Discrete Response Model Lecture 4

datascience@berkeley

Example

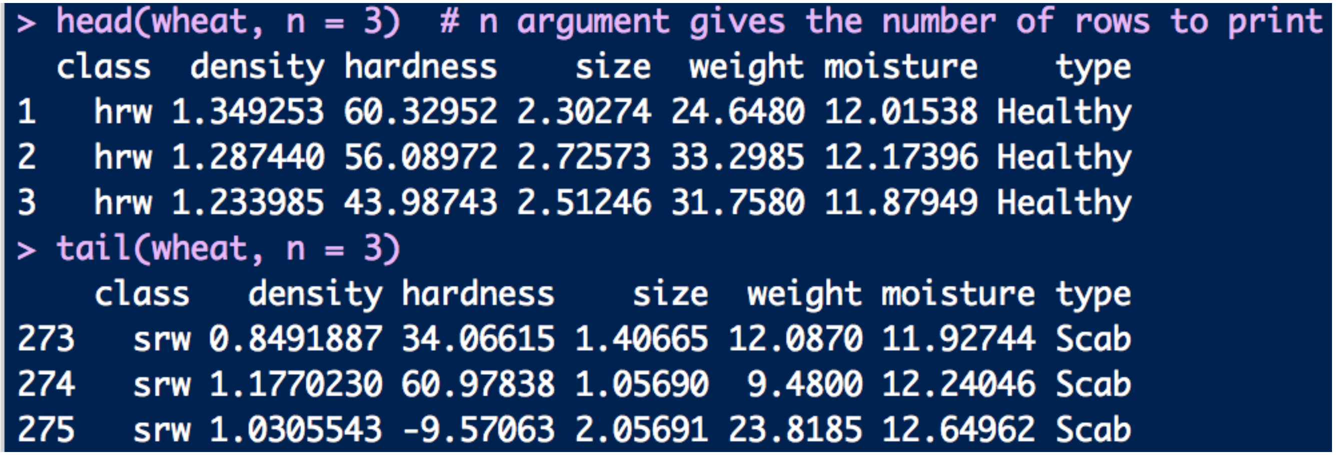
Example

- Wheat producers want to identify kernels that are in poor condition after being harvested.
- To facilitate this identification process, categorization systems have been developed to partition kernels into different categories (see Martin et al., 1998).
- For this example, we will look at the categories of "Healthy," "Sprout," or "Scab."
- In summary:
  - Healthy is the preferred condition because these kernels have not been damaged.
  - Sprout is less preferred than healthy because they have reduced weight and poorer flour quality.
  - Scab is less preferred than healthy because they come from plants that have been infected by a disease and have undesirable qualities in their appearance.

