# NTHU STAT 5410 - Linear Models Assignment 5 Report

105061110 周柏宇

1.

(a) Since , where and . The weighted least squares estimator is and its variance is . The reason why we should use WLS rather than OLS is that the variance of violets the assumption of equal variance from the Gauss-Markov theorem, which guarantees the variance of will be the smallest among all the linear unbiased estimators. However, we can fix the issue by fitting a transformed version of the original model , where . Then for the transformed model is exactly the for the original model.

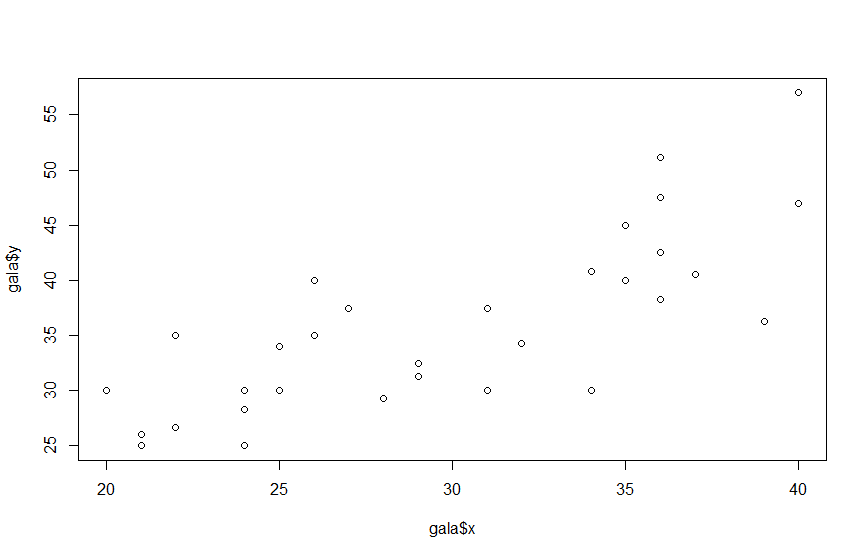
(b) In this case, , where and . The best linear unbiased estimator (BLUE) of β is and its variance is .

2.

(a)

> gala <- read.table("C:/Users/Thomas/Downloads/Linear\_models/hw5/E6.10.txt", header=T)

> plot(gala$x, gala$y)



Despite that the responses are averaged, we still can get a feeling of unequal variance among the observations due to their duplicates. If we assume the raw responses (i.e. ) are of constant variance . Then the sample variance will be the diagonal of the covariance matrix:

, where and . Therefore, .

> n <- gala[,2]

> x <- gala[,3]

> y <- gala[,4]

> WLS <- lm(y ~ x, weights=n)

> summary(WLS)

Call:

lm(formula = y ~ x, weights = n)

Weighted Residuals:

Min 1Q Median 3Q Max

-20.278 -7.661 -0.680 4.543 33.219

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.2932 4.5903 0.500 0.621

x 1.1319 0.1475 7.676 1.46e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.01 on 30 degrees of freedom

Multiple R-squared: 0.6626, Adjusted R-squared: 0.6514

F-statistic: 58.93 on 1 and 30 DF, p-value: 1.458e-08

(b) OLS estimators

> OLS <- lm(y ~ x)

> summary(OLS)

Call:

lm(formula = y ~ x)

Residuals:

Min 1Q Median 3Q Max

-10.2223 -3.0279 -0.6581 3.7721 10.7442

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.0325 4.5137 1.336 0.191

x 1.0056 0.1475 6.819 1.45e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.07 on 30 degrees of freedom

Multiple R-squared: 0.6079, Adjusted R-squared: 0.5948

F-statistic: 46.5 on 1 and 30 DF, p-value: 1.449e-07

Since the summary in WLS shows the weighted residuals as well as weighted RSE and weighted R-squared, we cannot compare the statistics directly. We need to use:

> WLS.RSS <- sum(WLS$residuals^2)

> WLS.RSS

[1] 790.0211

> OLS.RSS <- sum(OLS$residuals^2)

> OLS.RSS

[1] 771.0631

The RSS for OLS is indeed smaller since OLS minimizes the rather than .

