Week 10: Advanced Discrete Data

Multinomial Choice, Count Data

Max H. Farrell
The University of Chicago Booth School of Business
Today we will look at **discrete** responses.

In Week 4 we covered binary: $Y = 0$ or $1$.

**Two extensions today:**

- More categories: $Y = 0, 1, 2, 3, 4$
  - Unordered: buy product A, B, C, D, or nothing
  - Ordered: rate 1–5 stars

- Count: $Y = 0, 1, 2, 3, 4, \ldots$
  - How many products bought in a month?

The goal varies depending on the type of response.

At heart, we will keep using **linear models**.
First, more categories:

\[ Y = 0, 1, 2, \ldots, M. \]

These categories can be:

- **Unordered**: Buy product A, B, C, or nothing
- **Ordinal**: Low, medium, high risk
- **Cardinal**: Rate 1-5 stars

Predict choices/actions based on characteristics

... then do classification into multiple categories.

We will see *multinomial* and *conditional* logistic regression, but there are others: original, nested, mixed, universal, ...  

- For count-type (e.g. stars) the *Poisson* model works too.
Discrete Choice

Units (firms, customers, etc) face a set of choices, labeled \{0, 1, 2, \ldots, M\}. Or we want to classify units into one of these categories.

Choose based on utility (happiness/profit/etc)

- Utility unit \(i\) gets from option \(m\): \(U_{i,m} = X_{i,m}\beta_m + \varepsilon_{i,m}\).
- Chooses \(m^*\) if \(U_{i,m^*}\) is the biggest.
- So we want to model probabilities like \(\mathbb{P}[U_{i,m^*} > U_{i,m}]\)

Gives rise to a set of log-odds, one for each category:

\[
\log \left( \frac{\mathbb{P}[Y_i = m \mid X_i]}{\mathbb{P}[Y_i = 0 \mid X_i]} \right) = \beta_{0,m} + \beta_{1,m}X_i.
\]

Just like regular (binary) logistic regression.

\[\Rightarrow\textbf{ Interpretation} \text{ is just as easy (hard?) as before}\]
Words of **warning!**

As usual, we’re skipping a lot of statistical details. But now, we’re skipping economic/substantive stuff too:

- What are we assuming about the agent’s behavior?
- Are the choices structured (e.g. ordered)?
- What are the random shocks $\varepsilon_{i,m}$? Independent over $m$?
- Are the $X$ variables specific to the product (e.g. price) or the person (e.g. income)?
- Should the intercepts and/or coefficients vary over the alternatives ($\beta_{0,m}, \beta_{1,m}$) or not?

Different assumptions are matched by different models, interpretation requires some care.

- **Multinomial logit** is a good all-purpose choice, especially for pure prediction.
Remember: The more complex the model, the more the assumptions matter!

It’s worth talking about the big one here:

The **Independence of Irrelevant Alternatives**

- If \( m^* \) is chosen from \( \{0, 1, 2, \ldots, M\} \), then \( m^* \) is chosen from any subset of \( \{0, 1, 2, \ldots, M\} \).
- The log-odds of \( m \) vs. 0 doesn’t depend on other options.
- So if we change/remove options from the choice set, the odds of choosing \( m \) vs. 0 do not change.

Does that seem realistic?

- Does it matter? Maybe not for classification.

What to do?

- For now, we’ll ignore the problem, but you can test if IIA holds, then run a different model if need be.
Example: Choice of cracker brand

```r
> names(cracker <- read.csv("cracker_choice.csv"))
> table(cracker$choice)
kleebler nabisco private sunshine
   226  1792  1035   239
> cracker$choice <- relevel(cracker$choice, "private")
```

Perfect data for relationship or inference questions:

- How does my price relate to choice?
- How do others’ prices relate to choice?
- Does an ad help? Anyone’s ad?

But the $X$’s are about the product, not the person:

- Can’t figure out what customer type buys a which brand.
- What would prediction/classification mean here?
Descriptives and Visualizations
Not as simple as linear regression (no lines to graph) but still an important step.

