Logistic Regression

References:

Applied Linear Statistical Models, Neter et al.

Categorical Data Analysis, Agresti

Slides prepared by Elizabeth Newton (MIT)

Logistic Regression

- Nonlinear regression model when response variable is qualitative.
- 2 possible outcomes, success or failure, diseased or not diseased, present or absent
- Examples: CAD (y/n) as a function of age, weight, gender, smoking history, blood pressure
- Smoker or non-smoker as a function of family history, peer group behavior, income, age
- Purchase an auto this year as a function of income, age of current car, age

Response Function for Binary Outcome

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$E\{Y_i\} = \beta_0 + \beta_1 X_i$$

$$P(Y_i = 1) = \pi_i$$

$$P(Y_i = 0) = 1 - \pi_i$$

$$E\{Y_i\} = 1(\pi_i) + 0(1 - \pi_i) = \pi_i$$

$$E\{Y_i\} = \beta_0 + \beta_1 X_i = \pi_i$$

Special Problems when Response is Binary

Constraints on Response Function

$0 \leq E\{Y\} = \pi \ = \leq 1$

Non-normal Error Terms

When $Y_i=1$: $\varepsilon_i = 1-\beta_0-\beta_1 X_i$

When $Y_i=0$: $\varepsilon_i = -\beta_0-\beta_1 X_i$

Non-constant error variance

$Var\{Y_i\} = Var\{\varepsilon_i\} = \pi_i(1-\pi_i)$

Logistic Response Function

$$E\{Y\} = \pi = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

$$\pi(1 + \exp(\beta_0 + \beta_1 X)) = \exp(\beta_0 + \beta_1 X)$$

$$\pi + \pi \exp(\beta_0 + \beta_1 X) = \exp(\beta_0 + \beta_1 X)$$

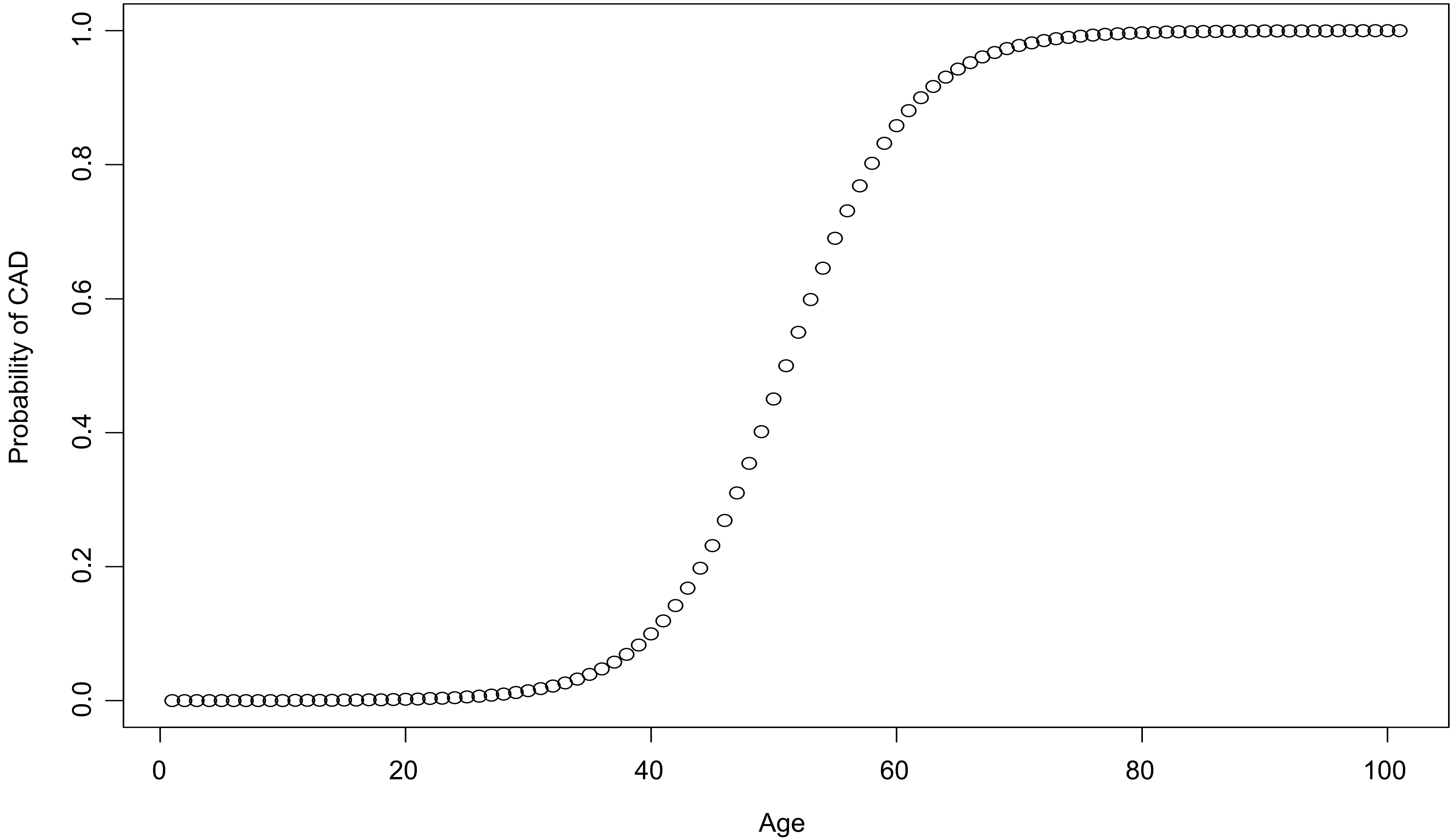
$$\pi = \exp(\beta_0 + \beta_1 X) - \pi \exp(\beta_0 + \beta_1 X)$$

$$\pi = (1 - \pi)\exp(\beta_0 + \beta_1 X)$$

$$\frac{\pi}{1 - \pi} = \exp(\beta_0 + \beta_1 X)$$

$$\log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X$$

Example of Logistic Response Function



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Properties of Logistic Response Function

