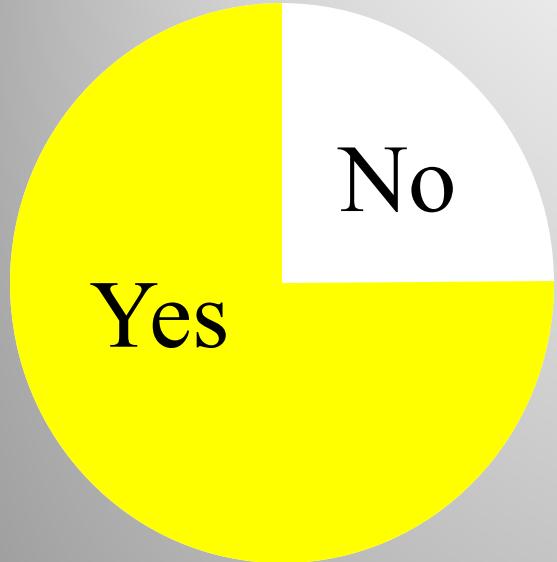


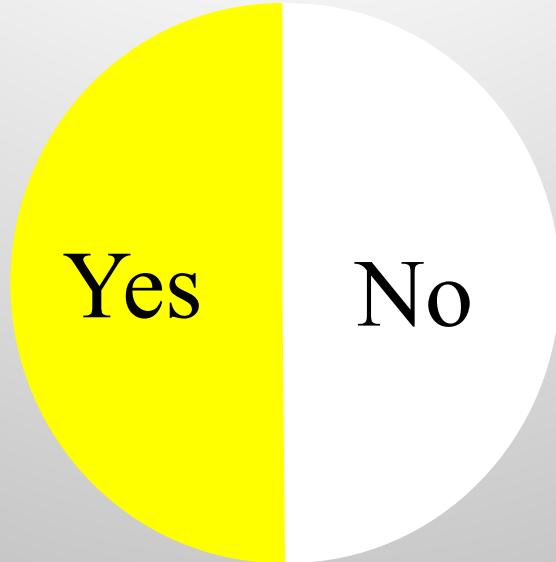
Likelihoods and Logistic Regression Diagnostics

SDS 291
April 22, 2020



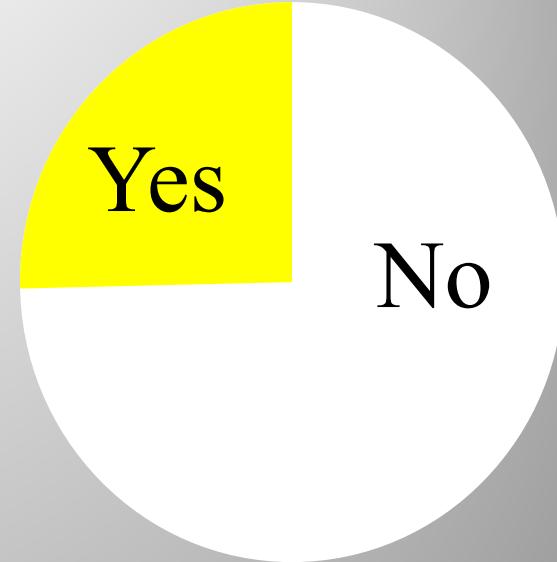
No

Yes



Yes

No



Yes

No

A

B

C

Estimating Parameters in Logistic Regression

Parameters are chosen to *maximize* the *likelihood* of the observed sample (maximum likelihood estimation).

If the i^{th} point is YES ($y_i = 1$), calculate $\hat{\pi}_i$.

If the i^{th} point is NO ($y_i = 0$), calculate $1 - \hat{\pi}_i$.

Likelihood:
$$L = \prod \hat{\pi}_i^{y_i} (1 - \hat{\pi}_i)^{1-y_i}$$

Here the estimated probabilities, $\hat{\pi}$, come from a model. What model is best?

