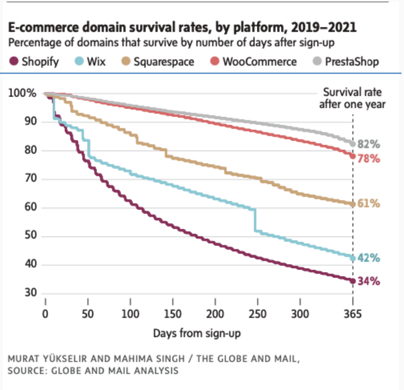
Methods of Applied Statistics I

STA2101H F LEC9101

Week 8

November 2 2022



Today

Start Recording

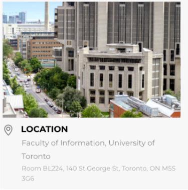
1. Upcoming events   No Class on November 9
2. Housekeeping
3. Recap
4. Observational studies and causality
5. Measures of risk
6. Generalized linear models
7. In the News

8. Office Hour Wednesday November 2: 4-5 pm in person; 7-8 pm on Zoom

- November 3 3.30-4.30 Statistical Sciences Seminar
  Room 9014, Hydro Building
  and online

  Alexandra Schmidt, McGill U
  "Modelling non-Gaussian spatio-temporal processes"

- November 10 9.00-6.00 CANSSI Ontario Statistical Software Conference
  BL224 140 St. George St.
  and online



Upcoming Part 2

London

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Home / News and events / Events / A celebration of 50 Years of the Cox model in memory of Sir David Cox

CONFERENCE *CENTRE FOR STATISTICAL METHODOLOGY series event*

A celebration of 50 Years of the Cox model in memory of Sir David Cox

Share

Photo shows Sir David Cox speaking at the Royal Statistical Society Conference. Photo credit: Royal Statistical Society.

Centre for Statistical Methodology Symposium

Where and when

Venue LSHTM, Keppel Street
London
WC1E 7HT
United Kingdom

Get Directions

Room John Snow Lecture Theatre and South Courtyard Café

Date Thursday 10 November 2022

Time 11:00 - 19:30

Date and time zone is UK

Admission

Registration required for in-person tickets. Free and open to all.

Contact

1972] 187

Regression Models and Life-Tables

By D. R. Cox

Imperial College, London

[Read before the Royal Statistical Society, at a meeting organized by the Research Section, on Wednesday, March 8th, 1972, Mr M. J. R. Healy in the Chair]

Summary

The analysis of censored failure times is considered. It is assumed that on each individual are available values of one or more explanatory variables. The hazard function (age-specific failure rate) is taken to be a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time. A conditional likelihood is obtained, leading to inferences about the unknown regression coefficients. Some generalizations are outlined.

Upcoming Part 2

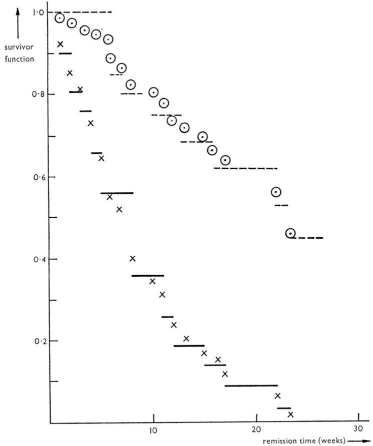
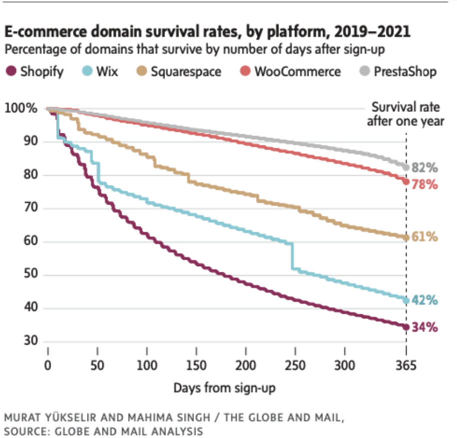


FIG. 1. Empirical survivor functions for data of Table 1. Product limit estimate, ------, sample 0 (6-MP); ———, sample 1 (control). Estimate constrained by proportionality: ⊙, sample 0; ×, sample 1. For clarity, the constrained estimates are indicated by the left ends of the defining horizontal lines.



