GLM II

Presented by Joseph O. Marker

Marker Actuarial Services, LLC
and
University of Michigan

2011 RPM Meeting
New Orleans, LA

Overall Goal:

- Go over concepts common to GLMs.
- Give ideas on how you can learn more about GLMs.
- Cover some modeling topics through GLM examples.

General Outline

Basic concepts, such as:
- Likelihood, deviance, canonical link, exponential family.

Develop concepts using the Poisson GLM

Illustrate modeling with a Gaussian (Normal) GLM
- Use this to look at modeling issues such as:
  - Validation and testing, specifically
  - Model comparison and use of training vs. testing data.

If time permits, special topics such as
- Survival models / censored data

Bibliography

[1] McCullagh, P., and J.A. Nelder, *Generalized Linear Models,* 2nd ed., Boca Raton, FL: Chapman & Hall/CRC, 1989.
Comments: Excellent reference on GLMs and Poisson GLMs and is often considered the definitive early work on generalized linear models.

[2] Faraway, Julian J., *Extending the Linear Model with R,* Boca Raton, FL: Chapman and Hall/CRC, 2006.
Comments: Textbook covering topics beyond linear models, such as GLMs, mixed-effects models, and non-parametric models. The book uses the programming language R and assumes knowledge of linear models.

[3] Klugman, Stuart, Harry Panjer, and Gordon Willmot, *Loss Models: From Data to Decisions,* Third ed., New York: John Wiley & Sons, Inc., 2008.
Comments: *(Included in the CAS Syllabus)* Standard Textbook for Exam 4

Setup for a GLM:

Data: Response variable Y and covariates $\{X_j$, j = 1, 2 ... $p\}$

Assume complete data and individual observations i=1, 2, ...n

Observation matrices $\boldsymbol{X}=(x_{ij}) = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{pmatrix}_{nxp}$ and $\boldsymbol{Y}= \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}_{nx1}$.

GLM common elements:

- Coefficients $\boldsymbol{\beta}^T = (\beta_1 \dots \beta_p)$ for which likelihood is maximized.
- The linear predictor $\boldsymbol{\eta}$, where $\eta_i = x_i^T \beta$.
- Link function g for which $g(\mu_i) = \eta_i$, where $\boldsymbol{\mu}$ is fitted value of $\mathbf{Y}$.

Poisson GLM:

Canonical link for the Poisson GLM is $g(\mu) = \ln(\mu)$.

The mean response $\mu = \exp(\eta_i) = \exp(x_i^T \beta)$.

Set up loglikelihood function for the i^{th} observation.

$L = (\Pr[Y = y_i \mid \mu]) = e^{-\mu} \frac{\mu^{y_i}}{y_i!}$, from which

$l(\beta) = y_i \ln(\mu) - \mu - \ln(y_i!) = \eta_i y_i - e^{\eta_i} - \ln(y_i!)$.

The total loglikelihood is $l(\beta) = \sum_{i=1}^{n} y_i x_i^T \beta - \exp(x_i^T \beta) - \ln(y_i!)$. (1)

Maximize $l(\beta)$ by setting all the partial derivatives $\frac{\partial l}{\partial \beta_j} = 0$.

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - x_{ij} \exp(x_i^T \beta) = 0, \text{ for j = 1, 2, ... } p.$$

We can rewrite this as $X^T y = X^T \hat{\mu}$ (2)

Important: This relationship holds only for the link function ln(μ), one reason we call this function the **canonical link**.[1] .

Comment [jm1]: When writing the general GLM, note the tables on p 55 of N&M. and Faraway pages 117 & 121.

There is a parallel in linear regression – we will discuss this later.

[1] This discussion follows [2], p. 57.

"Deviance" is the GLM criterion for goodness of fit. [2]

Deviance = Δl between the given model and "saturated" model.
Analogous to "residual sum of squares" in regression.

Illustrate using Poisson GLM
Rewrite (1) as $l(\beta) = \sum_{i=1}^{n} y_i \ln(\mu_i) - \mu_i - \ln(y_i!) \triangleq l(\boldsymbol{y}; \boldsymbol{\mu})$
Saturated model:
Think of a model with # parameters = n = number observations.
$l(\boldsymbol{y}; \boldsymbol{\mu})$ is maximized when $\boldsymbol{\mu} = \boldsymbol{y}$.
Saturated model has $\boldsymbol{\mu} = \boldsymbol{y}$ and $l = l(\boldsymbol{y}; \boldsymbol{y})$

Residual deviance = $2\,[l(\boldsymbol{y}; \boldsymbol{y}) - l(\boldsymbol{y}; \boldsymbol{\mu})]$
This is a "likelihood ratio" statistic.

