

Discrete Response Model

Lecture 4

datascience@berkeley

Estimation and Inference

Estimation

- Parameters are estimated using maximum likelihood estimation.
- For a sample of size m , the likelihood function is simply the product of m **multinomial distributions with probability parameters π_j** , given the i.i.d. assumption.
- Iterative numerical procedures are used then to find the parameter estimates.
- The **polr()** function from the **MASS** package is used.


It is especially important to have the levels of the **categorical response ordered** in the desired way when using `polr()`; otherwise, the ordering of the levels of Y will not be correctly taken into account.

The covariance matrix for the parameter estimates follows from using standard likelihood procedures outlined in the appendix of the text.

Inference

Suppose there is one explanatory variable ($p = 1$), and the hypotheses of interest are

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$


- If the null hypothesis is not rejected, this says that the log-odds comparing $P(Y \leq j)$ to $P(Y > j)$ do not depend on this specific explanatory variable. In the case of one categorical explanatory variable X , this is equivalent to independence between X and Y .
- If the null hypothesis is rejected, the ordering of the log-odds comparing $P(Y \leq j)$ to $P(Y > j)$ holds; i.e., the log-odds progressively grow larger or smaller depending on the sign of β_1 .

Example

As an example, imagine that **the presence of disease** makes sprouted kernels more desirable than scab kernels.

scab (Y = 1) < sprout (Y = 2) < healthy (Y = 3)

```
> levels(wheat$type)
[1] "Healthy" "Scab"      "Sprout"
> wheat$type.order<-factor(wheat$type, levels = c("Scab",
  "Sprout", "Healthy"))
> #head(wheat) #excluded to save space
> levels(wheat$type.order)
[1] "Scab"      "Sprout"    "Healthy"
```

We order it ourselves

Estimate the following model:

$$\text{logit}[P(Y \leq j)] = \beta_{j0} + \beta_1 X_1 + \dots + \beta_6 X_6$$

for $j = 1, 2$

Example

```
library(package = MASS)
mod.fit.ord<-polr(formula = type.order ~ class + density
  + hardness + size + weight + moisture, data = wheat,
  method = "logistic")
summary(mod.fit.ord)
```

Re-fitting to get Hessian

Call: `polr(formula = type.order ~ class + density + hardness +`
`|size + weight + moisture, data = wheat, method = "logistic")`

Coefficients:

	Value	Std. Error	t value
<u>classsrw</u>	0.17370	0.391764	0.4434
<u>density</u>	13.50534	1.713009	7.8840
<u>hardness</u>	0.01039	0.005932	1.7522
size	-0.29253	0.413095	-0.7081
weight	0.12721	0.029996	4.2411
moisture	-0.03902	0.088396	-0.4414

Intercepts:

	Value	Std. Error	t value
<u>Scab Sprout</u>	17.5724	2.2460	7.8237
<u>Sprout Healthy</u>	20.0444	2.3395	8.5677

Residual Deviance: 422.4178

AIC: 438.4178

Example

Because the actual model estimated by polr() is

$$\text{logit}(P(Y \leq j)) = \beta_{j0} - \eta_1 x_1 - \dots - \eta_p x_p$$

where $-\eta_r$ is β_r in our notation. Thus, we will always need to change the sign of the estimated parameter given by polr(). The estimated model is:

$$\begin{aligned} \text{logit}(\hat{P}(Y \leq j)) = & \hat{\beta}_{j0} - 0.17\text{SRW} - 13.51\text{density} \\ & - 0.01\text{hardness} + 0.29\text{size} \\ & - 0.13\text{weight} + 0.04\text{moisture} \end{aligned}$$

where $\hat{\beta}_{10} = 17.57$ and $\hat{\beta}_{20} = 20.04$.

Example

The "t value" column in the coefficients table provides the Wald statistic for testing

$$H_0: \beta_r = 0 \text{ vs. } H_a: \beta_r \neq 0$$

for $r = 1, \dots, 6$, and the Anova() function provides the corresponding LRTs:

```
> library(package = car) #If not done already
```

```
> Anova(mod.fit.ord)
```

```
Analysis of Deviance Table (Type II tests)
```

```
Response: type.order
```

	<u>LR</u>	<u>Chisq</u>	<u>Df</u>	<u>Pr(>Chisq)</u>
class	0.197	1	0.65749	
density	98.437	1	< 2.2e-16 ***	
hardness	3.084	1	0.07908 .	
size	0.499	1	0.47982	
weight	18.965	1	1.332e-05 ***	
moisture	0.195	1	0.65872	

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Example

```
> pi.hat.ord<-predict(object = mod.fit.ord, type = "probs")
> head(pi.hat.ord)
```

