robmixglm: An R Package for Robust Analysis Based on Mixtures

Ken J. Beath
Macquarie University

Abstract

A difficulty in statistical analysis is the presence of outliers, or observations which fit poorly to the assumed statistical model. Robust methods may be used to down weight these observations. One method is to assume that the data consists of standard and outlier observations with different distributions and fit these as a mixture model. For generalized linear models the outliers are assumed to be from an overdispersed model, constructed either by including a random effect in the linear predictor or where the distribution includes a scale parameter varying it. The use of the **robmixglm** R package is demonstrated on three examples, demonstrating how outliers may be identified. An advantage of this approach is that it is likelihood based, allowing all the associated methods. This is demonstated using a package for model selection.

Keywords: robust analysis, mixture model, outlier detection, R.

1. Introduction

A common problem in statistics is that some observations, known as outliers, are unusual for the parametric model used and may result in inaccurate parameter estimates as these observations are given excessive weighting. A simple method is to remove the outliers, but this has the effect of biasing the standard errors downwards, as this method may remove observations that are only slightly unusual, but not in fact incorrect, resulting in a reduction in the estimated error variance. A better method is to down weight observations based on their influence (Hampel, Ronchetti, Rousseeuw, and Stahel 1986), for example using the residuals, and this has been applied to a large number of models. A method for generalised linear models (GLM) is described in Cantoni and Ronchetti Cantoni and Ronchetti (2001) and Cantoni Cantoni (2004), which also incorporates an additional weighting to reduce the influence of high leverage points.

An alternative method of performing robust model fitting is to use a heavy-tailed distribution, as this will allow for outlier observations, giving the advantage of allowing likelihood based inference with an associated increase in efficiency. An addditional advantage occurs when the parameters of the distribution describing the shape are estimated, as this avoids the need for prespecified tuning parameters. An early Bayesian approach for linear models was to use a mixture of normals (Box and Tiao 1968; Abraham and Box 1978) with a pre-specified proportion of outliers. Aitkin and Tunnicliffe Wilson (1980) showed how models for mixtures of standard and outlier observations can be fitted using the EM algorithm, with a model allowing the mean to vary between the two groups. (Lange, Little, and Taylor 1989) used a

t distribution in place of a normal distribution, allowing for the degree of contamination by including the degrees of freedom ν as a parameter. Lange and Sinsheimer (1993) compared a number of error distributions including the t-distribution and mixture of normals (contaminated normal) on a number of datasets. While the t-distribution methods have the advantage of a lower number of parameters required, the mixture based methods have the advantage of greater flexibility and allowing for the identification of outliers.

The mixture based approach is difficult for most generalised linear models (McCullagh and Nelder 1989), as the scale parameter is fixed. Beath (2017) extended the mixture based approach to generalized linear models by modifying the outlier class to include in the linear predictor an additional normally distributed random effect, which is therefore an overdispersed generalised linear model as described by Aitkin Aitkin (1996). For distributions which have an estimated scale parameter, for example the negative binomial and normal, the outlier distribution can be specified using a different scale parameter for the standard and outlier components. The **robmixglm** package in R ({R Core Team} 2021) was developed to fit these models and is available at `https://CRAN.R-project.org/package=robmixglm`.

An advantage of the mixture-based approach is that it is model-based, and so standard likelihood procedures may be used. An example of this is model selection, where it is required to reduce the number of predictors to the minimum required to adequately fit the data. This may be performed simply for reasons of parsinonomy, as a simpler model will be easier to understand. It also has the advantage of removing some covariates that are highly correlated. It does have the disadvantage that it may produce a spurious improvement in fit, especially when the number of covariates is large compared to number of observations. There are a number of ways of avoiding this problem, for example dividing the data into training and validation data sets. For a general introduction see James, Witten, Hastie, and Tibshirani (2013, Chapter 6).

For robust linear models using influence functions there are two available methods (Heritier, Cantoni, Copt, and Victoria-Feser 2009, Section 3.4.5): Robust AIC and Robust Mallow's C_p, however these have not been extended to generalized linear models. For robust genearalized linear models the only published method is that of Cantoni and Ronchetti (2001) which uses a robust quasi-deviance. This can then be used to compare models with and without a given predictor, and a stepwise procedure is then performed to obtain the final model. This has the disadvantage that there is no penalty for increasing number of predictors in the model.

The combination of the robust mixture method and packages for model selection allow the use of two main methods: complete subset regression and step wise regression. In complete subset regression models are fitted for all possible subsets of the covariates, then based on some fitting criteria the best is chosen. This has the disadvantage of possibly long execution time, but guaranteeing that the best fitting subset is found. For stepwise regression, starting with a specified model, models of greater or lesser complexity are fitted, with models varying by only one covariate at each step. The best model based on a fitting criteria is chosen and the process repeated. If backward then only smaller models are allowed, for forward larger models and forward/backward both. The disadvantage of this method is that it may not find the best model, but it may be considerably faster.

There are two other packages available in R for robust generalised linear models, both based on influence functions. The function `glmRob()` using routines from Marazzi (1993, Chapter 10) and methods in Kunsch, Stefanski, and Carroll (1989) in package **robust** is restricted to bi-

nomial and Poisson models, and the function `glmrob()` (Cantoni 2004) in package **robustbase** uses the method described in Cantoni and Ronchetti (2001).

