Generalized Linear Mixed-Effects Models

Reconsideration of the Plant Fungus Example

Consider again the experiment designed to evaluate the effectiveness of an anti-fungal chemical on plants.

A total of 60 plant leaves were randomly assigned to treatment with 0, 5, 10, 15, 20, or 25 units of the anti-fungal chemical, with 10 plant leaves for each amount of anti-fungal chemical.

All leaves were infected with a fungus.

Following a two-week period, the leaves were studied under a microscope, and the number of infected cells was counted and recorded for each leaf.

Our Previous Generalized Linear Model

We initially considered a Poisson Generalized LM with a log link function and a linear predictor ($\boldsymbol{x}_i'\boldsymbol{\beta}$) that included an intercept and a slope coefficient on the amount of anti-fungal chemical applied to a leaf:

$y_i \sim \text{Poisson}(\lambda_i)$,
$\log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$,
$\boldsymbol{x}_i' = [1, x_i]$, $\boldsymbol{\beta} = [\beta_0, \beta_1]'$,
$y_1, \ldots, y_n$ independent.

Overdispersion

We found evidence of overdispersion indicated by greater variation among counts than would be expected based on the estimated mean count.

We considered a quasi-likelihood approach to account for the overdispersion, but another strategy would be to specify a model for the data that allows for variation in excess of the mean.

A New Model for the Infection Counts

Let $\ell_i \sim N(0, \sigma^2_\ell)$ denote a random effect for the ith leaf.

Suppose $\log(\lambda_i) = \beta_0 + \beta_1 x_i + \ell_i$ and $y_i|\lambda_i \sim \text{Poisson}(\lambda_i)$.

Finally, suppose $\ell_1, \ldots, \ell_n$ are independent and that $y_1, \ldots, y_n$ are conditionally independent given $\lambda_1, \ldots, \lambda_n$.

The Lognormal Distribution

If $\log(v) \sim N(\mu, \sigma^2)$, then v is said to have a *lognormal distribution*.

The mean and variance of a lognormal distribution are

$$E(v) = \exp(\mu + \sigma^2/2)$$

and

$$\text{Var}(v) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2).$$

Conditional Expectation and Variance

For random variables u and v,

$$E(u) = E(E(u|v))$$

and

$$\text{Var}(u) = E(\text{Var}(u|v)) + \text{Var}(E(u|v)).$$

A Lognormal Mixture of Poisson Distributions

Suppose $\log(v) \sim N(\mu, \sigma^2)$ and $u|v \sim \text{Poisson}(v)$.

Then the unconditional distribution of u is a lognormal mixture of Poisson distributions,

$$E(u) = E(E(u|v)) = E(v) = \exp(\mu + \sigma^2/2), \text{ and}$$

$$\begin{aligned}
\text{Var}(u) &= E(\text{Var}(u|v)) + \text{Var}(E(u|v)) = E(v) + \text{Var}(v) \\
&= \exp(\mu + \sigma^2/2) + \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2) \\
&= \exp(\mu + \sigma^2/2) + (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2) \\
&= E(u) + (\exp(\sigma^2) - 1)[E(u)]^2.
\end{aligned}$$

$E(y_i)$ and $\text{Var}(y_i)$ in Our New Model for Infection Counts

$$E(y_i) = E(E(y_i|\lambda_i)) = E(\lambda_i) = \exp(\beta_0 + \beta_1 x_i + \sigma_\ell^2/2).$$

$$\begin{aligned}
\text{Var}(y_i) &= E(\text{Var}(y_i|\lambda_i)) + \text{Var}(E(y_i|\lambda_i)) \\
&= E(\lambda_i) + \text{Var}(\lambda_i) \\
&= E(y_i) + (\exp(\sigma_\ell^2) - 1)[E(y_i)]^2.
\end{aligned}$$

Thus, when $\sigma_\ell^2 > 0$, $\text{Var}(y_i) > E(y_i)$.

The Probability Mass Function of y_i

For $y \in \{0, 1, 2, \ldots\}$,

$$\begin{aligned} f_i(y) &= Pr(y_i = y) = \int_0^\infty Pr(y_i = y|\lambda_i = \lambda) h(\lambda; \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma_\ell^2) d\lambda \\ &= \int_0^\infty \frac{\lambda^y \exp(-\lambda)}{y!} h(\lambda; \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma_\ell^2) d\lambda \\ &= \int_0^\infty \frac{\lambda^y \exp(-\lambda)}{y!} \frac{1}{\lambda\sqrt{2\pi\sigma_\ell^2}} \exp\left\{\frac{-(\log(\lambda) - x_i'\boldsymbol{\beta})^2}{2\sigma_\ell^2}\right\} d\lambda, \end{aligned}$$

where $h(\lambda; \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma_\ell^2)$ is the lognormal density of λ_i.

The Probability Mass Function of y_i (continued)

There is no closed-form expression for $f_i(y)$, the probability mass function of y_i.

The integral in $f_i(y)$ must be approximated using numerical methods.

The `R` function `glmer` in the package `lme4` uses the Laplace approximation (by default) to approximate the integral, but `glmer` also permits the use of the more general integral approximation method known as adaptive Gauss-Hermite quadrature.

The Log Likelihood

The log likelihood is $\ell(\boldsymbol{\beta}, \sigma^2_\ell \mid \boldsymbol{y}) = \sum_{i=1}^n \log\{f_i(y_i)\}$.

Let $\tilde{f}_i(\cdot)$ denote the approximation of $f_i(\cdot)$. Then the approximate log likelihood is

$$\tilde{\ell}(\boldsymbol{\beta}, \sigma^2_\ell \mid \boldsymbol{y}) = \sum_{i=1}^n \log\{\tilde{f}_i(y_i)\},$$

which can be maximized over $\boldsymbol{\beta}$ and σ^2_ℓ using numerical methods to obtain MLEs of $\boldsymbol{\beta}$ and σ^2_ℓ and an estimated inverse Fisher information matrix.

