Chapter 7

Limited dependent variables: classification

modeling binary, nominal, ordinal and count data

Modeling binary variables

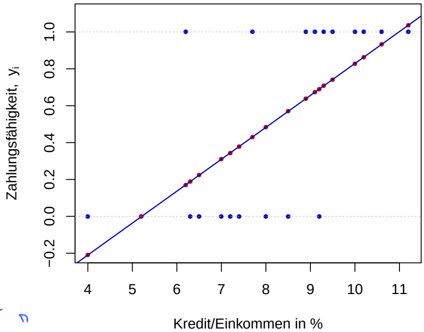
Practical question: a bank should decide about granting loans to new clients, i.e. forecast of the solvency

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

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).

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

Universität Augsburg



Linear Prob.-model

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

$E(Z) = \sum z_i \cdot P(Z = z_i)$ - expectation for a discrete RV.

expectation = probability of being insolvent

Note:

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

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

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

$$u_i = \begin{cases} 1 - \beta_0 - \beta_1 X \\ 0 - \beta_0 - \beta_1 X \end{cases}$$

Universität Augsburg

$F_z(z) = P(Z \le z)$

Transition to Logit/Probit

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

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

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

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

one can always compare two good or two bad clients

$$\begin{aligned} P(Y_i = 1|\boldsymbol{X}_i) &= P(Y_i^* > 0|\boldsymbol{X}_i) = P(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i > 0|\boldsymbol{X}_i) \\ &= P(-u_i < \boldsymbol{X}_i'\boldsymbol{\beta}|\boldsymbol{X}_i) = F(\boldsymbol{X}_i'\boldsymbol{\beta}), \end{aligned}$$

$\to$ cdf of the residuals u_i

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

Universität Augsburg

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

Logistic regression

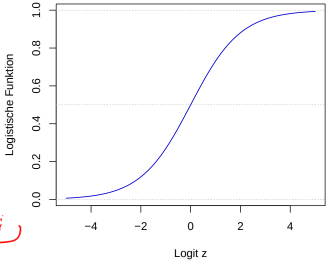
Idea: transformation with the logistic function

Logistic model

$$P(Y_i = 1|\boldsymbol{X}_i) = \frac{1}{1+e^{-z_i}};$$

for logits z_i it holds

$$z_i = \underbrace{\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}}_{\text{aggregated factor which determines your solvency.}} \cancel{+ u_i}$$



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

$y = \beta_0 + \beta_1 X_1 \ldots + \beta_j X_j + u. \Rightarrow$ cannot be visualized nicely

Universität Augsburg

Note : interpretation of the parameters is different than in case of LR

- The effect of a change in x_{ik} depends on $\boldsymbol{x}_i$ ab:

$$\frac{\partial P(Y_i=1)}{\partial x_{ik}} = \frac{\partial y_i}{\partial x_{ik}} = \frac{\partial F(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial x_{ik}} = \underbrace{\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}})^2}}_{>0}\beta_k$$

- The direction of change in $P(Y_i = 1)$ that arises due to a change in x_{ik} has the same sign as β_k

  $\beta_k > 0 \Rightarrow$ Prob. of insolvency ↗ with x_{ik}

  $\beta_k < 0 \Rightarrow$ Prob of insolvency ↘ with x_{ik}.

- If we have interactions, then the it is even more difficult:

$$\begin{aligned} z_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}x_{i2} + \beta_3 x_{i2} \\ \frac{\partial F(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial x_{i1}} &= \frac{e^{z_i}}{(1+e^{z_i})^2}\cdot(\beta_1 + \beta_2 x_{i2}) \underset{<0}{\overset{>0}{\text{or}}} \Rightarrow \end{aligned}$$

even the signs are not informative.

- Odds are helpful for interpretation (logit only!)

Odd $= 2 \Rightarrow$ Prob of insolvent is twice the prob of being solvent $\Rightarrow P(Y=1) = 2/3$ $P(Y=0) = 1/3$.

