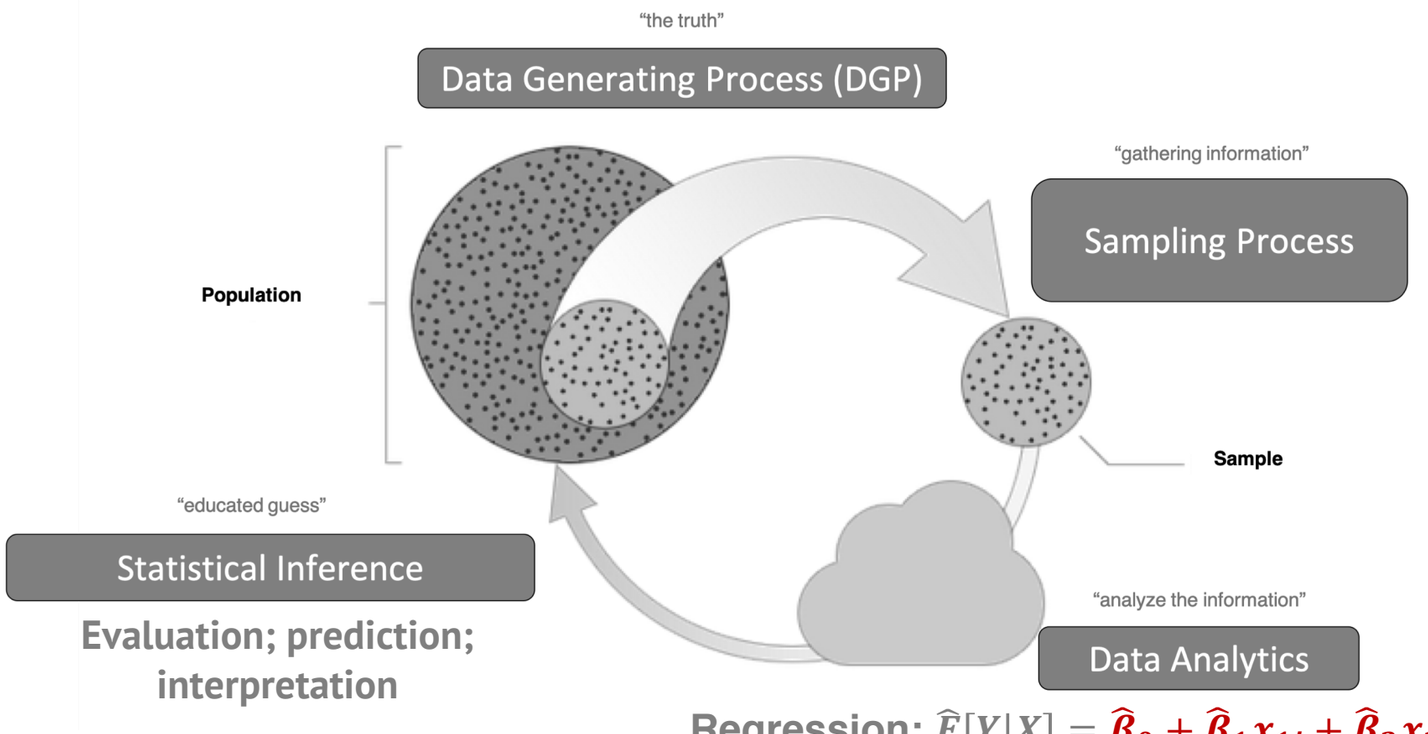
URP 6931. Introduction to Urban Analytics

Lecture 05: Logistic regression

Instructor: Shenhao Wang
Assistant Professor, Director of Urban AI Lab
Department of Urban and Regional Planning
University of Florida

