

Goodness of Fit in Logistic Regression

David M. Roche

April 13, 2021

Goodness of Fit for Logistic Regression

Collection of Binomial Random Variables

Suppose that we have k samples of n 0/1 variables, as with a binomial $\text{Bin}(n, p)$, and suppose that $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_k$ are the sample proportions. We know that

$$E(\hat{p}) = p$$

$$V(\hat{p}) = p(1 - p)/n$$

- If $\bar{p} = \text{Ave}(\hat{p}_i)$ then if the distribution really is binomial, we should have that the sample variance s^2 of the \hat{p}_i should be close to $\bar{p}(1 - \bar{p})/n$. If it is not, then there is something wrong.
- The sample variance can be as small as 0 if all the \hat{p}_i are the same, and is largest if some of the \hat{p}_i are 0 and the remainder are 1.

- For example, suppose that $k = 20$ and $n = 50$, If $p = 0.1$, then $\bar{p} \sim 0.1$ and $s^2 \sim p(1 - p)/n = (0.1)(0.9)/50 = 0.0018$.
- If 5 of the sample proportions are 1 and 45 are 0, then $\bar{p} = 0.1$ but $s^2 = [(5)(0.90)^2 + (45)((0.1)^2)] / 39 = 0.0918$, which is a factor of 50 too big.
- If the variance is too big, then either the distribution is not binomial, or we need more predictors (we have only one in this example, the intercept).

The deviance is

$$D = 2 \sum [y_i \ln(y_i/\hat{\mu}_i) + (n - y_i) \ln((n - y_i)/(n - \hat{\mu}_i))]$$

If we have k groups from a single binomial distribution, then $\hat{\mu}_i = n\bar{p}$. The expression

$$y_i \ln(y_i/n\bar{p}) + (n - y_i) \ln((n - y_i)/(n - n\bar{p}))$$

is like

$$(\hat{p}_i - \bar{p})^2 = (y_i - n\bar{p})^2/n^2$$

in that both get larger as the difference between the observed and expected get larger.

Residual Deviance

- Suppose we have k groups and n observations. The (residual) deviance of a model is the difference between the minus twice the log likelihood of that model and that of the saturated model that fits each group with its own proportion.
- So we could consider the deviance of the given model as a likelihood ratio test of whether the given model is satisfactory; that is, whether it can be shown that adding more variables helps the predictions.

- If our model has q predictors (counting categorical variables as one less than the number of levels and an intercept), then the difference from the saturated model is $k - q - 1$, and we could compare the deviance to a χ^2_{k-q-1} which has mean $k - q - 1$.
- If the deviance is too big, then something is wrong: Omitted predictors? Not binomial?

```
> summary(hyp.glm)
glm(formula = hyp.tbl ~ smoking + obesity + snoring, family = binomial)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.37766	0.38018	-6.254	4e-10	***
smokingYes	-0.06777	0.27812	-0.244	0.8075	
obesityYes	0.69531	0.28509	2.439	0.0147	*
snoringYes	0.87194	0.39757	2.193	0.0283	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 14.1259 on 7 degrees of freedom
Residual deviance: 1.6184 on 4 degrees of freedom

- Residual deviance: 1.6184 on 4 degrees of freedom
- The residual deviance is not too large, so we don't appear to have a problem.
- $\Pr(\chi_4^2 < 1.6184) = 0.20$ so it is not too small either.

Deviance for Grouped Data

- When data are entered as groups with disease/notdisease, then R uses the definition of deviance comparing it to a model saturated by groups.
- In the hypertension data, there are 8 groups and deviance is relative to an 8df model like Smoking*Obesity*Snoring.

Deviance for Ungrouped Data

- If the data are given in observation form with 0/1 response, then R uses a definition of deviance relative to an observation-saturated model where each response is perfectly predicted.
- This means that the deviance is just minus twice the log likelihood.
- We can still use the deviance test when the analysis is grouped.

```
> main.model <- glm(CHD~CAT+SMK+HPT,family=binomial,evans)
> full.model <- glm(CHD~CAT*SMK*HPT,family=binomial,evans)
> anova(main.model,full.model,test="Chisq")
```

Analysis of Deviance Table

Model 1: CHD ~ CAT + SMK + HPT

Model 2: CHD ~ CAT * SMK * HPT

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	605	414.05			
2	601	404.92	4	9.1367	0.05777 .