- Look at the price of one option when others are chosen

```
> colMeans(cracker[cracker$choice!="sunshine",2:5])
pr.private  pr.keebler  pr.nabisco  pr.sunshine
   68.01   112.53   107.74     96.44
> colMeans(cracker[cracker$choice=="sunshine",2:5])
pr.private  pr.keebler  pr.nabisco  pr.sunshine
   68.90   113.40   110.22     86.26
```
The price of all the options when Sunshine is chosen:
Final warning: Different packages/platforms do things differently, and it’s not always clear right away what or why.

First two Google results for “r multinomial logistic regression”:


   We use the `multinom` function from the `nnet` package to estimate a multinomial logistic regression model. There are other functions in other R packages capable of multinomial regression. We chose the `multinom` function because it does not require the data to be reshaped (as the `mlogit` package does).


   Multinomial logistic regression can be implemented with `mlogit()` from `mlogit` package and `multinom()` from `nnet` package. We will use the latter for this example.

So choose randomly? Or for convenience?

No! Pick a model (common coefficients? person or option specific $X$’s? . . . ) and then make sure you are getting it. `mlogit` works well and is very flexible, 

...but has some startup cost
Different slopes:

$$\log \left( \frac{\Pr[Y_i = m \mid X_i]}{\Pr[Y_i = 0 \mid X_i]} \right) = \beta_0 + \beta_{1,m} X_i$$

(aka *multinomial logit*)

```r
fit1 <- mlogit(choice ~ 1 | pr.private + pr.keebler + pr.nabisco +
+ pr.sunshine - 1, data=mlogit.data(cracker, choice="choice",
+ shape="wide"), reflevel="private"
) > summary(fit1)
```

Common slopes:

$$\log \left( \frac{\Pr[Y_i = m \mid X_i]}{\Pr[Y_i = 0 \mid X_i]} \right) = \beta_{0,m} + \beta_1 X_i$$

(aka *conditional logit*)

```r
fit2 <- mlogit(choice ~ pr, data=mlogit.data(cracker, choice="choice",
+ shape="wide", varying=c(2:5)), reflevel="private"
) > summary(fit2)
```
Interpretation

Exponentiate coefficients for interpretation

Single-slope model is easy:

```r
> exp(coef(fit2))

keebler:(intercept)  nabisco:(intercept)  sunshine:(intercept)
        0.9922845         7.1344553         0.5718713

pr
        0.9661083
```

▶ If the price of any name brand goes up by $0.05, then the odds of buying go down by a factor of $0.966^5 = 0.84$.

▶ Or, the odds of buying the store brand increase by $1/0.84, \approx 20\%$
Interpretation

Exponentiate coefficients for interpretation

Different slopes model is also interesting:

\[
\text{> exp(matrix(coef(fit1), .......)}
\]

\[
\begin{array}{cccc}
\text{pr.private} & \text{pr.keebler} & \text{pr.nabisco} & \text{pr.sunshine} \\
\text{keebler} & 1.02 & 0.957 & 1.012 & 1.007 \\
\text{nabisco} & 1.02 & 1.020 & 0.973 & 0.999 \\
\text{sunshine} & 1.01 & 1.023 & 1.007 & 0.942 \\
\end{array}
\]

▶ Why is \( b > 1 \) for \text{pr.private} always?
▶ Own price coefficient < 1?
▶ What if Keebler lowers their price by $0.05?
  → people \( 1/0.957^5 \approx 25\% \) more likely to buy Keebler
  → Nabisco sales go down by \( 1/1.020^5 \approx 10\% \)
  ... versus the store brand
Question: Do advertisements increase sales?

