36-402/608 Homework #12 Solutions 4/22

1. Lymphoma and radiation (34 points)

   Read problem 19.14 on page 574. Using ex1914.csv, load the data into R using this code:

```
lymph = read.csv("ex1914.csv")
lymphA = array(t(cbind(lymph$survive,lymph$died)),
               dim=c(2,2,17),
               dimnames = list(
                 outcome=c("survived","died"),
                 group=c("radiation","no"),
                 months=lymph$months[seq(1,by=2,length=17)]))
lymphA = aperm(lymphA,c(2,1,3))
```

   Perform the Woolf test (see HO20), and turn in the p-value and its interpretation.

```
woolf <- function(x) {
  x <- x + 1 / 2
  k <- dim(x)[3]
  or <- apply(x, 3, function(x) (x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
  w <- apply(x, 3, function(x) 1 / sum(1 / x))
  1 - pchisq(sum(w * (log(or) - weighted.mean(log(or), w)) ^ 2), k - 1)
}
woolf(lymphA)  # 0.946
```

   With a large p-value of 0.946, we retain the null hypothesis of equal odds ratios across the different months, so we can proceed to use the Mantel-Haenszel test.

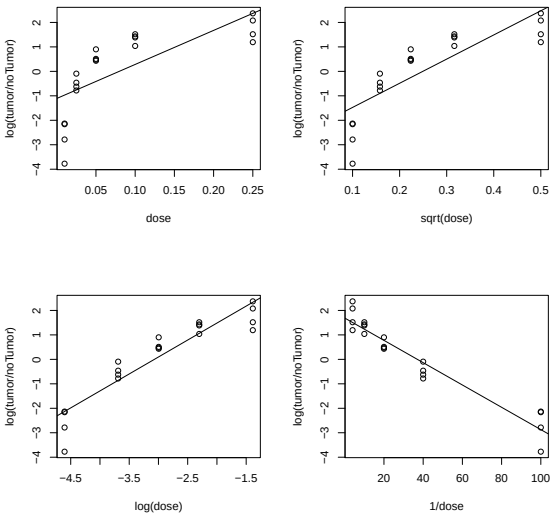
   Perform the Mantel-Haenszel test (two-sided). Turn in the M-H p-value, the estimate of the common odds ratio and its CI, and a careful statement of how these are interpreted.

```
mantelhaen.test(lymphA)
# Mantel-Haenszel X-squared = 2.3938, df = 1, p-value = 0.1218
# alternative hypothesis: true common odds ratio is not equal to 1
# 95 percent confidence interval:
#  0.1762224 1.1012146
# sample estimates:
# common odds ratio
#         0.4405209
```

With a p-value of 0.12, we retain the null hypothesis that the common odds ratio equals 1. As the text states, this can be interpreted as no evidence that the survival curves differ for those with and without radiation. At each month the odds of surviving for those in the radiation group is 0.18 to 1.10 times the odds of surviving in the no radiation group.

2. Trout tumors (33 points)

   Read problem 21.16 and load the data from ex2116.csv. Add the variable "noTumors" to the data frame.



   The EDA plots shown here demonstrate that the inverse of the dose is a good transformation to achieve linearity with the log odds of getting a tumor. Add an "invDose" variable to the data frame.

   Perform binomial logistic regression to model the log odds of a tumor vs. the inverse dose. Save the glm() object for later use.

```
trout = read.csv("ex2116.csv")
names(trout)=casefold(names(trout))
trout$noTumor = trout$total - trout$tumor
trout$invDose = 1/trout$dose
trlr = glm(cbind(tumor,noTumor)~invDose, trout, family="binomial")
```

   Perform quasibinomial logistic regression for the same model. Using the code on page 6 of HO21, obtain a p-value for the test with the null hypothesis of no overdispersion. Turn in the p-value.

```
qtr=glm(cbind(tumor,noTumor)~invDose, trout, family="quasibinomial")
```

```
1 - pchisq(summary(qtr)$dispersion * trlr$df.residual, trlr$df.residual)
# 0.0007922711  (p-value for overdispersion)
```

Based on the over-dispersion p-value, we must use the quasibinomial results to correct for over-dispersion (extra-binomial variation). Turn in the summary() for this model and an interpretation of $\exp(b_x)$ where x is the inverse dose. (You do not need to try to "undo" the meaning of the transformation.) Also roughly state in what way the quasibinomial results change our conclusions about the effects of dose on tumor production.

