## Simple (One Predictor) Linear Regression

Needed libraries

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.7  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

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

library(GGally)

## Warning: package 'GGally' was built under R version 3.5.2

##   
## Attaching package: 'GGally'

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

library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 3.5.2

Read-in the data. Before doing this make sure that you have placed the CreditData.csv file (downloadable from Canvas) in your projects working directory.

credit <- read\_csv("CreditData.csv")

## Parsed with column specification:  
## cols(  
## AnnualIncome = col\_double(),  
## HouseholdSize = col\_integer(),  
## YrsEdAfterHS = col\_integer(),  
## HrWkTV = col\_integer(),  
## AnnualCharges = col\_double()  
## )

Examine the structure and summary of the dataset

str(credit) #all variables numeric

## Classes 'tbl\_df', 'tbl' and 'data.frame': 5896 obs. of 5 variables:  
## $ AnnualIncome : num 21.8 65.5 54.2 73.7 110.4 ...  
## $ HouseholdSize: int 4 7 3 6 7 8 5 8 1 3 ...  
## $ YrsEdAfterHS : int 5 3 2 0 5 3 4 5 4 1 ...  
## $ HrWkTV : int 29 46 18 44 39 39 40 27 15 3 ...  
## $ AnnualCharges: num 10024 11249 6115 9786 12634 ...  
## - attr(\*, "spec")=List of 2  
## ..$ cols :List of 5  
## .. ..$ AnnualIncome : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ HouseholdSize: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ YrsEdAfterHS : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ HrWkTV : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ AnnualCharges: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## ..$ default: list()  
## .. ..- attr(\*, "class")= chr "collector\_guess" "collector"  
## ..- attr(\*, "class")= chr "col\_spec"

summary(credit) #notice 896 missing rows

## AnnualIncome HouseholdSize YrsEdAfterHS HrWkTV   
## Min. : 5.40 Min. :1.000 Min. :0.000 Min. : 0.0   
## 1st Qu.: 45.30 1st Qu.:3.000 1st Qu.:1.000 1st Qu.:15.0   
## Median : 65.50 Median :5.000 Median :3.000 Median :30.0   
## Mean : 67.61 Mean :4.534 Mean :2.525 Mean :29.9   
## 3rd Qu.: 90.40 3rd Qu.:6.000 3rd Qu.:4.000 3rd Qu.:45.0   
## Max. :134.20 Max. :8.000 Max. :5.000 Max. :60.0   
## NA's :896 NA's :896 NA's :896 NA's :896   
## AnnualCharges   
## Min. : 0   
## 1st Qu.: 6926   
## Median :11168   
## Mean :11351   
## 3rd Qu.:15724   
## Max. :32204   
## NA's :896

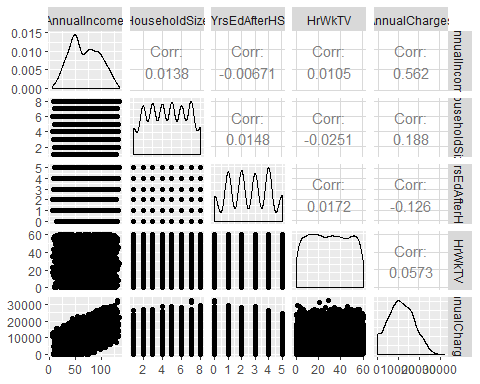
Get rid of missing data rows

credit <- credit %>% drop\_na() #delete any row with an NA value  
str(credit) #check structure after the drop

## Classes 'tbl\_df', 'tbl' and 'data.frame': 5000 obs. of 5 variables:  
## $ AnnualIncome : num 21.8 65.5 54.2 73.7 110.4 ...  
## $ HouseholdSize: int 4 7 3 6 7 8 5 8 1 3 ...  
## $ YrsEdAfterHS : int 5 3 2 0 5 3 4 5 4 1 ...  
## $ HrWkTV : int 29 46 18 44 39 39 40 27 15 3 ...  
## $ AnnualCharges: num 10024 11249 6115 9786 12634 ...

Our Y (response) variable in the dataset is “AnnualCharges”. Let’s look at ggparis plot for visualization and correlation.

ggpairs(credit)



The best variable (by correlation and confirmed by visualization) to predict AnnualCharges appear to be AnnualIncome (correlation = 0.562 and there is an intuitive increase in charges as income increases).

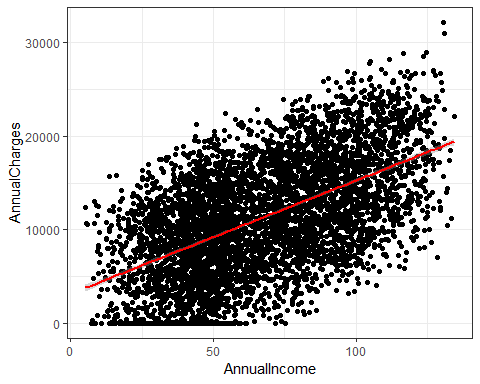
Build a regression model with AnnualIncome to predict AnnualCharges.

mod1<-lm(AnnualCharges ~ AnnualIncome, credit) #create linear regression model  
summary(mod1) #examine the model

##   
## Call:  
## lm(formula = AnnualCharges ~ AnnualIncome, data = credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12284.4 -3938.1 14.4 3947.9 13232.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3146.361 185.193 16.99 <2e-16 \*\*\*  
## AnnualIncome 121.355 2.529 47.98 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5027 on 4998 degrees of freedom  
## Multiple R-squared: 0.3153, Adjusted R-squared: 0.3152   
## F-statistic: 2302 on 1 and 4998 DF, p-value: < 2.2e-16

Is this a good model?

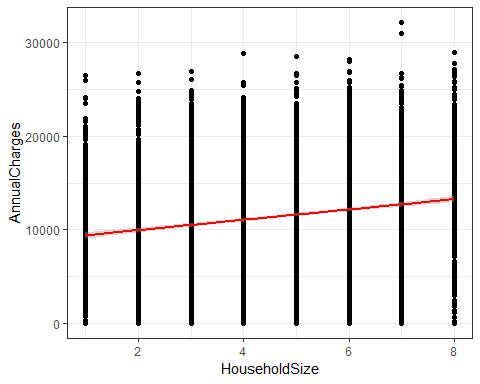
ggplot(credit,aes(x=AnnualIncome,y=AnnualCharges))+geom\_point()+geom\_smooth(method="lm", color="red")+ theme\_bw()



mod2<-lm(AnnualCharges ~ HouseholdSize, credit) #create linear regression model  
summary(mod2) #examine the model

##   
## Call:  
## lm(formula = AnnualCharges ~ HouseholdSize, data = credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13260.4 -4374.8 -105.1 4144.9 19494.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8853.15 202.92 43.63 <2e-16 \*\*\*  
## HouseholdSize 550.91 40.71 13.53 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5967 on 4998 degrees of freedom  
## Multiple R-squared: 0.03535, Adjusted R-squared: 0.03516   
## F-statistic: 183.2 on 1 and 4998 DF, p-value: < 2.2e-16

ggplot(credit,aes(x=HouseholdSize,y=AnnualCharges))+geom\_point()+geom\_smooth(method="lm", color="red")+ theme\_bw()



Is this a good model?

R-squared value is pretty poor. The HouseholdSize variable is significant (p-value is < 0.05)and has an intuitive sign. Note: As datasets increase in size, it is very easy for the predictor vairable to be significant.