Review lecture 04

Regression: recovering $E[Y|X]$



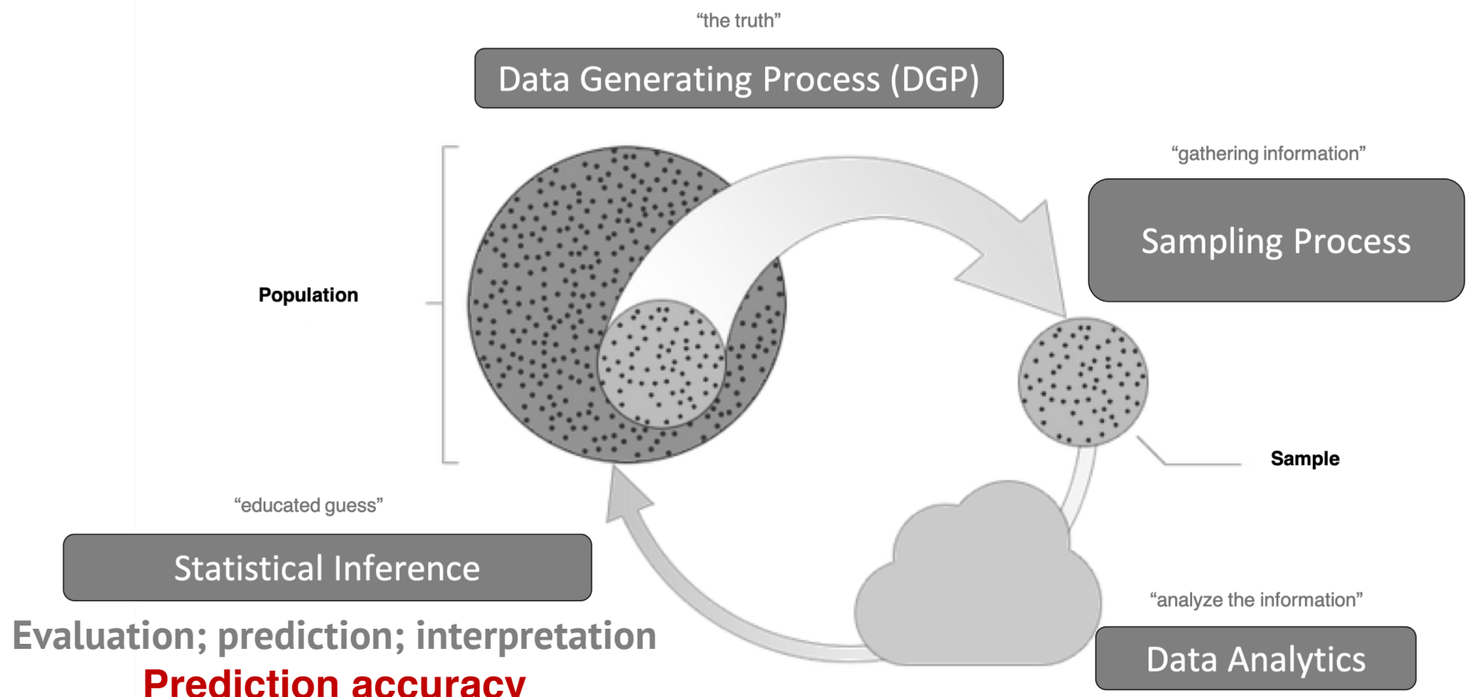
i.i.d.
Y: continuous

Evaluation; prediction; interpretation

Regression: $\widehat{E}[Y|X] = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_k x_{ki}$
Estimation/Training/Learning: OLS

Preview lecture 05

Regression: recovering $E[Y|X] = P(Y = 1|X)$



i.i.d.
Y: {0, 1}

Evaluation; prediction; interpretation
Prediction accuracy

Regression: $\hat{E}[Y|X] = \sigma(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki})$ – link function
Estimation/Training/Learning: Maximum Likelihood Estimation

Lecture 05. Logistic regression

1	2	3	4	5
Logistic regression as Bernoulli distribution	Understanding logistic regression: link function and the meaning of β	Maximum likelihood estimation (MLE)	Theoretical assumptions, statistical properties, and practice (old/new)	Prediction, evaluation, and back to the general diagram

Part 1. Logistic regression as Bernoulli distribution

It is still the **conditional** mean function

Power and critiques in multivariate linear regressions

Power

1. Linear-in-parameter models are powerful.
2. Simple interpretation: partial effect.
3. Statistical test.

Invalid (or not so valid) critiques

1. Linear - too simple
2. About causality.
3. The multicollinearity between independent variables

Valid critiques

1. A1 – omitted variable bias (about data)
2. A2 – function misspecification problem (about model)
3. A3-6. Other potential problems in sampling, etc.

However, one more valid critique in linear regressions is:

About the value range of the dependent variable

- When $y_i = \beta_0 + \beta_1 x_{1i} + \ ... \beta_k x_{ki} + u$, then y_i can take any value from $-\infty$ to $+\infty$. And y_i has to be a cardinal value.

However, many urban applications require us to use nominal, ordinal values, or cardinal values with limited value range as the dependent variables.

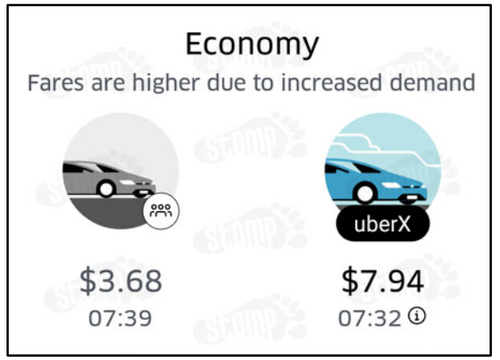
- Travel mode choice: non-auto vs. automobile {0, 1}
- Low vs. high property values {0, 1}
- Automobile counts in each household {0, 1, 2, 3, 4, 5+}
- Ratio of public transit usage in a census tract [0, 1]

Today, we focus on the binary discrete variable {0, 1} as the dependent variable.

A large number of examples using discrete dependent variables in urban applications

Travel mode choice

Uber single passenger vs. ride-sharing

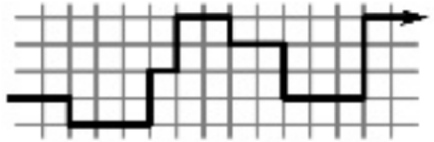


Residential & job location choice

Auto ownership

Trip purpose

Route choice



etc.

When Y is a binary discrete variable, we still model the conditional mean function

Y is a binary random variable (e.g. auto vs. others)

$$Y_i \in \{0, 1\}$$

Then $Y_i \sim B(p)$. It is a Bernoulli distribution similar to a coin flip.

$$E[Y_i] = P(Y_i = 1) = p$$

Conditioning on X:

$$\boldsymbol{E[Y_i|X_i] = P(Y_i = 1|X_i) = f(X_i)}$$

Notes

1. This is the essence about logistic regression: $E[Y_i|X_i] = \boldsymbol{P(Y_i = 1|X_i) = f(X_i)}$.
2. Here we adopt a "coin flip" approach to analyze the discrete outputs.
3. When we see a data set {0, 1, 0, 1, 1, 1, 0, 1, 0, 1}, then Bernoulli distribution is the default distributional assumption (lec02). The only difference lies in the unconditional vs. conditional analysis.
4. When the dependent variable is binary discrete, then modeling the conditional mean function $E[Y_i|X_i]$ is the same as modeling the conditional probability function $P(Y_i|X_i)$.

Doubts about the “coin flip” analogy

Why do I feel counter-intuitive about this coin flip approach (logistic regression)?

Use an example. Travel mode choice in the morning commuting: non-automobile: 0; automobile: 1.

Q1. I commute to work every day using automobiles. It is totally deterministic and I have never flipped a coin.

Q2. When I am in Boston, I commute to work via public transit. When I am in Miami, I commute to work via driving. Both are not only deterministic but also context-dependent. Why is the decision process a coin flip?

Q3. I drive to school during the weekdays and stay at home during the weekends. It is my routine, and how could this consistent routine be modeled as a coin flip?

Q4. In fact, I don’t have any routine and just follow my random emotion to ride whatever to school. How could any analytical approach capture my precious emotions?

