# Fitting Generalized Linear Models

Last time, we introduced the elements of the GLIM:

- The response $y$, with distribution

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

  where $\theta$ is the canonical parameter. from which we showed that $E(y) = \mu = b'(\theta)$ and $\text{Var}(y) = a(\phi)b''(\theta)$.

- The linear predictor

$$\eta = x^T\beta,$$

  where $x$ is a vector of covariates and $\beta$ is to be estimated.

- The link function, which connects $\eta$ to $\mu$,

$$g(\mu) = \eta.$$

In this notation, the subscript $i$ has been suppressed.

In many cases, $a(\phi)$ will have the form

$$a(\phi) = \phi/w,$$

where $\phi$ is the dispersion parameter and $w$ is a known weight.

**Example: normal response.** Under the normal model $y \sim N(\mu, \sigma^2)$, the log-density is

$$\begin{aligned} \log f &= -\frac{1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2}\log\left(2\pi\sigma^2\right) \\ &= \frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right). \end{aligned}$$

Therefore, the canonical parameter is $\theta = \mu$, and the remaining elements are $b(\theta) = \mu^2/2$ and $\phi = \sigma^2$, $a(\phi) = \phi$.

In a heteroscedastic model $y \sim N(\mu, \sigma^2/w)$, where $w$ is a known weight, we would have $\phi = \sigma^2$ and $a(\phi) = \phi/w$.

**Example: binomial response.** If $y \sim n^{-1}\text{Bin}(n, \pi)$, then the log-probability mass function is

$$\begin{aligned} \log f &= \log n! - \log(ny)! - \log(n(1-y))! \\ &\quad + ny\log\pi + n(1-y)\log(1-\pi) \end{aligned}$$

$$= \frac{y \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)}{1/n} + c,$$

where $c$ doesn't involve $\pi$. Therefore, the canonical parameter is

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

the $b$-function is

$$b(\theta) = -\log(1-\pi) = \log\left(1+e^{\theta}\right),$$

the dispersion parameter is $\phi = 1$, and $a(\phi) = \phi/n$.

**Canonical link.** We showed that the canonical parameter for $y \sim N(\mu, \sigma^2)$ is $\theta = \mu$, and the canonical parameter for $ny \sim \text{Bin}(n, \pi)$ is $\theta = \text{logit}(\pi)$. Notice that the most commonly used link function for a normal model is $\eta = \mu$, and the most commonly used link function for the binomial model us $\eta = \text{logit}(\pi)$. This is no accident. When $\eta = \theta$, we say that the model has a canonical link. Canonical links have some nice properties, which we will discuss.

It is often convenient to use a canonical link. But convenience does not imply that the data actually conform to it. It will be important to check the

appropriateness of the link through diagnostics, whether or not the link is canonical.

**Fitting generalized linear models via Fisher scoring.** ML estimation for $\beta$ may be carried out via Fisher scoring,

$$\beta^{(t+1)} = \beta^{(t)} + \left[-E\, l''(\beta^{(t)})\right]^{-1} l'(\beta^{(t)}),$$

where $l$ is the loglikelihood function for the entire sample $y_1, \ldots, y_N$.

Temporarily changing the notation, we will now let $l$, $l'$ and $l''$ denote the contribution of a single observation $y_i = y$ to the loglikelihood and its derivatives. We do this for simplicity, understanding that the corresponding functions for the entire sample are obtained by summing the contributions over the sample units $i = 1, \ldots, N$.

Ignoring constants, the loglikelihood is

$$l(\theta; y) = \frac{y\theta - b(\theta)}{a(\phi)}.$$

The contribution of $y = y_i$ to the $j$th element of the

score vector is

$$\frac{\partial l}{\partial \beta_j} = \left(\frac{\partial l}{\partial \theta}\right)\left(\frac{\partial \theta}{\partial \mu}\right)\left(\frac{\partial \mu}{\partial \eta}\right)\left(\frac{\partial \eta}{\partial \beta_j}\right).$$

The first factor is

$$\frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} = \frac{y-\mu}{a(\phi)}.$$

Because $\mu = b'(\theta)$, the second factor is

$$\frac{\partial \theta}{\partial \mu} = \frac{1}{b''(\theta)} = \frac{1}{V(\mu)} = \frac{a(\phi)}{\mathrm{Var}(y)},$$

where $V(\mu) = b''(\theta)$ is the variance function that we discussed last time.