---

[2] From [1], p. 24

Why use deviance (a.k.a. "scaled deviance")?

Important in hypothesis testing.

In comparing a larger model to a smaller nested model,

difference in scaled deviance is asymptotically χ^2
with degrees of freedom equal to
difference in number of identifiable parameters.[3]

This is the "likelihood ratio test".

You can treat GLM deviance residuals
similarly to the way you treat "residuals" in Linear Models

[3] [2], p. 121

Generalized Linear Models (GLMs):

A GLM has two components.

1. Distribution of Y in the exponential family, i.e.,
   $f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$, where f is density or point mass.

2. Link function relating the mean response to the linear predictor.

Here a, b, and c are functions. If ϕ is known, the model has **canonical parameter** θ.
We can show that $E(Y) = b'(\theta)$ and $var(Y) = b''(\theta)a(\phi)$.[4]

Comment [jm2]: N& M, p 28, Faraway, p 115

[4] [1], pp. 28-29

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right],$$

Comment [jm3]: N& M, p 28, Faraway, p 115

Application to Poisson and Normal GLM

Poisson:

$f(y;\theta,\phi) = \exp(y \ln \mu - \mu - \ln y!)$,

so that $\theta = \ln \mu$, $b(\theta) = \exp(\theta)$, $a(\phi)=1$, and $c(y, \phi) = \ln y!$.

Note $EY = \exp(\theta) = \mu$ and $\mathrm{Var}[Y] = \exp(\theta) = \mu$.

Normal: $f(y;\theta,\phi) =$

$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) = \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)\right],$$

so that $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $\phi = \sigma^2$, $a(\phi)=\phi$, $c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \ln(2\pi\phi)\right)$.

$EY = b'(\theta) = \theta = \mu$, and $\mathrm{var}(Y) = 1\phi = \sigma^2$

Canonical links for Poisson and Normal GLM.

Canonical link is the one for which linear predictor η equals θ.

For the Poisson Model, $\eta = \ln(\mu) \equiv g(\mu)$ as we showed earlier.

For the Normal GLM, $\eta = \theta = \mu$.

Equivalence of Normal GLM and Linear Regression.

Normal equations from linear regression result in $X^T y = X^T X \beta$.

However, the fitted value $\hat{\mu}$ equals $X\beta$ for multiple linear regression.

We showed earlier (equation (2)) that $X^T y = X^T \hat{\mu}$ for the GLM canonical link function.

$\therefore$ Result from GLM = Result from linear regression.

Moreover, the Normal GLM deviance equals $\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$, which equals the residual sum of squares from linear regression.

Other GLMs.

Besides the Poisson and Normal, there are three other GLMs:
Binomial (or Logistic), Gamma, and Inverse Gaussian. [5]

"Almost" GLMs.

Faraway, on p. 116 of [2], notes:

"Some other densities, such as the negative binomial and the Weibull distribution, are not members of the exponential family, but they are sufficiently close that the GLM can be fit with some modifications."

[5] For summary information about all the GLMs, see [1] p. 30, or [2] pp. 117 and 121.

We look at a GLM example.

- Response variable is "logloss".
- Covariates are x1, x2, and x3.
- Model using a GLM of the "family" Gaussian.
  - This produces the same predictions as linear regression.
- However, the same modeling statements apply to other GLMs.
- The data is stored in "data2"—

  Think of this as an observation matrix.

Explore data2:

```
> sapply(data2,mean)

      x1       x2       x3  logloss
2.999184 3.002408 1.996825 6.942337


>
> sapply(data2,sd)         #### uses n-1 in denominator

       x1        x2        x3   logloss
0.9967996 0.4991438 0.9989666 1.8556773
>
```

Correlation matrix:

```
> cor(data2[,1:4])

                x1         x2          x3   logloss
x1       1.0000000 0.69736525 -0.10496951 0.9107591
x2       0.6973653 1.00000000  0.09393076 0.8130073
x3      -0.1049695 0.09393076  1.00000000 0.1422006
logloss  0.9107591 0.81300734  0.14220057 1.0000000
>
```

For model validation, split the data into two pieces:
Training dataset – 80% of the observations.
Testing dataset – 20% of the observations

All modeling is done using the training data.

Training data is the subset defined by data2[data2$train,].

Let's look at histograms and densities:

```
> par(mfrow=c(2,1))    ####   Histogram and Density
> hist(logloss[data2$train],main="Histogram of logloss",
       freq=FALSE)
> plot(density(logloss[data2$train]),
        main="Density estimate of logloss")
```

Figure 1 shows the results.

Figure 2 shows plots the response logloss against the covariates.

