Poisson Regression with OverDispersion

Prof Boehm Vock

Learning Goals: - Explore differences between quasipoisson and negative binomial regression - Use graph of mean vs variance to choose between Poisson, QuasiPoisson, and Negative Binomial model. - Identify pros and cons of quasipoisson and negative binomial modeling approaches

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data(NMES1988, package = "AER")
NMES1988 <- as_tibble(NMES1988)

NMES1988
```

```
## # A tibble: 4,406 x 19
##    visits nvisits ovisits novisits emergency hospital health  chronic adl
##     <int>   <int>   <int>    <int>     <int>    <int> <fct>     <int> <fct>
##  1      5       0       0        0         0        1 average       2 normal
##  2      1       0       2        0         2        0 average       2 normal
##  3     13       0       0        0         3        3 poor          4 limited
##  4     16       0       5        0         1        1 poor          2 limited
##  5      3       0       0        0         0        0 average       2 limited
##  6     17       0       0        0         0        0 poor          5 limited
##  7      9       0       0        0         0        0 average       0 normal
##  8      3       0       0        0         0        0 average       0 normal
##  9      1       0       0        0         0        0 average       0 normal
## 10      0       0       0        0         0        0 average       0 normal
## # i 4,396 more rows
## # i 10 more variables: region <fct>, age <dbl>, afam <fct>, gender <fct>,
## #   married <fct>, school <int>, income <dbl>, employed <fct>, insurance <fct>,
## #   medicaid <fct>
```

```
help(NMES1988, package = "AER")
```

I"ll filll in the answers in the first section later, but I did fill in starting with #9 Consider the variable `hospital` as the response variable.

1. Conduct some EDA to determine which of the following variables may be potentially related to the number of hospitalizations: health, chronic, adl, income.

2. Which of our potential explanatory variables are related to *each other*? (We still need to think about issues of multicollinearity!)
3. Fit four separate Poisson regression models, each with a single predictor (health, chronic, adl, income) with `hospital` visits as the response. Which variable is by itself the best predictor? How do you know?
4. Consider again the same list of variables: health, chronic, adl, and income, but this time use the response variable of `visits`. Do some EDA to determine which variables might be related to the response variable of `visits` (number of physician office visits).

In class

In this section, we use `hospital` as the response variable.

5. Fit two different models with `hospital` as response that you think might be "good." Which one is better? How do you know?
6. Interpret the intercept and at least one slope from your model.
7. Check the plot of deviance residuals vs fitted values for evidence of nonlinearity.
8. Is there evidence of overdispersion?

In the next section, we will use `visits` as the response variable, and start with the following model:

```
visits1 <- glm(visits ~ health + chronic + adl + income,
               family = poisson,
               data = NMES1988)
summary(visits1)
```

