Binary Regression

GH Chapter 5, ISL Chapter 4

January 31, 2017

Seedling Survival

Tropical rain forests have up to 300 species of trees per hectare, which leads to difficulties when studying processes which occur at the community level. To gain insight into species responses, a sample of seeds were selected from a suite of eight species selected to represent the range of regeneration types which occur in this community.

Name	Size	Cotyledon type
Ardisia	3	H
C. biflora	7	H
Gouania	1	E
Hirtella	8	H
Inga	4	H
Maclura	2	E
C. racemosa	6	H
Strychnos	5	E

Size = 1 smallest to 8 largest
E = Epigeal - cotyledons
H = Hypogeal - seed food reserves

Experimental Design

This representative community was then placed in experimental plots manipulated to mimic natural conditions

- 8 PLOTs: 4 in forest gaps, 4 in understory conditions
- Each plot split in half: mammals were excluded from one half with a CAGE
- 4 subplots within each CAGE/NO CAGE
- 6 seeds of each SPECIES plotted in each SUBPLT
- 4 LITTER levels applied to each SUBPLT
- LIGHT levels at forest floor recorded
- SURV an indicator of whether they germinated and survived was recorded

Which variables are important in determining whether a seedling will survive? Are there interactions that influence survival probabilities?

Modeling Survival

Distribution for Survival of a single Seedling is a Bernoulli random variable

$$\mathrm{E}[\mathrm{SURV}_i \mid \textit{covariates}] = \pi_i$$

How should we relate covariates to probability of survival?
For example, probability of survival may depend on whether there was a CAGE to prevent animals from eating the seedling or LIGHT levels.

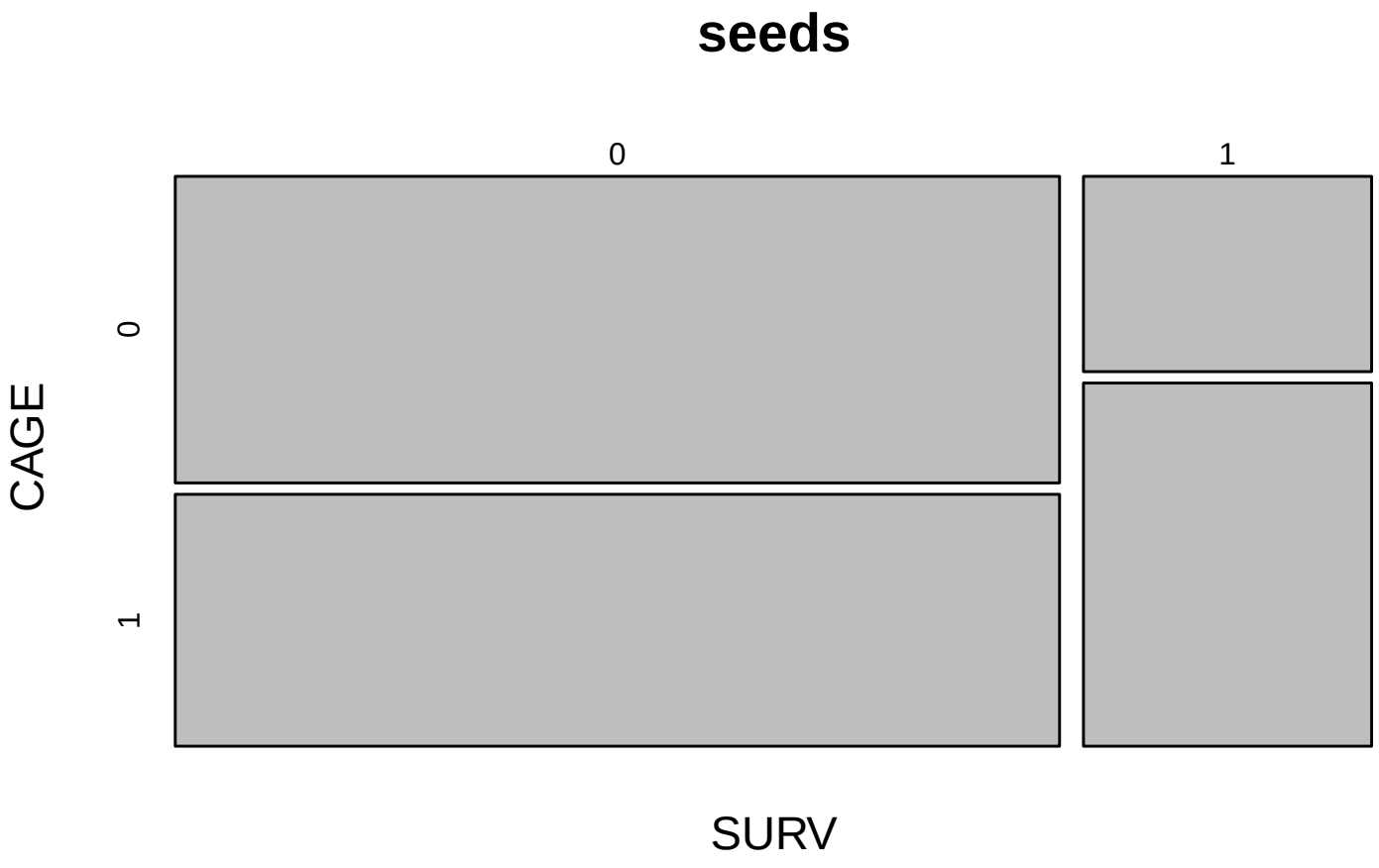
- Naive approach: Regress SURV on CAGE and LIGHT

