Two Examples of Logistic Regression

1 Budworms

This data set is analyzed by Venables and Ripley also, pp. 230 ff. *Check page ref's in current edition.*

The data is the result of an experiment on the toxicity of the tobacco budworm to doses of particular drug, that the moths are beginning to show resistance to. Batches of 20 moths of each sex were exposed for 3 days to the chemical and the number of dead (or disabled) moths in each group were recorded. The data is,

```
Dose Response Gender
1       1       M
2       4       M
4       9       M
8       13      M
16      18      M
32      20      M
1       0       F
2       2       F
4       6       F
8       10      F
16      12      F
32      16      F
```

The doses were in micro-grams, and the dose level when up by powers of two. It is a judgement call as to whether you will try to model with dose as the response or $\log_2(\text{dose})$. We will try both.

```
> bud <- read.table("budworm.dat", head=T)
> bud
   Dose Response Gender
 1    1        1      M
 2    2        4      M
 3    4        9      M
 4    8       13      M
 5   16       18      M
 6   32       20      M
 7    1        0      F
 8    2        2      F
 9    4        6      F
10    8       10      F
11   16       12      F
12   32       16      F
> attach(bud,pos=1)
> N <- 20
```

First, lets plot the data, and see if anything obvious shows up.

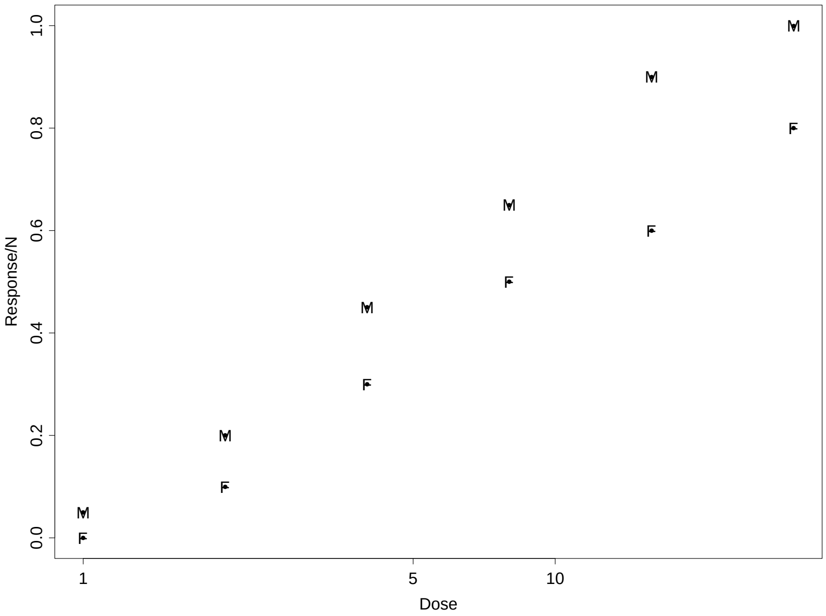

Figure 1: Frequency of budworm death (mortality) vs. log-dose of pesticide.

```
> plot( Dose, Response/N)
> points( Dose[Gender=="F"], (Response/N)[Gender=="F"], pch="F")
> points( Dose[Gender=="M"], (Response/N)[Gender=="M"], pch="M")
```

The same plot on the log(Dose) scale, is shown in Figure 1.

```
> plot( Dose, Response/N, log="x")
> points( Dose[Gender=="F"], (Response/N)[Gender=="F"], pch="F")
> points( Dose[Gender=="M"], (Response/N)[Gender=="M"], pch="M")
```

Given that we have more than 0-1 data, it makes a lot of sense to at least try to plot the (observed) logits; Figure 2 does this for both dose and log(dose) of the pesticide.

```
> par(mfrow=c(2,1))
> plot( Dose, logit(Response/N), log="x")
> title("Logit vs Log(Dose)")
> plot( Dose, logit(Response/N))
> title("Logit vs Dose")
```

This plot makes it quite clear, that the log scale has a much better chance of working.

So, proceed to fit the model. We surely want a model with different slopes for the two sexes

```
> ldose <- log(Dose)
> SF <- cbind(Response, N - Response)
```

