Lock-In and Unobserved Preferences in Server Operating System Adoption: A Case of Linux vs. Windows

Seung-Hyun Hong\textsuperscript{1}  Leonardo Rezende \textsuperscript{2}

\textsuperscript{1}University of Illinois at Urbana-Champaign

\textsuperscript{2}PUC-Rio and University of Illinois at Urbana-Champaign

April 2007
Motivation

- Well-known two explanations for Windows dominance:
Motivation

- Well-known two explanations for Windows dominance:
  - High switching costs and lock-in (so, potential inefficiency)
  - Superior quality (so, potential efficiency)
Motivation

- Well-known two explanations for Windows dominance:
  - High switching costs and lock-in (so, potential inefficiency)
  - Superior quality (so, potential efficiency)
- But not much econometric evidence on either explanation, possibly due to...
Motivation

- Well-known two explanations for Windows dominance:
  - High switching costs and lock-in (so, potential inefficiency)
  - Superior quality (so, potential efficiency)

- But not much econometric evidence on either explanation, possibly due to:
  - Lack of detailed micro data
  - Difficulty in identification between state dependence and unobserved heterogeneity
Our Approach and Findings

- Our Approach

Use detailed establishment-level data on operating systems.
Propose a new identification strategy (with weak assumptions).

Summary of Results
Without controlling for unobserved heterogeneity, we find a very strong positive correlation between the current choice and the previous choice in server operating systems. Once we account for unobserved preferences (potentially for product quality), we find little evidence for state dependence in both Windows adoption and Linux adoption.
Our Approach

- Use detailed establishment-level data on operating systems.
- Propose a new identification strategy (with weak assumptions).
Our Approach and Findings

- **Our Approach**
  - Use detailed establishment-level data on operating systems.
  - Propose a new identification strategy (with weak assumptions).

- **Summary of Results**
Our Approach and Findings

- **Our Approach**
  - Use detailed establishment-level data on operating systems.
  - Propose a new identification strategy (with weak assumptions).

- **Summary of Results**
  - **Without controlling for unobserved heterogeneity**, we find a very **strong positive correlation** between the current choice and the previous choice in server operating systems.
Our Approach and Findings

- **Our Approach**
  - Use detailed establishment-level data on operating systems.
  - Propose a new identification strategy (with weak assumptions).

- **Summary of Results**
  - Once we account for unobserved preferences (potentially for product quality), we find little evidence for state dependence in both Windows adoption and Linux adoption.
1. Data

- **Computer Intelligence Technology Database: 2000-2004**
  - Yearly survey of over 100,000 establishments in the U.S. (but not random sample).
  - Unbalanced panel data on detailed IT use of establishments.
1. Data

- **Computer Intelligence Technology Database: 2000-2004**
  - Yearly survey of over 100,000 establishments in the U.S. (but not random sample).
  - Unbalanced panel data on detailed IT use of establishments.
  - Segment-level (server, pc, and non-pc) information on operating systems (but not individual computer-level):
    - **windows**: Windows 95, 98, 2000, 2003, ME, NT, XP
    - **linux**: various versions of Linux (Red-Hat, SuSE, etc.)
    - **other**: Mac OS X, various proprietary Unix (Solaris, etc.)
1. Data

Table 1: Shares of Establishments with Each Operating System
(samples with any server operating system and with up-to-date info)

<table>
<thead>
<tr>
<th>Year</th>
<th>server.windows</th>
<th>server.linux</th>
<th>server.other</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>0.91</td>
<td>0.07</td>
<td>0.27</td>
</tr>
<tr>
<td>2001</td>
<td>0.88</td>
<td>0.10</td>
<td>0.26</td>
</tr>
<tr>
<td>2002</td>
<td>0.90</td>
<td>0.12</td>
<td>0.23</td>
</tr>
<tr>
<td>2003</td>
<td>0.91</td>
<td>0.14</td>
<td>0.22</td>
</tr>
<tr>
<td>2004</td>
<td>0.91</td>
<td>0.13</td>
<td>0.20</td>
</tr>
</tbody>
</table>
2. Basic Framework

