Continuous Response Variables and GLMs

Statistics 149

Spring 2006

Gamma Distribution

One form of the gamma density is

$$f(y;\nu,\lambda) = \frac{y^{\nu-1}e^{-\lambda y}}{\lambda^{\nu}\Gamma(\nu)}$$

Under this parametrization

$$E[Y] = \nu\lambda = \mu \qquad \mathrm{Var}(Y) = \nu\lambda^2 = \frac{\mu^2}{\nu}$$

Based on this, we can reparameterize this distribution giving the density function,

$$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)}\frac{\nu}{\mu}\left(\frac{\nu y}{\mu}\right)^{\nu-1} e^{-y\nu/\mu}$$

One feature of this distribution is that it has a constant coefficient of variation

$$CV(Y) = \frac{\sigma}{\mu} = \frac{1}{\sqrt{\nu}}$$

So instead of a constant standard deviation, as with the normal distribution, we have a constant relative standard deviation.

This fits into the generalized linear model framework nicely with

- $g(\mu_i) = X_i\beta$
- $\mathrm{Var}(y_i) = \phi\mu_i^2$, where $\phi = \frac{1}{\nu}$

Note that this can easily be extended to a weighted situation. In this case, the variance satisfies

$$\mathrm{Var}(y_i) = \phi\frac{\mu_i^2}{w_i}$$

where the w_i are known weights.

This situation could occur when the observed y_is are averages of w_i observations.

Note that the gamma distribution is also skewed

$$E[(Y-\mu)^3] = \frac{2\mu^3}{\nu^2} > 0$$

It can be shown, that as $\nu \to \infty$ the gamma distribution approaches a normal distribution.

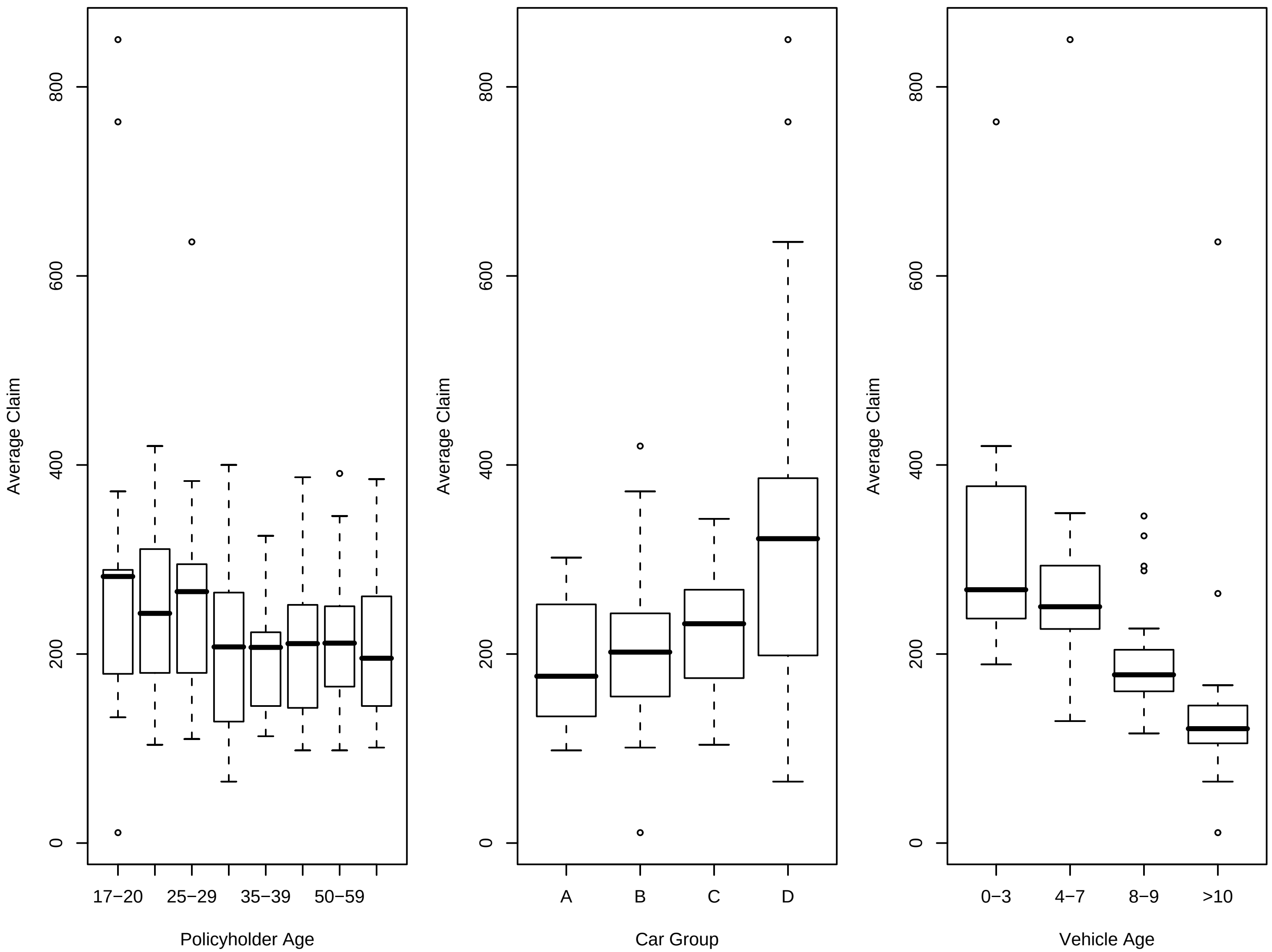
So instead of transforming your ys, assuming a gamma distribution can help with skewness problems, particularly when $\sigma \propto \mu$.

