Regression and Survival Analysis

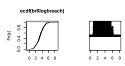
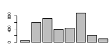
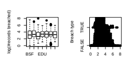
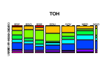
Tyler Moore

Computer Science & Engineering Department, SMU, Dallas, TX

Lecture 15–16

Notes

Guide to exploring data

Type of Data	Exploration	Statistics	RByEx
1 numerical variable		one way t-test, Wilcoxon test	6.3
1 categorical variable		–	3.1
# categories=2	–	`prop.test`	6.2
1 categorical, 1 numerical		anova, Permutation	10
# categories=2	–	2-way t, Wilcoxon test, Perm.	6.4
2 categorical variables		χ^2 test	3.2–3.5

Notes

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

Statistical Method	Response Variable	Explanatory Variable
Odds ratios	Binary (case/control)	Categorical variables (1 at a time)
Linear regression	Numerical	One or more variables (numerical or categorical)
Logistic regression	Binary	One or more variables (numerical or categorical)
Survival analysis	Time to event	One or more variables (numerical or categorical)

Notes

Linear regression

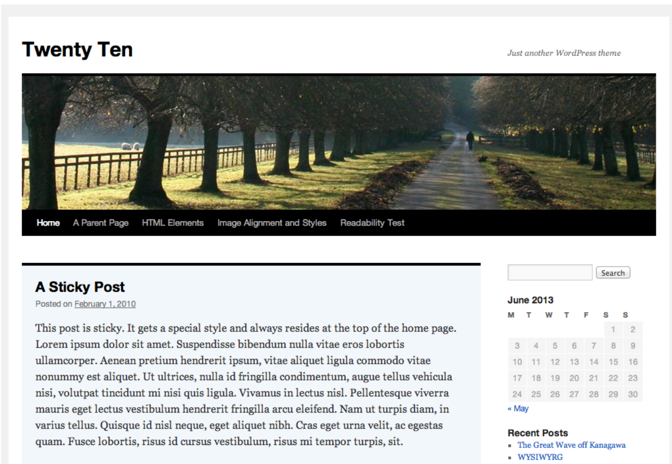
- Suppose the values of a numerical variable Y depend on the values of another variable X.

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

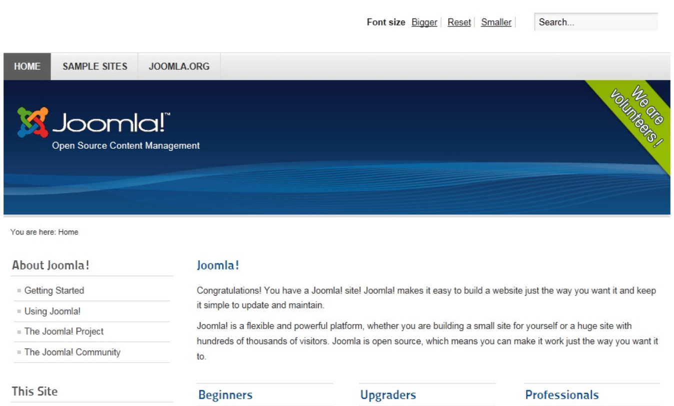
- If that dependence is linear then we can use linear regression to estimate the best-fit values of the constants c_0 and c_1 that minimize the error values for all the values $y_i \in Y$.
- For more info see "R by Example" Ch. 7.1–7.3

Notes

Notes



Notes



Notes

Notes

Dataset for linear regression example

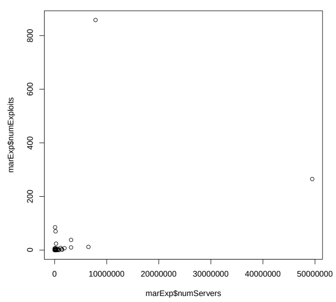
- Suppose you hypothesize that the popularity of a CMS platform influences the number of exploits made available
- We can use linear regression to test for such a relationship

generatorType	CMSmarketShare	numExploits
blogger	3.5	10
concrete5	0.1	1
contao	0.2	1
datalife engine	1.5	3
discuz	1.3	8
drupal	7.2	12

- Code: `http://lyle.smu.edu/~tylerm/courses/econsec/code/exregress.R`
- Data: `http://lyle.smu.edu/~tylerm/courses/econsec/data/eims.csv`

Notes

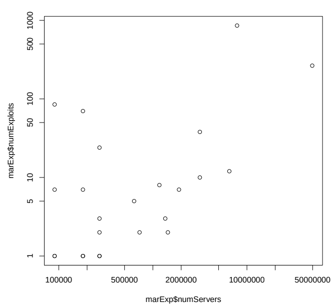
Scatter plot




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


Notes

Scatter plot (log-transformed)




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

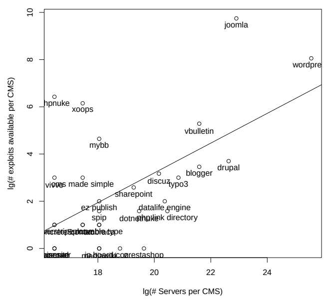

Notes

Linear regression

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

Call:
lm(formula = lgExploits ~ lgServers, data = marExp2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9692 -1.0655 -0.6013  0.5555  5.4554

