Logistická regrese LR

1. Zaměření metody LR

Navržena v 60tých letech jako alternativní postup k MNČ, když je závisle proměnná binární (medicina), kde $y$ představuje přítomnost (1) nebo nepřítomnost (0) choroby.

Jde o klasifikaci, když není splněna normalita víctrozměrného modelu.

Krajní podíly užití logistického regresního modelu za rozlišného proložení.
Příklad 8.2 Významnost sledovaných znaků ovlivňujících ústup leukemie

Lee (1980) publikoval data o leukemii pacientů a o ovlivnění jejího ústupu. Cílem je nalézt znaky, které jsou v navrženém logistickém regresním modelu statisticky významné k ovlivnění ústupu leukemie.

- **Data:** Pacientů (řádků) \( n = 27 \) a znaků (sloupců) \( m = 7 \).

Závisle proměnnou je **REMIT:** zda se objeví (1) či neobjeví (0) ústup leukemie.

Nezávisle proměnnými jsou u pacientů naměřené hodnoty 6 znaků:

- **CELL** značí celuliritu, buněčnost sraženiny kostní dřeně,
- **SMEAR** značí skvrnu diferenčního procenta napadení,
- **INFIL** značí procento infiltrátu kostní dřeně buňkou leukemie,
- **LI** je procento označeného indexu leukemických buněk kostní dřeně,
- **BLAST** je absolutní počet napadení v periferní krvi,
- **TEMP** značí nejvyšší naměřenou teplotu před začátkem léčby.

<table>
<thead>
<tr>
<th>Index</th>
<th>REMI</th>
<th>CELL</th>
<th>SMEAR</th>
<th>INFIL</th>
<th>LI</th>
<th>BLAST</th>
<th>TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.8</td>
<td>0.83</td>
<td>0.66</td>
<td>1.9</td>
<td>1.1</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>27</td>
<td>0</td>
<td>1</td>
<td>0.73</td>
<td>0.73</td>
<td>0.7</td>
<td>0.398</td>
<td>0.986</td>
</tr>
</tbody>
</table>
Řešení: Byly užity programy NCSS2000 [67], MINITAB [86] a STATISTICA [102].

Podmínky výpočtu:

Závisle proměnná: REMISS

Model obsahuje 6 nezávisle proměnných: CELL | SMEAR | INFIL | LI | BLAST | TEMP.

Objektů (řádků): 27
Znaků (sloupců): 7

Volba proměnných a výstavba modelu:

Třída závisle proměnné může nabývat hodnot, 0 a 1.
Počet je součet četností závisle proměnné pro každou třídu.
Řádky přináší počet objektů v každé třídě tak, jak byl vyčíslen z hodnot nezávisle proměnných.
Prior apriorní pravděpodobnost v každé třídě zadaná předem uživatelem.
Aktuálně vs. Predikce, $R^2$ značí hodnotu $R^2$, kterou obdržíme z regrese, kde závisle proměnná ve třídě bude lineární funkci predikované pravděpodobnosti této třídy.
% Správně klasifikováno přináší procento objektů v této třídě, které byly správně klasifikovány logistickým regresním modelem.

<table>
<thead>
<tr>
<th>REMISS</th>
<th>Třída</th>
<th>Počet</th>
<th>Řádky</th>
<th>Aktuálně</th>
<th>% Správně</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>18</td>
<td>18</td>
<td>0.66667</td>
<td>0.38775</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>0.33333</td>
<td>0.38775</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data: \( n = 27, p = 6 \),

<table>
<thead>
<tr>
<th>Index</th>
<th>REMISS</th>
<th>CELL</th>
<th>SMEAR</th>
<th>INFIL</th>
<th>LI</th>
<th>BLAST</th>
<th>TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.8</td>
<td>0.83</td>
<td>0.66</td>
<td>1.9</td>
<td>1.1</td>
<td>0.996</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.9</td>
<td>0.36</td>
<td>0.32</td>
<td>1.4</td>
<td>0.74</td>
<td>0.992</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.8</td>
<td>0.88</td>
<td>0.7</td>
<td>0.8</td>
<td>0.176</td>
<td>0.982</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0.87</td>
<td>0.87</td>
<td>0.7</td>
<td>1.053</td>
<td>0.986</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.9</td>
<td>0.75</td>
<td>0.68</td>
<td>1.3</td>
<td>0.519</td>
<td>0.982</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0.65</td>
<td>0.65</td>
<td>0.6</td>
<td>0.519</td>
<td>0.982</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.95</td>
<td>0.97</td>
<td>0.92</td>
<td>1</td>
<td>1.23</td>
<td>0.992</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0.95</td>
<td>0.87</td>
<td>0.83</td>
<td>1.9</td>
<td>1.354</td>
<td>1.02</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0.45</td>
<td>0.45</td>
<td>0.8</td>
<td>0.322</td>
<td>0.999</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0.95</td>
<td>0.36</td>
<td>0.34</td>
<td>0.5</td>
<td>0</td>
<td>1.038</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0.85</td>
<td>0.39</td>
<td>0.33</td>
<td>0.7</td>
<td>0.279</td>
<td>0.988</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0.7</td>
<td>0.76</td>
<td>0.53</td>
<td>1.2</td>
<td>0.146</td>
<td>0.982</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0.8</td>
<td>0.46</td>
<td>0.37</td>
<td>0.4</td>
<td>0.38</td>
<td>1.006</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0.2</td>
<td>0.39</td>
<td>0.08</td>
<td>0.8</td>
<td>0.114</td>
<td>0.99</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>0.9</td>
<td>0.9</td>
<td>1.1</td>
<td>1.037</td>
<td>0.99</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>0.84</td>
<td>0.84</td>
<td>1.9</td>
<td>2.064</td>
<td>1.02</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>0.65</td>
<td>0.42</td>
<td>0.27</td>
<td>0.5</td>
<td>0.114</td>
<td>1.014</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0.75</td>
<td>0.75</td>
<td>1</td>
<td>1.322</td>
<td>1.004</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>0.5</td>
<td>0.44</td>
<td>0.22</td>
<td>0.6</td>
<td>0.114</td>
<td>0.99</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
<td>0.63</td>
<td>0.63</td>
<td>1.1</td>
<td>1.072</td>
<td>0.986</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>1</td>
<td>0.33</td>
<td>0.33</td>
<td>0.4</td>
<td>0.176</td>
<td>1.01</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>0.9</td>
<td>0.93</td>
<td>0.84</td>
<td>0.6</td>
<td>1.591</td>
<td>1.02</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>0.95</td>
<td>0.32</td>
<td>0.3</td>
<td>1.6</td>
<td>0.531</td>
<td>1.002</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
<td>0.6</td>
<td>1.7</td>
<td>0.964</td>
<td>0.988</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>1</td>
<td>0.69</td>
<td>0.69</td>
<td>0.9</td>
<td>0.398</td>
<td>0.986</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>1</td>
<td>0.69</td>
<td>0.69</td>
<td>0.9</td>
<td>0.398</td>
<td>0.986</td>
</tr>
<tr>
<td>27</td>
<td>0</td>
<td>1</td>
<td>0.73</td>
<td>0.73</td>
<td>0.7</td>
<td>0.398</td>
<td>0.986</td>
</tr>
</tbody>
</table>
Iterační přibližení odhadů parametrů:

zvyšuje logit metodou maximální věrohodnosti:

<table>
<thead>
<tr>
<th></th>
<th>0. iter.</th>
<th>1. iter.</th>
<th>2. iter.</th>
<th>3. iter.</th>
<th>4. iter.</th>
<th>5. iter.</th>
<th>6. iter.</th>
<th>7. iter.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Úsek</td>
<td>0.0000</td>
<td>-18.7941</td>
<td>-43.8181</td>
<td>-64.4856</td>
<td>-66.9072</td>
<td>-59.9205</td>
<td>-58.1082</td>
<td>-58.0387</td>
</tr>
<tr>
<td>LI</td>
<td>0.0000</td>
<td>0.7895</td>
<td>0.4785</td>
<td>-1.7390</td>
<td>-8.4324</td>
<td>-19.9753</td>
<td>-24.3896</td>
<td>-24.6605</td>
</tr>
<tr>
<td>CELL</td>
<td>0.0000</td>
<td>5.8190</td>
<td>7.8289</td>
<td>6.5724</td>
<td>-0.7276</td>
<td>-14.0188</td>
<td>-18.9941</td>
<td>-19.2925</td>
</tr>
<tr>
<td>TEMP</td>
<td>0.0000</td>
<td>-5.9853</td>
<td>-8.4942</td>
<td>-7.5008</td>
<td>0.0725</td>
<td>14.0932</td>
<td>19.2909</td>
<td>19.6001</td>
</tr>
<tr>
<td>SMEAR</td>
<td>0.0000</td>
<td>-2.1173</td>
<td>-2.8763</td>
<td>-3.3905</td>
<td>-3.6410</td>
<td>-3.7879</td>
<td>-3.8874</td>
<td>-3.8959</td>
</tr>
<tr>
<td>INFIL</td>
<td>0.0000</td>
<td>-0.0030</td>
<td>-0.0152</td>
<td>-0.0980</td>
<td>-0.1186</td>
<td>-0.1456</td>
<td>-0.1513</td>
<td>-0.1511</td>
</tr>
<tr>
<td>BLAST</td>
<td>0.0000</td>
<td>20.6740</td>
<td>47.3696</td>
<td>71.2305</td>
<td>80.4766</td>
<td>84.6743</td>
<td>87.2300</td>
<td>87.4331</td>
</tr>
<tr>
<td><strong>Logit</strong></td>
<td><strong>-18.7150</strong></td>
<td><strong>-12.3083</strong></td>
<td><strong>-11.3620</strong></td>
<td><strong>-11.0574</strong></td>
<td><strong>-10.9449</strong></td>
<td><strong>-10.8812</strong></td>
<td><strong>-10.8753</strong></td>
<td><strong>-10.8753</strong></td>
</tr>
</tbody>
</table>

Logit představuje rozhodčí kritérium, pomocí kterého se rozhodne, zda dotyčný parametr logistický model zlepší nebo zhorší.
Průběh postupného krokového zavádění parametrů
do logistického regresního modelu

*\( R^2 \) modelu* je odhad koeficientu determinace \( R^2 \) pro logistický model proložený daty. *Změna v \( R^2 \)* udává hodnotu, která se přičte k celkové \( R^2 \), když se tento parametr přidá do modelu. Změny \( R^2 \) mají vesměs kladné znaménko, zavedení dalšího parametru do modelu způsobí zvýšení hodnoty \( R^2 \) čili zlepšení modelu až na konečnou 36.719 %.

\[
R^2 = \frac{\chi^2(df)}{\chi^2(df) + n - p - 1},
\]

kde \( df \) značí stupně volnosti u \( \chi^2 \)-testu a jsou rovny počtu nezávisle proměnných.