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

	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

Universität Augsburg

Estimation of the parameters: logit

The parameters are estimated using ML:

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

This simplifies to → to get rid of products and get sums.

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}=\sum_{i=1}^{n}\left[y_i-\frac{e^{z_i}}{1+e^{z_i}}\right]\boldsymbol{x}_i=\underbrace{\sum_{i=1}^{n}\left[y_i-\left(1-\frac{1}{1+e^{z_i}}\right)\right]\boldsymbol{x}_i=\mathbf{0}}_{\text{system of nonlinear equations}}$$

→ numerically

which leads to the 1st order conditions:

$$\sum_{i=1}^{n}\hat{p}_i\boldsymbol{x}_i=\sum_{i=1}^{n}y_i\boldsymbol{x}_i$$

Universität Augsburg

Estimation of the parameters: probit

Again ML:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \underbrace{\Phi(\boldsymbol{x}_i'\boldsymbol{\beta})}_{P(Y_i=1)}{}^{y_i} \cdot \underbrace{\big(1 - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})\big)}_{P(Y_i=0)}{}^{1-y_i} \longrightarrow max, \text{ w.r.t. } \beta_0, \ldots, \beta_K.$$

This simplifies to

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \left[y_i \frac{\phi(z_i)}{\Phi(z_i)} + (1 - y_i) \frac{-\phi(z_i)}{1 - \Phi(z_i)} \right] \boldsymbol{x}_i = \mathbf{0}.$$

It can be shown that (a general ML result):

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, E\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]\right)$$

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))`

in LR! $\hat{\sigma}^2(X'X)^{-1}$

Example: a data set with 700 observations   Here: Slide 439..

```
  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.434785   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
---
```

⇒ 0.05 ⇒ age irrelevant.

$\hat{\beta}_1 = 0{,}121 > 0$ ⇒ if debtinc increases, then Prob. of insolvency increases.

$\hat{\beta}_2 = -0{,}16 < 0$ ⇒ years with current employer increase (less chances of loosing job) ⇒ Prof of insolvency ↓.

	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

⇒ insignificant.

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

$e^{\hat{\beta}_i} = 1{,}12$ ⇒

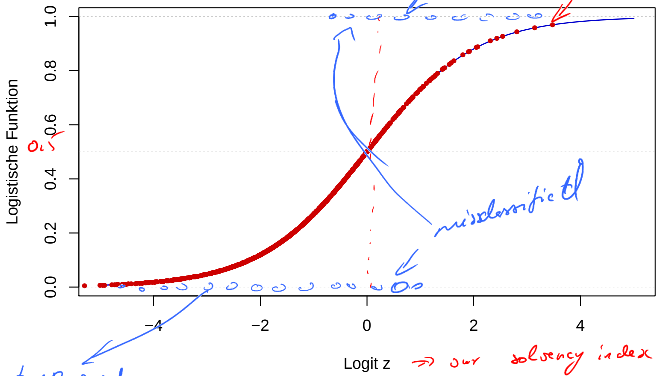
$$\frac{\overbrace{P(Y=1 \mid debtinc+1) / P(Y=0 / debtinc+1)}^{2{,}26}}{\underbrace{P(Y=1 \mid debtinc) / P(Y=0 / debtinc)}_{=2}} = 1{,}13.$$

Forecasts:

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

true bad guys

our forecasts, lie on the logistic function



true good

→ our solvency index.

$z < 0 \Rightarrow \hat{Y} = 0 \Leftarrow P(Y=1) < 0.5$

$z > 0 \Rightarrow \hat{Y} = 1 \Leftarrow P(Y=1) > 0.5$

Universität Augsburg

Goodness of the model

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

$P(Y=1)$ = fract. of 1 in sample

- pseudo-R^2:
  - Let LL_0 be the Log-Likelihood of the null model $(b_1 = \cdots = b_k = 0)$

    no features

log-lik without any features ⇒ prob. of getting the sample at hand.
$\left(\frac{N_1}{N}\right)^{N_1} \cdot \left(\frac{N_0}{N}\right)^{N_0}$

$$
\begin{aligned}
LL_0 &= \sum_{i=1}^{N} y_i \log(N_1/N) + \sum_{i=1}^{N} (1-y_i) \log(1 - N_1/N) \\
&= N_1 \log(N_1/N) + N_0 \log(N_0/N)
\end{aligned}
$$

↙ our model.

