Spend on Petrol by Income Analysis

Fuel tax debates

So, there’s currently a vibrant debate on a small New Zealandish corner of Twitter about a petrol tax coming into effect in Auckland today, and the different impacts of such taxes on richer and poorer households.

The Government has released analysis from the Stats NZ Household Expenditure Survey showing higher petrol consumption per household for higher income households (and hence paying more of the new tax). Sam Warburton, an economist with the New Zealand Institute, argues in response that poorer households have older, often larger, less efficient vehicles, leading to higher fuel costs per kilometre. This means a fuel tax will not only result in poor people paying more as a percentage of their income (as any sales tax on basic commodities will do), but paying more *per kilometre*. Further, poor people are more likely to live out of the big urban centres, and when in an urban area are less likely to be well serviced by public transport.

Mr Warburton also argues that the results from the Household Expenditure Survey are misleading because they include “people who can’t afford or otherwise don’t own cars”. His own analysis from the Ministry of Transport’s Household Travel Survey shows that households including two or more children, with Māori, with unemployed people, or in the poorer regions all pay more tax per kilometre.

I’m totally convinced by Mr Warburton on the argument that poorer people are paying more tax per kilometre, which makes a fuel tax a particularly regressive tax. Even paying the same rate per kilometre would be regressive because transport costs are a higher proportion of income for poorer people; so more tax *per kilometre* is really rubbing salt into the wound. I’m not convinced though by the suggestion that they will pay more of the new tax even in absolute terms; I’m inclined to trust the Household Economic Survey on this one.

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. 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 - 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.

Both these graphics - and most of those that follow - also use John and Draper’s modulus transform that I wrote about in some of my first ever posts on this blog. 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.

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)

We [made the data available](https://twitter.com/Economissive/status/1028861304306425856) for the below analysis

One of the issues of controversy about a levy like this is whether it will lead to “price spreading” – fuel companies absorbing some of the extra tax in Auckland and increasing prices in other regions. A relatively small number of firms make retail pricing decisions about fuel in New Zealand so it’s plausible that imperfect competition is making this possible. I had a look at the data to see if the intervention of the fuel levy in New Zealand’s biggest city can be seen to impact on fuel pricing in the rest of the country.

To cut to the chase, this graphic exemplifies my approach and results:

We see that after the spike at the time the tax was introduced, fuel prices in other regions have converged somewhat on Auckland’s prices (particularly when considering the relative change happening before the tax). The impact of the tax is still clearly felt much more strongly in Auckland than anywhere else (as of course would be expected – the question at issue is whether *anywhere* else would be impacted at all). More on that later.

**The data**

First, let’s look at the data Sam’s collected. The Twitter thread linked to above provides an Excel workbook on Dropbox. There’s a worksheet for each region (which are defined similarly, but not identically, to New Zealand’s official Regions) as well as one describing the source. For each region we have data on fuel prices for combinations of Company, Date, fuel type (eg diesel, 91 octane, etc) and Region. If we plot all the data other than liquid petroleum gas (which has particularly sparse observations), it looks like this:

“Companies” have been sorted in order of increasing average price for that graphic, but fairly crudely (ie not taking into account different mixes of fuel type by Company).

We can see we have more data for 91 octane petrol than the other types. For the rest of this post I’ll be focusing on just 91 octane.

Here’s the R code to tidy up the data to this point and draw the graphic. It assumes you’ve manually downloaded the Excel workbook to your working folder (I’m currently working with severely restricted internet, so couldn’t experiment in automating that process.)

library(tidyverse)

library(scales)

library(openxlsx)

library(nlme)

# Import data:

sn <- getSheetNames("fuel price data.xlsx")

sn <- sn[sn != "Source"]

fuel\_orig <- list()

for(i in 1:length(sn)){

tmp <- read.xlsx("fuel price data.xlsx", sheet = sn[i], cols = 1:7,

detectDates = TRUE, na.strings = c("NA", "n/a"))

tmp[ , "region"] <- sn[i]

fuel\_orig[[i]] <- tmp

}

# Combine into a single data frame

fuel\_df <- do.call("rbind", fuel\_orig)

# some useful extra information:

south\_island <- c("Canterbury", "Nelson", "Otago", "Southland", "West Coast")

big\_four <- c("CALTEX", "Z ENERGY", "BP", "MOBIL")