(See Exercise 9.9 and likelihood-play.xls)

A Simple(?) Max Likelihood Problem

If the i^{th} point is YES ($y_i = 1$), calculate $\hat{\pi}$.

If the i^{th} point is NO ($y_i = 0$), calculate $1 - \hat{\pi}$.

Likelihood:
$$L = \prod \hat{\pi}^{y_i} (1 - \hat{\pi})^{1-y_i}$$

Suppose $n = 20$ and there are 14 total successes.

$$L = \hat{\pi}^{14} (1 - \hat{\pi})^6$$

Calculus shows that the best estimate is

$$\hat{\pi} = \frac{14}{20}$$

Evaluating Overall Effectiveness

1. Likelihood of the sample

Output below gives *Log-likelihood*

- * Always negative
- * Smaller -2Log-likelihood is better

Automated

```
'Log Lik.' -67.81383 (df=2)
```

logLik(logitmod)

Manual

```
[1] -67.81383
```

```
pi<- fitted.values(logitmod)
likelihood <- ifelse(Pulse$Gender == 1, pi, 1 - pi)
log(prod(likelihood))
```

Example: Predict Gender via Weight

```
call:  
glm(formula = Gender ~ Wgt, family = binomial, data = Pulse)  
  
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-3.04279 -0.39608 -0.01075  0.53321  2.27350  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) 16.67791   2.20912   7.550 4.37e-14 ***  
Wgt         -0.10962   0.01458  -7.519 5.52e-14 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 321.00 on 231 degrees of freedom  
Residual deviance: 155.15 on 230 degrees of freedom  
AIC: 159.15  
  
Number of Fisher scoring iterations: 6  
  
'log Lik.' -77.57435 (df=2)
```

Hgt

'log Lik.' -67.81383 (df=2)

Evaluating Overall Fit

2. Test for overall fit

(Similar to regression ANOVA)

t.s. = G = improvement in $-2\ln(L)$ over a model with just a constant term

Compare to χ^2 with k d.f.

↗ # predictors

```
Likelihood ratio test
```

```
Model 1: Gender ~ Hgt
```

```
Model 2: Gender ~ 1
```

```
#df LogLik Df  chisq Pr(>chisq)
```

```
1    2   -67.814
```

```
2    1  -160.500  -1 185.37 < 2.2e-16 ***
```

```
---
```

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

```
require(lmtest)
```

```
lrtest(logitmod)
```

Deviance in R Output

```
call:  
glm(formula = Gender ~ Hgt, family = binomial, data = Pulse)  
  
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-2.77443 -0.34870 -0.05375  0.32973  2.37928  
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) 64.1416    8.3694   7.664 1.81e-14 ***  
Hgt          -0.9424    0.1227  -7.680 1.60e-14 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 321.00 on 231 degrees of freedom  
Residual deviance: 135.63 on 230 degrees of freedom  
AIC: 139.63  
  
Number of Fisher Scoring iterations: 6
```

-2LL

Constant Model

This Model

How much better do we do with height than with nothing?

$$G = 321.00 - 135.63 = 185.37$$

Test for Overall Model—Logistic

Is there something effective in the model?

$$H_0: \beta_1 = 0$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0$$

Same odds
for all X

$$H_1: \beta_1 \neq 0$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Odds are linear
function of X

$$t.s. = G = -2 \ln(L_0) - (-2 \ln(L_1)) \quad \text{Compare to } \chi^2_1.$$

Improvement in $-2\ln(L)$ when using linear function of X .

