Homework 05
Joseph Blubaugh
jblubau1@tamu.edu
STAT 659-700

3.11

a)

```r
A = c(8, 7, 6, 6, 3, 4, 7, 2, 3, 4); B = c(9, 9, 8, 14, 8, 13, 11, 5, 7, 6)

# a)
log(mean(B)) - log(mean(A)); round(exp(0.5877867), 1) == (mean(B)/mean(A))
```

```
[1] 0.5877867

[1] TRUE
```

b) Equation: $1.6094 + .5878x$ β has a multiplicative effect so the expected increase in defects from treatment B is $exp(.587) = 1.8$

```r
dta = data.frame(
  Y = c(A, B),
  X = c(rep("A", 10), rep("B", 10))
)

(mdl = glm(Y ~ X, family = poisson(link = "log"), data = dta))
```

```
Call:  glm(formula = Y ~ X, family = poisson(link = "log"), data = dta)

Coefficients:
(Intercept)           XB
     1.6094       0.5878

Degrees of Freedom: 19 Total (i.e. Null);  18 Residual
Null Deviance:      27.86
Residual Deviance: 16.27    AIC: 94.35
```

c) With a small pvalue we reject the null hypothesis that the treatments are the same

```r
x = 27.86 - 16.27; 1 - pchisq(x, 1)
```

```
[1] 0.0006630741
```

d) $.5878 \pm 1.96(.1764) = (.242, .993)$, $exp(.242, .993) = (1.27, 2.54)$

3.12

The likelihood ratio test using the deviance $16.26 - 14.435 = 1.832, 1 - pchisq(1.832, 1) = .175$ show that the coating thickness effect is insignificant. A 95% confidence interval for the coating parameter is $exp(-.2296 + c(-1, 1) * 1.96 * .1701) = (.569, 1.10)$.

```
dta$X2 = c(rep(0, 5), rep(1, 5), rep(0, 5), rep(1, 5))

mdl2 = glm(Y ~ X + X2, family = poisson(link = "log"), data = dta)
summary(mdl2)
```

```
Call:
glm(formula = Y ~ X + X2, family = poisson(link = "log"), data = dta)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.2952  -0.6785  -0.2688   0.6776   1.6307  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   1.7177     0.1602  10.719  < 2e-16 ***
XB            0.5878     0.1764   3.332 0.000861 ***
X2           -0.2296     0.1701  -1.349 0.177246    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 27.857  on 19  degrees of freedom
Residual deviance: 14.435  on 17  degrees of freedom
AIC: 94.517

Number of Fisher Scoring iterations: 4
```

3.13

a) The prediction equation is: $-.4284 + .5893(weight)$

```
crabs = read.csv("crabs.csv")
mdl = glm(satell ~ weight, family = poisson(link = "log"), data = crabs)
summary(mdl)
```

```

Call:
glm(formula = satell ~ weight, family = poisson(link = "log"),
    data = crabs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9307  -1.9981  -0.5627   0.9298   4.9992

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.42841    0.17893  -2.394   0.0167 *
weight       0.58930    0.06502   9.064   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 560.87  on 171  degrees of freedom
AIC: 920.16

Number of Fisher Scoring iterations: 5
```

b) $exp(-.4284 + .5893 * 2.44) = 2.74$

c) $exp(.5893 + \pm 1.96 * .06542) = exp(.461, .716) = (1.58, 2.04)$ We expect a 1 kg increase in crab weight to increase the number of satelites by 1.8

d) The Wald test concludes that weight has a significant effect on the number of satellites

```
library(aod); wald.test(b = coef(mdl), Sigma = vcov(mdl), Terms = 2)
```

```
Wald test:
----------

Chi-squared test:
X2 = 82.2, df = 1, P(> X2) = 0.0
```

e) The likelihood ratio test also concludes that weight has a significant effect on the number of satellites. $1 - pchisq(623.79 - 560.87, 1) =< .001$

