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: \((Y_i, x_i) i = 1, \ldots, n\) \(Y_i\) independent
\(Y_i = 1\) or 0
\(x_i \in \mathbb{R}^p\) covariates (known)

Model: \(p(x_i) := P(Y_i = 1|X_i = x_i)\)
\(\Rightarrow P(Y_i = 0|X_i = x_i) = 1 - p(x_i)\)

How to specify \(p(x_i)\)? We need \(p(x_i) \in [0, 1]\).
## Example: Survival on the Titanic

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Passenger Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>PClass</td>
<td>Passenger Class</td>
</tr>
<tr>
<td>Age</td>
<td>Age of Passenger</td>
</tr>
<tr>
<td>Sex</td>
<td>Gender of Passenger</td>
</tr>
</tbody>
</table>
| Survived     | Survived=1 means Passenger survived  
               | Survived=0 means Passenger did not survive |
Model hierarchy

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

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

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

Model 4) \( p(x) = F_0(x^t \beta) \)
\[ F_0 \text{ known cdf} \]
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 $\beta$ 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

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

\[ \psi \text{ link parameter} \]

\[ \text{joint estimation of } \beta \text{ and } \psi \text{ is needed} \]

Example:

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

\[ h(\eta, \psi) = \begin{cases} \frac{(\eta + 1)\psi_1 - 1}{\psi_1} & \eta \geq 0 \\ -\frac{(-\eta + 1)\psi_2 - 1}{\psi_2} & \eta < 0 \end{cases} \]

\[ \psi = (1, 1) \quad \text{corresponds to logistic regression} \]

\[ \psi = (\psi_1, 1) \quad \text{Right tail family} \]

\[ \psi = (1, \psi_2) \quad \text{Left tail family} \]

Both tail family
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:

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

\[ \Rightarrow P(Y_i | X_i = x_i) = P(U_i \leq x_i^t \beta) = F(x_i^t \beta) \]
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

\[ Y_i | X_i = x_i \sim \text{binary}(p(x_i)) \text{ independent} \]

\[ p(x_i) = P(Y_i = 1 | X_i = x_i) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \]

Binary response can be extended to binomial response:

\[ Y_i \sim \text{bin}(n_i, p(x_i)) \text{ ind.} \]

\[ \Rightarrow P(Y_i = y_i | X_i = x_i) = \binom{n_i}{y_i} p(x_i)^y_i (1 - p(x_i))^{n_i-y_i}, \]

i.e. \[ \left\{ \frac{Y_i}{n_i} \right\} \] is a GLM with canonical link.
Explorative data analysis (EDA) for binomial regression data

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

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

<table>
<thead>
<tr>
<th>covariates</th>
<th>qualitative / categorical</th>
<th>dichotomous (2 levels)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>polytomous ((J) levels)</td>
<td>ordinal ((J) levels)</td>
</tr>
<tr>
<td></td>
<td>quantitative</td>
<td></td>
</tr>
</tbody>
</table>

©(Claudia Czado, TU Munich)
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
> 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 \( p(x) \)

Dichotomous covariate.

<table>
<thead>
<tr>
<th>Status</th>
<th>Gender</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>not survived</td>
<td>female</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>709</td>
</tr>
<tr>
<td>survived</td>
<td>female</td>
<td>308</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>142</td>
</tr>
</tbody>
</table>

Data

Want to estimate

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 − ( p(0) )</td>
</tr>
<tr>
<td>1</td>
<td>( p(0) )</td>
</tr>
</tbody>
</table>

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

\( o := \frac{p}{1-p} \) “odds of success”, \( p = \) success probability

\( \text{logit}(p) := \log(o) = \log \left( \frac{p}{1-p} \right) \) Log odds
Influence of single covariate on $p(x)$

\[
\begin{align*}
\text{logit}(p(1)) &= \log \left( \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \right) = \log \left( e^{\beta_0 + \beta_1} \right) = \beta_0 + \beta_1 \\
\text{logit}(p(0)) &= \log(e^{\beta_0}) = \beta_0 \\
\psi &:= \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))} \quad \text{“odds ratio”} \\
p(1) \approx 0, p(0) \approx 0 &\Rightarrow \psi \approx \frac{p(1)}{p(0)} \quad \text{relative risk}
\end{align*}
\]
Odds ratio as dependency measure

<table>
<thead>
<tr>
<th>Data</th>
<th>Conditional distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>( X )</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( a )</td>
<td>( b )</td>
</tr>
<tr>
<td>( c )</td>
<td>( d )</td>
</tr>
</tbody>
</table>

\[ p(j) = P(Y = 1 | X = j), \quad j = 0, 1 \text{ conditional distribution.} \]

