STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 7

1. In the following situations, identify whether the described random variable plausibly has a Bernoulli, Binomial, Poisson, Normal or some other type of probability distribution. Briefly justify your answer.

   (a) Let Y be the number of iPhone owners out of a random sample of 100 that download an app today.

   Solution: Binomial. This is a discrete variable and can be viewed as counting the number of successes/failures in 100 independent trials.

   (b) Let Y be the number of babies born on a single day in Canberra.

   Solution: Poisson. This is a discrete variable counting the number of events in a fixed interval. It is reasonable to think that these events occur independently of the time since the last event.

   (c) Let Y be the initial weight (in kilograms) of a randomly selected male enrolling in the Biggest Loser diet.

   Solution: Normal. Reasonable to think that weights are symmetrically distributed around a mean and more likely to be close to that mean than far from it. Distribution could be skewed as well.

   (d) Let Y be the initial weight (in kilograms) of a randomly selected person on the Biggest Loser diet.

   Solution: Other. Would expect two modes, one for men and one for women.

   (e) Let Y be 1 if a randomly selected person who saw the movie The Wolverine liked it, and 0 if not.

   Solution: Bernoulli. Y is either 0 or 1.

   (f) Let Y be 1 if you flip a coin and it lands tails, and 2 if it lands heads.

   Solution: Other. Y follows a Bernoulli random variable plus a constant of 1.

2. This question concerns the choice of link functions.

   (a) Suppose a probability distribution has an unknown mean μ that is restricted to be greater than 1. Why is $g(\mu) = \log(\mu)$ NOT a sensible link function? What would be a more reasonable link function?

$f(-1,1) \Rightarrow (0,1)$

$\overset{+1}{(0,2)} \times \frac{1}{2} \Rightarrow (0,1)$

Solution: If μ is always greater than 1, then $g(\mu) = \log(\mu) > 0$. This is not sensible because the range of $x\beta$ is the entire real line and we would like the ranges of $g(\mu)$ and $x\beta$ to match up. A more reasonable link function would be $\log(\mu - 1)$.

(b) Suppose that a probability distribution has an unknown mean μ that is restricted to be between -1 and 1. What would be a reasonable link function for μ?

Solution: A reasonable link function would be $g(\mu) = \text{logit}\left(\frac{\mu+1}{2}\right)$ $\Rightarrow (0,1)$

3. The data file `Heart.txt` is located on Wattle and contains data for 99 individuals ordered in increasing age groups. For each age group (`age`), the total number of individuals at that age (`ssize`) and the number of subjects at that age who have symptoms of heart disease (`disease`) are given.

(a) Fit an unweighted least-squares empirical logit model to this data. In other words, fit the model

$$\log\left(\frac{\texttt{disease}/\texttt{ssize}}{1 - \texttt{disease}/\texttt{ssize}}\right) = \beta_0 + \beta_1\texttt{age} + \epsilon,$$

where ϵ is assumed to be normal with mean zero and constant variance. Note that, since many of the observed proportions are either zero or one, you will need to modify them to, say, 0.05 or 0.95, respectively, before taking the logit transformation. Also, investigate how important the choices of these arbitrary values are by re-fitting the regression after having set the zero and one values to 0.005 and 0.995.