$$\hat{\pi}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{CAGE}_i + \hat{\beta}_2 \mathrm{LIGHT}_i$$

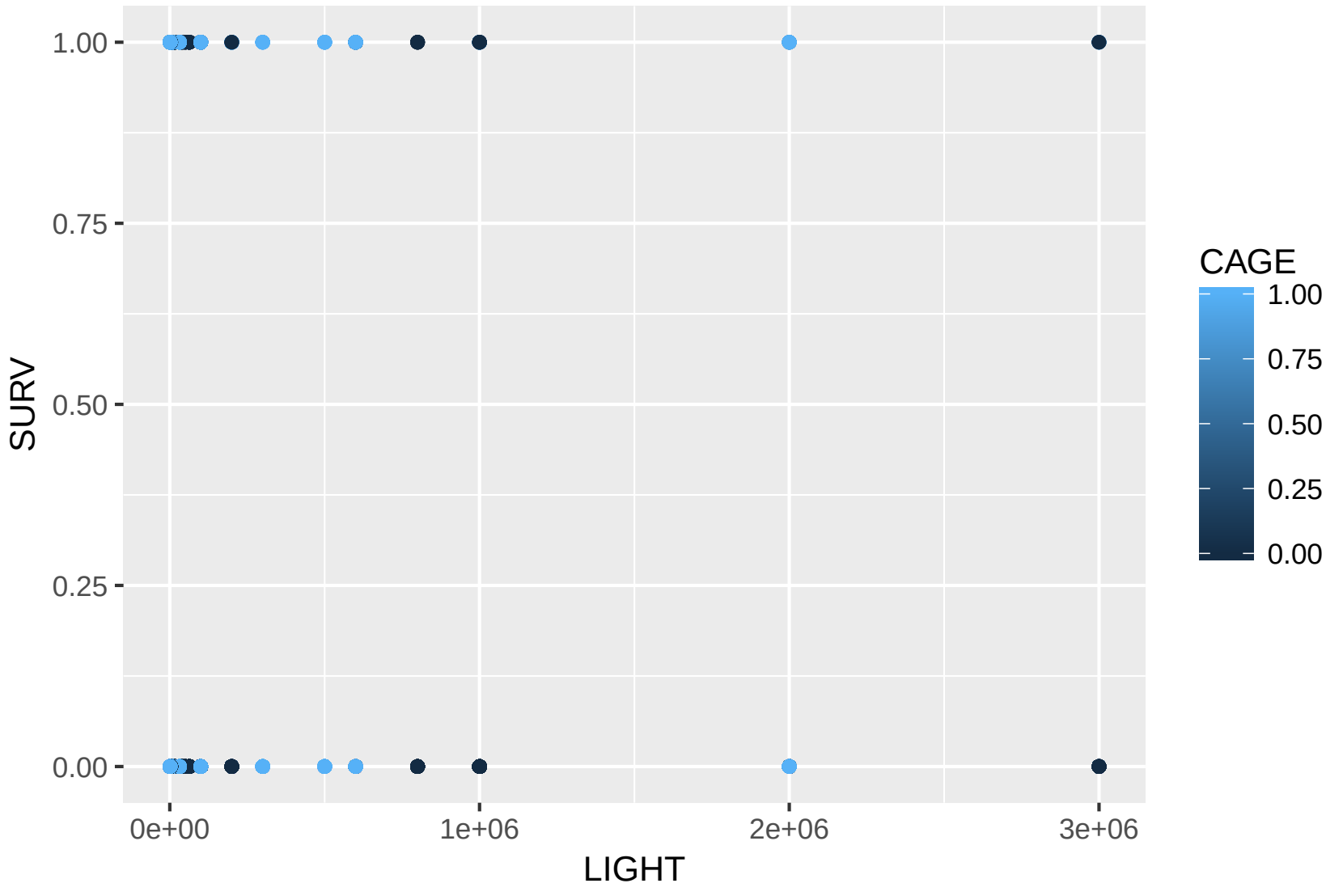
- Problems:
  - Fitted values of probabilities are not constrained to (0, 1)
  - Variances are not constant $\pi_i(1 - \pi_i)$ under Bernoulli model
- Unbiased?

