robmixglm: An R package for robust analysis using mixtures

Ken Beath
Macquarie University
Australia

1 Introduction

1.1 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. The standard group consists of subjects from a standard generalised linear model (GLM), and the outlier group consists of subjects from an overdispersed generalised linear model (Aitkin, 1996) obtained by incorporating a normally distributed random effect into the linear predictor. In a standard generalised linear model we have 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. For the robust model with class $c_i = 1$ for standard and $c_i = 2$ for outliers, and the normally distributed random effect $\lambda_i \sim N\left(0, \tau^2\right)$, 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. One advantage of the modell is that it is not restricted to GLMs, but can be applied to any model with a linear predictor.

1.2 Outlier Probability

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:

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

1.3 Outlier Test

A difficulty with a hypothesis test for the presence of outliers is that the null hypothesis is for a parameter on the edge of the parameter space, that is $\pi_2 = 0$. A consequence is that the likelihood ratio test no longer has the asympptotic 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 a p-value.

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, Chapter 6) are available Akaike's Information Criteria (AIC) where $AIC = -2LL + 2n_{par}$ and Bayesian Information Criteria (BIC) where $BIC = 2LL + \log(n_{obs})\, n_{par}$, 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.

1.4 `robmixglm` function

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

Table 1: `robmixglm` Families

2 Brain versus Body Weight

This data comprises the average brain and body weights for 28 land animals (Rousseeuw and Leroy, 1987). Of interest is to find if there is a relationship between brain and body mass and any deviations from this relationship. The

data are obtained from the MASS package. Given the right skewness of the data, it is first log-transformed for both variables.

```
> library(MASS)
> data(Animals)
> Animals$logbrain <- log(Animals$brain)
> 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 fitted models.

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

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

            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
Tau-sq       9.97408
```

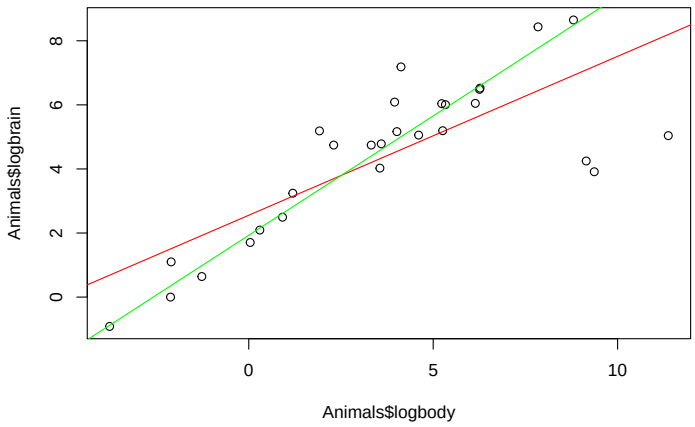

Figure 1: Observed and Fitted for Brain versus Body Weight

```
Sigma-sq     0.14977
---
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 30% outliers.Comparing AIC these are lower for the robust model, indicating a better fit. There is a large decrease in s^2 for the robust model, decreasing from $1.532^2 = 2.347$ down to 0.14977, with a corresponding increase in the value of the test statistics. The lines for each fitted model can then be plotted as shown in Figure 1.

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

```
> 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)))
> print(aitable)

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

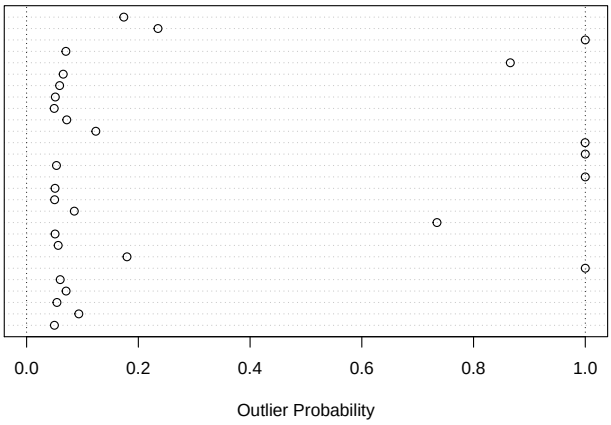

Figure 2: Outlier Probabilities for Brain versus Body Weight

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.

```
> outlierTest(brainbody.glm.rob, showProgress=FALSE)

p value 0.0050
```

This again shows clearly that there are outliers present. The outlying observations can be identified by plotting the posterior probability of being in the outlier class against the observation, as shown in Figure 2. Outliers can be identified as having an outlier probability of greater than 0.9.

