SAGE Research Methods Foundations: An Encyclopaedia

Analysis of proportions

Maarten L. Buis

Introduction

The purpose of using proportions is to make observations comparable by standardizing them. For example, we may look at the expenditure of municipalities on law enforcement. In this case it makes sense to compare the proportion of the total budget spend on law enforcement, in order to more meaningfully compare large and small cities. Proportions can enter an analysis as dependent or independent variables. An analysis may involve a single proportion – for example, the proportion of a municipal budget spend on law enforcement – or multiple proportions – for example, the proportions of a municipal budget spend on law enforcement, urban planning, social work, and other.

A single proportion as a dependent variable is hard to analyze using linear regression, as the upper and lower bound of the proportion will result in non-linearity of effects. Moreover, these bounds will typically result in heteroscedasticity. There are two main strategies for modeling proportions. The first strategy uses Maximum Likelihood to model the proportion as a beta distribution, whereby the conditional mean and variance depends on the explanatory variables using link functions, like the logit link function. The second strategy dispenses with the necessity to specify a distribution for the proportion, by maximizing the quasi-likelihood, and only model the conditional mean using a link function.

Multiple proportions that add to one as dependent variables have the additional challenge that these variables are mutually dependent; if you spent an extra minute a day watching television, then that minute cannot be spent on other activities. So the proportions tend to be negatively correlated. Both solutions to single proportion problem can be generalized to the multiple proportion case: The multivariate generalization of the beta distribution is the Dirichlet distribution, and one can use quasi-likelihood with a multinomial logit link function to let the conditional means depend on explanatory variables. Dirichlet regression implies the negative

correlation between the dependent variables under independence, while quasi-likelihood models tread the correlation as a nuisance parameter. So both classes of models may work well when one is interested in the influence of other explanatory variables. For example, how does the political orientation of the city council influence the proportions of the budget spent on various categories. However, these two classes of models are less suited when one is interested in the correlation between the proportions. For example, can a city spend a smaller part of his budget on policing if a city spends a larger part of his budget on social projects. A common strategy for such problems is to use a multivariate normal distribution on a transformation of the proportions.

The mutual dependence of proportions also poses a challenge when proportions are added as explanatory variables. Effects of explanatory variables are often interpreted as the expected change in the explained variable for a unit change in the explanatory variable while keeping all other variables constant. This latter part is logically impossible when adding multiple proportions as explanatory variables.

Proportions

The purpose of a proportion is to make observations comparable, even when there are wild differences in their base. For example, ten employees participating in a firm's pension plan means something very different when the firm has ten employees or a thousand employees. To make these firms comparable the number of employees participating is divided by the total number of employees to create a proportion. Such a proportion has a logical lower bound of zero, as it is impossible that less than nobody participates in the pension plan. Similarly, there is the logical upper bound of one, as it is impossible that more than everybody participates. Moreover, for a set of mutually exclusive and exhaustive categories – for example, the firm's pension plan, any other pension plan, no pension plan – the proportions in each of these categories have to add to one.

Sometimes variables are explicitly defined as proportions, but sometimes that is more implicit. For example, consider the following survey question from the General Social Survey: "On the average day, about how many hours do you personally watch television?". The variable that results from such a survey question will have many of the same properties and challenges as a proportion. It would have a lower bound (zero hours per day) and an upper bound (twenty-four hours per day). If we have a set of mutually exclusive and exhaustive activities, then the number of hours per day spent on these activities would have to add up to twenty-four.

Proportions often happen at an aggregate level, for example the proportion of non-western immigrants in a neighborhood. When analyzing these proportions, one has to be careful to remain on that aggregate level. If one finds a positive association between the number of cheap houses in a neighborhood and the proportion of non-western immigrants, then it is very tempting to conclude that non-western immigrants are attracted to cheap houses, that is, make a statement on the individual level rather than the neighborhood level. This would be an ecological fallacy. (Robinson, 1950) Association on the individual level can be very different from the association on the aggregate level, even with a different sign. If one is interested in the lower level effects, then the best solution is to get lower level data. If that is not an option, then one can look into ecological inference (King, 1997; King, Tanner, & Rosen, 2004).

Explaining a single proportion

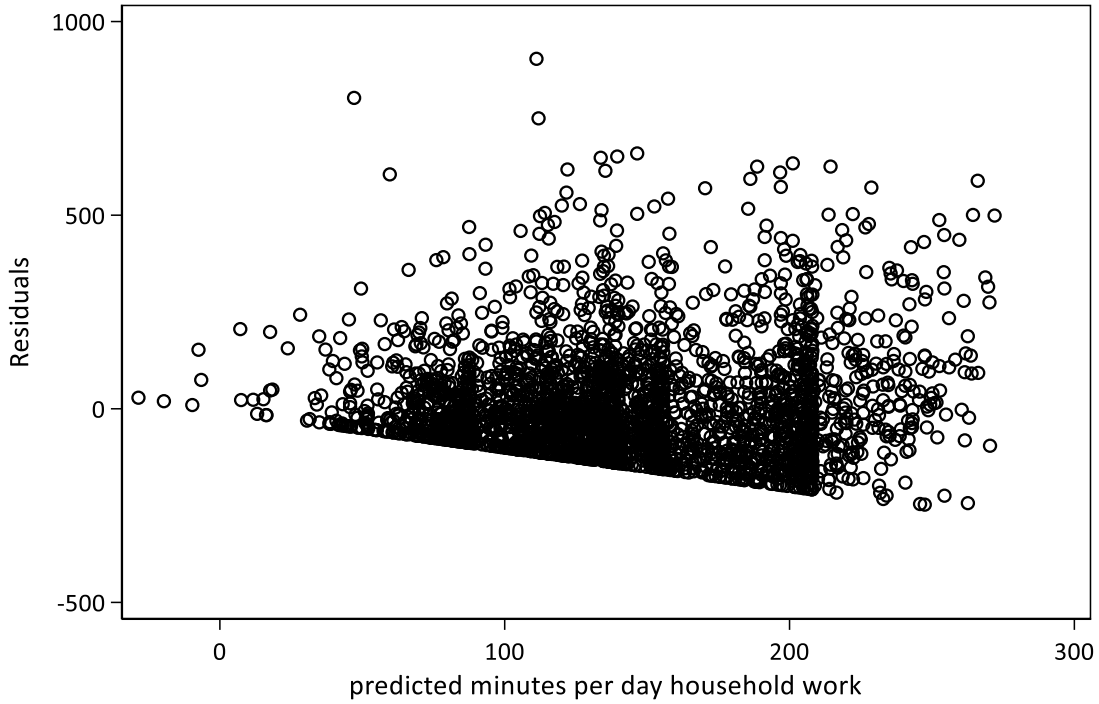
A proportion is a continuous variable. So it makes sense to start with what for many researchers is the default method for explaining continuous variables: linear regression. However, in that case two problems are likely to occur. First, the lower and upper bound are likely to impose a pattern on the residuals, leading to heteroscedasticity. Second, linear regression does not enforce the lower and upper bound, making it possible to get impossible predictions, that is, predicted proportions less than zero or more than one.

