Data Cleaning / Exploration and Feature Engineering

John Murdzek, FSA
Senior Experience Studies Actuary
Genworth Financial

Joe Long
Assistant Actuary and Data Scientist
Milliman, Minneapolis

March 21, 2018

ILTCI

Disclaimer

- The views and opinions expressed in this presentation are those of the presenter and do not reflect the official policy or position of Genworth Financial, Inc. or any of its subsidiaries.

Limitations

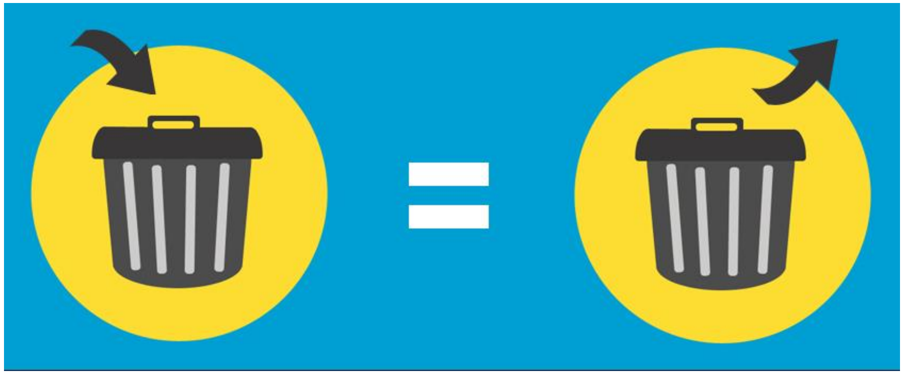
This presentation is intended for informational purposes only. It reflects the opinions of the presenter, and does not represent any formal views held by Milliman.

Milliman makes no representations or warranties regarding the contents of this presentation.

Milliman does not intend to benefit or create a legal duty to any recipient of this presentation.

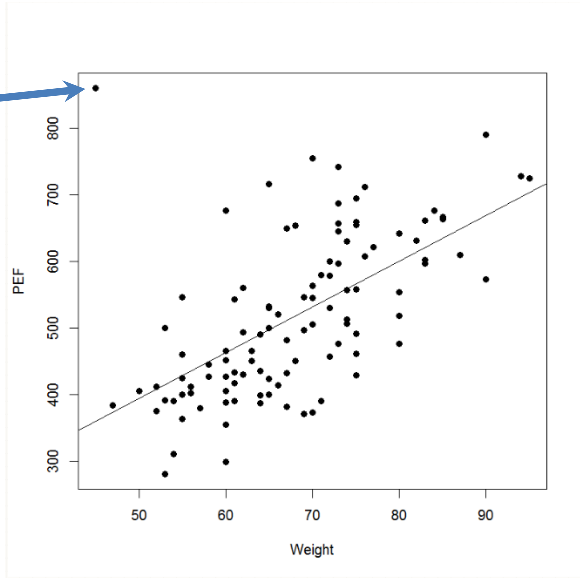
Overview

- Now we have data let’s start modeling right? Nope….
- Garbage in is garbage out, make sure your data is in good shape
- ~80% of your time will be spent here



Getting started

- It's important to make sure your data are clean and ready to go for your modeling project.
- Have you viewed your driver and response variables?
- Does your data have any outliers, or blanks?

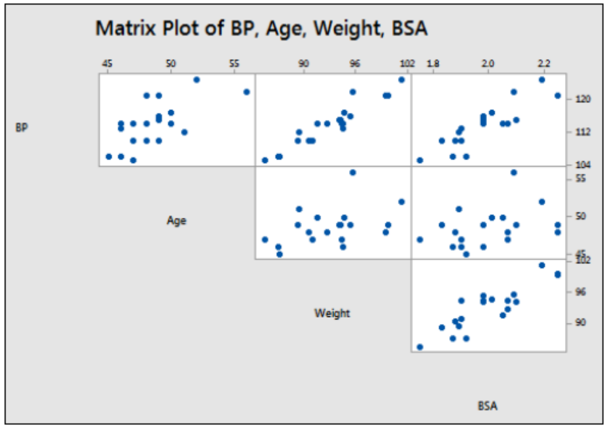


Source: http://stats.stackexchange.com/questions/194783/extreme-values-in-the-data

[Run R 0100: 1.0.1 – 1.0.7]

Data relationships

- If you notice patterns in your data that are strange find out why and try and fix them.
- Have you viewed relationships amongst your driver and response variables?
- Are some driver variables so related that they do not appear to add any value?



