Analysis: Association between sodium consumption and life expectancy at age 30

Kellie Ottoboni

June 22, 2016

1 Introduction

If salt is bad for you, then we'd expect

1. ceteris paribus, salt would be negatively associated with life expectancy.
2. salt would be a strong predictor of life expectancy after accounting for other important factors.

We assess these points in a variety of ways below. We find that neither point is satisified. Sodium consumption is not significantly negatively correlated with life expectancy, and it isn't a strong predictor of life expectancy after accounting for alcohol consumption, cigarette consumption, and per capita GDP.

On the other hand, we'd expect these other three factors to be more strongly associated with life expectancy. Alcohol and cigarettes are known to cause negative health outcomes, both short-term and long-term. Indeed, our analyses confirm these hypotheses: alcohol and smoking are negatively associated with and are strong predictors of life expectancy. Thus, we have confidence that the methods we explore below are able to detect an effect when the effect is actually present.

2 Data cleaning

Data preprocessing includes removing unwanted variables (we keep alcohol consumption and per capita cigarette consumption only) and imputing the missing values of the predictors. In particular:

- We merge two data sources. One contains the economic variables, life expectancy, alcohol, and sodium consumption. The other contains the annual number of cigarettes smoked per capita.
- There is no data for overall alcohol consumption levels in 2010, so we impute them. A country's population average alcohol consumption is estimated by taking a weighted average of the country's male and female alcohol consumption, where weights are the proportion of males and females in the population.
- On the other hand, there is no sex-specific alcohol consumption data for 1990. For each country, we use the proportion of alcohol consumption in 2010 attributable to males and females to estimate the male and female sex-specific alcohol consumption in 1990, respectively, using the formula

$$\text{Male ETOH in 1990} \approx \text{Overall ETOH in 1990} \times \frac{\text{Male ETOH in 2010}}{\text{Imputed overall ETOH in 2010}}$$

  A similar formula applies to females.

- Taiwan is missing all of its alcohol data, so we impute them. We impute the missing 2010 sex-specific alcohol consumption for Taiwan using a linear regression of sex-specific alcohol consumption on male and female life expectancy, male and female salt consumption, per capita GDP, and per capita annual cigarette consumption. Then we impute population average alcohol consumption by taking a weighted average of male and female alcohol consumption, using the proportion of the population that is male and female as the weights. We impute the missing 1990 overall alcohol consumption for Taiwan in the same way, regressing overall alcohol consumption on the predictors. Then we use the proportion of alcohol consumption in 2010 attributable to males and females to estimate the male and female sex-specific alcohol consumption in 1990, respectively.

3 Correlations

3.1 Raw Correlations

First, we report the raw correlations between life expectancy at age 30 and each of the predictors. This does not control for correlations between predictors. It is the crudest measure. Table 1 shows these correlations. For alcohol, smoking, and per capita GDP, the sign of the correlations are consistent across all four cuts of the data and are consistent with our expectations. Sodium is negatively associated with life expectancy for males but positively associated for females.

	Male		Female	
	Differenced	Cross-sectional	Differenced	Cross-sectional
etoh	-0.73	-0.36	-0.55	-0.01
smoking	-0.41	-0.11	-0.2	-0.14
Na	-0.41	-0.11	0.04	0.1

Table 1: Raw correlations with life expectancy at age 30.

3.2 Permutation Tests

We'd like to assess how significantly correlated each variable is with life expectancy, accounting for the other confounding variables. We do this by conducting stratified permutation tests. This data is observational, not randomized, so observations are not exchangeable without some assumptions. We use the idea of a balancing score from Rosenbaum and Rubin (1983). A balancing score is a function $b(X)$ of the covariates X that makes treatment assignment T conditionally independent of covariates:

$$X \perp\!\!\!\perp T \mid b(X).$$

If we stratify on a balancing score, we may permute observations within strata.

