Machine Learning

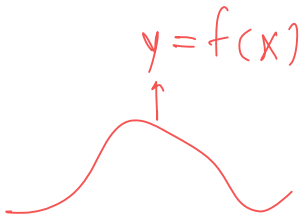
Lecture 4: Regression

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

BIG DATA EXPERIENCE CENTER

Lecture 3 - Topics



- Linear regression, constraints
- Poisson regression
- Survival analysis

Lab

https://drive.google.com/file/d/12lm4MO-xYqFZb-gp_VChhDncPjg_PPet/view?usp=sharing

Simple linear regression

- Simple linear regression model has one input x and one output y.
- The relationship can be explain as the following equation

$$f(\boldsymbol{x}) = w_0 + w_1 x$$

- Mathematically, parameters are obtained by least square method

Multiple regression (GLM)

- Multiple linear regression model structure is exactly the same as the linear regression

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$

- Mathematically, parameters are obtained by least square method

Multiple regression with interaction

- Adding interaction terms to a regression model can greatly expand understanding of the relationships among the variables in the model
- This occurs when two or more variables depend on one another for the outcome

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + \cdots$$

Feature selection via shrinkage (Regularization)

- Instead of explicitly selecting features, in some approaches we can bias the learning process towards using a small number of features
- Key ideas: objective function has two parts
  - Term representing error minimization
  - Term that shrinks parameters toward 0

Shrinkage and regression

- Consider the case of linear regression

$$y = f(x) = w_0 + \sum_{i=1}^{n} w_i x_i$$

- The standard approach minimizes sum squared error

$$E(w) = \sum_{d \in D} \left(y^{(d)} - f\left(x^d\right) \right)^2$$

$$= \sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{i=1}^{n} w_i x_i^{(d)} \right)^2$$

Ridge regression and the Lasso

ได้ weight เล็ก

- Ridge regression adds a penalty term, the L2 norm of the weights

$$E(w) = \sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{i=1}^{n} w_i x_i^{(d)} \right)^2 + \lambda \sum_{i=1}^{n} w_i^2$$

- The Lasso method adds a penalty term, the L_1 norm of the weights

$$E(w) = \sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{i=1}^{n} w_i x_i^{(d)} \right)^2 + \lambda \sum_{i=1}^{n} |w_i|$$

Lasso optimization

- Lasso stands for “least absolute shrinkage and selection operator

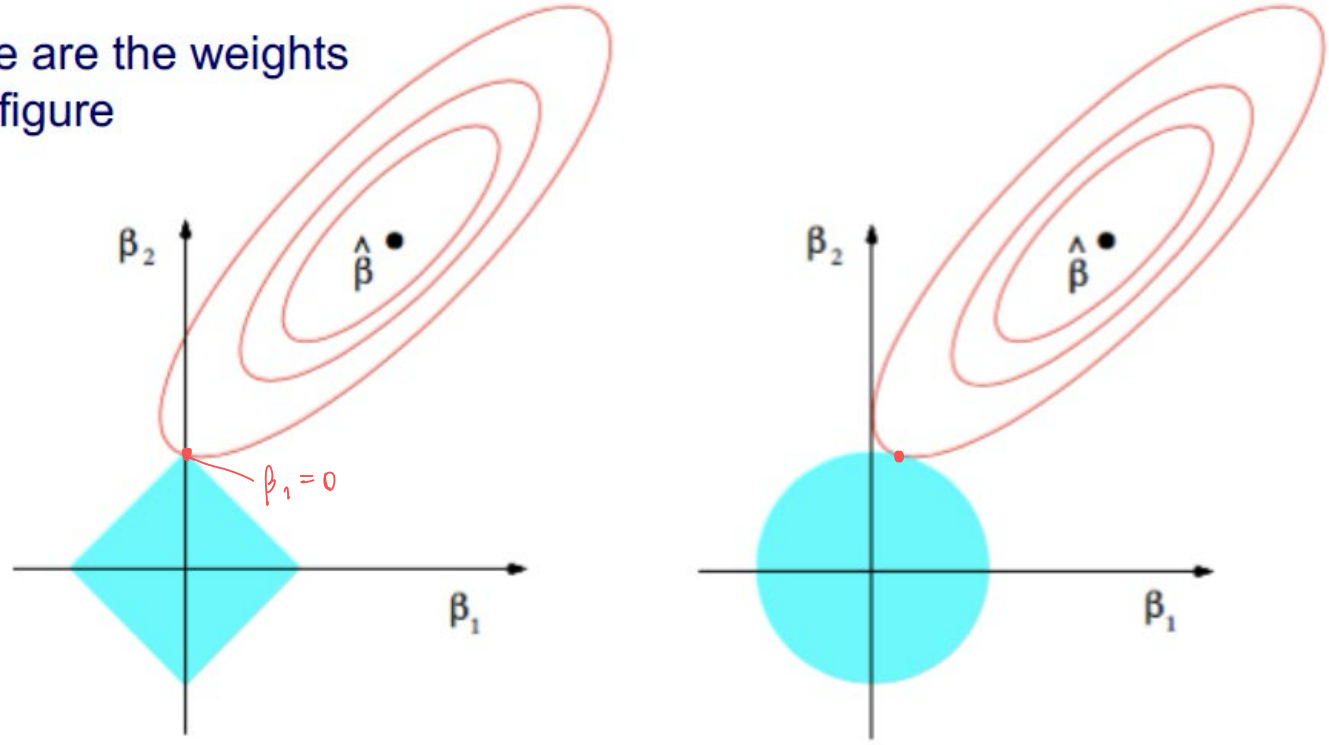
$$E(w) = \sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{i=1}^{n} w_i x_i^{(d)} \right)^2 + \lambda \sum_{i=1}^{n} |w_i|$$

