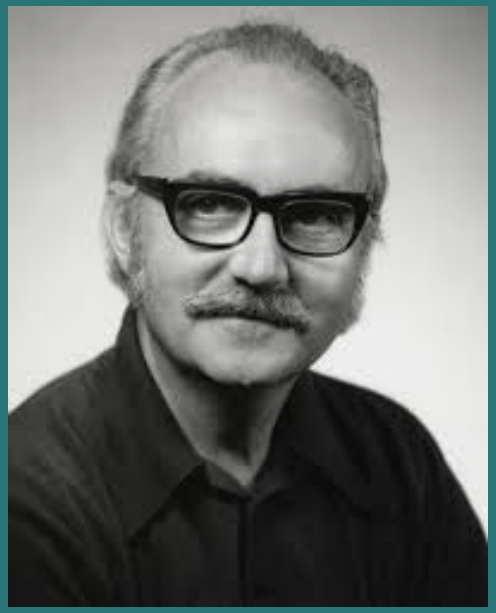
Discrete Choice

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



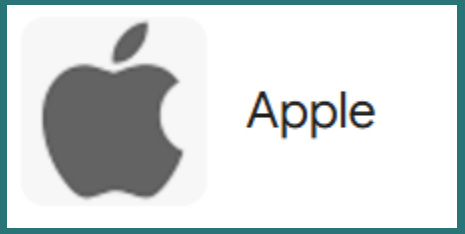
Question:

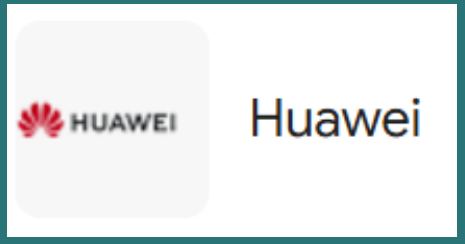
How do machines recognize hand-written digits?

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?





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

香港大学校长开什么牌子的车?

港聞 / 01偵查

港大校長座駕換車　張翔指定寶馬i7　205萬元豁免招標惹爭議

撰文：勞顯亮

出版：2023-09-12 07:00　更新：2023-09-12 13:37

102

基于环保考虑，大学的校园设施管理部门建议以纯电动车代替燃油车。在比较了市面上多款电动车及混合动力型车的资料后，该部门与校长室挑选了以上两款，安排校长及两个部门的同事于同一天参与试车，并归纳了各人的反馈意见。校园设施管理部门同时比较了两款车的性能表现、车厢容量、外观、价格，亦考虑了大学的需要。

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。

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 的期关系？这里，我们的回归方程是这样的:

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

Issues with linear regression

Suppose that your regression result is:

$$Y_i = 0.4 + 0.1 \times Age_i + 0.2 \times Female_i$$

Suppose that a person's age is 25 and gender is male, you predict that his $Y_i = 0.65$, that is, the person is likely to buy from you.

线性回归的问题

假设你的回归结果是这样的：

$$Y_i = 0.4 + 0.1 \times Age_i + 0.2 \times Female_i$$

假设一个人的年龄是 25，性别是男性，根据回归公式我们的预测是 $Y_i = 0.65$, 换句话说，这个消费者有可能消费我们的产品。

Issues with linear regression

Suppose that your regression result is:

$$Y_i = 0.4 + 0.1 \times Age_i + 0.3 \times Female_i$$

Suppose that another person's age is 40 and gender is female, you predict that her $Y_i = 1.1$.

How would you interpret this result? Will she apply for your credit card 1.1 times? It does not make any sense!

线性回归的问题

假设你的回归结果是这样的：

$$Y_i = 0.4 + 0.1 \times Age_i + 0.2 \times Female_i$$

假设一个人的年龄是 40，性别是女性，根据回归公式我们的预测是 $Y_i = 0.65$。

你该如何解释这个结果？她一定会消费我们的产品？她会消费1.1次？

What should we do?

Instead of predicting the value of Y_i directly, we can predict the probability that Y_i is equal to 1, i.e., we want to predict $\Pr[Y_i = 1]$.

