ORF 307: Lecture 11

Linear Programming: Chapter 12
Regression

Robert J. Vanderbei

March 28, 2019

Slides last edited on April 2, 2019
Outline

• Means and Medians

• Least Squares Regression

• Least Absolute Deviation (LAD) Regression

• LAD via LP

• Average Complexity of Parametric Self-Dual Simplex Method
Consider 1995 Adjusted Gross Incomes on Individual Tax Returns:

<table>
<thead>
<tr>
<th>Individual</th>
<th>AGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_1$</td>
<td>$25,462$</td>
</tr>
<tr>
<td>$b_2$</td>
<td>$45,110$</td>
</tr>
<tr>
<td>$b_3$</td>
<td>$15,505$</td>
</tr>
<tr>
<td>$b_{m-1}$</td>
<td>$33,265$</td>
</tr>
<tr>
<td>$b_m$</td>
<td>$75,420$</td>
</tr>
</tbody>
</table>

Real summary data is shown on the next slide...
Table 1. – 2014, Individual Income Tax Returns

Monetary amounts in 3rd column are in thousands of dollars

<table>
<thead>
<tr>
<th>Size of adjusted gross income</th>
<th>Number of returns</th>
<th>Adjusted gross income</th>
</tr>
</thead>
<tbody>
<tr>
<td>All returns</td>
<td>148,606,578</td>
<td>9,771,035,412</td>
</tr>
<tr>
<td>No adjusted gross income</td>
<td>2,034,138</td>
<td>-197,690,795</td>
</tr>
<tr>
<td>$1 under $5,000</td>
<td>10,262,509</td>
<td>26,379,097</td>
</tr>
<tr>
<td>$5,000 under $10,000</td>
<td>11,790,191</td>
<td>89,719,121</td>
</tr>
<tr>
<td>$10,000 under $15,000</td>
<td>12,289,794</td>
<td>153,830,822</td>
</tr>
<tr>
<td>$15,000 under $20,000</td>
<td>11,331,450</td>
<td>197,774,439</td>
</tr>
<tr>
<td>$20,000 under $25,000</td>
<td>10,061,750</td>
<td>226,042,578</td>
</tr>
<tr>
<td>$25,000 under $30,000</td>
<td>8,818,876</td>
<td>241,769,583</td>
</tr>
<tr>
<td>$30,000 under $40,000</td>
<td>14,599,675</td>
<td>507,486,039</td>
</tr>
<tr>
<td>$40,000 under $50,000</td>
<td>11,472,714</td>
<td>513,959,724</td>
</tr>
<tr>
<td>$50,000 under $75,000</td>
<td>19,394,648</td>
<td>1,191,956,661</td>
</tr>
<tr>
<td>$75,000 under $100,000</td>
<td>12,825,769</td>
<td>1,111,626,170</td>
</tr>
<tr>
<td>$100,000 under $200,000</td>
<td>17,501,251</td>
<td>2,361,756,261</td>
</tr>
<tr>
<td>$200,000 under $500,000</td>
<td>4,978,534</td>
<td>1,419,776,711</td>
</tr>
<tr>
<td>$500,000 under $1,000,000</td>
<td>834,981</td>
<td>562,622,816</td>
</tr>
<tr>
<td>$1,000,000 under $1,500,000</td>
<td>180,446</td>
<td>217,426,739</td>
</tr>
<tr>
<td>$1,500,000 under $2,000,000</td>
<td>77,065</td>
<td>132,463,053</td>
</tr>
<tr>
<td>$2,000,000 under $5,000,000</td>
<td>109,475</td>
<td>326,511,879</td>
</tr>
<tr>
<td>$5,000,000 under $10,000,000</td>
<td>26,579</td>
<td>181,943,504</td>
</tr>
<tr>
<td>$10,000,000 or more</td>
<td>16,733</td>
<td>505,681,010</td>
</tr>
</tbody>
</table>

Means and Medians

Median:

\[ \hat{x} = b^{1+m} \approx \$35,270. \]

