

Heteroscedasticity

Heteroscedasticity is a violation of an OLS assumption. However, it is a property of a model, not a problem. Heteroscedasticity is common for cross-sectional data.

Impact

OLS is a consistent estimator of even in the presence of heteroscedasticity.

The data set is adopted from Greene (2000). EXPEND is the average monthly credit card expenditure; AGE is age in years + 12ths of a year; INCOME is divided by 10,000; RENT is a dummy variable of the own rent.

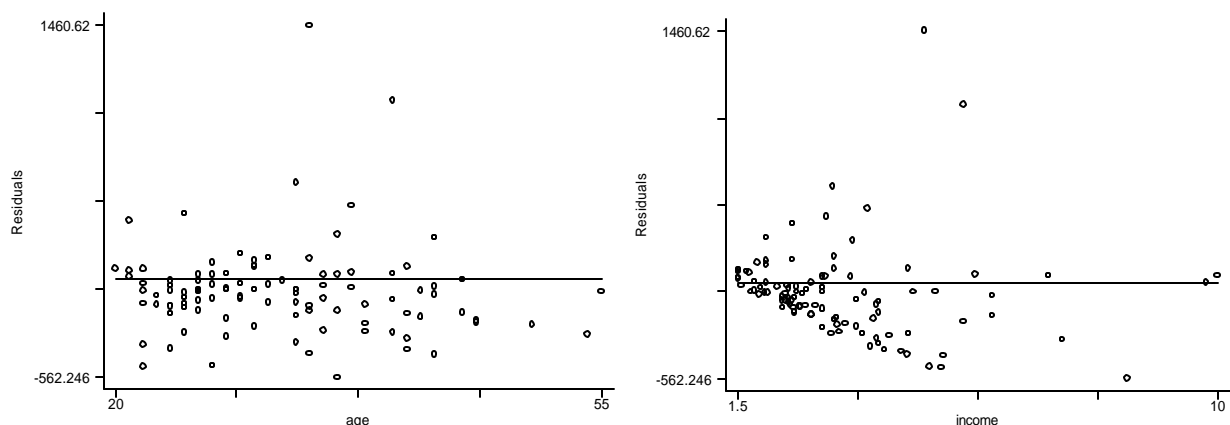
```
. regress expend age rent income inc_sq if expend~=0
```

| Source | SS | df | MS | | | |
|----------|------------|----|------------|-----------------|--------|--|
| Model | 1749357.01 | 4 | 437339.252 | Number of obs = | 72 | |
| Residual | 5432562.03 | 67 | 81083.0153 | F(4, 67) = | 5.39 | |
| Total | 7181919.03 | 71 | 101153.789 | Prob > F = | 0.0008 | |
| | | | | R-squared = | 0.2436 | |
| | | | | Adj R-squared = | 0.1984 | |
| | | | | Root MSE = | 284.75 | |

| | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|--------|-----------|-----------|-------|-------|----------------------|
| age | -3.081814 | 5.514717 | -0.56 | 0.578 | -14.08923 7.925606 |
| rent | 27.94091 | 82.92232 | 0.34 | 0.737 | -137.5727 193.4546 |
| income | 234.347 | 80.36595 | 2.92 | 0.005 | 73.93593 394.7581 |
| inc_sq | -14.99684 | 7.469337 | -2.01 | 0.049 | -29.9057 -.0879859 |
| _cons | -237.1465 | 199.3517 | -1.19 | 0.238 | -635.0541 160.7611 |

Detecting Heteroscedasticity (Greene 507-511)

There are two ways of detecting heteroscedasticity. One is to draw a plot of residuals and independent variables. The left graph illustrates relatively constant variance over AGE, while the right shows that the variance increase as income increases.



The other way is to conduct statistical tests, such as White's general test, Goldfeld-Quant, test, Breusch-Pagan/Godfrey test, and Glesjer Test.

1. White's General Test (1980)

White's general test does not require any specific assumption about the nature of the heteroscedasticity. It may simply identify some other specification errors (omitting relevant variables or including irrelevant variables). The test cannot detect all possible heteroscedasticity; consequently, it is nonconstructive. If the null hypothesis is rejected, we may run OLS with White's heteroscedasticity consistent estimator.¹

$$H_0: \sigma_i^2 = \sigma^2 \text{ for all } i$$

$$\text{Var}(\mathbf{b}) = \sigma^2 [X'X]^{-1} [X'\Omega X] [X'X]^{-1} \text{ is estimated by } \text{Var}(b) = [X'X]^{-1} \left[\sum e_i^2 x_i x_i' \right] [X'X]^{-1}$$

First, run OLS to get residuals and compute e_i^2 .