  - Let LL_v be the Log-Likelihood of the full model (with all variables)
  - Let LL_s be the Log-Likelihood of the saturated model (model with perfect fit, here $LL_s = 0$)

Prob. = 1 ⇒ log-prob = 0.

$LL_v \longrightarrow$ best model LL_s

$LL_v \longrightarrow$ worst one LL_0.

Universität Augsburg

- Deviance: $D = -2 \cdot LL_v$ (close 0)
- Cox and Snell R^2: $1 - (L_0/L_v)^{2/n}$
- Mc-Faddens-R^2: $1 - LL_v/LL_0$ (starting from 0.4)

```
    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 626.49  on 695  degrees of freedom
> library("DescTools")
> PseudoR2(logit.model, which="all")
       McFadden     McFaddenAdj        CoxSnell      Nagelkerke
     0.22113546      0.20870328      0.22438958      0.32849894
     AldrichNelson VeallZimmermann          Effron
     0.07170058      0.27698049      0.24041479
McKelveyZavoina            Tjur             AIC             BIC
     0.38471667      0.24430962    636.49076003    659.24616170
     logLik          logLik0             G2
     -313.24538001   -402.18210244    177.87344485
```

→ not a very good model $R^2 \approx 20-25\%$

Universität Augsburg

- Classification table

total number of truly bad guys

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

total number of truly good guys.

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

good guys who got no credit
Bank looses interest on the credit and good client.

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

$\to$ get money and do not pay back

$\rightsquigarrow$ $(479+72)/700 =$ 78,71% are correctly predicted. $\Rightarrow$ in 79% of decisions the decision was correct! $\Rightarrow$ looks nice.

But: there are 73,86% solvent clients in the sample. $\Rightarrow$ the above 79% look not so good now.

Quaestion: is the threshold 0.5 a good choice?

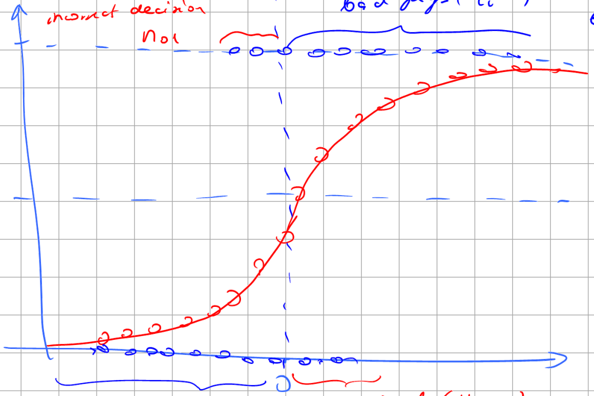
Universität Augsburg

bad predicted as good
incorrect decision
n_{01}

bad guys ($Y_i=1$) and forecasted as bad ($\hat{P}(Y=1)>0.5$)

⇒ correct decision

⇒ n_{11} such guys.



good forecasted as good ⇒ correct decision ⇒ n_{00} such guys.

good ($Y_i=0$) predicted as bad

incorrect decision ⇒ n_{10} such guys.

Client: toss 2 coins

Head Head
Head Tail
Tail Head ⇒ good

Tail Tail ⇒ bad

⇒ will give a correct classification in 75 % of cases, because there are 74% of good clients.

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.

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

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

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

Universität Augsburg

- Sensitivity$= 72/(72 + 111) = 0.39$ - only 39% of insolvent clients are classified as insolvent
- Specifity$= 479/(479 + 38) = 0.92$ - 92% of solvent clients are classified as solvent

⇒ our logit is good in identifying good clients, but bad in identifying bad clients ⇒ really dangerous.

$$
\begin{aligned}
\text{PPV or PV+} &= \frac{n_{11}}{n_{1\cdot}} = \frac{n_{11}}{n_{11} + n_{10}} \\
\text{NPV or PV-} &= \frac{n_{00}}{n_{0\cdot}} = \frac{n_{00}}{n_{01} + n_{00}}
\end{aligned}
$$

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.

