Proj1markdown

Jared Yu

November 13, 2017

#set working directory to where the file can be found  
setwd("C:/Users/jyqq9/Desktop/STA 108/Project 1")  
  
#read the data file and place in a variable  
mydata = read.table("UN.txt", header = T) #header=F if no first row.  
  
#Display some of the data  
head(mydata)

## Locality Fertility PPgdp  
## 1 Afghanistan 6.80 98  
## 2 Albania 2.28 1317  
## 3 Algeria 2.80 1784  
## 4 Angola 7.20 739  
## 5 Argentina 2.44 7163  
## 6 Australia 1.70 18788

#Display all of the data  
mydata

## Locality Fertility PPgdp  
## 1 Afghanistan 6.80 98  
## 2 Albania 2.28 1317  
## 3 Algeria 2.80 1784  
## 4 Angola 7.20 739  
## 5 Argentina 2.44 7163  
## 6 Australia 1.70 18788  
## 7 Austria 1.28 23260  
## 8 Azerbaijan 2.10 695  
## 9 Bahamas 2.29 14856  
## 10 Bahrain 2.66 12012  
## 11 Bangladesh 3.46 345  
## 12 Barbados 1.50 9255  
## 13 Belgium 1.66 22351  
## 14 Belize 3.15 3123  
## 15 Benin 5.66 361  
## 16 Bermuda 1.67 44579  
## 17 Bhutan 5.02 241  
## 18 Bolivia 3.82 935  
## 19 Botswana 3.70 2872  
## 20 Brazil 2.21 2888  
## 21 Brunei 2.48 12435  
## 22 Burkina.Faso 6.68 203  
## 23 Burundi 6.80 107  
## 24 Cambodia 4.77 233  
## 25 Cameroon 4.61 557  
## 26 Canada 1.48 22385  
## 27 Cape.Verde 3.30 1259  
## 28 Central.African.Rep 4.92 242  
## 29 Chad 6.65 127  
## 30 Chile 2.35 3992  
## 31 China 1.83 918  
## 32 Hong.Kong 1.00 23499  
## 33 Macao 1.10 14281  
## 34 Colombia 2.62 1900  
## 35 Comoros 4.90 278  
## 36 Congo 6.29 779  
## 37 Cook.Islands 3.50 4388  
## 38 Costa.Rica 2.28 4148  
## 39 Cote.dIvoire 4.73 637  
## 40 Croatia 1.65 4558  
## 41 Cuba 1.55 2545  
## 42 Cyprus 1.90 11449  
## 43 Czech.Rep 1.16 5501  
## 44 Dem.Rep.Congo 6.70 138  
## 45 Denmark 1.77 30265  
## 46 Djibouti 5.70 819  
## 47 Dominican.Rep 2.71 2500  
## 48 Ecuador 2.76 1425  
## 49 Egypt 3.29 1390  
## 50 El.Salvador 2.88 2189  
## 51 Equatorial.Guinea 5.89 3940  
## 52 Eritrea 5.43 177  
## 53 Estonia 1.22 4010  
## 54 Ethiopia 6.14 90  
## 55 Fiji 2.88 2046  
## 56 Finland 1.73 23456  
## 57 France 1.89 21990  
## 58 Fr.Guiana 3.33 7737  
## 59 Fr.Polynesia 2.44 13891  
## 60 Gabon 3.99 3379  
## 61 Gambia 4.70 300  
## 62 Germany 1.35 22418  
## 63 Ghana 4.11 265  
## 64 Greece 1.27 10727  
## 65 Guadeloupe 2.10 10323  
## 66 Guatemala 4.41 1717  
## 67 Guinea 5.82 375  
## 68 Guinea-Bissau 7.10 174  
## 69 Guyana 2.31 936  
## 70 Haiti 3.98 431  
## 71 Honduras 3.72 960  
## 72 Hungary 1.20 5209  
## 73 Iceland 1.95 27281  
## 74 India 3.01 467  
## 75 Indonesia 2.35 678  
## 76 Iran 2.33 5645  
## 77 Ireland 1.90 26725  
## 78 Israel 2.70 18816  
## 79 Italy 1.23 18928  
## 80 Jamaica 2.36 2990  
## 81 Japan 1.32 32540  
## 82 Jordan 3.57 1726  
## 83 Kazakhstan 1.95 1441  
## 84 Kenya 4.00 367  
## 85 Kiribati 3.80 468  
## 