To tackle this, we have to decide if coefficients should be common or not:

```r
> fit1.ad <- mlogit(choice ~ 1 | pr.private + pr.keebler + pr.nabisco +
+ pr.sunshine + ad.private + ad.keebler + ad.nabisco + ad.sunshine
+ - 1, data=mlogit.data(cracker, choice="choice",
+ shape="wide"), reflevel="private")

> fit2.ad <- mlogit(choice ~ ad | pr.private + pr.keebler + pr.nabisco +
+ pr.sunshine - 1, data=mlogit.data(cracker, choice="choice",
+ shape="wide", varying=c(6:9)), reflevel="private")

> fit3.ad <- mlogit(choice ~ pr + ad, data=mlogit.data(cracker, choice="choice"
+ shape="wide", varying=c(2:9)), reflevel="private")
```

You have to narrow the question, and make assumptions.

- Do customers respond to ads for brands the same way?
- Does Keebler running an ad increases Nabisco sales?
- Do ads in general raise awareness of crackers?
Question: Do advertisements increase sales?

Try Model 2: common slope on ad, different price effects

```r
> exp(coef(fit2.ad)[1])
adTRUE
1.129938
> exp(matrix(coef(fit2.ad)[2:13], nrow=3, ......
            pr.private pr.keebler pr.nabisco pr.sunshine
  keebler   1.02   0.957   1.011   1.008
  nabisco   1.02   1.019   0.974   0.999
  sunshine  1.01   1.022   1.006   0.944
```

Ads do increase\(^1\) sales of name brand crackers!

- Raise awareness/desire overall?
- But why do name brands benefit more than store brand?
- Even store brand ads contribute to this.

\(^1\) I meant “are associated with”
Question: Do advertisements increase sales?

Try Model 1:

```r
> exp(matrix(coef(fit1.ad)[1:12], ....
   pr.private pr.keebler pr.nabisco pr.sunshine
keebler  1.02  0.961  1.004  1.008
nabisco  1.02  1.020  0.971  0.998
sunshine 1.01  1.015  1.012  0.941
> exp(matrix(coef(fit1.ad)[13:24], ....
   ad.privateTRUE ad.keeblerTRUE ad.nabiscoTRUE ad.sunshineTRUE
keebler  1.07  2.51  0.892  1.11
nabisco  1.09  2.17  1.250  1.05
sunshine 1.08  1.29  2.094  1.22
```

▶ How much more likely are people to buy *keebler* (vs. store brand) if they run an *ad*?
   → \(2.5\times\) as likely! (OR multiplied by 2.51.)

▶ *Keebler* ad makes people 2.17 as likely to buy *Nabisco*?

▶ Others are more or less sensitive.

▶ Price effects are similar to before, reassuring pattern
Prediction and Classification

Predictions are still predicted probabilities, but for each level:

```r
> mlogit.probs <- fitted(fit1.ad, type="probabilities")
> dim(mlogit.probs)
[1] 3292 4
> head(mlogit.probs)
private   keebler   nabisco   sunshine
1 0.4436256 0.22767567 0.2891718 0.03952697
2 0.3174992 0.06503254 0.5886932 0.02877505
3 0.1707861 0.03133935 0.3193414 0.47853312
4 0.2384031 0.05639767 0.6888423 0.01635692
```

← each row sums to one as expected

Classification

- May not make sense in some discrete choice applications, but useful in general
- Predict what people buy, what risk class they fall into, …
Classification

We can’t use a simple cut-off any more:

what if two categories have \( \hat{P}[Y = m | X] > c \)?

The simplest, and most widely used idea, is to use the highest predicted probability:

\[
\hat{Y}_f = m^* \\
\Leftrightarrow \quad \hat{P}[Y_i = m^* | X_i] > \hat{P}[Y_i = m | X_i], \quad m = 0, \ldots, M
\]

Easy to find in R:

\[
> \text{colnames(mlogit.probs)[apply(mlogit.probs,1,which.max)]}
\]

Check your results!

\[
> \text{table(predicted.category,cracker$choice)}
\]

\( \Leftrightarrow \) What went wrong?
Poisson regression is a GLM designed for count data. How?

