

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

Kellie Ottoboni

May 20, 2016

1 Introduction

Idea: if salt is bad for you, then we'd expect

1. *ceteris paribus*, salt would be negatively associated with life expectancy and the association would be statistically significant.
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 satisfied. 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.

2 Data cleaning

Data preprocessing includes removing unwanted variables (we keep per capita GDP, 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.
- We impute the missing 2010 sex-specific alcohol consumption for Taiwan using a linear regression of male and female life expectancy, male and female salt consumption, per capita GDP, and per capita annual cigarette consumption on sex-specific alcohol 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/female as the weights.
- We impute the missing 1990 overall alcohol consumption for Taiwan in the same way as before, regressing predictors on overall alcohol consumption. 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 TODO 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	-0.74	-0.36	-0.55	-0.02
smoking	-0.41	-0.11	-0.18	-0.14
pc_gdp	0.49	0.73	0.37	0.71
Na	-0.41	-0.11	0.05	0.1

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.

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:

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

This means that treatment assignment is unconfounded given the generalized propensity score. Practically speaking, within strata of units with the same $r(T, X)$, treatment assignment is exchangeable. This suggests a two step 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. Create a regression tree to predict the GPS R using the covariates X , so leaves of the tree form strata with similar GPSs.
3. Do a stratified permutation test for correlation between treatment and outcomes.

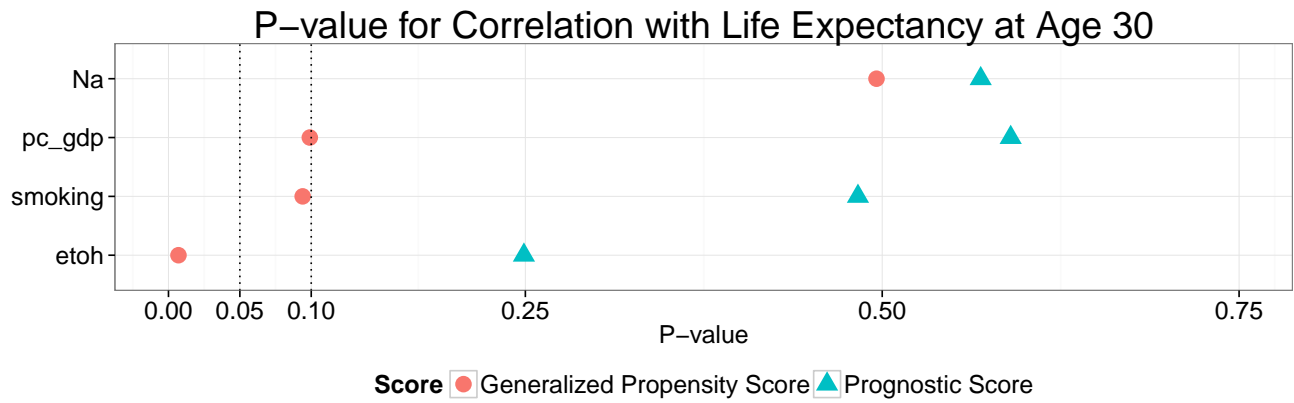


Figure 1: P-values for males, differenced (2010-1990).

