Clustered data models - Exercises 1

(Problems from 1 to 5 have been taken from Agresti, 2015, and problem 6 from Faraway, 2016)

1. (7.2 in Agresti) Suppose y_i are independent Poisson variates, with $\mu = \mathrm{E}(y_i),\ i = 1, ..., n$. For testing $H_0 :\ \mu = \mu_0$, show that the likelihood ratio statistic simplifies to

$$-2(L_0 - L_1) = 2[n(\mu_0 - \bar{y}) + n\bar{y}\log(\bar{y}/\mu_0)].$$

(Here, $L_0 = \log p(\mathbf{y}; \mu_0)$ and $L_1 = \log p(\mathbf{y}; \mu)$ denote log-likelihood functions.)

Solution. In the general case, the log-likelihood is given by

$$L_1 = \log \prod_{i=1}^{n} \frac{\mu^{y_i}}{y_i!} e^{-\mu} = \sum_{i=1}^{n} [y_i \log(\mu) - \mu - \log(y_i!)].$$

This is maximized by differentiating with respect to μ, setting the derivative equal to 0, and solving μ:

$$\frac{\partial L_1}{\partial \mu} = \frac{1}{\mu} \sum y_i - n = 0.$$

Thus $\hat{\mu} = \bar{y}$. In the restricted case, $\mu = \mu_0$:

$$L_0 = \sum_{i=1}^{n} [y_i \log(\mu_0) - \mu_0 - \log(y_i!)]$$

Thus, the likelihood ratio statistic becomes

$$\begin{aligned} -2(L_0 - L_1) &= -2[L(\mu = \mu_0) - L(\mu = \bar{y})] \\ &= -2 \left\{ \sum_{i=1}^{n} [y_i \log(\mu_0) - \mu_0 - \log(y_i!)] - \sum_{i=1}^{n} [y_i \log(\bar{y}) - \bar{y} - \log(y_i!)] \right\} \\ &= 2[n(\mu_0 - \bar{y}) + n\bar{y}\log(\bar{y}/\mu_0)]. \end{aligned}$$

2. (5.10 in Agresti) The calibration problem is that of estimating x_0 at which $P(y = 1) = \pi_0$ for some fixed π_0, such as 0.5. For the logistic model with a single explanatory variable, explain why a confidence interval for x_0 is the set of x values for which

$$|\hat{\beta}_0 + \hat{\beta}_1 x - \mathrm{logit}(\pi_0)| / [\mathrm{Var}(\hat{\beta}_0) + x^2 \mathrm{Var}(\hat{\beta}_1) + 2x\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)]^{1/2} < z_{\alpha/2}.$$

How could you invert a likelihood-ratio test to form an interval?

Solution. The value x_0 is the solution of the equation $\beta_0 + \beta_1 x = \mathrm{logit}(\pi_0)$. We can interpret this equation as the null hypothesis. In the alternative hypothesis, there are no restrictions on β_0 and β_1. The values of x for which the hypothesis is not rejected, are included in the confidence interval of x_0.

In a Wald-type test, we can interpret $\hat{\beta}_0 + \hat{\beta}_1 x$ as a test statistic and standardize it under the null hypothesis. If the null hypothesis is true, the standardized statistic approximately follows the standard normal distribution:

$$z = \frac{\hat{\beta}_0 + \hat{\beta}_1 x - \mathrm{logit}(\pi_0)}{\sqrt{\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)}} \sim N(0, 1).$$

The values of x, for which $|z| < z_{\alpha/2}$, are included in the confidence region. Noting that $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) = \mathrm{Var}(\hat{\beta}_0) + x^2 \mathrm{Var}(\hat{\beta}_1) + 2x\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$, we obtain what is claimed in the exercise.

We can also determine the confidence interval by inverting the likelihood ratio test. The test statistic is given by $-2(L_0 - L_1)$, where L_0 is the maximized log-likelihood function under the null hypothesis $\beta_0 + \beta_1 x = \text{logit}(\pi_0)$ and L_1 is that under no restrictions. When the null hypothesis is true, the test statistic approximately follows the chi squared distributed with 1 degree of freedom because there is one parameter restriction. The values of x for which the test is not rejected are included in the confidence interval. (All this should be done numerically – for example, by coding an R function – because there are no closed-form solutions.)