<table>
<thead>
<tr>
<th>Krok výstavby</th>
<th>Parametr zaveden</th>
<th>Logit</th>
<th>Dosažená ( R^2 )</th>
<th>Změna v ( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Úsek</td>
<td>-17.18588</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>2</td>
<td>LI</td>
<td>-13.03648</td>
<td>0.24144</td>
<td>0.24144</td>
</tr>
<tr>
<td>3</td>
<td>CELL</td>
<td>-12.17036</td>
<td>0.29184</td>
<td>0.05040</td>
</tr>
<tr>
<td>4</td>
<td>TEMP</td>
<td>-10.97669</td>
<td>0.36130</td>
<td>0.06946</td>
</tr>
<tr>
<td>5</td>
<td>SMEAR</td>
<td>-10.92900</td>
<td>0.36407</td>
<td>0.00277</td>
</tr>
<tr>
<td>6</td>
<td>INFIL</td>
<td>-10.87752</td>
<td>0.36707</td>
<td>0.00300</td>
</tr>
<tr>
<td>7</td>
<td>BLAST</td>
<td>-10.87533</td>
<td>0.36719</td>
<td>0.00013</td>
</tr>
</tbody>
</table>

Vysvětlení: Z počátečního logitu \( \ln L_{(1)} = -17.18588 \) až do terminace \( \ln L_{(1)} = -10.87533 \) pro \( REMISS = 0 \).
**Test postaveného logistického regresního modelu**

vychází z $H_0$: „Směrnice všech parametrů jsou nulové“

proti alternativní $H_A$: „.... nejsou nulové“.

**Kritérium $G = D$ (model bez proměnné) – $D$ (model s proměnnými) =

$[(−2) \times (−17.18588)] − [(−2) \times (−10.87533)] = 12.621$

pro 6 stupňů volnosti dává vypočtenou hladinu významnosti $P = 0.049463$

menší než $\alpha = 0.05$,

a proto je $H_0$ o nulovosti všech parametrů v modelu zamítnuta.

**Vysvětlení:** rozdíl deviancí $G$ testuje,

zda všechny regresní parametry $\beta$, kromě úseku $\beta_0$ jsou rovny nule.

**Slovy:** $D$ je dvojnásobek rozdílu mezi logaritmem pravděpodobnosti nalezeného logistického regresního modelu

a modelu, ve kterém je pouze úsek.

**Spočtená hladina významnosti $P = 0.049463$ je menší než $\alpha = 0.05$ a ukazuje, že směrnice parametrů $b_i$ v modelu nejsou rovny nule a navržený model je proto statisticky významný.**

<table>
<thead>
<tr>
<th>$R^2$ modelu</th>
<th>$df$</th>
<th>Odchylka $D$</th>
<th>Spočtená hladina významnosti $P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3869</td>
<td>6</td>
<td>12.62</td>
<td>0.049463</td>
</tr>
</tbody>
</table>
Odhad regresních parametrů a test jejich významnosti:

*Parametr* $b_i$ značí název $\beta_i$ pro nezávisle proměnnou, úsek je značen $\beta_0$.
*Odhad* $b_i$ značí směrnici regresního parametru $\beta_i$, vyčíslený minimalizací.
*Směrodatná odchylka* je odhadem přesnosti odhadovaného parametru.
*χ²* pro nulovost parametru $\beta = 0$ udává Pearsonovo testační kritérium $\chi^2$ pro 1 stupeň volnosti k testování $H_0: \beta_i = 0$ vs. $H_A: \beta_i \neq 0$, vyčíslí se dle Waldova kritéria $W^2_{a,i} = \left[ b_i / s(b_i) \right]^2$.
*P* značí spočtenou hladinu významnosti pro nalezené odhady: je-li $P$ menší než zvolená $\alpha = 0.05$, je nulovost $\beta_i = 0$ zamítnuta a odhad parametru $b_i$ je statisticky významný.

**Vysvětlení:** až na BLAST a LI jsou všechny znaky statisticky významné.

<table>
<thead>
<tr>
<th>Parameter $b_i$</th>
<th>Odhad $b_i$</th>
<th>Směrodatná odchylka pro $\beta_i = 0$</th>
<th>hladina $P$</th>
<th>Dolní 95% mez</th>
<th>Horní 95% mez</th>
</tr>
</thead>
<tbody>
<tr>
<td>B0: Úsek</td>
<td>-58.03849</td>
<td>1.90413</td>
<td>30.480</td>
<td>0.00000</td>
<td>-61.77051</td>
</tr>
<tr>
<td>B3: INFIL</td>
<td>19.60126</td>
<td>1.33668</td>
<td>14.664</td>
<td>0.00000</td>
<td>16.98142</td>
</tr>
<tr>
<td>B4: LI</td>
<td>-3.89596</td>
<td>2.35303</td>
<td>1.656</td>
<td>0.09778</td>
<td>-8.50782</td>
</tr>
<tr>
<td>B5: SMEAR</td>
<td>-19.29357</td>
<td>1.40647</td>
<td>13.718</td>
<td>0.00000</td>
<td>-22.05020</td>
</tr>
<tr>
<td>B6: TEMP</td>
<td>87.43390</td>
<td>1.89150</td>
<td>46.225</td>
<td>0.00000</td>
<td>83.72663</td>
</tr>
<tr>
<td>B1: BLAST</td>
<td>-0.15109</td>
<td>1.85979</td>
<td>-0.081</td>
<td>0.93525</td>
<td>-3.79620</td>
</tr>
</tbody>
</table>
Nalezený logistický regresní model:

Dle statistické významnosti odhadů parametrů byl stanoven logistický regresní model pro $REMISS = 0$:

$$-58.04 - 24.66 \text{CELL} + 19.60 \text{INFIL} - 3.90 \text{LI}$$
$$- 19.29 \text{SMEAR} + 87.43 \text{TEMP} + 0.15 \text{BLAST}.$$
*Pearsonův test dobré shody kvality těsnosti proložení, porovnává dané a nalezené četnosti.*

Data jsou roztříděna dle svých odhadnutých pravděpodobností od nejmenší do největší z 10 tříd:

1. **třída** obsahuje data s nejmenší vyčíslenou pravděpodobností a
10. **třída** s největší pravděpodobností.

**Vysvětlení:** Model zde prokládá data dobře, dané a vypočtené četnosti jsou si podobné.

**Hosmerova–Lemeshowova statistika** (= Pearsonovo $\chi^2$-kritérium) nabývá hodnoty spočtené hladiny významnosti $P = 0.477$, větší než $\alpha = 0.05$ a proto dobré proložení.

<table>
<thead>
<tr>
<th>Třída</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Celkem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nízký, nezměněný puls (1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dáno</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Nalezeno</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.3</td>
<td>0.6</td>
<td>0.9</td>
<td>0.8</td>
<td>1.6</td>
<td>2.1</td>
<td>2.6</td>
<td></td>
</tr>
<tr>
<td><strong>Vysoký, změněný puls (0)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dáno</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Nalezeno</td>
<td>2</td>
<td>3</td>
<td>2.9</td>
<td>1.7</td>
<td>2.4</td>
<td>2.1</td>
<td>1.2</td>
<td>1.4</td>
<td>0.9</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Celkem</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>27</td>
</tr>
</tbody>
</table>
ROC graf (tabulka)

K určení prahového bodu $P_c$ při klasifikaci objektů do dvou tříd, v události 1 a v neudálosti 0, se použijí křivky ROC.

Objekt se zařadí do té třídy, kam se nejvíce přiblížuje svou vyčíslenou pravděpodobností.

Z tabulky ROC je zřejmé, co nastane, když vybereme rozličné hodnoty pro prahový dělicí bod pravděpodobnosti $P_c$.

Při zařazování objektů do třídy může nastat jedna ze čtyř možností:

**Sloupec A, značený $N(1|1)$:** objekt je správně zařazen do této třídy 1 a pochází ze třídy 1.

**Sloupec C, značený $N(0|1)$:** objekt je nesprávně zařazen do opačné třídy 0 a pochází přitom ze třídy 1.

**Sloupec B, značený $N(1|0)$:** objekt je nesprávně zařazen do třídy 1 a pochází přitom z opačné třídy 0.

**Sloupec D, značený $N(0|0)$:** objekt je správně zařazen do opačné třídy 0 a pochází přitom z opačné třídy 0.
## Tabulka ROC pro REMISS = 0

| $P_C$ | $N(1|1)$ | $N(1|0)$ | $N(0|1)$ | $N(0|0)$ | Senzitivita | Specificita | Senzitivita | Podíl správně zařazených |
|-------|-----------|-----------|-----------|-----------|--------------|-------------|--------------|--------------------------|
| 0.05  | 18        | 9         | 0         | 0         | 1.000        | 0.000       | 1.000        | 0.667                   |
| 0.10  | 18        | 8         | 0         | 1         | 1.000        | 0.111       | 1.111        | 0.704                   |
| 0.15  | 17        | 8         | 1         | 1         | 0.944        | 0.111       | 1.056        | 0.667                   |
| 0.20  | 17        | 8         | 1         | 1         | 0.944        | 0.111       | 1.056        | 0.667                   |
| 0.25  | 17        | 7         | 1         | 2         | 0.944        | 0.222       | 1.167        | 0.704                   |
| 0.30  | 17        | 6         | 1         | 3         | 0.944        | 0.333       | 1.278        | 0.741                   |
| 0.35  | 17        | 4         | 1         | 5         | 0.944        | 0.556       | 1.500        | 0.815                   |
| 0.40  | 16        | 4         | 2         | 5         | 0.889        | 0.556       | 1.444        | 0.778                   |
| 0.45  | 15        | 4         | 3         | 5         | 0.833        | 0.556       | 1.389        | 0.741                   |
| 0.50  | 15        | 4         | 3         | 5         | 0.833        | 0.556       | 1.389        | 0.741                   |
| 0.55  | 15        | 3         | 3         | 6         | 0.833        | 0.667       | 1.500        | 0.778                   |
| 0.60  | 15        | 2         | 3         | 7         | 0.833        | 0.778       | 1.611        | 0.815                   |
| 0.65  | 15        | 1         | 3         | 8         | 0.833        | 0.889       | 1.722        | 0.852                   |
| 0.70  | 15        | 1         | 3         | 8         | 0.833        | 0.889       | 1.722        | 0.852                   |
| 0.75  | 12        | 1         | 6         | 8         | 0.667        | 0.889       | 1.556        | 0.741                   |
| 0.80  | 11        | 0         | 7         | 9         | 0.611        | 1.000       | 1.611        | 0.741                   |
| 0.85  | 8         | 0         | 10        | 9         | 0.444        | 1.000       | 1.444        | 0.630                   |
| 0.90  | 8         | 0         | 10        | 9         | 0.444        | 1.000       | 1.444        | 0.630                   |
| 0.95  | 8         | 0         | 10        | 9         | 0.444        | 1.000       | 1.444        | 0.630                   |

Plocha pod křivkou ROC: 0.87963
Počet zařazených objektů do obou tříd, do třídy 0 nebo do třídy 1, je vyčíslován pro různé hodnoty prahového dělicího bodu $P_c$.