These problems are illustrated using an example analysis of 2017 American Time Use Survey (U.S. Bureau of Labor Statistics 2017). Respondents are asked to report the activities they performed in the previous day. The sample is restricted to respondents who work and were asked to report on a non-holiday weekday. In this example analysis we are trying to explain the minutes spent on unpaid household work (cleaning, cooking, child care, care for adults in the household, etc.). This variable has a lower bound of zero, and upper bound of 24×60=1,440 minutes, and when one adds the times spent on all the activities, then that adds up to 1,440 for each respondent. So even though it is not explicitly called a proportion, it has all the main characteristics of a proportion. This variable is explained with a linear regression model with the variables usual hours per week doing paid work, the respondents sex and marital status, the work status of the partner, age and age squared. These results are not shown, but instead a plot of the residuals versus the fitted values is shown in Figure 1, as this is a clear illustration of the problems that often occur when using linear regression on a proportion.

Ideally, there should be no pattern in this graph. This is clearly not the case: There is a clear downward sloping lower bound visible. This can be explained by the fact that no observations can have a negative value on the number of minutes spent on household work, so if one predicts 10 minutes spent on household work, then the residual cannot be less than -10. Figure 1 also shows heteroscedasticity; the variance of the residuals is much larger at higher predicted values.

If the predicted values would have started to approach the maximum, the variance would have decreased again. This heteroscedasticity is a consequence of these lower and upper bounds. The second problem can be seen by looking at the predicted values: some of them are negative, which is impossible.

Analysis of proportions, Figure 1 Residual versus predicted value plot for a linear regression



Source: Created by author for this entry using data from (U.S. Bureau of Labor Statistics 2017).

Maximum Likelihood

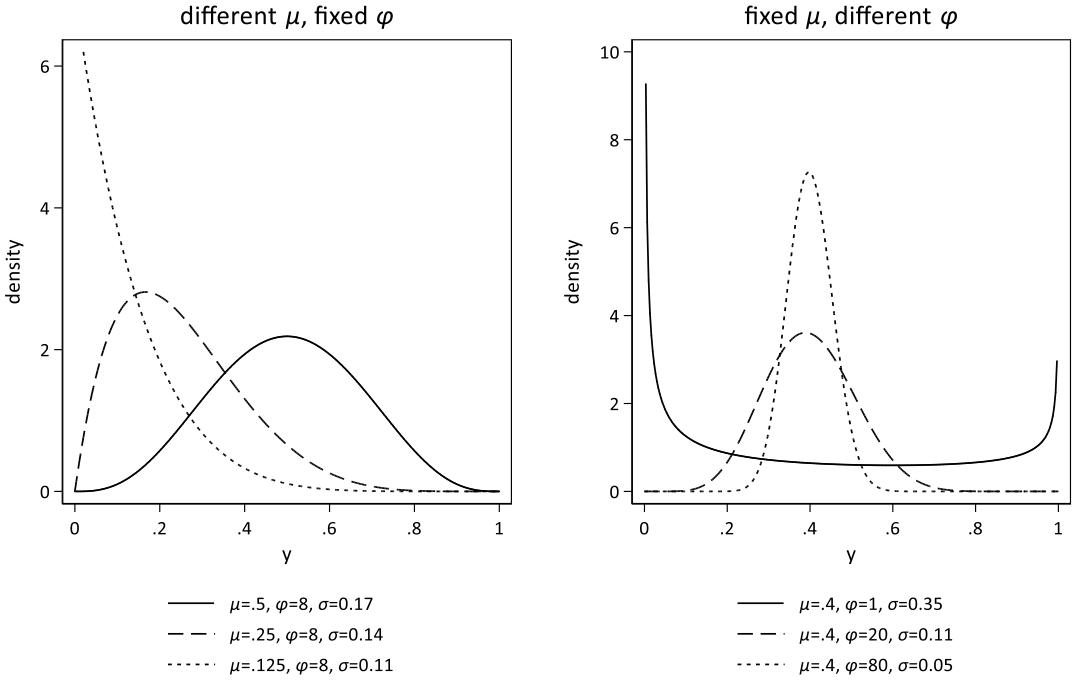
One common way of analyzing proportions is to do a regression analysis while assuming that the proportion follows a beta distribution. The beta distribution is a fairly flexible distribution for variables bounded between zero and one. It can be parameterized in terms of a location parameter μ and a scale parameter φ. The probability density function is:

$$f(y|\mu,\varphi) = \frac{\Gamma(\varphi)}{\Gamma(\mu\varphi)\Gamma\big((1-\mu)\varphi\big)} y^{\mu\varphi-1}(1-y)^{(1-\mu)\varphi-1}$$

The $\Gamma(\cdot)$ is the gamma function, which can be thought of as a generalization of the factorial function for non-integers. The mean of this distribution is the parameter μ, and must thus remain between zero and one. The parameter φ is one of the determinants of the variance, and must be

larger than 0. The variance of the beta distribution is $\mu(1-\mu)\frac{1}{1+\varphi}$ and consequently the standard deviation is $\sqrt{\mu(1-\mu)\frac{1}{1+\varphi}}$. This means that the variance is both a function of the predicted mean and the parameter φ. The variance decreases as φ increases. Also, the variance increases as the predicted mean gets closer to .5 and decreases as mean approaches either the upper or lower bound. The different shapes possible with a beta distribution are illustrated in Figure 2. This figure also shows how the standard deviation (σ) depends on both the μ and the φ.

