Harvard Department of Government 2003
Faraway Chapters 2-3, Binomial Data

JEFF GILL
Visiting Professor, Fall 2024

Overview

- We will create a regression model for dichotomous outcome variables: vote/not-vote, war/no-war, pass/fail, etc.
- Note that this is different than having dichotomous explanatory variables.
- Remember that regression is really conditional average, $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$, which does not have the same implications for $0/1$ outcomes on the LHS.
- Consider the probability that a single case has a 0 or a 1 as the outcome:

$$\pi_i = p(Y_i) = p(Y = 1|\mathbf{X} = \mathbf{x}_i), \quad \text{where} \quad \pi \in [0{:}1].$$

- So:

$$\mathbb{E}(Y_i|\mathbf{x}_i) = (\pi_1)(1) + (1 - \pi_i)(0) = \pi_i.$$

  (recall that for discrete RV $\mathbb{E}(A) = \sum_{\text{over events}} P(\text{A}) \times A$)

- This means that we are *estimating* an underlying probability value for given levels of a vector of explanatory variable values.

Legal Background

- We are interested in the impact of the defendant's race in judge or jury decisions to impose the death penalty versus life in prison for convicted murders.
- Most studies focus on southern states, including Georgia.
- Before 1972 Georgia (plus other states) gave juries wide discretion in deciding whether to impose the death penalty on defendants convicted of death-eligible murder offenses.

Legal Background

- In *Furman v. Georgia*, the U.S. Supreme Court struck down this feature of Georgia's capital sentencing procedure and by implication invalidated the death penalty across the U.S.
- This 5-4 decision stated that capital sentencing based on the relatively unguided discretion of juries violates the "cruel and unusual punishment" clause of the 8th Amendment, because it permits juries to impose the irreversible sentence of death on some defendants while other juries can impose the sentence of life imprisonment under similar circumstances.
- Interesting, there was no majority opinion produced in the case.
- But Justice Stewart wrote "For, of all the people convicted of rapes and murders in 1967 and 1968, many just as reprehensible as these, the petitioners are among a capriciously selected random handful upon whom the sentence of death has in fact been imposed."

Legal Background

- After the Furman decision, Georgia, amended their death penalty statute to meet the new Furman guidelines, which were approved by the Supreme Court.
- After the defendant was convicted of a capital crime (the first part of the bifurcated trial proceeding), there is an second hearing at which the jury received additional evidence in aggravation and mitigation.
- In order for the defendant to be made eligible for the death penalty, the jury must first determine the existence at least one of ten aggravating factors.
- Passing this hurdle, the jury then evaluates all trial evidence including mitigating evidence and additional aggravating evidence.
- This is called a *non-weighing* scheme because the jury is not required to weigh the statutory aggravating factors against mitigating evidence before imposing a death sentence.

Study Background

- David C. Baldus, Charles Pulaski, and George Woodworth (eg. the Baldus study) looked at the potential disparity in the imposition of the death sentence in Georgia based on the race of the murder victim and the race of the defendant.
- This is actually two studies, the second one examining about 762 cases with a murder conviction in Georgia from March 1973 to December 1979.
- The data contains 160 variables, including legal background, crime description, and demographics.
- From the 1970 US Census 1,187,149/4,589,575 or about 26% of Georgia residents were black.
- The death penalty was imposed:

22% cases of Black defendant, White victim
8% cases of White defendant and White victim
1% of cases of Black defendant and Black victim
3% of cases of White defendant and Black victim

Study In the Legal Setting

- The Baldus study was cited in the US Supreme Court in McClesky v. Kemp (1987), in which a black defendant (McClesky) was sentenced to death for killing a white police officer in Georgia.
- The central argument was that the sentence violated the Equal Protection clause of the 14th Amendment, since statistically he stood was more likely to get the death penalty since the victim was white.
- The Court (5-4) rejected McClesky's argument, on the grounds that statistical trends did not effectively *prove* the existence of discrimination among the jury who decided his particular case (Justice Powell).
- Justice Powell later told his biographer that McCleskey was the biggest mistake in his career and that if he could to do it over again, he would rule the that death penalty always unconstitutional (Jeffries 1994).
- McClesky was executed in 1991.

Some Citations (Or Why This Example is Important In This Context)

- Imbens & Rubin, *New Palgrave Dictionary of Economics* 2008.
- Greiner & Rubin, *Review of Economics and Statistics* 2011.
- Petrie & Coverdill, *Social Problems* 2010.
- Angrist, Imbens, and Rubin, *Journal of the American Statistical Association* 1996.
- Hundreds of law review articles.

Data Manipulation: Potential Explanatory Variables

- Hispanic and "other" removed from cases for clarity ($n_r = 45$).
- `race`: 0=white ($n_w = 297$), 1=black ($n_b = 463$)
- `educatn`: 1=middle school or lower, 2=some high school, 3=high school degree
- `employm`: 0=unemployed, 1=employed
- `SES`: 0=not low wage, 1=low wage
- `married`: 0=unmarried, 1=married
- `num.chld`: defendant's number of children (1-9+)
- `military`: -1=not honorable or not general discharge, 0=no military, 1=honorable, general, or currently serving
- `pr.arrst`: number of prior arrests
- `pr.incrc`: record shows prior incarceration in Georgia
- `plea`: 0="not guilty," 1="guilty"

Data Manipulation: Potential Explanatory Variables

- `defense`: 1=retained, 2=appointed
- `dp.sght`: did prosecution seek death penalty, yes=2, no=1
- `jdge.dec`: did judge take death penalty issue away from jury, 0=unknown, 1=yes, 0=no
- `pen.phse`: was there a penalty trial, 1=yes, 0=no
- `did.appl`: did the defendant appeal, 1=yes, 0=no
- `out.appl`: 1=conviction and dp affirmed, 2=conviction affirmed dp changed to life, 3=conviction reversed, 4=conviction and life affirmed, 5=conviction only reversed, 6=conviction affirmed life modified, 9=no appeal
- `vict.age`: 1=12 or less, 0=13 or more
- `vict.sex`: 1=male, 2=female
- `vict.rel`: 0=non-family, 1=family
- `vict.st1`: 1=police or judicial official, 0=otherwise
- `specialA`: 1=special/cruel circumstances, 0=otherwise

Data Manipulation: Potential Explanatory Variables

- `methodA`: 1=gun, 2=knife, 3=blunt object, 4=beating, 5=fractures, 8=hand strangulation, 10=rope/garrote, 14=drowning, 21=buried alive, 24=other
- `num.kill`: 1, 2, or 3.
- `num.prps` number of co-perpetrators in addition to defendant
- `def.age`: defendant's age according to: $1(<= 16), 2(17-20), 3(21-25), 4(26-35), 5(36-50), 6(> 50)$
- `aggrevat`: one or more aggravated method of killing
- `bloody`: bloody murder involved
- `fam.lov`: family or lover dispute
- `insane`: insanity defense used
- `mitcir`: one or more mitigating circumstances
- `num.depr`: number of depraved circumstances in murder
- `rape`: rape involved

Data Manipulation, Restriction, Matching, and Outcome Variable

- `sentence`: 0=life sentence(325), 1=death penalty (127)
- Pre/Post-Furman Breakdown:

	Not-DP	DP
Pre	112	44
Post	494	112

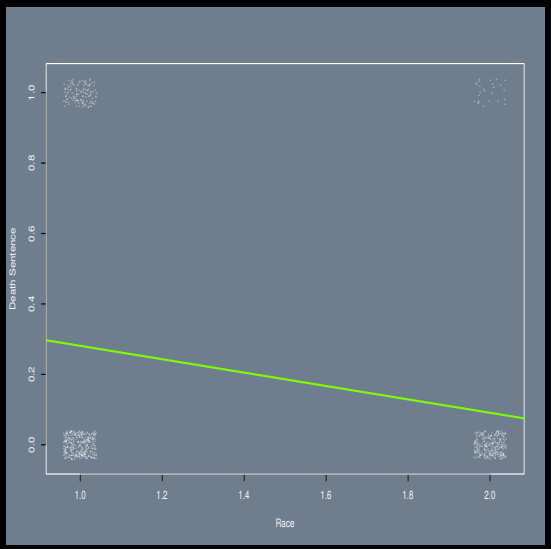
- `vict.rac`: victim's race (white=454, black=287)
- Load from:

```
baldus <- read.table(
    "https://jeffgill.org/wp-content/uploads/2024/07/baldus.clean2_.txt",
    header=TRUE)
```

Death Penalty Example

- Naïve linear-probability model:

```
lmod <- lm(sentence ~ vict.rac, data=baldus)
summary(lmod)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28097 -0.28097 -0.09091 -0.09091  0.90909

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.47104    0.04391  10.727  < 2e-16
vict.rac    -0.19006    0.02986  -6.365 3.43e-10

Residual standard error: 0.3952 on 736 degrees of freedom
Multiple R-squared:  0.05217,        Adjusted R-squared:  0.05089
F-statistic: 40.51 on 1 and 736 DF,  p-value: 3.432e-10

par(col.axis="white",col.lab="white",col.sub="white",col="white",
    bg="slategray")
plot(jitter(sentence,0.2) ~ jitter(vict.rac,0.2), baldus, pch="+",
     xlab="Race", ylab="Death Sentence", cex=0.35)
abline(lmod, col="lawngreen",lwd=3)
```