```
##
## Call:
## glm(formula = visits ~ health + chronic + adl + income, family = poisson,
##     data = NMES1988)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.420052   0.011953 118.804  < 2e-16 ***
## healthpoor       0.235449   0.017921  13.138  < 2e-16 ***
## healthexcellent -0.352553   0.030304 -11.634  < 2e-16 ***
## chronic          0.163730   0.004495  36.422  < 2e-16 ***
## adllimited       0.062647   0.015650   4.003 6.25e-05 ***
## income           0.007347   0.002044   3.594 0.000326 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 24365  on 4400  degrees of freedom
## AIC: 37152
##
## Number of Fisher Scoring iterations: 5
```

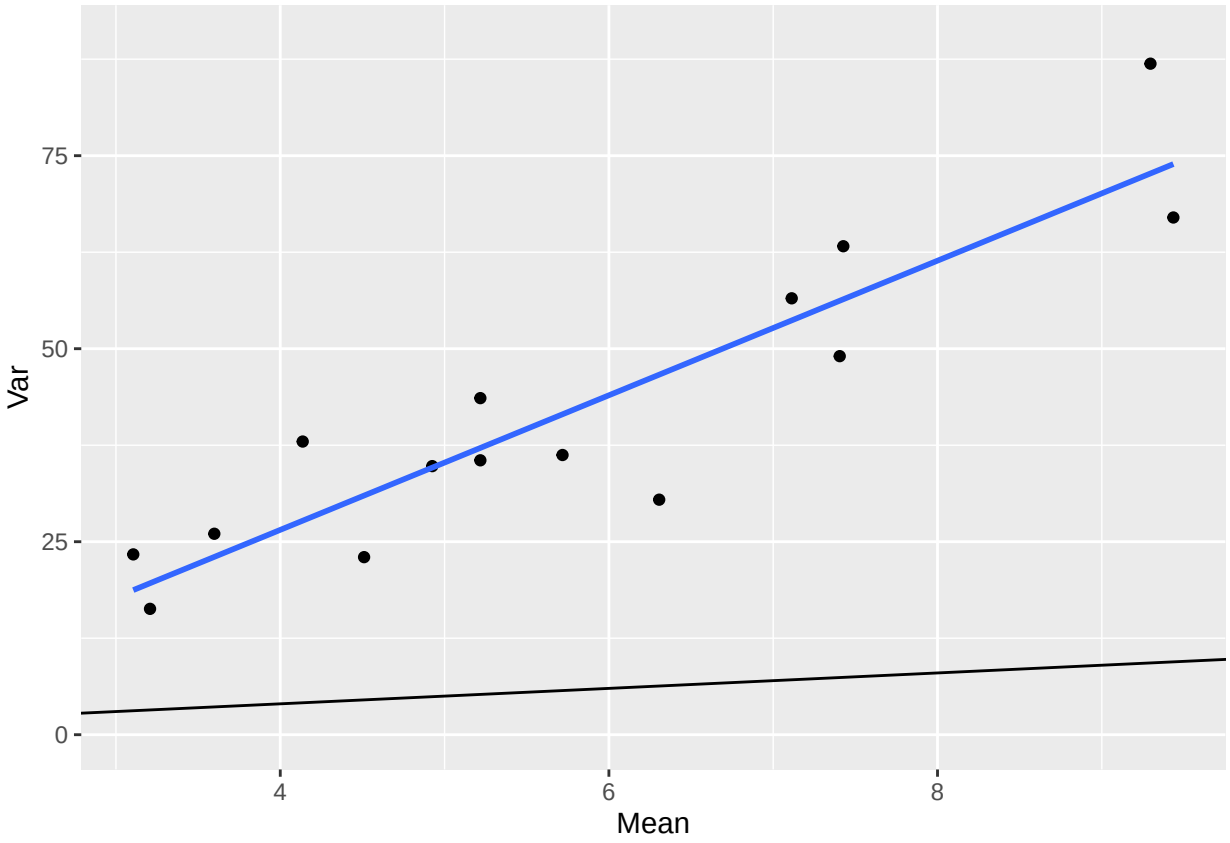
9. Is there evidence of overdispersion? How do you know?

Yes, residual deviance (24365) is much larger than df (4400)

10. Create a plot of mean vs variance. For this example, if you have a mutliple Poisson regression model you can do your grouping based on the predicted values. (see code below)

```
NMES1988 %>%
  mutate(pred = predict(visits1),
         grouping = cut_number(pred, 15)) %>%
  group_by(grouping) %>%
  summarize(Mean = mean(visits),
            Var = var(visits )) %>%
  ggplot(aes(Mean, Var)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  geom_abline(slope = 1, intercept = 0) +
  coord_cartesian(ylim = c(0 , 90))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



- Does the mean vs variance plot suggest overdispersion?

YES

- Is the relationship of mean and variance linear, or curved?

LINEAR

Negative binomial vs. Quasi Poisson

In Quasi Poisson, we assume

$$Var(Y) = \phi\lambda$$

This is a LINEAR relationship between mean and variance. We can handle under or overdispersion by estimating ϕ as less than or greater than 1.