Analysis of Proportions, Figure 2 Various shapes for the beta distribution



Source: Created by author for this entry.

A regression type model is created by replacing the parameter μ by a function of the explanatory variables and their effects like in linear regression, the so called linear predictor. To make sure that the predicted means respect the lower and upper bound, the linear predictor is usually transformed with a link function. (Ferrari & Cribari-Neto, 2004; Kieschnick & McCullough, 2003; Paolino, 2001; Smithson & Verkuilen, 2006). Common choices for the link function are the logit and the probit, but others are possible as well (Simas, Barreto-Souza, & Rocha, 2010).

In this model the variance also changes when the explanatory variables change, as the

variance is a function of that mean. The variance will decline as the mean approaches either the upper or lower bound. This is in line with the pattern of heteroscedasticity commonly found in proportions. One can further improve the model for the heteroscedasticity by also replacing φ with a function of explanatory variables (Smithson & Verkuilen, 2006), typically using a log link function as φ must be larger than 0. This is often a advisable because if one does not do that the model for μ tries to do two things: model how the conditional mean depends on the explanatory variables and how the conditional variance depends on the explanatory variables. If there is some heteroscedasticity that is not captured by the variance function of the beta distribution, then that will introduce bias in the estimates. (Meaney & Moineddin, 2014)

A more formal representation of this model is given in equations (1) till (3). Equation (1) specifies that some dependent variable y_i, which is bounded between 0 and 1, follows a beta distribution. However, the subscript *i* for the parameters indicate that the parameters differ from observation to observation. Equation (2) determines how the parameter μ_i depends the explanatory variables x_i, and equation (3) determines how the parameter φ_i depends on the explanatory variables z_i. The functions $f(\cdot)$ and $g(\cdot)$ are link functions. The argument of the function $f(\cdot)$ is bounded between zero and one and the argument of the function $g(\cdot)$ is bounded to be larger than zero. Both functions will turn its argument into a number that can take any value on the real number line. A typical example of $f(\cdot)$ would be the logit link function $f(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$, while a typical example of $g(\cdot)$ would be the log link function $g(\varphi_i) = \ln(\varphi_i)$. Common link functions for the mean are given in Table 1. The sets of explanatory variables x_i and z_i, could be the same variables, partially the same, or completely different.

$$y_i \sim Beta(\mu_i, \varphi_i) \tag{1}$$

$$f(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} = x_i \beta \tag{2}$$

$$g(\varphi_i) = \theta_0 + \theta_1 z_{1i} + \theta_2 z_{2i} + \cdots + \theta_k z_{ki} = z_i \theta \tag{3}$$

One of the nice properties of the logit link function is that the exponentiated coefficients can be interpreted directly. Continuing the example, an exponentiated coefficients shows the factor by which the relative proportion of time spent on doing household work changes for a unit change in the explanatory variable. The relative proportion is the proportion of time spent on household work divided by one minus the proportion of time spent on household work. In other words, the proportion of time spent on household work divided by the proportion of time spent on other things. As the total is the same for both the numerator and the denominator, it drops out, and one can also interpret it as the expected number of minutes doing household work for every minute doing something else.

Regardless of the link function one can interpret the model by computing marginal effects

after estimating the model. A marginal effect tells us by how much the mean changes for a unit change in an explanatory variable, that is, it is a slope or first derivative. The formula for the marginal effect depends on the link function. To get the marginal effect ($\partial\mu/\partial x$), one can use the chain rule: $\partial\mu/\partial x = \partial\mu/\partial x\beta \times \partial x\beta/\partial x$. The first part is given in the last column of Table 1. If the explanatory variable is added linearly, then the latter part is just β. One complication is that the marginal effect depends on the values on all explanatory variables, so each observation in the data will have its own marginal effect. In order to still be able to report "the" marginal effect as one number one typically computes the marginal effect for each observation in the data and report the mean of these marginal effect. This is typically referred to as the average marginal effect. Another complication is that the slope at one point on the regression line is not the best description of the effect of a categorical variable. Instead one typically reports discrete differences. Say the categorical explanatory variable is the respondent's biological sex, then one would compute for each observation the predicted proportions assuming that person was male, and the predicted proportion assuming that person was female. The difference between these two is the effect of being female, which differs from person to person depending on the other explanatory variables. So to report one effect, one reports the average of these effects.

Table 1 Common link functions for the mean in beta regression, the corresponding inverse link function, and the first dervative of the inverse link function with respect $x_i\beta$

Link function name	Link function $g(\mu_i) = x_i\beta$	Inverse link function $\mu_i = g^{-1}(x_i\beta)$	First derivative $\partial\mu_i/\partial x_i\beta$
Logit	$\ln\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{e^{x_i\beta}}{1+e^{x_i\beta}}$	$\mu_i(1-\mu_i)$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(x_i\beta)$	$\phi(x_i\beta)$
Log-log	$-\ln(-\ln(\mu_i))$	$\exp(-\exp(-x_i\beta))$	$-\mu_i\ln(\mu_i)$
Complementary log-log	$\ln(-\ln(1-\mu_i))$	$1-\exp(-\exp(x_i\beta))$	$(\mu_i-1)\ln(1-\mu_i)$

$\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function or quantile function of the standard normal distribution, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and $\phi(\cdot)$ is the probability density function of the standard normal distribution.

