Statistics & Experimental Design with R

Barbara Kitchenham
Keele University

General Linear Models

Logistic and Poisson Regression

Logistic Regression

- Predicts a categorical response variable from one or more explanatory variables
- Usually a binomial response variable
  - Used to predict module fault-proneness
  - Probability of project failing
  - Model is

$$log_e\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{j} \beta_j X_j$$

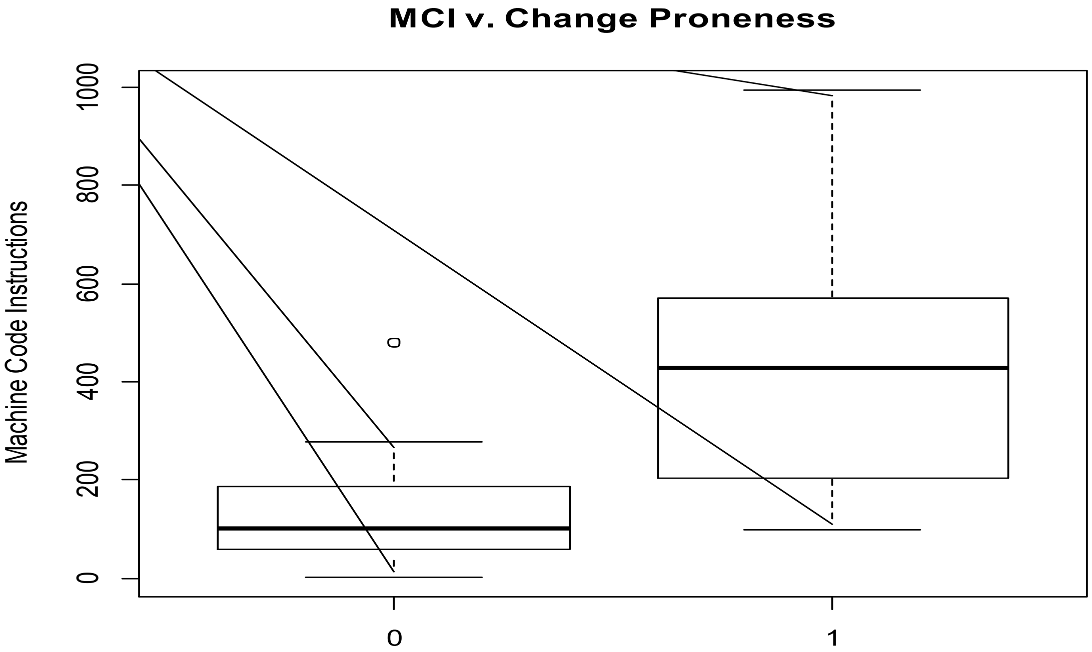
  - Outcome variable is the log odds also called logit
  - If it is equally likely that an object does or does not have a property the odds=1 and logit=0

General Linear Models (GLM)

- Ordinary regression and logistic regression
  - Both examples of linear models
- R uses the general linear modelling function glm() to handle logistic and Poisson regression
- GLM fits models of the form $g(\mu_Y) = \beta_0 + \sum_{i=1}^{j} \beta_j X_j$
- Where $g(\mu_Y)$ is a function of the conditional mean called the link function
- Link function for the binomial is the logit
- R Function is
  - glm(y~x1+x2+x3, family=binomial(link="logit"), data=mydata)

Example

- Data set with counts of changes
- More than two changes during development labelled
  - Change Prone (18 of 40 modules) i.e. Prior Probability=0.45



Logarithmic Regression Results

- If you have non-significant variables in a model, you can fit a reduced model
  - Compare the two fits using R function anova()
    - anova(reducedfit,fullfit,test="Chisq")
    - Chi-squared not significant suggests reduced fit better
    - Works if reducedfit is a subset of fullfit
  - Also check AIC values
- Check for "overdispersion"
  - Residual_Deviance /Residual_df
    - Means that variation is larger than expected given the model being fitted
    - Allows for heteroscedasticity
    - Problem if larger than 1, 34.369/38<1 for example

