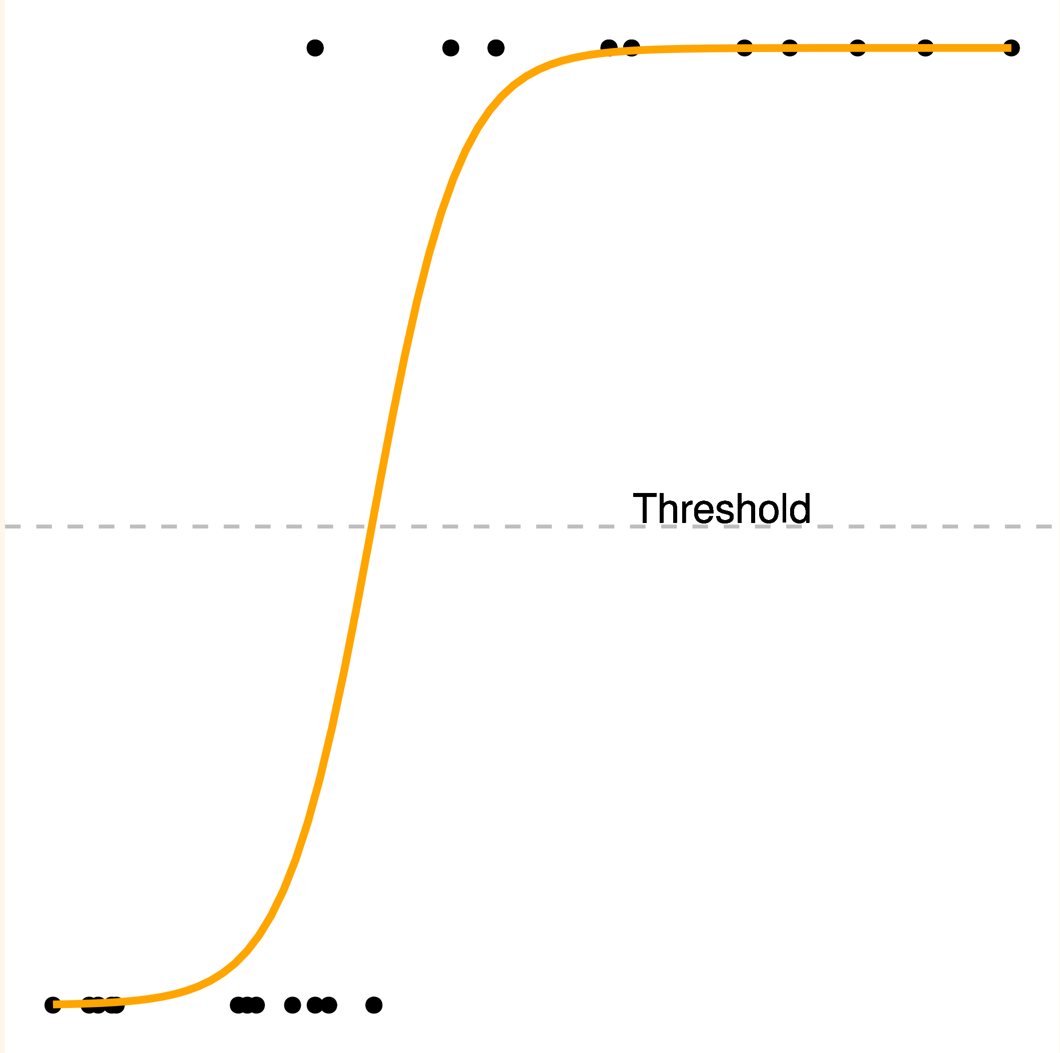
Logistic regression for binary responses: Diagnostics

Stat 230

May 16 2022

Overview



Today:

Checking log-odds linearity: log-odds plots

Residuals and Case influence stats

Model Assumptions

$$Y_i \mid X_i \overset{indep.}{\sim} \text{Bern}(\pi(X_i))$$

$$\eta_i = \log\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}$$

Independence

- how are data collected?!
- Are the cases naturally clustered together in a way that isn't accounted for in the model?
- More on this in binomial logistic regression!