Second, find all the possible and unique combinations of $X \otimes X$ (that is, include an intercept, original terms, squared terms, and interactions of the two terms, then remove the duplicates)

Fourth, regress e_i^2 on the result of $X \otimes X$ (an intercept should be included)

Fifth, compute $nR^2 \sim \chi^2(P-1)$, where P is the number of regressors, including an intercept.

```
. predict e, residual
. gen ee=e^2
. regress ee age rent income inc_sq age_sq inc_4sq age_rent age_inc age_inc2 inc_rent
inc2_rent inc_3sq if expend~=0
```

| Source | SS | df | MS | | | |
|----------|------------|----|------------|-----------------|---------|--|
| Model | 1.1055e+12 | 12 | 9.2121e+10 | Number of obs = | 72 | |
| Residual | 4.4492e+12 | 59 | 7.5411e+10 | F(12, 59) = | 1.22 | |
| Total | 5.5547e+12 | 71 | 7.8235e+10 | Prob > F = | 0.2905 | |
| | | | | R-squared = | 0.1990 | |
| | | | | Adj R-squared = | 0.0361 | |
| | | | | Root MSE = | 2.7e+05 | |

| ee | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|-----------|-----------|-----------|-------|-------|----------------------|
| age | 5366.344 | 48893.84 | 0.11 | 0.913 | -92470.01 103202.7 |
| rent | 812033.2 | 991630.4 | 0.82 | 0.416 | -1172215 2796281 |
| income | -2021686 | 1053558 | -1.92 | 0.060 | -4129850 86477.94 |
| inc_sq | 669051.3 | 365666.4 | 1.83 | 0.072 | -62645.53 1400748 |
| age_sq | -424.0867 | 627.4593 | -0.68 | 0.502 | -1679.63 831.4564 |
| inc_4sq | 3762.711 | 2277.356 | 1.65 | 0.104 | -794.2669 8319.69 |
| age_rent | 4661.719 | 14424.6 | 0.32 | 0.748 | -24201.83 33525.27 |
| age_inc | 11499.82 | 15614.26 | 0.74 | 0.464 | -19744.25 42743.89 |
| age_inc2 | -1093.311 | 1568.054 | -0.70 | 0.488 | -4230.98 2044.358 |
| inc_rent | -510190.5 | 469792.7 | -1.09 | 0.282 | -1450244 429862.6 |
| inc2_rent | 51834.86 | 61799.82 | 0.84 | 0.405 | -71826.29 175496 |

¹ PROC REG; MODEL expend = age rent income inc_sq /ACOV; RUN;

```
inc_3sq | -86804.79  51162.56  -1.70  0.095  -189180.8  15571.26
_cons   |  1637379   1290978   1.27  0.210  -945862.3  4220620
```

```
. display "The P value is " chi2tail(12, 14.328)
The P value is .28025503
```

$nR^2 = 72 * .1990 = 14.328 \sim \chi^2(13-1)$. Since 14.328 is less than the critical value 21.03, the null hypothesis is not rejected at the five percent significance level. Note that the p value of 14.328 with $df=12$ is .2803.

2. Goldfeld-Quandt Test (1965)

Goldfeld-Quandt test assumes that the observation can be divided into two groups (the first group with large variances and the second with small variances), then check whether or not disturbance variances of the groups are different systematically. So, we have to identify a variable to be used to separate data. F distribution requires that disturbance variances are normally distributed.

$$H_1: \sigma_i^2 = \sigma^2 x_i^2$$

First, sort the observation based on x_i

Second, separate the observations into two groups so that the first group has a bigger variance.

Third, run OLS separately to estimate $e_1'e_1$ (SSE_1) and $e_2'e_2$ (SSE_2).

Fourth, compute $\frac{e_1'e_1/(n_1 - K)}{e_2'e_2/(n_2 - K)} \sim F(n_1 - K, n_2 - K)$, where K is the number of regressors

including an intercept. Keep in mind that $e_1'e_1$ should be greater than $e_2'e_2$ for the test.