```
> summary(main.model)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.0324	0.3056	-9.924	< 2e-16 ***
CAT	0.8055	0.2963	2.719	0.00655 **
SMK	0.7098	0.2969	2.391	0.01681 *
HPT	0.5956	0.2844	2.094	0.03623 *

Null deviance: 438.56 on 608 degrees of freedom
Residual deviance: 414.05 on 605 degrees of freedom

- This is a test of whether we should add all of the interactions. The result is not significant, as a test of goodness of fit.
- But (see below) there can still be additional predictors that are important, in this case both by significance test and AIC.

```
> add1(main.model, full.model, test="Chisq")
```

```
Single term additions
```

```
Model:
```

```
CHD ~ CAT + SMK + HPT
```

	Df	Deviance	AIC	LRT	Pr(>Chi)
<none>		414.05	422.05		
CAT:SMK	1	413.65	423.65	0.4018	0.526167
CAT:HPT	1	406.74	416.74	7.3149	0.006838 **
SMK:HPT	1	414.04	424.04	0.0140	0.905891

Goodness of Fit for Uncategorized Data

- The procedure above works only if the number of groups in which the predictors are the same is small compared to n .
- A commonly used procedure if there are continuous predictors is the Hosmer-Lemeshow goodness of fit test.
- This works poorly if there are too many ties, has low statistical power, but may be useful when almost all the observations have distinct predictors.

- Order the data by the predicted values and cut into classes of equal size, say 10.
- Calculate observed and expected cases in each group.
- Use χ^2 test as usual from $(O - E)^2/E$.
- This can be done using `hoslem.test()` from the `ResourceSelection` package in R.
- This is very commonly used, but has low power, and interpretation in case of rejection can be difficult.

```
> library(ResourceSelection)
ResourceSelection 0.2-6    2016-02-15
Warning message:
package 'ResourceSelection' was built under R version 3.2.5
> mod2.glm <- glm(CHD~CAT+CHL+SMK+HPT,family=binomial,evans)
> hoslem.test(mod2.glm$y,fitted(mod2.glm))
```

Hosmer and Lemeshow goodness of fit (GOF) test

```
data:  mod2.glm$y, fitted(mod2.glm)
X-squared = 1.4748, df = 8, p-value = 0.9931
```

Note that the model omits interactions we know are important, but still passes the HL test.

Model Checking and Diagnostics

Linear Regression

- In linear regression, the major assumptions in order of importance:
- **Linearity:** The mean of y is a linear (in the coefficients) function of the predictors.
- **Independence:** Different observations are statistically independent.
- **Constant Variance:** The residual variance is the same for each observation.
- **Normality:** The error distribution is normal.

Diagnostics

Linear Regression

- Plot residuals vs. fitted values
- Plot residuals vs. predictors
- Look for influential observations with $dffits$ and $dfbeta$. These are observations that have a large effect on the coefficients.
- We can use many of these techniques in logistic regression.

Model Checking and Diagnostics

Logistic Regression

- In logistic regression, the major assumptions in order of importance:
- **Linearity:** The logit of the mean of y is a linear (in the coefficients) function of the predictors.
- **Independence:** Different observations are statistically independent.
- **Variance Function:** The variance of an observation with mean p is $p(1 - p)/n$.
- **Binomial:** The error distribution is binomial.

Diagnostics for Grouped Logistic Regression

- Deviance test for goodness of fit.
- Plot deviance residuals vs. fitted values. In this case, there are as many residuals and fitted values as there are distinct categories.
- Plot dfffits vs. fitted values. This is the scaled change in the predicted value of point i when point i itself is removed from the fit. This has to be the whole category in this case.
- All this works well automatically only when the data are given to R in aggregated form.

```
> summary(main.model)
```

```
Call:
```

```
glm(formula = CHD ~ CAT + SMK + HPT, family = binomial, data = evans)
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-0.8185	-0.5721	-0.4325	-0.3068	2.4817

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.0324	0.3056	-9.924	< 2e-16	***
CAT	0.8055	0.2963	2.719	0.00655	**
SMK	0.7098	0.2969	2.391	0.01681	*
HPT	0.5956	0.2844	2.094	0.03623	*

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 438.56  on 608  degrees of freedom  
Residual deviance: 414.05  on 605  degrees of freedom  
AIC: 422.05
```

```
Number of Fisher Scoring iterations: 5
```

```

> evans.cat1 <- aggregate(cbind(CHD,1-CHD,1)~CAT+SMK+HPT,FUN=sum,data=evans)

> print(evans.cat1)
  CAT SMK HPT CHD  V2  V3
1   0   0   0   5 117 122
2   1   0   0   1   5   6
3   0   1   0  15 193 208
4   1   1   0   7  11  18
5   0   0   1   4  51  55
6   1   0   1   7  32  39
7   0   1   1  20  82 102
8   1   1   1  12  47  59

> res <- as.matrix(evans.cat1)[,4:5]
> evans.cat1.glm <- glm(res~CAT+SMK+HPT,family=binomial,data=evans.cat1)

