Heteroskedasticity

- Heteroskedasticity means that the variance of the errors is not constant across observations.

- In particular the variance of the errors may be a function of explanatory variables.

- Think of food expenditure for example. It may well be that the "diversity of taste" for food is greater for wealthier people than for poor people. So you may find a greater variance of expenditures at high income levels than at low income levels.

- Heteroskedasticity may arise in the context of a “random coefficients model.
- Suppose for example that a regressor impacts on individuals in a different way

$$Y_i = a + (b_1 + e_i) X_{i1} + u_i$$

$$Y_i = a + b_1 X_{i1} + e_i X_{i1} + u$$

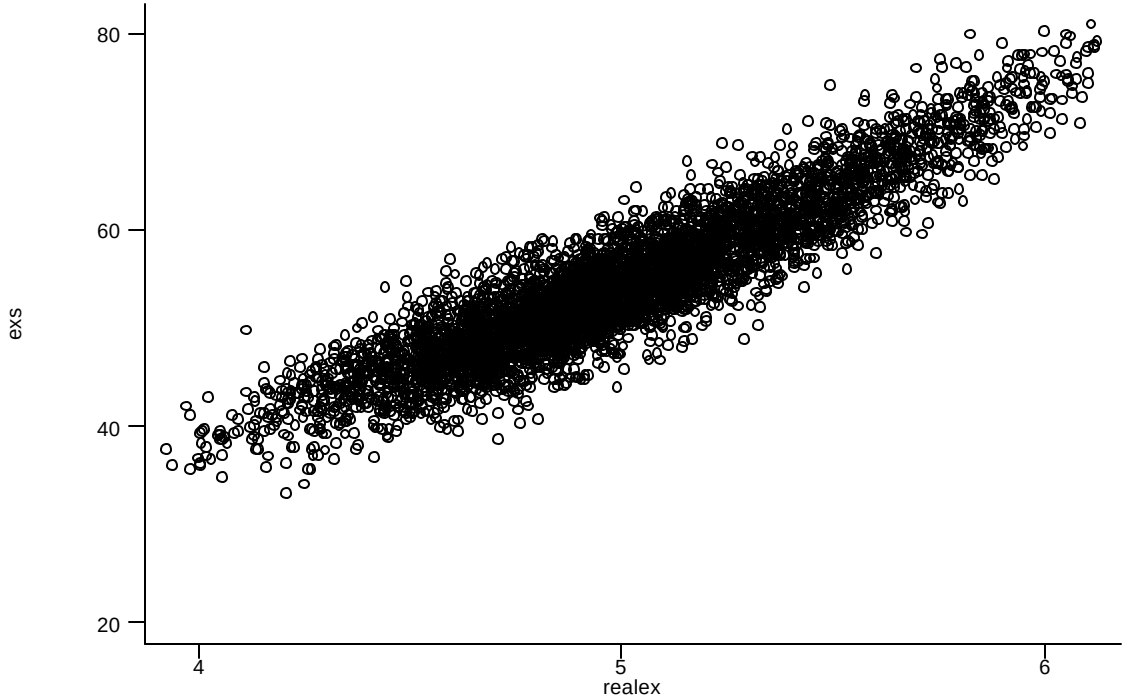
- Assume for simplicity that ε and u are independent.
- Assume that ε and X are independent of each other.
- Then the error term has the following properties:

$$E(\varepsilon_i X_i + u_i \mid X) = E(\varepsilon_i X_i \mid X) + E(u_i \mid X) = E(\varepsilon_i \mid X) X_i = 0$$

