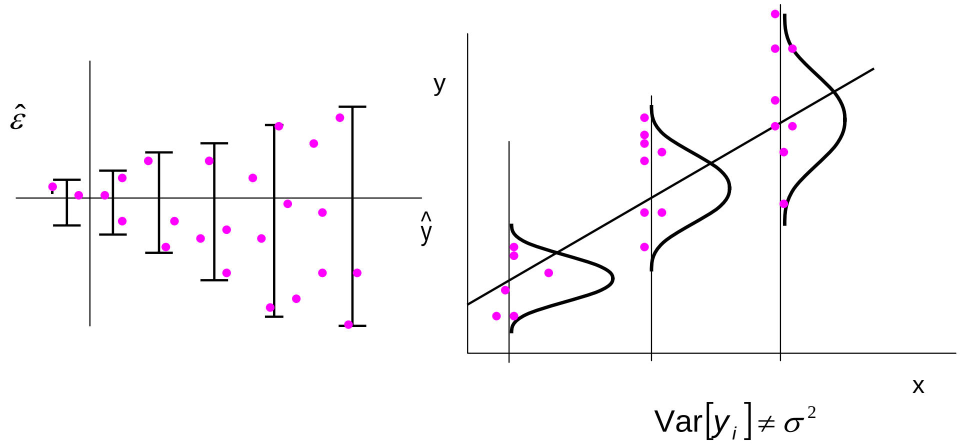
Chapter 2

Generalized Least squares

Contents

1 Introduction

In both ordinary least squares and maximum likelihood approaches to parameter estimation, we made the assumption of constant variance, that is the variance of an observation is the same regardless of the values of the explanatory variables associated with it, and since the explanatory variables determine the mean value of the observation, what we assume is that the variance of the observation is unrelated to the mean.



There are many real situations in which this assumption is inappropriate. In some cases the measurement system used might be a source of variability, and the size of the measurement error is proportional to the measured quantity. Other times this occurs when errors are correlated. Also, when the underlying distribution is continuous, but skewed, such as lognormal, gamma, etc., the variance is not constant, and in many cases variance is a function of the mean.
An important point is that the constant variance is linked to the assumption of normal distribution for the response.

When the assumption of constant variance is not satisfied a possible solution is to transform the data (for example taking *log* of the response variable and/or the explanatory variables) to achieve constant variance. Another approach is based on *generalized or weighted least squares* which is an modification of ordinary least squares which takes into account the inequality of variance in the observations. Weighted least squares play an important role in the parameter estimation for generalized linear models.

2 Generalized and weighted least squares

2.1 Generalized least squares

Now we have the model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad E[\boldsymbol{\varepsilon}] = 0 \quad Var[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{V}$$

where $\boldsymbol{V}$ is a known $n \times n$ matrix. If $\boldsymbol{V}$ is diagonal but with unequal diagonal elements, the observations y are uncorrelated but have unequal variance, while if $\boldsymbol{V}$ has non-zero off-diagonal elements, the observations are correlated.

If we estimate $\boldsymbol{\beta}$ by ordinary least squares, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{y}$, the estimator is not optimal. The solution is to transform the model to a new set of observations that satisfy the constant variance assumption and use least squares to estimate the parameters.
Since $\sigma^2\boldsymbol{V}$ is a covariance matrix, $\boldsymbol{V}$ is a symmetric non-singular matrix, therefore $\boldsymbol{V} = \boldsymbol{K}'\boldsymbol{K} = \boldsymbol{K}\boldsymbol{K}$, and $\boldsymbol{K}$ is called the *squared root* of $\boldsymbol{V}$. We define

$$\boldsymbol{z} = \boldsymbol{K}^{-1}\boldsymbol{y} \quad \boldsymbol{B} = \boldsymbol{K}^{-1}\boldsymbol{X} \quad \boldsymbol{g} = \boldsymbol{K}^{-1}\boldsymbol{\varepsilon} \Rightarrow \boldsymbol{z} = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{g} \tag{2.1}$$

