When should meta-regressions be robust?

Abstract

The purpose of this paper is to analyse appeals to robustness in the literature on meta-regressions. I will show that there are several different motivations for conducting a sensitivity analysis and appealing to robustness. In particular, sensitivity analyses in economic meta-regressions do not always aim to increase the epistemic credibility of some empirical results. One must thus distinguish between epistemically relevant and subject-matter relevant robustness analyses. The epistemically relevant robustness invariably requires that some results are indeed robust. In such cases, successful robustness provides assurance that some inference does not depend on the auxiliary details. In certain different circumstances, however, the epistemic credibility of the exercise does not hinge on whether or not the results turn out to be robust. This is why one must distinguish between particular motivations of studying the robustness of estimates, and the fact of robustness, in other words whether some estimates actually turn out to be robust.

**1 Introduction**

Meta-analysts in economics often consider meta-regression as a particular method of studying the robustness of findings. ’Meta-analysis is a systematic analysis of the robustness of empirical results to replications’ (Doucouliagos and Paldam 2013). They thus regard meta-regression analysis as a way of analysing and making explicit the difference that the various specifications make to the estimates. ‘The central task of meta-analysis is to filter out systematic biases, largely due to misspecification and selection, already contained in economics research’ (Stanley and Doucouliagos 2012, p. 13). When Stanley and Jarrell (1989) introduced meta-regression into economics, they took it to be providing a method of analysing the problem of specification searches.

In economics, the problems in primary empirical studies commonly concern omitted variable bias and multicollinearity: Since the true data generating process is typically unknown, different control variables are commonly used in studies of the same empirical question in economics. Furthermore, the research design is often viewed as the author’s main contribution. The easier it has become to run regressions with different model specifications, the easier it has become for empirical economists to search for a model specification that yields a statistically significant result which is consistent with what they were predisposed to believe.

Edward Leamer’ (1978; 1983; 1985) solution to such problems consisted in requiring that all econometric results be tested for robustness. This advice has been influential in the practice of econometrics: Top economics journals now require robustness tests (sensitivity analysis) as a matter of course for any econometric paper. There are, however some arguments that diminish the force of inferential robustness that are not applicable for e.g., derivational robustness.[[1]](#footnote-1) In particular, if an econometric model is correct, then it is not necessarily the case that the model should be robust with respect to various irrelevant regressors in the specification. It is well known that omitting relevant variables leads to biased estimates and including irrelevant ones decreases the precision of the estimates. Including irrelevant variables may also introduce bias if they are, by chance, correlated with the included covariates.

Given the centrality of the problem, and the problems with the initial appeal to robustness, at least three further approaches have emerged that endeavour to solve the problem of choosing a model specification. First, the general-to-specific modelling and the associated automated specification searches try to find the correct specification by running a number of specification tests (see Campos, Ericsson, and Hendry 2005 for a review). Second, experimental and quasi-experimental designs are increasingly used in economics, and such approaches are taken to provide a solution to the problem of specification (Angrist and Pischke 2010). In an experimental design, sampling from large experimental and control groups mitigates the need for control variables. Consequently, the design and the effect studied are relatively easy to standardize. Third, meta-analysts may study the sensitivity of results to alternative model specifications by running meta-regressions.

By including multiple studies based on the same data set, a meta-regression would be better able to identify and to estimate the sensitivity of our empirical knowledge to model (mis)- specification. Aside from testing whether a given area of literature contains a genuine effect, estimating the sensitivity of our empirical knowledge to specific model specifications is the purpose of meta-regression analysis (Stanley 2002, p. 228).

Most economic meta-analyses conduct meta-regressions, that is, they provide a statistical regression analysis on a set of estimates derived from econometric primary studies. The primary studies themselves are usually based on observational data.

The purpose of this paper is to analyse appeals to robustness in the literature on meta-regressions. I will show that there are several different motivations for conducting a sensitivity analysis and appealing to robustness. In particular, sensitivity analyses in economic meta-regressions do not always aim to increase the epistemic credibility of some empirical results. One must thus distinguish between *epistemically relevant* and *subject-matter relevant robustness analyses*. The epistemically relevant robustness invariably requires that some results are indeed robust. In such cases, successful robustness provides assurance that some inference does not depend on the auxiliary details. In certain different circumstances, however, the epistemic credibility of the exercise does not hinge on whether or not the results turn out to be robust. This is why one must distinguish between particular motivations of studying the robustness of estimates, and the fact of robustness, in other words whether some estimates actually turn out to be robust. The main contribution of this paper is to show that there is widespread confusion because robustness is simultaneously taken to be an epistemically relevant tool for ‘quality control’ (Bown and Sutton 2010) and as a way of obtaining subject-matter relevant knowledge.[[2]](#footnote-2) We will soon see that instances of the confusion can be found in econometrics, in meta-regression, and in the methodological writings on them.

The rationale for testing the robustness of econometric results can be interpreted either as an attempt to show that the inferences do not depend on particular auxiliary assumptions or as a humble way of expressing the limits of our knowledge. Leamer, in particular, takes for granted that we can never know what the true DGP is in most econometric studies, and the ‘extreme bounds’ thus reflect the uncertainty of estimates (2010). Leamer’s (1978, 1985) extreme bounds analysis is one particular method of conducting sensitivity analysis on functional forms. The basic idea is to derive the relevant estimates for all the linear combinations of variables that could potentially be relevant. The extreme bounds refer to the most extreme parameter values obtained in this way. In this paper, I will refer to various ways of checking whether estimates differ as a result of a change in some research dimension as sensitivity analysis or robustness analysis. I thus take them to be methodological tools that have wider application than some particular method such as Leamer’s extreme bounds.

Experimental designs try find the correct specification by randomizing the data collection in the right kind of way, obviating the need to specify models on the basis of theoretical restrictions or statistical considerations. General-to-specific methodology tries hard to find the one true specification with statistical means. Meta-analysts study the sensitivity of results to alternative model specifications by running meta-regressions.

The four approaches are thus proposing different kinds of solutions. Such differences derive, in part, from the more or less explicit methodological disagreements between different empirical economists. The general-to-specific methodology is firmly grounded in the idea that there is a single best model specification and the task is to find it, while Leamer and many meta-regression specialists who emphasise robustness deny that one could ever find out what such a specification is. This is why I discuss a more general version of such an argument, which I will refer to as the ‘argument from superior methods’, in Section 3. This argument is analysed in econometrics and meta-regression. Before that, it is necessary to discuss the absence of agreeing estimates, ‘heterogeneity’, in meta-regression in Section 2. These two sections provide the necessary background for discussing a case on meta-analyses on alcohol own-price demand elasticities in Section 4. The purpose of presenting a case study is to show how one can distinguish between different motivations and kinds of inferential robustness.

**2 Heterogeneity and robustness**

Consider now one of Woodward’s critiques of inferential robustness:

There is also the difficulty that the observations used in the regression may not be drawn from a single common population or data generating process and/or that the regression parameters may not be homogenous across different countries. In all of these cases, finding that a coefficient on some variable is apparently robust under alternative specifications of a regression equation is consistent with the coefficient mis‐describing the actual influence of that variable (Woodward 2006).

Woodward is pointing to circumstances under which robustness of parameter coefficients fails to provide any epistemic benefits. Yet appeals to robustness are widespread in econometrics and also in economic meta-regressions:

What is the best MRA [meta-regression analysis] model of economic research and how can we decide? In our view this is the wrong question to ask. The right question is: what research dimensions are robust to MRA model specification? However, econometric training will cause many researchers to look for the ‘best’ model (Stanley and Doucouliagos 2012, p. 102).

