A Generalized Linear Model for Binomial Response Data

Copyright ©2017 Dan Nettleton (Iowa State University)

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

As an example, consider 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.

```
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
```

One way to analyze this dataset would be to convert the binomial counts and totals into Bernoulli responses.

For example, the first line of the data set could be converted into 9 ones and $87 - 9 = 78$ zeros. Each of these 87 observations would have dose 0.01 as their explanatory variable value.

We could then use the logistic regression modeling strategy for Bernoulli response as described before.

A simpler and equivalent way to deal with this data is to consider a logistic regression model for the binomial counts directly.

A Logistic Regression Model for Binomial Count Data

For all $i = 1, \ldots, n$,

$$y_i \sim \text{binomial}(m_i, \pi_i),$$

where m_i is a known number of trials for observation i,

$$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})},$$

and $y_1, \ldots, y_n$ are independent.

The Binomial Distribution

Recall that for $y_i \sim \text{binomial}(m_i, \pi_i)$, the probability mass function of y_i is

$$P(y_i = y) = \begin{cases} \binom{m_i}{y} \pi_i^y (1-\pi_i)^{m_i - y} & \text{for } y \in \{0, \ldots, m_i\} \\ 0 & \text{otherwise} \end{cases},$$

$$E(y_i) = m_i \pi_i, \quad \text{and} \quad \text{Var}(y_i) = m_i \pi_i (1 - \pi_i).$$

The Binomial Log Likelihood

The binomial log likelihood function is

$$\begin{aligned}
\ell(\boldsymbol{\beta} \mid \boldsymbol{y}) &= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \ \log(1-\pi_i) \right] \\
&\qquad + \text{constant} \\
&= \sum_{i=1}^{n} [y_i \ \boldsymbol{x}_i'\boldsymbol{\beta} - m_i \ \log(1+\exp\{-\boldsymbol{x}_i'\boldsymbol{\beta}\})] \\
&\qquad + \text{constant.}
\end{aligned}$$

The function $\ell(\boldsymbol{\beta} \mid \boldsymbol{y})$ can be maximized over $\boldsymbol{\beta} \in \mathbb{R}^p$ using Fisher's scoring method to obtain an MLE $\hat{\boldsymbol{\beta}}$.

We can compare the fit of a logistic regression model to what is known as a *saturated model*.

The saturated model uses one parameter for each observation.

In this case, the saturated model has one free parameter π_i for each y_i.

Logistic Regression Model	Saturated Model
$y_i \sim \text{binomial}(m_i, \pi_i)$	$y_i \sim \text{binomial}(m_i, \pi_i)$
$y_1, \ldots, y_n$ independent	$y_1, \ldots, y_n$ independent
$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$ for some $\boldsymbol{\beta} \in \mathbb{R}^p$	$\pi_i \in [0, 1]$ for $i = 1, \ldots, n$ with no other restrictions
p parameters	n parameters

For all $i = 1, \ldots, n$,

the MLE of π_i under the logistic regression model is

$$\hat{\pi}_i = \frac{\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})},$$

and the MLE of π_i under the saturated model is

$$y_i/m_i.$$

Then the likelihood ratio statistic for testing the logistic regression model as the reduced model vs. the saturated model as the full model is

$$2\sum_{i=1}^{n}\left[y_i\ \log\left(\frac{y_i/m_i}{\hat{\pi}_i}\right)+(m_i-y_i)\ \log\left(\frac{1-y_i/m_i}{1-\hat{\pi}_i}\right)\right].$$

This statistic is sometimes called

the *Deviance Statistic*,

the *Residual Deviance*,

or just the the *Deviance*.

A Lack-of-Fit Test

When n is suitably large and/or $m_1, \ldots, m_n$ are each suitably large, the Deviance Statistic is approximately χ^2_{n-p} if the logistic regression model is correct.

Thus, the Deviance Statistic can be compared to the χ^2_{n-p} distribution to test for lack of fit of the logistic regression model.

Deviance Residuals

The term

$$d_i \equiv \text{sign}(y_i/m_i - \hat{\pi}_i)\sqrt{2\left[y_i \log\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - m_i\hat{\pi}_i}\right)\right]}$$

is called a *deviance residual*.

Note that the residual deviance statistic is the sum of the squared deviance residuals ($\sum_{i=1}^{n} d_i^2$).

Pearson's Chi-Square Statistic

Another lack of fit statistic that is approximately χ^2_{n-p} under the null is Pearson's Chi-Square Statistic:

$$\begin{aligned} X^2 &= \sum_{i=1}^{n} \left(\frac{y_i - \widehat{E}(y_i)}{\sqrt{\widehat{\text{Var}}(y_i)}} \right)^2 \\ &= \sum_{i=1}^{n} \left(\frac{y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}} \right)^2 . \end{aligned}$$

Pearson Residuals

The term

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

is known as a *Pearson residual*.

