GLM, heteroskedasticity-consistent estimator

$$\text{Asy. Var}\left[\boldsymbol{b}\right] = \frac{1}{n}\left(\mathsf{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}\mathsf{E}\left[\boldsymbol{x}_i\sigma^2\omega_i\boldsymbol{x}_i'\right]\left(\mathsf{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}.$$

In practice, the two expected values are unobserved: we do not have the information of the entire population. Furthermore, we do not observe σ^2, and we do not know the form of $\boldsymbol{\Omega}$ and hence ω_i. We want to estimate them. But why? The reason is about to get clear.

GLM, heteroskedasticity-consistent estimator

We know that the first expected value

$$\left(\mathsf{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}$$

is equal to

$$\left(\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1},$$

which we can estimate with

$$\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

GLM, heteroskedasticity-consistent estimator

We know that the second expected value

$$\mathrm{E}\left[\boldsymbol{x}_i \sigma^2 \omega_i \boldsymbol{x}_i'\right]$$

is equal to

$$\operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \sigma^2 \omega_i \boldsymbol{x}_i'.$$

How can we estimate it?

GLM, heteroskedasticity-consistent estimator

With certain assumptions on $\boldsymbol{x}_i$, and using the LLN (Greene, Theorems D.4 through D.9),

$$\text{plim}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\sigma^2\omega_i\boldsymbol{x}_i' = \text{plim}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i'.$$

Furthermore, since $\boldsymbol{b}$ is a consistent estimator of $\boldsymbol{\beta}$, e_i $(= y_i - x_i b)$ is a consistent estimator of ε_i. Hence,

$$\text{plim}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i' = \text{plim}\frac{1}{n}\sum_{i=1}^{n}e_i^2\boldsymbol{x}_i\boldsymbol{x}_i'.$$

The last term can be estimated with

$$\frac{1}{n}\sum_{i=1}^{n}e_i^2\boldsymbol{x}_i\boldsymbol{x}_i'.$$

GLM, heteroskedasticity-consistent estimator

These results mean that we can estimate

$$\text{Asy. Var}\left[\boldsymbol{b}\right] = \frac{1}{n}\left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}\mathrm{E}\left[\boldsymbol{x}_i\sigma^2\omega_i\boldsymbol{x}_i'\right]\left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}$$

by

$$\text{Est. Asy. Var}\left[\boldsymbol{b}\right] = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} e_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

Dropping the $\frac{1}{n}$ terms,

$$\text{Est. Asy. Var}\left[\boldsymbol{b}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sum_{i=1}^{n} e_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

GLM, heteroskedasticity-consistent estimator

$$\text{Est. Asy. Var}\left[\boldsymbol{b}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

is called the heteroskedasticity-consistent estimator of the variance of $\boldsymbol{b}$. We said that the t and F statistics are not valid if we use

$$\text{Est. Var}\left[\boldsymbol{b} \mid \boldsymbol{X}\right] = s^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

But they are valid if we use the HCE. They are then called the heteroskedasticity-consistent t and F statistics. HCE is powerful. $\boldsymbol{\Omega}$ is often unknown. HCE does not need to figure out $\boldsymbol{\Omega}$. We can use the HCE to make inference on $\boldsymbol{\beta}$. We only need to keep in mind that the HCE, and the test statistics that make use of the HCE, require a large n. We also do not need that the errors are normal!

GLM, heteroskedasticity-consistent estimator

To calculate the HCE in R or MATLAB, we recast

$$\text{Est. Asy. Var}\left[\boldsymbol{b}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

as

$$\text{Est. Asy. Var}\left[\boldsymbol{b}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}' diag\left(e_1^2, \ldots, e_n^2\right) \boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1},$$

where

$$diag\left(e_1^2, \ldots, e_n^2\right) = \begin{bmatrix} e_1^2 & 0 & \ldots & 0 \\ 0 & e_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & e_n^2 \end{bmatrix}.$$

GLM, heteroskedasticity-consistent estimator

How to interpret the HEC?

GLM, heteroskedasticity-consistent estimator

Start with the estimator of the variance of $\boldsymbol{b}$ under homoskedasticity:

$$
\begin{aligned}
\text{Est. Var}\left[\boldsymbol{b} \mid \boldsymbol{X}\right] &= s^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \\
&= \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \\
&= \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' diag\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{n-K},\ldots,\frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}\right)\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.
\end{aligned}
$$

We can move $\boldsymbol{e}'\boldsymbol{e}$ across the matrices because it is a scalar.

GLM, heteroskedasticity-consistent estimator

Under homoskedasticity:

$$\text{Est. Var}\left[\boldsymbol{b} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}' diag\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}, \ldots, \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}\right) \boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

Across the diagonal, the elements are same!

Under heteroskedasticity:

$$\text{Est. Asy. Var}\left[\boldsymbol{b}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}' diag\left(e_1^2, \ldots, e_n^2\right) \boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1},$$

Across the diagonal, the elements are different! You are accounting for heteroskedasticity!

GLM, heteroskedasticity-consistent estimator

```
. regress wage educ

      Source |       SS           df       MS      Number of obs   =       997
-------------+----------------------------------   F(1, 995)       =    251.46
       Model |  7842.35455         1  7842.35455   Prob > F        =    0.0000
    Residual |  31031.0745       995  31.1870095   R-squared       =    0.2017
-------------+----------------------------------   Adj R-squared   =    0.2009
       Total |   38873.429       996  39.0295472   Root MSE        =    5.5845

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.135645   .0716154    15.86   0.000     .9951106     1.27618
       _cons |  -4.860424   .9679821    -5.02   0.000    -6.759944   -2.960903
------------------------------------------------------------------------------
```

GLM, heteroskedasticity-consistent estimator

```
. regress wage educ, robust

Linear regression                               Number of obs     =        997
                                                F(1, 995)         =     178.66
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2017
                                                Root MSE          =     5.5845
```

wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf. Interval]	
educ	1.135645	.0849627	13.37	0.000	.9689186	1.302372
_cons	-4.860424	1.078429	-4.51	0.000	-6.976681	-2.744167