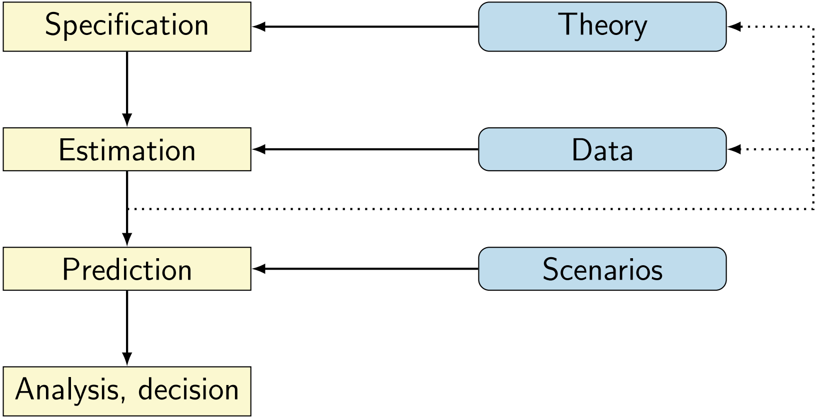
Mathematical modeling

Contingency table and linear regression

Michel Bierlaire

Introduction to transportation systems

Model development



Example: hypothesis

Theory: hypothesis

- Transportation mode share depends on trip purpose.
- Why? Convenience, flexibility, comfort

Example: data collection

Sample

- 2000 travelers randomly selected.
- traveler: somebody who made a trip yesterday.

Questions

Consider a trip that you have made yesterday.

1. What was the purpose of your trip? Work/leisure/other.
2. Did you use public transportation? Yes/no.

Example: data collection

Results

Contingency table:

	Work	Leisure	Others
PT	172	191	150
Not PT	345	648	494

Synthetic data generated from Microcensus 2015.

Example: model

Model

- Y: transportation mode. Qualitative with $\mathcal{A}=${public transport, others}.
- X: trip purpose. Qualitative with $\mathcal{A}=${work, leisure, others}.
- Model: $Y|X$.

Example: model parameters

$Y|\text{work}$

- $\theta_1 = p_{Y|\text{work}}(\text{PT})$
- $p_{Y|\text{work}}(\text{not PT}) = 1 - \theta_1$

$Y|\text{leisure}$

- $\theta_2 = p_{Y|\text{leisure}}(\text{PT})$
- $p_{Y|\text{leisure}}(\text{not PT}) = 1 - \theta_2$

$Y|\text{others}$

- $\theta_3 = p_{Y|\text{others}}(\text{PT})$
- $p_{Y|\text{others}}(\text{not PT}) = 1 - \theta_3$

Estimation of the parameters

One observation

- Consider a traveler traveling by public transport for work.
- What is the probability that our model correctly predicts this observation?

$$\Pr(Y = \mathrm{PT} | X = \mathrm{work}) = \theta_1.$$

All observations in one cell

- Consider all 172 travelers in the sample traveling by public transport for work.
- What is the probability that our model correctly predicts all these observations?
  θ_1^{172}

Estimation of the parameters

All observations

- Consider all travelers in the sample.
- What is the probability that our model correctly predicts all these observations? $\theta_1^{172}(1-\theta_1)^{345}\theta_2^{191}(1-\theta_2)^{648}\theta_3^{150}(1-\theta_3)^{494}$.