Mean:

\[ \bar{x} = \frac{1}{m} \sum_{i=1}^{m} b_i \]

\[ = \frac{9,771,035,412,000}{148,606,578} \]

\[ = \$65,751. \]
\[ \bar{x} = \arg\min_{x} \sum_{i=1}^{m} (x - b_i)^2. \]

Proof:

\[
\begin{align*}
  f(x) &= \sum_{i=1}^{m} (x - b_i)^2 \\
  f'(x) &= \sum_{i=1}^{m} 2(x - b_i) \\
  f'(&\bar{x}) = 0 \quad \implies \quad \bar{x} = \frac{1}{m} \sum_{i=1}^{m} b_i \\
  \lim_{x \to \pm\infty} f(x) &= +\infty \quad \implies \quad \bar{x} \text{ is a minimum}
\end{align*}
\]
Median’s Connection with Optimization

\[ \hat{x} = \arg\min_x \sum_{i=1}^{m} |x - b_i|. \]

Proof:

\[ f(x) = \sum_{i=1}^{m} |x - b_i| \]

\[ f'(x) = \sum_{i=1}^{m} \text{sgn} (x - b_i) \]

where \( \text{sgn}(x) = \begin{cases} 
1 & x > 0 \\
0 & x = 0 \\
-1 & x < 0 
\end{cases} \)

\[ = (\# \text{ of } b_i\text{'s smaller than } x) - (\# \text{ of } b_i\text{'s larger than } x). \]

