Business Statistics

Multinomial Regression

Zhanrui Cai
Assistant Professor in Analytics and Innovation

Review of Logistic Regression

- Used for classification.

- Model the probability that X belongs to each category in C:

$$Logit\ p(X) = \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Estimate β by MLE.

- e^{β_k} is explained as odds ratio for the variable X_k.

- Hypothesis testing $H_0: \beta_k = 0$ from R output.

Nominal and Ordinal Responses

- Nominal response
  - Red, green, blue
  - Yes, no
  - Sick, healthy
- Ordinal response
  - Young, middle aged, old
  - Dislike very much, dislike, no opinion, like, like very much

Example: Car Preferences

- In a study of motor vehicle safety, 150 men and 150 women were interviewed to rate how important air conditioning and power steering were to them when they were buying a car.

		Response		
Sex	Age	Unimportant	Import	Very Import
Women	18-23	26 (58%)	12 (27%)	7 (16%)
	24-40	9 (20%)	21 (47%)	15 (33%)
	> 40	5 (8%)	14 (23%)	41 (68%)
Men	18-23	40 (62%)	17 (26%)	8 (12%)
	24-40	17 (39%)	15 (34%)	12 (27%)
	> 40	8 (20%)	15 (37%)	18 (44%)
Total		105	94	101

Nominal Logistic Regression

- Choose one category as the reference category, say the 1st category
- Define the logits for the other categories as

$$\text{logit}(\pi_j) \equiv \log(\frac{\pi_j}{\pi_1}) = x^T\beta_j, \quad \text{for } j = 2, \ldots, J.$$

- The joint density is

$$f(\mathbf{y}|n) = (\pi_1)^{y_1} \ldots (\pi_J)^{y_J} \frac{n!}{y_1! \ldots y_J!},$$

which leads to the following likelihood function,

$$l(\beta|\mathbf{y}, n) \propto \prod_{j=2}^{J} (\frac{\pi_j}{\pi_1})^{y_j} = \exp(\sum_j y_j x^T \beta_j).$$

Nominal Logistic Regression Estimate

- Given MLE, we have

$$\hat{\pi}_j = \hat{\pi}_1 \exp(x^T \hat{\beta}_j), \quad \text{for } j = 2, \ldots, J.$$

- Since the probabilities add up to 1, we have

$$\hat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp(x^T \hat{\beta}_j)}$$

$$\hat{\pi}_j = \frac{\exp(x_j^T \hat{\beta}_j)}{1 + \sum_{j=2}^{J} \exp(x^T \hat{\beta}_j)}, \quad \text{for } j = 2, \ldots, J.$$

- Changing the reference category won't change the above probabilities.

Example: Car Preference

Define the following three dummy variables,

- x_1: the indicator of men;
- x_2: the indicator of age 24-40 years;
- x_3: the indicator of age $>$ 40 years.

The model is then

$$\log(\frac{\pi_j}{\pi_1}) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3, \quad j = 2, 3.$$

R Command

```
> car <- data.frame(res.unim=c(26, 9, 5, 40, 17, 8),
          res.im=c(12, 21, 14, 17, 15, 15),
          res.veim=c(7, 15, 41, 8, 12, 18),
          sex=c(rep("F", 3), rep("M",3)),
          age=rep(c("18-23", "24-40", ">40"), 2))
> car
  res.unim res.im res.veim sex   age
1       26     12        7   F 18-23
2        9     21       15   F 24-40
```

```
> library(nnet)  ### special library containing ‘‘multinom’’
> options(contrasts=c("contr.treatment", "contr.poly"))
> car.mult <- multinom(cbind(res.unim, res.im, res.veim)~sex+age,
                                  data=car)
> summary(car.mult)
Coefficients:
  (Intercept)       sexM age24-40   age>40
2  -0.5907992 -0.3881301 1.128268 1.587709
3  -1.0390726 -0.8130202 1.478104 2.916757

Std. Errors:
  (Intercept)      sexM  age24-40    age>40
2   0.2839756 0.3005115 0.3416449 0.4028997
3   0.3305014 0.3210382 0.4009256 0.4229276
```

Estimation Results

The estimated coefficients are

$\hat{\beta}_{02} = -0.591, \hat{\beta}_{12} = -0.388, \hat{\beta}_{22} = 1.128, \hat{\beta}_{32} = 1.588,$

$\hat{\beta}_{03} = -1.039, \hat{\beta}_{13} = -0.813, \hat{\beta}_{23} = 1.478, \hat{\beta}_{33} = 2.917.$

