Generalized Linear Model (GLM)

Emma Li

March 27, 2017

Compare Linear Regression & GLM

Linear Regression: E(Y) = μ = X * β	GLM: g(E(Y)) = g(μ) = X * β
* Linear relationship between X and E(Y)	* Linear relationship between X and g(E(Y))
* Multivariate normality	
* No or little multicollinearity	* No or little multicollinearity
* No auto-correlation	* No auto-correlation
* Error terms have similar variances	* Error terms have similar variances

Review GLM Examples

- Distributions in Exponential Family: Normal, Bernoulli, Binomial, Poisson, Negative Binomial, Gamma, Tweedie, Exponential, etc.

1. Y is count (e.g. claim count): Poisson distribution
2. Y is binary (e.g., loss or no loss): Bernoulli distribution

- Link Functions

1. Poisson distribution: log function = ln(λ) = X*β
2. Bernoulli distribution: logit function = ln(p/(1-p)) = X*β

R Packages

1. stats: glm() is used to fit generalized linear models
2. insuranceData: 'A Collection of Insurance Datasets Useful in Risk Classification in Non-life Insurance'

- Inputs: glm(formula, family, data, . . .)
- Outputs: coefficients, p values, residuals, fitted values, summary, . . .

Case Study

One dataset in package 'insuranceData' is called 'dataCar' and it is based on one-year vehicle insurance policies taken out in 2004 or 2005.

Variables	Descriptions
numclaims	number of claim
claimcst0	claim amount
veh_body	vehicle type
veh_age	vehicle age

Variables	Descriptions
gender	driver gender
area	location
agecat	driver age

Summary and Graphs

```r
library("insuranceData")
data(dataCar)
```

```r
dim(dataCar)
```

```
## [1] 67856    11
```

```r
colnames(dataCar)
```

```
##  [1] "veh_value" "exposure"  "clm"       "numclaims" "claimcst0"
##  [6] "veh_body"  "veh_age"   "gender"    "area"      "agecat"
## [11] "X_OBSTAT_"
```

Summary and Graphs

```r
head(dataCar)
```

```
##   veh_value  exposure clm numclaims claimcst0 veh_body veh_age gender area
## 1      1.06 0.3039014   0         0         0    HBACK       3      F    C
## 2      1.03 0.6488706   0         0         0    HBACK       2      F    A
## 3      3.26 0.5694730   0         0         0      UTE       2      F    E
## 4      4.14 0.3175907   0         0         0    STNWG       2      F    D
## 5      0.72 0.6488706   0         0         0    HBACK       4      F    C
## 6      2.01 0.8542094   0         0         0    HDTOP       3      M    C
##   agecat             X_OBSTAT_
## 1      2 01101    0    0    0
## 2      4 01101    0    0    0
## 3      2 01101    0    0    0
## 4      2 01101    0    0    0
## 5      2 01101    0    0    0
## 6      4 01101    0    0    0
```

Summary and Graphs

```r
str(dataCar)
```

```
## 'data.frame':    67856 obs. of  11 variables:
##  $ veh_value: num  1.06 1.03 3.26 4.14 0.72 2.01 1.6 1.47 0.52 0.38 ...
##  $ exposure : num  0.304 0.649 0.569 0.318 0.649 ...
##  $ clm      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ numclaims: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ claimcst0: num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ veh_body : Factor w/ 13 levels "BUS","CONVT",..: 4 4 13 11 4 5 8 4 4 4 ...
##  $ veh_age  : int  3 2 2 2 4 3 3 2 4 4 ...
##  $ gender   : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 2 1 1 ...
##  $ area     : Factor w/ 6 levels "A","B","C","D",..: 3 1 5 4 3 3 1 2 1 2 ...
##  $ agecat   : int  2 4 2 2 2 4 4 6 3 4 ...
##  $ X_OBSTAT_: Factor w/ 1 level "01101    0    0    0": 1 1 1 1 1 1 1 1 1 1 ...
```

