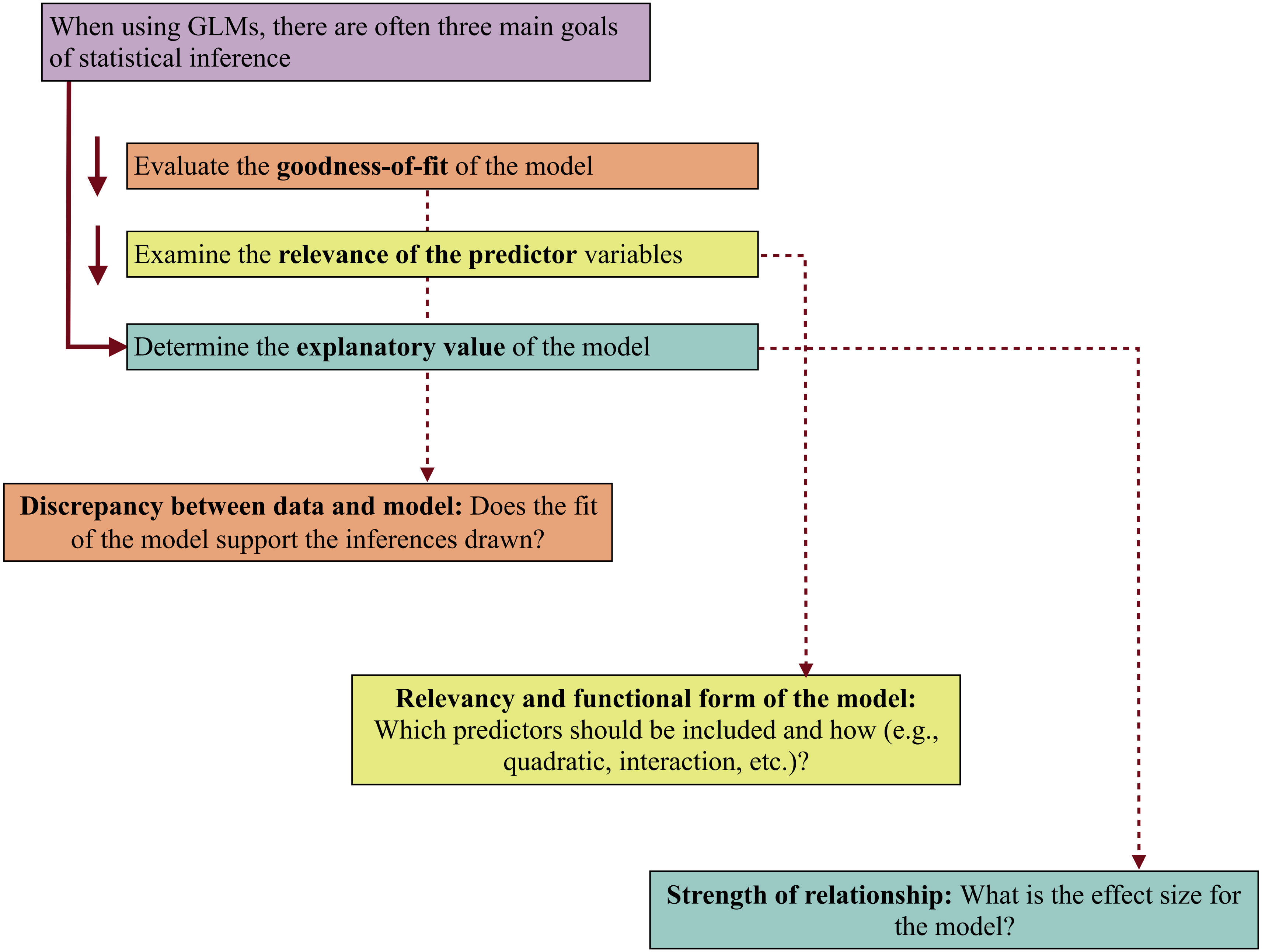
Model Fit for Logistic Regression

Andrew Zieffler
Department of Educational Psychology



When using GLMs, there are often three main goals of statistical inference

Evaluate the **goodness-of-fit** of the model

Examine the **relevance of the predictor** variables

Determine the **explanatory value** of the model

Discrepancy between data and model: Does the fit of the model support the inferences drawn?

Relevancy and functional form of the model: Which predictors should be included and how (e.g., quadratic, interaction, etc.)?

Strength of relationship: What is the effect size for the model?

Chi-Squared Test of Independence

Phi Coefficient: Measure of Association between Two Dichotomous Variables

	Fracture		
Mother fracture	No	Yes	
No	334	41	375
Yes	101	24	125
	435	65	500

$$\hat{\Phi} = \frac{(334 \times 24) - (101 \times 41)}{\sqrt{(375)(125)(435)(65)}}$$

$$\hat{\Phi} = 0.1064388$$

The phi coefficient shows positive association between the two variables.

Is the relationship statistically significant? Or does it differ from 0 within what we would expect because of sampling error?

$$H_0 : \Phi = 0$$

Pearson's χ^2 test of independence is used to examine whether two categorical variables have a relationship (or whether they are unrelated...independent)

Under the assumption of **no relationship** between x and y the theoretical distribution of responses in the cells is based completely on the table margins

	Fracture		
Mother fracture	No	Yes	
No			375
Yes			125
	435	65	500

Based on the table margins (marginal distributions) we can compute the **expected** cell counts

$$\text{Expected} = \frac{\text{Row total} \times \text{Column total}}{\text{Table total}}$$

For example, the expected number of subjects who had a fracture and also had a mother with a hip fracture (Yes, Yes) is

$$\text{Expected} = \frac{125 \times 65}{500}$$

$$\text{Expected} = 16.25$$

Observed and Expected (in parentheses) Cell Counts

	Fracture		
Mother fracture	No	Yes	
No	**334** (326.25)	**41** (48.75)	375
Yes	**101** (108.75)	**24** (16.25)	125
	435	65	500

The difference between the observed and expected cell counts give us an indication of how close the observed distribution matches the theoretical independent distribution

$$\text{Observed} - \text{Expected}$$

For example, in the (Yes, Yes) cell

$$24 - 16.25 = 7.75$$

The observed data show deviation from the independence...at least in this cell

Compute (O – E) for each cell

$$
\begin{aligned}
334 - 326.25 &= 7.75 \\
41 - 48.75 &= -7.75 \\
101 - 108.75 &= -7.75 \\
24 - 16.25 &= 7.75
\end{aligned}
$$

The observed data show deviation from the independence in each cell.