Two models

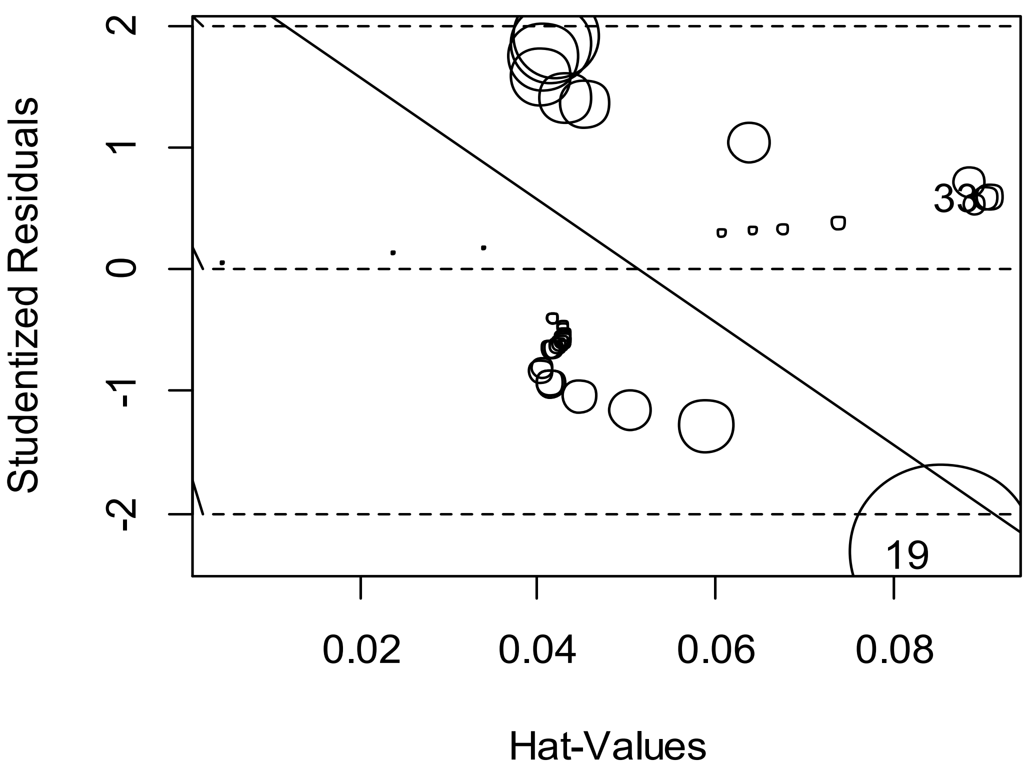
Coefficient	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.192	1.1933	-2.675	0.00747 **
MCI	0.02264	0.01127	2.008	0.04461 *
Loc	0.02184	0.01530	1.427	0.15346
Called	0.10769	0.2095	0.514	0.60731
Data	0.28992	0.4873	0.595	0.55189

Residual deviance: 31.200 on 35 degrees of freedom AIC=41.2

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.4899	0.7649	3.255	0.00113 **
MCI	0.009782	0.003156	3.100	0.00194 **

Residual deviance: 34.369 on 38 degrees of freedom AIC: 38.369

Influence Plot



Analysis of Deviance

Model 1: CngProne ~ MCI					
Model 2: CngProne ~ MCI + Loc + Called + Data					
	Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	38	34.369			
2	35	31.200	3	3.1693	0.3663

Confusion Matrix

- Assigned to most probable category
- How good is assignment?
  - Chi-squared test = 14.43 (p=0.000146)
  - Correlation=0.6
- Should use a Bayesian approach if you have unequal prior probabilities for the categories

Assigned	Actual		Total
	Change-Prone	Not Change-Prone	
Change-Prone	12	2	14
Not Change-Prone	6	20	26
Totals	18	22	40