86 S.Korea 2.02 8955  
## 87 Kuwait 2.66 16782  
## 88 Kyrgyzstan 2.64 306  
## 89 Laos 4.78 324  
## 90 Latvia 1.10 3212  
## 91 Lebanon 2.18 5087  
## 92 Lesotho 3.84 419  
## 93 Liberia 6.80 256  
## 94 Libya 3.02 5099  
## 95 Liechtenstein 1.64 34504  
## 96 Lithuania 1.25 3442  
## 97 Luxembourg 1.73 43041  
## 98 Madagascar 5.70 278  
## 99 Malawi 6.10 129  
## 100 Malaysia 2.90 3748  
## 101 Maldives 5.33 1947  
## 102 Mali 7.00 200  
## 103 Malta 1.77 9245  
## 104 Marshall.Is 3.68 1938  
## 105 Martinique 1.90 10723  
## 106 Mauritania 5.79 353  
## 107 Mauritius 1.95 3787  
## 108 Mexico 2.50 6150  
## 109 Micronesia 3.80 2215  
## 110 Mongolia 2.42 417  
## 111 Morocco 2.75 1145  
## 112 Mozambique 5.63 196  
## 113 Namibia 4.56 1639  
## 114 Nepal 4.26 226  
## 115 Netherlands 1.72 23785  
## 116 Neth.Antilles 2.05 12149  
## 117 New.Caledonia 2.45 15750  
## 118 New.Zealand 2.01 13185  
## 119 Nicaragua 3.75 489  
## 120 Niger 8.00 176  
## 121 Nigeria 5.42 435  
## 122 Norway 1.80 36445  
## 123 Oman 4.96 7421  
## 124 Pakistan 5.08 418  
## 125 Palau 3.00 6179  
## 126 Panama 2.70 3391  
## 127 Papua.New.Guinea 4.09 545  
## 128 Paraguay 3.84 1286  
## 129 Peru 2.86 2053  
## 130 Philippines 3.18 924  
## 131 Poland 1.26 4657  
## 132 Portugal 1.45 10944  
## 133 Puerto.Rico 1.89 19083  
## 134 Qatar 3.22 30493  
## 135 Reunion 2.30 9188  
## 136 Russia 1.14 2139  
## 137 Rwanda 5.74 205  
## 138 Saint.Kitts.and.Nevis 2.41 8426  
## 139 Saint.Lucia 2.27 4994  
## 140 St.Vincent/Grenadines 2.23 2940  
## 141 Samoa 4.12 1402  
## 142 Sao.Tome.and.Principe 3.99 312  
## 143 Saudi.Arabia 4.53 7724  
## 144 Senegal 4.97 479  
## 145 Serbia.and.Montenegro. 1.65 1008  
## 146 Seychelles 2.00 7850  
## 147 Sierra.Leone 6.50 164  
## 148 Singapore 1.36 20755  
## 149 Slovakia 1.28 3767  
## 150 Slovenia 1.14 9463  
## 151 Solomon.Islands 4.42 760  
## 152 Somalia 7.25 110  
## 153 South.Africa 2.61 2550  
## 154 Spain 1.15 14234  
## 155 Sri.Lanka 2.01 827  
## 156 Sudan 4.39 376  
## 157 Suriname 2.45 1965  
## 158 Swaziland 4.54 1204  
## 159 Sweden 1.64 23680  
## 160 Switzerland 1.41 34449  
## 161 Syria 3.32 4976  
## 162 Tajikistan 3.06 172  
## 163 Thailand 1.93 1858  
## 164 Macedonia 1.90 1723  
## 165 Timor-Leste 3.85 438  
## 166 Togo 5.33 273  
## 167 Tonga 3.71 1284  
## 168 Trinidad.and.Tobago 1.55 6817  
## 169 Tunisia 2.01 2077  
## 170 Turkey 2.43 2136  
## 171 Turkmenistan 2.70 1263  
## 172 Uganda 7.10 239  
## 173 United.Arab.Emirates 2.82 19816  
## 174 United.Kingdom 1.60 24186  
## 175 Tanzania 5.11 263  
## 176 USA 2.11 34788  
## 177 Uruguay 2.30 5514  
## 178 Uzbekistan 2.44 418  
## 179 Vanuatu 4.13 1085  
## 180 Venezuela 2.72 5009  
## 181 Viet.Nam 2.30 416  
## 182 Yemen 7.01 431  
## 183 Zambia 5.64 345  
## 184 Zimbabwe 3.90 703

