STAT 224 Lecture 14
Chapter 7 Weighted Least Squares

Yibi Huang

Unequal Variance

- The linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i,$$

  where the random errors are iid $N(0, \sigma^2)$.

- What if the ε_i's are indep. w/ unequal var $N(0, \sigma_i^2)$?
- The ordinary least squares (OLS) estimates for β_j's remain unbiased, but no longer have the minimum variance.
- **Weighted Least Squares (WLS)** fixes the problem of heteroscedasticity
- As seen in Chapter 6, we can also cope with heteroscedasticity by transforming the response; but sometime such a transformation is not available

Weighted Least Squares

For the model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad \text{where } \varepsilon_i\text{'s are indep. w/ } \mathrm{var}(\varepsilon_i) = \sigma_i^2,$$

the Weighted Least Squares method finding estimates for β's by minimizing

$$L(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2}{\sigma_i^2}.$$

- In OLS, $\sigma_i^2 = \sigma^2$ for all i, equivalent to minimize $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$
- In WLS, we focus
  - more on minimizing errors of obs. w/ smaller variances (more accurate), and
  - less on minimizing errors of obs. w/ larger variances (less accurate)

How to Estimate the Unknown Unequal Variance σ_i^2

There would be too many parameters to estimate
if each observation has its own parameter σ_i^2 of variance
since we can estimate at most n parameters w/ n observations

- Parameters of OLS: $\beta_0, \beta_1, \dots, \beta_p, \sigma^2$
- Parameters of WLS: $\beta_0, \beta_1, \dots, \beta_p, \sigma_1^2, \dots, \sigma_n^2$

Need prior knowledge about the variances σ_i^2. We'll focus on the case when σ_i^2's are **inversely proportional to** some *weights* w_i

$$\sigma_i^2 = \sigma^2 / w_i \quad i = 1, 2, \dots, n$$

where the *weights* $w_1, w_2, \dots, w_n$ are **known positive numbers** and σ^2 is unknown. In this case, WLS is equivalent to minimize

$$\sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

Weighted Least Squares (WLS) Estimates for β's (May Skip)

The WLS estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ that minimize the weighted sum of squares

$$\sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

is

$$\widehat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$$

where

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \ \mathbf{W} = \begin{pmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_n \end{pmatrix}_{n \times n}$$

and $\mathbf{W}$ is an $n \times n$ matrix with $(w_1, w_2, \ldots, w_n)$ on the diagonal and 0 elsewhere.

Standard Errors of WLS Estimates for β's (May Skip)

Under the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2 / w_i,$$

the covariance matrix of $\widehat{\boldsymbol{\beta}}_{WLS}$ is

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_{WLS}) = \sigma^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$$

The unknown variance parameter σ^2 is estimated by

$$\widehat{\sigma}^2 = \text{MSE} = \frac{\text{SSE}}{n-p-1}, \quad \text{where SSE} = \sum_i w_i (y_i - \widehat{y}_i)^2,$$

where the fitted values are

$$\widehat{y}_i = \widehat{\beta}_{0,WLS} + \sum_{j=1}^{p} \widehat{\beta}_{j,WLS} x_{ij}, \quad i = 1, \ldots, n$$

The s.e. of the WLS estimate $\widehat{\beta}_{j,WLS}$ for β_j is

$$\sqrt{\widehat{\sigma}^2 \times (j\text{th diagonal element of the matrix } (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1})}.$$

Tests and CIs for β_j for WLS

The t-statistic

$$\frac{\widehat{\beta}_{j,WLS} - \beta_j^0}{s.e.(\widehat{\beta}_{j,WLS})} \sim t_{n-p-1}, \quad \text{under } \mathrm{H}_0\text{: } \beta_j = \beta_j^0$$

and the t-CI

$$\widehat{\beta}_{j,WLS} \pm t_{(n-p-1,\alpha/2)} s.e.(\widehat{\beta}_{j,WLS})$$

for β_j's for WLS

can be used in the same way as those for OLS.