	Work	Leisure	Others
PT	172	191	150
Not PT	345	648	494

Estimation of the parameters

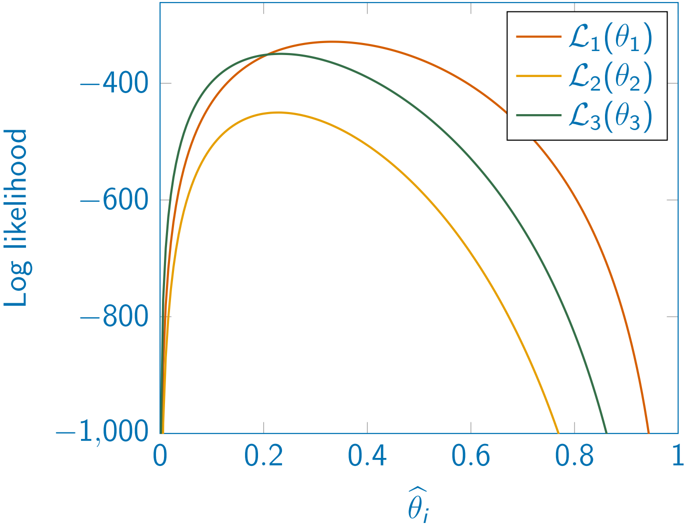
Likelihood function

$$\mathcal{L}^*(\theta_1,\theta_2,\theta_3) = \theta_1^{172}(1-\theta_1)^{345}\theta_2^{191}(1-\theta_2)^{648}\theta_3^{150}(1-\theta_3)^{494}.$$

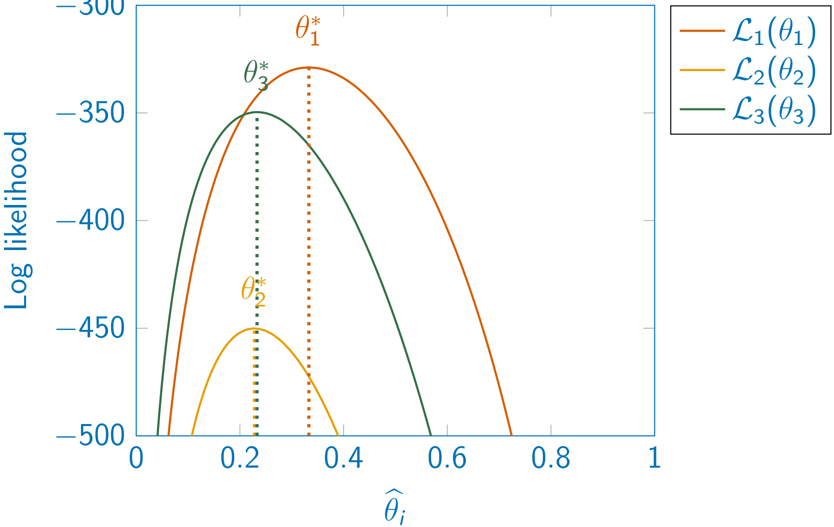
Log likelihood function

$$\begin{gathered}\mathcal{L}(\theta_1,\theta_2,\theta_3) = \\ 172\log\theta_1 + 345\log(1-\theta_1) + \\ 191\log\theta_2 + 648\log(1-\theta_2) + \\ 150\log\theta_3 + 494\log(1-\theta_3) = \\ \mathcal{L}_1(\theta_1) + \mathcal{L}_2(\theta_2) + \mathcal{L}_3(\theta_3).\end{gathered}$$

Maximum likelihood estimation



Maximum likelihood estimation



Maximum likelihood estimation

Estimates

- $\widehat{\theta}_1 = 0.333$
- $\widehat{\theta}_2 = 0.228$
- $\widehat{\theta}_3 = 0.233$

Data

	Work	Leisure	Others
PT	172	191	150
Not PT	345	648	494
Total	517	839	644
	$\frac{172}{517} =$ 0.333	$\frac{191}{839} =$ 0.228	$\frac{150}{644} =$ 0.233

