**Influencing Factors of Ghana Agricultural Profits Analysis**

**By Group 07:**

**Siqi Zhang**

**Huibo Jia**

**Qiyu Ye**

**Mark Russeff**

ECON5100 Case Study -Influencing Factors of Ghana Agricultural Profits Analysis

**Abstract**

This paper attempts to explore what determines agricultural profits in Ghana. The data for the study was sourced from the fourth round of the Ghana Living Standard Survey (GLSS 4). Six models were estimated to determine the effect of household member information (sex, education, age), characteristics of local area, household livestock, household equipment, household crop, and community market and transport on agricultural profits. Our analysis reveals that forest ecological zone has more profit, public market has a negative effect on profit, and different type of root crops could influence profit in reverse way.

**Keywords:** Agricultural Profit, Education, Descriptive Analysis, Regression, Diagnostic

**1. Introduction**

Agriculture industry, as one of the major sectors of Ghanaian economy, employs majority of the populace. As it stands, about 2.7 million households in Ghana own or operate a farm or keep livestock. More than half of households, which cultivate crops hire labor for their operations. The major crops, in terms of sales, are cocoa, maize, groundnuts/peanuts, and rice. About 2 and a half million households process crops or fish for sale, with the major responsibility for this activity falling on women. (Ghana Living Standards Survey, Report of The Fourth Round (GLSS 4) ,2000). The main purpose of this project is to analyze what determines agricultural profits in Ghana to help the ACME corporation get an idea of whom to target for their sales efforts because they are planning moving into agricultural inputs in Ghana. For example, household educational attainment, the characteristics of the local area and so on.

The rest of the paper is organized as follows: the second section talks about variables selection based on documentation of GLSS4; the methodology and hypothesis are explained in section three; section four presents the results and discussion and this is followed by conclusions.

**2. Literature Review**

**2.1 Basic information of agriculture in Ghana**

Based on the review of Ghana Living Standards Survey Report of The Fourth Round(GLSS4)，which was conducted between April 1998 and March 1999, we noted that agriculture is still the major industry in Ghana and most of the households make their lives relying on agriculture. This report identifies critical work from the Ghana Living Standard Survey to aid in defining sustainable agricultural productivity and shows the main results of the Fourth Round of the Ghana Living Standards Survey (GLSS 4) with labor force module. This survey covered a period of 12 months (April 1998 to March 1999) and about seven years after the GLSS 3 conducted between 1991 and 1992. The methodology of this survey was based on that of the third round with some minimal modifications. Demographic characteristics of the population, education, health, employment and time use, migration, housing conditions, household agriculture and non-farm businesses are the major topics included in this survey.

**2.2 Datasets and variables selection**

After reading all the documentation related to GLSS4, especially the The Estimation of Components of Households Incomes And Expenditures (Aggregate), Ghana Living Standards Survey Report of The Fourth Round(GLSS4), and Data User’s Guide (G4USERSG), and from our understanding of the definitions of all the variables, we decided to use AGRI1 as our dependent variable(Agricultural Profits). Then we extracted variables as determinants of agricultural profits from household basic information, household member information, household member education information, household livestock count and types, household agricultural equipment type, household havested crop count and types, household havested root count and types, as well as characteristics of local area, including community, transport, market information.

**3. Methodology**

**3.1 Data Type and Source**

Data for the study was sourced from the Ghana Living Standard Survey (GLSS 4) collected between April 1998 and March 1999. It took household as a key social and economic unit, provides valuable insights into living conditions in Ghana. The GLSS 4 is a nation-wide survey which collected detailed information on a variety of topics, including demographic characteristics of the population, education, health, employment and time use, migration, housing conditions, household agriculture and non-farm businesses. A representative nationwide sample of more than 5,998 households, containing over 25,000 persons, was covered in GLSS 4. Detailed information was collected on all aspects of living conditions, including health, education, employment, housing, agricultural activities, the operation of non-farm establishments, remittances, savings, and credit and assets. It also contains information on farm levels, household level characteristics and socio- demographic characteristics.

**3.2 Analysis Steps and Explanations**

After reading through the documentation, we determined a few data files that contain education, characteristics of local area, household livestock/crop/equipment, and community market are the ones we want to use to analyze the influencing factors of Ghana agricultural profits. Then we tidy and aggregate all data onto household level and fit linear regression model over different combination of variables.

Here is a detailed analysis steps and explanations with code, output and graphs written in R notebook: [Analysis\_Steps\_and\_Explanations](Analysis_Steps_and_Explanations.pdf)

**3.3 Models and Hypothesis**

In order to achieve the objectives of this paper, we used multiple regression method to build models. Six different models were estimated. The first is an unrestricted model for all features we found. The other five models are restricted models for different combinations of variables separately.

**Model 1: unrestricted model UR**

*profit ~ reslan + ez + loc2 + loc5 + loc3 + female + age + avgAge +*

*maxAge + minAge + educ + market + transport + livstcdTypeCount +*

*livstcd1 + livstcd2 + livstcd3 + livstcd4 + livstcd5 + livstcd6 +*

*livstcd7 + livstcd8 + livstcd9 + livstcd10 + livstcd11 +*

*livstcd12 + equipTypeCount + eqcdown21 + eqcdown22 + eqcdown31 +*

*eqcdown51 + eqcdown61 + eqcdown62 + eqcdown63 + eqcdown64 +*

*eqcdown65 + cropTypeCount + cropcd0 + cropcd1 + cropcd2 +*

*cropcd3 + cropcd4 + cropcd5 + cropcd6 + cropcd8 + cropcd9 +*

*cropcd10 + cropcd11 + cropcd12 + cropcd13 + cropcd14 + cropcd15 +*

*cropcd16 + cropcd17 + cropcd18 + cropcd19 + cropcd20 + cropcd21 +*

*cropcd22 + cropcd23 + cropcd24 + cropcd25 + cropcd26 + cropcd27 +*

*cropcd28 + cropcd29 + cropcd31 + cropcd32 + cropcd33 + cropcd34 +*

*cropcd35 + rootTypeCount + rootcd0 + rootcd5 + rootcd6 +*

*rootcd7 + rootcd8 + rootcd9 + rootcd11 + rootcd14 + rootcd16 +*

*rootcd18 + rootcd19 + rootcd20 + rootcd21 + rootcd22 + rootcd25 +*

*rootcd26 + rootcd27 + rootcd29 + rootcd30 + rootcd31 + rootcd33 +*

*rootcd34 + rootcd35 + rootcd36*

**Hypothesis:**

*reslanAkan =reslanEwe=reslanGaAdangbe =reslanDagbani =reslanHausa reslanOther=reslanUnknown=ezForest=ezSavannah =loc2Rural= loc5RuralCoastal =loc5RuralForest=loc5RuralSavannah =loc3Rural =femaleTRUE =age=avgAge =maxAge=minAge= educBasicEducation =educSecondaryEducation =educTertiaryEducation =educOther =marketTRUE =transportTRUE =livstcdTypeCount =livstcd1=livstcd2=livstcd3=livstcd4 =livstcd5=livstcd6=livstcd7=livstcd8=livstcd9=livstcd10=livstcd11=livstcd12 =equipTypeCount=eqcdown21=eqcdown22 =eqcdown31=eqcdown51=eqcdown61 =eqcdown62=eqcdown63=eqcdown64=eqcdown65=cropTypeCount=cropcd0 =cropcd1=cropcd2=cropcd3=cropcd4=cropcd5=cropcd6=cropcd8=cropcd9 =cropcd10=cropcd11=cropcd12=cropcd13=cropcd14=cropcd15=cropcd16= cropcd17=cropcd18=cropcd19=cropcd20=cropcd21=cropcd22=cropcd23= cropcd24=cropcd25=cropcd26=cropcd27=cropcd28=cropcd29=cropcd31= cropcd32=cropcd33=cropcd34=cropcd35=rootTypeCount =rootcd0=rootcd5= rootcd6=rootcd7=rootcd8=rootcd9=rootcd11=rootcd14=rootcd16= rootcd18= rootcd19=rootcd20=rootcd21=rootcd22=rootcd25=rootcd26=rootcd27=rootcd29 =rootcd30=rootcd31=rootcd33=rootcd34=rootcd35 = rootcd36 = 0*