**Senzitivita** je podíl správně zařazených objektů do první třídy A/(A+C).

**Specificita** je podíl správně zařazených objektů do opačné třídy D/(B+D).

(Pro optimální hodnotu prahového bodu $P_c$ dosahuje suma senzitivity + specificity maximální hodnoty).

**Podíl správně zařazených objektů:** vybere se taková hodnota prahového dělicího bodu $P_c$, pro kterou existuje právě maximální podíl správně zařazených objektů.

Nejspolehlivější pravidlo maximální plochy $AUC$ pod křivkou $ROC$ ukazuje, že všechny objekty jsou správně zařazeny, čím více se plocha blíží jedné nebo 100 %.
Tabulka $ROC$ pro REMISS = 1

| $P_C$ | $N(1|1)$ | $N(1|0)$ | $N(0|1)$ | $N(0|0)$ | Senzitivita | Specificita | Sensitivita | Podíl správně zařazených |
|-------|----------|----------|----------|----------|-------------|-------------|-------------|-------------------------|
| 0.05  | 9        | 10       | 0        | 8        | 1.000       | 0.444       | 1.444       | 0.630                   |
| 0.10  | 9        | 10       | 0        | 8        | 1.000       | 0.444       | 1.444       | 0.630                   |
| 0.15  | 9        | 10       | 0        | 8        | 1.000       | 0.444       | 1.444       | 0.630                   |
| 0.20  | 9        | 7        | 0        | 11       | 1.000       | 0.611       | 1.611       | 0.741                   |
| 0.25  | 8        | 6        | 1        | 12       | 0.889       | 0.667       | 1.556       | 0.741                   |
| 0.30  | 8        | 3        | 1        | 15       | 0.889       | 0.833       | 1.722       | 0.852                   |
| 0.35  | 8        | 3        | 1        | 15       | 0.889       | 0.833       | 1.722       | 0.852                   |
| 0.40  | 7        | 3        | 2        | 15       | 0.778       | 0.833       | 1.611       | 0.815                   |
| 0.45  | 6        | 3        | 3        | 15       | 0.667       | 0.833       | 1.500       | 0.778                   |
| 0.50  | 5        | 3        | 4        | 15       | 0.556       | 0.833       | 1.389       | 0.741                   |
| 0.55  | 5        | 3        | 4        | 15       | 0.556       | 0.833       | 1.389       | 0.741                   |
| 0.60  | 5        | 2        | 4        | 16       | 0.556       | 0.889       | 1.444       | 0.778                   |
| 0.65  | 5        | 1        | 4        | 17       | 0.556       | 0.944       | 1.500       | 0.815                   |
| 0.70  | 3        | 1        | 6        | 17       | 0.333       | 0.944       | 1.278       | 0.741                   |
| 0.75  | 2        | 1        | 7        | 17       | 0.222       | 0.944       | 1.167       | 0.704                   |
| 0.80  | 1        | 1        | 8        | 17       | 0.111       | 0.944       | 1.056       | 0.667                   |
| 0.85  | 1        | 1        | 8        | 17       | 0.111       | 0.944       | 1.056       | 0.667                   |
| 0.90  | 1        | 0        | 8        | 18       | 0.111       | 1.000       | 1.111       | 0.704                   |
| 0.95  | 0        | 0        | 9        | 18       | 0.000       | 1.000       | 1.000       | 0.667                   |

Plocha pod křivkou $ROC$: 0.87963
Klasifikace objektů logistickým modelem:

Predikce je určena z logistického regresního modelu při užití klasifikačního prahového bodu.

Procento správně klasifikovaných objektů je procento z celkového počtu, které padne na diagonálu tabulky.

<table>
<thead>
<tr>
<th>Dáno závisle proměnnou</th>
<th>Nalezeno predikcí logistickým modelem</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ne</td>
</tr>
<tr>
<td>četnost</td>
<td>15</td>
</tr>
<tr>
<td>řádkové procento</td>
<td>83.33</td>
</tr>
<tr>
<td>sloupcové procento</td>
<td>78.95</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>četnost</td>
<td>4</td>
</tr>
<tr>
<td>řádkové procento</td>
<td>44.44</td>
</tr>
<tr>
<td>sloupcové procento</td>
<td>21.05</td>
</tr>
<tr>
<td></td>
<td>19</td>
</tr>
<tr>
<td>četnost</td>
<td>19</td>
</tr>
<tr>
<td>řádkové procento</td>
<td>70.37</td>
</tr>
<tr>
<td>sloupcové procento</td>
<td>100</td>
</tr>
</tbody>
</table>

Procento správně klasifikovaných = 74.07
a) Predikovaná klasifikace:

*Daná třída* značí třídu, do které objekt skutečně patří.
*Nalezená třída* značí třídu vypočtenou dle parametrů logistického regresního modelu.
*Logistické skóre* je odhad pravděpodobnosti, že objekt patří do třídy *Ne*.
*Reziduum* představuje rozdíl mezi vypočtenou hodnotou logistického skóre a binárním indexem skutečné třídy. Vyčísli se: za index třídy *Ne* se dosadí 0 a za index třídy *Ano* dosadí 1.

<table>
<thead>
<tr>
<th>Řádek</th>
<th>Daná třída</th>
<th>Nalezená třída</th>
<th>Logistické skóre</th>
<th>Reziduum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ano (1)</td>
<td>Ano (1)</td>
<td>0.79</td>
<td>0.209</td>
</tr>
<tr>
<td>2</td>
<td>Ano (1)</td>
<td>Ne (0)</td>
<td>0.435</td>
<td>0.565</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>27</td>
<td>Ne (0)</td>
<td>Ne (0)</td>
<td>0.279</td>
<td>-0.279</td>
</tr>
</tbody>
</table>

Závěrečná tabulka klasifikovaných objektů

na základě navrženého logistického modelu.

<table>
<thead>
<tr>
<th>Nalezeno</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dáno</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>Celkem</td>
</tr>
<tr>
<td>Procento správně klasifikovaných 74.1 %</td>
</tr>
<tr>
<td>Řádek</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td>27</td>
</tr>
</tbody>
</table>

Daná třída určuje zadanou skutečnou třídu. Nalezená třída představuje nalezenou třídu na základě logistického regresního modelu. Logistické skóre je odhad pravděpodobnosti, že objekt patří do třídy Ne. Reziduum představu je to rozdíl mezi Logistickým skóre a indexem skutečné třídy. Index třídy Ne je 0 a index třídy Ano je 1.
b) *Chybně klasifikované objekty:*

Jsou zde zobrazeny pouze chybně zařazené řádky.

<table>
<thead>
<tr>
<th>Řádek</th>
<th>Daná třída</th>
<th>Nalezená třída</th>
<th>Skóre</th>
<th>Reziduum</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Ano (1)</td>
<td>Ne (0)</td>
<td>0.435</td>
<td>0.565</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>26</td>
<td>Ano (1)</td>
<td>Ne (0)</td>
<td>0.461</td>
<td>0.539</td>
</tr>
</tbody>
</table>
Grafy kvality logistické regresní analýzy.

Graf prahové operační charakteristiky ROC
vystihuje správnost diagnostického testu,
zda logistickým modelem vypočtené Ano nebo Ne je správné.

Na ose $y$ se vynáší senzitivita a na ose $x$ hodnota „1 minus specificita“. Křivky pomohou nalézt nejlepší hodnotu dělicího bodu.

Obr. 8.3a Graf ROC pro obě hodnoty REMISS = 0 a 1.

Obr. 8.3b Graf ROC pro hodnotu REMISS = 0.
Graf prahové operační charakteristiky ROC vystihuje správnost diagnostického testu, zda logistickým modelem vypočtené Ano nebo Ne je správné.

Na ose $y$ se vynáší senzitivita a na ose $x$ hodnota „1 minus specificita“. Křivky pomohou nalézt hodnotu dělicího bodu $P_C$ ke klasifikaci objektů.

Když leží dělicí bod $P_C$ v levém horním rohu grafu, je dosaženo nejvhodnější hodnoty a dochází k nejlepší klasifikaci objektů.

Kvalita nalezeného logistického modelu se posuzuje také dle plochy $AUC$ pod křivkou $ROC$: Čím více se $AUC$ blíží jedné nebo 100 %, tím je klasifikace objektů lepší.
Graf podílu správně zařazených objektů v závislosti na $P_C$

je velice užitečný graf k určení nejlepší hodnoty prahového dělicího bodu $P_C$.

Na ose $y$ je procento správně zařazených objektů a na ose $x$ hodnoty pravděpodobností dělicího bodu $P_C$ v jednotkách vyčíslované pravděpodobnosti.

Obr. 8.3d Graf závislosti obou křivek podílu správně zařazených objektů na $P_C$ (cutoff).
Nalezený logistický regresní model:

Dle statistické významnosti odhadů parametrů byl stanoven logistický regresní model pro $REMIS = 0$:

$$-58.04 - 24.66 \text{CELL} + 19.60 \text{INFIL} - 3.90 \text{LI}$$
$$- 19.29 \text{SMEAR} + 87.43 \text{TEMP} + 0.15 \text{BLAST}.$$ 

○ Závěr:

1. Byl nalezen logistický regresní model znaků $CELL$, $SMEAR$, $INFIL$, $LI$, $BLAST$, a $TEMP$, které významně ovlivňují znak ústupu leukemie $REMIS$. 

2. Znaky $LI$ a $BLAST$ nejsou statisticky významné.

3. Z $ROC$ byl odhadnut prahový dělicí bod pravděpodobnosti, dle kterého se objekty spolehlivě zařadí do dvou tříd ústupu a neústupu leukemie.
Logistická regrese LR se liší od lineární regrese: predikuje pravděpodobnost události, která se buď stala (1) nebo nestala (0).