```

```
> summary(evans.cat1.glm)
```

```
Call:
```

```
glm(formula = res ~ CAT + SMK + HPT, family = binomial, data = evans.cat1)
```

```
Deviance Residuals:
```

1	2	3	4	5	6	7	8
-0.2685	0.5256	-0.8950	2.0789	-0.2128	0.2638	1.2263	-1.4307

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.0324	0.3056	-9.924	< 2e-16 ***
CAT	0.8055	0.2963	2.719	0.00655 **
SMK	0.7098	0.2969	2.391	0.01681 *
HPT	0.5956	0.2844	2.094	0.03623 *

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 33.6416 on 7 degrees of freedom  
Residual deviance: 9.1367 on 4 degrees of freedom  
AIC: 45.737
```

```
Number of Fisher Scoring iterations: 4
```

```
> summary(main.model)
```

Call:

```
glm(formula = CHD ~ CAT + SMK + HPT, family = binomial, data = evans)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.8185	-0.5721	-0.4325	-0.3068	2.4817

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.0324	0.3056	-9.924	< 2e-16	***
CAT	0.8055	0.2963	2.719	0.00655	**
SMK	0.7098	0.2969	2.391	0.01681	*
HPT	0.5956	0.2844	2.094	0.03623	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 438.56 on 608 degrees of freedom
Residual deviance: 414.05 on 605 degrees of freedom
AIC: 422.05

Number of Fisher Scoring iterations: 5

The goodness of fit test is to compare 9.1367, the residual deviance, with a χ_4^2 .

```
> pchisq(deviance(evans.cat1.glm),4,lower=F)
[1] 0.05777162
```

We know that the CAT:HPT interaction is significant, which is somewhat indicated by the relatively high value of the residual deviance.

```
> summary(glm(res~CAT+SMK+HPT+CAT:HPT,family=binomial,data=evans.cat1))
```

Call:

```
glm(formula = res ~ CAT + SMK + HPT + CAT:HPT, family = binomial,  
     data = evans.cat1)
```

Deviance Residuals:

1	2	3	4	5	6	7	8
0.10972	-0.38331	-0.06311	0.18549	-0.74483	0.78093	0.40560	-0.54343

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.2032	0.3227	-9.925	< 2e-16 ***
CAT	1.9958	0.4941	4.039	5.37e-05 ***
SMK	0.6655	0.2981	2.232	0.02560 *
HPT	1.0246	0.3213	3.189	0.00143 **
CAT:HPT	-1.6750	0.6007	-2.789	0.00529 **

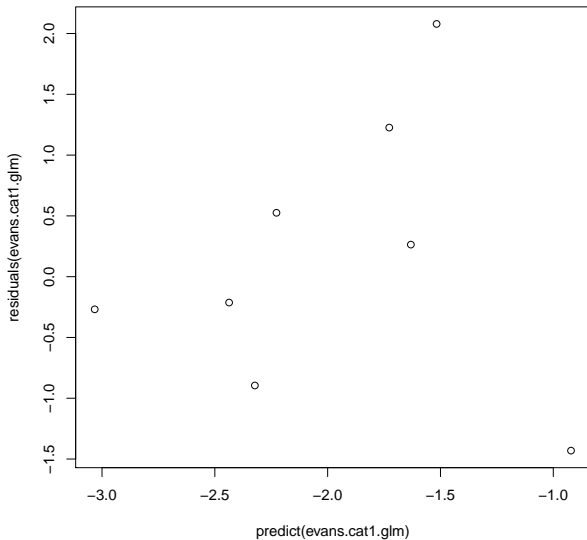
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

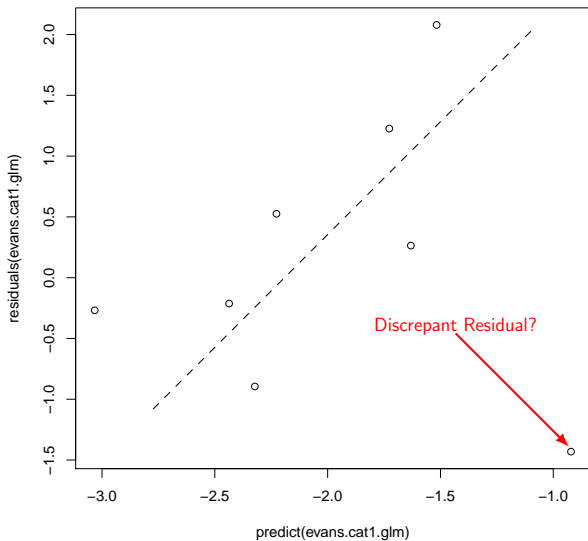
(Dispersion parameter for binomial family taken to be 1)

