Case-control studies, regression and survival analysis

Tyler Moore

CSE 7338
Computer Science & Engineering Department, SMU, Dallas, TX

Lectures 6–7

Outline

Case-control studies
Regression and survival analysis

Guide to exploring data

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Exploration</th>
<th>Statistics</th>
<th>RByEx</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 numerical variable</td>
<td>![bar chart] one way t-test</td>
<td>Wilcoxon test</td>
<td>6.1</td>
</tr>
<tr>
<td>1 categorical variable</td>
<td>![bar chart]</td>
<td>–</td>
<td>3.1</td>
</tr>
<tr>
<td>2 categories</td>
<td>prop.test</td>
<td>6.2</td>
<td></td>
</tr>
<tr>
<td>1 categorical, 1 numerical</td>
<td>![bar chart]</td>
<td>anova, Permutation</td>
<td>10</td>
</tr>
<tr>
<td>2 categories</td>
<td>2-way t, Wilcoxon, Perm.</td>
<td>6.4</td>
<td></td>
</tr>
</tbody>
</table>

Guide to analyzing data

- After visual exploration and any descriptive statistics, you may want to investigate relationships between variables more closely
- In particular, you can investigate how one or more explanatory (aka independent) variables influences response (aka dependent) variables

<table>
<thead>
<tr>
<th>Statistical Method</th>
<th>Response Variable</th>
<th>Explanatory Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds ratio</td>
<td>Binary (case/control)</td>
<td>Categorical variables (1 at a time)</td>
</tr>
<tr>
<td>Linear regression</td>
<td>Numerical</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Binary</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
<tr>
<td>Survival analysis</td>
<td>Time to event</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
</tbody>
</table>
Case-control studies

Case-control studies and cybercrime

In a perfect world, we could measure security using randomized controlled experiments similar to medicine.

But most security data is observational — we can’t select subjects and apply treatments to a subset.

Instead, we can observe that some targets are victimized, while other vulnerable targets are not.

Crucially, this observation happens after the fact (if at all).

Case-control study method is ideal for identifying risk factors when all you have is observational data.

Case-control study design

Case-control study design: smoking and lung cancer
### The odds ratio

<table>
<thead>
<tr>
<th></th>
<th>Case (afflicted)</th>
<th>Control (not afflicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>$p_{11}$</td>
<td>$p_{10}$</td>
</tr>
<tr>
<td>Not exposed</td>
<td>$p_{01}$</td>
<td>$p_{00}$</td>
</tr>
</tbody>
</table>

Odds ratio: $\frac{p_{11} \cdot p_{00}}{p_{10} \cdot p_{01}}$

---

### A word on odds ratios

- **Defining odds**
  - Suppose we have an event with two possible outcomes: success ($S$) and failure ($\bar{S}$).
  - The probability of each occurring happens with $p_s$ and $p_{\bar{S}} = 1 - p_s$.
  - The odds of the event are given by $\frac{p_s}{1 - p_s}$.

- **Defining odds ratios**
  - Suppose now there are two events $A$ and $B$, both of which can occur (with probabilities $p_A$ and $p_B$).
  
  
  \[
  \text{odd's ratio} = \frac{\text{odds}(A)}{\text{odds}(B)} = \frac{\frac{p_A}{1 - p_A}}{\frac{p_B}{1 - p_B}} = \frac{p_A \times (1 - p_B)}{(1 - p_A) \times p_B}
  \]

---

### Odds ratio example

- Adapted from http://www.ats.ucla.edu/stat/stata/faq/oratio.htm

- Suppose that 7 of 10 male applicants to engineering school are admitted, compared to 4 of 40 female applicants:
  - $p_{\text{male acc.}} = 0.7$, $p_{\text{male rej.}} = 1 - 0.7 = 0.3$
  - $p_{\text{female acc.}} = 0.6$, $p_{\text{female rej.}} = 1 - 0.4 = 0.6$
  - $\text{OR} = \frac{0.7}{0.3} = 2.33$
  - $\text{OR} (\text{female acc.}) = \frac{0.6}{0.4} = 1.5$

- Hence, we can say that the odds of a male applicant being admitted are 3.5 times stronger than for a female applicant.

---

### Case-control study: spear phishing and academic specialty

- Population: Malware spam recipients
- Case: Targeted email
- Control: Un-targeted email
- Exposed: Academic Subject
- Not Exposed: Other Subjects
- Exposed: Academic Subject
- Not Exposed: Other Subjects
Odds ratios for academic subjects in spear phishing study

<table>
<thead>
<tr>
<th>Subject Code</th>
<th>Subject</th>
<th>Odds Ratio</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Medicine &amp; Dentistry</td>
<td>0.15</td>
<td>(0.03 - 0.87)</td>
</tr>
<tr>
<td>B</td>
<td>Social Sciences</td>
<td>0.14</td>
<td>(0.84 - 2.80)</td>
</tr>
<tr>
<td>C</td>
<td>Vocational &amp; Technical Sciences</td>
<td>0.49</td>
<td>(0.19 - 1.30)</td>
</tr>
<tr>
<td>D</td>
<td>Business &amp; Administrative Studies</td>
<td>0.77</td>
<td>(0.17 - 3.49)</td>
</tr>
<tr>
<td>E</td>
<td>Arts &amp; Humanities</td>
<td>2.68</td>
<td>(1.19 - 5.72)</td>
</tr>
<tr>
<td>F</td>
<td>Computer &amp; Information Sciences</td>
<td>1.46</td>
<td>(0.60 - 3.56)</td>
</tr>
<tr>
<td>G</td>
<td>Language &amp; Linguistics</td>
<td>0.15</td>
<td>(0.02 - 1.40)</td>
</tr>
<tr>
<td>H</td>
<td>Health Sciences</td>
<td>0.50</td>
<td>(0.30 - 0.82)</td>
</tr>
<tr>
<td>I</td>
<td>Social Sciences</td>
<td>2.00</td>
<td>(0.97 - 4.14)</td>
</tr>
<tr>
<td>J</td>
<td>Business Administration</td>
<td>2.80</td>
<td>(1.33 - 5.92)</td>
</tr>
</tbody>
</table>

Notes

What do illicit online pharmacies have to do with phishing?

