Chapter 7

Modeling binary, nominal and count data

Modeling binary variables

Practical question: a bank should decide about granting loans to new clients, i.e. forecast of the solvency

$$Y_i = \begin{cases} 0, & \text{the client } i \text{ is solvent} \\ 1, & \text{the client } i \text{ is insolvent} \end{cases}$$

X_{1i}	–	debt-to-income ratio ($\times 100$);
X_{2i}	–	years with the current employer;
X_{3i}	–	other debts (in 1000 Euro);
X_{4i}	–	age (in years).

x_1 large $\Rightarrow$ higher chances of being insolvent

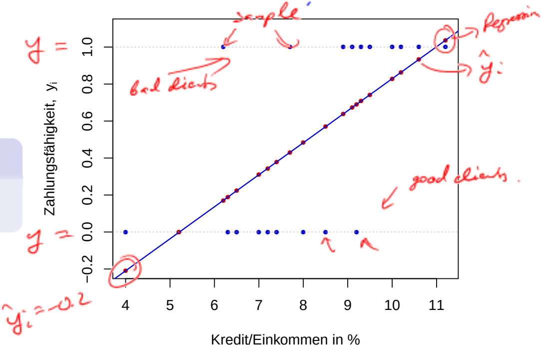
x_1 small $\Rightarrow$ lower chances of insolvency.

Question: can we use a linear regression model for binary variales? $\rightsquigarrow$ linear probability model

$$\hat{\beta} = (X'X)^{-1}X'\underline{y} \leftarrow \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} \Rightarrow \text{not a problem can be estimated.}$$

Universität Augsburg

scatter plot: on two horizontal lines ⇒ unusual



Linear Prob.-model

$$Y_i = \beta_0 + \beta_1 \cdot X_i + u_i$$

{0,1}.   debt/income.

Note:

- $(+)$ the forecast $\hat{Y}_i$ can be seen as probability

  expectation for discreet RV

$$E(Y_i|X_i) = 1 \cdot P(Y_i = 1|X_i) + 0 \cdot P(Y_i = 0|X_i) = p_i$$

  ⇒ $\hat{Y}_i$ can be seen as the probability of being insolvent!

- $(-)$ $\hat{Y}_i$ may lie outside of [0,1] ⇒ not a probability.
- $(-)$ R^2 is useless as a goodness-of-fit measure ($R^2=1$ cannot be attained)
- $(-)$ the residuals are not normally distributed Y-Binary ⇒ u Binary ⇒ not normal.
- $(-)$ $Var(Y_i|X_i) = p_i(1-p_i) \neq const \rightsquigarrow$ heteroscedastic

⇒ $Var(u_i) \neq$

Universität Augsburg

Transition to Logit/Probit

$\{0,1\}$

Let Y_i be the observed binary variable and Y_i^* the corresponding unobserved metric variable. For Y_i^* it holds:

continuous ⇒ LR is OK

$$Y_i^* = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i.$$

a continuous variable

Example: Y_i^* is an unobserved solvency of the client i with

$$\underline{Y_i = 1} \text{ if } Y_i^* > 0 \text{ and } \underline{Y_i = 0} \text{ if } Y_i^* \leq 0.$$

debt/income

$$\begin{aligned} P(Y_i = 1|\boldsymbol{X}_i) &= P(Y_i^* > 0|\boldsymbol{X}_i) = P(\cancel{\boldsymbol{X}}_i'\boldsymbol{\beta} + u_i > 0|\boldsymbol{X}_i) \\ &= \underbrace{P(-u_i < \boldsymbol{X}_i'\boldsymbol{\beta}|\boldsymbol{X}_i)}_{\text{definition of the cdf}} = F(\boldsymbol{X}_i'\boldsymbol{\beta}), \end{aligned}$$

⇒ $P(Y=1)$ = to the cdf of the residuals u evaluated at $X'\beta$.

where $F(\cdot)$ is the cdf of the residuals.

- $F(z) = \frac{1}{1+e^{-z}}$ - the cdf of the logistic distribution $\rightsquigarrow$ logit
- $F(z)$ - the cdf of the normal distribution $\rightsquigarrow$ probit

$$P(Y_i = 1|X_i) = F(X_i'\beta)$$

Universität Augsburg

Logistic regression

$\lim_{z\to\infty} \frac{1}{1+e^{-z}} = 1$

Idea: transformation with the logistic function

F (or cdf of the logistic distribution)

Logistic model

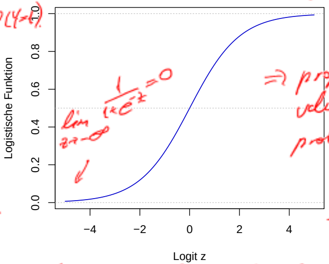
$$p_i = P(Y_i = 1|\boldsymbol{X}_i) = \frac{1}{1+e^{-z_i}}; \quad z_i \to X_i'\beta$$

for logits z_i it holds

$$z_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} \cancel{+ u_i}$$

$P(Y=1)$

$\lim_{z\to-\infty} \frac{1}{1+e^{-z}} = 0$



$\Rightarrow$ proper values for probabilities.

z_i

important: no residuals here!!! $P(Y_i=1)$ is a probability, so no RV's allowed on the right-hand-side.