WLS When σ_i is Proportional to x_i

If σ_i is Proportional to Some Predictor x_i

Suppose the variance of the ith observation

$$\sigma_i^2 = \mathrm{Var}(\varepsilon_i) = \sigma^2 x_i^2$$

is known to be proportional to some value $x_i > 0$, where $\sigma^2 > 0$ is an unknown constant

- Since σ^2 is a constant, this is equivalent to use the weights

$$w_i = \frac{1}{x_i^2}.$$

- Thus we minimize:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} \frac{1}{x_i^2}(y_i - \beta_0 - \beta_1 x_i)^2.$$

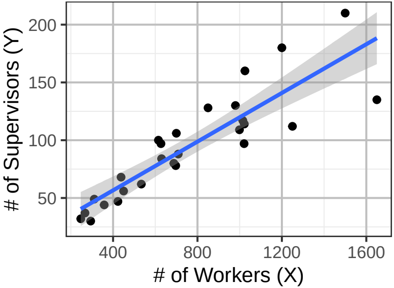
Supervisor/Employee Data (p.176)

Data: http://www.stat.uchicago.edu/~yibi/s224/data/P176.txt

X = # of Supervised Workers

Y = # of Supervisors in 27 Industrial Establishments

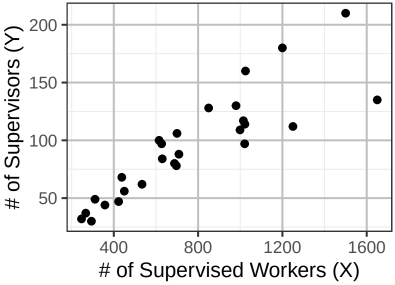
```
supvis = read.table("P176.txt", h=T)
library(ggplot2)
ggplot(supvis, aes(x=X, y=Y))+geom_point()+geom_smooth(method='lm')+
  labs(x="# of Workers (X)", y="# of Supervisors (Y)")
```



Supervisor/Employee Data — WLS Approach

As the variance of Y is proportional to X, we can use WLS with weight $w_i = 1/x_i^2$.

The `lm()` command can also fit WLS models. One just need to specify the *weights* in addition.




```
summary(lm(Y ~ X, data=supvis, weights=1/X^2))
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.803296   4.569745   0.832    0.413
X           0.120990   0.008999  13.445 6.04e-13 ***
---
Residual standard error: 0.02266 on 25 degrees of freedom
Multiple R-squared:  0.8785,    Adjusted R-squared:  0.8737
F-statistic: 180.8 on 1 and 25 DF,  p-value: 6.044e-13
```

Example: CIs for β_j in WLS

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.803296   4.569745   0.832    0.413
X           0.120990   0.008999  13.445 6.04e-13 ***
```

For the Supervisor/Employees Data, the 95% CI for β_1 is

$$\begin{aligned}\widehat{\beta}_{j,WLS} \pm t_{(n-p-1,\alpha/2)} s.e.(\widehat{\beta}_{j,WLS}) &\approx 0.12099 \pm 2.0595 \times 0.008999 \\ &\approx (0.1025, 0.1395)\end{aligned}$$

as $t_{(n-p-1,\alpha/2)} = t_{(25,0.025)}$ =`qt(1-0.025, df=25)` ≈ 2.0595.

<u>Interpretation</u>: Need to hire 10.25 to 13.95 more supervisors on average for every extra 100 workers, at 95% confidence.

The CI for β's can also be found using `confint()`.

```
confint(lm(Y ~ X, data=supvis, weights=1/X^2))
              2.5 %  97.5 %
(Intercept) -5.6083 13.2149
X            0.1025  0.1395
```

Sum of Squares and Multiple R^2 for WLS

- SST = $\sum_i w_i(y_i - \bar{y}_w)^2$, where $\bar{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i}$
- SSR = $\sum_i w_i(\widehat{y}_i - \bar{y}_w)^2$
- SSE = $\sum_i w_i(y_i - \widehat{y}_i)^2$
- SST = SSR + SSE remains valid
- df of SS: same as for OLS
- Multiple R^2 = SSR/SST
  - cannot compare the Multiple R^2 of a WLS model and a OLS model since SSR and SST are calculated differently
- MSE = SSE$/(n-p-1) = \widehat{\sigma}^2$
- `Residual standard error: 0.02266` gives $\sqrt{\text{MSE}}$
- The estimate for $\sigma_i^2 = \text{Var}(\varepsilon_i) = \sigma^2/w_i$ is MSE$/w_i$