- This is equivalent to the following constrained optimization problem

$$E(w) = \sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{i=1}^{n} w_i x_i^{(d)} \right)^2 \quad \text{subject to} \quad \sum_{i=1}^{n} |w_i| \leq t$$

Ridge regression and Lasso

β's are are the weights in this figure



- สีฟ้าวาดขึ้นจาก norm(λ)
- สีแดงมาจาก error
- หาจุดที่ balance กัน ระหว่างทั้งสอง (สัมผัสกัน)

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*

Figure from Hastie et al., The Elements of Statistical Learning, 2008

Feature selection via shrinkage

- Lasso (L1) tends to make many weights 0, inherently performing feature selection
- Ridge regression (L2) shrinks weights but isn't as biased towards selecting features
- L1 and L2 penalties can be used with other learning methods (logistic regression, neural nets, SVMs, etc.)
- Both can help avoid overfitting by reducing variance
- There are many variants with somewhat different biases
  - elastic net: includes L1 and L2 penalties

↳ รวมร่าง

As an optimization problem, binary class ℓ_2 penalized logistic regression minimizes the following cost function:

$$\text{l1\_ratio} = 0 \qquad \min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

Similarly, ℓ_1 regularized logistic regression solves the following optimization problem:

$$\text{l1\_ratio} = 1 \qquad \min_{w,c} \|w\|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

Elastic-Net regularization is a combination of ℓ_1 and ℓ_2, and minimizes the following cost function:

$$0 <= \text{l1\_ratio} <= 1 \qquad \min_{w,c} \frac{1-\rho}{2} w^T w + \rho \|w\|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1),$$

l1_ratio : ***float or None, optional (default=None)***

The Elastic-Net mixing parameter, with `0 <= l1_ratio <= 1`. Only used if `penalty='elasticnet'`. Setting ``l1_ratio=0`` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For `0 < l1_ratio <1`, the penalty is a combination of L1 and L2.

Embedded approach
Logistic regression

```
[45] lr = LogisticRegression(penalty = 'l1', C=0.1, solver='liblinear')
     lr.fit(X_train, y_train)
```

```
LogisticRegression
LogisticRegression(C=0.1, penalty='l1', solver='liblinear')
```