3. (7.32 in Agresti) For the horseshoe crab data, the negative binomial modeling shown in the R output first treats color as nominal-scale and then in quantitative manner, with the category numbers as scores. Interpret the result of the likelihood ratio test comparing the two models. For the simpler model, interpret the color effect and interpret results of the likelihood test of the null hypothesis of no color effect.

```
library(MASS)
fit.nb.color <- glm.nb(y ~ factor(color), data = Crabs) # Using Crabs.dat file
summary(fit.nb.color)
##
## Call:
## glm.nb(formula = y ~ factor(color), data = Crabs, init.theta = 0.8018786143,
##     link = log)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.4069     0.3526   3.990 6.61e-05 ***
## factor(color)2  -0.2146     0.3750  -0.572    0.567
## factor(color)3  -0.6061     0.4036  -1.502    0.133
## factor(color)4  -0.6913     0.4508  -1.533    0.125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.8019) family taken to be 1)
##
##     Null deviance: 199.23  on 172  degrees of freedom
## Residual deviance: 194.00  on 169  degrees of freedom
## AIC: 772.3
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.802
##           Std. Err.:  0.136
##
##  2 x log-likelihood:  -762.296
```

```
fit.nb.color2 <- glm.nb(y ~ color, data = Crabs) # Using color scores 1,2,3,4
summary(fit.nb.color2)
##
## Call:
## glm.nb(formula = y ~ color, data = Crabs, init.theta = 0.798572811,
##     link = log)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.7045     0.3095   5.507 3.66e-08 ***
```

```
## color        -0.2689     0.1225  -2.194   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.7986) family taken to be 1)
##
##     Null deviance: 198.77  on 172  degrees of freedom
## Residual deviance: 193.94  on 171  degrees of freedom
## AIC: 768.68
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.799
##           Std. Err.:  0.136
##
##  2 x log-likelihood:  -762.679
```

```
anova(fit.nb.color2, fit.nb.color)
## Likelihood ratio tests of Negative Binomial Models
##
## Response: y
##           Model     theta Resid. df    2 x log-lik.   Test    df LR stat.
## 1         color 0.7985728       171       -762.6794
## 2 factor(color) 0.8018786       169       -762.2960 1 vs 2     2 0.383383
##     Pr(Chi)
## 1
## 2 0.8255615
```

```
1 - pchisq(767.409 - 762.679, df = 172 - 171) # LR test vs. null model
## [1] 0.02964089
```

Solution. The likelihood ratio test does not reject the null hypothesis. Thus, the simpler model appears to be sufficient. The coefficient (-0.2689) is negative, so that the number of male satellites is smaller for darker (older) female crabs. When one moves to the next darkness category, the expected number of male satellites decreases by $100(1 - \exp(-0.2689))\% \approx 24\%$. The null hypothesis of no color effect is rejected. The p-value 0.02964089 (likelihood ratio test) is close to the p-value of the color effect 0.0282 (Wald test).

4. (7.33 in Agresti) For the horseshoe crab data, the following output shows a zero-inflated negtive binomial model using quantitative color for the zero component. Interpret results, and compare with the NB2 model fitted in the previous exercise with quantitative color. Can you conduct a likelihood-ratio test comparing them? Why or why not?

```
library(pscl)
summary(zeroinfl(y ~ 1 | color, dist = "negbin", data = Crabs))
##
## Call:
## zeroinfl(formula = y ~ 1 | color, data = Crabs, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.2246 -0.8186 -0.1712  0.5465  3.7220
##
```

```
## Count model coefficients (negbin with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.46324    0.06892  21.231  < 2e-16 ***
## Log(theta)   1.47997    0.35114   4.215  2.5e-05 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.7520     0.6658  -4.133 3.58e-05 ***
## color         0.8023     0.2389   3.358 0.000785 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 4.3928
## Number of iterations in BFGS optimization: 14
## Log-likelihood:  -363 on 4 Df
```

Solution. The coefficient for the color (0.8023) is positive, so the probability of a zero count (no male satellites) is larger for darker female crabs. The odds ratio for a zero count is $\exp(0.8023) = 2.231$ when comparing two consecutive color categories. For lightest crabs, the probability is $\text{logit}^{-1}(-2.7520 + 0.8023) = 0.1246$ and for darkest $\text{logit}^{-1}(-2.7520 + 4 \cdot 0.8023) = 0.6123$.

