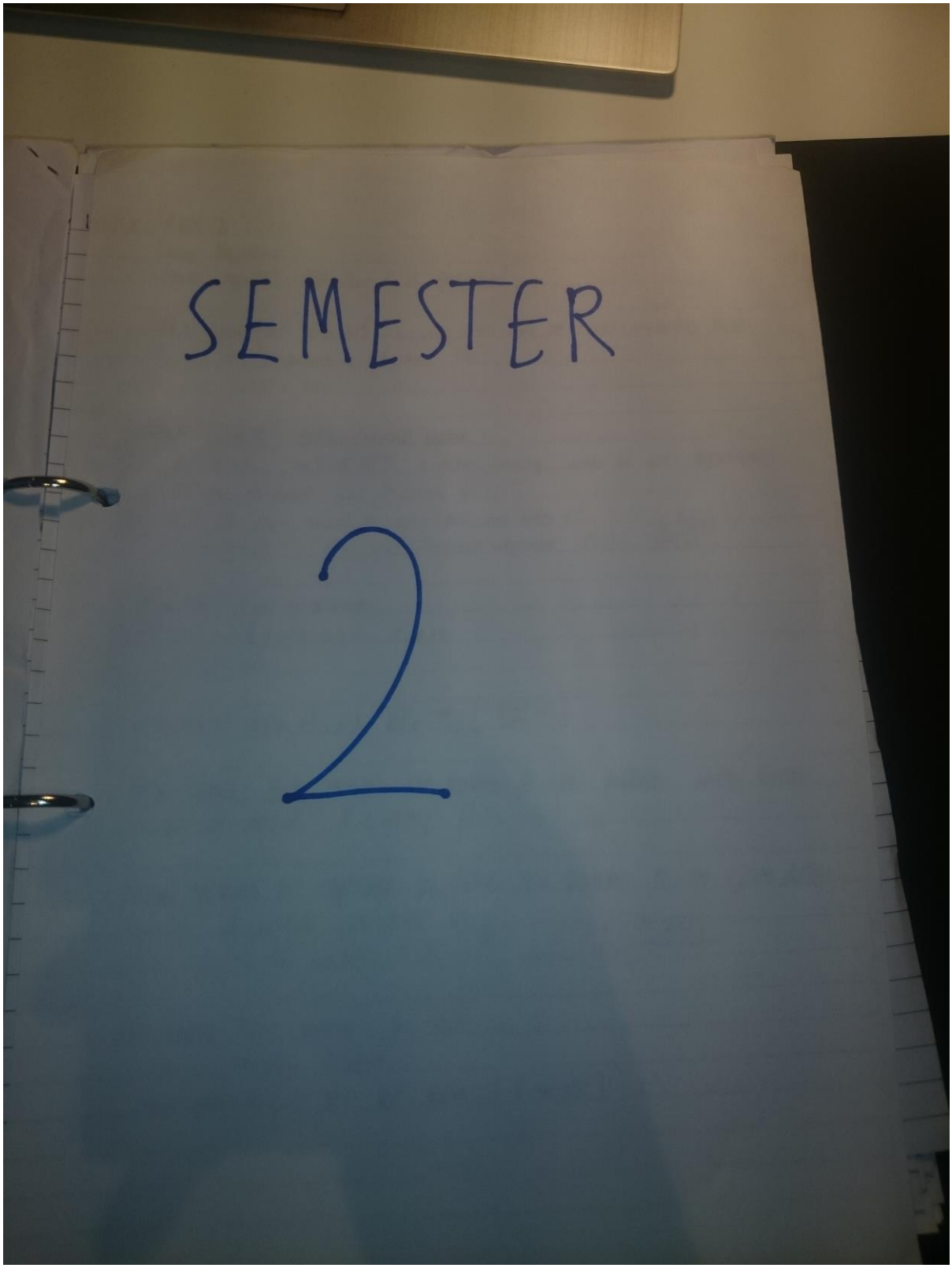

SEMESTER
2

19/6/16 ALSM 2

Linear Regression

Linear regression and how it is generalised

Set of observation $\{(y_i, x_i)\}_{i=1,\dots,n}$ such that the below relationship holds

$\forall i$: $y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_n x_{ki} + \varepsilon_i$

