Lecture 6. Generalized Linear Models

Suzer-Gurtekin Adapted from Michael Elliott's and James Wagner's Lecture Notes

February 2025

Overview

Regression Models

We recently looked at logistic regression.

Noticed differences with linear regression.

But there were also similarities.

Can we find common features?

Regression Models

Linear models of the form:

$E(Y_i) = \mu_i = x_i^T \beta$; where $Y_i \sim N(\mu_i, \sigma^2)$

where the random variables Y_i independent are the basis of most analyses of continuous data

The transposed vector x_i^T represents the *i*th row of the design matrix X

Generalizations of these example to the relationship between a continuous response and several explanatory variables to the relationship between a continuous response and several explanatory variables (multiple regression) and comparisons of more than two means (analysis of variance) are also of this form

Regression Models

This form of linear models can be generalized:

1. Response variables have distributions other than Normal distribution (may even be categorical)
    1. Recognition of "nice" properties of the Normal distribution shared by a wider class of distributions" **Exponential Family of Distributions**
    2. Relationship between response and predictors don't need be of the simple linear form

       Extension of numerical methods to estimate parameters β from $E(Y_i) = \mu_i = x_i^T\beta$; where $Y_i \sim N(\mu_i, \sigma^2)$ to the situation where there is some nonlinear function relating $E(Y_i) = \mu_i$ to linear component, $x_i^T\beta$:

       $g(\mu_i) = x_i^T\beta$ **Link function**

Regression Models

What are the key assumptions of the Gaussian (normal) linear model?

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_p x_{pi} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$

1) Sample is a random sample from the population of interest.
2) Mean model is correctly specified (x_i could be polynomial or other transformation of underlying covariates of interest to allow for linearity in the β).
3) Error terms are independent.
4) Error terms have equal variances.
5) Error terms are normally distributed.

Regression Models

In practice, none of these assumptions are ever met exactly (Nester 1996, Applied Statistics):

- "All models are wrong"
- "All variables are correlated"
- "Variances are never equal"
- "No data are normally distributed," etc.

But assumptions may be met approximately so that assumptions are still useful, or some may be dropped with minimal effect on results or interpretation.

Regression Models

How can we extend classical linear regression to handle situations that violate assumptions?

- For continuous data that is non-normally distributed, use Box-Cox method to determine transformations of the form

$$Y' = \begin{cases} \left. Y^{\lambda} - 1 \right/ \lambda \left[\prod Y_i \right]^{(\lambda-1)/n} & \lambda \neq 0 \\ \ln Y \left[\prod Y_i \right]^{1/n} & \lambda = 0 \end{cases}$$

so that the resulting Y' is approximately normal.

Regression Models

How can we extend classical linear regression to handle situations that violate assumptions?

- For correlated or non-homoscedastic data, can use properties of the multivariate normal to model correlation structure

  For non-normal distributions, consider methods that will provide consistent estimates of β in the absence of independence

Regression Models

Ex: Repeated measures or clustered data

- Repeated observations from the same individual (T-cell counts in HIV-infected individuals)
- Observations sampled together from a unit where subjects within the unit are more alike than subjects across units (persons within a vehicle in a passenger vehicle crash, individuals within a geographic region)

What assumptions of the Gaussian linear model are violated for repeated measures data?

- Independence of observations

Regression Models

When might the Gaussian linear regression model assumptions be violated?

Ex: Binary data.

- Presence or absence of disease
- Reached threshold level of a continuous outcome

$Y_i = \{0,1\}$

$Y_i \mid X_i = x_i \sim \text{Bernoulli}(\pi_i),$

where $\pi_i = \pi_i(x_i)$, a function of covariates

Regression Models

What assumptions of the Gaussian linear model are violated for dichotomous outcomes?

- Normality
- Constant variance, since
  $Var(Y_i \mid x_i) = \pi_i(1-\pi_i) = E(Y_i \mid x_i)\left(1 - E(Y_i \mid x_i)\right)$
- Linear model: need to constrain $0 \le E(Y_i \mid x_i) \le 1$

Regression Models

Ex: Count data

- Rare outcomes
- Cell counts in tables

$$Y_i \mid X_i = x_i \sim POISSON(\mu_i),$$

where $\mu_i = \mu_i(x_i)$, a function of covariates

What assumptions of the Gaussian linear model are violated for count data?

- Normality (holds approximately as $\mu \to \infty$)
- Constant variance, since $Var(Y_i \mid x_i) = \mu_i = E(Y_i \mid x_i)$

Regression Models

How can we extend classical linear regression to handle situations that violate assumptions?