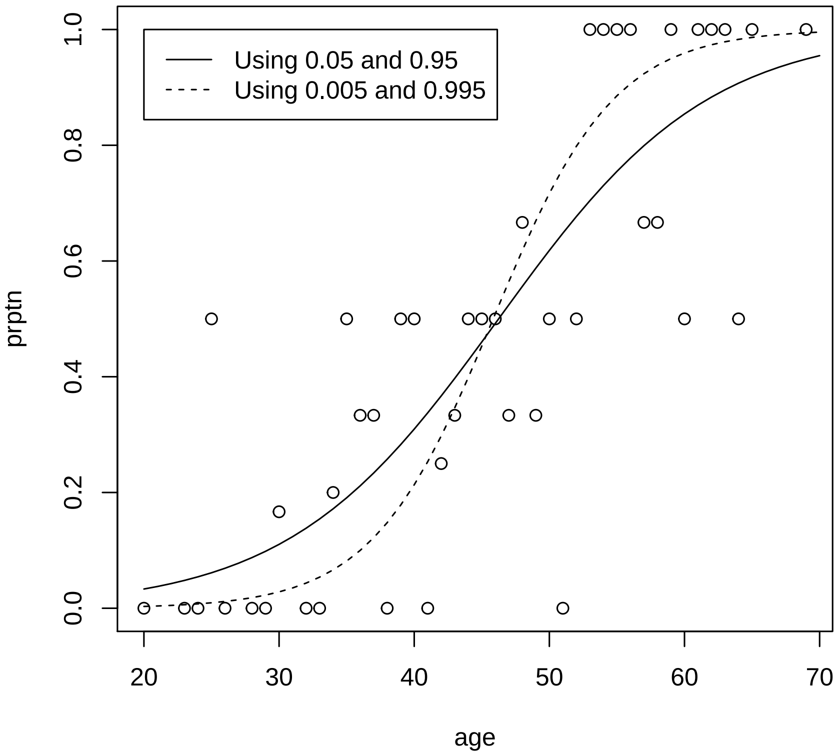
Solution: The required R code is:

```
> hrt <- read.table("Heart.txt", header=TRUE)
> attach(hrt)
> names(hrt)

[1] "age"     "ssize"   "disease"

> prptn <- disease/ssize
> nwprp <- ifelse(prptn==0, 0.05, prptn)
> nwprp <- ifelse(nwprp==1, 0.95, nwprp)
> lgtdse <- log(nwprp/(1-nwprp))
> hrt.reg1 <- lm(lgtdse~age)
> coef(hrt.reg1)

(Intercept)         age
 -5.9408796   0.1284622

> nwprp2 <- ifelse(prptn==0, 0.005, prptn)
> nwprp2 <- ifelse(nwprp2==1, 0.995, nwprp2)
> lgtdse2 <- log(nwprp2/(1-nwprp2))
> hrt.reg2 <- lm(lgtdse2~age)
> coef(hrt.reg2)
```

```
(Intercept)          age
-10.2403737    0.2233812
```

So, the arbitrary values we choose for the zero and one proportions seem to have a noticable effect on the parameter estimates. In addition, a plot of the two fitted curves (not necessary to answer the question) also shows the dramatic difference, the required code is below:

```
> plot(age, prptn)
> ages <- 20:70
> invlogit <- function(x) exp(x)/(1+exp(x))
> lines(ages,invlogit(coef(hrt.reg1)[1]+coef(hrt.reg1)[2]*ages), lty=1)
> lines(ages,invlogit(coef(hrt.reg2)[1]+coef(hrt.reg2)[2]*ages), lty=2)
> legend(20, 1, lty=1:2, c("Using 0.05 and 0.95", "Using 0.005 and 0.995"))
```



(b) The Delta method estimate of variance for $g(Y)$ is given by:

$$\mathbb{V}g(Y) \doteq \left[g'(\mathbb{E}Y)\right]^2 \mathbb{V}Y.$$

Use this fact, and the knowledge that Y is binomially distributed, to find the variance of the empirical logit transformed proportions. In addition, use these weights to fit a weighted least-squares empirical logit to the data. In other words, fit the models in (a), but use weighted least-squares with the appropriate weights. [NOTE: You will have to use the iterative scheme outlined in class.]