Problems with the Linear-Probability Approach

- Allows predictions outside of $[0:1]$.
- Deceptive sense of fit:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.47104    0.04391  10.727  < 2e-16
vict.rac    -0.19006    0.02986  -6.365 3.43e-10
```

- Wrong distributional implications:

$$Y_i = \alpha + \beta x_i + \epsilon \;\Rightarrow\; \pi_i = \alpha + \beta x_i,$$

but since $Y_i \in \{0, 1\}$, then ϵ_i is dichotomous not normally distributed:

$$\epsilon_i = 1 - \mathbb{E}[Y_i] = 1 - (\alpha + \beta x_i) = 1 - \pi_i$$

or

$$\epsilon_i = 0 - \mathbb{E}[Y_i] = 0 - (\alpha + \beta x_i) = -\pi_i$$

Problems with the Linear-Probability Approach

- The expectation is okay:

$$\mathbb{E}[\epsilon_i] = \mathbb{E}[Y_i - \alpha - \beta x_i] = \pi_i - \pi_i = 0,$$

but the variance is wrong:

$$\mathrm{Var}[\epsilon_i] = \mathbb{E}[\epsilon_i^2] - (\mathbb{E}[\epsilon_i])^2 = \mathbb{E}[\epsilon_i^2],$$

which turns out to be:

$$\begin{aligned}
\mathbb{E}[\epsilon_i^2] &= \sum_{i=0}^{1} \epsilon_i^2 p(\epsilon_i) \\
&= (1-\pi_i)^2(\pi_i) + (-\pi_i)^2(1-\pi_i) \\
&= (1-\pi_i)[(1-\pi_i)(\pi_i) + \pi_i^2] \\
&= (1-\pi_i)\pi_i \\
&= \pi_i - \pi_i^2.
\end{aligned}$$

This is a quadratic form and is therefore heteroscedastic, especially near zero and one.

Ad Hoc "Fix:" Constrained Linear-Probability Model

- Fix π artificially:

$$\pi = \begin{cases} 0 & 0 > \alpha + \beta x \\ \alpha + \beta x & 0 \leq \alpha + \beta x \leq 1 \\ 1 & \alpha + \beta x > 1 \end{cases}$$

- The hardest part is finding a criterion for 0 1 on the x-axis (corner points).
- The effect of x is difficult to interpret.
- This is a difficult estimation problem.
- The abrupt changes are substantively unreasonable (do not have derivatives).

New Conceptual Model

- Start with the linear predictor $\boldsymbol{\eta} = \alpha + \beta\mathbf{x}$.
- Now let's specify a link function that relates the linear additive RHS component to the expected value of the nonlinear LHS component:

$$\pi_i = g^{-1}(\eta_i) = p(\alpha_i + \beta_i x) \;\Rightarrow\; g(\pi_i) = \eta_i = \alpha_i + \beta_i x.$$

- Objectives for $g^{-1}()$:
  - ▷ smooth on [0:1]
  - ▷ For a positive effect of $\mathbf{x}_i$ on π_i:
    - $g^{-1} \rightarrow 0$ as $x_i \rightarrow, -\infty$
    - $g^{-1} \rightarrow 1$ as $x_i \rightarrow, +\infty$.
  - ▷ For a negative effect of $\mathbf{x}_i$ on π_i:
    - $g^{-1} \rightarrow 1$ as $x_i \rightarrow, -\infty$
    - $g^{-1} \rightarrow 0$ as $x_i \rightarrow, +\infty$.

New Conceptual Model

- There are two common solutions for $g^{-1}()$.
- Logit:

$$\Lambda(\eta_i) = [1 + \exp(-\eta_i)]^{-1}$$

- Probit:

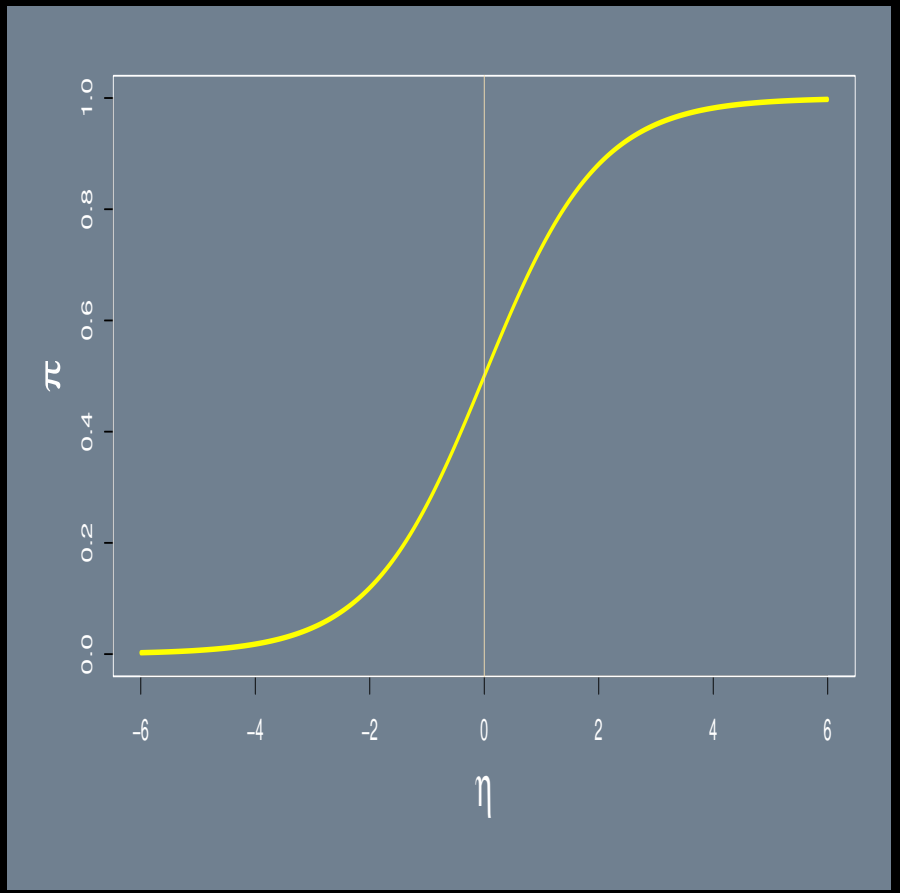
$$\Phi(\eta_i) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\eta_i} \exp[-\frac{1}{2}\eta_i^2] d\eta_i$$

- These are sometimes given in $g()$ form: $\Phi^{-1}(\pi_i)$ and $\Lambda^{-1}(\pi_i) = \text{logit}(\pi_i) = \log\left(\frac{p_i}{1-p_i}\right)$.
- Less common is the cloglog function:

$$g(\mu) = -\log\left(-\log(1-\mu)\right) \qquad g^{-1}(\eta) = 1 - \exp\left(-\exp(\eta)\right)$$

Latent Variable Justification

- Humans make dichotomous decisions from smooth preference structures, but we only see discrete choices in the data.
- The Index Function (Utility) model states that if *benefits* - *costs* = U is greater than zero then the choice should be a one, and vice-versa.



Latent Variable Justification

- Utility model states: $U_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ (subsume the constant into the vector), and $p(U_i > 0) = p(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i > 0) = p(\boldsymbol{\epsilon}_i > -\mathbf{x}_i\boldsymbol{\beta})$.
- Political Example:
  - U^R, the utility of voting for the Republican candidate
  - U^D, the utility of voting for the Democratic candidate
  - direction is arbitrary, so pick $Y = 1$ the decision to vote for the Republican candidate
  - Define the two utility functions in regression terms:

$$U_i^R = \mathbf{x}_i\boldsymbol{\beta}_R + \boldsymbol{\epsilon}_{iR} \qquad U_i^D = \mathbf{x}_i\boldsymbol{\beta}_D + \boldsymbol{\epsilon}_{iD}$$

  - So now:

$$\begin{aligned} p(Y_i = 1|\mathbf{x}_i) &= p(U_i^R > U_i^D) \\ &= p(\mathbf{x}_i\boldsymbol{\beta}_R + \boldsymbol{\epsilon}_{iR} > \mathbf{x}_i\boldsymbol{\beta}_D + \boldsymbol{\epsilon}_{iD}|\mathbf{x}_i) \\ &= p(\mathbf{x}_i[\boldsymbol{\beta}_R - \boldsymbol{\beta}_D] + \boldsymbol{\epsilon}_{iR} - \boldsymbol{\epsilon}_{iD} > 0) \\ &= p(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon} > 0) \end{aligned}$$

    which is just 1-CDF.

Binomial Regression Model

- If Y_i for $i = 1, \ldots, n$ is iid binomial $B(n_i, p_i)$, then:

$$p(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

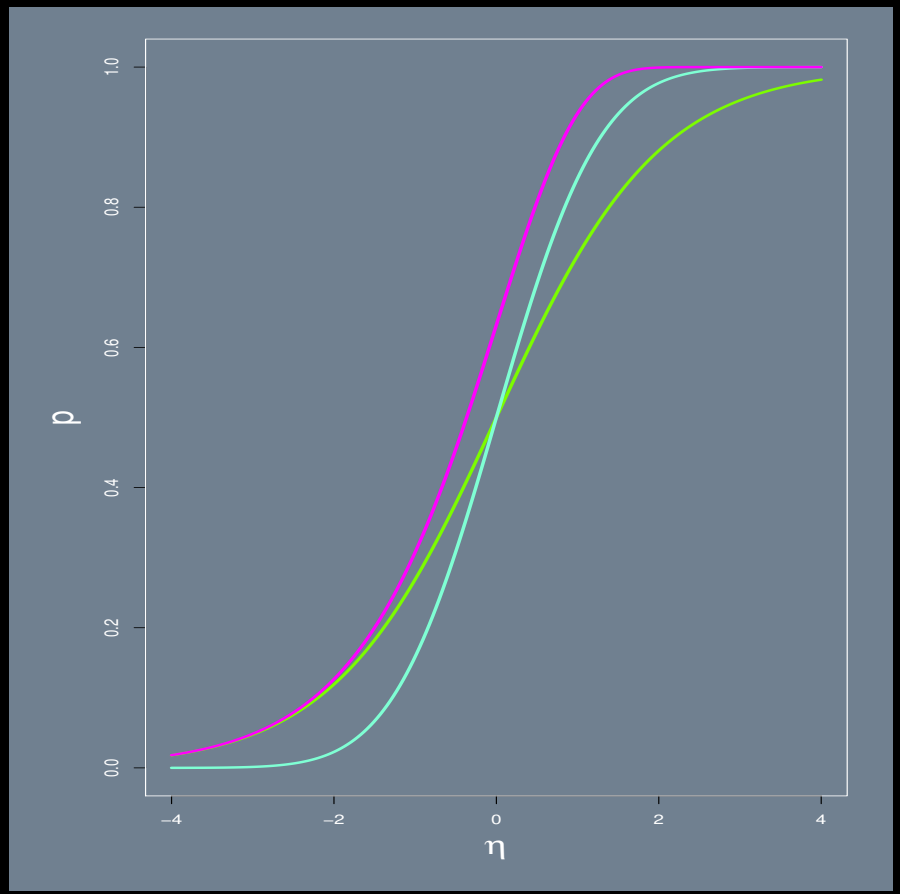
- Further suppose that these are affected by the same q predictors (covariates, explanatory variables), $x_{i1}, \ldots, x_{iq}$.
- The tool that connects these predictors to p is the linear predictor:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}.$$

- We still need a link function, $\eta_i = g(p_i)$, that is not an identity ($\eta_i = p_i$) since we need $0 \leq p_i \leq 1$.

Binomial Link Functions

- Logit (logistic): $\eta = \log\left(\frac{p}{1-p}\right)$, $p = \frac{\exp(\eta)}{1+\exp(\eta)}\,[1+\exp(-\eta)]$.
- Probit: $\eta = \Phi^{-1}(p)$, $p = \Phi(\eta)$.
- Complementary log-log: $\eta = \log(-\log(1-p))$, $p = 1-\exp(-\exp(\eta))$.