These are very common critiques in urban planning against any analytical approach.

How do you think about these critiques?

Answers to the doubts

Urban analytics formalize and generalize traditional planning approaches and your intuition

A: The power lies in the **conditioning**.

Formalize the intuition: [Math notations for the four questions]

Generalize the intuition, e.g. conditioning on what; distributional assumptions

Part 2. Understanding logistic regression

The key is this function: $E[y_i|x_i] = \boldsymbol{P}(\boldsymbol{y_i = 1}|\boldsymbol{x_i}) = \boldsymbol{\sigma}(\boldsymbol{\beta' x_i})$ – the logistic link function

Running example: y_i represents high vs. low property values. x_{1i} is still household income.

Change the notation first

$$\beta_0 + \beta_1 x_{1i} + \ \ldots + \beta_k x_{ki} = \boldsymbol{\beta' x_i}$$

then, $\boldsymbol{P(y_i = 1|x_i) = \sigma(\beta' x_i) = \sigma(\beta_0 + \beta_1 x_{1i} + \ \ldots + \beta_k x_{ki})}$

But what is exactly $\boldsymbol{\sigma(\beta' x_i)}$?

$\sigma(\beta' x_i)$: the link function $\sigma()$

- $\sigma(\beta' x_i)$ is used to generalize linear regressions by connecting the linear model $\beta' x_i$ to the dependent variable through a link function $\sigma(\beta' x_i)$.

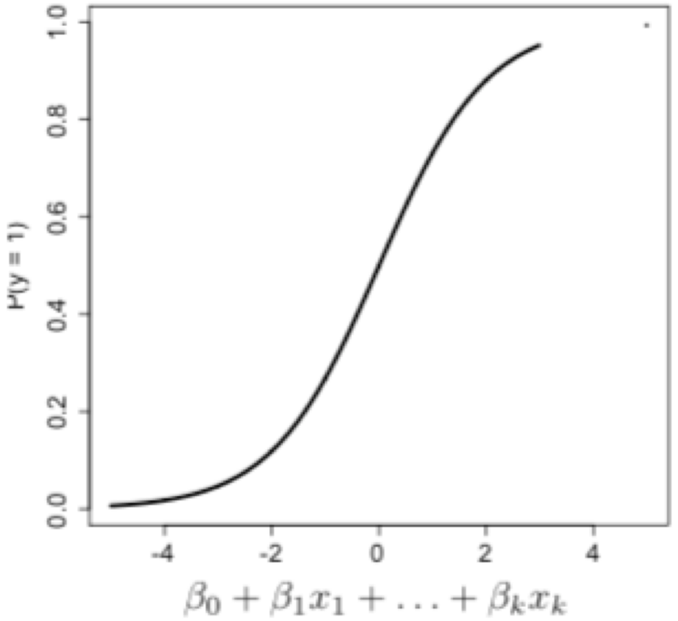
$$P(y_i = 1|x_i) = \sigma(\beta_0 + \beta_1 x_{1i} + \ \ldots + \beta_k x_{ki})$$

- Our goal is to find a function to map any value to a valid probability between zero and one. (Remember the output from a simple linear regression is invalid)
- One such σ function is called logistic link function, which is also the name of the regression. With its help, we can model the probability in the Bernoulli distribution.

$$\sigma(\beta_0 + \beta_1 x_{1i} + \ \ldots \beta_k x_{ki}) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_{1i} + \ \ldots \beta_k x_{ki})]}$$

Understanding the logistic link function

$$P(y_i = 1|x_i) = \sigma(\boldsymbol{\beta}'\boldsymbol{x_i}) = \frac{1}{1+ e^{-(\boldsymbol{\beta}'\boldsymbol{x_i})}}$$



Notes

- Inputs: $\boldsymbol{\beta}'\boldsymbol{x_i}$; outputs: $\sigma(\boldsymbol{\beta}'\boldsymbol{x_i})$.
- X-axis: $\boldsymbol{\beta}'\boldsymbol{x_i}$; Y-axis: $\sigma(\boldsymbol{\beta}'\boldsymbol{x_i})$.
- If we only look at this logistic link function, its form is $\sigma(t) = \frac{1}{1+ e^{-t}}$
- Range on the x-axis: any real value
- Range on the y-axis: value only between zero and one.
- Why? (1) $\boldsymbol{\beta}'\boldsymbol{x_i} = \mathbf{0}$; (2) $\boldsymbol{\beta}'\boldsymbol{x_i} > \mathbf{0}$; (3) $\boldsymbol{\beta}'\boldsymbol{x_i} < \mathbf{0}$.
- This specific logistic form is a bit arbitrary; however, it serves our goal by generating a valid probability value as the output.

Logistic regression in the example

Example

- y_i: low vs. high property values {0, 1}
- x_{1i}: household income; other x_{ji}: control variables

Model specification

$$P(y_i = 1|x_i) = \sigma(\boldsymbol{\beta' x_i}) = \frac{1}{1+ e^{-(\boldsymbol{\beta' x_i})}} = \frac{1}{1+ e^{-(\beta_0+\beta_1 x_{1i}+ \ldots \beta_k x_{ki})}}$$

Meaning of the model outputs

- $P(y_i = 1|x_i) = \sigma(\boldsymbol{\beta' x_i})$: probability of purchasing a property with high value (choosing 1)

$$\sigma(\boldsymbol{\beta' x_i}) = \frac{1}{1 + e^{-(\boldsymbol{\beta' x_i})}}$$

- $P(y_i = 0|x_i) = 1 - \sigma(\boldsymbol{\beta' x_i})$: probability of purchasing a property with low value (choosing 0)

$$1 - \sigma(\boldsymbol{\beta' x_i}) = \frac{e^{-(\boldsymbol{\beta' x_i})}}{1 + e^{-(\boldsymbol{\beta' x_i})}}$$

How to interpret β in logistic regression?

Review. In linear regression: β represents the partial effect ($\beta_j = \partial E[y_i|x_i]/\partial x_{ij}$). One unit increase in x_{ij} is associated with β_j unit change in y_i.

