University of Western Sydney
200045 Quantitative Project
–
A GLM-Based Method for Investigating Unclaimed Money

Anjali Sharma, Chen Zhong. Supervisor: Glenn Stone

2015-10-23

Contents

1 Acknowledgements

We would like to thank Dr. Glenn Stone, for his for his invaluable and practical suggestions throughout the many road blocks that appeared during this project. His willingness to give his time so generously has been very much appreciated.

2 Declaration:

- I hold a copy of this report, which I can produce if the original is lost or damaged.
- I hereby certify that no part of the report has been copied from any other student's work or from any other source except where due acknowledgement is made in the report.
- No part of the report has been written for me by any other person except where such collaboration was authorised by the unit coordinator.
- I am aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism. A copy may be retained in a database for future plagiarism checking.

3 Introduction

This research takes an in-depth look at both the monetary value and the number of accounts containing unclaimed money across NSW. Unclaimed money is the money held by the NSW state government that is given by an entity but remained uncollected by the recipient for some amount of time. The period of inactivity has been defined as 7 years for accounts held before December 2012, 3 years for accounts held between December 2012 and May 2015, and 7 years for those held after May 2015.

The question we have addressed is the effect of the area, organisation type, year account was opened and year account was classified on the number of accounts as well as the monetary value.

3.1 Why research unclaimed money?

Unclaimed Money is a public data set that has attracted little interest due to its lack of profitability for those without a claim to any of the money.

Some analysis takes place through businesses which locate claimants and offer claiming services in exchange for a percentage of the claim. However, this ignores the vast amount of money held in small amounts across many accounts.

The Australian Securities and Investments Commission is tasked with locating and informing claimants of monies but have expressed lack of resources, and an overwhelming number of accounts, due to law changes.

3.2 Approach

Data is available for unclaimed money for 380,671 accounts, going back to the year 1900. Methodology involves data visualization using Tableau, a drag and drop tool used for visual exploration, as well as R for merging, cleaning, sorting and GLM analysis.

- The individual parts of the project are the modelling for the amounts [Chen] and the modelling for the number of accounts [Anjali]
  - [Anjali] Counting accounts means weighting areas by their population to obtain meaningful frequencies
  - [Anjali] Trialing and testing different models, testing the significance of each variable, and forming conclusions from the frequency table.
  - [Chen] Amounts requires working out the distribution it has with respect to each variable and using the appropriate link function to create a meaningful model.
  - [Chen] Finding the right distribution is going to be a challenge due to the amount of distributions covered by GLM and their similarity.

3.2.1 Extracting Postcodes using regex function

Data collation has so far involved file conversion, merging the 26 alphabetical files into one, and extracting postcodes using a regex function The regex pattern was sourced using Australia Post postcode area definitions. PO Box and Locked Bag numbers were being included in the regex, therefore all 4 digits numbers after the words PO Box or Locked Bag had to be removed before extracting the data.

Data that was thrown out was defined as a sequence of 4 digits arriving after the word “box” or “bag” regardless of letter case. The regex pattern to define the postcode is a sequence of or-statements, with one statement for each state of territory. The boundaries of each digit are defined between square brackets, with curly brackets indicating the number of times to repeat the last square bracket’s contents. A loop was used to write the outcome of the pattern to a new variable:

```r
#Duplicate addresses
address = UM$Owner.Address

#Chuck out PO Boxes
address = gsub("box [0-9][0-9][0-9][0-9]", "", address, ignore.case = TRUE)
#Chuck out Locked Bags
address = gsub("bag [0-9][0-9][0-9][0-9]", "", address, ignore.case = TRUE)

#Pattern for extracting postcodes (line-break for printing only)
AU <- "(0[289][0-9]{2})|([1345689][0-9]{3})|(2[0-8][0-9]{2})|(290[0-9])|
       (291[0-4])|(7[0-4][0-9]{2})|(7[8-9][0-9]{2})"

#Identifies AU Postcode
AUPC <- ifelse(grepl(AU, UM$Owner.Address), 1, 0)

#Prints AU Postcode
x = regexpr(AU, address)
UM$Postcode <-  substring(address, x, x + attr(x, "match.length") - 1)
print(UM$Postcode)

#Creates new file with newly extracted data
write.csv(UM, file = "UMPCs.csv")

#Displays addresses according to postcode extraction

#If no postcode found, prints address
ifelse(UM$Postcode == "", as.character(UM$Owner.Address), "")

#If postcode found, prints address
ifelse(UM$Postcode == "", "", as.character(UM$Owner.Address))
```

3.2.2 Classifying Postcodes into categories

Postcodes needed to be categorised into larger areas to allow sufficient data in each of the categories to facilitate analysis of the data overall. The postcode ranges were sourced online:

Area	Postcode Range
Canberra CBD	2600, 2601, 2610
Canberra	2601 - 2609
Rest of ACT	2611 - 2620
Sydney CBD	1100 - 1299, 2000, 2001, 2007, 2009
Sydney Metro	2002 - 2006, 2008, 2010 - 2234
Riverina Area	2640 - 2660
Wollongong	2500 - 2534
Newcastle	2265 - 2333
Northern Rivers	2413 - 2484
Rest of NSW	2235 - 2412, 2485 - 2999
Melbourne CBD	3000 - 3006, 3205, 8000 - 8399
Melbourne Metro	3007 - 3204, 3206, 3207
Rest of VIC	3208 - 3999
Brisbane CBD	4000, 4001, 4003, 9000 - 9015
Brisbane Metro	4002, 4004 - 4207, 4300 - 4305, 4500 - 4519
Gold Coast	4208 - 4287
Sunshine Coast	4550 - 4575
Rest of QLD	4288 - 4299, 4306 - 4499, 4520 - 4549, 4576 - 4999
Adelaide CBD	5000, 5001, 5004, 5005, 5810, 5839, 5880 - 5889
Adelaide Metro	5002, 5003, 5006 - 5199
Rest of SA	5200 - 5749, 5825 - 5854
Perth CBD	6000, 6001, 6004, 6827, 6830 - 6832, 6837 - 6849
Perth Metro	6002, 6003, 6005 - 6199
Rest of WA	6200 - 6826, 6828, 6829, 6833 - 6836, 6850 - 6999
Hobart CBD	7000, 7001
Hobart Metro	7002 - 7099
Rest of TAS	7100 - 7999
Darwin Metro	0800 - 0832
Rest of NT	0833 - 0899

We classified postcodes by creating a function with a number of if-statements:

```r
PC2Area = function(PC) {
  if(is.na(PC))
    return("Unknown")
  if( PC %in% c(2600, 2601, 2610) )
    return("Canberra CBD")

  ...


  if( PC >= 0833 && PC <= 0899 )
    return("Rest of NT")
  return("XXXX")
}

## To check temporary variable 'Area' holds all results
Area = sapply(UM$Postcode, PC2Area)

## Shows first few unmatched postcodes
head(UM$Postcode[Area=="XXXX"])

## Write new column 'Area'
UM$Area = sapply(UM$Postcode, PC2Area)

write.csv(UM, file = "UM.csv")
```

3.2.3 Categorising Organisations

The 'Organisation' variable contains 4049 factors, meaning 4049 different companies comprise the entire set of 380,671 accounts. The frequency in which each organisation appears in the data set varies wildly, from 1 to 8837 times. Since we have no way to automatically categorise them, the most efficient method to convert this variable into something that can be analysed, is to categorise a certain number of the organisations which occur most frequently. To decide on this number, we created lists for come numbers of the top occurring companies and worked out the percentile of accounts each list of companies covered. We made the number of top companies increase in increments of 10.

```
##    NumEntries NumTopOrg percentage
## 1         268        70  0.7052632
## 2         276        80  0.7263158
## 3         283        90  0.7447368
## 4         289       100  0.7605263
## 5         295       110  0.7763158
## 6         299       120  0.7868421
## 7         304       130  0.8000000
## 8         307       140  0.8078947
## 9         311       150  0.8184211
## 10        314       160  0.8263158
```

Categories for the top 100 companies were collated by hand using company websites and information available on ASX (Australian Share Index). The sum of the number of accounts increased by about 1% for every 10 companies added above 100, also with diminishing return. whereas leading up to 100, the number of accounts was increasing by about 2%. We decided to use top 100 occurring companies because it covers 76% of the data. It is practical for the time we have.

We manually categorised these top 100 companies into 16 categories. We chose the 16 categories based on the nature of the business. We used two methods to determine the categories. First, we searched for the exact organisation name online and went to their official page,if there was one. We used the official page to determine the type of business. Second, we went to the ASX website and again searched the organisation name. If the organisation was registed on the ASX, we then extracted the category of each organisation from that organisation's page (from the 'Details' tab) on the ASX website.

This table is our final result of categorising each of the top 100 most frequently occurring organisations.

Organisation	Type
QANTAS AIRWAYS LTD	Airline
GRAINCORP LTD	Beverage and Food
GOODMAN FIELDER LTD	Beverage and Food
COCA-COLA AMATIL LTD	Beverage and Food
LION NATHAN LTD	Beverage and Food
USANA AUST PTY LTD	Beverage and Food
MACQUARIE UNI	Education
THE UNI OF NSW	Education
AMP LTD	Financial Services
AMP LTD/AMP	Financial Services
WESTPAC BANKING CORP	Financial Services
VEDA ADVANTAGE LTD	Financial Services
ING AUST LTD	Financial Services
AMERICAN EXPRESS AUST LTD	Financial Services
CAPITAL FIN AUST LTD	Financial Services
CHALLENGER LTD/CGF	Financial Services
CWEALTH SECURITIES LTD	Financial Services
ESANDA FIN CORP LTD	Financial Services
TOYOTA FIN AUST LTD	Financial Services
AMP GROUP FIN SERVS LTD	Financial Services
CHALLENGER LTD	Financial Services
TAB LTD	Gaming
TABCORP HLDGS LTD & TABCORP HLDGS LTD	Gaming
ARISTOCRAT LEISURE LTD	Gaming
OFFICE OF REAL EST SERVICES	Government
CITY OF SYDNEY	Government
PARRAMATTA CITY CNCL	Government
FAIRFIELD CITY CNCL & WOLLONDILLY SHIRE CNCL	Government
BORAL LTD	Industrial
RINKER GROUP PTY LTD	Industrial
OFFSET ALPINE PRINTING GROUP PTY LTD	Industrial
ONESTEEL LTD	Industrial
CSR LTD	Industrial
BRAMBLES INDUSTS LTD	Industrial
CSR LIMITED	Industrial
INTOLL GROUP	Industrial
MBF AUST LTD	Insurance
INS AUST GROUP	Insurance
INS AUST GROUP LTD	Insurance
INS AUST GROUP LTD/IAG	Insurance
GIO GEN LTD	Insurance
THE HOSPITALS CONTRIBUTION FUND OF AUST LTD	Insurance
QBE INS (AUST) LTD	Insurance
SUNCORP	Insurance
NIB HEALTH FUNDS	Insurance
ALLIANZ INS AUST LTD	Insurance
CWEALTH INS LTD	Insurance
ING IND FUND	Insurance
ACE INS LTD	Insurance
CGU INSURANCE LTD	Insurance
ECORP LTD	IT

Organisation	Type
APN NEWS & MEDIA LTD	Media Company
CNSLD MEDIA HLDGS LTD	Media Company
TEN NETWORK HLDGS	Media Company
FAIRFAX MEDIA LTD	Media Company
ARRIUM LTD	Mining
AUN	Mining
GENWORTH FIN MORTGAGE INS PTY LTD	Mortgage Lending
GALILEE SOLICITORS	Mortgage Lending
QBE LENDERS MORTGAGE INSCE LTD	Mortgage Lending
NATIONAL LENDING SOLUTIONS	Mortgage Lending
CAMBRIDGE CREDIT CORP LTD	Property Management
MIRVAC REAL EST INVEST TRUST	Property Management
GENERAL PROPERTY TRUST	Property Management
LEND LEASE GROUP/LLC	Property Management
MIRVAC REAL EST INVEST TRUST (FORMERLY MERIDIAN)	Property Management
LEND LEASE CORP LTD	Property Management
WESTFIELD GROUP	Property Management
WESTFIELD GROUP/WDC	Property Management
MIRVAC GROUP	Property Management
INVESTA PROPERTY GROUP	Property Management
STOCKLAND CORP LTD	Property Management
WOOLWORTHS LTD	Retailer
WOOLWORTHS LIMITED	Retailer
DAVID JONES LIMITED	Retailer
WOOLWORTHS LTD/WOW	Retailer
METCASH LTD	Retailer
AMP SUPER SAVINGS TRUST	SuperFund
AUST'N ELIGIBLE ROLLOVER FUND	SuperFund
SUPERTRACE ELIGIBLE ROLLOVER FUND	SuperFund
AMP ELIGIBLE ROLLOVER FUND	SuperFund
UNIVERSAL SUPER SCHEME FUND	SuperFund
STATE SUPER	SuperFund
BT FUNDS MGMT	SuperFund
CLUB PLUS SUPER PTY LTD	SuperFund
COLONIAL MUTUAL LIFE ASSUR LTD	SuperFund
AON ELIGIBLE ROLLOVER FUND	SuperFund
LEGAL & GEN SUPERTRACE	SuperFund
AUST PRIMARY SUPER FUND	SuperFund
SINGTEL OPTUS	Utility
ENERGY AUST (NCLE)	Utility
AUSGRID	Utility
THE AUSTRALIAN GAS LIGHT COMPANY	Utility
AGL ENERGY LTD	Utility
ORIGIN ENERGY LTD	Utility
ENERGY AUST	Utility
COUNTRY ENERGY	Utility
INTEGRAL ENERGY	Utility
ORIGIN	Utility
JACK GREEN (INTNL) PTY LTD	Utility

