Supplementary Online Content


eTable 1. HCUP Definitions for hospital size.

eTable 2. Numbers of patients in study and perforation rates for appendicitis. The numbers represent the absolute counts of cases in each category found in the HCUP databases for 2002-2008 for acute and perforated appendicitis.

eTable 3. Mean values representing appendicitis perforation rates for each of the factors assessed that could influence appendiceal perforation.

eTable 4. Logistic regression for perforated appendicitis on all the variables that might influence pediatric appendicitis perforation rates.

eAppendix

This supplementary material has been provided by the authors to give readers additional information about their work.
eTable 1. HCUP Definitions for hospital size.

<table>
<thead>
<tr>
<th>Location and Teaching Status</th>
<th>Hospital Bedsize</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small</td>
</tr>
<tr>
<td><strong>NORTHEAST REGION</strong></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>1-49</td>
</tr>
<tr>
<td>Urban, nonteaching</td>
<td>1-124</td>
</tr>
<tr>
<td>Urban, teaching</td>
<td>1-249</td>
</tr>
<tr>
<td><strong>MIDWEST REGION</strong></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>1-29</td>
</tr>
<tr>
<td>Urban, nonteaching</td>
<td>1-74</td>
</tr>
<tr>
<td>Urban, teaching</td>
<td>1-249</td>
</tr>
<tr>
<td><strong>SOUTHERN REGION</strong></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>1-39</td>
</tr>
<tr>
<td>Urban, nonteaching</td>
<td>1-99</td>
</tr>
<tr>
<td>Urban, teaching</td>
<td>1-249</td>
</tr>
<tr>
<td><strong>WESTERN REGION</strong></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>1-24</td>
</tr>
<tr>
<td>Urban, nonteaching</td>
<td>1-99</td>
</tr>
<tr>
<td>Urban, teaching</td>
<td>1-199</td>
</tr>
</tbody>
</table>
eTable 2. Numbers of patients in study and perforation rates for appendicitis. The numbers represent the absolute counts of cases in each category found in the HCUP databases for 2002-2008 for acute and perforated appendicitis.

<table>
<thead>
<tr>
<th>Age &lt; 18</th>
<th>White</th>
<th>Black</th>
<th>Latino</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonperforated Appendicitis</td>
<td>41538</td>
<td>4264</td>
<td>17370</td>
<td>32990</td>
</tr>
<tr>
<td>Perforated Appendicitis</td>
<td>15097</td>
<td>2343</td>
<td>9995</td>
<td>13919</td>
</tr>
<tr>
<td>Perforation rate (%)</td>
<td>26.7</td>
<td>35.5</td>
<td>36.5</td>
<td>29.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age &gt;= 18</th>
<th>White</th>
<th>Black</th>
<th>Latino</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonperforated Appendicitis</td>
<td>143332</td>
<td>14901</td>
<td>36384</td>
<td>92423</td>
</tr>
<tr>
<td>Perforated Appendicitis</td>
<td>58990</td>
<td>5783</td>
<td>12414</td>
<td>35927</td>
</tr>
<tr>
<td>Perforation rate (%)</td>
<td>29.2</td>
<td>28.0</td>
<td>25.4</td>
<td>28.0</td>
</tr>
</tbody>
</table>
eTable 3. Mean values representing appendicitis perforation rates for each of the factors assessed that could influence appendiceal perforation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.4025</td>
<td>0.3775</td>
<td>0.3893</td>
</tr>
<tr>
<td>Age</td>
<td>11.8488</td>
<td>11.7948</td>
<td>10.0047</td>
</tr>
<tr>
<td>Age*Age</td>
<td>154.5046</td>
<td>155.0643</td>
<td>120.5535</td>
</tr>
<tr>
<td>Lower Middle Class</td>
<td>0.2150</td>
<td>0.2396</td>
<td>0.2447</td>
</tr>
<tr>
<td>Upper Middle Class</td>
<td>0.2492</td>
<td>0.1897</td>
<td>0.2074</td>
</tr>
<tr>
<td>Wealthy</td>
<td>0.3755</td>
<td>0.1712</td>
<td>0.1803</td>
</tr>
<tr>
<td>Missing income information</td>
<td>0.0196</td>
<td>0.0209</td>
<td>0.0200</td>
</tr>
<tr>
<td>Midwest</td>
<td>0.1706</td>
<td>0.1135</td>
<td>0.0155</td>
</tr>
<tr>
<td>South</td>
<td>0.3656</td>
<td>0.5390</td>
<td>0.4197</td>
</tr>
<tr>
<td>West</td>
<td>0.1660</td>
<td>0.0730</td>
<td>0.4100</td>
</tr>
<tr>
<td>Medicare</td>
<td>0.0020</td>
<td>0.0026</td>
<td>0.0011</td>
</tr>
<tr>
<td>Private</td>
<td>0.7368</td>
<td>0.4358</td>
<td>0.3630</td>
</tr>
<tr>
<td>Medicaid</td>
<td>0.2031</td>
<td>0.4932</td>
<td>0.5347</td>
</tr>
<tr>
<td>Self Pay</td>
<td>0.1459</td>
<td>0.1707</td>
<td>0.1199</td>
</tr>
<tr>
<td>Emergent Admission</td>
<td>0.6385</td>
<td>0.7686</td>
<td>0.4841</td>
</tr>
<tr>
<td>Urgent Admission</td>
<td>0.1734</td>
<td>0.1117</td>
<td>0.1082</td>
</tr>
<tr>
<td>Elective Admission</td>
<td>0.0828</td>
<td>0.0605</td>
<td>0.0580</td>
</tr>
<tr>
<td>Medium Bed Size</td>
<td>0.2543</td>
<td>0.2291</td>
<td>0.3334</td>
</tr>
<tr>
<td>Large Bed Size</td>
<td>0.5688</td>
<td>0.6297</td>
<td>0.5443</td>
</tr>
<tr>
<td>Bed Size Missing</td>
<td>0.0003</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Non Federal Hospital</td>
<td>0.0709</td>
<td>0.0601</td>
<td>0.0616</td>
</tr>
<tr>
<td>Non profit Hospital</td>
<td>0.1922</td>
<td>0.1513</td>
<td>0.2320</td>
</tr>
<tr>
<td>Private-Investor Owned</td>
<td>0.0788</td>
<td>0.0818</td>
<td>0.1143</td>
</tr>
<tr>
<td>Combined Ownership Category</td>
<td>0.0535</td>
<td>0.0068</td>
<td>0.0108</td>
</tr>
<tr>
<td>Teaching Hospital</td>
<td>0.4203</td>
<td>0.6243</td>
<td>0.5424</td>
</tr>
<tr>
<td>Urban Hospital</td>
<td>0.8387</td>
<td>0.9225</td>
<td>0.9528</td>
</tr>
<tr>
<td>Annual Discharges (AD)</td>
<td>1.9010</td>
<td>2.5439</td>
<td>21.1794</td>
</tr>
<tr>
<td>AD*AD</td>
<td>6.4119</td>
<td>10.3462</td>
<td>0.1229</td>
</tr>
<tr>
<td>Annual Appendectomy Volume (AAV)</td>
<td>2.3068</td>
<td>2.6698</td>
<td>0.1178</td>
</tr>
<tr>
<td>AAV*AAV</td>
<td>9.1602</td>
<td>11.5438</td>
<td>0.1384</td>
</tr>
<tr>
<td>year=2002</td>
<td>0.1220</td>
<td>0.1211</td>
<td>0.1404</td>
</tr>
<tr>
<td>year=2003</td>
<td>0.1179</td>
<td>0.1442</td>
<td>0.1298</td>
</tr>
<tr>
<td>year=2004</td>
<td>0.1340</td>
<td>0.1354</td>
<td>0.1194</td>
</tr>
<tr>
<td>year=2005</td>
<td>0.1296</td>
<td>0.1256</td>
<td>0.1046</td>
</tr>
<tr>
<td>year=2006</td>
<td>0.1207</td>
<td>0.1062</td>
<td>0.3390</td>
</tr>
<tr>
<td>year=2007</td>
<td>0.1104</td>
<td>0.1212</td>
<td>0.0000</td>
</tr>
<tr>
<td>year=2008</td>
<td>0.1351</td>
<td>0.1311</td>
<td>0.0000</td>
</tr>
<tr>
<td>Sample size</td>
<td>56,564</td>
<td>6,590</td>
<td>29,486</td>
</tr>
</tbody>
</table>

