Bioassay
Generalized Linear Models - Part I

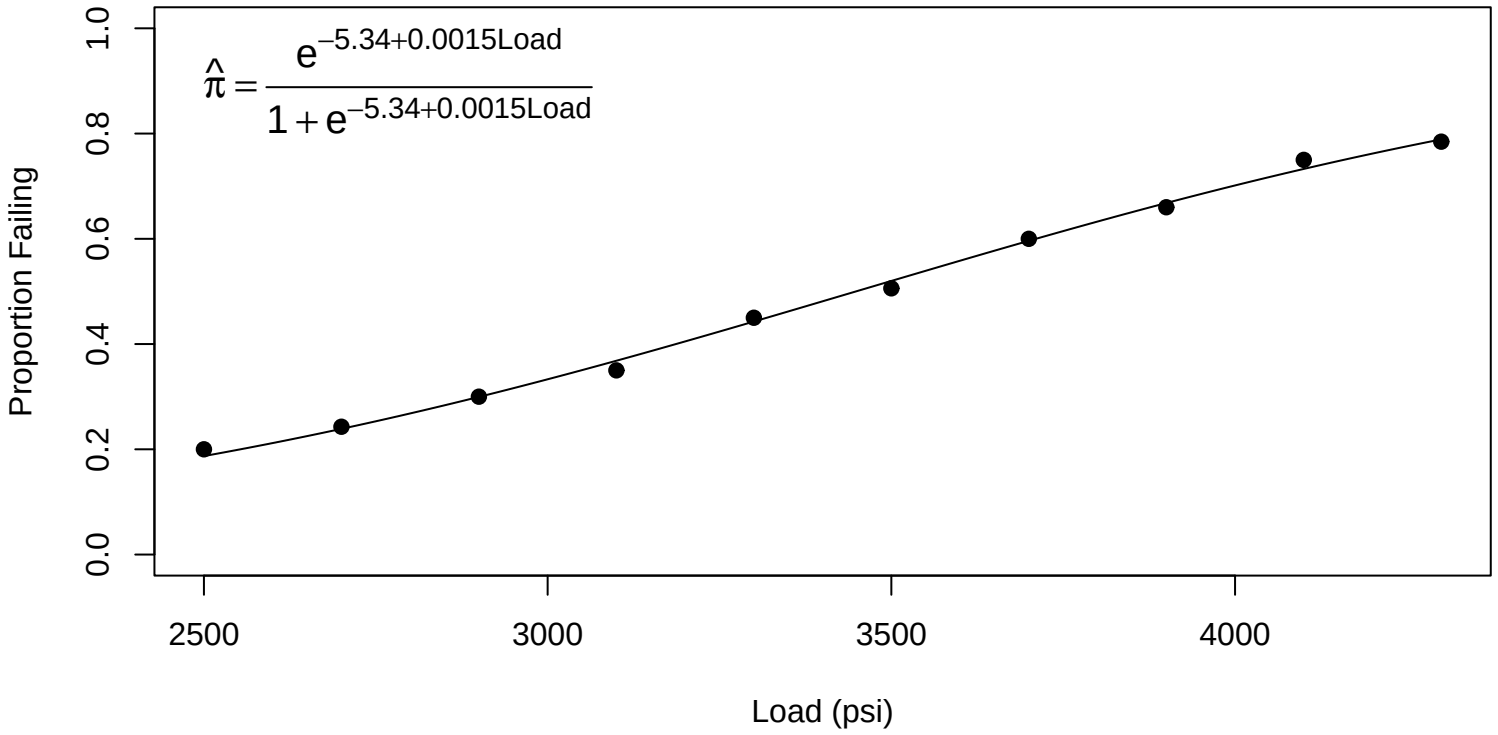
Statistics 149

Spring 2006

Bioassay - Effective/Lethal Doses

Example: Aircraft fasteners

A study was conducted to investigate the effect of pressure loads on the compressive strength of alloy fasteners used in aircraft construction.



