228.371 - Statistical Modelling for Engineers and Technologists

Week 9. Logistic Regression

Dr. Daniel Walsh
IIMS 3.07 x 41032
d.c.walsh@massey.ac.nz

Semester One - 2015

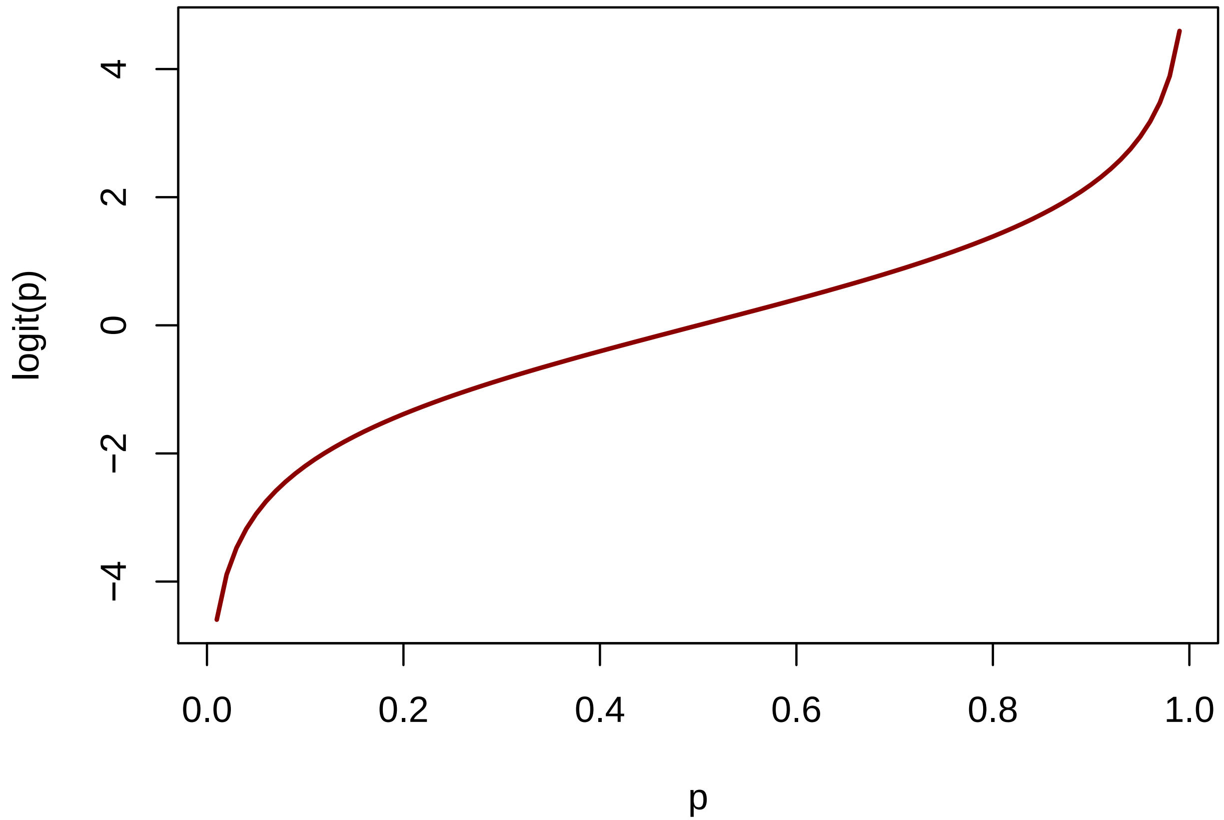
Logit function

$$g(\mathbb{E}(y_i)) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_q x_{qi}$$

Note that when y_i is 0/1 $E(Y_i)$ is the proportion p of 1s.