© 2011 American Medical Association. All rights reserved.
eTable 4. Logistic regression for perforated appendicitis on all the variables that might influence pediatric appendicitis perforation rates.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef. Est.</th>
<th>Std. Err.</th>
<th>Odds Ratio</th>
<th>95% Conf. Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>0.2851</td>
<td>0.0288</td>
<td>1.33</td>
<td>1.257 1.407</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.1565</td>
<td>0.0182</td>
<td>1.169</td>
<td>1.128 1.212</td>
</tr>
<tr>
<td>Asian</td>
<td>0.2366</td>
<td>0.0485</td>
<td>1.267</td>
<td>1.152 1.393</td>
</tr>
<tr>
<td>American Indian</td>
<td>0.3813</td>
<td>0.0822</td>
<td>1.464</td>
<td>1.246 1.72</td>
</tr>
<tr>
<td>Race-Other</td>
<td>0.1009</td>
<td>0.0347</td>
<td>1.106</td>
<td>1.033 1.184</td>
</tr>
<tr>
<td>Race-Missing</td>
<td>-0.00341</td>
<td>0.0173</td>
<td>0.997</td>
<td>0.963 1.031</td>
</tr>
<tr>
<td>Female</td>
<td>-0.0389</td>
<td>0.0126</td>
<td>0.962</td>
<td>0.938 0.986</td>
</tr>
<tr>
<td>Age</td>
<td>-0.1759</td>
<td>0.0072</td>
<td>0.839</td>
<td>0.827 0.851</td>
</tr>
<tr>
<td>Age*Age</td>
<td>0.00435</td>
<td>0.000343</td>
<td>1.004</td>
<td>1.004 1.005</td>
</tr>
<tr>
<td>Lower Middle Class</td>
<td>-0.0241</td>
<td>0.0189</td>
<td>0.976</td>
<td>0.941 1.013</td>
</tr>
<tr>
<td>Upper Middle Class</td>
<td>-0.0575</td>
<td>0.0196</td>
<td>0.944</td>
<td>0.909 0.981</td>
</tr>
<tr>
<td>Wealthy</td>
<td>-0.1528</td>
<td>0.0208</td>
<td>0.858</td>
<td>0.824 0.894</td>
</tr>
<tr>
<td>Missing income information</td>
<td>-0.0519</td>
<td>0.0462</td>
<td>0.949</td>
<td>0.867 1.039</td>
</tr>
<tr>
<td>SelfPay</td>
<td>0.2003</td>
<td>0.0231</td>
<td>1.222</td>
<td>1.168 1.278</td>
</tr>
<tr>
<td>AdmEmergent</td>
<td>0.2254</td>
<td>0.0222</td>
<td>1.253</td>
<td>1.199 1.309</td>
</tr>
<tr>
<td>AdmUrgent</td>
<td>0.3587</td>
<td>0.0285</td>
<td>1.431</td>
<td>1.354 1.514</td>
</tr>
<tr>
<td>AdmElective</td>
<td>-0.2763</td>
<td>0.1185</td>
<td>0.759</td>
<td>0.601 0.957</td>
</tr>
<tr>
<td>bedsizemed</td>
<td>-0.1014</td>
<td>0.0219</td>
<td>0.904</td>
<td>0.866 0.943</td>
</tr>
<tr>
<td>bedsizelarge</td>
<td>0.0781</td>
<td>0.0227</td>
<td>1.081</td>
<td>1.034 1.13</td>
</tr>
<tr>
<td>Self Pay</td>
<td>0.0807</td>
<td>0.0199</td>
<td>1.084</td>
<td>1.043 1.127</td>
</tr>
<tr>
<td>Emergent Admission</td>
<td>0.00849</td>
<td>0.0262</td>
<td>1.009</td>
<td>0.958 1.062</td>
</tr>
<tr>
<td>Urgent Admission</td>
<td>0.0918</td>
<td>0.0282</td>
<td>1.096</td>
<td>1.037 1.158</td>
</tr>
<tr>
<td>Elective Admission</td>
<td>0.072</td>
<td>0.0328</td>
<td>1.075</td>
<td>1.008 1.146</td>
</tr>
<tr>
<td>Medium Bed Size</td>
<td>-0.0224</td>
<td>0.0211</td>
<td>0.978</td>
<td>0.938 1.019</td>
</tr>
<tr>
<td>Large Bed Size</td>
<td>0.0291</td>
<td>0.026</td>
<td>1.029</td>
<td>0.978 1.083</td>
</tr>
<tr>
<td>Bed Size Missing</td>
<td>-0.1242</td>
<td>0.2736</td>
<td>0.883</td>
<td>0.517 1.51</td>
</tr>
<tr>
<td>Non Federal Hospital</td>
<td>0.00118</td>
<td>0.0337</td>
<td>1.001</td>
<td>0.937 1.069</td>
</tr>
<tr>
<td>Non profit Hospital</td>
<td>-0.0328</td>
<td>0.0275</td>
<td>0.968</td>
<td>0.917 1.021</td>
</tr>
<tr>
<td>Private-Investor Owned</td>
<td>-0.0536</td>
<td>0.0329</td>
<td>0.948</td>
<td>0.889 1.011</td>
</tr>
<tr>
<td>Combined Ownership Category</td>
<td>0.0513</td>
<td>0.0415</td>
<td>1.053</td>
<td>0.97 1.142</td>
</tr>
<tr>
<td>Teaching Hospital</td>
<td>0.1952</td>
<td>0.0236</td>
<td>1.216</td>
<td>1.161 1.273</td>
</tr>
<tr>
<td>Urban Hospital</td>
<td>0.1313</td>
<td>0.0287</td>
<td>1.14</td>
<td>1.078 1.206</td>
</tr>
<tr>
<td>Annual Discharges (AD)</td>
<td>-0.0108</td>
<td>0.0196</td>
<td>0.989</td>
<td>0.952 1.028</td>
</tr>
<tr>
<td>AD*AD</td>
<td>0.00219</td>
<td>0.00223</td>
<td>1.002</td>
<td>0.998 1.007</td>
</tr>
<tr>
<td>Annual Appendectomy Volume (AAV)</td>
<td>0.0151</td>
<td>0.0128</td>
<td>1.015</td>
<td>0.99 1.041</td>
</tr>
<tr>
<td>AAV*AAV</td>
<td>-0.00309</td>
<td>0.00137</td>
<td>0.997</td>
<td>0.994 1</td>
</tr>
<tr>
<td>year=2002</td>
<td>0.00209</td>
<td>0.0249</td>
<td>1.002</td>
<td>0.954 1.052</td>
</tr>
<tr>
<td>year=2003</td>
<td>0.00965</td>
<td>0.0253</td>
<td>1.01</td>
<td>0.961 1.061</td>
</tr>
<tr>
<td>year=2004</td>
<td>-0.0218</td>
<td>0.025</td>
<td>0.978</td>
<td>0.932 1.028</td>
</tr>
<tr>
<td>year=2005</td>
<td>0.00981</td>
<td>0.0248</td>
<td>1.01</td>
<td>0.962 1.06</td>
</tr>
<tr>
<td>year=2006</td>
<td>-0.0514</td>
<td>0.0255</td>
<td>0.95</td>
<td>0.904 0.999</td>
</tr>
<tr>
<td>year=2007</td>
<td>-0.0451</td>
<td>0.0255</td>
<td>0.956</td>
<td>0.909 1.005</td>
</tr>
<tr>
<td>year=2008</td>
<td>-0.0868</td>
<td>0.0261</td>
<td>0.917</td>
<td>0.871 0.965</td>
</tr>
</tbody>
</table>