Solution: We know that $\mathbb{V}Y = \frac{1}{n}\mu(1-\mu)$, where $\mu = \mathbb{E}Y$ and is the expected proportion of n people with symptoms of heart disease and simple differentiation shows that:

$$g'(\mu) = \frac{d}{d\mu}\ln\left(\frac{\mu}{1-\mu}\right) = \frac{1}{\mu(1-\mu)}.$$

So, $\mathbb{V}g(Y) \doteq 1/[n\mu(1-\mu)]$ and therefore we need to use weights of the form:

$$w_i^2 = \frac{1}{\mathbb{V}g(Y)} \doteq n\mu(1-\mu) \doteq n\widehat{Y}(1-\widehat{Y}).$$

The R commands to run the required iterative scheme are:

```
> ft1 <- exp(fitted(hrt.reg1))/(1+exp(fitted(hrt.reg1)))
> wgt <- ssize*ft1*(1-ft1)
> hrt.reg1 <- lm(lgtdse~age, weights=wgt)
> coef(hrt.reg1)

(Intercept)          age
 -5.9199644    0.1266409
```

Repeat until coefficient estimates stop changing.

```
> for(i in 1:100){
+   ft1 <- exp(fitted(hrt.reg1))/(1+exp(fitted(hrt.reg1)))
+   wgt <- ssize*ft1*(1-ft1)
+   hrt.reg1 <- lm(lgtdse~age, weights=wgt)
+ }
> coef(hrt.reg1)

(Intercept)          age
 -5.9320970    0.1269402
```

Using the `lgtdse2` (i.e., the logit transforms based on the modified proportions using 0.005 and 0.995 instead of 0.05 and 0.95), the weighted estimates are $b_{w,0} = -9.757401$ and $b_{w,1} = 0.2099463$, which are still substantially different.

(c) Fit a logistic regression to the data. Compare your parameter estimates to those from (a) and (b).

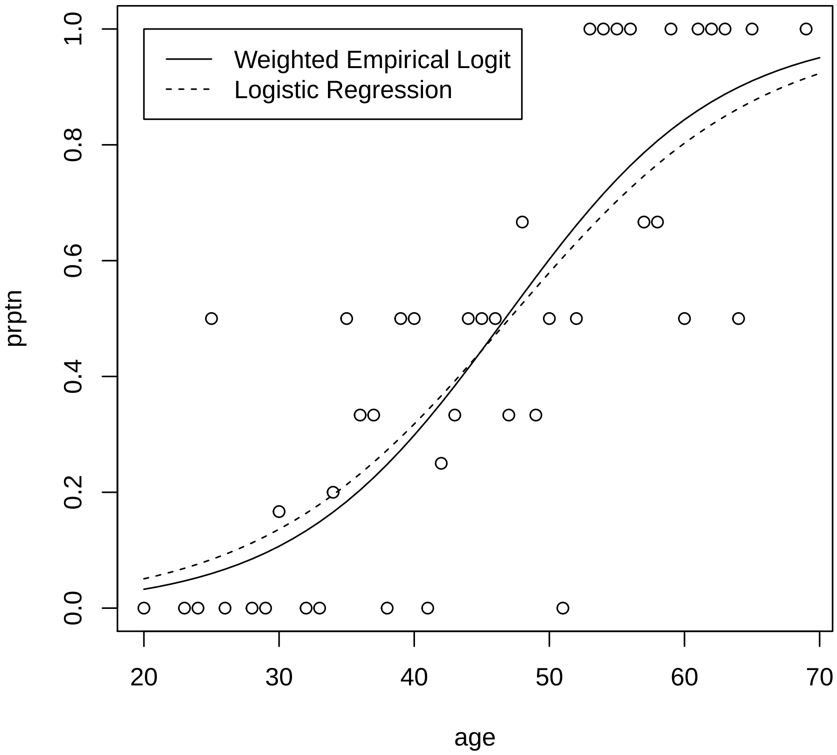
Solution: The required *R* commands are:

```
> hrt.glm <- glm(prptn~age, family=binomial, weights=ssize)
> coef(hrt.glm)
```

```
(Intercept)         age
 -5.0992974   0.1083923
```

