GLMs

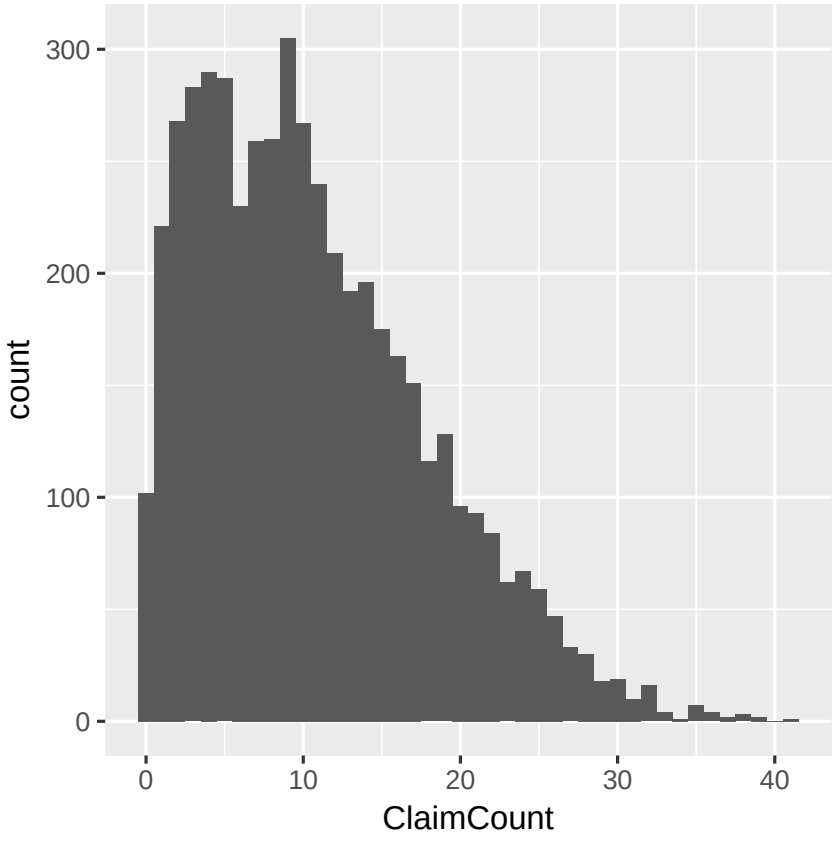
Brian A. Fannin

August 23, 2017

Fit a sample

Data

Claim counts for 5,000 policies.



How would you fit this data?

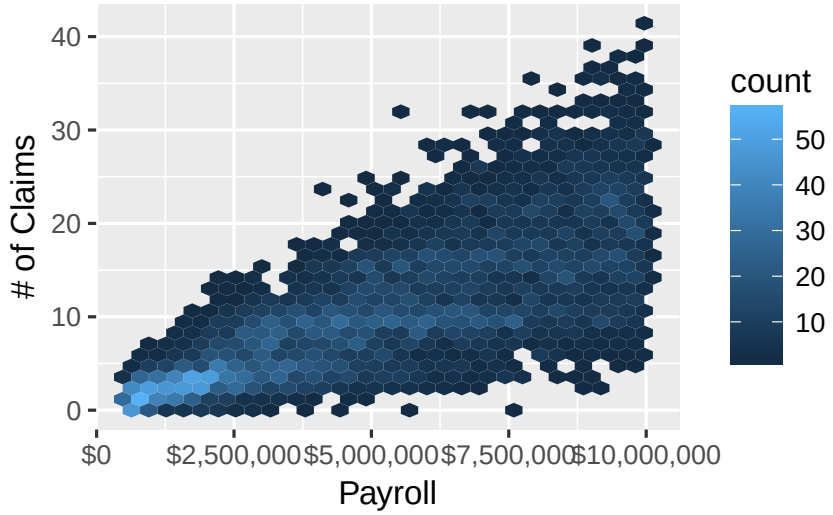
Things we can do when fitting a sample

- Pick a distribution
- Normal, lognormal, gamma, etc
- Transform data
- Often taking the log.
- Pick a fit method
- Maximum likelihood
- Least squares
- Minimum bias