For this to yield a valid test, we essentially need to model the balancing score correctly. This relies on the assumption that we account for all causal variables in X. We don't believe that we have included everything that may affect life expectancy at age 30. However, we would like to proceed with the method. For robustness, we try two different balancing scores: the generalized propensity score and the prognostic score. If results using the two balancing scores are qualitatively similar, then we have some assurance that the assumption may not be badly violated.

3.3 Generalized Propensity Score

Hirano and Imbens (2004) introduce the idea of a *generalized propensity score* for continuous treatments. They assume *weak unconfoundedness*:

Assumption 1 (Weak unconfoundedness)**.** $Y(t) \perp\!\!\!\perp T \mid X$ *for all* $t \in T$.

This essentially means that we've accounted for all potential confounders in X. Then, they define the generalized propensity score:

Definition 1 (Generalized propensity score)**.** *Let $r(t, x)$ be the conditional density of the treatment given the covariates:*

$$r(t, x) = f_{T|X}(t|x).$$

Then the generalized propensity score is $R = r(T, X)$.

The generalized propensity score has the following mechanical property:

$$X \perp\!\!\!\perp \mathbb{I}\{T = t\}|r(t, X)$$

This means that treatment assignment is unconfounded given the generalized propensity score. Practically speaking, within strata of units with the same $r(T, X)$, the level of treatment for each country amounts to an arbitrary labelling. This suggests a procedure to conduct a permutation test for the effect of treatment:

1. Estimate the generalized propensity score using a linear probability model for treatment given covariates. Formally, $T \sim N(X\beta, \sigma^2)$. Then, for unit i, the GPS estimate is

$$R_i = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(T_i - x_i'\hat{\beta})^2}{2\hat{\sigma}^2}\right)$$

   Note, we could fit distributions other than the normal by maximum likelihood if desired.

2. Stratify observations on the estimated GPS. We use k-means to partition the GPS into bins. We choose the number of strata k to have smallest average silhouette width; this will be data-dependent.

3. Do a stratified permutation test for correlation between treatment and outcomes.

3.4 Prognostic Score

We do the same thing using the prognostic score, the predicted outcome using no information on the treatment (Hansen, 2008). We use CART to predict the outcome and create strata.

3.5 Balance Tests

There are several ways we may run into trouble using this method. First, we must estimate the generalized propensity score and prognostic score. If either model is misspecified (which in this case, it likely is), then the score which we use to stratify may result in grouping the wrong countries. Second, we must choose how to stratify. We do this in a somewhat arbitrary way. For the generalized propensity score, we divide the countries by quintiles of the estimated scores. For the prognostic score, we use the leaves of the regression tree that we use to estimate the score. Both methods have arbitrary tuning parameters.

Both scores are balancing scores, so that strata formed using them should be "balanced" on all confounding variables. That is, there should be no correlation between the treatment level and any of the covariates within strata. To test this, we use the same stratified permutation test of correlation that we use to test the association between treatment and outcome, except here we apply it to treatment and each covariate.

Tables 2–5 show the results of these balance tests, for both male and female data and both balancing scores. For the most part, the stratification acheives balance: most of the p-values are large. However, when using the generalized propensity score to stratify and test for association between salt and life expectancy, the strata do not balance the covariates. The p-values for association between sodium consumption and alcohol and smoking are near or exactly zero for both males and females, and the p-value for per capita GDP is zero for males. Thus, the permutation tests based on this stratification method may not be valid and their results should be taken with a grain of salt (pun intended).

3.6 Results

Figure 1 shows the results for males using the differenced data (2010-1990). Based on the tests using the generalized propensity score stratification, sodium, and alcohol are all significantly associated with life expectancy at age 30. The p-values for association are all above 10% when considering the tests using prognostic score stratification.

Figure 2 shows the results for females using the differenced data (2010-1990). At the 5% level, no variable is significantly associated with life expectancy using either stratification method. Based on prognostic score stratification, alcohol has a p-value less than 10% and based on generalized propensity score stratification, sodium has a p-value less than 10%. Recall that for females, the correlation between sodium and life expectancy is actually positive; thus this result says that sodium has a nonsignificant association with *increases* in life expectancy at age 30.

