.35      8    0.1363
---------------------------------------------------
```

As noted before, White's general test is a special case of the Breusch-Pagan test, where the assumption of normally distributed errors has been relaxed (to do this, use the `iid` option of `hettest`) and an auxiliary variable list (i.e. the Xs, the Xs squared and the cross-product terms) is specified:

```
. quietly reg  income educ jobexp
. hettest  educ jobexp educ2 jobexp2 jobed, iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: educ jobexp educ2 jobexp2 jobed

         chi2(5)      =     8.98
         Prob > chi2  =   0.1100
```

Read on your own: Here (roughly) is how `imtest` works.

- First, you run a regression.
- Then, `imtest` uses the `predict` command to compute the residuals (e).
- The residuals are then squared.
- `imtest` creates new variables that are equal to the X-squared terms and the cross-products of the Xs with each other. So, for example, if X1 is the only variable, `imtest` computes $X1^2$. If you had X1 and X2, `imtest` would compute $X1^2$, $X2^2$, and X1*X2. If you had three Xs, `imtest` would compute $X1^2$, $X2^2$, $X3^2$, X1*X2, X1*X3, and X2*X3.
- The squared residuals are then regressed on the original Xs, the squared Xs, and the cross-product terms. The test statistic = $N * R^2$. If the null is true, i.e. there is no heteroskedasticity, the test statistic has a chi-square distribution with K*(K+3)/2 degrees of freedom. [NOTE: with 0/1 dichotomies, the squared terms are the same as the non-squared terms, which will cause some terms to drop out, reducing the d.f.]
- If there is no heteroskedasticity, then the test statistic should be insignificant.
- Conversely, if there is heteroskedasticity, then the test statistic will be significant.

- `imtest` also produces some additional tests for skewness and kurtosis that I won't discuss here.

Here is how you could interactively do the same thing that `imtest` is doing.

```
. quietly reg  income educ jobexp

. * Compute the residual
. predict e if e(sample), resid

. * Square the residual
. gen esquare = e^2

. * Compute squares and cross-products.
. gen educ2 = educ ^2
. gen jobexp2 = jobexp ^2
. gen jobed =  jobexp * educ

. * Regress the squared residuals on all the X terms
. reg  esquare educ jobexp educ2 jobexp2 jobed

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =    2.28
       Model |  11.2749039     5  2.25498078           Prob > F      =  0.1030
    Residual |  13.8427774    14  .988769814           R-squared     =  0.4489
-------------+------------------------------           Adj R-squared =  0.2521
       Total |  25.1176813    19  1.32198323           Root MSE      =  .99437

------------------------------------------------------------------------------
     esquare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0077052    .222315    -0.03   0.973    -.4845234     .469113
      jobexp |  -.4139054   .3654719    -1.13   0.276    -1.197765    .3699538
       educ2 |  -.0108245   .0092205    -1.17   0.260    -.0306005    .0089515
     jobexp2 |    .014225   .0096959     1.47   0.164    -.0065707    .0350208
       jobed |   .0160536    .014215     1.13   0.278    -.0144344    .0465417
       _cons |   3.020161   3.162498     0.95   0.356    -3.762723    9.803045
------------------------------------------------------------------------------