plot of SURV & CAGE

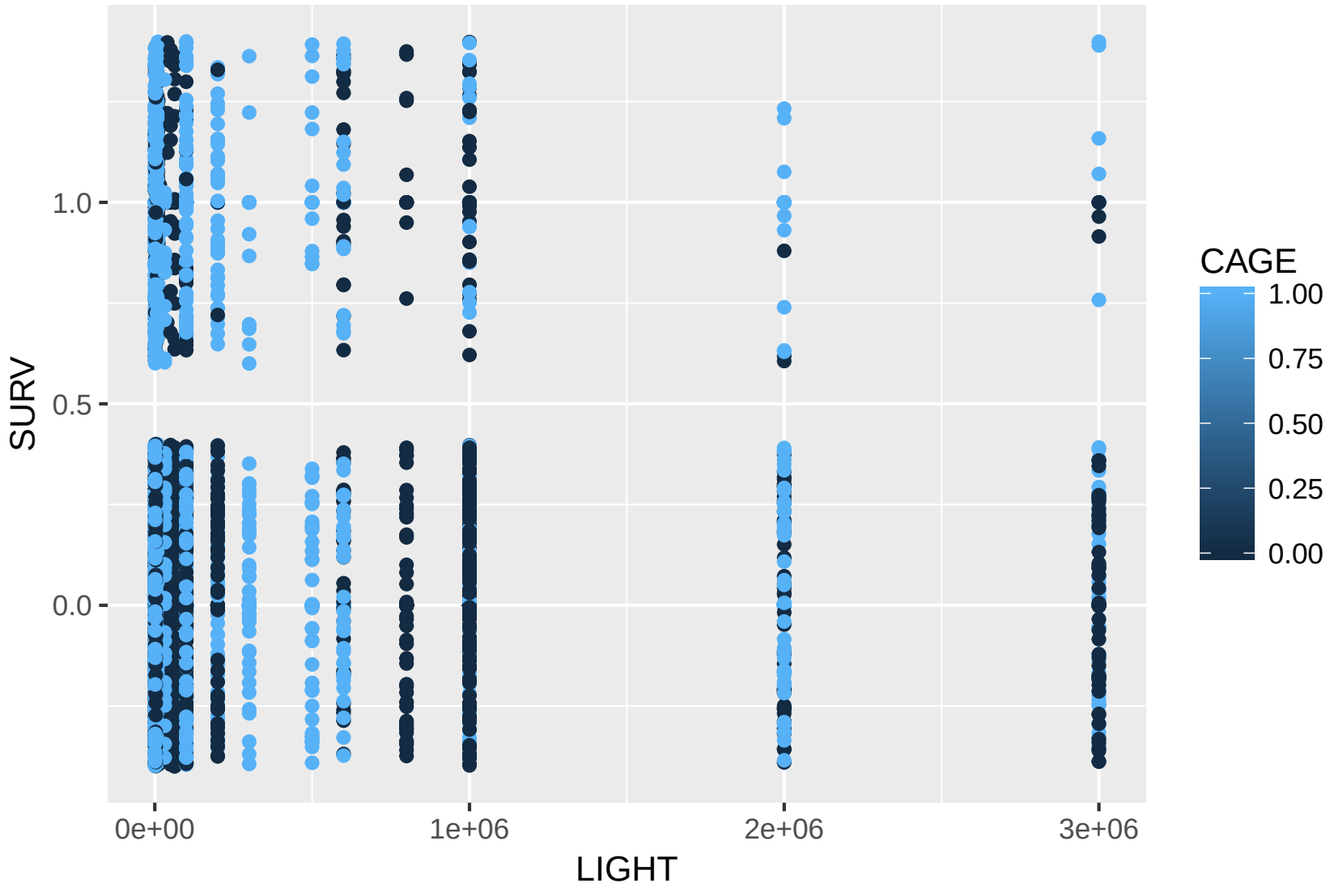
```
seeds = read.table("seeds.txt", header=TRUE)
mosaicplot(SURV ~ CAGE, data=seeds)
```