The third factor, $\partial\mu/\partial\eta$, will depend on the link function. The fourth factor is $\partial\eta/\beta_j = x_{ij}$, where $x_{ij}$ is the $j$th element of the covariate vector $x_i = x$ for the $i$th observation. Putting it all together, we have

$$\frac{\partial l}{\partial \beta_j} = \frac{y-\mu}{\mathrm{Var}(y)}\left(\frac{\partial \mu}{\partial \eta}\right) x_{ij}.$$

If we are using the canonical link $\eta = \theta$, then $\partial\mu/\partial\eta = \partial\mu/\partial\theta = b''(\theta)$, so the score becomes

$$\frac{\partial l}{\partial \beta_j} = \frac{y-\mu}{\mathrm{Var}(y)}\, b''(\theta)\, x_{ij} = \frac{y-\mu}{a(\phi)}\, x_{ij}.$$

To find the expected second derivatives, we can use

the property

$$\begin{aligned} -E\left(\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right) &= E\left[\left(\frac{\partial l}{\partial\beta_j}\right)\left(\frac{\partial l}{\partial\beta_k}\right)\right] \\ &= E\left(\frac{y-\mu}{\mathrm{Var}(y)}\right)^2\left(\frac{\partial\mu}{\partial\eta}\right)^2 x_{ij}x_{ik} \\ &= \frac{1}{\mathrm{Var}(y)}\left(\frac{\partial\mu}{\partial\eta}\right)^2 x_{ij}x_{ik}. \end{aligned}$$

With the canonical link, this becomes

$$E\left(\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right) = -\frac{b''(\theta)}{a(\phi)}\,x_{ij}x_{ik}.$$

But under the canonical link, the *actual* second derivative is

$$\begin{aligned} \frac{\partial^2 l}{\partial\beta_j\partial\beta_k} &= \frac{\partial}{\partial\beta_k}\left(\frac{\partial l}{\partial\theta}\right)\left(\frac{\partial\theta}{\partial\beta_j}\right) \\ &= \left(\frac{\partial l}{\partial\theta}\right)\left(\frac{\partial^2\theta}{\partial\beta_j\partial\beta_k}\right) + \left(\frac{\partial\theta}{\partial\beta_j}\right)\left(\frac{\partial^2 l}{\partial\theta^2}\right)\left(\frac{\partial\theta}{\partial\beta_k}\right). \\ &= 0 + \left(\frac{\partial^2 l}{\partial\theta^2}\right)x_{ij}x_{ik}. \end{aligned}$$

Also, we saw last time that

$$\frac{\partial^2 l}{\partial\theta^2} = -\frac{b''(\theta)}{a(\phi)},$$

so with the canonical link, the actual second

derivatives equal the observed second derivatives, and Fisher scoring is the same thing as Newton-Raphson. Under the canonical link, the second derivatives are constant with respect to the data $y$.

**Putting it together.** For an arbitrary link, we have just shown that

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \frac{y-\mu}{\mathrm{Var}(y)}\left(\frac{\partial \mu}{\partial \eta}\right) x_{ij}, \\ -E\left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right) &= \frac{1}{\mathrm{Var}(y)}\left(\frac{\partial \mu}{\partial \eta}\right)^2 x_{ij}x_{ik}. \end{aligned}$$

