# QUESTION 4.1

profsal <- read.csv('../data/prof\_sal\_third\_qu.csv', header=TRUE)

# here I just use 1/(numberof samples) as the weights

profsal$weights <- 1/profsal$sample

profsal

years sample thirdq weights

1 0 17 101300 0.05882353

2 2 33 111303 0.03030303

3 4 19 98000 0.05263158

4 6 25 124000 0.04000000

5 8 18 128475 0.05555556

6 12 60 117410 0.01666667

7 17 58 115825 0.01724138

8 22 31 134300 0.03225806

9 28 34 128066 0.02941176

10 34 19 164700 0.05263158

model.profsal <- lm(thirdq~years, data=profsal, weights=profsal$weights)

# lm seems to work just fine when passing in an additional vecor for weights

summary(model.profsal)

Call:

lm(formula = thirdq ~ years, data = profsal, weights = profsal$weights)

Weighted Residuals:

Min 1Q Median 3Q Max

-3148.9 -1528.5 -576.5 1766.0 2924.2

Coefficients:

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

(Intercept) 103925.6 5232.1 19.863 4.3e-08 \*\*\*

years 1517.9 305.8 4.964 0.0011 \*\*

---

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

Residual standard error: 2243 on 8 degrees of freedom

Multiple R-squared: 0.7549, Adjusted R-squared: 0.7242

F-statistic: 24.64 on 1 and 8 DF, p-value: 0.001102

# so the weighted least squares model is:

# thirdq = 1,517.9\*(years of experience) + 103,925.6

six.years <- 1517.9\*(6) + 103925.6

six.years

# the estimate for six years of experience is $113,003

> # QUESTION 5.1

> overdue <- read.table('../data/overdue.txt', header=TRUE)

> overdue$type <- c(rep(1,48),rep(0,48))

> head(overdue)

LATE BILL type

1 16 79 1

2 47 264 1

3 22 97 1

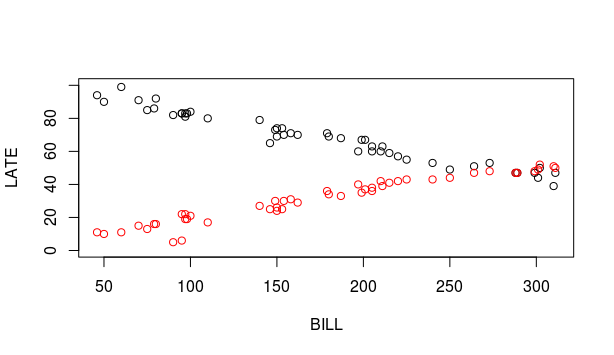
4 47 289 1

5 47 288 1

6 21 100 1

> plot(overdue$BILL[overdue$type==0],overdue$LATE[overdue$type==0],col=c("black"),ylab="LATE",xlab="BILL",ylim=c(0, 100))

> points(overdue$BILL[overdue$type==1],overdue$LATE[overdue$type==1],col=c("red"))

# visually, it looks like the type variable will play a role in the final model

> model.all <- lm(LATE~BILL, data=overdue)

> summary(model.all)

Call:

lm(formula = LATE ~ BILL, data = overdue)

Residuals:

Min 1Q Median 3Q Max

-45.846 -17.212 -0.793 19.007 47.774

Coefficients:

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

(Intercept) 51.98390 5.96405 8.716 9.84e-14 \*\*\*

BILL -0.01264 0.03128 -0.404 0.687

---

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

Residual standard error: 23.72 on 94 degrees of freedom

Multiple R-squared: 0.001734, Adjusted R-squared: -0.008885

F-statistic: 0.1633 on 1 and 94 DF, p-value: 0.687

# without considering type, the R squared is very low, indicating that this is not a very good model when we only consider the bill amount

> model.sep <-lm(LATE~BILL+type, data=overdue)

> summary(model.sep)

# as a quick test to tell if the type variable will be significant, we add in the variable to the model.

# with an R squared of about .63, it is much better

Call:

lm(formula = LATE ~ BILL + type, data = overdue)

Residuals:

Min 1Q Median 3Q Max

-27.7637 -11.4760 0.4037 12.4812 29.0765

Coefficients:

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

(Intercept) 70.68182 3.91286 18.064 <2e-16 \*\*\*

BILL -0.01264 0.01901 -0.665 0.508

type -37.39583 2.94375 -12.703 <2e-16 \*\*\*

---

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

Residual standard error: 14.42 on 93 degrees of freedom

Multiple R-squared: 0.635, Adjusted R-squared: 0.6272

F-statistic: 80.91 on 2 and 93 DF, p-value: < 2.2e-16

> model.new <- lm(LATE~BILL+type+type:BILL,data=overdue)

> summary(model.new)

Call:

lm(formula = LATE ~ BILL + type + type:BILL, data = overdue)

Residuals:

Min 1Q Median 3Q Max

-12.1211 -2.2163 0.0974 1.9556 8.6995

Coefficients:

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

(Intercept) 101.758184 1.198504 84.90 <2e-16 \*\*\*

BILL -0.190961 0.006285 -30.38 <2e-16 \*\*\*

type -99.548561 1.694940 -58.73 <2e-16 \*\*\*

BILL:type 0.356644 0.008888 40.12 <2e-16 \*\*\*

---

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

Residual standard error: 3.371 on 92 degrees of freedom

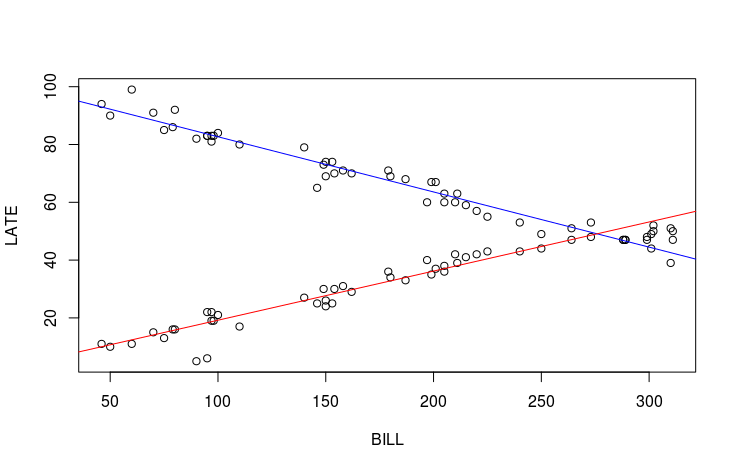
Multiple R-squared: 0.9803, Adjusted R-squared: 0.9796

F-statistic: 1524 on 3 and 92 DF, p-value: < 2.2e-16

> plot(LATE~BILL,data=overdue)

> abline(model.new,col=c("blue"))

> abline(101.75-99.55,-0.19+0.36,col=c("red"))



> # the test to check if there is a difference in the mean overdue dates based on the type of account

> model.anova <- lm(LATE~type, data=overdue)

> anova(model.anova)

Analysis of Variance Table

Response: LATE

Df Sum Sq Mean Sq F value Pr(>F)

type 1 33563 33563 162.34 < 2.2e-16 \*\*\*

Residuals 94 19434 207

---

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

# the p value is much less than 0.05, so it is significant

# This means that the mean number of overdue days is not equal for commercial and residental

> # QUESTION 5.2

> houston <- read.csv("../data/HoustonChronicle.csv",header=TRUE)

> head(houston)

District X.Repeating.1st.Grade X.Low.income.students Year County

1 Alvin 4.1 49.7 2004 Brazoria

2 Alvin 5.8 41.1 1994 Brazoria

3 Angleton 7.1 44.2 2004 Brazoria

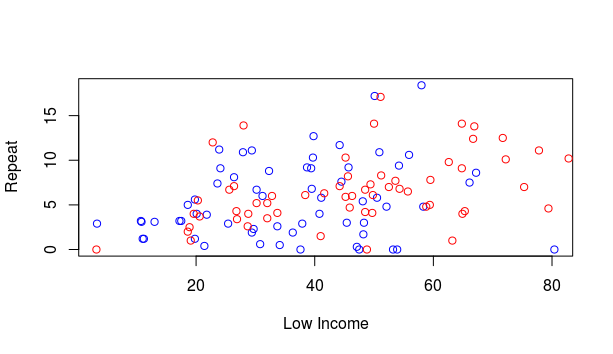
4 Angleton 6.7 30.2 1994 Brazoria

5 Brazosport 7.3 49.4 2004 Brazoria

6 Brazosport 2.6 33.7 1994 Brazoria

> plot(x=houston$X.Low.income.students, y=houston$X.Repeating.1st.Grade)