The Logistic Reg Max Likelihood Problem

Likelihood:
$$L = \prod \hat{\pi}_i^{y_i} (1 - \hat{\pi}_i)^{1-y_i}$$

where

$$\hat{\pi}_i = \frac{e^{\beta_0 + \beta_1 * x_i}}{1 + e^{\beta_0 + \beta_1 * x_i}}$$

For the putting data this is:

$$L = \left[\frac{e^{\beta_0 + \beta_1 * 3}}{1 + e^{\beta_0 + \beta_1 * 3}} \right]^{84} * \left[\frac{1}{1 + e^{\beta_0 + \beta_1 * 3}} \right]^{17} * \left[\frac{e^{\beta_0 + \beta_1 * 4}}{1 + e^{\beta_0 + \beta_1 * 4}} \right]^{88} * \left[\frac{1}{1 + e^{\beta_0 + \beta_1 * 4}} \right]^{31} * etc$$

Maximize this with respect to β_0 and β_1 .

Example: Golf Putts

Length	3	4	5	6	7
Made	84	88	61	61	44
Missed	17	31	47	64	90
$\hat{\pi}$	0.826	0.730	0.605	0.465	0.330

When Length = 3,
the model predicts

$$\frac{e^{3.256 - 0.5666 \cdot 3}}{1 + e^{3.256 - 0.5666 \cdot 3}} = 0.826$$

The Likelihood
component for
Length = 3 is
 $0.826^{84} 0.174^{17}$

Length	3	4	5	6	7
Made	84	88	61	61	44
Missed	17	31	47	64	90
$\hat{\pi}$	0.826	0.730	0.605	0.465	0.330

Combining for all the data, the Likelihood is:

$$L = 0.826^{84} 0.174^{17} 0.730^{88} 0.270^{31} \dots 0.330^{44} 0.670^{90}$$

$$\ln(L) = 84 \ln(0.826) + 17 \ln(0.174) + \dots + 90 \ln(0.670)$$

$$\ln(L) = -359.9$$

Coefficients are chosen to make $\ln(L)$ as large as possible.

Topic 11.2

Binary logistic regression

Goodness-of-fit tests

Saturated model

Logistic diagnostics

Overdispersion

Concordance

Pseudo- R^2

Recall: Evaluating Overall Fit (1 of 2)

1. Likelihood of the sample

Output gives ***residual deviance*** = $-2\log(L)$

- * Smaller $-2\log(L)$ is better
- * Coefficients are chosen to minimize $-2\log(L)$

For Putts1 model: $-2\log(L) = 719.89$

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	3.25684	0.36893	8.828	<2e-16	***
Length	-0.56614	0.06747	-8.391	<2e-16	***

Null deviance: 800.21 on 586 degrees of freedom
Residual deviance: 719.89 on 585 degrees of freedom

Recall: Evaluating Overall Fit (2 of 2)

2. Test for overall fit

G = improvement in $-2\log(L)$ over a model with just a constant term (null deviance)

Compare to χ^2 with k d.f. (chi-square)

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

predictors

```
Null deviance: 800.21 on 586 degrees of freedom
Residual deviance: 719.89 on 585 degrees of freedom
```

$$G = 800.2 - 719.9 = 80.3$$

`> 1-pchisq(80.3,1)`

P-value ≈ 0 , Reject H_0

[1] 0

Test Overall Model—Logistic

Is the predictor (or predictors) effective?

$$H_0: \beta_1 = 0$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0$$

Same odds
for all x

$$H_a: \beta_1 \neq 0$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x$$

Odds are linear function
of x (or multiple x_i)

$$G = -2 \log(L_0) - (-2 \log(L)) \quad \text{Compare to } \chi_k^2$$

Improvement in $-2\log(L)$ when using linear
function of x (or multiple x_i 's).

Goodness of Fit (Logistic)

Could we do much better with some *other* function of X ?

$$H_0: \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X$$

Odds are linear
function of X

$$H_a: \log\left(\frac{\pi}{1 - \pi}\right) = f(X)$$

vs.

