There is a relationship between weight and height in US adults, as seen in the US Centers for Disease Control and prevention (CDC) Behavioral Risk Factor Surveillance System, an annual telephone survey of around 400,000 interviews per year. In particular, I tested the widely-circulated claim that Body Mass Index (BMI) exaggerates the “fatness” of tall people because empirically (it is claimed), weight is proportional to height to the power of 2.5, not to 2.0 as implied by the formula for calculating BMI. I showed that in 2018 in the BRFSS at least, weight is actually proportional to height to the power of around 1.6 (after adjusting for race and sex). So if anything, the BMI calculation exaggerates the weight of *short* people, not tall people, on the logic of the original critique.

I have a couple of follow-up points today:

* whether it is best to to logarithm-transform a response variable in a particular model, or model it on its original scale and use a logarithm link function in a generalized linear model
* whether socioeconomic class is a confounding variable (richer people are taller, and also healthier in general, so perhaps the relatively low weights of tall people measured by the BMI are caused by the better health of rich people, not by a problem in the measurement).

Let’s deal with those one at a time.

**Log transform or log link?**

Our height and weight data looks a bit like this (limiting ourselves to a random sub-sample of the full survey to better illustrate a point without over-plotting):

There is a very slight curvature (not really visible) in the relationship between height and weight, but more importantly (if one is considering estimating a model by ordinary least squares) there is a clear “fanning out” effect – as the expected value of weight increases, so does its variance. In intro stats and econometrics courses, one is trained to recognise this effect as heteroskedasticity, which violates one of the assumptions needed for ordinary least squares to be an optimal fitting algorithm and certain inferences from its results to be valid.

The workaround usually taught in those same courses is to take logarithmic transforms of both variables. With many metrics common in the social sciences, this will turn a skewed distribution into one that at least resembles a Gaussian (‘Normal’) one. At the same time this conveniently turns your underlying model into one where the linear coefficients are elasticities, which have a nice interpretation (as exploited by me last week in examining how closely BMI matches real weight-height relationship):

Here’s the code for those two charts (assumes you have run last week’s code first), and doing some setup for later:

#-------------prep - small additions to last week--------

# Extra packages:

library(broom)

library(ggExtra)

# Set some reference levels for factor variables for better interpretation and stable estimates.

# Also take the chance to explicitly code some factors as missing so we get estimates for those

# people rather than dropping out of the model altogether:

llcp\_releveled <- llcp\_all %>%

mutate(hhinc = fct\_explicit\_na(hhinc),

education = fct\_explicit\_na(education),

race = fct\_explicit\_na(race),

age = fct\_explicit\_na(age)) %>%

mutate(hhinc = fct\_relevel(hhinc, "$25k to $35k"),

education = fct\_relevel(education, "Grade 12 GED"))

# specify survey design for use in invers-probability weighting in later GLM estimation

llcp\_svy <- svydesign(~psu, weights = ~survey\_weight, data = llcp\_releveled)

#----scatter plot charts----------

p20 <- llcp\_small %>%

ggplot(aes(x = height, y = weight)) +

geom\_jitter(size = 1, alpha = 0.2) +

geom\_smooth(method = "gam", se = FALSE) +

labs(title = "A non-linear relationship, and heteroskedastic response variable",

subtitle = "A standard challenge for the assumptions justifying ordinary least squares as an estimation method.",

caption = the\_caption,

x = "Height (m)",

y = "Weight (kg)")

ggMarginal(p20, fill = "steelblue", alpha = 0.5, colour = "white")

p22 <- p20 +

scale\_x\_log10() +

scale\_y\_log10() +

labs(title = "Log-log transformations can fix two problems at once",

subtitle = "Marginal densities are now symmetrical and closer to Normal; a linear relationship now shows elasticity.",

x = "Height (m) (logarithmic scale)",

y = "Weight (kg) (logarithmic scale)")

ggMarginal(p22, fill = "steelblue", alpha = 0.5, colour = "white")

In fact, this log transformation approach, in my opinion, is fine. But an alternative approach, which some argue we should always follow, is to model the response variable in its original metric (kg, not log kg), explicitly dealing with the increasing variance (ie dropping the “constant variance” assumption behind the usual rationale of ordinary least squares). We can do this while keeping the functional relationship of linearity in the logarithms of the two variables (and hence the elasticity interpretation of coefficients) by moving from a model of:

E(log(y)) = Xb

(which is the “log transform” approach), to:

log(E(y)) = Xb

(which is the “log link function” approach, as used in a Generalized Linear Model).

Where X is a matrix of explanatory variables that includes (in this case) the logarithm of height.