To estimate the probabilities, consider the preferences of women ($x_1 = 0$) aged 18-23 ($x_2 = x_3 = 0$). For this group,

$$\log(\frac{\hat{\pi}_2}{\hat{\pi}_1}) = -0.591, \frac{\hat{\pi}_2}{\hat{\pi}_1} = 0.5539,$$

$$\log(\frac{\hat{\pi}_3}{\hat{\pi}_1}) = -1.039, \frac{\hat{\pi}_3}{\hat{\pi}_1} = 0.3538,$$

$$\hat{\pi}_1 = 1/(1 + 0.5539 + 0.3538) = 0.524,$$

$$\hat{\pi}_2 = 0.290, \hat{\pi}_3 = 0.186.$$

Hierarchical or Nested Responses

Data on live births with deformations of the central nervous system in south Wales.

```
> cns
           Area NoCNS An Sp Other Water      Work
1       Cardiff  4091  5  9     5   110 NonManual
2       Newport  1515  1  7     0   100 NonManual
3       Swansea  2394  9  5     0    95 NonManual
4    GlamorganE  3163  9 14     3    42 NonManual
5    GlamorganW  1979  5 10     1    39 NonManual
6    GlamorganC  4838 11 12     2   161 NonManual
7     MonmouthV  2362  6  8     4    83 NonManual
8 MonmouthOther  1604  3  6     0   122 NonManual
9       Cardiff  9424 31 33    14   110    Manual
10      Newport  4610  3 15     6   100    Manual
11      Swansea  5526 19 30     4    95    Manual
12   GlamorganE 13217 55 71    19    42    Manual
13   GlamorganW  8195 30 44    10    39    Manual
14   GlamorganC  7803 25 28    12   161    Manual
15    MonmouthV  9962 36 37    13    83    Manual
16 MonmouthOther 3172  8 13     3   122    Manual
```

CNS: Variables

- NoCNS: no central nervous system (CNS) malformation.
- An, Sp and Other: three categories of various malformation.
- Water: water hardness
- Work: the type of work performed by the parents.

Hierarchical Response Model

- We can consider a multinomial logit model with four response categories.
- However, the category NoCNS dominates the result.
- Better to perform a hierarchical response model.
  - A binomial model of CNS vs. NoCNS: whether a malfunction has occurred
  - A multinomial model of the three CNS categories: given a malfunction has occurred, what type of malfunction?

CNS: Conclusion

- Binomial model
  - Both Water and Work have significant effect on the probability of having a malformation.
- Multinomial model with three CNS categories
  - Both have no effect of distinguishing the three malformations.
- Multinomial model with NoCNS included
  - Both are significant, but this is misleading as mainly driven by the large NoCNS category.

Ordinal Logistic Regression

- Ordinal responses are common in marketing research, opinion polls, and so on where soft measures are common.

- Cumulative logit model

$$\log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} = x^T \beta_j.$$

- Special case: Proportional odds model

$$\log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}.$$

Example: Car Preference

The following proportional odds model was fitted to the data:

$$\log\left(\frac{\pi_1}{\pi_2+\pi_3}\right) = \beta_{01} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

$$\log\left(\frac{\pi_1+\pi_2}{\pi_3}\right) = \beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

which leads to the estimates below,

$$\beta_{01} = 0.044, \beta_{02} = 1.655, \beta_1 = 0.576,$$

$$\beta_2 = -1.147, \beta_3 = -2.232.$$

Question: Calculate probabilities for women aged 18-23.

R Command

```
> library(MASS)
> freq <- c(car$res.unim, car$res.im, car$res.veim)
> res <- c(rep(c("unim", "im", "veim"), c(6,6,6)))
> res <- factor(res, levels=c("unim", "im", "veim"), ordered=T)
> car.ord <- data.frame(res=res, sex=rep(car$sex, 3),
                        age=rep(car$age, 3), freq=freq)
> car.polr <- polr(res~sex+age, data=car.ord, weights=freq)
```

```
> car.polr
Coefficients:
      sexM   age24-40     age>40
-0.5762219  1.1470976  2.2324560

Intercepts:
   unim|im    im|veim
0.04353746 1.65497620

Residual Deviance: 581.2956
AIC: 591.2956
```

Case Study: Data Analytics at Chow Tai Fook

- Chow Tai Fook (CTF) is one of the world's largest jewellers, with a retail network of over 3,100 points of sale (POS) globally.