$$g(p) = \text{logit}\,(p) = \log\left(\frac{p}{1-p}\right).$$

Logit function



$$\lim_{p \to 1} \text{logit}\,(p) = \infty$$

$$\lim_{p \to 0} \text{logit}\,(p) = -\infty$$

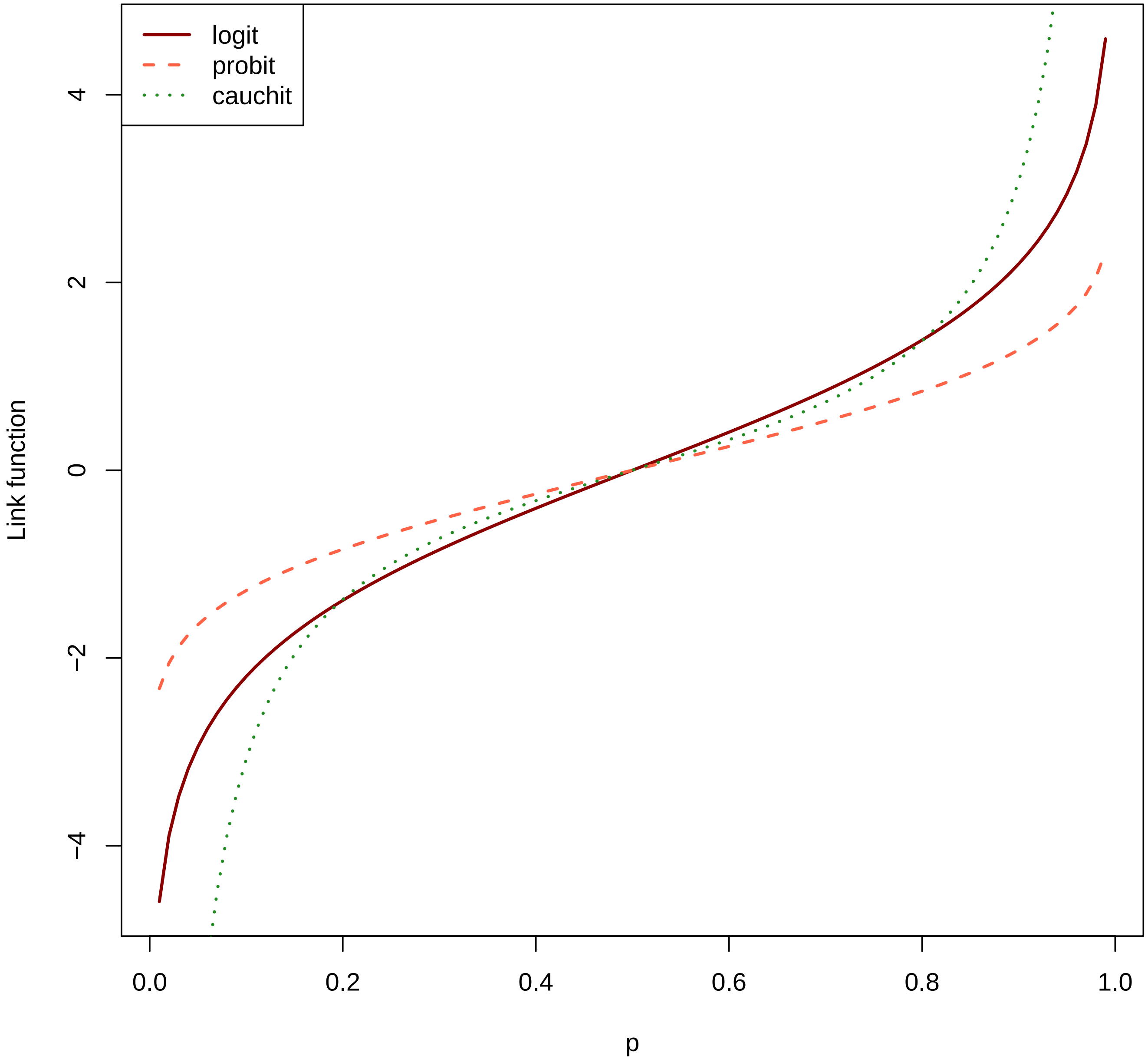
Link Functions

- g is called the link function
- For binomial data, g needs to map (0,1) onto the real line.
- There are other options besides the logit.
- Note that all inverse cumulative distribution functions for random variables on the real line do this.

Inverse CDF

- Recall that the CDF of a probability distribution $F(x) = \Pr(X \leq x)$.
- Thus $F(x)$ takes a real number and produces a number between 0 and 1; $F^{-1}(p)$ takes a number between 0 and 1 and gives back a number between $-\infty, \infty$
- We will use the inverse normal CDF (probit) denoted $\Phi^{-1}$and inverse Cauchy CDF (cauchit; equivalent to a t distribution with df$=$1).
- These functions do not have closed form and are computed numerically (R functions `qnorm (p); qt (p,df=1)`)

Link functions

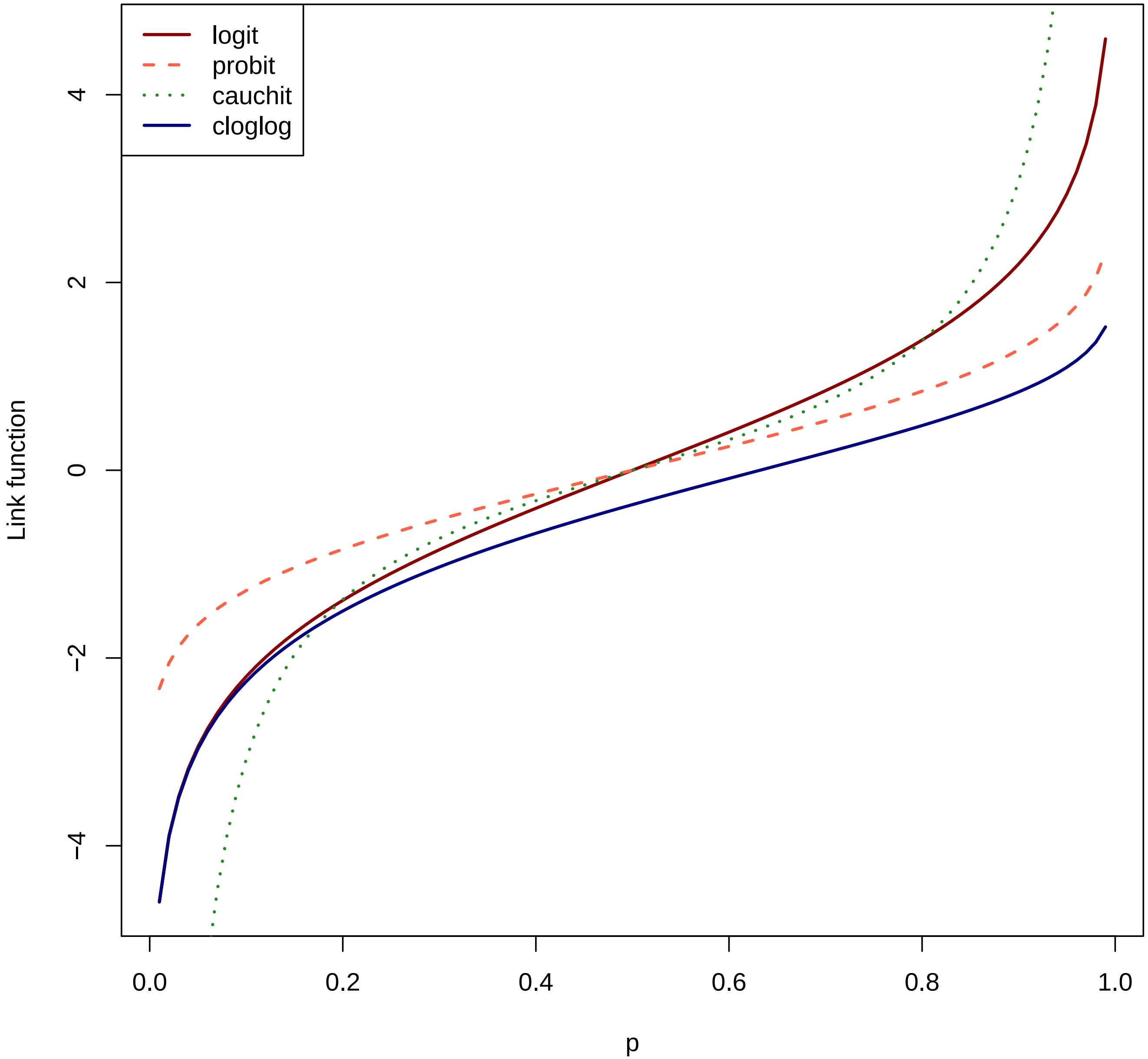


Complementary log-log function

$$g(p) = \log(-\log(1-p))$$

- Unlike other link functions, not symmetric around 0.5
- Developed for dilution series of bacterial cultures.