In both those formulae, E() represents the “Expected value”. Another way of thinking about this is that in the first version, the random residuals have a multiplicative relationship to y, whereas in the second they have an additive relationship. That may or may not help.

I’m not sure where I come down on this, although I like the elegance of always modelling your variable in its own units (ie the second approach, with a link function). In my blog post last week I used the transformation approach, and it worried me enough that I wanted to re-specify and fit my model as a generalized linear model with a a logarithm link function and variance proportional to the expected value of the response.

Here’s how I refit these two versions of a model of weight on height (plus sex, race and age):

#------------------Different error structure---

model1 <- svyglm(log(weight) ~ log(height) + sex + race + age, design = llcp\_svy)

model12 <- svyglm(weight ~ log(height) + sex + race + age, design = llcp\_svy,

family = quasi(link = "log", variance = "mu"))

tibble(

Variable = names(coef(model1)),

`Log - Log Gaussian` = coef(model1),

`Quasi GLM with log link and mu variance` = coef(model12)

)

Which gets us these results for the point estimates of the various coefficients in the linear predictor:

| **Variable** | **Log – Log Gaussian** | **Quasi GLM with log link and mu variance** |
| --- | --- | --- |
| (Intercept) | 3.5230877 | 3.5188870 |
| log(height) | 1.5484749 | 1.6139666 |
| sexMale | 0.0575332 | 0.0449635 |
| raceAmerican Indian or Alaskan Native | 0.0410786 | 0.0433961 |
| raceAsian | -0.1206134 | -0.1252266 |
| raceBlack or African American | 0.0512117 | 0.0514042 |
| raceHispanic, Latino/a, or Spanish origin | 0.0128204 | 0.0122439 |
| raceOther | -0.0085671 | -0.0042133 |
| racePacific Islander | 0.0317032 | 0.0371853 |
| race(Missing) | 0.0001800 | 0.0014521 |
| ageAge 65 or older | -0.0108261 | -0.0148626 |
| age(Missing) | -0.0400688 | -0.0443465 |

There are some small differences, but not material (at least for my purposes). The coefficients in the “Log-Log Gaussian” column differ slightly from those reported last week because I am now explicitly coding some people as “missing” data on particular variables and keeping them in the data (with an extra coefficient for each “missing” effect) rather than dropping them as I did last week.

Note that even my log-log transform model isn’t actually using ordinary least squares, but inverse-probability weighting (this is what svyglm() does under the hood in both cases) to take into account the complex survey origin of the data. That’s an important issue for working with this data, but not for the general point about obtaining similar results when modelling E(log(y)) to what one gets when modelling log(E(y)).

**Socioeconomic class as a confounding variable**

Having disposed of that model specification and estimation issue, what of the second point that was worrying me? Here’s what Nicholas Erskine wrote in response to last week’s post:

I smell confounding here. Height is correlated with socioeconomic status. Taller people on average having lower BMI scores could just be telling you that taller people are healthier on average

Well, I’d actually spotted some issues related this and thought about it myself beforehand, and the evidence that I’m not making that up is that my original post included code that prepared variables on household income and level of education for the survey respondents. Analysis of that issue didn’t make the cut for the blog post though. Let’s look at this whole height-weight-socioeconomic status relationship thing now.

I’m going to do this by fitting a couple of models to understand the relationship of my two key socioeconomic variables (household income and level of education) and:

* height (controlling for demographics)
* weight (controlling for demographics and height)

The second of these models will let me see if my estimate of the “elasticity” of height and weight stays at around 1.6, as per my previous blog post, rather than the 2.5 claimed by Trevethen.

**Socioeconomic variables relationship to *height***

Here are the 95% confidence intervals for the ‘impact’ on height of various variables. The scare quotes around ‘impact’ are because I believe my socioeconomic measurements of status *today* for my respondents are standing in as proxies for a confounder such as “socioeconomic status while growing up”. I don’t believe I there is information available in my dataset on parents’ income and education level.

What we see is that people with higher incomes and education levels are indeed taller, after controlling for age, race and sex. I’ve deliberately fit this model on the original scale with an identity link function for interpretability, so we can say (for example) that when controlling for all other factors, men are nearly 15cm on average taller than women; Americans aged 65+ around 2cm shorter than those age 18-64; etc.

**Socioeconomic variables and height’s relationship to *weight***

So, what happens when we use our best available proxy for socioeconomic status in our regression of weight on height, which I used last week to estimate an elasticity of weight on height of around 1.6? We find that as Nicholas suggested, higher socioeconomic status is associated with lower weight (which in today’s USA, can be seen as a proxy for better health).