> plot(houston$X.Low.income.students[houston$Year==1994],houston$X.Repeating.1st.Grade[houston$Year==1994],col=c("blue"),ylab="Repeat",xlab="Low Income")

> points(houston$X.Low.income.students[houston$Year==2004],houston$X.Repeating.1st.Grade[houston$Year==2004],col=c("red"))

> # PART A

> grade.model1 <- lm(X.Repeating.1st.Grade ~ X.Low.income.students, data = houston)

> summary(grade.model1)

# The fit appears good with a significant slope.

Call:

lm(formula = X.Repeating.1st.Grade ~ X.Low.income.students, data = houston)

Residuals:

Min 1Q Median 3Q Max

-8.9845 -2.5072 -0.4184 1.8505 11.1067

Coefficients:

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

(Intercept) 2.91419 0.83836 3.476 0.000709 \*\*\*

X.Low.income.students 0.07550 0.01823 4.141 6.47e-05 \*\*\*

---

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

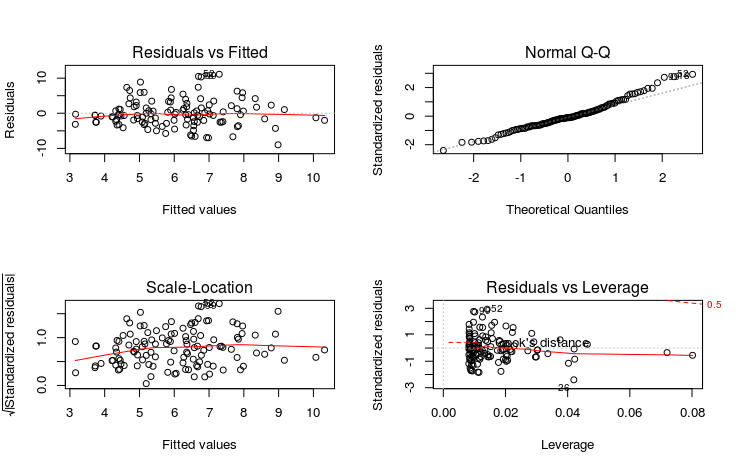
Residual standard error: 3.821 on 120 degrees of freedom

Multiple R-squared: 0.125, Adjusted R-squared: 0.1177

F-statistic: 17.14 on 1 and 120 DF, p-value: 6.472e-05

> par(mfrow=c(2,2))

> plot(grade.model1)



> # PART B

> grade.model2 <- lm(X.Repeating.1st.Grade ~ Year, data = houston)

> summary(grade.model2)

Call:

lm(formula = X.Repeating.1st.Grade ~ Year, data = houston)

Residuals:

Min 1Q Median 3Q Max

-6.6787 -2.6537 -0.6262 2.5750 12.9262

Coefficients:

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

(Intercept) -234.78689 146.20650 -1.606 0.111

Year 0.12049 0.07314 1.647 0.102

Residual standard error: 4.039 on 120 degrees of freedom

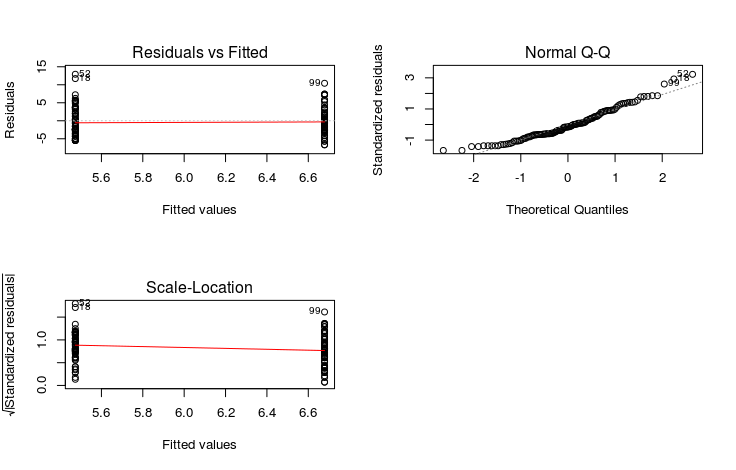
Multiple R-squared: 0.02212, Adjusted R-squared: 0.01397