Note that the Pearson statistic is the sum of the squared Pearson residuals ($\sum_{i=1}^{n} r_i^2$).

Residual Diagnostics

For large m_i values, both d_i and r_i should be approximately distributed as standard normal random variables if the logistic regression model is correct.

Thus, either set of residuals can be used to diagnose problems with model fit by, e.g., identifying outlying observations.

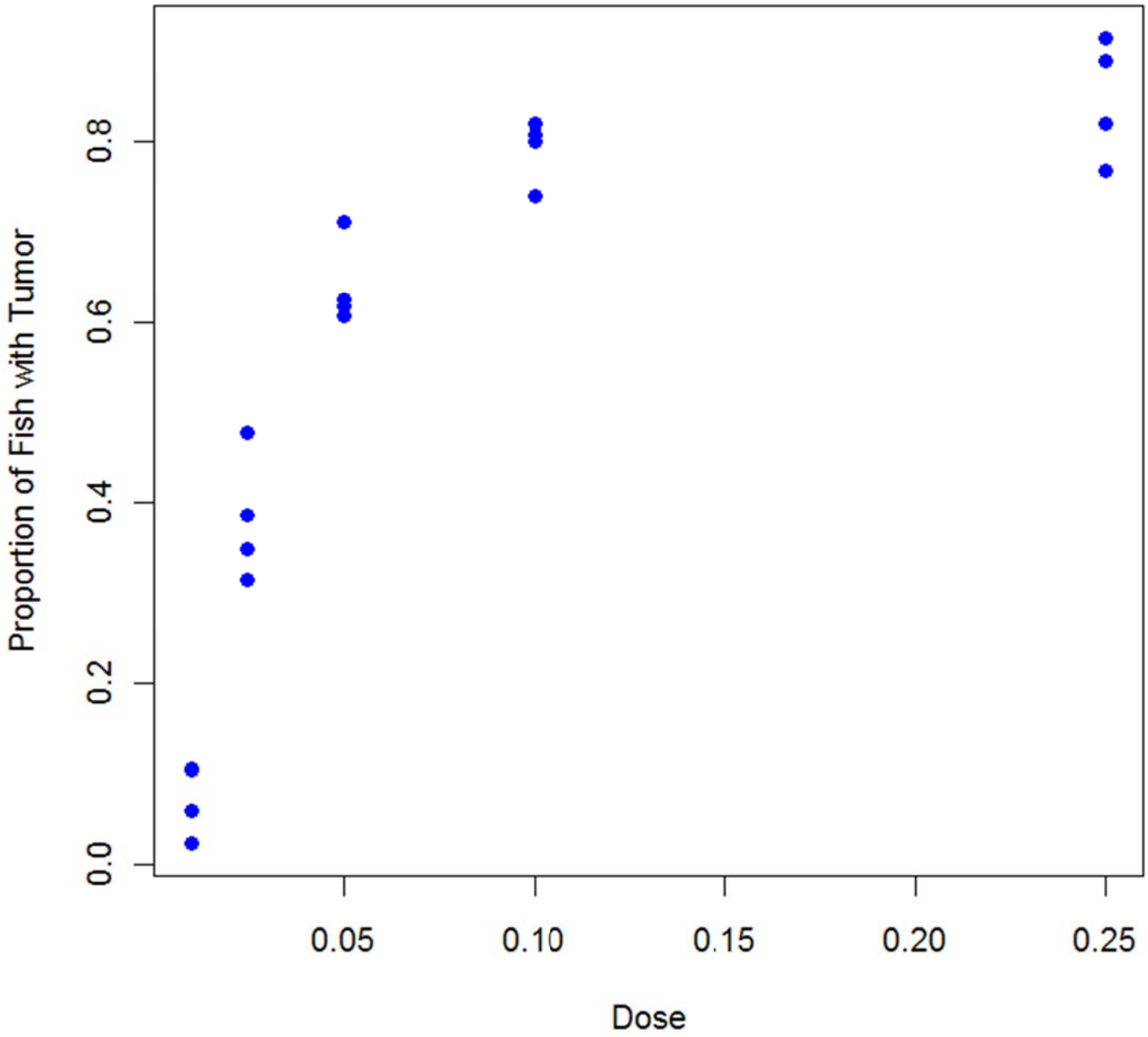
Strategy for Inference

1. Find the MLE for β using the method of Fisher Scoring, which results in an iterative weighted least squares approach.

2. Obtain an estimate of the inverse Fisher information matrix that can be used for Wald type inference concerning β and/or conduct likelihood ratio based inference of reduced vs. full models.

```
#Let's plot observed tumor proportions
#for each tank.

plot(d$dose,d$tumor/d$total,col=4,pch=19,
      xlab="Dose",
      ylab="Proportion of Fish with Tumor")
```