The p-values do not, in general, show close agreement between the two stratification methods. Stratification on the generalized propensity score tends to give smaller p-values than stratification on prognostic scores. Aside from this, the results suggest that after controlling for the other variables in the dataset, we did not identify a statistically significant association between increases in sodium consumption and decreases in life expectancy. This does not give evidence for the "salt is bad" hypothesis.

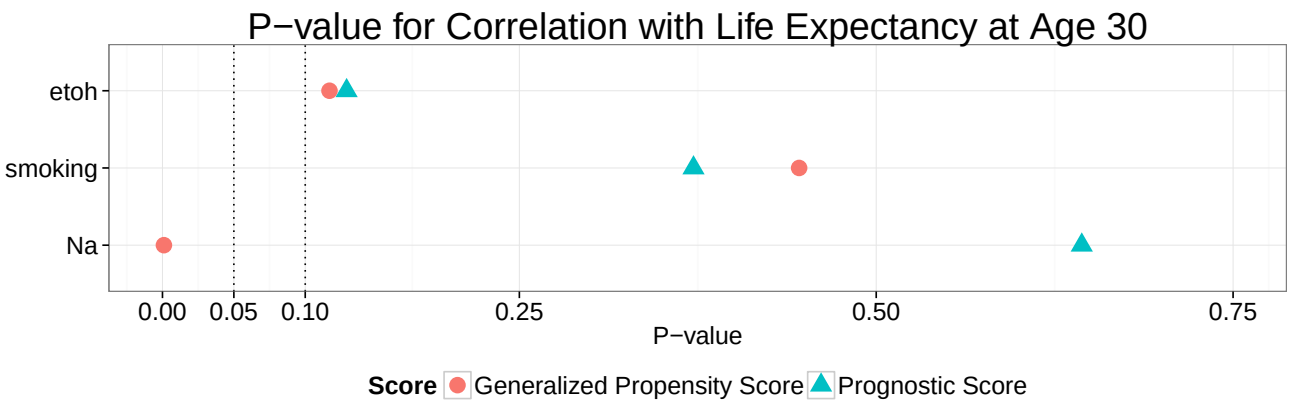

Figure 1: P-values for males, differenced (2010-1990).

	etoh	smoking	Na
etoh		0.42	0.14
smoking	0.28		0.17
Na	0.00	0.04	

Table 2: P-values from balance tests within strata formed by GPS, for male data. Rows are the notional treatment and columns are covariates.

	etoh	smoking	Na
etoh		0.63	0.75
smoking	0.62		0.14
Na	0.71	0.18	

Table 3: P-values from balance tests within strata formed by prognostic score, for male data. Rows are the notional treatment and columns are covariates.

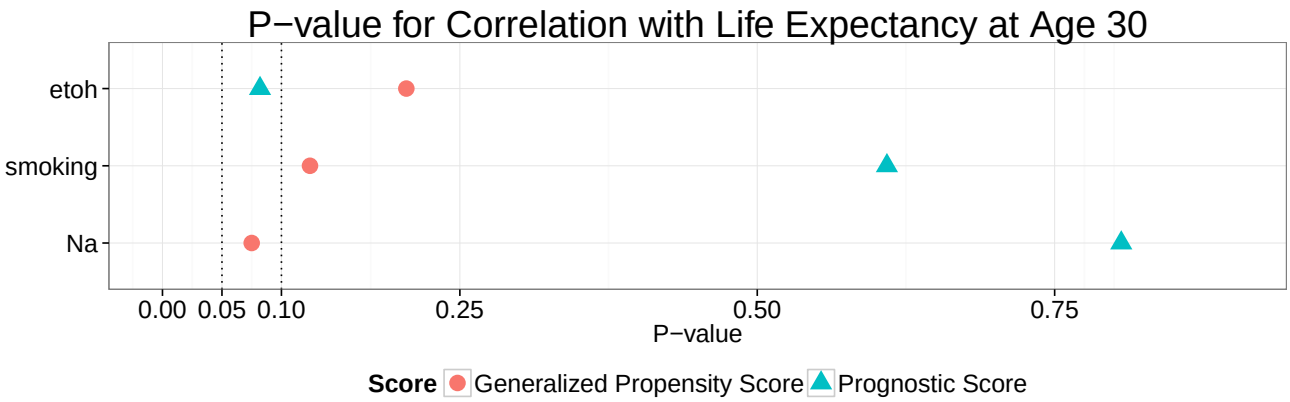

Figure 2: P-values for females, differenced (2010-1990).