$\log(\pi/(1-\pi))$=logit transformation, log odds

$\pi/(1-\pi)$ = odds

Logit ranges from $-\infty$ to ∞ as x varies from $-\infty$ to ∞

Likelihood Function

$$P(Y_i = 1) = \pi_i$$

$$P(Y_i = 0) = 1 - \pi_i$$

$$pdf : f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}, \mathrm{Y_i} = 0,1; \mathrm{i} = 1,2...\mathrm{n}$$

Since $\mathrm{Y_i}$ *a*re independent, joint pdf is;

$$\mathrm{g}(\mathrm{Y_i}...Y_n) = \Pi_{i=1}^{n} f_i(Y_i) = \Pi_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$\log \mathrm{g}(\mathrm{Y_i}...Y_n) = \sum_{i=1}^{n}[Y_i \log(\frac{\pi_i}{1-\pi_i})] + \sum_{i=1}^{n}\log(1-\pi_i)$$

Likelihood Function (continued)

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

$$1-\pi_i = \frac{1}{1+\exp(\beta_0 + \beta_1 X_i)}$$

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log[1+\exp(\beta_0 + \beta_1 X_i)]$$

Likelihood for Multiple Logistic Regression

$$\log L(\beta) = \sum_j (\sum_i y_i X_{ij})\beta_j - \sum_i \log[1 + \exp(\sum_j \beta_j x_{ij})]$$

$$\frac{\partial L}{\partial \beta_k} = \sum_i y_i x_{ik} - \sum_i x_{ik} [\frac{\exp(\sum_j \beta_j x_{ij})}{1 + \exp(\sum_j \beta_j x_{ij})}]$$

$$\text{Likelihood Equations}: \sum_i y_i x_{ik} = \sum_i x_{ik} [\frac{\exp(\sum_j \beta_j x_{ij})}{1 + \exp(\sum_j \beta_j x_{ij})}] = \sum_i \hat{\pi}_i x_{ik}$$

$$X' y = X' \hat{y}$$

Solution of Likelihood Equations

No closed form solution

Use Newton-Raphson algorithm

Iteratively reweighted least squares (IRLS)

Start with OLS solution for β at iteration t=0, β^0

$\pi_i^t=1/(1+\exp(-X_i'\beta^t))$

$\beta^{(t+1)}=\beta^t + (XVX)^{-1} X'(y-\pi^t)$

Where $V=\text{diag}(\pi_i^t(1-\pi_i^t))$

Usually only takes a few iterations

Interpretation of logistic regression coefficients

- $\text{Log}(\pi/(1-\pi))=X\beta$
- So each β_j is effect of unit increase in X_j on log odds of success with values of other variables held constant
- Odds Ratio=$\exp(\beta_j)$

Example: Spinal Disease in Children Data

SUMMARY:

The kyphosis data frame has 81 rows representing data on 81 children who have had corrective spinal surgery. The outcome Kyphosis is a binary variable, the other three variables (columns) are numeric.

ARGUMENTS:

Kyphosis

a factor telling whether a postoperative deformity (kyphosis) is "present" or "absent" .

Age

the age of the child in months.

Number

the number of vertebrae involved in the operation.

Start

the beginning of the range of vertebrae involved in the operation.

SOURCE:

John M. Chambers and Trevor J. Hastie, *Statistical Models in S,* Wadsworth and Brooks, Pacific Grove, CA 1992, pg. 200.