Odds are *any*
function of X

Use $\hat{\pi}$'s from logistic model Use \hat{p} 's for each value of X

$$\chi^2 = -2 \log(L) - (-2 \log(L_{\hat{p}}))$$

Improvement in $-2 \log(L)$ when using
sample \hat{p} 's at each X instead of linear
model

Compare to χ^2
d.f. = # X values –
2
(requires multiple
cases at each X)

Example: Golf Putts

Length	3	4	5	6	7
Made	84	88	61	61	44
Missed	17	31	47	64	90
$\hat{\pi}$	0.826	0.730	0.605	0.465	0.330
\hat{p}	0.831	0.739	0.565	0.488	0.328

$$-2 \log(L) = 719.89$$

$$-2 \log(L_{\hat{p}}) = 718.82$$

Change in $-2\log(L)$ is $X^2 = 719.89 - 718.82 =$

Compare to χ^2_3

```
1-pchisq(1.07, 3)
```

```
[1] 0.7843207
```

Do not reject $H_0 \rightarrow$ The logistic model is probably adequate.

Finding $-2\log(L_{\hat{p}})$ with R (1 of 2)

Trick: Treat X as a categorical variable

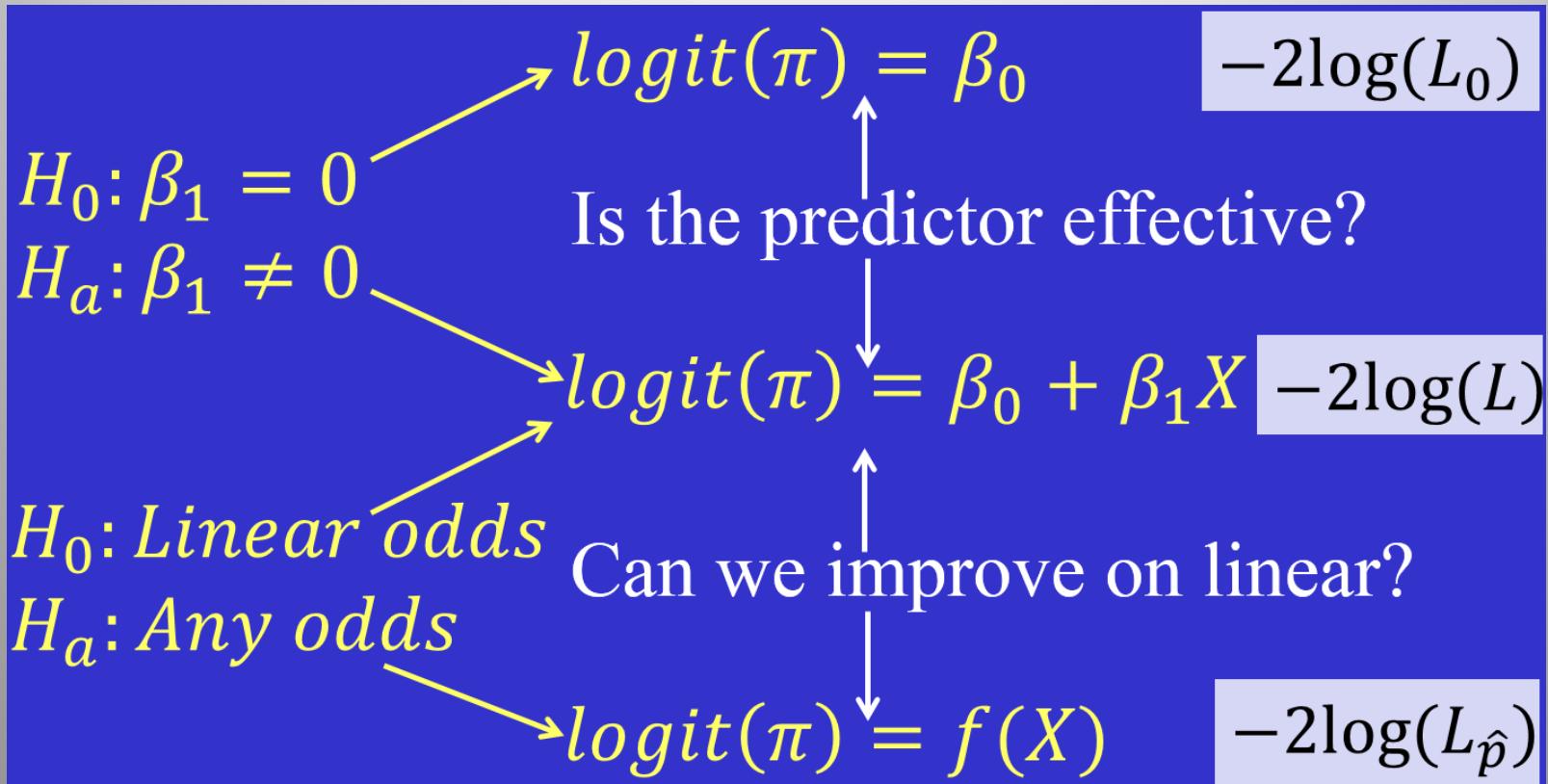
```
> saturated=glm(Made~factor(Length), family=binomial, data=Putts1)
> summary(saturated)

             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.5976    0.2659   6.007 1.89e-09 ***
factor(Length) 4 -0.5543    0.3382  -1.639   0.101
factor(Length) 5 -1.3369    0.3292  -4.061 4.90e-05 ***
factor(Length) 6 -1.6456    0.3205  -5.134 2.84e-07 ***
factor(Length) 7 -2.3132    0.3234  -7.154 8.46e-13 ***

Null deviance: 800.21 on 586 degrees of freedom
Residual deviance: 718.82 on 582 degrees of freedom
```

