

LINEAR MODEL

SP03 9 - 2

An assumption dealing with the disturbance term u_i ...

HETEROSKEDASTICITY

In: $Y_i = \beta_0 + \beta_1 X_i + u_i$

Assumption: That the errors u_i are *homoskedastic*, conditional on the X's

- That is, $\text{Var}(u_i) \equiv E(u_i^2) = \sigma^2$ for all i
- In matrix terms (again), that $E(\mathbf{U}\mathbf{U}') = \sigma^2 \mathbf{I}$

If this does not hold (but all other assumptions do), we say that the errors are ***heteroskedastic***...

- I.e. $\text{Var}(u_i) = \sigma_i^2$

- In matrix notation, $E(\mathbf{U}\mathbf{U}') = \sigma^2 \mathbf{\Omega}$, where

$$\mathbf{\Omega} = \begin{bmatrix} 1/\lambda_1 & 0 & \dots & 0 \\ 0 & 1/\lambda_2 & \dots & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & 1/\lambda_N \end{bmatrix}$$

- * Note that the off-diagonal elements are all zero, means...
- * ...we're assuming *no autocorrelation*.

Note a couple things about heteroskedasticity...

- It is a ***population*** characteristic...
 - * That is, it's a problem with the population of data (vs. multicollinearity, which is a sample issue)
 - * This means that just adding more data won't solve the problem...
- It is a problem ***conditional*** on the values of X and β
 - * This means that it is linked very closely to model specification
 - * More on that in a bit...

WHY might one encounter heteroscedasticity?

Book answers:

1) ***Learning Models***

- * The effect of learning/training on performance
- * One not only gets better at doing something, but the variance decreases
- * This implies that as the dependent variable gets larger, the variance gets smaller
 - + E.g. errors on a typing test...
 - + Or education and the accuracy of answers to informational questions on a survey

2) “Discretionary Income”: I think of this as the “**bounds**” effect...

- * As discretionary income increases, both the absolute level and the variability in spending will increase
 - + B/c there is more to spend, there is the *possibility* of spending a lot more, or a lot less (i.e., greater variance)
- In political science, cases where the variability is “bounded” by the mean (level)
 - + E.g. Number of dissents in Supreme Court cases, by year (1800-1992)
 - * It’s a count, so it can only vary to the extent that the values are themselves larger
 - * As dependent variable increases, variance increases...
 - * In fact, a count variable is often modeled as a *Poisson process*, where the mean and the variance are always equal...
 - + Of presidential approval (percentage approve/disapprove)
 - * This is a percentage, and is bounded by zero and 100
 - * What happens as the approval gets very high/low? (A: variance decreases)

3) Variation in data collection techniques, reducing ***measurement error*** in pooled observations

- * E.g. CATI systems, different RA’s, etc...
- * Also: *Aggregating* data over groups which are different in size
- * Can make a big difference...

4) Presence/influence of **outliers**

- * Observations that are substantially “different” from the others
- * Can influence the effect of the X’s on Y, and therefore the variability of the u’s across different observations *i*

Note that if $\sigma_i^2 = \sigma^2$ (or, alternately, $\Omega = \mathbf{I}$) the two formulas are the same...

In general:

- Estimates of $\text{Var}(\beta)_{\text{HET}}$ will be **inefficient** (i.e., won't have smallest variance)
 - * Our confidence intervals will be "too wide"
 - * In this case, we increase the possibility of a Type II error
- Since the traditional formula for $\text{Var}(\beta)$ is a *biased* estimate of $\text{Var}(\beta)_{\text{HET}}$, they will be biased, *generally downwards*...
 - * This means that our confidence intervals will be too narrow
 - * Increases the odds of making a Type I error
 - * This is more likely to happen...

Intuitively, this is because we're ignoring information contained in the residuals

Think of a population with two "types" one where the conditional variance of Y (i.e. $\text{Var}(u)$) is much greater than the other...

- E.g. one where $Y = 2 + 2X + 2$
the other $Y = 2 + 2X + 12$
- The linear relationship is the same for the two types...
- Since there's less error variation in the first type, we can be more "sure" of the relationship (in MSE terms) by using those observations
- If we sample from these two types, we'd like to "weight" the observations of the second type less, since we're "less sure" about them

BUT

- OLS treats all observations the same
 - * In other words, it "weights" all the (squared) errors equally
 - * Ignores the greater or lesser information about the relationship

More on this idea of weighting when we get to WLS/GLS on Tuesday...