```
ruler <- seq(-4,4,length=200)
postscript("Class.MLE/faraway.ch2.fig3.ps")
par(col.axis="white",col.lab="white",col.sub="white",
    col="white", bg="slategray",cex.lab=2,mar=c(6,6,2,2))
plot(ruler,exp(ruler)/(1+exp(ruler)),type="l",lwd=3,
     col="lawngreen",ylim=c(0,1),
     xlab=expression(eta),ylab="p")
lines(ruler,pnorm(ruler),lwd=3,col="aquamarine")
lines(ruler,1-exp(-exp(ruler)),lwd=3,col="magenta")
dev.off()
```

Binomial Treatment of the Death Penalty Data

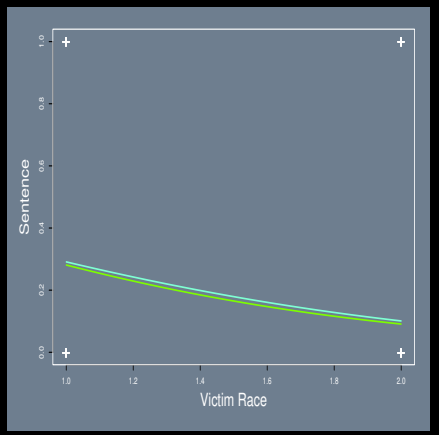
- Consider the Baldus data as a binomial experiment:

```
t(cbind(baldus$sentence,baldus$vict.rac))[,1:12]
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
[1,]    0    0    0    0    0    0    1    1    0     0     0     1
[2,]    1    2    1    2    2    1    1    1    2     2     2     2
```

Death Penalty Example

- Now that we have a reasonable set of assumptions, the regression model is estimated with *maximum likelihood* (note the binomial treatment of the data):

```
logitmod <- glm(sentence ~ vict.rac, family=binomial, baldus)
summary(logitmod)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.4233     0.2934   1.443    0.149
vict.rac     -1.3629     0.2307  -5.907 3.49e-09

probitmod <- glm(sentence ~ vict.rac,
                 family=binomial(link=probit), data=baldus)
summary(probitmod)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1753     0.1628   1.076    0.282
vict.rac     -0.7552     0.1213  -6.224 4.86e-10

ilogit <- inv.logit <- function(Xb)  1/(1+exp(-Xb))

postscript("CLASSES/Class.MLE/Images/baldus.fig2.ps")
par(col.axis="white",col.lab="white",col.sub="white",
    col="white", bg="slategray",cex.lab=2,mar=c(6,6,2,2))
plot(sentence ~ vict.rac, baldus, pch="+",
     ylab="Sentence", xlab="Victim Race", cex=2)
x <- seq(1,2,length=100)
lines(x,ilogit(0.4233-1.3629*x),col="lawngreen",lwd=3)
lines(x,pnorm(0.1753-0.7552*x)+0.01,col="aquamarine",lwd=3)
dev.off()
```



Binomial Model Estimation

- Define a likelihood function for observed iid y_i, where $i = 1, \ldots, n$ from $f(y|p)$.
- Then the *joint distribution* of these observed data is:

$$p(y_1, y_2, \ldots, y_n) = p(y_1|\boldsymbol{\beta}, \mathbf{x}_1) f(y_2|\boldsymbol{\beta}, \mathbf{x}_2) \cdots f(y_n|\boldsymbol{\beta}, \mathbf{x}_n) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}, \mathbf{x}_i).$$

- If we consider that p is really the unknown and the y_i are known, then it makes sense to think of this joint function as a function that reveals something about $\boldsymbol{\beta}$.
- Denote it $L(\boldsymbol{\beta}|\mathbf{x}, \mathbf{y})$, which is called a likelihood function.

Binomial Model Estimation

- More precisely, we can incorporate the information that Y can only be 0 or 1:

$$L(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}) = \prod_{y_i=0}[1-F(\mathbf{X}_i\boldsymbol{\beta})]\prod_{y_i=1}[F(\mathbf{X}_i\boldsymbol{\beta})]$$

$$= \prod_{i=1}^{n}[1-F(\mathbf{X}_i\boldsymbol{\beta})]^{1-y_i}[F(\mathbf{X}_i\boldsymbol{\beta})]^{y_i}$$

$$\ell(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n}[(1-y_i)\log(1-F(\mathbf{X}_i\boldsymbol{\beta})) + y_i\log(F(\mathbf{X}_i\boldsymbol{\beta}))]$$

- The log-likelihood is concave to the x-axis for common choices of $F()$, and produces coefficient estimates that are distributed student's-t.
- Generally with the binomial setup it is easier to think in terms of the CDF, $F()$, rather than the PDF, $f()$, since the former directly describes the S-curve of theoretical interest.

Binomial Model MLE

- The gradient is given by:

$$G = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \left[\frac{y_i f_i}{F_i} + (1 - y_i) \frac{-f + i}{1 - F_i} \right] \mathbf{x}_i$$

- The Hessian is given by:

$$H = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \ell(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \frac{f_i^2}{F_i(1 - F_i)} \mathbf{x}_i \mathbf{x}_i'$$

- The Variance-Covariance Matrix is calculated as:

$$VC_{\boldsymbol{\beta}} = E\left[-H^{-1}\right]$$

Common Forms

- Probit, where $\phi_i = \phi_i(\mathbf{x}_i\boldsymbol{\beta})$ and $\Phi_i = \Phi_i(\mathbf{x}_i\boldsymbol{\beta})$:

$$G = \sum_{y=0} \frac{-\phi_i}{1-\Phi_i}\boldsymbol{\beta}\mathbf{x}_i + \sum_{y_i=1} \frac{\phi_i}{\Phi_i}\boldsymbol{\beta}\mathbf{x}_i$$

$$H = \left\{ \sum_{i=0} \left[-\frac{-\phi_i^2}{(1-\Phi_i)^2} + \frac{\mathbf{x}_i\boldsymbol{\beta}\phi_i}{1-\Phi_i} \right] + \sum_{i=1} \left[-\frac{\mathbf{x}_i\boldsymbol{\beta}\phi_i}{\Phi_i} - \phi_i^2 \right] \right\} \mathbf{x}_i\mathbf{x}_i'$$

$$VC_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\phi_i^2}{\Phi_i(1-\Phi_1)}\mathbf{x}_i\mathbf{x}_i'$$

- Logit, where $\Lambda_i = 1/[1+\exp(\mathbf{X}_i\boldsymbol{\beta})]$:

$$G = \sum_{i=1}^{n} (y_i - \Lambda_i)\mathbf{x}_i \qquad H = \sum_{i=1}^{n} \left\{ -\Lambda_i(1-\Lambda_i) \right\} \mathbf{x}_i\mathbf{x}_i'$$

$$VC_{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} \left\{ \Lambda_i(1-\Lambda_i) \right\} \mathbf{x}_i\mathbf{x}_i' \right]^{-1}$$

Interpretation of Individual Binomial $\boldsymbol{\beta}$ Results

- sign of the parameter estimate
- predicted/fitted values
- marginal effects, including first differences
- derivative methods
- Note $\text{logit}(\boldsymbol{\beta}) \approx \frac{\pi}{\sqrt{3}}\text{probit}(\boldsymbol{\beta})$
- Wald (t-tests) for significance:

$$W = (R\hat{\boldsymbol{\beta}} - q)\left[R(VC_{\hat{\boldsymbol{\beta}}})R'\right]^{-1}(R\hat{\boldsymbol{\beta}} - q)$$

for $H_0 : R\hat{\boldsymbol{\beta}} = q$ (commonly $R = 1, q = 0$, so that $W \sim F_{df=J,n-K}$. (where J is the number of restrictions stipulated in R). For individual coefficients, this reduces to:

$$W_k = (\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k / VC_{\hat{\boldsymbol{\beta}}}[k,k])^{\frac{1}{2}} \sim t_{df=n-k}$$

(where $n \times k$ is the dimension of the $\mathbf{X}$ matrix).

- Note that the F-test is more robust than the t-test (Hauck-Donner effect, JASA 1977).

Percent Predicted Correctly

- Compares actual against predicted in a 2-by-2 table:

		Prediction 0	1
Data	0	correct	incorrect
	1	incorrect	correct

- But wait! These models do not produce predicted 0/1 values, for instance:

```
round(logitmod$fitted.values,3)
    1     2     3     4     5     6     7     8     9    10    11    12
0.281 0.091 0.281 0.091 0.091 0.281 0.281 0.281 0.091 0.091 0.091 0.091
```

from the Bernoulli treatment.

Percent Predicted Correctly

- The naïve criteria:

$$p_i = 1 \text{ if, } F(\mathbf{x}_i\boldsymbol{\beta}) > 0.5 \qquad p_i = 0 \text{ if, } F(\mathbf{x}_i\boldsymbol{\beta}) < 0.5$$

- Create the table:

```
ppc <- cbind(baldus$sentence, round(logitmod$fitted.values,3))
( naive <- matrix(c(
    nrow(ppc[(ppc[,1] == 0) & (ppc[,2] < 0.5),])/nrow(ppc),
    nrow(ppc[(ppc[,1] == 0) & (ppc[,2] > 0.5),])/nrow(ppc),
    nrow(ppc[(ppc[,1] == 1) & (ppc[,2] < 0.5),])/nrow(ppc),
    nrow(ppc[(ppc[,1] == 1) & (ppc[,2] > 0.5),])/nrow(ppc)),
    byrow=TRUE,ncol=2) )
```

```
          [,1] [,2]
[1,] 0.7926829    0
[2,] 0.2073171    0
```

Percent Predicted Correctly

- Better criteria: mean of $\hat{y}_i$or a substantive/theoretical point.

```
ppc <- cbind(baldus$sentence, round(logitmod$fitted.values,3))
( mean.criteria<- matrix(c(
    nrow(ppc[(ppc[,1] == 0) & (ppc[,2] < mean(ppc[,1])),])/nrow(ppc),
    nrow(ppc[(ppc[,1] == 0) & (ppc[,2] > mean(ppc[,1])),])/nrow(ppc),
    nrow(ppc[(ppc[,1] == 1) & (ppc[,2] < mean(ppc[,1])),])/nrow(ppc),
    nrow(ppc[(ppc[,1] == 1) & (ppc[,2] > mean(ppc[,1])),])/nrow(ppc)),
    byrow=TRUE,ncol=2) )

           [,1]      [,2]
[1,] 0.35230352 0.4403794
[2,] 0.03523035 0.1720867
```

Binomial Model Comparison

- Compare two models, one with ℓ parameters and one with s parameters such that $\ell > s$ and every parameter in the s set is also in the ℓ set: nesting.
- Denote the first as $L(p|\mathbf{y}, \mathbf{X}_L) = L_L$ and the second as $L(p|\mathbf{y}, \mathbf{X}_S) = L_S$.
- A tool for comparing these models is the likelihood ratio statistic:

$$LRT = 2\log\frac{L_L}{L_S} = 2(\log(L_L) - \log(L_S)) = -2\log\frac{L_S}{L_L} = -2(\log(L_S) - \log(L_L)).$$

- This is distributed asymptotically χ^2 with degrees of freedom the difference between the number of parameters in the two models.
- Tail values support the nesting values, meaning that the restricted values are not supported.

Binomial Model Comparison

- The most extreme case of L_L fits a "covariate" to every datapoint as an indicator function, and is thus a regression model where every datapoint is a separate inference.
- This is called the saturated model and provides no data-reduction and no modeling value, but serves as a reference point.
- For the binomial model, the saturated model can be described by $\hat{p}_i = y_i/n_i$, which is the number of success over the number of trials for the i th case (frequently $n_i = 1$).
- Another reference point is a model that uses β_0 only and is called a *mean model*.
- Thus any model we specify "lives" between these two extremes of model fit.
- Residuals in the nonlinear regression sense are called deviances to distinguish them from the assumptions in linear models.

Binomial Model Comparison

- So it should be clear that:

$$\sum D_{\text{saturated model}} < \sum D_{\text{our specified model}} < \sum D_{\text{mean model}}$$

- For the binomial model, the LRT reduces to a ratio of the saturated model to the specified model, given by:

$$D = 2\sum_{i=1}^{n}\{y_i \log(y_i/\hat{y}_i) + (n_i - y_i)\log((n_i - y_i)/(n_i - \hat{y}_i))\},$$

  where $\hat{y}_i$ are the fitted values from the smaller (specified) model.

- The mean model provides a large value of D called the *null deviance*.

- D for assessing a model with p covariates is asymptotically distributed χ^2_{n-p}, where $n - p$ is the degrees of freedom.

- Returning the Death Penalty example ($n = 23$), I left off the following information before: NEED