The AIC is 768.68 for the NB2 model and $-2 \cdot -363 + 2 \cdot 4 = 734$ for the zero-inflated model, so the latter model fits better. One cannot use the likelihood ratio test to compare the model, since neither of them is nested in the other. The first model does not contain the zero inflation part and the second model does not have color as an explanatory variable in the linear part.

5. (5.32 in Agresti) For the horseshoe crab dataset, let $y = 1$ if a female crab has at least one satellite, and let $y = 0$ if a female crab does not have any satellites. Fit a main-effects logistic model using color and weight as explanatory variables. Interpret and show how to conduct inference about the color and weight effects. Next, allow interaction between color and weight in their effects on y, and test whether this model provides a significantly better fit.

Solution.

```
Crabs <- read.table("~/tyo/opetus/clustered/data/Crabs.dat", header=TRUE)
Crabs$y2 <- ifelse(Crabs$y==0, 0, 1)
model1 <- glm(y2 ~ weight + color, family = binomial, data = Crabs)
summary(model1)
##
## Call:
## glm(formula = y2 ~ weight + color, family = binomial, data = Crabs)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.0316     1.1161  -1.820   0.0687 .
## weight        1.6531     0.3825   4.322 1.55e-05 ***
## color        -0.5142     0.2234  -2.302   0.0213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 190.27  on 170  degrees of freedom
```

```
## AIC: 196.27
##
## Number of Fisher Scoring iterations: 4
```

```
model1b <- glm(y2 ~ weight + factor(color), family = binomial, data = Crabs)
anova(model1b)
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y2
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            172     225.76
## weight         1  30.0214       171     195.74 4.273e-08 ***
## factor(color)  3   7.1949       168     188.54   0.06594 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2 <- glm(y2 ~ weight*color, family = binomial, data = Crabs)
model2b <- glm(y2 ~ weight*factor(color), family = binomial, data = Crabs)
anova(model2b)
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y2
##
## Terms added sequentially (first to last)
##
##
##                      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                  172     225.76
## weight                1  30.0214       171     195.74 4.273e-08 ***
## factor(color)         3   7.1949       168     188.54   0.06594 .
## weight:factor(color)  3   6.8860       165     181.66   0.07562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model2)
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y2
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           172     225.76
```

```
## weight        1  30.0214       171     195.74 4.273e-08 ***
## color         1   5.4684       170     190.27   0.01936 *
## weight:color  1   0.0791       169     190.19   0.77851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

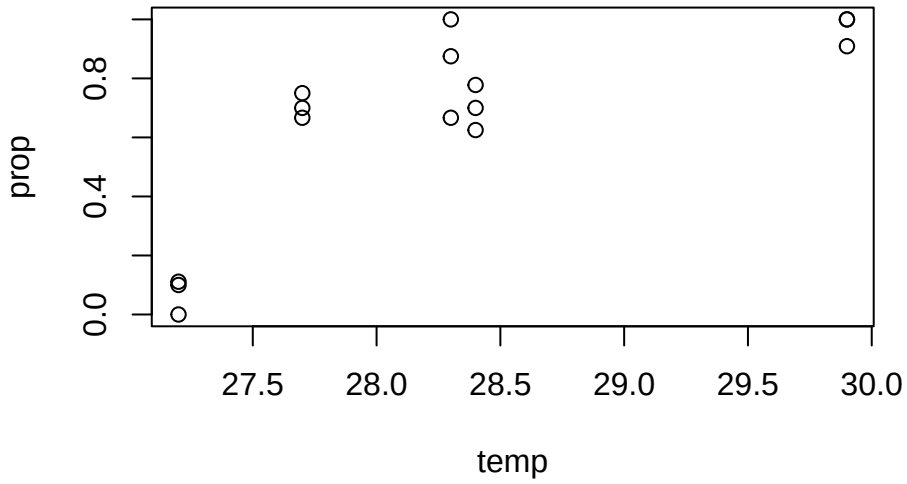
The color is a significant predictor when it is considered quantitative. Interpretation: the odds ratio for a female crab for her having at least one male satellite is $\exp(1.6531) = 5.2231$ when the comparison is made with a crab that weighs 1 kg less. Further, the odds ratio is $\exp(-0.5142) = 0.5980$ when the crab is compared with a crab belonging to the closest lower darkness category.

The interaction effect is not significant, irrespective of whether the color is modeled as categorical or quantitative.

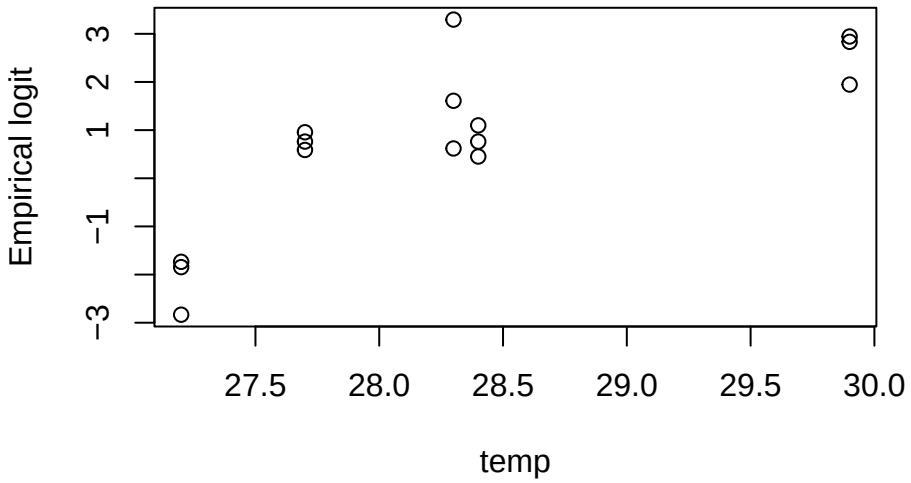
6. (3.2 in Faraway) Incubation temperature can affect the sex of turtles. An experiment was conducted with three independent replicates for each temperature and the number of male and female turtles born was recorded and can be found in the `turtle` dataset, included in the R package `faraway`.

(a) Plot the proportion of males against the temperature. Comment on the nature of the relationship.