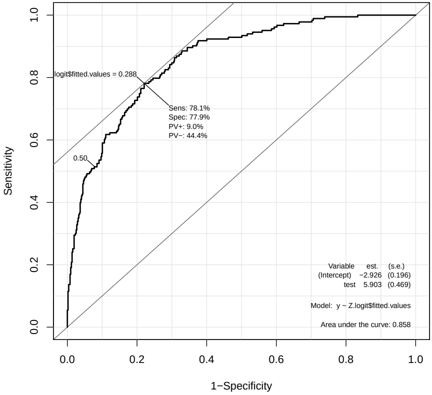
(PPV-positive predicted value, NPV-negative predicted value)

- PPV$= 72/(72 + 38) = 0.65$ - only 65% of all as insolvent classified clients are really insolvent
- NPV$= 479/(479 + 111) = 0.81$ - 81% of all as solvent classified clients are really solvent

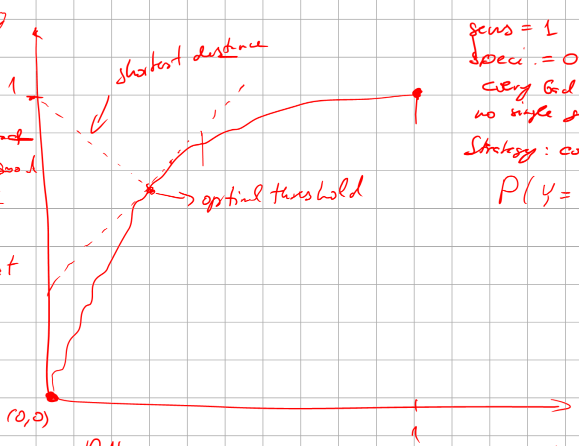
ROC-curve: senstitivity values as a function of specifity

- 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



sensit = 1
specif = 1
every bad as bad
every good as good
=> only correct decision.
"oracle" point



sens = 1
speci. = 0
every bad as bad
no single good as good
Strategy: every is a bad client
$P(Y=1) \geq 0$ ← instead of 0.5

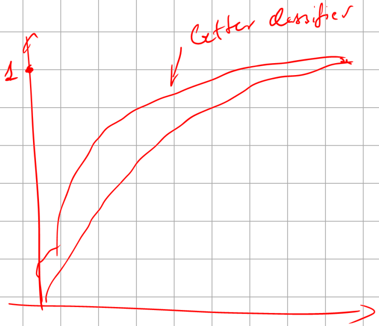
$\text{sensitivity} = 0 = \dfrac{n_{11}}{n_{\cdot 1}}$

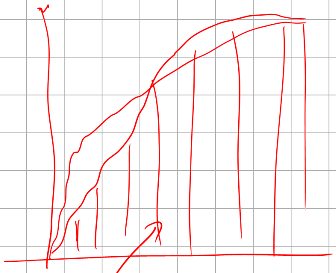
$\Rightarrow n_{11} = 0$

$\text{specificity} = 1$

=> no single bad as bad
every good as good.

=> Strategy: every client is a good client => $P(Y=1) > 1$ ← instead of 0.5





compare the areas under the curve.

optimal according to ROC.