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
```


Notes

Best-fit linear regression




```
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)
```

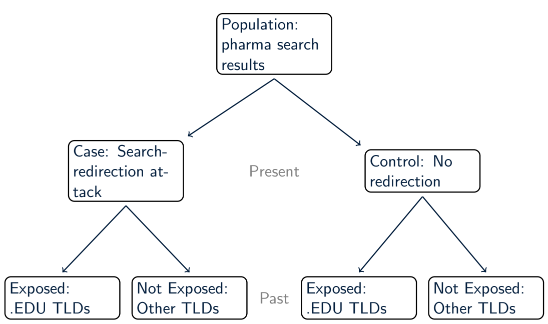

Notes

Illicit online pharmacies

- What do illicit online pharmacies have to do with phishing?
- Both make use of a similar criminal supply chain
  1. **Traffic**: hijack web search results (or send email spam)
  2. **Host**: compromise a high-ranking server to redirect to pharmacy
  3. **Hook**: affiliate programs let criminals set up website front-ends to sell drugs
  4. **Monetize**: sell drugs ordered by consumers
  5. **Cash out**: no need to hire mules, just take credit cards!
- For more: `http://lyle.smu.edu/~tylerm/usenix11.pdf`

Notes

Case-control study: search-redirection attacks



Notes

Case-control study: search-redirection attacks

R code: `http://lyle.smu.edu/~tylerm/courses/econsec/code/pharmaOdds.R`

Data format:

Date	Search Engine	Search Term	Pos.	URL	Domain	Redirects?	TLD
2011-11-03	Google	20 mg ambien overdose	1	http://products.sanofi.us/ambien/ambien.pdf	sanofi.us	False	other
2011-11-03	Google	20 mg ambien overdose	2	http://swift.sonoma.edu/education/newton/newtonsLaws/?20-mg-ambien-overdose	sonoma.edu	False	.EDU
2011-11-03	Google	20 mg ambien overdose	3	http://ambienoverdose.org/about-2/	ambienoverdose.org	False	.ORG
2011-11-03	Google	20 mg ambien overdose	4	http://answers.yahoo.com/question/index?qid=20090712025803AA10g8Z	yahoo.com	False	.COM
2011-11-03	Google	20 mg ambien overdose	5	http://en.wikipedia.org/wiki/Zolpidem	wikipedia.org	False	.ORG
2011-11-03	Google	20 mg ambien overdose	6	http://blocsonic.com/blog	blocsonic.com	False	.COM
2011-11-03	Google	20 mg ambien overdose	7	http://dinarvets.com/forums/index.php?/user/39154-ambien-side-effects/page	dinarvets.com	False	.COM
2011-11-03	Google	20 mg ambien overdose	8	http://nemo.mwd.hartford.edu/mwd08/images/?20-mg-ambien-overdose	hartford.edu	True	.EDU
2011-11-03	Google	20 mg ambien overdose	9	http://www.formspring.me/AmbienCheapOn	formspring.me	False	other
2011-11-03	Google	20 mg ambien overdose	11	http://www.drugs.com/pro/zolpidem.html	drugs.com	False	.COM
2011-11-03	Google	20 mg ambien overdose	12	http://www.engineer.tamuk.edu/departments/ieen/images/ambien.html	tamuk.edu	False	.EDU
2011-11-03	Bing	20 mg ambien overdose	1	http://answers.yahoo.com/question/index?qid=20090712025803AA10g8Z	yahoo.com	False	.COM
2011-11-03	Bing	20 mg ambien overdose	2	http://www.healthcentral.com/sleep-disorders/h/20-mg-ambien-overdose.html	healthcentral.com	False	.COM
2011-11-03	Bing	20 mg ambien overdose	3	http://ambien20mg.com/	ambien20mg.com	False	.COM
2011-11-03	bing	20 mg ambien overdose	4	http://www.chacha.com/question/will-20-mg-of-ambien-cr-get-you-high	chacha.com	True	.COM
2011-11-03	bing	20 mg ambien overdose	5	http://www.rxlist.com/ambien-drug.htm	rxlist.com	True	.COM
2011-11-03	Bing	20 mg ambien overdose	6	http://www.drugs.com/pro/zolpidem.html	drugs.com	False	.COM
2011-11-03	Bing	20 mg ambien overdose	7	http://answers.yahoo.com/question/index?qid=20111024222432AARFvPB	yahoo.com	False	.COM
2011-11-03	Bing	20 mg ambien overdose	8	http://en.wikipedia.org/wiki/Zolpidem	wikipedia.org	False	.ORG
2011-11-03	Bing	20 mg ambien overdose	9	http://www.thefullwiki.org/Sertraline	thefullwiki.org	False	.ORG
2011-11-03	bing	20 mg ambien overdose	10	http://www.rxlist.com/edluar-drug.htm	rxlist.com	True	.COM
2011-11-03	Bing	20 mg ambien overdose	11	http://www.formspring.me/ambienpill	formspring.me	False	other
2011-11-03	Bing	20 mg ambien overdose	12	http://ambiendosage.net/	ambiendosage.net	False	.NET

Notes

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

Statistical Method	Response Variable	Explanatory Variable
Odds ratios	Binary (case/control)	Categorical variables (1 at a time)
Linear regression	Numerical	One or more variables (numerical or categorical)
Logistic regression	Binary	One or more variables (numerical or categorical)
Survival analysis	Time to event	One or more variables (numerical or categorical)

Notes

Odds ratios for case-control study