Retail Network in Mainland China
中國內地零售網絡

As at 31 March 2019 於2019年3月31日

CHOW TAI FOOK JEWELLERY 周大福珠寶 2,769 POS 零售點		HEARTS ON FIRE 3 POS 零售點	159 SIS/CIS 店中店/店內專櫃
ARTRIUM 周大福藝堂 2 POS 零售點		T MARK 4 POS 零售點	500 CIS 店內專櫃
JEWELRIA 周大福薈館 32 POS 零售點	CTF Watch 周大福鐘錶 113 POS 零售點	SOINLOVE 27 POS 零售點	MONOLOGUE 38 POS 零售點

Total POS in Mainland China 中國內地總零售點
2,988 POS 零售點

Retail Network in Hong Kong, Macau and other markets
香港、 澳門及其他市場零售網絡

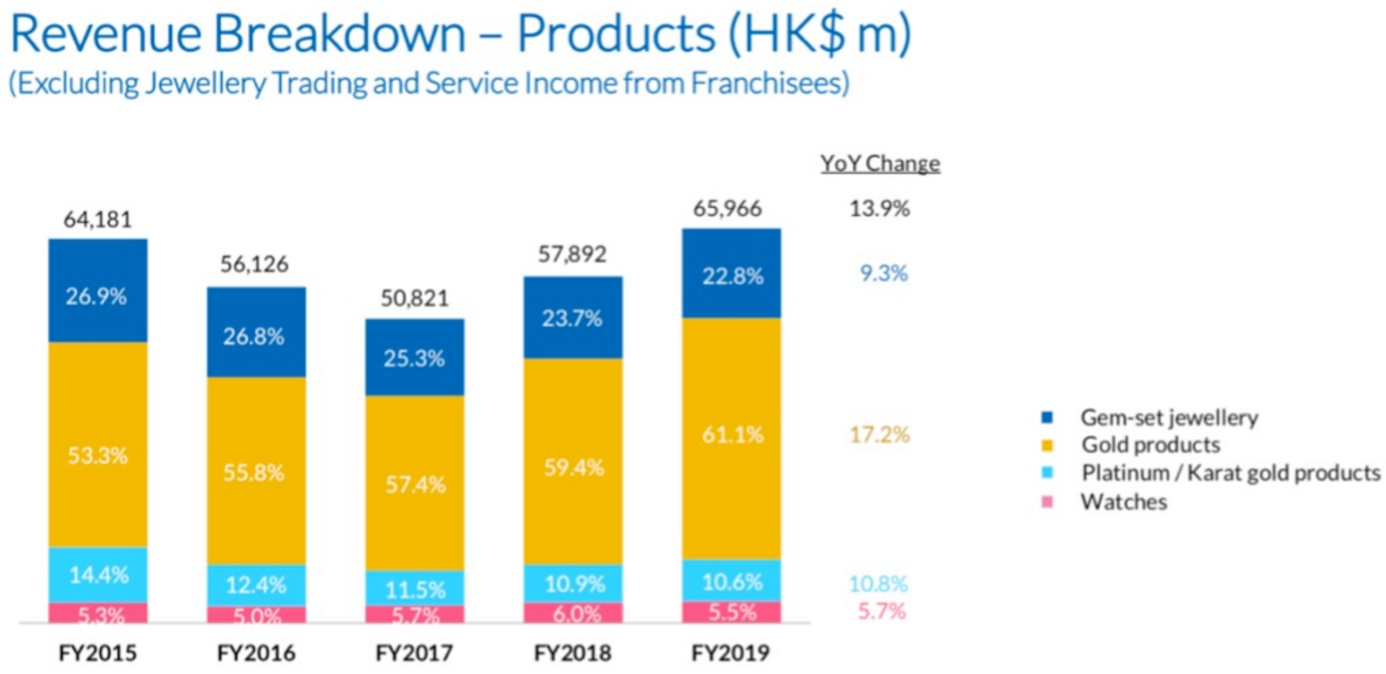
As at 31 March 2019 於2019年3月31日

CHOW TAI FOOK JEWELLERY 周大福珠寶 Hong Kong and Macau 香港及澳門 98 POS 零售點	Other markets 其他市場 31 POS 零售點	HEARTS ON FIRE 15 POS 零售點	27 SIS/CIS 店中店/店內專櫃
ARTRIUM 周大福藝堂 1 POS 零售點		T MARK 1 POS 零售點	55 CIS 店內專櫃

Total POS in Hong Kong, Macau and other markets
香港、澳門及其他市場總零售點
146 POS 零售點

Case Study: Data Analytics at Chow Tai Fook

- CTF offered products in four major categories, including
  - gem-set jewellery,
  - gold products,
  - platinum/karat gold products,
  - watches.