Link functions



Example: `fir.txt`

```
fir <- read.table (file="Data/fir.txt", header=TRUE)
m1 <- glm( y ~ log(dia), data=fir, family=binomial(link=logit)   )  ## Default
m2 <- glm( y ~ log(dia), data=fir, family=binomial(link=probit)  )
m3 <- glm( y ~ log(dia), data=fir, family=binomial(link=cauchit) )
m4 <- glm( y ~ log(dia), data=fir, family=binomial(link=cloglog) )
```

Compare with AIC or deviance (smaller is better).

Example: `fir.txt`

- Compare with AIC or equivalently deviance (smaller is better). Note degrees of freedom/dimension of model are not changing.
- Frequently there is not much difference between models. Logit has many useful features (interpretation in terms of odds, case-control equivalence for β_1 etc.)
- $\exp((AIC_{min} - AIC_i)/2)$ can be interpreted as the relative probability that the i^{th} model minimises the (estimated) information loss.
- If δ is the AIC difference, $\delta \leq 2$ is not worth mentioning, $2 \leq \delta \leq 6$ is weak, $\delta \geq 6$ is warrants consideration.

Model Comparison

```
M <- rbind (c(m1$aic, m1$deviance),
            c(m2$aic, m2$deviance),
            c(m3$aic, m3$deviance),
            c(m4$aic, m4$deviance))
rownames (M) <- c(m1$call$family, m2$call$family,
                  m3$call$family, m4$call$family)
colnames (M) <- c("AIC", "Deviance")
print (cbind (M, (M[,1] - min(M[,1]))/2))
```

```
                             AIC Deviance
binomial(link = logit)   659.2420 655.2420 0.3188929
binomial(link = probit)  660.3883 656.3883 0.8920381
binomial(link = cauchit) 658.6042 654.6042 0.0000000
binomial(link = cloglog) 673.0692 669.0692 7.2324930
```

Here, exclude cloglog model; others roughly equivalent. Not worth changing from logit based on fit alone.

Poisson

- Recall the Poisson distribution, used for counts of (relatively) rare events in time intervals of a fixed length.

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, 3\ldots$$

- Mean λ, var$=\lambda$, sd$=\sqrt{\lambda}$.
- As for the binomial, the possible values of Y (non negative integers) make linear regression unsuitable
- Also like binomial, variance changes with mean.

Poisson

Use glm framework:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots \beta_q x_{qi}$$

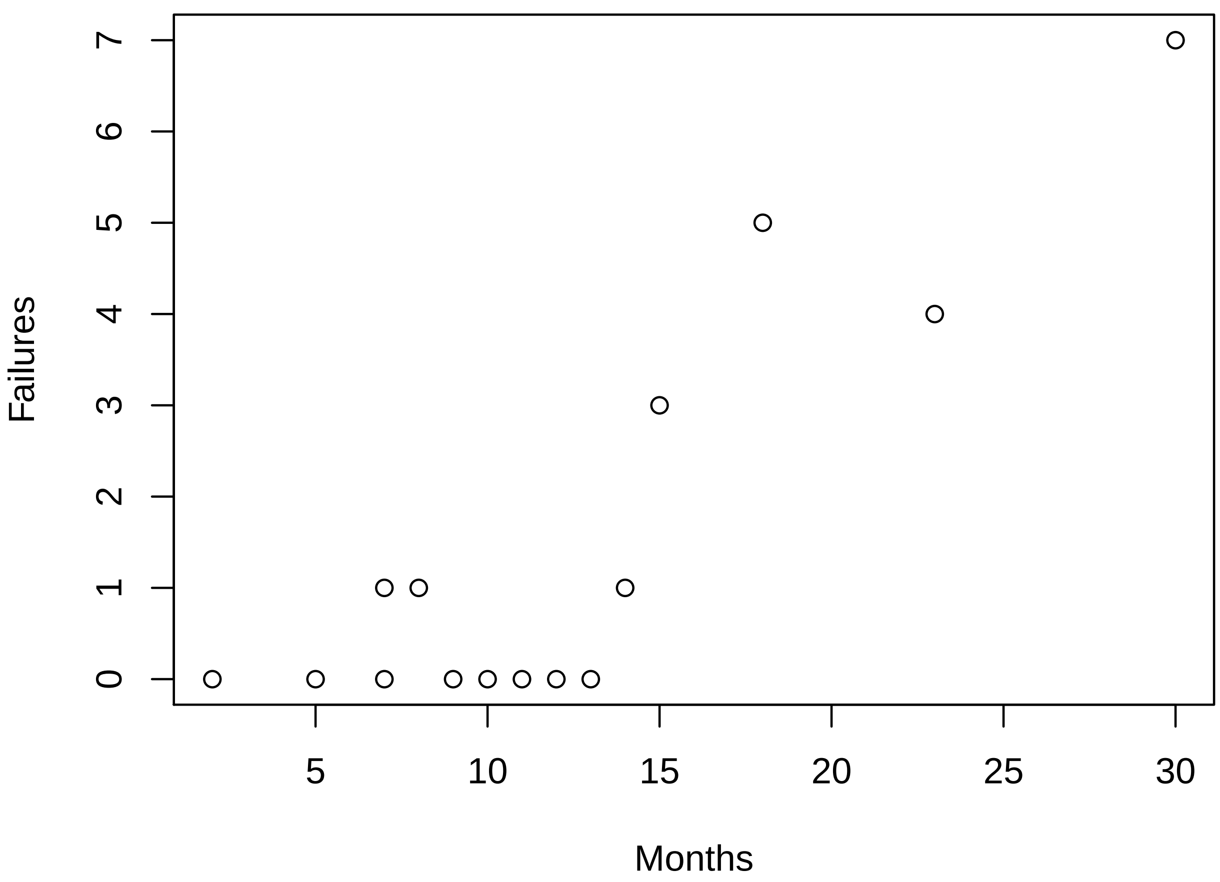
y_i is then Poisson with mean λ.

Link is log function.