Remember the Poisson distribution:

- Integers only
- One parameter: $\lambda > 0$
- $\lambda =$ Mean and Variance

Extensions: Negative Binomial Model, Zero-Inflated Models
**Example:** Number of take-over bids (**NUMBIDS**) received by 126 US firms during 1978–85.

Covariates include

- Continuous variables: **BIDPREM**, **INSTHOLD**, **SIZE**
- Indicators: **LEGLREST**, **FINREST**, **REGULATN**, **WHTKNIGHT**

Looks like Poisson!
Browsing the data

More bids on average for the indicators being “Yes”:

<table>
<thead>
<tr>
<th>Indicator</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEGLREST</td>
<td>1.46</td>
<td>2.11</td>
</tr>
<tr>
<td>REALREST</td>
<td>1.69</td>
<td>1.96</td>
</tr>
<tr>
<td>FINREST</td>
<td>1.70</td>
<td>2.08</td>
</tr>
<tr>
<td>WHTKNIGHT</td>
<td>1.18</td>
<td>2.12</td>
</tr>
<tr>
<td>REGULATN</td>
<td>1.67</td>
<td>1.91</td>
</tr>
</tbody>
</table>

Pattern in the continuous variables?

- Bid price / stock price before bid
- % held by institutions
- Total assets
Poisson regression results:

> poisson.fit <- glm(NUMBIDS ~ BIDPREM+INSTHOLD+SIZE+LEGLREST + +REALREST+FINREST+ WHTKNGHT+REGULATN, family="poisson")
> summary(poisson.fit)

Coefficients:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 1.10489  | 0.53443    | 2.067   | 0.03870  *|
| BIDPREM        | -0.76982 | 0.37676    | -2.043  | 0.04103  *|
| INSTHOLD       | -0.13180 | 0.40027    | -0.329  | 0.74195  |
| SIZE           | 0.03428  | 0.01865    | 1.839   | 0.06598  .|
| LEGLREST       | 0.25191  | 0.15005    | 1.679   | 0.09318  .|
| REALREST       | -0.02024 | 0.17726    | -0.114  | 0.90909  |
| FINREST        | 0.09190  | 0.21688    | 0.424   | 0.67176  |
| WHTKNGHT       | 0.49821  | 0.15822    | 3.149   | 0.00164  **|
| REGULATN       | 0.01178  | 0.16297    | 0.072   | 0.94237  |

⇒ The expected log(NUMBIDS) increases by 0.498 ≈ 1/2 if a White Knight is involved. What?
Just like logistic regression, exponentiating aids interpretation.

```r
> exp(coef(poisson.fit))

(Intercept)    BIDPREM    INSTHOLD     SIZE    LEGLREST
 3.0188810   0.4630954   0.8765198  1.0348739  1.2864796
REALREST    FINREST    WHTKNIGHT   REGULATN
 0.9799633   1.0962533  1.6457787  1.0118518
```

⇒ The incidence rate is 1.65 times higher if a White Knight is involved.

The incidence rate ratio (IRR, or risk ratio) is *almost* like the odds ratio from logistic regression: $\beta_j > 0 \Leftrightarrow \text{IRR} > 1$.

- The number of bids (the count we’re modeling) is expected to be 65% higher if a White Knight is involved.
- For a one-unit increase in the % held by institutions, number of bids falls by $1 - 0.88 = 12\%$. 
The End

Whew! We made it!

Thanks for all your hard work. I know it’s been tough.

I hope everyone got something (a lot?) out of it.

- Good luck!
- Keep working on projects.
- Feedback please!
What we did

W1: Introduction, The simple linear regression (SLR) model
W2: Inference for SLR
W3: Multiple linear regression (MLR)
W4: Regression diagnostics and data transformations
W5: MLR: Causal inference, model building I
W6: Midterm
W7: MLR: model building II, data mining
W8: Generalized Linear Models, classification
W9: Time series models and autoregression
W10: Advanced topics in dependence (VAR, Clustering, Panels)
W11: Final exam and Project due