- For outcomes that are members of the exponential family, use e.f. distribution properties to determine transformations of expectations $\theta_i = E(Y_i \mid X_i = x_i)$ that allow predictors to be modeled linearly:
  - Normal: $\theta_i = x_i^T \beta$
  - Binary: $\ln \frac{\theta_i}{1-\theta_i} = x_i^T \beta$
  - Count: $\ln \theta_i = x_i^T \beta$

Regression Models

This form of linear models can be generalized:

1. Response variables have distributions other than Normal distribution (may even be categorical)
    1. Recognition of "nice" properties of the Normal distribution shared by a wider class of distributions" **Exponential Family of Distributions**
    2. Relationship between response and predictors don't need be of the simple linear form

       Extension of numerical methods to estimate parameters β from $E(Y_i) = \mu_i = x_i^T\beta$; where $Y_i \sim N(\mu_i, \sigma^2)$ to the situation where there is some nonlinear function relating $E(Y_i) = \mu_i$ to linear component, $x_i^T\beta$:

       $g(\mu_i) = x_i^T\beta$ **Link function**

Regression Models

This has been done!

Regression models reformulated as **generalized linear models**

Two requirements:

1. Distribution of Outcome Variable is in the Exponential Family
2. Link function

Exponential Family

Many common distributions are exponential family:

- Normal
- Binomial
- Poisson

Exponential Family

Exponential Family is any distribution that can be written in the following form:

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

In this setup, θ is a location parameter and ϕ is the scale or dispersion parameter.

Exponential Family

For example, we can rewrite the Normal distribution (on the left) in this exponential family form (on the right):

$$\frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] = exp\left[\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right)\right]$$

"Translate" the pieces:

$\theta = \mu$

$\phi = \sigma^2$

$a(\phi) = \phi$

$b(\theta) = \mu^2/2$

$c(y, \phi) = -(y^2/\phi + log(2\pi\phi))/2$

Exponential Family

We can do the same for the binomial distribution:

$$\binom{n}{y}\mu^y(1-\mu)^{n-y} = exp\left(ylog\frac{\mu}{1-\mu} + nlog(1-\mu)\right) + log\binom{n}{y}$$

$\theta = log\frac{\mu}{1-\mu}$
$b(\theta) = nlog(1 + exp\theta)$
$c(y, \phi) = log\binom{n}{y}$

Link Function

The other requirement is the existence of a function that links the outcome to a linear predictor:

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

Link Function

We saw this with logistic regression:

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

$$log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}$$

In this case, the link function is $\eta = g(\mu) = log\ \frac{\mu}{1-\mu}$

Link Function

We can back-transform with the inverse of this function:

$$y = g^{-1}(\eta)$$

We saw this in the logistic regression setting as:

$$Pr(Y_i = 1) = logit^{-1}\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}\right)$$

$$Pr(Y_i = 1) = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}\right)}} = \frac{e^{\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}\right)}}{1 + e^{\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}\right)}}$$

Link Function

Family	Link	Variance Function
Normal	$\eta = \mu$	1
Poisson	$\eta = log\mu$	μ
Binomial	$\eta = log(\mu/(1 - \mu))$	$\mu(1 - \mu)$

Fitting a GLM

The normal model can be fit using Ordinary Least Squares.

Not true for every type of generalized linear model.

Start from maximum likelihood.

Here, $a_i(\phi) = \phi/w_i$:

$$logL(\theta_i, \phi|y_i) = w_i \left[\frac{y_i\theta_i - b(\theta_i)}{\phi}\right] + c(y_i, \phi)$$

In the normal case, we can derive an estimator that maximizes this.
For other distributions, this is not possible.

Fitting a GLM

More often, model is fit using numerical optimization.

Newton-Raphson is an algorithm that will do this sort of optimization.

Can be equivalent to Iteratively Reweighted Least Squares (IRWLS).

Fitting a GLM

Start from the normal linear model $Y = X\beta + \varepsilon$.

Suppose $Var(Y) \propto f(\hat{\eta})$, where $\hat{y} = \hat{\eta} = X\hat{\beta}$.

We would do a weighted regression where the weights w_i are $\frac{1}{w_i} = f(\hat{\eta})$.

A procedure for fitting this model starts with $w_i = 1 \forall i$, estimates $\hat{\beta}$, use this estimate to compute new w_i, re-estimate $\hat{\beta}$, and so on until the estimates converge.

