Taking back control

Ralf Martin

# Wage regression

Let’s look at a new example: regressions of hourly wages on years of education (note this is US data)

library(haven)   
getwd()

## [1] "C:/Users/rmart/github/datastorieshub/code"

#load dataset   
data <- read.csv("https://www.dropbox.com/s/9agc2vmamfztlel/WAGE1.csv?dl=1")  
head(data)

## X wage educ exper tenure nonwhite female married numdep smsa northcen south  
## 1 1 3.10 11 2 0 0 1 0 2 1 0 0  
## 2 2 3.24 12 22 2 0 1 1 3 1 0 0  
## 3 3 3.00 11 2 0 0 0 0 2 0 0 0  
## 4 4 6.00 8 44 28 0 0 1 0 1 0 0  
## 5 5 5.30 12 7 2 0 0 1 1 0 0 0  
## 6 6 8.75 16 9 8 0 0 1 0 1 0 0  
## west construc ndurman trcommpu trade services profserv profocc clerocc  
## 1 1 0 0 0 0 0 0 0 0  
## 2 1 0 0 0 0 1 0 0 0  
## 3 1 0 0 0 1 0 0 0 0  
## 4 1 0 0 0 0 0 0 0 1  
## 5 1 0 0 0 0 0 0 0 0  
## 6 1 0 0 0 0 0 1 1 0  
## servocc lwage expersq tenursq  
## 1 0 1.131402 4 0  
## 2 1 1.175573 484 4  
## 3 0 1.098612 4 0  
## 4 0 1.791759 1936 784  
## 5 0 1.667707 49 4  
## 6 0 2.169054 81 64

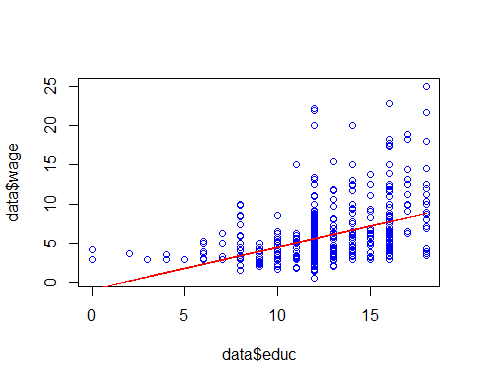
summary(data)

## X wage educ exper   
## Min. : 1.0 Min. : 0.530 Min. : 0.00 Min. : 1.00   
## 1st Qu.:132.2 1st Qu.: 3.330 1st Qu.:12.00 1st Qu.: 5.00   
## Median :263.5 Median : 4.650 Median :12.00 Median :13.50   
## Mean :263.5 Mean : 5.896 Mean :12.56 Mean :17.02   
## 3rd Qu.:394.8 3rd Qu.: 6.880 3rd Qu.:14.00 3rd Qu.:26.00   
## Max. :526.0 Max. :24.980 Max. :18.00 Max. :51.00   
## tenure nonwhite female married   
## Min. : 0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median : 2.000 Median :0.0000 Median :0.0000 Median :1.0000   
## Mean : 5.105 Mean :0.1027 Mean :0.4791 Mean :0.6084   
## 3rd Qu.: 7.000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :44.000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## numdep smsa northcen south   
## Min. :0.000 Min. :0.0000 Min. :0.000 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000   
## Median :1.000 Median :1.0000 Median :0.000 Median :0.0000   
## Mean :1.044 Mean :0.7224 Mean :0.251 Mean :0.3555   
## 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:0.750 3rd Qu.:1.0000   
## Max. :6.000 Max. :1.0000 Max. :1.000 Max. :1.0000   
## west construc ndurman trcommpu   
## Min. :0.0000 Min. :0.00000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :0.0000 Median :0.00000 Median :0.0000 Median :0.00000   
## Mean :0.1692 Mean :0.04563 Mean :0.1141 Mean :0.04373   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.00000 Max. :1.0000 Max. :1.00000   
## trade services profserv profocc   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.2871 Mean :0.1008 Mean :0.2586 Mean :0.3669   
## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## clerocc servocc lwage expersq   
## Min. :0.0000 Min. :0.0000 Min. :-0.6349 Min. : 1.0   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.: 1.2030 1st Qu.: 25.0   
## Median :0.0000 Median :0.0000 Median : 1.5369 Median : 182.5   
## Mean :0.1673 Mean :0.1407 Mean : 1.6233 Mean : 473.4   
## 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.: 1.9286 3rd Qu.: 676.0   
## Max. :1.0000 Max. :1.0000 Max. : 3.2181 Max. :2601.0   
## tenursq   
## Min. : 0.00   
## 1st Qu.: 0.00   
## Median : 4.00   
## Mean : 78.15   
## 3rd Qu.: 49.00   
## Max. :1936.00

