**The USA Consumer Expenditure Survey**

One good thing about this debate was it motivated me to do something that’s been on my list for a while, which is to get my toes in the water with the USA Bureau of Labor Statistics’ impressive [Consumer Expenditure Survey](https://www.bls.gov/cex/home.htm). Huge amounts of confidentialised microdata from this mammoth program are freely available for download – without even filling in any forms! This makes it suitable to use in a blog post in a way I couldn’t with the New Zealand or Australian Household Expenditure Surveys (both of which have Confidentialised Unit Record Files for researchers, but with restrictions that get in the way of quick usage in public like this).

Big caveat for what follows – literally I looked at this survey for the first time today, and it is *very complex*. Just for starters, the Consumer Expenditure Survey really [comprises two surveys](https://www.bls.gov/cex/) – an Interview Survey for “major and/or recurring items” and the Diary Survey for “minor or frequently purchased items”. It is very possible that I am using not-the-best variables. Feedback welcomed.

**Densities of income and fuel spend**

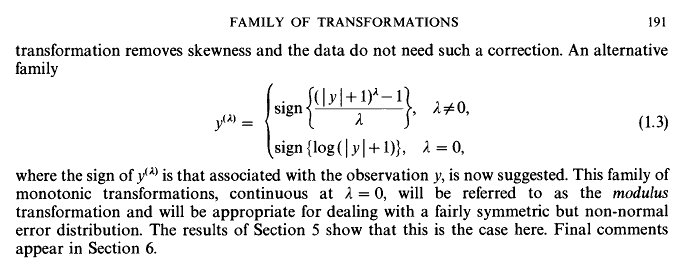
Let’s get started by downloading the data. Here’s a couple of graphs of what I think are the main variables of interest here. These draw on:

* gasmocq “Gasoline and motor oil this quarter”
* fincbtxm “Total amount of family income – imputed or collected” (in the past 12 months? although the data dictionary only implies this)
* fam\_size “Number of Members in CU” (ie in responding household)
* bls\_urbn “Is this CU located in an urban or a rural area”

Both these graphics are showing a quantity divided by household size to get a simple estimate of amount per person. A better approach would be to used equivalised figures, taking into account economies of scale for larger households and different cost structures for different age groups, but it probably would have taken me all morning just to work out a safe way of doing that so I’ve stuck with the simpler method.

It’s a great way to effectively visualise heavily skewed data that has zero and negative values as well as positive, a common occurrence with economic variables. The helper functions for the transformation are available in my bag-of-misc R package frs available from GitHub.

Code Chunks – Creating a scale transformation



I hope I’m not breaching copyright by posting that snippet; I’ll certainly take it down if anyone complains. I agree with Wickham that this useful piece of work should be much more widely known (and am sheepish myself for not knowing it) so hats off to the JSTOR for having it available at all.

John and Draper also show that the methods Box and Cox use to determine parameters of the boxcox transformation can be used to get an optimal value of their tuning parameter lambda. I’m not too worried about that just yet as I’m only using this for visualising the distribution; it might be something I want to investigate though when I move to modelling the data.

## **Implementing a transformation for easy re-use**

John and Draper’s transformation is slightly different from mine, has some nice properties and is definitely much more thoroughly thought through. It’s more complex though, and deserves a proper defined implementation for re-use. There’s no better way of doing this than creating a new transformation using the platform provided by Wickham’s scales package, very appropriate given where tip came from. Here’s how I try that:

library(scales)

*# John and Draper's modulus transformation*

modulus\_trans **<-** **function**(lambda){

trans\_new("modulus",

transform **=** **function**(y){

**if**(lambda **!=** 0){

yt **<-** **sign**(y) **\*** (((**abs**(y) **+** 1) **^** lambda **-** 1) **/** lambda)

} **else** {

yt **=** **sign**(y) **\*** (**log**(**abs**(y) **+** 1))

}

**return**(yt)

},

inverse **=** **function**(yt){

**if**(lambda **!=** 0){

y **<-** ((**abs**(yt) **\*** lambda **+** 1) **^** (1 **/** lambda) **-** 1) **\*** **sign**(yt)

} **else** {

y **<-** (**exp**(**abs**(yt)) **-** 1) **\*** **sign**(yt)

}

**return**(y)

}

)

}

It hasn’t been peer reviewed or anything so use at your own risk. Please let me know if you find anything wrong with it.