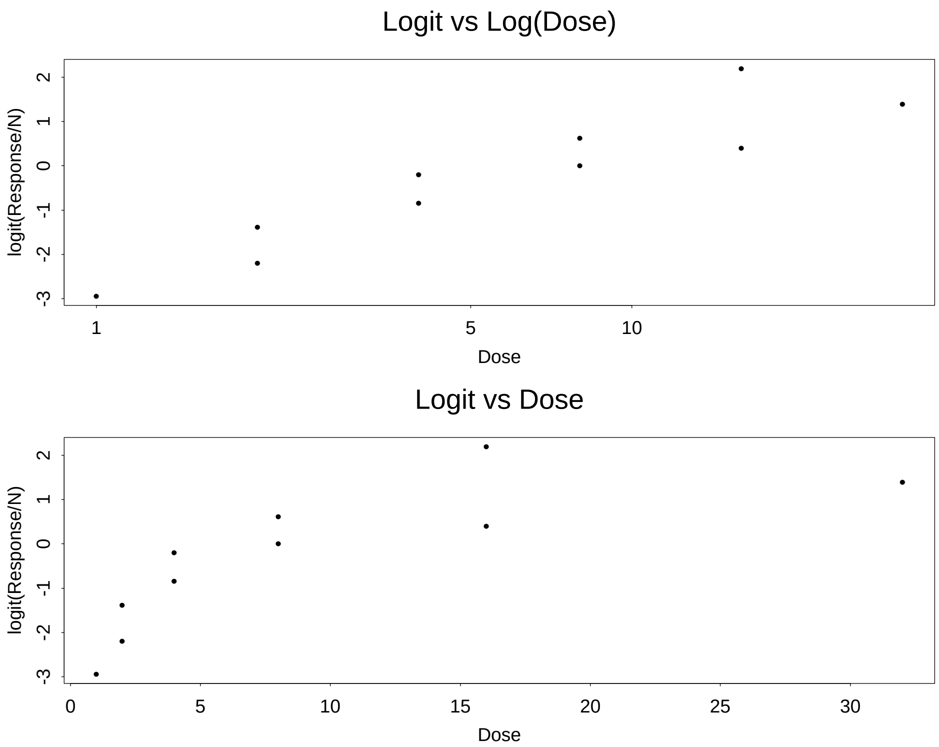

Figure 2: Logit-mortality vs Dose and Log-dose of pesticide.

```
> mymod <- glm(SF ~ Gender * ldose, family=binomial)
> summary(mymod)

 ..........
Coefficients:
                  Value Std. Error    t value
 (Intercept) -2.9935414  0.5525295 -5.4178852
      Gender  0.1749865  0.7781556  0.2248733
       ldose  1.3071341  0.2410133  5.4234939
Gender:ldose  0.5091459  0.3894474  1.3073547


    Null Deviance: 124.8756 on 11 degrees of freedom

Residual Deviance: 4.993727 on 8 degrees of freedom

> anova(mymod,test="Chisq")
Analysis of Deviance Table

Binomial model

Response: SF
```

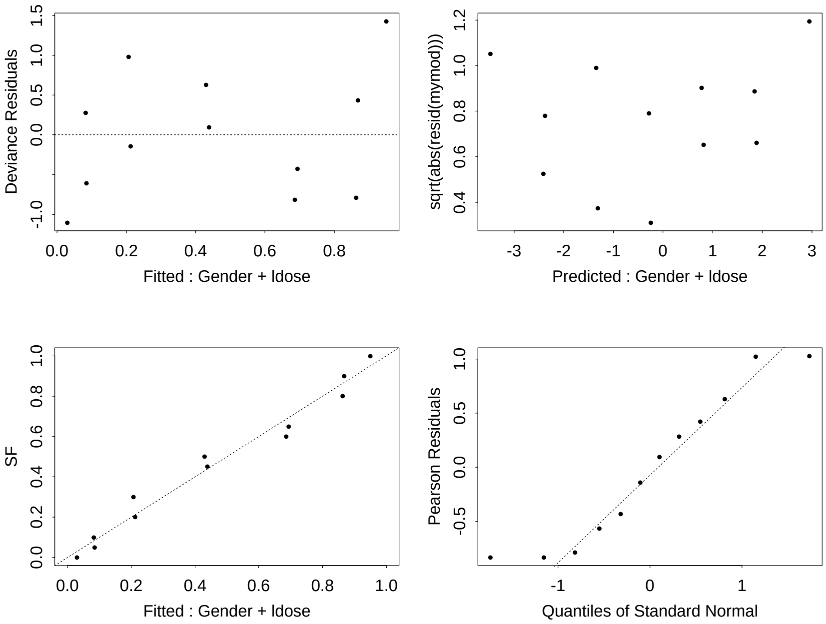

Figure 3: Standard Splus residual diagnostic plots for `glm(SF ~ Gender + ldose, family=binomial)`.

```
Terms added sequentially (first to last)
             Df Deviance Resid. Df Resid. Dev   Pr(Chi)
        NULL                    11   124.8756
      Gender  1   6.0770        10   118.7986 0.0136955
       ldose  1 112.0415         9     6.7571 0.0000000
Gender:ldose  1   1.7633         8     4.9937 0.1842088
>
```

Now, what is going on here? `Gender` seems to have no effect (based on the t-value), but the deviance is quite large? Even more, our plots seem to imply what looks like a gender effect. The best way to find out is to distrust the parameter stuff (it can get easily confused), and try fitting a model without the interaction. This makes sense anyway, because there is little/no evidence of an interaction.

```
> anova(mymod <- glm(SF ~ Gender + ldose, family=binomial))

Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev
  NULL                    11   124.8756
Gender  1   6.0770        10   118.7986
 ldose  1 112.0415         9     6.7571
> summary(mymod)
Coefficients:
                Value Std. Error   t value
(Intercept) -3.473154  0.4682939 -7.416612
```

```
      Gender  1.100743  0.3557226  3.094385
       ldose  1.535336  0.1890119  8.122959

    Null Deviance: 124.8756 on 11 degrees of freedom

Residual Deviance: 6.757064 on 9 degrees of freedom
```

That is a little more like it, and seems to be a VERY good fitting model. We should at least look at some diagnostics, and some other plots (see Figure 3), but unless they show something very odd, we are done.

```
> plot(mymod)
```

Figure 3 looks very good. We should finish by plotting the data and fitted curves on the same graph, and that is left as an exercise for you (or look it up in the Venables and Ripley). Venables and Ripley also go on to talk about LD50, and how to estimate it.

2 An Extended Example: Stroke

2.1 Background

The data are from a prospective study whose purpose was to characterize the spontaneous recovery of language functioning in patients during the first 3 to 4 months following a stroke. Details of the design and results of this study may be found in Holland et al. (1985) and Holland, Greenhouse, Fromm, and Swindell (1989). Briefly, patients admitted to a hospital with the diagnosis of stroke and who met the study inclusion criteria were seen daily (6 days/week) for 15 minutes beginning 48 to 72 hours poststroke and continuing until hospital discharge. The daily visits consisted of semistructured conversational interactions and systemic scoring of the patient's communication behavior. At hospital discharge, patients were formally tested using the Western Aphasia Battery (WAB). If the WAB indicated aphasic dysfunction (AQ $<$ 93.8), the patient was seen and assessed at 1 month postdischarge. If this second WAB demonstrated continued difficulty the patient was assessed again at 2 months postdischarge. The data set to be discussed here consists of 61 patients who had unilateral (right- or left-hemisphere) stroke, ischemic or hemorrhagic etiology, and impaired language functioning at time of hospital discharge. We will use these data to investigate the correlates of language recovery during the two month period post-hospital discharge. Definitions and the codes for the variables of interest are presented in Table 2.

