2019R1 Discrete Data Analysis (STAT5107) Assignment 3

Yiu Chung WONG 1155017920

```
set.seed(5107);
```

2.

$$log(\pi) = \alpha + \boldsymbol{\beta} X + \epsilon$$

* For binary predictor, this is Relative Risk regression. * The coefficient is the log relative risk. * It has a log link function for the binomial (or Bernoulli) outcome. * The log-binomial regression does not respect the natural parameter constraints; * It does not ensure that predicted probabilities are mapped to the [0,1] range. * e.g. for predictors that take a positive value, the resulting π would be greater than 1.

3.

For

$$f(y; k, p) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} p^k (1-p)^y \qquad \text{for } y = 0, 1, 2, \ldots$$

Then it can be rewritten in exponential form as:

$$\begin{aligned}
f(y; k, p) &= \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \exp\left[\ln(p^k (1-p)^y)\right] \\
&= \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \exp\left[k \ln(p) + y \ln(1-p)\right] \\
&= \exp\left[k \ln(p)\right] \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \exp\left[y \ln(1-p)\right]
\end{aligned}$$

where

$$\begin{aligned}
a(p) &= exp\left[k \ln(p)\right] \\
b(y) &= \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \\
Q(p) &= \ln(1-p)
\end{aligned}$$

4a.

```
x_i <- 10000;
pi_i <- -.0003 + .0304*x_i;
P <- 100*pi_i/x_i;
```

The estimated proportion vote for Buchanan in 2000 was roughly 3.039997% of that for for Perot in 1996.

4b.

```
pi_i_real <- .0079;
x_i <- .0774;
pi_i_predict <- -.0003 + .0304*x_i;
```

- The outcome is 3.8481023 times than what the linear relation would have predicted. This suggests anonymity.

4c.

```
pi_logit = (1 + exp(-(-7.164 + 12.219*x_i)))^-1
```

- π_i is 0.0019888.
- The outcome is 3.9723094 times than what the logistic regression would have predicted. This suggestes that it is an outlier.

5.