Log-odds linearity

- quantitative predictors are linearly related to the log odds of success

EDA for logistic models: Empirical log odds plot

$$\eta_i = \log\left(\frac{\pi\left(X_i\right)}{1-\pi\left(X_i\right)}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}$$

Plot the empirical (sample) log-odds against the predictor and look for linearity. Get empirical log odds for binned (grouped) data

1. Group cases into groups with similar predictor values using `ntile`
2. Compute (`summarize`) the proportion of successes within each group (`group_by`)

$$\tilde{\pi}_{emp} = \frac{\text{number of successes in the group}}{\text{group size}}$$

3. Compute (`summarize`) the log odds of success in the group, within each group.

$$\text{logit}_{emp} = \ln\left(\frac{\tilde{\pi}_{emp}}{1-\tilde{\pi}_{emp}}\right)$$

Example: BWCA

1999 windstorm in **northern MN**

- what factors are associated with a tree blow down?

sample of 659 trees

- $y = 1$ means the tree died during the storm
- D: tree diameter (inches)

```
library(dplyr)
blowBF <- read.csv("https://raw.githubuserconte
mean(blowBF$y) # proportion died
[1] 0.353566
```

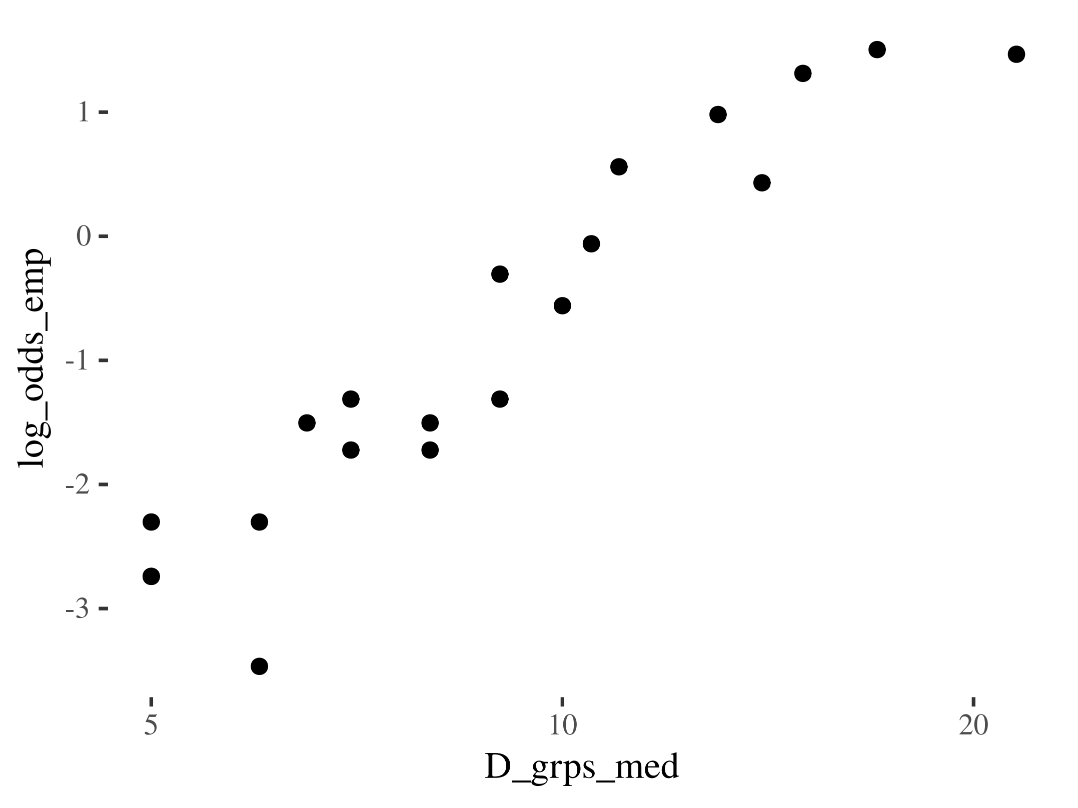
```
table(blowBF$status) # character response

    died survived
     233      426
```

```
glimpse(blowBF)
Rows: 659
Columns: 5
$ X      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1
$ D      <dbl> 9, 11, 9, 9, 5, 8, 8, 6, 8, 5, 2
$ S      <dbl> 0.0242120, 0.0305947, 0.0305947,
$ y      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ status <chr> "survived", "survived", "survive
```

