Problem 1

In a study of nesting horseshoe crabs, each female horseshoe crab had a male crab attached to her in her nest. The study investigated factors that affect whether the female crab had any other males, called satellites, residing near her. Explanatory variables that are thought to affect this included the female crab's color (C), spine condition (S), carapace width (W) and weight (Wt). The response outcome for each female crab is her number of satellites (Sa). There are 173 females in this study. Data are provided in the crab.txt.

(a) Fit a Poisson model (M1) with log link with W as the single predictor. Check the goodness of fit and interpret your model.

(b) Fit a model (M2) with W and Wt as predictors. Compare it with the model in (a). Interpret your results.

(c) Check over dispersion in M2. Interpret the model after adjusting for over dispersion.

Solution:

(a)

$$log(E(y)) = \beta_0 + \beta_1 W = -3.305 + 0.164W$$

We get a deviance statistic 567.88, or Pearson chi-squared statistic 544.157, with df = 173-2=171. The corresponding p-value is very closed to 0. Hence the model does not fit the data well. We should probably consider overdispersion issues.

Interpretation: Carapace width is a significant predictor (p-value < 0.001). With each unit of increase in carapace width, expected count of satellites increases by a factor of $exp(0.164) = 1.178$.

(b)

$$log(E(y)) = \beta_0 + \beta_1 W + \beta_2 Wt = -1.29 + 0.046W + 0.447Wt$$

We get a deviance statistic 559.89, or Pearson chi-squared statistic 536.59, with df = 173-3=170. The corresponding p-value is very closed to 0. Hence the model also does not fit the data well. Compared with model in a), weight variable is significant in model b) (Wald or deviance statistic should be reported). However, the model fit is still not good based on GOF test.

(c) From the half-normal plot using residuals, we are certain that over dispersion exists since the plot deviates from slope 1. The over dispersion parameter can be estimated by $G/(n-p) = 536.6/(173-3) = 3.16$. The model after adjusting for covariates has the same coefficient as the original model. Using adjusted Pearson chi-squared statistic 163 with df=173-3=170, we get p-value $0.64 > 0.05$. Hence we do not have enough evidence to show the model does not fit the data well.

Problem 2

Researchers examined a large number of fish to determine the prevalence of parasites. The dataset (parasite.txt) includes the variables Intensity (i.e., the number of parasites), Area (a categorical variable), Year (to be treated as categorical), and Length of the fish.

(a) Fit a Poisson model with log link to the data with area, year, and length as predictors. Interpret each model parameter.

(b) Test for goodness of fit of the model in (a) and state conclusions.

(c) Researchers suspect that there may be two strains of fish, one that is susceptible to parasites and one that is not. Without knowing which fish are susceptible, this could be regarded as a zero-inflated model. Building on the model in (a) (using the same predictors), fit an appropriate model to the data that can account for extra zeros. Provide an interpretation for each model parameter in terms of the problem.

Solution:

(a) On average, the intensity of parasites for fish in area 2 is 0.809 times the intensity of parasites for fish in area 1, adjusting for year and length of the fish.
On average, the intensity of parasites for fish in area 3 is 0.890 times the intensity of parasites for fish in area 1, adjusting for year and length of the fish.
On average, the intensity of parasites for fish in area 4 is 4.08 times the intensity of parasites for fish in area 1, adjusting for year and length of the fish.
On average, the intensity of parasites for fish in year 2000 is 1.95 times the intensity of parasites for fish in year 1999, adjusting for area and length of the fish.
On average, the intensity of parasites for fish in year 2001 is 0.80 times the intensity of parasites for fish in year 1999, adjusting for area and length of the fish.
On average, one unit increase in the length of the fish is associated with 2.8% (1-0.972) decrease in the intensity of parasites for fish, adjusting for area and year.

(b) The deviance goodness of fit test statistic for the model is 19153. Compare this statistic to a chi-squared distribution with 1184 degrees of freedom and it yields a p-value that is extremely closed to 0. Hence, the model is not a good fit to the data.

