Logisztikus regresszió

2018. október 26.

Néhány példa

Mi a valószínűsége egy adott betegségnek a páciens bizonyos megfigyelt jellemzői (pl. nem, életkor, laboreredmények, BMI stb.) alapján?

Mely genetikai polimorfizmusok hajlamosítanak egy adott betegségre vagy védenek attól? *Mi a betegség kialakulásának esélye egy adott polimorfizmus hordozása esetén?*

Randomizált kontrollált kísérletben a kezelés hatásosságát szeretnénk mérni, potenciális zavaró változók hatásának kiszűrésével. *Meggyógyul-e a beteg, ha az adott kezelést kapja?*

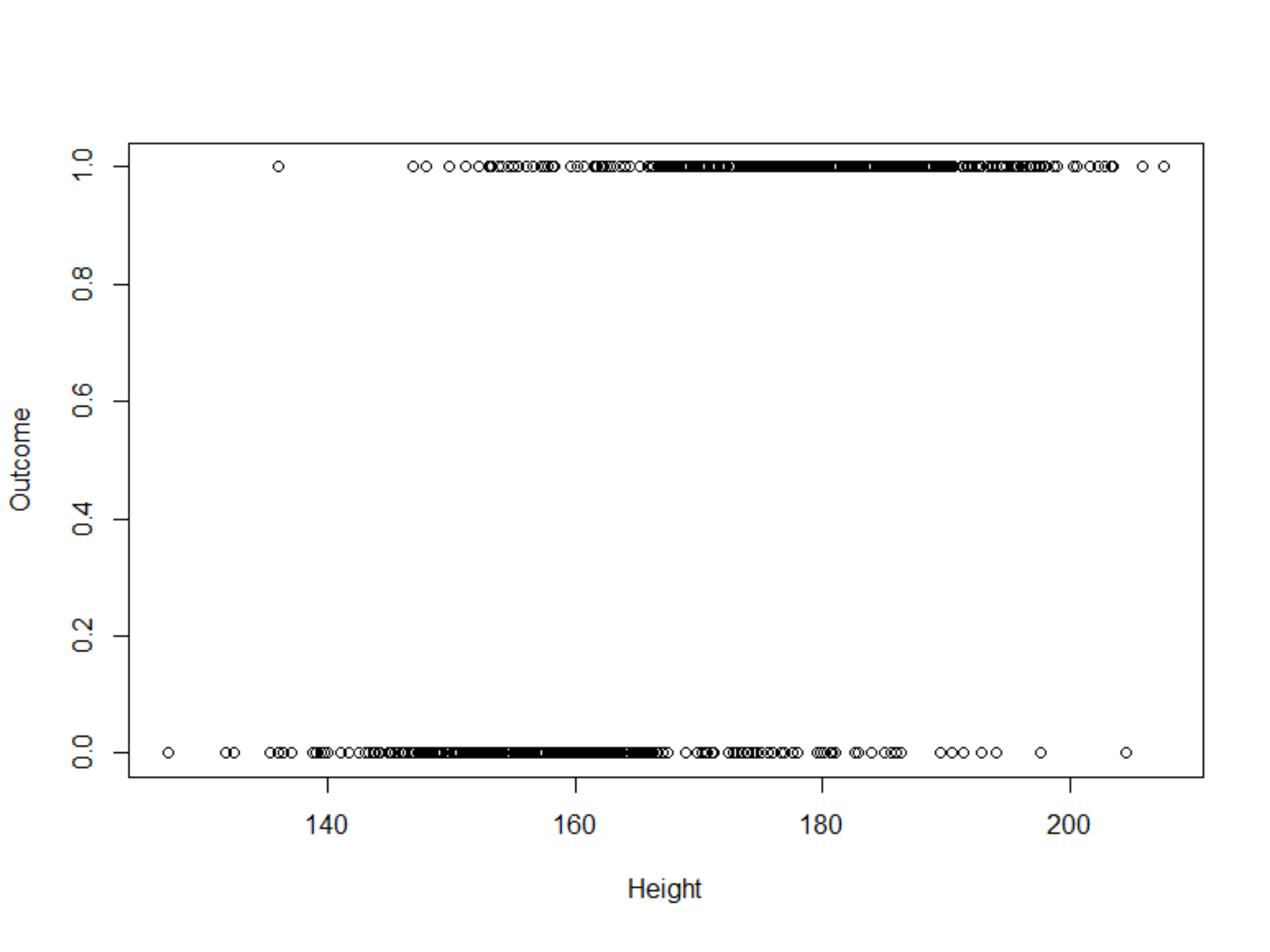
Logisztikus regressziós modell

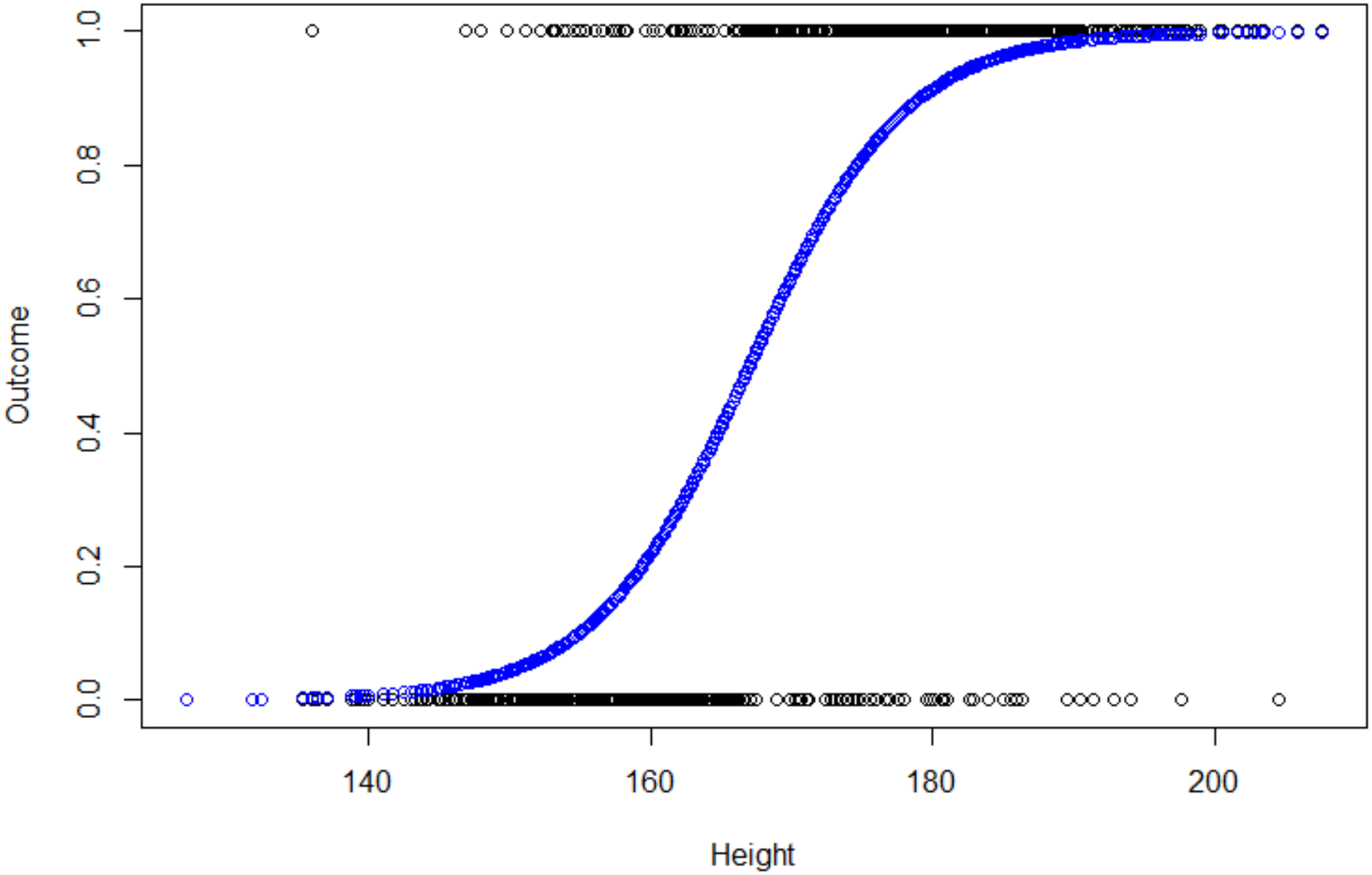
p = esemény valószínűsége

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

$$\frac{p}{1-p} = e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$


Outcome
0.0
0.2
0.4
0.6
0.8
1.0
140
160
180
200
Height


Outcome
0.0
0.2
0.4
0.6
0.8
1.0
140
160
180
200
Height

Logisztikus regressziós modell értelmezése

$$odds(X_1, \dots, X_n) = \frac{p}{1-p} = e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}$$