plot of SURV versus LIGHT and CAGE



plot of SURV versus LIGHT and CAGE jittered



Logistic Regression

To build in the necessary constraints that the probabilities are between 0 and 1 convert to log-odds or "logits"

- Odds of survival: $\pi_i/(1-\pi_1)$

$$\text{logit}(\pi_i) \stackrel{\text{def}}{=} \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{CAGE}_i + \beta_2 \text{LIGHT}_i = \eta_i$$

- η_i is the linear predictor
- logit is the *link* function that relates the mean π_i to the linear predictor η_i
- Generalized Linear Models (GLMs)
- Find Maximum Likelihood Estimates (optimization problem)

Logits

To convert from the linear predictor η to the mean π, use the inverse transformation:

- log odds (SURV$= 1) = \eta$
- odds (SURV $= 1) =$ exp$(\eta) = \omega$
- $\pi =$ odds$/(1 +$ odds$) = \omega/(1 + \omega)$
- $\omega = \pi/(1 - \pi)$

Can go in either direction

Interpretation of Coefficients

$$\omega_i = \exp(\beta_0 + \beta_1 \text{CAGE}_i + \beta_2 \text{LIGHT}_i)$$

- When all explanatory variables are 0 (CAGE= 0, LIGHT= 0), the odds of survival are $\exp(\beta_0)$
- The ratio of odds (or odds ratio) at $X_j = A$ to odds at $X_j = B$, for fixed values of the other explanatory variables is

$$\text{Odds ratio} = \frac{\omega_A}{\omega_b} = \exp(\beta_j(A - B))$$

$$\text{Odds ratio} = \frac{\omega_A}{\omega_b} = \exp(\beta_j) \text{ if } A - B = 1$$

$$\text{Odds}(X_j = A) = \exp(\beta_j) \cdot \text{Odds}(X_j = B)$$

- Coefficients are log odds ratios

R Code

- use `glm()` rather than `lm()`
- model formula as before
- need to specify family (and link if not default)

```
seeds = read.table("seeds.txt", header=TRUE)
seeds.glm0 = glm(SURV ~ 1, data=seeds, family=binomial)
seeds.glm1 = glm(SURV ~  CAGE + LIGHT,
                 family = binomial,
                 data=seeds)
```