The remainder of this paper is organised as follows: Section 2 describes the theory for the models, Section 3 describes the functions, Section 4 contains examples and Section 5 a summary.

2. Model

2.1. Robust Mixture Model

Package **robmixglm** implements the method of Beath (2017). This assumes that data consists of a mixture of two types of observations: standard and outlier, where the standard group consists of subjects from a standard generalised linear model (GLM). The outlier group may be constructed in either of two ways:

1. They are from an overdispersed generalised linear model as described by Aitkin (1996), obtained by incorporating a normally distributed random effect into the linear predictor.

2. For distributions, for example gaussian and negative binomial, where the overdispersion is determined by a scale parameter, the outlier group has greater dispersion through choice of a different scale parameter than the standard group.

The basis of a generalised linear model is that the distribution of the data Y_i, is from the linear exponential family. The relationship between the conditional mean $\mu_i = E[y_i|x_i]$ and the covariates is through the link function $g(\mu_i) = \mathbf{x}_i^T\beta$ (McCullagh and Nelder 1989, p. 27), where $\mathbf{x}_i$ is a vector of covariates for observation i with the first element 1 corresponding to the intercept, and $\mathbf{x}_i^T\beta$ is referred to as the linear predictor. The overdispersed model described by Aitkin (1996) is constructed by including an individual level random effect. For the robust model with class $c_i = 1$ for standard and $c_i = 2$ for outliers, and the normally distributed random effect $\lambda_i \sim N(0, \tau^2)$, the link function is

$$g(\mu_i|c_i, \lambda_i) = \begin{cases} \mathbf{x}_i^T\beta, & c_i = 1 \\ \mathbf{x}_i^T\beta + \lambda_i, & c_i = 2 \end{cases}$$

with the proportion of standard observations and outliers π_1, π_2 respectively, where $\pi_1 + \pi_2 = 1$ and these are assumed constant over $\mathbf{x}_i$. Estimates of the parameters are obtained through a GEM algorithm, with the marginal likelihood for the outlier class obtained by integration over the random effect using Gauss-Hermite quadrature. One advantage of the model is that it is not restricted to GLMs, but can be applied to any model with a linear predictor.

2.2. Outlier Probability

Identification of outliers can be performed using the posterior probability of membership of the outlier class. Given an observed outcome y_i then $f_1(y_i)$ and $f_2(y_i)$ are the values of the density functions for the standard and outlier points respectively, evaluated at the maximum likelihood estimates. Then the probability that the subject is in class 2, the outlier class, is

(McLachlan and Peel 2000, Section 2.8.1):

$$P\left(c_i = 2|y_i\right) = \frac{\hat{\pi}_2 f_2\left(y_i\right)}{\hat{\pi}_1 f_1\left(y_i\right) + \hat{\pi}_2 f_2\left(y_i\right)}$$

2.3. Outlier Test

As the method is likelihood based a test can be performed for the presence of outliers, equivalent to a test that the proportion of outliers is zero, that is $\pi_2 = 0$. A difficulty with this test is that the null hypothesis is for a parameter on the edge of the parameter space, so that the likelihood ratio test no longer has the asymptotic chi-square distribution under the null hypothesis. This requires that the null distribution is simulated, known as the Bootstrap Likelihood Ratio Test (BLRT) (McLachlan 1987) or equivalently a parametric bootstrap (Davison and Hinkley 1997, Section 4.2). The observed test statistic is then compared to the simulated distribution to obtain the p-value for the test.

An alternative to the BLRT is to use an information criteria, which has the advantage of being much faster but is not as reliable as the BLRT. The basis of an information criteria is a function of the log likelihood penalised by the number of parameters in the model. Two information criteria (McLachlan and Peel 2000, SEction 6.8) are available; Akaike's Information Criteria (AIC) where $AIC = -2LL + 2n_{par}$ and Bayesian Information Criteria (BIC) where $BIC = 2LL + \log\left(n_{obs}\right) n_{par}$, where LL is the log likelihood for the fitted model, n_{par} is the number of parameters in the model and n_{obs} is the number of observations. Of the two, BIC has been preferred by a number of authors, for example Fraley and Raftery (1998), for determining the number of components in a mixture model.

3. Description of the functions

The basic function is `robmixglm(formula,family,offset,data)` where the parameters have the same meaning as for the `glm()` function. The parameter `family` is a string describing the error distribution and link for the generalised linear model. Valid families are shown in Table 1.

family	error distn.	link
gausssian	gaussian or normal	identitly
binomial	binomial	logit
poisson	Poisson	log
truncpoisson	truncated Poisson	log
gamma	gamma	log
nbinom	negative binomial	log

Table 1: **robmixglm** Families

Two main methods are supplied: `outlierProbs()`, which extracts the posterior probabilities of being an outlier and has an associated `plot()` method, and `outlierTest()`, which performs a BLRT for the presence of outliers.

