Untitled

Akshay Kulkarni 001445074

November 21, 2018

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

# Loading all the required libraries   
  
library(tidyverse)

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

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 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(RODBC)  
library(ineq) # for Lc and Gini  
library(ggplot2)  
library(scales)

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

library(showtext) # for fonts

## Loading required package: sysfonts

## Loading required package: showtextdb

library(stringr) # for str\_wrap  
library(grid)  
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(ggcorrplot)  
library(modelr)  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(leaps)  
library(nnet)  
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

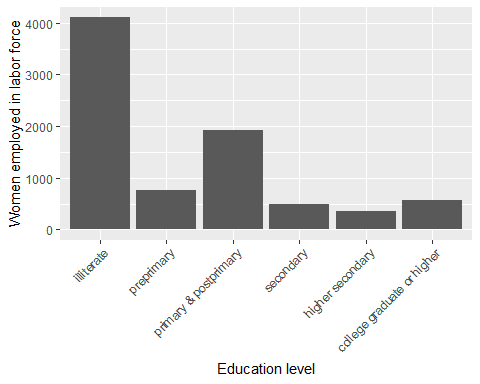
##   
## Attaching package: 'randomForest'

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

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(dplyr)

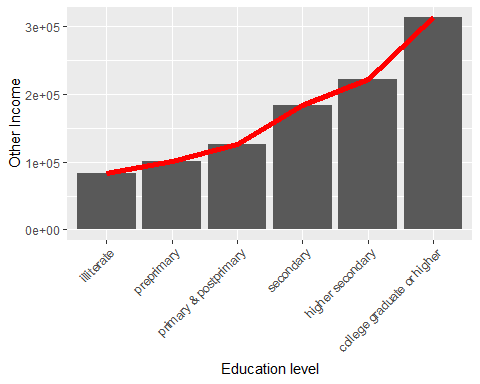
load("C:/Users/Akshay/Desktop/ICPSR\_36151/DS0003/36151-0003-Data.rda")  
  
  
  
women <- da36151.0003  
  
women <- filter(women, EW6 >= 25 & EW6 <= 59)  
women <- filter(women, RO6 == "(1) Married 1" | RO6 == "(0) married, spouse absent")  
women <- women %>% filter(!is.na(GR46), GR46 != '-')  
  
women <- dplyr::select(women, HHID, PERSONID, IDHH, IDPERSON, EW5, EW6, EW8, EW9, EW10, GR46, ID11, ID13,   
 HHEDUC, HHEDUCM, HHEDUCF, NCHILDM, NCHILDF, SPED6, SPED2, SPED3, SPRO5, INCCROP, INCAGPROP,   
 INCANIMAL, INCAG, INCBUS, INCBUSCALC, INCOTHER, INCOME, INCOMEPC, INCNONAG, INCAGLAB, INCSALARY,   
 INCNREGA,RSUNEARN, GR48, GR46B,GR46A)  
  
women <- mutate(women, education = ifelse(EW8 == "(00) none 0", "illiterate",   
 ifelse(EW8 == "(01) 1st class 1" | EW8 == "(02) 2nd class 2" | EW8 == "(03) 3rd class 3" | EW8 == "(04) 4th class 4", "preprimary",   
 ifelse(EW8 == "(05) 5th class 5" | EW8 == "(06) 6th class 6" | EW8 == "(07) 7th class 7" | EW8 == "(08) 8th class 8" | EW8 == "(09) 9th class 9", "primary & postprimary",   
 ifelse(EW8 == "(10) Secondary 10" | EW8 == "(11) 11th Class 11", "secondary",   
 ifelse(EW8 == "(12) High Secondary 12" | EW8 == "(13) 1 year post-secondary 13" | EW8 == "(14) 2 years post-secondary 14", "higher secondary",   
 ifelse(EW8 == "(15) Bachelors 15" | EW8 == "(16) Above Bachelors 16", "college graduate or higher", "-")))))))  
View(head(women, 20))  
  
women$education <- factor(women$education, levels = c("illiterate","preprimary" ,"primary & postprimary" , "secondary", "higher secondary", "college graduate or higher"))  
ggplot(data=women %>% filter(!is.na(education), !is.na(GR46B), GR46B == '(1) Yes 1'), aes(x = education)) + geom\_bar() +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +   
 labs(x = "Education level", y = "Women employed in labor force")



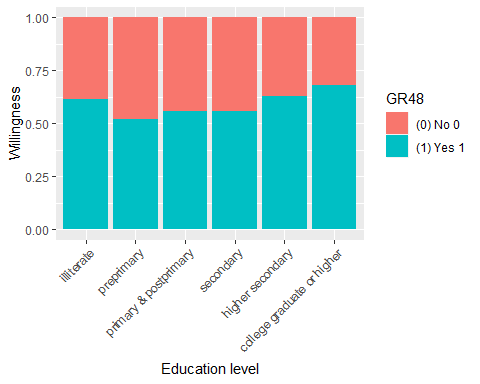
summary <- women %>%  
 filter(!is.na(education), !is.na(RSUNEARN)) %>%  
 group\_by(education) %>%  
 summarise(other\_household\_income = mean(RSUNEARN))  
  
summary

## # A tibble: 6 x 2  
## education other\_household\_income  
## <fct> <dbl>  
## 1 illiterate 83016.  
## 2 preprimary 101302.  
## 3 primary & postprimary 126467.  
## 4 secondary 183716.  
## 5 higher secondary 222812.  
## 6 college graduate or higher 313270.

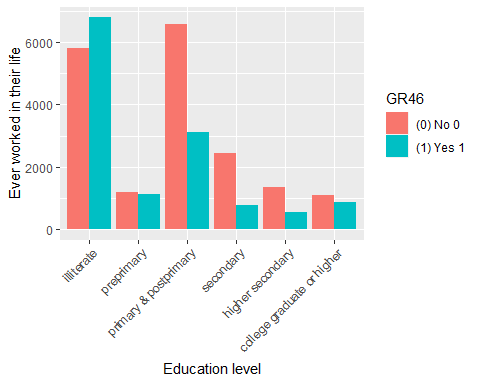
ed\_prop <- prop.table(table(women$education))  
  
ggplot(data = summary) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +   
 geom\_histogram(aes(x= education, y = other\_household\_income,group = 1),stat="identity") +  
 geom\_line(aes(x= education, y = other\_household\_income,group = 1),stat="identity",color="red",lwd=2) +  
 labs(x = "Education level", y = "Other Income")



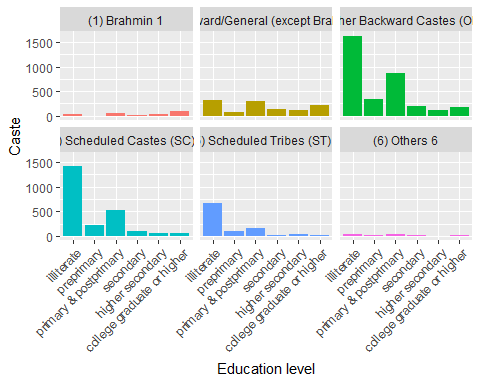
ggplot(data = women %>% filter(!is.na(education), !is.na(GR48))) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +   
 geom\_bar(aes(x= education, fill= GR48), position="fill") +  
 labs(x = "Education level", y = "Willingness")



ggplot(data = women %>% filter(!is.na(education), !is.na(GR46))) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +   
 geom\_bar(aes(x= education, fill= GR46), position="dodge") +  
 labs(x = "Education level", y = "Ever worked in their life")



ggplot(data = women %>% filter(!is.na(education), !is.na(ID13), !is.na(GR46B), GR46B == '(1) Yes 1')) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +   
 geom\_bar(aes(x= education, fill = ID13),show.legend = FALSE) +   
 facet\_wrap(~ID13) +  
 labs(x = "Education level", y = "Caste")



