Investigating the Effects of Household Factors on Agricultural Value of Production in Tanzania

Yifan Chen yffch@ucdavis.edu

Yuyao Nie yynie@ucdavis.edu

Department of Statistics and Biostatistics, University of California, Davis

STAT 206: Statistical Methods for Research

Prof. Jie. Peng

Dec 9 2024

Abstract

Agricultural production occupies an important part in sub-Saharan Africa. As a main source of income in Tanzania, it influences the growth of the economy and local employment. Although the government of Tanzania has tried several ways to push this sector, it still does not perform well considering the vulnerability of the agricultural sector to frequently changing conditions nowadays, including extreme weather. This article explores the effects of households' socio-economic status on the agricultural value of production at the household level in the regions of Kongwa and Kiteto, two representative drought-stricken regions in Tanzania. The article examines the importance of household factors, such as household income, age of household heads, gender of household heads, household size, etc. The finding shows that efforts needed to be taken by local governments to increase the households' farm sizes, improve the households ability to access off-farm income, and shorten households' distance to the capital market.

I. Introduction

This dataset is obtained from the household survey conducted by the International Institute of Tropical Agriculture in 2022. The survey was to collect data on the agricultural performance of households in regions of Kongwa and Kiteto, two representative drought-stricken regions in Tanzania. This dataset includes 578 observations and 22 variables, containing detailed information on agricultural households in terms of socio-economic factors and informational access. A more detailed description of specific factors included in the study can be found in Table 1 in the appendix.

In this study, we want to explore factors that influence the agricultural value of production of households in drought-stricken regions in Tanzania. Knowing what factors influence the agricultural value of production of households in drought-stricken regions in Tanzania could help local governments in policy design in the response of increasing durability of local households against frequently changing circumstances like extreme weather or climate change. In addition, factors that affect the value of production of households in drought-stricken regions in Tanzania may be applied in other drought-stricken areas in Africa. Therefore, the same policies can also be used in those regions to help improve the value of production of households.

II. Methods and Results

(i) Exploratory data analysis

First, determine the type of each variable in preparation for model building. Considering NA values would cause larger errors in the model, we drop observations with NA values and deal with the remaining 544 effective cases. In addition, based on the eye observation of the dataset, we identify the significant variance among different quantitative variables, so we choose to standardize quantitative variables in predictors to minimize model errors.

Then, use a histogram or a pie chart to see the distribution of each variable. Quantitative variables HDDS, Ageh and Yearsvillage follow nearly normal distributions, and the rest of the quantitative variables are all right-skewed. Qualitative variables District and nonfarm_income are nearly symmetric, and the rest of the qualitative variables are all skewed.

Figure 1: Histograms of quantitative variables

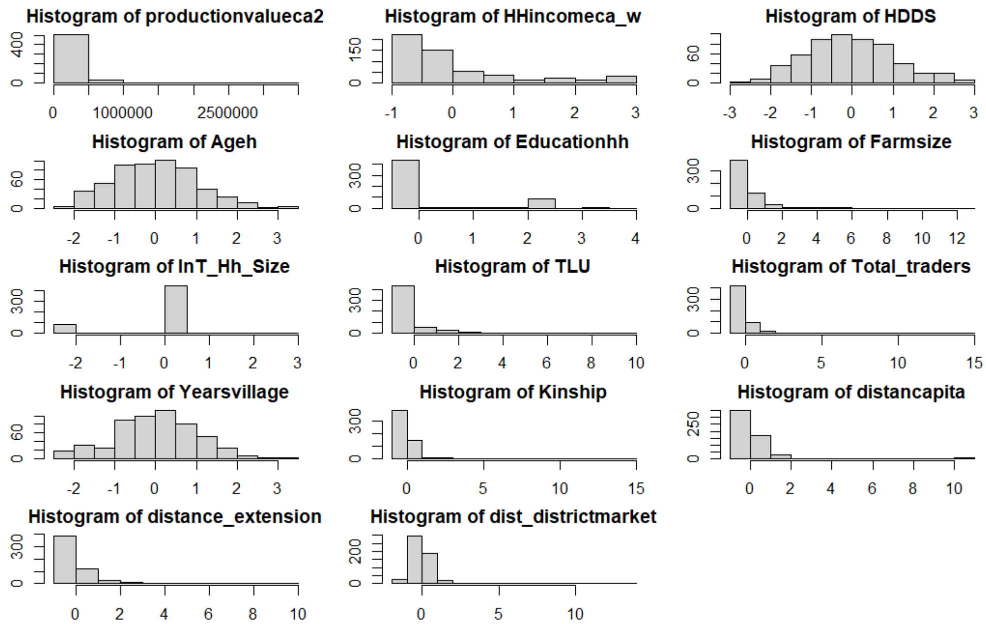
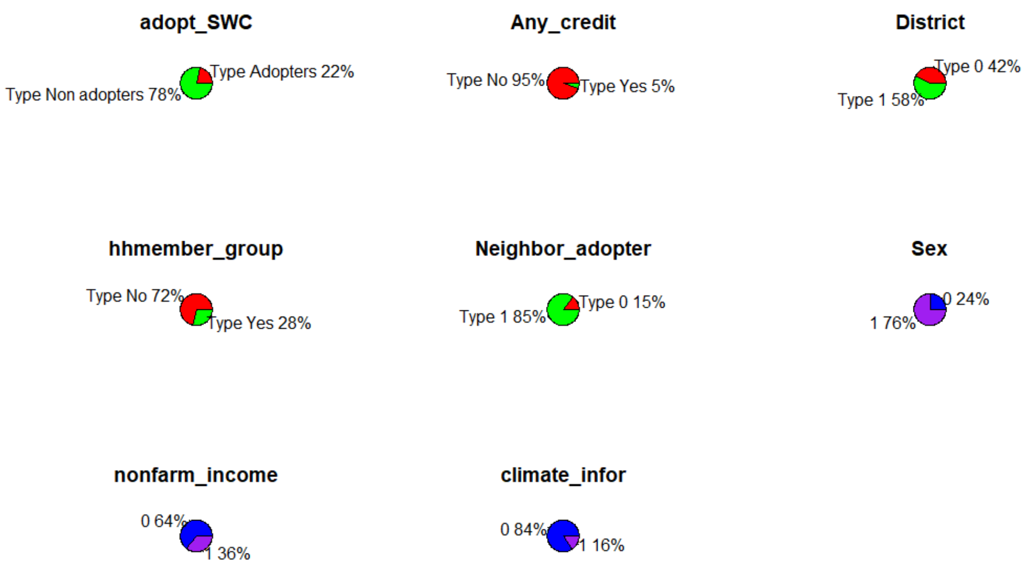


Figure 2: Pie charts with percentage of qualitative variables



Finally, use a scatter plot matrix among quantitative variables with the lower panel showing correlation coefficients and side-by-side box plots to find relationships among variables. There are most non-linear correlations among quantitative variables. By eyeball observation over boxplot, we see obvious differences in Any_credit, District, sex, non-farm income, and whether access to climate information.

Figure 3: Scatter plot matrix of quantitative variables

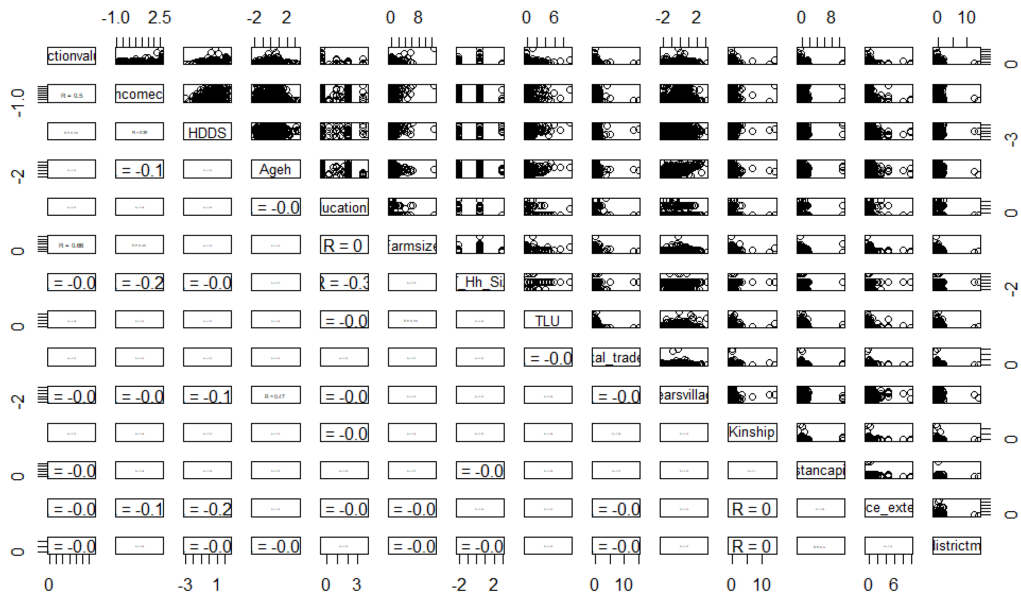
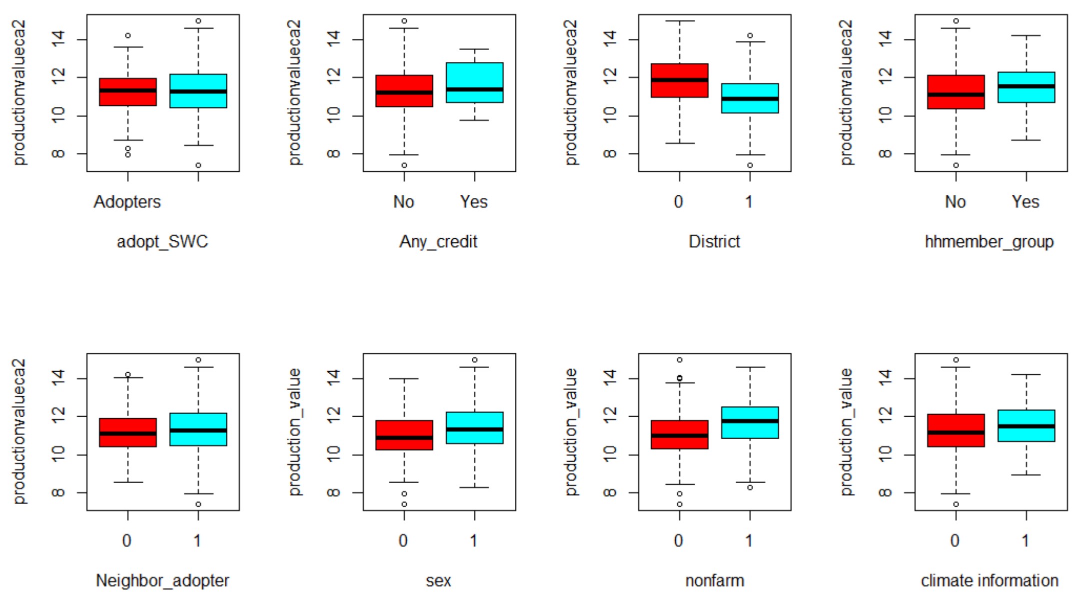