```
> 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  1.0000000        NA        NA
    .EDU  5.8390966 5.5363269 6.1591917
    .GOV  0.4311855 0.3064817 0.5882604
    .NET  0.5946029 0.5568593 0.6342355
    .ORG  2.8811488 2.7971838 2.9674615
    other 1.3437113 1.2809207 1.4090669
```


Notes

Odds ratios for case-control study

```
> pr.tldodds$p.value
         two-sided
Predictor        midp.exact
    .COM                 NA
    .EDU  0.000000000000000
    .GOV  0.000000009212499
    .NET  0.000000000000000
    .ORG  0.000000000000000
    other 0.000000000000000
         two-sided
Predictor                                                           fisher.exact
    .COM                                                                      NA
    .EDU  0.00000000000000000000000000000000000000000000000000000000000000000000
    .GOV  0.00000001116730951558381248266507181077233923360836342908442020416260
    .NET  0.00000000000000000000000000000000000000000000000000000000000003109266
    .ORG  0.00000000000000000000000000000000000000000000000000000000000000000000
    other 0.00000000000000000000000000000002254063153941187904769660716762484880
         two-sided
Predictor                                                   chi.square
    .COM                                                            NA
    .EDU  0.000000000000000000000000000000000000000000000000000000000000
    .GOV  0.000000150899123313924415716095442548116967174109959159977734
    .NET  0.000000000000000000000000000000000000000000000000000000017562
    .ORG  0.000000000000000000000000000000000000000000000000000000000000
    other 0.000000000000000000000000000000000390896706121527347442976835
```


Notes

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).

$$\begin{aligned}
\text{odd's ratio} &= \frac{\text{odds}(A)}{\text{odds}(B)} \\
&= \frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}} \\
&= \frac{p_A \times (1-p_B)}{(1-p_A) \times p_B}
\end{aligned}$$

Notes

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 10 female applicants
  - $p_{\text{male acc.}} = 0.7$, $p_{\text{male rej.}} = 1 - 0.7 = 0.3$
  - $p_{\text{female acc.}} = 0.4$, $p_{\text{female rej.}} = 1 - 0.4 = 0.6$
  - $p_{\text{odds(male acc.)}} = \frac{0.7}{0.3} = 2.33$
  - $p_{\text{odds(female acc.)}} = \frac{0.4}{0.6} = 0.667$
  - $OR = \frac{2.33}{0.667} = 3.5$
- Hence, we can say that the odds of a male applicant being admitted are 3.5 times stronger than for a female applicant.

Notes

Back to the case-control study: how to interpret the odds ratios?

```
> 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  1.0000000        NA        NA
    .EDU  5.8390966 5.5363269 6.1591917
    .GOV  0.4311855 0.3064817 0.5882604
    .NET  0.5946029 0.5568593 0.6342355
    .ORG  2.8811488 2.7971838 2.9674615
    other 1.3437113 1.2809207 1.4090669
```


Notes

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

Statistical Method	Response Variable	Explanatory Variable
Odds ratios	Binary (case/control)	Categorical variables (1 at a time)
Linear regression	Numerical	One or more variables (numerical or categorical)
Logistic regression	Binary	One or more variables (numerical or categorical)
Survival analysis	Time to event	One or more variables (numerical or categorical)

Notes

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:

$$\log \frac{p_{redir}}{1-p_{redir}} = c_0 + c_1\, \text{Position}_1 + c_2\, \text{TLD}_2 + \epsilon$$

Notes

Logistic regression in action

- Code: http://lyle.smu.edu/~tylerm/courses/econsec/code/pharmaLogit.R

```
> 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:
    Min       1Q   Median       3Q      Max
-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 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)
```


Notes

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)
$N
[1] 175795

$R2
[1] 0.07736148
```


Notes

Obtaining the odds ratios

Recall the logistic regression equation

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

Exponentiate coefficients to get interpretable odds ratios

```
> coef(pr.logit)
(Intercept)     tld.EDU     tld.GOV     tld.NET     tld.ORG    tldother
 -1.8351654   1.7645946  -0.8451420  -0.5199959   1.0581945   0.2953898
> #get odds ratios for the coefficients plus 95% CI
> 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
tld.EDU     5.8392049 5.5364431 6.1584001
tld.GOV     0.4294964 0.3053796 0.5858515
tld.NET     0.5945230 0.5568118 0.6341472
tld.ORG     2.8811645 2.7972246 2.9675454
tldother    1.3436501 1.2808599 1.4090019
```


Notes

Logistic regression #2: TLD and search result position

```
> 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)
```