If \( m \) is odd:
Regression
### Parametric Self-Dual Simplex Method: Data

<table>
<thead>
<tr>
<th>Name</th>
<th>$m$</th>
<th>$n$</th>
<th>iters</th>
<th>Name</th>
<th>$m$</th>
<th>$n$</th>
<th>iters</th>
</tr>
</thead>
<tbody>
<tr>
<td>25fv47</td>
<td>777</td>
<td>1545</td>
<td>5089</td>
<td>nesm</td>
<td>646</td>
<td>2740</td>
<td>5829</td>
</tr>
<tr>
<td>80bau3b</td>
<td>2021</td>
<td>9195</td>
<td>10514</td>
<td>recipe</td>
<td>74</td>
<td>136</td>
<td>80</td>
</tr>
<tr>
<td>adlittle</td>
<td>53</td>
<td>96</td>
<td>141</td>
<td>sc105</td>
<td>104</td>
<td>103</td>
<td>92</td>
</tr>
<tr>
<td>afiro</td>
<td>25</td>
<td>32</td>
<td>16</td>
<td>sc205</td>
<td>203</td>
<td>202</td>
<td>191</td>
</tr>
<tr>
<td>agg2</td>
<td>481</td>
<td>301</td>
<td>204</td>
<td>sc50a</td>
<td>49</td>
<td>48</td>
<td>46</td>
</tr>
<tr>
<td>agg3</td>
<td>481</td>
<td>301</td>
<td>193</td>
<td>sc50b</td>
<td>48</td>
<td>48</td>
<td>53</td>
</tr>
<tr>
<td>bandm</td>
<td>224</td>
<td>379</td>
<td>1139</td>
<td>scagr25</td>
<td>347</td>
<td>499</td>
<td>1336</td>
</tr>
<tr>
<td>beaconfd</td>
<td>111</td>
<td>172</td>
<td>113</td>
<td>scagr7</td>
<td>95</td>
<td>139</td>
<td>339</td>
</tr>
<tr>
<td>blend</td>
<td>72</td>
<td>83</td>
<td>117</td>
<td>scfxm1</td>
<td>282</td>
<td>439</td>
<td>531</td>
</tr>
<tr>
<td>bnl1</td>
<td>564</td>
<td>1113</td>
<td>2580</td>
<td>scfxm2</td>
<td>564</td>
<td>878</td>
<td>1197</td>
</tr>
<tr>
<td>bnl2</td>
<td>1874</td>
<td>3134</td>
<td>6381</td>
<td>scfxm3</td>
<td>846</td>
<td>1317</td>
<td>1886</td>
</tr>
<tr>
<td>boeing1</td>
<td>298</td>
<td>373</td>
<td>619</td>
<td>scorpion</td>
<td>292</td>
<td>331</td>
<td>411</td>
</tr>
<tr>
<td>boeing2</td>
<td>125</td>
<td>143</td>
<td>168</td>
<td>scrs8</td>
<td>447</td>
<td>1131</td>
<td>783</td>
</tr>
<tr>
<td>bore3d</td>
<td>138</td>
<td>188</td>
<td>227</td>
<td>scsd1</td>
<td>77</td>
<td>760</td>
<td>172</td>
</tr>
<tr>
<td>brandy</td>
<td>123</td>
<td>205</td>
<td>585</td>
<td>scsd6</td>
<td>147</td>
<td>1350</td>
<td>494</td>
</tr>
<tr>
<td>czprob</td>
<td>689</td>
<td>2770</td>
<td>2635</td>
<td>scsd8</td>
<td>397</td>
<td>2750</td>
<td>1548</td>
</tr>
<tr>
<td>d6cube</td>
<td>403</td>
<td>6183</td>
<td>5883</td>
<td>sctap1</td>
<td>284</td>
<td>480</td>
<td>643</td>
</tr>
<tr>
<td>degen2</td>
<td>444</td>
<td>534</td>
<td>1421</td>
<td>sctap2</td>
<td>1033</td>
<td>1880</td>
<td>1037</td>
</tr>
<tr>
<td>degen3</td>
<td>1503</td>
<td>1818</td>
<td>6398</td>
<td>sctap3</td>
<td>1408</td>
<td>2480</td>
<td>1339</td>
</tr>
<tr>
<td>e226</td>
<td>162</td>
<td>260</td>
<td>598</td>
<td>seba</td>
<td>449</td>
<td>896</td>
<td>766</td>
</tr>
<tr>
<td>Name</td>
<td>$m$</td>
<td>$n$</td>
<td>iters</td>
<td>Name</td>
<td>$m$</td>
<td>$n$</td>
<td>iters</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>------------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>etamacro</td>
<td>334</td>
<td>542</td>
<td>1580</td>
<td>share1b</td>
<td>107</td>
<td>217</td>
<td>404</td>
</tr>
<tr>
<td>fffff800</td>
<td>476</td>
<td>817</td>
<td>1029</td>
<td>share2b</td>
<td>93</td>
<td>79</td>
<td>189</td>
</tr>
<tr>
<td>finnis</td>
<td>398</td>
<td>541</td>
<td>680</td>
<td>shell</td>
<td>487</td>
<td>1476</td>
<td>1155</td>
</tr>
<tr>
<td>fit1d</td>
<td>24</td>
<td>1026</td>
<td>925</td>
<td>ship04l</td>
<td>317</td>
<td>1915</td>
<td>597</td>
</tr>
<tr>
<td>fit1p</td>
<td>627</td>
<td>1677</td>
<td>15284</td>
<td>ship04s</td>
<td>241</td>
<td>1291</td>
<td>560</td>
</tr>
<tr>
<td>forplan</td>
<td>133</td>
<td>415</td>
<td>576</td>
<td>ship08l</td>
<td>520</td>
<td>3149</td>
<td>1091</td>
</tr>
<tr>
<td>ganges</td>
<td>1121</td>
<td>1493</td>
<td>2716</td>
<td>ship08s</td>
<td>326</td>
<td>1632</td>
<td>897</td>
</tr>
<tr>
<td>greenbea</td>
<td>1948</td>
<td>4131</td>
<td>21476</td>
<td>ship12l</td>
<td>687</td>
<td>4224</td>
<td>1654</td>
</tr>
<tr>
<td>grow15</td>
<td>300</td>
<td>645</td>
<td>681</td>
<td>ship12s</td>
<td>417</td>
<td>1996</td>
<td>1360</td>
</tr>
<tr>
<td>grow22</td>
<td>440</td>
<td>946</td>
<td>999</td>
<td>sierra</td>
<td>1212</td>
<td>2016</td>
<td>793</td>
</tr>
<tr>
<td>grow7</td>
<td>140</td>
<td>301</td>
<td>322</td>
<td>standata</td>
<td>301</td>
<td>1038</td>
<td>74</td>
</tr>
<tr>
<td>israel</td>
<td>163</td>
<td>142</td>
<td>209</td>
<td>standmps</td>
<td>409</td>
<td>1038</td>
<td>295</td>
</tr>
<tr>
<td>kb2</td>
<td>43</td>
<td>41</td>
<td>63</td>
<td>stocfor1</td>
<td>98</td>
<td>100</td>
<td>81</td>
</tr>
<tr>
<td>lotfi</td>
<td>134</td>
<td>300</td>
<td>242</td>
<td>stocfor2</td>
<td>2129</td>
<td>2015</td>
<td>2127</td>
</tr>
<tr>
<td>maros</td>
<td>680</td>
<td>1062</td>
<td>2998</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A Regression Model for Algorithm Efficiency

