Winter Institute in Data Science and Big Data

Generalized Linear Models

JEFF GILL
Distinguished Professor
Departments of Government, and Mathematics & Statistics
Center for Data Science
American University

Overview

- Previously we have seen “closed form” estimators for quantities of interest, such as $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

- Moving to nonlinear models for categorical and limited support outcomes requires a more flexible process.

- Maximum Likelihood Estimation (Fisher 1922, 1925) is a classic method that finds the value of the estimator “most likely to have generated the observed data, assuming the model specification is correct.”

- There is both an abstract idea to absorb and a mechanical process to master.

More Background

- Suppose we care about some political phenomenon $\mathbf{Y}$, and determine that it has distribution $f()$.
- The stochastic component is:
$$\mathbf{Y} \sim f(\mu, \tau).$$
- The systematic component is:
$$\mu = g^{-1}(\mathbf{X}\boldsymbol{\beta})$$
- This setup is very general and covers all of the nonlinear regression models we will cover.
- You have seen the linear model in a similar form before:
$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$
$$\boldsymbol{\epsilon}_i = N(0, \sigma^2).$$

More Background

- But now we are going to think of it in this more general way, for example:

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \mathbf{X}_i\boldsymbol{\beta}. \end{aligned}$$

- An even more general way specifies a link function:

$$\begin{aligned} g(\mathbf{Y}_i) &= \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \\ \mathbf{Y}_i &= g^{-1}(\mathbf{X}_i\boldsymbol{\beta}) + \boldsymbol{\epsilon}_i \\ \mathbf{Y}_i &= g^{-1}(\mu_i) + \boldsymbol{\epsilon}_i \end{aligned}$$

- We typically write this in expected value terms:

$$\mathbb{E}[Y|\mathbf{X}, \boldsymbol{\beta}] = \boldsymbol{\mu}$$

The Likelihood Function

- Assume that:
$$x_1, x_1, \ldots, x_n \sim \text{ iid } f(x|\theta),$$
where θ is a parameter that is critical to the data generation process (DGP).

- Since these values are independent, the joint distribution of the observed data is just the product of their individual PDF/PMFs:
$$f(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

- But once we observe the data $\mathbf{x}$ is fixed.

- It is θ that is unknown, so rewrite the joint distribution function according to:
$$f(\mathbf{x}|\theta) = L(\theta|\mathbf{x}).$$

- Note that this is a purely *notational* change, nothing is different mathematically.

The Likelihood Function

- Fisher (1922) justifies this because at this point we know $\mathbf{x}$.

$$f(\mathbf{x}|\theta) \longrightarrow L(\theta|\mathbf{x}).$$

- A semi-Bayesian justification works as follows, we want to perform:

$$p(\mathbf{x}|\theta) = \frac{p(\mathbf{x})}{p(\theta)}p(\theta|\mathbf{x}).$$

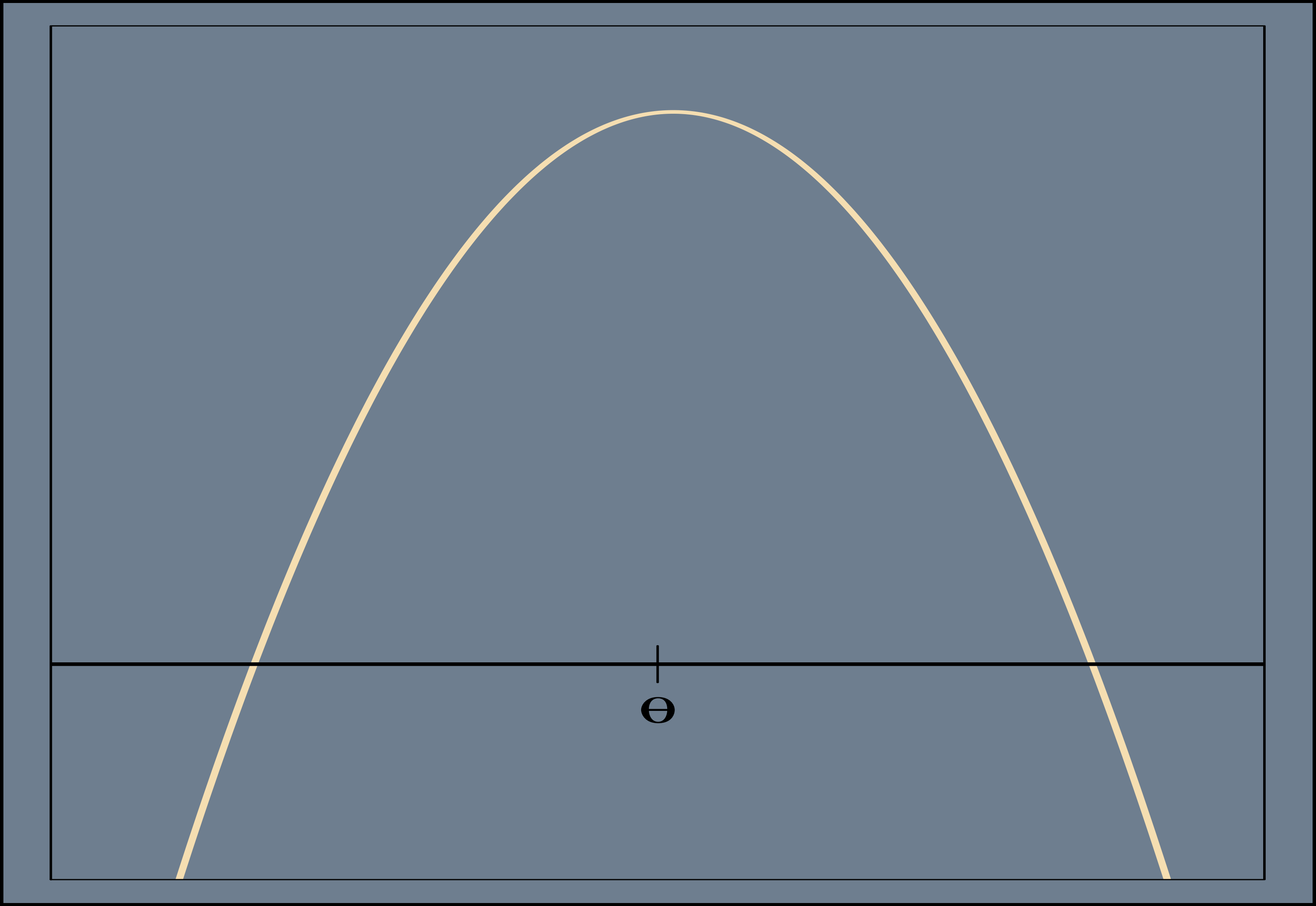
  but $p(\mathbf{x}) = 1$ since the data has already occurred, and if we put a finite uniform prior on θ over its finite allowable range (support), then $p(\theta) = 1$.

- Therefore:

$$p(\mathbf{x}|\theta) = \frac{1}{1}p(\theta|\mathbf{x}) = p(\theta|\mathbf{x}).$$

- The only caveat here is the finiteness of the support of θ.

Generic Likelihood Function Illustration



Poisson MLE

- Start with the Poisson PMF for x_i:

$$p(X = x_i) = f(x_i|\theta) = \frac{e^{-\theta}\theta^{x_i}}{x_i!},$$

which requires the assumptions: non-concurrence of arrivals, the number of arrivals is proportion to the time of study, this rate is constant over the time, and there is no serial correlation of arrivals.

- The likelihood function is created from the joint distribution:

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-\theta}\theta^{x_1}}{x_1!}\frac{e^{-\theta}\theta^{x_2}}{x_2!}\cdots\frac{e^{-\theta}\theta^{x_n}}{x_n!} = e^{-n\theta}\theta^{\sum x_i}\left(\prod_{i=1}^{n} x_i!\right)^{-1}.$$

- Suppose we have the data: $\mathbf{x} = \{5, 1, 1, 1, 0, 0, 3, 2, 3, 4\}$, then the likelihood function is:

$$L(\theta|\mathbf{x}) = \frac{e^{-10\theta}\theta^{20}}{207360},$$

which is the probability of observing *this* exact sample.

Poisson MLE

- It is often easier to deal the logarithm of the MLE:

$$\log L(\theta|\mathbf{x}) = \ell(\theta|\mathbf{x}) = \log\left(e^{-n\theta}\theta^{\sum x_i}\left(\prod_{i=1}^{n} x_i!\right)^{-1}\right) = -n\theta + \sum_{i=1}^{n} x_i \log(\theta) - \log\left(\prod_{i=1}^{n} x_i!\right).$$

- For our small example this is:

$$\ell(\theta|\mathbf{x}) = -10\theta + 20\log(\theta) - \underbrace{\log(207360)}_{12.242}.$$

- Importantly, for the family of functions that we will use the likelihood function and the log-likelihood function have the same mode (maximum of the function) for θ.

- They are both guaranteed to be concave to the x-axis.

Obtaining the Poisson MLE

- Freshman calculus: where is the maximum of the function? At the point when first derivative of the function equals zero.

- So take the first derivative, set it equal to zero, and solve.

- $\frac{d}{d\theta}\ell(\theta|\mathbf{x}) \equiv 0$ is called the likelihood equation.

- For the example:

$$\ell(\theta|\mathbf{x}) = -10\theta + 20\log(\theta) - \underbrace{\log(207360)}_{12.242}.$$

Taking the derivative, and setting equal to zero:

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = -10 + 20\theta^{-1} \equiv 0,$$

so that $20\theta^{-1} = 10$, and therefore $\hat{\theta} = 2$ (note the hat).

Obtaining the Poisson MLE

- More generally:

$$\begin{aligned}\ell(\theta|\mathbf{x}) &= -n\theta + \sum_{i=1}^{n}\log(\theta) - \log\left(\prod_{i=1}^{n} x_i!\right)\\ \frac{d}{d\theta}\ell(\theta|\mathbf{x}) &= -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i \equiv 0\\ \hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{\mathbf{x}}\end{aligned}$$

- It is *not* true that the MLE is always the data mean.