Let $\underset{\sim}{\theta} = \begin{bmatrix} \underset{\sim}{\beta} \\ \sigma^2_\ell \end{bmatrix}$

```
> library(lme4)
>
> leaf=factor(1:60)
>
> o=glmer(y~x+(1|leaf),family=poisson(link = "log"))
>
> summary(o)

Generalized linear mixed model fit by maximum likelihood
 Family: poisson ( log )
Formula: y ~ x + (1 | leaf)

     AIC      BIC    logLik  deviance
 410.7283 417.0113 -202.3642 404.7283
```

$-2\ell(\hat{\underset{\sim}{\theta}}) + 2(3)$

$-2\ell(\hat{\underset{\sim}{\theta}}) + 3\log(60)$

$\ell(\hat{\underset{\sim}{\theta}})$

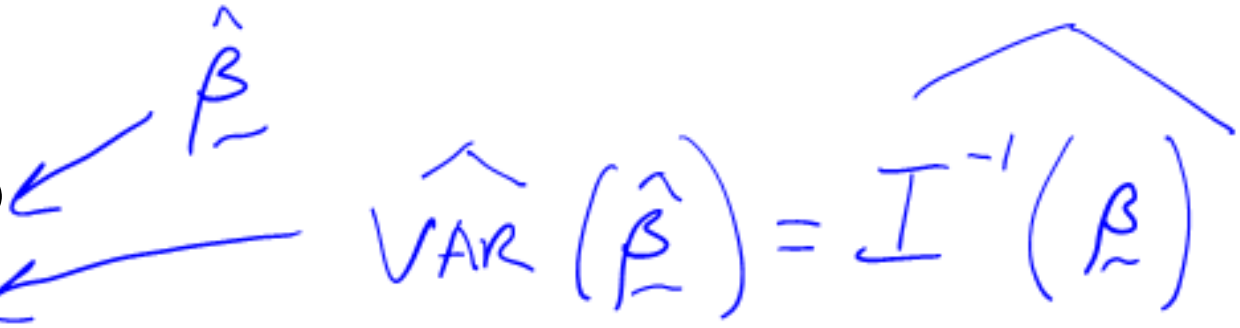
$-2\ell(\hat{\underset{\sim}{\theta}})$

```
Random effects:
 Groups Name        Variance Std.Dev.
 leaf   (Intercept) 0.4191   0.6474
Number of obs: 60, groups: leaf, 60

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.14916    0.15964   25.99   <2e-16 ***
x           -0.15704    0.01252  -12.54   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
  (Intr)
x -0.788
```

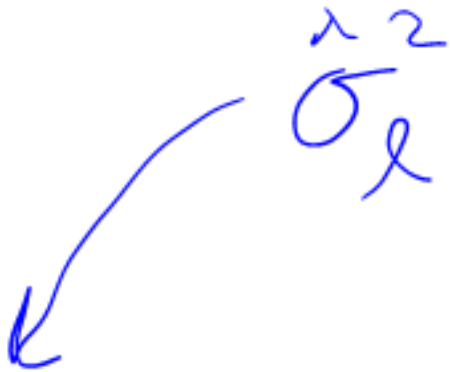
$\hat{\sigma}^2_\ell$ $\hat{\sigma}_\ell$

$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$

$$\frac{\widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{\widehat{Var}(\hat{\beta}_0)\,\widehat{Var}(\hat{\beta}_1)}}$$



```
> b=fixef(o)
> v=vcov(o)
> b
(Intercept)           x
  4.1491634  -0.1570416


> v
2 x 2 Matrix of class "dpoMatrix"
             (Intercept)             x
(Intercept)  0.025484034 -0.0015742367
x           -0.001574237  0.0001567947
```




```
> sigma.sq.leaf=unlist(VarCorr(o))
>
> sigma.sq.leaf
     leaf
0.4191069
```

Conditional vs. Marginal Mean

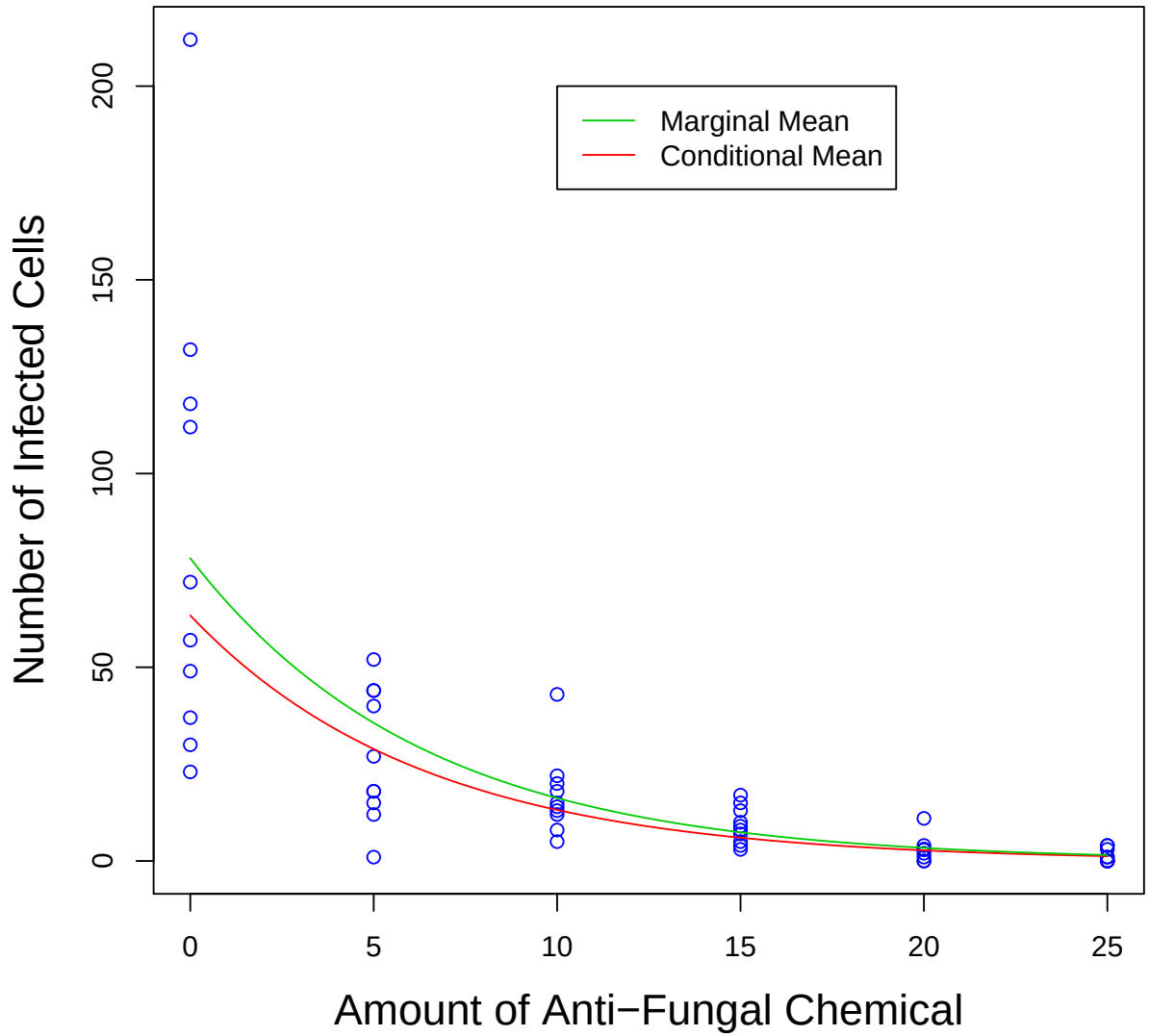
The conditional mean of y_i given $\ell_i = 0$ is