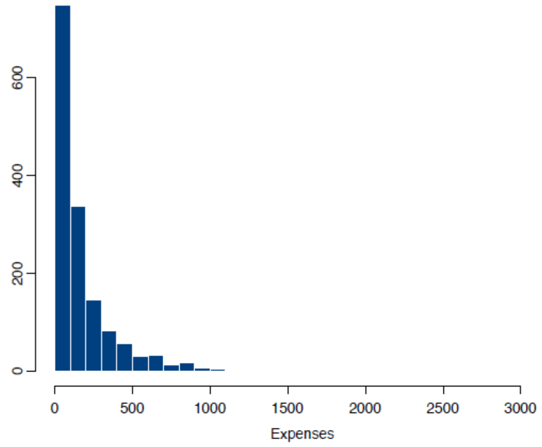
Source: https://onlinecourses.science.psu.edu/stat501/node/347

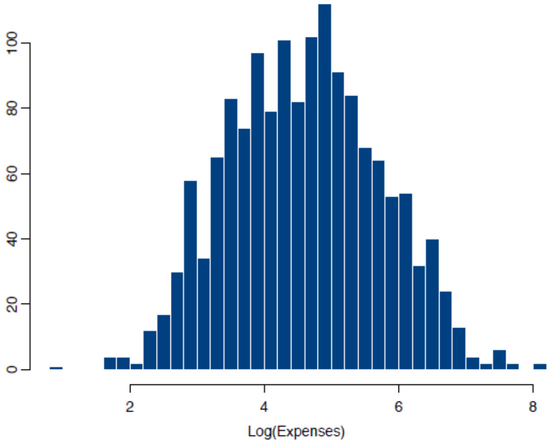
Moment of truth

- Sometimes your data may not be good enough quality to use in a modeling project.
- Are the data granular enough for your modeling purpose?
- Are the data credible enough? Can industry data help out?
- Do you find too many outliers, blanks, and / or other suspicious patterns?
- Do you find that certain relationships you expect to be true, are violated by simple views of your data? **[Run R 0100: 1.0.8 – 1.0.13]**

Variable transformations

- Data is clean – Can we make any additional variables?
- Start simple – We can always circle back.
- Would any transformations of your variables better serve you?





Source: http://www.kenbenoit.net/courses/ME104/logmodels2.pdf

Base level characteristics

- The most populous part of your data is what can be considered the "base level" characteristic.
- For example, if there are more claim terminations for females than males, then females would be the desirable base level characteristic of the gender driver variable.
- This concept will become more important in the next section.

"The base level should not be sparse. … Any level which is not sparse is an appropriate base level."

– Piet de Jong and Gillian Z. Heller, "Generalized Linear Models for Insurance Data"

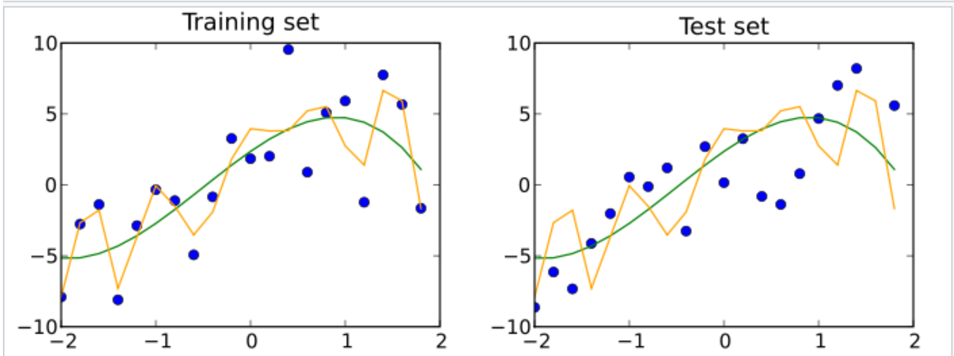
Key data fields

- Key data fields
  - GroupIndicator
  - Cov_Type_Bucket
  - ClaimDuration
  - Gender
  - ClaimType
  - Exposure
  - Terminations
  - Incurred_Year
  - IncurredAgeBucket
  - Region
  - Diagnosis_Category
  - TQ_Status
  - Infl_Rider_Bucket
  - Max_Ben_Bucket