We need a measure of the overall deviation in **all** cells from independence

$$\sum (\mathrm{O} - \mathrm{E}) = 0$$

The sum is always zero because we are computing a residual (for the cell)...and sums of residuals are zero

In order to sum these together, we need to compute the squared residual

$$(\mathrm{O} - \mathrm{E})^2$$

Lastly, these are scaled by dividing the squared residual by the expected cell count

$$\frac{(\mathrm{O} - \mathrm{E})^2}{\mathrm{E}}$$

Compute (O – E)²/E for each cell

$$\frac{(334 - 326.25)^2}{326.25} = 0.184$$

$$\frac{(41 - 48.75)^2}{48.75} = 1.232$$

$$\frac{(101 - 108.75)^2}{108.75} = 0.552$$

$$\frac{(24 - 16.25)^2}{16.25} = 3.696$$

Under H_0, $\hat{\chi}^2$ is distributed as

$$\sim \chi^2_{(r-1)(c-1)}$$

For our example,

$$\sim \chi^2_1$$

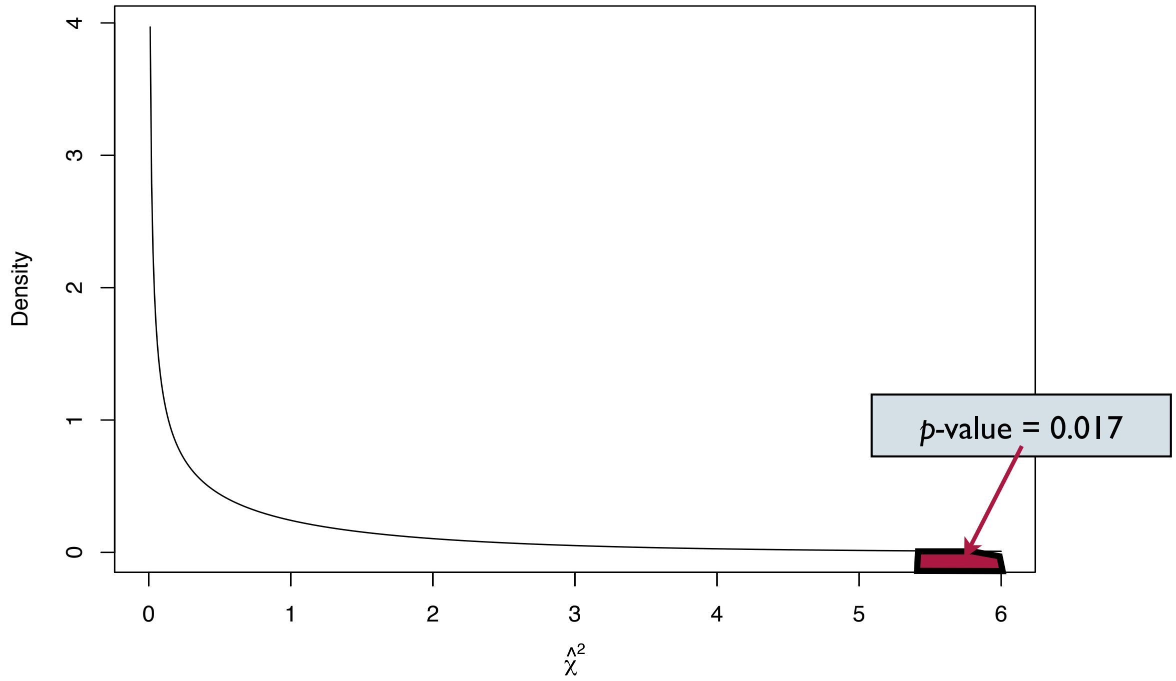
$$\sum \frac{(\mathrm{O} - \mathrm{E})^2}{\mathrm{E}} = 5.664$$

This is the estimated chi-squared value

If the two variables have no relationship (i.e., are independent) their chi-squared value would be 0.

We expect that, even if in the population there is no relationship (chi-squared = 0), that a random sample drawn from that population would have a non-zero chi-squared value because of sampling error.

Is 5.664 within the expected sampling error of 0?



p-value = 0.017

Using R to Carry Out Chi-Squared Test

```
> library(gmodels)
> CrossTable(glow$fracture, glow$momfrac, format = "SPSS", chisq = TRUE)
```

```
   Cell Contents
|-------------------------|
|                   Count |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|
```

```
Total Observations in Table:  500

              | glow$momfrac
glow$fracture |          0  |          1  | Row Total |
--------------|-------------|-------------|-----------|
            0 |       334   |        41   |      375  |
              |     0.184   |     1.232   |           |
              |   89.067%   |   10.933%   |   75.000% |
              |   76.782%   |   63.077%   |           |
              |   66.800%   |    8.200%   |           |
--------------|-------------|-------------|-----------|
            1 |       101   |        24   |      125  |
              |     0.552   |     3.696   |           |
              |   80.800%   |   19.200%   |   25.000% |
              |   23.218%   |   36.923%   |           |
              |   20.200%   |    4.800%   |           |
--------------|-------------|-------------|-----------|
 Column Total |       435   |        65   |      500  |
              |   87.000%   |   13.000%   |           |
--------------|-------------|-------------|-----------|
```

The GLOW data were examined to determine whether there is a relationship between having a fracture and whether or not a subject's mother had a hip fracture. The sample suggested a positive relationship ($\Phi = 0.106$). This relationship was also found to be statistically reliable, $\chi^2(1) = 5.66$, $p = 0.017$.