Logitová transformace vede na sigmoidální vztah mezi závisle proměnnou $y$ a vektorem nezávisle proměnných $x$.

Při velmi nízkých hodnotách $x$ se pravděpodobnost proměnné $y$ blíží k nule. Při vysokých hodnotách $x$ se blíží k jedné.

Logistická regrese používá kategorickou závisle proměnnou zatímco lineární regrese užívá pouze spojitou vysvětlovanou proměnnou.
Logitová transformace vychází z poměru šancí či naděje.

Dle typu závisle proměnné y se rozlišují:

**Binární logistická regrese**: binární závisle proměnná nabývá pouze dvou hodnot, například přítomnost-absence, muž-žena. Vektor nezávisle proměnných $\mathbf{x}$ obsahuje jednu či více spojitých proměnných (prediktory) nebo diskrétních, kategorických (faktory).

**Ordinální logistická regrese**: ordinální závisle proměnná nabývá tři a více možných stavů, např. silný nesouhlas, nesouhlas, souhlas, silný souhlas. Vektor $\mathbf{x}$ nezávisle proměnných obsahuje jak prediktory tak i faktory.

**Nominální logistická regrese**: nominální závisle proměnná o více než třech úrovních, např. mezi kterými je definována pouze odlišnost. Vektor $\mathbf{x}$ může obsahovat jak prediktory, tak i faktory.
2. Logistický regresní model

V LR potřebujeme vědět, zda se událost stala (1) nebo nestala (0).

Jde o dichotomickou hodnotu 0 - 1 závisle proměnné \( y \), ze které se predikuje odhad pravděpodobnosti, že se událost stala (1) či nestala (0).

Je-li predikovaná pravděpodobnost větší než 0.50, pak se událost stala (1), je-li menší než 0.50, pak se nestala (0).

Postup LR porovnává pravděpodobnost události odehrané \( L_{(1)} \) vůči pravděpodobnosti události neodehrané \( L_{(0)} = 1 - L_{(1)} \).

Využijeme pravděpodobnostní poměr \( L_{(1)}/L_{(0)} \), ve kterém pravděpodobnost \( L_{(1)} \) je vyjádřena logistickou funkcí

\[
L_{(1)} = \frac{1}{1 + e^{C - z}}
\]
2. Logistický regresní model

V LR potřebujeme vědět, zda se událost stala (1) nebo nestala (0).

Jde o dichotomickou hodnotu 0 - 1 závisle proměnné $y$, ze které se predikuje odhad pravděpodobnosti, že se událost stala (1) či nestala (0).

Je-li predikovaná pravděpodobnost větší než 0.50, pak se událost stala (1), je-li menší než 0.50, pak se nestala (0).

Postup LR porovnává pravděpodobnost události odehrané $L_{(1)}$ vůči pravděpodobnosti události neodehrané $L_{(0)} = 1 - L_{(1)}$.

Využijeme pravděpodobnostní poměr $L_{(1)}/L_{(0)}$, ve kterém pravděpodobnost $L_{(1)}$ je vyjádřena logistickou funkcí

$$L_{(1)} = \frac{1}{1 + e^{c - z}}$$
**Pravděpodobnostní poměr** (zvaný “poměr šancí”) je vyjádřen

\[
\frac{L_{(1)}}{L_{(0)}} = e^{a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_p x_p}
\]

kde odhadované koeficienty \(a_0, a_1, a_2, \ldots, a_p\) jsou míry změny poměru obou pravděpodobností \(L_{(1)}/L_{(0)}\).

Poměr je lineární funkcí diskriminační funkce o \(p\) nezávisle proměnných

\[
Z = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_p x_p
\]

Po zlogaritmování a úpravě vyjde

\[
C - Z = \ln \left( \frac{L_{(0)}}{L_{(1)}} \right)
\]

kde \(C\) je absolutní člen \(a_0\).
Dle klasifikačního postupu je

\[ L_{(0)} = P(G = 1 \mid x) \] a
\[ L_{(1)} = P(G = 1 \mid x) = 1 - P(G = 1 \mid x). \]

a po úpravách bude

\[
\ln \left( \frac{L_{(1)}}{L_{(0)}} \right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p
\]

kde \( b_0 = -C + a_0, \ b_i = a_i \) pro \( i = 1, \ldots, p. \)

**Například:** ve sportu řekneme, že tým má šanci 3:1. Tvrzení říká, že favorizovaný tým má pravděpodobnost vítězství \( \frac{3}{3 + 1} = \frac{3}{4} = 0.75. \)

Platí tedy pravděpodobnostní poměr

\[
\frac{L_{(1)}}{L_{(0)}} = \frac{0.75}{1 - 0.75} = \frac{3}{1}.
\]
Aposteriorní pravděpodobnost $P(G = j \mid x)$ zařazení do $j$-té kategorie: logistický model lze rozšířit na případ $K$ tříd, a předpokládat, že aposteriorní pravděpodobnost $P(G = j \mid x)$ zařazení do $j$-té kategorie bude

$$\ln \frac{P(G = 1 \mid x)}{P(G = K \mid x)} = b_{1, \ 0} + b_1^T x$$

$$\ln \frac{P(G = 2 \mid x)}{P(G = K \mid x)} = b_{2, \ 0} + b_2^T x$$

$$...$$

$$\ln \frac{P(G = K-1 \mid x)}{P(G = K \mid x)} = b_{K-1, \ 0} + b_K^T x$$
Po zpětné transformaci vychází

\[
P(G = j \mid x) = \frac{\exp (b_{j,0} + b_{j}^T x)}{1 + \sum_{l=1}^{K-1} \exp (b_{l,0} + b_{l}^T x)}
\]

a

\[
P(G = K \mid x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp (b_{l,0} + b_{l}^T x)}
\]

Označíme pravděpodobnost

\[
P(G = K \mid x) = p_k(x, \mathbf{b}),
\]

aby se zvýraznilo, že jde o funkci regresních parametrů

\[
\mathbf{b} = [b_{1,0}, b_{1}, b_{2,0}, b_{2}, \ldots, b_{K-1,0}, b_{K-1}].
\]

(Pro \(K = 2\) přechází tento model na standardní logistický model pro binární vysvětlovanou proměnnou \(y = G\)).
Odhady parametrů:

Pro odhad parametrů logistických modelů se používá metoda maximální věrohodnosti.

Přítomnost v první třídě \( y = 1 \) je pro \( G = 1 \).
Nepřítomnost v první třídě \( y = 0 \) je pro \( G = 2 \) čili přítomnost ve druhé třídě.

**Výchozí data:** vektor \( y \) rozměru \( n \times 1 \) a matice \( X \) rozměru \( n \times m \).

Pro \( i \)-tý objekt má \( y_i \) hodnotu buď 0, nebo 1 a \( x_i^T \) je \( i \)-tý řádek matice \( X \).

Označme

\[
p(x, b) = p_1(x, b) \quad \text{a} \quad 1 - p(x, b) = p_2(x, b)
\]

a za předpokladu binomického rozdělení \( y \) lze zapsat logaritmus věrohodnostní funkce ve tvaru

\[
\ln L(b) = \sum_{i=1}^{n} \{ y_i \ln p(x_i, b) + (1 - y_i) \ln (1 - p(x_i, b)) \}
\]

\[
= \sum_{i=1}^{n} \{ y_i b^T x_i - \ln (1 + \exp(b^T x_i)) \}
\]

kde \( b^T = \{ b_0, b_1 \} \) a předpokládá se, že první sloupec matice \( X \) obsahuje pouze jedničky (absolutní člen).
Metoda odhadu parametrů:

Pro maximalizaci \( \ln L(b) \) se využívá nulity prvních derivací

\[
J = \frac{d \ln L(b)}{db} = \sum_{i=1}^{n} x_i (y_i - p(x_i, b)) = 0
\]

Jde o soustavu \( m + 1 \) nelineárních rovnic vzhledem k \( b \). Řešení soustavy nelineárních rovnic využívá Newtonův-Raphsonovův algoritmus, který vyžaduje matici druhých derivací (hessiánu)

\[
H = \frac{d^2 L(b)}{db \, db^T} = -\sum_{i=1}^{n} x_i x_i^T p(x_i, b) (1 - p(x_i, b))
\]

Newtonova-Raphsonova metoda je iterativní, takže výsledkem \( j \)-té iterace je zpřesněný odhad

\[
b_{(j+1)} = b_{(j+1)} - H_{(j)}^{-1} J_{j}
\]

kde pro vektor pravděpodobnosti \( p \) rozměru \( n \times 1 \) s prvky \( p(x_i, b_{(j)}) \), a diagonální matici vah \( W \) rozměru \( n \times m \) s prvky lze psát

\[
J_{(j)} = X^T (y - p) \quad \text{a} \quad H = -X^T W X
\]
Interpretace regresních koeficientů

Žádné předpoklady o $x$ neexistují a $x$ mohou být jak diskrétní (faktory) tak i spojité veličiny (prediktory).

Předpoklad říká, že logit $\ln(L_{(1)}/L_{(0)})$ je lineární funkci nezávisle proměnných.

Pro $\ln(L_{(1)}/L_{(0)})$ se užívá termín logit nebo-li logit transformace pravděpodobnosti.

Logistický model se nazývá vícenásobný logistický regresní model (krátké logit) a koeficienty $b_i$ jsou interpretovány jako regresní koeficienty.

Logit lze ale také upravit: dosazením za $L_{(1)} - (1 - L_{(0)})$ dostaneme

$$L_{(0)} = \frac{1}{1 + \exp[-(b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p)]]}$$

Obecně:

Kladné znaménko koeficientu $b_i$ zvyšuje pravděpodobnost $L_{(0)}$ a záporné znaménko tuto pravděpodobnost snižuje.
Diskuse koeficientu $b_i$:

1) Je-li $b_i$ kladné, funkce $\exp$ je větší než 1 a pravděpodobnostní poměr $(L_{(1)}/L_{(0)})$ se bude zvyšovat.

   Zvýšení se objeví, když predikovaná pravděpodobnost odehrané události $L_{(1)}$ se zvýší a predikovaná pravděpodobnost neodehrané události $L_{(0)}$ se sníží.
   Proto má model vyšší predikovanou pravděpodobnost odehrané události $L_{(1)}$.

2) Je-li $b_i$ záporné, je funkce $\exp$ menší než 1 a pravděpodobnost se sníží.

3) Pro koeficient rovný nule vede funkce $\exp$ k hodnotě 1 čili k žádné změně pravděpodobnosti.
Test významnosti regresních koeeficientů

Logistická regrese umožňuje testovat významnost koeeficientů čili ověřit, že regresní koeeficient se liší od nuly.