3.14

a) The prediction equation is: $-.8647 + .7603(weight)$ The negative binomial model has a lower AIC score than the poisson model so there is evidence that the negative binomial model fits the data better. The poisson model also has a large deviance to df ratio indicating that the model does not fit the data very well.

```
library(MASS)

mdl.nb = glm.nb(satell ~ weight, data = crabs)
summary(mdl.nb)
```

```

Call:
glm.nb(formula = satell ~ weight, data = crabs, init.theta = 0.9310592338, 
    link = log)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.8394  -1.4122  -0.3247   0.4744   2.1279  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -0.8647     0.4048  -2.136   0.0327 *  
weight        0.7603     0.1578   4.817 1.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.9311) family taken to be 1)

    Null deviance: 216.43  on 172  degrees of freedom
Residual deviance: 196.16  on 171  degrees of freedom
AIC: 754.64

Number of Fisher Scoring iterations: 1


              Theta:  0.931 
          Std. Err.:  0.168 

 2 x log-likelihood:  -748.644 
```

b) $exp(.7603 + c(-1, 1) * 1.96 * .1578) = (1.57, 2.91)$. The negative binomial model has variance equal to the variance of the poisson model plus the dispersion factor which affects the range of the the interval estimates. The negative binomial model will always produce wider intervals than the poisson model when the dispersion paramter is > 0.

3.15

a)

```
exp(-2.38 + 1.733); exp(-2.38)
```

```
[1] 0.5236143
```

```
[1] 0.09255058
```

b)

```
exp(1.733 + c(-1, 1) * 1.96 * .147)
```

```
[1] 4.241343 7.546773
```

c) The negative binomial models confidence intervals are more believable because the variance for blacks and whites is much larger than the sample means so the poisson model is not really appropriate.

d) The negative binomial model has a large dispersion parameter indicating that the variance is much larger than the mean probably due to the fact that the data set has a lot of 0s, much more more than would be expected under the poisson distribution.

3.17

Model: $log(\mu) = \alpha$
μ Estimate: $log(\mu) = .495/38.7 = .0127$
Standard Error: $\sqrt{.0127} = .1126$

Model with offset: $log(\mu) = .0127 + log(38.7) = 3.67$
95% CI: $3.67 \pm 1.96 * .1126 = (3.44, 3.89) \rightarrow exp(3.44, 3.89) = (31.5, 48.9)$

The 95% confidence interval for the expected rate of injuries per 1000 driving years is 31.5 - 48.9

3.18

a) The number of arrests is correlated with the number of attendees so it might be reasonable to assume that the number of attendees times an overall rate would approximate the number of arrests.

$$\begin{aligned} E(Y) =& \mu(Attendence) \\ log(E(Y)/Attendence) =& log(\mu) + log(Attendence) \end{aligned}$$

b) μ is the intercept: 3.64, the expected number of arrests are $exp(3.64) = 38.1$

```
dta = data.frame(
  Attendence = c(404,286,443,169,222,150,321,189,258,223,211,215,108,
                 210,224,211,168,185,158,429,226,150,148 ),
  Arrests = c(308,197,184,149,132,126,110,101,99,81,79,78,68,67,60,
              57,55,44,38,35,29,20,19)
)

(mdl = glm(Arrests ~ 0 + Attendence, family = poisson(link = "log"),
           data = dta))
```

