Weighted Least Squares Example

M/STAT 501: Fall 2022

Professor Ratings

Bleske-Rechek and Fritsch (2011) analyzed a data set of the ratings of 366 instructors at one large campus in the Midwest. Each instructor in the data had at least 10 ratings over a several year period. Students provided ratings from 1 (worst) to 5 (best). These data are built into R in the `alr4` library.

```
library(alr4)
data(Rateprof)
```

Let Y_{ij} be the quality rating of the ith instructor by the jth student, $j = 1, \ldots, n_i$, and $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ be the mean quality rating for the ith instructor. Similarly, let x_{1ij} and x_{2ij} be the easiness and helpfulness ratings, respectively, of the ith instructor by the jth student, with mean easiness and mean helpfulness for the ith instructor denoted by $\bar{x}_{1i}$ and $\bar{x}_{2i}$. Note that the data set only reports mean ratings, not individual student's ratings.

Do in class:

1. Assume $E(Y_{ij}|X) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}$ and $Var(Y_{ij}|X) = \sigma^2$. Derive the expression for $E(\bar{Y}_i|X)$ and $Var(\bar{Y}_i|X)$.

$$\begin{aligned} E(\bar{Y}_i|X) &= \frac{1}{n_i}\sum_{j=1}^{n_i} E(Y_{ij}|X) \\ &= \frac{1}{n_i}\sum_{j=1}^{n_i} (\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}) \\ &= \beta_0 + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i} \end{aligned}$$

2. If we fit a linear model to $E(\bar{Y}_i|X)$, would it meet the constant variance assumption? Explain why or why not.

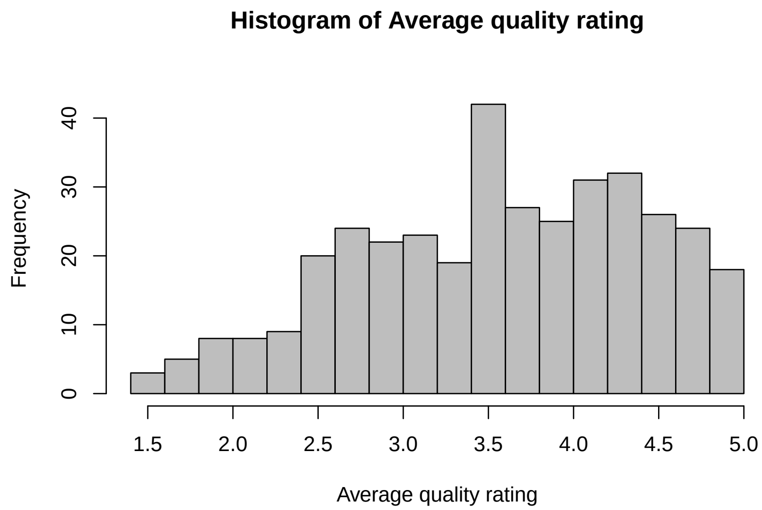
   No. The variance of Y_i depends on n_i. Assuming Y_{ij} is independent of Y_{ik} for $j \neq k$,

$$Var(\bar{Y}_i|X) = \frac{1}{n_i^2}\sum_{j=1}^{n_i} Var(Y_{ij}|X) = \frac{\sigma^2}{n_i}$$

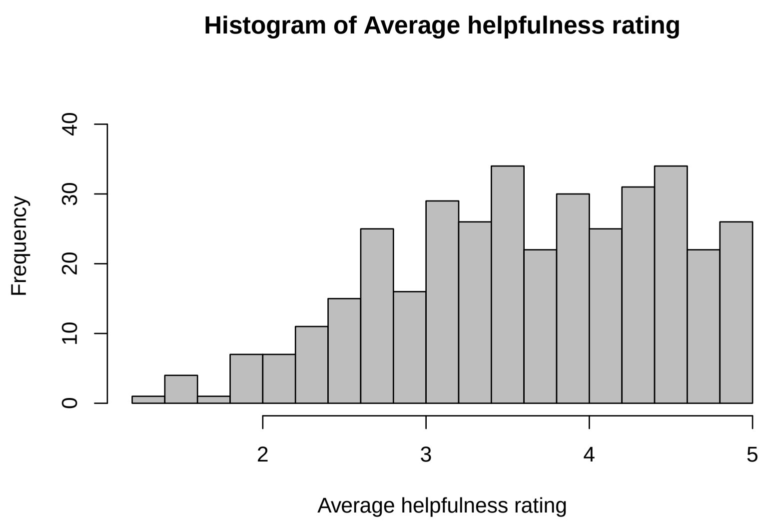
3. Let $\mathbf{Y} = \begin{pmatrix} \bar{Y}_1 & \bar{Y}_2 & \cdots & \bar{Y}_{366} \end{pmatrix}'$ be the response vector for this data set with variance-covariance matrix $\text{Var}(\mathbf{Y}) = \sigma^2 \mathbf{\Omega}$. Write out the elements in the first four rows and first four columns of $\mathbf{\Omega}$, i.e., report the 4×4 matrix that consists of elements in rows $1-4$ and columns $1-4$.