So far we have looked at questions along the lines of

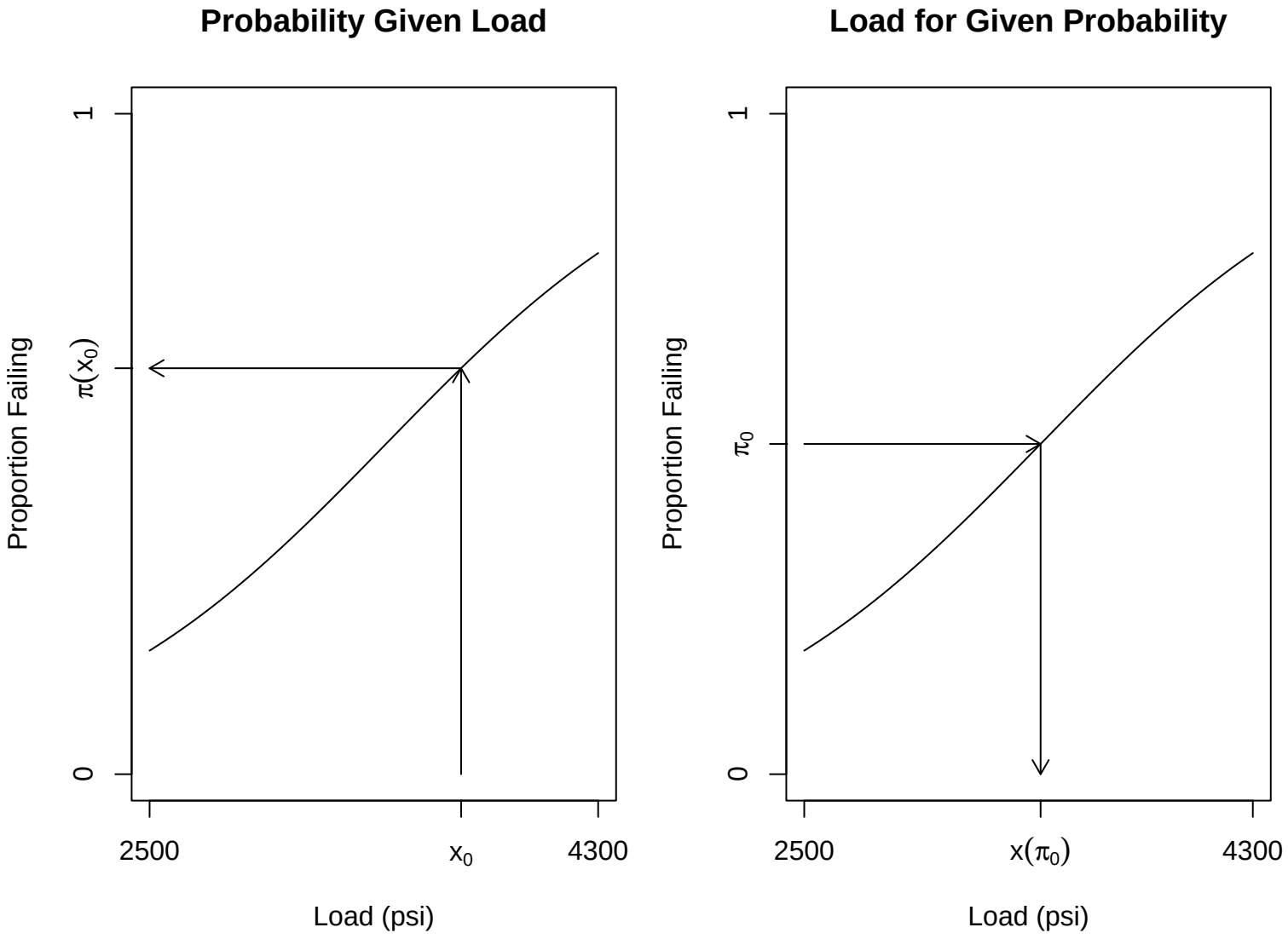
> What is the probability that a fastener will fail for a given load x_0?

By logistic regression, the estimate of this is

$$\hat{\pi}(x_0) = \frac{e^{-5.34+0.0015x_0}}{1+e^{-5.34+0.0015x_0}}$$

Instead, suppose we are interested in the question

> What load will give a probability of failure of 25%? 50%? π_0 (in general)?


Probability Given Load
Proportion Failing
1
$\pi(x_0)$
0
2500
x_0
4300
Load (psi)
Load for Given Probability
Proportion Failing
1
π_0
0
2500
$x(\pi_0)$
4300
Load (psi)

If we knew β, we need to solve

$$\pi_0 = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

or equivalently

$$\text{logit}(\pi_0) = \beta_0 + \beta_1 x$$

for x, which yields

$$x(\pi_0) = \frac{\text{logit}(\pi_0) - \beta_0}{\beta_1}$$

Usually we won't know β, but we can estimate, giving an estimate of x of

$$\hat{x}(\pi_0) = \frac{\text{logit}(\pi_0) - \hat{\beta}_0}{\hat{\beta}_1}$$

For example

$$\begin{aligned}\hat{x}(0.25) &= \frac{\text{logit}(0.25) - (-5.3397)}{0.0015484} \\ &= \frac{\log \frac{0.25}{0.75} + 5.3397}{0.0015484} = 2738.961\end{aligned}$$

So we would expect as 25% failure rate to occur at around 2739 psi.

$$\begin{aligned}\hat{x}(0.5) &= \frac{\text{logit}(0.5) - (-5.3397)}{0.0015484} \\ &= \frac{0 + 5.3397}{0.0015484} = 3448.460\end{aligned}$$

A 50% failure rate is estimate to occur at 3448 psi.