Example: Valve failure

Number of valve failures after a number of months.

```
Failures <- c( 5, 3, 0, 1, 4, 0, 0, 1, 0, 0, 0, 1, 0, 7, 0)
Months   <- c(18,15,11,14,23,10, 5, 8, 7,12,13, 7, 2,30, 9)
plot (Failures ~ Months)
```

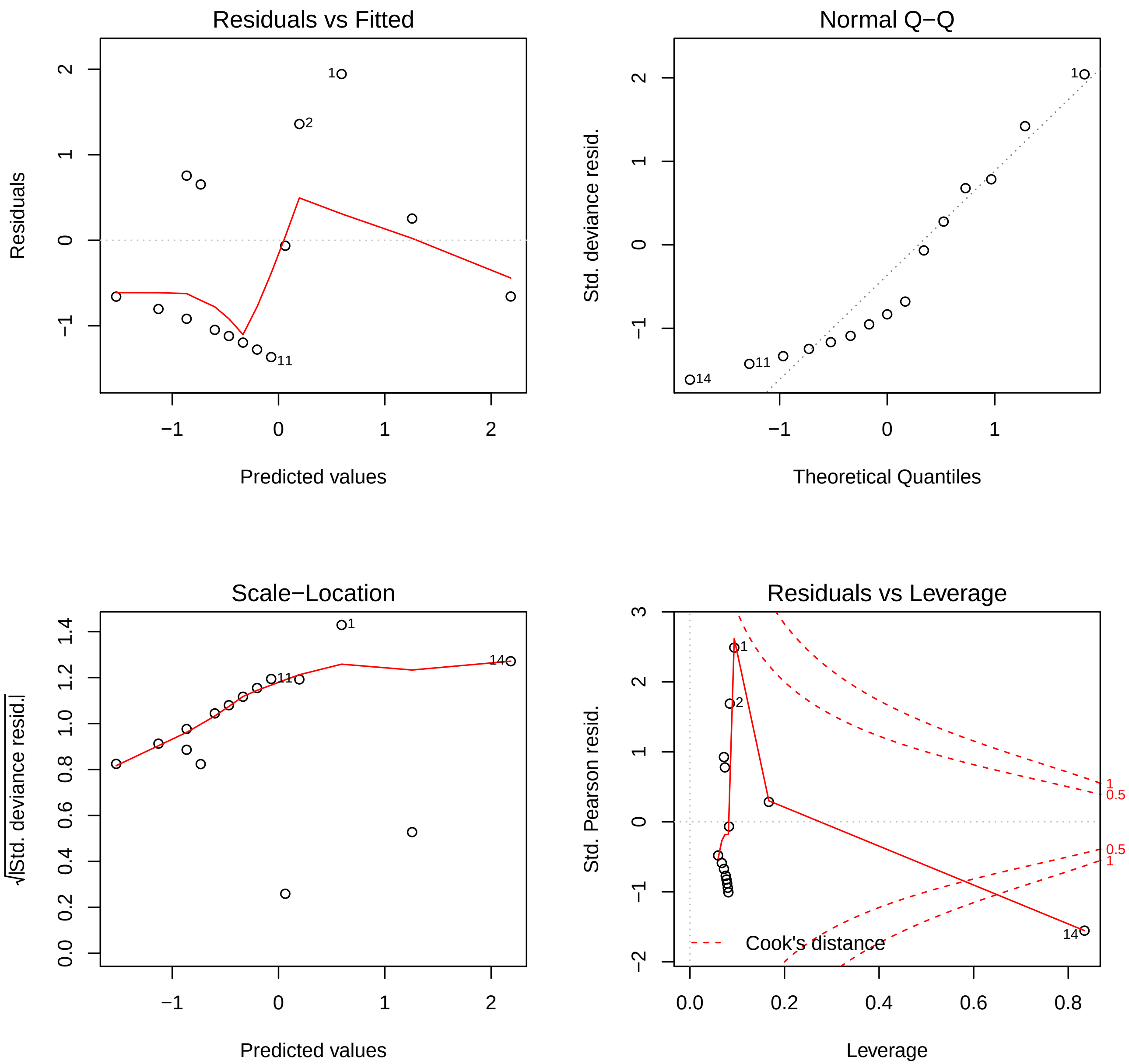


Example: Valve failure

```
m1 <- glm ( Failures ~ Months, family="poisson")
summary (m1)
```

```
Call:
glm(formula = Failures ~ Months, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3662  -1.0839  -0.6590   0.4532   1.9438

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.79238    0.57000  -3.145  0.00166 **
Months       0.13256    0.02496   5.312 1.09e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 44.167  on 14  degrees of freedom
Residual deviance: 16.334  on 13  degrees of freedom
AIC: 39.88

Number of Fisher Scoring iterations: 5
```

Diagnostic plots



Does an additional term help?

```
m2 <- glm ( Failures ~ Months + I(Months^2), family="poisson")
summary (m2)
```

```
Call:
glm(formula = Failures ~ Months + I(Months^2), family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4328  -0.8836  -0.4083   0.5240   1.3650

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.581523   1.836187  -2.495   0.0126 *
Months       0.459974   0.192534   2.389   0.0169 *
I(Months^2) -0.008119   0.004625  -1.755   0.0792 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 44.167  on 14  degrees of freedom
Residual deviance: 12.850  on 12  degrees of freedom
AIC: 38.396

Number of Fisher Scoring iterations: 6
```

Model comparison

```
anova(m1, m2, test="Chisq")
Analysis of Deviance Table

Model 1: Failures ~ Months
Model 2: Failures ~ Months + I(Months^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        13     16.334
2        12     12.850  1   3.4835  0.06198 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: `coal.csv`

- Response is number of fractures in a coal seam above another coal seam that has been mined.
- Predictors are:
  `InnerBurden`, distance between seams (feet),
  `PctExtraction`, % extraction of the lower, previously mined seam,
  `LowerHeight`, lower seam height (feet), and
  `Time`, time that the mine has been in operation (years).

```
coal <- read.csv ("Data/coal.csv", header=TRUE)
summary (coal)
```

```
  Fractures       InnerBurden      PctExtraction    LowerHeight          Time
Min.   :0.000   Min.   : 11.0   Min.   :50.00   Min.   : 36.00   Min.   : 0.000
1st Qu.:1.000   1st Qu.: 65.0   1st Qu.:65.00   1st Qu.: 42.00   1st Qu.: 0.875
Median :2.000   Median :132.5   Median :80.00   Median : 51.00   Median : 5.000
Mean   :2.227   Mean   :169.2   Mean   :75.93   Mean   : 56.64   Mean   : 7.273
3rd Qu.:3.250   3rd Qu.:195.0   3rd Qu.:85.00   3rd Qu.: 60.50   3rd Qu.:10.000
Max.   :5.000   Max.   :900.0   Max.   :90.00   Max.   :165.00   Max.   :35.000
```

Coal data

```
m1 <- glm ( Fractures ~ InnerBurden + PctExtraction + LowerHeight + Time,
            family=poisson, data=coal)
m2 <- step(m1)
summary(m2)
```

```
...
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.7403253  0.9799904  -3.817 0.000135 ***
InnerBurden   -0.0015217  0.0008216  -1.852 0.063991 .
PctExtraction  0.0629242  0.0122905   5.120 3.06e-07 ***
Time          -0.0296676  0.0154712  -1.918 0.055163 .
```

Model Selection

AIC has included some parameters that are not significant based on the Wald test. Does the (preferred) likelihood ratio test agree?