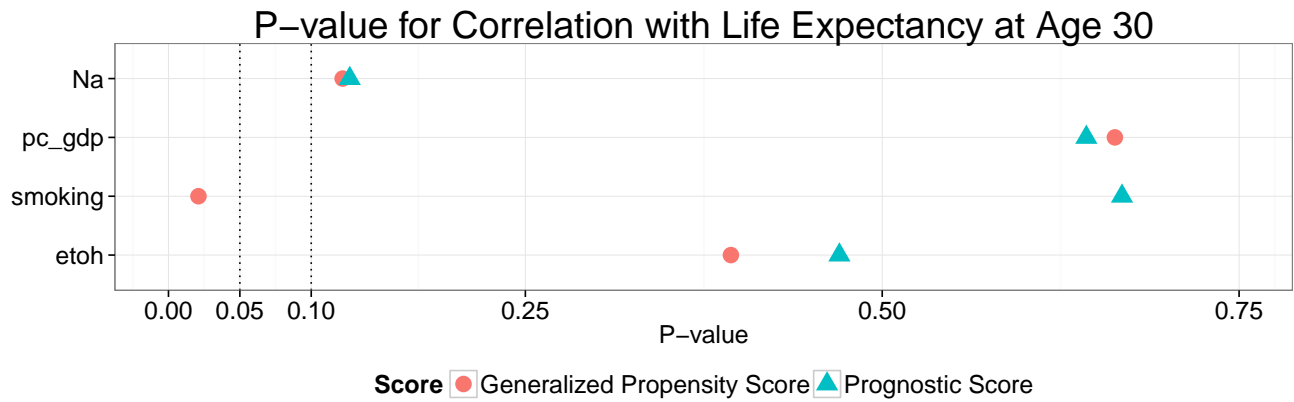


Figure 2: P-values for females, differenced (2010-1990).

3.4 Prognostic Score

elaborate We could do the same thing using the prognostic score (Hansen, 2008). This is essentially what we did for model-based matching, but we don't need to subtract off the prediction.

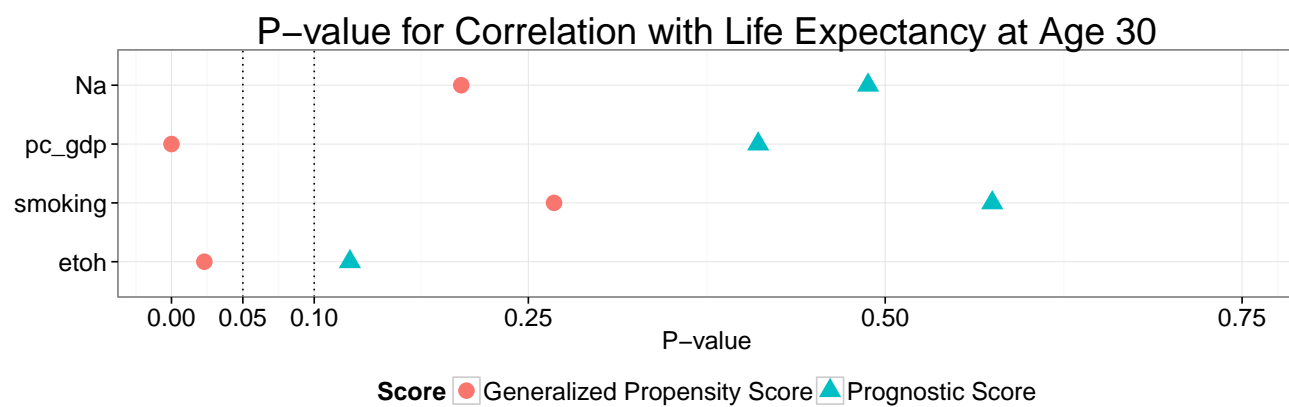


Figure 3: P-values for males, cross-sectional.

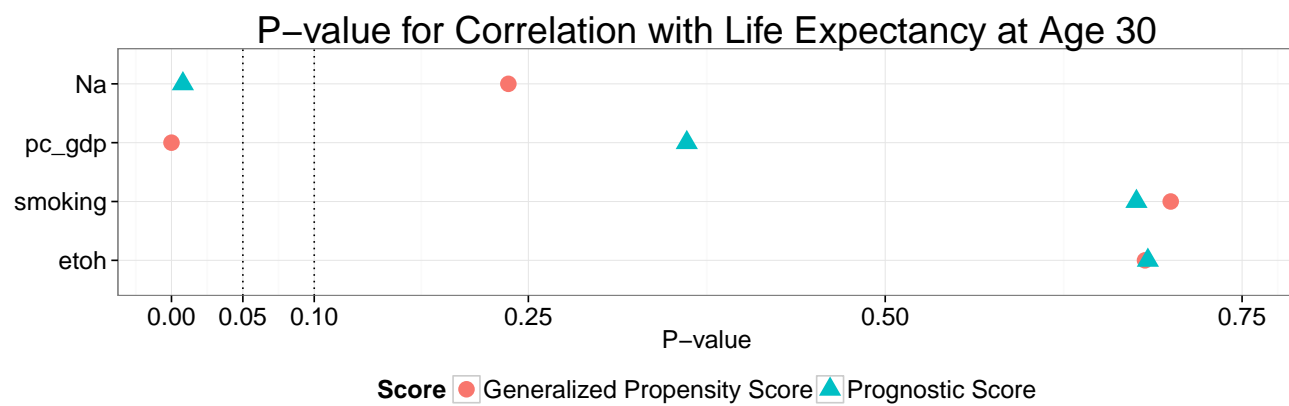


Figure 4: P-values for females, cross-sectional.

