

Data Platform

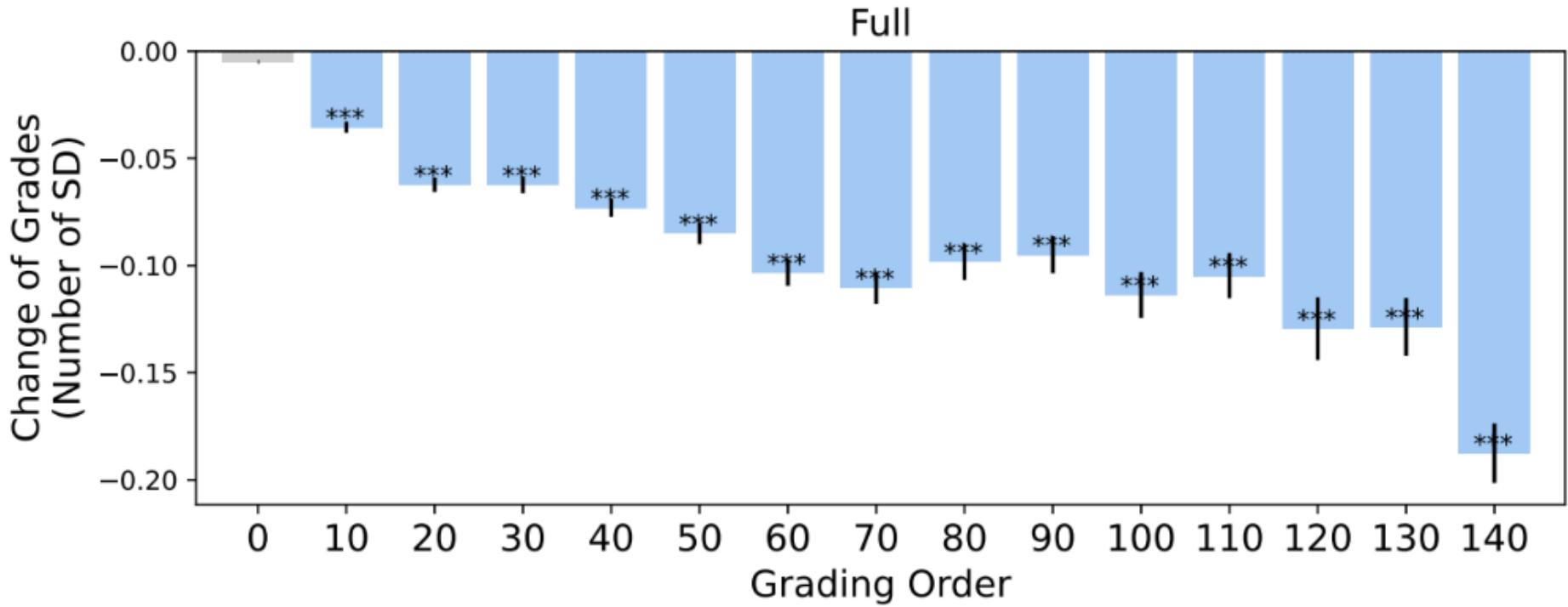
We will be using a data platform for your data project

- Please sign up an account at <http://acrc.hku.hk/> using your HKU email address.
- Please add your coursepack using the link <https://www.acrc.hku.hk/enrol/1000012200>

Order of Presentation

Should I present early?

Figure 3 Assignment Grades and Grading Order: Full Sample



If you are graded earlier, then you are likely to receive a higher grade! But why is this so?

Possible Reasons:

- Fatigue (“审美疲劳”)
- Higher expectations?
- Can detect mistakes more easily? ...

