Supplemental Material for

“Models with transformed variables: interpretation and software”

by J. Barrera-Gómez and X. Basagaña.
A Adjusted measures: setting the value of $K$

Suppose that we are interested in estimating the effect of an explanatory variable $X$ on the mean of a response variable $Y$, $\mathbb{E}(Y)$, based on the multiple linear regression model

$$Y = \beta_0 + \beta X + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad (1)$$

where $X_2, \ldots, X_p$ are other explanatory variables, and the unknown parameters $\beta_0$, and $\beta, \beta_2, \ldots, \beta_p$, are the intercept of the model and the regression coefficients, respectively; and $\epsilon$ is the normally distributed error term. Focusing only on the relationship between $X$ and $Y$ while keeping all the other covariates constant, model (1) can be written as

$$Y = \beta X + K + \epsilon, \quad (2)$$

where

$$K = \beta_0 + \beta_2 X_2 + \cdots + \beta_p X_p. \quad (3)$$

In order to describe the relationship between $X$ and $\mathbb{E}(Y)$ in (2), the value of $K$ should be set. Let $i = 1, \ldots, n$ be the identifier for the $i$th individual in the dataset used to fit model (1), and $X_{ji}$ the observed value of $X_j$ on the $i$th individual, $j = 2, \ldots, p$. Thus, the expected mean of $Y$, under model (1), for a given value $X = x_0$, computed in a population with values of the remaining variables $X_j$ equal to those observed in the $i$th individual is

$$\mathbb{E}(Y|X = x_0\{X_j=X_{ji}\}) = \beta x_0 + K_i, \quad (4)$$

where, from (3), $K_i = \beta_0 + \beta_2 X_2i + \cdots + \beta_p X_pi$. Computing (4) for all $i$ and then, averaging over all subjects, it results in

$$\mathbb{E}(Y|X = x_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y|X = x_0\{X_j=X_{ji}\}) = \frac{1}{n} \sum_{i=1}^{n} (\beta x_0 + K_i) = \beta x_0 + \frac{\sum_{i=1}^{n} K_i}{n} = \beta x_0 + \bar{K},$$

where

$$\bar{K} = \frac{\sum_{i=1}^{n} K_i}{n} = \frac{\sum_{i=1}^{n} (\beta_0 + \beta_2 X_2i + \cdots + \beta_p X_pi)}{n} = \beta_0 + \beta_2 \bar{X}_2 + \cdots + \beta_p \bar{X}_p. \quad (5)$$

If $X_j$ is binary, then $\bar{X}_j$ is its observed proportion in the dataset.

Thus, setting $K = \bar{K}$ as in (5) is equivalent to saying that measures and effects are calculated for an average individual in the population$^1$ and they can be interpreted as adjusted measures and adjusted effects. The software we provide automatically handles the calculation of $K$.

B Adjusted measures in a linear regression model under log transformations

For a log transformed $Y$ and an untransformed $X$, the estimated median (or geometric mean) of $Y$ associated to a value $X = x$ is

$$\hat{M}[Y|X = x] = e^{\beta x + \bar{K}}.$$
For an untransformed $Y$ and a log transformed $X$, the estimated mean of $Y$ associated to a value $X = x$ is

$$\hat{E}[Y|X = x] = \hat{\beta}\log(x) + \hat{K}.$$ 

For a log transformed $Y$ and a log transformed $X$, the estimated median or geometric mean of $Y$ associated to a value $X = x$ is

$$\hat{M}[Y|X = x] = e^{\hat{\beta}\log(x) + \hat{K}}.$$ 

### C Considerations on log and power transformations

All log transformations in this paper consider the natural logarithm (i.e., basis $e$), which is the default when log transforming variables using standard software. Some researchers use other bases in order to obtain coefficients with a predefined interpretation. For instance, $\log_2(X)$ is sometimes used to directly obtain the effect of a 2-fold increase in $X$. It is important to point out that using other logarithm bases or using the methodology described in this paper lead to exactly the same solution.

Also, it should be noted that the log transformation works only for positive variables. In the case of power transformations, there are values of the exponent that are incompatible with nonpositive variables. In addition, if a power transformation is applied to the response variable, the assumed bijectivity does not hold in some cases (e.g., quadratic transformation with both negative and positive values of the variable). Our software automatically deals with these constraints.

### D Relationship between $X$ and $Y$ changes in a linear model with logarithmic transformations

See eFigure 1 (page 3).

### E Guidance on choosing values for $c$ and $q$

The additive change in $X$, $c$, should be set taking into account the observed range of $X$. Otherwise, the computed effect size could be unrealistic, either too large and not reflecting a plausible change, or too small and reflecting a negligible increment that precludes an easy interpretation. One possible choice is setting $c$ at a value equivalent to the observed interquartile range of $X$. Thus, denoting $Q_1$ and $Q_3$ as the first and third quartiles, respectively, we can set $c = Q_3 - Q_1$. Selecting such type of change is common in order to ensure that the change in $X$ is plausible in the study population and it is helpful if one is not familiar with the units of $X$. If $X$ is binary, then necessarily $c = 1$.

Regarding the relative change in $X$, $q$, for instance, if $q = 2$, we get the effect for a 2-fold increase in $X$; or if $q = 10$, the effect is for a 10-fold increase in $X$. Multiplying $X$ by $q$ is equivalent to an $r = 100(q-1)\%$ change in $X$, which may be a more convenient way to express relative changes when $q < 2$. Thus, if $q = 1.5$, we get the effect for an $r = 50\%$ increase in $X$. The relative change in $X$, $q$, should be set taking into account the observed range of $X$. Thus, similarly to the setting of the additive change $c$, we recommend $q = Q_3/Q_1$, where $Q_1$ and $Q_3$ are the first and third quartile of $X$, respectively.
**eFigure 1:** Relationship between $X$ and $Y$ changes in a linear model with logarithmic transformations. Arrows in the plot area characterize an additive or multiplicative change in $X$ and the resulting additive or multiplicative change in $Y$. In a linear model without transformations (A), an additive change in $X$ implies an additive variation in $Y$. In the example, adding 1 unit to $X$ leads to an additive increase of 2 units in $Y$. In a model with log transformed $X$ (B), a multiplicative change in $X$ implies an additive change in $Y$. In the example, doubling $X$ leads to an additive increase of 1 unit in $Y$. In a model with log transformed $Y$ (C), an additive change in $X$ implies a multiplicative change in $Y$. In the example, adding 1 unit to $X$ leads to doubling $Y$. In a model with both $X$ and $Y$ log transformed (D), a multiplicative change in $X$ implies a multiplicative change in $Y$. In the example, doubling $X$ leads to a tripling $Y$. 
F  Approximate interpretation of the regression coefficient \( \beta \) of interest under log transformation in linear models, for a 1 unit or a 1% increase in the explanatory variable.

For the special case of small additive (c), relative (q) or percent (r) changes in \( X \), depending on the case, instead of calculating the effect size using the corresponding formula, one could use an approximation based on the first-order Taylor series expansion resulting in the effect of \( X \) being approximately proportional to (or linear in) c, q or r. Thus, eTable 1 shows this approximation in the case of log transformation in linear, logistic and Poisson regression models, for a 1 unit (c = 1) or a 1% (r = 1; equivalently, \( q = 1.01 \)) increase in the explanatory variable. Such results provide a direct interpretation of the coefficient \( \beta \) of interest.

The effect of \( X \) on \( Y \) while keeping all the other covariates constant, as we consider here, is known as marginal effect in those cases in which the change in \( X \) is very small (infinitesimal).

### eTable 1: Approximate interpretation of the regression coefficient \( \beta \) under linear, logistic and Poisson 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Log</th>
<th>Interpretation*</th>
<th>Approximation error$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>none</td>
<td>( \beta ) units change in the mean of ( Y ) for unit increase in ( X )</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>100% change in the median$^c$ of ( Y ) for unit increase in ( X )</td>
<td>(&lt; 10% \text{ if }</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>( \beta/100 ) units change in the mean of ( Y ) for 1% increase in ( X )</td>
<td>0.5% for any ( \hat{\beta} )</td>
</tr>
<tr>
<td></td>
<td>X, Y</td>
<td>( \beta )% change in the median$^c$ of ( Y ) for 1% increase in ( X )</td>
<td>(&lt; 10% \text{ if }</td>
</tr>
<tr>
<td>Logistic</td>
<td>none</td>
<td>( 1 + \hat{\beta} ) odds ratio of ( Y ) for unit increase in ( X )</td>
<td>(&lt; 9% \text{ if } \hat{\beta} \in [0, 0.5]; &lt; 2% \text{ if } \hat{\beta} \in [0, 0.2])</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>( 1 + \hat{\beta}/100 ) odds ratio of ( Y ) for 1% increase in ( X )</td>
<td>(&lt; 3% \text{ if }</td>
</tr>
<tr>
<td>Poisson</td>
<td>none</td>
<td>100% change in the mean of ( Y ) for unit increase in ( X )</td>
<td>(&lt; 5% \text{ if } \hat{\beta} \in [0, 0.1])</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>( \beta )% change in the mean of ( Y ) for 1% increase in ( X )</td>
<td>(&lt; 10% \text{ if }</td>
</tr>
</tbody>
</table>

$^a$: (1 - \( \alpha \))\% confidence interval is obtained when replacing \( \hat{\beta} \) by \( \hat{\beta} \pm z_{1-\alpha/2 \text{se}(\hat{\beta})} \). See eAppendix I for details.

$^b$: Percentage error relative to the true value of the effect.

$^c$: Equivalently, geometric mean.

G  Measures and effects in linear regression models with power transformations

Unlike in the case of log transformations, under power transformations the effect of \( X \) on \( Y \) cannot be summarized using appropriate additive or relative changes in \( X \) or \( Y \).