Example

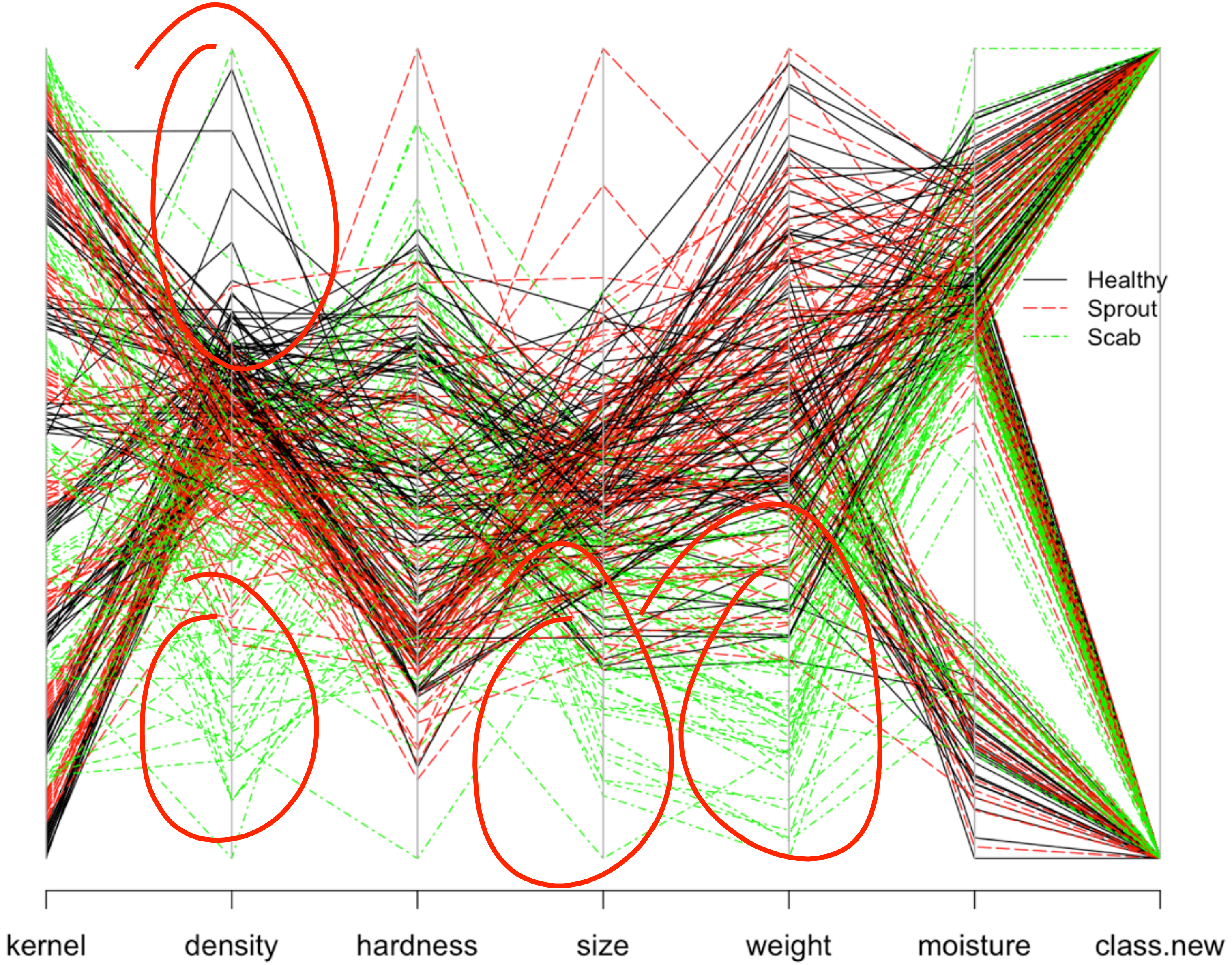
- Ideally, it would be preferred to make these categorizations for each kernel through using an automated process.
- To test a new system out, **275 wheat kernels** were classified by human examination (assumed to be perfect). The automated system uses information about the class of the wheat kernel (soft red winter or hard red winter) and measurements for density, hardness, size, weight, and moisture for the kernel.