These two quotations provide the motivation for this paper. I will argue that while Woodward is right about the fact that the robustness of estimates with respect to the specification is perfectly consistent with the coefficient mis-describing the influence of that variable, he gives a misleading impression that robustness of estimates should always be relevant for establishing the credibility of the inferences. Indeed, the point of the case study is to show that there is a huge number of research dimensions with respect to which meta-regression models may legitimately fail to be robust.

Stanley and Doucouliagos’ are explicitly criticizing the idea that the argument from superior methods should be applied to the choice of the MRA specification. Yet, as we will soon see, the very same authors are using the argument from superior methods in a later publication (Stanley and Doucouliagos 2016). I do not, however, charge them with inconsistency. The point is rather that, although the MRA specification ought to be robust with respect to some issues, this requirement should not apply to all research dimensions. For example, they would surely not be willing to say that results from a MRA with and without correcting for publication bias ought to be robust. If they are guilty of any methodological confusion, it is using the notion of robustness for both epistemically and subject-matter relevant purposes, without clearly indicating which interpretation one should adopt in the different contexts.

Such examples raise a more general question: What are the circumstances under which robustness enhances an econometric inference? Note that when Stanley (in the 2002 quote above) talks about studying the sensitivity of regressions, he is not committed to the idea that the results ought to be robust. The idea is rather that studying the sensitivity of results may be valuable even when the results are not robust: the point of conducting a meta-analysis is to provide information about *why* the estimates vary, information that is helpful in better appreciating the subject matter of the primary studies. Indeed, the term ‘meta-regression’ is only used for studies that run a meta-regression model in which the effects of different moderator variables on the estimates are explicitly studied. Merely calculating a weighted average for some set of estimates does not count as meta-regression even though it does count as meta-analysis.

Information on which research dimensions affect the estimates cannot be gleaned merely from primary studies because they are based on a given set of data and a given econometric method and model. Whether or not sensitivity or lack of sensitivity is epistemically crucial for the inferences thus depends on the details of the kind of inferential robustness that is in play.

Econometricians already implicitly and explicitly acknowledge the distinction between different methods of investigating robustness, and the fact of robustness because the methods are usually referred to as ‘sensitivity analysis’, and the primary reason for acceptable lack of factual robustness, ‘heterogeneity’ of data, has a well-defined meaning. Econometric estimates exhibit heterogeneity for various reasons. Sampling error and errors in measuring the variables are sufficient to explain why estimates are practically never identical from two econometric studies. In meta-analysis such within-study heterogeneity is supplemented with between-study heterogeneity that derives from several different sources: choice of data (individual or aggregate, different countries, different time-periods), functional form, method of regression (OLS, linear, generalized, GMM, AIDS, Rotterdam etc.), type of regression (time series, panel, cross section) author, data source, and so on.[[3]](#footnote-3) A meta-regression may show sensitivity or lack of sensitivity with respect to all these factors.

The epistemic problem lies in the difficulty of figuring out what kind of heterogeneity is prevalent. To see this, consider a debate that occurred some years ago in the journal Evaluation Review. There are standard statistical tests to determine the degree of heterogeneity of a set of estimates. The I2 (Higgins and Thompson 2002) statistic has become the most commonly used one.[[4]](#footnote-4) These tests provide estimates of the proportion of between-study heterogeneity in total variability.

The debate started when Patsopoulos et al. (2008) proposed algorithms for reducing heterogeneity by reducing I2. Their proposal consists of a sensitivity analysis with respect to the studies included in a meta-regression. Their algorithm thus tampers with the selection of studies into a meta-regression by seeking a set of studies that minimizes I2. Higgins (2008) responded by pointing out that it makes no sense to try to minimize I2 because it simply provides a descriptive statistic of the proportion of heterogeneity that is attributable to between-study variability. The total variability (τ2) is a point estimate of among-study variance of true effects. Since τ2 depends on within-study precisions (i.e., sample sizes of primary studies), so must I2. Thus, I2 can be large simply because the primary studies have large sample size and the within-study variation is small. This is why maximizing the within-study variability via selecting studies with *small* sample size is a likely consequence of minimizing I2. It is obviously rather easy to agree with Higgins. Rücker et al. (2008) recommend that the decision whether or not to pool results in meta-analysis should not be based solely on I2 because it typically increases with increasing sample size.

In economics, meta-analysts can often explain why there is heterogeneity between studies because several possible sources are usually apparent and testable (Nelson and Kennedy 2009). The problem is just that there are usually not enough data to obtain statistically significant estimates for dummies (in the meta-regression) that represent the variations in studies. However this problem is not specific to economics (see e.g., Prady et al. 2014), and if anything, economic meta-regressions usually have more studies with different specifications than other sciences for the reasons mentioned in the introduction. In theory, it is possible to distinguish between heterogeneity as a true characteristic of the population and as an artefact of the different methods used to estimate effect sizes in the primary studies (Lipsey 2009). But deciding whether heterogeneity derives from methodological differences or from the fact that the primary studies do not study exactly the same concept is difficult (Doucouliagos and Paldam 2013).

Meta-analysts distinguish between fixed (FEM) and random effects (REM) models.[[5]](#footnote-5) The former presuppose that the estimates from the various primary studies concern a given single effect. The idea is thus that all the studies concern the same population. Random effect models, in contrast, presume that each primary study is estimating the value in its own study population and these populations are randomly sampled from the overall population. The use of a fixed-effects model is usually recommended if there is low heterogeneity (variability) in the individual studies contributing to the meta-analysis, and random-effects models are to be used when the contributing studies have heterogeneous estimates of the outcome measure being combined.

Both kinds of models are often tested for robustness, but the interpretation of the results of sensitivity analyses is different. In a RE model the lack of robustness (of the meta-model) provides us with further information on how the effect depends on various factors. In a FEM model the lack of robustness means that the meta-model is fragile, and some authors (e.g., Nelson and Kennedy 2009) recommend ‘re-thinking the analysis’ in such a case. Knowing when to re-think the analysis presupposes that the analyst is able to say with sufficient confidence that the primary studies really study the same effect, and that the differences cannot be attributed to differences in e.g., the functional forms or the included variables. Thus, in principle the circumstances under which robustness provides the results with epistemic support are known.

However, the criteria for selecting the right kind of meta-model, FEM or REM (or ‘mixed’) are clearly a matter of some controversy among meta-analysts. While FE models are commonly taken to be reasonable when the purpose of the meta-analysis is providing an accurate estimate of the size of some effect in some population, they should not be used for out-of-sample predictions. In contrast, REM models are more useful for e.g., policy because they are taken to incorporate information on how the effect size depends on various circumstances. But the purpose of the investigation is not the only relevant factor for deciding which model to use. REM can only be used if the sample size for the meta-analysis is sufficiently large, while using a FEM is always possible. Sometimes meta-analysts appeal to the aforementioned formal tests of heterogeneity in order to choose between the two. However, given the arguments of Higgins and Rücker et al., methodologically aware meta-analysts seem to think that one should not proceed by choosing the appropriate model on the basis of a formal heterogeneity test. In addition to the problems pointed out by Higgins, heterogeneity in estimates may derive from many different sources, and meta-analysts cannot usually be sure whether it is because the primary studies are comparing ‘apples and oranges’, because of publication bias, or because the primary studies contain badly measured data.[[6]](#footnote-6)

Some meta-analysts claim that that they cannot decide qua statisticians whether the primary studies are theoretically sufficiently homogeneous to justify FEM (e.g., Borenstein et al. 2009). Higgins (2008) says that there is in principle no limit to how heterogeneous the data may be allowed to be for a meta-analysis. This choice is rather made on the basis of the subject matter studied, and thus depends on the case at hand. Given that it is widely acknowledged that data for meta-regressions in economics always feature considerable heterogeneity (e.g., Stanley 2013), the difficulty of figuring out the right kind of meta-regression model is always present. Since being able to decide whether FEM or REM is appropriate requires context-specific knowledge of the empirical issue at hand, meta-analysts must invest considerable energy into studying the primary literature before they can choose the statistical meta-model in an informed way.

