# GLMs - Diagnostics

Bruce A Craig

Department of Statistics
Purdue University

# Outline

- Diagnostics approach
- Types of Residuals
- Other Diagnostics
- Example

Reading: Faraway Ch. 8, Agresti Ch. 5-6, KNNL Ch. 14

# Diagnostics

- Always important to check model conditions
  - If not adaquately met, inference cannot be trusted
- Diagnostic approach to GLMs mirror those for LMs
- Not all methods applicable but some have been adapted
- Challenges
  - Var($Y$) is usually not constant
  - Skewness of EFDs
  - Discreteness of some EFDs

# Linear Model Approach

- Given that model is typically fit using IRWLS, we can consider results from last iteration as if weighted LM
- Helps us derive model-checking diagnostics
  - residuals
  - leverage / influence measures
- Weighted LM set-up
  - $Z = X\beta + \epsilon, \quad E[\epsilon] = 0, \quad var(\epsilon) = \phi \; Var\{z_i^{(k)}\}$
  - $\hat{\beta} = (X'WX)^{-1}X'WZ$
  - $\text{Cov}(\hat{\beta}) = (X'WX)^{-1}$
  - $H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$
- Be careful with dispersion parameter $\phi$. Not needed in IRWLS but needed to get proper SEs here

# Residuals: Part I

- Response residual: $r_{i,R} = y_i - \hat{\mu}_i$
  - Analogue to simple residual in regression
  - Do not have constant variance
  - In R: `residuals(glmfit, type="response")`
- Working residual: $r_{i,W} = Z_i - \hat{\eta}_i$
  - Add $x_{ij}\hat{\beta}_j$ for partial residuals (plot vs $X_j$)
  - Can be used to assess model misfit
  - In R: `residuals(glmfit, type="working")`
    `residuals(glmfit, type="partial")`
- Pearson residual: $r_{i,P} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$
  - $\sum r_{i,P}^2 = X^2$
  - In R: `residuals(glmfit, type="pearson")`

# Residuals: Part II

- Standardized Pearson residual: $r_{i,SP} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\ V(\hat{\mu}_i)\ (1-h_{ii})}}$
  - $h_{ii}$ is the $i$th diagonal element of $H$
  - Have constant variance if $V(\mu)$ correctly specified
  - Analogue to studentized residuals
  - Useful for detecting variance misspecification
- In R: `library(boot); glm.diag(glmfit)$rp`

# Residuals: Part III

- Deviance residual: $r_{i,D} = sign(y_i - \hat{\mu}_i)\sqrt{d_i}$
  - $d_i$ is Case $i$'s contribution to the model deviance
  - $\sum r_{i,D}^2 = D(\hat{\beta})$
- Standardized deviance residual: $r_{i,SD} = \frac{r_{i,D}}{\sqrt{\hat{\phi}(1-h_{ii})}}$
- Deviance residuals more Normal (or at least less skewed) than Pearson residuals
  - Not when $y$ is binary!
- When less skewed, may be better than Pearson residuals for outlier detection
- In R: $r_D$: `residuals(glmfit, type="deviance")` and $r_{DS}$: `library(boot); glm.diag(glmfit)$rd`

# Residuals: Part IV

- Jackknife residual: approximated by

$$r_{i,J} = sign(y_i - \hat{\mu}_i)\sqrt{(1 - h_{ii})r_{i,SD}^2 + h_{ii}r_{i,SP}^2}$$

  - represents the diff in $D$ when $i$th response omitted
  - like deleted studentized residuals
  - is in between $r_{i,SD}$ and $r_{i,SP}$
  - usually closer to $r_{i,SD}$ since $h_{ii}$ small
  - a good choice for diagnostics
- In R: `library(boot); glm.diag(glmfit)$res` or `rstudent(glmfit)`

# Randomized Quantile Residuals

- We are comfortable assessing residuals from Normal dists
- Residuals from GLMs can be highly non-Normal, making it difficult to interpret plots
- Dunn and Smyth (1996) proposed randomized quantile residuals

$$r_{i,q} = \Phi^{-1}(F(Y_i|\hat{\mu}_i, \hat{\phi}))$$

  where $F$ is the cdf of the EFD
- Premise: If $Y_i$ are from EFD, then $F(Y_i) \sim U(0,1)$ and $r_{i,q} \sim N(0,1)$
- If $Y_i$ are from EFD, then variation due to estimation of $\mu$ and $\phi$

