Solution to HW4

Problem 3.11

(a) Under model $\log(\mu) = \alpha + \beta x$, $x = 1$ for treatment B and $x = 0$ for treatment A, we have:

$$\log(\mu_A) = \alpha + \beta \times 0 = \alpha, \quad \log(\mu_B) = \alpha + \beta \times 1 = \alpha + \beta.$$

So $\beta = \log(\mu_B) - \log(\mu_A) = \log(\mu_B/\mu_A)$, or equivalently, $e^{\beta} = \mu_B/\mu_A$.

(b) We used the following SAS program to fit the above model assuming data to follow a Poisson distributions for each treatment:

```
data trta;
  input y @@;
  x=0;
  datalines;
  8 7 6 6 3 4 7 2 3 4
  ;

data trtb;
  input y @@;
  x=1;
  datalines;
  9 9 8 14 8 13 11 5 7 6
  ;

data prob3_11; set trta trtb;
run;

proc genmod;
  model y = x / dist=poi link=log type3;
run;
```

```
*****************************************************************************
                     Criteria For Assessing Goodness Of Fit

        Criterion                     DF           Value        Value/DF

        Deviance                      18         16.2676          0.9038
        Scaled Deviance               18         16.2676          0.9038
        Pearson Chi-Square            18         16.0444          0.8914
        Scaled Pearson X2             18         16.0444          0.8914

                Analysis Of Maximum Likelihood Parameter Estimates

                          Standard      Wald 95%              Wald
 Parameter  DF  Estimate     Error   Confidence Limits  Chi-Square  Pr > ChiSq

 Intercept   1    1.6094    0.1414    1.3323    1.8866      129.51      <.0001
 x           1    0.5878    0.1764    0.2421    0.9335       11.11      0.0009
 Scale       0    1.0000    0.0000    1.0000    1.0000

                       LR Statistics For Type 3 Analysis

                                            Chi-
                    Source           DF     Square    Pr > ChiSq

                    x                 1      11.59        0.0007
```

The fitted equation is

$$\log(\widehat{\mu}) = 1.6094 + 0.5878x.$$

Interpretation of $\widehat{\beta} = 0.5878$: $\widehat{\mu}_B/\widehat{\mu}_A = e^{0.5878} = 1.8$. Compared to treatment A, treatment B produced wafers with the average number of imperfections 80% more than treatment A.

(c) Since $H_0 : \mu_A = \mu_B$ is equivalent to $H_0 : \beta = 0$. The Wald test is $\chi^2 = (0.5878/0.1764)^2 = 11.11$ with P-value $= P(\chi^2_1 \geq 11.11) = 0.0009$.

Note: In order to get the LRT for $H_0 : \beta = 0$, we either have to use option `type3` in the `model` statement or run the null (model without `x`) and then calculate the LRT. The LRT is $\chi^2 = 11.59$, basically the same as the Wald test.

(d) A 95% Wald CI for β is $[0.2421, 0.9335]$, so a 95% Wald CI for μ_B/μ_A is $[e^{0.2421}, e^{0.9335}] = [1.27, 2.54]$.

Note: If a LR CI for μ_B/μ_A is preferred, then we can get a LR CI for β using the option `lrci` in the `model` statement. Exponentiating both ends gives a LR CI for μ_B/μ_A.

(e) The Pearson estimate of the possible over-dispersion parameter is $\widehat{\phi} = 0.8914$, indicating no over-dispersion. Therefore, it is reasonable to assume the Poisson distribution for the data give each treatment.

Note: Here the GLM is a saturated model (perfect fit to the data). Therefore, we can use the Pearson Chi-square statistic or the Deviance divided by the *df* to see if there is an over-dispersion.

(f) The SAS program and (part of) output for the model without treatment is

```
proc genmod;
  model y = / dist=poi link=log;
run;
****************************************************************************
                    Criteria For Assessing Goodness Of Fit

        Criterion                    DF           Value        Value/DF

        Deviance                     19         27.8570          1.4662
        Scaled Deviance              19         27.8570          1.4662
        Pearson Chi-Square           19         27.7143          1.4586
        Scaled Pearson X2            19         27.7143          1.4586
```

There is a strong evidence of over-dispersion. For example, the Pearson estimate of the over-dispersion parameter is $\widehat{\phi} = 1.46$, much larger than 1. Therefore, it is very unlikely that the pooled data is from a Poisson distribution.