F-statistic: 2.714 on 1 and 120 DF, p-value: 0.1021

> plot(grade.model2)

hat values (leverages) are all = 0.01639344

and there are no factor predictors; no plot no. 5



> # there is no visually discernible difference between the two populations

> t.test(houston$X.Repeating.1st.Grade~houston$Year)

# a quick t test also shows that with p value of over .10, there is not a significant difference in the two populations

Welch Two Sample t-test

data: houston$X.Repeating.1st.Grade by houston$Year

t = -1.6474, df = 117.81, p-value = 0.1021

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.6533051 0.2434691

sample estimates:

mean in group 1994 mean in group 2004

5.473770 6.678689

> # PART C

> grade.model3 <- lm(X.Repeating.1st.Grade ~ Year + X.Low.income.students, data = houston)

> summary(grade.model3)

Call:

lm(formula = X.Repeating.1st.Grade ~ Year + X.Low.income.students,

data = houston)

Residuals:

Min 1Q Median 3Q Max

-8.6768 -2.5451 -0.4769 1.6624 11.3469

Coefficients:

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

(Intercept) -73.54333 145.12258 -0.507 0.613256

Year 0.03831 0.07272 0.527 0.599274

X.Low.income.students 0.07248 0.01917 3.782 0.000245 \*\*\*

---

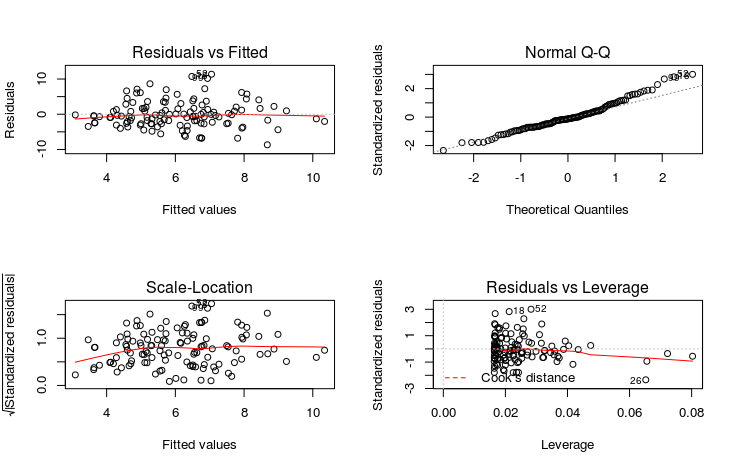
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.832 on 119 degrees of freedom

Multiple R-squared: 0.127, Adjusted R-squared: 0.1124

F-statistic: 8.659 on 2 and 119 DF, p-value: 0.0003083

> plot(grade.model3)



> grade.model4 <- lm(X.Repeating.1st.Grade ~ X.Low.income.students + Year + X.Low.income.students:Year, data = houston)

> summary(grade.model4)

Call:

lm(formula = X.Repeating.1st.Grade ~ X.Low.income.students +

Year + X.Low.income.students:Year, data = houston)

Residuals:

Min 1Q Median 3Q Max

-8.1606 -2.6121 -0.5576 1.7495 11.6014

Coefficients:

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

(Intercept) 80.950443 352.012893 0.230 0.819

X.Low.income.students -3.734749 7.898023 -0.473 0.637

Year -0.038956 0.176109 -0.221 0.825

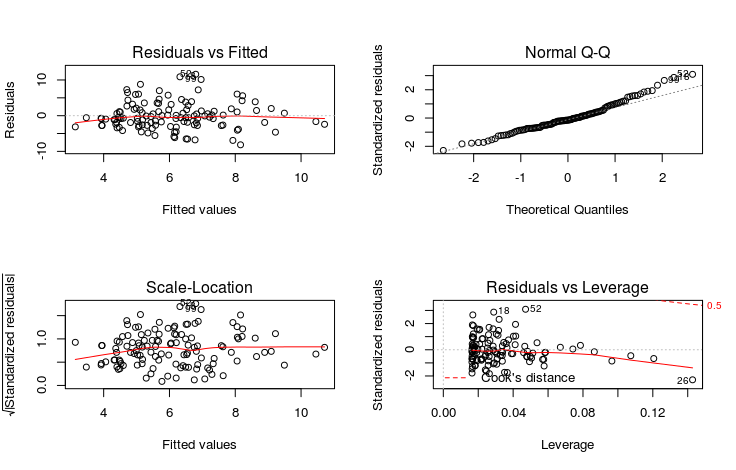
X.Low.income.students:Year 0.001903 0.003949 0.482 0.631