A number of observations in the middle of the sample may be omitted in order to increase the power of the test by highlighting the difference. But no more than a third of the observations should be dropped (Harvey and Phillips 1974), since smaller degree of freedom may diminish the power of the test.

```
. regress expend age rent income inc_sq if expend==0 & size==1
```

| Source | SS | df | MS | Number of obs = | 36 |
|----------|------------|----|------------|-----------------|---------|
| Model | 355333.826 | 4 | 88833.4566 | F(4, 31) = | 0.56 |
| Residual | 4894130.09 | 31 | 157875.164 | Prob > F = | 0.6915 |
| Total | 5249463.92 | 35 | 149984.683 | R-squared = | 0.0677 |
| | | | | Adj R-squared = | -0.0526 |
| | | | | Root MSE = | 397.34 |

| expend | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|--------|-----------|-----------|-------|-------|----------------------|
| age | -1.940401 | 11.89225 | -0.16 | 0.871 | -26.1948 22.314 |
| rent | -52.82816 | 163.0564 | -0.32 | 0.748 | -385.3838 279.7275 |
| income | 250.1355 | 226.814 | 1.10 | 0.279 | -212.4547 712.7257 |
| inc_sq | -16.11413 | 17.82175 | -0.90 | 0.373 | -52.46183 20.23358 |
| _cons | -259.1084 | 637.735 | -0.41 | 0.687 | -1559.777 1041.561 |

```
. regress expend age rent income inc_sq if expend==0 & size ==0
```

| Source | SS | df | MS | Number of obs = | 36 |
|----------|------------|----|------------|-----------------|--------|
| Model | 73163.5275 | 4 | 18290.8819 | F(4, 31) = | 1.74 |
| Residual | 326247.258 | 31 | 10524.1051 | Prob > F = | 0.1668 |
| | | | | R-squared = | 0.1832 |
| | | | | Adj R-squared = | 0.0778 |
| Total | 399410.786 | 35 | 11411.7367 | Root MSE = | 102.59 |

| expend | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|--------|-----------|-----------|-------|-------|----------------------|
| age | -4.137398 | 2.670104 | -1.55 | 0.131 | -9.583111 1.308314 |
| rent | 108.8721 | 45.24256 | 2.41 | 0.022 | 16.59927 201.1449 |
| income | 16.88556 | 475.5545 | 0.04 | 0.972 | -953.0142 986.7853 |
| inc_sq | 3.693402 | 107.1857 | 0.03 | 0.973 | -214.9133 222.3001 |
| _cons | 153.1303 | 501.5151 | 0.31 | 0.762 | -869.7166 1175.977 |

```
. disp "The P value is " Ftail(31, 31, 15.001)
The P value is 1.377e-11
```

$\frac{e_1'e_1/(n_1 - K)}{e_2'e_2/(n_2 - K)} = \frac{4,894,130/(36 - 5)}{326,247/(36 - 5)} = 15.001 \sim F(31,31)$. The statistic is greater than the critical value of the five percent level, 1.822; therefore, the null hypothesis is rejected ($p < .0000$.)

3. Breusch-Pagan/Godfrey Lagrange Multiplier Test (1979)

This test assumes that disturbance variances vary with a set of regressors, not a single regressor. The LM test is known to be sensitive to the assumption of normality.

$H_1: \sigma_i^2 = \sigma^2 f(\mathbf{a}_0 + \mathbf{a}'z_i)$, where z_i is a vector of independent variables (or a subset of X) that are suspected of causing heteroscedasticity. The alternative hypothesis states that heteroscedasticity depends on z_i with specific functional forms. The null hypothesis is $\mathbf{a} = 0$: homoscedasticity.

First, regress y_i on X_i to get $e'e$ (SSE) and residuals.

Second, calculate $g_i = \frac{e_i^2}{e'e/n} - 1 = \frac{e_i^2}{SSE/n} - 1$ ²

Third, regress g on z_i to get SSR.

Fourth, compute $LM = \frac{1}{2} g'Z(Z'Z)^{-1}Z'g = \frac{SSR}{2}$, where Z is a n by $(j+1)$ matrix and j is the number of variables in z_i .

Finally, conclude using $LM \sim \chi^2(j)$. Under the null hypothesis of homoscedasticity, LM is asymptotically distributed as chi-squared with degree of freedom equal to the number of variable in z_i .

