4/13/2010 36-402/608 ADA-II H. Seltman

Breakout #22: Poisson Regression

In R, Poisson regression is performed using
`result = glm(y ~ x..., data=my.dtf, family=poisson)`
where "y" is a count, and "x..." is any prediction formula.

As usual `summary(result)` has the standard errors and p-values, as well as AIC (as `$aic`).

The `glm()` object has a `$deviance` component that can be used for the likelihood ratio test. E.g., to compare glm objects named "full" and "reduced" use:

```
p.val = 1 - pchisq(reduced$deviance - full$deviance, reduced$df.res - full$df.res)
```

Sometimes a reasonable alternative to Poisson regression is linear regression on a transformed outcome in the form of square root of counts. If the residual plots look OK, you can go with that model.

Use `family=quasipoisson` to check for under/over dispersion.

The analysis shown here is from problem 24 of chapter 22 of The Sleuth and represents valve characteristics and number of failures from a nuclear reactor. See the factor coding statements to get some idea of the valve characteristics. The number of failures is modeled as Poisson.

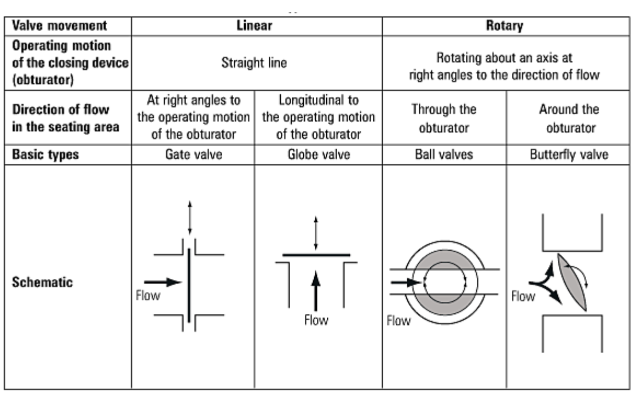
It is important to separate Poisson regression problems into those where the different units studied have equal exposure (in time or space) vs. those with unequal exposure. The latter can only be modeled if the extent of exposure is recorded also.

E.g., if $\log(\mu_i|x_i) = \beta_0 + \beta_1 x_i$ for "unit" exposure, then for exposure t_i we expect $\log(\mu_i/t_i|x_i) = \beta_0 + \beta_1 x_i$ which implies $\log(\mu_i) - \log(t_i) = \beta_0 + \beta_1 x_i$ which implies $\log(\mu_i) = \beta_0 + \beta_1 x_i + 1.0 * \log(t_i)$. In other words we can use the usual Poisson regression model if we include log(exposure) as an explanatory variable with a fixed, known coefficient of 1.0. This is done in R (and other programs) by setting the "offset" to log(exposure).

```
valve=read.csv("ex2224.csv")
names(valve)=casefold(names(valve))
summary(valve)
#      system         operator         valve            size
#  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
#  1st Qu.:3.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:1.250
#  Median :3.000   Median :2.500   Median :4.000   Median :2.000
#  Mean   :3.422   Mean   :2.189   Mean   :3.856   Mean   :1.967
#  3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:2.000
#  Max.   :5.000   Max.   :4.000   Max.   :6.000   Max.   :3.000
```

```
#      mode          failures          time
#  Min.   :1.000   Min.   : 0.000   Min.   : 1.000
#  1st Qu.:1.000   1st Qu.: 0.000   1st Qu.: 1.000
#  Median :2.000   Median : 0.000   Median : 2.500
#  Mean   :1.578   Mean   : 1.611   Mean   : 4.344
#  3rd Qu.:2.000   3rd Qu.: 2.000   3rd Qu.: 4.000
#  Max.   :2.000   Max.   :23.000   Max.   :36.000
valve$operator=factor(valve$operator, labels=c("air","solenoid","motor","Manual"))
valve$system=factor(valve$system, labels=c("contain","nuclear","power","safety","aux"))
valve$valve=factor(valve$valve, labels=c("ball","Butterfly","diaphragm","gate",
                   "Globe","Dir"))
valve$mode=factor(valve$mode, labels=c("closed","open"))
valve$sizeGroup=factor(valve$size, labels=c("small","medium","large"))

nrow(valve) # 90
valve[1:5,]
#    system operator     valve size   mode failures time sizeGroup
# 1 contain    motor      gate    3 closed        2    4     large
# 2 contain    motor      gate    3   open        2    4     large
# 3 contain    motor     Globe    1 closed        1    2     small
# 4 nuclear      air Butterfly    2   open        0    2    medium
# 5 nuclear      air diaphragm    2 closed        0    2    medium
```



