Generalized Linear Models

QIU Hongxiang (David)

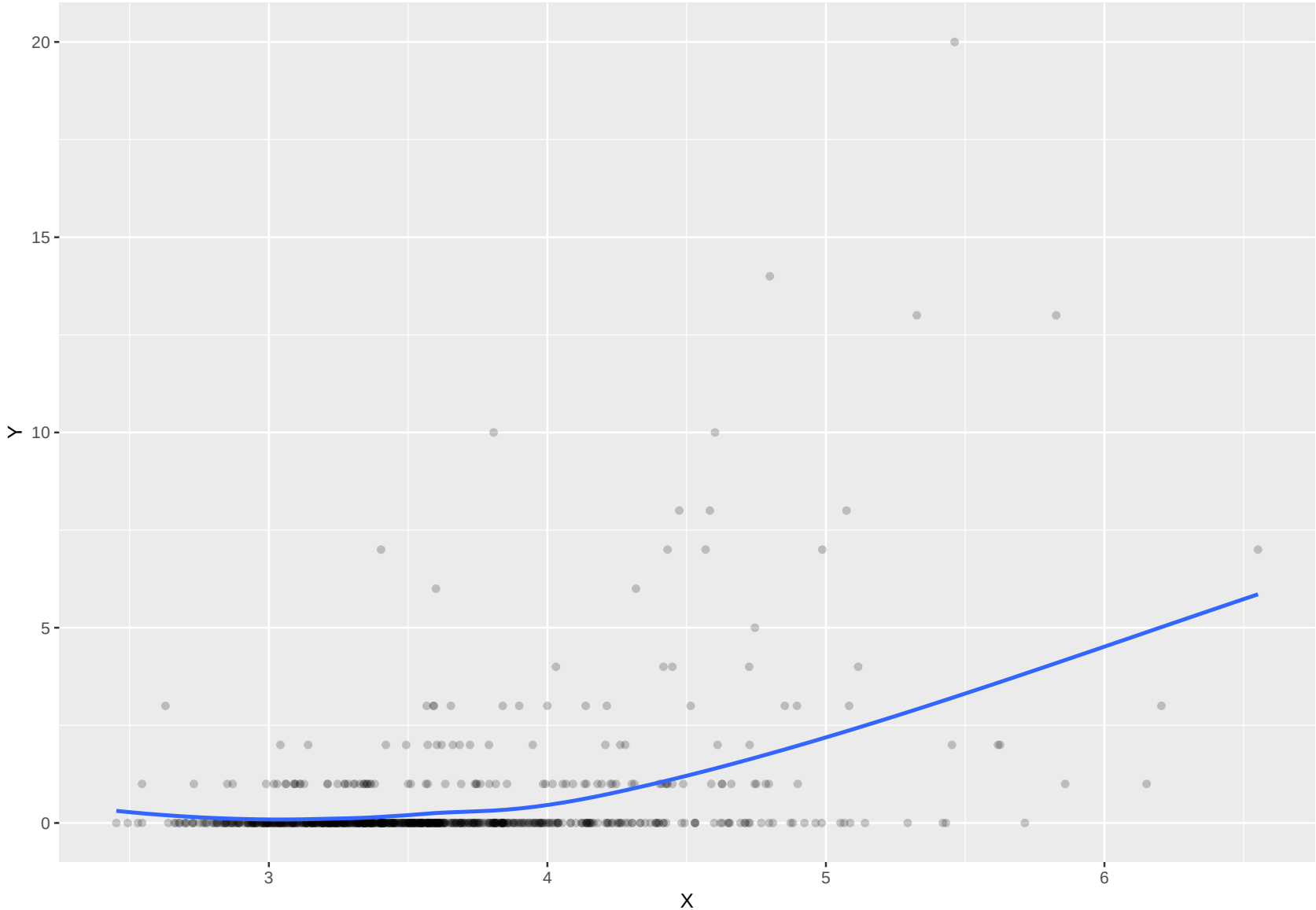
Wednesday, February 27, 2019

2. Generalized Linear Models

- Poisson regression
- Generalized linear models
- Linear regression as a GLM
- Estimating equations and robust SE
- Estimating relative risk/risk ratio (RR)
- Confidence intervals and Z-tests
- Wald tests
- Checking assumptions of GLM

Poisson regression: motivating example

In the "Teeth" data, suppose we are interested in modeling the relationship between the number of teeth extracted due to periodontal disease in a 1-year period (response) and the measure of severity of periodontal disease (predictor).