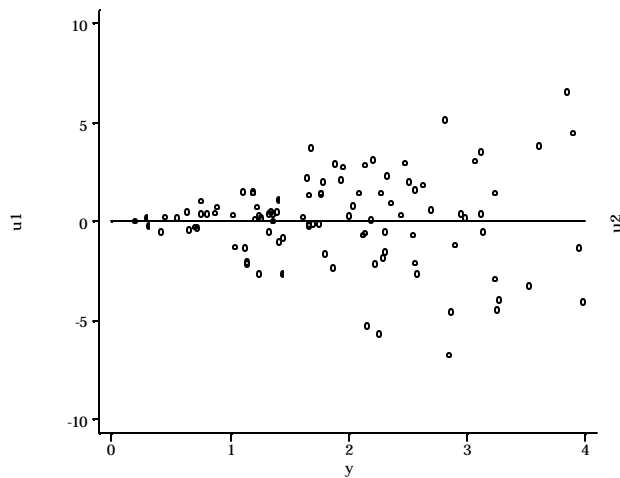
HOW TO DETECT HETEROSKEDASTICITY???

As with autocorrelation, most (all) tests are based on the observed \hat{u}_i 's...

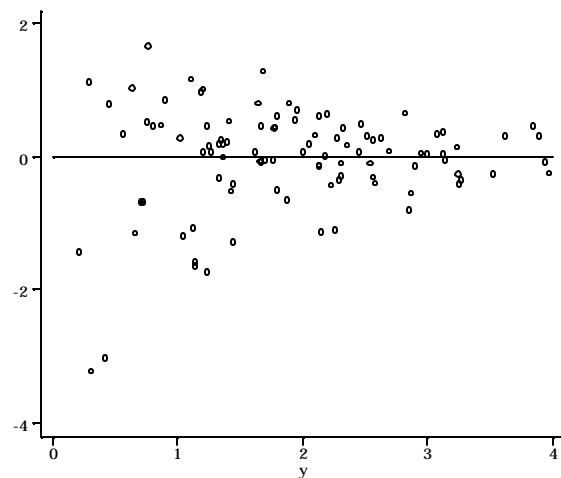
Graphical Methods

1) LOOK AT THE BLOODY RESIDUALS

- Since we're interested in the variance of the u 's (i.e. $E(u_i^2)$), *plot* the \hat{u}_i^2
 - * Can plot them against \hat{y} from the regression
 - + Tells if the mean/variance of \hat{u} varies with the predicted value if y
 - + If not, the pattern will be simply spherical/random
 - + Other patterns:



Var(\hat{u}) increases with Y



Var(\hat{u}) decreases with Y

- Another alternative is to *plot them against one or more of the X variables*
 - * Again, to see if the variance of the residuals varies systematically with any of the RHS variables
 - * Look for similar patterns as above...

As with autocorrelation, graphical methods can be quite powerful for suggesting the presence and/or form of heteroskedasticity...

More Formal Statistical Tests

All are based on the idea that the residuals vary according to some function of the predicted dependent or independent variable(s).

Park Test:

- Regress the (log of the) squared residuals on the (log of an) independent variable
- If the coefficient is statistically significant, it is evidence of heteroskedasticity...
- Can do this for each of the RHS variables
- A good place to start, but there are better ways...

Glejser Test:

- Regress the absolute values of the residuals $|\hat{u}_i|$ on some function of the independent variable
 - * E.g. X_i , or $1/X_i$, or X_i^{-1} , or $1/X_i^{-1}$...
- If the value of the estimated coefficient for this regression is statistically significant, its evidence of heteroskedasticity
- Test is only useful for *large samples*; only suggestive in small samples...

Breusch-Pagan-Godfrey Test:

- Run OLS, obtain \hat{u}_i^2
- Divide each of the squared residuals by (RSS/N) (this is one estimate of σ^2); call these p_i (sometimes called *generalized residuals*)
- Regress the p_i against all the m variables you might think are causing the heteroskedasticity...
- Take the ESS of this regression, and divide it by two
 - * This statistic follows a chi-square distribution with $(m-1)$ degrees of freedom
 - * Rejecting the null hypothesis indicates heteroskedasticity
- This is also a *large-sample test*; only suggestive in small samples...

White's Test:

- Regress the \hat{u}_i^2 on all the independent variables, their squares, and their cross-products (for a total of l regressors)
- Multiply the R^2 from this regression times N
- This statistic is distributed as chi-square, with l degrees of freedom
 - * Rejecting the null hypothesis is evidence of heteroskedasticity...

There are other tests as well...

- Some based on “ranking the observations” according to \hat{u}_i or X_i

E.g.:

Spearman Rank Correlation Test:

- Comparing the “rank” of observations based on \hat{u}_i with that based on \hat{Y}_i or X_i

Goldfeld-Quandt Test

- Taking out “middle observations” and running separate auxiliary regressions on the extreme values
- Then can do an F-test on the ratio of the two ESS's; if the ratio is near 1.0, this is evidence against heteroskedasticity

Lots of other tests as well

- A widely-studied area...

WHAT TO DO IF YOU'VE GOT IT???

Next time...