#Set x and y to mydata's height and weight  
x=mydata$PPgdp  
y=mydata$Fertility  
  
#Display some of the data for mydata  
str(x)

## int [1:184] 98 1317 1784 739 7163 18788 23260 695 14856 12012 ...

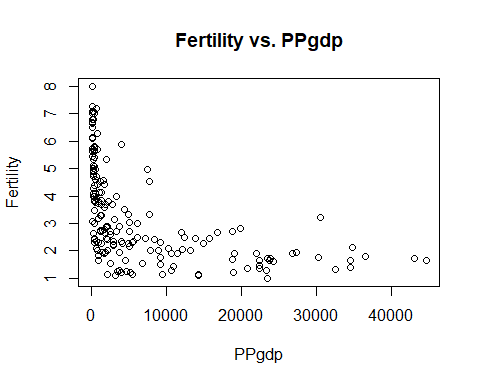
str(y)

## num [1:184] 6.8 2.28 2.8 7.2 2.44 1.7 1.28 2.1 2.29 2.66 ...

str(mydata)

## 'data.frame': 184 obs. of 3 variables:  
## $ Locality : Factor w/ 184 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Fertility: num 6.8 2.28 2.8 7.2 2.44 1.7 1.28 2.1 2.29 2.66 ...  
## $ PPgdp : int 98 1317 1784 739 7163 18788 23260 695 14856 12012 ...

#1  
#plot x and y, label the axes  
plot(x,y,xlab="PPgdp",ylab="Fertility",main="Fertility vs. PPgdp")



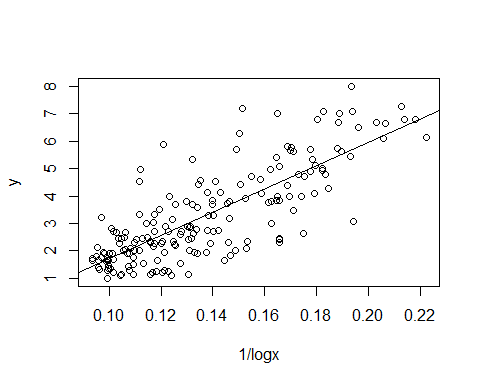
#2  
#create variables for the log transformations  
xlog = 1/log(x)  
  
plot(xlog, y, xlab="1/logx",ylab="y")  
  
#3a  
#xbar and ybar  
ybar=mean(y)  
xbar=mean(xlog)  
  
#n, the number of variables  
n=length(x)  
  
#betahat1 and betahat0  
betahat1=sum((xlog-xbar)\*(y-ybar))/sum((xlog-xbar)^2)  
betahat0=ybar-betahat1\*xbar  
betahat0

## [1] -2.560345

betahat1

## [1] 42.57107

#plot a straight line for the data using slope and intercept  
abline(a = betahat0, b = betahat1)



#yhat  
yhat = betahat0+betahat1\*xlog  
  
#SSR  
SSR = sum((yhat-ybar)^2)  
  
#SSE  
SSE = sum((y-yhat)^2)  
  
#SSTO  
SSTO = sum((y-ybar)^2)  
  
#R^2 and rsquared  
R2 = SSR/SSTO  
R2

## [1] 0.6360678

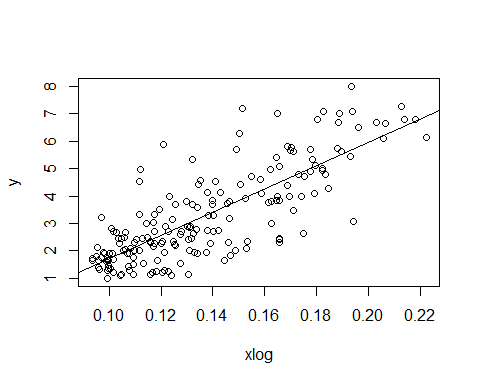
rsquared = (1 - SSE/SSTO)  
rsquared

## [1] 0.6360678

#3b  
#lm function  
model = lm(y~xlog)  
summary(model)