Estimators

My data

	Work	Leisure	Others
PT	172	191	150
Not PT	345	648	494

My estimate: θ_1 =33.3%

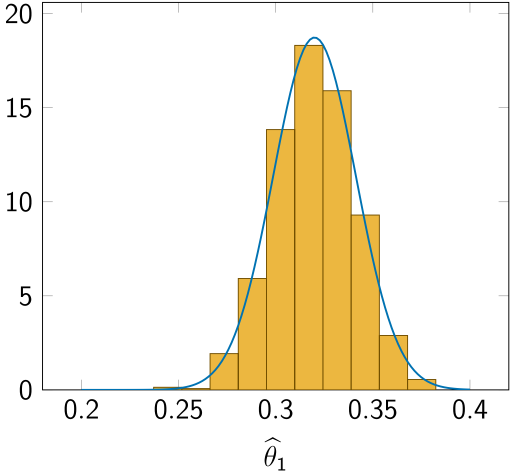
My colleague's data

	Work	Leisure	Others
PT	168	207	140
Not PT	317	677	491

Her estimate: θ_1 =34.6%

Estimators

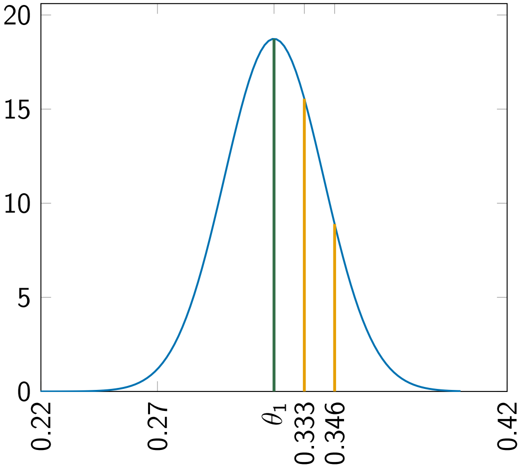
Perform the same analysis with 1000 different samples



Estimators

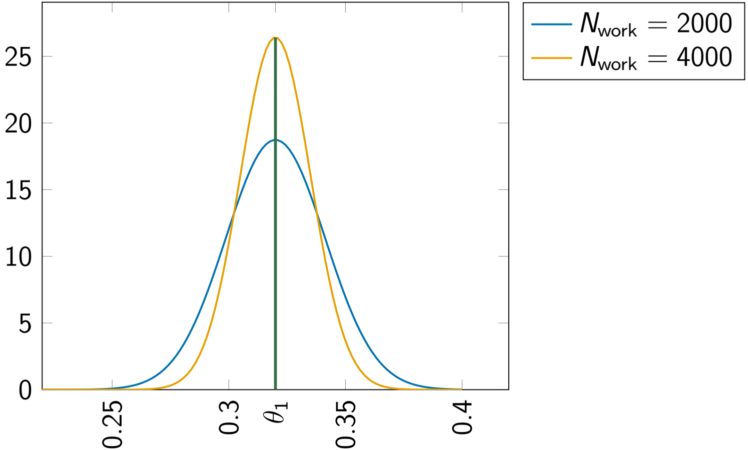
Estimators are distributed. They are random variables.

True value: $\theta_1 = 0.320$. My estimate: 0.333. Her estimate: 0.346



Estimators

The larger the sample, the lower the variance of the estimator.



Maximum likelihood estimation

Likelihood function

Probability that our model reproduces the N observations:

$$\mathcal{L}^*(\theta) = \prod_{n=1}^{N} \Pr(Y = y_n, X = x_n; \theta) = \prod_{n=1}^{N} \Pr(Y = y_n | X = x_n; \theta) \Pr(X = x_n)$$

Log likelihood function

In practice, we take the log:

$$\mathcal{L}(\theta) = \log \mathcal{L}^*(\theta) = \sum_{n=1}^{N} \log \Pr(Y = y_n | X = x_n; \theta) + \log \Pr(X = x_n)$$

Maximum likelihood estimation

Estimator

$$\widehat{\theta} = \operatorname{argmax}_\theta \mathcal{L}(\theta) = \sum_{n=1}^{N} \log \Pr(Y = y_n | X = x_n; \theta) + \cancel{\log \Pr(X = x_n)}$$

Asymptotic properties

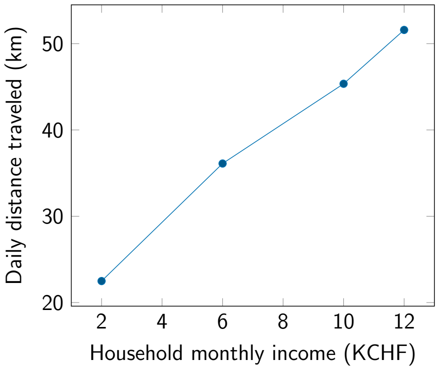
- Consistency.
- Efficiency.
- Normally distributed.