Valve movement	Linear		Rotary	
Operating motion of the closing device (obturator)	Straight line		Rotating about an axis at right angles to the direction of flow	
Direction of flow in the seating area	At right angles to the operating motion of the obturator	Longitudinal to the operating motion of the obturator	Through the obturator	Around the obturator
Basic types	Gate valve	Globe valve	Ball valves	Butterfly valve
Schematic	Flow	Flow	Flow	Flow

Question 1: What would have happened in our analyses if we hadn't used factor()? Forgetting to treat (unordered) categorical variables to factors is one of the most common causes of a bad analysis in R. To do so causes R to make the arbitrary factor levels (e.g., valve types) ordered and equally spaced.

```
v1=glm(failures~system+operator+valve+size+mode, data=valve,
       family=poisson, offset=log(time))
summary(v1)
# Coefficients:    Estimate Std. Error z value Pr(>|z|)
# (Intercept)      -5.60059    0.90151  -6.212 5.22e-10 ***
# systemnuclear     0.84550    0.53347   1.585  0.11299
# systempower       0.81323    0.50904   1.598  0.11013
# systemsafety      0.88612    0.55155   1.607  0.10814
# systemaux        -0.05361    0.57345  -0.093  0.92552
# operatorsolenoid  0.74251    0.57904   1.282  0.19973
# operatormotor    -1.08856    0.25138  -4.330 1.49e-05 ***
# operatorManual   -2.31326    0.47677  -4.852 1.22e-06 ***
# valveButterfly    0.64644    0.74999   0.862  0.38872
# valvediaphragm    0.57828    0.78207   0.739  0.45965
# valvegate         3.12242    0.59809   5.221 1.78e-07 ***
# valveGlobe        1.81486    0.60894   2.980  0.00288 **
# valveDir          1.04705    0.94094   1.113  0.26581
# size              1.03790    0.18381   5.647 1.64e-08 ***
# modeopen         -0.05197    0.18286  -0.284  0.77624
# (Dispersion parameter for poisson family taken to be 1)
#     Null deviance: 385.53  on 89  degrees of freedom
# Residual deviance: 210.69  on 75  degrees of freedom
# AIC: 345.03
```

Question 2: What are our preliminary conclusions about valve failure? What specifically does the intercept tell us? What are some reasons that this model might be inadequate?

Compared to air operated valves (arbitrary baseline), motor and Manual operated valves have statistically significantly fewer failures. Compared to ball valves, gate and globe values have more failures. There is a significant effect of valve size that needs closer inspection.

The intercept is the expected log of the failure rate for small, normally closed, air-operated ball valves use for containment.

Size might be non-linear. We might need important interactions. We might have over-dispersion.

```
v1SG=glm(failures~system+operator+valve+sizeGroup+mode, data=valve,
         family=poisson, offset=log(time))
summary(v1SG)
# Coefficients:    Estimate Std. Error z value Pr(>|z|)
# (Intercept)      -3.76867    0.81935  -4.600 4.23e-06 ***
# systemnuclear     0.91556    0.53184   1.721  0.08516 .
# systempower       1.01881    0.50548   2.016  0.04385 *
# systemsafety      1.22309    0.55518   2.203  0.02759 *
# systemaux         0.33292    0.58408   0.570  0.56869
# operatorsolenoid  0.70437    0.56669   1.243  0.21389
# operatormotor    -1.19261    0.24851  -4.799 1.59e-06 ***
# operatorManual   -2.47233    0.47660  -5.187 2.13e-07 ***
# valveButterfly    0.18533    0.76105   0.244  0.80761
# valvediaphragm    0.60674    0.78107   0.777  0.43727
# valvegate         2.95894    0.60010   4.931 8.19e-07 ***
# valveGlobe        1.79318    0.61040   2.938  0.00331 **
# valveDir          1.00891    0.93009   1.085  0.27803
# sizeGroupmedium  -0.01219    0.28340  -0.043  0.96568
# sizeGrouplarge    1.61457    0.32104   5.029 4.93e-07 ***
# modeopen         -0.20934    0.19033  -1.100  0.27138
# (Dispersion parameter for poisson family taken to be 1)
#     Null deviance: 385.53  on 89  degrees of freedom
# Residual deviance: 195.68  on 74  degrees of freedom
# AIC: 332.02
```

Question 3: What's different and which model is more appropriate?