Embedded approach
Logistic regression

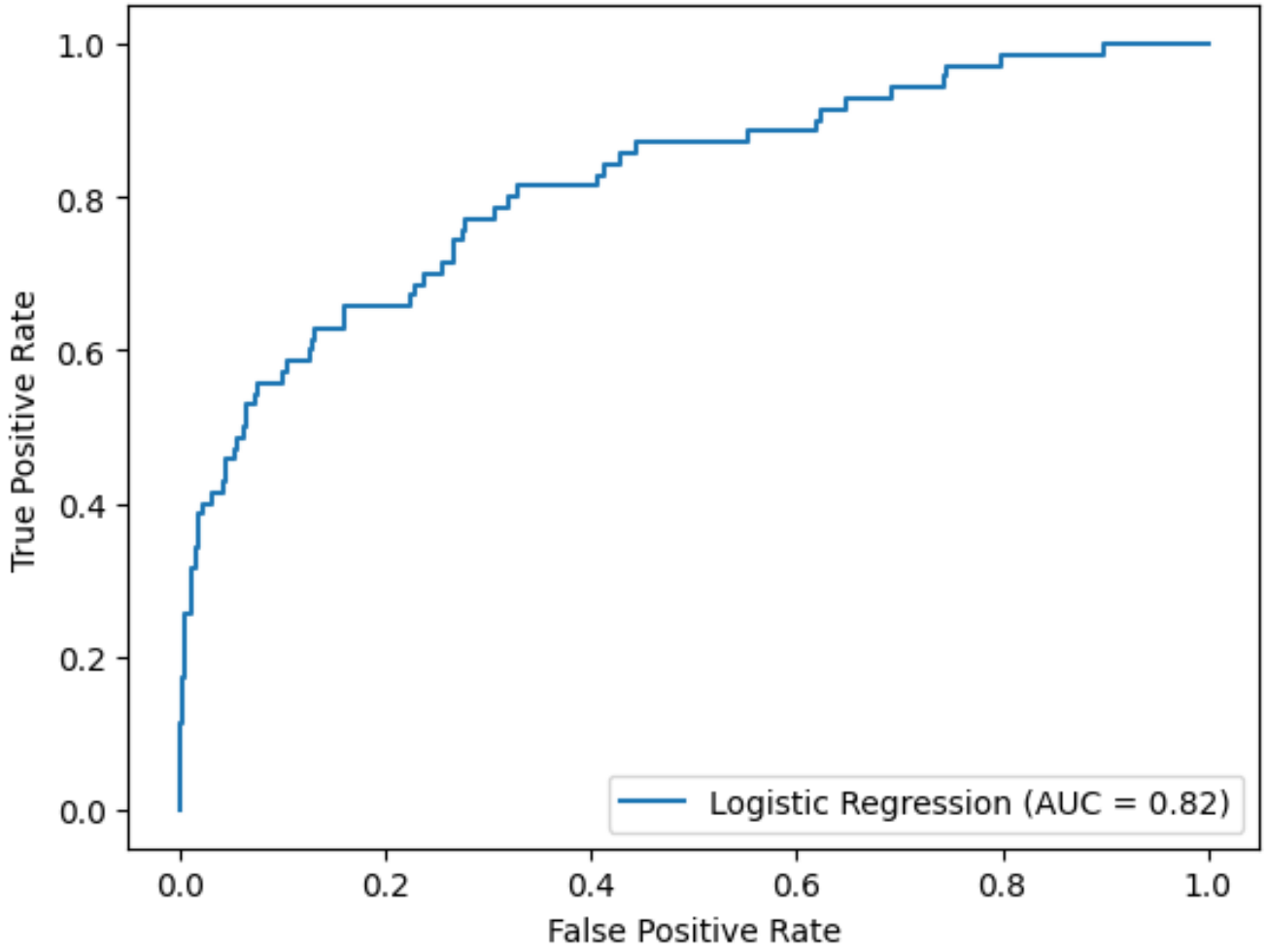
```
[46] lr.coef_[0]
```

```
array([ 4.39645027e-01,  0.00000000e+00,  0.00000000e+00,  3.64882336e-01,
        0.00000000e+00,  0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
        0.00000000e+00,  0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
        0.00000000e+00,  0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
        0.00000000e+00,  0.00000000e+00,  0.00000000e+00,  5.38658322e-01,
        1.08934507e+00, -1.10260038e-02, -3.53481268e-04,  3.35990896e-02,
        0.00000000e+00, -2.10427843e-01,  2.69803017e-03, -2.26244263e-01,
        0.00000000e+00, -1.45808259e-01, -1.18404056e-04,  7.08198896e-06,
        1.36897240e-01,  0.00000000e+00,  1.39256563e-01, -1.07618808e-01,
       -1.17913518e-01, -1.94316308e-02, -3.42778865e-02, -7.65126263e-02,
        3.73647156e-02, -1.03708133e-01,  1.20789097e-01, -6.54389495e-02])
```