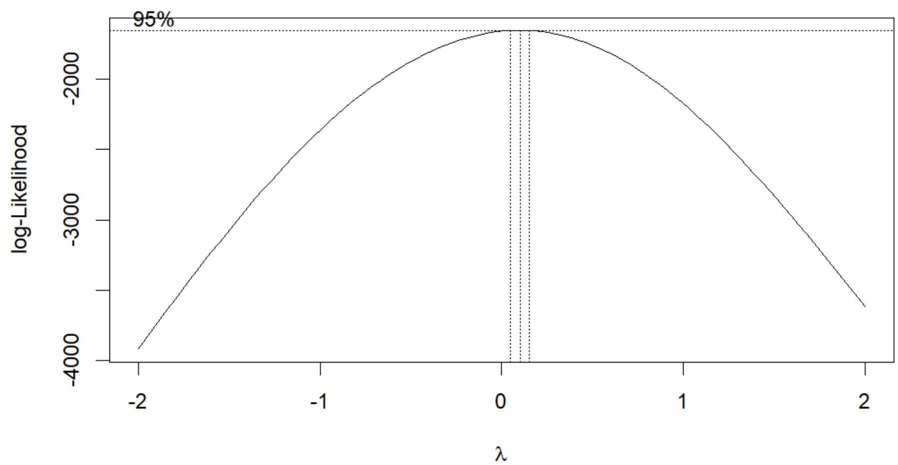
Figure 4: productionvalueca2 side-by-side box plots



(ii) Preliminary model investigation

We use 70% of the dataset as training data to build the model and use the other 30% as validation data to perform model validation. To begin the preliminary fitting step, we first consider fitting a first-order model with the value of production as the response variable, and the rest of the 21 variables as independent variables. Then use the Box-Cox procedure to find whether the response variable needs transformation.

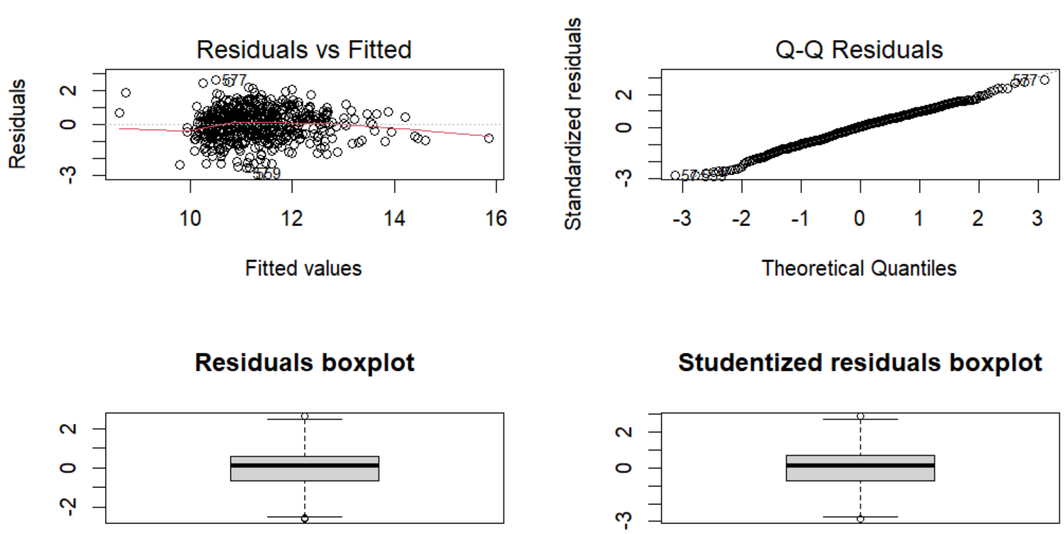
Figure 5: Box-cox procedure plot



Because λ is nearly equal to zero, the Box-Cox procedure suggests logarithmic transformation of the response variable. Therefore, we use logarithmic transformation of the response variable to replace the original response variable.

Fit the first-order model, then draw residual plots to determine whether the preliminary fit is enough. From these residual plots, there appears to be linearity in the regression relation, and no influential outliers exist.

Figure 6: Residual plots



To determine whether high multicollinearity exists, we use VIF. After summarizing VIF among quantitative variables, we found the mean of VIF is 1.268 so it is small enough, which means that any collinearity is not severe enough to adversely affect the regression coefficients or their interpretation, so any interaction terms and high-order power terms do not need to be included in the model.

Based on these preliminary fits, we decide to use logarithmic transformation of the value of production as the response variable and not include any interaction terms or high order power terms because of the low VIF value.

So we could treat the first-order model as the full model, and all X variables of the full model are in the potential pool of X variables for subsequent analysis.

(iii) Model selection

We assume the full model is a correct model, then candidate models are sub-models that contain a subset of 21 independent variables. To choose a good sub-model, we use a forward stepwise procedure, which consists of one forward selection step followed by one backward elimination step, using AIC as the selection criterion.

Finally, the selected sub-model, found by the forward stepwise procedure, includes nine independent variables: HHincomeca_w, District, acess_nonfarm, Farmsize, distancapita, TLU, lnT_Hh_Size, Yearsvillage, and hhmember_group.

(iv) Model validation

Through comparing quantitative box plots between training data and validation data, we find that they have similar distributions, so training data and validation data are alike.

From ANOVA of stepfit, we find that AIC has nearly no decrease when adding hhmember_group to the model. Therefore, we consider two models: one is the selected sub-model from the model selection step; and the other is the selected sub-model without hhmember_group. Then we test their model validation.

For internal validation, we consider using Cp and Pressp to check the validity of a model by using training data. According to Table 2, we find that Cp from these two models is nearly equal to p, and their SSEp are reasonably close to Pressp, which means there is no significant model bias and no severe over-fitting by the model.

	Cp	p	Pressp	SSEp
Model1	5.42319	10	305.8743	287.6405
Model2	5.360875	9	305.7893	289.1657

Table 2: The criterion values of internal validation

For external validation, we consider using MSPEv to check consistency in estimation by using validation data. According to Table 3, we find that MSPEv of these two models is similar, indicating that they have similar predictive ability. In addition, MSPEv is not much larger than MSEp so there is no severe over-fitting by the model.

	MSPEv	Pressp/n	MSEp
Model1	1.086345	0.8049325	0.7774067
Model2	1.104016	0.8047088	0.7794225

Table 3: The criterion values of external validation

Based on the principle of parsimony, which encourages the adoption of simpler models provided they sufficiently explain the observed phenomena for avoiding unnecessary complexity, we choose the selected sub-model without hhmember_group as the final model. Finally, we refit this model using the entire dataset.

III. Conclusions and Discussion

Now we know, HHincomeca_w, District, acess_nonfarm, Farmsize, distancapita, TLU, lnT_Hh_Size, Yearsvillage, would affect the value of production. While our results can be generalized to arid regions in Tanzania as we only consider two typical drylands in Tanzania, other sub-Saharan regions can benefit from the result because they have similar climate conditions.

The local governments could take this information into consideration. For example, given the positive correlation between acess_nonfarm and the response variable, the local government could establish more information platforms for their residents to find additional income opportunities. Another possible government action can be investment in the capital sector, effectively converting unused land into agricultural land.

Coefficients	Intercept	HHincomeca_w	District1	acess_nonfarm1	Farmsize
Estimate	11.44483	0.42935	-0.55694	0.46407	0.27952

Coefficients	distancapita	TLU	lnT_Hh_Size	Yearsvillage
Estimate	-0.15857	0.12658	-0.07142	-0.08349

Table 4: Coefficients of final model

The article mainly focuses on the household level and only includes two representative regions of arid areas in Tanzania. For further investigation, more data on sub-Saharan regions needed to be collected to generalize the result in a broader perspective. Also, along with sustainable development goals, the focus of study can therefore be shifted to other sustainable technologies such as organic farming and drip irrigation that are helpful for arid areas. In addition, areas that suffer from other extreme weather, including floods, can also be considered in the next step.