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

```
#Let's fit a logistic regression model
#dose is a quantitative explanatory variable.
o=glm(cbind(tumor,total-tumor)~dose,
      family=binomial(link=logit),
      data=d)

summary(o)

Call:
glm(formula = cbind(tumor, total - tumor) ~ dose,
family = binomial(link = logit),
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.3577  -4.0473  -0.1515   2.9109   4.7729
```

$y_i \quad m_i - y_i$

$y_i \sim \text{Binomial}(m_i, \pi_i)$

$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{dose}_i$

SOME PRETTY EXTREME RESIDUALS

Coefficients: $\hat{\beta}$ Wald Test Stats & p-values

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.86705	0.07673	-11.30	<2e-16	***
dose	14.33377	0.93695	15.30	<2e-16	***

(Dispersion parameter for binomial family taken to be 1)

20 - 1

Null deviance: 667.20 on 19 degrees of freedom

Residual deviance: 277.05 on 18 degrees of freedom

AIC: 368.44

20 - 2

Number of Fisher Scoring iterations: 5

$\hat{\ell}_S$ = Maximized Saturated Log Likelihood

$\hat{\ell}_0$ = Maximized Null Log Likelihood

$\hat{\ell}$ = Maximized Fitted Model Log Likelihood

$2\hat{\ell}_S - 2\hat{\ell}_0 = 667.20$

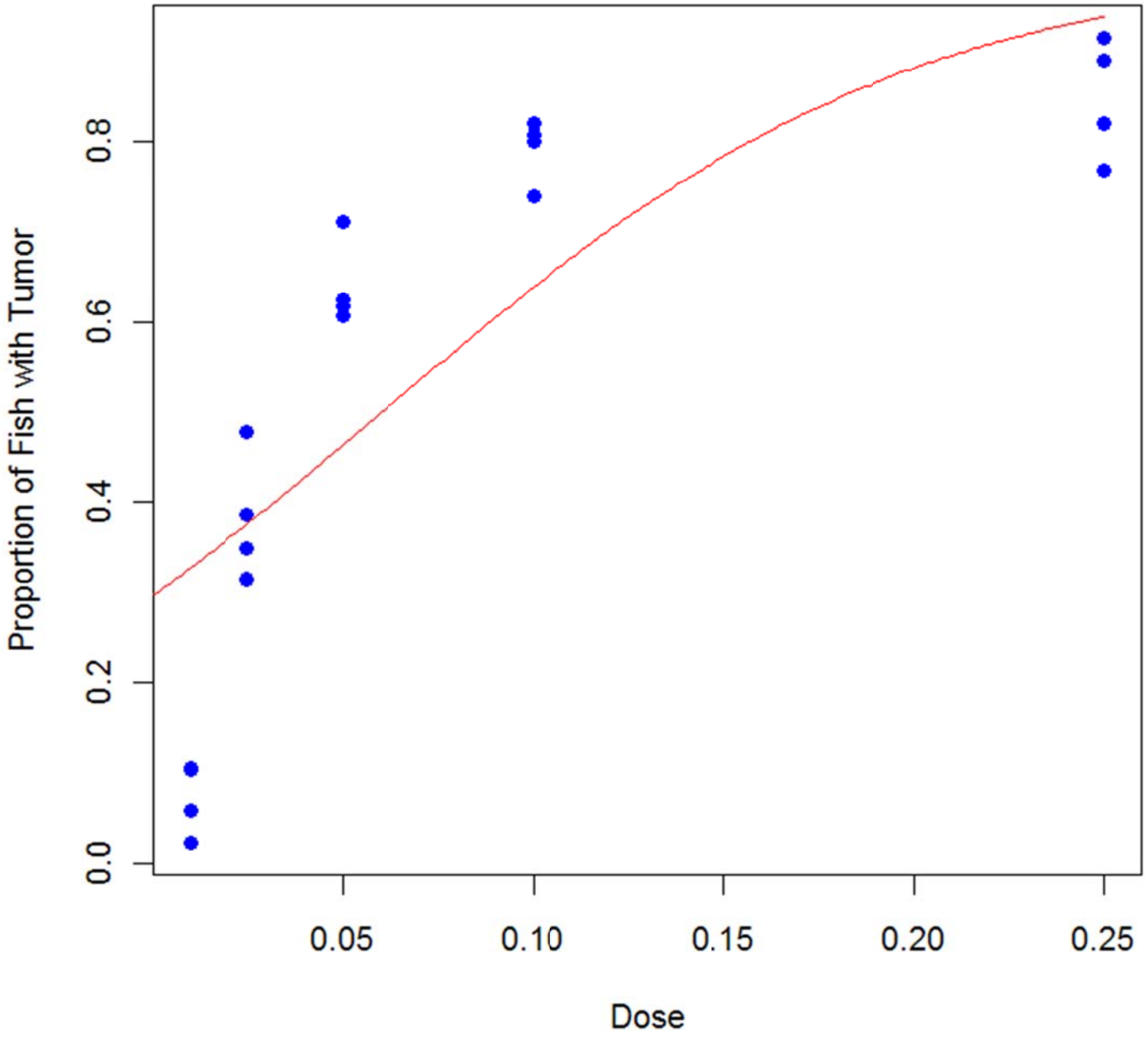
$2\hat{\ell}_S - 2\hat{\ell} = 277.05$

```
#Let's plot the fitted curve.

b=coef(o)
u=seq(0,.25,by=0.001)
xb=b[1]+u*b[2]
pihat=1/(1+exp(-xb))
lines(u,pihat,col=2,lwd=1.3)
```

