

# Lecture 4.

## Logistic Regression Example

Lecturer: Tetiana Marynych, [marynychtetyana@gmail.com](mailto:marynychtetyana@gmail.com)

## Binary logistic regression: Seedling Survival example

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

## Modeling Survival

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

$$E[SURVi \mid X_i] = p_i$$

- 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.

## Modeling Survival

▶ Naive approach: Regress SURV on CAGE and LIGHT

$$\hat{p}_i = \hat{\beta}_0 + \hat{\beta}_1 CAGE_i + \hat{\beta}_2 LIGHT_i$$

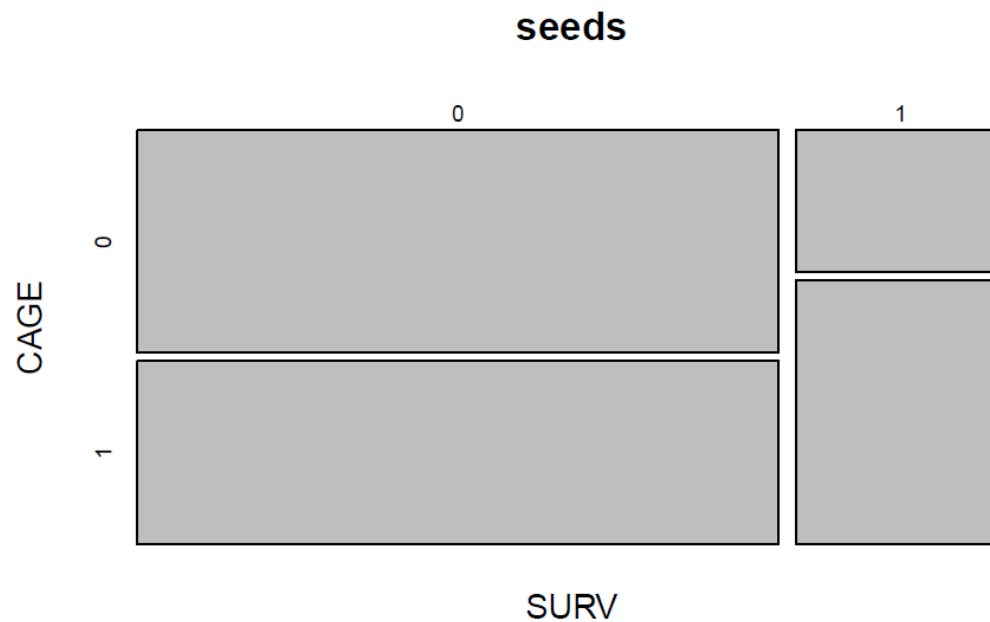
▶ Problems:

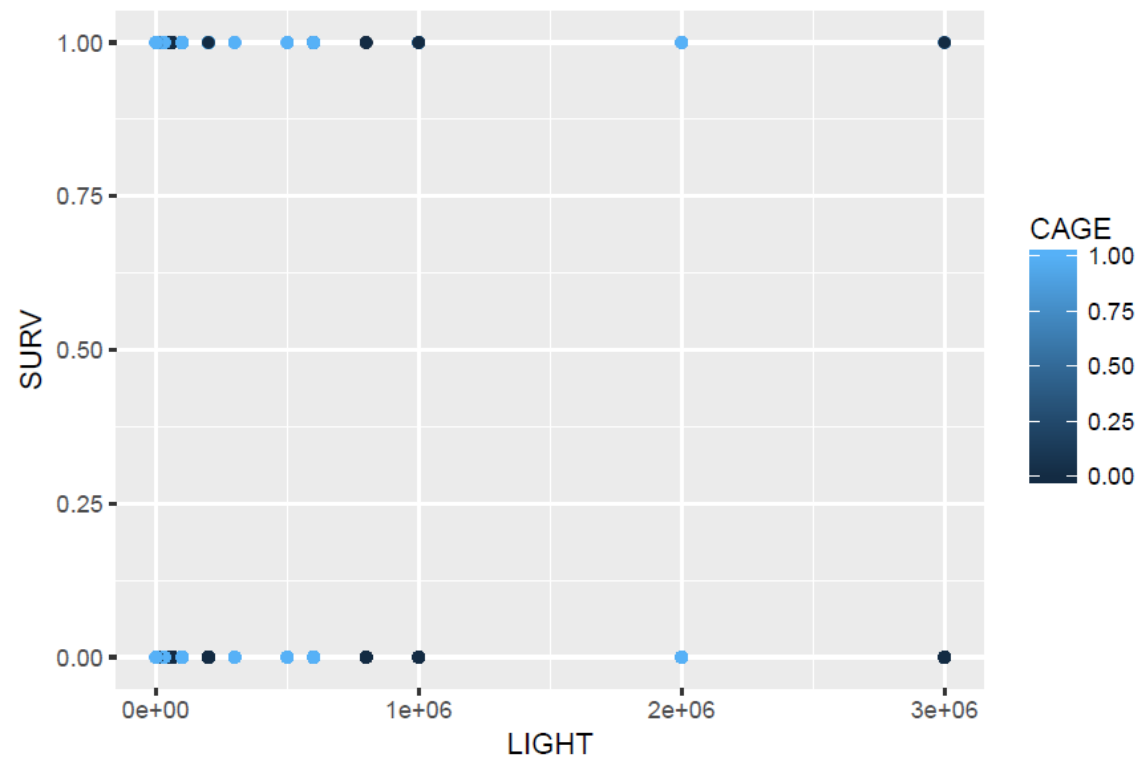
- Fitted values of probabilities are not constrained to (0, 1)
- Variances are not constant  $p_i(1 - p_i)$  under Bernoulli model