The practical solution to such difficulties seems to be that some meta-analysts run both FE and RE models, thereby testing their meta-models for robustness (Greenhouse and Iyengar 2009). The very existence of such practices testifies that meta-analysts acknowledge the difficulty of finding out the true meta-analytic functional form. This, in turn, means that if different meta-regressions on the same effect reach different conclusions, they are likely to do so because different authors have different priors for how the results should look like. Given that the meta-analysts’ understanding of the primary literature affects their judgments concerning the methodology for the meta-regressions, it is not surprising that Stegenga (2011) gives examples of meta-analyses (in the biomedical sciences) that reach different conclusions about what is supposedly the same effect in the primary literature.

**3 The argument from superior methods**

Robert Hudson (2013) argues that it is not sensible to use information from a measuring device if the researchers know that there is superior device in the sense that it provides more accurate measurements. Thus, whether or not the two devices provide the same measurements, that is, whether the results are robust does not really matter because in the case of a discrepancy, the researchers should always use the measurement from the more reliable device. In this section, I will discuss similar arguments in econometrics. Before setting off, however, it is important to clarify that since Hudson’s argument concerns measurement robustness whereas this paper is about inferential robustness, I do not have anything to say about it in Hudson’s experimental context. Furthermore, of the three arguments presented below, only one (that from Hoover and Perez) can be taken to be framed as an argument against robustness. The point of discussing these arguments is to see the circumstances under which lack of robustness is not epistemically harmful: if we can use the argument from superior methods for evaluating two different estimates for the same concept, then there is a good reason for the estimates to be non-robust.

The core message of Hudson’s argument is that the scientists must always work as hard as they can to find the best method. There is some reason to be concerned about whether such principles are always followed in meta-analysis before resorting to robustness considerations, given that one finds comments like this: ‘We believe that sensitivity analyses, as currently performed, are usually an invitation to post hoc data dredging with few or no rules in the game. This reduces their inferential reliability’ (Patsopoulos, Evangelou, and Ioannidis 2009, p. 1740).

An introductory textbook on econometrics provides the following argument for never using a rather unsophisticated (OLS) estimator for standard errors:

If the homoscedasticity-only[[7]](#footnote-7) and heteroscedasticity-robust standard errors are the same, nothing is lost by using the heteroscedasticity-robust standard errors; If they differ, however, then you should use to more reliable ones that allow for heteroscedasticity. The simplest thing, then, is always to use heteroscedasticity-robust standard errors (Stock and Watson 2012, p. 203).

Heteroscedasticity implies that the estimates of standard errors will be biased unless they are estimated with estimators that take it into account. ‘Robust standard errors’ are thus estimated with an estimator that takes heteroscedasticity into account (White 1980).

In this case, we know that the heteroscedasticity-robust estimators are superior because we can show analytically that they yield the same correct results as the homoscedasticity-only estimators when the data are homoscedastic. Data for meta-regressions consist of pairs of estimates and their variances (or standard errors or sample sizes). They are always heteroscedastic because the primary studies have different sample sizes. The data would be homoscedastic only if the variances of all the primary studies were uniform – which is obviously not to be expected.[[8]](#footnote-8)

Hoover and Perez (2004) also appeal to the idea that one method is superior. They compare different methods of selecting a functional form for a regression: Leamer’s (1978, 1985) extreme bounds as applied by Levine & Renelt (1992), Sala-i-Martin’s (1997) method, and the general-to-specific algorithm that they developed (Hoover and Perez 1999). They show that the general-to-specific algorithm performs the best, in the sense of finding the ‘true’ functional form most often. Extreme bounds rejects too many variables and Sala-i-Martin’s method too few. The argument is based on being able to know what the true functional form is: the simulation generates the data as if it came from a data generating process that a particular functional form captures because the authors themselves fixed the ‘true’ functional form.

Stanley and Doucouliagos (2016) use this strategy to argue that their WLS-MRA (unrestricted weighted least squares meta-regression analysis) estimator performs better, in the sense of giving the correct estimates, than the standard FEM and REM models when there is publication selection bias. The critical part of the argument shows that WLS-MRA is practically as good as REM and FEM even when there is no publication selection bias.

Each of these arguments is valid, given what they endeavour to establish, viz., the superiority of a particular method (i.e., estimator), and given that the researchers are indeed able to demonstrate the superiority of one method. Yet, such arguments could be used to debunk using robustness analysis in general only if it were always possible to thus demonstrate the superiority of one method. Thus, although Hoover and Perez’ simulation argument shows that Leamer’s extreme bounds sensitivity analysis is not the best way of *finding* the correct model specification, it may still be the case that robustness of results with respect to the details of model specification is desirable.

Econometricians typically do not know what the true functional form should be. One perfectly good reason why the estimates should not be too sensitive to the specification is that, ‘if the estimates of the coefficients of interest are numerically similar across the alternative specifications, then this provides evidence that the estimates from your base specification [one based on economic theory and expert knowledge] are reliable. If, on the other hand, the estimates of the coefficients of interest change substantially across specifications, this often provides evidence that the original specification had omitted variable bias’ (Stock and Watson 2012, p. 275). Subtracting regressors from the true specification is obviously an instance in which the true specification should not be robust. But the true specification is usually robust with respect to adding irrelevant regressors: they do not bias the parameter estimates of the dependent variable. As long as such regressors do not determine the dependent variable (this is what being irrelevant means), they can even be correlated with the independent variables. Only the variances of the regressors will be affected.

Thus, even though robustness in Leamer’s sense is not an adequate guide to model specification, and even though it is neither necessary nor sufficient for a regressor to belong to the data-generating process (Hoover & Perez 2004, p. 790), it may still provide relevant information on whether the specification could be free from obvious weaknesses.

Leamer expresses the reason for sensitivity analyses as follows:

The point of the sensitivity analyses that I have been advocating begins with the admission that the historical data are compatible with countless alternative data-generating models. If there is one, the best we can do is to get close; we are never going to know it (Leamer 2010, p. 38).

He views his own method of extreme bounds as a way of expressing the degree of uncertainty that is appropriate for econometric results: given that the different model specifications generate different estimates, the honest thing to do is to report the extremes of what one can get from all possible linear combinations of the regressors.

Woodward (2006) and also Aldrich (2006) argue that the true functional form must be included in the set of specifications that are tested in a robustness analysis: However precise are the estimates from a mis-specified model, they will be biased. For the reasons already mentioned, however, the data for meta-analyses in economics typically consists of a mixture of estimates that derive from correctly and incorrectly specified models. Thus the very idea that there could be non-biased estimates from a meta-regression model does not seem sensible. Furthermore, given widespread heterogeneity, we can also ask: with respect to which population could estimates be non-biased?