Notes

Logistic regression #2: TLD and search result position

```
> exp(cbind(OR = coef(pr.logit2), confint(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 5.5852012 6.2149893
tld.GOV        0.4314497 0.3067092 0.5886711
tld.NET        0.5878939 0.5505610 0.6271261
tld.ORG        2.8629455 2.7793345 2.9489947
tldother       1.3503082 1.2870831 1.4161226
resultPosition 1.0181977 1.0168021 1.0195962
> NagelkerkeR2(pr.logit2) #compute pseudo R^2 on logistic regression
$N
[1] 175795

$R2
[1] 0.08329341
```


Notes

Logistic regression #3: TLD, position, search engine

```
> 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
```


Notes

Logistic regression #3: TLD, position, search engine

```
> exp(cbind(OR = coef(pr.logit3), confint(pr.logit3)))
Waiting for profiling to be done...
                           OR      2.5 %     97.5 %
(Intercept)        0.07567444 0.0731465 0.07827858
tld.EDU            4.48253465 4.2449618 4.73330372
tld.GOV            0.42582135 0.3022669 0.58201442
tld.NET            0.65109897 0.6094052 0.69495871
tld.ORG            2.48399513 2.4099342 2.56025578
tldother           1.37590197 1.3107099 1.44382462
resultPosition     1.01877252 1.0173601 1.02018796
searchEnginegoogle 2.29579645 2.2348606 2.35850810
> NagelkerkeR2(pr.logit3) #compute pseudo R^2 on logistic regression
$N
[1] 175795

$R2
[1] 0.1166546
```


Notes

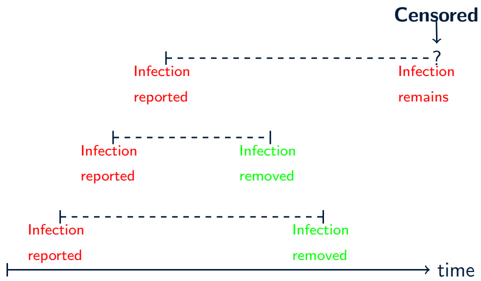
Guide to analyzing data

- 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

Statistical Method	Response Variable	Explanatory Variable
Odds ratios	Binary (case/control)	Categorical variables (1 at a time)
Linear regression	Numerical	One or more variables (numerical or categorical)
Logistic regression	Binary	One or more variables (numerical or categorical)
Survival analysis	Time to event	One or more variables (numerical or categorical)

Notes

Survival analysis



Notes

Censored data happens a lot

- Real-world situations
  - 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`

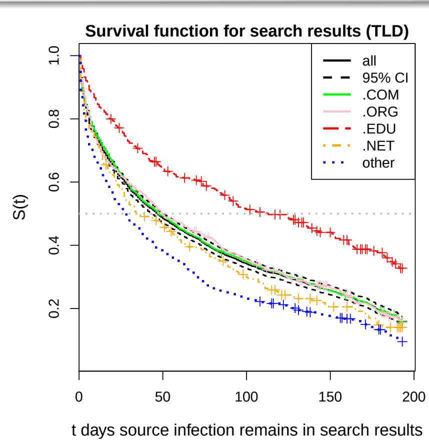
Notes

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)
  - Analagous to χ^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

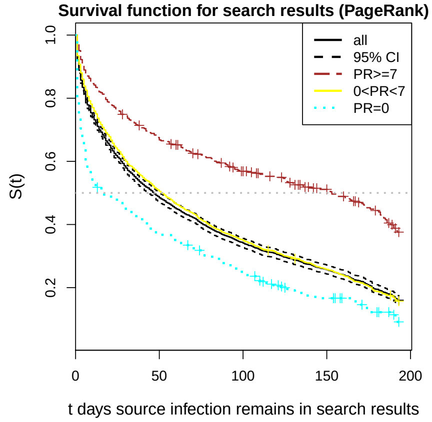
Notes

Pharmacy redirection duration by TLD



Notes

Pharmacy redirection duration by PageRank



Notes

Statistics disentangle effect of TLD, PageRank on duration

Cox-proportional hazard model $h(t) = \exp(\alpha + \text{PageRank}x_1 + \text{TLD}x_2)$				
	coef.	exp(coef.)	Std. Err.)	Significance
PageRank	-0.079	0.92	0.0094	$p < 0.001$
`.edu`	-0.26	0.77	0.084	$p < 0.001$
`.net`	0.10	1.1	0.081	
`.org`	0.055	1.1	0.052	
other TLDs	0.34	1.4	0.053	$p < 0.001$
log-rank test: Q=159.6, $p < 0.001$				

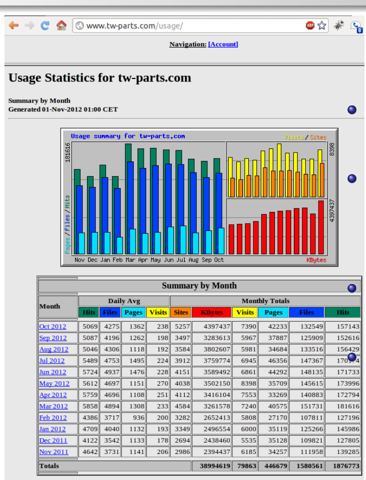
Notes

Phishing website recompromise

- Full paper: `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 ≥ 7 days apart
- 83% of phishing websites with recompromises ≥ 7 days apart are placed in different directories on the server