Poisson regression: motivating example

Linear regression might be fine, but

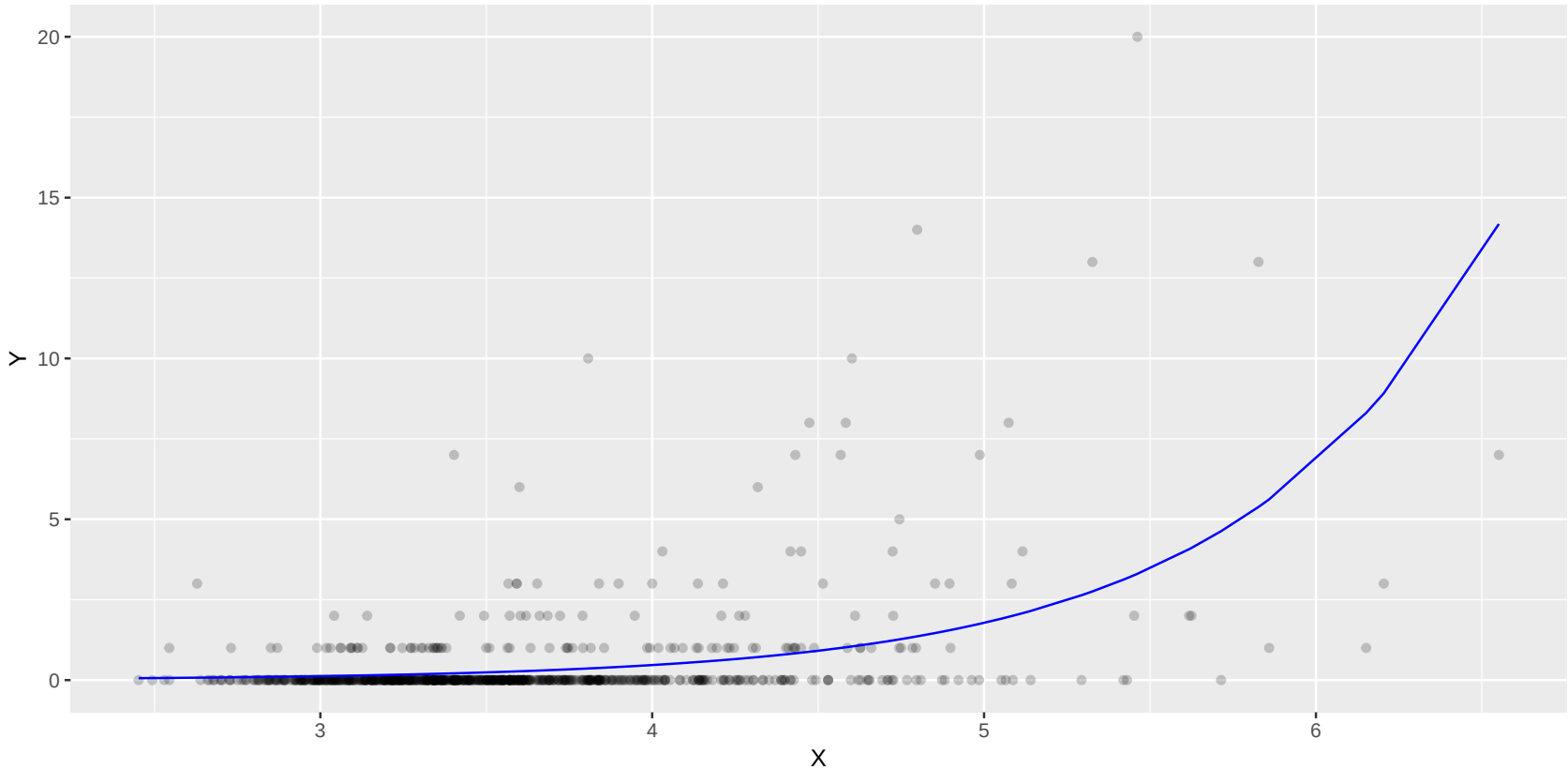
- Y is a count, so $E[Y|X]$ is always positive. Linear regression does not account for this. Besides, the classical linear regression assumes normal error, which is impossible here.
- There seems a nonlinear relationship.