```
m3 <- glm( Fractures ~ PctExtraction + Time, family=poisson, data=coal)
m4 <- glm( Fractures ~ PctExtraction + InnerBurden, family=poisson, data=coal)
anova(m2,m3,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: Fractures ~ InnerBurden + PctExtraction + Time
Model 2: Fractures ~ PctExtraction + Time
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        40     38.089
2        41     41.952 -1  -3.8637  0.04934 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m2,m4,test="Chisq")
```

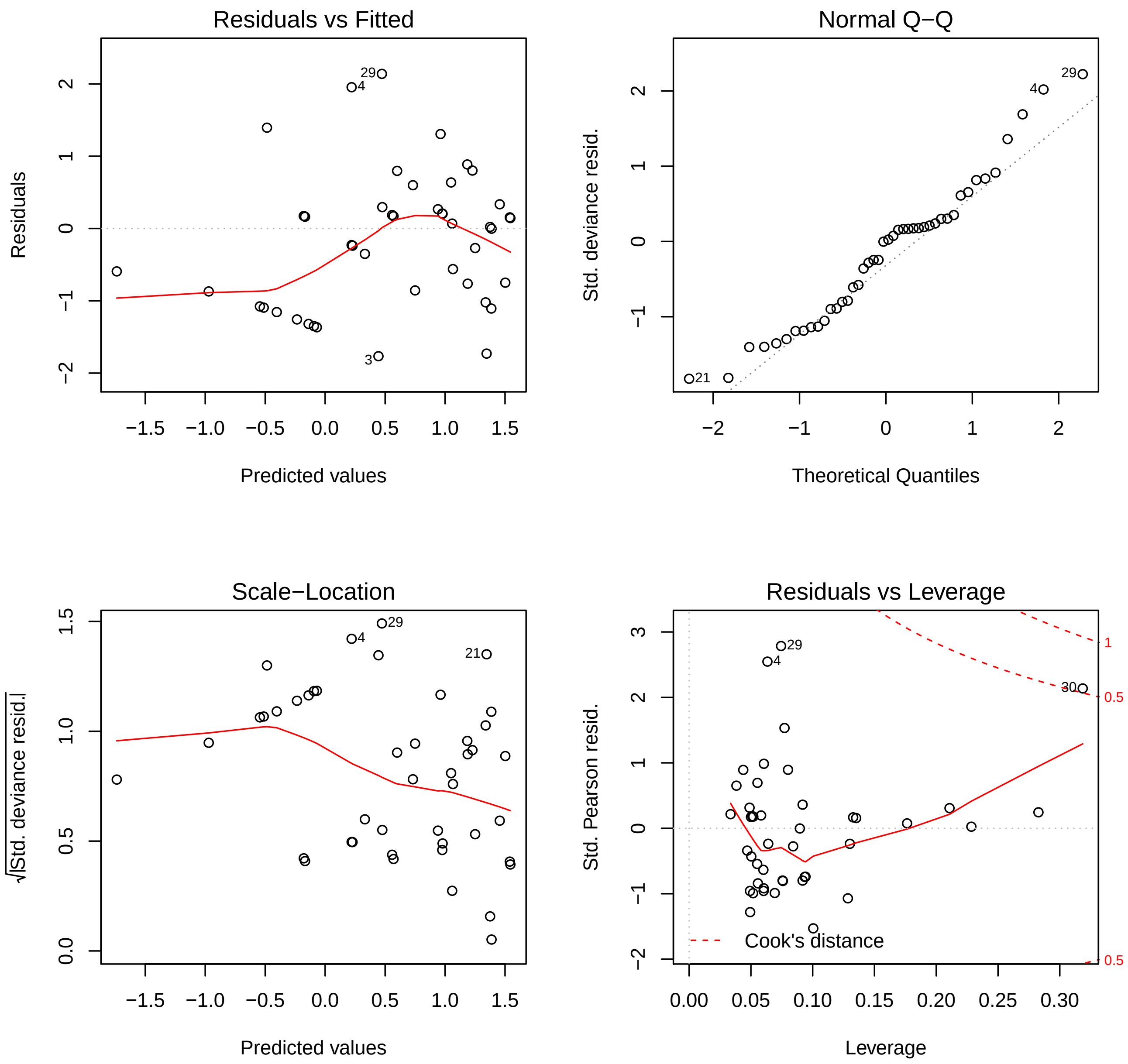
```
Analysis of Deviance Table

Model 1: Fractures ~ InnerBurden + PctExtraction + Time
Model 2: Fractures ~ PctExtraction + InnerBurden
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        40     38.089
2        41     42.094 -1  -4.0052  0.04536 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Adequacy

LRT suggest the three variable model. Does this provide a good fit?

```
m2$df.resid
[1] 40
m2$deviance
[1] 38.08851
pchisq(38.089,40, lower=FALSE)
[1] 0.5565525
## Large p-value indicates good fit
```

Diagnostic Plots: `coal.csv : m2`



Model Adequacy

No serious problems, but if we wanted to check out some of the numbered points we could do it like this:

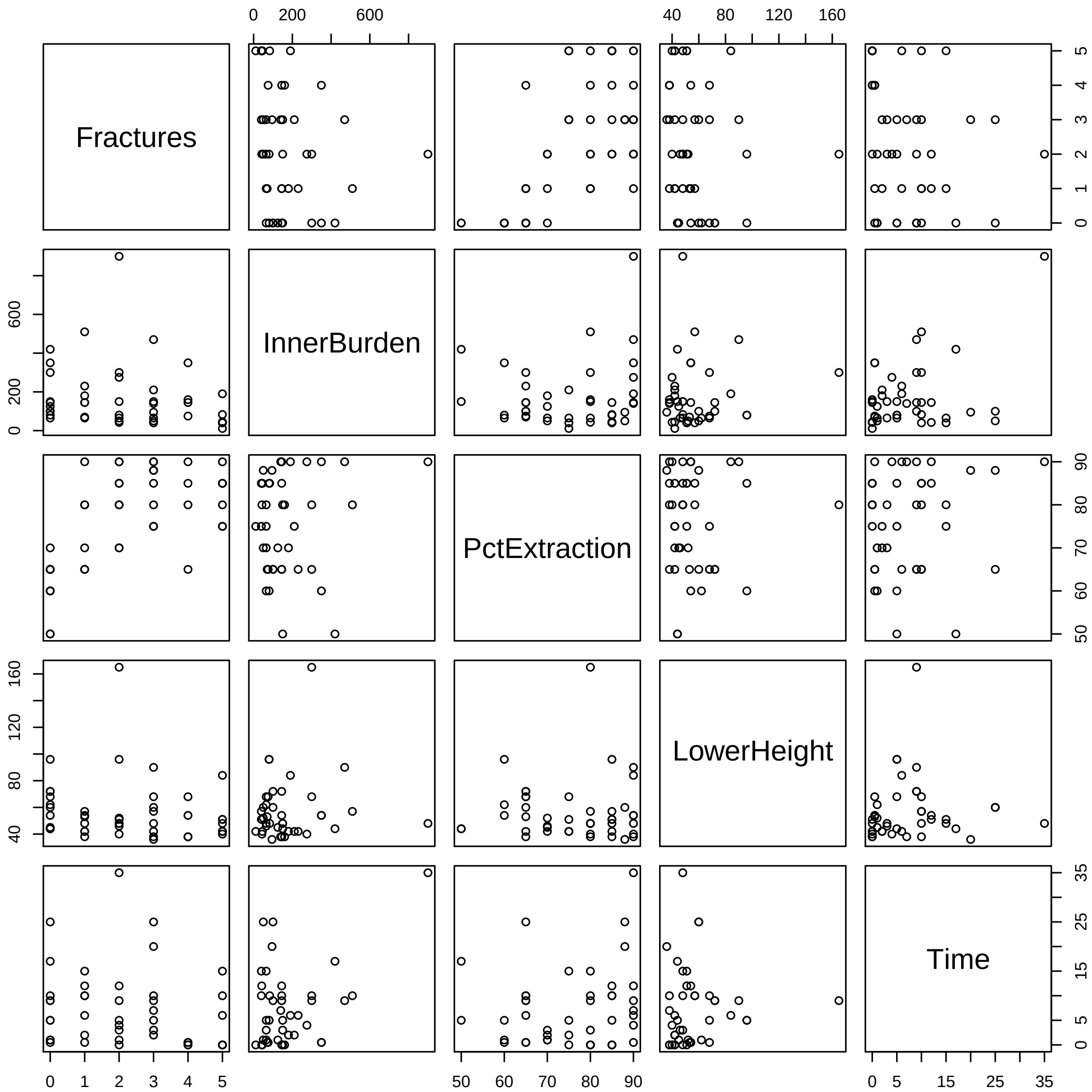
```
coal[c(3,4,21,29,30),]
```

```
   Fractures InnerBurden PctExtraction LowerHeight Time
3          0         125            70          45  1.0
4          4          75            65          68  0.5
21         1         145            90          54 12.0
29         5          40            75          51 15.0
30         2         900            90          48 35.0
```

```
pairs(coal)
```

Comparison with the pairs plot shows point 30 is the oldest mine in the sample and has an unusually high Inner Burden. This leads to a moderately (but not critically) large Cook's distance.

Pairs Plot: `coal.csv`



Example: aircraft.txt

Aircraft damage data.

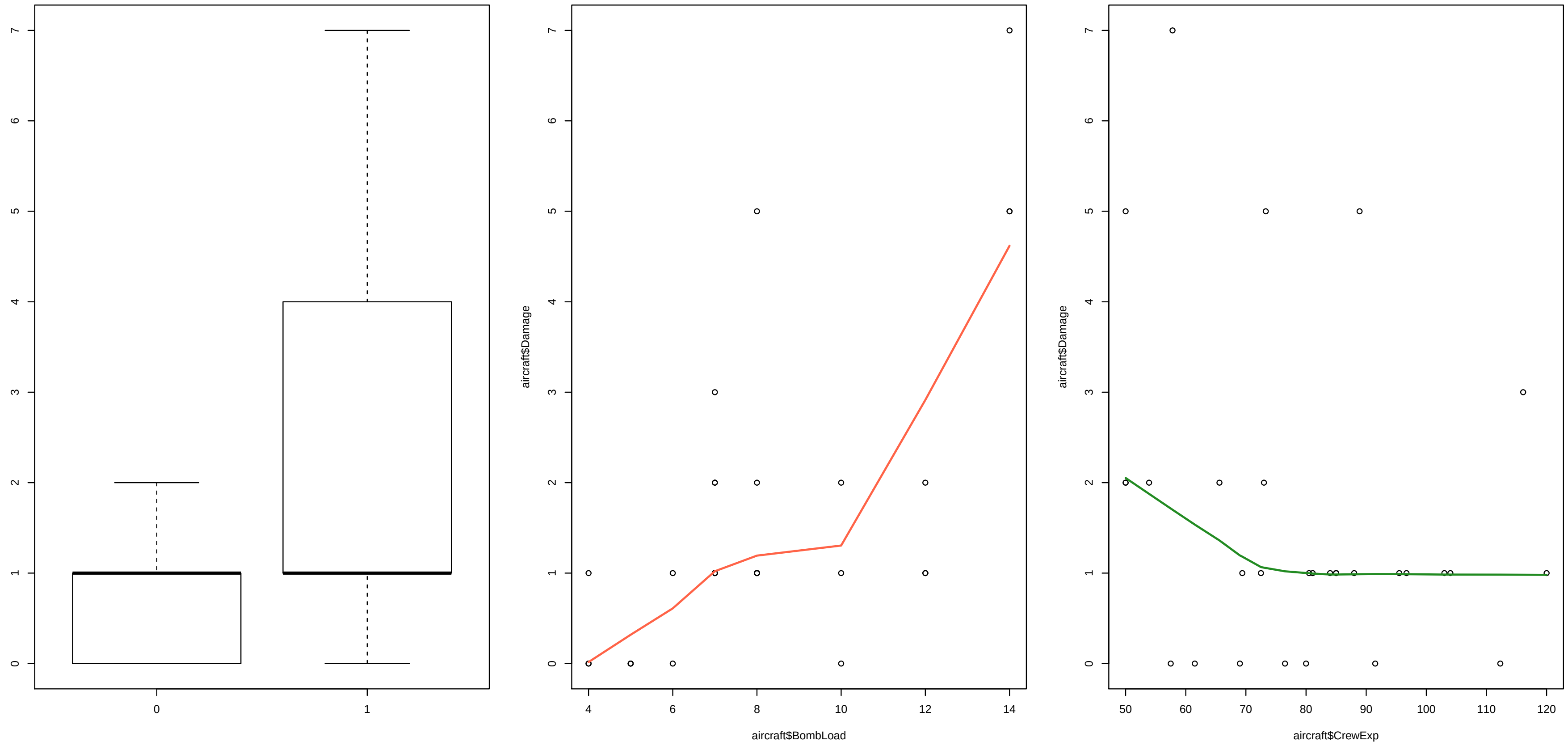
```
aircraft <- read.table ("Data/aircraft.txt", header=TRUE)
summary (aircraft)
```

```
     Damage          Aircraft        BombLoad          CrewExp      
 Min.   :0.0    Min.   :0.0    Min.   : 4.000   Min.   : 50.00  
 1st Qu.:1.0    1st Qu.:0.0    1st Qu.: 6.250   1st Qu.: 66.45  
 Median :1.0    Median :0.5    Median : 7.500   Median : 80.25  
 Mean   :1.6    Mean   :0.5    Mean   : 8.133   Mean   : 79.72  
 3rd Qu.:2.0    3rd Qu.:1.0    3rd Qu.:10.000   3rd Qu.: 90.85  
 Max.   :7.0    Max.   :1.0    Max.   :14.000   Max.   :120.00  
```

```
cor (aircraft)
```

```
             Damage  Aircraft    BombLoad     CrewExp
Damage    1.0000000 0.4639468  0.66382721 -0.26265437
Aircraft  0.4639468 1.0000000  0.70545702  0.21804716
BombLoad  0.6638272 0.7054570  1.00000000 -0.02244671
CrewExp  -0.2626544 0.2180472 -0.02244671  1.00000000
```