# ````````````````````````````````````````````````````````````````````````````````````  
  
  
  
set.seed(1)  
women\_par <- resample\_partition(women, c(train = 0.8, test = 0.2))  
  
  
  
# Converting partitions into tibbles  
  
train\_partition <- as\_tibble(women\_par$train)  
test\_partition <- as\_tibble(women\_par$test)  
  
  
  
# Checking for factors with single/ 1 levels ( DROP = Factor has only one level)  
  
(l <- sapply(women, function(x) is.factor(x)))

## HHID PERSONID IDHH IDPERSON EW5 EW6   
## FALSE FALSE TRUE TRUE TRUE FALSE   
## EW8 EW9 EW10 GR46 ID11 ID13   
## TRUE FALSE TRUE TRUE TRUE TRUE   
## HHEDUC HHEDUCM HHEDUCF NCHILDM NCHILDF SPED6   
## TRUE TRUE TRUE FALSE FALSE TRUE   
## SPED2 SPED3 SPRO5 INCCROP INCAGPROP INCANIMAL   
## TRUE TRUE FALSE FALSE FALSE FALSE   
## INCAG INCBUS INCBUSCALC INCOTHER INCOME INCOMEPC   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## INCNONAG INCAGLAB INCSALARY INCNREGA RSUNEARN GR48   
## FALSE FALSE FALSE FALSE FALSE TRUE   
## GR46B GR46A education   
## TRUE TRUE TRUE

m <- women[, l]  
ifelse(n <- sapply(m, function(x) length(levels(x))) == 1, "DROP", "NODROP")

## IDHH IDPERSON EW5 EW8 EW10 GR46 ID11   
## "NODROP" "NODROP" "NODROP" "NODROP" "NODROP" "NODROP" "NODROP"   
## ID13 HHEDUC HHEDUCM HHEDUCF SPED6 SPED2 SPED3   
## "NODROP" "NODROP" "NODROP" "NODROP" "NODROP" "NODROP" "NODROP"   
## GR48 GR46B GR46A education   
## "NODROP" "NODROP" "NODROP" "NODROP"

# No factors in dataset with single levels which mightve caused contrast error while running the model.  
  
  
  
# GLM / Logistic Regression  
  
fit <- glm(GR46 ~(education), data = women\_par$train, family = binomial(link= "logit")) # Single predictor variable  
  
fit1 <- glm(GR46 ~ EW8+ID11+ID13+RSUNEARN+INCOME, data = women\_par$train, family = binomial(link= "logit"))  
  
  
summary(fit)

##   
## Call:  
## glm(formula = GR46 ~ (education), family = binomial(link = "logit"),   
## data = women\_par$train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.2415 -1.0680 -0.8376 1.1147 1.6805   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.14938 0.01994 7.490 6.87e-14  
## educationpreprimary -0.17516 0.05046 -3.471 0.000518  
## educationprimary & postprimary -0.90695 0.03155 -28.745 < 2e-16  
## educationsecondary -1.28229 0.04963 -25.835 < 2e-16  
## educationhigher secondary -1.01657 0.05932 -17.137 < 2e-16  
## educationcollege graduate or higher -0.41219 0.05469 -7.536 4.83e-14  
##   
## (Intercept) \*\*\*  
## educationpreprimary \*\*\*  
## educationprimary & postprimary \*\*\*  
## educationsecondary \*\*\*  
## educationhigher secondary \*\*\*  
## educationcollege graduate or higher \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 34514 on 25407 degrees of freedom  
## Residual deviance: 33121 on 25402 degrees of freedom  
## (1 observation deleted due to missingness)  
## AIC: 33133  
##   
## Number of Fisher Scoring iterations: 4

summary(fit1)

##   
## Call:  
## glm(formula = GR46 ~ EW8 + ID11 + ID13 + RSUNEARN + INCOME, family = binomial(link = "logit"),   
## data = women\_par$train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -8.4904 -0.9394 -0.5676 1.0347 7.6834   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -5.959e-01 7.691e-02 -7.749  
## EW8(01) 1st class 1 2.323e-01 2.273e-01 1.022  
## EW8(02) 2nd class 2 -1.375e-01 1.066e-01 -1.290  
## EW8(03) 3rd class 3 1.329e-02 9.178e-02 0.145  
## EW8(04) 4th class 4 -2.228e-01 7.818e-02 -2.849  
## EW8(05) 5th class 5 -5.412e-01 5.094e-02 -10.624  
## EW8(06) 6th class 6 -5.955e-01 9.218e-02 -6.461  
## EW8(07) 7th class 7 -6.557e-01 6.471e-02 -10.132  
## EW8(08) 8th class 8 -7.646e-01 5.967e-02 -12.814  
## EW8(09) 9th class 9 -9.325e-01 6.206e-02 -15.026  
## EW8(10) Secondary 10 -9.288e-01 5.701e-02 -16.293  
## EW8(11) 11th Class 11 -5.537e-01 1.411e-01 -3.924  
## EW8(12) High Secondary 12 -6.648e-01 7.158e-02 -9.287  
## EW8(13) 1 year post-secondary 13 1.255e-02 2.739e-01 0.046  
## EW8(14) 2 years post-secondary 14 -2.623e-01 2.230e-01 -1.176  
## EW8(15) Bachelors 15 -6.076e-02 8.055e-02 -0.754  
## EW8(16) Above Bachelors 16 7.889e-01 1.131e-01 6.975  
## ID11(2) Muslim 2 -6.139e-01 4.813e-02 -12.755  
## ID11(3) Christian 3 2.176e-01 9.465e-02 2.299  
## ID11(4) Sikh 4 -6.626e-02 9.821e-02 -0.675  
## ID11(5) Buddhist 5 -2.424e-01 1.701e-01 -1.425  
## ID11(6) Jain 6 -7.198e-01 3.363e-01 -2.141  
## ID11(7) Tribal 7 2.128e-01 2.334e-01 0.912  
## ID11(8) Others 8 2.086e-01 4.521e-01 0.461  
## ID11(9) None 9 -3.611e-01 9.733e-01 -0.371  
## ID13(2) Forward/General (except Brahmin) 2 3.330e-01 7.909e-02 4.211  
## ID13(3) Other Backward Castes (OBC) 3 8.241e-01 7.596e-02 10.850  
## ID13(4) Scheduled Castes (SC) 4 1.341e+00 7.894e-02 16.989  
## ID13(5) Scheduled Tribes (ST) 5 1.610e+00 9.032e-02 17.821  
## ID13(6) Others 6 9.025e-01 1.418e-01 6.363  
## RSUNEARN -1.945e-05 6.522e-07 -29.821  
## INCOME 1.593e-05 6.013e-07 26.493  
## Pr(>|z|)   
## (Intercept) 9.29e-15 \*\*\*  
## EW8(01) 1st class 1 0.30690   
## EW8(02) 2nd class 2 0.19707   
## EW8(03) 3rd class 3 0.88489   
## EW8(04) 4th class 4 0.00438 \*\*   
## EW8(05) 5th class 5 < 2e-16 \*\*\*  
## EW8(06) 6th class 6 1.04e-10 \*\*\*  
## EW8(07) 7th class 7 < 2e-16 \*\*\*  
## EW8(08) 8th class 8 < 2e-16 \*\*\*  
## EW8(09) 9th class 9 < 2e-16 \*\*\*  
## EW8(10) Secondary 10 < 2e-16 \*\*\*  
## EW8(11) 11th Class 11 8.71e-05 \*\*\*  
## EW8(12) High Secondary 12 < 2e-16 \*\*\*  
## EW8(13) 1 year post-secondary 13 0.96344   
## EW8(14) 2 years post-secondary 14 0.23962   
## EW8(15) Bachelors 15 0.45062   
## EW8(16) Above Bachelors 16 3.06e-12 \*\*\*  
## ID11(2) Muslim 2 < 2e-16 \*\*\*  
## ID11(3) Christian 3 0.02148 \*   
## ID11(4) Sikh 4 0.49989   
## ID11(5) Buddhist 5 0.15423   
## ID11(6) Jain 6 0.03231 \*   
## ID11(7) Tribal 7 0.36200   
## ID11(8) Others 8 0.64444   
## ID11(9) None 9 0.71064   
## ID13(2) Forward/General (except Brahmin) 2 2.55e-05 \*\*\*  
## ID13(3) Other Backward Castes (OBC) 3 < 2e-16 \*\*\*  
## ID13(4) Scheduled Castes (SC) 4 < 2e-16 \*\*\*  
## ID13(5) Scheduled Tribes (ST) 5 < 2e-16 \*\*\*  
## ID13(6) Others 6 1.97e-10 \*\*\*  
## RSUNEARN < 2e-16 \*\*\*  
## INCOME < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 34456 on 25360 degrees of freedom  
## Residual deviance: 29725 on 25329 degrees of freedom  
## (48 observations deleted due to missingness)  
## AIC: 29789  
##   
## Number of Fisher Scoring iterations: 6

