Programovanie v prostredí

VII. Regresia a všeobecné lineárne modely

Aleš Kozubík

Katedra Matematických Metód a Operačnej Analýzy

19.12.2019

Čo si ukážeme

- Miery asociácie dát (štatistickej závislosti)
- Test pre koreláciu
- Jednoduchá lineárna regresia

Miery asociácie (štatistickej závislosti)

Kovariancia je mierou lineárnej závislosti dvoch spojitých premenných. Je závislá od jednotiek merania.
Ak máme náhodný výber dvojíc $(x_1, y_1), \ldots, (x_n, y_n)$, definujeme

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

a kovarianciu S_{xy} ako

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Miery asociácie (štatistickej závislosti)

Pearsonov korelačný koeficient je mierou závislosti nezávislou od jednotiek merania. Definujeme

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2, \quad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \overline{y})^2$$

Potom

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}}.$$

Miery asociácie (štatistickej závislosti)

Spearmanov poradový korelačný koeficient je neparametrickou alternatívou k Pearsonovmu koeficientu, meria lineárnu aj nelineárnu závislosť.

Ak máme náhodný výber dvojíc $(x_1, y_1), \ldots, (x_n, y_n)$, a nech $R_1, \ldots, R_n$ sú poradia hodnôt $x_1, \ldots, x_n$ a $S_1, \ldots, S_n$ poradia hodnôt $y_1, \ldots, y_n$ v usporiadaných výberoch.

Spearmanov poradový korelačný koeficient sa potom definuje ako korelačný koeficient dvojíc $(R_1, S_1), \ldots, (R_n, S_n)$.

Výpočet kovariancie a korelácie

Kovariancia a korelácia dvoch premenných

Pre výpočet kovariancie používame funkciu `cov()` a pre výpočet korelácie funkciu `cor()`.

Nech náhodný výber z dvoch premenných je v podobe vektorov `x` a `y`. Potom použijeme `cov(x,y,)` resp. `cor(x,y)`.

```
> cov(trees$Height,trees$Volume)
[1] 62.66
> cor(trees$Height,trees$Volume)
[1] 0.5982497
```

Výpočet kovariancie a korelácie

Kovariančná a korelačná matica

Pomocou funkcií `cov()` a `cor()` môžeme vypočítať aj celú kovariančnú resp. korelačnú maticu pre celú množinu dát.

```
> cov(trees)
             Girth    Height     Volume
Girth    9.847914 10.38333   49.88812
Height  10.383333 40.60000   62.66000
Volume  49.888118 62.66000 270.20280
> cor(trees)
             Girth     Height     Volume
Girth   1.0000000 0.5192801 0.9671194
Height  0.5192801 1.0000000 0.5982497
Volume  0.9671194 0.5982497 1.0000000
```

Výpočet kovariancie a korelácie

Kovariančná a korelačná matica – poznámky

Ak počítame kovarianciu resp. koreláciu dvoch premenných, obidva vektory musia mať rovnakú dĺžku.

Ak sa v dátovej množine vyskytujú neznáme hodnoty, je potrebné ako argument `use` funkcie `cov()` resp. `cor()` nastaviť hodnotu `"pairwise"`. Inak by kovariančná matica obsahovala taktiež neurčené hodnoty.

```
> x<-c(1,2,3,4,5)
> y<-c(1,2,3)
> y[5]<-5
> cov(x,y,use="pairwise")
[1] 2.916667
```

Výpočet kovariancie a korelácie

Kovariančná a korelačná matica – poznámky

Všetky hodnoty, ktoré vstupujú do výpočtu musia byť numerické, inak dôjde ku chybe.

```
> cov(iris)
Error: is.numeric(x) || is.logical(x) is not TRUE
> cov(iris[-c(4,5)])
              Sepal.Length Sepal.Width Petal.Length
Sepal.Length    0.6856935  -0.0424340    1.2743154
Sepal.Width    -0.0424340   0.1899794   -0.3296564
Petal.Length    1.2743154  -0.3296564    3.1162779
```

Výpočet kovariancie a korelácie

Spearmanov koeficient poradovej korelácie

Pre výpočet Spearmanovho koeficientu poradovej korelácie môžeme opäť použiť funkciu `cor()`, je však potrebné nastaviť parameter `method` na hodnotu `"spearman"`

```
> cor(trees$Height,trees$Volume,method="spearman")
[1] 0.5787101
```

Test pre korelačný koeficient

Test významnosti pre koreláciu určuje, či je korelácia štatisticky významná. Ide teda o test nulovej hypotézy $H_0 : \rho = 0$.