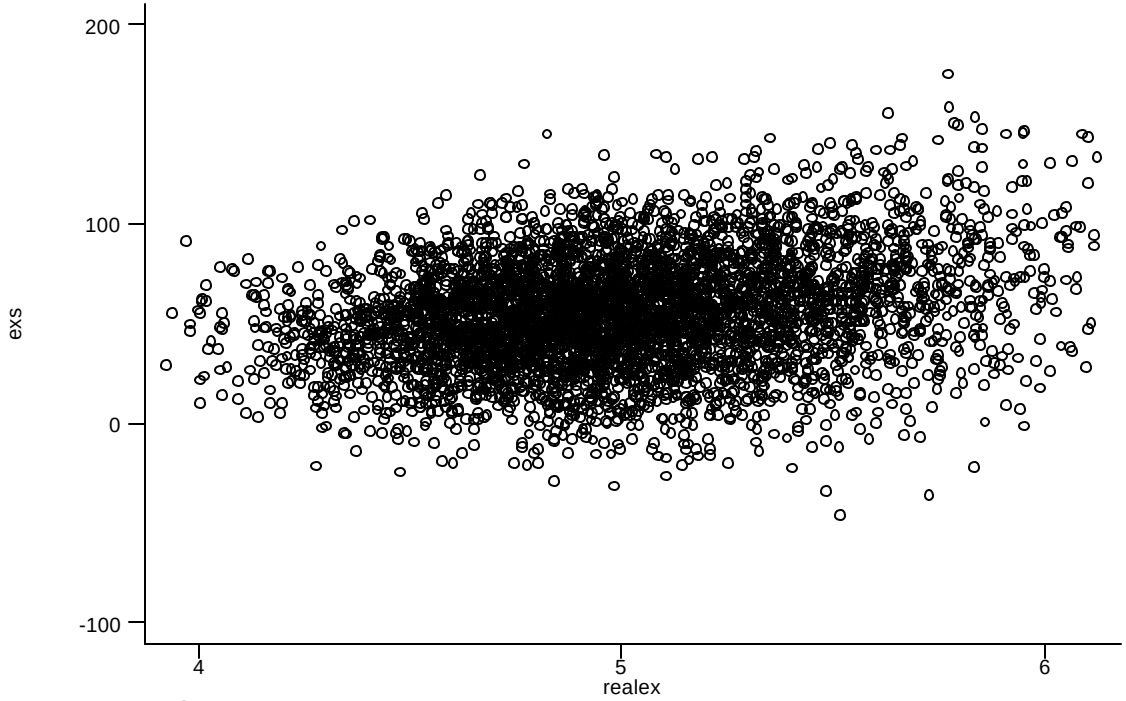
$$Var(\varepsilon_i X_i + u_i \mid X) = Var(\varepsilon_i X_i \mid X) + Var(u_i \mid X) = X_i^2 \sigma_\varepsilon^2 + \sigma_u^2$$

- Where σ_ε^2 is the variance of ε

In both scatter diagrams the (average) slope of the underlying relationship is the same.



Scatter Diagram HOMOSKEDASTICITY



Scatter Diagram RANDOM COEFFICIENTS MODEL HETEROSKEDASTICITY

Implications of Heteroskedasticity

- Assuming all other assumptions are in place, the assumption guaranteeing unbiasedness of OLS **is not violated**. Consequently **OLS is unbiased** in this model

- However the assumptions required to prove that OLS is efficient are violated. Hence **OLS is not BLUE** in this context

- We can devise an efficient estimator by reweighing the data appropriately to take into account of heteroskedasticity

- If there is heteroskedasticity in our data and we ignore it then the **standard errors of our estimates will be incorrect**

- However, if all the other assumptions hold our **estimates will still be unbiased.**

- Since the standard errors are incorrect **inference may be misleading**

Correcting the Standard errors for Heteroskedasticity of unknown kind - The Eicker-White procedure

- If we suspect heteroskedasticity but we do not know its precise form we can still compute our standard errors in such a way that the are **robust to the presence of heteroskedasticity**

- This means that they will be correct whether we have heteroskedasticity or not.

- The procedure is justified for large samples.

. replace exs = 1 + (10+5*invnorm(uniform()))*rr + 3*invnorm(uniform())

(4785 real changes made)

$$Y_i = 1 + (10 + v) * X + u$$

. regr exs rr, robust

Regression with robust standard errors Number of obs = 4785

F(1, 4783) = 295.96

Prob > F = 0.0000

R-squared = 0.0679

Root MSE = 26.933

exs	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
rr	10.06355	.5849706	17.20	0.000	8.916737	11.21036
_cons	1.262277	3.063608	0.41	0.680	-4.743805	7.268359

. replace exs = 1 + (10+0*invnorm(uniform()))*rr + 3*invnorm(uniform())