In general, the estimated median of \( Y \) associated to a given value of the explanatory variable \( X = x \) is

\[
\hat{M}[Y|X = x] = f^{-1}(\hat{\beta}\hat{x} + \hat{K}).
\]  

(6)

The additive change in the median of \( Y \), associated to varying \( X \) from \( X = u_1 \) to \( X = u_2 \), is

\[
\Delta \hat{M} = \hat{M}[Y|X = u_2] - \hat{M}[Y|X = u_1] = f^{-1}(\hat{\beta}\hat{u}_2 + \hat{K}) - f^{-1}(\hat{\beta}\hat{u}_1 + \hat{K}).
\]  

(7)
In addition, the percent change in the median of $Y$ can also be considered,

$$
\Delta \tilde{M}_\% = 100 \frac{\tilde{M}[Y|X = u_2] - \tilde{M}[Y|X = u_1]}{\tilde{M}[Y|X = u_1]} = 100 \left[ \frac{f^{-1}(\hat{\beta}u_2 + \hat{K})}{f^{-1}(\hat{\beta}u_1 + \hat{K})} - 1 \right] \%.
$$

(8)

In the cases with an untransformed $Y$, measures and effects can be interpreted in terms of the mean of $Y$, $E$. In general, the investigator is free to choose additive (using $c = u_2 - u_1$) or relative (using $q = u_2/u_1$ or $r = 100(u_2/u_1 - 1)\%$) changes for $X$, and to interpret the resulting changes in $Y$ as additive or relative (equations (7) or (8), respectively). In practice, though, it is often the case that one of the choices leads to easier interpretations than the others. Since it is difficult to generalize, we recommend creating tables with the four possible combinations, and then picking the one that results in the easiest interpretation. These tables should report the effects for several values of $u_1$, that is to say, for several basal values of $X$ along the observed range. To report realistic effects, the number and position of the points where the effect is computed should be set coherently with the value of $c$ or $r$. In general, if $X$ is binary, then necessarily $u_1 = 0$ and $u_2 = 1$ in (7) and (8), resulting in the effect associated to changing $X$ from its reference level to the alternative level. If $X$ is categorical with more than two levels, as many effects as alternative levels of $X$ should be computed, each of them being analogous to the case of a binary $X$. Our software facilitates the creation of such tables.

H Illustrative examples

H.1 Log transformation in the response variable: $\log(Y)$ vs. $X$

Consider the evaluation of the association between the intima media thickness of the carotid artery (IMT), measured in mm, $Y$, and age, in years, $X$. Variable $Y$ was log transformed to achieve normality. eFigure 2(a) shows the graphical relationship between $X$ and $Y$ before and after the transformation. Thus, consider the model

$$
\hat{Y} = \log(Y) = \hat{\beta}_0 + \beta X + \epsilon.
$$

(9)

Fitting model (9) resulted in $\hat{\beta}_0 = -0.878$ and $\hat{\beta} = 9.5 \cdot 10^{-3}$, with standard errors $\hat{\text{se}}(\hat{\beta}_0) = 1.8 \cdot 10^{-2}$ and $\hat{\text{se}}(\hat{\beta}) = 3.1 \cdot 10^{-4}$, respectively. The log transformation in $Y$ precludes the direct interpretation of the coefficients above. Next, we show how to interpret the age effect on the IMT in the natural scale of the variables.

For a log transformed $Y$ and an untransformed $X$, the estimated median (or geometric mean) of $Y$ associated to a value $X = x$ is, under model (9),

$$
\tilde{M}[Y|X = x] = e^{\hat{\beta}x + \hat{K}},
$$

(10)

where, in this case, $\hat{K} = \hat{\beta}_0$.

With regard to the effect of $X$, it can be reported as a measure which is independent of the values of both $X$ and $\hat{K}$. Indeed, an additive change of $c$ units in $X$ results in a relative change in the median or geometric mean of $Y$ equal to

$$
\tilde{\Delta}M_\% = 100 \frac{\tilde{M}[Y|X = x + c] - \tilde{M}[Y|X = x]}{\tilde{M}[Y|X = x]} = 100(e^{\hat{\beta}c} - 1)\%,
$$

(11)
**eFigure 2:** Three examples of log transformation in a linear model. A-B: intima media thickness (IMT) and age. C-D: birth weight and cord serum cotinine. E-F: cat allergen levels in the home, measured in the living room and in the bed mattress. Representations are in both the original (left column) and transformed (right columns) scales. In all three examples, synthetic data were simulated to emulate true data patterns observed in the cited studies.
when fixing the remaining explanatory variables, if any. The additive change in $X$, $c$, should be set taking into account the observed range of $X$ (see Section E). For instance, the first and third quartile of the variable age were 51 and 66 years, respectively. Thus, an additive change in $X$ equivalent to the interquartile range is $c = 66 - 51 = 15$ years. Using formula (11), we obtain $\Delta M_{\%} = 100(e^{0.0095\times 15} - 1)\% \approx 15.4\%$. Therefore, a 15-year increase in age is associated to a 15.4% increase in the median or geometric mean of the IMT. To obtain a confidence interval (see Section I), one first need to compute $\hat{\beta} \pm z_{1-\alpha/2}\hat{\text{se}}(\hat{\beta})$. For a 95% confidence interval, $\alpha = 0.05$ and $z_{1-\alpha/2} = 1.96$, and therefore $\hat{\beta} \pm z_{1-\alpha/2}\hat{\text{se}}(\hat{\beta}) = 0.0095 \pm 1.96 \cdot 0.00031$. Thus, the interval is $100(e^{0.0095\pm 1.96 \cdot 0.00031} - 1)\% = (14.3\%, 16.4\%)$.

For the special case of small changes in $X$, $|c| \ll 2/|\hat{\beta}|$, instead of calculating (11), one could use an approximation based on the first-order Taylor series expansion (see Section F) resulting in the effect of $X$ being approximately proportional to $c$, and report that a $c$ units change in $X$ leads to a $100\hat{\beta}c\%$ change in $Y$. The underestimation error is not greater than 5% if $c$ is not greater than $0.35/\hat{\beta}$, or 2% if $c$ is not greater than $0.2/\hat{\beta}$. Thus, for small values of $\hat{\beta}$, the quantity $100\hat{\beta}$ can be interpreted directly as the percent change in $Y$ for 1 unit increase in $X$, with error not greater than 2% if $|\hat{\beta}| < 0.2$, or not greater than 5% if $|\hat{\beta}| < 0.35$. In the IMT example, the result $100\hat{\beta} = 0.95 \approx 1$ can be directly interpreted as the percent increase in the IMT for 1 year increase in the age. One can get the exact result using (11) with $c = 1$.

H.2 Log transformation in the explanatory variable in a linear regression model: $Y$ vs. $\log(X)$

Consider now the evaluation of the association between birth weight, in grams, $Y$, and cord serum cotinine, in ng/ml, $X$. Figure 2(b) shows how the distribution of cotinine levels is more homogeneous in the logarithmic scale. Thus, consider the model

$$Y = \beta_0 + \beta \tilde{X} + \epsilon,$$

(12)

where $\tilde{X} = \log(X)$ is the log transformed $X$. When the linear model (12) was fitted, we obtained $\hat{\beta}_0 = 3406.5$ and $\hat{\beta} = -80$, with standard errors $\hat{\text{se}}(\hat{\beta}_0) = 42.6$ and $\hat{\text{se}}(\hat{\beta}) = 14.9$, respectively. Next, we show how to interpret the results in the natural scale of the variables.

For an untransformed $Y$ and a log transformed $X$, the estimated mean of $Y$ associated to a value $X = x$ is, under model (12),

$$\hat{E}[Y|X = x] = \hat{\beta} \log(x) + \hat{K},$$

where, in this case, $\hat{K} = \hat{\beta}_0$.

The effect of $X$ can be reported as a measure which is independent of the values of both $X$ and $\hat{K}$. Specifically, multiplying $X$ by a factor $q$ results in an additive change in the mean of $Y$ equal to

$$\Delta \hat{E} = \hat{E}[Y|X = qx] - \hat{E}[Y|X = x] = \hat{\beta} \log(q),$$

(13)

when fixing the remaining explanatory variables, if any. For instance, if $q = 2$, we get the effect for a 2-fold increase in $X$; or if $q = 10$, the effect is for a 10-fold increase in $X$. Multiplying $X$ by $q$ is equivalent to an $r = 100(q - 1)/q$ change in $X$, which may be a more convenient way to express relative changes when $q < 2$. Thus, if $q = 1.5$, we get the effect for an $r = 50\%$ increase in $X$. The relative change in $X$, $q$, should be set taking into account the observed range of $X$ (see Section E).
Thus, \( q \) can be set at a value equivalent to an interquartile ratio in cotinine. The first and third quartiles are 3.23 ng/ml and 39.23 ng/ml, respectively, and therefore \( q = 39.23/3.23 = 12.1 \). Then, from (13), \( \Delta E = -80 \log(12.1) \approx -200 \) g. Thus, the estimated effect of 12.1-fold increase in the cotinine level is a decrease of 200 g in the mean birth weight. A 95% confidence for this effect is (-273 g, -126 g). Alternatively, by exploring eFigure 2(b), one can see that several 10-fold changes occur in the population and choose the more common number \( q = 10 \), in which case \( \Delta E = -184 \) g, the effect of a 10-fold increase in the cotinine level.

For the special case of small changes in \( X \), that is to say \( q \) close to 1 or \( r \) close to 0, the first-order approximation of the Taylor series expansion of (13) (see Section F) results in that the effect of \( X \) is approximately proportional to \( r \) and one can report that an \( \% \) change in \( X \) leads to a change of \( \beta r/100 \) units in \( Y \). The overestimation error is not greater than 5% if \( r \) is not greater than 10%, or 2.5\% if \( r \) is not greater than 5\%. Thus, the quantity \( \hat{\beta}/100 \) can be interpreted directly as the change in \( Y \) for a 1\% increase in \( X \). This interpretation is not always useful. For instance, in the cotinine example, the result \( \beta \cdot 10/100 = -8 \) can be interpreted as the decrease in the mean birth weight, in grams, for a 10\% increase in cotinine levels, which is a negligible effect.

H.3 Log transformation in both the response and the explanatory variables in a linear regression model: \( \log(Y) \) vs. \( \log(X) \)

Consider an epidemiological study to assess the association between cat allergen levels (\( Fel d I \)) in the bed mattress, \( X \), and in the living room, \( Y \), in homes of study participants, taking into account cat ownership, \( C \).\(^7\) eFigure 2(c) shows that the log scale in both variables \( X \) and \( Y \) linearizes the relationship and also corrects apparent extreme values. Thus, we consider the model

\[
\hat{Y} = \log(Y) = \beta_0 + \beta \hat{X} + \beta_2 C + \epsilon, \quad (14)
\]

where \( \hat{X} = \log(X) \), and \( C = 1 \) in homes with cat and \( C = 0 \) otherwise. Fitting model (14) resulted in \( \hat{\beta}_0 = -0.054 \), \( \hat{\beta} = 0.639 \) and \( \hat{\beta}_2 = 1.527 \), with standard errors \( \hat{\text{se}}(\hat{\beta}_0) = 0.117 \), \( \hat{\text{se}}(\hat{\beta}) = 0.045 \) and \( \hat{\text{se}}(\hat{\beta}_2) = 0.231 \), respectively. Next, we show how to interpret the relationship in the natural scale of the variables.