General Steps

- This process is import to us:

  1. Identify the PMF or PDF.
  2. Create the likelihood function from the joint distribution of the observed data.
  3. Change to the log for convenience.
  4. Take the first derivative with respect to the parameter of interest.
  5. Set equal to zero.
  6. Solve for the MLE.

Poisson Example in R

```
# POISSON LIKELIHOOD AND LOG-LIKELIHOOD FUNCTION
llhfunc<-function(X,p,do.log=TRUE) {
        d <- rep(X,length(p))
        q.vec <- rep(length(y.vals),length(p)); p.vec <- rep(p,q.vec)
        print(q.vec)
        d.mat <- matrix(dpois(d,p.vec,log=do.log),ncol=length(p))
        print(d.mat)
        if (do.log==TRUE) apply(d.mat,2,sum)
        else apply(d.mat,2,prod)
}
```

Poisson Example in R

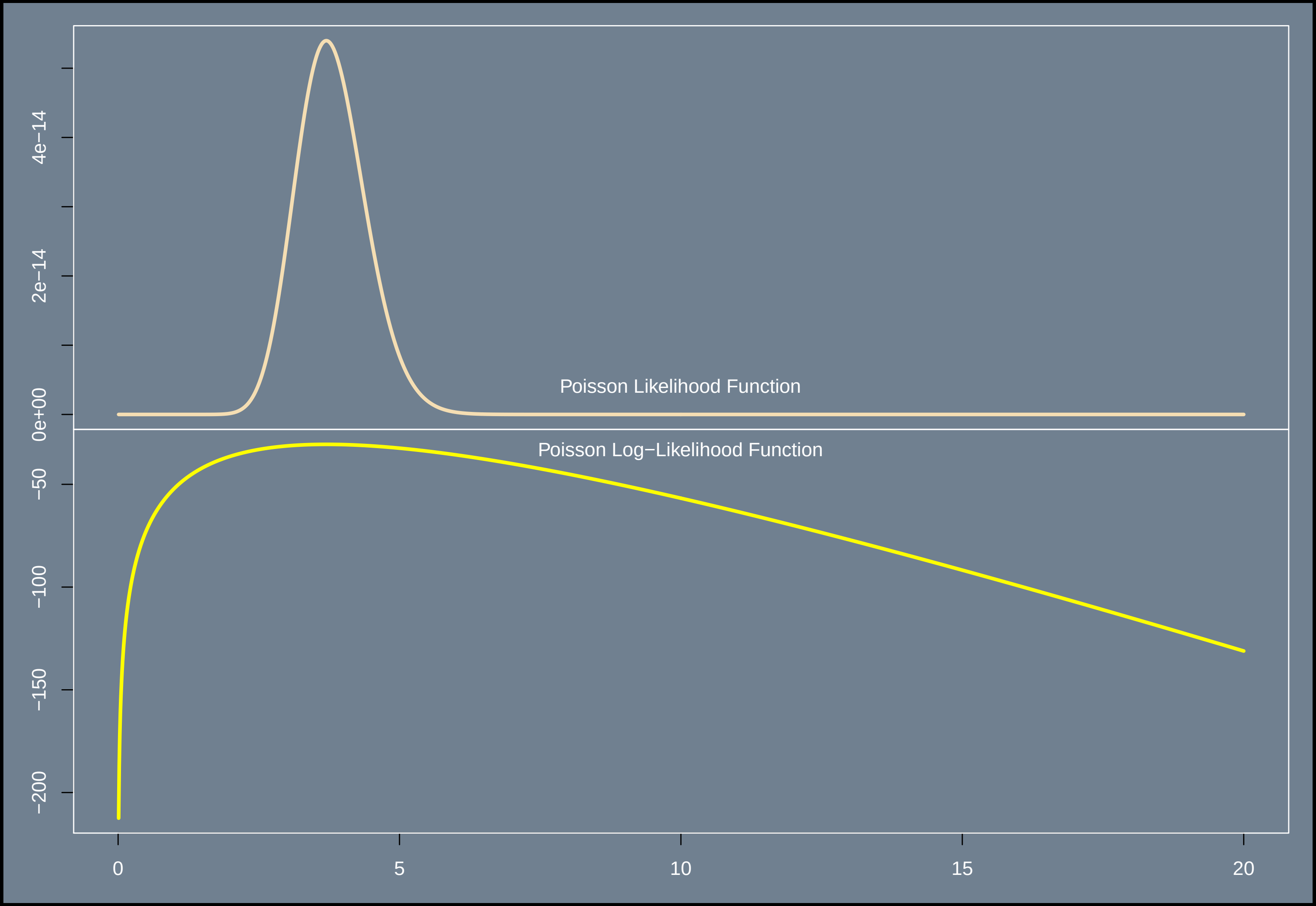
```
# HERE'S A TEST FUNCTION
y.vals<-c(1,3,1,5,2,6,8,11,0,0)
llhfunc(y.vals,c(4,30))
[1] 10 10
         [,1]    [,2]
 [1,] -2.6137 -26.599
 [2,] -1.6329 -21.588
 [3,] -2.6137 -26.599
 [4,] -1.8560 -17.782
 [5,] -1.9206 -23.891
 [6,] -2.2615 -16.172
 [7,] -3.5142 -13.395
 [8,] -6.2531 -10.089
 [9,] -4.0000 -30.000
[10,] -4.0000 -30.000
[1]  -30.666 -216.114
```

Poisson Example in R

```
# USE THE R CORE FUNCTION FOR OPTIMIZING, par=STARTING VALUES,
# control=list(fnscale=-1) INDICATES A MAXIMIZATION, bfgs=QUASI-NEWTON ALGORITHM
mle <- optim(par=1,fn=llhfunc,X=y.vals,control=list(fnscale=-1),method="BFGS")

# MAKE A PRETTY GRAPH OF THE LOG AND NON-LOG VERSIONS
ruler <- seq(from=.01, to=20, by= .01)
poison.ll <- llhfunc(y.vals,ruler)
poison.l <- llhfunc(y.vals,ruler,do.log=FALSE)

par(oma=c(3,3,1,1),mar=c(0,0,0,0),mfrow=c(2,1))
plot(ruler,poison.l,col="wheat",type="l",xaxt="n",lwd=3)
text(mean(ruler),mean(poison.l),"Poisson Likelihood Function")
plot(ruler,poison.ll,col="yellow",type="l",lwd=3)
text(mean(ruler),mean(poison.ll)/2,"Poisson Log-Likelihood Function")
```


Poisson Likelihood Function
Poisson Log−Likelihood Function
0e+00
2e−14
4e−14
−50
−100
−150
−200
0
5
10
15
20

Measuring the Uncertainty of the MLE

- The first derivative measures slope and the second derivative measures “curvature” of the function at a given point.

- The more peaked the function is at the MLE, the more “certain” the data are about this estimator.

- The square root of the negative inverse of the expected value of the second derivative is the SE of the MLE.

- In multivariate terms for vector $\boldsymbol{\theta}$, we take the negative inverse of the expected *Hessian*.

- Poisson example:

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

$$\frac{d^2}{d\theta^2}\ell(\theta|\mathbf{x}) = \frac{d}{d\theta}\left(\frac{d}{d\theta}\ell(\theta|\mathbf{x})\right) = -\theta^{-2}\sum_{i=1}^{n} x_i$$

- The expected value (estimate) of θ is the MLE, so:

$$SE(\hat{\theta}) = \frac{\hat{\theta}^2}{\sum_{i=1}^{n} x_i} = \frac{\bar{\mathbf{x}}^2}{n\bar{\mathbf{x}}} = \frac{\bar{\mathbf{x}}}{n}.$$

Multivariable MLE

- Now $\boldsymbol{\theta}$ is a vector of coefficients to be estimated (eg. regression).
- The Score Function is:

  $$\dot{\ell}(\boldsymbol{\theta}|\mathbf{x}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta}|\mathbf{x})$$

  which we use to get the MLE $\hat{\boldsymbol{\theta}}$.
- The Hessian Matrix is:

  $$\mathbf{H} = \frac{\partial^2\ell(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$$

  which we use to get the SE of the MLE.
- The information matrix is:

  $$\mathbf{I} = -\mathbb{E}(f)\left[\left.\frac{\partial^2\ell(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\hat{\boldsymbol{\theta}}}\right] \equiv \mathbb{E}(f)\left[\left.\frac{\partial\ell(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}}\frac{\partial\ell(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}'}\right|_{\hat{\boldsymbol{\theta}}}\right]$$

  where the equivalence of these forms is called the *information equality*.
- The variance-covariance of $\hat{\boldsymbol{\theta}}$ is produced by:

  $$\boldsymbol{\Sigma} = \mathbf{I}^{-1}$$

Properties of the MLE (Birnbaum 1962)

- Consistency:

$$\text{plim}\hat{\theta} = \theta.$$

- Asymptotic Normality:

$$\hat{\theta} \underset{a}{\sim} N\left(\theta, I(\theta)^{-1}\right) \quad \text{where } I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\right].$$

- Asymptotic Efficiency: no other estimator has lower variance, the variance of the MLE meets the Crámer-Rao Lower Bound.

- Invariance To Reparameterization:

$$\gamma = c(\theta) \implies \hat{\gamma} = c(\hat{\theta}).$$

Dichotomous 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.