Note that these values are reasonable close to those for the empirical logit regressions using the "corrected" values of 0.05 and 0.95. However, we did not have to worry at all about fixing this problem, as it was done automatically within the *IRLS* algorithm of `glm`, (and all such fixes can be shown to lead to the same parameter estimates for the IRLS algorithm). A plot of the two fitted curves (not necessary to answer the question) shows the similarity:

```
> plot(age, prptn)
> ages <- 20:70
> invlogit <- function(x) exp(x)/(1+exp(x))
> lines(ages, invlogit(coef(hrt.reg1)[1]+coef(hrt.reg1)[2]*ages), lty=1)
> lines(ages, invlogit(coef(hrt.glm)[1]+coef(hrt.glm)[2]*ages), lty=2)
> legend(20, 1, lty=1:2, c("Weighted Empirical Logit", "Logistic Regression"))
```



$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0+\beta_1 age$$

$$\frac{\pi}{1-\pi} = exp(\beta_0+\beta_1 age)$$

(d) Using the results of the model in (c), by what factor has your odds of having symptoms of heart disease increased once you are 10 years older? Does your answer to this question depend on what age you currently are? [HINT: Recall that your odds of having symptoms at a particular age is just $\pi(\texttt{age})/[1-\pi(\texttt{age})]$.]

Solution: We know that our logistic model says that:

$$\begin{aligned} odds(\texttt{age}) &= \frac{\pi(\texttt{age})}{1-\pi(\texttt{age})} = \exp(\beta_0+\beta_1\texttt{age}) \\ odds(\texttt{age}+10) &= \frac{\pi(\texttt{age}+10)}{1-\pi(\texttt{age}+10)} = \exp[\beta_0+\beta_1(\texttt{age}+10)]. \end{aligned}$$

Therefore, the factor by which your odds increases is just

$$\frac{odds(\texttt{age}+10)}{odds(\texttt{age})} = \exp[\beta_0+\beta_1(\texttt{age}+10)-\beta_0-\beta_1\texttt{age}] = \exp(10\beta_1).$$

So, our estimate of the factor of increase is $\exp(10\widehat{\beta}_1) = 2.956$. Note that this does not depend on the initial value of `age`, so the model states that every 10 years you age increases your odds of heart disease symptoms by a factor of about 3. [NOTE: We should ask ourselves if this structural feature of the model makes sense for the actual situation at hand, and if not, then we should consider changing our model.]

4. The data for the anaesthetic depth example shown in lectures is stored in the file `AnsthcSum.txt` on Wattle. In Tutorial 5, we saw that the estimate for the 50%-response concentration, $x_{0.5}$, was given by

$$x_{0.5} = \frac{g(0.5)-\widehat{\beta}_0}{\widehat{\beta}_1},$$

where g was either the logistic, probit or complementary log-log link function.

(a) Calculate a 95% confidence interval for this quantity for each of these three link functions. [NOTE: Recall that for binomial data we assume that $\phi = 1$, since we have included the factors $1/n_i$ into the weights.]

Solution: The required confidence intervals are for the function of the parameters:

$$h(\beta) = \frac{g(0.5)-\beta_0}{\beta_1},$$

which means that

$$\frac{\partial h(\beta)}{\partial \beta} = \begin{pmatrix} -\frac{1}{\beta_1} \\ \frac{\beta_0-g(0.5)}{\beta_1^2} \end{pmatrix}.$$