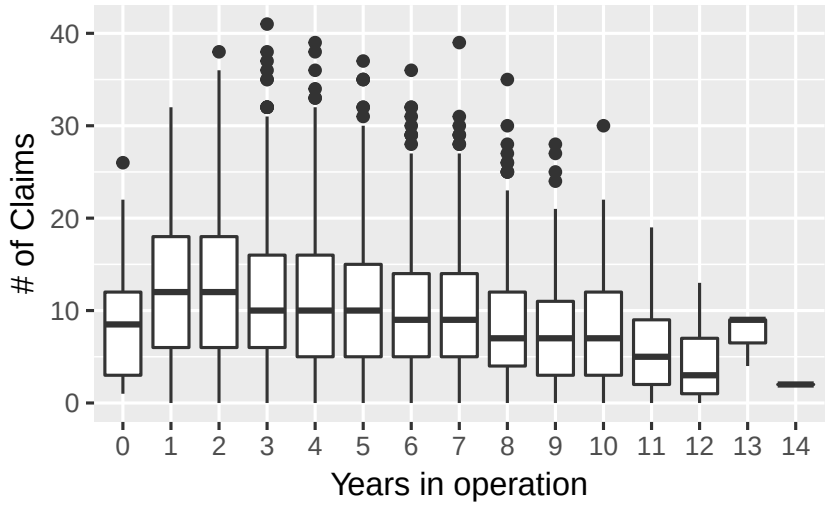
- Assess quality of fit
- r-squared, penalized r-squared
- F-stat
- Likelihood, penalized likelihood

Add predictors

Number of claims ~ Payroll

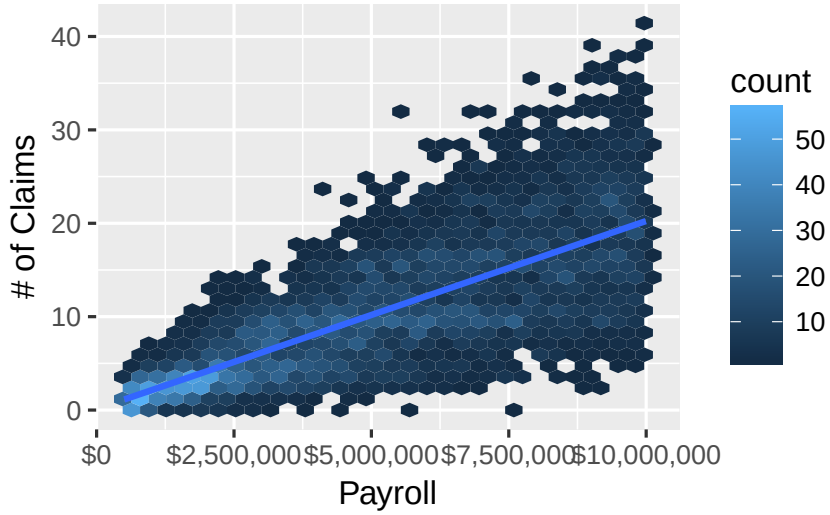


Number of claims ~ Years in operation



Number of claims ~ # of EEs

How about a linear fit?



What's wrong with this?

- Heteroskedastic
- Does it really capture the mean?

Another distribution makes more sense

But how do we do that? If only we had a linear model that was a bit more general ...

GLMs

Recall OLS Assumptions

Warning: I play fast and loose with the difference between the response variable and the error term and probably lots of other statistical things. I'm not classically trained. I play by ear.

OLS Assumptions

- Linear relationship between response and predictors: y ~ 1 + x1 + x2
- Errors are normally distributed
- Errors are uncorrelated
- Errors are homoskedastic

More general assumptions

- Relationship is between response and *transformed* linear combination of predictors
- Errors need not be normally distributed
- Distributions have some constraints

Mathematically

$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ $\mu_i = g^{-1}(\eta_i)$

$g(x)$ is the "link" function. I've seen η called the systematic component.

I don't know why the expectation is equal to the inverse of the link function. It makes my head hurt.

Models require us to specify two things

1. The distribution
2. The "link" function

Distribution restrictions

Must be one of the exponential family of functions.

$f(y; \theta, \phi) = exp[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)]$

Note this *doesn't* include the lognormal. That's OK; we can always do a log transform of our data and fit a normal.