mod1 <- lm(wage ~ educ, data)  
summary(mod1)

##   
## Call:  
## lm(formula = wage ~ educ, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.3396 -2.1501 -0.9674 1.1921 16.6085   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.90485 0.68497 -1.321 0.187   
## educ 0.54136 0.05325 10.167 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.378 on 524 degrees of freedom  
## Multiple R-squared: 0.1648, Adjusted R-squared: 0.1632   
## F-statistic: 103.4 on 1 and 524 DF, p-value: < 2.2e-16

plot(data$educ, data$wage, col = "blue")  
lines(data$educ, mod1$fitted.values, col = "red")



summary(lm(exper ~ educ , data))

##   
## Call:  
## lm(formula = exper ~ educ, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.652 -9.971 -2.971 9.125 30.625   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 35.4615 2.6279 13.494 < 2e-16 \*\*\*  
## educ -1.4682 0.2043 -7.187 2.3e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.96 on 524 degrees of freedom  
## Multiple R-squared: 0.08973, Adjusted R-squared: 0.08799   
## F-statistic: 51.65 on 1 and 524 DF, p-value: 2.295e-12

summary(lm(wage ~ educ + exper, data))

##   
## Call:  
## lm(formula = wage ~ educ + exper, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.5532 -1.9801 -0.7071 1.2030 15.8370   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.39054 0.76657 -4.423 1.18e-05 \*\*\*  
## educ 0.64427 0.05381 11.974 < 2e-16 \*\*\*  
## exper 0.07010 0.01098 6.385 3.78e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.257 on 523 degrees of freedom  
## Multiple R-squared: 0.2252, Adjusted R-squared: 0.2222   
## F-statistic: 75.99 on 2 and 523 DF, p-value: < 2.2e-16

summary(lm(educ ~ female,data))

##   
## Call:  
## lm(formula = educ ~ female, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.3175 -0.7883 -0.3175 1.6825 5.6825   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.7883 0.1668 76.652 <2e-16 \*\*\*  
## female -0.4709 0.2410 -1.953 0.0513 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.762 on 524 degrees of freedom  
## Multiple R-squared: 0.00723, Adjusted R-squared: 0.005335   
## F-statistic: 3.816 on 1 and 524 DF, p-value: 0.05129

summary(lm(wage ~ educ+female,data))

##   
## Call:  
## lm(formula = wage ~ educ + female, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.9890 -1.8702 -0.6651 1.0447 15.4998   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.62282 0.67253 0.926 0.355   
## educ 0.50645 0.05039 10.051 < 2e-16 \*\*\*  
## female -2.27336 0.27904 -8.147 2.76e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.186 on 523 degrees of freedom  
## Multiple R-squared: 0.2588, Adjusted R-squared: 0.256   
## F-statistic: 91.32 on 2 and 523 DF, p-value: < 2.2e-16

summary(lm(wage ~ educ+exper+female,data))

##   
## Call:  
## lm(formula = wage ~ educ + exper + female, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.3856 -1.9652 -0.4931 1.1199 14.8217   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.73448 0.75362 -2.302 0.0218 \*   
## educ 0.60258 0.05112 11.788 < 2e-16 \*\*\*  
## exper 0.06424 0.01040 6.177 1.32e-09 \*\*\*  
## female -2.15552 0.27031 -7.974 9.74e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.078 on 522 degrees of freedom  
## Multiple R-squared: 0.3093, Adjusted R-squared: 0.3053   
## F-statistic: 77.92 on 3 and 522 DF, p-value: < 2.2e-16

library(haven) # make sure libraries such as this are installed. If not go to Tools -> Install Packages  
 df=read.csv("https://www.dropbox.com/s/g1w75gkw7g91zef/foreigners.csv?dl=1")  
 df['crimesPc']=df$crimes11/df$pop11  
   
   
 library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

df=df%>%merge(read.csv("https://www.dropbox.com/s/gwq2wmmxr8s3v7t/foreigners\_more.csv?dl=1"),  
 by="area") # Getting further variables  
  