- Both make use of a similar criminal supply chain
  - **Traffic**: hijack web search results (or send email spam)
  - **Host**: compromise a high-ranking server to redirect to pharmacy
  - **Hook**: affiliate programs let criminals set up website front-ends to sell drugs
  - **Monetize**: sell drugs ordered by consumers
  - **Cash out**: no need to hire mules, just take credit cards!

For more: [http://lyle.smu.edu/~tylerm/usenix11.pdf](http://lyle.smu.edu/~tylerm/usenix11.pdf)

Case-control study: search-redirection attacks

**Population**: pharma search results

**Case**: Search-redirection attack

**Control**: No redirection

<table>
<thead>
<tr>
<th>Exposed: EDU TLDs</th>
<th>Not Exposed: Other TLDs</th>
<th>Past</th>
<th>Exposed: EDU TLDs</th>
<th>Not Exposed: Other TLDs</th>
<th>Present</th>
</tr>
</thead>
</table>


Data format:

<table>
<thead>
<tr>
<th>Date</th>
<th>Search Engine</th>
<th>Search Term</th>
<th>Pos.</th>
<th>URL Domain</th>
<th>Redirects?</th>
<th>TLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011-11-03</td>
<td>Google</td>
<td>20 mg ambien overdose</td>
<td>1</td>
<td><a href="http://products.sanofi.us/ambien/ambien.pdf">http://products.sanofi.us/ambien/ambien.pdf</a></td>
<td>false</td>
<td>.EDU</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Google</td>
<td>20 mg ambien overdose</td>
<td>2</td>
<td><a href="http://swift.sonoma.edu/education/newton/newtonsLaws/?20-mg-ambien-overdose">http://swift.sonoma.edu/education/newton/newtonsLaws/?20-mg-ambien-overdose</a></td>
<td>false</td>
<td>.EDU</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Google</td>
<td>20 mg ambien overdose</td>
<td>3</td>
<td><a href="http://ambienoverdose.org/about-2/">http://ambienoverdose.org/about-2/</a></td>
<td>false</td>
<td>.ORG</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Google</td>
<td>20 mg ambien overdose</td>
<td>4</td>
<td><a href="http://answers.yahoo.com/question/index?qid=20090712025803AA10g8Z">http://answers.yahoo.com/question/index?qid=20090712025803AA10g8Z</a></td>
<td>false</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Google</td>
<td>20 mg ambien overdose</td>
<td>6</td>
<td><a href="http://blocsonic.com/blog">http://blocsonic.com/blog</a></td>
<td>false</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Google</td>
<td>20 mg ambien overdose</td>
<td>8</td>
<td><a href="http://nemo.mwd.hartford.edu/mwd08/images/?20-mg-ambien-overdose">http://nemo.mwd.hartford.edu/mwd08/images/?20-mg-ambien-overdose</a></td>
<td>true</td>
<td>.EDU</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Google</td>
<td>20 mg ambien overdose</td>
<td>9</td>
<td><a href="http://www.formspring.me/AmbienCheapOn">http://www.formspring.me/AmbienCheapOn</a></td>
<td>false</td>
<td>.OTHER</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Google</td>
<td>20 mg ambien overdose</td>
<td>10</td>
<td><a href="http://www.drugs.com/pro/zolpidem.html">http://www.drugs.com/pro/zolpidem.html</a></td>
<td>false</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Google</td>
<td>20 mg ambien overdose</td>
<td>11</td>
<td><a href="http://engineer.tamuk.edu/departments/ieen/images/ambien.html">http://engineer.tamuk.edu/departments/ieen/images/ambien.html</a></td>
<td>false</td>
<td>.EDU</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>1</td>
<td><a href="http://answers.yahoo.com/question/index?qid=20111024222432AARFvPB">http://answers.yahoo.com/question/index?qid=20111024222432AARFvPB</a></td>
<td>false</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>2</td>
<td><a href="http://www.healthcentral.com/sleep-disorders/h/20-mg-ambien-overdose.html">http://www.healthcentral.com/sleep-disorders/h/20-mg-ambien-overdose.html</a></td>
<td>false</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>3</td>
<td><a href="http://ambien20mg.com/">http://ambien20mg.com/</a></td>
<td>false</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>4</td>
<td><a href="http://www.chacha.com/question/will-20-mg-of-ambien-cr-get-you-high">http://www.chacha.com/question/will-20-mg-of-ambien-cr-get-you-high</a></td>
<td>true</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>5</td>
<td><a href="http://www.rxlist.com/ambien-drug.htm">http://www.rxlist.com/ambien-drug.htm</a></td>
<td>true</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>6</td>
<td><a href="http://answers.yahoo.com/question/index?qid=20111122222232FQRvPB">http://answers.yahoo.com/question/index?qid=20111122222232FQRvPB</a></td>
<td>false</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>7</td>
<td><a href="http://en.wikipedia.org/wiki/Zolpidem">http://en.wikipedia.org/wiki/Zolpidem</a></td>
<td>false</td>
<td>.ORG</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>8</td>
<td><a href="http://www.thefullwiki.org/Sertraline">http://www.thefullwiki.org/Sertraline</a></td>
<td>false</td>
<td>.ORG</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>9</td>
<td><a href="http://www.rxlist.com/edluar-drug.htm">http://www.rxlist.com/edluar-drug.htm</a></td>
<td>true</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>10</td>
<td><a href="http://www.formspring.me/ambienpill">http://www.formspring.me/ambienpill</a></td>
<td>false</td>
<td>.OTHER</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>11</td>
<td><a href="http://ambiendosage.net/">http://ambiendosage.net/</a></td>
<td>false</td>
<td>.COM</td>
</tr>
<tr>
<td>2011-11-03</td>
<td>Bing</td>
<td>20 mg ambien overdose</td>
<td>12</td>
<td><a href="http://ambiendosage.net/">http://ambiendosage.net/</a></td>
<td>false</td>
<td>.NET</td>
</tr>
</tbody>
</table>
Case-control studies
Odds ratios for case-control study

```r
> library(epitools)
> pr.tldodds<-oddsratio(pr$tld,pr$redirects,verbose=T)
> pr.tldodds$measure
odds ratio with 95% C.I.
Predictor estimate lower upper
.COM 1.000000 NA NA
.EDU 5.839096 5.563529 6.159197
.GOV 0.431855 0.306481 0.5882504
.NET 0.594602 0.556859 0.6342355
.ORG 2.881149 2.797383 2.9674615
other 1.343711 1.280920 1.4090669
```