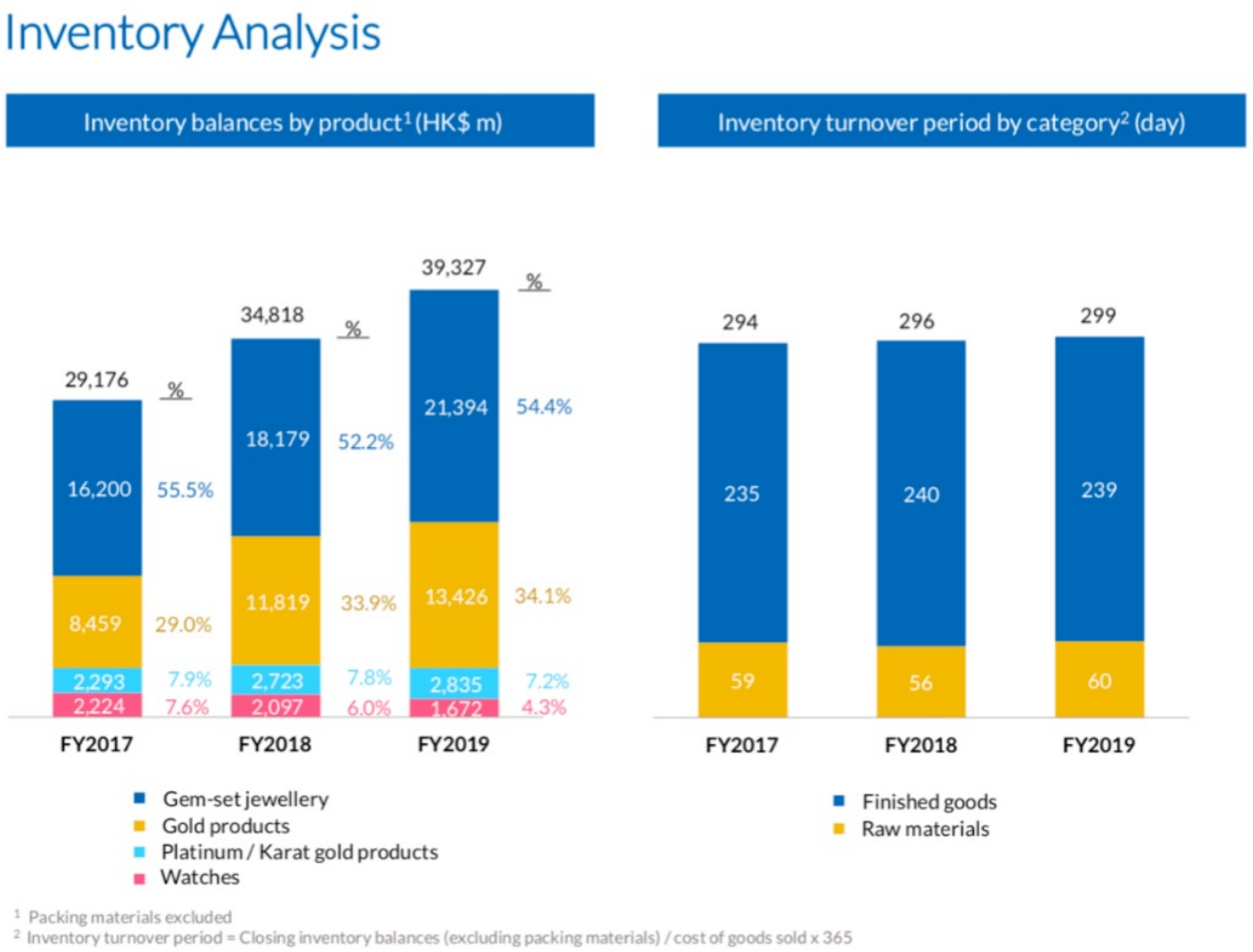


% of revenue	1H2018	2H2018	1H2019	2H2019
Gem-set jewellery	24.2%	23.4 %	23.4%	22.2%
Gold products	57.8%	60.5%	60.5%	61.5%
Platinum / Karat gold products	11.3%	10.6%	10.5%	10.7%
Watches	6.7%	5.4%	5.6%	5.5%

Case Study: Data Analytics at Chow Tai Fook

- Unlike fast-moving consumer goods, the high value of inventory and slow turnover of individual SKUs in the jewellery industry made good inventory management the key to healthy profitability.

- Our goal: predict consumers' choice of products, to help with inventory management.