For a log transformed \( Y \) and a log transformed \( X \), the estimated median or geometric mean of \( Y \) associated to a value \( X = x \) is, under model (14),

\[
\hat{M}[Y|X = x] = e^{\hat{\beta}\log(x)+\hat{K}}, \quad (15)
\]

where, in the cat example, \( \hat{K} = \hat{\beta}_0 + \hat{\beta}_2 C \). The proportion of homes with cat was \( \bar{C} = 0.14 \) and, therefore, the mean value of \( \hat{K} \) was \( -0.05 + 1.53 \cdot 0.14 = 0.16 \) (see Section A), which, used in (15), provides the curve for the estimated adjusted median or geometric mean of \( Y \) as a function of \( X \).

The effect of \( X \) can be reported as a measure which is independent of the values of both \( X \) and \( \hat{K} \). Indeed, multiplying \( X \) by a factor \( q \) results in a percent change in the median or geometric mean of \( Y \) equal to

\[
\Delta M_{\%} = 100 \frac{\hat{M}[Y|X = qx] - \hat{M}[Y|X = x]}{\hat{M}[Y|X = x]} = 100(q^{\hat{\beta}} - 1)\% \quad (16)
\]

when fixing the remaining explanatory variables, if any. The multiplicative factor \( q \) is equivalent to a percent change \( r = 100(q - 1)\% \).
Applying this to the cat allergen example, we first can set \( q = Q_3/Q_1 \), similarly to what we did in the previous example. The first and third quartiles of the allergen level in the mattress are 0.064 \( \mu g/g \) and 0.557 \( \mu g/g \), respectively, and therefore \( q = 0.557/0.064 = 8.76 \) or, equivalently, \( r = 100(8.76 - 1)\% = 776\% \). Then, \( \hat{\Delta}M_{\%} = 100(8.76^{0.639} - 1)\% = 300\% \). Thus, each 8.76-fold or, equivalently, a 776\% increase in allergen level in the mattress is associated with a 300\% increase in the median or geometric mean of the allergen level in the living room. The 95\% confidence interval is (231\%, 384\%). Alternatively, by exploring eFigure 2(c), one can see that several 10-fold changes occur in the population and choose \( q = 10 \), in which case \( \hat{\Delta}M_{\%} = 336\% \), the increase in the allergen level in the living room associated with a 10-fold increase in allergen level in the mattress.

For the special case of small changes in \( X \), that is to say \( q \) close to 1 or \( r \) close to 0, the first-order approximation of the Taylor series expansion of (16) (see Section F) results in that the effect of \( X \) is approximately proportional to \( r \) and one can report that an \( r\% \) change in \( X \) leads to a \( \beta r\% \) change in the median or geometric mean of \( Y \). This approximation is suitable if \( |r| \ll \frac{200}{|\beta| - 1} \). Taking \( r = 1 \), one obtains that \( \hat{\beta} \) can be directly interpreted as the percent change in the median or geometric mean of \( Y \) for a 1\% increase in \( X \). This measure is often known as elasticity. This is valid for values of \( \hat{\beta} \) such that \( |\hat{\beta}| \ll 200 \). In the cat allergen example, \( |\hat{\beta}| = 0.639 \ll 200 \), and therefore we can interpret \( \hat{\beta} = 0.639 \) as the elasticity, that is to say, that 1\% increase in the median or geometric mean of allergen level in the mattress is associated to a 0.64\% increase in the allergen level in the living room.

**H.4 Power transformation in both the response and the explanatory variables in a linear regression model: \( 1/Y^2 \) vs. \( 1/\sqrt{X} \)**

Consider the modeling of the association between blood levels of triglycerides, \( X \), and glucose, \( Y \), both measured in mg/dl, taken from a study of cardiovascular health.\(^5\) eFigure 7(a) shows the relationship between triglycerides and glucose levels in the original scale, in which residuals of a linear regression model showed non-linearity, non-normality and several influential observations. In order to linearize the relationship, we consider the model

\[
\hat{Y} = \frac{1}{Y^2} = \beta_0 + \beta \hat{X} + \epsilon,
\]

where \( \hat{X} = 1/\sqrt{X} \). Thus, this case corresponds to \( f(Y) = 1/Y^2 \), which was accepted to be normally distributed, and \( g(X) = 1/\sqrt{X} \). Fitting the model resulted in \( \hat{\beta}_0 = 0.54 \cdot 10^{-4} \) (mg/dl)\(^{-2} \) and \( \hat{\beta} = 5.72 \cdot 10^{-4} \) (mg/dl)\(^{-3/2} \), with standard errors, \( \hat{se}(\hat{\beta}_0) = 8.41 \cdot 10^{-6} \) (mg/dl)\(^{-2} \) and \( \hat{se}(\hat{\beta}) = 7.98 \cdot 10^{-5} \) (mg/dl)\(^{-2} \), respectively. Next, we show how to interpret the triglycerides effect on the glucose levels in the natural scale of the variables.

The estimated median of \( Y \) associated to a given value of the explanatory variable \( X = x \) is

\[
\hat{M}[Y|X = x] = f^{-1}(\hat{\beta} \hat{x} + \hat{K}).
\]  \hspace{1cm} (17)

In this example, \( f^{-1}(Y) = 1/\sqrt{Y} \) so, using (17), the equation for the estimated median of the glucose level for a given triglycerides level, \( X = x \), becomes

\[
\hat{M}[Y|X = x] = \frac{1}{\sqrt{\hat{\beta}_0 + \hat{\beta} \hat{x}}} = \frac{100}{\sqrt{0.54 + 5.72/\sqrt{x}}} \text{ mg/dl,}
\]

9
which can then be plotted as in eFigure 7(a).

Apart from having a graphical description of the curve \( \hat{M}(X) \), one may be interested in providing measures of effect. In eFigure 7(a) one can see that the curve increases faster at low levels of \( X \) than at high levels, where the curve is flatter. Unlike in the case of log transformations, this behavior cannot be summarized using appropriate additive or relative changes in \( X \) or \( Y \) (see Section G). Specifically, the change in the median of glucose level associated to a change in triglycerides level from \( u_1 \) to \( u_2 \), becomes (see Section G)

\[
\Delta \hat{M} = \frac{100}{\sqrt{0.54 + 5.72/\sqrt{u_2}}} - \frac{100}{\sqrt{0.54 + 5.72/\sqrt{u_1}}} \text{mg/dl}, \tag{18}
\]

In addition, the percent change in the median of \( Y \) can also be considered,

\[
\Delta \hat{M}_\% = 100 \frac{\hat{M}[Y|X = u_2] - \hat{M}[Y|X = u_1]}{\hat{M}[Y|X = u_1]} \% = 100 \left[ \frac{f^{-1}(\beta \hat{u}_2 + \hat{K})}{f^{-1}(\beta \hat{u}_1 + \hat{K})} - 1 \right] \%, \tag{19}
\]

which, in the glucose example, becomes

\[
\Delta \hat{M}_\% = 100 \left( \sqrt{\frac{0.54 + 5.72/\sqrt{u_1}}{0.54 + 5.72/\sqrt{u_2}}} - 1 \right) \%. \tag{20}
\]

Confidence intervals can be obtained as explained in Section I. In order to explore the effect of triglycerides level on glucose level, and following Section G, we first compute the 2.5th and 97.5th
**eTable 2:** Estimated median of glucose level ($\hat{M}(x)$) and estimated changes both additive and percent ($\hat{\Delta M}$ and $\hat{\Delta M}_\%$, respectively), associated to a $c = 50$ mg/dl additive increase in triglycerides level, $x$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x + 50$</th>
<th>$\hat{M}(x)$ (95% CI)</th>
<th>$\hat{\Delta M}$ (95% CI)</th>
<th>$\hat{\Delta M}_%$ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100</td>
<td>86.0 (83.9, 88.4)</td>
<td>8.7 (6.3, 10.9)</td>
<td>10.1 (7.3, 13.0)</td>
</tr>
<tr>
<td>100</td>
<td>150</td>
<td>94.7 (93.3, 96.3)</td>
<td>4.8 (3.3, 6.5)</td>
<td>5.1 (3.5, 6.8)</td>
</tr>
<tr>
<td>150</td>
<td>200</td>
<td>99.5 (97.2, 102.0)</td>
<td>3.2 (2.9, 3.5)</td>
<td>3.2 (2.7, 3.7)</td>
</tr>
<tr>
<td>200</td>
<td>250</td>
<td>102.8 (99.6, 106.3)</td>
<td>2.4 (1.6, 3.4)</td>
<td>2.3 (1.6, 3.2)</td>
</tr>
</tbody>
</table>

**eTable 3:** Estimated median of glucose level ($\hat{M}(x)$) and estimated changes both additive and percent ($\hat{\Delta M}$ and $\hat{\Delta M}_\%$, respectively), associated to an $r = 50\%$ increase in triglycerides level, $x$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x + 50% x$</th>
<th>$\hat{M}(x)$ (95% CI)</th>
<th>$\hat{\Delta M}$ (95% CI)</th>
<th>$\hat{\Delta M}_%$ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.0</td>
<td>75.0</td>
<td>86.0 (83.9, 88.4)</td>
<td>5.2 (3.8, 6.4)</td>
<td>6.0 (4.4, 7.6)</td>
</tr>
<tr>
<td>75.0</td>
<td>112.5</td>
<td>91.2 (89.7, 92.8)</td>
<td>5.0 (3.5, 6.5)</td>
<td>5.5 (3.9, 7.1)</td>
</tr>
<tr>
<td>112.5</td>
<td>168.8</td>
<td>96.2 (94.5, 97.9)</td>
<td>4.7 (3.3, 6.3)</td>
<td>4.9 (3.4, 6.5)</td>
</tr>
<tr>
<td>168.8</td>
<td>253.1</td>
<td>100.9 (98.2, 103.8)</td>
<td>4.4 (3.0, 6.2)</td>
<td>4.4 (3.0, 5.9)</td>
</tr>
</tbody>
</table>

percentiles of triglycerides level, which were 47 mg/dl and 313 mg/dl, respectively. Thus, we can report the effects (18) and (20) between pairs of consecutive values of $u_1 = 50, 100, 150, 200$ and 250 mg/dl, that is to say, for a $c = 50$ mg/dl additive increase, and also between pairs of consecutive values: 50, 75, 112.5, 168.8 and 253.1 mg/dl, that is, for an $r = 50\%$ increase. Results are shown in eTable 2 and eTable 3, respectively. Exploring these tables, one can see that the more easily interpretable effect appears to be the additive change in the median of $Y$ associated to a percent change in $X$ ($\hat{\Delta M}$). Indeed, in eTable 3, one can see that, for any given value of the triglycerides level along the observed range, a 50% increase in triglycerides level is associated to around a 4.8 mg/dl increase in the median glucose level. For the other measures in eTables 2 and 3, the effect is more dependent on the basal value of the triglycerides level. Our software facilitates the creation of tables similar to eTables 2 and 3.