$$OR = \frac{odds(X_1 + 1)}{odds(X_1)} = \frac{e^{(\beta_0 + \beta_1 (X_1 + 1) + \cdots + \beta_n X_n)}}{e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}} =$$

$$= \frac{e^{(\beta_1 (X_1 + 1))}}{e^{(\beta_1 X_1)}} = e^{\beta_1}$$

Általánosított lineáris modellek

$$E(Y) = \mu = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)$$

g: link függvény

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential Gamma	real: $(0, +\infty)$	Exponential-response data, scale parameters	Inverse	$\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences			
Categorical	integer: $[0, K)$ K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1	outcome of single K-way occurrence			
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

https://en.wikipedia.org/wiki/Generalized_linear_model

Példa adat

```
> df <- read.table("logreg_df.txt", header=T, quote="", sep="\t", as.is=F)
```

```
> head(df)
  Age Gender Social Stress Height Weight Outcome.1 Outcome.2 Outcome.3 Outcome.4
1  65      1      3 Normal  170.0   64.3         1         0         1         0
2  56      0      2 Normal  168.2   60.8         1         0         0         1
3  67      1      1 Normal  190.9   83.4         1         0         1         1
4  40      1      3 Normal  165.9   76.7         0         1         0         0
5  11      1      1   High  175.0   72.1         0         0         1         1
6  70      0      1   High  184.0   71.6         1         0         0         0
> summary( df )
      Age            Gender          Social            Stress       Height          Weight         Outcome.1     
 Min.   :10.00   Min.   :0.000   Min.   :0.000   ExtremeLow : 23   Min.   :131.5   Min.   : 38.80   Min.   :0.000  
 1st Qu.:26.00   1st Qu.:0.000   1st Qu.:1.000   Low        :126   1st Qu.:160.1   1st Qu.: 64.90   1st Qu.:0.000  
 Median :39.00   Median :1.000   Median :2.000   Normal     :689   Median :170.2   Median : 72.90   Median :0.000  
 Mean   :40.15   Mean   :0.521   Mean   :2.017   High       :139   Mean   :170.3   Mean   : 72.77   Mean   :0.448  
 3rd Qu.:55.00   3rd Qu.:1.000   3rd Qu.:3.000   ExtremeHigh: 23   3rd Qu.:180.6   3rd Qu.: 80.10   3rd Qu.:1.000  
 Max.   :70.00   Max.   :1.000   Max.   :4.000                     Max.   :213.7   Max.   :101.10   Max.   :1.000  
   Outcome.2       Outcome.3       Outcome.4     
 Min.   :0.000   Min.   :0.000   Min.   :0.000  
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000  
 Median :0.000   Median :1.000   Median :1.000  
 Mean   :0.317   Mean   :0.509   Mean   :0.555  
 3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :1.000   Max.   :1.000   Max.   :1.000  
```