$-2\log(L_{\hat{p}})$ for saturated model

Finding $-2\log(L_{\hat{p}})$ with R (2 of 2)



Can extend these to models with multiple predictors.

Unusual Points

Recall: In ordinary regression, we use residual plots, leverage, and Cook's D as tools to identify unusual points.

Similar tools for logistic regression:

- Pearson residuals
- Deviance residuals
- Standardized residuals
- Leverage (same as in ordinary regression)
- Cook's D

Example: Pulse Data

Y = Sex (0 = Male, 1 = Female)

X = Height (inches)

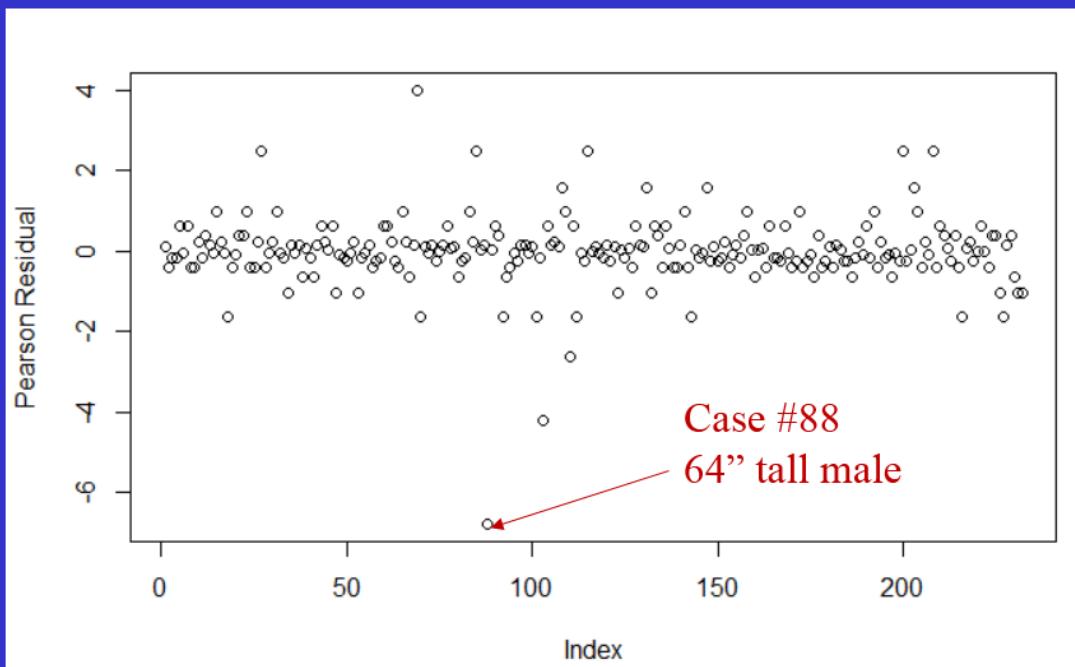
```
> lmodSex=glm(Sex~Hgt,family=binomial,data=Pulse)
> summary(lmodSex)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416    8.3694   7.664 1.81e-14 ***
Hgt          -0.9424    0.1227  -7.680 1.60e-14 ***
---
Null deviance: 321.00 on 231 degrees of freedom
Residual deviance: 135.63 on 230 degrees of freedom
```