# ````````````````````````````````````````````````````````````````````````````````````  
  
# FITTING DIFFERENT GLM MODELS  
  
#fit1 <- glm(GR46 ~ EW8+SPED2+INCOME+ID11+ID13, data = women\_par$train, family = binomial(link = "logit"))  
#  
# fit2 <- glm(GR46 ~ (education), data = women\_par$train, family = binomial(link="logit"))  
#   
# fit3 <- glm(GR46 ~ education + INCOME, data = women, family = binomial())  
#   
# fit4 <- glm(GR46 ~ education + RSUNEARN + EW6 + SPED6 , data = women, family = binomial())  
#   
# fit5 <- glm(GR46 ~ education + RSUNEARN + EW6 + SPED6 + ID13, data = women, family = binomial())  
#  
# fit6 <- glm(GR46 ~ EW8 + RSUNEARN + EW6 + SPED6 + ID13 + ID11, data = women\_par$train, family = binomial(link = "logit"))  
  
# ````````````````````````````````````````````````````````````````````````````````````  
  
  
  
  
# Stepwise generalized logistic regression linear regression to check if if adding diff variables improves the model or not.  
  
lm.one <- glm(GR46 ~ education,data=train\_partition,family = binomial(link= "logit"),na.action = na.roughfix)  
  
lm.all <- glm(GR46 ~ EW5+ EW6+ education+ EW9+ EW10+ID11+ ID13+HHEDUC+ HHEDUCM+ HHEDUCF+SPED6+ SPED2+ SPED3+ INCOME+RSUNEARN,data=train\_partition,family = binomial(link= "logit"),na.action = na.roughfix)  
  
step(lm.one,scope=list(upper=lm.all, lower= lm.one), direction = "forward",trace = 1)

## Start: AIC=33134.01  
## GR46 ~ education  
##   
## Df Deviance AIC  
## + ID13 5 31767 31789  
## + HHEDUC 16 32350 32394  
## + SPED6 16 32352 32396  
## + HHEDUCM 16 32414 32458  
## + RSUNEARN 1 32493 32507  
## + ID11 8 32527 32555  
## + HHEDUCF 16 32764 32808  
## + SPED2 1 32827 32841  
## + SPED3 2 32830 32846  
## + INCOME 1 32868 32882  
## + EW5 10 32859 32891  
## + EW10 4 33093 33113  
## + EW6 1 33113 33127  
## <none> 33122 33134  
## + EW9 1 33121 33135  
##   
## Step: AIC=31789.28  
## GR46 ~ education + ID13  
##   
## Df Deviance AIC  
## + HHEDUC 16 31175 31229  
## + SPED6 16 31191 31245  
## + HHEDUCM 16 31241 31295  
## + RSUNEARN 1 31294 31318  
## + HHEDUCF 16 31484 31538  
## + ID11 8 31519 31557  
## + SPED3 2 31545 31571  
## + SPED2 1 31551 31575  
## + EW5 10 31539 31581  
## + INCOME 1 31599 31623  
## + EW10 4 31748 31778  
## <none> 31767 31789  
## + EW6 1 31766 31790  
## + EW9 1 31766 31790  
##   
## Step: AIC=31228.89  
## GR46 ~ education + ID13 + HHEDUC  
##   
## Df Deviance AIC  
## + ID11 8 30839 30909  
## + RSUNEARN 1 30896 30952  
## + EW5 10 30959 31033  
## + SPED6 16 30989 31075  
## + SPED2 1 31102 31158  
## + INCOME 1 31107 31163  
## + SPED3 2 31116 31174  
## + EW6 1 31124 31180  
## + HHEDUCF 16 31118 31204  
## + EW10 4 31154 31216  
## + HHEDUCM 16 31139 31225  
## <none> 31175 31229  
## + EW9 1 31174 31230  
##   
## Step: AIC=30909.31  
## GR46 ~ education + ID13 + HHEDUC + ID11  
##   
## Df Deviance AIC  
## + RSUNEARN 1 30568 30640  
## + EW5 10 30626 30716  
## + SPED6 16 30621 30723  
## + SPED2 1 30749 30821  
## + INCOME 1 30771 30843  
## + SPED3 2 30777 30851  
## + EW6 1 30799 30871  
## + HHEDUCF 16 30785 30887  
## + HHEDUCM 16 30798 30900  
## + EW10 4 30824 30902  
## + EW9 1 30835 30907  
## <none> 30839 30909  
##   
## Step: AIC=30640.13  
## GR46 ~ education + ID13 + HHEDUC + ID11 + RSUNEARN  
##   
## Df Deviance AIC  
## + INCOME 1 29517 29591  
## + SPED6 16 30365 30469  
## + EW5 10 30406 30498  
## + SPED2 1 30482 30556  
## + EW6 1 30517 30591  
## + SPED3 2 30516 30592  
## + HHEDUCF 16 30513 30617  
## + EW10 4 30550 30630  
## + EW9 1 30561 30635  
## + HHEDUCM 16 30534 30638  
## <none> 30568 30640  
##   
## Step: AIC=29590.73  
## GR46 ~ education + ID13 + HHEDUC + ID11 + RSUNEARN + INCOME  
##   
## Df Deviance AIC  
## + SPED6 16 29319 29425  
## + EW5 10 29408 29502  
## + SPED2 1 29432 29508  
## + SPED3 2 29470 29548  
## + EW6 1 29478 29554  
## + EW9 1 29510 29586  
## + EW10 4 29504 29586  
## + HHEDUCF 16 29482 29588  
## <none> 29517 29591  
## + HHEDUCM 16 29487 29593  
##   
## Step: AIC=29425.37  
## GR46 ~ education + ID13 + HHEDUC + ID11 + RSUNEARN + INCOME +   
## SPED6  
##   
## Df Deviance AIC  
## + EW5 10 29236 29362  
## + HHEDUCF 16 29270 29408  
## + EW6 1 29306 29414  
## + SPED3 2 29310 29420  
## + EW10 4 29307 29421  
## + SPED2 1 29315 29423  
## + EW9 1 29316 29424  
## <none> 29319 29425  
## + HHEDUCM 16 29302 29440  
##   
## Step: AIC=29362.37  
## GR46 ~ education + ID13 + HHEDUC + ID11 + RSUNEARN + INCOME +   
## SPED6 + EW5  
##   
## Df Deviance AIC  
## + HHEDUCF 16 29189 29347  
## + SPED3 2 29223 29353  
## + EW10 4 29224 29358  
## + SPED2 1 29234 29362  
## <none> 29236 29362  
## + EW9 1 29236 29364  
## + EW6 1 29236 29364  
## + HHEDUCM 16 29223 29381  
##   
## Step: AIC=29346.53  
## GR46 ~ education + ID13 + HHEDUC + ID11 + RSUNEARN + INCOME +   
## SPED6 + EW5 + HHEDUCF  
##   
## Df Deviance AIC  
## + SPED3 2 29175 29337  
## + EW10 4 29177 29343  
## + SPED2 1 29186 29346  
## <none> 29189 29347  
## + EW6 1 29187 29347  
## + EW9 1 29188 29348  
## + HHEDUCM 16 29170 29360  
##   
## Step: AIC=29337.19  
## GR46 ~ education + ID13 + HHEDUC + ID11 + RSUNEARN + INCOME +   
## SPED6 + EW5 + HHEDUCF + SPED3  
##   
## Df Deviance AIC  
## + EW10 4 29164 29334  
## + SPED2 1 29173 29337  
## <none> 29175 29337  
## + EW6 1 29174 29338  
## + EW9 1 29175 29339  
## + HHEDUCM 16 29156 29350  
##   
## Step: AIC=29333.53  
## GR46 ~ education + ID13 + HHEDUC + ID11 + RSUNEARN + INCOME +   
## SPED6 + EW5 + HHEDUCF + SPED3 + EW10  
##   
## Df Deviance AIC  
## + SPED2 1 29161 29333  
## <none> 29164 29334  
## + EW6 1 29162 29334  
## + EW9 1 29163 29335  
## + HHEDUCM 16 29145 29347  
##   
## Step: AIC=29333.3  
## GR46 ~ education + ID13 + HHEDUC + ID11 + RSUNEARN + INCOME +   
## SPED6 + EW5 + HHEDUCF + SPED3 + EW10 + SPED2  
##   
## Df Deviance AIC  
## <none> 29161 29333  
## + EW6 1 29160 29334  
## + EW9 1 29161 29335  
## + HHEDUCM 16 29142 29346