	Scab	Sprout	Healthy
1	0.03661601	0.2738502	0.6895338
2	0.03351672	0.2576769	0.7088064
3	0.08379891	0.4362428	0.4799583
4	0.01694278	0.1526100	0.8304472
5	0.11408176	0.4899557	0.3959626
6	0.02874814	0.2308637	0.7403882

```
> head(predict(object = mod.fit.ord, type = "class"))
[1] Healthy Healthy Healthy Healthy Sprout Healthy
Levels: Scab Sprout Healthy
```

The estimated probability of being healthy for the first observation is

$$\hat{\pi}_{\text{healthy}} = 1 - \frac{e^{20.04 - 0.17 \times 0 + 0.04 \times 12.02}}{1 + e^{20.04 - 0.17 \times 0 + 0.04 \times 12.02}} = 0.6895$$

Example

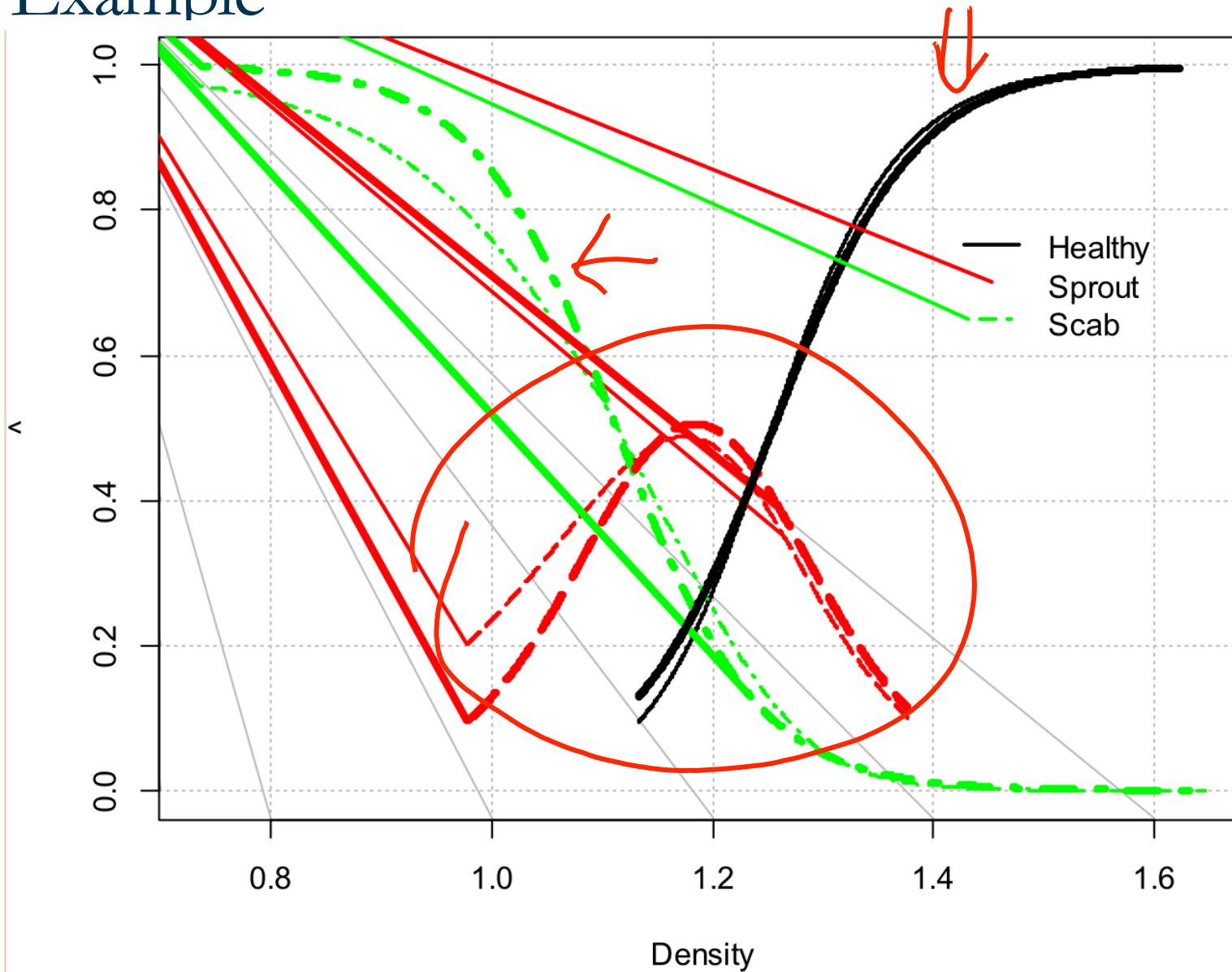
When there is only one explanatory variable in the model, we can easily examine the estimated probabilities through a plot.

The model using only density is

$$\text{logit}(\hat{P}(Y \leq j)) = \hat{\beta}_{j0} - 15.64\text{density}$$

$$\hat{\beta}_{10} = 17.41 \text{ and } \hat{\beta}_{20} = 19.63.$$

Example



Berkeley

SCHOOL OF
INFORMATION