**Model 2 : restricted model TOP with top features from agricultural characteristics information**

*profit ~ cropcd25 + rootcd18 + rootcd20 + rootTypeCount + eqcdown61 +*

*rootcd36 + livstcd5 + equipTypeCount + rootcd6 + rootcd7 +*

*livstcd10 + rootcd30 + rootcd5 + rootcd8 + livstcd2*

**Hypothesis:**

*cropcd25 = rootcd18 = rootcd20 = rootTypeCount = eqcdown61 = rootcd36 = livstcd5 = equipTypeCount = rootcd6 = rootcd7 = livstcd10 = rootcd30 = rootcd5 = rootcd8 = livstcd2 = 0*

**Model 3 : restricted model R1 without correlated variables**

*profit ~ reslan + ez + age + market + livstcd5 + livstcd6 + livstcd7 +*

*livstcd10 + equipTypeCount + eqcdown61 + cropcd5 + cropcd8 +*

*cropcd11 + cropcd25 + cropcd29 + rootcd7 + rootcd18 + rootcd20 +*

*rootcd27 + rootcd33 + rootcd36*

**Hypothesis:**

*reslanAkan = reslanEwe = reslanGaAdangbe = reslanDagbani = reslanHausa = reslanOther = reslanUnknown = ezForest = ezSavannah = educBasicEducation = educSecondaryEducation = educTertiaryEducation = educOther = marketTRUE = livstcd5 = livstcd6 = equipTypeCount = cropTypeCount = cropcd8 = cropcd11 = cropcd25 = rootcd8 = rootcd18 = rootcd20 = rootcd27 = rootcd33 = 0*

**Model 4 : restricted model R2 with only education and local characteristic variables**

*profit ~ educ + ez + loc2 + loc5 + loc3 + market + transport*

**Hypothesis:**

*educBasicEducation = educSecondaryEducation = educTertiaryEducation = educOther = ezForest = ezSavannah = loc2Rural = loc5RuralCoastal = loc5RuralForest = loc5RuralSavannah = loc3Rural = marketTRUE = transportTRUE = 0*

**Model 5 : restricted model R3 removing loc5 and loc3 from R2**

*profit ~ educ + ez + loc2 + market + transport*

**Hypothesis:**

*educBasicEducation = educSecondaryEducation = educTertiaryEducation = educOther = ezForest = ezSavannah = loc2Rural = marketTRUE = transportTRUE = 0*

**Model 6 : restricted model R4 with education \* age**

*profit ~ educ \* age + female + ez + loc2 + market + transport*

**Hypothesis:**

*educBasicEducation = educSecondaryEducation = educTertiaryEducation = educOther = age = femaleTRUE = ezForest = ezSavannah = loc2Rural = marketTRUE=transportTRUE=educBasicEducation:age=educSecondaryEducation:age = educTertiaryEducation:age = educOther:age = 0*

**3.4 Regression Analysis**

The F-statistics and regression diagnostics are used to determine whether the separate models or unrestricted model fits the data.