Observed Data:

\[ t = \# \text{ of iterations} \]
\[ m = \# \text{ of constraints} \]
\[ n = \# \text{ of variables} \]

Model:

\[ t \approx 2^\alpha (m + n)^\beta \]

Linearization: Take logs:

\[ \log t = \alpha \log 2 + \beta \log(m + n) + \epsilon \]
\[ \uparrow \]
\[ \text{error} \]
Solve several instances (say $N$ of them):

\[
\begin{bmatrix}
\log t_1 \\
\log t_2 \\
\vdots \\
\log t_N \\
\end{bmatrix} = \begin{bmatrix}
\log 2 & \log(m_1 + n_1) \\
\log 2 & \log(m_2 + n_2) \\
\vdots & \vdots \\
\log 2 & \log(m_N + n_N) \\
\end{bmatrix} \begin{bmatrix}
\alpha \\
\beta \\
\end{bmatrix} + \begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_N \\
\end{bmatrix}
\]

In matrix notation:

\[b = Ax + \epsilon\]

**Goal:** find $x$ that “minimizes” $\epsilon$. 
Least Squares Regression

**Euclidean Distance:**  \( \|x\|_2 = \left(\sum_i x_i^2\right)^{1/2} \)

**Least Squares Regression:**  \( \bar{x} = \text{argmin}_x \| b - Ax \|_2^2 \)

**Calculus:**

\[
f(x) = \| b - Ax \|_2^2 = \sum_i \left( b_i - \sum_j a_{ij} x_j \right)^2
\]

\[
\frac{\partial f}{\partial x_k}(\bar{x}) = \sum_i 2 \left( b_i - \sum_j a_{ij} \bar{x}_j \right) (-a_{ik}) = 0, \quad k = 1, 2, \ldots, n
\]

Rearranging,

\[
\sum_i a_{ik} b_i = \sum_i \sum_j a_{ik} a_{ij} \bar{x}_j, \quad k = 1, 2, \ldots, n
\]

In matrix notation,

\[
A^T b = A^T A \bar{x}
\]

Assuming \( A^T A \) is invertible,

\[
\bar{x} = \left( A^T A \right)^{-1} A^T b
\]
Least Absolute Deviation Regression

**Manhattan Distance:** \( \| x \|_1 = \sum_i |x_i| \)

**Least Absolute Deviation Regression:** \( \hat{x} = \arg\min_x \| b - Ax \|_1 \)

**Calculus:**

\[
f(x) = \| b - Ax \|_1 = \sum_i \left| b_i - \sum_j a_{ij} x_j \right|
\]

\[
\frac{\partial f}{\partial x_k}(\hat{x}) = \sum_i \frac{b_i - \sum_j a_{ij} \hat{x}_j}{\left| b_i - \sum_j a_{ij} \hat{x}_j \right|} (-a_{ik}) = 0, \quad k = 1, 2, \ldots, n
\]

Rearranging,

\[
\sum_i \frac{a_{ik} b_i}{\epsilon_i(\hat{x})} = \sum_i \sum_j \frac{a_{ik} a_{ij} \hat{x}_j}{\epsilon_i(\hat{x})}, \quad k = 1, 2, \ldots, n
\]

In matrix notation,

\[
A^T E(\hat{x}) b = A^T E(\hat{x}) A \hat{x}, \quad \text{where } E(\hat{x}) = \text{Diag}(\epsilon(\hat{x}))^{-1}
\]

Assuming \( A^T E(\hat{x}) A \) is invertible,

\[
\hat{x} = \left( A^T E(\hat{x}) A \right)^{-1} A^T E(\hat{x}) b
\]
An implicit equation.