3.2.4 Adding population data for each area

We downloaded a file from ABS website (File name 3218.0 Regional Population Growth, Australia) which was released at 11:30 am 31/03/2015. We used the latest data from 2014, and categorised the postcode into 21 areas. We matched all 21 postcode areas to the corresponding ABS data:

PC AREA	ABS AREA	2014 POP.
Canberra CBD	Total Canberra Inner City	2823
Canberra	GREATER CANBERRA	79183
Rest of ACT	REST OF ACT	306813
Sydney CBD	Total Sydney Inner City	203774
Sydney Metro	GREATER SYDNEY	4636854
Riverina Area	Riverina Area	158144
Wollongong	Wollongong	296845
Newcastle	Newcastle	368131
Northern Rivers	Richmond tweed	242116
Rest of NSW	REST OF NSW	1612608
Melbourne CBD	Total Melbourne City	122190
Melbourne Metro	GREATER MELBOURNE	4318138
Rest of VIC	REST OF VIC	1401339
Brisbane CBD	Total Brisbane Inner	65542
Brisbane Metro	GREATER BRISBANE	2209018
Gold Coast	Total Gold Coast	560266
Sunshine Coast	Sunshine Coast	335874
Rest of QLD	REST OF QLD	1869515
Adelaide CBD	Total Adelaide City	22690
Adelaide Metro	GREATER ADELAIDE	1281941
Rest of SA	REST OF SA	381083
Perth CBD	Total Perth City	108216
Perth Metro	GREATER PERTH	1912987
Rest of WA	REST OF WA	552186
Hobart CBD	Total Hobart Inner	50757
Hobart Metro	GREATER HOBART	168486
Rest of TAS	REST OF TAS	295519
Darwin Metro	Total Darwin City	26281
Rest of NT	REST OF NT	78412

We found the individual area population for the CBD of each state and territory in the ABS file and subtracted it from the total population of the corresponding state to uncover the presented population size.

The aim is not to forecast any changes in population based on area, but to use the ratio of population between each area to gain perspective of how much the number of accounts depends on population or other factors. Therefore, population data is only needed for the one year, 2014. If there was more time, we could have used the ratios from the appropriate year corresponding to the year the account was classified as unclaimed. This would have allowed us to model the frequency of each area over time.

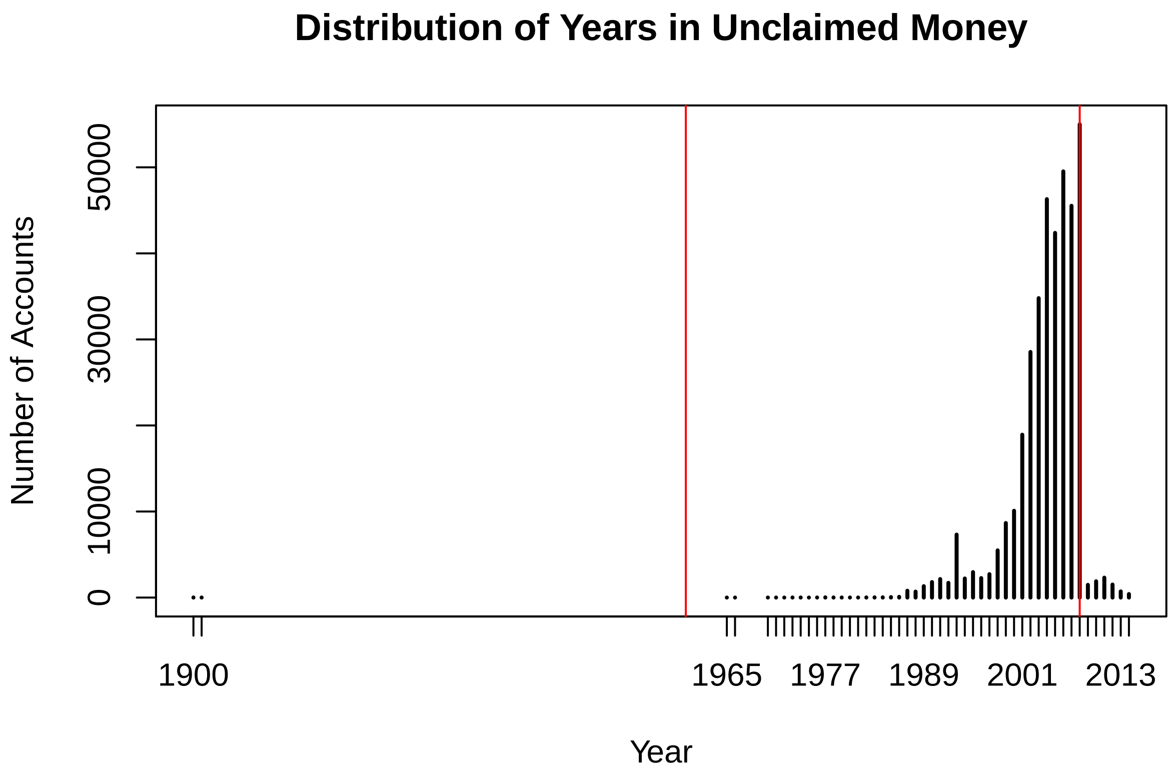
3.2.5 Truncating the data

The date in the original data (‘Date’), which indicates when the accounts were classified as unclaimed, was in DD/MM/YYYY format. We only need the year for our model, since the time the account was created (‘Year’) was in YYYY format only. Below is the code that we used to extract the year from the ‘Date’ column in the original data, which we then rewrote as the variable ‘Classified’.

```
dates = as.Date(UM$Date, format="%d-%b-%y")
UM$Classified= as.numeric(format(dates,"%Y"))
write.csv(UM, file = "UM.csv", row.names = FALSE)
```

Then we plotted the distribution of the years, and found two issues.

```
plot(table(UM$Year), ylab = "Number of Accounts", xlab = "Year", main = "Distribution of Years in Uncla
abline(v = 1960, col = "red")
abline(v = 2008, col = "red")
```



Firstly, the entries from 1900 and 1901 had potentially high leverage. We decide to remove them, since there were only 19 entries before 1960, which is an insignificant percentage of the total 385,000 entries. Secondly, there seems to be a sharp drop from 2008, indicating a different distribution to the majority of the data. Taking these facts into account in our analysis would require special treatment and due to time constraints, we chose to use the data between 1960 and 2008.

Here we omitted any rows which contain factors that are NA, or not applicable and only kept those variables that are useful to our model. UML is the subset that we decided to keep.

```
UM2 = na.omit(UM[,c("Amount", "Year", "Organisation.Type", "Area", "Classified")])
UML = subset(UM2, Year > 1960 & Year < 2009)
```

4 Statistical Theory

4.1 Concept

Linear models assume a linear relationship between the mean of the response and the predictors, as well as a Normal distribution of the response around this mean. Generalised linear models (GLM) generalise this in two ways:

1. The mean of the response is related to the linear combination of the linear predictors via the link-function.
2. The distribution of the response around this mean can be any distribution from the exponential family

4.2 Matrix Formulation of Linear Models

Equation for each observation of the response variable Y from 1 to n:

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \cdots + \beta_p X_{1p} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_{21} + \cdots + \beta_p X_{2p} + \varepsilon_2 \\
Y_3 &= \beta_0 + \beta_1 X_{31} + \cdots + \beta_p X_{3p} + \varepsilon_3 \\
\vdots \\
Y_n &= \beta_0 + \beta_1 X_{n1} + \cdots + \beta_p X_{np} + \varepsilon_n
\end{aligned}$$

Turn these simultaneous equations into matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & \dots & \dots & X_{1p} \\ 1 & X_{21} & \dots & \dots & \dots & X_{2p} \\ 1 & X_{31} & \dots & \dots & \dots & X_{3p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & \dots & \dots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Turn these matrix into a single equation:

$$Y = X\beta + \varepsilon$$

Here, Y is the vector of each observation of the response variable, X is the matrix of each of the predictor variables which explain the mean of the response, β is the vector of the coefficients which define the contribution of each predictor variable and ε is the vector of the i^{th} row vector of the noise variables assumed to be Normal in linear models.

4.3 Estimating β when the response is Normally Distributed

Least squares is one criterion commonly used for estimating the $\hat{\beta}$. Least squares can be seen as the maximum likelihood method when the data is Normally distributed. The idea behind maximum likelihood is to choose $\hat{\beta}$ to maximise the "chance" of getting the data we actually observed. To explain the method of least squares, we will first start with a simple linear model:

This is the formula for simple linear regression:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

Given a sample $(x_i, y_i), i = 1, 2, ..., n$, the likelihood of β is defined as:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} f(y_i; \beta)$$

The function f is the probability density function of y, and depends on β and x. For the Normal distribution:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(y - \beta_0 - \beta_1 x)^2}{2\sigma^2}$$

The maximum likelihood method chooses β to maximise $\mathcal{L}(\beta)$. It is easy to work with the log-likelihood which has the same maxima, since log is a monotone function.

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \log f(y_i; \beta)$$

Substituting the Normal probability density function for $f(y_i; \beta)$:

$$\begin{aligned}
\mathcal{L}(\beta) = \sum_{i=1}^{n} \log f(y_i; \beta) &= \sum_{i=1}^{n} \log[\frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(y_i - \beta_0 - \beta_1 x)^2}{2\sigma^2}] \\
&= \sum_{i=1}^{n} [\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x)^2] \\
&= \sum_{i=1}^{n} [\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x)^2] \\
&= -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2
\end{aligned}$$

Note that there is a negative sign in front of $\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$. Therefore maximising the likelihood of β can be accomplished by minimising $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$, which is the least squares criteria.

4.4 The Generalised Linear Model

The linear model makes the assumption that Y is of Normal distribution with the two parameters below:

$$Y \sim N(\mu, \sigma^2)$$

And, $\mu = x^t \beta$.

GLM relaxes these assumptions to any distribution in the exponential family, which is discussed further in the next section. And, $\mu = g^{-1}(x^t \beta)$, for some g, meaning $x^t \beta$ is not directly a parameter, but instead some transformation g can be made to the parameter μ to obtain $x^t \beta$. This g is known as the link function.

In terms of our project, the linear model may be applicable to analysing the amounts in each account, since amounts are non-binary, non-count data which is possibly Normal. The linear model is not applicable to analysing the number of counts, since that is count data, involving only discrete numbers and no negatives. Therefore, generalised linear models are required for this project.

GLM also implies a relationship between the variance of Y and the mean μ, where ϕ is a dispersion parameter:

$$Var(Y_i) = \phi Var(\mu_i)$$

4.5 Exponential family

GLM handles distributions in the exponential family. The exponential family of statistical models that GLM uses takes the following general form:

$$f(y|\theta, \phi) = \exp(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi))$$

θ and ϕ are some parameters. ‘a’, ‘b’ and ‘c’ are functions that are also distinct for each member of the exponential family, and can therefore be used to specify the member of the exponential family.

The exponential family distribution has mean and variance (See McCullagh and Nelder 1994, pg. 29):

$$\mathbf{E}(Y) = \mu = b'(\theta)$$
$$\mathbf{Var}(Y) = b"(\theta)a(\phi)$$

The Binomial and Poisson distributions are examples of the exponential distribution family that may be relevant to our project.

These are the major distributions and their canonical link functions. For certain link functions, the relationship $\theta = g(\mu)$ holds. In this case, g is called the canonical link function.

Family	C.Link	Variance Function
Normal	μ	1
Poisson	$\log(\mu)$	μ
Binomial	$\log(\frac{\mu}{1-\mu})$	$\frac{\mu}{1-\mu}$
Gamma	$\frac{1}{\mu}$	μ^2

Canonical links lead to desirable statistical properties of the generalised linear model, and hence tend to be used by default. However there is no prior reason why this link should give the best fit when modelling. How these canonical link functions are derived is shown for two non-Normal distributions from the exponential family in the next two sections.