**4. Results and discussion**

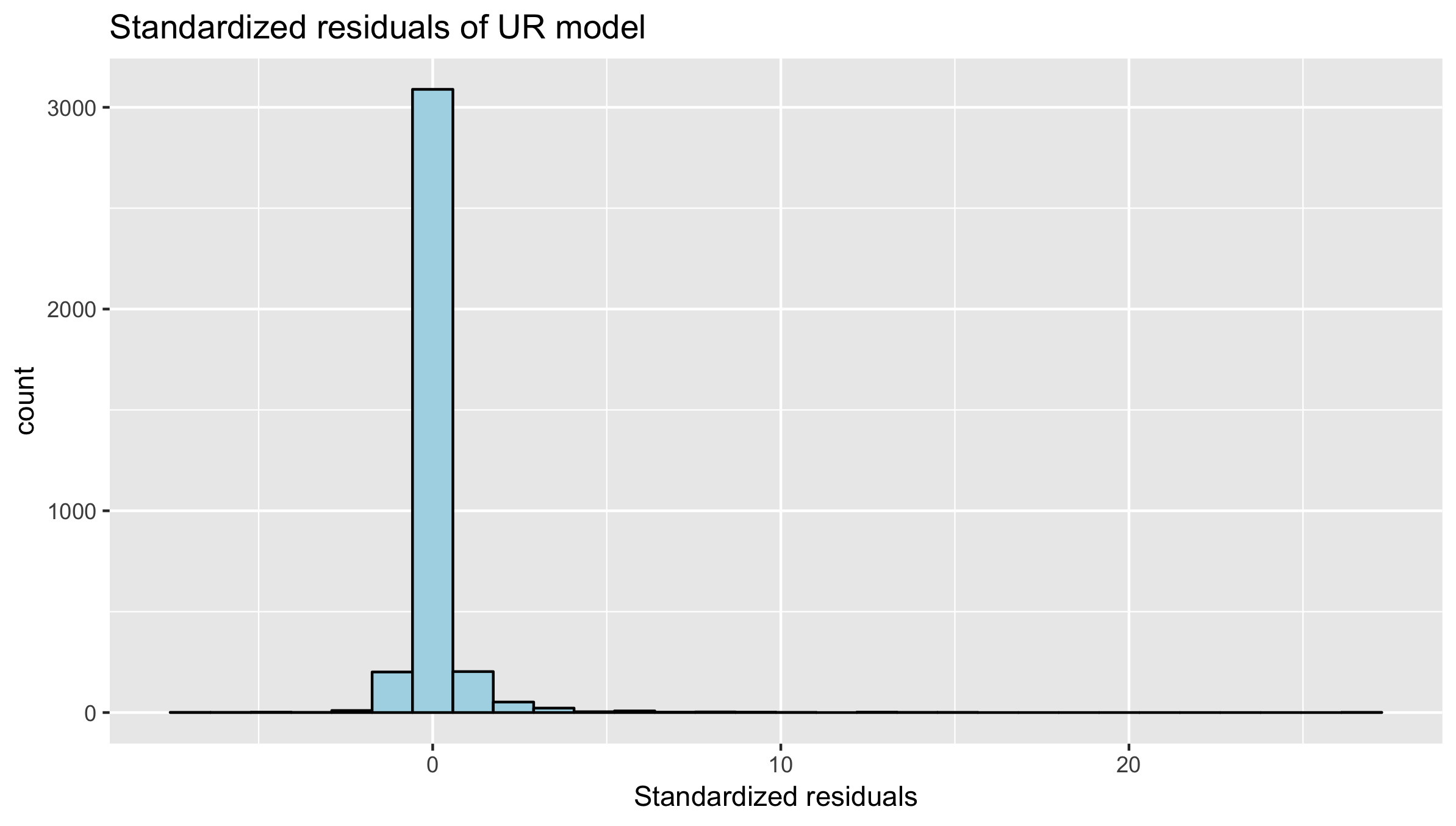
**4.1 Descriptive Statistics**

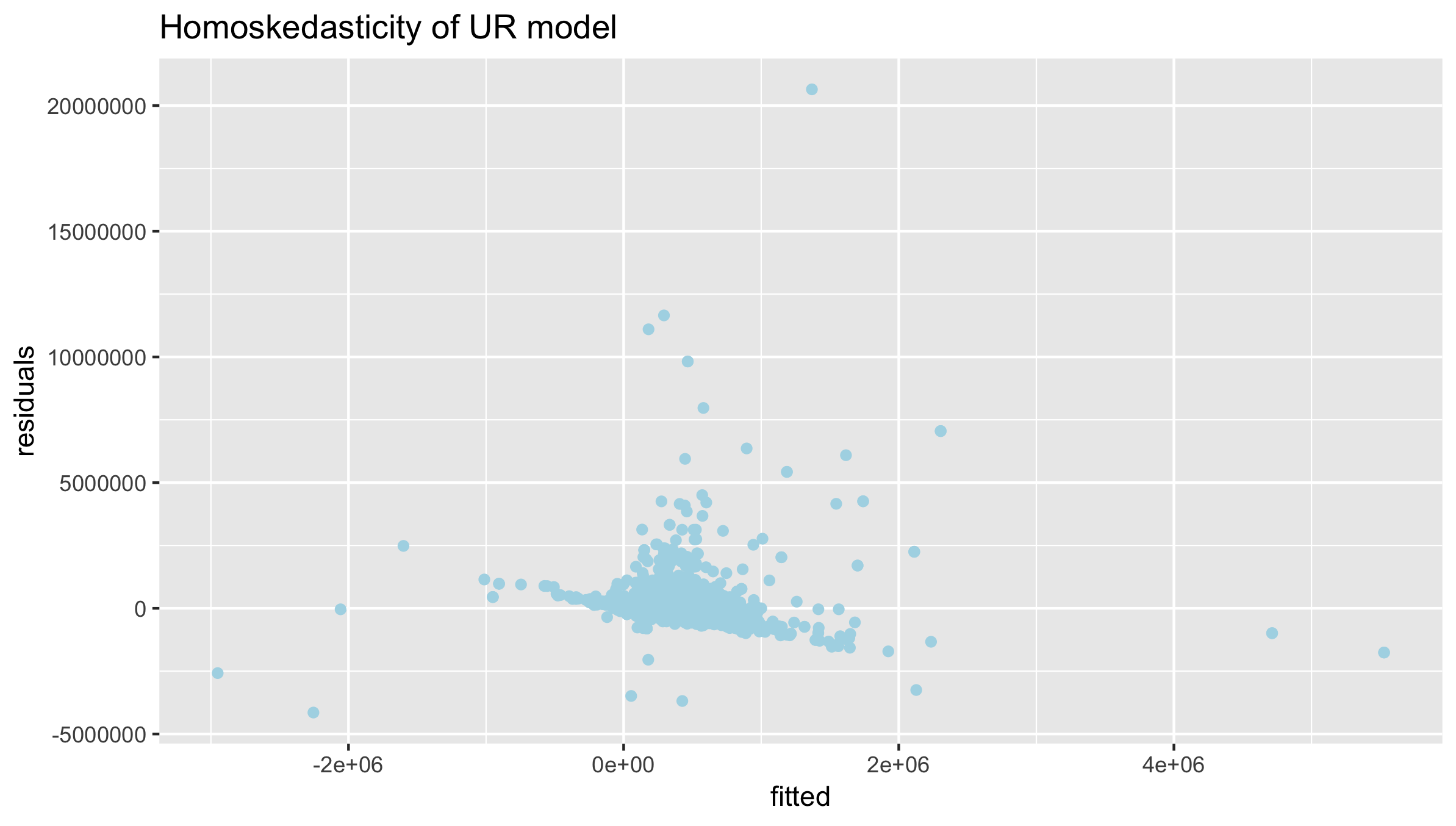
In general, the average household agriculture profit in Ghana is about 314881.5 cedi per acre, when disaggregated, the averages are 314,737.2 and 324,043 per acre for rural and urban households respectively. Also, about 84% of household agriculture profit per acre is generated by household living in rural area. These values have very high standard deviation indicating the presence of outliers.

From our unrestricted model i.e. hh\_profit\_model\_ur, we can see that livstcd5, eqcdown61, rootcd18, rootcd20, rootcd27, rootce36 are statistically significant, which means households that own pigs as livestock, own outboard motor as agriculture equipment, or grow root crops such as cassava, cocoyam, eggplant and pawpaw can affect agriculture profit significantly. It should be noted that eqcdown61 and rootcde27 both have negative effect on our training target, which means household that own outboard motor or grow eggplant as root crops can generate decrease in household profit per acre. When disaggregated, the unrestricted model for rural area shows that same factors as in unrestricted model that affect rural household agriculture profit significantly, however, in unrestricted model for urban area, there are no variables showing strong statistically significance.