Vykonáva sa pomocou funkcie `cor.test()`, ktorá má takéto argumenty:

- `data1`, `data2` sú množiny dát (vektory), ktorých závislosť chceme posudzovať,
- `method` implicitne Pearsonov korelačný koeficient, ale môžeme zadať `method=spearman`,
- `conf.level` pre hladinu spoľahlivosti, implicitne 0.95,
- `alt` pre určenie obojstrannej resp. jednostrannej alternatívy.

Test pre korelačný koeficient

Príklad

Využijeme zabudovanú databázu `trees` a otestujeme závislosť medzi obvodom kmeňa stromu a jeho objemom.

```
> cor.test(trees$Girth,trees$Volume)
     Pearsons product-moment correlation
data:  trees$Girth and trees$Volume
t = 20.478, df = 29, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9322519 0.9841887
sample estimates:
      cor
0.9671194
```

Všeobecný lineárny model

Všeobecný lineárny model má tvar

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \varepsilon,$$

kde:

- Y je vysvetlovaná premenná,
- X_i sú vysvetľujúve náhodné premenné,
- β_i sú koeficienty, ktoré je potrebné odhadnúť,
- ε je chybový člen, predpokladá sa, že má normované normálne rozdelenie.

Jednoduchá lineárna regresia

Hodnoty náhodnej premennej Y sa vysvetľujú pomocou jedinej vysvetľujúcej náhodnej premennej X.

Základný tvar sa teda zjednoduší na

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Odhady koeficientov β_0 a β_1 sa určujú pomocou metódy najmenších štvorcov, t.j.

$$\min_{\beta_0, \beta_1} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Určenie parametrov lineárneho modelu v R

Pre určenie odhadu parametrov lineárneho modelu používame funkciu `lm(resp~var1,dataset)`, ktorá má tieto argumenty:

- `resp` vysvetľovaná premenná, resp z anglického *response*,
- `var1` vysvetľujúca premenná,
- `dataset` množina vstupných dát.

Poznámka: Pre „čistú“ lineárnu závislosť $y = \beta_1 X$ použijeme funkciu v tvare `lm(resp~-1+var1,dataset)`.

Jednoduchý lineárny model - príklad

Na základe meraní zaznamenaných v databáze `trees` chceme nájsť lineárnu závislosť medzi objemom stromu a obvodom jeho kmeňa.

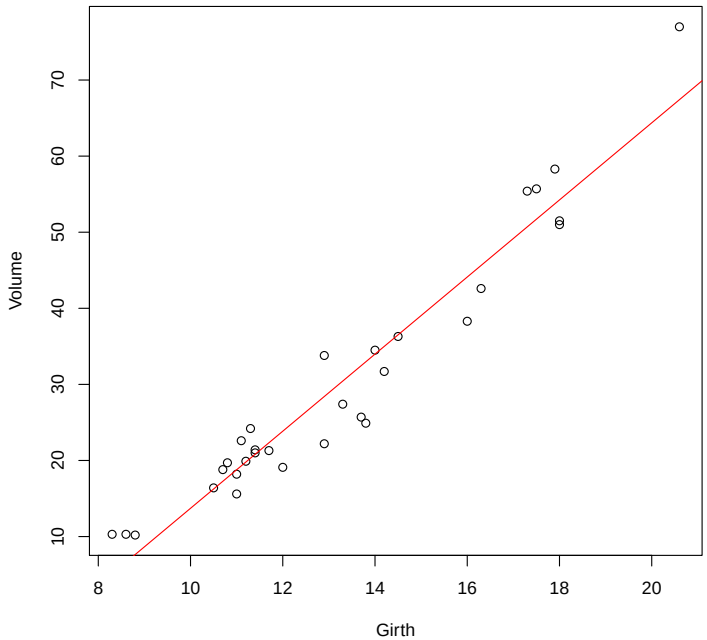
```
> lm(Volume~Girth,trees)

Call:
lm(formula = Volume ~ Girth, data = trees)
Coefficients:
(Intercept)        Girth
    -36.943        5.066
```

Je teda $V = -36.943 + 5.066 \cdot G$.