- Reduced-form version of establishment $i$’s net benefit of choosing operating system $j$ at year $t$:

$$\pi_{ijt} = \gamma_{jt} + \sum_{k} \beta_{jk} y_{ik(t-1)} + \alpha_{j} x_{it} + \eta_{ij} + \epsilon_{ijt},$$

where $y_{ijt} \in \{\text{server.windows}_{it}, \text{server.linux}_{it}, \text{server.other}_{it}\}$. 
2. Basic Framework

- Reduced-form version of establishment $i$’s net benefit of choosing operating system $j$ at year $t$:

$$\pi_{ijt} = \gamma_{jt} + \sum_k \beta_{jk} y_{ik(t-1)} + \alpha_j x_{it} + \eta_{ij} + \epsilon_{ijt},$$

where $y_{ijt} \in \{\text{server.windows}_{it}, \text{server.linux}_{it}, \text{server.other}_{it}\}$.

- Assume that $\epsilon_{ijt}|H^t_i \sim N(0, \sigma_t)$, where $H^t_i = (H_{i0}, H_{i1}, \ldots, H_{it})$, and $H_{it} = (\text{server.linux}_{i(t-1)}, \text{server.windows}_{i(t-1)}, \text{server.other}_{i(t-1)}, x_{it})$.

$$\Rightarrow \Pr(y_{ijt} = 1|H^t_i) = \Phi \left( \frac{\gamma_{jt} + \sum_k \beta_{jk} y_{ik(t-1)} + \alpha_j x_{it} + \eta_{ij}}{\sigma_t} \right)$$
2. Basic Framework

- Reduced-form version of establishment $i$’s net benefit of choosing operating system $j$ at year $t$:

$$\pi_{ijt} = \gamma_{jt} + \sum_k \beta_{jk} y_{ik(t-1)} + \alpha_j x_{it} + \eta_{ij} + \epsilon_{ijt},$$

where $y_{ijt} \in \{\text{server.windows}_it, \text{server.linux}_it, \text{server.other}_it\}$.

- Assume that $\epsilon_{ijt}|H^t_i \sim \mathcal{N}(0, \sigma_t)$, where $H^t_i = (H_{i0}, H_{i1}, \ldots, H_{it})$, and $H_{it} = (\text{server.linux}_{i(t-1)}, \text{server.windows}_{i(t-1)}, \text{server.other}_{i(t-1)}, x_{it})$.

$$\Rightarrow \quad \Pr(y_{ijt} = 1|H^t_i) = \Phi \left( \frac{\gamma_{jt} + \sum_k \beta_{jk} y_{ik(t-1)} + \alpha_j x_{it} + \eta_{ij}}{\sigma_t} \right)$$

- Unobserved heterogeneity $\eta_{ij}$ generates spurious state dependence between $y_{ijt}$ and $y_{ij(t-1)}$. 
3. Potential Other Approaches to Account for \( \eta_{ij} \) in Discrete Choice Panel Data Models

- **Random effects:**
  - Impose known distributional assumptions for \( G(\eta_i) \) and integrate \( \eta_i \) out.
  - But cannot simply do it because \( \eta_i \) is correlated with \( y_{i(t-1)} \).
3. Potential Other Approaches to Account for $\eta_{ij}$ in Discrete Choice Panel Data Models

- **Random effects:**
  - Impose known distributional assumptions for $G(\eta_i)$ and integrate $\eta_i$ out.
  - But cannot simply do it because $\eta_i$ is correlated with $y_{i(t-1)}$.
  - Instead, consider the likelihood for establishment $i$ given by

  $$L_i = \int \prod_{t=1}^{T} \Pr(y_{it} | H_{i}^{t-1}, \eta_i) f(y_{i0} | \eta_i) dG(\eta_i).$$

  - But need to specify $f(y_{i0} | \eta_i)$ (i.e. initial condition problem).
3. Potential Other Approaches to Account for $\eta_{ij}$ in Discrete Choice Panel Data Models

- **Fixed effects:**
  - Do not impose any distributional assumption on $\eta_i$, but use a conditional logit approach to “difference out” $\eta_i$. 