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:

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

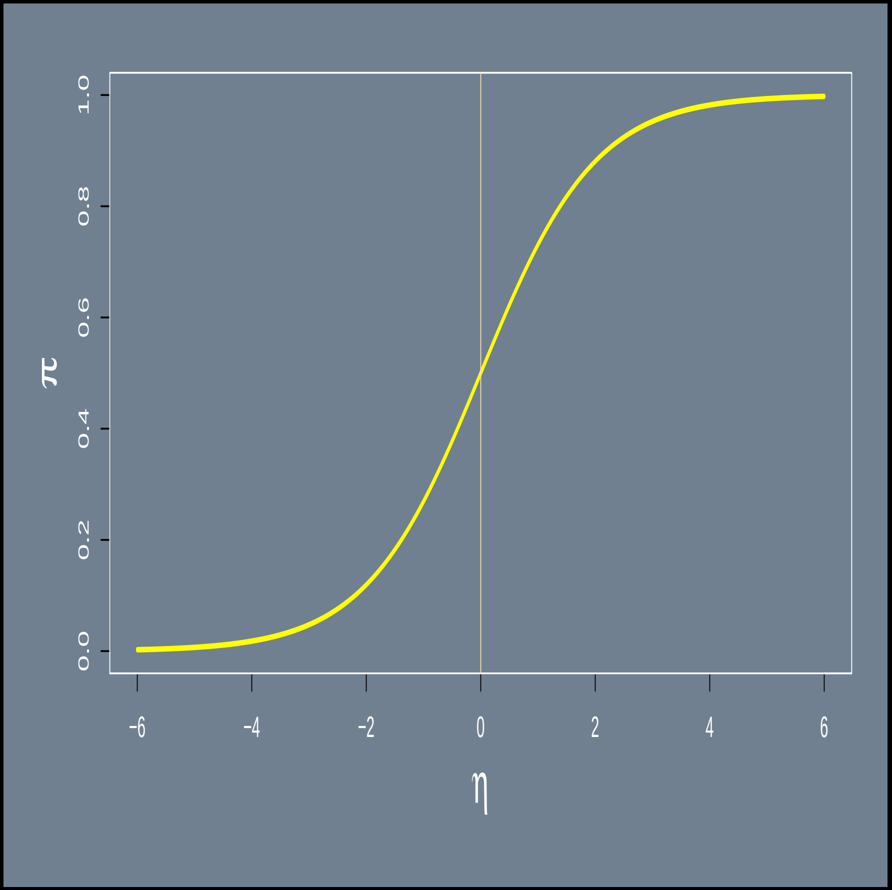
- These are sometimes given in $g()$ form: $\Phi^{-1}(\pi_i)$ and $\Lambda^{-1}(\pi_i) = \text{logit}(\pi_i) = \log\left(\frac{p_i}{1-p_i}\right)$.
- Less common is the cloglog function:

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

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:

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

  - ▷ 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), $x_{i1}, \ldots, x_{iq}$.

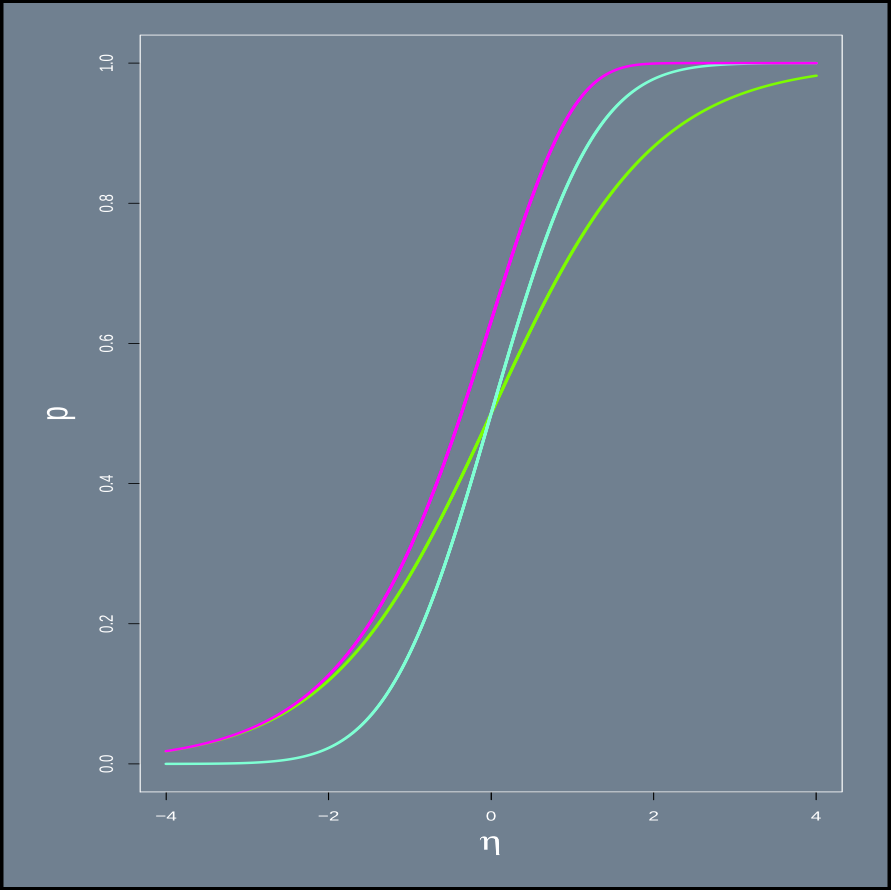
- The tool that connects these predictors to p is the linear predictor:

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

- 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 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(logitmod2$fitted.values,3)
    1     2     3     4     5     6     7     8     9    10    11    12    13
0.939 0.859 0.829 0.603 0.430 0.375 0.375 0.375 0.322 0.274 0.230 0.230 0.230
   14    15    16    17    18    19    20    21    22    23
0.230 0.158 0.130 0.086 0.086 0.069 0.069 0.045 0.036 0.023
```

from the Bernoulli treatment.

Percent Predicted Correctly

- The naïve criteria:

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

- Create the table:

```
ppc <- cbind(orings2$damage, round(logitmod2$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,] 0.69565 0.00000
[2,] 0.13043 0.17391
```

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

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 Challenger example ($n = 23$), I left off the following information before:

```
summary(logitmod)
    :
    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
```

Binomial Model Comparison

- Formal tests:
  - Specified model versus saturated model:

    ```
    pchisq(deviance(logitmod),df.residual(logitmod),lower=FALSE)
    0.71641
    ```

    which is not in the χ^2_{21} tail, so it is statistically "close" to the saturated model and therefore a good fit.

  - Mean model versus saturated model:

    ```
    pchisq(38.9,22,lower=FALSE)
    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 temperature) versus mean model ($D_S - D_L$):

    ```
    pchisq(38.9-16.9,1,lower=FALSE)
    2.7265e-06
    ```

    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.
  - We could also have done a Wald test on temperature:

    ```
                Estimate Std. Error z value Pr(>|z|)
    temp         -0.2162     0.0532   -4.07  4.8e-05
    ```

    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: $\hat{\beta}_j \pm z^{\alpha/2} se(\hat{\beta}_j)$.

- Low-tech method:

```
c(-0.2162-1.96*0.0532,-0.2162+1.96*0.0532)
-0.32047 -0.11193
```

- Hi-tech method:

```
summary(logitmod)$coefficients[,1]
        + qnorm(0.975) * t(c(-1,1) %o% summary(logitmod)$coefficients[,2])
(Intercept)  5.20243 18.12355
temp        -0.32046 -0.11201
```

- Profile likelihood version (accounts for covariance):