In the case when we are counting successes, the level where we should get 50% successes is often referred to as the $ED50$ (Effective dose). Similarly, the $ED90$ would be the level where we would expect 90% successes.

In the case where are counting failures, particularly deaths, people often talk about LD (lethal dose) levels, particularly $LD50$s. This terminology is particularly common in toxicology.

Given that we can estimate $x(\pi_0)$ it would also be nice to get a confidence interval for it.

One approach would be to calculate the standard error of $\hat{x}(\pi_0)$ and to use the interval

$$\hat{x}(\pi_0) \pm z^*_{\alpha/2} SE(\hat{x}(\pi_0))$$

This can be done as the SE can be determined.

For simplicity, lets just consider the $ED50$ from now on as the formulas are easier to deal with. Let $\hat{x} = -\hat{\beta}_0/\hat{\beta}_1$ be the estimate of the ED50. It can be shown that

$$\mathrm{Var}(\hat{x}) \approx \frac{\nu_{00} - 2\hat{x}\nu_{01} + \hat{x}^2\nu_{11}}{\hat{\beta}_1^2}$$

where $\nu_{00} = \mathrm{Var}(\hat{\beta}_0)$, $\nu_{11} = \mathrm{Var}(\hat{\beta}_1)$, and $\nu_{01} = \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$.

However this particular interval tends not to work well as the asymptotic normality often isn't good. Instead another approach is more popular.

Fieller Intervals

Suppose we are interested in estimating the quantity $x_0 = -\beta_0/\beta_1$ where β_0 and β_1 are estimated by $\hat{\beta}_0$ and $\hat{\beta}_1$ and these estimates are assumed to be normally distributed with means β_0 and β_1, variances ν_{00} and ν_{11} and covariance ν_{01} (our setup, at least asymptotically).

Now consider the random variable

$$\psi = \hat{\beta}_0 + x_0\hat{\beta}_1$$

Then

$$E[\psi] = \beta_0 + x_0\beta_1 = \beta_0 + -\frac{\beta_0}{\beta_1}\beta_1 = 0$$

and

$$V(x_0) = \mathrm{Var}(\psi) = \nu_{00} + 2x_0\nu_{01} + x_0^2\nu_{11}$$

Then a confidence set for x_0 is given by the set of x satisfying

$$\left| \frac{\hat{\beta}_0 + x\hat{\beta}_1}{\sqrt{V(x)}} \right| \leq z^*_{\alpha/2}$$

or equivalently

$$(\hat{\beta}_0 + x\hat{\beta}_1)^2 \leq z^{*2}_{\alpha/2} V(x)$$

This involves solving the quadratic equation

$$(\hat{\beta}_1^2 - z^{*2}_{\alpha/2}\nu_{11})x^2 - (2\nu_{01}z^{*2}_{\alpha/2} - 2\hat{\beta}_0\hat{\beta}_1)x + \hat{\beta}_0^2 - \nu_{00}z^{*2}_{\alpha/2} = 0$$

For the fastener example, a 95% confidence interval is (3342.314, 3556.656) (the estimate is 3448.460)

```
> fieller(coef(fasten.glm), vcov(fasten.glm))
fasten.load fasten.load
   3342.314    3556.656

# this is a function I wrote since writing out the math is ugly
```

This procedure won't necessarily give an interval. The result could be an interval, a semi-infinite interval, or the complement of an interval. In fact it may not give an interval (I think).