4. Examples

4.1. Brain versus Body Weight

This data gives the average brain and body weights for 28 land animals (Rousseeuw and Leroy 1987) which was obtained from the **MASS** package. Of interest is to find if there is a relationship between brain and body mass and any deviations from this relationship. Given the right skewness of the data, it is first log-transformed for both variables.

```
R> library("MASS")
R> data(Animals)
R> Animals$logbrain <- log(Animals$brain)
R> Animals$logbody <- log(Animals$body)
```

First is fitted a standard linear model, and then the robust model. If AIC or BIC are to be used to compare the models, then it is important to use `glm()` rather than `lm()`, as otherwise the log likelihoods are not comparable with those from `robmixglm()`, thus preventing comparison of AIC and BIC between the standard and robust models.

```
R> brainbody.glm <- glm(logbrain~logbody, data=Animals)
R> summary(brainbody.glm)

Call:
glm(formula = logbrain ~ logbody, data = Animals)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2890  -0.6763   0.3316   0.8646   2.5835

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.55490    0.41314   6.184 1.53e-06 ***
logbody      0.49599    0.07817   6.345 1.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 2.345692)

    Null deviance: 155.427  on 27  degrees of freedom
Residual deviance:  60.988  on 26  degrees of freedom
AIC: 107.26

Number of Fisher Scoring iterations: 2

R> brainbody.glm.rob <- robmixglm(logbrain~logbody, data=Animals)
R> summary(brainbody.glm.rob)
```

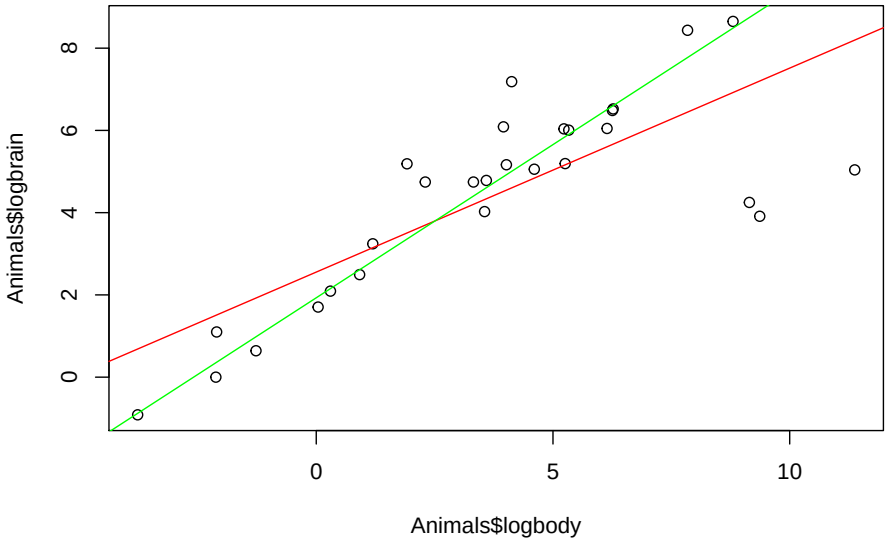

Figure 1: Observed and Fitted for Brain versus Body Weight

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.92968    0.16567   11.65   <2e-16 ***
logbody        0.74495    0.02895   25.73   <2e-16 ***
Outlier p.     0.29842
Sigma-sq       0.14977
Sigma-sq Out. 10.12383
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     logLik      AIC      BIC
  -41.09157 92.18313 98.84415
```

The robust model estimates that there are about 0 % outliers. There is a large decrease in s^2 for the robust model, decreasing from 2.35 down to 0.15 showing the effect of reducing the influence of the outliers. The coefficient `Sigma-sq Out.` gives the error variance for the outlier group, and is much higher than for the standard group. The lines for each fitted model can then be plotted using the `abline()` function as shown in Figure 1. An alternative method is to use the `predict()` method, which allows the predictions to be further transformed.

```
R> plot(Animals$logbody, Animals$logbrain)
R> abline(brainbody.glm, col="red")
R> abline(brainbody.glm.rob, col="green")
```

As a rough guide to which is the appropriate model we can compare AIC and BIC for the two models, which can be extracted as follows using the standard `AIC()` and `BIC()` functions.

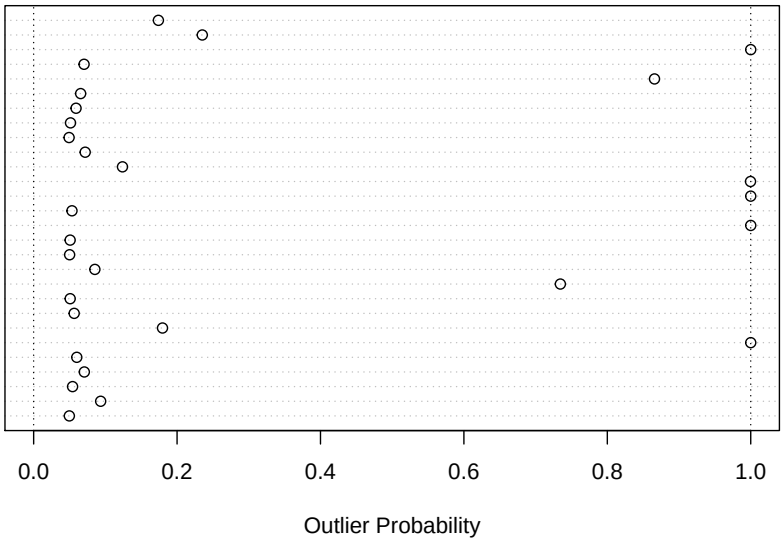

Figure 2: Outlier Probabilities for Brain versus Body Weight

```
R> aitable <- data.frame(model=c("Standard", "Robust"),
+    aic=c(AIC(brainbody.glm), AIC(brainbody.glm.rob)),
+    bic=c(BIC(brainbody.glm), BIC(brainbody.glm.rob)))
R> print(aitable)

     model       aic       bic
1 Standard 107.25779 111.25440
2   Robust  92.18313  98.84415
```

This shows clearly the better fit of the robust model with lower AIC and BIC. The presence of outliers can also be tested using `outlierTest()`, performing a bootstrap likelihood ratio test (BLRT), for a more accurate result than comparing information criteria.