Problem 3.13

(a) We can fit a GLM to the crab data with the log link using the following SAS program

```
title "Model 1: Analysis of crab data using Poisson dist with log link";
proc genmod data=crab;
  model satell = weight / dist=poi link=log scale=pearson;
run;
```

```
**********************************************************************
                    Criteria For Assessing Goodness Of Fit

        Criterion                      DF           Value        Value/DF

        Deviance                       171        560.8664         3.2799
        Scaled Deviance                171        178.9679         1.0466
        Pearson Chi-Square             171        535.8957         3.1339
        Scaled Pearson X2              171        171.0000         1.0000

              Analysis Of Maximum Likelihood Parameter Estimates

                           Standard       Wald 95%             Wald
 Parameter  DF  Estimate      Error   Confidence Limits  Chi-Square  Pr > ChiSq

 Intercept   1   -0.4284     0.3168   -1.0493    0.1924        1.83      0.1762
 weight      1    0.5893     0.1151    0.3637    0.8149       26.21      <.0001
 Scale       0    1.7703     0.0000    1.7703    1.7703
```

The fitted model is

$$\log(\widehat{\mu}) = -0.4284 + 0.5893wt,$$

where `wt` is the weight (in kg) of a female crab.

(b) When wt=2.44kg, the estimated mean of Y is

$$\widehat{\mu} = e^{-0.4284+0.5893\times 2.44} = 2.74.$$

(c) The interpretation of $\widehat{\beta} = 0.5893$: for female crabs that are 1kg heavier, their mean number of satellites increases by $e^{0.5893} - 1 = 0.80 = 80\%$ (or multiplicatively by 1.80).

95% Wald CI for β: [0.3637, 0.8149] (taken from the output)

95% Wald CI for e^{β}: $[e^{0.3637}, e^{0.8149}] = [1.44, 2.26]$.

(d) The Wald test (taking into account over-dispersion) is $\chi^2 = (0.5893/0.1151)^2 = 26.21$. The P-value < 0.001. We reject the null that the number of satellites of a female crab is independent of her weight.

(e) We fit two models assuming no over-dispersion and get the LRT under $LRT_0 = 2(71.9524 - 35.9898) = 71.9252$. The Pearson estimate of the over-dispersion under the full model is $\widehat{\phi} = 3.1339$. So the correct LRT test is

$$LRT_1 = \frac{LRT_0}{\widehat{\phi}} = \frac{71.9252}{3.1339} = 22.95.$$

Under $H_0 : \beta = 0$, the corrected $LRT_1 \sim \chi^2_1$. The conclusion is similar to the Wald test. We reject the null that the number of satellites of a female crab is independent of her weight.

Note: This correct LRT statistic can be obtained using option `type3` in the `model` statement together with option `scale=pearson`.

Problem 3.18

(a) We plot the ratio of the number of arrest to the total attendance against the total attendance in the following

```
options ls=80 ps=25 nodate;

data prob3_18;
  input att arrest;
  logatt = log(att);
  ratio = arrest/att;
  cards;
  404 308
  286 197
  443 184
  169 149
  222 132
  150 126
  321 110
  189 101
  258 99
  223 81
  211 79
  215 78
  108 68
  210 67
  224 60
  211 57
  168 55
  185 44
  158 38
  429 35
  226 29
  150 20
  148 19
;

proc plot;
  plot ratio*att;
run;
********************************************************************************
             Plot of ratio*att.  Legend: A = 1 obs, B = 2 obs, etc.
  1.0 +
      |
      |           A   A
ratio |                                                              A
      |                                      A
      |   A
      |                         A
  0.5 +                   A
      |                                                                      A
      |               A       AA A      A           A
      |                       B  A
      |             A    A
      |           B              A
      |                                                                   A
  0.0 +
      --+---------+---------+---------+---------+---------+---------+---------+-
       100       150       200       250       300       350       400       450

                                          att
```

We see in this plot that there is no obvious pattern between Y/t and t. So it is reasonable to fit the model $E(Y)/t = \mu$, or equivalently, $E(Y) = \mu t$. This model is also equivalent to $\log\{E(Y)/t\} = \alpha$, where $\alpha = \log(\mu)$.