$$\hat{\pi}_i = \frac{\exp(\underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}})}{1+\exp(\underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}})} = \frac{1}{\exp(-\underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}})+1}$$

$$= \left[1+\exp(-\underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}})\right]^{-1}$$


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

```
#Let's use a reduced versus full model
#likelihood ratio test to test for
#lack of fit relative to the
#saturated model.
1-pchisq(deviance(o),df.residual(o))
[1] 0
```

$2\hat{l}_s - 2\hat{l}$ 277.05 $n - p = 20 - 2$ 18

```
#We could try adding higher-order
#polynomial terms, but let's just
#skip right to the model with dose
#as a categorical variable.
```

```
d$dosef=gl(5,4)
d
```

gl function generates levels

	dose	tumor	total	dosef
1	0.010	9	87	1
2	0.010	5	86	1
3	0.010	2	89	1
4	0.010	9	85	1
5	0.025	30	86	2
6	0.025	41	86	2
7	0.025	27	86	2
8	0.025	34	88	2
9	0.050	54	89	3
10	0.050	53	86	3
11	0.050	64	90	3
12	0.050	55	88	3
13	0.100	71	88	4
14	0.100	73	89	4
15	0.100	65	88	4
16	0.100	72	90	4
17	0.250	66	86	5
18	0.250	75	82	5
19	0.250	72	81	5
20	0.250	73	89	5

```
o=glm(cbind(tumor,total-tumor)~dosef,
      family=binomial(link=logit),
      data=d)

summary(o)

Call:
glm(formula = cbind(tumor, total - tumor) ~ dosef,
family = binomial(link = logit),
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0966  -0.6564  -0.1015   1.0793   1.8513
```

$y_i \sim \text{BINOMIAL}(\pi_i)$

$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \theta_{\text{dosef}_i}$

5 SUCCESS PROBABILITIES, 1 FOR EACH DOSE

$\theta_1 = \beta_1$ (INTERCEPT)

$\theta_2 = \beta_1 + \beta_2$

$\theta_3 = \beta_1 + \beta_3$

$\theta_4 = \beta_1 + \beta_4$

$\theta_5 = \beta_1 + \beta_5$

$[1+\exp(-\beta_1)]^{-1}$

$[1+\exp(-\beta_1-\beta_2)]^{-1}$

$\vdots$

$[1+\exp(-\beta_1-\beta_5)]^{-1}$

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.5557     0.2076 -12.310   <2e-16 ***
dosef2        2.0725     0.2353   8.809   <2e-16 ***
dosef3        3.1320     0.2354  13.306   <2e-16 ***
dosef4        3.8900     0.2453  15.857   <2e-16 ***
dosef5        4.2604     0.2566  16.605   <2e-16 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 667.195  on 19  degrees of freedom
Residual deviance:  25.961  on 15  degrees of freedom
AIC: 123.36