² Omitting -1 does not affect the LM test, although it produces different estimators.

```
. gen g=ee/(5432562/72) - 1
. regress g income inc_sq if expend~=0
```

| Source | SS | df | MS | Number of obs = | 72 |
|----------|------------|----|------------|-----------------|--------|
| Model | 83.8406129 | 2 | 41.9203064 | F(2, 69) = | 3.24 |
| Residual | 891.858779 | 69 | 12.9254895 | Prob > F = | 0.0451 |
| Total | 975.699392 | 71 | 13.742245 | R-squared = | 0.0859 |
| | | | | Adj R-squared = | 0.0594 |
| | | | | Root MSE = | 3.5952 |

| g | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|--------|-----------|-----------|-------|-------|----------------------|
| income | 2.261001 | .9534358 | 2.37 | 0.021 | .3589488 4.163054 |
| inc_sq | -.1876192 | .0918291 | -2.04 | 0.045 | -.3708132 -.0044251 |
| _cons | -5.020459 | 2.053893 | -2.44 | 0.017 | -9.117863 -.9230545 |

```
. display "The P value is " chi2tail(2, 41.9203)
The P value is 7.891e-10
```

$LM = \frac{SSR}{2} = \frac{83.4806}{2} = 41.9203 \sim \chi^2(2)$. Since the statistic is greater than the critical value 5.99, the null hypothesis is rejected at the five percent level ($p < .0000$).

4. Glesjer Test (1969)

Glesjer test is a Wald test. This test considers some specific formulations of the disturbance variances. Again z_i is a vector of independent variables (or a subset of X) that are suspected of causing heteroscedascity.

$Var(e_i) = \sigma^2[\mathbf{s}'z_i]$, then regress e_i^2 on z_i

$Var(e_i) = \sigma^2[\mathbf{s}'z_i]^2$, then regress $|e_i|$ on z_i

$Var(e_i) = \sigma^2 \exp[\mathbf{s}'z_i]$, then regress $\log |e_i|$ on z_i

$H_0: \mathbf{b}_j = 0$ for $j=1 \dots J$ (excluding an intercept)

$JF \sim \chi^2(j)$, where j is the number of variable in z_i and F is the F score of OLS for each formulation.

```
. regress ee income inc_sq if expend~=0
```

| Source | SS | df | MS | Number of obs = | 72 |
|----------|------------|----|------------|-----------------|---------|
| Model | 4.7731e+11 | 2 | 2.3865e+11 | F(2, 69) = | 3.24 |
| Residual | 5.0774e+12 | 69 | 7.3585e+10 | Prob > F = | 0.0451 |
| Total | 5.5547e+12 | 71 | 7.8235e+10 | R-squared = | 0.0859 |
| | | | | Adj R-squared = | 0.0594 |
| | | | | Root MSE = | 2.7e+05 |

| ee | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|----|-------|-----------|---|------|----------------------|
| | | | | | |

| | | | | | | |
|--------|-----------|----------|-------|-------|-----------|----------|
| income | 170597.6 | 71938.88 | 2.37 | 0.021 | 27083.5 | 314111.8 |
| inc_sq | -14156.29 | 6928.713 | -2.04 | 0.045 | -27978.69 | -333.883 |
| _cons | -303352.7 | 154970.9 | -1.96 | 0.054 | -612511.1 | 5805.705 |

$2(3.24) = 6.48 \sim \chi^2(2)$. The statistic is greater than the critical value 5.99; so reject the null hypothesis at the five percent level ($p < .0392$).

. regress abs_e income inc_sq if expend~=0

| Source | SS | df | MS | Number of obs = | 72 |
|----------|------------|----|------------|-----------------|--------|
| Model | 582525.302 | 2 | 291262.651 | F(2, 69) = | 6.93 |
| Residual | 2898094.21 | 69 | 42001.3653 | Prob > F = | 0.0018 |
| Total | 3480619.51 | 71 | 49022.81 | R-squared = | 0.1674 |
| | | | | Adj R-squared = | 0.1432 |
| | | | | Root MSE = | 204.94 |

| abs_e | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|--------|-----------|-----------|-------|-------|----------------------|
| income | 196.4062 | 54.35002 | 3.61 | 0.001 | 87.98084 304.8315 |
| inc_sq | -17.0083 | 5.234662 | -3.25 | 0.002 | -27.45117 -6.565437 |
| _cons | -261.0439 | 117.0809 | -2.23 | 0.029 | -494.6139 -27.47391 |

$2(6.93) = 13.86 \sim \chi^2(2)$. Reject the null hypothesis at the five percent level ($p < .0010$).