Can be solved using *successive approximations*:

\[
\begin{align*}
x^0 &= 0 \\
x^1 &= \left( A^T E(x^0) A \right)^{-1} A^T E(x^0) b \\
x^2 &= \left( A^T E(x^1) A \right)^{-1} A^T E(x^1) b \\
&\quad \vdots \\
x^{k+1} &= \left( A^T E(x^k) A \right)^{-1} A^T E(x^k) b \\
&\quad \vdots \\
\hat{x} &= \lim_{k \to \infty} x^k
\end{align*}
\]
Least Absolute Deviation Regression via LP

First of Two Methods

\[ \min \sum_{i} \left| b_i - \sum_{j} a_{ij} x_j \right| \]

Equivalent Linear Program:

\[ \min \sum_{i} t_i \]
\[ -t_i \leq b_i - \sum_{j} a_{ij} x_j \leq t_i \quad i = 1, 2, \ldots, m \]
param m;
param n;

set I := {1..m};
set J := {1..n};

param A {I,J};
param b {I};

var x{J};
var t{I};

minimize sum_dev: sum {i in I} t[i];

subject to lower_bound {i in I}:
    -t[i] <= b[i] - sum {j in J} A[i,j]*x[j];

subject to upper_bound {i in I}:
    b[i] - sum {j in J} A[i,j]*x[j] <= t[i];
Least Absolute Deviation Regression via LP

Second of Two Methods

\[ \min \sum_i \left| b_i - \sum_j a_{ij}x_j \right| \]

Equivalent Linear Program:

\[ \min \sum_i (t_i^+ + t_i^-) \]

\[ t_i^+ - t_i^- = b_i - \sum_j a_{ij}x_j \quad i = 1, 2, \ldots, m \]

\[ t_i^+, t_i^- \geq 0 \]
AMPL Model

param m;
param n;

set I := {1..m};
set J := {1..n};

param A {I,J};
param b {I};

var x{J};
var t_plus{I} >= 0;
var t_minus{I} >= 0;

minimize sum_dev:
    sum {i in I} (t_plus[i] + t_minus[i]);

subject to t_def {i in I}:
    t_plus[i] - t_minus[i] = b[i] - sum {j in J} A[i,j]*x[j];

https://vanderbei.princeton.edu/307/regression/regr2.txt
https://vanderbei.princeton.edu/307/regression/namemniter
Thought experiment:

- \( \mu \) starts at \( \infty \).
- In reducing \( \mu \), there are \( n + m \) barriers.
- At each iteration, one barrier is passed—the others move about randomly.
- To get \( \mu \) to zero, we must on average pass half the barriers.
- Therefore, on average the algorithm should take \( \frac{m + n}{2} \) iterations.

Using 69 real-world problems from the Netlib suite...

Least Squares Regression:

\[
\begin{bmatrix}
\hat{\alpha} \\
\hat{\beta}
\end{bmatrix} = \begin{bmatrix}
-1.03561 \\
1.05152
\end{bmatrix} \implies T \approx 0.488(m + n)^{1.052}
\]

Least Absolute Deviation Regression:

\[
\begin{bmatrix}
\hat{\alpha} \\
\hat{\beta}
\end{bmatrix} = \begin{bmatrix}
-0.9508 \\
1.0491
\end{bmatrix} \implies T \approx 0.517(m + n)^{1.049}
\]
A log–log plot of $T$ vs. $m + n$ and the $L^1$ and $L^2$ regression lines.
Parametric Self−Dual Simplex Method

\[ \text{iters} = 0.4165(m + n)^{0.9759} \]

https://vanderbei.princeton.edu/307/python/psd_simplex_pivot.ipynb
The Parametric Self-Dual Simplex Method has been studied, with the number of pivots and iters given by:

\[ \text{iters} = 1.4880 \min(m, n)^{1.3434} \]

This formula can be found in the code at:

https://vanderbei.princeton.edu/307/python/psd_simplex_pivot.ipynb
\[
\text{iters} = 1.571 \min(m, n)^{1.3333}
\]