Summary and Graphs

```r
summary(dataCar)
```

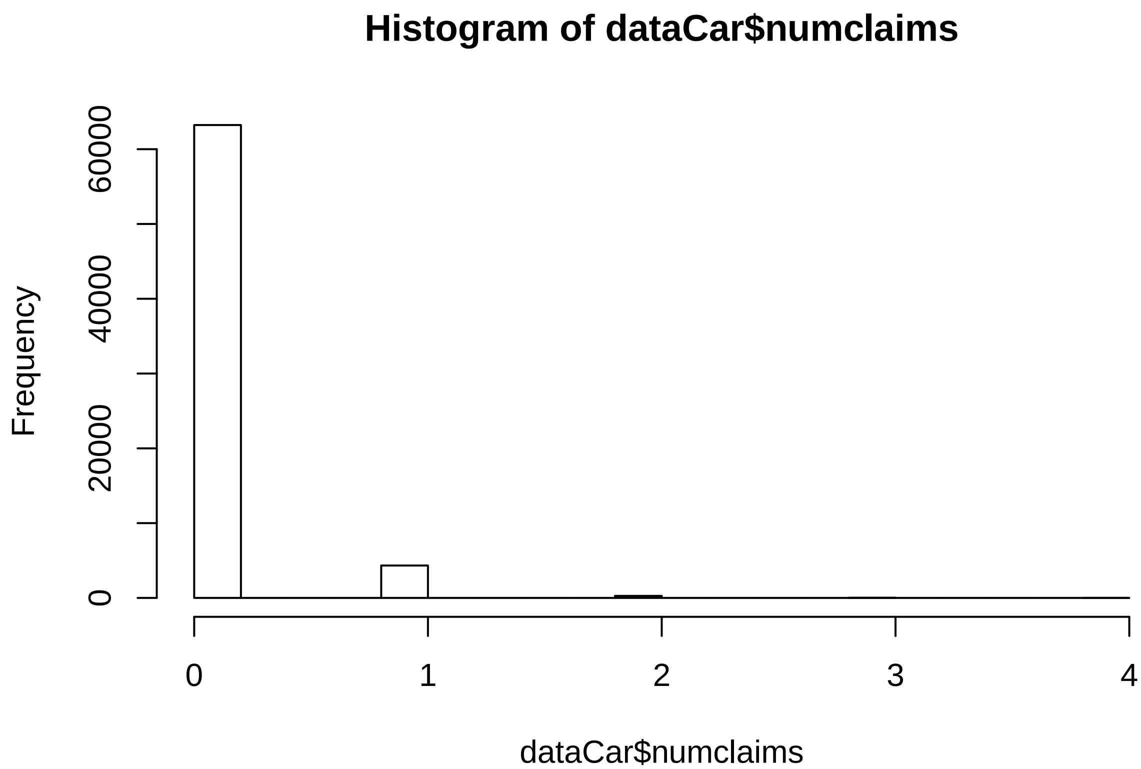
```
##    veh_value         exposure             clm            numclaims
##  Min.   : 0.000   Min.   :0.002738   Min.   :0.00000   Min.   :0.00000
##  1st Qu.: 1.010   1st Qu.:0.219028   1st Qu.:0.00000   1st Qu.:0.00000
##  Median : 1.500   Median :0.446270   Median :0.00000   Median :0.00000
##  Mean   : 1.777   Mean   :0.468651   Mean   :0.06814   Mean   :0.07276
##  3rd Qu.: 2.150   3rd Qu.:0.709103   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :34.560   Max.   :0.999316   Max.   :1.00000   Max.   :4.00000
##
##    claimcst0         veh_body        veh_age      gender    area
##  Min.   :    0.0   SEDAN  :22233   Min.   :1.000   F:38603   A:16312
##  1st Qu.:    0.0   HBACK  :18915   1st Qu.:2.000   M:29253   B:13341
##  Median :    0.0   STNWG  :16261   Median :3.000             C:20540
##  Mean   :  137.3   UTE    : 4586   Mean   :2.674             D: 8173
##  3rd Qu.:    0.0   TRUCK  : 1750   3rd Qu.:4.000             E: 5912
##  Max.   :55922.1   HDTOP  : 1579   Max.   :4.000             F: 3578
##                    (Other): 2532
##      agecat                     X_OBSTAT_
##  Min.   :1.000   01101    0    0    0:67856
##  1st Qu.:2.000
##  Median :3.000
##  Mean   :3.485
##  3rd Qu.:5.000
##  Max.   :6.000
##
```

Summary and Graphs

```r
table(dataCar$numclaims)
```

```
##
##     0     1     2     3     4
## 63232  4333   271    18     2
```