$= \beta^T x_i + \varepsilon_i$

Where $y_i \in \mathbb{R}$ outcome/response variable

$x_i = (1, x_{1i}, \dots, x_{ki}) \in \mathbb{R}^{k+1}$ a vector collating values of the explanatory variable associated with the outcome y_i

ε_i is the noise, residual or error associated with y_i

$\varepsilon_i \sim N(0, \sigma^2)$ $\forall i$ $i \neq j \Rightarrow$ residual independent $P(\varepsilon) = \frac{\exp\left(\frac{-\varepsilon^2}{2\sigma^2}\right)}{\sqrt{2\pi}\,\sigma}$

$P(y_i | \beta, x_i) \Rightarrow$ pdf or distribution

$\mathbb{E}[y_i] = \mathbb{E}[x_i^T \beta + \varepsilon_i] = x_i^T \beta + 0$

$Var[y_i] = \sigma^2$

$P(y_i | \beta, x_i) = N(y_i, \beta^T x_i, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[\frac{-(y_i - \beta^T x_i)^2}{2\sigma^2}\right]$

Best parameter β are estimated that maximize the joint probability density function of all the residuals : $\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\, p(\varepsilon_1, \dots, \varepsilon_n)$

Because residuals are independent and follow same distribution P_ε, the joint density function of residual) corresponds to

$p(\varepsilon_1, \dots, \varepsilon_n) = \prod_{i=1}^{n} P_\varepsilon(\varepsilon_i) = \prod \frac{\exp(-\frac{1}{2}\varepsilon_i^2)}{\sqrt{2\pi}\,\sigma} = \prod \frac{\exp\left[\frac{-(y_i - \beta^T x_i)^2}{2\sigma^2}\right]}{\sqrt{2\pi}\,\sigma}$

Can redefine linear regression $\forall i$ that :

- y_i is normally distributed
- With mean $\mathbb{E}[y_i] = \beta^T x_i$ and Variance $\mathbb{E}[(y_i - \beta^T x)^2] = \sigma^2$

19/01/

Generalized Linear Models (GLM)

- Will generalize these premises used for linear regression as follows:

- The probability density function $P_{y|x,\beta}$ is a member of the exponential family of distributions. Other distributions available to deal when outcome y is not an element of $\mathbb{R}$ but for instance is binary.

- Expectation of y given X and β is defined by:
  $\mathbb{E}[y] = \int_{\mathcal{Y}} y \, P_{y|x,\beta}(y|x,\beta)\, dy$ with $\mathcal{Y}$ domain of definition for outcome y

Now related to the explanatory variable(s) with a link function g such that $g[\mathbb{E}[y]] = \beta^T x$

For instance, in the case of linear regression, the link function g that we used is the identity function g defined for $z \in \mathbb{R} \Rightarrow g(z) = z \in \mathbb{R}$

g is a link function that is bijective and its inverse g^{-1} exists.
In general, this function maps the space of the expectation $\mathbb{E}[y]$ to the space $\mathbb{R}$ where $\beta^T x$ takes its value.

Probability Distributions

The pdf P_y of random variable y has the following properties:

- $\forall y \in \mathcal{Y}$ $P_y(y) \geq 0$: function P_y is positive for all possible outcomes $\mathcal{Y}$ used as space of all possible outcomes.

- The function P_y integrate to 1 on the space of all possible outcomes such that:
  - when y is a continuous r.v $\int_{\mathcal{Y}} P_y(y)\, dy = 1$
  - when y is discrete r.v $\sum_{y \in \mathcal{Y}} P_y(y) = 1$

2 19/01/16 ALSM2

Exponential Family of Distribution

- The distribution belongs to the exponential family if it can be written as:

$P_{y|\theta}(y|\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$

where a,b,c,d are known functions. If $a(y) = y$, then the distribution is said to be in canonical form

Expectation

- y continuous $\mathbb{E}[y] = \int_{\mathbb{R}} y\, P_y(y)\, dy$
- y Discrete $\mathbb{E}[y] = \sum_{y \in \mathbb{N}} y\, P_y(y)$