Splitting the datasets

- Split data into training and validation data.
- What proportion of your data can you afford to segment out of your training data analysis?
- Can you afford to also make a test dataset?

[Run R 0150]



A training set (left) and a test set (right) from the same statistical population are shown as blue points. Two predictive models are fit to the training data. Both fitted models are plotted with both the training and test sets. In the training set, the MSE of the fit shown in orange is 4 whereas the MSE for the fit shown in green is 9. In the test set, the MSE for the fit shown in orange is 15 and the MSE for the fit shown in green is 13. The orange curve severely overfits the training data, since its MSE increases by almost a factor of four when comparing the test set to the training set. The green curve overfits the training data much less, as its MSE increases by less than a factor of 2.

Source: https://en.wikipedia.org/wiki/Test_set

Predictive Analytics Workshop

GLM / Poisson Modeling

John Murdzek, FSA
Senior Experience Studies Actuary
Genworth Financial

March 21, 2018

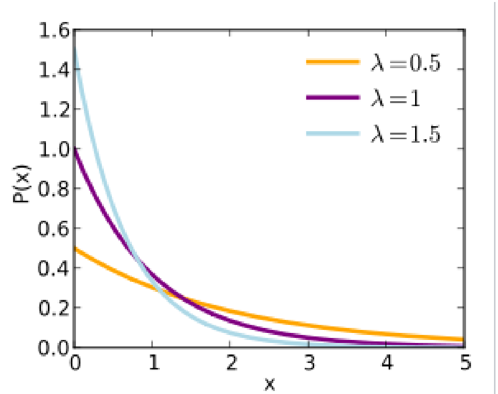
ILTCI

Disclaimer

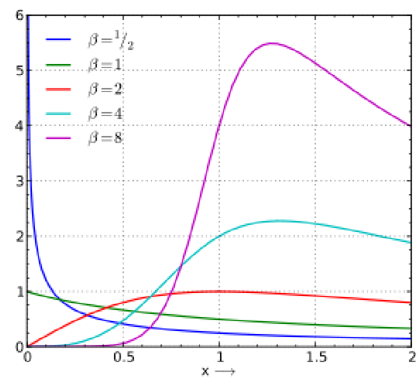
- The views and opinions expressed in this presentation are those of the presenter and do not reflect the official policy or position of Genworth Financial, Inc. or any of its subsidiaries.

Overview

- Start simple to get a feel for your data.
- Do you have some *a priori* notions as to what relationships should exist in your dataset?
- Do you have a sense as to how the hazard rate function might look?
- R can help you fit models based on your answers to the above.



Exponential failure density functions. Each of these has a (different) constant hazard function (see text).



Hazard function $h(t)$ plotted for a selection of log-logistic distributions.

Source: https://en.wikipedia.org/wiki/Failure_rate

Poisson GLM data preparation

- Initial test GLMs can be run to determine which variables are contributing most to explaining the variation of the CTRs.
- But first the data need to be prepared to make the Poisson regression possible.
- The data preparation steps condense what was done in the pre-work and create “factor” variables which are important to the GLM code.

[Run R 0240: 2.4.1 – 2.4.10]

Simple Poisson GLM results

- Fit gender, situs, and claim duration
- The residuals can help indicate over-dispersion.
- The "Estimates" are the GLM coefficients. Their exponentiated values produce the CTR of any effect that is desired.
- The standard errors and p-values tell you whether the estimate is significant.