Pearson Residuals

$$\text{Pearson residual} = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

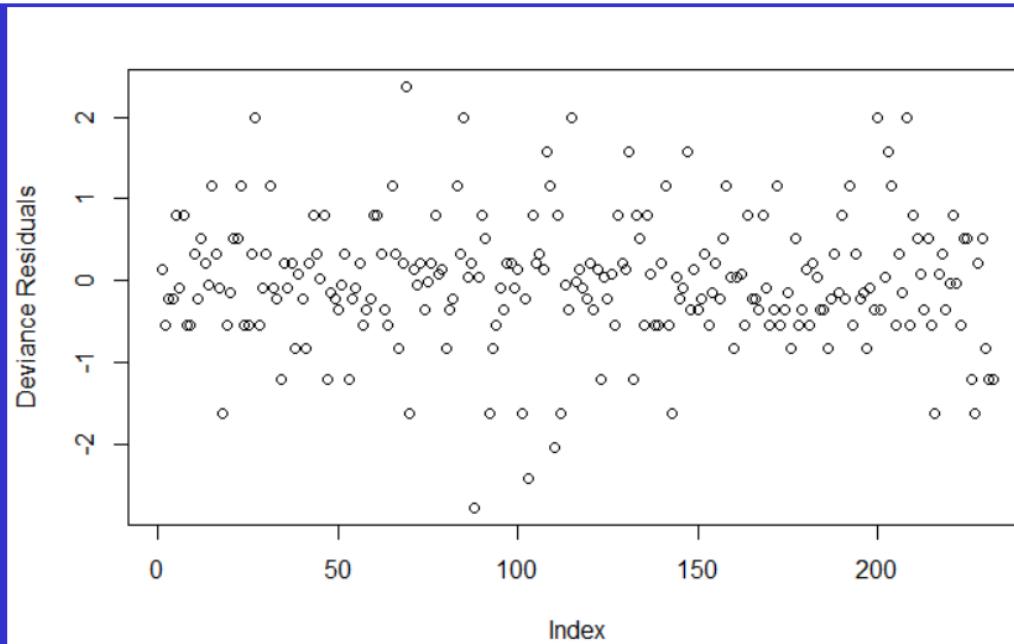
```
> plot(residuals(lmodSex, "pearson"), ylab="Pearson Residual")
```



Deviance Residuals

$$\text{Deviance residual} = \begin{cases} \sqrt{-2\log(\hat{\pi}_i)} & \text{if "yes"} \\ -\sqrt{-2\log(1 - \hat{\pi}_i)} & \text{if "no"} \end{cases}$$