```
R> outlierTest(brainbody.glm.rob)

p value 0.0050
```

This again shows strong evidence that there are outliers present. The outlying observations can be identified by plotting the posterior probability, obtained using `outlierProbs()`, of being in the outlier class against the observation, as shown in Figure 2. As a simple guide, the outliers can be identified as having an outlier probability of greater than 0.9.

```
R> plot(outlierProbs(brainbody.glm.rob))
```

It appears that there are 5 outliers, with a possible other. These can be printed out as follows.

```
R> print(data.frame(Animals,
+    outlierprob=as.numeric(outlierProbs(brainbody.glm.rob)))
+    [outlierProbs(brainbody.glm.rob) > 0.8,])
```

```
                  body  brain logbrain   logbody outlierprob
Dipliodocus   11700.00   50.0 3.912023  9.367344   1.0000000
Human            62.00 1320.0 7.185387  4.127134   0.9999969
Triceratops    9400.00   70.0 4.248495  9.148465   1.0000000
Rhesus monkey     6.80  179.0 5.187386  1.916923   0.9996808
Chimpanzee       52.16  440.0 6.086775  3.954316   0.8658141
Brachiosaurus 87000.00  154.5 5.040194 11.373663   1.0000000
```

The 3 outliers on the lower side of the fitted line are dinosaurs, as would be expected as reptiles usually have smaller brains, and on the high side are humans, rhesus monkeys and possibly chimpanzees, again as would be expected as apes have relatively larger brains. We can produce plots of residual versus fitted for both the the standard and robust models, as shown in Figure 3. With the robust model the outliers are much more obvious. This comes about for two reasons: with the robust model the estimate of the residual variance is much lower and the fitted line is no longer dragged towards the outliers, so the residuals are increased.

```
R> resdata <- data.frame(
+    model=factor(rep(1:2, each=dim(Animals)[1]),
+      labels=c("Standard", "Robust")),
+    fitted=c(fitted(brainbody.glm), fitted(brainbody.glm.rob)),
+    residual=c(residuals(brainbody.glm), residuals(brainbody.glm.rob)))
R> xyplot(residual~fitted|model, data=resdata)
```

4.2. Carrot Damage

This is analysis of an experiment to determine the dose-response for insecticide on carrot fly on carrots conducted at the National Vegetable Research Station Phelps (1982), with the analysis presented in that paper including an offset which will be ignored here. This data has been previosly analysed in Williams (1987) and McCullagh and Nelder (1989), to demonstrate techniques for detecting outliers. Of interest is that observation 14 appears to be an outlier. We obtain the data from the **robustbase** package.

```
R> library(robustbase)
R> data(carrots)
```

Fitting the two models:

```
R> carrots.glm <- glm(cbind(success, total-success)~logdose+factor(block),
+      family="binomial", data=carrots)
R> summary(carrots.glm)
```

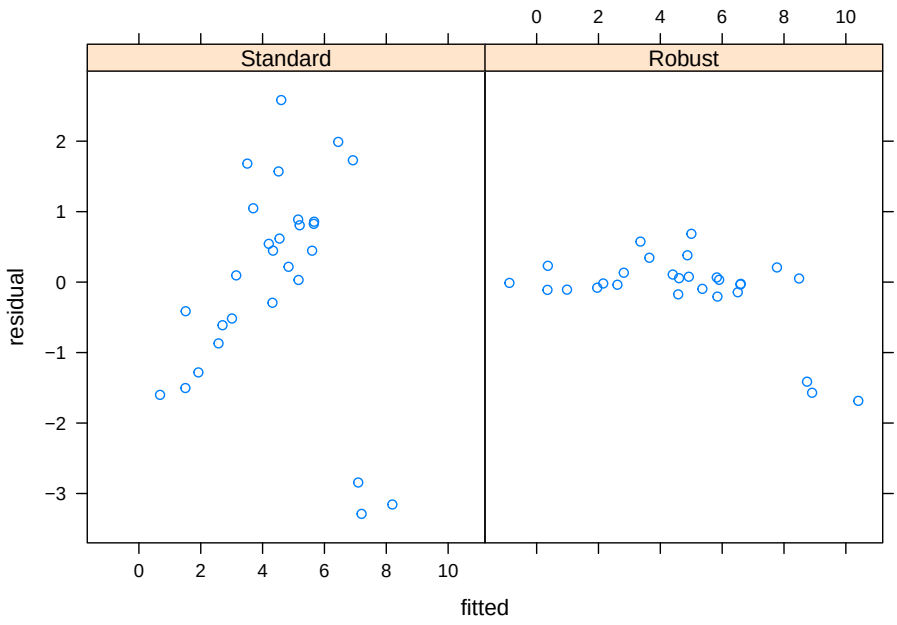

Figure 3: Residual versus Fitted for Brain versus Body Weight

