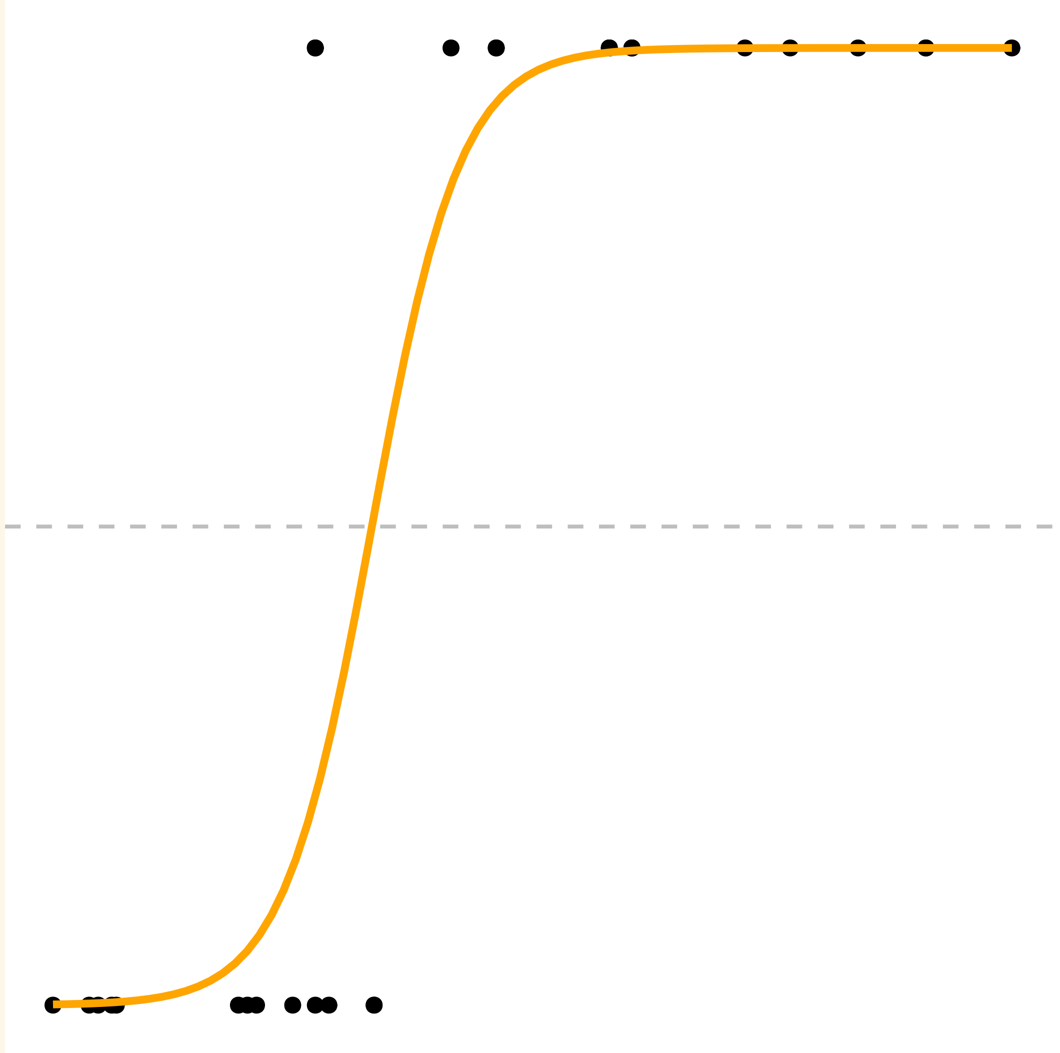
Quasi-binomial model

Stat 230

May 27 2022

Overview



Today:

Quasi models

Fitting in R

GLM model

- GLM assumes $Y \mid x \sim$ some probability distribution where

$$E(Y \mid x) = f\left(x; \beta_0, \ldots, \beta_p\right) \quad (\text{ kernel mean })$$
$$V(Y \mid x) = \sigma^2(x) \quad (\text{ some function that may depend on x})$$