Example: Car Insurance Claims (McCullagh and Nelder, section 8.4.1)

The data involve average claims for damage for privately owned and comprehensively insured vehicles in 1975. The averages given are in pounds sterling, adjusted for inflation (data reported in 1980). Three factors are thought likely to affect the average claim.

- Policyholder's age (`policy`): 17-20, 21-24, 25-29, 30-34, 35-39, 40-49, 50-59, 60+ (8 levels)
- Car group (`group`): A, B, C, and D (4 levels)
- Vehicle age (`vehicle`): 0-3, 4-7, 8-9, 10+ (4 levels)

The number of claims m_{ijk} on which each average is based varies widely from 0 to 434. Since they vary widely, they should be included as weights in any analysis.


Average Claim
0
200
400
600
800
17−20
25−29
35−39
50−59
Policyholder Age
A
B
C
D
Car Group
0−3
4−7
8−9
>10
Vehicle Age

Based on this plot there are some clears patterns that stand out (at least to me)

- Drivers under the age of 30 tend to have higher claims
- Claims tend to increase as car group goes from A to D
- Older cars tend to have lower claims

An early analysis by (Baxter et al, 1980) fit the normal based model

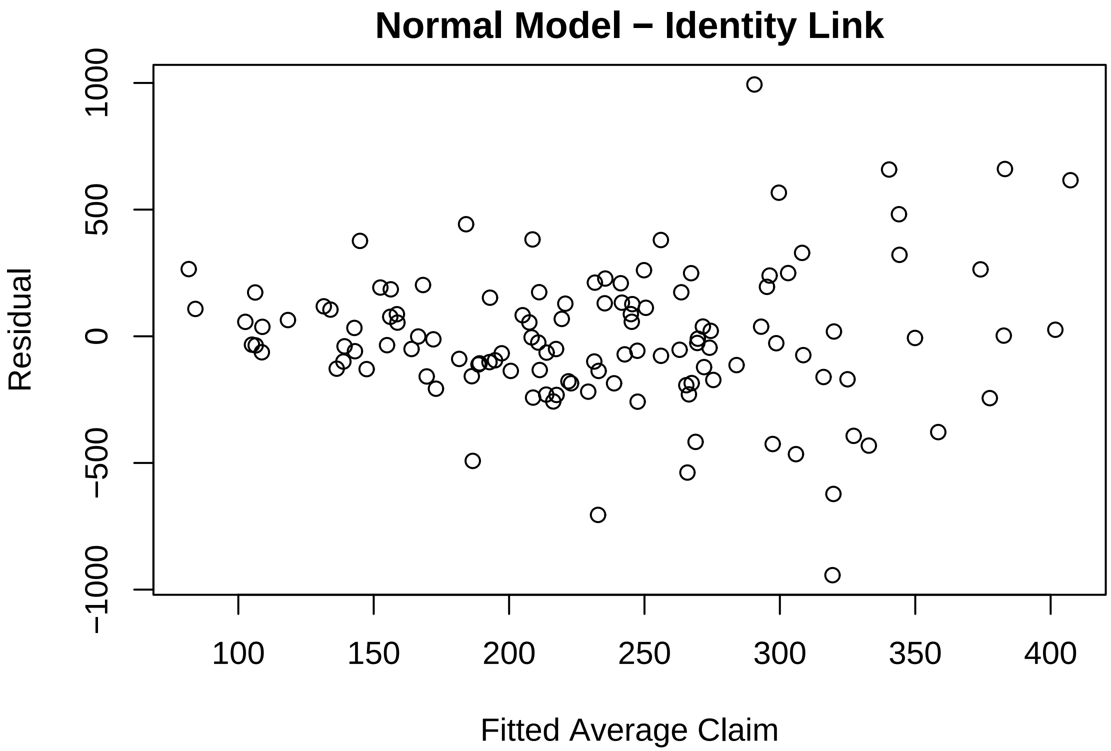
```
Call: glm(formula = claim ~ policy + group + vehicle,
     family = gaussian(), data = claims, weights = m,
     subset = m > 0)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-942.93  -136.69   -26.45   129.48   993.89
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    298.666     31.579   9.458 7.39e-16 ***
policy21-24     -5.596     33.944  -0.165 0.869359
policy25-29    -24.639     31.920  -0.772 0.441858
policy30-34    -33.225     31.719  -1.047 0.297195
policy35-39    -87.888     31.637  -2.778 0.006441 **
policy40-49    -66.987     31.112  -2.153 0.033515 *
policy50-59    -63.347     31.249  -2.027 0.045085 *
policy60+      -63.147     31.572  -2.000 0.047973 *
