Interpretation of linear, logistic and Poisson regression models with transformed variables and its implementation in the R package tlm

Jose Barrera-Gómez[a]

jbarrera@creal.cat

[a]Centre for Research in Environmental Epidemiology
http://www.creal.cat

Barcelona, October 24, 2013

Introduction

- Variables in a linear regression model are frequently transformed (e.g., homogeneity of variance, normality of errors, linearization, homogeneity of predictors).
- Researchers in health sciences are familiar with such transformations but less is known on how to interpret and report the effects in the original scale of the variables.
- The logarithmic transformation is especially important (e.g., adequacy of the lognormal distribution to describe ferritin, calcium, immunoglobulin, triglyceride or cotinine levels).

Aims

- Illustrate the interpretation of effects, in the original scale, under a linear model with transformed variables.
- Pay particular attention to the logarithmic transformation but also consider other transformations.
- Consider transformations of the explanatory variable in the logistic and Poisson regression models.
- Provide the R package tlm, which produces both numerical and graphical outputs.

Concepts

Linear model

Supose that we are interested in estimating the effect of an explanatory variable X on a response variable Y, based on the multiple linear regression model

$$\left.\begin{array}{l} \mathbb{E}(\tilde{Y}) = \beta\tilde{X} + K \\ K = \beta_0 + \beta_2 X_2 + \cdots + \beta_p X_p \\ \tilde{Y} \text{ and } \tilde{X} \text{ are transformations of } Y \text{ and } X \end{array}\right\} \quad (1)$$

Assumptions

1. Monotonic bijective transformations. Specifically,

$$\tilde{Y} = f_a(Y) \text{ and } \tilde{X} = f_b(Y), \text{ where}$$

$$f_p(U) = \begin{cases} \log(U) & \text{if } p = 0 \\ U^p & \text{if } p \neq 0 \end{cases}, \quad U > 0. \quad (2)$$

2. The modeled variable, $\tilde{Y}$ (or Y if the response variable is untransformed), is normally distributed conditional on the explanatory variables, and therefore, symmetric.

Transforming means

The generalized mean

If we calculate the (arithmetic) mean in the transformed space and then undo the transformation, we obtain the generalized mean:

$$f_p^{-1}(\overline{f_p(Y)}) = \left(\frac{\sum_i Y_i^p}{n}\right)^{1/p}.$$

Particular cases

Harmonic mean if $p = -1$
Geometric mean if $p = 0$ (log)
Arithmetic mean if $p = 1$ (no transformation)
Quadratic mean if $p = 2$

The median

Under the assumption of symmetry ($\leftarrow$ normality) and the family $f_p()$, the generalized mean is equal to the median.

Intepretation of linear models with transformed response

Under assumptions required for a linear model fitting and the family of transformations $f_p()$:

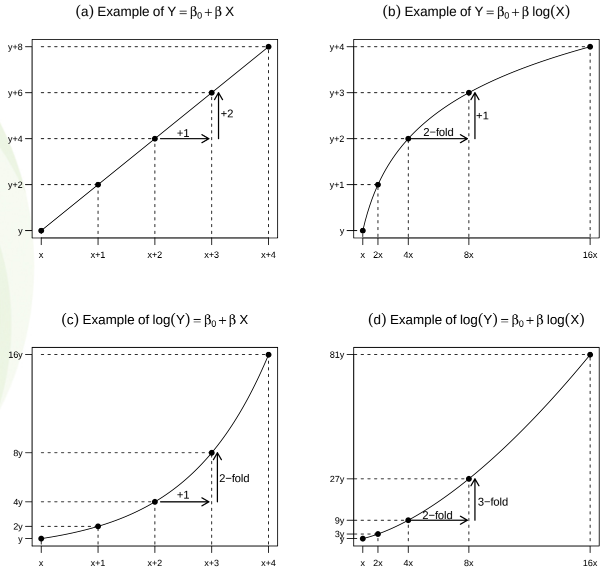
If no transformations

- Measure for the position of $Y|X$ and effect of X on Y: expected (adjusted) mean.
- Effect size does not depend on X (nor K): $\Delta X = 1 \Rightarrow \Delta\mathbb{E}(Y) = \beta$. Additive-additive relationship.