Odds ratios for case-control study

```r
> pr.tldodds$p.value
ten-sided
Predictor midp.exact
.COM NA
.EDU 0.000000000000000
.GOV 0.0000000000000000
.NET 0.0000000000000000
.ORG 0.0000000000000000
other 0.0000000000000000

ten-sided
Predictor fisher.exact
.COM NA
.EDU 0.00000000000000000000000000000000000000000000000000000000000000000000
.GOV 0.0000011173591553812466570181077239286036342904420046260
.NET 0.00000000000000000000000000000000000000000000000000000000000003109266
.ORG 0.00000000000000000000000000000000000000000000000000000000000000000000
other 0.00000000000000000000000000000000000000000000000000000000000000000000

ten-sided
Predictor chi.square
.COM NA
.EDU 0.000000000000000000000000000000000000000000000000000000000000
.GOV 0.000000150899123313924415716095442548116967174095815997734
.NET 0.0000000000000000000000000000000000000000000000000000000000017562
.ORG 0.000000000000000000000000000000000000000000000000000000000000
other 0.000000000000000000000000000000000000000000000000000000000000039069706125273474297635
```

How to interpret the odds ratios?

```r
> library(epitools)
> pr.tldodds<-oddsratio(pr$tld,pr$redirects,verbose=T)
> pr.tldodds$measure
odds ratio with 95% C.I.
Predictor estimate lower upper
.COM 1.000000 NA NA
.EDU 5.839096 5.563529 6.159197
.GOV 0.431855 0.306481 0.5882504
.NET 0.594602 0.556859 0.6342355
.ORG 2.881149 2.797383 2.9674615
other 1.343711 1.280920 1.4090669
```

Regression and survival analysis
Guide to analyzing data

- After visual exploration and any descriptive statistics, you may want to investigate relationships between variables more closely.
- In particular, you can investigate how one or more explanatory (aka independent) variables influences response (aka dependent) variables.

<table>
<thead>
<tr>
<th>Statistical Method</th>
<th>Response Variable</th>
<th>Explanatory Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds ratio</td>
<td>Binary (case/control)</td>
<td>Categorical variables (1 at a time)</td>
</tr>
<tr>
<td>Linear regression</td>
<td>Numerical</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Binary</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
<tr>
<td>Survival analysis</td>
<td>Time to event</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
</tbody>
</table>
Suppose the values of a numerical variable $Y$ depend on the values of another variable $X$.

$$Y = c_0 + c_1 X + \epsilon$$

If that dependence is linear then we can use linear regression to estimate the best-fit values of the constants $c_0$ and $c_1$ that minimize the error values for all the values $y_i \in Y$.

For more info see “R by Example” Ch. 7.1–7.3
Regression and survival analysis

### Dataset for linear regression example

- Suppose you hypothesize that the popularity of a CMS platform influences the number of exploits made available.
- We can use linear regression to test for such a relationship.

<table>
<thead>
<tr>
<th>generatorType</th>
<th>CMSmarketShare</th>
<th>numExploits</th>
</tr>
</thead>
<tbody>
<tr>
<td>blogger</td>
<td>3.5</td>
<td>10</td>
</tr>
<tr>
<td>concrete5</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>contao</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>datalife engine</td>
<td>1.5</td>
<td>3</td>
</tr>
<tr>
<td>discuz</td>
<td>1.3</td>
<td>8</td>
</tr>
<tr>
<td>drupal</td>
<td>7.2</td>
<td>12</td>
</tr>
</tbody>
</table>

- Data: [http://lyle.smu.edu/~tylerm/courses/econsec/data/eims.csv](http://lyle.smu.edu/~tylerm/courses/econsec/data/eims.csv)

### Scatter plot

```r
plot(y=marExp$numExploits,x=marExp$numServers)
```

### Scatter plot (log-transformed)

```r
plot(y=marExp$numExploits,x=marExp$numServers, log = 'xy')
```
Linear regression

```r
> reg <- lm(lgExploits ~ lgServers, data = marExp2)
> summary(reg)
```

**Call:**

```r
lm(formula = lgExploits ~ lgServers, data = marExp2)
```

**Residuals:**

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.9692</td>
<td>-1.0655</td>
<td>-0.6013</td>
<td>0.5555</td>
<td>5.4554</td>
</tr>
</tbody>
</table>

**Coefficients:**

|        | Estimate | Std. Error | t value | Pr(>|t|) |
|--------|----------|------------|---------|---------|
| (Intercept) | -9.4067  | 3.1924     | -2.947  | 0.006280 ** |
| lgServers   | 0.6304   | 0.1681     | 3.750   | 0.000784 *** |

**Signif. codes:*** 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

**Residual standard error:** 2.091 on 29 degrees of freedom

**Multiple R-squared:** 0.3266, Adjusted R-squared: 0.3034

**F-statistic:** 14.07 on 1 and 29 DF, **p-value:** 0.0007842

---

Best-fit linear regression

![Best-fit linear regression](image)

```r
plot(y = marExp2$lgExploits, x = marExp2$lgServers,
     xlab = "lg(# Servers per CMS)",
     ylab = "lg(# exploits available per CMS)"
)

text(x = marExp2$lgServers, y = marExp2$lgExploits - 0.3,
     lab = marExp2$generatorType)
abline(reg$coef)
```

Guide to analyzing data

- After visual exploration and any descriptive statistics, you may want to investigate relationships between variables more closely.
- In particular, you can investigate how one or more explanatory (aka independent) variables influence response (aka dependent) variables.