Normal / Gaussian Distribution

$y \in \mathbb{R}$ $P_{y|\theta}[y|\theta,\sigma^2] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(y-\theta)^2}{2\sigma^2}\right]$ $\theta \in \mathbb{R}$ $\sigma \in \mathbb{R}^+$ $\mathbb{E}[y] = \theta$

$$\exp\left[\frac{-(y-\mu)^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right]$$

$$\exp\left[\frac{-y^2}{2\sigma^2} + \frac{2y\mu}{2\sigma^2} + \underbrace{\frac{\mu^2}{2\sigma^2}} - \log(\sqrt{2\pi}\sigma)\right]$$

$d(y)$ $a(y)b(\theta)$ $c(\theta)$

Poisson Distribution

Expresses the probability of a given number of events y occuring in a fixed interval of time and/or space (and/or fixed total population size)

$$P_{y|\theta}(y|\theta) = \frac{\theta^y \exp(-\theta)}{y!} \quad y \in \mathbb{N} \quad \theta \in \mathbb{R}^+$$

$$\mathbb{E}[y] = \sum_{y \in \mathbb{N}} y \frac{\theta^y}{y!} \exp(-\theta) = \theta$$

$\exp(-\theta)\theta \sum_{y=1}^{\infty} \frac{\theta^{y-1}}{(y-1)!}$ taylor expansion

$= \exp(\theta)\exp(\theta) = \theta$

θ is the parameter used to plug in information from explanatory variable x

$\theta \Rightarrow (x^T\beta)$ somehow associated

$= \frac{1}{x!} \exp[x \log\lambda - \lambda]$

$= \frac{1}{x!} \exp(x \log\lambda) \exp(-\lambda)$

$d(y) \quad a(y)b(\theta) \quad c(\theta)$

Bernulli

y is binary variable, $P_{y|\theta}[y=1|\theta] = \theta \quad P_{y|\theta}(y=0|\theta) = 1-\theta$

$P_{y|\theta} = \theta^y (1-\theta)^{1-y} \quad y \in [0,1] \quad \theta \in [0,1] \quad \mathbb{E}[y] = \theta$

Binomial

Response y is # success in n trials, proportion θ real number 0,1

$P_{y|\theta}(y|\theta) = \frac{n!}{(n-y)!y!} \theta^y (1-\theta)^{n-y} \Rightarrow \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \mathbb{E}[y] = n\theta$

Exponential

$P_{y|\theta}(y|\theta) = \theta \exp(-\theta y) \quad y \in \mathbb{R}^+ \quad \theta \in \mathbb{R}^{+*}$ (and not ∪) $\quad \mathbb{E}[y] = \frac{1}{\theta}$

$\exp(-\theta y + \log\theta)$

$b(\theta)a(y) \quad c(\theta) \quad A(y)=0$

$\int_0^{+\infty} P(y|\theta)\,dy = 1 \quad \int_0^{\infty} \theta \exp(-\theta y)\,dy : [-\exp^{-\theta y}]_0^{\infty}$

$-\exp(-\infty) + \exp(0) = 0+1 = 1$

Weibull

$P_{y|\lambda,\theta}(y|\lambda,\theta) = \lambda\theta y^{\lambda-1} \exp[-\theta, y^{\lambda}] \quad y \in \mathbb{R}^+ \quad \theta \in \mathbb{R}^{+*} \quad \lambda \in \mathbb{R}^{+*}$

Expectation is $\mathbb{E}[y] = \left(\frac{1}{\theta}\right)^{1/\lambda} \Gamma\left(1+\frac{1}{\lambda}\right)$

$\Gamma(u) = \int_0^{+\infty} s^{u-1} \exp(-s)\,ds$

Exp is a special case of weibull with $\lambda = 1$

GLM

1. We have collated independently a set of responses y_i aswell as the values for some explanatory variables stored in vector x_i i.e have observations $\{(y_i, x_i)\}_{i=1,\dots,n}$

2. Response y_i has a distribution $p_{y|\theta}(\theta_i|\theta_i)$ that is a member of exponential family, indexed by parameter θ and related to expectation of response $\mathbb{E}[y_i]$

3. Model constructed by linking expectation of response $\mathbb{E}[y_i]$ with the linear predictor $x_i^T\beta$ $\quad g(\mathbb{E}[y_i]) = x_i^T\beta$
$\mathbb{E}[y_i] = g^{-1}(x_i^T\beta)$