Zobrazenie regresnej priamky

Regresnú priamku doplníme do grafu nameraných hodnôt pomocou funkcie `abline()`




```
> plot(Volume~Girth,trees)
> abline(lm(Volume~Girth,trees)
+ ,col="red")
```

Sprístupnenie analýzy modelu

Výsledok si uložíme ako objekt, napr. `model<-lm()`. Sú tak dostupné najmä o ukazovatele presnosti vyrovnania modelom. (fit of the model)

- Koeficient determinácie R^2,ktorý charakterizuje, aká časť variability je vysvetlená modelom,
- Prispôsobený R^2, obdoba koeficientu determinácie súpravou vzhľadom na počet členov v modeli.
- Test významnosti pre koeficienty, testuje hypotézu, či koeficient je rôzny od nuly a prispieva tak ku vysvetľujúcej schopnosti modelu. pokiaľ je menší ako 0.05, odporúča sa daný faktor z modelu vyradiť.
- F-test určuje, či je model štatisticky významne lepším prediktorom než stredná hodnota.

Sprístupnenie analýzy modelu

Pre súhrnné štatistiky použijeme funkciu `summary()`. Dostaneme:

- kvantily pre chyby,
- odhady koeficientov so smerodajnou odchýlkou a testami významnosti,
- stredná kvadratická odchýlka reziduí $\sqrt{\frac{\sum e_i^2}{n-2}}$,
- koeficient determinácie R^2 (bežný aj prispôsobený)
- F-test.

Sprístupnenie analýzy modelu

Náš príklad so stromami

```
> model<-lm(Volume~Girth,trees)
> summary(model)

Call:
lm(formula = Volume ~ Girth, data = trees)
Residuals:
    Min      1Q  Median      3Q     Max
-8.065  -3.107   0.152   3.495   9.587
```

Významné kvantily reziduálnych chýb.

Sprístupnenie analýzy modelu

Náš príklad so stromami

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
```

Koeficient, pre každý z nich stredná kvadratická odchýlka, hodnota testovacej štatistiky a p-hodnota. Obidva sú s vysokou pravdepodobnosťou odlišné od nuly.

Sprístupnenie analýzy modelu

Náš príklad so stromami

```
Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353,    Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Koeficient determinácie $R^2 = 0.9353$, teda model vysvetľuje až 93.53% variability. F-test, p-hodnota potvrdzuje vysokú spoľahlivosť pre predikčné schopnosti modelu oproti odhadu pomocou strednej hodnoty.

Lineárny model, ďalšie funkcie

Funkcia `coef(model)` dáva ako odpoveď koeficienty modelu.

Funkcia `confint(model)` dáva ako odpoveď intervaly spoľahlivosti prekoeficienty modelu. Implicitná hladina spoľahlivosti je 0.95, pre jej zmenu treba zadať argument `level=hodnota`.

Lineárny model, ďalšie funkcie

Náš príklad so stromami

```
> coef(model)
(Intercept)       Girth
 -36.943459    5.065856
```

```
> confint(model,level=0.99)
                0.5 %     99.5 %
(Intercept) -46.21910 -27.667821
Girth         4.38399   5.747723
```

Polynomiálna regresia

Regresná funkcia je v tvare

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots \beta_n x^n + \varepsilon.$$

Použijeme opäť funkciu `lm()`, argumenty budú ale `resp`, `var`, `I(var^2)`, `I(var^3)` atď., a `dataset`.

Mocniny `I(var^2)`, `I(var^3)` musia byť zapuzdrené do `I()`, vzhľadom na špeciálny význam symbolov `^` a `*`.

Polynomiálna regresia

Náš prípad so stromami ako regresia pomocou kvadratickej funkcie

```
> lm(Volume~Girth+I(Girth^2),trees)
Call:
lm(formula = Volume ~ Girth + I(Girth^2), data = trees)
Coefficients:
(Intercept)        Girth   I(Girth^2)
    10.7863      -2.0921       0.2545
```

Model má teda tvar $V = 10.7863 - 2.0921G + 0.2545G^2$.