Here’s that function in use, re-creating the density plot of income from my previous post, with the zero and negative values showing up nicely but without the crampedness of showing the income on an unadjusted scale. I haven’t done polishing necessary to have axis labels at the various modal points, which is what I would do for a serious use. This code assumes the existence of a database called nzis11, created as described in this post.

library(RODBC)

library(showtext)

library(dplyr)

library(ggplot2)

*# comnect to database*

PlayPen **<-** odbcConnect("PlayPen\_prod")

sqlQuery(PlayPen, "use nzis11")

*# load fonts*

font.add.google("Poppins", "myfont")

showtext.auto()

inc **<-** sqlQuery(PlayPen, "select \* from vw\_mainheader")

ggplot(inc, aes(x **=** income)) **+**

geom\_density() **+**

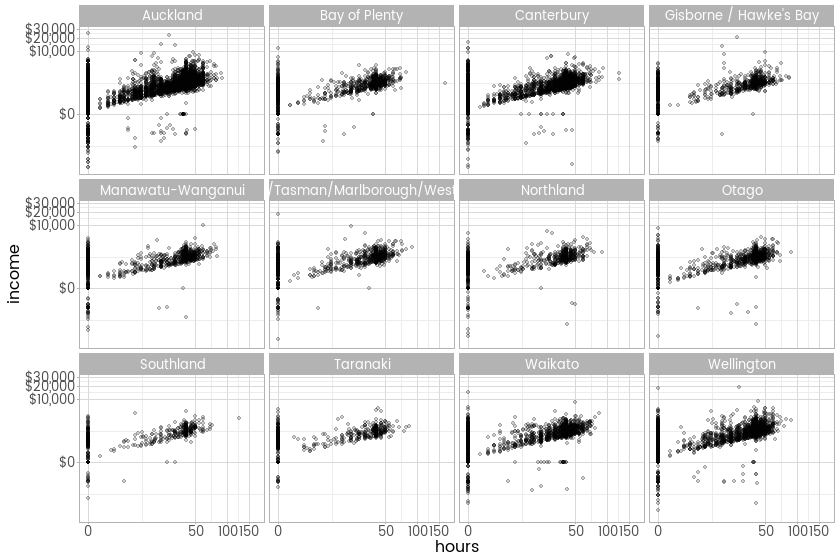
geom\_rug() **+**

scale\_x\_continuous(trans **=** modulus\_trans(lambda **=** 0.25), label **=** dollar) **+**

theme\_minimal(base\_family **=** "myfont")

## **Income (including negatives) and hours worked**

So the great advantage of a statistical transformation within the scales paradigm is that you never need to worry again about the transformation, you just make it an argument to scale\_x\_continuous (or any continuous scale, including colour and size, if you want). I’ll finish this latest step in the journey with New Zealand’s Income Survey simulated unit record file from Statistics New Zealand by showing the new transformation in action on two axes at once, with a plot showing the relationship between hours worked and income earned by region, this time with the negative incomes left in.



ggplot(inc, aes(x **=** hours, y **=** income)) **+**

facet\_wrap(**~**region) **+**

geom\_point(alpha **=** 0.2) **+**

scale\_x\_continuous(trans **=** modulus\_trans(0.25)) **+**

scale\_y\_continuous(trans **=** modulus\_trans(0.25), label **=** dollar) **+**

theme\_light(base\_family **=** "myfont")

### **Late addition - comparison to the untransformed version**

As requested in the comments, here’s the untransformed version

ggplot(inc, aes(x **=** hours, y **=** income)) **+**

facet\_wrap(**~**region) **+**

geom\_point(alpha **=** 0.2) **+**

scale\_y\_continuous(label **=** dollar) **+**

theme\_light(base\_family **=** "myfont")

Here’s code to do that download, prepare the data for later and draw graphs:

# mandatory reading: https://www.bls.gov/cex/pumd\_novice\_guide.pdf

library(haven)

library(tidyverse)

library(scales)

library(frs) # for modulus transforms of scales

library(MASS) # for rlm. Watch out for dplyr::select and MASS::select clash.

library(ggExtra)

library(survey)

# download the latest year's data (caution - 71MB):

dir.create("bls-cd")

download.file("https://www.bls.gov/cex/pumd/data/stata/intrvw16.zip",

destfile = "bls-cd/intrvw16.zip", mode = "wb")

# download the data dictionary

download.file("https://www.bls.gov/cex/pumd/ce\_pumd\_interview\_diary\_dictionary.xlsx",

destfile = "bls-cd/ce\_pumd\_interview\_diary\_dictionary.xlsx", mode = "wb")

unzip("bls-cd/intrvw16.zip", exdir = "bls-cd")

# The FMLI files have characteristics, income, weights, and summary expenditure

# the numbers in file names refer to year and quarter. So intrvw16/fmli162.dta is

# 2016, second quarter

# Import 2017 first quarter:

fmli171 <- read\_dta("bls-cd/intrvw16/fmli171.dta")

# looks like no attributes associated with the various columns; have to use the Excel data dictionary

str(attributes(fmli171))

attr(fmli171, "labels")

attr(fmli171, "label")

