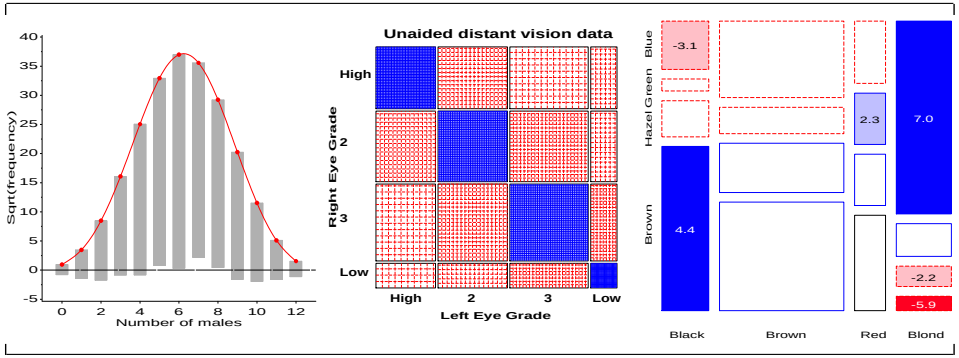
Visualizing Categorical Data with SAS and R

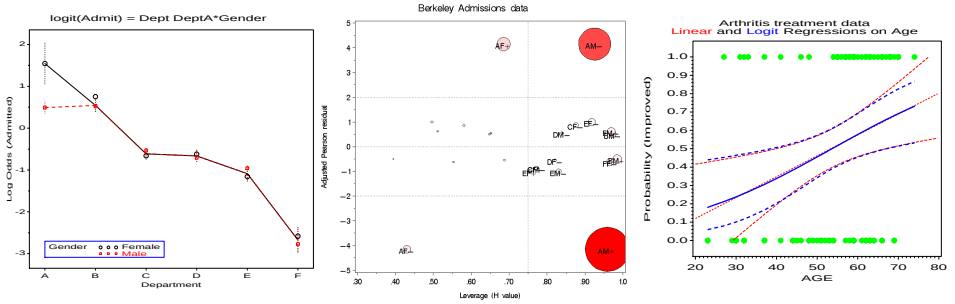
Michael Friendly

York University

Short Course, 2012
Web notes: `datavis.ca/courses/VCD/`



Part 4: Model-based methods for categorical data



Topics:

- Logit models
  - Plots for logit models
  - Diagnostic plots for generalized linear models
- Logistic regression models
  - Logistic regression: Binary response
  - Model plots
  - Effect plots for generalized linear models
  - Influence measures and diagnostic plots

Modeling approaches: Overview

Association models

- **Loglinear models**

(Contingency table form)

[Admit][GenderDept]
[AdmitDept][GenderDept]
[AdmitDept][AdmitGender][GenderDept]

- **Poisson GLMs**

(Frequency data frame)

Freq ~ Admit + Gender*Dept
Freq ~ Admit*Dept + Gender*Dept
Freq ~ Admit*Dept + Admit*Gender + Gender*Dept

- **Ordered variables**

Freq ~ right+left+Diag(right,left)
Freq ~ right+left+Symm(right,left)

Response models

- **Binary response**
- Categorical predictors: Logit models

logit(Admit) ~ 1
logit(Admit) ~ Dept
logit(Admit) ~ Dept + Gender

- Continuous/mixed predictors: Logistic regression models

Pr(Admit) ~ Dept + Age + GRE

- **Polytomous response**
- Ordinal: proportional odds model

Improve ~ Age + Sex + Treatment

- General: multinomial model

WomenWork ~ Kids + HusbandInc

Logit models

For a *binary* response, each loglinear model is equivalent to a logit model (logistic regression, with categorical predictors)

- e.g., Admit $\perp$ Gender $|$ Dept (conditional independence $\equiv$ [AD][DG])

$$\log\, m_{ijk} = \mu + \lambda_i^A + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{AD} + \lambda_{jk}^{DG}$$

So, for admitted ($i = 1$) and rejected ($i = 2$), we have:

$$\log m_{1jk} = \mu + \lambda_1^A + \lambda_j^D + \lambda_k^G + \lambda_{1j}^{AD} + \lambda_{jk}^{DG} \qquad (7)$$

$$\log m_{2jk} = \mu + \lambda_2^A + \lambda_j^D + \lambda_k^G + \lambda_{2j}^{AD} + \lambda_{jk}^{DG} \qquad (8)$$

Thus, subtracting (7)-(8), terms not involving Admit will cancel:

$$\begin{aligned} L_{jk} &= \log m_{1jk} - \log m_{2jk} = \log(m_{1jk}/m_{2jk}) = \text{log odds of admission} \\ &= (\lambda_1^A - \lambda_2^A) + (\lambda_{1j}^{AD} - \lambda_{2j}^{AD}) \\ &= \alpha + \beta_j^{\text{Dept}} \qquad \text{(renaming terms)} \end{aligned}$$

where,

- α: overall log odds of admission
- β_j^{Dept}: effect on admissions of department,
- associations among predictors are assumed, but don't appear in the logit model.

Logit models

Other loglinear models have similar, simpler forms as logit models, where only the relations of the response to the predictors appear in the equivalent logit model.

- Admit ⊥ Gender ⊥ Dept (mutual independence ≡ [A][D][G])

$$\begin{aligned} \log m_{ijk} &= \mu + \lambda_i^A + \lambda_j^D + \lambda_k^G \\ \equiv L_{jk} &= (\lambda_1^A - \lambda_2^A) = \alpha \qquad \text{(constant log odds)} \end{aligned}$$

- Admit ⊥ Gender | Dept, except for Dept. A

$$\begin{aligned} \log m_{ijk} &= \mu + \lambda_i^A + \lambda_j^D + \lambda_k^G + \lambda_{ij}^{AD} + \lambda_{jk}^{DG} + \delta_{(j=1)}\lambda_{ik}^{AG} \\ \equiv L_{jk} &= \log(m_{1jk}/m_{2jk}) = \alpha + \beta_j^{\text{Dept}} + \delta_{(j=1)}\beta^{\text{Gender}} \end{aligned}$$

where,

- β_j^{Dept}: effect on admissions for department j,
- $\delta_{(j=1)}\beta^{\text{Gender}}$: 1 df term for effect of gender in Dept. A.

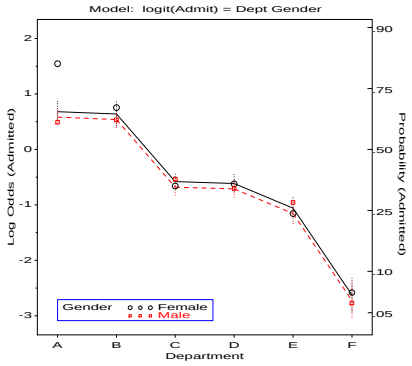
Logit models: Overview

- **Fitting procedures**
  - `PROC CATMOD`, `PROC LOGISTIC`
  - `PROC GENMOD / dist=poisson`
  - SPSS: Logistic regression, Loglinear → Logit, Generalized Linear Models
  - R: glm(), gnm()
- **Visualization procedures**
  - `CATPLOT` macro - plot predicted, observed log odds from `CATMOD`
  - `INFLGLIM` macro - influence plots for generalized linear models
  - `HALFNORM` macro - half-normal plot of residuals for generalized linear models
- **SAS craft**
  - All SAS procedures → output dataset with obs., fitted values, residuals, diagnostics, etc.
  - New model → new output dataset
  - Plotting steps remain the same
  - Similar ideas for SPSS, R

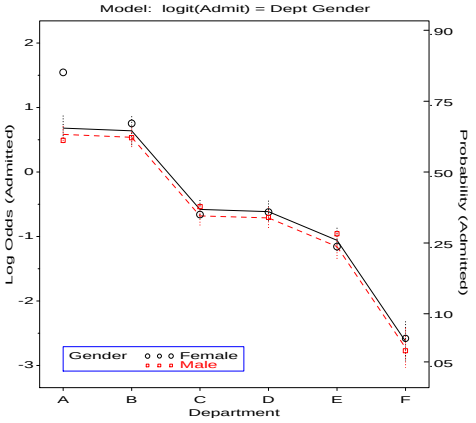
Plots for logit models

- Fit: `PROC CATMOD`; plot: `CATPLOT` macro
- Model: Admit ∼ Gender + Dept ↔ loglinear [AD] [AG] [DG]

```
proc catmod order=data data=berkeley;
   weight freq;
   response / out=predict;
   model admit = dept gender / ml;
%catplot(data=predict, xc=dept, class=gender,
    type=FUNCTION, z=1.96, legend=legend1);
```



Plots for logit models



- Plots observed and predicted on the logit scale (type=FUNCTION)

⇒ Main effects model— parallel profiles

- Probabilities on a separate scale (added below)

Logit models: details

- **Model:** Admit ∼ Gender + Dept ↔ [AD] [AG] [DG]

catberk2.sas ···

```
%include catdata(berkeley);
proc catmod order=data
            data=berkeley;
   weight freq;
   response / out=predict;
   model admit = dept gender / ml;
  run;
```

PROC CATMOD output: Overall tests and goodness of fit

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	1	262.49	<.0001
dept	5	534.78	<.0001
gender	1	1.53	0.2167
Likelihood Ratio	5	20.20	0.0011

- No effect of Gender; big effect of Dept
- LR test (vs. saturated model): Model doesn't fit well— Why? How to modify?

