Overview of Generalized Linear Models

Logistic regression is a special case of *a generalized linear model*

Logistic Regression

The logistic function maps an input from the unit range (0,1) to the real line:

$$logit(x) = \log\left(\frac{x}{1-x}\right)$$

More importantly, the inverse-logit function maps a continous variable to the unit range (0,1)

$$logit(x)^{-1} = \frac{\exp(x)}{1+\exp(x)}$$

.

The `qlogis` (for logit) and `plogis` (inverse-logit) functions in R can be used for this calculation. For instance `plogis(1) =` 0.7310586.

Formally, the inverse-logistic function is used as part of the GLM:

$$y \sim Bernoulli \tag{1}$$

$$Pr(y_i = 1) = \pi_i = logit^{-1}(X\underline{\beta}) \tag{2}$$

Note there is not an ϵ term in this model. The randomness comes from the Bernoulli distribution.

Recall the `beer` dataset, but now instead of trying to model consumption, lets consider whether a day is a weekday or weekend.

```
beer <- read_csv('http://math.montana.edu/ahoegh/Data/Brazil_cerveja.csv') %>%
  mutate(consumed = consumed - mean(consumed))
```

```
bayes_logistic <- stan_glm(weekend ~ consumed, data = beer,
                           family = binomial(link = "logit"), refresh = 0)
```

Now how to interpret the model coefficients?

```
bayes_logistic
```

```
## stan_glm
##  family:       binomial [logit]
##  formula:      weekend ~ consumed
##  observations: 365
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) -1.2    0.1
## consumed     0.3    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Interpreting the coefficients can be challenging due to the non-linear relationship between the outcome and the predictors.

Predictive interpretation

One way to interpret the coefficients is in a predictive standpoint. For instance, consider an day with average consumption, then the probability of a weekend would be `invlogit(-1.2 + 0.3 * 0)` = 0.23, where as the probability of a day with 10 more liters of consumption (relative to an average day) would have a weekend probability of `invlogit(-1.2 + 0.3 * 10)` = 0.86

Of course, we should always think about uncertainty, so we can extract simulations from the model.

`posterior_linpred` was useful with regression, but need `posterior_epred` here

```
new_data <- data.frame(consumed = c(0,10))
posterior_sims <- posterior_epred(bayes_logistic, newdata = new_data)
summary(posterior_sims)
```

```
##        1                2
##  Min.   :0.1445   Min.   :0.6897
##  1st Qu.:0.2073   1st Qu.:0.8473
##  Median :0.2245   Median :0.8726
##  Mean   :0.2249   Mean   :0.8691
##  3rd Qu.:0.2425   3rd Qu.:0.8960
##  Max.   :0.3267   Max.   :0.9579
```

It can also be useful to consider predictions of an individual data point. *This is how you would conduct posterior predictive checks.*

```
new_obs <- posterior_predict(bayes_logistic, newdata = new_data)
head(new_obs)
```

```
##      1 2
## [1,] 0 0
## [2,] 0 1
## [3,] 0 1
## [4,] 0 0
## [5,] 0 1
## [6,] 0 1
```

```
colMeans(new_obs)
```

```
##       1       2
## 0.22975 0.85925
```

odds ratios and log odds

logistic regression can be re-written as

$$
\begin{aligned}
y &\sim Bernoulli && (3)\\
\log\left(\frac{Pr[y=1|X]}{Pr[y=0|X]}\right) &= \beta_0 + \beta_1 x && (4)\\
\log\left(\frac{Pr[y=1|X]}{1-Pr[y=1|X]}\right) &= \beta_0 + \beta_1 x && (5)\\
& && (6)
\end{aligned}
$$

Thus, a one unit change in x increases the log odds of y by a factor of β_1

Furthermore, logistic regression can also re-written as

$$
\begin{aligned}
y &\sim Bernoulli && (7)\\
\log\left(\frac{Pr[y=1|X]}{Pr[y=0|X]}\right) &= \beta_0 + \beta_1 x && (8)\\
\frac{Pr[y=1|X]}{1-Pr[y=1|X]} &= \exp\left(\beta_0 + \beta_1 x\right) && (9)\\
& && (10)
\end{aligned}
$$

Then consider $\exp \beta_1$

$$
\begin{aligned}
\exp(\beta_1) &= \frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1(x))} && (11)\\
&= \frac{Pr[y=1|X=x+1]/Pr[y=0|X=x+1]}{Pr[y=1|X=x]/Pr[y=0|X=x]} && (12)
\end{aligned}
$$

hence, this can be interpreted as an odds ratio

Interpretation of log odds and odds ratios can be difficult; however, interpreting the impact on probabilities requires setting other parameter values and the change is non-linear (different change in probability for a one unit change in a predictor).

Model Comparison

We can use cross validation in the same manner a standard linear models.

```r
loo(bayes_logistic)
```

```
##
## Computed from 4000 by 365 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo   -168.8 10.4
## p_loo         1.9  0.2
## looic       337.6 20.8
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```r
temp_model <- stan_glm(weekend~max_tmp, data = beer, refresh=0)
loo(temp_model)
```

```
##
## Computed from 4000 by 365 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo   -230.0  9.2
## p_loo         2.4  0.2
## looic       460.0 18.3
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```