Notes

The Webalizer



Month	Daily Avg Hits	Daily Avg Files	Daily Avg Pages	Daily Avg Visits	Monthly Totals Sites	Monthly Totals KBytes	Monthly Totals Visits	Monthly Totals Pages	Monthly Totals Files	Monthly Totals Hits
Oct 2012	5069	4275	1362	238	5257	4397437	7390	42233	132549	157143
Sep 2012	5087	4196	1262	198	3497	3283613	5967	37887	125909	152616
Aug 2012	5046	4306	1118	192	3584	3802607	5981	34684	133516	156429
Jul 2012	5489	4753	1495	224	3912	3759774	6945	46356	147367	[illegible]
Jun 2012	5724	4937	1476	228	4151	3589492	6861	44292	148135	171733
May 2012	5612	4697	1151	270	4038	3502150	8398	35709	145615	173996
Apr 2012	5759	4696	1108	251	4112	3416104	7553	33269	140883	172794
Mar 2012	5858	4894	1308	233	4584	3261578	7240	40575	151731	181616
Feb 2012	4386	3717	936	200	3282	2652413	5808	27170	107811	127196
Jan 2012	4709	4040	1132	193	3349	2496554	6000	35119	125266	145986
Dec 2011	4122	3542	1133	178	2694	2438460	5535	35128	109821	127805
Nov 2011	4642	3731	1141	206	2986	2394437	6185	34257	111958	139285
Totals						**38994619**	**79863**	**446679**	**1580561**	**1876773**

- 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

Notes

Types of evil search

- Vulnerability searches: `phpizabi v0.848b c1 hfp1` (unrestricted file upload vuln.), `inurl: com_juser` (arbitrary PHP execution vuln.)
- Compromise searches: `allintitle: welcome paypal`
- Shell searches: `intitle: ''index of'' r57.php`, `c99shell drwxrwx`

Search type	Websites	Phrases	Visits
Any evil search	204	456	1 207
Vulnerability search	126	206	582
Compromise search	56	99	265
Shell search	47	151	360

Notes

One phishing website compromised using evil search

Notes

One phishing website compromised using evil search

1: 2007-11-30 10:31:33 phishing URL reported: `http://chat2me247.com/stat/q-mono/pro/www.lloydstsb.co.uk/lloyds_tsb/logon.ibc.html`		
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
6: 2007-12-04 21:14:06 phishing URL reported: `http://chat2me247.com/seasalter/www.usbank.com/online_banking/index.html`		
7: 2007-12-04	phpizabi v0.415b r3	1 hit

Notes

Let's work with the data

R code: `http://lyle.smu.edu/~tylerm/courses/econsec/code/surviveEvil2.R`

Data format:

TLD	1st Compromise	2nd Compromise	# days	Censored	Evil searches?
com	2008-01-28	2008-03-31	63	0	TRUE
com	2007-11-23	2008-03-31	129	0	TRUE
IP	2008-01-16	2008-03-31	75	0	TRUE
com	2008-01-16	2008-03-31	75	0	TRUE
com	2007-10-28	2007-11-06	8	1	TRUE
com	2008-01-20	2008-03-31	71	0	TRUE
jp	2007-11-12	2008-03-31	140	0	TRUE
nu	2008-01-31	2008-03-31	60	0	TRUE
net	2007-12-27	2008-03-31	95	0	TRUE
com	2008-02-08	2008-03-31	52	0	TRUE
IP	2007-12-07	2008-01-07	31	1	TRUE
IP	2008-01-29	2008-03-31	62	0	TRUE
com	2007-10-22	2007-11-14	22	1	TRUE
com	2008-01-22	2008-03-31	69	0	TRUE

Notes

Step 1: Create a survival object

```
#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  IP
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$lt,event=webzlt$censor,type='right')
```

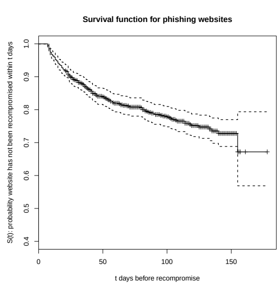

Notes

Working with survival objects

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

Notes

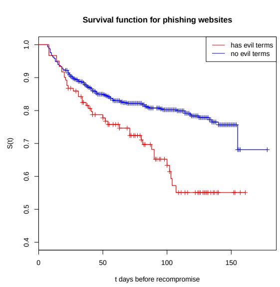
#1: Empirically estimate survival probability overall




```
S.all<-Surv(time=webzlt$lt,event=webzlt$censor,type='right')
surv.all<-survfit(S.all~1)
plot(surv.all,xlab='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)
```


Notes

#2: Emp. estimate survival prob. for 1 cat. var.




```
S.all<-Surv(time=webzlt$lt,event=webzlt$censor,type='right')
surv.evil<-survfit(S.all~webzlt$hasevil)
plot(surv.evil,xlab='t days before recompromise',
       ylab='S(t)',ylim=c(0.4,1), lwd=1.5,col=c('blue','red'),
       main='Survival function for phishing websites')
legend("topright",legend=c("has evil terms","no evil terms"),
       col=c("red","blue"),lty=1)
```


Notes

#2: Emp. estimate survival prob. for 1 cat. var.