4.6 Binomial Distribution

For a binomial distribution, the canonical link function is derived as follows:

$$\begin{aligned}
\log(f(y|\theta,\phi)) = \log(\binom{n}{y}\mu^y(1-\mu)^{n-y}) &= \log\binom{n}{y}\mu^y(1-\mu)^{n-y}) \\
&= \log\binom{n}{y} + \log(\mu^y) + \log((1-\mu)^{n-y}) \\
&= \log\binom{n}{y} + y\log(\mu) + (n-y)\log(1-\mu) \\
&= \log\binom{n}{y} + y\log(\mu) + n\log(1-\mu) - y\log(1-\mu) \\
&= \log\binom{n}{y} + y\log(\frac{\mu}{1-\mu}) + n\log(1-\mu)
\end{aligned}$$

Therefore, the canonical link function is given by

$$\theta = \log(\frac{\mu}{(1-\mu)})$$

Note: log is base e.

4.7 Poisson Distribution

For a Poisson distribution, the link function is derived as follows:

$$\begin{aligned}
\log(f(y|\theta,\phi)) = \log(\frac{e^{-\mu}\mu^y}{y!}) &= \log(\frac{e^{-\mu}\mu^y}{y!}) \\
&= \log(e^{-\mu}) + \log(\mu^y) - \log(y!) \\
&= -\mu + \log(\mu^y) - \log(y!) \\
&= \log(\mu^y) - \mu - \log(y!) \\
&= y\log(\mu) - \mu - \log(y!)
\end{aligned}$$

Therefore, the canonical link function is given by

$$\theta = \log(\mu)$$

This link function is the most natural choice for the Poisson distribution, since the mean μ for Poisson is always above zero. The obvious choice is $\mu = e^{X\beta}$. Writing the equation for the link function as a function of the mean μ gives $X\beta = \log(\mu)$ to make certain that the mean μ is positive. This will have the effect of making addition within X affect μ multiplicatively.

It is important to note that the link functions mentioned are the canonical link functions, and are not the only option, or always the best option. The canonical link is the most obvious choice, for the corresponding distribution, mathematically and computationally. Ultimately, the link function used to create our final model will be chosen based on the best fit with our data.

For the Poisson distribution, maximum likelihood is used for estimation.

4.8 Maximum Likelihood for the Poisson Distribution

For Poisson, β is estimated using maximum likelihood.

The probability density function of a Poisson distribution is:

$$f(y) = \exp^{-\mu} \frac{\mu^y}{y!}$$

The maximum likelihood equation in this case would be:

$$\mathcal{L}(\beta) \prod_{i=1}^{n} \exp^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!}$$

We then log-transform f to turn the formula into a sum, instead of a product:

$$\begin{aligned}
\prod_{i=1}^{n} \exp^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!} &= \sum_{i=1}^{n} \log[\exp^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!}] \\
&= \sum_{i=1}^{n} [-\mu_i + y_i \log(\mu_i) - \log(y_i!)]
\end{aligned}$$

If we then replace μ with $g^{-1}(X\beta)$:

$$\sum_{i=1}^{n} [-g^{-1}(X\beta) + y_i \log(g^{-1}(X\beta)) - \log(y_i!)]$$

Since $\log(y_i)$ is not dependent on β, it does not need to be estimated. So our maximum likelihood formula becomes:

$$\mathcal{L}(\beta) \sum_{i=1}^{n} [-g^{-1}(X\beta) + Y_i \log(g^{-1}(X\beta))]$$

5 Analysis

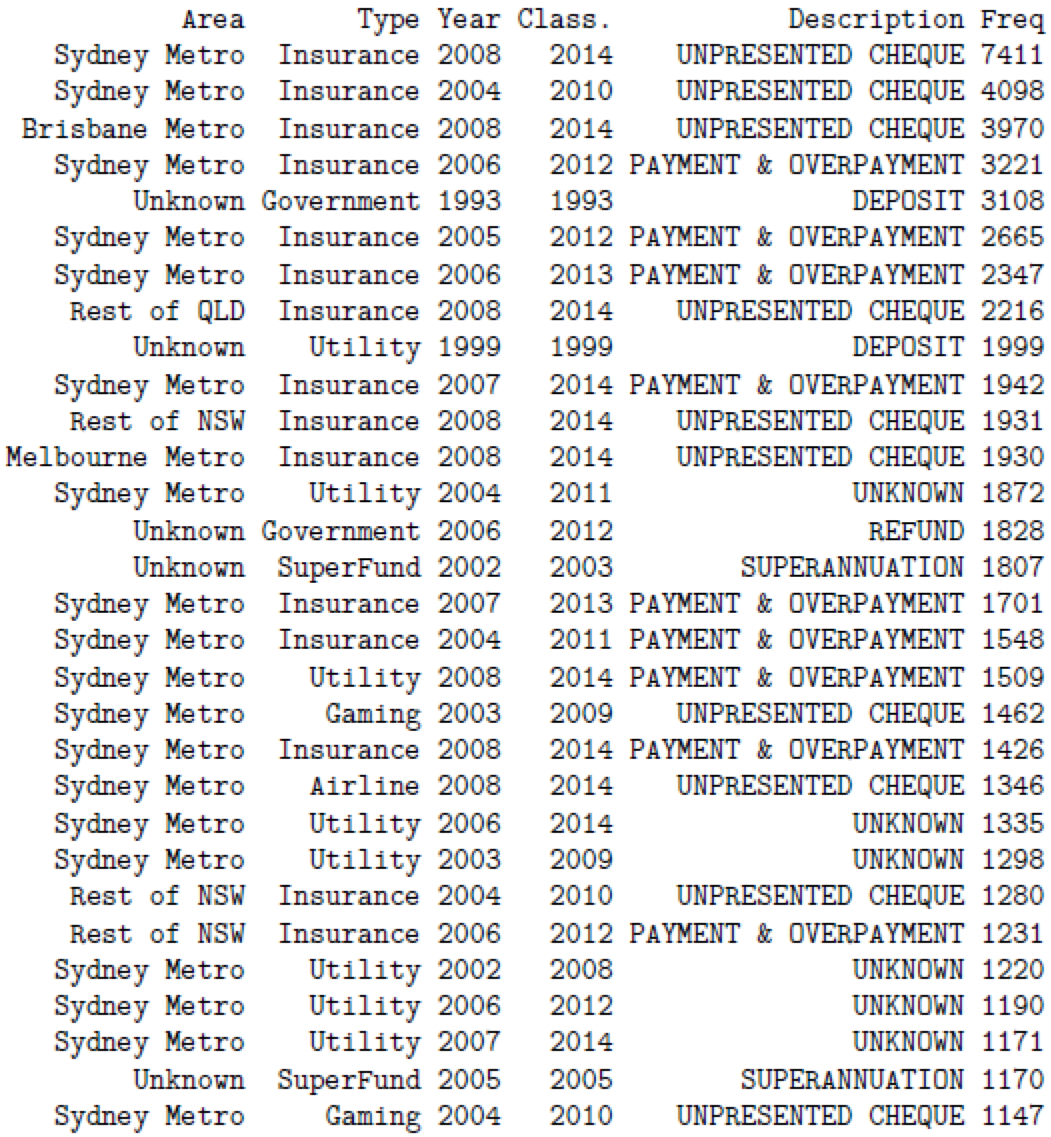
5.1 Analysing the number of accounts with unclaimed money [Anjali]

5.1.1 Tabulating the data

The first problem with analysing the number of accounts is that there is no response variable for this in the data in its current form. To analyse the counts, and which variables correlate with count, we first create a table of frequencies. This table was then sorted by the frequency variable (Freq) in descending order, to find the combination of categories that most occur in the data.

```r
CountsUM = as.data.frame(xtabs(~ Area + Organisation.Type + Year + Classified + Description, data=UM))

SortedCounts = CountsUM[order(-CountsUM$Freq), ]
```

This table shows the frequency of each intersection of the 5 variables, as shown below:



```
          Area       Type Year Class.          Description Freq
  Sydney Metro  Insurance 2008   2014    UNPRESENTED CHEQUE 7411
  Sydney Metro  Insurance 2004   2010    UNPRESENTED CHEQUE 4098
Brisbane Metro  Insurance 2008   2014    UNPRESENTED CHEQUE 3970
  Sydney Metro  Insurance 2006   2012 PAYMENT & OVERPAYMENT 3221
       Unknown Government 1993   1993               DEPOSIT 3108
  Sydney Metro  Insurance 2005   2012 PAYMENT & OVERPAYMENT 2665
  Sydney Metro  Insurance 2006   2013 PAYMENT & OVERPAYMENT 2347
   Rest of QLD  Insurance 2008   2014    UNPRESENTED CHEQUE 2216
       Unknown    Utility 1999   1999               DEPOSIT 1999
  Sydney Metro  Insurance 2007   2014 PAYMENT & OVERPAYMENT 1942
   Rest of NSW  Insurance 2008   2014    UNPRESENTED CHEQUE 1931
Melbourne Metro Insurance 2008   2014    UNPRESENTED CHEQUE 1930
  Sydney Metro    Utility 2004   2011               UNKNOWN 1872
       Unknown Government 2006   2012                REFUND 1828
       Unknown  SuperFund 2002   2003         SUPERANNUATION 1807
  Sydney Metro  Insurance 2007   2013 PAYMENT & OVERPAYMENT 1701
  Sydney Metro  Insurance 2004   2011 PAYMENT & OVERPAYMENT 1548
  Sydney Metro    Utility 2008   2014 PAYMENT & OVERPAYMENT 1509
  Sydney Metro     Gaming 2003   2009    UNPRESENTED CHEQUE 1462
  Sydney Metro  Insurance 2008   2014 PAYMENT & OVERPAYMENT 1426
  Sydney Metro    Airline 2008   2014    UNPRESENTED CHEQUE 1346
  Sydney Metro    Utility 2006   2014               UNKNOWN 1335
  Sydney Metro    Utility 2003   2009               UNKNOWN 1298
   Rest of NSW  Insurance 2004   2010    UNPRESENTED CHEQUE 1280
   Rest of NSW  Insurance 2006   2012 PAYMENT & OVERPAYMENT 1231
  Sydney Metro    Utility 2002   2008               UNKNOWN 1220
  Sydney Metro    Utility 2006   2012               UNKNOWN 1190
  Sydney Metro    Utility 2007   2014               UNKNOWN 1171
       Unknown  SuperFund 2005   2005         SUPERANNUATION 1170
  Sydney Metro     Gaming 2004   2010    UNPRESENTED CHEQUE 1147
```

From this table, it can be seen that unpresented cheques from insurance organisations in Sydney Metro, opened in 2008, and classified as unclaimed in 2014, have the highest number of accounts in the data, 7411 accounts, almost double the amount of the second most frequent category. Insurance organisations in Sydney Metro make up 6 of the top 10 entrues in the table

Aside from metro areas, the rest of Queensland makes up a large chunk of accounts in unpresented cheques issued in 2008, probably due to a series of storms and floods which occurred in the area at that time. These were classified in 2014 as unclaimed.

One mysterious category, that does not fit the general trend of insurance and utility in metro areas, is the one with unknown area, involving deposits from the government that were issued and classified in 1993. This could be a case of mistaken data entry, as there is not enough time within one year to classify a deposit as unclaimed, according to the law. Further investigation of the original data in this category reveals that government organisation issuing these deposits was the Office of Real Estate Services, on the 20th of January 1993, and that no postcodes were recorded, so the area could not be classified in those cases.

On the 29th of October, 1999, Energy Australia issues 1999 deposits, for which the postcodes are unknown, which were classified as unclaimed in 1999, also. The organisation name is listed as “Energy Aust (NCLE)” which indicated that these may be deposits for the Newcastle International Sports Centre sponsored by

Energy Australia in 2001, for which some funding has yet to be claimed.

These kinds of chunks in the data reveal an underlying structure which is not continuous, and probably not suited to GLMs, but at this point, it is too late to try something else.