Estimates

```
library(xtable)
xtable(summary(seeds.glm1)$coef,
       digits=c(0, 4, 4, 1,-2))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.4373	0.0709	-20.3	3.02E-91
CAGE	0.7858	0.0875	9.0	2.79E-19
LIGHT	-0.0000	0.0000	-5.3	9.09E-08

- Coefficient for the dummy variable CAGE$=0.79$
- If CAGE increases by 1 unit (No CAGE to CAGE) the odds of survival change by $\exp(.79) = 2.2$
- The odds of survival in a CAGE are 2.2 times higher than odds of survival in the open.

Confidence Intervals

- MLEs are approximately normally distributed (large samples)
  - mean β_j
  - estimated variance $\text{SE}(\beta_j)^2$
- Asymptotic posterior distribution for β_j is $N(\hat{\beta}_j, \text{SE}(\beta_j)^2)$
- $(1-\alpha)100\%$ CI based on normal theory:

$$\hat{\beta}_j \pm Z_{\alpha/2}\text{SE}(\beta_j)$$

- 95% CI for coefficient for CAGE:

$$0.7858 \pm 1.96 * 0.7858 = (0.62, 0.96)$$

- Exponentiate to obtain interval for odds ratio:
  $\exp(0.62), \exp(0.96) = (1.85, 2.607)$

The odds of survival in a CAGE are 1.85 to 2.607 times higher than odds of survival in the open (with confidence 0.95).

Deviance

The concept of Deviance replaces Sum-of-Squares in GLMs

- residual deviance = -2 log likelihood at MLEs

$$-2\sum_i y_i \log(\hat{\pi}_i) + (1-y_i)\log(1-\hat{\pi}_i)$$

$$\log(\hat{\pi}_i/(1-\hat{\pi}_i)) = \hat{\beta}_0 + \text{CAGE}\ \hat{\beta}_1 + \text{LIGHT}\ \hat{\beta}_2$$

- null deviance = residual deviance under model with constant mean (Total Sum of Squares in Gaussian)
- analysis of deviance
- change in (residual) deviance has an asymptotic χ^2 distribution with degrees of freedom based on the change in number of parameters

Analysis of Deviance Table

```
anova(seeds.glm0, seeds.glm1, test="Chi")

## Analysis of Deviance Table
##
## Model 1: SURV ~ 1
## Model 2: SURV ~ CAGE + LIGHT
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3071     3426.1
## 2      3069     3299.0  2   127.09 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

OverDispersion/Lack of Fit?