Now let $\hat{y}_i = 1$ if $\widehat{P(Y_i = 1|\boldsymbol{X}_i)} > 0.288$ and 0 else.

```
> confusionMatrix(as.numeric(Z.logit$fitted.values>0.288), y)
Confusion Matrix and Statistics
          Reference
Prediction   0   1
         0 402  40
         1 115 143
               Accuracy : 0.7786
                 95% CI : (0.746, 0.8088)
    No Information Rate : 0.7386
    P-Value [Acc > NIR] : 0.008201
            Sensitivity : 0.7776
            Specificity : 0.7814
         Pos Pred Value : 0.9095
         Neg Pred Value : 0.5543
```

40 => bad as good (was 111) => good

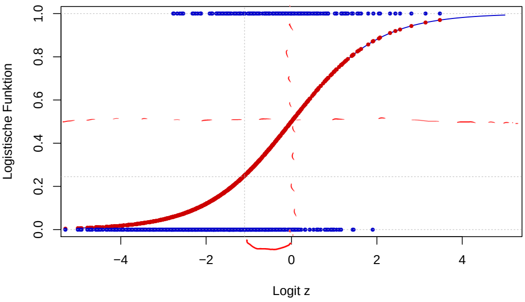
115 => good as bad (was 72) => bad

=> (402+143)/700. almost as before

=> much better => model is better in identifying bad clients.

=> worse => worse in classifying good clients.

the model is significantly better than assigning a const.

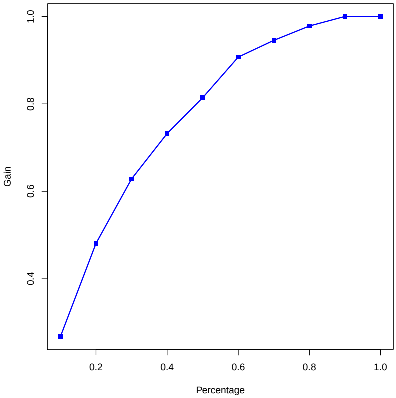


Universität Augsburg

Gain-curve:

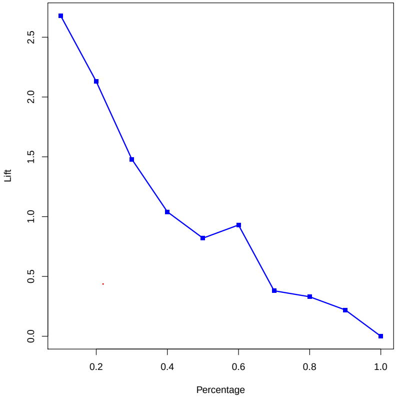
- If Gain equals 40% for 20%, this implies that if 20 % of the clients are classified as insolvent, then the algorithm will detect 40% of really insolvent clients .
- The digonal shows a model-free classification: If 20 % of the clients are classified as insolvent, then the algorithm detects 20% of really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

R: gains-package



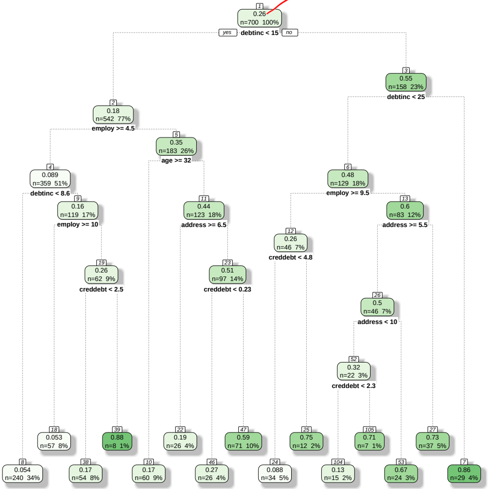
Lift curve: how much better is the predictive model compared to the model free classification?

- 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.



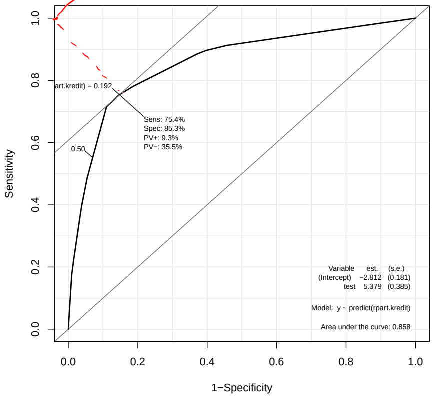
The CART method can be applied to binary data: classification trees