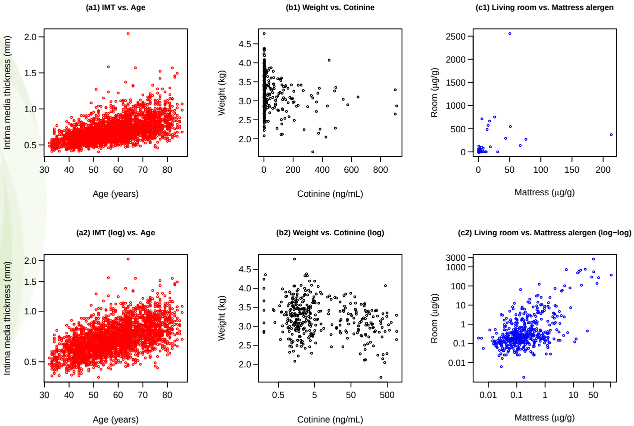
Under transformations $f_p()$

- Measure for the position of $Y|X$ and effect of X on Y: expected (adjusted) median or generalized mean (geometric, harmonic or quadratic in some cases).
- Effect size does depend on X (and K): it can not be summarized by (a function of) β.
  Exception: log transformation in Y and/or X...

Linear models with log transformations



Linear models with log transformations



(a): intima media thickness (IMT) and age. (b): birth weight and cord serum cotinine. (c): cat allergen levels in the home, measured in the living room and in the bed mattress.

Intepretation of linear models with log transformations

Interpretation and size of the (adjusted) effect of X on Y under linear models with log transformed variables. In the transformed model, β is the regression coefficient associated to X.

Model	Log	Effect size†	Effect interpretation
Linear	none	$\widehat{\Delta E} = \hat{\beta} c$	Additive change in the mean of Y when adding $c^{\S}$ units to X
	Y	$\widehat{\Delta M}_{\%} = 100 \left(e^{\hat{\beta} c} - 1 \right) \%$	Relative change in the median$^{\sharp}$ of Y when adding $c^{\S}$ units to X
	X	$\widehat{\Delta E} = \hat{\beta} \log(q)$	Additive change in the mean of Y when multiplying X by q
	X, Y	$\widehat{\Delta M}_{\%} = 100 \left(q^{\hat{\beta}} - 1 \right) \%$	Relative change in the median$^{\sharp}$ of Y when multiplying X by q
Logistic	none	$\widehat{\mathrm{OR}} = e^{\hat{\beta} c}$	Odds ratio for Y when adding $c^{\S}$ units to X
	X	$\widehat{\mathrm{OR}} = q^{\hat{\beta}}$	Odds ratio for Y when multiplying X by q
Poisson	none	$\widehat{\Delta E}_{\%} = 100(e^{\hat{\beta} c} - 1)\%$	Relative change in the mean of Y when adding $c^{\S}$ units to X
	X	$\widehat{\Delta E}_{\%} = 100(q^{\hat{\beta}} - 1)\%$	Relative change in the mean of Y when multiplying X by q

$\dagger$: $(1-\alpha)\%$ confidence interval is obtained when replacing $\hat{\beta}_i$ by $\hat{\beta}_i \pm z_{1-\alpha/2}\,\widehat{\mathrm{se}}(\hat{\beta}_i)$.

$\sharp$: Equivalently, geometric mean.

$\S$: If X is binary, $c = 1$.

Intepretation of linear models with log transformations

Approximate interpretation of the regression coefficient β under linear models with log transformed variables as the effect for a 1 unit or a 1% increase in the quantitative explanatory variable of interest, X. The last column indicates the error in the approximation.

Log	Interpretation†	Approximation error$^{\sharp}$
none	$\hat{\beta}$ units change in the mean§ of Y for unit increase in X	none
Y	$100\hat{\beta}$% change in the median§ of Y for unit increase in X	$< 10\%$ if $\lvert\hat{\beta}\rvert < 0.2$; $< 5\%$ if $\lvert\hat{\beta}\rvert < 0.1$
X	$\hat{\beta}/100$ units change in the mean of Y for 1% increase in X	0.5% for any $\hat{\beta}$
$X, Y^{\flat}$	$\hat{\beta}$% change in the median§ of Y for 1% increase in X	$< 10\%$ if $\lvert\hat{\beta}\rvert < 20$; $< 5\%$ if $\lvert\hat{\beta}\rvert < 10$

$\dagger$: $(1-\alpha)$% confidence interval is obtained when replacing $\hat{\beta}_i$ by $\hat{\beta}_i \pm z_{1-\alpha/2}\,\widehat{\text{se}}(\hat{\beta}_i)$.