$$E(y_i|\ell_i = 0) = \exp(\beta_0 + \beta_1 x_i).$$

The marginal mean of y_i is

$$E(y_i) = \exp(\beta_0 + \beta_1 x_i + \sigma_\ell^2/2) = \exp(\beta_0 + \beta_1 x_i)\exp(\sigma_\ell^2/2).$$

```
plot(x,y,xlab="Amount of Anti-Fungal Chemical",
     ylab="Number of Infected Cells",col=4,cex.lab=1.5)
lines(xgrid,exp(b[1]+b[2]*xgrid),col=2)
lines(xgrid,exp(b[1]+b[2]*xgrid+sigma.sq.leaf/2),col=3)
legend(10,200,c("Marginal Mean","Conditional Mean"),
       lty=1,col=c(3,2))
```


Marginal Mean
Conditional Mean
Number of Infected Cells
0
50
100
150
200
0
5
10
15
20
25
Amount of Anti−Fungal Chemical

Generalized Linear Mixed-Effects Models

The model for the infection counts is a special case of a Generalized Linear Mixed-Effects Model (GLMM):

For $i = 1, \ldots, n$, $y_i|\mu_i$ has a distribution in the exponential dispersion family with mean μ_i, and $y_1, \ldots, y_n$ are conditionally independent given $\mu_1, \ldots, \mu_n$.

For some link function $g(\cdot)$,

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u}, \text{ where } \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G})$$

and $\boldsymbol{G}$ is a variance matrix of known form that may depend on unknown parameters (e.g., variance components).

In our model for the infection counts, we have . . .

conditional Poisson distributions,

$\mu_i = \lambda_i$,

$g(\cdot) = \log(\cdot)$,

$\boldsymbol{z}_i'$ = the ith row of the $n \times n$ identity matrix,

$\boldsymbol{u} = [\ell_1, \ldots, \ell_n]'$, and

$\boldsymbol{G} = \sigma_\ell^2 \boldsymbol{I}$.

Reconsider again the experiment designed to evaluate the effectiveness of an anti-fungal chemical on plants.

Suppose the 60 plant leaves used in the experiment were obtained by selecting two leaves from each of 30 plants.

Ten leaves obtained from five randomly selected plants were assigned to each treatment (0, 5, 10, 15, 20, or 25 units of the anti-fungal chemical).

Two weeks after fungal infection, the leaves were studied under a microscope, and the number of infected cells was counted and recorded for each leaf.

```
> d=data.frame(plant,leaf,x,y)
> head(d)
  plant leaf x   y
1     1    1 0  30
2     1    2 0  57
3     2    3 0  23
4     2    4 0 118
5     3    5 0 212
6     3    6 0 132
> tail(d)
   plant leaf  x y
55    28   55 25 3
56    28   56 25 1
57    29   57 25 4
58    29   58 25 0
59    30   59 25 4
60    30   60 25 0
```

An Updated Generalized Linear Mixed-Mixed Model

All is as in the previous model on slide 5 except that now we have

$$\log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u}, \text{ where}$$

$\boldsymbol{z}_i'$ is the ith row of $\boldsymbol{Z} = [\boldsymbol{I}_{30\times 30} \otimes \boldsymbol{1}_{2\times 1}, \boldsymbol{I}_{60\times 60}]$ and

$$\boldsymbol{u} = \begin{bmatrix} p_1 \\ \vdots \\ p_{30} \\ \ell_1 \\ \vdots \\ \ell_{60} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_p^2\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_\ell^2\boldsymbol{I} \end{bmatrix} \right).$$

PLANT RANDOM EFFECTS

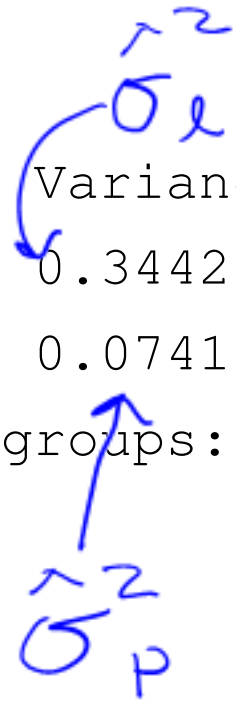
LEAF RANDOM EFFECTS

$$\underset{\sim}{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \sigma^2_p \\ \sigma^2_\ell \end{bmatrix}$$

```
> o=glmer(y~x+(1|plant)+(1|leaf),
          family=poisson(link = "log"))
> summary(o)

Generalized linear mixed model fit by maximum likelihood
 Family: poisson ( log )
Formula: y ~ x + (1 | plant) + (1 | leaf)

     AIC      BIC   logLik  deviance
 412.2036 420.5810 -202.1018 404.2036
```

$$-2\ell(\hat{\underset{\sim}{\theta}}) + 4\log(60)$$

```
Random effects:
 Groups Name        Variance Std.Dev.
 leaf   (Intercept) 0.34426  0.5867
 plant  (Intercept) 0.07414  0.2723
Number of obs: 60, groups: leaf, 60; plant, 30
```



```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.15125    0.17157   24.20   <2e-16 ***
x           -0.15725    0.01324  -11.88   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
  (Intr)
x -0.791
```

```
> b=fixef(o)
> v=vcov(o)
> vc=unlist(VarCorr(o))
> b
(Intercept)           x
  4.1512534  -0.1572525
> v
2 x 2 Matrix of class "dpoMatrix"
             (Intercept)             x
(Intercept)  0.029436558 -0.0017978827
x           -0.001797883  0.0001753472
> vc
     leaf     plant
0.34425551 0.07413645
```