<table>
<thead>
<tr>
<th>Statistical Method</th>
<th>Response Variable</th>
<th>Explanatory Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds ratio</td>
<td>Binary (case/control)</td>
<td>Categorical variables (1 at a time)</td>
</tr>
<tr>
<td>Linear regression</td>
<td>Numerical</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Binary</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
<tr>
<td>Survival analysis</td>
<td>Time to event</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
</tbody>
</table>

Logistic regression

- Suppose we wanted to examine how a numerical variable (e.g., position in search results) affects a binary response variable (e.g., whether the URL redirects or not).
- We can’t use the odds ratios from case-control studies because that requires a categorical variable.
- Suppose that we’d also like to examine how both position in search results and TLD affect whether a URL redirects.
- For these cases, we need a logistic regression

  \[
  \log \frac{p}{1-p} = c_0 + c_1 x_1 + c_2 x_2 + \epsilon
  \]

  So for the example above considering position and TLD:

  \[
  \log \frac{p_{redirect}}{1-p_{redirect}} = c_0 + c_1 \text{Position}_1 + c_2 \text{TLD}_2 + \epsilon
  \]
Illicit online pharmacies

What do illicit online pharmacies have to do with phishing?
Both make use of a similar criminal supply chain

Traffic: hijack web search results (or send email spam)
Host: compromise a high-ranking server to redirect to pharmacy
Hook: affiliate programs let criminals set up website front-ends to sell drugs
Monetize: sell drugs ordered by consumers
Cash out: no need to hire mules, just take credit cards!

For more: http://lyle.msu.edu/~tylerm/usenix11.pdf

Logistic regression in action


```
> pr.logit <- glm(redirects ~ tld, data=pr, family=binomial(link = "logit"))
> summary(pr.logit)
```

```
Call:  
glm(formula = redirects ~ tld, family = binomial(link = "logit"),
     data = pr)

Deviance Residuals: 
-1.1476 -0.5442 -0.5442 -0.5442 2.3438

Coefficients: 
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.835165 0.008626 -212.75 < 0.0000000000000002 ***
tld.EDU 1.764595 0.027159 64.97 < 0.0000000000000002 ***
tld.GOV -0.845142 0.165381 -5.11 0.000000322 ***
tld.NET -0.519996 0.033165 -15.68 < 0.0000000000000002 ***
tld.ORG 1.058195 0.015079 70.18 < 0.0000000000000002 ***
tldother 0.295390 0.024323 12.14 < 0.0000000000000002 ***
```

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 165287 on 175794 degrees of freedom
Residual deviance: 156797 on 175789 degrees of freedom
AIC: 156809

Number of Fisher Scoring iterations: 4

```
> NagelkerkeR2(pr.logit)
$R2
[1] 0.07736148
```

Logistic regression in action (ctd.)

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 165287 on 175794 degrees of freedom
Residual deviance: 156797 on 175789 degrees of freedom
AIC: 156809

Number of Fisher Scoring iterations: 4

```
> NagelkerkeR2(pr.logit)
$R2
[1] 0.07736148
```

Obtaining the odds ratios

Recall the logistic regression equation

\[
\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon
\]

Exponentiate coefficients to get interpretable odds ratios

```
> exp(cbind(OR = coef(pr.logit), confint(pr.logit)))
Waiting for profiling to be done...

OR 2.5% 97.5%
(Intercept) 0.1595871 0.1569062 0.1623025
-1.8351654 0.0086262 -212.75 < 0.0000000000000002 ***
tld.EDU 5.8392049 5.5336431 6.1584001 1.7645946 0.0271590 64.97 < 0.0000000000000002 ***
tld.GOV 0.4294964 0.3053796 0.5858515 -0.8451420 0.1653811 -5.11 0.000000322 ***
tld.NET 0.5945230 0.5568118 0.6341472 -0.5199960 0.0331651 -15.68 < 0.0000000000000002 ***
tld.ORG 2.8811645 2.7973226 2.9675454 1.0581945 0.0150790 70.18 < 0.0000000000000002 ***
tldother 1.3436501 1.2808599 1.4090019 0.2953900 0.0243229 12.14 < 0.0000000000000002 ***
```
> pr.logit2 <- glm(redirects ~ tld + resultPosition, data=pr, family=binomial(link = 'logit'))
> summary(pr.logit2)

Call:  
  glm(formula = redirects ~ tld + resultPosition, family = binomial(link = "logit"), 
  data = pr)

Deviance Residuals:  
   Min 1Q Median 3Q Max  
-1.2680 -0.5968 -0.5355 -0.4757 2.4268

Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -2.14012   0.01497  -142.920 < 0.0000000000000002 ***  
tld.EDU       1.77355   0.02726   65.072 < 0.0000000000000002 ***  
tld.GOV      -0.84060   0.16587   -5.068  0.000000402 ***  
tld.NET      -0.53121   0.03321  -15.993 < 0.0000000000000002 ***  
tld.ORG       1.05185   0.01512   69.587 < 0.0000000000000002 ***  
tldother      0.30033   0.02437   12.322 < 0.0000000000000002 ***  
resultPosition 0.01803   0.00070   25.762 < 0.0000000000000002 ***  

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 165287 on 175794 degrees of freedom
Residual deviance: 156129 on 175788 degrees of freedom
AIC: 156143
Number of Fisher Scoring iterations: 5

> exp(coef(pr.logit2))

Waiting for profiling to be done...

NagelkerkeR2(pr.logit2) #compute pseudo R^2 on logistic regression

OR  2.5 % 97.5 %
(Intercept) 0.1176407 0.1142316 0.1211375
 tld.EDU 5.8917404 5.5885212 6.2149893
 tld.GOV 0.4314497 0.3067092 0.5886711
 tld.NET 0.5878939 0.5505610 0.6271261
 tld.ORG 2.8602465 2.7793345 2.9498947
 tldother 1.3503082 1.2870831 1.4161226
resultPosition 1.0181977 1.0168021 1.0195962
searchEnginegoogle 2.2957964 2.2348606 2.3585081

> NagelkerkeR2(pr.logit3) #compute pseudo R^2 on logistic regression

$N
[1] 175795

$R2
[1] 0.08329341

> pr.logit3 <- glm(redirects ~ tld + resultPosition + searchEngine, data=pr, family=binomial(link = 'logit'))
> summary(pr.logit3)

