Lecture 3: Binary and binomial regression models

Claudia Czado

TU München

Overview

- Model classes for binary/binomial regression data
- Explorative data analysis (EDA) for binomial regression data
  - main effects
  - interaction effects

Binary regression models

Data: $(Y_i, \mathbf{x_i})\ i = 1, \ldots, n$ Y_i independent

$Y_i = 1$ or 0

$\mathbf{x_i} \in \mathbb{R}^p$ covariates (known)

Model: $p(\mathbf{x_i}) := P(Y_i = 1 | \mathbf{X_i} = \mathbf{x_i})$

$\Rightarrow P(Y_i = 0 | \mathbf{X_i} = \mathbf{x_i}) = 1 - p(\mathbf{x_i})$

How to specify $p(\mathbf{x_i})$? We need $p(\mathbf{x_i}) \in [0, 1]$.

Example: Survival on the Titanic

Source: `http://www.encyclopedia-titanica.org/`

Name	Passenger Name
PClass	Passenger Class
Age	Age of Passenger
Sex	Gender of Passenger
Survived	Survived=1 means Passenger survived
	Survived=0 means Passenger did not survive

Model hierarchy

Model 1) $p(\mathbf{x}) = F(\mathbf{x})$ $F \in \{F : \mathbb{R}^p \to [0,1]\},\ F$ unknown

Model 2) $p(\mathbf{x}) = F(\boldsymbol{x_i^t \beta})$ $F \in \{F : \mathbb{R} \to [0,1]\},\ F$ unknown

Model 3) $p(\mathbf{x}) = F(\boldsymbol{x_i^t \beta})$ $F \in \{F : \mathbb{R} \to [0,1]\ cdf\},\ F$ unknown

Model 4) $p(\mathbf{x}) = F_0(\boldsymbol{x_i^t \beta})$ F_0 known cdf

Model properties

Model 1: - simple interpretation of covariate effects not possible
- estimation of p-dimensional F difficult
  $\rightarrow$ smoothing methods
  (O'Sullivan, Yandell, and Raynor (1986), Hastie and Tibshirani (1999))

Model 2: - estimation of F now one dimensional, but additional estimation for β needed
- Interpretation of covariate effects remains difficult

Model 3: - Since cdf's are monotone, covariate effects are easily interpretable
- Different classes for cdf's F can be chosen:

Parametric Approach: Link Families

$$\mathcal{F} = \{F(\cdot, \psi),\ \psi \in \Psi,\ F(\cdot, \psi)\ cdf,\ F(\cdot, \cdot) \text{ known}\}$$

$\rightarrow$ ψ link parameter
$\rightarrow$ joint estimation of $\boldsymbol{\beta}$ and ψ is needed

Example:

(Czado 1997)

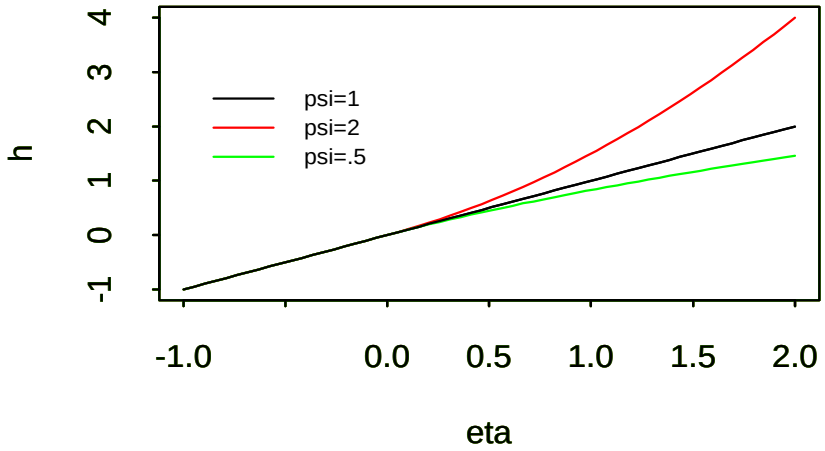
$$F(\eta, \boldsymbol{\psi}) = \frac{e^{h(\eta, \boldsymbol{\psi})}}{1+e^{h(\eta, \boldsymbol{\psi})}}$$

$$h(\eta, \boldsymbol{\psi}) = \begin{cases} \frac{(\eta+1)^{\psi_1}-1}{\psi_1} & \eta \geq 0 \\ -\frac{(-\eta+1)^{\psi_2}-1}{\psi_2} & \eta < 0 \end{cases} \qquad \boldsymbol{\psi} = (\psi_1, \psi_2)$$