. * Test stat = N * R-squared
. display 20*.4489
8.978
```

Other Tests. There are lots of other tests for heteroskedasticity. They make different assumptions about the form of heteroskedasticity, whether or not error terms are normally distributed, etc. The readings go over some of these and if you give the command **`findit heteroskedasticity`** from within Stata you'll come up with more options. Also, Appendix A discusses the Goldfeldt-Quant test, which is somewhat antiquated, but which you may occasionally come across in your reading.

Dealing with heteroscedasticity

1. Respecify the Model/Transform the Variables. As noted before, sometimes heteroscedasticity results from improper model specification. There may be subgroup differences. Effects of variables may not be linear. Perhaps some important variables have been left out of the model. *If these are problems deal with them first!!!* Don't just launch into other

techniques, such as WLS, because they don't get to the heart of the problem.

Incidentally, Allison says (p. 128) "My own experience with heteroskedasticity is that is has to be pretty severe before it leads to serious bias in the standard errors. Although it is certainly worth checking, I wouldn't get overly anxious about it."

Warning: In general, I agree, with the qualifier that heteroskedasticity that results from model mis-specification is something to be concerned about. Indeed, I would say in such cases the problem isn't really heteroskedasticity, it is model mis-specification; fix that problem and the heteroskedasticity may go away. HOWEVER, by checking for heteroskedasticity, you may be able to identify model specification problems. So, *I would probably be a little more anxious about heteroskedasticity than Allison implies.*

2. Use Robust Standard Errors. Stata (but not SPSS) includes options with most routines for estimating *robust standard errors* (you'll also hear these referred to as Huber/White estimators or sandwich estimators of variance). As noted above, heteroskedasticity causes standard errors to be biased. OLS assumes that errors are both independent and identically distributed; robust standard errors relax either or both of those assumptions. Hence, when heteroskedasticity is present, robust standard errors tend to be more trustworthy.

With regards to the related problem of error terms not being independent: Survey data are often collected using clustering, e.g. a sample will be drawn of areas and individuals within those areas will then be selected for inclusion in the sample. Also, some groups will be deliberately oversampled, e.g. your sampling scheme might be set up to include a disproportionately large number of minorities, in order to ensure that you have enough minorities to do subsample analyses of them. Such strategies can lead to violations of the assumption that the error terms are independent of each other (since people sampled in clusters tend to be more similar than people sampled totally at random). There is a good discussion of this in "Sampling Weights and Regression Analysis" by Winship and Radbill, Sociological Methods and Research, V. 23, # 2, Nov 1994 pp. 230-257.

Another example (given by Hamilton in *Statistics with Stata, Updated for Stata 9*, pp. 258-259: 51 college students were asked to rate the attractiveness of photographs of unknown men and women. The procedure was reported 4 times, producing 204 sets of records. Hamilton says "It seems reasonable to assume that disturbances (unmeasured influences on the ratings) were correlated across the repetitions by each individual. Viewing each participant's four rating sessions as a cluster should yield more realistic standard errors." In this case you add the `cluster(id)` option to the regression command, where id is the id number of the subject.

As Allison points out, the use of robust standard errors does *not* change coefficient estimates, but (because the standard errors are changed) the test statistics will give you reasonably accurate p values. The use of Weighted Least Squares (described next) will also correct the problem of bias in the standard errors, and will also give you more efficient estimates (i.e. WLS will give you estimates that have the smallest possible standard errors). But, WLS requires more assumptions

and is more difficult to implement, so robust standard errors seem to be a more common and popular method for dealing with issues of heteroskedasticity.

Actually, it is not quite correct to say that SPSS can not compute robust standard errors; I did see a web page once (that took several pages to print out) that showed how to do it. But, with Stata, robust standard errors can usually be computed via the addition of two parameters, `robust` and `cluster`. The `robust` option relaxes the assumption that the errors are identically distributed, while `cluster` relaxes the assumption that the error terms are independent of each other. For example, rerunning the above regression with the `robust` option, we get

```
. reg  income educ jobexp, robust

Regression with robust standard errors                Number of obs =      20
                                                      F(  2,    17) =   48.15
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.8450
                                                      Root MSE      =  4.0743