It follows that the score vector for the entire data set $y_1, \ldots, y_N$ can be written as

$$\frac{\partial l}{\partial \beta} = X^T A(y-\mu),$$

where $X = (x_1, \ldots, x_N)^T$,

$$\begin{aligned} A &= \mathrm{Diag}\left[\left[\mathrm{Var}(y_i)\right]^{-1}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right], \\ &= \mathrm{Diag}\left[\mathrm{Var}(y_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)\right]^{-1}, \end{aligned}$$

and $y$ and $\mu$ now denote the entire vectors

$$y = (y_1, \ldots, y_N)^T,$$

$$\mu \quad = \quad (\mu_1, \ldots, \mu_N)^T.$$

The expected Hessian matrix becomes

$$-E\left(\frac{\partial^2 l}{\partial\beta\partial\beta^T}\right) = X^T W X,$$

where

$$\begin{aligned} W &= \text{Diag}\left[\left[\text{Var}(y_i)\right]^{-1}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right] \\ &= \text{Diag}\left[\text{Var}(y_i)\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2\right]^{-1}. \end{aligned}$$

An iteration of Fisher scoring is then

$$\beta^{(t+1)} = \beta^{(t)} + \left(X^T W X\right)^{-1} X^T A\left(y-\mu\right),$$

where $W$, $A$ and $\mu$ are calculated from $\beta^{(t)}$.

**Iteratively reweighted least squares.** Recall that a heteroscedastic normal model is fit by weighted least squares (WLS),

$$\hat{\beta} = \left(X^T W X\right)^{-1} X^T W y,$$

where $y$ is the response and $W$ is the diagonal matrix of weights, which is equivalent to OLS regression of

$W^{1/2}y$ on $W^{1/2}X$.

We can arrange the step of Fisher scoring to make it resemble WLS. First, rewrite it as

$$\beta^{(t+1)} = \left(X^T W X\right)^{-1}\left[X^T W X\beta^{(t)} + X^T A\left(y-\mu\right)\right].$$

Then note that $X\beta = (\eta_1, \ldots, \eta_N)^T = \eta$. Also, note that $A$ and $W$ are related by

$$A = W\left(\frac{\partial \eta}{\partial \mu}\right),$$

where $\partial\eta/\partial\mu = \text{Diag}(\partial\eta_i/\partial\mu_i)$. Therefore, we can write it as

$$\beta^{(t+1)} = \left(X^T W X\right)^{-1} X^T W z,$$

where

$$\begin{aligned} z &= \eta + \left(\frac{\partial \eta}{\partial \mu}\right)(y-\mu) \\ &= (z_1, \ldots, z_N)^T, \end{aligned}$$

where $z_i = \eta_i + (\partial\eta_i/\partial\mu_i)(y_i - \mu_i)$. In the GLIM literature, $z_i$ is often called the **adjusted dependent variate** or the **working variate**. Fisher scoring can therefore be regarded as iteratively reweighted least squares (IRWLS) carried out on a transformed version

of the response variable.

At each cycle, we

- use the current estimate for $\beta$ to calculate a new working variate $z$ and a new set of weights $W$, and then
- regress $z$ on $X$ using weights $W$ to get the updated $\beta$.

Viewing Fisher scoring as IRWLS makes it easy to program this algorithm as a macro in any statistical package (even Minitab!) capable of WLS.

Viewing Fisher scoring as IRWLS has an additional advantage: It provides an excellent basis for us to derive model-checking diagnostics. The diagnostics that are commonly used in regression—plotting residuals versus fitted values, leverage and influence measures, etc.—have obvious analogues in GLIM's when we view the fitting procedure as IRWLS.

**Covariance matrix.** The final value for $(X^TWX)^{-1}$ upon convergence is the estimated covariance matrix for $\hat{\beta}$. The diagonal elements of this matrix provide

the squared SE's for the estimated coefficients.

**What about the dispersion parameter?** Recall that the variance of $y_i$ is usually of the form $V(\mu_i)\phi/w_i$, where $V$ is the variance function, $\phi$ is the dispersion parameter, and $w_i$ is a known weight. In this case, $\phi$ cancels out of the IRWLS procedure and $\hat{\beta}$ itself is the same under any assumed value for $\phi$. So, we could actually remove $\phi$ from the $W$ matrix. But we have to be careful, because the assumed value for $\phi$ must be put back in to get a correct estimated covariance matrix for $\hat{\beta}$.