Note: Alternatively we may use the CDF $\Phi(z_i)$ of $N(0,1)$ $\rightsquigarrow$ probit-model

$$P(Y_i = 1|X_i) = \Phi(z_i)$$

Universität Augsburg

Estimation of the parameters

The parameters are estimated using ML:

$y_i = 1$ $\qquad y_i = 0$

$$L = \prod_{i=1}^{n} \underbrace{\left(\frac{1}{1+e^{-z_i}}\right)^{y_i}}_{P(Y_i=1)} \cdot \underbrace{\left(1 - \frac{1}{1+e^{-z_i}}\right)^{1-y_i}}_{P(Y_i=0)} \longrightarrow max, \text{ w.r.t. } \beta_0, \ldots, \beta_k.$$

Note:

- In contrary to the LR the estimation is always numeric.
- Likelihood-Ratio tests can be used to check the significance of the parameters.
- R: `glm(y ~ X,data=data, family=binomial(logit))`

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.434735   0.482326  -2.975  0.00293 **
debtinc      0.121391   0.019023   6.381 1.76e-10 ***
employer    -0.161795   0.023742  -6.815 9.44e-12 ***
debts        0.093460   0.045045   2.075  0.03801 *
age         -0.004397   0.014212  -0.309  0.75701
---
```

significant

insignificant

$\beta_0, \dots \beta_4$ $\Rightarrow$ age has no impact on the probability of being solvent / insolvent.

Interpretation: LR : x_i changes by 1 $\Rightarrow$ y changes by β_i

Here : difficult x_i changes by 1 $\Rightarrow$ $P(Y=1)$ – changes by ???

Example: a data set with 700 observations

odds = 3 ⇒ the probability of being insolvent is 3 times higher than probability of being solvent.

x_i changes by 1, then odds changes by e^{β_i}

	debtinc	employer	debts	age
$\hat{\beta}_i$	0.121*	-0.162*	0.093*	-0.004
$e^{\hat{\beta}_i}$	1.129	0.851	1.098	0.996

(*) - significant with $\alpha = 0.05$

$\frac{1}{1+e^{-z}}$

Odds of the logistic regression

$$\text{Odds} = \frac{P(Y=1|\boldsymbol{X})}{P(Y=0|\boldsymbol{X})} = e^z$$

$\Rightarrow 1 - \frac{1}{1+e^{-z}} = \frac{1+e^{-z}-1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}}$

	Logit (z)	Odds	$P(Y=1\|\boldsymbol{X})$
⇒ $\beta > 0$	rises by β	rises by $e^\beta > 1$	rises
⇒ $\beta < 0$	falls by β	falls by $e^\beta < 1$	falls

⇒ debtinc and debts ⇒ increase the prob. of insolvency.

⇒ the years with the current employer decrease the prob. of insolvency.

⇒ age is insignificant.

Universität Augsburg

Forecasts:

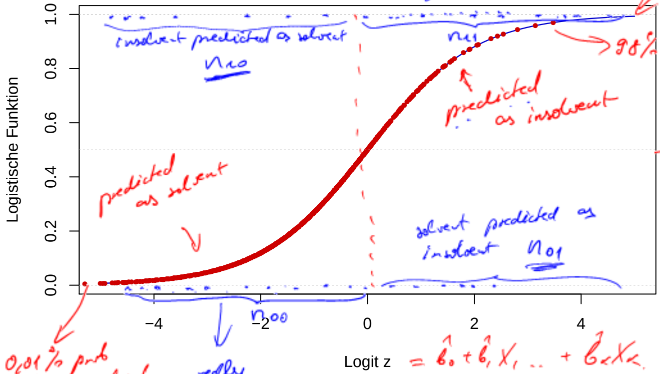
plug-in estimates β's

predicted probability of being insolvent

$$P(\widehat{Y_0 = 1|\boldsymbol{X}_0}) = \hat{\pi}_0 = \frac{1}{1 + e^{-\hat{b}_0 - \hat{b}_1 X_{10} - \ldots - \hat{b}_k X_{k0}}}.$$

really insolvent.

logistic line



98% prob of being insolvent.

0.5

0,01% prob of being insolvent.

really solvent.

$= \hat{b}_0 + \hat{b}_1 X_1 \ldots + \hat{b}_k X_k$

! in contrary to a LR: here we can visualize the forecasts because of z.

Universität Augsburg

Goodness of the model

$R^2 = 1$ cannot be attained

Problem: classical measures, such as R^2, cannot be used $\rightsquigarrow$ pseudo-R^2; classification tables; graphical measures (ROC-curve)

assume we do not have any X-Variables; 75% good; 25% bad clients $\Rightarrow$ for a new client there is a prob. of 75% of being good.

- pseudo-R^2:
  - Let LL_0 be the Log-Likelihood of the null model ($b_1 = \cdots = b_k = 0$)
  - Let LL_v be the Log-Likelihood of the full model (with all variables)
  - Let LL_s be the Log-Likelihood of the $\underbrace{\text{saturated model}}_{\text{oracle model}}$ (model with perfect fit, here $LL_s = 0$)
  - Deviance: $D = -2 \cdot LL_v$ (close 0)
  - Mc-Faddens-R^2: $1 - LL_v/LL_0$ (starting from 0.4)

slide 392

our / worst model

LL_0