# Randomized Quantile Residuals

- If $F$ not continuous, a more general definition is needed
- Let $a_i = F(\max(0, Y_i - 1)|\hat{\mu}_i, \hat{\phi})$ and $b_i = F(Y_i|\hat{\mu}_i, \hat{\phi})$
- The randomized quantile residual is

$$r_{i,q} = \Phi^{-1}(u_i)$$

  where $u_i \sim U(a_i, b_i)$
- Due to this randomness, the residuals will vary each plot to plot but the general patterns in them should remain
- Does not address influence but could "leave out" case to get $\hat{\mu}_{(i)}$ and $\hat{\phi}_{(i)}$ but far more computation

# Cook's Distance and Leverage

- The Cook's distance statistics: $D_i = \frac{(\hat{\beta}_{(i)}-\hat{\beta})' X' W X (\hat{\beta}_{(i)}-\hat{\beta})}{p \ \hat{\phi}}$
  - $\hat{\beta}_{(i)}$ is an estimate of $\beta$ when excluding case $i$
  - $p$ is the number of parameters
  - Measures the standardized change in linear predictor when the $i$th case is deleted
    - a standardized sum of squared $\Delta\beta$
  - Requires $n$ maximizations but can be approximated by a one-step procedure
- In R: `library(boot); glm.diag(glmfit)$cook` or `cooks.distance(glmfit)`

# Plots to Consider

- Can plot residuals versus $\hat{\mu}$ or $\hat{\eta}$ with the latter prefered
- Partial residuals plots can be used to assess model structure
- Can also plot $\hat{\eta}$ versus $z$ for linearity
- Faraway considers the use of half-Normal plots to look for unusual values
- `glm.diag.plots` provides a set of diagnostic plots
  - Jackknife residuals versus linear predictor
  - Normal scores plot of standardized deviance residuals
  - Cook's distance by case
  - Cook's distance against $h_i/(1-h_i)$

# Example

```
###  Poisson distribution with log link
set.seed(612)

##Generate data
x1 = rnorm(200,0,1.5)
x1sq = x1*x1
x2 = rnorm(200,0,1.5)
eta = 2.0 + .1*x1 + 0.05*x1sq - .1*x2
y = rpois(200,exp(eta))

##Fit using incorrect model
bc = glm(y~x1+x2,family=poisson(link="log"))

##Fit using model with x1squared
bc1 = glm(y~x1+x1sq+x2,family=poisson(link="log"))
```

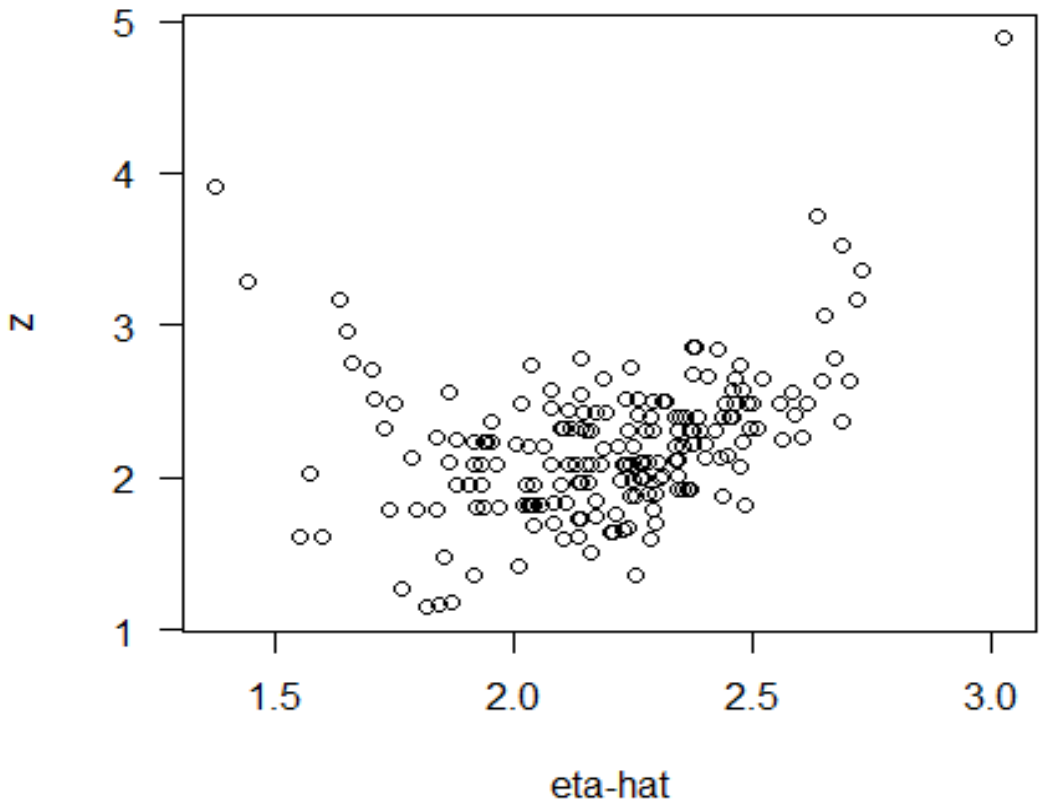
# Example - Misfit model

```
##Get the mu-hats
bcmu = predict(bc,type = "response")
##Get the eta-hats
bceta = predict(bc)

##Get the transformed y's. Linearized responses
##Because log link deta/dmu = 1/mu
z = bceta+(y-bcmu)/bcmu

###Response residual
rr = residuals(bc,type="response")

###Working residual
rw = residuals(bc,type="working")

###Partial residuals
rpartial = residuals(bc,type="partial")
```

