STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 6

1. (Ramsey and Schafer, 2013, Chap 20, Ex 11). The file `oring.csv` contains data on the launch temperatures (degrees Farenheit) and an indicator of O-ring failure for 24 space shuttle launches prior to the space shuttle `Challenger` disaster of Jan 27, 1986. The night before the disaster, scientists warned the shuttle should not be launched because of predicted cold weather. Fuel seal problems, which had been encountered in earlier flights, were suspected of being associated with low temperatures.

   (a) Fit the logistic regression model of `Failure` (1 for failure) on `Temperature`. Report the estimated coefficients and their standard errors.

   Solution: The following commands can do the job.

```
> oring <- read.csv("oring.csv", header=T)
> attach(oring)
> names(oring)

[1] "TEMP"    "FAILURE"

> m1 <- glm(FAILURE ~ TEMP, family = binomial(link="logit"))
> summary(m1)

Call:
glm(formula = FAILURE ~ TEMP, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2125  -0.8253  -0.4706   0.5907   2.0512

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.87535    5.70291   1.907   0.0565 .
TEMP        -0.17132    0.08344  -2.053   0.0400 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 28.975  on 23  degrees of freedom
Residual deviance: 23.030  on 22  degrees of freedom
AIC: 27.03

Number of Fisher Scoring iterations: 4
```

The fitted logistic regression model for the probability p, of failure given temperature is:

$$logit(p) = 10.8753 - 0.1713 temp$$

with the estimated standard errors being 5.70 and 0.083 for the intercept and the slope, respectively.

(b) Test whether the coefficient of `Temperature` is 0, using Wald's test. Report a one-sided p-value (the alternative hypothesis is that the coefficient is negative; odds of failure decrease with increasing temperature).

Solution: The following commands can do the job.

```
> pnorm(-0.17132/0.08344)

[1] 0.02002602
```

One-sided p-value = 0.0200. Reject null hypothesis that coefficient of temperature is 0.

(c) Test whether the coefficient of `Temperature` is 0, using the drop in deviance test.

Solution: The following commands can do the job.

```
> pchisq(summary(m1)$null.deviance -summary(m1)$deviance,
+        summary(m1)$df.null-summary(m1)$df.residual, lower.tail=FALSE)

[1] 0.01476632
```

Drop in deviance = 5.9441 with 1 d.f. gives p-value = 0.0148. This is a two-sided p-value for the coefficient. We may want to use a one-sided alternative hypothesis (i.e., $H_A : \beta_{Temperature} < 0$). The corresponding one-sided p-value is approximately 0.0074. (Chi-square distribution is not symmetric.)

(d) Give a 95% confidence interval for the coefficient of `Temperature`.

Solution: The following commands can do the job.

```
> c(-0.17132-1.96*0.08344, -0.17132+1.96*0.08344)

[1] -0.3348624 -0.0077776
```

The 95% confidence interval extends from -0.03348 to -0.0078.

(e) What is the estimated logit of failure probability at 31°F? What is the estimated probability of failure?

Solution: The following commands can do the job.

```
> pred.31 <- as.numeric(c(1,31) %*% coef(m1))
> pred.31

[1]  5.564414

> invlogit <- function(x){1/(1+exp(-x))}
> invlogit(pred.31)

[1]  0.9961828

> # use predict function in R to get predicted value
> # and std. error of prediction
> newdata = data.frame(TEMP = 31)
> # prediction on probability of FAILURE scale
> pred.failure = predict(m1, newdata = newdata,
+ type = "response", se.fit = TRUE)
> # prediction at TEMP=31
> pred.failure$fit

        1
0.9961828

> # 95% CI for predicted prob at TEMP=31
> c(pred.failure$fit - 1.96*pred.failure$se.fit,
+ pred.failure$fit + 1.96*pred.failure$se.fit)

        1         1
0.9728114 1.0195542

> # prediction on log(odds) scale
> pred.logit = predict(m1, newdata = newdata, type = "link", se.fit = TRUE)
> invlogit(pred.logit$fit)

        1
0.9961828

> # 95% CI for predicted prob at TEMP=31
> c(invlogit(pred.logit$fit - 1.96 * pred.logit$se.fit),
+ invlogit(pred.logit$fit + 1.96 * pred.logit$se.fit))

        1         1
0.3585385 0.9999918

> ##-- which confidence interval estimate would you report?
```

logit(failure probability) = 5.564414; the estimated probability of failure = 0.9962.

(f) Why must the answer to part (e) be treated cautiously?

Solution: It represents an extrapolation beyond the range of the available data.

2. (Ramsey and Schafer, 2013, Chap 20, Data Problem 15). A study examined the association between nesting locations of the Northern Spotted Owl and availability of mature forests. Wildlife biologists identified 30 nest sites. The researchers selected 30 other sites at random coordinates in the same forest. On the basis of aerial photographs, the percentage of mature forests was measured in various rings around each of the 60 sites. The data is in the file `owl.csv`.

(a) Apply two-sample t-tests to these data to see whether the percentage of mature forest is larger at nest sites than at random sites.