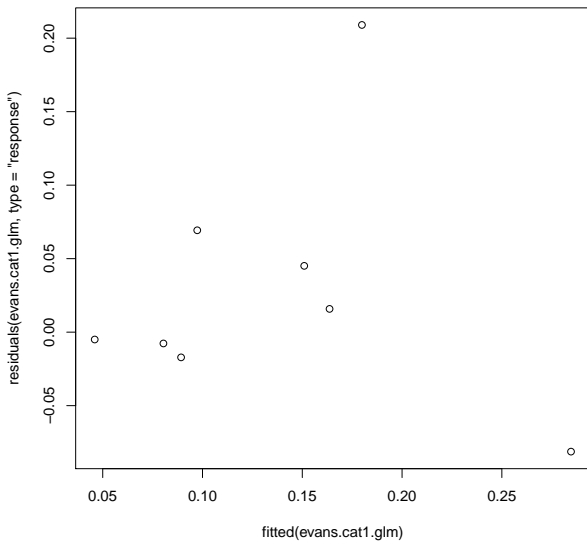
Null deviance: 33.6416 on 7 degrees of freedom
Residual deviance: 1.8218 on 3 degrees of freedom
AIC: 40.422

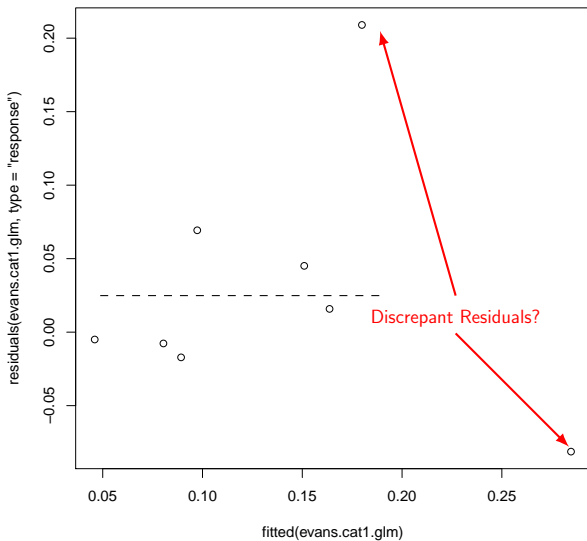
```
> pdf("evanscat1res1.pdf")
> plot(predict(evans.cat1.glm),residuals(evans.cat1.glm))
> dev.off()
windows
  2
> pdf("evanscat1res2.pdf")
> plot(fitted(evans.cat1.glm),residuals(evans.cat1.glm,type="response"))
> dev.off()
```

The first is on the scale of the linear predictor, the second on the $[0, 1]$ scale. Note that the last point $(1, 1, 1)$ has a discordant residual.



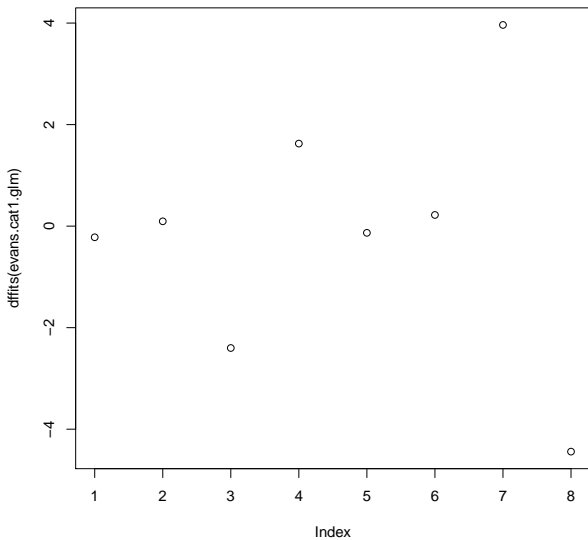


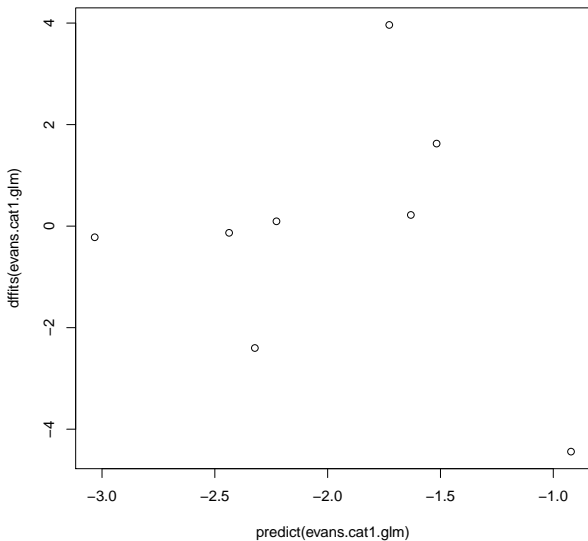




```
> pdf("evanscat1dff1.pdf")
> plot(dffits(evans.cat1.glm))
> dev.off()

> pdf("evanscat1dff2.pdf")
> plot(predict(evans.cat1.glm),dffits(evans.cat1.glm))
> dev.off()
```