© 2011 American Medical Association. All rights reserved.
eAppendix

Conceptual explanation of Blinder-Oaxaca decomposition as applied to logistic regression

Data with binary outcomes (alive or dead etc.) are often assessed statistically by logistic regression. Logistic regression finds the best linear relationship between the logarithm of the ratio of the probability of an event happening (p) to it not happening (1-p). Using data obtained from the accompanying paper, we performed some simple regression and decomposition analyses to illustrate the results of logistic regression analysis with subsequent gap decomposition.

Using the 2001-2008 HCUP database, we identified 27,343 Latino children with appendicitis who experienced a 37% perforation rate. In the 56,564 white children the perforation rate was 27%. Thus, the racial gap in perforation rates to be explained was 37% - 27% or 10% or 0.10. In our analysis we found that age was a major determinate of perforation rates and performed regression analysis using age as a predictor variable for illustrative purposes.

Logistic regression of the odds for perforation as a function of age stratified by ethnicity yields:

\[
\ln \left( \frac{P_{\text{WP}}}{1 - P_{\text{WP}}} \right) = -0.0048 - 0.0556 \times \text{Age} \\
\ln \left( \frac{P_{\text{LP}}}{1 - P_{\text{LP}}} \right) = 0.5488 - 0.1090 \times \text{Age}
\]

where \( P_{\text{WP}} \) is the probability for white children appendiceal perforation and \( P_{\text{LP}} \) is that for Latino children.

Graphically:
ages and the gap narrows as the children get older. Age contributes to the gap in Latino and white perforated appendicitis rates. The following diagram illustrates how gap decomposition analysis breaks down the gap into explained and unexplained components:

The upper red line represents the regression line for Latino children perforated appendicitis rates as a function of age and the lower blue line that for white children. The purple dotted line denotes the log ratio of the probability of perforation to no perforation for Latinos on the y-axis (-0.4543) and the mean perforation rate for Latinos on the x-axis (9.2 years old). The green dashed-dotted line represents the log
ratio of the probability of perforation to no perforation for whites on the y-axis (-0.9629) and the mean perforation rate for whites on the x-axis (11.0 years old).

Solving the equation

\[
\ln\left(\frac{p}{1-p}\right)
\]

for Latinos at the mean age for perforation yields an expected probability of perforation of 0.39 and for whites of 0.28. The gap to be explained is 0.11 or 11 percentage points (slightly different than the overall gap in the entire population because of the age-adjustment in the regression analysis).

Gap decomposition in this example finds that amount of the racial difference in perforation rates that can be explained by age. It is performed by entering the age values for white children into the Latino regression equation. This is denoted by the black dashed line which represents entering the mean age for perforation for whites (11.0 years old) into the Latino regression equation. When this is done

\[
\ln\left(\frac{p}{1-p}\right)
\]

is -0.6505. Solving for \( p \) yields a probability of 0.34. Subtracting this value from the probability of perforation for Latinos (0.38) yields a gap of 0.04. 0.04/(0.39-0.28) is 36%. This implies that 36% of the gap between Latino and white perforated appendicitis rates can be explained by the age effect. The remaining 64% is unexplained by this particular regression analysis.

This example is a simplified illustration of decomposition analysis for a single variable. In practice, multivariate regression is used to facilitate the simultaneous assessment of many variables and allow for covariate adjustment. More rigorous mathematical explanations of decomposition can be found in (1;2) and its specific application to logistic regression (3-5).
Technical Description of the Blinder-Oxaca decomposition as applied to logistic regression

Nonlinear Decomposition Method

For a linear regression, the standard Blinder-Oaxaca decomposition of the white/black gap in the average value of the dependent variable, $Y$, can be expressed as:

\[
(1.1) \quad \overline{Y}_W - \overline{Y}_B = \left[ \overline{X}^W \hat{\beta}^W - \overline{X}^B \hat{\beta}^B \right] + \left[ \overline{X}^B ( \hat{\beta}^W - \hat{\beta}^B ) \right]
\]

where $\overline{X}^j$ is a row vector of average values of the independent variables and $\hat{\beta}^j$ is a vector of coefficient estimates for race $j$. Following Fairlie (4), the decomposition for a nonlinear equation, such as $Y = F(X \hat{\beta})$, can be written as:

\[
(1.2) \quad \overline{Y}_W - \overline{Y}_B = \left[ \sum_{i=1}^{N^W} \frac{F(X_i^W \hat{\beta}^W)}{N^W} - \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^B)}{N^B} \right] + \left[ \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^W)}{N^B} - \sum_{i=1}^{N^W} \frac{F(X_i^W \hat{\beta}^B)}{N^B} \right]
\]

where $N^j$ is the sample size for race $j$. This alternative expression for the decomposition is used because $\overline{Y}$ does not necessarily equal $F(\overline{X} \hat{\beta})$. In both (1.1) and (1.2), the first term in brackets represents the part of the racial gap that is due to group differences in distributions of $X$, and the second term represents the part due to differences in the group processes determining levels of $Y$. The second term also captures the portion of the racial gap due to group differences in unmeasurable or unobserved endowments. Similar to most previous studies applying the decomposition technique, we do not focus on this "unexplained" portion of the gap because of the difficulty in interpreting results (Jones and Cain for more discussion (6;7)).
To calculate the decomposition, define $\bar{Y}^j$ as the average probability of the binary outcome of interest for race $j$ and $F$ as the cumulative distribution function from the logistic distribution. Alternatively, for a probit model $F$ would be defined as the cumulative distribution function from the standard normal distribution.

An equally valid method of calculating the decomposition is to use the minority coefficient estimates, $\hat{\beta}^M$, as weights for the first term and the white distributions of the independent variables, $\bar{X}^w$, as weights for the second term. This alternative method of calculating the decomposition often provides different estimates, which is the familiar index problem with the Blinder-Oaxaca decomposition technique. A third alternative is to weight the first term of the decomposition expression using coefficient estimates from a pooled sample of the two groups (see Oaxaca and Ransom (8) for an example). We follow this approach to calculate the decompositions. In particular, we use coefficient estimates from a logit regression that includes a sample of all racial groups.

Using the pooled coefficients from a sample of all racial groups has the advantage over using the white coefficients because it captures the determinants for all groups and are more precisely estimated (because of the larger sample and more heterogeneity of patients). They are also preferred over the minority coefficients because they are less likely to be influenced by discrimination. The goal of the decomposition is to estimate how much differences in patient or hospital level factors explain of the racial gap in health care outcomes given a non-discriminatory environment.