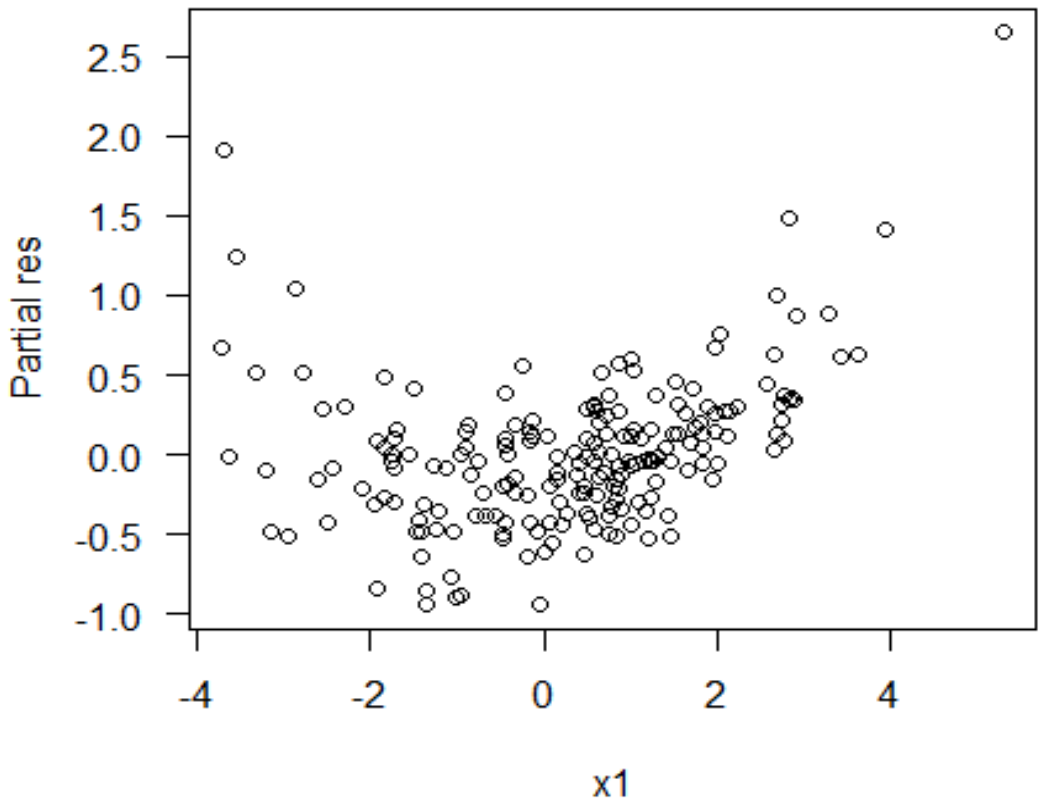
# Example - Misfit model

```
###Check eta-hat versus z's
plot(bceta,z)

###Partial residual plots
plot(x1,rpartial[,1])
plot(x2,rpartial[,2])

###Pearson residual
rp = residuals(bc,type="pearson")

###Standardized Pearson residual
rsp = glm.diag(bc)$rp

###Check constant variance
plot(bceta,rsp,xlab="Linear Predictor",ylab="Residual",las=1)

###Deviance Residual
rd = residuals(bc,type="deviance")
```