  
   
 reg1=lm(crimesPc~b\_migr11,df)  
   
   
 reg1=lm(crimesPc~pop11+b\_migr11,df)  
 summary(reg1)

##   
## Call:  
## lm(formula = crimesPc ~ pop11 + b\_migr11, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.6243 -0.4052 -0.1253 0.2347 13.8304   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.124e+00 1.018e-01 11.034 < 2e-16 \*\*\*  
## pop11 -1.033e-06 5.078e-07 -2.034 0.0428 \*   
## b\_migr11 4.105e-02 5.335e-03 7.694 1.77e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9627 on 321 degrees of freedom  
## (96 observations deleted due to missingness)  
## Multiple R-squared: 0.1561, Adjusted R-squared: 0.1508   
## F-statistic: 29.68 on 2 and 321 DF, p-value: 1.483e-12

summary(reg1)

##   
## Call:  
## lm(formula = crimesPc ~ pop11 + b\_migr11, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.6243 -0.4052 -0.1253 0.2347 13.8304   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.124e+00 1.018e-01 11.034 < 2e-16 \*\*\*  
## pop11 -1.033e-06 5.078e-07 -2.034 0.0428 \*   
## b\_migr11 4.105e-02 5.335e-03 7.694 1.77e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9627 on 321 degrees of freedom  
## (96 observations deleted due to missingness)  
## Multiple R-squared: 0.1561, Adjusted R-squared: 0.1508   
## F-statistic: 29.68 on 2 and 321 DF, p-value: 1.483e-12

summary(lm(crimesPc~b\_migr11+pop11,df))

##   
## Call:  
## lm(formula = crimesPc ~ b\_migr11 + pop11, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.6243 -0.4052 -0.1253 0.2347 13.8304   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.124e+00 1.018e-01 11.034 < 2e-16 \*\*\*  
## b\_migr11 4.105e-02 5.335e-03 7.694 1.77e-13 \*\*\*  
## pop11 -1.033e-06 5.078e-07 -2.034 0.0428 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9627 on 321 degrees of freedom  
## (96 observations deleted due to missingness)  
## Multiple R-squared: 0.1561, Adjusted R-squared: 0.1508   
## F-statistic: 29.68 on 2 and 321 DF, p-value: 1.483e-12

summary(lm(b\_migr11~pop11,df))

##   
## Call:  
## lm(formula = b\_migr11 ~ pop11, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.039 -5.187 -2.698 1.225 40.835   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.240e+00 9.530e-01 6.548 2.18e-10 \*\*\*  
## pop11 3.088e-05 4.883e-06 6.326 8.02e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.863 on 337 degrees of freedom  
## (81 observations deleted due to missingness)  
## Multiple R-squared: 0.1061, Adjusted R-squared: 0.1035   
## F-statistic: 40.01 on 1 and 337 DF, p-value: 8.024e-10

summary(lm(crimesPc~b\_migr11+pop11+urate2011,df))

##   
## Call:  
## lm(formula = crimesPc ~ b\_migr11 + pop11 + urate2011, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.98367 -0.29360 -0.07496 0.18189 2.92237   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.461e-01 8.058e-02 9.259 < 2e-16 \*\*\*  
## b\_migr11 2.242e-02 2.875e-03 7.799 9.54e-14 \*\*\*  
## pop11 -3.273e-07 2.804e-07 -1.167 0.244   
## urate2011 5.527e-02 9.504e-03 5.815 1.50e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5047 on 311 degrees of freedom  
## (105 observations deleted due to missingness)  
## Multiple R-squared: 0.2674, Adjusted R-squared: 0.2603   
## F-statistic: 37.84 on 3 and 311 DF, p-value: < 2.2e-16

summary(lm(b\_migr11~urate2011,df))