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

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

```
> 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
```

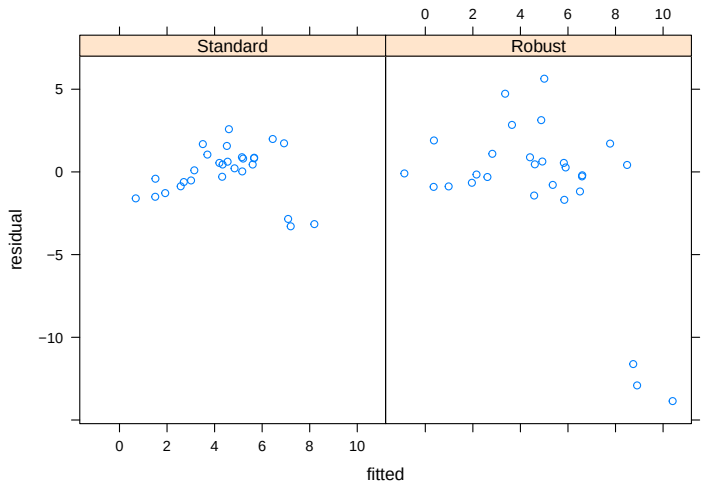

Figure 3: Residual versus Fitted for Brain versus Body Weight

```
Rhesus monkey     6.80  179.0 5.187386  1.916923   0.9996808
Chimpanzee       52.16  440.0 6.086775  3.954316   0.8658167
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 retiles usually have smaller brains, and on the high side are humans, rhesus monkeys and possibly chimpanzees, again as would be expected as apes have 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.

```
> 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)))
> xyplot(residual~fitted|model, data=resdata)
```

3 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). The analysis presented in that paper included an offset which

will be ignored here. Of interest is that observation 14 appears to be an outlier. This data has been previosly analysed in Williams (1987) and McCullagh and Nelder (1989), to demonstrate techniques for detecting outliers. We obtain the data from the `robustbase` package.

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

Fitting the two models:

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

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
```

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

               Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.4609     0.8372   2.940  0.00329 **
logdose          -2.0632     0.4416  -4.672 2.99e-06 ***
factor(block)B2   0.1765     0.2808   0.628  0.52971
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
```

To compare the results of the two models we can extract the coefficients and place them in a table:

```
> 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))
> print(carrot.results, quote=FALSE)

                 StdEst  StdSE     Stdp  RobEst  RobSE     Robp
(Intercept)      2.0226 0.6501 0.001863  2.4609 0.8372 0.003286
logdose         -1.8174 0.3439  < 1e-04 -2.0632 0.4416  < 1e-04
factor(block)B2  0.3009 0.1991 0.130733  0.1765 0.2808 0.529715
factor(block)B3 -0.5424 0.2318 0.019286 -0.5305 0.2709 0.050249
```

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.

```
> outlierTest(carrots.robustmix, showProgress=FALSE)

p value 0.0040

> 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. This shows the outlier and also that there is no systematic variation.

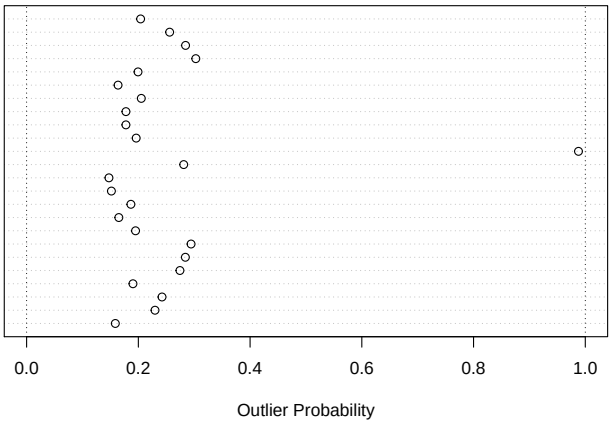

Figure 4: Outlier Probabilities for Carrot Damage

```
> plot(1:dim(carrots)[1], carrots$success/carrots$total,
+   xlab="Observation", ylab="Proportion")
> points(1:dim(carrots)[1], fitted(carrots.glm), pch=2, col="red")
> points(1:dim(carrots)[1], fitted(carrots.robustmix), pch=3, col="blue")
```

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

4 Diabetes Data

This data was from a study of the prevalence of cardiovascular risk factors such as obesity and diabetes for African Americans (Willems et al., 1997). The data are from Heritier et al. (2009), and are slightly modified from Harrell (2015). Data was available for 403 subjects screened for diabetes, reduced to 372 after removal of cases with missing data. The data are part of the `robmixglm` package. Fit the standard and robust models:

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

Call:
glm(formula = glyhb ~ age + gender + bmi + waisthip + frame,
```

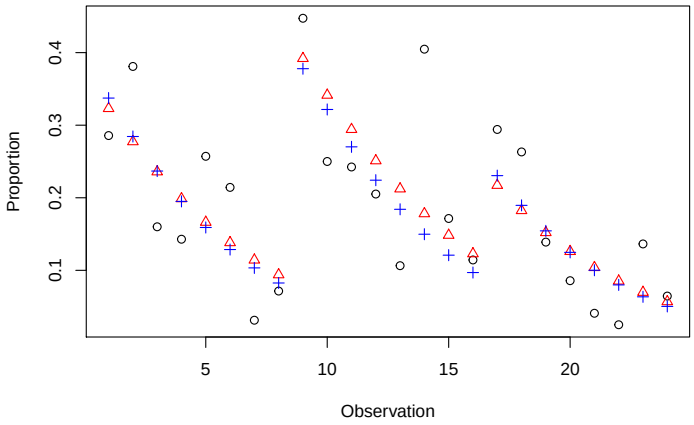

Figure 5: Observed and Fitted for Carrot Damage Models

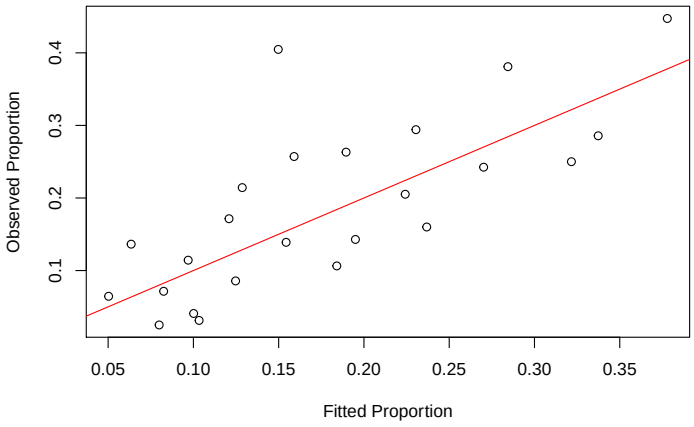

Figure 6: Observed versus Fitted for Carrot Damage

```
    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

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

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.002330   0.559404   5.367 8.01e-08 ***
age          0.013899   0.002585   5.376 7.63e-08 ***
gendermale   0.018244   0.090133   0.202   0.8396
bmi          0.010404   0.007077   1.470   0.1415
waisthip     1.056508   0.575110   1.837   0.0662 .
framemedium -0.052746   0.109855  -0.480   0.6311
framesmall  -0.184365   0.137615  -1.340   0.1803
Outlier p.   0.235691
Tau-sq      20.235727
Sigma-sq     0.340610
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     logLik      AIC      BIC
  -630.1581 1280.316 1319.505
```

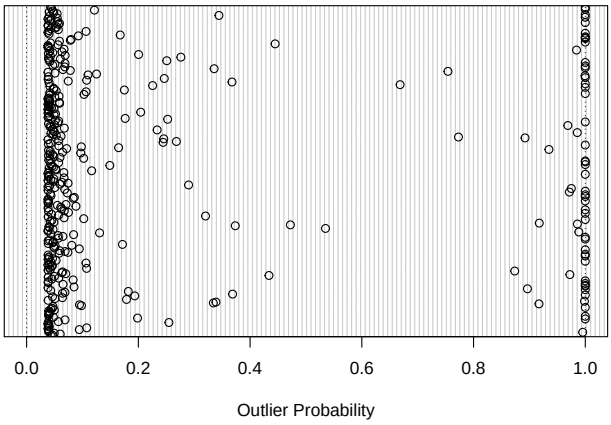

Figure 7: Outlier Probabilities for Diabetic Data

Test for outliers and plot the outlier probabilities in Figure 7.

```
> outlierTest(diabdata.robustmix, showProgress=FALSE)

p value 0.0010

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

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

It is often of interest to simplify a model. This may be performed simply for reasons of parsinonomy, as a simpler model will be easier to understand. It also has the adavantage of removing some covariates that are highly correlated. Removing the covariates will result in a reduction in the standard errors and a consequential decrease in p-values. It does have the disadvantage that it may produce a spurious improvement in fit, especially when the number of covariates compared to 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 et al. (2013, Chapter 6). The advantage of `robmixglm` is that it is likelihood based, so can be used a part of any method that requires a likelihood.