groupB          -2.462      9.384  -0.262 0.793489
groupC          34.184     10.026   3.410 0.000913 ***
groupD         108.660     12.235   8.881 1.51e-14 ***
vehicle4-7     -24.206      6.690  -3.618 0.000452 ***
vehicle8-9     -76.752     11.121  -6.901 3.56e-10 ***
vehicle>10    -126.635     14.746  -8.588 6.95e-14 ***
---
(Dispersion parameter for gaussian family taken to be 82604.51)
```

This analysis agrees with the patterns suggested in the box plots.

When looking at the residual plot,



the constant CV assumption $\sigma \propto \mu$ doesn't look unreasonable, as the plot has a rough megaphone shape.

When fitting a gamma model, the log likelihood has the form (assuming ν a known constant)

$$l(\beta) = \sum_{i=1}^{n} \nu(-y_i/\mu_i - \log \mu_i)$$

If there are weighted observations, as there are in the example, the log likelihood gets adjusted to

$$l(\beta) = \sum_{i=1}^{n} w_i \nu(-y_i/\mu_i - \log \mu_i)$$

The usual form for the deviance (assuming weighted observations) is

$$X^2 = -2 \sum_{i=1}^{n} \nu w_i \left(\log \frac{y_i}{\hat{\mu}_i} - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)$$

This will take the value 0 when there is a perfect fit (i.e. $\hat{\mu}_i = y_i$)

The canonical link for the gamma is the reciprocal function (`link="inverse"`). This can be seen from the log likelihood function

$$l(\beta) = \sum_{i=1}^{n} \nu(-y_i/\mu_i - \log \mu_i)$$

One potential problem with this link function is that it can lead to negative fitted $\hat{\mu}_i$. One approach to dealing with this problem is to constrain the $\hat{\beta}$ to give positive $\hat{\mu}$s in the ranges of Xs of interest.

An approach is to use the log link (`link="log"` in **R**), as this will enforce $\hat{\mu}_i > 0$, which is a requirement of the gamma distribution.

The third link available in **R** is the identity link (`link="identity"`).