But in logistic regression, this partial effect is not very straightforward.

Math derivation.

$$\frac{\partial E[y_i|x_i]}{\partial x_{ij}} = \frac{\partial}{\partial x_{ij}}\left[\frac{1}{1+\exp(-\boldsymbol{\beta'x_i})}\right]$$

After a few steps (Calculus I), you can see the result:

$$\frac{\partial E[y_i|x_i]}{\partial x_{ij}} = \beta_j * \frac{1}{1+\exp(-\boldsymbol{\beta'x_i})} * \frac{\exp(-\boldsymbol{\beta'x_i})}{1+\exp(-\boldsymbol{\beta'x_i})} = \beta_j * \sigma(\boldsymbol{\beta'x_i}) * (1 - \sigma(\boldsymbol{\beta'x_i}))$$

Since $E[y_i|x_i] = P(y_i|x_i) = \sigma(\boldsymbol{\beta'x_i})$, we can rewrite the formula as:

$$\frac{\partial \sigma(\boldsymbol{\beta'x_i})}{\partial x_{ij}} = \beta_j \sigma(\boldsymbol{\beta'x_i})(1 - \sigma(\boldsymbol{\beta'x_i}))$$

How to interpret β in logistic regression?

Since $E[y_i|x_i] = P(y_i = 1|x_i) = \sigma(\boldsymbol{\beta}'\boldsymbol{x_i})$, we can rewrite the formula as:

$$\frac{\partial \sigma(\boldsymbol{\beta}'\boldsymbol{x_i})}{\partial x_{i1}} = \beta_1 \, \sigma(\boldsymbol{\beta}'\boldsymbol{x_i})(1 - \sigma(\boldsymbol{\beta}'\boldsymbol{x_i}))$$

Dissecting this formula

- Intuition, σ measure the probability of purchasing high-value houses, as opposed to low-value ones. x_{i1} is household income.
- Then the partial effect $\partial\sigma/\partial x_{ij}$ measures how the probability of purchasing high-value houses is associated with one unit change in household income.
- This partial effect is nonlinearly related to the coefficient β_j because of the scaling factor $\sigma\,(1-\sigma)$
- However, the partial effect $\partial\sigma/\partial x_{i1}$ and β_1 have the same sign – both positive or both negative.
- As a result, it might not be a best practice to directly interpret β in logistic regression, but people use β to evaluate the sign of the correlation and statistical significance in practice.

The link function $\sigma(t)$ for Generalized Linear Regression

Output is always positive

See whiteboard

Output is always negative

See whiteboard

Output is small nominal values {0, 1, 2, 3, 4}

See whiteboard

Output is always between zero and one.

See whiteboard

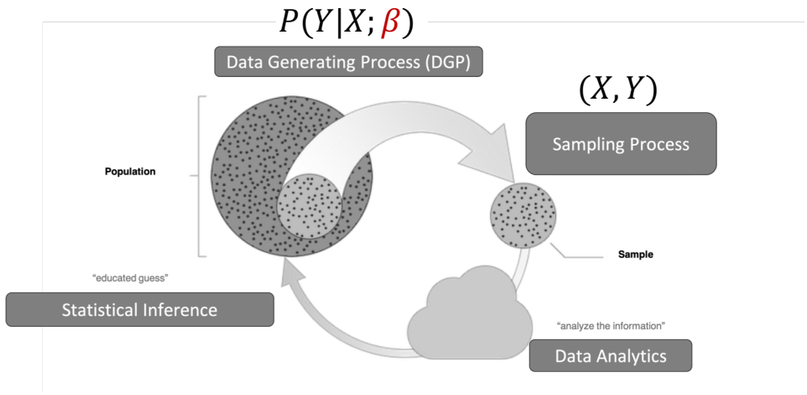
Part 3. Maximum likelihood estimation (MLE)

- Q: How to estimate the β coefficients from the logistic regression $P(y_i = 1|x_i) = \frac{1}{1+ e^{-(\boldsymbol{\beta' x_i})}}$?
- A: We use MLE, and MLE is much more widely used than OLS. You will see it repeatedly in all the analytical paradigms, e.g. machine learning & deep learning

What is the Intuition of MLE?

We choose the parameters that are mostly likely to generate the observed data points.

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}}\, \boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\beta})$$



Intuition about Maximum Likelihood Estimation (MLE)

Example 1. Bernoulli Distribution.

You observe the following data points: {0, 1, 0, 1, 1, 1, 0, 1, 0, 1}

Question: Which distribution is most likely to represent the true data generating process?

1. Bernoulli(0.2)
2. Bernoulli(0.4)
3. Bernoulli(0.6)

An intuitive step:

1. Specify a parameter p for the Bernoulli distribution Bernoulli(p)
2. Then comparing the probability of seeing the data set (counting the order).
    1. $0.2^6 * 0.8^4 = 2.62 * 10^{-5}$
    2. $0.4^6 * 0.6^4 = 53.08 * 10^{-5}$
    3. $0.6^6 * 0.4^4 = 119.44 * 10^{-5}$

It is most likely that we observe such a data set from a Bernoulli distribution with $p = 0.6$

Intuition about Maximum Likelihood Estimation (MLE)

Example 1. Gaussian Distribution.

You observe the following data points:

$$\{0.51, 1.01, -0.42, 1.12, -1.21, 0.02, 0.35, 1, -0.22, 2.00\}$$

Question: Which distribution is most likely to represent the true data generating process?