##   
## Call:  
## lm(formula = b\_migr11 ~ urate2011, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.693 -6.169 -2.860 1.549 42.717   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.6865 1.5206 4.397 1.49e-05 \*\*\*  
## urate2011 0.5890 0.1827 3.224 0.00139 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.25 on 326 degrees of freedom  
## (92 observations deleted due to missingness)  
## Multiple R-squared: 0.03091, Adjusted R-squared: 0.02793   
## F-statistic: 10.4 on 1 and 326 DF, p-value: 0.00139

summary(lm(crimesPc~b\_migr11+pop11+medianage,df))

##   
## Call:  
## lm(formula = crimesPc ~ b\_migr11 + pop11 + medianage, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.3870 -0.3786 -0.1240 0.1837 14.0674   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.131e+00 8.862e-01 3.533 0.000472 \*\*\*  
## b\_migr11 2.907e-02 7.465e-03 3.894 0.000120 \*\*\*  
## pop11 -1.509e-06 5.462e-07 -2.764 0.006045 \*\*   
## medianage -4.439e-02 1.947e-02 -2.280 0.023275 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9565 on 320 degrees of freedom  
## (96 observations deleted due to missingness)  
## Multiple R-squared: 0.1696, Adjusted R-squared: 0.1618   
## F-statistic: 21.78 on 3 and 320 DF, p-value: 7.329e-13

summary(lm(crimesPc~b\_migr11+pop11+urate2011+medianage,df))

##   
## Call:  
## lm(formula = crimesPc ~ b\_migr11 + pop11 + urate2011 + medianage,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8873 -0.2680 -0.0783 0.1434 3.1754   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.689e+00 4.855e-01 7.599 3.57e-13 \*\*\*  
## b\_migr11 5.446e-03 3.879e-03 1.404 0.16130   
## pop11 -8.656e-07 2.793e-07 -3.099 0.00212 \*\*   
## urate2011 4.016e-02 9.320e-03 4.309 2.20e-05 \*\*\*  
## medianage -6.305e-02 1.027e-02 -6.138 2.55e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4774 on 310 degrees of freedom  
## (105 observations deleted due to missingness)  
## Multiple R-squared: 0.3468, Adjusted R-squared: 0.3383   
## F-statistic: 41.14 on 4 and 310 DF, p-value: < 2.2e-16

An alternative strategy: Unemployment in 2004 can’t be affected by the surge in migration after 2004

summary(lm(crimesPc~b\_migr11+pop11+medianage+urate2004,df %>% filter(crimesPc<15)))

##   
## Call:  
## lm(formula = crimesPc ~ b\_migr11 + pop11 + medianage + urate2004,   
## data = df %>% filter(crimesPc < 15))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.9334 -0.3021 -0.0885 0.1659 3.1744   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.037e+00 5.105e-01 7.907 4.44e-14 \*\*\*  
## b\_migr11 2.124e-03 4.075e-03 0.521 0.60257   
## pop11 -8.958e-07 2.911e-07 -3.077 0.00227 \*\*   
## medianage -6.787e-02 1.086e-02 -6.252 1.31e-09 \*\*\*  
## urate2004 4.623e-02 1.825e-02 2.534 0.01178 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5011 on 316 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.3226, Adjusted R-squared: 0.314   
## F-statistic: 37.62 on 4 and 316 DF, p-value: < 2.2e-16

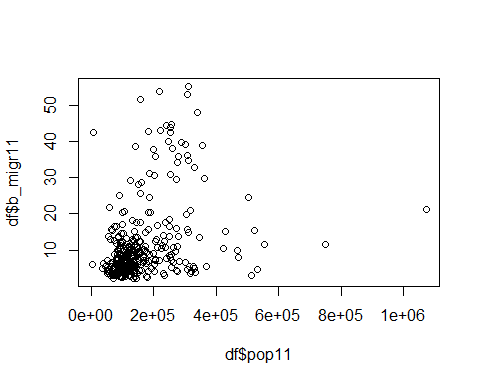
summary(lm(crimesPc~b\_migr11+urate2011+medianage,df))