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  5.664604     d.f. =  1     p =  0.01731063
```

Assumptions of the Chi-Squared Test

- **The data are randomly sampled from a fixed population**
- **The overall sample is sufficiently large *and* each of the expected cell counts are sufficiently large.**
  - This is required to be sure the sampling distribution is approximately distributed as chi-squared
  - All expected cell counts ≥ 1 (i.e., no empty cells)
  - Each of the expected cell counts in a 2x2 table is ≥ 5
  - For non 2x2 tables, 80% of the expected cell counts ≥ 5
- **Independence of the observations**

For small sample sizes, the chi-squared test can produce poor results. Several solutions have been proposed including:

- Yate's continuity correction and
- Fisher's exact test

If the independence assumption is violated, a test that allows for correlated observations should be used (e.g., McNemar's test)

The rule-of-thumb that the minimum expected count ≥ 5 appears to have been an arbitrary choice (probably written by Fisher), Campbell (2007) has provided better advice for expected cell counts based on the research design used to obtain the contingency table.

Campbell, I. (2007). Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine, 26*, 3661–3675.
See also http://www.iancampbell.co.uk/twobytwo/background.htm.

Goodness of Fit

The Chi-squared test can also used to evaluate the fit between data and model.

- Do demographic data collected on a survey (observed data) follow the same distribution as would be expected from the Census (model)?
- Do the residuals (observed data) follow a particular theoretical model (e.g., normal)?

When used in this manner the Chi-squared test is referred to as a **goodness-of-fit** test.

The form of the χ^2 goodness-of-fit test is the same as the χ^2 test of independence, namely

$$\sum \frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$$

where the expected counts are based on the theoretical model.

Note that the χ^2 test of independence is really a goodness-of-fit test. The theoretical model is that of independence, and the expected counts are based on that model.

There are other methods to evaluate goodness-of-fit aside from the χ^2 test. One method, is the likelihood ratio test for goodness-of-fit.

$$G = 2\sum \mathrm{O} \times \ln\left(\frac{\mathrm{O}}{\mathrm{E}}\right)$$

If there is perfect fit between model and data, O = E. G is then equal to 0. Higher values of G indicate mis-fit.

G ~ χ^2 with $df = (k - 1)$ where k is the number of values the variable takes.

For an r x c table, $df = (r - 1)(c - 1)$

$$G = 2 \times \left[334 \times \ln\left(\frac{334}{326.25}\right) + 41 \times \ln\left(\frac{41}{48.75}\right) + 101 \times \ln\left(\frac{101}{108.75}\right) + 24 \times \ln\left(\frac{24}{16.25}\right)\right]$$

$$G = 5.270$$

```
> O = c(334, 41, 101, 24)
> E = c(326.25, 48.75, 108.75, 16.25)

# LRT
> G = 2 * sum(O * log(O/E))
> G
[1] 5.269774

# Compute p-value
> pchisq(G, df = 1, lower = FALSE)
[1] 0.02169884
```

The GLOW data were examined to determine whether there is a relationship between having a fracture and whether or not a subject's mother had a hip fracture. The sample suggested a positive relationship ($\Phi = 0.106$). Using the likelihood ratio test, this relationship was also found to be statistically reliable, $\chi^2(1) = 5.27$, $p = 0.022$.

Model Goodness-of-Fit

The Deviance

In order to to assess a model's goodness-of-fit, we need a measure of the discrepancy between the observed data and the fitted model.

In linear regression the SSE (or $SS_{\text{Residuals}}$) gave a measure of this fit (or discrepancy)

$$\sum (y_i - \hat{y}_i)^2$$

In a likelihood framework, the **deviance** measures how closely the predicted values from the fitted model match the observed values from the raw data.

$$\mathcal{D}(\boldsymbol{\beta}; \mathbf{y}) = N \cdot \ln\left(2\pi\sigma_\epsilon^2\right) + \frac{1}{\sigma_\epsilon^2} \cdot \sum (y_i - \hat{y}_i)^2$$

Note that the deviance (see Unit 2) includes the $SS_{\text{Residuals}}$

In generalized linear models, the deviance is analogous to the SSE in linear models.

In linear regression, to **evaluate a predictor**, a model that includes the predictor is compared to a reduced model that does not include the predictor

To evaluate predictors in generalized linear models, the procedure parallels that for linear models.

Measure to compare the models is the $SS_{\text{Residuals}}$

Measure to compare the models is the deviance

For better fitting models, the $SS_{\text{Residuals}}$ is smaller

For better fitting models, the deviance is smaller

Comparing Models using the Deviance

To compare models, we can examine the **difference in their deviance** measures.

$$\mathcal{D}_{\text{Model Reduced}} - \mathcal{D}_{\text{Model Full}}$$

A reduced model will always have a larger deviance (more misfit) than a fuller model.

$$(-2 \times \ln \mathcal{L}_{\text{R}}) - (-2 \times \ln \mathcal{L}_{\text{F}})$$

$$-2 \times (\ln \mathcal{L}_{\text{R}} - \ln \mathcal{L}_{\text{F}})$$

$$-2 \times \ln \left(\frac{\mathcal{L}_{\text{R}}}{\mathcal{L}_{\text{F}}} \right)$$

We can substitute –2 times the log-likelihood for both models in to this difference

$$\mathcal{D}(\boldsymbol{\beta}; \mathbf{y}) = -2 \times \ln \left[\mathcal{L}(\boldsymbol{\beta}; \mathbf{y}) \right]$$

The difference in deviances is –2 times a **likelihood ratio**. This is again called *G*.

G ~ χ^2 with *df* = difference in the estimated number of parameters between the two models

$$G = -2 \times \ln\left(\frac{\mathcal{L}_\mathrm{R}}{\mathcal{L}_\mathrm{F}}\right)$$

To evaluate a predictor, we compute G based on the likelihood for the model when that predictor is included (full model) and when it is not (reduced model)

Under the null-hypothesis

$$H_0 : \beta_1 = 0$$

$$G \sim \chi^2_{df_R - df_F}$$

$$ln\left(\frac{\pi_1}{1-\pi_i}\right) = \beta_0 + \beta_1(\text{momfrac}) + \epsilon_i$$

full model

$$ln\left(\frac{\pi_1}{1-\pi_i}\right) = \beta_0 + \epsilon_i$$

reduced model

```
# Fit full model
> glm.f = glm(fracture ~ momfrac, data = glow, family = binomial(link = "logit"))

# Fit reduced model
> glm.r = glm(fracture ~ 1, data = glow, family = binomial(link = "logit"))

# Compute analysis of deviance
> anova(glm.r, glm.f, test = "LRT")

Analysis of Deviance Table

Model 1: fracture ~ 1
Model 2: fracture ~ momfrac
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       499     562.34
2       498     557.07  1   5.2698   0.0217 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This is exactly the results we obtained from the LRT test of goodness-of-fit.

The deviance for a model can be computed using the deviance() function.

```
# Compute deviance for the full model
> deviance(glm.f)

[1] 557.0654
```

The deviance is also given in the summary() results

```
> summary(glm.f)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1960     0.1136 -10.532   <2e-16 ***
momfrac       0.6605     0.2810   2.351   0.0187 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 557.07  on 498  degrees of freedom
AIC: 561.07

Number of Fisher Scoring iterations: 4
```

In certain cases, the deviance from a single model can be used to evaluate that model's fit

Saturated Model

A baseline for model fit

Saturated model is a model that includes as many parameters as there are data points.

- *n* = 2; a model with two parameters (e.g., a model with a slope and intercept) would be a saturated model
- *n* = 3; a model with three parameters (e.g., a quadratic model) would be a saturated model

Saturated model will **always have perfect fit** to the observed data (think about *n* = 2 and a line).