The parameter ϕ only affects the estimation of variance/standard error, so the predicted β values will be the same as the ordinary Poisson model.

In Negative binomial, we let $E(Y) = \mu$. The overdispersion paramter r is used in the relationship

$$Var(Y) = \mu + \mu^2/r$$

The paramter r must be greater than 0. Thus this model only handles OVER dispersion, and a quadratic (curved) relationship between mean and variance is assumed.

The entire likelihood is different in Negative binomial compared to Poisson, and unusually large and small values (big residuals) have different weight. This means that the estimated β values will be different than the ordinary Poisson model, although our interpretation of them is the same.

11. When the mean vs variance relationship is LINEAR, quasipoisson can work. When the mean vs variance is QUADRATIC (curved) we use the negative binomial model. Below, we fit both the quasipoisson and the negative binomial model. Compare the coefficient estimates and standard errors.

```
visits1quas <- glm(formula = visits ~ health + chronic + adl + income, family = quasipoisson,
    data = NMES1988)

visits1nb <- MASS::glm.nb(visits ~ health + chronic + adl + income,
               data = NMES1988)

summary(visits1quas)$coef  %>% round(4)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.4201     0.0320 44.3119   0.0000
## healthpoor        0.2354     0.0480  4.9004   0.0000
## healthexcellent  -0.3526     0.0812 -4.3392   0.0000
## chronic           0.1637     0.0121 13.5849   0.0000
## adllimited        0.0626     0.0420  1.4931   0.1355
## income            0.0073     0.0055  1.3405   0.1802
```

```
summary(visits1nb)$coef %>% round(4)
```

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.3611     0.0291 46.7006   0.0000
## healthpoor        0.2530     0.0507  4.9905   0.0000
## healthexcellent  -0.3438     0.0622 -5.5252   0.0000
## chronic           0.1886     0.0124 15.2676   0.0000
## adllimited        0.0829     0.0413  2.0074   0.0447
## income            0.0104     0.0053  1.9589   0.0501
```

NOTE: we should test these borderline p-values with drop in deviance instead!

```
# DO NOT EDIT THIS CHUNK
visits1nb2 <- MASS::glm.nb(visits ~ health + chronic + adl,
               data = NMES1988)
anova(visits1nb2, visits1nb, test = "Chisq")
```

```
## Likelihood ratio tests of Negative Binomial Models
```

```
##
## Response: visits
##                              Model    theta Resid. df    2 x log-lik.   Test
## 1          health + chronic + adl 1.125855      4401       -24568.10
## 2 health + chronic + adl + income 1.126974      4400       -24564.48 1 vs 2
##      df LR stat.    Pr(Chi)
## 1
## 2     1 3.623041 0.05698462
```

```
summary(visits1nb2)
```

```
##
## Call:
## MASS::glm.nb(formula = visits ~ health + chronic + adl, data = NMES1988,
##     init.theta = 1.125854515, link = log)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.38958    0.02549  54.518  < 2e-16 ***
## healthpoor       0.24692    0.05066   4.874 1.09e-06 ***
## healthexcellent -0.33606    0.06213  -5.409 6.32e-08 ***
## chronic          0.18845    0.01236  15.250  < 2e-16 ***
## adllimited       0.07636    0.04121   1.853   0.0639 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.1259) family taken to be 1)
##
##     Null deviance: 5481.6  on 4405  degrees of freedom
## Residual deviance: 5041.6  on 4401  degrees of freedom
## AIC: 24580
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.1259
##           Std. Err.:  0.0306
##
##  2 x log-likelihood:  -24568.1030
```

```
visits1nb3 <- MASS::glm.nb(visits ~ health + chronic,
               data = NMES1988)

summary(visits1nb3)
```

```
##
## Call:
## MASS::glm.nb(formula = visits ~ health + chronic, data = NMES1988,
##     init.theta = 1.124729448, link = log)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.39808    0.02518  55.520  < 2e-16 ***
## healthpoor       0.27272    0.04873   5.597 2.19e-08 ***
## healthexcellent -0.34140    0.06210  -5.498 3.84e-08 ***
```