The multiple logistic regression model will allow us to investigate simultaneously the effect of a number of different variables on the likelihood of spontaneous language recovery. A feature of the logistic regression model is that it allows us to use explanatory variables that are either quantitative or discrete. Using the variables defined and coded in Table 2 for the stroke recovery study, the specification of the deterministic component of the multiple linear logistic regression model on the logit scale is

$$ln(\tfrac{\pi}{1-\pi}) = \beta_0 + \beta_1 \cdot (SIDE) + \beta_2 \cdot (GENDER) + \beta_3 \cdot (AGE) + \beta_4 \cdot (LOSD) + \beta_5 \cdot (WAB1). \qquad (1)$$

Note that if we were to dichotomize length of stay, age and WAB1 then a multiple cross-classification table of these data would consist of $2^6 = 64$ cells or subgroups and that with 61 patients such a table would have numerous sparse or empty cells which would be a problem for a chi-square analysis. An advantage of a statistical model such as the one given in equation (1), however, is that the model provides a structure that smooths the fit of the probability of a response to all the cells in such a table.

```
Side   Gender   Age   LOSD   WAB1   Status

  0      1       74    17    92.5     1
  1      1       75    30    87.8     0
  0      1       72    17    65.9     0
  0      1       60    21    86.4     1
  0      0       69    11    90.0     1
  0      0       62    13    6.2      0
```

Table 1: Example of what the data file 'stroke.dat' looks like

2.2 Interpretation of Regression Coefficients

The output below presents part of the S-Plus output listing the estimates of the regressions coefficients and their standard errors. We see that the coefficients for gender, length-of-stay, and the WAB score at discharge are all significantly related to the probability of spontaneous language recovery at the 0.05 level. The coefficient for length-of-stay is negative indicating that the longer patients are in the hospital the smaller the chance of spontaneous language recovery relative to patients who have shorter lengths-of-stay even after controlling for the effect of the WAB1 score. The coefficient for WAB1 is positive as we would expect. Males, after controlling for the other variables in the model, are $exp(2.88) = 17.8$ times as likely to recover spontaneous language function as females. A 95% confidence interval for β_2, the adjusted log odds ratio of recovery for males relative to females, however, is large and ranges from 0.47 to 5.29. We also note that side of stroke and subject's age do not have a statistically significant association with spontaneous language recovery after controlling for the effects of the other variables. Details about these tests can be found in Hosmer and Lemeshow (1989, Chapter 5).

```
> stroke <- read.table("stroke.dat", head=T)
> stroke[1:10,]
   Side Gender Age LOSD WAB1 Status
 1    0      0  79   27 18.2      0
 2    0      0  82   17 86.0      0
 3    1      0  72   42 46.3      0
 4    0      0  59   23 35.6      0
 5    0      0  60   28  0.0      0
 6    1      1  75   12 81.9      1
 7    0      0  75   53 56.0      0
 8    0      0  85   46 17.3      0
 9    0      0  79   58 74.9      0
10    0      1  82   34 18.2      0

> mymod <- glm(Status ~ Side + Gender + Age + LOSD +WAB1, family=binomial,
+ data=stroke)

> summary(mymod)

Call: glm(formula = Status ~ Side + Gender + Age + LOSD + WAB1,
```

Variable	*Description*	*Values/Codes*
STATUS	Status at last observation	0 = Not Recovered 1 = Recovered
SIDE	Side of lesion	0 = Left 1 = Right
GENDER		0 = Female 1 = Male
AGE	Age at stroke onset	In years
LOSD	Length of hospital stay	In days
WAB1	WAB at hospital discharge	Range: 0 - 100

Table 2: Variables and Codes from the Stroke Recovery Study

```
        family = binomial, data = stroke)
Deviance Residuals:
       Min         1Q       Median            3Q      Max
 -1.317524 -0.2999725 -0.003778562 -0.0002212266 2.862154

Coefficients:
                    Value Std. Error    t value
(Intercept) -11.072759221 6.28633143 -1.7614024
       Side  -1.369026127 1.10318267 -1.2409786
     Gender   2.876569185 1.25325078  2.2952862
        Age  -0.006938071 0.03062231 -0.2265691
       LOSD  -0.196137069 0.09595028 -2.0441532
       WAB1   0.164467235 0.07172681  2.2929674

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 65.71913 on 60 degrees of freedom

Residual Deviance: 28.9878 on 55 degrees of freedom

Number of Fisher Scoring Iterations: 7
```

```
Correlation of Coefficients:
       (Intercept)       Side     Gender         Age       LOSD
  Side  0.2553975
Gender -0.5007016  -0.2793548
   Age -0.3373404  -0.1265137  0.2006563
  LOSD  0.1199385   0.2046374 -0.5309336 -0.1894954
  WAB1 -0.9292116  -0.3055970  0.5278826  0.0616042 -0.2969773
> anova(mymod)
Analysis of Deviance Table

Binomial model

Response: Status

Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev
  NULL                    60   65.71913
  Side  1  0.10011        59   65.61902
Gender  1  3.36010        58   62.25892
   Age  1  1.45449        57   60.80444
  LOSD  1 12.25425        56   48.55018
  WAB1  1 19.56238        55   28.98780
>
```

2.3 Analysis of the Fit of the Model

Figure 4 presents a residual analysis for this logistic regression.