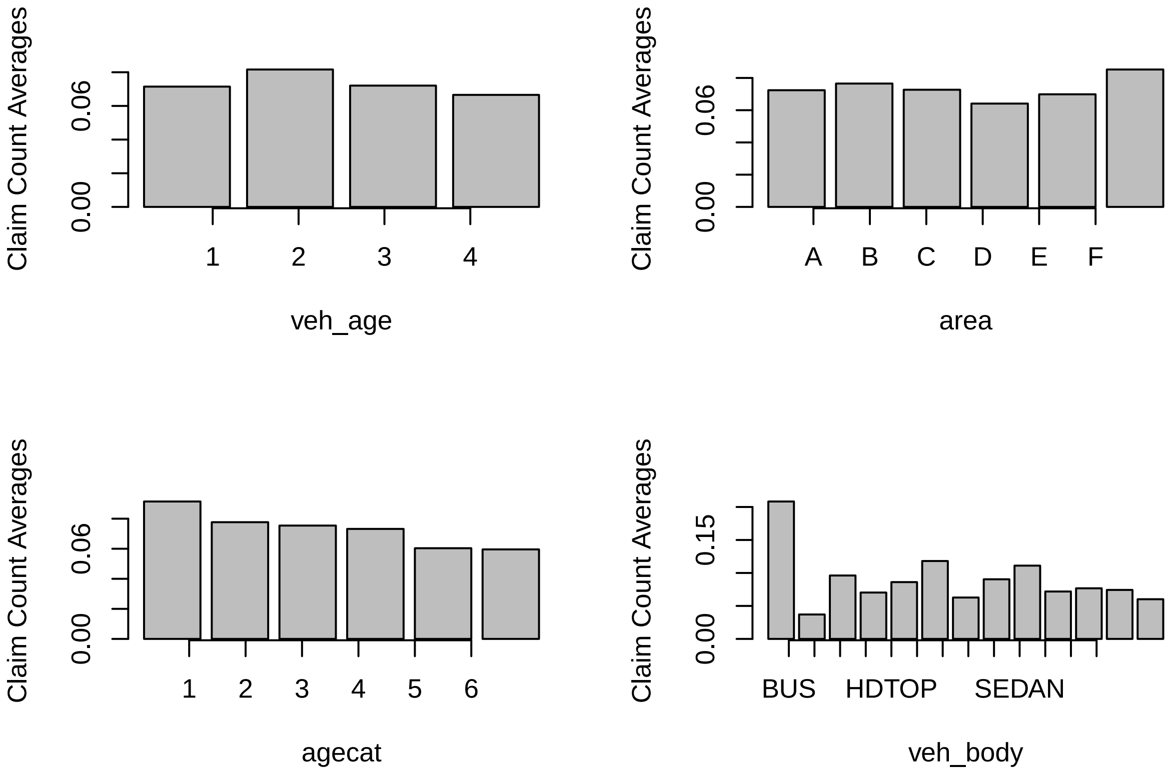
```r
hist(dataCar$numclaims)
```



Basic Data Manipulation

```
avg <- function(x) {
  dat <- aggregate(dataCar$numclaims, by = list(dataCar[, x]), FUN = mean)
  barplot(dat$x, xlab = x, ylab = "Claim Count Averages")
  axis(side=1, at=1:nrow(dat), labels=dat$Group.1)
}

par(mfrow=c(2,2))
  avg(x = "veh_age")
  avg(x = "area")
  avg(x = "agecat")
  avg(x = "veh_body")
```



Basic Data Manipulation

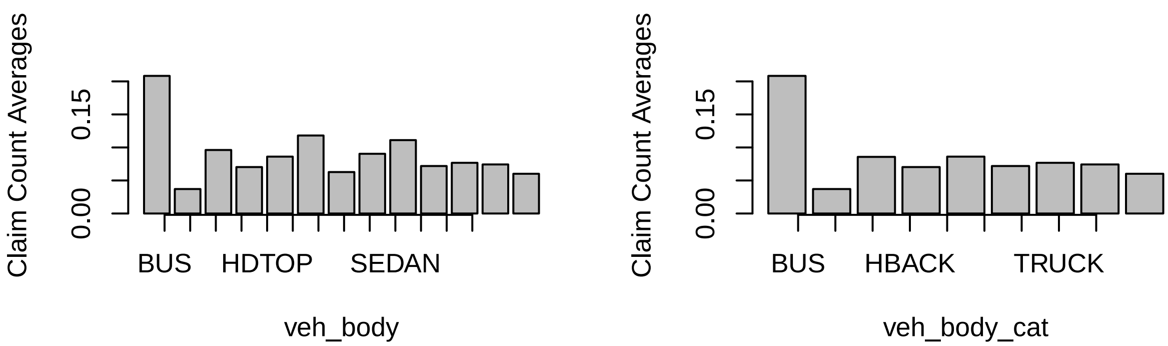
```
summaries<-function(x) {
  means<-aggregate(dataCar$numclaims, by = list(dataCar[, x]), FUN = mean)
  lengths<-aggregate(dataCar$numclaims, by = list(dataCar[, x]), FUN = length)
  means_lengths<-merge(means,lengths,by="Group.1"  )
  colnames(means_lengths)<-c(x,'numclaims average','count')
  means_lengths[order(means_lengths[,"numclaims average"]),]
}
summaries("veh_body")
```

```
##    veh_body numclaims average count
## 2     CONVT        0.03703704    81
## 13      UTE        0.06018317  4586
## 7     MIBUS        0.06276151   717
## 4     HBACK        0.07031457 18915
## 10    SEDAN        0.07187514 22233
## 12    TRUCK        0.07428571  1750
## 11    STNWG        0.07674805 16261
## 5     HDTOP        0.08613046  1579
## 8     PANVN        0.09042553   752
## 3     COUPE        0.09615385   780
## 9     RDSTR        0.11111111    27
## 6     MCARA        0.11811024   127
## 1       BUS        0.20833333    48
```

```r
library("plyr")
dataCar$veh_body_cat<-mapvalues(dataCar$veh_body
                        , from = c("COUPE", "MCARA", "MIBUS", "PANVN", "RDSTR")
                        , to = c("Others","Others","Others","Others","Others"))
summaries("veh_body_cat")
```

```
##   veh_body_cat numclaims average count
## 2        CONVT        0.03703704    81
## 9          UTE        0.06018317  4586
## 3        HBACK        0.07031457 18915
## 6        SEDAN        0.07187514 22233
## 8        TRUCK        0.07428571  1750
## 7        STNWG        0.07674805 16261
## 5       Others        0.08572618  2403
## 4        HDTOP        0.08613046  1579
## 1          BUS        0.20833333    48
```

```r
par(mfrow=c(2,2))
  avg(x = "veh_body")
  avg(x = "veh_body_cat")
```