4. Generate plots to investigate the distributions and relationships between the three variables of interest.

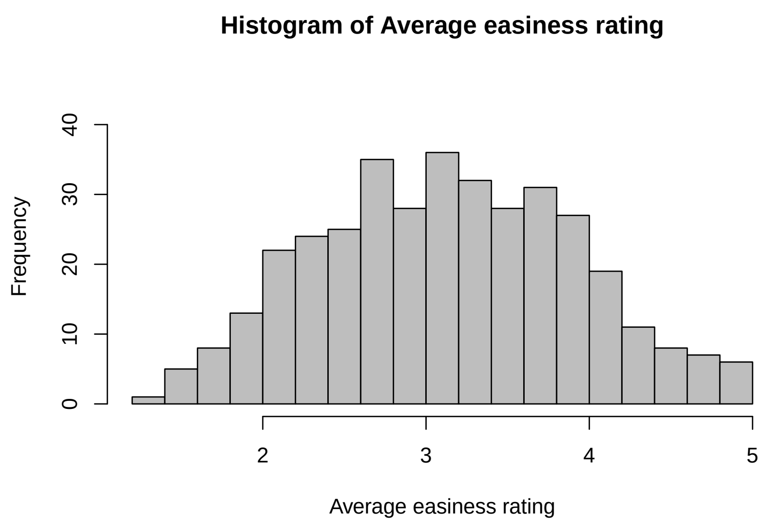
```
hist(Rateprof$quality, breaks=15, col="gray",
   xlab="Average quality rating",
   main ="Histogram of Average quality rating", ylim=c(0,45))
```



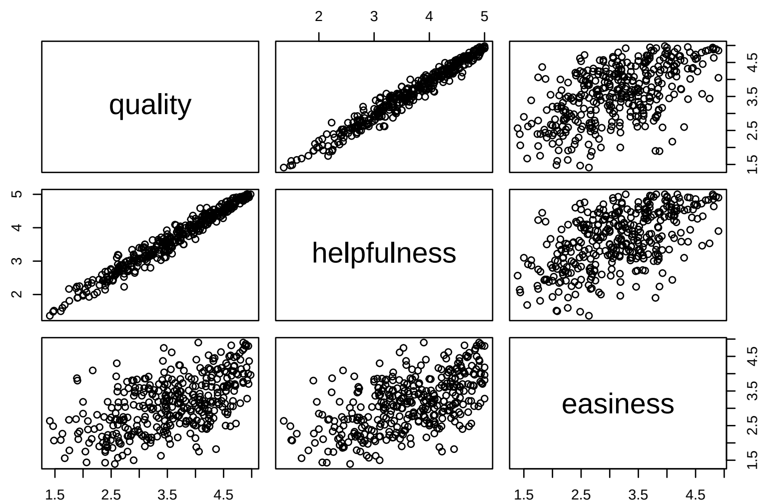

```
hist(Rateprof$helpfulness, breaks=15, col="gray",
   xlab="Average helpfulness rating",
   main ="Histogram of Average helpfulness rating", ylim=c(0,45))
```




```
hist(Rateprof$easiness, breaks=15, col="gray",
   xlab="Average easiness rating",
   main ="Histogram of Average easiness rating", ylim=c(0,45))
```




```
plot(Rateprof[,c(8,9,11)])
```




```
cor(Rateprof[,c(8,9,11)])
```

```
##               quality helpfulness  easiness
## quality     1.0000000   0.9810314 0.5651154
## helpfulness 0.9810314   1.0000000 0.5635184
## easiness    0.5651154   0.5635184 1.0000000
```

5. Fit the weighted least squares model. Write the equation of the fitted model, and interpret each of the three fitted coefficients in context of the problem.

```
mod <- lm(quality ~ easiness + helpfulness, weights = numRaters, data = Rateprof)
summary(mod)
```

```
##
## Call:
## lm(formula = quality ~ easiness + helpfulness, data = Rateprof,
##     weights = numRaters)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -2.88556 -0.50405  0.07072  0.49018  2.53349
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.05189    0.04010  -1.294    0.197
## easiness     0.01287    0.01288   0.999    0.318
## helpfulness  0.98674    0.01188  83.093   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8417 on 363 degrees of freedom
## Multiple R-squared:  0.965,  Adjusted R-squared:  0.9648
## F-statistic:  5001 on 2 and 363 DF,  p-value: < 2.2e-16
```

6. Fit the ordinary least squares fit to these data. How do the coefficients change? How does the residual standard error change? Why?

```r
mod.OLS <- lm(quality ~ easiness + helpfulness, data = Rateprof)
summary(mod.OLS)
```

```
##
## Call:
## lm(formula = quality ~ easiness + helpfulness, data = Rateprof)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51900 -0.09828  0.01194  0.09409  0.50101
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.009125   0.041047   0.222    0.824
## easiness    0.019387   0.013224   1.466    0.143
## helpfulness 0.965372   0.012210  79.063   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1623 on 363 degrees of freedom
## Multiple R-squared:  0.9626, Adjusted R-squared:  0.9624
## F-statistic:  4677 on 2 and 363 DF,  p-value: < 2.2e-16
```