Just as one can look at the marginal effect on the average proportion, one can also look at the marginal effect on the standard deviation. The standard deviation of the beta distribution

was $\sqrt{\mu_i(1-\mu_i)\frac{1}{1+\varphi_i}}$, so to get the effect of an explanatory variable on the standard deviation one computes the first derivative of that function with respect to that explanatory variable. In general, the marginal effect on the standard deviation is then: $\frac{1}{2sd_i(1+\varphi_i)}\left[(1-2\mu_i)\frac{\partial\mu_i}{\partial x}+\frac{\mu_i(1-\mu_i)}{1+\varphi_i}\frac{\partial\varphi_i}{\partial x}\right]$, where $\frac{\partial\mu_i}{\partial x}$ and $\frac{\partial\varphi_i}{\partial x}$ depend on the link functions chosen for the mean and the scale parameter. For example, if we use the logit link function for the mean and the log link function for the scale parameter, then the marginal effect of an explanatory variable on the standard deviation is $\frac{var_i}{2sd_i}\left[(1-2\mu_i)\beta-\frac{\theta}{(1+\varphi_i)z_i\theta}\right]$, where var_i is the predicted variance for observation *i*, and sd_i is the square root of that predicted variance. However, most statistical packages have dedicated commands for computing marginal effects, like the `margins` command in Stata, the `margins` macro in SAS, or the `margins()` library in R. So, usually one uses those commands rather than deriving these formulas and filling them out by hand.

Table 2 shows an example analysis of the time spent on household work to illustrate how to interpret the results from such a model. The dependent variable is the proportion of time in a day spent on doing household work. It is the number of minutes spent doing household work divided by 1,440 (the number of minutes in a day). The model uses the logit link for the mean function and the log link for the scale function. This means that the exponentiated parameters can be interpreted directly. For the mean equation The constant is the relative proportion for the group that has the value zero on all explanatory variables. The usual hours per week worked was centered at 40, so the constant refers to men who work 40 hours per week. This group is expected to have spent 0.08 minutes on household work for every minute they spent on other things. Working an hour per week longer decreases the minutes spent on household work per minute spent on other things by a factor of 0.99 or (0.99-1)×100%= - 1%. Being a female increases the relative proportion 71%.

The third column in Table 2 shows the average marginal effects on the time spent doing household work. In this case it makes sense to multiply the marginal effects by 1,440 to get the marginal effects in terms of minutes per day. So working an hour per week more reduces the time spent on household work by 1.3 minutes per day, and women tend to do 67 minutes per day more household work than men.

The exponentiated parameters for the φ equation represent the factor by which this parameter changes for a unit change in the explanatory variable. So for a men working 40 hours a week predicts a φ of 6, and this φ increases by 1% for every hour per week a person works longer. A larger φ means a lower standard deviation, so the group of people working long hours are more homogenous (smaller standard deviation) than the group of people working short hours. The φ also increases by 1% if someone is a female, but this change is far from significant. The marginal effect on the standard deviation (σ) is shown in the last column. The standard

deviation decreases by one minute for every hour per week a person works longer. This is due to two effects: First, the negative effect of working hours on the mean proportion spent on household work, means that people working long hours will be closer to the lower bound, and this will reduce the standard deviation. Second the positive effect of working hours on φ will reduce that variance even more for people with longer working hours. The standard deviation is substantially bigger for women than for men, but this difference in standard deviation is almost entirely driven by the difference in mean.

Table 2 Results form a beta regression explaining the proportion of a day spent on household work

		$\exp(\beta)$		Marginal effect[a] (in minutes per day)	
		μ	φ	μ	σ
Usual hours per week work		0.99*** (0.00)	1.01* (0.01)	-1.30*** (0.00)	-1.12*** (0.00)
Respondent's Gender	male	(reference)	(reference)	(reference)	(reference)
	female	1.71*** (0.00)	1.01 (0.88)	67.18*** (0.00)	33.47*** (0.00)
Constant		0.08*** (0.00)	6.10*** (0.00)		
Observations		2784			

Source: Created by author for this entry using data from (U.S. Bureau of Labor Statistics 2017).

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[a] discrete changes are used for categorical variables

A key issue with beta regression is that it is only defined for values *between but not including* zero and one. So this model cannot handle proportions of exactly zero or one. In the example above, there are a substantial number of respondents that did not do any household work the previous day. The example actually used a transformed proportion suggested by Smithson and Verkuilen (2006) to "nudge" those zeros a bit in. If y is the original proportions with zeros and/or ones and N is the total number of observations, then the transformed proportion is $\frac{y(N-1)+0.5}{N}$.

An alternative way of dealing with exact zeros and ones is to estimate a zero one inflated beta regression (Cook, Kieschnick, & McCullough, 2008; Ospina & Ferrari, 2010, 2012). This will

estimate the probabilities of having the value zero and/or one as separate processes. The logic is that we can often think of proportions of zeros or ones as being qualitatively different and generated through a different process as the other proportions. For example, we can think of doing no household work at all is something very different from doing a tiny bit of household work.

The zero one inflated beta distribution consists of three parts:

- a probability that the dependent variable is zero, $Pr(y = 0)$. This probability could be made dependent on explanatory variables using any model for a binary dependent variable. For example, it can be modeled as a logistic regression.
- a probability that the dependent variable is one, $Pr(y = 1)$. This probability can be modeled in the same way as the $\Pr(\mathrm{y} = 0)$.
- the distribution of the dependent variable given that it is between zero and one, which is modeled as a beta distribution as in equations (1) till (3).

This means that $y_i$is assumed to be distributed as:

$$f(y_i) = \Pr(y_i = 0) + \Pr(y_i = 1) + (1 - \Pr(y_i = 0)) \times (1 - \Pr(y_i = 1)) \times Beta(\mu_i, \varphi_i)$$

If there are no exact ones in the data, then this model simplifies to a zero inflated beta:

$$f(y_i) = \Pr(y_i = 0) + (1 - \Pr(y_i = 0)) \times Beta(\mu_i, \varphi_i)$$

Similarly, if there are no exact zeros in the data, the model simplifies to a one inflated beta:

$$f(y_i) = \Pr(y_i = 1) + (1 - \Pr(y_i = 1)) \times Beta(\mu_i, \varphi_i)$$

Continuing the example started above, Table 3 shows the results of modeling the proportion of the day spent on household work as a zero inflated beta. The beta part is modeled as before, that is, a logit link function for μ_i and a log link function for φ_i. The probability of a zero, that is, spending no time at all on household work, is also modeled with a logit link function. There are now three sets of parameters: one for the mean, one for the probability of choosing zero, and one for the scale parameter φ_i. When exponentiated all three sets of parameters can be interpreted directly. Starting with the μ equation: The constant means that a male who works 40 hours per week and does some household work is expected to spend 0.10 minutes on household work for every minute it spends on something else. This ratio decreases by 1% for every hour the person works longer, and increases by 47% if the respondent is female. The second column of Table 3 gives the odds and odds ratios for doing no household work at all. The odds of doing no household work for men that work 40 hours per week is 0.24 (the constant), that is, we expect in that group to find 0.24 persons who have done no household work for every person that has done some household work. This odds increases (a non-significant) 1% for every

hour one works longer and decreases by 68% if the respondent is a female. The φ parameter for men who work 40 hours per week is 11 and increases by 1% for every hours someone works longer and decreases a non-significant 16% when the respondent is female.