Nula zde značí, že pravděpodobnostní poměr \( L_{(1)}/L_{(0)} \) se nemění a pravděpodobnost tím pádem není ovlivněna.

Studentův \( t \)-test k vyšetření statistické významnosti jednotlivých regresních koeeficientů.

Pro velké výběry lze užít Waldovo testační kritérium \( W_{a,i} = (b_i/s(b_i))^2 \), které vyčísluje statistickou významnost pro odhady regresních koeeficientů stejně jako ve vícenásobné regresi.
Waldova statistika $W_{a,i}$ má $\chi^2$-rozdělení s 1 stupněm volnosti a představuje čtverec poměru odhadu regresního koeficientu a jeho směrodatné odchylky $W_{a,i} = (b_i/s(b_i))^2$.

Pro kategorické proměnné má $W_{a,i}$ počet stupňů volnosti roven o 1 méně než je počet kategorií.

Waldova statistika $W_a$ má ale jednu nežádoucí vlastnost. Když je absolutní hodnota regresního koeficientu $b_i$ veliká a odhad je ho směrodatné odchylky $s(b_i)$ je také veliký, je výsledkem příliš malá hodnota testačního kritéria $W_{a,i}$, která vede k selhání zamítnutí nulové hypotézy, že regresní koeficient je nulový. Proto, je -li regresní koeficient veliký, neúžijeme Waldova kritéria.
Parciální korelace

Je obtížné určit příspěvek jednotlivých proměnných.

Příspěvek každé proměnné závisí také na ostatních proměnných v logistickém modelu.

K vyšetření parciální korelace mezi závisle proměnnou a každou nezávisle proměnnou se užívá korelační koeficient $R_i$, (v intervalu od -1 do +1).

1) **Kladné hodnoty** $R_i$: když roste hodnota $R_i$, zvyšuje se pravděpodobnost objektu “v události” $L_{(1)}$.

2) **Záporné hodnoty** $R_i$: naopak snižuje se pravděpodobnost objektů “v události” $L_{(1)}$.

3) **Malé hodnoty** $R_i$: proměnná má malý vliv na model.
Kategorické proměnné

Jednou z důležitějších výhod logistického modelu je možnost užívat i kategorické nezávislé proměnné $x$, zvané faktory.

Za faktor lze použít numerickou, textovou nebo datumovou hodnotu, zvanou úroveň nebo referenční hladina.

Interpretace odhadovaných regresních koeficientů je relativní vůči této hladině.

Nejjednodušší situací je jediný faktor $x$ se dvěma možnými hodnotami, například, deprese u 143 žen a 101 mužů je ovlivněna pohlavím, kde faktor pohlaví má dvě úrovni: pro muže je $x = 0$ a pro ženy je $x = 1$. 
Je-li objektem žena, pak pravděpodobnostní poměr, že je žena v depresi, je například 40/143. Podobně je tento poměr u mužů například 10/101. Pravděpodobnostní poměr, že jedinec je v depresi bude

\[
\frac{L_{(1)}}{L_{(0)}} = \frac{40/143}{10/101} = 2.825.
\]

A proto šance žen čili pravděpodobnostní poměr žen nacházet se v depresi je 2.825krát větší než šance mužů.

A podobně můžeme vyčíslit také šanci "nebýt v depresi":

\[
\frac{L_{(0)}}{L_{(1)}} = \frac{143/40}{101/10} = 0.354.
\]

Pojmu šance se hodně využívá v biomedikálních aplikacích. Je mírou spojení binární proměnné, jako je faktor risku výskytu dané události, například nemoci.
Kategorická proměnná čili faktor má dvě úrovně, tj. $x = 0$ značící muže a $x = 1$ značící ženy a logistickou rovnici pak bude

$$L_{(1)} = \frac{1}{1 + e^{-a - b \cdot x}}$$

odhad parametru $a = -2.313$ a odhad $b = 1.039$.

Odhad $b$ představuje přirozený logaritmus pravděpodobnostního poměru žen a mužů, $1.039 = \ln 2.825$, a proto pravděpodobnostní poměr $e^b = e^{1.039} = 2.825$.

Odhad $a$ je přirozený logaritmus pravděpodobnostního poměru mužů ($x = 0$) nebo $-2.313 = \ln 10/101$. Existuje-li pouze jedna dichotomní proměnná, není potřebné provádět logistickou regresní analýzu.

Přibližný interval spolehlivosti pro pravděpodobnostní poměr jako pro binární proměnnou se vypočte užitím odhadu směrnice $b$ a odhadu je jí směrodatné odchylky.

Například 95% interval spolehlivosti pro pravděpodobnostní poměr se vyčísli jako $\exp(b \pm 1.96 \cdot s(b))$. 
3. Volba proměnných

Například úloha logistické předpovědi infarktu:

Data jsou z dlouhodobého sledování z počátku zdravých pacientů, u kterých byla dlouhodobě provedena opakovaná měření. Několik jedinců bylo postiženo infarktem, několik ne.

Byl sledován výběr nezávisle proměnných, které by mohly odhalit blížící se infarkt.

Výběr účinných nezávisle proměnných byl předem lékaři vytypován.

Častěji však uživatel předem neví nic o nezávisle proměnných.

Proměnné $x$ jsou nejprve vyšetřovány, která je nejvíce spjata z dichotomní závisle proměnnou.
Studentův $t$-test významnosti jednotlivých parametrů: užívá se dostatečně vysoká hladina významnosti, například $\alpha = 0.15$, aby užitečná nezávisle proměnná nemohla být odstraněna.

Vyšetření zredukuje počet nezávisle proměnných na 10 či ještě méně.

Pak nastoupí kroková logistická regresní analýza: jde o test, zda proměnná $x_i$ zlepší predikční schopnost modelu. Postupy a jejich kritéria jsou užita k rozhodování, kolik proměnných $x_i$ a které je třeba užít.

Testy v dopředu krokové analýze jsou postaveny na $\chi^2$-statistice: velká hodnota $\chi^2$ nebo malá spočtená hladina významnosti $P$ ukazují, že nezávisle proměnná by měla být zařazena do proměnných.

Nalezená velká hodnota $\chi^2$ ukazuje, že proměnné jsou užitečné.
4. Těsnost proložení logistickým modelem

Před analýzou je třeba posoudit, zda nejsou odlehlé hodnoty. Rozptylové diagramy snadno odhalí odlehlé body.
Proměnné nemusí být normálně rozděleny.
Regresní diagnostika s analýzou vlivných bodů odhalí $O$ a $E$.

Logistická křivka má esovitý tvar a vystihuje logistický model, který je vzhledem ke koefficientům $b$ nelineární.

Po linearizační transformaci budou koeeficienty představovat směrnice u proměnných lineárního regresního modelu.
Mírou těsnosti proložení navrženého modelu daty je hodnota pravděpodobnosti \( L_{(1)} \), že se událost uskuteční. Místo veličiny \( L_{(1)} \) se používá tzv. odchylka, deviance \( D = -2 \ln L_{(1)} \) čili \( D = -2LL \).

\( D \) představuje míru těsnosti proložení dat logistickým regresním modelem:

1) Dobrý model vede k vysoké pravděpodobnosti objektů v události \( L_{(1)} \), což přeformulováno do veličiny \(-2 \ln L_{(1)}\) poskytuje malou hodnotu blízkou nule.

2) Minimální hodnotou pro \(-2 \ln L_{(1)}\) je nula, při které je dosaženo naprosto perfektní těsnosti proložení.

Rozdíl v odchylce je definován vztahem

\[ G = D(\text{model bez proměnné}) - D(\text{model s proměnnou}) \]

čili

\[ G = -2 \ln \frac{\text{pravděpodobnost modelu bez proměnné}}{\text{pravděpodobnost modelu s proměnnou}} \]

Veličina \( G \) proto odpovídá věrohodnostnímu poměru.
Těsnost proložení: spočívá porovnání experimentálních hodnot \( E \) s vypočtenými \( V \):

**Pearsonův test dobré shody \( \chi^2 \)** se užije, když model platí:
- Velká hodnota \( \chi^2 \) indikuje špatné proložení modelu.
- Malé hodnoty vypočtené hladiny významnosti \( P \) indikují špatné proložení modelu.

Nejužívanější způsoby posouzení těsnosti proložení:
Nejužívanější způsoby posouzení těsnosti proložení:

a) **Klasický Pearsonův přístup** začíná s identifikováním různých kombinací hodnot proměnných v regresním modelu, tj. vzorů.

Například dvě dichotomní proměnné, (pohlaví a zaměstnání) vedou na 4 kombinace: muž zaměstnán, muž nezaměstnán, žena zaměstnána, žena nezaměstnána.

- Pro každou kombinaci vyčíslíme počet \( E \) experimentálních hodnot jednotlivců (objektů) ve třídě I a II.

- Podobně pro každého jednotlivce vypočteme pravděpodobnost, že se nachází ve třídě I a ve třídě II logistickou regresní analýzou.

- Suma těchto pravděpodobností pro daný vzor se označí \( V \).

- Testační statistika testu dobré shody \( \chi^2 \) se vyčísli jako

\[
\chi^2_{exp} = \sum_{i=1} 2E \left( \ln \frac{E}{V} \right)
\]

kde suma se provede přes všechny odlišné vzory.

- Rezidua se sledují právě pro tyto odlišné vzory.
b) Hosmerův-Lemeshowův test dobré shody byl navržen v 1982. Pearsonův $\chi^2$-test dobré shody k redukci v logaritmech hodnoty pravděpodobnosti je mírou sledování zlepšení těsnosti zavedením jedné či více nezávisle proměnných.

Základní model, který je podobný výpočtu sumy čtverců při použití pouze průměrů, poskytuje nulovou linii k porovnání.

Vedle $\chi^2$-testu existuje několik $R^2$-podobných měr k posouzení těsnosti proložení, obdoba koeeficientu determinace ve vícnásobné regresi.

"Pseudo $R^2$" v logistické regresi pro logitový model se vypočte dle

$$R^2_{logit} = \frac{2 \ln L_{mul} - (-2 \ln L_{model})}{-2 \ln L_{mul}} = - \frac{D_{model} + D_{mul}}{D_{mul}}$$
c) **Metoda klasifikačních matic**, vyvinutých v diskriminační analýze slouží k vyhodnocení predikční schopnosti v pojmech zařazení do třídy.

Pravděpodobnost zařazení do třídy I je vypočtena pro každého jednotlivce (objekt) ve výběru a výsledný počet je uspořádán vzestupně. Pravděpodobnosti jsou pak rozděleny do 10 skupin (decily).