How to do that? We want to find out a function f such that

$$\Pr[Y_i = 1] \approx f(X_i)$$

Next, we will look for such a function f.

我们该怎么办?

与其直接预测 Y_i 的具体数值, 我们尝试预测 Y_i 等于 1 的概率, 即我们想预测 $\Pr[Y_i = 1]$.

我们需要找到一个这样的函数 f :

$$\Pr[Y_i = 1] \approx f(X_i)$$

现在，我们一起找找这样的函数 f.

What should we do?

How to do that? We want to find out a function f such that

$$\Pr[Y_i = 1] \approx f(X_i)$$

Here, we need to impose some restrictions on the function f:

1. $f(X) \geq 0$ for all X: probabilities are nonnegative.
2. $f(X) \leq 1$ for all X: probabilities are no more than 100%.
3. $f(X)$ is either increasing or decreasing with X.

我们该怎么办?

上一步我们说到我们希望找到这样的函数 f

$$\Pr[Y_i = 1] \approx f(X_i)$$

但并不是所有的函数 f 都可以。我们希望的函数需要满足一些条件:

1. $f(X) \geq 0$ for all X: 概率不能是负数。
2. $f(X) \leq 1$ for all X: 概率不能超过 100%.
3. $f(X)$ 对于 X 是递增或者递减函数.

What should we do?

$$\Pr[Y_i = 1] \approx f(X_i)$$

Here, we need to impose some restrictions on function f:

1. $f(X) \geq 0$ for all X: probabilities are nonnegative.
2. $f(X) \leq 1$ for all X: probabilities are no more than 100%.
3. $f(X)$ is either increasing or decreasing with X.

Can you propose such a function f? Any ideas?

我们该怎么办?

上一步我们说到我们希望找到这样的函数 f

$$\Pr[Y_i = 1] \approx f(X_i)$$

但并不是所有的函数 f 都可以。我们希望的函数需要满足一些条件:

1. $f(X) \geq 0$ for all X: 概率不能是负数。
2. $f(X) \leq 1$ for all X: 概率不能超过 100%.
3. $f(X)$ 对于 X 是递增或者递减函数.

你能找出这样的一个函数吗? 试试看!