then, using results in section 7 of Review in Matrix Algebra,

$$\begin{aligned} E[\boldsymbol{g}] &= \boldsymbol{K}^{-1}E[\boldsymbol{\varepsilon}] = 0 \\ Var[\boldsymbol{g}] &= Var[\boldsymbol{K}^{-1}\boldsymbol{\varepsilon}] = \boldsymbol{K}^{-1}Var[\boldsymbol{\varepsilon}]\boldsymbol{K}^{-1} = \sigma^2\boldsymbol{K}^{-1}\boldsymbol{V}\boldsymbol{K}^{-1} = \sigma^2\boldsymbol{K}^{-1}\boldsymbol{K}\boldsymbol{K}\boldsymbol{K}^{-1} = \sigma^2\boldsymbol{I}, \end{aligned}$$

and so we are under the assumptions of ordinary least squares. The least squares function is

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\boldsymbol{z} - \boldsymbol{B}\boldsymbol{\beta})'(\boldsymbol{z} - \boldsymbol{B}\boldsymbol{\beta}) = (\boldsymbol{K}^{-1}\boldsymbol{y} - \boldsymbol{K}^{-1}\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{K}^{-1}\boldsymbol{y} - \boldsymbol{K}^{-1}\boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{K}^{-1}\boldsymbol{K}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

Taking the partial derivative with respect to $\boldsymbol{\beta}$ and setting it to 0, we get:

$$(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{y} \quad \text{normal equations}$$

The **generalized least squares estimator** of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \underbrace{(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{V}^{-1}}_{(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'}\boldsymbol{y}$$

and

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{V}^{-1}E[\boldsymbol{y}] = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta} \\ Var[\hat{\boldsymbol{\beta}}] &= \sigma^2(\boldsymbol{B}'\boldsymbol{B})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{K}^{-1}\boldsymbol{K}^{-1}\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \end{aligned}$$

Again, under normal theory, the generalized least squares estimators are the maximum likelihood estimators since the log-likelihood function is:

$$L \propto -\ln(\sigma^2) - \frac{1}{2}\ln|\boldsymbol{V}| - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

The analysis of variance table is:

Source	df	SS	MS	F
Regression	$p-1$	$SS_R = \boldsymbol{y}'\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$	$MSR = SS_R/(p-1)$	$\frac{MSR}{MSE}$
Error	$n-p$	$SS_E = \boldsymbol{y}'\boldsymbol{V}^{-1}\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$	$MSE = SS_E/(n-p)$	
Total	$n-1$	$SS_T = \boldsymbol{y}'\boldsymbol{V}^{-1}\boldsymbol{y}$	$MST = SS_T/(n-1)$	

2.2 Weighted least squares

Some times the errors are uncorrelated, but have unequal variance. In this case we use *weighted least squares*. The covariance matrix of $\boldsymbol{\varepsilon}$ has the form

$$\sigma^2\boldsymbol{V} = \begin{bmatrix} 1/w_1 & & & 0 \\ & 1/w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{bmatrix}$$

Let $\boldsymbol{W} = \boldsymbol{V}^{-1}$, $\boldsymbol{W}$ is also diagonal, with elements w_i. The **weighted least squares estimator** of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$$

Note that observations with large variances get smaller weights than observations with smaller variances.
Examples of possible weights are:

- Error proportional to a predictor $Var(\varepsilon_i) \propto x_i$ suggests $w_i = x_i^{-1}$.
- When an observation y_i is an average of several, n_i, observations at that point of the explanatory variable, then, $Var(y_i) = \sigma^2/n_i$ suggests $w_i = n_i$.

