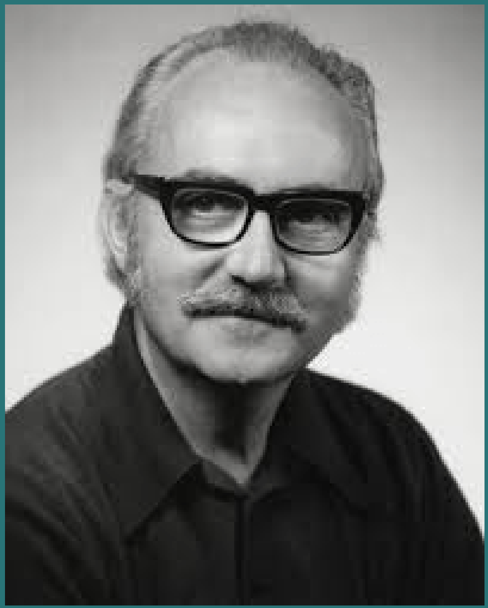
# Discrete Choice

## 离散选择模型

*All models are wrong, some are useful.*

所有的模型都是错的，但有些是有用的。

--- George Box



Question:

How do machines recognize hand-written digits?

机器是如何识别手写数字的?

What brand is my smartphone?

我的手机是什么品牌的?

Apple

Xiaomi

Huawei

OPPO

What is the brand of the HKU president’s car?

香港大学校长开什么牌子的车?

How do users choose among different banks?

用户是如何选择银行的?

中国建设银行
China Construction Bank

ICBC

中國銀行
BANK OF CHINA

Daniel McFadden’s model became so popular, and he won the Nobel Prize in Economics in 2000 for “his development of theory and methods for analyzing discrete choice.”

Daniel McFadden 建立了研究个人选择的模型。他的模型非常成功，为他赢得了2000年的诺贝尔经济学奖，颁奖词是“开创了研究离散选择模型的理论和方法。”

# Modelling Consumer Choice

Human-beings always need to make choices, from your marriage choice to buying a bottle of milk.

While individuals can make choices in their own ways, as consumer analysts, we do want to understand how consumers make their choices.

# 消费者行为建模

人的一生离不开各种各样的选择，小到买哪个品牌的牛奶，大到如何选择自己的婚姻和职业。张雪峰?

每个人都用他们自己的方式进行各种选择。但是，我们还是希望尽可能的理解他们是如何进行选择的，并预测消费者的选择。

# Imaging that you are a bank manager.


You want to understand how consumers choose between different credit card companies when applying for credit cards. In this way, you can understand which are really your potential clients, and you can target on these consumers better.

# 假设你是银行的经理...


你想知道大家是怎么去选择信用卡的，这样可以帮助你分析你的潜在客户，并改善你的产品来吸引更多的客户。

# Your data is as follows...

For each consumer, you know his or her demographics (e.g., gender, age), occupation, income, geographic location, credit histories, etc. These are your independent variables.

You also know which credit card they applied to, e.g., Citibank, HSBC, BOC, American Express, ... or none of the above. This is your dependent variable.

You task: Building a model that predicts the dependent variable using your independent variables.

# 你的数据如下...

你知道每个消费者的人口统计学信息 (例如年龄，性别，民族), 职业，收入，地理位置，信用历史等等，你可以把他们作为你的自变量(解释变量)。

你也知道每个消费者申请了哪个银行的信用卡(建设银行，中国银行，工商银行，汇丰银行等)，而这是你的因变量(被解释变量)。

你的任务：建立你的统计学模型，通过自变量预测你的因变量。

# What would you do?

你会怎么做?

# Let us start with something simpler.

Now, you want to predict whether or not a consumer applies for your company's credit card. Here, the dependent variable $Y_i$ is YES or NO. For simplicity, let $Y_i = 1$ for YES and $Y_i = 0$ for NO.

For each individual, the independent variables again include demographics, occupation, income, location, etc. We use $X_i$ to denote the independent variables.

Our task: Predict $Y_i$ using $X_i$.

# 我们先来点简单的...

我们考虑一个简单的问题，我们只分析消费者有没有申请建设银行的信用卡。这里，你的被解释变量 $Y_i$ 的取值是 YES 或者 NO. 简单起见，我们用 $Y_i = 1$ 代表 YES，用 $Y_i = 0$ 代表 NO.

我们仍然知道每个消费者的人口统计学信息 (例如年龄，性别，民族), 职业，收入，地理位置，信用历史等等，并且用 $X_i$ 表示这些被解释变量。

你的任务：用 $X_i$ 来预测 $Y_i$.

# What should you do?

Our task: Predict $Y_i$ using $X_i$, where $Y_i \in \{0, 1\}$.

Question: Can we use linear regression to analyze the relationship between $Y_i$ and $X_i$, that is, we use the following linear model:

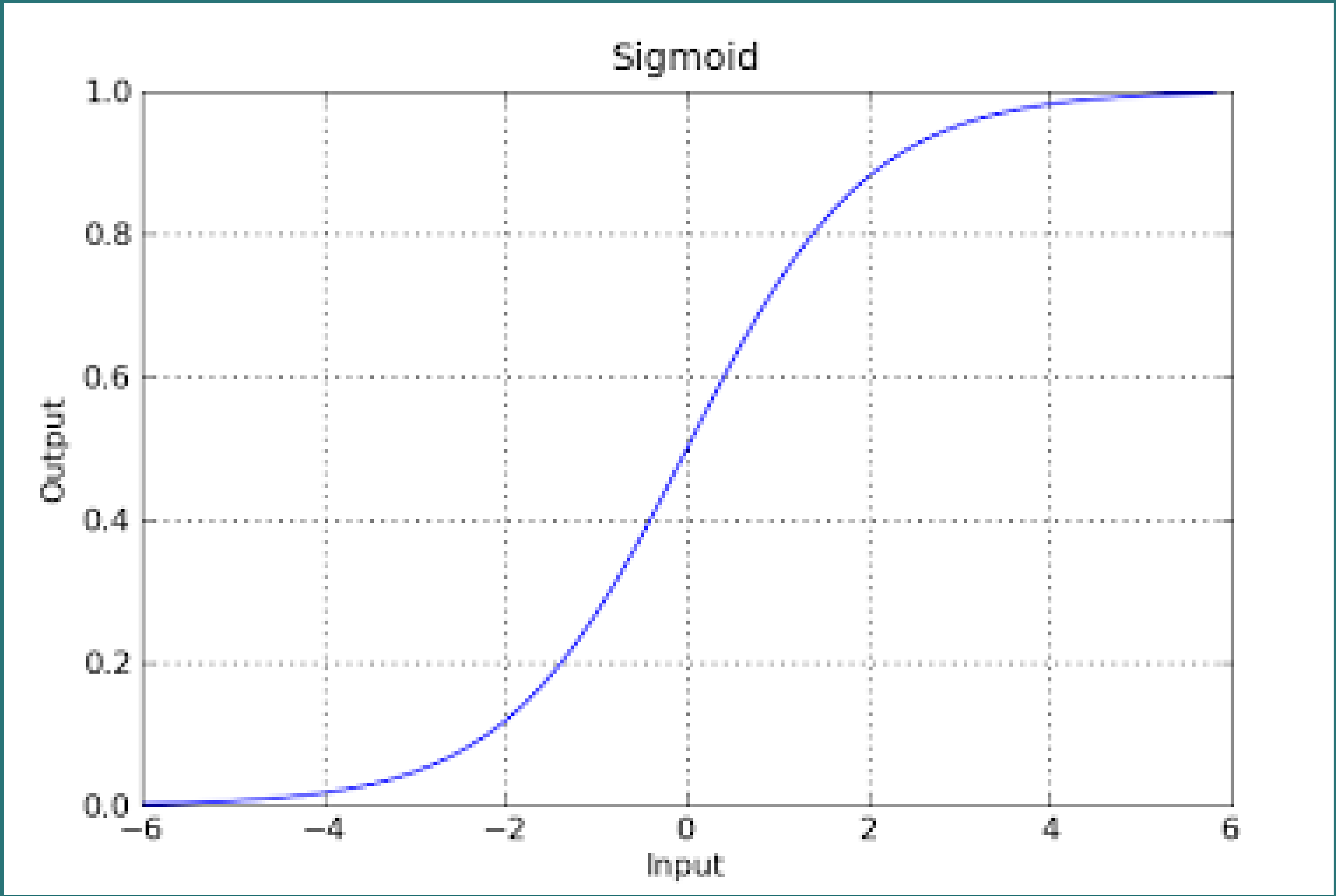
$$Y_i = \alpha + \beta X_i$$

# 我们该怎么做?

我们的任务：用 $X_i$ 来预测 $Y_i$，其中 $Y_i \in \{0,1\}$.

问题: 我们能不能用简单的线性回归来预测$Y_i$ 和 $X_i$ 的关系？这里，我们的回归方程是这样的:

$$Y_i = \alpha + \beta X_i$$




The logistic function 逻辑函数

As an illustration, we first load the following dataset in R.

我们看看下面的数据：

```
library(readr)
mydata <- read.csv("https://ximarketing.github.io/data/banking.csv")
head(mydata)
```

The data reads as follows:

具体的数据是这样的：

| | age | job | previous | success |
|---|---|---|---|---|
| 1 | 44 | blue-collar | 0 | 0 |
| 2 | 53 | technician | 0 | 0 |
| 3 | 28 | management | 2 | 1 |
| 4 | 39 | services | 0 | 0 |
| 5 | 55 | retired | 1 | 1 |
| 6 | 30 | management | 0 | 0 |

| | age | job | previous | success |
|---|---|---|---|---|
| 1 | 44 | blue-collar | 0 | 0 |
| 2 | 53 | technician | 0 | 0 |
| 3 | 28 | management | 2 | 1 |
| 4 | 39 | services | 0 | 0 |
| 5 | 55 | retired | 1 | 1 |
| 6 | 30 | management | 0 | 0 |

The data is about the outcome of a marketing campaign in a Portuguese banking that promotes a term deposit to their clients. Success denotes the final outcome of the campaign (1 = success, 0 = failure).

- Job includes admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown.
- Previous denotes the number of previous contacts with the client.

| | age | job | previous | success |
|---|---|---|---|---|
| 1 | 44 | blue-collar | 0 | 0 |
| 2 | 53 | technician | 0 | 0 |
| 3 | 28 | management | 2 | 1 |
| 4 | 39 | services | 0 | 0 |
| 5 | 55 | retired | 1 | 1 |
| 6 | 30 | management | 0 | 0 |

这是一家葡萄牙银行一次营销活动的数据。Success 代表这次营销活动对这名消费者是否成功 (1 = 成功, 0 = 失败).

- Job 包括 admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown.
- Previous 代表之前公司联系过消费者的次数.

```r
result <- glm(success ~ age + factor(job) + previous, data
              = mydata, family = "binomial")
summary(result)
```

Next, we build up a logistic regression model using success to be the dependent variable, independent variables include age, job, and number of previous contacts.