Housekeeping

- Project – see course web page for outline and marking scheme

- Homework

  marking
  - HW7 due Nov 4 (Friday)
  - HW8 posted Nov 2/3/4 due Nov 16 (Wednesday)
  - HW9 posted Nov 16/17 due Nov 23 (Wednesday)
  - HW10 (Last) posted Nov 23/24 due Dec 1 (Wednesday)

- Syllabus – see course web page for updated syllabus
  - nonparametric regression (ELM-2 Ch.14, ELM-1 Ch.11)
  - survival data analysis (SM Ch.5.4, 10.8)
  - analysis of categorical responses (ELM-2 Ch. 6,7, ELM-1 Ch.5)
  - random effects and mixed models (ELM2 Ch.10, ELM-1 Ch.8)
  - longitudinal data analysis (ELM-2 Ch.11, ELM-1 Ch.9)

Recap

- likelihood function inference Cheatsheet
- Maximum Likelihood Estimate $\hat{\theta}$ and estimated cov matrix $\{-\ell''(\hat{\theta})\}^{-1} = j(\hat{\theta})^{-1}$
- Likelihood ratio test and nested models $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\}$
- Application to binomial: regression model and saturated model
- Residual deviance as a test of model fit
- Pearson's χ^2 correction

$$\sum_{i=1}^{m} \left[\left\{ \frac{y_i - n_i p_i(\hat{\beta})}{n_i p_i(\hat{\beta})} \right\}^2 + \left\{ \frac{n_i - y_i - n_i(1 - p_i(\hat{\beta}))}{n_i\{1 - p_i(\hat{\beta})\}} \right\}^2 \right] = \ldots =$$

Overdispersion

"Boxes of trout eggs were buried at five different stream locations and retrieved at 4 different times. The number of surviving eggs was recorded. The box was not returned to the stream."

J. Hinde, C.G.B. Demétrio / Computational Statistics & Data Analysis 27 (1998) 151–170 159

Table 3
Trout egg data

Location	Survival period (weeks)			
in stream	4	7	8	11
1	89/94	94/98	77/86	141/155
2	106/108	91/106	87/96	104/122
3	119/123	100/130	88/119	91/125
4	104/104	80/97	67/99	111/132
5	49/93	11/113	18/88	0/138

- $Y_i \sim Bin(n_i, p_i) \Rightarrow \mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = n_i p_i(1 - p_i)$
- variance is determined by the mean

- ```
  bmod <- glm(cbind(survive,total-survive) ~ location + period, family = binomial,
              data = troutegg)

  summary(bmod)
   Null deviance: 1021.469  on 19  degrees of freedom
  ## Residual deviance:   64.495  on 12  degrees of freedom
  ## AIC: 157.03
  ```
- quasi-binomial: $\mathrm{E}(Y_i) = n_i p_i, \quad \mathrm{Var}(Y_i) = \phi n_i p_i(1 - p_i)$
- estimate ϕ? over-dispersion parameter
- usually use $X^2/(n-p)$, where

