-

-

Problem 1 Dog mimics

In a recent experiment, researchers looked at whether dogs prefer people who mimic them. They had 30 dogs, and presented them with two researchers: one mimicked the dog, and the other did not. They then looked at which researcher the dog preferred. Their thesis was that dogs would prefer the human who mimicked them, so the proportion of trials where the dog preferred the mimicker should be larger than 50%. We want to estimate the actual proportion of trials where the dog preferred the mimicker from the experimental data. We can assume that the number of dogs preferring the mimicker follows a binomial distribution.

a) What are the parameters of the binomial distribution?

b) What assumptions do we need to make when using this distribution?

c) How reasonable are these assumptions for this experiment?

In the experiment, 16 dogs out of 30 preferred the mimicker.

d) What is the estimate for the preference?

Below we plot the likelihood curve (Figure 1).

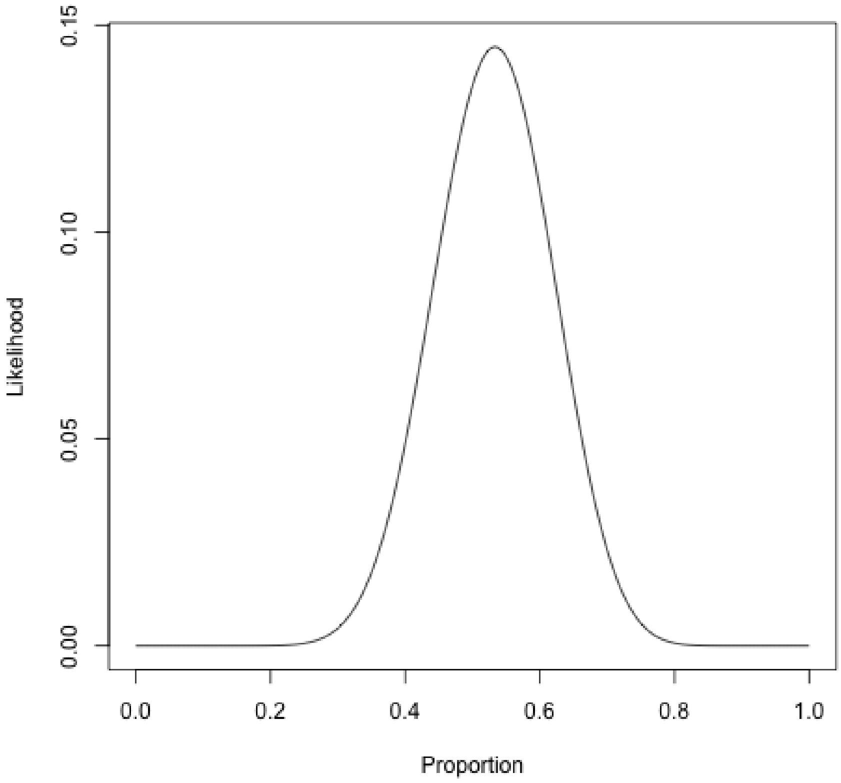

Figure 1: Plot of the likelihood for dog mimic data.

e) What is plotted on the y-axis against p, $Pr(n = 16\ |N = 30\ , p)$ or $Pr(p|n = 16\ , N = 30\ )$?

f) We want a confidence interval for the probability. Give one way we could we calculate an exact confidence interval for our estimate of the proportion of success?

We can calculate an approximate confidence interval by assuming a normal distribution. The standard error for a binomial distribution is $\sqrt{p(1-p)/N}$.

g) What is the standard error for this data?

h) What is the approximate 95% confidence interval?

i) Do the data show any evidence for dogs preferring mimics?

Problem 2 Circadian rhythms

It is well known that light has a strong effect on human circadian rhythm, e.g. sleeping patterns. In 1998 Campbell and Murphy claimed that the human circadian clock can be reset not just by light to the eyes, but by light applied to the back of the knee. In 2002 Wright and Czeisler decided to test this theory, they weren't completely convinced by the first study.