Polynomiálna regresia

Súhrnné údaje získame podobne ako pri lineárnej regresii:

```
> model<-lm(Volume~Girth+I(Girth^2),trees)
> summary(model)


Call:
lm(formula = Volume ~ Girth + I(Girth^2), data = trees)

Residuals:
     Min       1Q   Median       3Q      Max
-5.4889  -2.4293  -0.3718   2.0764   7.6447
```

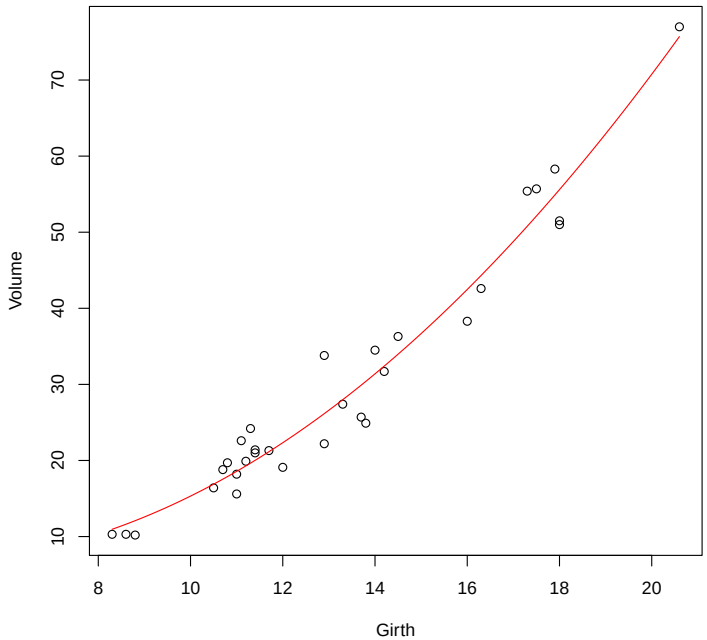
Polynomiálna regresia

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78627   11.22282   0.961 0.344728
Girth       -2.09214    1.64734  -1.270 0.214534
I(Girth^2)   0.25454    0.05817   4.376 0.000152

Residual standard error: 3.335 on 28 degrees of freedom
Multiple R-squared:  0.9616,    Adjusted R-squared:  0.9588
F-statistic: 350.5 on 2 and 28 DF,  p-value: < 2.2e-16
```

Zobrazenie regresnej krivky

Regresnú krivku doplníme do grafu nameraných hodnôt pomocou funkcie `curve()`




```
> plot(Volume~Girth,trees)
> curve(
10.78627-2.09214*x+0.25454*x^2,
col="red",add=T)
```

Multilineárny regresný model

Regresná funkcia má všeobecný tvar

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \varepsilon.$$

Použijeme opäť funkciu `lm()`, avšak s viacerými argumentmi `var1,var2,...` zodpovedajúcimi vysvetľujúcim premenným.

V rámci viacnásobnej lineárnej regresie môžeme brať do úvahy aj vzájomné interakcie jednotlivých faktorov. Tieto interakcie sa potom vyznačujú rôznymi oddeľovačmi premenných `var1, var2,...`.

Multilineárny regresný model - interakcie

Symboly používané vo vzťahoch premenných

- + oddeľuje vysvetľujpe premenné,
- : vyznačuje interakciu medzi dvomi premennými,
- * zkratka,pre všetky interakcie, `x*y*z` expanduje na `x+y+z+x:y+x:z+y:z+x:y:z`
- ^ interakcie do daného stupňa, `(x+y+z)^2` expanduje na `x+y+z+x:y+x:z+y:z`.

Multilineárny regresný model

Príklad

Využijeme databázu `mtcars`, hľadáme závislosť spotreby od počtu valcov, hmotnosti a výkonu vozidla.

```
> lm(mpg~cyl+wt+hp,data=mtcars)
Call:
lm(formula = mpg ~ cyl + wt + hp, data = mtcars)
Coefficients:
(Intercept)          cyl           wt           hp
   38.75179     -0.94162     -3.16697     -0.01804
```

Výsledný model

$$mpg = 38.75 - 0.94cyl - 3.17wt - 0.02hp.$$

Multilineárny regresný model

Príklad

Využijeme databázu `mtcars`, hľadáme závislosť spotreby od hmotnosti,výkonu a interakcie medzi nimi