. regress log_e income inc_sq if expend~=0

| Source | SS | df | MS | Number of obs = | 26 |
|----------|------------|----|------------|-----------------|--------|
| Model | 5.74091017 | 2 | 2.87045509 | F(2, 23) = | 2.34 |
| Residual | 28.1991211 | 23 | 1.22604874 | Prob > F = | 0.1187 |
| Total | 33.9400313 | 25 | 1.35760125 | R-squared = | 0.1691 |
| | | | | Adj R-squared = | 0.0969 |
| | | | | Root MSE = | 1.1073 |

| log_e | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|--------|-----------|-----------|-------|-------|----------------------|
| income | .8644331 | .4129358 | 2.09 | 0.048 | .0102103 1.718656 |
| inc_sq | -.0852633 | .0394105 | -2.16 | 0.041 | -.16679 -.0037365 |
| _cons | 3.085854 | .8636933 | 3.57 | 0.002 | 1.299169 4.87254 |

$2(12.07) = 24.14 \sim \chi^2(2)$. Reject the null hypothesis at the five percent level ($p < .0000$).

5. Groupwise Heteroscedasticity

This test is the Likelihood Ratio test, which uses maximum likelihood estimators of \mathbf{s}^2 and \mathbf{s}_g^2 :
 $s^2 = e'e/n$ and $s_g^2 = e_g'e_g/n_g$. Therefore, normality makes the Likelihood Ratio test powerful.

H_0 : groups have the same variance. $H_1: \mathbf{s}_i^2 = \mathbf{s}^2 x_i^2$ (Goldfeld-Quandt test)

First, run a pooled OLS to get $e'e$ (SSE)

Second, separate groups using a independent variable.

Third, run OLS for separated subsamples to get $e_g'e_g$ (SSE_g).

Finally, compute $LR = -2(\ln L_F - \ln L_G) = n(\ln s^2) - \sum [n_g(\ln s_g^2)] \sim \chi^2(G-1)$, where G is the number of groups.

$$s^2 = e'e/n = 5,432,562/72 = 75,452.25, \ln s^2 = 11.2313$$

$$s_1^2 = e_1'e_1/n_1 = 285,617/28 = 10,200.607, \ln s_1^2 = 9.2302$$

$$s_2^2 = 770,440/24 = 32,101.667, \ln s_2^2 = 10.3767$$

$$s_3^2 = 3,686,290/20 = 184,314.5, \ln s_3^2 = 12.1244$$

$$LR = 72(11.2313) - 28(9.2302) - 24(10.3767) - 20(12.1244) = 58.6756 \sim \chi^2(3-1)$$

The statistic 58.6756 is greater than the critical value 5.99, allowing us to reject the null hypothesis at the five percent significance level ($p < .0000$).

```
. regress expend age rent income inc_sq if expend~=0 & group==1
```

| Source | SS | df | MS | Number of obs = | 28 |
|----------|------------|----|------------|-----------------|--------|
| Model | 69889.6905 | 4 | 17472.4226 | F(4, 23) = | 1.41 |
| Residual | 285617.288 | 23 | 12418.1429 | Prob > F = | 0.2631 |
| | | | | R-squared = | 0.1966 |
| | | | | Adj R-squared = | 0.0569 |
| Total | 355506.978 | 27 | 13166.9251 | Root MSE = | 111.44 |

| expend | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|--------|-----------|-----------|-------|-------|----------------------|
| age | -3.587137 | 4.177876 | -0.86 | 0.399 | -12.22973 5.055458 |
| rent | 109.8732 | 51.19141 | 2.15 | 0.043 | 3.975676 215.7707 |
| income | 441.4945 | 920.5038 | 0.48 | 0.636 | -1462.713 2345.702 |
| inc_sq | -101.0931 | 222.6113 | -0.45 | 0.654 | -561.5996 359.4134 |
| _cons | -277.1786 | 902.3146 | -0.31 | 0.761 | -2143.759 1589.401 |

```
. regress expend age rent income inc_sq if expend~=0 & group==2
```