##   
## Call:  
## lm(formula = crimesPc ~ b\_migr11 + urate2011 + medianage, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.9690 -0.2528 -0.0703 0.1431 3.1876   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.169745 0.461898 6.862 3.67e-11 \*\*\*  
## b\_migr11 0.005737 0.003931 1.459 0.145442   
## urate2011 0.036038 0.009351 3.854 0.000141 \*\*\*  
## medianage -0.053052 0.009886 -5.366 1.57e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4839 on 311 degrees of freedom  
## (105 observations deleted due to missingness)  
## Multiple R-squared: 0.3265, Adjusted R-squared: 0.32   
## F-statistic: 50.26 on 3 and 311 DF, p-value: < 2.2e-16

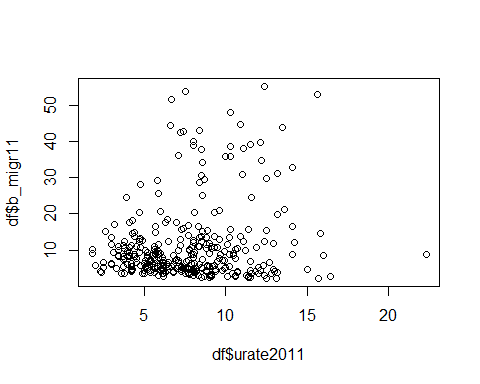
summary(lm(crimesPc~b\_migr11+urate2004+medianage,df))

##   
## Call:  
## lm(formula = crimesPc ~ b\_migr11 + urate2004 + medianage, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.0193 -0.2801 -0.0955 0.1499 3.1917   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.520392 0.488534 7.206 4.23e-12 \*\*\*  
## b\_migr11 0.002952 0.004120 0.717 0.4742   
## urate2004 0.037707 0.018277 2.063 0.0399 \*   
## medianage -0.058060 0.010515 -5.522 7.00e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5078 on 317 degrees of freedom  
## (99 observations deleted due to missingness)  
## Multiple R-squared: 0.3023, Adjusted R-squared: 0.2957   
## F-statistic: 45.78 on 3 and 317 DF, p-value: < 2.2e-16

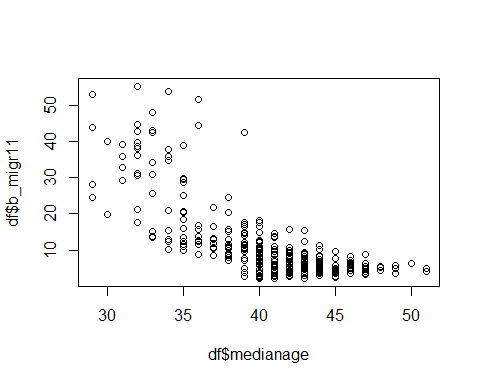
plot(df$pop11,df$b\_migr11)



plot(df$urate2011,df$b\_migr11)



plot(df$medianage,df$b\_migr11)



#df[df$pop11>600000,"area"]

How about differentiating by different groups? Muslims’ get a lot of hate by English neo nazis

reg2=lm(crimesPc~b\_migr11+mus\_sh+pop11+medianage+urate2011,df)  
summary(reg2)

##   
## Call:  
## lm(formula = crimesPc ~ b\_migr11 + mus\_sh + pop11 + medianage +   
## urate2011, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8951 -0.2700 -0.0780 0.1408 3.1797   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.729e+00 4.903e-01 7.606 3.44e-13 \*\*\*  
## b\_migr11 5.940e-03 3.965e-03 1.498 0.13509   
## mus\_sh -5.206e-03 8.458e-03 -0.615 0.53870   
## pop11 -8.385e-07 2.830e-07 -2.963 0.00329 \*\*   
## medianage -6.419e-02 1.045e-02 -6.144 2.48e-09 \*\*\*  
## urate2011 4.091e-02 9.408e-03 4.349 1.86e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4778 on 309 degrees of freedom  
## (105 observations deleted due to missingness)  
## Multiple R-squared: 0.3476, Adjusted R-squared: 0.337   
## F-statistic: 32.92 on 5 and 309 DF, p-value: < 2.2e-16