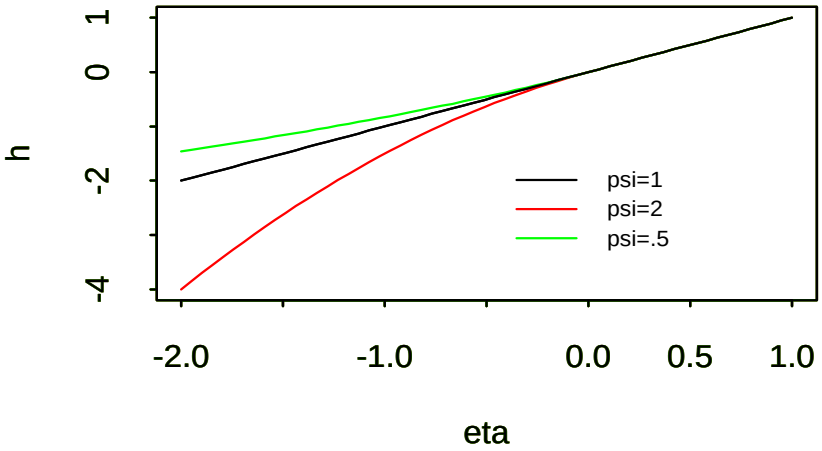
Both tail family

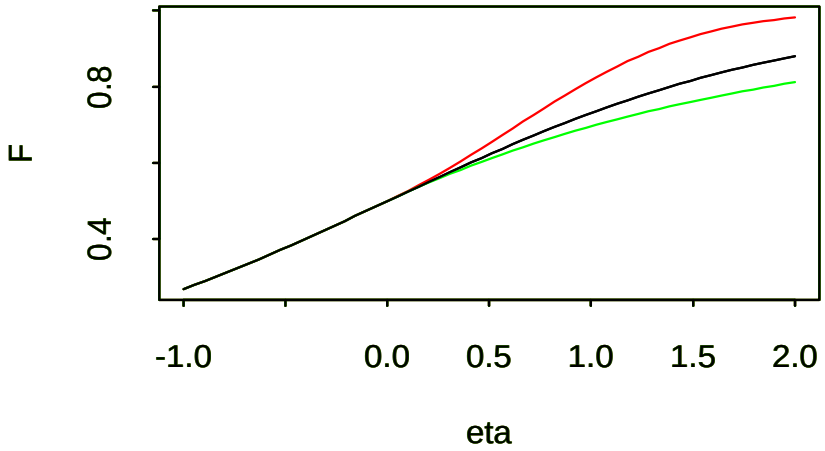
$\boldsymbol{\psi} = (1, 1)$ corresponds to logistic regression
$\boldsymbol{\psi} = (\psi_1, 1)$ Right tail family
$\boldsymbol{\psi} = (1, \psi_2)$ Left tail family

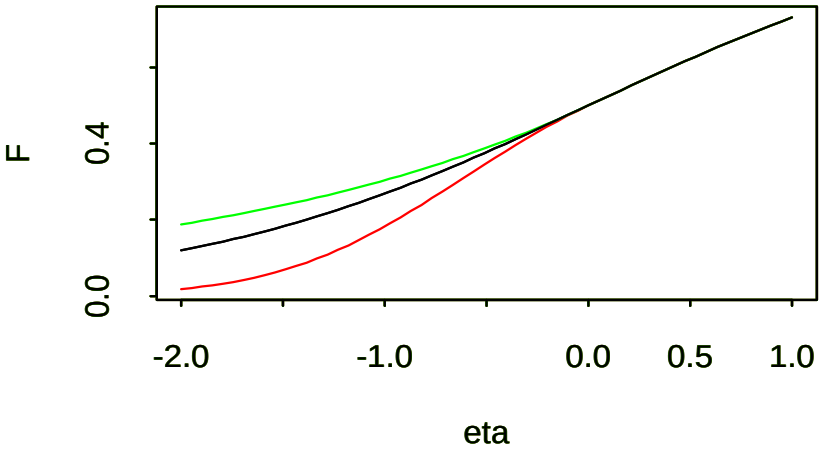
Right Tail Modification



Left Tail Modification







Nonparametric Approach

- Klein and Spady (1993)

- Bayesian approach: need a prior for the class of cdf's, i.e. a stochastic process such as the Dirichlet process. Markov Chain Monte Carlo (MCMC) methods are required to estimate the posterior distribution (see Newton, Czado, and Chappell (1996))

Restriction to cdf's can be justified by the threshold approach:

$$Y_i = 1 \Leftrightarrow \boldsymbol{x_i^t \beta} \geq U_i \quad \text{where } U_i \sim F \ i.i.d.$$

$$\Rightarrow P(Y_i|\mathbf{X_i} = \mathbf{x_i}) = P(U_i \leq \boldsymbol{x_i^t \beta}) = F(\boldsymbol{x_i^t \beta})$$

Model 4:

- Most common and simplest model, however gives not always the best fit (link misspecification)

- **Examples:**

  - $F(\eta) = \frac{e^\eta}{1+e^\eta}$ logistic regression
  - $F(\eta) = \Phi(\eta)$ probit regression
  - $F(\eta) = 1 - \exp\{-exp\{\eta\}\}$ complementary log-log regression

Logistic regression

$$Y_i|\mathbf{X_i} = \mathbf{x_i} \sim \text{ binary}(p(\mathbf{x_i})) \text{ independent}$$

$$p(\mathbf{x_i}) = P(Y_i = 1|\mathbf{X_i} = \mathbf{x_i}) = \frac{e^{\boldsymbol{x_i^t\beta}}}{1+e^{\boldsymbol{x_i^t\beta}}}$$

Binary response can be extended to binomial response:

$$\begin{aligned} & Y_i \sim \text{ bin}(n_i, p(\mathbf{x_i})) \text{ ind.} \\ \Rightarrow \quad & P(Y_i = y_i|\mathbf{X_i} = \mathbf{x_i}) = \left(\begin{array}{c} n_i \\ y_i \end{array} \right) p(\mathbf{x_i})^{y_i}(1 - p(\mathbf{x_i}))^{n_i - y_i}, \end{aligned}$$

i.e. $\left\{\frac{Y_i}{n_i}\right\}$ is a GLM with canonical link.