Plots for logit models: Output data set

PROC CATMOD output data set: observed & predicted, probabilities & logits

dept	gender	admit	_TYPE_	_OBS_	_PRED_	_SEPRED_
A	Male		FUNCTION	0.492	0.582	0.069
A	Male	Admit	PROB	0.621	0.642	0.016
A	Male	Reject	PROB	0.379	0.358	0.016
A	Female		FUNCTION	1.544	0.682	0.099
A	Female	Admit	PROB	0.824	0.664	0.022
A	Female	Reject	PROB	0.176	0.336	0.022
B	Male		FUNCTION	0.534	0.539	0.086
B	Male	Admit	PROB	0.630	0.631	0.020
B	Male	Reject	PROB	0.370	0.369	0.020
B	Female		FUNCTION	0.754	0.639	0.116
B	Female	Admit	PROB	0.680	0.654	0.026
B	Female	Reject	PROB	0.320	0.346	0.026
...						
F	Male		FUNCTION	-2.770	-2.724	0.158
F	Male	Admit	PROB	0.059	0.062	0.009
F	Male	Reject	PROB	0.941	0.938	0.009
F	Female		FUNCTION	-2.581	-2.625	0.158
F	Female	Admit	PROB	0.070	0.068	0.010
F	Female	Reject	PROB	0.930	0.932	0.010

This contains both the observed and fitted logit values (_TYPE_='FUNCTION') and probabilities (_TYPE_='PROB')

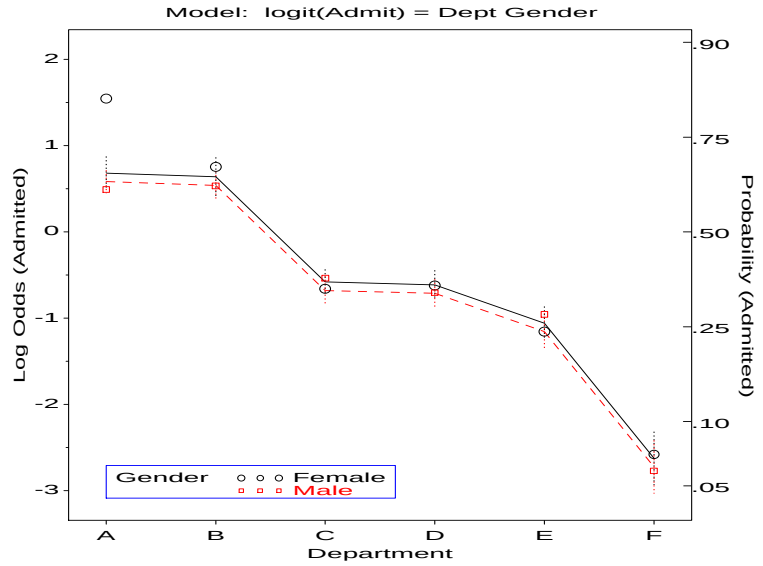
CATPLOT macro

- Plot logit values (_TYPE_='FUNCTION') or probabilities (_TYPE_='PROB')
- With PSCALE macro, can plot on logit scale, with probability scale on right.

··· catberk2.sas

```
%pscale(lo=-4, hi=3, anno=pscale);

title 'Model:  logit(Admit) = Dept Gender'
   a=-90 'Probability (Admitted)';
axis1 order=(-3 to 2) offset=(4)
      label=(a=90 'Log Odds (Admitted)');
axis2 label=('Department') offset=(4);
%catplot(data=predict, class=gender, xc=dept,
    type=FUNCTION,     /* plot logit values          */
    z=1.96,            /* show 1.96 x SE -> 95% CI */
    anno=pscale);      /* add probability scale     */
```


CATPLOT macro

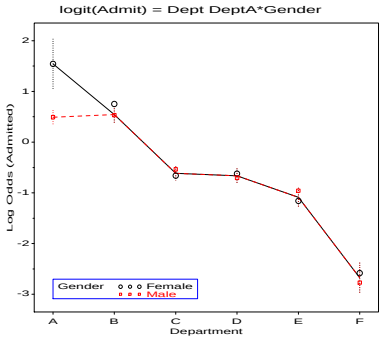


→ no effect of Gender, except in Dept A (Females more likely admitted!)

Fitting and graphing other models

- Change MODEL statement → new fitted values
- Plotting step remains the same
- Admit ⊥ Gender | Dept, except for Dept. A ↔ Admit ∼ Dept + $\delta_{j=1}$ Gender

```
proc catmod order=data data=berkeley;
   response / out=predict;
   model admit = dept dept1AG / ml;
%catplot(data=predict, xc=dept, class=gender,
    type=FUNCTION, z=1.96, legend=legend1);
```



Fitting and graphing other models: details

- Model: Admit ⊥ Gender | Dept, except for Dept. A
- Need to define a dummy variable for effect of Gender in Dept. A

catberk6.sas ···

```
1  %include catdata(berkeley);
2  data berkeley;
3     set berkeley;
4     *-- Dummy variable for Gender in Dept A;
5     dept1AG = (gender='F') *  (dept=1);
6     format dept dept.;
7
8  proc catmod order=data
9               data=berkeley;
10    weight freq;
11    population dept gender;
12    direct dept1AG;
13    response / out=predict;
14    model admit = dept dept1AG / ml;
15   run;
16    ...
```


Fitting and graphing other models:details

PROC CATMOD output:

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	1	291.22	<.0001
dept	5	571.45	<.0001
dept1AG	1	16.04	<.0001
Likelihood Ratio	5	2.68	0.7489

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		-0.6685	0.0392	291.22	<.0001
dept	A	1.1606	0.0705	271.21	<.0001
	B	1.2113	0.0802	227.95	<.0001
	C	0.0528	0.0687	0.59	0.4426
	D	0.00358	0.0727	0.00	0.9607
	E	-0.4210	0.0871	23.34	<.0001
dept1AG		1.0521	0.2627	16.04	<.0001

Fits well! How to interpret?

Fitting and graphing other models: details

PROC CATMOD: observed and predicted logits:

··· catberk6.sas ···

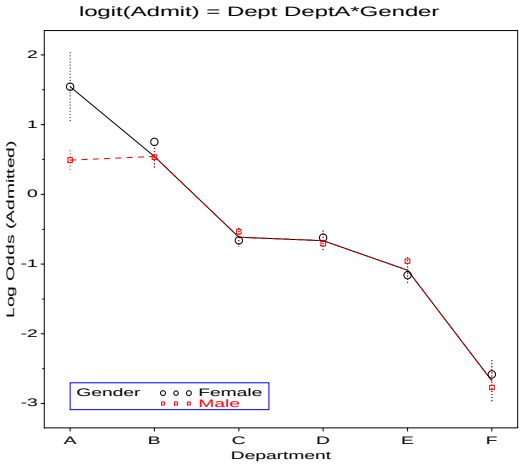
```
17 proc print data=predict;
18   id dept gender;
19   var _obs_ _pred_ _sepred_;
20   format _numeric_  6.3 dept dept.;
21   where(_type_='FUNCTION');
```

dept	gender	_OBS_	_PRED_	_SEPRED_
A	M	0.492	0.492	0.072
A	F	1.544	1.544	0.253
B	M	0.534	0.543	0.086
B	F	0.754	0.543	0.086
C	M	-0.536	-0.616	0.069
C	F	-0.660	-0.616	0.069
D	M	-0.704	-0.665	0.075
D	F	-0.622	-0.665	0.075
E	M	-0.957	-1.090	0.095
E	F	-1.157	-1.090	0.095
F	M	-2.770	-2.676	0.152
F	F	-2.581	-2.676	0.152

Fitting and graphing other models: details

catberk6.sas

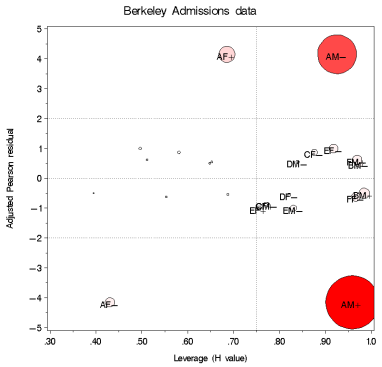
```
22 title 'logit(Admit) = Dept DeptA*Gender';
23 %catplot(data=predict, x=dept, class=gender,
24    type=FUNCTION,      /* plot the log odds */
25    z=1.96);            /* 95% error bars    */
```



Diagnostic plots for Generalized Linear Models

INFLGLIM macro: Influence plots for generalized linear models (Williams, 1987)

- Fit: PROC GENMOD; calculates additional diagnostic measures (Hat value, Cook's D, etc.)
- Plot: measures of residual (GY=$\Delta\chi^2$, χ^2 residual) vs. leverage (GX=hat value), bubble size (area, radius) $\sim$ Cook's D.
- $\rightarrow$ which cells have undue impact on fitted model?



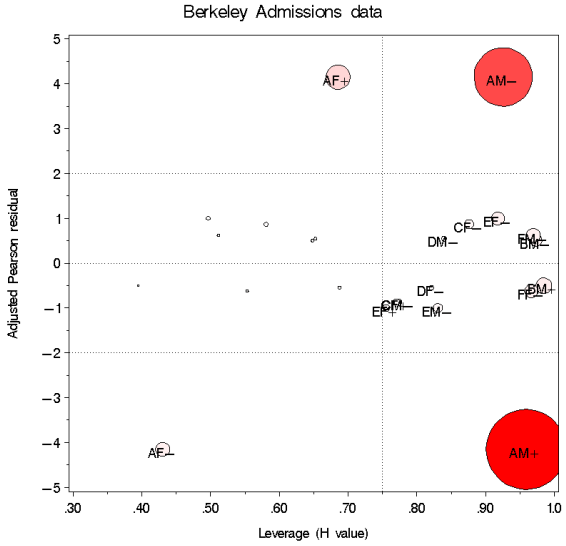
INFLGLIM macro: Example

- Berkeley data, model $[AD][GD] \leftrightarrow L_{ij} = \alpha + \beta_j^{\text{Dept}}$

genberk1.sas

```
1  %include catdata(berkeley);
2  *-- make a cell ID variable, joining factors;
3  data berkeley;
4     set berkeley;
5     cell = trim(put(dept,dept.)) ||
6            gender ||
7            trim(put(admit,yn.));
8
9  %inflglim(data=berkeley,
10    class=dept gender admit,
11    resp=freq,
12    model=admit|dept gender|dept,
13    dist=poisson,
14    id=cell,
15    gx=hat, gy=streschi);
```


INFLGLIM macro: Example



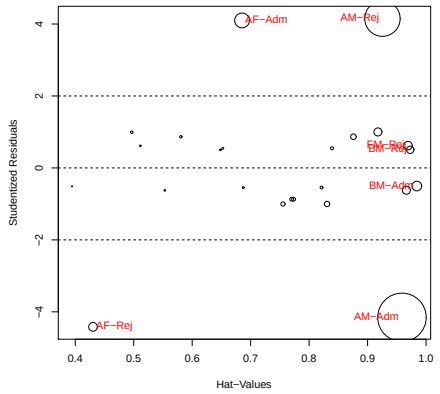
- All cells which do not fit ($|r_i| > 2$) are for department A.
- Males applying to dept A have large leverage $\Rightarrow$ large influence (Cook's D)

Influence plots in R

The `influencePlot()` function in the `car` package gives similar plots:

berkeley-diag.R

```
berkeley <- as.data.frame(UCBAdmissions)
 ...
berk.mod <- glm(Freq ~ Dept * (Gender+Admit), data=berkeley,
                family="poisson")
influencePlot(berk.mod, id.n=3, id.col="red")
```



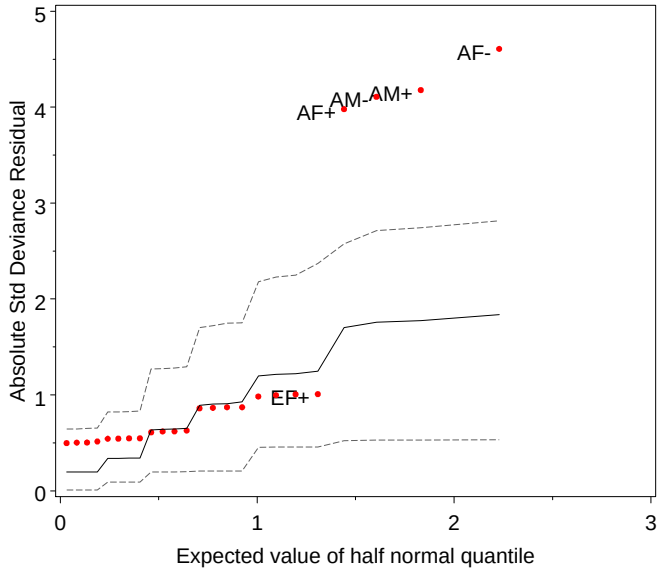
Diagnostic plots for Generalized Linear Models

`HALFNORM` macro: Half-normal plot of residuals (Atkinson, 1981)

- Plot ordered *absolute* residuals, $|r|_{(i)}$ vs. expected normal values, $|z|_{(i)}$
- Standard normal confidence envelope not suitable for GLMs
- Simulate reference 'line' and envelope with simulated confidence intervals

··· genberk1.sas

```
%halfnorm(data=berkeley,
   class=dept gender admit,
   resp=freq,
   model=dept|gender dept|admit,
   dist=poisson, id=cell);
```

- Points with largest |residual| labeled
- The model fits well, except in department A.