Logisztikus regresszió R-ben

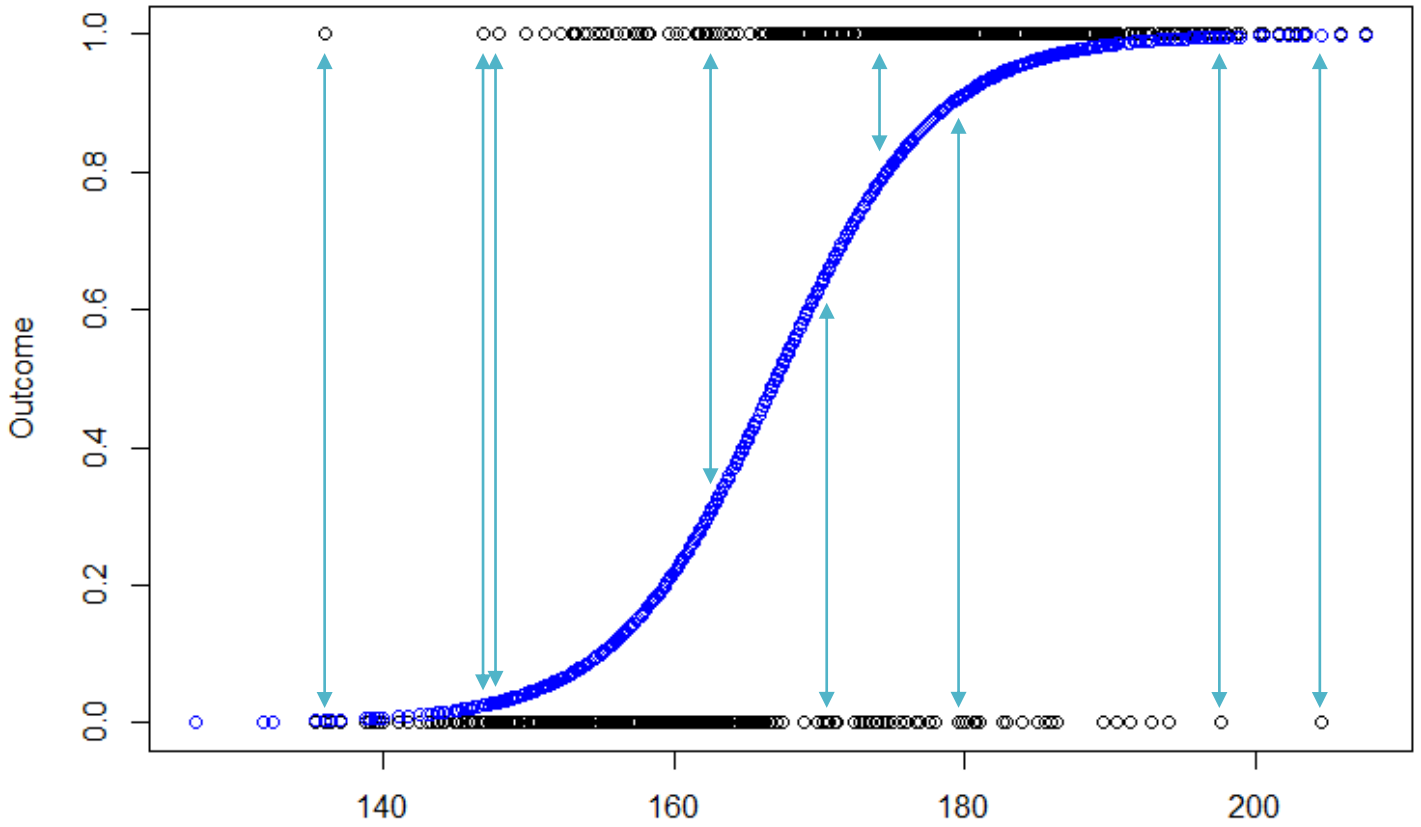
```
> fit <- glm( Outcome.1 ~ Age + Gender + Social + Stress + Height + Weight,
              data = df,
              family = "binomial" )

> summary( fit )
```

Eredmény

```
> fit <- glm( Outcome.1 ~ Age + Gender + Social + Stress + Height + Weight, data = df, family = "binomial" )
> summary( fit )

Call:
glm(formula = Outcome.1 ~ Age + Gender + Social + Stress + Height +
    Weight, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5018  -0.7286  -0.4101   0.6520   2.3577

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.794e+01  7.841e+02   0.023    0.982
Age                5.180e-02  5.275e-03   9.820   <2e-16 ***
Gender             1.987e-01  2.720e-01   0.731    0.465
Social             8.076e-03  6.738e-02   0.120    0.905
StressLow         -1.613e+01  7.841e+02  -0.021    0.984
StressNormal      -1.926e+01  7.841e+02  -0.025    0.980
StressHigh        -1.645e+01  7.841e+02  -0.021    0.983
StressExtremeHigh -2.502e-01  1.091e+03   0.000    1.000
Height            -8.213e-03  8.098e-03  -1.014    0.310
Weight            -9.250e-03  9.944e-03  -0.930    0.352
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1375.46  on 999  degrees of freedom
Residual deviance:  931.43  on 990  degrees of freedom
AIC: 951.43

Number of Fisher Scoring iterations: 16
```