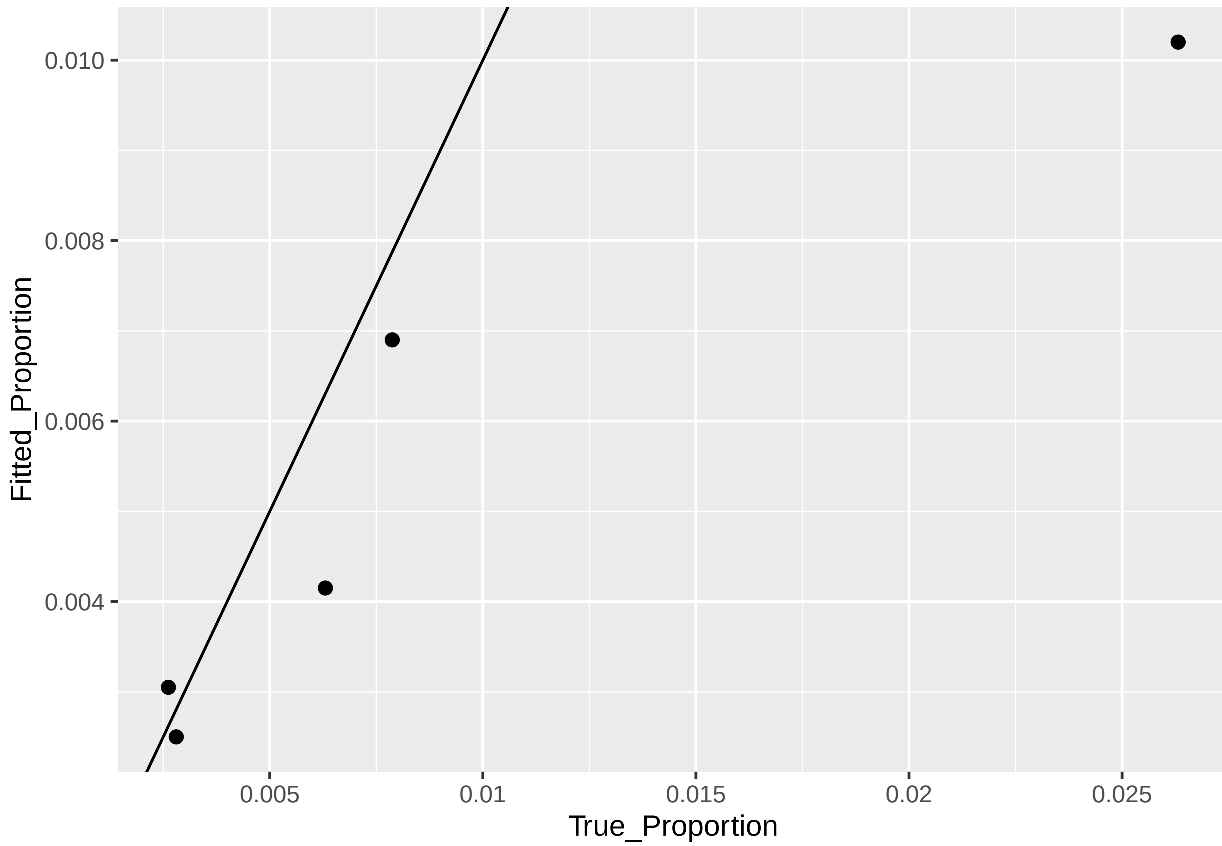
```
table <- matrix(c(17066, 48, 14464, 38, 788, 5, 126, 1, 37, 1), 2);
p_table <- prop.table(table, margin = 2);

scores <- c(0, .5, 1.5, 4.0, 7.0);
alpha <- .0025;
beta <- .0011;
linear_relation <- function(x, a = alpha, b = beta){a + b*x};
predicted_p <- linear_relation(scores);

fitting <- matrix(c(p_table[2,], predicted_p),
                  byrow = TRUE,
                  nrow = 2, dimnames = list(
                    c("True_Proportion", "Fitted_Proportion"),
                    c("0", "<1", "1-2", "3-5", ">=6")));
```

- For non-drinker ($x = 0$), $\hat{\pi}(x) = .0025$
- For every step increase in drinking level, the probability increase by 0.0011.
- The CI for the coefficient incluses 0; there is a chance that alcohol has no effect on congenital sex organ malformations.

```
t_fitting = as.data.frame(t(fitting))
ggplot(data = t_fitting, aes(x = True_Proportion, y = Fitted_Proportion)) +
  geom_point(size = 2) +
  geom_abline() +
  scale_x_continuous(breaks = seq(0, .03, by = .005), labels = seq(0, .03, by = .005));
```



* This is a poor fitting.

```
get_relative_risk <- function(x, a){
  array = x/x[a];
  array[a:length(array)];
  };
relative_risk <- get_relative_risk(predicted_p, 1)
```

- Relative Risks compared to no drinking are: 1, 1.22, 1.66, 2.76, 4.08

6a.

```
alpha <- -.4284;
beta <- .5893;
crabs_model <- function(x) alpha + beta * x;
weight <-2.44;
expected_Y <- exp(crabs_model(weight));
```

- On average, a 2.44 kg female crab has 2.7442066 satellites.

6b.

```
se <- .0650;
log_CI <- qnorm(c(.025, .975), mean = beta, sd = se);
CI <- exp(log_CI);
```

- On average, for every kg of weight increase, the number of satellites increase by 1.8027261.

6c.

```
df <- 171;
cutoff <- qchisq(p = .05, df = df, lower.tail = FALSE);
w <- (beta - 0)^ 2 / se^2;
```

- H_0 = coefficient not sigificantly different from zero.
- A chi-square distribution with 171 degrees of freedom has a cutoff value at 202.5125774 ($\alpha = .05$); the Wald test yields a test statistic at 82.1951456. Hence there is not enough evidence to reject H_0.