Week 7

Jaymeen Gandhi

November 4, 2019

require(tidyverse)

## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

poverty=read\_delim('D:/Into data mining, visulization(Prof.yegin)/Poverty.txt', delim='\t')

## Parsed with column specification:  
## cols(  
## State = col\_character(),  
## MetropolitanResidence = col\_double(),  
## White = col\_double(),  
## Graduates = col\_double(),  
## Poverty = col\_double(),  
## SingleFemaleHousehold = col\_double()  
## )

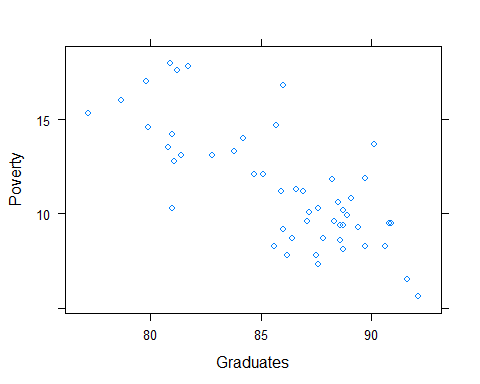
poverty

## # A tibble: 51 x 6  
## State MetropolitanReside~ White Graduates Poverty SingleFemaleHouse~  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Alabama 55.4 71.3 79.9 14.6 14.2  
## 2 Alaska 65.6 70.8 90.6 8.3 10.8  
## 3 Arizona 88.2 87.7 83.8 13.3 11.1  
## 4 Arkansas 52.5 81 80.9 18 12.1  
## 5 California 94.4 77.5 81.1 12.8 12.6  
## 6 Colorado 84.5 90.2 88.7 9.4 9.6  
## 7 Connecticut 87.7 85.4 87.5 7.8 12.1  
## 8 Delaware 80.1 76.3 88.7 8.1 13.1  
## 9 District of C~ 100 36.2 86 16.8 18.9  
## 10 Florida 89.3 80.6 84.7 12.1 12   
## # ... with 41 more rows

require(lattice)

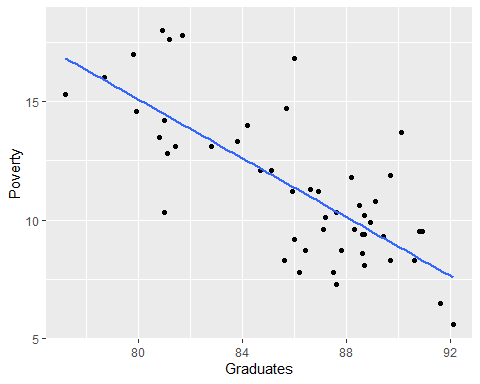
## Loading required package: lattice

xyplot(Poverty~Graduates, data = poverty)



ggplot(data=poverty)+geom\_point(aes(x=Graduates, y=Poverty))+geom\_smooth(method='lm', aes(x=Graduates, y=Poverty) ,fill=NA)

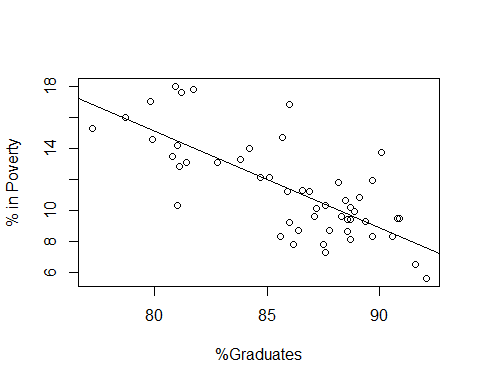
## `geom\_smooth()` using formula 'y ~ x'



lm ( Poverty ~ Graduates,data=poverty)

##   
## Call:  
## lm(formula = Poverty ~ Graduates, data = poverty)  
##   
## Coefficients:  
## (Intercept) Graduates   
## 64.7810 -0.6212

plot(Poverty ~ Graduates, data = poverty, ylab =  
"% in Poverty",  
xlab =  
"%Graduates")  
lm\_pov\_grad = lm(poverty$Poverty ~ poverty$Graduates)  
abline(lm\_pov\_grad)



lm(formula = Graduates ~ Poverty , data = poverty)

