Generalized linear model

Peter Smits

Regression

> *... a method that summarizes how the average values of a numerical outcome variable vary over subpopulations defined by linear functions of predictors. [...] Regression can be used to predict an outcome given a linear function of these predictors, and regression coefficients can be thought of as comparisons across predicted values or as comparisons among averages in the data.*

Gelman and Hill, 2007, p.31

Linear regression

Written out

N is number of observations. K is number of predictors plus one. y is a length N vector of observations. X is a $N \times K$ matrix of predictors (and a column of 1s). β is a length K vector of regression coefficients (including intercept).

$y \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, $\beta_k \in \mathbb{R}$ for $k = 1, ..., K$.

$$\begin{aligned} y_i &\sim \mathcal{N}(\mu_i, \sigma) \\ \mu_i &= X_i \beta \end{aligned}$$

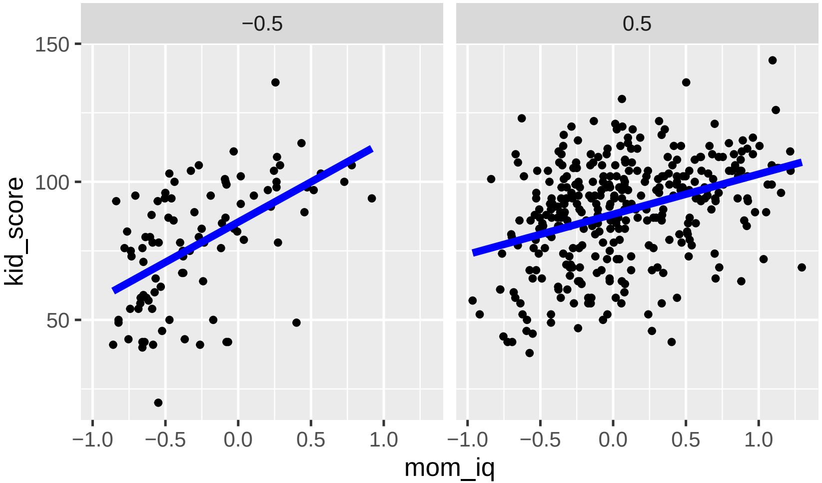
for $i = 1, ..., N$.

Fitting a regression model

```
kid_iq <- read.dta(here::here('ARM_Data', 'child.iq', 'kidiq.dta')) %>%
  as_tibble() %>%
  clean_names()

# feature processing
kid_iq <-
  kid_iq %>%
  dplyr::select(-c(mom_age, mom_work)) %>%
  mutate_at(vars(-kid_score), ~ arm::rescale(., binary.inputs = '-0.5,0.5')) %>%
  mutate(mom_hsXmom_iq = mom_hs * mom_iq)

model_kidiq <- lm(kid_score ~ ., data = kid_iq)

tidy(model_kidiq) %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	86.83	1.21	71.56	0.00
mom_hs	2.84	2.43	1.17	0.24
mom_iq	21.80	2.43	8.96	0.00
mom_hsXmom_iq	-14.53	4.87	-2.99	0.00

Inspecting a regression model



Interpreting regression parameters

- The intercept can only be interpreted assuming zero values for the other predictors.
- If predictors are mean centered, the intercept is the average value of the response when all predictors are at their mean.
- Coeffcient β is the expected difference in y between two observations that differ by 1 in a single predictor.
- σ is standard deviation of dispersion around μ (i.e. $X\beta$).

Key assumptions

In order from most to least important...

1. Validity
2. Additivity and linearity
3. Independence of errors
4. Equal variance of errors
5. Normality of errors

Logistic regression

Written out

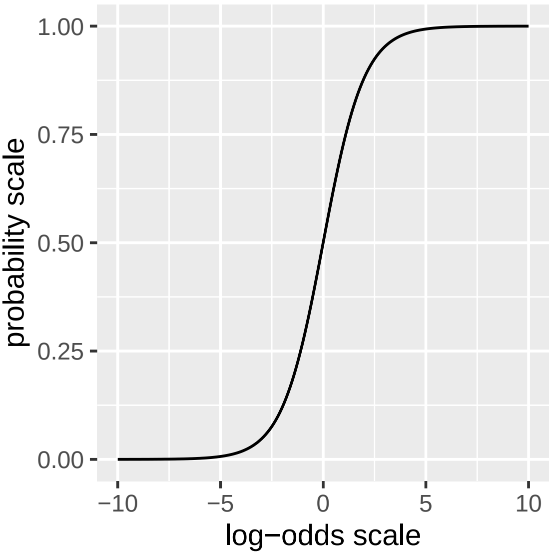
N is number of observations. K is number of predictors plus one. y is a length N vector of observations. X is a $N \times K$ matrix of predictors (and a column of 1s). β is a length K vector of regression coefficients (including intercept).

$y \in 0, 1$, $\theta \in [0, 1]$, $\beta_k \in \mathbb{R}$ for $k = 1, ..., K$, and $\text{logit}(p) = \log(p/1 - p)$ and $\text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$.

$$y_i \sim \text{Bernoulli}(\theta_i)$$
$$\theta_i = \text{logit}^{-1}(X_i\beta)$$

for $i = 1, ..., N$.

Logistic function $\text{logit}(p) = \frac{p}{1-p}$



Fitting a logistic regression model