5.1.2 Modelling with Poisson

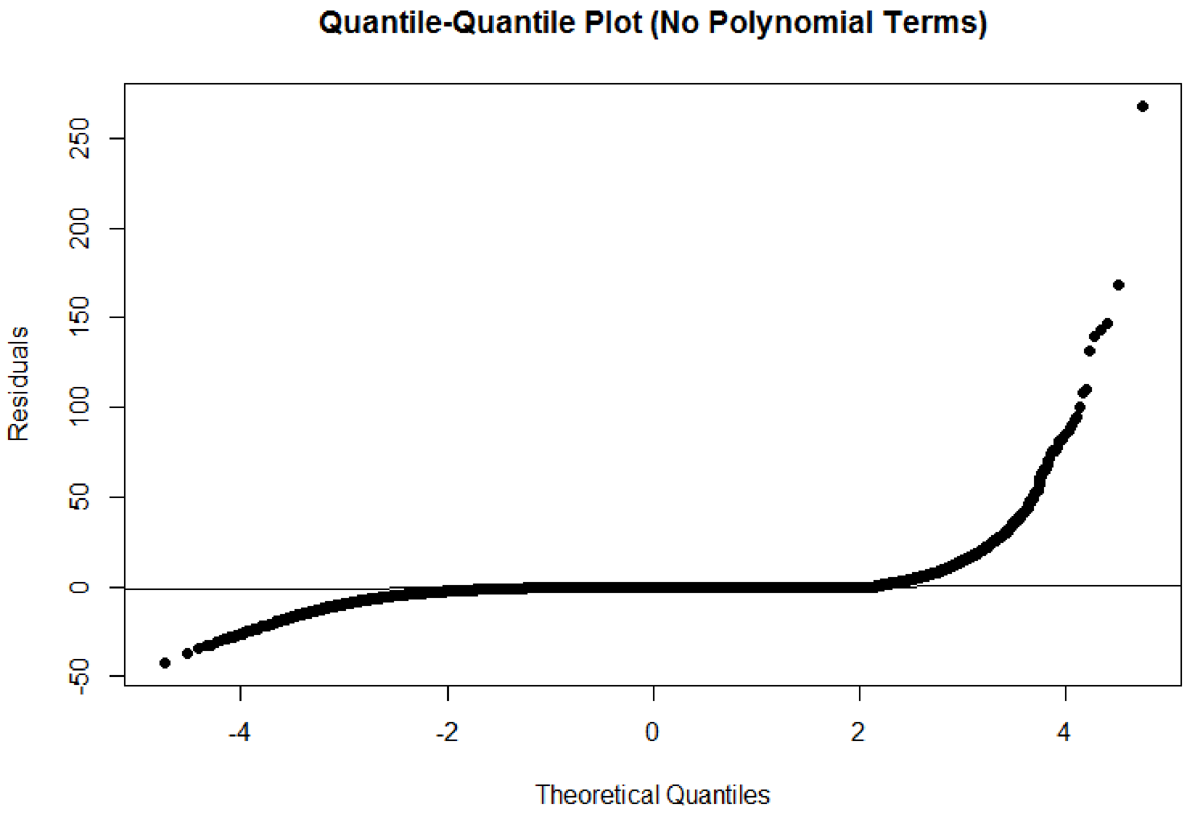
So, we make the 'Year' and 'Classified' variables continuous, to keep the model at a workable size. If each year was treated as a factor, independent of the other years, that could solve part of the problems with underlying structures in our data, but that requires many more parameters and much more RAM than we have access to at the moment.

```
#Using a slightly different counts table to model:
Counts = as.data.frame(xtabs(~Area+Organisation.Type+Year+Classified, data=UML))

Counts$Year = as.numeric(Counts$Year)
Counts$Classified = as.numeric(Counts$Classified)
```

One other thing to be dealt with when modelling is that, at a level of factors, we now have the opposite problem than we had before. Before, in the original data, we had no zero values, since we had no data on those accounts which never contained any unclaimed money. Now, in the frequency table, we have an abundance of zeroes, in those intersections of categories which contain no accounts. We no longer have a need for a zero-truncated Poisson model, and we can use the Poisson model as it is usually used.

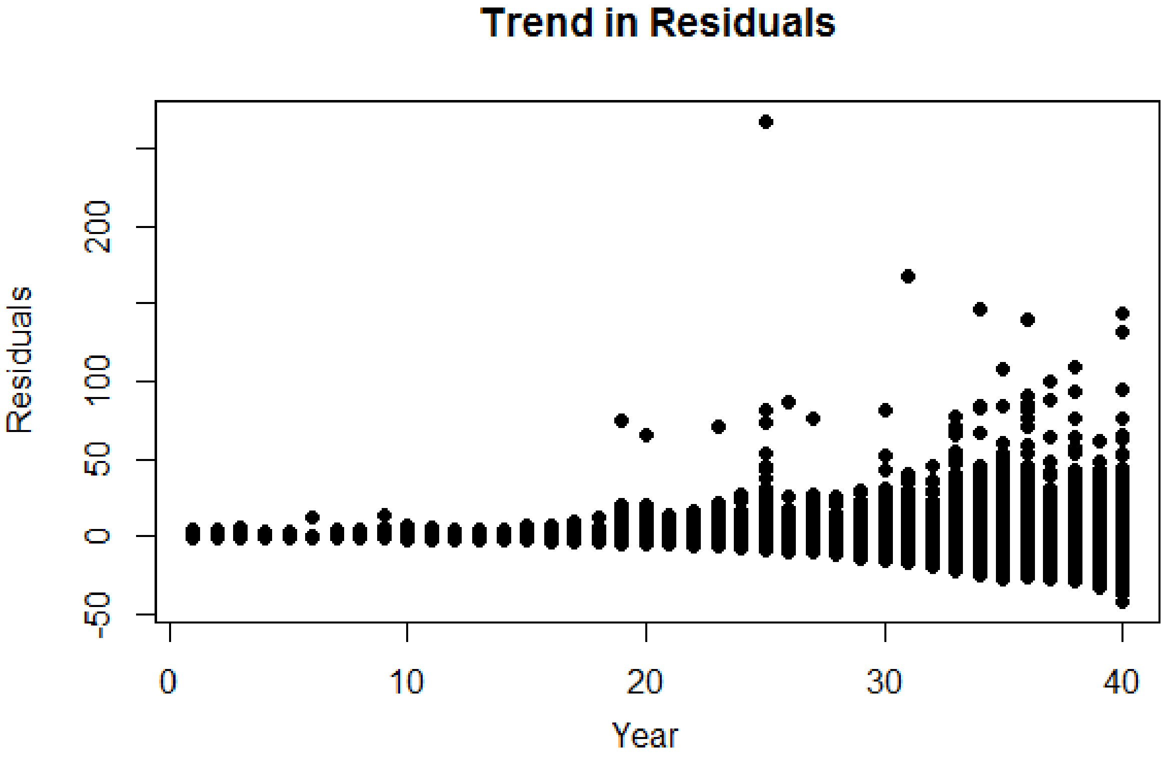
```
PoissonAccounts <- glm(Freq ~ Area+Organisation.Type+Year+Classified,
                       family = poisson(link = "log"), data = Counts)
```

```
qqnorm(residuals(PoissonAccounts), pch=16, ylab="Residuals",
       xlab = "Theoretical Quantiles", main = "Quantile-Quantile Plot (No Polynomial Terms)")
qqline(residuals(PoissonAccounts))
```



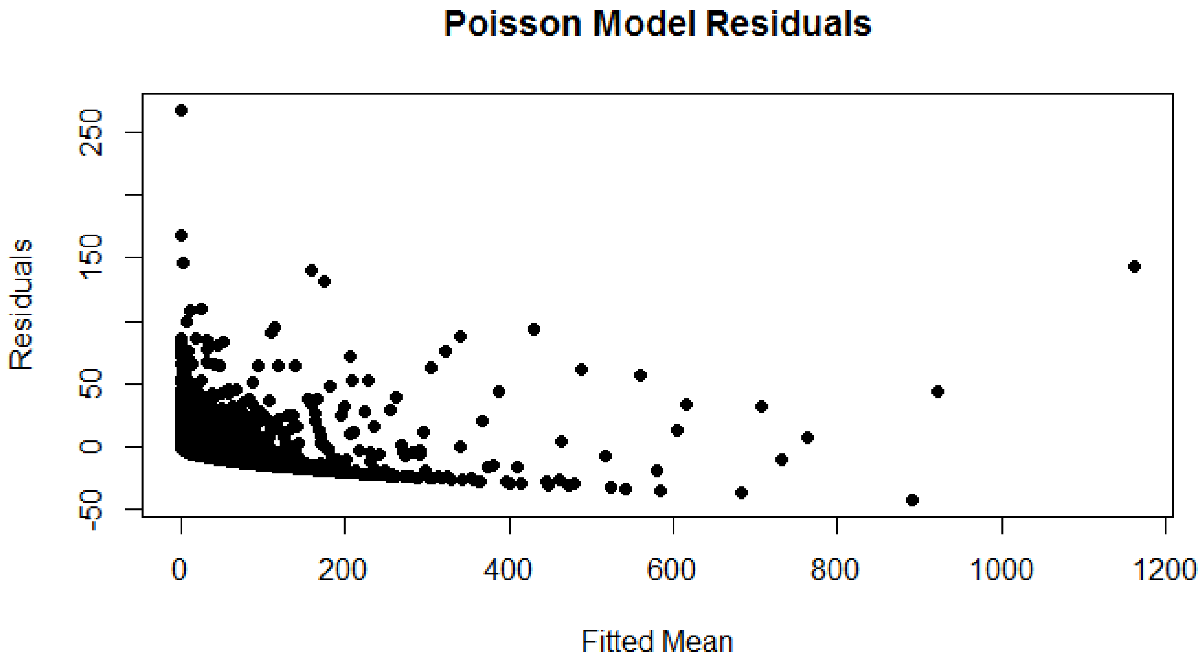
In the plot above, a curvature can be seen in the fit of this model, meaning something significant is affecting the original data that we have not taken into account in this model. In the plot below, the curve looks like it is fanning out as the year increases, showing overdispersion. The downward curve at the bottom of the graph can be attributed to this fanning out also, as it fans out until it hits the minimum a residual can be, which is the negative of the mean.

```r
plot(residuals(PoissonAccounts) ~ Counts$Year, pch=16, ylab="Residuals",
     xlab="Year",  main = "Trend in Residuals")
```



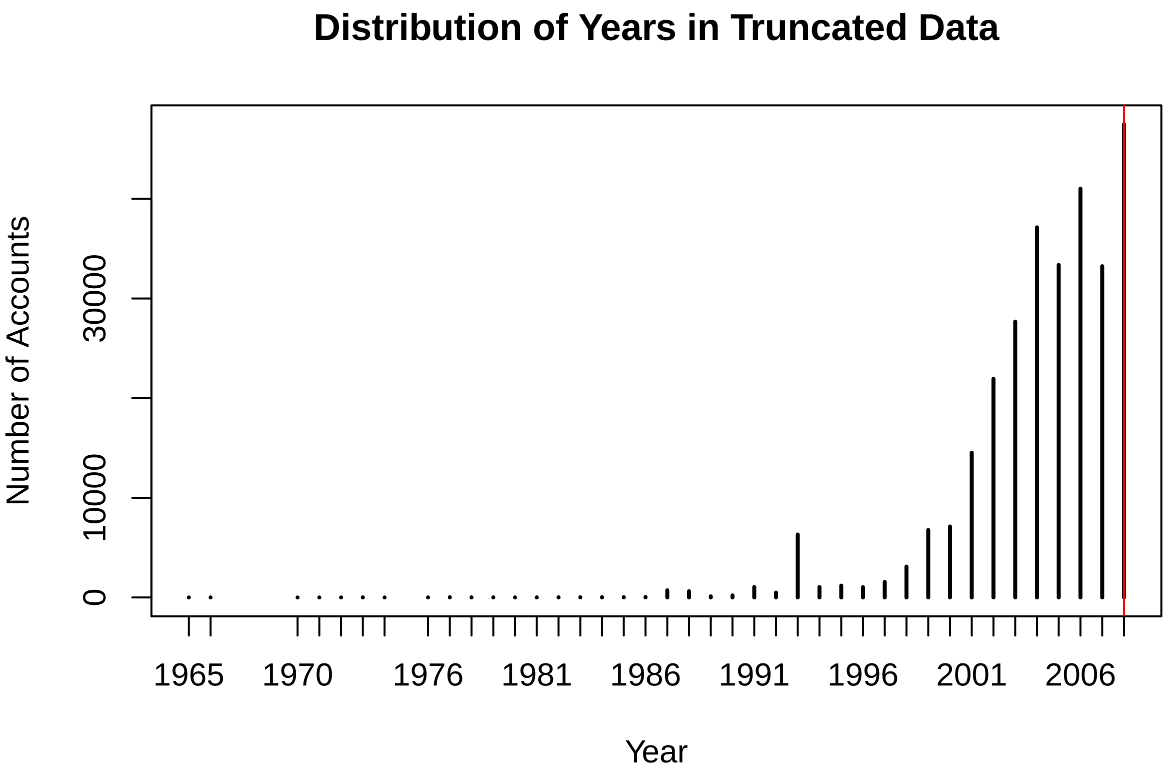
This shows that the curve increases with the variable ‘Year’, indicating that there may be a polynomial relationship between year and the data. This can also be seen in the following plot of the distribution of years after truncating the data.

```
plot(residuals(PoissonAccounts) ~ fitted(PoissonAccounts), pch=16, ylab="Residuals",
     xlab="Fitted Mean",  main = "Poisson Model Residuals")
```



There is a trend in the residuals compared to the fitted means. The residuals decrease as the fitted mean increases.

```
plot(table(UML$Year), ylab = "Number of Accounts", xlab = "Year", main = "Distribution of Years in Trunc
abline(v = 1960, col = "red")
abline(v = 2008, col = "red")
```