The first term in (1.2) provides an estimate of the contribution of racial differences in the entire set of independent variables to the racial gap in the dependent variable. Estimation of the total contribution is relatively simple as one only needs to calculate two sets of predicted probabilities and take the difference between the average values of the two. Identifying the contribution of group
differences in specific variables to the racial gap, however, is not as straightforward. To simplify, first assume that \( N_B = N_W \) and that there exists a natural one-to-one matching of black and white observations. Using coefficient estimates from a logit regression for a pooled sample, \( \hat{\beta} \), the independent contribution of \( X_1 \) to the racial gap can then be expressed as:

\[
(1.3) \quad \frac{1}{N_B} \sum_{i=1}^{N_B} F(\hat{\alpha} + X_{i1}^{B}\hat{\beta}_1^* + X_{i2}^{B}\hat{\beta}_2^*) - F(\hat{\alpha} + X_{i1}^{W}\hat{\beta}_1^* + X_{i2}^{W}\hat{\beta}_2^*).
\]

Similarly, the contribution of \( X_2 \) can be expressed as:

\[
(1.4) \quad \frac{1}{N_B} \sum_{i=1}^{N_B} F(\hat{\alpha} + X_{i1}^{B}\hat{\beta}_1^* + X_{i2}^{B}\hat{\beta}_2^*) - F(\hat{\alpha} + X_{i1}^{W}\hat{\beta}_1^* + X_{i2}^{W}\hat{\beta}_2^*).
\]

The contribution of each variable to the gap is thus equal to the change in the average predicted probability resulting from sequentially switching the white characteristics to black characteristics one variable or set of variables at a time.\(^iv\) A useful property of this technique is that the sum of the contributions from individual variables will be equal to the total contribution from all of the variables evaluated with the full sample.

In practice, the sample sizes of the two groups are rarely the same and a one-to-one matching of observations from the two samples is needed to calculate (1.3) and (1.4). In this example, it is likely that the black sample size is substantially smaller than the white sample size. To address this problem, first use the pooled coefficient estimates to calculate predicted probabilities, \( \hat{Y}_i \), for each black and white observation in the sample. Next, draw a random subsample of whites with a sample size equal to \( N_B \) and randomly match it to the full black sample. The decomposition estimates obtained from this
procedure depend on the randomly chosen subsample of whites. Ideally, the results from the
decomposition should approximate those from matching the entire white sample to the black sample. A
simple method of approximating this hypothetical decomposition is to draw a large number of random
subsamples of whites, match each of these random subsamples of whites to the black sample, and
calculate separate decomposition estimates. The mean value of estimates from the separate
decompositions is calculated and used to approximate the results for the entire white sample. We
recommend using 1000 random subsamples of whites to calculate these means.

This non-linear technique has broader applications for identifying the causes of racial, gender,
geographical or other categorical differences in any binary dependent variable in which a logit or probit
model is used. SAS programs are available at people.ucsc.edu/~rfairlie/decomposition, and Stata
programs are available by entering "ssc install fairlie" in Stata. R and SAS code used for the current paper
is included below.
Reference List


---

i Note that the Blinder-Oaxaca decomposition is a special case of (4.2).

ii A useful property of the logit regression that includes a constant term is that the average of the predicted probabilities must equal the proportion of ones in the sample. In contrast, the predicted probability evaluated at the means of the independent variables is not necessarily equal to the proportion of ones, and in the sample used below it is larger because the logit function is concave for values greater than 0.5.

iii A black dummy variable is included in estimating the logit model with the pooled sample of blacks and whites, but is not used to calculate the decomposition.

iv Unlike in the linear case, the independent contributions of $X_1$ and $X_2$ depend on the value of the other variable. This implies that the choice of a variable as $X_1$ or $X_2$ (or the order of switching the distributions) is potentially important in calculating its contribution to the racial gap.
Sample SAS Code used for this manuscript

******************************************************************************;
* Example of Non-Linear Decomposition Technique for Logit Model *
* Updated on 9/3/10 *
* Used in: *
* Fairlie, Robert W. 1999 "The Absence of the African-American Owned *
* Business: An Analysis of the Dynamics of Self-Employment," *
* Decomposition Technique to Logit and Probit Models," Journal of *
* Economic and Social Measurement, 30(4): 305-316. *
******************************************************************************;

libname sasdata 'C:\FilesUTS\Disparity\SASrun';
options obs=max;
options nolabel ls=75 ps=140;
title 'Home Computer - Pooled 2 (All Races) Coefficients';

* specify number of simulations; *
* NOTE: change this to 1000 for final run; *
%let numsims=1;

* define race variables to be included only in pooled logits; *
%let r=6;
%let racevars=black Hispanic API AMIAN RaceOther raceMISSING; *
%let racevars=black Hispanic;

* specify number and names of independent variables; *
%let k=37;

%let vars=
FEMALE
AGE
agesqr
middleclass1
middleclass2
rich
zipincMISSING
Midwest
South
West
* define categories of variables for decomposition;  
%let group1= female ;  
%let group2= age agesqr;  
%let group3= middleclass1 middleclass2 rich zipincMISSING;  
%let group4= Midwest South West;  
%let group5= Medicare Private Medicaid SelfPay;  
%let group6= AdmEmergent AdmUrgent AdmElective;  
%let group7= bedsizemed bedsizelarge bedsizemiss;  
%let group8= controlgov_nonfed controlpriv_notprf controlpriv_invest controlpriv_collapsed hosp_teaching;  
%let group9= loc_urban;  
%let group10= total_disc total_disc_sqr numAnyAppy numAnyAppy_sqr;  

%let mvars=  
mFEMALE  
mAGE  
magesqr  
mmiddleclass1  
middleclass2  
mrich  
mzipincMISSING  
mMidwest  
mSouth  
mWest
mMedicare
mPrivate
mMedicaid
mSelfPay
mAdmEmergent
mAdmUrgent
mAdmElective
mbedsizemed
mbedsizelarge
mbedsizemiss
mcontrolgov_nonfed
mcontrolpriv_notprf
mcontrolpriv_invest
mcontrolprivCollapsed
mhosp_teaching
mloc_urban
mtotal_disc
mtotal_disc_sqr
mnumAnyAppy
mnumAnyAppy_sqr
my2002
my2003
my2004
my2005
my2006
my2007
my2008

* define categories of variables for decomposition;
%let mgroup1= mfemale ;
%let mgroup2= mage magesqr;
%let mgroup3= mmiddleclass1 mmiddleclass2 mrich mzipincMISSING;
%let mgroup4= mMidwest mSouth mWest;
%let mgroup5= mMedicare mPrivate mMedicaid mSelfPay;
%let mgroup6= mAdmEmergent mAdmUrgent mAdmElective;
%let mgroup7= mbedsizemed mbedsizelarge mbedsizemiss;
%let mgroup8= mcontrolgov_nonfed mcontrolpriv_notprf mcontrolpriv_invest
mcontrolprivCollapsed mhosp_teaching;
%let mgroup9= mloc_urban;
%let mgroup10= mtotal_disc mtotal_disc_sqr mnumAnyAppy mnumAnyAppy_sqr;

%let wvars=
wFEMALE
wAGE
wagesqr
wmiddleclass1
wmiddleclass2
wrich
wzipincMISSING
wMidwest
wSouth
wWest
* define categories of variables for decomposition;
%let wgroup1= wfemale ;
%let wgroup2= wage wagesqr;
%let wgroup3= wmiddleclass1 wmiddleclass2 wrich wzipincMISSING;
%let wgroup4= wMidwest wSouth wWest;
%let wgroup5= wMedicare wPrivate wMedicaid wSelfPay;
%let wgroup6= wAdmEmergent wAdmUrgent wAdmElective;
%let wgroup7= wbedsizemed wbedsizelarge wbedsizemiss;
%let wgroup8= wcontrolgov_nonfed wcontrolpriv_notprf wcontrolpriv_invest
wcontrolpriv_collapsed whosp_teaching;
%let wgroup9= wloc_urban;
%let wgroup10= wtotal_disc wtotal_disc_sqr wnumAnyAppy wnumAnyAppy_sqr;