LINEAR MODEL

SP03 10 - 1

What do we do about heteroskedasticity?

Let's reconsider the basic model with heteroskedasticity:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\text{Var}(u_i) = \sigma_i^2$$

We said before we want to use the variance information to “weight” observations

- Give those with large variances less weight
- This with smaller variances should “count more”...

One way to do this is to divide each side of the equation by σ_i :

$$Y_i/\sigma_i = \beta_0(1/\sigma_i) + \beta_1(X_i/\sigma_i) + u_i/\sigma_i$$

Now the variance of the “error term” is

$$\begin{aligned} E(u_i/\sigma_i)^2 &= 1/\sigma_i^2 E(u_i)^2 \\ &= \sigma_i^2/\sigma_i^2 \\ &= 1 \end{aligned}$$

In other words, its homoskedastic!

- This means that the estimator of the “rescaled” variables is BLUE!

What does that estimator look like?

Define $w_i = 1/\sigma_i$ as the “weight” given to observation i ...

Then the above equation is:

$$w_i Y_i = \beta_0 w_i + \beta_1 w_i X_i + w_i u_i$$

Recall that the idea is to minimize the sum of squared errors...

We can write the sum of squared (homoskedastic) errors as:

$$\begin{aligned} \sum (w_i u_i)^2 &= \sum w_i (Y_i - \beta_0 - \beta_1 X_i)^2 \\ &= \sum w_i (Y_i^2 + \beta_0^2 + \beta_1^2 X_i^2 - 2\beta_0 Y_i - 2\beta_1 X_i Y_i - 2\beta_0 \beta_1 X_i) \end{aligned}$$

To minimize this, we take the derivatives with respect to the two coefficients β_0 and β_1 , and set them equal to zero...

$$\begin{aligned}\frac{\partial \Sigma(w_i u_i)^2}{\partial \beta_0} &= 2 \Sigma w_i (\beta_0 + \beta_1 X_i - Y_i) \\ &= 2 \Sigma w_i (Y_i - \beta_0 - \beta_1 X_i) (-1) \\ \frac{\partial \Sigma(w_i u_i)^2}{\partial \beta_1} &= 2 \Sigma w_i (-X_i Y_i + \beta_0 X_i - \beta_1 X_i^2) \\ &= 2 \Sigma w_i (Y_i - \beta_0 - \beta_1 X_i) (-X_i)\end{aligned}$$

Setting these equations equal to zero yields the normal equations for the **generalized least squares** (GLS) model:

$$\begin{aligned}\Sigma w_i Y_i &= \beta_0 \Sigma w_i + \beta_1 \Sigma w_i X_i \\ \Sigma w_i X_i Y_i &= \beta_0 \Sigma w_i X_i + \beta_1 \Sigma w_i X_i^2\end{aligned}$$

- This is sometimes also called *weighted least squares* (WLS)
- The normal equations for GLS are very similar to those for OLS we did a few weeks ago...
- In fact, **GLS is just OLS on variables that have been transformed to meet the OLS assumptions**, as is the case here

In the matrix context, we have:

$$b = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{Y}$$

- Note that both “terms” ($\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$) are multiplied by the inverse of the diagonal matrix of “weights” associated with each observation
- This is identical to what we’re doing above in the two-variable case

The GLS/WLS approach is **very** flexible, and has some nice properties...

- If we think that the nature of the heteroskedasticity is related to one of the X variables in some way, we can “weight” the observations accordingly...

Examples:

- If $\mathbf{Var}(u_i) = s^2 X_i^2$ (Variance of u is proportional to X_i^2)

then

* Divide each side of the equation by X_i :

$$Y_i/X_i = \beta_0(1/X_i) + \beta_1(X_i/X_i) + u_i/X_i$$

* Now $\text{Var}(u_i/X_i) = E(u_i/X_i)^2 = \sigma^2$: homoskedastic

* Simply estimate this transformed equation using OLS

* “Get back” to the original coefficients by multiplying by X_i

- Likewise, if $\mathbf{Var}(u_i) = s^2 X_i$ (Variance of u is proportional to X_i)

then

* Divide each side of the equation by $\sqrt{X_i}$:

$$Y_i/\sqrt{X_i} = \beta_0(1/\sqrt{X_i}) + \beta_1(X_i/\sqrt{X_i}) + u_i/\sqrt{X_i}$$

* Now $\text{Var}(u_i/\sqrt{X_i}) = E(u_i/\sqrt{X_i})^2 = 1/X_i E(u_i)^2 = \sigma^2$: homoskedastic

* Simply estimate this transformed equation using OLS

* “Get back” to the original coefficients by multiplying by $\sqrt{X_i}$

A more common example is where we have aggregated data across the various *is*, in which the variance of u is proportional to N_i :

$$\mathbf{Var}(u_i) = s^2 N_i$$

- E.g., state averages, where there are large differences in state populations

- If this is the case, we want to weight the observations by

How does one do this in Stata?