Other R functions

- Robust Logistic Regression
  - glmRob() in "robust" package
- Mulitnomial Regression
  - If the response variable has more than two unordered categories
  - Use mlogit() in the "mlogit" package
- Ordinal logistic regression
  - If the response variable is a set of unordered categories
  - Use lrm() in the "rms" package

Poisson Regression

- Used for Y-variables that are counts of rare occurrences
- In this case the family=poisson and link="log"
- For Poisson variables mean=variance
  - For Changes mean=3.05, variance=5.33
  - Should check whether significant overdispersion

Example Results

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.384296	0.1996	1.925	0.0542 .
MCI	0.005799	0.001437	4.036	5.44e-05 ***
Loc	-0.005256	0.002056	2.557	0.0106 *
Called	0.07015	0.032400	2.165	0.0304 *
Data	-0.09041	0.075082	-1.204	0.2286

Residual deviance: 21.572 on 35 degrees of freedom, AIC: 142.18

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3033	0.1885	1.609	0.108
MCI	0.0058	0.001444	4.018	5.87e-05 ***
Loc	-0.005825	0.002002	-2.910	0.0036 **
Called	0.05138	0.02806	1.831	0.0671 ·

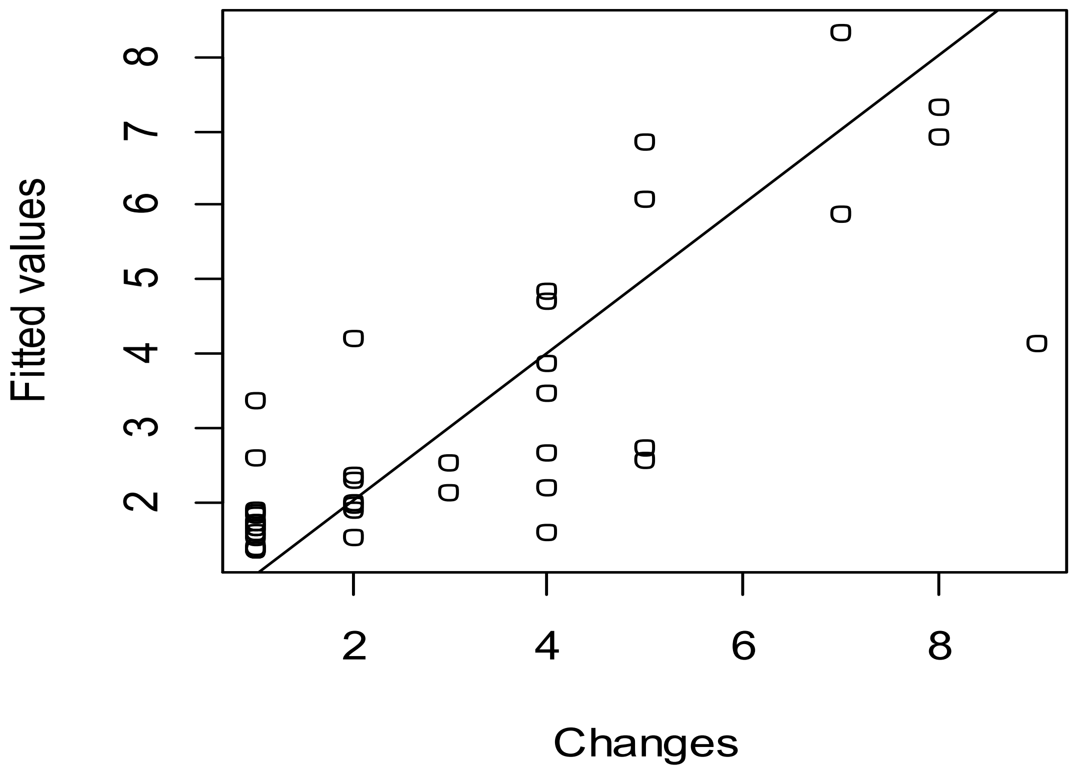
Residual deviance: 23.037 on 36 degrees of freedom, AIC: 141.64