$$f(X) = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$$

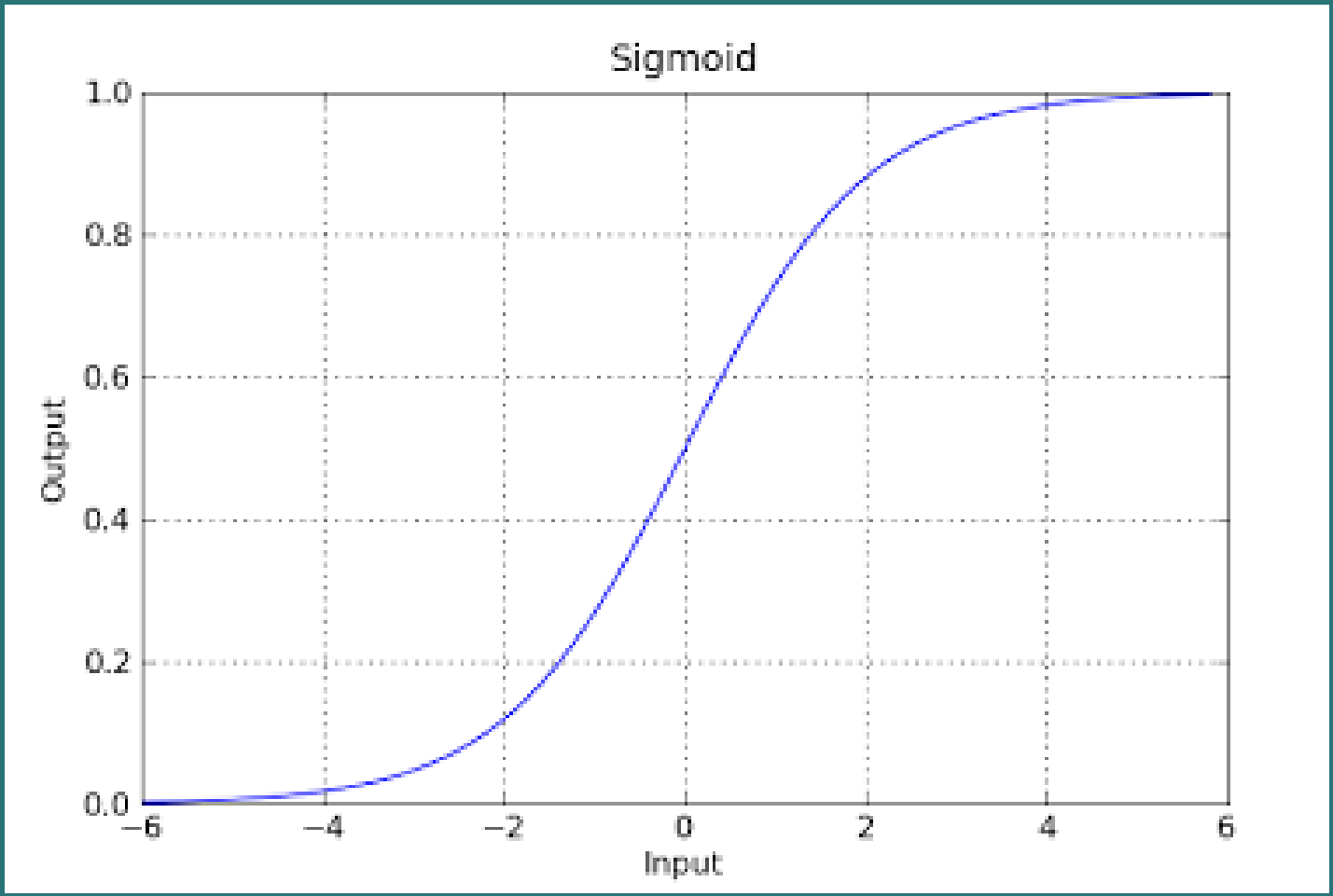
Note: $\exp(x) = e^x$ is the exponential function.

When $\beta > 0$, $f(X)$ increases with X; when $\beta < 0$, $f(X)$ decreases with X.

$$f(X) = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$$

注释: $\exp(x) = e^x$ 是指数函数.

当$\beta > 0$ 时, $f(X)$ 是 X 的增函数; 当$\beta < 0$ 时, $f(X)$ 是 X 的减函数.



The logistic function 逻辑函数

Our task

We already know the values X_i and Y_i for each individual i. We would like to find the values of α and β to approximate the relationship between X_i and Y_i:

$$\Pr(Y_i = 1) \approx \frac{\exp(\alpha + \beta X_i)}{1 + \exp(\alpha + \beta X_i)}$$

This is done via maximum likelihood estimation.

我们的任务

现在，对于每一个消费者 i， 我们都知道对应的 X_i 和 Y_i 的数值. 我们希望进一步找到 α 和 β 的数值来近似 X_i 和 Y_i 之间的关系:

$$\Pr(Y_i = 1) \approx \frac{\exp(\alpha + \beta X_i)}{1 + \exp(\alpha + \beta X_i)}$$

