

Contents

- 1 Weighted Least Squares and Aitken Model
- 2 Robust Regression

An Overview

- Weight least squares
- Ridge regression (lec 13)
- Robust regression
- Nonparametric regression (STAT740, STAT570)
- Bootstrap (STAT600)

1 Weighted Least Squares and Aitken Model

2 Robust Regression

THE AITKEN MODEL

- $\mathbf{y} = \mathbf{X} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{V})$ with **known** nonsingular variance-covariance matrix \mathbf{V}
 - ▶ Identical to the Gauss-Markov Linear Model except that $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{V}$ instead of $\sigma^2 \mathbf{I}$.
 - ▶ The Normal Theory Aitken Model adds an assumption of normality:
 $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{V})$
 - ▶ Observations are now correlated, or have unequal variances.
 - ▶ The correlations, or unequal variances, follow a known pattern

Examples

- Analysis of averages: The data to be analyzed are averages of unequal numbers of observations. y_i is an average of n_i observations; $\text{Var}(Y_i) = \sigma^2/n_i$; that is the first 4 rows and columns of $\text{Var}(\epsilon)$ is $\sigma^2 \text{diag}(1/n_1, 1/n_2, 1/n_3, 1/n_4)$.
- Analysis of data on a pedigree (genetic relationships among parents, children, grandchildren, etc). Then, genetic correlations between parents and children, among children are all known $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ then the first 4 rows and columns of $\text{Var}(\boldsymbol{\epsilon})$ are

$$\sigma^2 \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$$

where ρ_{ij} is the known genetic correlation among individuals i, j .

- Regression on data collected over time: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $X_i =$ year like 1990, 1991, \dots . Assume errors follow an autoregressive process, the first 4 rows and columns of $\text{Var}(\epsilon)$ are

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

where ρ_{ij} is the known genetic correlation among individuals i, j .

- Split plot experiment with two experimental units: growth chamber and pot for example. Then two error variances σ_c^2 and σ_p^2 , if two pots per chamber, the first 4 rows and columns of $\text{Var}(\epsilon)$ are

$$\sigma_c^2 \begin{bmatrix} 1+k & 1 & 0 & 0 \\ 1 & 1+k & 0 & 0 \\ 0 & 0 & 1+k & 1 \\ 0 & 0 & 1 & 1+k \end{bmatrix}$$

where $k = \sigma_p^2 / \sigma_c^2$ is known.

Review of Spectral Decomposition Theorem

- The Spectral Decomposition Theorem: any positive definite symmetric matrix V can be written as $V = UDU'$ where D is a diagonal matrix of eigenvalues and U is a matrix of orthonormal eigenvectors that $UU' = I$
 - ▶ For any positive definite symmetric V , there exists a nonsingular symmetric matrix denoted by $V^{1/2}$ that $V^{1/2}V^{1/2} = V$
 - ▶ Given U and D for which $UDU' = V$ then $V^{1/2} = U\sqrt{D}U'$
 - ▶ $V^{1/2}$ can be viewed as the “square root” of a matrix (verify $V^{1/2}V^{1/2} = V$ (HW))
 - ▶ Define $V^{-1/2}$ as $(V^{-1})^{1/2}$ and compute it as $V^{-1/2} = U(1/\sqrt{D})U'$ where matrix $1/\sqrt{D}$ is diagonal matrix with elements $1/\sqrt{D_{ii}}$

Converting an Aitken model to GM model

- Data model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{V})$

- ▶ Pre-multiply all terms by $\mathbf{V}^{-1/2}$ then

$$\mathbf{V}^{-1/2}\mathbf{V}^{-1/2}\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1/2}\boldsymbol{\varepsilon}$$

- ▶ This leads to a regression model $\mathbf{Z} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\delta}$ with

$$\mathbf{Z} = \mathbf{V}^{-1/2}\mathbf{y}, \quad \mathbf{W} = \mathbf{V}^{-1/2}\mathbf{X}, \quad \boldsymbol{\delta} = \mathbf{V}^{-1/2}\boldsymbol{\varepsilon}$$

- ▶ Motivation: what is $\text{Var}(\boldsymbol{\delta})$?

- ▶ After transformation, we have a Gauss-Markov Model.