Example: aircraft.txt



Example: aircraft.txt

Stepwise selection from empty model.

```
m0 <- glm ( Damage ~ 1, data=aircraft, family=poisson)
m1 <- step( m0, scope = ~ Aircraft*BombLoad*CrewExp, direction="both")
summary (m1)
```

```
Call:
glm(formula = Damage ~ BombLoad + CrewExp, family = poisson,
    data = aircraft)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.370166   0.799445  -0.463   0.6433
BombLoad     0.209737   0.045333   4.627 3.72e-06 ***
CrewExp     -0.014024   0.008225  -1.705   0.0882 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 50.860  on 29  degrees of freedom
Residual deviance: 25.508  on 27  degrees of freedom
AIC: 87.817
```

Example: aircraft.txt

Stepwise selection from full model.

```
m0 <- glm ( Damage ~  Aircraft*BombLoad*CrewExp, data=aircraft, family=poisson)
m2 <- step( m0, scope = ~ Aircraft*BombLoad*CrewExp, direction="both")
summary (m2)
```

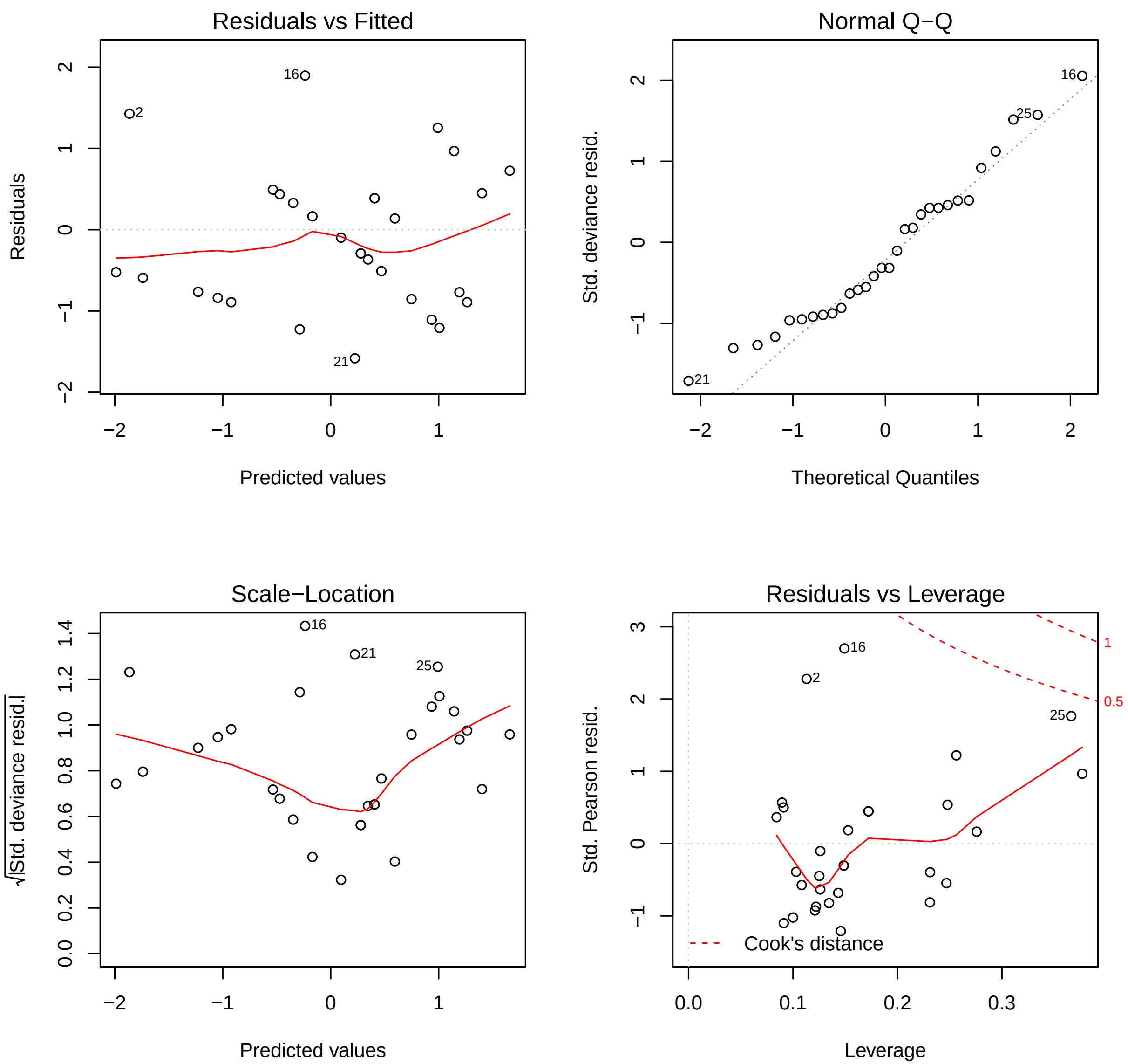
```
Call:
glm(formula = Damage ~ Aircraft + BombLoad + CrewExp + Aircraft:BombLoad,
    family = poisson, data = aircraft)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.746339   1.907928  -1.439   0.1500
Aircraft           3.504680   2.039303   1.719   0.0857 .
BombLoad           0.568818   0.263155   2.162   0.0307 *
CrewExp           -0.016580   0.008061  -2.057   0.0397 *
Aircraft:BombLoad -0.436066   0.272867  -1.598   0.1100
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 50.860  on 29  degrees of freedom
Residual deviance: 21.841  on 25  degrees of freedom
AIC: 88.151
```