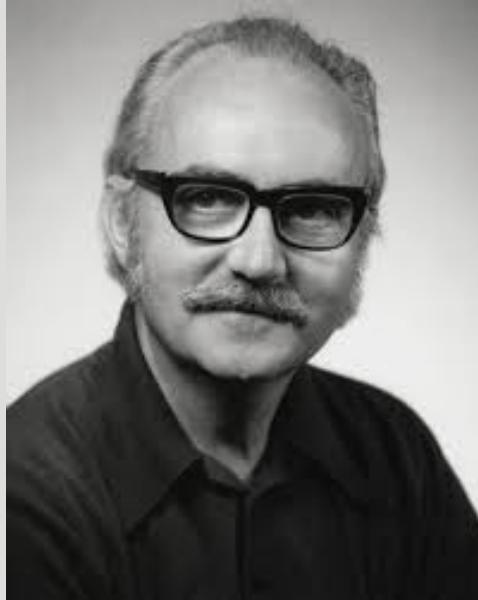
Are there any groups willing to present early?

Word Cloud

Discrete Choice Models

“ All models are wrong, some are useful.

--- George Box



No model can capture the full complexity of the real world,
yet a simplified model can still give good predictions or
insight within a relevant range.

How does AI recognize hand-written digits?



What brand is my smartphone?



oppo

I am using a Xiaomi phone.

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



中午12:21

HD R 4G 5G 67%



不想吃萨莉亚了

关注



You can find his car at the
Knowles Building carpark.

hku我是老大
不是)
#一键查询留学生精神状态

12-04 广东

共 157 条评论



爱评论的人运气都不差

@



-士

这是专门拉翔的

2天前

广东

回复



住在哪边

会说多说 简直是天才



2天前 中国香港 回复



23

说点什么...



263



35



157



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 developed a model to understand our choices. His 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.”

Modelling Consumer Choice

Human beings are constantly making choices—from deciding whom to marry to choosing a bottle of milk.

While individuals make decisions in their own ways, we want to understand how consumers make those choices.

Imaging that you are a bank manager...