Explorative data analysis (EDA) for binomial regression data

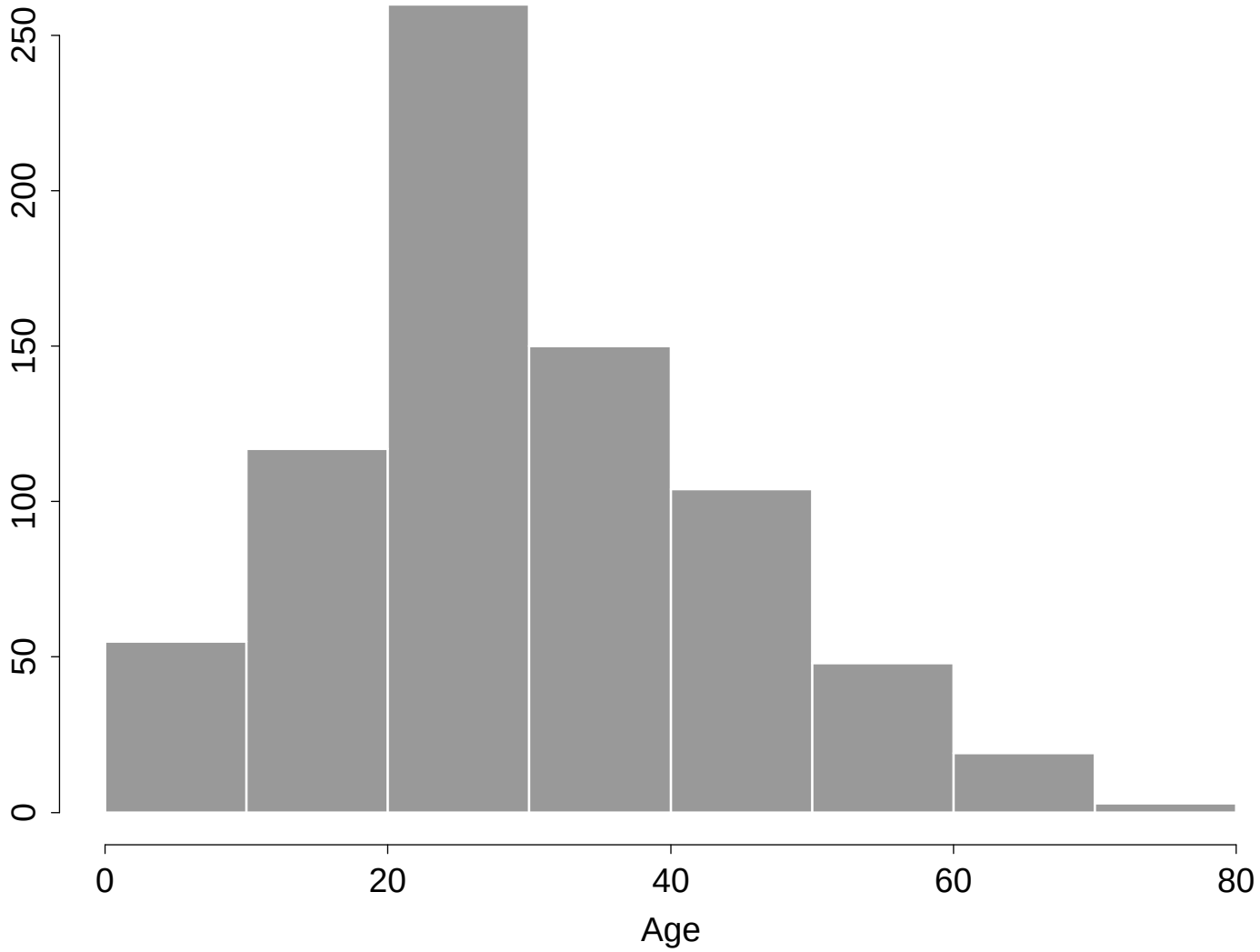
Data: $(Y_i, \mathbf{x_i})$, $\mathbf{x_i} = (x_{i1}, \ldots, x_{ik})$ k potentially important covariates.

Problem: Variable selection.
With many covariates one needs screening methods, such as EDA.

covariates	qualitative / categorical	dichotomous	(2 levels)
		polytomous	(J levels)
		ordinal	(J levels)
	quantitative		

Example: Titanic data summaries

```
> attach(titanic)
> table(PClass)
 1st 2nd 3rd
 322 280 711
> table(Sex)
 female male
    462  851
> table(Survived)
   0   1
 863 450
```


0
50
100
150
200
250
0
20
40
60
80
Age

```
> table(Survived,PClass)
  1st 2nd 3rd
0 129 161 573
1 193 119 138
> table(Survived,Sex)
  female male
0    154  709
1    308  142
```

Third Class and male passengers survived less often then other class or female passengers.

Influence of single covariate on $p(\mathbf{x})$

Dichotomous covariate.

Data

Status	Gender	
	female	male
not survived	154	709
survived	308	142

Want to estimate

Y	X	
	0	1
0	$1-p(0)$	$1-p(1)$
1	$p(0)$	$p(1)$

Logistic model: $p(x) = P(Y=1|X=x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ $\qquad x = 0, 1$

o:= $\frac{p}{1-p}$ "odds of success", p = success probability

$logit(p) := \log(o) = \log\left(\frac{p}{1-p}\right)$ Log odds

Influence of single covariate on $p(\mathbf{x})$

$$logit(p(1)) = \log\left(\frac{e^{\beta_0+\beta_1 x}/(1+e^{\beta_0+\beta_1 x})}{1/(1+e^{\beta_0+\beta_1 x})}\right) = \log\left(e^{\beta_0+\beta_1}\right) = \beta_0+\beta_1$$

$$logit(p(0)) = \log(e^{\beta_0}) = \beta_0$$

} linear in $x = 0, 1$

$\psi := \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}$ "odds ratio"

$p(1) \approx 0, p(0) \approx 0 \Rightarrow \psi \approx \frac{p(1)}{p(0)}$ – relative risk

Odds ratio as dependency measure

Data

Y	X	
	0	1
0	a	b
1	c	d

Conditional distribution

Y	X	
	0	1
0	$1-p(0)$	$1-p(1)$
1	$p(0)$	$p(1)$

$p(j) = P(Y = 1|X = j) \quad j = 0, 1$ conditional distribution.

Want to see how $p(j)$ is changing.

$\psi = \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}$ measures change in conditional distributions.

