Heteroskedasticity is the violation of **Assumption 5** (the error term has a constant variance).

Pure Heteroskedasticity

Tends to be seen in cross-sectional data more than time series data. Tends to be seen when there is a lot of variation in the dependent variable.

Heteroskedasticity that is a function of the error term of a correctly specified regression equation.

Assumption 5 is the assumption of homoskedasticity:

$$Var(\varepsilon_i) = \sigma^2, \qquad i = 1,2,\dots,n$$

If this assumption holds, the error term observations are all being drawn from the same distribution (with mean zero and variance σ^2).

If this assumption is not satisfied we have heteroskedasticity:

$$Var(\varepsilon_i) = \sigma_i^2, \qquad i = 1,2,\dots,n$$

There are many ways to specify the $Var(\varepsilon_i)$.

Most common form of heteroskedasticity is where the variance of the error term is related to an exogenous variable Z_i:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$Var(\varepsilon_i) = \sigma^2 Z_i^2$$

Z_i may or may not be in the regression equation as an independent variable. It is usually a measure of the observation's size. Z_i is called a **proportionality factor**.

Impure Heteroskedasticity

This type of heteroskedasticity is caused by a specification error such as an omitted variable.

The Consequences of Heteroskedasticity

1. Pure heteroskedasticity does not cause bias in the regression coefficient estimates.

2. Heteroskedasticity causes OLS to no longer be a minimum variance estimator.

3. Heteroskedasticity causes the estimated variances of the regression coefficients to be biased, leading to unreliable hypothesis testing. The *t*-statistics will actually appear to be more significant than they really are.

Testing for Heteroskedasticity

Plotting the residuals is always a good first step.

The Park Test

Consider the regression equation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

and suppose we believe

$$Var(\varepsilon_i) = \sigma^2 Z_i^2.$$

Step 1: Compute the residuals from the OLS estimation of

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Step 2: Estimate the auxiliary regression

$$\ln\left(e_i^2\right) = \alpha_0 + \alpha_1 \ln(Z_i) + u_i$$

Step 3: Test $H_0: \alpha_1 = 0$ against $H_A: \alpha_1 \neq 0$ using a t-test

Problem with this test: might not be able to identify Z

The White Test

Most useful test. Consider the regression equation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

This test does not assume a particular form for the heteroskedasticity.

Step 1: Compute the residuals from the OLS estimation of

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Step 2: Estimate the auxiliary regression

$$e_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{1i}^2 + \alpha_4 X_{2i}^2 + \alpha_5 X_{1i} X_{2i} + u_i$$

Include all the explanatory variables, their squares and their cross-products.

Step 3: Test the overall significance of this equation using the test statistic nR^2 which follows a chi-square distribution with degrees of freedom equal to the number of explanatory variables in the auxiliary regression. The n is the sample size and the R^2 is the R^2 from the auxiliary regression. Table B-8 gives critical values for the chi-square distribution. If the value of your test statistic is greater than the critical value, you reject the null hypothesis.

Remedies for Heteroskedasticity

As always, make sure there is no obvious specification error.

1. Weighted Least Squares

Consider the regression equation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

and suppose we have tested and found support for

$$Var(\varepsilon_i) = \sigma^2 Z_i^2.$$

We need to transform this equation with heteroskedasticity to one that is homoskedastic.

The regression equation can be re-written as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + Z_i u_i$$

where $Var(u_i) = \sigma^2$.

If we transform the equation by dividing both sides by Z_i we obtain a new regression equation that is homoskedastic:

$$\frac{Y_i}{Z_i} = \beta_0 \frac{1}{Z_i} + \beta_1 \frac{X_{1i}}{Z_i} + \beta_2 \frac{X_{2i}}{Z_i} + u_i$$

OLS is now BLUE.

2. Heteroskedasticity-Corrected Standard Errors

Adjust the standard errors of the estimated regression coefficients but not the estimates themselves since they are still unbiased. These standard errors are called **White Heteroskedasticity-Consistent Standard Errors**.

3. Redefine the variables

Switching from a linear model to a double-log model might do it.

Example

Imagine a rock band hires us to evaluate their revenues from going on tour. Let's suppose we collect data for the band's most recent tour in 50 US states.

We set up the following regression model

$$\text{revenues}_i = \beta_0 + \beta_1 \text{advertising}_i + \beta_2 \text{stadium}_i + \beta_3 \text{CD}_i + \beta_4 \text{radio}_i + \beta_5 \text{weekend}_i + \varepsilon_t$$

The ticket price is always the same so it is not included in the model.