Number of Fisher Scoring iterations: 4
```

$\hat{\beta}_1$ $\hat{\beta}_2$ $\hat{\beta}_3$ $\hat{\beta}_4$ $\hat{\beta}_5$

$2\hat{\ell}_S - 2\hat{\ell}_0$ $20-1$

$2\hat{\ell}_S - 2\hat{\ell}$ $20-5$

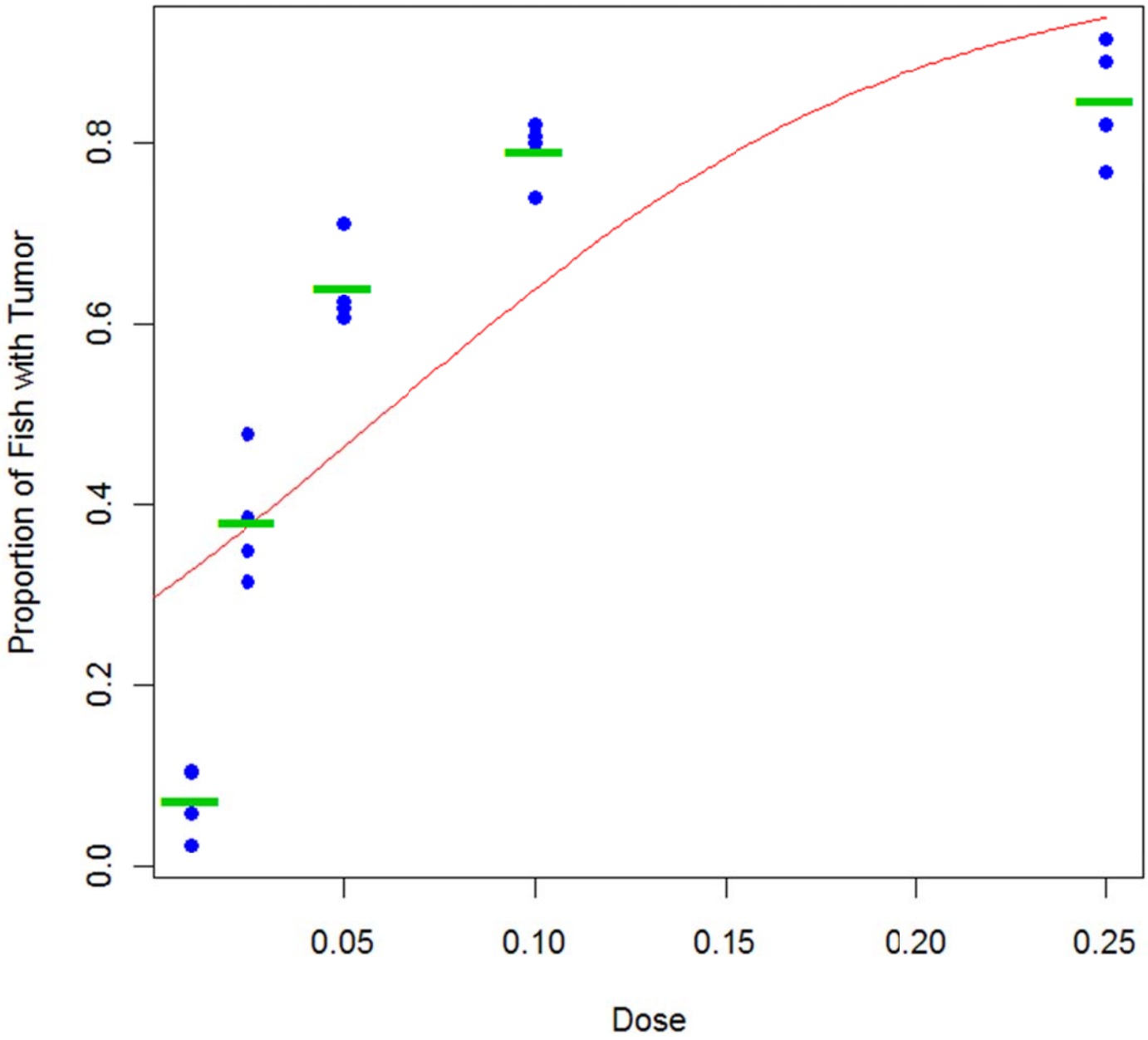
```
#Let's add the new fitted values to our plot.

fitted(o)
```

$$\hat{\pi}_i = \left[1 + \exp\left(-\underline{x}_i' \hat{\underline{\beta}}\right) \right]^{-1}$$

```
        1         2         3         4         5         6         7
0.07204611 0.07204611 0.07204611 0.07204611 0.38150289 0.38150289 0.38150289
        8         9        10        11        12        13        14
0.38150289 0.64022663 0.64022663 0.64022663 0.64022663 0.79154930 0.79154930
       15        16        17        18        19        20
0.79154930 0.79154930 0.84615385 0.84615385 0.84615385 0.84615385

points(d$dose,fitted(o),pch="_",cex=3,col=3)
```


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

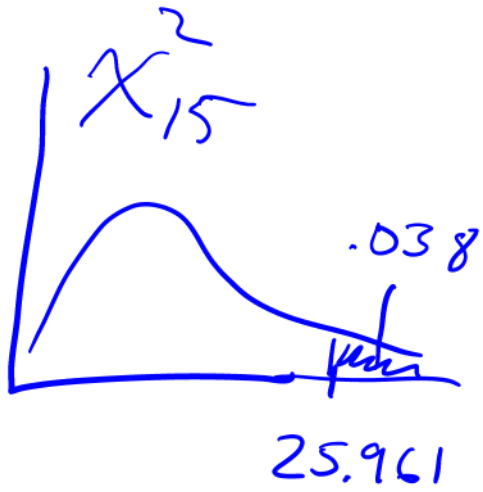
```
#The fit looks good, but let's formally
#test for lack of fit.
```

$25.961 = 2\hat{\ell}_S - 2\hat{\ell}$ $\quad n - p = 20 - 5 = 15$

```
1-pchisq(deviance(o),df.residual(o))
[1] 0.03843272

#There is still a significant lack of fit
#when comparing to the saturated model.

