Model choice

To test significance single variable could use

\[
\frac{\hat{\beta}_i}{\sqrt{(X^TWX)_{ii}^{-1}}} \xrightarrow{D} N(0, 1)
\]

(cf F v. t-test) The following Likelihood ratio test treats groups of variables (greater power, by NP, so preferable even for single variable).

The scaled residual deviance \( D(y) \)

\[
D(y) = 2(\ell(\hat{\theta}^{(s)}; y) - \ell(\hat{\beta}; y)).
\]

\( \hat{\theta}_i^{(s)} = \arg \max_\theta f(y_i|\theta_i) \) deviance relative to saturated model 1 parameter per datum.

Nested models \( Q \supset P \) dimensions \( q > p \), LRT statistic is \( \Lambda = D^{(P)}(y) - D^{(Q)}(y), \)

\[
D^{(P)}(y) - D^{(Q)}(y) \sim \chi^2(q - p)
\]

approximately, at large \( n \).
Example: budworm moths, test for difference male/female. Allow different intercept and slope in linear predictor.

\[
\begin{align*}
\text{numdead} & \sim 1 + \text{ldose} + \text{sex} + \text{ldose:sex} \\
\log \left( \frac{\mu_i}{m - \mu_i} \right) & = \beta_1 + g_i \beta_2 + x_i \beta_3 + x_i g_i \beta_4 \\
\text{numdead} & \sim 1 + \text{ldose} \\
\log \left( \frac{\mu_i}{m - \mu_i} \right) & = \beta_1 + x_i \beta_3
\end{align*}
\]
> bw.glm1<-glm(cbind(numdead,numalive)~ldose*sex,
            family=binomial,data=bw)
> summary(bw.glm1)
...
   Null deviance: 124.8756 on 11 d.o.f.
   Residual deviance: 4.9937 on 8 d.o.f.
...
> bw.glm0<-glm(cbind(numdead,numalive)~ldose,
            family=binomial,data=bw)
> bw.glm0$deviance
[1] 16.98403

16.984 - 4.9937 = 11.9903 with \( n = 12, q = 4, p = 2 \) and since \( D' - D \sim \chi^2(q - p) \) p-value

> 1-pchisq(11.99,2)
[1] 0.002491177

strong evidence against the null, in favor of gender being explanatory.
Test for no relation between mean response $\mu_i$ and linear predictor $\eta_i = \beta_1 + x_2\beta_2 + ... + x_p\beta_p$ uses

$$D^{(0)}(y) - D(y) \sim \chi^2(p - 1)$$

where $D(y)$ is (scaled residual) deviance for $p$ variable model and $D^{(0)}(y)$ is the deviance for the model with $\beta_2 = ... = \beta_p = 0$ (the null deviance).

\[
\text{numdead} \sim 1 + \text{ldose} + \text{sex} + \text{ldose:sex} \\
\text{numdead} \sim 1
\]

Example: test for no linear relation to mean, using the null deviance, $D^{(0)} = 124$ so $124 - 5 = 119$ with $n = 12, q = 4, p = 1$ and since $D' - D \sim \chi^2(q - p)$ p-value $1 - \text{pchisq}(119, 3)$ will be tiny so reject hypothesis of no linear relation.
Model choice... Digging a bit deeper: the interactions change the odds for mortality.

```r
> bw.glm1<-glm(cbind(numdead,numalive)~ldose*sex,
  data=bw,family=binomial)
> summary(bw.glm1)
...
Coefficients:
  Estimate  Std. Error   z value  Pr(>|z|)
(Intercept)  -2.9935     0.5527    -5.416    6.09e-08
 ldose        0.9060     0.1671     5.422     5.89e-08
 sexM         0.1750     0.7783     0.225     0.822
 ldose:sexM   0.3529     0.2700     1.307     0.191
...
Residual deviance: 4.9937 on 8 D.O.F.
```

\[
O(x) = \frac{\pi(x)}{1 - \pi(x)} = \exp(\beta_1 + \beta_2 g + \beta_3 x)
\]
\[
= \begin{cases} 
\exp(-3 + 0.17 + 0.9x + 0.35x) & \text{Male}, \\
\exp(-3 + 0.9x) & \text{Female}.
\end{cases}
\]