2.3 Iteratively reweighted least squares

Sometimes we will have prior information on the weights w_i, others we might find, looking at residual plots, that the variability is a function of one or more explanatory variables. In these cases we have to estimate the weights, perform the analysis, re-estimate the weights again based on these results and perform the analysis again. This procedure is called *iteratively reweighted keast squares* (IRLS). This method is also applied in generalized linear models as we will see in the next chapter.
To give an example of IRLS, suppose that $Var(\varepsilon_i) = \gamma_0 + \gamma_1 x_1$:

1. Start with $w_i = 1$
2. Use least squares to estimate $\boldsymbol{\beta}$
3. use the residuals to estimate $\boldsymbol{\gamma}$, perhaps by regressing $\hat{\boldsymbol{\varepsilon}}^2$ on $\boldsymbol{x}_1$
4. Recompute the weights and go to 2.

Continue until convergence. More details on the effect of the method on $\hat{\beta}$ can be found in Ruppert and Carroll (1988)

3 Examples

The following examples are taken from Chapter 5 of Faraway (2002)

3.1 Generalized least squares: The Longley data

The original source of the data is Longley (1967). The response variable is the number of people employed, yearly from 1947 to 1962 and the explanatory variables are GNP and Population.
Fit the linear model:

```
> data(longley)
> g<-lm(Employed~GNP+Population,data=longley)
> summary(g,cor=T)

Call:
lm(formula = Employed ~ GNP + Population, data = longley)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80899 -0.33282 -0.02329  0.25895  1.08800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.93880   13.78503   6.452 2.16e-05
GNP          0.06317    0.01065   5.933 4.96e-05
Population  -0.40974    0.15214  -2.693   0.0184

Residual standard error: 0.5459 on 13 degrees of freedom
Multiple R-Squared: 0.9791,     Adjusted R-squared: 0.9758
F-statistic: 303.9 on 2 and 13 DF,  p-value: 1.221e-11

Correlation of Coefficients:
           (Intercept) GNP
GNP         0.98
Population -1.00       -0.99
```

What do you notice?

In data collected over time such as this, errors could be correlated. Assuming that errors take an autoregressive form:

$$\varepsilon_{i+1} = \rho\varepsilon_i + \delta_i \quad \delta_i \sim N(0, \tau^2)$$

We can estimate ρ by means of the sample correlation of residuals:

```
> cor(g$res[-1],g$res[-16])
[1] 0.3104092
```

A model with autoregressive error has covariance matrix $\boldsymbol{V}_{ij} = \rho^{|i-j|}$. Assuming that ρ is know and equal to 0.3104092. Then, $\boldsymbol{V}$ is computed as,

```
> V<-diag(16)
> V<-0.3104092^abs(row(V)-col(V))
```

and the generalized least squares estimate $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{y}$ is,

```
> X<-model.matrix(g)
> V.inv<-solve(V)
> beta<-solve(t(X)%*%V.inv%*%X)%*%t(X)%*%V.inv%*%longley$Empl
> beta

                    [,1]
  (Intercept) 94.89887752
  GNP          0.06738948
  Population  -0.47427391
```

The standard error of $\hat{\boldsymbol{\beta}}$, $\sqrt{Var(\hat{\boldsymbol{\beta}})} = \sqrt{\sigma^2(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}}$ is

```
> res<-longley$Empl-X%*%beta
> sig<-sum(res^2)/g$df
> sqrt(diag(solve(t(X)%*%V.inv%*%X))*sig)
(Intercept)         GNP  Population
14.15760467  0.01086675  0.15572652
```

Another way to fit this model would be to use the `lm()` function but with the model in equation (2.1)

```
> K<-chol(V)
> K.inv<-solve(t(K))
> B<-K.inv%*%X
> z<-K.inv%*%longley$Empl
> lm(z~B-1)$coeff
B(Intercept)         BGNP  BPopulation
 94.89887752   0.06738948  -0.47427391
```

In practice, we do not know the value of ρ, and so we would estimate ρ again from the data

```
 cor(res[-1],res[-16])
[1] 0.3564161
```

and fit the model again, and iterate until convergence.