# which columns are numeric, candidates to hold income and expenditure non-binned data:

formats <- sapply(fmli171, function(x){attr(x, "format.stata")})

formats[grepl("[0-9]g", formats)]

# Bit of basic processing for convenience. Make a new data frame called d with some labels we need later,

# and the per person variables calculated in one place.

# First, the code labels for income classes:

inclasses <- c("Less than $5,000", "$5,000 to $9,999", "$10,000 to $14,999", "$15,000 to $19,999",

"$20,000 to $29,999", "$30,000 to $39,999", "$40,000 to $49,999", "$50,000 to $69,999",

"$70,000 and over")

inclass\_lu <- data\_frame(

inclass = paste0("0", 1:9),

inclass\_ch = ordered(inclasses, levels = inclasses)

)

# create the working version of the data frame:

d <- fmli171 %>%

mutate(gas\_pp = gasmocq / fam\_size \* 4,

inc\_pp = fincbtxm / fam\_size,

gas\_p\_inc = gasmocq \* 4 / fincbtxm,

no\_gas = as.integer(gas\_pp == 0),

bls\_urbn\_ch = ifelse(bls\_urbn == 1, "Urban", "Rural"),

) %>%

left\_join(inclass\_lu, by = "inclass")

# parameter for how much to transform the dollar scales:

lambda <- 0.2

ggplot(d, aes(x = gas\_pp, colour = bls\_urbn\_ch)) +

geom\_density() +

scale\_x\_continuous(label = dollar,

trans = modulus\_trans(lambda),

breaks = modulus\_breaks(lambda)) +

geom\_rug() +

ggtitle("Spend on petrol has a spike at $0 and a skewed distribution beyond that",

"Unweighted, 2017 quarter 1. Horizontal scale has been transformed.") +

labs(caption = "Source: USA Bureau of Labor Statistics Consumer Expenditure Survey",

colour = "",

x = "Expenditure per person on gasoline and motor oil this quarter x 4")

# some looks at various variables as described in the Excel data dictionary:

summary(fmli171$cuincome) # total income; not present

summary(fmli171$earnincx) # earnings before tax; not present

table(fmli171$earncomp) # composition of earners

table(fmli171$cutenure) # housing tenure (6 categories)

table(fmli171$inclass) # income bracket

summary(fmli171$inc\_rank) # Weighted cumulative percent ranking based on total current income before taxes (for complete income reporters)

summary(fmli171$othrincm) # other income

summary(fmli171$ffrmincx) # farm income, not present

summary(fmli171$fincbtax) # total amount of family income before taxes in the past 12 months

summary(fmli171$finlwt21) # final calibrated weight

ggplot(d, aes(x = inc\_pp, colour = bls\_urbn\_ch)) +

geom\_density() +

geom\_rug() +

ggtitle("After imputation, income has a skewed distribution with no spike at zero",

"Unweighted, 2017 quarter 1. Horizontal scale has been transformed.") +

scale\_x\_continuous(label = dollar,

trans = modulus\_trans(lambda),

breaks = modulus\_breaks(lambda)) +

labs(caption = "Source: USA Bureau of Labor Statistics Consumer Expenditure Survey",

colour = "",

x = "Family income per person before taxes in the past 12 months")

**Relationship of income and fuel spend**

Here are some different ways of looking at the relationship between household income and the amount spent on fuel. They all show a lot of variation between individual households, but significant evidence of a material positive relationship between the two.

First, here’s a graph that tries to combine the binned “income classification” with the ranking of the household on income(ie its quantile if you like). The categories aren’t as neat as might be expected, I’m pretty sure because of these variables representing different states of imputation:

The collection of people who don’t spend any money on petrol drags the regression line downwards, but it’s the slope that counts; definitely upwards. The higher ranked a household is on income, the more they spend per person on gasoline and motor oil.

BTW, note that the points in these plots are different sizes. The size is mapped to the calibrated survey weight indicating how many people in the US population each sample point is representing; this is a good starting point for trying to represented weighted data in a scatter plot.

I’m wary of using quantiles or rankings in this sort of analysis; I don’t see much or any gain over other transformations and new risks and interpretability problems are introduced. Perhaps more usefully, here are some straightforward scatterplots of income per person and vehicle fuel expenditure per person:

No doubt about that strong relationship; poorer households spend less on fuel (and nearly everything else, of course, although that’s not shown) than do richer households.