- 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
```


Notes

#3: Regression with survival prob. as response variable

```
S.all<-Surv(time=webzlt$lt,event=webzlt$censor,type='right')
evil.ph <- coxph( S.all ~ webzlt$hasevil, method="breslow")
summary(evil.ph)
> summary(evil.ph)
Call:
coxph(formula = Surv(webzlt$lt, webzlt$censor) ~ webzlt$hasevil,
    method = "breslow")

  n= 867, number of events= 181

                     coef 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
```


Notes

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**
- R code: `http://lyle.smu.edu/~tylerm/data/bitcoin/bitcoinExScript.R`

Notes

Notes

Linode hackers escape wi ×

www.theregister.co.uk/2012/03/02/linode_bitcoin_heist/

Login | Sign up

The Register®

Data Center Cloud Software Networks **Security** Policy Business Jobs Hardware Science Bootnotes

Print | Like 28 | Tweet 131 | Alert

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, 2nd March 2012 17:05 GMT

Updated Popular web host Linode has been hacked by cyber-thieves who made off with a stash of bitcoins worth $71,000 (£44,736) in real money.

The crooks pulled off the heist after obtaining admin passwords for Linode's network gear. Having infiltrated its systems, the thieves proceeded to target several Bitcoin-related servers, stealing $15k (£9.45k) from one merchant and more than 10,000 bitcoins ($56k, £35k) from Bitcoinica, a trading exchange for the digital currency. Bitcoinica 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, Bitcoinica is also among the services affected.**

Notes

Hacker steals $250k in Bi ×

arstechnica.com/tech-policy/2012/09/hacker-steals-250k-in-bitcoins-from-online-exchange-bitfloor/

ars technica

MAIN MENU MY STORIES: 25 FORUMS SUBSCRIBE NOW

LAW & DISORDER / CIVILIZATION & DISCONTENTS

Hacker steals $250k in Bitcoins from online exchange Bitfloor

Irreversible transactions make Bitcoin security a high-stakes business.

by Timothy B. Lee - Sept 4 2012, 8:20pm CDT

INTERNET CRIME 88

The future of the up-and-coming Bitcoin exchange Bitfloor was thrown into question Tuesday when the company's founder reported that someone had compromised 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 sold all of his Bitcoins in a firesale that temporarily pushed the price down to zero. The site

Notes

The largest Bitcoin excha ×

www.reddit.com/r/Bitcoin/comments/1b8mtn/the_largest_bitcoin_exchange_in_brazil_gets/

MY SUBREDDITS FRONT - ALL - RANDOM | PICS - FUNNY - POLITICS - GAMING - ASKREDDIT - WORLDNEWS - VIDEOS - IAMA - TODAYILEARNED - WTF - AWW

BITCOIN comments related

138

The largest Bitcoin exchange in Brazil gets hacked: depositors are not guaranteed to get their money back (self.Bitcoin)

submitted 1 day ago by avsa

Disclaimer: I'm not associated in any form 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 only – bitcoin exchange in Brazil, has been offline for almost a week now. For the first few days there was no communication, 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'l do my best to translate what I got.

As far as I understood, someone hacked his "redeem code" feature, being able to generate false credits in the system. Then during the night the hacker moved out all his credit into bitcoins, leaving MercadoBitcoin without enough BTC to pay back all the other depositors.

Mercado hasn't revealed how much was robbed or more details than that, but has said he will try to pay back what he can, in that order:

1. Withdrawals in Reais that were requested before the attack
2. Deposits in Reais that hadn't been credited yet
3. Current balances in Reais
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.

Notes

https://bitcointalk.org/index.php?topic=5441.msg1413156#msg1413156

M4v3R
Hero Member

Posts: 524

Ignore

Re: BitMarket.Eu - ownership changed (in a way)
December 21, 2012, 08:53:16 AM #518

Hello all. I'm terrible sorry for not responding to this earlier. A mix of personal issues with searching for a solution prevented me from it.

Unfortunately, I have very bad news. I cannot currently proccess your withdrawals. The situation is very complicated and it's all my fault, that's why I feel terrible about it. I tried to make this up, to keep the site afloat and somehow recover the funds, but it's not possible anymore. Right now there are 1786 BTC pending withdrawal, which I can't honor...