$\sharp$: Percentage error relative to the true value of the effect.

$\S$: Equivalently, geometric mean.

$\flat$: Also valid in the Poisson regression model with log transformed X.

Intepretation of linear models with other transformations

Expected adjusted median of the response Y

$$\hat{M}(x) = f_a^{-1}(\hat{\beta} f_b(x) + \widehat{\bar{K}}),$$

where

$$\widehat{\bar{K}} = \hat{\beta}_0 + \hat{\beta}_2 \bar{X}_2 + \cdots + \hat{\beta}_p \bar{X}_p.$$

Expected adjusted effect of X ($X = u_1 \rightarrow X = u_2$) on the median of the response Y

- Additive change in the median of Y:

$$\widehat{\Delta M} = \hat{M}(u_2) - \hat{M}(u_1) = f_a^{-1}(\hat{\beta} f_b(u_2) + \widehat{\bar{K}}) - f_a^{-1}(\hat{\beta} f_b(u_1) + \widehat{\bar{K}}).$$

- Percent change in the median of Y:

$$\widehat{\Delta M}_{\%} = 100 \frac{\hat{M}(u_2) - \hat{M}(u_1)}{\hat{M}(u_1)} \% = 100 \left[\frac{f_a^{-1}(\hat{\beta} f_b(u_2) + \widehat{\bar{K}})}{f_a^{-1}(\hat{\beta} f_b(u_1) + \widehat{\bar{K}})} - 1 \right] \%,$$

For binary X, $u_1 = 0$ and $u_2 = 1$, and $b = 1$.

Software implementation

The R package tlm

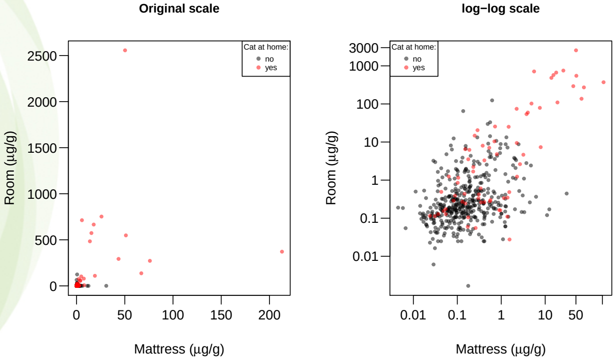
We are developing the R package tlm which allows to interpret and display adjusted effects both graphically and numerically.

Main functions

- `tlm`: fits the model in the transformed space.
  - Specific methods `print` and `summary` provide additional information on the transformations done.
  - Specific method `plot` (original space, transformed space and graphical diagnosis).
- `predict`: computes the expected adjusted median of Y (or the adjusted mean of $f_a(Y)$, in the transformed space) as a function of X. Confidence intervals are based on parametric bootstrap.
- `effectInfo`: provides information about how to interpret effects in the original scale.
- `effect`: computes the expected change in the adjusted median of Y associated to a given change in X.

Example: cat allergen levels in the home

Cat allergen levels measured in the living room (Y) and in the bed mattress (X):



Example: cat allergen levels in the home

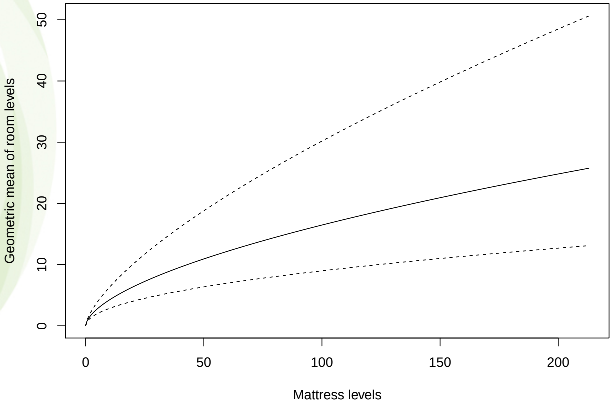
```
> head(cat)

     id       bed      room cat     logbed    logroom
1 30001 0.3781000  8.004959 Yes -0.9725966  2.0800612
2 30002 0.2546667  0.216320  No -1.3677996 -1.5309965
3 30004 0.2888511 20.437040 Yes -1.2418439  3.0173489
4 30005 0.0718000  0.384000  No -2.6338708 -0.9571127
5 30006 0.0916053  0.192640  No -2.3902661 -1.6469321
6 30007 0.0860870  0.103600  No -2.4523969 -2.2672179

> library(tlm)

> catmodel <- tlm(y = logroom, x = logbed, z = cat, ypow = 0, xpow = 0, data = cat)
> catmodel

Linear regression fitted model in the transformed space
-------------------------------------------------------


Transformations:
   In the response variable: log
   In the explanatory variable: log

Call:
lm(formula = logroom ~ logbed + cat, data = cat)

Coefficients:
(Intercept)       logbed       catYes
    -0.1282       0.5913       1.4296
```