Logistic regression models

Response variable

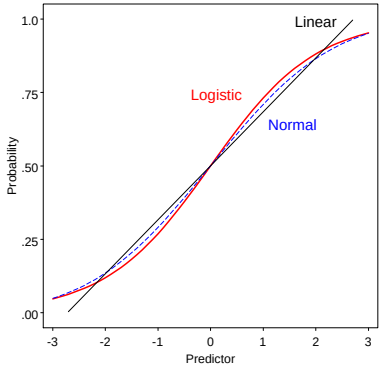
- Binary response: success/failure, vote: yes/no
- Binomial data: x successes in n trials (grouped data)
- Ordinal response: none $<$ some $<$ severe depression
- Polytomous response: vote Liberal, Tory, NDP, Green

Explanatory variables

- Quantitative regressors: age, dose
- Transformed regressors: $\sqrt{\text{age}}$, log(dose)
- Polynomial regressors: age^2, age^3, $\cdots$
- Categorical predictors: treatment, sex
- Interaction regessors: treatment $\times$ age, sex $\times$ age

Logistic regression models: Binary response

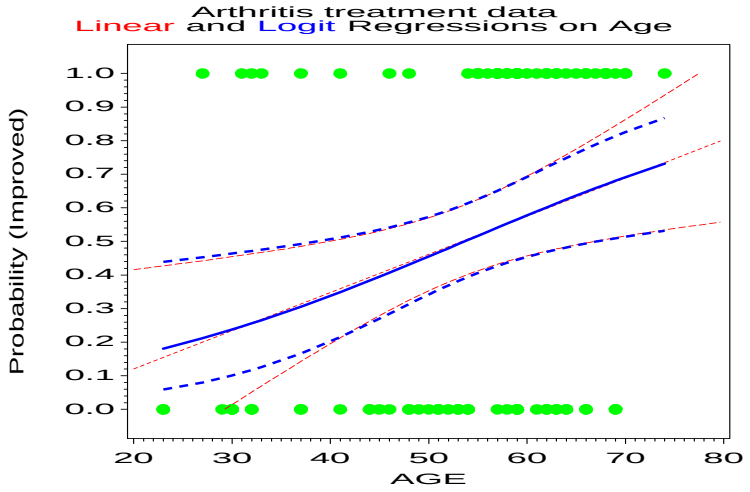
- For a binary response, $Y \in (0,1)$, want to predict $\pi = \Pr(Y = 1 \,|\, x)$
- Linear regression will give predicted values outside $0 \le \pi \le 1$
- Logistic model:
  - $\text{logit}(\pi_i) \equiv \log[\pi/(1-\pi)]$ avoids this problem
  - logit is interpretable as "log odds" that $Y = 1$
- Probit (normal transform) model $\rightarrow$ similar predictions, but is less interpretable



Logistic regression models: Binary response

Quantitative predictor: Linear and Logit regression on age

- Except in extremes, linear and logistic models give similar predicted values



Logistic regression models: Binary response

- For a binary response, $Y \in (0,1)$, let $\mathbf{x}$ be a vector of p regressors, and π_i be the probability, $\Pr(Y = 1 \,|\, \mathbf{x})$.
- The logistic regression model is a linear model for the *log odds*, or *logit* that $Y = 1$, given the values in $\mathbf{x}$,

$$\begin{aligned} \text{logit}(\pi_i) \equiv \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha + \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} \\ &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \end{aligned}$$

- An equivalent (non-linear) form of the model may be specified for the probability, π_i, itself,

$$\pi_i = \{1 + \exp(-[\alpha + \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}])\}^{-1}$$

- The logistic model is a *linear model* for the log odds, but also a *multiplicative* model for the odds of "success,"

$$\frac{\pi_i}{1-\pi_i} = \exp(\alpha + \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}) = \exp(\alpha)\exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta})$$

so, increasing x_{ij} by 1 increases $\text{logit}(\pi_i)$ by β_j, and multiplies the odds by e^{β_j}.

Logistic regression models: Binary response

Fitting

`PROC LOGISTIC` (or `ROBUST` macro— M-estimation)

- Data:
  - Frequency form (from `PROC FREQ`)— when all predictors are discrete
  - Case form— when any predictors are quantitative
- Models:
  - `CLASS` statement (V7+)— no need for dummy variables
    - discrete predictors
    - can specify *order* and *parameterization* (effect, polynomial, reference cell)
  - `MODEL` statement— allows `GLM` syntax, e.g.,

```
proc logistic;
  class Sex Treat;
  model Better = Sex | Treat | Age @2;
```

  - $\Rightarrow$ `Better = Sex Treat Age Sex*Treat Sex*Age Treat*Age`

Logistic regression models: Binary response

Visualization

- Goal: *see* and *understand* the data and fitted model
- LOGODDS macro: Plot observed responses, fitted and smoothed probabilities
- Model plots:
  - OUTPUT statement →
    - fitted $\hat{\pi}_i$, lower/upper $(1-\alpha)$ CI, and/or
    - fitted logit, $(\alpha + \mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}) \pm z_{1-\alpha/2} se(\text{logit})$
  - Plot with standard procedures (PROC GCHART, GPLOT)
  - Utility macros (BARS, LABEL, POINTS, PSCALE, etc.) for custom displays
- Effect plots— plot hierarchical subset of effects, averaging over those not included.
- INFLOGIS macro: Influence plots for logistic regression models
- ADDVAR macro: Added variable plots for new predictors or transformations of old

Example: Arthritis treatment data

- Predictors: Sex, Treatment (treated, placebo), Age
- Response: improvement (none, some, marked)
  - Consider first as binary response: None vs. (Some or Marked)='Better'
- Data in case form:

arthrit.sas

```
data arthrit;
   length treat $7. sex $6. ;
   input id treat $ sex $ age improve @@ ;
   case = _n_;
   better  = (improve > 0);  *-- Make binary response;
datalines ;
57 Treated Male   27 1   9 Placebo Male   37 0
46 Treated Male   29 0  14 Placebo Male   44 0
77 Treated Male   30 0  73 Placebo Male   50 0
  ... (observations omitted)
56 Treated Female 69 1  42 Placebo Female 66 0
43 Treated Female 70 1  15 Placebo Female 66 1
                        71 Placebo Female 68 1
                         1 Placebo Female 74 2
;
```


LOGODDS macro: Empirical logit plots

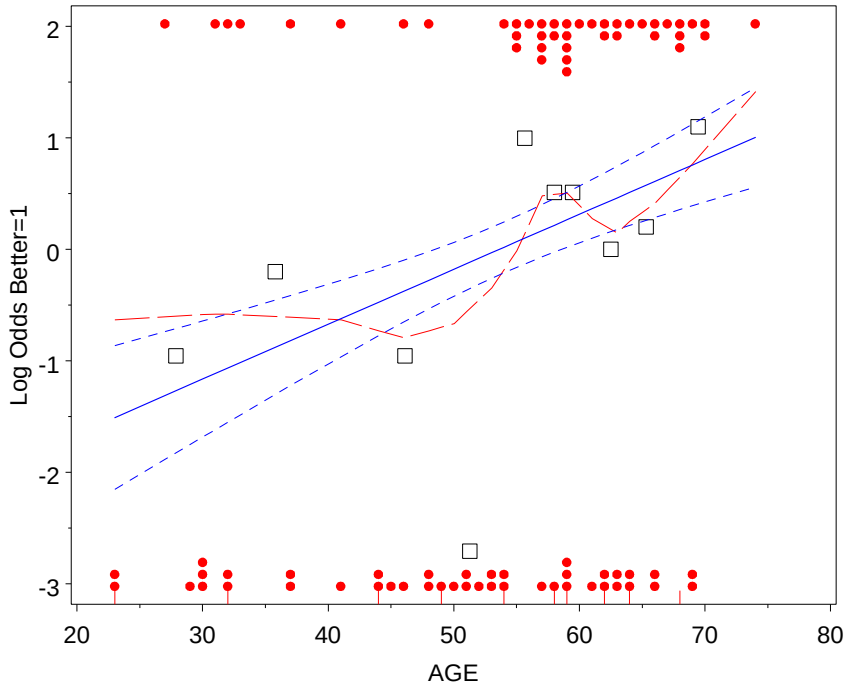
Problems with visualizing discrete outcomes:

- **Linearity**: Is a linear relation realistic?
- **Smoothing**: Discrete data often requires smoothing to see!