IV. Appendices

Variable	Type	Definition
adopt_SWC	Qualitative	1= Adopters of soil water conservation technologies, 0= Otherwise
productionvalueca2	Quantitative	Agricultural Value of production
HHincomeca_w	Quantitative	Household income per capita
HDDS	Quantitative	Household dietary diversity score
Ageh	Quantitative	Age of the household head
Sexhh	Qualitative	Sex of the household head
Educationhh	Quantitative	Education of the household head
Farmsize	Quantitative	Farm size in hectares
lnT_Hh_Size	Quantitative	Household size
Any_credit	Qualitative	1= Credit access, 0= Otherwise
TLU	Quantitative	Tropical livestock unit
Total_traders	Quantitative	Number of traders
Yearsvillage	Quantitative	Year in the village
Kinship	Quantitative	Number of friends and relatives
acess_nonfarm	Qualitative	Access to off farm income
climate_infor	Qualitative	Household had access to climatic information
distancapita	Quantitative	Distance to capital market
distance_extension	Quantitative	Distance extension office
dist_districtmarket	Quantitative	Distance to the main market
District	Qualitative	1= Kongwa district, 0= Otherwise
hhmember_group	Qualitative	1= Household is the member to a farmers' organization, 0= Otherwise
Neighbor_adopter	Qualitative	1= Neighbour/friend is an adopter of SWCT, 0= Otherwise

Table 1. Description of all variables

```r
# Load data
SWC <- read.csv("D:/R/STA 206/SWC_data.csv", header = TRUE)
# definite factor variable: adopt_SWC, Any_credit, District, hhmember_group,
Neighbor_adopter
SWC$adopt_SWC <- as.factor(SWC$adopt_SWC)
SWC$Any_credit <- as.factor(SWC$Any_credit)
```

```
SWC$District <- as.factor(SWC$District)
SWC$hhmember_group <- as.factor(SWC$hhmember_group)
SWC$Neighbor_adopter <- as.factor(SWC$Neighbor_adopter)
SWC$Sexhh <- as.factor(SWC$Sexhh)
SWC$acess_nonfarm <- as.factor(SWC$acess_nonfarm)
SWC$climate_infor <- as.factor(SWC$climate_infor)
sapply(SWC,class)
```

```
##                HHID           adopt_SWC  productionvalueca2
HHincomeca_w
##           "integer"            "factor"           "integer"
"integer"
##                HDDS                Ageh               Sexhh
Educationhh
##           "integer"           "integer"            "factor"
"integer"
##            Farmsize         lnT_Hh_Size          Any_credit
TLU
##           "integer"           "integer"            "factor"
"integer"
##       Total_traders        Yearsvillage             Kinship
acess_nonfarm
##           "integer"           "integer"           "integer"
"factor"
##       climate_infor        distancapita  distance_extension
dist_districtmarket
##            "factor"           "integer"           "integer"
"integer"
##            District      hhmember_group    Neighbor_adopter
##            "factor"            "factor"            "factor"
```

```
# find missing value
which(SWC == '')
```

```
## [1]  1065  1159  1160  6285  6379  6380 12665 12759 12760
```

```
# from visual look, row 485, 579, 580 have no value so drop them
SWC <- SWC[c(1:484,486:578),]
# find missing value
which(SWC == '')
```

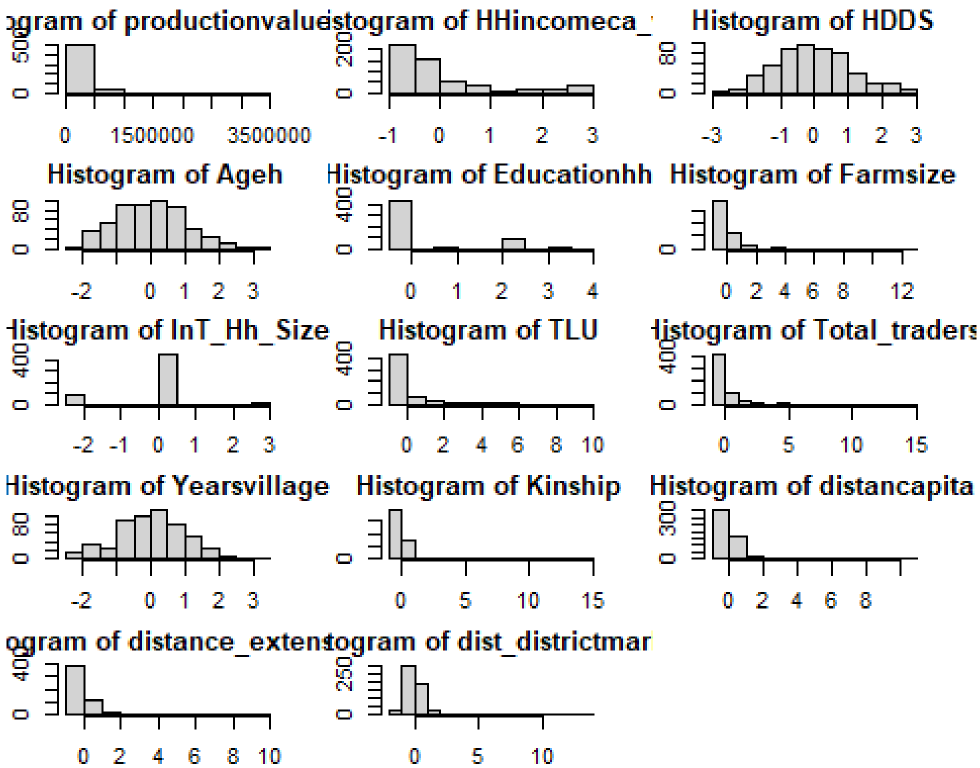
```
## integer(0)
```

```
# drop old class ''
SWC$adopt_SWC <- droplevels(SWC$adopt_SWC)
SWC$Any_credit <- droplevels(SWC$Any_credit)
SWC$District <- droplevels(SWC$District)
SWC$hhmember_group <- droplevels(SWC$hhmember_group)
SWC$Neighbor_adopter <- droplevels(SWC$Neighbor_adopter)
# drop ID
drops <- c("HHID")
```

```r
SWC <- SWC[,!(names(SWC)%in%drops)]
# drop NA
SWC <- na.omit(SWC)
SWC[,c(3:5,7:9,11:14,17:19)] <- scale(SWC[,c(3:5,7:9,11:14,17:19)])

summary(SWC)
```

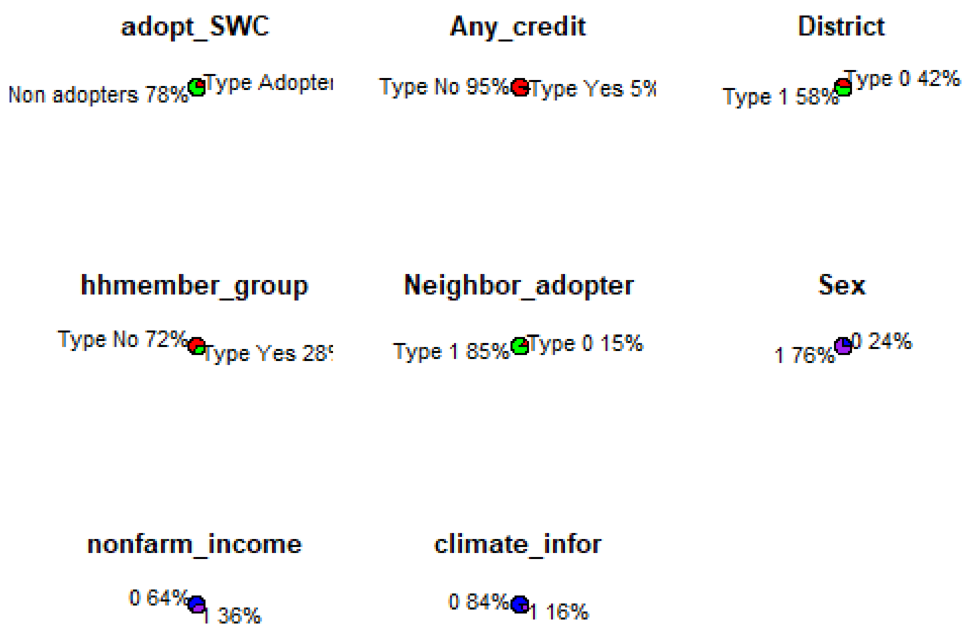
```
##          adopt_SWC   productionvalueca2  HHincomeca_w         HDDS
##  Adopters    :118   Min.   :   1600    Min.   :-0.9176   Min.   :-2.9625
##  Non adopters:426   1st Qu.:  35584    1st Qu.:-0.7033   1st Qu.:-0.6503
##                     Median :  76630    Median :-0.3742   Median :-0.1879
##                     Mean   : 171103    Mean   : 0.0000   Mean   : 0.0000
##                     3rd Qu.: 193608    3rd Qu.: 0.2799   3rd Qu.: 0.7370
##                     Max.   :3325556    Max.   : 2.6071   Max.   : 2.5867
##       Ageh           Sexhh     Educationhh         Farmsize
##  Min.   :-2.18198   0:131   Min.   :-0.4774   Min.   :-0.62098
##  1st Qu.:-0.67678   1:413   1st Qu.:-0.4774   1st Qu.:-0.40485
##  Median :-0.07471           Median :-0.4774   Median :-0.18872
##  Mean   : 0.00000           Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.67789           3rd Qu.:-0.4774   3rd Qu.: 0.02741
##  Max.   : 3.38725           Max.   : 3.7917   Max.   :12.56302
##   lnT_Hh_Size     Any_credit      TLU              Total_traders
##  Min.   :-2.178   No :517    Min.   :-0.40232   Min.   :-0.43936
##  1st Qu.: 0.381   Yes: 27    1st Qu.:-0.40232   1st Qu.:-0.30463
##  Median : 0.381              Median :-0.40232   Median :-0.21481
##  Mean   : 0.000              Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.381              3rd Qu.:-0.03694   3rd Qu.:-0.03517
##  Max.   : 2.940              Max.   : 9.46282   Max.   :14.83019
##   Yearsvillage        Kinship          acess_nonfarm climate_infor
##  Min.   :-2.1220   Min.   :-0.65405   0:347         0:459
##  1st Qu.:-0.7015   1st Qu.:-0.40471   1:197         1: 85
##  Median : 0.0680   Median :-0.15538
##  Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.7191   3rd Qu.: 0.09396
##  Max.   : 3.1459   Max.   :14.30618
##   distancapita      distance_extension dist_districtmarket District
##  Min.   :-0.96667   Min.   :-0.7070    Min.   :-1.1690     0:230
##  1st Qu.:-0.49363   1st Qu.:-0.4151    1st Qu.:-0.3741     1:314
##  Median :-0.01662   Median :-0.2691    Median :-0.1046
##  Mean   : 0.00000   Mean   : 0.0000    Mean   : 0.0000
##  3rd Qu.: 0.22188   3rd Qu.: 0.1250    3rd Qu.: 0.4343
##  Max.   :10.95459   Max.   : 9.8031    Max.   :13.9072
##  hhmember_group Neighbor_adopter
##  No :389        0: 79
##  Yes:155        1:465
##
##
##
##
```

```
# draw histogram for quantitative variables
par(mar = c(2, 2, 2, 2))
par(mfrow = c(5,3))
for(i in c(2:5,7:9,11:14,17:19)){
hist(SWC[, i], main=paste("Histogram of", names(SWC)[i]), xlab =
names(SWC)[i])}
par(mfrow = c(1,1))
```