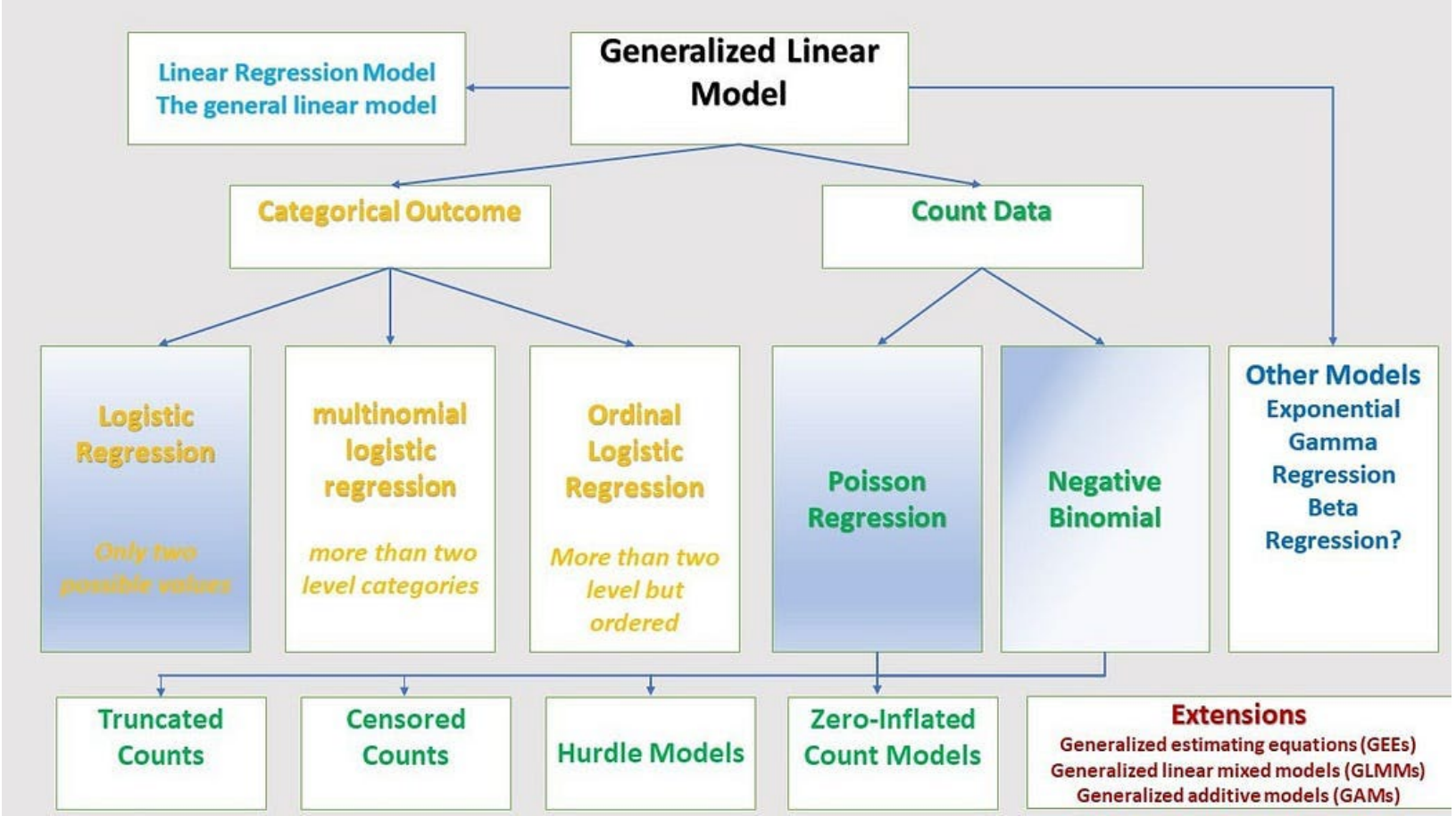
```
X_train.columns[lr.coef_[0]>0.001]
```

```
Index(['BusinessTravel_Travel_Frequently', 'Department_Sales',
       'MaritalStatus_Single', 'OverTime_Yes', 'DistanceFromHome',
       'HourlyRate', 'NumCompaniesWorked', 'PerformanceRating',
       'YearsAtCompany', 'YearsSinceLastPromotion'],
      dtype='object')
```

Embedded approach
Logistic regression



Generalized linear model GLM



https://medium.com/@sahin.samia/a-comprehensive-introduction-to-generalized-linear-models-fd773d460c1d

Poisson Regression

Poisson regression

$y = f(x)$

odd ratio

$$\log\left(\frac{p}{1-p}\right) = f(x)$$

$$\frac{p}{1-p} = e^{f(x)}$$

$$p = e^{f(x)} - p \cdot e^{f(x)}$$

$$p = \frac{e^{f(x)}}{1+e^{f(x)}} = \frac{1}{1+e^{-f(x)}}$$

∴ sigmoid

So far we have looked at 2 kinds of regression

- Linear regression (simple, multiple) → ทำนายปริมาณ
  - Continuous response, normally distributed with constant variance
  - Mean a linear function of the covariates
- "Logistic regression" → โอกาสในการเกิดขึ้น
  - Response (number of successes in n trials) has a binomial distribution Bin(n, p)
  - Mean is *np* where log-odds of *p* is a linear function of the covariates

Poisson regression (cont)

- Now we consider “Poisson regression”
  - Response is a count, assumed to have a Poisson distribution with (positive) mean μ
  - Assume that log μ is a linear function of the covariates
  - Alternatively,

    μ = exp(linear function of covariates)
- Poisson is a standard distribution when response is a count.

Poisson distribution

$$\Pr(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

Count Y can have values 0, 1, 2, . . .

(thus, mean μ must be positive)

The Poisson Regression Model

- The response Y with covariates $x_1, \ldots, x_k$ has a Poisson distribution, with mean μ

- Mean μ is related to the covariates by

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$ ~ คล้าย lin reg. $y = m_1x_1 + m_2x_2 + \ldots$

$$\mu = e^{(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}$$

- As for logistic regression, the parameters are estimated by maximum likelihood

Interpretation of β's

- If $\beta_j > 0$, mean ***increases*** with x_j
- If $\beta_j < 0$, mean ***decreases*** with x_j