	etoh	smoking	Na
etoh		0.66	0.98
smoking	0.36		0.02
Na	0.00	0.00	

Table 4: P-values from balance tests within strata formed by GPS, for female data. Rows are the notional treatment and columns are covariates.

	etoh	smoking	Na
etoh		0.14	0.85
smoking	0.14		0.01
Na	0.84	0.01	

Table 5: P-values from balance tests within strata formed by prognostic score, for female data. Rows are the notional treatment and columns are covariates.

4 Variable importance

In his seminal paper on random forests, Leo Breiman introduced a variable importance measure based on permutations. The idea is that if a variable is important in a regression, then perturbing it will worsen the predictive performance. On the other hand, if we perturb a variable and the predictions remain relatively good, then the variable is not important to the model. We perturb the variables by permuting them, breaking the association between the feature and all other variables and the outcome.

Though the idea of permutation variable importance came from random forests, it is sufficiently general that it can apply to any predictive model. We can use several metrics for this: we report the "absolute importance," the original root mean squared error (RMSE) minus average permuted RMSE, and the "normalized importance," the absolute importance divided by the original RMSE. Large values indicate more importance.

4.1 Results

We compare several prediction methods: CART, random forests, OLS, and OLS with interaction terms. We run each function using default tuning parameters. Alcohol is consistently the most important predictor for males, while neither smoking nor sodium is consistently least important across the methods. In the female data, alcohol is the most important predictor and smoking is the least important.

	Absolute			Normalized		
	etoh	smoking	Na	etoh	smoking	Na
CART	0.695	0	0.037	0.472	0	0.025
Random Forest	0.767	0.368	0.329	1.039	0.499	0.446
OLS	0.889	0.057	0	0.716	0.046	0
OLS with interactions	1.13	0.161	0.171	1.066	0.152	0.161

Table 6: Variable importance for males, differenced (2010-1990)

	Absolute			Normalized		
	etoh	smoking	Na	etoh	smoking	Na
CART	0.337	0	0.197	0.355	0	0.208
Random Forest	0.506	0.228	0.267	0.991	0.446	0.523
OLS	0.475	0.002	0.129	0.52	0.002	0.141
OLS with interactions	0.662	0.121	0.204	0.807	0.148	0.249

Table 7: Variable importance for females, differenced (2010-1990)

4.2 Alternative Variable Importance for Random Forests

The default variable importance measure in `randomForest` is different from what we described above. Instead of permuting the predictor of interest and evaluating the full model on this perturbed dataset, the function does the permutation procedure for each individual tree used in growing the random forest. The final importance measure is the average decrease in mean squared error across trees in the forest.

Figure 3 shows these measures when we train a random forest with the default R settings. Note that we report root mean squared error for the overall random forest predictor in the previous section, so the scales of the two measures are not comparable. The order of variable importance is the same for both the male and female datasets: alcohol is substantially more important than smoking and sodium, but smoking is slightly more important than sodium.