1. Gaussian(0.0, 1.0)
2. Gaussian(100.0, 1.0)
3. Gaussian(-100.0, 1.0)

An extension of your intuition:

1. Specify a parameter μ and σ^2 for the Gaussian distribution N(μ, σ^2)
2. Then search for all possible μ and σ^2 so that the $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\sigma}}^{\boldsymbol{2}}$ are most likely to give you the observed data points.
3. Mathematically it is represented as: $\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \boldsymbol{P}(\boldsymbol{Y}; \boldsymbol{\beta})$
4. When we extend it to the regression tasks, we need to use conditioning:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\beta})$$

Estimating logistic regression with MLE

Given i.i.d. training data $\{(x_1, y_1), \dots, (x_N, y_N)\}$, we compute the likelihood of seeing such data as a function of β.

$$l(\beta) = \prod_{i:y_i=1} P(y_i = 1|x_i) \prod_{i:y_i=0} (1 - P(y_i = 1|x_i))$$

By using $P(y_i = 1|x_i) = \sigma(\boldsymbol{\beta' x_i})$, we obtain

$$l(\beta) = \prod_{i:y_i=1} \sigma(\boldsymbol{\beta' x_i}) \prod_{i:y_i=0} (1 - \sigma(\boldsymbol{\beta' x_i}))$$

Taking log transformation on both sides, we get

$$L(\beta) = \sum_{i:y_i=1} \log[\sigma(\boldsymbol{\beta' x_i})] + \sum_{i:y_i=0} \log[1 - \sigma(\boldsymbol{\beta' x_i})]$$

Because log transformation is monotonic, the $\hat{\beta}$ that maximizes $l(\beta)$ is the same as the one that maximizes $L(\beta)$. In most of the practice, we only work on $L(\beta)$, which is called log-likelihood.

Estimating logistic regression with MLE

We seed to identify the $\hat{\beta}$ that maximizes the log-likelihood.

$$L(\beta) = \sum_{i:y_i=1} \log[\sigma(\boldsymbol{\beta}'\boldsymbol{x_i})] + \sum_{i:y_i=0} \log[1 - \sigma(\boldsymbol{\beta}'\boldsymbol{x_i})]$$

Or you take negative on both sides, the formula above is equivalent to minimizing:

$$L_{CE}(\beta) = -\sum_{i:y_i=1} \log[\sigma(\boldsymbol{\beta}'\boldsymbol{x_i})] - \sum_{i:y_i=0} \log[1 - \sigma(\boldsymbol{\beta}'\boldsymbol{x_i})]$$

which is called the cross-entropy loss in machine learning.

Notes

- For the purpose of this course, it is more important to know the logic of MLE rather than remembering the math details.
- The process of finding the optimum $\hat{\beta}$ from $L(\beta)$ is way beyond the scope of this class. Typically people use convex optimization methods to estimate $\hat{\beta}$. e.g. BHHH.
- In Python, you could write one line of script to obtain $\hat{\beta}$.
- If I have to choose two most important performance metrics from the statistical and machine learning traditions, I will choose log-likelihood and cross-entropy loss. Please remember the two metrics.

Part 4. Theoretical assumptions, statistical properties, and practice

I make this part quite similar to the previous lecture to facilitate your understanding.
Disclaimer: but the true theoretical foundations of MLE are much more complicated.

About the assumptions from linear regression

Linear regression

- **Assumption 1. No omitted variables (about data)**
- **Assumption 2. Linearity in Parameters (about model)**
- **Assumption 3. Independent and identically distributed random sampling**
- **Assumption 4. No perfect collinearity**
- **Assumption 5. Homoskedasticity**
- **Assumption 6. Normality**

Notes

- Roughly speaking, the assumptions are still critical for deriving the statistical properties of the estimators in MLE.
- One small change is a stronger assumption in A2. No function misspecification because we need to be correct in $\mathrm{E}[\mathrm{y_i}|\mathrm{x_i}] = \sigma(\beta' x_i)$. Both $\sigma()$ and $\beta' x_i$ need to be correct.

Assumptions in logistic regression

Disclaimer: the formal regularity conditions in MLE for a logistic regression are much more complicated, but here I try to translate the regularity conditions and connect them to linear regression.

Logistic regression

- **Assumption 1. No omitted variables (about data)**
- **Assumption 2. Correct model specification in $\sigma(\beta' x_i)$ (about model)**
- **Assumption 3. Independent and identically distributed random sampling**
- **Assumption 4. No perfect collinearity**
- **Assumption 5. Homoskedasticity**
- **Assumption 6. Normality**

Under these assumptions, we can derive the statistical properties of the estimators in MLE.

Statistical properties

Theorem

Under Assumptions 1-5, the vector of the maximum likelihood estimators $\widehat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$, asymptotically follows a multivariate normal distribution

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{MLE}} \sim N(\boldsymbol{\beta}, -\frac{1}{N} E[H(\boldsymbol{\beta})]^{-1})$$

Notes

- $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$ asymptotically follows the normal distribution, which is the same as the results from the linear regression.
- Each element of $\widehat{\boldsymbol{\beta}}$ (i.e. $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k$) is normally distributed, and $\hat{\beta}$ is an unbiased estimator of β as $\mathrm{E}[\hat{\beta}] = \beta$. (same as linear regression)
- However, the variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$ is a bit more complicated, which typically relies on numerical methods to solve the optimization problems.

z-value for logistic regression

Theorem

Given Assumptions I-V, the MLE $\hat{\beta}_j$ is asymptotically normally distributed:

$$Z = \frac{\hat{\beta}_j - \beta_j}{SE[\hat{\beta}_j]} \sim N(0,1)$$

What matters here?

- Except for the name, the process is exactly the same as linear regression.
- We can obtain a distribution of $\hat{\beta}_j$ for statistical test.
- We can test the $\hat{\beta}_j$ one by one.
- We can use Python (or R, Stata, etc.) to compute this distribution.

Using the z-value as a test statistic

Null hypothesis (H_0): $\beta_j = c$. (typically c = 0).

- Compute the z–value as $z = \frac{\widehat{\beta}_j - c}{SE[\widehat{\beta}_j]}$
- Compare the z–value to the critical value $z_{\alpha/2}$ (typically 2) for the α level test (typically $\alpha = 0.05$; then confidence level is 0.95), which under the null hypothesis satisfies

$$P\left(-t_{\alpha/2} \leq z \leq t_{\alpha/2}\right) = 1 - \alpha$$

- Decide whether the realized value of z is unusually large given the known distribution of the test statistic.
- Finally, either declare that we reject H_0 or not, or report the p-value.
- P-value. It measures the probability of observing a z-value at least as extreme as one we observe assuming that H_0 is true.
- We will see the z-value and p-value in logistic regression in the statistical report from Python.