Fitting the example data with the canonical inverse link gives

```
Call:
glm(formula = claim ~ policy + group + vehicle, family = Gamma(),
    data = claims, weights = m, subset = m > 0)

Deviance Residuals:
      Min          1Q      Median          3Q         Max
-3.275886  -0.486974  -0.008689   0.588895   3.285718
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.411e-03  4.179e-04   8.161 6.30e-13 ***
policy21-24  1.014e-04  4.362e-04   0.232 0.816659
policy25-29  3.500e-04  4.124e-04   0.849 0.397929
policy30-34  4.623e-04  4.106e-04   1.126 0.262639
policy35-39  1.370e-03  4.192e-04   3.268 0.001447 **
policy40-49  9.695e-04  4.046e-04   2.396 0.018281 *
policy50-59  9.164e-04  4.079e-04   2.247 0.026687 *
policy60+    9.201e-04  4.157e-04   2.213 0.028954 *
groupB       3.765e-05  1.687e-04   0.223 0.823772
groupC      -6.139e-04  1.700e-04  -3.611 0.000463 ***
groupD      -1.421e-03  1.806e-04  -7.867 2.84e-12 ***
vehicle4-7   3.663e-04  1.009e-04   3.632 0.000430 ***
vehicle8-9   1.651e-03  2.268e-04   7.281 5.45e-11 ***
vehicle>10   4.154e-03  4.423e-04   9.391 1.05e-15 ***
---
(Dispersion parameter for Gamma family taken to be 1.209015)
```

```
    Null deviance: 649.87  on 122  degrees of freedom
Residual deviance: 124.78  on 109  degrees of freedom
AIC: 84702
```

Note that this analysis agrees with the basic pattern seen in the boxplots of the data. With the inverse link

$$\hat{\mu}_{ijk} = \frac{1}{\hat{\mu}_0 + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_k}$$

a negative $\hat{\alpha}$, $\hat{\beta}$, or $\hat{\gamma}$ leads to increasing $\hat{\mu}$.

So the switching of signs here from the normal based analysis earlier is to be expected and consistent with it.

Estimating ϕ and ν

As mentioned before

$$\mathrm{Var}(y_i) = \phi\mu_i^2 = \frac{\mu_i^2}{\nu}$$

So any inference will need to account for this parameter.

As with the quasi likelihood analyzes with binomial and Poisson like data, we can use the Pearson residuals to estimate ϕ as

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)^2 \stackrel{n\to\infty}{\longrightarrow} \phi$$

where p is the number of parameters estimated.

This is the estimate that **R** gives. From this, we can estimate the coefficient of variation for a single observation by

$$\widehat{CV}(y_i) = \sqrt{\hat{\phi}}$$

Its possible to base estimates of ϕ on the deviance, X^2, instead of the Pearson statistic. However they tend to work less well. You can get into problems with y_i very close to 0 (it blows up). Also there are problems with consistency of estimators, particularly with estimating the coefficient of variation $\sqrt{\phi}$.

However the method of moment estimate based on X_p^2 will lead to consistent estimators of ϕ and $\sqrt{\phi}$.

For the example, $\hat{\phi} = 1.21$, $\widehat{CV}(y_i) = 1.1$.

Note that $\nu = \frac{1}{\phi}$ is the standard shape parameter of the gamma distribution. For the example $\hat{\nu} = 0.83$, suggesting that each observation looks roughly exponential, which is the case when $\nu = 1$.

Inference Procedures

Inference in this case is similar to before. First, if ϕ is known

$$z = \frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta})} \overset{approx.}{\sim} N(0,1)$$

Since ϕ isn't usually known and thus estimated, inference on individual βs is based on t_{n-p} distributions, similarly to the quasi-binomial and quasi-Poisson analyzes discussed earlier.

Also, as before, since there is more uncertainty since ϕ is unknown, using a heavier tailed distribution is not unreasonable.

For examining multiple βs, i.e. comparing models, we again will mimic the approach taken in the quasi-likelihood analyzes.

As before, if ϕ is known,

$$X^2(\text{Reduced Model}) - X^2(\text{Full Model}) \overset{approx.}{\sim} \phi\chi^2_{df_1}$$

where df_1 is the difference in the number of parameters fit in the two models. However since ϕ isn't known, inference will be based on

$$F = \frac{(X^2(\text{Reduced Model}) - X^2(\text{Full Model}))/df_1}{X_p^2(\text{Full Model})/df_2}$$
$$= \frac{(X^2(\text{Reduced Model}) - X^2(\text{Full Model}))/df_1}{\hat{\phi}}$$

where $df_2 = n - p$ is the degrees of the freedom for the residual deviance. This should be compared to and F_{df_1,df_2} distribution.

