STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

Chapter 5: WLS and LOF

Weighted Least Square
Lack of Fit

Weighted Least Squares (WLS)

- relax the assumption $\mathrm{Var}(Y|X) = \sigma^2$
- change to $\mathrm{Var}(Y|X = x_i) = \mathrm{Var}(e_i) = \frac{\sigma^2}{w_i}$
  where $w_1, \cdots, w_n$ are known positive numbers
- in matrix form, the model becomes

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{e} \qquad \mathrm{Var}(\mathbf{e}) = \sigma^2 \boldsymbol{W}^{-1},$$

  where $\boldsymbol{W}$ is a diagonal matrix with elements $w_1, \cdots, w_n$
- the estimator $\boldsymbol{\beta}$ is defined as the minimizer of

$$\begin{aligned} RSS(\boldsymbol{\beta}) &= \sum_i w_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 \\ &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

WLS Solution

- the WLS solution is $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}$
- this can be obtained using results from OLS
- more precisely, transform the WLS problem into an OLS problem
- first we calculate

$$
\begin{aligned}
\mathrm{Var}(\boldsymbol{W}^{1/2}\mathbf{e}) &= \boldsymbol{W}^{1/2}\mathrm{Var}(\mathbf{e})\boldsymbol{W}^{1/2} \\
&= \boldsymbol{W}^{1/2}(\sigma^2\boldsymbol{W}^{-1})\boldsymbol{W}^{1/2} \\
&= \boldsymbol{W}^{1/2}(\sigma^2\boldsymbol{W}^{-1/2}\boldsymbol{W}^{-1/2})\boldsymbol{W}^{1/2} \\
&= \sigma^2(\boldsymbol{W}^{1/2}\boldsymbol{W}^{-1/2})(\boldsymbol{W}^{-1/2}\boldsymbol{W}^{1/2}) \\
&= \sigma^2\mathbf{I}
\end{aligned}
$$

WLS Solution - con't

- multiply $\boldsymbol{W}^{1/2}$ to the regression model

$$\boldsymbol{W}^{1/2}\boldsymbol{Y} = \boldsymbol{W}^{1/2}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}^{1/2}\mathbf{e}$$

- define $\mathbf{Z} = \boldsymbol{W}^{1/2}\mathrm{Y}, \mathrm{M} = \boldsymbol{W}^{1/2}\boldsymbol{X}$ and $\mathbf{d} = \boldsymbol{W}^{1/2}\mathbf{e}$, then

$$\mathbf{Z} = \mathrm{M}\boldsymbol{\beta} + \mathbf{d}$$

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathrm{M}'\mathrm{M})^{-1}\mathrm{M}'\mathbf{Z} \\
&= \left((\boldsymbol{W}^{1/2}\boldsymbol{X})'(\boldsymbol{W}^{1/2}\boldsymbol{X})\right)^{-1}(\boldsymbol{W}^{1/2}\boldsymbol{X})'(\boldsymbol{W}^{1/2}\boldsymbol{Y}) \\
&= (\boldsymbol{X}'\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}\boldsymbol{Y}) \\
&= (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y})
\end{aligned}$$

WLS: Other Remarks

- how to determine the weights?
- sometimes the weights $w_1, \cdots, w_n$ are known

  (i) if y_i is the average of n_i observations, then
  $\mathrm{Var}(y_i) = \frac{\sigma^2}{n_i}$ and $w_i = n_i$

  (ii) if y_i is the total of n_i observations, then $\mathrm{Var}(y_i) = n_i\sigma^2$
  and $w_i = \frac{1}{n_i}$
- collapse data by predictor values (sufficient statistic)
- sometimes W may depend on unknown parameters, and the choice could be subjective or based on some criteria

Lack of Fit (LOF)