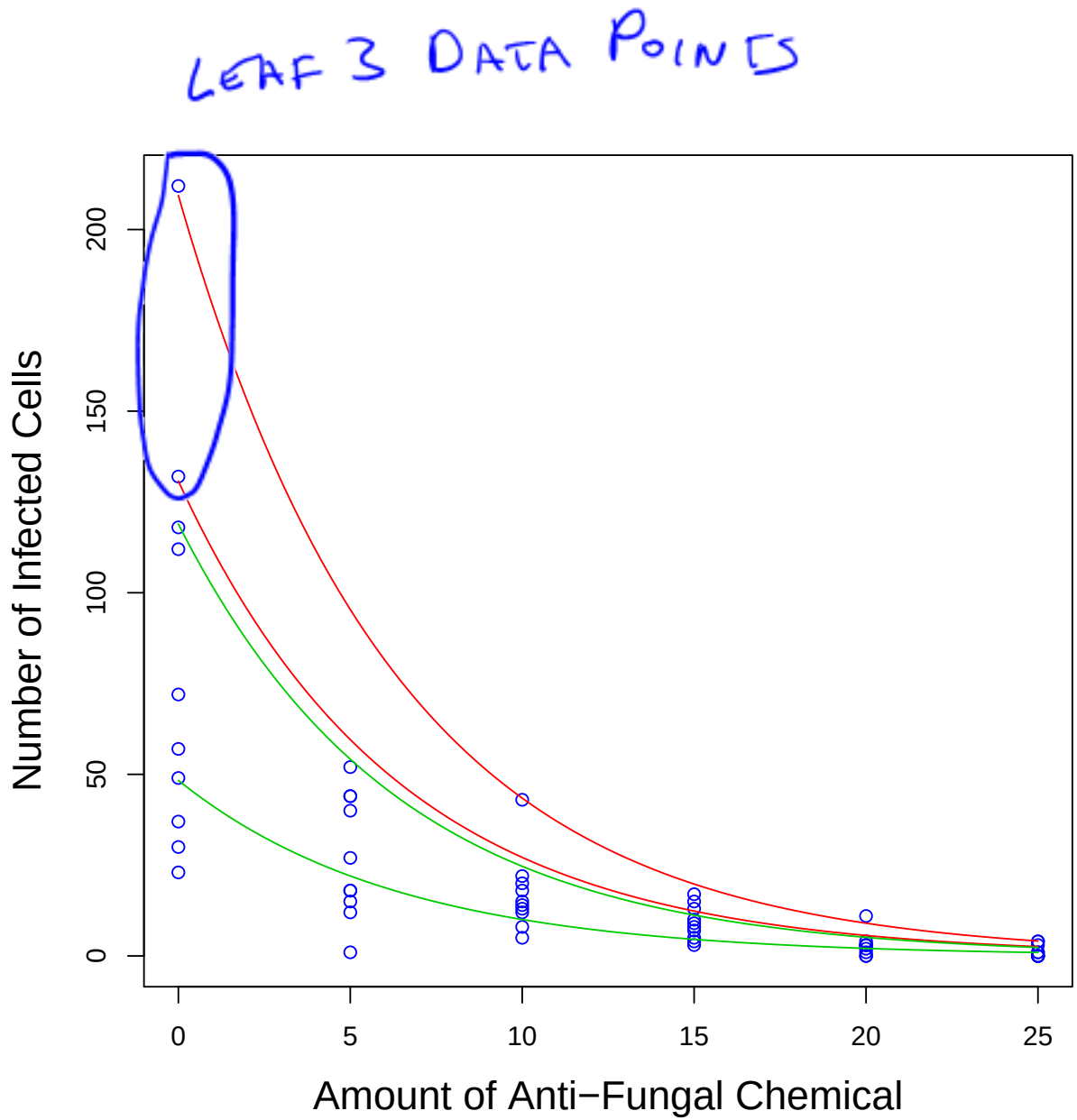
$\hat{\beta}$

$\widehat{VAR}(\hat{\beta})$

$\tilde{\sigma}^2_\ell$ $\hat{\sigma}^2_p$

APPROXIMATE BLUPs OF PLANT AND LEAF RANDOM EFFECTS

```
uhat=ranef(o)
uplant=unlist(uhat$plant)
uleaf=unlist(uhat$leaf)

plot(x,y,xlab="Amount of Anti-Fungal Chemical",
     ylab="Number of Infected Cells",col=4,cex.lab=1.5)

lines(xgrid,exp(b[1]+b[2]*xgrid+uplant[3]+uleaf[5]),col=2)
lines(xgrid,exp(b[1]+b[2]*xgrid+uplant[3]+uleaf[6]),col=2)
lines(xgrid,exp(b[1]+b[2]*xgrid+uplant[30]+uleaf[59]),col=3
lines(xgrid,exp(b[1]+b[2]*xgrid+uplant[30]+uleaf[60]),col=3
```


LEAF 3 DATA POINTS
Number of Infected Cells
0
50
100
150
200
0
5
10
15
20
25
Amount of Anti−Fungal Chemical