$$X^2 = \sum \frac{(y_i - n_i\hat{p}_i)^2}{n\hat{p}_i(1 - \hat{p}_i)}$$

⟶ overdisp.Rmd
  ```

- the estimation of over-dispersion, and use of *t*- and *F*-tests, is approximate
- there isn't a binomial model with this structure
- but it is sometimes a handy fudge

- a more formal approach is to find a more flexible distribution for responses that are binary, or proportions
- for example, the beta distribution on $(0, 1)$ has two parameters ELM-2 §3.6

$$f(y \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 < y < 1$$

- $$E(Y) = \mu = \frac{\alpha}{\alpha+\beta}, \qquad \text{var}(Y) = \frac{\mu(1-\mu)}{1+\alpha+\beta} = \frac{\mu(1-\mu)}{1+\phi}, \quad \phi = \alpha + \beta$$
- $\text{logit}(\mu_i) = x_i^T\beta$, etc. $1/(1+\phi)$ is now the overdispersion parameter

Measures of risk

- see posted handout on case-control studies
- consider for simplicity binomial responses with a single binary covariate:

$$\text{logit}(p_i) \sim \beta_0 + \beta_1 z_i, \quad i = 1, \dots, n$$

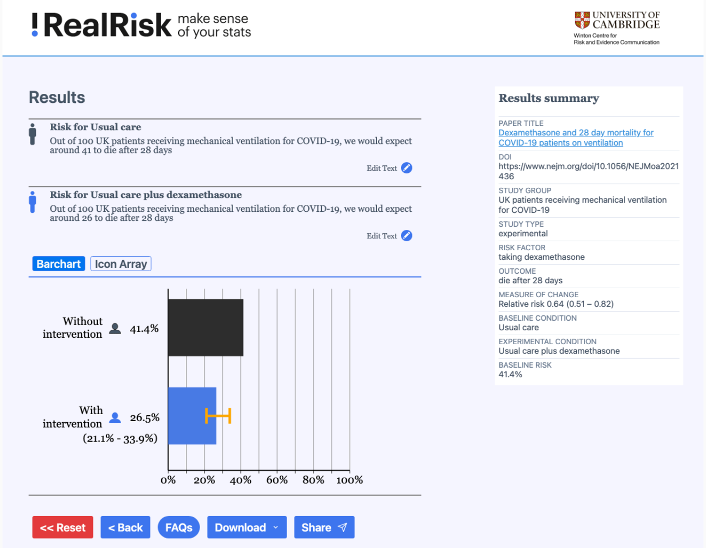
- no difference between groups $\iff$ odds-ratio $\equiv 1 \iff \beta_1 = 0$
- odds ratio of 3 or more is considered "large"

... Measures of risk

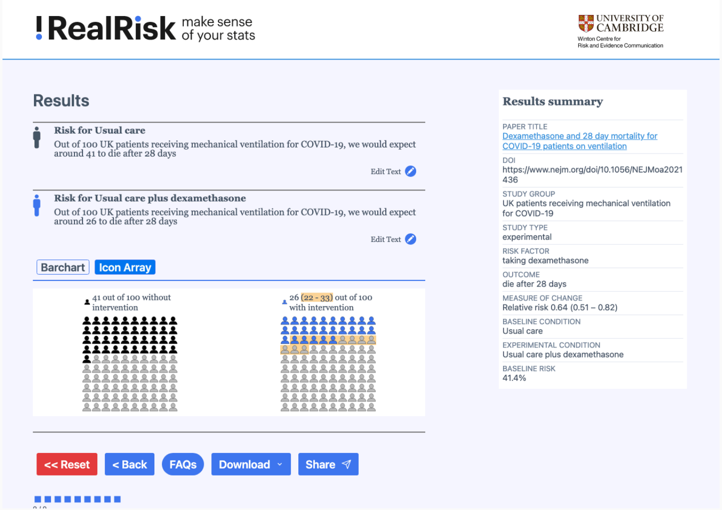
- we might be interested in risk ratio $\frac{p_1}{p_0}$ instead of odds ratio $\frac{p_1(1-p_0)}{p_0(1-p_1)}$
- also called relative risk
- if p_1 and p_0 are both small, ($y = 1$ is rare), then

$$\frac{p_1}{p_0} \approx \frac{p_1(1-p_0)}{p_0(1-p_1)}$$

- sometimes p_1/p_0 can be large but if p_1 and p_0 are both small the risk difference $p_1 - p_0$ might also be very small
- in order to estimate the difference we need to know the baseline risk p_0
- bacon sandwiches www.youtube.com/watch?v=4szyEbU94ig
- risk calculator https://realrisk.wintoncentre.uk/p1



Odds ratio 0.64; baseline risk 41.4%



Odds ratio 0.64; baseline risk 41.4%

Odds ratio 2.91; baseline risk 1/1000

Whether we sample prospectively or retrospectively, the odds ratio is the same

	Lung cancer	
	1	0
	cases	controls
smoke = 1 (yes)	688	650
smoke = 0 (no)	21	59
	709	709

$$\text{retro: } OR = \frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{650 \times 21} = 2.97$$

$$\text{prosp: } OR = \frac{\{688/(688+650)\}/\{650/(688+650)\}}{21/(21+59)/\{59/(21+59)\}} = \frac{688 \times 59}{650 \times 21} = 2.97$$

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of some feature of a defined population

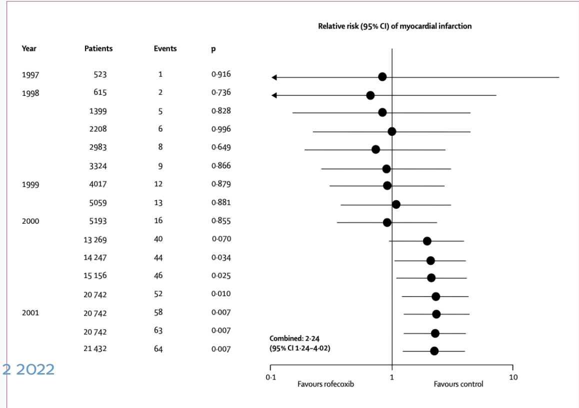
  could in principle be found exactly
- tracking across time of such features
- study of a relationship between features, where individuals may be examined
  - at a single time point
  - at several time points for different individuals
  - at different time points for the same individual
- census
- meta-analysis: statistical assessment of a collection of studies on the same topic