这里，我们用一种最大似然估计的方法。

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

我们看看下面的数据：

```
1 library(readr)
2 mydata <- read.csv("https://ximarketing.github.io/data/banking.csv")
3 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 代表之间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 $\Pr[Y = 1 \mid X]$, 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 Φ as

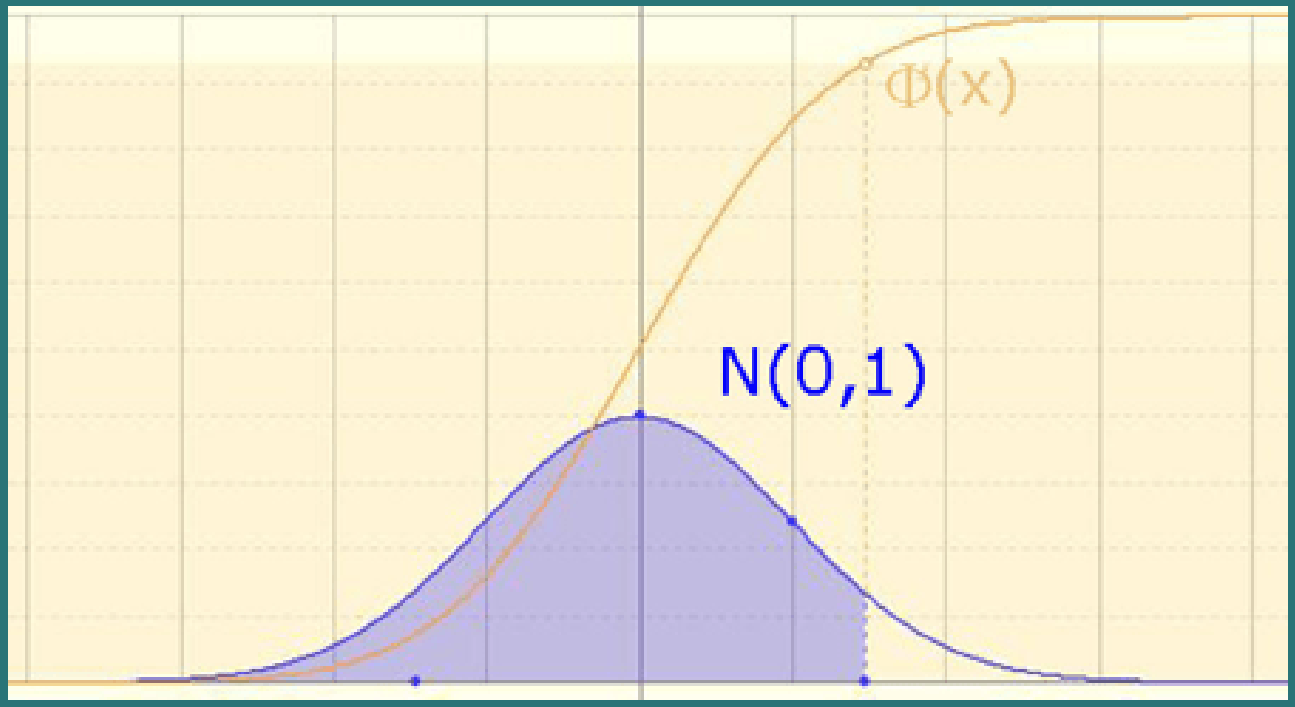
$$\Phi(z) = \Pr[v \leq z].$$

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 Φ 定义为

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

Probit Regression



Probit Regression

```r
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?

Logistics vs. Probit

They are similar models that yield similar (though not identical) inferences.

- Logistic regression is more popular in healthcare.
- Probit regression is more popular in political science.

But in most situations, it does not matter which method you choose to go with. Working with either will be fine.

Logistics vs. Probit

这两个模型会给你类似，但稍有不同的结果

- 逻辑回归常常被用在医疗分析上；
- Probit 回归在政治分析领域更加流行。

但在大多数情况下，你可以随便选择其中的一个模型，而具体的选择往往是一种习惯而不是基于任何理论。

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 $Y_i \in \{A, B, \dots\}$ 来表示.

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, $Y_i \in \{A, B, \dots\}$.

Idea: Instead of predicting Y_i directly, we predict the probability $\Pr[Y_i = A], \Pr[Y_i = B], \dots$

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 $Y_i \in \{A, B, \dots\}$ 来表示.

思路：与其直接预测 Y_i，我们不如预测这些概率： $\Pr[Y_i = A], \Pr[Y_i = B], \dots$

Suppose that consumers have three choices, A, B, C.

Now, given X_i, we would like to come up with three functions $f_A(X_i)$, $f_B(X_i)$ and $f_C(X_i)$, such that

$$\Pr[Y_i = A] \approx f_A(X_i),$$
$$\Pr[Y_i = B] \approx f_B(X_i),$$
$$\Pr[Y_i = C] \approx f_C(X_i).$$

假设每个消费者有三个选择, 我们记为A, B, C.