You want to understand how consumers choose between different credit cards. In this way, you can understand who your potential clients are, and can target on these consumers better.

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, AMEX, ... or none of the above. This is your dependent variable.

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

What would you do?

Let us start with something simpler.

Now, you want to predict whether or not a consumer applies for your own 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 .

What should you do?

Our simplified 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 run the following linear model:*

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

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.

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!

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 look for such a function 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 .

What should we do?

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

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

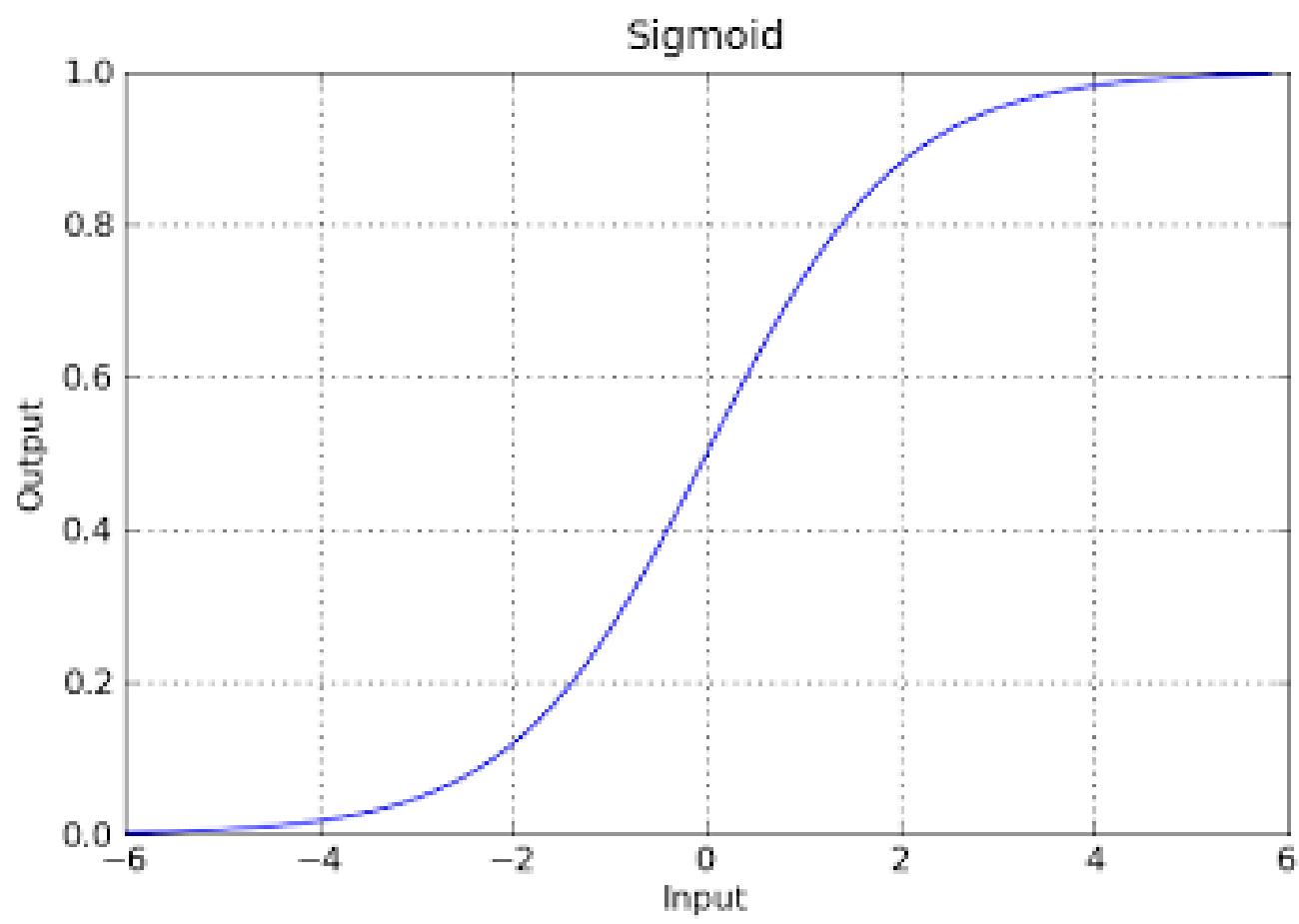
1. $f(X) \geq 0$ for all X : probabilities are nonnegative.
2. $f(X) \leq 1$ for all X : probabilities are no more than 100%.
3. $f(X)$ is either increasing or decreasing with X .