**4.1 Regression diagnostics:**

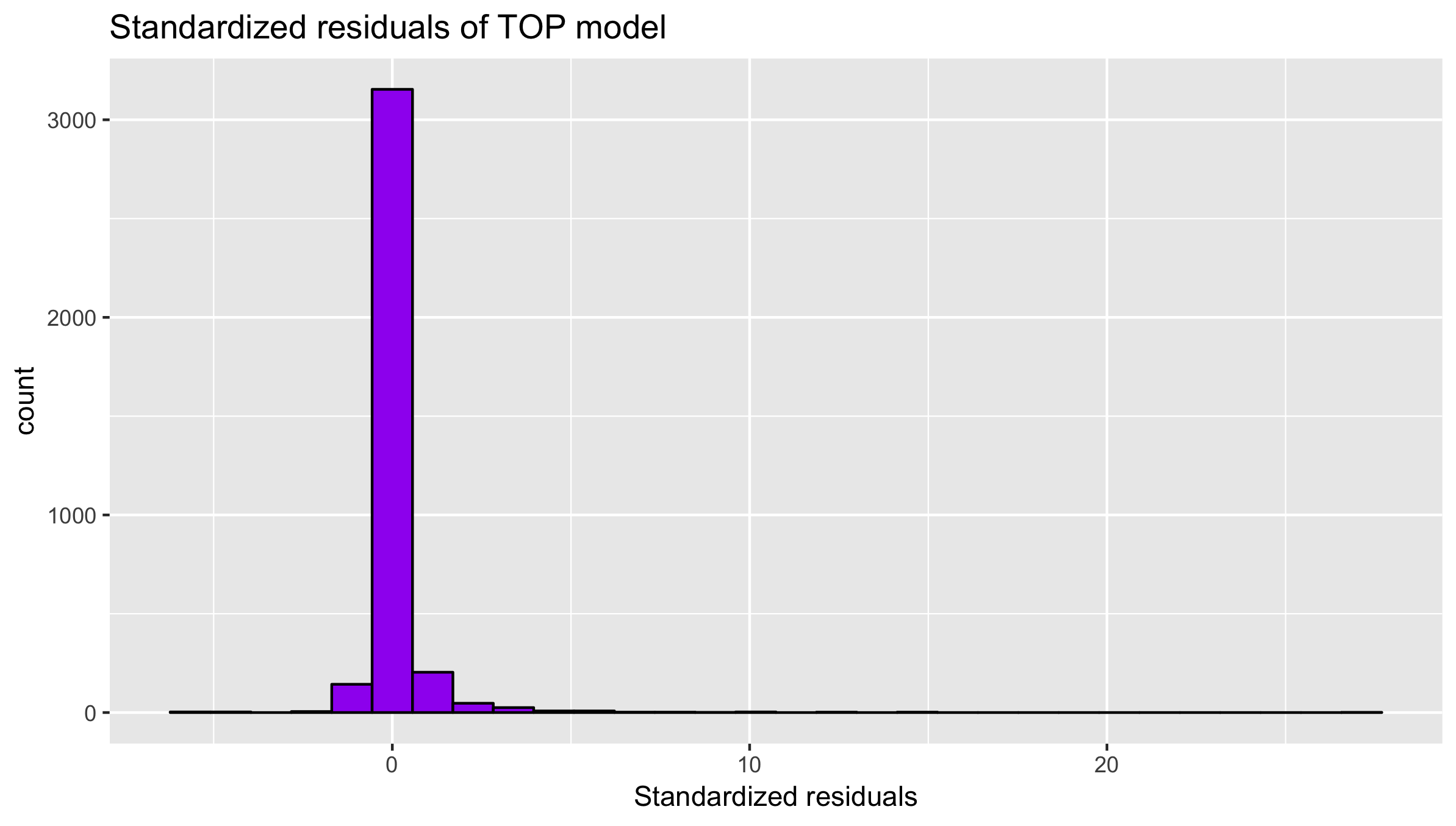
**Model 1 UR:**

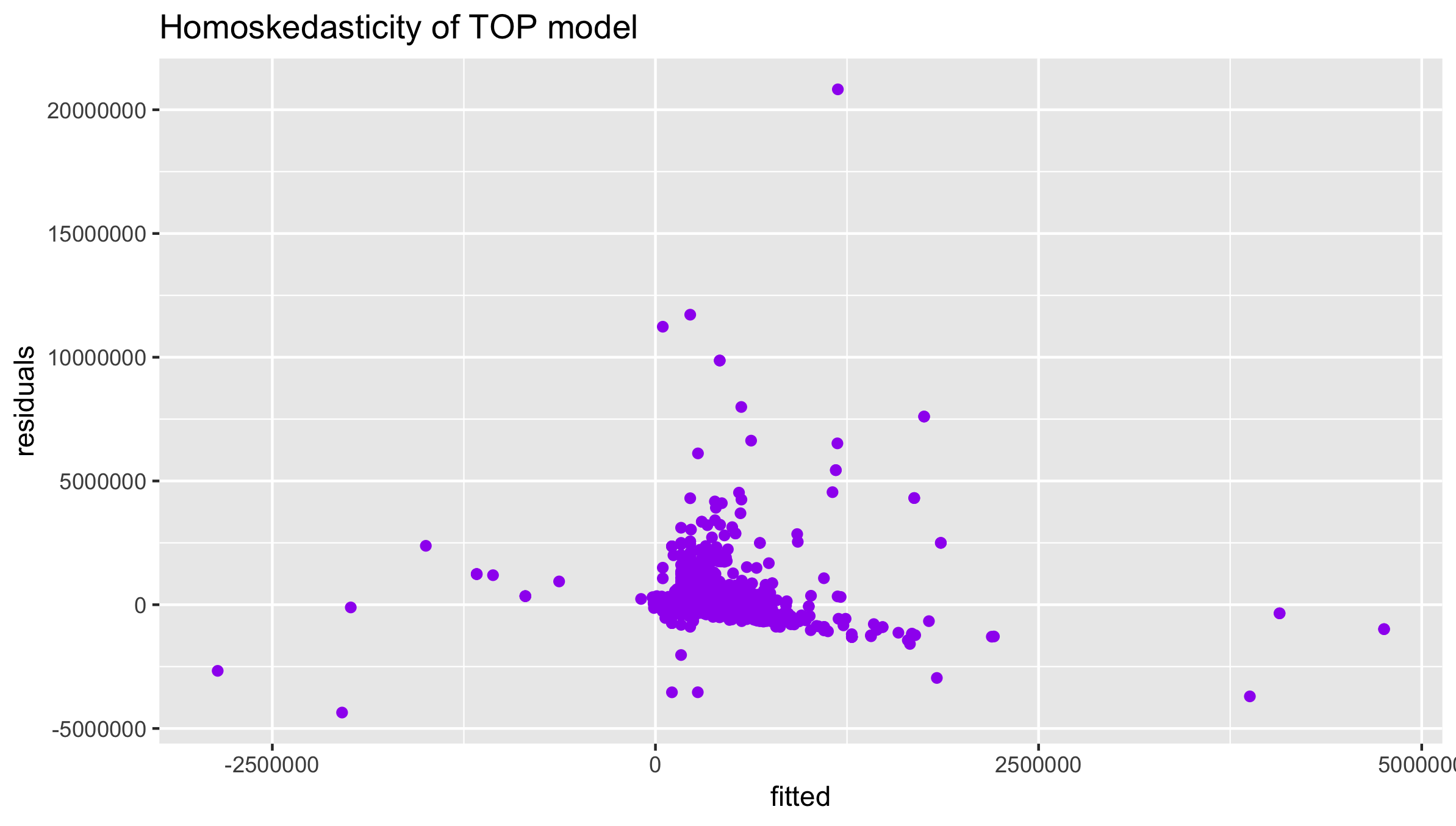
****

****

The plots illustrate an approximately normal distribution of residuals produced by Model 1, which is symmetric, hence the Model 1 is fit.