Figure 4 shows the partial dependency plots for each variable. We use the random forest to estimate the marginal effect of each variable on changes in life expectancy at age 30 from 1990 to 2010, holding the

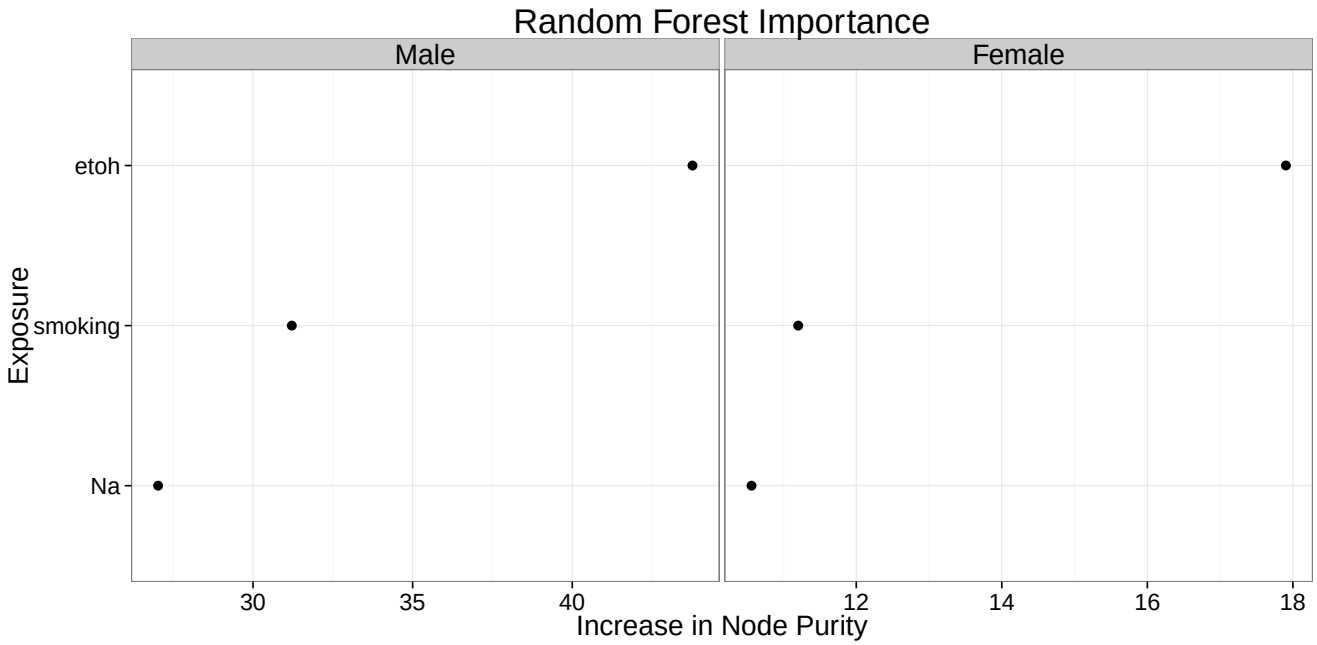

Figure 3: Variable importance measure from the extttrandomForest package.

other variables fixed at their observed values. These figures suggest that each variable has a large effect (steep positive or negative slope, graphically) only locally. For instance, decreases in cigarettes per capita seem to have no association with life expectancy, but any increase in cigarettes per capita has a negative association with life expectancy. The findings for sodium are less clear. It appears that for males, only increases in sodium consumption over 0.2 grams per day are associated with decreases in life expectancy. For females, any increase in sodium consumption is associated with increased life expectancy. Compared to the other variables, the magnitude of changes in life expectancy according to sodium consumption are smaller; the curves for alcohol and per capita GDP fill the entire plot frame, while the curves for sodium span half of that range.

We've considered variable importance in several manners, and all confirm that sodium is a less important predictor than alcohol. Because the dataset is small and parts of the covariate space may be sparse, we should be careful not to place too much weight in these results. Furthermore, one must bear in mind that importance may depend on the choice of model. The four models that we've considered here show similar patterns of variable importance, giving us confidence in the relative magnitudes of each variable's predictive power.