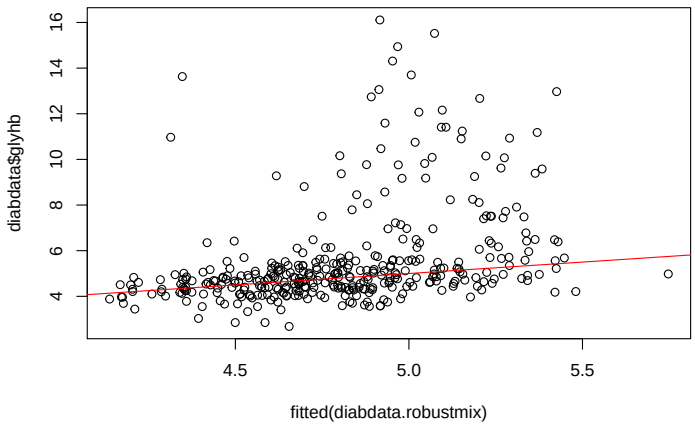

Figure 8: Observed versus Fitted for Diabetic Data

Two main methods exist: 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 taking a long time, but guaranteeing that the best fitting subset is found. For step wise 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 back wise 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.

The `step` function, a simplified version of `stepAIC` described in Venables and Rispley (1999), allows for step wise model selection based on the AIC statistic. Here we use the default of 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.

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

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.582896   0.474757   5.440 5.31e-08 ***
```

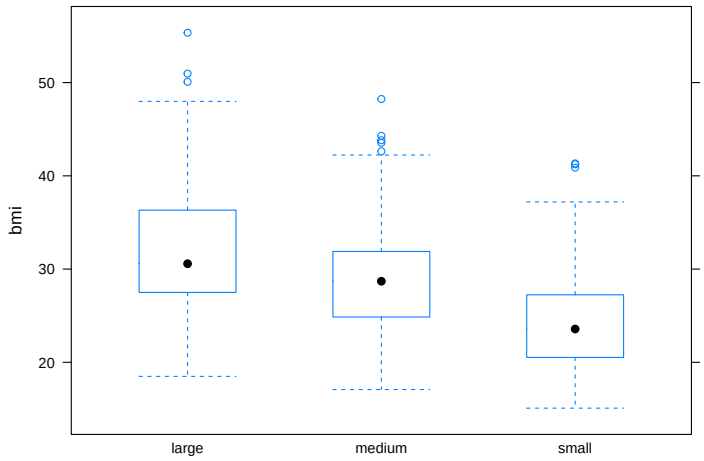

Figure 9: BMI by Frame

```
age           0.014560   0.002537   5.739 9.54e-09 ***
bmi           0.015162   0.005771   2.627  0.00861 **
waisthip      1.267878   0.541608   2.341  0.01923 *
Outlier p.    0.236816
Tau-sq       20.131911
Sigma-sq      0.342148
---
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`. The reason is the correlation between BMI and frame, as shown in Figure 9.

```
> library(lattice)
> bwplot(bmi~frame, data=diabdata)
```

References

M. Aitkin. A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6:251–262, 1996. ISSN 0960-3174. doi: 10.1007/BF00140869.

K. J. Beath. A mixture-based approach to robust analysis of generalised linear models. *Journal of Applied Statistics*, 4763:1–13, 2017. ISSN 13600532. doi: 10.1080/02664763.2017.1414164. URL https://doi.org/10.1080/02664763.2017.1414164.

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

C. Fraley and A. E. Raftery. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *Computer Journal*, 41(8):578–588, 1998.

F. E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression and Survival Analysis*. Springer, 2nd edition, 2015. ISBN 9781441929181.

S. Heritier, E. Cantoni, S. Copt, and M.-P. Victoria-Feser. *Robust Methods in Biostatistics*. John Wiley, Chichester:United Kingdom, 2009.

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

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, London, 2nd edition, 1989.

G. J. McLachlan. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied Statistics*, 36(3):318–324, 1987.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

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

P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.

W. N. Venables and B. D. Rispley. *Modern Applied Statistics with S-PLUS*. Springer-Verlag, New York, third edition, 1999.

J. P. Willems, J. T. Saunders, D. E. Hunt, and J. B. Schorling. Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *Southern Medical Journal*, 90:814–820, 1997.

D. A. Williams. Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions. *Journal of the Royal Statistical Society. Series C.*, 36 (2):181–191, 1987. doi: 10.1002/jae.l.