Observations 1:16 of kyphosis data set

- `kyphosis[1:16,]`

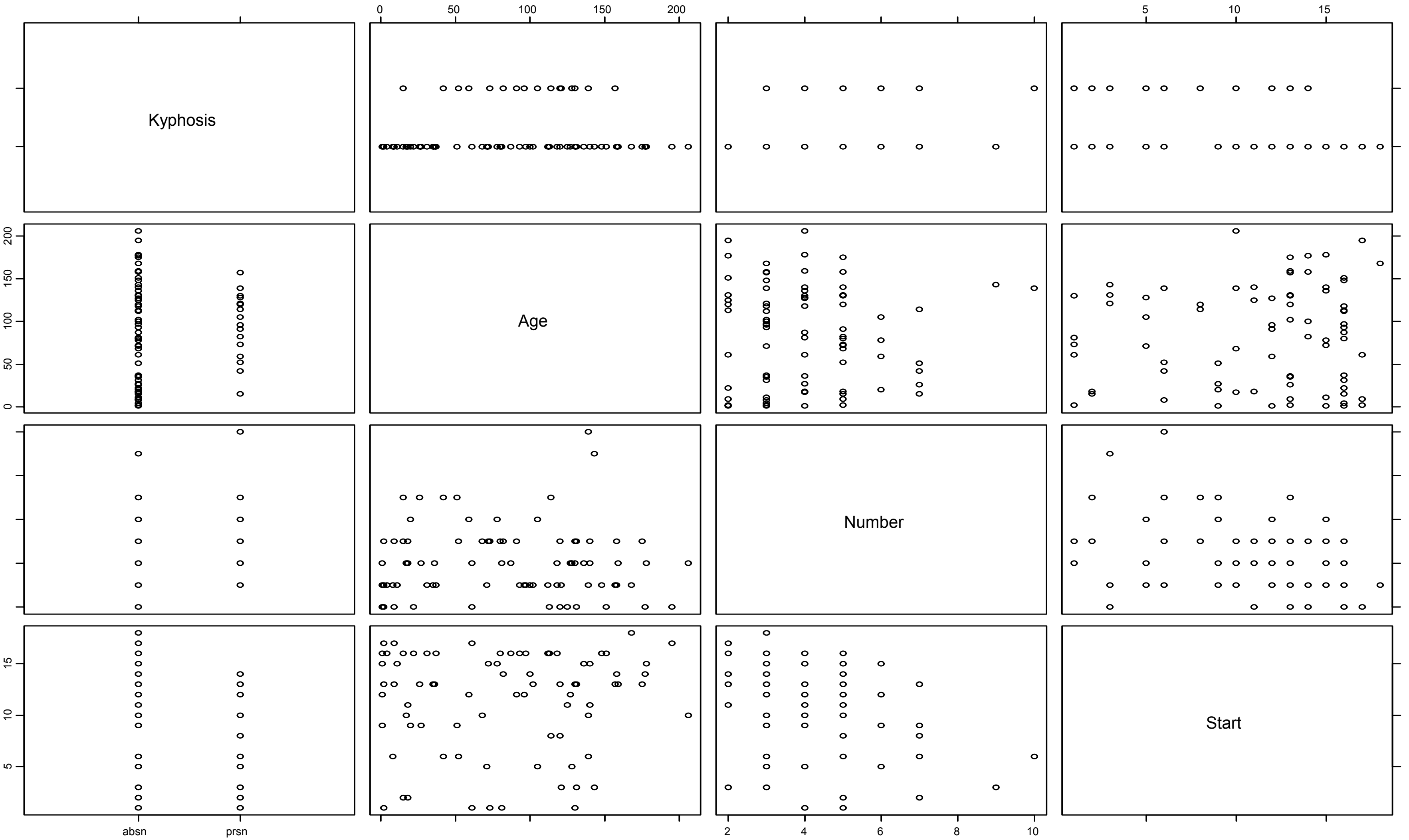
```
   Kyphosis Age Number Start
 1   absent  71      3     5
 2   absent 158      3    14
 3  present 128      4     5
 4   absent   2      5     1
 5   absent   1      4    15
 6   absent   1      2    16
 7   absent  61      2    17
 8   absent  37      3    16
 9   absent 113      2    16
10  present  59      6    12
11  present  82      5    14
12   absent 148      3    16
13   absent  18      5     2
14   absent   1      4    12
16   absent 168      3    18
```

Variables in kyphosis

- summary(kyphosis)

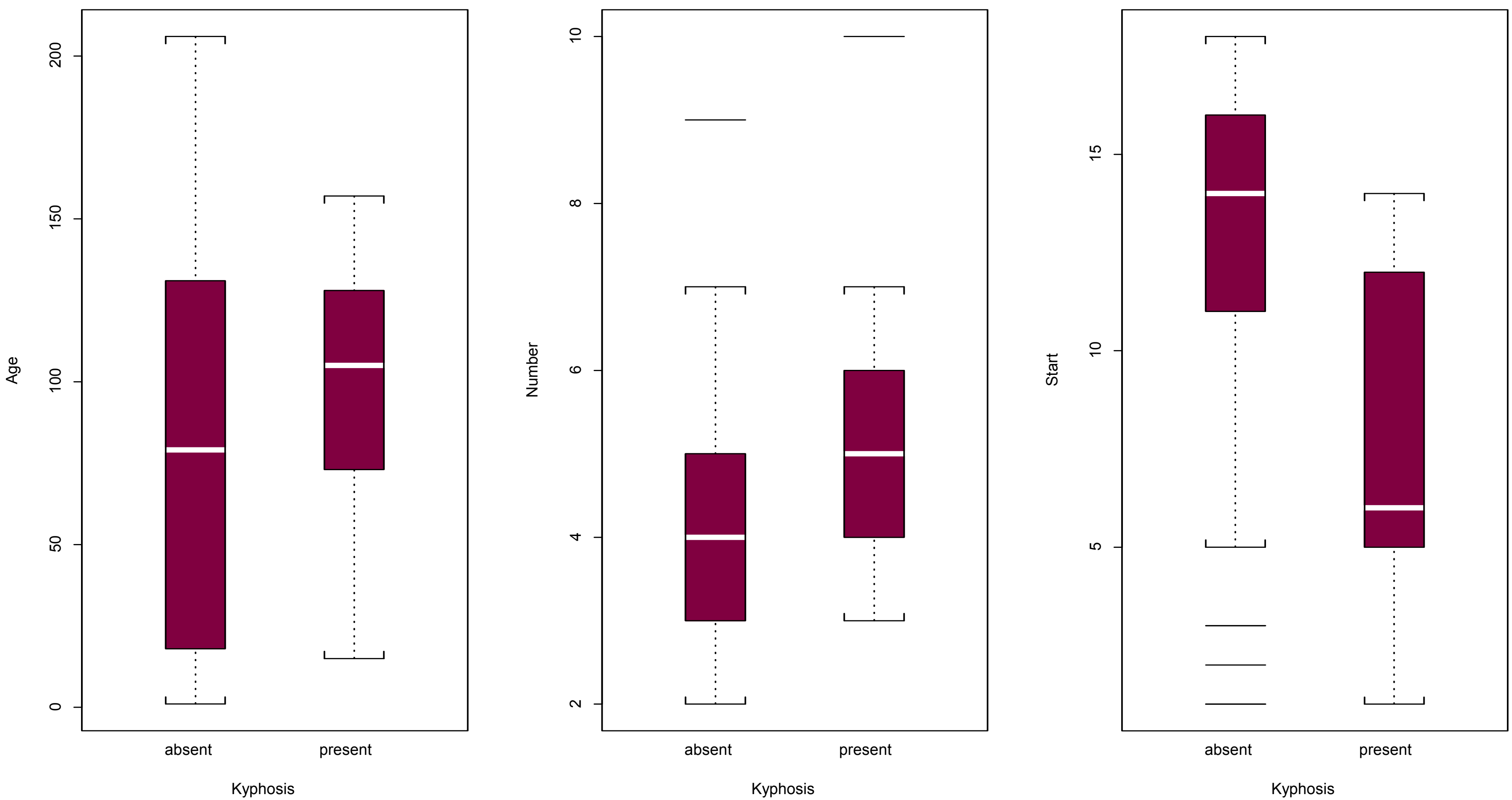
```
   Kyphosis          Age                Number              Start
 absent:64      Min.:  1.00       Min.: 2.000       Min.: 1.00
present:17   1st Qu.: 26.00    1st Qu.: 3.000    1st Qu.: 9.00
              Median: 87.00     Median: 4.000     Median:13.00
                Mean: 83.65       Mean: 4.049       Mean:11.49
             3rd Qu.:130.00    3rd Qu.: 5.000    3rd Qu.:16.00
                Max.:206.00       Max.:10.000       Max.:18.00
```