Linearity in empirical log odds vs. (median) diameter?

```
ggplot(blowBF_empLO, aes(x=D_grps_med, y=log_odds_emp)) +
 geom_point() +
scale_x_log10()  # log is better
```



Residuals in logistic models

- **Response:** response minus estimated mean

$$r_i = y_i - \hat{\pi}\left(X_i\right)$$

- **Pearson:** response residuals standardized based on the binomial SD:

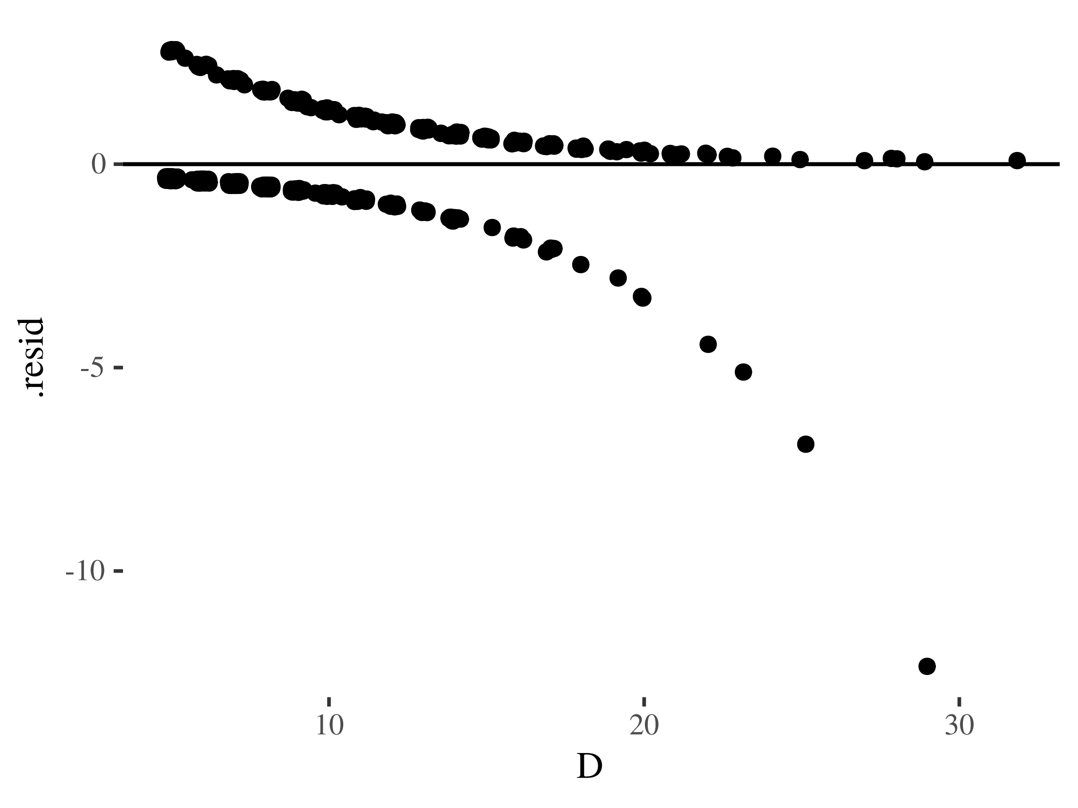
$$pr_i = \frac{y_i - \hat{\pi}\left(X_i\right)}{\sqrt{\hat{\pi}\left(X_i\right)\left(1 - \hat{\pi}\left(X_i\right)\right)}}$$

- **Deviance:** each case's contribution to the residual deviance

$$\mathrm{Dres}_i = \mathrm{sign}(y_i - \hat{\pi}\left(X_i\right))\sqrt{2\left[y_i \ln\left(\frac{y_i}{\hat{\pi}\left(X_i\right)}\right) + (1 - y_i)\ln\left(\frac{1 - y_i}{1 - \hat{\pi}\left(X_i\right)}\right)\right]}$$