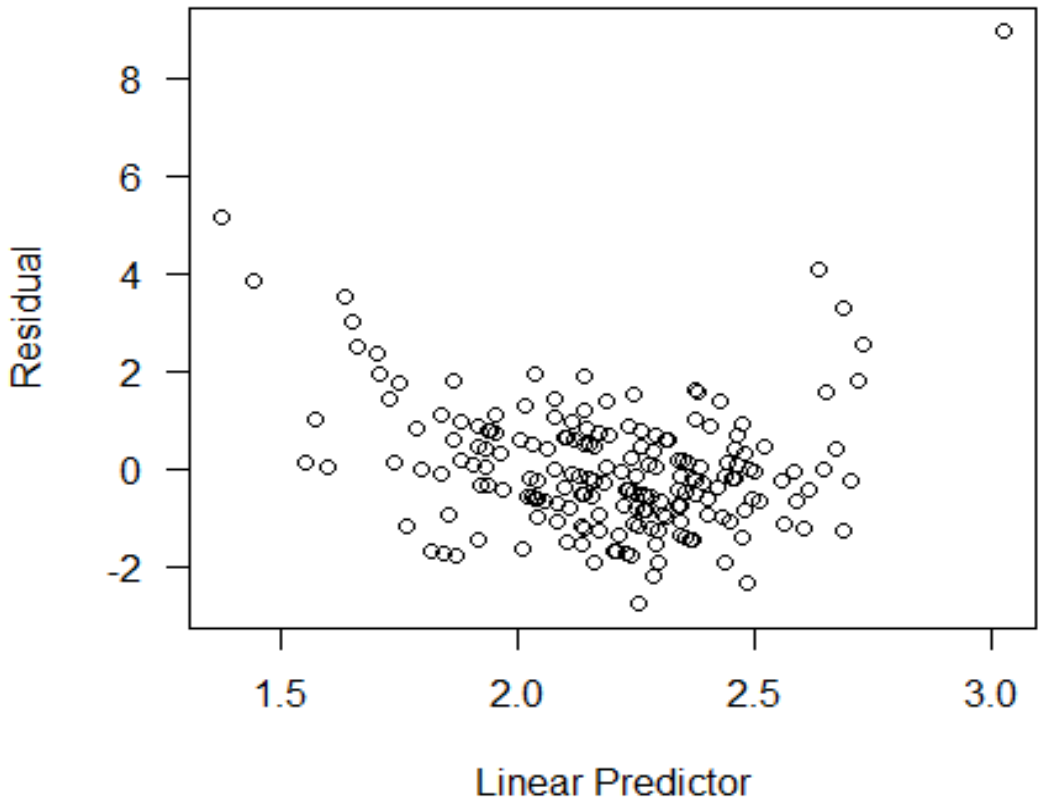
# Checking the linearity



# Partial Residual Plot



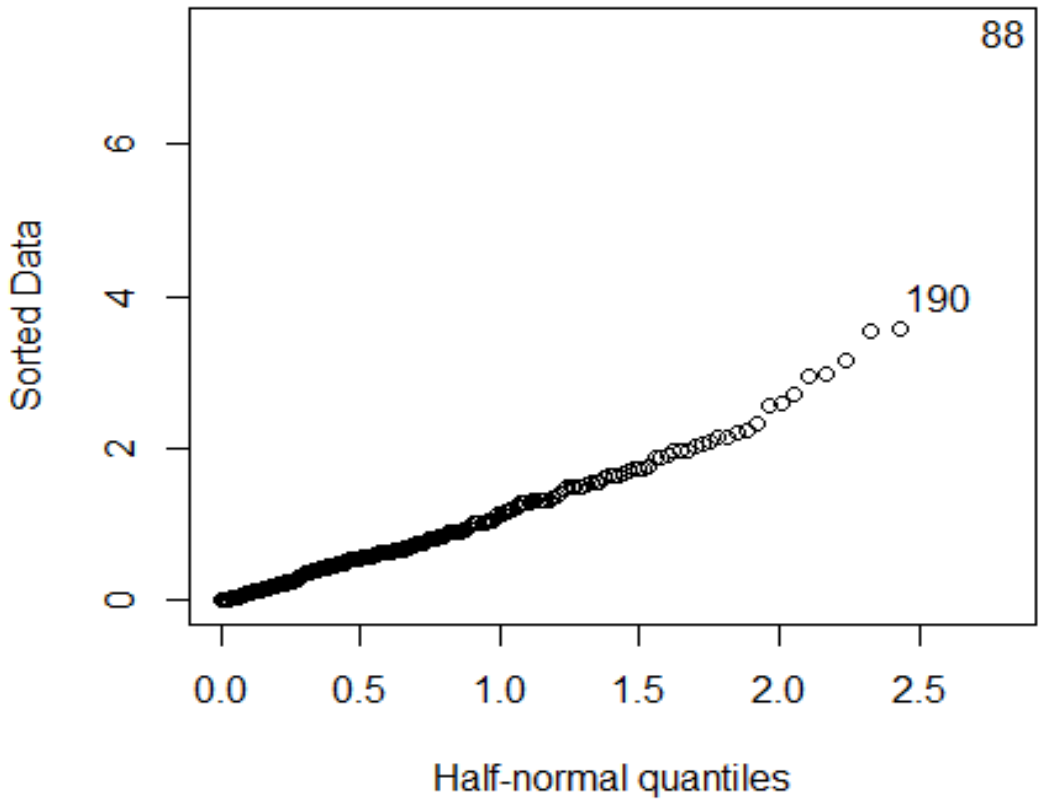
# Explaining all the variance?