$\psi = 1 \Leftrightarrow Y$ and X are independent

Unstructured model

$$\hat{p}^{obs}(\mathbf{x}) := \frac{\text{number of obs. with } Y = 1 \text{ and } \mathbf{X} = \mathbf{x}}{\text{number of obs. with } \mathbf{X} = \mathbf{x}} \qquad x = 0, 1$$

$$\hat{p}^{obs}(1) = \tfrac{d}{b+d} \quad \hat{p}^{obs}(0) = \tfrac{c}{a+c} \Rightarrow \hat{\psi}^{obs} = \frac{\hat{p}^{obs}(1)/(1-\hat{p}^{obs}(1))}{\hat{p}^{obs}(0)/(1-\hat{p}^{obs}(0))} = \frac{da}{bc}$$

$$\Rightarrow \widehat{\log(\psi)}^{obs} = \log(\hat{\psi}^{obs}) \quad \text{(est. log odds ratio)}$$

$$\widehat{Var}(\widehat{\log(\psi)}^{obs}) \approx \left(\tfrac{1}{a} + \tfrac{1}{b} + \tfrac{1}{c} + \tfrac{1}{d}\right) \quad \text{(est. var. of } \widehat{\log(\psi)}^{obs}\text{)}$$

$100(1-\alpha)\%\ CI$ for $\log\psi$: $\quad \log\hat{\psi}^{obs} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\log(\hat{\psi}^{obs}))}$

$100(1-\alpha)\%\ CI$ for ψ : $\quad \left(e^{\log\hat{\psi}^{obs} - z_{\alpha/2}\sqrt{\widehat{Var}(\log(\hat{\psi}^{obs}))}}, e^{\log\hat{\psi} + z_{\alpha/2}\sqrt{\widehat{Var}(\log(\hat{\psi}^{obs}))}}\right)$

Example: Survival on the Titanic

$$Y = \begin{cases} 1 & \text{survived} \\ 0 & \text{not survived} \end{cases} \qquad X = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$$

Y	X	
	0	1
0	154	709
1	308	142

$\hat{p}(0) = \frac{308}{154+308} = 0.67$ — 67% of females have survived

$\hat{p}(1) = \frac{142}{709+142} = 0.17$ — 17% of males have survived

$\hat{o}(0) = \frac{\hat{p}(0)}{1-\hat{p}(0)} = \frac{0.67}{1-0.67} = 2$ — Women survived twice as often as not to survive

$\hat{o}(1) = \frac{\hat{p}(1)}{1-\hat{p}(1)} = 0.2 = \frac{1}{5}$ — Men did not survive 5 times as often as to survive

$\Rightarrow \quad \hat{\psi} = \frac{\hat{o}(1)}{\hat{o}(0)} = \frac{0.2}{2} = 0.1 = \frac{1}{10}$ — Women had 10 times higher odds to survive compared to men

Polytomous covariate

Y	Category of X		
	1	$\cdots$	J
0	$1-p(1)$	$\cdots$	$1-p(J)$
1	$p(1)$	$\cdots$	$p(J)$

Nominal categories

(Example: car marks: BMW, VW, Ford: *unordered*)

Model: $p(j) := P(Y = 1|X = j) = \frac{e^{\beta_0+\beta_1 I_1(i)+\cdots+\beta_{J-1} I_{J-1}(i)}}{1+e^{\beta_0+\beta_1 I_1(i)+\cdots+\beta_{J-1} I_{J-1}(i)}}$

$$I_j(i) = \begin{cases} 1 & x_i = j \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \ldots, J-1 \qquad \textit{dummy coding}$$

$$\Rightarrow \quad p(J) = \frac{e^{\beta_0}}{1+e^{\beta_0}} \qquad \rightarrow \qquad \beta_0 \quad \text{parametrizes} \quad logit(p(J))$$

Only $J-1$ dummy variables are used to avoid a non full rank design matrix.

$$\Rightarrow \psi_j := \frac{p(j)/(1-p(j))}{p(J)/(1-p(J))} = e^{\beta_j} \quad \forall j = 1, \ldots, J-1$$

J is reference category.
If $\psi_1 = \ldots = \psi_{J-1}$: constant odds ratio

Example: Survival on the Titanic

Consider Pclass as nominal covariate

Y	Pclass		
	1	2	3
0	129	161	573
1	193	119	138
$\hat{p}^{obs}(j)$	$\frac{193}{129+193} = 0.60$	0.42	0.19
$\hat{o}^{obs}(j)$	$\frac{0.6}{1-0.6} = 1.50$	0.72	0.23
$logit(\hat{p}^{obs}(j))$	0.41	-0.33	-1.50
$\hat{\psi}^{obs}(j)$	$\frac{1.5}{0.23} = 6.50$	3.10	

First (second) class passengers had a 6.5 (3.1) times higher odds to survive compared to third class passengers $\longrightarrow$ dependence between class and survival status

Ordinal categories

Examples: marks: A, B, C, D, E; age groups.

Ordinal categories result often from grouping quantitative data.

Two coding possible:

-use dummy variables as with nominal categories

-use scores

Example:

	Age groups			
	$20-34$	$35-44$	$45-54$	$55-64$
$s(j)$ *scores (means)*	27	39.5	49.5	59.5

$$p(j) = P(Y=1|X=j) = \frac{e^{\beta_0+\beta_1 s(j)}}{1+e^{\beta_0+\beta_1 s(j)}} \Rightarrow \log\left(\frac{p(j)}{1-p(j)}\right) = \beta_0 + \beta_1 s(j)$$

If there is no functional relationship between j and $\hat{p}^{obs}(j)$ (or $\log\left(\frac{\hat{p}^{obs}(j)}{1-\hat{p}^{obs}(j)}\right)$) then a dummy coding is more appropriate.

Quantitative covariates

Binomial model: $Y_i|X_i = x_i \sim bin(n_i, p(x_i))$ independent
For a logistic model we have

$$logit(p(x_i)) = \beta_0 + \beta_1 x_i$$

This model is appropriate if $logit(\hat{p}^{obs}(x_i))$ linear in x.