Earlier this year, I had this "genius" idea which led me to making a fatal mistake. I thought I could provide a hedge fund service for Bitmarket users. There were other sites providing this service so I guesses that it could be successful. I had experience in trading before, all I needed is a platform. And there was one - Bitcoinica. I was so convinced with this idea (and sooo wrong in hindsight) that for a while I kept majority of "offline" Bitmarket funds there. What I didn't expect was that one day it could just dissapear - taking all the money with it. What's worse, the funds were shorted when it happened (converted to USD and sold) - and after Bitcoinica dissapeared BTC price rose by about 250% until now. So while there is still chance to recover the funds (there is an appointed liquidator assigned to this case and I've already sent in claims) it will be not enough to cover all people's funds. For the record - there are ~~20161~~ 18787.72139217 BTC missing (edit: I subtracted my funds that also were deposited on Bitmarket), and Bitcoinica claims total for around 50K USD (the exact amount is uncertain because the liquidators haven't yet stated at what rate they will liquidate positions).

Sadly, I alone, I'm out of options. I don't have own money to pay for this loss (Bitmarket never made any real profit and I make up for a living by part-time web/mobile programming). The options for making this up for everyone as I see are:

- find an investor (or investors) that is willing to cover at least part of is debt. I would transfer all rights to the website software, servers and database to him and also work as a technician, possibly also implementing features he'd wanted. If you reading this have the funds necessary to make this work, PLEASE contact me on this.
- freeze all current funds and "start over" trading with explicit fees, implementing much-needed features like rating system and others. All profits from the fees would go directly to a fund for repaying the debt. I'm afraid that this option

Notes

Notes

Data collection methodology

- Data sources
  1. Daily transaction volume data on 40 exchanges converting into 33 currencies from `bitcoincharts.com`
  2. Checked for closure, mention of security breaches and whether investors were repaid on Bitcoin Wiki and forums
  3. 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

Notes

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)

Notes

18 closed Bitcoin currency exchanges

Exchange	Origin	Dates Active	Daily vol.	Closed?	Breached?	Repaid?	AML
BitcoinMarket	US	4/10 – 6/11	2454	yes	yes	–	34.3
Bitomat	PL	4/11 – 8/11	758	yes	yes	yes	21.7
FreshBTC	PL	8/11 – 9/11	3	yes	no	–	21.7
Bitcoin7	US/BG	6/11 – 10/11	528	yes	yes	no	33.3
ExchangeBitCoins.com	US	6/11 – 10/11	551	yes	no	–	34.3
Bitchange.pl	PL	8/11 – 10/11	380	yes	no	–	21.7
Brasil Bitcoin Market	BR	9/11 – 11/11	0	yes	no	–	24.3
Aqoin	ES	9/11 – 11/11	11	yes	no	–	30.7
Global Bitcoin Exchange	?	9/11 – 1/12	14	yes	no	–	27.9
Bitcoin2Cash	US	4/11 - 1/12	18	yes	no	–	34.3
TradeHill	US	6/11 – 2/12	5082	yes	yes	yes	34.3
World Bitcoin Exchange	AU	8/11 – 2/12	220	yes	yes	no	25.7
Ruxum	US	6/11 – 4/12	37	yes	no	yes	34.3
btctree	US/CN	5/12 – 7/12	75	yes	no	yes	29.2
btcex.com	RU	9/10 – 7/12	528	yes	no	no	27.7
IMCEX.com	SC	7/11 – 10/12	2	yes	no	–	11.9
Crypto X Change	AU	11/11 – 11/12	874	yes	no	–	25.7
Bitmarket.eu	PL	4/11 – 12/12	33	yes	no	no	21.7

Notes

22 open Bitcoin currency exchanges

Exchange	Origin	Dates Active	Daily vol.	Closed?	Breached?	Repaid?	AML
bitNZ	NZ	9/11 – pres.	27	no	no	–	21.3
ICBIT Stock Exchange	SE	3/12 – pres.	3	no	no	–	27.0
WeExchange	US/AU	10/11 – pres.	2	no	no	–	30.0
Vircurex	US?	12/11 – pres.	6	no	yes	–	27.9
btc-e.com	BG	8/11 – pres.	2604	no	yes	yes	32.3
Mercado Bitcoin	BR	7/11 – pres.	67	no	no	–	24.3
Canadian Virtual Exchange	CA	6/11 – pres.	832	no	no	–	25.0
btcchina.com	CN	6/11 – pres.	473	no	no	–	24.0
bitcoin-24.com	DE	5/12 – pres.	924	no	no	–	26.0
VirWox	DE	4/11 – pres.	1668	no	no	–	26.0
Bitcoin.de	DE	8/11 – pres.	1204	no	no	–	26.0
Bitcoin Central	FR	1/11 – pres.	118	no	no	–	31.7
Mt. Gox	JP	7/10 – pres.	43230	no	yes	yes	22.7
Bitcurex	PL	7/12 – pres.	157	no	no	–	21.7
Kapiton	SE	4/12 – pres.	160	no	no	–	27.0
bitstamp	SL	9/11 – pres.	1274	no	no	–	35.3
InterSango	UK	7/11 – pres.	2741	no	no	–	35.3
Bitfloor	US	5/12 – pres.	816	no	yes	no	34.3
Camp BX	US	7/11 – pres.	622	no	no	–	34.3
The Rock Trading Company	US	6/11 – pres.	52	no	no	–	34.3
bitme	US	7/12 – pres.	77	no	no	–	34.3
FYB-SG	SG	1/13 – pres.	3	no	no	–	33.7

Notes

What factors affect whether an exchange closes?

- We hypothesize three variables affect survival time for a Bitcoin exchange
  1. **Average daily transaction volume** (positive)
  2. **Experiencing security breach** (negative)
  3. **AML/CFT compliance** (negative)
- Since lifetimes are censored, we construct a Cox proportional hazards model:

$$h_i(t) = h_0(t)\exp(\beta_1 \log(\text{Daily vol.})_i + \beta_2 \text{Breached}_i + \beta_3 \text{AML}_i).$$

Notes

R code: Cox proportional hazards model