Call:  
  glm(formula = redirects ~ tld + resultPosition + searchEngine, family = binomial(link = "logit"), 
  data = pr)

Deviance Residuals:  
   Min 1Q Median 3Q Max  
-1.3270 -0.6539 -0.4812 -0.3956 2.5988

Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -2.5813149  0.0172986 -149.221 < 0.0000000000000002 ***  
tld.EDU       1.5001887  0.0277776  54.007 < 0.0000000000000002 ***  
tld.GOV      -0.8537354  0.1666852  -5.122  0.000000303 ***  
tld.NET      -0.4290936  0.0335099  -12.805 < 0.0000000000000002 ***  
tld.ORG       0.9098682  0.0154358   58.945 < 0.0000000000000002 ***  
tldother     0.3191095  0.0246746   12.933 < 0.0000000000000002 ***  
resultPosition 0.0185985  0.0007081   26.265 < 0.0000000000000002 ***  
searchEnginegoogle 0.8310798  0.0137375   60.497 < 0.0000000000000002 ***

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 165287 on 175794 degrees of freedom
Residual deviance: 152322 on 175787 degrees of freedom
AIC: 152338
Number of Fisher Scoring iterations: 5

> exp(coef(pr.logit3))

Waiting for profiling to be done...

OR  2.5 % 97.5 %
(Intercept) 0.07567444 0.0731465 0.07827858
 tld.EDU 4.48254365 4.2498184 4.73303972
 tld.GOV 0.45823195 0.3229690 0.60014422
 tld.NET 0.65108987 0.6049052 0.69496781
 tld.ORG 2.83939875 2.7499342 2.93058788
 tldother 1.37590197 1.3107099 1.44382462
resultPosition 0.319095 0.246746 1.29334000
searchEnginegoogle 2.93778465 2.8360600 3.03808100

> NagelkerkeR2(pr.logit3) #compute pseudo R^2 on logistic regression

$N
[1] 175795

$R2
[1] 0.1166546
Guide to analyzing data

- After visual exploration and any descriptive statistics, you may want to investigate relationships between variables more closely.
- In particular, you can investigate how one or more explanatory (aka independent) variables influence response (aka dependent) variables.

<table>
<thead>
<tr>
<th>Statistical Method</th>
<th>Response Variable</th>
<th>Explanatory Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds ratio</td>
<td>Binary (case/control)</td>
<td>Categorical variables (1 at a time)</td>
</tr>
<tr>
<td>Linear regression</td>
<td>Numerical</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Binary</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
<tr>
<td>Survival analysis</td>
<td>Time to event</td>
<td>One or more variables (numerical or categorical)</td>
</tr>
</tbody>
</table>

Survival analysis

Censored data happens a lot

- Real-world situation:
  - Life expectancy
  - Criminal recidivism rates
- Cybercrime applications:
  - Measuring time to remove X (where X = malware, phishing, scam website, ...)
  - Measuring time to compromise
  - Measuring time to re-infection
- Best resource I found on survival analysis in R: [http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf](http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf)

Survival analysis (package `survival` in R)

- Key challenge: estimating probability of survival when some data points survive at the end of the measurement.
  - Solution: use the Kaplan-Meier estimator to compute probabilities that account for samples still alive (`survfit` in R).
- Common question: Are survival functions split over categorical variables statistically different?
  - Use the log-rank test (`survdiff` in R).
  - Analogous to \( \chi^2 \) test.
- Cox-proportional hazard model (coxph in R) is a more sophisticated way to see how multiple variables affect the hazard rate.
  - Hazard function \( h(t) \): expected number of failures during the time period \( t \).
Regression and survival analysis

Pharmacy redirection duration by TLD

![Survival function for search results (TLD)](image)

Survival function for search results (TLD)

- all
- 95% CI
- .COM
- .ORG
- .EDU
- .NET
- other

1 days source infection remains in search results

Regression and survival analysis

Pharmacy redirection duration by PageRank

![Survival function for search results (PageRank)](image)

Survival function for search results (PageRank)

- all
- 95% CI
- PR≥7
- 0<PR<7
- PR=0

1 days source infection remains in search results

Regression and survival analysis

Statistics disentangle effect of TLD, PageRank on duration

Cox-proportional hazard model

\[ h(t) = \exp(\alpha + \text{PageRank}x_1 + \text{TLD}x_2) \]

<table>
<thead>
<tr>
<th></th>
<th>coef.</th>
<th>exp(coef.)</th>
<th>Std. Err.</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>-0.079</td>
<td>0.92</td>
<td>0.0094</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>.edu</td>
<td>-0.260</td>
<td>0.77</td>
<td>0.084</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>.net</td>
<td>0.100</td>
<td>1.1</td>
<td>0.081</td>
<td></td>
</tr>
<tr>
<td>.org</td>
<td>0.055</td>
<td>1.1</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>other TLDs</td>
<td>0.340</td>
<td>1.4</td>
<td>0.053</td>
<td>p &lt; 0.001</td>
</tr>
</tbody>
</table>

log-rank test: \( Q=159.6, p < 0.001 \)

Regression and survival analysis

Phishing website recompromise

- Full paper: [http://lyle.smu.edu/~tylerm/cs81.pdf](http://lyle.smu.edu/~tylerm/cs81.pdf)
- What constitutes recompromise?
  - If one attacker loads two phishing websites on the same server a few hours apart, we classify it as one compromise.
  - If the phishing pages are placed into different directories, it is more likely two distinct compromises.
- For simplicity, we define website recompromise as distinct attacks on the same host occurring \( \geq 7 \) days apart.
- 83% of phishing websites with recompromises \( \geq 7 \) days apart are placed in different directories on the server.
The Webalizer

Web page usage statistics are sometimes set up by default in a world-readable state. We automatically checked all sites reported to our feeds for the Webalizer package, revealing over 2,486 sites from June 2007–March 2008. 1,320 (53%) recorded search terms obtained from ‘Referrer’ header in the HTTP request. Using these logs, we can determine whether a host used for phishing had been discovered using targeted search.