##   
## Call:  
## lm(formula = y ~ xlog)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6499 -0.5764 -0.0475 0.5886 3.3154   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.5603 0.3363 -7.612 1.41e-12 \*\*\*  
## xlog 42.5711 2.3869 17.835 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.031 on 182 degrees of freedom  
## Multiple R-squared: 0.6361, Adjusted R-squared: 0.6341   
## F-statistic: 318.1 on 1 and 182 DF, p-value: < 2.2e-16

plot(xlog,y)  
#plot fitted line  
abline(model$coefficients)



#3c  
#matrix manipulation  
head(mydata)

## Locality Fertility PPgdp  
## 1 Afghanistan 6.80 98  
## 2 Albania 2.28 1317  
## 3 Algeria 2.80 1784  
## 4 Angola 7.20 739  
## 5 Argentina 2.44 7163  
## 6 Australia 1.70 18788

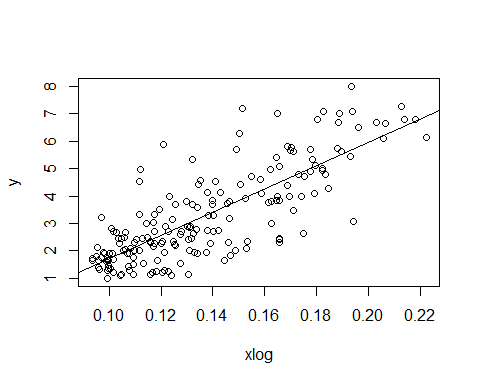
X=as.matrix(cbind(rep(1,n),1/log(mydata[,3])))  
View(X)  
str(X)

## num [1:184, 1:2] 1 1 1 1 1 1 1 1 1 1 ...

XTX = t(X)%\*%X  
XTXinv = solve(XTX)  
Y = as.matrix(mydata[,2])  
View(Y)  
XTY = t(X)%\*%Y  
betahatMatrix = XTXinv%\*%XTY  
betahatMatrix

## [,1]  
## [1,] -2.560345  
## [2,] 42.571070

plot(xlog,y)  
#draw fitted matrix line  
abline(betahatMatrix)



Yhat = X%\*%betahatMatrix  
head(Yhat)

## [,1]  
## [1,] 6.724578  
## [2,] 3.366205  
## [3,] 3.125947  
## [4,] 3.884643  
## [5,] 2.235485  
## [6,] 1.765555

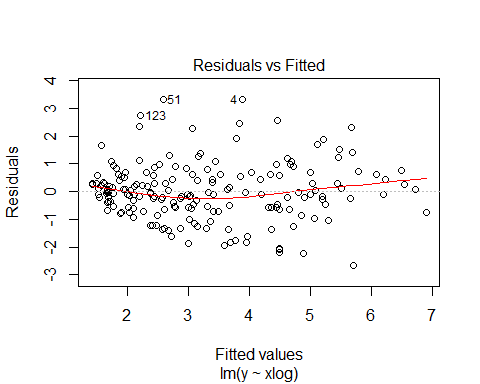
res = as.vector(Y - Yhat)  
head(res)

## [1] 0.07542232 -1.08620457 -0.32594674 3.31535670 0.20451514 -0.06555487

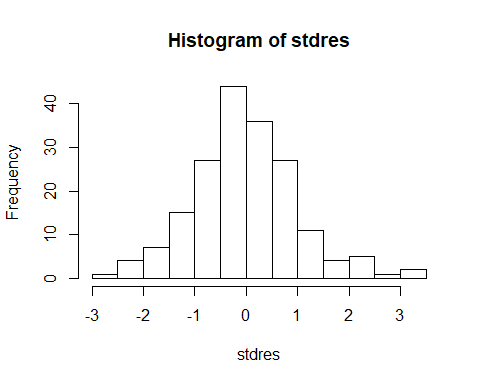
SSE = res%\*%res  
  
SSTO = sum((Y-mean(Y))^2)  
  