**Example: Normal regression.** Suppose that $y_i \sim N(\mu_i, \sigma^2/w_i)$ where $w_i$ is a known weight, and let $\eta_i = g(\mu_i) = x_i^T\beta$ for some link function $g$. In this case, $\phi = \sigma^2$ and the variance function is constant. If we use a log link,

$$\log \mu_i \;=\; x_i^T\beta,$$

then $\partial\eta_i/\partial\mu_i = 1/\mu_i$, the weight matrix is

$$W \;=\; \text{Diag}\left[\frac{w_i\mu_i^2}{\sigma^2}\right],$$

and the working variate is

$$\begin{aligned} z_i &= \eta_i + \frac{y_i - \mu_i}{\mu_i} \\ &= x_i^T\beta + \frac{y_i - \exp(x_i^T\beta)}{\exp(x_i^T\beta)}. \end{aligned}$$

We do not need to assume anything about $\sigma^2$ to find $\hat{\beta}$, but we do need an estimate to get a covariance matrix for $\hat{\beta}$. The traditional estimate would be

$$\hat{\sigma^2} = \frac{1}{N-p}\sum_{i=1}^{N} w_i \left(y_i - \hat{\mu}_i\right)^2,$$

where $\hat{\mu}_i = \exp(x_i^T\hat{\beta})$. This is not exactly unbiased, nor is it the ML estimate (the ML estimate uses $N$ in the denominator).

If we use the identity link $\mu_i = x_i^T\beta$, then $\partial\eta_i/\partial\mu_i = 1$, $W = \text{Diag}(w_i\sigma^{-2})$, and $z_i = y_i$. Neither $z_i$ nor $W$ depends on the current estimate of $\beta$, and the procedure reduces to a single iteration of WLS. In this case,

$$\hat{\sigma^2} = \frac{1}{N-p}\sum_{i=1}^{N} w_i \left(y_i - \hat{\mu}_i\right)^2$$

with $\hat{\mu}_i = x_i^T \hat{\beta}$ is exactly unbiased.

**Example: Binomial regression.** In previous lectures, we described the ML fitting procedure for logistic regression and binomial models with arbitrary link functions. Now let's re-create the procedure using our new notation for GLIM's.

If $ny_i \sim \text{Bin}(n_i, \mu_i)$, then

$$\begin{aligned} \text{Var}(y_i) &= \frac{\mu_i(1-\mu_i)}{n_i} \\ &= \frac{\phi}{n_i}\,\mu_i(1-\mu_i) \end{aligned}$$

for $\phi = 1$ (no over- or underdispersion). The variance function is

$$V(\mu_i) = \mu_i(1-\mu_i).$$

Under a logit link

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^T\beta,$$

we have

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i(1-\mu_i)},$$

the weight matrix becomes

$$W = \text{Diag}\left[\, n_i\mu_i(1-\mu_i)\,\right],$$

and the working variate is

$$\begin{aligned} z_i &= \eta_i + \frac{y_i - \mu_i}{\mu_i(1-\mu_i)} \\ &= x_i^T\beta + \frac{y_i - \text{expit}(x_i^T\beta)}{\text{expit}(x_i^T\beta)(\,1 - \text{expit}(x_i^T\beta)\,)}. \end{aligned}$$

Notice that in this new notation, $y_i$ is the observed *proportion* of successes in $n_i$ trials, rather than the actual number of successes.

**Generalized linear modeling software.**
Generalized linear modeling is now a standard part of modern statistical packages. In R, the relevant function is `glm()`. In SAS, you can use PROC GENMOD. As with ordinary linear regression software, you need to declare the response variable $y$ and the predictors $x$. In addition, however, you also need to declare the **distributional family.** The most common distributional families are normal (Gaussian), binomial, and Poisson. When you select the distributional family, you are actually selecting the variance function. After selecting the family, you also need to select the **link function**.

If you do not explicitly choose a link function, the

software will, by default, use the canonical link for the given distributional family. If you do not specify a family, the software will use the Gaussian or normal family. Therefore, the software will, by default, fit a normal linear regression.