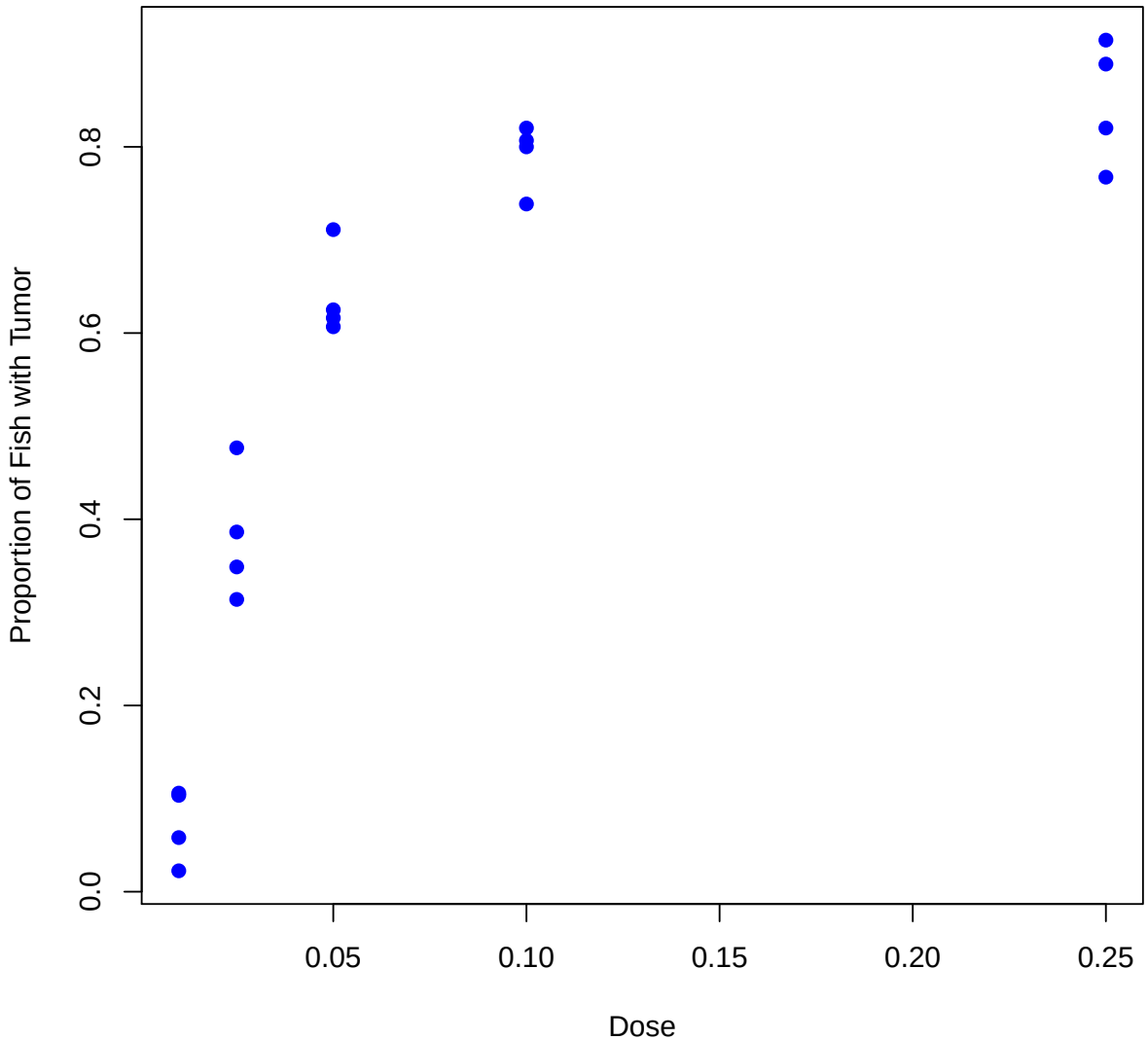
Now suppose that instead of a conditional Poisson response, we have a conditional binomial response for each unit in an experiment or an observational study.

As an example, consider again the trout data set discussed on page 669 of *The Statistical Sleuth*, 3rd edition, by Ramsey and Schafer.

Five doses of toxic substance were assigned to a total of 20 fish tanks using a completely randomized design with four tanks per dose.

For each tank, the total number of fish and the number of fish that developed liver tumors were recorded.

```
d=read.delim("http://dnett.github.io/S510/Trout.txt")
d
    dose tumor total
1  0.010     9    87
2  0.010     5    86
3  0.010     2    89
4  0.010     9    85
5  0.025    30    86
6  0.025    41    86
7  0.025    27    86
8  0.025    34    88
9  0.050    54    89
10 0.050    53    86
11 0.050    64    90
12 0.050    55    88
13 0.100    71    88
14 0.100    73    89
15 0.100    65    88
16 0.100    72    90
17 0.250    66    86
18 0.250    75    82
19 0.250    72    81
20 0.250    73    89
```


Proportion of Fish with Tumor
0.0
0.2
0.4
0.6
0.8
0.05
0.10
0.15
0.20
0.25
Dose

A GLMM for the Tumor Data

Let $m_i =$ the number of fish in tank i.

Let $y_i =$ the *proportion* of fish in tank i with tumors.

Suppose $y_i|\pi_i \overset{ind}{\sim} \text{binomial}(m_i, \pi_i)/m_i$.

Then $E(y_i|\pi_i) = m_i\pi_i/m_i = \pi_i$.

Suppose $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u}$, where ...

A GLMM for the Tumor Data (continued)

$$\boldsymbol{x}_i' = \begin{cases} [1,0,0,0,0] & \text{if dose} = 0.010 \\ [1,1,0,0,0] & \text{if dose} = 0.025 \\ [1,0,1,0,0] & \text{if dose} = 0.050 \\ [1,0,0,1,0] & \text{if dose} = 0.100 \\ [1,0,0,0,1] & \text{if dose} = 0.250 \end{cases},$$

$\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]'$,

$\boldsymbol{z}_i$ is the ith row of $\boldsymbol{I}_{20\times 20}$, and

$\boldsymbol{u} = [u_1, \ldots, u_{20}]' \sim N(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I})$.

A GLMM for the Tumor Data (continued)

Alternatively, we could introduce two subscripts ($i = 1, 2, 3, 4, 5$ for dose and $j = 1, 2, 3, 4$ for tank nested within dose) and rewrite the same model as

$$y_{ij}|\pi_{ij} \overset{ind}{\sim} \text{binomial}(m_{ij}, \pi_{ij})/m_{ij}$$

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \delta_i + u_{ij}$$

$$u_{11}, u_{12}, \ldots, u_{53}, u_{54} \overset{iid}{\sim} N(0, \sigma_u^2)$$

```
> d$dosef=gl(5,4)
> tank=factor(1:20)
> o=glmer(cbind(tumor,total-tumor)~dosef+(1|tank),
+         family=binomial(link="logit"),nAGQ=20,
+         data=d)
```

DEFAULT nAGQ = 1 IS LAPLACE APPROXIMATION.
INTEGER nAGQ > 1 IS NUMBER OF POINTS PER AXIS FOR GAUSS-HERMITE APPROXIMATION TO $\ell(\underset{\sim}{\theta})$.
nAGQ ↑ ⟹ ACCURACY ↑, SPEED ↓

```
> b=fixef(o)
> v=vcov(o)
> vc=unlist(VarCorr(o))

> b
(Intercept)       dosef2       dosef3       dosef4       dosef5
  -2.559544    2.075196    3.136824    3.896338    4.269113

> round(v,3)
5 x 5 Matrix of class "dpoMatrix"
            (Intercept) dosef2 dosef3 dosef4 dosef5
(Intercept)       0.046 -0.046 -0.046 -0.046 -0.046
dosef2           -0.046  0.060  0.046  0.046  0.046
dosef3           -0.046  0.046  0.060  0.046  0.046
dosef4           -0.046  0.046  0.046  0.065  0.046
dosef5           -0.046  0.046  0.046  0.046  0.071

> vc
      tank
0.009590674
```