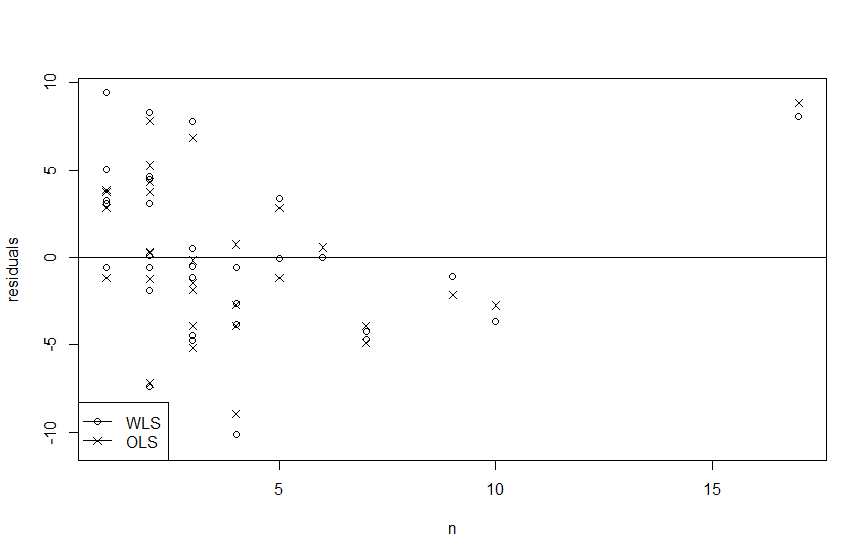
The residuals plot against n:

> plot(n, WLS$residuals, pch=1, ylab="residuals")

> points(n, OLS$residuals, pch=4)

> legend("bottomleft", legend = c("WLS","OLS"), pch = c(1,4), lty = c(1,1))

> abline(a=0, b=0)



We can see that for larger n (or lower sample variance), the residuals of WLS are slightly closer to zero compared to those of OLS.

(c)

> length(unique(x))

[1] 17

> length(x)

[1] 32

We can see that with 32 observations there are only 17 unique x. As the plot shows, we have several duplicates on some x. With these duplicates, we can estimate the pure error.

> sm <- lm(y ~ factor(x), weights=n) # saturated model

> summary(sm)

Call:

lm(formula = y ~ factor(x), weights = n)

Weighted Residuals:

Min 1Q Median 3Q Max

-17.638 -3.500 0.000 2.089 10.789

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 30.000 7.939 3.779 0.00182 \*\*

factor(x)21 -4.643 8.218 -0.565 0.58045

factor(x)22 0.850 8.575 0.099 0.92235

factor(x)24 -2.233 8.369 -0.267 0.79320

factor(x)25 3.333 8.575 0.389 0.70295

factor(x)26 8.333 9.167 0.909 0.37772

factor(x)27 7.500 9.723 0.771 0.45249

factor(x)28 -0.700 8.487 -0.082 0.93536

factor(x)29 1.900 8.421 0.226 0.82453

factor(x)31 3.750 8.876 0.422 0.67867

factor(x)32 4.300 8.487 0.507 0.61978

factor(x)34 9.257 8.487 1.091 0.29262

factor(x)35 13.000 8.697 1.495 0.15571

factor(x)36 18.483 8.103 2.281 0.03757 \*

factor(x)37 10.500 8.327 1.261 0.22657

factor(x)39 6.300 8.876 0.710 0.48874

factor(x)40 19.000 8.697 2.185 0.04520 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.939 on 15 degrees of freedom

Multiple R-squared: 0.8939, Adjusted R-squared: 0.7807

F-statistic: 7.898 on 16 and 15 DF, p-value: 0.0001177

> anova(WLS, sm)

Analysis of Variance Table

Model 1: y ~ x

Model 2: y ~ factor(x)

Res.Df RSS Df Sum of Sq F Pr(>F)

1 30 3006.20

2 15 945.47 15 2060.7 2.1796 0.07132 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

With the result of ANOVA suggests, we reject that the model is lack of fit.

3.