Basic Data Manipulation

```r
summaries("veh_age")
```

```
##   veh_age numclaims average count
## 4       4        0.06655056 18948
## 1       1        0.07146936 12257
## 3       3        0.07206938 20064
## 2       2        0.08163019 16587
```

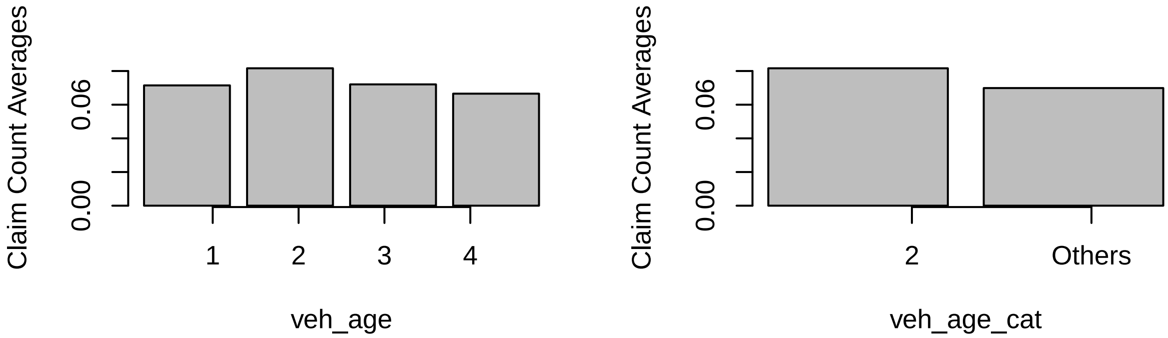
```r
dataCar$veh_age_cat<-factor(ifelse(dataCar$veh_age =='2','2','Others'))
aggregate(dataCar$numclaims, by = list(dataCar[, "veh_age_cat"]), FUN = mean)
```

```
##   Group.1          x
## 1       2 0.08163019
## 2  Others 0.06988629
```

```r
summaries("veh_age_cat")
```

```
##   veh_age_cat numclaims average count
## 2      Others        0.06988629 51269
## 1           2        0.08163019 16587
```

```r
par(mfrow=c(2,2))
  avg(x = "veh_age")
  avg(x = "veh_age_cat")
```



Basic Data Manipulation

```r
summaries("area")
```

```
##   area numclaims average count
## 4    D        0.06411354  8173
## 5    E        0.06985792  5912
## 1    A        0.07240069 16312
## 3    C        0.07268744 20540
## 2    B        0.07653099 13341
## 6    F        0.08524315  3578
```

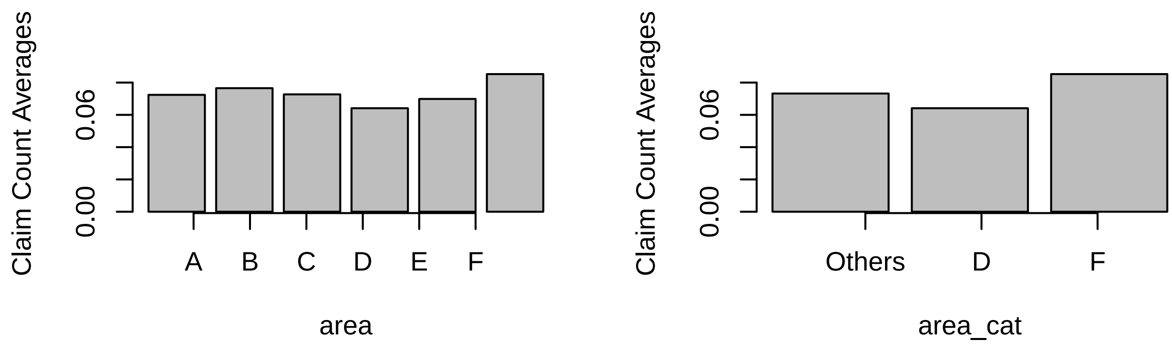
```r
dataCar$area_cat<-mapvalues(dataCar$area
                            , from = c("A","B","C","E")
                            , to = c("Others","Others","Others","Others"))
aggregate(dataCar$numclaims, by = list(dataCar[, "area_cat"]), FUN = mean)
```

```
##   Group.1          x
## 1  Others 0.07321986
## 2       D 0.06411354
## 3       F 0.08524315
```

```r
summaries("area_cat")
```

```
##   area_cat numclaims average count
## 1        D        0.06411354  8173
## 3   Others        0.07321986 56105
## 2        F        0.08524315  3578
```

```r
par(mfrow=c(2,2))
  avg(x = "area")
  avg(x = "area_cat")
```