- Unit increase in x_j changes the mean by a factor of $\exp(\beta_j)$
  (like the odds in logistic regression)

no negative value

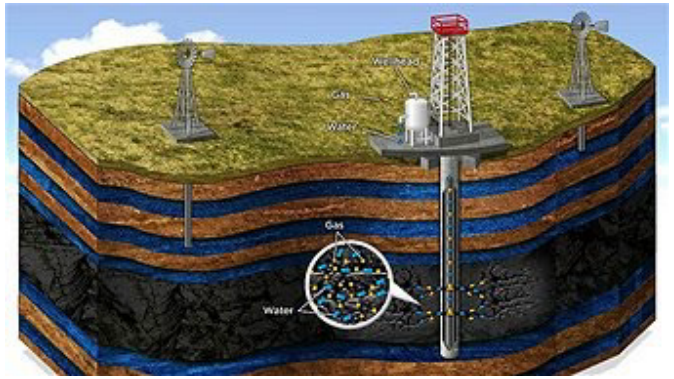
Estimation of β's

- To estimate the β's , we use the method of *maximum likelihood,* as in logistic regression
- Basic idea:
  - Using the ***Poisson*** distribution, we can work out the probability of getting any particular set of responses y.
  - In particular, we can work out the probability of getting the data we actually observed – this is the likelihood
  - Choose β's to maximise this probability (or, equivalently, the log-likelihood)

Example: Mining accident data

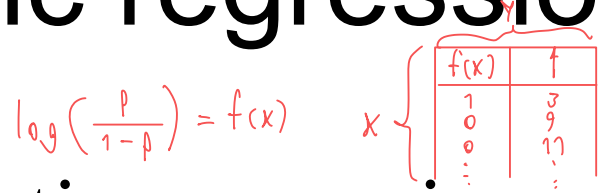
- This example features the number of accidents per mine in a 3 month period in 44 coal mines in West Virginia. The variables are
  - COUNT: the number of accidents (response)
  - INB: inner burden thickness
  - EXTRP: percentage of coal extracted from mine
  - AHS: the average height of the coal seam in the mine
  - AGE: the age of the mine

In .ipynb file: ถ้าค่าช่วง CI ไม่คร่อม 0 แสดงว่า feature มีผลต่อ y
ประกอบกับ $P>|z| = 0$
∴ EXTRP มีผลต่อ accidents อย่างมีนัยสำคัญ