```
## chronic          0.19148    0.01222  15.676  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.1247) family taken to be 1)
##
##     Null deviance: 5477.9  on 4405  degrees of freedom
## Residual deviance: 5041.8  on 4402  degrees of freedom
## AIC: 24582
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.1247
##           Std. Err.:  0.0306
##
##  2 x log-likelihood:  -24571.6350
```

```
visits1nb31 <- MASS::glm.nb(visits ~ health + chronic,
               data = NMES1988)

summary(visits1nb31)
```

```
##
## Call:
## MASS::glm.nb(formula = visits ~ health + chronic, data = NMES1988,
##     init.theta = 1.124729448, link = log)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.39808    0.02518  55.520  < 2e-16 ***
## healthpoor       0.27272    0.04873   5.597 2.19e-08 ***
## healthexcellent -0.34140    0.06210  -5.498 3.84e-08 ***
## chronic          0.19148    0.01222  15.676  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.1247) family taken to be 1)
##
##     Null deviance: 5477.9  on 4405  degrees of freedom
## Residual deviance: 5041.8  on 4402  degrees of freedom
## AIC: 24582
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.1247
##           Std. Err.:  0.0306
##
##  2 x log-likelihood:  -24571.6350
```

```
anova(visits1nb3, visits1nb2, test = "Chisq")
```

```
## Likelihood ratio tests of Negative Binomial Models
##
```

```
## Response: visits
##                     Model    theta Resid. df    2 x log-lik.   Test    df
## 1       health + chronic 1.124729      4402       -24571.63
## 2 health + chronic + adl 1.125855      4401       -24568.10 1 vs 2     1
##   LR stat.    Pr(Chi)
## 1
## 2 3.532098 0.06019159
```

```
# EDA Suggests no relationship as well.
NMES1988 %>%
  group_by(adl) %>%
  summarize(mean(visits), sd(visits), median(visits))
```

```
## # A tibble: 2 x 4
##   adl     `mean(visits)` `sd(visits)` `median(visits)`
##   <fct>            <dbl>        <dbl>            <int>
## 1 normal            5.39         6.33                4
## 2 limited           7.27         8.06                5
```