One relatively easy method for the analysis of the fit of the logistic regression model is based on calculating the predicted probability of spontaneous language recovery based on the model in (1) compared to the observed probability of language recovery Specifically, substitute the estimated regression coefficients given in above into the multiple logistic regression equation in (1). For each patient in the study calculate his or her individual predicted probability of language recovery by calculating the linear combination of each patient's individual set of explanatory variables with the estimated regression coefficients and use the inverse of the logit transformation to solve for the probability of language recovery, $\hat{\pi}$. For example, consider a patient whose stroke was on the left side (SIDE=0), was a 61 year old male (GENDER=1, AGE=61), stayed in the hospital 19 days (LOSD=19), and whose WAB at discharge was 81.6 (WAB1=81.6). The estimated probability of spontaneous language recovery based on the model for such a patient is

$$\hat{\pi} = \frac{1}{1 + e^{-(-11.1-1.37\cdot(0)+2.88\cdot(1)-0.0069\cdot(61)-0.196\cdot(19)+0.165\cdot(81.6))}} = 0.77.$$

This value can be interpreted to mean that 77 patients out of 100 with the above characteristics can be expected to recover normal language functioning spontaneously within the first 2 months post-hospital discharge. To study the goodness of fit of the model, we obtain the frequency distribution of the values of $\hat{\pi}$, and determine, say, the quintiles of the distribution. The observed number of patients who had normal language functioning are tallied within each quintile, and the expected number of patients who had normal language functioning (predicted by the model) is determined by summing the $\hat{\pi}$'s for all the patients in the quintile.

Table 3 presents the expected and observed number of patients who had normal language functioning for each quintile of risk of language recovery. Comparisons of the expected number predicted by the model to the observed number suggests that this model fits the data reasonably well.

A test for whether there is a significant difference between the observed and expected numbers in Table 3 can be constructed based on a Pearson chi-square test. This chi-square test would have degrees-of-freedom equal to the number of groups (e.g., 5 for the number of quintiles) minus 2 (Hosmer and Lemeshow, 1989, chapter 5). In our example, this chi-square statistics is equal to 3.60 with 3 degrees-of-freedom and has a non-significant p-value=0.69, indicating that the model adequately fits the data. A small p-value would indicate a significant lack of fit of the model.

	Quintile of risk	Expected Number	Observed Number
Highest	5	10.7	11
	4	3.1	2
	3	0.2	1
	2	0.0	0
Lowest	1	0.0	0
Total		14.0	14

Table 3: Analysis of the fit of the logit model in Equation (1)

2.4 How to plot these data

Look at the stroke data. This is explicitly 0–1 data, and it is very hard to think about plotting it. A simple plot of, say, `Status` against `WAB1`, is shown in Figure 5, and it not all that informative. Even if I jitter the points a little, it is still not very easy to see any trends or patterns.

```
> plot(WAB1, Status)
> plot(WAB1, Status+0.1*runif(length(Status)))
```

A somewhat more useful approach is to plot the data and add some sort of smooth curve that will allow you to at least look at some of the trends in the data. The particular approach I took with these data was to superimpose a (black-box) smooth curve generated by loess. The various pictures in Figure 6 result from different smoothing parameters.

```
> scatter.smooth(WAB1, Status,span=0.2)
> title("Span = 0.2")
> scatter.smooth(WAB1, Status,span=0.35)
> title("Span = 0.35")
> scatter.smooth(WAB1, Status,span=0.5)
> title("Span = 0.5")
> scatter.smooth(WAB1, Status,span=0.65)
> title("Span = 0.65")
```

This is still not perfect, but you certainly get the sense that the probability goes up, probably more or less monotonely with WAB1.

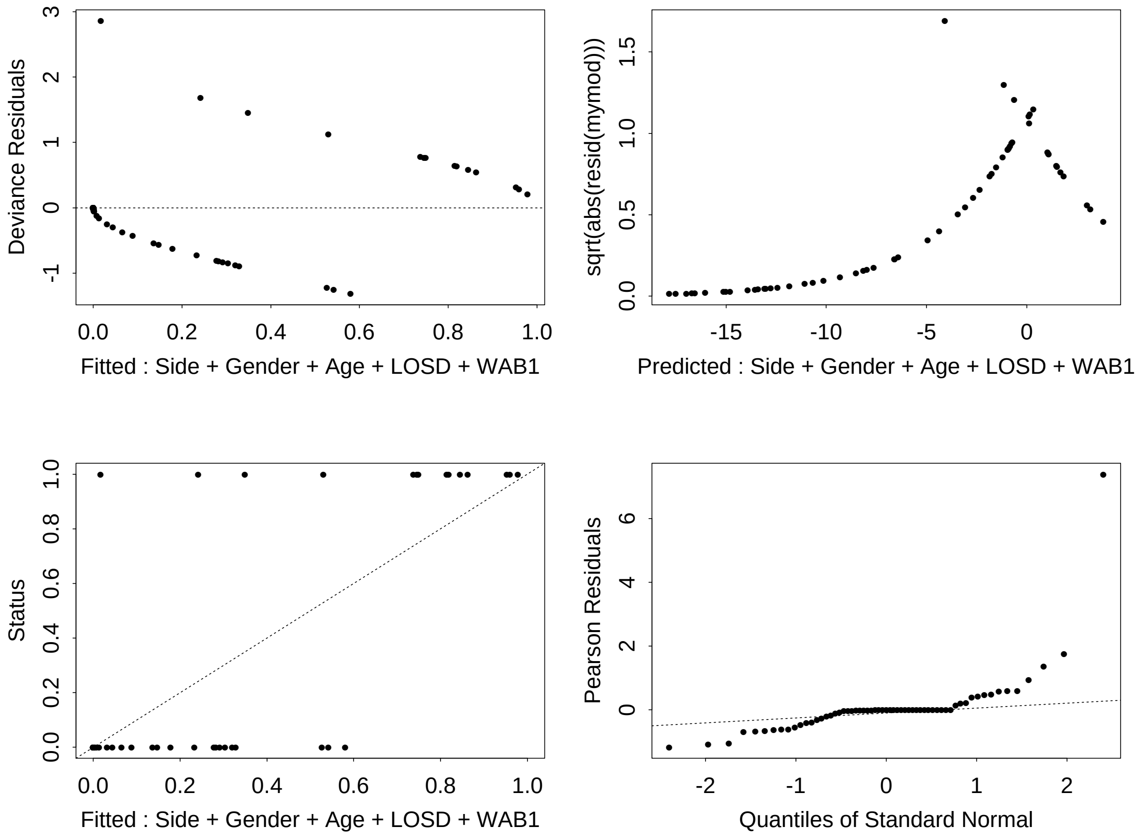

Figure 4: Splus's standard residual plots for `glm(Status ~ Side + Gender + Age + LOSD +WAB1, family=binomial, + data=stroke)`.