Note that because “job” is not a number, we treat it as a fixed effect by enclosing it within a factor bracket.

```r
result <- glm(success ~ age + factor(job) + previous, data
              = mydata, family = "binomial")
summary(result)
```

接下来，我们建立一个逻辑回归模型。其中，我们的因变量为营销活动是否成功，而自变量包括年龄，工作类型，和之前联系的次数。

注意工作本身不是一个数字，因此我们把工作当做一个固定效应来进行处理。

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -2.217305   0.074546 -29.744  < 2e-16 ***
age                       0.001890   0.001776   1.064 0.287149    
factor(job)blue-collar   -0.625683   0.051213 -12.217  < 2e-16 ***
factor(job)entrepreneur  -0.416913   0.099843  -4.176 2.97e-05 ***
factor(job)housemaid     -0.257722   0.109806  -2.347 0.018922 *  
factor(job)management    -0.174238   0.067648  -2.576 0.010005 *  
factor(job)retired        0.667628   0.078456   8.510  < 2e-16 ***
factor(job)self-employed -0.188631   0.093062  -2.027 0.042670 *  
factor(job)services      -0.487867   0.066121  -7.378 1.60e-13 ***
factor(job)student        0.879372   0.086922  10.117  < 2e-16 ***
factor(job)technician    -0.168579   0.050048  -3.368 0.000756 ***
factor(job)unemployed     0.093801   0.097300   0.964 0.335027    
factor(job)unknown       -0.150583   0.181433  -0.830 0.406558    
previous                  0.879022   0.024831  35.401  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

How to interpret these results? 怎么解释这些结果?

We look at the estimates and the p-value (significance).

Age is not significant; it means whether a client accepts your promotion has little to do with his or her age.

Previous is significant and positive, meaning that getting a deal is easier when you have more previous interaction with the client.

Lastly, which types of jobs are more likely to accept your promotion? Retired and student. On the other hand, blue-collar, services, and entrepreneurs are unlikely to be convinced.

年龄不显著：这表示营销的效果跟消费者的年龄没什么关系。

Previous 显著为正，这表示我们之前跟消费者接触的次数越多，这次营销活动就越容易成功。

最后，哪些行业的人更乐于接受我们的营销？是退休和学生。另一方面，蓝领，服务业者和创业者更不容易被忽悠成功。

# Probit regression

# Probit 回归

# Probit Regression

In logistic regression, we adopt the logistic function to estimate $\Pr\left[Y = 1 \,\middle|\, X\right]$, which satisfies the properties that we listed. However, the logistic function is not the only function that satisfies those properties. Now, we introduce another function that can also make predictions about binary outcomes.

# Probit 回归

在逻辑回归中，我们选择逻辑函数来描述关系$\Pr\left[Y=1 \,\middle|\, X\right]$，而这一函数满足我们之前提出的全部三个要求。但是，逻辑函数并不是唯一满足三个要求的函数。这里，我们再介绍一个新的函数，它同样满足这三个要求。

# Probit Regression

Here, we use the cumulative distribution function of the standard normal distribution. Mathematically, suppose that $v \sim N(0, 1)$ is a standard normal random variable, then we can define the cumulative distribution function $\Phi$ as

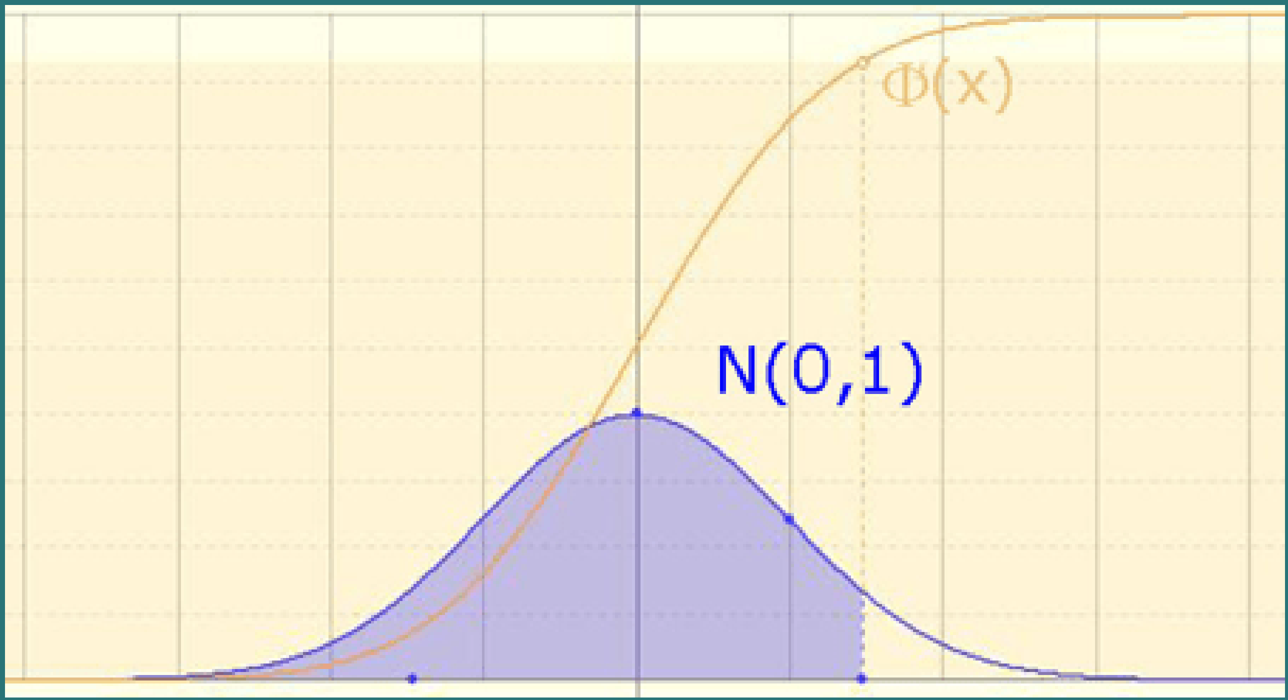
$$\Phi(z) = \Pr[v \leq z].$$

# Probit 回归

这里，我们选择的函数是标准正态分布的累计分布函数。它的数学定义是这样的：假如随机变量 $v \sim N(0,1)$ 服从标准正态分布, 那么它的累计分布函数 $\Phi$ 定义为

$$\Phi(z) = \Pr[v \leq z].$$

# Probit Regression



# Probit Regression

```
library(readr)
mydata <- read.csv("https://ximarketing.github.io/data/banking.csv")
head(mydata)
probit   <- glm(success ~ age + factor(job) + previous, data
                = mydata, family = binomial(link = "probit"))
summary(probit)
```