Honoré and Kyriazidou (2000) extend it to the case with predetermined variables. But identification requires to observe $(y_{i1}, y_{i2}, y_{i3}, y_{i4}) = (0, 0, 1, 1)$ and $(y_{i1}, y_{i2}, y_{i3}, y_{i4}) = (0, 1, 0, 1)$.
3. Potential Other Approaches to Account for $\eta_{ij}$ in Discrete Choice Panel Data Models

- Fixed effects:
  - Do not impose any distributional assumption on $\eta_i$, but use a conditional logit approach to “difference out” $\eta_i$.
  - Honoré and Kyriazidou (2000) extend it to the case with predetermined variables. But identification requires to observe $(y_{i1}, y_{i2}, y_{i3}, y_{i4}) = (0, 0, 1, 1)$ and $(y_{i1}, y_{i2}, y_{i3}, y_{i4}) = (0, 1, 0, 1)$. 

---

S. Hyun Hong (UIUC) and Leo Rezende (PUC-Rio and UIUC)
4. Our Approach and Identification

- Follow the method developed by Arellano and Carrasco (JoE, 2003) (hence, AC method).

- **Key identifying assumption**: the quality of product $j$ is not fully observed by establishments, so that they revise their assessment of quality, based on previous experiences. Instead of $\eta_{ij}$, $E(\eta_{ij}|H_i^t)$ enters the net benefit function, s.t.

\[
\pi_{ijt} = \gamma_{jt} + \sum_k \beta_{jk} y_{ik(t-1)} + \alpha_j x_{it} + E(\eta_{ij}|H_i^t) + \epsilon_{ijt},
\]
4. Our Approach and Identification

- **Key idea of the AC method**
  - Rewrite the adoption probability conditional on the history:

\[
p_{ijt} \equiv \Pr(y_{ijt} = 1 | H^t_i) = \Phi \left( \frac{\gamma_{jt} + \sum_k \beta_{jk} y_{ik(t-1)} + \alpha_j x_{it} + E(\eta_{ij} | H^t_i)}{\sigma_t} \right)
\]
4. Our Approach and Identification

- **Key idea of the AC method**
  - Rewrite the adoption probability conditional on the history:
    \[
    p_{ijt} \equiv \Pr(y_{ijt} = 1 | H^t_i) = \Phi \left( \frac{\gamma_{jt} + \sum_k \beta_{jk} y_{ik(t-1)} + \alpha_j x_{it} + E(\eta_{ij} | H^t_i)}{\sigma_t} \right)
    \]
  - Invert the equation above:
    \[
    E(\eta_{ij} | H^t_i) = \Phi^{-1}(p_{ijt}) \sigma_t - \gamma_{jt} - \sum_k \beta_{jk} y_{ik(t-1)} - \alpha_j x_{it}.
    \]
4. Our Approach and Identification

- **Key idea of the AC method**
  - Rewrite the adoption probability conditional on the history:

\[
p_{ijt} \equiv \Pr(y_{ijt} = 1|H_i^t) = \Phi \left( \frac{\gamma_{jt} + \sum_k \beta_{jk} y_{ik(t-1)} + \alpha_j x_{it} + E(\eta_{ij}|H_i^t)}{\sigma_t} \right)
\]

- Invert the equation above:

\[
E(\eta_{ij}|H_i^t) = \Phi^{-1}(p_{ijt})\sigma_t - \gamma_{jt} - \sum_k \beta_{jk} y_{ik(t-1)} - \alpha_j x_{it}.
\]

- Note that \(E[E(\eta_{ij}|H_i^{t+1})|H_i^t] = E(\eta_{ij}|H_i^t)\).
4. Our Approach and Identification

- Key idea of the AC method
  - Rewrite the adoption probability conditional on the history:
    \[
    p_{ijt} \equiv \Pr(y_{ijt} = 1|H_i^t) = \Phi\left(\frac{\gamma_{jt} + \sum_k \beta_{jk} y_{ik(t-1)} + \alpha_j x_{it} + E(\eta_{ij}|H_i^t)}{\sigma_t}\right)
    \]
  - Invert the equation above:
    \[
    E(\eta_{ij}|H_i^t) = \Phi^{-1}(p_{ijt})\sigma_t - \gamma_{jt} - \sum_k \beta_{jk} y_{ik(t-1)} - \alpha_j x_{it}.
    \]
  - Note that \(E[E(\eta_{ij}|H_i^{t+1})|H_i^t] = E(\eta_{ij}|H_i^t)\).
  - Moment condition is then given by \(E(\nu_{ijt}|H_i^t) = 0\), where
    \[
    \nu_{ijt} \equiv E(\eta_{ij}|H_i^{t+1}) - E(\eta_{ij}|H_i^t).
    \]
    \[
    \Rightarrow \quad E(H_i^t \nu_{ijt}) = 0
    \]
4. Our Approach and Identification