```
summary(logitmod)
    :
    Null deviance: 753.32  on 737  degrees of freedom
Residual deviance: 711.11  on 736  degrees of freedom
```

Binomial Model Comparison

- Formal tests:
  - Specified model versus saturated model:

    ```
    pchisq(deviance(logitmod),df.residual(logitmod),lower=FALSE)
    [1] 0.7385095
    ```

    which is not in the χ^2_{21} tail, so it is statistically "close" to the saturated model and therefore a good fit.
  - Mean (null) model versus saturated model (not an important test):

    ```
    pchisq(753.32,737,lower=FALSE)
    [1] 0.014489
    ```

    which is in the χ^2_{22} tail, so it is statistically "far" from the saturated model and therefore not a good fit.
  - Specified model (with victim's race) versus mean model ($D_S - D_L$):

    ```
    pchisq(753.32-711.11,1,lower=FALSE)
    [1] 8.197975e-11
    ```

    which is in the χ^2_{22} tail, so L_S is statistically "far" from L_L.

Binomial Model Comparison

- Cautions:
  - The approximation of D to a χ^2 distributed statistic is poor for small n_i and "lumpy" distribution of n_i as well.
  - Most texts recommend $n_i \geq 5,\ \forall i$, but this is just a rule-of-thumb (and all rules-of-thumb in statistics are dumb).
  - We could also have done a Wald test on temperature:

```
            Estimate Std. Error z value Pr(>|z|)
vict.rac     -1.3629     0.2307  -5.907 3.49e-09
```

    but differences of deviances are usually more accurate than tests on a single deviance.
  - When Wald provides significant results but a deviance comparison doesn't (the Hauck-Donner effect).

Binomial Model Comparison

- Confidence interval for the j th coefficient: $\hat{\beta}_j \pm z^{\alpha/2} se(\hat{\beta}_j)$.
- Regular method:

```
summary(logitmod)$coefficients[,1]
        + qnorm(0.975) * t(c(-1,1) %o% summary(logitmod)$coefficients[,2])
(Intercept) -0.5750885 0.5750885
vict.rac    -0.4522436 0.4522436
```

- Profile likelihood version (accounts for covariance):

```
library(MASS)
confint(logitmod)
Waiting for profiling to be done...
                2.5 %     97.5 %
(Intercept) -0.145398  1.0077408
vict.rac    -1.834010 -0.9260507
```

Real Example: Model of Vote Choice 1994 American National Election Study

	Parameter Estimate	Standard Error	z-statistic	p-value
Choice Parameters				
Intercept	-1.116	0.387	-2.882	0.004
Democratic Support for Clinton	-0.015	0.008	-1.943	0.052
Republican Support for Clinton	0.030	0.011	2.701	0.007
Democratic Crime Concern	0.044	0.009	4.960	0.000
Republican Crime Concern	0.007	0.009	0.699	0.485
Democratic Gvt. Help Disadv.	0.029	0.011	2.698	0.007
Republican Gvt. Help Disadv.	-0.006	0.013	-0.438	0.661
Democratic Gvt. Spending	0.114	0.025	4.633	0.000
Republican Gvt. Spending	-0.100	0.025	-4.030	0.000
Democratic Federal Healthcare	0.031	0.008	3.670	0.000
Republican Federal Healthcare	-0.017	0.010	-1.691	0.091
Democratic Ideology Entropy	0.104	0.131	0.794	0.427
Republican Ideology Entropy	0.303	0.068	4.437	0.000
Party Identification Scale	0.368	0.028	13.158	0.000

Goodness of Fit Test: $LRT = 359.3869, p < 0.0001$ for $\chi^2_{df=19}$
Percent Correctly Classified: 78.66% (using the "naive criteria")

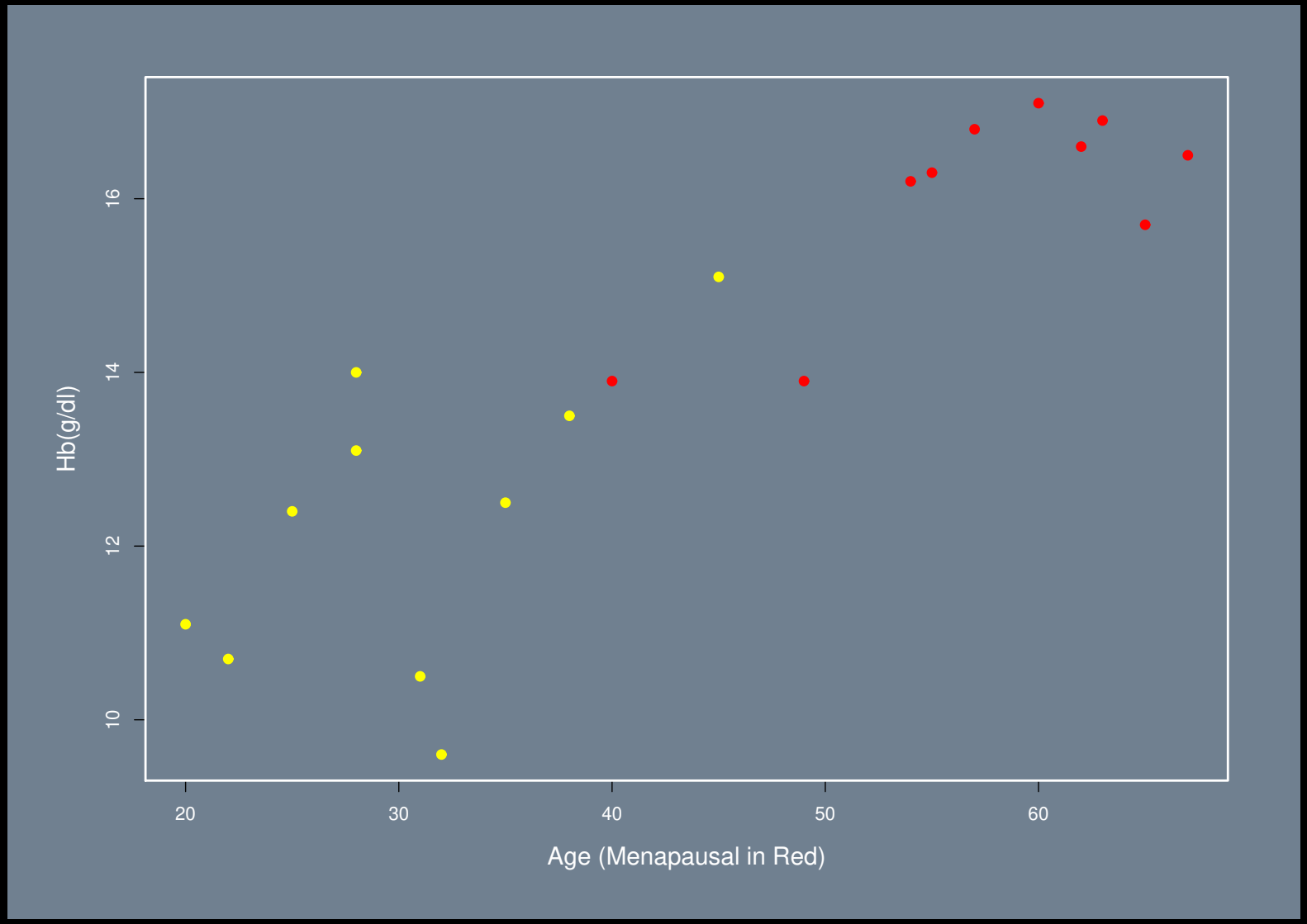
Example: Anemia

- Consider again a study of anemia in women in a given clinic in St. Louis where 20 cases are chosen at random from the full study to get the data here.
- From a blood sample we get:
  - hemoglobin level (Hb) in grams per deciliter (12–15 g/dl is normal in adult females)
  - packed cell volume (PCV) in percent of blood volume that is occupied by red blood cells (also called hematocrit, Ht or HCT, or erythrocyte volume fraction, EVF). 38% to 46% is normal in adult females.
- We also have:
  - age in years
  - menopausal (0=no, 1=yes)
- There is an obvious endogeneity problem in modeling Hb(g/dl) versus PCV(%).

Anemia Data

Subject	Hb(g/dl)	PCV(%)	Age	Menopausal
1	11.1	35	20	0
2	10.7	45	22	0
3	12.4	47	25	0
4	14.0	50	28	0
5	13.1	31	28	0
6	10.5	30	31	0
7	9.6	25	32	0
8	12.5	33	35	0
9	13.5	35	38	0
10	13.9	40	40	1
11	15.1	45	45	0
12	13.9	47	49	1
13	16.2	49	54	1
14	16.3	42	55	1
15	16.8	40	57	1
16	17.1	50	60	1
17	16.6	46	62	1
18	16.9	55	63	1
19	15.7	42	65	1
20	16.5	46	67	1

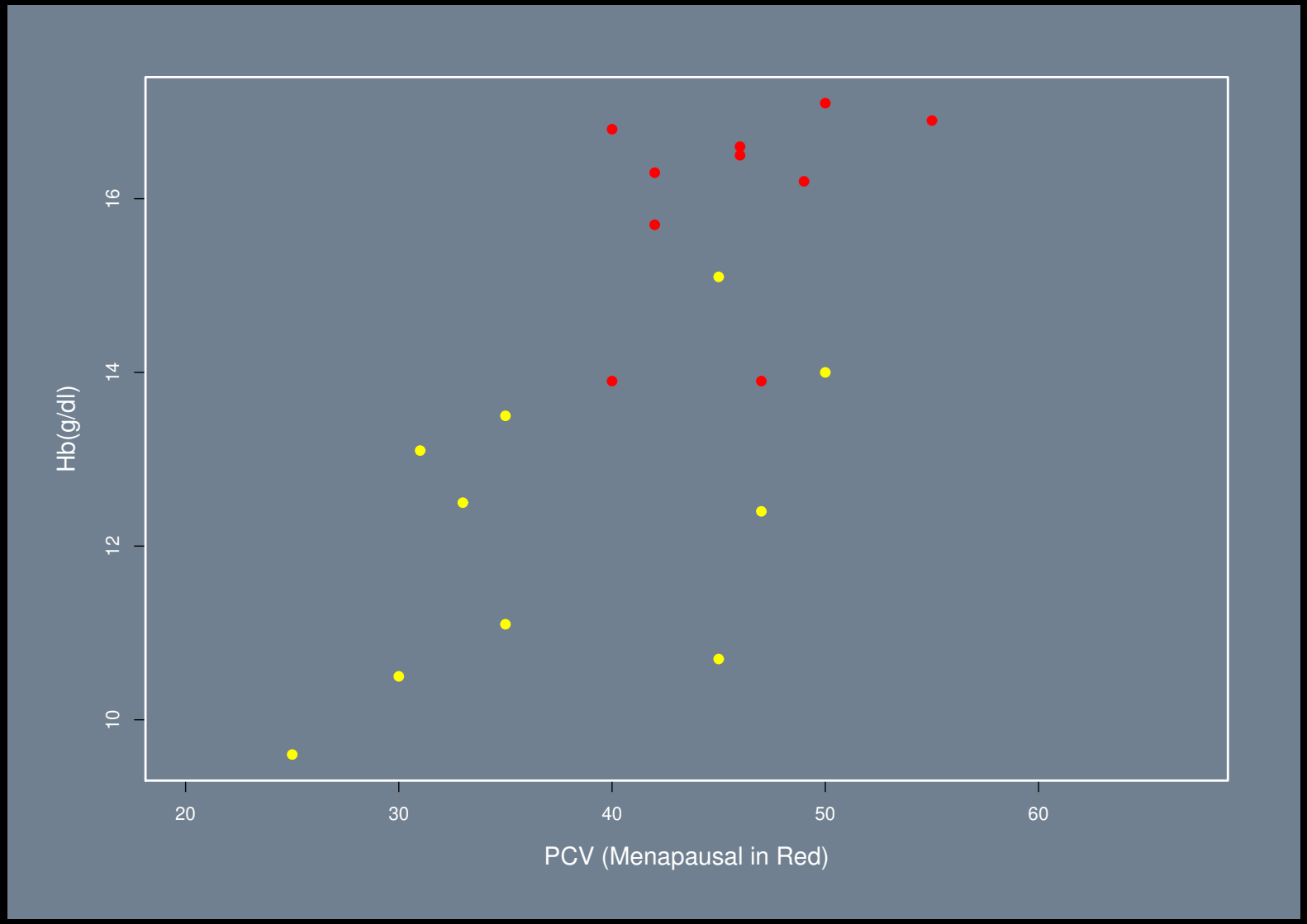
Scatterplot of the Anemia Data



Scatterplot of the Anemia Data