#The problem is over dispersion, otherwise
#known in this case as extra binomial variation.
```



Overdispersion

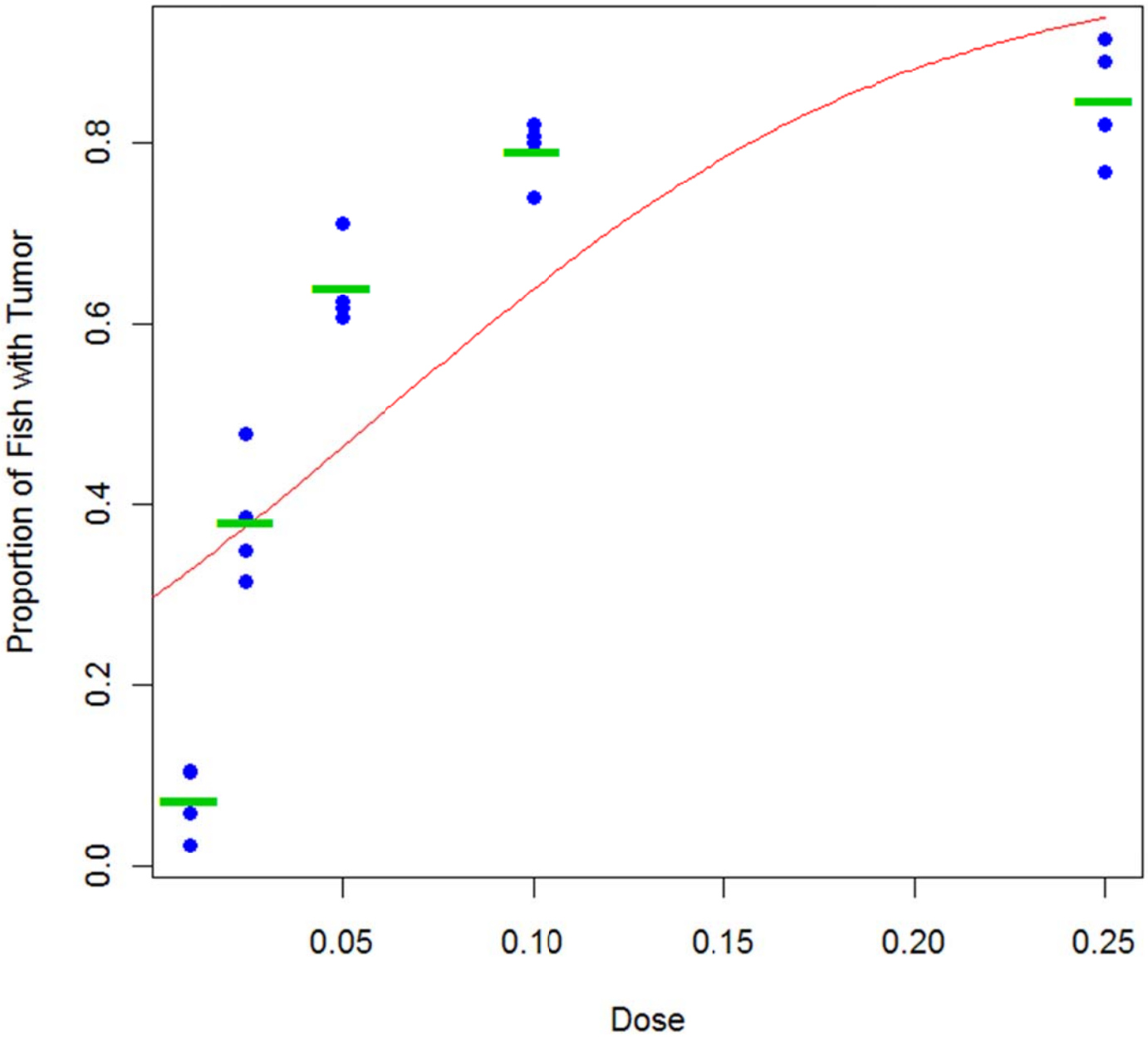
In the Generalized Linear Models framework, its often the case that $\text{Var}(y_i)$ is a function of $E(y_i)$.

That is the case for logistic regression where

$$\begin{aligned} \text{Var}(y_i) &= m_i\pi_i(1-\pi_i) = m_i\pi_i - \frac{(m_i\pi_i)^2}{m_i} \\ &= E(y_i) - [E(y_i)]^2/m_i. \end{aligned}$$

Thus, when we fit a logistic regression model and obtain estimates of the mean of the response, we get estimates of the variance of the response as well.

If the variability of our response is greater than we should expect based on our estimates of the mean, we say that there is *overdispersion*.


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

If either the Deviance Statistic or the Pearson Chi Square Statistic suggests a lack of fit that cannot be explained by other reasons (e.g., poor model for the mean or a few extreme outliers), overdispersion may be the problem.

Quasi-Likelihood Inference

If there is overdispersion, a *quasi-likelihood* approach may be used.

In the binomial case, we make all the same assumptions as before except that we assume

$$\text{Var}(y_i) = \phi \, m_i \, \pi_i(1 - \pi_i)$$

for some unknown dispersion parameter $\phi > 1$.

The dispersion parameter ϕ can be estimated by

$$\hat{\phi} = \frac{\sum_{i=1}^{n} d_i^2}{n-p}$$

or

$$\hat{\phi} = \frac{\sum_{i=1}^{n} r_i^2}{n-p}.$$

All analyses are as before except that

1. The estimated variance of $\hat{\boldsymbol{\beta}}$ is multiplied by $\hat{\phi}$.
2. For Wald type inferences, the standard normal null distribution is replaced by t with $n - p$ degrees of freedom.
3. Any test statistic T that was assumed χ^2_q under H_0 is replaced with $T/(q\hat{\phi})$ and compared to an F distribution with q and $n - p$ degrees of freedom.

These changes to the inference strategy in the presence of overdispersion are analogous to the changes that would take place in normal theory Gauss-Markov linear model analysis if we switched from assuming σ^2 were known to be 1 to assuming σ^2 were unknown and estimating it with MSE.
(Here ϕ is like σ^2 and $\hat{\phi}$ is like MSE.)

Whether there is overdispersion or not, all the usual ways of conducting generalized linear models inference are approximate except for the special case of normal theory linear models.