Example

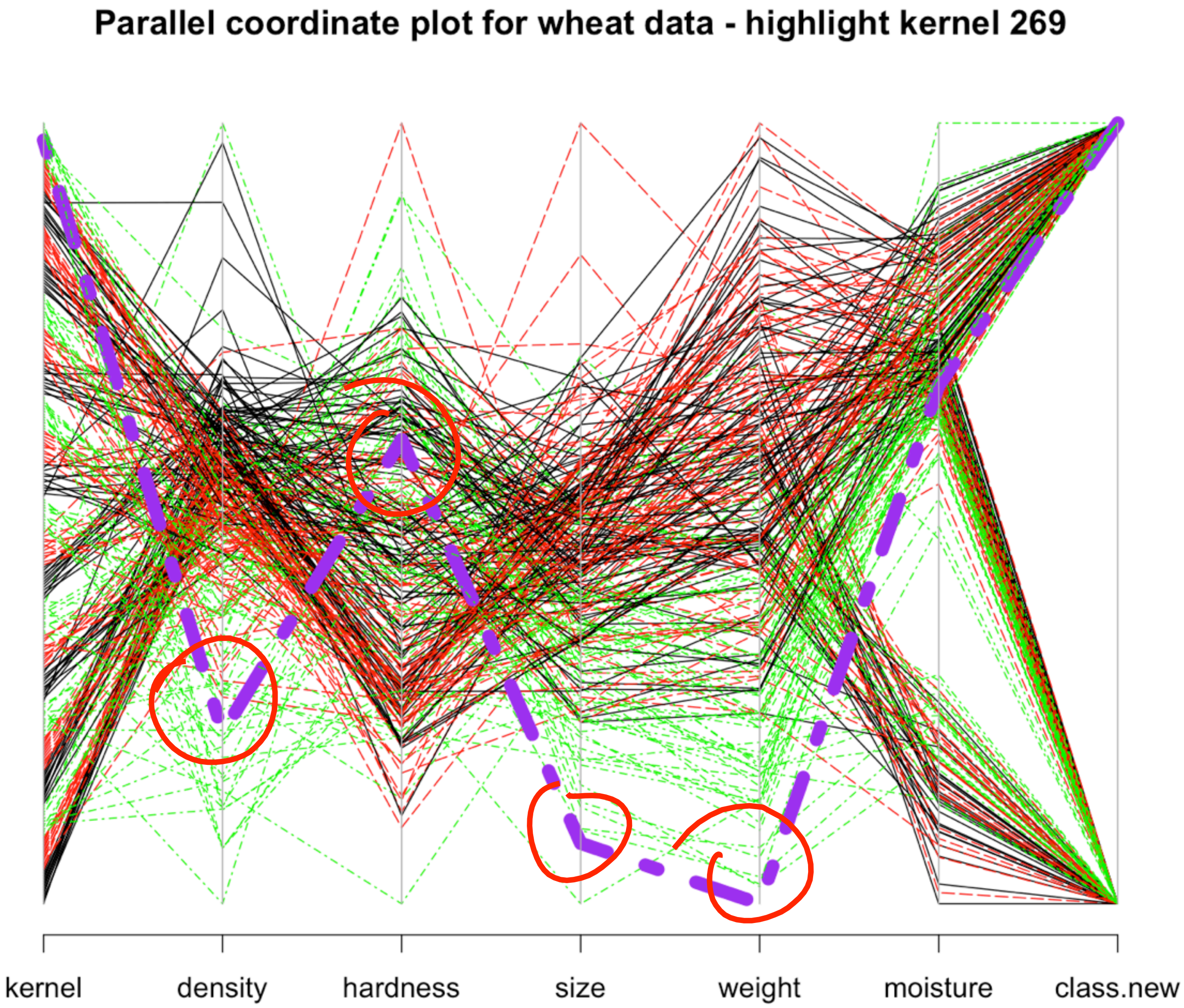


```
> head(wheat, n = 3)  # n argument gives the number of rows to print
  class  density hardness    size  weight moisture    type
1   hrw 1.349253 60.32952 2.30274 24.6480 12.01538 Healthy
2   hrw 1.287440 56.08972 2.72573 33.2985 12.17396 Healthy
3   hrw 1.233985 43.98743 2.51246 31.7580 11.87949 Healthy
> tail(wheat, n = 3)
    class   density hardness    size  weight moisture type
273   srw 0.8491887 34.06615 1.40665 12.0870 11.92744 Scab
274   srw 1.1770230 60.97838 1.05690  9.4800 12.24046 Scab
275   srw 1.0305543 -9.57063 2.05691 23.8185 12.64962 Scab
```