Types of evil search

- Vulnerability searches: phpizabi v0.848b cl hp1 (unrestricted file upload vuln.), inurl: com_juser (arbitrary PHP execution vuln.)
- Compromise searches: allintitle: welcome paypal
- Shell searches: intitle: ‘’index of’’ r57.php, c99shell druxrwx

<table>
<thead>
<tr>
<th>Search type</th>
<th>Websites</th>
<th>Phrases</th>
<th>Visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any evil search</td>
<td>204</td>
<td>456</td>
<td>1,207</td>
</tr>
<tr>
<td>Vulnerability search</td>
<td>126</td>
<td>206</td>
<td>582</td>
</tr>
<tr>
<td>Compromise search</td>
<td>56</td>
<td>99</td>
<td>265</td>
</tr>
<tr>
<td>Shell search</td>
<td>47</td>
<td>151</td>
<td>360</td>
</tr>
</tbody>
</table>

One phishing website compromised using evil search

2: 2007-11-30  no evil search term  0 hits
3: 2007-12-01  no evil search term  0 hits
4: 2007-12-02  phpizabi v0.415b r3  1 hit
5: 2007-12-03  phpizabi v0.415b r3  1 hit
7: 2007-12-04  phpizabi v0.415b r3  1 hit
Let's work with the data

**R code:**
```
```

**Data format:**
```
<table>
<thead>
<tr>
<th>TLD</th>
<th>1st Compromise</th>
<th>2nd Compromise</th>
<th># days</th>
<th>Censored</th>
<th>Evil searches?</th>
</tr>
</thead>
<tbody>
<tr>
<td>com</td>
<td>2008-01-28</td>
<td>2008-03-31</td>
<td>63</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>com</td>
<td>2007-11-23</td>
<td>2008-03-31</td>
<td>129</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>IP</td>
<td>2008-01-16</td>
<td>2008-03-31</td>
<td>75</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>com</td>
<td>2008-01-16</td>
<td>2008-03-31</td>
<td>75</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>com</td>
<td>2007-10-28</td>
<td>2007-11-06</td>
<td>8</td>
<td>1</td>
<td>TRUE</td>
</tr>
<tr>
<td>com</td>
<td>2008-01-20</td>
<td>2008-03-31</td>
<td>71</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>jp</td>
<td>2007-11-12</td>
<td>2008-03-31</td>
<td>140</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>nu</td>
<td>2008-01-31</td>
<td>2008-03-31</td>
<td>60</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>net</td>
<td>2007-12-27</td>
<td>2008-03-31</td>
<td>95</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>com</td>
<td>2008-02-08</td>
<td>2008-03-31</td>
<td>52</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>IP</td>
<td>2007-12-07</td>
<td>2008-01-07</td>
<td>31</td>
<td>1</td>
<td>TRUE</td>
</tr>
<tr>
<td>IP</td>
<td>2008-01-29</td>
<td>2008-03-31</td>
<td>62</td>
<td>0</td>
<td>TRUE</td>
</tr>
<tr>
<td>com</td>
<td>2007-10-22</td>
<td>2007-11-14</td>
<td>22</td>
<td>1</td>
<td>TRUE</td>
</tr>
<tr>
<td>com</td>
<td>2008-01-22</td>
<td>2008-03-31</td>
<td>69</td>
<td>0</td>
<td>TRUE</td>
</tr>
</tbody>
</table>
```

### Step 1: Create a survival object

```r
# Remember the definition of censored
# 0 = has not been recompromised
# 1 = has been recompromised
> head(webzlt)
  dom startdate  enddate lt  censored hasevil tld
1 com 2008-01-28 2008-03-31 63 0 TRUE com
2 com 2007-11-23 2008-03-31 129 0 TRUE com
3 IP 2008-01-16 2008-03-31 75 0 TRUE com
4 com 2008-01-16 2008-03-31 75 0 TRUE com
5 com 2007-10-28 2007-11-06 8 1 TRUE com
6 com 2008-01-20 2008-03-31 71 0 TRUE com
> S.all<-Surv(time=webzlt$lt, event=webzlt$censored, type='right')
```

### Working with survival objects

1. Empirically estimate survival probability overall
   - Supply `survfit` with a constant right-hand side formula
   - E.g.:
     ```r
     surv.all<-survfit(S.all~1)
     ```

2. Empirically estimate survival probability compared to single categorical variable
   - Supply `survfit` with a constant categorical variable in right-hand side of formula
   - E.g.:
     ```r
     survfit(S.all~webzlt$hasevil)
     ```

3. Regression with survival probability as response variable
   - Supply `survfit` with a constant categorical variable in right-hand side of formula
   - E.g.:
     ```r
     coxph(S.all~webzlt$hasevil, method="breslow")
     ```

### #1: Empirically estimate survival probability overall

```r
S.all<-Surv(time=webzlt$lt, event=webzlt$censored, type='right')
surv.all<-survfit(S.all~1)
plot(surv.all, slab="t days before recompromise",
    ylab="S(t): probability website has not been recompromised within t days",
ylim=c(0.4,1), main="Survival function for phishing websites", lwd=1.5)
```
#2: Emp. estimate survival prob. for 1 cat. var.

Regression and survival analysis

Is the difference between survival probabilities across categories statistically significant?

> survdiff(S.all~webzlt$hasevil)

Call:
survdiff(formula = S.all ~ webzlt$hasevil)

    N Observed Expected (O-E)^2/E (O-E)^2/V
webzlt$hasevil=FALSE 746 140 156.7 1.79 13.4
webzlt$hasevil=TRUE 121 41 24.3 11.55 13.4

Chisq= 13.4 on 1 degrees of freedom, p= 0.000249

#3: Regression with survival prob. as response variable

Regression and survival analysis

Summary:

summary(evil.ph)


> summary(evil.ph)

Call:
coxph(formula = Surv(webzlt$lt, webzlt$censor) ~ webzlt$hasevil, method = "breslow")
n= 867, number of events= 181

 coefficients exp(coef) se(coef)      z  Pr(>|z|)
webzlt$hasevilTRUE 0.6393  1.8951  0.1778 3.595 0.000325 ***

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

exp(coef) exp(-coef) lower .95 upper .95
webzlt$hasevilTRUE 1.895  0.5277  1.337  2.685