Over (or under) dispersion

But what if our model's mean function is well modeled, but our variance is off

$Y \mid x \sim$ some probability distribution where

$E(Y \mid x) = f\left(x; \beta_0, \ldots, \beta_p\right)$ (kernel mean)

$V(Y \mid x) \neq \sigma^2(x)$ (some function that could involve x)

E.g. in a binomial logistic model, this means the variance of our response doesn't equal that of a binomial probability model:

$$\begin{aligned} Y \mid x &\sim \operatorname{Binom}(m_i, \pi\left(x_i\right)) \\ E(Y \mid x) &= m_i \pi\left(x_i\right) \\ V\left(Y_i \mid x_i\right) &\neq m_i \pi\left(x_i\right)\left(1-\pi\left(x_i\right)\right) \end{aligned}$$

Estimating the dispersion parameter ψ

- For a GLM, the dispersion parameter ψ ("psi") is estimated from the deviance G^2 from the regular GLM:

$$\hat{\psi} = \frac{G^2}{n - (p + 1)}$$

- $\hat{\psi} > 1$: overdispersion (responses are more variable than expected)
- $\hat{\psi} < 1$: underdispersion (responses are less variable than expected)
- e.g. for a quasi-binomial model, G^2 is the (residual) deviance from a regular binomial logistic model.

Estimating with a quasi-GLM

- Parameter estimates for β are from the regular GLM model.
- e.g. $\hat{\beta}$ from a regular binomial logistic model
- Quasi model Standard errors for $\hat{\beta}'$ s are adjusted versions of the regular GLM SE:

$$SE_{quasi}\left(\hat{\beta}_i\right) = \sqrt{\hat{\psi}} SE_{GLM}\left(\hat{\beta}_i\right)$$

- e.g. for a quasi-binomial model, $SE_{\text{binom}}\left(\hat{\beta}_i\right)$ are the usual SE from a regular binomial logistic model.

Inference with a quasi-GLM

- Conduct "z"-inference (Wald tests/CI) using SEs equal to $SE_{quasi}\left(\hat{\beta}_i\right)$
- Compare quasi-binomial models using a F-test stat equal to

$$F = \frac{\left(G^2_{\text{reduced}} - G^2_{\text{full}}\right)/(\#\text{ terms tested })}{\hat{\psi}}$$

using an F-distribution with degrees of freedom equal to the number of terms tested and $n-(p+1)$. G^2 is the model deviance from fitting the usual binomial model for two competing models.

R: Quasi-binomial model

- A quasi-binomial model is fit with

```
glm(y/m ~ x1 + x2, family = quasibinomial, weights = m, data = mydata)
```

- Model comparisons with a quasi-binomial model are done with anova:

```
anova(red_quasi, full_quasi, test = "F")
```

Example: Rake data

The USGS monitors submersed aquatic vegetation (SAV) in the Mississippi by using a long-handled rake (from a boat) to pull SAV from the river bottom.

```
RakeData <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/RakeData.c
glimpse(RakeData)
```

```
Rows: 27
Columns: 6
$ X         <int> 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 20, …
$ SiteRake  <int> 6, 6, 2, 6, 5, 6, 6, 2, 2, 4, 6, 6, 1, 6, 6, 2, 6, 6, 6, 0, …
$ SiteM     <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, …
$ SiteBiom  <dbl> 287.5731, 118.0538, 118.0381, 1064.3720, 254.8879, 225.2125,…
$ SiteDepth <dbl> 0.1000000, 0.4000000, 0.6000000, 0.8000000, 1.0833333, 0.500…
$ SiteSub   <chr> "sand", "silt", "silt", "sand", "silt", "sand", "silt", "san…
```

Data description

- Cases $=27$ sites that contain SAV
- $\mathrm{m}=$ SiteM $=6$ locations (quadrats) raked per site
- $\mathrm{Y}=$ SiteRake $=\#$ locations with SAV detected per site
- $\mathrm{X}=$ total site biomass, average water depth, substrate (soil type: silt or sand)
- $\pi(X)=$ Probability the rake detects SAV at a site with predictors X

Binomial logistic regression

```r
rake_glm <- glm(SiteRake/SiteM ~ log(SiteBiom+1) + SiteDepth + SiteSub,
                family = binomial,
                weights = SiteM,
                data = RakeData)
```