Pro každý naměřený počet jednotlivců ve třídě I je vyčíslen počet \( E \). Užitím logické regrese jsou pro jedince v každém decilu vypočteny počty \( V \). Pak se vyčísli Pearsonova \( \chi^2 \)-statistika testu dobré shody

\[
\chi^2_{\text{exp}} = \sum_{i=1}^{n} \frac{(E - V)^2}{V}
\]

kde sumace se provede přes obě třídy a 10 decilů.

Velká hodnota \( \chi^2 \) nebo malá hodnota \( P \) indikují, že proložení není dobré.
5. Kvalita vyhodnocení logistickou regresí

Třídíme objekty do tříd, musíme nalézt prahový bod pravděpodobnosti $P_c$: objekt je "v události", když pravděpodobnost události větší nebo rovna hodnotě $P_c$.

Graf prahové operační charakteristiky ROC k detekci signálu, když signál nebylo vždy možné správně přijmout.

Na ose $y$ je procento správně zařazených objektů "v události" nazvané pozitivní podíl (a v lékařském výzkumu nazývané citlivost).

Na ose $x$ je procento nesprávně zařazených objektů nazvané falešný podíl nebo v lékařském výzkumu "1 minus specificita" (v lékařském výzkumu nazývané senzitivita zařazených krys pro správně zařazené krysy a specificita krys pro falešně zařazené krysy).
1) Horní křivka v grafu ROC představu je výtečnou predikci: i pro malé hodnoty podílu nesprávně zařazených objektů se získá vysoké procento správně zařazených objektů, které skutečně jsou “v události”.

2) Střední křivka je skutečná křivka při uvažování malého počtu nezávisle proměnných, - třeba dvou. Vysoké procento (80 %) objektů správně zařazených v události je v poměru k 65 % chybně zařazených v události na nepřijatelné hladině.

3) Dolní hypotetická křivka, (přímka) odpovídá nahodilým výsledkům, například házení mincí. Blízkost střední křivky k dolní ukazuje , že je potřeba buď volit jinou, anebo přidat ještě další nezávisle proměnnou, abychom získali lepší model, i když je ale tento model statisticky významný na spočtené hladině $P = 0.009$. 
(a) Vybereme prahový bod na dolní části křivky grafu ROC a nechceme mít příliš mnoho objektů, zařazených jako "v události", bude se nazývat **přísný práh**. **Nevýhoda:** je ztráta mnoha objektů, které jsou "v události".

(b) Vybereme prahový bod na horní části křivky grafu ROC a chceme mít hodně objektů zařazených jako "v události", bude se nazývat **nedbalý práh**. **Nevýhoda:** sice velmi málo objektů "v události" bude ztraceno ale mnoho objektů "v neudálosti" bude chybně označeno jako "v události".

Křivky v grafu ROC **musí procházet body (0, 0) a (1, 1).**
Maximální plocha pod křivkou je \( \text{dna} \) čili 100%.
Numerická hodnota velikosti plochy bude blízká 1, když predikce modelu bude výtečná.
Když bude plocha blízká hodnotě 0.5, bude predikce modelu špatná.

Křivka ROC je proto užitečná při rozhodování, který ze dvou logistických modelů vybrat: lepší model dosáhne větší plochy pod křivkou ROC ale také větší výšky prahového bodu na křivce ROC.

Většina programů vybírá logistický model podle kritéria největší plochy pod křivkou ROC.
6. Aplikace logistické regrese

Modelu vícnásobné logistické regrese se často užívá k odhadu pravděpodobnosti jisté události, která se přihodí danému objektu.

K určení logistického regresního modelu je třeba výběru dat, ve kterém každý objekt, jedinec byl sledován v uvedeném časovém období a hodnoty závažných proměnných byly od začátku pečlivě zaznamenávány.
Výběr dat může být uskutečněn dvojím způsobem:


Na datech prvního podvýběru se vyčíslí logistický regresní model, který pak může být aplikován na člena druhého podvýběru.

Předpokládá se, že původní výběr je v ustáleném stavu, tzn. neobjevily se žádné podstatné změny, které by pozměnily vztah mezi nezávisle proměnnými a výskytem události.
2. Případ řídícího výběru: spočívá v získání dvou náhodných výběrů: první výběr, ve kterém se událost objeví, a druhý výběr, ve kterém se událost neobjeví.

Hodnoty predikovaných proměnných se musí získat retrospektivním způsobem, z minulých záznamů nebo ze vzpomínek.

Konstanta $a$ musí být nastavena tak, aby vyjadřovala pravý poměr objektu v události.
Existují důležité požadavky, které je třeba respektovat:

1. Model předpokládá, že logaritmus pravděpodobnostního poměru je lineárně závislý na nezávislých proměnných. Nesplnění by mělo být předem prověřeno buď užitím měr těsnosti proložení, nebo jinými způsoby. To může vyžadovat transformaci dat.

2. Výpočty jsou často časově náročné, a proto by měl uživatel rozumně redukovat počet proměnných.

3. Logistická regrese by se neměla užívat k vyhodnocení faktorů risku v dlouhodobých studiích, ve kterých jsou je dnotlivé studie rozličné délky.

4. Regresní koeficienty pro nezávisle proměnnou v logistickém regresním modelu závisí na ostatních proměnných, zařazených do logistického modelu. Koeficienty pro stejnou nezávisle proměnnou, když se použijí různé výběry proměnných, mohou být zcela odlišné.

5. Je-li užita seřazená analýza, kterákoliv proměnná pro seřání nemůže být použita jako nezávisle proměnná.

Příklad 8.5 **Významnost znaků intenzivní péče pacientů na JIP**
Sledováno 200 pacientů na jednotce intenzivní péče JIP. Cílem bylo postavit logistický regresní model k predikci pravděpodobnosti přežití pacienta a vyšetřit významnost rizikových faktorů úmrtnosti.

**Zdrojová matice:**

20 znaků pro 200 pacientů

*ID* je kód pacienta.

*STA* je kód přežití: 0 = přežije, 1 = nepřežije.

*AGE* vyjadřuje věk v letech.

*RACE* značí rasu: 1 = bílá, 2 = černá, 3 = ostatní.

*SER* značí obsluhu na JIP: 0 = běžná lékařská, 1 = operativní, chirurgická.

*CAN* udává, zda rakovina je akutní problém tohoto pacienta: 0 = není, 1 = ane je.

*CRN* značí existenci chronické ledvinové poruchy: 0 = není, 1 = ano je.

*INF* zda pacient dostal infekci až po vstupu na JIP: 0 = ne, 1 = ano.

*CPR* vyšetření předcházel přijetí na JIP: 0 = ne, 1 = ano.

*SYS* udává systolický krevní tlak při přijetí na JIP v mm Hg.

*HRA* je krevní pule při přijetí na JIP v pulsech/min.

*PRE* značí přežití v poslední půl roce: 0 = ne, 1 = ano.

*TYP* značí způsob přijetí na JIP: 0 = dle volby, 1 = v mimořádné nouzi.

*FRA* je důvod přijetí na JIP jako dlouhá kost, šíje, jednoduchá plocha, zlomenina kyčle: 0 = ne, 1 = ano.

*PO2* je uvolněný kyslík z krve: 0 pro *PO2* ≥ 60 a 1 pro *PO2* < 60.

*PH* udává pH krve: 0 pro *PH* ≥ 7.25 a 1 pro *PF* < 7.25.

*PCO* značí CO₂ v krvi: 0 pro *PCO* ≤ 45 a 1 pro *PCO* > 45.

*BIC* vyjadřuje diuhličitan v krvi: 0 pro *BIC* ≤ 18 a 1 pro *BIC* > 18.

*CRE* značí creatinin v krvi: 0 pro *CRE* ≤ 2.0 a 1 pro *CRE* > 2.0.

*LOC* značí stav vědomí při přijetí na JIP: 0 = žádné kóma, 1 = hluboké omámení, 2 = kóma.
řešení: Byly užity programy NCSS2000 [67], MINITAB [86] a STATISTICA [102].

Podmínky výpočtu:

Závislé proměnné: STA
Nezávislé proměnné: AGE|SEX|RACE|SER|CAN|CRN|INF|CPR |SYS|HRA |PRE|TYP|FRA|PO2|PH|PCO|BIC|CRE|LOC

Objektů (řádků): 200
Znaků (sloupců): 20

Volba proměnných a průběh výstavby modelu:
Po ukončeném iteračním přibližení.

<table>
<thead>
<tr>
<th>Třída</th>
<th>Řádky</th>
<th>Prior</th>
<th>Aktuálně vs. Predikce, $R^2$</th>
<th>Správně klasifikováno [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>160</td>
<td>0.80000</td>
<td>0.35837</td>
<td>98.125</td>
</tr>
<tr>
<td>1</td>
<td>40</td>
<td>0.20000</td>
<td>0.35837</td>
<td>40.000</td>
</tr>
<tr>
<td>Celkem</td>
<td>200</td>
<td></td>
<td></td>
<td>86.500</td>
</tr>
</tbody>
</table>
## Výstavba logistického regresního modelu

obsahuje postupné iterační zjišťování směrnic regresních parametrů až po dosažení termínačního kritéria pro $STA = 0$.