**4 Alcohol own-price demand elasticities**

I will now briefly describe a case from economics. My aim is to illustrate how different judgments concerning which studies should be included in a meta-regression lead to differences in results. I am aware that, from the meta-analysts’ point of view, looking at a particular case to study heterogeneity and robustness may be pointless if one acknowledges that the methodology of meta-analysis itself does not tell us how closely the different studies focus on the same empirical concept. However, my aim was to find the easiest possible case in terms of the similarity of the empirical effect across the primary studies. In economics, there are several meta-analyses on at least foreign aid effectiveness, minimum wage legislation, and alcohol demand elasticities. I selected the last, hoping to be able to look at a theoretically simple case where the estimates could at least in principle concern the same concept. Here is what one meta-analysis on alcohol demand elasticities posits about this issue:

A frequent criticism of meta-analyses is that they combine ‘apples and oranges’; that is, they combine results from studies that differ in important ways. Our sample of studies is conceptually very well-integrated, but diverse in terms of units analyzed, treatments (i.e. size of tax or price change evaluated), outcome measures, settings, time and specific statistical models. On the last issue, a purist would argue that results from models with differing sets of covariates cannot be combined with the methods described here (moreover, methods to address this issue have not yet been developed). (Wagenaar, Salois, and Komro 2009, p. 187)

Alcohol own price elasticity is the percentage change in alcohol consumption that results from a one per cent increase in its price. The reason why there are many econometric studies on this particular parameter is that taxes on alcohol are widely considered to be an effective way of reducing its consumption and the associated harms (intoxication, traffic accidents due to drunk driving, liver cirrhosis etc.). Basic economic theory, the ‘law of demand’, implies that demand should decrease as price increases, so that the elasticities ought to be negative. However, policymakers also need quantitative information on elasticities because taxation is an effective way to curb consumption only if the elasticities are sufficiently large in absolute terms.

I will now describe the differences between four recently published meta-analyses on alcohol own-price elasticities: (Fogarty 2010; Gallet 2007; Nelson 2013b; Wagenaar, Salois, and Komro 2009). Let us start with a summary of the results. Most studies report separate elasticities for three types of beverage: beer, wine and spirits.[[9]](#footnote-9)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Beer | Wine | Spirits | Total alc. |
| Gallet | -.36 | -.70 | -.68 | -.5 |
| Fogarty | -.33 | -.55 | -.76 | na. |
| Wagenaar | -.46 | -.69 | -.80 | -.5 |
| Nelson trimmed samples FE | -.26 | -.34 | -.49 | -.46 |
| Nelson trimmed samples RE | -.35 | -.58 | -.64 | -.58 |
| Nelson publication bias corrected  | -.29 | -.46 | -.54 | -.49 |
| Alcohol own-price demand elasticities from four meta-analyses[[10]](#footnote-10) |

With some justification from their respective results, Wagenaar et al. conclude that taxation is an effective way to control alcohol consumption, and Nelson concludes that it isn’t. There are several kinds of reasons for the differences. First, there are differences in MRA methodology. Wagenaar et al. use a random effects model. They justify the choice by the results of a formal heterogeneity test that showed the data to be heterogeneous. Fogarty provides a comparison of FEM and REM, showing that FEM results yield lower elasticities for beer and spirits but vice versa for wine. Nelson also uses both FEM and REM, with consistently lower elasticities for FE (i.e., lower elasticities from studies that give more weight to studies with a large sample size). Gallet’s estimates are based on nonweighted medians. This is tantamount to giving the same weight to every sample size. Finally, Nelson is the only one to test for publication bias and attempt to correct for it.

Another class of differences derives from the inclusion criteria for primary studies used by the different meta-analyses. Gallet and Wagenaar et al. include studies that report tax rather than price elasticities, and Fogarty does not tell whether they are included. Gallet, Wagenaar et al. and (apparently) Fogarty include individual-level (survey) studies. Nelson’s basic search resulted in 578 primary studies. He excluded 223 studies on the grounds that they are alcohol harms studies or give individual-level data. He excluded an additional 58 studies that are in Gallet, Fogarty and Wagenaar et al.: 12 because the studies report tax elasticities, 9 report old data, 8 have missing standard errors 9 report only income elasticities, 10 provide brand/firm/survey data, 3 are duplicate studies, 7 use linear models or are excluded for other reasons. Nelson also has 64 studies not included in Fogarty or Gallet, and 135 not in Wagenaar et al. These are either brand new, or somewhat older but unpublished or published but the previous authors did not find them.

There are a large number of different reasons why different econometric regression models may yield different estimates even when the effect is as well-defined as the alcohol own-price elasticity. The estimates may be robust or fail to be robust with respect to the dataset chosen and the years it covers, the method of regression, and the functional form of the regression equation. Meta-regression estimates may be robust or fail to be robust with respect to the set of primary studies included, the meta-regression method used, in particular whether a FE or a RE model is used, the moderator variables included, and a large number of context-specific factors. Finally, given the widespread practice of conducting sensitivity analyses in primary studies, many of them contain several estimates. If the meta-analyst includes all the estimates in the meta-regression, this creates a massive independence problem in the data[[11]](#footnote-11). The meta-analyst thus has to decide how and how many estimates are selected.

In the case of alcohol own-price elasticities context-specific factors that may generate heterogeneity include the type of beverage (beer, wine or spirits or ‘alcohol’), whether the data are from individual-level or aggregate-level studies, whether the demand-estimations differentiate between on-premise and off-premise sales, whether the data are from monopoly sellers or not, whether the estimates are differentiated based on sex, age, or drinking habits (binge vs. moderate drinkers), whether compensated or non-compensated (Hicksian or Marshallian) elasticity estimates are studied, whether short-term or long-term elasticities are studied, whether brand-level data are included, whether alcohol tax elasticities are used as proxy for demand elasticities, and whether zero-observations (i.e., abstainers) are taken into account in the econometric model. In addition to these factors, ‘alcohol’ is measured in different ways in different studies: the alcohol elasticity estimates may vary because the beverage categories are not exactly the same in all countries (e.g., alcopops and/or cider may be together with beer or wine), the alcohol content of different drinks are based on imprecise measures that are different in different countries, and alcohol content may be calculated as ethanol-equivalent units or, more commonly, estimated from the per capita demand models.

I have listed such differences in some detail because the researchers have information on how most methodological choices will affect the results even before they conduct their meta-analysis and meta-regression. Presumably, however, readers who are not familiar with this particular primary literature do not have such knowledge. I will now try to show that even though the studies on alcohol own-price demand elasticities concern a well-defined concept, robustness with respect to the aforementioned differences is not to be expected. However, this should not be interpreted as an argument against the epistemic utility of robustness. The point is rather that only particular kinds of study characteristics are relevant for the epistemic value of robustness: When the argument from superior methods can be employed, one should indeed employ it. Whether or not one should use it depends on how secure the superiority judgments are. When it cannot be employed, one may have reason to resort to robustness. If one does resort to robustness for this reason, the motivation to do so is different from the case in which a researcher uses robustness for epistemic purposes and hopes the results to be similar (robust) because the difference concerns an auxiliary assumption, just like in derivational robustness (Kuorikoski, Lehtinen, and Marchionni 2010). Here the motivation is rather to express one’s uncertainty about what the correct model is.

Consider now the differences in inclusion criteria that were explicitly mentioned in our studies on alcohol.

*Tax elasticities*. If markets are competitive, economics suggests that tax pass-through rates should be unitary but Kenkel (2005) estimated pass-through rates of an alcohol tax hike in Alaska: 10 per cent rise in taxes resulted in about a 20 percent rise in prices. Thus, using tax elasticity as proxy for price elasticity are highly likely to inflate price elasticity estimates. However, one must also take into account the finding that the pass through rate may well be *less than one* in those cases in which taxes are reduced rather than increased.

*Individual-level data*. Elasticities from individual-level data tend to be higher than from aggregate data (but see Nelson 2013b). Using individual data may thus inflate price elasticity estimates compared to using only aggregate level data.

*Old data*. Old data tends to yield higher elasticities than new because they often used single-equation models. Furthermore, some authors (Fogarty) argue that there is a decreasing trend in alcohol demand elasticities. Using old data may thus inflate elasticities.

*Brand data*. Brand data tends to yield higher elasticities than other kinds of data for obvious reasons: Heineken is a substitute for Urquell. Using brand or firm data thus also inflates elasticities.