Example: aircraft.txt

Very different models produced. Anova, AIC suggest smaller model

```
anova(m1, m2, test="Chisq")
Analysis of Deviance Table

Model 1: Damage ~ BombLoad + CrewExp
Model 2: Damage ~ Aircraft + BombLoad + CrewExp + Aircraft:BombLoad
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        27     25.508
2        25     21.841  2   3.6662   0.1599
```

Diagnostic Plots: `aircraft.csv : m1`



OK for both models.

Example: aircraft.txt

However, interaction between bombload and plane type is conceptually appealing; (sequential) testing makes Aircraft type and look very important.

```
anova(m2,test="Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: Damage

Terms added sequentially (first to last)


                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                  29     50.860
Aircraft            1  12.5580        28     38.302 0.0003945 ***
BombLoad            1   9.9503        27     28.352 0.0016083 **
CrewExp             1   3.5524        26     24.799 0.0594579 .
Aircraft:BombLoad   1   2.9581        25     21.841 0.0854472 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: aircraft.txt

Variance inflation factors indicate problems with the larger model, due to the interaction

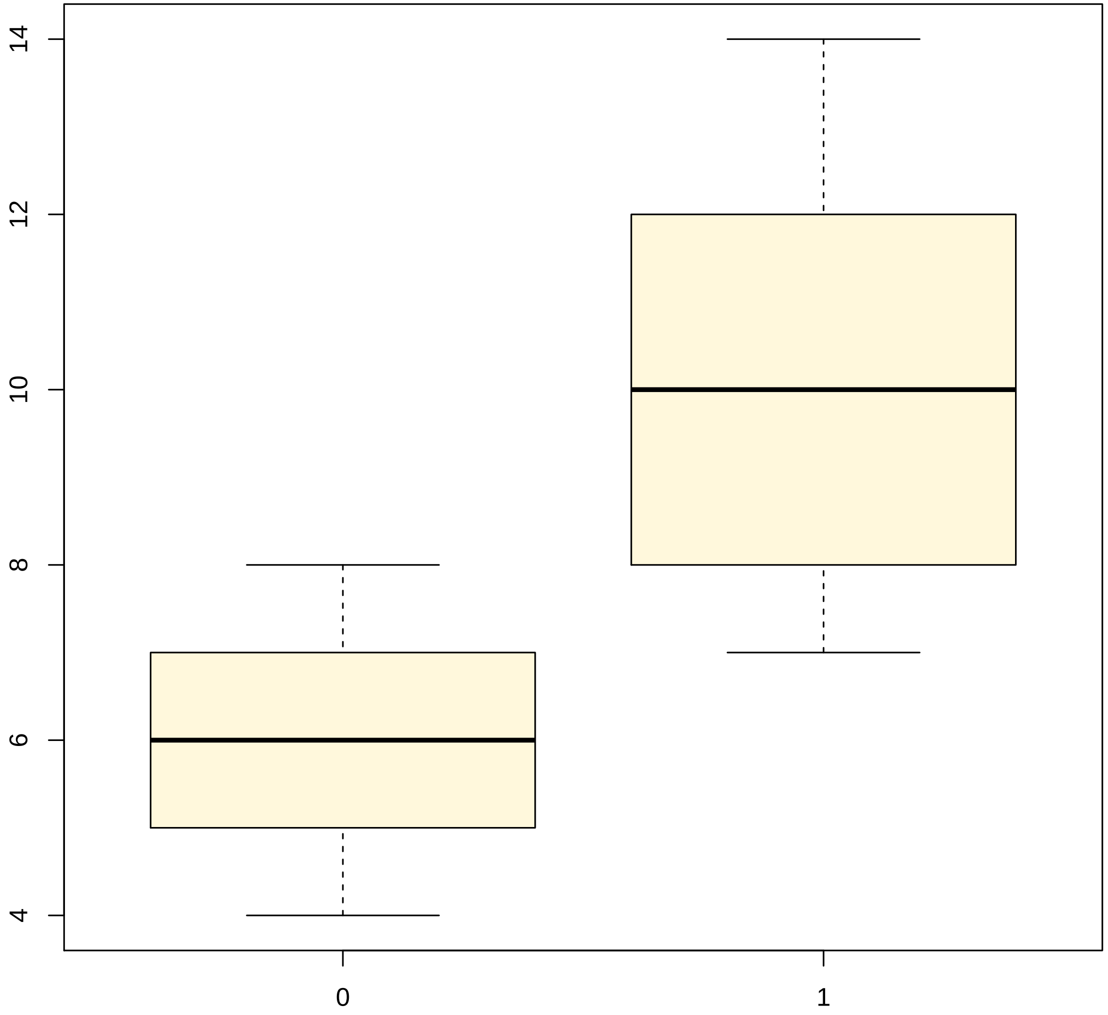
```
library("car")
vif(m2)
```

```
       Aircraft          BombLoad           CrewExp Aircraft:BombLoad
      37.429876         27.852066          1.045448        102.092677
```

```
vif(m3)
```

```
PctExtraction          Time
     1.036499      1.036499
```

Example: `aircraft.txt` - Bombload and Aircraft



Example: aircraft.txt - Final model

Boxplot indicates aircraft type and Bombload are strongly related, indicating why aircraft type initially appeared important.

```
 anova(m1, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: Damage

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        29     50.860
BombLoad  1  22.3381        28     28.522 2.286e-06 ***
CrewExp   1   3.0146        27     25.508   0.08252 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 pchisq(28.52, 28, lower=FALSE)
[1] 0.4371648
 ## Large p-value indicates overall fit is good
```