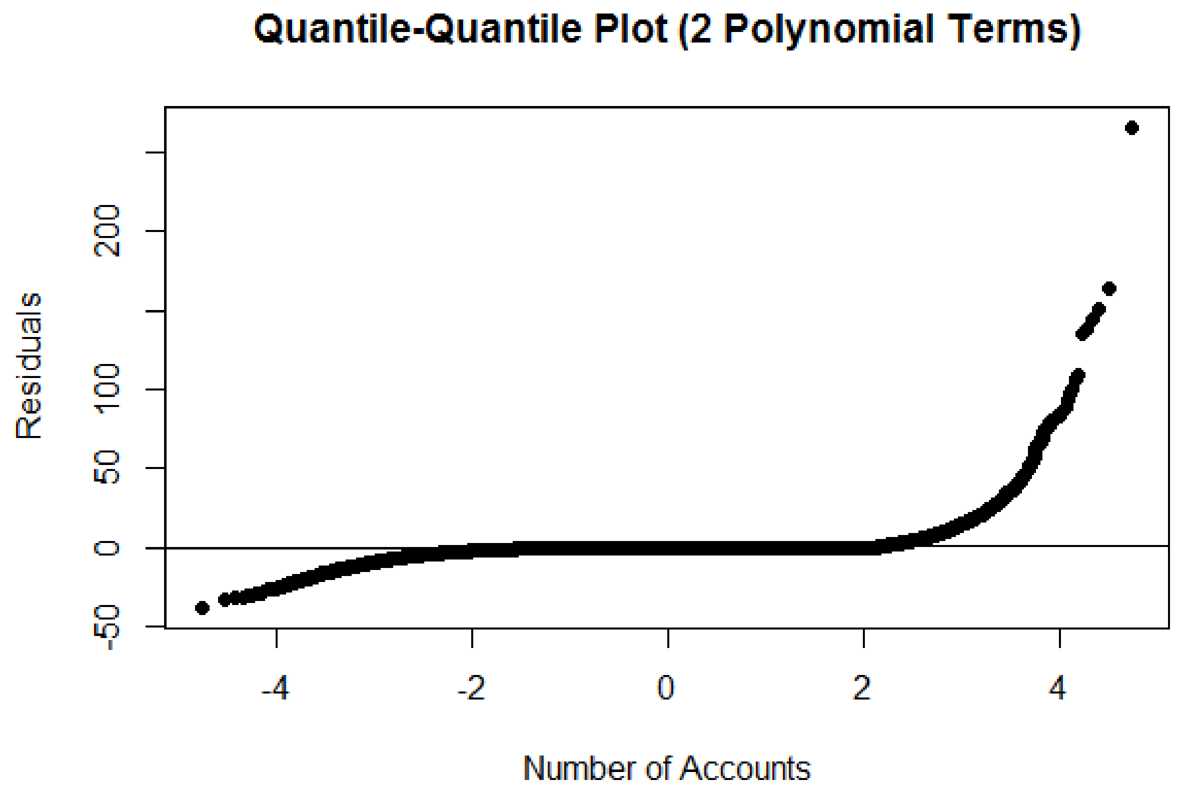
We can still see some problems in the later years where there appears to be dipping from the top, which again suggests some order of polynomials may be needed to accurately fit a model to these detailed structures.

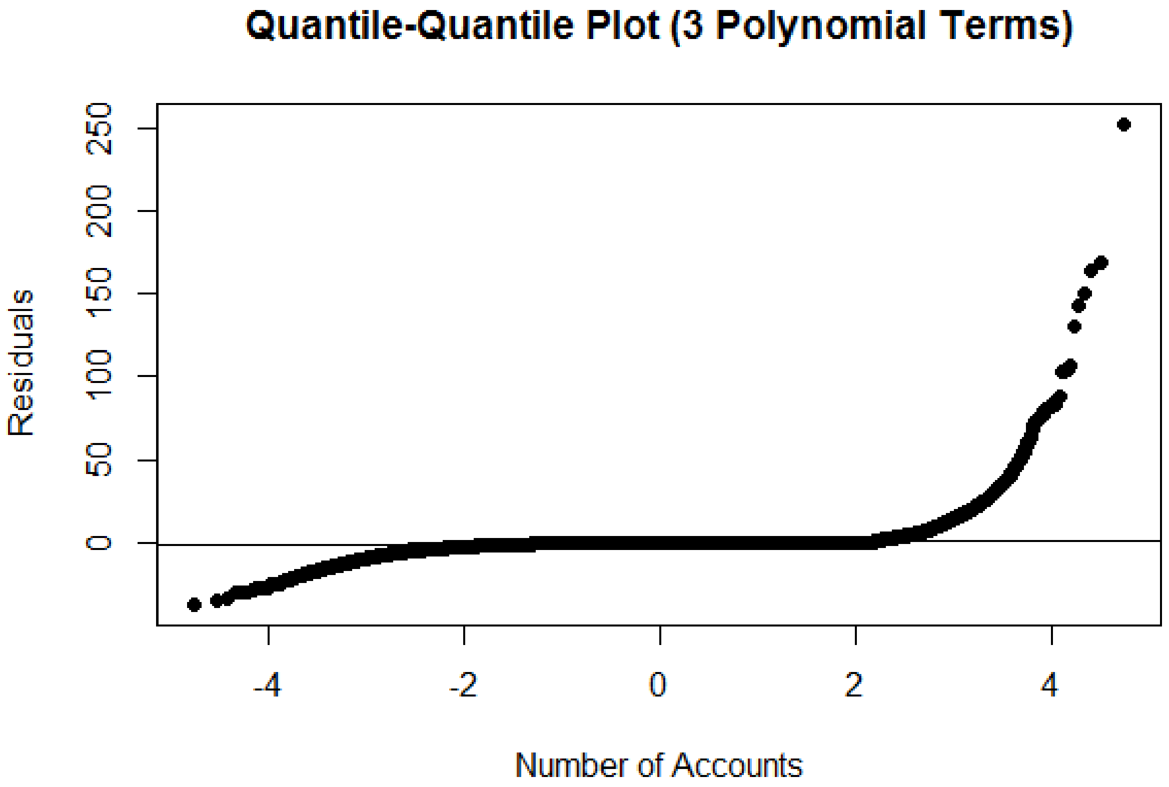
Here are two Poisson models, with different polynomial terms. Each contains the same four variables, and a log link. The difference is that each uses a different power in the polynomial transformation.

```
PoissonAccounts.poly2 <- glm(Freq ~ Area+Organisation.Type+poly(Year,2)+poly(Classified,2),
                             family = poisson(link = "log"), data = Counts)

PoissonAccounts.poly3 <- glm(Freq ~ Area+Organisation.Type+poly(Year,3)+poly(Classified,3),
                             family = poisson(link = "log"), data = Counts)
```

We must plot the residuals of each model for diagnostics.

```
qqnorm(residuals(PoissonAccounts.poly2), pch=16, ylab="Residuals",
       xlab = "Number of Accounts", main = "Quantile-Quantile Plot (2 Polynomial Terms)")
qqline(residuals(PoissonAccounts.poly2))

qqnorm(residuals(PoissonAccounts.poly3), pch=16, ylab="Residuals",
       xlab = "Number of Accounts", main = "Quantile-Quantile Plot (3 Polynomial Terms)")
qqline(residuals(PoissonAccounts.poly3))
```



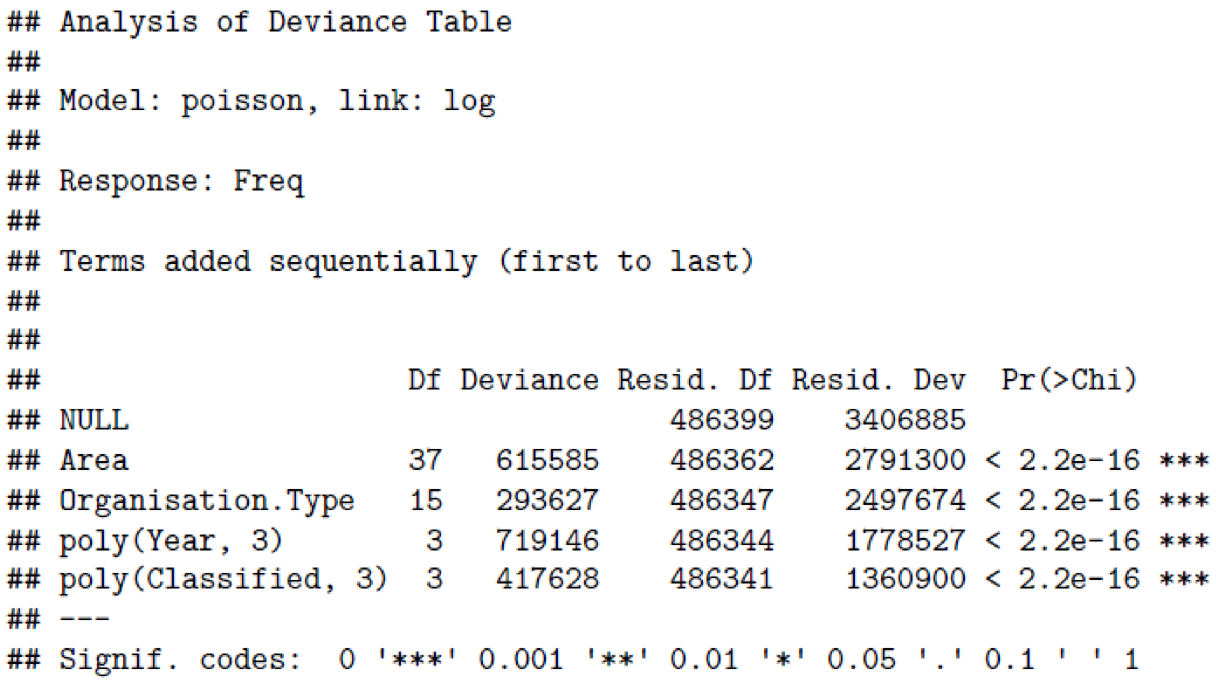


Although these graphs look the same, the polynomial transformation of power 3 is a slightly better fit, both waver at the ends, showing a curvature to the graph that has not been accounted for, either in the variables affecting the original data, or in the model itself.

It should be noted at this point, that the reason the population of each area was not used as a predictor in these models, was because the variable ‘Population’ overlaps with the variable ‘Area’, so any influence population has on the response, the area will have that same influence. The only use of the population data would have been if the right model was found, the population would have given us insight into what the coefficients of each area tell us.

Many other kinds of models were tried, including the Negative Binomial, Normal, Inverse Gaussian, and Anscombe transformations to try to find a fit. Poisson is still the most successful, and an ANOVA will show that despite the bad fit, the variables we have considered are all highly significant.

```
anova(PoissonAccounts.poly3, test = "LRT")
```

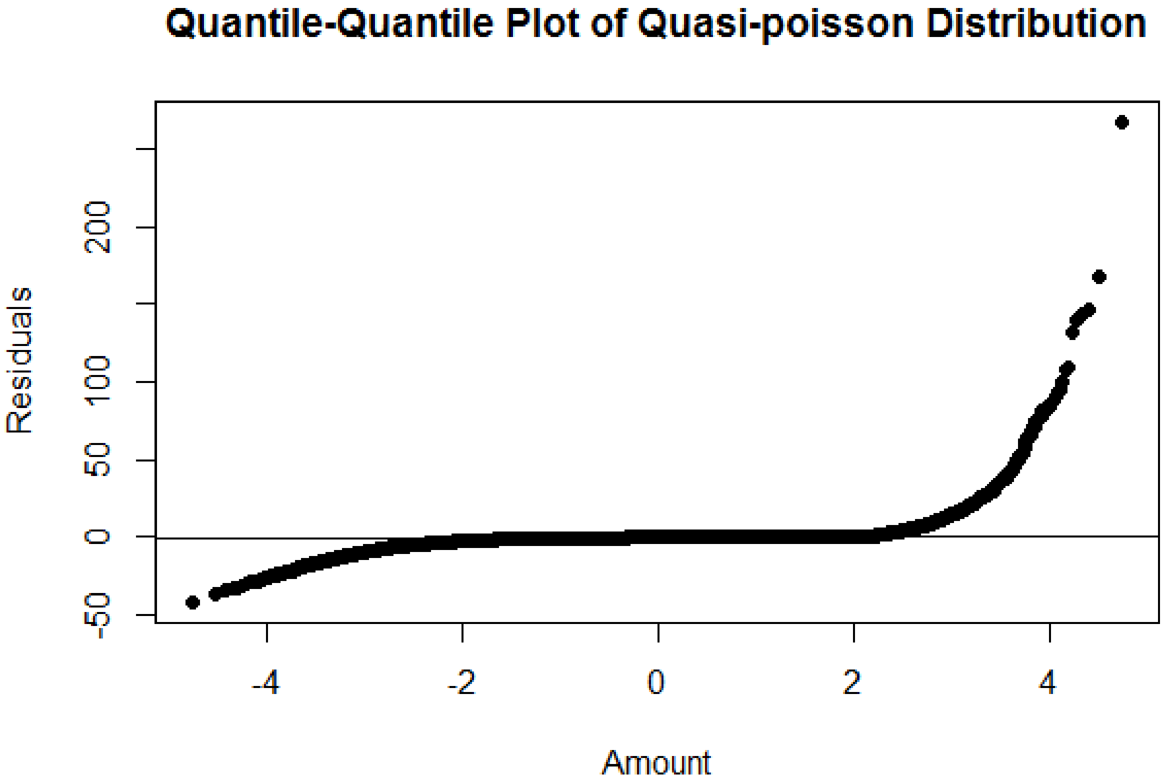


```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Freq
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                              486399    3406885
## Area                37   615585    486362    2791300 < 2.2e-16 ***
## Organisation.Type   15   293627    486347    2497674 < 2.2e-16 ***
## poly(Year, 3)        3   719146    486344    1778527 < 2.2e-16 ***
## poly(Classified, 3)  3   417628    486341    1360900 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The last method we will try to take into account the over dispersion in this model, is a quasi-Poisson model. This model uses a parameter for the dispersion, a parameter which does not exist in the Poisson distribution. There is no formal density that has this mean-variance relationship. Its purpose is to fit models with these previously discussed problems.