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

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

\[ \psi = 1 \iff Y \text{ and } X \text{ are independent} \]
Unstructured model

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

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

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

\[ \hat{\text{Var}}(\hat{\log(\psi)^{obs}}) \approx (\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}) \quad \text{(est. var. of } \hat{\log(\psi)}^{obs}) \]

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

100(1 - \alpha)\% CI for \psi : \quad (e^{\log \hat{\psi}^{obs} - z_{\alpha/2} \sqrt{\hat{\text{Var}}(\log(\hat{\psi}^{obs}))}}, e^{\log \hat{\psi} + z_{\alpha/2} \sqrt{\hat{\text{Var}}(\log(\hat{\psi}^{obs}))}})
Example: Survival on the Titanic

\[ Y = \begin{cases} 1 & \text{survived} \\ 0 & \text{not survived} \end{cases} \quad X = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases} \]

<table>
<thead>
<tr>
<th></th>
<th>( Y )</th>
<th>( X )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>154</td>
<td>709</td>
</tr>
<tr>
<td>1</td>
<td>308</td>
<td>142</td>
</tr>
</tbody>
</table>

\[ \hat{p}(0) = \frac{308}{154 + 308} = 0.67 \]

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

\[ \hat{\psi} = \frac{\hat{o}(1)}{\hat{o}(0)} = \frac{0.2}{2} = 0.1 = \frac{1}{10} \]

Women had 10 times higher odds to survive compared to men
### Polytomous covariate

<table>
<thead>
<tr>
<th>$Y$</th>
<th>Category of $X$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$1$</td>
</tr>
<tr>
<td>0</td>
<td>$1 - p(1)$</td>
</tr>
<tr>
<td>1</td>
<td>$p(1)$</td>
</tr>
</tbody>
</table>
Nominal categories

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

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

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

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

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 \( \psi_1 = \ldots = \psi_{J-1} : \text{constant odds ratio} \)
Example: Survival on the Titanic

Consider Pclass as nominal covariate

<table>
<thead>
<tr>
<th>Y</th>
<th>Pclass</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>129</td>
</tr>
<tr>
<td>1</td>
<td>193</td>
</tr>
</tbody>
</table>

\[
\hat{p}^{obs}(j) = \frac{193}{129+193} = 0.60 \quad 0.42 \quad 0.19
\]

\[
\hat{o}^{obs}(j) = \frac{0.6}{1-0.6} = 1.50 \quad 0.72 \quad 0.23
\]

\[
logit(\hat{p}^{obs}(j)) = 0.41 \quad -0.33 \quad -1.50
\]

\[
\hat{\psi}^{obs}(j) = \frac{1.5}{0.23} = 6.50 \quad 3.10
\]

First (second) class passengers had a 6.5 (3.1) times higher odds to survive compared to third class passengers \(\rightarrow\) 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

<table>
<thead>
<tr>
<th>Example:</th>
<th>Age groups</th>
<th>$s(j)$ scores (means)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20 – 34</td>
<td>35 – 44</td>
</tr>
<tr>
<td></td>
<td>45 – 54</td>
<td>55 – 64</td>
</tr>
</tbody>
</table>

$s(j)$ scores (means) 27 39.5 49.5 59.5

$p(j) = P(Y = 1|X = j) = \frac{e^{\beta_0+\beta_1s(j)}}{1+e^{\beta_0+\beta_1s(j)}} \Rightarrow \log \left( \frac{p(j)}{1-p(j)} \right) = \beta_0 + \beta_1 s(j)$

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:** $Y_i | X_i = x_i \sim \text{bin}(n_i, p(x_i))$ independent

For a logistic model we have

$$\text{logit}(p(x_i)) = \beta_0 + \beta_1 x_i$$

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

**Problem:**

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

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

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

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.

```r
> 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")
}
```
```r
> titanic.main()

Main Effects for Sex

<table>
<thead>
<tr>
<th></th>
<th>female</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp. logit</td>
<td>0.69</td>
<td>-1.61</td>
</tr>
<tr>
<td>n</td>
<td>462.00</td>
<td>851.00</td>
</tr>
</tbody>
</table>

Main Effects for PClass

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp. logit</td>
<td>0.4</td>
<td>-0.3</td>
<td>-1.42</td>
</tr>
<tr>
<td>n</td>
<td>322.0</td>
<td>280.0</td>
<td>711.0</td>
</tr>
</tbody>
</table>

Main Effects for Age

<table>
<thead>
<tr>
<th></th>
<th>0.17+ thru 20</th>
<th>20.00+ thru 25</th>
<th>25.00+ thru 32</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp. logit</td>
<td>0.01</td>
<td>-0.57</td>
<td>-0.64</td>
</tr>
<tr>
<td>n</td>
<td>171.00</td>
<td>139.00</td>
<td>157.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>32.00+ thru 43</th>
<th>43.00+ thru 71</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp. logit</td>
<td>-0.4</td>
<td>-0.22</td>
</tr>
<tr>
<td>n</td>
<td>140.0</td>
<td>148.00</td>
</tr>
</tbody>
</table>
```
For quantitative covariates a grouped variable using quintiles are used.