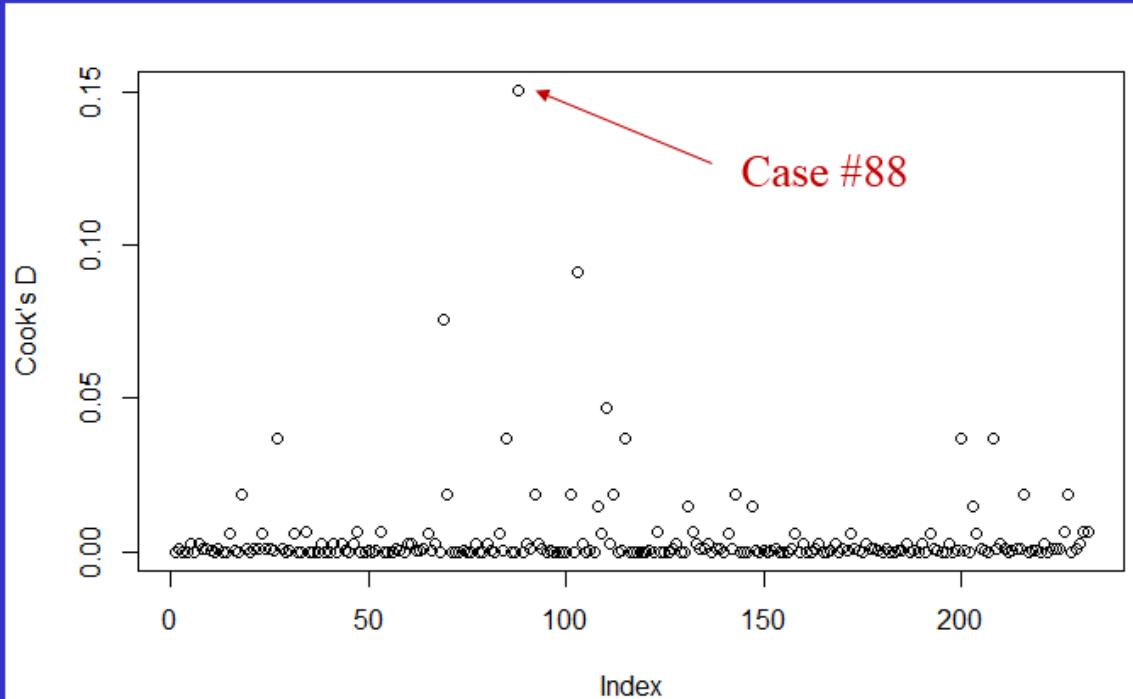
```
> plot(residuals(lmodSex), ylab="Deviance Residual")
```



Cook's Distance

As in ordinary regression, uses both residuals and leverage.

```
> plot(cooks.distance(lmodSex) , ylab="Cook's D")
```



Measuring Overall Fit: Pseudo-R²

Recall for ordinary regression:

$$R^2 = 1 - \frac{SSE}{SStotal}$$

One way to mimic this for logistic regression is to use $-2\log(L)$ in place of sums-of-squares

variance

$$\text{pseudo } R^2 = 1 - \frac{-2 \log(L)}{-2 \log(L_0)} = 1 - \frac{\log(L)}{\log(L_0)}$$

i.e., We compare the residual and null deviances.

Example: ICU Data (3 of 3)

For SysBP as a predictor of Survive:

```
> lmodBP=glm(Survive~SysBP,family=binomial,data=ICU)
> summary(lmodBP)
---
    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 191.34  on 198  degrees of freedom

> 1-summary(lmodBP)$deviance/summary(lmodBP)$null.deviance
0.04409365
```

For Pulse as a predictor of Survive:

```
> lmodPulse=glm(Survive~Pulse,family=binomial,data=ICU)
> summary(lmodPulse)
---
    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 199.96  on 198  degrees of freedom

> 1-summary(lmodPulse)$deviance/summary(lmodPulse)$null.deviance
[1] 0.001001958
```

Additional Concepts

These are totally optional...