Another example

Data: Swiss Microcensus 2015.
Source: ARE.

Theory: hypothesis

- Daily distance traveled depends on income
- Why? Individuals with different incomes have different socio-professional activities. They also have different access to mobility.



Another example: data

Monthly income (KCHF)	Daily distance (km)
2	22.49
6	36.11
10	45.35
12	51.59

Another example: model

Model

- Y: daily distance (km). Continuous.
- X: monthly income (KCHF). Continuous.
- Model: $Y|X$.

Specification: linear regression

$$Y|(X = x_n) = \theta_1 x_n + \xi_n, \quad \text{where } \xi_n \sim N(\theta_0, \theta_2^2).$$

$Y|(X = x_n) = \theta_1 x_n + \theta_0 + \theta_2 \xi_n$
$\xi_n \sim N(0, 1)$, i.i.d. across observations, θ_0 and θ_2 constant across observations
$f_{\xi_n}(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$.
θ_0, θ_1 and θ_2 are parameters to estimate.

Another example: model

Probability density function

$$f_{Y|x_n}(z;\theta_0,\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2^2}} \exp\left(-\frac{1}{2}\left(\frac{z-\theta_1 x_n-\theta_0}{\theta_2}\right)^2\right)$$

Estimation of the parameters

One observation

$$x_1 = 2, y_1 = 22.49.$$

What is the probability that our model correctly predicts this observation? Zero!
We have a continuous random variable

Log likelihood function

Intuition

Somehow, the pdf plays the role of the probability.

One observation

$$x_1 = 2, y_1 = 22.49.$$

Contribution to the log likelihood:

$$\log f_{Y|x_1}(y_1; \theta_0, \theta_1, \theta_2) =$$

$$\cancel{-\frac{1}{2}\log(2\pi)} - \log(\theta_2) - \frac{1}{2\theta_2^2}(y_1 - \theta_1 x_1 - \theta_0)^2$$

Log likelihood function

All observations

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log f_{Y|x_n}(y_n; \theta) = -N \log(\theta_2) - \frac{1}{2\theta_2^2} \sum_{n=1}^{N} (y_n - \theta_1 x_n - \theta_0)^2$$

Maximum likelihood estimation

$$\max_{\theta} \mathcal{L}(\theta) = -N \log(\theta_2) - \frac{1}{2\theta_2^2} \sum_{n=1}^{N} (y_n - \theta_1 x_n - \theta_0)^2$$

Least squares

Procedure

- Denote $\theta_2 = \sigma$.
- Fix σ and solve for θ_0 and θ_1.
- Then solve for σ.

Fix σ

$$\max_{\theta} \mathcal{L}(\theta) = \cancel{-N \log(\sigma)} - \frac{1}{2\cancel{\sigma^2}} \sum_{n=1}^{N} (y_n - \theta_1 x_n - \theta_0)^2$$

becomes

$$\min_{\theta} \frac{1}{2} \sum_{n=1}^{N} (y_n - \theta_1 x_n - \theta_0)^2$$

Solution: $\widehat{\theta}_0$ and $\widehat{\theta}_1$.