```
Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 626.49  on 695  degrees of freedom
```

$LL_0 - LL_v$

$R^2 = \frac{626}{804} = 0.75 > 0.4.$

- Classification table

really insolvent | predicted to be insolvent | predicted to be solvent

		predicted		
		$\hat{Y}=1$	$\hat{Y}=0$	
truth	$y=1$	$n_{11}=TP$	$n_{01}=FN$	$n_{\cdot 1}=P$
	$y=0$	$n_{10}=FP$	$n_{00}=TN$	$n_{\cdot 0}=N$
		$n_{1\cdot}$	$n_{0\cdot}$	

really solvent

Let $\hat{y}_i = 1$ if $\widehat{P(Y_i = 1|\boldsymbol{X}_i)} \geq 0.5$ and 0 else.

$n_{10} = 38 \Rightarrow$ solvent predicted as insolvent
$n_{01} = 111 \Rightarrow$ insolvent predicted as solvent

	$\hat{Y}=1$	$\hat{Y}=0$
$Y=1$	72	111 $\Rightarrow$ wrong
$Y=0$	38	479

$\Rightarrow$ correct decisions

$n_{11} = 72$ – the number of insolvent, who are predicted to be insolvent.
$n_{00} = 479$ – solvent predicted as solvent.

$\rightsquigarrow$ $(479+72)/700 = 78{,}71\%$ are correctly predicted. $\Rightarrow$ in 79% of cases we made a correct decision with our logit.

But: there are 73,86% solvent clients in the sample.

$\frac{38+479}{700} \approx 75\% \Rightarrow$ toss 2 coins $\rightarrow$ N N $\Rightarrow$ insolvent (25%); N H, H H, H N $\Rightarrow$ solvent. (75%) $\Rightarrow$ 75% of correct decision with 2 coins.

Quaestion: is the threshold 0.5 a good choice? $\Rightarrow$ objective is to find another threshold 30%? 70%?

Universität Augsburg

Goodness of the model and the choice of the threshold

- ROC (*receiver operating characteristics*), Lift and Gain curves are used to visualize and to quantify the goodness of the classification algorithms.

number of insolvent classified as insolvent

$$
\begin{aligned}
\text{sensitivity} &= \frac{n_{11}}{n_{\cdot 1}} = \frac{n_{11}}{(n_{11} + n_{01})} \\
\text{specifity} &= \frac{n_{00}}{n_{\cdot 0}} = \frac{n_{00}}{(n_{10} + n_{00})}
\end{aligned}
$$

number of insolvent clients

solvent as solvent

number of solvent

Sensitivity: the fraction of correctly classified 1-values among all true 1-objects.

Specifity: the fraction of correctly classified 0-values among all true 0-object.

⇒ 61% of insolvent are classified as solvent ⇒ these get money from the bank, but cannot pay it back!!!

- Sensitivity= $\overset{n_{11}}{72}/(\overset{n_{11}}{72}+\overset{n_{01}}{111})=0.39$ - only 39% of insolvent clients are classified as insolvent
- Specifity= $\overset{n_{00}}{479}/(\overset{n_{00}}{479}+\overset{n_{10}}{38})=0.92$ - 92% of solvent clients are classified as solvent

⇒ only 8% of solvent are classified as insolvent.

⇒ that our model works well for solvent clients, but has a very low detection rate for insolvent!!!

Aim: increase sensitivity of the model. (this will obviously decrease the specifity).

number of clients predicted as being insolvent.

$$\text{PPV or PV+} = \frac{n_{11}}{n_{1\cdot}} = \frac{n_{11}}{(n_{11}+n_{10})}$$

$$\text{NPV or PV-} = \frac{n_{00}}{n_{0\cdot}} = \frac{n_{00}}{(n_{01}+n_{00})}$$

predicted solvent clients.

PPV: the fraction of correctly classified 1-values among all objects classified as 1.

NPV: the fraction of correctly classified 0-values among all objects classified as 0.

(PPV-positive predicted value, NPV-negative predicted value)

⇒ 35% are in fact solvent, but classified as insolvent
→ will get no credit, despite of being able to pay it back
(bank looses good clients).

- PPV= $\overset{n_{11}}{72}/(\overset{n_{11}}{72}+\overset{n_{10}}{38}) = 0.65$ - only 65% of all as insolvent classified clients are really insolvent