Parallel Coordinate Plot



Parallel Coordinate Plot



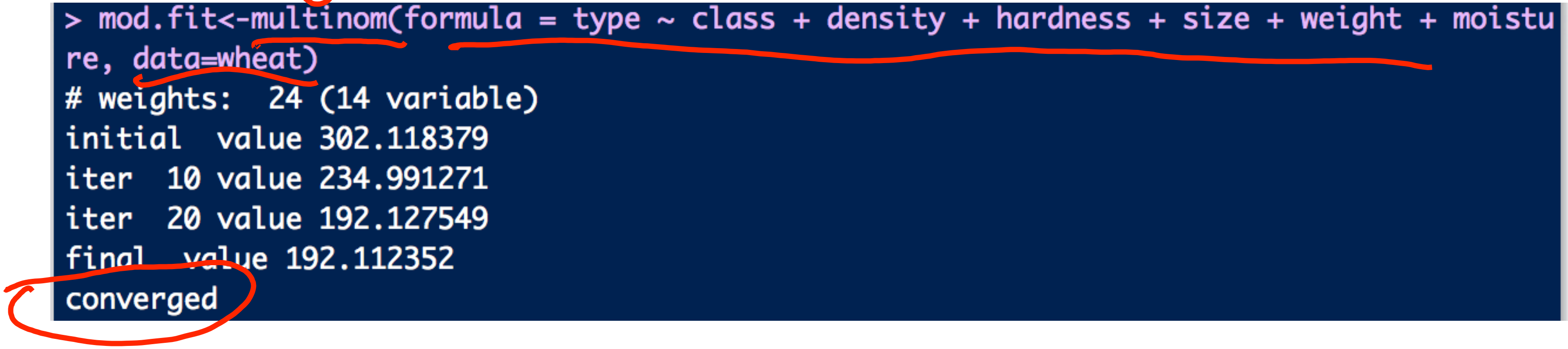
A Multinomial Logistic Regression Model

Consider the following model

$$\log(\pi_j/\pi_1) = \beta_{j0} + \beta_{j1}x_1 + \ldots + \beta_{j6}x_6 \text{ for } j = 2, 3$$