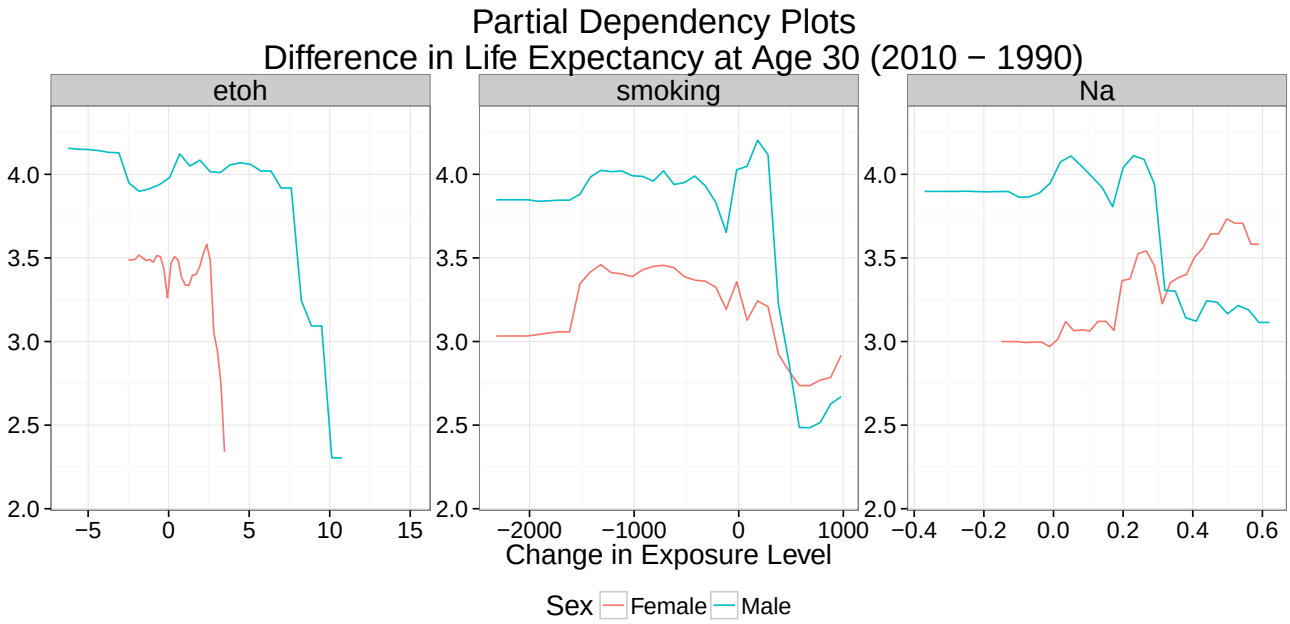

Figure 4: Partial dependency plots for each of the four variables. They display the estimated marginal effect of the variable on change in life expectancy at age 30 from 1990 to 2010.

5 Linear models

First, we look at the data cross-sectionally. To account for unobserved confounding variables, we add a fixed effect term for each country. The key assumption (aside from the usual linear model/response schedule assumptions) is that these country-specific effects do not change over time; all that changes is the treatment over time. This amounts to estimating a different intercept for each country. We need to correct the standard errors when we estimate this way by using the Huber-White Sandwich estimator. Let the subscript i denote country and t denote time period. The model is

$$Y_{it} = \alpha_i + \lambda_t + \beta_1 \text{ETOH}_{it} + \beta_2 \text{SMOKE}_{it} + \gamma \text{Na}_{it} + \varepsilon_{it}$$

α_i is the fixed effect for country i. λ_t is a time effect.

We're interested in the parameter γ. If salt were detrimental to health, we'd expect γ to be negative.

It is numerically equivalent to estimate the model using the differenced data, subtracting each country's data point in 1990 from its data in 2010. The model then becomes

$$\Delta Y_{it} = \Delta\lambda_t + \beta_1 \Delta\text{ETOH}_i + \beta_2 \Delta\text{SMOKE}_i + \gamma \Delta\text{Na}_i + \Delta\varepsilon_i$$

This eliminates the need to estimate fixed effects, making estimation more efficient. However, this model assumes constant additive treatment effects. In particular,

1. Any country that increases its Na+ consumption by 1 unit will shift its life expectancy by γ.
2. Trends in life expectancy would be the same in all countries in the absence of treatment (salt consumption).

We report usual OLS standard errors as well as Huber-White robust standard errors.