**H.5 Log transformation in the explanatory variable in a logistic regression model**

The logistic regression model for a binary response $Y$, and log transformed quantitative explanatory variable $X$, is

$$\log \left( \frac{P(Y = 1)}{P(Y = 0)} \right) = \beta \tilde{X} + K,$$

(21)

where $\tilde{X} = \log(X)$. As is known, under the logistic regression model for an untransformed $X$, the odds ratio of $Y$, OR, associated to a one unit increase in $X$, is $e^\beta$. Instead, under model (21), the OR associated to a relative change in $X$, that is, multiplying $X$ by a factor $q$ (or, equivalently, changing $X$ an $r = 100(q - 1)\%$), is $e^{\beta \log(q)}$, which is equivalent to

$$\hat{\text{OR}} = q^\beta,$$

(22)
which is independent of the remaining explanatory variables in the model, if any. For relatively small values of \( q \) or \( r \) \(|q - 1| \ll \frac{2}{|\hat{\beta} - 1|}\) or \( r \ll \frac{200}{|\hat{\beta} - 1|} \), respectively), one can use the first-order approximation of the Taylor series expansion of the effect (see Section F), \( \hat{\text{OR}} \approx 1 + (q - 1)\hat{\beta} \) or, equivalently, \( \hat{\text{OR}} \approx 1 + \hat{\beta} r/100 \). This approximation underestimates the effect (see Section F).

Revisiting the cotinine example (see Section H.2), suppose we are now interested in the association between low birth weight, defined as weight lower than 2500 g, \( Y \), and cotinine level, \( X \). After log transforming \( X \), we considered the model

\[
\log \left( \frac{P(Y = 1)}{P(Y = 0)} \right) = \beta_0 + \tilde{\beta} X, \tag{23}
\]

where \( \tilde{X} = \log(X) \). Fitting model (23) resulted in \( \hat{\beta} = 0.331 \), with standard error \( \hat{\text{se}}(\hat{\beta}) = 0.129 \), which leads to a 95% confidence interval (0.078, 0.583). The first and third quartiles of \( X \) were 3.23 ng/mL and 39.23 ng/mL, respectively. Thus, setting \( q = 39.23/3.23 = 12.1 \), equation (22) results in \( \hat{\text{OR}} = 2.28 \). That is, an interquartile range increase (or a 12.1-fold increase) in cotinine level is associated to an odds ratio of low birth weight equal to 2.28 (i.e., a 128% increase in the odds of low birth weight). Using the 95% confidence interval for \( \hat{\beta} \) above, the corresponding confidence interval for the effect was (1.22, 4.29).

H.6 Log transformation in the explanatory variable in a Poisson regression model

The Poisson regression model for a count response \( Y \), and log transformed quantitative explanatory variable \( X \), is

\[
\log(E) = \beta \tilde{X} + K, \tag{24}
\]

where \( E \) is the mean of \( Y \) and \( \tilde{X} = \log(X) \). As is known, under the Poisson regression model for an untransformed \( X \), the percent change in the mean of \( Y \), associated to a one unit increase in \( X \), is \( 100(e^{\beta} - 1)\% \). Instead, under model (24), the percent change in the mean of \( Y \) associated to a relative change in \( X \), that is, multiplying \( X \) by a factor \( q \) (or, equivalently, increasing \( X \) an \( r = 100(q - 1)\% \)), is \( 100(e^{\beta \log(q)} - 1)\% \), which is equivalent to

\[
\Delta E\% = 100(q^{\hat{\beta}} - 1)\%, \tag{25}
\]

which is independent of the remaining explanatory variables in the model, if any. For relatively small values of \( q \) or \( r \) \(|q - 1| \ll \frac{2}{|\hat{\beta} - 1|}\) or \( r \ll \frac{200}{|\hat{\beta} - 1|} \), respectively), one can use the first-order approximation of the Taylor series expansion of the effect (see Section F), \( \Delta E\% \approx 100\hat{\beta}(q - 1) \) or, equivalently, \( \Delta E\% \approx \hat{\beta} r \). For instance, if \( r = 1\% \), one can directly interpret \( \hat{\beta} \) as the approximate percent change in the mean of \( Y \) associated to a 1% increase in \( X \). The relative error in this approximation is lower than 10% if \(|\hat{\beta}| < 20\) and lower than 5% if \(|\hat{\beta}| < 10\) (see Section F).

Suppose a hypothetical study evaluating the association between the yearly number of asthma attacks, \( Y \), and the cat allergen levels, \( X \). Cat allergen levels were log transformed and model (24) was fitted resulting in \( \hat{\beta} = 0.092 \), with standard error \( \hat{\text{se}}(\hat{\beta}) = 0.017 \), and thus, a 95% confidence interval (0.06, 0.13). The first and third quartiles of cat allergen levels were 0.064 \( \mu \text{g/g} \) and 0.557
Thus, setting \( q = 0.557/0.064 = 8.76 \), equation (25) results in \( \Delta E\% = 22.1\% \). That is, an interquartile range (or a 8.76-fold) increase in allergen levels was associated to a 22.1% increase in the mean yearly number of asthma attacks. The corresponding confidence interval for the effect was (13.9%, 32.6%).

I Confidence intervals for measures and effects

Under logarithm transformations in the response variable, the explanatory variable, or both, a 100(1\( - \alpha \))% confidence interval for measures and effects can be obtained by just replacing \( \hat{\beta} \) by its confidence interval limits, \( \hat{\beta} \pm z_{1-\alpha/2} \hat{\sigma}(\hat{\beta}) \), where \( z_{1-\alpha/2} \) is the 100(1\( - \alpha/2 \))th percentile of the standard normal distribution. However, our software uses the \( t \) distribution, which is more appropriate for small sample sizes, instead of the normal distribution. It should be noted that, although these distributions provide the shortest possible confidence interval in the transformed scale, there is no assurance that the resulting intervals in the original space are also the shortest possible.

Under power transformations, a confidence interval for the median of the response variable (or the mean, if the response variable is not transformed) can be obtained by first computing a confidence interval for the linear predictor on the transformed scale, using the estimates of the \( \beta \) coefficients in the model and their estimated covariances, and then applying the inverse transformation, if any, to the previous interval limits. In the case of effects, a confidence interval can be obtained using nonparametric bootstrap. Essentially, this method consists in resampling from the original data, estimating the effects in the new dataset, and repeating the process \( N \) times (e.g., \( N = 999 \)). One can then use the percentiles of the distribution obtained with the resampling process to calculate confidence intervals.

J R and Stata software usage

In this section, we show how to reproduce results in the examples of the paper using statistical software. Specifically, we do it using both R, in section J.1, and Stata, in section J.2. Using our R package, confidence intervals for measures and effects are computed by transforming the endpoints of the intervals in the transformed scale when it is possible, while non-parametric bootstrap is used otherwise. In Stata, when transforming the endpoints is not possible, the delta method is used. As it can be seen, in the cases of the examples presented in the paper, results differ slightly in general.

J.1 The R package tlm

In this section, we illustrate the usage of our tlm package (version 0.1.2) for R (R Foundation for Statistical Computing, Vienna, Austria). The package tlm facilitates the computation and presentation of the effects described in the paper. The package performs the model fitting in the transformed space and provides the effects, both numerically and graphically, on the original scale. It also computes confidence intervals for measures and effects presented here, based on the \( t \)-distribution or bootstrap resampling, depending on the case (eAppendix I). The package and a user’s guide are available at http://cran.r-project.org/web/packages/tlm/.
**J.1.1 Example 1: intima media thickness and age**

Once the package has been downloaded and installed, we should load it in our R session:

```r
> library(tlm)
```

Then, data can be loaded with

```r
> data(imt)
```

The `imt` data were simulated to emulate true data patterns observed in a real study. Data contains the age, in years, and the intima media thickness of the carotid artery (IMT), in mm, for 2784 adults. In addition, the logarithm of IMT was also computed. Below one can see the first rows of the dataset and a summary.

```r
> head(imt)
```

<table>
<thead>
<tr>
<th>age</th>
<th>imt</th>
<th>logimt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>53</td>
<td>-0.41723048</td>
</tr>
<tr>
<td>2</td>
<td>67</td>
<td>-0.08168583</td>
</tr>
<tr>
<td>3</td>
<td>57</td>
<td>-0.04710877</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>-0.32489555</td>
</tr>
<tr>
<td>5</td>
<td>73</td>
<td>-0.30029063</td>
</tr>
<tr>
<td>6</td>
<td>57</td>
<td>-0.26940367</td>
</tr>
</tbody>
</table>

```r
> summary(imt)
```

<table>
<thead>
<tr>
<th>age</th>
<th>imt</th>
<th>logimt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>32.00</td>
<td>-0.9794</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>51.00</td>
<td>-0.4609</td>
</tr>
<tr>
<td>Median</td>
<td>59.00</td>
<td>-0.3206</td>
</tr>
<tr>
<td>Mean</td>
<td>58.48</td>
<td>-0.3196</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>66.00</td>
<td>-0.1834</td>
</tr>
<tr>
<td>Max.</td>
<td>86.00</td>
<td>0.4567</td>
</tr>
</tbody>
</table>

Further information about data is available with the help function:

```r
> ?imt
```

Suppose that we are interested in the effect of age on IMT, under a linear regression model with log transformation in the response variable, IMT. The model can be fitted as follows:

```r
> modimt <- tlm(y = logimt, x = age, data = imt, ypow = 0)
```

where `ypow = 0` indicates that the response variable is already log transformed. The fitted model results in:

```r
> modimt
```
Linear regression fitted model in the transformed space
-------------------------------------------------------

Transformations:
In the response variable: log

Call:
lm(formula = logimt ~ age, data = imt)

Coefficients:
(Intercept)    age
  -0.877692  0.009543

Further information on the fitted model is available using the `summary` function:

```r
> summary(modimt)
```

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

Transformations:
In the response variable: log

Call:
lm(formula = logimt ~ age, data = imt)

Residuals:
  Min     1Q   Median     3Q    Max
-0.57891 -0.11792  0.00142  0.11879  0.61968

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.877692   0.018398  -47.71   <2e-16 ***
age          0.009543   0.000309   30.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1778 on 2782 degrees of freedom
Multiple R-squared:  0.255, Adjusted R-squared:  0.2547
F-statistic: 952.2 on 1 and 2782 DF,  p-value: < 2.2e-16

The function `effectInfo` provides information on interpreting the relationship between age ($X$) and IMT ($Y$):

```r
> effectInfo(modimt)
```

The effect of $X$ on $Y$ can be summarized with a single number as follows:

- Change in $X$: additive of $c$ units
- Type of effect on $Y$: percent change in the geometric mean of $Y$
- Effect size: $100 \times [\exp(c \times \text{beta}) - 1] \%$
beta coefficient estimate:

|        | Estimate | Std. Error | t value | Pr(>|t|) |
|--------|----------|------------|---------|----------|
| age    | 0.009543334 | 0.0003092751 | 30.8571 | 4.531758e-180 |

Further details can be obtained using effect().