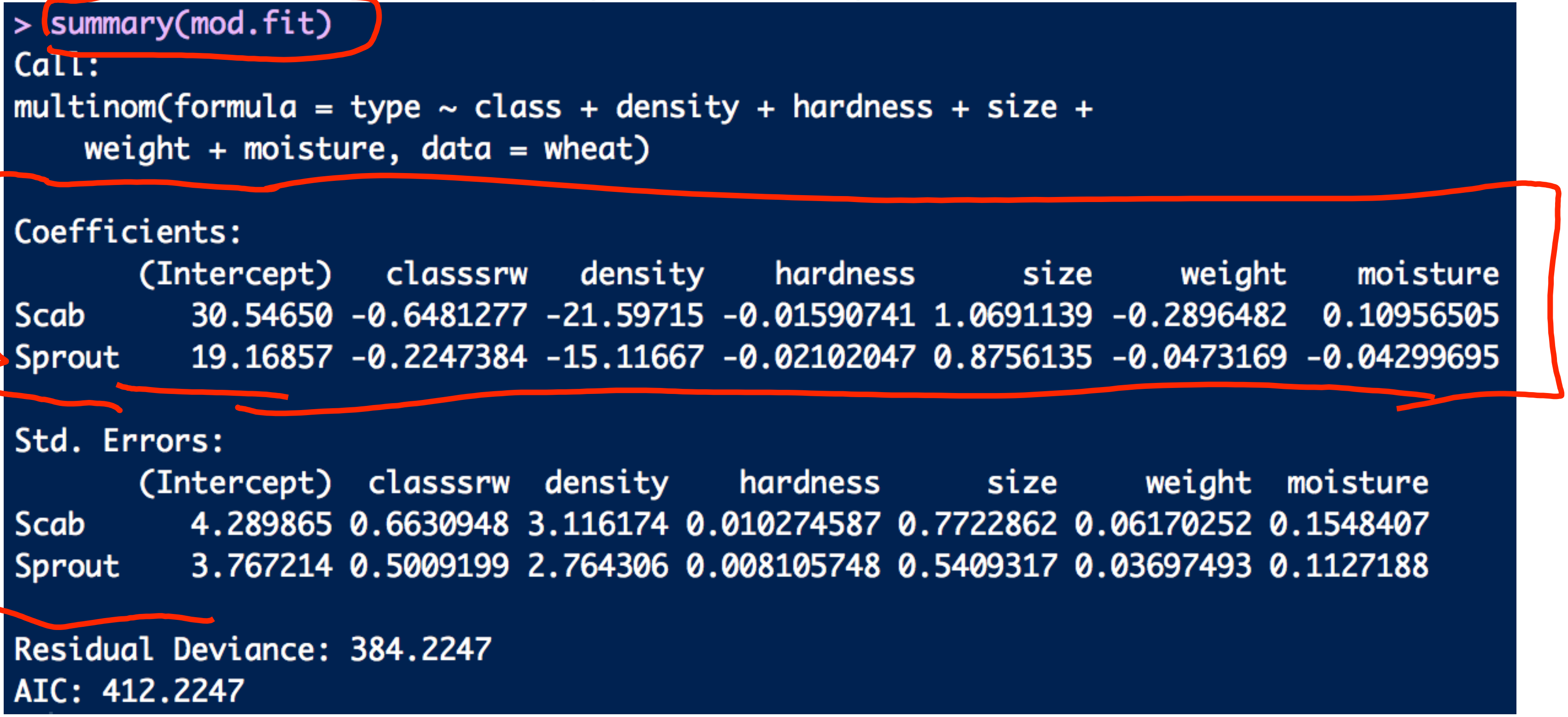
```
> levels(wheat$type)   #Shows the 3 categories
[1] "Healthy" "Scab"     "Sprout"
```

Thus, j = 1 is healthy, j = 2 is scab, and j = 3 is sprout.

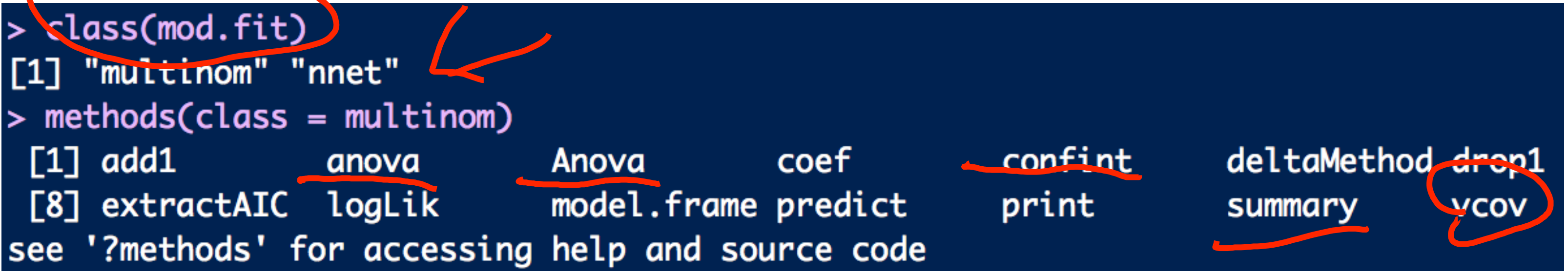


```
> mod.fit<-multinom(formula = type ~ class + density + hardness + size + weight + moistu
re, data=wheat)
# weights:  24 (14 variable)
initial  value 302.118379
iter  10 value 234.991271
iter  20 value 192.127549
final  value 192.112352
converged
```