Effect sizes

- Meta-analyses combine the results from many different studies
- it is helpful if the coefficient estimates are all on the same scale

- Example: Jüni et al., 2004 Rofecoxib trials

online



... Effect sizes

- Several 'effect estimates' have been proposed
- in the context of these meta-analyses

- Cohen's *d* is a difference in means, divided by an estimate of the standard deviation of the difference
  not the standard error

- relative risks, or odds-ratios, for $0, 1$ explanatory variables are already on a standardized scale
  related to probabilities

- A-level maths paper referred to standardized estimates of β after logistic regression
- this might be a re-scaling of the covariates (math ability, etc.) to standardized units
  ??

To understand how Cohen's *d* for two independent groups is calculated, let's first look at the formula for the *t*-statistic:

$$t = \frac{\overline{M}_1 - \overline{M}_2}{\text{SD}_{\text{pooled}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Here $\overline{M}_1 - \overline{M}_2$ is the difference between the means, and $\text{SD}_{\text{pooled}}$ is the pooled standard deviation (Lakens, 2013), and n1 and n2 are the sample sizes of the two groups that are being compared. The *t*-value is used to determine whether the difference between two groups in a *t*-test is statistically significant (as explained in the chapter on *p*-values. The formula for Cohen's *d*_ is very similar:

$$d_s = \frac{\overline{M}_1 - \overline{M}_2}{\text{SD}_{\text{pooled}}}$$

As you can see, the sample size in each group (n_1 and n_2) is part of the formula for a *t*-value, but it is not part of the formula for Cohen's *d* (the pooled standard deviation is computed by

Which reminds me

Hünermand & Louw

On the Nuisance of Control Variables in Regression Analysis

Paul Hünermund
Copenhagen Business School, Kilevej 14A, Frederiksberg, 2000, DK.
phu.si@cbs.dk

Beyers Louw
Maastricht University, Tongersestraat 53, 6211 LM Maastricht, NL.
jb.louw@maastrichtuniversity.nl

September 28, 2022

Which reminds me

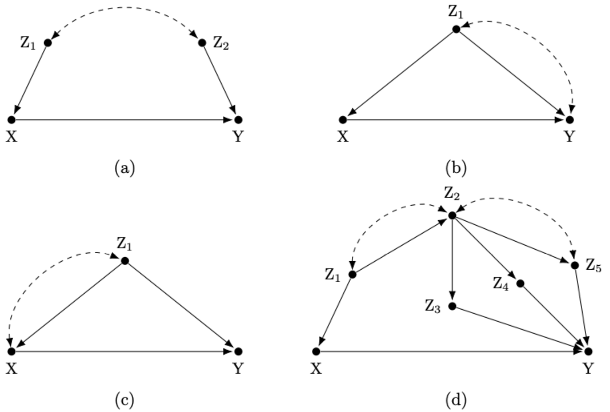


Figure 1: Examples of causal diagrams with valid control variable Z_1

can estimate causal effect of X on Y by controlling for Z_1, but cannot estimate causal effect of Z_1 on Y

Binary/Binomial responses, extras

- with binary data, may get complete separation of 1s and 0s
- leading to likelihood function not maximized at finite β ELM-2 2.7

- sometimes binary responses can be thought of as an indicator for the size of a latent variable Z, ELM-2 4.1
- i.e. $Y = 1 \iff Z > c$ for some fixed c
- distribution of Z sometimes called a tolerance distribution

- could be, e.g. $Z \sim N(0, 1)$, then $Y = 1$ with probability
- if $Z \sim \mathit{Logistic}$, then $Y = 1$ with probability $\frac{\exp(y-\mu)/\sigma}{1+\exp(y-\mu)/\sigma}$

