CH 14: More Logistic Regression

Odds Ratios

If there are two outcomes, with probabilities p and $1-p$, then $\frac{p}{1-p}$ is called odds.

If the two probabilities are equal then the odds would be $\frac{1/2}{1/2} = 1$. *If the odds are 2 (or 1/2), this corresponds to p = 2/3 and q = 1/3.*

An odds ratio is the result of dividing two odds:

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

.

an odds ratio of two corresponds to a change in odds, rather than a change in probabilities associated with events 1 and 2.

logistic regression can be re-written as

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

$$\log\left(\frac{Pr[y=1|X]}{Pr[y=0|X]}\right) = \beta_0 + \beta_1 x \tag{2}$$

$$\log\left(\frac{Pr[y=1|X]}{1 - Pr[y=1|X]}\right) = \beta_0 + \beta_1 x \tag{3}$$

$$\tag{4}$$

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

$$
\begin{align}
y &\sim Bernoulli \tag{5} \\
\log\left(\frac{Pr[y=1|X]}{Pr[y=0|X]}\right) &= \beta_0 + \beta_1 x \tag{6} \\
\frac{Pr[y=1|X]}{1 - Pr[y=1|X]} &= \exp\left(\beta_0 + \beta_1 x\right) \tag{7} \\
& \tag{8}
\end{align}
$$

Then consider $\exp \beta_1$

$$
\begin{align}
\exp(\beta_1) &= \frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1(x))} \tag{9} \\
&= \frac{Pr[y=1|X=x+1]/Pr[y=0|X=x+1]}{Pr[y=1|X=x]/Pr[y=0|X=x]} \tag{10}
\end{align}
$$

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).