In some cases, you will also need to specify **weights**. The weights $w_i$. These are known factors that are inversely proportional to the variance of $y$. In a binomial model, the $n_i$'s will be the weights. If no weights are given, the software will assume that all weights are 1.

Here's an example of how to use the `glm()` function in R.

```
> ########################################################################
> # Example: Logistic regression in R using glm()
> #
> # Enter the Berkeley graduate admission data
>
> dept <- c("A","A","B","B","C","C","D","D","E","E","F","F")
> sex  <- c("M","F","M","F","M","F","M","F","M","F","M","F")
> accept <- c(512, 89, 353, 17, 120, 202, 139, 131,  53,  94,  22,  24)
> reject <- c(313, 19, 207,  8, 205, 391, 278, 244, 138, 299, 351, 317)
>
> # Define the response as the proportion of successes
> n <- accept + reject
> y <- accept/n
>
>
> # change dept and sex to factors
> dept <- factor(dept)
> sex <- factor(sex)
```

```
>
> # use the contrasts() function to see what effects will be created
> contrasts(dept)
  B C D E F
A 0 0 0 0 0
B 1 0 0 0 0
C 0 1 0 0 0
D 0 0 1 0 0
E 0 0 0 1 0
F 0 0 0 0 1
> contrasts(sex)
  M
F 0
M 1
>
>
> # fit the model with main effects only
> result <- glm( y ~ dept + sex, family=binomial(link="logit"),
+    weights=n)
> summary(result)

Call:
glm(formula = y ~ dept + sex, family = binomial(link = "logit"),
    weights = n)

Deviance Residuals:
      1        2        3        4        5        6        7        8
-1.2536   3.7319  -0.0575   0.2777   1.2357  -0.9116   0.1180  -0.1227
      9       10       11       12
 1.2076  -0.8424  -0.2148   0.2125

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.67913    0.09908   6.854 7.18e-12 ***
deptB       -0.04362    0.10984  -0.397    0.691
deptC       -1.26090    0.10661 -11.827  < 2e-16 ***
deptD       -1.28782    0.10576 -12.177  < 2e-16 ***
deptE       -1.73751    0.12609 -13.780  < 2e-16 ***
deptF       -3.30527    0.16997 -19.447  < 2e-16 ***
sexM        -0.09673    0.08081  -1.197    0.231
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```
    Null deviance: 876.572  on 11  degrees of freedom
Residual deviance:  20.225  on  5  degrees of freedom
AIC: 103.17

Number of Fisher Scoring iterations: 4

>
>
> # now fit the final model from Lecture 13
> deptA <- 1*(dept=="A")
> deptB <- 1*(dept=="B")
> deptC <- 1*(dept=="C")
> deptD <- 1*(dept=="D")
> deptE <- 1*(dept=="E")
> deptF <- 1*(dept=="F")
> deptA.male <- 1*( (dept=="A")&(sex=="M") )
>
> # Note: the "-1" notation removes the intercept
> result <- glm( y ~ -1 + deptA + deptB + deptC + deptD +
+     deptE + deptF + deptA.male,
+     family=binomial(link="logit"),  weights=n)
> summary(result)

Call:
glm(formula = y ~ -1 + deptA + deptB + deptC + deptD + deptE +
    deptF + deptA.male, family = binomial(link = "logit"), weights = n)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.0000   0.0000  -0.1041   0.4978   0.6950  -0.5177  -0.3270   0.3435
      9       10       11       12
 0.8120  -0.5754  -0.4341   0.4418

Coefficients:
           Estimate Std. Error z value Pr(>|z|)
deptA       1.54420    0.25272   6.110 9.94e-10 ***
deptB       0.54286    0.08575   6.330 2.44e-10 ***
deptC      -0.61569    0.06916  -8.902  < 2e-16 ***
deptD      -0.65925    0.07496  -8.794  < 2e-16 ***
deptE      -1.08950    0.09535 -11.427  < 2e-16 ***
deptF      -2.67565    0.15243 -17.553  < 2e-16 ***
deptA.male -1.05208    0.26271  -4.005 6.21e-05 ***
---
```

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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1105.6870  on 12  degrees of freedom
Residual deviance:    2.6085  on  5  degrees of freedom
AIC: 85.552

Number of Fisher Scoring iterations: 3
```

Now here's the same thing using PROC GENMOD. In the model statement, we use the "event/trial" syntax. Note that, by default, SAS creates effect codes for the CLASS variables.