However, in practice, you only need to remember this statement in logistic regression

Statement: "With 95% level of confidence, we observe a statistically significant relationship between y and x."

Find the large z-value (z > 2)
Find the **small p-value (p < 0.05)**

- Different from linear regression, we cannot argue that one unit increase in x is associated with $\hat{\beta}$ unit change in y. However, you can still argue for the statistical significance between x_{ij} and y_i and compute the marginal effects in Python.
- Similar to linear regressions, we are more worried about the violation of A1 and A2 in logistic regressions. That is why people always enrich the models.

An example in practice

Logit Regression Results

Dep. Variable:	property_value_discrete	No. Observations:	4167
Model:	Logit	Df Residuals:	4164
Method:	MLE	Df Model:	2
Date:	Tue, 07 Feb 2023	Pseudo R-squ.:	0.3953
Time:	17:31:00	Log-Likelihood:	-1745.9
converged:	True	LL-Null:	-2887.1
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-6.1168	0.199	-30.666	0.000	-6.508	-5.726
inc_median_household	0.0001	3.44e-06	31.228	0.000	0.000	0.000
households	1.276e-05	4.15e-05	0.307	0.759	-6.86e-05	9.41e-05

Logit Regression Results

Dep. Variable:	property_value_discrete	No. Observations:	4167
Model:	Logit	Df Residuals:	4154
Method:	MLE	Df Model:	12
Date:	Tue, 07 Feb 2023	Pseudo R-squ.:	0.5052
Time:	17:37:13	Log-Likelihood:	-1428.7
converged:	True	LL-Null:	-2887.1
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-7.0535	1.194	-5.907	0.000	-9.394	-4.713
inc_median_household	7.972e-05	4.73e-06	16.841	0.000	7.04e-05	8.9e-05
households	4.422e-05	4.92e-05	0.900	0.368	-5.21e-05	0.000
travel_driving_ratio	-3.2319	1.205	-2.683	0.007	-5.593	-0.871
travel_pt_ratio	13.1837	2.058	6.406	0.000	9.150	17.217
travel_taxi_ratio	15.1983	6.287	2.417	0.016	2.876	27.520
travel_work_home_ratio	-1.5354	1.748	-0.879	0.380	-4.961	1.890
edu_higher_edu_ratio	7.4498	0.584	12.756	0.000	6.305	8.594
household_size_avg	-0.0086	0.013	-0.653	0.514	-0.035	0.017
vacancy_ratio	-0.5773	0.454	-1.271	0.204	-1.468	0.313
rent_median	0.0019	0.000	9.352	0.000	0.001	0.002
race_white_ratio	1.6642	0.299	5.566	0.000	1.078	2.250
race_asian_ratio	-9.7624	1.865	-5.235	0.000	-13.418	-6.107

Process

1. Check R Square.
2. Check coefficients
3. Check z and p values.
4. Interpret coefficients.
5. Keep enriching the model
6. Write a report.

e.g. “With 95% level of confidence, we observe a statistically significant relationship between property value categories and income, after controlling for all the other variables. The controlling variables include travel, household size, and racial composition.”

The relationship is non-linear, so we cannot easily conclude the quantitative association between x and y by reading the coefficients. However, it is not hard to compute in Python.

Part 5. Prediction, evaluation, and back to the general diagram

Prediction with a threshold Value

The outcome of a logistic regression model is a probability

Often, we want to make a binary prediction.

- e.g. Does this person prefer high vs. low-value properties?

We can do this using a *threshold value t.*

- If $P(High\ value\ property\ =\ 1) \geq\ t$, predict high-value property.
- If $P(High\ value\ property\ =\ 1) <\ t$, predict low-value property.

People often use t = 0.5 as a threshold, however, it is often found by testing a group of thresholds.

Evaluating the model

Compare actual outcomes to predicted outcomes using a confusion matrix.

	Predicted = 0	Predicted = 1
Actual = 0	True Negatives (TN)	False Positives (FP)
Actual = 1	False Negatives (FN)	True Positives (TP)

Evaluating the model

Confusion matrix

	Predicted = 0	Predicted = 1
Actual = 0	True Negatives (TN)	False Positives (FP)
Actual = 1	False Negatives (FN)	True Positives (TP)

Developing metrics from the confusion matrix

Prediction accuracy = (TN + TP)/N
Sensitivity = TP/(TP + FN)
Specificity = TN/(TN + FP)
Prediction error rate = (FP + FN)/N
False negative error rate = FN/(TP + FN)
False positive error rate = FP/(TN + FP)
(N = number of observations)

Notes: the best evaluation metric depends on contexts. However, prediction accuracy is the most commonly used metric in machine learning community (not necessarily so in statistics).

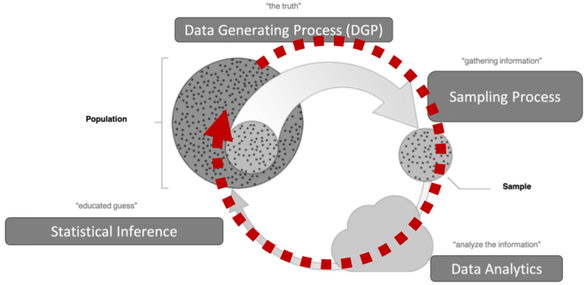
Back to the general diagram

Univariate Linear Regression

1. Establish the goal (DGP)
   e.g. recovering $E[Y|X]$
2. Make modeling assumptions (e.g. i.i.d.)
   e.g. $E[Y|X] = \beta_0 + \beta_1 x_i$
3. Estimate the model by minimizing an objective
   e.g. $\underset{\beta_0,\beta_1}{\text{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - x_i\beta_1)^2$
4. Examine the performance
   e.g. R^2
5. Use the model (interpretation, prediction, etc.)
   e.g. $\hat{\beta}_0, \hat{\beta}_1$

Logistic regression

1. Establish the goal (DGP)
   e.g. recovering $E[Y|X] = P(Y = 1|X)$
2. Make modeling assumptions (e.g. i.i.d.)
   e.g. $E[Y|X] = \sigma(\beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki})$
3. Train the model by minimizing an objective
   e.g. $\underset{\beta}{argmax}$ log-likelihood
4. Examine the performance
   e.g. log-likelihood, accuracy, confusion matrix
5. Use the model (interpretation, prediction, etc.)
   e.g. $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k$



Part 6. Urban applications with statistical methods (including both linear and logistic regressions)

About project: nearly every urban application has some statistical practices. If you have not found it, try harder!