```
Call:
glm(formula = cbind(success, total - success) ~ logdose + factor(block),
    family = "binomial", data = carrots)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9200  -1.0215  -0.3239   1.0602   3.4324

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.0226     0.6501   3.111  0.00186 **
logdose          -1.8174     0.3439  -5.285 1.26e-07 ***
factor(block)B2   0.3009     0.1991   1.511  0.13073
factor(block)B3  -0.5424     0.2318  -2.340  0.01929 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.344  on 23  degrees of freedom
Residual deviance: 39.976  on 20  degrees of freedom
AIC: 128.61

Number of Fisher Scoring iterations: 4
```

```
R> carrots.robustmix <- robmixglm(cbind(success, total-success)~logdose+
+      factor(block), family="binomial", data=carrots)
R> summary(carrots.robustmix)

                Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.4609     0.8367   2.941  0.00327 **
logdose          -2.0632     0.4414  -4.674 2.95e-06 ***
factor(block)B2   0.1765     0.2808   0.628  0.52970
factor(block)B3  -0.5305     0.2709  -1.958  0.05025 .
Outlier p.        0.2482
Tau-sq            0.4509
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     logLik      AIC      BIC
  -57.91094 127.8219 134.8902
```

There are an estimated 0 % outliers. The value of τ^2 of 0.45 indicates the additional overdispersion required to fit the outlier observations. To compare the results of the two models we can extract the coefficients and place them in a table, with standard model results prefixed with "Std" and robust model results with "Rob":

```
R> carrot.results <- data.frame(
+    StdEst=format(summary(carrots.glm)$coefficients[1:4, 1],
+      digits=4),
+    StdSE=format(summary(carrots.glm)$coefficients[1:4, 2],
+      digits=4),
+    Stdp=format.pval(summary(carrots.glm)$coefficients[1:4, 4],
+      digits=4, eps=0.0001),
+    RobEst=format(summary(carrots.robustmix)$coefficients[1:4, 1],
+      digits=4),
+    RobSE=format(summary(carrots.robustmix)$coefficients[1:4, 2],
+      digits=4),
+    Robp=format.pval(summary(carrots.robustmix)$coefficients[1:4, 4],
+      digits=4, eps=0.0001))
R> print(carrot.results, quote=FALSE)

                 StdEst  StdSE     Stdp  RobEst  RobSE     Robp
(Intercept)      2.0226 0.6501 0.001863  2.4609 0.8367 0.003269
logdose         -1.8174 0.3439  < 1e-04 -2.0632 0.4414  < 1e-04
factor(block)B2  0.3009 0.1991 0.130733  0.1765 0.2808 0.529701
factor(block)B3 -0.5424 0.2318 0.019286 -0.5305 0.2709 0.050248
```

Test for outliers and plot the outlier probabilities in Figure 4. This shows clearly that observation 14, with an outlier probability close to one, is the only outlier. However there are a number of observations that have at least a moderate probability of being an outlier.

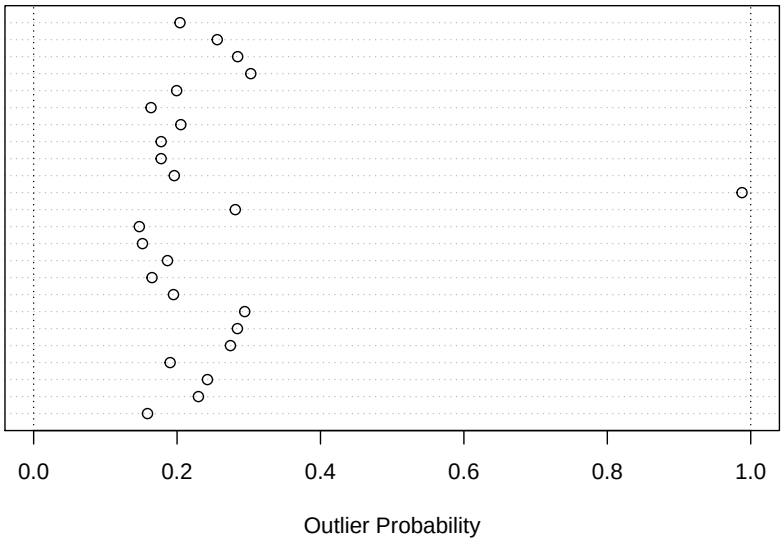

Figure 4: Outlier Probabilities for Carrot Damage

```
R> outlierTest(carrots.robustmix)

p value 0.0050

R> plot(outlierProbs(carrots.robustmix))
```

A plot incorporating the observed and predicted for both models is shown in Figure 5. This shows clearly again that observation 14 is the outlier observation. Observed versus fitted is shown in Figure 6, and shows the outlier and also that there is no systematic variation.