(c) Interpretation of the coefficients in the count model part:
For the fish that is susceptible to parasites, the intensity of parasites for

fish in area 2 is on average 1.31 times the intensity of parasites for fish in area 1, adjusting for year and length of the fish.

For the fish that is susceptible to parasites, the intensity of parasites for fish in area 3 is on average 1.16 times the intensity of parasites for fish in area 1, adjusting for year and length of the fish.

For the fish that is susceptible to parasites, the intensity of parasites for fish in area 4 is on average 2.57 times the intensity of parasites for fish in area 1, adjusting for year and length of the fish.

For the fish that is susceptible to parasites, the intensity of parasites for fish in year 2000 is on average 1.48 times the intensity of parasites for fish in year 1999, adjusting for area and length of the fish.

For the fish that is susceptible to parasites, the intensity of parasites for fish in year 2001 is on average 0.96 times the intensity of parasites for fish in year 1999, adjusting for area and length of the fish.

For the fish that is susceptible to parasites, one unit increase in the length of the fish is expected to be associated with 4.0% (1-0.96) decrease in the intensity of parasites for fish, adjusting for area and year.

Interpretation of the coefficients in the binomial model part:

The odds ratio of having no parasite is expected to be 2.05 for fish in area 2 compared to area 1, adjusting for year and length of the fish.

The odds ratio of having no parasite is expected to be 1.93 for fish in area 3 compared to area 1, adjusting for year and length of the fish.

The odds ratio of having no parasite is expected to be 0.36 for fish in area 4 compared to area 1, adjusting for year and length of the fish.

The odds ratio of having no parasite is expected to be 0.47 in year 2000 compared to year 1999, adjusting for area and length of the fish.

The odds ratio of having no parasite is expected to be 1.58 in year 2001 compared to year 1999, adjusting for area and length of the fish.

On average, one unit increase in the length of fish is associated with 1% (1-0.99) decrease in the odds of having no parasite, adjusting for area and year.

```
# problem 1
# a)
crab = read.table("crab.txt",header=T)
fit1 <- glm(Sa ~ W, family=poisson(link=log), data = crab)
summary(fit1)

# Goodness of fit test
D = sum(residuals(fit1, type="deviance")^2)
G = sum(residuals(fit1, type="pearson")^2)
1-pchisq(D,173-2)

# b)
```

```
fit1b <- glm(Sa ~ W+Wt, family=poisson(link=log), data = crab)
summary(fit1b)

# Goodness of fit test
D = sum(residuals(fit1b, type="deviance")^2)
G = sum(residuals(fit1b, type="pearson")^2)
1-pchisq(D,173-3)

# c)
# test over-dispersion (half normal plot)
res=residuals(fit1b,type=’pearson’)
plot(qnorm((173+1:173+0.5)/(2*173+1.125)),sort(abs(res)),xlab=’Expected Half-Normal Order St
abline(a=0,b=1)
abline(a=0,b=sqrt(phi),lty=2)

# estimate over-dispersion parameter
res.p1=residuals(fit1b,type=’pearson’)  # exactly the same as pearson residual for wave.glm3
G1=sum(res.p1^2) # calc dispersion param based on full model
phi=G1/(173-3)
phi # 3.16

# fit model with constant over-dispersion
summary(fit1b,dispersion=phi)
# goodness of fit
pval=1-pchisq(G1/phi,170)
pval # fit is ok

# Problem 2
# a)
parasite = read.table("parasite.txt",header=T)
parasite$Intensity
unique(parasite$Prevalence)

fit2 <- glm(Intensity ~ factor(Area) + factor(Year) + Length,
            family=poisson(link=log), data=parasite)
summary(fit2)

exp(fit2$coefficients)

# b)
# Goodness of fit test
D = sum(residuals(fit2, type="deviance")^2)
G = sum(residuals(fit2, type="pearson")^2)
pchisq(deviance(fit2),df.residual(fit2), lower.tail = F)

# c)
```

```
library(pscl)
fit2c <- zeroinfl(Intensity ~ factor(Area) +  factor(Year) + Length, data = parasite)
summary(fit2c)
round(exp(fit2c$coefficients$count),2)
round(exp(fit2c$coefficients$zero),2)
```