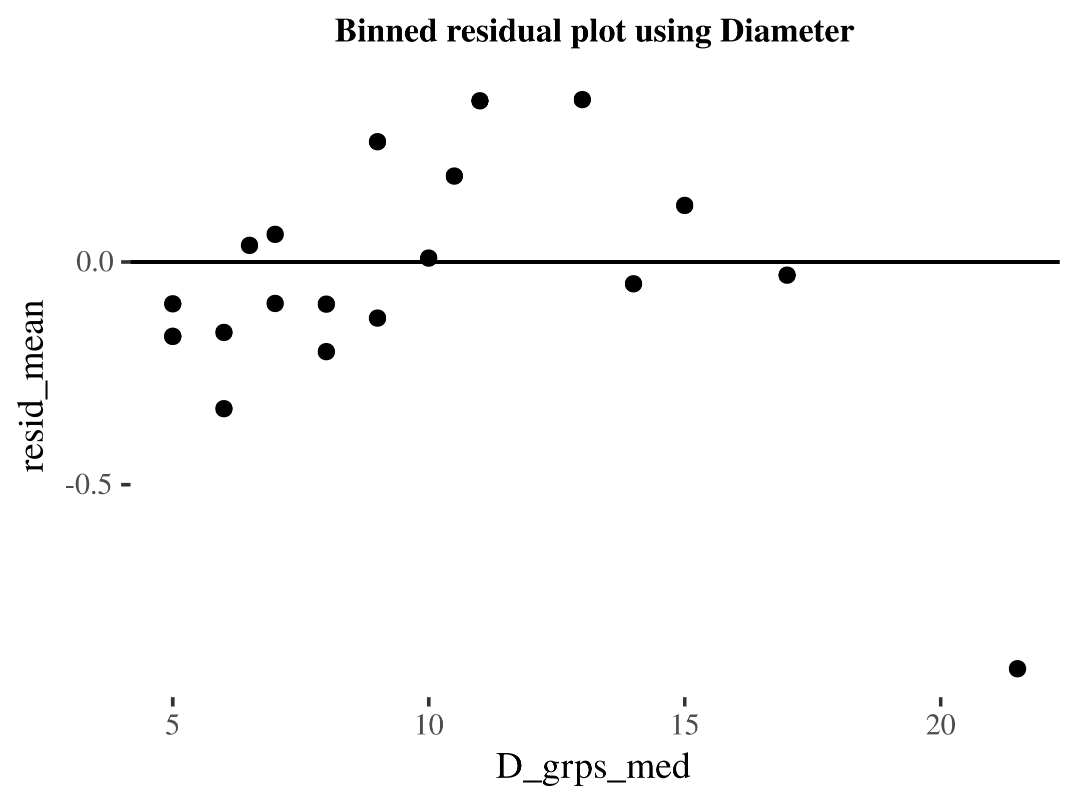
Interpretation of binary model's residual plots

```r
fir_glm1 <- glm(y ~ D, family=binomial, data=blowBF)
blowBF_aug1 <- augment(fir_glm1, data=blowBF,
                       type.residuals="pearson")
ggplot(blowBF_aug1, aes(x=D, y=.resid)) +
geom_jitter(height = .05) +
geom_hline(yintercept = 0)
```



Binned response residuals

```
ggplot(blowBF_resid1, aes(x=D_grps_med, y=resid_mean)) +
 geom_point() +
 geom_hline(yintercept = 0) +
 labs(title="Binned residual plot using Diameter")+
  theme(plot.title = element_text(hjust=0.5, size=9, face='bol
```



Low diameter cases overestimated and higher diameters underestimated

Case influence in logistic models

leverage and Cook's distance are used with logistic models, just as they are with regular linear models.

Use the usual R commands:

- `plot(my_glm, which = 5)` (or type 4)
- `ggResidpanel::resid_panel(my_glm, plots =c("cookd", "lev"))`

Example: BWCA analysis

```
# y = 1 means died
fir_glm2 <- glm(y ~ log(D), family=binomial, data=blowBF)
tidy(fir_glm2, conf.int=TRUE)
# A tibble: 2 × 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)    -7.89     0.633     -12.5 9.92e-36    -9.18     -6.69
2 log(D)          3.26     0.276      11.8 3.02e-32     2.74      3.82
```

- The odds of death as a function of diameter?
- Effect of doubling tree diameter?