```
cor(NMES1988$visits, NMES1988$income)
```

```
## [1] -0.004951069
```

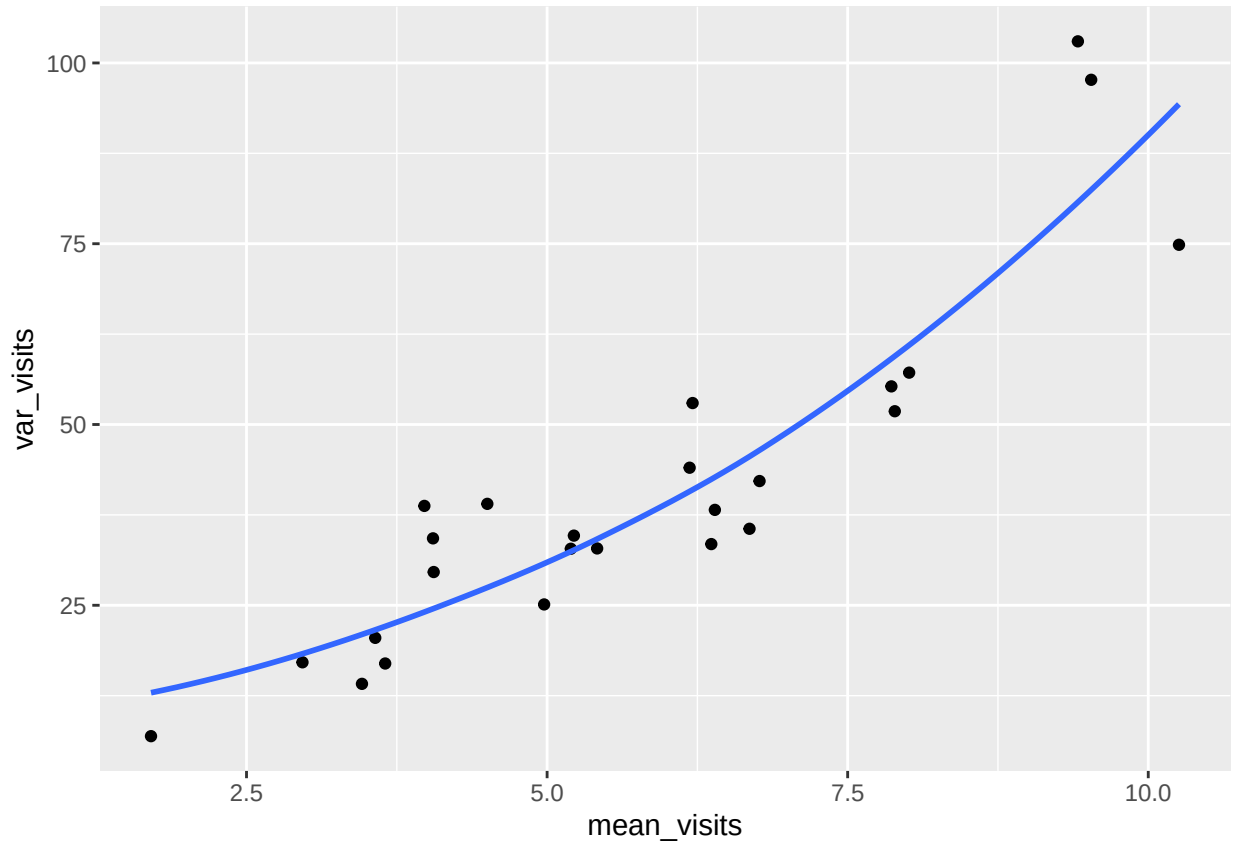
What you should see above: The linear relationship of mean and variance suggests a quasi poisson model would be appropriate. If you really want to have a likelihood based model (for example, if you want to do lots of LRT/DDTs to compare nested models), you might still go with the negative binomail. Though the assumption is a quadratic relationship, it can do ok for a linear relationship too. In this case, either model is justifiable, and we come to similar overall conclusions.

12. A different group of researchers proposes the following model instead, and create the mean vs variance plot. Which type of overdispersed model seems more appropriate: quasipoisson or negative binomial?

```
visits2 <- glm(formula = visits ~ insurance + health + chronic +
    afam + school + age,
    family = poisson,
    data = NMES1988)

# CUT/GROUP by the PREDICTED VALUES
NMES1988 %>%
  mutate(pred = predict(visits2),
         group = cut_number(pred, 25)) %>%
  group_by(group) %>%
  summarize(mean_visits = mean(visits),
            var_visits = var(visits)) %>%
  ggplot(aes(mean_visits, var_visits)) +
  geom_point() +
  geom_smooth(span = 2, se = FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The relationship of mean and variance is clearly quadratic!

13. Here again we fit both. What differences do you notice?

```
visits2quas <- glm(formula = visits ~ insurance + health + chronic +
    afam + school + age,
    family = quasipoisson,
    data = NMES1988)

visits2nb <- MASS::glm.nb(formula = visits ~ insurance + health + chronic +
    afam + school + age,
    data = NMES1988)


summary(visits2quas)$coef %>% round(4)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.2158     0.2132  5.7018   0.0000
## insuranceyes      0.2004     0.0459  4.3694   0.0000
## healthpoor        0.3222     0.0462  6.9750   0.0000
## healthexcellent  -0.3846     0.0799 -4.8132   0.0000
## chronic           0.1685     0.0118 14.2680   0.0000
## afamyes          -0.0397     0.0583 -0.6803   0.4964
## school            0.0248     0.0049  5.0208   0.0000
## age              -0.0265     0.0267 -0.9929   0.3208
```

```
summary(visits2nb)$coef %>% round(4)
```