Fitting a GLM

Follow a similar procedure for other GLMs.

Regress $g(y)$ on X.

Use $Var(g(y))$ to form weights.

However, $Var(g(y))$ may not make sense for some GLMs (e.g. binomial).

Fitting a GLM

Need some method to linearize $g(y)$.

Let $\eta = g(\mu)$. Now do a one-step expansion:

$$\begin{aligned} g(y) &\approx g(\mu) + (y - \mu)g'(\mu) \\ &= \eta + (y - \mu)\frac{d\eta}{d\mu} \\ &\equiv z \end{aligned}$$

$$\hat{Var}(z) = \left(\frac{d\eta}{d\mu}\right)^2 Var(\hat{\mu})$$

Which gives us a weight for next step: $\hat{Var}(z) = \frac{1}{w}$

Fitting a GLM

This gives us a procedure, IRWLS, to estimate GLMs:

1. Set initial estimates $\hat{\eta}_0$ and $\hat{\mu}_0$.
2. Form the initial z: $z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\frac{d\eta}{d\mu}\,|_{\hat{\eta}_0}$
3. Form the initial w: $\frac{(1)}{w_0} = \left(\frac{d\eta}{d\mu}\right)^2 |_{\hat{\eta}_0}\, V(\hat{\mu}_0)$.
4. Re-estimate β to get $\hat{\eta}_1$.
5. Repeat steps 2, 3, and 4 until convergence.

Fitting a GLM

Let's apply this to an actual dataset. Bliss (1935) looked at number of insects dying at different levels of insecticide concentration. Adapted from Faraway (2006).

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

Is this a restrospective or a prospective design?

In-class exercise 1

	dead	alive
Conc > 0	73	47
Conc = 0	2	28

Consider this is a prospective design:

- Compute relative risk
- Comment on the odds ratio

Consider this is a retrospective (case-control) design:

- Compute relative risk
- Comment on the odds ratio

In-class exercise 1

	dead	alive
Conc > 0	73	47
Conc = 0	2	28

Consider this is a prospective design:

- Compute relative risk
- Comment on the odds ratio

Consider this is a retrospective (case-control) design:

- Compute relative risk
- Comment on the odds ratio

Fitting a GLM

Let's apply this to an actual dataset. Bliss (1935) looked at number of insects dying at different levels of insecticide concentration. Adapted from Faraway (2006).

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

If it is a retrospective design what element would you change it to make it a prospective design?
If it is a prospective design what element would you change it to make it a prospective design?

Notation

Table: General Classification of a **Sample** by Risk Factor and Disease Status

Risk Factor Classification	Disease Classification		
	+(present)	-(absent)	Total at Risk
+(present)	a	b	a+b
-(absent)	c	d	c+d
Total	a+c	b+d	n

Incidence

The **incidence proportion estimator is** the proportion of the sample that developed a condition during a specified time period. The following are estimators for the incidence proportion:

$$I_{Exposed} = \frac{a}{a+b}$$

$$I_{Unexposed} = \frac{c}{c+d}$$

$$I_{Pop} = \frac{a+c}{n}$$

Relative Risk

The relative risk is often used to compare the incidence proportions across groups:

$$RR = \frac{I_{Exposed}}{I_{Unexposed}} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Relative risk is sometimes also used to compare incidence rates or even prevalence.

Odds Ratio

In this new notation:

$$OR = \frac{\frac{a}{a+b} / \frac{b}{a+b}}{\frac{c}{c+d} / \frac{d}{c+d}} = \frac{ad}{bc}$$

Prospective Studies

For these studies, the following estimator is used for the incidence rate:

$$E\left\{\frac{a}{a+b}\right\} = \frac{A}{A+B}$$

Retrospective or Case-Control Studies

Under this design, $\frac{a}{a+b}$ is *not* an unbiased estimator the population incidence.

This should make sense as those are set by the *design*, and not by their rate of occurrence in the population.

Notation

Table: General Classification of a **Sample** by Risk Factor and Disease Status

Risk Factor Classification	Disease Classification		
	+(present)	-(absent)	Total at Risk
+(present)	a	b	a+b
-(absent)	c	d	c+d
Total	a+c	b+d	n

Retrospective or Case-Control Studies

We can estimate slightly different quantities:

$$E\left\{\frac{a}{a+c}\right\} = \frac{A}{A+C}$$

$$E\left\{\frac{b}{b+d}\right\} = \frac{B}{B+D}$$

For example, the proportion of persons with cancer (*cases*) that smoke. And the proportion of persons without cancer (*controls*) that smoke.