So, to find the confidence intervals:

```
> ansth <- read.table("AnsthcSum.txt", header=TRUE)
> ansth.prbt <- glm(prptn~conc, family=binomial(link=probit),
+                    weights=ssize, data=ansth)
> est <- -as.vector(coef(ansth.prbt))[1]/as.vector(coef(ansth.prbt))[2]
> dh <- c(-1/as.vector(coef(ansth.prbt))[2],
+         as.vector(coef(ansth.prbt))[1]/as.vector(coef(ansth.prbt)[2]^2))
> sd <- sqrt(t(dh)%*%summary(ansth.prbt)$cov.unscaled%*%dh)
> upper <- est + qt(0.975,4)*sd
> lower <- est - qt(0.975,4)*sd
> c(lower, est, upper)
```

```
[1] 0.9323877 1.1604613 1.3885350
```

```
> ansth.lgt <- glm(prptn~conc, family=binomial(link=logit),
+                   weights=ssize, data=ansth)
> est <- -as.vector(coef(ansth.lgt))[1]/as.vector(coef(ansth.lgt))[2]
> dh <- c(-1/as.vector(coef(ansth.lgt))[2],
+         as.vector(coef(ansth.lgt))[1]/as.vector(coef(ansth.lgt)[2]^2))
> sd <- sqrt(t(dh)%*%summary(ansth.lgt)$cov.unscaled%*%dh)
> upper <- est + qt(0.975,4)*sd
> lower <- est - qt(0.975,4)*sd
> c(lower, est, upper)
```

```
[1] 0.9276443 1.1620175 1.3963908
```

```
> ansth.cll <- glm(prptn~conc,family=binomial(link=cloglog),
+                   weights=ssize, data=ansth)
> g05 <- log(-log(1-0.5))
> est <- (g05-as.vector(coef(ansth.cll))[1])/as.vector(coef(ansth.cll))[2]
> dh <- c(-1/as.vector(coef(ansth.cll))[2],
+         (as.vector(coef(ansth.cll))[1]-g05)/as.vector(coef(ansth.cll)[2]^2))
> sd <- sqrt(t(dh)%*%summary(ansth.cll)$cov.unscaled%*%dh)
> upper <- est + qt(0.975,4)*sd
> lower <- est - qt(0.975,4)*sd
> c(lower, est, upper)
```

```
[1] 0.8870842 1.1267824 1.3664806
```

(b) Now, calculate 95% confidence intervals for the parameters β_0 and β_1.

Solution: The required *R* commands:

```
> sds <- sqrt(diag(summary(ansth.prbt)$cov.unscaled))
> upper <- coef(ansth.prbt)+(qt(0.975,4)*sds)
> lower <- coef(ansth.prbt)-(qt(0.975,4)*sds)
> cbind(lower, upper)
```

```
                lower      upper
(Intercept)  0.2012634  7.5146070
conc        -6.4123182 -0.2366499
```

```
> sds <- sqrt(diag(summary(ansth.lgt)$cov.unscaled))
> upper <- coef(ansth.lgt)+(qt(0.975,4)*sds)
> lower <- coef(ansth.lgt)-(qt(0.975,4)*sds)
> cbind(lower, upper)
```

```
                 lower      upper
(Intercept)  -0.2462593 13.1836086
conc        -11.2408417  0.1073182
```

```
> sds <- sqrt(diag(summary(ansth.cll)$cov.unscaled))
> upper <- coef(ansth.cll)+(qt(0.975,4)*sds)
> lower <- coef(ansth.cll)-(qt(0.975,4)*sds)
> cbind(lower, upper)
```

```
                 lower      upper
(Intercept)  0.03931748  7.4238661
conc        -7.08938672 -0.1846083
```

(c) Using the confidence intervals from (b), we might suggest an alternative confidence interval for the 50%-response concentration be constructed from the largest and smallest values that the 50%-response concentration can take for parameter values within the intervals of (b). Use the logistic link model to investigate this alternative method. Do you think that this alternative method will generally provide reasonable intervals?

Solution: For the logistic link model, the intervals for β_0 and β_1 both contain 0, meaning that the largest value for $-\beta_0/\beta_1$ is positively infinite and the smallest value of this quantity is negatively infinite! Clearly this seemingly logical approach does not work at all in this case, and indeed it will tend to do very poorly in most cases. There is a well-documented geometrical reason for this, but it is based on some higher level mathematics and so we will not delve into it here.