Table 3 Exponentiated parameters for a zero-inflated beta model on the proportion of time spent on household work

		μ	$\Pr(y = 0)$	φ
Usual hours per week work		0.99***	1.01	1.01**
		(0.00)	(0.20)	(0.01)
Respondent's Gender	male	1.00	1.00	1.00
		(.)	(.)	(.)
	female	1.47***	0.32***	0.84
		(0.00)	(0.00)	(0.06)
Constant		0.10***	0.24***	11.48***
		(0.00)	(0.00)	(0.00)
Observations		2784		

Source: Created by author for this entry using data from (U.S. Bureau of Labor Statistics 2017).

Exponentiated coefficients; *p*-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Instead of the coefficients one can also look at the marginal effects. Now different marginal effects can be of interest. Table 4 shows three of these: the effect on the overall mean (including zero), the effect on the probability of doing no household work, and the effect on the mean given that one does at least some household work. These are shown in Table 4. Working an hour longer reduces the average time spend on household work by 1.3 minutes, has no noticeable effect on the probability of doing no household work, and the average time spend on household work given that one does some household work also increases by 1.3 minutes. Overall females spend 68 minutes more on household work than men, they are 12 percentage points less likely to do no household work and if they do at least some household work then they spend 55 minutes more than men.

Table 4 Marginal effects (the means are in minutes per day) for zero inflated beta model

		Overall mean	Pr(y=0)	Mean if y > 0
Usual hours per week work		-1.28***	0.00	-1.31***
		(0.00)	(0.20)	(0.00)
Respondent's Gender	male	(reference)	(reference)	(reference)
	female	68.04***	-0.12***	55.49***
		(0.00)	(0.00)	(0.00)
Observations		2784		

Source: Created by author for this entry using data from (U.S. Bureau of Labor Statistics 2017).

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Quasi-likelihood

A major alternative to models based on the beta distribution are models estimated by maximizing a quasi-likelihood function. Quasi-likelihood models for proportions have been around for a long time (Wedderburn, 1974), but have popularized more recently by Papke and Wooldridge (1996). For creating a quasi-likelihood function, we need a so-called mean function and a variance function. The mean function relates the predicted mean to the explanatory variables. Any inverse link function that can be used for the beta distribution can also be used as a mean function for a quasi-likelihood function. The variance function captures how the variance depends on the predicted mean. These two functions are combine to form the quasi-likelihood function. The strength of these models is that they don't require a correct specification of the entire distribution, but only require a correct specification of the mean function. Even the variance function does not have to be correct.

The most common quasi-likelihood model for fractional data is the fractional logit model. It uses the logit inverse link function (see Table 1) for the mean function and the variance function is $var(\mu_i) = \mu_i(1 - \mu_i)$. A convenient feature of this model is that it can be estimated using regular logistic regression models with robust standard errors (Papke & Wooldridge, 1996). Fractional logit models have two advantages over the models based on the beta distribution.

First, it is more robust than the beta distribution. In a beta regression an error in the specification of the φ part, will also result in a bias of the estimates in the μ part of the model. For a fractional logit only a correctly specified conditional mean is necessary. (Meaney & Moineddin, 2014; Papke & Wooldridge, 1996) However, the fractional logit only models the conditional mean. As a consequence, it cannot answer questions about other characteristics, like

are men more homogenous (smaller variance) than women when it comes to the amount of time spent on household work. If that is the question of interest, then a model for the entire distribution, like a beta regression model, is necessary. (Smithson & Verkuilen, 2006)

Second, a fractional logit model can include observations with exact zeros or ones. The predicted mean cannot become zero or one, but the observations can. The logic used to include such observations is very different from the logic used in a zero one inflated beta. In a fractional logit model, a zero proportion is just an indication of a small conditional mean. Applied to the example we have used thus far: If one usually spends a very small fraction of the day on household work, then it can happen that on a day one spends no time on household work. So the difference between a zero and a small proportion is assumed to be gradual rather than a completely different process.

Table 5 shows the results of a fractional logit model on the proportion of the day spent on household work, that is, the explanatory variables determine the conditional mean via the logit link function from Table 1. The results are very similar to the beta regression model from Table 2. A male working 40 hours is expected to spend 0.08 minutes on household work for every minute spend on other things, this ratio increases by 1% if the respondent works an hour longer, and increases by 75% when the respondent is female. This corresponds to 1.3 minutes less time for household work for every hour the respond work longer, and 71 minutes more household work per day for women compared to men.