- NPV= $\overset{n_{00}}{479}/(\overset{n_{00}}{479}+\overset{n_{01}}{111}) = 0.81$ - 81% of all as solvent classified clients are really solvent

⇒ 19% of classified as solvent are insolvent
bank looses the whole credits

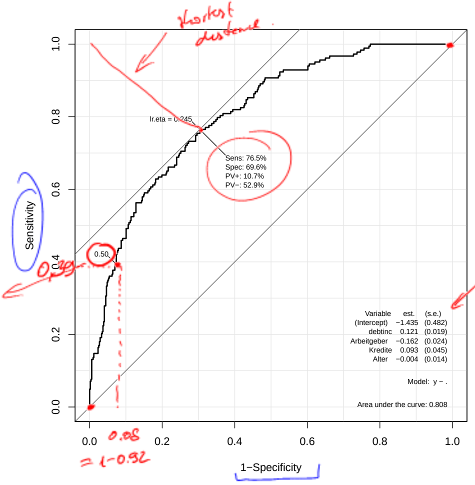
ROC-curve: senstitivity values as a function of specifity

- The steeper the function, the better the algorithm. ROC-value is the square under the curve.
- If the curve is close to the diagonal, then the algorithm is as good as random assignments.

R: `roc`-function from the `pROC`-package

optimal threshold = 0.245.
⇒ every client with prob. of insolvency ≥ 0.245 is classified as insolvent, and else as solvent.

natural threshold



the same parameters as before.

Universität Augsburg

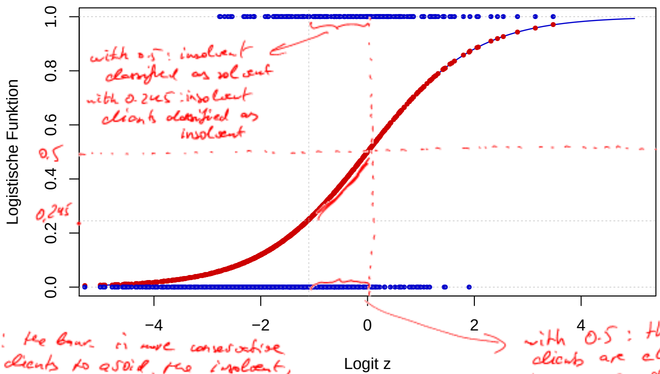
Now let $\hat{y}_i = 1$ if $\widehat{P(Y_i = 1|\boldsymbol{X}_i)} > 0.245$ and 0 else.

72 111
38 479.

479 ↘ 359
⇒ worse job in detecting solvent clients
spec : 92% ↘ 69,6%

	$\hat{Y}=1$	$\hat{Y}=0$
$Y=1$	140	43
$Y=0$	158	359

72 ↗ 140 ⇒ better job in detecting insolvent
sens : 39% ↗ 76.5%.
⇒ price
sens ↗ ⇒ spec ↘
and other way around.



new threshold: the bank is more conservative
⇒ reject more clients to avoid the insolvent, but also takes the risk of rejecting good clients

with 0.5 : these good clients are classified as good
with 0.245 : these good are classified as bad.

Gain-curve:

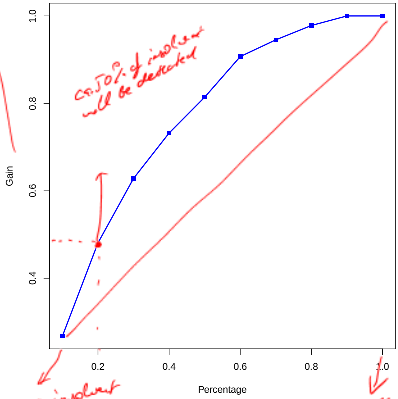
- If Gain equals 40% for 20%, this implies that if 20 % of the clients are classified as insolvent, then the algorithm will detect 40% of really insolvent clients .
- The digonal shows a model-free classification: If 20 % of the clients are classified as insolvent, then the algorithm detects 20% of really insolvent clients.
- The steeper the curve, the better (with a single kink).

R: gains-package

% of detected insolvent clients.

optimal fraction of clients to be classified as insolvent.

ca. 50% of insolvent will be detected



% of clients classified as insolvent

none as insolvent (friendly bank)

all as insolvent (unfriendly bank)

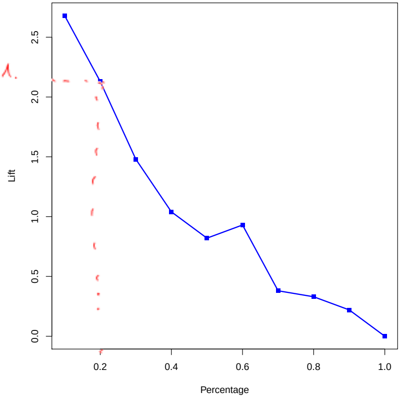
Universität Augsburg

Lift curve: how much better is the predictive model compared to the model free classification?

- If lift is 2.1 for 20%, this implies that the model detection rate of insolvent clients is larger by 2.1 compared to model-free classification.

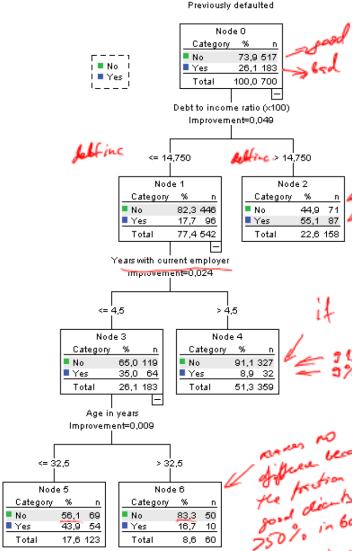
the larger the values on Y-Axis the better your model is compared to "tossing coins".

2.1.



The CART method can be applied to binary data: classification trees

CART ⇒ applied to a binary y.



44.9% of good
55.1% of bad
⇒ terminal node ⇒ every client with debtinc > 14,75 is classified as insolvent

if debtinc < 14,75 & employer > 4.5 ⇒ solvent

91% of solvent
9% of insolvent

no difference because the fraction of good clients > 50% in both
⇒ forecast is the same!

Classification

Observed	Predicted: No	Predicted: Yes	Predicted: Percent Correct
No	446	71	86,3%
Yes	96	87	47,5%
Overall Percentage	77,4%	22,6%	76,1%

Growing Method: CRT

Dependent Variable: Previously defaulted

worse than the logit

(but with threshold of 50%)

sens = $\frac{87}{96+87}$ < 0.45

Universität Augsburg

Modelling nominal data

Practical question:

- Choice of the political party depending of the characteristics of the voters;
- Choice of a product brand depending on the characteristics of the client;

4 categories (not ordered)

Example:

mode	–	"car", "air", "train", oder "bus"