```
library(faraway)
turtle$prop <- turtle$male/(turtle$male+turtle$female)
plot(prop ~ temp, turtle)
```




```
plot(log((male+0.5)/(female+0.5)) ~ temp, turtle, ylab = "Empirical logit")
```



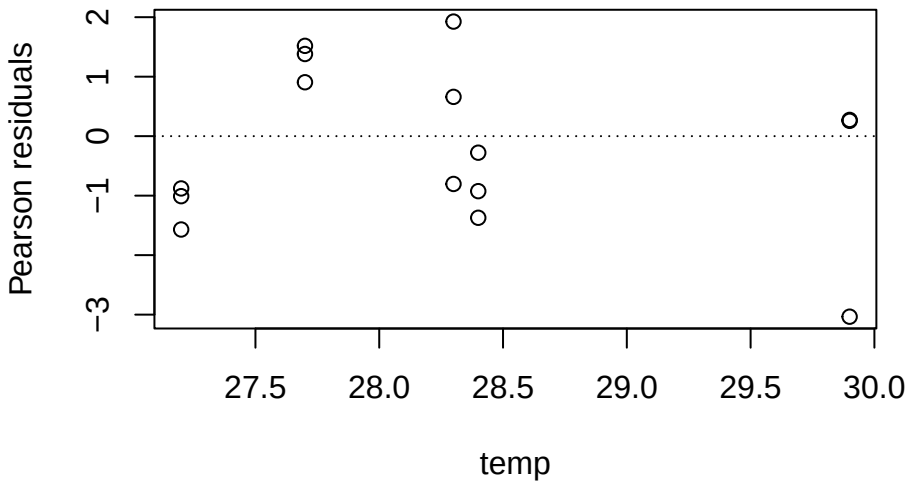

```
# The proportion of males grows with temperature but the relationship is nonlinear.
```

(b) Fit a binomial response model with a linear term in temperature. Does this model fit the data?