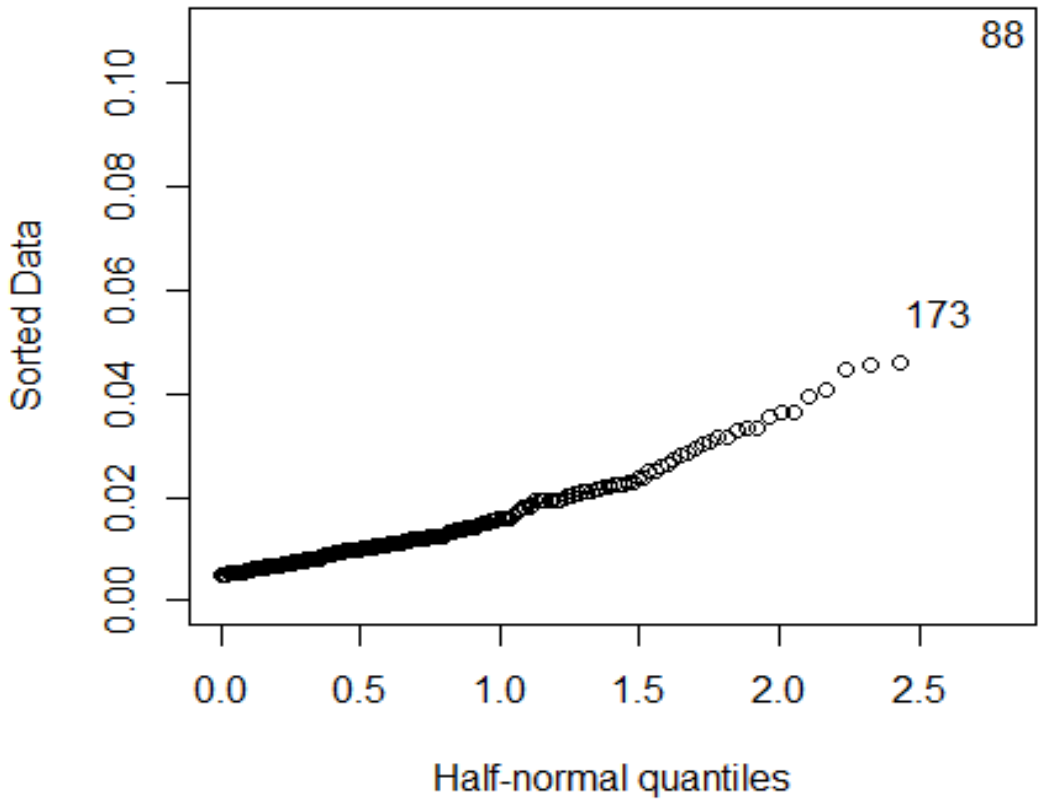
# Example- Misfit model

```
###Residual plot
plot(bceta,rsd,xlab="Linear Predictor",ylab="Residual",las=1)

###Jackknife residual
rj = rstudent(bc)

###Residual plot
plot(bceta,rj,xlab="Linear Predictor",ylab="Residual",las=1)

###Check for unusual points
halfnorm(rj)

###Leverages
lev = glm.diag(bc)$h
halfnorm(lev)

###Cook's Distance
cooksd = cooks.distance(bc)
halfnorm(cooksd)
```