choice	–	decision ⇒ the selected transport
wait	–	wait~~n~~ing time, 0 for "car"
vcost	–	variable costs
travel	–	time
gcost	–	total costs
income	–	income
size	–	number of persons

different values for different transportations modes

characterise the person, not the transport.
→ the same

fastest → (100)

longest → (417)

	individual	mode	choice	wait	vcost	travel	gcost	income	size
1	1	air	no	69	59	100	70	35	1
2	1	train	no	34	31	372	71	35	1
3	1	bus	no	35	25	417	70	35	1
4	1	car	yes	0	10	180	30	35	1
5	2	air	no	64	58	68	68	30	2
6	2	train	no	44	31	354	84	30	2

Multinomial logit model

logit.

For the simple logit model it holds:

same βs

$$P(Y=1|\boldsymbol{x}) = \frac{exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+exp(\boldsymbol{x}'\boldsymbol{\beta})} = \frac{1}{1+e^{-z}} = \frac{e^{z}}{1+e^{z}}$$

$$\ln\left(\frac{P(Y=0|\boldsymbol{x})}{P(Y=1|\boldsymbol{x})}\right) = \boldsymbol{x}'\boldsymbol{\beta} = z.$$

→ linear combination of variables defines ln(odds)

For the k categories of Y we define:

category specific.

$k=4$
$r=1..3$

three such models/equations in our case.

$$\ln\left(\frac{P(Y=r|\boldsymbol{x})}{P(Y=k|\boldsymbol{x})}\right) = \boldsymbol{x}'\boldsymbol{\beta}_r, \qquad r=1,\ldots,k-1$$

→ last category is the reference category (air)

with

only / all categories ⇒ the model links all categories

$$P(Y=r|\boldsymbol{x}) \quad = \quad \frac{exp(\boldsymbol{x}'\boldsymbol{\beta}_r)}{1+\sum_{s=1}^{k-1} exp(\boldsymbol{x}'\boldsymbol{\beta}_s)}, \quad r=1,\ldots,k-1$$

$$P(Y=k|\boldsymbol{x}) \quad = \quad \frac{1}{1+\sum_{s=1}^{k-1} exp(\boldsymbol{x}'\boldsymbol{\beta}_s)}.$$

to guarantee that $\sum_{r=1}^{k} P(Y=r|x) = 1$

One category, i.e. the k-th, is the reference category.

Universität Augsburg

Note:

- Estimation via ML assuming independence of the observations. This is a questionable assumption:
  - similar categories; $\Rightarrow$ train → fast train, local train; air → lowcost, Lufthansa and Co.; bus → private (small), large. $\Rightarrow$ subcategories can be highly correlated; bus → red, non-red
  - odds do not depend on other categories, etc.
  - Solution: Hausmann/McFadden test
- Goodness-of-fit, tests as for logit.

$$\ln\left(\frac{P(Y=r|x)}{P(Y=k|x)}\right) = x'\beta_r$$

there is interrelation between categories

depends only on the r-th category!

Universität Augsburg

Global and category specific variables

$x'_r \alpha_r \Rightarrow$ non identifiable

$$\boldsymbol{x}'\boldsymbol{\beta}_r \quad \mapsto \quad \boldsymbol{x}'_{glob}\boldsymbol{\beta}^*_r + \boldsymbol{x}'_{spec,r}\boldsymbol{\alpha}$$

- Global variables (income, number of persons) do not depend on the categories and have individual parameters for each category: $\boldsymbol{x}'_{glob}\boldsymbol{\beta}^*_r$. $\Rightarrow$ income doesn't depend on category $\Rightarrow$ the parameter has to depend on the category.
  The sign of the parameters cannot be interpreted.
- The category specific variables (waiting time, costs) depend on the categories and are evaluated relatively to the reference category.
  $\Rightarrow$ waiting depends on its own on the category $\Rightarrow$ no sense to make the parameter depend on the category too!
  $$(\boldsymbol{x}_{spec,r} - \boldsymbol{x}_{spec,k})'\boldsymbol{\alpha} \quad \text{or} \quad \boldsymbol{x}'_{spec,r}\boldsymbol{\alpha}$$
  The sign of the parameters can be interpreted.

Let *gcost* and *wait* be category specific and *income* and *size* are global variables. The reference category is *air*.

global