The LOGODDS macro:

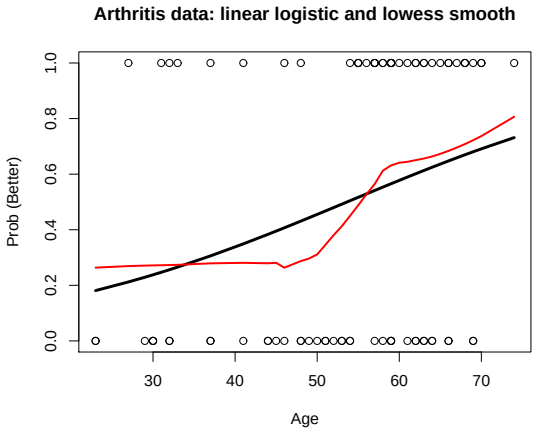
- Show the data: Plot (0/1) responses [stacked or jittered]
- Divide X into groups (e.g., deciles), emprical logit, log $\left(\frac{y_i+1/2}{n_i-y_i+1/2}\right)$, for each
- Linear logistic regression, plus smoothed curve (LOWESS macro)

```
%include catdata(arthrit);
%logodds(data=arthrit,
    x=age, y=Better,  /* vars to plot               */
    smooth=0.5,       /* LOWESS smoothing parameter */
    plot=logit);      /* plot on logit scale        */
```

Smoothing the binary observations

Can also use direct smoothing:



- SAS: `PROC LOESS`, `lowess` macro; R: `lowess()`
- There is a hint that the relation may be non-linear
- But data is thin at the extremes

PROC LOGISTIC: Model fitting and plotting

- Specify ordering of response levels (`order=` or `descending` options)
- Specify parameterizations for `CLASS` variables
- `OUTPUT` statement to get fitted logits and probabilities

glogist1c.sas ···

```
1
2 proc logistic data=arthrit descending;
3    class sex (ref=last) treat (ref=first) / param=ref;
4    model better = sex  treat  age;
5    output out=results
6           p=prob l=lower u=upper
7           xbeta=logit stdxbeta=selogit / alpha=.33;
```

The output includes:

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	6.2576	0.0124
treat	1	10.7596	0.0010
age	1	5.5655	0.0183

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.5033	1.3074	11.8649	0.0006
sex	Female	1	1.4878	0.5948	6.2576	0.0124
treat	Treated	1	1.7598	0.5365	10.7596	0.0010
age		1	0.0487	0.0207	5.5655	0.0183

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex Female vs Male	4.427	1.380	14.204
treat Treated vs Placebo	5.811	2.031	16.632
age	1.050	1.008	1.093

Parameter estimates (reference cell coding):

- $\beta_1 = 1.49 \Rightarrow$ Females $e^{1.49}$=4.43 $\times$ more likely better than Males
- $\beta_2 = 1.76 \Rightarrow$ Treated $e^{1.76}$=5.81 $\times$ more likely better than Placebo
- $\beta_3 = 0.0487 \Rightarrow$ odds ratio=1.05 $\Rightarrow$ odds of improvement increase 5% each year. Over 10 years, odds of improvement $= e^{10 \times 0.0486} = 1.63$, a 63% increase.

PROC LOGISTIC: Full-model plots

Full-model plots display the fitted (predicted) values over *all combinations* ofpredictors:

- Plot fitted values from the dataset specified on the `OUTPUT` statement
- Plot either predicted probabilities or logits
- Confidence intervals or standard errors allow showing error bars

The first few observations from the `results` dataset:

id	sex	treat	age	better	prob	lower	upper	logit	selogit
57	Male	Treated	27	1	0.194	0.103	0.334	-1.427	0.758
9	Male	Placebo	37	0	0.063	0.032	0.120	-2.700	0.725
46	Male	Treated	29	0	0.209	0.115	0.350	-1.330	0.728
14	Male	Placebo	44	0	0.086	0.047	0.152	-2.358	0.658
77	Male	Treated	30	0	0.217	0.122	0.357	-1.281	0.713
73	Male	Placebo	50	0	0.112	0.065	0.188	-2.066	0.622
...									

- `prob`– predicted probabilities, with CI (`lower` ,`upper`)
- `logit`– predicted logit, with standard error `selogit`

PROC LOGISTIC: Full-model plots

Basic plots:

- Plot either logit or probability vs. one predictor (continuous or most levels)
- Separate curves for one factor (= `factor`)
- Separate panels for all others (`BY` statement)

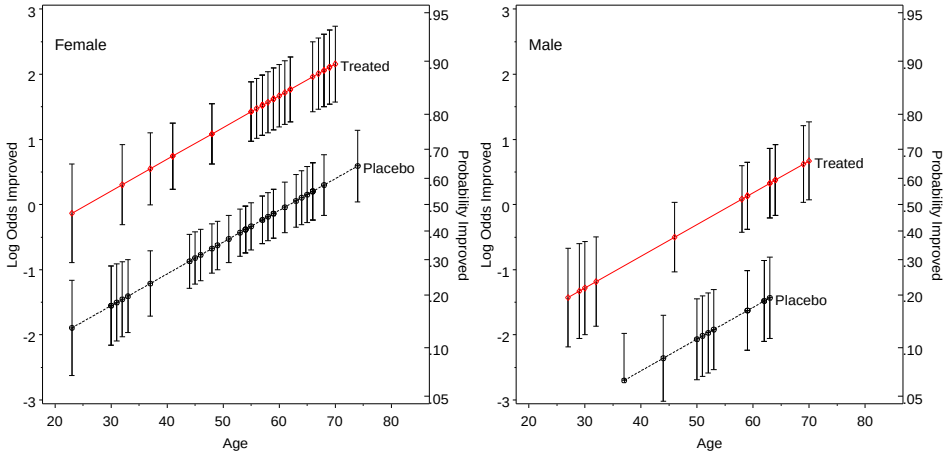
```
1  proc gplot data=results;
2     plot (logit prob) * age = treat;      /* separate curves */
3     by sex;                               /* separate panels */
4     symbol1 v=circle i=join l=3 c=black; /* placebo */
5     symbol2 v=dot    i=join l=1 c=red;   /* treated */
```

- `SYMBOL` statement— define the point value (`v=`), interpolate option (`i=`), line style (`l=`), color (`c=`), etc.

PROC LOGISTIC: Model plots

Enhanced plots:

- Plot on logit scale, with probability scale at right (`PSCALE` macro)
- Show 67% error bars $\approx \pm 1$ se (`BARS` macro)
- Custom legend and panel labels (`LABEL` macro)



PROC LOGISTIC: Full-model plots

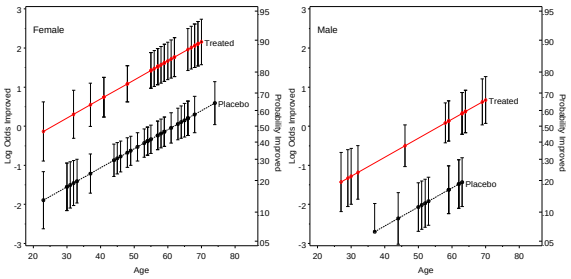
Enhanced plots:

··· glogist1c.sas ···

```
9    *-- Error bars, on logit scale;
10 %bars(data=results, var=logit,
11    class=age, cvar=treat, by=age,
12    barlen=selogit, out=bars);
13
14   *-- Custom legends and panel labels;
15 %label(data=results, y=logit, x=age, xoff=1, cvar=treat,
16    by=sex, subset=last.treat, out=label1, pos=6, text=treat);
17 %label(data=results, y=2.5, x=20, size=2,
18    by=sex, subset=first.sex, out=label2, pos=6, text=sex);
19
20   *-- Probability scales at right;
21 %pscale(out=pscale,
22    byvar=sex, byval=%str('Female','Male'));
23
24   *-- Join ANNOTATE datasets;
25 data bars;
26    set label1 label2 bars pscale;
27 proc sort;
28    by sex;
```


··· glogist1c.sas

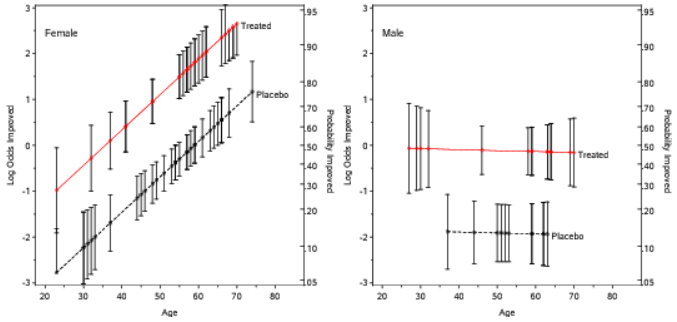
```
30 title ' '
31    h=1.8 a=-90 'Probability Improved'    /* right axis label   */
32    h=2.5 a=-90 ' ';                      /* extra space        */
33 goptions hby=0;                          /* suppress BY values */
34 proc gplot data=results;
35    plot logit * age = treat /
36         vaxis=axis1 haxis=axis2 hm=1 vm=1
37         nolegend anno=bars frame;
38    by sex;
39    axis1 label=(a=90 'Log Odds Improved')
40         order=(-3 to 3);
41    axis2 order=(20 to 80 by 10) offset=(2,6);
42    symbol1 v=+ i=join l=3 c=black;
43    symbol2 v=- i=join l=1 c=red;
44    label age='Age';
45 run;
```



Models with interactions

- **Plotting fitted values**
  - Only need to change the MODEL statement
  - Output dataset automatically incorporates all model terms
  - Plotting steps remain *exactly* the same

```
proc logistic data=arthrit descending;
   class sex (ref=last) treat (ref=first) / param=ref;
   model  better = treat sex  | age @2;;
   output out=results p=prob l=lower u=upper
          xbeta=logit stdxbeta=selogit / alpha=.33;
```



Effect plots: basic ideas

Show a given effect (and low-order relatives) controlling for other model effects.

Data

	x1	x2	sex	x1x2	y	yhat
1	1	1	F	1	4.73	4.46
2	2	1	M	0	6.10	5.55
3	3	1	F	-1	4.32	4.34
4	1	1	F	1	4.84	4.46
5	2	1	F	0	4.73	4.40
...	...	...	...	...	...	...
29	2	2	M	0	6.10	6.15
30	3	2	F	1	6.71	7.14

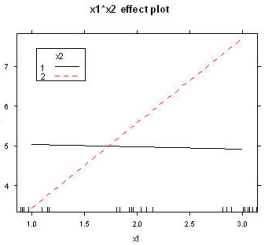
- Fit data: $\mathbf{X}\hat{\boldsymbol{\beta}} \Rightarrow \hat{\mathbf{y}}$
- Score data $\mathbf{X}^* \hat{\boldsymbol{\beta}} \Rightarrow \hat{\mathbf{y}}^*$
  - plot vars: vary over range
  - control vars: fix at means