$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)'$

$= \widehat{VAR}(\hat{\beta})$

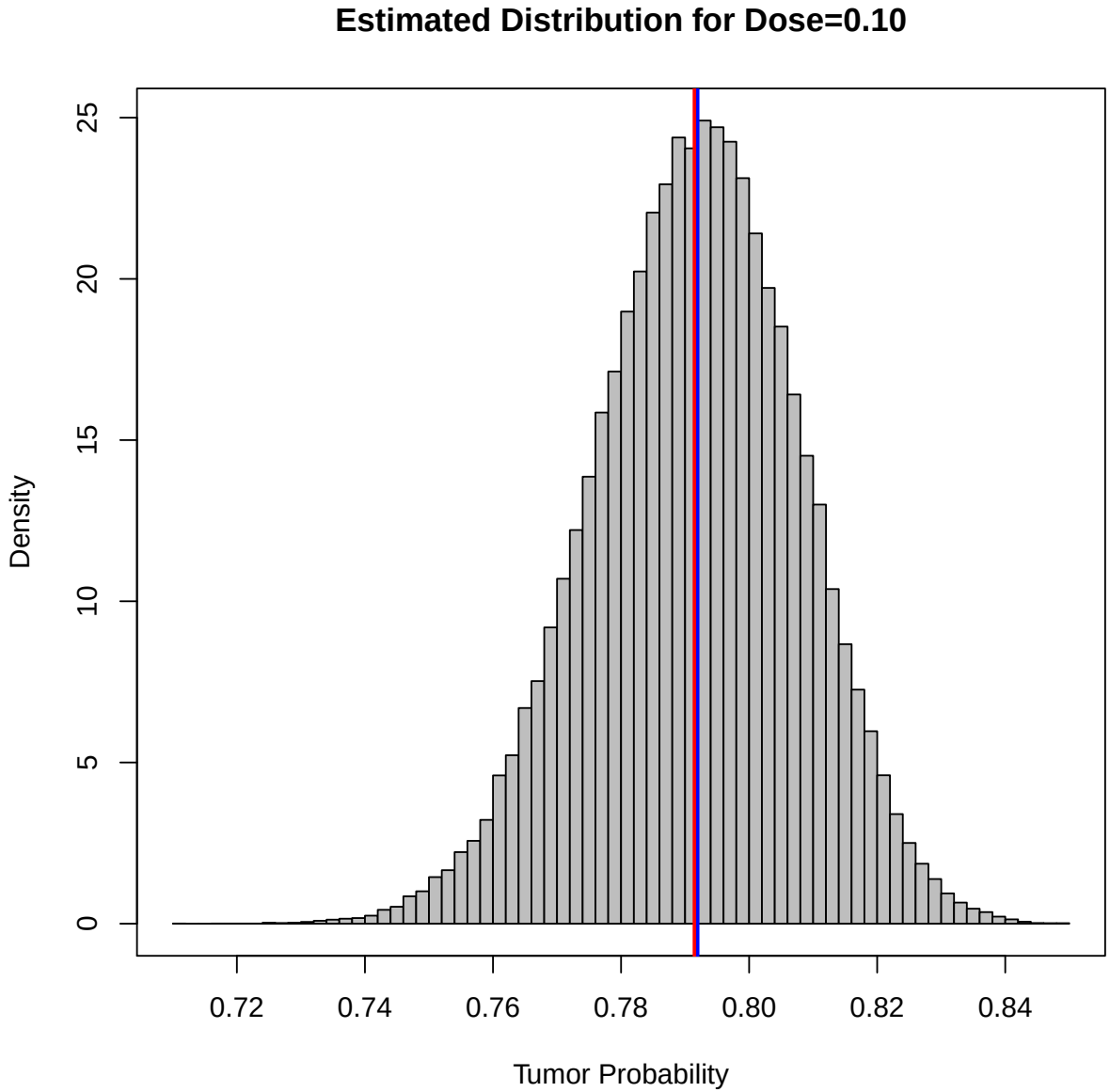
$\hat{\sigma}^2_t$

```
> #Estimated tumor probability for fish in a
> #"typical" tank (tank random effect = 0)
> #treated with 0.10 units (dose 4) is
>
> 1/(1+exp(-(b[1]+b[4])))
(Intercept)
  0.7919621
```

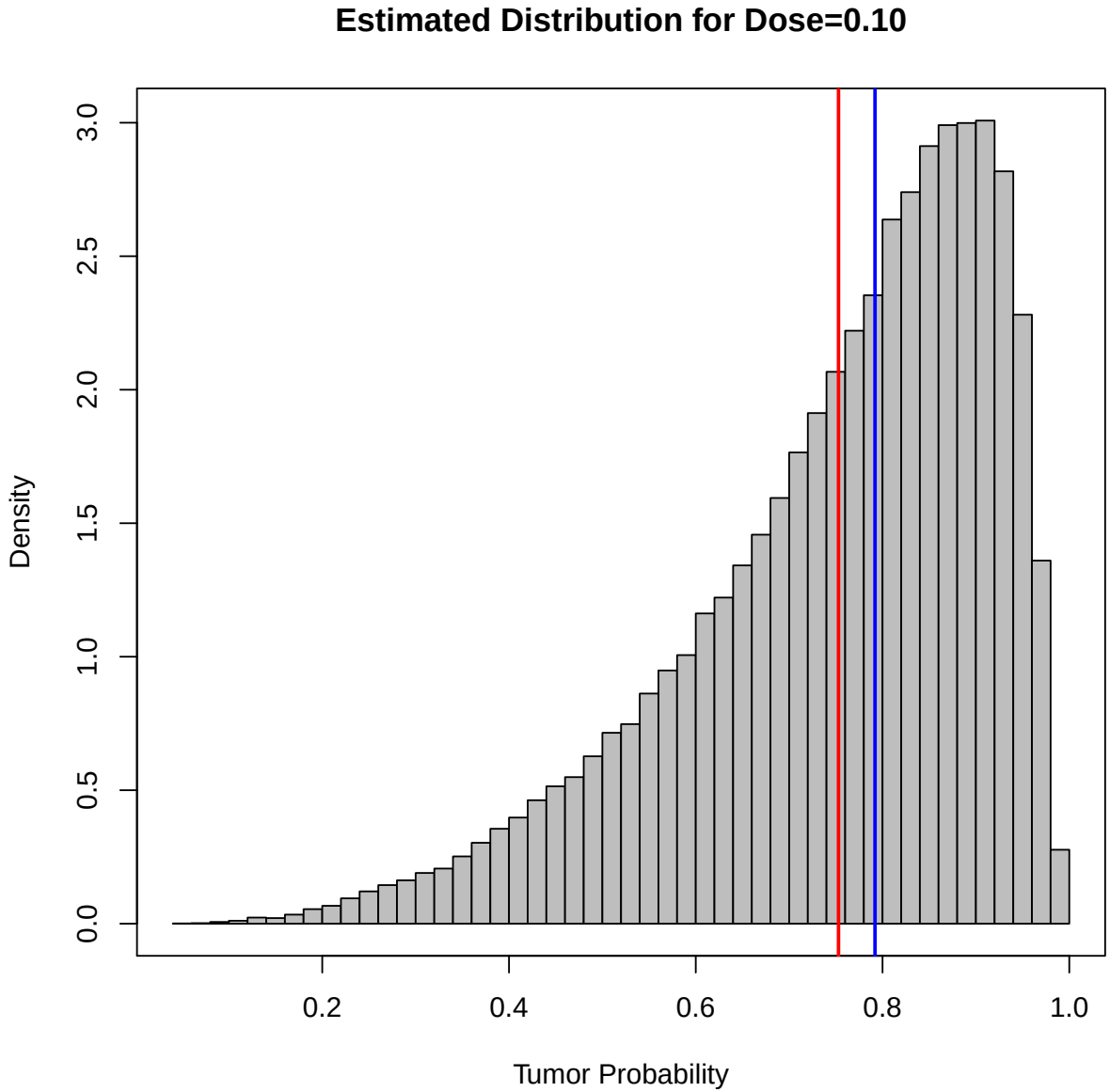
$$= \frac{\exp(\hat{\beta}_1 + \hat{\beta}_4)}{1+\exp(\hat{\beta}_1 + \hat{\beta}_4)}$$