On the other hand, there’s equally no doubt that poorer households spend more on fuel as a proportion of their income:

Here’s the code for those four graphics:

#-----------income quantile v fuel scatter plot-------------------

ggplot(d, aes(y = gas\_pp, x = inc\_rank, size = finlwt21)) +

geom\_point(aes(colour = inclass\_ch), alpha = 0.2) +

geom\_smooth(method = "rlm", aes(weight = finlwt21)) +

scale\_y\_continuous(label = dollar,

trans = modulus\_trans(lambda),

breaks = modulus\_breaks(lambda)) +

labs(x = "Income quantile") +

ggtitle("Spend on petrol increases as income quantile of the household increases",

"Blue line shows robust regression using M estimator. Vertical axis is transformed.") +

labs(caption = "Source: USA Bureau of Labor Statistics Consumer Expenditure Survey",

colour = str\_wrap("Income class of household based on income before taxes", 20),

y = "Expenditure per person on gasoline and motor oil this quarter") +

theme(legend.position = "right") +

guides(colour = guide\_legend(override.aes = list(size = 4, alpha = 1))) +

scale\_size\_area(guide = "none")

#-----------scatter plots----------------

p <- ggplot(d, aes(x = inc\_pp, y = gas\_pp, size = finlwt21)) +

geom\_point(alpha = 0.3) +

geom\_smooth(method = "rlm") +

scale\_x\_continuous("Income per person in household in past 12 months",

label = dollar,

trans = modulus\_trans(lambda),

breaks = modulus\_breaks(lambda)) +

scale\_y\_continuous(label = dollar,

trans = modulus\_trans(lambda),

breaks = modulus\_breaks(lambda)) +

theme(legend.position = "none") +

ggtitle("Spend on petrol increases as income of the household increases",

"Blue line shows robust regression using M estimator. Both axes are transformed.") +

labs(caption = "Source: USA Bureau of Labor Statistics Consumer Expenditure Survey",

y = "Expenditure per person on gasoline and motor oil this quarter x 4")

ggMarginal(p, type = "density", fill = "grey", colour = NA)

ggMarginal(p %+% filter(d, gas\_pp > 0 & inc\_pp > 0),

type = "density", fill = "grey", colour = NA)

#-----------proportion spent on fuel-------------

d %>%

filter(inc\_pp > 1000) %>%

ggplot(aes(x = inc\_pp, y = gas\_p\_inc, size = finlwt21)) +

geom\_point(alpha = 0.3) +

scale\_x\_continuous("Income per person in household in past 12 months",

label = dollar,

trans = modulus\_trans(lambda),

breaks = modulus\_breaks(lambda)) +

coord\_cartesian(ylim = c(0, 1)) +

scale\_y\_continuous() +

geom\_smooth(se = FALSE) +

theme(legend.position = "none") +

ggtitle("Poorer households spend more on petrol proportionately than do richer households") +

labs(caption = "Source: USA Bureau of Labor Statistics Consumer Expenditure Survey",

y = "Expenditure per person on gasoline and motor oil\nas a proportion of income")

**Who is likely to spend nothing on petrol at all?**

Finally, I was interested in who spends nothing on petrol at all. This relates to Mr Warburton’s argument that the New Zealand Household Economic Survey if flawed for these purposes because the average spend on petrol includes people who have been priced out of vehicles altogether. In fact, with the US data, there is a very strong negative relationship between household income and the probability of spending zero on gasoline and motor oil.:

However, as the previous scatterplots showed, removing either or both of the zero income and zero fuel spend cases from the US Consumer Expenditure survey doesn’t serve to remove the relationship between income and gasoline spend.

Finally, here’s the code for this last bit of analysis:

p1 <- d %>%

ggplot(aes(x = inc\_pp, y = no\_gas, size = finlwt21)) +

geom\_point(alpha = 0.1) +

geom\_smooth(method = "glm", method.args = list(family = "binomial")) +

scale\_x\_continuous("Income per person in household in past 12 months",

label = dollar,

trans = modulus\_trans(lambda),

breaks = modulus\_breaks(lambda)) +

labs(y = "Probability of spending $0 on gasoline and motor oil this quarter\n") +

theme(legend.position = "none") +

labs(caption = "Source: USA Bureau of Labor Statistics Consumer Expenditure Survey 2017 Q1") +

ggtitle("Poorer households are more likely to spend nothing on petrol at all",

"Blue line shows logistic regression; points show American households")

print(p1)

p2 <- p1 %+% filter(d, inc\_pp > 0) +

ggtitle("", "Effect holds even if restricted to households with positive income")

print(p2)

# Check out the relationship with slightly more formal modelling than just on the fly in the graphic.

# crude approximation of the survey design, that takes just the primary sampling units and weights

# into account but ignores the stratification and post-stratification calibration (this will be

# conservative for inference so that's ok).

# Also, better would be to transform `inc\_pp`, or use a gam, or something. This will do for now!

dd <- svydesign(~psu, data = d, weights = ~finlwt21)

mod <- svyglm(no\_gas ~ inc\_pp, design = dd, family = "quasibinomial")

summary(mod)