0,26 => % of bad clients



Universität Augsburg

oracle



$P(Y=1) \geq 0{,}192 \Rightarrow$ Bad

$< 0{,}192 \Rightarrow$ good

```
> confusionMatrix(as.numeric(predict(rpart.kredit)
       >0.192), y)
Confusion Matrix and Statistics
          Reference
Prediction   0   1
         0 420  40
         1  97 143
               Accuracy : 0.8043
                 95% CI : (0.7729, 0.8331)
    No Information Rate : 0.7386
    P-Value [Acc > NIR] : 2.822e-05
                  Kappa : 0.5395
 Mcnemar's Test P-Value : 1.715e-06
            Sensitivity : 0.8124
            Specificity : 0.7814
         Pos Pred Value : 0.9130
         Neg Pred Value : 0.5958
```

better than logit.

Test of

$H_0: n_{01} = n_{10} = 0$

$\Rightarrow$ „boosting"

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;

Example:

mode	–	“car”, “air”, “train”, oder “bus”
choice	–	decision
wait	–	waiting time, 0 for “car”
vcost	–	variable costs
travel	–	time
gcost	–	total costs
income	–	income
size	–	number of persons

	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

why car
wife's car

Universität Augsburg

Multinomial logit model

For the simple logit model it holds:

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

Odd $= e^{z}$

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

For the k categories of Y we define:

ln(Odd) as above

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

category specific vector of parameters.

reference category

with

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

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

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;
  - odds do not depend on other categories, etc.
  - Solution: Hausmann/McFadden test
- Goodness-of-fit, tests as for logit.

Global and category specific variables

category specific.

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

global

- Global variables (income, number of persons) do not depend on the categories and have individual parameters for each category: $x'_{glob}\beta^*_r$.
  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.

$$(x_{spec,r} - x_{spec,k})'\alpha \quad \text{or} \quad x'_{spec,r}\alpha$$

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*.

$\frac{P(Y=\text{train})}{P(Y=\text{air})} \rightarrow$ if income $\nearrow$

$\Rightarrow$ with higher income there is a higher chance of taking a plane than a train.

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

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 ***
---

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

more people $\Rightarrow$ higher chances for train, car, bus.

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
```

largest prob
→ car

train.

Universität Augsburg

Ordered response models

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

up to now: $Y_i^* > 0 \Rightarrow Y_i = 1$

$Y_i^* < 0 \Rightarrow Y_i = 0$

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 $\gamma_M = \infty$.

γ_1 γ_2 γ_3 $\Rightarrow$ for 4 categories

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

Universität Augsburg

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
```

Standardized.

without intercept.

in logit only β's

here: β's and γ's.

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

Universität Augsburg

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

$x'\beta + u \leq \gamma \Rightarrow u \leq \gamma - x'\beta$

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

The log-likelihood function is then given by

$$
LL(\boldsymbol{\beta}|\boldsymbol{X}) = \sum_{i=1}^{N} \log P(Y_i = y_i|\boldsymbol{x}_i) \rightarrow max, \text{ w.r.t. } \boldsymbol{\beta}, \gamma\text{'s}
$$

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

For. logit: $\prod \left(\frac{1}{1+e^{-t}}\right) \cdot \left(1 - \frac{1}{1+e^{-t}}\right)$

Example:

$\beta_1 = 2{,}14 > 0 \Rightarrow$ increasing market value increases the probab. to move to a better category.

β's

```
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
```

$\beta_2 = -0{,}7 \Rightarrow$ high dividends indicate a lower quality company.

small ⇒ insignificant

1   2   3.

−0,03   1,01

1st category

→ the 3rd category.

Universität Augsburg

Modelling for count data

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.

Poisson distribution

The Poisson distribution is frequently used to model rare events

# of events.

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

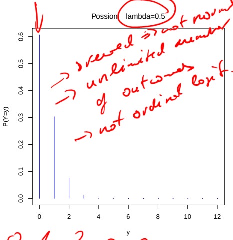
with the intensity parameter λ. It fulfils the *equidispersion*-condition:

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

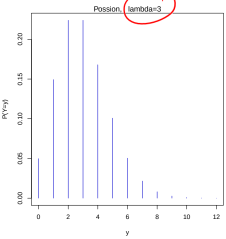
→ expected # of events per period of time.