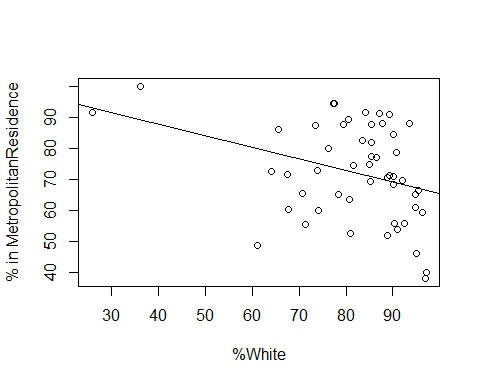
##   
## Call:  
## lm(formula = Graduates ~ Poverty, data = poverty)  
##   
## Coefficients:  
## (Intercept) Poverty   
## 96.2022 -0.8979

my\_model=lm(formula = Poverty ~ Graduates , data = poverty)

summary(my\_model)

##   
## Call:  
## lm(formula = Poverty ~ Graduates, data = poverty)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.1624 -1.2593 -0.2184 0.9611 5.4437   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.78097 6.80260 9.523 9.94e-13 \*\*\*  
## Graduates -0.62122 0.07902 -7.862 3.11e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.082 on 49 degrees of freedom  
## Multiple R-squared: 0.5578, Adjusted R-squared: 0.5488   
## F-statistic: 61.81 on 1 and 49 DF, p-value: 3.109e-10

plot(MetropolitanResidence ~ White, data = poverty, ylab =  
"% in MetropolitanResidence",  
xlab =  
"%White")  
lm\_pov\_grad = lm(poverty$MetropolitanResidence ~ poverty$White)  
abline(lm\_pov\_grad)

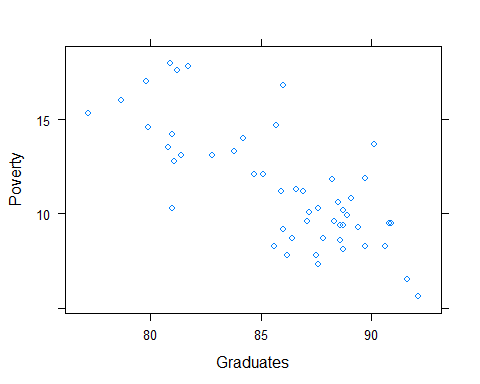


my\_model2=lm(formula = MetropolitanResidence ~ White , data = poverty)

summary(my\_model2)

##   
## Call:  
## lm(formula = MetropolitanResidence ~ White, data = poverty)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.173 -9.877 1.254 11.066 21.467   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 103.0094 12.2262 8.425 4.31e-11 \*\*\*  
## White -0.3764 0.1475 -2.551 0.0139 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.5 on 49 degrees of freedom  
## Multiple R-squared: 0.1173, Adjusted R-squared: 0.09925   
## F-statistic: 6.509 on 1 and 49 DF, p-value: 0.0139

require(lattice)  
xyplot(Poverty~Graduates , data = poverty)



cor(poverty$Graduates,poverty$Poverty)

## [1] -0.7468583

my\_model=lm(Poverty~Graduates,data = poverty)  
summary(my\_model)

##   
## Call:  
## lm(formula = Poverty ~ Graduates, data = poverty)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.1624 -1.2593 -0.2184 0.9611 5.4437   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.78097 6.80260 9.523 9.94e-13 \*\*\*  
## Graduates -0.62122 0.07902 -7.862 3.11e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.082 on 49 degrees of freedom  
## Multiple R-squared: 0.5578, Adjusted R-squared: 0.5488   
## F-statistic: 61.81 on 1 and 49 DF, p-value: 3.109e-10

##poverty=64.78-0.62\*Graduates  
##if graduation rate increase by 1 unit than it decrases the poverty by 0.62 units.

my\_model2=lm(Poverty~White,data = poverty)  
summary(my\_model2)

##   
## Call:  
## lm(formula = Poverty ~ White, data = poverty)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1581 -1.9846 -0.4046 2.2057 6.6698   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.9715 2.5113 6.758 1.57e-08 \*\*\*  
## White -0.0688 0.0303 -2.270 0.0276 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.978 on 49 degrees of freedom  
## Multiple R-squared: 0.09518, Adjusted R-squared: 0.07672   
## F-statistic: 5.155 on 1 and 49 DF, p-value: 0.02762