Problem:

If $Y_i = 0$ or $Y_i = n_i$ we have

$$\log\left(\frac{\hat{p}^{obs}(x_i)}{1-\hat{p}^{obs}(x_i)}\right) = \log\left(\frac{Y_i/n_i}{1-Y_i/n_i}\right) = \log\left(\frac{Y_i}{n_i-Y_i}\right) \quad \text{undefined}$$

Consider therefore

empirical logits: $l_{x_i} := \log\left(\frac{Y_i+1/2}{n_i-Y_i+1/2}\right)$

Bernoulli model: $Y_i|X_i = x_i \sim bern(p(x_i))$ independent

$$\Rightarrow l_{x_i} = \begin{cases} \log\left(\frac{3/2}{1/2}\right) = \log(3) \approx 1.1 \\ \log\left(\frac{1/2}{3/2}\right) = \log(1/3) \approx -1.1 \end{cases}$$

Need smoothing to interpret the plot of x_i versus l_{x_i}

Other approach: group data to achieve an binomial response with an ordinal covariate. Proceed as before.

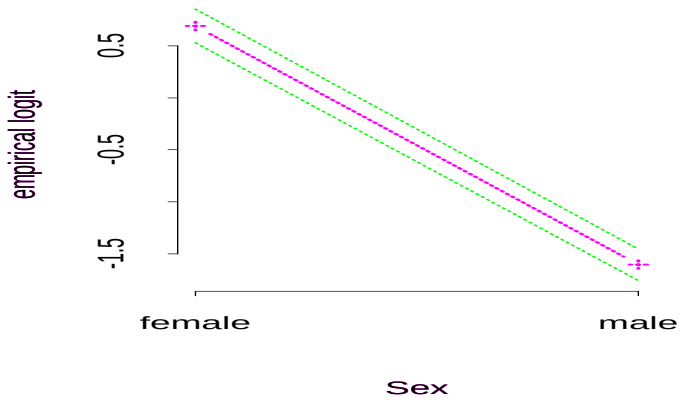
Titanic EDA for each covariate

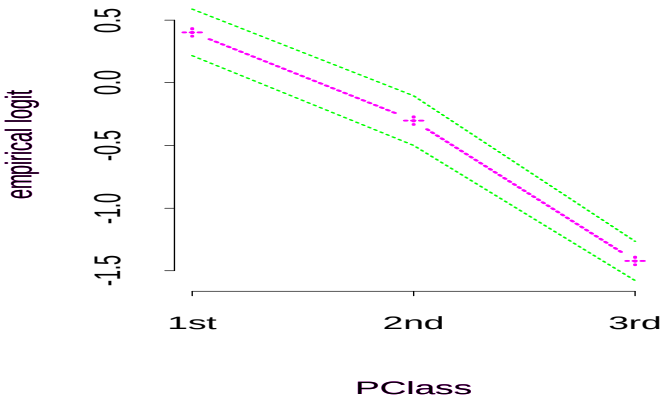
The Splus function `main1.plot()` calculates empirical logits and plots them together with pointwise 95% Confidence limits.

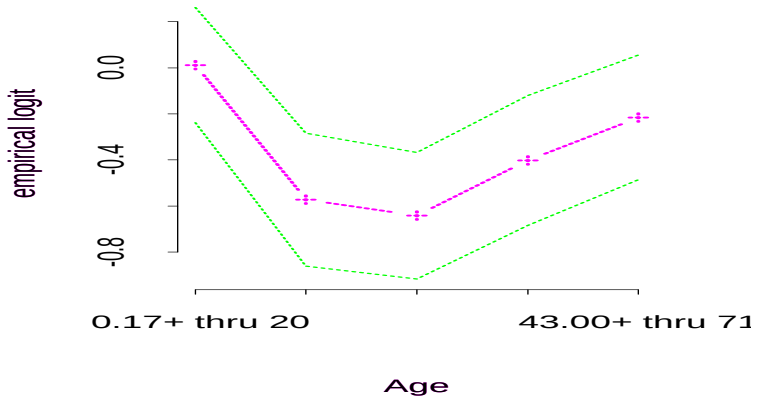
```
> titanic.main    # Splus code
function(ps = F)
{
  Age.cut <- cut(Age, breaks = quantile(Age, probs =
              c(0, 0.2, 0.4, 0.6, 0.8, 1), na.rm = T))
  if(ps == T) {
   ps.options(colors = ps.colors.rgb[c("black", "cyan",
    "magenta", "green", "MediumBlue", "red"),  ],
    horizontal = F)
   postscript(file = "titanic.main.ps")
  }
  par(mfrow = c(2, 2))
  main1.plot(Survived, Sex, "Sex")
  main1.plot(Survived, PClass, "PClass")
  main1.plot(Survived, Age.cut, "Age")
}
```

```
> titanic.main()
Main Effects for  Sex
           female    male
emp. logit   0.69   -1.61
         n 462.00  851.00

Main Effects for  PClass
             1st    2nd     3rd
emp. logit   0.4   -0.3   -1.42
         n 322.0  280.0  711.00

Main Effects for  Age
            0.17+ thru 20 20.00+ thru 25 25.00+ thru 32
emp. logit           0.01          -0.57          -0.64
         n         171.00         139.00         157.00
           32.00+ thru 43 43.00+ thru 71
emp. logit           -0.4          -0.22
         n           140.0         148.00
```

For quantitative covariates a grouped variable using quintiles are used.