# Make long, thin, tidy version:

fuel\_tidy <- fuel\_df %>%

select(-LPG) %>%

gather(fueltype, value, -Company, -Date, -region) %>%

filter(!is.na(value)) %>%

mutate(island = ifelse(region %in% south\_island, "South", "North"),

company\_type = ifelse(Company %in% big\_four, "Big Four", "Smaller")) %>%

mutate(region = fct\_reorder(region, as.numeric(as.factor(island))),

Company = fct\_reorder(Company, value))

# Overview graphic:

fuel\_tidy %>%

ggplot(aes(x = Date, y = value, colour = Company)) +

facet\_grid(fueltype~region, scales = "free\_y") +

geom\_point(size = 0.8) +

scale\_y\_continuous("Price per litre at the pump", label = dollar) +

labs(x = "Date in 2018",

caption = "Source: pricewatch.co.nz, collated by @Economissive") +

ggtitle("Petrol prices in New Zealand over several months in mid 2018") +

guides(colour = guide\_legend(override.aes = list(size=5))) +

scale\_colour\_brewer(palette = "Set1")

**Regional comparisons with Auckland**

I tried a couple of different ways of comparing prices in individual regions with those in Auckland. I think this graphic is probably the most informative and straightforward:

The grey line in the background of each facet represents Auckland’s price; the shaded blue rectangle is the post-tax period (ie 1 July 2018 and onwards). The grey shaded area shows the difference between the given region’s price and that of Auckland.

We can see a lot of regional variation here, and an interesting pattern with three (or maybe even all) of the South Island regions experiencing price declines in June then picking up in July. Of course, the 11.5 cent increase in price in Auckland is very obvious in the grey lines and shading. Later on I’ll be using time series intervention analysis on this data; this is an approach commonly used in evaluating the impact of evaluations. If we were only after the direct impact, there would be no need to do any statistical tests beyond this graphic above; the big spike in prices hits you between the eyes, and there is no doubt about the discontinuity in Auckland’s prices on 1 July! The question, of course, is how sustained that impact is, and whether it bled into secondary impacts in other regions.

Here’s a second graphic that tries to visually simplify what’s going on, by calculating a single line of the ratio of prices in each region to those in Auckland. I think what it gains in visual simplicity (less lines and shading) it loses in clear interpretability. In particular, it’s not possible to tell from this graphic what changes in the graphic come from changes in Auckland, and which come from changes in the comparison region. That’s not a deal-breaker for using a graphic like this, but it does strongly suggest we should also include the first one, with the rawer average prices per region plainly shown without transformation, for context.

Note that even after the introduction of the Auckland-specific fuel tax, 91 octane petrol in several regions still costs more than in Auckland. The regions with prices higher than Auckland in the most recent data in the collection are West Coast, Otago, Nelson, Canterbury and Wellington.

Here’s the R code for those two graphics:

# construct two convenient summaries of the data, different ways of comparing regions to Auckland:

p91 <- fuel\_tidy %>%

filter(fueltype == "91") %>%

group\_by(region, island, Date) %>%

summarise(value = mean(value, tr = 0.2)) %>%

ungroup()

p91\_rel <- p91 %>%

group\_by(Date) %>%

mutate(Auckland = value[region == "Auckland"]) %>%

filter(! region %in% c("Auckland", "Wairarapa")) %>%

mutate(perc\_of\_auck = value / Auckland)

# Plot showing original price data

ggplot() +

# annoying trick necessary here to draw a semi-transparent background rectangle:

geom\_rect(data = data.frame("hello world"),

xmin = as.Date("2018-07-01"), xmax = Inf, ymin = -Inf, ymax = Inf, fill = "blue", alpha = 0.1) +

# now draw the actual data:

geom\_ribbon(data = p91\_rel, aes(x = Date, ymin = Auckland, ymax = value), fill = "grey", alpha = 0.5) +

geom\_line(data = p91\_rel, aes(x = Date, y = Auckland), colour = "grey50") +

geom\_line(data = p91\_rel, aes(x= Date, y = value, colour = island), size = 1.2) +

facet\_wrap(~region, ncol = 3) +

scale\_y\_continuous("Price of 91 octane petrol compared to in Auckland\n", label = dollar) +

labs(x = "2018; grey line shows Auckland",

caption = "Source: pricewatch.co.nz, collated by @Economissive")