Basic Data Manipulation

```r
summaries("agecat")
```

```
##   agecat numclaims average  count
## 6      6         0.05956927  6547
## 5      5         0.06035768 10736
## 4      4         0.07319785 16189
## 3      3         0.07541067 15767
## 2      2         0.07766990 12875
## 1      1         0.09143156  5742
```

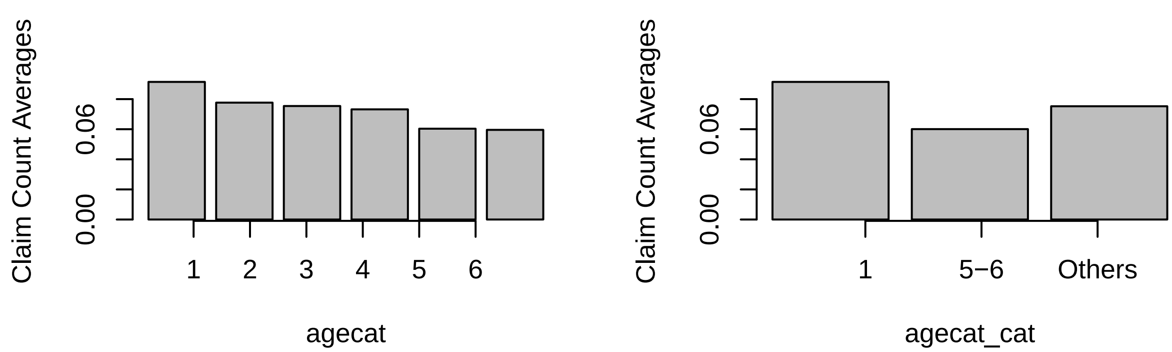
```r
dataCar$agecat_cat<-factor(ifelse(dataCar$agecat =='1','1'
                          ,ifelse(dataCar$agecat %in% c('5','6'),'5-6','Others')))
aggregate(dataCar$numclaims, by = list(dataCar[, "agecat_cat"]), FUN = mean)
```

```
##   Group.1          x
## 1       1 0.09143156
## 2     5-6 0.06005902
## 3  Others 0.07526042
```

```r
summaries("agecat_cat")
```

```
##   agecat_cat numclaims average  count
## 2        5-6         0.06005902 17283
## 3     Others         0.07526042 44831
## 1          1         0.09143156  5742
```

```r
par(mfrow=c(2,2))
  avg(x = "agecat")
  avg(x = "agecat_cat")
```



1. Poisson Regression

```
formulas<-"numclaims ~ veh_body_cat"
poisson_reg1 <- glm(formulas, data =dataCar, family=poisson)
summary(poisson_reg1)
```

```
##
## Call:
## glm(formula = formulas, family = poisson, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6455  -0.3918  -0.3791  -0.3750   4.8766
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.5686     0.3162  -4.960 7.03e-07 ***
## veh_body_catCONVT    -1.7272     0.6583  -2.624 0.008695 **
## veh_body_catOthers   -0.8880     0.3238  -2.742 0.006102 **
## veh_body_catHBACK    -1.0862     0.3174  -3.422 0.000622 ***
## veh_body_catHDTOP    -0.8833     0.3276  -2.696 0.007022 **
## veh_body_catSEDAN    -1.0642     0.3172  -3.355 0.000794 ***
## veh_body_catSTNWG    -0.9986     0.3175  -3.145 0.001659 **
## veh_body_catTRUCK    -1.0312     0.3282  -3.142 0.001676 **
## veh_body_catUTE      -1.2417     0.3219  -3.857 0.000115 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26768  on 67855  degrees of freedom
## Residual deviance: 26734  on 67847  degrees of freedom
## AIC: 36186
##
## Number of Fisher Scoring iterations: 6
```

```
dataCar <- within(dataCar, veh_body_cat <- relevel(veh_body_cat, ref = 'Others'))
poisson_reg1 <- glm(formulas, data =dataCar, family=poisson)
summary(poisson_reg1)
```

```
##
## Call:
## glm(formula = formulas, family = poisson, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6455  -0.3918  -0.3791  -0.3750   4.8766
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.456597   0.069673 -35.259  < 2e-16 ***
## veh_body_catBUS    0.887981   0.323812   2.742 0.006102 **
## veh_body_catCONVT -0.839240   0.581539  -1.443 0.148982
## veh_body_catHBACK -0.198179   0.074875  -2.647 0.008126 **
```

```
## veh_body_catHDTOP  0.004705   0.110487   0.043 0.966033
## veh_body_catSEDAN -0.176228   0.074028  -2.381 0.017287 *
## veh_body_catSTNWG -0.110630   0.075204  -1.471 0.141273
## veh_body_catTRUCK -0.143240   0.112012  -1.279 0.200972
## veh_body_catUTE   -0.353766   0.092074  -3.842 0.000122 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26768  on 67855  degrees of freedom
## Residual deviance: 26734  on 67847  degrees of freedom
## AIC: 36186
##
## Number of Fisher Scoring iterations: 6
```