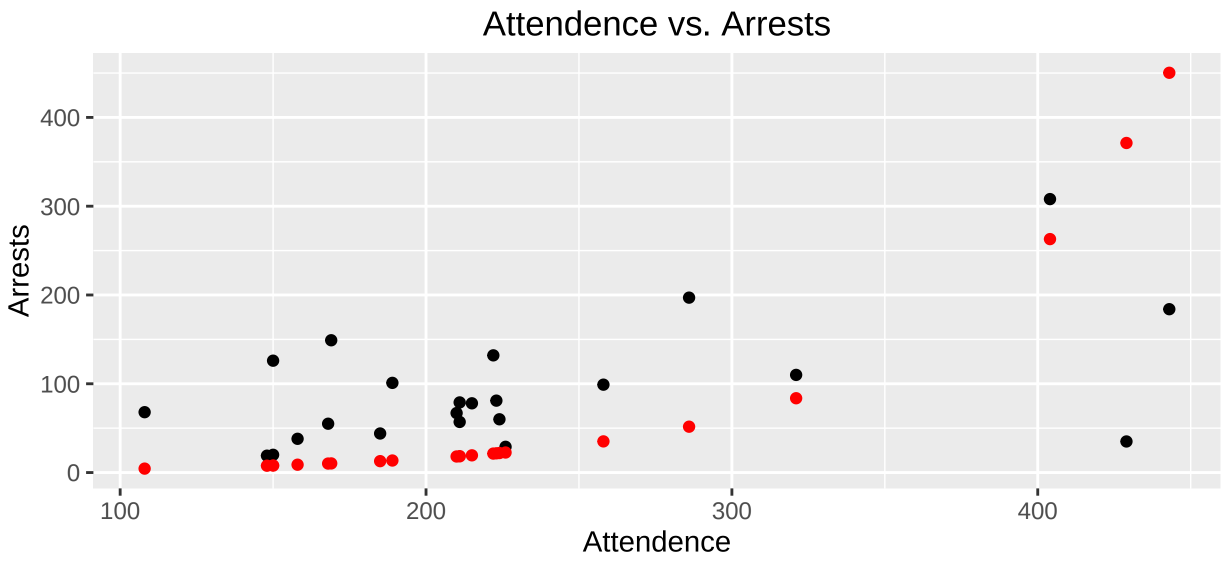
```
Call:  glm(formula = Arrests ~ 0 + Attendence, family = poisson(link = "log"),
    data = dta)

Coefficients:
Attendence
   0.01379

Degrees of Freedom: 23 Total (i.e. Null);  22 Residual
Null Deviance:      16080
Residual Deviance: 3453     AIC: 3597
```

c) There are many records that would be considered extreme, however the most extreme are 1, 2, 3, 20, 22, 23

```
library(ggplot2)

dta$Prediction = exp(predict(mdl, dta))

## Pearson Residuals
dta$rP = with(dta,
              (Arrests - mean(Arrests)) /
                (sqrt(mean(Arrests) * ( 1 - hatvalues(mdl))))
)

dta
```

```
   Attendence Arrests Prediction         rP
1         404     308 262.963075 24.8180326
2         286     197  51.652794 10.9083914
3         443     184 450.296397 12.1384011
4         169     149  10.286857  5.8283506
5         222     132  21.367030  4.0700289
6         150     126   7.915453  3.4392403
7         321     110  83.702450  1.8127017
8         189     101  13.554375  0.8445886
9         258      99  35.105757  0.6394749
10        223      81  21.663768 -1.2346424
11        211      79  18.359292 -1.4418400
12        215      78  19.400606 -1.5460713
13        108      68   4.435091 -2.5809611
14        210      67  18.107817 -2.6892125
15        224      60  21.964627 -3.4191990
16        211      57  18.359292 -3.7288966
17        168      55  10.145954 -3.9321528
18        185      44  12.826853 -5.0760537
19        158      38   8.838824 -5.6964993
20        429      35 371.228006 -7.1974380
21        226      29  22.578938 -6.6446794
22        150      20   7.915453 -7.5645233
23        148      19   7.700095 -7.6681713
```

```
ggplot(dta) +
  geom_point(aes(x = Attendence, y = Arrests), color = "black") +
  geom_point(aes(x = Attendence, y = Prediction), color = "red") +
  ggtitle("Attendence vs. Arrests")
```



d) The dispersion parameter is very high which would indicate that the poisson model is not a good fit. Furthermore the deviance/df measure is very high as well also supporting that the poisson model method is inappropriate.

```
mdl = glm.nb((Arrests/Attendence) ~ 1, data = dta)
summary(mdl)
```

```

Call:
glm.nb(formula = (Arrests/Attendence) ~ 1, data = dta, init.theta = 54926.29517,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8910  -0.8723  -0.8603   0.8103   0.8545

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9046     0.3278   -2.76  0.00578 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(54926.3) family taken to be 1)

    Null deviance: 16.84  on 22  degrees of freedom
Residual deviance: 16.84  on 22  degrees of freedom
AIC: 35.104

Number of Fisher Scoring iterations: 1

              Theta:  54926
          Std. Err.:  2456813
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -31.104
```