For example their won't be a value of x satisfying

$$(\hat{\beta}_0 + x\hat{\beta}_1)^2 \leq z^{*2}_{\alpha/2} V(x)$$

if

$$(2\nu_{01} z^{*2}_{\alpha/2} - 2\hat{\beta}_0\hat{\beta}_1)^2 < 4(\hat{\beta}_1^2 - z^{*2}_{\alpha/2}\nu_{11})(\hat{\beta}_0^2 - \nu_{00} z^{*2}_{\alpha/2})$$

(essentially no interval)

In addition, if

$$\hat{\beta}_1^2 - z_{\alpha/2}^{*2}\nu_{11} < 0$$

the values satisfying

$$(\hat{\beta}_0 + x\hat{\beta}_1)^2 \leq z_{\alpha/2}^{*2}V(x)$$

will satisfy $(-\infty, L) \cup (U, \infty)$ where L and U are the roots of the quadratic equation to be solved.

However if $\hat{\beta}_1$ is large relative to $SE(\hat{\beta})$, then this procedure should give a narrow interval for $x(\pi_0)$. Intuitively this makes sense, as large changes in x will lead to large changes in $\pi(x)$.

In the general case of confidence sets for $x(\pi_0)$, the procedure gets changed to finding xs satisfying

$$\left| \frac{\hat{\beta}_0 + x\hat{\beta}_1 - \text{logit}(\pi_0)}{\sqrt{V(x)}} \right| \leq z_{\alpha/2}^*$$

Comparing Linear and Logistic Regression

So far we've seen two different types of regression, linear regression (General Linear Model), and Logistic Regression. Lets look at the similarity between the two settings

Feature	Linear Regression	Logistic Regression
Random Component	$Y_i \vert X_i \overset{ind}{\sim} N(\mu(X), \sigma^2)$	$Y_i \vert X_i \overset{ind}{\sim} Bin(1, \pi(X_i))$
Systematic Component	$\eta(X_i) = X_i\beta$	$\eta(X_i) = X_i\beta$
Link	$\mu(Y_i \vert X_i) = \eta(X_i)$	$\text{logit}(\mu(Y_i \vert X_i)) = \eta(X_i)$

In both settings we have a distributional assumption about the response variable, a linear predictor involving covariates, and a relationship between the mean of the distribution and the linear predictor.

Generalized Linear Model

We can extend this comparison to a more general situation, giving the Generalize Linear Model. This involves the following 4 pieces.

1. Distribution: What is the distribution of the response variable y. Often taken to be a member of the exponential family.

2. Linear predictor: $\eta = X\beta$

3. Link function $g(\cdot)$: Relates the linear predictor to the mean of the outcome variable

$$g(\mu) = \eta = X\beta \qquad \mu = g^{-1}(\eta) = g^{-1}(X\beta)$$

   $g(x)$ needs to be a continuous, monotonic function of x.

In logistic regression we have

$$\text{logit}(\mu) = \log \frac{\mu}{1-\mu} = \eta$$

So the logit function is the link function in this case.

4. Dispersion parameter ϕ: Some distributions have an additional parameter dealing with the the spread of the distribution. The form of this usually depends on the relationship between the mean and the variance. With some distributions, this is fixed (e.g. Poisson or binomial), while with others it is an additional parameter to the modeled and estimated (e.g. normal or gamma).

Exponential Family of Distributions

While the earlier structure can be used in many situations, this structure works well for a particular class of distributions known as the exponential family.

The class includes the normal (Gaussian), binomial, Poisson, Gamma, Hypergeometric, and Inverse Gaussian.

Distributions in this class have a density (or mass) function of the form

$$f(y;\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\}$$

As we will see soon, the parameter θ (known as the canonical parameter) relates to the mean (and higher moments) of the distribution, and ϕ relates to the dispersion (variance) of the distribution.

For example, the normal distribution belongs to this class as

$$f(y;\theta,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\}$$
$$= \exp\{(y\mu - \mu^2/2)/\sigma^2 - (y^2/\sigma^2 + \log(2\pi\sigma^2))/2\}$$

so that $\theta = \mu$ and $\phi = \sigma^2$ and

$$a(\phi) = \phi, \qquad b(\theta) = \theta^2/2, \qquad c(y,\phi) = (y^2/\sigma^2 + \log(2\pi\sigma^2))/2$$

Distributions in this class have similar forms for their means and variances. It can be shown that

$$\mu(Y) = b'(\theta) \qquad \mathrm{Var}(Y) = b''(\theta)a(\phi)$$

The function $b(\theta)$ is the cumulant function (related to the log of the moment generating / characteristic function)

The function $b''(\theta)$ is known as the variance function. This depends on the canonical parameter (and thus the mean). This function, considered as a function of μ, will be written $V(\mu)$.