```
link
a specification for the model link function. This can be a name/expression,
a literal character string, a length-one character vector,
or an object of class "link-glm" (such as generated by make.link)
provided it is not specified via one of the standard names given next.

The gaussian family accepts the links (as names) identity, log and inverse;
the binomial family the links logit, probit, cauchit,
(corresponding to logistic, normal and Cauchy CDFs respectively)
log and cloglog (complementary log-log);
the Gamma family the links inverse, identity and log;
the poisson family the links log, identity, and sqrt;
and the inverse.gaussian family the links 1/mu^2, inverse, identity and log.
```

Generalized linear models

`glm` has several options for `family`

```
binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```

Each of these is a member of the class of generalized linear models

Generalized: distribution of response is not assumed to be normal

Linear: some transformation of $E(y_i)$ is of the form $x_i^T\beta$ link function

Generalized linear models: theory

- $f(y_i; \mu_i, \phi_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i; \phi_i)\}$
- $E(y_i \mid x_i) = b'(\theta_i) = \mu_i$ defines μ_i as a function of θ_i
- $g(\mu_i) = x_i^T\beta = \eta_i$ links the n observations together via covariates
- $g(\cdot)$ is the link function; η_i is the linear predictor
- $\text{Var}(y_i \mid x_i) = \phi_i b''(\theta_i) = \phi_i V(\mu_i)$
- $V(\cdot)$ is the variance function

Examples

- Normal: $f(y_i;\mu_i,\sigma^2) = \frac{1}{\sqrt(2\pi)\sigma}\exp\{-\frac{1}{2\sigma^2}(y_i-\mu_i^2)\}$
  $$= \exp\{\frac{y_i\mu_i - (1/2)\mu_i^2}{\sigma^2} - (1/2)\log\sigma^2 - y_i^2/2\sigma^2 - (1/2)\log\sqrt(2\pi)\}$$
  $$\phi_i = \sigma^2, \quad \theta_i = \mu_i, \quad b(\mu_i) = \mu_i^2/2, b'(\mu_i) = \mu_i, b''(\mu_i) = 1$$
- Binomial: $f(r_i;p_i) = \binom{m_i}{r_i} p_i^{r_i}(1-p_i)^{m_i-r_i}; \quad y_i = r_i/m_i$
  $$= \exp[m_i y_i \log\{p_i/(1-p_i)\} + m_i\log(1-p_i) + \log\binom{m_i}{m_i y_i}]$$
  $$\phi_i = 1/m_i, \quad \theta_i = \log\{p_i/(1-p_i)\}, \quad b(p_i) = -\log(1-p_i), \quad p_i = E(y_i)$$
- ELM (§8.1/6.1) uses $a_i(\phi)$ in place of ϕ_i, later $a_i(\phi) = \phi/w_i$;
  SM uses ϕ_i, later (p. 483) $\phi_i = \phi a_i$

Family	Canonical link	Variance function	ϕ_i
Normal	$\eta = \mu$	1	σ^2
Binomial	$\eta = \log\{\mu/(1-\mu)\}$	$\mu(1-\mu)$	$1/m_i$
Poisson	$\eta = \log(\mu)$	μ	1
Gamma	$\eta = 1/\mu$	μ^2	$1/\nu$
Inverse Gaussian	$\eta = 1/\mu^2$	μ^3	ξ

Gamma: $f(y_i;\mu_i,\nu) = \dfrac{1}{\Gamma(\nu)}\left(\dfrac{\nu}{\mu_i}\right)^{\nu} y_i^{\nu-1}\exp(-\dfrac{\nu}{\mu_i})y_i$

$$= \exp[-\frac{\nu}{\mu_i}y_i - \nu\log(\frac{1}{\mu_i}) + (\nu-1)\log(y_i) + \nu\log(\nu) - \log\{\Gamma(\nu)\}]$$

$$= \exp\{\nu(\frac{y_i}{-\mu_i} - \log(\frac{1}{\mu_i}) + (\nu-1)\log(y_i) - \log\Gamma(\nu) + \nu\log(\nu)\}$$

