Methodology: labor-market measures for likelier Trump or Clinton supporters
Jed Kolko, Indeed chief economist

This document explains how we estimated labor-market measures for likelier Trump and Clinton supporters. These estimates first appeared in the New York Times Upshot section on Sept. 5, 2017.

We replicate the headline jobs numbers that the Bureau of Labor Statistics (BLS) regularly produces from the monthly Current Population Survey (CPS) of households. The CPS does not ask respondents whom they voted for or which political party they belong to. Instead, we used a publicly available survey about voting, the Cooperative Congressional Election Study, and built a model to predict each respondent's presidential vote based on basic demographic information that is also available in the CPS.

The methodology has three steps:

1. Building a predictive model of voting behavior
2. Calculating labor-market series for likelier supporters of Trump and Clinton
3. Seasonally adjusting the labor-market series

The classification-tree procedure in Step 1 and the seasonal adjustment in Step 3 were done in R; the remainder was done in Stata.

Step 1: building a predictive model of voting behavior

The predictive model of the 2016 vote uses the Cooperative Congressional Election Study, which has a sample of more than 40,000 respondents who voted in the 2016 presidential election for either Donald Trump or Hillary Clinton.

The model predicts a respondent's 2016 vote based on seven demographic variables: age, race (including Hispanic origin), educational attainment, sex, marital status, presence of dependent children in the household, and state of residence. These variables were chosen because of their availability in the CPS and other publicly available datasets. To note:

- State is the smallest geography that is comprehensively available in CPS microdata. The CCES does report ZIP code as well, which was not incorporated into the model because it is not available in the CPS.
- Income, employment status and other economic-outcome variables were not included as model inputs because the ultimate research goal is to estimate economic-outcome variables.

The model uses a hybrid CART-probit approach: The binary dependent variable of the 2016 presidential vote is regressed on a set of categorical dummies for seven demographic variables as well as dummies for the terminal nodes from a classification tree based on the same seven demographic variables. The inclusion of terminal-node dummies from a classification tree incorporates the most important interaction effects among the seven demographic variables.

The classification tree shows that the most important predictor of how someone voted was race (including Hispanic origin), followed by education. A variant of this approach is described here.

A probit regression of the 2016 vote on the categorical dummies for the seven demographic variables yields only a pseudo r-squared of .16 (this is the approach taken in recent academic work on partisanship, economic confidence and household spending that built a predictive model from the same voting survey). A probit regression of the 2016 vote on categorical dummies for the terminal nodes from a classification tree yields a pseudo r-squared of .17. The full model, with both sets of dummies, yields a pseudo r-squared of .18.

The output of the model is the likelihood that a respondent voted for Trump or Clinton for an individual year of age (18-95), five categories of race (including Hispanic origin), five levels of educational attainment, sex, marital status, presence of dependent children in the household, and state of residence -- a total of 795,600 possible demographic combinations.

Step 2: Calculating labor-market series by predicted voting behavior

The labor-market measures are based on the CPS basic monthly public-use microdata files, downloaded from the Census FTP site and read using the NBER Stata do-files and dictionaries. For every CPS respondent 18 and older from 2008 to the present, their likelihood of being a Trump or Clinton supporter is estimated by matching their demographics to the model in Step 1. The CPS basic monthly survey does not ask respondents about voting behavior or political party affiliation.

Estimates for the U-3 and U-6 unemployment rates, the prime-age employment-population ratio, and the prime-age labor-force participation rate follow the BLS definitions and weights (PWCMPWGT in the Census FTP files). Our estimates match the seasonally unadjusted series published by the BLS, with only occasional divergences attributable to confidentiality perturbations in the public-use files and differences in rounding procedures.

Estimates for median wages are calculated using the BLS filters for full-time wage and salary workers in the outgoing rotation groups, using the recoded usual weekly earnings variable (PRERNWA) and the appropriate weight (PWORWGT). A key difference from standard BLS practice is the calculation of the median. For wage, earning and income data, median calculations may be affected by "heaping" -- the bunching of responses around round numbers -- such that medians can either stay at the same round number over time or make large jumps between round numbers. While the BLS resolves this by interpolating, we construct a pseudo-median that equals the mean of earnings observations within the 40th-60th percentiles of the relevant weighted distribution. For median wages, a three-month trailing average is reported.

For all measures -- labor-force indicators and median wages -- the estimates for likelier Trump and Clinton supporters are calculated by weighting the measure by the likelihood that each

respondent is a supporter of each candidate. In effect, PWCMPWGT or PWORWGT was multiplied by the likelihood that the respondent is a Trump supporter in order to get the likelier Trump supporter measure, and multiplied by one minus that likelihood in order to get the likelier Clinton supporter measure.

The output of this step is two data series from January 2010 to the present for each labor market indicator: one for likelier Trump supporters and one for likelier Clinton supporters.

Step 3: Seasonally adjusting the labor-market series

The final step is to seasonally adjust all of the labor-market series. This was done with the Census Bureau's X-13 seasonal adjustment software, using the seasonal package in R. Median earnings are adjusted with a log transformation; all other measures are adjusted without transformation.

All of the data series are adjusted directly, even for indicators that the BLS adjusts indirectly. Data series for likelier Trump and Clinton supporters were adjusted individually; the seasonal pattern for a given indicator can differ for the two voter bases.

Final points

The average of the seasonally adjusted series for likelier Trump and Clinton supporters might not equal the official BLS published seasonally adjusted series for an indicator. This is because of the above seasonal adjustment issues; the restriction to adults age 18 and over; and the fact that the total voter-likelihood weights for Clinton and Trump are not equal to each other.

The survey of voting behavior, the CCES, asks respondents about their employment status in the pre-election phase of the survey, though the question wording does not appear to be directly comparable to the CPS question. The unemployment rate among respondents was 8.3 percent for subsequent actual Trump voters versus 9.5 percent for subsequent actual Clinton voters. The difference of 1.2 percent lies between the estimated higher rate for likelier Clinton supporters relative to likelier Trump supporters of 1.0 percent for U-3 and 2.0 percent for U-6 in October 2016 based on CPS data and the voting model.