- Since the number of model parameters is the same as the sample size, there is **no gain in parsimony** when we use the model as a proxy for the data
- The saturated model also shifts as we obtain new data, so there is **no generalizability** for the model (the saturated model *overfits* the data)

The saturated model in the GLOW example has 500 parameters

In GLMs, the saturated model has the same distribution and link function as the model of interest, but $g(\mu_i) = \psi_i$ for $i = 1, 2, 3, \ldots, N$

In other words, the saturated model allows a different mean response for each subject and thus has N parameters.

Inference for the Logistic Regression Model

Saturated Model

Law of Likelihood

"Within the framework of a statistical model, a particular set of data supports one statistical hypothesis better than another if the likelihood of the first hypothesis on the data exceeds the likelihood of the second hypothesis"

If the fitted model is adequate, then we would expect that

$$\mathcal{L}(\text{Fitted Model}) \sim \mathcal{L}(\text{Saturated Model})$$

and

$$\frac{\mathcal{L}(\text{Fitted Model})}{\mathcal{L}(\text{Saturated Model})} \sim 1$$

Since, the deviance measures how closely the predicted values from the fitted model match the observed values from the raw data, one way to assess the adequacy of a particular fitted model is to compare its deviance with the deviance from the saturated model.

Saturated Model

Recall the likelihood is the joint probability of observing the data that were collected.

$$\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{y}) = 1$$

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = 0$$

$$\mathcal{D}(\boldsymbol{\beta}; \boldsymbol{y}) = 0$$

$$G = \mathcal{D}(\text{Fitted Model}) - \mathcal{D}(\text{Saturated Model})$$

$$G = \mathcal{D}(\text{Fitted Model}) - 0$$

$$G = \mathcal{D}(\text{Fitted Model})$$

The difference in deviance has a known distribution and therefore can be used for hypothesis testing.

If the fitted model describes the data as well (or nearly as well) as the saturated model,

$$H_0 : \mathcal{D}(\text{Fitted Model}) - \mathcal{D}(\text{Saturated Model}) = 0$$

or

$$H_0 : \Gamma = 0$$

then

$$G \sim \chi^2_{N-p}$$

This suggest we expect a *G* or deviance near zero if the fitted model adequately fits the data.

The above deviance test compares the proposed model to the most general (saturated) model, and hence asks the question, can we use a more parsimonious model to describe the data as well as the most general model does?

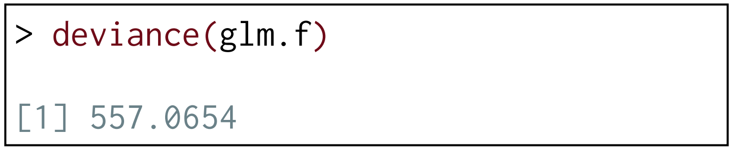
The χ^2 distribution is correct for fixed *df* only when *n* is large.

The *df* for the saturated model is based on the sample size, *n*. And as *n*→∞ the *df* is not fixed (*df*→∞)

For categorical predictors—**grouped data**—we can assume a fixed *df* (since it is based on the number of values the predictor can take on), even as *n*→∞. No matter how large the sample, there are only a finite number of values the predictor can take.

For most GLMS, there is no benchmark distribution we can use to evaluate the deviance and get a *p*-value.

The best we can do is point out that a large deviance suggests misfit and a small deviance suggests better fit.



```
> deviance(glm.f)

[1] 557.0654
```

Although the deviance may suggest misfit, it is hard to evaluate (is this large misfit?) and also does not suggest *why* the misfit occurs.

To evaluate the degree of misfit, we fit the "worst" fitting model to see how bad the deviance can get.

To evaluate why there is misfit, we examine the model residuals.

Evaluating the Degree of Misfit

"Worst" fitting model

The "worst" fitting model is an intercept-only model

```
> deviance(glm.r)

[1] 562.3351
```

The deviance for the model that uses `momfrac` as a predictor improves the deviance from 562.3 to 557.1, a difference of 5.26.

This difference is statistically significant as already shown by the LRT.

```
> summary(glm.f)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -1.1960     0.1136 -10.532   <2e-16 ***
momfrac       0.6605     0.2810   2.351   0.0187 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 557.07  on 498  degrees of freedom
AIC: 561.07

Number of Fisher Scoring iterations: 4
```

The `summary()` function's output also includes the deviance for the "worst" fitting model.

Table 3

Results of fitting a series of logistic regression models predicting fractures for $n = 500$ subjects. Coefficient-level inference includes the Wald statistic and p-value for the one-degree-of-freedom χ^2.

	Model A		Model B	
Predictor	B	SE	B	SE
Mother had a hip fracture (1 = Yes, 0 = No)	0.66	0.28		
	2.35 ($p = .019$)			
Age			0.05	0.01
			4.55 ($p < .001$)	
(Intercept)	–1.20	0.11	–4.78	0.83
	–10.53 ($p < .001$)		–5.78 ($p < .001$)	
Model evaluation				
Deviance	557.07		541.06	

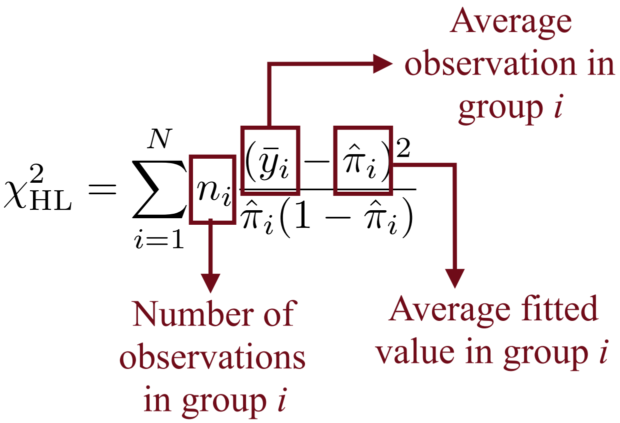
The deviance for Model B is smaller than the deviance for Model A, indicating better fit.

Hosmer–Lemeshow Test

Hosmer and Lemeshow (1980) proposed a goodness-of-fit test for logistic models

- Order the fitted values from smallest to largest
- Form *g* equally sized groups with group 1 composed of the *N/g* observations with the smallest fitted values, group 2 composed of the observations having the next smallest fitted values, etc.
- Compute

$$\chi^2_{\text{HL}} = \sum_{i=1}^{N} n_i \frac{(\bar{y}_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}$$



$$\chi^2_{\text{HL}} = \sum_{i=1}^{N} \boxed{n_i} \frac{(\boxed{\bar{y}_i} - \boxed{\hat{\pi}_i})^2}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

Average observation in group *i*
Number of observations in group *i*
Average fitted value in group *i*

The Hosmer–Lemeshow test statistic has a complicated distribution, but it can asymptotically be approximated by a χ^2-distribution with *g*–2 degrees of freedom.