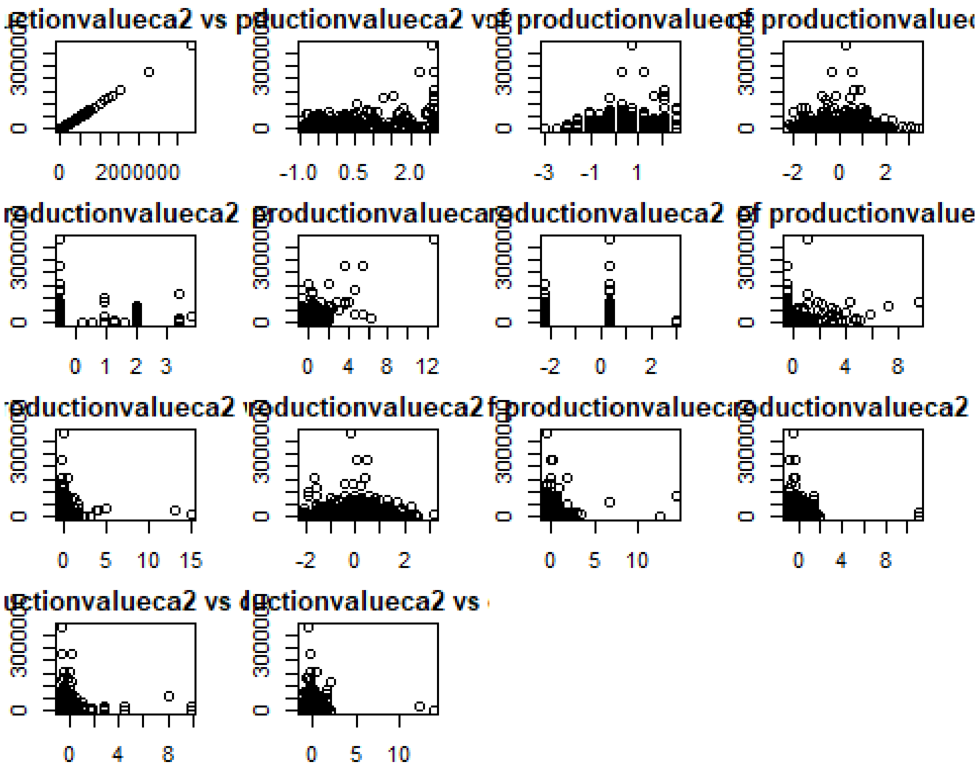



```
# draw pie charts for categorical variables
# define function for adopt_SWC
n <- nrow(SWC)
lbls1 <- c("Adopters","Non adopters")
pct1 <- round(100*table(SWC$adopt_SWC)/n)
lab1 <- paste('Type',lbls1,pct1,sep=' ')
lab1 <- paste(lab1,'%',sep='')
# define function for Any_credit
lbls2 <- c("No","Yes")
pct2 <- round(100*table(SWC$Any_credit)/n)
lab2 <- paste('Type',lbls2,pct2,sep=' ')
lab2 <- paste(lab2,'%',sep='')
# define function for District
lbls3 <- c("0","1")
pct3 <- round(100*table(SWC$District)/n)
lab3 <- paste('Type',lbls3,pct3,sep=' ')
lab3 <- paste(lab3,'%',sep='')
# define function for hhmember_group
```

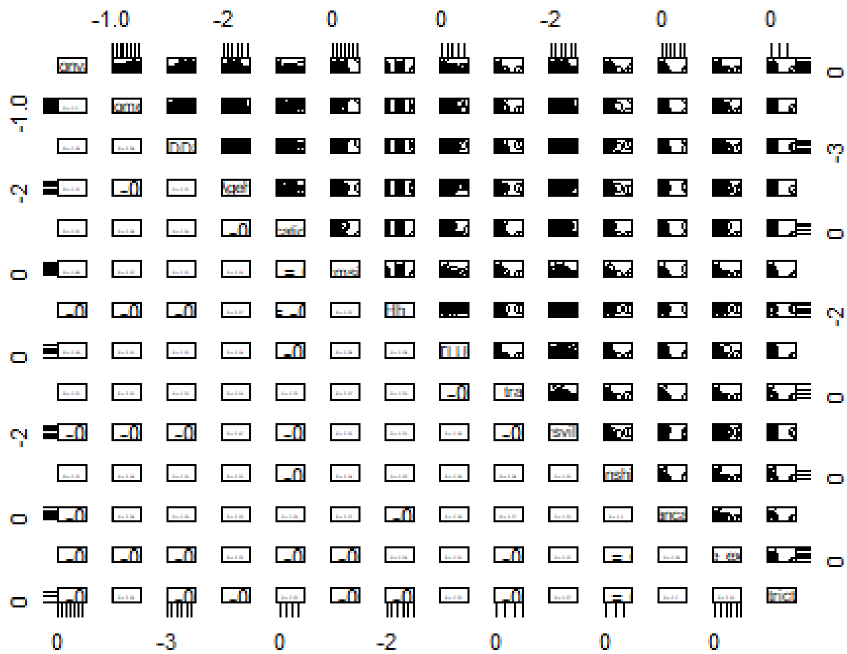
```r
lbls4 <- c("No","Yes")
pct4 <- round(100*table(SWC$hhmember_group)/n)
lab4 <- paste('Type',lbls4,pct4,sep=' ')
lab4 <- paste(lab4,'%',sep='')
# define function for Neighbor_adopter
lbls5 <- c("0","1")
pct5 <- round(100*table(SWC$Neighbor_adopter)/n)
lab5 <- paste('Type',lbls5,pct5,sep=' ')
lab5 <- paste(lab5,'%',sep='')
par(mfrow = c(3,3))
# draw pie chart for adopt_SWC
pie(table(SWC$adopt_SWC),labels=lab1,col=c('red','green'),
    main='adopt_SWC')
# draw pie chart for Any_credit
pie(table(SWC$Any_credit),labels=lab2,col=c('red','green'),
    main='Any_credit')
# draw pie chart for District
pie(table(SWC$District),labels=lab3,col=c('red','green'),
    main='District')
# draw pie chart for hhmember_group
pie(table(SWC$hhmember_group),labels=lab4,col=c('red','green'),
    main='hhmember_group')
# draw pie chart for Neighbor_adopter
pie(table(SWC$Neighbor_adopter),labels=lab5,col=c('red','green'),
    main='Neighbor_adopter')
## pie chart of sex
lbls6 <- c('0','1')
pct6 <- round(100*table(SWC$Sexhh)/n)
lab6 <- paste(lbls6,pct6)
lab6 <- paste(lab6,'%',sep='')
pie(table(SWC$Sexhh),labels=lab6,col=c('blue','purple'),
main='Sex')
## pie chart of access to nonfarm income
lbl7 <- c('0','1')
pct7 <- round(100*table(SWC$acess_nonfarm)/n)
lab7 <- paste(lbl7,pct7)
lab7 <- paste(lab7,'%',sep='')
pie(table(SWC$acess_nonfarm),labels=lab7,col=c('blue','purple'),
main='nonfarm_income')
## pie chart of climate information
lbl8 <- c('0','1')
pct8 <- round(100*table(SWC$climate_infor)/n)
lab8 <- paste(lbl8,pct8)
lab8 <- paste(lab8,'%',sep='')
pie(table(SWC$climate_infor),labels=lab8,col=c('blue','purple'),
main='climate_infor')
par(mfrow = c(1,1))
```



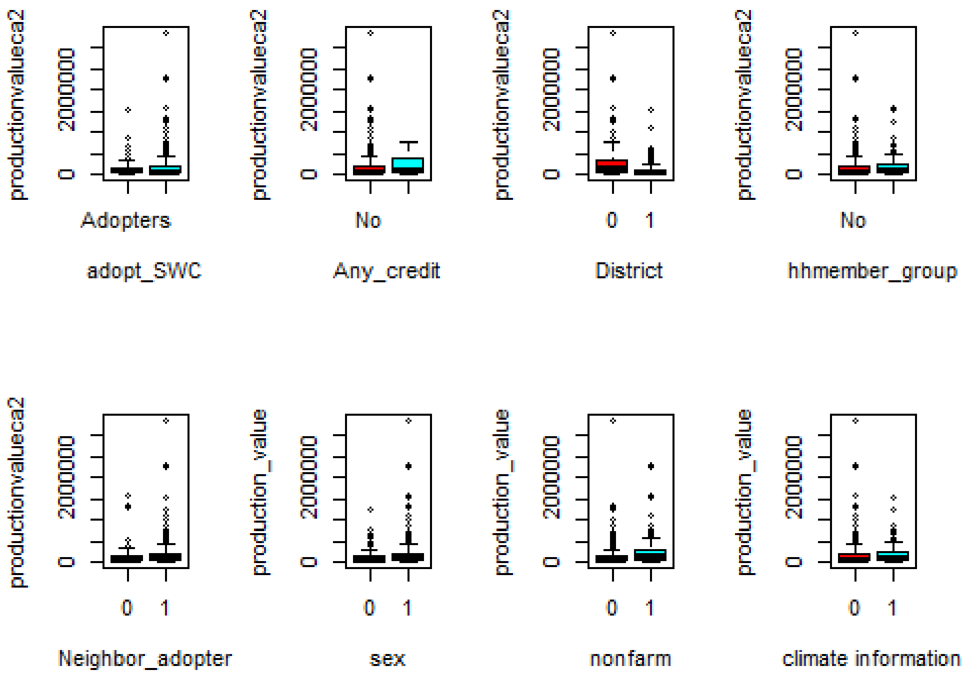

```r
# draw scatter plot between the response variable and the quantitative
predictors
par(mfrow = c(4, 4))
par(mar = c(2, 2, 2, 2))
for(i in c(2:5,7:9,11:14,17:19)){
  plot(SWC[,i],SWC[,2],main = paste("scatter plot of productionvalueca2 vs",
names(SWC)[i]))}
par(mfrow = c(1, 1))
```




```r
# draw scatter plot matrix among quantitative variable with the lower panel
showing correlation coefficients
# define function
panel.cor <- function(x, y){
  #usr <- par("usr")
  #on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y, use="complete.obs"), 2)
  txt <- paste0("R = ", r)
  cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}
pairs(~productionvalueca2+HHincomeca_w+HDDS+Ageh+Educationhh+Farmsize+lnT_Hh_
Size+TLU+Total_traders+Yearsvillage+Kinship+distancapita+distance_extension+d
ist_districtmarket, data=SWC, lower.panel = panel.cor)
```