Odds ratio for male against female mortality is now \( \exp(0.17 + 0.35x) \) a function of \( x = \text{ldose} \).
sex and sex:ldose were not separately significant. This is correlation (here is \( \text{cov}(\widehat{\beta}) = I^{-1} \))

```r
> cs<-summary(bw1.glm)$cov.scaled; v<-diag(cs)
> round(correlation.matrix<-cs/sqrt(v%*%t(v)),2)

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>ldose</th>
<th>sexM</th>
<th>ldose:sexM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.00</td>
<td>-0.91</td>
<td>-0.71</td>
<td>0.56</td>
</tr>
<tr>
<td>ldose</td>
<td>-0.91</td>
<td>1.00</td>
<td>0.65</td>
<td>-0.62</td>
</tr>
<tr>
<td>sexM</td>
<td>-0.71</td>
<td>0.65</td>
<td>1.00</td>
<td>-0.89</td>
</tr>
<tr>
<td>ldose:sexM</td>
<td>0.56</td>
<td>-0.62</td>
<td>-0.89</td>
<td>1.00</td>
</tr>
</tbody>
</table>
```

Make ldose \( \perp \) with intercept and sexM \( \perp \) ldose:sexM.

```r
> glm(cbind(numdead,numalive)~sex*I(ldose-3),
   data=bw,family=binomial)

... Coefficients:

|                  | Estimate | Std. Err | z val | Pr(>|z|) |
|------------------|----------|----------|-------|----------|
| (Intercept)      | -0.2754  | 0.2305   | -1.195| 0.23215  |
| sexM             | 1.2337   | 0.3770   | 3.273 | 0.00107  |
| I(ldose-3)       | 0.9060   | 0.1671   | 5.422 | 5.89e-08 |
| sexM:I(ldose-3)  | 0.3529   | 0.2700   | 1.307 | 0.19117  |

Null deviance: 124.8756 on 11 DOF
Residual deviance: 4.9937 on 8 DOF
...
A little more model selection to further demonstrate deviance based tests. Test to drop the interaction.
H0: numdead~sex+ldose, H1: numdead~sex*ldose

```r
> bw.glm<-glm(cbind(numdead,numalive)~sex+ldose, 
  data=bw,family=binomial)
> summary(bw.glm)
...
Coefficients:

  Estimate  Std. Error   z value  Pr(>|z|)
(Intercept)  -3.4732     0.4685    -7.413   1.23e-13
  sexM        1.1007     0.3558     3.093   0.00198
  ldose       1.0642     0.1311     8.119   4.70e-16
...
Residual deviance:  6.757  on  9  DOF
...
> 1-pchisq(6.757-4.9937,1)
[1] 0.1842134
```

Accept numdead~sex+ldose. We rejected numdead~ldose and numdead~1 earlier. This model gives a better fit to the data than the alternatives we have allowed. But... is it any good?
Goodness of fit for a GLM (a first look)

The saturated model has no constraints on the GLM parameters $\theta$

$$\ell(\theta^{(s)}; y) = \sum_i y_i \theta_i^{(s)} - \kappa(\theta^{(s)}) + c(y_i; \phi)$$

The test $H_0 : \theta = \theta(\beta)$ against $H_1 : \theta = \theta^{(s)}$ is a test for goodness of fit, i.e part of the fit diagnostics.

Recall scaled residual deviance $D(y)$

$$D(y) = 2(\ell(\hat{\theta}^{(s)}; y) - \ell(\hat{\beta}; y)).$$

$\hat{\theta}_i^{(s)} = \arg \max_{\theta} f(y_i | \theta_i)$ deviance relative to saturated model 1 parameter per datum.

Since $D(y)$ is the LRT statistic comparing the saturated model ($n$ parameters) with the model of interest ($p$ parameters) $D(Y) \sim \chi^2(n-p)$ approximately, as $\hat{\theta}^{(s)} \rightarrow \theta^{(s)}$ and $\hat{\beta} \rightarrow \beta$. 

Example: The deviance for `numdead~ldose+sex` is $D(y) = 6.757$, with $n = 12$ and $p = 3$ so $n - p = 9$ DOF. Testing with this as the null and the saturated as alternative gives a $p$-value