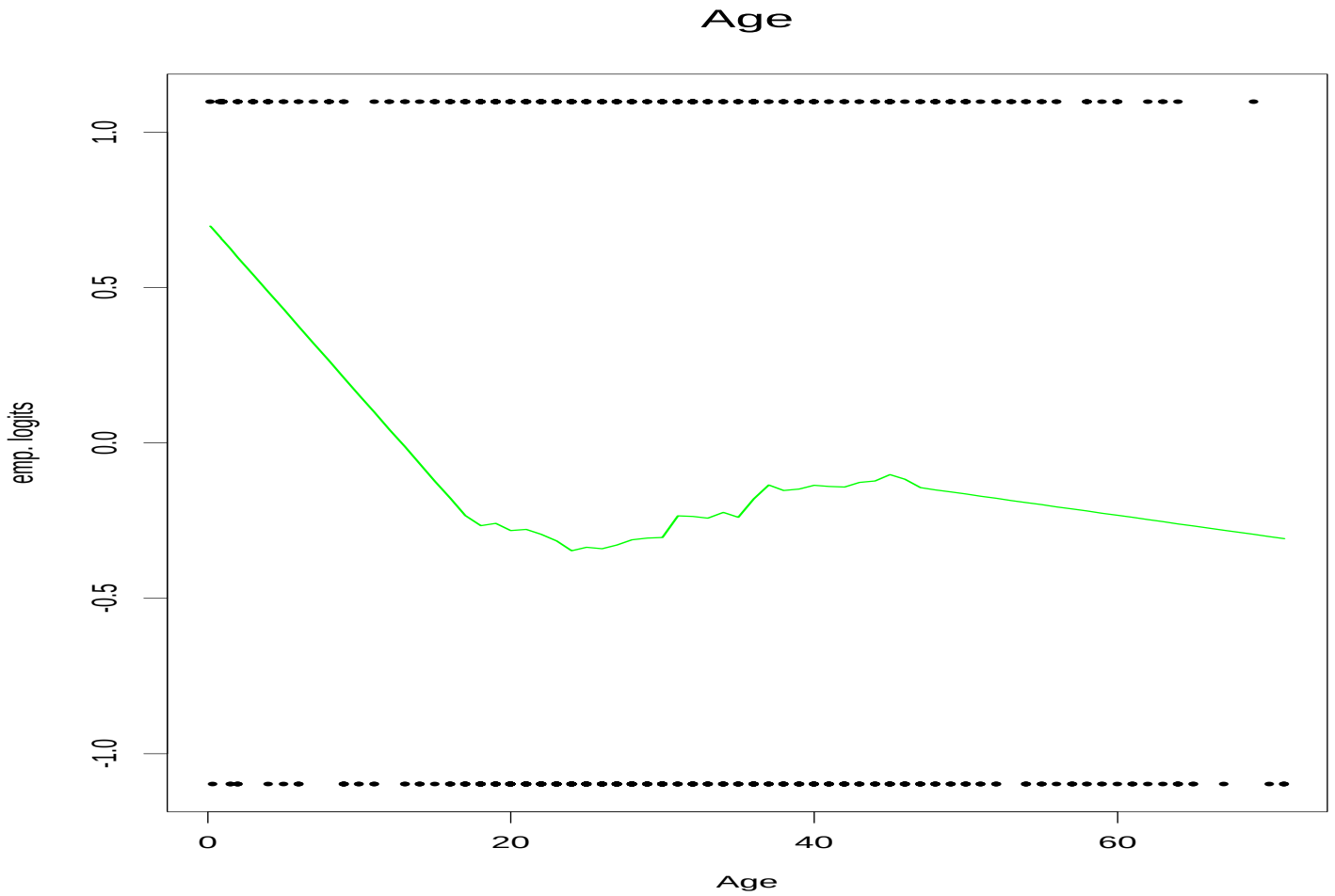






Quadratic Effect of Age?

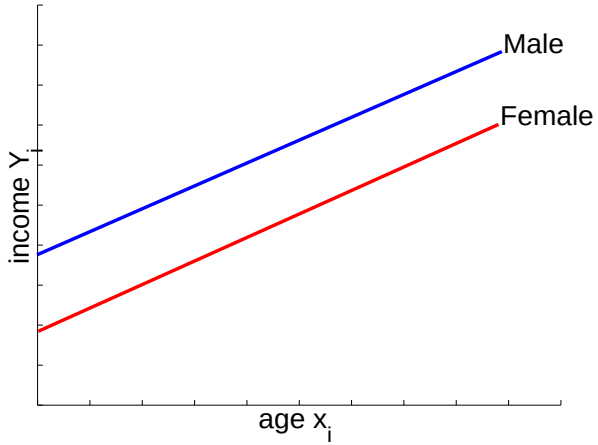
Smoothed empirical logits for binary Responses:



Indicates nonlinear Age Effect, but maybe not quadratic

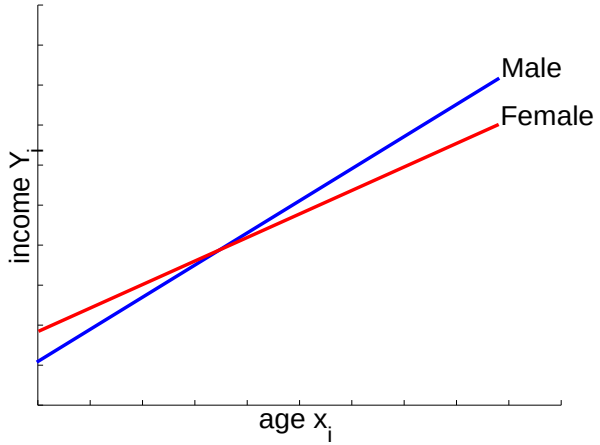
Influence on $p(\mathbf{x})$ of several covariates

Linear models: one quantitative/one dichotomous



No interaction: difference in income independent of age

same slopes, different intercepts



Interaction: difference in income dependent of age

different slopes + intercepts

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 x_i \cdot D_i + \epsilon_i \qquad D_i = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$$

i male: $Y_i = \beta_0 + \beta_1 x_i + \beta_2 + \beta_3 x_i + \epsilon_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i + \epsilon_i$

i female: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Testing for interaction: $H_0 : \beta_3 = 0 \qquad H_1 : \beta_3 \neq 0$

If second covariable is polytomous with J levels use

$$
\begin{aligned}
D_{1i} &= \begin{cases} 1 & \text{obs. } i \text{ has category } 1 \\ 0 & \text{otherwise} \end{cases} \\
&\vdots \\
D_{(J-1)i} &= \begin{cases} 1 & \text{obs. } i \text{ has category } J-1 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

For interactions add terms $x_i D_{1i}, \ldots, x_i D_{(k-1)i}$. Note category J is the reference category here.