\[ \text{Sex \ empirical logit} \]

\[ \text{PClass \ empirical logit} \]

\[ \text{Age \ empirical logit} \]

Quadratic Effect of Age?
Smoothed empirical logits for binary Responses:

Indicates nonlinear Age Effect, but maybe not quadratic
Influence on $p(x)$ of several covariates

Linear models: one quantitative/one dichotomous

\[ Y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 x_i \cdot D_i + \epsilon_i \]

\[ D_i = \begin{cases} 
1 & \text{male} \\
0 & \text{female} 
\end{cases} \]

\begin{align*}
\text{i male:} & \quad 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 \\
\text{i female:} & \quad Y_i = \beta_0 + \beta_1 x_i + \epsilon_i 
\end{align*}

Testing for interaction: \( H_0 : \beta_3 = 0 \quad \text{H}_1 : \beta_3 \neq 0 \)
If second covariable is polytomous with $J$ levels use

$$D_{1i} = \begin{cases} 1 & \text{obs. } i \text{ has category 1} \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

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

For interactions add terms $x_iD_{1i}, \ldots, x_iD_{(k-1)i}$. 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

<table>
<thead>
<tr>
<th></th>
<th>PClass.1st</th>
<th>PClass.2nd</th>
<th>PClass.3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex.female</td>
<td>2.65</td>
<td>1.95</td>
<td>-0.50</td>
</tr>
<tr>
<td>Sex.male</td>
<td>-0.71</td>
<td>-1.76</td>
<td>-2.02</td>
</tr>
</tbody>
</table>

Cell Sizes

<table>
<thead>
<tr>
<th></th>
<th>PClass.1st</th>
<th>PClass.2nd</th>
<th>PClass.3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex.female</td>
<td>143</td>
<td>107</td>
<td>212</td>
</tr>
<tr>
<td>Sex.male</td>
<td>179</td>
<td>173</td>
<td>499</td>
</tr>
</tbody>
</table>
### Interaction Effects for Sex and Age

**Empirical Logit**

<table>
<thead>
<tr>
<th>Age Range</th>
<th>Sex</th>
<th>Coefficient</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17+ thru 20</td>
<td>female</td>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>Age 20.00+ thru 25</td>
<td>male</td>
<td>-0.68</td>
<td>-1.56</td>
</tr>
<tr>
<td>Age 25.00+ thru 32</td>
<td>female</td>
<td>0.81</td>
<td>1.50</td>
</tr>
<tr>
<td>32.00+ thru 43</td>
<td>male</td>
<td>-1.38</td>
<td>-1.65</td>
</tr>
<tr>
<td>Age 43.00+ thru 71</td>
<td>female</td>
<td>1.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>-1.58</td>
<td></td>
</tr>
</tbody>
</table>

**Cell Sizes**

<table>
<thead>
<tr>
<th>Age Range</th>
<th>Sex</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17+ thru 20</td>
<td>female</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>90</td>
</tr>
<tr>
<td>Age 20.00+ thru 25</td>
<td>female</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>88</td>
</tr>
<tr>
<td>Age 25.00+ thru 32</td>
<td>female</td>
<td>46</td>
</tr>
<tr>
<td>32.00+ thru 43</td>
<td>male</td>
<td>51</td>
</tr>
<tr>
<td>Age 43.00+ thru 71</td>
<td>female</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>90</td>
</tr>
</tbody>
</table>
# Interaction Effects for PClass and Age

**Empirical Logit**

<table>
<thead>
<tr>
<th>Age</th>
<th>PClass.1st</th>
<th>PClass.2nd</th>
<th>PClass.3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17+ thru 20</td>
<td>1.77</td>
<td>0.90</td>
<td>-0.78</td>
</tr>
<tr>
<td>20.00+ thru 25</td>
<td>1.05</td>
<td>-0.60</td>
<td>-1.13</td>
</tr>
<tr>
<td>25.00+ thru 32</td>
<td>0.39</td>
<td>-0.43</td>
<td>-1.38</td>
</tr>
<tr>
<td>32.00+ thru 43</td>
<td>0.56</td>
<td>-0.25</td>
<td>-1.57</td>
</tr>
<tr>
<td>43.00+ thru 71</td>
<td>0.08</td>
<td>-0.88</td>
<td>-0.90</td>
</tr>
</tbody>
</table>
Interaction Effects are present since lines are nonparallel
Smoothed Logits for Binary Responses:

Age and Sex

Age and PClass
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