A third option is to use the `nlme()` library, which contains the `gls()` function. We can use it to fit this model,

```
> library(nlme)
> g<-gls(Employed~GNP+Population,correlation=corAR1(form=~Year),data=longley)
> summary(g)
Generalized least squares fit by REML
  Model: Employed ~ GNP + Population
  Data: longley
       AIC      BIC    logLik
  44.66377 47.48852 -17.33188

Correlation Structure: AR(1)
 Formula: ~Year
 Parameter estimate(s):
      Phi
0.6441692

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 101.85813 14.198932  7.173647  0.0000
GNP           0.07207  0.010606  6.795485  0.0000
Population   -0.54851  0.154130 -3.558778  0.0035

 Correlation:
           (Intr) GNP
GNP         0.943
Population -0.997 -0.966

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.5924564 -0.5447822 -0.1055401  0.3639202  1.3281898

Residual standard error: 0.689207
Degrees of freedom: 16 total; 13 residual
```

The final estimate for ρ is 0.644. However, if we check the confidence intervals

```
intervals(g)
Approximate 95% confidence intervals

 Coefficients:
                  lower         est.       upper
(Intercept) 71.18320440 101.85813280 132.5330612
GNP          0.04915865   0.07207088   0.0949831
Population  -0.88149053  -0.54851350  -0.2155365
attr(,"label")
[1] "Coefficients:"
```

```
Correlation structure:
         lower      est.     upper
Phi -0.4430373 0.6441692 0.9644866
attr(,"label")
[1] "Correlation structure:"


 Residual standard error:
    lower      est.     upper
0.2477984 0.6892069 1.9169062
```

we see that it is not significantly different from 0, and therefore we can ignore it.

3.2 Weighted least squares: The proton data

This example is from an experiment aimed to study the interaction of certain kinds of elementary particles on collision with proton targets. The experiment was designed to test certain theories about the nature of strong interaction. The cross-section (crossx) variable is belived to be linearly related to the inverse of the energy (in the data set this variable appears already inverted). At each level of momentum, a very large number of of observations were taken so that it is possible to accurately estimate the estandard deviation of the response (sd)

```
> library(faraway)
> data(strongx)
> strongx
   momentum energy crossx sd
1         4  0.345    367 17
2         6  0.287    311  9
3         8  0.251    295  9
4        10  0.225    268  7
5        12  0.207    253  7
6        15  0.186    239  6
7        20  0.161    220  6
8        30  0.132    213  6
9        75  0.084    193  5
10      150  0.060    192  5
```

First, we fit the model without weights,

```
> gu<-lm(crossx~energy,data=strongx)
> summary(gu)

Call:
lm(formula = crossx ~ energy, data = strongx)
```