Deviance residuals

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5018  -0.7286  -0.4101   0.6520   2.3577
```



$$d_i = s_i\sqrt{-2\{y_i log(p_i) + (1-y_i)log(1-p_i)\}}$$

$$s_i = \begin{Bmatrix} 1, ha\ y_i = 1 \\ -1, ha\ y_i = 0 \end{Bmatrix}$$

Koefficiensek

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.794e+01  7.841e+02   0.023    0.982
Age               5.180e-02  5.275e-03   9.820   <2e-16 ***
Gender            1.987e-01  2.720e-01   0.731    0.465
Social            8.076e-03  6.738e-02   0.120    0.905
StressLow        -1.613e+01  7.841e+02  -0.021    0.984
StressNormal     -1.926e+01  7.841e+02  -0.025    0.980
StressHigh       -1.645e+01  7.841e+02  -0.021    0.983
StressExtremeHigh -2.502e-01  1.091e+03   0.000    1.000
Height           -8.213e-03  8.098e-03  -1.014    0.310
Weight           -9.250e-03  9.944e-03  -0.930    0.352
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Estimate = β_i

Std. Error = A koefficiens standard hibája

Z value = Estimate / Std. Error

Pr(>|z|) = A kapott „Z value” mennyire extrém standard normális eloszlást feltételezve

A modell jósága

```
    Null deviance: 1375.46  on 999  degrees of freedom
Residual deviance:  931.43  on 990  degrees of freedom
AIC: 951.43

Number of Fisher Scoring iterations: 16
```

Null deviance = Mennyire magyarázza az adatainkat az üres modell

Residual deviance = Mennyire magyarázza az adatainkat a változó(ka)t tartalmazó modell

AIC = Akaike information criterion

Kategorikus változó referenciaértéke

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.794e+01  7.841e+02   0.023    0.982
Age                5.180e-02  5.275e-03   9.820   <2e-16 ***
Gender             1.987e-01  2.720e-01   0.731    0.465
Social             8.076e-03  6.738e-02   0.120    0.905
StressLow         -1.613e+01  7.841e+02  -0.021    0.984
StressNormal      -1.926e+01  7.841e+02  -0.025    0.980
StressHigh        -1.645e+01  7.841e+02  -0.021    0.983
StressExtremeHigh -2.502e-01  1.091e+03   0.000    1.000
Height            -8.213e-03  8.098e-03  -1.014    0.310
Weight            -9.250e-03  9.944e-03  -0.930    0.352
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary( df$Stress )
 ExtremeLow         Low      Normal        High ExtremeHigh
         23         126         689         139          23
> df$Stress2 <- relevel(df$Stress, ref = "Normal")
> summary( df$Stress2 )
     Normal  ExtremeLow         Low        High ExtremeHigh
        689          23         126         139          23
```

```
Call:
glm(formula = Outcome.1 ~ Age + Stress2 + Gender + Social + Height +
    Weight, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5018  -0.7286  -0.4101   0.6520   2.3577

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.320753   1.465846  -0.901    0.368
Age                 0.051801   0.005275   9.820   <2e-16 ***
Stress2ExtremeLow  19.259747 784.070503   0.025    0.980
Stress2Low          3.128637   0.285380  10.963   <2e-16 ***
Stress2High         2.812234   0.256385  10.969   <2e-16 ***
Stress2ExtremeHigh 19.009565 758.768772   0.025    0.980
Gender              0.198700   0.271994   0.731    0.465
Social              0.008076   0.067384   0.120    0.905
Height             -0.008213   0.008098  -1.014    0.310
Weight             -0.009250   0.009944  -0.930    0.352
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1375.46  on 999  degrees of freedom
Residual deviance:  931.43  on 990  degrees of freedom
AIC: 951.43
```