* prepare original data for program;
data temp;
  set sasdata.tempORIG;
  mergeobs=1;
* define dependent variable;
  *y=hcomp;
* delete observations with any missing values for dep or indep vars;
  *if y=. or hsgrad=. or inc1015=. then delete;
* define sample - e.g. only keep working-age adults for this run;
  * if age<25 or age>55 then delete;

© 2011 American Medical Association. All rights reserved.
* define subset of data for estimating coefficients;
* use this to estimate coefficients using only one group e.g. white coefficients;
* currently set to use all groups (pooled sample) to estimate coefficients;

```sas
data temp2;
   set temp;

* create minority sample with minority variable names;
* define which minority group is used;
data minority (keep=ym &mvars mergeobs);
   set temp;
   array varsa(&k) &vars;
   array mvarsa(&k) &mvars;
   ym=y;
   do i=1 to &k;
      mvarsa(i)=varsa(i);
   end;
   if black=1 then output;

* create full white sample with white variable names;
data white (keep=yw &wvars mergeobs);
   set temp;
   array varsa(&k) &vars;
   array wvarsa(&k) &wvars;
   yw=y;
   do i=1 to &k;
      wvarsa(i)=varsa(i);
   end;
   if white=1 then output;

* print out full sample means;
proc means data=minority;
   title2 'Minority Means';
proc means data=white;
   title2 'White Means - Full Sample';

* calculate means of dependent variables for full sample;
* these values are used to calculate the total gap in the decomposition;
proc means data=minority noprint;
   var ym;
   output out=ymdata mean=ymfull;
proc means data=white noprint;
   var yw;
   output out=ywdata mean=ywfull;

* estimate logit model to obtain coefficients;
* set for pooled or specific group sample above;
proc logistic data=temp2 outest=orgcoefs covout descending;
   model y=&racevars &vars / link=logit;
   title2 'Logit for Coefficients';

* remove race dummies from coefficient dataset;
* only need this for pooled estimates;
data coefs (drop=&racevars _link_ _type_ _status_ _name_ _lnlike_);
set orgcoefs;
mergeobs=1;
if _n_=1; /* coefs are in first row */

* calculate predicted probabilities for both samples;
data white;
merge white coefs;
by mergeobs;
array coefsa(&k) &vars;
array wvarsa(&k) &wvars;
xbeta=intercept;
do i=1 to &k;
   xbeta=xbeta+wvarsa(i)*coefsa(i);
end;
wordprob=exp(xbeta)/(1+exp(xbeta));
data minority;
merge minority coefs;
by mergeobs;
array coefsa(&k) &vars;
array mvarsa(&k) &mvars;
xbeta=intercept;
do i=1 to &k;
   xbeta=xbeta+mvarsa(i)*coefsa(i);
end;
mordprob=exp(xbeta)/(1+exp(xbeta));

* sort minority data by predicted probabilities for later matching;
proc sort data=minority;
   by mordprob;

* create empty starting dataset for simulations;
data means2;
   set _null_;

* create macro for simulations;
%macro simulate;
%do i=1 %to &numsims;

* first, delete white observations to match black sample size;
data white1;
   set white;
   random1=ranuni(&i);
   proc sort data=white1;
      by random1;
data white2 (drop=ym);
   merge minority (keep=ym) white1;
   if ym=. then delete; /* deletes extra white observations */

* second, reorder random white subsample by predicted probabilities;
proc sort data=white2;
  by wordprob;

* third, merge datasets together for matching;
data combined;
  merge white2 minority;

* calculate decomposition components;
data combined;
  set combined;
  one=1;
  array coefsa(&k) &vars;
  * define distribution switches as arrays;
    array x0a(&k) &wgroup1 &wgroup2 &wgroup3 &wgroup4 &wgroup5 &wgroup6 &wgroup7 &wgroup8 &wgroup9 &wgroup10 &wgroup11;
    array x1a(&k) &mgroup1 &wgroup2 &wgroup3 &wgroup4 &wgroup5 &wgroup6 &wgroup7 &wgroup8 &wgroup9 &wgroup10 &wgroup11;
    array x2a(&k) &mgroup1 &mgroup2 &wgroup3 &wgroup4 &wgroup5 &wgroup6 &wgroup7 &wgroup8 &wgroup9 &wgroup10 &wgroup11;
    array x3a(&k) &mgroup1 &mgroup2 &mgroup3 &wgroup4 &wgroup5 &wgroup6 &wgroup7 &wgroup8 &wgroup9 &wgroup10 &wgroup11;
    array x4a(&k) &mgroup1 &mgroup2 &mgroup3 &mgroup4 &wgroup5 &wgroup6 &wgroup7 &wgroup8 &wgroup9 &wgroup10 &wgroup11;
    array x5a(&k) &mgroup1 &mgroup2 &mgroup3 &mgroup4 &mgroup5 &wgroup6 &wgroup7 &wgroup8 &wgroup9 &wgroup10 &wgroup11;
    array x6a(&k) &mgroup1 &mgroup2 &mgroup3 &mgroup4 &mgroup5 &mgroup6 &wgroup7 &wgroup8 &wgroup9 &wgroup10 &wgroup11;
    array x7a(&k) &mgroup1 &mgroup2 &mgroup3 &mgroup4 &mgroup5 &mgroup6 &mgroup7 &wgroup8 &wgroup9 &wgroup10 &wgroup11;
    array x8a(&k) &mgroup1 &mgroup2 &mgroup3 &mgroup4 &mgroup5 &mgroup6 &mgroup7 &mgroup8 &wgroup9 &wgroup10 &wgroup11;
    array x9a(&k) &mgroup1 &mgroup2 &mgroup3 &mgroup4 &mgroup5 &mgroup6 &mgroup7 &mgroup8 &mgroup9 &wgroup10 &wgroup11;
    array x10a(&k) &mgroup1 &mgroup2 &mgroup3 &mgroup4 &mgroup5 &mgroup6 &mgroup7 &mgroup8 &mgroup9 &mgroup10 &wgroup11;
    array x11a(&k) &mgroup1 &mgroup2 &mgroup3 &mgroup4 &mgroup5 &mgroup6 &mgroup7 &mgroup8 &mgroup9 &mgroup10 &mgroup11;

  xb0=intercept;
  xb1=intercept;
  xb2=intercept;
  xb3=intercept;
  xb4=intercept;
  xb5=intercept;
  xb6=intercept;
  xb7=intercept;
  xb8=intercept;
  xb9=intercept;
  xb10=intercept;
  xb11=intercept;

* perform white to black variable distribution switches;
do i=1 to &k;
    xb0=xb0+x0a(i)*coefs(i);
    xb1=xb1+x1a(i)*coefs(i);
    xb2=xb2+x2a(i)*coefs(i);
    xb3=xb3+x3a(i)*coefs(i);
    xb4=xb4+x4a(i)*coefs(i);
    xb5=xb5+x5a(i)*coefs(i);
    xb6=xb6+x6a(i)*coefs(i);
    xb7=xb7+x7a(i)*coefs(i);
    xb8=xb8+x8a(i)*coefs(i);
    xb9=xb9+x9a(i)*coefs(i);
    xb10=xb10+x10a(i)*coefs(i);
    xb11=xb11+x11a(i)*coefs(i);
end;

* calculate various predicted probabilities;
  pred0=exp(xb0)/(1+exp(xb0));
  pred1=exp(xb1)/(1+exp(xb1));
  pred2=exp(xb2)/(1+exp(xb2));
  pred3=exp(xb3)/(1+exp(xb3));
  pred4=exp(xb4)/(1+exp(xb4));
  pred5=exp(xb5)/(1+exp(xb5));
  pred6=exp(xb6)/(1+exp(xb6));
  pred7=exp(xb7)/(1+exp(xb7));
  pred8=exp(xb8)/(1+exp(xb8));
  pred9=exp(xb9)/(1+exp(xb9));
  pred10=exp(xb10)/(1+exp(xb10));
  pred11=exp(xb11)/(1+exp(xb11));

* calculate various pdfs for standard error calculations;
  fhat0=pred0*(1-pred0);
  fhat1=pred1*(1-pred1);
  fhat2=pred2*(1-pred2);
  fhat3=pred3*(1-pred3);
  fhat4=pred4*(1-pred4);
  fhat5=pred5*(1-pred5);
  fhat6=pred6*(1-pred6);
  fhat7=pred7*(1-pred7);
  fhat8=pred8*(1-pred8);
  fhat9=pred9*(1-pred9);
  fhat10=pred10*(1-pred10);
  fhat11=pred11*(1-pred11);

* create intercept component to derivatives;
  dc1db0=fhat0*one-fhat1*one;
  dc2db0=fhat1*one-fhat2*one;
  dc3db0=fhat2*one-fhat3*one;
  dc4db0=fhat3*one-fhat4*one;
  dc5db0=fhat4*one-fhat5*one;
  dc6db0=fhat5*one-fhat6*one;
* calculate contribution derivatives (delta method);
  array dc1dba(&k) dc1db1-dc1db&k;
  array dc2dba(&k) dc2db1-dc2db&k;
  array dc3dba(&k) dc3db1-dc3db&k;
  array dc4dba(&k) dc4db1-dc4db&k;
  array dc5dba(&k) dc5db1-dc5db&k;
  array dc6dba(&k) dc6db1-dc6db&k;
  array dc7dba(&k) dc7db1-dc7db&k;
  array dc8dba(&k) dc8db1-dc8db&k;
  array dc9dba(&k) dc9db1-dc9db&k;
  array dc10dba(&k) dc10db1-dc10db&k;
  array dc11dba(&k) dc11db1-dc11db&k;