##   
## Call: glm(formula = GR46 ~ education + ID13 + HHEDUC + ID11 + RSUNEARN +   
## INCOME + SPED6 + EW5 + HHEDUCF + SPED3 + EW10 + SPED2, family = binomial(link = "logit"),   
## data = train\_partition, na.action = na.roughfix)  
##   
## Coefficients:  
## (Intercept)   
## -3.457e-01   
## educationpreprimary   
## 1.702e-01   
## educationprimary & postprimary   
## -2.022e-01   
## educationsecondary   
## -1.398e-01   
## educationhigher secondary   
## 3.463e-01   
## educationcollege graduate or higher   
## 1.197e+00   
## ID13(2) Forward/General (except Brahmin) 2   
## 2.381e-01   
## ID13(3) Other Backward Castes (OBC) 3   
## 6.970e-01   
## ID13(4) Scheduled Castes (SC) 4   
## 1.166e+00   
## ID13(5) Scheduled Tribes (ST) 5   
## 1.405e+00   
## ID13(6) Others 6   
## 7.511e-01   
## HHEDUC(01) 1st class 1   
## 7.056e-02   
## HHEDUC(02) 2nd class 2   
## 1.530e-01   
## HHEDUC(03) 3rd class 3   
## -1.639e-01   
## HHEDUC(04) 4th class 4   
## 9.504e-02   
## HHEDUC(05) 5th class 5   
## -9.154e-02   
## HHEDUC(06) 6th class 6   
## 1.259e-01   
## HHEDUC(07) 7th class 7   
## -1.631e-01   
## HHEDUC(08) 8th class 8   
## -3.350e-02   
## HHEDUC(09) 9th class 9   
## -5.691e-02   
## HHEDUC(10) Secondary 10   
## -1.971e-01   
## HHEDUC(11) 11th Class 11   
## -3.664e-02   
## HHEDUC(12) High Secondary 12   
## -3.424e-01   
## HHEDUC(13) 1 year post-secondary 13   
## -2.386e-01   
## HHEDUC(14) 2 years post-secondary 14   
## -5.419e-01   
## HHEDUC(15) Bachelors 15   
## -4.445e-01   
## HHEDUC(16) Above Bachelors 16   
## -4.736e-01   
## ID11(2) Muslim 2   
## -7.588e-01   
## ID11(3) Christian 3   
## 2.428e-01   
## ID11(4) Sikh 4   
## -9.457e-02   
## ID11(5) Buddhist 5   
## -2.183e-01   
## ID11(6) Jain 6   
## -6.581e-01   
## ID11(7) Tribal 7   
## 2.235e-01   
## ID11(8) Others 8   
## 2.101e-01   
## ID11(9) None 9   
## -4.444e-01   
## RSUNEARN   
## -1.736e-05   
## INCOME   
## 1.481e-05   
## SPED6(01) 1st class 1   
## 1.493e-01   
## SPED6(02) 2nd class 2   
## 5.276e-02   
## SPED6(03) 3rd class 3   
## 1.931e-01   
## SPED6(04) 4th class 4   
## -1.040e-01   
## SPED6(05) 5th class 5   
## -1.705e-01   
## SPED6(06) 6th class 6   
## -5.422e-01   
## SPED6(07) 7th class 7   
## -2.449e-01   
## SPED6(08) 8th class 8   
## -3.746e-01   
## SPED6(09) 9th class 9   
## -4.089e-01   
## SPED6(10) Secondary 10   
## -5.744e-01   
## SPED6(11) 11th Class 11   
## -3.634e-01   
## SPED6(12) High Secondary 12   
## -5.024e-01   
## SPED6(13) 1 year post-secondary 13   
## -6.565e-01   
## SPED6(14) 2 years post-secondary 14   
## -1.649e-01   
## SPED6(15) Bachelors 15   
## -6.345e-01   
## SPED6(16) Above Bachelors 16   
## -3.496e-01   
## EW5(02) Wife 2   
## 2.078e-01   
## EW5(03) Daughter 3   
## 2.816e-01   
## EW5(04) Daughter-in-law 4   
## -1.049e-01   
## EW5(05) Grandchild 5   
## 9.356e+00   
## EW5(06) Mother 6   
## -3.712e-01   
## EW5(07) Sister 7   
## 2.103e-01   
## EW5(08) Mother-in-law 8   
## 2.466e+00   
## EW5(09) Niece 9   
## 1.601e-01   
## EW5(10) Sister in law 10   
## -3.269e-01   
## EW5(11) Other relatives 11   
## -1.507e-01   
## HHEDUCF(01) 1st class 1   
## -7.424e-02   
## HHEDUCF(02) 2nd class 2   
## -1.725e-01   
## HHEDUCF(03) 3rd class 3   
## -2.423e-02   
## HHEDUCF(04) 4th class 4   
## -3.185e-01   
## HHEDUCF(05) 5th class 5   
## -1.047e-01   
## HHEDUCF(06) 6th class 6   
## -2.091e-01   
## HHEDUCF(07) 7th class 7   
## -1.046e-01   
## HHEDUCF(08) 8th class 8   
## -2.729e-01   
## HHEDUCF(09) 9th class 9   
## -3.700e-01   
## HHEDUCF(10) Secondary 10   
## -2.418e-01   
## HHEDUCF(11) 11th Class 11   
## -1.341e-01   
## HHEDUCF(12) High Secondary 12   
## -2.970e-01   
## HHEDUCF(13) 1 year post-secondary 13   
## -2.776e-01   
## HHEDUCF(14) 2 years post-secondary 14   
## -5.447e-02   
## HHEDUCF(15) Bachelors 15   
## -4.512e-01   
## HHEDUCF(16) Above Bachelors 16   
## 8.004e-02   
## SPED3(1) Little 1   
## -1.542e-01   
## SPED3(2) Fluent 2   
## -1.124e-01   
## EW10(2) Good 2   
## 7.903e-02   
## EW10(3) OK 3   
## 6.704e-02   
## EW10(4) Poor 4   
## -7.210e-02   
## EW10(5) Very poor 5   
## -2.074e-01   
## SPED2(1) Yes 1   
## -1.123e-01   
##   
## Degrees of Freedom: 25408 Total (i.e. Null); 25323 Residual  
## Null Deviance: 34520   
## Residual Deviance: 29160 AIC: 29330