```r
> 1-pchisq(6.757,9)
[1] 0.6624024
```

so no evidence for misfit here.

As an example with mild misfit `numdead~ldose` had residual deviance 16.984 on 10 DOF for $p = 0.07$ and mild evidence for misfit. We already know the cause is the need for the explanatory variable gender.
This goodness of fit test needs $\tilde{\theta}^{(s)} \rightarrow \theta^{(s)}$.
This is not large $n!$ - adds components to $\theta^{(s)}$

Example: Binomial model $Y_i \sim \text{Binomial}(m, \pi_i)$, $\mu_i = m\pi_i$;

$$\ell(\theta^{(s)}; y) = \sum_i y_i \theta_i^{(s)} - m \log(1 + \exp(\theta_i^{(s)}))$$

MLE for $\theta_i^{(s)}$ is $\hat{\theta}(s) = \log(y_i/(m - y_i))$ so

$$\ell(\hat{\theta}(s); y) = \sum_i y_i \log \left( \frac{y_i}{m - y_i} \right) - m \log \left( \frac{m_i}{m - y_i} \right)$$

If $x_i \hat{\beta}_i = \log(\hat{\mu}_i/(m - \hat{\mu}_i))$ then

$$\ell(\hat{\beta}; y) = \sum_i y_i \log \left( \frac{\hat{\mu}_i}{m - \hat{\mu}_i} \right) - m \log \left( \frac{m_i}{m - \hat{\mu}_i} \right)$$

so

$$D(y) = 2 \sum_i y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) + (m - y_i) \log \left( \frac{m - y_i}{m - \hat{\mu}_i} \right)$$

and $D \sim \chi^2(n - p)$ as $m \rightarrow \infty$ at fixed $n$. 
GLM Diagnostics (residuals)

Compare response $Y_i$ with fitted values $\hat{\mu}_i$. Residuals $Y_i - \hat{\mu}_i$.

```r
> bw$numdead-20*fitted(bw.glm)
     1     2     3     4
-0.7060152 -0.2557709 0.2139042 -0.8857031 ...
... 12
... -1.2776412
```

`fitted(bw.glm)` is `predict(bw.glm, type='response')` This is the same kind of visualisation we made before.
Several natural definitions for GLM residuals: The working residuals are $g'(\mu)(y - \mu)$. Misfit in the space of the linear predictor is

$$g(y_i) - x_i\hat{\beta} \simeq (g(\mu_i) - x_i\beta) + g'(\mu_i)(y_i - \mu_i) = g'(\mu_i)(y_i - \mu_i)$$

The IRLS iteration was

$$z_i = x_i\beta + g'(\mu_i)(y_i - \mu_i)$$
$$\beta' = (X^TWX)^{-1}X^TWz$$

these are $e_i = z_i - x_i\beta$, residuals of IRLS algorithm, with $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$.

The Cook’s distance is defined in the space of the linear predictor:

$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{-i})^T(X^TWX)(\hat{\beta} - \hat{\beta}_{-i})}{p\phi}$$

Compare CD $C_i$ in our NLM at known $\phi = \sigma^2$

$$\frac{(\hat{y} - \hat{y}_{-i})^T(\hat{y} - \hat{y}_{-i})}{p\sigma^2} = \frac{(\hat{\beta} - \hat{\beta}_{-i})^T(X^TX)(\hat{\beta} - \hat{\beta}_{-i})}{p\sigma^2}$$
We are doing NLM style diagnostics on the last (weighted) regression in IRLS algorithm.