Table 5 Exponentiated coefficients and marginal effects (in terms of minutes per day) of a fractional logit model

		$\exp(\beta)$	Marginal effect
Usual hours per week work		0.99^{***}	-1.34^{***}
		(0.00)	(0.00)
Respondent's Gender	male	(reference)	(reference)
	female	1.75^{***}	70.65^{***}
		(0.00)	(0.00)
Constant		0.08^{***}	
		(0.00)	
	Observations	2784	

Source: Created by author for this entry using data from (U.S. Bureau of Labor Statistics 2017).

p-values in parentheses

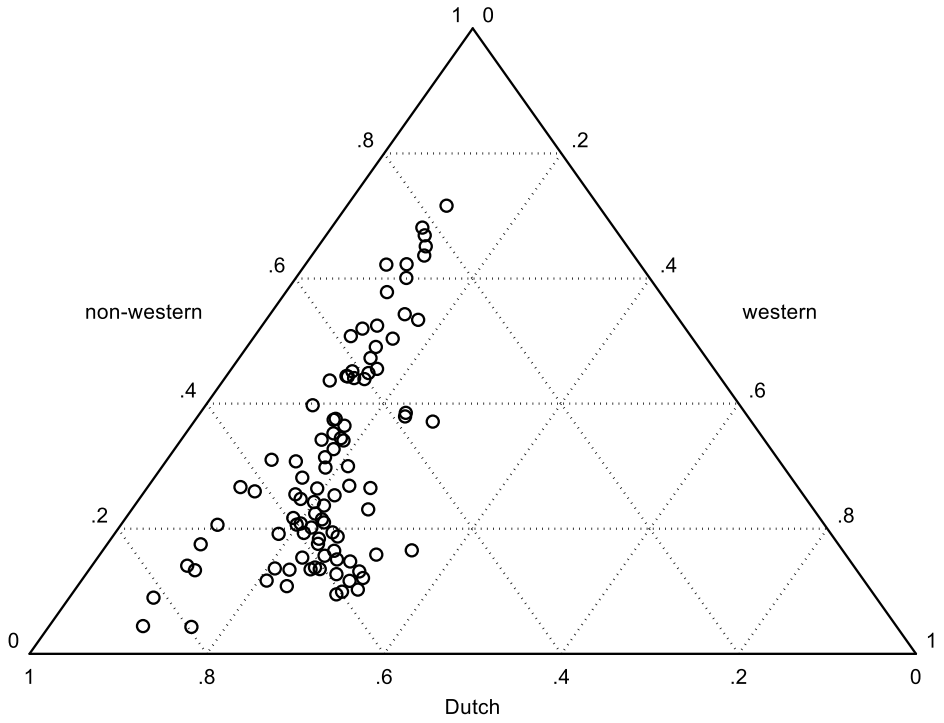
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Explaining multiple proportions

The model becomes more complex when there are multiple proportions that all add up to one. For example, the composition of neighborhoods in terms of the immigration background of the residents. Before one starts modeling, it is a good idea to first look at the distribution in the data. The challenge is that more variables means a higher dimensional graph. However, one of the proportions is redundant: If we have three proportions adding up to one, and we know two of those, then we also know the third. So three proportions can be plotted in two dimensions. This gives rise to a common graph for plotting three proportions: the triangular plot. An example is Figure 3, which shows the compositions of different neighborhoods in the city of Amsterdam in the Netherlands. There are three sets of grid lines. To find which gridline belongs to which axis, one looks at the axis that crosses it at the value zero. The grid lines parallel to this crossing axis are the relevant grid lines. So if one is looking for the proportion of western non-Dutch migrants in a neighborhood, then the axis crossing it at zero is the upward sloping axis (non-western immigrant), so the upward sloping grid lines are the relevant gridlines for western non-Dutch migrants. The horizontal gridlines are relevant to the proportion of non-western immigrants, and the downward sloping gridlines for the proportion of non-immigrants (Dutch). It shows that the proportion of western non-Dutch immigrants remains between about 10% and a bit less than 40%, the proportion of Dutch residents remains between a bit less than 20% and a bit more than 80%, and the proportion of non-western immigrants between less than 10% and a bit less than

80%.

Analysis of proportions, Figure 3 Proportions of residents with non-western, western but non-Dutch, and Dutch backgrounds in different neighborhoods in Amsterdam.



Source: Created by author for this entry using data from (OIS Amsterdam 2017).

Maximum Likelihood and Maximum Quasi-likelihood

Because the proportions have to add to one, the different proportions are related: If the share of non-western immigrant increase in a neighborhood, then at least one of the other shares has to decrease. A model for the entire distribution of these proportions should allow for that correlation. Moreover, this interrelationship between proportions means that an explanatory variable influencing one proportion, automatically also influences the remaining proportions as well, and a model should take that into account. One candidate for such a model is the Dirichlet distribution, which is a multivariate generalization of the Beta distribution. For three proportions (y_1, y_2, and y_3, where $y_1 = 1 - y_2 - y_3$) the probability density function is shown below. It has two (in general: the number of proportions minus one) location parameters μ and one scale

parameter φ.

$$f(y_1, y_3, y_3|\mu, \varphi) = \frac{\Gamma(\varphi)}{\Gamma(\mu_2\varphi)\Gamma(\mu_3\varphi)\Gamma\big((1-\mu_2-\mu_3)\varphi\big)} {y_1}^{(1-\mu_2-\mu_3)\varphi-1} {y_2}^{\mu_2\varphi-1} {y_3}^{\mu_3\varphi-1}$$

The means, variances and covariances are shown below. It shows that the means add up to one, the variances are very similar to the variances of the beta distribution, and the covariances are all negative.

$$mean(y_k) = \begin{cases} \mu_k & \text{if } k>1 \\ 1 - \sum_{k=2}^{K} \mu_k & \text{if } k = 1 \end{cases}$$

$$var(y_k) = \mu_k(1-\mu_k)\frac{1}{1+\varphi}$$

$$cov\big(y_i, y_j\big) = -\mu_i\mu_j\frac{1}{1+\varphi}$$

The link function that has mainly been used to include explanatory variables is the multinomial link function for the means and the log link function for the scale parameter (φ). The inverse multinomial link function is:

$$\mu_k = \begin{cases} \frac{e^{x_i\beta_k}}{1+\sum_{j=2}^{K} e^{x_i\beta_j}} & \text{if } k>1 \\ \frac{1}{1+\sum_{j=2}^{K} e^{x_i\beta_j}} & \text{if } k = 1 \end{cases}$$

The Multinomial logit link function can be used to create a quasi-likelihood model for multiple proportions. It can be estimated by maximizing the log likelihood function of a regular multinomial logit using the proportions as dependent variables and using robust standard errors. (Mullahy, 2015) The strengths and weaknesses of the multinomial fractional logit compared to the Dirichlet regression are analogous to the strength and weaknesses of the fractional logit compared to beta regression: The fractional multinomial logit is a more robust than Dirichlet regression because the fractional multinomial logit does not try to model the variances and covariances of the proportions. This also means that the fractional multinomial logit can only be used to answer questions concerning the conditional means and not about other properties of the distribution like the variances. Moreover, the fractional multinomial logit can include fractions of exactly zero or one, while Dirichlet regression cannot.