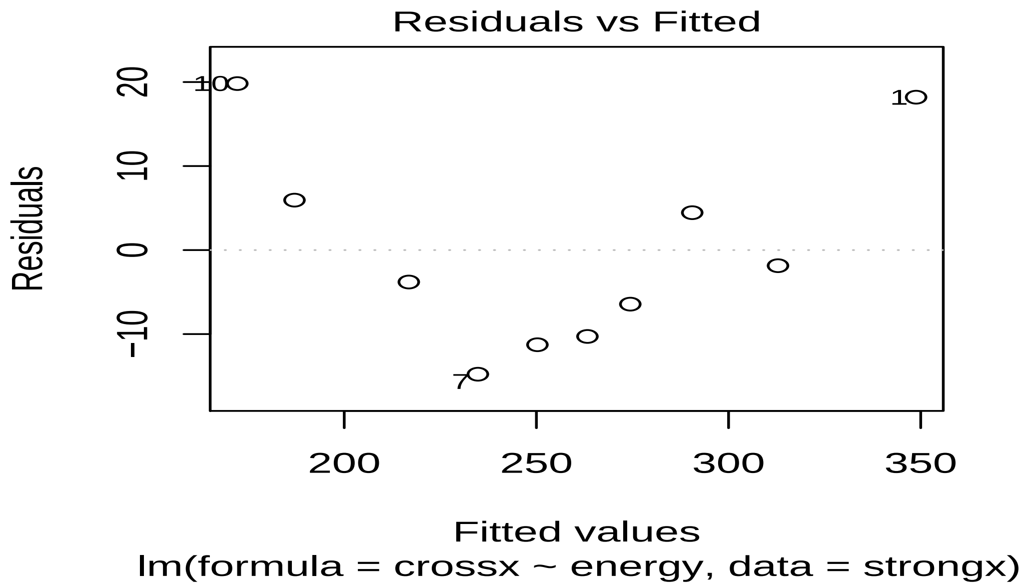


Figure 1: Plot of residuals versus fitted values

```
Residuals:
    Min      1Q  Median      3Q     Max
-14.773  -9.319  -2.829   5.571  19.818

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   135.00      10.08   13.40 9.21e-07 ***
energy        619.71      47.68   13.00 1.16e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.69 on 8 degrees of freedom

Multiple R-Squared: 0.9548,     Adjusted R-squared: 0.9491
F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.165e-06
```

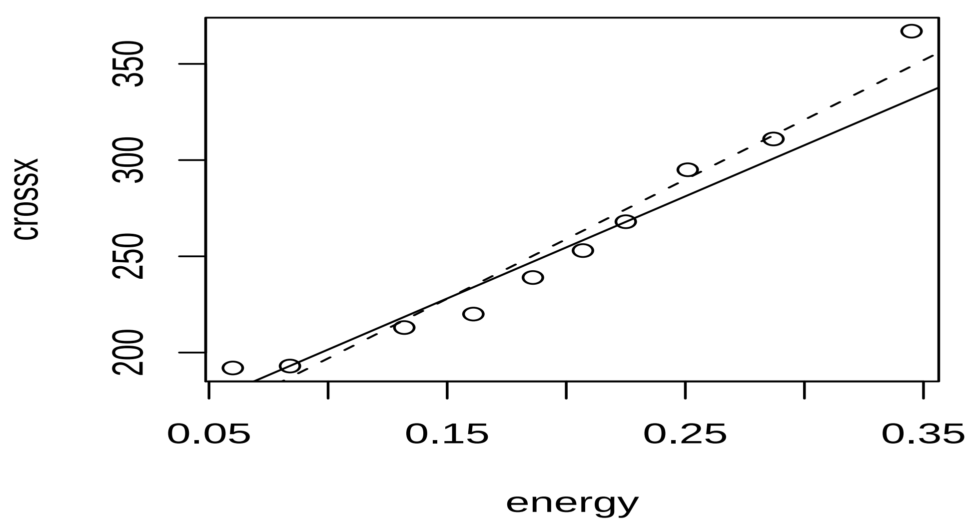
Now, we fit me model with weights

```
> g<-lm(crossx~energy,data=strongx,weights=sd^-2)
> summary(g)

Call:
lm(formula = crossx ~ energy, data = strongx, weights = sd^-2)
```

```
Residuals:
       Min         1Q     Median         3Q        Max
-2.323e+00 -8.842e-01  1.266e-06  1.390e+00  2.335e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  148.473      8.079   18.38 7.91e-08 ***
energy       530.835     47.550   11.16 3.71e-06 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 1.657 on 8 degrees of freedom

Multiple R-Squared: 0.9397,     Adjusted R-squared: 0.9321
F-statistic: 124.6 on 1 and 8 DF,  p-value: 3.710e-06
```

Compare the value of $\hat{\sigma}$, and the two fits in the following picture

```
> plot(crossx~energy,data=strongx)
> abline(g)
> abline(gu,lty=2)
```



Bibliography

Faraway, J. (2002). *Practical Regression and Anova Using R (electronic book)*. http://www.stat.lsa.umich.edu/ faraway/book/.

Longley, J. (1967). An appraisal of least squares programs from the point of view of the user. *Journal of the American Statistical Association*, 62:819–841.

Ruppert, D. and Carroll, R. (1988). *Transformation and Weightening in Regression*. Chapman and Hall, London and New York.