A Multinomial Logistic Regression Model

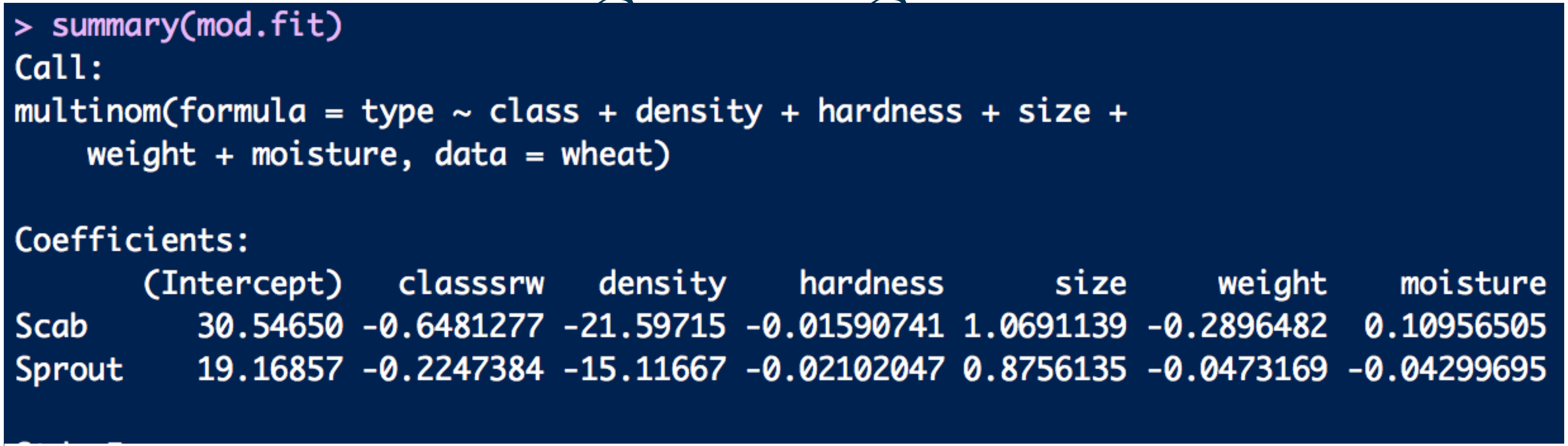


```
> summary(mod.fit)
Call:
multinom(formula = type ~ class + density + hardness + size +
    weight + moisture, data = wheat)

Coefficients:
        (Intercept)   classsrw   density    hardness      size     weight    moisture
Scab       30.54650 -0.6481277 -21.59715 -0.01590741 1.0691139 -0.2896482  0.10956505
Sprout     19.16857 -0.2247384 -15.11667 -0.02102047 0.8756135 -0.0473169 -0.04299695

Std. Errors:
        (Intercept)  classsrw  density    hardness      size     weight  moisture
Scab       4.289865 0.6630948 3.116174 0.010274587 0.7722862 0.06170252 0.1548407
Sprout     3.767214 0.5009199 2.764306 0.008105748 0.5409317 0.03697493 0.1127188

Residual Deviance: 384.2247
AIC: 412.2247
```



```
> class(mod.fit)
[1] "multinom" "nnet"
> methods(class = multinom)
 [1] add1        anova       Anova       coef        confint     deltaMethod drop1
 [8] extractAIC  logLik      model.frame predict     print       summary     vcov
see '?methods' for accessing help and source code
```