Large values of the statistic indicate lack of fit, but the test does not indicate why (e.g., model mis-specification).

The test has only moderate power since the grouping is based on a model which is assumed to be true.

The test is highly influenced by the number of groups and Kuss (2002) suggested that the test may be highly unstable.

```
> library(MKmisc)
> HLgof.test(fit = fitted(glm.b), obs = glow$fracture)

$C

     Hosmer-Lemeshow C statistic

data:  fitted(glm.b) and glow$fracture
X-squared = 8.8399, df = 8, p-value = 0.356


$H

     Hosmer-Lemeshow H statistic

data:  fitted(glm.b) and glow$fracture
X-squared = 3.5434, df = 8, p-value = 0.8958
```

The large *p*-value does not indicate lack-of-fit.

```
> HLgof.test(fit = fitted(glm.f), obs = glow$fracture)

Warning messages:
1: In HLgof.test(fit = fitted(glm.f), obs = glow$fracture)
:
  Found only 1 different groups for Hosmer-Lemesho C
statistic.
2: In pchisq(chisq, param) : NaNs produced
```

Problems computing the HL test often indicate that there are not enough unique predictor values. Need to add other predictors to the model.

One better measure of fit is **Akiake's Information Criteria** (AIC).

- Basis in framework known as Kullback–Leibler information theory (Kullback & Leibler, 1951)
- AIC is also estimate of predictive accuracy
- Ability of model to predict new data
- This is often of interest to applied researchers

Penalty guards against improved fit just by adding predictors

$$AIC = \text{deviance} + \boxed{2 \cdot K}$$

where K is the number of estimated parameters (number of fixed effects in model)

- Smaller deviance = better fit (same is true for AIC)
- If worthless predictors are added (i.e., deviance does not improve much), AIC will increase because 2K will go up
- If non-worthless predictors are added, 2K will still go up, but the decrease in deviance will outweigh this

```
> AIC(glm.b)

[1] 545.0613
```

Some details about the AIC to bear in mind

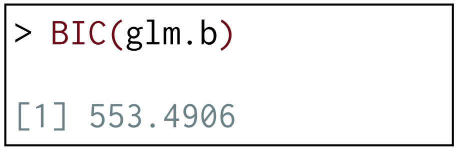
- AIC values are not standardized
- AIC values may be positive or negative
- AIC values may be very large or very small
- AIC does not offer a method of statistical testing (i.e., the term statistically significant should never be used with AIC)
- AIC offers a method of rank ordering the models in a given set of candidate models (nested or not)
- AIC is affected by sample size and should not be used to compare models across studies
- AIC is further changed when he response variable changes so it should not be used to compare models where the response is transformed

Another measure of fit is **Schwatrz's (Bayesian) Information Criteria** (BIC).

$$BIC = \text{deviance} + \ln(N) \cdot K$$

where K is the number of estimated parameters (number of fixed effects in model) and N is the sample size.

Penalty guards against improved fit just by adding predictors and increasing sample size



```
> BIC(glm.b)

[1] 553.4906
```

Interpreting differences in BIC across models

(Raftery, 1995):

- 0–2: Weak evidence
- 2–6: Positive evidence
- 6–10: Strong evidence
- >10: Very strong evidence

Information Criteria: AIC and BIC

Each information criterion "penalizes" the log-likelihood statistic for "excesses" in the structure of the current model

- The AIC penalty accounts for the number of parameters in the model.
- The BIC penalty goes further and also accounts for sample size.

Smaller values of AIC & BIC indicate better fit

Models need not be nested, but datasets must be the same.

Disputing evidence...two candidate models

Careful: Gelman & Rubin (1995) declare these statistics and criteria to be "off-target and only by serendipity manage to hit the target"

Table 3

Results of fitting a series of logistic regression models predicting fractures for $n = 500$ subjects. Coefficient-level inference includes the likelihood ratio statistic and p-value for the one-degree-of-freedom χ^2.

	Model A		Model B	
Predictor	B	SE	B	SE
Mother had a hip fracture (1 = Yes, 0 = No)	0.66	0.28		
	$X^2 = 5.27$			
	($p = 0.022$)			
Age			0.05	0.01
			$X^2 = 21.27$	
			($p < 0.001$)	
(Intercept)	–1.20	0.11	–4.78	0.83
Model evaluation				
Deviance	557.07		541.06	
AIC	561.07		545.06	
BIC	569.49		553.49	
HL			3.54 ($df = 8$)	
			($p = 0.896$)	

Model B has a lower AIC than Model A

Difference in BIC ~16. This would be very strong evidence, according to Rafferty, that Model B fits better than Model A.

Examining Residuals

Assumptions for the Logistic Regression Model

There are very few assumptions for the logistic regression model.

- The model has been correctly specified (all meaningful predictors have been included)
- Errors are independent
- Linearity between the predictors and the logits
- Large sample size for tests and CIs

GLM Residuals

Residuals from a GLM are **not always** defined the same as OLS residuals

$$e_i \xcancel{= Y_i - \hat{Y}_i}$$

R produces *five* different types of residuals for GLMs.

- Deviance residuals (default)
- Pearson residuals
- Working residuals
- Response residuals
- Partial residuals

The deviance can be expressed as the sum of N terms—one for each observation

$$\mathcal{D}(\boldsymbol{\beta};\mathbf{y}) = \sum D_i$$

Deviance residual is then defined as $\sqrt{D_i}$

The sign of the residual is the same as the sign of

$$Y_i - \hat{\mu}_i$$

The **Response residual** is the OLS defined residual

$$e_i = Y_i - \hat{\mu}_i$$

Response residuals ignore the non-constant variance that is innate to the GLM models and are not recommended for diagnostic purposes

Standardizing the response residual according to the variance function V(μ) gives the **Pearson residual**