But after controlling for all these things, weight is *still* proportional to height to the power of 1.6, not 2 and definitely not 2.5. So I’m pretty confident to say that what we’re seeing (with lower BMIs of tall people, by the standard calculation) is not a result of confounding socioeconomic status. In other words, last week’s conclusions stand up.

There’s some other interesting interpretive points from both these models about how these various factors interrelate but I’ll not go into them just now.

**Modelling code**

Here’s the code for fitting those two models and drawing those charts. Features to note include:

* use of Thomas Lumley’s survey package and in particular svyglm() so we get estimates and standard errors appropriate for the complex survey nature of the data
* use of the “quasi” family and explicitly chosen link and variance functions to fit my various model structure and interpetability needs
* using thick segments and no dot points, to focus the eye on confidence intervals rather than point estimates
* a bit of mucking around to make the different levels of factors (such as age) have visual location and colour in common while still retaining their natural sequencing
* explicit coding of the missing levels of variables with significant observations who otherwise would be dropped out; and use of transparency to avoid these estimates taking too much attention in the charts
* use of a function for some repetitive transformation work to turn model output into something needed for the chart

#---------------Prep for using in all models with these particular variables---------

#' Function to extract and relabel effects from a model. Very specific to today's problem.

get\_ready <- function(m){

# remove intercept:

d <- suppressWarnings(tidy(confint(m))[-1, ]) %>%

# remove height - as a continuous variable it doesn't work well

filter(!grepl("log(height)", .rownames, fixed = TRUE)) %>%

rename(variable = .rownames) %>%

mutate(var\_type = case\_when(

grepl("^education", variable) ~ "Education",

grepl("^hhinc", variable) ~ "Household Income",

grepl("^race", variable) ~ "Race",

grepl("^sex", variable) ~ "Sex",

grepl("^age", variable) ~ "Age"

)) %>%

mutate(variable = gsub("^education", "", variable),

variable = gsub("^hhinc", "", variable),

variable = gsub("^race", "", variable),

variable = gsub("^sex", "", variable),

variable = gsub("^age", "", variable)) %>%

mutate(var\_seq = 1:n())

}

#-----------------income and education impacting on height-------------

# First, is height really dependent on (or at least correlated with) income and education?

model13 <- svyglm(height ~ sex + race + hhinc + education + age, design = llcp\_svy,

family = quasi(link = "identity", variance = "mu"))

d13 <- get\_ready(model13)

p13 <- d13 %>% ggplot(aes(x = X2.5..\*100, xend = X97.5..\*100,

y = var\_seq, yend = var\_seq,

colour = var\_type,

alpha = I(variable == "(Missing)"))) +

geom\_vline(colour = "black", xintercept = 0)+

geom\_segment(size = 3) +

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

scale\_alpha\_manual(values = c(1, 0.2), guide = "none") +

labs(x = "95% confidence interval of average difference in height (cm)\nCompared to a white, 18-64yrs, non-Hispanic woman with Year 12 education, household income $25k-$35k",

colour = "",

y = "",

caption = the\_caption,

title = "Income and education are related to height even after controlling for sex, age and race",

subtitle = "The 'effects' of today's income and education are likely to actually be indicators of socioeconomic status while young,

impacting on both height and economic outcomes.") +

scale\_y\_continuous(breaks = d13$var\_seq, labels = d13$variable) +

theme(panel.grid.minor = element\_blank())

print(p13)

#---------------Height / weight---------

model14 <- svyglm(weight ~ log(height) + sex + race + hhinc + education + age,

design = llcp\_svy,

family = quasi(link = "log", variance = "mu"))

d14 <- get\_ready(model14)

p14 <- d14 %>% ggplot(aes(x = exp(X2.5..), xend = exp(X97.5..),

y = var\_seq, yend = var\_seq,

colour = var\_type,

alpha = I(variable == "(Missing)"))) +

geom\_segment(size = 3) +

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

scale\_alpha\_manual(values = c(1, 0.2), guide = "none") +

geom\_vline(colour = "black", xintercept = 1) +

annotate("text", x= 0.92, y = 9,

label = bquote(~weight%prop%height^1.6~"(not shown on chart)")

) +

labs(x = "95% confidence interval of impact on weight (expressed as a multiplier)\nCompared to a white non-Hispanic woman with Year 12 education, household income $25k-$35k",

colour = "",

y = "",

caption = the\_caption,

title = "Income and education are related to weight after controlling for sex, height and race",

subtitle = "The effects are complex, but the groups with most education and highest income have the lower weights.") +

scale\_y\_continuous(breaks = d14$var\_seq, labels = d14$variable) +

theme(panel.grid.minor = element\_blank())

print(p14)