For example, the R statements that produce the first panel are:

```
smoothScatter(x=data2$x1[data2$train],
     y=data2$logloss[data2$train],
     main="Logloss vs. x1 ",
     nrpoints = 100, pch = ".", cex = 1, col = "black",
     xlab = "x1", ylab = "logloss",
     )
> temp1 <- lm(logloss~x1,data=data2[data2$train,])
> abline(temp1,col="red")
```

Note positive linear relationship of response vs. each covariate.

Now we model using the Gaussian GLM with no interactions:

```
modelnull <-  glm(logloss  ~ 1 ,   #### intercept only
     data = data2[data2$train,],
     family = gaussian,          ### normal or Gaussian glm
     x=T)

model2 <- glm(logloss  ~  x1 + x2 + x3 ,
     data = data2[data2$train,],
     family = gaussian,              ### normal or Gaussian glm
     x=T)
```

"Modelnull" is for reference in comparing models.

The simple statement "plot(model2)" produces Figure 3.

First panel plots residuals versus fitted values
   --- notice the upturn at both ends.

The "summary" function in R generates diagnostics for the GLM

```
> summary(model2,correlation=T)

glm(formula = logloss ~ x1 + x2 + x3,
  family = gaussian,
  data = data2[data2$train,], x = T)

........   (output omitted)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.045014   0.025663  -40.72   <2e-16 ***
x1           1.361782   0.005790  235.18   <2e-16 ***
x2           1.060133   0.011484   92.32   <2e-16 ***
x3           0.360594   0.004131   87.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2560862)

    Null deviance: 54577.5  on 15999  degrees of freedom
Residual deviance:  4096.4  on 15996  degrees of freedom
AIC: 23616
```

All variables are significant.

There is an ANOVA function in R – we will illustrate this later.

We add the interaction variables to model 2.

```
model3 <- glm(logloss ~ x1 + x2 + x3 + x1*x2 + x1*x3+x2*x3,
     data = data2[data2$train,],
     family = gaussian,      ### normal or Gaussian glm
     x=T)
```

Statement "plot(model3)" produces Figure 4.

Panel 1 plots Model 3 residuals versus fitted values.

Its curve of best fit is horizontal, unlike that for Model 2..

Partial output from `summary(model3,correlation=F)`

```
...........
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.079148   0.080422   0.984  0.32505
x1           1.025129   0.022866  44.832  < 2e-16 ***
x2           0.637106   0.032137  19.825  < 2e-16 ***
x3           0.330515   0.024761  13.348  < 2e-16 ***
x1:x2        0.123240   0.006584  18.718  < 2e-16 ***
x1:x3       -0.017160   0.005414  -3.170  0.00153 **
x2:x3        0.027229   0.010804   2.520  0.01173 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interactions are significant, but only x1*x2 is "really" significant.

Figure 5 shows the density for the Model 3 residuals.

Density similar to a normal r.v. with mean 0, and sigma < 1.

To compare nested models, use R's the **"anova"** function.:

```
anova(modelnull, model2,model3,test="Chi")

Analysis of Deviance Table

Model 1: logloss ~ 1
Model 2: logloss ~ x1 + x2 + x3
Model 3: logloss ~ x1 + x2 + x3 + x1 * x2 + x1 * x3 + x2 * x3
   Resid.Df Resid. Dev Df Deviance        P(>|Chi|)
1     15999      54577
2     15996       4096  3    50481   < 2.2e-16 ***
3     15993       4007  3       89   < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Using this criterion, the "full model" is indicated.

Model selection using AIC:

The AIC (Aikaike Information Criterion) is a way to rank models.

Let p be the number of covariates in a model.

AIC = − 2 * (maximum loglikelihood) + kp.

Smaller AIC is better.

For true AIC, k=2.

This number determines the "penalty" for adding variables.

Other methods have harsher penalties (higher k)
for adding variables (e.g., BIC).

"R" (in the MASS library) has a stepwise model selection method using AIC:

```
AIC1 <- stepAIC(model3,
  scope=list (upper = ~ x1+x2+x3+x1*x2+x1*x3+x2*x3+x1*x2*x3),
  k = 2, trace=T)
Start:  AIC=23270.14
logloss ~ x1 + x2 + x3 + x1*x2 + x1*x3 + x2*x3

           Df Deviance     AIC
<none>          4007.2   23270    (note: Start w/ Model 3)
+ x1:x2:x3  1   4007.2   23272
- x2:x3     1   4008.8   23275
- x1:x3     1   4009.7   23278
- x1:x2     1   4095.0   23615
```

The upper model is Model 3 with 3-way interaction added.
Lower model is the null model by default.

AIC is lowest for model3. Three-way interaction is not indicated.

Getting rid of `x2*x3` and `x1*x3` should be investigated.