1. Poisson Regression

```r
formulas<-"numclaims ~ veh_age_cat"
dataCar <- within(dataCar, veh_age_cat <- relevel(veh_age_cat, ref = 'Others'))
poisson_reg2 <- glm(formulas, data =dataCar, family=poisson)
summary(poisson_reg2)
```

```
##
## Call:
## glm(formula = formulas, family = poisson, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4041  -0.3739  -0.3739  -0.3739   4.9515
##
## Coefficients:
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  -2.66089    0.01671 -159.276  < 2e-16 ***
## veh_age_cat2  0.15533    0.03190    4.869 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26768  on 67855  degrees of freedom
## Residual deviance: 26745  on 67854  degrees of freedom
## AIC: 36184
##
## Number of Fisher Scoring iterations: 6
```

1. Poisson Regression

```r
formulas<-"numclaims ~ area_cat"
dataCar <- within(dataCar, area_cat <- relevel(area_cat, ref = 'Others'))
```

```r
poisson_reg3 <- glm(formulas, data =dataCar, family=poisson)
summary(poisson_reg3)
```

```
##
## Call:
## glm(formula = formulas, family = poisson, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4129  -0.3827  -0.3827  -0.3827   4.9144
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -2.61429    0.01560 -167.559   <2e-16 ***
## area_catD   -0.13281    0.04639   -2.863   0.0042 **
## area_catF    0.15204    0.05935    2.562   0.0104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26768  on 67855  degrees of freedom
## Residual deviance: 26752  on 67853  degrees of freedom
## AIC: 36193
##
## Number of Fisher Scoring iterations: 6
```

1. Poisson Regression

```r
formulas<-"numclaims ~ agecat_cat"
dataCar <- within(dataCar, agecat_cat <- relevel(agecat_cat, ref = 'Others'))
poisson_reg4 <- glm(formulas, data =dataCar, family=poisson)
summary(poisson_reg4)
```

```
##
## Call:
## glm(formula = formulas, family = poisson, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4276  -0.3880  -0.3880  -0.3466   5.0705
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)   -2.58680    0.01722 -150.257  < 2e-16 ***
## agecat_cat1    0.19464    0.04692    4.149 3.35e-05 ***
## agecat_cat5-6 -0.22563    0.03549   -6.357 2.06e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

```
##     Null deviance: 26768  on 67855  degrees of freedom
## Residual deviance: 26698  on 67853  degrees of freedom
## AIC: 36139
##
## Number of Fisher Scoring iterations: 6
```

1. Poisson Regression

```r
formulas<-"numclaims ~ veh_body_cat+veh_age_cat+area_cat+agecat_cat"
dataCar <- within(dataCar, agecat_cat <- relevel(agecat_cat, ref = 'Others'))
poisson_reg5 <- glm(formulas, data =dataCar, family=poisson)
summary(poisson_reg5)
```

```
##
## Call:
## glm(formula = formulas, family = poisson, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7609  -0.3973  -0.3788  -0.3494   5.0553
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.455364   0.070404 -34.875  < 2e-16 ***
## veh_body_catBUS    0.853238   0.324157   2.632  0.00848 **
## veh_body_catCONVT -0.870229   0.581556  -1.496  0.13455
## veh_body_catHBACK -0.214739   0.075125  -2.858  0.00426 **
## veh_body_catHDTOP -0.006637   0.110925  -0.060  0.95229
## veh_body_catSEDAN -0.160530   0.074291  -2.161  0.03071 *
## veh_body_catSTNWG -0.120589   0.075658  -1.594  0.11097
## veh_body_catTRUCK -0.157103   0.112529  -1.396  0.16268
## veh_body_catUTE   -0.370031   0.092471  -4.002 6.29e-05 ***
## veh_age_cat2       0.158271   0.032063   4.936 7.97e-07 ***
## area_catD         -0.121573   0.046717  -2.602  0.00926 **
## area_catF          0.105021   0.061053   1.720  0.08540 .
## agecat_cat1        0.204307   0.047198   4.329 1.50e-05 ***
## agecat_cat5-6     -0.216944   0.035804  -6.059 1.37e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26768  on 67855  degrees of freedom
## Residual deviance: 26627  on 67842  degrees of freedom
## AIC: 36090
##
## Number of Fisher Scoring iterations: 6
```

2. Logistic Regression

```r
dataCar$numclaims_bin <- ifelse(dataCar$numclaims == 0, 0, 1)
table(dataCar$numclaims_bin)
```