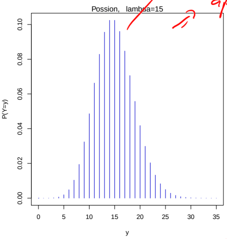
in 60% of cases you don't go to a doctor

like Normal.
→ apply LR.



→ skewed → not normal
→ unlimited number of outcomes
→ not ordinal log f.





0, 0, 1, 0 0, 1, 2, 0, 0 . . .

2, 3, 4, 3, 3 0, . . . 7

0, 14, 15, 20, 10, 15 . . .

Universität Augsburg

Poisson regression model

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

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

>0

expected number of visits for person i depends on individual characteristics in x_i.

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

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

$$\ln\left(\frac{(h(x_i'\beta))^{y_i} \cdot e^{-h(x_i'\beta)}}{y_i!}\right)$$

Universität Augsburg

Goodness of the model

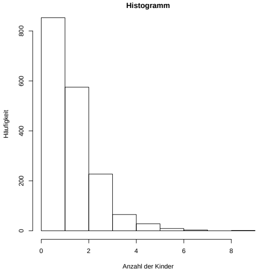
To measure the goodness of the model we use deviance, i.e. the difference between the log-likelihood for the actual observations (perfect/saturiertes model) and the log-likelihood for the predicted values:

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

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

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



```
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
```

Poisson regression.

$\exp(x'b)$

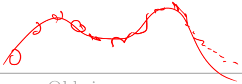
$E(\#\text{ of kids}) = \exp\{\beta_0 + \beta_1 \cdot age + \beta_2 \cdot age^2 + \dots + \beta \cdot dur + \beta\, dur^2 + \beta \cdot I(god = 2) + \beta \cdot I(god = 3) + \dots \beta \cdot I(god = 6) . +$

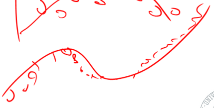
test using GLH

God = 1. 6

$y = \dots \beta \cdot God + .$

$\Rightarrow$ will





Universität Augsburg

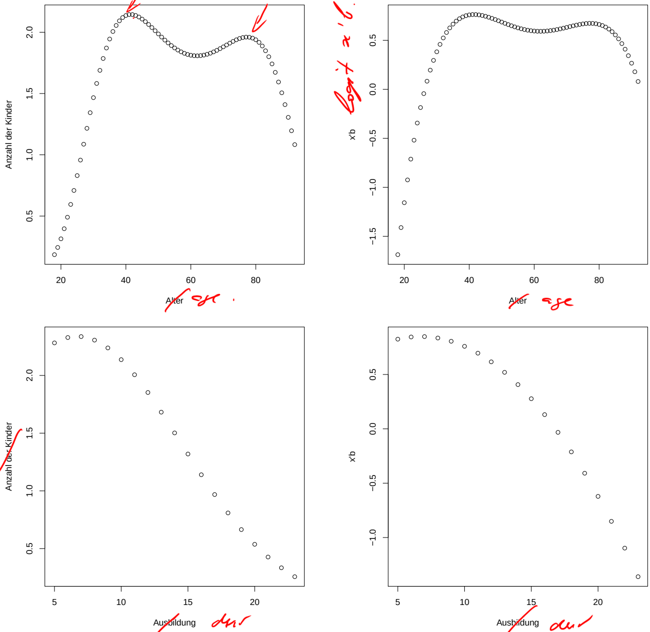
1930.

Baby Bo.. 1960 1920.

# of kids

$exp(x'\hat{\beta})$

# of kids.



Universität Augsburg

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 $Var(Y_i) = \lambda_i + \lambda_i^2/\nu$.

```
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)

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
```

related to ν $\Rightarrow$ allows E and Var to be different.