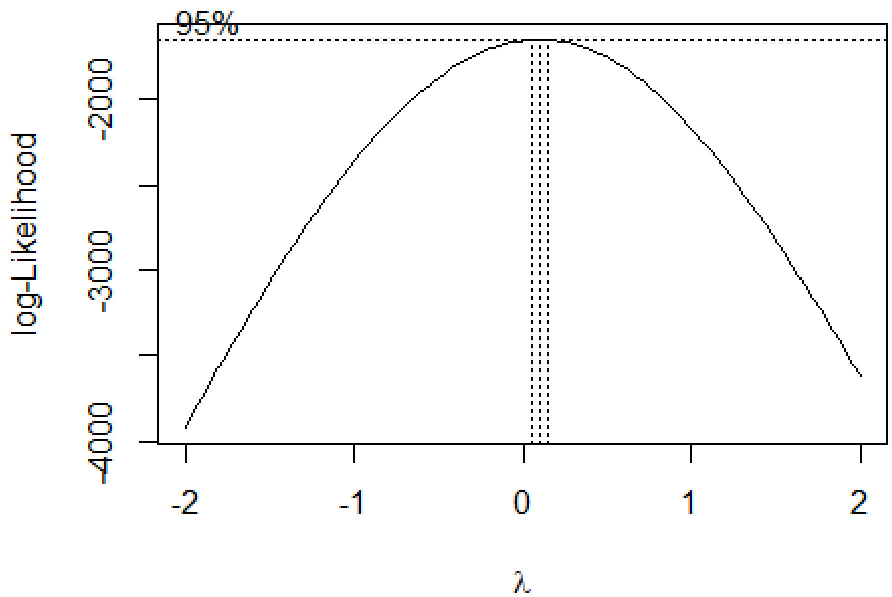
```r
# analyze factor variable
par(mfrow = c(2,4))
boxplot(SWC$productionvalueca2~SWC$adopt_SWC,
xlab='adopt_SWC',ylab='productionvalueca2',col=rainbow(2))
boxplot(SWC$productionvalueca2~SWC$Any_credit,
xlab='Any_credit',ylab='productionvalueca2',col=rainbow(2))
boxplot(SWC$productionvalueca2~SWC$District,
xlab='District',ylab='productionvalueca2',col=rainbow(2))
boxplot(SWC$productionvalueca2~SWC$hhmember_group,
xlab='hhmember_group',ylab='productionvalueca2',col=rainbow(2))
boxplot(SWC$productionvalueca2~SWC$Neighbor_adopter,
xlab='Neighbor_adopter',ylab='productionvalueca2',col=rainbow(2))
## production value versus sex
boxplot(SWC$productionvalueca2~SWC$Sexhh,
xlab='sex',ylab='production_value',col=rainbow(2))
## production value versus access_nonfarm
boxplot(SWC$productionvalueca2~SWC$acess_nonfarm,
xlab='nonfarm',ylab='production_value',col=rainbow(2))
## production value versus climate infor
boxplot(SWC$productionvalueca2~SWC$climate_infor,
xlab='climate information',ylab='production_value',col=rainbow(2))
```



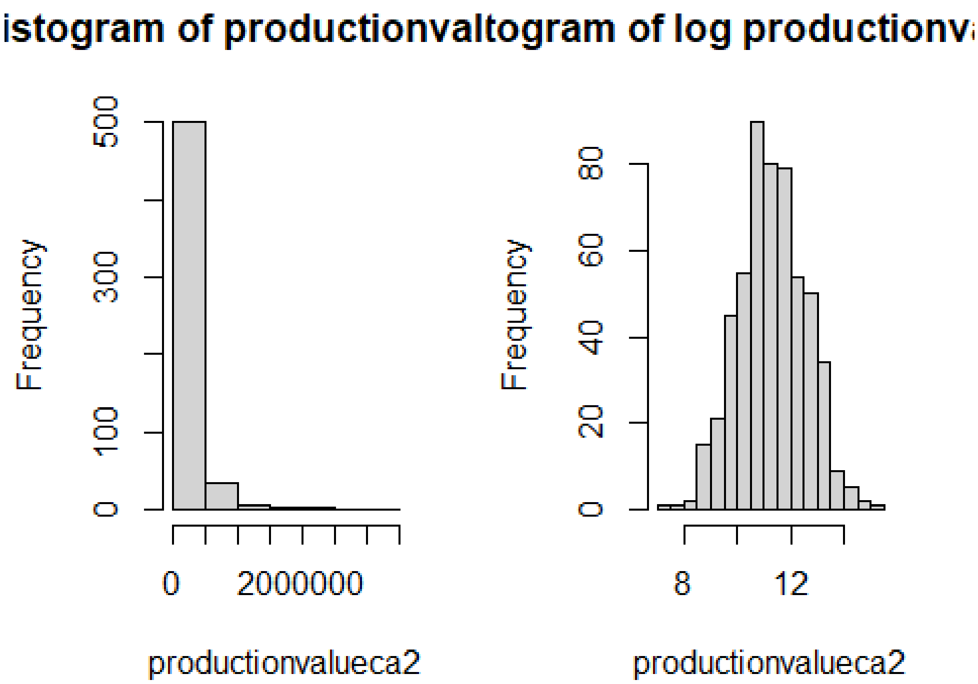

```r
par(mfrow = c(1,1))
```

preliminary fit

```r
library(MASS)
boxcox(lm(productionvalueca2~., data = SWC))
```



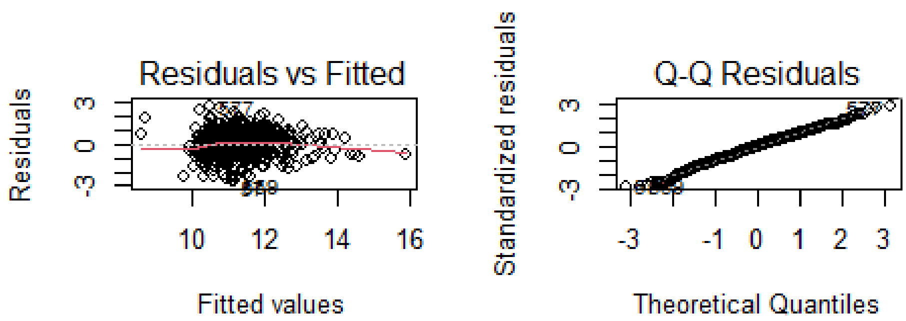

```r
# box-cox procedure suggests logarithm transformation of the response
variable
par(mfrow = c(1,2))
# draw histogram of productionvalueca2
hist(SWC[, 2], main=paste("Histogram of", names(SWC)[2]), xlab =
names(SWC)[2])
# draw histogram for log(productionvalueca2)
transform <- log(SWC[, 2])
hist(transform, main=paste("Histogram of log", names(SWC)[2]), xlab =
names(SWC)[2])
```

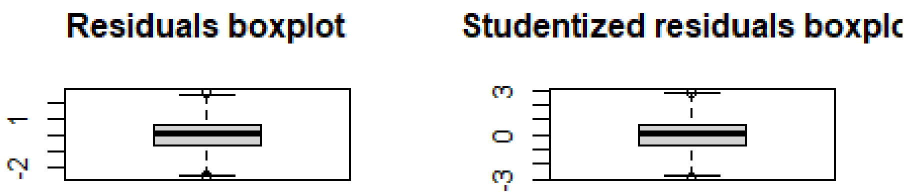



```
par(mfrow = c(1,1))
# log-transform appears to be normal distribution
# replace productionvalueca2 with log-transform then make productionvalueca2
as response variable
SWC$productionvalueca2 <- transform

fit <- lm(productionvalueca2~., data = SWC)
summary(fit)

##
## Call:
## lm(formula = productionvalueca2 ~ ., data = SWC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59684 -0.65329  0.07412  0.61248  2.62558
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             11.39587    0.19329  58.958  < 2e-16 ***
## adopt_SWCNon adopters -0.03263    0.10986  -0.297   0.7665
## HHincomeca_w             0.37125    0.05005   7.418 4.86e-13 ***
## HDDS                     0.10080    0.04585   2.199   0.0283 *
## Ageh                    -0.04521    0.04800  -0.942   0.3467
## Sexhh1                   0.02260    0.10151   0.223   0.8239
## Educationhh             -0.01290    0.04264  -0.302   0.7624
## Farmsize                 0.26224    0.04678   5.605 3.37e-08 ***
```