```
##
##     0     1
## 63232  4624
```

```r
formulas<-"numclaims_bin ~ veh_body_cat"
dataCar <- within(dataCar, veh_body_cat <- relevel(veh_body_cat, ref = 'Others'))
logistic_reg1 <- glm(formulas, data =dataCar, family=binomial)
summary(logistic_reg1)
```

```
##
## Call:
## glm(formula = formulas, family = binomial, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6444  -0.3870  -0.3719  -0.3707   2.5674
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.46081    0.07578 -32.473  < 2e-16 ***
## veh_body_catBUS     0.99447    0.37748   2.634 0.008427 **
## veh_body_catCONVT -0.79729    0.59296  -1.345 0.178757
## veh_body_catHBACK -0.17570    0.08118  -2.164 0.030441 *
## veh_body_catHDTOP  0.04971    0.11885   0.418 0.675728
## veh_body_catSEDAN -0.18274    0.08043  -2.272 0.023078 *
## veh_body_catSTNWG -0.09353    0.08162  -1.146 0.251835
## veh_body_catTRUCK -0.14803    0.12120  -1.221 0.221933
## veh_body_catUTE   -0.35091    0.09910  -3.541 0.000398 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33767  on 67855  degrees of freedom
## Residual deviance: 33733  on 67847  degrees of freedom
## AIC: 33751
##
## Number of Fisher Scoring iterations: 5
```

2. Logistic Regression

```r
formulas<-"numclaims_bin ~ veh_age_cat"
dataCar <- within(dataCar, veh_age_cat <- relevel(veh_age_cat, ref = 'Others'))
logistic_reg2 <- glm(formulas, data =dataCar, family=binomial)
summary(logistic_reg2)
```

```
##
```

```
## Call:
## glm(formula = formulas, family = binomial, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3973  -0.3685  -0.3685  -0.3685   2.3340
##
## Coefficients:
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  -2.65577    0.01783 -148.917  < 2e-16 ***
## veh_age_cat2  0.15641    0.03432    4.558 5.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33767  on 67855  degrees of freedom
## Residual deviance: 33746  on 67854  degrees of freedom
## AIC: 33750
##
## Number of Fisher Scoring iterations: 5
```

2. Logistic Regression

```r
formulas<-"numclaims_bin ~ area_cat"
dataCar <- within(dataCar, area_cat <- relevel(area_cat, ref = 'Others'))
logistic_reg3 <- glm(formulas, data =dataCar, family=binomial)
summary(logistic_reg3)
```

```
##
## Call:
## glm(formula = formulas, family = binomial, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4037  -0.3770  -0.3770  -0.3770   2.3673
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -2.60862    0.01670 -156.171  < 2e-16 ***
## area_catD   -0.13079    0.04925   -2.656  0.00791 **
## area_catF    0.14234    0.06445    2.209  0.02721 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33767  on 67855  degrees of freedom
## Residual deviance: 33754  on 67853  degrees of freedom
## AIC: 33760
##
## Number of Fisher Scoring iterations: 5
```

2. Logistic Regression

```r
formulas<-"numclaims_bin ~ agecat_cat"
dataCar <- within(dataCar, agecat_cat <- relevel(agecat_cat, ref = 'Others'))
logistic_reg4 <- glm(formulas, data =dataCar, family=binomial)
summary(logistic_reg4)
```

```
##
## Call:
## glm(formula = formulas, family = binomial, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4251  -0.3817  -0.3817  -0.3415   2.3962
##
## Coefficients:
##                Estimate Std. Error  z value Pr(>|z|)
## (Intercept)    -2.58298    0.01848 -139.764  < 2e-16 ***
## agecat_cat1     0.22434    0.05048    4.444 8.83e-06 ***
## agecat_cat5-6 -0.22965    0.03774   -6.085 1.16e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33767  on 67855  degrees of freedom
## Residual deviance: 33698  on 67853  degrees of freedom
## AIC: 33704
##
## Number of Fisher Scoring iterations: 5
```

2. Logistic Regression

```r
formulas<-"numclaims_bin ~ veh_body_cat+veh_age_cat+area_cat+agecat_cat"
dataCar <- within(dataCar, agecat_cat <- relevel(agecat_cat, ref = 'Others'))
logistic_reg5 <- glm(formulas, data =dataCar, family=binomial)
summary(logistic_reg5)
```

```
##
## Call:
## glm(formula = formulas, family = binomial, data = dataCar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7560  -0.3908  -0.3737  -0.3423   2.6610
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.46085    0.07660 -32.125  < 2e-16 ***
## veh_body_catBUS     0.96181    0.37863   2.540 0.011077 *
## veh_body_catCONVT -0.83061    0.59320  -1.400 0.161447
## veh_body_catHBACK -0.19287    0.08150  -2.366 0.017961 *
```