They constructed an experiment and measured daily cycles of melatonin (a hormone that regulates sleep-wake cycles) in 22 people. Each of these people were randomly assigned to one of three treatments; three hours of light to the eyes only, three hours of light to the knees only, and no light.

This produced data on:

- Melatonin shift in hours (**shift**), (negative = production later, positive = earlier production).
- Treatment group; control, eyes or knees (**treatment**).

10 rows of the data are shown in Figure 2.

```
treatment shift
1         control        0.53
2         control        0.36
3         control        0.20
10         knee        0.31
11         knee        0.03
12         knee        -0.29
20         eyes        -1.52
21         eyes        -2.04
22         eyes        -2.83
```

Figure 2: Top 10 rows of circadian rhythm data.

a) Write a biological question you could answer with this data.

b) Would you choose an lm() or a glm() to address this question, why?

The figures below show some plots relating to the model fit of a linear model of this data.

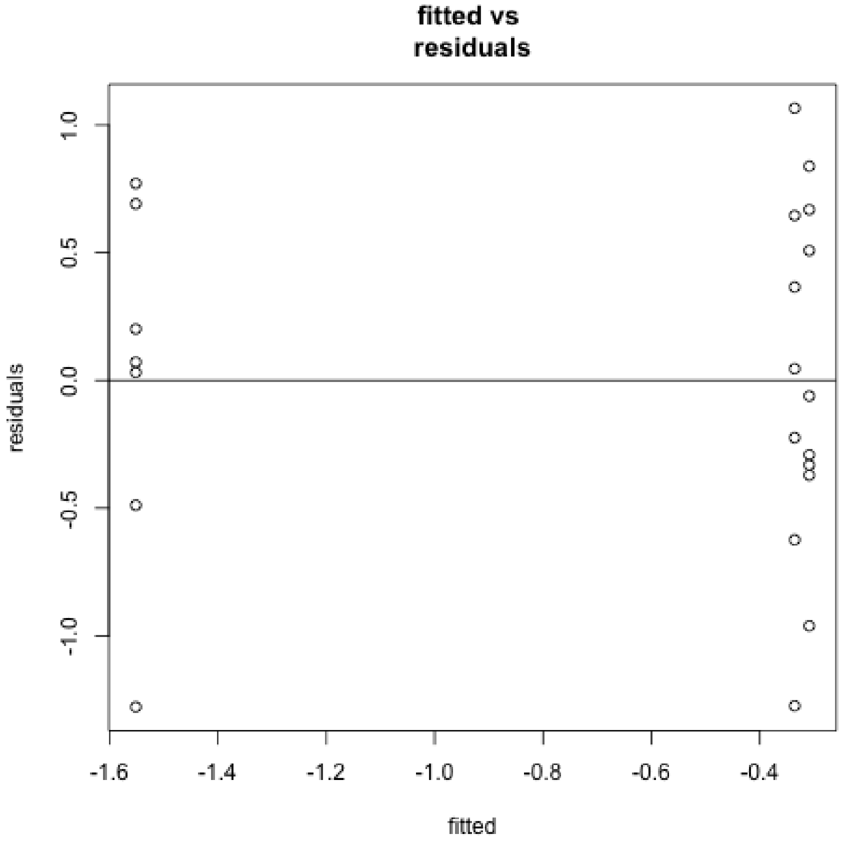

Figure 3: Plot of fitted versus residuals for model1.

c) Look at Figure 3, which assumption does this plot test? and how well does this model meet this assumption?