Survival analysis

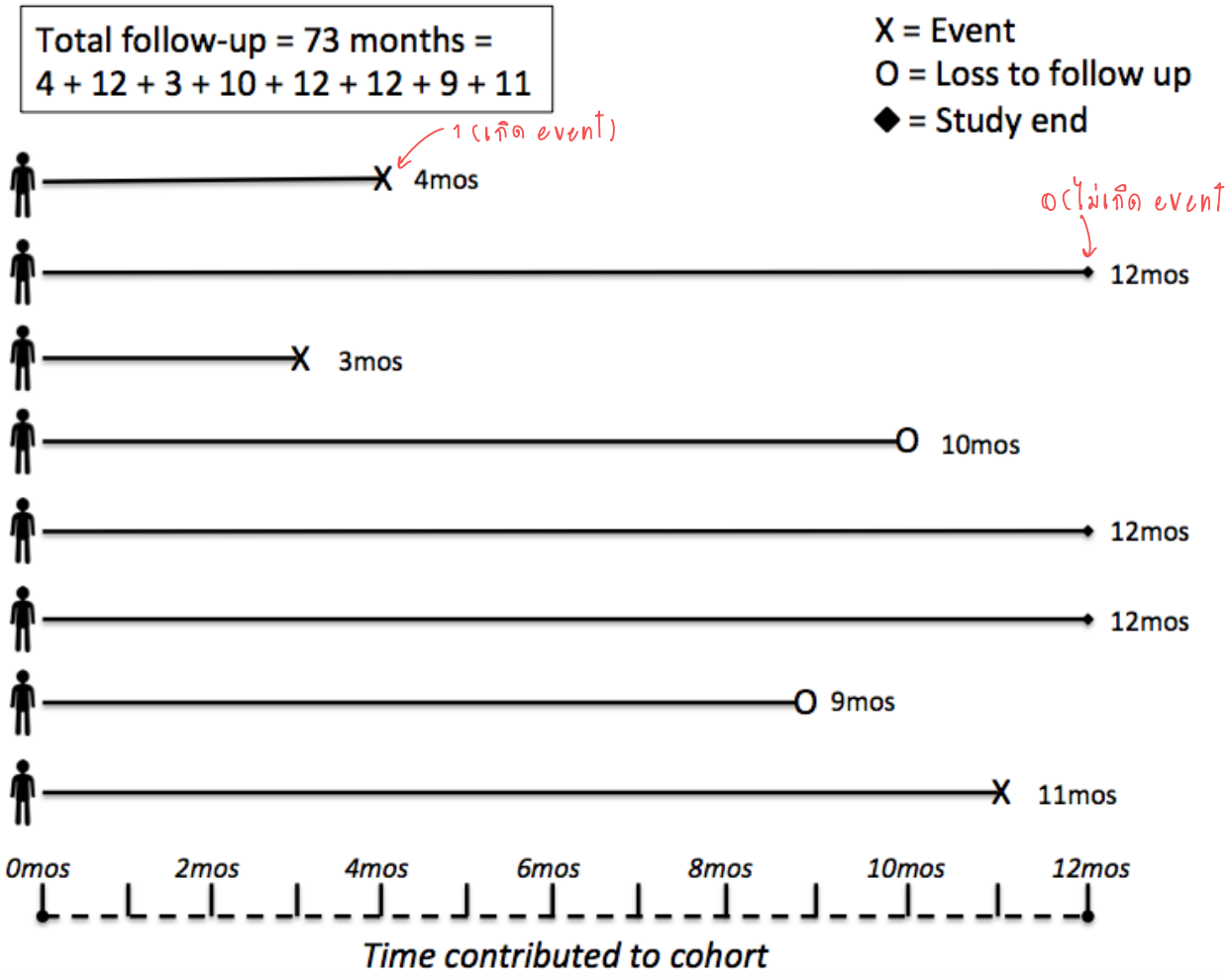
Logistic regression vs time



- In logistic regression, we were interested in studying how risk factors were associated with presence or absence of disease.
- Sometimes, we are interested in how a risk factor or treatment affects time to disease or some other event.
- In these cases, logistic regression is not appropriate.

Survival analysis

- Survival analysis is used to analyze data in which the time until the event is of interest.
- The response is often referred to as a failure time, survival time, or event time.



Example:

- Time until tumor recurrence
- Time until a machine part fails
- Time until mobile phone recharge
- Time until the next credit card usage

The survival time response

- Usually continuous
- May be incompletely determined for some subjects
  - For some subjects we may know that their survival time was at least equal to some time t. บอกได้เวลาก่อนหยุดอย่างน้อย t
  - Whereas, for other subjects, we will know their exact time of event.
- Incompletely observed responses are **censored**
- Is always≥0.

ช่วงเวลาไม่ได้กิจกรรมที่สนใจ
ex. ไม่เข้าเกม 1 สัปดาห์, หมอนัดแล้วไม่ไป

Analysis issue

- If there is no censoring, standard linear regression procedures could be used.
- However, these may be inadequate because
  - Time to event is restricted to be positive and has a skewed distribution.
  - The probability of surviving past a certain point in time may be of more interest than the expected time of event.
  - The hazard function, used for regression in survival analysis, can lend more insight into the failure mechanism than linear regression.

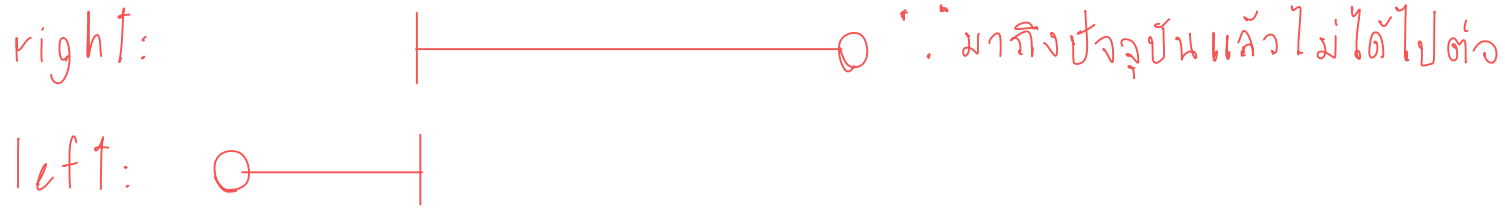
Censoring เก็บข้อมูลต่อไม่ได้

- Censoring is present when we have some information about a subject's event time, but we don't know the exact event time.
- For the analysis methods we will discuss to be valid, censoring mechanism must be independent of the survival mechanism.

Reasons censoring might occur

- A subject does not experience the event before the study ends
- A person is lost to follow-up during the study period
- A person withdraws from the study

These are all examples of **right-censoring.**



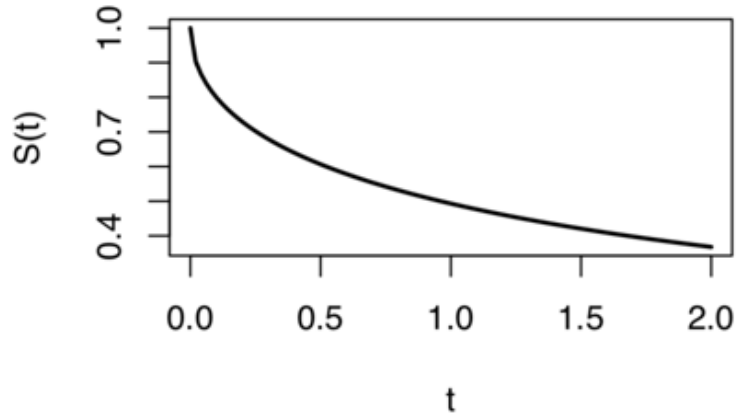
Terminology and notation

- T denotes the response variable, $T \geq 0$.
- The survival function is

$$S(t) = Pr(T > t) = 1 - F(t).$$

$S(t)$: survival functional → prob. ที่น่าจะอยู่รอดอย่างน้อยถึงเวลานั้น

$F(t)$: cumulative distribution of failure events



ช่วงเวลาที่คาดว่าจะอยู่รอด (อย่างน้อย)

Survival function

- The survival function gives the probability that a subject will survive past time *t*.
- As t ranges from 0 to ∞ , the survival function has the following properties
  - It is non-increasing
  - At time t= 0, S(t) = 1. In other words, the probability of surviving past time 0 is 1.
  - At time t=∞, S(t)=S(∞)=0. As time goes to infinity, the survival curve goes to 0.
- In theory, the survival function is smooth.
- In practice, we observe events on a discrete time scale (days, weeks, etc.).

prob. ณ เวลาใดเวลาหนึ่ง

- The hazard function, $h(t)$, is the instantaneous rate at which events occur, given no previous events.