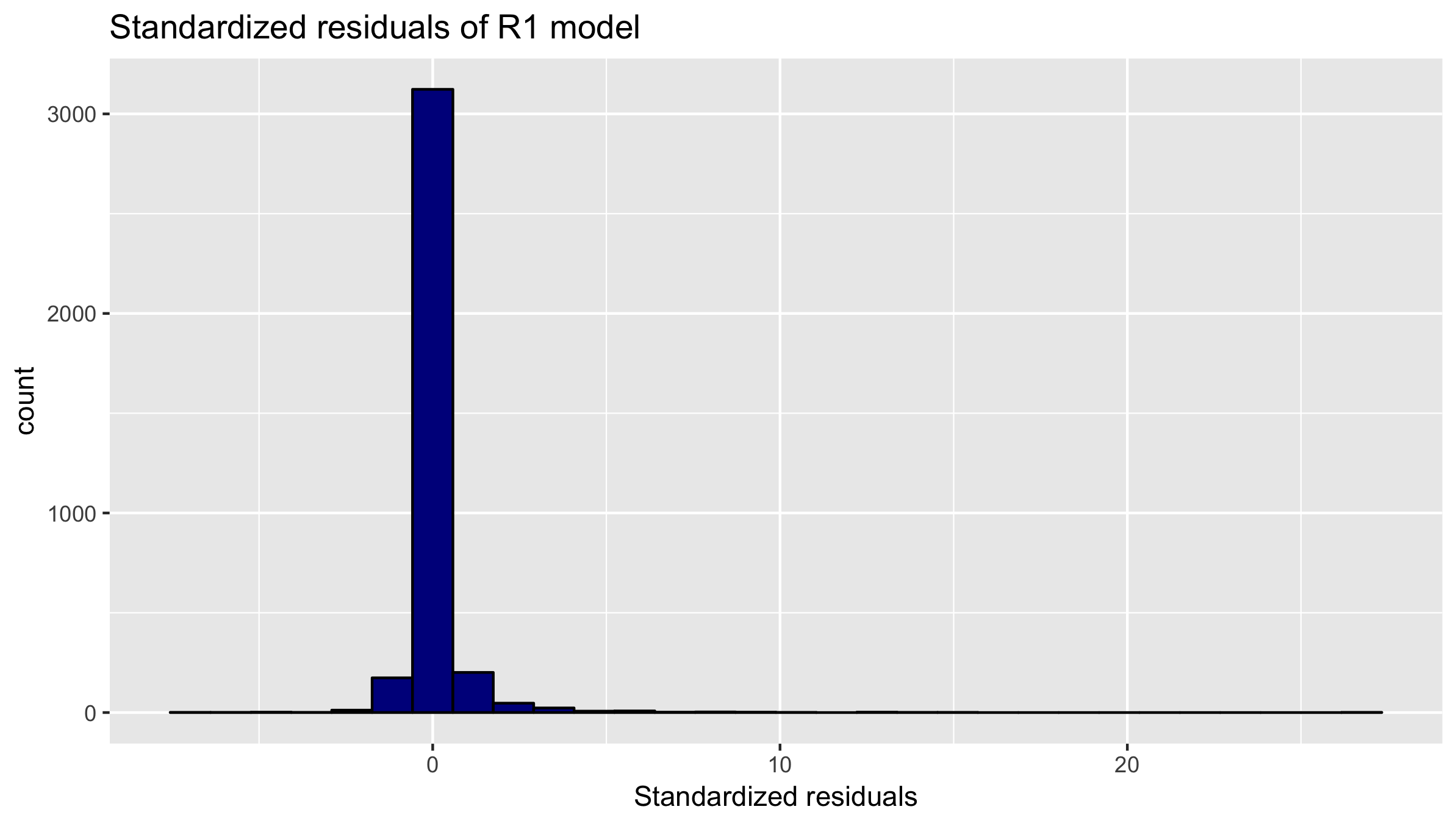
**Model 2 TOP:**

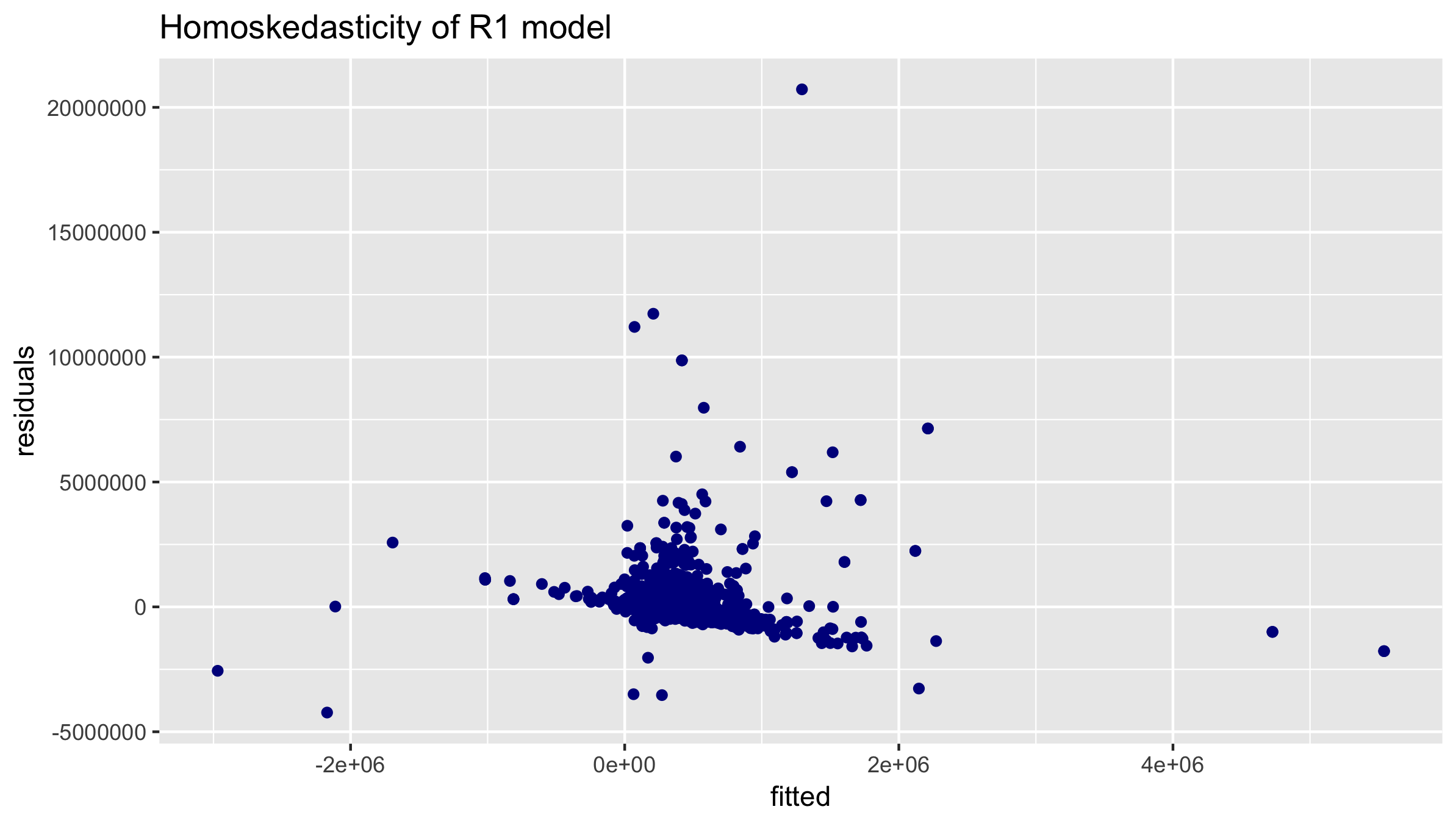
****

****

Model 2 has the similar distribution as Model 1, which is also fit but can be improved as well.