# Trying multinomial regression   
  
fit1m <- multinom(GR46 ~ EW8+EW9+EW10+SPED2+INCOME+ID11+ID13, data = train\_partition,hess=TRUE)

## # weights: 38 (37 variable)  
## initial value 16505.913811   
## iter 10 value 14837.871086  
## iter 20 value 14610.607847  
## iter 30 value 14561.240824  
## iter 40 value 14537.597916  
## final value 14537.312773   
## converged

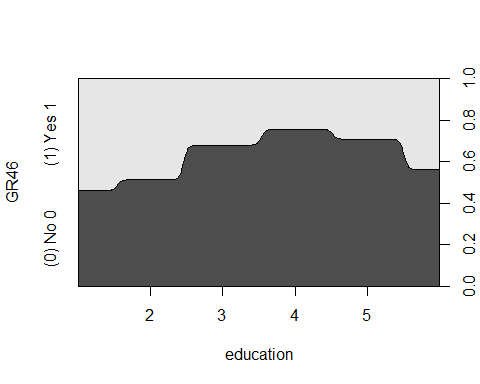
# Caluculating prediction values ( to be run everytime after running the model)  
   
pred <- as\_tibble(predict.glm(fit1, test\_partition,na.action = na.pass))  
  
  
# Converting predictions to match the levels of the response variable  
  
pred\_fc <- mutate(pred, value = ifelse(pred > 0.5,"(1) Yes 1","(0) No 0"))  
p\_class <- factor(pred\_fc$value, levels = levels(women$GR46))  
  
  
  
# Displaying Confusion Matrix   
  
Cmatrix <- confusionMatrix( p\_class, test\_partition$GR46, dnn = c("Prediction", "Reference"))  
Cmatrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction (0) No 0 (1) Yes 1  
## (0) No 0 3365 1778  
## (1) Yes 1 319 883  
##   
## Accuracy : 0.6695   
## 95% CI : (0.6578, 0.6811)  
## No Information Rate : 0.5806   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.2654   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9134   
## Specificity : 0.3318   
## Pos Pred Value : 0.6543   
## Neg Pred Value : 0.7346   
## Prevalence : 0.5806   
## Detection Rate : 0.5303   
## Detection Prevalence : 0.8106   
## Balanced Accuracy : 0.6226   
##   
## 'Positive' Class : (0) No 0   
##

cat("The F Score is",Cmatrix$byClass["F1"])

## The F Score is 0.7624334

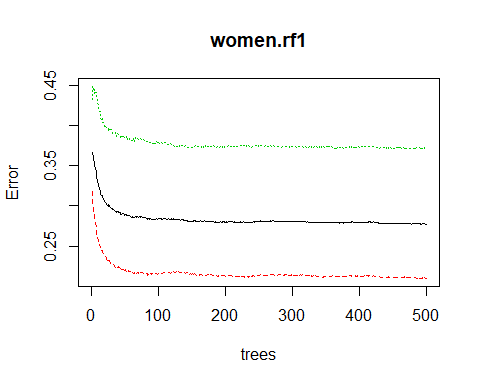
cdplot(GR46 ~ education, data=women)



# Running randomForest on "women" data.  
  
women.rf1=randomForest(GR46 ~EW6+ education+ SPED6+ SPRO5+ INCOME+ RSUNEARN, data = train\_partition, na.action = na.roughfix, mtry=6)  
women.rf1

##   
## Call:  
## randomForest(formula = GR46 ~ EW6 + education + SPED6 + SPRO5 + INCOME + RSUNEARN, data = train\_partition, mtry = 6, na.action = na.roughfix)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 6  
##   
## OOB estimate of error rate: 27.72%  
## Confusion matrix:  
## (0) No 0 (1) Yes 1 class.error  
## (0) No 0 11714 3107 0.209635  
## (1) Yes 1 3937 6651 0.371836

plot(women.rf1)



# Calculating predictions for randomForest  
  
cl <- as\_tibble(predict(women.rf1, test\_partition))  
p\_cl <- factor(cl$value, levels = levels(women$GR46))  
  
  
# Displaying COnfusion Matrix  
  
Cmatrix1 <- confusionMatrix( p\_cl,test\_partition$GR46, dnn = c("Prediction", "Reference"))  
Cmatrix1

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction (0) No 0 (1) Yes 1  
## (0) No 0 2847 1011  
## (1) Yes 1 823 1641  
##   
## Accuracy : 0.7099   
## 95% CI : (0.6985, 0.7211)  
## No Information Rate : 0.5805   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.3984   
## Mcnemar's Test P-Value : 1.262e-05   
##   
## Sensitivity : 0.7757   
## Specificity : 0.6188   
## Pos Pred Value : 0.7379   
## Neg Pred Value : 0.6660   
## Prevalence : 0.5805   
## Detection Rate : 0.4503   
## Detection Prevalence : 0.6102   
## Balanced Accuracy : 0.6973   
##   
## 'Positive' Class : (0) No 0   
##

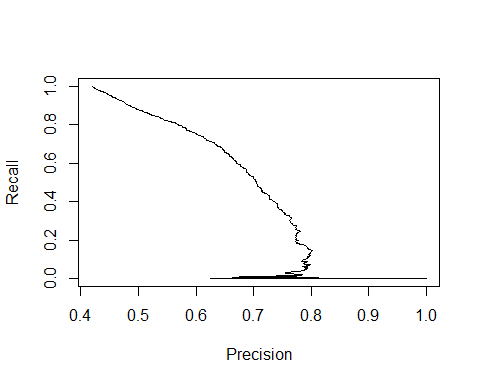
cat("The F Score is",Cmatrix1$byClass["F1"])