Concordance= 0.539 (se = 0.013)
Rsquare= 0.013 (max possible= 0.932)
Likelihood ratio test= 11.43 on 1 df, p=0.0007219
Wald test = 12.92 on 1 df, p=0.0003246
Score (logrank) test = 13.37 on 1 df, p=0.000256

One more survival example: Bitcoin currency exchanges

- Bitcoin is a digital crypto-currency
- Decentralization is a key feature of Bitcoin’s design
- Yet an extensive ecosystem of 3rd-party intermediaries now supports Bitcoin transactions: currency exchanges, escrow services, online wallets, mining pools, investment services, ...
- Most risk Bitcoin holders face stems from interacting with these intermediaries, who act as de facto central authorities
- We focus on risk posed by failures of currency exchanges
Linode hackers escape with $70K in daring bitcoin heist

Compromised servers ransacked for digital cash

By John Leyden • Get more from this author
Posted in Security, 3rd March 2012 17:15 GMT

Updated Popular public host Linode has been hacked by cyber-thieves who made off with a stash of bitcoins worth $71,000 (444,796) in real money.

The crooks puffed up the heist after obtaining admin passwords for Linode’s network gear. Having infiltrated its systems, the thieves proceeded to target several bitcoin-related servers, siphoning $11k (69.49k) from one merchant and more than 10,000 bitcoins ($56k, £35k) from Bitcointica, a trading exchange for the digital currency.

Bitcointica has promised to reimburse customers for any losses. It said in a statement:

Many of you have heard that several bitcoin services were victims of a recent Linode security breach today. Unfortunately, Bitcointica is also among the services affected.

Hacker steals $250k in Bitcoins from online exchange Bitfloor

Irreversible transactions make Bitcoin a high-stakes business.

by Timothy B. Lee • Sept 4, 2012, 8:23pm EDT

The future of the up-and-coming Bitcoin exchange Bitfloor was thrown into question Tuesday when the company’s founder reported that someone had compromising his servers and made off with about 24,000 Bitcoins, worth almost a quarter-million dollars. The exchange no longer has enough cash to cover all of its deposits, and it has suspended its operations while it considers its options.

Bitfloor is not the first Bitcoin service brought low by hackers. Last year, the most popular Bitcoin exchange, Mt. Gox, suspended operations for a week after an attacker compromised a user account and stole all of his Bitcoins in a flurry that temporarily pushed the price down to zero. The site

Disclaimer: I’m not associated in any way with Mercado Bitcoin other than having done trades there. Luckily for me I didn’t have any money there at the moment.

Mercado Bitcoin: The largest and – according to bogus articles in Brazil – the largest bitcoin exchange in Brazil, has been offline for almost a week now. For the first three days there was no communications, but the owner just sent an email to all accounts explaining he was hacked. I haven’t seen it posted anywhere in English so I do my best to translate what I got.

As far as I understand, someone hacked his “redem code” feature, being able to generate fake credits in the system. And during the night the hacker moved out all his credit in Bitcoins, leaving Mercado Bitcoin without enough BTC to pay back all the other depositors.

Mercado heard that how much was robbed or more details than that, but has said he will try to pay back what he can. In this order:

1. Withdrawals in Reais that were requested before the attack
2. Deposits in Reais that had been credited yet
3. Current balances in Reals
4. Current balances in Bitcoins

Meaning that depending on how much was left, bitcoin balances will only be given back if he is able to pay back all the money (in Reais) to other creditors, and even that money isn’t fully guaranteed.
Regression and survival analysis

Data collection methodology

- **Data sources**
  - Daily transaction volume data on 40 exchanges converting into 33 currencies from bitcoincharts.com
  - Checked for closure, mention of security breaches and whether investors were repaid on Bitcoin Wiki and forums
  - To assess impact of pressure from financial regulators, we identified each exchange’s country of incorporation and used a World Bank index on compliance with anti-money laundering regulations

- **Key measure: exchange lifetime**
  - Time difference between first and last observed trade
  - We deem an exchange closed if no transactions are observed at least 2 weeks before data collection finished

Some initial summary statistics

- 40 Bitcoin currency exchanges opened since 2010
- 18 have subsequently closed (45% failure rate)
- Median lifetime is 381 days
- 45% of closed exchanges did not reimburse customers
- 9 exchanges were breached (5 closed)
### Regression and survival analysis

#### 18 closed Bitcoin currency exchanges

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BitcoiniMarket</td>
<td>US</td>
<td>4/10 – 6/11</td>
<td>2454</td>
<td>yes</td>
<td>yes</td>
<td>–</td>
<td>34.3</td>
</tr>
<tr>
<td>Bitstamp</td>
<td>PL</td>
<td>4/11 – 6/11</td>
<td>758</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>17.7</td>
</tr>
<tr>
<td>FreshBTC</td>
<td>PL</td>
<td>6/11 – 9/11</td>
<td>3</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>9.3</td>
</tr>
<tr>
<td>Bittrex</td>
<td>US/BG</td>
<td>4/11 – 10/11</td>
<td>528</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>33.3</td>
</tr>
<tr>
<td>ExchangeBitcoins</td>
<td>US</td>
<td>6/11 – 10/11</td>
<td>551</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>34.3</td>
</tr>
<tr>
<td>Bittrex.com</td>
<td>PL</td>
<td>8/11 – 10/11</td>
<td>380</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>21.7</td>
</tr>
<tr>
<td>Brazil Bitcoin</td>
<td>BR</td>
<td>9/11 – 11/11</td>
<td>0</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>24.3</td>
</tr>
<tr>
<td>Agor3</td>
<td>ES</td>
<td>9/11 – 1/12</td>
<td>11</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>30.7</td>
</tr>
<tr>
<td>Global Bitcoin Exchange</td>
<td>7</td>
<td>9/11 – 1/12</td>
<td>14</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>27.9</td>
</tr>
<tr>
<td>BitOnCash</td>
<td>US</td>
<td>4/11 – 1/12</td>
<td>18</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>34.3</td>
</tr>
<tr>
<td>TradeHill</td>
<td>US</td>
<td>6/11 – 2/12</td>
<td>5062</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>34.3</td>
</tr>
<tr>
<td>World Bitcoin Exchange</td>
<td>AU</td>
<td>8/11 – 2/12</td>
<td>220</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>25.7</td>
</tr>
<tr>
<td>Rosum</td>
<td>US</td>
<td>6/11 – 4/12</td>
<td>37</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>34.3</td>
</tr>
<tr>
<td>bitcoint</td>
<td>US/CN</td>
<td>5/12 – 7/12</td>
<td>75</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>29.2</td>
</tr>
<tr>
<td>bitcex</td>
<td>RU</td>
<td>9/11 – 7/12</td>
<td>528</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>27.7</td>
</tr>
<tr>
<td>IMEXX.com</td>
<td>SC</td>
<td>7/11 – 10/12</td>
<td>2</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>19.9</td>
</tr>
<tr>
<td>Crypto X Change</td>
<td>AU</td>
<td>11/11 – 11/12</td>
<td>874</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>25.7</td>
</tr>
<tr>
<td>Bitmarket.eu</td>
<td>PL</td>
<td>4/11 – 12/12</td>
<td>33</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>21.7</td>
</tr>
</tbody>
</table>