Score data

	x1	x2	sex	x1:x2	y	yhat*
31	1	1	0.5	1	NA	5.030
32	2	1	0.5	2	NA	4.971
33	3	1	0.5	3	NA	4.912
34	1	2	0.5	2	NA	3.437
35	2	2	0.5	4	NA	5.574
36	3	2	0.5	6	NA	7.710

plot vars control vars

plot



Effect plots for generalized linear models: Details

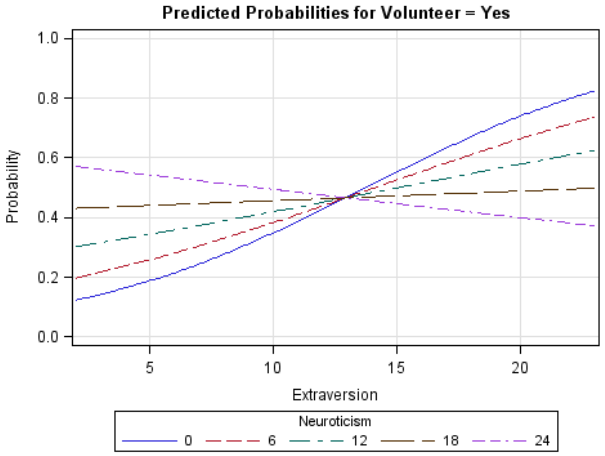
- For simple models, full model plots show the complete relation between response and *all predictors*.
- Fox (1987)— For complex models, often wish to plot a specific main effect or interaction (including lower-order relatives)— *controlling for other effects*
  - Fit full model to data with linear predictor (e.g., logit) $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ and link function $g(\boldsymbol{\mu}) = \boldsymbol{\eta} \rightarrow$ estimate $\mathbf{b}$ of $\boldsymbol{\beta}$ and covariance matrix $\widehat{V(\mathbf{b})}$ of $\mathbf{b}$.
  - Vary each predictor in the term over its' range
  - Fix other predictors at "typical" values (mean, median, proportion in the data)
  - $\rightarrow$ "effect model matrix," $\mathbf{X}^*$
  - Calculate fitted effect values, $\hat{\boldsymbol{\eta}}^* = \mathbf{X}^*\mathbf{b}$.
  - Standard errors are square roots of $\text{diag}(\mathbf{X}^* \widehat{V(\mathbf{b})} \mathbf{X}^{*\mathsf{T}})$
  - Plot $\hat{\boldsymbol{\eta}}^*$, or values transformed back to scale of response, $g^{-1}(\hat{\boldsymbol{\eta}}^*)$.
- *Note*: This provides a general means to visualize interactions in *all* linear and generalized linear models.

Effect plots software

- General method
  - Create a grid of values for predictors in the effect (EXPGRID macro)
  - Fix other predictors at "typical" values (mean, median, proportion in the data)
  - Concatenate grid with data
  - Fit model $\rightarrow$ output data set $\rightarrow$ fitted values in the grid
  - Standard errors automatically calculated
  - Plot fitted values in the grid
- EFFPLOT macro
  - Works with PROC REG, PROC GLM, PROC LOGISTIC, PROC GENMOD
  - Uses MEANPLOT macro to do the plotting
  - Some limitations – can't plot correct standard errors
- SAS 9.3 ODS Graphics
  - Several procedures now do effects-like plots: LOGISTIC, GLM, GLIMMIX
  - Easy; PROC LOGISTIC quite flexible
- R: effects package
  - Most general: Handles linear models (lm()), generalized linear models (glm()), multinomial (multinom()) and proportional-odds (polr()) models.
  - allEffects(model) calculates effects for all high-order terms in model
  - plot(allEffects(model)) plots them

Effect plots: Example

- Cowles and Davis (1987)— Volunteering for a psychology experiment
  - Predictors: Sex, Neuroticism, Extraversion
  - → strong interaction, Neuroticism × Extraversion

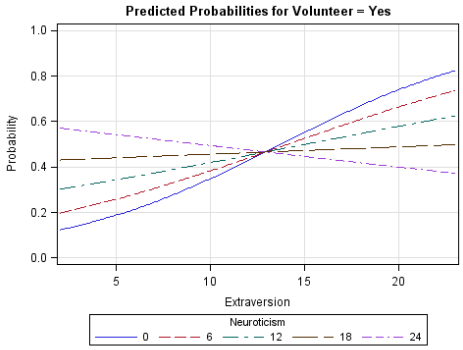


Fit computed at Sex=1.5

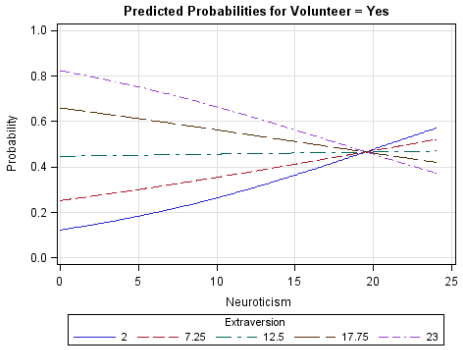
Effect plots: SAS 9.3 ODS Graphics

cowles-logistic-eff.sas

```
1 proc logistic data=cowles outest=parm  descending ;
2     class Sex;
3     model Volunteer = Sex   Extraver | Neurot / lackfit ;
4     effectplot slicefit(x=Extraver sliceby=Neurot) / at(sex=1.5) noobs;
5     effectplot slicefit(x=Neurot sliceby=Extraver) / at(sex=1.5) noobs;
6     effectplot contour(x=Neurot y=Extraver) / at(sex=1.5) noobs;
7     run;
```



Fit computed at Sex=1.5

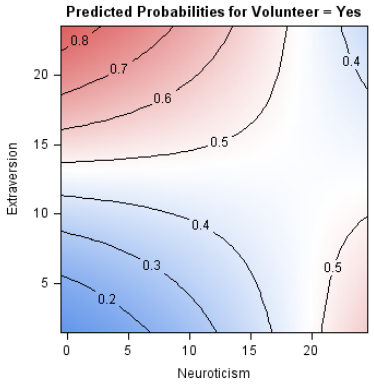


Fit computed at Sex=1.5

Effect plots: SAS 9.3 ODS Graphics

cowles-logistic-eff.sas

```
1 proc logistic data=cowles outest=parm  descending ;
2     class Sex;
3     model Volunteer = Sex   Extraver | Neurot / lackfit ;
4     effectplot contour(x=Neurot y=Extraver) / at(sex=1.5) noobs;
5     run;
```

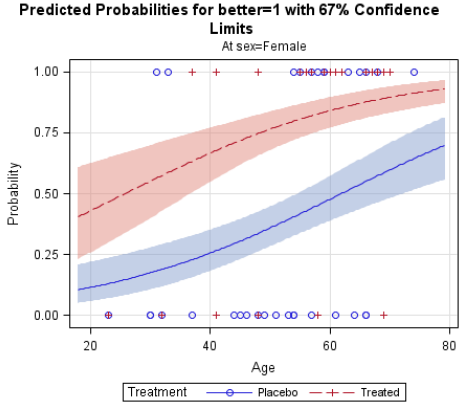


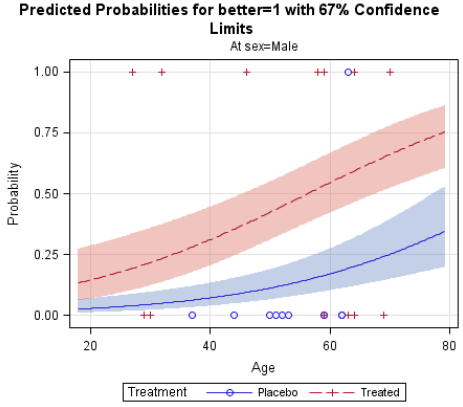
Fit computed at Sex=1.5

SAS 9.2: ODS Graphics

arthritis-logistic-ods.sas

```
1 %include catdata(arthrit);
2 ods graphics on;
3 proc logistic data=arthrit descending
4       plots(only)=(effect(plotby=sex sliceby=treat showobs clband alpha=0.33));
5    class sex (ref=last) treat (ref=first) / param=ref;
6    model  better = sex  treat  age / clodds=wald;
7 run;
8 ods graphics off;
```





Effect plots with the effects package in R

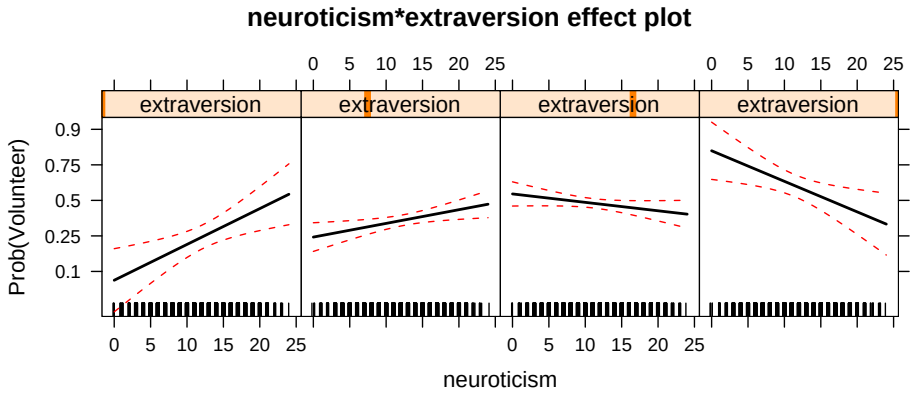
```
> library(effects)  ## load the effects package
> data(Cowles)
> mod.cowles <- glm(volunteer ~ sex + neuroticism*extraversion,
+     data=Cowles, family=binomial)
> summary(mod.cowles)
```

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.358207   0.501320  -4.704 2.55e-06 ***
sexmale                -0.247152   0.111631  -2.214  0.02683 *
neuroticism             0.110777   0.037648   2.942  0.00326 **
extraversion            0.166816   0.037719   4.423 9.75e-06 ***
neuroticism:extraversion -0.008552   0.002934  -2.915  0.00355 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1933.5  on 1420  degrees of freedom
Residual deviance: 1897.4  on 1416  degrees of freedom
AIC: 1907.4
```


Effect plots with the effects package in R

Calculate effects for all model terms, plot neuro:extra:

```
> eff.cowles <- allEffects(mod.cowles,
+      xlevels=list(neuroticism=0:24,
+                   extraversion=seq(0, 24, 8)))
>
> plot(eff.cowles, 'neuroticism:extraversion', ylab="Prob(Volunteer)",
+     ticks=list(at=c(.1,.25,.5,.75,.9)), layout=c(4,1), aspect=1)
```



Extended example: Arrests for Marihuana Possession

Context & background

- In Dec. 2002, the *Toronto Star* examined the issue of racial profiling, by analyzing a data base of 600,000+ arrest records from 1996-2002.
- They focused on a subset of arrests for which police action was discretionary, e.g., simple possession of small quantities of marijuana, where the police could:
  - Release the arrestee with a summons— like a parking ticket
  - Bring to police station, hold for bail, etc.— harsher treatment
- Response variable: `released` – `Yes, No`
- Main predictor of interest: skin-`colour` of arrestee (black, white)

Extended example: Arrests for Marihuana Possession

Data

Control variables:

- `year`, `age`, `sex`
- `employed`, `citizen` – `Yes, No`
- `checks` — Number of police data bases (previous arrests, previous convictions, parole status, etc.) in which the arrestee's name was found.

