Notes on Harrell: Regression Modeling Strategies, 2nd edition

Nicholas J. Cox, Durham University, July 2022

p.i Fienberg

p.viii course notes ... cover basic regression and many other topics

p.1 prefer 'indicator variable' to 'dummy variable' [other examples, not listed here] (for some history and my prejudices, see Section 2 of Cox and Schechter (2019); I have horror stories of misplaced outrage when a speaker's remark that "gender is just a dummy variable" was wildly misunderstood by qualititative researchers)

p.5 Dichotomizing

p.5 footnote c the "cost" of a false positive [extra space needed]

p.5 footnote c prefer 'use' to 'utilize' [other examples, not listed here]

p.5 footnote c unknown Y

p.19 females or males

p.19 value of age

p.19 semi-colon before 'hence power and precision will suffer'

p.20 12 studies were included

p.20 which happen to optimize

p.25 Figure mixes red and green? (also pp.27, 151, 156, 204, 213, 245, 253, 265, 295, 296, 300, 317, 349, 369, 372, 373, 380, 485, 511)

p.27 Table 2.3 k

p.49 all variables that either

p.50 1. are related

p.51 other subjects' values

p.53 real predicted values[c].

p.54 Extreme amounts of missing data do

p.60 footnote h distribution.

p.77 than to the number in the "final" model [not 'rather than']

p.80 vice versa [no hyphen]

p.83 identity function [not 'identify function']

p.85 Figure 4.3 units of measurement on x axes? (p.154 Figure 7.3; p.246 Figure 10.10)

p.86 Figure 4.4 smaller symbols (also p.597 Figure 21.1)

p.86 Figure 4.4 caption delete 'lower' twice

p.101 note 18 insert space before 'See [390]'

p.115 not only will quantify overfitting nicely, but also can be used to shrink predicted values

p.121 parsimonious

p.122 note 9 in a sample

p.124 ad hoc [no hyphen]

p.127 centered on

p.128 sentence on function `ns` is awkwardly written

p.128 delete comma after 'keep track of transformation parameters'

p.138 why log(cholesterol +10)

p.140 tertiles delimit tertile bins, classes, or groups (pp.150, 422 quartiles) (p.373 sextiles (better than 6-tiles)) (see Section 6 of Cox (2018) for more) (p.561 is careful on this)

p.142 for 'etal' read 'et al.'

p.143 Stata calls 'tall and thin' a long layout

p.144 line 4 baseline.

p.144 different from

p.152 Table at bottom −1756.95 not -1756.95 (also p.356 −0.68 not -0.68) (also p.375 −2.4 not -2.4)

pp.153, 154, 174, 200, 243, 244, 256, 265, 278, 279, 298, 303, 336, 342, 374, 375, 382 similar points about − not -

p.159 'the treatment effect will the difference in slopes' ?words missing

p.160 is problematic

p.162 etc. spike histograms now discussed at Cox (2021)

p.180 interquartile

p.197 in comparison with (also pp.204, 285, 359, 364, 370, 372, 374, 375)

p.233 $n = 96$

p.245 units mg %? (also pp.248, 250, 252, 253)

p.266 plot the predicted response against the predictor (I thought everyone used the convention that you plot y versus or against x, until I met exceptions: see discussion at https://stats.stackexchange.com/questions/146533/versus-vs-how-to-properly-use-this-word-in-data-analysis) (also p.386)

p.272 last line discussion of

p.273 note 15 Newson [not Newsom]

p.285 Figure 11.5 better square? (also p.301 Figure 12.7; p.356 Figure 14.11; p.469 Figure 19.11)

p.319 'distinct' better than 'unique': see Section 2 of Cox and Longton (2008) for more (also pp.361–363, 483)

p.322 steps (2) and (3)

p.365 been neither diagnosed nor treated

p.395 $|x_2 - 0.5|$

p.412 more nearly normally distributed

p.420 note 7 for 'significantly more efficient' read 'much more efficient'

p.429 last line containing, say,

p.462 some stray text in Figure 19.6 caption

p.489 Figure 20.3 caption refers to solid and dashed lines, but both look solid to me (also p.490 Figure 20.4; p.492 Figure 20.5)

p.536 to use the `rms` package

p.537 Unix is retrospectively what UNIX should have been called

p.539 [6] Agresti 3rd edition 2013

p.539 [12] publisher is SAGE (also p.548 [200] ff.)

p.540 [30] and the Support Investigators

p.541 [36] 33: 517–535

p.543 [90] For the GUSTO-I Investigators

p.544 [103] [109] less use of upper case in paper titles (also p.554 [317]; p.557 [399]; p.567 [623])

p.544 [116] SUPPORT Investigators

p.546 [146] R.B. D'Agostino, Jr

p.546 [153] in R

p.546 [164] A.R.T. Donders (also p.560 [462] [463]; p.568 [628])

p.549 [216] Minnesota

p.552 [277] London

p.552 [284] Epidemiology

p.555 [346] Carpenter

p.556 [375] American Journal of Epidemiology

p.558 [430] Stefanski

p.560 [473] 2: 45–64

p.568 [635] 2002

p.575 Hoeffding D also 81

p.582 variogram also 154

I added index entries

calibration plot

cumulative distribution function

dichotomies, against

dot chart

logit scale

quantile plot

rug plot

spike histogram

Cox, N.J. 2018. Speaking Stata: From rounding to binning. *Stata Journal* 18: 741–754.

https://journals.sagepub.com/doi/pdf/10.1177/1536867X1801800311

Cox, N.J. 2021. Stata tip 141: Adding marginal spike histograms to quantile and cumulative distribution plots. *Stata Journal* 21: 838–846.

https://journals.sagepub.com/doi/pdf/10.1177/1536867X211045583

Cox, N.J. and Longton, G.M. 2008. Speaking Stata: Distinct observations. *Stata Journal* 8: 557–568.

https://journals.sagepub.com/doi/pdf/10.1177/1536867X0800800408

Cox, N.J. and Schechter, C.B. 2019. Speaking Stata: How best to generate indicator or dummy variables. *Stata Journal* 19: 246–259.

https://journals.sagepub.com/doi/pdf/10.1177/1536867X19830921