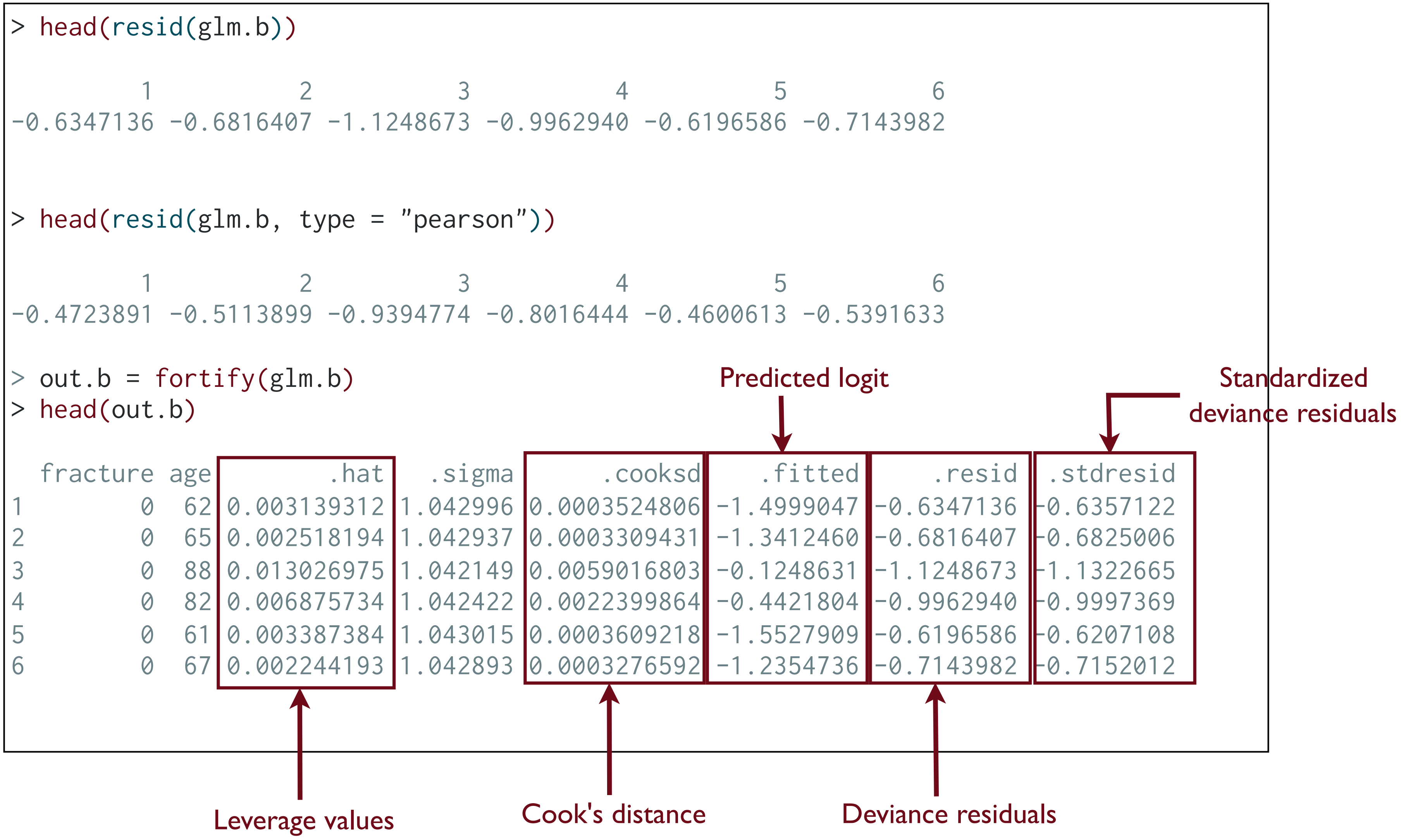
$$e_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Since the Pearson residuals are often skewed, they are not recommended for diagnostic purposes

The **Working residual** is the OLS defined residual based on the last iteration of the IRLS estimation process

The **Partial residual** is a matrix of the working residuals computed from omitting each predictor in the model

Examine Residuals



```
> head(resid(glm.b))

        1         2         3         4         5         6 
-0.6347136 -0.6816407 -1.1248673 -0.9962940 -0.6196586 -0.7143982 


> head(resid(glm.b, type = "pearson"))

        1         2         3         4         5         6 
-0.4723891 -0.5113899 -0.9394774 -0.8016444 -0.4600613 -0.5391633 

> out.b = fortify(glm.b)
> head(out.b)

  fracture age        .hat   .sigma      .cooksd    .fitted     .resid  .stdresid
1        0  62 0.003139312 1.042996 0.0003524806 -1.4999047 -0.6347136 -0.6357122
2        0  65 0.002518194 1.042937 0.0003309431 -1.3412460 -0.6816407 -0.6825006
3        0  88 0.013026975 1.042149 0.0059016803 -0.1248631 -1.1248673 -1.1322665
4        0  82 0.006875734 1.042422 0.0022399864 -0.4421804 -0.9962940 -0.9997369
5        0  61 0.003387384 1.043015 0.0003609218 -1.5527909 -0.6196586 -0.6207108
6        0  67 0.002244193 1.042893 0.0003276592 -1.2354736 -0.7143982 -0.7152012
```

Predicted logit

Standardized deviance residuals

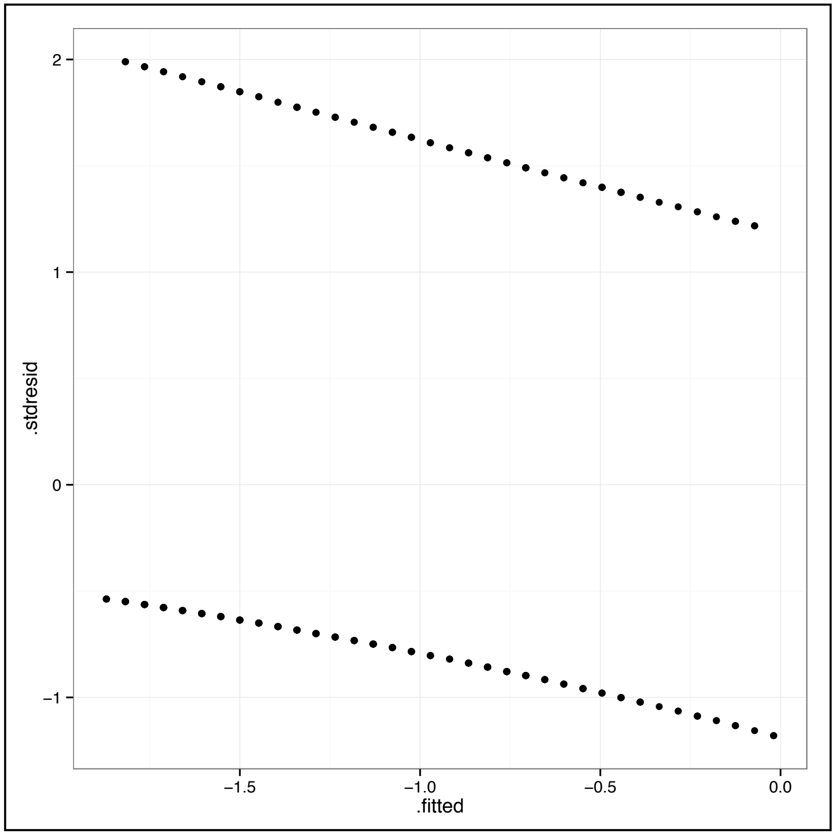
Leverage values

Cook's distance

Deviance residuals

Examine Residuals

```
> ggplot(data = out.b, aes(x = .fitted, y = .stdresid)) +
    geom_point() +
    theme_bw()
```