##povert=16.975-0.06\*whites  
##if graduation rate increase by 1 unit than it decrases the poverty by 0.06 units  
##Not a good model chech r squared is 0.09 ie 9%  
##When the population of white is 0 than 16.95 times will increase the poverty.

my\_multi\_model=lm(Poverty~Graduates +White,data = poverty)  
summary(my\_multi\_model)

##   
## Call:  
## lm(formula = Poverty ~ Graduates + White, data = poverty)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.794 -1.246 -0.178 1.069 5.062   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.94522 6.73197 9.647 8.10e-13 \*\*\*  
## Graduates -0.59378 0.08050 -7.376 1.95e-09 \*\*\*  
## White -0.03089 0.02158 -1.431 0.159   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.06 on 48 degrees of freedom  
## Multiple R-squared: 0.5759, Adjusted R-squared: 0.5582   
## F-statistic: 32.59 on 2 and 48 DF, p-value: 1.147e-09

##poverty=64.94-0.59Graduates-0.03whites  
##if graduation rate increase by 1 unit than it decrases the poverty by 0.59 units and the whites be constant.  
##poverty=a\*graduates+b\*whites+c\*blacks

my\_multi\_model2=lm(Poverty~Graduates +White +MetropolitanResidence +SingleFemaleHousehold,data = poverty)  
summary(my\_multi\_model2)

##   
## Call:  
## lm(formula = Poverty ~ Graduates + White + MetropolitanResidence +   
## SingleFemaleHousehold, data = poverty)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.289 -1.506 -0.323 1.235 4.449   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 66.47653 12.58990 5.280 3.41e-06 \*\*\*  
## Graduates -0.55471 0.10491 -5.288 3.33e-06 \*\*\*  
## White -0.04814 0.03306 -1.456 0.152   
## MetropolitanResidence -0.05632 0.01955 -2.881 0.006 \*\*   
## SingleFemaleHousehold 0.05054 0.24330 0.208 0.836   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.934 on 46 degrees of freedom  
## Multiple R-squared: 0.6416, Adjusted R-squared: 0.6104   
## F-statistic: 20.58 on 4 and 46 DF, p-value: 8.884e-10

## A unit increase in metroplRES will decrease the poverty by 0.05 , if every other variable is kept constant  
##For significance check the last column ie p value.SingleFemHos is the less significant and also check the stars in p value the one without star has to be removed.

my\_multi\_model3=lm(Poverty~Graduates +White +MetropolitanResidence,data = poverty)  
summary(my\_multi\_model3)

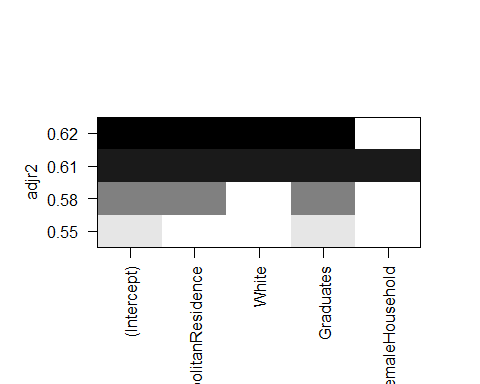
##   
## Call:  
## lm(formula = Poverty ~ Graduates + White + MetropolitanResidence,   
## data = poverty)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.2346 -1.4785 -0.3699 1.2153 4.5579   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 68.72202 6.38893 10.756 2.89e-14 \*\*\*  
## Graduates -0.56972 0.07527 -7.569 1.13e-09 \*\*\*  
## White -0.05333 0.02148 -2.483 0.01665 \*   
## MetropolitanResidence -0.05553 0.01898 -2.926 0.00528 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.915 on 47 degrees of freedom  
## Multiple R-squared: 0.6412, Adjusted R-squared: 0.6183   
## F-statistic: 28 on 3 and 47 DF, p-value: 1.553e-10

##This will be the best model since Rsquared is 60% and p value is very low.

require(leaps)

## Loading required package: leaps

lps=regsubsets(Poverty~MetropolitanResidence+White+Graduates+SingleFemaleHousehold, data=poverty, method='backward')   
plot(lps, scale="adjr2" )



## Here we see the relationship of poverty with various other attributes.