------------------------------------------------------------------------------
             |               Robust
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.933393   .2006214     9.64   0.000     1.510119    2.356667
      jobexp |   .6493654   .1701214     3.82   0.001     .2904407     1.00829
       _cons |  -7.096855   3.365609    -2.11   0.050    -14.19767    .0039603
------------------------------------------------------------------------------
```

Comparing the results with the earlier regression, you note that none of the coefficient estimates changed, but the standard errors and hence the t values are a little different. Had there been more heteroskedasticity in these data, we would have probably seen bigger changes.

In some cases, Stata will use robust standard errors whether you explicitly ask for them or not. For example, if you have used clustering when collecting your data (and you tell Stata this via use of the `cluster` parameter) the error terms will not be independent. Hence, if you ever wonder why you are getting robust standard errors when you did not ask for them, it is probably because robust standard errors are more appropriate for what you are doing. (Indeed, with more complicated analyses, Stata will often surprise you by doing things you don't expect; if you can figure out why, you will learn a lot about advanced statistics!)

Caution: Do not confuse robust standard errors with robust regression. Despite their similar names, they deal with different problems:

Robust standard errors address the problem of errors that are not independent and identically distributed. The use of robust standard errors will not change the coefficient estimates provided by OLS, but they will change the standard errors and significance tests.

Robust regression, on the other hand, deals with the problem of outliers in a regression. Robust regression uses a weighting scheme that causes outliers to have less impact on the estimates of regression coefficients. Hence, robust regression generally will produce different coefficient estimates than OLS does.

NOTE: For Weighted Least Squares, I will only cover Case II in class.

3. Use Weighted Least Squares. A more difficult option (but superior when you can make it work right) is the use of weighted least squares. Generalized Least Squares [GLS] is a technique that will always yield estimators that are BLUE when either heteroscedasticity or serial correlation are present. OLS selects coefficients that minimize the sum of squared regression residuals, i.e.

$$\Sigma(Y_j - \hat{Y}_j)^2$$

GLS minimizes a *weighted* sum of squared residuals. In the case of heteroscedasticity, observations expected to have error terms with large variances are given a smaller weight than observations thought to have error terms with small variances. Specifically, coefficients are selected which minimize

$$\sum_{j=1}^{N} \frac{(Y_j - \hat{Y}_j)^2}{VAR(\varepsilon_j)}$$

OLS is a special case of GLS, when the variance of all residuals is the same for all cases. The smaller the error variance, the more heavily the case is weighted. Intuitively, this makes sense: the observations with the smallest error variances should give the best information about the position of the true regression line.

GLS estimation can be a bit complicated. However, under certain conditions, estimators equivalent to those generated by GLS can be obtained using a Weighted Least Squares (WLS) procedure utilizing OLS regression on a transformed version of the original regression model.

See Appendix B for a discussion of the rarely used WLS Case I, where the error variances are somehow miraculously known.

WLS CASE II. A far more common case is when we think the error variances vary directly with an IV. For example, suppose we think there is heteroscedasticity in which the standard deviation of the error term is linearly related to X1, i.e.

$$\sigma_i = CX_{1i}$$

In other words, the larger X is, the more error variability there will be.

[OPTIONAL] Conceptually, we then do two things. First, we divide each IV and DV by X1 (or whichever X is believed to be causing heteroscedasticity). We then do a regression with the transformed variables. Specifically, we compute

$Y^* = Y/X1$
$X0^* = 1/X1$
$X1^* = X1/X1 = 1$
$X2^* = X2/X1$
etc. for all the X's.

We then estimate the model

$Y^* = \alpha X0^* + \beta_1 + \beta_2 X2^* + ... + \beta kXK^* + \varepsilon^*$

With this equation, note that (1) the transformed errors will once again be homoscedastic (2) The only "tricky" thing is matching the coefficients up correctly. The intercept for the transformed equation is the estimate of β_1 in the original equation. The coefficient for X0* is the intercept of the original equation. That is, the intercept and first regression coefficient are "flipped."

Using Stata for WLS Case II. The `aweights` parameter (analytical weights) in Stata provides one means for doing WLS (this corresponds to SPSS method 2, which I describe below). Take the X which you think is causing the heteroscedasticity, and proceed as follows:

```
. gen inveduc = (1/educ)^2

. reg income  educ jobexp [aw = inveduc]
(sum of wgt is   4.4265e-01)

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   86.20
       Model |  1532.21449     2  766.107244           Prob > F      =  0.0000
    Residual |  151.090319    17  8.88766581           R-squared     =  0.9102
-------------+------------------------------           Adj R-squared =  0.8997
       Total |  1683.30481    19  88.5949898           Root MSE      =  2.9812