```
postscript("Class.PreMed.Stats/Images/anaemia1.fig.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
        col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(anaemia$Age[anaemia$Menapause==0],anaemia$Hb[anaemia$Menapause==0],
        pch=19,col="yellow",
        xlim=range(anaemia$Age),ylim=range(anaemia$Hb),
        xlab="Age (Menapausal in Red)",ylab="Hb(g/dl)")
points(anaemia$Age[anaemia$Menapause==1],anaemia$Hb[anaemia$Menapause==1],
        pch=19,col="red")
dev.off()
```

Scatterplot of the Anaemia Data



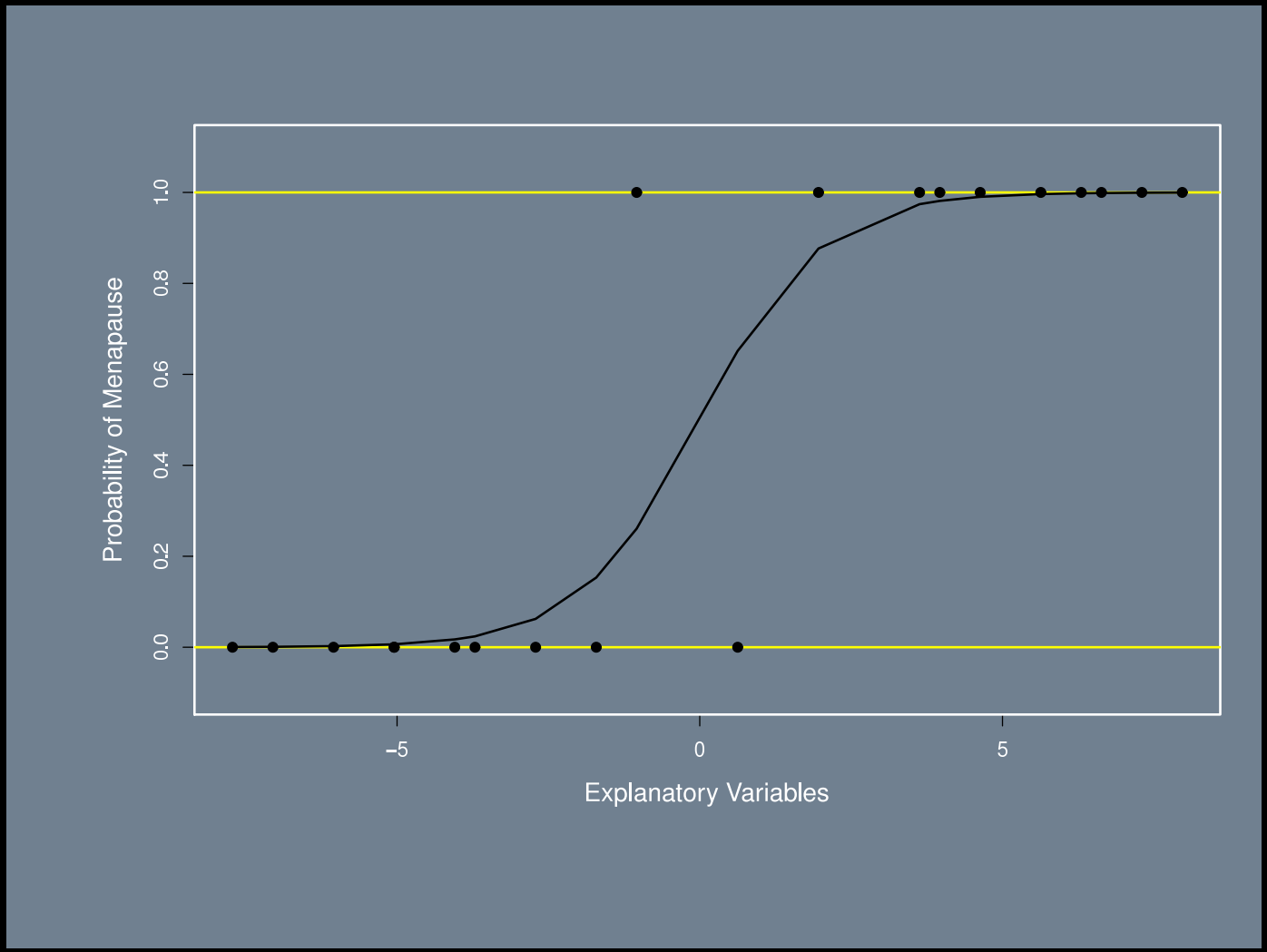
Scatterplot of the Anaemia Data

```
postscript("Class.PreMed.Stats/Images/anaemia2.fig.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
        col.sub="white",col="white",bg="slategray", cex.lab=1.3)
plot(anaemia$PCV[anaemia$Menapause==0],anaemia$Hb[anaemia$Menapause==0],
        pch=19,col="yellow",
        xlim=range(anaemia$Age),ylim=range(anaemia$Hb),
        xlab="PCV (Menapausal in Red)",ylab="Hb(g/dl)")
points(anaemia$PCV[anaemia$Menapause==1],anaemia$Hb[anaemia$Menapause==1],
        pch=19,col="red")
dev.off()
```

Logistic Regression: Anaemia Example

```
summary( glm(Menapause~Age, data=anaemia, family=binomial(link=logit)) )
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.45227  -0.13139  -0.00176   0.09818   1.63990

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -14.395      7.462   -1.93    0.054
Age            0.334      0.174    1.92    0.055

    Null deviance: 27.7259  on 19  degrees of freedom
Residual deviance:  5.7632  on 18  degrees of freedom
```

Logistic Illustration



Logistic Illustration

```
inv.logit <- function(mu)  log(mu/(1-mu))
logit <- function(Xb)  1/(1+exp(-Xb))
ana.logit <- glm(Menapause ~ Age, data=anaemia, family=binomial(link=logit))
postscript("Class.PreMed.Stats/Images/logit.anaemia1.fig.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
        col.sub="white",col="white",bg="slategray",
        cex.lab=1.3,oma=c(4,2,2,2))
xbeta <- as.matrix(cbind(rep(1,length=nrow(anaemia)),anaemia$Age))
        %*% coef(ana.logit)
plot(range(xbeta),c(-0.1,1.1),type="n",xlab="Explanatory Variables",
        ylab="Probability of Menapause")
abline(h=c(0,1),col="yellow")
x <- seq(from=min(xbeta),to=max(xbeta),length=100)
points(xbeta,anaemia$Menapause,col="black",pch=19)
lines(xbeta,logit(xbeta),col="black")
dev.off()
```

Logit Model for Survey Responses in Scotland

- These data come from the British General Election Study, Scottish Election Survey, 1997 (ICPSR Study Number 2617).
- These data contain 880 valid cases, each from an interview with a Scottish national after the election.
- Our outcome variable of interest is their party choice in the UK general election for Parliament where we collapse all non-Conservative party choices (abstention, Labour, Liberal Democrat, Scottish National, Plaid Cymru, Green, Other, Referendum) to one category, which produces 104 Conservative votes.
- For probit, $\sigma^2 = 1$ to establish the scale and provide an intuitive (standard) probit metric.

Logit Model for Survey Responses in Scotland, Explanatory Variables

- POLITICS, which asks how much interest the respondent has in political events (increasing scale: none at all, not very much, some, quite a lot, a great deal).
- READPAP, which asks about daily morning reading of the newspapers (yes=1 or no=0).
- PTYTHNK, how strong that party affiliation is for the respondent (categorical by party name).
- IDSTRNG (increasing scale: not very strong, fairly strong, very strong).
- TAXLESS asks if "it would be better if everyone paid less tax and had to pay more towards their own healthcare, schools and the like" (measured on a five point increasing Likert scale).
- DEATHPEN asks whether the UK should bring back the death penalty ((measured on a five point increasing Likert scale).
- LORDS queries whether the House of Lords should be reformed (asked as *remain as is* coded as zero and *change is needed* coded as one).
- SCENGBEN asks how economic benefits are distributed between England and Scotland with the choices: England benefits more $= -1$, neither/both lose $= 0$, Scotland benefits more $= 1$.

Logit Model for Survey Responses in Scotland, Explanatory Variables

- INDPAR asks which of the following represents the respondent's view on the role of the Scottish government in light of the new parliament: (1) Scotland should become independent, separate from the UK and the European Union, (2) Scotland should become independent, separate from the UK but part of the European Union, (3) Scotland should remain part of the UK, with its own elected parliament which has some taxation powers, (4) Scotland should remain part of the UK, with its own elected parliament which has no taxation powers, and (5) Scotland should remain part of the UK without an elected parliament.
- SCOTPREF1 asks "should there be a Scottish parliament within the UK? (yes=1, no=0).
- RSEX, the respondent's sex.
- RAGE, the respondent's age.
- RSOCCLA2, the respondents social class (7 category ascending scale).
- TENURE1, whether the respondent rents (0) or owns (1) their household.
- PRESBm a categorical variable for church affiliation, measurement of religion is collapsed down to one for the dominant historical religion of Scotland (Church of Scotland/Presbyterian) and zero otherwise and designated

Logit Model for Survey Responses in Scotland

- Run a probit model for the conservative/not-conservative outcome with these covariates:
- Results give across two slides...