Retrospective or Case-Control Studies

If the sample sizes are large, then the estimated odds ratio

$$OR = \frac{\frac{a}{a+b} / \frac{b}{a+b}}{\frac{c}{c+d} / \frac{d}{c+d}} = \frac{ad}{bc}$$

is a *consistent* estimator.

We also know that for rare conditions, the odds ratio and relative risk are approximately equal. This gives us a way to estimate these quantities in case-control studies.

In class exercise 2

Convert the Bliss dataset to case level data
Fit a GLM and compare the regression coefficient estimates to those from the model that we fit
Upload dataset Bliss_caselevel_cont.csv
Visualize data
Fit a logistic regression on dead by concentration (CONC3-continous variable).
Print out the residual deviance
Compare the null model to the model with the single predictor "Conc3"
Add a square term for CONC3 into the model
Compare nested model to larger model
Make a recommendation between two models based on the analysis of deviance table
Examine the Pearson residuals
Examine the deviance residuals
Examine the working residuals
Compute the difference in the coefficients when each case is dropped

Fitting a GLM

First, let's look at the answer we get using logistic regression and the `glm` function.

```
> logitmod<-glm(cbind(dead,alive)~conc,family=binomial,data=bliss)
> summary(logitmod)$coef
             Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -2.323790  0.4178878 -5.560798 2.685438e-08
conc         1.161895  0.1814158  6.404598 1.507665e-10
```

The `glm` function is performing the procedure we will emulate.

Fitting a GLM

Starting values:

```
y<-bliss$dead/30
mu<-y
```

Then do the first iteration:

```
eta<-logit(mu)
z<-eta+(y-mu)/(mu*(1-mu))
w<-30*mu*(1-mu)
linmod<-lm(z~conc,weights=w,bliss)
```

Fitting a GLM

After the first iteration, coefficient estimates:

```
> coef(linmod)
(Intercept)        conc
  -2.302462    1.153587
```

Recall the logistic output:

```
> logitmod<-glm(cbind(dead,alive)~conc,family=binomial,data=bliss)
> summary(logitmod)$coef
             Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -2.323790  0.4178878 -5.560798 2.685438e-08
conc         1.161895  0.1814158  6.404598 1.507665e-10
```

Close, but now we have updated weights.

Fitting a GLM

Repeat for several iterations:

```
for (i in 1:5){
  eta<-linmod$fit
  mu<-ilogit(eta)
  z<-eta+(y-mu)/(mu*(1-mu))
  w<-30*mu*(1-mu)
  linmod<-lm(z~bliss$conc,weights=w)
  cat(i,coef(linmod),"\n")
}
```

Fitting a GLM

Compare results to `glm` package results:

```
> summary(linmod)$coef        #Coefficients correct, StdErr's wrong
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -2.323790 0.14621445 -15.89302 0.0005416213
bliss$conc   1.161895 0.06347542  18.30464 0.0003557462
> summary(logitmod)$coef
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -2.323790  0.4178878 -5.560798 2.685438e-08
conc         1.161895  0.1814158  6.404598 1.507665e-10
```

Fitting a GLM

Recall that $\hat{Var}(\hat{\beta}) = (X^T W X)^{-1}\hat{\phi}$.

For binomial distribution, $\phi = 1$. We can carry out that calculation:

```
xm<-model.matrix(linmod)
wm<-diag(w)
sqrt(diag(solve(t(xm) %*% wm %*% xm)))
```

Which yields:

```
(Intercept)  bliss$conc
  0.4178878   0.1814158
```

Fitting a GLM

Recall that for the normal model $\phi = \sigma^2$.

So that: $\hat{Var}(\hat{\beta}) = (X^T W X)^{-1} \hat{\sigma}^2$.

Therefore, $\frac{(X^T W X)^{-1} \hat{\sigma}^2}{\hat{\sigma}^2} = (X^T W X)^{-1}$

```
> summary(linmod)$coef[,2]/summary(linmod)$sigma
(Intercept)   bliss$conc
  0.4178878    0.1814158
```

Hypothesis Tests

We've already seen Likelihood Ratio Tests in several contexts.

GLM's link these notions together in a general framework.

Often testing a saturated model (the data predict themselves) against a model with fewer predictors.

Hypothesis Tests

Can write the LRT in this general way: $2\,(log(y, \phi \mid y) - log(\hat{\mu}, \phi \mid y))$

Or, we can rewrite this in exponential family form:

$$\sum_i^n 2w_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i))/\phi$$

This is called the "scaled deviance" since it is divided by the ϕ parameter.