Validation:

We ran the model using the "training data".
The remaining data is the "test data".

Model 3 generates fitted values and residuals for the training data.
Model coefficients can generate predicted values for the test data.
Define "residuals" on the test data: *residual* = *actual* – *predicted*.
Test data residuals should be similar to those for training data.

We compare the residuals for the training and test data in Figure 6

Validating a model is more complex than just performing this test.

Orthogonal Covariates:

When covariates are correlated , orthogonalizing is useful.

With orthogonal covariates, the ANOVA table is independent of the order the variables are brought into the model.

How to orthogonalize:

First obtain the covariance matrix covar for the training data.

```
           x1         x2          x3
x1  0.9820935 0.34435952 -0.10579856
x2  0.3443595 0.24880551  0.04424759
x3 -0.1057986 0.04424759  1.00123295
```

The function `chol(covar)` gives the Cholesky upper triangular matrix U for which covar = $U^T U$.

Covariates V = XU^{-1} have covariance = identity matrix.

Run model4, like model 3 except with orthogonal covariates V:

The coefficients using `summary(model4)` are:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.321669   0.075557  -4.257 2.08e-05 ***
v1           1.456315   0.022573  64.517  < 2e-16 ***
v2           0.323552   0.013422  24.107  < 2e-16 ***
v3           0.329205   0.025164  13.083  < 2e-16 ***
v1:v2        0.042609   0.003995  10.666  < 2e-16 ***
v1:v3       -0.008176   0.003972  -2.059   0.0396 *
v2:v3        0.008224   0.003981   2.066   0.0389 *
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The variables v1*v3 and v2*v3 are less significant than the corresponding variables in model3.

Orthogonalizing improves the model,
but the first variable v1 is a multiple of x1.

Principal components choose orthogonal variables,
the first one being the linear combination of x1, x2, and x3 that explains the most variance.

Other topics:

Censored variables:

You cannot use a GLM because GLMs assume exact values rather than ranges.

Survival models help with this situation.

Simulation:

Once you find a good model, use it to simulate outcomes.

The CAS has a free public simulation model that allows one to:

- simulate number of claims using many claim count distributions.
- simulate size of claim using many severity distributions.
- simulate the reserve change process.
- produce loss development triangles.