Binomial logistic regression

```
summary(rake_glm)
```

```

Call:
glm(formula = SiteRake/SiteM ~ log(SiteBiom + 1) + SiteDepth + 
    SiteSub, family = binomial, data = RakeData, weights = SiteM)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.2516  -0.3863   0.8381   1.0709   1.7355  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -1.8528     0.8319  -2.227   0.0259 *  
log(SiteBiom + 1)   0.7475     0.1157   6.461 1.04e-10 ***
SiteDepth          -1.2472     0.8175  -1.526   0.1271    
SiteSubsilt         0.4691     0.4545   1.032   0.3020    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118.36  on 26  degrees of freedom
Residual deviance:  50.44  on 23  degrees of freedom
AIC: 82.189

Number of Fisher Scoring iterations: 4
```

Goodness-of-fit

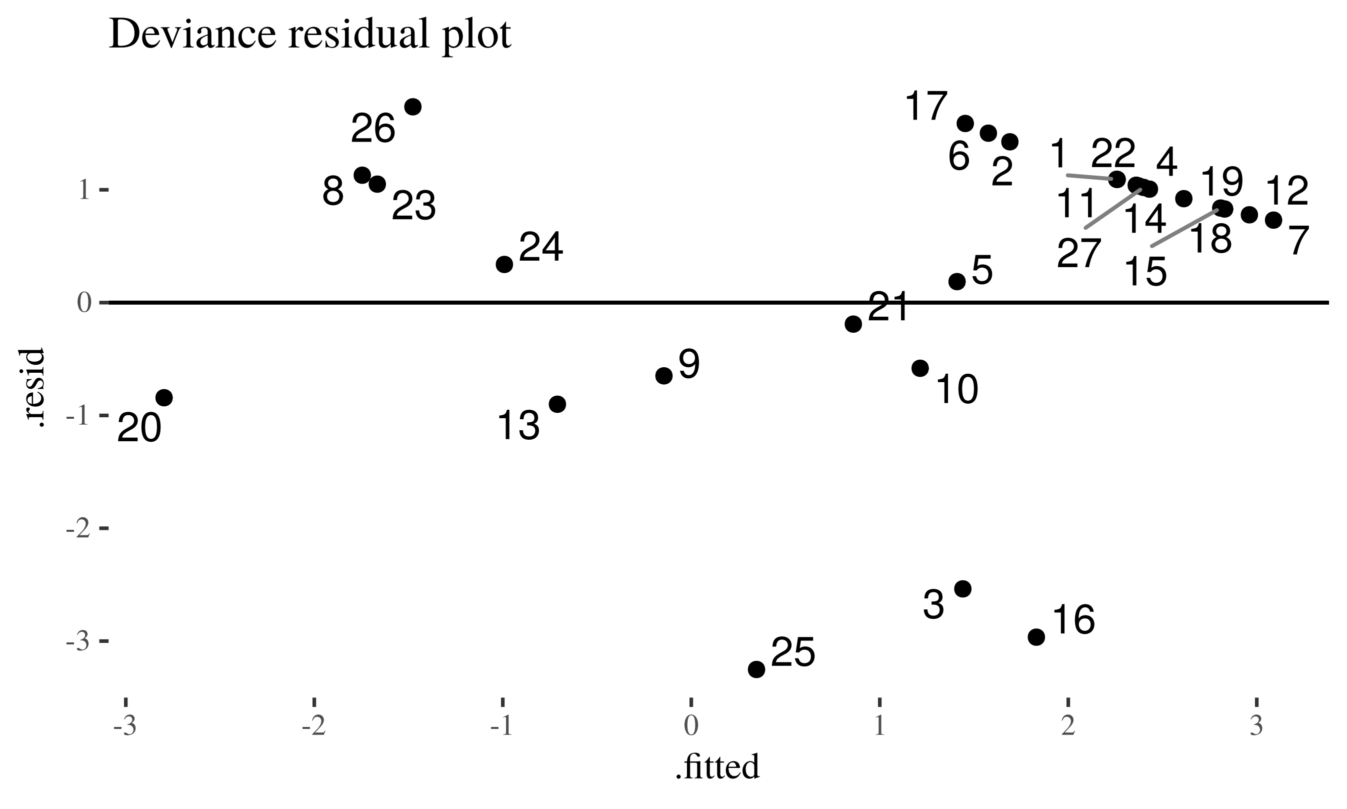
$$\begin{aligned} H_0 &: \text{ logistic model} \\ H_A &: \text{ saturated model} \end{aligned}$$

```r
1 - pchisq(50.44, df = 23)
```