The "deviance" is not divided by the ϕ parameter. Deviance and Scaled Deviance are the same for Poisson and Binomial since $\phi = 1$.

Hypothesis Tests

GLM	Deviance
Normal	$\sum_i^n (y_i - \hat{\mu}_i^2)$
Poisson	$2\sum_i^n [y_i log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)]$
Binomial	$2\sum_i^n \left[y_i log(\frac{y_i}{\hat{\mu}_i}) + (m - y_i) log\left(\frac{(m-y_i)}{(m-\hat{\mu}_i)}\right)\right]$

For the Poisson model, the second term is usually zero if an intercept is included in the model. This should look familiar.

The binomial model is specified here as $y_i \sim Bin(m, p_i)$ and $\mu_i = mp_i$.

Hypothesis Tests

Two kinds of hypothesis tests for which we can use the deviance:

1. Goodness of fit – does the model fit the data?
2. Nested models – does the smaller model fit the data as well?

Hypothesis Tests

Recall a similar statistic, Pearson's X^2:

$$X^2 = \sum_i^n \frac{(y_i - \hat{\mu}_i)^2}{Var(\hat{\mu}_i)}$$

Which we wrote for a two-way table as:

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

Under the assumption that the model is correct, the Deviance and Pearson's X^2 both have χ^2 distributions with degrees of freedom determined by the number of parameters.

Hypothesis Tests

For comparing a larger to smaller model (nested), the difference in the Deviance has an asymptotic χ^2 distribution.

Larger model has p parameters.
Smaller model has q parameters ($p > q$).

With an estimate of ϕ we can compute an F-statistic of the form:

$$\frac{(G_q^2 - G_p^2)/(df_q - df_p)}{\hat{\phi}}$$

where $\hat{\phi} = X^2/(n - p)$.

Hypothesis Tests

Let's look at the example we have been using – concentrations of insecticide and dead bugs.

Goodness of fit test. Compare the model estimates to the observed values.

Output gives us the Deviance of the model and the degrees of freedom:

```
> summary(logitmod)
<...>
    Null deviance: 64.76327  on 4  degrees of freedom
Residual deviance:  0.37875  on 3  degrees of freedom
```

Hypothesis Tests

We can compare this to the χ^2 distribution to get a p-value.

```
> 1-pchisq(deviance(logitmod),df.residual(logitmod))
[1] 0.9445968
```

Hypothesis Tests

We can also test nested models. The null model is nested within the model with `conc` as a predictor:

Here is the R code (we have seen this before):

```
> anova(logitmod,test="Chi")
Analysis of Deviance Table
Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     4     64.763
conc  1   64.385         3      0.379 1.024e-15 ***
```

The model with the predictor fits the data significantly better than the model without.

Hypothesis Tests

This test does generalize to any model nested within another.

For example, we could estimate a model with *conc* and $conc^2$

```
> logitmod2<-glm(cbind(dead,alive)~conc+ I(conc^2),family=binomial,bliss)
> logitmod2

Coefficients:
(Intercept)          conc     I(conc^2)
   -2.49589       1.41018       -0.06117
```

Hypothesis Tests

Then compare the model with just *conc* to *conc* and $conc^2$:

```
> anova(logitmod,logitmod2,test="Chi")
Analysis of Deviance Table

Model 1: cbind(dead, alive) ~ conc
Model 2: cbind(dead, alive) ~ conc + I(conc^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         3    0.37875
2         2    0.19549  1  0.18325   0.6686
```

In this case, adding the predictor does not improve the fit.

Residuals

In linear regression with normal outcome, the residuals ($\hat{\varepsilon} = y - \hat{\mu}$) are very useful for diagnostics.

This is a problem for other GLMs, which may have non-constant variance.

For example, for both Poisson and Binomial, the variance is a function of the mean.

Residuals

Need alternative residuals for GLMs.

One alternative is a Pearson residual:

$$r_P = \frac{y - \hat{\mu}}{\sqrt{Var(\hat{\mu})}}$$

The square of this term, when summed over all cases, adds up to the

Pearson X^2. That is, $\sum r_P^2 = X^2$.

This is comparable to the standardized residual in linear regression.