Can you propose such a function f ? Any ideas?

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

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 .



The logistic function

A **video** explaining logistic function

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. Check [here](#) if you want to know more details.

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



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

The data reads as follows:

| | 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 (real) data is about the outcome of a marketing campaign in a Portuguese bank 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 interactions with the client. m deposit

```
# run a logistic regression, dv is success, iv is age, previous, data is mydata  
model <- glm(success ~ age + previous, data = mydata, family = binomial)
```



```
1 result <- glm(success ~ age + factor(job) + previous, data  
2                      = mydata, family = "binomial")  
3 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.

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?

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 interactions 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.

Using GPT 5:

Run a logistic regression. The dependent variable is success. The independent variables are age, previous, and job. Show the results.



banking.csv
电子表格

★ Key Findings

1. Significant predictors

- **previous** (number of previous contacts)
 - $\beta = 0.8790$, $p < 0.001$
 - Strongest predictor: more previous contacts → higher probability of success.
- **Job categories with *higher* probability of success (vs. baseline job category):**
 - **student**, $\beta = 0.8794$, $p < 0.001$
 - **retired**, $\beta = 0.6676$, $p < 0.001$
- **Job categories with *lower* probability of success:**
 - blue-collar, services, technician, self-employed, housemaid, etc.
 - All statistically significant except unemployed and unknown.

2. Not significant

- **age** ($\beta = 0.0019$, $p = 0.287$)
Age does **not** significantly predict success here.
- Some job categories (unemployed, unknown) are also not significant.

Probit regression

Probit Regression

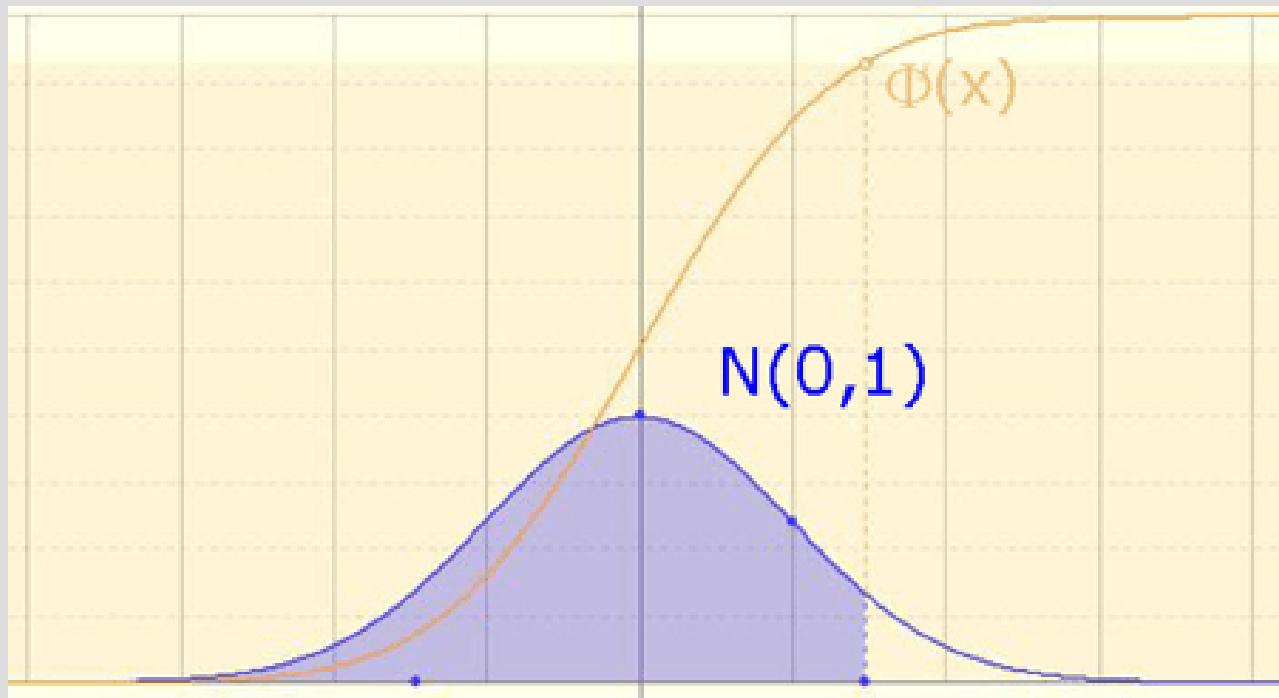
In logistic regression, we adopt the logistic function to estimate $\Pr [Y = 1 \mid X]$, which satisfies the properties that we listed. However, the logistic function is not the only function that satisfies those properties. Now, we introduce another function that can also make predictions about binary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

Probit 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 Regression





Probit Regression

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

Using GPT 5:

Run a Probit regression. The dependent variable is success. The independent variables are age, previous, and job. Show the results.



banking.csv

电子表格

less

 复制代码

| | coef | std_err | z | P> z | [0.025 | 0.975] |
|-------------------|---------|---------|---------|--------|---------|---------|