```
library(MASS)
confint(logitmod)
Waiting for profiling to be done...
               2.5 %   97.5 %
(Intercept)  5.57520 18.73760
temp        -0.33266 -0.12018
```

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: Anaemia

- Consider again the study of anaemia in women in a given clinic where 20 cases are chosen at random from the full study to get the data here.

- From a blood sample we get:
  - haemoglobin 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.

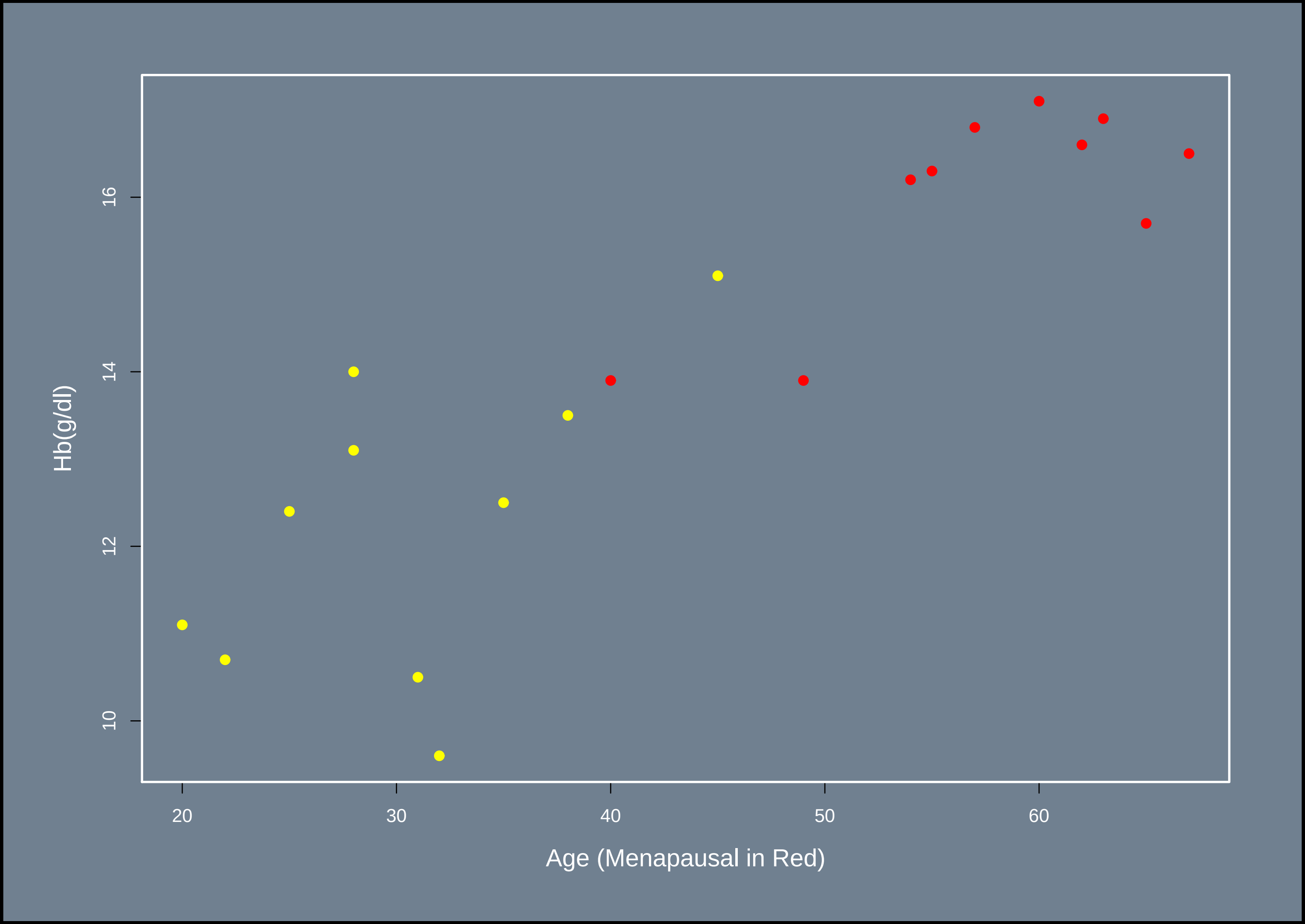
- We also have:
  - age in years
  - menopausal (0=no, 1=yes)

- There is an obvious endogeneity problem in modeling Hb(g/dl) versus PCV(%).

Anaemia 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

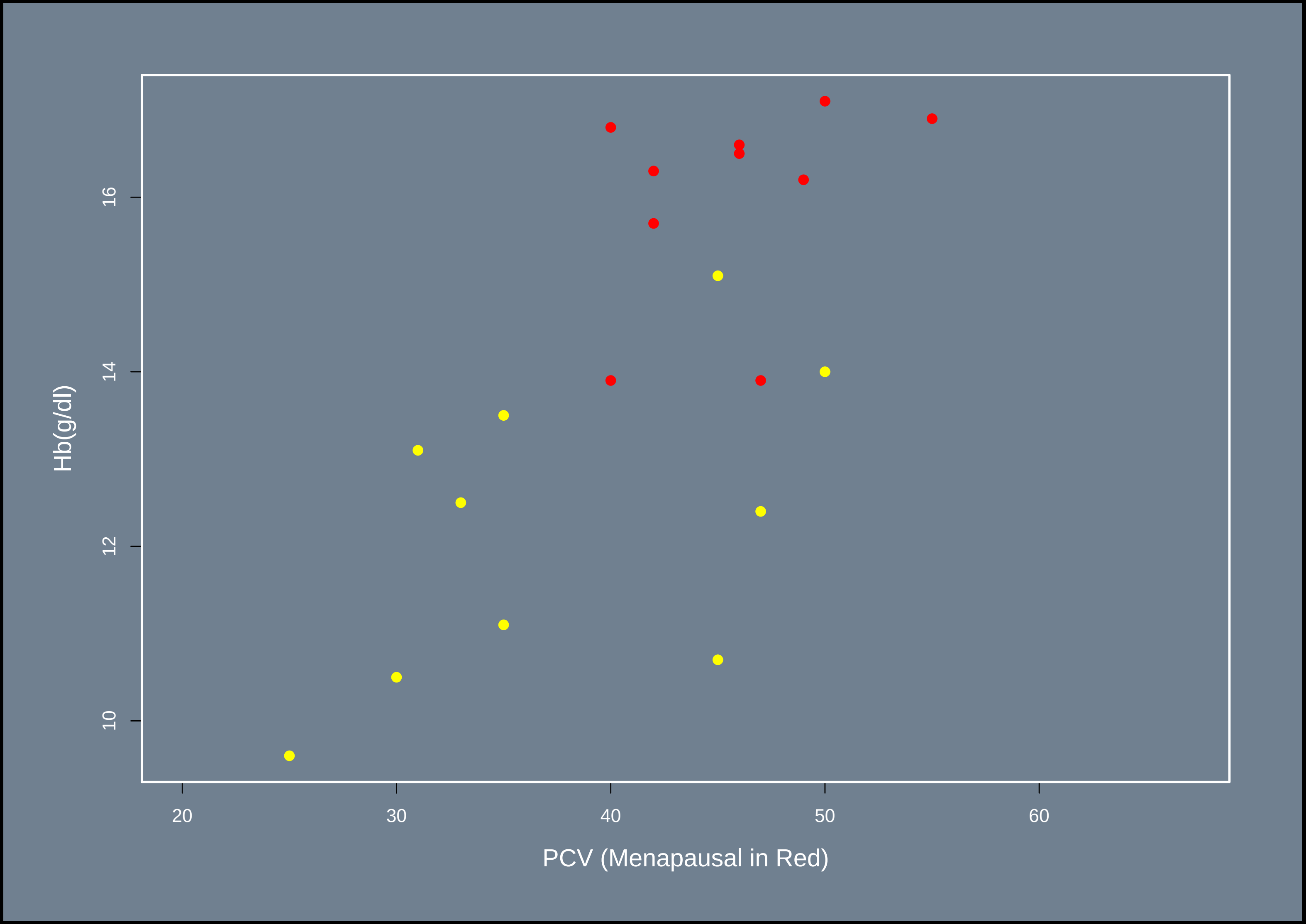
Scatterplot of the Anaemia Data



Scatterplot of the Anaemia 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()
```