[Run R 0240: 2.4.11 – 2.4.12]

```
+                                    chisq_stat = statistic ** 2)
> summary(results_ps)

Call:
glm(formula = SumTerminations ~ relevel(Gender.f, "Female") +
    relevel(ClaimType.f, "HHC") + relevel(ClmDurBucket.f, "Mos1-3"),
    family = poisson, data = ctr_data_ps, offset = LnExposure)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2132  -0.6157  -0.4565  -0.1808   4.1074

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -2.48972    0.03009 -82.745  < 2e-16
relevel(Gender.f, "Female")Male         0.29472    0.01845  15.970  < 2e-16
relevel(ClaimType.f, "HHC")ALF         -0.34411    0.02386 -14.422  < 2e-16
relevel(ClaimType.f, "HHC")NH           0.11367    0.02038   5.578 2.44e-08
relevel(ClmDurBucket.f, "Mos1-3")Mo4-12 -1.03715   0.03157 -32.851  < 2e-16
relevel(ClmDurBucket.f, "Mos1-3")Yrs2-6 -1.40913   0.03048 -46.226  < 2e-16
relevel(ClmDurBucket.f, "Mos1-3")Yrs7+  -1.44508   0.05953 -24.274  < 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: 59074  on 69418  degrees of freedom
Residual deviance: 56591  on 69412  degrees of freedom
AIC: 80812

Number of Fisher Scoring iterations: 6

>
```

Poisson GLM breakout exercise

- Include some new variables in the GLM from the list below, in addition to Gender, ClaimType, and ClmDurBucket:
  - IncurredAgeBucket.f
  - Region.f
  - Diagnosis_Category.f
  - TQ_Status.f
  - Infl_Rider_Bucket.f
  - Max_Ben_Bucket.f

- Spend 15 - 20 minutes determining which variables you might want to incorporate

Poisson GLM variable selection

- Age is important and significant.
- Region appears not to be as helpful.
- The diagnosis categories could probably use some grouping.
- Tax qualification appears not to add too much value.
- The inflation options appear not to be too different from one another.
- Benefit period is important and could use some grouping.