```
Residual standard error: 0.02266 on 25 degrees of freedom
Multiple R-squared:  0.8785,    Adjusted R-squared:  0.8737
F-statistic: 180.8 on 1 and 25 DF,  p-value: 6.044e-13
```

F-tests for WLS

If two WLS models are nested and use the same weights, then we can compare them using the ANOVA F-statistic

$$F = \frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(\text{dfE}_{reduced} - \text{dfE}_{full})}{\text{MSE}_{full}}$$

$$\sim F_{\text{dfE}_{reduced}-\text{dfE}_{full},\text{dfE}_{full}} \quad \text{under } \text{H}_0\text{: reduced model is correct}$$

```
lmwls = lm(Y ~ X, data=supvis, weights=1/X^2)
lmwls2 = lm(Y ~ X + I(X^2), data=supvis, weights=1/X^2)
anova(lmwls,lmwls2)
Analysis of Variance Table

Model 1: Y ~ X
Model 2: Y ~ X + I(X^2)
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     25 0.0128
2     24 0.0116  1   0.00124 2.58   0.12
```

Residuals for WLS in R

- `model$res` give the raw residuals $e_i = y_i - \hat{y}_i$, which are NOT adjusted by weights
- `hatvalues(model)` gives the *leverage* h_{ii}, which is the = ith diagonal element of the *hat matrix*

$$H = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$$

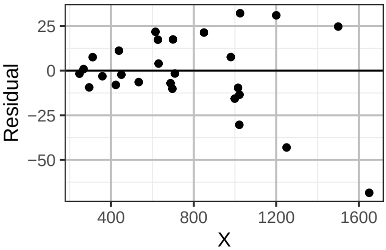
- $\mathrm{Var}(e_i) = \sigma_i^2(1 - h_{ii})$ where h_{ii} = leverage, and $\sigma_i^2 = \sigma^2/w_i$
- `rstandard(model)` gives internally Studentized residuals

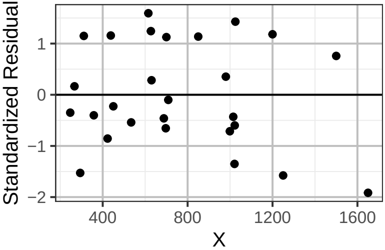
$$r_i = \frac{e_i}{\sqrt{\widehat{\sigma}_i^2(1 - h_{ii})}} \sim \text{ approx. } N(0, 1), \quad \text{where } \widehat{\sigma}_i^2 = \mathrm{MSE}/w_i$$

  which are weight-adjusted
- `rstudent(model)` gives externally Studentized residuals

Residual Plots

