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

Kellie Ottoboni

May 21, 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 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.

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 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	-0.74	-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

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 maeks 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 densityof 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 $r(T, X)$, 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.

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 Results

Figures 1 and 2 show the results for males, differenced and cross-sectionally, respectively. In both, per capita GDP and alcohol are both significantly associated with life expectancy at age 30. Smoking is weakly (at the 10% level) associated with life expectancy at age 30 in the differenced data. The p-values for sodium are all above 10%.

Figures 3 and 4 show the results for females, differenced and cross-sectionally, respectively. The results are less conclusive here. When looking at the differenced data and stratifying on GPS, smoking is significantly associated with life expectancy. When looking at the cross-sectional data, sodium and per capita GDP are significantly associated with life expectancy. Recall that for females, the correlation between sodium and life expectancy is actually positive; thus this result says that sodium has a statistically significant 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 GPS tends to give smaller p-values than the other stratification. I am not sure how much confidence we should place in this analysis. Taken at face value, however, it says 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.

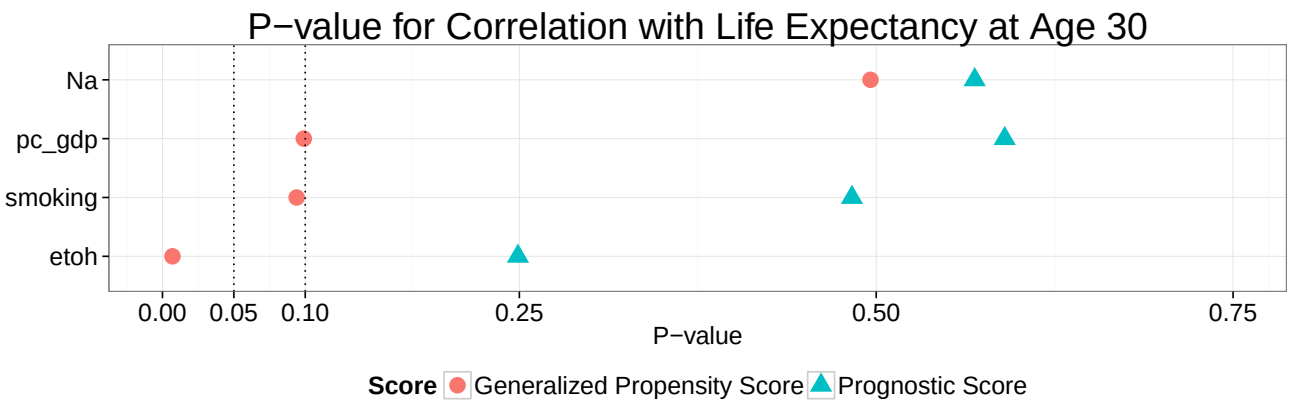

Figure 1: P-values for males, differenced (2010-1990).

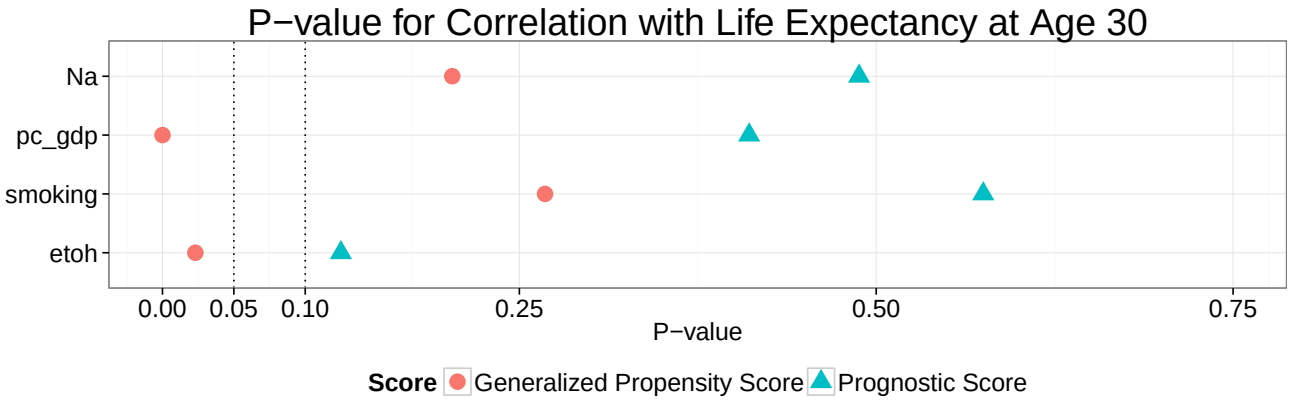

Figure 2: P-values for males, cross-sectional.