```
## veh_body_catHDTOP  0.04098    0.11940   0.343 0.731415
## veh_body_catSEDAN -0.16582    0.08076  -2.053 0.040043 *
## veh_body_catSTNWG -0.10095    0.08214  -1.229 0.219099
## veh_body_catTRUCK -0.16210    0.12184  -1.330 0.183370
## veh_body_catUTE   -0.36724    0.09958  -3.688 0.000226 ***
## veh_age_cat2       0.15866    0.03452   4.597 4.30e-06 ***
## area_catD         -0.12005    0.04964  -2.418 0.015599 *
## area_catF          0.08939    0.06632   1.348 0.177728
## agecat_cat1        0.23396    0.05082   4.604 4.14e-06 ***
## agecat_cat5-6     -0.21965    0.03809  -5.767 8.05e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 33767  on 67855  degrees of freedom
## Residual deviance: 33635  on 67842  degrees of freedom
## AIC: 33663
##
## Number of Fisher Scoring iterations: 5
```

Model Selection

Poisson Regression

- Nested Models

```
anova(poisson_reg1,poisson_reg5, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: numclaims ~ veh_body_cat
## Model 2: numclaims ~ veh_body_cat + veh_age_cat + area_cat + agecat_cat
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     67847      26734
## 2     67842      26627  5   106.29 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(poisson_reg2,poisson_reg5, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: numclaims ~ veh_age_cat
## Model 2: numclaims ~ veh_body_cat + veh_age_cat + area_cat + agecat_cat
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     67854      26745
## 2     67842      26627 12   117.77 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(poisson_reg3,poisson_reg5, test="Chisq")
```

```
## Analysis of Deviance Table
```

```
##
## Model 1: numclaims ~ area_cat
## Model 2: numclaims ~ veh_body_cat + veh_age_cat + area_cat + agecat_cat
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     67853      26752
## 2     67842      26627 11   124.74 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(poisson_reg4,poisson_reg5, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: numclaims ~ agecat_cat
## Model 2: numclaims ~ veh_body_cat + veh_age_cat + area_cat + agecat_cat
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     67853      26698
## 2     67842      26627 11   70.921 8.162e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Non-Nested Models

```
AICs<-c(poisson_reg1$aic,poisson_reg2$aic,poisson_reg3$aic,poisson_reg4$aic)
AICs
```

```
## [1] 36186.37 36183.86 36192.83 36139.01
```

Select 2 Poisson Regressions

Model Selection

Logistic Regression

- Nested Models

```
anova(logistic_reg1,logistic_reg5, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: numclaims_bin ~ veh_body_cat
## Model 2: numclaims_bin ~ veh_body_cat + veh_age_cat + area_cat + agecat_cat
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     67847      33733
## 2     67842      33635  5   98.905 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(logistic_reg2,logistic_reg5, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: numclaims_bin ~ veh_age_cat
## Model 2: numclaims_bin ~ veh_body_cat + veh_age_cat + area_cat + agecat_cat
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     67854      33746
```

```
## 2     67842      33635 12   111.89 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(logistic_reg3,logistic_reg5, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: numclaims_bin ~ area_cat
## Model 2: numclaims_bin ~ veh_body_cat + veh_age_cat + area_cat + agecat_cat
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     67853      33754
## 2     67842      33635 11   119.13 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(logistic_reg4,logistic_reg5, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: numclaims_bin ~ agecat_cat
## Model 2: numclaims_bin ~ veh_body_cat + veh_age_cat + area_cat + agecat_cat
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     67853      33698
## 2     67842      33635 11   63.237 2.308e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Non-Nested Models

```r
AICs<-c(logistic_reg1$aic,logistic_reg2$aic,logistic_reg3$aic,logistic_reg4$aic)
AICs
```

```
## [1] 33751.47 33750.45 33759.70 33703.80
```

Select 2 Logistic Regressions

Model Selection

Poisson Regression vs Logistic Regression

- You cannot use likelihood-based statistics like AIC to compare across models with different likelihood functions.
- Difference in likelihood functions will account for the differences in the AIC probably more than differences in fit.

1. Poisson regression: Poisson function
2. Logistic regression: Bernoulli function.

- Recommend broader approaches to choose the model
  - Predicted outcomes. MSE, cross-validation, etc.
  - Intuitive interpretation of coefficients

```r
MSEs<-c(mean(poisson_reg4$residuals^2),mean(poisson_reg5$residuals^2)
       ,mean(logistic_reg4$residuals^2),mean(logistic_reg5$residuals^2))
MSEs
```

```
## [1] 14.84110 15.03182 15.95505 16.15452
```

What model would you choose?

Reference

https://en.wikipedia.org/wiki/Logistic_regression#Maximum_likelihood_estimation

https://en.wikipedia.org/wiki/Poisson_regression

https://cran.r-project.org/web/packages/insuranceData/insuranceData.pdf

http://stats.stackexchange.com/questions/139201/model-selection-can-i-compare-the-aic-from-models-of-count-data-between-