Statistical results from Aitken model

- Goal: estimate $\mathbb{E}(\mathbf{y})$
- Under the GM model, the best estimate of $\mathbb{E}(\mathbf{Z})$ is

- To estimate $\mathbb{E}(\mathbf{y}) = \mathbf{V}^{1/2}\mathbb{E}(\mathbf{Z})$, we use

$$\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

- Goal: estimate $C\beta$
- $C\beta$ is estimable if C can be written as a linear combination of rows of W , that is $C = AW$
- if $C\beta$ estimable, the BLUE is the ordinary least squares (OLS) estimator using Z and W

- The estimator

$$C(X'V^{-1}X)^{-1}X'V^{-1}y := C\hat{\beta}_G$$

is called Generalized Least Squares (GLS) estimator.

- Goal: estimate $\text{Var}(\mathbf{C}\boldsymbol{\beta}_G)$
- Following the approach earlier

Generalized Least Squares

- $\hat{\beta}_G = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ is a solution to the Aitken Equations:

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

which follow from the Normal Equations

$$\begin{aligned} \mathbf{W}'\mathbf{W}\mathbf{b} &= \mathbf{W}'\mathbf{Z} \\ \Rightarrow \mathbf{X}'\mathbf{V}^{-1/2}\mathbf{V}^{-1/2}\mathbf{X}\mathbf{b} &= \mathbf{X}'\mathbf{V}^{-1/2}\mathbf{V}^{-1/2}\mathbf{y} \\ \Rightarrow \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\mathbf{b} &= \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \end{aligned}$$

- Solving the Normal Equations is equivalent to

$$\min_{\beta \in \mathbb{R}^p} (\mathbf{Z} - \mathbf{W}\beta)'(\mathbf{Z} - \mathbf{W}\beta)$$

- ▶ Now

$$\begin{aligned} (\mathbf{Z} - \mathbf{W}\beta)'(\mathbf{Z} - \mathbf{W}\beta) &= (\mathbf{V}^{-1/2}\mathbf{y} - \mathbf{V}^{-1/2}\mathbf{X}\beta)'(\mathbf{V}^{-1/2}\mathbf{y} - \mathbf{V}^{-1/2}\mathbf{X}\beta) \\ &= (\mathbf{y} - \mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta), \text{ is called Generalized Least Squares problem} \\ &\text{(GLS)} \end{aligned}$$

- ▶ Thus, $\hat{\beta}_G = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ is a solution to this GLS

- When \mathbf{V} is diagonal, the term “Weighted Least Squares” (WLS) is commonly used instead of GLS

- ▶ Define \mathbf{D} = diagonal matrix of inverse weights, $D_{ii} = 1/w_i$; then

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{D}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^n w_i (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$$

- ▶ When observations have **unequal** variances, then $\text{Var}(\mathbf{y}) = \text{diag}(\sigma^2)$ and w_i is proportional to $1/\sigma_i^2$
- An unbiased estimator of σ^2 is

$$\hat{\sigma}_G^2 := \frac{\mathbf{Z}'(\mathbf{I} - \mathbf{P}_w)\mathbf{Z}}{n - \text{rank}(\mathbf{W})} = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G)' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G)}{n - k}$$

(STAT640 for more details)

Summary of Aitken model results

- Aitken model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{V})$
 - 1 equivalent to GM model: $\mathbf{Z} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\delta}$ by premultiplying $\mathbf{V}^{-1/2}$
 - 2 estimate $\mathbb{E}(\mathbf{y}) = \mathbf{V}^{-1/2}\mathbb{E}(\mathbf{Z})$ by $\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$
 - 3 estimate $\mathbf{C}\boldsymbol{\beta}$ by $\mathbf{C}\hat{\boldsymbol{\beta}}_G = \mathbf{C}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$
 - 4 estimate $\text{Var}(\mathbf{C}\boldsymbol{\beta})$ by $\hat{\sigma}_G^2\mathbf{C}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{C}'$
 - 5 estimate $\text{Var}(\boldsymbol{\delta})$ by

$$\hat{\sigma}_G^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G)' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G)}{n - k}$$

, which means estimating $\text{Var}(\mathbf{y})$ by $\hat{\sigma}_G^2\mathbf{V}$

- If add normality, all inferential results for GM linear model follow

What if V misspecified?

- Data follow Aitken model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{V})$ but accidentally analyzed using $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I})$
- Results stated; derivations can be found in the textbook or most Econometrics textbooks
 - ▶ $\mathbb{E}(\mathbf{y})$ unbiased, so $\mathbf{C}\hat{\boldsymbol{\beta}}$ unbiased
 - ▶ OLS estimates not as efficient as GLS, $\text{Var}(\mathbf{C}\hat{\boldsymbol{\beta}}) > \text{Var}(\mathbf{C}\hat{\boldsymbol{\beta}}_G)$ (as you have seen in Exam 1)
 - ▶ $\text{Var}(\mathbf{C}\hat{\boldsymbol{\beta}})$ is NOT $\sigma^2 \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$, the bias can be considerable
 - ▶ Assume \mathbf{X} full rank so

All inference is suspect, unless $(\mathbf{X}'\mathbf{V}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$ is close to \mathbf{I}

What is V not known?