Scatter plot matrix kyphosis data set

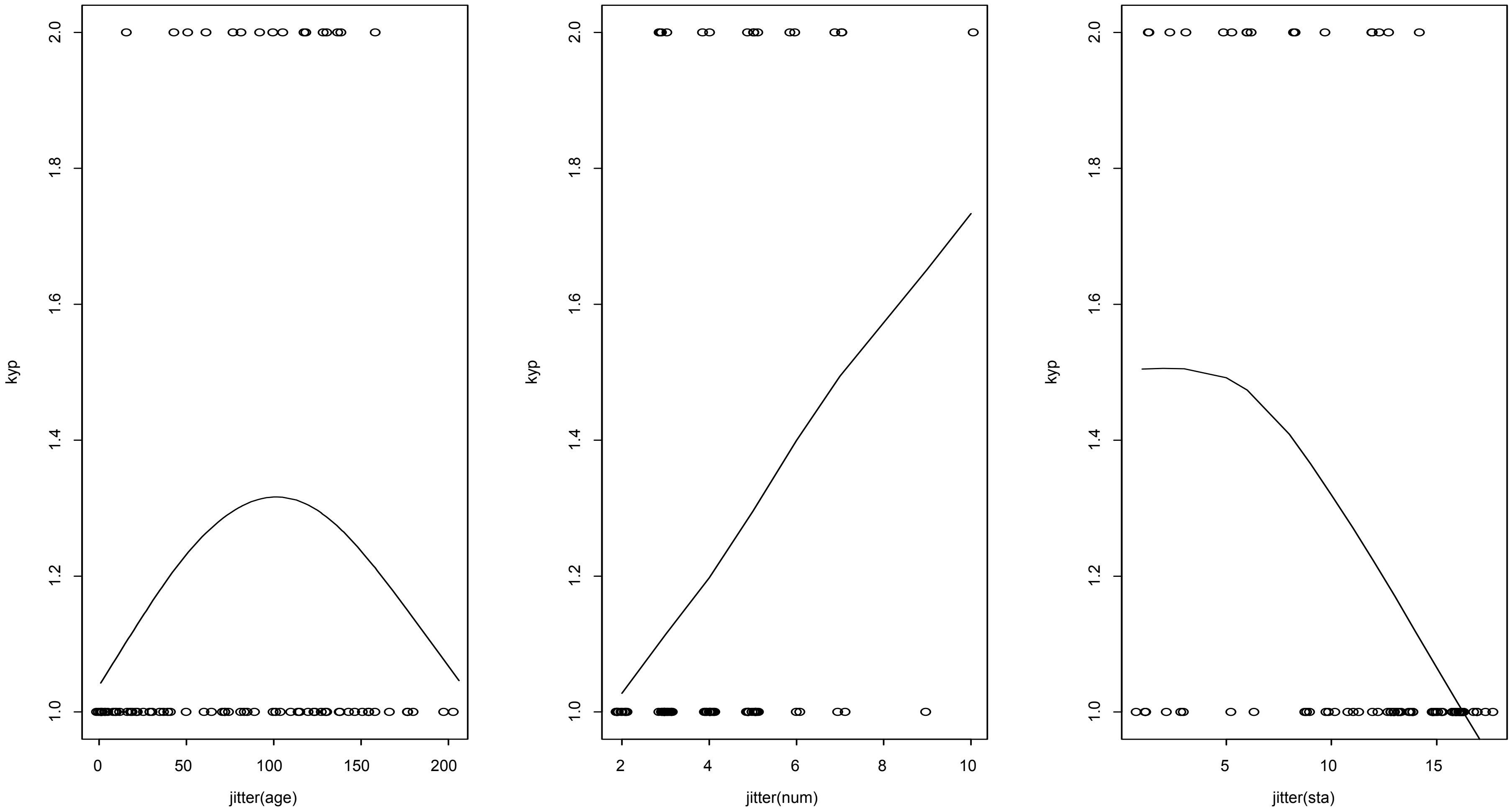


This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Boxplots of predictors vs. kyphosis



Smoothing spline fits, df=3



Summary of glm fit

```
Call: glm(formula = Kyphosis ~ Age + Number + Start,
  family = binomial, data = kyphosis)

Deviance Residuals:
       Min         1Q     Median         3Q     Max
 -2.312363 -0.5484308 -0.3631876 -0.1658653 2.16133

Coefficients:
                  Value Std. Error   t value
(Intercept) -2.03693225 1.44918287 -1.405573
        Age  0.01093048 0.00644419  1.696175
     Number  0.41060098 0.22478659  1.826626
      Start -0.20651000 0.06768504 -3.051043
```

Summary of glm fit

```
 Null Deviance: 83.23447 on 80 degrees of freedom

Residual Deviance: 61.37993 on 77 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:
       (Intercept)        Age     Number
   Age -0.4633715
Number -0.8480574   0.2321004
 Start -0.3784028  -0.2849547  0.1107516
```