[Run R 0240: 2.4.14]

```
Call:
glm(formula = SumTerminations ~ relevel(Gender.f, "Female") +
    relevel(ClaimType.f, "HHC") + relevel(IncurredAgeBucket.f,
    "80 to 84") + relevel(ClmDurBucket.f, "Mos1-3") + relevel(Region.f,
    "Unknown") + relevel(Diagnosis_Category.f, "Unknown") + relevel(TQ_Status.f,
    "U") + relevel(Infl_Rider_Bucket.f, "Unknown") + relevel(Max_Ben_Bucket.f,
    "Unknown"), family = poisson, data = ctr_data_ps, offset = LnExposure)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3828  -0.6052  -0.4258  -0.1796   4.0314

Coefficients:
                                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                       -3.0637613  0.5048068  -6.069 1.29e-09
relevel(Gender.f, "Female")Male                                    0.3168230  0.0187292  16.916  < 2e-16
relevel(ClaimType.f, "HHC")ALF                                    -0.2261186  0.0259880  -8.701  < 2e-16
relevel(ClaimType.f, "HHC")NH                                      0.2058818  0.0211961   9.713  < 2e-16
relevel(IncurredAgeBucket.f, "80 to 84")60 to 64                   0.3104159  0.0497131   6.244 4.26e-10
relevel(IncurredAgeBucket.f, "80 to 84")65 to 69                   0.1359209  0.0385241   3.528 0.000418
relevel(IncurredAgeBucket.f, "80 to 84")70 to 74                   0.0840069  0.0301986   2.782 0.005406
relevel(IncurredAgeBucket.f, "80 to 84")75 to 79                   0.0431267  0.0258587   1.668 0.095360
relevel(IncurredAgeBucket.f, "80 to 84")85 to 89                   0.0416842  0.0286104   1.457 0.145128
relevel(IncurredAgeBucket.f, "80 to 84")90+                        0.0522087  0.0447138   1.168 0.242960
relevel(IncurredAgeBucket.f, "80 to 84")LT 60                      0.3130447  0.0546117   5.732 9.91e-09
relevel(ClmDurBucket.f, "Mos1-3")Mo4-12                           -0.8780505  0.0319685 -27.466  < 2e-16
relevel(ClmDurBucket.f, "Mos1-3")Yrs2-6                           -1.1326129  0.0313270 -36.154  < 2e-16
relevel(ClmDurBucket.f, "Mos1-3")Yrs7+                            -1.1449662  0.0606209 -18.887  < 2e-16
relevel(Region.f, "Unknown")Mid-West                               0.4501381  0.0636294   7.074 1.50e-12
relevel(Region.f, "Unknown")Northeast                              0.4850839  0.0650068   7.462 8.52e-14
relevel(Region.f, "Unknown")South                                  0.5135551  0.0623401   8.238  < 2e-16
relevel(Region.f, "Unknown")West                                   0.4365562  0.0637572   6.847 7.53e-12
relevel(Diagnosis_Category.f, "Unknown")Alzheimer's               -0.4626133  0.0549960  -8.412  < 2e-16
relevel(Diagnosis_Category.f, "Unknown")Arthritis                 -0.3469784  0.0623161  -5.568 2.58e-08
relevel(Diagnosis_Category.f, "Unknown")Cancer                     1.0092363  0.0572849  17.618  < 2e-16
relevel(Diagnosis_Category.f, "Unknown")Circulatory               -0.0382711  0.0620208  -0.617 0.537189
relevel(Diagnosis_Category.f, "Unknown")Congenital                 0.5163128  0.2107075   2.450 0.014271
relevel(Diagnosis_Category.f, "Unknown")Diabetes                  -0.3727338  0.1043621  -3.572 0.000355
relevel(Diagnosis_Category.f, "Unknown")Digestive System           0.1708742  0.0989065   1.728 0.084054
relevel(Diagnosis_Category.f, "Unknown")Endocrine/Immunity System -0.1164194  0.1361622  -0.855 0.392548
relevel(Diagnosis_Category.f, "Unknown")Genitourinary System       0.0009937  0.1030113   0.010 0.992303
relevel(Diagnosis_Category.f, "Unknown")Hypertension              -0.4392938  0.1362949  -3.223 0.001268
relevel(Diagnosis_Category.f, "Unknown")Ill-Defined And Miscellaneous Conditions -0.3472433  0.0718307  -4.834 1.34e-06
relevel(Diagnosis_Category.f, "Unknown")Injury                     0.0532918  0.0608328   0.876 0.381010
relevel(Diagnosis_Category.f, "Unknown")Mental                    -0.6377687  0.1445785  -4.411 1.03e-05
relevel(Diagnosis_Category.f, "Unknown")Nervous System And Sense Organs -0.4544683  0.0626192  -7.258 3.94e-13
relevel(Diagnosis_Category.f, "Unknown")Pregnancy Disorders        0.1637342  0.3811007   0.430 0.667461
relevel(Diagnosis_Category.f, "Unknown")Respiratory                0.1404179  0.0693015   2.026 0.042746
relevel(Diagnosis_Category.f, "Unknown")Skin And Subcutaneous Tissue -0.2243730  0.1703392  -1.317 0.187767
relevel(Diagnosis_Category.f, "Unknown")Stroke                    -0.4473993  0.0599936  -7.457 8.82e-14
relevel(TQ_Status.f, "U")N                                        -0.3063513  0.0498924  -6.140 8.24e-10
relevel(TQ_Status.f, "U")Q                                        -0.3504302  0.0239963 -14.604  < 2e-16
relevel(Infl_Rider_Bucket.f, "Unknown")GPO                         0.4926762  0.5004115   0.985 0.324849
relevel(Infl_Rider_Bucket.f, "Unknown")Inflation                   0.4844602  0.5003174   0.968 0.332892
relevel(Infl_Rider_Bucket.f, "Unknown")None                        0.5038136  0.5004384   1.007 0.314058
relevel(Max_Ben_Bucket.f, "Unknown")1 to 2                        -0.1304479  0.0552113  -2.363 0.018142
relevel(Max_Ben_Bucket.f, "Unknown")3 to 4                        -0.2401490  0.0525437  -4.570 4.87e-06
relevel(Max_Ben_Bucket.f, "Unknown")5 +                           -0.3205315  0.0579384  -5.532 3.16e-08
relevel(Max_Ben_Bucket.f, "Unknown")LT 1                           0.1875213  0.0998139   1.879 0.060284
relevel(Max_Ben_Bucket.f, "Unknown")Unlimited                     -0.4693533  0.0535686  -8.762  < 2e-16
```

Poisson GLM variable selection, continued

[Run R 0240: 2.4.15]

- A Type III analysis can also be helpful in understanding if an entire driver variable is important and significant.