Examples: CAV, policy, location, economic growth, urbanization, and housing

Adoption of shared autonomous vehicles (SAV)

Reference: Krueger, R., et al. (2016). "Preferences for shared autonomous vehicles." Transportation research part C: emerging technologies **69**: 343-355.

Dependent variable: $y \in \{0, 1\}$ (Adoption vs. no adoption)

Independent variables: socio-demographics (income, age, etc.) and travel attributes

Research question: Who will adopt shared autonomous vehicles in the future?

Method: logistic regression.

Table 6
Individual-specific coefficients as estimated for model specification 5.

Coefficient	SAV without DRS		SAV with DRS	
	Estimate	*p*-value	Estimate	*p*-value
Gender (reference = female and other)				
Male	0.21	0.16	0.01	0.93
Age (reference = 30 to 49 years old)				
18–23 years old	0.08	0.78	0.30	0.33
24–29 years old	0.26	0.28	0.63	0.01
50–64 years old	0.13	0.59	−0.02	0.95
65–84 years old	−0.43	0.10	0.01	0.98
Income (reference = 599 AUD/week or less)				
600–1249 AUD/week	0.01	0.97	−0.20	0.26
1250 AUD/week or more	−0.33	0.12	−0.30	0.21
Presence of children (persons aged 17 years old or younger) in the household (reference = no)				
Yes	−0.11	0.52	−0.31	0.12
Car availability (reference = yes)				
No	−0.21	0.42	0.13	0.61
Means of transportation (reference = bicycling or walking)				
Car as driver incl. motorbike/scooter, carsharing	0.69	0.05	0.24	0.53
Car as passenger incl. taxi	0.35	0.36	0.68	0.07
PT incl. PT only and PT and car combined	0.91	0.04	0.49	0.27

Policy choice of lottery vs. auction

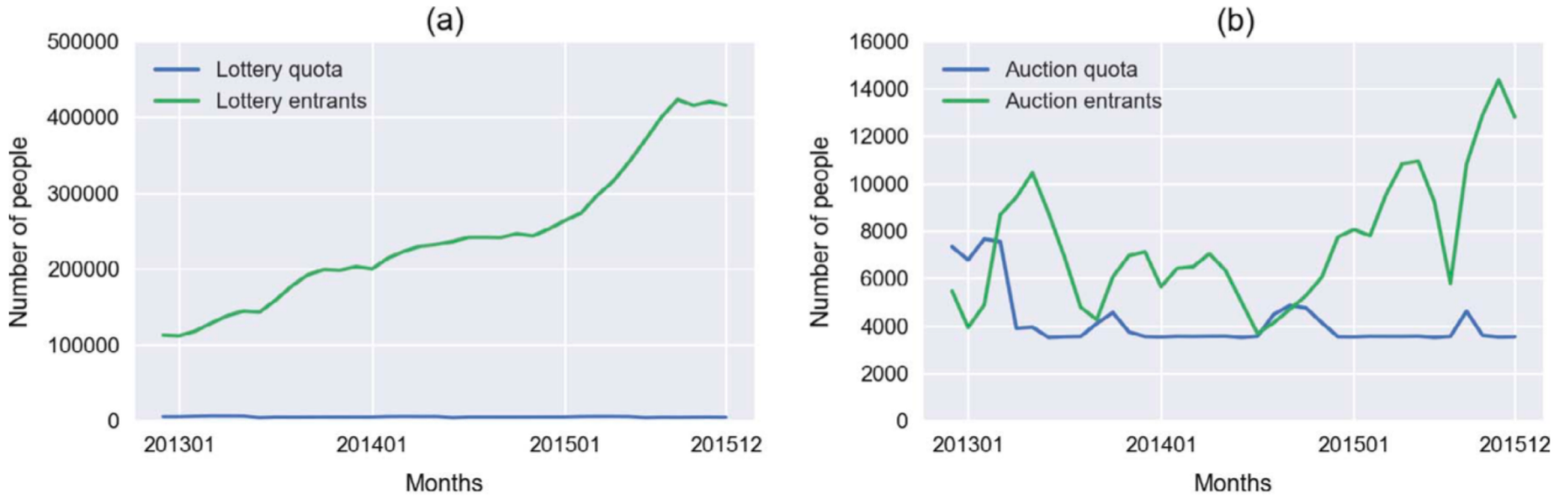
Reference: Wang, S. and J. Zhao (2017). "The distributional effects of lotteries and auctions—License plate regulations in Guangzhou." Transportation Research Part A: Policy and Practice **106**: 473-483.

Dependent variable: $y \in \{0, 1\}$ (auction vs. lottery)

Independent variables: socio-demographics (income, age, etc.)

Research question: Who chose lotteries vs. auctions in winning car license plate?

Method: logistic regression.



Residential location choice

Reference: Waddell, P. (1993). "Exogenous workplace choice in residential location models: is the assumption valid?" Geographical Analysis **25**(1): 65-82.

Dependent variable: $y \in \{0, 1\}$ (location A vs. B)

Independent variables: socio-demographics (income, age, etc.) and travel attributes

Research question: Who chose location A vs. B as their residential locations?

Method: logistic regression (a bit more advanced).