<table>
<thead>
<tr>
<th></th>
<th>0. iter.</th>
<th>1. iter.</th>
<th>2. iter.</th>
<th>3. iter.</th>
<th>4. iter.</th>
<th>5. iter.</th>
<th>6. iter.</th>
</tr>
</thead>
<tbody>
<tr>
<td>B0: Úsek</td>
<td>2.383</td>
<td>3.910</td>
<td>4.962</td>
<td>5.364</td>
<td>5.413</td>
<td>5.414</td>
<td></td>
</tr>
<tr>
<td>B1: AGE</td>
<td>-0.018</td>
<td>-0.035</td>
<td>-0.048</td>
<td>-0.052</td>
<td>-0.052</td>
<td>-0.052</td>
<td></td>
</tr>
<tr>
<td>B2: SEX</td>
<td>0.216</td>
<td>0.388</td>
<td>0.506</td>
<td>0.550</td>
<td>0.555</td>
<td>0.555</td>
<td></td>
</tr>
<tr>
<td>B3: RACE0</td>
<td>0.030</td>
<td>0.029</td>
<td>0.012</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>B4: SER</td>
<td>0.184</td>
<td>0.359</td>
<td>0.489</td>
<td>0.540</td>
<td>0.545</td>
<td>0.545</td>
<td></td>
</tr>
<tr>
<td>B5: CAN</td>
<td>-1.025</td>
<td>-1.884</td>
<td>-2.488</td>
<td>-2.728</td>
<td>-2.757</td>
<td>-2.758</td>
<td></td>
</tr>
<tr>
<td>B6: CRN</td>
<td>-0.021</td>
<td>0.027</td>
<td>0.076</td>
<td>0.098</td>
<td>0.102</td>
<td>0.102</td>
<td></td>
</tr>
<tr>
<td>B7: INF</td>
<td>-0.099</td>
<td>-0.087</td>
<td>0.004</td>
<td>0.051</td>
<td>0.056</td>
<td>0.056</td>
<td></td>
</tr>
<tr>
<td>B8: CPR</td>
<td>-0.262</td>
<td>-0.601</td>
<td>-0.864</td>
<td>-0.966</td>
<td>-0.978</td>
<td>-0.978</td>
<td></td>
</tr>
<tr>
<td>B9: SYS</td>
<td>0.004</td>
<td>0.007</td>
<td>0.010</td>
<td>0.011</td>
<td>0.012</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>B10: HRA0</td>
<td>0.002</td>
<td>0.003</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>B11: PRE0</td>
<td>-0.434</td>
<td>-0.724</td>
<td>-0.876</td>
<td>-0.924</td>
<td>-0.929</td>
<td>-0.929</td>
<td></td>
</tr>
<tr>
<td>B12: TYP0</td>
<td>-0.879</td>
<td>-1.733</td>
<td>-2.411</td>
<td>-2.705</td>
<td>-2.744</td>
<td>-2.744</td>
<td></td>
</tr>
<tr>
<td>B13: FRA0</td>
<td>-0.273</td>
<td>-0.669</td>
<td>-1.007</td>
<td>-1.137</td>
<td>-1.152</td>
<td>-1.152</td>
<td></td>
</tr>
<tr>
<td>B14: PO20</td>
<td>-0.191</td>
<td>-0.308</td>
<td>-0.361</td>
<td>-0.383</td>
<td>-0.387</td>
<td>-0.387</td>
<td></td>
</tr>
<tr>
<td>B16: PCO0</td>
<td>0.951</td>
<td>1.911</td>
<td>2.738</td>
<td>3.122</td>
<td>3.172</td>
<td>3.173</td>
<td></td>
</tr>
<tr>
<td>B17: BIC0</td>
<td>0.166</td>
<td>0.444</td>
<td>0.683</td>
<td>0.782</td>
<td>0.794</td>
<td>0.794</td>
<td></td>
</tr>
<tr>
<td>B18: CRE0</td>
<td>-0.212</td>
<td>-0.243</td>
<td>-0.237</td>
<td>-0.234</td>
<td>-0.233</td>
<td>-0.233</td>
<td></td>
</tr>
<tr>
<td>B19: LOC0</td>
<td>-1.386</td>
<td>-1.986</td>
<td>-2.460</td>
<td>-2.678</td>
<td>-2.706</td>
<td>-2.706</td>
<td></td>
</tr>
</tbody>
</table>

*Logit*  -138.63  -74.500  -66.632  -64.955  -64.802  -64.800  -64.800
Stanovení parametrů logistického modelu pro $STA = 0$ a pro $STA = 1$. Jsou vyznačeny statisticky významné parametry.

<table>
<thead>
<tr>
<th></th>
<th>$STA = 0$</th>
<th></th>
<th>$STA = 1$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
<td>Logit</td>
<td>$\chi^2$</td>
<td>$P$</td>
</tr>
<tr>
<td><strong>B0: Úsek</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B1: AGE</strong></td>
<td>1</td>
<td>-70.726</td>
<td>11.852</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>B2: SEX</strong></td>
<td>1</td>
<td>-65.432</td>
<td>1.264</td>
<td>0.261</td>
</tr>
<tr>
<td><strong>B3: RACE</strong></td>
<td>1</td>
<td>-64.800</td>
<td>0.000</td>
<td>0.993</td>
</tr>
<tr>
<td><strong>B4: SER</strong></td>
<td>1</td>
<td>-65.256</td>
<td>0.913</td>
<td>0.339</td>
</tr>
<tr>
<td><strong>B5: CAN</strong></td>
<td>1</td>
<td>-68.831</td>
<td>8.062</td>
<td>0.005</td>
</tr>
<tr>
<td><strong>B6: CRN</strong></td>
<td>1</td>
<td>-64.809</td>
<td>0.018</td>
<td>0.893</td>
</tr>
<tr>
<td><strong>B7: INF</strong></td>
<td>1</td>
<td>-64.805</td>
<td>0.011</td>
<td>0.916</td>
</tr>
<tr>
<td><strong>B8: CPR</strong></td>
<td>1</td>
<td>-65.274</td>
<td>0.949</td>
<td>0.330</td>
</tr>
<tr>
<td><strong>B9: SYS</strong></td>
<td>1</td>
<td>-65.955</td>
<td>2.310</td>
<td>0.129</td>
</tr>
<tr>
<td><strong>B10: HRA</strong></td>
<td>1</td>
<td>-64.877</td>
<td>0.154</td>
<td>0.695</td>
</tr>
<tr>
<td><strong>B11: PRE</strong></td>
<td>1</td>
<td>-65.856</td>
<td>2.113</td>
<td>0.146</td>
</tr>
<tr>
<td><strong>B12: TYP</strong></td>
<td>1</td>
<td>-69.984</td>
<td>10.369</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>B13: FRA</strong></td>
<td>1</td>
<td>-65.403</td>
<td>1.206</td>
<td>0.272</td>
</tr>
<tr>
<td><strong>B14: PO2</strong></td>
<td>1</td>
<td>-64.902</td>
<td>0.205</td>
<td>0.651</td>
</tr>
<tr>
<td><strong>B15: PH</strong></td>
<td>1</td>
<td>-66.915</td>
<td>4.230</td>
<td>0.040</td>
</tr>
<tr>
<td><strong>B16: PCO</strong></td>
<td>1</td>
<td>-68.421</td>
<td>7.242</td>
<td>0.007</td>
</tr>
<tr>
<td><strong>B17: BIC</strong></td>
<td>1</td>
<td>-65.194</td>
<td>0.789</td>
<td>0.375</td>
</tr>
<tr>
<td><strong>B18: CRE</strong></td>
<td>1</td>
<td>-64.823</td>
<td>0.047</td>
<td>0.829</td>
</tr>
<tr>
<td><strong>B19: LOC</strong></td>
<td>1</td>
<td>-76.005</td>
<td>22.411</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Paretův graf Studentova $t$-koeficientu pro STA
vyšetření významnosti parametrů logistického regresního modelu dat ICU
Odhad regresních parametrů a test jejich významnosti.

Půltočně jsou označeny statisticky významné parametry nalezené Waldovým testem.

<table>
<thead>
<tr>
<th>Parametr</th>
<th>Odhad ( b_i )</th>
<th>Směrodatná odchylka ( s )</th>
<th>Wald. test ( W )</th>
<th>( P )</th>
<th>Dolní mez ( C_D )</th>
<th>Horní mez ( C_H )</th>
</tr>
</thead>
<tbody>
<tr>
<td>B0: Úsek</td>
<td>5.414</td>
<td>2.122</td>
<td>6.511</td>
<td>0.011</td>
<td>1.255</td>
<td>9.572</td>
</tr>
<tr>
<td>B1: AGE</td>
<td>-0.052</td>
<td>0.017</td>
<td>9.203</td>
<td>0.002</td>
<td>-0.086</td>
<td>-0.019</td>
</tr>
<tr>
<td>B2: SEX</td>
<td>0.555</td>
<td>0.501</td>
<td>1.226</td>
<td>0.268</td>
<td>-0.427</td>
<td>1.537</td>
</tr>
<tr>
<td>B3: RACE</td>
<td>0.005</td>
<td>0.528</td>
<td>0.000</td>
<td>0.993</td>
<td>-1.031</td>
<td>1.040</td>
</tr>
<tr>
<td>B4: SER</td>
<td>0.545</td>
<td>0.575</td>
<td>0.898</td>
<td>0.343</td>
<td>-0.582</td>
<td>1.673</td>
</tr>
<tr>
<td>B5: CAN</td>
<td>-2.758</td>
<td>0.980</td>
<td>7.911</td>
<td>0.005</td>
<td>-4.679</td>
<td>-0.836</td>
</tr>
<tr>
<td>B6: CRN</td>
<td>0.102</td>
<td>0.762</td>
<td>0.018</td>
<td>0.894</td>
<td>-1.392</td>
<td>1.596</td>
</tr>
<tr>
<td>B7: INF</td>
<td>0.056</td>
<td>0.535</td>
<td>0.011</td>
<td>0.916</td>
<td>-0.992</td>
<td>1.104</td>
</tr>
<tr>
<td>B8: CPR</td>
<td>-0.978</td>
<td>0.984</td>
<td>0.989</td>
<td>0.320</td>
<td>-2.906</td>
<td>0.950</td>
</tr>
<tr>
<td>B9: SYS</td>
<td>0.012</td>
<td>0.008</td>
<td>2.161</td>
<td>0.142</td>
<td>-0.004</td>
<td>0.027</td>
</tr>
<tr>
<td>B10: HRA</td>
<td>0.004</td>
<td>0.009</td>
<td>0.152</td>
<td>0.697</td>
<td>-0.015</td>
<td>0.022</td>
</tr>
<tr>
<td>B11: PRE</td>
<td>0.929</td>
<td>0.629</td>
<td>2.184</td>
<td>0.139</td>
<td>2.162</td>
<td>0.303</td>
</tr>
<tr>
<td>B12: TYP</td>
<td>-2.744</td>
<td>0.995</td>
<td>7.603</td>
<td>0.006</td>
<td>-4.695</td>
<td>-0.794</td>
</tr>
<tr>
<td>B13: FRA</td>
<td>-1.152</td>
<td>0.999</td>
<td>1.329</td>
<td>0.249</td>
<td>-3.111</td>
<td>0.807</td>
</tr>
<tr>
<td>B14: PO2</td>
<td>-0.387</td>
<td>0.851</td>
<td>0.206</td>
<td>0.650</td>
<td>-2.055</td>
<td>1.281</td>
</tr>
<tr>
<td>B15: PH</td>
<td>-2.415</td>
<td>1.231</td>
<td>3.849</td>
<td>0.050</td>
<td>-4.828</td>
<td>-0.002</td>
</tr>
<tr>
<td>B16: PCO</td>
<td>3.173</td>
<td>1.386</td>
<td>5.241</td>
<td>0.022</td>
<td>0.456</td>
<td>5.889</td>
</tr>
<tr>
<td>B17: BIC</td>
<td>0.794</td>
<td>0.916</td>
<td>0.751</td>
<td>0.386</td>
<td>-1.002</td>
<td>2.590</td>
</tr>
<tr>
<td>B18: CRE</td>
<td>-0.233</td>
<td>1.075</td>
<td>0.047</td>
<td>0.828</td>
<td>-2.341</td>
<td>1.874</td>
</tr>
<tr>
<td>B19: LOC</td>
<td>-2.706</td>
<td>0.752</td>
<td>12.955</td>
<td>0.000</td>
<td>-4.180</td>
<td>-1.233</td>
</tr>
</tbody>
</table>
Nalezený logistický regresní model.