Overdispersion (1 of 2)

Sometimes seemingly binomial data, for which the variance should be predictable, actually deviate from being binomial.

Try the Putts3 data (short form):

```
> lmodover=glm(cbind(Made, Missed) ~ Length, family=binomial, data=Putts3)
> summary(lmodover)

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.25684   0.36893   8.828 <2e-16 ***
Length      -0.56614   0.06747  -8.391 <2e-16 ***
-----
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 87.1429 on 4 degrees of freedom
Residual deviance: 6.8257 on 3 degrees of freedom
```

If Residual deviance is well above the d.f., then we may have “extra-binomial” variation.

Overdispersion (2 of 2)

What to do if we suspect overdispersion?

Change the **family** from **binomial** to **quasibinomial**

```
> lmodQ=glm(cbind(Made,Missed) ~Length,family=quasibinomial,data=Putts3)
> summary(lmodQ)
```

	Estimate	Std. Error	t value	Pr(> t)	P-values not as extreme
(Intercept)	3.2568	0.5552	5.866	0.00988 **	
Length	-0.5661	0.1015	-5.576	0.01139 *	

(Dispersion parameter for quasibinomial family taken to be 2.264714)

```
Null deviance: 87.1429 on 4 degrees of freedom
Residual deviance: 6.8257 on 3 degrees of freedom
```

Original Putts2 Data

```
> lmodPutt2=glm(cbind(Made, Missed) ~ Length, family=binomial, data=Putts2)
> summary(lmodPutt2)
```

	Estimate	Std. Error	z value	Pr(> z)							
(Intercept)	3.25684	0.36893	8.828	<2e-16 ***							
Lengths	-0.56614	0.06747	-8.391	<2e-16 ***							

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 81.3865 on 4 degrees of freedom
Residual deviance: 1.0692 on 3 degrees of freedom

Not much concern with overdispersion.

Measuring Overall Fit: Concordance

Basic idea:

- Consider a case where the response is “yes” and the model gives a predicted yes proportion of $\hat{\pi}_i$.
- Look at a case with a “no” response that has predicted yes proportion of $\hat{\pi}_j$.
- If $\hat{\pi}_i > \hat{\pi}_j$, we say that pair of cases is concordant (good; a “yes” should have a higher $\hat{\pi}$ than a “no”)
- If $\hat{\pi}_i < \hat{\pi}_j$, we say that pair of cases is discordant.
- Compare every “yes” case to every “no” case and find the proportion of pairs that are concordant.

Example: ICU Data (1 of 3)

Consider a model to predict **Survive** for ICU patients using blood pressure (**SysBP**). The data have 160 cases that survive and 40 cases that did not survive

```
> lmodBP=glm(Survive~SysBP,family=binomial,data=ICU)
> summary(lmodBP)

      Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.777195   0.757375 -1.026  0.30481
SysBP        0.017019   0.006001  2.836  0.00456 **

---
Null deviance: 200.16 on 199 degrees of freedom
Residual deviance: 191.34 on 198 degrees of freedom
```

There are $160 \times 40 = 6400$ possible pairs

	Concordant	Discordant	Ties
Number	3960	2287	153
Percent	61.9%	35.7%	2.4%

Example: ICU Data (2 of 3)

Compare the 61.9% concordance for using SYSBP as the predictor to a model using Pulse rate.

```
> lmodPulse=glm(Survive~Pulse,family=binomial,data=ICU)
> summary(lmodPulse)

              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.679129  0.679863  2.470   0.0135 *
Pulse        -0.002941  0.006552 -0.449   0.6535
---
Null deviance: 200.16 on 199 degrees of freedom
Residual deviance: 199.96 on 198 degrees of freedom
```

	Concordant	Discordant	Ties
Number	3252	2973	175
Percent	50.8%	46.5%	2.7%

Pulse is a weaker predictor than SysBP.