```
> library(effects)
> data(Arrests)
> some(Arrests)
```

	released	colour	year	age	sex	employed	citizen	checks
915	No	Black	2001	35	Male	Yes	Yes	4
1568	Yes	White	2002	21	Male	Yes	Yes	0
2981	Yes	White	2000	23	Male	Yes	Yes	2
3381	Yes	Black	1998	23	Male	No	Yes	2
3516	Yes	White	2002	22	Male	Yes	Yes	0
4128	No	White	2001	29	Male	Yes	Yes	1
4142	Yes	Black	1998	23	Male	Yes	Yes	3
4634	Yes	White	2001	18	Male	Yes	Yes	0
4732	Yes	White	1999	21	Male	Yes	Yes	3
5183	Yes	White	1999	19	Male	Yes	Yes	0

Extended example: Arrests for Marihuana Possession

Model

To allow possibly non-linear effects of `year`, we treat it as a factor:

```
> Arrests$year <- as.factor(Arrests$year)
```

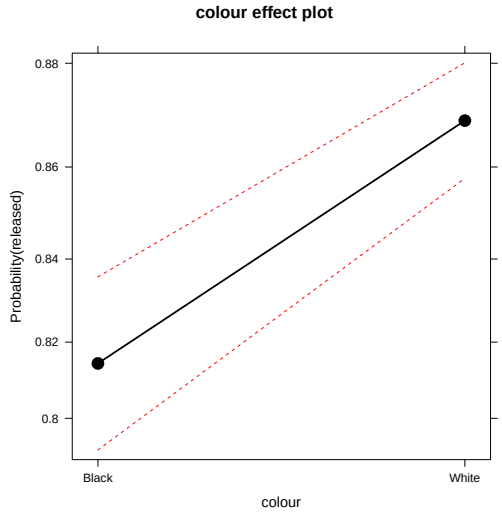
Logistic regression model with all main effects, plus interactions of `colour:year` and `colour:age`

```
> arrests.mod <- glm(released ~ employed + citizen + checks + colour *
+     year + colour * age, family = binomial, data = Arrests)
> Anova(arrests.mod)

Analysis of Deviance Table (Type II tests)

Response: released
            LR Chisq Df Pr(>Chisq)
employed      72.673  1  < 2.2e-16 ***
citizen       25.783  1  3.820e-07 ***
checks       205.211  1  < 2.2e-16 ***
colour        19.572  1  9.687e-06 ***
year           6.087  5  0.2978477
age            0.459  1  0.4982736
colour:year   21.720  5  0.0005917 ***
colour:age    13.886  1  0.0001942 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Effect plots: colour

Evidence for different treatment of blacks and whites ("racial profiling"), controlling (adjusting) for other factors

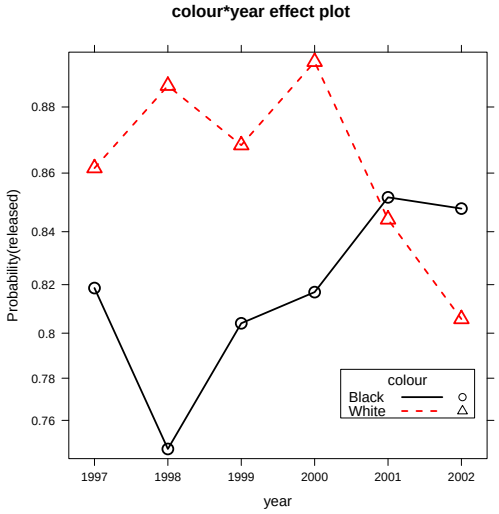
```
> plot(effect("colour", arrests.mod), multiline = FALSE, ylab = "Probability(release
```



Effect plots: Interactions

The story turned out to be more nuanced than reported by the *Toronto Star*, as shown in effect plots for interactions with colour.

```
> plot(effect("colour:year", arrests.mod), multiline = TRUE, ...)
```

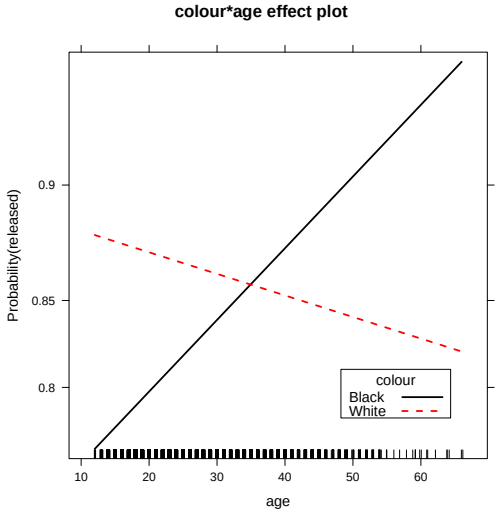


- Up to 2000, strong evidence for differential treatment of blacks and whites
- Also evidence to support Police claim of effect of training to reduce racial effects in treatment

Effect plots: Interactions

The story turned out to be more nuanced than reported by the *Toronto Star*, as shown in effect plots for interactions with colour.

```
> plot(effect("colour:age", arrests.mod), multiline = TRUE, ...)
```

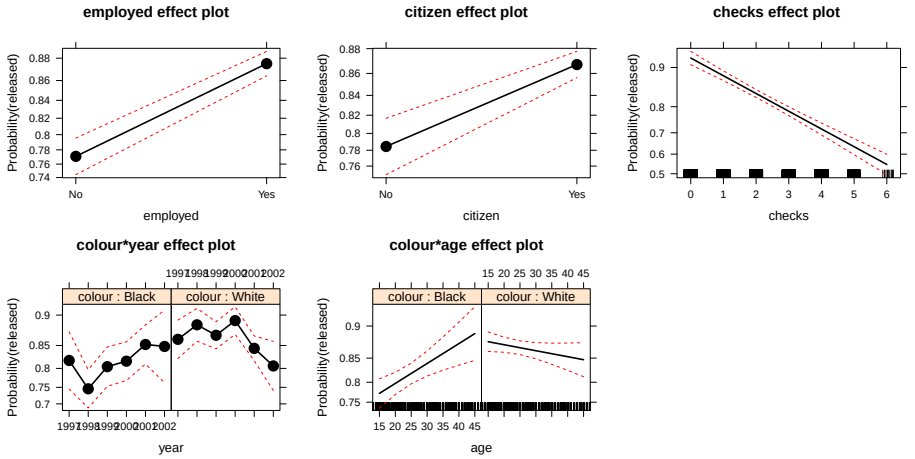


- Opposite age effects for blacks and whites:
- Young blacks treated *more* harshly than young whites
- Older blacks treated *less* harshly than older whites

Effect plots: allEffects

All model effects can be viewed together using `plot(allEffects(mod))`

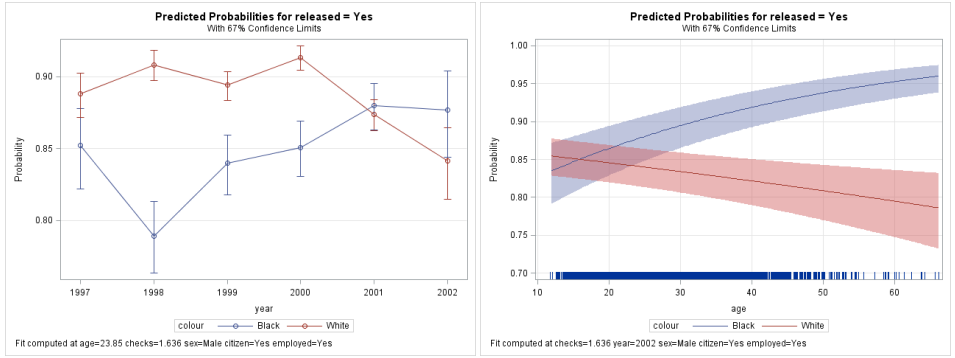
```
> arrests.effects <- allEffects(arrests.mod, xlevels = list(age = seq(15,
+     45, 5)))
> plot(arrests.effects, ylab = "Probability(released)", ask = FALSE)
```



Effect plots: SAS

Arrests-logistic.sas

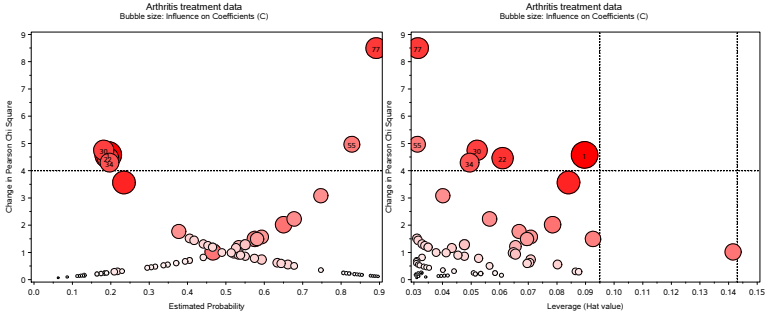
```
1 proc logistic data=arrests descending;
2    class colour year sex citizen employed;
3    model released =  colour|year colour|age  sex  employed  citizen  checks;
4    effectplot interaction (x=year sliceby=colour) / clm alpha=0.33 noobs;
5    effectplot slicefit (x=age sliceby=colour) / clm alpha=0.33 obs(fringe jitter);
6    run;
```



NB: These plots are computed at average levels of quantitative variables, but at reference levels of class variables: Sex=Male, citizen=Yes, employed=Yes

Influence measures and diagnostic plots

- **Leverage**: Potential impact of an individual case $\sim$ distance from the centroid in space of predictors
- **Residuals**: Which observations are poorly fitted?
- **Influence**: Actual impact of an individual case $\sim$ leverage $\times$ residual
  - **C**, **CBAR** – analogs of Cook's D in OLS $\sim$ standardized change in regression coefficients when *i*-th case is deleted.
  - **DIFCHISQ**, **DIFDEV** – $\Delta\chi^2$ when *i*-th case is deleted.