| | *Dependent variable:* | |
|---|---|---|
| | success | |
| | *logistic* | *probit* |
| | (1) | (2) |
| age | 0.002 | 0.0004 |
| | (0.002) | (0.001) |
| factor(job)blue-collar | -0.626*** | -0.312*** |
| | (0.051) | (0.026) |
| factor(job)entrepreneur | -0.417*** | -0.205*** |
| | (0.100) | (0.050) |
| factor(job)housemaid | -0.258** | -0.138** |
| | (0.110) | (0.057) |
| factor(job)management | -0.174** | -0.090** |
| | (0.068) | (0.035) |
| factor(job)retired | 0.668*** | 0.391*** |
| | (0.078) | (0.044) |
| factor(job)self-employed | -0.189** | -0.093* |
| | (0.093) | (0.048) |
| factor(job)services | -0.488*** | -0.247*** |
| | (0.066) | (0.033) |
| factor(job)student | 0.879*** | 0.497*** |
| | (0.087) | (0.050) |
| factor(job)technician | -0.169*** | -0.092*** |
| | (0.050) | (0.026) |
| factor(job)unemployed | 0.094 | 0.046 |
| | (0.097) | (0.053) |
| factor(job)unknown | -0.151 | -0.084 |
| | (0.181) | (0.095) |
| previous | 0.879*** | 0.494*** |
| | (0.025) | (0.014) |
| Constant | -2.217*** | -1.273*** |
| | (0.075) | (0.039) |
| Observations | 41,188 | 41,188 |
| Log Likelihood | -13,462.880 | -13,460.430 |
| Akaike Inf. Crit. | 26,953.770 | 26,948.860 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

# Logistic vs. Probit

Question: Which one makes more sense?

你觉得哪个结果更合理？

The next question: What should we do when consumers have more than two choices?

当消费者有多于两个选择的时候，我们该怎么办?

More specifically, let us consider the following problem.

Each consumer $i$ has his or her own information, which is measured by the independent variable $X_i$. The dependent variable is a choice made by the consumer, $Y_i \in \{A, B, \dots\}$.

现在，让我们考虑下面的问题:

我们知道每个消费者 $i$ 的个人信息, 而这些信息将成为我们的自变量 $X_i$. 因变量是每个消费者具体的选择, 我们用 $Y_i \in \{A, B, \ldots\}$ 来表示.

```r
install.packages("foreign")
install.packages("nnet")
install.packages("stargazer")

library(foreign)
library(nnet)
library(stargazer)
```

We install and load several packages for multinomial logit regression.

我们需要安装几个包来帮我们实现多项式逻辑回归模型(MNL模型)。

We first load the data from the Internet. 我们读取数据：

```r
mydata <- read.csv("https://ximarketing.github.io/data/bankchoice.csv")
head(mydata, n = 20)
```

Here is the data... 数据是这样的：

| | Choice | Age | Female | Income | Education | Job |
|---|---|---|---|---|---|---|
| 1 | CCB | 62 | 0 | 7 | 2 | Industry |
| 2 | CCB | 34 | 1 | 4 | 5 | Retired |
| 3 | CCB | 68 | 0 | 5 | 2 | Industry |
| 4 | CCB | 60 | 0 | 3 | 2 | Education |
| 5 | ICBC | 18 | 0 | 6 | 2 | Industry |
| 6 | CCB | 18 | 0 | 4 | 3 | Student |
| 7 | CCB | 51 | 0 | 7 | 3 | Education |
| 8 | CCB | 25 | 1 | 3 | 2 | Unemployed |
| 9 | BOC | 42 | 1 | 4 | 4 | Education |
| 10 | CCB | 71 | 1 | 5 | 3 | Service |
| 11 | BOC | 23 | 0 | 4 | 5 | Student |
| 12 | BOC | 30 | 0 | 2 | 5 | Retired |

Here, we want to predict how individuals choose among the four major banks, ICBC, CCB, BOC and ABC.

The independent variables include the followings:

Age: Age of the consumer

Female: Whether or not the consumer is female (Female = 1)

Income: Income level from 1 (lowest) to 7 (highest)

Education: Education level from 1 (lowest) to 5 (highest)

Job: the job of the consumer, including Finance, Service, Education, Industry, Government, Unemployed, Retired

这里，我们想要预测个人在四大银行（工商银行、建设银行、中国银行和农业银行）之间的选择。

自变量包括以下内容：

- 年龄：消费者的年龄
- 性别：消费者是否为女性（女性 = 1）
- 收入：收入水平，从 1（最低）到 7（最高）
- 教育：教育水平，从 1（最低）到 5（最高）
- 职业：消费者的职业，包括金融、服务、教育、工业、政府、失业、退休

We use the multinom function to perform multinomial logit regression: 我们用multinom函数进行MNL模型分析

```r
result <- multinom(formula = Choice ~ Age + Female +
                   Income + Education + factor(Job), data = mydata)
result
```

Oh, the results do not read nicely... 结果看起来不那么友好...

```
Coefficients:
     (Intercept)          Age      Female    Income Education factor(Job)Finance factor(Job)Government
BOC    0.5323798 -0.033020498 -0.08978578 0.6148213 1.2732409          4.298487             0.5149483
CCB    1.1092104 -0.018040309 -0.51117331 0.8868991 0.8061882          4.027794             0.8415596
ICBC  -0.5025889 -0.003286372 -0.37033922 0.5203970 0.2076094          5.033164             2.0033345
     factor(Job)Industry factor(Job)Retired factor(Job)Service factor(Job)Student factor(Job)Unemployed
BOC             2.439599         -2.5722926        -0.04396783         -1.0233035             0.2331643
CCB             2.154324         -3.1883673        -0.33249448         -1.4232623            -0.8674983
ICBC            4.733630         -0.6089712         1.28617434          0.7010564             1.2056360
```

No worries, let's try the stargazer function.