# ratio of Auckland prices to others

ggplot() +

geom\_hline(yintercept = 1, colour = "grey50") +

geom\_rect(data = data.frame("hello world"),

xmin = as.Date("2018-07-01"), xmax = Inf, ymin = -Inf, ymax = Inf, fill = "blue", alpha = 0.1) +

geom\_line(data = p91\_rel, aes(x= Date, y = perc\_of\_auck, colour = island)) +

facet\_wrap(~region, ncol = 3) +

scale\_y\_continuous("Price of 91 octane petrol as a percentage of in Auckland\n", label = percent) +

labs(x = "2018",

caption = "Source: pricewatch.co.nz, collated by @Economissive")

**Modelling the impact of an intervention over time**

When it came to directly addressing our question of interest regarding price spreading, I opted to group all non-Auckland regions together and compare average prices there with those in Auckland. There are better ways of modelling this that make full use of the granular data available (mostly involving mixed effects models, and more complex ways of representing the trend over time than linearly; and they would certainly take into account weighting from the spread in population over regions) but they come with big costs in complexity that I don’t have time for right now. Plus, the difference-of-averages method struck me as the easiest way to interpret and communicate, not to mention think about, the question of whether prices were converging back towards eachother after the initial shock of the addition of the tax. This leads me to the graphic I showed earlier in this post:

The linear regression lines shown in that graphic are a simplified version of the formal statistical model we want to fit and use to test our hypothesis. We’re looking for evidence that the slope of the post-tax line is materially less than the slope of the pre-tax line; in other words, is the gap in pricing between Auckland and other regions declining after the initial 11.5 cent shock of the tax.

I defined this as a simple linear model, but fit it using generalized least squares with time series residuals (auto-regressive moving average of order (1, 1)). This is straightforward to specify and fit using the gls function in Pinheiro, Bates et al’s nlme package, but there are other ways of doing it too.

This results in the following:

|  |  |
| --- | --- |
|  | |
|  | *Dependent variable:* |
|  |  |
|  | value |
|  | |
| Date | 0.001\*\* |
|  | (0.0003) |
|  |  |
| post\_tax | 19.159\*\* |
|  | (7.804) |
|  |  |
| Date:post\_tax | -0.001\*\* |
|  | (0.0004) |
|  |  |
| Constant | -10.470\*\* |
|  | (4.946) |
|  |  |
|  | |
| Observations | 93 |
| Log Likelihood | 330.591 |
| Akaike Inf. Crit. | -647.182 |
| Bayesian Inf. Crit. | -629.761 |
|  | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |

…which simply confirms what is visually obvious, that there is indeed statistically significant evidence of the slope changing direction downwards after the tax is introduced. In other words, we do have evidence consistent with some degree of “spreading” taking place. After the initial clean shock of the introduction of the tax, prices in Auckland and in the rest of the country are indeed converging somewhat; although nowhere near as much as the full cost of the tax.

This effect holds whether I use all of New Zealand as the comparison point or just the South Island (which has less competition in fuel retailers) or just the North Island, although in the latter case the effect is not as strong (as can be seen in the graphic). It also doesn’t seem to matter whether we use all available prices, or just those of the “big four” companies that are present in all regions.

We can’t say for sure the effect comes from introducing the tax from just looking at the numbers. Drawing that conclusion would require carefully considering any other possible causality options. For example, one driver of the pattern we’re seeing is clearly that prices in Canterbury, Nelson and Otago stopped declining and started rising slightly in July. What are other plausible causes of that pattern? Understanding and considering such alternative theories would need more knowledge of the fuel market in New Zealand than I have, so I’ll leave it to others to debate that. All I can safely conclude is what I wrote above:

after the spike caused by the tax, fuel prices in Auckland and in the rest of the country are converging somewhat (although much less than the full cost of the tax), and plausibly this is because of companies’ price adjustments down in Auckland and up elsewhere to spread the cost of the tax over a broader base.

