Lecture 16: Weighted Least Squares

Ailin Zhang

2023-06-20

Recap: Addressing Violations of Assumptions

We transform variables (including predictors and responses):

- to solve the non-linearity problem
  - Common strategies: log()
- to reduce skewness (non-normal errors)
  - Power transformation
- to solve the non-constant variability problem
  - Variance-Stabilizing Transformation
  - Box-Cox Method

We have touched this topic when introducing polynomial models and ordinal categorical predictors

Unequal Variance

- For linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

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

- What if the ϵ_i's are independent with unequal variances: $N(0, \sigma_i^2)$?
- The ordinary least squares (OLS) estimates for β_j's remain unbiased, but no longer have the minimum variance.
- **Weighted Least Squares (WLS)** will fix the problem of heteroscedasticity

Weighted Least Squares

For the model, $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$ where ϵ_i's are independent with $\mathrm{Var}(\epsilon_i) = \sigma_i^2$

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

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

    - we focus more on minimizing errors of obs. with smaller variances (more accurate), and
    - we focus less on minimizing errors of obs. with larger variances (less accurate)

- In OLS, $\sigma_i^2 = \sigma^2$ for all i, minimizing L is equivalent to minimize $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$

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 with n observations
  - Parameters of WLS (at most): $\beta_0, \beta_1, \ldots, \beta_p, \sigma_1^2, \ldots, \sigma_n^2$
- 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

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

  where the weights $w_1, w_2, \ldots, w_n$ 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 Estimates

- The WLS estimate of $(\beta_0, \beta_1, \ldots, \beta_p)$ that minimizes the weighted sum of squares: $\sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$
- Matrix Notation:

$$RSS = (Y - X\beta)^T W (Y - X\beta)$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n1} & \ldots & x_{np} \end{bmatrix}, W = \begin{bmatrix} w_1 & & & & \\ & w_2 & & & \\ & & w_3 & & \\ & & & \ddots & \\ & & & & w_n \end{bmatrix}$$

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

Weighted Least Squares Estimates

$$RSS(\beta) = (Y - X\beta)^T W (Y - X\beta)$$

- $\hat{\beta}_{WLS} = (X^T W X)^{-1} X^T W Y$
- Derivation # 1: Treat the regression model as OLS:

$$W^{1/2} Y = W^{1/2} X\beta + W^{1/2}\epsilon$$

- Derivation # 2: Matrix Algebra

$$\frac{\partial RSS}{\partial \beta} =$$

Statistical Properties of WLS Estimator

- $\hat{\beta}_{WLS}$ is an unbiased estimator of β

  $E(\hat{\beta}_{WLS}) = E((X^TWX)^{-1}X^TWY) = (X^TWX)^{-1}X^TWX\beta = \beta$

- $\mathrm{Var}(\hat{\beta}_{WLS}) = \sigma^2(X^TWX)^{-1}$

- Estimation of σ^2

  $\hat{\sigma^2} = \dfrac{RSS}{n-p-1}$, where $RSS = (Y - X\hat{\beta})^TW(Y - X\hat{\beta})$

- The fitted value $\hat{y}_i$ $\hat{y}_i = {x_i}^T\hat{\beta}_{WLS}$

- The s.e. of the WLS estimate is
  $\sqrt{\hat{\sigma^2} \times (\text{jth diagonal element of the matrix } (X^TWX)^{-1})}$

Hypothesis Tests and CIs for β_j

- The t-statistic: $\dfrac{\hat{\beta}_{j,WLS} - \beta_j^0}{s.e.(\hat{\beta}_{j,WLS})} \sim t_{n-p-1}$ under $H_0 : \beta_j = \beta_j^0$ can be used the same way as OLS
- $1 - \alpha$ Confidence Interval:

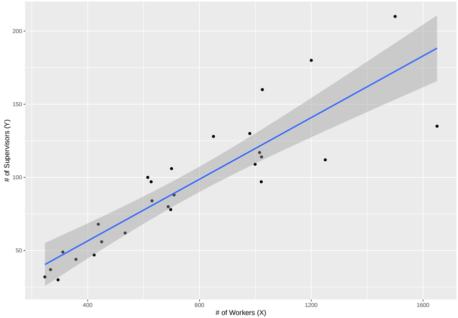
  $\hat{\beta}_{j,WLS} \pm t_{(n-p-1,\alpha/2)} s.e.(\hat{\beta}_{j,WLS})$