别担心，我们可以用stargazer函数来分析

```r
stargazer(result, type="html",
out="result.html")
getwd()
```

Now, our results are nicely summarized in the table on the right-hand side:

What does it mean?

结果在我们的右表，它说明了什么？

| | *Dependent variable:* | | |
|---|---|---|---|
| | BOC | CCB | ICBC |
| | (1) | (2) | (3) |
| Age | $-0.033^{***}$ | $-0.018^{***}$ | $-0.003$ |
| | (0.002) | (0.002) | (0.002) |
| Female | $-0.090$ | $-0.511^{***}$ | $-0.370^{***}$ |
| | (0.077) | (0.076) | (0.079) |
| Income | $0.615^{***}$ | $0.887^{***}$ | $0.520^{***}$ |
| | (0.028) | (0.028) | (0.029) |
| Education | $1.273^{***}$ | $0.806^{***}$ | $0.208^{***}$ |
| | (0.038) | (0.038) | (0.039) |
| factor(Job)Finance | $4.298^{***}$ | $4.028^{***}$ | $5.033^{***}$ |
| | (0.053) | (0.051) | (0.076) |
| factor(Job)Government | $0.515^{**}$ | $0.842^{***}$ | $2.003^{***}$ |
| | (0.246) | (0.244) | (0.262) |
| factor(Job)Industry | $2.440^{***}$ | $2.154^{***}$ | $4.734^{***}$ |
| | (0.518) | (0.518) | (0.525) |
| factor(Job)Retired | $-2.572^{***}$ | $-3.188^{***}$ | $-0.609^{***}$ |
| | (0.145) | (0.143) | (0.167) |
| factor(Job)Service | $-0.044$ | $-0.332^{*}$ | $1.286^{***}$ |
| | (0.188) | (0.187) | (0.209) |
| factor(Job)Student | $-1.023^{***}$ | $-1.423^{***}$ | $0.701^{***}$ |
| | (0.161) | (0.159) | (0.182) |
| factor(Job)Unemployed | $0.233$ | $-0.867^{***}$ | $1.206^{***}$ |
| | (0.182) | (0.181) | (0.202) |
| Constant | $0.532^{***}$ | $1.109^{***}$ | $-0.503^{**}$ |
| | (0.200) | (0.198) | (0.221) |
| Akaike Inf. Crit. | 81,191.300 | 81,191.300 | 81,191.300 |
| *Note:* | $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ | | |

| | *Dependent variable:* | | |
|---|---|---|---|
| | BOC | CCB | ICBC |
| | (1) | (2) | (3) |
| Age | -0.033*** | -0.018*** | -0.003 |
| | (0.002) | (0.002) | (0.002) |
| Female | -0.090 | -0.511*** | -0.370*** |
| | (0.077) | (0.076) | (0.079) |
| Income | 0.615*** | 0.887*** | 0.520*** |
| | (0.028) | (0.028) | (0.029) |
| Education | 1.273*** | 0.806*** | 0.208*** |
| | (0.038) | (0.038) | (0.039) |
| factor(Job)Finance | 4.298*** | 4.028*** | 5.033*** |
| | (0.053) | (0.051) | (0.076) |
| factor(Job)Government | 0.515** | 0.842*** | 2.003*** |
| | (0.246) | (0.244) | (0.262) |
| factor(Job)Industry | 2.440*** | 2.154*** | 4.734*** |
| | (0.518) | (0.518) | (0.525) |
| factor(Job)Retired | -2.572*** | -3.188*** | -0.609*** |
| | (0.145) | (0.143) | (0.167) |
| factor(Job)Service | -0.044 | -0.332* | 1.286*** |
| | (0.188) | (0.187) | (0.209) |
| factor(Job)Student | -1.023*** | -1.423*** | 0.701*** |
| | (0.161) | (0.159) | (0.182) |
| factor(Job)Unemployed | 0.233 | -0.867*** | 1.206*** |
| | (0.182) | (0.181) | (0.202) |
| Constant | 0.532*** | 1.109*** | -0.503** |
| | (0.200) | (0.198) | (0.221) |
| Akaike Inf. Crit. | 81,191.300 | 81,191.300 | 81,191.300 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | | |

Here, we take ABC as benchmark and compare other routes against it. Alternatively, you can view the parameters for ABC to be equal to 0.

Age: When consumer is older, she/he is less likely to choose BOC/CCB compared with ABC.

Female: Compared with CCB, females are more willing to choose ABC.

| | Dependent variable: | | |
|---|---|---|---|
| | BOC | CCB | ICBC |
| | (1) | (2) | (3) |
| Age | -0.033*** | -0.018*** | -0.003 |
| | (0.002) | (0.002) | (0.002) |
| Female | -0.090 | -0.511*** | -0.370*** |
| | (0.077) | (0.076) | (0.079) |
| Income | 0.615*** | 0.887*** | 0.520*** |
| | (0.028) | (0.028) | (0.029) |
| Education | 1.273*** | 0.806*** | 0.208*** |
| | (0.038) | (0.038) | (0.039) |
| factor(Job)Finance | 4.298*** | 4.028*** | 5.033*** |
| | (0.053) | (0.051) | (0.076) |
| factor(Job)Government | 0.515** | 0.842*** | 2.003*** |
| | (0.246) | (0.244) | (0.262) |
| factor(Job)Industry | 2.440*** | 2.154*** | 4.734*** |
| | (0.518) | (0.518) | (0.525) |
| factor(Job)Retired | -2.572*** | -3.188*** | -0.609*** |
| | (0.145) | (0.143) | (0.167) |
| factor(Job)Service | -0.044 | -0.332* | 1.286*** |
| | (0.188) | (0.187) | (0.209) |
| factor(Job)Student | -1.023*** | -1.423*** | 0.701*** |
| | (0.161) | (0.159) | (0.182) |
| factor(Job)Unemployed | 0.233 | -0.867*** | 1.206*** |
| | (0.182) | (0.181) | (0.202) |
| Constant | 0.532*** | 1.109*** | -0.503** |
| | (0.200) | (0.198) | (0.221) |
| Akaike Inf. Crit. | 81,191.300 | 81,191.300 | 81,191.300 |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