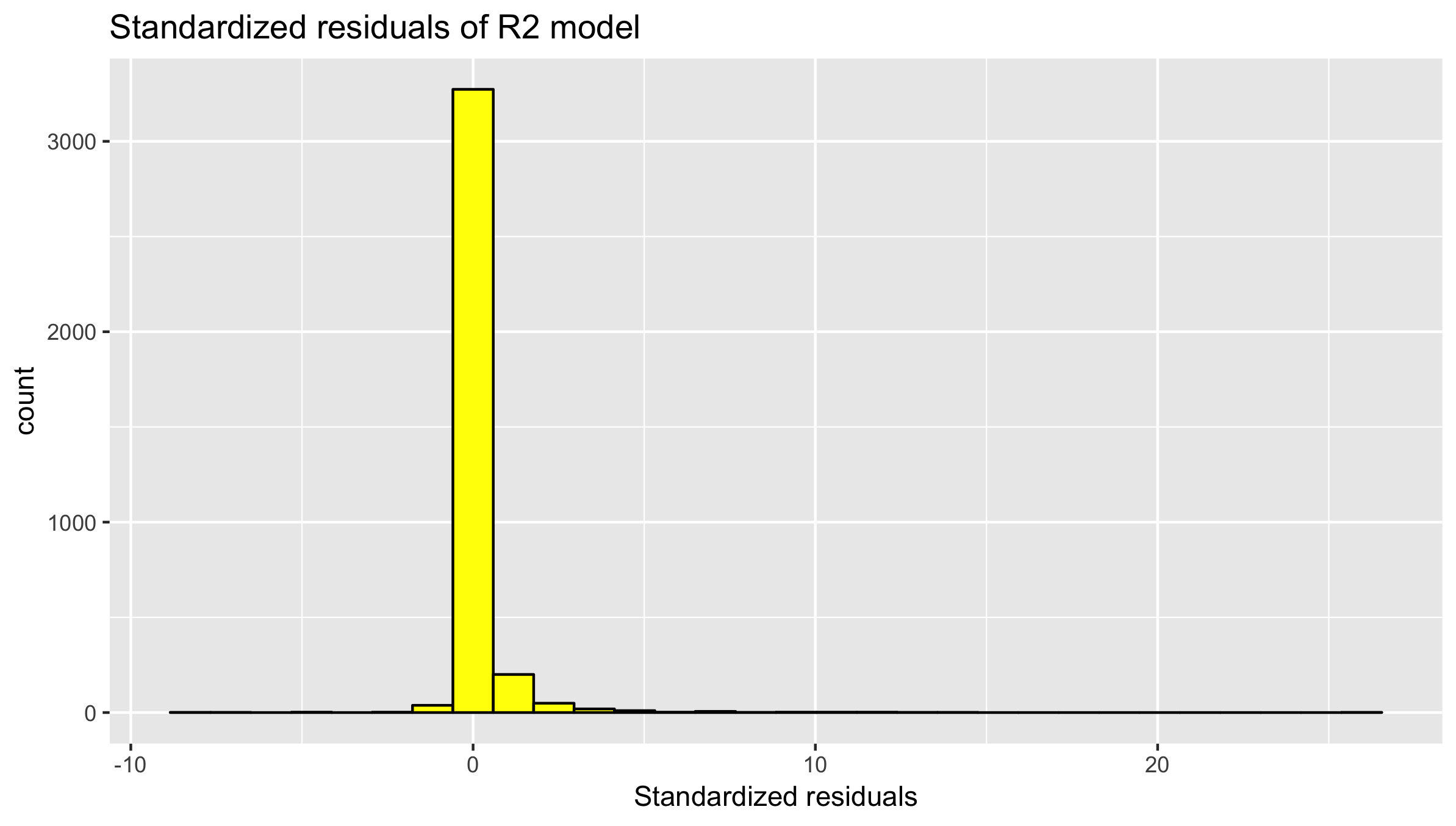
**Model 3 R1:**

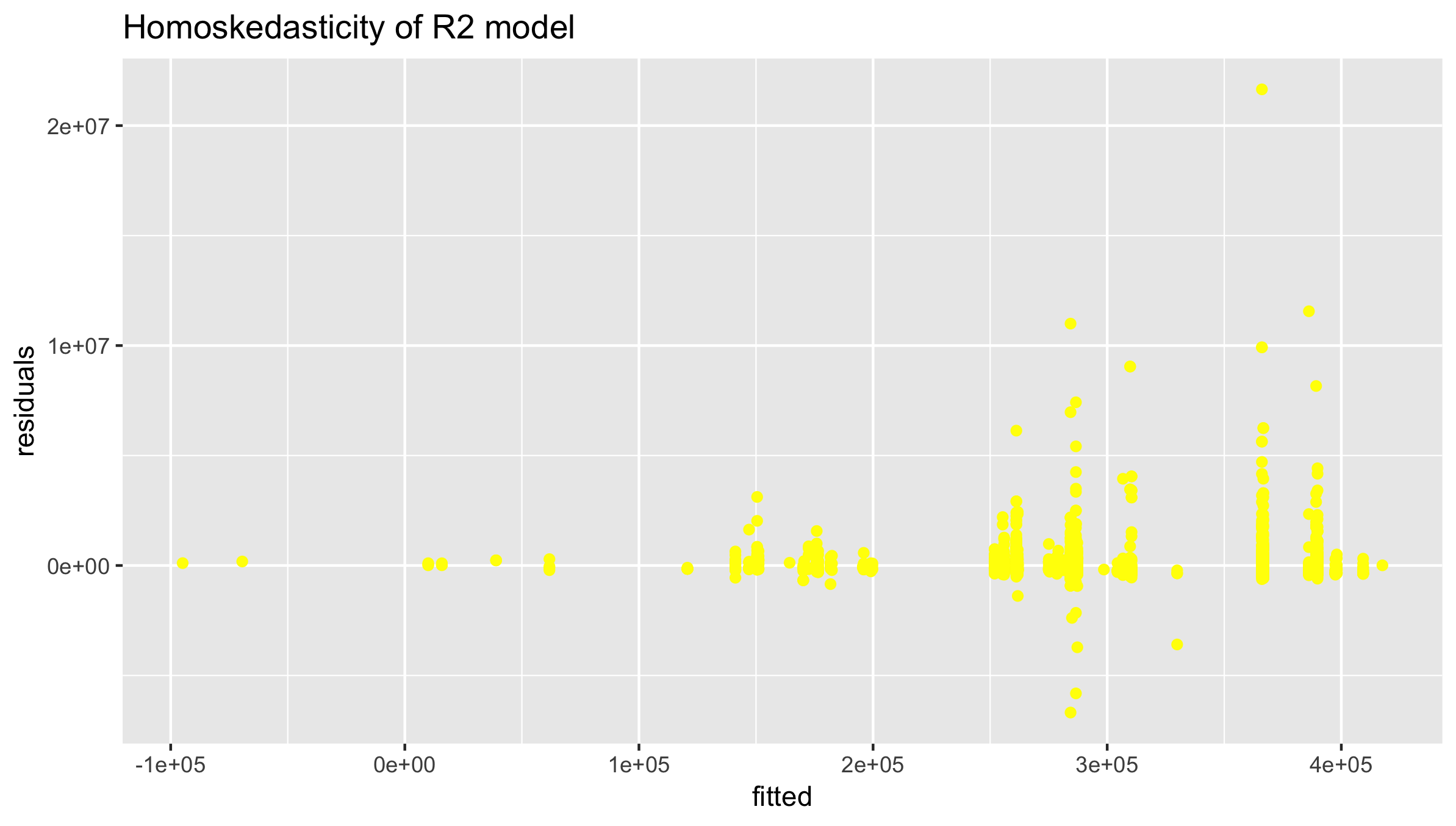
****

****

Model 3 has the similar distribution as Model 1 and Model2, indicating there is no big improvement from the former models.