```
#Let's estimate the dispersion parameter.

phihat=deviance(o)/df.residual(o)
phihat
[1] 1.730745

#We can obtain the same estimate by using
#the deviance residuals.

di=residuals(o,type="deviance")
sum(di^2)/df.residual(o)
[1] 1.730745

#We can obtain an alternative estimate by
#using the Pearson residuals.

ri=residuals(o,type="pearson")
phihat=sum(ri^2)/df.residual(o)
phihat
```

$$\frac{2\hat{\ell}_s - 2\hat{\ell}}{n-p} = \frac{25.961}{15}$$

$$\frac{\sum_{i=1}^{n} d_i^2}{n-p} = \frac{25.961}{15}$$

$$\hat{\phi} = \frac{\sum_{i=1}^{n} r_i^2}{n-p}$$

```
[1] 1.671226
```

$= \hat{\phi}$

```
#Now we will conduct a quasilikelihood analysis
#that accounts for overdispersion.

oq=glm(cbind(tumor,total-tumor)~dosef,
       family=quasibinomial(link=logit),
       data=d)

summary(oq)

Call:
glm(formula = cbind(tumor, total - tumor) ~ dosef,
family = quasibinomial(link = logit),
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0966  -0.6564  -0.1015   1.0793   1.8513
```

SAME AS BEFORE

$\hat{\beta}$ IS SAME AS BEFORE

INCREASED BY MULTIPLICATIVE FACTOR OF $\sqrt{\hat{\phi}}$

DF = n - p = 15

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.5557     0.2684  -9.522 9.48e-08 ***
dosef2        2.0725     0.3042   6.814 5.85e-06 ***
dosef3        3.1320     0.3043  10.293 3.41e-08 ***
dosef4        3.8900     0.3171  12.266 3.20e-09 ***
dosef5        4.2604     0.3317  12.844 1.70e-09 ***

(Dispersion parameter for quasibinomial family taken
to be 1.671232)
```