```
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$dailyvol) -0.17396      0.84   0.0719 -2.4185 0.016
amlsv$HackedTRUE      0.85685      2.36   0.5715  1.4992 0.130
amlsv$All             0.00411      1.00   0.0421  0.0978 0.920

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


Notes

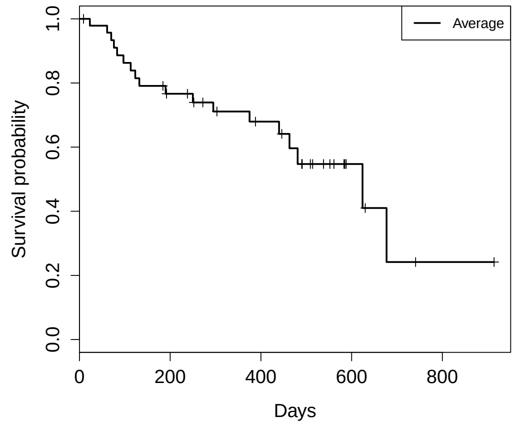
Cox proportional hazards model: results

		coef.	exp(coef.)	Std. Err.)	Significance
$\log(\text{Daily vol.})_i$	β_1	-0.173	0.840	0.072	$p = 0.0156$
Breached_i	β_2	0.857	2.36	0.572	$p = 0.1338$
AML_i	β_3	0.004	1.004	0.042	$p = 0.9221$
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)

Notes

Survival probability for Bitcoin exchanges



Notes

R code: Survival probability for Bitcoin exchanges

```
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"))
```

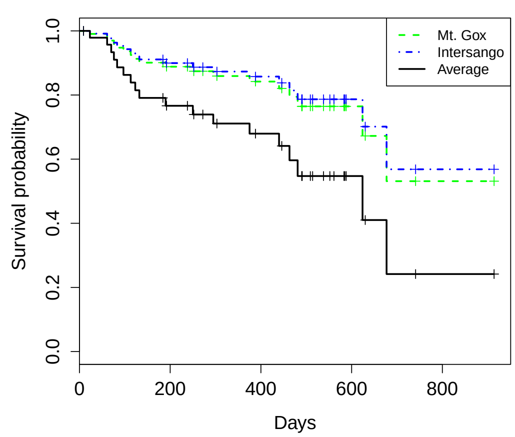

Notes

Reminder: data frame structure

```
> 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$dailyvol) -0.17396      0.84   0.0719 -2.4185 0.016
amlsv$HackedTRUE      0.85685      2.36   0.5715  1.4992 0.130
amlsv$All             0.00411      1.00   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 Exchnage   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
Brasil Bitcoin Market      0.1896721  FALSE 24.330
Canadian Virtual Exchange 832.3611224  FALSE 25.000
btcchina.com             472.6303602  FALSE 24.000
bitcoin-24.com           923.6339683  FALSE 26.000
```


Notes

High-volume exchanges have better chance to survive



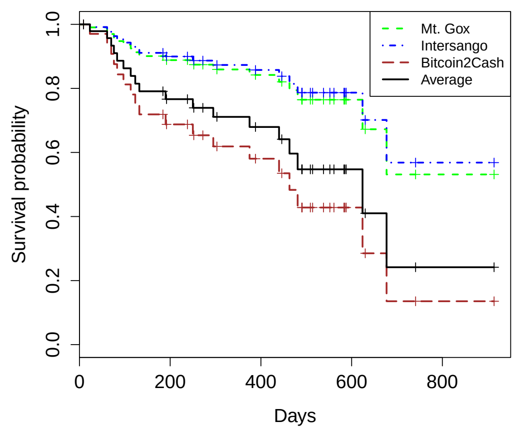
Notes

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

```
coxplots<-survfit(cox.vh,newdata=amlsv)
par(mar=c(4.1,4.1,0.5,0.5))
plot(coxplots[15],col="green",lty="dashed",lwd=2,
xlab="Days",
ylab="Survival probability",
cex.lab=1.3,
cex.axis=1.3
)  #Mt Gox
lines(coxplots[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"))
```

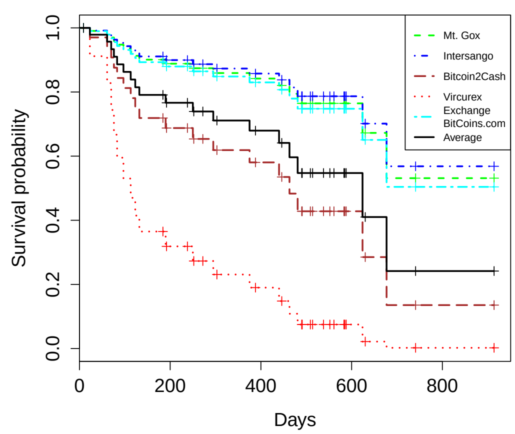

Notes

Low-volume exchanges have worse chance to survive



Notes

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



Notes

Notes