Poisson regression

Poisson regression is typically used for count responses. The motivation is to take into account that the mean has to be positive and to model the count response by the Poisson distribution.
Poisson regression uses a log transform on the mean to ensure the mean is positive:

$$\log(\mu) \equiv \log(E[Y]) = \beta_0 + \beta_1 X$$

The fitted model:



Interpretation of Poisson regression coefficients

$$\log(\mu) \equiv \log(E[Y]) = \beta_0 + \beta_1 X$$

β_0 is $\log \mu$ when the predictor is 0, i.e. $X = 0$.
Translating to the original scale:

$$\beta_0 = \log(\mu(0)) \implies \mu(0) = e^{\beta_0}.$$

Therefore, e^{β_0} is the mean response when the predictor is 0.
β_1 is the difference in the log mean (log "rate" in terms of Poisson distribution) per unit difference in X:

$$\beta_1 = \log(\mu(X+1)) - \log(\mu(X))$$

Exponentiating both sides:

$$e^{\beta_1} = \mu(X+1)/\mu(X)$$

So e^{β_1} is the ratio of mean number of teeth extracted per unit difference in severity measure.

The fitted model for Teeth data

```
coef(summary(fit))
```

```
##               Estimate Std. Error   z value      Pr(>|z|)
## (Intercept) -6.108946 0.25036107 -24.40054 1.687647e-131
## X            1.337260 0.05538261  24.14585 8.255438e-129
```

The estimated intercept is -6.11. The exponentiated intercept is 0.002.

```
exp(coef(fit)[1])
```

```
## (Intercept)
## 0.002222893
```

The estimated slope is 1.34. The exponentiated slope is 3.81:

```
exp(coef(fit)[2])
```

```
##        X
## 3.808594
```

Note that the default SEs are not valid unless we really believe each Y_i is Poisson distributed with mean $\exp(\beta_0 + \beta_1 X_i)$ (or believe the Poisson mean-variance relationship $\text{var}(Y_i) = E[Y_i] = \exp(\beta_0 + \beta_1 X_i)$).

Generalized Linear Models (GLM)

All GLMs are defined in terms of two things:

1. Mean model: the scale on which the mean is analyzed (link function)
2. The mean-variance relationship (variance function)[1]

For logistic regression the link function is the logit function. For Poisson regression, we analyzed the mean response on the log scale; thus, the link function is the log function.
We can use (almost) any mathematical function as a link function. It is usually denoted by g. Thus in a general GLM, the relationship between mean response and the covariates is written as

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

If we use the subscript i to index the observations in the sample, this is written as

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{p-1,i}.$$

[1] Traditionally, this is the *random component*: the distribution of Y given its mean μ. In most cases, only the mean and variance of this distribution determines the GLM. (More on this later.)

The Variance Function

Poisson regression assumes a Poisson mean-variance relationship, i.e., $\text{var}(Y) = \mu$. The rationale for this variance-mean relationship comes from the property of the Poisson distribution that its variance is equal to its mean.
Logistic regression is based on the known form of the variance for a Bernoulli distribution: $\text{var}(Y) = \mu(1-\mu) = p(1-p)$.
We can use (almost) any mathematical function to describe the mean-variance relationship. The function is called the **variance function** and is usually denoted by v. Then the mean-variance relationship in a GLM is written as

$$\text{var}(Y) = v(\mu).$$

Sometimes there is a scale parameter σ^2. Then the mean-variance relationship in a GLM is written as

$$\text{var}(Y) = \sigma^2 v(\mu),$$

where σ^2 is the scale (or "dispersion") parameter and v is the variance function. With the observation index this is written as

$$\text{var}(Y_i) = \sigma^2 v(\mu_i).$$

Linear regression as a GLM

Recall the "error" notation for the multiple linear regression model.
Linearity:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots \beta_{p-1} X_{p-1,i} + \epsilon_i, \text{ with } E[\epsilon_i] = 0, \ i = 1, 2, \ldots, n,$$