For example, we can compare the main effects model with the model with containing all 2-way interactions (assuming inverse link) as follows

```
> anova(claims.inv, claims.inv2, test='F')
Analysis of Deviance Table

Model 1: claim ~ policy + group + vehicle
Model 2: claim ~ (policy + group + vehicle)^2
  Resid. Df Resid. Dev  Df Deviance      F Pr(>F)
1       109    124.783
2        58     65.585  51   59.198 1.0487 0.4285
```

In this case, there is little evidence that including the 2-way interactions improves the fit. The only interaction that looks somewhat interesting is the `policy:group`, and it doesn't look to be significant

```
> anova(claims.inv, claims.inv3, test='F')
Analysis of Deviance Table

Model 1: claim ~ policy + group + vehicle
Model 2: claim ~ policy + group + vehicle + policy:group
  Resid. Df Resid. Dev  Df Deviance    F Pr(>F)
1       109    124.783
2        88     90.749  21   34.034 1.43 0.1265
```

If we were to check the main effects, they would all appear to be important. Though not quite right since the design isn't quite balanced to the the empty cells, and thus the necessary contrasts aren't orthogonal, the ANOVA table which follows shows the basic pattern.

```
> anova(claims.inv, test='F')
Analysis of Deviance Table

Model: Gamma, link: inverse

Response: claim

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                      122     649.87
policy   7    82.18       115     567.69  9.7101 2.373e-09 ***
group    3   228.31       112     339.38 62.9462 < 2.2e-16 ***
vehicle  3   214.60       109     124.78 59.1672 < 2.2e-16 ***
```

Residual Analysis and Model Checking

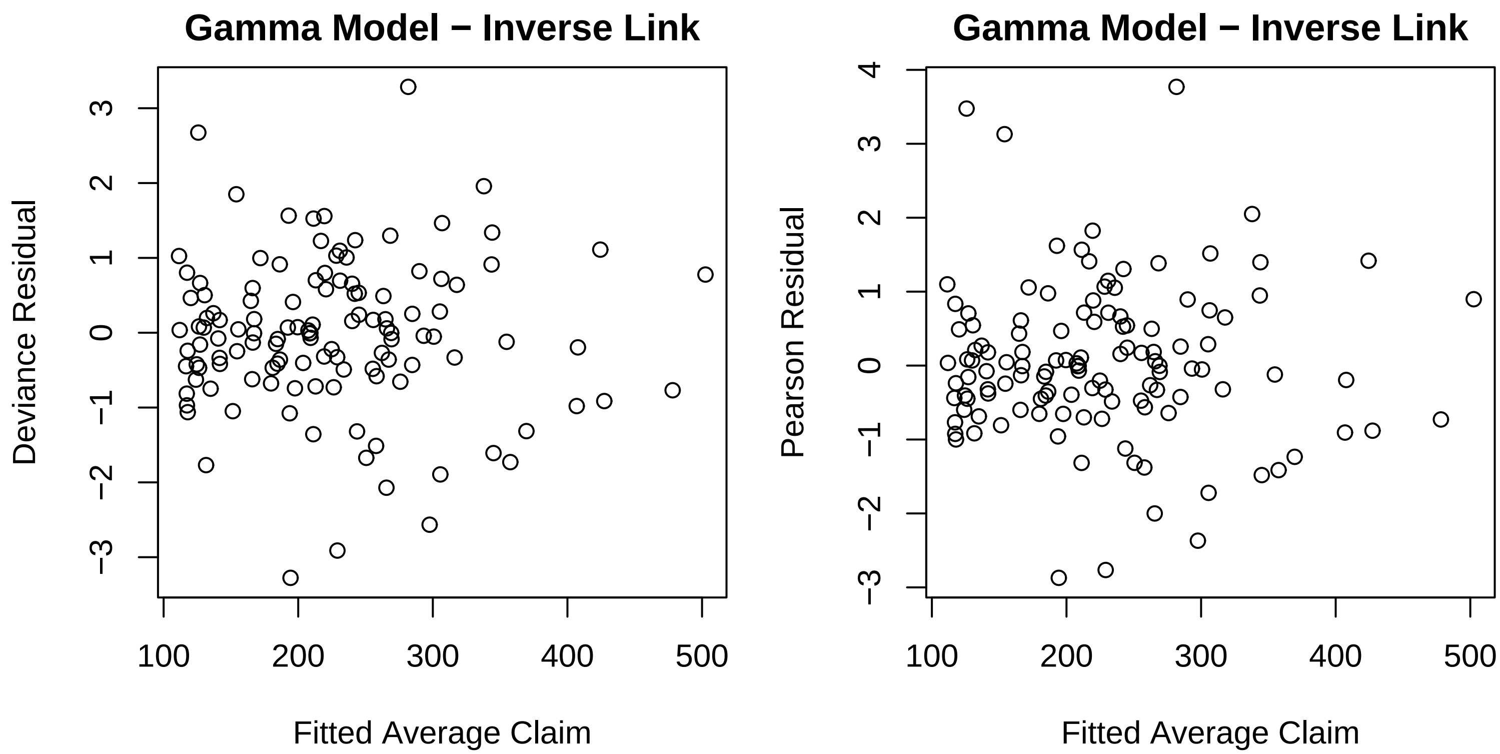
The deviance residuals for the gamma model are