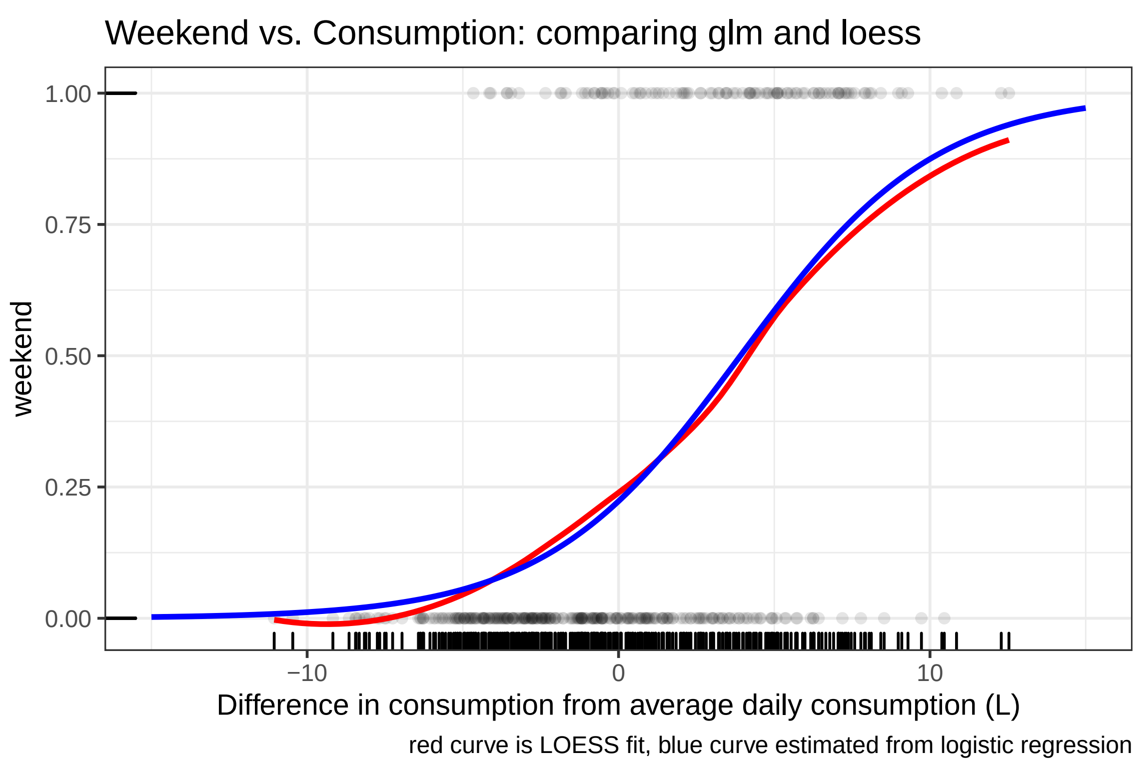
Data visualization

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

```
## Parsed with column specification:
## cols(
##   consumed = col_double(),
##   precip = col_double(),
##   max_tmp = col_double(),
##   weekend = col_double()
## )
```

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

beer %>% ggplot(aes(y = weekend, x = consumed)) +
  geom_point(alpha = .1) +
  geom_smooth(formula = 'y~x', method = 'loess', color = 'red', se = F) +
  geom_rug() + ggtitle('Weekend vs. Consumption: comparing glm and loess') +
  theme_bw() + xlab('Difference in consumption from average daily consumption (L)') +
  geom_line(inherit.aes = F, data = tibble(temp = seq(-15,15, by = .1),
          y = plogis(coef(bayes_logistic)['(Intercept)'] + coef(bayes_logistic)['consumed']*temp)),
           aes(x=temp, y=y), color = 'blue',lwd = 1) +
  labs(caption = 'red curve is LOESS fit, blue curve estimated from logistic regression')
```



Model interpretation

```
bayes_logistic
```

```
## stan_glm
##  family:       binomial [logit]
##  formula:      weekend ~ consumed
##  observations: 365
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) -1.2    0.2
## consumed     0.3    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

- (Intercept): *we can interpret this term with all other predictors constant - another good reason to standardize variables. Hence with an average daily consumption, the probability of the day being a weekend is logit*$^{-1}(-1.3) =$ *0.21. (with minimal uncertainty)*

- consumed: *for each additional unit of consumption, the the log-odds of being a weekend increase by about 0.3 or the odds ratio of being a weekend increases by about* $\exp(0.3) =$ *1.35 or the probability of a weekend increases from 0.21 to 0.27 if consumption increases from 0 to 1. (with minimal uncertainty)*

The last interpretation of the consumed, suggests that scaling variables can also be useful. Then you can state as consumed goes from 0 (the average) to 1 (one standard deviation greater than average) the probability of being a weekend increases from – to –.