Residuals

- Response Residuals: $y_i - \pi_i$

- Pearson Residuals: $(y_i - \pi_i)/\text{sqrt}(\pi_i(1-\pi_i))$

- Deviance Residuals: $\text{sqrt}(-2\log(|1-y_i-\pi_i|))$

Model Deviance

- Deviance of fitted model compares log-likelihood of fitted model to that of saturated model.
- Log likelihood of saturated model=0

$$DEV = -2\sum_{i=1}^{n} Y_i \log(\hat{\pi}_i) + (1 - Y_i)\log(1 - \hat{\pi}_i)$$

$$d_i = sign(Y_i - \hat{\pi}_i)\{-2[Y_i \log(\hat{\pi}_i) + (1 - Y_i)\log(1 - \hat{\pi}_i)]\}^{1/2}$$

$$\sum_i d_i^2 = DEV$$

Covariance Matrix

```
> x<-model.matrix(kyph.glm)

> xvx<-t(x)%*%diag(fi*(1-fi))%*%x

> xvx
            (Intercept)         Age     Number      Start
(Intercept)    9.620342    907.8887   43.67401   86.49845
        Age  907.888726 114049.8308 3904.31350 9013.14464
     Number   43.674014   3904.3135  219.95353  378.82849
      Start   86.498450   9013.1446  378.82849 1024.07328

> xvxi<-solve(xvx)
> xvxi
             (Intercept)            Age       Number        Start
(Intercept)  2.101402986 -0.00433216784 -0.2764670205 -0.0370950612
        Age -0.004332168  0.00004155736  0.0003368969 -0.0001244665
     Number -0.276467020  0.00033689690  0.0505664221  0.0016809996
      Start -0.037095061 -0.00012446655  0.0016809996  0.0045833534
> sqrt(diag(xvxi))
[1] 1.44962167 0.00644650 0.22486979 0.06770047
```

Change in Deviance resulting from adding terms to model

```
> anova(kyph.glm)
Analysis of Deviance Table

Binomial model

Response: Kyphosis

Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev
  NULL                    80   83.23447
   Age  1  1.30198        79   81.93249
Number  1 10.30593        78   71.62656
 Start  1 10.24663        77   61.37993
```

Summary for kyphosis model with age^2 added

```
Call: glm(formula = Kyphosis ~ poly(Age, 2) + Number
  + Start, family = binomial, data = kyphosis)

Deviance Residuals:
       Min         1Q    Median          3Q      Max
 -2.235654 -0.5124374 -0.245114 -0.06111367 2.354818

Coefficients:
                    Value Std. Error   t value
  (Intercept)  -1.6502939 1.40171048 -1.177343
poly(Age, 2)1   7.3182325 4.66933068  1.567298
poly(Age, 2)2 -10.6509151 5.05858692 -2.105512
       Number   0.4268172 0.23531689  1.813798
        Start  -0.2038329 0.07047967 -2.892080
```

Summary of fit with age^2 added

```
 Null Deviance: 83.23447 on 80 degrees of freedom

Residual Deviance: 54.42776 on 76 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:
              (Intercept) poly(Age, 2)1 poly(Age,
  2)2     Number
poly(Age, 2)1 -0.2107783
poly(Age, 2)2  0.2497127  -0.0924834
       Number -0.8403856   0.3070957    -0.0988896
        Start -0.4918747  -0.2208804     0.0911896
  0.0721616
```

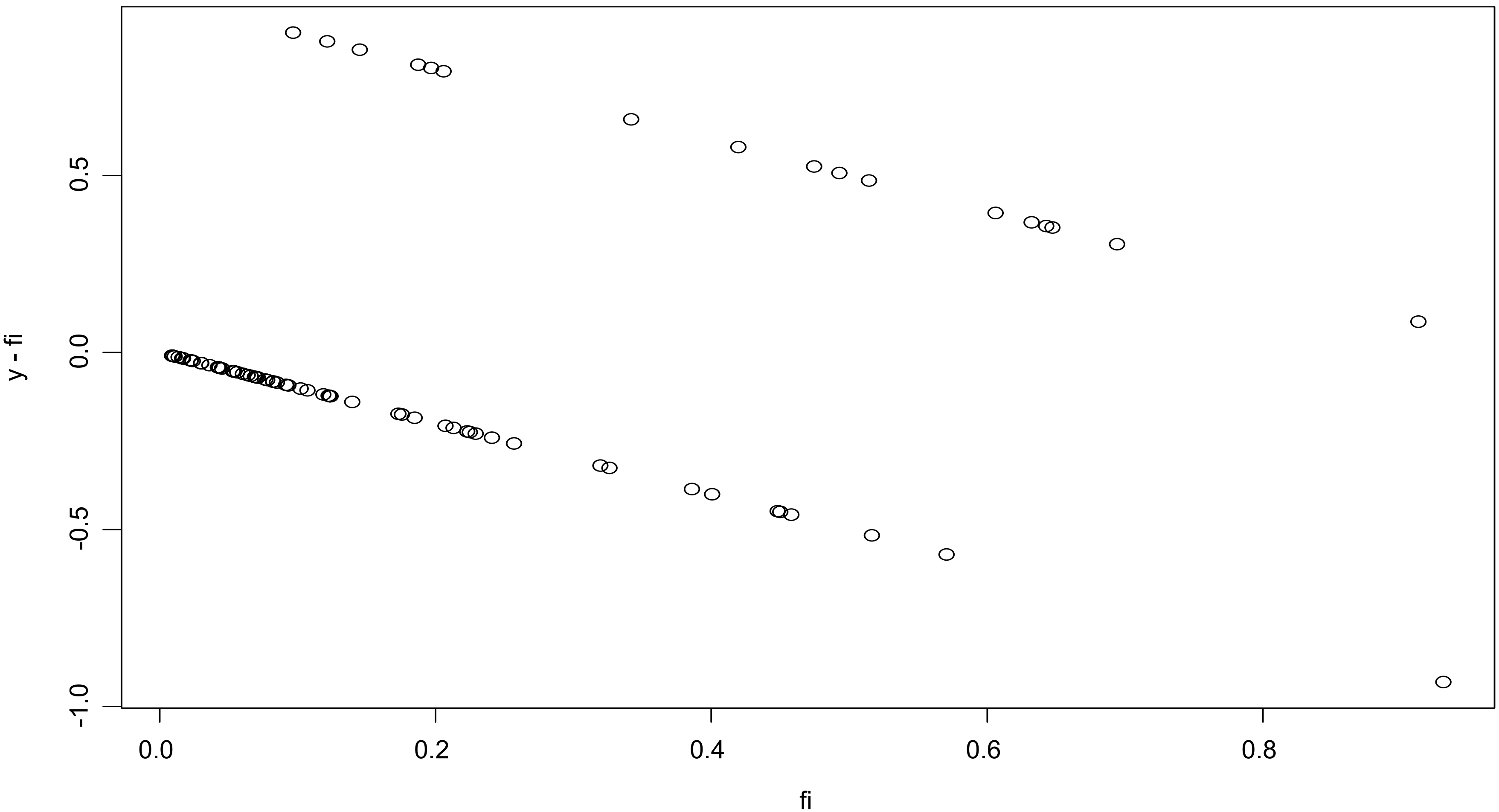
Analysis of Deviance

```
> anova(kyph.glm2)
Analysis of Deviance Table

Binomial model

Response: Kyphosis

Terms added sequentially (first to last)
             Df Deviance Resid. Df Resid. Dev
        NULL                    80   83.23447
poly(Age, 2)  2 10.49589        78   72.73858
      Number  1  8.87597        77   63.86261
       Start  1  9.43485        76   54.42776
```