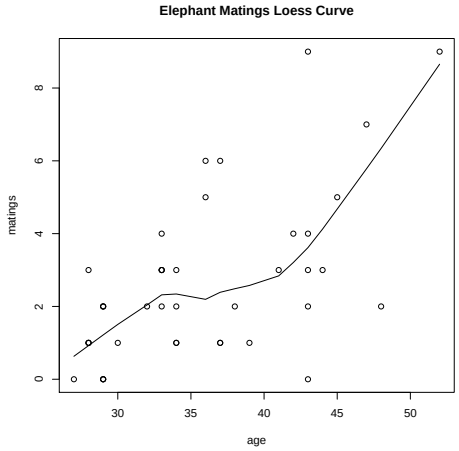
```
summary(qtr)
# Coefficients:   Estimate Std. Error t value Pr(>|t|)
# (Intercept)    1.636552   0.135625   12.07 4.62e-10 ***
# invDose       -0.046664   0.004069  -11.47 1.04e-09 ***
# (Dispersion parameter for quasibinomial family taken to be 2.390596)
#     Null deviance: 667.195  on 19  degrees of freedom
# Residual deviance:  41.809  on 18  degrees of freedom

exp(qtr$coef[2]) # 0.954
# Optional CI:
tmp = summary(qtr)$coef[2,]
round(exp(tmp[1]+c(-1,1)*1.96*tmp[2]), 3)  #  0.947 0.962
```

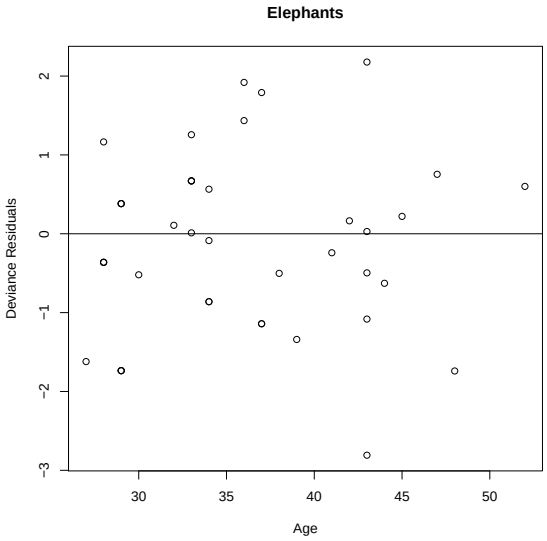
As the inverse of the dose goes up by one unit, the odds of having a tumor drop by 4.6% (95% CI=[3.8,5.3]), i.e., they are multiplied by 0.954.

3. Mating Elephants (33 points) Turn in R code and a brief summary of your conclusions for the mating elephant problem on page 645 using data from case2201.csv.

   An EDA plot is shown here.



Include a residual plot, as well as a check for extra-Poisson variation using family=quasipoisson with your results.



The residual plot looks fine.

```
mate = read.csv("case2201.csv")
names(mate)=casefold(names(mate))
m1 = glm(matings ~ age, mate, family="poisson")
qm1 = glm(matings ~ age, mate, family="quasipoisson")
1 - pchisq(summary(qm1)$dispersion * m1$df.residual, m1$df.residual)
#  0.2308694  (so overdispersion is not supported by the evidence)

summary(m1)
# Coefficients: Estimate Std. Error z value Pr(>|z|)
# (Intercept)   -1.58201    0.54462  -2.905  0.00368 **
# age            0.06869    0.01375   4.997 5.81e-07 ***
# (Dispersion parameter for poisson family taken to be 1)
#     Null deviance: 75.372  on 40  degrees of freedom
# Residual deviance: 51.012  on 39  degrees of freedom
# AIC: 156.46

exp(m1$coef[2]) # 1.07
```

I conclude that there is a significant increase in the odds of mating (1.07 times as much) for each year older an elephant gets in the age range studied.