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

- The cumulative hazard describes the accumulated risk up to time t, $H(t) = \int_0^t h(u)du$.

If we know any one of the functions $S(t)$, $H(t)$, or $h(t)$, we can derive the other two functions.

$$h(t) = -\frac{\partial \log(S(t))}{\partial t}$$

$$H(t) = -\log(S(t))$$

$$S(t) = \exp(-H(t))$$

Survival data

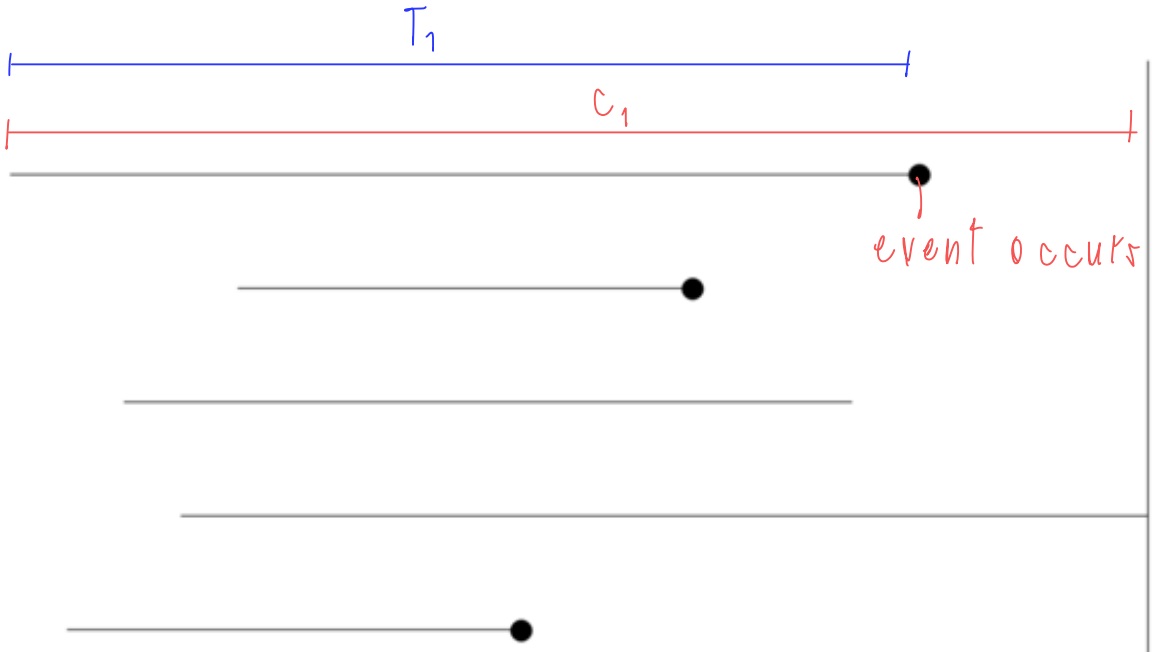
How do we record and represent survival data with censoring?

- T_i denotes the response for the ith subject.
- Let C_i denote the censoring time for the ith subject
- Let δ_i denote the event indicator

$$\delta_i = \begin{cases} 1 \text{ if the event was observed } (T_i \leq C_i) \\ 0 \text{ if the response was censored } (T_i > C_i). \end{cases}$$

- The observed response is $Y_i = \min(T_i, C_i)$.

Example



เริ่มต้น → ปลายเส้น

เริ่มต้น → ช่วงเวลาที่สนใจ

T_i	C_i	Y_i	δ_i
80	100	80	1
40	80	40	1
74+	74	74	0
85+	85	85	0
40	95	40	1

Termination of study

Estimating $S(t)$ and $H(t)$

If we are assuming that every subject follows the same survival function (no covariates or other individual differences), we can easily estimate $S(t)$.

- We can use nonparametric estimators like the Kaplan-Meier estimator

คล้าย histogram

- We can estimate the survival distribution by making parametric assumptions
  - exponential
  - Weibull
  - Gamma
  - log-normal

Non-parametric estimation of S

- When no event times are censored, a non-parametric estimator of $S(T)$ is $1 - F_n(t)$, where $F_n(t)$ is the empirical cumulative distribution function.

- When some observations are censored, we can estimate $S(t)$ using the Kaplan-Meier product-limit estimator.