With the more appropriate coding of the three sizes as categorical, we don't make the assumption of linearity with size, and this assumptions is rejected by seeing that bMedium is very different from bLarge. We conclude that large valves have a higher failure rate than medium and small valves which do not have a significantly different failure rate from each other. Note that we now see some significant "system" effects; this can happen in an unbalanced, observational study as opposed to a designed experiment.

```
v1q=glm(failures~system+operator+valve+sizeGroup+mode, data=valve,
        family=quasipoisson, offset=log(time))
summary(v1q)
# Coefficients:    Estimate Std. Error t value Pr(>|t|)
# (Intercept)      -3.76867    1.74297  -2.162   0.0338 *
# systemnuclear     0.91556    1.13136   0.809   0.4210
# systempower       1.01881    1.07528   0.947   0.3465
# systemsafety      1.22309    1.18100   1.036   0.3037
# systemaux         0.33292    1.24248   0.268   0.7895
# operatorsolenoid  0.70437    1.20549   0.584   0.5608
# operatormotor    -1.19261    0.52864  -2.256   0.0270 *
# operatorManual   -2.47233    1.01385  -2.439   0.0171 *
# valveButterfly    0.18533    1.61895   0.114   0.9092
# valvediaphragm    0.60674    1.66153   0.365   0.7160
# valvegate         2.95894    1.27657   2.318   0.0232 *
# valveGlobe        1.79318    1.29848   1.381   0.1714
# valveDir          1.00891    1.97853   0.510   0.6116
# sizeGroupmedium  -0.01219    0.60286  -0.020   0.9839
# sizeGrouplarge    1.61457    0.68294   2.364   0.0207 *
# modeopen         -0.20934    0.40488  -0.517   0.6067
# (Dispersion parameter for quasipoisson family taken to be 4.525197)
#     Null deviance: 385.53  on 89  degrees of freedom
# Residual deviance: 195.68  on 74  degrees of freedom
# AIC: NA

1 - pchisq(summary(v1q)$dispersion * v1SG$df.res, v1SG$df.res) # 0
exp(-2.47233) # 0.084
```

Question 3: How do we know that the Poisson (variance=mean) model is inadequate? What do you conclude about valve failure after adjusting for extra-Poisson variation?

The p-value rejects a dispersion of 1. With appropriately wider confidence intervals and bigger p-value we now conclude that the significant effects are operator (motor and manual both better than air), valve (gate worse than ball) and size (large worse than small).

```
valve$operMan = factor(as.numeric(valve$operator),
                       levels=c(4,1,2,3),
                       labels=c("Manual","air","solenoid","motor"))
v1qM=glm(failures~system+operMan+valve+sizeGroup+mode, data=valve,
         family=quasipoisson, offset=log(time))
summary(v1qM)
# ...
# operManair       2.47233    1.01385   2.439  0.01715 *
# operMansolenoid  3.17669    1.55816   2.039  0.04505 *
# operManmotor     1.27971    1.06563   1.201  0.23362
```

Question 4: How does the code work to change the baseline for operator? What different conclusions can we now justify?

The “levels” changes the order so that the old level “4” is now first and the baseline. We must be very careful to give the “labels” in the order that really does match the “ levels” or our results will be misleading. Now we can specifically say that air and solenoid valves are worse than manual values while motor shows no statistically significant difference with manual , where before we couldn’t compare solenoid or motor operation to manual operation.

```
v1q9=glm(failures~system+mode+valve+operMan*sizeGroup, data=valve,
         family=quasipoisson, offset=log(time))
summary(v1q9)
# Coefficients: (1 not defined because of singularities)
#                                   Estimate Std. Error t value Pr(>|t|)
# ...
# operManair                         -0.4566     1.3939  -0.328   0.7442
# operMansolenoid                   -15.4342  2570.1681  -0.006   0.9952
# operManmotor                       -2.6733     1.9116  -1.398   0.1665
# sizeGroupmedium                    -2.5676     2.3771  -1.080   0.2838
# sizeGrouplarge                     -2.9967     2.4031  -1.247   0.2166
# operManair:sizeGroupmedium          2.3734     2.4639   0.963   0.3388
# operMansolenoid:sizeGroupmedium    18.8008  2570.1692   0.007   0.9942
# operManmotor:sizeGroupmedium        3.3524     2.8746   1.166   0.2475
# operManair:sizeGrouplarge           4.3476     2.4829   1.751   0.0844 .
# operMansolenoid:sizeGrouplarge          NA         NA      NA       NA
# operManmotor:sizeGrouplarge         5.7416     2.8598   2.008   0.0486 *
```

Question 5: The above (partial) results are for the only significant 2-way interaction. Why does one line have NA? How could you explain the significant interaction to a client?

The NA is because there were no large valves operated with solenoids in the dataset. The combination of large values with motor control (and perhaps large valves with air control) is associated with an increased failure rate.