The Difference Between Link Functions and Data Transformations

Log transformation

During the last classe, we learned:

- Log transformations of dependente variable are often recommended for skewed data.
- Log transformations might help us meet the linear regression assumptions.
- How do we interpret the coefficients, if we fit a linear model with a log-transformed dependent variable?

Geometric mean ratio

LN(Length of hospital stay) = 1.515619 + 0.012892 x PRISM

Every 1 unit difference in PRISM is associated with a $e^{0.012892}$ = 1,013 fold change in length of stay

OR

$(e^{0.012892} - 1) * 100 = 1.3$

For every one-unit increase in the PRISM, the length of hospital stay increases by about 1.3%.

Generalized Linear Models (GLM)

- **Generalized linear models** are called *generalized linear* because they connect an outcome to its predictors in a linear way.

- **Link function** is the function used to make this connection:

$$g(Y) = a + b_1\ x_1 + b_2\ x_2 + \ldots + b_k\ x_k$$

Y is the estimated value of predicted, mean or expected value of dependent value wich folows a known distribution (link function)

- The logistic regression is a GLM with *logit* link.

GLM – Logistic regression

Binary outcome

Logistic regression – **link function:** *logit*

The **glm()** command is designed to perform GLM on, for example, binary outcome data (logistic regression):

```
glm(data$outcome ~ data$predictor1, data$predictor2, data = data, family = "binomial")
```

$$logit(p) = \alpha + \beta x \qquad y \sim Binom(Trials, p)$$

GLM – Logistic regression

The data in the score2013.sav file refer to admissions to pediatric intensive care units in several Portuguese hospital units.

The minimum systolic tension of the first 12 hours of hospitalization and mechanical ventilation at some point in the first hour of hospitalization are potential predictors of mortality in the ICU. Study the individual association of each of them with mortality.

Binary outcome:
"vivo" or "Falecido"
(alive or dead)

ventil	TAS12	outcome
Não	87	Vivo
Não	102	Vivo
Sim	74	Vivo
Sim	97	Falecido
Não	78	Vivo
Sim	115	Vivo
Sim	88	Vivo
Sim	76	Vivo
Sim	107	Vivo
Sim	120	Vivo
Sim	99	Vivo
Sim	60	Vivo
Não	69	Vivo
Não	83	Vivo
Sim	92	Vivo
Não	126	Vivo
Sim	95	Vivo

GLM - Logistic regression

```
> logistic3 <- glm(data2$outcome ~ data2$ventil + data2$TAS12, data = data2, family = "binomial")
> summary(logistic3)

Call:
glm(formula = data2$outcome ~ data2$ventil + data2$TAS12, family = "binomial",
    data = data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1592  -0.3947  -0.2907  -0.2080   3.0196

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.526274   0.275370  -1.911  0.05598 .
data2$ventilSim  0.580226   0.204934   2.831  0.00464 **
data2$TAS12     -0.048460   0.004834 -10.024  < 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: 948.95  on 1767  degrees of freedom
Residual deviance: 816.67  on 1765  degrees of freedom
AIC: 822.67

Number of Fisher Scoring iterations: 6

> exp(cbind(OR = coef(logistic3), confint(logistic3)))
Waiting for profiling to be done...
                       OR     2.5 %    97.5 %
(Intercept)     0.5908025 0.3415328 1.0069762
data2$ventilSim 1.7864429 1.2024859 2.6908286
data2$TAS12     0.9526958 0.9435877 0.9616648
```

ventil	TAS12	outcome
Não	87	Vivo
Não	102	Vivo
Sim	74	Vivo
Sim	97	Falecido
Não	78	Vivo
Sim	115	Vivo
Sim	88	Vivo
Sim	76	Vivo
Sim	107	Vivo
Sim	120	Vivo
Sim	99	Vivo
Sim	60	Vivo
Não	69	Vivo
Não	83	Vivo
Sim	92	Vivo
Não	126	Vivo
Sim	95	Vivo

GLM – Poisson regression

Outcome: counts

Poisson regression – **link function:** *natural logarithm (ln)*

The glm() command is designed to perform GLM on, for example, count outcome data (poisson Regression):

glm(data$outcome ~ data$predictor1, data$predictor2, data = data, family = "poisson")

Assumption: rate of the event among individuals with the same explanatory variables is constant over the whole study period.

$$\ln(r) = a + b_1\ x_1 + b_2\ x_2 + \ldots + b_k\ x_k$$

GLM – Poisson regression

ASSUMPTIONS:

- The observations should be independente
- The rate should be constant
- Mean=variance (right skewed)

GLM – Poisson regression

The exponential of a particular coefficient is the estimated relative rate associated with the respective variable.