Use **iterative reweighted least squares (IRLS)**:

- 1 Fit the model using ordinary least squares or GLS with a *working covariance* V^0 .
- 2 Estimate the standard deviation function $\{\sigma_i : i = 1, \dots, n\}$ by using the residuals $\{e_i : i = 1, \dots, n\}$ and denote the estimate as $\{\hat{s}_i : i = 1, \dots, n\}$; or more generally estimate \hat{V}^1 using $\text{Cov}(\hat{\varepsilon})$
- 3 Obtain the estimated weights using

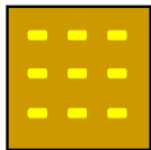
$$\hat{w}_i = \frac{1}{\hat{s}_i^2} \quad \text{or} \quad (\hat{V}^1)^{-1}$$

- 4 Compute the weighted least squares estimate $\hat{\beta}_G$ using \hat{w}_i or $(\hat{V}^1)^{-1}$.
- 5 Repeat Steps 2–4 until the estimates are convergent.

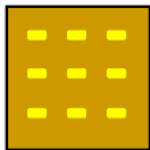
An example

- Experiment Researchers were interested in comparing the dry weight of maize seedlings from two different genotypes. For each genotype, nine seeds were planted in each of four trays. The eight trays in total were randomly positioned in a growth chamber. Three weeks after the emergence of the first seedling, emerged seedlings were harvested from each tray and weighed together after drying to obtain one weight for each tray. Although nine seeds were planted in each tray, fewer than nine seedlings emerged in many of the trays. Thus, weights were recorded on a per seedling basis, and the number of seedlings that emerged in each tray was also recorded.

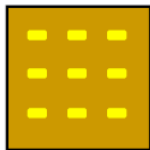
A



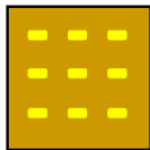
A



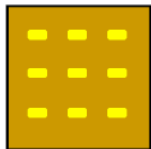
B



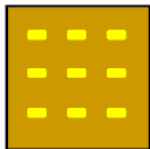
B



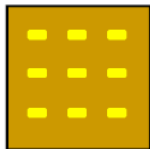
B



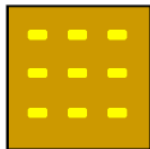
A



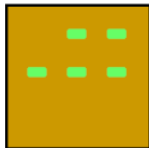
B



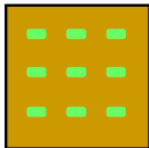
A



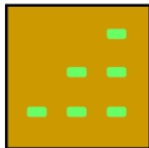
A



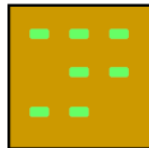
A



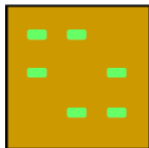
B



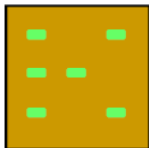
B



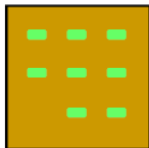
B



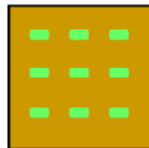
A



B



A



```

> d=read.delim("http://www.stat.colostate.edu/~riczw/teach/STAT540_F15
/computing/lec12/SeedlingDryWeight.txt")
> d
  Genotype Tray AverageWeightPerSeedling NumberOfSeedlings
1         A   1                10                5
2         A   2                18                9
3         A   3                14                6
4         A   4                19                9
5         B   5                13                6
6         B   6                10                7
7         B   7                15                6
8         B   8                 9                8
> y=d[,3]
> geno=d[,1]
> count=d[,4]
> X=matrix(model.matrix(~geno),nrow=8)
> X
      [,1] [,2]
[1,]  1  0
[2,]  1  0
[3,]  1  0
[4,]  1  0
[5,]  1  1
[6,]  1  1
[7,]  1  1
[8,]  1  1

```

```
> V=diag(1/count)

> # Compute  $V^{-.5}$ 
> A=diag(sqrt(count))

> # In general, we could compute  $V^{-.5}$  as follows:
> e=eigen(V)
> A=e$vectors%*%diag(1/sqrt(e$values))%*%t(e$vectors)

> # Now transform y and X to z and W.
> z=A%*%y
> W=A%*%X
```

```
> o=lm(z~W-1)
```

```
> summary(o)
```