Predicting Customers' Choice

- Data were sourced from 2 channels:
  - The basic information on each SKU, such as price and weight (e.g. categorical variables r_info1, r_info3, common_info1).
  - POS data on the location and timing of each purchase, as well as the daily traffic captured in retail stores (e.g. branch indicator and customer arrival count).

- We looked at purchases of 35 products from 3 CTF branches.

- Given features a customer wants, which of the 35 products the customer is more likely to buy?
  - Use multinomial regression model?
  - Nominal or ordinal?

R Command

- Training data: 20050112 - 20061103
- Testing data: 20061104 - 20070103

```
# Training/Testing set split
training <- my_data %>% filter(baseDate < "2006-11-04")
nrow(training)
```

```
## [1] 7980
```

```
testing <- my_data %>% filter(baseDate >= "2006-11-04")
nrow(testing)
```

```
## [1] 3080
```

```
training_selected <- training[training$mode == TRUE,]

training_multinom <- training
training_multinom$purchaseID <- rep(training_selected$productID, each=35)

fit.multinom <- multinom(purchaseID ~ r_info1_11 + r_info1_111_126 +
                    r_info3_1 + r_info3_23 + r_info3_4567 + common_info1_0 +
                    ct + branchID_16 + branchID_26, data=training_multinom)
```

Performance of Multinomial Regression

- Accuracy of the predicted top 3 choices

  = percentage that 3 products with the highest predicted probability / logit contains the true choice

```r
train_result_df <- predict(fit.multinom, newdata=training_selected,
                           type="probs")
train_correct_cnt_1 <- 0
for (i in 1:length(train_actual)){
  array_1 <- sort(desc(train_result_df[i,]))[1:3]
  name_1 <- names(array_1)
  if (train_actual[i] %in% name_1){
    train_correct_cnt_1 = train_correct_cnt_1 + 1
  }
}
train_correct_cnt_1
```

- Training Accuracy = 43.4%
- Testing Accuracy = 27.3%
- Much better than random guess accuracy = 1/35 = 2.9%

Multinomial Logistic Regression Summary

- Extension of logistic regression to more than two categories
  - Include logistic regression as special case when only two categories

- Nominal or ordinal

- Estimate β by MLE

- e^{β_k} is explained as odds ratio for the variable X_k
  - Between which two categories

Model Selection in Logistic Regression

Model Selection in Logistic Regression

- Previous techniques can be applied.
- Build models: best subset regression, stepwise regression, Lasso, ridge.
- Select variables/ tunning parameters: AIC, BIC, Cross-validation.

MLE

- Recall the likelihood:

$$l(\beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} \left(1 - p(x_i)\right)$$

- where

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}.$$

Lasso + Ridge

- Lasso:

$$-2\log\big(l(\beta)\big) + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$\max_{\beta_0,\beta}\left\{\sum_{i=1}^{N}\left[y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\right]-\lambda\sum_{j=1}^{p}|\beta_j|\right\}.$$

- Ridge:

$$-2\log\big(l(\beta)\big) + \lambda\sum_{j=1}^{p}\beta_j^2$$

- Use R function glmnet to fit, change family to be binomial.

Linear Discriminant Analysis

- Suppose that we model each class density as multivariate Gaussian.

- For data x in class k, we assume its density to be

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k)}.$$

- Linear discriminant analysis (LDA) assumes that when comparing two classes k and l,

$$
\begin{aligned}
\log \frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} &= \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell} \\
&= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k+\mu_\ell)^T \mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell) \\
&\quad + x^T \mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell),
\end{aligned}
$$

- which is a linear function in x.

Estimation

- In practice we do not know the parameters of the Gaussian distributions, and will need to estimate them using our training data:

  - $\hat{\pi}_k = N_k/N$, where N_k is the number of class-k observations;
  - $\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$;
  - $\hat{\mathbf{\Sigma}} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N-K)$.

Estimation

- The LDA rule classifies to class 2 if

$$x^T \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2^T \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^T \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_1 + \log(N_1/N) - \log(N_2/N)$$

Logistic Regression or LDA?

- It seems that the models (for log odds) are the same.
- The difference lies in the way the linear coefficients are estimated.
- The logistic regression model is more general, in that it makes less assumptions.

Logistic Regression or LDA?

- The logistic regression model leaves the marginal density of X as an arbitrary density function P(X), and fits the parameters of P(G|X) by maximizing the likelihood.

- For LDA, we fit the parameters by maximizing the full log-likelihood.

- In practice these assumptions are never correct, and often some of the components of X are qualitative. It is generally felt that logistic regression is a safer, more robust bet than the LDA.