- The table shows the impact to Deviance and AIC for removing a driver variable.

```
> drop1(results_ps,~.,test="Chisq")
Single term deletions

Model:
SumTerminations ~ relevel(Gender.f, "Female") + relevel(ClaimType.f,
    "HHC") + relevel(IncurredAgeBucket.f, "80 to 84") + relevel(ClmDurBucket.f,
    "Mos1-3") + relevel(Region.f, "Unknown") + relevel(Diagnosis_Category.f,
    "Unknown") + relevel(TQ_Status.f, "U") + relevel(Infl_Rider_Bucket.f,
    "Unknown") + relevel(Max_Ben_Bucket.f, "Unknown")
                                         Df Deviance   AIC     LRT  Pr(>Chi)
<none>                                        53708 78008
relevel(Gender.f, "Female")               1    53987 78285  279.11 < 2.2e-16 ***
relevel(ClaimType.f, "HHC")               2    53976 78272  268.05 < 2.2e-16 ***
relevel(IncurredAgeBucket.f, "80 to 84")  7    53770 78056   61.79 6.627e-11 ***
relevel(ClmDurBucket.f, "Mos1-3")         3    54790 79084 1081.68 < 2.2e-16 ***
relevel(Region.f, "Unknown")              4    53775 78067   67.01 9.696e-14 ***
relevel(Diagnosis_Category.f, "Unknown") 18    56001 80264 2292.62 < 2.2e-16 ***
relevel(TQ_Status.f, "U")                 2    53918 78213  209.51 < 2.2e-16 ***
relevel(Infl_Rider_Bucket.f, "Unknown")   3    53710 78004    1.83     0.608
relevel(Max_Ben_Bucket.f, "Unknown")      5    53916 78206  207.89 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

- All driver variables are statistically significant, except for the inflation rider variable.

Poisson GLM variable selection, continued

- Variable selection decisions are potentially subjective ones. For this analysis, the following decisions are made:
  - The inflation rider driver does not contribute much explanatory power, and can therefore be eliminated.
  - Diagnosis and benefit period should probably be condensed to a more manageable set of values, and should be kept in the model.
  - The tax qualified status and region drivers are eliminated because the coefficients are so similar.

Poisson GLM smoothing

- The calendar year variable's pattern is somewhat erratic, so it might make sense to smooth it out.
- A logarithmic variable is used in its place.
- This also gives the model more parsimony (lower AIC), and the potential to extrapolate by calendar year.

```
> results_ps <- glm(SumTerminations ~
+                      relevel(Gender.f, "Female")
+                    + relevel(ClaimType.f,"HHC")
+                    + relevel(Diagnosis2.f,"Non-Mental")
+                    + relevel(BP2.f,"Non-Life")
+                    + relevel(IncurredAgeBucket.f,"80 to 84")
+                    + relevel(ClmDurBucket.f,"Mos1-3")
+                    + LnIncYr,
+                    data = ctr_data_ps,
+                    family = poisson,
+                    offset = LnExposure
+                    )
> summary(results_ps)

Call:
glm(formula = SumTerminations ~ relevel(Gender.f, "Female") +
    relevel(ClaimType.f, "HHC") + relevel(Diagnosis2.f, "Non-Mental") +
    relevel(BP2.f, "Non-Life") + relevel(IncurredAgeBucket.f,
    "80 to 84") + relevel(ClmDurBucket.f, "Mos1-3") + LnIncYr,
    family = poisson, data = ctr_data_ps, offset = LnExposure)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5788  -0.9097  -0.4351   0.4983   3.7309