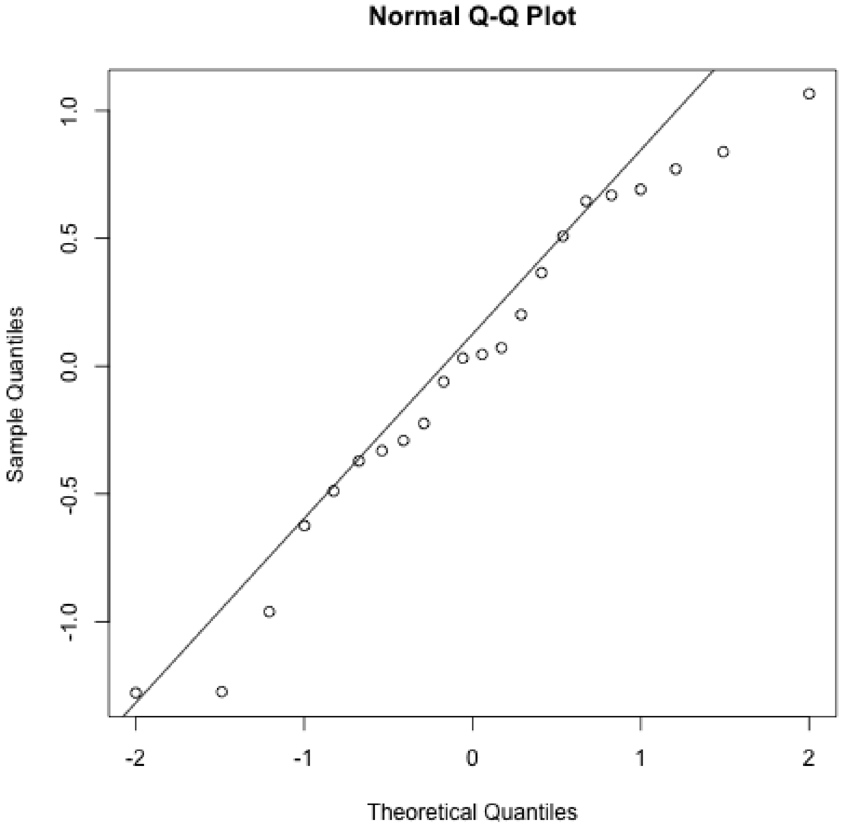

Figure 4: Normal QQ plot for model1.

d) Look at Figure 4, which assumption does this plot test? and how well does this model meet this assumption?

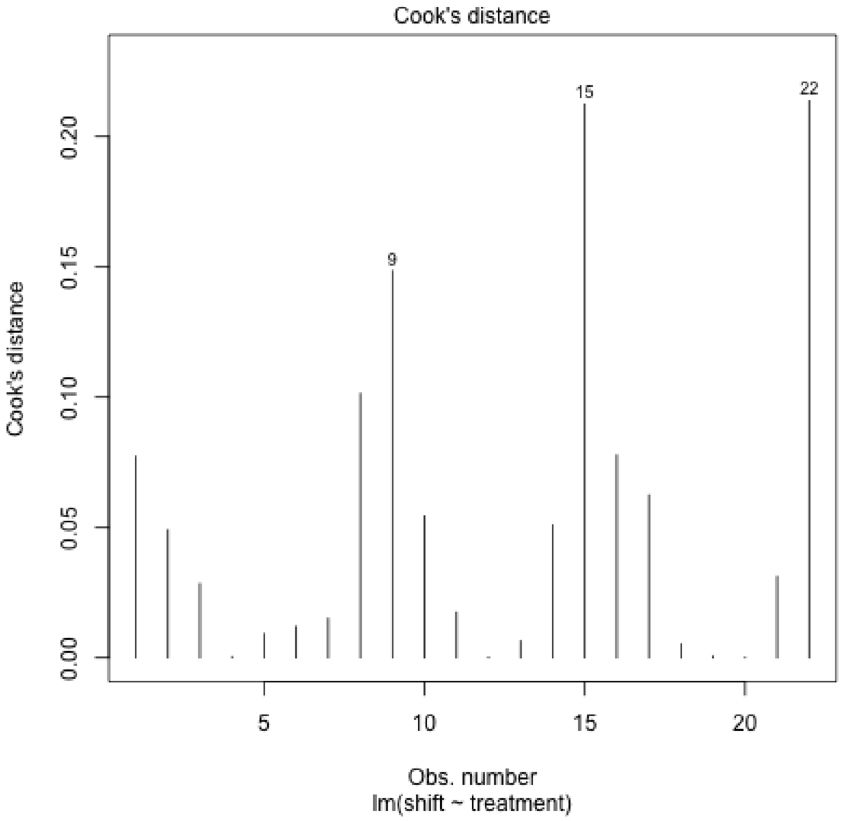

Figure 5: Cook's distance plot for model1.