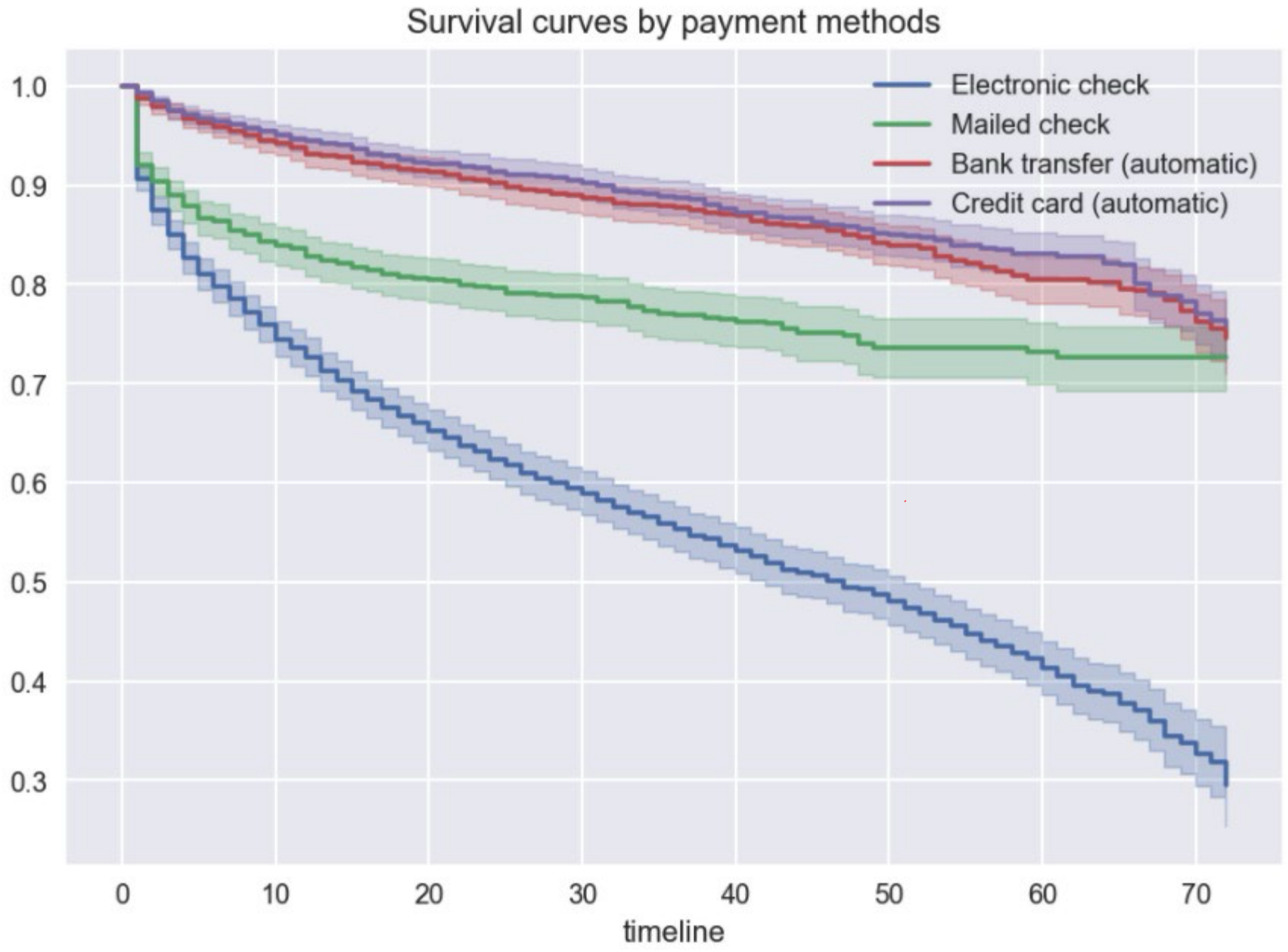
t	No. subjects at risk	Deaths	Censored	Cumulative survival
59	26	1	0	$25/26 = 0.962$
115	25	1	0	$24/25 \times 0.962 = 0.923 = \frac{24}{26}$
156	24	1	0	$23/24 \times 0.923 = 0.885$
268	23	1	0	$22/23 \times 0.885 = 0.846$
329	22	1	0	$21/23 \times 0.846 = 0.808$
353	21	1	0	$20/21 \times 0.808 = 0.769$
365	20	0	1	$20/20 \times 0.769 = 0.769$
377	19	0	1	$19/19 \times 0.769 = 0.769$
421	18	0	1	$18/18 \times 0.769 = 0.769$
431	17	1	0	$16/17 \times 0.769 = 0.688$
⋮				⋮
⋮				⋮

คนรอด

chain rule

y

t	S_i
10	1
27	0
50	0
⋮	⋮
200	0


Survival curves by payment methods
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
10
20
30
40
50
60
70
timeline

End of Lecture 4