## Modeling Survival

plot of SURV & CAGE

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





## 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:  $\frac{p_i}{1-p_i}$

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{CAGE}_i + \beta_2 \text{LIGHT}_i = X\beta = \eta_i$$

- $\eta_i$  is the linear predictor
- logit is the link function that relates the mean probability  $p_i$  to the linear predictor  $\eta_i$
- Generalized Linear Models, GLMs
- Find Maximum Likelihood Estimates (optimization problem)

## Logits

To convert from the linear predictor  $\eta_i$  to the mean  $p_i$  use the inverse transformation:

$$\log Odds(SURV = 1) = X\beta = \eta$$

$$Odds(SURV = 1) = \exp(X\beta) = \exp(\eta) = \omega$$

$$p_i = \frac{Odds}{1 + Odds} = \frac{\omega}{1 + \omega}$$

$$\omega = \frac{p_i}{1 - p_i}$$

Can go in either direction



## Interpretation of the Coefficients

$$\omega_i = \exp(\beta_0 + \beta_1 CAGE_i + \beta_2 LIGHT_i)$$

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

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

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

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

- 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)
```

## R Summary Interpretation

```
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  $\exp(0.79) = 2.2$
- The odds of survival in a CAGE are 2.2 times higher than odds of survival in the open.

## Confidence Intervals, CI

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

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

- ✓ 95% CI for coefficient CAGE:

$$0.7858 \pm 1.96 * 0.0875 = (0.62; 0.96)$$

- ✓ 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).

## Significance of the Coefficients

- ✓ Check hypothesis  $H_0: \beta_j = 0$

$$Z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \text{ or } p\text{-value: } \Pr(>|Z|)$$

- ✓ Check overall model adequacy:

- Check hypothesis  $H_0: \beta_0 = \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$
- Likelihood Ratio Statistics or Deviance:

$$\Lambda^* = -2(l(\hat{\beta}^0) - l(\hat{\beta}))$$

where  $l(\hat{\beta})$  - log likelihood ratio of the full model;

$l(\hat{\beta}^0)$  - log likelihood ratio of the null reduced model (without explanatory variables, only with the intercept).

## Deviance

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

▣ residual deviance =  $-2 \log$  likelihood at MLEs

$$-2 \sum_{i=1}^n [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$$

$$\log(\hat{p}_i / (1 - \hat{p}_i)) = \hat{\beta}_0 + \hat{\beta}_1 CAGE_i + \hat{\beta}_2 LIGHT_i$$

▣ null deviance = residual deviance under model with constant mean (Total Sum of Squares in Gaussian)

▣ analysis of deviance:  $\sigma^2 = deviance(null) - deviance(full)$

▣ change in (residual) deviance has an asymptotic  $\chi^2$  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
```

## Lack of Fit

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

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