```
scot.mat <- read.table("http://jeffgill.org/data/scotland.dat",sep=",",header=TRUE)
Y        <- as.numeric(scot.mat[,1])
X        <- as.matrix(scot.mat[,2:ncol(scot.mat)])
glm.out  <- glm(Y ~ X, family=binomial(link=probit))
```

Logit Model for Survey Responses in Scotland, Results (not in order)

```
summary(glm.out)

Call:
glm(formula = Y ~ X[, -1], family = binomial(link = probit))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.223  -0.287  -0.120  -0.022   3.598

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.38  on 879  degrees of freedom
Residual deviance: 338.98  on 864  degrees of freedom
AIC: 371

Number of Fisher Scoring iterations: 8
```

Logit Model for Survey Responses in Scotland, Results (not in order)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8032	0.5655	-1.42	0.1555
X[, -1]POLITICS	0.1999	0.0777	2.57	0.0101
X[, -1]READPAP	0.2626	0.1840	1.43	0.1536
X[, -1]PTYTHNK	-0.5765	0.0928	-6.21	5.3e-10
X[, -1]IDSTRNG	0.2114	0.0775	2.73	0.0064
X[, -1]TAXLESS	0.1059	0.0736	1.44	0.1501
X[, -1]DEATHPEN	0.0817	0.0578	1.41	0.1573
X[, -1]LORDS	-0.4267	0.1597	-2.67	0.0075
X[, -1]SCENGBEN	0.3279	0.1107	2.96	0.0031
X[, -1]SCOPREF1	-0.9728	0.1889	-5.15	2.6e-07
X[, -1]RSEX	0.3785	0.1712	2.21	0.0270
X[, -1]RAGE	0.0118	0.0043	2.74	0.0062
X[, -1]RSOCCLA2	-0.1218	0.0582	-2.09	0.0363
X[, -1]TENURE1	0.4634	0.1808	2.56	0.0104
X[, -1]PRESB	-0.1417	0.1675	-0.85	0.3975
X[, -1]IND.PAR	0.2500	0.1925	1.30	0.1940

Percent Predicted Correctly

```
scot.pred <- scot.out$fitted.values
scot.pred[scot.pred < 0.5] <- 0
scot.pred[scot.pred > 0.5] <- 1
table(scot.pred,scot.mat$VOTE)

scot.pred   0   1
        0 750  50
        1  26  54

sum(diag(table(scot.pred,scot.mat$VOTE)))/nrow(scot.mat)
[1] 0.91364
```

Percent Predicted Correctly

```
mean(scot.pred)
[1] 0.09091
scot.pred <- scot.out$fitted.values
scot.pred[scot.pred < mean(scot.pred)] <- 0
scot.pred[scot.pred > mean(scot.pred)] <- 1
table(scot.pred,scot.mat$VOTE)

scot.pred   0   1
        0 663  11
        1 113  93

sum(diag(table(scot.pred,scot.mat$VOTE)))/nrow(scot.mat)
[1] 0.85909
```

Tolerance Distribution (related to IRT)

- A student taking a test has aptitude $T \sim N(\mu, \sigma^2)$, which we would like to measure.
- A particular question has difficulty d_i, and the student will get it right if $d_i < T$.
- Consider d_i to be fixed, so that the probability that the student gets the question *wrong* is:

$$p_i = p(T \leq d_i) = \Phi^{-1}\left(\frac{d_i - \mu}{\sigma}\right),$$

and from rearranging:

$$\begin{aligned}\Phi(p_i) &= \frac{d_i - \mu}{\sigma} \\ &= -\mu/\sigma + d_i/\sigma \\ &= \beta_0 + \beta_1 d_i\end{aligned}$$

meaning that this is really a probit regression model with a *tolerance distribution* for T.

- Not much more here except that this shows the connection between a normal assumption and probit regression.

Tabular Analysis of Binary Outcomes

- Binary outcomes are often called *events*, meaning they either happened or didn't.
- Usually these are labeled 0 and 1, where the one denotes "happened."
- Sometimes the 1 is called a "success."
- These are only labels and switching the assignment never changes the construction or reliability of the statistical model.
- Tables of events have a very specific construction:

2×2 Contingency Table

	Experimental-Manipulation		
Outcome	Treatment	Control	Row Total
Positive	a	b	$a+b$
Negative	c	d	$c+d$
Column Total	$a+c$	$b+d$	

- Hypothesized relationships are usually down the primary diagonal of the table.

Odds and Odds Ratios

- Odds of an event is the ratio of the probability of an event *happening* to the probability of the event *not happening*:

$$Odds = \frac{p}{1-p},$$

  where p is the probability of the event.

- Odds Ratio compares the odds of an event under treatment to odds under control:

$$OR = \frac{\left(\frac{p_T}{1-P_T}\right)}{\left(\frac{P_C}{1-P_C}\right)} = \frac{\frac{\frac{a}{a+c}}{1-\frac{a}{a+c}}}{\frac{\frac{b}{b+d}}{1-\frac{b}{b+d}}} = \frac{\frac{\frac{a}{a+c}}{\frac{a+c}{a+c}-\frac{a}{a+c}}}{\frac{\frac{b}{b+d}}{\frac{b+d}{b+d}-\frac{b}{b+d}}} = \frac{\left(\frac{a}{c}\right)}{\left(\frac{b}{d}\right)} = \frac{ad}{bc}.$$

- For rare events, the odds and probability are close since $a \ll c$, so $a/c \approx a/(a+c)$, and the OR is close to the RR ($RR \approx \frac{p_T}{p_C}$).
- Nicely, the OR for failure is just the inverse of the OR for success (symmetry).

Interpreting Odds

- Some people prefer to think in terms of *odds* rather than probability:

$$o = \frac{p}{1-p} = \frac{p(y=1)}{p(y=0)} \qquad\qquad p = \frac{o}{1+o}$$

  where 0 is obviously on the support $(0 : \infty)$.

- This is essentially how logit works since:

$$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

- So if x_2 is held constant, then a one-unit change in x_1 gives a β_1 change in the log-odds of success (or a $\exp(\beta_1)$ change in the odds).

- Relatedly, if p_1 is the probability of success under condition 1 and p_2 is the probability of success under condition 2, then the relative risk is simply:

$$RR = \frac{p_1}{p_2}$$

Example: Cohort Study of Adolescents

- A random sample of size 2437, asking about cannabis and psychotic symptoms up to 4 years later(!).
- Summary table (Henquet, et al. 2005):

Cannabis Use and Psychosis

	Cannabis	*No Cannabis*	Total
Event	82	342	424
No Event	238	1775	2013
Total	320	2117	2437

- Thus the odds ratio for psychosis is:

$$OR = \frac{ad}{bc} = \frac{82 \times 1775}{342 \times 238} = 1.79.$$

- Since psychosis is a relatively rare event, this close to the relative risk:

$$RR = \frac{p_T}{p_C} = \frac{\left(\frac{82}{320}\right)}{\left(\frac{342}{2117}\right)} = 1.59.$$

Interpreting Odds, Respiratory Disease

- Respiratory Disease in < 1 year-olds:

```
library(MASS); data(babyfood)
xtabs(disease/(disease+nondisease)~sex+food,babyfood)

       Bottle   Breast    Suppl
Boy  0.168122 0.095142 0.129252
Girl 0.125000 0.066810 0.125984

mdl <- glm(cbind(disease,nondisease) ~ sex + food, family=binomial,babyfood)
summary(mdl)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.613      0.112  -14.35  < 2e-16
sexGirl       -0.313      0.141   -2.22    0.027
foodBreast    -0.669      0.153   -4.37  1.2e-05
foodSuppl     -0.173      0.206   -0.84    0.401

Null deviance: 26.37529  on 5  degrees of freedom
Residual deviance:  0.72192  on 2  degrees of freedom
```

Interpreting Odds, Respiratory Disease

- The interaction model is the saturated model for these data since $k - 1$ degrees of freedom gets consumed by 1 sex and 2 food categories.
- A deviance (not Wald) test for each of the main effects relative to the full is done with:

```
drop1(mdl,test="Chi")

Single term deletions
Model:
cbind(disease, nondisease) ~ sex + food
       Df Deviance  AIC  LRT Pr(Chi)
<none>          0.7 40.2
sex     1       5.7 43.2  5.0   0.026
food    2      20.9 56.4 20.2 4.2e-05
```

where the LRTs show strong evidence for inclusion.

Interpreting Odds, Respiratory Disease

- Coefficient interpretations:
  - `foodBreast -0.669`, so $\exp(-0.669) = 0.51222$, meaning that breast feeding reduces the odds of respiratory disease to 51% of bottle only feeding (the reference).
  - Computing a confidence interval on the log-odds scale (better coverage properties for categorical variables):

```
exp(c(-0.669-1.96*0.153,-0.669+1.96*0.153))
0.37951 0.69134
                                          or:
library(MASS); exp(confint(mdl))
Waiting for profiling to be done...
              2.5 %  97.5 %
(Intercept) 0.15920 0.24743
sexGirl     0.55362 0.96292
foodBreast  0.37819 0.68952
foodSuppl   0.55554 1.24643
```

More Measures of Goodness of Fit

- Pearson's X^2 is intended to look like the Sum of Squares Error:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

and for successes: $E_i = n_i \hat{p}_i$ $O_i = y_i$, but for failures $E_i = n_i(1 - \hat{p}_i)$ $O_i = n_i - y_i$, giving:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1 - \hat{p}_i)}$$

courtesy of some unpleasant algebra.

- We can also define an individual Pearson residual:

$$r_i^p = (y_i - n_i \hat{p}_i) / \sqrt{\text{var}(y_i)}$$

which means that:

$$X^2 = \sum_{i=1}^{n} (r_i^p)^2$$

More Measures of Goodness of Fit

- New dataset, insects dying at differing levels of insecticide concentration.