Summary

Model:

$$\mathbb{E}(y_i) \quad = \quad \mu_i; \qquad g(\mu_i) = x_i^T\beta; \qquad \text{Var}(y_i) = \phi_i V(\mu_i) \qquad \phi_i = a_i\phi$$

Estimation:

$$\hat{\beta} \quad = \quad (X^TWX)^{-1}X^TWz; \quad z = X\beta + W^{-1}u; \qquad z(\beta) = X\beta + W^{-1}(\beta)u(\beta)$$

Variance:

$$\text{Var}(\hat{\beta}) \quad \doteq \quad (X^TWX)^{-1}$$

W is diagonal

On pp. 118-119 of ELM, this iteration is carried out in `R` on the `bliss` data

Summary 2

$$\hat{\beta} = (X^T W X)^{-1} X^T W z; \quad z = X\beta + W^{-1}u; \qquad z(\beta) = X\beta + W^{-1}(\beta)u(\beta)$$

$$\text{Var}(\hat{\beta}) \doteq (X^T W X)^{-1} \qquad W \text{ is diagonal}$$

$$W_{ii} =$$

$$u_i =$$

Note $\hat{\beta}$ is free of ϕ because of W and W^{-1}, but $\text{Var}(\hat{\beta})$ depends on ϕ

Warning: in ELM W is defined slightly differently (no ϕ), so he has $\text{Var}(\hat{\beta}) = (X^T W X)^{-1}\hat{\phi}$

Summary 2

$$\hat\beta = (X^TWX)^{-1}X^TWz; \quad z = X\beta + W^{-1}u; \qquad z(\beta) = X\beta + W^{-1}(\beta)u(\beta)$$

$$\text{Var}(\hat\beta) \doteq (X^TWX)^{-1} \qquad W \text{ is diagonal}$$

$$W_{ii} = \frac{1}{\phi a_i\{g'(\mu_i)\}^2 V(\mu_i)}$$

$$u_i = \frac{y_i - \mu_i}{\phi a_i g'(\mu_i)V(\mu_i)}$$

Note $\hat\beta$ is free of ϕ because of W and W^{-1}, but $\text{Var}(\hat\beta)$ depends on ϕ

Warnings

1. in ELM W is defined slightly differently (no ϕ), so he writes $\widehat{\text{Var}}(\hat\beta) = (X^TWX)^{-1}\hat\phi$
2. ELM uses w_i where SM uses $1/a_i$

Analysis of data using GLMs: overview

- choose a model, often based on type of response — or on mean/variance relationship
- fit a model, using maximum likelihood estimation — convergence (almost) guaranteed
- inference for individual coefficients $\hat{\beta}_j$ from `summary`
- inference for groups of coefficients by analysis of deviance

- estimation of ϕ based on Pearson's Chi-square

typo in ELM p.121: cross out $=$ var($\hat{\mu}$)

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

- analysis of deviance: see p. 121 (near bottom) — likelihood ratio tests
- diagnostics: same as for `lm` — ELM p.124; SM p.477
  - residuals: deviance or Pearson; can be standardized — ELM likes 1/2 normal plots
  - influential observations: uses hat matrix — `SMPracticals` has very good GLM diagnostics