The function effect provides as default the expected change in IMT for an additive change in age equal to the interquartile range:

> effect(modimt)

Computing effects...

Adjusted percent change in the geometric mean of the response variable for a 'c' units additive change in the explanatory variable equivalent to the interquartile range:

<table>
<thead>
<tr>
<th>c</th>
<th>Estimate</th>
<th>lower95%</th>
<th>upper95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>15.39029</td>
<td>14.3454</td>
</tr>
</tbody>
</table>

For further information on interpreting the effect use effectInfo().

A graphical representation of the relationship between age and IMT is available using the plot function. For instance, the code:

> plot(modimt, observed = T, xname = "Age (years)", yname = "IMT (mm)"")

provides eFigure 4, which represents the expected geometric mean (or median) of IMT (and a 95% confidence interval) as a function of the age. The argument observed controls whether observations are shown in the plot (default is FALSE). The argument type controls if the plot should be shown in the original space of the variables (default) or in the transformed space, or if it should be a model diagnostics plot. Further information on the usage of the plot function is available using

> ?tlm

### J.1.2 Example 2: birth weight and cord serum cotinine

Data can be loaded with

> data(cotinine)

The cotinine data were simulated to emulate true data patterns observed in a real study. Data contains the birth weight, in grams, and the cord serum cotinine level in the mother, in ng/ml, in 351 newborns. In addition, the logarithm of cotinine levels was also computed. Data also contains a binary variable indicating whether the birth weight was low (defined as lower than 2500 g). Further information about data is available using ?cotinine. Below one can see the first rows of the dataset and a summary.

> head(cotinine)
**eFigure 4:** Expected geometric mean (or median) of IMT (and a 95% confidence interval) as a function of the age.

<table>
<thead>
<tr>
<th>cotinine</th>
<th>logcotinine</th>
<th>weight</th>
<th>underweight</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5.1584035</td>
<td>1.640627</td>
<td>3626</td>
</tr>
<tr>
<td>5</td>
<td>0.2909473</td>
<td>-1.234613</td>
<td>3672</td>
</tr>
<tr>
<td>11</td>
<td>4.1119142</td>
<td>1.413889</td>
<td>3779</td>
</tr>
<tr>
<td>12</td>
<td>3.0037959</td>
<td>1.099877</td>
<td>3540</td>
</tr>
<tr>
<td>14</td>
<td>5.9240779</td>
<td>1.779025</td>
<td>3179</td>
</tr>
<tr>
<td>17</td>
<td>7.3854370</td>
<td>1.999510</td>
<td>2494</td>
</tr>
</tbody>
</table>

> `summary(cotinine)`

<table>
<thead>
<tr>
<th>cotinine</th>
<th>logcotinine</th>
<th>weight</th>
<th>underweight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>0.200</td>
<td>Min.</td>
<td>-1.609</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>3.234</td>
<td>1st Qu.</td>
<td>1.174</td>
</tr>
<tr>
<td>Median</td>
<td>7.385</td>
<td>Median</td>
<td>2.000</td>
</tr>
<tr>
<td>Mean</td>
<td>39.431</td>
<td>Mean</td>
<td>2.353</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>39.226</td>
<td>3rd Qu.</td>
<td>3.669</td>
</tr>
<tr>
<td>Max.</td>
<td>910.000</td>
<td>Max.</td>
<td>6.813</td>
</tr>
</tbody>
</table>

Suppose that we are interested in the effect of cotinine level on birth weight, under a linear regression model with log transformation in the explanatory variable, cotinine. The model can be fitted as follows:
> modcot <- tlm(y = weight, x = logcotinine, data = cotinine, xpow = 0)

where \(x_{\text{pow}} = 0\) indicates that the explanatory variable is already log transformed. The fitted model provides the following results:

> summary(modcot)

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

Transformations:
  In the explanatory variable: log

Call:
  lm(formula = weight ~ logcotinine, data = cotinine)

Residuals:
  Min 1Q Median 3Q Max
  -1390.64 -280.11 -2.95 300.47 1422.31

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 3406.49 | 42.59 | 79.978 | <2e-16 *** |
| logcotinine | -80.00 | 14.95 | -5.351 | 1.58e-07 *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 450.1 on 349 degrees of freedom
Multiple R-squared: 0.07583, Adjusted R-squared: 0.07318
F-statistic: 28.64 on 1 and 349 DF, p-value: 1.585e-07

The function `effectInfo` provides information on interpreting the relationship between cotinine level (\(X\)) and weight (\(Y\)):

> effectInfo(modcot)

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 \(r = 100 \times (q - 1)\%\) to \(X\))
- Type of effect on \(Y\): additive change in the mean of \(Y\)
- Effect size: \(\beta \times \log(q)\) units of \(Y\)

  beta coefficient estimate:

  | Estimate | Std. Error | t value | Pr(>|t|) |
  |----------|------------|---------|----------|
  | logcotinine | -80.00108 | 14.94986 | -5.351292 | 1.584903e-07 |

Further details can be obtained using `effect()`.

The function `effect` provides as default the expected change in birth weight for a relative (or percent) change in cotinine level equal to the interquartile ratio:

> effect(modcot)
Computing effects...

Adjusted additive change in the mean of the response variable for an 
'r'\% change in the explanatory variable equivalent to the interquartile 
ratio:

\begin{verbatim}
  r  Estimate  lower95%  upper95%
  1  1112.878  -199.649  -273.027
\end{verbatim}

For further information on interpreting the effect use effectInfo().

For a 10-fold increase in the cotinine level, the effect is

\begin{verbatim}
> effect(modcot, q = 10)
\end{verbatim}

Computing effects...

Adjusted additive change in the mean of the response variable for an 
'r' = 900\% change in the explanatory variable:

\begin{verbatim}
  r  Estimate  lower95%  upper95%
  1  900     -184.2093  -251.9126
\end{verbatim}

For further information on interpreting the effect use effectInfo().

The code

\begin{verbatim}
> plot(modcot, xname = "Cotinine (ng/ml)", yname = "birth weight (g)"
\end{verbatim}

provides eFigure 5, which represents the expected mean of birth weight as a function of the cotinine level.

J.1.3 Example 3: cat allergen levels in bed mattress and in living room

Data can be loaded with

\begin{verbatim}
> data(feld1)
\end{verbatim}

The feld1 data were simulated to emulate true data patterns observed in a real study. Data contains the cat allergen levels (Fel d 1), in \(\mu g/g\), measured in both the bed mattress and the living room of 471 homes. In addition, the logarithm of both variables was calculated. Data also contains a binary variable indicating whether there was a domestic cat at home. Further information about data is available using ?feld1. Below one can see the first rows of the dataset and a summary.

\begin{verbatim}
> head(feld1)
\end{verbatim}
**eFigure 5:** Expected mean of birth weight (and a 95% confidence interval) as a function of the cotinine level.

```
mattress  room logmattress logroom cat
1  0.66504894 0.26758221 -0.4078946 -1.3183284 yes
2  0.02723504 0.16796992 -3.6032509 -1.7839704 no
3  0.16773827 0.76728932 -1.7853504 -0.2648913 no
4  0.01391101 0.05480368 -4.2750750 -2.9039979 no
5  0.04216982 0.11931900 -3.1660504 -2.1259547 yes
6  1.44212520 17.29388484 0.3661179 2.8503530 yes
> summary(feld1)

mattress  room logmattress logroom cat
  Min. : 0.0030 Min. : 0.0017 Min. :-5.7975 Min. :-6.3565
  1st Qu.: 0.0636 1st Qu.: 0.1076 1st Qu.:-2.7559 1st Qu.:-2.2295
  Median : 0.1773 Median : 0.3081 Median :-1.7297 Median :-1.1774
  Mean : 3.1976 Mean : 27.5263 Mean :-1.5942 Mean :-0.8528
  3rd Qu.: 0.5565 3rd Qu.: 1.1713 3rd Qu.:-0.5861 3rd Qu.: 0.1581
  cat
  no :403
  yes: 68
```
Suppose that we are interested in the association between the allergen levels in the bed mattress (considered as the explanatory variable) and in the living room (considered as the response variable), under a linear regression model with log transformation in both variables. The model can be fitted as follows:

```r
> modcat <- tlm(y = logroom, x = logmattress, z = cat, data = feld1, ypow = 0, xpow = 0)
```

where `ypow = 0` and `xpow = 0` indicate that both the response and the explanatory variables are already log transformed, respectively. The argument `z = cat` indicates that the binary variable `cat` is included in the model as another explanatory variable. The fitted model provides the following results:

```r
> summary(modcat)
```

Linear regression fitted model in the transformed space
-------------------------------------------------------
Transformations:
    In the response variable: log
    In the explanatory variable: log
Call:
    lm(formula = logroom ~ logmattress + cat, data = feld1)
Residuals:
    Min     1Q   Median     3Q    Max
   -5.9172 -1.1228  -0.0631   0.9464   6.0440
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.05412    0.11651  -0.465  0.642
logmattress  0.63936    0.04454  14.354  < 2e-16 ***
catyes      1.52747    0.23098   6.613  1.03e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.68 on 468 degrees of freedom
Multiple R-squared:  0.4192,  Adjusted R-squared:  0.4167
F-statistic: 168.9 on 2 and 468 DF,  p-value: < 2.2e-16

The function `effectInfo` provides information on interpreting the relationship between mattress \(X\) and living room \(Y\) allergen levels:

```r
> effectInfo(modcat)
```

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 \(r = 100 \times (q - 1)\%\) to \(X\))
- Type of effect on \(Y\): percent change in the geometric mean of \(Y\)
- Effect size: \(100 \times (q^{\beta} - 1)\%\)
beta coefficient estimate:
  Estimate Std. Error  t value   Pr(>|t|)
logmattress  0.6393565 0.04454319 14.35363 5.610118e-39

Further details can be obtained using effect().

The function `effect` provides as default the expected change in living room allergen level for a relative (or percent) change in mattress allergen level equal to the interquartile ratio, adjusted by averaging among homes with and without domestic cat:

```r
> effect(modcat)
```

Computing effects...