Indeed, in all model specifications that we run for the male data, salt has a negative effect on life expectancy at age 30. In the full model for males (column (6) in Table 8), a one unit increase in sodium consumption is associated with a decrease in life expectancy of 0.024 years. The sign of the association is

positive for females: after controlling for alcohol and smoking, (column (6) in Table 9), a one unit increase in sodium consumption is associated with a 2.18 year increase in life expectancy (significant at 10% level). The sign of β_1 and β_2, the coefficients corresponding to alcohol and smoking, are negative for both males and females, and the coefficient for alcohol is significant at the 1% level.

Table 8: Male, differenced regressions (2010-1990)

	OLS	Robust	OLS	Robust	OLS	Robust
	(1)	(2)	(3)	(4)	(5)	(6)
Na	-3.498^{**}	-3.498^{**}	-0.186	-0.186	-0.024	-0.024
	(1.363)	(1.514)	(1.232)	(0.920)	(1.236)	(0.982)
etoh			-0.255^{***}	-0.255^{***}	-0.239^{***}	-0.239^{***}
			(0.051)	(0.065)	(0.053)	(0.066)
smoking					-0.0004	-0.0004
					(0.0003)	(0.0003)
Constant	4.197^{***}	4.197^{***}	4.099^{***}	4.099^{***}	3.884^{***}	3.884^{***}
	(0.335)	(0.220)	(0.256)	(0.238)	(0.319)	(0.344)

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 9: Female, differenced regressions (2010-1990)

	OLS	Robust	OLS	Robust	OLS	Robust
	(1)	(2)	(3)	(4)	(5)	(6)
Na	0.322	0.322	2.118*	2.118*	2.180*	2.180*
	(1.258)	(1.305)	(1.088)	(1.113)	(1.125)	(1.129)
etoh			−0.419***	−0.419***	−0.410***	−0.410***
			(0.095)	(0.118)	(0.101)	(0.123)
smoking					−0.0001	−0.0001
					(0.0003)	(0.0002)
Constant	3.192***	3.192***	2.947***	2.947***	2.897***	2.897***
	(0.336)	(0.224)	(0.275)	(0.270)	(0.328)	(0.305)

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

6 R and package versions used

```
sessionInfo()

## R version 3.2.0 (2015-04-16)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.9.5 (Mavericks)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] cluster_2.0.3       rpart_4.1-10        stargazer_5.2
##  [4] plm_1.5-12          lmtest_0.9-34       zoo_1.7-12
##  [7] sandwich_2.3-4      randomForest_4.6-12 Hmisc_3.17-0
## [10] Formula_1.2-1       survival_2.38-3     lattice_0.20-33
## [13] ggplot2_1.0.1       reshape2_1.4.1      dplyr_0.4.3
## [16] xtable_1.8-0        knitr_1.11
##
## loaded via a namespace (and not attached):
##  [1] splines_3.2.0       colorspace_1.2-6    mgcv_1.8-9
##  [4] nloptr_1.0.4        foreign_0.8-66      DBI_0.3.1
##  [7] RColorBrewer_1.1-2  plyr_1.8.3          stringr_1.0.0
## [10] MatrixModels_0.4-1  munsell_0.4.2       gtable_0.1.2
## [13] bdsmatrix_1.3-2     evaluate_0.8        latticeExtra_0.6-26
## [16] SparseM_1.7         quantreg_5.19       pbkrtest_0.4-2
## [19] parallel_3.2.0      highr_0.5.1         proto_0.3-10
## [22] Rcpp_0.12.2         acepack_1.3-3.3     scales_0.3.0
## [25] formatR_1.2.1       lme4_1.1-10         gridExtra_2.0.0
## [28] digest_0.6.8        stringi_1.0-1       tools_3.2.0
## [31] magrittr_1.5        lazyeval_0.1.10     car_2.1-0
## [34] MASS_7.3-45         Matrix_1.2-2        assertthat_0.1
## [37] minqa_1.2.4         R6_2.1.1            nnet_7.3-11
## [40] nlme_3.1-122
```