Types of Residuals in Logistic Regression

- In linear regression, the residual is always $y - \hat{y}$.
- In logistic regression we have multiple types, partly because we have multiple scales.
- The deviance is the sum of $y_i \ln(y_i/\hat{\mu}_i) + (n - y_i) \ln((n - y_i)/(n - \hat{\mu}_i))$, which is always positive and lives on the χ^2 scale.
- The deviance residual is the signed square root of the deviance contribution, positive if $y > \hat{y}$ and negative otherwise.
- When $y = 1$, all the residuals are positive and when $y = 0$ they are all negative.

Types of Residuals in Logistic Regression

- Pearson and response residuals are on the response scale

$$r = \frac{p - \hat{p}}{\sqrt{\hat{p}(1 - \hat{p})/n}}$$

- This is approximately standard normal if n is large.
- If the data are not grouped, then

$$r_{\text{response}} = y - \hat{y} \qquad r_{\text{Pearson}} = \frac{y - \hat{y}}{\sqrt{\hat{y}(1 - \hat{y})}}$$

Types of Residuals in Logistic Regression

- The **partial residual** is useful for assessing the linearity of the relationship between a quantitative variable and the response.
- The partial residual for observation i and predictor j is

$$r_{ij} = \hat{\beta}_j x_{ij} + \frac{y_i - \hat{y}_i}{\hat{y}_i(1 - \hat{y}_i)}$$

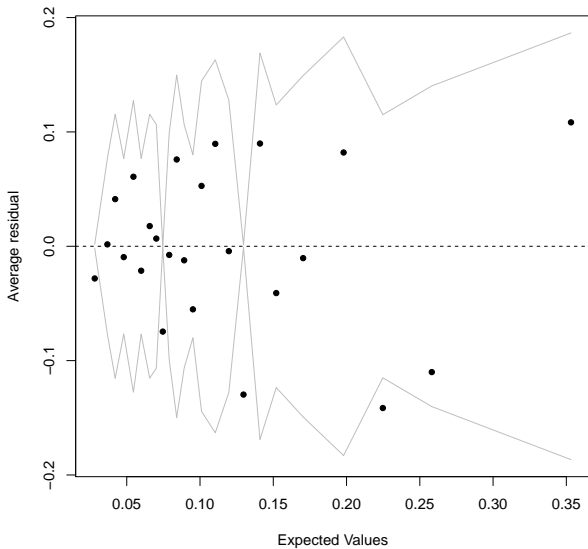
- The second term on the RHS is called the **working residual** and is related to the algorithm that minimizes the deviance.

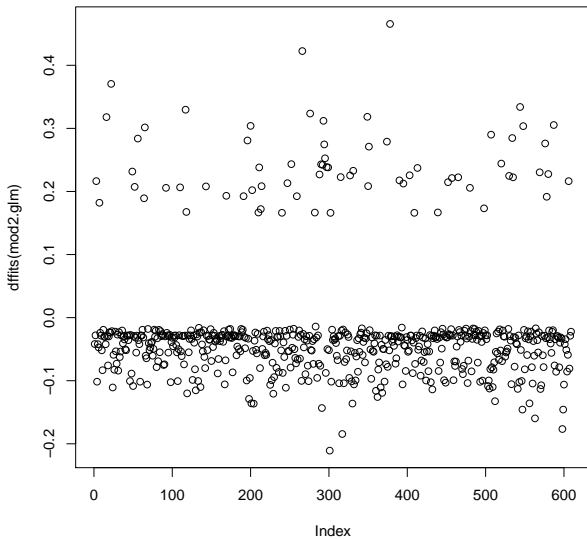
Diagnostics for Ungrouped Logistic Regression