$\hat{\phi}$

SAME AS BEFORE

```
    Null deviance: 667.195  on 19  degrees of freedom
Residual deviance:  25.961  on 15  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

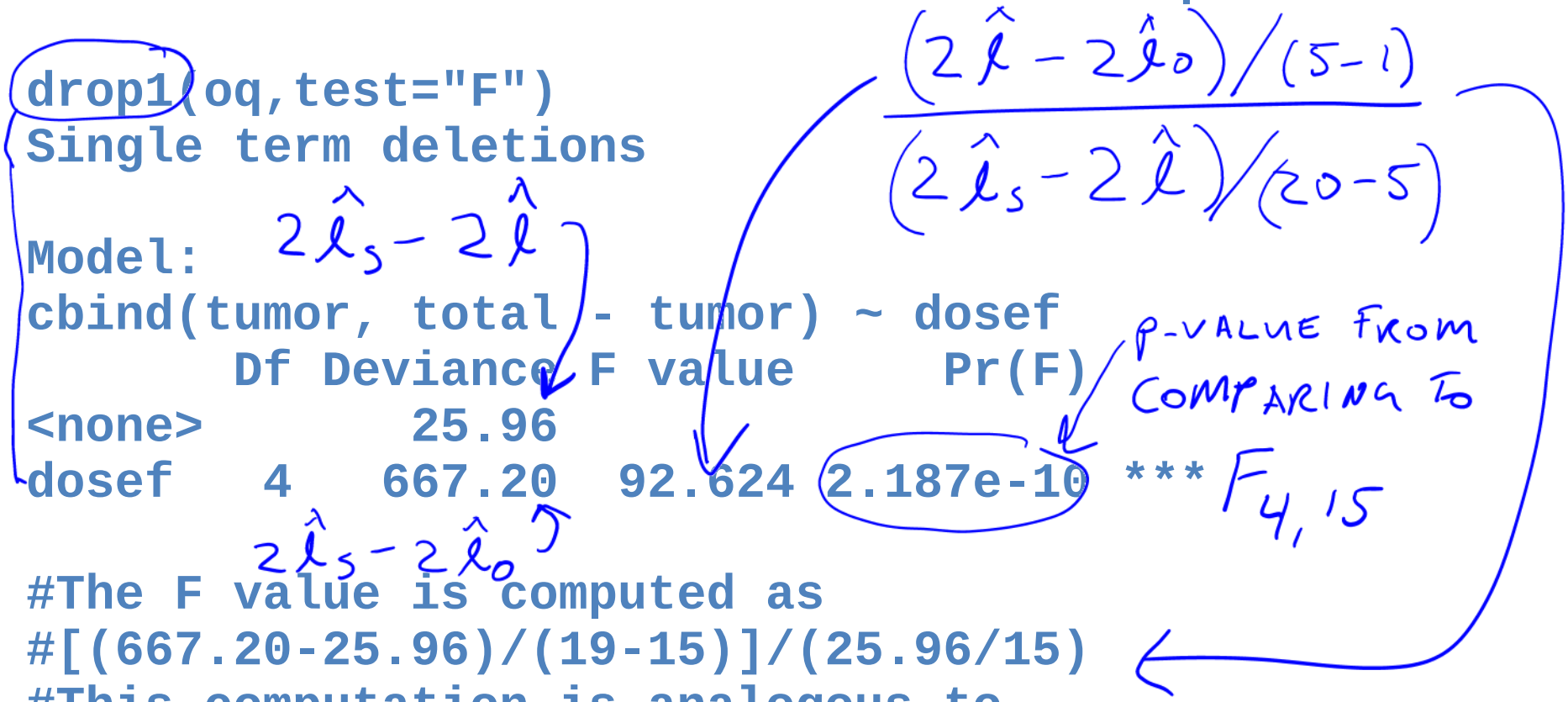
```
#Test for the effect of dose on the response.

drop1(oq,test="F")
Single term deletions

Model:
cbind(tumor, total - tumor) ~ dosef
       Df Deviance F value    Pr(F)
<none>       25.96
dosef   4   667.20  92.624 2.187e-10 ***

#The F value is computed as
#[(667.20-25.96)/(19-15)]/(25.96/15)
#This computation is analogous to
#[(SSEr-SSEf)/(DFr-DFf)]/(SSEf/DFf)
#where deviance is like SSE.

#There is strong evidence that
#the probability of tumor formation
#is different for different doses
#of the toxicant.
```



```
#Let's test for a difference between
#the top two doses.

b=coef(oq)
b
(Intercept)      dosef2      dosef3      dosef4      dosef5
  -2.555676    2.072502    3.132024    3.889965    4.260424

v=vcov(oq)
v
            (Intercept)      dosef2      dosef3      dosef4      dosef5
(Intercept)   0.0720386 -0.07203860 -0.07203860 -0.0720386 -0.0720386
dosef2       -0.0720386  0.09250893  0.07203860  0.0720386  0.0720386
dosef3       -0.0720386  0.07203860  0.09259273  0.0720386  0.0720386
dosef4       -0.0720386  0.07203860  0.07203860  0.1005702  0.0720386
dosef5       -0.0720386  0.07203860  0.07203860  0.0720386  0.1100211
```

$$\sqrt{\widehat{Var}(c'\hat{\beta})} = \sqrt{c'\,\widehat{Var}(\hat{\beta})\,c} = SE$$

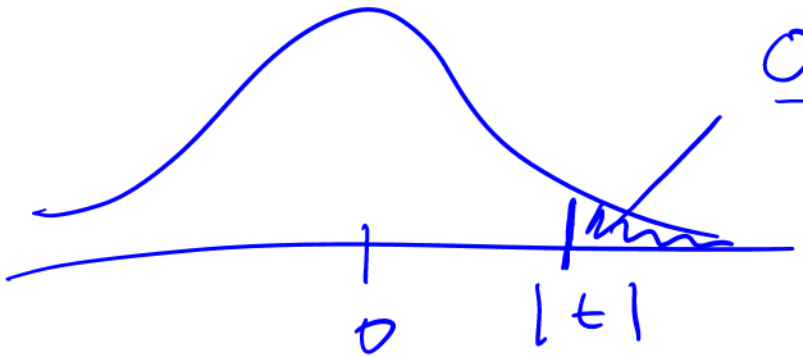
```
se=sqrt(t(c(0,0,0,-1,1))%*%v%*%c(0,0,0,-1,1))

tstat=(b[5]-b[4])/se
```

$$t = \frac{\hat{\beta}_5 - \hat{\beta}_4}{SE}$$

```
pval=2*(1-pt(abs(tstat),df.residual(oq)))

pval
0.1714103
```



$$\frac{0.1714103}{2}$$