Figure 1

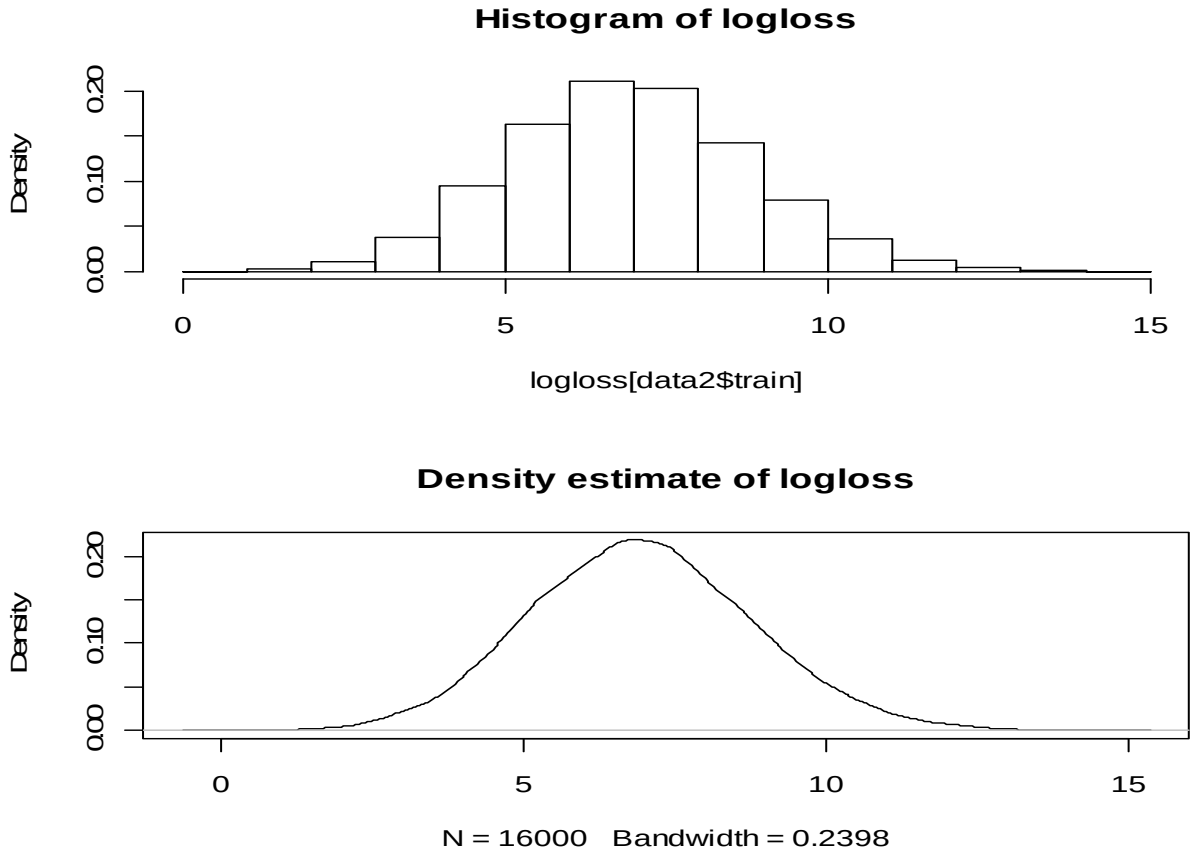

Histogram of logloss
Density
0.00
0.10
0.20
0
5
10
15
logloss[data2$train]
Density estimate of logloss
Density
0.00
0.10
0.20
0
5
10
15
N = 16000 Bandwidth = 0.2398

Figure 2

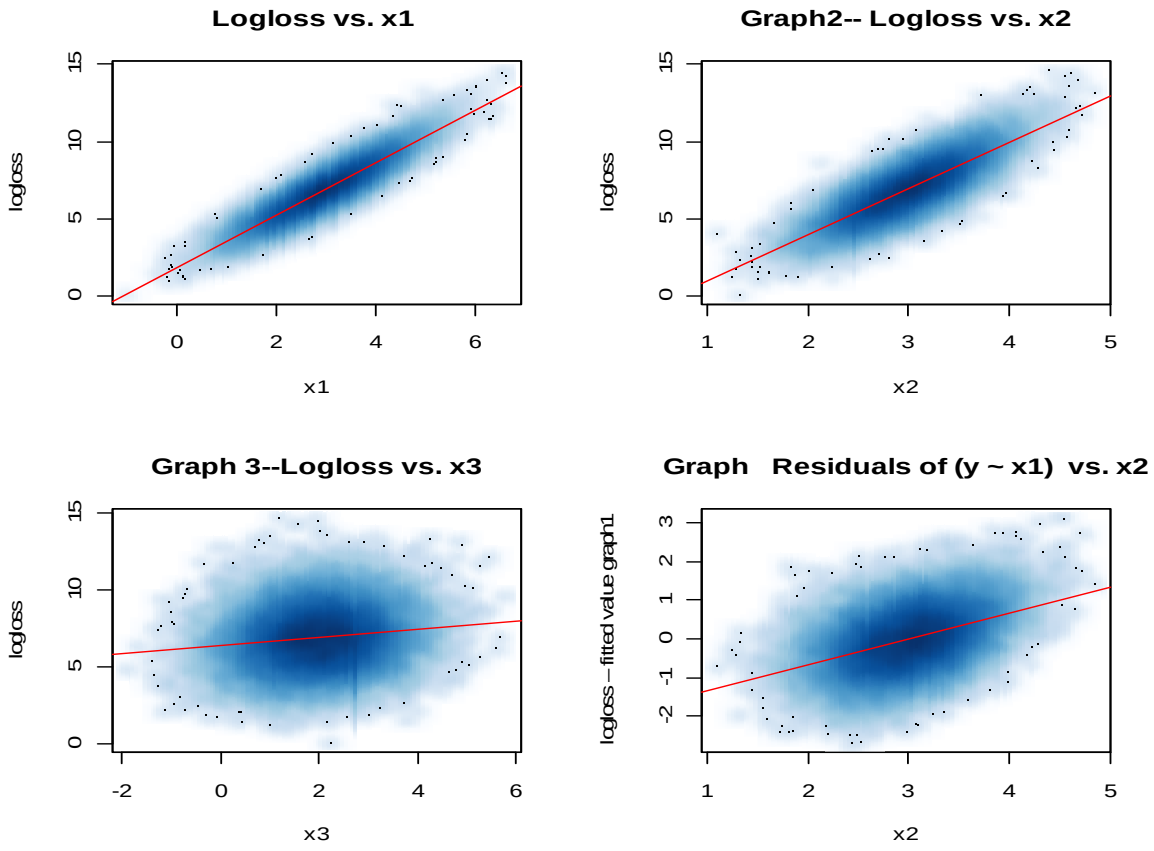

Logloss vs. x1
logloss
x1
Graph2-- Logloss vs. x2
logloss
x2
Graph 3--Logloss vs. x3
logloss
x3
Graph Residuals of (y ~ x1) vs. x2
logloss – fitted value graph1
x2