(b) The above model is equivalent to

$$\log\{E(Y)\} = \log(t) + \alpha.$$

This is a GLM with the log link and an offset $\log(t)$ but without a covariate. Assuming a Poisson distribution for Y, we got $\widehat{\alpha} = -0.9103 (SE = 0.022)$ so $\widehat{\mu} = e^{-0.9103} = 0.4024$. The following is the program (option `r` in `model` statement made a request to output residuals).

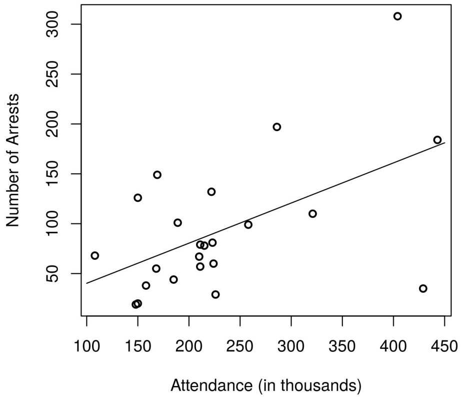
```
proc genmod;
  model arrest = / dist=poi offset=logatt link=log r;
run;
*************************************************************************
                     Criteria For Assessing Goodness Of Fit

        Criterion                      DF           Value        Value/DF

        Deviance                       22        669.4458         30.4294
        Scaled Deviance                22        669.4458         30.4294
        Pearson Chi-Square             22        658.4846         29.9311
        Scaled Pearson X2              22        658.4846         29.9311

              Analysis Of Maximum Likelihood Parameter Estimates

                           Standard       Wald 95%             Wald
 Parameter  DF  Estimate      Error   Confidence Limits   Chi-Square  Pr > ChiSq

 Intercept   1   -0.9103     0.0216   -0.9527   -0.8679      1769.91      <.0001
 Scale       0    1.0000     0.0000    1.0000    1.0000


                                                   Std        Std
                     Raw    Pearson   Deviance   Deviance    Pearson  Likelihood
 Observation    Residual   Residual   Residual   Residual   Residual    Residual

           1   145.42577  11.405531  10.136615  10.545888  11.866039   10.652121
           2   81.910324  7.6352007  6.9246837   7.119132  7.8496006   7.1603899
           3   5.7317253  0.4292876  0.4270174  0.4460357   0.448407   0.4462341
           4   80.992464  9.8212348  8.4702274  8.6083758  9.9814179   8.6554493
           5   42.664657  4.5139485  4.2115954  4.3025302  4.6114115   4.3158917
           6   65.638282  8.4484377  7.3605021  7.4667607   8.570402   7.5001789
           7   -19.17408  -1.687045  -1.731598  -1.786458  -1.740494   -1.783712
           8   24.944235  2.8602508  2.7220537  2.7718491  2.9125743   2.7769823
           9   -4.822155  -0.473256  -0.476992  -0.489025  -0.485195    -0.48884
          10   -8.737755  -0.922385  -0.937996  -0.958343  -0.942393   -0.957678
          11   -5.908817  -0.641245  -0.648907  -0.662202  -0.654383   -0.661893
          12   -8.518463  -0.915813  -0.931494  -0.950952  -0.934944   -0.950309
          13   24.539563   3.722372  3.4354303  3.4709226  3.7608288   3.4770622
          14   -17.50641  -1.904374  -1.976552  -2.016851  -1.943201   -2.013989
          15   -30.14017  -3.174581  -3.381978   -3.45568  -3.243763      -3.447
          16   -27.90882  -3.028761  -3.222669  -3.288697  -3.090816   -3.281058
          17   -12.60512  -1.533054  -1.584829  -1.610521  -1.557906   -1.608882
          18   -30.44612  -3.528669  -3.822862  -3.891274  -3.591817   -3.881226
          19   -25.58101  -3.208145  -3.470251  -3.523082  -3.256985   -3.515452
          20   -137.6345  -10.47523   -12.7891  -13.33951  -10.92606    -13.1609
          21   -61.94499  -6.495563  -7.589358  -7.756275  -6.638423   -7.711983
          22   -40.36172  -5.195039   -6.04471  -6.131974  -5.270037   -6.109286
          23    -40.5569  -5.255314  -6.139955  -6.227387  -5.330148   -6.204128
```

The interpretation of $\widehat{\mu} = 0.4024$: For every 10,000 attendance, on average there were approximately 4 arrests.