4. Link function g is a monotonic differentiable function (ensures) inverse g^{-1} exists)

5. Estimate β by $\hat{\beta}$ = argmax likelihood or argmax posterior probability.

6. $\mathbb{E}[y] = g^{-1}(x^T\hat{\beta})$

Linear regression: Natural link function g is the identity $\theta \in \mathbb{R}$
Poisson " : " " is the log $\theta \in \mathbb{R}^{+*}$
Binomial " : " " is logit funcln $\theta \in [0,1]$ $g(\theta) = \ln\left[\frac{\theta}{1-\theta}\right]$
Survival Analysis : " " will be log function

These proposed link functions relate to the function $b(\theta)$ defined for distributions in canonical form in exp family of distributions. Other link functions can be used

Survival Analysis - weibull, exponential $y \in \mathbb{R}^+$
Linear Regression - Normal $y \in \mathbb{R}$
Poisson Regression - Poisson $y \in \mathbb{N}$
Logistic Regression - Binomial $y \in (0, 1, \dots n)$

likelihood = cost function corresponds to joint dist of all values given parameters needed

$P(y_1, y_n | \theta_1, \theta_n) = \prod_{i=1}^{n} P(y_i|\theta_i)$ independence of response variable y_i only concerned with θ_i, other θ's don't give info on y_i

$\forall i \quad y_i \sim P_{y|\theta}(y|\theta)$

likelihood $= P(y_1, y_n | \theta_1, \theta_n) = \prod_{i=1}^{N} P_{y|\theta}(y_i|\theta_i)$
Take log likelihood: $\sum_{i=1}^{N} \log(P_{y|\theta}(y|\theta))$

$P_{y|\theta}(y|\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$
$\text{Log}(L) = \sum_{i=1}^{N} a(y_i)b(\theta_i) + c(\theta_i) + d(y_i)$

each $\theta_i \Rightarrow g^{-1}(x_i^T \beta)$ link function
$\beta = \text{argmax } \log(L)$

6

26/01/16 ALSM 2

LOGISTIC REGRESSION

Binomial Dist: $P(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \quad y \in (0,1,\dots n)$

$P(y|\theta) = \exp\left[\log\left[\binom{n}{y}\theta^y(1-\theta)^{n-y}\right]\right]$

$= \exp\left[\log\binom{n}{y} + y\log(\theta) + (n-y)\log(1-\theta)\right]$

$= \exp\left[\log\binom{n}{y} + y\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right]$

$\quad\quad d(y) \quad\quad a(y)\ b(\theta) \quad\quad c(\theta)$

Value of θ that maximises $P(y|\theta)$?

$\frac{d\,P(y|\theta)}{d(\theta)} = \binom{n}{y}\left[y\theta^{y-1}(1-\theta)^{n-y} - \theta^y(n-y)(1-\theta)^{n-y-1}\right]$

$= \binom{n}{y}\left[\theta^{y-1}(1-\theta)^{n-y-1}\right]\left[y(1-\theta) - \theta(n-y)\right]$

$y(1-\theta) - \theta(n-y) = 0$

$y = n\theta \quad$ or $\quad \theta = y/n$

Expected Value

$\mathbb{E}[y] = \sum_{y=0}^{n} y\binom{n}{y}(\theta^y)(1-\theta)^{n-y}$

$= \sum_{y=1}^{n} y\binom{n}{y}|\theta|^y(1-\theta)^{n-y}$ can't have $y=0$ or else have 0.

$y\binom{n}{y} = \frac{y\,n!}{(n-y)!\,y!} = \frac{y\cdot n\cdot(n-1)!}{(n-1-(y-1))!\,y(y-1)!} = n\binom{n-1}{y-1}$

$= n\theta \underbrace{\sum_{y=0}^{n-1}\binom{n-1}{y-1}\theta^{y-1}(1-\theta)^{(n-1-(y-1))}}_{=\mathrm{Bin}(\theta,m)} = n\theta$