- **Two-step Estimation** (for a discrete vector of $H^t_i$):
  1. Estimate $p_{ijt}(H^t_i) \equiv \Pr(y_{ijt} = 1|H^t_i)$ nonparametrically.
  2. Replace $p_{ijt}$ with $\hat{p}_{ijt}$, and use a GMM based on

$$E \left\{ d_{ih}^{t-1} \left[ \sigma_t \Phi^{-1}(p_{ijt}) - \sigma_{t-1} \Phi^{-1}(p_{ij(t-1)}) - \Delta \gamma_t - \sum_k \beta_{jk} \Delta y_{i_k(t-1)} - \alpha_j \Delta x_{it} \right] \right\} = 0,$$

where $d_{ih}^t = \mathbb{I}\{H^t_i = \phi^t_h\}$ and $\phi^t_h$ denote a value that $H^t_i$ can take.
4. Our Approach and Identification

- **Two-step Estimation** (for a discrete vector of $H_i^t$):
  1. Estimate $p_{ijt}(H_i^t) \equiv \Pr(y_{ijt} = 1|H_i^t)$ nonparametrically.
  2. Replace $p_{ijt}$ with $\hat{p}_{ijt}$, and use a GMM based on

$$E \left\{ d_{ih}^{t-1} \left[ \sigma_t \Phi^{-1}(p_{ijt}) - \sigma_{t-1} \Phi^{-1}(p_{ij(t-1)}) - \Delta \gamma_t - \sum_k \beta_{jk} \Delta y_{ik(t-1)} - \alpha_j \Delta x_{it} \right] \right\} = 0,$$

where $d_{ih}^t = \mathbb{I}\{H_i^t = \phi_h^t\}$ and $\phi_h^t$ denote a value that $H_i^t$ can take.

- **Note on our implementation**
  - Estimate the model using relatively homogeneous subsamples, instead of including many covariates.
  - Use four sets of complete panel data:
5. Results

Table 2: **Windows Adoption** using 2000-2003 complete panel

<table>
<thead>
<tr>
<th>Variable</th>
<th>Probit</th>
<th>AC Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate (1)</td>
<td>Estimate (2)</td>
</tr>
<tr>
<td>A. NAICS 31-32:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>server.linux_{t-1}</td>
<td>-0.5008</td>
<td>-0.5263</td>
</tr>
<tr>
<td>server.windows_{t-1}</td>
<td>2.185</td>
<td>2.092</td>
</tr>
<tr>
<td>server.other_{t-1}</td>
<td>-0.4742</td>
<td>-0.5019</td>
</tr>
<tr>
<td>pc.linux_{t-1}</td>
<td>0.2031</td>
<td>0.2373</td>
</tr>
<tr>
<td>pc.windows_{t-1}</td>
<td>0.5063</td>
<td>0.6117</td>
</tr>
<tr>
<td>pc.other_{t-1}</td>
<td>0.1175</td>
<td>0.0831</td>
</tr>
<tr>
<td>non-pc.windows_{t-1}</td>
<td>0.0394</td>
<td>-0.2629</td>
</tr>
<tr>
<td>non-pc.other_{t-1}</td>
<td>-0.0244</td>
<td>-0.0958</td>
</tr>
<tr>
<td>γ2001</td>
<td>-0.4670</td>
<td>-0.5999</td>
</tr>
<tr>
<td>γ2002</td>
<td>-0.5218</td>
<td>-0.6655</td>
</tr>
<tr>
<td>γ2003</td>
<td>-0.2289</td>
<td>-0.3337</td>
</tr>
<tr>
<td>Additional control</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
5. Results

