# Lawrence Nagbe

## BAN 502 LINREG Practice

### March 7, 2023

Loading Libraries for use in data analysis

library(tidymodels)

## Warning: package 'tidymodels' was built under R version 4.2.2

## ── Attaching packages ────────────────────────────────────── tidymodels 1.0.0 ──

## ✔ broom 1.0.3 ✔ recipes 1.0.5  
## ✔ dials 1.1.0 ✔ rsample 1.1.1  
## ✔ dplyr 1.1.0 ✔ tibble 3.1.8  
## ✔ ggplot2 3.4.1 ✔ tidyr 1.3.0  
## ✔ infer 1.0.4 ✔ tune 1.0.1  
## ✔ modeldata 1.1.0 ✔ workflows 1.1.3  
## ✔ parsnip 1.0.4 ✔ workflowsets 1.0.0  
## ✔ purrr 1.0.1 ✔ yardstick 1.1.0

## Warning: package 'broom' was built under R version 4.2.2

## Warning: package 'dials' was built under R version 4.2.2

## Warning: package 'dplyr' was built under R version 4.2.2

## Warning: package 'ggplot2' was built under R version 4.2.2

## Warning: package 'infer' was built under R version 4.2.2

## Warning: package 'parsnip' was built under R version 4.2.2

## Warning: package 'purrr' was built under R version 4.2.2

## Warning: package 'recipes' was built under R version 4.2.2

## Warning: package 'rsample' was built under R version 4.2.2

## Warning: package 'tidyr' was built under R version 4.2.2

## Warning: package 'tune' was built under R version 4.2.2

## Warning: package 'workflows' was built under R version 4.2.2

## Warning: package 'workflowsets' was built under R version 4.2.2

## Warning: package 'yardstick' was built under R version 4.2.2

## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ purrr::discard() masks scales::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ recipes::step() masks stats::step()  
## • Learn how to get started at https://www.tidymodels.org/start/

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.2

## Warning: package 'readr' was built under R version 4.2.2

## Warning: package 'stringr' was built under R version 4.2.2

## Warning: package 'forcats' was built under R version 4.2.2

## Warning: package 'lubridate' was built under R version 4.2.2

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ forcats 1.0.0 ✔ readr 2.1.4  
## ✔ lubridate 1.9.2 ✔ stringr 1.5.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ readr::col\_factor() masks scales::col\_factor()  
## ✖ purrr::discard() masks scales::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ stringr::fixed() masks recipes::fixed()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ readr::spec() masks yardstick::spec()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(GGally)

## Warning: package 'GGally' was built under R version 4.2.2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 4.2.2

library(gridExtra) # cleans the plot and makes it readable

## Warning: package 'gridExtra' was built under R version 4.2.2

##   
## Attaching package: 'gridExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine

Read into the dataset

library(readr)  
Credit <- read\_csv("CreditData.csv")

## Rows: 5000 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (5): AnnualIncome, HouseholdSize, YrsEdAfterHS, HrWkTV, AnnualCharges  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View the structucre and summary of the dataset before analysis

str(Credit)

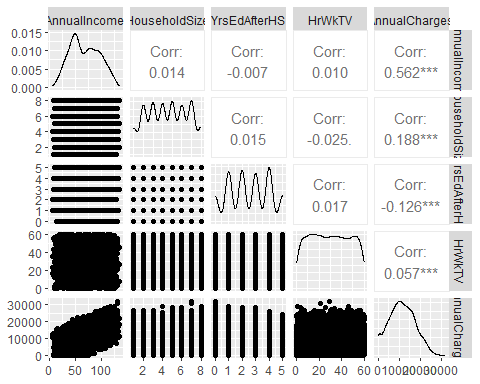
## spc\_tbl\_ [5,000 × 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ AnnualIncome : num [1:5000] 21.8 65.5 54.2 73.7 110.4 ...  
## $ HouseholdSize: num [1:5000] 4 7 3 6 7 8 5 8 1 3 ...  
## $ YrsEdAfterHS : num [1:5000] 5 3 2 0 5 3 4 5 4 1 ...  
## $ HrWkTV : num [1:5000] 29 46 18 44 39 39 40 27 15 3 ...  
## $ AnnualCharges: num [1:5000] 10024 11249 6115 9786 12634 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. AnnualIncome = col\_double(),  
## .. HouseholdSize = col\_double(),  
## .. YrsEdAfterHS = col\_double(),  
## .. HrWkTV = col\_double(),  
## .. AnnualCharges = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

summary(Credit)