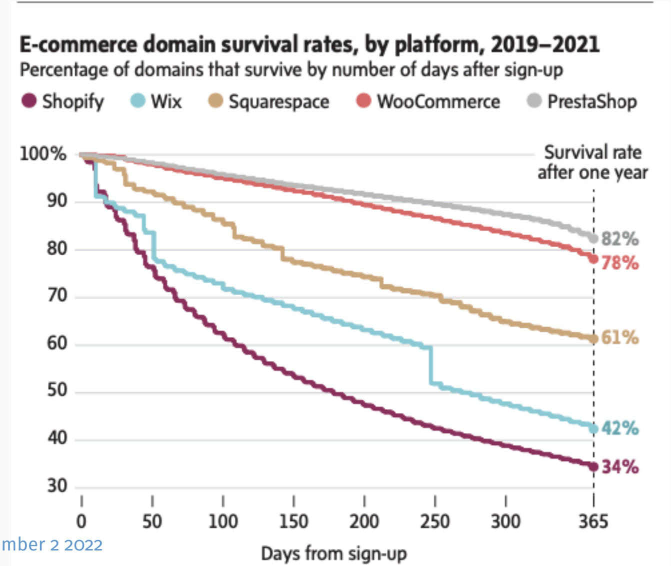
`glm.diag`, `plot.glm.diag`

In the News


Report on Business
SATURDAY, OCTOBER 22, 2022 | GLOBEANDMAIL.COM
IAN McGUGAN
The question investors can no longer ignore: How do you recession-proof a portfolio?
ENERGY
Chemical engineer Peter Guthrie to be Alberta's next energy minister
PERSONAL FINANCE
Don't worry, young adults, CPP and EI will be there when you need them
INVESTING
Equity strategist buys health care stocks, moves away from banks
COVER STORY
Back to basics
Almost a year into his tenure at Rogers Communications, CEO Tony Staffieri shares what dominates his focus: service and performance. Andrew Willis reports
How Europe is trying to build a future free of fossil fuels during an energy crisis
MARKETS
COMPANIES
Shopify has a growing problem with customer retention, analysis reveals
SPORTS
Bryan Trottier's memoir shows a gentle macho positivity, Cathal Kelly says
Toronto Blue Jays agree to three-year deal with manager John Schneider
Women's World Cup draw to set the stage for 32-team tournament next year
WHAT'S THE


SATURDAY, OCTOBER 22, 2022 | THE GLOBE AND MAIL
REPORT ON BUSINESS | B7
Shopify: Company's customer survival rate is substantially lower than its rivals, data show
Shopify's survival rate
34% of stores survive after 365 days
Survival rate of Shopify stores by year of sign-up
Percentage of Shopify Plus domains that survive
E-commerce domain survival rates, by platform, 2019–2021
Stores migrating from Shopify to other platforms
WooCommerce
Custom Cart
Wix
Square Online
Yampi
cartX
Tiendanube
SHOPLINE
Shoplazza
GoDaddy Online Store

Shopify



E-commerce domain survival rates, by platform, 2019–2021
Percentage of domains that survive by number of days after sign-up
Shopify
Wix
Squarespace
WooCommerce
PrestaShop
Survival rate after one year
100%
90
80
70
60
50
40
30
82%
78%
61%
42%
34%
0
50
100
150
200
250
300
365
Days from sign-up

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

Check for updates

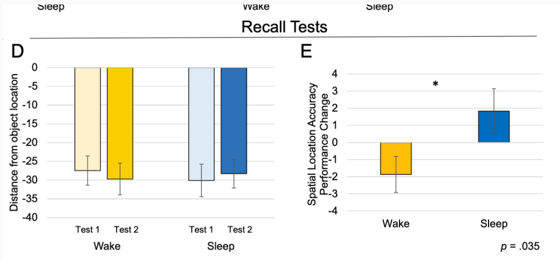
Sleep facilitates spatial memory but not navigation using the Minecraft Memory and Navigation task

Katharine C. Simon[a,1], Gregory D. Clemenson[b], Jing Zhang[a], Negin Sattari[a], Alessandra E. Shuster[a], Brandon Clayton[a], Elisabet Alzueta[c], Teji Dulai[c], Massimiliano de Zambotti[c], Craig Stark[b], Fiona C. Baker[c,d], and Sara C. Mednick[a]

Edited by Thomas Albright, Salk Institute for Biological Studies, La Jolla, CA; received February 11, 2022; accepted August 4, 2022

Sleep facilitates hippocampal-dependent memories, supporting the acquisition and mainte-


A
Environment 1
Environment 2
Environment 3
Environment 4


Sleep
Wake
Sleep
Recall Tests
D
0
-5
-10
-15
-20
-25
-30
-35
-40
Distance from object location
Test 1
Test 2
Test 1
Test 2
Wake
Sleep
E
4
3
2
1
0
-1
-2
-3
-4
Spatial Location Accuracy Performance Change
*
Wake
Sleep
p = .035