Figure 3 – Plot information for Model 2

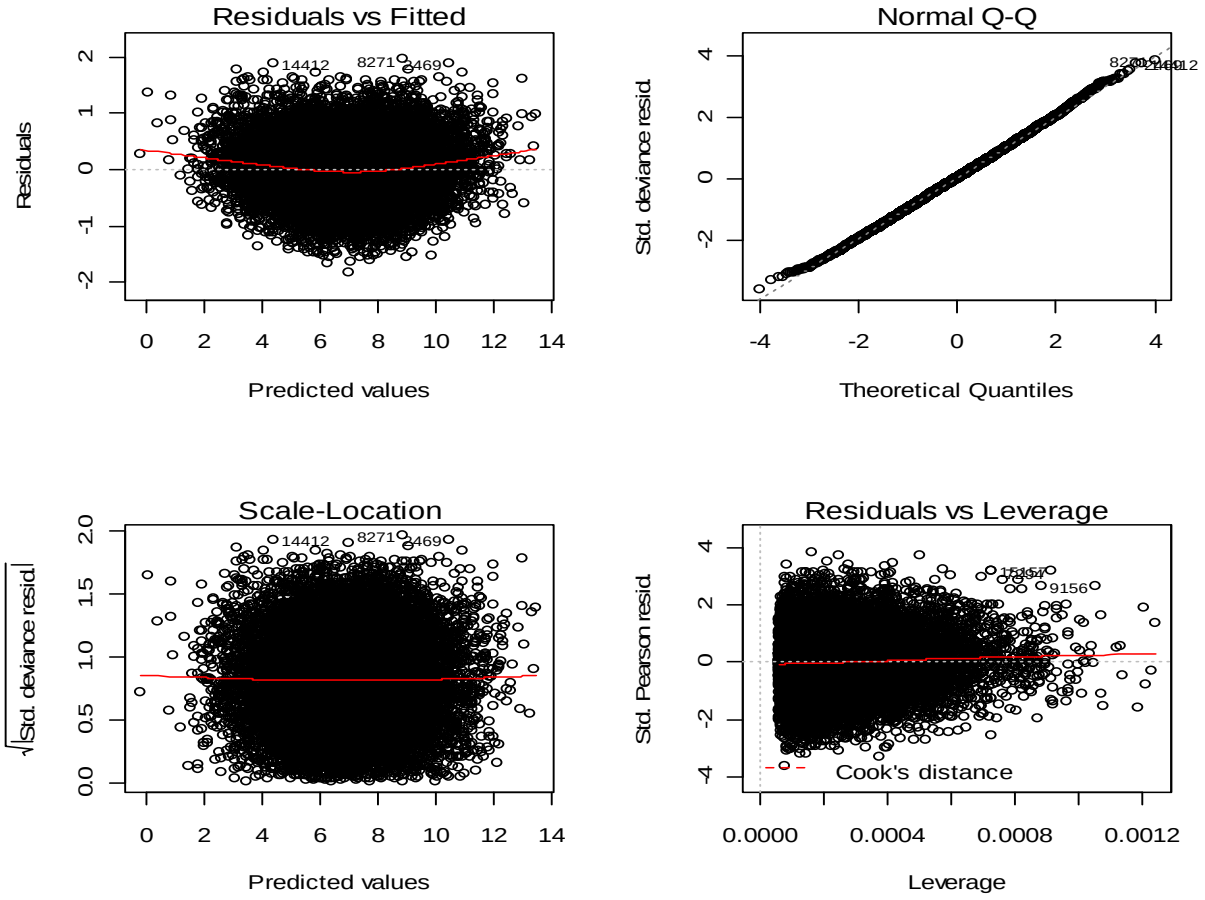


Figure 4 – Plot Information for Model 3