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.9802     0.1980  4.9511   0.0000
## insuranceyes      0.2329     0.0414  5.6220   0.0000
## healthpoor        0.3693     0.0489  7.5572   0.0000
## healthexcellent  -0.3840     0.0618 -6.2145   0.0000
## chronic           0.1953     0.0121 16.1237   0.0000
## afamyes          -0.0439     0.0521 -0.8426   0.3994
## school            0.0262     0.0045  5.7665   0.0000
## age              -0.0074     0.0249 -0.2969   0.7665
```

```
summary(visits2quas)
```

```
##
## Call:
## glm(formula = visits ~ insurance + health + chronic + afam +
##     school + age, family = quasipoisson, data = NMES1988)
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.215793   0.213228   5.702 1.26e-08 ***
## insuranceyes      0.200411   0.045867   4.369 1.27e-05 ***
## healthpoor        0.322214   0.046196   6.975 3.52e-12 ***
## healthexcellent -0.384621   0.079909  -4.813 1.53e-06 ***
## chronic           0.168543   0.011813  14.268  < 2e-16 ***
## afamyes          -0.039656   0.058295  -0.680    0.496
## school            0.024779   0.004935   5.021 5.35e-07 ***
## age              -0.026528   0.026719  -0.993    0.321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 6.952921)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 23868  on 4398  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

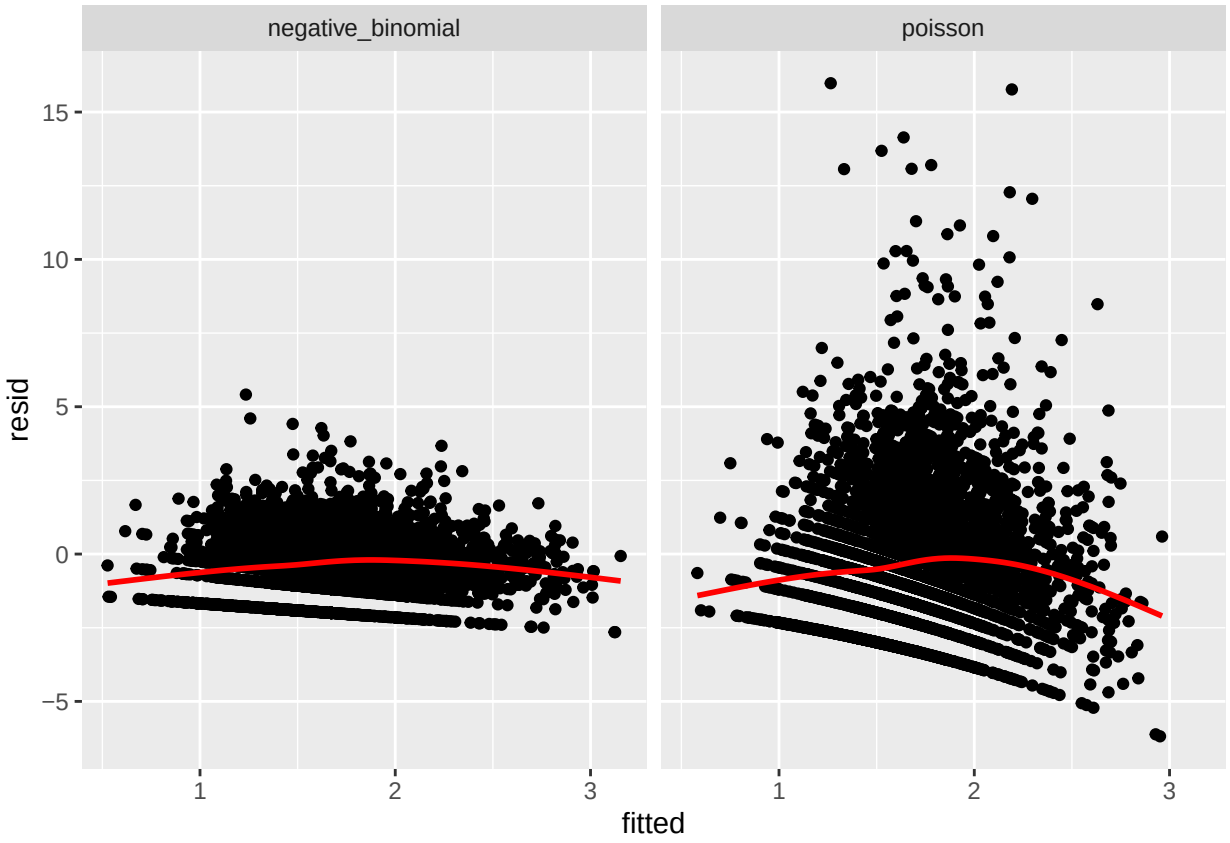
The difference in results between the two models is bigger. We should trust the Negative Binomial results here!