(4785 real changes made)

$$Y_i = 1 + (10 + v) * X + u$$

. regr exs rr

```
      Source |       SS       df       MS              Number of obs =    4785
-------------+------------------------------           F(  1,  4783) =27346.97
       Model |  250067.192     1  250067.192           Prob > F      =  0.0000
    Residual |   43736.894  4783  9.14423876           R-squared     =  0.8511
-------------+------------------------------           Adj R-squared =  0.8511
       Total |  293804.086  4784  61.4138976           Root MSE      =  3.0239

------------------------------------------------------------------------------
         exs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          rr |   10.00641   .0605095   165.37   0.000     9.887787    10.12504
       _cons |   .8871864   .3266196     2.72   0.007     .2468618    1.527511
------------------------------------------------------------------------------
```

- To see how we can do this lets go back to the derivation of the variance for the estimator of the slope coefficient in the simple two variable regression model (**lecture 3)**
- We had that

$$E[(\hat{b} - b)^2 \mid X] =$$

$$\frac{1}{\left[\sum_{i=1}^{N}(X_i - \overline{X})^2\right]^2}\left(\left\{\sum_{j=1}^{N}\sum_{i=1}^{N}(X_i - \overline{X})(X_j - \overline{X})E[(u_i - \overline{u})(u_j - \overline{u}) \mid X]\right\}\right) =$$

$$\frac{1}{\left[\sum_{i=1}^{N}(X_i - \overline{X})^2\right]^2}\left(\left\{\sum_{i=1}^{N}(X_i - \overline{X})^2 E[(u_i - \overline{u})^2 \mid X]\right\}\right)$$

- The problem arises because $E[(u_i - \bar{u})^2 \mid X]$ is no longer a constant (s^2).

- The variance of the residual changes from observation to observation. Hence in general we can write $E[(u_i - \bar{u})^2 \mid X] = s_i^2$

- We gave an example in the random coefficients model how this can arise. In that case the variance depended on X_i

The Variance of the slope coefficient estimated by OLS when there is heteroskedasticity

$$E[(\hat{b}-b)^2 \mid X] =$$

$$\frac{1}{\left[\sum_{i=1}^{N}(X_i-\overline{X})^2\right]^2}\left(\left\{\sum_{i=1}^{N}(X_i-\overline{X})^2 s_i^2]\right\}\right)$$

The Eicker-White formula

- To estimate this variance we can replace the s_i^2 for each observation by the squared OLS residual for that observation

$$\hat{u}_i^2 = Y_i - \hat{a} - \hat{b}X_i$$

- Thus we estimate the variance of the slope coefficient by using

$$\hat{Var}(\hat{b}) = \frac{\left(\left\{ \sum_{i=1}^{N} (X_i - \overline{X})^2 \hat{u}_i^2] \right\} \right)}{\left[NVar(X) \right]^2}$$

Summary of steps for estimating the variance of the slope coefficients in a way that is robust to the presence of Heteroskedasticity

- Estimate regression model by OLS.
- Obtain residuals.
- Use residuals in formula of previous page.
- A similar procedure can be adapted for the multiple regression model.

Serial Correlation or Autocorrelation

- We have assumed that the errors across observations are not correlated: **Assumption 3**
- We now consider relaxing this assumption in a specific context: With data aver time
- Suppose we have time series data: I.e. we observe *(Y,X)* sequentially in regular intervals over time. (GDP, interest rates, Money Supply etc.).
- We use *t* as a subscript to emphaisize that the observationas are over time only.

The model

- Consider the regression $Y_t = a + bX_t + u_t$

- When we have serial correlation the errors are correlated over time.
- For example a large negative shock to GDP in one period may signal a negative shock in the next period.
- One way to capture this is to use an **Autoregressive model** for the residuals, i.e.

$$u_t = r\,u_{t-1} + v_t$$

- In this formulation the error this period depends on the error in the last period and on an **innovation** v_t.
- v_t is assumed to satisfy all the classical assumptions Assumption 1 to Assumption 3.

- We consider the **stationary autoregressive** case only in which the effect of a shock eventually dies out. This will happens if

$$-1 < r < 1$$

- To see this substitute out one period back to get

$$u_t = r^2 u_{t-2} + r v_{t-1} + v_t$$

- And so on to get

$$u_t = r^k u_{t-k} + v_t + r v_{t-1} + r^2 v_{t-2} + r^3 v_{t-3} + \ldots + r^{k-1} v_{t-(k-1)}$$

- Thus a shock that occurs n periods back has an impact of r^n

Implications of serial correlation

- Under serial correlation of the stationary type OLS is unbiased if the other assumptions are still valid (In particular Assumption 1)
- OLS is no longer efficient (Conditions for the Gauss Markov theorem are violated).
- If we ignore the presence of serial correlation and we estimate the model using OLS, the variance of our estimator will be incorrect and inference will not be valid.