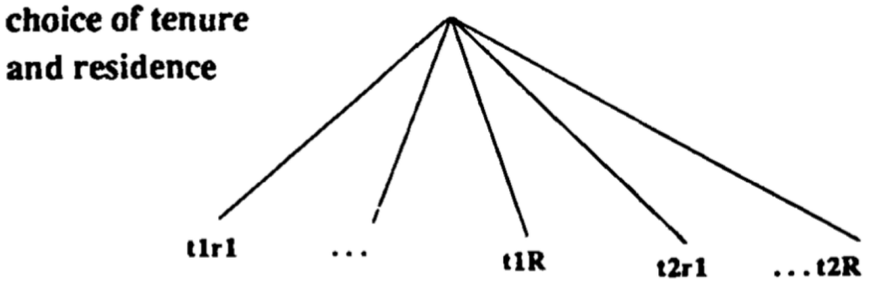


FIG. 3. Two-level Nested Logit Specification

where

$$P_n(r|wt) = \frac{e^{(\bar{V}_r + \bar{V}_{tr} + \bar{V}_{wr} + \bar{V}_{wtr})\mu^r}}{\sum_{r' \in R_{nwt}} e^{(\bar{V}_{r'} + \bar{V}_{tr'} + \bar{V}_{wr'} + \bar{V}_{wtr'})\mu^r}}; \qquad (7)$$

Economic growth

Reference: Glaeser, E. L., et al. (1995). "Economic growth in a cross-section of cities." Journal of monetary Economics **36**(1): 117-143.

Dependent variable: y (economic growth)

Independent variables: baseline GDP, unemployment rate, education, population, etc.

Research question: What factors can predict economic growth?

Method: linear regression (a bit more advanced).

TABLE VI: CITY GROWTH AND EDUCATION
Independent Variable: Log of Growth Rate (1960-1990)

Dependent Variable:	1 City Popula-tion	2 City Popula-tion	3 City Popula-tion	4 SMSA* Popula-tion	5 City Income
Intercept	0.819	-1.108	1.104	0.422	15.664
Log (Population 1960)	-0.042 (0.024)	0.121 (0.248)	-0.040 (0.024)	0.038 (0.025)	-0.012 (0.009)
Per Capita Income 1960 ($1000)	-0.212 (0.085)	-0.223 (0.087)	-0.235 (0.093)	-0.177 (0.136)	-0.155 (0.034)
Unemployment Rate 1960	-0.044 (0.017)	-0.044 (0.017)	-0.042 (0.017)	-0.044 (0.017)	-0.018 (0.0065)
Manufacturing Share 1960	-0.353 (0.276)	-0.322 (0.281)	-0.300 (0.284)	-0.686 (0.259)	-0.144 (0.105)
Median Years of Schooling 1960	0.080 (0.035)	0.264 (0.280)		0.059 (0.038)	0.024 (0.013)
Median. Schooling in 1960 times Log (Population 1960)		-0.015 (0.023)			
Percent of Population with 12-15 Years of Schooling			0.014 (0.006)		

Suburbanization

Reference: Baum-Snow, N. (2007). "Did highways cause suburbanization?" The quarterly journal of economics **122**(2): 775-805.

Dependent variable: y (population growth in suburban areas)

Independent variables: number of highways, and other controls (income, population, etc.).

Research question: Did highway cause suburbanization?

Method: linear regression (a bit more advanced).

Table IV: Long-Difference Regressions of the Determinants of Constant Geography Central City Population Growth, 1950-1990
Large MSAs in 1950

	Change in Log Population in Constant Geography Central Cities					
	OLS3	IV1	IV2	IV3	IV4	IV5
Change in Number of Rays	-0.059	-0.030	-0.106	-0.123	-0.114	-0.101
	(0.014)**	(0.022)	(0.032)**	(0.029)**	(0.026)**	(0.046)*
1950 Central City Radius	0.080		0.111	0.113	0.106	0.125
	(0.014)**		(0.023)**	(0.023)**	(0.023)**	(0.021)**
Change in Simulated Log Income	0.084			0.048	-6.247	-0.137
	(0.378)			(0.417)	(6.174)	(0.480)
Change in Log of MSA Population	0.363			0.424	0.374	0.405
	(0.082)**			(0.094)**	(0.079)**	(0.108)**
Change in Gini Coeff of Simulated Income					-23.416	
					(23.266)	
Log 1950 MSA Population						-0.062
						(0.062)
Constant	-0.640	-0.203	-0.359	-0.588	4.580	-0.611
	(0.260)*	(0.078)*	(0.076)**	(0.281)*	(5.091)	(0.265)*
Observations	139	139	139	139	139	139
R-squared	0.39	0.00	0.01	0.30	0.33	0.37

Housing

Reference: Baum-Snow, N. and M. E. Kahn (2000). "The effects of new public projects to expand urban rail transit." Journal of public economics **77**(2): 241-263.

Dependent variable: y (housing/rental values)

Independent variables: proximity to transit lines, and other controls.

Research question: Did transit accessibility increase housing values?

Method: linear regression (a bit more advanced).

Table 5
Housing capitalization of transit[a]

	Change in census tract median rental price, 1980 to 1990		Change in census tract median home price, 1980 to 1990	
	City dummies and central city dummy but no demographic controls	City dummies and central city dummy with demographic controls	City dummies and central city dummy but no demographic controls	City dummies and central city dummy with demographic controls
Change in distance to transit	−15.75	−10.31	−5741	−2582
	(3.99)	(3.03)	(2928)	(2383)
Change in distance to transit squared	0.32	0.21	109	24
	(0.16)	(0.12)	(114)	(92)
Observations	3546	3499	3410	3369
Adj. R^2	0.477	0.616	0.261	0.579

Problem Set 1. Practicing statistical regressions

Same data structure as the lab sessions, but its context is in **Illinois**.

Deadline: **Feb 21**

Part 1. Descriptive analysis (5pts)

Part 2. Univariate regression (5pts)

Part 3. Multivariate regression (5pts)