Example: BWCA analysis

```
tidy(fir_glm2, conf.int=TRUE)
# A tibble: 2 × 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)    -7.89     0.633     -12.5 9.92e-36    -9.18     -6.69
2 log(D)          3.26     0.276      11.8 3.02e-32     2.74      3.82
```

- The odds of death as a function of diameter:

$$\widehat{odds}(D) = e^{-7.89+3.26\ln(D)} = e^{-7.89}D^{3.26}$$

Example: BWCA analysis

```
tidy(fir_glm2, conf.int=TRUE)
# A tibble: 2 × 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)    -7.89     0.633     -12.5 9.92e-36    -9.18     -6.69
2 log(D)          3.26     0.276      11.8 3.02e-32     2.74      3.82
```

Doubling diameter is associated with a 9.58-fold increase in the odds of death (95% CI 6.68 to 14.12).

$$m^{\hat{\beta}_1} = 2^{3.26} = 9.58$$

```
2^3.26 # estimate
[1] 9.57983
2^2.74 # lower bound
[1] 6.680703
2^3.82 # upper bound
[1] 14.12325
```

Example: BWCA analysis

```
anova(fir_glm2)
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev
NULL                     658     856.21
log(D)  1   200.97       657     655.24
```

Example: BWCA with status response

- status's second level is survived
- a model with status as the response will model the probability of survival!
  - `glm` will want this `as.factor` in order to fit the model!

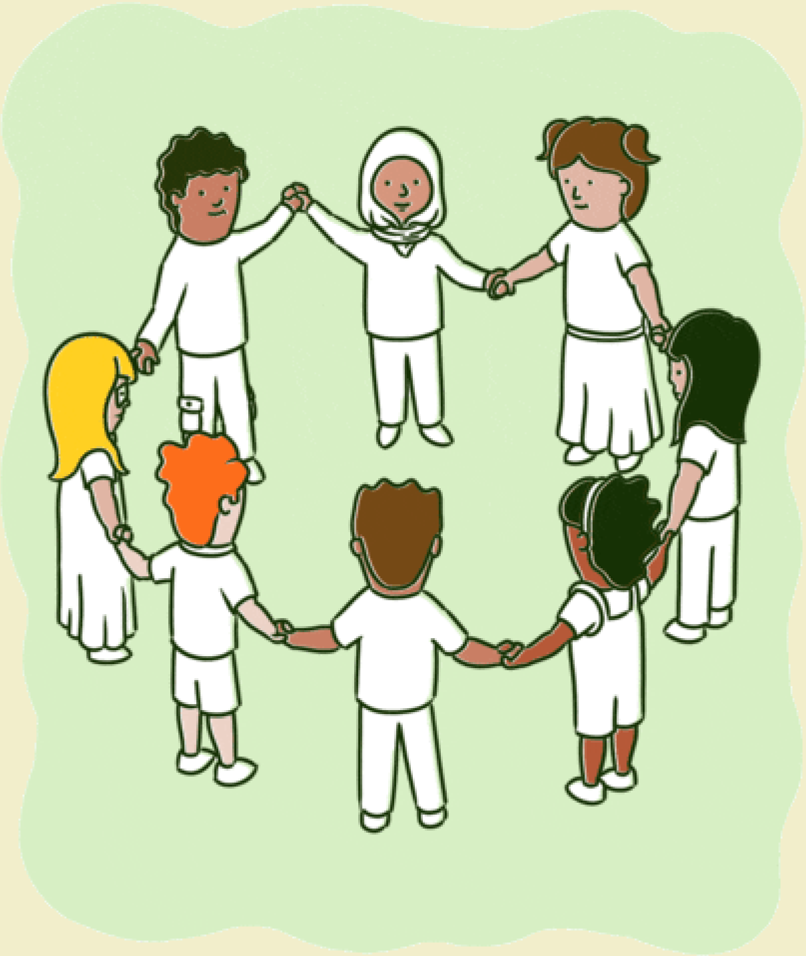
```
table(blowBF$status)

    died survived 
     233      426 
```

```
fir_glm2_status <- glm(as.factor(status) ~ log(D), family=binomial, data=blowBF)
tidy(fir_glm2_status, conf.int=TRUE)
# A tibble: 2 × 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)     7.89     0.633      12.5 9.92e-36     6.69      9.18
2 log(D)         -3.26     0.276     -11.8 3.02e-32    -3.82     -2.74
```

Your Turn 1

05:00



- Go over to the in class activity file
- Go over the class activity in your group