Lots of folks get very excited about this formula. I don't. I can never remember it and I never feel as though I need to. If you like this formula, you'll see it often, but you won't see it any more today.

Canonical links

Distribution	Link	
binomical	logit	$g(x) = \frac{exp(x)}{1+exp(x)}$
gaussian	identity	$g(x) = x$
poisson	log	$g(x) = ln(x)$
Gamma	inverse	$g(x) = 1/x$

Very easy to program

A linear model:

```
fit_lm <- lm(ClaimCount ~ Payroll, data = dfGLM)
```

A GLM:

```
fit_glm <- glm(ClaimCount ~ Payroll + YearsInOperation +
    NumberOfEmployees, data = dfGLM, family = "poisson")
```

Programmatic differences:

- Must indicate the family
- Must provide the link, though only if we're using something non-canonical

Summary

```
##
## Call:
## glm(formula = ClaimCount ~ Payroll + YearsInOperation + NumberOfEmployees,
##     family = "poisson", data = dfGLM)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6316  -0.8639  -0.0898   0.6735   3.8710
##
## Coefficients:
##                      Estimate Std. Error
## (Intercept)         8.629e-01  1.833e-02
## Payroll             1.993e-07  1.714e-09
## YearsInOperation   -4.921e-02  2.042e-03
## NumberOfEmployees   4.961e-04  7.988e-06
##                    z value Pr(>|z|)
## (Intercept)          47.07   <2e-16 ***
## Payroll             116.23   <2e-16 ***
```

```
## YearsInOperation   -24.09   <2e-16 ***
## NumberOfEmployees   62.11   <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: 25650.1  on 4999  degrees of freedom
## Residual deviance:  6544.9  on 4996  degrees of freedom
## AIC: 26065
##
## Number of Fisher Scoring iterations: 4
```

Predictions

Measuring fit quality

Measuring fit quality

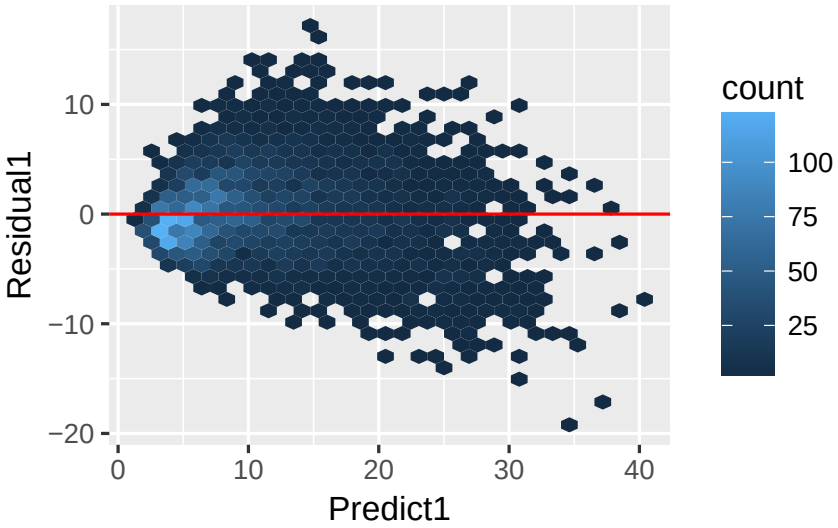
Comparing models typically involves comparison of the likelihood. Note that - comparable to r^2 - more parameters will *always* give better fit metrics, unless we're penalizing for extra parameters. AIC and BIC do this. In the formulas below, p is the number of parameters and L is the (conditional) likelihood. Lower is better.

$AIC = 2[-ln(L) + p]$ $BIC = -2L + p * ln(n)$

Deviance

- Null deviance is similar to sum of squares in OLS.
- Reduction in residual deviance suggests a better model. Again, adding parameters will *always* reduce residual deviance. Simple > complex

Residuals



Offset

The offset is a kind of scaling factor that should not be included as a predictor. Comparable to the notion of exposure in insurance pricing.