```r
> plot(fitted.values(bw.glm), bw.glm$residuals)
> plot(cooks.distance(bw.glm))
```
Need residuals on a fixed scale (identify misfit)

Bad points add to deviance:

\[ \ell_i(\theta; y_i) = \frac{y_i \theta - \kappa(\theta_i)}{\phi} + c(y_i; \phi), \]

and

\[ d_i = -2 \ell_i(\hat{\beta}; y_i) + 2 \ell_i(\hat{\theta}_i^{(s)}; y_i), \]

\[ D(y) = \sum_{i=1}^{n} d_i. \]

Deviance residuals

\[ r_i = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i}. \]

sign: too large or too small compared to the fitted mean?

[compare NLM with \( \sigma^2 = 1 \) known, then \( \ell = -\text{RSS}/2\sigma^2 \) and \( D = \text{RSS} \), so \( r_i \) is equivalent to \( e_i \) and \( D = \sum_i r_i^2 \), also \( D \) and \( \text{RSS}/\sigma^2 \) are approximately \( \sim \chi^2(n - p) \)]
> summary(bw.glm)
...
Deviance Residuals:
    Min 1Q Median 3Q Max
-1.10540 -0.65343 -0.02225 0.48471 1.42944

Coefficients:
...

> summary(bw.glm)$deviance.resid
       1         2         3         4         5         6         7         8         9        10        11        12
-0.6087798 -0.1407910 -0.8068011 -0.7816145 -0.6401939 -0.7492806 -0.6087798 -0.1407910 ... -0.7873883

Standardised deviance residuals $h_{ii} = H_{ii}$, $H$ from IRLS,

$$r'_i = \text{sign}(y_i - \hat{\mu}_i) \frac{\sqrt{d_i}}{\sqrt{\phi} \sqrt{1 - h_{ii}}}.$$ 

If $D(y) \sim \chi^2(n - p)$ is good then $r' \sim N(0, I_n)$ approximately.

For the budworm moth analysis, numdead~sex+ldose standardised deviance residuals
> #bw.glm is the glm output for numdead~sex+ldose
> plot(fitted.values(bw.glm),rstandard(bw.glm))

No evidence for misfit, no large standardised deviance residuals. There is no obvious sign of a trend in the residuals. This is often a sign of a problem with the link function.
To see what happens in a poorer model fit