14. We can also examine the fitted vs deviance residuals for the poisson and the negative binomial models. Why are the deviance residuals so much smaller for the negative binomial model? How might we adjust our visualization for a better comparison?

```
fitdeviance <- data.frame(fitted = predict(visits2),
                          resid = resid(visits2),
                    model = "poisson") %>%
  bind_rows(data.frame(fitted = predict(visits2nb),
                          resid = resid(visits2nb),
                    model = "negative_binomial"))

ggplot(fitdeviance, aes(x = fitted, y = resid)) +
  geom_point() +
  geom_smooth(se = FALSE, color = "red", span = 2) +
  facet_wrap(~model)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

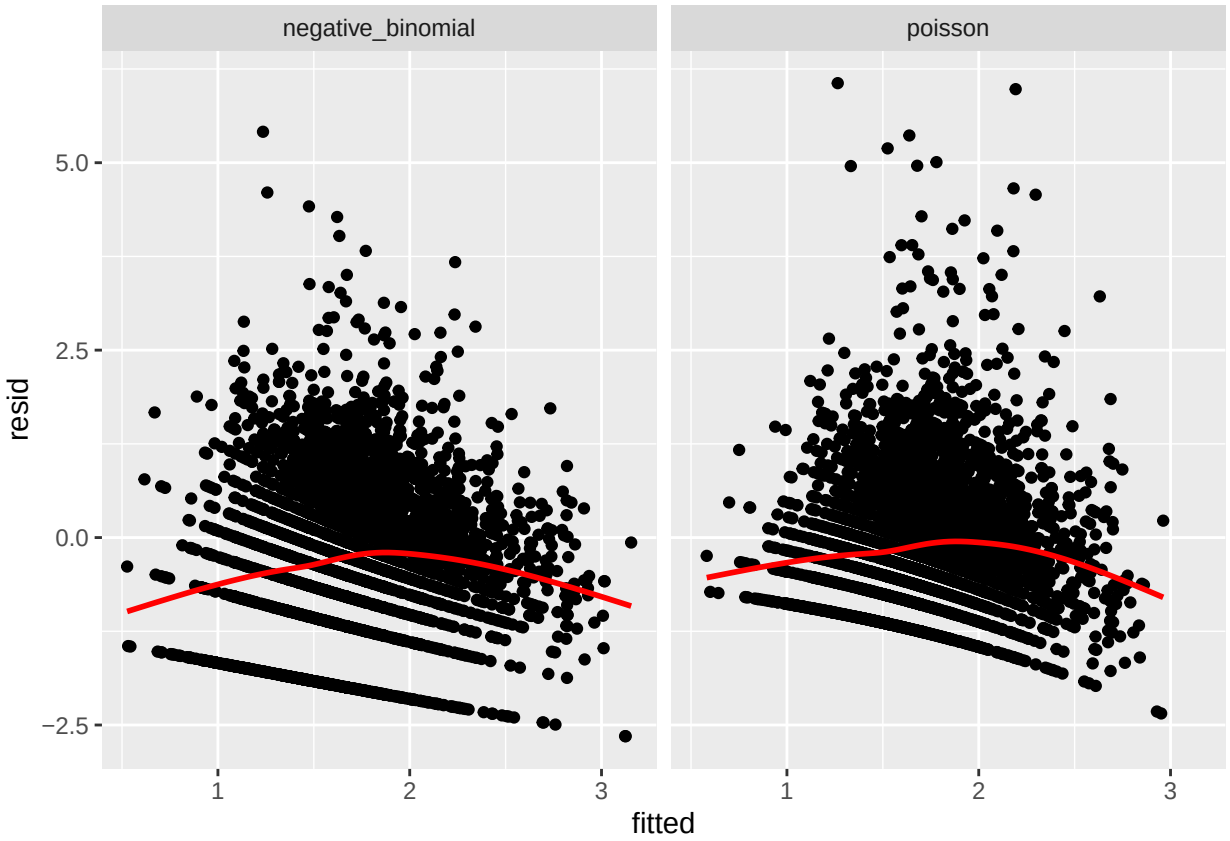


The poisson are not really scaled properly by variance. Divide those deviance residuals by sqrt(phi).