Interakció vizsgálata

```
> fit <- glm( Outcome.2 ~ Gender + Social, data = df, family = "binomial" )
> summary( fit )

Call:
glm(formula = Outcome.2 ~ Gender + Social, family = "binomial",
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6671  -0.7192  -0.4625   0.8970   2.3685

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.14295    0.21858 -14.379  < 2e-16 ***
Gender       0.40035    0.15414   2.597  0.00939 **
Social       0.96139    0.07479  12.854  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1249.2  on 999  degrees of freedom
Residual deviance: 1024.6  on 997  degrees of freedom
AIC: 1030.6
```

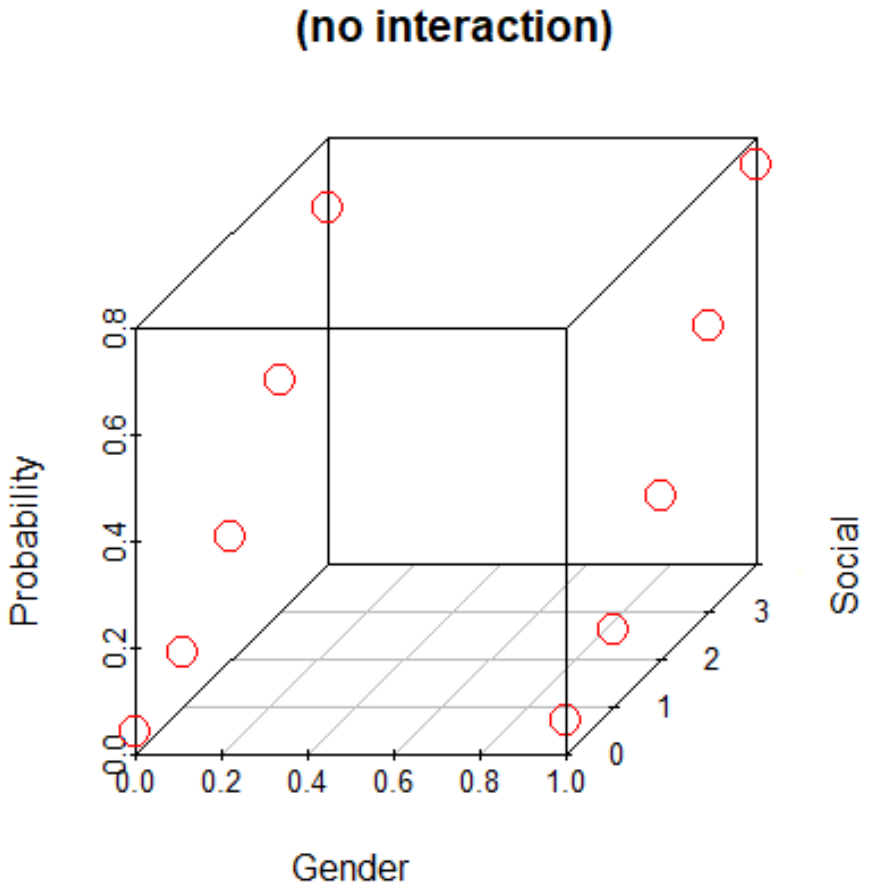
Interakció vizsgálata

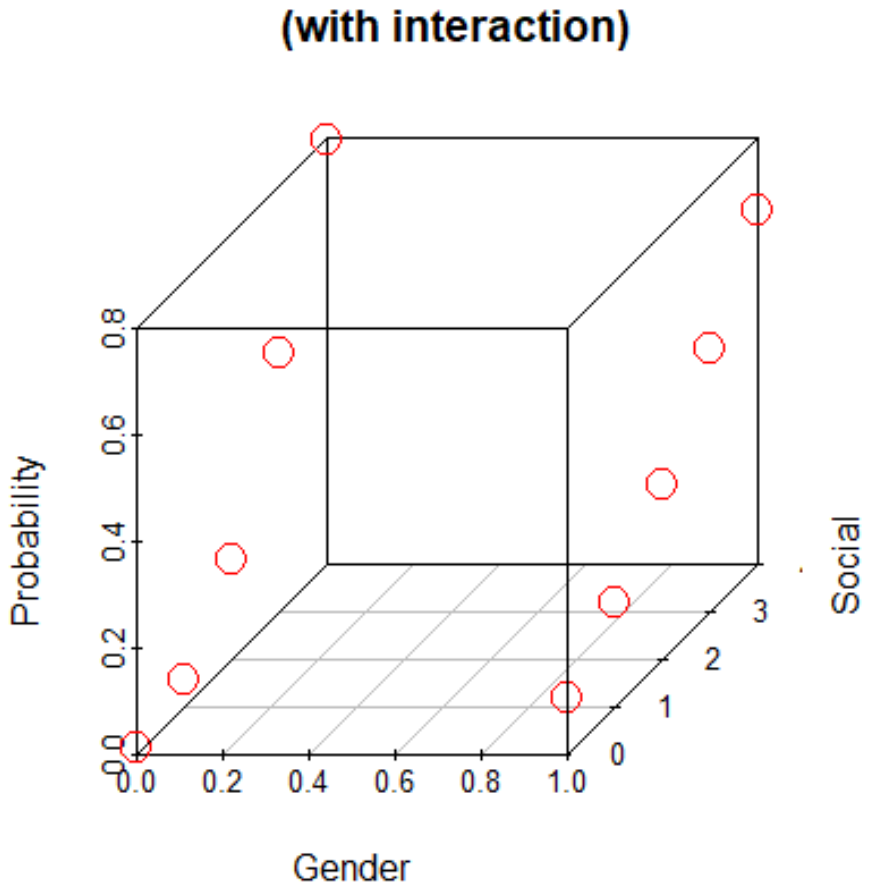
```
> fit <- glm( Outcome.2 ~ Gender * Social, data = df, family = "binomial" )
> summary( fit )

Call:
glm(formula = Outcome.2 ~ Gender * Social, family = "binomial",
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7838  -0.6549  -0.4735   0.9077   2.4311

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.3233     0.3915 -11.042  < 2e-16 ***
Gender         2.1913     0.4528   4.839 1.30e-06 ***
Social         1.4216     0.1416  10.040  < 2e-16 ***
Gender:Social -0.7202     0.1671  -4.310 1.63e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1249.2  on 999  degrees of freedom
Residual deviance: 1004.2  on 996  degrees of freedom
AIC: 1012.2
```

Interakció vizsgálata

```
fit.ni <- glm( Outcome.2 ~ Gender + Social, data = df, family = "binomial" )