```
## [1] 0.002024634
```

Surprising result if model were true.



## Goodness of Fit

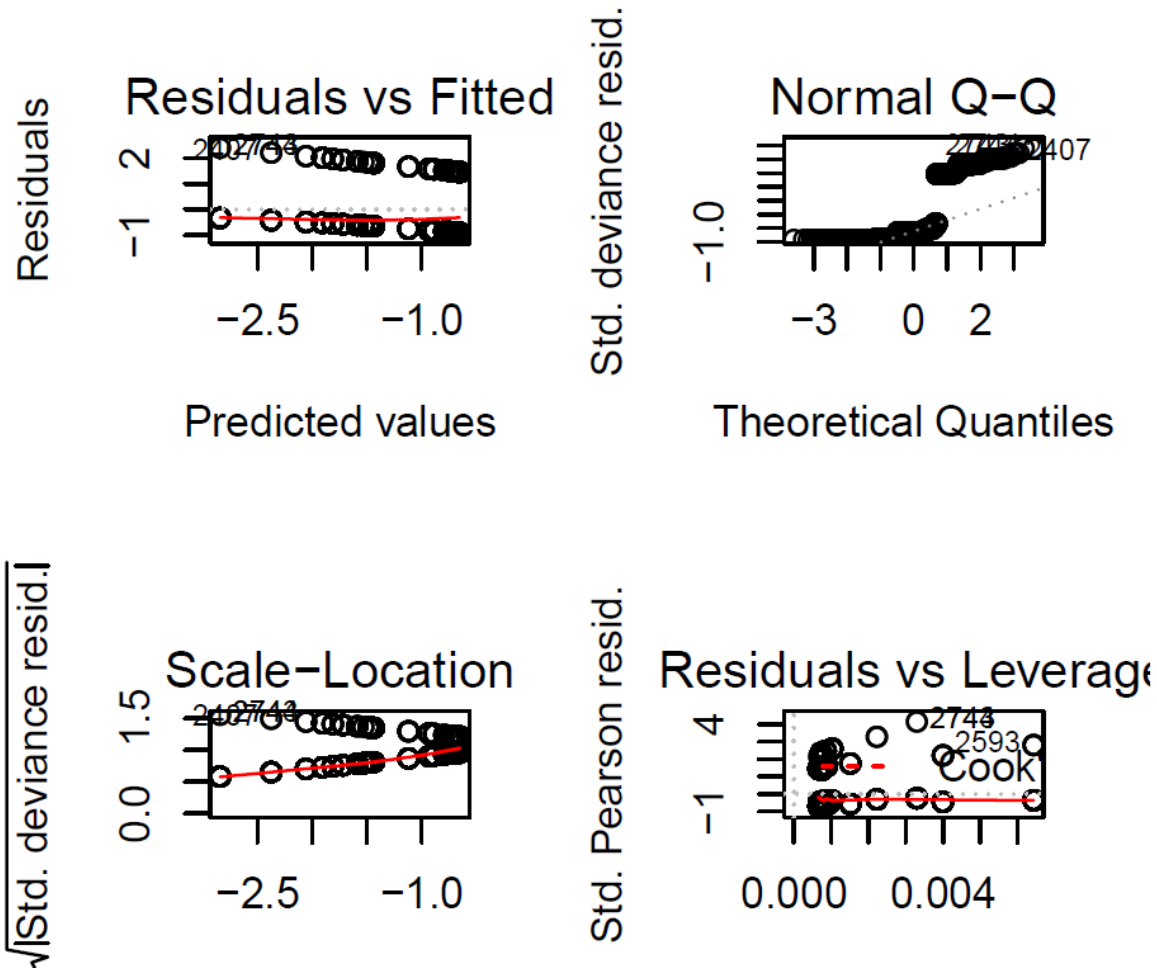
### 1. Coefficient of Determination for GLM:

$$\text{Log likelihoods - } R^2 = 1 - \frac{l(\hat{\beta})}{l(\hat{\beta}^0)}$$

$$\text{Deviances - } R^2 = 1 - \frac{\text{Resid.Dev.Full}}{\text{Resid.Dev.Null}}$$

### 2. Diagnostics with residuals plots: `plot(seeds.glm1)`

## Diagnostic Plots: `plot(seeds.glm1)`



## 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(>Ch
## NULL			3071	3426.1	
## SPECIES	7	572.20	3064	2853.9	< 2.2e-
## CAGE	1	109.29	3063	2744.6	< 2.2e-
## log(LIGHT)	1	15.30	3062	2729.3	9.191e-
## factor(LITTER)	3	7.92	3059	2721.4	0.047

```
...
```

## Interactions?

- The presence of a CAGE may be more important for survival for some species than others - implies an interaction
- The odds of survival / Cage compared to odds of survival / no Cage depend on SPECIES
- Fit model with up to 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

---

## R Code 2: Analysis of Deviance table

```
# Estimate the null (reduced) model and full model
```

```
> fit.glm0 = glm(Y ~ 1, data=quality, family="binomial")
```

```
> fit.glm = glm(Y ~ X1 + X2, data=quality, family="binomial")
```

```
# Lack of fit if residual deviance larger than expected
```

```
> anova(fit.glm0, fit.glm, test="Chi") # Deviance analysis for 2 models
```

```
> anova(fit.glm) # Deviance analysis for one model (by default the null model is  
the intercept-only model)
```

```
# Residual deviance p-value
```

```
> pchisq(fit.glm$deviance, fit.glm$df.residual, lower=FALSE)
```

```
# Diagnostics residual plot
```

```
> plot(fit.glm)
```