There is also a set of reasons for explainable differences in estimates that were not, or so it seems at least, used as exclusion criteria.

*Kinds of elasticities.* Compensated (Hicksian) demand elasticities are obviously smaller than non-compensated (Marshallian) ones (e.g., Fogarty 2010, p. 450). More generally, the elasticities depend on the exact set of controls used in the primary studies; cigarette smoking, etc.

*Off-premise vs. on-premise*. In absolute terms, data collected from off-premise sales tend to yield lower elasticities than from on-spot consumption, especially with spirits (Sousa 2014).

*Sex*. Women tend to have higher elasticities than men (e.g, Nelson 2014c).

*Age.* Young adults tend to have higher elasticities than old (e.g., Xuan et al. 2016).

*Abstainers taken into account.* Given that alcohol consumption data usually have a number of zero-observations due to abstainers, the better quality studies use methods that take this into account. Nelson excludes a study for this reason in a systematic review on heavy drinking (Nelson 2013a) but there is no information on this issue in the four papers referred to above.

*Short-run and long-run.* Long-run elasticities tend to be higher(Xu and Chaloupka 2011).

Frequent vs. non-frequent data: frequent data yield more elastic estimates.

It is perfectly reasonable to expect that the elasticities are not robust with respect to variations in all and any of these factors. From this point of view, it is not surprising that the meta-regressions did not reach similar conclusions. What is interesting about these differences is that, when the authors employed different inclusion criteria, they affect the elasticities in the same direction: Nelson excluded more studies that systematically inflate the elasticities.

These elasticities differ for perfectly good reasons, but in some other cases the differences point to differences in particular modelling assumptions that should not affect the inferences. As James Stock put it:

Far more attention needed to be paid to identification of the causal effect of interest, and econometric inference should not hinge on subsidiary modeling assumptions. First, credible identification of key causal effects or parameters, and second, statistical inference that is robust to subsidiary modeling assumptions have guided much of applied and theoretical econometric research (Stock 2010, p. 84).

The epistemic issue with robustness thus concerns the subsidiary (or ‘auxiliary’) assumptions just like in derivational robustness. A relatively easy way to recognize them is to ask: what kind of lack of robustness would make you distrust the meta-analytic (or more generally, econometric) inference? For example, meta-analysts usually assume that the sampling distribution for RE models is normal (see Greenhouse and Iyengar 2009). It is clear that if the estimates depend heavily on this distributional assumption, then the inferences based on the meta-regression are fragile in a way that matters for their epistemic credibility. Ruhm et al. (2012) provide an example from alcohol elasticity research in which the results ought to be robust but they are not. They find that the alcohol demand elasticities derived from the most commonly used dataset (ACCRA) are highly unstable: plausible specifications yield either no or extremely large responsiveness to price. Such demonstrations of failure of robustness diminish the credibility of primary as well as meta results from this particular dataset.

Thus far the philosophical discussion on robustness in econometrics has mostly concerned the issue of whether the estimates can legitimately depend on the functional form of or the included variables in the regression equation. But the specification of the functional form cannot really be an auxiliary issue because if it does not correspond to the data generating process, the estimates derived on its basis are known to be biased, and many econometricians are willing to say that such estimates are meaningless.

Let us now go back and ask which of these meta-analyses can best justify their methodological and inclusion/exclusion choices? Asking this question forces us to reflect on the purpose of meta-analyses. Possible purposes include summarising a given empirical literature on a particular parameter and showing how the results depend on various methodological choices by the primary researchers. From this latter point of view, it might well be acceptable to include every study on alcohol elasticities that reports the estimates reasonably clearly.[[12]](#footnote-12)

Given that some exclusion choices were made, summarizing this empirical research must have also been the aim of at least some of the alcohol-demand meta-analysts. Such an exercise makes most sense against the background of using taxes as a policy instrument. Thus, a health policy maker wants to know how much demand will decrease if taxes are raised by x per cent. From this perspective, it is clearly inadmissible to include individual-level, old, brand or tax elasticities as proxy for demand elasticities or several estimates from a single study, because estimates based on such data will provide misleading information. On the other hand, Gallet does not make any statements on how alcohol policy should be designed other than noting that since different groups of people have different elasticities, this should be taken into account. Fogarty and Gallet’s studies clearly tried to explain what kind of factors affect alcohol demand elasticities rather than come up with a ‘correct’ meta-estimate, whereas Wagenaar et al. and Nelson did. Wagenaar et al. did not conduct a meta-regression at all, and Nelson explicitly stated that he tries to summarize estimates in one of his alcohol demand meta-analyses (2014b, p. 181).

It thus seems that there is no such thing as a purpose-independent correct meta-analysis. The fact that Gallet and Fogarty’s exclusion criteria and estimates were different from Nelson’s is thus acceptable. However, given that Nelson and Wagenaar et al. indicated that they tried to obtain a correct meta-estimate, their exclusion criteria should have been the same. I am not sure whether it is acceptable to appeal to the argument from superior methods here. Although the exclusion criteria seem to indicate that Nelson’s study is superior, there are several other research dimensions to consider before one could conclude that Nelson’s study is superior in all respects: we must make choices on calculating an unweighted mean or using a RE or a FE model, taking into account publication bias, and dealing with correlated data explicitly.

Let us start with *publication bias*. It is commonly acknowledged to arise for two reasons: first, it is difficult to publish results that are statistically not significant. Second, if the background theory strongly supports certain kinds of results, then evidence which is inconsistent with the theory is less likely to be published. Stanley and Doucouliagos argue that ’the more strongly economists agree on a phenomenon, the greater its empirical distortion. For example, we find that own-price elasticities are often inflated.’ (2012, p. 145). This need not always be a matter of ideology. For example, consider a Master’s thesis that studied alcohol demand elasticities with the so-called Almost Ideal Demand System (AIDS) specification.[[13]](#footnote-13) This specification dictates consistency with the law of demand. When the author found *positive* elasticities for a certain group of beverages, he candidly admitted that this should not have happened, and decided not to report those elasticity estimates at all – they are meaningless given the AIDS specification. Positive elasticities are thus not reported simply because they may constitute proof of the falsity of the specification.

There are standard methods for detecting publication bias in meta-analysis such as funnel graphs. They are scatter plots that depict effect sizes against sample size. The usual interpretation is that an asymmetric funnel graph implies publication bias. The interpretive problem is that heterogeneity may also lead to funnel plot asymmetry if it induces a correlation between study sizes and intervention effects. Funnel plots could thus be symmetrical even in the presence of publication bias. If effect size estimates are related to standard errors, the random effects estimate will be pulled more towards findings from smaller studies than the fixed effect estimate will be (Sterne et al. 2011). This is perhaps one reason why FE models continue to be used even when the data are rather heterogeneous: random effects models can have undesirable consequences. In any case, given that significant publication bias has been found in alcohol elasticity research, it seems safe to conclude that it should have been investigated. Disregarding the problem of correlated data by including all estimates from the primary studies obviously biases the estimates. Ignoring the problem is clearly unacceptable.

The choice between REM and FEM is a more complicated issue. RE and FE models are weighted regressions which effectively homogenize the data.[[14]](#footnote-14) This is the reason why Nelson, like all minimally sophisticated meta-analysts, excluded primary studies that do not contain information on the variances of the estimates: the inverses of the variances are used as weights.