summary( fit.ni )


fit.wi <- glm( Outcome.2 ~ Gender * Social, data = df, family = "binomial" )

summary( fit.wi )


library(scatterplot3d)

attach(df)

par(mfrow=c(1,2))

scatterplot3d(Gender, Social, predict(fit.ni,type = "response"), color="red",
cex.symbols = 2, zlab = "Probability", main="(no interaction)")

scatterplot3d(Gender, Social, predict(fit.wi,type = "response"), color="red",
cex.symbols = 2, zlab = "Probability", main="(with interaction)")

par(mfrow=c(1,1))
```

Interakció vizsgálata





Melyik modell a helyes?

AIC, Akaike information criterion

Statisztikai modellek relatív jóságának mérése, információelméleti alapon

- Mennyi információt veszítünk, ha az adott modellel reprezentáljuk azt a folyamatot, amely az adatot generálta?
- Relatív mérőszám, modellek összehasonlítása
- Paraméterek számával büntet => túlilleszkedés elkerülése

$$AIC = 2k - 2\ln(L)$$

k: paraméterek száma

L: Likelihood, az adat valószínűsége az adott modellt feltéve

AIC használata

AIC_1, AIC_2, AIC_3, …, AIC_R *R számú modell*

AIC_{min} : minimális AIC érték

Annak valószínűsége, hogy az *i*. modell esetén legkisebb az információveszteség (azaz ez a modell a legjobb):

$$P_{best=i} = \exp\left(\frac{1}{2}(AIC_{min} - AIC_i)\right)$$

Melyik modell a helyes? Példa.

```
fit.ni <- glm( Outcome.2 ~ Gender + Social, data = df, family = "binomial" )
fit.wi <- glm( Outcome.2 ~ Gender * Social, data = df, family = "binomial" )
```