在这里，我们将农业银行（ABC）作为基准，并将其他银行与其进行比较。或者，您可以将ABC的参数视为等于0。

- 年龄：当消费者年龄较大时，与农业银行相比，她/他选择中国银行或建设银行的可能性较小。
- 性别：与建设银行相比，女性更倾向于选择农业银行。

The complete code is here:

```r
library(foreign)
library(nnet)
library(stargazer)
mydata <- read.csv("https://ximarketing.github.io/data/bankchoice.csv")
head(mydata)
result <- multinom(formula = Choice ~ Age + Female +
                     Income + Education + factor(Job),
                     data = mydata)
result
stargazer(result, type="html", out="result.html")
```

We first load the data from the Internet. 我们读取数据：

```
mydata <-
read.csv("https://ximarketing.github.io/data/multinomial_route_choice.csv")
head(mydata)
```

Here is the data... 数据是这样的：

| | Choice | Flow | Distance | Seat_belt | Passengers | Age | Male | Income | Fuel_efficiency |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Arterial | 460 | 48 | 0 | 0 | 2 | 0 | 1 | 28 |
| 2 | Rural | 440 | 44 | 0 | 0 | 2 | 0 | 1 | 28 |
| 3 | Freeway | 130 | 61 | 0 | 0 | 2 | 0 | 1 | 28 |
| 4 | Arterial | 595 | 59 | 1 | 0 | 2 | 1 | 2 | 27 |
| 5 | Rural | 515 | 70 | 1 | 0 | 2 | 1 | 2 | 27 |
| 6 | Freeway | 340 | 87 | 1 | 0 | 2 | 1 | 2 | 27 |

Here, we want to predict how individuals choose the route when driving. The dependent variable is the chosen route, which can be arterial, rural, and freeway.

The independent variables include the followings:

Flow: A measure of traffic flow (how busy the traffic is).

Distance: The distance of the planned trip.

Seat_belt: whether the driver wears seat belt.

Passengers: Number of passengers carried.

Age: Age group of the driver.

Male: Whether the driver is male or not.

Income: Income level of the driver.

Fuel_efficiency: Fuel efficiency level of the vehicle.

这里，我们希望分析司机是如何选择道路的。每个司机的选项是我们的因变量，包括主干道(arterial)，乡间公路(rural)和高速路(freeway)三种选择。

自变量包括以下内容：

流量 Flow: 当前道路的繁忙情况.

里程 Distance: 需要驾驶路段的里程

安全带 Seat_belt: 司机有没有系安全带

乘客 Passengers: 车上有多少乘客.

年龄 Age: 司机的年龄.

男性 Male: 司机是否是男性.

收入 Income: 司机的收入水平.

燃油效率 Fuel_efficiency: 车辆的燃油效率.

We use the multinom function to perform multinomial logit regression: 我们用multinom函数进行MNL模型分析

```r
result <- multinom(formula = Choice ~ Flow + Distance +
                     Seat_belt + Passengers + Age + Male +
                     Income + Fuel_efficiency, data = mydata)
result
```

Oh, the results do not read nicely... 结果看起来不那么友好...

```
Coefficients:
        (Intercept)         Flow   Distance  Seat_belt Passengers         Age        Male
Freeway   13.673284 -0.049143703  0.1362782 -0.8924558  0.4775758  0.17728498  0.06331663
Rural      7.558223 -0.008436186 -0.0455514 -0.3451560  0.1436887 -0.06181751 -0.04244764
            Income Fuel_efficiency
Freeway -0.5430466     -0.06321059
Rural    0.1319585     -0.01778424
```

No worries, let's try the stargazer function.

别担心，我们可以用stargazer函数来分析

```
stargazer(result, type="html", out="result.html")
```

| | *Dependent variable:* | |
|---|---|---|
| | Freeway | Rural |
| | (1) | (2) |
| Flow | -0.049*** | -0.008*** |
| | (0.006) | (0.001) |
| Distance | 0.136*** | -0.046*** |
| | (0.031) | (0.014) |
| Seat_belt | -0.892 | -0.345 |
| | (0.663) | (0.319) |
| Passengers | 0.478 | 0.144 |
| | (0.454) | (0.275) |
| Age | 0.177 | -0.062 |
| | (0.310) | (0.157) |
| Male | 0.063 | -0.042 |
| | (0.638) | (0.302) |
| Income | -0.543 | 0.132 |
| | (0.379) | (0.144) |
| Fuel_efficiency | -0.063 | -0.018 |
| | (0.068) | (0.038) |
| Constant | 13.673*** | 7.558*** |
| | (0.158) | (1.390) |
| Akaike Inf. Crit. | 419.424 | 419.424 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

Now, our results are nicely summarized in the table on the right-hand side:

What does it mean?

结果在我们的右表，它说明了什么？

| | *Dependent variable:* | |
|---|---|---|
| | Freeway | Rural |
| | (1) | (2) |
| Flow | -0.049*** | -0.008*** |
| | (0.006) | (0.001) |
| Distance | 0.136*** | -0.046*** |
| | (0.031) | (0.014) |
| Seat_belt | -0.892 | -0.345 |
| | (0.663) | (0.319) |
| Passengers | 0.478 | 0.144 |
| | (0.454) | (0.275) |
| Age | 0.177 | -0.062 |
| | (0.310) | (0.157) |
| Male | 0.063 | -0.042 |
| | (0.638) | (0.302) |
| Income | -0.543 | 0.132 |
| | (0.379) | (0.144) |
| Fuel_efficiency | -0.063 | -0.018 |
| | (0.068) | (0.038) |
| Constant | 13.673*** | 7.558*** |
| | (0.158) | (1.390) |
| Akaike Inf. Crit. | 419.424 | 419.424 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

Here, we take arterial as the benchmark and compare other routes against it. Alternatively, you can view the parameters for arterial to be equal to zero.