```
data(bliss)
bliss
```

```
  dead alive conc
1    2    28    0
2    8    22    1
3   15    15    2
4   23     7    3
5   27     3    4
```

```
modl <- glm(cbind(dead,alive) ~ conc, family=binomial, data=bliss)
sum(residuals(modl,type="pearson")^2)
```

```
0.36727
```

```
deviance(modl)
```

```
0.37875
```

Proportion of Deviance Explained

- Meant to be like the R^2 measure for linear models (Nagelkerke 1991).
- Definition:

$$R^2 = \frac{1 - (\hat{L}_0/\hat{L})^{2/n}}{1 - \hat{L}_0} = \frac{1 - \exp((D - D_{null})/n)}{1 - \exp(D_{null}/n)}$$

  where $\hat{L}_0$ is the maximized likelihood under the null.
- Implementation:

```
(1-exp((modl$dev-modl$null)/150))/(1-exp(-modl$null/150))
```

```
0.99532
```

Prediction and Effective Doses

- We want to predict an outcome, in this case the probability of success, for levels of the explanatory variables: $g^{-1}(\hat{\eta}) = g^{-1}(x_0\hat{\beta})$.
- Returning to the insect data, predict the response at a dose of 2.5:

```
modl <- glm(cbind(dead,alive) ~ conc, family=binomial,data=bliss)
lmodsum <- summary(modl)
x0 <- c(1,2.5)
( eta0 <- sum(x0*coef(modl)) )
```

```
0.58095
```

```
ilogit(eta0)
```

```
0.64129
```

meaning that 64% are predicted die at this level.

Prediction and Effective Doses

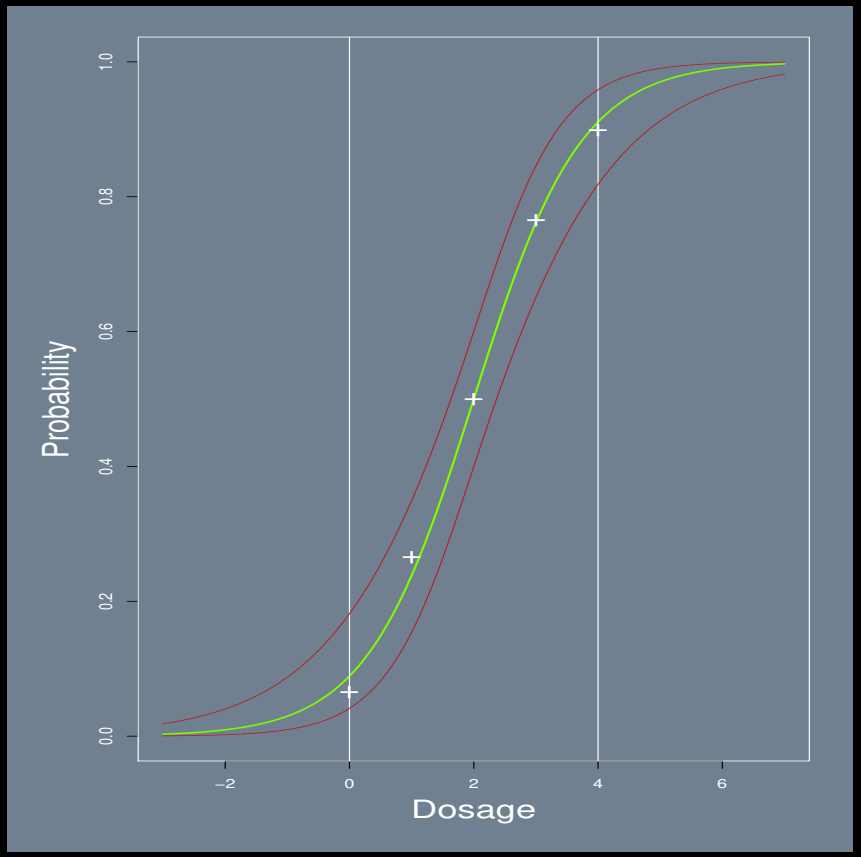
- We also want a measure of uncertainty around this prediction in the form of a 95% CI:

```
( cm <- lmodsum$cov.unscaled )

            (Intercept)      conc
(Intercept)    0.174630 -0.065823
conc          -0.065823  0.032912

( se <- sqrt( t(x0) %*% cm %*% x0) )

0.2263

ilogit(c(eta0-1.96*se,eta0+1.96*se))

0.53430 0.7358
```

Confidence Bands for Effective Doses

- Here is a better tool...

```
ruler <- seq(-3,7,length=200)
predicts <- predict(modl,newdata=data.frame(conc=ruler),se=TRUE)

ci <- cbind( ilogit(predicts$fit-qnorm(0.975)*predicts$se.fit),
             ilogit(predicts$fit+qnorm(0.975)*predicts$se.fit) )

postscript("Class.MLE/faraway.ch2.fig5.ps")
par(col.axis="white",col.lab="white",col.sub="white",
    col="white", bg="slategray",cex.lab=2,mar=c(6,6,2,2))
plot(ruler,ilogit(predicts$fit),xlab="Dosage",ylab="Probability",
     type="l",lwd=2, col="lawngreen")
abline(v=c(0,4),col="white")
lines(ruler,ci[,1],col="firebrick")
lines(ruler,ci[,2],col="firebrick")
points(0:4,bliss$dead/(bliss$alive+bliss$dead),pch="+",cex=2)
dev.off()
```



- Where the indicated points are from the original data values (5 concentrations).

LD50 For Political Scientists

- Sometimes we would like to go backwards: what levels of x product a certain probability?
- One common question: what is the effective dose required to get a prediction of 50% killed (LD50)?
- For a logit link this is just:

$$\begin{aligned} p(y=1|x) = \frac{1}{2} &= [1+\exp(-\beta_0 - \beta_1 x)]^{-1} \\ 2 &= 1+\exp(-\beta_0 - \beta_1 x) \\ 0 &= -\beta_0 - \beta_1 x \\ \widehat{\text{LD50}} &= -\hat{\beta}_0/\hat{\beta}_1 \end{aligned}$$

- Returning to the Bliss data (typo here in my copy of the book):

```
(ld50 <- -modl$coef[1]/modl$coef[2])
(Intercept)
          2
```

LD50 For Political Scientists

- The variance of a function of a random variable $\hat{\theta}$ can often be obtained by the *delta method*:

$$\mathrm{Var} g(\hat{\theta}) \approxeq g'(\hat{\theta}) \mathrm{Var}(\hat{\theta}) g'(\hat{\theta})$$

- Here:

$$\mathrm{ld50} = \hat{\theta}.$$

- So:

$$\frac{d}{d\hat{\beta}_1} g(\hat{\theta}) = \frac{d}{d\hat{\beta}_1}(-\hat{\beta}_0/\hat{\beta}_1) = -1/\hat{\beta}_2$$

  and:

$$\frac{d}{d\hat{\beta}_2} g(\hat{\theta}) = \frac{d}{d\hat{\beta}_2}(-\hat{\beta}_0/\hat{\beta}_1) = \hat{\beta}_0/\hat{\beta}_2$$

- Executing:

```
dr <- c(-1/modl$coef[2],modl$coef[1]/modl$coef[2]^2)
( sqrt(dr %*% lmodsum$cov.unscaled %*% dr)[,] )
[1] 0.17844
```

Overdispersion in Dichotomous Choice Models

- If we meet the described assumptions, then the two times the residual (summed) deviance is approximately χ^2 with $n - p$ degrees of freedom.
- However, sometimes we are in the tail of this distribution not because we have chosen the wrong explanatory variables, but because of:
  - outliers,
  - sparse data,
  - overdispersion: $\text{Var}(Y) \gg mp(1-p)$, where m is the size of the binomial trial group (often denoted n_i when there are differences).
- Underdispersion is rare.
- Typical causes of overdispersion:
  - variation in p across binomial trials (violates iid assumption),
  - unmeasured clustering in the data,
  - dependence between trials (which can come from clustering).
- One diagnostic: plot $\hat{\mu}$ versus $(y - \hat{\mu})^2$.

Overdispersion in Dichotomous Choice Models

- In regular models $\sigma^2 = \phi = 1$, and `R` even reminds us of this assumption.
- A test for $\phi > 1$ can be constructed by modifying the Pearson statistic according to:

$$\hat{\sigma}^2 = X^2/(n-k) = \frac{1}{n-k}\sum_{i=1}^{n}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}.$$

- Then the variance of the coefficient variance is adjusted with:

$$\widehat{\text{Var}}\hat{\boldsymbol{\beta}} = \hat{\sigma}^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1},$$

  where $\mathbf{W} = \text{diag}(mp(1-p))$ (the coefficient estimate is still unbiased).
- This added uncertainty replaces the chi-square model comparison with an approximate F-test:

$$F \approx \frac{D_{small} - D_{large}}{\widehat{\text{Var}}\hat{\boldsymbol{\beta}}(df_{small} - df_{large})}.$$

Overdispersion Example: boxes of trout eggs buried in 5 places

```
data(troutegg)
ftable(xtabs(cbind(survive,total) ~ location+period, troutegg))
                survive total
location period
1        4           89    94
         7           94    98
         8           77    86
         11         141   155
2        4          106   108
         7           91   106
         8           87    96
         11         104   122
3        4          119   123
         7          100   130
         8           88   119
         11          91   125
4        4          104   104
         7           80    97
         8           67    99
         11         111   132
5        4           49    93
         7           11   113
         8           18    88
         11           0   138
```

Overdispersion Example: boxes of trout eggs buried in 5 places

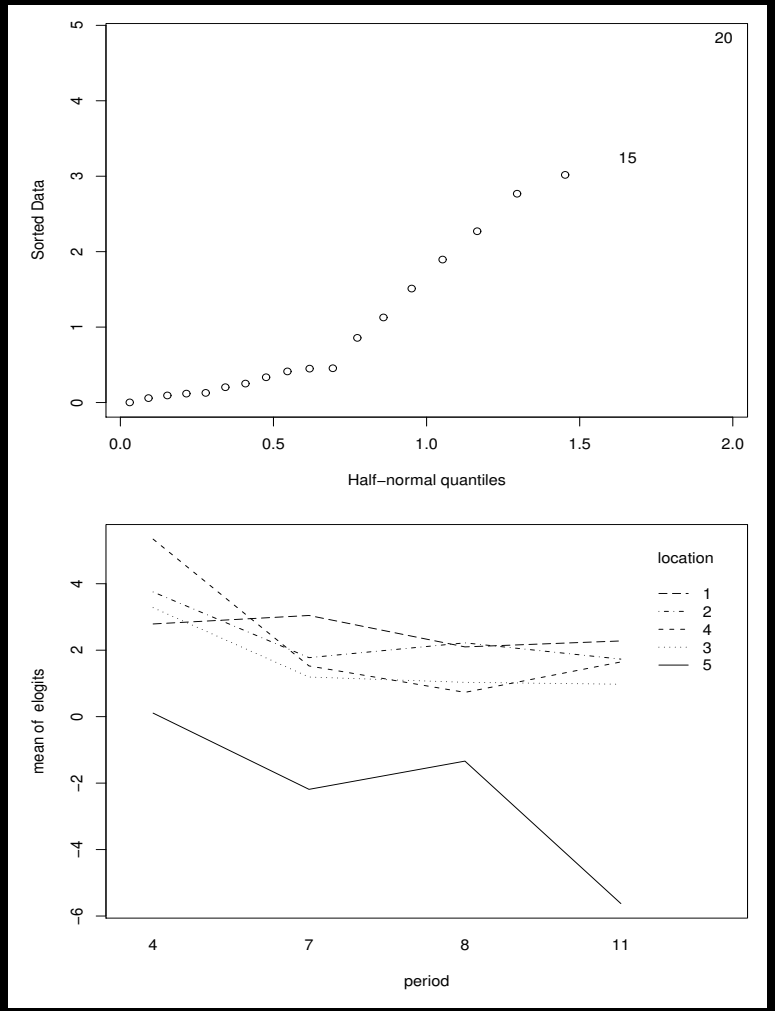
```
bmod <- glm(cbind(survive,total-survive) ~ location+period, family=binomial,troutegg)
bmod

Coefficients:
(Intercept)    location2    location3    location4    location5      period7      period8
      4.636       -0.417       -1.242       -0.951       -4.614       -2.170       -2.326
   period11
     -2.450

Degrees of Freedom: 19 Total (i.e. Null);  12 Residual
Null Deviance:            1020
Residual Deviance: 64.5         AIC: 157
```

- Since 64.5 is way into the tail of a χ^2_{12} distribution, we know to be worried.

Overdispersion Example: boxes of trout eggs buried in 5 places

- Sparseness? No, `min(troutegg$total)` returns 86.
- Outliers? No, `halfnorm(residuals(bmod))` shows no problems.
- Specification error? No, an interaction plot of the *empirical logits* $(\log(y+0.5)-\log(m-y+0.5))$ shows no major relationships.