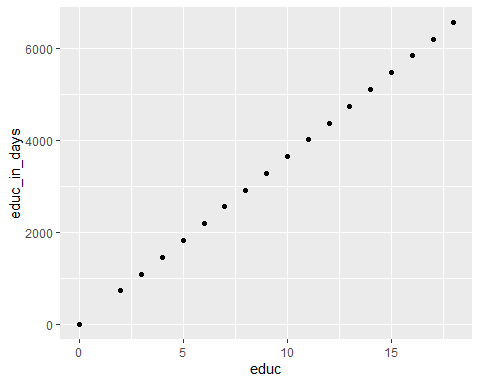
# Multicolinearity

## Perfect colinearity

library(ggplot2)  
data=data %>% mutate(educ\_in\_days=educ\*365)  
cor(data %>% select(educ,educ\_in\_days))

## educ educ\_in\_days  
## educ 1 1  
## educ\_in\_days 1 1

ggplot(data,aes(x=educ,y=educ\_in\_days))+geom\_point()



reg2=lm(wage~female+educ+educ\_in\_days,data)  
reg2 %>% summary()

##   
## Call:  
## lm(formula = wage ~ female + educ + educ\_in\_days, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.9890 -1.8702 -0.6651 1.0447 15.4998   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.62282 0.67253 0.926 0.355   
## female -2.27336 0.27904 -8.147 2.76e-15 \*\*\*  
## educ 0.50645 0.05039 10.051 < 2e-16 \*\*\*  
## educ\_in\_days NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.186 on 523 degrees of freedom  
## Multiple R-squared: 0.2588, Adjusted R-squared: 0.256   
## F-statistic: 91.32 on 2 and 523 DF, p-value: < 2.2e-16

## What about imperfect colinearity?

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

evenmore=read.csv( "https://www.dropbox.com/s/pwotro2ghawkppg/foreign\_evenmore.csv?dl=1")  
df=df%>% inner\_join(evenmore,by="area")  
  
  
  
rr=lm(crimesPc~b\_migr11+urate2011+  
 pop11+  
 shxage0t17+   
 shxage18t29+shxage30t44+shxage45t64+meanage,df %>% filter(crimesPc<150))  
rr%>% summary()

##   
## Call:  
## lm(formula = crimesPc ~ b\_migr11 + urate2011 + pop11 + shxage0t17 +   
## shxage18t29 + shxage30t44 + shxage45t64 + meanage, data = df %>%   
## filter(crimesPc < 150))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.9592 -0.2153 -0.0735 0.1329 3.1625   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.089e+00 3.394e+01 0.238 0.81180   
## b\_migr11 1.646e-04 5.270e-03 0.031 0.97510   
## urate2011 3.746e-02 9.443e-03 3.967 9.07e-05 \*\*\*  
## pop11 -8.774e-07 2.736e-07 -3.206 0.00149 \*\*   
## shxage0t17 -6.445e-02 3.035e-01 -0.212 0.83198   
## shxage18t29 -5.900e-03 2.483e-01 -0.024 0.98106   
## shxage30t44 -2.058e-02 1.833e-01 -0.112 0.91064   
## shxage45t64 -8.662e-02 1.187e-01 -0.730 0.46614   
## meanage -6.790e-02 4.269e-01 -0.159 0.87372   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4645 on 306 degrees of freedom  
## (9 observations deleted due to missingness)  
## Multiple R-squared: 0.3896, Adjusted R-squared: 0.3737   
## F-statistic: 24.41 on 8 and 306 DF, p-value: < 2.2e-16

cor(df %>% select(shxage0t17,   
 shxage18t29,  
 shxage30t44,  
 shxage45t64,meanage),use="complete.obs")

## shxage0t17 shxage18t29 shxage30t44 shxage45t64 meanage  
## shxage0t17 1.00000000 0.01257122 0.3281674 -0.2878871 -0.5427118  
## shxage18t29 0.01257122 1.00000000 0.5810169 -0.9006728 -0.8061182  
## shxage30t44 0.32816735 0.58101695 1.0000000 -0.7408842 -0.8229079  
## shxage45t64 -0.28788711 -0.90067279 -0.7408842 1.0000000 0.8938519  
## meanage -0.54271181 -0.80611820 -0.8229079 0.8938519 1.0000000