Flow: When there is a high flow, drivers are very less likely to choose freeway, and a bit less likely to choose rural compared with arterial.

Distance: When distance is long, drivers are more likely to choose freeway and less likely to choose rural route...

| | Dependent variable: | |
|---|---|---|
| | Freeway | Rural |
| | (1) | (2) |
| Flow | -0.049*** | -0.008*** |
| | (0.006) | (0.001) |
| Distance | 0.136*** | -0.046*** |
| | (0.031) | (0.014) |
| Seat_belt | -0.892 | -0.345 |
| | (0.663) | (0.319) |
| Passengers | 0.478 | 0.144 |
| | (0.454) | (0.275) |
| Age | 0.177 | -0.062 |
| | (0.310) | (0.157) |
| Male | 0.063 | -0.042 |
| | (0.638) | (0.302) |
| Income | -0.543 | 0.132 |
| | (0.379) | (0.144) |
| Fuel_efficiency | -0.063 | -0.018 |
| | (0.068) | (0.038) |
| Constant | 13.673*** | 7.558*** |
| | (0.158) | (1.390) |
| Akaike Inf. Crit. | 419.424 | 419.424 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

这里，我们将主干道为基准，将其它道路与主干道进行比较，并将对应的系数设为0.

Flow: 当车流量很大的时候，司机最不愿意驾驶高速路，最愿意驾驶主干道。

Distance: 当里程长的时候，司机最愿意驾驶高速路，最不愿意驾驶乡村公路。

The complete code is here:

```r
library(foreign)
library(nnet)
library(stargazer)
mydata <-
read.csv("https://ximarketing.github.io/data/multinomial_route_choice.csv"
)
head(mydata)
result <- multinom(formula = Choice ~ Flow + Distance +
                     Seat_belt + Passengers + Age + Male +
                     Income + Fuel_efficiency, data = mydata)
result
stargazer(result, type="html", out="result.html")
```

Back to the Question:

回到之前的问题:

How do machines recognize hand-written digits?

机器是如何识别手写数字的?

## Back to the Question:

How do machines recognize hand-written digits?

Absolutely, there are many sophisticated algorithms for handwriting recognition such as convolutional neural networks. But in the early stage, scientists just use the multinomial logit model to perform the task.

Input: Handwriting in pixels.

Output: $Y_i \in \{0, 1, \ldots, 9\}$

Absolutely, there are many sophisticated algorithms for handwriting recognition such as convolutional neural networks. But in the early stage, scientists just use the multinomial logit model to perform the task.

有很多好的算法可以帮助我们识别手写数字，比如卷积神经网络。但在早期阶段，计算机学家们使用的还是MNL模型。

Input 输入: Handwriting in pixels 一个一个像素的图像.

Output 输出: $Y_i \in \{0, 1, \ldots, 9\}$ 十个数字之一

# Conditional Logit Model

## 条件Logit模型

In multinomial logit model, a person chooses among a few alternatives. The decision hinges on the decision maker's personal features, not the features of the alternatives. In our previous example, the route decision hinges on features such as distance, age, which are constant across all alternatives.

In conditional logit model, a person chooses among a few alternatives. The decision hinges on the alternatives' features, not the feature of the individuals.

在 multinomial logit model 模型, 一个人从几个选项中做出选择。 这个选择取决于这个人的个人特征而不是这些选项的特征，例如，选择基于这个人的年龄，性别等个人特征。

在 conditional logit model 模型, 一个人从几个选项中做出选择。这个选择取决于这个选项的特征而不是这个人的特征。例如，选择基于这个选项的价格，质量，颜色等。

Example:

Consumers choose among three computers, A, B, and C.

1. If the choices are based on consumers' age, gender, education etc, then we use the multinomial logit model.
2. If the choices are based on the price, quality of the computers, then we use the conditional logit model.

举例:

消费者从三个电脑品牌中选择一个, A, B, 和 C.

1. 如果选择是基于消费者的年龄，性别，职业等信息，那么我们选择的模型是 multinomial logit model.
2. 如果选择是基于每个电脑的价格，质量，服务等，那么我们选择的模型是 conditional logit model.

```r
install.packages("survival")
library(survival)
library(stargazer)
mydata = read.csv("https://ximarketing.github.io/data/conjoint.csv")
head(mydata)
```

| id | interest | downpayment | rebate | speed | choice |
|---|---|---|---|---|---|
| 1 | 3.75 | 40 | 0.15 | 0.5 | 0 |
| 1 | 4.00 | 25 | 0.15 | 1.0 | 0 |
| 1 | 3.75 | 25 | 0.00 | 1.0 | 1 |
| 2 | 3.50 | 20 | 0.10 | 0.5 | 1 |
| 2 | 3.75 | 25 | 0.30 | 1.5 | 0 |
| 2 | 3.75 | 20 | 0.30 | 1.0 | 0 |

| id | interest | downpayment | rebate | speed | choice |
|---|---|---|---|---|---|
| 1 | 3.75 | 40 | 0.15 | 0.5 | 0 |
| 1 | 4.00 | 25 | 0.15 | 1.0 | 0 |
| 1 | 3.75 | 25 | 0.00 | 1.0 | 1 |
| 2 | 3.50 | 20 | 0.10 | 0.5 | 1 |
| 2 | 3.75 | 25 | 0.30 | 1.5 | 0 |
| 2 | 3.75 | 20 | 0.30 | 1.0 | 0 |

Consumer 1 (id = 1) chooses between three offers:

| Interest Rate | Down Payment | Rebate | Speed (Months) | Choice |
|---|---|---|---|---|
| 3.75% | 40% | 0.15% | 0.5 | NO |
| 4.00% | 25% | 0.15% | 1.0 | NO |
| 3.75% | 25% | 0% | 1.0 | YES |

| id | interest | downpayment | rebate | speed | choice |
| --- | --- | --- | --- | --- | --- |
| 1 | 3.75 | 40 | 0.15 | 0.5 | 0 |
| 1 | 4.00 | 25 | 0.15 | 1.0 | 0 |
| 1 | 3.75 | 25 | 0.00 | 1.0 | 1 |
| 2 | 3.50 | 20 | 0.10 | 0.5 | 1 |
| 2 | 3.75 | 25 | 0.30 | 1.5 | 0 |
| 2 | 3.75 | 20 | 0.30 | 1.0 | 0 |