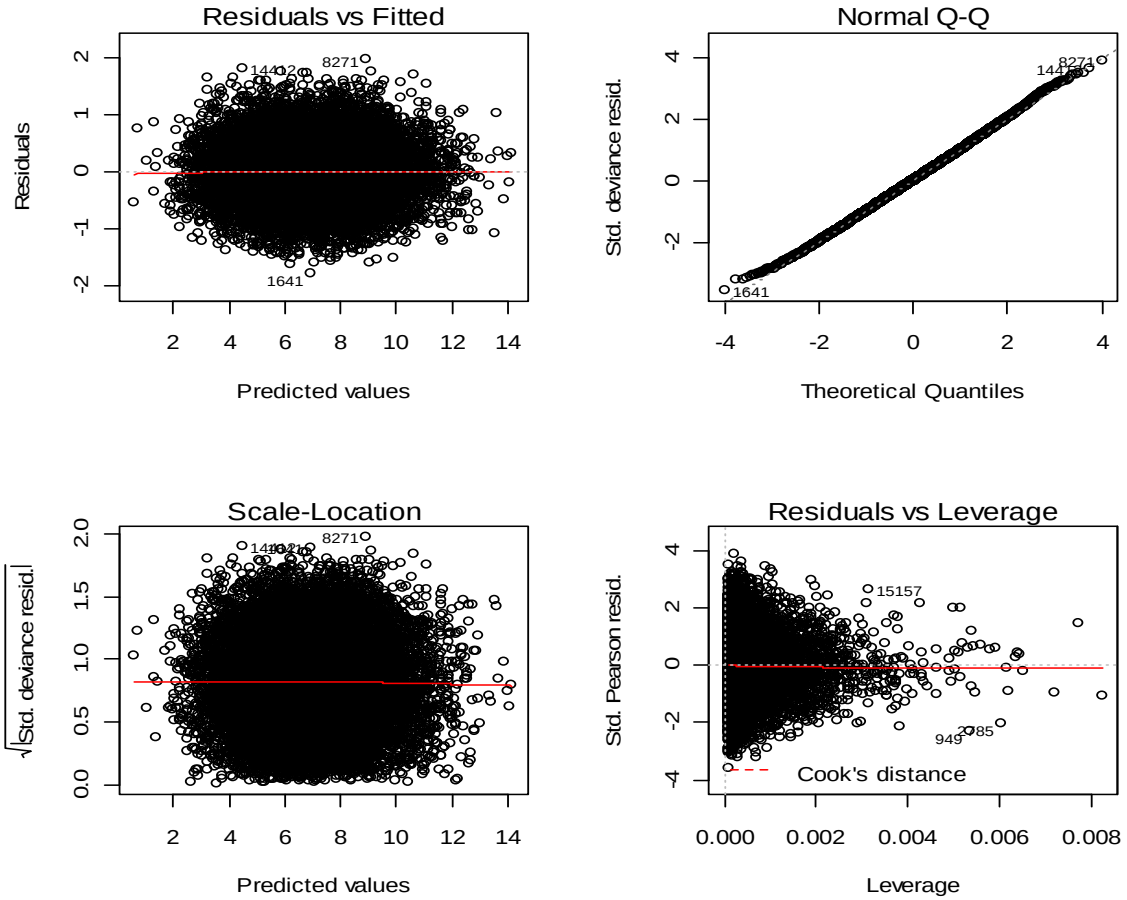


Figure 5 – Model 3 Residuals

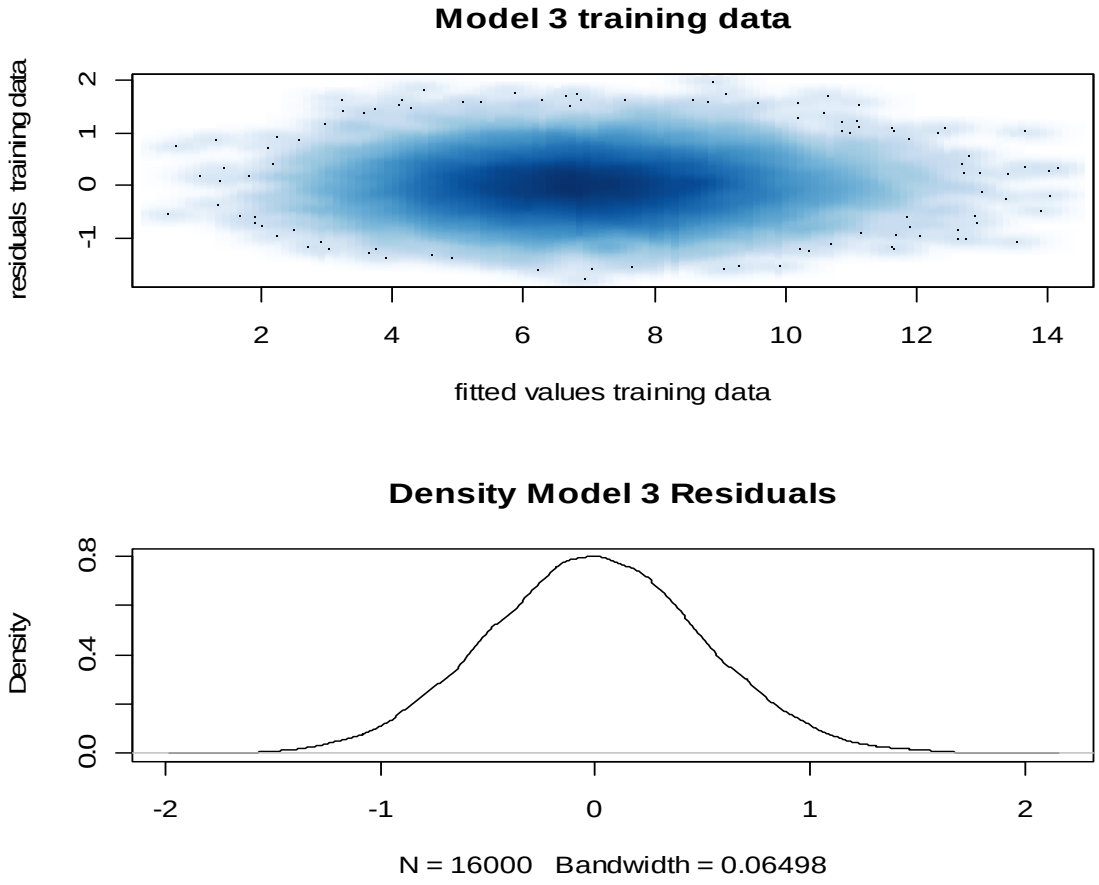