Lack of Fit

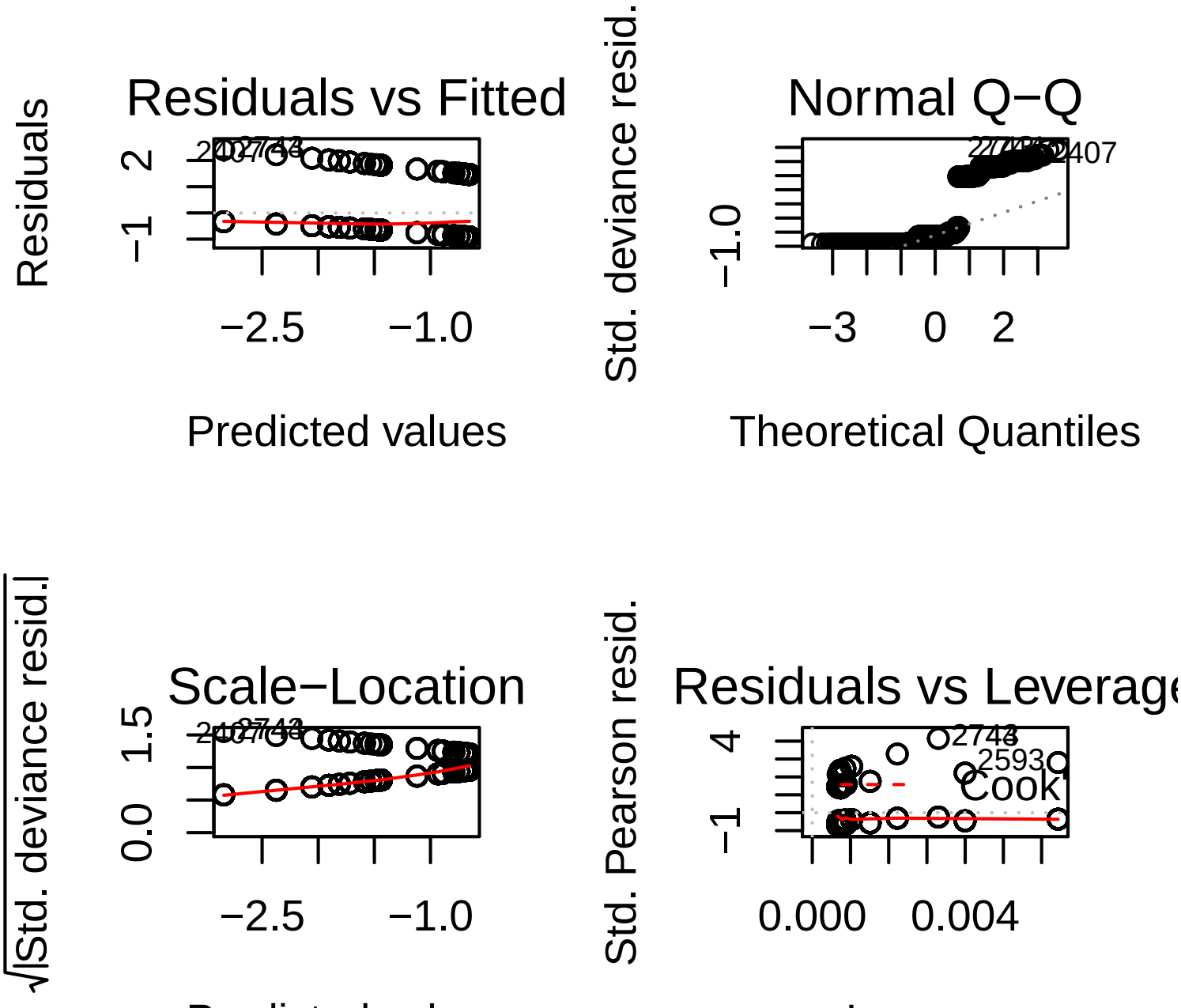
- Lack of fit if residual deviance larger than expected
- no variance needed to compare
- Residual Deviance has a χ^2 with $n - p$ df
- p-value $= P(\chi^2_{n-p} >$ observed deviance)

```
pchisq(seeds.glm1$deviance, seeds.glm1$df.residual,
       lower=FALSE)

## [1] 0.002024634
```