Coefficients:
                                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                                    -2.25398    0.11854 -19.015  < 2e-16
relevel(Gender.f, "Female")Male                 0.28387    0.01850  15.345  < 2e-16
relevel(ClaimType.f, "HHC")ALF                 -0.26751    0.02429 -11.013  < 2e-16
relevel(ClaimType.f, "HHC")NH                   0.14036    0.02064   6.799 1.05e-11
relevel(Diagnosis2.f, "Non-Mental")Mental      -0.39533    0.02117 -18.674  < 2e-16
relevel(BP2.f, "Non-Life")Lifetime             -0.20134    0.02127  -9.467  < 2e-16
relevel(IncurredAgeBucket.f, "80 to 84")60 to 64  0.34336    0.04849   7.081 1.43e-12
relevel(IncurredAgeBucket.f, "80 to 84")65 to 69  0.17628    0.03752   4.698 2.63e-06
relevel(IncurredAgeBucket.f, "80 to 84")70 to 74  0.09227    0.02966   3.111  0.00186
relevel(IncurredAgeBucket.f, "80 to 84")75 to 79  0.03985    0.02572   1.550  0.12124
relevel(IncurredAgeBucket.f, "80 to 84")85 to 89  0.06284    0.02847   2.207  0.02732
relevel(IncurredAgeBucket.f, "80 to 84")90+       0.07111    0.04426   1.607  0.10816
relevel(IncurredAgeBucket.f, "80 to 84")LT 60     0.38168    0.05366   7.113 1.14e-12
relevel(ClmDurBucket.f, "Mos1-3")Mo4-12        -1.00810    0.03163 -31.872  < 2e-16
relevel(ClmDurBucket.f, "Mos1-3")Yrs2-6        -1.38078    0.03096 -44.597  < 2e-16
relevel(ClmDurBucket.f, "Mos1-3")Yrs7+         -1.46500    0.06334 -23.130  < 2e-16
LnIncYr                                        -0.07028    0.04154  -1.692  0.09067 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10567.7  on 6221  degrees of freedom
Residual deviance:  7525.7  on 6205  degrees of freedom
AIC: 16995

Number of Fisher Scoring iterations: 6

>
```

[Run R 0240: 2.4.16 – 2.4.20]

Poisson GLM cross-terms

- The introduction of cross-terms can also help improve the model fit.
- Cross-terms are introduced between:
  - Calendar year / Age
  - Calendar year / Duration
- The model fit is improved, and the extra parameter “expense” appears to be justified (lower AIC).

Console ~/R Projects/Examples/Output/

```
+                  + relevel(Diagnosis2.f,"Non-Mental")
+                  + relevel(BP2.f,"Non-Life")
+                  + LnIncYr * relevel(IncurredAgeBucket.f,"80 to 84")
+                  + LnIncYr * relevel(ClmDurBucket.f,"Mos1-3"),
+                  data = ctr_data_ps,
+                  family = poisson,
+                  offset = LnExposure
+                  )
> summary(results_ps)

Call:
glm(formula = SumTerminations ~ relevel(Gender.f, "Female") + 
    relevel(ClaimType.f, "HHC") + relevel(Diagnosis2.f, "Non-Mental") + 
    relevel(BP2.f, "Non-Life") + LnIncYr * relevel(IncurredAgeBucket.f, 
    "80 to 84") + LnIncYr * relevel(ClmDurBucket.f, "Mos1-3"), 
    family = poisson, data = ctr_data_ps, offset = LnExposure)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.4506  -0.9082  -0.4417   0.5102   3.7714  