- F-test from ANOVA could only tell if the regression model (i.e. slope in simple linear regression) helps explaining or not
- but it does not tell if the explanation is enough
- that is, any lack of fit
- main idea behind the "Lack of Fit Test":
  - if the model is good, then $\mathrm{E}(\hat{\sigma}^2) \approx \sigma^2$ unbiased $\Rightarrow$ model is correct
  - if the model is "not enough", then $\hat{\sigma}^2$ will be estimating something bigger than σ^2 (why?)
- so we could compare σ^2 and $\hat{\sigma}^2$

Lack of Fit - con't

- Lack of Fit Test: two cases:
  1. σ^2 known
  2. σ^2 unknown
- σ^2 known, if there no lack of fit (NH), assuming normal error,

$$X^2 = \frac{RSS}{\sigma^2} = \frac{(n-(p+1))\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{(n-(p+1))}$$

- this actually becomes a hypothesis test, p-value is $P(X^2 \geq X^2_{obs} |$ no lack of fit$)$

Lack of Fit, σ^2 unknown

- what do we do if σ^2 is unknown?
- estimate it!
- but we need to estimate it in a “model-free” manner: not use any model
- we can do it if we have repeated measurements at some x_i’s, otherwise NOT!
- we call these repeated measurements replicates, denoted by y_{ij}, $j = 1, \ldots, n_i$, corresponding to x_i

Sum of Squares for Pure Error

- for example, if we have 3 replicates at x_i, then we can calculate the sample variance of these 3 observations
- and use it as an estimate of σ^2 (at x_i)
- since we assume Var$(y_{ij}|x_i) = \sigma^2$ is constant at all x_i's
- if we have replicates at more values of x_i, then we can pool them together to get a better estimate of σ^2
- this involves the calculation of SS_{pe}, sum of squares for pure error

Computation of Pure Error

- Table 5.4 An Illustration of the Computation of Pure Error

x_i	y_{ij}	$\overline{y}_i$	$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$	$\hat{\sigma}$	df
1	2.55				
1	2.75	2.6233	0.0243	0.1102	2
1	2.57				
2	2.40	2.4000	0	0	0
3	4.19	4.4450	0.1301	0.3606	1
3	4.70				
4	3.81				
4	4.87	4.0325	2.2041	0.8571	3
4	2.93				
4	4.52				
			2.3585		6

Computation of Pure Error - con't

- $SS_{pe} = 0.0243 + \cdots + 2.2041 = 2.3585$ with 6 df
- similar to "pooled sample variance", the pure error estimate of σ^2 is

$$\hat{\sigma}^2_{pe} = SS_{pe}/df_{pe} = 2.3585/6 = 0.3931$$

- as similar to $SYY = SS_{reg} + RSS$, we split RSS as
- $RSS = SS_{lof} + SS_{pe}$
  SS_{lof}: sum of squares due to lack of fit $(\bar{y}_i \Rightarrow \beta_0 + \beta_1 x_i)$
  SS_{pe}: sum of squares due to pure error $(y_{ij} \Rightarrow \bar{y}_i)$
- implied by SS_{pe} is a saturated model

Decomposition: $RSS = SS_{pe} + SS_{lof}$

$$
\begin{aligned}
RSS_{ols} &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - \hat\beta_0 - \hat\beta_1 x_i)^2 \\
&= \sum_{i}\sum_{j}(y_{ij} - \bar y_i + \bar y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 \\
&= \sum_{i}\sum_{j}(y_{ij} - \bar y_i)^2 + \sum_{i} n_i(\bar y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 \\
&\qquad + 2\sum_{i=1}^{n}\left[\sum_{j=1}^{n_i}(y_{ij} - \bar y_i)\right]\underline{(\bar y_i - \hat\beta_0 - \hat\beta_1 x_i)} \\
&= \sum_{i}\sum_{j}(y_{ij} - \bar y_i)^2 + \sum_{i} n_i(\bar y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 \\
&= SS_{pe} + SS_{lof} = SS_{pe} + RSS_{wls}.
\end{aligned}
$$

doesn't depend on j

Lack of Fit, σ^2 unknown

- obtained from R function "pureErrorAnova" in "alr3"