REVENUES: revenue from each concert in dollars
ADVERTISING: advertising expenditures for each concert in dollars
STADIUM: maximum capacity of each stadium for each concert
CD = number of cd's sold in concert area six months prior to show
RADIO = index of how often the rock band's songs were played on the radio in each concert area (this variable ranges from 1 (rarely) to 5 (all the time))
WEEKEND = 1 if concert is held on a Friday or Saturday night, 0 otherwise

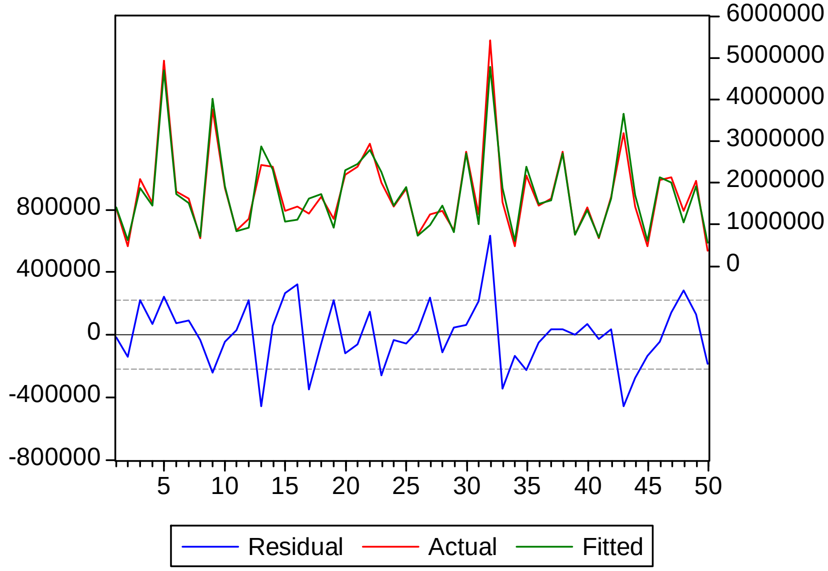
OLS regression:

Dependent Variable: REVENUES
Method: Least Squares
Sample: 1 50
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ADVERTISING	3.147334	1.328637	2.368843	0.0223
STADIUM	34.66051	7.888484	4.393811	0.0001
CD	8.299202	6.049464	1.371891	0.1771
RADIO	300425.7	70633.17	4.253323	0.0001
WEEKEND	356003.5	84215.38	4.227298	0.0001
C	73215.34	70909.63	1.032516	0.3075

R-squared	0.958248	Mean dependent var	1753187.
Adjusted R-squared	0.953504	S.D. dependent var	1018119.
S.E. of regression	219536.3	Akaike info criterion	27.54859
Sum squared resid	2.12E+12	Schwarz criterion	27.77803
Log likelihood	-682.7147	F-statistic	201.9707
Durbin-Watson stat	1.930626	Prob(F-statistic)	0.000000

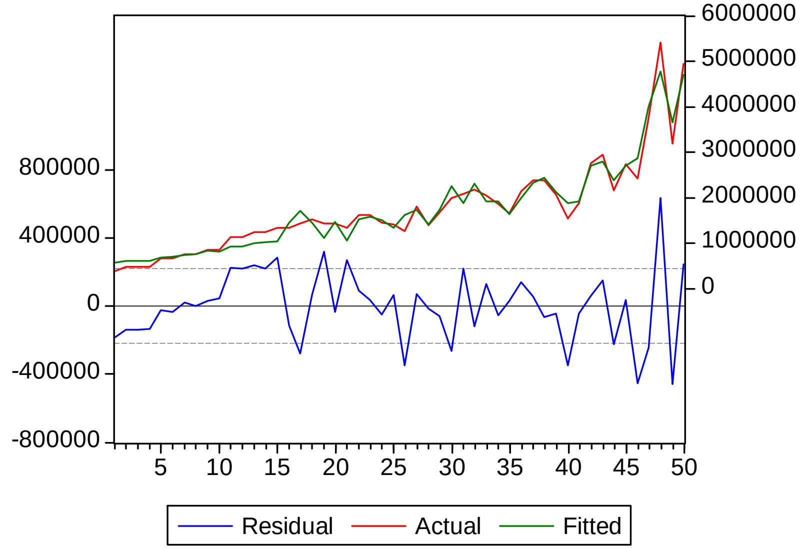
Plot of residuals against the order the observations were recorded:



Using this graph, heteroskedasticity does not appear to be a problem. But thinking about this problem more carefully, you realized that the 50 concert states vary significantly in terms of size and that this may cause the error term variance to be proportional to each state's population (i.e. this is the Z).

So....you go ahead and attach each state's population to the data set.

Let's now order the data from low population states to high population states and plot the same residuals again



There may be heteroskedasticity but not clear-cut.

Need to formally test.