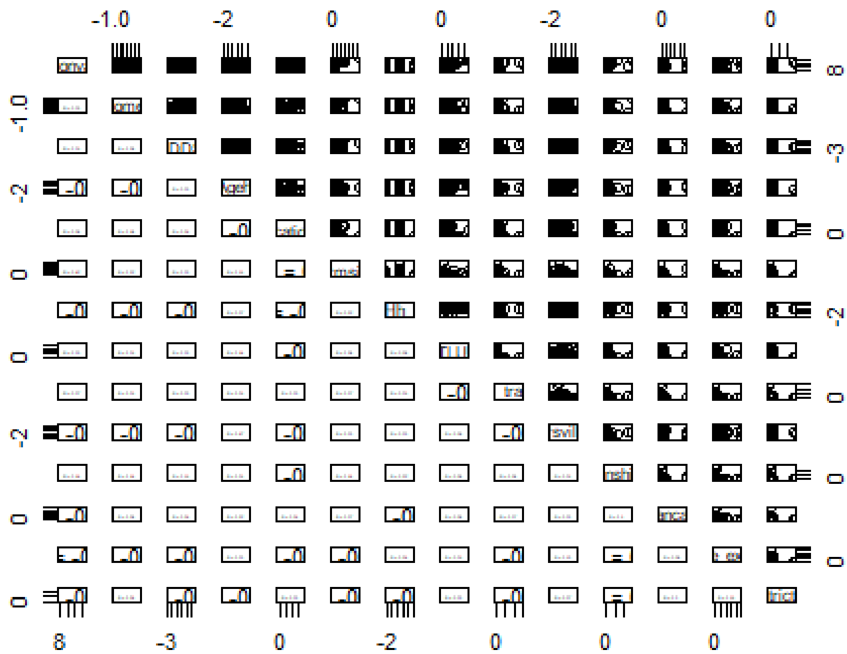
```
## lnT_Hh_Size           -0.09332    0.04483  -2.082   0.0379 *
## Any_creditYes          0.07436    0.19057   0.390   0.6966
## TLU                    0.10763    0.04382   2.456   0.0144 *
## Total_traders          0.04950    0.04157   1.191   0.2343
## Yearsvillage          -0.05685    0.04812  -1.182   0.2379
## Kinship                0.05810    0.04176   1.391   0.1647
## acess_nonfarm1         0.47512    0.08704   5.458 7.44e-08 ***
## climate_infor1         0.03987    0.11967   0.333   0.7392
## distancapita          -0.17464    0.04331  -4.032 6.34e-05 ***
## distance_extension     0.01261    0.04194   0.301   0.7639
## dist_districtmarket    0.01289    0.04288   0.301   0.7638
## District1             -0.59108    0.09350  -6.322 5.55e-10 ***
## hhmember_groupYes      0.15507    0.10097   1.536   0.1252
## Neighbor_adopter1      0.02219    0.11831   0.188   0.8513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.93 on 522 degrees of freedom
## Multiple R-squared:  0.4652, Adjusted R-squared:  0.4437
## F-statistic: 21.62 on 21 and 522 DF,  p-value: < 2.2e-16
```

```r
##residual plot
par(mfrow=c(2,2))
plot(fit,which=1)
plot(fit,which=2)
boxplot(fit$residuals, main="Residuals boxplot")
library(MASS)
boxplot(studres(fit), main="Studentized residuals boxplot")
```



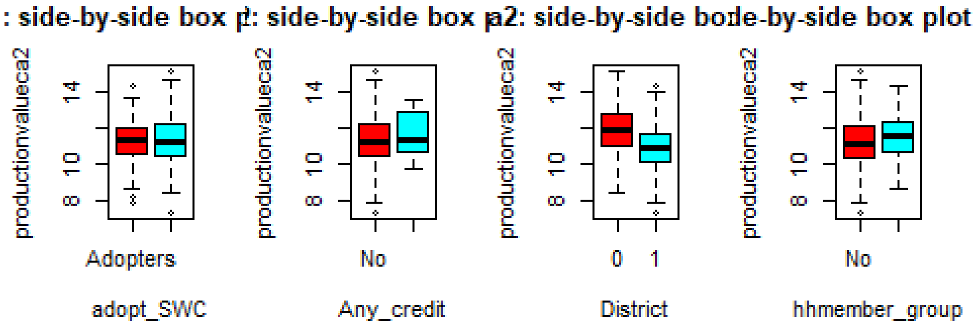


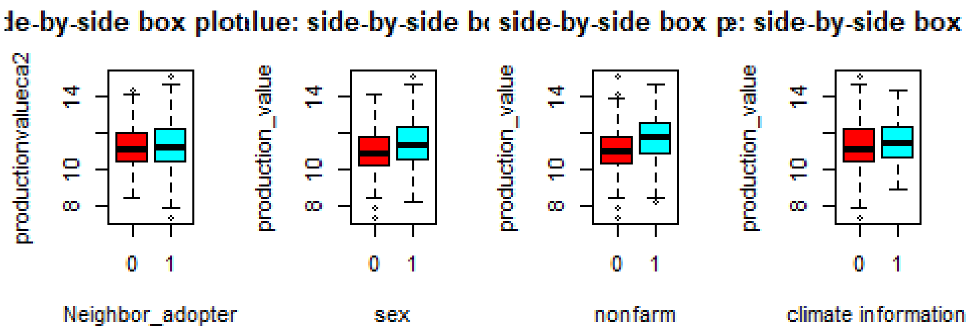

```
# draw scatter plot matrix among quantitative variable with the lower panel
showing correlation coefficients
# define function
panel.cor <- function(x, y){
  #usr <- par("usr")
  #on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y, use="complete.obs"), 2)
  txt <- paste0("R = ", r)
  cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}
pairs(~productionvalueca2+HHincomeca_w+HDDS+Ageh+Educationhh+Farmsize+lnT_Hh_
Size+TLU+Total_traders+Yearsvillage+Kinship+distancapita+distance_extension+d
ist_districtmarket, data=SWC, lower.panel = panel.cor)
```



```
# analyze factor variable
par(mfrow = c(2,4))
boxplot(SWC$productionvalueca2~SWC$adopt_SWC,main='productionvalueca2:
side-by-side box plot by adopt_SWC level',
xlab='adopt_SWC',ylab='productionvalueca2',col=rainbow(2))
boxplot(SWC$productionvalueca2~SWC$Any_credit,main='productionvalueca2:
side-by-side box plot by Any_credit level',
xlab='Any_credit',ylab='productionvalueca2',col=rainbow(2))
boxplot(SWC$productionvalueca2~SWC$District,main='productionvalueca2:
side-by-side box plot by District level',
xlab='District',ylab='productionvalueca2',col=rainbow(2))
boxplot(SWC$productionvalueca2~SWC$hhmember_group,main='productionvalueca2:
side-by-side box plot by hhmember_group level',
xlab='hhmember_group',ylab='productionvalueca2',col=rainbow(2))
boxplot(SWC$productionvalueca2~SWC$Neighbor_adopter,main='productionvalueca2:
side-by-side box plot by Neighbor_adopter level',
xlab='Neighbor_adopter',ylab='productionvalueca2',col=rainbow(2))
## production value versus sex
boxplot(SWC$productionvalueca2~SWC$Sexhh,main='Production value: side-by-side
box plot by gender',
xlab='sex',ylab='production_value',col=rainbow(2))
## production value versus access_nonfarm
boxplot(SWC$productionvalueca2~SWC$acess_nonfarm,main='Production value:
side-by-side box plot by acess_nonfarm',
xlab='nonfarm',ylab='production_value',col=rainbow(2))
## production value versus climate infor
```

```
boxplot(SWC$productionvalueca2~SWC$climate_infor,main='Production value:
side-by-side box plot by climate_infor',
xlab='climate information',ylab='production_value',col=rainbow(2))
```



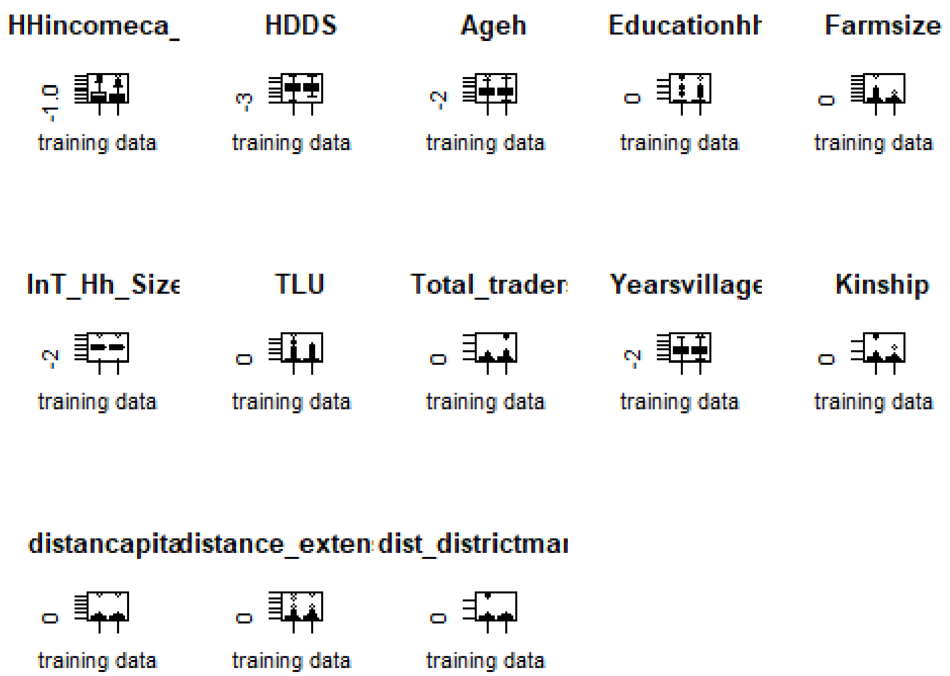



```
par(mfrow = c(1,1))

# determine whether multicollinearity(interaction terms and/or high order
terms needed)
# VIF summary
summary(diag(solve(cor(SWC[,c(2:5,7:9,11:14,17:19)]))))

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.029   1.126   1.187   1.268   1.374   1.677

# because VIF is small enough, so any collinearity is not severe enough to
adversely affect the regression coefficients or their interpretation
# Base on these preliminary fits, we decided to use log(productionvalueca2)
as the response variable; and not include any interaction terms and high
order power terms because of low VIF value
```

model selectioin

```
## split the data
SWC_S = SWC[complete.cases(SWC),]
set.seed(253)
n_s = nrow(SWC_S) # number of cases in SWC_S (366)
index_s = sample(1:n_s, size = n_s*0.7, replace = FALSE)
```