Nalezený model obsahuje pouze statisticky významné parametry s jejich odhadnutými směrnicemi pro $STA = 0$:

$$5.414 - 0.052 \text{AGE} + 0.555 \text{SEX} + 0.005 \text{RACE} + 0.545 \text{SER} - 2.758 \text{CAN} + 0.102 \text{CRN} + 0.056 \text{INF} - 0.978 \text{CPR} + 0.012 \text{SYS} + 0.004 \text{HRA} - 0.929 \text{PRE} - 2.744 \text{TYP} - 1.152 \text{FRA} - 0.387 \text{PO2} - 2.415 \text{PH} + 3.173 \text{PCO} + 0.794 \text{BIC} - 0.233 \text{CRE} - 2.706 \text{LOC}.$$
Kvalita modelu a těsnost proložení dat.

*Odchylka* (deviance) \( D = -2 \ln L = -2LL \) je rovna \( D = -64.800 \).

**Test rozdílu** mezi nalezeným logistickým regresním modelem a modelem úseku se všemi ostatními parametry \( b_i \) nulovými je založen na *rozdílu v odchylce* \( G = D (\text{model bez proměnné}) - D (\text{model s proměnnou}) = 70.561. \)

Při 19 stupních volnosti je \( P < 0.001 \), což dokazuje, že alespoň jeden z regresních parametrů je různý od nuly, protože \( P \) je nižší než zvolená hladina \( \alpha = 0.05 \).

**Testy dobré shody** se svými poměrně vyššími hodnotami spočtené hladiny významnosti \( P = 0.040 \) a 0.998 ukazují, že navržený model dobře prokládá data.

<table>
<thead>
<tr>
<th>Použitý test</th>
<th>( \chi^2 )</th>
<th>( df )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearsonův</td>
<td>214.487</td>
<td>180</td>
<td>0.040</td>
</tr>
<tr>
<td>Rozdílový</td>
<td>129.6</td>
<td>180</td>
<td>0.998</td>
</tr>
<tr>
<td>Hosmerův–Lemeshowův</td>
<td>5.726</td>
<td>8</td>
<td>0.678</td>
</tr>
<tr>
<td>Brownovy testy: obecný alternativní</td>
<td>1.407</td>
<td>2</td>
<td>0.495</td>
</tr>
<tr>
<td>Brownovy testy: symetrický alternativní</td>
<td>1.063</td>
<td>1</td>
<td>0.302</td>
</tr>
</tbody>
</table>
**Klasifikace objektů logistickým modelem.**

Tabulka klasifikovaných objektů na základě navrženého logistického regresního modelu.

<table>
<thead>
<tr>
<th>Vypočteno</th>
<th>0</th>
<th>1</th>
<th>Celkově</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dáno</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>157</td>
<td>3</td>
<td>160</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
<td>16</td>
<td>40</td>
</tr>
<tr>
<td>Celkem</td>
<td>81</td>
<td>19</td>
<td>200</td>
</tr>
</tbody>
</table>

Procento správně klasifikovaných: 86.5 %

**Kvalita modelu logistické regrese.**

<table>
<thead>
<tr>
<th>Test odchylky (deviance)</th>
<th>df</th>
<th>Kritérium</th>
<th>Kritérium/df</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>180</td>
<td>129.600</td>
<td>0.720</td>
</tr>
<tr>
<td>Pearsonův $\chi^2$-test</td>
<td>180</td>
<td>214.487</td>
<td>1.192</td>
</tr>
<tr>
<td>Logit</td>
<td>180</td>
<td>−64.800</td>
<td></td>
</tr>
</tbody>
</table>
### Tabulka ROC pro STA = 0

| $P_C$ | N(1|1) | N(1|0) | N(0|1) | N(0|0) | Senzitivita | Specificita | Podíl správně zařazených zařazených |
|-------|-------|-------|-------|-------|-------------|-------------|-----------------------------------|
| 0.050 | 160   | 34    | 0     | 6     | 1.000       | 0.150       | 1.150                             | 0.830                           |
| 0.100 | 160   | 33    | 0     | 7     | 1.000       | 0.175       | 1.175                             | 0.835                           |
| 0.150 | 159   | 30    | 1     | 10    | 0.993       | 0.250       | 1.243                             | 0.845                           |
| 0.200 | 159   | 30    | 1     | 10    | 0.993       | 0.250       | 1.243                             | 0.845                           |
| 0.250 | 158   | 29    | 2     | 11    | 0.987       | 0.275       | 1.262                             | 0.845                           |
| 0.300 | 158   | 27    | 2     | 13    | 0.987       | 0.325       | 1.312                             | 0.855                           |
| 0.350 | 158   | 26    | 2     | 14    | 0.987       | 0.350       | 1.337                             | 0.860                           |
| 0.400 | 158   | 25    | 2     | 15    | 0.987       | 0.375       | 1.362                             | 0.865                           |
| 0.450 | 158   | 25    | 2     | 15    | 0.987       | 0.375       | 1.362                             | 0.865                           |
| 0.500 | 157   | 24    | 3     | 16    | 0.981       | 0.400       | 1.381                             | 0.865                           |
| 0.550 | 155   | 23    | 5     | 17    | 0.968       | 0.425       | 1.393                             | 0.860                           |
| 0.600 | 153   | 20    | 7     | 20    | 0.956       | 0.500       | 1.456                             | 0.865                           |
| 0.650 | 146   | 18    | 14    | 22    | 0.912       | 0.550       | 1.462                             | 0.840                           |
| 0.700 | 142   | 15    | 18    | 25    | 0.887       | 0.625       | 1.512                             | 0.835                           |
| 0.750 | 135   | 13    | 25    | 27    | 0.843       | 0.675       | 1.518                             | 0.810                           |
| 0.800 | 124   | 8     | 36    | 32    | 0.775       | 0.800       | 1.575                             | 0.780                           |
| 0.850 | 105   | 6     | 55    | 34    | 0.656       | 0.850       | 1.506                             | 0.695                           |
| 0.900 | 88    | 3     | 72    | 37    | 0.550       | 0.925       | 1.475                             | 0.625                           |
| 0.950 | 70    | 2     | 90    | 38    | 0.437       | 0.950       | 1.387                             | 0.540                           |

**Plocha pod křivkou ROC:** 0.846
| $P_C$ | N(1|1) | N(1|0) | N(0|1) | N(0|0) | Senzitivita | Specificita | +Specificita | Podíl správně zařazených |
|------|------|------|------|------|---------|---------|-----------|-----------------|
| výsek | A    | B    | C    | D    | A/(A+C) | D/(B+D) | Senzitivita | Specificita |
| 0.050 | 38   | 90   | 2    | 70   | 0.950   | 0.437   | 1.387     | 0.540         |
| 0.100 | 37   | 72   | 3    | 88   | 0.925   | 0.550   | 1.475     | 0.625         |
| 0.150 | 34   | 55   | 6    | 105  | 0.850   | 0.656   | 1.506     | 0.695         |
| 0.200 | 32   | 36   | 8    | 124  | 0.800   | 0.775   | 1.575     | 0.780         |
| 0.250 | 27   | 25   | 13   | 135  | 0.675   | 0.843   | 1.518     | 0.810         |
| 0.300 | 25   | 18   | 15   | 142  | 0.625   | 0.887   | 1.512     | 0.835         |
| 0.350 | 22   | 14   | 18   | 146  | 0.550   | 0.912   | 1.462     | 0.840         |
| 0.400 | 20   | 7    | 20   | 153  | 0.500   | 0.956   | 1.456     | 0.865         |
| 0.450 | 17   | 5    | 23   | 155  | 0.425   | 0.968   | 1.393     | 0.860         |
| 0.500 | 16   | 3    | 24   | 157  | 0.400   | 0.981   | 1.381     | 0.865         |
| 0.550 | 15   | 2    | 25   | 158  | 0.375   | 0.987   | 1.362     | 0.865         |
| 0.600 | 15   | 2    | 25   | 158  | 0.375   | 0.987   | 1.362     | 0.865         |
| 0.650 | 14   | 2    | 26   | 158  | 0.350   | 0.987   | 1.337     | 0.860         |
| 0.700 | 13   | 2    | 27   | 158  | 0.325   | 0.987   | 1.312     | 0.855         |
| 0.750 | 11   | 2    | 29   | 158  | 0.275   | 0.987   | 1.262     | 0.845         |
| 0.800 | 10   | 1    | 30   | 159  | 0.250   | 0.993   | 1.243     | 0.845         |
| 0.850 | 10   | 1    | 30   | 159  | 0.250   | 0.993   | 1.243     | 0.845         |
| 0.900 | 7    | 0    | 33   | 160  | 0.175   | 1.000   | 1.175     | 0.835         |
| 0.950 | 6    | 0    | 34   | 160  | 0.150   | 1.000   | 1.150     | 0.830         |

Plocha pod křivkou $ROC$: 0.857
Grafy kvality logistické regresní analýzy.

**Graf prahové operační charakteristiky ROC** vystihuje správnost diagnostického testu, zda logistickým modelem vypočtené Ano nebo Ne je správné.

**Graf podílu správně zařazených objektů v závislosti na $P_C$** je velice užitečný graf k určení nejlepší hodnoty prahového dělicího bodu $P_C$.

![Diagram 1](image1)

![Diagram 2](image2)

Obr. 8.8a Graf ROC pro $STA = 0$ a 1.

○ **Závěr:** V logistickém regresním modelu je z původních 18 znaků nezávisle proměnných statisticky významných 6 znaků (pro proměnnou $STA = 0$), a to $AGE, CAN, TYP, PH, PCO$ a $LOC$. 
Statistické zpracování experimentálních dat
Potřebná literatura ke studiu v roce 2006

2004, 970 stran, 690,- Kč
2005, 500 stran, 450,- Kč
2006, 980 stran, 600,- Kč
Děkuji za pozornost!

http://meloun.upce.cz