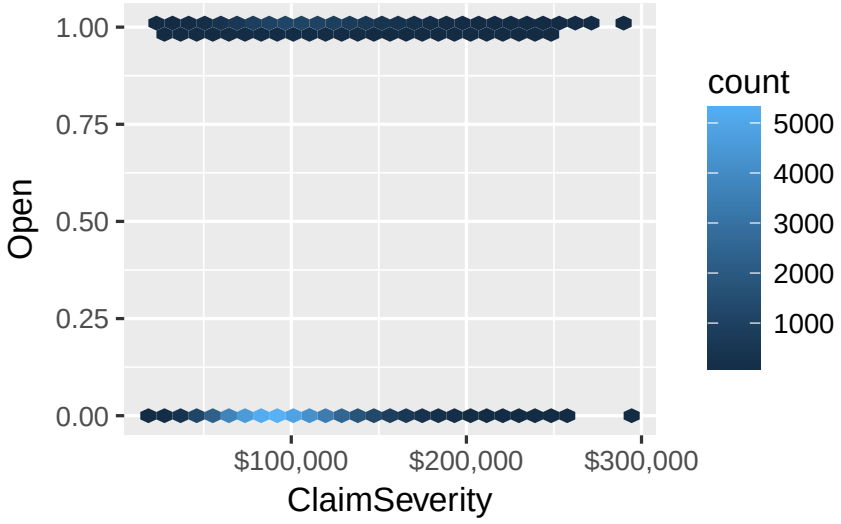
Compare these two models

```
fit_1 <- glm(ClaimCount ~ 1 + Payroll, data = dfGLM,
    family = "poisson")

fit_2 <- glm(ClaimCount ~ 1, data = dfGLM, family = "poisson",
    offset = log(Payroll))

fit_1$aic
## [1] 30611.27
fit_2$aic
## [1] 29325.83
coef(fit_1)
##  (Intercept)      Payroll
## 1.184594e+00 1.982046e-07
coef(fit_2)
## (Intercept)
##   -13.10503
```

Fit for the second model is better, because payroll isn't really a *predictor* of loss. It is a scaling element for exposure. Think the number of deaths by heart disease in Manhattan vs. number of deaths by heart disease in a rural town.

Logistic regression (if time permits)

Binomial



Fitting a logistic

```
##
## Call:
## glm(formula = Open ~ ClaimSeverity, family = "binomial", data = dfBinomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5636  -0.6921  -0.6061  -0.5146   2.1991
##
## Coefficients:
##                 Estimate Std. Error z value
## (Intercept)   -2.612e+00  3.735e-02  -69.93
## ClaimSeverity  1.189e-05  3.346e-07   35.53
##               Pr(>|z|)
## (Intercept)     <2e-16 ***
## ClaimSeverity   <2e-16 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53704  on 53555  degrees of freedom
## Residual deviance: 52436  on 53554  degrees of freedom
```

```
## AIC: 52440
##
## Number of Fisher Scoring iterations: 4
```

The logistic function

Transforms the real number range to a number between zero and one.

$f(\alpha) = \frac{exp(\alpha)}{exp(\alpha)+1}$

Watch that link function!