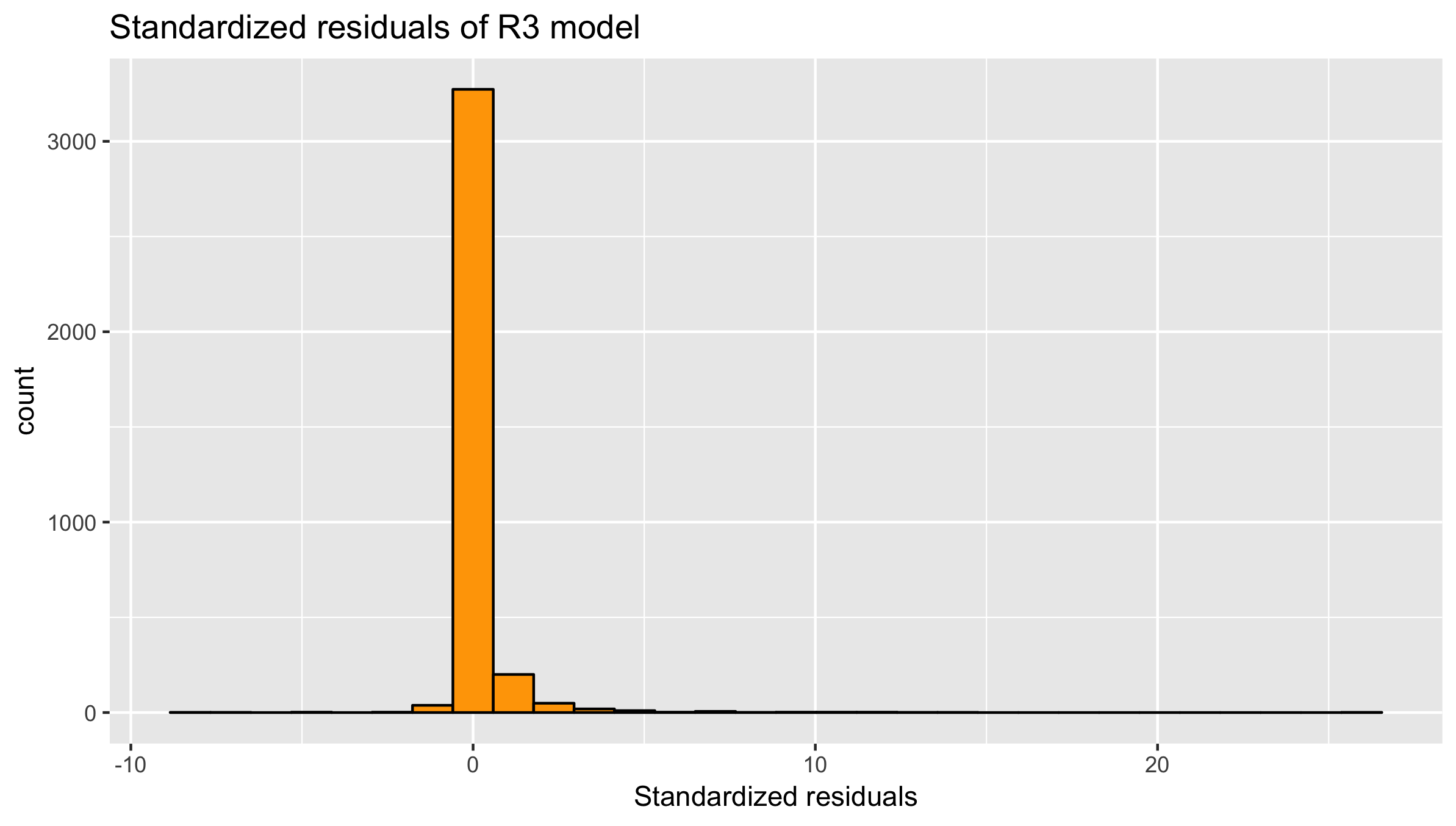
**Model 4 R2:**

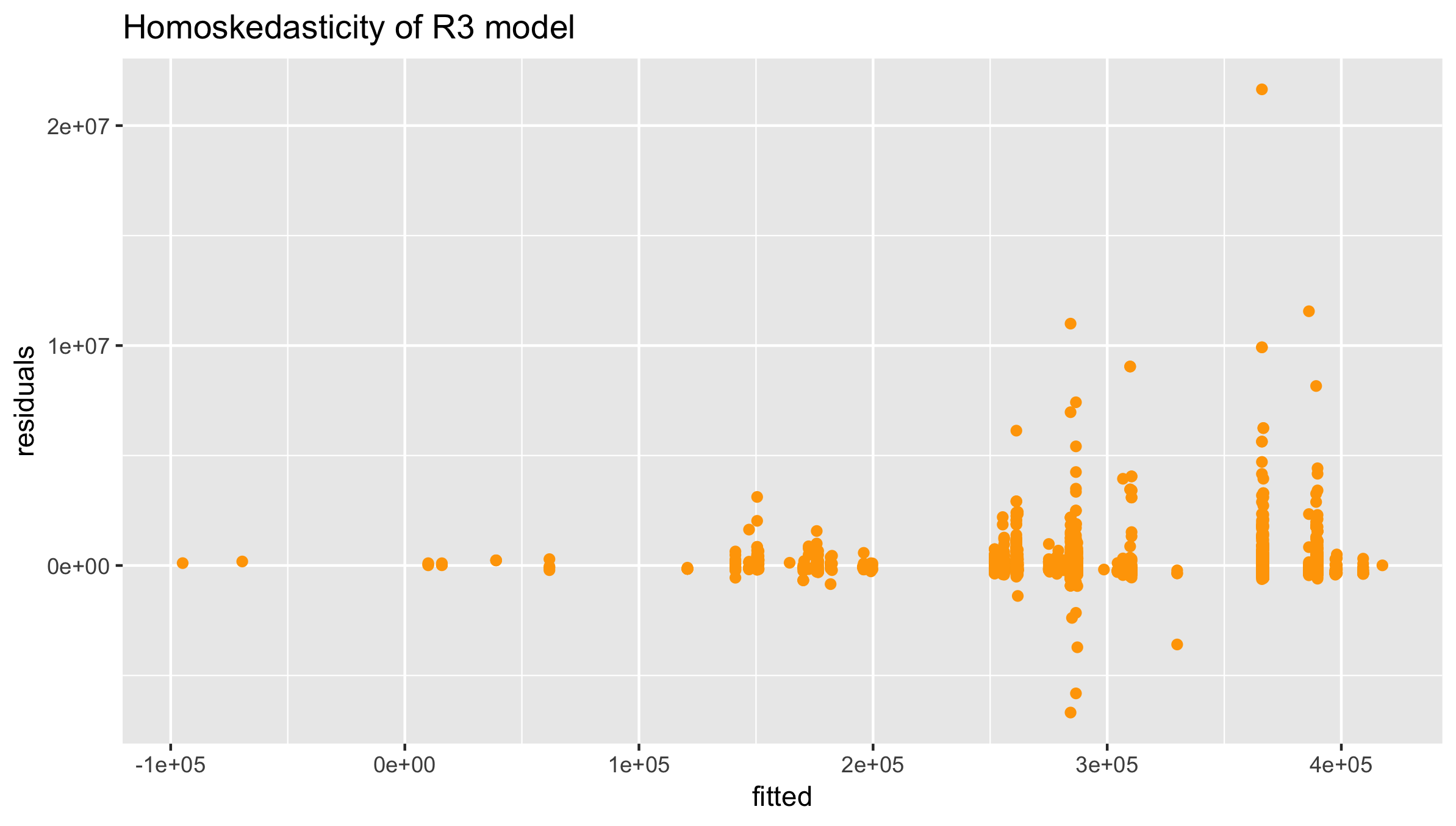
****

****

The plot shows a y-axis unbalanced residual , which means this model can be made significantly more accurate.