现在，给定消费者的信息 X_i, 我们想找到三个函数 $f_A(X_i), f_B(X_i)$ 和 $f_C(X_i)$, 使得

$$\begin{aligned}
\Pr[Y_i = A] &\approx f_A(X_i), \\
\Pr[Y_i = B] &\approx f_B(X_i), \\
\Pr[Y_i = C] &\approx f_C(X_i).
\end{aligned}$$

As before, we place a few restrictions on these functions:

1. The probabilities must be nonnegative, i.e., $f_j(X_i) \geq 0$
2. Probabilities cannot exceed 1, i.e., $f_j(X_i) \leq 1$
3. Probabilities are monotone with X_i
4. Now, we have a new constraint: all the probabilities must add up to 100%, i.e.,

$$f_A(X_i) + f_B(X_i) + f_C(X_i) = 1.$$

Any ideas for the functions?

但是我们不能随便找三个函数。他们必须满足一定条件：

1. 所有的概率都不能是负数，即 $f_j(X_i) \geq 0$
2. 所有的概率都不能超过1，即 $f_j(X_i) \leq 1$
3. 所有的概率都是 X_i 的单调函数
4. 最后，我们还有一个额外的要求：消费者必须在这三个选项中选择一个，即

$$f_A(X_i) + f_B(X_i) + f_C(X_i) = 1.$$

你能找到这样的一组函数吗？

$$f_A(X_i) = \frac{\exp(\alpha_A + \beta_A X_i)}{\exp(\alpha_A + \beta_A X_i) + \exp(\alpha_B + \beta_B X_i) + \exp(\alpha_C + \beta_C X_i)}$$

$$f_B(X_i) = \frac{\exp(\alpha_B + \beta_B X_i)}{\exp(\alpha_A + \beta_A X_i) + \exp(\alpha_B + \beta_B X_i) + \exp(\alpha_C + \beta_C X_i)}$$

$$f_C(X_i) = \frac{\exp(\alpha_C + \beta_C X_i)}{\exp(\alpha_A + \beta_A X_i) + \exp(\alpha_B + \beta_B X_i) + \exp(\alpha_C + \beta_C X_i)}$$

They satisfy all the constraints! 它们符合左右的条件！

We need to estimate the values of α's and β's.

```r
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：

```
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

```
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")
```

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	

	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。但在早起阶段，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
library(survival)
library(stargazer)
mydata = read.csv("https://ximarketing.github.io/data/conjoint.csv")
head(mydata)
```

	id	price	storage	ram	cpu	choice
1	1	400	512	4	3.6	1
2	1	400	256	8	2.8	0
3	1	300	128	4	5.0	0
4	2	500	256	2	5.0	0
5	2	300	512	8	2.8	0
6	2	400	512	4	3.6	1

	id	price	storage	ram	cpu	choice
1	1	400	512	4	3.6	1
2	1	400	256	8	2.8	0
3	1	300	128	4	5.0	0
4	2	500	256	2	5.0	0
5	2	300	512	8	2.8	0
6	2	400	512	4	3.6	1

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

1. Price = 400, Storage = 512 GB, RAM = 4 GB, CPU = 3.6 GHz
2. Price = 400, Storage = 256 GB, RAM = 8 GB, CPU = 2.8 GHz
3. Price = 300, Storage = 128 GB, RAM = 4 GB, CPU = 5.0 GHz

And this consumer chooses the first computer (choice = 1)

	id	price	storage	ram	cpu	choice
1	1	400	512	4	3.6	1
2	1	400	256	8	2.8	0
3	1	300	128	4	5.0	0
4	2	500	256	2	5.0	0
5	2	300	512	8	2.8	0
6	2	400	512	4	3.6	1

消费者 1 (id = 1) 从三个电脑中选择一个:

1. Price = 400, Storage = 512 GB, RAM = 4 GB, CPU = 3.6 GHz
2. Price = 400, Storage = 256 GB, RAM = 8 GB, CPU = 2.8 GHz
3. Price = 300, Storage = 128 GB, RAM = 4 GB, CPU = 5.0 GHz

这个消费者选择了第一个电脑 (choice = 1)