Call:

```
lm(formula = z ~ W - 1)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.648	-5.619	-0.100	6.422	8.690

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
W1	16.103	1.650	9.763	6.64e-05 ***
W2	-4.622	2.376	-1.946	0.0997 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.883 on 6 degrees of freedom

Multiple R-squared: 0.959, Adjusted R-squared: 0.9454

F-statistic: 70.21 on 2 and 6 DF, p-value: 6.882e-05

```
> # Because V is diagonal in this case, we can alternatively analyze using lm and th
> # weights argument.
> o2=lm(y~geno,weights=count)
> summary(o2)
```

Call:

```
lm(formula = y ~ geno, weights = count)
```

Weighted Residuals:

Min	1Q	Median	3Q	Max
-13.648	-5.619	-0.100	6.422	8.690

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	16.103	1.650	9.763	6.64e-05 ***
genoB	-4.622	2.376	-1.946	0.0997 .

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 8.883 on 6 degrees of freedom

Multiple R-squared: 0.3868, Adjusted R-squared: 0.2847

F-statistic: 3.785 on 1 and 6 DF, p-value: 0.09966

```
> # The unweighted (OLS) analysis is inferior in this case. The OLS estimator of beta is still
> # unbiased, but it's variance is larger than that of the GLS estimator. OLS inferences
> # regarding beta are not in general, valid.
```

```
> o3=lm(y~geno)
> summary(o3)
```

Call:

```
lm(formula = y ~ geno)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.250	-2.000	0.000	2.875	3.750

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	15.250	1.750	8.714	0.000126 ***
genoB	-3.500	2.475	-1.414	0.207031

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.5 on 6 degrees of freedom

Multiple R-squared: 0.25, Adjusted R-squared: 0.125

F-statistic: 2 on 1 and 6 DF, p-value: 0.207

1 Weighted Least Squares and Aitken Model

2 Robust Regression

Robust regression

- Look for a reason to remove an influential outlier.
- A good option when many regression models must be fit in some automated way.
- Option 1: Least Absolute Deviations (LAD)– Also called \mathcal{L}_1 regression
 - ▶ Minimize L_1 with respect to β instead of least square where

$$L_1 = \sum_i |e_i|$$

- ▶ Not squaring the residuals dampens the impact of large residuals.
- ▶ Asymptotic theory does exist
- ▶ Estimated regression coefficients may not be unique (intuitively, why?)

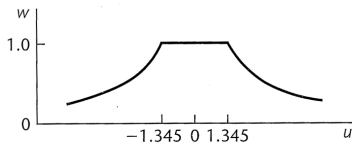
IRLS Robust Regression

- Option 2: Use a set of weights to reduce the impact of outliers
 - ▶ Weights vary inversely with the magnitude of the residual
- “Iteratively Reweighted Least Squares” method estimates the weights and regression coefficients simultaneously (in iterative steps)
 - ▶ Choose a weight function for weighting cases
 - ▶ Choose starting weights
 - ▶ Apply WLS to obtain the working residuals
 - ▶ Use these residuals to revise the weights
 - ▶ Return to WLS step
 - ▶ Continue this cycle until convergence

Weight functions

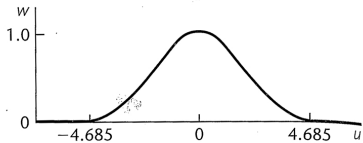
(a) Huber Weight Function

$$w = \begin{cases} 1 & |u| \leq 1.345 \\ 1.345/|u| & |u| > 1.345 \end{cases}$$



(b) Bisquare Weight Function

$$w = \begin{cases} [1 - (u/4.685)^2]^2 & |u| \leq 4.685 \\ 0 & |u| > 4.685 \end{cases}$$



Details

- Starting values:
 - ▶ For Huber, use OLS residuals to calculate weights
 - ▶ For bisquare, use Huber residuals to calculate weights
- Scaled residuals. Both methods operate on the scaled residuals:
 - ▶ Estimate “median absolute deviation”
$$MAD = \text{median}\{|e_i - \text{median}(e_i)|\} / 0.6745$$
 to estimate σ
 - ▶ Scaled residuals are $u_i = e_i / MAD$
- Stopping rule:
 - ▶ Stop when weights change very little, or
 - ▶ Stop when residuals change very little, or
 - ▶ Stop when $\hat{\beta}$ changes very little, or
 - ▶ Stop when fitted values change very little.