```
ggplot(supvis, aes(x=X, y=lmwls$res)) + geom_point() +
  ylab("Residual") + geom_hline(yintercept=0)
ggplot(supvis, aes(x=X, y=rstandard(lmwls))) + geom_point() +
  ylab("Standardized Residual") + geom_hline(yintercept=0)
```





- The raw residuals are not weight-adjusted
  The residual plot is still funnel-shaped
- To see if the weights are chosen properly to fix the heteroscedastic problem, plot standardized or studentized residuals and see if the points scatter evenly around the zero line

Confidence/Prediction Intervals for WLS Models in R

Note that *weights* must be provided for prediction or the intervals computed won't be correct.

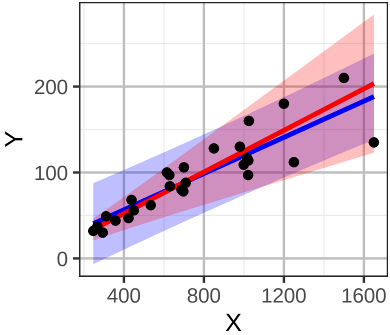
```
predict(lmwls, data.frame(X=1200), weights=1/1200^2,
        interval="confidence")
  fit   lwr   upr
1 149 134.3 163.7
predict(lmwls, data.frame(X=1200), weights=1/1200^2,
        interval="prediction")
  fit   lwr   upr
1 149 91.07 206.9
```

- At 95% confidence, industrial establishments with 1200 workers require 134.26 to 163.72 supervisors on average
- At 95% confidence, an industrial establishment that has 1200 workers is predicted to have 91.07 to 206.91 supervisors

95% Prediction Intervals — OLS v.s. WLS

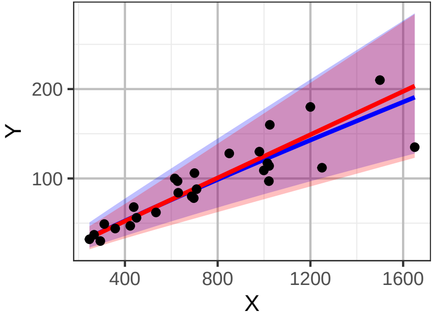
- Blue: OLS: `lm(Y ~ X, data=supvis)`
- Red: WLS: `lm(Y ~ X, data=supvis, weights=1/X^2)`

- Closer to points with smaller variance is the WLS line (red) than the OLS line (blue)
- WLS Prediction intervals reflect the variability of observations increases w/ X



WLS Model v.s. OLS Model w/ Transformation

- Blue: OLS model `log(Y) ~ log(X)`
- Red: WLS model `Y ~ X`



The OLS model w/ transformation `log(Y) ~ log(X)` (blue) and the WLS model `Y ~ X` (red) give nearly identical predicted values and prediction intervals. Both models are adequate.

WLS: Group Means with Varying Sample Sizes

Group Means with Varying Sample Sizes

Here is another scenario to use WLS.

$$y_i^{(j)} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{ip} + \varepsilon_i^{(j)}, \quad \varepsilon_i^{(j)} \sim N(0,\ \sigma^2)$$

- n_i observations $y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(n_i)}$ with identical predictor values: $x_{i1}, \ldots, x_{ip}$
- Only the group mean $\bar{y}_i = \sum_{j=1}^{n_i} y_i^{(j)}/n_i$ is recorded. The original values $y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(n_i)}$ are not available
- The variance of each individual $y_i^{(j)}$ is σ^2.
- The variance of a group mean $\bar{y}_i$ is $\sigma_i^2 = \mathrm{Var}(\bar{y}_i) = \dfrac{\sigma^2}{n_i}$, i.e.,

$$\bar{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \bar{\varepsilon}_i, \quad \bar{\varepsilon}_i \sim N(0,\ \sigma^2/n_i)$$

- Hence, the WLS weights are

$$w_i = n_i \quad \text{since} \quad \mathrm{Var}(\bar{\varepsilon}_i) = \frac{\sigma^2}{w_i} = \frac{\sigma^2}{n_i}.$$

Example: Travel-Chicago Data

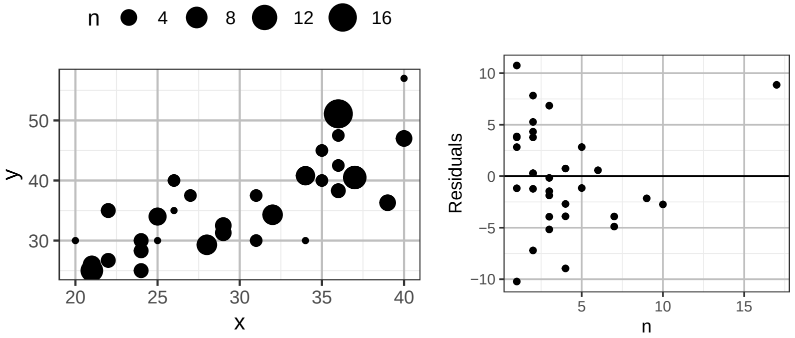
n	1	1	7	3	2	4	4	3	1	1	...	3
x	26	40	32	36	27	39	29	22	34	25	...	24
y	35	57	34.3	38.3	37.5	36.3	31.3	35	30	30	...	25.0

Data: http://www.stat.uchicago.edu/~yibi/s224/data/ChiBus.txt

- Each case is a pair of zones in the city of Chicago
- x = travel times, computed from bus timetables augmented by walk times from zone centers to bus-stops (assuming a walking speed of 3 mph) and expected waiting times for the bus (= half of the time between successive buses).
- y = average travel times as reported to the U.S. Census Bureau by n travelers.
- n = number of travelers/observations for each case

Data from Exercise 6.8 on p.129, Regression Analysis– Theory, Methods, and Applications by Ashish Sen, Muni Srivastava, 1990

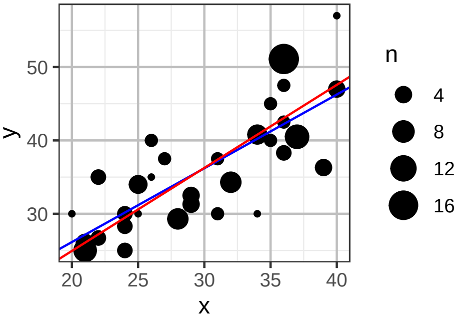
```r
ggplot(chibus, aes(x=x, y=y, size=n)) + geom_point() +
  theme(legend.position="top")
chi.ols = lm(y ~ x, data=chibus)
ggplot(chibus, aes(x=n, y=chi.ols$res)) + geom_point() +
  ylab("Residuals") + geom_hline(yintercept=0)
```



The scatterplot (left) looks fine but the residual plot (right) for the naive OLS model `lm(y ~ x, data=chibus)` shows that magnitude of residuals decreases as n increases.

OLS Line v.s. WLS Line