Several considerations speak against using FEM. First, given that the analyses of Gallet, Fogarty and Wagenaar et al. included such a wide variety of studies, it simply strains credulity to argue that there is a single population of studies that exhibit a fixed effect: alcohol demand elasticity. Second, as noted above, FE models can be legitimately used if the purpose of the meta-analysis is not to make conditional or out-of-sample predictions but it also strains credulity to argue that the sole purpose of Nelson and Wagenaar et al. is to summarize an empirical literature, given that the authors know that their studies will be used in public discussions concerning tax policy. On the other hand, if the publication bias is sufficiently severe, REM may yield rather misleading estimates. Yet, some meta-analysts working in different substantive fields such as medicine often recommend using only RE models because the data collection methods and the methodologies with which they are analysed always differ so much that one never comes to the judgment that the effect sizes (elasticities in economics) concern the same theoretical concept. One may use FEM simply because the sample size is not large enough for a REM. There is an obvious trade-off between inclusiveness and heterogeneity of data, and Nelson (2015b) is aware of it. Recall that alcohol was selected precisely because it seemed to provide a situation in which the primary studies concern a uniform theoretical concept. Thus, if FEM cannot be used here, it is doubtful whether it could be used anywhere in economics. But it seems to be widely used in all subject areas of economic meta-analyses. Is the choice between not conducting meta-regression at all if the sample sizes are too small, and using REM when they are not?

Whatever the verdict on which meta-model is best in this case, it is clear that the robustness or lack of robustness of the estimates with respect to the choice between REM and FEM has very little epistemic punch. It is perfectly understandable that the estimates may differ for this reason. The fact that two of the four authors provide estimates for both models gives another example of the motivation that Leamer had for conducting robustness analyses: given that the authors do not know what the correct model is, they provide estimates for both and let the readers decide which model to believe more strongly.

It may now be worthwhile to see the latest exchanges between Nelson and his critics. He published two systematic reviews on heavy and binge drinking in 2013 and 2015, respectively. These two papers do not conduct meta-analyses but rather provide qualitative reviews. The first of these (Nelson 2013a) took the analysis of Ayyagari et al. (2013) as a starting point. They argue that attention to drinker heterogeneity is critical in welfare analyses because higher taxes could well fail to reduce alcohol-related externalities in a substantial manner. The objective of Nelson’ study was to ‘test the robustness of this result by conducting a review of empirical studies on price (or tax) response of adult drinkers, ages 26 and older’ (p. 267). Xuan et al. (2016) criticized the latter study (Nelson 2015a) as follows.

Nelson’s approach failed to account for the heterogeneity of the effect sizes and sample variation due to different studies, and potential publication bias. Dr. Nelson stated that due to “diversity of models and results,” quantitative coefficient estimates for a meta-analysis were not collected. Dr. Nelson’s reference to “diversity of models and results” is the exact reason why quantitative meta-analysis is necessary to obtain a common effect estimate in the first place.

In other words, they criticized Nelson for failing to conduct a meta-analysis. Nelson (2016) responded that a qualitative approach was chosen ‘because primary studies report findings for numerous heterogeneous outcomes and employ different data sets and statistical methods’. This response is noteworthy because the 19 studies that Nelson selected for the review concerned price elasticities for binge drinking only – a concept that is much more specific than alcohol price elasticity. Given that the most inclusive study (Nelson 2013b) includes estimates from a FE model, Nelson’s choice of a systematic review reflects the judgement that individual data are fundamentally different from register data, and that individual data are so heterogeneous that estimates based on them they cannot be amalgamated at all.

Nelson (2016) also launched a scathing critique of Wagenaar et al. (2009):

Past meta-analyses of alcohol prices and alcohol-related harms are deeply flawed by incomplete data sets; lack of comparative quantitative measures, improper weighting of average effect-sizes, lack of controls for publication bias, lack of adjustments for non-independent observations, failure to properly employ meta-regression techniques.

Although I agree with Nelson that the meta-analyses on alcohol demand elasticities by the other authors seem to make various questionable assumptions, I would still hesitate to appeal to the argument from superior methods here. After all, this is an ongoing dispute and it would be foolhardy to ignore some choices that he has made. For example, in (Nelson 2015a) he excludes all studies that concerned young adults up to 26 years. This is very odd, given that it is known that young adults engage in binging more often than other age groups, and given that primary studies seem to indicate that young adults have higher elasticities than other age groups.[[15]](#footnote-15)

On the other hand, he does analyse young adults (ages 18-26) in (Nelson 2014c). More generally, Nelson has used the results of the same search of alcohol elasticity studies in August and September 2012 in a large number of publications, splitting up the sample in various different ways, depending on whether the individual meta-analyses concern the overall population or youth or women or heavy or binge drinkers, and on whether the data concern beer or wine and spirits, or affordability (Nelson 2014a). Such a research strategy raises the following question: Why not conduct one meta-regression on alcohol demand elasticities, regressing all these variables with the largest possible sample, and another that endeavours to find the best overall estimate, perhaps excluding individual-level data? From Nelson’s own point of view, the answer is clear: the overall elasticity provides one piece of information about elasticity, but since tax policy is an effective way of curbing alcohol consumption only if it is able to target heavy drinkers too, showing that heavy drinking men who drink beer are the least responsive to price is important if one wants to establish that taxes are not an effective instrument. Yet, this does not change the fact that Nelson could have included individual-level data in the most inclusive study, and could perhaps have obtained, from a single meta-regression, results similar to those that are now scattered in several publications.

The point of discussing such details about alcohol research is to show that the argument from superior methods fails to provide a clear choice of the primary studies. Meta-analysts’ disagreements are usually found in the exclusion criteria. The problem is precisely that deciding which studies are of sufficiently good quality to be eligible in a meta-analysis requires judgment, and different scholars make different judgments.[[16]](#footnote-16) The disagreement about the superior methods in the primary study level is transferred to the meta-level battles on inclusion and exclusion criteria.

Wagenaar et al. (2009) included 10 primary studies on ‘heavy alcohol use’, of which 2 are included in 19 studies in Nelson’s study (2013a) of heavy drinking! All these studies are indeed included in Nelson’s (2015a) study of binge drinking.

**Conclusions**

As Matt & Cook (2009) argue, the main contribution of meta-analysis is identifying the realm of application of a claim, that is, studying whether an association holds with specific populations of persons, settings, times and ways of varying the cause or measuring the effect; holds across different populations of people, settings, times, and ways of operationalizing a cause and effect; and can even be extrapolated to other populations of people, settings, times, causes, and effects than those studied to date. Given that the main goal is studying generalizability of empirical knowledge, there are a large number of methodological choices and context-specific issues with respect to which meta-regressions may legitimately fail to be robust. I analysed meta-regression in particular, in order to show how studying the robustness of results with respect to various issues may be valuable even though the analysis ends up with the judgment that the results are not robust. Meta-regression is thus a prime example of a method of studying robustness in such a way that the outcomes of the investigation need not be robust in order to be epistemically credible.

On the other hand, if an empirical result is indeed robust across different people, setting etc., then the robustness does provide relevant information on how widely the result can be applied. In other words, my aim is definitely not to downplay the importance of subject-matter relevant robustness of results. It is important for making judgments about the generalizability of the results. Furthermore, if the results in some field are highly robust to changes in methodological as well as substantial factors, then distinguishing between epistemically relevant and subject-matter relevant robustness does not have much importance because, as long as the results do not depend on some undetected confounder common to all the studies, we can apply the results, extrapolate and generalize etc. without much concern for their validity. It is just that economics is not a science where this could happen particularly often.

Distinguishing between different motivations for studying robustness and the fact of robustness or non-robustness allows us to see when estimates ought to be robust and when their non-robustness does not matter for the credibility of the empirical methods. There are three main motivations for studying the robustness of findings. One may want to express one’s ignorance of the correct method, to study how a regularity depends on substantively different circumstances or to show that the empirical estimates do not depend on irrelevant modelling choices. Lack of robustness is problematic if the differences in the studies concern substantively irrelevant characteristics.