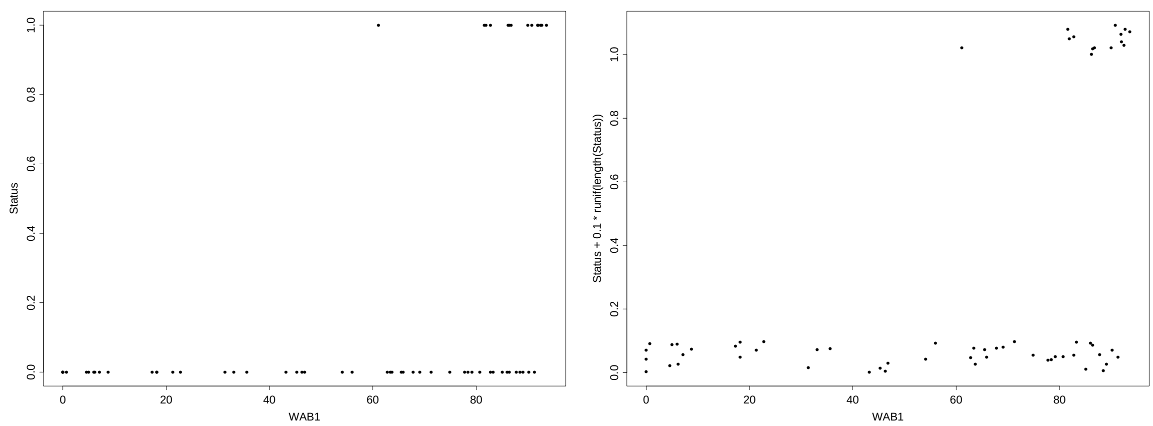

Figure 5: Stroke status (0/1) vs WAB at intake (WAB1) (unjittered on the left; jittered on the right).

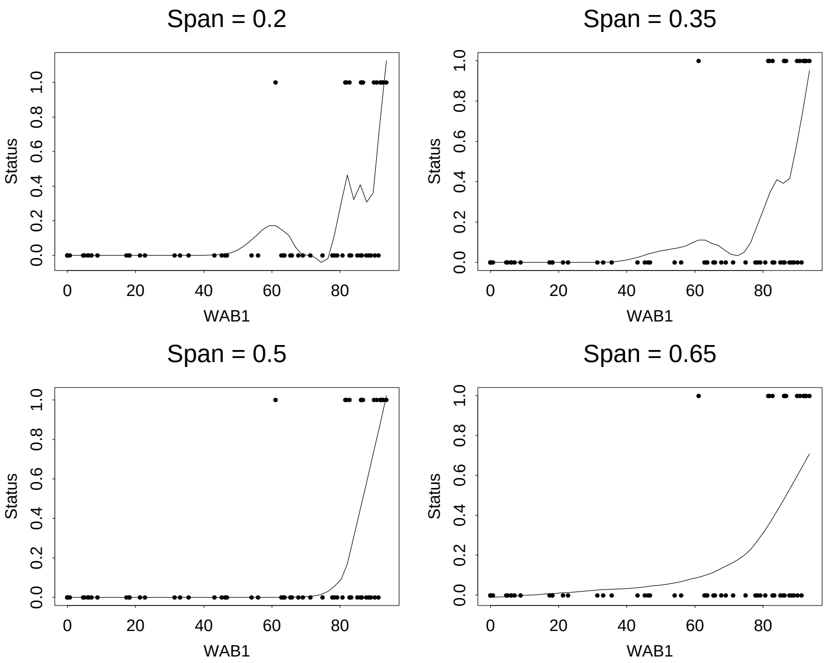

Figure 6: Various loess smooths of stroke status vs. WAB1.

Compare this to the similar picture(s) for Age (Figure 7).

```
> for (i in c(0.2, 0.35, .5, 0.65)){
+ scatter.smooth(Age, Status, span=i)
+ title(paste("Span = ", i))
+ }
```

If anything, there is a bit of evidence for some type of quadratic effect of Age, but the probable story is that Age is not that important, **on its own**. We should always remember that variables that do not seem to be important (from these marginal plots), may be quite important when paired with other variables (that is viewed conditionally).

With that hint about the Age variable, I re-ran the logistic regression model for these data and obtained,

```
> mymod <- glm(Status ~ Side + Gender + LOSD +WAB1 + Age + Age^2, family=bin>
> summary(mymod)

Call: glm(formula = Status ~ Side + Gender + LOSD + WAB1 + Age + Age^2, family =
        binomial)
Deviance Residuals:
       Min          1Q         Median             3Q       Max
 -1.317304 -0.2637303 -0.0002506098 -6.014239e-07 2.770847

Coefficients:
                   Value   Std. Error   t value
(Intercept) -42.13652444 17.240544854 -2.444037
```