Estimating with serial correlation

- Define a **lag** of a variable to be its past value. Thus *Xt-1* denotes the value of *X* one period ago. The period may be a year, or a month or whatever is the interval of sampling (day or minute in some financial applications)
- Write:

$$Y_t = a + bX_t + u_t$$

$$r\,Y_t = r\,a + r\,bX_t + r\,u_t$$

- Subtract the second from the first to get

$$Y_t - rY_{t-1} = (a - ra) + bX_t - rbX_{t-1} + (u_t - ru_{t-1})$$

$$Y_t - rY_{t-1} = (a - ra) + b(X_t - rX_{t-1}) + v_t$$

- Now suppose we knew r
- Then we could construct the variables

$$Y_t - rY_{t-1} \; and \; (X_t - rX_{t-1})$$

- Then the regression with these transformed variables satisfies the Assumptions 1-4.

- Thus, according to the Gauss Markov theorem if we estimate *b* with these variables we will get an efficient estimator.

- This procedure is called **Generalised Least Squares (GLS)**.
- However we cannot implement it directly because we do not know r

A two step procedure for estimating the regression function when we have Autocorrelation

- Step 1: Regress *Yt* on *Yt-1*, *Xt* and *Xt-1*. The coefficient of *Yt-1* will be an estimate of r
- Construct $Y_t - \hat{r}Y_{t-1}$ *and* $(X_t - \hat{r}X_{t-1})$

- Step 2. Run the Regression using OLS to obtain *b:*

$$Y_t - \hat{r}Y_{t-1} = a^* + b(X_t - \hat{r}X_{t-1}) + v_t$$

- This procedure is called **Feasible GLS**

Summary

- When we know r GLS is BLUE

- When r has to be estimated in a first step then **feasible GLS** is efficient in large samples only.
- In fact in small samples feasible GLS will be generally biased. However in practice it works well with reasonably sized samples.

EXAMPLE: Estimating the AR coefficient in the error term (rho) and transforming the model to take into account of serial correlation.

regr lbp lpbr lpsmr lryae lag* **Log Butter Purchases Monthly data**

Source | SS df MS **Number of obs = 50 one observation lost by lagging**

---

log butter purchases lbp | Coef. Std. Err. t P>|t| [95% Conf. Interval]

---+---

Log price of butter	lpbr	-.6269146	.2779184	-2.26	0.029	-1.187777	-.0660527
Log price of margarine	lpsmr	-.2295241	.5718655	-0.40	0.690	-1.383595	.9245473
Log real income	lryae	.8492604	.4972586	1.71	0.095	-.154248	1.852769
One month Lag of the above							
	laglpbr	.4854572	.271398	1.79	0.081	-.062246	1.033161
	laglpsmr	.6630088	.5499639	1.21	0.235	-.4468633	1.772881
	laglryae	-.7295632	.5246634	-1.39	0.172	-1.788377	.3292504
Lag of dependent variable:							
Estimate of rho	**laglbp**	**.6138367**	**.1160545**	5.29	0.000	.3796292	.8480441
	_cons	2.815675	.8810168	3.20	0.003	1.037711	4.593639

. regr lbprho lpbrrho lpsmrrho lryaerho

```
      Source |       SS       df       MS                      Number of obs =      50
-------------+------------------------------                   F(  3,    46) =    4.72
       Model |  .051787788     3  .017262596                   Prob > F      =  0.0059
    Residual |  .168231703    46  .003657211                   R-squared     =  0.2354
-------------+------------------------------                   Adj R-squared =  0.1855
       Total |  .220019492    49  .004490194                   Root MSE      =  .06047
```

All variables now have been constructed as X(t)-0.61X(t-1)

```
------------------------------------------------------------------------------
      lbprho |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lpbrrho |   -.724766   .2255923    -3.21   0.002     -1.17886   -.2706722
    lpsmrrho |   .4980802    .396111     1.26   0.215    -.2992498     1.29541
    lryaerho |   .8608964   .4937037     1.74   0.088    -.1328776     1.85467
       _cons |   2.026532   .3107121     6.52   0.000     1.401101    2.651963
------------------------------------------------------------------------------
```