If second covariate is quantitative use

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} \cdot x_{i2} + \epsilon_i$$

to model interaction.

Discovering interactions in logistic regression:

- Since the logits should be linear in the covariates one can look for non parallel lines when empirical logits are used.

- Confidence bands should be considered, when assessing non parallelity.

EDA of interaction effects for the Titanic data

```
> titanic.inter()
Interaction Effects for  Sex and PClass
Empirical Logit

           PClass.1st PClass.2nd PClass.3rd
Sex.female       2.65       1.95      -0.50
  Sex.male      -0.71      -1.76      -2.02

Cell Sizes

           PClass.1st PClass.2nd PClass.3rd
Sex.female        143        107        212
  Sex.male        179        173        499
```

Interaction Effects for Sex and Age

Empirical Logit

	Age. 0.17+ thru 20	Age.20.00+ thru 25
Sex.female	0.80	0.77
Sex.male	-0.68	-1.56
	Age.25.00+ thru 32	Age.32.00+ thru 43
Sex.female	0.81	1.50
Sex.male	-1.38	-1.65
	Age.43.00+ thru 71	
Sex.female	1.93	
Sex.male	-1.58	

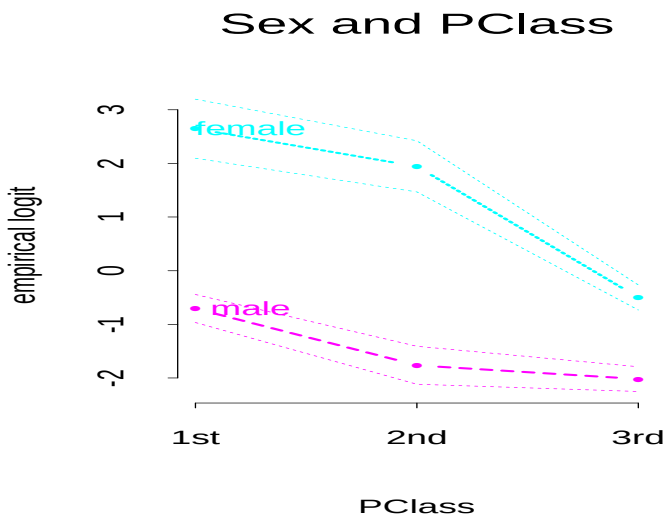
Cell Sizes

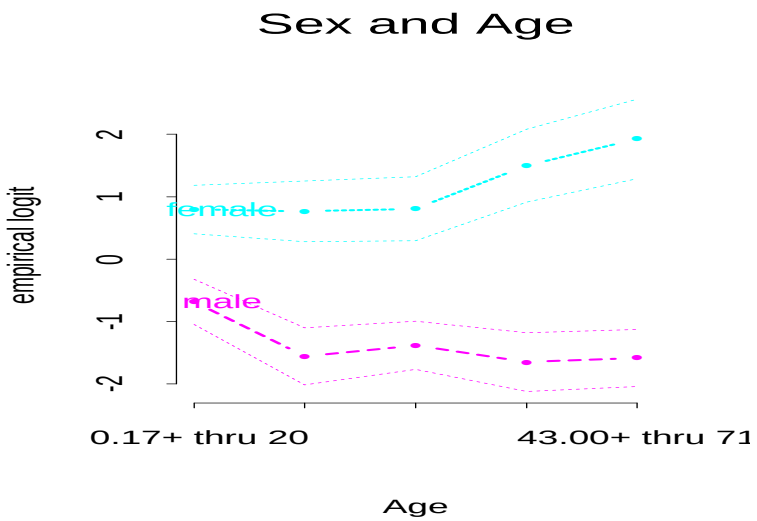
	Age. 0.17+ thru 20	Age.20.00+ thru 25
Sex.female	81	51
Sex.male	90	88
	Age.25.00+ thru 32	Age.32.00+ thru 43
Sex.female	46	51
Sex.male	111	89
	Age.43.00+ thru 71	
Sex.female	58	
Sex.male	90	

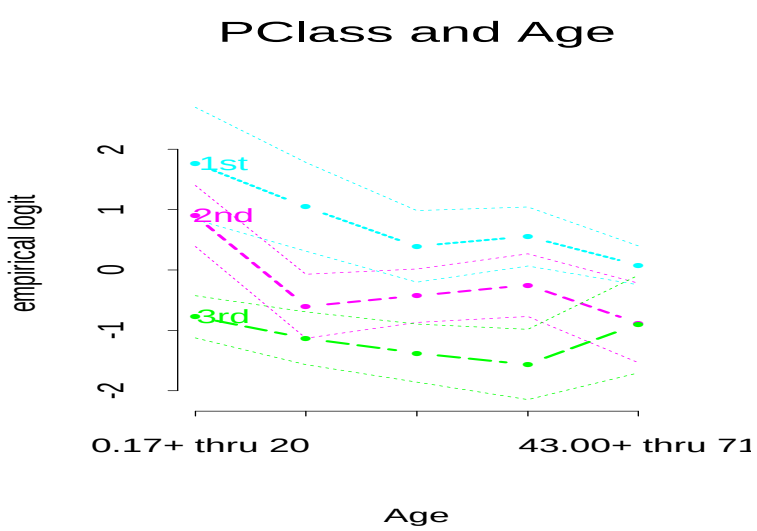
Interaction Effects for PClass and Age

Empirical Logit

	Age. 0.17+ thru 20	Age.20.00+ thru 25
PClass.1st	1.77	1.05
PClass.2nd	0.90	-0.60
PClass.3rd	-0.78	-1.13
	Age.25.00+ thru 32	**Age.32.00+ thru 43**
PClass.1st	0.39	0.56
PClass.2nd	-0.43	-0.25
PClass.3rd	-1.38	-1.57
	Age.43.00+ thru 71	
PClass.1st	0.08	
PClass.2nd	-0.88	
PClass.3rd	-0.90	