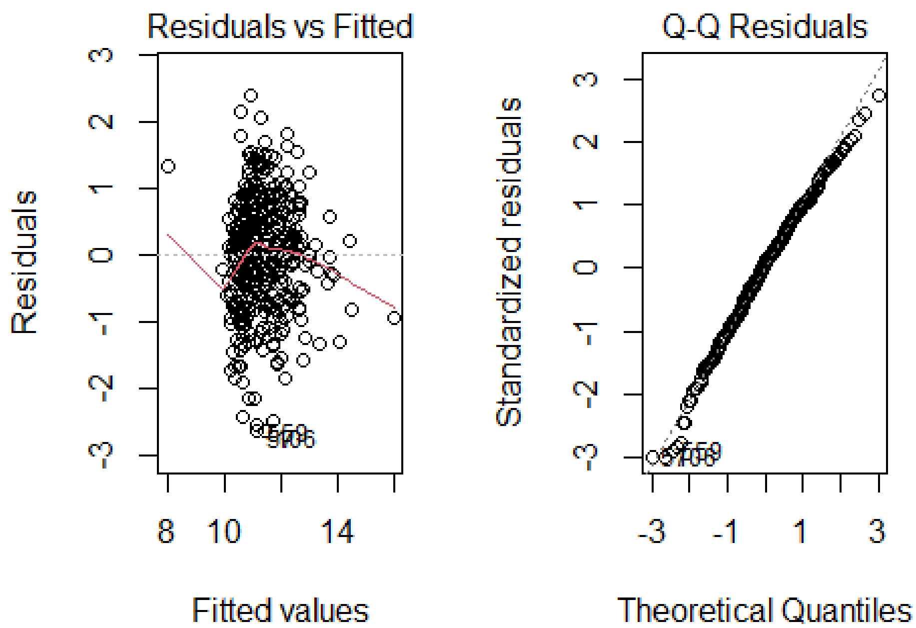
```r
SWC_C  = SWC_S[index_s,] # get the training data set.
SWC_V  = SWC_S[-index_s,] # the remaining 183 cases form the validation set.
n_c <- nrow(SWC_C)

## check if training data and validation data are alike
par(mar = c(5, 4, 4, 2))
par(mfrow = c(3,5))
for (col_name in c('HHincomeca_w', 'HDDS', 'Ageh', 'Educationhh',
'Farmsize','lnT_Hh_Size',  'TLU', 'Total_traders', 'Yearsvillage', 'Kinship',
'distancapita','distance_extension', 'dist_districtmarket')){
  boxplot(SWC_C[, col_name], SWC_V[, col_name], main = col_name, names =
c('training data', 'validation data'))
}
par(mfrow = c(1,1))
```




```r
## we found that they have similar distribution

fit0 = lm(productionvalueca2 ~ 1, data = SWC_C)
fit1 = lm(productionvalueca2 ~ ., data = SWC_C) # fit the training data
library(MASS)
## stepwise
step_f = stepAIC(fit0, scope = list(upper = fit1, lower = ~1), trace = 0,
direction = "both", k = 2)

step_f$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## productionvalueca2 ~ 1
##
## Final Model:
## productionvalueca2 ~ HHincomeca_w + District + acess_nonfarm +
##     Farmsize + distancapita + TLU + lnT_Hh_Size + Yearsvillage +
##     hhmember_group
##
##
##                Step Df   Deviance Resid. Df Resid. Dev       AIC
## 1                                       379   568.3932 155.00406
## 2    + HHincomeca_w  1 159.538763       378   408.8545  31.81145
## 3        + District  1  53.466984       377   355.3875 -19.44576
## 4   + acess_nonfarm  1  22.284792       376   333.1027 -42.05375
## 5        + Farmsize  1  17.985394       375   315.1173 -61.14599
## 6    + distancapita  1  13.900558       374   301.2168 -76.28964
## 7             + TLU  1   4.036365       373   297.1804 -79.41614
## 8     + lnT_Hh_Size  1   4.726657       372   292.4537 -83.50862
## 9    + Yearsvillage  1   3.287997       371   289.1657 -85.80508
## 10 + hhmember_group  1   1.525236       370   287.6405 -85.81474
```

```
## use residual plot to check if the model is adequate
par(mfrow=c(1,2))
plot(step_f, which = 1:2)
```



model validation

```
# Internal validation
MSE_full <- anova(fit1)["Residuals",3]
SSE <- anova(step_f)["Residuals",2]
MSE <- anova(step_f)["Residuals",3]
p <- length(step_f$coefficients)
Cp <- SSE/MSE_full - (n_c - 2*p)
press <- sum(step_f$residuals^2/(1-influence(step_f)$hat)^2)
# External validation
fit_v <- lm(step_f, data = SWC_V) # Model fs1 on validation data
summary(step_f)

##
## Call:
## lm(formula = productionvalueca2 ~ HHincomeca_w + District + acess_nonfarm
+
##     Farmsize + distancapita + TLU + lnT_Hh_Size + Yearsvillage +
##     hhmember_group, data = SWC_C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.61788 -0.60654  0.07506  0.61535  2.38904
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       11.51155    0.08634 133.321  < 2e-16 ***
## HHincomeca_w       0.42373    0.05274   8.034 1.27e-14 ***
## District1         -0.66051    0.09906  -6.668 9.48e-11 ***
## acess_nonfarm1     0.49367    0.09606   5.139 4.47e-07 ***
## Farmsize           0.25031    0.05023   4.984 9.60e-07 ***
## distancapita      -0.20798    0.04715  -4.411 1.35e-05 ***
## TLU                0.12201    0.04862   2.510   0.0125 *
## lnT_Hh_Size       -0.11108    0.04550  -2.441   0.0151 *
## Yearsvillage      -0.10465    0.04730  -2.213   0.0275 *
## hhmember_groupYes  0.14687    0.10485   1.401   0.1621
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8817 on 370 degrees of freedom
## Multiple R-squared:  0.4939, Adjusted R-squared:  0.4816
## F-statistic: 40.13 on 9 and 370 DF,  p-value: < 2.2e-16
```

```
summary(fit_v)
```

```
##
## Call:
## lm(formula = step_f, data = SWC_V)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54894 -0.61903  0.05013  0.63825  2.76932
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       11.07534    0.15908  69.623  < 2e-16 ***
## HHincomeca_w       0.41150    0.09339   4.406 1.96e-05 ***
## District1         -0.34474    0.18455  -1.868 0.063660 .
## acess_nonfarm1     0.40809    0.18164   2.247 0.026084 *
## Farmsize           0.39627    0.10337   3.834 0.000184 ***
## distancapita      -0.05842    0.07799  -0.749 0.454996
## TLU                0.12863    0.08418   1.528 0.128557
## lnT_Hh_Size        0.03089    0.09946   0.311 0.756564
## Yearsvillage      -0.07541    0.08250  -0.914 0.362133
## hhmember_groupYes  0.33919    0.17852   1.900 0.059294 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.01 on 154 degrees of freedom
## Multiple R-squared:   0.42, Adjusted R-squared:  0.3861
## F-statistic: 12.39 on 9 and 154 DF,  p-value: 1.179e-14
```

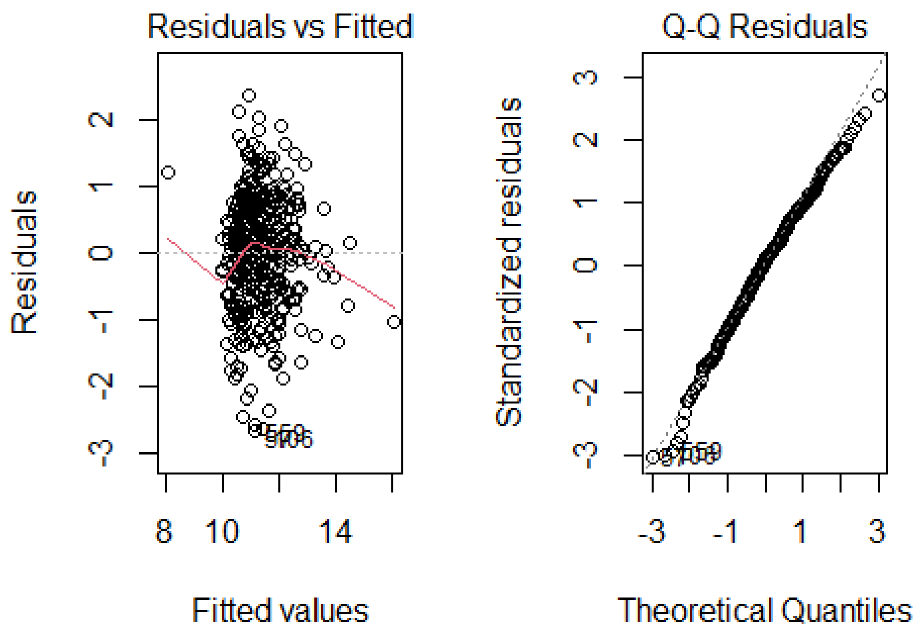
```
# percent change in parameter estimation
round(abs(coef(step_f) - coef(fit_v))/abs(coef(step_f))*100, 3)
```

```
##       (Intercept)      HHincomeca_w         District1    acess_nonfarm1
##             3.789             2.886            47.807            17.336
```

```
##          Farmsize      distancapita               TLU       lnT_Hh_Size
##            58.310            71.911             5.422           127.806
##      Yearsvillage hhmember_groupYes
##            27.940           130.953

sd <- summary(step_f)$coefficients[,"Std. Error"]
sd_v <- summary(fit_v)$coefficients[,"Std. Error"]
# percent change in standard errors
round(abs(sd - sd_v)/sd*100, 3)

##       (Intercept)       HHincomeca_w         District1    acess_nonfarm1
##            84.234            77.074            86.293            89.099
##          Farmsize      distancapita               TLU       lnT_Hh_Size
##           105.806            65.426            73.145           118.594
##      Yearsvillage hhmember_groupYes
##            74.431            70.255

# mean squared prediction error
pred = predict.lm(step_f, SWC_V[,-2])
mspe = mean((pred - SWC_V[,2])^2)

step_f2 <- lm(productionvalueca2 ~ HHincomeca_w + District + acess_nonfarm +
    Farmsize + distancapita + TLU + lnT_Hh_Size + Yearsvillage , data =
SWC_C)
# Internal validation
SSE2 <- anova(step_f2)["Residuals",2]
MSE2 <- anova(step_f2)["Residuals",3]
p2 <- length(step_f2$coefficients)
Cp2 <- SSE2/MSE_full - (n_c - 2*p2)
press2 <- sum(step_f2$residuals^2/(1-influence(step_f2)$hat)^2)
# External validation
fit_v2 <- lm(step_f2, data = SWC_V) # Model fs1 on validation data
summary(step_f2)

##
## Call:
## lm(formula = productionvalueca2 ~ HHincomeca_w + District + acess_nonfarm
+
##     Farmsize + distancapita + TLU + lnT_Hh_Size + Yearsvillage,
##     data = SWC_C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65013 -0.57739  0.07595  0.64119  2.36370
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.54995    0.08198 140.884  < 2e-16 ***
## HHincomeca_w    0.43370    0.05232   8.289 2.12e-15 ***
## District1      -0.65360    0.09907  -6.597 1.44e-10 ***
## acess_nonfarm1  0.48847    0.09611   5.082 5.92e-07 ***
```

```
## Farmsize        0.25180    0.05028   5.008 8.53e-07 ***
## distancapita   -0.20295    0.04707  -4.311 2.08e-05 ***
## TLU             0.12721    0.04854   2.621  0.00913 **
## lnT_Hh_Size    -0.10667    0.04545  -2.347  0.01946 *
## Yearsvillage   -0.09654    0.04700  -2.054  0.04069 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8828 on 371 degrees of freedom
## Multiple R-squared:  0.4913, Adjusted R-squared:  0.4803
## F-statistic: 44.78 on 8 and 371 DF,  p-value: < 2.2e-16

summary(fit_v2)

##
## Call:
## lm(formula = step_f2, data = SWC_V)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70731 -0.67670  0.04158  0.64421  2.66274
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.17576    0.15130  73.867  < 2e-16 ***
## HHincomeca_w    0.42990    0.09366   4.590 9.11e-06 ***
## District1      -0.30258    0.18475  -1.638   0.1035
## acess_nonfarm1  0.34128    0.17970   1.899   0.0594 .
## Farmsize        0.42012    0.10347   4.061 7.75e-05 ***
## distancapita   -0.04205    0.07817  -0.538   0.5914
## TLU             0.13640    0.08479   1.609   0.1097
## lnT_Hh_Size     0.04754    0.09991   0.476   0.6348
## Yearsvillage   -0.06593    0.08304  -0.794   0.4285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 155 degrees of freedom
## Multiple R-squared:  0.4064, Adjusted R-squared:  0.3758
## F-statistic: 13.27 on 8 and 155 DF,  p-value: 1.652e-14

# percent change in parameter estimation
round(abs(coef(step_f2) - coef(fit_v2))/abs(coef(step_f2))*100, 3)

##    (Intercept)   HHincomeca_w      District1 acess_nonfarm1       Farmsize
##          3.240          0.876         53.706         30.132         66.850
##   distancapita            TLU    lnT_Hh_Size   Yearsvillage
##         79.279          7.223        144.572         31.708

sd2 <- summary(step_f2)$coefficients[,"Std. Error"]
sd_v2 <- summary(fit_v2)$coefficients[,"Std. Error"]
```