$$Dres_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{-2\left(\log\left(\frac{y_i}{\hat{\mu}_i}\right) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)}$$

The Pearson residuals are

$$Pres_i = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$$

As before, these can be used to check for outliers and adequacy of the mean model. For the example,



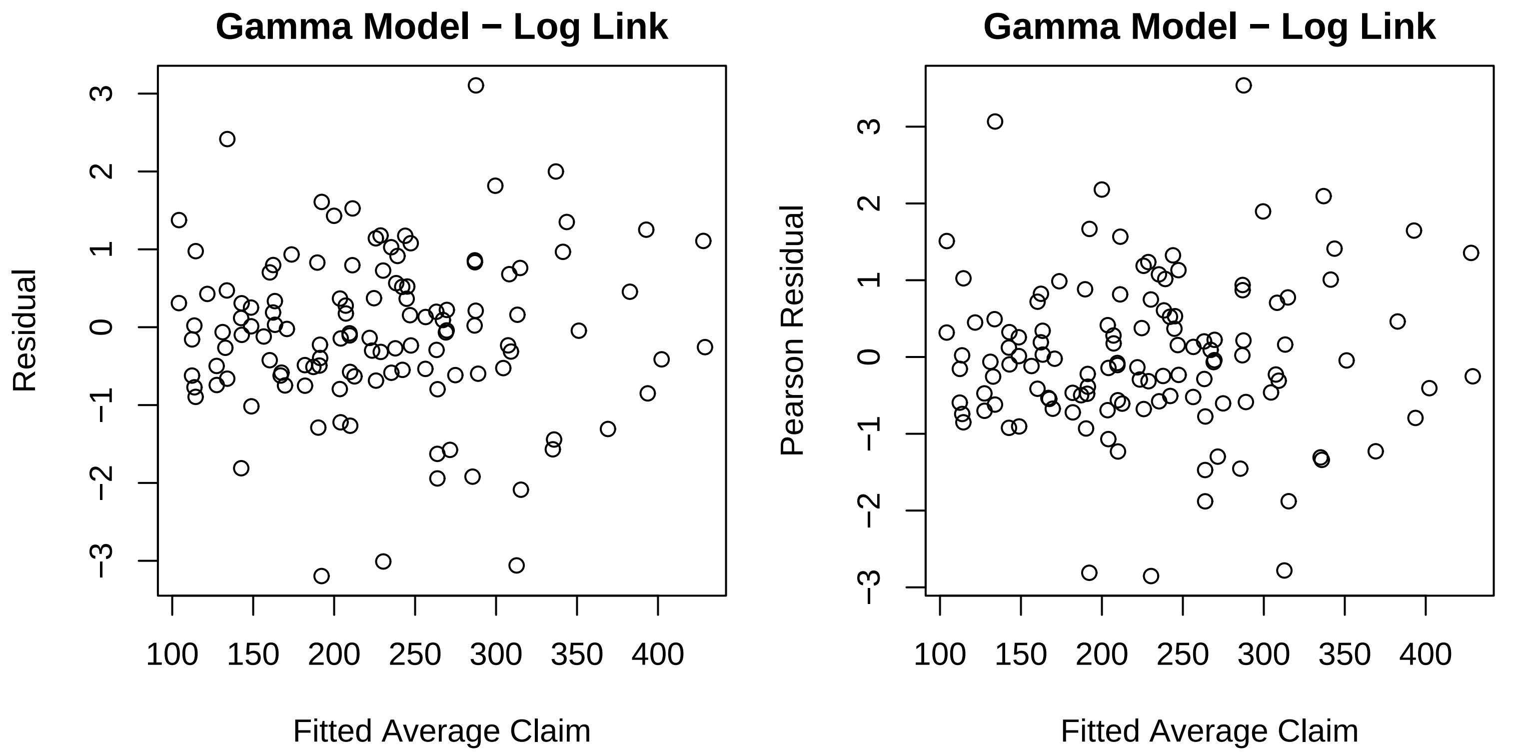
Doesn't look too bad, though maybe we over corrected on the variance (slight funnel shape). Don't see any obvious curvature, which would suggest that