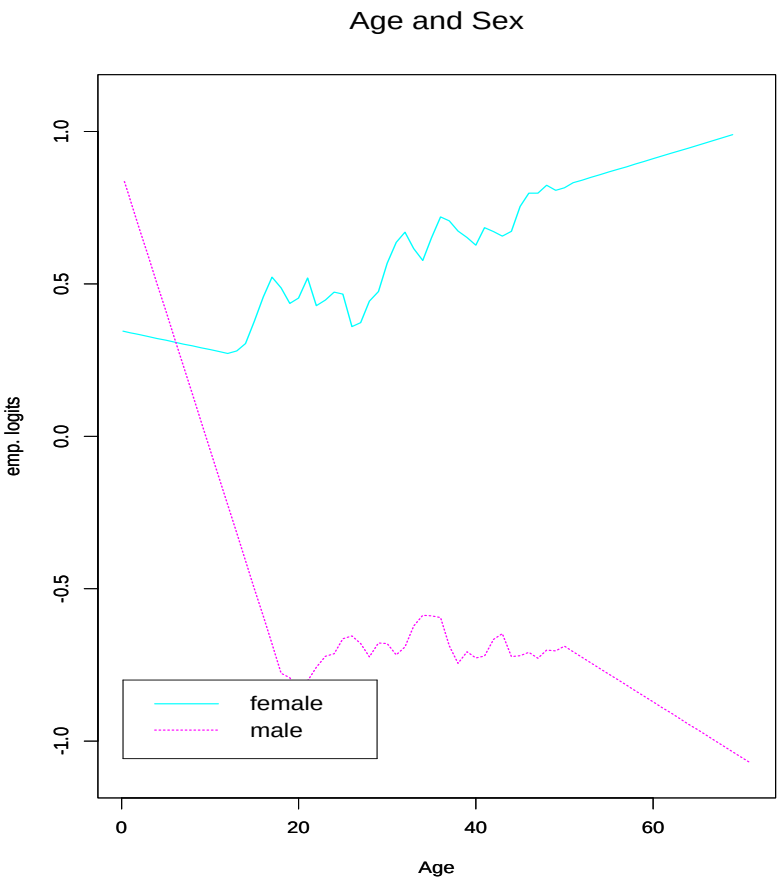


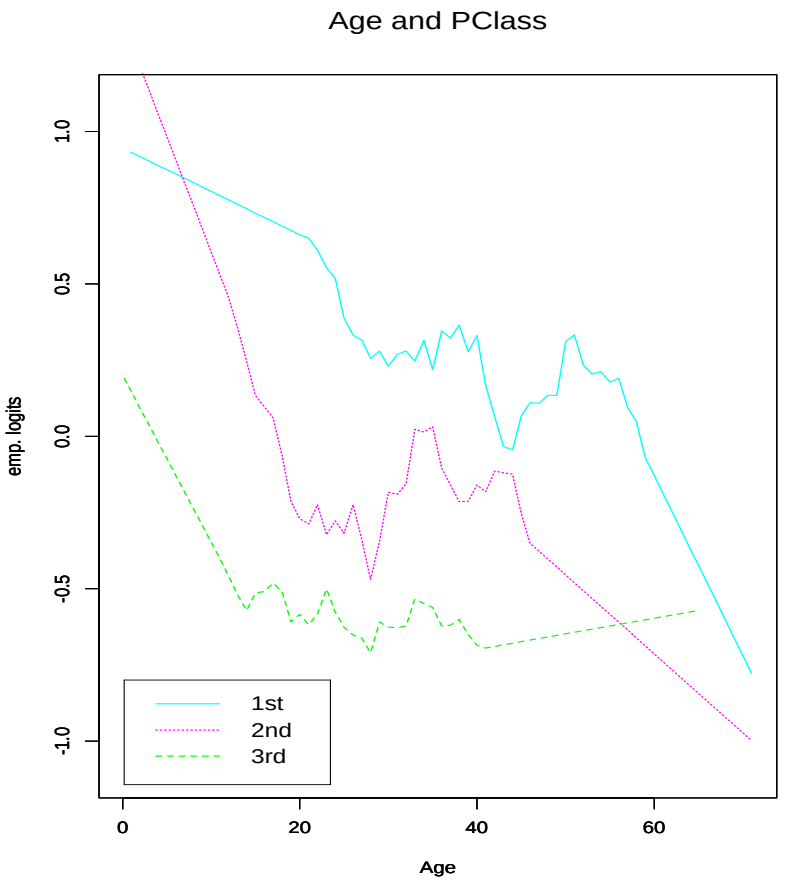




Interaction Effects are present since lines are nonparallel

Smoothed Logits for Binary Responses:





Final notes to EDA in logistic regression

- EDA is only a screening methods. Hypotheses generated by the EDA have to be verified with partial deviance tests.

- Binomial models are needed to assess the fit with a residual deviance test.

References

Czado, C. (1997). On selecting parametric link transformation families in generalized linear models. *J. Statist. Plann. Inference 61*, 125–139.

Hastie, T. and R. Tibshirani (1999). *Generalized additive models (2nd edition)*. London: Chapman & Hall.

Klein, R. and R. Spady (1993). An efficient semi parametric estimator for binary response models. *Econometrica 61*, 387–421.

Newton, M., C. Czado, and R. Chappell (1996). Bayesian inference for semi parametric binary regression. *JASA 91*, 142–153.

O'Sullivan, F., B. Yandell, and W. Raynor (1986). Automatic smoothing of regression functions in generalized linear models. *JASA 81*, 96–103.