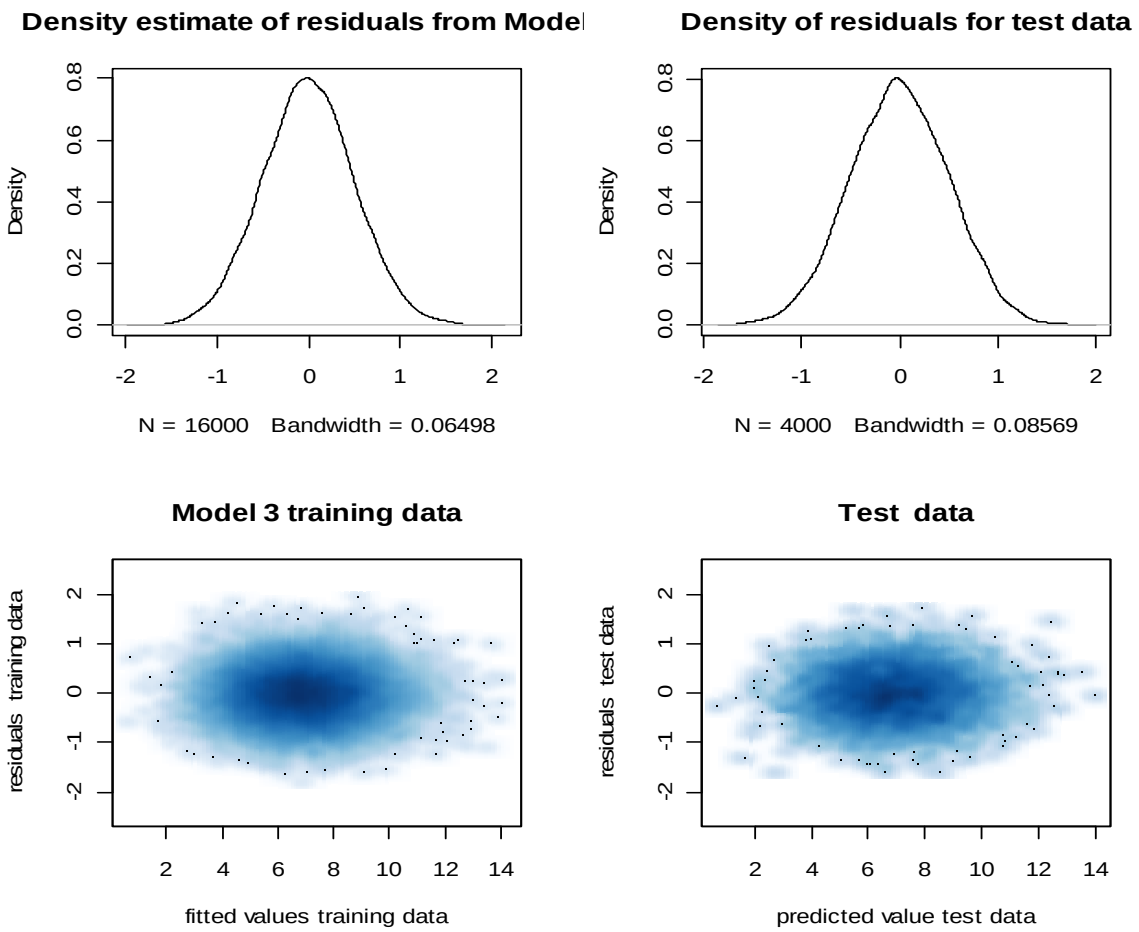


Figure 6 – Residuals for Training and Test Data



THE END