Residual standard error: 3.845 on 118 degrees of freedom

Multiple R-squared: 0.1288, Adjusted R-squared: 0.1066

F-statistic: 5.813 on 3 and 118 DF, p-value: 0.0009689

> plot(grade.model4)



> # There is not much to show for the association between the low-income students and failure rates between the years

> anova(grade.model3,grade.model4)

Analysis of Variance Table

Model 1: X.Repeating.1st.Grade ~ Year + X.Low.income.students

Model 2: X.Repeating.1st.Grade ~ X.Low.income.students + Year + X.Low.income.students:Year

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

1 119 1747.8

2 118 1744.4 1 3.4351 0.2324 0.6307

> anova(grade.model1,grade.model4)

Analysis of Variance Table

Model 1: X.Repeating.1st.Grade ~ X.Low.income.students

Model 2: X.Repeating.1st.Grade ~ X.Low.income.students + Year + X.Low.income.students:Year

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

1 120 1751.9

2 118 1744.4 2 7.512 0.2541 0.7761

# It is reasonable to not consider the Year in the model, but including low-income is better than not having it.

# The year does not influence the model, but there is an association between low-income students and higher grade repetitions.

> # QUESTION 5.3

> latour <- read.table("../data/Latour.txt", header=TRUE)

> head(latour)

Vintage Quality EndofHarvest Rain

1 1961 5 28 0

2 1962 4 50 0

3 1963 1 53 1

4 1964 3 38 0

5 1965 1 46 1

6 1966 4 40 0

> # PART A

> latour.model1 <- lm(Quality ~ EndofHarvest + Rain + Rain:EndofHarvest, data = latour)

> summary(latour.model1)

Call:

lm(formula = Quality ~ EndofHarvest + Rain + Rain:EndofHarvest,

data = latour)

Residuals:

Min 1Q Median 3Q Max

-1.6833 -0.5703 0.1265 0.4385 1.6354

Coefficients:

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

(Intercept) 5.16122 0.68917 7.489 3.95e-09 \*\*\*

EndofHarvest -0.03145 0.01760 -1.787 0.0816 .

Rain 1.78670 1.31740 1.356 0.1826

EndofHarvest:Rain -0.08314 0.03160 -2.631 0.0120 \*

---

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

Residual standard error: 0.7578 on 40 degrees of freedom

Multiple R-squared: 0.6848, Adjusted R-squared: 0.6612

F-statistic: 28.97 on 3 and 40 DF, p-value: 4.017e-10

> latour.model2 <- lm(Quality ~ EndofHarvest + Rain, data = latour)

> summary(latour.model2)

Call:

lm(formula = Quality ~ EndofHarvest + Rain, data = latour)

Residuals:

Min 1Q Median 3Q Max

-1.4563 -0.7366 0.1430 0.6413 1.7652

Coefficients:

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

(Intercept) 6.14633 0.61896 9.930 1.8e-12 \*\*\*

EndofHarvest -0.05723 0.01564 -3.660 0.000713 \*\*\*

Rain -1.62219 0.25478 -6.367 1.3e-07 \*\*\*

---

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

Residual standard error: 0.8107 on 41 degrees of freedom

Multiple R-squared: 0.6303, Adjusted R-squared: 0.6123

F-statistic: 34.95 on 2 and 41 DF, p-value: 1.383e-09

> anova(latour.model2,latour.model1)

# the impact that rain causes on the harvest and wine quality is significant

Analysis of Variance Table

Model 1: Quality ~ EndofHarvest + Rain

Model 2: Quality ~ EndofHarvest + Rain + Rain:EndofHarvest

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

1 41 26.945

2 40 22.971 1 3.9749 6.9218 0.01203 \*

---

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

> # PART B

> coefficients(latour.model1)

(Intercept) EndofHarvest Rain EndofHarvest:Rain

5.16121899 -0.03144552 1.78669768 -0.08313781

> # i

> -1/coefficients(latour.model1)

(Intercept) EndofHarvest Rain EndofHarvest:Rain

-0.1937527 31.8010292 -0.5596918 12.0282216

> # without rain, the days delayed give a point decrease of 31.8

> # ii

> -1/(coefficients(latour.model1)[2] + coefficients(latour.model1)[4])

EndofHarvest

8.727273

> # with rain, the days delayed give a point decrease of 8.7