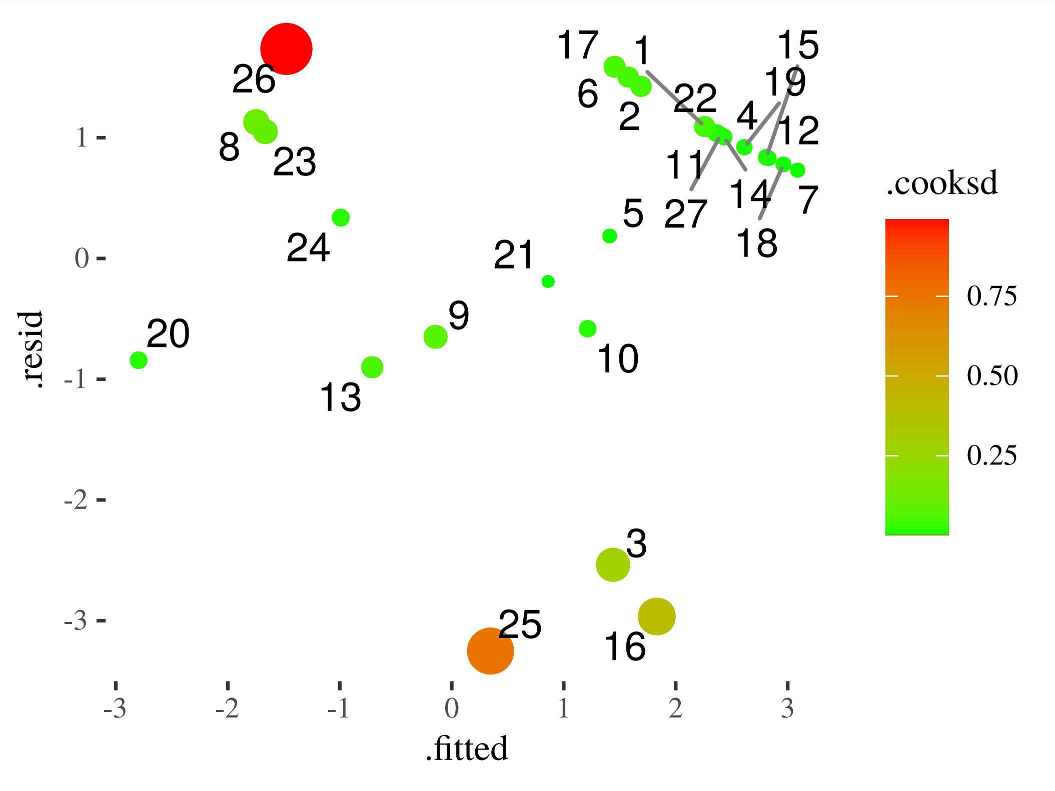
```
[1] 0.0008063917
```

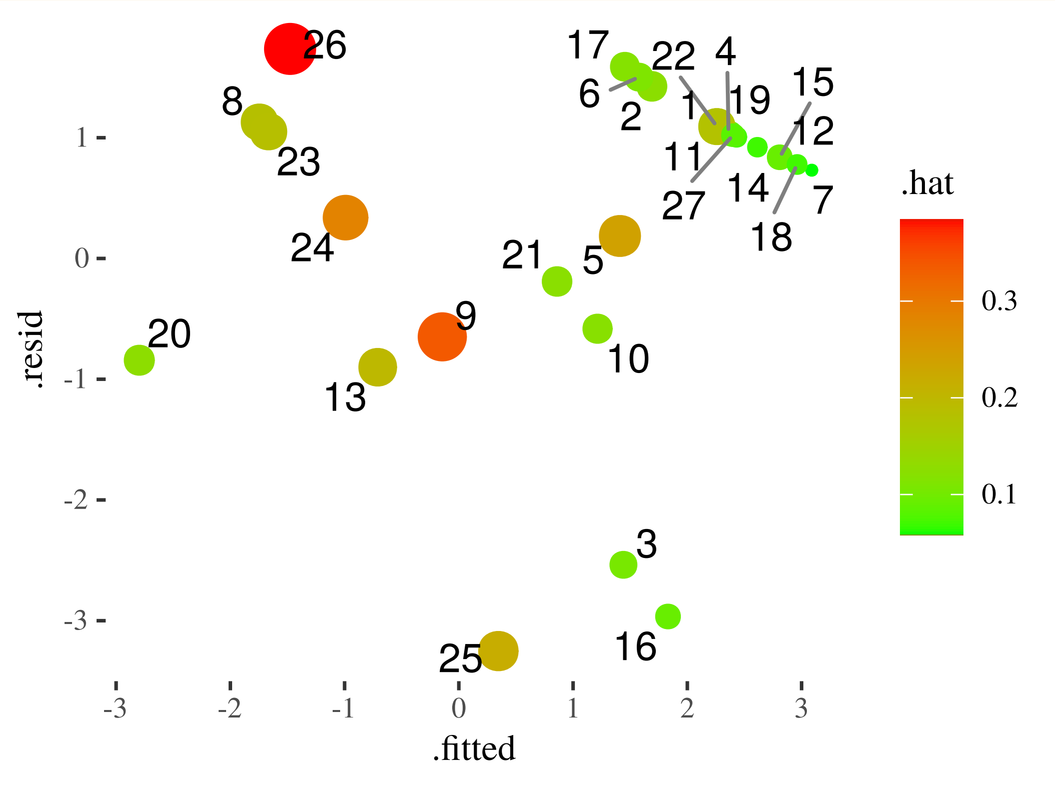
The GOF p-value for this model is 0.0008, which suggests that there is enough evidence to say the model is not adequate

Residual analysis



Leverage and Cook's distance





Estimate the dispersion parameter

```
    Null deviance: 118.36  on 26  degrees of freedom
Residual deviance:  50.44  on 23  degrees of freedom
```

$$\hat{\psi} = \frac{50.44}{23} \approx 2.1$$

The dispersion parameter is estimated as 2.14

Fit the quasi-binomial model

```r
rake_quasi_glm <- glm(SiteRake/SiteM ~ log(SiteBiom+1) + SiteDepth + SiteSub,
                      family = quasibinomial,
                      weights = SiteM,
                      data = RakeData)
```

Quasi-binomial logistic regression

```
summary(rake_quasi_glm)
```

```
Call:
glm(formula = SiteRake/SiteM ~ log(SiteBiom + 1) + SiteDepth + 
    SiteSub, family = quasibinomial, data = RakeData, weights = SiteM)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.2516  -0.3863   0.8381   1.0709   1.7355  

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -1.8528     1.2168  -1.523 0.141474    
log(SiteBiom + 1)   0.7475     0.1692   4.417 0.000199 ***
SiteDepth          -1.2472     1.1958  -1.043 0.307814    
SiteSubsilt         0.4691     0.6648   0.706 0.487527    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 2.139552)

    Null deviance: 118.36  on 26  degrees of freedom
Residual deviance:  50.44  on 23  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Standard errors

Binomial Model

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -1.8528     0.8319  -2.227   0.0259 *  
log(SiteBiom + 1)  0.7475     0.1157   6.461 1.04e-10 ***
SiteDepth         -1.2472     0.8175  -1.526   0.1271    
SiteSubsilt        0.4691     0.4545   1.032   0.3020    
```