Influence measures and diagnostic plots

`PROC LOGISTIC`: printed output with the `influence` option

```
1 proc logistic data=arthrit descending;
2    model better = sex treat age / influence;
```

The LOGISTIC Procedure

Regression Diagnostics

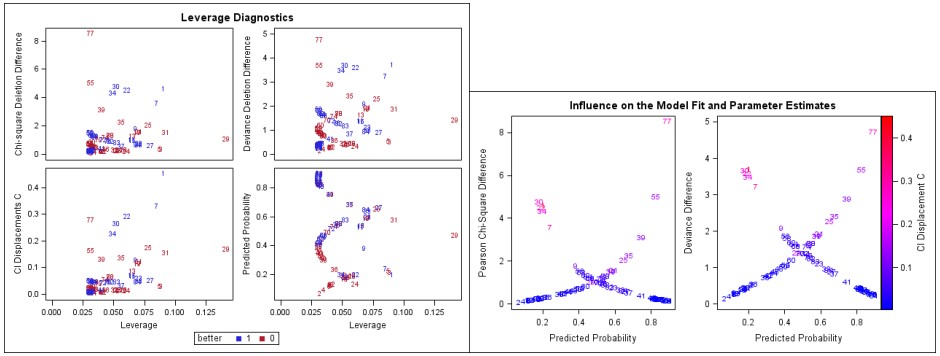
Case Number	Covariates: Sex Female	Covariates: Treatment Treated	Covariates: Age	Pearson Residual	Deviance Residual	Hat Matrix Diagonal	Intercept DfBeta	sexFemale DfBeta	treatTreated DfBeta	age DfBeta	Confidence Interval Displacement C
1	0	1.0000	27.0000	2.0415	1.8124	0.0897	0.5585	-0.3369	0.1096	-0.5016	0.4510
2	0	0	37.0000	-0.2593	-0.3607	0.0310	-0.0410	0.0344	0.0260	0.0255	0.00222
3	0	1.0000	29.0000	-0.5143	-0.6851	0.0876	-0.1361	0.0865	-0.0289	0.1203	0.0278
4	0	0	44.0000	-0.3075	-0.4251	0.0341	-0.0450	0.0464	0.0350	0.0221	0.00346
5	0	1.0000	30.0000	-0.5270	-0.7001	0.0865	-0.1369	0.0894	-0.0303	0.1200	0.0288
6	0	0	50.0000	-0.3559	-0.4884	0.0386	-0.0460	0.0595	0.0449	0.0145	0.00529
7	0	1.0000	32.0000	1.8072	1.7034	0.0840	0.4505	-0.3113	0.1086	-0.3869	0.3272
8	0	0	51.0000	-0.3647	-0.4998	0.0396	-0.0458	0.0620	0.0468	0.0126	0.00570
9	0	1.0000	46.0000	1.2848	1.3963	0.0668	0.1889	-0.2337	0.0985	-0.1158	0.1266

Too much output, doesn't highlight unusual cases, ...

Influence measures and diagnostic plots

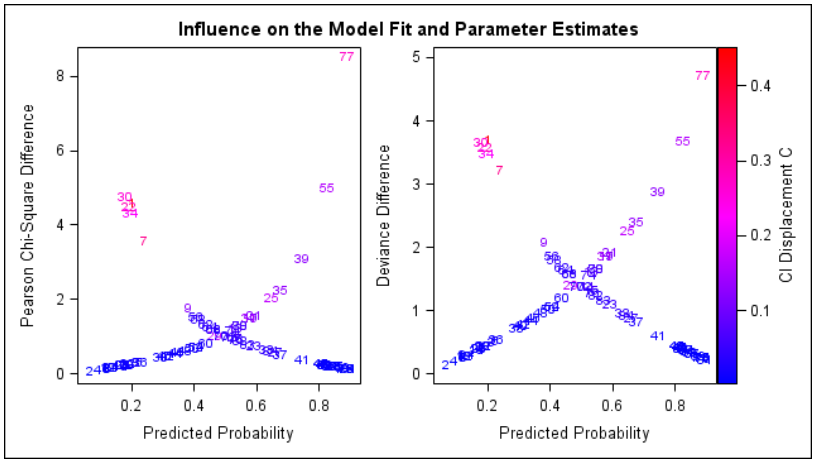
PROC LOGISTIC: plotting diagnostic measures with the `plots` option

```
proc logistic data=arthrit descending
        plots(only label)=(leverage dpc);
   class sex (ref=last) treat (ref=first) / param=ref;
   model  better = sex treat age ;
run;
```



Influence measures and diagnostic plots: Influence plots

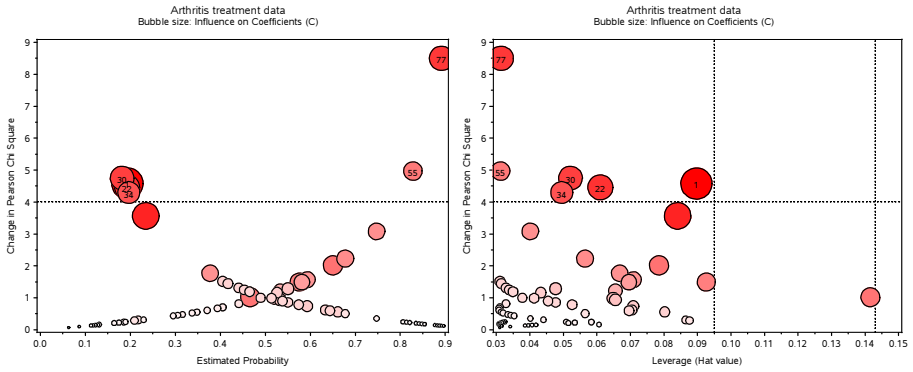
The option `plots(label)=dpc` gives plots of $\Delta\chi^2$ (DIFCHISQ, DIFDEV) vs. $\widehat{p}$
Points are colored according to the influence measure C.



The two bands of points correspond to better $= \{0, 1\}$

INFLOGIS macro

- Specialized version of `INFLGLIM` macro for logistic regression
- Plots a measure of change in χ^2 (DIFCHISQ or DIFDEV) vs. predicted probability or leverage.
- Bubble symbols show actual influence (C or CBAR)
- Shows standard cutoffs for "large" values
- Flexible labeling of unusual cases



INFLOGIS macro: Example

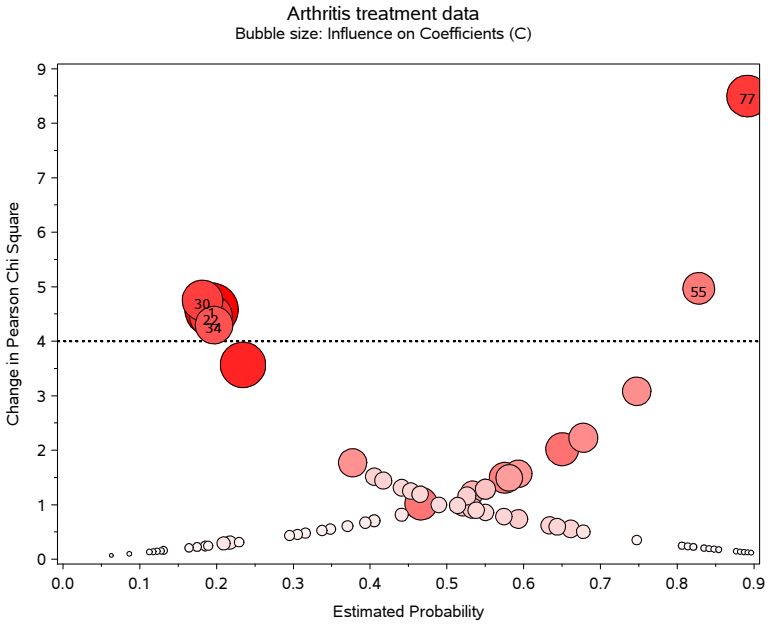
logist1b.sas

```
%include data(arthrit);
%inflogis(data=arthrit,
   class=sex treat,       /* CLASS variables */
   y=better,              /* response        */
   x=sex treat age,       /* predictors      */
   id=case,               /* case ID         */
   gy=DIFCHISQ,           /* graph ordinate  */
   gx=PRED HAT,           /* graph abscissas */
   loptions=descending);
```

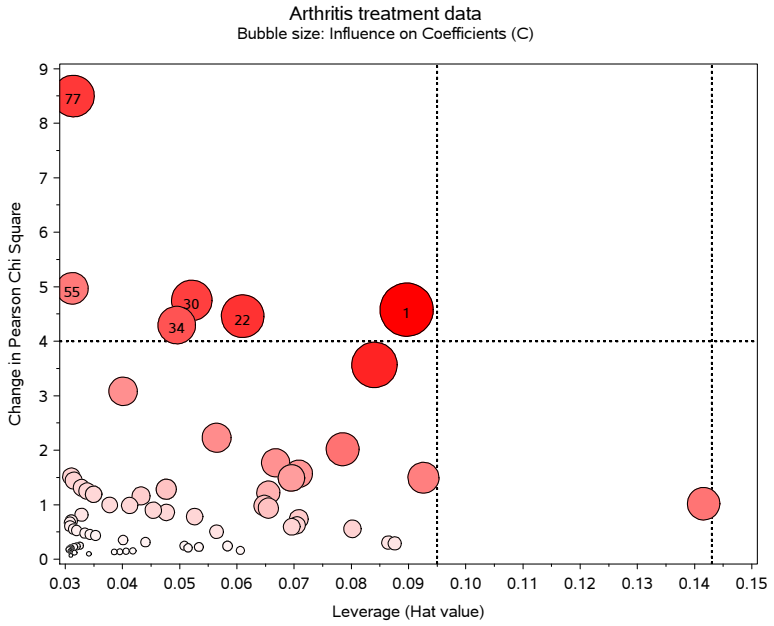
Printed output lists cases with "large" leverage, residual or influence:

case	better	sex	treat	age	pred	hat	difchisq	difdev	c
1	1	Male	Treated	27	.806	.09	4.578	3.695	0.451
22	1	Male	Placebo	63	.807	.06	4.460	3.565	0.290
30	1	Female	Placebo	31	.818	.05	4.749	3.657	0.261
34	1	Female	Placebo	33	.803	.05	4.296	3.464	0.224
55	0	Female	Treated	58	.172	.03	4.970	3.676	0.160
77	0	Female	Treated	69	.108	.03	8.498	4.712	0.276

INFLOGIS macro: Example



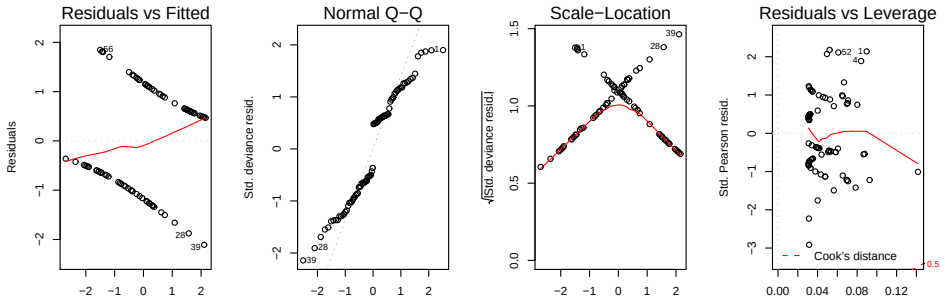
INFLOGIS macro: Example



Diagnostic plots in R

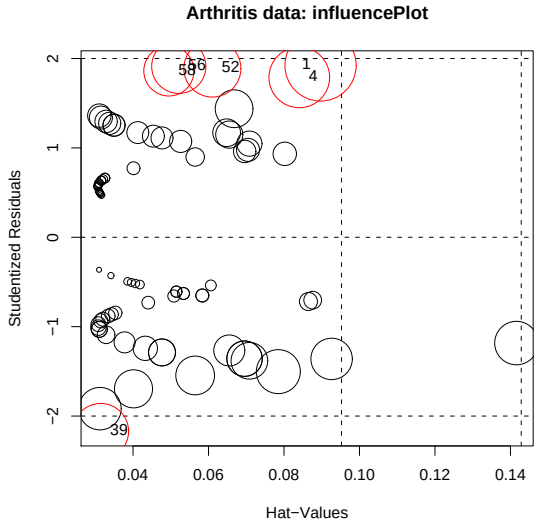
In R, plotting a `glm` object gives the "regression quartet"

```
arth.mod1 <- glm(Better ~ Age+Sex+Treatment,data=Arthritis,
                 family='binomial')
plot(arth.mod1)
```