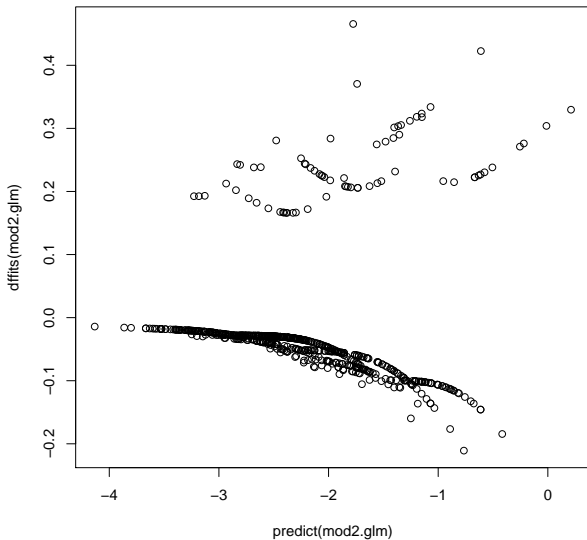
- Possible HL test for goodness of fit
- Plot deviance residuals vs. fitted values. We can either group the fitted values as in the HL test using the, `binnedplot` function in the `arm` package or smooth the plot with `lowess`.
- Plot partial residuals for each quantitative variable vs. the value of the variable.
- Plot `dffits` vs. fitted values.
- Plot `dfbetas` vs. index and/or fitted value for each quantitative variable. This is the change in the coefficient of variable j when point i is removed.

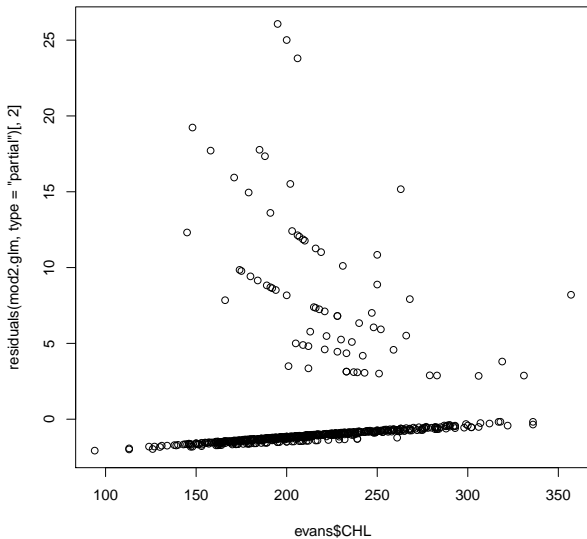
```
> mod2.glm <- glm(CHD~CAT+CHL+SMK+HPT,family=binomial,evans)
> binnedplot(fitted(mod2.glm),residuals(mod2.glm,type="response"))
> plot(dffits(mod2.glm))
> plot(predict(mod2.glm),dffits(mod2.glm))
> which(dffits(mod2.glm) > .3)
 16  22  65 117 200 266 276 293 349 378 544 548 587
 16  22  65 117 200 266 276 293 349 378 544 548 587
> plot(evans$CHL,residuals(mod2.glm,type="partial")[,2])
> plot(dfbeta(mod2.glm)[,1])
> plot(dfbeta(mod2.glm)[,2])
> plot(dfbeta(mod2.glm)[,3])
> plot(dfbeta(mod2.glm)[,4])
> plot(dfbeta(mod2.glm)[,5])
```

Binned residual plot

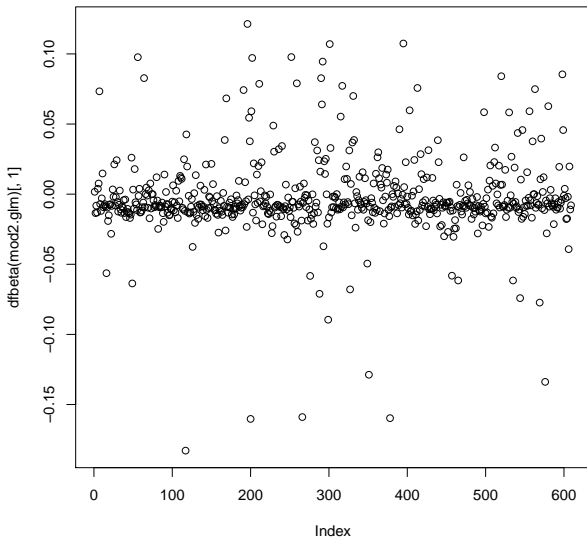








- Curvature in the partial residual plot for CHL may indicate non-linearity.
- This is supported by the curvature in the dffits plot vs. predicted values.