3.19

a) the deviance to df ratio for the intercept only model is low so the poisson model does a pretty good job fitting to the data, however when compared to the model with year as a predictor variable the ratio is even lower indicating that time plays a part in the number of train collisions and so the rate does not remain constant over time.

```
  year  km train.collisions train.road.collisions collisions
1 1975 436                5                     2          7
2 1976 426                2                    12         14
3 1977 425                1                     8          9
4 1978 430                2                     4          6
5 1979 426                3                     3          6
6 1980 430                2                     2          4
```

```
(mdl.1 = glm(collisions ~ 1, family = poisson(link = "log"), data = dta))
```

```

Call:  glm(formula = collisions ~ 1, family = poisson(link = "log"),
    data = dta)

Coefficients:
(Intercept)
      1.769

Degrees of Freedom: 28 Total (i.e. Null);  28 Residual
Null Deviance:      39.39
Residual Deviance: 39.39    AIC: 143.8
```

```
(mdl.2 = glm(collisions ~ year, family = poisson(link = "log"), data = dta))
```

```

Call:  glm(formula = collisions ~ year, family = poisson(link = "log"),
    data = dta)

Coefficients:
(Intercept)         year
   70.02294     -0.03434

Degrees of Freedom: 28 Total (i.e. Null);  27 Residual
Null Deviance:      39.39
Residual Deviance: 25.7     AIC: 132.1
```

b) `(-.0337)^2 / (.013^2) = 6.72` which is larger than the critical value of 3.84 so we conclude that $\beta \neq 0$

c) `1 - exp(c(-.06, -.08)) = (.058, .076)`, each year, the number of train collisions is expected to decrease between 5.8-7.6%.

3.20

a) For all age groups except the oldest group, the ratio of smokers dieing from coronary issues is higher than non smokers. The highest rate difference is for the youngest group and it steadily declines from there.

Table 1: Table continues below

Age	Non.Smoker.Yrs	Smoker.Yrs	Death.Non.Smoker	Death.Smoker
35-44	18793	52407	2	32
45-54	10673	43248	12	104
55-64	5710	28612	28	206
65-74	2585	12663	28	186
75-84	1462	5317	31	102

Non.Smoker.ratio	Smoker.ratio	ratio
0.0001064	0.0006106	5.738
0.001124	0.002405	2.139
0.004904	0.0072	1.468
0.01083	0.01469	1.356
0.0212	0.01918	0.9047

b) Model: $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ikk}$, where α are the levels of the age groups and β is an indicator for smoking. $\alpha_1 = \beta_1 = 0$ where μ is mean response of the 35-44 age group of non-smokers. The model assumes constant rates by age because there is no interaction term specified in the model.

c) The oldest group indicates that smokers are less likely to die of coronary death than non-smokers. Model: $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha_i\beta_j) + e_{ikk}$

$$\begin{aligned} y_{20} =& \mu + \alpha_2 + \beta_0 + (\alpha_2\beta_0) \\ =& \mu + (\text{35-44}) + 0 + ((\text{35-44})0) \\ =& \mu + (\text{35-44}) \end{aligned}$$

$$\begin{aligned} y_{21} =& \mu + \alpha_2 + \beta_1 + (\alpha_2\beta_1) \\ =& \mu + (\text{35-44}) + \beta_1 + ((\text{35-44})\beta_1) \end{aligned}$$