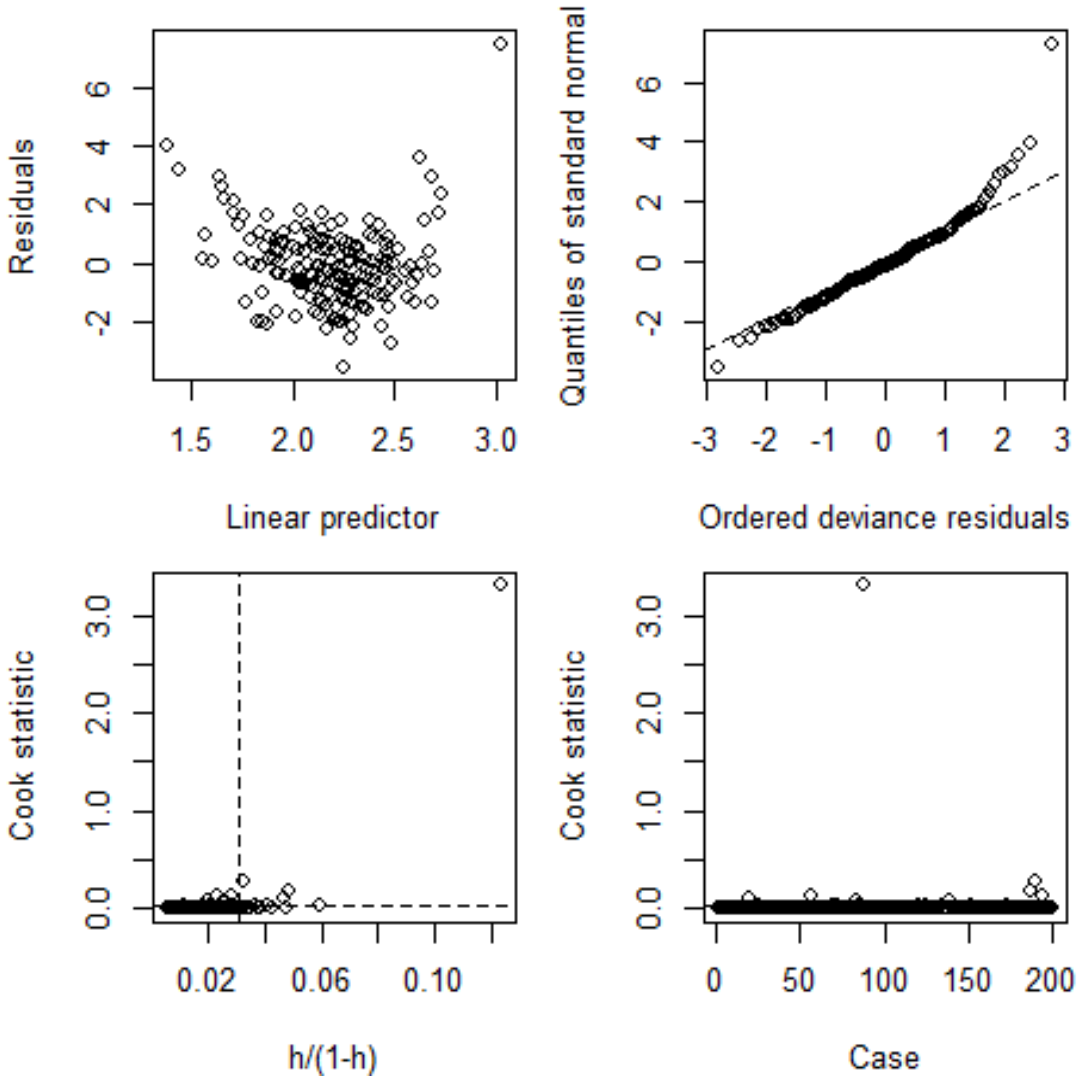
# Unusual residuals?