Example: cat allergen levels in the home

```
> summary(catmodel)

Linear regression fitted model in the transformed space
-------------------------------------------------------

Transformations:
   In the response variable: log
   In the explanatory variable: log

Call:
lm(formula = logroom ~ logbed + cat, data = cat)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2453 -0.9784 -0.0585  0.7937  5.4805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1282     0.1183  -1.083    0.279
logbed        0.5913     0.0483  12.242  < 2e-16 ***
catYes        1.4296     0.2199   6.500 2.06e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.555 on 468 degrees of freedom
Multiple R-squared:  0.3853, Adjusted R-squared:  0.3827
F-statistic: 146.7 on 2 and 468 DF,  p-value: < 2.2e-16
```

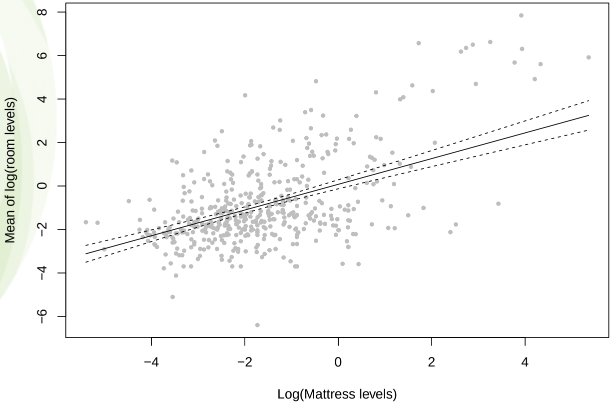
Example: cat allergen levels in the home

```
> plot(catmodel, xname = "Mattress levels", yname = "room levels")
```



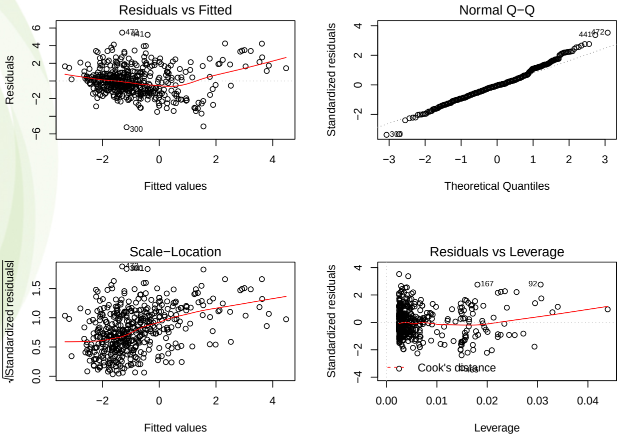
Example: cat allergen levels in the home

```
> plot(catmodel, xname = "Mattress levels", yname = "room levels", type = "transform",
+       observed = T)
```



Example: cat allergen levels in the home

```
> plot(catmodel, type = "diagnosis")
```



Example: cat allergen levels in the home

```
> predict(catmodel)  # Default: 10 points in arithmetic progression in the given space
Estimated adjusted geometric mean of the response variable (original space):

        logbed    Estimate    lower95%    upper95%
1    0.0045004  0.04428898  0.03007509  0.06522055
2   23.6706670  7.02358298  4.37193843 11.28348871
3   47.3368336 10.58127923  6.18328755 18.10743384
4   71.0030003 13.44793578  7.57129183 23.88588115
5   94.6691669 15.94152405  8.74040687 29.07555596
6  118.3353335 18.18996420  9.76974271 33.86729902
7  142.0015001 20.26055599 10.69984419 38.36412211
8  165.6676668 22.19406780 11.55478448 42.62966967
9  189.3338334 24.01748931 12.35024206 46.70676004
10 213.0000000 25.74982227 13.09710344 50.62595329
```