```
R> plot(1:dim(carrots)[1], carrots$success/carrots$total,
+    xlab="Observation", ylab="Proportion")
R> points(1:dim(carrots)[1], fitted(carrots.glm), pch=2, col="red")
R> points(1:dim(carrots)[1], fitted(carrots.robustmix), pch=3, col="blue")
R> legend(20,4,legend=c("Observed","Standard","Robust"),pch=c(1,2,3),col=c("black","red","


R> plot(fitted(carrots.robustmix), carrots$success/carrots$total,
+    xlab="Fitted Proportion", ylab="Observed Proportion")
R> abline(a=0.0, b=1.0, col="red")
```

4.3. Diabetes Data

This data is from a study of the prevalence of cardiovascular risk factors such as obesity and diabetes for African Americans (Willems, Saunders, Hunt, and Schorling 1997), and were

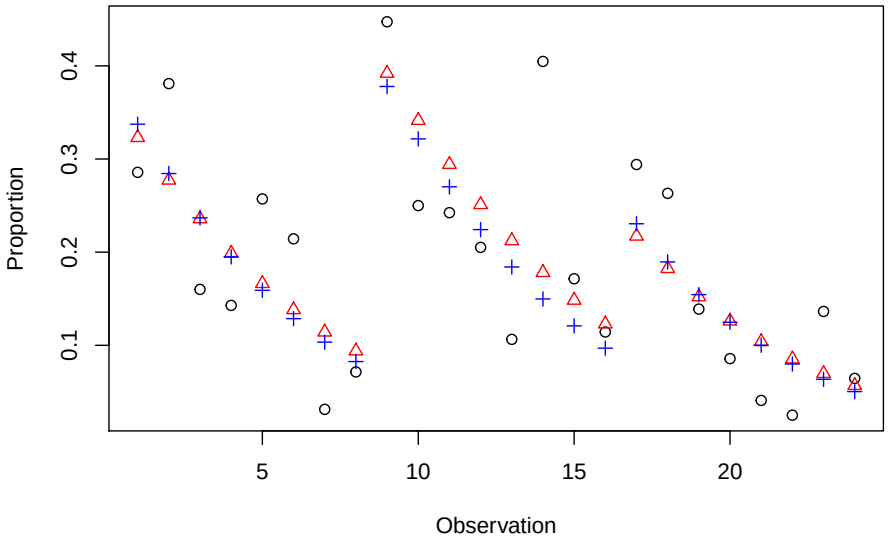

Figure 5: Observed and Fitted for Carrot Damage Models

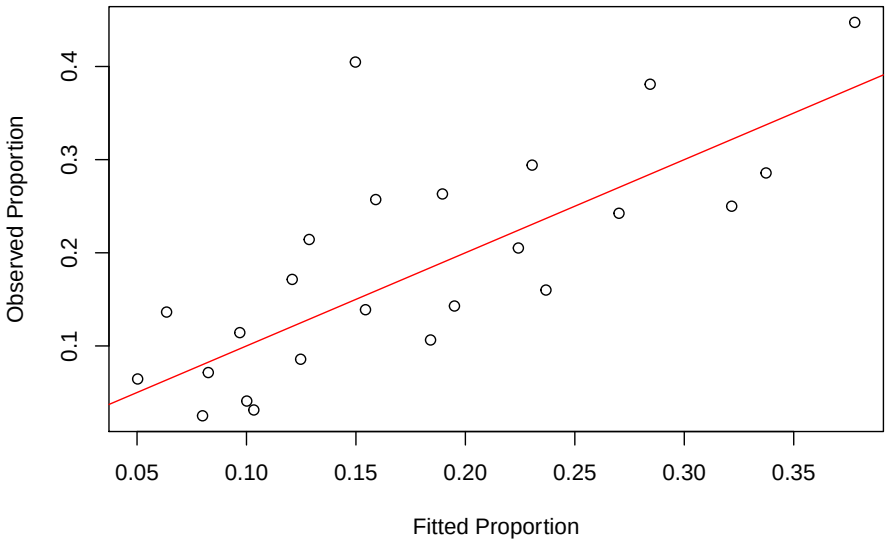

Figure 6: Observed versus Fitted for Carrot Damage

obtained from Heritier *et al.* (2009) and are included in the **robmixglm** package. Data was available for 403 subjects screened for diabetes, reduced to 372 after removal of cases with missing data. The aim is to predict the level of glycosated haemoglobin from the risk factors. It actually appears that a gamma model with log link is better, but I have kept the gaussian to be consistent with Heritier *et al.* It is also quite likely that there is another predictor that is missing from the analysis. Fit the standard and robust models:

```
R> diabdata.glm <- glm(glyhb~age+gender+bmi+waisthip+frame,
+                                  data=diabdata)
R> summary(diabdata.glm)

Call:
glm(formula = glyhb ~ age + gender + bmi + waisthip + frame,
    data = diabdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2195  -1.1379  -0.4676   0.2614  10.2285

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.340044   1.563959  -0.217   0.8280
age          0.041324   0.007136   5.791 1.51e-08 ***
gendermale   0.063536   0.256950   0.247   0.8048
bmi          0.039969   0.019888   2.010   0.0452 *
waisthip     3.163880   1.687404   1.875   0.0616 .
framemedium  0.115422   0.289920   0.398   0.6908
framesmall  -0.049235   0.365635  -0.135   0.8930
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4.307101)

    Null deviance: 1830.8  on 371  degrees of freedom
Residual deviance: 1572.1  on 365  degrees of freedom
AIC: 1607.8