```
result<-clogit(choice ~ price + cpu +
                 ram + storage + strata(id), data=mydata)
summary(result)
```

```
            coef  exp(coef)   se(coef)      z Pr(>|z|)
price   -0.0038226  0.9961847  0.0004281 -8.929   <2e-16 ***
cpu      0.4974295  1.6444886  0.0378409 13.145   <2e-16 ***
ram      0.1486753  1.1602962  0.0070257 21.162   <2e-16 ***
storage  0.0055173  1.0055325  0.0002284 24.159   <2e-16 ***
```

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

	Dependent variable:
	choice
price	-0.003***
	(0.0004)
cpu	0.366***
	(0.027)
ram	0.138***
	(0.007)
storage	0.005***
	(0.0002)
Observations	6,000
R^2	0.198
Max. Possible R^2	0.519
Log Likelihood	-1,537.106
Wald Test	790.650*** (df = 4)
LR Test	1,320.236*** (df = 4)
Score (Logrank) Test	1,155.273*** (df = 4)
Note:	*p<0.1; **p<0.05; ***p<0.01

When price increases, the computer is less likely to be chosen; when CPU, RAM or Storage increases, the computer is more likely to be chosen.

当价格变高，一个电脑更难被选择；当CPU或者RAM或者硬盘变大，这个电脑更容易被选择。

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
price	-0.0038226	0.9961847	0.0004281	-8.929	<2e-16	***
cpu	0.4974295	1.6444886	0.0378409	13.145	<2e-16	***
ram	0.1486753	1.1602962	0.0070257	21.162	<2e-16	***
storage	0.0055173	1.0055325	0.0002284	24.159	<2e-16	***

The coefficient for price is -0.0038 and the coefficient for RAM is 0.1486. Because $0.1486/0.0038 = 38.8$, it suggests that a 1GB increase in RAM is equivalent to a \$38.8 decrease in price. Or put differently, 1 GB RAM is worth \$38.8 to an average consumer.

价格的系数是-0.0038而内存的系数是0.1486. 因为$0.1486/0.0038 = 38.8$，这说明内存增加1GB的优势相当于价格降低38.8元带来的优势。换句话说，1GB的内存对于消费者的价值是38.8元。

The complete code is here:

```
library(survival)
library(stargazer)
mydata = read.csv("https://ximarketing.github.io/data/conjoint.csv")
head(mydata)
result<-clogit(choice ~ price + cpu +
                 ram + storage + 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 PCs available in the market:

- (1) Price = 400, CPU = 3.6 GHz, RAM = 4 GB, Storage = 512 GB
- (2) Price = 280, CPU = 3.2 GHz, RAM = 4 GB, Storage = 256 GB

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

假设有以下两款电脑进入市场，我们分析他们的占有率

- (1) Price = 400, CPU = 3.6 GHz, RAM = 4 GB, Storage = 512 GB
- (2) Price = 280, CPU = 3.2 GHz, RAM = 4 GB, Storage = 256 GB

```r
library(survival)
library(stargazer)
mydata = read.csv("https://ximarketing.github.io/data/conjoint.csv")
head(mydata)
result<-clogit(choice ~ price + cpu +
                 ram + storage + strata(id), data=mydata)

coef_price <- coef(result)["price"]
coef_cpu <- coef(result)["cpu"]
coef_ram <- coef(result)["ram"]
coef_storage <- coef(result)["storage"]

price1 <- 400; cpu1 <- 3.6; ram1 <- 4; storage1 <- 512
price2 <- 280; cpu2 <- 3.2; ram2 <- 4; storage2 <- 256

d1 <- exp(price1 * coef_price + cpu1 * coef_cpu + ram1 * coef_ram +
storage1 * coef_storage)
d2 <- exp(price2 * coef_price + cpu2 * coef_cpu + ram2 * coef_ram +
storage2 * coef_storage)

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

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