Kyphosis data, 16 obs, with fit and residuals

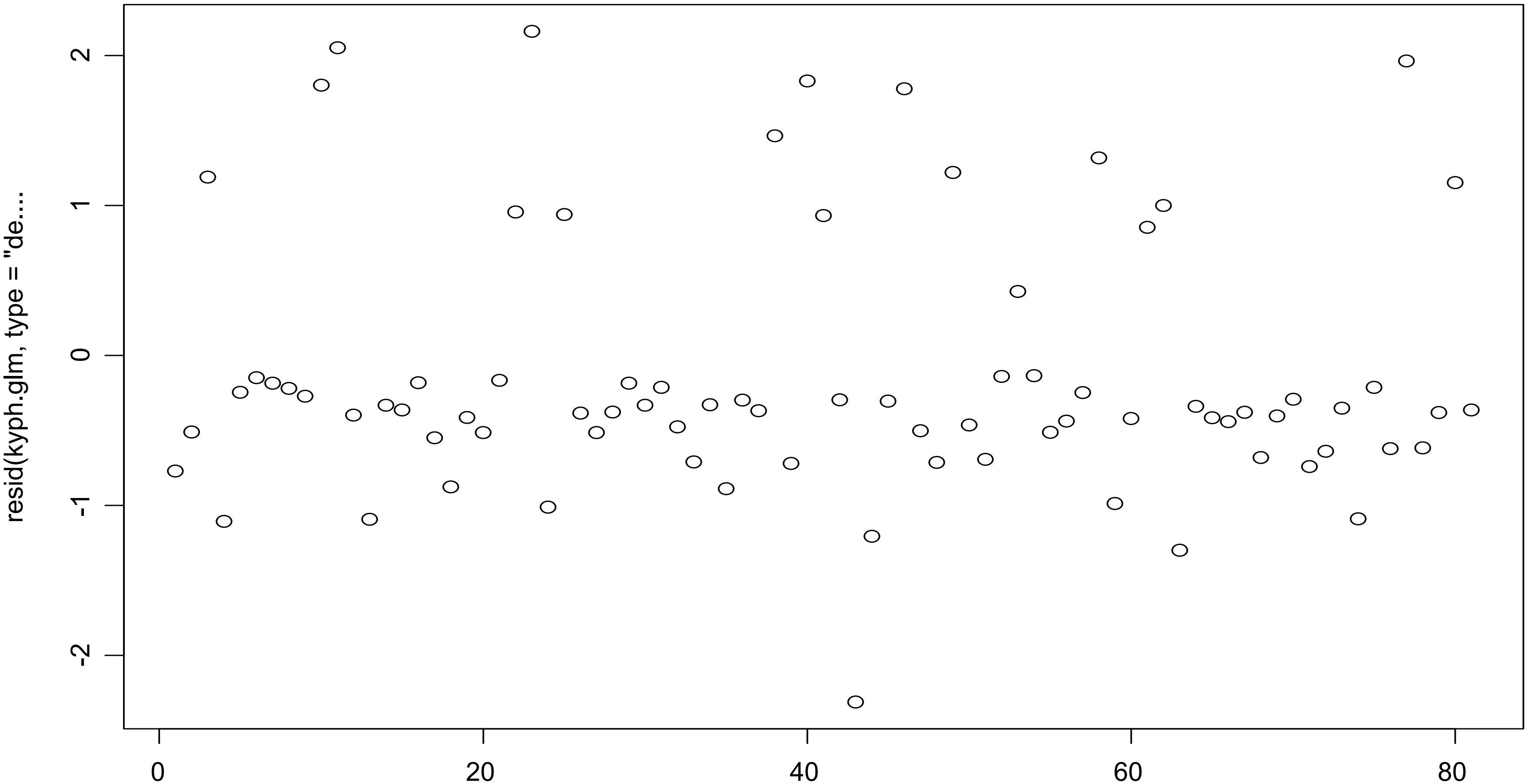
```
cbind(kyphosis,round(p,3),round(rr,3),round(rp,3),round(rd,3))[1:16,]
   Kyphosis Age Number Start   fit     rr     rp     rd
 1   absent  71      3     5 0.257 -0.257 -0.588 -0.771
 2   absent 158      3    14 0.122 -0.122 -0.374 -0.511
 3  present 128      4     5 0.493  0.507  1.014  1.189
 4   absent   2      5     1 0.458 -0.458 -0.919 -1.107
 5   absent   1      4    15 0.030 -0.030 -0.175 -0.246
 6   absent   1      2    16 0.011 -0.011 -0.105 -0.148
 7   absent  61      2    17 0.017 -0.017 -0.131 -0.185
 8   absent  37      3    16 0.024 -0.024 -0.157 -0.220
 9   absent 113      2    16 0.036 -0.036 -0.193 -0.271
10  present  59      6    12 0.197  0.803  2.020  1.803
11  present  82      5    14 0.121  0.879  2.689  2.053
12   absent 148      3    16 0.076 -0.076 -0.288 -0.399
13   absent  18      5     2 0.450 -0.450 -0.905 -1.094
14   absent   1      4    12 0.054 -0.054 -0.239 -0.333
16   absent 168      3    18 0.064 -0.064 -0.261 -0.363
17   absent   1      3    16 0.016 -0.016 -0.129 -0.181
```