## The F Score is 0.7563762

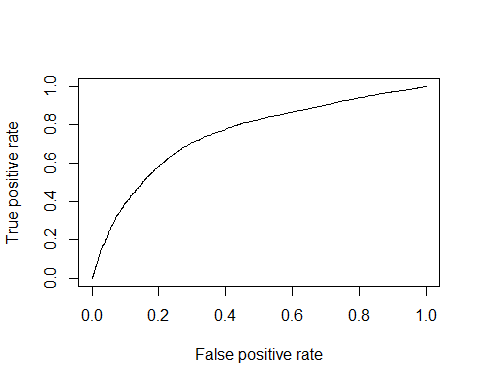
importance(women.rf1)

## MeanDecreaseGini  
## EW6 1511.9517  
## education 633.4431  
## SPED6 1589.5760  
## SPRO5 1561.9494  
## INCOME 3428.3759  
## RSUNEARN 3621.4756

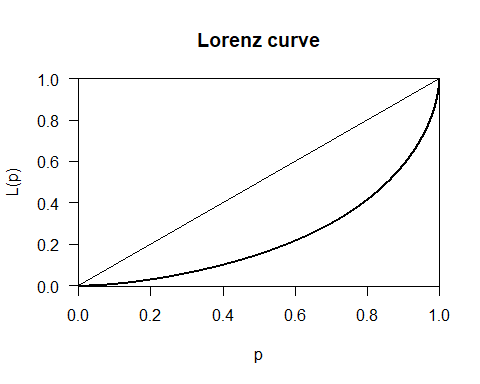
# ````````````````````````````````````````````````````````````````````````````````````  
  
# Plotting Precision , Recall and TPR/FPR graph (ROC)   
  
OOB.votes <- predict (women.rf1,test\_partition,type="prob")  
OOB.pred <- OOB.votes[,2]  
  
pred.obj <- prediction (OOB.pred,test\_partition$GR46)  
  
RP.perf <- performance(pred.obj, "rec","prec")  
plot (RP.perf)



ROC.perf <- performance(pred.obj, "tpr","fpr")  
plot (ROC.perf)



library(tidyverse)  
library(dplyr)  
  
# Loading DS1 Dataframe   
  
load("C:/Users/Akshay/Desktop/ICPSR\_36151/DS0001/36151-0001-Data.rda")  
load("C:/Users/Akshay/Desktop/ICPSR\_36151/DS0010/36151-0010-Data.rda")  
  
ds1 <- as\_tibble(da36151.0001)  
  
ds1 <- dplyr::select(ds1, HHID, NF5, NF25, NF45, INCCROP, INCAGPROP, INCANIMAL, INCAG, INCBUS,   
 INCOTHER, INCEARN, INCBENEFITS, INCOME, INCREMIT, INCOMEPC, WS3NM, WS4, WS5,   
 WS7MONTHS, WSEARN, WS12, WSEARNAGLAB, WSEARNNONAG, WSEARNSALARY, WSEARNNREGA,   
 RSUNEARN, INCNONAG, INCAGLAB, INCSALARY, INCNREGA, INCNONNREGA, HHEDUC, HHEDUCM,   
 HHEDUCF, MG10)  
  
ds10 <- as\_tibble(da36151.0010)  
  
# Loading IHDS 2005 data from tsv files (no .rda available)  
  
ihds2005 <- as\_tibble(read.delim('C:/DMP/22626-0001-Data.tsv',sep = "\t"))  
  
  
  
# preliminary LC and GIni Plots   
  
x <- ineq(ds1$INCOME,type= "Gini")  
  
y <- Lc(ds1$INCOME)  
  
plot(y)



# Plotting Gini Curves for 2004-05 and 2011-12 IHDS Data.  
  
  
font.add.google("Poppins", "myfont")

## 'font.add.google()' is now renamed to 'font\_add\_google()'  
## The old version still works, but consider using the new function in future code

showtext.auto()

## 'showtext.auto()' is now renamed to 'showtext\_auto()'  
## The old version still works, but consider using the new function in future code

Income05 <- ihds2005$income  
Income <- ds1$INCOME  
  
lorenz05 <- Lc(Income05)  
lorenz\_df1 <- data.frame(prop\_pop = lorenz05$p, income = lorenz05$L) %>%  
 mutate(prop\_equality = prop\_pop)  
  
ineq(Income05,type="Gini")

## [1] 0.5302466

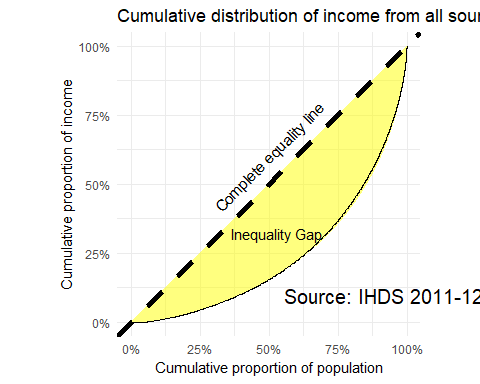
ineq(Income,type="Gini")

## [1] 0.5448205

lorenz11 <- Lc(Income)  
lorenz\_df2 <- data.frame(prop\_pop = lorenz11$p, income = lorenz11$L) %>%  
 mutate(prop\_equality = prop\_pop)  
  
p1 <- ggplot(lorenz\_df1, aes(x = prop\_pop, y = income)) +  
 geom\_ribbon(data=lorenz\_df2,aes(ymax = prop\_equality, ymin = income), fill = "yellow",alpha=0.5)+  
 geom\_line() +  
 geom\_abline(slope = 1, xintercept = 0, type="l", lty=2,lwd=2) +  
 scale\_x\_continuous("Cumulative proportion of population", label = percent) +  
 scale\_y\_continuous("Cumulative proportion of income", label = percent) +  
 theme\_minimal(base\_family = "myfont") +  
 coord\_equal() +  
 annotate("text",0.53, 0.32, label = "Inequality Gap", family = "myfont") +  
 annotate("text", 0.5 , 0.6, label = "Complete equality line", angle = 45, family = "myfont") +   
 ggtitle (  
 str\_wrap("Cumulative distribution of income from all sources", 200))

## Warning: Ignoring unknown parameters: xintercept, type

print(p1)  
  
grid.text("Source: IHDS 2011-12", 0.8, 0.23,   
 gp = gpar(fontfamily = "myfont", fontsize = 15))



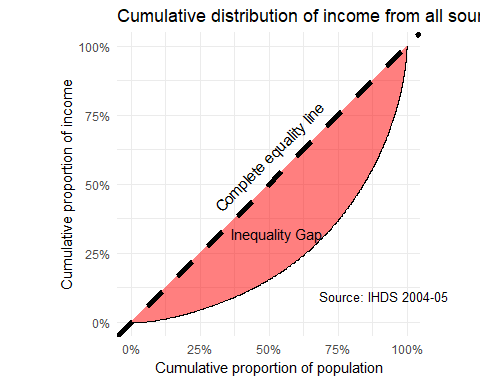
ggsave("IncomeEquality.jpeg")

## Saving 5 x 4 in image

font.add.google("Poppins", "myfont")  
showtext.auto()  
  
  
p2 <- ggplot(lorenz\_df1, aes(x = prop\_pop, y = income)) +  
 geom\_ribbon(aes(ymax = prop\_equality, ymin = income), fill = "red",alpha=0.5) +  
 geom\_line() +  
 geom\_abline(slope = 1, xintercept = 0, type="l", lty=2,lwd=2) +  
 scale\_x\_continuous("Cumulative proportion of population", label = percent) +  
 scale\_y\_continuous("Cumulative proportion of income", label = percent) +  
 theme\_minimal(base\_family = "myfont") +  
 coord\_equal() +  
 annotate("text",0.53, 0.32, label = "Inequality Gap", family = "myfont") +  
 annotate("text", 0.5 , 0.6, label = "Complete equality line", angle = 45, family = "myfont") +   
 ggtitle (  
 str\_wrap("Cumulative distribution of income from all sources", 200))