```
options linesize=72;

data admissions;
   input dept $ sex $ reject accept;
   n = accept + reject;
   cards;
DeptA Male   313 512
DeptA Female  19  89
DeptB Male   207 353
DeptB Female   8  17
DeptC Male   205 120
DeptC Female 391 202
DeptD Male   278 139
DeptD Female 244 131
DeptE Male   138  53
DeptE Female 299  94
DeptF Male   351  22
DeptF Female 317  24
;

proc genmod data=admissions;
   class dept sex;
   model accept/n = dept sex / dist=binomial link=logit;
   run;
```

Relevant portions of the SAS output:

The GENMOD Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.ADMISSIONS |
| Distribution | Binomial |
| Link Function | Logit |
| Response Variable (Events) | accept |
| Response Variable (Trials) | n |

| | |
|---|---|
| Number of Observations Read | 12 |
| Number of Observations Used | 12 |
| Number of Events | 1756 |
| Number of Trials | 4526 |

Class Level Information

| Class | Levels | Values |
|---|---|---|
| dept | 6 | DeptA DeptB DeptC DeptD DeptE DeptF |
| sex | 2 | Female Male |

Criteria For Assessing Goodness Of Fit

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 5 | 20.2251 | 4.0450 |
| Scaled Deviance | 5 | 20.2251 | 4.0450 |
| Pearson Chi-Square | 5 | 18.8317 | 3.7663 |
| Scaled Pearson X2 | 5 | 18.8317 | 3.7663 |
| Log Likelihood | | -2594.4532 | |

Algorithm converged.

Analysis Of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| Intercept | | 1 | -2.7229 | 0.1577 | -3.0319 | -2.4138 | 298.19 |

Analysis Of Parameter Estimates

| Parameter | Pr > ChiSq |
|---|---|
| Intercept | <.0001 |

Analysis Of Parameter Estimates

| Parameter | | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square |
|---|---|---|---|---|---|---|---|
| dept | DeptA | 1 | 3.3053 | 0.1700 | 2.9721 | 3.6384 | 378.17 |
| dept | DeptB | 1 | 3.2616 | 0.1788 | 2.9113 | 3.6120 | 332.89 |
| dept | DeptC | 1 | 2.0444 | 0.1679 | 1.7153 | 2.3734 | 148.31 |
| dept | DeptD | 1 | 2.0174 | 0.1699 | 1.6845 | 2.3504 | 141.01 |
| dept | DeptE | 1 | 1.5678 | 0.1804 | 1.2141 | 1.9214 | 75.49 |
| dept | DeptF | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| sex | Female | 1 | 0.0967 | 0.0808 | -0.0617 | 0.2551 | 1.43 |
| sex | Male | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . |
| Scale | | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | |

Analysis Of Parameter Estimates

| Parameter | | Pr > ChiSq |
|---|---|---|
| dept | DeptA | <.0001 |
| dept | DeptB | <.0001 |
| dept | DeptC | <.0001 |
| dept | DeptD | <.0001 |
| dept | DeptE | <.0001 |
| dept | DeptF | . |
| sex | Female | 0.2313 |
| sex | Male | . |
| Scale | | |

NOTE: The scale parameter was held fixed.

**Diagnostics.** We have shown that the Fisher scoring algorithm for a GLIM can be written as IRWLS,

$$\beta^{(t+1)} = \left(X^T W X\right)^{-1} X^T W z,$$

where

$$W = \text{Diag}\left[\text{Var}(y_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2\right]^{-1}$$

is the matrix of weights and

$$z = \eta + \left(\frac{\partial \eta}{\partial \mu}\right)(y - \mu)$$

is the working variate. Now we will appeal to the interpretation as IRWLS to suggest diagnostic techniques to check the appropriateness of the model. This material is derived from Chapter 12 of McCullagh and Nelder (1989).

**Residuals.** The Pearson residual is defined as

$$r = \frac{y_i - \hat{\mu}}{\sqrt{\hat{\text{Var}}(y)}},$$

where $\hat{\mu}$ is the ML estimate for $\mu$, and

$$\hat{\text{Var}}(y) = a(\phi)V(\hat{\mu})$$

is the estimated variance of $y$.

If we write the deviance as $D = \sum_{i=1}^{N} d_i$ where $d_i$ is the contribution of the $i$th unit, then the deviance residual is