Least squares

Solve for σ

Residuals:

$$z = \sum_{n=1}^{N} (y_n - \widehat{\theta}_1 x_n - \widehat{\theta}_0)^2$$

Solve

$$\max_{\sigma} -N \log(\sigma) - \frac{1}{2\sigma^2} z.$$

Derivative: $-N\sigma^{-1} + \sigma^{-3} z = 0$
Derivative: $\sigma^2 = z/N$

Least squares: statistical note

Maximum likelihood estimator of the variance

$$\widehat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (y_n - \widehat{\theta}_1 x_n - \widehat{\theta}_0)^2$$

Biased estimator

Unbiased estimator of the variance

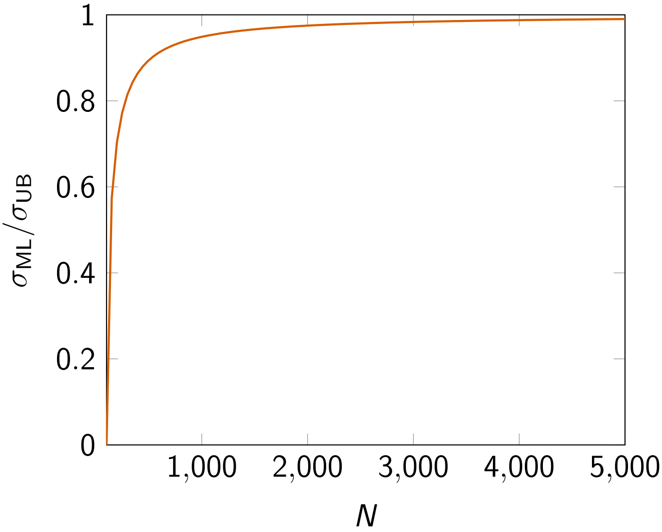
- Let $K = 2$ be the number of parameters.

$$\widehat{\sigma}^2 = \frac{1}{N - K} \sum_{n=1}^{N} (y_n - \widehat{\theta}_1 x_n - \widehat{\theta}_0)^2$$

- Not important if N is significantly larger than K.

Least squares: statistical note

$K = 100$



Linear regression and least squares

Model with $X \in \mathbb{R}^{K-1}$

$$Y|(X = x) = \sum_{k=1}^{K-1} \theta_k x_k + \theta_0 + \sigma\varepsilon$$

Regression curve of Y on x

$$\mathrm{E}[Y|X = x] = \sum_{k=1}^{K-1} \theta_k x_k + \theta_0$$

Linear regression and least squares

Least square estimates

$$\widehat{\theta} = \operatorname{argmin} \frac{1}{2} \sum_{n=1}^{N} (y_n - \sum_{k=1}^{K-1} \theta_k X_k - \theta_0)^2$$

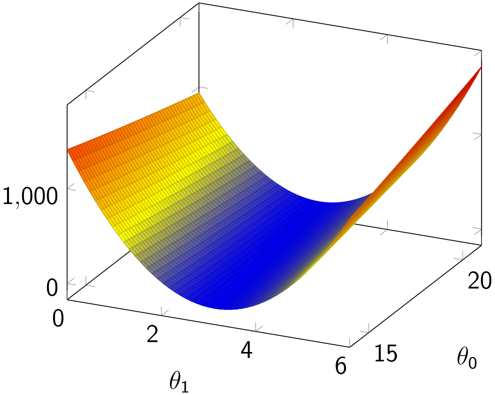
Estimation of the variance

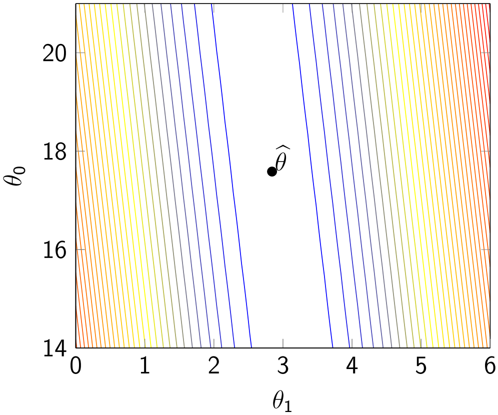
$$(\widehat{\sigma})^2 = \frac{1}{N} \sum_{n=1}^{N} (y_n - \sum_{k=1}^{K-1} \widehat{\theta}_k X_k - \widehat{\theta}_0)^2$$

or

$$(\widehat{\sigma})^2 = \frac{1}{N-K} \sum_{n=1}^{N} (y_n - \sum_{k=1}^{K-1} \widehat{\theta}_k X_k - \widehat{\theta}_0)^2$$