A joint hypothesis test could also be a good idea

library("car")  
  
rr%>% vif()

## b\_migr11 urate2011 pop11 shxage0t17 shxage18t29 shxage30t44   
## 4.493034 1.278986 1.346358 471.751245 1556.049097 348.477568   
## shxage45t64 meanage   
## 180.462107 2271.595826

# Joint hypothesis tests

Testing multiple restrictions at once; e.g. does age really not matter in the regression above?

library("car")  
 linearHypothesis(rr, c("shxage0t17 =0" ,  
 "shxage18t29=0",  
 "shxage30t44=0",  
 "shxage45t64=0",  
 "meanage=0"  
 ) )

## Linear hypothesis test  
##   
## Hypothesis:  
## shxage0t17 = 0  
## shxage18t29 = 0  
## shxage30t44 = 0  
## shxage45t64 = 0  
## meanage = 0  
##   
## Model 1: restricted model  
## Model 2: crimesPc ~ b\_migr11 + urate2011 + pop11 + shxage0t17 + shxage18t29 +   
## shxage30t44 + shxage45t64 + meanage  
##   
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 311 79.228   
## 2 306 66.011 5 13.217 12.254 7.689e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

wage1 <- read.csv("https://www.dropbox.com/s/9agc2vmamfztlel/WAGE1.csv?dl=1")  
 wage\_educ\_exper <- lm(wage ~ educ + exper, wage1)  
   
   
 #< by transforming model  
   
 wage1["educPexper"] <- wage1$educ+wage1$exper  
 head(wage1)

## X wage educ exper tenure nonwhite female married numdep smsa northcen south  
## 1 1 3.10 11 2 0 0 1 0 2 1 0 0  
## 2 2 3.24 12 22 2 0 1 1 3 1 0 0  
## 3 3 3.00 11 2 0 0 0 0 2 0 0 0  
## 4 4 6.00 8 44 28 0 0 1 0 1 0 0  
## 5 5 5.30 12 7 2 0 0 1 1 0 0 0  
## 6 6 8.75 16 9 8 0 0 1 0 1 0 0  
## west construc ndurman trcommpu trade services profserv profocc clerocc  
## 1 1 0 0 0 0 0 0 0 0  
## 2 1 0 0 0 0 1 0 0 0  
## 3 1 0 0 0 1 0 0 0 0  
## 4 1 0 0 0 0 0 0 0 1  
## 5 1 0 0 0 0 0 0 0 0  
## 6 1 0 0 0 0 0 1 1 0  
## servocc lwage expersq tenursq educPexper  
## 1 0 1.131402 4 0 13  
## 2 1 1.175573 484 4 34  
## 3 0 1.098612 4 0 13  
## 4 0 1.791759 1936 784 52  
## 5 0 1.667707 49 4 19  
## 6 0 2.169054 81 64 25

#>  
   
 summary( lm(wage ~ educPexper + exper, wage1))

##   
## Call:  
## lm(formula = wage ~ educPexper + exper, data = wage1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.5532 -1.9801 -0.7071 1.2030 15.8370   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.39054 0.76657 -4.423 1.18e-05 \*\*\*  
## educPexper 0.64427 0.05381 11.974 < 2e-16 \*\*\*  
## exper -0.57418 0.05159 -11.129 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.257 on 523 degrees of freedom  
## Multiple R-squared: 0.2252, Adjusted R-squared: 0.2222   
## F-statistic: 75.99 on 2 and 523 DF, p-value: < 2.2e-16

library("car")  
 linearHypothesis(wage\_educ\_exper, c("educ =0" ,"exper=0") )

## Linear hypothesis test  
##   
## Hypothesis:  
## educ = 0  
## exper = 0  
##   
## Model 1: restricted model  
## Model 2: wage ~ educ + exper  
##   
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 525 7160.4   
## 2 523 5548.2 2 1612.2 75.99 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1