This can be proven based on the standard results (under certain regularity conditions)

$$E\left[\frac{\partial l}{\partial \theta}\right] = 0$$

and

$$E\left[\frac{\partial^2 l}{\partial \theta^2}\right] + E\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = 0$$

where $l(\theta, \phi; y) = \log f(y; \theta, \phi)$ is the log-likelihood function.

The function $a(\phi)$ is often of the form

$$a(\phi) = \frac{\phi}{w}$$

where ϕ, also denoted by σ^2 and called the dispersion parameter, is constant over observations and w is a known *prior weight* that varies from observation to observation.

For example, if observations are the average of m iid normal observations

$$a(\phi) = \frac{\sigma^2}{m}$$

	Normal	Poisson	Binomial	Gamma	Inverse Gaussian
Notation	$N(\mu, \sigma^2)$	$P(\mu)$	$Bin(m, \pi)/m$	$G(\mu, \nu)$	$IG(\mu, \sigma^2)$
Range of y	$(-\infty, \infty)$	$0, 1, 2, \ldots$	$0, \frac{1}{m}, \frac{2}{m}, \ldots, 1$	$(0, \infty)$	$(0, \infty)$
ϕ	σ^2	1	$\frac{1}{m}$	$\frac{1}{\nu}$	σ^2
$b(\theta)$	$\frac{\theta^2}{2}$	e^θ	$\log(1 + e^\theta)$	$-\log(-\theta)$	$-\sqrt{-2\theta}$
$\mu(\theta)$	θ	e^θ	$\frac{e^\theta}{1+e^\theta}$	$\frac{-1}{\theta}$	$\frac{-1}{\sqrt{-2\theta}}$
$\theta(\mu)$	identity	log	logit	reciprocal	$\frac{1}{\mu^2}$
$V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2	μ^3

The function $\mu(\theta)$ is the inverse canonical link.

The function $\theta(\mu)$ is known as the canonical link.

Link Functions

Changing the link functions allows for different relationships between the response and predictor variables. The choice of link function $g(\cdot)$ should be made so that the relationship between the transformed mean and the predictor variables is linear.

Note transforming the mean via the link function is different from transforming the data by the same function.

You will end up with different models, except in special situations, usually involving linear transformations.

For example consider the two models

1. Transform data: $E[\log Y_i|X_i] = X_i\beta$

   $\log Y_i|X_i, \beta \sim N(X_i\beta, \sigma^2)$ or equivalently $Y_i|X_i, \beta \sim logN(X_i\beta, \sigma^2)$

$$E[Y_i|X_i, \beta] = \exp\left(X_i\beta + \frac{\sigma^2}{2}\right)$$

   and

$$\text{Var}(Y_i|X_i, \beta) = \exp(2(X_i\beta + \sigma^2))(\exp(\sigma^2) - 1)$$

2. Transform mean: $\log \mu(Y_i|X_i) = X_i\beta$

   $Y_i|X_i, \beta \sim N(\mu_i, \sigma^2)$ where $\log \mu_i = X_i\beta, \mu_i = \exp(X_i\beta)$ (normal model with log link)

$$E[Y_i|X_i, \beta] = \exp(X_i\beta)$$

   and

$$\text{Var}(Y_i|X_i, \beta) = \sigma^2$$

The first model has a different mean and the variability depends on X where as the variability in the second model does not depend on X.

When choosing a link function, you often need to consider the plausible values of the mean of the distribution.

For example, with binomial data, the success probability must be in [0,1]. However $X\beta$ can take values on $(-\infty, \infty)$.

Thus you can get into trouble with binomial data with the model $\mu = X\beta$ (identity link).

Possible choices include

- Logit link (logistic regression):

$$g(\mu) = \log \frac{\mu}{1-\mu}$$

- Probit link (probit regression):

$$g(\mu) = \Phi^{-1}(\mu) \qquad (\text{Standard Normal Inverse CDF})$$

- Complementary Log-Log link

$$g(\mu) = \log(-\log(\mu))$$

All of these happen to be quantile functions for different distributions.

Thus the inverse link functions are CDFs

- Logit link:

$$g^{-1}(\eta) = \frac{e^{\eta}}{1 + e^{\eta}} \qquad (\text{Standard Logistic})$$

- Probit link:

$$g^{-1}(\eta) = \Phi(\eta) \qquad (N(0, 1))$$

- Complementary Log-Log link:

$$g^{-1}(\eta) = e^{-e^{\eta}} \qquad (\text{Gumbel})$$

Thus in this case any distribution defined on $(-\infty, \infty)$ could be the basis for a link function, but these are the popular ones. One other choice that is used are based on t_ν distributions as they have some robustness properties.