Plot of response residual vs. fit

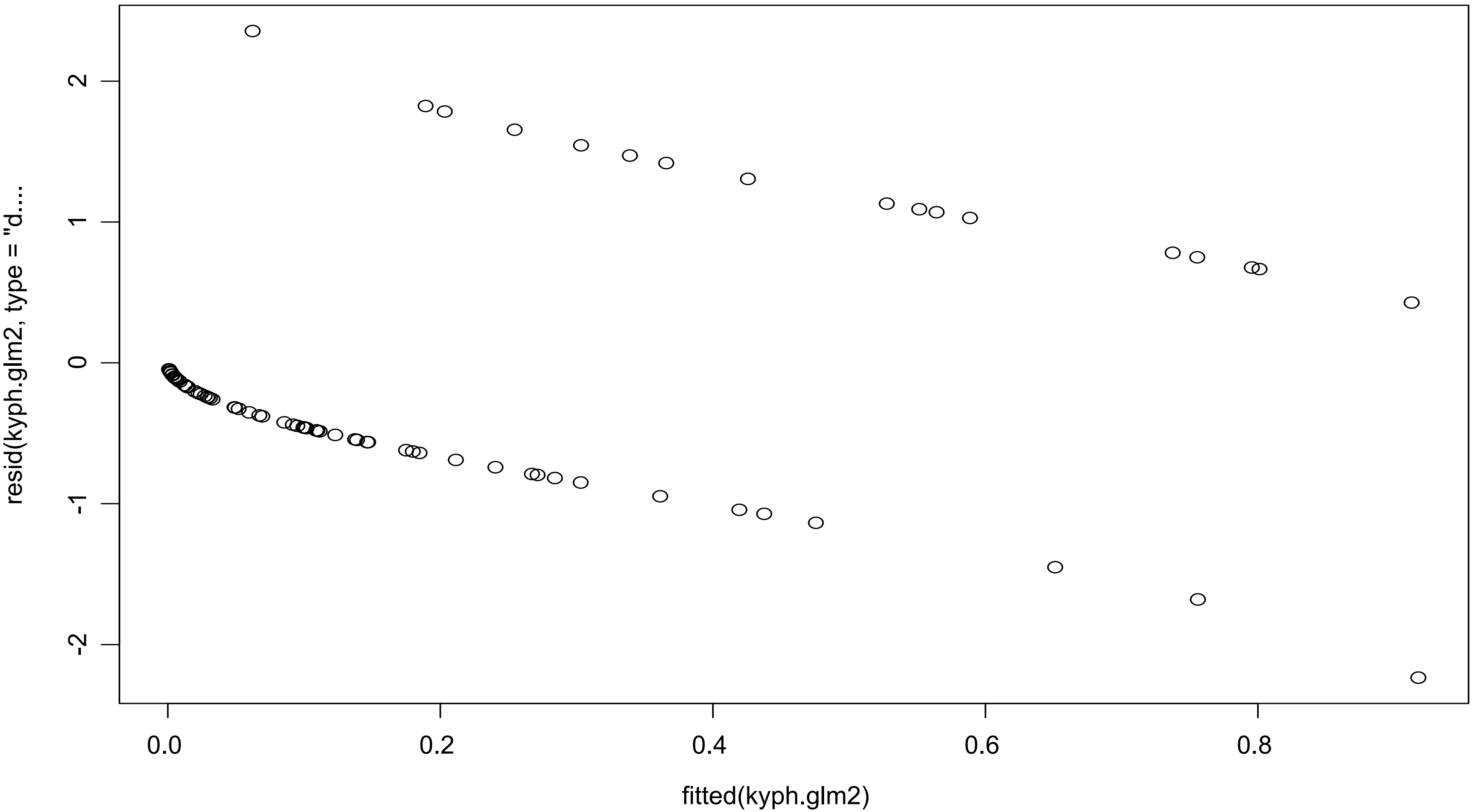


This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Plot of deviance residual vs. index