**References**

Aldrich, J. H. (2006). When are inferences too fragile to be believed? *Journal of Economic Methodology,* V13, 161-177.

Angrist, J. D. & Pischke, J. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives,* 24, 3-30.

Ayyagari, P., Deb, P., Fletcher, J., Gallo, W. & Sindelar, J. L. (2013). Understanding heterogeneity in price elasticities in the demand for alcohol for older individuals. *Health Economics,* 22, 89-105.

Borenstein, M., Hedges, L. V., Higgins, J. P. T.& Rothstein, H. (2009). Criticisms of meta-analysis. *Introduction to Meta-Analysis* John Wiley & Sons.

Bown, M. J. & Sutton, A. J. (2010). Quality Control in Systematic Reviews and Meta-analyses. *European Journal of Vascular and Endovascular Surgery,* 40, 669-677.

Boyle, K. J., Kaul, S.& Parmeter, C. F. (2015). Meta-analysis: Econometric advances and new perspectives toward data synthesis and robustness. In R. J. Johnston, J. Rolfe and R. S. Rosenberger (Eds.), *The Economics of Non-Market Goods and Resources: Benefit Transfer of Environmental and Resource Values* Dordrect: Springer.

Campos, J., Ericsson, N.R., and Hendry, D.F. (2005) *General-to-Specific Modeling: An Overview and Selected Bibliography* Board of Governors of the Federal Reserve System International Finance Discussion Papers no. 838: .

Doucouliagos, H. (2016). Meta-regression analysis: Producing credible estimates from diverse evidence. *IZA world of labor,* 320, 1-10.

Doucouliagos, H. & Paldam, M. (2013). The Robust Result in Meta-analysis of Aid Effectiveness: A Response to Mekasha and Tarp. *The Journal of Development Studies,* 49, 584-587.

Fogarty, J. (2010). The demand for beer, wine and spirits: A survey of the literature. *Journal of Economic Surveys,* 24, 428-478.

Gallet, C. A. (2007). The demand for alcohol: a meta-analysis of elasticities. *Australian Journal of Agricultural and Resource Economics,* 51, 121-135.

Greenhouse, J. B. & Iyengar, S. (2009).
Sensitivity analysis and diagnostics. In H. M. Cooper, L. V. Hedges and J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* New York: Russell Sage Foundation.

Halme, P., Toivanen, T., Honkanen, M., Kotiaho, J. S., Mönkkönen, M. & Timonen, J. (2010). Flawed Meta-Analysis of Biodiversity Effects of Forest Management. *Conservation Biology,* 24, 1154-1156.

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International journal of epidemiology,* 37, 1158-1160.

Higgins, J. P. T. & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine,* 21, 1539-1558.

Hoover, K. D. & Perez, S. J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Economics Journal,* 2, 167-191.

Hoover, K. D. & Perez, S. J. (2004). Truth and robustness in cross-country growth regressions. *Oxford Bulletin of Economics and Statistics,* 66, 765-798.

Hudson, R. (2013) *Seeing Things: The Philosophy of Reliable Observation.* New York: Oxford University Press.

Kenkel, D. (2005). Are Alcohol Tax Hikes Fully Passed Through to Prices? Evidence from Alaska, American Economic Review, 95, 2, 273-277.

Kuorikoski, J., Lehtinen, A. & Marchionni, C. (2010). Economic modelling as robustness analysis. *British Journal for the Philosophy of Science,* 61, 541-567.

Leamer, E. E. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York ; Chichester: Wiley.

Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *The American Economic Review,* 73, 31-43.

Leamer, E. E. (1985). Sensitivity Analyses Would Help. *The American Economic Review,* 75, 308-313.

Leamer, E. E. (2010). Tantalus on the Road to Asymptopia. *The Journal of Economic Perspectives,* 24, 31-46.

Levine, R. & Renelt, D. (1992). A Sensitivity Analysis of Cross-Country Growth Regressions. *The American Economic Review,* 82, 942-963.

Lintonen, T., Karlsson, T., Nevalainen, J. & Konu, A. (2013). Alcohol policy changes and trends in adolescent drinking in Finland from 1981 to 2011. *Alcohol and alcoholism,* 48, 620-626.

Lipsey, M. W. (2009). Identifying interesting variables and analysis opportunities. In H. M. Cooper, L. V. Hedges and J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* New York: Russell Sage Foundation.

Matt, G. E. & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. M. Cooper, L. V. Hedges and J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* New York: Russell Sage Foundation.

Nelson, J. P. (2013a). Does heavy drinking by adults respond to higher alcohol prices and taxes? A survey and assessment. *Economic Analysis & Policy,* 43, 265-291.

Nelson, J. P. (2013b). Meta-analysis of alcohol price and income elasticities – with corrections for publication bias. *Health Economics Review,* 3, 1-10.

Nelson, J. P. (2013c). Robust Demand Elasticities for Wine and Distilled Spirits: Meta-Analysis with Corrections for Outliers and Publication Bias. *Journal of Wine Economics,* 8, 294-317.

Nelson, J. P. (2014a). Alcohol affordability and alcohol demand: cross-country trends and panel data estimates, 1975-2008. *Alcoholism: Clinical and Experimental Research,* 38, 1167-1175.

Nelson, J. P. (2014b). Estimating the price elasticity of beer: Meta-analysis of data with heterogeneity, dependence, and publication bias. *Journal of Health Economics,* 33, 180-187.

Nelson, J. P. (2014c). Gender differences in alcohol demand: a systematic review of the role of prices and taxes. *Health Economics,* 23, 1260-1280.

Nelson, J. P. (2015a). Binge drinking and alcohol prices: A systematic review of age-related results from econometric studies, natural experiments and field studies. *Health Economics Review,* 5, 1-13.

Nelson, J. P. (2015b). Meta-analysis: Statistical methods. In R. J. Johnston, J. Rolfe and R. S. Rosenberger (Eds.), *The Economics of Non-Market Goods and Resources: Benefit Transfer of Environmental and Resource Values* Dordrect: Springer.

Nelson, J. P. (2016). Reply to the critics on "binge drinking and alcohol prices". *Health Economics Review,* 6, 1-3.

Nelson, J. P. & Kennedy, P. E. (2009). The Use (and Abuse) of Meta-Analysis in Environmental and Natural Resource Economics: An Assessment. *Environ Resource Econ,* 42, 345-377.

Paillet, Y., Bergès, L., Hjältén, J., Ódor, P., Avon, C., Bernhardt-Römermann, M., Bijlsma, R., De Bruyn, L., Fuhr, M., Grandin, U., Kanka, R., Lundin, L., Luque, S., Magura, T., Matesanz, S., Mészáros, I., Sebastià, M. -., Schmidt, W., Standovár, T., Tóthmérész, B., Uotila, A., Valladares, F., Vellak, K. & Virtanen, R. (2010). Compromises in Data Selection in a Meta-Analysis of Biodiversity in Managed and Unmanaged Forests: Response to Halme et al. *Conservation Biology,* 24, 1157-1160.

Patsopoulos, N. A., Evangelou, E. & Ioannidis, J. P. (2008). Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. *International journal of epidemiology,* 37, 1148-1157.

Patsopoulos, N. A., Evangelou, E. & Ioannidis, J. P. (2009). Heterogeneous views on heterogeneity. *International journal of epidemiology,* 38, 1740-1742.

Prady, S. L., Burch, J., Crouch, S. & MacPherson, H. (2014). Problems caused by heterogeneity in meta-analysis: a case study of acupuncture trials. *Acupuncture in Medicine,* 32, 56-61.