## 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   
## AnnualCharges   
## Min. : 0   
## 1st Qu.: 6926   
## Median :11168   
## Mean :11351   
## 3rd Qu.:15724   
## Max. :32204

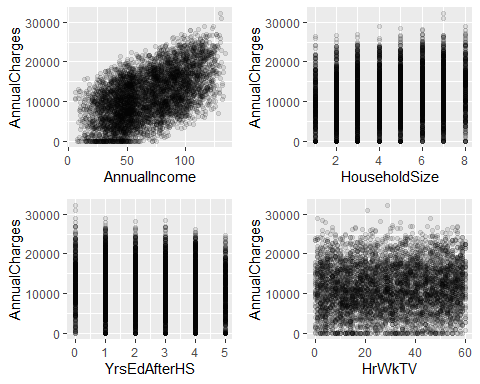
Using ggpairs to visualize the dataset and compares correlations of the variables

ggpairs(Credit)



Alternative to the the ggpairs visualization is the alpha plot for visualization purposes

p1 = ggplot(Credit, aes(x=AnnualIncome,y=AnnualCharges)) + geom\_point(alpha=0.1) #changing alpha is helpful when many points may overlap  
p2 = ggplot(Credit, aes(x=HouseholdSize,y=AnnualCharges)) + geom\_point(alpha=0.1)  
p3 = ggplot(Credit, aes(x=YrsEdAfterHS,y=AnnualCharges)) + geom\_point(alpha=0.1)  
p4 = ggplot(Credit, aes(x=HrWkTV,y=AnnualCharges)) + geom\_point(alpha=0.1)  
grid.arrange(p1,p2,p3,p4,ncol=2)



Building a Regression Model with AnnualIncome to Predict AnnualCharges, using non tidymodels approach

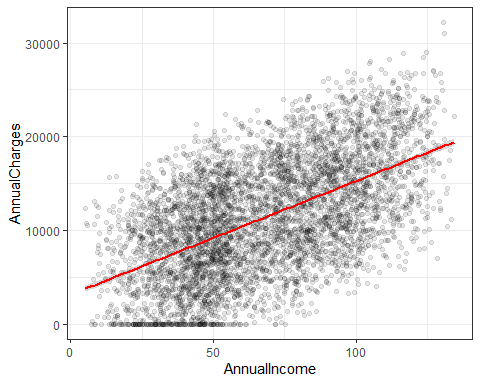
mod1 = lm(AnnualCharges ~ AnnualIncome, Credit)  
summary(mod1) #examining the model: Is this a good model?

##   
## 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

Plot the model

ggplot(Credit, aes(x=AnnualIncome,y=AnnualCharges)) + geom\_point(alpha=0.1) + geom\_smooth(method = "lm", color = "red") + theme\_bw()

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



Building similar model but in Tidymodels, let’s use recipe

Credit\_simple = recipe(AnnualCharges ~ AnnualIncome, Credit)  
Credit\_simple

##

## ── Recipe ──────────────────────────────────────────────────────────────────────

##

## ── Inputs

## Number of variables by role

## outcome: 1  
## predictor: 1

Next, we specify the kind of model we are building

lm\_model = #gives the model type a name  
 linear\_reg() %>% #specifies that we are doing linear regression  
 set\_engine("lm") #specifies the specific type of linear tool we want to use

Next, we will combine the recipe and the model with a workflow

lm\_wflow =  
 workflow() %>%  
 add\_model(lm\_model) %>%  
 add\_recipe(Credit\_simple)

Next, we fit (execute) the workflow on our dataset

lm\_fit = fit(lm\_wflow, Credit)

Finally, let’s see the output of the lm\_fit

summary(lm\_fit$fit$fit$fit) # three fits, the actual fits is embeded in the object

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## 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