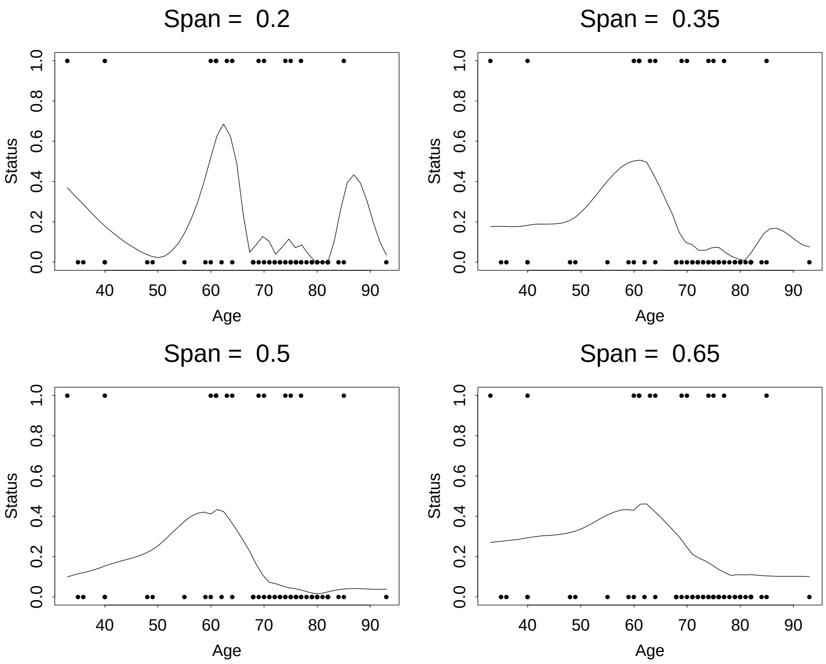

Figure 7: Various loess smooths of stroke status vs. age.

```
       Side  -1.70612965  1.320176507 -1.292350
     Gender   4.03384776  1.847166535  2.183803
       LOSD  -0.28388626  0.130491084 -2.175522
       WAB1   0.24891576  0.104667163  2.378165
        Age   0.92453559  0.421765743  2.192059
   I(Age^2)  -0.00791642  0.003578441 -2.212254

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 65.71913 on 60 degrees of freedom

Residual Deviance: 21.7621 on 54 degrees of freedom

> anova(mymod)
Analysis of Deviance Table

Binomial model

Response: Status

Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev
    NULL                    60   65.71913
    Side  1  0.10011        59   65.61902
  Gender  1  3.36010        58   62.25892
```

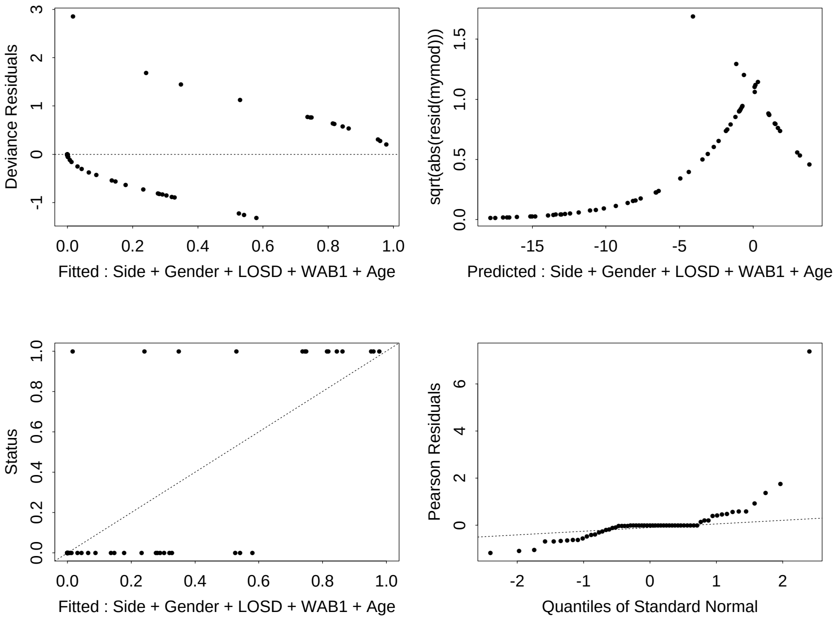

Figure 8: Residual diagnostic plots for `glm(Status ~ Side + Gender + Age + LOSD +WAB1, family=binomial, + data=stroke)`.

```
    LOSD  1 13.46451        57   48.79441
    WAB1  1 19.75552        56   29.03889
     Age  1  0.05110        55   28.98780
I(Age^2)  1  7.22570        54   21.76210
```

This is somewhat surprising (because I didn't think there was an Age effect). Could we have seen the same trend from a residual analysis of these data?

Let us go back to the original model and look at residuals (Figure 8). There is at least one large (Pearson) residual, and it is the same point as the large (deviance) residual. What about a plot of (deviance) residual against `Age` (when `Age` is in the model).

```
> mymod <- glm(Status ~ Side + Gender + LOSD +WAB1 + Age, family=binomial)
> scatter.smooth(Age, (residuals(mymod)))
> identify(Age, (residuals(mymod)))
[1] 59
> scatter.smooth(Age[-59], (residuals(mymod))[-59])
```

We see that observation `59` is the one with the large residual. If I plot residuals of my model against age (the lower plot in Figure 9), the quadratic trend seems to go away, at least a bit.

Note that this shows that observation #59 is highly influentialon the coefficient of `Age^2` in the model. This should show up in Cook's D or DFBETA plots as well. Observation `59` is now suspect, so let us remove that observation, and see what the resulting model looks like.