Note: We can directly fit $E(Y) = \mu t$. This is a GLM with the identity link but without an intercept, which can be fit using the following SAS program. The results will be the same:

```
proc genmod;
  model arrest = att / dist=poi link=identity noint r;
run;
```

(c) The plot is given in the following:



The teams with absolute standardized Pearson residuals greater than (or close to) 2 are: 11.9 for observation 1, 7.8 for observation 2, 10 for observation 4, 4.6 for observation 5, 8.6 for observation 6, 2.9 for observation 8, 3.8 for observation 13, -1.9 for observation 14, -3.2 for observation 15, -3.1 for observation 16, -3.6 for observation 18, -3.3 for observation 19, -11 for observation 20, -6.6 for observation 21, -5.3 for observation 22, -5.3 for observation 23.

Note: We got too many outliers because the data does not have a Poisson distribution. Specifically, there is a lot of over-dispersion in the data. After we take into account the over-dispersion, there are only two outliers, observation 1 and observation 20.

(d) Assuming Y to have a negative binomial distribution, we got $\widehat{\alpha} = -0.9052$ with $SE = 0.12$, and the dispersion parameter is estimated to be $\widehat{D} = 0.32 > 0$. Even though we got almost the same estimate for α, the standard error estimate of $\widehat{\alpha}$ by assuming a negative binomial distribution is much larger than the SE by assuming a Poisson distribution (0.022). The estimated dispersion parameter indicates that $\text{var}(Y|x) = \mu + 0.32\mu^2$. Therefore, it is not appropriate to assume a Poisson distribution for the data. We need to use over-dispersion Poisson or negative binomial distribution for Y in making inference.

Problem 3.20

(a) The death rates (defined as the number of deaths per 1000 person-years) for smokers and non-smokers and their ratio for each age group:

```
data prob3_20;
  input age smoke pyear death;
  pyear=pyear/1000;
  smoke1=(smoke=1);
  smoke0=(smoke=0);
  drate = death/pyear;
  age1=(age=1);
  age2=(age=2);
  age3=(age=3);
  age4=(age=4);
  age5=(age=5);
  logpy = log(pyear);
  cards;
  1 0 18793 2
  2 0 10673 12
  3 0 5710 28
  4 0 2585 28
  5 0 1462 31
  1 1 52407 32
  2 1 43248 104
  3 1 28612 206
  4 1 12663 186
  5 1 5317 102
  ;

 data smoke; set prob3_20;
   if smoke=1;
   drate1 = drate;
   drop drate;
run;

data nonsmoke; set prob3_20;
   if smoke=0;
   drate0 = drate;
   drop drate;
run;

data new; merge smoke nonsmoke;
  lograte0=log(drate0);
  lograte1=log(drate1);
  rratio = drate1/drate0;
run;

proc print;
  var age drate1 drate0 rratio;
run;
*************************************************************
```

```
                 Obs    age     drate1      drate0      rratio

                   1      1      0.6106      0.1064    5.73755
                   2      2      2.4047      1.1243    2.13881
                   3      3      7.1998      4.9037    1.46824
                   4      4     14.6885     10.8317    1.35606
                   5      5     19.1838     21.2038    0.90473
```

where age=1 is for the first age group (35-44), age=2 for the second age group (45-54), etc, `drate1` and `drate0` are death rates for smokers and non-smokers, `rratio` is the ratio between `drate1` and `drate0`. From this table, we see that the death rates for both smokers and non-smokers increase as age increases, that smokers have a higher death rate than non-smokers for each age group, except probably the last age group (75-84). However, the magnitude of the difference in terms of ratio of death rates declines as age increases.