#### Regression and survival analysis

### What factors affect whether an exchange closes?

- We hypothesize three variables affect survival time for a Bitcoin exchange
  - **Average daily transaction volume** (positive)
  - **Experiencing security breach** (negative)
  - **AML/CFT compliance** (negative)

Since lifetimes are censored, we construct a Cox proportional hazards model:

\[
h_i(t) = h_0(t) \exp(\beta_1 \log(Daily\ vol_i) + \beta_2\ Breached_i + \beta_3\ AML_i).\]

### R code: Cox proportional hazards model

```r
cox.vh <- coxph(Surv(time=amlsv$lifetime,event=amlsv$censored,type='right') ~ log2(amlsv$dailyvol) + amlsv$Hacked + amlsv$All, method="breslow")
> cox.vh

Call: coxph(formula = Surv(time = amlsv$lifetime, event = amlsv$censored, type = "right"), ~ log2(amlsv$dailyvol) + amlsv$Hacked + amlsv$All, method = "breslow")

coef exp(coef) se(coef)     z     p
log2(amlsv$dailyvol) -0.17396 0.843 0.0719 -2.4185 0.016
amlsv$HackedTRUE 0.85685 2.360 0.5715 1.4992 0.130
amlsv$All 0.00411 1.000 0.0421 0.0978 0.920

Likelihood ratio test=6.28 on 3 df, p=0.0988, n=40, number of events= 18
```

### Notes
Cox proportional hazards model: results

<table>
<thead>
<tr>
<th></th>
<th>coef.</th>
<th>exp(coef.)</th>
<th>Std. Err.</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(Daily vol.)</td>
<td>$\beta_1$</td>
<td>-0.173</td>
<td>0.840</td>
<td>0.072</td>
</tr>
<tr>
<td>Breached,</td>
<td>$\beta_2$</td>
<td>0.857</td>
<td>2.36</td>
<td>0.572</td>
</tr>
<tr>
<td>AML</td>
<td>$\beta_3$</td>
<td>0.004</td>
<td>1.004</td>
<td>0.042</td>
</tr>
</tbody>
</table>

**log-rank test:** $Q=7.01$ ($p = 0.0715$), $R^2 = 0.145$

- Higher daily transaction volumes associated with longer survival times (statistically significant)
- Experiencing a breach associated with shorter survival times (not quite statistically significant)

Survival probability for Bitcoin exchanges

![Graph showing survival probability over days for Bitcoin exchanges.](image)

R code: Survival probability for Bitcoin exchanges

```r
par(mar=c(4.1,4.1,0.5,0.5))
plot(survfit(cox.vh),col="black",lty="solid",lwd=2,
     xlab="Days",
     ylab="Survival probability",
     cex.lab=1.3,
     cex.axis=1.3)
legend("topright",legend=c("Average"),col=c("black"),lwd=2,lty=c("solid"))
```

Reminder: data frame structure

```r
> cox.vh
Call:
coxph(formula = Surv(time = amlsv$lifetime, event = amlsv$censored, type = "right") ~ log2(amlsv$dailyvol) + amlsv$Hacked + amlsv$All, method = "breslow")

coef exp(coef) se(coef)  z  p
log2(amlsv$dailyvol)  -0.17396 0.8400 0.0719 -2.4185 0.016
amlsv$HackedTRUE 0.85685 2.3624 0.5715 1.4992 0.130
amlsv$All 0.00411 1.0041 0.0421 0.0978 0.920

Likelihood ratio test=6.28 on 3 df, p=0.0988 n= 40, number of events= 18

> head(amlsv[,c('dailyvol','Hacked','All')],10)
dailyvol Hacked All
Global Bitcoin Exchange 13.7413402 FALSE 27.866
Vircurex 5.6135567 TRUE 27.866
Crypto X Change 874.2331200 FALSE 25.670
World Bitcoin Exchange 220.0284211 TRUE 25.670
btc-e.com 2603.7702724 TRUE 32.330
Mercado Bitcoin 67.0104275 FALSE 24.330
Brazilian Bitcoin Market 0.1896721 FALSE 24.330
Canadian Virtual Exchange 832.3611224 FALSE 25.000
```
Regression and survival analysis

High-volume exchanges have better chance to survive

R code: High-volume exchanges have better chance to survive

R code:
```r
cplots<-survfit(cox.vh,newdata=amlsv)
par(mar=c(4.1,4.1,0.5,0.5))
plot(cplots[15],col="green",lty="dashed",lwd=2,
     xlab="Days",
     ylab="Survival probability",
     cex.lab=1.3,
     cex.axis=1.3)

#Mt Gox
lines(cplots[28],col="blue",lty="dotdash",lwd=2) #Intersango
lines(survfit(cox.vh),lwd=2) #Mean

legend("topright",legend=c("Mt. Gox","Intersango","Average"),
       col=c("green","blue","black"),lwd=2,
       lty=c("dashed","dotdash","solid"))
```

Regression and survival analysis

Low-volume exchanges have worse chance to survive

Yet some lower-risk exchanges collapse, high-risk survive

Notes