```
model1 <- glm(prop ~ temp, weights = male + female, family = binomial, data = turtle)
summary(model1)
```

```
##
## Call:
## glm(formula = prop ~ temp, family = binomial, data = turtle,
##     weights = male + female)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -61.3183    12.0224  -5.100 3.39e-07 ***
## temp          2.2110     0.4309   5.132 2.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.508  on 14  degrees of freedom
## Residual deviance: 24.942  on 13  degrees of freedom
## AIC: 53.836
##
## Number of Fisher Scoring iterations: 5
```

```
plot(residuals(model1, type = "pearson") ~ temp, turtle, ylab = "Pearson residuals")
abline(h=0, lty = 3)
```




```
1-pchisq(24.942, df = 13)
```

```
## [1] 0.02349208
```

```
# The model does not fit well. The residual deviance is about twice the degrees of freedom.
# Further, the residuals are not zero-centered for all values of the predictor.
```

(c) Check for outliers.

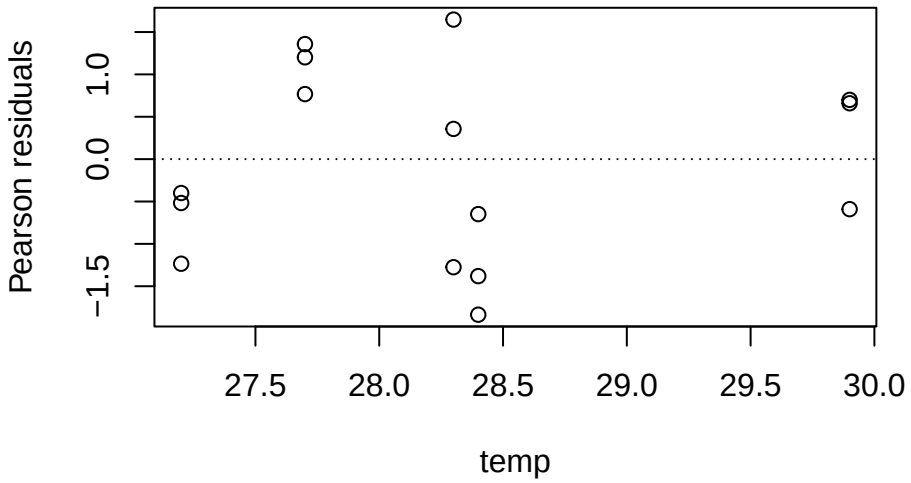
No outlier can be observed (see the first figure).

(d) Add a quadratic term in temperature. Is this additional term a significant predictor of the response? Does the quadratic model fit the data?

```
model2 <- glm(prop ~ temp + I(temp^2), weights = male + female, family = binomial, data = turtle)
summary(model2)
```

```
##
## Call:
## glm(formula = prop ~ temp + I(temp^2), family = binomial, data = turtle,
##     weights = male + female)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -677.5950   268.7984  -2.521   0.0117 *
## temp          45.9173    18.9169   2.427   0.0152 *
## I(temp^2)     -0.7745     0.3327  -2.328   0.0199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.508  on 14  degrees of freedom
## Residual deviance: 20.256  on 12  degrees of freedom
## AIC: 51.15
##
## Number of Fisher Scoring iterations: 4
```

```
plot(residuals(model2, type = "pearson") ~ temp, turtle, ylab = "Pearson residuals")
abline(h=0, lty = 3)
```




```
1-pchisq(20.256, df = 12)
```

```
## [1] 0.06239564
```

```
# The quadratic term is significant.
# According to the chi squared test, the fit is ok.
# The residual figure indicates that there is still room for improvement.
```

(e) There are three replicates for each value of temperature. Assuming independent binomial variation, how much variation would be expected in the three proportions observed? Compare this to the observed variation in these proportions. Do they approximately agree or is there evidence of greater variation?

```
# For this question and the next, I aggregate the dataset. There are several ways
# to do this in R; here, I use tapply.

turtle2 <- data.frame(temp = unique(turtle$temp), male = tapply(turtle$male, turtle$temp, sum),
                      female = tapply(turtle$female, turtle$temp, sum))
turtle2$n <- turtle2$male+turtle2$female
turtle2$prop <- turtle2$male/turtle2$n

turtle2$exp.var <- turtle2$prop*(1-turtle2$prop)/(turtle2$n/3)# expected variance in proportions.
# On average, there are turtle2$n/3 (about 10) turtles in each cluster

