Summary of Regression Modeling Concepts

Contents

Configuring R

Functions from these packages will be used throughout this document:

```
library(conflicted) # check for conflicting function definitions
# library(printr) # inserts help-file output into markdown output
library(rmarkdown) # Convert R Markdown documents into a variety of formats.
library(pander) # format tables for markdown
library(ggplot2) # graphics
library(ggfortify) # help with graphics
library(dplyr) # manipulate data
library(tibble) # `tibble`s extend `data.frame`s
library(magrittr) # `%>%` and other additional piping tools
library(haven) # import Stata files
library(knitr) # format R output for markdown
library(tidyr) # Tools to help to create tidy data
library(plotly) # interactive graphics
library(dobson) # datasets from Dobson and Barnett 2018
library(parameters) # format model output tables for markdown
library(haven) # import Stata files
library(latex2exp) # use LaTeX in R code (for figures and tables)
library(fs) # filesystem path manipulations
library(survival) # survival analysis
library(survminer) # survival analysis graphics
library(KMsurv) # datasets from Klein and Moeschberger
library(parameters) # format model output tables for
library(webshot2) # convert interactive content to static for pdf
library(forcats) # functions for categorical variables ("factors")
library(stringr) # functions for dealing with strings
library(lubridate) # functions for dealing with dates and times
library(broom) # Summarizes key information about statistical objects in tidy tibbles
library(broom.helpers) # Provides suite of functions to work with regression model 'broom::tidy()' t
```

Here are some R settings I use in this document:

```
rm(list = ls()) # delete any data that's already loaded into R

conflicts_prefer(dplyr::filter)
ggplot2::theme_set(
  ggplot2::theme_bw() +
        # ggplot2::labs(col = "") +
    ggplot2::theme(
      legend.position = "bottom",
      text = ggplot2::element_text(size = 12, family = "serif")))

knitr::opts_chunk$set(message = FALSE)
options('digits' = 6)

panderOptions("big.mark", ",")
pander::panderOptions("table.emphasize.rownames", FALSE)
pander::panderOptions("table.split.table", Inf)
conflicts_prefer(dplyr::filter) # use the `filter()` function from dplyr() by default
legend_text_size = 9
run_graphs = TRUE
```

1 We use different probability models for different data types

- Binary outcomes: Bernoulli models
- Event rate outcomes: Poisson/Negative binomial models
- Time-to-event outcomes: Survival models
- Catch-all: Gaussian models

2 We use different link functions to connect these models with covariates

- Bernoulli models: logit link
- Count models: log link + offset
- Survival models: log link
- Gaussian models: identity link

Figure 1 shows how the various models we have studied have analogous structures: each row uses model-specific transformations to connect the modeled quantity to a linear predictor $\eta(\tilde{x}) = \tilde{x} \cdot \tilde{\beta} + \eta_0$.

Linear (Gaussian):

$$\underset{\text{mean}}{\underset{\sim}{\mu}} \underset{\overset{=}{\leftarrow}}{\overset{=}{\rightarrow}} \underset{\substack{\text{linear}\\ \text{predictor}}}{\underset{\sim}{\eta}}(\tilde{x}) = \tilde{x} \cdot \tilde{\beta} + \mu(0)$$

Logistic (Binomial):

$$\underset{\text{mean}}{\underset{\sim}{\mu}} \underset{\pi \cdot n}{\overset{\frac{\mu}{n}}{\rightarrow}} \underset{\text{probability}}{\underset{\sim}{\pi}} \underset{\frac{\omega}{1+\omega}}{\overset{\frac{\pi}{1-\pi}}{\rightarrow}} \underset{\text{odds}}{\underset{\sim}{\omega}} \underset{\exp\{\eta\}}{\overset{\log\{\omega\}}{\rightarrow}} \underset{\text{log-odds}}{\underset{\sim}{\eta}} = \tilde{x} \cdot \tilde{\beta} + \eta(0)$$

Count (Poisson):

$$\underset{\text{mean}}{\underset{\sim}{\mu}} \underset{\lambda \cdot t}{\overset{\mu/t}{\rightarrow}} \underset{\text{rate}}{\underset{\sim}{\lambda}} \underset{\exp\{\eta\}}{\overset{\log\{\lambda\}}{\rightarrow}} \underset{\text{log-rate}}{\underset{\sim}{\eta}} = \tilde{x} \cdot \tilde{\beta} + \log\{\lambda(0)\}$$

Survival (Cox PH):

$$\underset{\text{mean}}{\underset{\sim}{\mu}} \underset{\int_{t=0}^{\infty} \mathrm{S}(t|\tilde{x})dt}{\leftarrow} \underbrace{\mathrm{S}(t|\tilde{x})}_{\text{survival}} \underset{\exp\{-\Lambda(t|\tilde{x})\}}{\overset{-\log\{\mathrm{S}(t|\tilde{x})\}}{\rightarrow}} \underbrace{\Lambda(t|\tilde{x})}_{\text{cumulative hazard}} \underset{\int_0^t \lambda(u|\tilde{x})\,du}{\overset{\Lambda'(t|\tilde{x})}{\rightarrow}} \underbrace{\lambda(t|\tilde{x})}_{\text{hazard}} \underset{\exp\{\eta(t|\tilde{x})\}}{\overset{\log\{\lambda(t|\tilde{x})\}}{\rightarrow}} \underbrace{\eta(t|\tilde{x})}_{\text{log-hazard}} = \tilde{x} \cdot \tilde{\beta} + \eta_0(t)$$

Figure 1: Parallel structures of GLMs and survival models. Each row shows the transformation chain from the modeled quantity to the linear predictor $\eta(\tilde{x}) = \tilde{x} \cdot \tilde{\beta} + \eta_0$, where η_0 denotes the model-specific intercept term, and $\eta_0(t)$ denotes the Cox PH baseline log-hazard term.

3 We use maximum likelihood estimation to fit models to data

- likelihood
- log-likelihood
- score function
- hessian

4 We use asymptotic normality of MLEs to quantify uncertainty about models

- observed information matrix
- expected information matrix
- standard error
- confidence intervals
- p-values

5 We use (log) likelihood ratios to compare models

Sometimes we adjust these comparisons for model size (AIC, BIC)

References