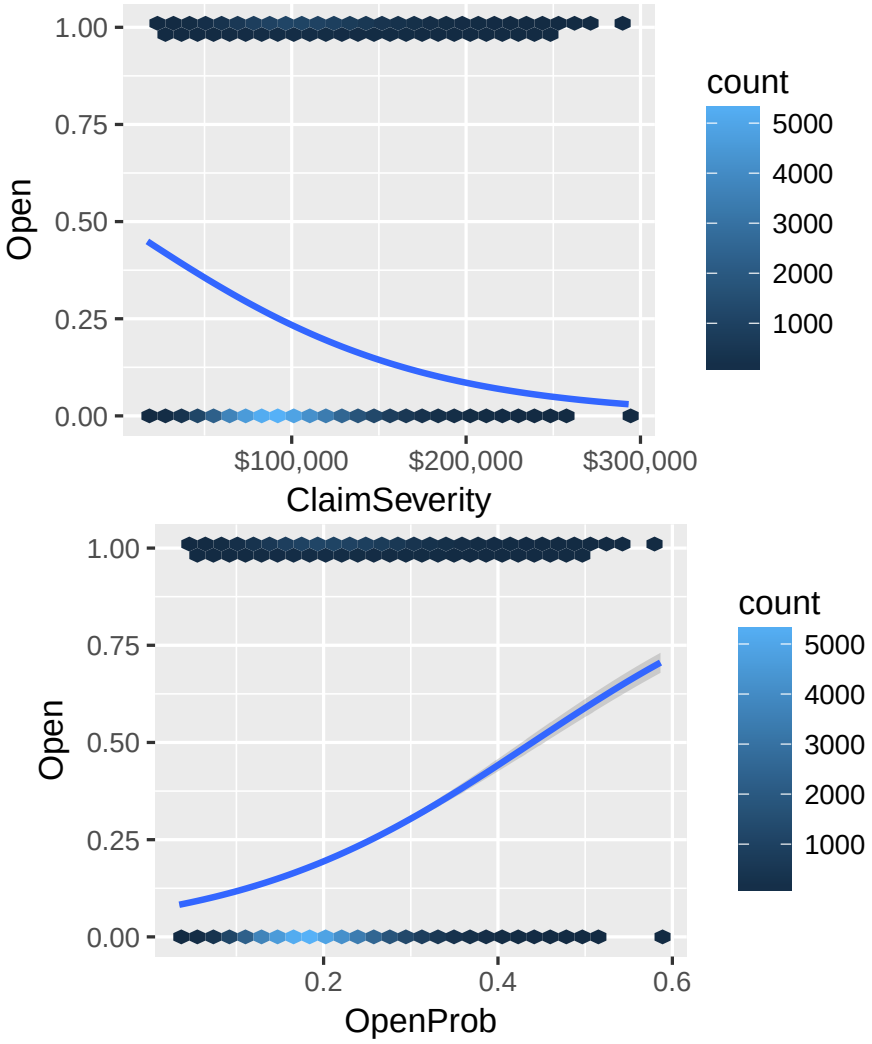
```
##
## Call:
## glm(formula = Open ~ 0 + ClaimSeverity, family = "binomial",
##     data = dfBinomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0920  -0.8070  -0.7139  -0.5616   2.6330
##
## Coefficients:
##                 Estimate Std. Error z value
## ClaimSeverity -1.186e-05  1.038e-07  -114.3
##               Pr(>|z|)
## ClaimSeverity   <2e-16 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74244  on 53556  degrees of freedom
## Residual deviance: 57845  on 53555  degrees of freedom
## AIC: 57847
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = Open ~ 1 + ClaimSeverity, family = "binomial",
##     data = dfBinomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.5636  -0.6921  -0.6061  -0.5146   2.1991
##
## Coefficients:
##                 Estimate Std. Error z value
## (Intercept)   -2.612e+00  3.735e-02  -69.93
## ClaimSeverity  1.189e-05  3.346e-07   35.53
##               Pr(>|z|)
## (Intercept)     <2e-16 ***
## ClaimSeverity   <2e-16 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53704  on 53555  degrees of freedom
## Residual deviance: 52436  on 53554  degrees of freedom
## AIC: 52440
##
## Number of Fisher Scoring iterations: 4

##
## Call:
## glm(formula = Open ~ 0 + ClaimSeverity, family = binomial(link = "identity"),
##     data = dfBinomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3319  -0.7053  -0.6116  -0.4849   2.4600
##
## Coefficients:
##                Estimate Std. Error z value
## ClaimSeverity 2.006e-06  1.702e-08   117.9
##               Pr(>|z|)
## ClaimSeverity   <2e-16 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:   Inf  on 53556  degrees of freedom
## Residual deviance: 52413  on 53555  degrees of freedom
## AIC: 52415
```

```
##
## Number of Fisher Scoring iterations: 3

##
## Call:
## glm(formula = Open ~ 1 + ClaimSeverity, family = binomial(link = "identity"),
##     data = dfBinomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3204  -0.7048  -0.6127  -0.4886   2.4387
##
## Coefficients:
##                Estimate Std. Error z value
## (Intercept)   3.400e-03  5.254e-03   0.647
## ClaimSeverity 1.973e-06  5.385e-08  36.633
##               Pr(>|z|)
## (Intercept)      0.518
## ClaimSeverity   <2e-16 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53704  on 53555  degrees of freedom
## Residual deviance: 52412  on 53554  degrees of freedom
## AIC: 52416
##
## Number of Fisher Scoring iterations: 3
```

Binomial w/fit



Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. `http://www.stat.columbia.edu/~gelman/arm/`.