```
> library("mlogit")
> mlogit(choice~wait+gcost|income+size, ...)
```

→ category specific

```
Coefficients :
                     Estimate Std. Error t-value  Pr(>|t|)
train:(intercept) -2.3115942  0.7525161 -3.0718 0.0021276 **
bus:(intercept)   -3.4504941  0.9064886 -3.8064 0.0001410 ***
car:(intercept)   -7.8913907  0.9880615 -7.9867 1.332e-15 ***
wait              -0.1013180  0.0112207 -9.0296 < 2.2e-16 ***
gcost             -0.0197064  0.0053844 -3.6599 0.0002523 ***
train:income      -0.0589804  0.0154532 -3.8167 0.0001352 ***
bus:income        -0.0277037  0.0169812 -1.6314 0.1027991
car:income        -0.0041153  0.0127301 -0.3233 0.7464866
train:size         1.3289497  0.3141683  4.2301 2.336e-05 ***
bus:size           1.0090796  0.3952899  2.5528 0.0106874 *
car:size           1.0392585  0.2665513  3.8989 9.663e-05 ***
---
```

single parameter for category specific variables

category specific parameters for global variables

```
Log-Likelihood: -176.77
McFadden R^2:  0.37705
Likelihood ratio test : chisq = 213.98 (p.value = < 2.22e-16)
```

→ the same goodness of fit measure as for Logit

With the estimated paremeters we can estimate the probabilities $P(Y_i = r|\boldsymbol{x}_i)$ for all r.

	air	train	bus	car
[1,]	0.2368302	0.00000000	0.24496423	0.5182056
[2,]	0.2083323	0.27785076	0.00000000	0.5138170
[3,]	0.0000000	0.12686485	0.23058033	0.6425548
[4,]	0.1151004	0.05063597	0.02141839	0.8128452
[5,]	0.3405917	0.20694648	0.05624436	0.3962174
[6,]	0.1316850	0.36965292	0.26144217	0.2372200

= 1

1 - sum of other

using formulas on the board

individual 1–5 ⇒ car
6 ⇒ train

Universität Augsburg

Ordered response models

Aim: Y can take M different ordered (!) values (credit ratings, grades, income classes, etc)

Using a single latent variable we can specify

$$\begin{aligned} Y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \\ Y_i &= j \quad \text{if} \quad \gamma_{j-1} < Y_i^* \leq \gamma_j \end{aligned}$$

for some unknown threshold values γ_j with $\gamma_0 = -\infty$ and $\gamma_M = \infty$.

Assuming logistic cdf for u_i we obtain ordered logit model and assuming normality we obtain ordered probit model.

Example: a (simplified) rating of companies - $Y = 1$ - lowest, $Y = 2$ - average, $Y = 3$ - highest

MARKET_VALUE	DIV_PER_SHR	TOTAL_DEBT	rating
-0.08911931	-0.08063048	-0.02276501	1
-0.09350059	-0.08148114	-0.14012151	3
-0.09452652	-0.08019333	-0.13869093	2
-0.09633656	-0.08090222	-0.13784721	2
-0.09254201	-0.08130392	-0.13822051	2
0.95192265	-0.06091170	13.33345544	3

$$
\begin{aligned}
Y^* &= \boldsymbol{x}'\boldsymbol{\beta} + u \\
Y &= 1 \quad \text{if} \quad y^* \leq \gamma_{1|2} \\
&= 2 \quad \text{if} \quad \gamma_{1|2} < y^* \leq \gamma_{2|3} \\
&= 3 \quad \text{if} \quad \gamma_{2|3} < y^*
\end{aligned}
$$

Logi

Assuming normal error terms we can state the corresponding prob's

$$
\begin{aligned}
P(Y_i \leq k|\boldsymbol{x}_i) &= P(Y_i^* \leq \gamma_{(k-1)|k}|\boldsymbol{x}_i) = \Phi(\gamma_{(k-1)|k} - \boldsymbol{x}_i'\boldsymbol{\beta}) \\
P(Y_i = 1|\boldsymbol{x}_i) &= P(Y_i^* \leq \gamma_{1|2}|\boldsymbol{x}_i) = \Phi(\gamma_{1|2} - \boldsymbol{x}_i'\boldsymbol{\beta}) \\
P(Y_i = 3|\boldsymbol{x}_i) &= P(Y_i^* > \gamma_{2|3}|\boldsymbol{x}_i) = 1 - \Phi(\gamma_{2|3} - \boldsymbol{x}_i'\boldsymbol{\beta}) \\
P(Y_i = 2|\boldsymbol{x}_i) &= P(\gamma_{1|2} < Y_i^* \leq \gamma_{2|3}|\boldsymbol{x}_i) = \Phi(\gamma_{2|3} - \boldsymbol{x}_i'\boldsymbol{\beta}) - \Phi(\gamma_{1|2} - \boldsymbol{x}_i'\boldsymbol{\beta})
\end{aligned}
$$

The log-likelihood function is then given by

$P(Y_1 = 1, Y_2 =$

$$
LL(\boldsymbol{\beta}|\boldsymbol{X}) = \sum_{i=1}^{N} \log P(Y_i = y_i|\boldsymbol{x}_i) \rightarrow max, \text{ w.r.t. } \boldsymbol{\beta}
$$

The inferences follow in a similar fashion as for the simple logit.

$P($

Example:

```
Coefficients:
                Value Std. Error t value
MARKET_VALUE  2.14118     0.6803  3.1474
DIV_PER_SHR  -0.70836     0.3189 -2.2212
TOTAL_DEBT    0.05553     0.1367  0.4063
Intercepts:
    Value   Std. Error t value
1|2 -0.0309  0.0789    -0.3916
2|3  1.0181  0.0879    11.5797
Residual Deviance: 1480.408
AIC: 1490.408
> ordlog.pred = predict(ordlog, type="probs")
> ordlog.pred
              1            2         3
   5.259901e-01 2.340771e-01 0.2399328
   5.298021e-01 2.330434e-01 0.2371545
   5.305567e-01 2.328364e-01 0.2366069
   5.313852e-01 2.326083e-01 0.2360065
   5.292958e-01 2.331818e-01 0.2375224
   5.453970e-02 8.685558e-02 0.8586047
```

γ
γ

>0.5

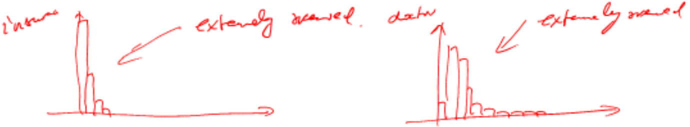
Modelling for count data

$Y = \{0,1,2,3\dots\}$ ⇒ a few values, but the set is not limited.
⇒ cannot apply ordered Logit ⇐ fixed number of categories

Practical questions:

- the number of claims by an insurance company per time period;
- the number of cosultations by a doctor per year ;
- the number of insolvent companies per time period;
- occurrences of a seldom disease per season; ⇒
- ....

Note: the modelling is particularly important for small values of the target variable (rare events) and the distribution is heavily skewed.



Universität Augsburg

Poisson distribution

The Poisson distribution is frequently used to model rare events

discrete

$$P(Y=y)=\begin{cases}\frac{\lambda^y}{y!}e^{-y}, & \text{for } y=0,1,2,\dots\\ 0, & \text{else,}\end{cases}$$

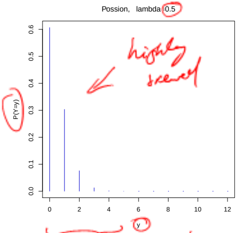
any natural number

with the intensity parameter λ. It fulfils the *equidispersion*-condition:

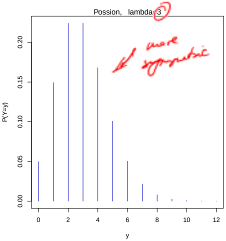
$$E(Y)=Var(Y)=\lambda$$

low intensity expected number of claims per year = 0.5

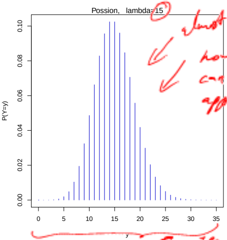
⇒ we expect 3 claims



highly skewed



more symmetric



almost symmetric normal distribution can be a good approximation

up to 5 different outcomes in the sample ⇒ Poisson

OLS for ≥ 5 possible outcomes.

LR: $y = \underline{\beta_0 + \beta_1 x_1 + \ldots + \beta_J x_J} + u. \Rightarrow E(Y|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_J x_J$

Poisson regression model

$E(Y) = \lambda$ plug x's into λ. $\lambda > 0$ - always $\Rightarrow \lambda = \beta_0 + \beta_1 x_1 + \ldots + \beta_J x_J$ - can be negative.

Let $Y_i, \boldsymbol{x}_i$ be independent realisations, while Y_i follows Poisson distribution with

link forces $x'\beta$ positive and guarantees positive λ's.

$$E(Y_i|\boldsymbol{x}_i) = h(\boldsymbol{x}_i'\boldsymbol{\beta}) = exp(\boldsymbol{x}_i'\boldsymbol{\beta}) = \lambda_i.$$

- The interpretation of the parameters follows as for the logit model.
- The parameters are estimated via ML:

$$LL(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln(h(\boldsymbol{x}_i'\boldsymbol{\beta})) - h(\boldsymbol{x}_i'\boldsymbol{\beta}) - \ln(y_i!) \longrightarrow max, \text{ w.r.t. } \boldsymbol{\beta}$$

$P(Y_1 = 2, Y_2 = 1, Y_3 = 0) = P(Y_1 = 2) \cdot P(Y_2 = 1) P(Y_3 = 0) =$

$= \frac{(h(x_1'\beta))^2}{2!} e^{-h(x_1'\beta)} \cdot \frac{(h(x_2'\beta))^1}{1!} \cdot e^{-h(x_2'\beta)} \cdot \frac{(h(x_3'\beta))^0}{0!} e^{-h(x_3'\beta)} \rightarrow$

$\rightarrow$ max w.r.t β $x = (\text{age}, \text{job}, \ldots)$.

Universität Augsburg

Goodness of the model

To measure the goodness of the model we use deviance, i.e. the difference between the log-likelihood for the actual observations (perfect/saturiertes model) and the log-likelihood for the predicted values:

$$D = -2\sum_{i=1}^{n}[LL_i(\hat{Y}_i) - LL_i(Y_i)] = 2\sum_{i=1}^{n}\left[Y_i \ln(Y_i/\hat{\lambda}_i)\right] \sim \chi^2_{n-p}$$

likelihood of your model (max on the previous slide)

likelihood of the oracle model.

R^2 cannot be applied, as for logit.

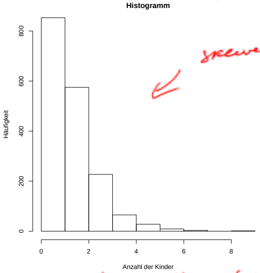
Example: number of children

child	-	number of children
age	-	age of the woman
dur	-	years at school/college
nation	-	nationality, 0 = german , 1 = else
god	-	trust in God: 1 = strong, ..., 6 = never thought about it
univ	-	university degree: 0 = no, 1 = yes ⇒ dummy.