```
ols.beta = lm(y ~ x, data=chibus)$coef
wls.beta = lm(y ~ x, data=chibus, weights=n)$coef
ggplot(chibus, aes(x=x, y=y, size=n)) + geom_point() +
  geom_abline(intercept= ols.beta[1], slope= ols.beta[2], col="blue") +
  geom_abline(intercept= wls.beta[1], slope= wls.beta[2], col="red")
```

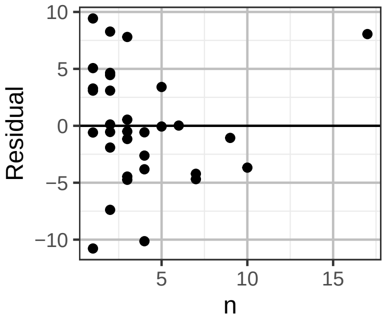


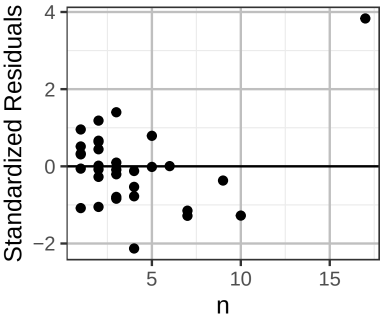
Blue line: OLS

Red line: WLS

Residual Plots of WLS

```
chi.wls = lm(y ~ x, data=chibus, weights=n)
ggplot(chibus, aes(x=n, y=chi.wls$res)) + geom_point() +
  ylab("Residual") + geom_hline(yintercept=0)
ggplot(chibus, aes(x=n, y=rstandard(chi.wls))) + geom_point() +
  ylab("Standardized Residuals") + geom_hline(yintercept=0)
```

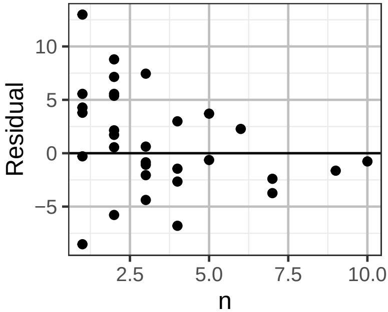


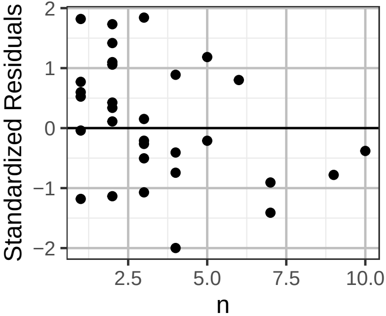


There is a potential outlier.

After Removing the Outlier

```
chibus2 = subset(chibus, n<17)
chi.wls2 = lm(y ~ x, data=chibus2, weights=n)
ggplot(chibus2, aes(x=n, y=chi.wls2$res)) + geom_point() +
  ylab("Residual") + geom_hline(yintercept=0)
ggplot(chibus2, aes(x=n, y=rstandard(chi.wls2))) + geom_point() +
  ylab("Standardized Residuals") + geom_hline(yintercept=0)
```





```r
# Model with the outlier
summary(chi.wls)$coef
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)    2.293     4.5903  0.4996 0.62101433061
x              1.132     0.1475  7.6764 0.00000001458
# Model without the outlier
summary(chi.wls2)$coef
            Estimate Std. Error t value       Pr(>|t|)
(Intercept)   7.4294     3.4747   2.138 0.041058924129
x             0.9146     0.1148   7.967 0.000000008721
```