$$= \frac{1}{1+\exp(-\hat{\beta}_1 - \hat{\beta}_4)}$$

```
> #Estimated distribution of tumor probabilities
> #for tanks treated with 0.10 units (dose 4):
>
> set.seed(5369)
> sim.tank.effects=rnorm(100000,0,sqrt(vc))
> sim.tumor.probs=1/(1+exp(-(b[1]+b[4]+sim.tank.effects)))
> hist(sim.tumor.probs,col="gray",probability=T,nclass=50,
+      ylab="Density",xlab="Tumor Probability",
+      main="Estimated Distribution for Dose=0.10")
> box()
> abline(v=1/(1+exp(-(b[1]+b[4]))),col='blue',lwd=2)
> abline(v=mean(sim.tumor.probs),col='red',lwd=2)
```


Estimated Distribution for Dose=0.10
Density
0
5
10
15
20
25
0.72
0.74
0.76
0.78
0.80
0.82
0.84
Tumor Probability

```
> #How would the picture change if the
> #tank standard deviation had been estimated
> #to be 1.0 instead of 0.0979?
>
> set.seed(5369)
> sim.tank.effects=rnorm(100000,0,1)
> sim.tumor.probs=1/(1+exp(-(b[1]+b[4]+sim.tank.effects)))
> hist(sim.tumor.probs,col="gray",probability=T,nclass=50,
+      ylab="Density",xlab="Tumor Probability",
+      main="Estimated Distribution for Dose=0.10")
> box()
> abline(v=1/(1+exp(-(b[1]+b[4]))),col='blue',lwd=2)
> abline(v=mean(sim.tumor.probs),col='red',lwd=2)
```


Estimated Distribution for Dose=0.10
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.2
0.4
0.6
0.8
1.0
Tumor Probability

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \delta_i + u_{ij}$$

$$\begin{aligned}
E(\pi_{ij}) &= E[\text{logit}^{-1}(\delta_i + u_{ij})] \\
&= E\left[\frac{1}{1+\exp\{-(\delta_i + u_{ij})\}}\right] \\
&\neq \frac{1}{1+\exp\{-(\delta_i + E[u_{ij}])\}} \\
&= \frac{1}{1+\exp(-\delta_i)} \\
&= \text{logit}^{-1}(\delta_i) \\
&= E(\pi_{ij}|u_{ij}=0).
\end{aligned}$$

Is the expected probability of a tumor the same for dose 4 and 5?

$$H_0 : E(\pi_{4j}) = E(\pi_{5j})$$

$$\iff \quad H_0 : E\left[\frac{1}{1+\exp\{-(\delta_4 + u_{4j})\}}\right] = E\left[\frac{1}{1+\exp\{-(\delta_5 + u_{5j})\}}\right]$$

$$\iff \quad H_0 : \delta_4 = \delta_5$$

$$\iff \quad H_0 : \delta_4 - \delta_5 = 0$$

Test of $H_0 : \delta_4 - \delta_5 = 0$

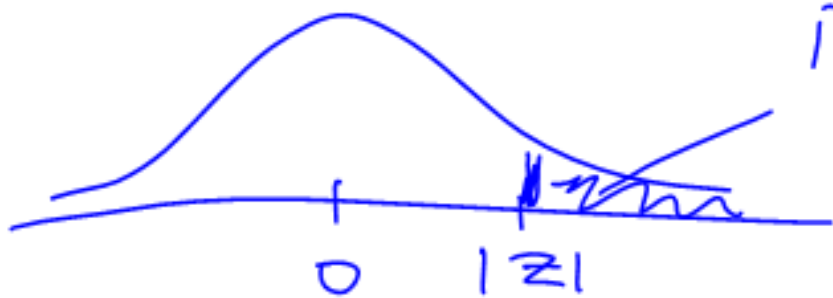
$\underset{\sim}{c}' = (0, 0, 0, 1, -1)$

$\underset{\sim}{c}'\hat{\underset{\sim}{\beta}} = \hat{\beta}_4 - \hat{\beta}_5$

```
> cc=c(0,0,0,1,-1)
> est=drop(t(cc)%*%b)
> est
[1] -0.372775
> se=drop(sqrt(t(cc)%*%v%*%cc))
> z.stat=est/se
> z.stat
[1] -1.763663
> p.value=2*(1-pnorm(abs(z.stat),0,1))
> p.value
[1] 0.07778872
```

$\sqrt{\underset{\sim}{c}' \widehat{VAR}(\hat{\underset{\sim}{\beta}})\underset{\sim}{c}}$

$z = \dfrac{\underset{\sim}{c}'\hat{\underset{\sim}{\beta}}}{\sqrt{\underset{\sim}{c}' \widehat{VAR}(\hat{\underset{\sim}{\beta}})\underset{\sim}{c}}}$



P-VALUE / 2

Confidence Interval for $\delta_4 - \delta_5$

$$\underset{\sim}{c}'\hat{\underset{\sim}{\beta}} \pm 2\sqrt{\underset{\sim}{c}'\widehat{VAR}(\hat{\underset{\sim}{\beta}})\underset{\sim}{c}}$$

```
> est+c(-2,2)*se
[1] -0.79550322  0.04995314
```

How should we interpret this confidence interval?