```r
# issue with row numbers
wells <- read.delim(here::here('ARM_Data', 'arsenic', 'wells.dat'),
                    sep = ' ') %>%
  as_tibble() %>%
  clean_names()

# feature processing
wells <-
  wells %>%
  dplyr::select(-c(assoc, educ)) %>%
  mutate_at(vars(-switch), ~ arm::rescale(., binary.inputs = '-0.5,0.5')) %>%
  mutate(distXarsenic = dist * arsenic)

model_wells <- glm(switch ~ ., data = wells)

tidy(model_wells) %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.58	0.01	64.86	0.00
arsenic	0.21	0.02	11.78	0.00
dist	-0.15	0.02	-8.34	0.00
distXarsenic	-0.04	0.04	-0.97	0.33

Inspecting a logistic regression model

Interpreting logistic regression parameters

- A regression coefficient describes the expected change in the response per unit difference in its predictor.
- However, the logit function introduced into our model creates a nonlinearity makes clear interpretation challenging.

The intercept of logistic regression

- As always, the intercept can only be interpreted assuming zero values for the other predictors.
- If predictors are mean centered, the intercept is the average value of logit(response) when all predictors are at their mean.
- If zero is not interesting, or not even in the model, must be evaluated at some other point.

Logistic regression coefficients near the mean of the data

- A difference of 1 in a predictor corresponds to expected change of β in the logit probability of the response.
- Can evaluate change in response at or near the mean value of predictor x.
  - Difference in $\Pr(y = 1)$ corresponding to adding 1 to x is $\text{logit}^{-1}(\beta\bar{x}) - \text{logit}^{-1}(\beta(\bar{x} + 1))$.
- Or use derivative of logistic curve at central value.
  - Differentiating $\text{logit}^{-1}(\alpha + \beta x)$ wrt x gives $\beta \exp(\alpha + \beta x)/(1 + \exp(\alpha + \beta x))^2$.
  - Calculate $\alpha + \beta x$ for central value.
  - Plug into derivative to give "change" in $\Pr(y = 1)$ per small unit of "change" in x.

Coefficients as odds ratios

- *Odds*: If two outcomes have the probabilities $(p, 1-p)$, then the odds of p is $p/(1-p)$. Odds 1 is equivalent to probability 0.5.
- *Odds ratio*: Ratio of two odds $(p_1/(1-p_1))/(p_2/(1-p_2))$
- Exponentiated logistic regression coefficients (i.e. $\exp(\beta)$) can be interepreted as odds ratios
- Odds are difficult to understand, odds ratios even harder.

The “divide by 4 rule”

The logistic curve is steepest at its center when $X\beta = 0$ so that $\text{logit}^{-1}(X\beta) = 0.5$.

The slope of the curve – the derivative of the logistic function – is maximized at this point and equals $\beta \exp(0)/(1 + \exp(0))^2 = \beta/4$.

Thus, $\beta/4$ is the *maximum difference* in $\Pr(y = 1)$ corresponding to a unit difference in x.

Poisson regression

Written out

N is number of observations. K is number of predictors plus one. y is a length N vector of observations. X is a $N \times K$ matrix of predictors (and a column of 1s). β is a length K vector of regression coefficients (including intercept).

$y \in \mathbb{N}$, $\lambda \in \mathbb{R}^+$, $\beta_k \in \mathbb{R}$ for $k = 1, ..., K$.

$$
\begin{aligned}
y_i &\sim \text{Poisson}(\lambda_i) \\
\lambda_i &= \exp(X_i \beta)
\end{aligned}
$$

for $i = 1, ..., N$.

Ordered categorical response

Written out

Unordered categorical response

Written out

Multiple outcome forms of logistic regression can be coded directly in Stan. For instance, suppose there are K possible outcomes for each output variable y_i. Also suppose that there is a D-dimensional vector x_i of predictors for y_{i}.

If N, $N > 0$, and if $\theta \in \mathbb{R}^N$ forms an N-simplex (i.e. has nonnegative entries summing to one), then for $y \in 1, ..., N$,

$$y_i \sim \text{Categorical}(\theta_i)$$
$$\theta_i = \text{softmax}(x_i \beta)$$

for $i = 1, ..., N$.

FYI, softmax is

$$\text{softmax}(y) \frac{\exp(y)}{\sum_{k=1}^{K} \exp(y_k)}$$

Robust regression