* create other variable components to derivatives;
  do i=1 to &k;
    dc1dba(i)=fhat0*x0a(i)-fhat1*x1a(i);
    dc2dba(i)=fhat1*x1a(i)-fhat2*x2a(i);
    dc3dba(i)=fhat2*x2a(i)-fhat3*x3a(i);
    dc4dba(i)=fhat3*x3a(i)-fhat4*x4a(i);
    dc5dba(i)=fhat4*x4a(i)-fhat5*x5a(i);
    dc6dba(i)=fhat5*x5a(i)-fhat6*x6a(i);
    dc7dba(i)=fhat6*x6a(i)-fhat7*x7a(i);
    dc8dba(i)=fhat7*x7a(i)-fhat8*x8a(i);
    dc9dba(i)=fhat8*x8a(i)-fhat9*x9a(i);
    dc10dba(i)=fhat9*x9a(i)-fhat10*x10a(i);
    dc11dba(i)=fhat10*x10a(i)-fhat11*x11a(i);
  end;

*****************************;
* calculate standard errors *;
*****************************;
* clean up coefficient/covariance dataset;
* calculate decomposition estimates to save and use for variance calculations;
proc means data=combined noprint;
  var yw ym pred0-pred11
output out=means1 mean=;

* create separate datasets to read into proc iml;
* NOTE: check to make sure variables are in the proper order to match to the covariance matrix;

  data cont1 (keep=dc1db0 dc1db1-dc1db&k)
  cont2 (keep=dc2db0 dc2db1-dc2db&k)
  cont3 (keep=dc3db0 dc3db1-dc3db&k)
  cont4 (keep=dc4db0 dc4db1-dc4db&k)
  cont5 (keep=dc5db0 dc5db1-dc5db&k)
  cont6 (keep=dc6db0 dc6db1-dc6db&k)
  cont7 (keep=dc7db0 dc7db1-dc7db&k)
  cont8 (keep=dc8db0 dc8db1-dc8db&k)
  cont9 (keep=dc9db0 dc9db1-dc9db&k)
  cont10 (keep=dc10db0 dc10db1-dc10db&k)
  cont11 (keep=dc11db0 dc11db1-dc11db&k);
  set means1;

  proc iml;
  use covar;
  read all var _num_ into V;
  use cont1;
  read all var _num_ into DC1DB;
  use cont2;
  read all var _num_ into DC2DB;
  use cont3;
  read all var _num_ into DC3DB;
  use cont4;
  read all var _num_ into DC4DB;
  use cont5;
  read all var _num_ into DC5DB;
  use cont6;
  read all var _num_ into DC6DB;
  use cont7;
  read all var _num_ into DC7DB;
  use cont8;
  read all var _num_ into DC8DB;
  use cont9;
  read all var _num_ into DC9DB;
use cont10;
read all var _num_ into DC10DB;
use cont11;
read all var _num_ into DC11DB;

* calculate standard error;
VAR1=DC1DB*V*t(DC1DB);
VAR2=DC2DB*V*t(DC2DB);
VAR3=DC3DB*V*t(DC3DB);
VAR4=DC4DB*V*t(DC4DB);
VAR5=DC5DB*V*t(DC5DB);
VAR6=DC6DB*V*t(DC6DB);
VAR7=DC7DB*V*t(DC7DB);
VAR8=DC8DB*V*t(DC8DB);
VAR9=DC9DB*V*t(DC9DB);
VAR10=DC10DB*V*t(DC10DB);
VAR11=DC11DB*V*t(DC11DB);

create vardata var {var1 var2 var3 var4 var5 var6 var7 var8 var9 var10
var11};
append;

* merge variance calculations from proc iml to decomp dataset;
data means1 (keep=yw ym pred0-pred11 var1-var11);
   merge means1 vardata;
* append latest simulation results to all previous simulation results;
data means2;
   set means2 means1;
%end;
%mend;
run;

* turn off notes because macro generates a lot of information;
* remove this option for debugging;
options nonotes;
* run simulation - note that it runs numsims times because of do loop above;
%simulate;
run;

* calculate contribution estimates from changes in predicted probabilities;
data means2;
   set means2;
   cont1=pred0-pred1;
   cont2=pred1-pred2;
   cont3=pred2-pred3;
cont4=pred3-pred4;
cont5=pred4-pred5;
cont6=pred5-pred6;
cont7=pred6-pred7;
cont8=pred7-pred8;
cont9=pred8-pred9;
cont10=pred9-pred10;
cont11=pred10-pred11;
cont12=pred11-pred12;
cont13=pred12-pred13;

* calculate means of decomposition runs;
proc means data=means2;
   title2 'Mean Values of Contribution Estimates from Simulations';
   var yw pred0 pred1-pred11 ym cont1-cont11 var1-var11;
   output out=meandecomp mean=;

* append the full sample means for ys and calculate percent contributions;
* dataset now has only one obs for the means;
data meandecomp;
   merge meandecomp ywdata ymdata;
   gap=ywfull-ymfull;
   perc1=cont1/gap;
   perc2=cont2/gap;
   perc3=cont3/gap;
   perc4=cont4/gap;
   perc5=cont5/gap;
   perc6=cont6/gap;
   perc7=cont7/gap;
   perc8=cont8/gap;
   perc9=cont9/gap;
   perc10=cont10/gap;
   perc11=cont11/gap;
   se1=sqrt(var1);
   se2=sqrt(var2);
   se3=sqrt(var3);
   se4=sqrt(var4);
   se5=sqrt(var5);
   se6=sqrt(var6);
   se7=sqrt(var7);
   se8=sqrt(var8);
   se9=sqrt(var9);
   se10=sqrt(var10);
   se11=sqrt(var11);

* format output for final decomposition table;
* outputs contribution estimates, gap percents and standard errors;
proc means data=meandecomp;
   title2 'Final Output for Table - Mean Values of Decomposition Runs';
   var ywfull ymfull gap
      cont1 se1 perc1 cont2 se2 perc2 cont3 se3 perc3 cont4 se4 perc4
      cont5 se5 perc5 cont6 se6 perc6

© 2011 American Medical Association. All rights reserved.
The following is an R implementation of the decomposition algorithm. This was adapted from the SAS code that is also included in this appendix.

The following steps should be followed to use this algorithm.
1. Create a comma separated variable (csv) data file with the variable names in the 1st row of the file. The file used in the R code below is Temporig. Note that R is a case sensitive language so that capital and small letters are not the same. The order in which the variables appear in the file's columns is not important. There can be extraneous columns of data that the program will ignore (i.e. columns of data for variables not used in the gap analysis).

The current version of the algorithm must have the dependent variable labeled y, the minority indicator variable labeled "Black" and the white variable labeled "White". Black should have fewer subjects than White. Data with missing variables will be ignored by this program.

2. Create a csv file named "Variables.csv" This should be a csv file. The first row must contain the variable names Variables in column 1, Groups in column 2, Race in column 3, Dependent in Column 4 and Comparison in Column 5.

The variables used in the gap analysis should be listed in Column 1. Reference variables are excluded (If female is compared to male with male being the reference, male is excluded). The groups that the individual variables are aggregated into are listed in column 2. The algorithm has not been tested for random group assignment. Groups should be ordered from 1 to the last group number, in order. Column 3 has all the race variable names in the file created in step #1, except White. The dependent variable is placed in row 2 of column 4 and the groups compared in column 5 with the larger group in row 2. In the current version of the algorithm, columns 4 and 5 are not used.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Groups</th>
<th>Race</th>
<th>Dependent</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEMALE</td>
<td>1</td>
<td>Black</td>
<td>y</td>
<td>White</td>
</tr>
<tr>
<td>AGE</td>
<td>2</td>
<td>Hispanic</td>
<td></td>
<td>Black</td>
</tr>
<tr>
<td>agesqr</td>
<td>2</td>
<td>API</td>
<td></td>
<td></td>
</tr>
<tr>
<td>middleclass1</td>
<td>3</td>
<td>AMIAN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Place these files in a directory. The directory used for this implementation was called C:/FilesUTS/Disparity/Rdecomp. Note that R uses forward and not backslashes for directory names. The directory is identified in steps