- Use `-reg-`, with the `[aweight]` option...

- e.g.: `.reg y x1 x2 x3 x4 [aweight=1/pop]`

One problem with GLS/WLS, however, is that we may not know σ_i^2 *at all*

- Or, analogously, we don't know Ω

Under these circumstances, there are a couple ways to proceed:

“Feasible Generalized Least Squares” (FGLS)

- Obtain a consistent *estimate* of the variance of the errors (i.e., the σ_i^2)
- For example, regressing the estimated \hat{u}_i^2 's on the variables, and generating predicted values
- Then substitute these (consistently) estimated error variances into the GLS weighting formulation
- FGLS is neither unbiased, nor efficient, nor (for that matter) even linear

BUT

- It IS asymptotically equivalent to GLS, meaning that it can be used if samples are quite large with good results

Alternatively, one can use **White's (1980) "heteroskedastic-consistent" covariance matrix estimate.**

- Sometimes referred to as Huber's (1967) method, or "robust" standard errors
- Recall that the heteroskedastic variance formula is $\text{Var}(\beta) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$
- Whereas one would normally have to come up with an estimate of \mathbf{W} , White figured out that it only necessary to consistently estimate $(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})$
- Used the estimated OLS residuals and the \mathbf{X} variables...
- In matrix notation, rather than the usual OLS:

$$\text{Var}(b) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

we use:

$$\text{Var}(b)_{\text{White}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{S}\hat{u}_i^2\mathbf{x}_i\mathbf{x}_i')(\mathbf{X}'\mathbf{X})^{-1}$$

where \mathbf{x}_i is the vector of independent variables for the i th observation and the \hat{u}_i^2 's are the estimated squared residuals from OLS estimation.

- The basic idea is that one calculates the variances using the observation-specific OLS-estimated residual variances.
- Including the RHS variables in the "weight" also takes care of much of any heteroskedasticity that might be related to those variables.
 - * Most statistical packages (but not SPSS...) will compute these s.e.'s for you automatically (in Stata, use the **-robust-** option).
- Points about the White method:
 - * It assumes that the investigator *does not know* the form of the heteroskedasticity (other alternatives for *grouped* data, or when more is known about the nature of the heteroskedasticity – **see notes at the end**).
 - * The usual t and F tests are now only **asymptotically** valid.
 - * These variance estimates are slightly worse than OLS estimates if the errors are truly homoskedastic,

BUT
 - * Are WAY better than OLS if the data are heteroskedastic in any way
 - * Also, they are *consistent*; that is, they become more accurate in increased sample sizes.
 - * In general, its not a bad idea to use these, especially if you have reason to suspect heteroskedasticity.

OTHER APPROACHES...

Downs & Rocke (1979)

- Simple article, suggesting how heteroskedasticity might in fact be of substantive interest...
- The idea is that we may actually be *interested in* the variance
- Especially true when variance has a substantively interesting interpretation or meaning
- Spawned a good deal of research in this area...

Gronke (1997)

- Presidents might care about variance, as well as the mean, in their presidential approval
 - * Would prefer smaller variances rather than larger...
 - * The variance might easily be a function of some independent variables, e.g. strength of party attachments
- Explicit parameterization of the variance term
- Model is:

$$\mathbf{Y} = \mathbf{b}\mathbf{X} + \mathbf{U}$$

$$\text{Var}(\mathbf{U}) = f(\mathbf{g}\mathbf{Z})$$

- Estimated via MLE (skip that part, for now...)
- The idea is that the variability in the estimates (and thus the precision with which we can talk about our “mean” parameters” will *also* depend on the independent variables
- Makes sense: we might be very sure that a strong partisan will be pro- or anti- Clinton, but less sure about an independent...

This is a MUCH BETTER way of dealing with heteroskedasticity than just correcting for it...

In Stata: `-regh-`...

Robust Standard Errors with “Clustering”... (MLE explanation...)

Standard ML variance estimate

$$V = \left(\frac{-\partial^2 \ln L}{\partial \beta^2} \right)^{-1}$$

General, “Robust” estimate:

$$V_R = V \sum_{i=1}^N \left[u_i' u_i \right] V$$

where u_i is the contribution of i to the scores $\partial \ln L / \partial \beta$, i.e., $\partial \ln L_i / \partial \beta$, evaluated at the estimated of β .

“Clustered” Estimate:

$$V_C = V \sum_{j=1}^{N_C} \left[\left(\sum_{i=1}^{n_j} u_{ij} \right)' \left(\sum_{i=1}^{n_j} u_{ij} \right) \right] V$$

where each of the N_C “clusters” $j = \{1, 2, \dots, N_C\}$ consists of n_j observations $i = \{1, 2, \dots, n_j\}$