Number of Fisher Scoring iterations: 2

R> diabdata.robustmix <- robmixglm(glyhb~age+gender+bmi+waisthip+frame,
+           data=diabdata)
R> summary(diabdata.robustmix)

               Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.002329   0.559352   5.368 7.98e-08 ***
age            0.013899   0.002585   5.376 7.62e-08 ***
gendermale     0.018244   0.090133   0.202   0.8396
```

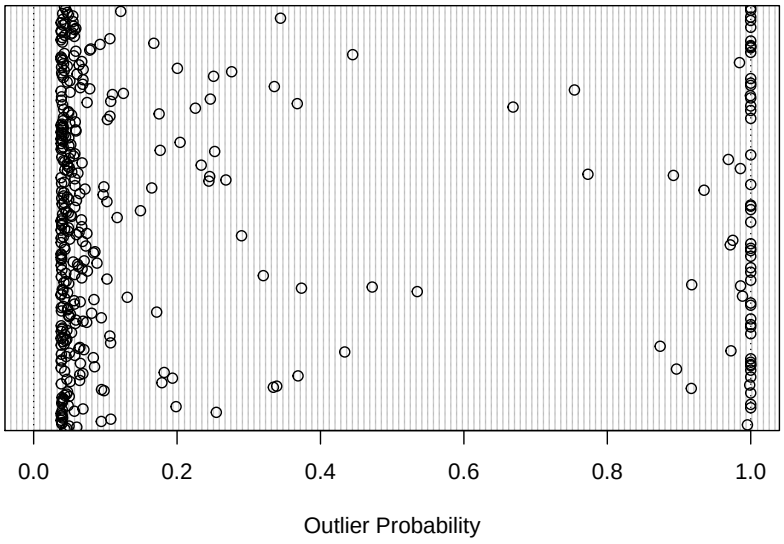

Figure 7: Outlier Probabilities for Diabetic Data

```
bmi             0.010404   0.007077   1.470   0.1415
waisthip        1.056509   0.575061   1.837   0.0662 .
framemedium    -0.052746   0.109855  -0.480   0.6311
framesmall     -0.184365   0.137613  -1.340   0.1803
Outlier p.      0.235690
Sigma-sq        0.340610
Sigma-sq Out. 20.576419
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     logLik      AIC      BIC
  -630.1581 1280.316 1319.505
```

Test for outliers and plot the outlier probabilities in Figure 7, which shows a large number of outliers. This may indicate there is a problem with the model used, possibly an incorrect distribution.

```
R> outlierTest(diabdata.robustmix)

p value 0.0010

R> plot(outlierProbs(diabdata.robustmix))
```

The observed versus fitted may be plotted as in Figure 8. This shows a generally increasing variance at higher predicted values and an increase in the mean, suggesting that there may

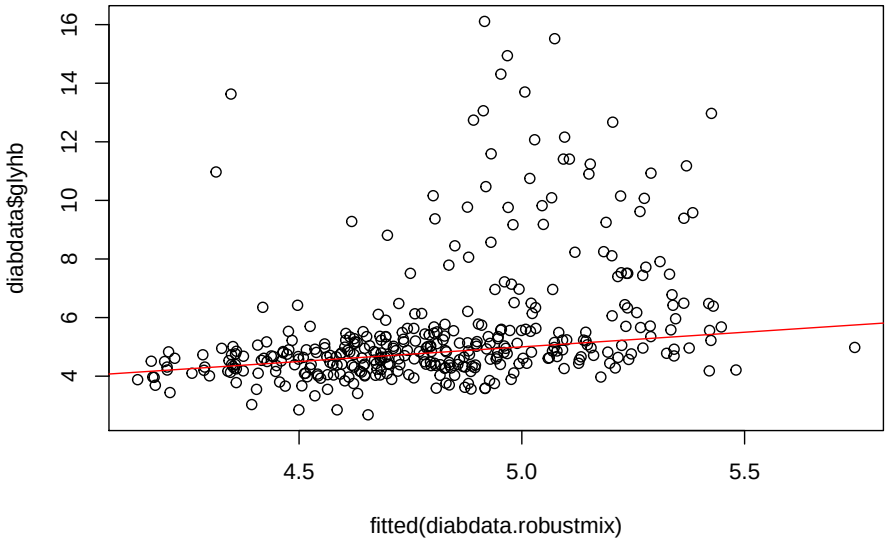

Figure 8: Observed versus Fitted for Diabetic Data using Robust Model

be alternative models, for example a gamma with log link, which may be a better fit to the data.

```
R> plot(fitted(diabdata.robustmix), diabdata$glyhb)
R> abline(a=0.0, b=1.0, col="red")
```

The `step()` function, a simplified version of `stepAIC()` described in Venables and Rispley (1999), allows for stepwise model selection based on the AIC statistic. Here we use the default setting of both backward and forward selection, and start with the full model. The first parameter of the function defines the models to be fitted, and the second defines the terms from which the model is selected. Further parameters are defined in the documentation for `step()`. The function produces a large amount of output, giving the AIC and change for each fitted model, so this has been removed using the `trace=FALSE` parameter. The final model fit is then printed. Note that in practice we would embed this code within a more rigorous analysis, using either a training and test dataset or cross-validation.