d) There are not enough degrees of freedom to fit an interaction term so the first model is better. The deviance/df ratio of the first model is low ~ .39, but the deviance in the second model is 0 because there is no error due to not having anymore degrees of freedom. Age appears to be the primary contributor on the count of coronary cases. The smoking variable appears to be insignificant. If we had more observations we could test the interaction and maybe see a different result.

```
(mdl.1 = glm(Rate ~ Age + Smoker, family = poisson(link = "log"), data = dta))
```

```
Call:  glm(formula = Rate ~ Age + Smoker, family = poisson(link = "log"), 
    data = dta)

Coefficients:
(Intercept)     Age45-54     Age55-64     Age65-74     Age75-84  
    -1.1004       1.5937       2.8261       3.5721       4.0312  
  Smokeryes  
     0.1441  

Degrees of Freedom: 9 Total (i.e. Null);  4 Residual
Null Deviance:      74.16 
Residual Deviance: 1.565    AIC: Inf
```

```
(mdl.2 = glm(Rate ~ Age*Smoker, family = poisson(link = "log"), data = dta))
```

```
Call:  glm(formula = Rate ~ Age * Smoker, family = poisson(link = "log"), 
    data = dta)

Coefficients:
       (Intercept)            Age45-54            Age55-64  
           -2.2403              2.3575              3.8303  
          Age65-74            Age75-84           Smokeryes  
            4.6228              5.2945              1.7470  
Age45-54:Smokeryes  Age55-64:Smokeryes  Age65-74:Smokeryes  
           -0.9868             -1.3630             -1.4424  
Age75-84:Smokeryes  
           -1.8472  

Degrees of Freedom: 9 Total (i.e. Null);  0 Residual
Null Deviance:      74.16 
Residual Deviance: 3.732e-15    AIC: Inf
```

Additional 1

The color and weight model had the lowest AIC of all of the models. All of the models had a deviance/df ratio over 3 indicating that the poisson regression model may not be appropriate.

Table 3: Partial SAS Output for color-weight model

Criterion	DF	Value	Value.DF
Deviance	168	551.8	3.28
Scaled Deviance	168	551.8	3.28
AIC	0	917.1	0

Additional 2

The negative binomial model with weight as the parameter is the best model because its AICc is the lowest.

Table 4: SAS Output of Model Comparison

Model	AICc
Poisson Weight	762.1
Poisson Intercept Only	767.2
Negative Binom Weight	740.3
Negative Binom Intercept Only	744.8

Code for Additional 2

```
proc genmod data=sasuser.crab;
model satell = weight / dist=zip link=log;
zeromodel;
title 'poisson weight';
run;

proc genmod data=sasuser.crab;
model satell = / dist=zip link=log;
zeromodel;
title 'poisson intercept only';
run;

proc genmod data=sasuser.crab;
model satell = weight / dist=zinb link=log;
zeromodel;
title 'negative binom weight';
run;

proc genmod data=sasuser.crab;
model satell = / dist=zinb link=log;
zeromodel;
```

```
title 'negative binom intercept only';
run;
```

Code for Additional 1

```
proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = / dist=poi link=log;
title 'intercept only';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = color / dist=poi link=log;
title 'color';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = spine / dist=poi link=log;
title 'spine';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = width / dist=poi link=log;
title 'width';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = weight / dist=poi link=log;
title 'weight';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = color spine / dist=poi link=log;
title 'color spine';
run;
```

```
proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = color width / dist=poi link=log;
title 'color width';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = color weight / dist=poi link=log;
title 'color weight';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = spine width / dist=poi link=log;
title 'spine width';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = spine weight / dist=poi link=log;
title 'spine weight';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = width weight / dist=poi link=log;
title 'width weight';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = color spine width / dist=poi link=log;
title 'color spine width';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = color spine weight / dist=poi link=log;
title 'color spine weight';
```

```
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = color width weight / dist=poi link=log;
title 'color width weight';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = spine width weight / dist=poi link=log;
title 'spine width weight';
run;

proc genmod data=sasuser.crab;
ods select Modelfit;
class color spine;
model satell = color spine width weight / dist=poi link=log;
title 'color spine width weight';
run;
```