Diagnostic plots in R

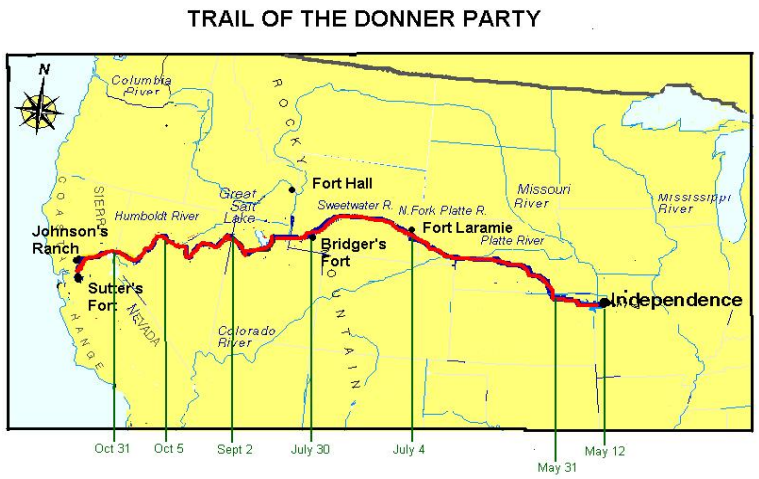
```
library(car)
influencePlot(arth.mod1)
```



Donner Party: A graphic tale of survival & influence

History:

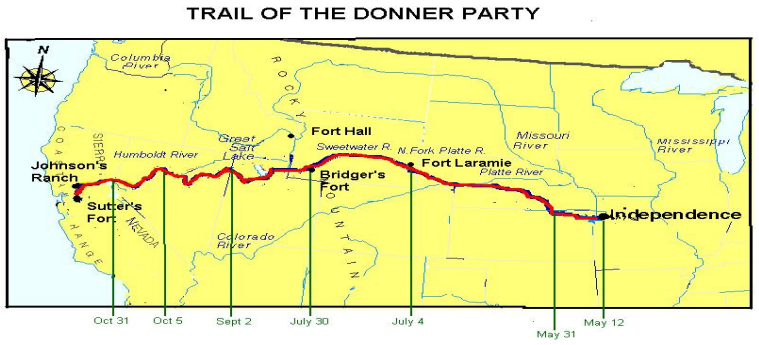
- Apr–May, 1846: Donner/Reed families set out from Springfield, IL to CA
- Jul: Bridger's Fort, WY, 87 people, 23 wagons



Donner Party: A graphic tale of survival & influence

History:

- "Hasting's Cutoff", untried route through Salt Lake Desert, Wasatch Mtns. (90 people)
- Worst recorded winter: Oct 31 blizzard— Missed by 1 day, stranded at "Truckee Lake" (now Donner's Lake, Reno)
  - Rescue parties sent out ("Dire necessity", "Forelorn hope", ...)
  - Relief parties from CA: 42 survivors (Mar–Apr, '47)



The Donner Party: Who lived and died?

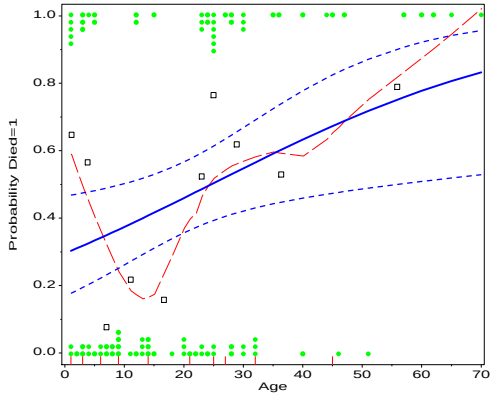
- Other analyses, e.g., (Ramsay and Schafer, 1997):
  - Log Odds (survive) $\sim$ linear with Age
  - Odds (survive $|$Women / survive $|$Men) $= 4.9$
  - (Ignored children)

NAME	AGE	MALE	SURVIVED	DEATH
Antoine	23	1	0	29DEC46
Breen, Edward	13	1	1	.
Breen, Margaret I.	1	0	1	.
Breen, James	5	1	1	.
Breen, John	14	1	1	.
Breen, Mary	40	0	1	.
Breen, Patrick	51	1	1	.
Breen, Patrick Jr.	9	1	1	.
Breen, Peter	3	1	1	.
Breen, Simon	8	1	1	.
Burger, Charles	30	1	0	27DEC46
Denton, John	28	1	0	26FEB47
Dolan, Patrick	40	1	0	27DEC46
Donner, Elitha Cumi	13	0	1	.
Donner, Eliza Poor	3	0	1	.
Donner, Elizabeth	45	0	0	14MAR47
Donner, Francis E.	6	0	1	.
Donner, George	62	1	0	18MAR47
Donner, George Jr.	9	1	1	.
...				

Empirical logit plots

- Is a linear logistic model satisfactory for these data?
- Discrete data often requires smoothing to see!

```
1 %logodds(data=donner, y=Died, x=Age, smooth=0.5);
```



$\Rightarrow$ relation with Age is quadratic: youngest and oldest most likely to perish.

Quadratic model?

- Fit: Pr(Death) $\sim$ Age + Age^2 + Male
- Statistical evidence for Age^2 equivocal:
  - Wald $\chi^2_{(1)} = 2.84, p = 0.09$; but
  - LR $G^2_{(1)} = 4.40, p = 0.03$. ...

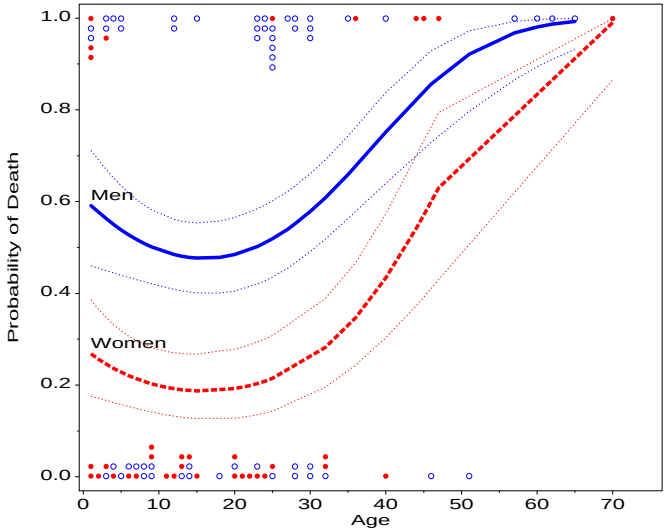
Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	-1.7721	0.5673	9.7588	0.0018
AGE	0.0168	0.0184	0.8355	0.3607
AGE2	0.00208	0.00123	2.8439	0.0917
MALE	1.3745	0.5066	7.3617	0.0067

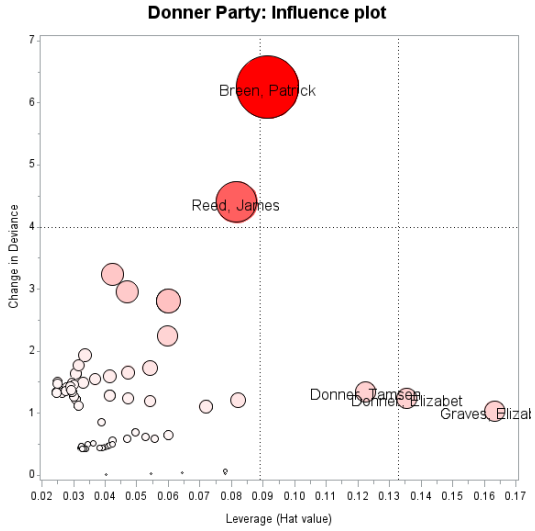
- Males: exp(1.3745) = 3.95 times as likely to die, controlling for Age, Age^2

Quadratic model?

- Visual evidence is persuasive (but the data are thin at older ages)



Who was influential?



Why are they influential?

NAME	Died	Age	M?	PRED	StuRes	Hat	DifDev	C
Breen, Patrick	0	51	1	.921	-2.365	.09	6.25	1.294
Reed, James	0	46	1	.856	-2.054	.08	4.40	0.575
Donner, Elizabeth	1	45	0	.571	1.139	.14	1.24	0.136
Donner, Tamsen	1	44	0	.541	1.183	.12	1.35	0.135
Graves, Elizabeth	1	47	0	.630	1.050	.16	1.04	0.137

- Patrick Breen, James Reed: Older men who survived
- Elizabeth & Tamsen Donner, Elizabeth Graves: Older women who survived
- Moral lessons of this story:
  - Don't try to cross the Donner Pass in late October; if you do, bring food
  - Plots of fitted models show *only* what is included in the model
  - Discrete data often need smoothing (or non-linear terms) to see the pattern
  - Always examine model diagnostics — preferably graphic

Summary: Part 4

- **Logit models**
  - Analogous to ANOVA models for a binary response
  - Equivalent to loglinear model, including interaction of all predictors
  - Fitting: SAS: `PROC CATMOD`, `PROC LOGISTIC`; R: `glm()`
  - Visualization: plot fitted logits (or probabilties) vs. factors (`CATPLOT` macro)
- **Logistic regression**
  - Analogous to regression models for a binary response
  - Coefficients: increment to log odds / ΔX; $\exp \beta \sim$ multiplier of odds per ΔX
  - Discrete responses: smoothing often useful
  - Visualization: plot fitted logits (or probabilties) vs. predictors
- **Effect plots**
  - Plot a main effect or interaction in the context of a more complex model
  - Shows that effect *controlling for* (averaged over) all other model effects
  - SAS: `EFFPLOT` macro; R: `effects` package
- **Influence & diagnostics**
  - Influence plots highlight *unusual* cases/cells — large impact on fitted model
  - Probability plots of residuals help to check model assumptions
  - SAS: `INFLGLIM` macro, `HALFNORM` macro; R: `plot(my.glm)`, `influencePlot(my.glm)`