## Warning: Ignoring unknown parameters: xintercept, type

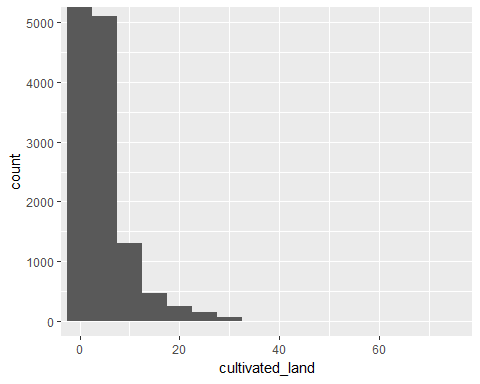
print(p2)  
  
grid.text("Source: IHDS 2004-05", 0.8, 0.23,   
 gp = gpar(fontfamily = "myfont", fontsize = 10))



# loading Agrarian Data   
  
load("C:/Users/Akshay/Desktop/ICPSR\_36151/DS0002/36151-0002-Data.rda")  
  
  
household\_data <- da36151.0002  
View(head(household\_data,20))  
  
# Sorting and Selecting relevant variables for the model  
  
agrarian <- dplyr::select(household\_data, HHID, ID14, FM2, FM3,FM4A,FM5A,FM6A,FM4B,FM5B, FM6B,FM4C,FM26A, FM26B,  
 FM5C, FM6C, FM40E, FM11A, FM11B, FM11C, FM27B, FM29, FM30, FM31, FM32, FM33, FM34, FM29RS, FM30RS, FM40E)  
  
# FM29RS - "Rupees spent last year on fertilizer and manure"  
# FM30RS - "Rs last year on herbicides and pesticides"  
# FM27A - "Hired farm labour days"  
# ID14 - "Main income source"  
# FM2 - "Local area unit name"  
# FM3 - "Local units/acre"  
# FM4A - "Owned kharif"  
# FM5A - "Rented in kharif"  
# FM6A - "Rented out kharif"  
# FM4B - "Owned rabi"  
# FM5B - "Rented in rabi"  
# FM6B - "Rented out rabi"  
# FM4C - "Owned summer"  
# FM5C - "Rented in summer"  
# FM6C - "Rented out summer"  
# FM40E - "Tractors/Tillers"  
# FM11A - "Cultivated kharif"  
# FM11B - "Cultivated rabi"  
# FM11C - "Cultivated summer"  
# FM26A - "Crop residue total value (rupees)"  
# FM26B - "Crop residue sold (rupees)"  
# FM27B - "Hired farm labour Rs"  
# FM29: - "Fertilizers Rs"  
# FM30: - "Pesticides Rs"  
# FM31: - "Irrigation water Rs"  
# FM32: - "Hired Equipment/Animals Rs"  
# FM33: - "Agriculture loan repayment Rs"  
# FM34: - "Farm miscellaneous Rs"  
  
  
agrarian <- agrarian %>%  
 mutate(cultivated\_land = (FM11A + FM11B + FM11C) /FM3)  
  
#agrarian <- na.omit(agrarian)  
  
agrarian <- filter(agrarian, cultivated\_land <= 30)  
  
summary(agrarian$cultivated\_land)

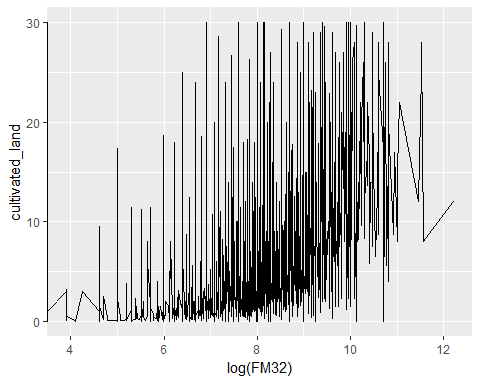
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.750 2.000 3.508 4.125 30.000

ggplot(agrarian) + geom\_histogram(aes(cultivated\_land), binwidth = 5) + coord\_cartesian(xlim = c(0, 75), ylim = c(0,5000))



#Partitioning data for training and Testing  
  
agrarian1 <- resample\_partition(agrarian, c(train = 0.8,test = 0.1, valid = 0.1))  
  
agrarian\_train <- as\_tibble(agrarian1$train)  
agrarian\_valid <- as\_tibble(agrarian1$valid)  
agrarian\_test <- as\_tibble(agrarian1$test)  
  
  
train=sample(1:nrow(agrarian\_train),nrow(agrarian\_train))  
  
  
#test = sample(1:nrow(agrarian\_test),nrow(agrarian\_test))   
#valid = sample(1:nrow(agrarian\_valid),nrow(agrarian\_valid))   
#agri.rf1=randomForest(cultivated\_land ~ FM27B+FM29+FM30+FM31+FM32+FM33+FM34, data = agrarian, subset = train,na.action = na.roughfix)  
  
  
  
# Running randomForest   
  
agri.rf1=randomForest(cultivated\_land ~ FM27B+FM29RS+FM30RS+FM31+FM32+FM34+FM26A, data = agrarian\_train,   
 na.action = na.roughfix, ntree = 200)  
  
ggplot(agrarian) + geom\_line(mapping = aes( x = log(FM32) , y = cultivated\_land))

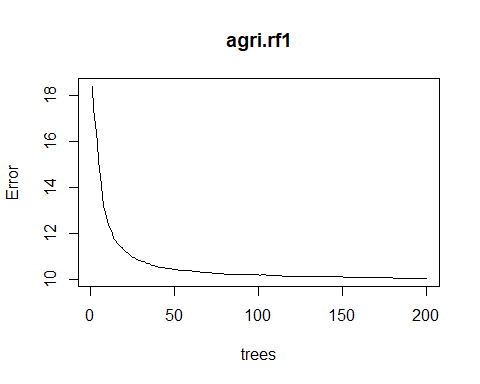
## Warning: Removed 2249 rows containing missing values (geom\_path).



agri.rf1

##   
## Call:  
## randomForest(formula = cultivated\_land ~ FM27B + FM29RS + FM30RS + FM31 + FM32 + FM34 + FM26A, data = agrarian\_train, ntree = 200, na.action = na.roughfix)   
## Type of random forest: regression  
## Number of trees: 200  
## No. of variables tried at each split: 2  
##   
## Mean of squared residuals: 10.05006  
## % Var explained: 52.86

plot(agri.rf1)

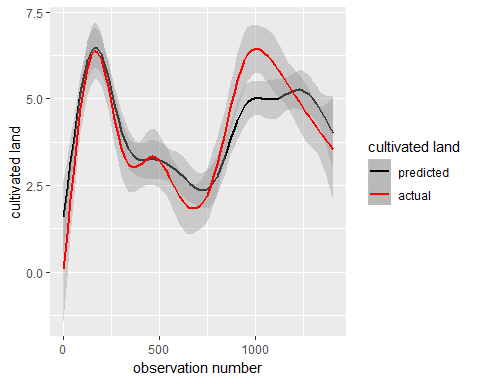


importance(agri.rf1)