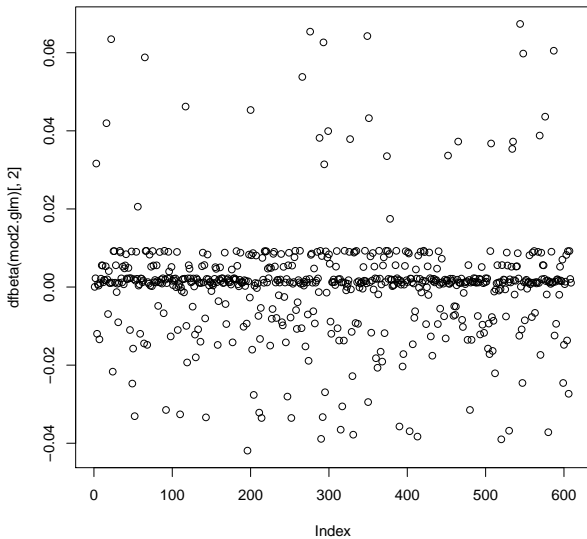


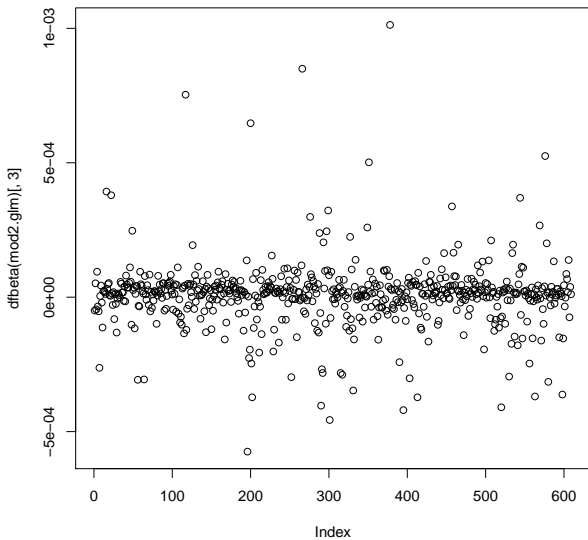
- There are 6 points with high influence for the intercept.
- Omission will increase the intercept.
- Most have high CAT, most are hypertensive, all have CHD.
- It would seem that omission of a CHD case would tend to decrease the intercept, but it increases instead.

```

> evans[order(dfbeta(mod2.glm)[,1])[1:6],]
      ID CHD CAT AGE CHL SMK ECG DBP SBP HPT CH  CC
117 2891  1  1  56 331  1  0 110 190  1  1 331
200 5131  1  1  52 306  1  0 108 178  1  1 306
378 12051 1  0  67 357  0  0  90 129  0  0  0
266 7051  1  1  67 319  0  0 104 182  1  1 319
576 18131 1  1  56 283  1  0 100 188  1  1 283
351 11361  1  1  76 279  1  0  96 136  1  1 279

```





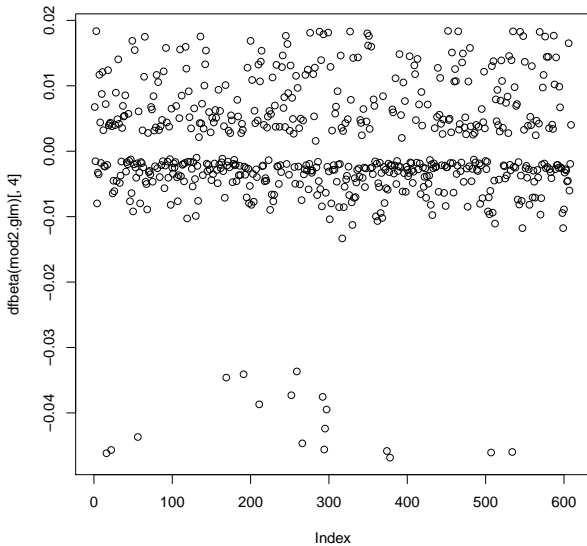
- There are 4 points with high influence for CHL.
- Omission will decrease the coefficient.
- All have CHD and very high CHL.

```
> head(sort(evans$CHL,decreasing=T))
```

```
[1] 357 336 336 331 322 319
```

```
> evans[order(dfbeta(mod2.glm)[,3],decreasing=T)[1:4],]
```

	ID	CHD	CAT	AGE	CHL	SMK	ECG	DBP	SBP	HPT	CH	CC
378	12051	1	0	67	357	0	0	90	129	0	0	0
266	7051	1	1	67	319	0	0	104	182	1	1	319
117	2891	1	1	56	331	1	0	110	190	1	1	331
200	5131	1	1	52	306	1	0	108	178	1	1	306