Independence:

$$\epsilon_1, \ldots, \epsilon_n \text{ independent}$$

Constant variance:

$$\text{var}(\epsilon_i) = \sigma^2, \ i = 1, \ldots, n$$

Normality or large sample size:

$$\epsilon_i \text{ Normally distributed or } n \text{ large}$$

The multiple linear regression model in GLM notation

To make the connection with GLMs we express the model in terms of the conditional distribution of Y given the X's and remove all mention of errors. Linearity:

$$\mu_i = E[Y_i|X_{1i}, \ldots, X_{p-1,i}] = \beta_0 + \beta_1 X_{1i} + \cdots \beta_{p-1} X_{p-1,i}, \quad i = 1, 2, \ldots, n,$$

Conditional independence of observations given covariate:

$$Y_1, \ldots, Y_n \text{ conditionally independent given } (X_{1i}, \ldots, X_{p-1,i})$$

Constant variance (conditional on covariates):

$$\text{var}(Y_i|X_{1i}, \ldots, X_{p-1,i}) = \sigma^2$$

Normality of conditional distribution of Y given covariates, or large sample size:

$$(Y_i|X_{1i}, \ldots, X_{p-1,i}) \text{ Normally distributed, or } n \text{ large}$$

The Poisson Regression Model

What are the analogous equations for the Poisson regression model?
Linearity **on the logarithmic scale**:

$$\log(\mu_i) = \log(E[Y_i|X_{1i}, \ldots, X_{p-1,i}]) = \beta_0 + \beta_1 X_{1i} + \cdots \beta_{p-1} X_{p-1,i}, \quad i = 1, 2, \ldots, n,$$

Conditional independence of observations given covariate

$$Y_1, \ldots, Y_n \text{ conditionally independent given } (X_{1i}, \ldots, X_{p-1,i})$$

Variance equal to mean:

$$\text{var}(Y_i|X_{1i}, \ldots, X_{p-1,i}) = \mu_i$$

Large sample size:

$$n \text{ must be large}$$

We could assume $(Y_i|X_{1i}, \ldots, X_{p-1,i})$ is Poisson distributed, but it is already in the mean-variance relationship. Unlike linear regression, n must be large even with this assumption.

The Logistic Regression Model

Linearity **on the log-odds (logit) scale**:

$$\log \frac{\mu_i}{1-\mu_i} = g(E[Y_i|X_{1i},\ldots,X_{p-1,i}]) = \beta_0+\beta_1 X_{1i}+\cdots\beta_{p-1}X_{p-1,i}, \quad i=1,2,\ldots,n,$$

Conditional independence of observations given covariate

$$Y_1,\ldots,Y_n \text{ conditionally independent given } (X_{1i},\ldots,X_{p-1,i})$$

Mean-Variance relationship determined by the Bernoulli variance function:

$$\text{var}(Y_i|X_{1i},\ldots,X_{p-1,i}) = \mu_i(1-\mu_i)$$

Large sample size:

$$n \text{ must be large}$$

Again, we could "assume" $(Y_i|X_{1i},\ldots,X_{p-1,i})$ is Bernoulli distributed (which we know must be true), but it is already in the mean-variance relationship. Unlike linear regression, n must be large even with this assumption.

Generalized Linear Model

Here is the general form for a GLM.

Linearity **on the scale of the link function**:

$$g(\mu_i) = g(E[Y_i|X_{1i}, \ldots, X_{p-1,i}]) = \beta_0 + \beta_1 X_{1i} + \cdots \beta_{p-1} X_{p-1,i}, \quad i = 1, 2, \ldots, n,$$

Conditional independence of observations given covariate

$$Y_1, \ldots, Y_n \text{ conditionally independent given } (X_{1i}, \ldots, X_{p-1,i})$$

Mean-Variance relationship determined by the variance function:

$$\text{var}(Y_i|X_{1i}, \ldots, X_{p-1,i}) = v(\mu_i) \text{ or } \sigma^2 v(\mu_i)$$

(or assume the distribution of $Y_i|X_{1i}, \ldots, X_{p-1,i}$)

Large sample size:

n must be large

Interpretations of coefficients in a GLM

By analogy with the interpretation of parameters in linear regression, we have the following interpretations for the coefficients in a GLM.

β_0 is $g(\mu)$ where μ is the mean response when all predictors are set to 0, i.e., $X_1 = X_2 = \cdots = X_{p-1} = 0$.

β_j is the difference in $g(\mu)$ per unit difference in X_j with *all other predictors held fixed*.

Because these interpretations are in terms of the scale of the link function it is usually best to interpret the parameters in terms of the original ("raw") scale of the response variable.