Adjusted percent change in the geometric mean of the response variable for an 'r'\% change in the explanatory variable equivalent to the interquartile ratio:

<table>
<thead>
<tr>
<th>r</th>
<th>Estimate lower95%</th>
<th>upper95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>775.6541</td>
<td>300.393</td>
</tr>
<tr>
<td>3</td>
<td>231.1347</td>
<td>384.137</td>
</tr>
</tbody>
</table>

For further information on interpreting the effect use `effectInfo()`.

For a 10-fold increase in mattress allergen level, the adjusted effect is

```r
> effect(modcat, q = 10)
```

Computing effects...

Adjusted percent change in the geometric mean of the response variable for an 'r' = 900\% change in the explanatory variable:

<table>
<thead>
<tr>
<th>r</th>
<th>Estimate lower95%</th>
<th>upper95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>900.0000</td>
<td>256.3902</td>
</tr>
<tr>
<td>3</td>
<td>433.1948</td>
<td>283.2010</td>
</tr>
</tbody>
</table>

For further information on interpreting the effect use `effectInfo()`.

The code

```r
> plot(modcat, xname = "Mattress levels", yname = "living room levels")
```

provides eFigure 6, which represents the expected geometric mean (or median) of the cat allergen levels in the living room as a function of the cat allergen levels in the bed mattress.
**eFigure 6:** Expected geometric mean (or median) of the cat allergen levels in the living room (and a 95% confidence interval) as a function of the cat allergen levels in the bed mattress. Both variables were measured in µg/g.

### J.1.4 Example 4: glucose and triglycerides levels in blood

Data can be loaded with

```r
> data(glucose)
```

The `glucose` data were simulated to emulate true data patterns observed in a real study. Data contains the blood levels of glucose and triglycerides, both in mg/dl, measured in 471 adults. In addition, the transformations $f(Y) = 1/Y^2$ and $g(X) = 1/\sqrt{X}$ were computed for glucose and triglycerides levels, respectively. Further information about data is available using `?glucose`. Below one can see the first rows of the dataset and a summary.

```r
> head(glucose)
```

```
     trigly  gluc2tri inv12tri inv2glu
 1  264     116  0.06154575 7.431629e-05
 2  151     123  0.08137885 6.609822e-05
 3   67      96  0.12216944 1.085069e-04
 4   73      86  0.17041115 1.352082e-04
 5  180     104  0.07453560 9.245562e-05
 6  130     114  0.08770580 7.694675e-05
```
Suppose that we are interested in the association between the triglycerides level (considered as the explanatory variable) and the glucose level (considered as the response variable), under a linear regression model with the transformations described above. The model can be fitted as follows:

```r
> modgluco <- tlm(y = inv2glu, x = inv12tri, data = glucose, ypow = -2, xpow = -1/2)
```

which provides the following results:

```r
> summary(modgluco)
```

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

Transformations:
  In the response variable: power, exponent = -2
  In the explanatory variable: power, exponent = -1/2

Call:
  lm(formula = inv2glu ~ inv12tri, data = glucose)

Residuals:
  Min 1Q Median 3Q Max

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
  (Intercept)  5.424e-05 8.409e-06  6.45   3.25e-10 ***
  inv12tri 5.715e-04 7.982e-05  7.16   3.92e-12 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.557e-05 on 398 degrees of freedom
Multiple R-squared:  0.1141,  Adjusted R-squared:  0.1119
F-statistic: 51.27 on 1 and 398 DF,  p-value: 3.915e-12

Under the transformations considered in this case, there is no summary effect:

```r
> effectInfo(modgluco)
```

The effect of X on Y cannot be summarized with a single number. Its behavior can be explored using effect().
Thus, the following code reproduces results in Table 3 of the paper:

```r
table3 <- effect(modgluco, x1 = 50, npoints = 4, c = 50)
table3

Computing effects...

Adjusted change in the median of the response variable when the explanatory
variable changes from x1 to x2 (confidence intervals based on 999 bootstrap samples):

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>EstimateDiff</th>
<th>lower95%</th>
<th>upper95%</th>
<th>EstimatePercent</th>
<th>lower95%</th>
<th>upper95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>8.703222</td>
<td>6.334019</td>
<td>10.907257</td>
<td>10.114803</td>
<td>7.250154</td>
<td>12.960938</td>
</tr>
<tr>
<td>3</td>
<td>150</td>
<td>3.235200</td>
<td>2.178028</td>
<td>4.439836</td>
<td>3.249826</td>
<td>2.215835</td>
<td>4.370665</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>2.397980</td>
<td>1.602089</td>
<td>3.372809</td>
<td>2.333003</td>
<td>1.597478</td>
<td>3.172078</td>
</tr>
</tbody>
</table>
```

Analogously, the following code reproduces results in Table 4:

```r
x0 <- 50 * 1.5^(0:4)
x0

[1] 50.000 75.000 112.500 168.750 253.125

> table4 <- effect(modgluco, x1 = x0)
table4

Adjusted change in the median of the response variable when the explanatory
variable changes from x1 to x2 (confidence intervals based on 999 bootstrap samples):

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>EstimateDiff</th>
<th>lower95%</th>
<th>upper95%</th>
<th>EstimatePercent</th>
<th>lower95%</th>
<th>upper95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.00</td>
<td>75.00</td>
<td>5.152950</td>
<td>3.813373</td>
<td>6.363338</td>
<td>5.988710</td>
<td>4.361787</td>
<td>7.553217</td>
</tr>
<tr>
<td>75.00</td>
<td>112.50</td>
<td>4.971764</td>
<td>3.542537</td>
<td>6.461374</td>
<td>5.451653</td>
<td>3.853272</td>
<td>7.140190</td>
</tr>
<tr>
<td>112.50</td>
<td>168.75</td>
<td>4.724024</td>
<td>3.260444</td>
<td>6.322945</td>
<td>4.912204</td>
<td>3.396935</td>
<td>6.507858</td>
</tr>
<tr>
<td>168.75</td>
<td>253.125</td>
<td>4.420430</td>
<td>2.976703</td>
<td>6.163127</td>
<td>4.381299</td>
<td>3.006967</td>
<td>5.941818</td>
</tr>
</tbody>
</table>
```

The code

```r
> plot(modgluco, xname = "Triglycerides (mg/dl)", yname = "glucose (mg/dl)")
```

provides eFigure 7, which represents the expected median of the glucose level as a function of the triglycerides level.
J.1.5 Example 5: low birth weight and cord serum cotinine

Revisiting section J.1.2, suppose that now we are interested in the effect of cotinine level on low birth weight (defined as birth weight lower than 2500 g), under a logistic regression model with log transformation in the explanatory variable, cotinine. The model can be fitted as follows:

```r
> modcot2 <- tlm(y = underweight, x = logcotinine, data = cotinine, xpow = 0, family = binomial)
```

where `xpow = 0` indicates that the explanatory variable is already log transformed and the argument `family = binomial` indicates that the regression model is logistic with logit link (default is `family = gaussian`, for the lineal regression model). The fitted model provides the following results:

```r
> summary(modcot2)
```

<table>
<thead>
<tr>
<th>Logistic regression fitted model in the transformed space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformations:</td>
</tr>
<tr>
<td>In the response variable: logit link for logistic regression</td>
</tr>
<tr>
<td>In the explanatory variable: log</td>
</tr>
</tbody>
</table>

**eFigure 7:** Expected median of the glucose levels (and a 95% confidence interval) as a function of the triglycerides level.
Call:
\texttt{glm(formula = underweight \sim \log\text{cotinine}, family = \text{binomial}, data = cotinine)}

Deviance Residuals:
\begin{tabular}{cccc}
Min & 1Q & Median & 3Q & Max \\
-0.7061 & -0.4186 & -0.3231 & -0.2821 & 2.6239 \\
\end{tabular}

Coefficients:
\begin{tabular}{cccc}
Estimate & Std. Error & z value & \text{Pr}(>|z|) \\
(Intercept) & -3.5146 & 0.4539 & -7.744 & 9.65e-15 *** \\
\log\text{cotinine} & 0.3306 & 0.1289 & 2.566 & 0.0103 * \\
\end{tabular}

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 175.09 on 350 degrees of freedom
Residual deviance: 168.45 on 349 degrees of freedom
AIC: 172.45

Number of Fisher Scoring iterations: 5

The function \texttt{effectInfo} provides information on interpreting the relationship between cotinine level (\(X\)) and low birth weight (\(Y\)):

```r
> effectInfo(modcot2)
```

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 \(r = 100 \times (q - 1)\%\) to \(X\))
- Type of effect on \(Y\): odds ratio of \(Y\)
- Effect size: \(q^{\beta}\)

\text{beta coefficient estimate:}
\begin{tabular}{cccc}
Estimate & Std. Error & z value & \text{Pr}(>|z|) \\
\log\text{cotinine} & 0.3306394 & 0.1288549 & 2.565982 & 0.01028842 \\
\end{tabular}

Further details can be obtained using \texttt{effect()}.

The function \texttt{effect} provides as default the odds ratio (OR) of low birth weight for a relative (or percent) change in the cotinine level equal to the interquartile ratio:

```r
> effect(modcot2)
```

Computing effects...

Adjusted odds ratio of the response variable for an \(r\)\% change in the explanatory variable equivalent to the interquartile ratio:

```
r Estimate lower95% upper95%
1 1.112878 2.282194 1.21516 4.286193
```

For further information on interpreting the effect use \texttt{effectInfo()}. 

27
The code

\[
> \text{plot(modcot2, xname = "Cotinine (ng/ml)", yname = "low birth weight")}
\]

provides eFigure 8, which represents the expected probability of low birth weight as a function of the cotinine level.

![Graph](image)

**eFigure 8**: Expected probability of low birth weight (and a 95% confidence interval) as a function of the cotinine level.

### J.2 Stata

In this section we illustrate the usage of existing commands in the Stata 12 software (Stata Corp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP) to obtain the results in the examples of the paper.

#### J.2.1 Example 1: intima media thickness and age

Loading data:

\[
. \text{use "imt.dta", clear}
\]

Calculate the interquartile range of age:
. summ age, det

-------------------------------------------------------------
Percentiles Smallest
1% 35 32
5% 40 32
10% 44 32 Obs 2784
25% 51 32 Sum of Wgt. 2784
50% 59 Largest
75% 66 86 Mean 58.48024
90% 73 86 Variance 118.8117
95% 77 86 Skewness -.0015562
99% 83 86 Kurtosis 2.513045

. di r(p75) - r(p25)
15

Fit the regression model:
. reg logimt age

Source | SS df MS Number of obs = 2784
-------------+------------------------------ F( 1, 2782) = 952.16
Model | 30.11429 1 30.11429 Prob > F = 0.0000
Residual | 87.9871804 2782 .031627311 R-squared = 0.2550
-------------+------------------------------ Adj R-squared = 0.2547
Total | 118.10147 2783 .042436748 Root MSE = .17784

------------------------------------------------------------------------------
logimt | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
age | .0095433 .0003093 30.86 0.000 .0089369 .0101498
_cons | -.8776919 .0183979 -47.71 0.000 -.9137667 -.8416171
------------------------------------------------------------------------------

In this case, the solution is basically an exponentiated coefficient, thus one can obtain a better solution with \texttt{lincom, eform} than with \texttt{nlcom}.