TABLE 5.5 Analysis of Variance for the Data in Table 5.4

F value is not 11.6247

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	1	4.5693	4.5693	11.6247	0.01433
Residuals	8	4.2166	0.5271		
Lack of fit	2	1.8582	0.9291	2.3638	0.17496
Pure error	6	2.3584	0.3931		

- $$F\text{-value} = \frac{SS_{lof}/df_{lof}}{SS_{pe}/df_{pe}}$$

- compare with $F(df_{lof}, df_{pe})$

Apple Shoots Data

- Y: # of stem units, X: days from dormancy
- a simple linear regression will do? partial data

Long Shoots

Day	*n*	$\overline{y}$	SD	*Len*
0	5	10.200	0.830	1
3	5	10.400	0.540	1
7	5	10.600	0.540	1
13	6	12.500	0.830	1
18	5	12.000	1.410	1
24	4	15.000	0.820	1
25	6	15.170	0.760	1
32	5	17.000	0.720	1
38	7	18.710	0.740	1
42	9	19.220	0.840	1

Apple Shoots Data - con't

TABLE 5.7 Regression for Long Shoots in the Apple Data

(a) WLS regression using day means

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.973754    0.314272   31.74   <2e-16
Day         0.217330    0.005339   40.71   <2e-16

Residual standard error: 1.929 on 20 degrees of freedom
Multiple R-Squared: 0.988

Analysis of Variance Table
          Df Sum Sq Mean Sq F value    Pr(>F)
Day        1 6164.3  6164.3  1657.2 < 2.2e-16
Residuals 20   74.4     3.7
```

(b) OLS regression of y on *Day*

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.973754    0.21630   56.11   <2e-16
Day         0.217330    0.00367   59.12   <2e-16
```

Residual standard error: ~~1.762~~ on 187 degrees of freedom
Multiple R-Squared: 0.949

```
Analysis of Variance Table
              Df Sum Sq Mean Sq F value    Pr(>F)
Regression     1 6164.3  6164.3  1657.2 < 2.2e-16
Residual     187  329.5     1.8
 Lack of fit  20   74.4     3.7    2.43    0.0011
 Pure error  167  255.1     1.5
```

Apple Shoots Data - con't

- WLS: use 22 daily means as response
  OLS: use 189 original # of stem units
- parameter estimates, SS_{reg} are the same, general conclusions are the same
- RSS_{wls} and RSS_{ols} are different
  $RSS_{wls} = 74.4$ with 20 d.o.f.
  $RSS_{ols} = SS_{pe} + SS_{lof} = 255.1 + 74.4 = 329.5$
- note $SS_{pe} = RSS_{ols} - RSS_{wls} = SYY_{ols} - SYY_{wls}$
- pure error test shows lack of fit, but such a large sample size ($n = 189$) can detect a small deviation that may not be scientifically or practically important

General F-testing

- NH: $\mathbf{Y} = \mathbf{X_1}\boldsymbol{\beta}_1 + \mathbf{e}$
  AH: $\mathbf{Y} = \mathbf{X_1}\boldsymbol{\beta}_1 + \mathbf{X_2}\boldsymbol{\beta}_2 + \mathbf{e}$
- in general, model in NH is a subset of the model in AH
- i.e., by setting some parameters in AH to 0
- $F = \frac{(RSS_{NH} - RSS_{AH})/(df_{NH} - df_{AH})}{RSS_{AH}/df_{AH}}$
- compare to critical value $F_{(\alpha, df_{NH} - df_{AH}, df_{AH})}$
  or compute p-value $P(F \geq F_{obs} | NH)$ with
  $F \sim F_{(df_{NH} - df_{AH}, df_{AH})}$ under NH