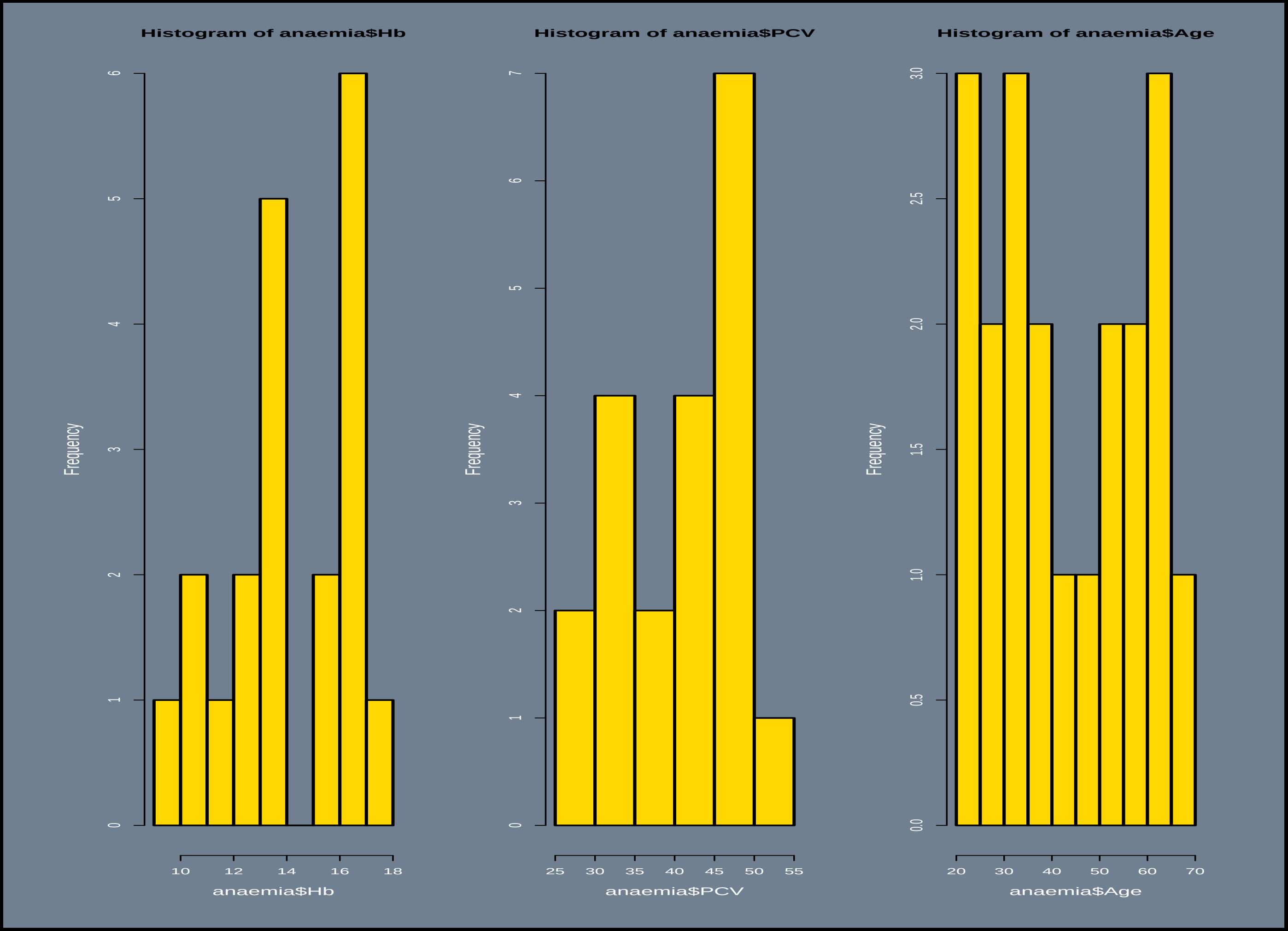
Scatterplot of the Anaemia Data



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()
```

Distribution of the Anaemia Data?



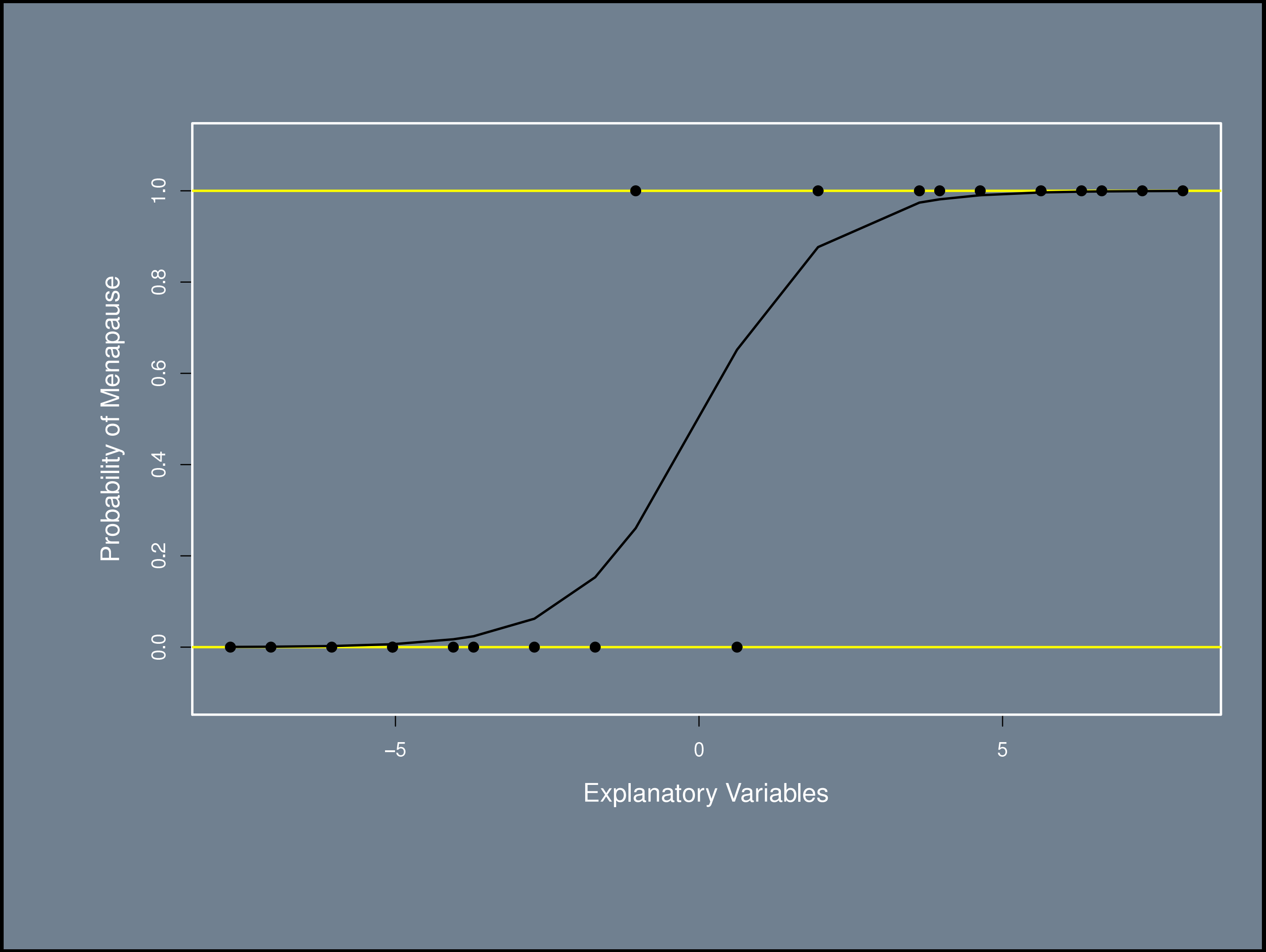
Logistic Regression: Anaemia Example

```
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
```

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
    ```

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:

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

The Poisson PMF

- Back to the probability mass function:

$$f(Y|\lambda) = \frac{(\lambda)^Y e^{-\lambda}}{Y!}, \qquad\qquad y = 0, 1, 2, \dots, \;\; \lambda > 0$$

  where λ is the "intensity parameter."

- This is the probability that exactly Y arrivals occur.

- Faraway's Galapagos Island data:

```
data(gala)
head(gala)
             Species Endemics  Area Elevation Nearest Scruz Adjacent
Baltra            58       23 25.09       346     0.6   0.6     1.84
Bartolome         31       21  1.24       109     0.6  26.3   572.33
Caldwell           3        3  0.21       114     2.8  58.7     0.78
Champion          25        9  0.10        46     1.9  47.4     0.18
Coamano            2        1  0.05        77     1.9   1.9   903.82
Daphne.Major      18       11  0.34       119     8.0   8.0     1.84
```

Poisson Assumptions

- **Infinitesimal Interval.** The probability of an arrival in the interval: $(t : \delta t)$ equals $\lambda\delta t + \circ(\delta t)$ where λ is the intensity parameter discussed above and $\circ(\delta t)$ is a time interval with the property: $\lim_{\delta t \to 0} \frac{\circ(\delta t)}{\delta t} = 0$. In other words, as the interval δt reduces in size towards zero, $\circ(\delta t)$ is negligible compared to δt. This assumption is required to establish that λ adequately describes the intensity or expectation of arrivals. Typically there is no problem meeting this assumption provided that the time measure is adequately granular with respect to arrival rates.

- **Non-Simultaneity of Events.** The probability of more than one arrival in the interval: $(t : \delta t)$ equals $\circ(\delta t)$. Since $\circ(\delta t)$ is negligible with respect to $\lambda\delta t$ for sufficiently small $\lambda\delta t$, the probability of simultaneous arrivals approaches zero in the limit.

- **I.I.D. Arrivals.** The number of arrivals in any two consecutive or non-consecutive intervals are independent and identically distributed. More specifically, $P(Y = y) \ \in \ (T_j : T_{j+1})$ does not depend on $P(Y = y) \ \in \ (T_k : T_{k+1})$ for any $j \neq k$.

Poisson Features

- The intensity parameter (λ) is both the mean and variance for a single Poisson distributed random variable.

- The intensity parameter is tied to a time interval, and rescaling time rescales the intensity parameter.

- Sums of independent Poisson random variables are themselves Poisson.

- We can also specifically model time by including it in the intensity parameter: $\lambda^* = \lambda t$.

Derivation of MLE

- PMF:

$$p(Y = y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

- Likelihood function:

$$L(\lambda|\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

- Log-likelihood function:

$$\ell(\lambda|\mathbf{y}) = -n\lambda + \log(\lambda)\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log(y_i!)$$

- MLE:

$$\frac{d}{d\lambda}\ell(\lambda|\mathbf{y}) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} y_i \equiv 0 \;\Rightarrow\; n\lambda = \sum_{i=1}^{n} y_i \;\Rightarrow\; \hat{\lambda} = \bar{y}$$

Derivation of the Variance

- Second derivative of the LL:

$$\frac{d^2}{d\lambda^2}\ell(\lambda|\mathbf{y}) = \frac{d}{d\lambda}\left(-n + \frac{1}{\lambda}\sum_{i=1}^{n} y_i\right) = -\lambda^{-2}\sum_{i=1}^{n} y_i,$$

  called the Hessian.

- Fisher Information:

$$FI = -E_\lambda\left[\frac{d^2}{d\lambda^2}\ell(\lambda|\mathbf{y})\right] = -E_\lambda\left[-\lambda^{-2}\sum_{i=1}^{n} y_i\right] = n\bar{y}E_\lambda\left[\lambda^{-2}\right] = \frac{n}{\bar{y}}$$

  since $E\lambda = \bar{y}$.

- Variance:

$$\text{Var}[\lambda] = (FI)^{-1} = \bar{y}/n.$$

Link Function for Poisson Regression

- Definition:

$$\log(\lambda_i) = \eta_i \;\; \Rightarrow \;\; \lambda_i = \exp(\eta_i) = \exp(\mathbf{X}_i\boldsymbol{\beta})$$

- Start with the substitution:

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\Big|_{\lambda_i=\exp(\mathbf{X}_i\boldsymbol{\beta})} = \prod_{i=1}^{n} e^{-\exp(\mathbf{X}_i\boldsymbol{\beta})}\exp(\mathbf{X}_i\boldsymbol{\beta})^{y_i}/y_i!$$

- Take the log:

$$\ell(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n}\left[-\exp(\mathbf{X}_i\boldsymbol{\beta}) + y_i(\mathbf{X}_i\boldsymbol{\beta}) - \log(y_i!)\right]$$

- Now take the first derivative:

$$\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n}\left[\exp(\mathbf{X}_i\boldsymbol{\beta})\mathbf{X}_j + \mathbf{y}_i\mathbf{X}_j\right], \qquad \forall j$$

- Or in full matrix terms: $\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\lambda}$, where $\hat{\lambda} = \mathbf{X}\hat{\boldsymbol{\beta}}$ (the normal equation for the Poisson model).

- Problem: there does not exist a closed form solution for $\hat{\boldsymbol{\beta}}$, so we use numerical methods.

Application: Poisson Model of Military Coups.

- Sub-Saharan Africa has experienced a disproportionately high proportion of regime changes due to the military takeover of government for a variety of reasons, including ethnic fragmentation, arbitrary borders, economic problems, outside intervention, and poorly developed governmental institutions.

- These data, selected from a larger set given by Bratton and Van De Walle (1994), look at potential causal factors for counts of military coups (ranging from 0 to 6 events) in 33 sub-Saharan countries over the period from each country's colonial independence to 1989.

- Seven explanatory variables are chosen here to model the count of military coups: **Military Oligarchy** (the number of years of this type of rule); **Political Liberalization** (0 for no observable civil rights for political expression, 1 for limited, and 2 for extensive); **Parties** (number of legally registered political parties); **Percent Legislative Voting**; **Percent Registered Voting**; **Size** (in one thousand square kilometer units); and **Population** (given in millions).

Application: Poisson Model of Military Coups.

- A generalized linear model for these data with the Poisson link function is specified as:

$$g^{-1}(\boldsymbol{\theta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \exp\left[\mathbf{X}\boldsymbol{\beta}\right] = \mathbb{E}[\mathbf{Y}] = \mathbb{E}[\textbf{Military Coups}].$$

- In this specification, the systematic component is $\mathbf{X}\boldsymbol{\beta}$, the stochastic component is $\mathbf{Y} = \textbf{Military Coups}$, and the link function is $\boldsymbol{\theta} = \log(\mathbf{M})$.

- We can re-express this model by moving the link function to the left-hand side exposing the linear predictor: $g(\mathbf{M}) = \log(\mathbb{E}[\mathbf{Y}]) = \mathbf{X}\boldsymbol{\beta}$ (although this is now a less intuitive form for understanding the outcome variable).

- The `R` language GLM call for this model is:

```
africa.out <- glm(MILTCOUP ~ MILITARY+POLLIB+PARTY93+PCTVOTE+PCTTURN
                           +SIZE*POP+NUMREGIM*NUMELEC, family=poisson).
```

- The new part is `family=poisson`, where poisson is not capitalized.

Application: Poisson Model of Military Coups.

	Parameter Estimate	Standard Error	95% Confidence Interval
(Intercept)	2.9209	1.3368	[0.3008: 5.5410]
Military Oligarchy	0.1709	0.0509	[0.0711: 0.2706]
Political Liberalization	-0.4654	0.3319	[-1.1160: 0.1851]
Parties	0.0248	0.0109	[0.0035: 0.0460]
Percent Legislative Voting	0.0613	0.0218	[0.0187: 0.1040]
Percent Registered Voting	-0.0361	0.0137	[-0.0629:-0.0093]
Size	-0.0018	0.0007	[-0.0033:-0.0004]
Population	-0.1188	0.0397	[-0.1965:-0.0411]
Regimes	-0.8662	0.4571	[-1.7621: 0.0298]
Elections	-0.4859	0.2118	[-0.9010:-0.0709]
(Size)(Population)	0.0001	0.0001	[0.0001: 0.0002]
(Regimes)(Elections)	0.1810	0.0689	[0.0459: 0.3161]

Application: Poisson Model of Military Coups.

- Note that the two interaction terms are specified by using the multiplication character. The iteratively weighted least squares algorithm converged in only four iterations using Fisher scoring, and the results are provided in the table.

- The model appears to fit the data quite well:
  - an improvement from the null deviance of 62 on 32 degrees of freedom to a residual deviance of 7.5 on 21 degrees of freedom
  - evidence that the model does not fit would be supplied by a model deviance value in the tail of a χ^2_{n-k} distribution
  - and nearly all the coefficients have 95% confidence intervals bounded away from zero and therefore appear reliable in the model.

Poisson GLM of Capital Punishment Data

The model is developed from the Poisson link function, $\boldsymbol{\eta} = \log(\boldsymbol{\mu})$, with the objective of finding the best $\boldsymbol{\beta}$ vector in:

$$
\begin{aligned}
\underbrace{g^{-1}(\boldsymbol{\eta})}_{17\times 1} &= g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) \\
&= \exp\left[\boldsymbol{X}\boldsymbol{\beta}\right] \\
&= \exp\left[\mathbf{1}\beta_0 + \mathbf{INC}\beta_1 + \mathbf{POV}\beta_2 + \mathbf{BLK}\beta_3 + \mathbf{CRI}\beta_4 + \mathbf{SOU}\beta_5 + \mathbf{DEG}\beta_6\right] \\
&= \mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{EXE}].
\end{aligned}
$$

```
dp.97 <- read.table("http://jeffgill.org/files/jeffgill/files/cpunish.dat_.txt",
    ‘header=TRUE)
PROPDEGREE <- matrix(apply(dp.97[,12:14],1,sum)/apply(dp.97[8:14],1,sum),
        nrow(dp.97),1,dimnames=list(dimnames(dp.97)[[1]],"PROPDEGREE"))
dp.97 <- cbind(dp.97,PROPDEGREE)
dp.out <- glm(EXECUTIONS ~ INCOME + PERPOVERTY + PERBLACK + log(VC100k96) + SOUTH
                                  + PROPDEGREE, family=poisson, data=dp.97)
```

Poisson GLM of Capital Punishment Data, 1997

		Median	Percent	Percent	Violent		Proportion
State	Executions	Income	Poverty	Black	Crime/100K	South	w/Degrees
Texas	37	34453	16.7	12.2	644	1	0.16
Virginia	9	41534	12.5	20.0	351	1	0.27
Missouri	6	35802	10.6	11.2	591	0	0.21
Arkansas	4	26954	18.4	16.1	524	1	0.16
Alabama	3	31468	14.8	25.9	565	1	0.19
Arizona	2	32552	18.8	3.5	632	0	0.25
Illinois	2	40873	11.6	15.3	886	0	0.25
South Carolina	2	34861	13.1	30.1	997	1	0.21
Colorado	1	42562	9.4	4.3	405	0	0.31
Florida	1	31900	14.3	15.4	1051	1	0.24
Indiana	1	37421	8.2	8.2	537	0	0.19
Kentucky	1	33305	16.4	7.2	321	0	0.16
Louisiana	1	32108	18.4	32.1	929	1	0.18
Maryland	1	45844	9.3	27.4	931	0	0.29
Nebraska	1	34743	10.0	4.0	435	0	0.24
Oklahoma	1	29709	15.2	7.7	597	0	0.21
Oregon	1	36777	11.7	1.8	463	0	0.25
	EXE	**INC**	**POV**	**BLK**	**CRI**	**SOU**	**DEG**

Source: United States Census Bureau, United States Department of Justice.

Poisson GLM of Capital Punishment Data

Table 1: Modeling Capital Punishment in the United States: 1997

	Coefficient	Standard Error	95% Confidence Interval
(Intercept)	-6.30665	4.17678	[-14.49299: 1.87969]
Median Income	0.00027	0.00005	[0.00017: 0.00037]
Percent Poverty	0.06897	0.07979	[-0.08741: 0.22534]
Percent Black	-0.09500	0.02284	[-0.13978: -0.05023]
log(Violent Crime)	0.22124	0.44243	[-0.64591: 1.08838]
South	2.30988	0.42875	[1.46955: 3.15022]
Degree Proportion	-19.70241	4.46366	[-28.45102:-10.95380]
Null deviance: 136.573, $df = 16$			Maximized $\ell()$: -31.7375
Summed deviance: 18.212, $df = 11$			AIC: 77.475

First Differences for Non-Linear Models

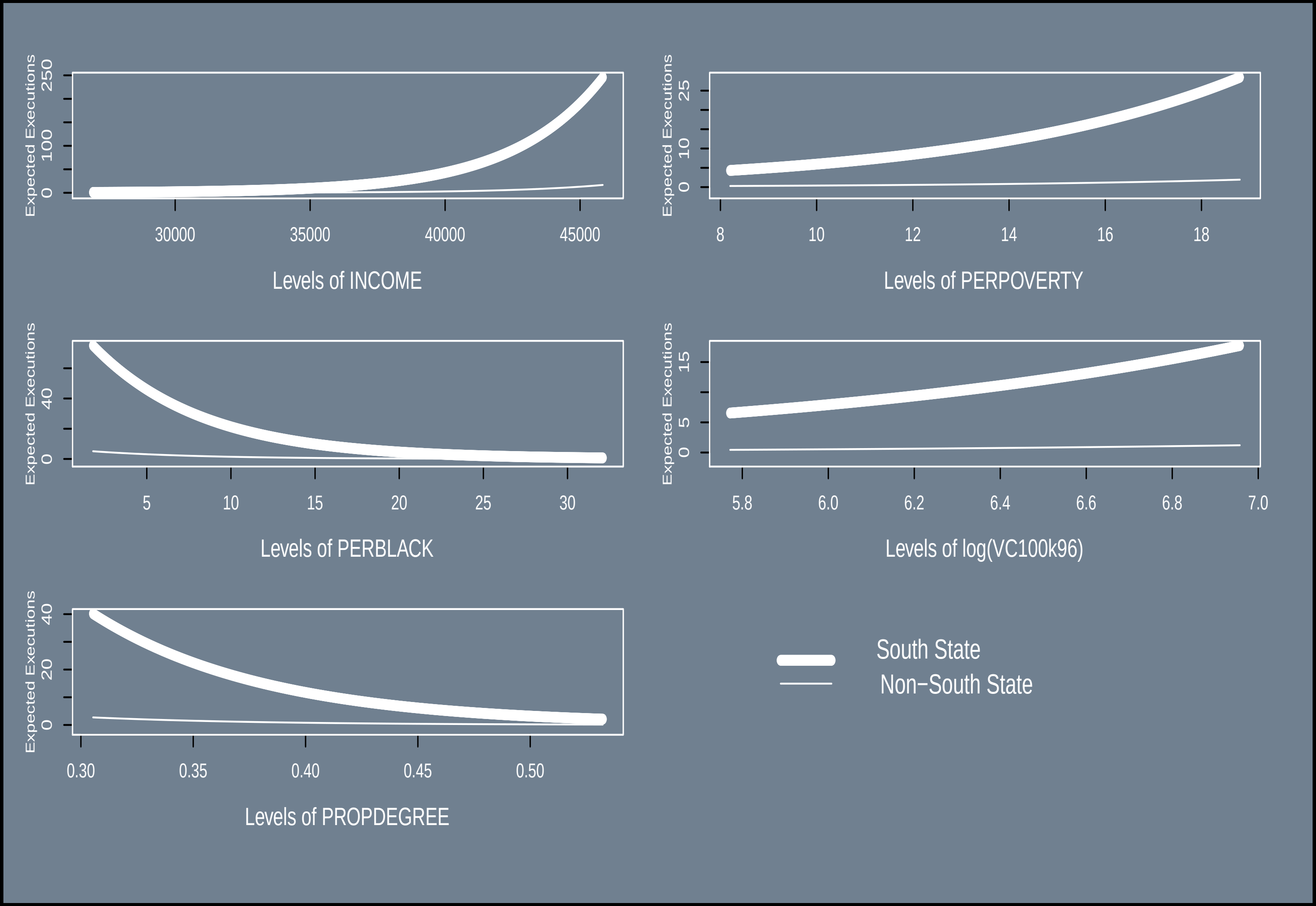
- We can no longer use “a one unit change in X gives a β change in Y.”
- Main idea:
  - pick one covariate of interest, $\mathbf{X}_q$
  - choose 2 levels of this variable, $\mathbf{X}_{1,q}$, $\mathbf{X}_{2,q}$
  - set all other covariates at their mean, $\bar{\mathbf{X}}_{-q}$
  - create two predictions by running these values through the link function:

$$\hat{Y}_1 = g^{-1}(\bar{\mathbf{X}}_{-q}\hat{\boldsymbol{\beta}}_{-q} + \mathbf{X}_{1,q}\hat{\boldsymbol{\beta}}_q)$$

$$\hat{Y}_2 = g^{-1}(\bar{\mathbf{X}}_{-q}\hat{\boldsymbol{\beta}}_{-q} + \mathbf{X}_{2,q}\hat{\boldsymbol{\beta}}_q)$$

  - Look at $\hat{Y}_1 - \hat{Y}_2$.
- For example:

```
dp.1 <- dp.2 <- c(1,apply(dp.97[,c(3,4,5,6,7,15)],2,mean))
dp.1[6] <- 0; dp.2[6] <- 1
y.1 <- exp(dp.1 %*% dp.out$coef); y.2 <- exp(dp.2 %*% dp.out$coef)
y.2 - y.1
```


Expected Executions
0
100
250
30000
35000
40000
45000
Levels of INCOME
Expected Executions
0
10
25
8
10
12
14
16
18
Levels of PERPOVERTY
Expected Executions
0
40
5
10
15
20
25
30
Levels of PERBLACK
Expected Executions
0
5
15
5.8
6.0
6.2
6.4
6.6
6.8
7.0
Levels of log(VC100k96)
Expected Executions
0
20
40
0.30
0.35
0.40
0.45
0.50
Levels of PROPDEGREE
South State
Non−South State

Poisson GLM of Capital Punishment, First Difference Code

```
X <- cbind(rep(1,nrow(dp.97)), as.matrix(dp.97[,3:5]), as.matrix(log(dp.97[,6])),
        as.matrix(dp.97[,7]), as.matrix(dp.97[,15]))
X.0 <- cbind(X[,1:5],rep(0,length=nrow(X)),X[,7])
dimnames(X.0)[[2]] <- names(dp.out$coefficients)
X.1 <- cbind(X[,1:5],rep(1,length=nrow(X)),X[,7])
dimnames(X.1)[[2]] <- names(dp.out$coefficients)

postscript("/Users/jgill/Class.MLE/glm.fig2.ps")
par(mfrow=c(3,2),mar=c(4,3,2,2),oma=c(3,1,1,1),col.axis="white",col.lab="white",
    col.sub="white",col="white",bg="slategray")
```

Poisson GLM of Capital Punishment, First Difference Code

```
for (i in 2:(ncol(X.0)-1))  {
  if (i==6) i <- i+1
  ruler  <- seq(min(X.0[,i]),max(X.0[,i]),length=1000)
  xbeta0 <- exp(dp.out$coefficients[-i]%*%apply(X.0[,-i],2,mean)
                + dp.out$coefficients[i]*ruler)
  xbeta1 <- exp(dp.out$coefficients[-i]%*%apply(X.1[,-i],2,mean)
                + dp.out$coefficients[i]*ruler)
  plot(ruler,xbeta0,type="l",xlab="",ylab="",
      ylim=c(min(xbeta0,xbeta1)-2,max(xbeta0,xbeta1)) )
  lines(ruler,xbeta1,type="b")
  mtext(outer=F,side=1,paste("Levels of",dimnames(X.0)[[2]][i]),cex=0.8,line=3)
  mtext(outer=F,side=2,"Expected Executions",cex=0.6,line=2)
}
plot(ruler[100:200],rep(ruler[400],101),bty="n",xaxt="n",yaxt="n",xlab="",ylab="",
        type="l",xlim=range(ruler),ylim=range(ruler))
lines(ruler[100:200],rep(ruler[600],101),type="b")
text(ruler[445],ruler[400],"Non-South State",cex=1.4)
text(ruler[390],ruler[700],"South State",cex=1.4)
dev.off()
```

Poisson GLM of Capital Punishment, Continued

Table 2: Residuals From Poisson Model of Capital Punishment

	Response	**Pearson**	**Working**	**Deviance**	**Anscombe**
Texas	1.70755431	0.28741478	0.04837752	0.28515874	0.28292493
Virginia	0.87407687	0.30671010	0.10762321	0.30136452	0.29629097
Missouri	4.59530299	3.86395636	3.24898061	2.86925916	2.27854829
Arkansas	0.26481208	0.13694108	0.07081505	0.13544624	0.13391171
Alabama	0.95958171	0.67097152	0.46916278	0.62736060	0.58874967
Arizona	0.95395198	0.93375106	0.91397549	0.82741022	0.74425671
Illinois	0.13924315	0.10197129	0.07467388	0.10084230	0.09963912
South Carolina	-0.38227185	-0.24752186	-0.16027167	-0.25478237	-0.26235519
Colorado	-0.95901329	-0.68428704	-0.48826435	-0.75706323	-0.84845827
Florida	-1.82216650	-1.08543456	-0.64657649	-1.25272634	-1.49557143
Indiana	-2.17726883	-1.21566195	-0.67880001	-1.42915840	-1.74185735
Kentucky	-2.31839936	-1.26926054	-0.69489994	-1.49593905	-1.83715998
Louisiana	-1.60160305	-0.99359914	-0.61640776	-1.13620002	-1.33738726
Maryland	0.10161119	0.10709684	0.11287657	0.10527242	0.10341466
Nebraska	0.07022962	0.07261924	0.07506941	0.07194451	0.07107841
Oklahoma	0.49917358	0.70406163	0.99304011	0.62019695	0.55401828
Oregon	-0.90510552	-0.65451282	-0.47330769	-0.72189767	-0.80517526

New and Old Ways to Look at Model Fit

- Approximation to Pearson's Statistic.

$$X^2 = \sum_{i=1}^{n} \mathbf{R}^2_{Pearson} = \sum_{i=1}^{n} \left[\frac{\mathbf{Y} - \boldsymbol{\mu}}{\sqrt{VAR[\boldsymbol{\mu}]}} \right]^2.$$

- If the sample size is sufficiently large, then $\frac{X^2}{a(\psi)} \sim \chi^2_{n-p}$ where n is the sample size, p is the number of explanatory variables including the constant, and $a(\psi)$ is the scale function that we'll see in Chapter 6.

- For the summed deviance with sufficient sample size it is also true that $D(\boldsymbol{\eta}, \mathbf{y})/a(\psi) \sim \chi^2_{n-p}$.

- Recall that it is also common to contrast this with the *null deviance*: the deviance function calculated for a model with no covariates (mean function only).

New and Old Ways to Look at Model Fit

- Akaike Information Criterion.
  minimizes the negative likelihood penalized by the number of parameters:

  $$\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}|\mathbf{y}) + 2p$$

  where $\ell(\hat{\boldsymbol{\beta}}|\mathbf{y})$ is the maximized model log likelihood value and p is the number of explanatory variables in the model (including the constant). (AIC has a bias towards models that overfit with extra parameters since the penalty component is obviously linear with increases in the number of explanatory variables, and the log likelihood often increases more rapidly.)

- Schwartz Criterion/Bayesian Information Criterion (BIC).

  $$\text{BIC} = -2\ell(\hat{\boldsymbol{\beta}}|\mathbf{y}) + p\text{log}(n)$$

  where n is the sample size.

- There is also a Deviance Information Criterion (DIC) used in Bayesian MCMC estimation.

Congressional Cosponsoring of Bills

- Fowler (2006) looks at patterns of sponsorship and cosponsorship in Congress from 1973 to 2004.

```
cosponsor <- read.table("http://jeffgill.org/files/jeffgill/files/fowler.dat_.txt", header=TRUE)
summary(cosponsor)
Congress        Period   Total.Sponsors   Total.Bills    Mean.Bills.Per.Leg
100th: 2   1974: 2   Min.   : 99.0   Min.   : 4188   Min.   : 15.00
101st: 2   1976: 2   1st Qu.:101.0   1st Qu.: 7278   1st Qu.: 19.00
102nd: 2   1978: 2   Median :267.0   Median : 7849   Median : 44.50
103rd: 2   1980: 2   Mean   :268.9   Mean   : 8814   Mean   : 47.62
104th: 2   1982: 2   3rd Qu.:437.0   3rd Qu.: 8832   3rd Qu.: 75.00
105th: 2   1984: 2   Max.   :442.0   Max.   :20994   Max.   :111.00

Mean.Cos.Per.Leg Mean.Cos.Per.Bill  Cos.Per.Leg      Mean.Dist         Senate
Min.   :121.0    Min.   : 2.000    Min.   : 49.0   Min.   :1.170   Min.   :0.0
1st Qu.:174.5    1st Qu.: 3.000    1st Qu.: 70.0   1st Qu.:1.300   1st Qu.:0.0
Median :260.0    Median : 4.000    Median : 80.5   Median :1.545   Median :0.5
Mean   :247.5    Mean   : 7.969    Mean   :101.5   Mean   :1.515   Mean   :0.5
3rd Qu.:303.8    3rd Qu.:14.250    3rd Qu.:143.2   3rd Qu.:1.673   3rd Qu.:1.0
Max.   :376.0    Max.   :19.000    Max.   :184.0   Max.   :1.950   Max.   :1.0
```

Application to Congressional Cosponsoring of Bills

- Look at summary statistics:

```
mean(cosponsor$Mean.Bills.Per.Leg)
[1] 47.625
var(cosponsor$Mean.Bills.Per.Leg)
[1] 828.24

mean(cosponsor$Mean.Cos.Per.Leg)
[1] 247.5
var(cosponsor$Mean.Cos.Per.Leg)
[1] 6134.7
```

- This is clear evidence of *overdispersion* in the original unconditional count data.

- We are actually more interested in overdispersion in the modeled counts, which are conditional on the form of the model specification including the link function and the collection of covariates.

Over/Under Dispersion

- For Poisson models the mean and the variance of a single random variable are assumed to be the same.

- For the likelihood function as a statistic, the variance is scaled by n.

- Overdispersion, $\text{Var}(Y) > \mathbb{E}(Y)$, is relatively common, whereas underdispersion, $\text{Var}(Y) < \mathbb{E}(Y)$ is rare.

- Biggest effect is to make the standard errors wrong.

- One diagnostic: plot $\hat{\mu}$ versus $(y - \hat{\mu})^2$.

- Solution: make μ a random variable rather than a fixed constant to be estimated, with a gamma distribution: $G[\mu\alpha, \alpha]$. So
$$\mathbb{E}[Y] = \mu \qquad\qquad \text{Var}[Y] = \frac{\mu}{\phi}$$

- This is called the "Poisson-Gamma" model and it means that Y is distributed *negative binomial.*

Negative Binomial

- Negative binomial distribution has the same sample space (i.e. on the counting measure) as the Poisson, but contains an additional parameter which can be thought of as gamma distributed and therefore used to model a variance function.

- Used by many to fit a count model with overdispersion.

- The binomial distribution measures the number of successes in a given number of fixed trials, whereas the negative binomial distribution measures *the number of failures, y before the r^{th} success.*

- An alternative but equivalent form,

$$f(y|r,p) = \binom{y-1}{r-1} p^r (1-p)^{y-r},$$

  measures the number of trials necessary to get r successes.

- An important application of the negative binomial distribution is in survey research design. If the researcher knows the value of p from previous surveys, then the negative binomial can provide the number of subjects to contact in order to get the desired number of responses for analysis.

Negative Binomial

- The PMF is:

$$f(Y|k,p) = \binom{y-1}{k-1} p^k (1-p)^{y-k}, \qquad y = 0, 1, 2, \ldots, \qquad 0 \le p \le 1.$$

- For this parameterization, we get:

$$\mathbb{E}[Y] = \mu, \qquad \text{Var}[Y] = \frac{\mu(1+\phi)}{\phi}.$$

- If ϕ (the dispersion parameter) is unknown, use the estimate:

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

- This gives an F-test for comparing models (big values implies a difference in models).

Negative Binomial

- There are two interpretations:
  - as a generalized Poisson,
  - with probability p, modeling the number of trials, Y, before the kth success (alternatively failure) where k is fixed in advance.
- For estimation, use `library(MASS)`, which has `glm.nb`.
- Note that there is also:

```
dnbinom(x, size, prob, mu, log = FALSE)
pnbinom(q, size, prob, mu, lower.tail = TRUE, log.p = FALSE)
qnbinom(p, size, prob, mu, lower.tail = TRUE, log.p = FALSE)
rnbinom(n, size, prob, mu)
```

Negative Binomial GLM, Congressional Activity: 1995

- Compare the number of bills assigned to committee in the first 100 days of the 103rd and 104th Houses as a function of the number of members on the committee, the number of subcommittees, the number of staff assigned to the committee, and a dummy variable indicating whether or not it is a high prestige committee.

- The model is developed with the link function:

$$\eta = g(\mu) = \log\left(\frac{\mu}{\mu + \frac{1}{k}}\right) \quad \longrightarrow \quad \mu = g^{-1}(\eta) = \frac{\exp(\eta)}{k(1 - \exp(\eta))},$$

  where $\eta = \mathbf{X}\boldsymbol{\beta}$, and $k \geq 1$ is the overdispersion term.

Negative Binomial GLM, Bills Assigned to Committed, First 100 Days

Committee	Size	Subcommittees	Staff	Prestige	Bills–103rd	Bills–104th
Appropriations	58	13	109	1	9	6
Budget	42	0	39	1	101	23
Rules	13	2	25	1	54	44
Ways and Means	39	5	23	1	542	355
Banking	51	5	61	0	101	125
Economic/Educ. Opportunities	43	5	69	0	158	131
Commerce	49	4	79	0	196	271
International Relations	44	3	68	0	40	63
Government Reform	51	7	99	0	72	149
Judiciary	35	5	56	0	168	253
Agriculture	49	5	46	0	60	81
National Security	55	7	48	0	75	89
Resources	44	5	58	0	98	142
Transport./Infrastructure	61	6	74	0	69	155
Science	50	4	58	0	25	27
Small Business	43	4	29	0	9	8
Veterans Affairs	33	3	36	0	41	28
House Oversight	12	0	24	0	233	68
Standards of Conduct	10	0	9	0	0	1
Intelligence	16	2	24	0	2	4

Model Code

```
committee.dat <-
 read.table("http://jeffgill.org/files/jeffgill/files/committe.dat_.txt",header=TRUE)

committee.poisson <- glm(BILLS104 ~ SIZE + SUBS * (log(STAFF)) + PRESTIGE +
        BILLS103, family=poisson, data=committee.dat)
1 - pchisq(summary(committee.poisson)$deviance,
           summary(committee.poisson)$df.residual)
[1] 0   # IN THE TAIL INDICATES OVERDISPERSION

committee.out <- glm.nb(BILLS104 ~ SIZE + SUBS * (log(STAFF)) + PRESTIGE +
        BILLS103, data=committee.dat)

resp <- resid(committee.out,type="response")
pears <- resid(committee.out,type="pearson")
working <- resid(committee.out,type="working")
devs <- resid(committee.out,type="deviance")
cbind(resp,pears,working,devs)
```

Negative Binomial GLM, Congressional Activity: 1995

```
                              resp    pears  working     devs
Appropriations            -7.38308 -0.99451 -0.55167 -1.22671
Budget                    -6.17325 -0.40931 -0.21161 -0.43997
Rules                     22.54158  1.98665  1.05048  1.56745
Ways_and_Means          -135.06135 -0.56848 -0.27560 -0.63081
Banking                   21.00117  0.40998  0.20194  0.38568
Economic_Educ_Oppor      -93.92104 -0.85695 -0.41757 -1.01572
Commerce                 -58.03818 -0.36306 -0.17639 -0.38675
International_Relations  -49.33480 -0.89295 -0.43918 -1.06810
Government_Reform         32.60986  0.57003  0.28018  0.52480
Judiciary                 27.80878  0.25343  0.12349  0.24378
Agriculture               24.21181  0.85168  0.42635  0.75680
National_Security         27.14348  0.87911  0.43881  0.77861
Resources                 26.13708  0.45893  0.22559  0.42884
TransInfrastructure       79.10378  2.10068  1.04226  1.64133
Science                  -34.35454 -1.12146 -0.55993 -1.43001
Small_Business           -12.50419 -1.14887 -0.60984 -1.48074
Veterans_Affairs         -14.18802 -0.66378 -0.33630 -0.75200
House_Oversight           16.14917  0.62009  0.31145  0.56716
Stds_of_Conduct            0.37836  0.44850  0.60864  0.40700
Intelligence             -13.58498 -1.43490 -0.77253 -2.05981
```

Modeling Bill Assignment – 104th House, Results

	Coefficient	Standard Error	95% Confidence Interval
(Intercept)	-6.80543	2.54651	[-12.30683:-1.30402]
Size	-0.02825	0.02093	[-0.07345: 0.01696]
Subcommittees	1.30159	0.54370	[0.12701: 2.47619]
log(Staff)	3.00971	0.79450	[1.29329: 4.72613]
Prestige	-0.32367	0.44102	[-1.27644: 0.62911]
Bills in 103rd	0.00656	0.00139	[0.00355: 0.00957]
Subcommittees:log(STAFF)	-0.32364	0.12489	[-0.59345:-0.05384]
Null deviance: 107.314, $df = 19$			Maximized $\ell()$: 10559
Summed deviance: 20.948, $df = 13$			AIC: 121130

Gamma Regression

- The Gamma GLM is used when the support of the outcome variable is $[0{:}\infty]$.

- Assume Y is distributed gamma indexed by two parameters: the shape parameter, and the inverse-scale parameter.

- The gamma distribution is most commonly written in “rate” format:

$$f(y|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}y^{\alpha-1}e^{-\beta y}, \qquad y,\alpha,\beta > 0.$$

- `R` uses as a default the “scale” format:

$$f(y|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^{-\alpha}y^{\alpha-1}e^{-y/\beta}, \qquad y,\alpha,\beta > 0.$$

Gamma Regression

- The canonical link for the gamma family variable μ, is $\theta = -\frac{1}{\mu}$.

- So $b(\theta) = \log(\mu) = \log\left(-\frac{1}{\theta}\right)$ with the restriction: $\theta < 0$. Therefore: $b(\theta) = -\log(-\theta)$.

- The χ^2 distribution is gamma$(\frac{\rho}{2}, \frac{1}{2})$ for ρ degrees of freedom, and the exponential distribution is gamma$(1, \beta)$.

Gamma GLM of Electoral Politics in Scotland

- On September 11, 1997 Scottish voters overwhelming (74.3%) approved the establishment of the first Scottish national parliament in nearly three hundred years.
- On the same ballot, the voters gave strong support (63.5%) to granting this parliament taxation powers.
- Data: 32 *Unitary Authorities* (also called council districts), U.K. government sources, includes 40 potential explanatory variables
- Used here: CouncilTax (COU), PerClaimantFemale (PCR), StdMortalityRatio (MOR), Active (ACT), GDP (GDP), Percentage5to15 (PER).

The model for these data using the gamma link function is produced by:

$$\begin{aligned}\underbrace{g^{-1}(\boldsymbol{\theta})}_{32\times 1} &= g^{-1}(\boldsymbol{X\beta})\\ &= -\frac{1}{\boldsymbol{X\beta}}\\ &= -\left[\mathbf{1}\beta_0 + \mathbf{COU}\beta_1 + \mathbf{PCR}\beta_2 + \mathbf{MOR}\beta_3 + \mathbf{ACT}\beta_4 + \mathbf{GDP}\beta_5\right]^{-1}\\ &= E[\mathbf{Y}] = E[\mathbf{YES}].\end{aligned}$$

The systematic component here is $\boldsymbol{X\beta}$, the stochastic component is $\mathbf{Y} = \mathbf{YES}$, and the link function is $\boldsymbol{\theta} = -\frac{1}{\boldsymbol{\mu}}$.

Gamma GLM

```
scotland.df <- read.table(
    "http://jeffgill.org/files/jeffgill/files/scotvote.dat_.txt",
    header=TRUE)

scottish.vote.glm <- glm((PerYesTax/100) ~ CouncilTax * PerClaimantFemale
                          + StdMortalityRatio + Active + GDP + Percentage5to15,
                          family=Gamma, data=scotland.df)

graph.summary(scottish.vote.glm)
```

Gamma GLM

```
Family: Gamma      Link function: inverse
                               Coef Std.Err. 0.95 Lower 0.95 Upper CIs:ZE+RO
(Intercept)                  -1.777    1.148     -4.026       0.473  |--o--|
CouncilTax                    0.005    0.002      0.002       0.008      |o|
PerClaimantFemale             0.203    0.053      0.099       0.308      |o|
StdMortalityRatio            -0.007    0.003     -0.012      -0.002      |o|
Active                        0.011    0.004      0.003       0.019      |o|
GDP                           0.000    0.000      0.000       0.000      |o|
Percentage5to15              -0.052    0.024     -0.099      -0.005      |o|
CouncilTax:PerClaimantFemale  0.000    0.000      0.000       0.000      |o|
N: 32    log-likelihood: 59.892    AIC: -111.784  Dispersion Parameter: 0.0035842
    Null deviance: 0.536 on 31 degrees of freedom
Residual deviance:  0.087 on 24 degrees of freedom
```