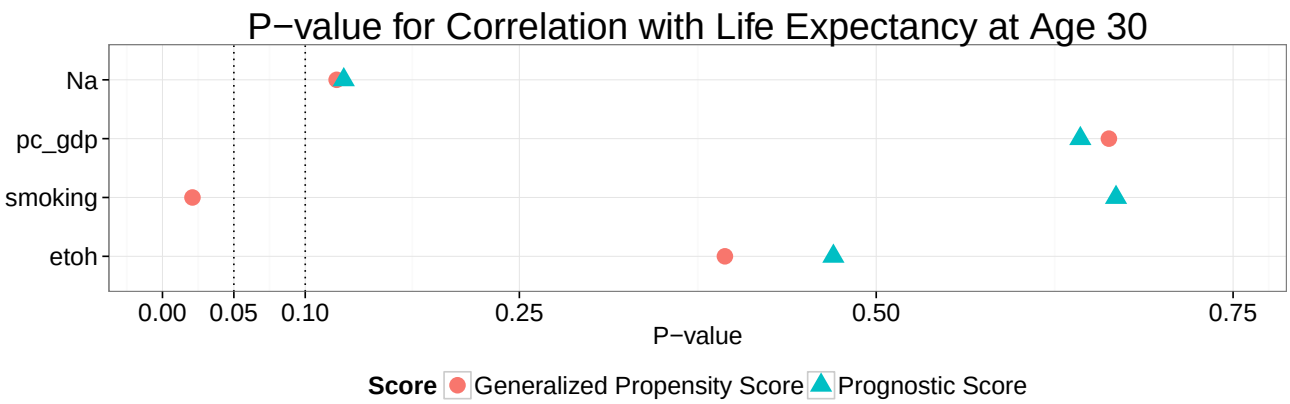

Figure 3: P-values for females, differenced (2010-1990).

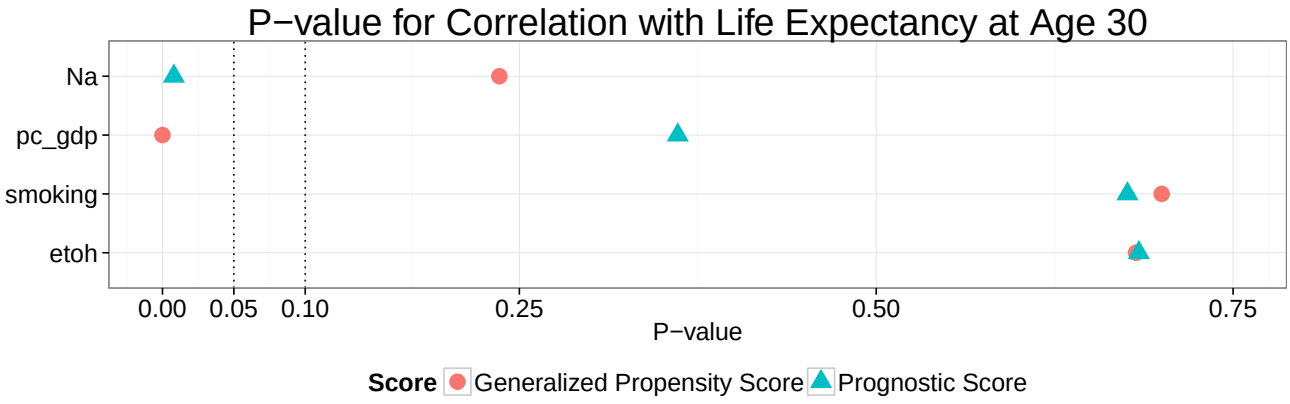

Figure 4: P-values for females, cross-sectional.