Results for $c = 15$
. lincom 15*age, eform

( 1) 15*age = 0

------------------------------------------------------------------------------
logimt | exp(b) Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
(1) | 1.153903 .0053531 30.86 0.000 1.143454 1.164447
------------------------------------------------------------------------------

29
With \texttt{nlcom}:
\begin{verbatim}
    . nlcom exp(15*_b[age])

    _nl_1:  exp(15*_b[age])
\end{verbatim}

\begin{verbatim}
| logimt | Coef. Std. Err.  t    P>|t|   [95% Conf. Interval] |
|--------|-----------------|-----------------|----------|-----------------------|
| _nl_1  | 1.153903 0.0053531 215.56 0.000 1.143406 1.164399 |
\end{verbatim}

In percent change:
\begin{verbatim}
    . nlcom 100*(exp(15*_b[age]) - 1)

    _nl_1:  100*(exp(15*_b[age]) - 1)
\end{verbatim}

\begin{verbatim}
| logimt | Coef. Std. Err.  t    P>|t|   [95% Conf. Interval] |
|--------|-----------------|-----------------|----------|-----------------------|
| _nl_1  | 15.39029 0.5353102 28.75 0.000 14.34064 16.43993 |
\end{verbatim}

We obtain the same point estimates with \texttt{nlcom}, with slightly different confidence interval. To obtain eFigure 9, first run \texttt{margins} to obtain predictions, then use \texttt{marginsplot}:
\begin{verbatim}
    . qui: margins , expression(exp(predict(xb))) atmeans at(age=(32(1)87))
    . marginsplot, recast(line) recastci(rarea) ciopts(color(*.3)) ytitle("Median of IMT")
\end{verbatim}

\subsection{Example 2: birth weight and cord serum cotinine}

Loading data:
\begin{verbatim}
    . use "cotinine.dta", clear
\end{verbatim}

Calculate the ratio of 3rd to 1st quartile of cotinine:
\begin{verbatim}
    . summ cotinine, det
\end{verbatim}

\begin{verbatim}
  cotinine
  ______________________________________________________
  Percentiles Smallest
  1%  .4545648    .2
  5%  1.046755   .2909473
  10% 1.61569   .3876726   Obs    351
  25% 3.230196   .4545648  Sum of Wgt.  351
  50% 7.385437   Largest
  75% 39.29175  435.7175
  90% 107.3762  531.1402  Variance  7492.561
  95% 180.5182  547.0098  Skewness  5.129338
  99% 435.7175   910  Kurtosis  39.55553
\end{verbatim}
eFigure 9: Expected geometric mean (or median) of IMT (and a 95% confidence interval) as a function of the age.

\[ \text{di } r(p75)/r(p25) \]
12.163892

Calculate log of \( Q_3/Q_1 \):

\[ \text{di } \log(12.163892) \]
2.4984719

Fit the regression model:

\[ \text{reg weight logcotinine} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 351</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5800790.57</td>
<td>1</td>
<td>5800790.57</td>
<td>F(  1,  349) = 28.64</td>
</tr>
<tr>
<td>Residual</td>
<td>70696087.2</td>
<td>349</td>
<td>202567.585</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R-squared = 0.0758</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.0732</td>
</tr>
<tr>
<td>Total</td>
<td>76496877.7</td>
<td>350</td>
<td>218562.508</td>
<td>Root MSE = 450.08</td>
</tr>
</tbody>
</table>
Results for $q = 12.163892$ need $\log(12.163892) = 2.50$:

```stata
lincom logcotinine* 2.50
( 1) 2.5*logcotinine = 0
```

To obtain eFigure 10, we need to obtain the predictions from `margins`, then store them in a new dataset, and generate the graph using that dataset:

```stata
qui: margins, atmeans at(logcotinine = (-1.6(.1)6.8))
matrix A = r(at), r(table)'
clear
svmat A, names(col)
gen cotinine = exp(logcotinine)
twoway (rarea ul ll cotinine, color(*.3)) (line b cotinine, lcolor(navy)), legend(off) /*
*/ ytitle(Mean of weight)
```

### J.2.3 Example 3: cat allergen levels in bed mattress and in living room

#### Loading data:

```stata
use "feld1.dta", clear
```

Calculate the ratio of 3rd to 1st quartile of mattress allergen levels:

```stata
summ mattress, det
```

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
<th>Largest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>.0049772</td>
<td>.5617408</td>
</tr>
<tr>
<td>5%</td>
<td>.011469</td>
<td>.6197836</td>
</tr>
<tr>
<td>10%</td>
<td>.0232505</td>
<td>.288531</td>
</tr>
<tr>
<td>25%</td>
<td>.0629056</td>
<td>.7615518</td>
</tr>
<tr>
<td>50%</td>
<td>.1773452</td>
<td>.3733373</td>
</tr>
<tr>
<td>75%</td>
<td>.5617408</td>
<td>.6197836</td>
</tr>
<tr>
<td>90%</td>
<td>2.288531</td>
<td>76.15518</td>
</tr>
<tr>
<td>95%</td>
<td>5.195854</td>
<td>373.3373</td>
</tr>
<tr>
<td>99%</td>
<td>60.33552</td>
<td>467.7814</td>
</tr>
</tbody>
</table>

Mean: 3.197642

Std. Dev.: 28.20656

Variance: 795.6099

Skewness: 14.4239

Kurtosis: 220.3523
eFigure 10: Expected mean of birth weight (and a 95% confidence interval) as a function of the cotinine level.

\[ \text{di } r(p75)/r(p25) \]

8.929896

Note that Stata calculates the percentiles in a slightly different way than R. In this particular case, the differences between Stata and R are noticeable, so we set \( q = 8.76 \) in order to obtain the same results than with R.

Fit regression model:

```
.reg logroom logmattress cat
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 471</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F( 2, 468) = 168.89</td>
<td>Prob &gt; F = 0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>953.08068</td>
<td>2</td>
<td>476.54034</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>1320.5095</td>
<td>468</td>
<td>2.8216015</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2273.59018</td>
<td>470</td>
<td>4.83742592</td>
<td></td>
</tr>
</tbody>
</table>

| logroom | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|---------|-------|-----------|---|-----|------------------|
|         |       |           |   |     |                  |
Effect of mattress levels for \( q = 8.76 \):  
\[
\text{nlcom } 100 \times (8.76^\_b[\text{logmattress}] - 1)
\]

To obtain eFigure 11, we need to obtain the predictions from margins, then store them in a new dataset, and generate the graph using that dataset:

\[
\text{qui: margins , expression(exp(predict(xb))) atmeans at(logmattress = (-5.8(.1)6.2))}
\]

\[
\text{matrix A= r(at), r(table)'}
\]

\[
\text{clear}
\]

\[
\text{svmat A,names(col)}
\]

\[
\text{gen mattress=exp(logmattress)}
\]

\[
\text{twoway (rarea ul ll mattress, color(*.3)) (line b mattress, lcolor(navy)), legend(off) /*}
\]
\[
\text{*/ ytitle(Median of room)}
\]

J.2.4 Example 4: glucose and triglycerides levels in blood

Loading data:

\[
\text{use "glucose.dta", clear}
\]

First we need to calculate the transformed value for the values of triglycerides where we want to calculate the effects:

\[
\text{di 50}^{-(-.5)}
\]

\[
.14142136
\]

\[
\text{di 100}^{-(-.5)}
\]

\[
.1
\]

\[
\text{di 150}^{-(-.5)}
\]
eFigure 11: Expected geometric mean (or median) of the cat allergen levels in the living room (and a 95% confidence interval) as a function of the cat allergen levels in the bed mattress. Both variables were measured in µg/g.

Fit the regression model:

```
  . reg inv2glu inv12tri
```
Supplementary material for “Models with transformed variables: interpretation and software”

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6.4884e-08</td>
<td>1</td>
<td>6.4884e-08</td>
<td>F( 1, 398) = 51.27</td>
</tr>
<tr>
<td>Residual</td>
<td>5.0368e-07</td>
<td>398</td>
<td>1.2655e-09</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>5.6857e-07</td>
<td>399</td>
<td>1.4250e-09</td>
<td>Adj R-squared = 0.1119</td>
</tr>
</tbody>
</table>

Root MSE = 3.6e-05

| inv2glu | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|---------|-------|-----------|-------|------|----------------------|-------------------|
| inv1tri | 0.0005715 | 0.0000798 | 7.16  | 0.000 | 0.0004146 0.0007285 |
| _cons   | 0.0000542 | 8.41e-06  | 6.45  | 0.000 | 0.0000377 0.0000708 |

Calculate the median response at the values of triglycerides 50, 100, 150, 200 and 250, using the transformed values computed above:

```
. margins , at(inv1tri = (.14142136 .1 .08164966 .07071068 .06324555)) atmeans /*
   */ expression((predict(xb))^(-0.5)) post
```

Adjusted predictions Number of obs = 400
Model VCE : OLS
Expression : (predict(xb))^(-0.5)

<table>
<thead>
<tr>
<th>1._at</th>
<th>inv1tri = 0.1414214</th>
</tr>
</thead>
<tbody>
<tr>
<td>2._at</td>
<td>inv1tri = 0.1</td>
</tr>
<tr>
<td>3._at</td>
<td>inv1tri = 0.0816497</td>
</tr>
<tr>
<td>4._at</td>
<td>inv1tri = 0.0707107</td>
</tr>
<tr>
<td>5._at</td>
<td>inv1tri = 0.0632455</td>
</tr>
</tbody>
</table>

| Delta-method |
|--------------|------------------|---------------|-----------|-------------------|
| Margin Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|---------------|-----|-------|---|-------------------|
| _at 1 | 86.0444 | 1.129963 | 76.15 | 0.000 | 83.82972 88.25909 |
| 2 | 94.74763 | 0.7631345 | 124.16 | 0.000 | 93.25191 96.24334 |
| 3 | 99.54992 | 1.21427 | 81.98 | 0.000 | 97.16999 101.9298 |
| 4 | 102.7851 | 1.699167 | 60.49 | 0.000 | 99.45481 106.1154 |
| 5 | 105.1831 | 2.115352 | 49.72 | 0.000 | 101.0371 109.3291 |

Having used the option post in the last command allows us to calculate combinations of those numbers. For example, we can compute the difference in medians:

```
. lincom _b[2._at]-_b[1._at]
```

36
We can also calculate the percent change in the median:

\[ \text{nlcom} \ 100 \times \left( \frac{\_b[2\_at]}{\_b[1\_at]} \right) - 1 \]

\[
\text{nlcom} \ 100 \times \left( \frac{\_b[2\_at]}{\_b[1\_at]} \right) - 1 \\
\text{nlcom} \ 100 \times \left( \frac{\_b[2\_at]}{\_b[1\_at]} \right) - 1
\]

| Coef.     | Std. Err. | z  | P>|z| | [95% Conf. Interval] |
|-----------|-----------|----|------|----------------------|
| _nl_1_    | 10.1148   | 1.37692 | 7.35 | 0.000               | 7.41609 | 12.81352 |
The results above replicate those in Table 3 of the paper (additive changes in $X$). In order to replicate Table 4 (relative changes in $X$), we need to repeat the same process changing the values of $X$.

Calculate the transformed value for the values of triglycerides where we want to calculate the effects:

```
. di 50^(-.5)
```

1.4142136

```
. di 75^(-.5)
```

1.1547005
Fit the regression model:

```
. reg inv2glu inv12tri
```

```
Source | SS       df       MS              Number of obs = 400
-------------+------------------------------ F( 1, 398) = 51.27
Model | 6.4884e-08 1 6.4884e-08 Prob > F = 0.0000
Residual | 5.0368e-07 398 1.2655e-09 R-squared = 0.1141
-------------+------------------------------ Adj R-squared = 0.1119
Total | 5.6857e-07 399 1.4250e-09 Root MSE = 3.6e-05

------------------------------------------------------------------------------
inv2glu | Coef.  Std. Err.     t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
inv12tri | .0005715  .0000798  7.16  0.000   .0004146   .0007285
_cons     | .0000542  8.41e-06  6.45  0.000   .0000377   .0000708
------------------------------------------------------------------------------
```

Calculate the median response at the values of triglycerides 50, 75, 112.5, 168.8 and 253.1, using the transformed values computed above:

```
. margins , at(inv12tri=(.14142136 .11547005 .0942809 .07696863 .06285704)) atmeans /*
 */ expression((predict(xb))^(-0.5)) post
```

```
Adjusted predictions Number of obs = 400
Model VCE : OLS
Expression : (predict(xb))^(-0.5)
1._at : inv12tri = .1414214
2._at : inv12tri = .1154701
3._at : inv12tri = .0942809
```
4. _at : inv12tri = .0769686
5. _at : inv12tri = .062857

|          | Delta-method | Margin Std. Err. | z  | P>|z| | [95% Conf. Interval] |
|----------|--------------|-----------------|----|------|----------------------|
| _at      |              |                 |    |      |                      |
| 1        | 86.0444      | 1.129963        | 76.15 | 0.000 | 83.82972 88.25909   |
| 2        | 91.19736     | .7734727        | 117.91 | 0.000 | 89.68138 92.71333  |
| 3        | 96.16912     | .849001         | 113.27 | 0.000 | 94.50511 97.83313  |
| 4        | 100.8965     | 1.403734        | 71.88  | 0.000 | 98.14522 103.6478  |
| 5        | 105.3125     | 2.139884        | 49.23  | 0.000 | 101.1202 109.5049  |

Having used the option `post` in the last command allows us to calculate combinations of those numbers. For example, we can compute the difference in medians:

```
. lincom _b[2._at] - _b[1._at]
```

( 1)  - 1bn._at + 2._at = 0

|          | Coef. Std. Err. | z  | P>|z| | [95% Conf. Interval] |
|----------|-----------------|----|------|----------------------|
| (1)      | 5.152952        | .6088722 | 8.46 | 0.000 | 3.959584 6.346319 |

```
. lincom _b[3._at] - _b[2._at]
```

( 1)  - 2._at + 3._at = 0

|          | Coef. Std. Err. | z  | P>|z| | [95% Conf. Interval] |
|----------|-----------------|----|------|----------------------|
| (1)      | 4.971765        | .6966349 | 7.14 | 0.000 | 3.606385 6.337144 |

```
. lincom _b[4._at] - _b[3._at]
```

( 1)  - 3._at + 4._at = 0

|          | Coef. Std. Err. | z  | P>|z| | [95% Conf. Interval] |
|----------|-----------------|----|------|----------------------|
| (1)      | 4.727369        | .7672872 | 6.16 | 0.000 | 3.223514 6.231225 |

40
. lincom _b[5._at] - _b[4._at]

( 1)  - 4._at + 5._at = 0

---------------------------------------------------------------------
| Coef. Std. Err.  z  P>|z|  [95% Conf. Interval]      |
|-----------------+-------------------------------------------------|
| (1)  |  4.416047  .813686  5.43 0.000  2.821252   6.010842 |

We can also calculate the percent change in the median:

. nlcom 100*((_b[2._at]/_b[1._at])-1)

  _nl_1:  100*((_b[2._at]/_b[1._at]) - 1)

---------------------------------------------------------------------
| Coef. Std. Err.  z  P>|z|  [95% Conf. Interval]      |
|-----------------+-------------------------------------------------|
|   _nl_1 |  5.988712  .7692816  7.78 0.000  4.480948   7.496476 |

. nlcom 100*((_b[3._at]/_b[2._at]) - 1)

  _nl_1:  100*((_b[3._at]/_b[2._at]) - 1)

---------------------------------------------------------------------
| Coef. Std. Err.  z  P>|z|  [95% Conf. Interval]      |
|-----------------+-------------------------------------------------|
|   _nl_1 |  5.451654  .7806555  6.98 0.000  3.921598   6.981711 |

. nlcom 100*((_b[4._at]/_b[3._at]) - 1)

  _nl_1:  100*((_b[4._at]/_b[3._at]) - 1)

---------------------------------------------------------------------
| Coef. Std. Err.  z  P>|z|  [95% Conf. Interval]      |
|-----------------+-------------------------------------------------|
|   _nl_1 |  4.915683  .7767377  6.33 0.000  3.393305   6.438061 |

. nlcom 100*((_b[5._at]/_b[4._at]) - 1)
Supplementary material for “Models with transformed variables: interpretation and software”

\[ \text{nl}_1: 100 \times \left( \frac{\text{b[5.}}{\text{b[4.}} - 1 \right) \]

---

| Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|-------|-----------|------|------|---------------------|
| \text{nl}_1 | 4.37681 | 0.7553513 | 5.79 | 0.000 | 2.896348 | 5.857271 |

---

To obtain eFigure 12, we need to obtain the predictions from margins, then store them in a new dataset, and generate the graph using that dataset.

Since we used the option post above, we need to run the regression model again:

\[ . \text{reg inv2glu inv12tri} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6.4884e-08</td>
<td>1</td>
<td>6.4884e-08</td>
<td>F( 1, 398) = 51.27</td>
</tr>
<tr>
<td>Residual</td>
<td>5.0368e-07</td>
<td>398</td>
<td>1.2655e-09</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>5.6857e-07</td>
<td>399</td>
<td>1.4250e-09</td>
<td>R-squared = 0.1141</td>
</tr>
</tbody>
</table>

---

| inv2glu | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|--------|-------|-----------|------|------|---------------------|
| inv12tri | .0005715 | .0000798 | 7.16 | 0.000 | .0004146 | .0007285 |
| \_cons | .0000542 | 8.41e-06 | 6.45 | 0.000 | .0000377 | .0000708 |

---

\[ . \text{qui: margins , at(inv12tri = (0.038(0.001)0.174)) atmeans expression((predict(xb))^(-0.5))} \]

\[ . \text{matrix A = r(at), r(table）} \]

\[ . \text{clear} \]

\[ . \text{svmat A,names(col)} \]

\[ . \text{gen triglycerides= inv12tri^-2} \]

\[ . \text{twoway (rarea ul ll triglycerides, color(*.3)) (line b triglycerides, lcolor(navy)), legend(off) /*} \]
\[ */ \text{ytitle(Median of glucose)} \]

\[ J.2.5 \text{ Example 5: low birth weight and cord serum cotinine} \]

Loading data:

\[ . \text{use "cotinine.dta", clear} \]

Calculate the ratio of 3rd to 1st quartile of cotinine:

\[ . \text{summ cotinine, det} \]
**eFigure 12:** Expected median of the glucose levels (and a 95% confidence interval) as a function of the triglycerides level.

### cotinine

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>.4545648</td>
</tr>
<tr>
<td>5%</td>
<td>1.046755</td>
</tr>
<tr>
<td>10%</td>
<td>1.61569</td>
</tr>
<tr>
<td>25%</td>
<td>3.230196</td>
</tr>
<tr>
<td>50%</td>
<td>7.385437</td>
</tr>
<tr>
<td>75%</td>
<td>39.29175</td>
</tr>
<tr>
<td>90%</td>
<td>107.3762</td>
</tr>
<tr>
<td>95%</td>
<td>180.5182</td>
</tr>
<tr>
<td>99%</td>
<td>435.7175</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Largest</th>
<th>Std. Dev.</th>
<th>Mean</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>.4545648</td>
<td>86.55958</td>
<td>39.43137</td>
<td>7492.561</td>
<td>5.129338</td>
<td>39.55553</td>
</tr>
</tbody>
</table>

\[ \text{di } r(p75)/r(p25) \]

12.163892

Fit the logistic regression model:
. logit underweight logcotinine

Iteration 0: log likelihood = -87.545697
Iteration 1: log likelihood = -84.441139
Iteration 2: log likelihood = -84.224014
Iteration 3: log likelihood = -84.223414

Logistic regression

| Coef. Std. Err.  | z   | P>|z|   | 95% Conf. Interval |
|------------------|-----|-------|------------------|
| underweight      |     |       |                  |
| logcotinine      | 0.3306394 | 0.1288639 | 2.57 | 0.010 | 0.0780707, 0.5832081 |
| _cons            | -3.514616 | 0.4539241 | -7.74 | 0.000 | -4.404291, -2.624941 |

Log likelihood = -84.223414

Pseudo R2 = 0.0379

------------------------------------------------------------------------------

Odds ratio for $q = 12.163892$:

. nlcom 12.163892^_b[logcotinine]

    _nl_1: 12.163892^_b[logcotinine]

------------------------------------------------------------------------------

| Coef. Std. Err.  | z   | P>|z|   | 95% Conf. Interval |
|------------------|-----|-------|------------------|
| _nl_1            | 2.284377 | 0.7354846 | 3.11 | 0.002 | 1.8428535, 3.7259 |

------------------------------------------------------------------------------

References