```
PoissonAccounts.6 <- glm(Freq ~ Area+Organisation.Type+Year+Classified,
                         family = quasipoisson(link = "log"), data = Counts)

anova(PoissonAccounts.6, test="F")

qqnorm(residuals(PoissonAccounts.6), pch=16, ylab="Residuals", xlab = "Amount",
       main = "Quantile-Quantile Plot of quasi-Poisson Distribution")
qqline(residuals(PoissonAccounts.6))

summary(PoissonAccounts.6)
```

```
## Analysis of Deviance Table
##
## Model: quasipoisson, link: log
##
## Response: Freq
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
## NULL                            486399    3406885
## Area              37   615585    486362    2791300  10.474 < 2.2e-16 ***
## Organisation.Type 15   293627    486347    2497674  12.324 < 2.2e-16 ***
## Year               1   705464    486346    1792209 444.139 < 2.2e-16 ***
## Classified         1   403614    486345    1388595 254.103 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



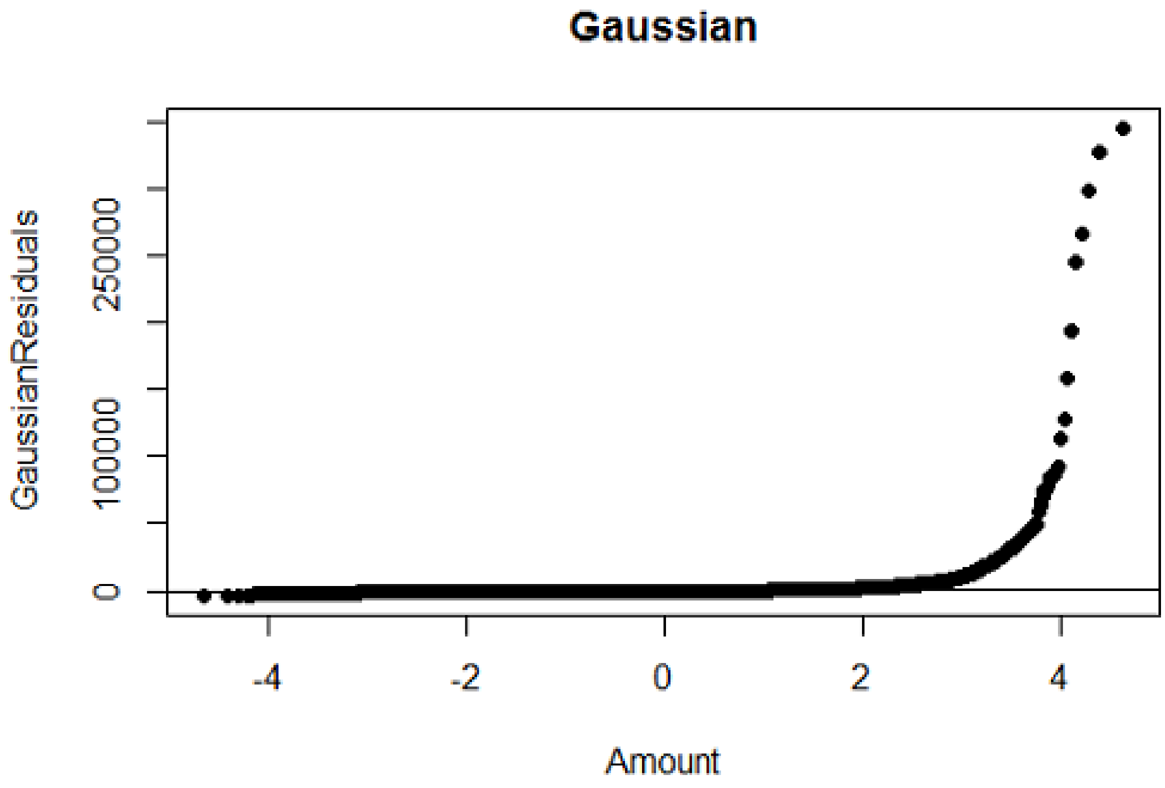
The quasi-Poisson model cannot fix the quantile-quantile plot, nor the residual trends with the fitted mean and year. What it did do, is change the test for the p-value, from a Chi-Square test, to an F test. The F test was not previously possible, because there was no parameter describing the dispersion. In fitting a quasi-Poisson model, we have also estimated the dispersion parameter, which is extremely large. This is seen in the output of "summary(PoissonAccounts.6)", which gives the output "Dispersion parameter for quasipoisson family taken to be 1588.388". This confirms over-dispersion is present in the model.

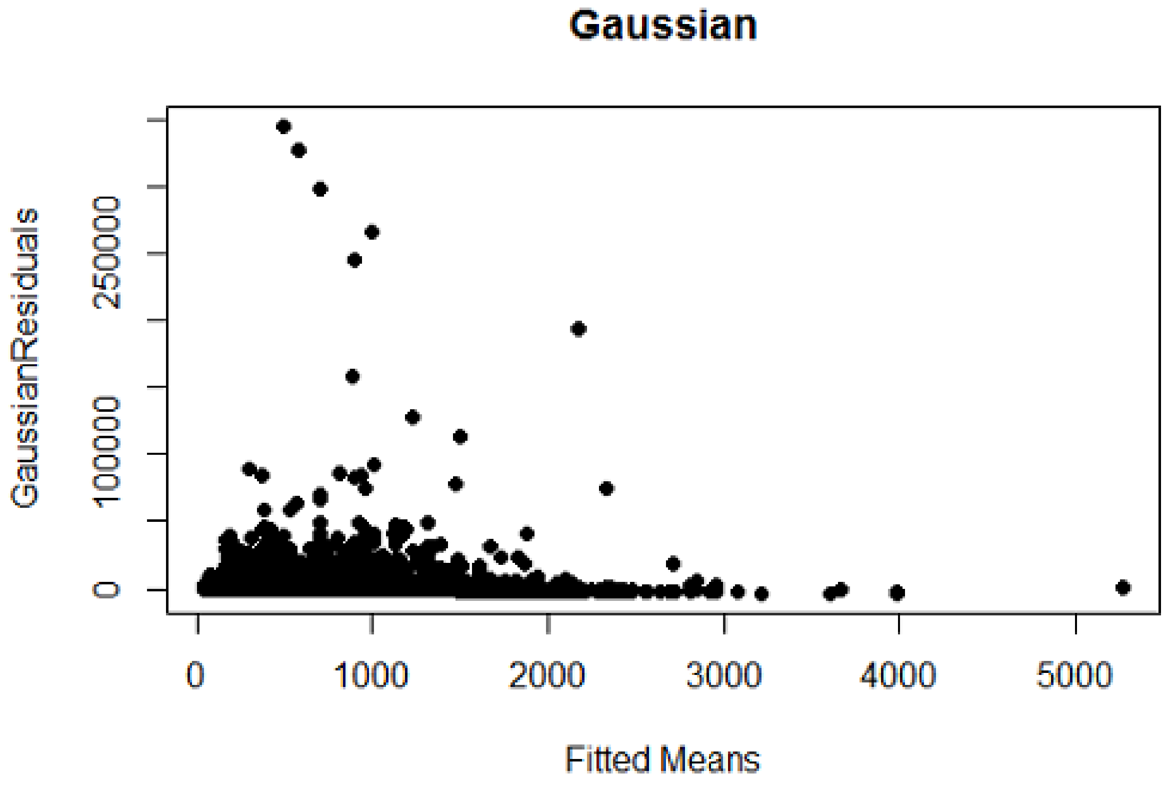
5.2 Analysing the amount within the accounts [Chen]

We will be using the log link function for all our general linear model, since a log link function will always guarantee a positive mean result. When analyzing the amount of money in an account, there will only be positive results.

5.2.1 GLM Gaussian fits with log link function

```
## GLM Gaussian fits with log link function
GaussianAmount = glm(Amount ~ Year + Organisation.Type + Area + Classified,
                     family = gaussian(link = "log"), data=UML)
qqnorm(residuals(GaussianAmount), pch=16, ylab="GaussianResiduals", xlab = "Amount",
       main="Gaussian")
qqline(residuals(GaussianAmount))
plot(residuals(GaussianAmount) ~ fitted(GaussianAmount), pch=16, ylab="GaussianResiduals",
     xlab="Fitted Means", main="Gaussian")
```


Gaussian
GaussianResiduals
0
100000
250000
-4
-2
0
2
4
Amount


Gaussian
GaussianResiduals
0
100000
250000
0
1000
2000
3000
4000
5000
Fitted Means

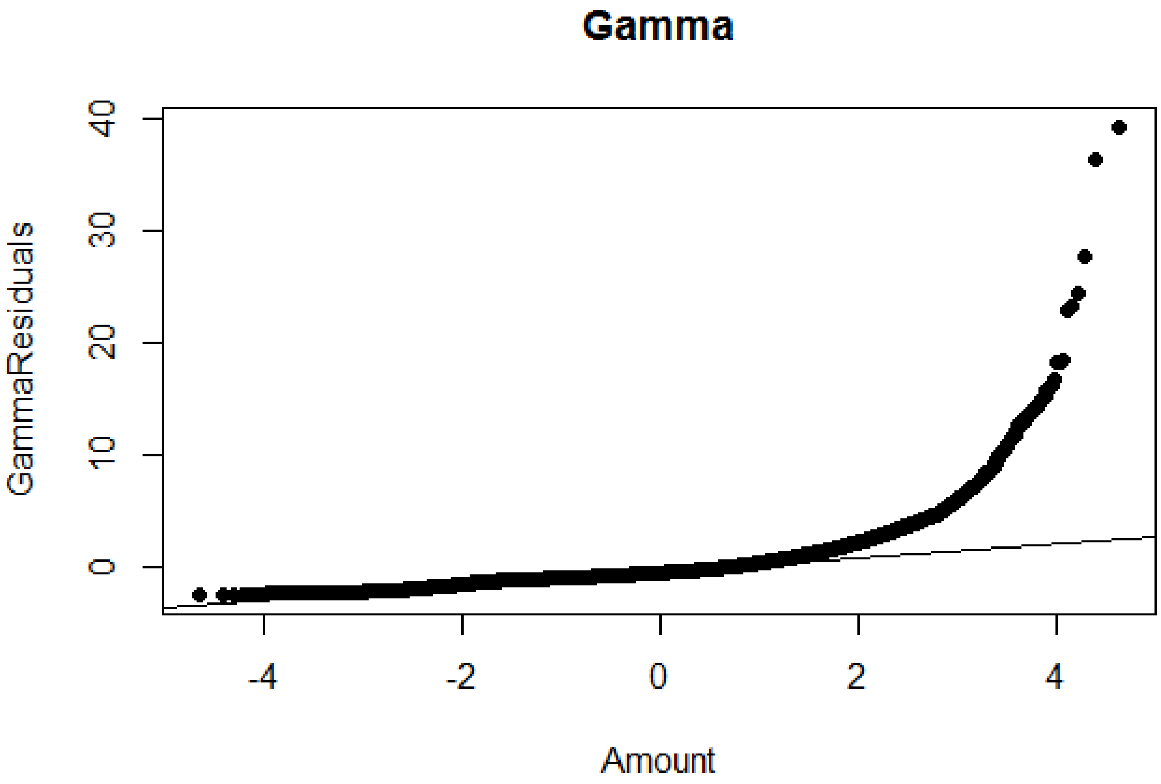
Deviance residuals of GLM model should look approximately Normally distributed regardless of its distribution. However it is not the case for the Gaussian model, which shows heavy skewing to the right, indicating that the residual of Gaussian model is not Normally distributed.