```
R> library("MASS")
R> diabdata.step <- step(diabdata.robustmix,
+    glyhb ~ age + gender + bmi + waisthip + frame,
+    trace = FALSE)
R> summary(diabdata.step)

               Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.582896   0.474758   5.440 5.31e-08 ***
age            0.014560   0.002537   5.739 9.52e-09 ***
```

```
bmi            0.015162   0.005771   2.627  0.00861 **
waisthip       1.267877   0.541610   2.341  0.01924 *
Outlier p.     0.236816
Sigma-sq       0.342148
Sigma-sq Out. 20.474058
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     logLik      AIC      BIC
  -631.4526 1276.905 1304.338
```

The resulting model has excluded `frame` and `gender`, and resulted in an increased level of evidence for `bmi` as a consequence of the correlation between it and `frame`.

5. Summary

It has been shown how the **robmixglm** package may be used to fit generalized linear models accounting for outliers. Features of the package allow for identification of outliers and a test for the presence of outliers, as a consequence of the use of a probability model, thus allowing construction of a likelihood function. Further example showed how standard R functions can be used to perform model selection. As the method allows for robustness with and model with a linear predictor then it is possible to extend the range of models fitted. One example is the zero truncated Poisson which is included in the package.

References

Abraham B, Box GEP (1978). "Linear models and spurious observations." *Applied Statistics*, **27**(2), 131–138. doi:10.2307/2346940.

Aitkin M (1996). "A general maximum likelihood analysis of overdispersion in generalized linear models." *Statistics and Computing*, **6**, 251–262. doi:10.1007/BF00140869.

Aitkin M, Tunnicliffe Wilson G (1980). "Mixture models, outliers, and the EM algorithm." *Technometrics*, **22**(3), 325–331. doi:10.2307/1268316.

Beath KJ (2017). "A mixture-based approach to robust analysis of generalised linear models." *Journal of Applied Statistics*, **4763**, 1–13. doi:10.1080/02664763.2017.1414164.

Box G, Tiao G (1968). "A Bayesian approach to some outlier problems." *Biometrika*, **55**(1), 119–129. doi:10.2307/2334456.

Cantoni E (2004). "Analysis of robust quasi-deviances for generalized linear models." *Journal of Statistical Software*, **10**, 1–9. doi:10.18637/jss.v010.i04.

Cantoni E, Ronchetti E (2001). "Robust inference for generalized linear models." *Journal of the American Statistical Association*, **96**(455), 1022–1030. ISSN 0162-1459. doi:10.1198/016214501753209004.

Davison A, Hinkley D (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Fraley C, Raftery AE (1998). "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis." *Computer Journal*, **41**(8), 578–588. doi:10.1093/comjnl/41.8.578.

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York.

Heritier S, Cantoni E, Copt S, Victoria-Feser MP (2009). *Robust Methods in Biostatistics*. John Wiley, Chichester:United Kingdom.

James G, Witten D, Hastie T, Tibshirani R (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York. ISBN 9781461471370.

Kunsch HR, Stefanski LA, Carroll RJ (1989). "Conditionally Unbiased Bounded-Influence Estimation in General Regression Models , With Applications to Generalized Linear Models." *Journal of the American Statistical Association*, **84**(406), 460–466. doi:10.1080/01621459.1989.10478791.

Lange K, Sinsheimer JS (1993). "Normal/Independent Distributions and Their Applications in Robust Regression." *Journal of Computational and Graphical Statistics*, **2**(2), 175–198. doi:10.2307/1390698.

Lange KL, Little RJA, Taylor JMG (1989). "Robust Statistical Modeling Using the t Distribution." *Journal of the American Statistical Association*, **84**(408), 881–896.

Marazzi A (1993). *Algorithms, Routines and S Functions for Robust Statistics*. Chapman and Hall, New York.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. 2nd edition. Chapman & Hall, London.

McLachlan GJ (1987). "On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture." *Applied Statistics*, **36**(3), 318–324. doi:DOI:10.2307/2347790.

McLachlan GJ, Peel D (2000). *Finite Mixture Models*. Wiley, New York.

Phelps K (1982). "Use of the complementary log-log function to describe dose-response relationships in insecticide evaluation field trials." In R Gilchrist (ed.), *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, pp. 155–163. Springer-Verlag, Berlin.

{R Core Team} (2021). "R: A Language and Environment for Statistical Computing." URL https://www.r-project.org/.

Rousseeuw PJ, Leroy AM (1987). *Robust Regression and Outlier Detection*. Wiley.

Venables WN, Rispley BD (1999). *Modern Applied Statistics with S-PLUS*. Third edition. Springer-Verlag, New York.

Willems JP, Saunders JT, Hunt DE, Schorling JB (1997). "Prevalence of coronary heart disease risk factors among rural blacks: A community-based study." *Southern Medical Journal*, **90**, 814–820. doi:10.1097/00007611-199708000-00008.

Williams DA (1987). "Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions." *Journal of the Royal Statistical Society. Series C.*, **36**(2), 181–191. doi:10.2307/2347550.

Affiliation:

Ken J. Beath
Department of Mathematics and Statistics
Faculty of Science and Engineering
Macquarie University NSW 2109
Australia
E-mail: ken.beath@mq.edu.au
URL: http://web.science.mq.edu.au/directory/listing/person.htm?id=kbeath