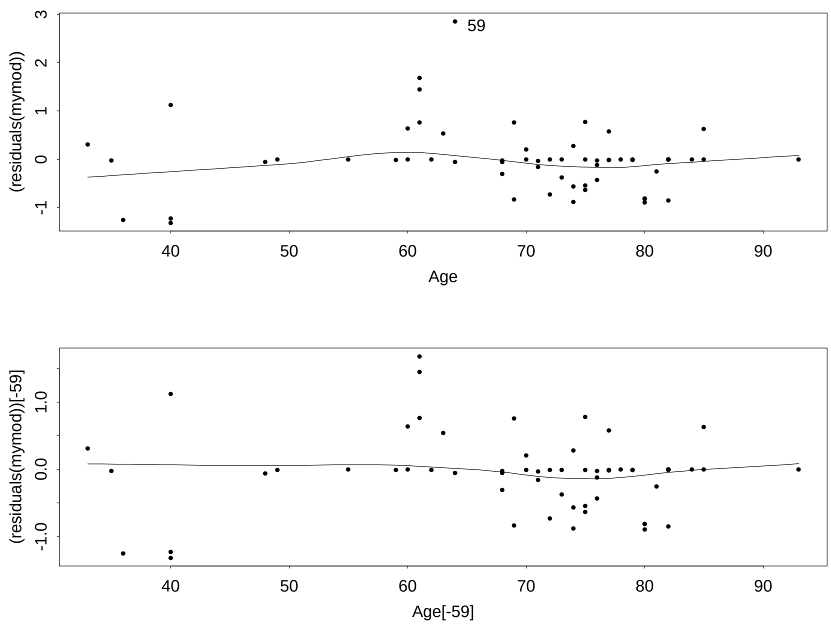

Figure 9: Loess smooths of residuals of `glm(Status ~ Side + Gender + LOSD +WAB1 + Age, family=binomial)`, with and without observation 59.

```
> WW <- rep(1, length(Side))
> WW[59] <- 0
> mymod <- glm(Status ~ Side + Gender + LOSD +WAB1 + Age,
+ weight=WW, family=binomial)
Warning messages:
  linear convergence not obtained in  10  iterations. in: glm.fitter(x = X, y =
        Y, w = w, start = start, offset = offset, family = family, ....
```

Oh-oh, what does that mean? (Basically, the optimization algorithm is converging slowly, which is an indication that there are some fitted probabilities that are very close to 0 or 1). We can try a few more iterations, via.

```
> mymod <- glm(Status ~ Side + Gender + LOSD +WAB1 + Age,
+ weight=WW, maxit=30,family=binomial)
```

Now, what does the summary and anova look like?

```
> summary(mymod)

Call: glm(formula = Status ~ Side + Gender + LOSD + WAB1 + Age, family = binomial,
        weights = WW, maxit = 30)
Deviance Residuals:
      Min           1Q        Median            3Q      Max
 -1.21685 -0.006138687 -2.107342e-08 -2.107342e-08 2.120409
```

```
Coefficients:
                    Value  Std. Error     t value
(Intercept) -46.034665461 23.60416082 -1.95027757
       Side  -7.479606389  4.79366143 -1.56031178
     Gender   7.889726328  4.28891089  1.83956406
       LOSD  -0.677382165  0.43418335 -1.56012928
       WAB1   0.662724777  0.34770876  1.90597667
        Age  -0.001289562  0.04096666 -0.03147832

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 62.71879 on 60 degrees of freedom

Residual Deviance: 13.32727 on 54 degrees of freedom

Number of Fisher Scoring Iterations: 10

Correlation of Coefficients:
       (Intercept)       Side     Gender       LOSD       WAB1
  Side  0.8469550
Gender -0.9187997  -0.8860396
  LOSD  0.7944690   0.9174434 -0.8555061
  WAB1 -0.9831416  -0.9061715  0.9317554 -0.8758313
   Age -0.0387868   0.0615410  0.0243350 -0.0022418 -0.0596445
Warning messages:
  1 rows with zero weights not counted in: summary(mymod)
```

If I now start adding the quadratic term in `Age`, I get myself into all sorts of trouble.

```
> formula(mymod)
Status ~ Side + Gender + LOSD + WAB1 + Age + Age^2
> anova(mymod)
Analysis of Deviance Table

Binomial model

Response: Status

Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev
    NULL                    60   62.71879
    Side  2  0.62858        58   62.09021
  Gender  1  2.34693        57   59.74328
    LOSD  1 13.67051        56   46.07277
    WAB1  1 32.74451        55   13.32826
     Age  1  0.00099        54   13.32727
```

```
I(Age^2)  1 13.32727        53    0.00000
Warning messages:
1: linear convergence not obtained in  10  iterations. in: glm.fit(x, y, w,
        offset = offset, family = family, maxit = control$maxit,  ....
2: linear convergence not obtained in  10  iterations. in: glm.fit(x, y, w,
        offset = offset, family = family, maxit = control$maxit,  ....
```

Whoa. What is going on here? We suddenly have a model that fits perfectly?

The summary looks even weirder?

```
> summary(mymod)

Call: glm(formula = Status ~ Side + Gender + LOSD + WAB1 + Age + Age^2, family =
        binomial, weights = WW, maxit = 50)
Deviance Residuals:
         Min            1Q        Median            3Q           Max
 -3.9065e-05 -2.107342e-08 -2.107342e-08 -2.107342e-08 3.281089e-05

Coefficients:
                   Value  Std. Error       t value
(Intercept) -2858.6085408 2227320.001 -0.0012834297
       Side  -164.0618144  165491.082 -0.0009913635
     Gender   273.5752941  237195.052  0.0011533769
       LOSD   -16.2464405   13633.473 -0.0011916582
       WAB1    24.1122280   17153.825  0.0014056473
        Age    39.8782980   35880.527  0.0011114190
   I(Age^2)    -0.3549147     302.566 -0.0011730157

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 62.71879 on 60 degrees of freedom

Residual Deviance: 0 on 53 degrees of freedom

Number of Fisher Scoring Iterations: 29

Correlation of Coefficients:
         (Intercept)       Side     Gender       LOSD       WAB1         Age
     Side  0.3661039
   Gender -0.8397332  -0.5132934
     LOSD  0.7676113   0.5867586 -0.7405370
     WAB1 -0.9678428  -0.5023038  0.8384746 -0.7822973
      Age -0.9537356  -0.2802197  0.7986261 -0.8085521  0.8643352
I(Age^2)  0.9585630   0.3372172 -0.8129406  0.8428222 -0.8863173 -0.9963403
Warning messages:
1: fitted values close to 0 or 1 in: family$deviance(mu, y, w, residuals = T)
2: 1 rows with zero weights not counted in: summary(mymod)
```

```
>
```

On reflection, it is clear what is going on. Look at the first plot of Status against WAB1 (Figure 5). Here we see, that absent the one observation in the middle (which is 59), we have an almost perfect discrimination between the two levels of Status. That is the clue to what is "broken". With the 7 parameters (anything like WAB1 squared would have done), we can find a function that perfectly discriminates between the two levels of status, and then maximum likelihood modelling breaks down.