```
> lm(mpg~wt+hp+wt*hp,data=mtcars)
Call:
lm(formula = mpg ~ wt + hp + wt * hp, data = mtcars)
Coefficients:
(Intercept)              wt              hp           wt:hp
   49.80842        -8.21662        -0.12010         0.02785
```

Výsledný model

$$mpg = 49.81 - 8.22wt - 0.12hp + 0.03 \cdot wt \cdot hp.$$

Jednoduché transformácie premenných

Priamo v lineárnych modeloch môžeme na vysvetľovanú premennú aplikovať transformačné funkcie.

Tak napríklad pomocou `lm(log(y)~var1,dataset)` získame regresný model pre hodnoty logaritmov premennej `y`.

Ak sa v definícii transformačnej funkcie objavujú symboly pre umocnennie `^`, je ich potrebné uzavrieť do `I()`. napríklad `lm(I(y^2)~var1,data=dataset)`.

Jednoduché transformácie premenných

Príklad

Využijeme zabudovanú databázu `women`,vyjadríme hmotnosť ako exponenciálnu funkciu výšky.

```
> lm(log(weight)~height,data=women)
Call:
lm(formula = log(weight) ~ height, data = women)
Coefficients:
(Intercept)       height
    3.27508      0.02518
```

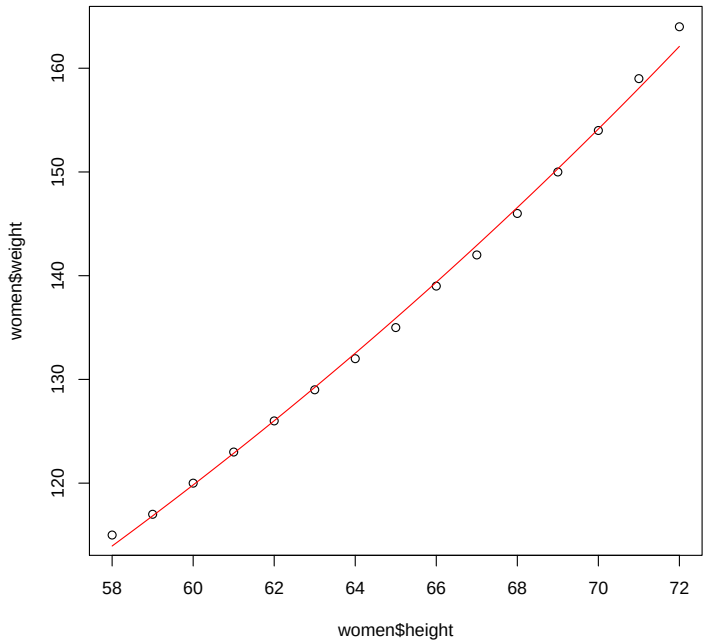
Výsledný model

$$w = \mathrm{e}^{3.27508+0.02518 \cdot h}.$$

Jednoduché transformácie premenných

Zobrazenie regresnej krivky

Regresnú krivku doplníme do grafu nameraných hodnôt pomocou funkcie `curve()`




```
> plot(women$height,women$weight)
> curve(exp(3.275+0.025184*x)
col="red",add=T)
```

Zahrnutie faktorových premenných

Do regresného modelu môžu byť zaradené aj premenné typu `factor`.

O tom, či je niektorá z premenných typu faktor sa presvedčíme pomocou funkcie `class()`

Pre ukážky použijeme súbor `people2.csv`, ktorý obsahuje (okrem iného) údaje o výške, rozpätí rúk a farbe očí.

Budeme skúmať závislosť výšky od rozpätia rúk a farby očí, ktorá je zrejme premennou typu faktor.

Zahrnutie faktorových premenných

Príklad

```
> people2<-read.csv("people2.csv")
> class(people2$Eye.Color)
[1] "factor"
> lm(Height~Hand.Span+Eye.Color, people2)
Call:
lm(formula=Height~Hand.Span+Eye.Color,data=people2)
Coefficients:
(Intercept)     Hand.Span  Eye.ColorBrown  Eye.ColorGreen
    82.8902        0.4456         -3.6233         -4.1924
```

Zahrnutie faktorových premenných

Čo znamená výsledok