For x_1, e^{b1} is the estimate relative rate associated with x_1.

e^{b1} is the estimated rate of outcome for x_1+1 relative to the estimated rate of outcome for x_1, while adjusting for all other variables.

GLM – Poisson regression

- A researcher conduct a clinical trial of 50 patients with epilepsy, 25 of whom were randomized to receive the new anti-epilepsy drug and 25 of whom received na old drug.
- Patient's age and group (new or old drug) were registered.
- The outcome consisted of counts of seizures occuring during the follow-up period of eight weeks.

group	age	seizures
new drug	26	1
new drug	23	2
new drug	23	3
new drug	26	5
new drug	34	4
new drug	28	2
new drug	21	2
new drug	33	3
new drug	31	2
new drug	31	3
new drug	28	4
new drug	37	5
new drug	32	6
new drug	20	7
new drug	31	4
new drug	37	3
new drug	32	2
new drug	21	1
new drug	31	1
new drug	27	1
new drug	32	2

GLM – Poisson regression

```
poisson <- glm(datapoisson$seizures ~ datapoisson$age + datapoisson$group,
data = datapoisson, family = "poisson")

summary(poisson)

exp(cbind(RR = coef(poisson), confint(poisson)))
```

GLM – Poisson regression

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.347514   0.678376   0.512  0.60846
datapoisson$age           0.003112   0.021099   0.148  0.88273
datapoisson$groupnew drug 0.561495   0.204292   2.748  0.00599 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                              RR     2.5 %   97.5 %
(Intercept)               1.415545 0.3667683 5.258693
datapoisson$age           1.003117 0.9626480 1.045765
datapoisson$groupnew drug 1.753292 1.1809467 2.637163
```

1.75 is the estimated rate of Seizures for new drug grup relative to the estimated rate of seizures for old drug group, while adjusting for age (for patients of same age)

So what is the difference between link functions and data transformations with the logarithm function?

The difference between link functions and data transformations

- Poisson regression makes use of a natural log link function as follows:

$$\ln(\mu_y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- There is not a direct linear relationship of the x variables to the average count, but there is a relationship: a function (ln) of the mean of y is related to a linear combination of x variables.

- The key thing to understand is that the natural log link function is a function of the mean of y, not the y values themselves.

The difference between link functions and data transformations

- Below is a linear model equation where the original dependent variable, y, has been natural log transformed.

$$\ln(y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon_i$$

- The natural log has been taken of *each individual value of y*, and that is being used as the dependent variable.

- We could also write it as follows, where we are modeling the mean of ln(y):

$$\mu_{\ln(y)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

The difference between link functions and data transformations

$\ln(\mu_y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ poisson

$\mu_{\ln(y)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ transformation of y

When we <u>transform the data</u> in a linear model, we are no longer claiming that y is <u>normally distributed</u> around a mean, given the x values — we are claiming that the new outcome variable, $\ln(y_i)$, is normally distributed.

In the case of the Poisson model the *link* function does not change the distribution of the actual observations (Poisson distributed).

The difference between link functions and data transformations

If you used data transformation in a linear model, you cannot simply take the exponent of the mean of ln(y) to get the mean of y

but

you can do this with a link function.

If you have specific values of x variables, you can calculate the predicted average count, μ_y based on those x values b_y inversing the natural log

$$\mu_y = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}$$

this makes generalized linear models so useful.

GLM – Poisson regression

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.347514   0.678376   0.512  0.60846
datapoisson$age           0.003112   0.021099   0.148  0.88273
datapoisson$groupnew drug 0.561495   0.204292   2.748  0.00599 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                              RR     2.5 %   97.5 %
(Intercept)               1.415545 0.3667683 5.258693
datapoisson$age           1.003117 0.9626480 1.045765
datapoisson$groupnew drug 1.753292 1.1809467 2.637163
```

1.75 is the estimated rate of Seizures for new drug grup relative to the estimated rate of seizures for old drug group, while adjusting for age (for patients of same age)