A Multinomial Logistic Regression Model



```
> summary(mod.fit)
Call:
multinom(formula = type ~ class + density + hardness + size +
    weight + moisture, data = wheat)

Coefficients:
       (Intercept)   classsrw   density    hardness      size     weight    moisture
Scab      30.54650 -0.6481277 -21.59715 -0.01590741 1.0691139 -0.2896482  0.10956505
Sprout    19.16857 -0.2247384 -15.11667 -0.02102047 0.8756135 -0.0473169 -0.04299695
```

$$\log(\underbrace{\hat{\pi}_{\text{scab}} / \hat{\pi}_{\text{healthy}}}) = 30.55 - 0.65\text{SRW} - 21.60\text{density}$$
$$-0.016\text{hardness} + 1.07\text{size} - 0.29\text{weight}$$
$$+0.11\text{moisture}$$

$$\log(\underbrace{\hat{\pi}_{\text{sprout}} / \hat{\pi}_{\text{healthy}}}) = 19.17 - 0.22\text{SRW} - 15.12\text{density}$$
$$-0.021\text{hardness} + 0.88\text{size} - 0.047\text{weight}$$
$$-0.043\text{moisture}$$

The Estimated Model

$$\log(\hat{\pi}_{scab} / \hat{\pi}_{healthy}) = 30.55 - 0.65SRW - 21.60density \\ -0.016hardness + 1.07size - 0.29weight \\ +0.11moisture$$

$$\log(\hat{\pi}_{sprout} / \hat{\pi}_{healthy}) = 19.17 - 0.22SRW - 15.12density \\ -0.021hardness + 0.88size - 0.047weight \\ -0.043moisture$$

- Notice how R forms an indicator variable for the class of the wheat ("classsrw" corresponds to SRW).
- Now that we have the estimated model, many of the basic types of analyses done in the last few lectures can be performed here!
- The R code used is very similar as well. Because of the similarity, we will assign those as take-home exercise.

Remarks

- The **mcprofile package** cannot be used for likelihood-ratio-based inference methods.
- Confidence intervals for π_j are more complicated to calculate than what we saw in Week 2.
  - The main reason is because Brian Ripley, the author of the nnet package, does not believe that one-at-a-time intervals should be calculated. For example, my program shows how to calculate one-at-a-time 95% intervals as

$$0.7376 < \pi_{\text{Healthy}} < 0.9728$$
$$-0.0067 < \pi_{\text{Scab}} < 0.0995$$
$$0.0143 < \pi_{\text{Sprout}} < 0.1825$$

    for the first observation.
- Of course, $\pi_{\text{Healthy}} + \pi_{\text{Scab}} + \pi_{\text{Sprout}} = 1$ needs to occur.
- If we added the upper limits from the intervals together, we have a total greater than 1! For this reason, Ripley advocates constructing a confidence region. However, this is much more difficult to calculate, and he does not provide any code (no one else provides any code, either) to calculate it for these types of models. More discussion is included in the text.