| Source | SS | df | MS | Number of obs = | 24 |
|----------|------------|----|------------|-----------------|--------|
| Model | 442608.435 | 4 | 110652.109 | F(4, 19) = | 2.73 |
| Residual | 770439.629 | 19 | 40549.4541 | Prob > F = | 0.0599 |
| | | | | R-squared = | 0.3649 |
| | | | | Adj R-squared = | 0.2312 |
| Total | 1213048.06 | 23 | 52741.2201 | Root MSE = | 201.37 |

| expend | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|--------|-----------|-----------|-------|-------|----------------------|
| age | -1.737986 | 6.713201 | -0.26 | 0.799 | -15.78888 12.31291 |
| rent | 183.5318 | 125.2312 | 1.47 | 0.159 | -78.58009 445.6438 |
| income | 3218.369 | 2103.145 | 1.53 | 0.142 | -1183.564 7620.303 |
| inc_sq | -458.8817 | 324.4732 | -1.41 | 0.173 | -1138.012 220.2485 |
| _cons | -5241.037 | 3451.723 | -1.52 | 0.145 | -12465.58 1983.502 |

. regress expend age rent income inc_sq if expend~=0 & group==3

| Source | SS | df | MS | | | |
|----------|------------|----|------------|-----------------|---------|--|
| Model | 550230.536 | 4 | 137557.634 | Number of obs = | 20 | |
| Residual | 3686289.51 | 15 | 245752.634 | F(4, 15) = | 0.56 | |
| Total | 4236520.05 | 19 | 222974.739 | Prob > F = | 0.6954 | |
| | | | | R-squared = | 0.1299 | |
| | | | | Adj R-squared = | -0.1022 | |
| | | | | Root MSE = | 495.73 | |

| expend | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|--------|-----------|-----------|-------|-------|----------------------|----------|
| age | -9.043037 | 20.57187 | -0.44 | 0.667 | -52.89095 | 34.80487 |
| rent | -265.185 | 274.1793 | -0.97 | 0.349 | -849.5842 | 319.2143 |
| income | 580.3217 | 503.2949 | 1.15 | 0.267 | -492.4259 | 1653.069 |
| inc_sq | -37.71664 | 36.07229 | -1.05 | 0.312 | -114.6029 | 39.16963 |
| _cons | -973.8416 | 1572.732 | -0.62 | 0.545 | -4326.04 | 2378.357 |

Correcting Heteroscedasticity

I. Aitken's Generalized Least Squares (GLS)

When \mathbf{s}_i^2 or Ω is known, $\hat{\mathbf{b}}_{GLS} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} Y$ is BLUE. (Aitken's theorem)

$$\Omega = C \Lambda C'; \quad \Omega^{-1} = C \Lambda^{-1} C' = C \Lambda^{-1/2} \Lambda^{-1/2} C';$$

$$\Omega = (P' P)^{-1}, \text{ where } P = C \Lambda^{-1/2} \text{ so that } \Omega^{-1} = P' P$$

In order to estimate GLS, divide every terms including intercept in the original model by \mathbf{s}_i (transformation of dependent and independent variables), then run OLS without an intercept.

$$\Omega = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & w_n \end{bmatrix}, \quad y_* = \begin{bmatrix} y_1/w_1 \\ y_2/w_2 \\ \dots \\ y_n/w_n \end{bmatrix}, \quad \text{and } X_* = \begin{bmatrix} x_{11}/w_1 \dots x_{k1}/w_1 \\ x_{12}/w_1 \dots x_{k1}/w_1 \\ \dots \\ x_{1n}/w_n \dots x_{kn}/w_n \end{bmatrix}$$

If $Var(\mathbf{e}_i) = \mathbf{s}_i^2 = \mathbf{s}^2 x_{ik}^2$, divide every terms by x_k , then run OLS without an intercept. This is the Weighted Least Squares (WLS) method. If \mathbf{s}_i^2 turns out proportional to a particular independent variable, divide every terms by $\sqrt{x_k}$, then run OLS without an intercept. Alternatively, compute $w_i = 1/x_k$, then run original OLS with the w_i weighted.³

Feasible GLS

When \mathbf{s}_i^2 or Ω is known, it should be estimated to get feasible GLS.

³ Although ANOVA table and (adjusted) R squares are different, two methods produce identical estimators.