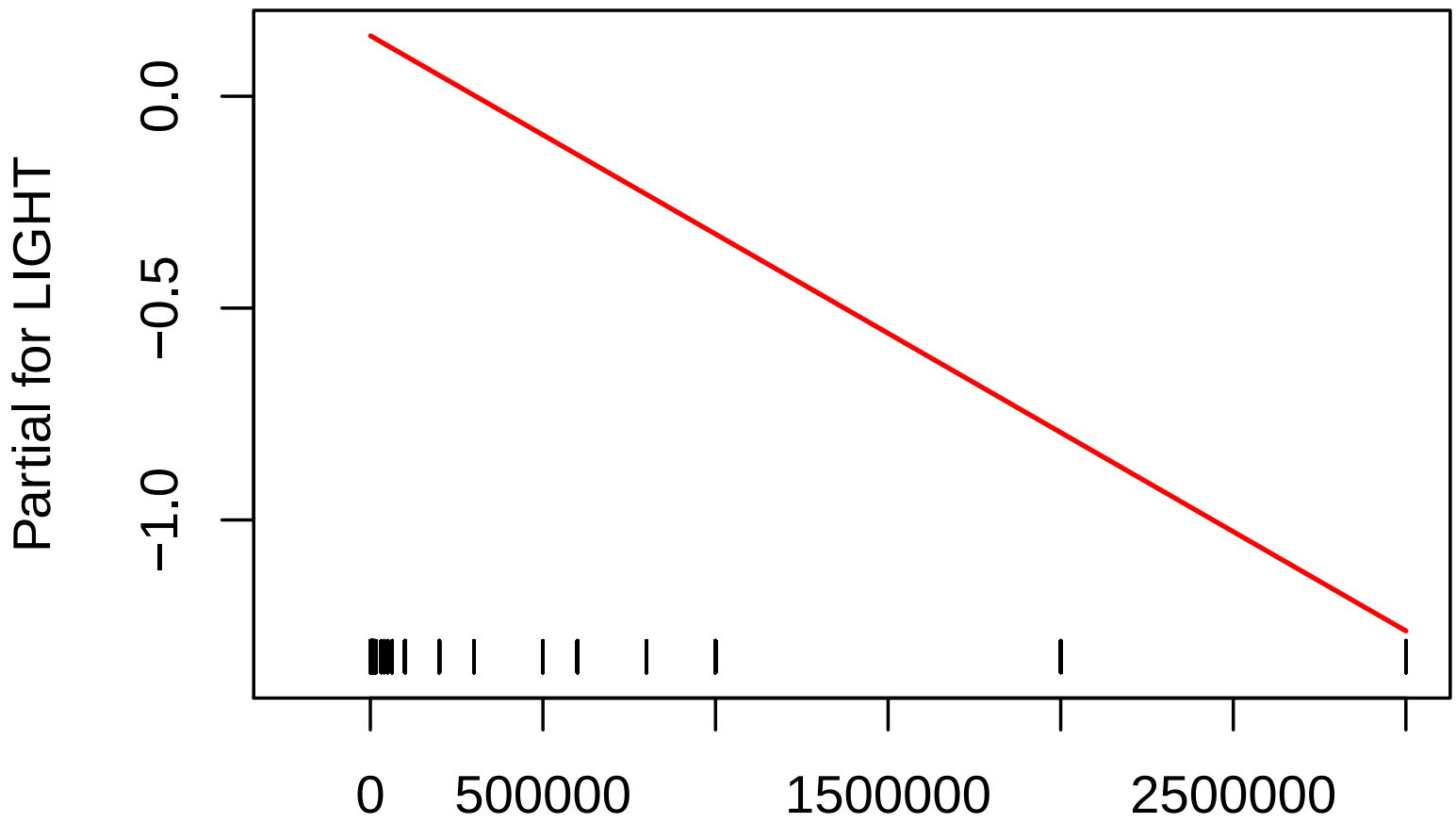
Surprising result if model were true.

Diagnostic Plots: plot(seeds.glm1)