```
elogits <- log((troutegg$survive+0.5)/
(troutegg$total-troutegg$survive+0.5))

with(troutegg,interaction.plot(period,
location,elogits))
```



Overdispersion Example: boxes of trout eggs buried in 5 places

- Estimating $\hat{\sigma}^2$ shows it to be much larger than 1:

```
(sigma2 <- sum(residuals(bmod,type="pearson")^2)/12)
5.3303
```

- Now do an F-test of the predictors using the new $\hat{\sigma}^2$:

```
drop1(bmod,test="Chi")
Single term deletions
scale:  5.3303

         Df Deviance AIC F value   Pr(F)
<none>             64 157
location  4      914 308    39.5 8.1e-07
period    3      229 182    10.2  0.0013
```

Overdispersion Example: boxes of trout eggs buried in 5 places

- And summarize the new results using the new value of $\hat{\sigma}^2$:

```
summary(bmod,dispersion=sigma2)

             Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.636      0.649    7.14  9.5e-13
location2      -0.417      0.568   -0.73    0.463
location3      -1.242      0.507   -2.45    0.014
location4      -0.951      0.528   -1.80    0.072
location5      -4.614      0.578   -7.99  1.4e-15
period7        -2.170      0.550   -3.94  8.1e-05
period8        -2.326      0.561   -4.15  3.4e-05
period11       -2.450      0.540   -4.53  5.8e-06

(Dispersion parameter for binomial family taken to be 5.3303)
```

Overdispersion Example: Teenage Conformance and Survival in a Social Group

```
data(tg)
ftable(xtabs(cbind(survive,total) ~ location+period, tg))
location period survive total   period survive total
1        4       89       94     7      94      98
         8       77       86    11     141     155
2        4      106      108     7      91     106
         8       87       96    11     104     122
3        4      119      123     7     100     130
         8       88      119    11      91     125
4        4      104      104     7      80      97
         8       67       99    11     111     132
5        4       49       93     7      11     113
         8       18       88    11       0     138
```

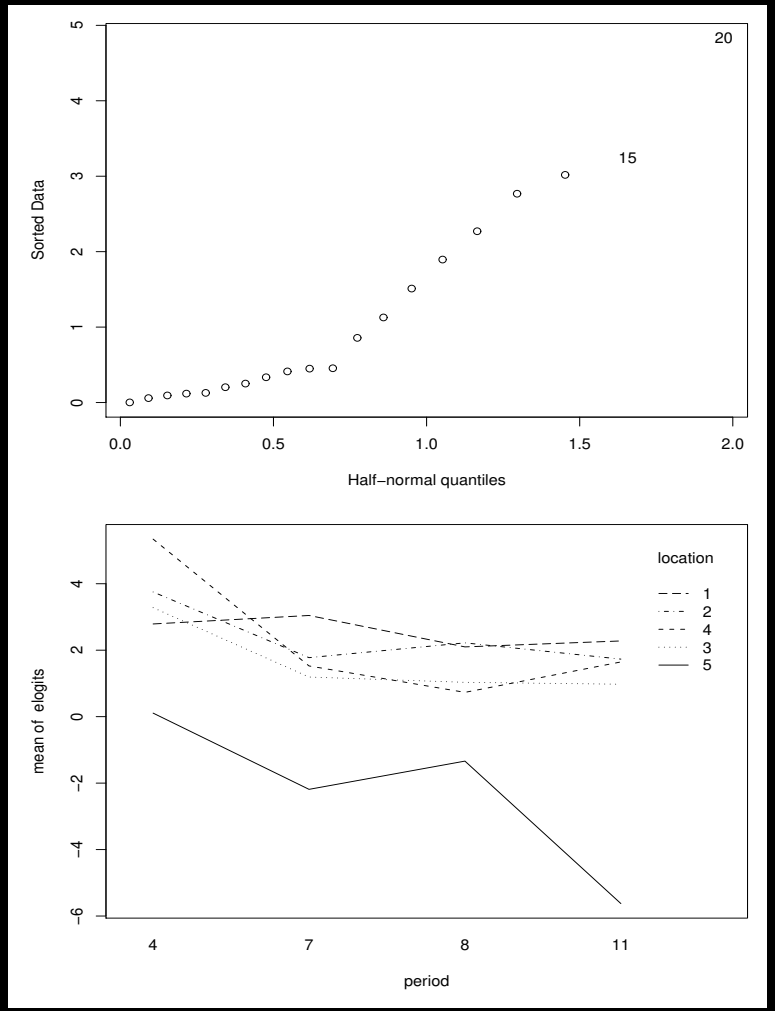
Overdispersion Example: Teenage Conformance and Survival in a Social Group

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8305  -0.3650  -0.0303   0.6191   3.2434

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.6358     0.2813  16.479  < 2e-16
location2    -0.4168     0.2461  -1.694   0.0903
location3    -1.2421     0.2194  -5.660 1.51e-08
location4    -0.9509     0.2288  -4.157 3.23e-05
location5    -4.6138     0.2502 -18.439  < 2e-16
period7      -2.1702     0.2384  -9.103  < 2e-16
period8      -2.3256     0.2429  -9.573  < 2e-16
period11     -2.4500     0.2341 -10.466  < 2e-16
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1021.469  on 19  degrees of freedom
Residual deviance:   64.495  on 12  degrees of freedom
AIC: 157.03
```

Overdispersion Example: Teenage Conformance and Survival in a Social Group

- Sparseness? No, `min(tg$total)` returns 86.
- Outliers? No, `halfnorm(residuals(teen.out))` shows no problems.
- Specification error? No, an interaction plot of the *empirical logits* $(\log(y+0.5)-\log(m-y+0.5))$ shows no major relationships.

```
elogits <- log((tg$survive+0.5)/
(tg$total-tg$survive+0.5))
with(tg,interaction.plot(period,
location,elogits))
```



Overdispersion Example: Teenage Conformance and Survival in a Social Group

- Estimating $\hat{\sigma}^2$ shows it to be much larger than 1:

```
(sigma2 <- sum(residuals(teen.out,type="pearson")^2)/12)
5.3303
```

- And summarize the new results using the new value of $\hat{\sigma}^2$:

```
summary(teen.out,dispersion=sigma2)

             Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.636      0.649    7.14  9.5e-13
location2      -0.417      0.568   -0.73    0.463
location3      -1.242      0.507   -2.45    0.014
location4      -0.951      0.528   -1.80    0.072
location5      -4.614      0.578   -7.99  1.4e-15
period7        -2.170      0.550   -3.94  8.1e-05
period8        -2.326      0.561   -4.15  3.4e-05
period11       -2.450      0.540   -4.53  5.8e-06

(Dispersion parameter for binomial family taken to be 5.3303)
```

Overdispersion Example: Teenage Conformance and Survival in a Social Group

- Another strategy for dealing with overdispersion in dichotomous outcome models is using Quasi-likelihood.
- This relaxes the form of the relevant likelihood function such that it relies on moments rather than full contributions.

```
teen.out.q <- glm(cbind(survive,total-survive) ~ location+period,
    family = quasibinomial(logit),data=tg)
```

- Note the modification to `family`.
- The results are the same, subject to algorithmic rounding from:

```
summary(teen.out.q)
```

Overdispersion Example: Teenage Conformance and Survival in a Social Group

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8305  -0.3650  -0.0303   0.6191   3.2434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.6358     0.6495   7.138 1.18e-05
location2    -0.4168     0.5682  -0.734 0.477315
location3    -1.2421     0.5066  -2.452 0.030501
location4    -0.9509     0.5281  -1.800 0.096970
location5    -4.6138     0.5777  -7.987 3.82e-06
period7      -2.1702     0.5504  -3.943 0.001953
period8      -2.3256     0.5609  -4.146 0.001356
period11     -2.4500     0.5405  -4.533 0.000686
(Dispersion parameter for quasibinomial family taken to be 5.330358)
    Null deviance: 1021.469  on 19  degrees of freedom
Residual deviance:   64.495  on 12  degrees of freedom
AIC: NA                        Number of Fisher Scoring iterations: 5
```

Beta Regression

- Beta regression is used for outcome variables that are $[0:1]$ rather than $\{0,1\}$.
- The link function is the beta distribution:
  - PDF: $\mathcal{BE}(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, 0 < \alpha, \beta$
  - $E[X] = \frac{\alpha}{\alpha+\beta}$.
  - $\text{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- It can also be used in a general for other bounded variables by rescaling.

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{(y-a)^{\alpha-1}(b-y)^{\beta-1}}{(b-a)^{\alpha+\beta-1}},$$

  where: $a < y < b$, and $\alpha, \beta > 0$.
- So the general form of the beta density operates on a support bounded by $[a{:}b]$, which are user-specified limits greater than zero, and it easily reduces to the standard form with a change of variable calculation, $X = \frac{Y-a}{b-a}$, so that $0 < x < 1$, but α and β are unchanged.
- It is easy to go back and forth between X and Y since the reverse change is also a linear form: $Y = (b-a)X + a$.

Beta Regression

- To make this distribution easy to specify as a GLM do the following transformations:

$$\mu = \frac{\alpha}{\alpha + \beta} \qquad \text{and} \qquad \phi = \alpha + \beta$$

- This means that:

$$\mathcal{E}[Y] = \mu \qquad \text{and} \qquad \text{Var}[Y] = \frac{\mu(1-\mu)}{1+\phi}$$

  which means that the link function can be logit, probit, or cloglog.

- Faraway uses an example from Simon Wood's nonparametric book:

```
data(mammalsleep, package="faraway"); library(mgcv)
( mammalsleep$pdr <- with(mammalsleep, dream/sleep) )
```

Beta Regression

```
 [1]         NA 0.24096386         NA         NA 0.46153846 0.07142857
 [7] 0.19796954 0.16129032 0.24827586 0.14432990 0.12000000 0.17948718
[13] 0.26213592         NA 0.25000000 0.00000000 0.38317757 0.11214953
[19] 0.21311475 0.33701657         NA 0.13157895 0.23611111         NA
[25] 0.24193548         NA 0.24637681 0.09756098 0.27586207         NA
[31]         NA 0.15384615 0.10050251 0.23750000 0.22641509 0.25000000
[37] 0.09848485 0.15625000 0.28865979 0.17816092         NA 0.10588235
[43] 0.08256881 0.13138686 0.22619048 0.10714286         NA 0.19696970
[49] 0.24489796 0.12500000 0.13636364 0.09259259         NA 0.15789474
[55]         NA 0.21359223 0.17293233 0.09259259 0.16455696 0.05825243
[61] 0.34020619         NA
```

Beta Regression

```
modb <- gam(pdr ~ log(body) + log(lifespan), family=betar, mammalsleep,
    na.action=na.omit) ### UGH! ###
summary(modb)

Family: Beta regression(8.927)
Link function: logit

Formula:
pdr ~ log(body) + log(lifespan)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.37795    0.37322   1.013    0.311
log(body)     0.26796    0.05513   4.860 1.17e-06
log(lifespan) -0.92266    0.16585  -5.563 2.65e-08


R-sq.(adj) =  -0.178   Deviance explained = 69.6%
-REML = -47.801  Scale est. = 1         n = 45
```