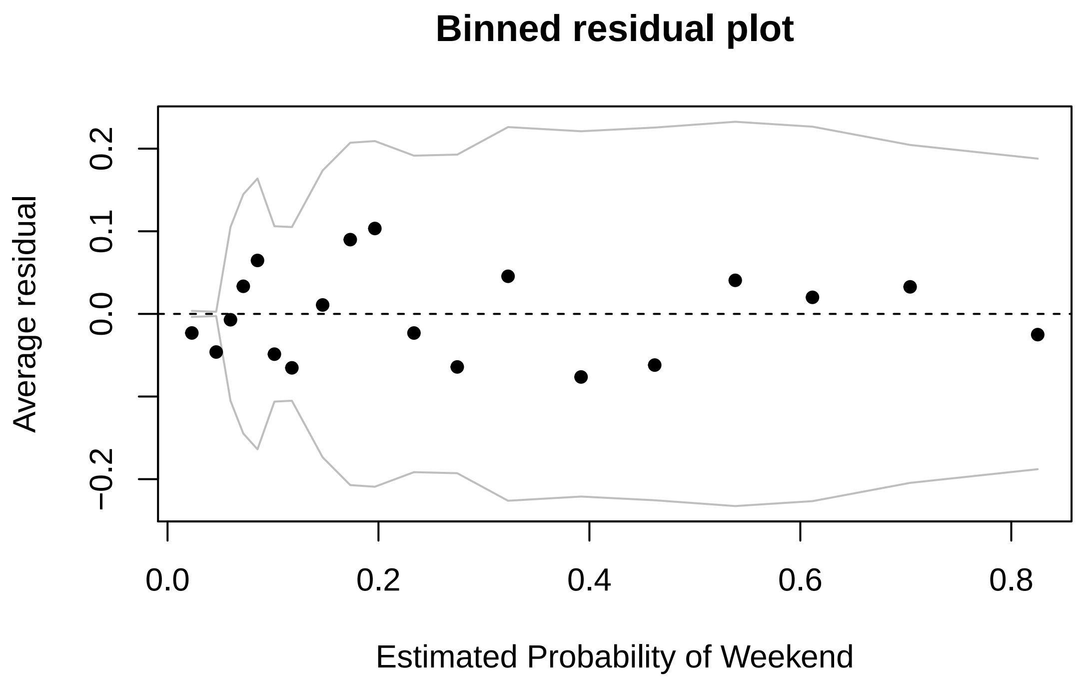
Residuals

Just as with standard regression models, *which by the way are a special case of glms, we can use residuals and posterior predictive distributions to evaluate model fit.*

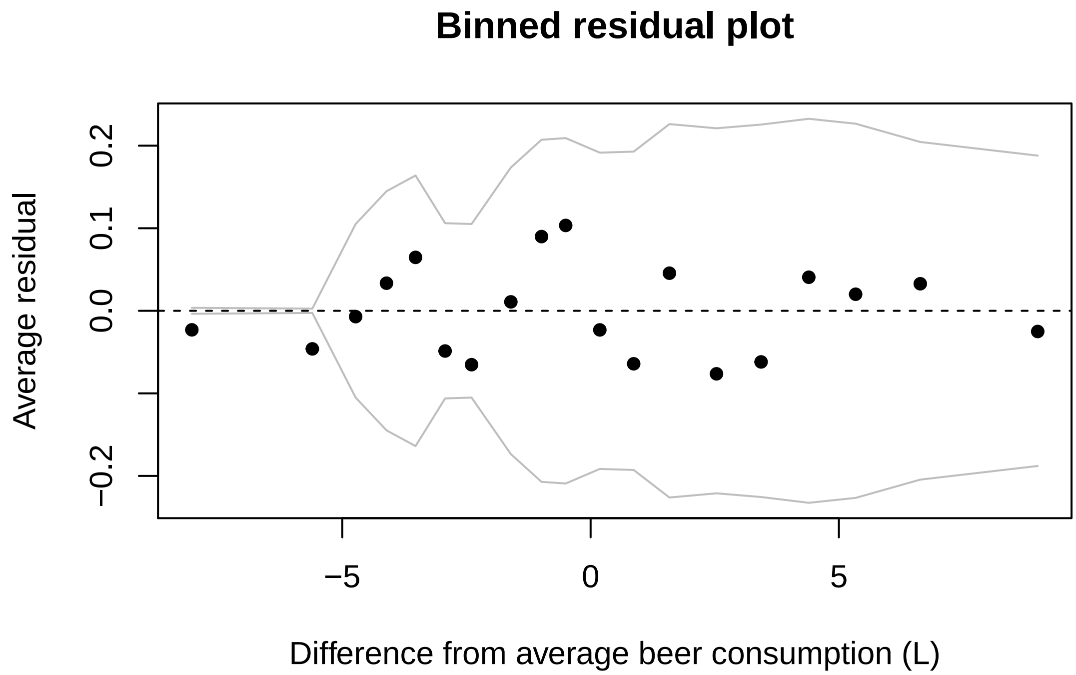
We can define a residual to be

$$
\begin{aligned}
r_i &= y_i - Exp[y_i|X_i] && (11)\\
&= y_i - logit^{-1}(X_i\beta_i) && (12)\\
&= \pi_i && (13)
\end{aligned}
$$

```r
binnedplot(predict(bayes_logistic,type = 'response'),resid(bayes_logistic),
           xlab = 'Estimated Probability of Weekend')
```




```r
binnedplot(beer$consumed,resid(bayes_logistic),
           xlab = 'Difference from average beer consumption (L)')
```



Binned residual plot
Average residual
0.2
0.1
0.0
−0.2
−5
0
5
Difference from average beer consumption (L)