Residuals

We can examine these for the insecticide model.

```
> residuals(logitmod,"pearson")
            1             2             3             4             5
-4.325234e-01  3.643729e-01 -3.648565e-15  6.414687e-02 -2.081068e-01

> sum((residuals(logitmod,"pearson"))^2)
[1] 0.3672674
```

Residuals

Another form of the residual is the deviance residual.

$$r_D = sign(y - \hat{\mu})\sqrt{d_i}$$

In the binomial context, this is:

$$d_i = 2y \log\left(\frac{y}{\hat{\mu}}\right) + 2(n-y)\log\left(\frac{n-y}{n-\hat{\mu}}\right)$$

As with the Pearson residuals, $\sum r_D^2 = Deviance = \sum d_i$.

Residuals

Let's look at these for the insecticide problem.

```
> residuals(logitmod) #these are the default residual
          1           2           3           4           5
-0.45101510  0.35969607  0.00000000  0.06430235 -0.20449347

> sum((residuals(logitmod))^2)
[1] 0.3787483
```

Residuals

We can also look at the residuals on the logit scale.

These are the “working residuals.”

Since we programmed an iterative routine, we can extract these from those results:

```
> residuals(linmod)
             1              2              3              4              5
-2.770876e-01  1.561410e-01 -1.520235e-16  2.748820e-02 -1.333195e-01
```

Also possible to extract them from the `glm` output:

```
> residuals(logitmod,"working")
             1              2              3              4              5
-2.770876e-01  1.561410e-01 -1.332268e-15  2.748820e-02 -1.333195e-01
```

Leverage and Influence

Recall from the linear model the "hat" matrix H, where $\hat{y} = Hy$.

The diagonals of this matrix, h_i, are known as the leverages.

Large leverages indicate potentially influential points.

Leverage and Influence

For GLMs, the matrix is a function of the predictors (**X**) and the weights developed through the iterative fitting.

$$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$$

Leverage and Influence

Another measure of influence is the difference in the coefficients when each case is dropped.

Can extract them from `glm` output.

`influence(logitmod)` will extract the leverages, the changes in the coefficients from deleting a case, and changes in the Pearson X^2 and Deviance from deleting a case.

Leverage and Influence

Here are the changes in coefficients:

```
> influence(logitmod)$coef
   (Intercept)          conc
1 -0.214001479  0.080663550
2  0.155671882 -0.047087302
3  0.000000000  0.000000000
4 -0.005841678  0.008417729
5  0.049263918 -0.036573429
```

Overdispersion

For some GLM's, the variance is a separate parameter.

For linear regression with normal outcome, mean and variance are estimated separately.

For some exponential family distributions, mean and variance are linked.

- Binomial variance: $Y_i \sim bin(n_i, p_i), Var(Y_i) = n_i p_i (1 - p_i)$
- Poisson variance: $Y_i \sim Poi(\lambda)$, $Var(Y_i) = \lambda$

Overdispersion

If the observed variance is larger than that implied by the model, this is described as **overdispersion**.

Given that the deviance has an asymptotic distribution χ^2_{n-p} then expected value of the deviance is $n - p$.

If the Deviance$>$ $n - p$, this *may* indicate overdispersion.

Overdispersion

This can be the result of other factors:

- Missing covariates
- Nonlinear functions of covariates
- Large outliers
- Small subgroups (a function of the model)

Overdispersion

Why does this occur?

- Variation among probabilities
- Correlations between binary responses

Two sides of the same coin.

Overdispersion

Recall that the estimated variance for GLMs is $\hat{Var}(\hat{\beta}) = (X^T W X)^{-1}\hat{\phi}$

For the normal model, the scale parameter is $\phi = \sigma^2$.

But for the Poisson and Binomial, $\phi = 1$.

Overdispersion is the situtation where $\phi > 1$.

Overdispersion

The recommended method is to estimate ϕ as $\hat{\phi} = \frac{(X^2)}{n-p}$.

This scale parameter is then used to modify the variance estimates (it's a multiplier).

Doesn't affect estimates of the coefficients.

Overdispersion

In R, use the results from GLM to calculate $\frac{X^2}{n-p}$:

```
d<-sum(logitmod$weights * logitmod$residuals^2)/logitmod$df.residual
```

Then, use summary on the model output – rescales the variances automatically.

```
summary(logitmod,dispersion=d,correlation=TRUE,symbolic.cor=TRUE)
```

Summary

- Exponential family distributions can be united under GLMs
- Need a link function: links response to the linear predictor
- Need special methods for estimation (iterative)
- Deviance is a generalization of variance
- Various hypothesis testing approaches and diagnostics generalized for GLMs
- Overdispersion is a problem unique to some distributional assumptions