Park Test

Dependent Variable: LOG(RESID^2)
Method: Least Squares
Sample: 1 50
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOG(POPULATION)	0.641170	0.310189	2.067030	0.0441
C	13.35592	4.672219	2.858582	0.0063

R-squared	0.081737	Mean dependent var	22.99185
Adjusted R-squared	0.062607	S.D. dependent var	2.285309
S.E. of regression	2.212615	Akaike info criterion	4.465406
Sum squared resid	234.9920	Schwarz criterion	4.541886
Log likelihood	-109.6351	F-statistic	4.272613
Durbin-Watson stat	1.396421	Prob(F-statistic)	0.044147

Weighted Least Squares

Dependent Variable: REVENUES
Method: Least Squares
Sample: 1 50
Included observations: 50
Weighting series: POPULATION

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ADVERTISING	4.945554	1.060738	4.662372	0.0000
STADIUM	24.57760	10.34730	2.375267	0.0220
CD	8.378595	5.582973	1.500741	0.1406
RADIO	393140.5	121455.4	3.236911	0.0023
WEEKEND	695840.4	146630.9	4.745524	0.0000
C	-262544.2	170660.9	-1.538397	0.1311

Weighted Statistics

R-squared	0.994305	Mean dependent var	2741657.
Adjusted R-squared	0.993658	S.D. dependent var	5163024.
S.E. of regression	411157.7	Akaike info criterion	28.80351
Sum squared resid	7.44E+12	Schwarz criterion	29.03295
Log likelihood	-714.0877	F-statistic	220.7719
Durbin-Watson stat	2.899802	Prob(F-statistic)	0.000000

Unweighted Statistics

R-squared	0.906210	Mean dependent var	1753187.
Adjusted R-squared	0.895552	S.D. dependent var	1018119.
S.E. of regression	329039.7	Sum squared resid	4.76E+12
Durbin-Watson stat	1.343793		

The White Test

White Heteroskedasticity Test:

F-statistic	6.036562	Probability	0.000007
Obs*R-squared	39.63334	Probability	0.003655

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 08/11/09 Time: 21:13
Sample: 1 50
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.89E+11	4.95E+10	3.820635	0.0006
ADVERTISING	5149460.	2482203.	2.074552	0.0467
ADVERTISING^2	35.36982	27.99974	1.263219	0.2162
ADVERTISING*STADIUM	-599.8550	276.1516	-2.172195	0.0379
ADVERTISING*CD	-178.4319	177.3528	-1.006084	0.3224
ADVERTISING*RADIO	2284099.	1453695.	1.571237	0.1266
ADVERTISING*WEEKEND	2436290.	1576309.	1.545566	0.1327
STADIUM	-15030148	9079681.	-1.655361	0.1083
STADIUM^2	1580.498	680.0162	2.324206	0.0271
STADIUM*CD	1731.925	904.7144	1.914333	0.0652
STADIUM*RADIO	-12192611	8764398.	-1.391152	0.1744
STADIUM*WEEKEND	-21643059	9259410.	-2.337412	0.0263
CD	10749212	6681980.	1.608687	0.1182
CD^2	274.1774	293.7963	0.933223	0.3582
CD*RADIO	-19563691	7208794.	-2.713865	0.0109
CD*WEEKEND	-1179236.	5923193.	-0.199088	0.8435
RADIO	-2.95E+11	9.51E+10	-3.105109	0.0041
RADIO^2	1.42E+11	5.27E+10	2.690739	0.0115
RADIO*WEEKEND	1.06E+11	4.94E+10	2.135779	0.0410
WEEKEND	5.14E+10	6.91E+10	0.743333	0.4631

Adjusted R-squared	0.661356	S.D. dependent var	7.12E+10
S.E. of regression	4.14E+10	Akaike info criterion	52.02069
Log likelihood	-1280.517	F-statistic	6.036562
Durbin-Watson stat	2.122886	Prob(F-statistic)	0.000007

White Heteroskedasticity-Consistent Standard Errors

Dependent Variable: REVENUES
Method: Least Squares
Sample: 1 50
Included observations: 50
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ADVERTISING	3.147334	1.948887	1.614939	0.1135
STADIUM	34.66051	10.02266	3.458216	0.0012
CD	8.299202	8.284117	1.001821	0.3219
RADIO	300425.7	80135.47	3.748973	0.0005
WEEKEND	356003.5	78030.18	4.562382	0.0000
C	73215.34	95545.02	0.766291	0.4476

R-squared	0.958248	Mean dependent var	1753187.
Adjusted R-squared	0.953504	S.D. dependent var	1018119.
S.E. of regression	219536.3	Akaike info criterion	27.54859
Sum squared resid	2.12E+12	Schwarz criterion	27.77803
Log likelihood	-682.7147	F-statistic	201.9707
Durbin-Watson stat	2.233807	Prob(F-statistic)	0.000000

Redefining Variables