$$\mu_{ijk} = \frac{1}{\mu_0 + \alpha_i + \beta_j + \gamma_k}$$

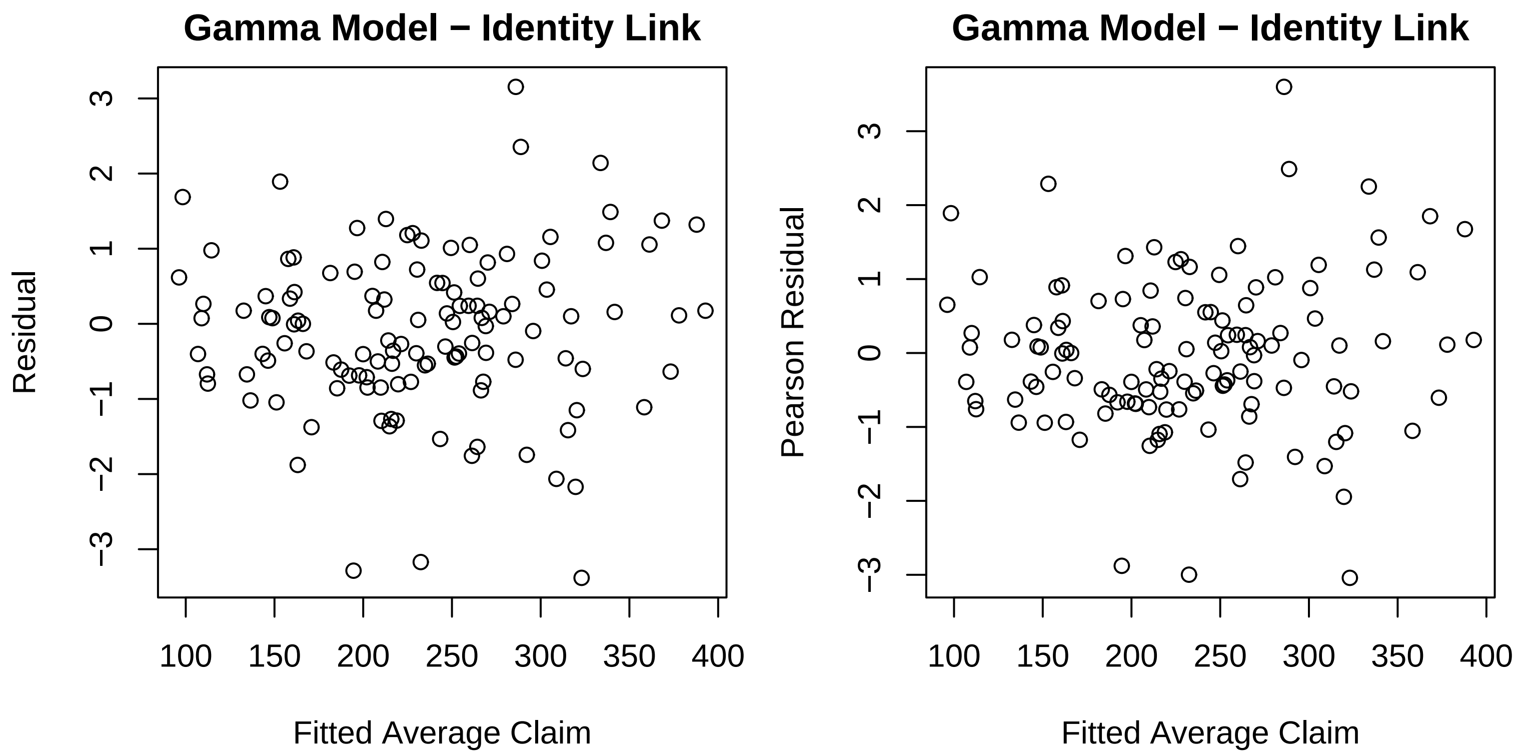
is a poor model.

