Consider a simple regression

\[ y = \beta_0 + \beta_1 x + u \]  

(1)

Under the assumption of homoscedasticity \( var(u|x) = \sigma^2 \) there are three implications:

(A) \( var(y|x) = \sigma^2 \)

(B) the variance of \( \hat{\beta}_1 \) is \( var(\hat{\beta}_1|x) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \). By default, stata command `reg y x` uses this formula to find se, t, etc.

(C) OLS is the best linear unbiased estimator (BLUE), a result called Gauss-Markov Theorem (covered in eco411).

Simply put, (A) implies that the variance of \( y \) remains constant across observations. It’s easy to think of a counterexample. Let \( y \) be consumption, and \( x \) be income. We believe the variation of \( y \) (measured by variance) can increase as \( x \) increases. In other words, there may be heteroskedasticity defined as

\[ var(u|x) = h(x_i) \neq constant \]  

(2)

So heteroskedasticity implies that the variance is function of \( x \), so varies from observation to observation.

(B) and (C) do not hold when heteroskedasticity is present

The fact that (B) fails in the presence of heteroskedasticity suggests we need a new formula for variance of \( \hat{\beta}_1 \). That new formula is [8.3] in the textbook:

\[ var_{\text{heteroskedasticity-robust}}(\hat{\beta}_1|x) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\hat{u}_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\hat{u}_i^2}{SST_x} \]  

(3)

Formula (3) is called heteroskedasticity-robust variance, and is valid no matter homoskedasticity is true or false. The stata command

`reg y x, r`

uses formula (3) to find the heteroskedasticity-robust standard error, t value, p value and confidence interval

Use command `reg y x, r` as long as the sample is large

The heteroskedasticity can be detected using either informal method or formal test. Informal method can be plotting the squared residual against \( x \). No pattern means homoskedasticity. Another informal method is summarizing squared residual by \( x \).

Exercise 1: can you think of other informal method?

Exercise 2: is the OLS estimate biased if heteroskedasticity is present?
One formal test for heteroskedasticity is Breusch-Pagan (BP) test, given by equation [8.16] in the textbook

\[ BP\ Test = nR^2_{\bar{u}_i} \]  

(4)

where \( R^2_{\bar{u}_i} \) denotes the R-squared of regressing squared residual onto the regressor. Read page 277 (5th edition).


The fact that (C) fails in the presence of heteroskedasticity suggests we may use a better (more efficient, with smaller variance) estimator, called \textit{generalized least squares} (GLS) estimator (or weighted least squares WLS in this context). We can show

\[ \text{var} \left( \frac{u_i}{\sqrt{h(x_i)}} \right) = \text{constant} \]  

(5)

Equation (5) shows the transformed or \textit{weighted} error term \( \frac{u_i}{\sqrt{h(x_i)}} \) is homoscedastic. So the idea of GLS is simple: we need to divide y and x by the square root of \( h(x_i) \), or equivalently, we need to weight y and x using the \textit{inverse} of \( h(x_i) \). Then apply the OLS using the transformed or weighted data. The result is GLS estimator. Since the weighted error term is homoscedastic,

\textit{GLS is BLUE when heteroskedasticity is present}

In practice, we need to estimate \( h(x_i) \). The so-called \textit{feasible} GLS (FGLS) estimator can be obtained in three-step procedure:

Step 1: regress the log squared residual on x, and save the fitted value, called ghat.

Step 2: get the exponential of ghat, called hhat

Step 3, let the weight (we) be the inverse of hhat, and use command

\[ \text{reg } y \times [w = \text{we}] \]

The point of the log-exponential operation is to ensure we get \textit{non-negative} estimate of \( h(x_i) \), which is variance

Exercise 4: what if in step 1 we regress residual on x? What is ghat? Can we take log of residual?
Example: use Smoke data and estimate the demand for cigarettes

.* OLS, example 8.7, equation (8.35) in textbook

.* reg cigs lincome lcigpric educ age agesq restaurn

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 807</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>8003.02506</td>
<td>6</td>
<td>1333.83751</td>
<td>F( 6, 800) = 7.42</td>
</tr>
<tr>
<td>Residual</td>
<td>143750.658</td>
<td>800</td>
<td>179.688322</td>
<td>R-squared = 0.0527</td>
</tr>
<tr>
<td>Total</td>
<td>151753.683</td>
<td>806</td>
<td>188.280003</td>
<td>Adj R-squared = 0.0456</td>
</tr>
</tbody>
</table>

| cigs | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|------|-------|-----------|---|------|---------------------|
| lincome | .8802682 | .7277832 | 1.21 | 0.227 | -.548322 to 2.308858 |
| lcigpric | -.7508586 | .5773363 | -.13 | 0.907 | -12.08355 to 10.58183 |
| educ | -.5014982 | .1670772 | -3.00 | 0.003 | -.8294597 to -.1735368 |
| age | .7706936 | .1601223 | 4.81 | 0.000 | .456384 to 1.085003 |
| agesq | -.0090228 | .001743 | -5.18 | 0.000 | -.0124443 to -.0056013 |
| restaurn | -.2825085 | .111794 | -2.54 | 0.011 | -.5007462 to -.6427078 |
| _cons | -3.639841 | .2407866 | -1.5 | 0.128 | -.50.90466 to 43.62497 |

Note: the standard error, t value, p-value and confidence interval are all WRONG if heteroskedasticity is present!

First signal for heteroskedasticity: the variance changes across restaurn:

.* by restaurn: sum uhatsq

-> restaurn = 0

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>uhatsq</td>
<td>608</td>
<td>194.081</td>
<td>385.2502</td>
<td>.0044481</td>
<td>4930.936</td>
</tr>
</tbody>
</table>

-> restaurn = 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>uhatsq</td>
<td>199</td>
<td>129.394</td>
<td>311.7096</td>
<td>.0000874</td>
<td>2718.679</td>
</tr>
</tbody>
</table>

Now consider the formal BP test:

.* reg uhatsq lincome lcigpric educ age agesq restaurn

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 807</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1495266.57</td>
<td>6</td>
<td>249211.0949</td>
<td>F( 6, 800) = 5.55</td>
</tr>
<tr>
<td>Residual</td>
<td>105559907</td>
<td>800</td>
<td>131194.883</td>
<td>R-squared = 0.0400</td>
</tr>
<tr>
<td>Total</td>
<td>109955173</td>
<td>806</td>
<td>134520.81</td>
<td>Adj R-squared = 0.0328</td>
</tr>
</tbody>
</table>

| uhatsq | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|--------|-------|-----------|---|------|---------------------|
| lincome | 24.63848 | 19.7218 | 1.25 | 0.212 | -14.07411 to 63.35107 |
| lcigpric | 60.97655 | 156.4487 | 0.39 | 0.697 | -246.1219 to 368.075 |
| educ | -2.384225 | 4.527535 | -0.53 | 0.599 | -11.27148 to 6.503025 |
| age | 19.41748 | 4.339068 | 4.48 | 0.000 | 10.90018 to 27.93478 |
| agesq | -.2147895 | .0472335 | -4.55 | 0.000 | -.3075058 to -.1220733 |
| restaurn | -1.18137 | .3012789 | -2.36 | 0.018 | -130.3204 to -12.04232 |
| _cons | -636.303 | 652.4946 | -0.98 | 0.330 | -1917.107 to 644.5005 |
Income has significant effect on the demand for cigarettes, price does not.