Let u denote the random effect associated with any particular randomly selected tank.

Let π_k denote the probability that a randomly selected fish in the randomly selected tank will develop a tumor if the tank is treated with dose $= k$ (for $k = 4, 5$).

Our model says $\log\left(\frac{\pi_4}{1-\pi_4}\right) = \delta_4 + u$ and $\log\left(\frac{\pi_5}{1-\pi_5}\right) = \delta_5 + u$. Thus,

$$\begin{aligned}
\delta_4 - \delta_5 &= (\delta_4 + u) - (\delta_5 + u) \\
&= \log\left(\frac{\pi_4}{1-\pi_4}\right) - \log\left(\frac{\pi_5}{1-\pi_5}\right) \\
&= \log\left(\frac{\pi_4}{1-\pi_4} \Big/ \frac{\pi_5}{1-\pi_5}\right),
\end{aligned}$$

which implies $\frac{\pi_5}{1-\pi_5} = \exp(\delta_5 - \delta_4)\frac{\pi_4}{1-\pi_4}$.

Confidence Interval for $\exp(\delta_5 - \delta_4)$

```
> exp(-est)
[1] 1.451758
>
> exp(-est+c(-2,2)*se)
[1] 0.951274 2.215556
```

$\exp(\underline{c}'\hat{\underline{\beta}})$

$\exp(\underline{c}'\hat{\underline{\beta}} \pm 2\sqrt{\underline{c}'\widehat{\text{VAR}}(\hat{\underline{\beta}})\underline{c}})$

For any given tank, the odds of tumor formation when dose $= 5$ are estimated to be 1.45 times the odds of tumor formation when dose $= 4$. An approximate 95% confidence interval for this within-tank multiplicative effect is $(0.95, 2.22)$.

Previous Analysis of the Tumor Data

- During our previous analysis of the tumor example, we fit a Generalized Linear Model with a different binomial success probability for each dose. (Call this the full model.)
- To test

$$H_0 : \textit{full model is adequate}$$

we examined the residual deviance statistic

$$-2 \log \Lambda_{f,s} = 2\hat{\ell}_s - 2\hat{\ell}_f,$$

where $\hat{\ell}_s$ and $\hat{\ell}_f$ are the log likelihood maximized under the saturated and full models, respectively.

Previous Analysis of the Tumor Data (continued)

- The residual deviance statistic $-2\log\Lambda_{f,s} = 2\hat{\ell}_s - 2\hat{\ell}_f$ is approximately distributed as $\chi^2_{n-p_f}$ under H_0, where $n = 20$ is the dimension of the saturated model parameter space and $p_f = 5$ is the dimension of the full model parameter space.
- Because the observed value of $-2\log\Lambda_{f,s} = 2\hat{\ell}_s - 2\hat{\ell}_f$ was unusually large for a $\chi^2_{n-p_f}$ random variable, we detected lack of fit.
- If the lack of fit is due to overdispersion, we now have two different strategies for managing overdispersion.

Strategy 1: Use a Quasi-Likelihood (QL) approach.

- Estimate an overdispersion parameter ϕ by

$$\hat{\phi} = \frac{2\hat{\ell}_s - 2\hat{\ell}_f}{n - p_f} = \frac{\sum_{i=1}^{n} d_i^2}{n - p_f} \text{ or } \hat{\phi} = \frac{\sum_{i=1}^{n} r_i^2}{n - p_f}.$$

- To test a reduced model (r) vs. the full model (f), compare

$$\frac{(2\hat{\ell}_f - 2\hat{\ell}_r)/(p_f - p_r)}{\hat{\phi}}$$

to an F distribution with $p_f - p_r$ and $n - p_f$ degrees of freedom.

Strategy 1: Quasi-Likelihood (QL) (continued)

- To test $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$, compare

$$\frac{(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'[\boldsymbol{C}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{C}']^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})/\text{rank}(\boldsymbol{C})}{\hat{\phi}}$$

  to an F distribution with $\text{rank}(\boldsymbol{C})$ and $n - p_f$ degrees of freedom.

Strategy 1: Quasi-Likelihood (QL) (continued)

- To test $H_0 : \boldsymbol{c}'\boldsymbol{\beta} = d$, compare

$$\frac{\boldsymbol{c}'\hat{\boldsymbol{\beta}} - d}{\sqrt{\hat{\phi}\,\boldsymbol{c}'\,\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\,\boldsymbol{c}}}$$

  to a t distribution with $n - p_f$ degrees of freedom.

Strategy 1: Quasi-Likelihood (QL) (continued)

- To obtain a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{c}'\boldsymbol{\beta}$, use

$$\boldsymbol{c}'\hat{\boldsymbol{\beta}} \pm t_{n-p_f,1-\alpha/2}\sqrt{\hat{\phi}\,\boldsymbol{c}'\,\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\,\boldsymbol{c}},$$

where $t_{n-p_f,1-\alpha/2}$ is the $1-\alpha/2$ quantile of the t distribution with $n-p_f$ degrees of freedom.

Strategy 1: Quasi-Likelihood (QL) (continued)

- If you use a QL approach, you are not fitting a new model.
- Rather, you are adjusting the inference strategy to account for overdispersion in the data relative to the original full model you fit.
- There is no point in re-testing for overdispersion once you decide to use the QL inference strategy.
- Data are still overdispersed relative to the fitted model, but the QL inference strategy adjusts for overdispersion to get tests with approximately the right size and confidence intervals with closer to nominal coverage rates (in theory).

Strategy 2: Fit a GLMM

- A GLMM with a random effect for each observation is one natural model for overdispersed data.
- We do not re-test for overdispersion once we decide to use a GLMM with a random effect for each observation because the model we are fitting allows for overdispersed data.
- Because GLMM inference relies on asymptotic normal and chi-square approximations, it may be more liberal (p-values smaller and confidence intervals narrower) than the QL approach, especially for small datasets.

Another Reason for Choosing a GLMM

- In the last description of the experiment to study the effects of an anti-fungal chemical, plants are the experimental units and leaves are the observational units.
- As was the case for experiments with normally distributed responses, a random effect for each experimental unit should be included in a model for the data when there is more than one observation per experimental unit.
- In the model for the infection count data, the plant random effects allow for correlation between the responses of leaves from the same plant.