There are a few outliers, mainly occurring with young drivers and car type D.

Now lets look at what happens with the other links available in **R**.

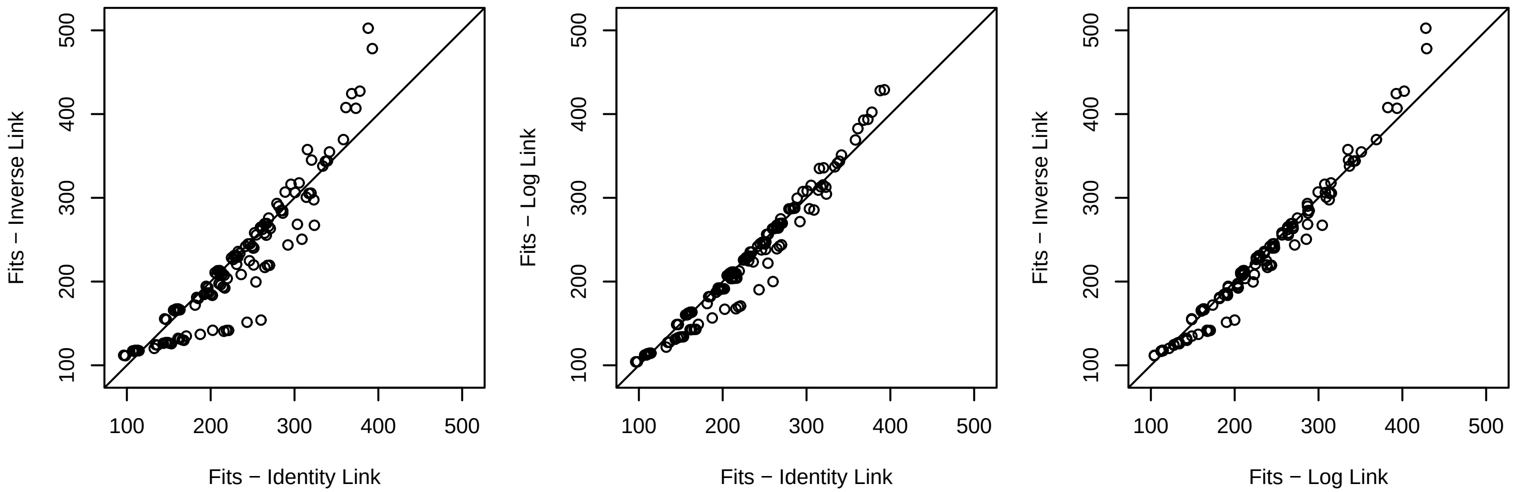


Looks about the same. If there was a problem with the inverse, there is probably a similar problem here.



Probably looks a bit better. The outliers still exist, but there is less of a funnel shape.

While the residual plots for the three different link functions look similar, there are difference, as can be seen by comparing the fitted values for the different models.



Note that the biggest difference in the fits occurs with inverse and identity links. This is not surprising as $\frac{1}{x}$ is a stronger transformation than $\log x$.

Checking for goodness of fit is more difficult in the case of continuous responses. First the goodness of fit type tests available in the binomial and Poisson cases aren't available here. The χ^2 distributional approximations don't work here since

- ϕ unknown
- Even is ϕ is known, its usually not a good approximation, since often there are few repeated observations.

Generally to check you need to look at residual plots and comparing models. For example, with the insurance claims data, adding the 2-way interaction terms doesn't give a significantly better fit.

If there are repeated observations (i.e. multiple observations with the same levels of the predictor values), we can do a bit better.

The idea is to fit a different mean for each unique combination of the predictor variables (the full model). This is compared to the model of interest (the reduced model) with the F test discussed earlier.

In the case of normal responses, this is just the standard lack of fit F test.

For the claims example, this test won't work since there are no repeated observations. If you try calculating the residual degrees of freedom you get 0. Also the estimate of $\hat{\phi}$ for the full model here is undefined $(= \frac{0}{0})$