Note that a link function doesn't have to have the property of mapping the range of the mean to $(-\infty, \infty)$. For example, we used it (sort of), in the soda bottle return example, though in that case it doesn't work well. Lets do it better by fitting a model corresponding to

$$\mu(Y_i/m|X_i) = \beta_0 + \beta_1 X_i = \mu_i$$
$$\mathrm{Var}(Y_i/m|X_i) = \phi\frac{\mu_i(1-\mu_i)}{m}$$

Trying to get $\phi = 1$ is a bit of work, so we aren't quite fitting a binomial model with the identity link.

(This is an example of a quasi-likelihood model.)

```
> return.ident.glm <- glm(ret.prop[,1] ~ deposit,
+    family=quasi(link="identity", variance="mu(1-mu)"))
> summary(return.ident.glm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0771708  0.0194416   3.969   0.0165 *
deposit     0.0275962  0.0009692  28.473 9.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasi family taken to be 0.005050623)

    Null deviance: 2.216343  on 5  degrees of freedom
Residual deviance: 0.020666  on 4  degrees of freedom
```

So the estimated probabilities of return are invalid when the deposit is outside the interval (2.80, 33.44) which is matches well with the range of the data.

In the binomial case, it can be reasonable if the success probabilities lie in the range (0.2, 0.8) for the levels of the predictor variables of interest.

For the example, the observed proportions range from 0.144 to 0.898, which goes outside this range.

Similarly, an inverse link function doesn't have to have to map $X\beta$ back to the whole range of the mean for a distribution.

For example, the log link will only give positive means ($\mu = e^{\eta}$). This can be an useful model with normal data, even though in general a normal mean can take any value.

Common Link Functions

The following are common link function choices for different distributions (all available in **R**).

- Normal (**R** calls this `gaussian`)
  - Identity: $g(\mu) = \mu$
  - Log: $g(\mu) = \log \mu$
  - Inverse: $g(\mu) = \frac{1}{\mu}$
- Binomial
  - Logit: $g(\mu) = \log \frac{\mu}{1-\mu} = \text{logit}(\mu)$
  - Probit: $g(\mu) = \Phi^{-1}(\mu)$
  - Cauchit: $g(\mu) = \tan(\pi(\mu - 1/2))$
  - Complementary Log-Log link: $g(\mu) = \log(-\log(1-\mu))$
  - Log: $g(\mu) = \log \mu$

- Poisson
  - Log: $g(\mu) = \log \mu$
  - Identity: $g(\mu) = \mu$
  - Square root: $g(\mu) = \sqrt{\mu}$
- Gamma
  - Inverse: $g(\mu) = \frac{1}{\mu}$
  - Log: $g(\mu) = \log \mu$
  - Identity: $g(\mu) = \mu$
- Inv-Normal
  - Inverse squared: $g(\mu) = \frac{1}{\mu^2}$
  - Inverse: $g(\mu) = \frac{1}{\mu}$
  - Log: $g(\mu) = \log \mu$
  - Identity: $g(\mu) = \mu$

The first link function mentioned for each distribution is the canonical link.

This is the link function that sets the transformed mean to the canonical parameter (i.e. $g(\mu) = \theta$)

So for the binomial setting

$$\text{logit}\left(\frac{e^{\theta}}{1+e^{\theta}}\right) = \theta$$

Dispersion Parameter

So far we have only discussed the mean function. However we also need to consider the variability of the data as well. For any distribution, we can consider the variance to be a function of the mean ($V(\mu)$) and a dispersion parameter (ϕ)

$$\text{Var}(Y) = \phi V(\mu)$$

The variance functions and dispersion parameters for the common distributions are

Distribution	$N(\mu, \sigma^2)$	$P(\mu)$	$Bin(m, \mu)/m$	$Gamma(\mu, \nu)$
$V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2
ϕ	σ^2	1	$\frac{1}{m}$	$\frac{1}{\nu}$

Note for the Gamma distribution, the form of these can depend on how the distribution is parameterized. In this case $\frac{1}{\nu}$ is the square of the coefficient of variation or the usual shape parameter and μ is the mean.

So when building models we need models for dealing with the dispersion in the data. Exactly how you want to do this will depend on the problem.

For now, for the normal, gamma, and inverse Gaussian we will estimate the dispersion parameter and for the binomial and Poisson we will treat it as a fixed, known constant.

Later we will look at binomial and Poisson cases where we will estimate a dispersion parameter.

(Actually the deposit example I fit earlier was an example of this.)