```
Coefficients:
(Intercept)     Hand.Span   Eye.ColorBrown   Eye.ColorGreen
    82.8902        0.4456          -3.6233          -4.1924
```

Model predikcie výšky v podľa rozpätia HS je:

$$
\begin{aligned}
v &= 82.89 + 0.45 \cdot HS && \text{pre osoby s modrými očami,} \\
v &= 82.89 + 0.45 \cdot HS - 3.6233 && \text{pre osoby s hnedými očami,} \\
v &= 82.89 + 0.45 \cdot HS - 4.1924 && \text{pre osoby so zelenými očami.}
\end{aligned}
$$

Zahrnutie faktorových premenných

Zmena referenčnej hodnoty faktora

V predchádzajúciom príklade bola ako referenčná hodnota farby očí modrá.

Referenčnú hodnotu je možné zmeniť pomocou funkcie `relevel()`.

Syntax je:

```
>dataset$variable<-relevel(dataset$variable,"reflevel")
```

kde `reflevel` je nová hodnota

Zahrnutie faktorových premenných

Zmena referenčnej hodnoty faktora - ukážka

```
>people2$Eye.Color<-relevel(people2$Eye.Color,"Green")
>lm(Height~Hand.Span+Eye.Color,people2)
Call:
lm(formula=Height~Hand.Span+Eye.Color,data=people2)
Coefficients:
(Intercept)     Hand.Span    Eye.ColorBlue   Eye.ColorBrown
 78.6978          0.4456          4.1924          0.5690
```

Logistická regresia

Využíva sa na predikciu pravdepodobnosti výskytu určitej udalosti, resp. závislej náhodnej premennej, ktorá nadobúda dichotomické hodnoty 0 alebo 1.

Nech $X_1, \ldots, X_n$ je náhodný výber, α a $\beta_k, k = 1, \ldots, n$ parametre modelu.

Očakávanú hodnotu závislej premennej Y potom predpokladáme v tvare

$$\mathbb{E}(Y) = \frac{1}{1 + \mathrm{e}^{\left(-\alpha + \sum_{k=1}^{n} \beta_k x_k\right)}}.$$

Logistická regresia

Príklad

Na základe databázy `mtcars` zostavíme model logistickej regresie pre predpovedanie, či má vozidlo automatickú alebo manuálnu prevodovku, v závislosti od výkonu `hp` a hmotnosti `weight`.

Model má teda podobu

$$\mathbb{P}(mt = 1) = \frac{1}{1 + e^{(-\alpha + \beta_1 \cdot hp + \beta_2 \cdot w)}}.$$

Použijeme funkciu `glm()`, ktorá vytvára zovšeobecnený lineárny model. Použijeme argument `family=binomial`.

Logistická regresia

Príklad

```
> am.model<-glm(am~hp+wt,data=mtcars,family=binomial)
> am.model
Call:glm(formula=am~hp+wt,family=binomial,data=mtcars)
Coefficients:
(Intercept)           hp           wt
   18.86630      0.03626     -8.08348
Degrees of Freedom: 31 Total (i.e. Null);  29 Residual
Null Deviance:      43.23
Residual Deviance: 10.06        AIC: 16.06
```

Logistická regresia

Predikcia podľa modelu

```
> newdata<-data.frame(hp=120,wt=2.8)
> predict(am.model,newdata,type="response")
        1
0.6418125
```

Pravdepodobnosť, že dané vozidlo bude mať manuálnu prevodovku je 0.6418125.

Logistická regresia

Testy významnosti pre koeficienty

```
> summary(am.model)
Call:
glm(formula=am~hp+wt,family=binomial,data=mtcars)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2537  -0.1568  -0.0168   0.1543   1.3449
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 18.86630    7.44356   2.535  0.01126
hp           0.03626    0.01773   2.044  0.04091
wt          -8.08348    3.06868  -2.634  0.00843
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 10.059  on 29  degrees of freedom
AIC: 16.059
```

Logistická regresia

Význam premenných

Niektoré nové premenné:

- `deviance`, s prívlastkom `null` udáva odchýlku ak by sa použil iba konštantný člen, s prívlastkom `residual` je zostatková odchýlka celého modelu. V našom prípade sa odchýlka výrazne zredukovala z 43.23 na hodnotu 10.059,
- Fisherov iteračný proces je druhom Newtonovej metody pri hľadaní maxima vierohodnosti. Číslo udáva počet iterácií, ktoré boli potrebné na vyrovnanie.
- `AIC` je Akaikeho Informačné kritérium, jeho číselná hodnota sama o sebe nemá reálny význam, slúži na porovnanie alternatívnych modelov. treba vyberať modely s nižším AIC.