## IncNodePurity  
## FM27B 46137.22  
## FM29RS 71392.99  
## FM30RS 41500.69  
## FM31 10540.73  
## FM32 32390.92  
## FM34 33707.71  
## FM26A 42644.67

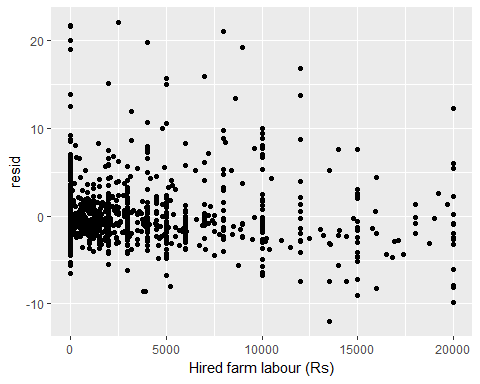
# Plotting Actual vs Predicted cultvated land values to visually test the model fit   
  
  
cl <- as\_tibble(predict(agri.rf1, na.omit(agrarian\_test)))  
test\_omit <- na.omit(agrarian\_test)  
x1 <- seq(1:nrow(test\_omit))  
df <- mutate(cl,arpita = row\_number())  
df1 <- mutate(test\_omit, no\_rows = row\_number())  
ggplot() + geom\_smooth(data = df, mapping = aes(x = arpita, y = value,color='black')) +   
geom\_smooth(data = df1, mapping = aes(x= no\_rows, y = df1$cultivated\_land,color='red')) +  
 scale\_colour\_manual(name = 'cultivated land',   
 values =c('black'='black','red'='red'), labels = c('predicted','actual')) +   
 labs(x = "observation number", y = "cultivated land")

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'  
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



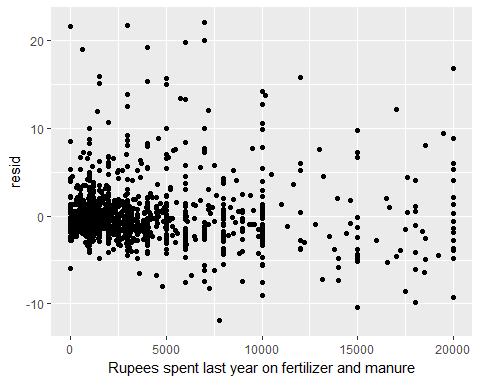
# Plotting Residuals against predictor variables  
  
agrarian\_test %>% add\_residuals(agri.rf1) %>%  
 ggplot(aes(x=FM27B, y=resid)) + geom\_point() + scale\_x\_continuous(limits = c(0,20000)) +   
 labs(x = "Hired farm labour (Rs)")

## Warning: Removed 439 rows containing missing values (geom\_point).



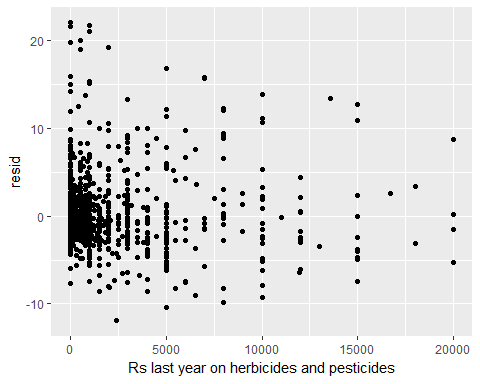
agrarian\_test %>% add\_residuals(agri.rf1) %>%  
 ggplot(aes(x=FM29RS, y=resid)) + geom\_point() + scale\_x\_continuous(limits = c(0,20000)) +   
 labs(x = "Rupees spent last year on fertilizer and manure")

## Warning: Removed 402 rows containing missing values (geom\_point).



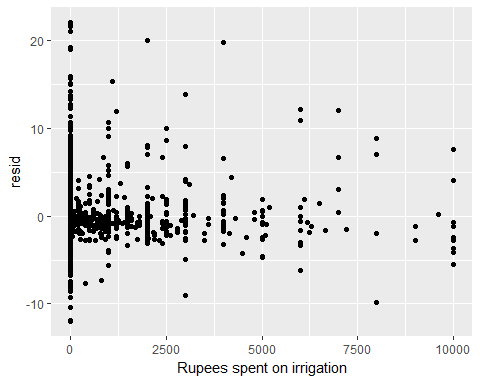
agrarian\_test %>% add\_residuals(agri.rf1) %>%  
 ggplot(aes(x=FM30RS, y=resid)) + geom\_point() + scale\_x\_continuous(limits = c(0,20000)) +   
 labs(x = "Rs last year on herbicides and pesticides")

## Warning: Removed 354 rows containing missing values (geom\_point).



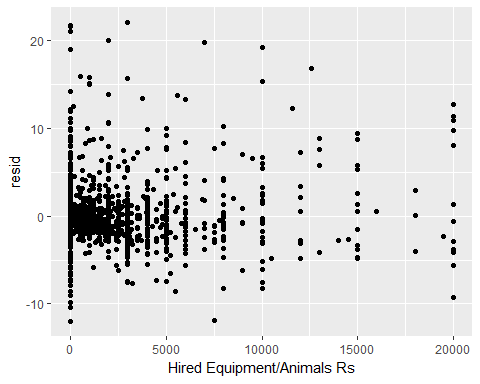
agrarian\_test %>% add\_residuals(agri.rf1) %>%  
 ggplot(aes(x=FM31, y=resid)) + geom\_point() + scale\_x\_continuous(limits = c(0,10000)) +   
 labs(x = "Rupees spent on irrigation")

## Warning: Removed 361 rows containing missing values (geom\_point).



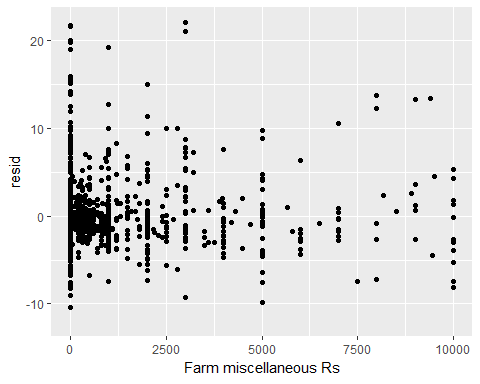
agrarian\_test %>% add\_residuals(agri.rf1) %>%  
 ggplot(aes(x=FM32, y=resid)) + geom\_point() + scale\_x\_continuous(limits = c(0,20000)) +   
 labs(x = "Hired Equipment/Animals Rs")

## Warning: Removed 359 rows containing missing values (geom\_point).



agrarian\_test %>% add\_residuals(agri.rf1) %>%  
 ggplot(aes(x=FM34, y=resid)) + geom\_point() + scale\_x\_continuous(limits = c(0,10000)) +   
 labs(x = "Farm miscellaneous Rs")

## Warning: Removed 415 rows containing missing values (geom\_point).



# total costs evaluations  
  
df\_test1 <- mutate(agrarian, money\_invested = FM29 + FM32 +   
 FM31 + FM30 + FM27B + FM34, unsold\_residue = FM26A - FM26B)%>%   
 filter(ID14=="(01) Cultivation 1") %>%  
 summarise(total\_money = sum(money\_invested, na.rm = TRUE),   
 Crop\_residue = sum(FM26A, na.rm = TRUE),   
 Residue\_sold = sum(FM26B, na.rm = TRUE),  
 unsold\_residue = sum(unsold\_residue, na.rm = TRUE))