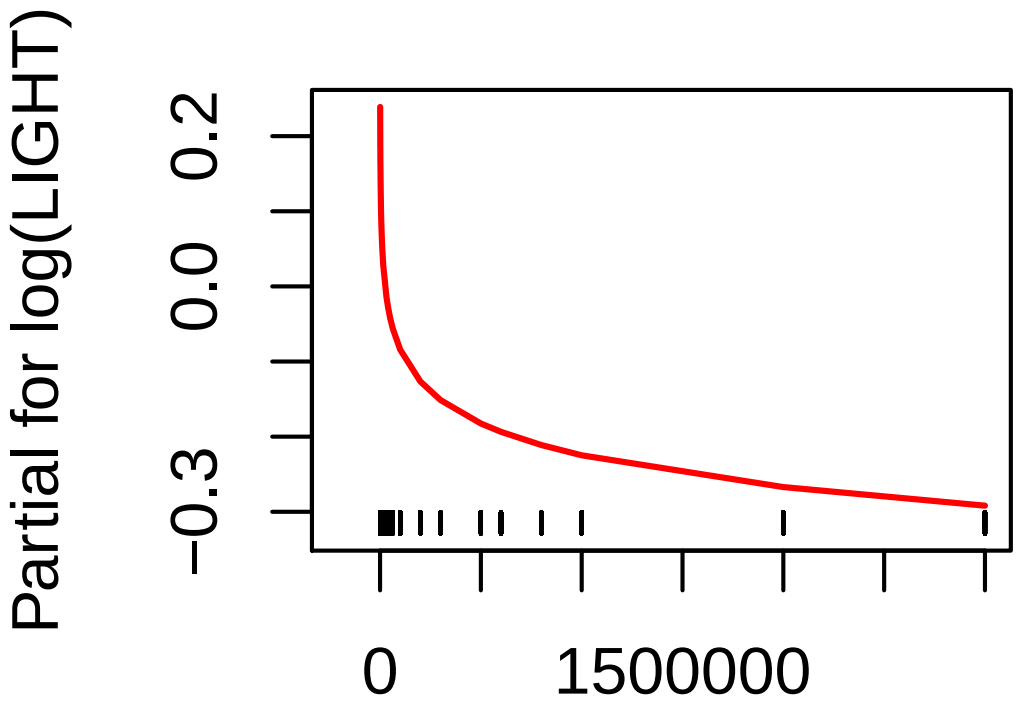
termplot

```
termplot(seeds.glm1, term='LIGHT', rug=T)
```



log(LIGHT)

```
seeds.glm2 = glm(SURV ~  CAGE + log(LIGHT),
                 family = binomial,
                 data=seeds)
termplot(seeds.glm2, term="log(LIGHT)", rug=T)
```



Other Variables

```
seeds.glm3 = glm(SURV ~  SPECIES + CAGE + log(LIGHT) +
                   factor(LITTER),  data=seeds, family=binomial)
xtable(summary(seeds.glm3)$coef)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.83	0.25	-3.29	0.00
SPECIESC. biflora	0.19	0.17	1.10	0.27
SPECIESC. racemosa	0.85	0.16	5.19	0.00
SPECIESGouania	-2.64	0.36	-7.32	0.00
SPECIESHirtella	1.14	0.16	7.03	0.00
SPECIESInga	0.90	0.16	5.55	0.00
SPECIESMaclura	-2.64	0.36	-7.32	0.00
SPECIESStrychnos	-1.19	0.22	-5.46	0.00
CAGE	0.93	0.10	9.59	0.00
log(LIGHT)	-0.09	0.02	-4.50	0.00
factor(LITTER)1	0.09	0.13	0.67	0.50
factor(LITTER)2	0.24	0.13	1.81	0.07
factor(LITTER)4	-0.14	0.14	-1.01	0.31

Analysis of Deviance

```
anova(seeds.glm3, test="Chi")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: SURV
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi
## NULL                             3071     3426.1
## SPECIES           7   572.20      3064     2853.9 < 2.2e-1
## CAGE              1   109.29      3063     2744.6 < 2.2e-1
## log(LIGHT)        1    15.30      3062     2729.3 9.191e-0
## factor(LITTER)    3     7.92      3059     2721.4   0.0476
## ---
```

Interactions?

The presence of a CAGE may be more important for survival for some species than others - implies an interation
The odds of survival | Cage compared to odds of survival | no Cage depend on SPECIES
Fit model with upto 4 way interactions:

```
seeds.glm4 = glm(SURV~SPECIES*CAGE*log(LIGHT)*LITTER,
                 data=seeds, family=binomial)
```

The analysis of deviance test suggests that there are three way interactions

Hierarchical Model

So far we have not taken into account all the sources of variation or information about the experimental design.

- SPECIES (size & cotyledon type) and LITTER are randomized to sub-plots. Expect that survival of seedlings in the same sub-plot may be related, which suggests a sub-plot random effect.
- sub-plots are nested within CAGE within plots (so expect that sub-plots in the same CAGE are correlated, as well as sub-plots within the same plot may have a similar survival.
- Plot characteristics may affect survival (light levels)

How to model?