To illustrate the use of these models Table 6 shows the results from a Dirichlet and fractional multinomial logit model of the composition of Amsterdam's neighborhoods. They both use the multinomial logit link function for the means and the Dirichlet model used the log link

function for the scale parameter. The explanatory variable is the average price of a house in thousands of euros per square meter, centered at 4000 euro per square meter (approximately the median price). So the constant refers to a neighborhood with median priced houses. The exponentiated coefficients are still interpretable in terms of relative proportion ratios, but now they are relative to the baseline proportion; in this case the non-western migrants. So, the constant in the western equation means that in such a neighborhood one expects to find 0.8 (Dirichlet) or 0.7 (fractional multinomial logit) western migrants for every non-western migrant. This ratio increases by 74% (Dirichlet) or 82% (fractional multinomial logit) if the average price increases by a 1000 euros per square meter. Similarly, one expects to find 2 (Dirichlet) or 1.9 (fractional multinomial logit) Dutch residents for every non-western resident. This ratio increases by 55% (Dirichlet) or 63% (fractional multinomial logit) if the average price increases by a 1000 euros per square meter.

The marginal effects in Table 6 show that the percentage of non-western migrants decrease by 9 (Dirichlet) or 10 (fractional multinomial logit) percentage points if the average price increases by a 1000 euros per square meter. The share of Western migrants increases by 4 percentage points if the average price increases by a 1000 euros per square meter, and the share of Dutch residents increases by 5 (Dirichlet) or 6 (fractional multinomial logit) for a thousand-euro increase in average price. Notice that the marginal effects add up to 0. This way the predicted proportions will add up to one.

Table 6 Exponentiated coefficients and marginal effects of a Dirichlet and multinomial fractional logit model

		Exp(beta)		Marginal effect	
		Dirichlet	Fractional multinomial logit	Dirchlet	Fractional multinomial logit
Non-western	Price			-0.09***	-0.10***
				(0.00)	(0.00)
	Constant				
Western	Price	1.74***	1.82***	0.04***	0.04***
		(0.00)	(0.00)	(0.00)	(0.00)
	Constant	0.79***	0.72***		
		(0.00)	(0.00)		
Dutch	Price	1.55***	1.63***	0.05***	0.06***
		(0.00)	(0.00)	(0.00)	(0.00)
	Constant	2.03***	1.94***		
		(0.00)	(0.00)		
φ	Price	2.05***			
		(0.00)			
	Constant	30.99***			
		(0.00)			
	N	95	95	95	95

Source: Created by author for this entry using data from (OIS Amsterdam 2017).
p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Modeling the inter-relationship between proportions

The Dirichlet and fractional multinomial logit models are good for modeling the impact of explanatory variables on the proportions, but they are not suitable for investigating the interrelationship between proportions. The fractional multinomial logit model ignores that completely by designating that covariance as nuisance parameters. The covariance structure in a Dirichlet distribution is best thought of as a structure that is as close as possible to independence. Normally independence implies a covariance of zero, but with proportions all covariances cannot be zero, instead they would tend to be negative. (Aitchison, 2003 [1986]) More precisely, the covariance present in a Dirichlet distribution captures independence in the following sense: If for each observation we draw a value from *k* independent gamma distributions, and transform those draws to proportions, then those proportions would follow a Dirichlet distribution. So the covariance structure implicit in the Dirichlet distribution is useful as a null-hypothesis of independence, but not for studying associations between proportions that deviate from

independence.

One solution that has been proposed by Aitchison (2003 [1986]) is to transform the proportions analogous to a multinomial link function, and assume that those transformed variables have a multivariate normal distribution. If we have three proportions y_1, y_2, and y_3, where $y_1 = 1 - y_2 - y_3$, then the transformation will create two new variables z_2 and z_3, such that $z_2 = \ln\left(\frac{y_2}{y_1}\right)$ and $z_3 = \ln\left(\frac{y_3}{y_1}\right)$. This model has not been used a lot in the social sciences. The biggest unsolved hurdle is how to transform the results of that model into a metric that is easy to interpret and communicate.

Proportions as explanatory variables

Multiple proportions can also show up as explanatory variables. For example, one could try to explain the Gross Domestic Product (GDP) per capita of countries with the proportion of the workforce employed in the primary (agriculture and mining), secondary (industry), and tertiary (service) sector. Often the effects of explanatory variables are interpreted as the influence of changing one variable while keeping all other variables constant. This cannot be true with multiple proportions: if a larger share of the workforce is employed in the tertiary sector than either the primary sector, or the secondary sector, or both have to decline.

In fact, if one tried to add all three proportions to the model, then most statistical software will either drop one of these proportions from the model or not estimate the model and return an error message saying that there is perfect multicolinearity. This is to be expected. With three proportions adding up to one, one of the proportions is redundant.

Consider the example in table 7. It shows the result of a quasi-likelihood model with a log link function explaining the GDP per capita with the composition of the labor force for non-oil producing countries in 2014. The GDP per capita comes from the Penn World Tables (Feenstra, Inklaar & Timmer 2015) and is measured as purchasing power parities (PPPs) in 2011 US dollars. The composition of the labor force originates from the World Development Indicators (World Bank Group 2017). The proportion of the workforce employed in the service sector is excluded from the model. This means that the exponentiated constant is the predicted GDP per capita for a country with no one employed in either agriculture or industry, and thus everybody employed in services. In such a country the model predicts a GDP per capita of about 80,000 dollars. A percentage point increase in the part of the labor force employed in agriculture *and* a corresponding percentage point decrease in the part of the labor force employed in services leads to a decrease in GDP per capita of 5%. Similarly, a percentage point increase in the share working in industry *and* a corresponding percentage point decrease in the share working in services leads to a decrease in GDP per Capita of 3%. Since the service sector is excluded from

the model, the coefficients have to be interpreted as the effect of an increase in one sector which is compensated by a corresponding decrease in the service sector alone. The model does allow for other scenarios, like a percentage point increase in agriculture and half a percentage point decrease in industry and services, but that is not how the coefficients are be interpreted. If one has only three proportions one can show the complete range effects by plotting predicted outcomes on a triangular plot, like in Figure 4.