GLM: `fracture ~ age`

Each value of *X* produce a unique y-hat value. *Y* – *Y-hat* has only two possible values for the residual at each *X*

Gelman and Hill (2011) suggest examining **binned residuals** from a logistic regression. To compute binned residuals,

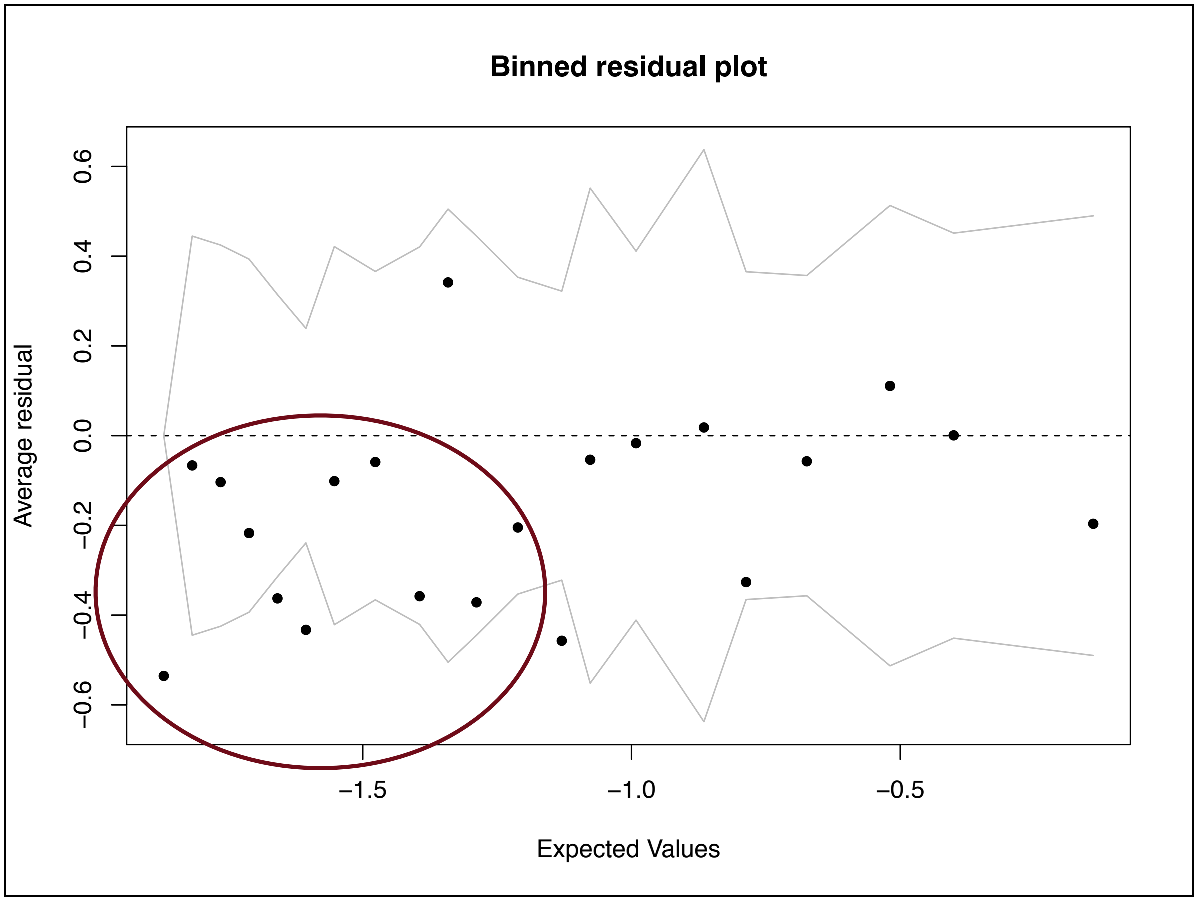
- In the residual plot form *g* bins, each having equal width (but not necessarily with equal number of observations)
- Compute the average residual within each bin
- Compute the average fitted value for each bin
- Plot the average residuals versus the average fitted value

```
> library(arm)
> binnedplot(x = out.b$.fitted, y = out.b$.resid)
```