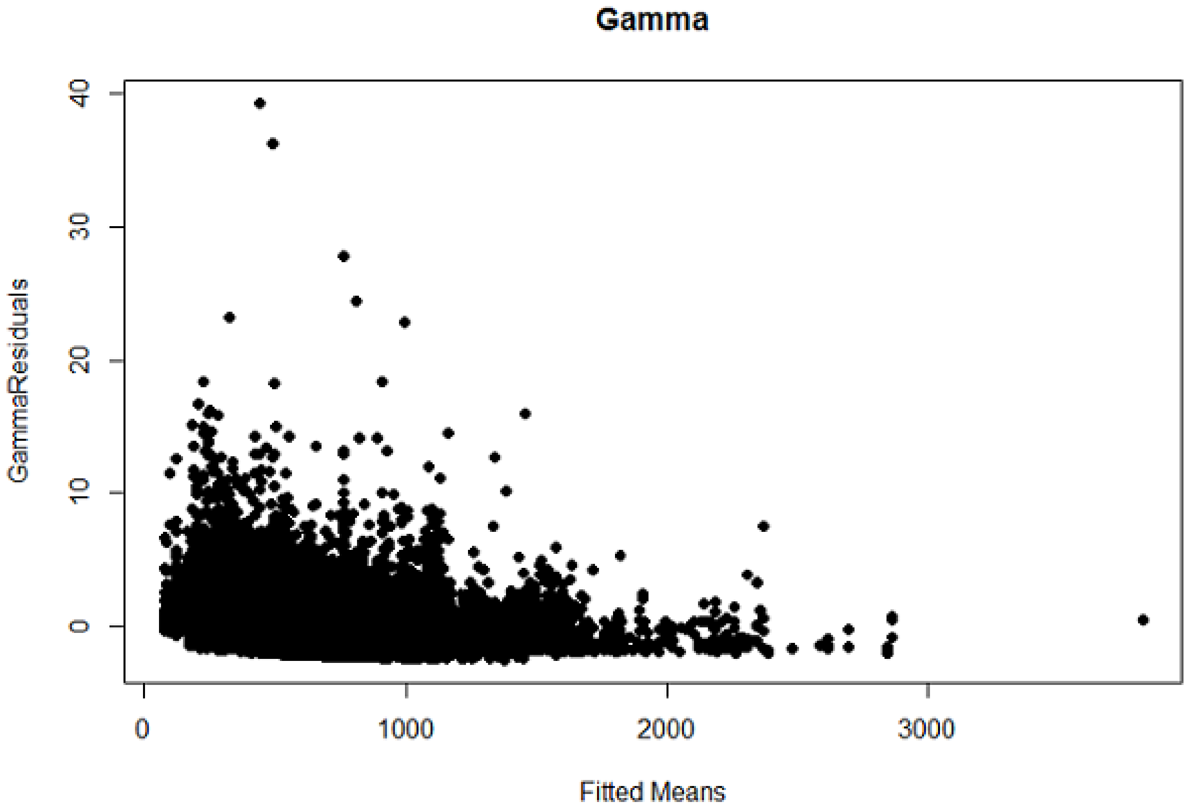
Non-constant variance was another problem for this fit. For an adequate regression analysis,you need to have a constant variance of the error terms, and they must have a mean of zero. The residual of Gaussian model appear to increase then decrease across the fitted values, so this indicate the variance in the error terms has a non-constant variance.

The conclusion is that the Gaussian distribution was not an adequate model, let's try the next distribution in the exponential family, the Gamma distribution.

5.2.2 GLM Gamma fits with log link function

```
## GLM Gamma fits with log link function
GammaAmount = glm(Amount ~ Year + Organisation.Type + Area + Classified ,
                  family = Gamma(link = "log"), data=UML)
qqnorm(residuals(GammaAmount), pch=16, ylab="GammaResiduals", xlab = "Amount",
       main="Gamma")
qqline(residuals(GammaAmount))
plot(residuals(GammaAmount) ~ fitted(GammaAmount), pch=16, ylab="GammaResiduals",
     xlab="Fitted Means", main="Gamma")
```





The residuals of the Gamma model was not Normally distributed. Also, the residual of the Gamma model appear to increase then decrease across the fitted values, therefore this indicate the variance in the error terms has a non-constant variance.

The conclusion was that the Gamma model was not an adequate model either. Next we attempted the inverse Gaussian distribution, unfortunately the inverse Gaussian distribution did not converge.

Although the binomial distribution and Poisson distribution are part of the exponential family, but they are both discrete probability distributions. Hence they were not suited to model the amount of money, which is continuous. After all those attempts, we were only left with the Tweedie distribution in the exponential family for GLM modelling.

5.2.3 GLM Tweedie fits with log link function

The Tweedie family of distributions has a defined variance function but with no closed form probability density function (except in special cases).

Tweedie GLMs require the selection of two parameters. A power in the link function and a power in the variance function.

The link function for Tweedie distribution:

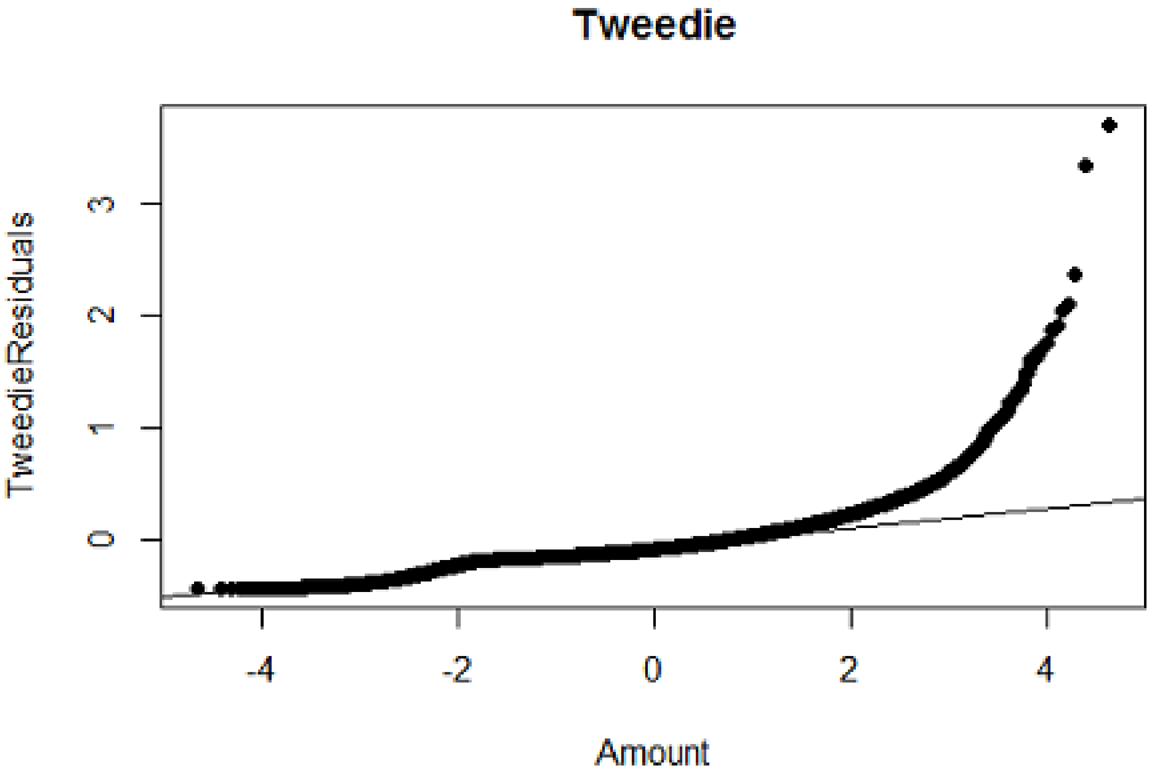
$$\mu^q = X\beta$$

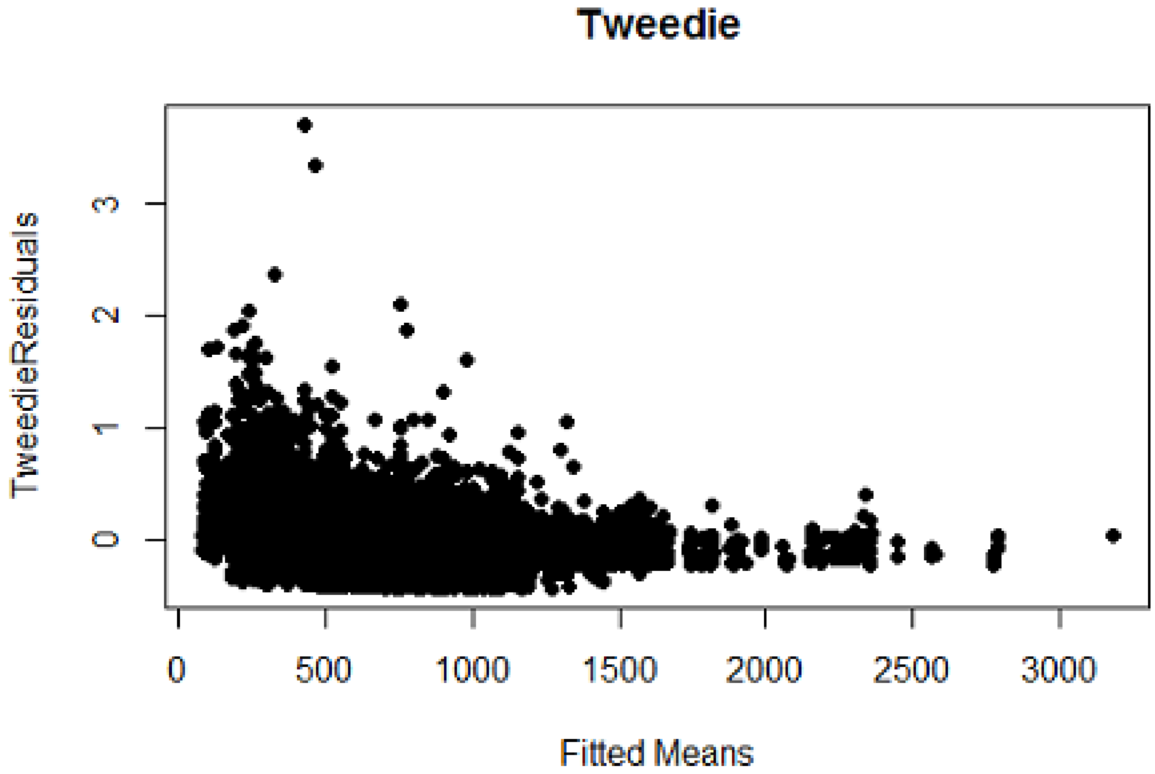
By convention, the log link is power (q) of zero in the Tweedie link function. The variance function for Tweedie distribution:

$$Var(Y) = \mu^p$$

The largest variance power of the Tweedie model for convergence was 2.7.

```
### It requires the statmod package to run the Tweedie GLM
require(statmod)
TweedieAmount = glm(Amount ~ Year + Organisation.Type + Area + Classified
                    ,family=tweedie(link.power=0, var.power=2.7), data=UML)
qqnorm(residuals(TweedieAmount), pch=16, ylab="TweedieResiduals", xlab = "Amount",
       main="Tweedie")
qqline(residuals(TweedieAmount))
plot(residuals(TweedieAmount) ~ fitted(TweedieAmount), pch=16, ylab="TweedieResiduals",
     xlab="Fitted Means", main="Tweedie")
```





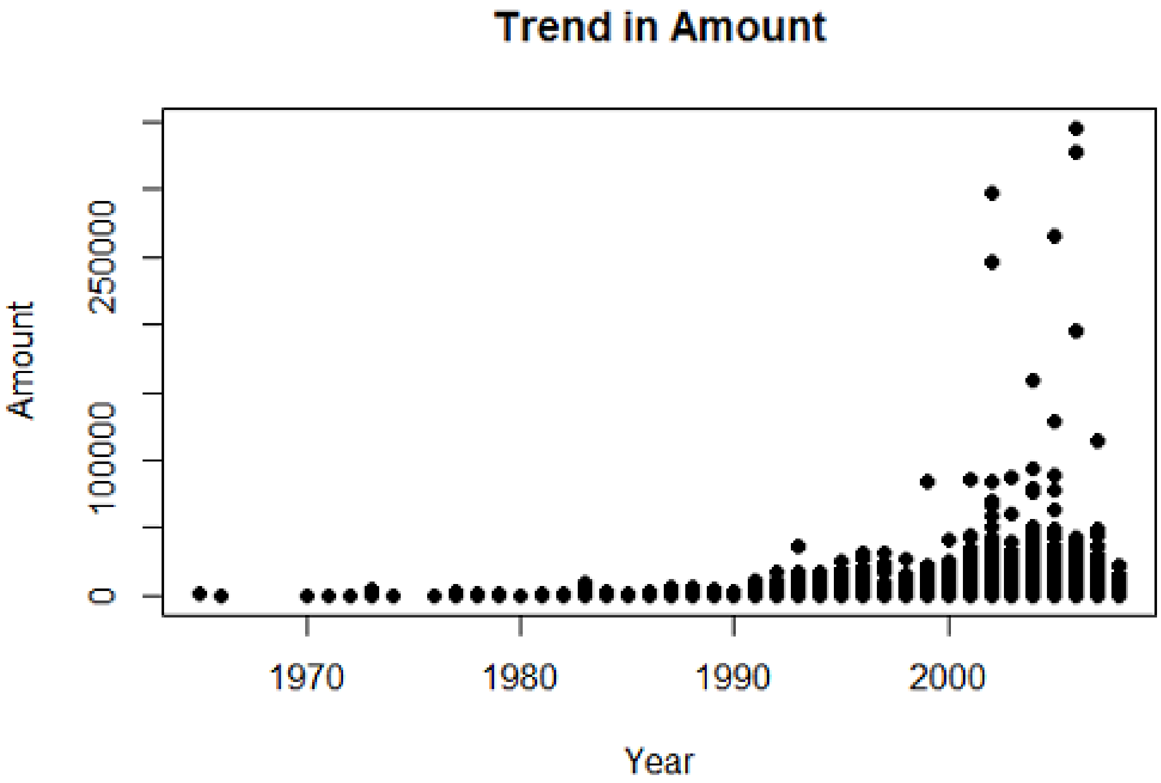
Again the residual of the Tweedie model was not Normally distributed. Also the residual of the Tweedie model appear to decrease across the fitted values, therefore this indicate the variance in the error terms has a non-constant variance.