```
wd = "C:/FilesUTS/Disparity/Rdecomp"
setwd(wd)
```

4. Copy and paste the program below into R. As the program runs various calculations and results are presented. The averages are listed and also written to a csv file named "averages". The gap between White and Black is
also listed as the program runs. The number of iterations must be designated:

numsims=1000

The sample program is set at 1000. You should reduce this to 1 to test your file structure and ensure there are no errors. A large file can take 15-30 minutes to run 1,000 simulations. One common error is to not have the variable label names be exactly the same between the 2 input files. We recommend running 1,000 simulations for the final gap analysis.

5. The program will create 2 csv files. One is called "averages". It contains the mean values for the variables used by the program in aggregate, for the White and for the Black variables. The other file is called "gap" It contains the program output. Each column represents a group numbered as they were in the Variables file. Group names are preceded by the letter "V". Gap outputs are provided and the actual gap and its standard error (SE). The percentage (not fraction) of the gap explained by the group is provided in the last line of the file. This is calculated by multiplying the gap in row 2 * 100 and dividing by the gap in probability between White and Black.

```r
library(Design)             ### Regression package Design must be installed into R for this program

wd = "C:/FilesUTS/Disparity/Rdecomp"
setwd(wd)

############ specify number of simulations HERE ############
############ NOTE: change this to 1000 for final run ############

numsims=1000

############ input CSV variable file  First column=names, 2nd=group numbers
############ 3rd column=race variables, 4th=dependent variable name
############ reference variables are placed in column 1 and assigned group=0
############ routine for identifying reference variable not written yet