# Unusual residuals?



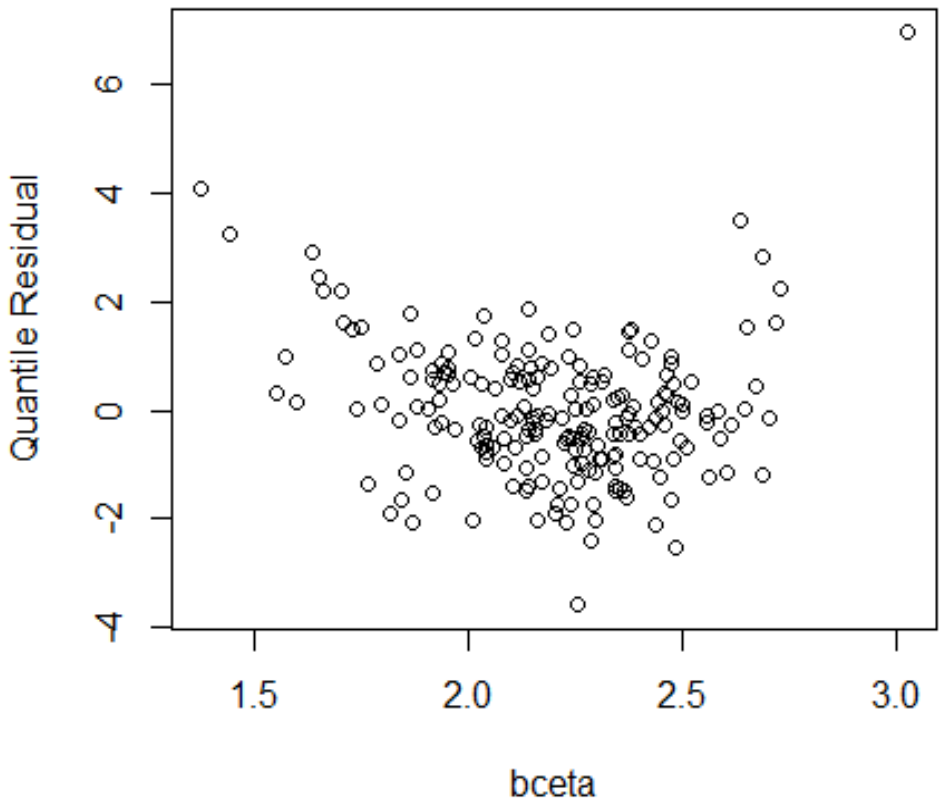
# Diagnostic Summary



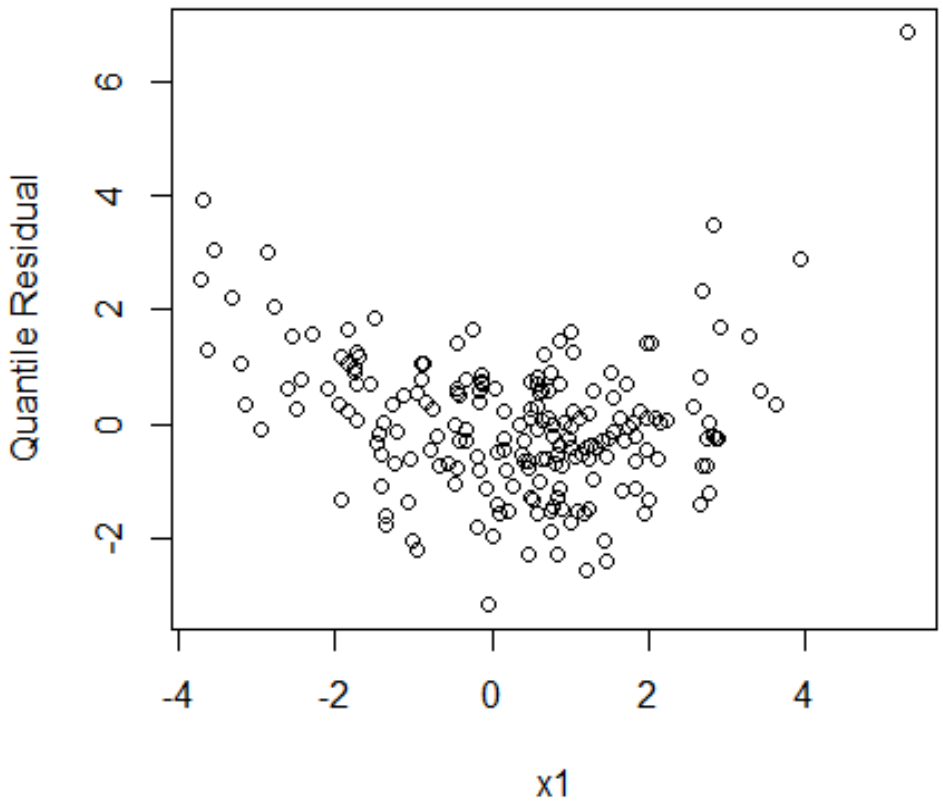
# Examining Quantile Residuals

```
#install.packages("statmod")
library(statmod)

###Residual plot
plot(bceta,qres.pois(bc),ylab="Quantile Residual")
plot(x1,qres.pois(bc),ylab="Quantile Residual")
plot(x2,qres.pois(bc),ylab="Quantile Residual")

###Assess Normality of the residuals
qqnorm(qres.pois(bc))
abline(a=0,b=1,col="red")
```