> gala <- read.table("C:/Users/Thomas/Downloads/Linear\_models/hw5/E6.11.txt", header=T)

> father <- gala[,1]

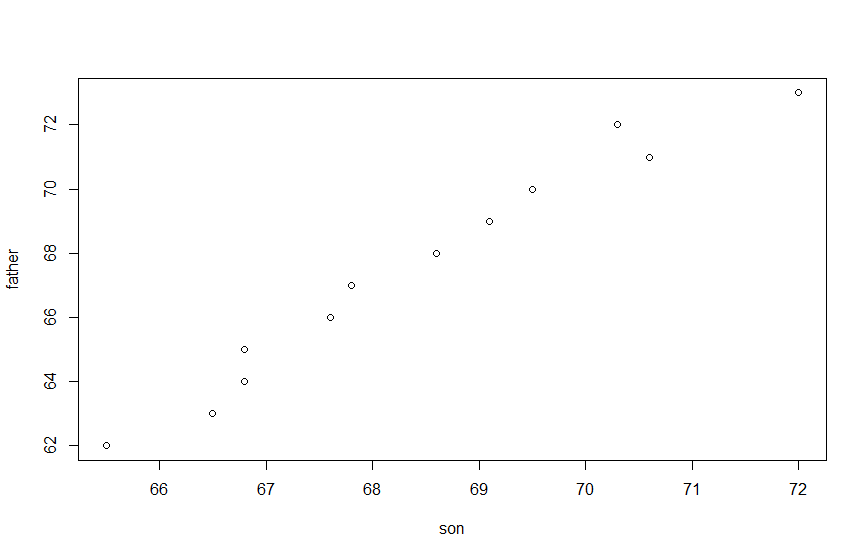
> son <- gala[,2]

> n\_father <- gala[,3]

Normally we would like to plot the height of son against the height of father due to the biological causation. If the number of sons for each height category were available, then we can recover the covariance matrix by setting its diagonal to the sample variance, which is inversely proportional to the number of kids used to do the average.

However, in this data set, we are provided with an extra variable – the number of fathers. In order to utilize the information, we plot

> plot(son, father)

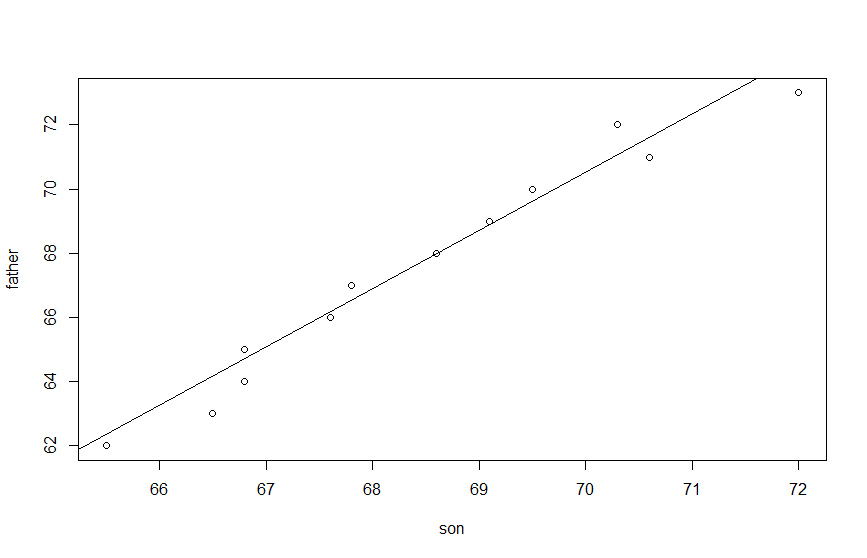


And fit a model with the number of fathers as weights

> fit1 <- lm(father ~ son, w=n\_father)

> abline(fit1)

> summary(fit1)



Call:

lm(formula = father ~ son, weights = n\_father)

Weighted Residuals:

Min 1Q Median 3Q Max

-2.8646 -2.4112 -0.2091 1.3596 2.6244

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -56.6269 7.7922 -7.267 2.70e-05 \*\*\*

son 1.8165 0.1138 15.957 1.93e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.125 on 10 degrees of freedom

Multiple R-squared: 0.9622, Adjusted R-squared: 0.9584

F-statistic: 254.6 on 1 and 10 DF, p-value: 1.926e-08

The fitted model is