```r
> bw.glm <- glm(cbind(numdead, numalive) ~ ldose, 
    data = bw, family = binomial)
> plot(fitted.values(bw.glm), rstandard(bw.glm))
```

Small data set but still some large misfit.
Contingency Tables

<table>
<thead>
<tr>
<th></th>
<th>recovered</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>treated</td>
<td>$r_t$</td>
<td>$s_t$</td>
</tr>
<tr>
<td>untreated</td>
<td>$r_u$</td>
<td>$s_u$</td>
</tr>
<tr>
<td></td>
<td>$r$</td>
<td>$s$</td>
</tr>
</tbody>
</table>

What difference does treatment make?
How was the data gathered?

1. **a** (prospective) Treat $n_t$ individuals and leave $n_u$ and count recoveries. $r_t$ is Binomial with $n_t$ trials (and $r_u$ with $n_u$).

   b (retrospective) Find $r$ recovered and $s$ remained sick and count numbers treated in each class. $r_t$ is Binomial with $r$ trials (and $s_t$ with $s$).

2. Take $n$ individuals and count numbers treated and recovered. Totals are multinomial.

3. Monitor patients at a hospital for a fixed time. Totals $r_t$, $d_t$, $r_u$, $d_u$ may be Poisson.
Example (from Dr Lunn’s 07 lectures)

<table>
<thead>
<tr>
<th>Sex</th>
<th>Smoking</th>
<th></th>
<th>Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>2059</td>
<td>857</td>
<td>2916</td>
</tr>
<tr>
<td>F</td>
<td>1130</td>
<td>1373</td>
<td>2503</td>
</tr>
<tr>
<td>Tot</td>
<td>3189</td>
<td>2230</td>
<td>5419</td>
</tr>
</tbody>
</table>

\[ Y_i \sim \text{Binomial}(n_i, \pi_i). \]

\[ x_1 = (1, 1), x_2 = (1, 0) \text{ (indicator=1 Male)}. \]

\[
\begin{align*}
\pi_1 & = \frac{\exp(\beta_1 + \beta_2)}{1 + \exp(\beta_1 + \beta_2)} \\
\pi_2 & = \frac{\exp(\beta_1)}{1 + \exp(\beta_1)}. \\
\end{align*}
\]

Test for effect due to 'Sex' is test for \( \beta_2 \neq 0 \) Does 'Sex' predict 'Smoker' (is odds ratio for Male/Female smoking significantly different from one)?
> (smk<-data.frame(S=c(2059,1130),NS=c(857,1373),
>       Sex=c('M','F')))
>   S  NS Sex
> 1 2059 857 M
> 2 1130 1373 F
> smk.glm<-glm(cbind(smk$S,smk$NS)~Sex,
>        family=binomial,data=smk)
> summary(smk.glm)
> ...
> Coefficients:
>   Estimate Std. Error z value Pr(>|z|)
> (Intercept)  -0.19478  0.04017  -4.849 1.24e-06
> SexM         1.07132  0.05715   18.747 < 2e-16
> ...
> Null deviance: 363.57 on 1 degrees of freedom
> Residual deviance: 0.00 on 0 dof
>
> Model is saturated. 'Sex' is explanatory. Odds ratio for smoking in males and females
> exp(1.07132) ≃ 2.92.
Note $\beta_2 = \log(\pi_1(1 - \pi_2)/\pi_2(1 - \pi_2))$

$\hat{\pi}_i = y_i/n_i$

$\exp(\hat{\beta}_2) = y_1(n_2 - y_2)/y_2(n_1 - y_1) = AD/BC$. 

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>F</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>Sex</td>
<td>Smoking</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>M</td>
<td>2059</td>
<td>857</td>
</tr>
<tr>
<td>F</td>
<td>1130</td>
<td>1373</td>
</tr>
<tr>
<td>Tot</td>
<td>m1=3189</td>
<td>m2=2230</td>
</tr>
</tbody>
</table>

\[
Y_i' \sim \text{Binomial}(m_i, \pi'_i).
\]

\[
x_1 = (1, 1), x_2 = (1, 0) \text{ (indicator=1 Smoker)}.
\]

\[
\pi'_1 = \frac{\exp(\beta'_1 + \beta'_2)}{1 + \exp(\beta'_1 + \beta'_2)}
\]

\[
\pi'_2 = \frac{\exp(\beta'_1)}{1 + \exp(\beta'_1)}.
\]

Does 'Smoker' predict gender?
> (smk <- data.frame(M = c(2059, 857), F = c(1130, 1373),
                  Smoke = c(T, F)))

<table>
<thead>
<tr>
<th>M</th>
<th>F</th>
<th>Smoke</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2059</td>
<td>1130</td>
</tr>
<tr>
<td>2</td>
<td>857</td>
<td>1373</td>
</tr>
</tbody>
</table>

> smk.glm <- glm(cbind(smk$M, smk$F) ~ Smoke,
                 family = binomial, data = smk)

> summary(smk.glm)

...  

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | -0.47132   | 0.04353 | -10.83   | <2e-16    |
| SmokeTRUE  | 1.07132    | 0.05715 | 18.75    | <2e-16    |

Notice $\hat{\beta}'_2 = \hat{\beta}_2$.

$$\exp(\hat{\beta}'_2) = y'_1(m_2 - y'_2)/y'_2(m_1 - y'_1) = AD/BC$$

The same property holds for larger (and $k$-way) tables of categorical variables and is one of the attractions of working with log-odds.