$$r \;=\; \text{sign}(y-\mu)\sqrt{d}.$$

For example, in a binomial model $y_i \sim \text{Bin}(n_i, \pi_i)$, the Pearson residual is

$$r_i \;=\; \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}},$$

and the deviance residual is $r_i = \text{sign}(y_i - n_i\hat{\pi}_i)\sqrt{d_i}$, where

$$d_i \;=\; 2\left\{ y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right\}.$$

(For computational purposes, interpret $0\log 0$ as 0.)

Deviance and the Pearson residuals behave something like the standardized residuals in linear regression. McCullagh and Nelder suggest that distributional properties of deviance residuals are a little closer to those of their linear regression counterparts, and they suggest using the deviance residuals in plots.

**Plotting residuals versus fitted values.** Plot the residuals on the vertical axis versus the linear predictor $\eta$ on the horizontal axis. As in linear

regression, we hope to see something like a "horizontal band" with mean $\approx 0$ and constant variance as we move from left to right.

- Curvature in the plot may be due to a wrong link function or the omission of a nonlinear (e.g. quadratic) term for an important covariate.
- Non-constancy of range suggests that the variance function may be incorrect.

For binary responses, this plot is not very informative; all the points will lie on two curves, one for $y = 0$ and the other for $y = 1$. However, the plot may still help us to find outliers (residuals greater than about 2 or 3 in the positive or negative direction).

**Plotting residuals versus individual covariates.** In the same way, we can also plot the residuals versus a single covariate. (If the model has only one predictor, this will be equivalent to the last plot.) Again, we hope to see something like a horizontal band. Curvature in this plot suggests that the $x$-variable in question ought to enter into the model in a nonlinear fashion—for example, we might add a quadratic term $x^2$ or try various transformations like

$\sqrt{x}$ or $\log x$.

**Absolute residuals versus fitted values.** Plotting $|r|$ versus the fitted values $\mu$ can reveal a problem with the variance function. If there is no trend, the variance function is probably okay. An increasing trend (positive slope) suggests that the variance function is increasing too slowly with the mean; for example, $V(\mu) = \mu$ might have to be replaced with $V(\mu) = \mu^2$. Within a particular parametric family (e.g., binomial or Poisson) we can't really change the variance function. However, we can with a quasilikelihood approach (we'll talk about that later).

What are the implications of an incorrect variance function? Recall that in OLS regression, heteroscedasticity has the following implications: the estimate $\hat{\beta}$ is still unbiased, but it is no longer efficient. For GLIM's the situation is similar.

If the variance function is **correct**, then

- $\hat{\beta}$ is asymptotically unbiased, normal and efficient, and
- the estimated covariance matrix for $\hat{\beta}$ from the Fisher scoring algorithm is a consistent estimate

of $\text{Var}(\hat{\beta})$.

In the variance function is **not correct**, then

- $\hat{\beta}$ is still asymptotically unbiased and normal, but
- $\hat{\beta}$ is not efficient, and
- the estimated covariance matrix for $\hat{\beta}$ is not consistent for $\text{Var}(\hat{\beta})$.

The last problem (inconsistency of the variance estimate) can be fixed by using the so-called **sandwich estimator**. We will learn about this later, when we talk more about quasilikelihood.