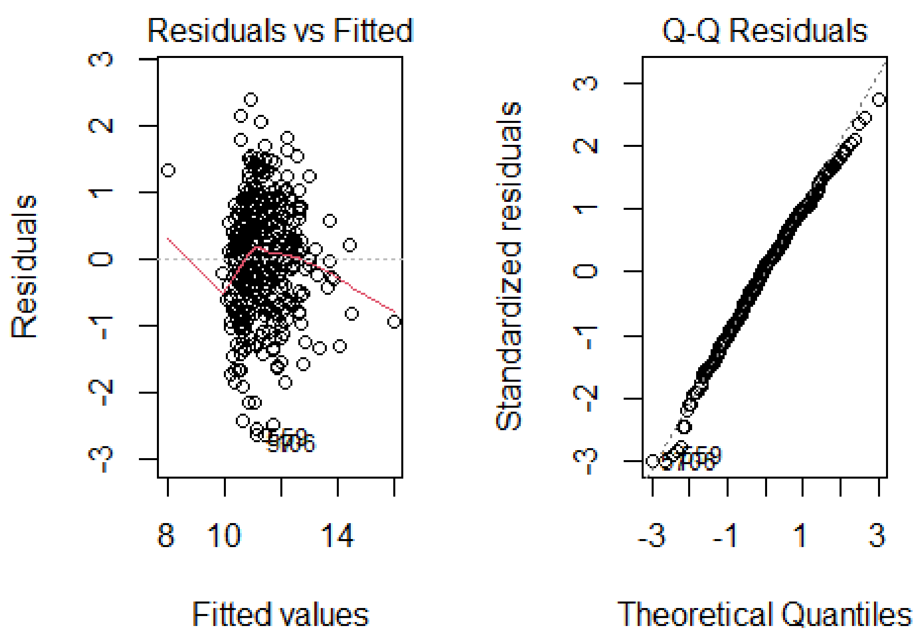
```r
# percent change in standard errors
round(abs(sd2 - sd_v2)/sd2*100, 3)

##    (Intercept)   HHincomeca_w      District1 acess_nonfarm1       Farmsize
##         84.548         79.004         86.482         86.976        105.775
##   distancapita            TLU    lnT_Hh_Size   Yearsvillage
##         66.060         74.674        119.815         76.672

# mean squared prediction error
pred2 = predict.lm(step_f2, SWC_V[,-2])
mspe2 = mean((pred2 - SWC_V[,2])^2)

# validation output
# first model
cat("The selected sub-model from model selection step","\n",
    'Cp:',Cp,'p:',p,"\n",
    'Pressp:',press,'SSEp:',SSE,"\n",
    'mspev:',mspe,'press/n:',press/n_c,'MSE:',MSE,"\n")

## The selected sub-model from model selection step
##  Cp: 5.42319 p: 10
##  Pressp: 305.8743 SSEp: 287.6405
##  mspev: 1.086345 press/n: 0.8049325 MSE: 0.7774067

# second model
cat("The selected sub-model without hhmember_group","\n",
    'Cp2:',Cp2,'p2:',p2,"\n",
    'Pressp2:',press2,'SSEp2:',SSE2,"\n",
    'mspe2v:',mspe2,'press2/n:',press2/n_c,'MSE2:',MSE2,"\n"
    )

## The selected sub-model without hhmember_group
##  Cp2: 5.360875 p2: 9
##  Pressp2: 305.7893 SSEp2: 289.1657
##  mspe2v: 1.104016 press2/n: 0.8047088 MSE2: 0.7794225

## use residual plot to check if the model is adequate
par(mfrow=c(1,2))
plot(step_f2, which = 1:2)
```




```
# refit final model using all data
final_fit <- lm(productionvalueca2 ~ HHincomeca_w + District + acess_nonfarm
+
    Farmsize + distancapita + TLU + lnT_Hh_Size + Yearsvillage , data = SWC)
summary(final_fit)

##
## Call:
## lm(formula = productionvalueca2 ~ HHincomeca_w + District + acess_nonfarm
+
##     Farmsize + distancapita + TLU + lnT_Hh_Size + Yearsvillage,
##     data = SWC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74237 -0.61378  0.09356  0.63272  2.55709
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.44483    0.07317 156.425  < 2e-16 ***
## HHincomeca_w    0.42935    0.04615   9.304  < 2e-16 ***
## District1      -0.55694    0.08845  -6.296 6.36e-10 ***
## acess_nonfarm1  0.46407    0.08542   5.433 8.44e-08 ***
## Farmsize        0.27952    0.04588   6.092 2.13e-09 ***
## distancapita   -0.15857    0.04037  -3.928 9.68e-05 ***
## TLU             0.12658    0.04249   2.979  0.00302 **
```

```
## lnT_Hh_Size    -0.07142    0.04229  -1.689  0.09186 .
## Yearsvillage   -0.08349    0.04114  -2.029  0.04291 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9323 on 535 degrees of freedom
## Multiple R-squared:  0.4491, Adjusted R-squared:  0.4409
## F-statistic: 54.52 on 8 and 535 DF,  p-value: < 2.2e-16
```

```
## draw residual plots
par(mfrow=c(1,2))
plot(step_f, which = 1:2)
```



V. Reference

1. Bjornlund, V., Bjornlund, H., & Van Rooyen, A. F. (2020). *Why agricultural production in sub-Saharan Africa remains low compared to the rest of the world – a historical perspective*. International Journal of Water Resources Development, 36(sup1), S20–S53. https://doi.org/10.1080/07900627.2020.1739512
2. Jayne, T. S., & Sanchez, P. A. (2021). *Agricultural productivity must improve in sub-Saharan Africa*. Science, 372(6546), 1045–1047. https://doi.org/10.1126/science.abf5413
3. Justin, Urassa. (2015). *Factors influencing maize crop production at household levels: A case of Rukwa Region in the southern highlands of Tanzania*. African Journal of

Agricultural Research.
https://www.researchgate.net/publication/283870463_Factors_influencing_maize_crop_production_at_household_levels_A_case_of_Rukwa_Region_in_the_southern_highlands_of_Tanzania

4. Mariamu, Abdallah. (2020). *Exploring factors affecting agricultural productivity in Tanzania: Policy implication for climate change* (By Professor Cho, Yoon Cheong & Professor Choi, Changyong). https://archives.kdischool.ac.kr/bitstream/11125/41440/1/Exploring%20factors%20affecting%20agricultural%20productivity%20in%20Tanzania.pdf
5. Kutner, Nachtsheim, Neter, and Li, (2005). *Applied Linear Statistical Models*. Boston : McGraw-Hill Irwin.
6. Khuble, K. (2024, February 7). Essay on organic farming. EDUCBA. https://www.educba.com/essay-on-organic-farming/
7. Wageningen University and Research. I. F. P. R.I , & I.I.T.A. (2021). Combined GIS and Africa RISING Baseline Evaluation Survey (ARBES) Farm Household data for FarmMatch, Tanzania [Dataset]. In Harvard Dataverse. https://doi.org/10.7910/dvn/z8wxqm