Solution: Our two-sample t-tests will compare the percentage of mature forest between nest and random sites at the same ring radius measurement.

```
> owl <- read.table("owl.csv", header=T, sep=",")
> attach(owl)
> names(owl)

[1] "SITE" "PCTRING1" "PCTRING2" "PCTRING3" "PCTRING4" "PCTRING5" "PCTRING6"
[8] "PCTRING7"

> owl.t.test<-function(){
+         for (i in 1:7){
+                 p.value[i]<-t.test(owl[,i+1][SITE=="Nest"],
+                 owl[,i+1][SITE=="Random"],
+                 alternative="greater")$p.value
+         }
+         names(p.value) <- colnames(owl)[-1]
+         return(p.value)
}

> p.value <- rep(0,7)
> owl.t.test()

     PCRING1      PCRING2      PCRING3      PCRING4      PCRING5      PCRING6
4.812467e-04 1.162899e-02 3.447056e-05 4.619681e-03 1.355561e-04 2.616569e-04
     PCRING7
2.481927e-01
```

The one-sided p-values for the t-tests which compare the mean percentage of old forest in ring k about owl sites to the mean percentage of old forest in ring k about random sites, are k = 1: 0.0004; k =2: 0.012; k=3: 0.00003; k=4: 0.0046; k=5: 0.00014; k=6: 0.00026; k=7: 0.25. We can see from most rings (except `PCTRING7`), that there does exist correlation between the maturity of the forest and owl nesting.

However, the t-test comparisons are not too helpful in answering the question of interest. Our real question of interest is to predict nesting locations for owls on the basis of the percentage of mature forest in the surrounding area. Also, the significance of outer rings may be a result of a correlation with the percentage of mature forests in smaller rings.

(b) Construct a binary response variable that indicates whether a site is a nest site. Use a logistic regression to investigate how large an area about the site has importance in distinguishing nest sites from random sites on the basis of mature forest.

Solution: What are some strategies for model building? A logical guideline is to build the model out of the centre, that is, include outer rings after inner rings have been allowed to enter the model. We could also explore interaction terms, but be careful, the sample size is small so there may be insufficient degrees of freedom to estimate the additional parameters.

The analysis can also be based on backward selection. Start off with the full model including all the covariates.

```
> site.ind<-ifelse(SITE=="Nest",1,0)
> m1 <- glm(site.ind ~ PCTRING1 + PCTRING2 + PCTRING3 + PCTRING4 + PCTRING5 +
+           PCTRING6 + PCTRING7 , family = binomial(link="logit"));
> summary(m1)$coefficients
```

```
               Estimate Std. Error    z value    Pr(>|z|)
(Intercept) -9.80304165 3.38934315 -2.8923131 0.003824166
PCTRING1     0.05708155 0.03713149  1.5372812 0.124224461
PCTRING2    -0.11729910 0.04989653 -2.3508470 0.018730735
PCTRING3     0.12180836 0.05198975  2.3429301 0.019132966
PCTRING4    -0.01694037 0.04276520 -0.3961251 0.692012765
PCTRING5     0.03295553 0.03905371  0.8438516 0.398752354
PCTRING6     0.10890760 0.06631097  1.6423769 0.100511925
PCTRING7    -0.05156848 0.03618735 -1.4250417 0.154145142
```

From the above, we observe that the last 4 covariates appear to be not significant, we thus build a model with only the first three covariates, with the output being given by

```
> m2 <- glm(site.ind ~ PCTRING1 + PCTRING2 + PCTRING3,
+           family = binomial(link="logit"))


> summary(m2)$coefficients
```

```
               Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -6.84619158 2.03830138 -3.358773 0.0007828936
PCTRING1     0.06305353 0.03228349  1.953120 0.0508053646
```

```
PCTRING2    -0.08832661 0.04148260 -2.129245 0.0332340297
PCTRING3     0.12149131 0.04339201  2.799854 0.0051125649

> reduce = m2
> full = m1
> drop_in_dev = reduce$deviance - full$deviance
> dif_in_df = reduce$df.residual - full$df.residual
> pchisq(drop_in_dev, dif_in_df, lower.tail = FALSE)

[1] 0.09909625

> exp(summary(m2)$coefficients[4,1]*5)

[1] 1.835756

> exp((summary(m2)$coefficients[4,1]-1.96*summary(m2)$coefficients[4,2])*5)

[1] 1.199872

> exp((summary(m2)$coefficients[4,1]+1.96*summary(m2)$coefficients[4,2])*5)

[1] 2.808634
```

A drop-in-deviance test comparing these two models gives a p-value of about 0.1, thus we might treat the second model as our final one. From this model, we can see that the percentage of mature forest of the three inner rings may possess some importance in distinguishing the nest sites from random sites.

Specifically, the percentage of old forest in ring 3 is significantly associated with the odds that the site is a nest site, even after accounting for the percentage of old forest within rings 1 and 2 (2-sided p-value = 0.0005).

Associated with an additional 5% of old forest in ring 3 is an estimated increase in the odds that the centre is a nest site of 83% (95% CI estimated 20% to 180%).

References

F. L. Ramsey and D. W. Schafer. The statistical sleuth: a course in methods of data analysis. Brooks/Cole, 2013.