# Expected standard deviations for different temperatures:
sqrt(turtle2$exp.var)
```

```
## [1] 0.08729713 0.16070051 0.10749677 0.15220775 0.06074429
```

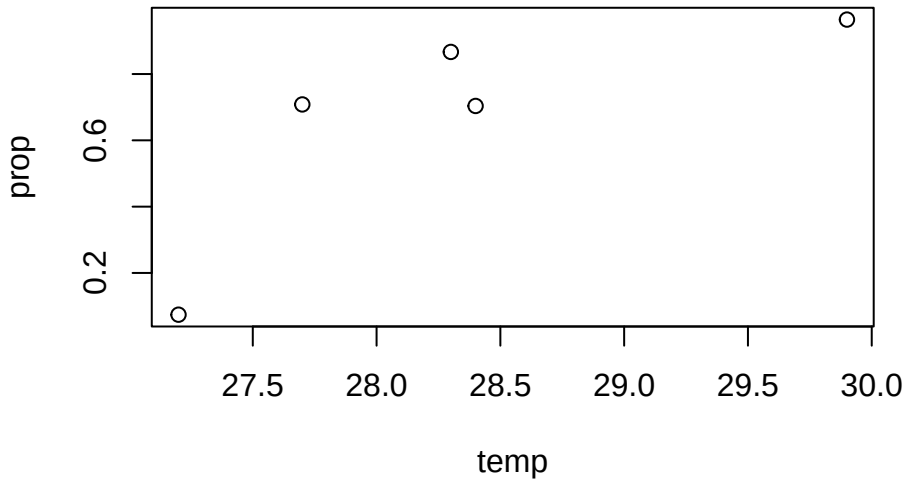
```
# empirical standard deviations for different temperatures
tapply(turtle$prop, turtle$temp, sd)
```

```
##       27.2       27.7       28.3       28.4       29.9
## 0.06119523 0.04194352 0.16839383 0.07639310 0.05248639
```

```
# There is no evidence of overdispersion
```

(f) If the three replicates are homogenous, they could be combined so that the dataset would have only five cases in total. Create this dataset and fit a model linear in temperature. Compare the fit seen for this model with that found in (b).

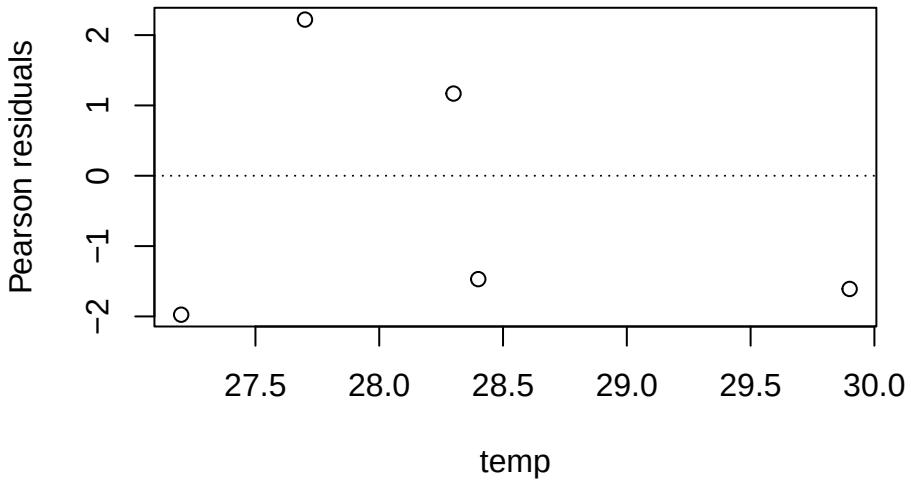
```
plot(prop ~ temp, turtle2)
```




```
model1b <- glm(prop ~ temp, family = binomial, weights = male + female, data = turtle2)
summary(model1b)
```

```
##
## Call:
## glm(formula = prop ~ temp, family = binomial, data = turtle2,
##     weights = male + female)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -61.3183    12.0224  -5.100 3.39e-07 ***
## temp          2.2110     0.4309   5.132 2.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 64.429  on 4  degrees of freedom
## Residual deviance: 14.863  on 3  degrees of freedom
## AIC: 33.542
##
## Number of Fisher Scoring iterations: 5
```

```
plot(residuals(model1b, type = "pearson") ~ temp, turtle2, ylab = "Pearson residuals")
abline(h=0, lty = 3)
```




```
1-pchisq(14.863, df=3)
```

```
## [1] 0.001937548
```

```
# The fit is even poorer here.
```