dat = read.csv("Variables.csv", header = T, na.strings = "x", fill = FALSE, colClasses = "character")
```

© 2011 American Medical Association. All rights reserved.
##### input CSV Data file

##### Black is name in data file that is used for MINORITY

##### White is name in data file that is used for Comparator

##### y is the dependent variable

temp= read.csv("Temporig.csv", header = T, na.strings = "x")

na.omit(temp) ### omit missing values

##### build DR character vector of Race names
numRace=0
for(i in 1:length(dat$Race))
  if(dat$Race[i]!="")numRace=numRace+1  ### find how many race variables there are

DR=vector(mode="character",length=numRace)
for(i in 1:numRace)
  DR[i]=dat$Race[i]  #### DR is a vector of race names

VarNamesTemp=colnames(temp)

##### Eliminate un-needed columns
z=temp[,!(VarNamesTemp %in% dat$Variables)]

#### Reorder matrix so that columns are ordered the same as in the dat$Variables vector
yz=colnames(z)
ydat=dat$Variables

zz=z

for(i in 1:length(ydat))
  for(j in 1:length(yz))
    if (ydat[i]==yz[j])
    {
      zz[,i]=z[,j]
      colnames(zz)[i]=colnames(z)[j]
    }

z=zz

#### Done reordering

DR2=DR[!(DR %in% "White") ]
DR2=DR2[!(DR2 %in% "Black") ]
zz = temp[, (colnames(temp) %in% DR2)]
y = temp$Y
Black = temp$Black
White = temp$White  #### y, White and Black need to be in the main file
temp = cbind(z, zz, Black, White, y)
intercept = rep(1, nrow(temp))
temp = cbind(intercept, temp)

######## Build minority and white array
minority = subset(temp, Black == 1)
white = subset(temp, White == 1)

######## Find mean values for variables
avg = mean(temp[, seq(1:length(colnames(temp)))])
mavg = mean(minority[, seq(1:length(colnames(minority)))])
wavg = mean(white[, seq(1:length(colnames(white)))])

ym = mean(minority[, "y"])
yw = mean(white[, "y"])
cat(" Minority Mean Response = ", ym, "\n", "White Mean Response = ", yw, "\n")
a = cbind(avg, mavg, wavg)
formatC(a, format="f", digits=3, big.mark=" ")
write.csv(a, "averages")  #### writes file "averages" to directory that has each variable's average values

######## Pooled regression
varsx = vector("integer", length(dat$Variables))  #### Create vars variable = regression coef
racex = vector("integer", numRace)
varsx = colnames(temp)
varsx = varsx[varsx != "intercept"]
for (i in 1: numRace)
    racex[i] = dat$Race[i]
varsx = subset(varsx, varsx != "y")
varsx = subset(varsx, varsx != "White")
vars = paste(varsx, collapse = "+")
racevars = paste(racex, collapse = "+")

#### estimate logit model to obtain coefficients
#### Logit for Coefficients
F1=as.formula(paste("y-",racevars,vars, sep = " + "))  ###need as.formula to use concated strings in lrm

zz=temp[,!(colnames(temp) %in% c("intercept"))]

pooled=lrm(F1,zz)
PooledCoef=as.vector(pooled$coef)

##### remove race dummies from coefficient dataset
orgcoefs=pooled$coefficients
for(i in 1:numRace)
orgcoefs = orgcoefs[!names(orgcoefs)==racex[i]]
TruncPooledCoef=as.vector(orgcoefs)

##### calculate predicted probabilities for both samples

Wbeta=vector("numeric",length(temp$AGE))
Mbeta=vector("numeric",length(temp$AGE))
wordprob=vector("numeric",length(temp$AGE))
mordprob=vector("numeric",length(temp$AGE))

wtemp=white
for(i in 1: numRace)
  wtemp[,racex[i]]=0

mtemp=minority
for(i in 1: numRace)
  mtemp[,racex[i]]=0

Wbeta=predict(pooled,newdata=wtemp)
wordprob=exp(Wbeta)/(1+exp(Wbeta))

Mbeta=predict(pooled,newdata=mtemp)
mordprob=exp(Mbeta)/(1+exp(Mbeta))

minority=cbind(minority,mordprob)
white=cbind(white,wordprob)

##### sort minority data by predicted probabilities for later matching
minority=minority[order(minority$mordprob),]  ## TEST 061711
white=white[order(white$wordprob),]  ## TEST 061711
### set up variables for macro

GrpVars=dat$Variables                                    ## List of variables for switches
NumVars=length(GrpVars)                          ## Number of variables
NumGrps=max(as.integer(dat$Groups))         ## Number of groups
NumGrps1= NumGrps+1                        ## One more than NumGrps-for loop indexing
NumVars1=NumVars+1                   ## 1 more than the number of variables to account for the intercept

xb=matrix(nrow=nrow(minority),ncol=NumGrps1)
xbTemp=matrix(nrow=nrow(minority),ncol=NumGrps1)
dcdb0=matrix(nrow=nrow(minority),ncol=NumGrps1)
dcdba=matrix(nrow=NumVars,ncol=NumGrps)
Mmatrix=array(diag(1,NumVars1,NumVars1),c(NumVars1,NumVars1,NumGrps1))    ##matrix for switchs
Wmatrix=array(diag(0,NumVars1,NumVars1),c(NumVars1,NumVars1,NumGrps1))
dcdba=array(0,c(nrow(minority),NumVars,NumGrps1))
dcdbaMeans=matrix(0,nrow=NumGrps,ncol=NumVars)
ywavgCUM=0
ymCUM=0
predMeansCUM=matrix(0,nrow=1,ncol=NumGrps1)
varCUM=matrix(0,nrow=1,ncol=NumGrps)

Mmatrix[,,1]=diag(0,NumVars1,NumVars1)         ## Subscript 0 in SAS is 1 in R code
Wmatrix[,,1]=diag(1,NumVars1,NumVars1)         ## matrix goes from 1 to NumGrps + 1
Wmatrix[1,1,]=1                                                         ## to include intercept
Mmatrix[1,1,]=0
GrpIndex=1

for(i in 2:NumGrps1)
{
    for(j in 2:NumVars1)
        if(GrpIndex < as.integer(dat$Groups[j-1]))
            {
            Mmatrix[j,j,i]=0  ## Mmatrix mutiplies minority data
            Wmatrix[j,j,i]=1  ## Wmatrix mutiplies White data
            }
        GrpIndex=GrpIndex+1
    }

### Need to build xb0 matrix (could use Mmatrix12 for Wmatrix and vise versa)

white2=white[(!(colnames(white) %in% "DR"))]
tempa=c("White", "wordprob")
white2=white2[(!(colnames(white2) %in% tempa)]
white2 = apply(white2, 2, as.numeric)

minority2 = minority[, !(colnames(minority) %in% DR)]
tempa = c("White", "y", "mordprob")
minority2 = minority2[, !(colnames(minority2) %in% tempa)]
minority2 = apply(minority2, 2, as.numeric)

#TruncPooledCoef = PooledCoef[1:NumVars]

##### create macro for simulations  ###########################
for(ii in 1:numsims)
{
    ###### START SIMULATION MACRO

white3 = white2[sample(length(white$AGE), length(minority$AGE)),] ###random sample of length minority
ywavg = mean(white["y"])
white3 = white3[, !(colnames(white3) %in% c("y"))]

for(i in 1:NumGrps1)
    xb[,i] = white3 %*% Wmatrix[,i] %*% TruncPooledCoef

for(i in 1:NumGrps1)
    xbTemp[,i] = minority2 %*% Mmatrix[,i] %*% TruncPooledCoef

xb = xb + xbTemp

pred = exp(xb)/(1+exp(xb))  # calculate various predicted probabilities
fhat = pred*(1-pred)        # calculate various pdfs for standard error calculations
for(i in 1:NumGrps)
    dcdb0[,i] = fhat[,i] - fhat[,i+1]

    #Calculate contribution derivative

white4 = white3[,colnames(white3) != "intercept"]  # no longer need intercept
minority3 = minority2[,colnames(minority2) != "intercept"]
Wmatrix1 = Wmatrix[2:NumVars1,2:NumVars1,1:NumGrps1]
Mmatrix1 = Mmatrix[2:NumVars1,2:NumVars1,1:NumGrps1]

for(i in 1:NumGrps)
{
    j = i+1
    x1 = (white4 %*% Wmatrix1[,i])
    x2 = (minority3 %*% Mmatrix1[,i])
    x3 = (x1+x2) * fhat[,i]
    x1 = (white4 %*% Wmatrix1[,j])
    x2 = (minority3 %*% Mmatrix1[,j])
    y3 = (x1+x2) * fhat[,j]

    #Calculate contribution derivative
## Calculate Standard Errors ##

```r
predMeans = colMeans(pred)
dcdb0Means = colMeans(dcdb0)
dcdb0Means = dcdb0Means[1:NumGrps]
for (i in 1:NumGrps)
    dcdbaMeans[,i] = colMeans(dcdb[,i])
dcdb = cbind(dcdb0Means, dcdbaMeans)
covar = pooled$var[, !(colnames(pooled$var) %in% DR)]
covar = covar[, !(rownames(covar) %in% DR),]
varD = dcdb %*% covar %*% t(dcdb)
var = diag(varD)
```

```r
ywavgCUM = rbind(ywavgCUM, ywavg)
yMCM = rbind(yMCM, ym)
predMeansCUM = rbind(predMeansCUM, predMeans)
varCUM = rbind(varCUM, var)
```

```r
## calculate contribution estimates from changes in predicted probabilities

ywavgCUM = ywavgCUM[2:nrow(ywavgCUM),]
yMCM = yMCM[2:nrow(yMCM),]
predMeansCUM = predMeansCUM[2:nrow(predMeansCUM),]
varCUM = varCUM[2:nrow(varCUM),]
cont = matrix(0, nrow = numsims, ncol = NumGrps)
for (i in 1:NumGrps)
    cont[,i] = predMeansCUM[,i] - predMeansCUM[,i+1]

gap = yw - ym
perc = 100 * colMeans(cont) / gap
## Provided as percent NOT fraction
se = sqrt(colMeans(varCUM))
FinalOut = rbind(colMeans(cont), se, perc, deparse.level = 1)

formatC(FinalOut, format = "f", digits = 6, big.mark = " ")
write.csv(FinalOut, "Gap Output")
```

```r
print("Observed white probability")
yw
print("calculated white probability")
mean(ywavgCUM)

print("Observed minority probability")
ym
print("calculated white probability")
```

© 2011 American Medical Association. All rights reserved.
mean(ymCUM)

print("GAP= (white - minority)")
gap

## STOP COPING HERE