Comparing Models

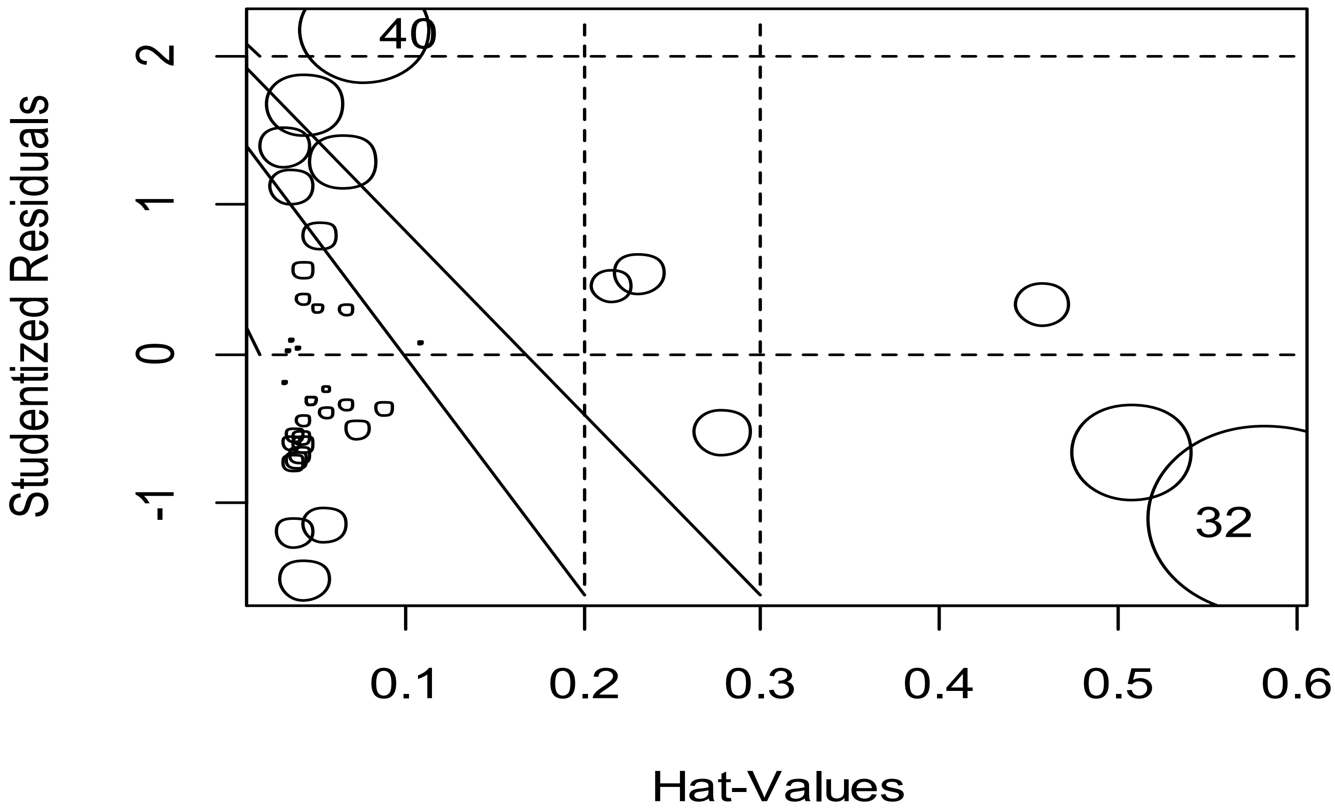
Analysis of Deviance Table
Model 1: Changes ~ MCI + Loc + Called
Model 2: Changes ~ MCI + Loc + Called + Data

Resid.	Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	36	23.037			
2	35	21.572	1	1.4643	0.2263

Changes v. Fitted values



Influence Plot for Poisson Model



GLMs

- R function make GLM easy to use
- No excuse for not using correct model
- Most useful diagnostics still available
  – But more difficult to interpret