Back to our example

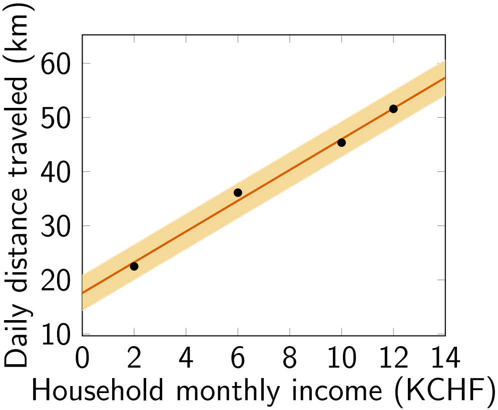




$\widehat{\theta}_0 = 17.6$, $\widehat{\theta}_1 = 2.84$, $\widehat{\sigma} = 0.896$ or $\widehat{\sigma} = 1.27$

Regression line

with 99% confidence interval



Linear regression: matrix form

K parameters, N observations

Data: $x \in \mathbb{R}^{N \times K}$, $y \in \mathbb{R}^N$

$$y = x\theta + \sigma\varepsilon$$

where

$$\theta \in \mathbb{R}^K, \sigma \in \mathbb{R}, \varepsilon \in \mathbb{R}^N \text{ and } \varepsilon \sim N(0, I).$$

Example: $K = 2$

$$\begin{pmatrix} 22.49 \\ 36.11 \\ 45.35 \\ 51.59 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 1 & 6 \\ 1 & 10 \\ 1 & 12 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} + \sigma \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}$$

Linear regression: matrix form

K parameters, N observations

Data: $x \in \mathbb{R}^{N \times K}$, $y \in \mathbb{R}^N$

$$y = x\theta + \sigma\varepsilon$$

where

$$\theta \in \mathbb{R}^K, \sigma \in \mathbb{R}, \varepsilon \in \mathbb{R}^N \text{ and } \varepsilon \sim N(0, I).$$

Normal equations and estimator

$$x^T x \widehat{\theta} = x^T y \iff \widehat{\theta} = (x^T x)^{-1} x^T y$$

Estimator of σ

$$\widehat{\sigma}^2 = \frac{1}{N}(y - x\theta)^T (y - x\theta) \text{ or } \widehat{\sigma}^2 = \frac{1}{N-K}(y - x\theta)^T (y - x\theta)$$

Example

$$y = \begin{pmatrix} 22.69 \\ 36.11 \\ 45.35 \\ 51.59 \end{pmatrix}, \; x = \begin{pmatrix} 1 & 2 \\ 1 & 6 \\ 1 & 10 \\ 1 & 12 \end{pmatrix}$$

Normal equations and estimator

$$\begin{pmatrix} 4 & 30 \\ 30 & 284 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} = \begin{pmatrix} 155.74 \\ 1334.22 \end{pmatrix} \iff \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} = \begin{pmatrix} 17.6 \\ 2.84 \end{pmatrix}$$

Example

Residuals

$$y - x\theta = \begin{pmatrix} -0.764746 \\ 1.487797 \\ -0.639661 \\ -0.083390 \end{pmatrix} \quad (y - x\theta)^T (y - x\theta) = 3.2145$$

ML estimator of σ

$$\sqrt{\frac{1}{4} \cdot 3.2145} = 0.896$$

Unbiased estimator of σ

$$\sqrt{\frac{1}{4-2} \cdot 3.2145} = 1.27$$

Summary

Variables

- Continuous
- Qualitative discrete
- Random

Model

- Causality $\neq$ correlation.
- Specification based on theory/hypotheses.

Parameter estimation

- Maximum likelihood estimation.

Models covered so far:

- X and Y are both discrete.
- X and Y are both continuous.

What if Y is discrete and X continuous?