Several options...

```
> predict(catmodel, x = quantile(cat$room, probs = 0:4/4))
> predict(catmodel, npoints = 100, space = "transformed", level = 0.99)
```

Example: cat allergen levels in the home

```
> effectInfo(catmodel)

The effect of X on Y can be summarized with a single number as follows:
 - Change in X: multiplicative of factor q (equivalently, adding an r = 100 * (q - 1)% to X)
 - Type of effect on Y: percent change in the geometric mean of Y
 - Effect size: 100 * (q^beta - 1)%
   beta coefficient estimate:
        Estimate Std. Error  t value     Pr(>|t|)
logbed 0.5913161 0.04830168 12.24214 4.354439e-30

Further details can be obtained using effect(), providing either the multiplicative ('q') or
the percent ('r') change in X, and the level for the confidence interval, 'level'.

> effect(catmodel)

Percent change in the geometric mean of Y when changing X
from the 1st to the 3rd quartile: 192.8697

95% confidence interval: (146.4713, 248.0028)
```

Several options...

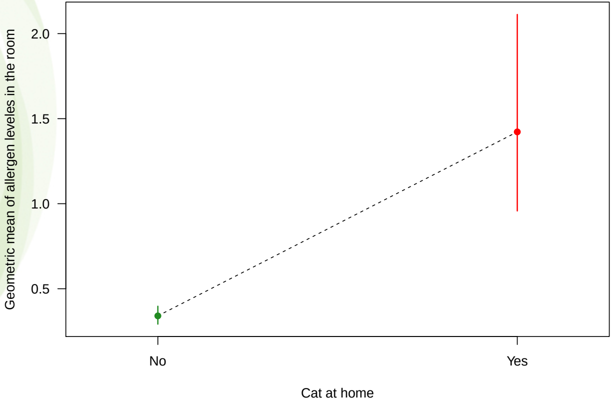
```
> effect(object, x1 = NULL, x2 = NULL, c = NULL, q = NULL, r = NULL, npoints = NULL,
+       level = 0.95, nboot = 5000, seed = 4321)
```

Example: cat allergen levels in the home

```
> catmodel2 <- tlm(y = logroom, x = cat, z = logbed, ypow = 0, data = cat)
> effectInfo(catmodel2)
The effect of X on Y can be summarized with a single number as follows:
 - Change in X: changing X from its reference, 'No', to the alternative level
 - Type of effect on Y: percent change in the geometric mean of Y
 - Effect size: 100 * [exp(beta) - 1]%
   beta coefficient estimate:
    Estimate   Std. Error      t value     Pr(>|t|)
1.429615e+00 2.199399e-01 6.500025e+00 2.061135e-10

Further details can be obtained using effect() and providing the level for the
confidence interval, 'level'.
> predict(catmodel2)
Estimated adjusted geometric mean of the response variable (original space):

  cat  Estimate  lower95%  upper95%
1  No 0.3405043 0.2919107 0.3971872
2 Yes 1.4223169 0.9575328 2.1127059
> effect(catmodel2)
Adjusted change in the geometric mean of the response variable when the
explanatory variable changes from its reference level, 'No', to an
alternative level. Confidence interval for the difference was computing
based on 5000 bootstrap samples:

          EstimateDiff  lower95% upper95% EstimatePercent lower95% upper95%
No -> Yes     1.081813 0.6127211 1.757564        317.7089 171.1285 543.5352

Further information about interpreting the effect using effectInfo()
```

Example: cat allergen levels in the home

```
> plot(catmodel2, xname = "Cat at home", yname = "allergen leveles in the room",
+       las = 1, col = c("forestgreen", "red"))
```



Centre for Research
in Environmental
Epidemiology

Parc de Recerca Biomèdica de Barcelona
Doctor Aiguader, 88
08003 Barcelona (Spain)
Tel. (+34) 93 214 70 00
Fax (+34) 93 214 73 02

info@creal.cat
www.creal.cat

Thanks!

Questions?