| const | -1.2732 | 0.0390 | -32.958 | 0.0000 | -1.3490 | -1.1974 |
| age | 0.0004 | 0.0009 | 0.463 | 0.6430 | -0.0012 | 0.0020 |
| previous | 0.4937 | 0.0139 | 35.465 | 0.0000 | 0.4664 | 0.5209 |
| job_blue-collar | -0.3122 | 0.0257 | -12.142 | 0.0000 | -0.3627 | -0.2617 |
| job_entrepreneur | -0.0364 | 0.0090 | -4.091 | 0.0000 | -0.0539 | -0.0189 |
| job_housemaid | -0.0246 | 0.0101 | -2.432 | 0.0150 | -0.0444 | -0.0048 |
| job_management | -0.0159 | 0.0063 | -2.528 | 0.0114 | -0.0283 | -0.0035 |
| job_retired | 0.0694 | 0.0078 | 8.925 | 0.0000 | 0.0541 | 0.0847 |
| job_self-employed | -0.0165 | 0.0085 | -1.936 | 0.0528 | -0.0331 | 0.0001 |
| job_services | -0.0438 | 0.0059 | -7.414 | 0.0000 | -0.0555 | -0.0321 |
| job_student | 0.0881 | 0.0088 | 9.959 | 0.0000 | 0.0710 | 0.1052 |
| job_technician | -0.0164 | 0.0047 | -3.510 | 0.0004 | -0.0256 | -0.0072 |
| job_unemployed | 0.0081 | 0.0093 | 0.868 | 0.3857 | -0.0101 | 0.0263 |
| job_unknown | -0.0149 | 0.0170 | -0.879 | 0.3792 | -0.0479 | 0.0181 |

| | <i>Dependent variable:</i> | |
|--------------------------|----------------------------|----------------------|
| | success | |
| | <i>logistic</i> | <i>probit</i> |
| | (1) | (2) |
| age | 0.002 (0.002) | 0.0004 (0.001) |
| factor(job)blue-collar | -0.626*** (0.051) | -0.312*** (0.026) |
| factor(job)entrepreneur | -0.417*** (0.100) | -0.205*** (0.050) |
| factor(job)housemaid | -0.258** (0.110) | -0.138** (0.057) |
| factor(job)management | -0.174** (0.068) | -0.090** (0.035) |
| factor(job)retired | 0.668*** (0.078) | 0.391*** (0.044) |
| factor(job)self-employed | -0.189** (0.093) | -0.093* (0.048) |
| factor(job)services | -0.488*** (0.066) | -0.247*** (0.033) |
| factor(job)student | 0.879*** (0.087) | 0.497*** (0.050) |
| factor(job)technician | -0.169*** (0.050) | -0.092*** (0.026) |
| factor(job)unemployed | 0.094 (0.097) | 0.046 (0.053) |
| factor(job)unknown | -0.151 (0.181) | -0.084 (0.095) |
| previous | 0.879*** (0.025) | 0.494*** (0.014) |
| Constant | -2.217*** (0.075) | -1.273*** (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?

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.

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

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

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 $\Pr[Y_i = A], \Pr[Y_i = B], \dots$

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

$$\Pr[Y_i = A] \approx f_A(X_i),$$

$$\Pr[Y_i = B] \approx f_B(X_i),$$

$$\Pr[Y_i = C] \approx f_C(X_i).$$

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?

$$f_A(X_i) = \frac{\exp(\alpha_A + \beta_AX_i)}{\exp(\alpha_A + \beta_AX_i) + \exp(\alpha_B + \beta_BX_i) + \exp(\alpha_C + \beta_CX_i)}$$

$$f_B(X_i) = \frac{\exp(\alpha_B + \beta_BX_i)}{\exp(\alpha_A + \beta_AX_i) + \exp(\alpha_B + \beta_BX_i) + \exp(\alpha_C + \beta_CX_i)}$$

$$f_C(X_i) = \frac{\exp(\alpha_C + \beta_CX_i)}{\exp(\alpha_A + \beta_AX_i) + \exp(\alpha_B + \beta_BX_i) + \exp(\alpha_C + \beta_CX_i)}$$

They satisfy all the constraints!

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