$AIC_{(\text{no interaction})} = 1030.6$

$AIC_{(\text{with interaction})} = 1012.2$

$AIC_{min} = AIC_{(\text{with interaction})}$

$P(AIC_{(\text{no interaction})} \text{ a helyes modell}) =$

$= \exp((AIC_{(\text{with interaction})} - AIC_{(\text{no interaction})})/2)$

$\sim= 0.0001$

Egymással összefüggő prediktorok

```
> summary( glm( Outcome.3 ~ Height, data = df, family = "binomial" ) )

Call:
glm(formula = Outcome.3 ~ Height, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1430  -1.0481   0.5816   1.0367   2.0055

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.55381    0.90922  -10.51   <2e-16 ***
Height       0.05635    0.00533   10.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.0  on 999  degrees of freedom
Residual deviance: 1254.7  on 998  degrees of freedom
AIC: 1258.7
```

Egymással összefüggő prediktorok

```
> summary( glm( Outcome.3 ~ Weight, data = df, family = "binomial" ) )

Call:
glm(formula = Outcome.3 ~ Weight, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0823  -1.0457   0.5664   1.0315   2.2167

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.115947   0.500095  -10.23   <2e-16 ***
Weight       0.070845   0.006816   10.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.0  on 999  degrees of freedom
Residual deviance: 1257.9  on 998  degrees of freedom
AIC: 1261.9
```

Egymással összefüggő prediktorok

```
> summary( glm( Outcome.3 ~ Height + Weight, data = df, family = "binomial" ) )

Call:
glm(formula = Outcome.3 ~ Height + Weight, family = "binomial",
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1491  -0.9733   0.5460   0.9552   2.4369

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.122394   0.924542 -10.949  < 2e-16 ***
Height        0.039178   0.005846   6.701 2.06e-11 ***
Weight        0.047992   0.007433   6.456 1.07e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386  on 999  degrees of freedom
Residual deviance: 1210  on 997  degrees of freedom
AIC: 1216
```

Egymással összefüggő prediktorok

```
> summary( glm( Outcome.3 ~ Gender + Height + Weight, data = df, family = "binomial" ) )

Call:
glm(formula = Outcome.3 ~ Gender + Height + Weight, family = "binomial",
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7141  -0.7703   0.7362   0.7736   1.7481

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.860133   1.303476  -1.427    0.154
Gender       2.025314   0.241496   8.387   <2e-16 ***
Height       0.002790   0.007297   0.382    0.702
Weight       0.005059   0.009025   0.561    0.575
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.0  on 999  degrees of freedom
Residual deviance: 1132.9  on 996  degrees of freedom
AIC: 1140.9
```

Egymással összefüggő prediktorok

```
> summary( glm( Outcome.3 ~ Gender, data = df, family = "binomial" ) )

Call:
glm(formula = Outcome.3 ~ Gender, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6525  -0.7631   0.7678   0.7678   1.6589

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0847     0.1052  -10.32   <2e-16 ***
Gender        2.1554     0.1454   14.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.0  on 999  degrees of freedom
Residual deviance: 1133.4  on 998  degrees of freedom
AIC: 1137.4
```

Egymással összefüggő prediktorok

Melyik modell helyes?
Melyik változó hatása szignifikáns?

Modell (Outcome.3 ~)	AIC	P(jobb modell, mint a legjobb)	Szignifikáns változó
Height	1258.7	~ 0%	Height
Weight	1261.9	~ 0%	Weight
Height + Weight	1216	~ 0%	Height, Weight
Gender + Height + Weight	1140.9	~ 17.4%	Gender
Gender	**1137.4**	legjobb	Gender

Hasznos függvények

```
summary(fit)        # display results
confint(fit)        # 95% CI for the coefficients
exp(coef(fit))      # exponentiated coefficients
exp(confint(fit))   # 95% CI for exponentiated coefficients
predict(fit, type="response")       # predicted values
residuals(fit, type="deviance")     # residuals
```

Feladat:
Outcome.4 változó vizsgálata

Mely változók szignifikánsak?

Melyik a legjobb modell?

Van interakció?

Táblázat előállítása:

Koefficiensek (exp)

95% konfidencia intervallum (exp)

p-érték

Modell értelmezése?