```
mean(children$child)
[1] 1.57297
> var(children$child)
[1] 1.552769
```

equidispersion property is more or less fulfilled.

$E(Y) = Var(Y) = \lambda$ ← for a Poisson distribution



skewed.

⇒ less possible outcomes

Universität Augsburg

```
glm(formula = child ~ age + I(age^2) + I(age^3) + I(age^4) +
    dur + I(dur^2) + nation + god + univ, family = poisson(link = log),
    data = children)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1514  -0.7559   0.0102   0.4832   3.6715

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.228e+01  1.484e+00  -8.277  < 2e-16 ***
age          9.359e-01  1.239e-01   7.553 4.26e-14 ***
I(age^2)    -2.490e-02  3.786e-03  -6.577 4.80e-11 ***
I(age^3)     2.842e-04  4.915e-05   5.781 7.42e-09 ***
I(age^4)    -1.180e-06  2.297e-07  -5.137 2.80e-07 ***
dur          1.118e-01  6.652e-02   1.680 0.092904 .
I(dur^2)    -8.328e-03  2.997e-03  -2.779 0.005454 **
nation1      5.686e-02  1.386e-01   0.410 0.681599
god2        -1.025e-01  5.903e-02  -1.736 0.082599 .
god3        -1.448e-01  6.780e-02  -2.136 0.032683 *
god4        -1.279e-01  7.088e-02  -1.805 0.071128 .
god5        -3.621e-02  6.695e-02  -0.541 0.588569
god6        -9.241e-02  7.505e-02  -1.231 0.218239
univ1        6.372e-01  1.729e-01   3.686 0.000228 ***
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2067.4  on 1760  degrees of freedom
Residual deviance: 1718.6  on 1747  degrees of freedom
AIC: 5196.8
```

age, age^2, age^3, age^4 ⇒ polynomial of order 4

⇒ education

⇒ minor significance.

⇒ education

$1 - \frac{1718}{2067} \approx$ 15–20% wrt dispersion

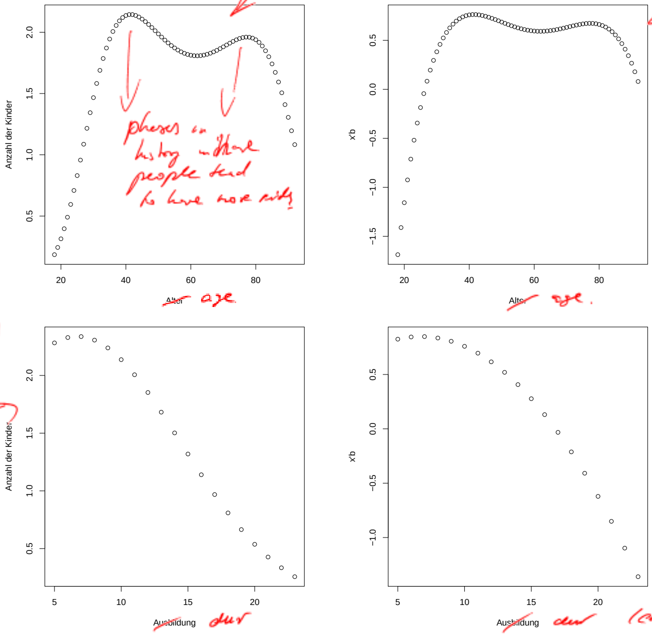
McFadden R^2

4th polynomial.

$\hat{y}$

logit

$x'\beta \to e^{x'\beta}$



$\hat{y}$

negative correlation between education & of kids.

(education)

Note: for the Poisson distribution it should hold $E(Y_i) = Var(Y_i) = \lambda_i$.

If this assumption is not fulfilled then we have *overdispersion/underdispersion*.

Solution: as an alternative we can use Quasi-Poisson- or the negative binomial distribution (negbin). Both distributions allow for different expectations and variances.

For negbin it holds:

$$P(Y_i|\boldsymbol{x}_i) = \frac{\Gamma(Y_i+\nu)}{\Gamma(\nu)\Gamma(Y_i+1)} \cdot \left(\frac{\lambda_i}{\lambda_i+\nu}\right)^{Y_i} \cdot \left(\frac{\nu}{\lambda_i+\nu}\right)^{\nu}$$

with $E(Y_i) = \lambda_i = exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ and $Var(Y_i) = \lambda_i + \lambda_i^2/\nu$.

$\nu \to +\infty$

negbin $\to$ Poisson.

allow the E and Var to be different (in the sample) generalized linear model

```
glm(formula = child ~ age + I(age^2) + I(age^3) + I(age^4) +
    dur + I(dur^2) + nation + god + univ, family = negative.binomial(theta = 1,
    link = log), data = children)
```

link function. (or exp)

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.56820  -0.50984  -0.01054   0.29990   1.90633

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.338e+01  1.267e+00 -10.555  < 2e-16 ***
age          1.022e+00  1.075e-01   9.502  < 2e-16 ***
I(age^2)    -2.730e-02  3.342e-03  -8.169 5.90e-16 ***
I(age^3)     3.126e-04  4.395e-05   7.113 1.65e-12 ***
I(age^4)    -1.302e-06  2.074e-07  -6.277 4.34e-10 ***
dur          1.269e-01  5.990e-02   2.118 0.034294 *
I(dur^2)    -9.577e-03  2.637e-03  -3.632 0.000289 ***
nation1      8.309e-02  1.349e-01   0.616 0.538128
god2        -1.186e-01  5.849e-02  -2.028 0.042743 *
god3        -1.681e-01  6.642e-02  -2.530 0.011483 *
god4        -1.563e-01  6.923e-02  -2.258 0.024075 *
god5        -3.273e-02  6.602e-02  -0.496 0.620135
god6        -1.205e-01  7.384e-02  -1.632 0.102848
univ1        7.749e-01  1.581e-01   4.900 1.04e-06 ***
---

(Dispersion parameter for Negative Binomial(1) family taken to be 0.3516262)

    Null deviance: 1023.1  on 1760  degrees of freedom
Residual deviance:  852.3  on 1747  degrees of freedom
AIC: 5911.9
```