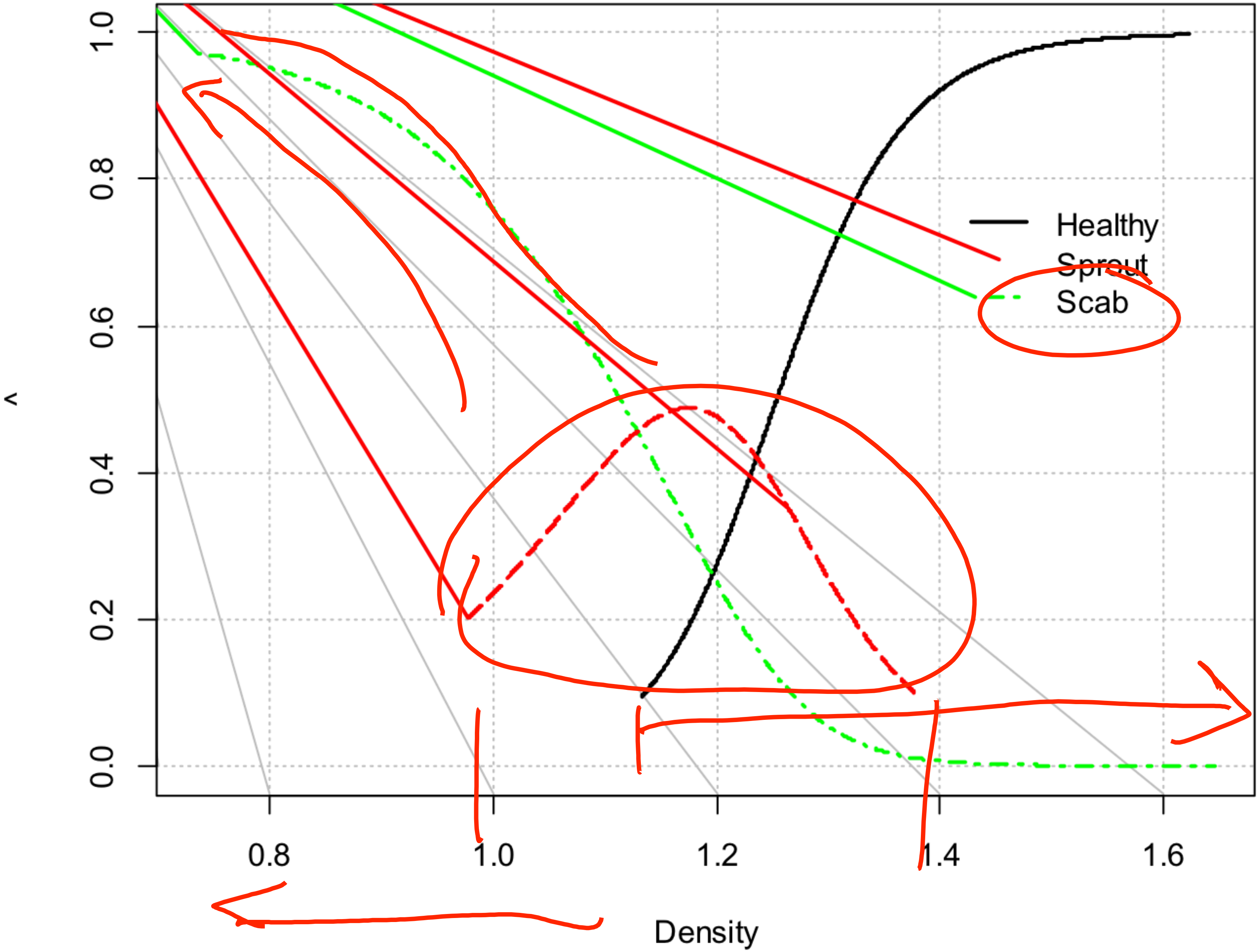
Visualize Estimated Probability

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

$$\log(\hat{\pi}_{scab} / \hat{\pi}_{healthy}) = 29.38 - 24.56\text{density}$$

and

$$\log(\hat{\pi}_{sprout} / \hat{\pi}_{healthy}) = 19.12 - 15.48\text{density}$$


1.0
0.8
0.6
0.4
0.2
0.0
^
0.8
1.0
1.2
1.4
1.6
Density
Healthy
Sprout
Scab

Estimated Probabilities

Below is a summary of the estimated probabilities for selected density values:

```
  density.values Healthy Scab Sprout
1            0.8    0.00 0.95   0.05
2            0.9    0.00 0.89   0.11
3            1.0    0.01 0.76   0.24
4            1.1    0.05 0.54   0.41
5            1.2    0.27 0.25   0.48
6            1.3    0.69 0.05   0.25
7            1.4    0.92 0.01   0.07
8            1.5    0.98 0.00   0.02
9            1.6    1.00 0.00   0.00
```

- The lines are drawn from the smallest to the largest observed density value for a wheat kernel condition.
- We see that the estimated scab probability is the largest for the smaller density kernels. The estimated healthy probability is the largest for the high-density kernels. For density levels in the middle, sprout has the largest estimated probability. The parallel coordinates plot displays similar findings where the density levels tend to follow the scab < sprout < healthy ordering.

Berkeley SCHOOL OF INFORMATION