Rsquaredmatrix = 1 - SSE/SSTO  
Rsquaredmatrix

## [,1]  
## [1,] 0.6360678

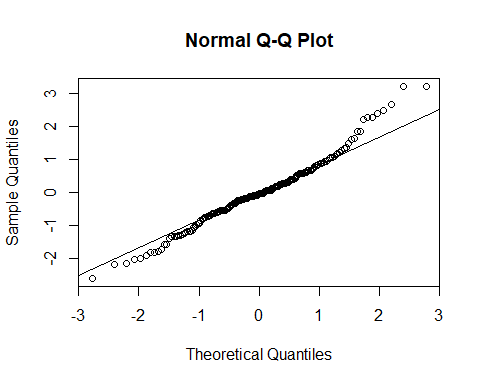
#4  
model = lm(y~xlog)  
plot(model, which=1)



#The dispersion of the residuals seem to be smaller at first, but they spread out as  
#fitted values increase. At the end they shrink again, but the change is not severe.  
stdres=rstandard(model)  
hist(stdres)



#The histogram appears unimodal, with the standard deviation of residuals  
#going to -3 and 3.   
qqnorm(stdres)  
qqline(stdres)



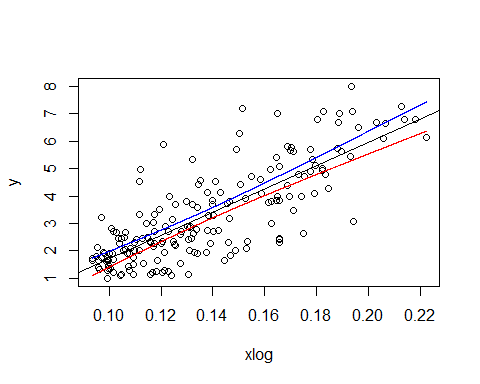
#The QQ plot lies mostly along the line, so there is a mostly normal distribution.  
  
#5  
#test whether beta1 = 0 at 0.05 significance level  
#H0: b1 = 0 v.s. H1: b1 =/= 0  
#T.S. t\* = (b1hat - 0)/SE(b1hat)  
summary(model)

##   
## Call:  
## lm(formula = y ~ xlog)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6499 -0.5764 -0.0475 0.5886 3.3154   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.5603 0.3363 -7.612 1.41e-12 \*\*\*  
## xlog 42.5711 2.3869 17.835 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.031 on 182 degrees of freedom  
## Multiple R-squared: 0.6361, Adjusted R-squared: 0.6341   
## F-statistic: 318.1 on 1 and 182 DF, p-value: < 2.2e-16

#for the lm model, the slope has a p-value of < 2\*10^-16, therefore the conclusion  
#is to reject the null hypothesis  
  
#6  
MSE = summary(model)$sigma^2  
Xh = 1/log(20000)  
Yh=betahat0+betahat1\*Xh  
Yh+c(-1,1)\*qt((1-0.01/2),n-2)\*sqrt(MSE\*(1/n + ((Xh-mean(xlog))^2)/sum((xlog-mean(xlog))^2)))

## [1] 1.438266 2.038231

#1.438266 2.038231  
  
#7  
xseq = seq(min(mydata[,3]), max(mydata[,3]), 0.1)  
xlogseq = 1/log(xseq)  
W = sqrt(2\*qf(1-0.05, 2, n-2))  
yseq=betahat0+betahat1\*xlogseq  
se.y.seq = sqrt(MSE\*(1/n + ((xlogseq-mean(xlog))^2)/sum((xlog-mean(xlog))^2)))  
low = yseq - W\*se.y.seq  
high = yseq + W\*se.y.seq  
  
plot(xlog, y)  
abline(betahat0,betahat1)  
lines(xlogseq,low, col="red")  
lines(xlogseq,high, col="blue")



#8  
MSE = summary(model)$sigma^2  
Xh = 1/log(25000)  
Yh=betahat0+betahat1\*Xh  
Yh+c(-1,1)\*qt((1-0.01/2),n-2)\*sqrt(MSE\*(1/n + 1 + ((Xh-mean(xlog))^2)/sum((xlog-mean(xlog))^2)))

## [1] -1.057710 4.344765

#-1.057710 4.344765