2.5 Influence of a point on the logistic regression coefficients: DFBETA

Pregibon's (1981) approximation to DFBETA is not pre-programmed into Splus, and for a small data set like this it is just easy to compute DFBETA's exactly. The basic idea of the following code is to fit the `glm` model with each data point in turn removed. The summary is then a plot of each of the coefficients, with as many points on the plot as there are data points. As in OLS regression, the plot is surprisingly effective and quickly shows us where the pathology in the data lies.

```
# Heavily annotated, for class

# Generate a vector of all ones, of length the number of observations.
W <-  rep(1,length(Side))
# Generate current weights.
WW <- W
# Fit the base model
mymod <- glm(Status ~ Gender + LOSD + Side + Age + WAB1, weight=WW,
   family=binomial)

# Save the coefficients, just in case
base.coef <- coefficients(mymod)

# Set up a matrix to hold the coefficients for each model with one
# point deleted.
infs <- matrix(0, nrow=length(Side), ncol=length(coefficients(mymod)))

# Loop through the data, setting one weight equal to 0.
# Save the coefficients in the infs matrix
for (i in 1:length(Side)){
        WW<-W; WW[i] <- 0;
        infs[i,] <- coefficients(update(mymod, . ~ .))
}
## Can't think of an easy, smart summary, so just plot the
## coefficients.

#Set up a 2 x 3 plotting frame
par(mfrow=c(2,3))

#Do the 6 plots.  After each plot, allow the user to
# identify the points which look interesting.
```

```
for(i in 1:6){
        plot(1:length(Side), infs[,i])
        title(names(coefficients(mymod))[i])
        identify(1:length(Side), infs[,i])
}
```

After running that code, and interacting with the various plots, you end up with the plot in Figure 10. There are other things we could have plotted (like following the deviance as the models changed), but this plot alone tells us much of the story.

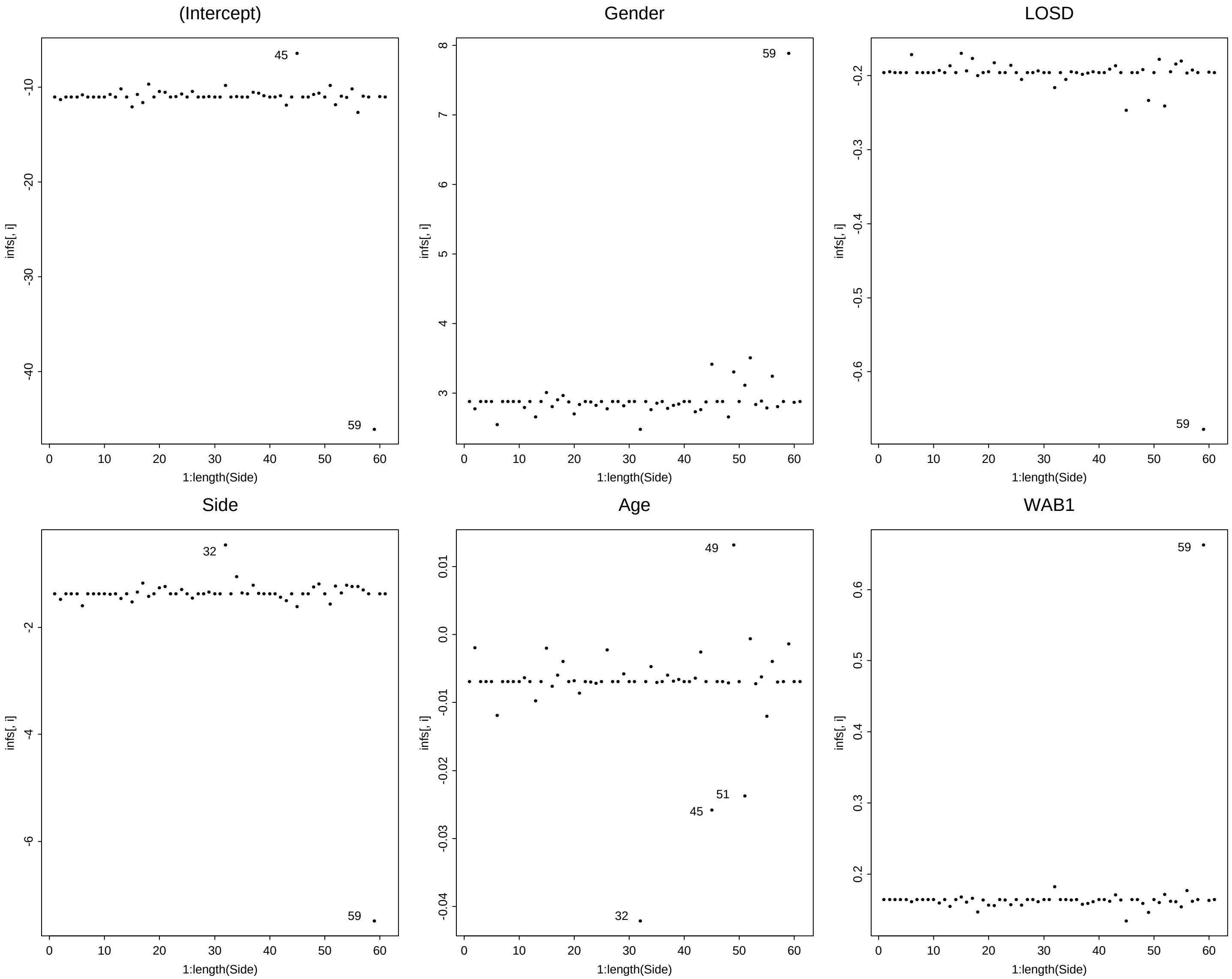

Figure 10: Exact DFBETA plots, obtained by refitting the model with each observation deleted in turn.