Table 7 the influence of the composition of the labor forces on GDP per capita using a quasi-likelihood model with a log link function

	exp(β)
agriculture	0.95***
	(0.00)
industry	0.97**
	(0.00)
Constant	79,987***
	(0.00)
Observations	155

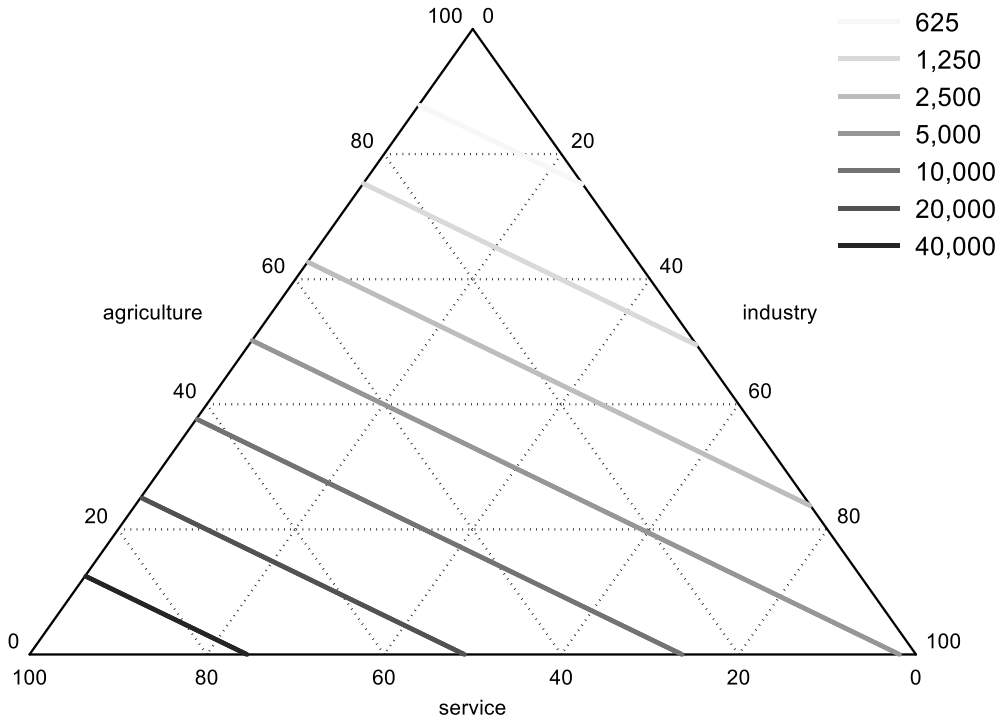
Source: Created by author for this entry using data from (Feenstra, Inklaar & Timmer 2015) and (World Bank Group 2017)

Exponentiated coefficients; *p*-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

With more proportions the triangle plot is not an option. In that case one can compare the predicted outcome in different scenarios. For example, compare the predicted GDP per capita of the USA and Germany, as they have similar proportions employed in agriculture but different proportions employed in industry and services. Then compare the predicted GDP per capita of Germany with China as they have similar proportions employed in industry, but different proportions in agriculture and services. That way one can create effects that represent more realistic scenarios than the ones where all changes are compensated by the service sector alone.

Analysis of proportions, Figure 4 Predicted GDP per Capita based on the composition of the labor force.



Source: Created by author for this entry using data from (Feenstra, Inklaar & Timmer 2015) and (World Bank Group 2017)

.

Further Readings

Papke, L. E., & Wooldridge, J. M. (2008). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics*, *145*(1-2), 121-133.

Pawlowsky-Glahn, V., & Egozcue, J. J. (2016). Spatial analysis of compositional data: a historical review. *Journal of Geochemical Exploration*, *164*, 28-32.

Ramalho, E. A., Ramalho, J. J., & Murteira, J. M. (2011). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, *25*(1), 19-68.

Rocha, A. V., & Simas, A. B. (2011). Influence diagnostics in a general class of beta regression models. *Test*, *20*(1), 95-119.

References

Aitchison, J. (2003 [1986]). *The Statistical Analysis of Compositional Data*. Caldwell, NJ: The Blackburn Press.

Cook, D. O., Kieschnick, R., & McCullough, B. D. (2008). Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance, 15*(5), 860-867.

Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2015), The next generation of the Penn World Table. *American Economic Review*, *105*(10), 3150-3182.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799-815.

Kieschnick, R., & McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling, 3*(3), 193-213.

King, G. (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton: Princeton University Press.

King, G., Tanner, M. A., & Rosen, O. (2004). *Ecological inference: New methodological strategies*. Cambridge: Cambridge University Press.

Meaney, C., & Moineddin, R. (2014). A Monte Carlo simulation study comparing linear regression, beta regression, variable-dispersion beta regression and fractional logit regression at recovering average difference measures in a two sample design. *BMC Medical Research Methodology, 14*(1), 14-35.

Mullahy, J. (2015). Multivariate fractional regression estimation of econometric share models. *Journal of Econometric Methods, 4*(1), 71-100.

Ospina, R., & Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers, 51*(1), 111.

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis, 56*(6), 1609-1623.

OIS Amsterdam (2017). *Stadsdelen in cijfers 2017.* Amsterdam: Gemeente Amsterdam.

Paolino, P. (2001). Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables. *Political Analysis, 9*(4), 325-346.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of applied econometrics, 11*(6), 619-632.

Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review, 15*(3), 351-357.

Simas, A. B., Barreto-Souza, W., & Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis, 54*(2), 348-366.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods, 11*(1), 54-71.

U.S. Bureau of Labor Statistics. (2017). *American time use survey*. Washington, DC: U.S. Bureau of Labor Statistics.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika, 61*(3), 439-447.

World Bank Group (2017) *World Development Indicators.* Washington, DC: World Bank Group.