用户1 (id = 1) 从下面三个按揭计划中做出选择：

| 按揭利率 | 首付 | 回赠 | 审批时间（月度） | 选择 |
| --- | --- | --- | --- | --- |
| 3.75% | 40% | 0.15% | 0.5 | NO |
| 4.00% | 25% | 0.15% | 1.0 | NO |
| 3.75% | 25% | 0% | 1.0 | YES |

```r
result<-clogit(choice ~ interest +  downpayment + rebate
               + speed + strata(id), data=mydata)
summary(result)
```

```
                 coef exp(coef)  se(coef)       z Pr(>|z|)    
interest    -1.185055  0.305729  0.097289 -12.181  < 2e-16 ***
downpayment -0.052922  0.948454  0.002336 -22.652  < 2e-16 ***
rebate       0.177522  1.194254  0.149303   1.189  0.23444    
speed       -0.117274  0.889341  0.039587  -2.962  0.00305 ** 
```

```r
stargazer(result, type="html", out="result.html")
```

| | *Dependent variable:* |
|---|---|
| | choice |
| interest | -1.185*** |
| | (0.097) |
| downpayment | -0.053*** |
| | (0.002) |
| rebate | 0.178 |
| | (0.149) |
| speed | -0.117*** |
| | (0.040) |
| Observations | 18,000 |
| $R^2$ | 0.039 |
| Max. Possible $R^2$ | 0.519 |
| Log Likelihood | -6,237.584 |
| Wald Test | 643.940*** (df = 4) |
| LR Test | 708.180*** (df = 4) |
| Score (Logrank) Test | 679.185*** (df = 4) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

When interest rate increases, the user is less likely to choose the plan; when down payment increases, the user is less likely to choose the plan; when approval takes longer time, the user is less likely to choose the plan.

当利率增加，首付变高，或者审批时间变长的情况下，用户选择按揭的概率降低。

| | coef | exp(coef) | se(coef) | z | Pr(>\|z\|) | |
|---|---|---|---|---|---|---|
| interest | -1.185055 | 0.305729 | 0.097289 | -12.181 | < 2e-16 | *** |
| downpayment | -0.052922 | 0.948454 | 0.002336 | -22.652 | < 2e-16 | *** |
| rebate | 0.177522 | 1.194254 | 0.149303 | 1.189 | 0.23444 | |
| speed | -0.117274 | 0.889341 | 0.039587 | -2.962 | 0.00305 | ** |

The coefficient for interest is -1.185 and the coefficient for speed is -0.1172. Because 0.1172/1.185=0.098, it suggests that a 1 month increase in approval time is equivalent to a 0.098% decrease in interest rate.

利率的系数是-1.185而审批时间系数是-0.1172. 因为0.1172/1.185=0.098，这说明审批时间增加一个月的代价相当于利率上涨0.098%元带来的代价。换句话说，一个月的审批时间对于消费者的价值是0.098%的利率。

The complete code is here:

```r
library(survival)
library(stargazer)
mydata = read.csv("https://ximarketing.github.io/data/conjoint.csv")
head(mydata)
result<-clogit(choice ~ interest +  downpayment + rebate
                + speed + strata(id), data=mydata)
summary(result)
stargazer(result, type="html", out="result.html")
```

# Predicting Market Share
# 预测市场占有率

Suppose that there are two plans available in the market:

| Interest | Down Payment | Rebate | Speed |
| --- | --- | --- | --- |
| 3.85% | 30% | 0.1% | 1 month |
| 4.25% | 25% | 0.25% | 0.5 months |

We can use our regression results to predict their market share, following the formula of conditional logit.

现在市场上有两种按揭产品，我们想预测它们的市场占有率

| 利率 | 首付 | 现金回赠 | 审批速度 |
|---|---|---|---|
| 3.85% | 30% | 0.1% | 1 个月 |
| 4.25% | 25% | 0.25% | 0.5 个月 |

```r
library(survival)
library(stargazer)
mydata = read.csv("https://ximarketing.github.io/data/conjoint.csv")
head(mydata)
result<-clogit(choice ~ interest +  downpayment +   rebate
               + speed + strata(id), data=mydata)
coef_interest <- coef(result)["interest"]
coef_downpayment <- coef(result)["downpayment"]
coef_rebate <- coef(result)["rebate"]
coef_speed <- coef(result)["speed"]

interest1 <- 3.85; downpayment1 <- 30; rebate1 <- 0.1; speed1 <- 1
interest2 <- 4.25; downpayment2 <- 25; rebate2 <- 0.25; speed2 <- 0.5

d1 <- exp(interest1 * coef_interest + downpayment1 * coef_downpayment +
            rebate1 * coef_rebate + speed1 * coef_speed)
d2 <- exp(interest2 * coef_interest + downpayment2 * coef_downpayment +
            rebate2 * coef_rebate + speed2 * coef_speed)

s1 <- d1/(d1+d2)
s2 <- d2/(d1+d2)
print(c(s1, s2))
```

## 讨论问题:

假设你要分析消费者如何选择银行(例如建设银行，工商银行，中国银行等)开立账户。但一个消费者可以选择多于一个银行开户。我们应该如何分析这个问题?

课后讨论问题：

如果我们知道消费者几个选项中最喜欢哪个选项，我们可以用离散选择模型来分析数据。但假如数据显示的是消费者最讨厌哪个选项而不是最喜欢哪个选项，我们还可以用离散选择模型分析这些数据吗？

请点击这个链接或者扫描二维码回答问题