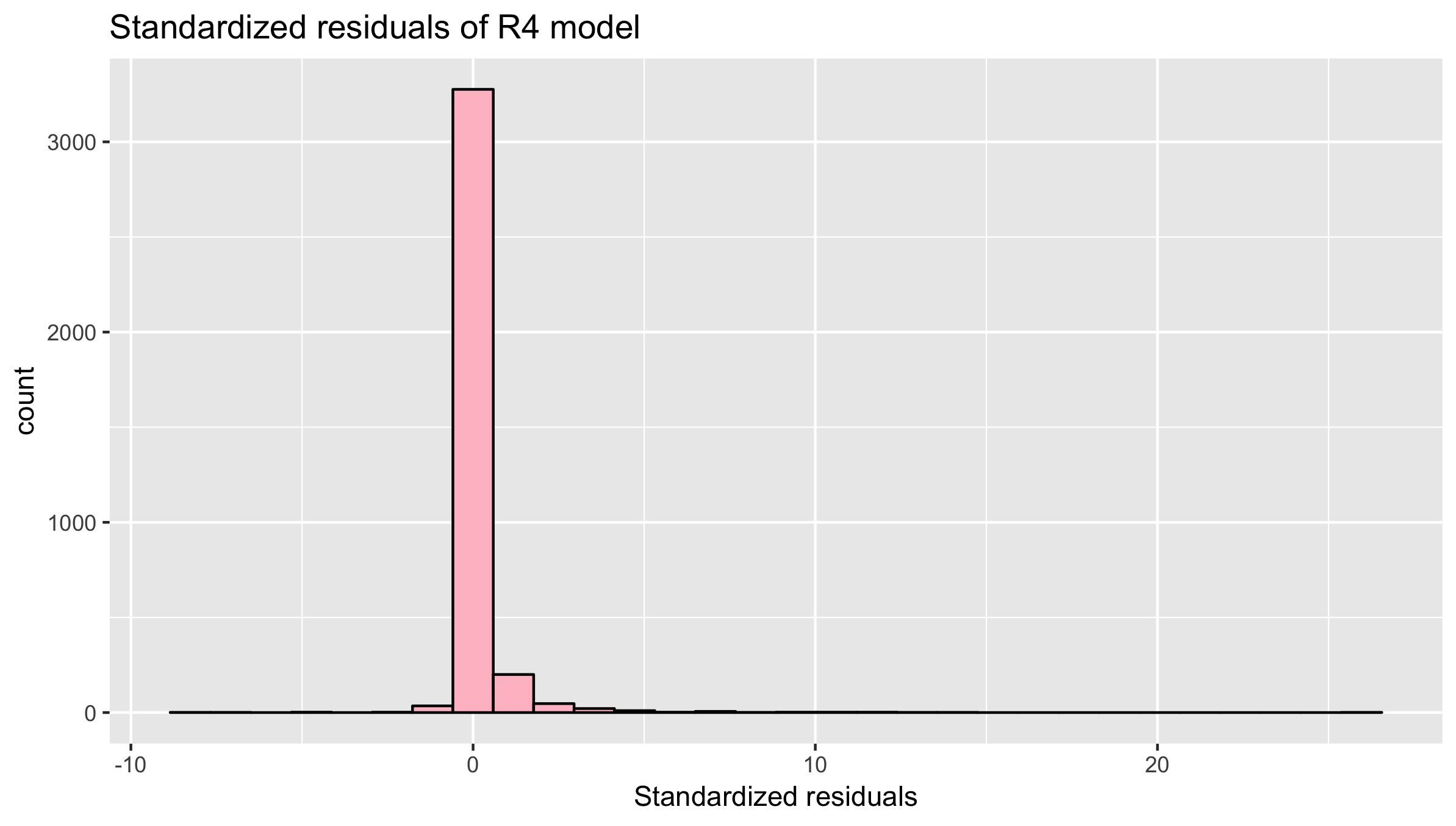
**Model 5 R3:**

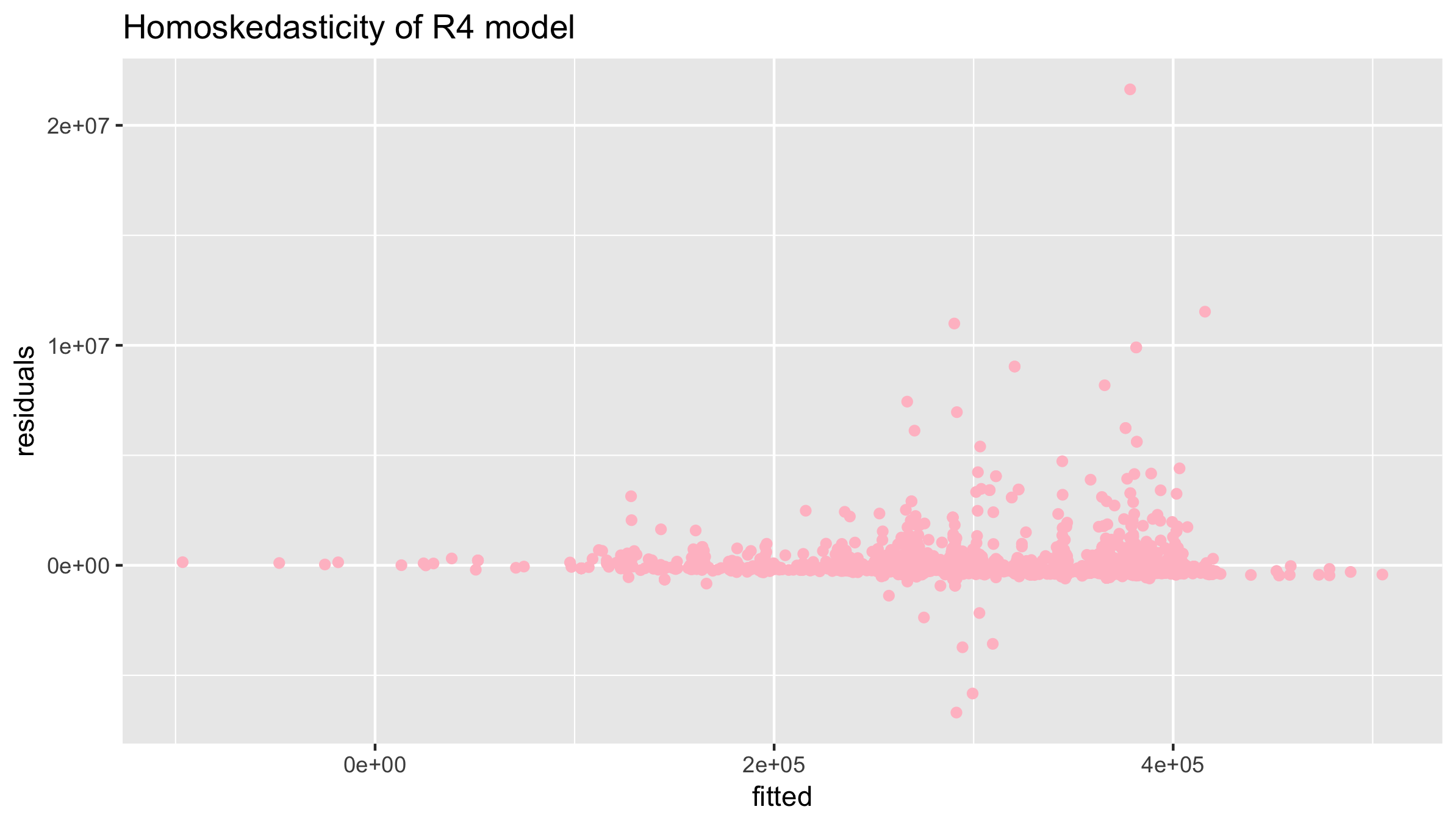
****

****

Same problem as Model 4, hence, this model can be made significantly more accurate.

**Model 6 R4:**

****

****

Model 6 plots show most of the predictions are a bit too high, and then some would be way too low and aren’t evenly distributed vertically. This almost always means the model can be made significantly more accurate. Most of the time we’ll find that the model was directionally correct but pretty inaccurate relative to an improved version.

**In general, Model 1 is more symmetric and bell-shaped than the other five models, which indicates the most fitted model among all the six models we built up.**

**5. Conclusions:**

**Appendix**

* 1. **Find correlations between features and training target (HH profit per acre)**

**3.3.1 For features already grouped by (clust, nh)**

To calculate the correlation between feature and training target, i.e. HH profit per acre, firstly we added features from aggregates data in batch mode. For features already grouped by (clust, nh), we added feature directly using the function below:

AddFeature <- function(data, feature){

all\_features <- data %>%

left\_join(feature, by=c("clust", "nh")) %>%

replace(., is.na(.), 0)

return (all\_features)

}

To add this kind of feature more efficiently, we developed a function so that we could add features from files in batch mode.

AddAggFeaturesFromFiles <- function(data, path, namePattern) {

files <- list.files(path, full.names = TRUE, pattern = namePattern)

for (i in 1:length(files)) {

file <- files[i]

data <- AddFeature(data, read\_dta(file))

}

return (data)

}

**3.3.2 For features not grouped by (clust, nh)**

In this situation, we aggregated features by (clust, nh), generate (mean, sum) of one feature, using the function shown below:

GetAggFeature <- function(data, dataCol){

aggData <- data %>%

select(c("clust", "nh", dataCol)) %>%

group\_by(clust, nh) %>%

summarise\_at(c(3), funs(mean, sum))

names(aggData)[3] <- paste(dataCol, colnames(aggData)[3], sep="\_")

names(aggData)[4] <- paste(dataCol, colnames(aggData)[4], sep="\_")

return (aggData)

}

After adding training target to all features list, 1 training target and 157 features were generated.

**3.3.3 Calculate the correlations**

In this part, we calculated the correlations in batch mode, and generated a new data frame *all\_correlations* with three columns. Column names are *index, colName*, and *correlation,* ordered by correlation in descending order. Below is the function we used.

findCorrelation <- function(a) {

df <- data.frame(index = (NA), colName=(NA), correlation = (NA))

for (i in 1:ncol(a)) {

correlationP <- cor(a[i], a[1])

row <- c(i, colnames(a[i]),correlationP)

df<- rbind(df, row)

}

df <- df %>%

filter(!is.na(colName))

df <- df[order(df$correlation, decreasing = T),]

return (df)

}

After applying above function to calculate the correlations, we selected top n features, and passed them to *lm()* parameter. Through the comparison, it is not hard to find that these features overlap the features that we used to calculate our training target. Thus, we later dropped these features.

Same method has been applied to raw data. We found that features such as livstcd5, livstcd6, cropcd8, cropcd11, rootcd8, rootcd18, etc. are statistic significant compared to other features, thus we trained model based on these features.