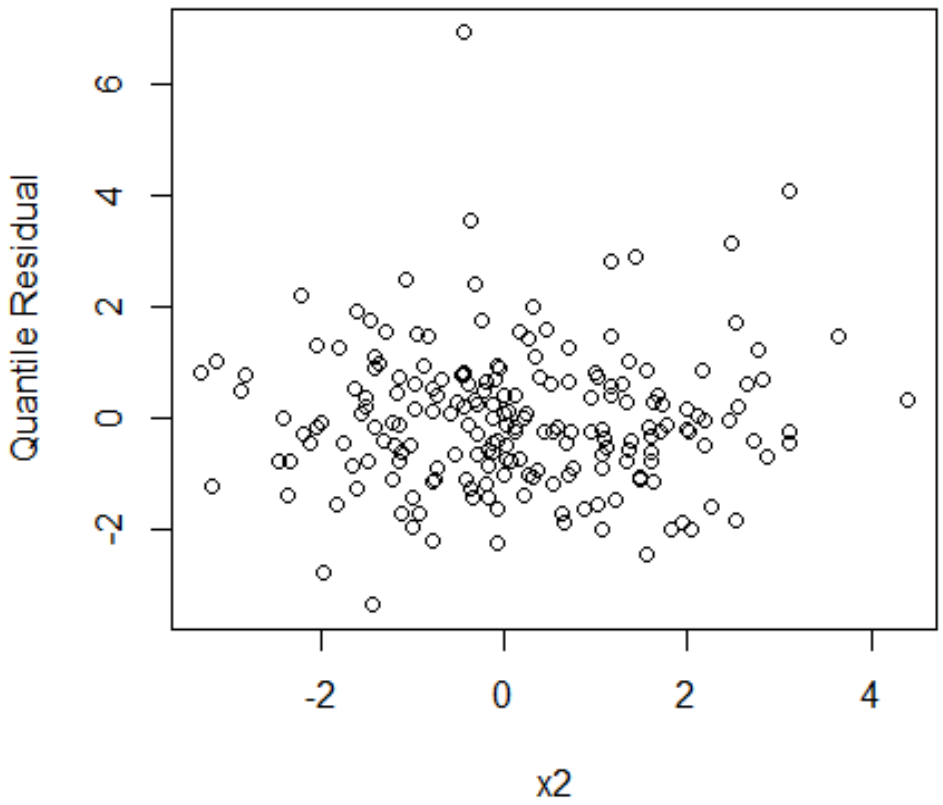
# Residual Plot



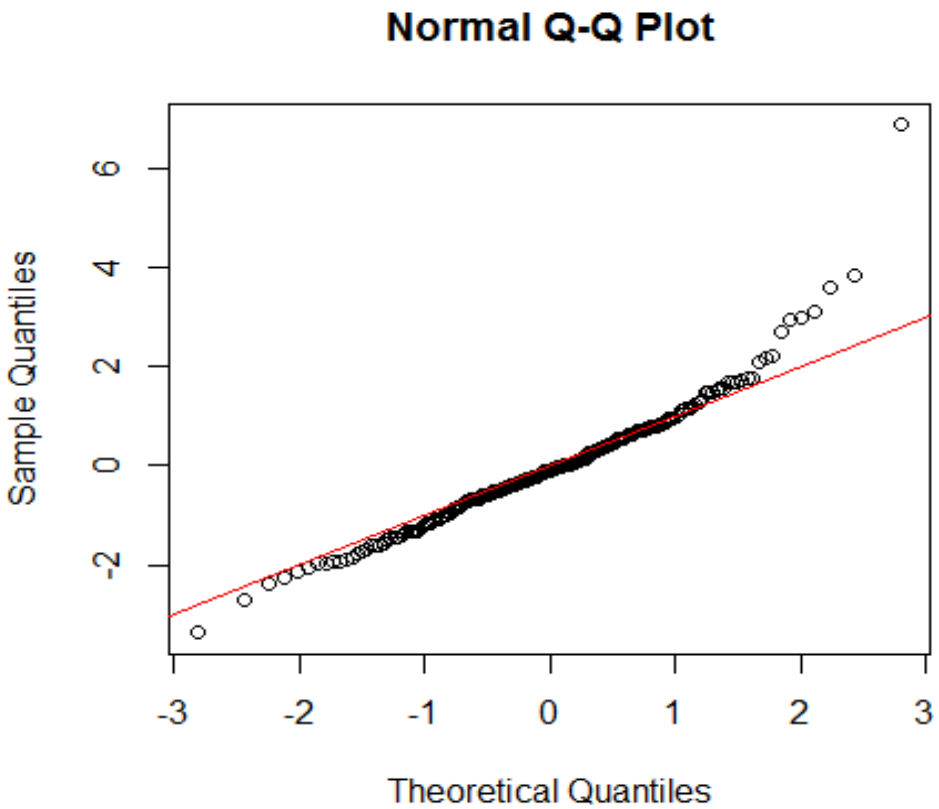
# Residual Plot



# Residual Plot



# QQPlot


Normal Q-Q Plot
Sample Quantiles
-2
0
2
4
6
-3
-2
-1
0
1
2
3
Theoretical Quantiles