The conclusion was that the Tweedie model was not an adequate model either. After every distribution within the exponential family was attempted for the GLM model and failed to come up with an adequate model. We conclude that GLM may not be right for this data set. By this stage, it was too late to try something else, except the linear regression on a transformed response.

5.2.4 Simple linear regression

Around the same time we discovered something significant is affecting the original data that we have not taken into account in fitting the model.

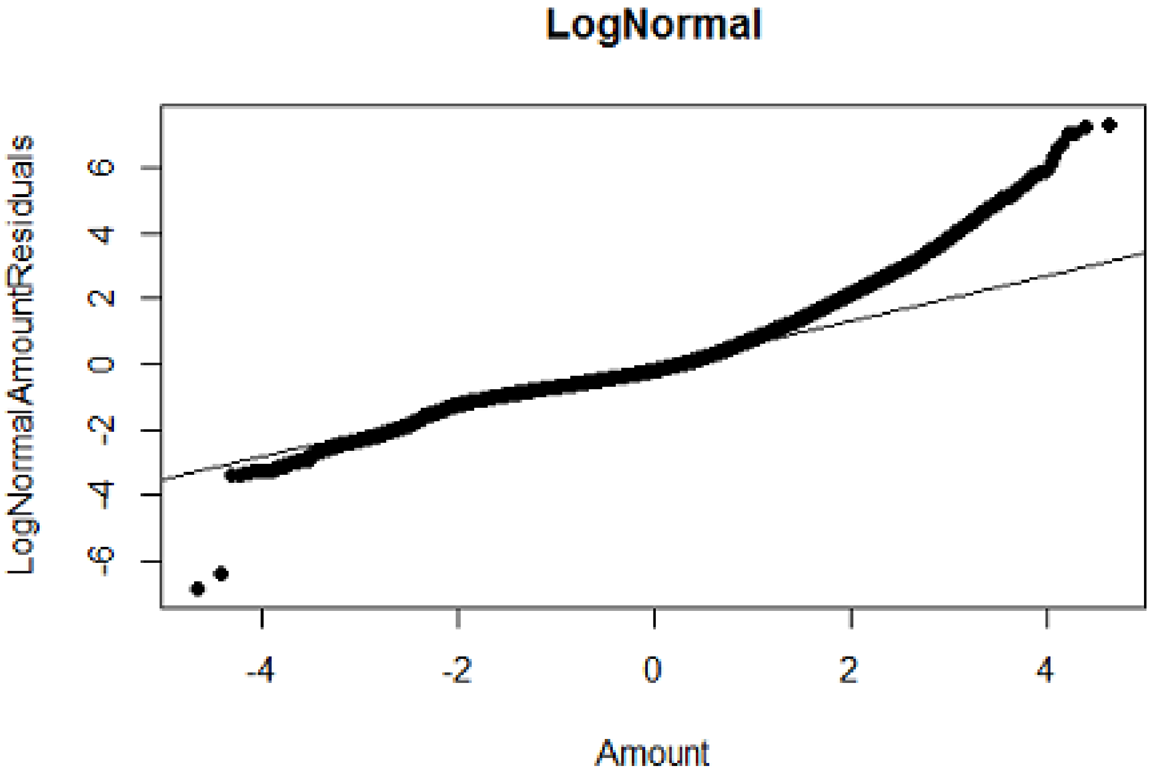
```r
plot(UML$Amount ~ UML$Year, pch=16, ylab="Amount",
     xlab="Year",  main = "Trend in Amount")
```

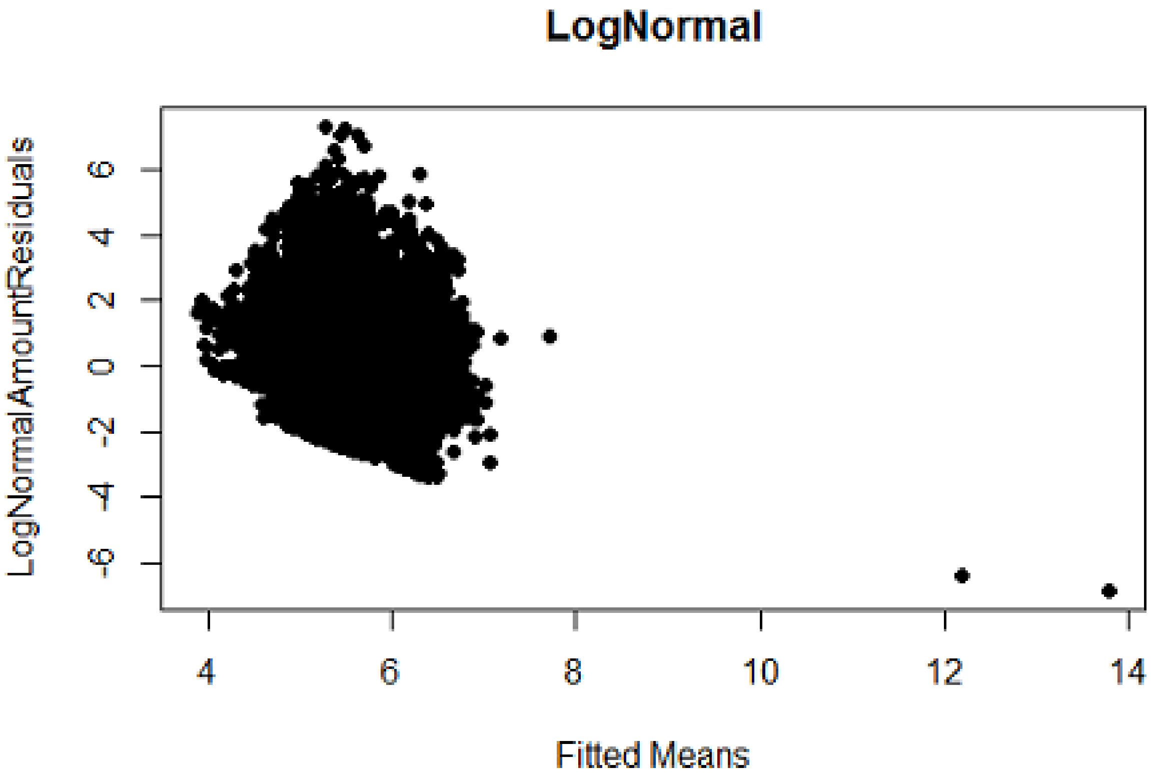


This plot shows that the response ‘Amount’ increases with the variable ‘Year’, indicating that there may be a polynomial relationship between year and the response ‘Amount’. Therefore fitting the Normal model with polynomial transformed variable poly (Year,4) and poly (Classified,4) may help to find an adequate model. The number four is the power of the polynomial transformation, which is the power of polynomial transformation that produced the best result.

We also log transformed the response ‘Amount’, because the transformation may be used to reduce skewness. To reduce right skewness, take the logarithms transformation, which was the problem we having. #We could not use logarithms transformation for all other GLM models due to we have used log link function.#

```
LogNormalAmount = lm(log(Amount) ~ poly(Year,4) + Organisation.Type + Area
                     + poly(Classified,4), data = UML)
qqnorm(residuals(LogNormalAmount), pch=16, ylab="LogNormalAmountResiduals",
       xlab = "Amount", main="LogNormal")
qqline(residuals(LogNormalAmount))
plot(residuals(LogNormalAmount) ~ fitted(LogNormalAmount), pch=16,
     ylab="LogNormalAmountResiduals", xlab="Fitted Means", main="LogNormal")
```


LogNormal
LogNormalAmountResiduals
-6
-4
-2
0
2
4
6
-4
-2
0
2
4
Amount



The log transformation of the response ‘Amount’ with with polynomial transformed variable poly (Year,4) and poly (Classified,4) was the best model we could produce. The residual may still look not Normally distributed, but it was better than any other GLM model. Also there was no trend in the residual, so this indicate variance in the error terms has a constant variance.

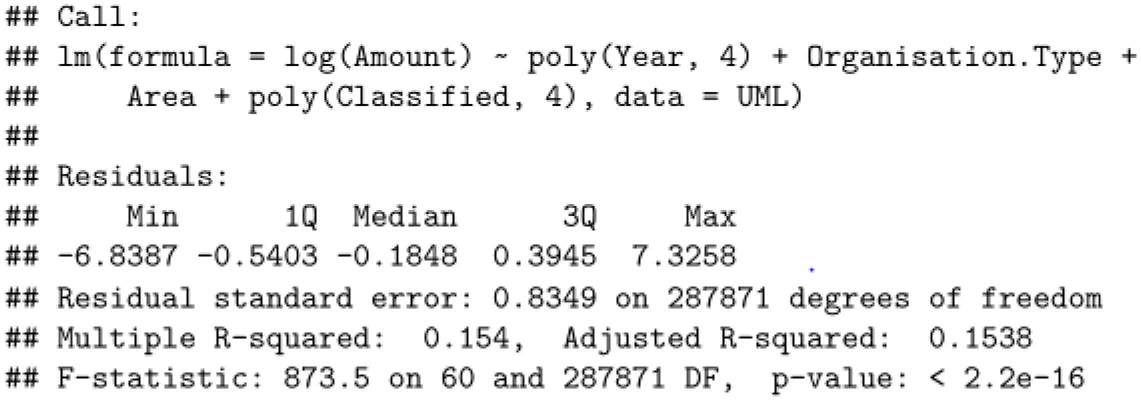
This was still far from a good fitted model. However due to time constraints this was the best we could do. Below is some conclusion base on this model:

```
anova(LogNormalAmount, test = "LRT")
```

```
## Analysis of Variance Table
##
## Response: log(Amount)
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## poly(Year, 4)             4  17117  4279.3 6139.24 < 2.2e-16 ***
## Organisation.Type        15   8094   539.6  774.10 < 2.2e-16 ***
## Area                     37   4312   116.5  167.18 < 2.2e-16 ***
## poly(Classified, 4)       4   7008  1751.9 2513.35 < 2.2e-16 ***
## Residuals            287871 200659     0.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analysis of Variance shows that all variables are significant with almost zero P-value.

```r
summary(LogNormalAmount)
```



```
## Call:
## lm(formula = log(Amount) ~ poly(Year, 4) + Organisation.Type +
##     Area + poly(Classified, 4), data = UML)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8387 -0.5403 -0.1848  0.3945  7.3258
## Residual standard error: 0.8349 on 287871 degrees of freedom
## Multiple R-squared:  0.154,  Adjusted R-squared:  0.1538
## F-statistic: 873.5 on 60 and 287871 DF,  p-value: < 2.2e-16
```

Adjusted R-squared of 0.1538 suggest the model is only explains 15% of the variability of the response data around its mean.

6 Conclusion

All the variables are significant for all models, for both the number of accounts and the amounts in those accounts. Despite this, no adequate model can be found. At this stage, we need to go back to investigate the original data, because there is clearly an underlying structure, particularly within the years, which we have not identified, which is significantly influencing our model. The end of this project, investigating public data using generalised linear models, has left us in the middle of a real world data analysis project. We have learned a lot about the practicality of GLMS, as well as their limits. Continuing on from this point would involve non-parametric methods.

7 References

Abs.gov.au,. '3218.0 - Regional Population Growth, Australia, 2013-14'. N.p., 2015. Web. 12 Aug. 2015.

Apps08.osr.nsw.gov.au,. 'NSW Office Of State Revenue'. N.p., 2015. Web. 30 July 2015. (https://www.apps08.osr.nsw.gov.au/erevenue/ucm/ucm_list.php)

Asx.com.au,. 'Home - Australian Securities Exchange - ASX'. N.p., 2015. Web. 17 Oct. 2015.

Faraway, Julian James. Extending Linear Models With R. Boca Raton, Fla.: Chapman & Hall/CRC, 2006. Print.

McCullagh, P, and John A Nelder. Generalized Linear Models. London: Chapman & Hall, 1994. Print.

Osr.nsw.gov.au,. 'About Unclaimed Money | Office Of State Revenue'. N.p., 2015. Web. 30 July 2015.