Revisit Supervisor Dataset

X = # of Supervised Workers

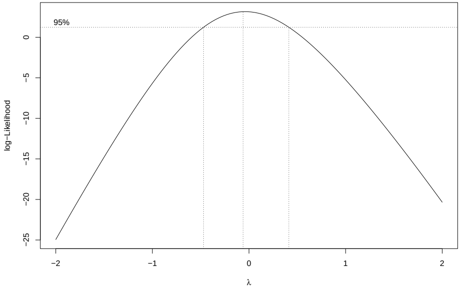
Y = # of Supervisors in 27 Industrial Establishments

```
supvis = read.table("supvis.txt", h=T)
ggplot(supvis, aes(x=X, y=Y))+geom_point()+geom_smooth(method=
  labs(x="# of Workers (X)", y="# of Supervisors (Y)")
```



Box-Cox Method for Supervisor/Employee Data

```
library(MASS)
boxcox(lm(Y ~ X + I(X^2), data=supvis))
```



- The middle dash line marks the optimal λ, the right and left dash line mark the 95% C.I. for the optimal λ.
- For the plot, we see the optimal λ is around 0.1, and the 95% C.I. contains 0. For simplicity, we pick $\lambda = 0$ and use $\log Y$.

σ_i is Proportional to Some Predictor x_i

Suppose the variance of the i th observation

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

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

- Let $w_i = \dfrac{1}{x_i^2}$
- 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 — WLS Approach

- 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))
```

```
##
## Call:
## lm(formula = Y ~ X, data = supvis, weights = 1/X^2)
##
## Weighted Residuals:
##       Min        1Q    Median        3Q       Max
## -0.041477 -0.013852 -0.004998  0.024671  0.035427
##
## Coefficients:
##             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 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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
```

CIs for β_j

```
##               Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) 3.8032958 4.569745381  0.8322774 4.131324e-01
## X           0.1209903 0.008998637 13.4454039 6.043603e-13
```

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

```
qt(1-0.025, df=25)
```

```
## [1] 2.059539
```

$\hat{\beta}_{j,WLS} \pm t_{(n-p-1,\alpha/2)} s.e.(\hat{\beta}_{j,WLS}) \approx 0.1209 \pm 2.0595 \times 0.008999 \approx (0.1025, 0.1395)$

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

CIs for β_j

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

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

```
##                  2.5 %     97.5 %
## (Intercept) -5.6082710 13.2148626
## X            0.1024573  0.1395233
```

Sum of Squares and Multiple R^2 for WLS

- $SST = SSR + SSE$ remains valid
- $SST = \sum_i w_i(y_i - \bar{y}_w)^2, \quad$ where $\bar{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i}$
- $SSR = \sum_i w_i(\hat{y}_i - \bar{y}_w)^2$
- $SSE = RSS = \sum_i (y_i - \hat{y}_i)^2$
- df of SS: same as 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
- $MSE = SSE/(n - p - 1) = \hat{\sigma}^2$
  - σ can be extracted from the summary table
  - The estimate of $\hat{\sigma}_i^2 = \sigma^2/w_i$

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
- H_0: 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 0.012842
## 2     24 0.011598  1 0.0012444 2.5751 0.1216
```

Residuals for WLS in R

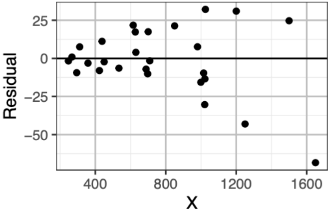
- `model$res` give the raw residuals $e_i = y_i - \hat{y}_i$, which are NOT adjusted by weights
- `hatvalues(model)` gives the leverage h_{ii}, which is the i th diagonal element of the hat matrix

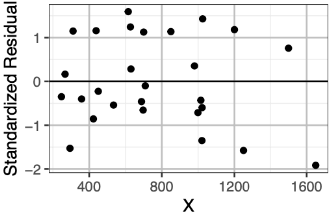
$$H = X(X^T WX)^{-1} X^T W$$

- $\mathrm{Var}(e_i) = \sigma_i^2(1 - h_{ii})$ where h_{ii} = leverage, and $\sigma_i^2 = \sigma^2 / w_i$
- `rstandard(model)` gives internally Standardized residuals
- `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
## 1 148.9917 134.2599 163.7235
```

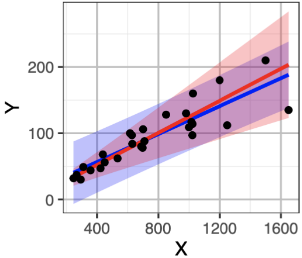
```
predict(lmwls, data.frame(X=1200), weights=1/1200^2,
        interval="prediction")
```

```
##        fit      lwr      upr
## 1 148.9917 91.07202 206.9113
```

- 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 with X