Rhodes, W. (2012). Meta-analysis: An introduction using regression models. *Evaluation Review,* 36, 24-71.

Rücker, G., Schwarzer, G., Carpenter, J. R. & Schumacher, M. (2008). Undue reliance on I2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology,* 8, 79-88.

Ruhm, C. J., Jones, A. S., McGeary, K. A., Kerr, W. C., Terza, J. V., Greenfield, T. K. & Pandian, R. S. (2012). What U.S. data should be used to measure the price elasticity of demand for alcohol? *Journal of Health Economics,* 31, 851-862.

Sala-i-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review,* 87, 178-183.

Sousa, J. (2014) *Estimation of Price Elasticities of Demand for Alcohol in the United Kingdom* HMRC working paper 16, <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/387513/HMRC_WorkingPaper_16_Alcohol_elasticities_final.pdf>: HMRC Working paper 16:.

Stanley, T. D. (2002). Meta-Analysis]: Response from T.D. Stanley. *The Journal of Economic Perspectives,* 16, 227-229.

Stanley, T. D. (2013). Does economics add up? An introduction to meta-regression analysis. *European Journal of Economics and Economic Policies,* 207-220.

Stanley, T. D. and Doucouliagos, H. (2012) *Meta-Regression Analysis in Economics and Business.* 1st ed. New York: Routledge.

Stanley, T. D. & Doucouliagos, H. (2016). Neither fixed nor random: weighted least squares meta-regression. *Research Synthesis Methods,* 1-24.

Stanley, T. D. & Jarrell, S. B. (1989). Meta-regression analysis: A quantitative method of literature surveys. *Journal of Economic Surveys,* 3, 161-170.

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences,* 42, 497-507.

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., R\ucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D. & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ,* 343, .

Stock, J. H. (2010). The Other Transformation in Econometric Practice: Robust Tools for Inference. *The Journal of Economic Perspectives,* 24, 83-94.

Stock, J. H. and Watson, M. W. (2012) *Introduction to Econometrics.* , global ed. Boston, Mass. ; London: Pearson.

Wagenaar, A. C., Salois, M. J. & Komro, K. A. (2009). Effects of beverage alcohol price and tax levels on drinking: a meta-analysis of 1003 estimates from 112 studies. 104, 179-190.

White, H. (1980). A heteroscedasticity consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica,* 48, 817-838.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology,* 13, 219-240.

Xu, X. & Chaloupka, F. J. (2011). The Effects of Prices on Alcohol Use and its Consequences. *Alcohol Research & Health,* 34, 236-245.

Xuan, Z., Babor, T. F., Naimi, T. S., Blanchette, J. G. & Chaloupka, F. J. (2016). Comment on "binge drinking and alcohol prices". *Health Economics Review,* 6, 1-4.

1. This terminology is used as in Woodward (2006): inferential robustness means the robustness of statistical models and inferences, and derivational robustness concerns theoretical modelling. [↑](#footnote-ref-1)
2. Boyle et al. (2015, p. 385) seem to be at least vaguely aware of the distinction: ‘One should not jump to the conclusion that the presence or absence of robustness implies desirable or undesirable outcomes’. [↑](#footnote-ref-2)
3. Some discussions on meta-regression take ‘heterogeneity’ to refer merely to such between-study differences. [↑](#footnote-ref-3)
4. It makes a correction to Cohcran’s Q which makes it independent of the number of studies in the meta-sample. [↑](#footnote-ref-4)
5. Their properties, similarities and differences are discussed in a large number of papers. Rhodes (2012) provides a particularly clear introduction. [↑](#footnote-ref-5)
6. Doucouliagos (2016, p. 8) thus recommends to present only meta-regression findings that are robust to all justifiable variation in how effects are measured. [↑](#footnote-ref-6)
7. Homoscedasticity means that the variance of the error term ui given the independent variable Xi across the observations i=1,…,n is a constant E(ui|Xi=x)=σu that does not depend on i. Otherwise the error term is heteroscedastic. [↑](#footnote-ref-7)
8. Although using heteroscedasticity-robust estimators is now just a matter of a click of a button, this was not the case 20 years ago. This means that, old primary studies are likely to have too high precision. [↑](#footnote-ref-8)
9. Estimates from Gallet and Fogarty refer to unweighted medians. Fogarty also reports means that are slightly higher than the ones reported here. The elasticities from Nelson are weighted means, and those from Wagenaar et al. mean values for standardized effect sizes r. Nelson provides essentially the same estimates for wine and spirits in (Nelson 2013c). In Nelson (2014b) he provides a beer elasticity estimate for a full sample (i.e., not based on trimmed samples) FE: -.23. Apparently the RE estimate is the same (-.35) with trimmed and full samples. Nelson presents a summary of his and other meta-analysts’ results in all these publications. I chose to say that the results from Wagenaar et al. are ‘weighted means’ because the estimates are weighed by precision, and this is the standard understanding of what ‘weighted’ means. Nelson’s tables, however, present them as ‘unweighted’ means, apparently because they use, according to Nelson, 105 estimates for beer from 47 studies (p. 182). Nelson uses one method (the cumulative one) to correct for publication bias in (Nelson 2013b) and two further ones in (2014b). In (2014b) he reports -.20 for FE and -.23 for RE obtained with ‘trim and fill’, -.14 to -.19 with ‘FAT-PET’ FE metaregressions, and -.23 and -.30 from cumulative FE and RE, respectively. [↑](#footnote-ref-9)
10. A reviewer to an earlier version of this paper complained that Gallet’s estimates are not from meta-analysis at all (despite the title of the paper), that Fogarty’s estimates are not based on meta-regression, that Wagenaar et al. use 292 studies instead of the 1003 indicated in the title of their paper and so on. While there may be justification to all of these criticisms, it is not the purpose of this paper to figure out who is right. For the purposes of this paper, what matters is that there are different numbers in the table. The question is whether it makes sense to put the results of these studies into the same table. My impression is that the reviewer thinks that only Nelson’s studies deserve to be mentioned as meta-regressions on alcohol own-price elasticities. Even if that were to be a correct judgment, the others do claim to conduct meta-analyses, and that is sufficient for being included in this table. [↑](#footnote-ref-10)
11. To appreciate the magnitude of the problem consider that, Gallet, for example, had 1172 estimated price elasticities from 132 primary studies. He ignored the problem by including all of them in his meta-regression. [↑](#footnote-ref-11)
12. Various primary studies provide relevant information on how using tax elasticities as proxy for demand elasticities, individual-level, old, and brand or firm data, affect the elasticities. In this sense, the meta-analysis is not necessary. [↑](#footnote-ref-12)
13. I cannot provide further information on this Master’s thesis because this would put my anonymity at risk. [↑](#footnote-ref-13)
14. The fact that meta-analyses in economics usually have access to variance information explains why they do not usually need to use heteroscedasticity-consistent estimators: they can deal with the problem with weighted regression instead. [↑](#footnote-ref-14)
15. However, Lintonen et al. (2013) provide some rather bizarre results on adolescent’s elasticities; their demand for alcohol sometimes plummets after major reductions in alcohol taxes or increasing availability. [↑](#footnote-ref-15)
16. Nelson clearly exercises such judgment but in some other fields the culture of conducting meta-analyses seems different. Consider, for example, the response by Paillet et al. (2010) to the critique (Halme et al. 2010) that they had included bad quality studies and several estimates from the same primary studies in their meta-analysis: ‘If we were to subjectively exclude from our analyses the papers we considered flawed (for methodological, statistical, or other reasons), our objectivity could be questioned. We therefore chose to include these studies in our analyses.’ Such a response reflects a refusal to consider the argument from superior methods at all. [↑](#footnote-ref-16)