Table 2: **Windows Adoption** using 2000-2003 complete panel

<table>
<thead>
<tr>
<th>Variable</th>
<th>Probit</th>
<th>AC Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate (1) S.E.</td>
<td>Estimate (2) S.E.</td>
</tr>
<tr>
<td><strong>B. NAICS 33: Manufacturing</strong></td>
<td>2.162 0.099</td>
<td>2.101 0.103</td>
</tr>
<tr>
<td><strong>C. NAICS 4: Wholesale, Retail, and Transportation</strong></td>
<td>2.334 0.146</td>
<td>2.293 0.152</td>
</tr>
<tr>
<td><strong>D. NAICS 51: Information</strong></td>
<td>2.249 0.102</td>
<td>2.198 0.104</td>
</tr>
<tr>
<td><strong>E. NAICS 52-53: Finance and Real Estate</strong></td>
<td>2.161 0.191</td>
<td>2.118 0.205</td>
</tr>
<tr>
<td><strong>F. NAICS 54-56: Professional and Technical Services</strong></td>
<td>2.141 0.128</td>
<td>2.075 0.135</td>
</tr>
<tr>
<td><strong>G. NAICS 61: Educational Services</strong></td>
<td>2.120 0.077</td>
<td>2.071 0.080</td>
</tr>
<tr>
<td><strong>H. NAICS 62: Health Care</strong></td>
<td>2.364 0.155</td>
<td>2.360 0.164</td>
</tr>
<tr>
<td><strong>I. NAICS 9: Public Administration</strong></td>
<td>2.276 0.090</td>
<td>2.232 0.093</td>
</tr>
<tr>
<td>Additional control</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Table 3: Linux Adoption using 2000-2003 complete panel

<table>
<thead>
<tr>
<th>Variable</th>
<th>Probit Estimate (1)</th>
<th>S.E. (1)</th>
<th>Estimate (2)</th>
<th>S.E. (2)</th>
<th>AC Method Estimate (3)</th>
<th>S.E. (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>server.linux_{t-1}</td>
<td>A. NAICS 31-32: Manufacturing 2.370 0.131</td>
<td>2.259 0.138</td>
<td>0.233 5.555</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B. NAICS 33: Manufacturing 2.307 0.082</td>
<td>2.190 0.086</td>
<td>0.792 2.655</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C. NAICS 4: Wholesale, Retail, and Transportation 2.185 0.126</td>
<td>2.078 0.132</td>
<td>0.772 3.720</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D. NAICS 51: Information 2.426 0.103</td>
<td>2.238 0.109</td>
<td>1.547 6.749</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E. NAICS 52-53: Finance and Real Estate 2.598 0.178</td>
<td>2.597 0.189</td>
<td>-1.432 8.833</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F. NAICS 54-56: Professional and Technical Services 2.366 0.097</td>
<td>2.219 0.101</td>
<td>-0.180 2.697</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G. NAICS 61: Educational Services 2.218 0.058</td>
<td>2.123 0.060</td>
<td>-0.058 1.370</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H. NAICS 62: Health Care 2.302 0.116</td>
<td>2.248 0.119</td>
<td>0.387 0.610</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I. NAICS 9: Public Administration 2.339 0.075</td>
<td>2.264 0.078</td>
<td>1.251 2.201</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional control</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6. Conclusion and Future Research

- **Main Findings**
  Once we account for unobserved preferences (potentially for product quality), we find little evidence for state dependence in both Windows adoption and Linux adoption.
6. Conclusion and Future Research

- **Main Findings**
  Once we account for unobserved preferences (potentially for product quality), we find little evidence for state dependence in both Windows adoption and Linux adoption.

- **One Implication of Findings**
  Likely that establishments in our data continued to use Windows (or Linux) for their server operating systems not because of lock-in, but because of better quality, the assessment of which can change depending on their experiences (thus, consistent with both Windows dominance and increasing popularity in Linux).
6. Conclusion and Future Research

- Alternative Interpretation of Findings
  
  The results also suggest that true state dependence and lock-in may not be measured properly by simply estimating the coefficient for $y_{i,j}(t-1)$.
6. Conclusion and Future Research

- Alternative Interpretation of Findings
  1. The results also suggest that true state dependence and lock-in may not be measured properly by simply estimating the coefficient for $y_{ij}(t-1)$.
  2. Moreover, unobserved preferences and quality may reflect potential network effects within the same segment, which calls for further modeling.