```
fitdeviance <- data.frame(fitted = predict(visits2),
                          resid = resid(visits2)/sqrt(6.95),
                    model = "poisson") %>%
  bind_rows(data.frame(fitted = predict(visits2nb),
                          resid = resid(visits2nb),
                    model = "negative_binomial"))

ggplot(fitdeviance, aes(x = fitted, y = resid)) +
  geom_point() +
  geom_smooth(se = FALSE, color = "red", span = 2) +
  facet_wrap(~model)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



15. Which model is preferred: visits1 or visits2?

```
visits1nb
```

```
##
## Call:  MASS::glm.nb(formula = visits ~ health + chronic + adl + income,
##     data = NMES1988, init.theta = 1.126974201, link = log)
##
## Coefficients:
##     (Intercept)       healthpoor  healthexcellent          chronic
##         1.36109          0.25298         -0.34383          0.18859
##      adllimited           income
##         0.08290          0.01043
##
## Degrees of Freedom: 4405 Total (i.e. Null);  4400 Residual
## Null Deviance:       5485
## Residual Deviance: 5041  AIC: 24580
```

```
visits2nb
```

```
##
## Call:  MASS::glm.nb(formula = visits ~ insurance + health + chronic +
##     afam + school + age, data = NMES1988, init.theta = 1.159808622,
##     link = log)
##
## Coefficients:
##     (Intercept)     insuranceyes       healthpoor  healthexcellent
##        0.980246         0.232883         0.369323        -0.384041
```

```
##         chronic          afamyes           school              age
##        0.195322        -0.043934         0.026170        -0.007384
##
## Degrees of Freedom: 4405 Total (i.e. Null);  4398 Residual
## Null Deviance:       5593
## Residual Deviance: 5039  AIC: 24480
```

We can use AIC if we use the negative binomial models. Visits2 has the lower AIC so is preferred. (Note however we still could probably make an even better model!)

Important notes:

- To compare two NESTED quasipoisson models, we can use anova(m1, m1, test = "F"). This an approximation to the likelihood ratio test. (Can't use LRT becuase it is not a true likelihood!)
- The negative binomial model uses a true likelihood. To compare two NESTED negative binomial models, we can use anova(m1, m1, test = "Chisq") just as with Poisson or other models! We can also compare AICs.
- The QuasiPoisson model also works in cases of underdispersion. We will see the Residual deviance as LESS than the df in the summary output and estimate $\phi < 1$.
- The Negative Binomial model can only handle OVERdispersion.
- It isn't advisable to compare the AICs of the Poisson to the Negative Binomial because the model structure is so different. We have learned to use the goodness of fit test to test for overdispersion. Other tests exist to directly compare negative binomial to poisson models, but they are beyond the scope of this class.