(b) The model for log rates (λ) having 4 parameters for age and one parameter for smoking is

$$\log(\lambda) = \beta_0 + \beta_1 smoke + \beta_2 age2 + \beta_3 age3 + \beta_4 age4 + \beta_5 age5$$

where `smoke` is the dummy variable for smokers, `age2` is the dummy variable for age group 2, `age3` is the dummy variable for age group 3, etc. For smokers and non-smokers in the same age group, we have

$$\log\{\lambda(smoke = 1, age) - \log\{\lambda(smoke = 0, age)\} = \beta_1 \Rightarrow \frac{\lambda(smoke = 1, age)}{\lambda(smoke = 0, age)} = e^{\beta_1}.$$

That is, this model implies that the ratios of death rates between smokers and non-smokers for all age group are the same (independent of age). However, we see in (a) that the sample ratio of death rates between smokers and non-smokers declines as age increases, indicating this model is not appropriate.

(c) Based on (a), we see that the death rates increase as age increase for both smokers and non-smokers, but the rates of increase are different (also reflected in the ratio of the death rates). This may indicate a log rate model with main effects of age and smoking and their interaction:

$$\log(\lambda) = \beta_0 + \beta_1 age + \beta_2 smoke + \beta_3 age \times smoke.$$

Based on this model, we have

$$\begin{aligned} & \log\{\lambda(smoke = 1, age) - \log\{\lambda(smoke = 0, age) = \beta_2 + \beta_3 age \\ \Rightarrow \quad & \log\{\lambda(smoke = 1, age)/\lambda(smoke = 0, age)\} = \beta_2 + \beta_3 age, \end{aligned}$$

a linear function of age.

(d) First we use the following SAS program to fit the model in (b) where `logpy` is the log of person-years (in 1000 person-years).

```
proc genmod data=prob3_20;
  model death = smoke age2 age3 age4 age5
       / dist=poi link=log offset=logpy;
run;
```

```
***************************************************************************
                         Criteria For Assessing Goodness Of Fit

        Criterion                    DF           Value         Value/DF

        Deviance                      4         12.1339           3.0335
        Scaled Deviance               4         12.1339           3.0335
        Pearson Chi-Square            4         11.1565           2.7891
        Scaled Pearson X2             4         11.1565           2.7891
```

```
              Analysis Of Maximum Likelihood Parameter Estimates

                              Standard      Wald 95%          Wald
Parameter  DF  Estimate      Error   Confidence Limits  Chi-Square  Pr > ChiSq

Intercept   1   -1.0116     0.1918   -1.3874   -0.6358       27.83      <.0001
smoke       1    0.3545     0.1074    0.1441    0.5650       10.90      0.0010
age2        1    1.4840     0.1951    1.1016    1.8664       57.86      <.0001
age3        1    2.6275     0.1837    2.2674    2.9876      204.53      <.0001
age4        1    3.3505     0.1848    2.9883    3.7127      328.72      <.0001
age5        1    3.7001     0.1922    3.3234    4.0769      370.54      <.0001
Scale       0    1.0000     0.0000    1.0000    1.0000
```

Based on this model (even though it is not appropriate), the estimated ratio of the death rates between smokers and non-smokers is $e^{0.3545} = 1.43$. That is, smokers are 43% more likely to die no matter what age group they are in.

With scores {1,2,3,4,5} assigned to the age group, we fit the model in (c) using the following SAS program:

```
proc genmod data=prob3_20;
  model death = age smoke age*smoke
       / dist=poi link=log offset=logpy;
run;
```

```
********************************************************************************
                   Criteria For Assessing Goodness Of Fit

        Criterion                     DF           Value        Value/DF

        Deviance                       6         59.8953          9.9825
        Scaled Deviance                6         59.8953          9.9825
        Pearson Chi-Square             6         56.1029          9.3505
        Scaled Pearson X2              6         56.1029          9.3505
              Analysis Of Maximum Likelihood Parameter Estimates

                              Standard      Wald 95%          Wald
Parameter  DF  Estimate      Error   Confidence Limits  Chi-Square  Pr > ChiSq

Intercept   1   -1.9594     0.3057   -2.5585   -1.3603       41.09      <.0001
age         1    1.0468     0.0774    0.8951    1.1986      182.76      <.0001
smoke       1    1.2837     0.3258    0.6450    1.9223       15.52      <.0001
age*smoke   1   -0.2490     0.0836   -0.4128   -0.0852        8.87      0.0029
Scale       0    1.0000     0.0000    1.0000    1.0000
```

Based on this model, we estimated that

$$\log\{\widehat{\lambda}(smoke = 1, age = 1)/\widehat{\lambda}(smoke = 0, age = 1)\} = 1.284 - 0.249 age.$$