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.795724   .1555495    11.54   0.000     1.467544    2.123905
       jobexp |   .4587992   .1628655     2.82   0.012      .115183    .8024155
       _cons |  -3.159669    1.94267    -1.63   0.122    -7.258345    .9390065
------------------------------------------------------------------------------
```

Note that both the coefficients and the standard errors are different from before. If we were confident that we were using the correct weights, this would be a superior solution to anything we have done before. If, however, the weights are wrong, we may have just made things worse! I actually prefer SPSS for WLS (at least for Case II), because I think it provides a better test of what weighting scheme is best.

See Appendix C for how to do this in SPSS.

4. Summary of recommendations for dealing with heteroskedasticity. In most instances, I would recommend option 1 (respecify the model or transform the variables) or option 2 (robust standard errors). Most of the examples I have seen using Stata take those approaches. However, in special cases, option 3 (WLS) can be the best. What makes WLS hard, though, is knowing what weights to use. The weights either have to be known for some reason or you have to have some sort of plausible theory about what the weights should be like, e.g. error terms go up as X goes up.

Appendix A: Goldfeldt-Quant (GQ) test.

[NOTE: This section is primarily included in case you come across this test in your readings. The tests we discuss above are now considered superior. Skim through this on your own.] The GQ test can be used when it is believed that the variance of the error term increases consistently or decreases consistently as X increases. The procedure is as follows:

- ✓ Order the data by the magnitude of the IV X which is thought to be related to the error variance.
- ✓ Omit the middle d observations. Typically, d might equal 20% of the sample size. This technically isn't necessary, but experience shows that this tends to improve the power of the test (i.e. makes it more likely you will reject the null hypothesis when it is false).
- ✓ Fit two separate regressions—one for the low values of X and one for the high. Each will involve (N - d)/2 cases, and [(N - d)/2 - 2] degrees of freedom.
- ✓ Calculate the residual sum of squares for each equation: SSE_{low} for the low X's, and SSE_{high} for the high X's.
- ✓ If you think the error variance is an *increasing* function of X, compute the following statistic:

$$F_{(N-d-4)/2,(N-d-4)/2} = \frac{SSE_{high}}{SSE_{low}}$$

If you think the error variance is a decreasing function of X, then reverse the numerator and denominator. This statistic assumes that the error process is normally distributed and there is no serial correlation. The F statistic should approximately equal 1 if the errors are homoscedastic. If the F value is greater than the critical value, we reject the null hypothesis of homoscedasticity.

- ✓ Note that this is a test of

$H_{0:}$: $\sigma_I^2 = \sigma_2^2 = \sigma 3^2 ... = \sigma_N^2$

H_A: $\sigma_I^2 = CX_i^2$ (Where C is some constant)

- ✓ You can easily modify the procedure when you have more than one IV in the model. Again, you order by *one* of the X variables (the one you think may be causing the problem). The other steps are the same, except that the F statistic is

$$F_{(N-d-2K-2)/2,(N-d-2K-2)/2} = \frac{SSE_{high}}{SSE_{low}}$$

(Note that in the bivariate regression case, K = 1, hence the first F test is just a special case of this one).

- ✓ To do this in SPSS, you would do something like the following:
  - ■ Run a frequencies on the X variable. Determine the "cutoff" values for the bottom 40% and the top 60%. Suppose, for example, that this conveniently worked out to be X = 40 and X = 60.

- REGRESSION VARIABLES = Y X/
  SELECT=X LE 40/
  DEPENDENT = Y/
  METHOD = ENTER X/

  REGRESSION VARIABLES = Y X/
  SELECT=X GE 60/
  DEPENDENT = Y/
  METHOD = ENTER X/

  Compute the F statistic by hand, using the Residual sums of squares from the two equations.

- Specific example: (Not a great one, but handy...) Using the income, education, jobexp example I've used before, I dropped the middle 8 cases and produced the following.:

```
-> * Select cases with educ <= 10.

-> REGRESSION /DESCRIPTIVES DEF
->            /VARIABLES EDUC JOBEXP INCOME
->            /SELECT=EDUC LE 10
->            /DEPENDENT INCOME
->            /ENTER EDUC JOBEXP.

* * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           2           388.59613        194.29806
Residual             3           113.01221         37.67074

F =       5.15780       Signif F =  .1069

-> * Select cases with educ >= 15.

-> REGRESSION /DESCRIPTIVES DEF
->            /VARIABLES EDUC JOBEXP INCOME
->            /SELECT=EDUC GE 15
->            /DEPENDENT INCOME
->            /ENTER EDUC JOBEXP.


Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           2           411.23135        205.61567
Residual             3            45.53699         15.17900

F =      13.54607       Signif F =  .0315
```

In this example, N = 20, d = 8, K = 2, SSE_{high} = 45.54, SSE_{Low} = 113.01. If there is heteroscedasticity, it is greater at lower values of education, ergo GQ = SSE_{low}/SSE_{high} = 113.01/45.54 = 2.48 with d.f. = 3,3. This is not significant at the .05 level.

✓ Here are the corresponding Stata Cards:

```
. use http://www.nd.edu/~rwilliam/stats2/statafiles/reg01.dta, clear
. reg income educ jobexp if educ <=10

      Source |       SS       df       MS              Number of obs =       6
-------------+------------------------------           F(  2,     3) =    5.16
       Model |  388.596126     2  194.298063           Prob > F      =  0.1069
    Residual |  113.012206     3  37.6707353           R-squared     =  0.7747
-------------+------------------------------           Adj R-squared =  0.6245
       Total |  501.608332     5  100.321666           Root MSE      =  6.1376

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.975563   .9958481     1.98   0.142     -1.19367    5.144796
      jobexp |   1.057993   .6990496     1.51   0.227    -1.166695    3.282681
       _cons |  -12.45698   10.21039    -1.22   0.310      -44.951    20.03704
------------------------------------------------------------------------------

. reg income educ jobexp if educ >=15

      Source |       SS       df       MS              Number of obs =       6
-------------+------------------------------           F(  2,     3) =   13.55
       Model |  411.231321     2   205.61566           Prob > F      =  0.0315
    Residual |  45.5369983     3  15.1789994           R-squared     =  0.9003
-------------+------------------------------           Adj R-squared =  0.8338
       Total |  456.768319     5  91.3536638           Root MSE      =   3.896

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   2.910157    .808277     3.60   0.037     .3378588    5.482455
      jobexp |   .6475437   .2526255     2.56   0.083    -.1564234    1.451511
       _cons |  -23.24063   13.03198    -1.78   0.173     -64.7142    18.23295
------------------------------------------------------------------------------
```

✓ GQ is not helpful if you think there is heteroscedasticity, but it isn't monotonic, i.e. you think extreme values of X at either end produce larger error variances (like the hourglass shape we saw before).

✓ GQ has been found to be reasonably powerful when we are able to correctly identify the variable to use in the sample separation. This does limit its generality, however; sometimes heteroskedasticity might be a function of several variables. The other tests we discuss are more flexible and also simpler.

Appendix B: WLS CASE I.

[OPTIONAL] Suppose the error variances are miraculously known. We then do the following:

Divide each variable by σ_j, i.e. compute

$$Y_i^* = \frac{Y_i}{\sigma_i}$$

$$X_{0i}^* = \frac{1}{\sigma_i}$$

$$X_{ki}^* = \frac{X_{ki}}{o_i} \text{ for all K X's}$$

You then estimate the model

$$Y_i^* = \alpha X_{0i}^* + \sum_{k=1}^{K} \beta_k X_{ki}^* + \varepsilon_i^*$$

which is equivalent to

$$\frac{Y_i}{\sigma_i} = \alpha \frac{1}{\sigma_i} + \sum_{k=1}^{K} \beta_k \frac{X_k}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}$$

Why does this work? Recall that (1) $VAR(\varepsilon_i) = \sigma_I^2$, (2) If C is a constant and X is a variable, $V(X/C) = V(X)/C^2$ (3) ergo, $VAR(\varepsilon_i/\sigma_I) = 1$. Ergo, the transformed error term is homoskedastic, and all the assumptions of OLS regression are met.

With this approach, note that there is no constant in the transformed equation. Rather, the coefficient for X0* corresponds to the intercept in the original equation.

Appendix C: Using SPSS for WLS Case II

[Read this section if your life depends on doing WLS with SPSS; otherwise just skim it.] There are at least three ways to do WLS in SPSS. Not all three options will work in all versions of SPSS; and, other statistical programs may or may not have similar options available. I'll therefore illustrate all 3 ways, going from the easiest to the hardest.

1. (EASIEST) By far the simplest approach is to use the SPSS WLS command (and SPSS seems to be easier than Stata in this respect). On the SPSSWIN menu, you will find this under Statistics/Regression/Weight Estimation.