```
1 library(foreign)
2 library(nnet)
3 library(stargazer)
```

We install and load several packages for multinomial logit regression.

THE MEASUREMENT OF URBAN TRAVEL DEMAND

Daniel McFADDEN*

Department of Economics, University of California, Berkeley, U.S.A.

Part of McFadden's data on route choice



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

Here is the data...

| | 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.

We use the multinom function to perform multinomial logit regression:



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

Oh, the results do not read nicely...

| 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.



```
1 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?

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

Note: *p<0.1; **p<0.05; ***p<0.01

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

Using GPT 5:

I want to run a multinomial logit model. The DV is Choice. All other variables are independent variables. Give me the results.



multinomial_route_choice.csv

电子表格

The complete code is here:



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

How does AI recognize hand-written digits?



How does AI 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: $Y_i \in \{0, 1, \dots, 9\}$

Conditional Logit Model

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

Example:

Consumers choose among three PC brands, 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.

Mortgage Data



```
1 library(survival)
2 library(stargazer)
3 mydata = read.csv("https://ximarketing.github.io/data/conjoint.csv")
4 head(mydata)
```

| | id | interest | downpayment | rebate | speed | choice | |
|---|-----------|-----------------|--------------------|---------------|--------------|---------------|---|
| 1 | 1 | 3.75 | | 40 | 0.15 | 0.5 | 0 |
| 2 | 1 | 4.00 | | 25 | 0.15 | 1.0 | 0 |
| 3 | 1 | 3.75 | | 25 | 0.00 | 1.0 | 1 |
| 4 | 2 | 3.50 | | 20 | 0.10 | 0.5 | 1 |
| 5 | 2 | 3.75 | | 25 | 0.30 | 1.5 | 0 |
| 6 | 2 | 3.75 | | 20 | 0.30 | 1.0 | 0 |

| | id | interest | downpayment | rebate | speed | choice |
|---|-----------|-----------------|--------------------|---------------|--------------|---------------|
| 1 | 1 | 3.75 | 40 | 0.15 | 0.5 | 0 |
| 2 | 1 | 4.00 | 25 | 0.15 | 1.0 | 0 |
| 3 | 1 | 3.75 | 25 | 0.00 | 1.0 | 1 |
| 4 | 2 | 3.50 | 20 | 0.10 | 0.5 | 1 |
| 5 | 2 | 3.75 | 25 | 0.30 | 1.5 | 0 |
| 6 | 2 | 3.75 | 20 | 0.30 | 1.0 | 0 |

Consumer 1 (id = 1) chooses between three mortgage plans:

| Interest Rate | Downpayment | Cash Rebate | Processing Time |
|----------------------|--------------------|--------------------|------------------------|
| 3.75% | 40% | 0.15% | 0.5 month |
| 4.00% | 25% | 0.15% | 1.0 month |
| 3.75% | 25% | 0% | 1.0 month |

This consumer ended up choosing the last plan (choice = 1).



```
1 result = clogit(choice ~ interest + downpayment + rebate  
2                      + speed + strata(id), data=mydata)  
3 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 | ** | | | | | |
| --- | | | | | | | | | | | |
| Signif. codes: | 0 | *** | 0.001 | ** | 0.01 | * | 0.05 | . | 0.1 | ' | 1 |



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

| <i>Dependent variable:</i> | |
|------------------------------|----------------------|
| | 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 mortgage is less likely to be chosen;
- When downpayment increases, the mortgage is less likely to be chosen;
- The effect of rebate is not significant;
- When processing time (speed) becomes longer, the mortgage is less likely to be chosen.

Using GPT 5:

I want to run a conditional logit model. id denotes the individual who is making the choice, choice is the person's choice (1 = choose, 0 = not choose), other variables are IVs. Show me the result.



conjoint.csv

电子表格

Conditional Logit Model Regression Results

| | | | |
|-----------------|------------------|-------------------|-------|
| Dep. Variable: | choice | No. Observations: | 18000 |
| Model: | ConditionalLogit | No. groups: | 6000 |
| Log-Likelihood: | -6237.6 | Min group size: | 3 |
| Method: | BFGS | Max group size: | 3 |
| Date: | Thu, 04 Dec 2025 | Mean group size: | 3.0 |
| Time: | 01:24:50 | | |

| coef | std err | z | P> z | [0.025 | 0.975] |
|------|---------|---|------|--------|--------|
| | | ↓ | | | |

Interpretation

- **interest** (-1.185, p<0.001): Higher interest rate makes a product *much less likely to be chosen*.
- **downpayment** (-0.053, p<0.001): Higher required downpayment reduces choice probability.
- **rebate** (0.178, ns): Positive sign, but *not statistically significant*.
- **speed** (-0.117, p=0.003): Higher speed (depending on what it represents in your design) decreases likelihood of being chosen.

Other Models

There are also many other models beyond ones we discuss in class:

- If your dependent variable is the number of units (e.g., X bottles of milk; Y individuals...), you can use **Poisson regression**.
- If your dependent variable is censored (e.g., you only observe those whose income is greater than 100K), you can use **Tobit model**.

<https://www.youtube.com/embed/i8tjLQUPc8Y?enablejsapi=1>