Analysis Methods for Salt and Mortality

Kellie Ottoboni

Draft May 16, 2016

- **Generalized matching** We may be able to do some form of our original idea.

  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 tree-based method, where leaves of the tree form strata with the same generalized propensity score.
  2. Do a stratified permutation test for correlation between treatment and outcomes.

  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.

  I suggest that we do both: a stratified permutation test based on the generalized propensity score and another based on the predicted prognostic score.

- **Ordinary linear regression**

  – We clearly don't want to assume selection on observables: alcohol consumption, GDP, and smoking levels clearly aren't the only factors that account for a country's life expectancy at age 30.

- This isn't a random experiment, so running a regression won't give us a causal estimate of anything.
- What we can do, however, is look at some measure of goodness of fit, with and without salt in the regression.

- **Breiman's variable importance**
  - What Breiman suggests (I need to find a reference – I think it's in the Two Cultures paper) is to compute goodness of fit (root mean squared prediction error) for the true model as a baseline. Then, permute the variable of interest, breaking any association between the variable and the other covariates and outcome. Refit the model using the permuted predictor and recompute the RMSE. Do this many times to get a distribution of RMSEs. TO DO: Is this a valid permutation test? Are things actually exchangeable? If not, maybe that's okay – we're not trying to control any error rate, we just want some measure of importance and a p-value can serve as one.
  - We can do this for each of the predictors, so we can see the relative importance of each in the model. For example, we'd hope to see a big decrease in RMSE when we perturb GDP and a relatively smaller decrease in RMSE when we perturb salt.
  - Of course, this will depend on the model we choose. We should do this procedure for a variety of models to give evidence that salt is consistently less important than other variables, independent of the choice of model. I suggest doing it for OLS, CART, random forests, TO DO: what else?

- **Econometric linear models**
  - Fixed effects: To account for unobserved confounding variables, we will 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.
  - Differenced estimator: We have measurements from 1990 and 2010 for each country. If we take the difference, we can estimate instead

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

    This eliminates the need for the 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).

   The second point makes more sense in the context of binary treatments than here (but perhaps there's a more sensible way of putting it that clarifies the point).

– For this to have a causal interpretation, we need salt consumption to be assigned to countries at random, conditional on GDP, ETOH, and smoking levels. We clearly don't have this. I'm wondering if it would still be useful to do this, with the major caveat that we're letting go of causal interpretations.