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 prediction error minus average permuted prediction error, and the “normalized importance,” the absolute importance divided by the original prediction error. Large values indicate more importance. Consistently for different cuts of the data and different methods, sodium intake appears to be the least important predictor. However, sodium is more important than smoking when we look at the cross-sectional datasets using OLS with interactions.

	Absolute				Normalized			
Variable	etoh	smoking	pc_gdp	Na	etoh	smoking	pc_gdp	Na
Random Forest	0.463	0.252	0.448	0.232	0.659	0.359	0.637	0.33
OLS	0.809	0.055	0.166	0.003	0.731	0.049	0.15	0.003
OLS with interactions	1.802	0.591	0.916	0.493	2.605	0.854	1.325	0.713

Table 1: Variable importance for males, differenced (2010-1990)

	Absolute				Normalized			
Variable	etoh	smoking	pc_gdp	Na	etoh	smoking	pc_gdp	Na
Random Forest	0.269	0.164	0.296	0.207	0.572	0.349	0.629	0.44
OLS	0.396	0	0.059	0.125	0.465	0	0.069	0.147
OLS with interactions	0.788	0.147	0.49	0.311	1.315	0.244	0.818	0.519

Table 2: Variable importance for females, differenced (2010-1990)

	Absolute				Normalized			
Variable	etoh	smoking	pc_gdp	Na	etoh	smoking	pc_gdp	Na
Random Forest	0.81	0.581	2.443	0.576	0.661	0.474	1.992	0.47
OLS	0.468	0.018	2.222	0.013	0.19	0.007	0.902	0.005
OLS with interactions	1.005	0.179	2.314	0.623	0.459	0.082	1.056	0.284

Table 3: Variable importance for males, cross-sectional

4.1 Conditional Variable Importance

We may want to do the conditional permutation tests in this paper: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-307>. They first use cart to partition the observations according to the other covariates, then permute within those groups instead of permuting everything. This is supposed to account for dependence between features.

	Absolute				Normalized			
Variable	etoh	smoking	pc_gdp	Na	etoh	smoking	pc_gdp	Na
Random Forest	0.441	0.42	1.794	0.402	0.556	0.529	2.262	0.506
OLS	0.043	0.002	1.472	0.05	0.024	0.001	0.807	0.028
OLS with interactions	0.583	0.141	1.69	0.274	0.371	0.09	1.076	0.175

Table 4: Variable importance for females, cross-sectional

5 Linear models

First, we look at conventional linear models for the differenced data. We report usual OLS standard errors as well as Huber-White robust standard errors. If salt were detrimental, then we'd expect to see a negative coefficient in the linear model. We do observe a negative coefficient for males, when we don't include any other variables in the model. However, as soon as we add alcohol to the model, the association disappears because increases in alcohol consumption explain a large portion of the decrease in life expectancy. We consider the full model including alcohol, smoking, and per capita GDP, in columns (7) and (8). For males, a one unit increase in sodium consumption from 1990 to 2010 is associated with an increase in life expectancy of 0.271 years. For females, it's associated with a 2.073 year increase in life expectancy between 1990 and 2010 (significant at 10% level).

Table 5: Male, Differenced

	OLS	Robust	OLS	Robust	OLS	Robust	OLS	Robust
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Na	-3.463** (1.326)	-3.463** (1.485)	-0.170 (1.177)	-0.170 (0.863)	-0.011 (1.180)	-0.011 (0.943)	0.271 (1.127)	0.271 (0.924)
etoh			-0.254*** (0.049)	-0.254*** (0.062)	-0.238*** (0.051)	-0.238*** (0.063)	-0.214*** (0.049)	-0.214*** (0.059)
smoking					-0.0004 (0.0003)	-0.0004 (0.0003)	-0.0003 (0.0003)	-0.0003 (0.0003)
pc_gdp							0.0001** (0.00003)	0.0001 (0.00004)
Constant	4.115*** (0.326)	4.115*** (0.211)	4.016*** (0.244)	4.016*** (0.226)	3.806*** (0.305)	3.806*** (0.331)	3.068*** (0.455)	3.068*** (0.629)

Notes:

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Next, we look at the data cross-sectionally instead of taking the difference over time. Again, if salt were detrimental to health, we'd expect it to have a negative coefficient in the linear model. We find the opposite: in all fixed effect specifications that we run, salt has a significant positive effect on life expectancy at age 30. In the OLS models, salt has no significant association with life expectancy. In the

Table 6: Female, Differenced

	OLS	Robust						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Na	0.356 (1.207)	0.356 (1.279)	2.080* (1.043)	2.080* (1.058)	2.110* (1.079)	2.110* (1.081)	2.073* (1.065)	2.073* (1.031)
etoh			-0.403*** (0.091)	-0.403*** (0.114)	-0.398*** (0.097)	-0.398*** (0.120)	-0.358*** (0.100)	-0.358*** (0.112)
smoking					-0.00004 (0.0002)	-0.00004 (0.0002)	-0.00001 (0.0002)	-0.00001 (0.0002)
pc_gdp							0.00003 (0.00002)	0.00003 (0.00003)
Constant	3.114*** (0.322)	3.114*** (0.211)	2.879*** (0.263)	2.879*** (0.247)	2.855*** (0.315)	2.855*** (0.286)	2.519*** (0.395)	2.519*** (0.440)

Notes:

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

full fixed effects model, for males, a one unit increase in sodium consumption is associated with an increase in life expectancy of 3.004 years (significant at 10% level). For females, it's associated with a 6.06 year increase in life expectancy (significant at 1% level).

Table 7: Male, Absolute

	Fixed Effects	Fixed Effects	Fixed Effects	OLS	Robust	Fixed Effects
	(1)	(2)	(3)	(4)	(5)	(6)
Na	4.576** (2.112)	8.436*** (2.566)	6.234** (2.859)	-0.509 (0.882)	-0.509 (0.851)	3.004* (1.509)
etoh		-0.316*** (0.113)	-0.170 (0.107)	-0.250*** (0.069)	-0.250*** (0.067)	-0.131* (0.066)
smoking			-0.003*** (0.001)	0.0003 (0.0004)	0.0003 (0.0004)	-0.001*** (0.0003)
pc_gdp				0.0002*** (0.00003)	0.0002*** (0.0001)	0.0002*** (0.00005)
Constant				44.289*** (3.527)	44.289*** (4.050)	

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] rpart_4.1-10      stargazer_5.2      plm_1.5-12
```


Table 8: Female, Absolute

	Fixed Effects	Fixed Effects	Fixed Effects	OLS	Robust	Fixed Effects
	(1)	(2)	(3)	(4)	(5)	(6)
Na	9.686*** (1.324)	11.219*** (0.940)	9.641*** (1.076)	0.933 (0.665)	0.933 (0.751)	6.060*** (1.231)
etoh		-0.605*** (0.135)	-0.407*** (0.146)	-0.156 (0.132)	-0.156 (0.132)	-0.247** (0.119)
smoking			-0.001*** (0.0004)	-0.0001 (0.0003)	-0.0001 (0.0003)	-0.001** (0.0003)
pc_gdp				0.0002*** (0.00002)	0.0002*** (0.00004)	0.0001*** (0.00004)
Constant				44.675*** (2.486)	44.675*** (3.065)	

Notes:

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

```
## [4] lmtest_0.9-34      zoo_1.7-12          sandwich_2.3-4
## [7] randomForest_4.6-12 Hmisc_3.17-0        Formula_1.2-1
## [10] survival_2.38-3     lattice_0.20-33     ggplot2_1.0.1
## [13] reshape2_1.4.1      dplyr_0.4.3         xtable_1.8-0
## [16] 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    codetools_0.2-14    evaluate_0.8
## [16] latticeExtra_0.6-26 SparseM_1.7          quantreg_5.19
## [19] pbkrtest_0.4-2     parallel_3.2.0      highr_0.5.1
## [22] proto_0.3-10       Rcpp_0.12.2         acepack_1.3-3.3
## [25] scales_0.3.0       formatR_1.2.1       lme4_1.1-10
## [28] gridExtra_2.0.0    digest_0.6.8        stringi_1.0-1
## [31] tools_3.2.0        magrittr_1.5         lazyeval_0.1.10
## [34] cluster_2.0.3      car_2.1-0           MASS_7.3-45
## [37] Matrix_1.2-2       assertthat_0.1      minqa_1.2.4
## [40] R6_2.1.1           nnet_7.3-11         nlme_3.1-122
```