Here’s the code for that final graphic and statistical modelling;

# Data on the difference between Auckland's average price and those in other areas:

diff\_data <- fuel\_tidy %>%

filter(fueltype == "91" & company\_type == "Big Four") %>%

group\_by(Date) %>%

summarise(auck\_v\_rest =

mean(value[region == "Auckland"]) -

mean(value[region != "Auckland"]),

auck\_v\_si =

mean(value[region == "Auckland"]) -

mean(value[island == "South"]),

auck\_v\_ni =

mean(value[region == "Auckland"]) -

mean(value[island == "North" & region != "Auckland"]),

) %>%

mutate(post\_tax = as.integer(Date >= as.Date("2018-07-01"))) %>%

gather(comparison, value, -Date, -post\_tax) %>%

mutate(comparison = case\_when(

comparison == "auck\_v\_si" ~ "Compared to South Island",

comparison == "auck\_v\_ni" ~ "Compared to rest of North island",

comparison == "auck\_v\_rest" ~ "Compared to all NZ except Auckland"))

# Graphic:

ggplot(diff\_data, aes(x = Date, y = value)) +

facet\_wrap(~comparison, ncol = 3) +

geom\_line() +

geom\_smooth(aes(group = post\_tax), method = "lm") +

scale\_y\_continuous("Average price of 91 octane petrol in Auckland\nminus average price in comparison area",

label = dollar) +

labs(x = "Date in 2018\nAverage prices have not been weighted by population or sales",

caption = "Source: pricewatch.co.nz, collated by @Economissive") +

ggtitle("Fuel prices in Auckland compared to three other comparison areas",

"Restricted to prices from BP, Caltex, Mobil and Z Energy")

# Modelling:

# first, make a convenient subset of the data (useful later in various tests and diagnostics):

D <- subset(diff\_data, comparison == "Compared to all NZ except Auckland")

# Fit model, taking care to specify time series residuals, which aren't as useful for inference

# as i.i.d. residuals and hence lead to more conservative inference:

model <- gls(value ~ Date \* post\_tax,

data = subset(diff\_data, comparison == "Compared to all NZ except Auckland"),

cor = corARMA(p = 1, q = 1))

# print-friendly summary of coefficieints

stargazer::stargazer(model, type = "html")

# more comprehensive summary (not shown in blog):

summary(model)

**Observations per region**

A topic of interest in fuel pricing debate in New Zealand is the number of companies present in each region, with a particular focus on the presence of Gull. In case of interest, here are the observations in the pricewatch data collected by Sam Warburton:

| **region** | **–** | **ALTERNATE FUEL** | **GULL** | **CHALLENGE** | **CALTEX** | **MOBIL** | **Z ENERGY** | **GAS ALLEY** | **BP** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Auckland | 206 | 143 | 349 | 251 | 279 | 372 | 279 | 279 | 372 |
| Bay of Plenty | 262 | 200 | 295 | 202 | 278 | 264 | 279 | 161 | 370 |
| Coromandel | 3 | 6 | 249 | 4 | 209 | 219 | 233 | 69 | 277 |
| Waikato | 0 | 0 | 228 | 227 | 276 | 302 | 278 | 220 | 367 |
| Hawke’s Bay | 0 | 45 | 217 | 16 | 261 | 228 | 255 | 132 | 345 |
| Northland | 0 | 0 | 205 | 52 | 274 | 229 | 256 | 268 | 275 |
| Manawatū-Whanganui | 0 | 108 | 193 | 120 | 275 | 265 | 277 | 176 | 371 |
| Taranaki | 0 | 184 | 178 | 188 | 243 | 88 | 274 | 188 | 301 |
| Wellington | 0 | 91 | 78 | 252 | 278 | 277 | 279 | 213 | 372 |
| East Coast | 0 | 41 | 71 | 158 | 221 | 122 | 145 | 104 | 211 |
| Central Plateau | 0 | 0 | 44 | 51 | 159 | 244 | 275 | 0 | 295 |
| Wairarapa | 0 | 1 | 0 | 0 | 2 | 42 | 3 | 2 | 103 |
| Canterbury | 0 | 262 | 0 | 279 | 279 | 278 | 279 | 245 | 372 |
| Nelson | 0 | 50 | 0 | 34 | 216 | 237 | 263 | 163 | 237 |
| Otago | 0 | 225 | 0 | 226 | 273 | 265 | 277 | 116 | 364 |
| Southland | 0 | 114 | 0 | 197 | 249 | 195 | 226 | 98 | 245 |
| West Coast | 0 | 115 | 0 | 227 | 119 | 121 | 115 | 65 | 242 |

That table was generated by:

library(knitr)

fuel\_tidy %>%

group\_by(region, Company) %>%

summarise(freq = n()) %>%

spread(Company, freq, fill = 0) %>%

arrange(desc(GULL)) %>%

kable