```
* 1. WLS THE EASIEST WAY.
TSET NEWVAR=NONE.
WLS income  WITH educ jobexp
  /SOURCE educ
  /POWER 2
  /CONSTANT
  /PRINT all.
```

Weighted Least Squares

```
Source variable..    EDUC                  POWER value =  2.000
Dependent variable.. INCOME

Multiple R           .95407
R Square             .91024
Adjusted R Square    .89968
Standard Error       .44351

            Analysis of Variance:

               DF   Sum of Squares      Mean Square

Regression      2        33.911556        16.955778
Residuals      17         3.343989          .196705

F =      86.19893       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta         T  Sig T

EDUC           1.795724     .155550    .872919    11.544  .0000
JOBEXP          .458799     .162866    .213008     2.817  .0119
(Constant)    -3.159670    1.942670               -1.626  .1222

Log-likelihood Function =  -58.345242
```

Note that, in this example, I constrained the power value to be 2. However, it is also legitimate to have something like

```
WLS income  WITH educ jobexp
  /SOURCE educ
  /POWER -2 to 2 by .5
  /CONSTANT
  /PRINT all.
```

This will try out a range of power values, and tell you which one is best. (In actuality, it turns out the best power is 0, meaning that you don't actually need WLS for this example. As far as I know, Stata does not offer anything comparable, but I could be wrong)

2. (NEXT MOST EASIEST) The next simplest approach is to compute a weighting variable and then use the SPSS Regression Regwgt parameter. (Compare this to Stata using aweights) Take the X which you think is causing the heteroscedasticity, and proceed as follows:

```
-> Compute InvEduc = (1/Educ)**2.
-> REGRESSION /DESCRIPTIVES DEF
->            /VARIABLES EDUC JOBEXP INCOME
->            /regwgt = InvEduc
->            /DEPENDENT INCOME
->            /ENTER EDUC JOBEXP.

          * * * *   M U L T I P L E   R E G R E S S I O N   * * * *
Weighted Least Squares -  Weighted By.. INVEDUC

            Mean  Std Dev  Label

EDUC       5.029     .681
JOBEXP    11.538     .650
INCOME    11.165    1.400

N of Cases =    20

Correlation:
               EDUC     JOBEXP     INCOME

EDUC          1.000       .277       .932
JOBEXP         .277      1.000       .455
INCOME         .932       .455      1.000

          * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Multiple R           .95407
R Square             .91024
Adjusted R Square    .89968
Standard Error       .44351

Analysis of Variance
                   DF      Sum of Squares      Mean Square
Regression          2            33.91156         16.95578
Residual           17             3.34399           .19671

F =      86.19893       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta         T  Sig T

EDUC           1.795724     .155550    .872919    11.544  .0000
JOBEXP          .458799     .162866    .213008     2.817  .0119
(Constant)    -3.159670    1.942670               -1.626  .1222
```

3. (HARDEST – Optional) Finally, the really hard way (but which will probably work with just about any statistical package) is to compute variables the way we described above.

```
-> * WLS the hard way.
-> COMPUTE newX0 = 1/Educ.
-> COMPUTE neweduc = 1.
```

```
-> COMPUTE newjob = jobexp/educ.
-> COMPUTE newinc = income/educ.

-> REGRESSION /DESCRIPTIVES DEF
->            /VARIABLES newx0 newjob newinc
->            /DEPENDENT newinc
->            /ENTER newx0 newjob.

* * * *   M U L T I P L E   R E G R E S S I O N   * * * *

           Mean  Std Dev  Label

NEWX0      .111     .101
NEWJOB    1.401    1.208
NEWINC    2.087     .531

N of Cases =    20

Correlation:
              NEWX0     NEWJOB     NEWINC

NEWX0         1.000       .856       .291
NEWJOB         .856      1.000       .528
NEWINC         .291       .528      1.000

           * * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Multiple R           .61306
R Square             .37584
Adjusted R Square    .30241
Standard Error       .44351

Analysis of Variance
                   DF      Sum of Squares      Mean Square
Regression          2             2.01360          1.00680
Residual           17             3.34399           .19671

F =       5.11831       Signif F =  .0182

------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta         T  Sig T

NEWX0         -3.159670    1.942670   -.602561    -1.626  .1222
NEWJOB          .458799     .162866   1.043643     2.817  .0119
(Constant)     1.795724     .155550               11.544  .0000
```

Note that, with all three approaches, you get the same metric coefficients and same T values. However, the R^2, F and the standardized coefficients are different (and wrong) in the last approach, reflecting the fact that transformed vars are being analyzed and the global hypothesis being tested is somewhat different. Ergo, if your statistical package permits it, you should try to use one of the other approaches.