Quasi-binomial Model

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -1.8528     1.2168  -1.523 0.141474    
log(SiteBiom + 1)  0.7475     0.1692   4.417 0.000199 ***
SiteDepth         -1.2472     1.1958  -1.043 0.307814    
SiteSubsilt        0.4691     0.6648   0.706 0.487527    
```

$$SE\left(\hat{\beta}_i^{quasi}\right) = \sqrt{2.139552} \times SE(\hat{\beta}_i)$$

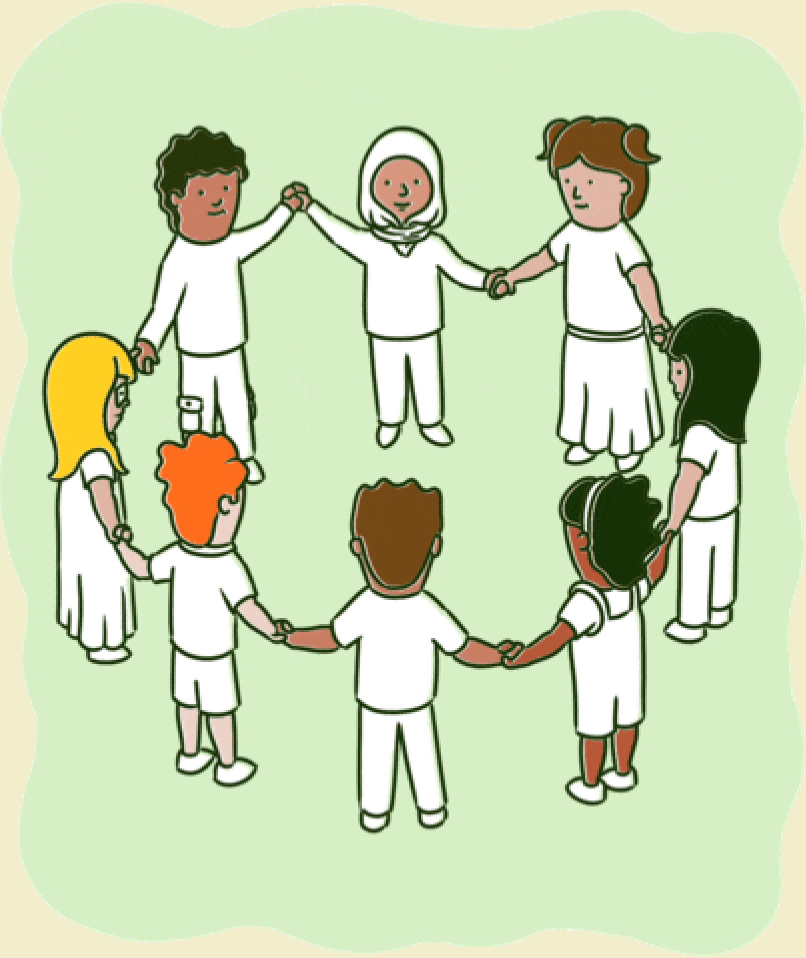
Quasi-binomial F-test

```
anova(rake_glm)
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: SiteRake/SiteM

Terms added sequentially (first to last)

                  Df Deviance Resid. Df Resid. Dev
NULL                                 26    118.357
log(SiteBiom + 1)  1   64.263        25     54.094
SiteDepth          1    2.571        24     51.522
SiteSub            1    1.083        23     50.440
```

Your Turn 1

05:00



Complete (1a-c):

Test the hypotheses (1d):

$$
\begin{aligned}
H_0 &: \beta_2 = \beta_3 = 0 \\
H_A &: \text{ at least one } \beta_2, \beta_3 \neq 0
\end{aligned}
$$

Compare models using anova with the F test in R

```
rake_quasi_glm_red <- update(rake_quasi_glm, ~ . - SiteDepth - SiteSub)
anova(rake_quasi_glm_red, rake_quasi_glm, test = "F")
Analysis of Deviance Table

Model 1: SiteRake/SiteM ~ log(SiteBiom + 1)
Model 2: SiteRake/SiteM ~ log(SiteBiom + 1) + SiteDepth + SiteSub
  Resid. Df Resid. Dev Df Deviance     F Pr(>F)
1        25     54.094
2        23     50.440  2   3.6542 0.854 0.4388
```

The p-value is from an F-distribution with 2 and 25 degrees of freedom. The test results suggest that neither depth nor substrate are statistically significant (F = 0.854, p-value=0.439).