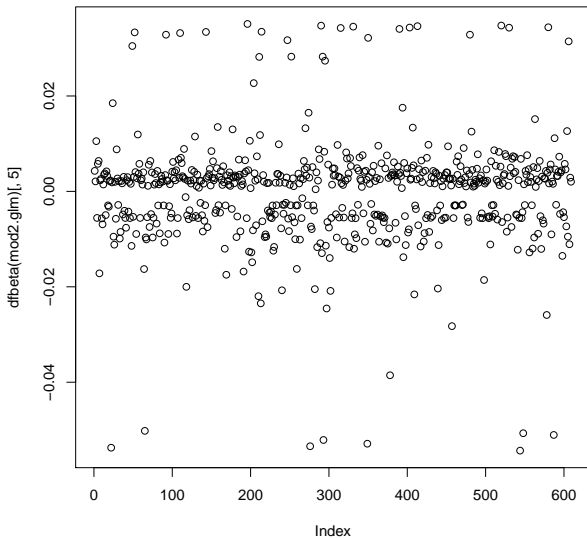


- There are 17 points with high influence for SMK.
- Omission will increase the coefficient.
- All have CHD and don't smoke. In fact, these points consist of all the subjects with CHD who don't smoke. Omission of even 1 has a high effect on the coefficient.

```

> evans[order(dfbeta(mod2.glm)[,4])[1:17],]
      ID CHD CAT AGE CHL SMK ECG DBP SBP HPT CH  CC
378 12051  1  0  67 357  0  0  90 129  0  0  0
16   283  1  1  51 259  0  1 102 135  1  1 259
507 15511  1  1  67 236  0  1 106 200  1  1 236
534 16481  1  1  69 230  0  1 100 170  1  1 230
374 11941  1  1  65 222  0  1  88 162  1  1 222
22   381  1  1  64 247  0  1  75 130  0  0 247
294 9201  1  1  63 213  0  1 156 256  1  1 213
266 7051  1  1  67 319  0  0 104 182  1  1 319
56   1061  1  1  46 166  0  1  76 162  1  1 166
295 9261  1  0  67 250  0  0 100 158  1  0  0
297 9601  1  0  45 263  0  0  86 132  0  0  0
211 5451  1  0  63 202  0  0 110 160  1  0  0
292 9101  1  0  67 188  0  1 102 168  1  0  0
252 6821  1  0  65 185  0  0 105 156  1  0  0
169 4551  1  0  54 206  0  1  76 142  0  0  0
191 4961  1  0  72 200  0  1  86 138  0  0  0
259 6931  1  0  56 195  0  1  94 150  0  0  0

```



- There are 8 points with high influence for the coefficient of hypertension.
- Omission will increase the coefficient.
- Only 71 cases of CHD out of 609, and only 28 are not hypertensive.

	Not Hypertensive	Hypertensive
No CHD	326	212
CHD	28	43


```

evans[order(dfbeta(mod2.glm)[,5])[1:8],]
      ID CHD CAT AGE CHL SMK ECG DBP SBP HPT CH  CC
544 16711  1  1  68 242  1  0  84 128  0  0 242
22   381  1  1  64 247  0  1  75 130  0  0 247
276  8721  1  1  64 233  1  0  94 140  0  0 233
349 11341  1  1  56 228  1  0  92 152  0  0 228
293  9191  1  1  56 221  1  1  78 154  0  0 221
587 18491  1  1  74 212  1  1  70 144  0  0 212
548 16871  1  1  58 209  1  1  94 140  0  0 209
65   1201  1  1  66 205  1  0  80 150  0  0 205

```

- All have CHD, all have high CAT, none are hypertensive, almost all smoke.
- Blood pressure is high “normal”.
- One would expect that omission of a CHD case without hypertension would decrease the coefficient, but this is affected by correlation of the predictors.

The Role of Diagnostics

- Diagnostics can be useful for identifying problems in a model or in the data.
- The Evans County data are already cleaned, but if there were erroneous observations, residual and leverage plots could identify them.

Overdispersion

- A common problem with logistic regression is overdispersion.
- This is when $V(\hat{p}) \gg p(1 - p)/n$
- This can happen if the true parameter p varies even when the covariates do not.
- We can/should then use the quasibinomial, in which $V(\hat{p}) = \theta p(1 - p)/n$

```
> summary(glm(CHD~CAT+CHL+SMK+HPT,family=quasibinomial,evans))
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-1.0066	-0.5276	-0.4102	-0.3108	2.5560

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-4.975282	0.797487	-6.239	8.3e-10	***
CAT	1.021916	0.313496	3.260	0.00118	**
CHL	0.008963	0.003289	2.725	0.00662	**
SMK	0.714577	0.301457	2.370	0.01808	*
HPT	0.483481	0.290735	1.663	0.09684	.

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for quasibinomial family taken to be 1.021549)
```

```
Null deviance: 438.56  on 608  degrees of freedom  
Residual deviance: 406.52  on 604  degrees of freedom  
AIC: NA
```

- In this case, there is no sign of overdispersion.
- Note that this can depend on the model as well as the data.
- Fitting the quasibinomial model is the best test of this.
- You should always check for overdispersion in a binomial (or Poisson) model.
- If there is overdispersion and you use a standard logistic regression, the inferences are wrong.

Homework: Due 4/22/2021

Try some of these diagnostic techniques on your model for the Evans County Data.