e) Look at Figure 5, which assumption does this plot test? and how well does this model meet this assumption?

```
Call:
lm(formula = shift ~ treatment, data = LightData)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.27857 -0.36125  0.03857  0.61147  1.06571 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -0.30875    0.24888  -1.241  0.22988   
treatmenteyes -1.24268    0.36433  -3.411  0.00293 **
treatmentknee -0.02696    0.36433  -0.074  0.94178   
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.7039 on 19 degrees of freedom
Multiple R-squared:  0.4342,	Adjusted R-squared:  0.3746 
F-statistic: 7.289 on 2 and 19 DF,  p-value: 0.004472

                   2.5 %     97.5 %
(Intercept)   -0.8296694  0.2121694
treatmenteyes -2.0052265 -0.4801306
treatmentknee -0.7895122  0.7355836
```

Figure 6: Summary and confint output of model0

f) What do the coefficients from model0 (Figure 6) represent mathematically?

g) What statistical conclusions can you draw from these results??

Problem 3 Clutch size

A decrease in number of eggs a bird produces (it's clutch size) the later in the year it breeds, has been observed across many bird species. The data below can be used to test whether this happens in great tits (kjøttmeis).

Researchers collected data on when the first egg appeared in great tit nests and then counted the total number of eggs that were laid. The data comes from a single year (1990).

This produced data on:

- The date the first egg appeared, in days since the 1st April 1990 (**LayDate**) - The total number of eggs counted in the nest (**ClutchSize**)

The first 10 rows of the data are shown in Figure 7.

```
LayDate ClutchSize
6043        25         12
6044        18         9
6045        28         6
6046        24         8
6047        16         9
6048        18         9
6049        19         9
6050        4         12
6051        27         8
6052        30         7
```

Figure 7: Top 10 rows of clutch size data.

To answer the question *Does the time eggs are laid influence the total clutch size?* researchers fit the model in Figure 8:

```
model1 <- glm(ClutchSize ~ LayDate, data = BirdData, family = poisson(link = "log"))
```

Figure 8: R code for model1.

a) Why did researchers choose the model in Figure 8 for this data?

```
Analysis of Deviance Table

Model 1: ClutchSize ~ LayDate
Model 2: ClutchSize ~ LayDate + I(LayDate^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       166     62.384
2       165     62.374  1 0.010234   0.9194
```

Figure 10: R code output of Figure 9.

b) Why has a log link function been used here?

Before going too far with an analysis, the researchers decide to test whether there is any evidence of a quadratic effect of lay date on clutch size. As a specific hypothesis this is: *Does lay date have a quadratic association with clutch size in great tits?*.

They run the following R code in Figure 9:

```
model2 <- glm(ClutchSize ~ LayDate + I(LayDate^2), data = BirdData, family = poisson(l

anova(model1, model2, test="LRT")
```

Figure 9: R code.

c) Look at the output of the test in Figure 9, is there any evidence of a quadratic effect of lay date? Explain your answer.

d) We have several variables (mean temperature, precipitation, and insect abundance) which might also affect clutch size. How would we select the variables that best explain the data?

```
deviance(model1)/df.residual(model1)

[1] 0.3758064
```

Figure 11: R code calculating deviance ratio.

e) Look at the code in Figure 11, what assumption of model1 is this code testing? and is this assumption met?

```
coef(model1)

(Intercept)     LayDate
2.35205677 -0.01051424

confint(model1)

                  2.5 %       97.5 %
(Intercept)  2.22715164  2.475657352
LayDate     -0.01730312 -0.003797819
```

Figure 12: Coefficients and confidence intervals for model1.

As temperatures warm, lay dates are getting earlier. This year, 2019, the first great tit laid an egg 5 days before the start of April.

f) Predict the clutch size for a bird with a lay date of -5, using the coefficients in Figure 12.

g) The mean clutch size in 1990 was 9 eggs. What implications could the prediction for 2019 have for the great tit population, relative to 1990? Give a maximum of 3 implications.

h) Give two problems with generating predictions such as those in f (that predict outside of the original dataset)?