For smokers and non-smokers in the first age group (35-44), the ratio of the death rates are:

$$\frac{\widehat{\lambda}(smoke = 1, age = 1)}{\widehat{\lambda}(smoke = 0, age = 1)} = e^{1.284-0.249\times 1} = 2.82.$$

That is, smokers from 35-44 are 182% more likely to die than non-smokers in the same age group. However, as age increases, this difference gets smaller and smaller. For example, for smokers and non-smokers in the last age group (75-84), the ratio of the death rates are:

$$\frac{\widehat{\lambda}(smoke = 1, age = 5)}{\widehat{\lambda}(smoke = 0, age = 5)} = e^{1.284-0.249\times 5} = 1.04.$$

That is, smokers from 75-84 are only 4% (probably not significant) more likely to die than non-smokers in the same age group.

The deviance for the model in (b) is $\chi^2_D = 12.14$ with $df = 4$. The deviance for the model in (c) is $\chi^2_D = 59.90$ with $df = 6$. These statistics indicate that neither model fits the data well (under Poisson distributional assumption for the data). But model (b) fits the data better than model (c).

(e) Including the quadratic age effect both for smokers and non-smokers, we got

```
proc genmod data=prob3_20;
  model death = age age*age smoke age*smoke age*age*smoke
       / dist=poi link=log offset=logpy;
run;
```

```
*****************************************************************************
                     Criteria For Assessing Goodness Of Fit

        Criterion                     DF           Value        Value/DF

        Deviance                       4          1.2459          0.3115
        Scaled Deviance                4          1.2459          0.3115
        Pearson Chi-Square             4          1.2143          0.3036
        Scaled Pearson X2              4          1.2143          0.3036

              Analysis Of Maximum Likelihood Parameter Estimates

                                      Standard   Wald 95% Confidence          Wald
  Parameter       DF    Estimate       Error           Limits           Chi-Square

  Intercept        1     -4.3550      0.9056    -6.1300     -2.5800          23.13
  age              1      2.6830      0.5440     1.6168      3.7492          24.32
  age*age          1     -0.2421      0.0774    -0.3939     -0.0903           9.78
  smoke            1      1.9742      0.9548     0.1028      3.8457           4.28
  age*smoke        1     -0.6572      0.5774    -1.7889      0.4745           1.30
  age*age*smoke    1      0.0511      0.0828    -0.1112      0.2134           0.38
  Scale            0      1.0000      0.0000     1.0000      1.0000
```

Under Poisson distributional assumption for the data, this model fits the data much better. The deviance statistic is $\chi^2_D = 1.2459$ with $df = 4$, P-value = 0.87.

Note: It seems that there is no interaction between age and smoking since both interaction terms are not significant. However, after we remove age*age*smoke from the model, age*smoke is significant (P-value=0.0015) with deviance $\chi^2_D = 1.64$ with $df = 5$ (good fit).

```
proc genmod data=prob3_20;
  model death = age age*age smoke age*smoke
       / dist=poi link=log offset=logpy;
run;
```

```
********************************************************************
                     Criteria For Assessing Goodness Of Fit

        Criterion                     DF           Value        Value/DF

        Deviance                       5          1.6358          0.3272
        Scaled Deviance                5          1.6358          0.3272
        Pearson Chi-Square             5          1.5506          0.3101
        Scaled Pearson X2              5          1.5506          0.3101
```

```
                    Analysis Of Maximum Likelihood Parameter Estimates

                               Standard      Wald 95%              Wald
Parameter   DF   Estimate        Error   Confidence Limits   Chi-Square   Pr > ChiSq

Intercept    1    -3.8841       0.4501   -4.7662    -3.0020        74.48       <.0001
age          1     2.3765       0.2079    1.9689     2.7841       130.61       <.0001
age*age      1    -0.1977       0.0274   -0.2513    -0.1440        52.17       <.0001
smoke        1     1.4410       0.3722    0.7115     2.1705        14.99       0.0001
age*smoke    1    -0.3076       0.0970   -0.4978    -0.1174        10.05       0.0015
Scale        0     1.0000       0.0000    1.0000     1.0000
```

Problem 3.22

(a) True; (b) False; (c) False.