ch_4 weighted least squares

Jack Wright

11/3/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at C:/Users/jwright/Documents/GitHub/business_analytics/Modern_approach_with_r
```

coping with nonconstant error variance *weighted least squares* (WLS)

straight line regression based on weighted least squares

consider standard regression model

y=beta+beta*x+e

e have mean 0 but variance σ^2/w_i when w_i is very large than the variance of e is close to zero. In this situation the estimates of regression parameters should be such that the fitted line at x_i should be very close to y_i (close to the data.

i think this means that the specific variance at e_i is a product of the variance of the system divided by its weight.

when w_i is very small, then the variance of e_i is very large.

We need to take into account the weights w_i when estimateing the regression parameters $\beta_0\ and\ \beta_1$

achieved by considering the following weighted version of the residual sum of squares

$$WRSS = \sum w_i(y_i - \hat{y_{wi}})^2 = \sum w_i(y_i - b_o - b_1x_1)^2$$

WRSS, the larger the value of w_i the more the ith case is taken into account.

to get the weighted least squares estimates, we seek the values of b_0 and b_1 to minimize WRSS.

MATH

weighted least squares estimate

$$\hat{\beta_{0W}} = \frac{\sum w_i y_i}{\sum w_i} - \hat{\beta_{1W}} \frac{\sum w_i x_i}{\sum w_i} = \bar{y_w} - \hat{\beta_{1w}} \bar{x_w}$$

EX:

developing bid on contract cleaning

develop a regression to model the relationship between number of rooms cleaned (Y) and number of crews(X) and predict the number of rooms that can be cleaned by 4 and 16 crews.

x-var: number of crews is discrete.

y-var: multiple measurements of the Y-variable at each value of x.

in this case it is possible to directly calculate the standard deviation of Y at each value of x.

so

$$Y_i = \beta_0 + \beta_1 x_1 + e_i$$

where e_i have mean - but variance $sigma^2/w_i$

in this case we take

$$w_i = \frac{1}{Standard\ Deviation(Y_i)^2}$$

(are we forcing the total variance to be 1? and calucluating what weights are needed to make this happen?)

```
file<-here('data','cleaningwtd.txt')
df<-read.table(file,header=TRUE)

summary(lm(formula=Rooms~Crews,weights=1/StdDev^2,data=df))
```

```
##
## Call:
## lm(formula = Rooms ~ Crews, data = df, weights = 1/StdDev^2)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43184 -0.82013  0.03909  0.69029  2.01030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8095     1.1158   0.725    0.471
## Crews         3.8255     0.1788  21.400   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9648 on 51 degrees of freedom
## Multiple R-squared:  0.8998, Adjusted R-squared:  0.8978
## F-statistic:   458 on 1 and 51 DF,  p-value: < 2.2e-16
```

I think we are forcing the variance to be 1… I keep putting in different varainces and the residual standard error is about the square of it.

Prediction intervals for weighted least squares

theres a function in R for predicting a weight, i guess it just uses the variance of the model.

leverage for weighted least squares

ith fitted or predicted value from weighted least squares (just uses the predicted betas based on the weights?)

reality check: if all weights are equal then WLS=LS

using least squares to calculate weighted least squares

if you multiply through $\sqrt(w_i)$ the weights will cancel out in the variance of the error, leaving you with just the regular unweighted variance.

So we can calculate the weighted least squares fit of the model (this is a multiple linear regression, hence 2 x's)

$$Y_{newi} = \sqrt(w_i)Y_i$$

$$x_{1NEWi} = \sqrt{w_i}$$

$$x_{2NEWi} = \sqrt{w_i}$$

$$e_{NEWi} = \sqrt{w_i}e_i$$

so basically you can account for the weights by baking them into a new model, modifying them by that squared weight

EX:

contract cleaning bid

recall

$$w_i = \frac{1}{Standard\ Deviation(Y_i)^2}$$

use the estimated or sample standard deviations from the data to produce weights

the results are the same if you calculate new values including the weights

Defining residuals for weighted least squares

use of weighted least squares

Used in an important special case when Y_i is the average or the median of n observatiions

$$Var(Y_i) \approx \frac{1}{n_i}$$

in this case we take the weight i to be n_i (number of observations at that level?)