Coefficients:
                                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                                           -0.07489    0.41238  -0.182 0.855893
relevel(Gender.f, "Female")Male                        0.28770    0.01851  15.543  < 2e-16
relevel(ClaimType.f, "HHC")ALF                        -0.26359    0.02432 -10.840  < 2e-16
relevel(ClaimType.f, "HHC")NH                          0.14420    0.02067   6.978 3.00e-12
relevel(Diagnosis2.f, "Non-Mental")Mental             -0.39268    0.02118 -18.542  < 2e-16
relevel(BP2.f, "Non-Life")Lifetime                    -0.20009    0.02127  -9.408  < 2e-16
LnIncYr                                               -0.88478    0.15352  -5.763 8.25e-09
relevel(IncurredAgeBucket.f, "80 to 84")60 to 64      -2.25650    0.54132  -4.169 3.07e-05
relevel(IncurredAgeBucket.f, "80 to 84")65 to 69      -2.52326    0.40877  -6.173 6.71e-10
relevel(IncurredAgeBucket.f, "80 to 84")70 to 74      -1.11712    0.31504  -3.546 0.000391
relevel(IncurredAgeBucket.f, "80 to 84")75 to 79      -0.21513    0.28366  -0.758 0.448203
relevel(IncurredAgeBucket.f, "80 to 84")85 to 89       0.43962    0.34214   1.285 0.198820
relevel(IncurredAgeBucket.f, "80 to 84")90+            0.73858    0.61341   1.204 0.228571
relevel(IncurredAgeBucket.f, "80 to 84")LT 60         -3.06430    0.65245  -4.697 2.65e-06
relevel(ClmDurBucket.f, "Mos1-3")Mo4-12               -2.70735    0.42624  -6.352 2.13e-10
relevel(ClmDurBucket.f, "Mos1-3")Yrs2-6               -3.07292    0.39676  -7.745 9.55e-15
relevel(ClmDurBucket.f, "Mos1-3")Yrs7+                -2.84736    0.53904  -5.282 1.28e-07
LnIncYr:relevel(IncurredAgeBucket.f, "80 to 84")60 to 64  0.98602    0.20343   4.847 1.25e-06
LnIncYr:relevel(IncurredAgeBucket.f, "80 to 84")65 to 69  1.03356    0.15498   6.669 2.58e-11
LnIncYr:relevel(IncurredAgeBucket.f, "80 to 84")70 to 74  0.46082    0.11995   3.842 0.000122
LnIncYr:relevel(IncurredAgeBucket.f, "80 to 84")75 to 79  0.09385    0.10772   0.871 0.383618
LnIncYr:relevel(IncurredAgeBucket.f, "80 to 84")85 to 89 -0.13673    0.12808  -1.068 0.285716
LnIncYr:relevel(IncurredAgeBucket.f, "80 to 84")90+      -0.23860    0.22663  -1.053 0.292421
LnIncYr:relevel(IncurredAgeBucket.f, "80 to 84")LT 60     1.30804    0.24528   5.333 9.67e-08
LnIncYr:relevel(ClmDurBucket.f, "Mos1-3")Mo4-12           0.62960    0.15781   3.990 6.62e-05
LnIncYr:relevel(ClmDurBucket.f, "Mos1-3")Yrs2-6           0.62937    0.14800   4.252 2.12e-05
LnIncYr:relevel(ClmDurBucket.f, "Mos1-3")Yrs7+            0.53373    0.22209   2.403 0.016251
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10567.7  on 6221  degrees of freedom
Residual deviance:  7394.7  on 6195  degrees of freedom
AIC: 16884

Number of Fisher Scoring iterations: 6

> |
```

[Run R 0240: 2.4.21]

Poisson GLM variation

- A Type I analysis can be run, which describes the variation explained by each driver variable in a sequential order. A different ordering might give you a different result.
- A Type III analysis can also be run, which describes reduction in the variation explained by removing one driver variable at a time. This is order-independent.

[Run R 0240: 2.4.22 – 2.4.25]

Poisson GLM over-fitting

- If we use too many parameters in the GLM, we will over-fit the model.
- The data will be fit “too well” by the model, and will probably not test well in a separate random sample.

[Run R 0240: 2.4.26 – 2.4.27]

Poisson GLM offset model

- We can replace the exposures (t) in the GLM with an expected basis, like an industry table (or benchmark).
- The GLM coefficients are now in reference to the benchmark.
- Courtesy of Milliman:

Standard model: $\ln(\mu) = \ln(t) + \beta_0 + \sum \beta_i x_i$

Update model: $\ln(\mu) = \ln(t * benchmark) + \sum \beta_i x_i$

$$\widehat{\mu}_j = t * benchmark * e^{\widehat{\beta_1} x_1} \times \cdots \times e^{\widehat{\beta_k} x_k}$$

[Run R 0250: 2.5.1 – 2.5.15]