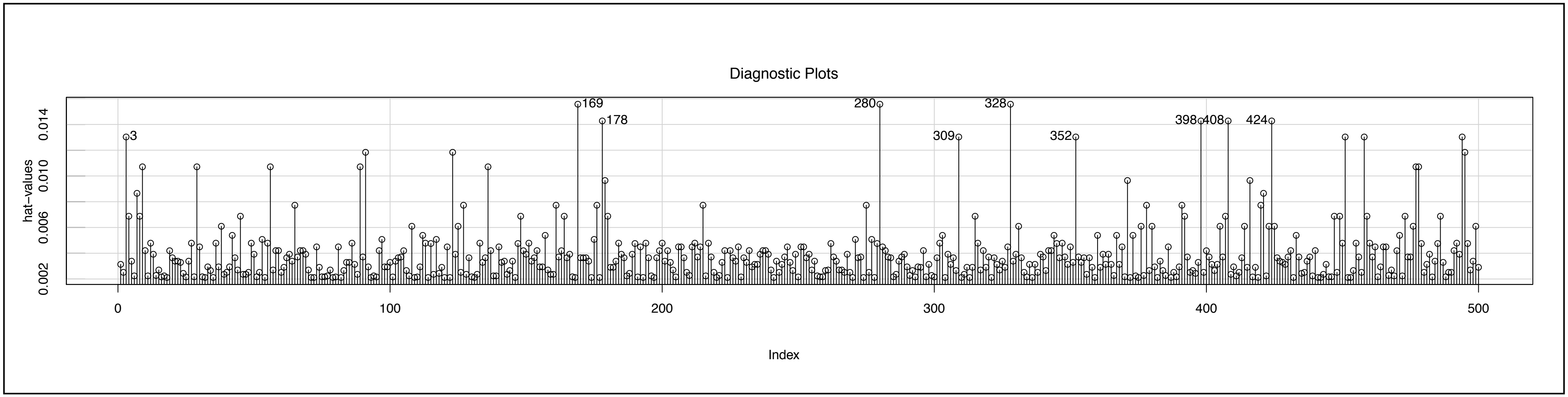
The number of bins is roughly

$$\sqrt{n}$$

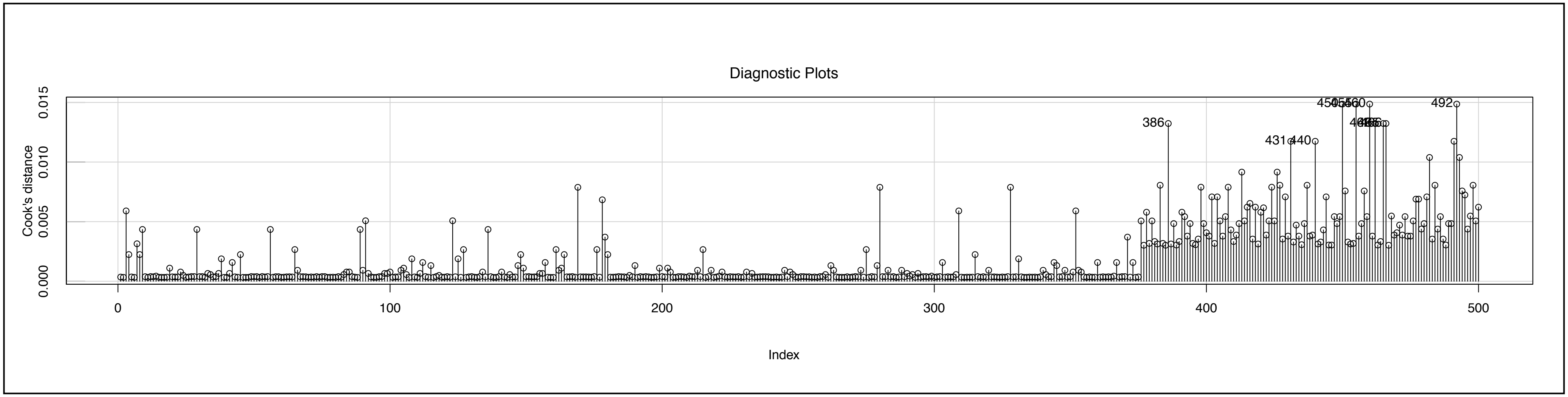
This can be changed by setting the argument `nclass=`



```
> library(car)
> influenceIndexPlot(glm.b, vars = "hat", id.n = 10)
```

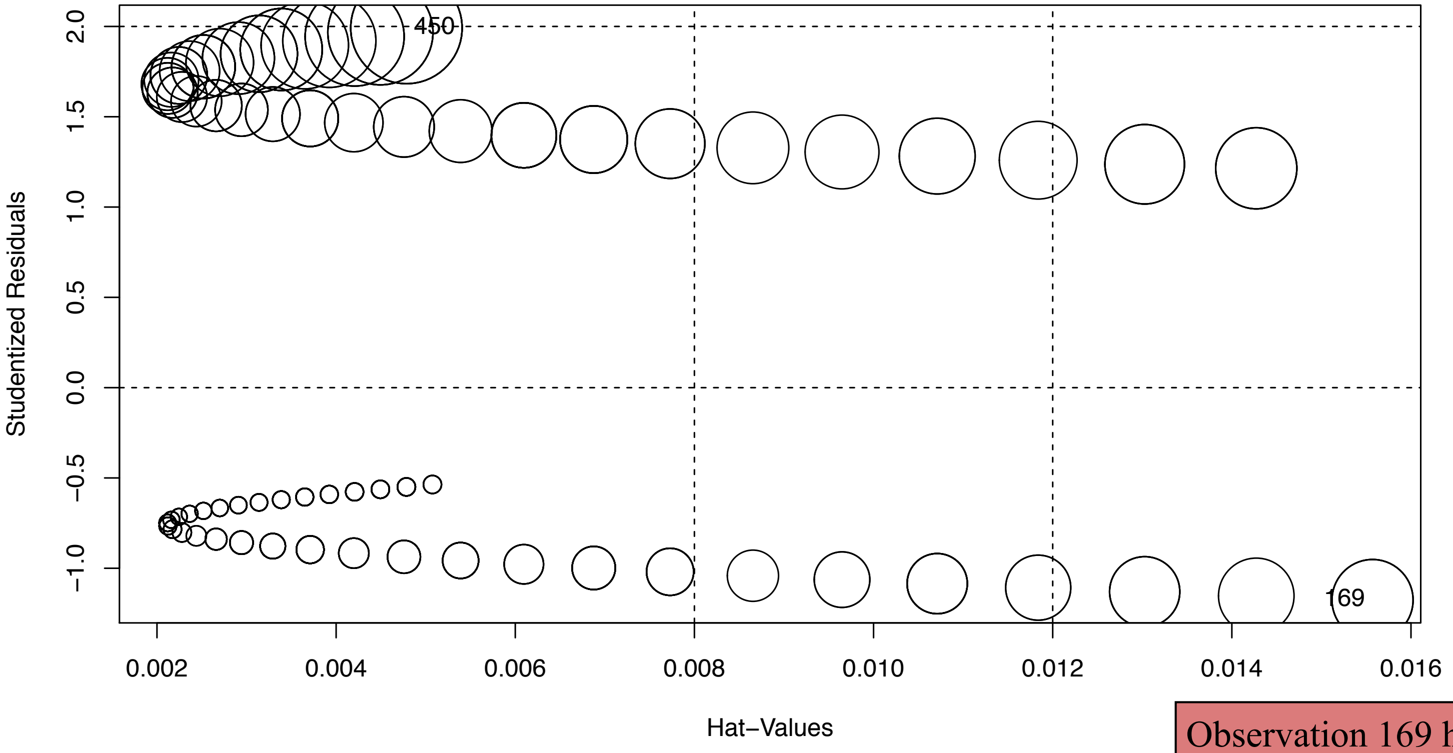


```
> influenceIndexPlot(glm.b, vars = "Cook", id.n = 10)
```



```
> influencePlot(glm.b, vars = "hat", id.n = 10)
```

Observation 460 has a high studentized residual value, and a large Cook's distance.



What happens if we remove these two observations?

Observation 169 has a high leverage value, and a large Cook's distance.

```
> glm.c =  update(glm.b, subset = - c(169, 450))
```

Table 3
Results of fitting a series of logistic regression models predicting fractures for $n = 500$ subjects. Coefficient-level inference includes the Wald statistic and p-value for the one-degree-of-freedom χ^2.

	Model B		Model C	
Predictor	B	SE	B	SE
Age	0.05	0.01	0.06	0.01
	4.55 ($p < .001$)		4.74 ($p < .001$)	
(Intercept)	–4.78	0.83	–4.99	0.84
	–5.78 ($p < .001$)		–5.95 ($p < .001$)	
Model evaluation				
Deviance	541.06		535.69	

There doesn't seem to be much difference in the size of the coefficients (nor their statistical reliability). The fit seems slightly improved.