Plot of deviance residuals vs. fitted value



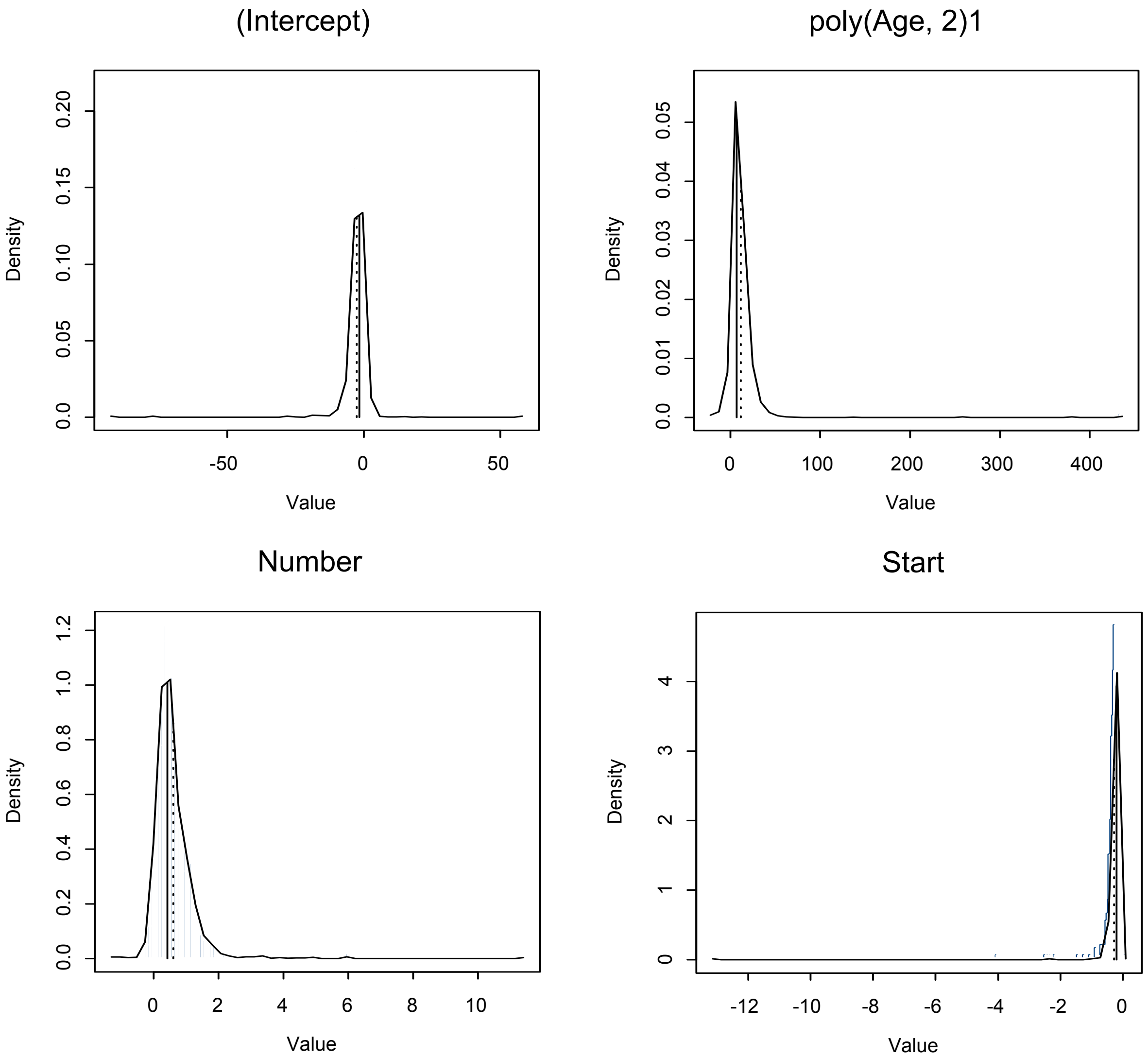
Summary of bootstrap for kyphosis model

```
Call:
bootstrap(data = kyphosis, statistic = coef(glm(Kyphosis ~
   poly(Age, 2) + Number + Start, family = binomial,
   data = kyphosis)), trace = F)

Number of Replications: 1000

Summary Statistics:
              Observed     Bias     Mean      SE
  (Intercept)  -1.6503 -0.85600  -2.5063  5.1675
poly(Age, 2)1   7.3182  4.33814  11.6564 22.0166
poly(Age, 2)2 -10.6509 -7.48557 -18.1365 37.6780
       Number   0.4268  0.17785   0.6047  0.6823
        Start  -0.2038 -0.07825  -0.2821  0.4593

Empirical Percentiles:
                   2.5%          5%       95%      97.5%
  (Intercept)  -8.52922  -7.247145   1.1760   2.27636
poly(Age, 2)1  -6.13910  -1.352143  27.1515  34.64701
poly(Age, 2)2 -48.86864 -38.993192  -4.9585  -4.13232
       Number  -0.07539  -0.003433   1.4756   1.82754
        Start  -0.58795  -0.470139  -0.1159  -0.08919
```

Summary of bootstrap (continued)

```
BCa Confidence Limits:
                    2.5%       5%        95%      97.5%
  (Intercept)  -6.4394  -5.3043   2.39707   3.56856
poly(Age, 2)1 -18.2205 -10.1003  18.34192  21.56654
poly(Age, 2)2 -24.2382 -20.3911  -1.75701  -0.19269
       Number  -0.7653  -0.1694   1.14036   1.27858
        Start  -0.3521  -0.3167  -0.03478   0.01461

Correlation of Replicates:
              (Intercept) poly(Age, 2)1 poly(Age, 2)2  Number   Start
  (Intercept)      1.0000       -0.4204        0.5082 -0.5676 -0.1839
poly(Age, 2)1     -0.4204        1.0000       -0.8475  0.4368 -0.6478
poly(Age, 2)2      0.5082       -0.8475        1.0000 -0.3739  0.5983
       Number     -0.5676        0.4368       -0.3739  1.0000 -0.4174
        Start     -0.1839       -0.6478        0.5983 -0.4174  1.0000
```

Histograms of coefficient estimates



E Newton

QQ Plots of coefficient estimates