$y_1 \sim p(\theta_1, n_1)$

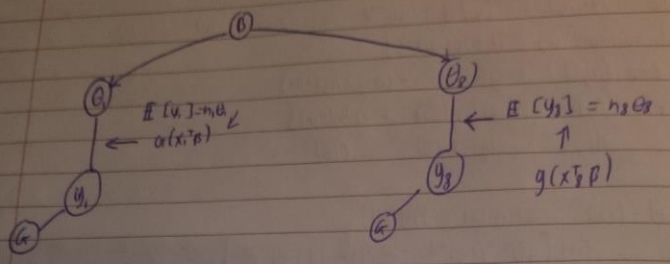
$y_2 \sim p(\theta_2, n_2)$

$\vdots$

$y_n \sim p(\theta_n, n_n)$

26/01/16

Saturated model / Unrestricted GLM



$L(\theta_1, \theta_2, \ldots, \theta_3) = P(y_1, \ldots, y_3 \mid \theta_1, \ldots, \theta_3)$

$= \prod_{i=1}^{3} P(y_i \mid \theta_i)$ independence

$= \prod_{i=1}^{3} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i - y_i}$

$(\hat{\theta}_1, \ldots, \hat{\theta}_3) = \text{argmax}\ L(\theta_1, \ldots, \theta_3)$

$\hat{\theta}_1 = y_1/n_1 \quad \ldots \quad \hat{\theta}_3 = y_3/n_3$

Can plot θ_i vs x_i

$\theta \in [0,1] \xrightarrow{g} x_i^T\beta \in \mathbb{R}$ has to be invertible to get back to g!

$\xleftarrow{g^{-1}} \beta_0 + \beta_1 x$

$g(\theta) = \log\left[\frac{\theta}{1-\theta}\right] = z \qquad z \in \mathbb{R}$

$[0,1] \to (-\infty, +\infty)$ called logit function

$\mathbb{R}$

$g^{-1}(z) = \frac{e^z}{1+e^z}$

Can also use the probit function Φ

$P(y \mid \beta, x) = \binom{n}{y} (g(\beta^T x))^y (1 - g(\beta^T x))^{n-y}$

$L(\beta) = \prod_{i=1}^{N} P(y_i \mid \beta x_i)$

$= \prod_{i=1}^{3} \binom{n_i}{y_i} [g(\beta^T x_i)]^{y_i} [1 - g(\beta^T x_i)]^{n_i - y_i}$ — logit / probit etc

26/01/16 ALSM2

Find β such that $d\mathcal{L}(\beta)/d(\beta) = 0$

NOTE: you can add a $\beta_2 X^2$ term if you want to add another dimension to model

Dist	Links	Possible Models
Binomial	Probit	$\beta_0 + \beta_1 X$
	Logit	$\beta_0 + \beta_1 X + \beta_2 X^2$
		β_0
		$\beta_1 X$
		$\beta_1 X + \beta_2 X^2$
		$\beta_0 + \beta_2 X^2$ etc

- Use AIC/BIC to determine "best" model
- Can calculate deviance to see if model is good

GLM Example

y_i : Patient Survive? 1: No 0: yes

$\{(x_i, y_i)\}$ $i = 1, \ldots, 40$

Response $y_i \sim$ Binomial $P(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$

$i = 1$ $P(y_1|\theta_1) = \binom{1}{y}\theta_1^y(1-\theta_1)^{1-y}$ (only 1 choice $\Rightarrow$ bernoulli)

$= \theta_1^{y_1}(1-\theta_1)^{1-y_1}$

$i = 2$ $= \theta_2^{y_2}(1-\theta_2)^{1-y_2}$

$i = 40$ $= \theta_{40}^{y_{40}}(1-\theta_{40})^{1-y_{40}}$

$E[y] = n\theta = \theta$ $n = 1$ (bernoulli)

$\frac{d\,L(y|\theta)}{d(\theta)} = 0$ $\Rightarrow \theta = y/n = y$

$L(\theta_1, \ldots, \theta_{40}) = \prod_{i=1}^{40} \theta_i^{y_i}(1-\theta_i)^{1-y_i}$ $\quad \dim(\vec{\theta}) = 40$

$(\theta_1^*, \ldots, \theta_{40}^*) = \operatorname{argmax} L(\theta_1, \ldots, \theta_{40})$

link function $= b(\theta) = g(\theta) = \log\left[\frac{\theta}{1-\theta}\right] \Rightarrow$ logit

GLM: $L(\beta) = \prod_{i=1}^{40} g^{-1}(\beta_0 + \beta_1 x_i)^{y_i}[1 - g^{-1}(\beta_0 + \beta_1 x_i)]^{1-y_i}$

$\dim(\beta) = 2$