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.

4.1 Results

In the differenced datasets, alcohol is the most important predictor. In the cross-sectional data, per capita GDP is the most important predictor. This makes sense: per capita GDP probably grows globally at similar rates, so it would add less information when looking at the differenced data.

Sodium intake appears to be the least important predictor for different cuts of the data and different methods. Exceptions include OLS with interactions in the male cross-sectional data, all models for the female differenced data, and OLS and OLS with interactions for the female cross-sectional data.

	Absolute				Normalized			
	etoh	smoking	pc_gdp	Na	etoh	smoking	pc_gdp	Na
Random Forest	0.454	0.278	0.391	0.235	0.634	0.388	0.545	0.328
OLS	0.829	0.042	0.172	0.003	0.75	0.038	0.156	0.002
OLS with interactions	1.819	0.579	0.873	0.512	2.631	0.837	1.262	0.741

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

	Absolute				Normalized			
	etoh	smoking	pc_gdp	Na	etoh	smoking	pc_gdp	Na
Random Forest	0.847	0.589	2.508	0.566	0.713	0.496	2.11	0.476
OLS	0.46	0.021	2.292	0.013	0.187	0.008	0.931	0.005
OLS with interactions	1	0.181	2.395	0.64	0.457	0.083	1.093	0.292

Table 3: Variable importance for males, cross-sectional

	Absolute				Normalized			
	etoh	smoking	pc_gdp	Na	etoh	smoking	pc_gdp	Na
Random Forest	0.314	0.168	0.262	0.199	0.669	0.357	0.557	0.423
OLS	0.392	0	0.062	0.118	0.461	0	0.073	0.139
OLS with interactions	0.745	0.15	0.455	0.305	1.243	0.25	0.759	0.509

Table 4: Variable importance for females, differenced (2010-1990)

	Absolute				Normalized			
	etoh	smoking	pc_gdp	Na	etoh	smoking	pc_gdp	Na
Random Forest	0.452	0.446	1.663	0.417	0.553	0.546	2.038	0.511
OLS	0.036	0.003	1.457	0.049	0.02	0.002	0.798	0.027
OLS with interactions	0.628	0.125	1.749	0.258	0.4	0.08	1.114	0.164

Table 5: Variable importance for females, cross-sectional

5 Linear models

5.1 Fixed Effects Model

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

$$Y_{it} = \alpha_i + \lambda_t + \beta_1 \text{GDP}_{it} + \beta_2 \text{ETOH}_{it} + \beta_3 \text{SMOKE}_{it} + \gamma \text{Na}_{it} + \varepsilon_{it}$$

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. 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 assoiation with life expectancy. In the full fixed effects model for males (column (6) in Table 6), 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 (column (6) in Table 7), it's associated with a 6.06 year increase in life expectancy (significant at 1% level).

5.2 Differenced Model

Next, we look at conventional linear models for the differenced data. We have measurements from 1990 and 2010 for each country. If we take the difference, we can estimate

$$\Delta Y_{it} = \Delta\lambda_t + \beta_1 \Delta\text{GDP}_i + \beta_2 \Delta\text{ETOH}_i + \beta_3 \Delta\text{SMOKE}_i + \gamma \Delta\text{Na}_i + \Delta\varepsilon_i$$

This eliminates the need for fixed effects. 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.

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 focus on the full model including alcohol, smoking, and per capita GDP. For males (column (8) in Table 8), 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 (column (8) in Table 9), it's associated with a 2.073 year increase in life expectancy between 1990 and 2010 (significant at 10% level).

Table 6: Male cross-sectional data

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

Notes:

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

Table 7: Female cross-sectional data

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

Notes:

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

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

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

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

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
##  [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     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     cluster_2.0.3
## [34] car_2.1-0           MASS_7.3-45         Matrix_1.2-2
## [37] assertthat_0.1      minqa_1.2.4         R6_2.1.1
## [40] nnet_7.3-11         nlme_3.1-122
```