Interpretation of GLM parameters on the scale of the response

If $\beta_0 = g(\mu)$ where μ is the mean response corresponding to all X's equal to 0, then $\mu = g^{-1}(\beta_0)$, where g^{-1} is the inverse function to g. For example, with Poisson regression, $g = \log$ so $g^{-1} = \exp$, and so the mean response for all X's being 0 in Poisson regression is $\exp(\beta_0)$. For logistic regression, we have $\mu = \text{expit}(\beta_0)$.

For the coefficient β_1, for simplicity consider the case with only one predictor. We have $\beta_1 = g(\mu(X+1)) - g(\mu(X))$. For Poisson regression we exponentiate to get an interpretation as a ratio of means (or "rate ratio), i.e., $\beta_1 = \log \mu(X+1) - \log \mu(X)$, so $\exp(\beta_1) = \mu(X+1)/\mu(X)$. For logistic regression, the exponentiatied coefficient is a ratio of odds ("odds-ratio").

For other GLMs the interpretation is determined by the form of the inverse-link function g^{-1}.

The 3 Most Commonly Used GLMs

Model	Link Function	Variance Function
Linear regression	μ ("Identity")	σ^2 ("constant")
Poisson regression	$\log \mu$ ("Log")	μ ("Poisson")
Logistic regression	$\log(\mu/(1-\mu))$ ("Logit")	$\mu(1-\mu)$ ("Bernoulli")

Fitting GLMs

The traditional method used for fitting a GLM to data is based on *maximum-likelihood* (ML). Recall that the ML method is based on postulating a probability distribution for the data and then choosing the parameter values (coefficients) that maximize the probability of the observed data.

1. For linear regression a normal distribution is used to derive the likelihood function, which leads to the least squares method, i.e., in this case, ML=LS.
2. For Poisson regression a Poisson distribution is used to derive the likelihood function.
3. For logistic regression a Bernoulli distribution is used.

Warning: The algorithms used to maximize the likelihood are not guaranteed to converge in general. Sometimes you will get a message saying the algorithm has failed to converge, but sometimes you will get no warning message but garbage results! But for these three GLMs, the algorithms are almost always guaranteed to converge.

Estimating Equations and Robust SEs

To calculate MLE, we can calculate the log-likelihood and maximize it.[2] We can further calculate its derivatives with respect to the β-coefficients, set them to 0 and solve the equation.[3] This approach is a special case of a more general method: estimating equation.

In the three most common GLMs, the corresponding estimating equations only depend on the link function and the mean-variance relationship. If "unusual" mean-variance relationship is used without assuming the conditional distribution $Y|X$, we use estimating equations to estimate β-coefficients.

If we assume the mean model is correct, then estimators $\hat{\beta}$ based on these estimating equations converge to the β-coefficients in the true mean model and are approximately normally distributed with large sample size *regardless of whether the assumed mean-variance relationship is true or not*. When the assumed var(Y) is wrong, the asymptotic variance is different, and we need robust SEs (also called sandwich SEs).[4]

The mean-variance relationship in the GLM can be regarded as "working mean-variance relationship" if we use robust SEs: we need a working var(Y) to estimate β, but we don't need it to be correct for statistical inference.

---

[2] E.g. gradient descent.

[3] E.g. Newton-Raphson

[4] Strictly speaking, robust SEs are valid when X is also random. When X is fixed (e.g. experiments), robust SE are usually a bit conservative. In practice we usually live with it.

Robust SEs (continued)

Robust SEs are based on a valid estimate of cov($\hat{\underline{\beta}}$):

```
library(sandwich)
(v<-vcovHC(fit))
```

```
##             (Intercept)           X
## (Intercept)  0.26599917 -0.06472412
## X           -0.06472412  0.01645752
```

The robust SE is just the square root of diagonal elements:

```
(robust.SE<-sqrt(diag(vcovHC(fit))))
```

```
## (Intercept)           X
##   0.5157511   0.1282868
```

Compare with default SEs:

```
coef(summary(fit))
```

```
##              Estimate Std. Error   z value      Pr(>|z|)
## (Intercept) -6.108946 0.25036107 -24.40054 1.687647e-131
## X            1.337260 0.05538261  24.14585 8.255438e-129
```

Estimating Relative Risk/Risk Ratio (RR) for Binary Responses

Suppose we want to estimate the "relative risk" (or "risk ratio") per unit difference in X, i.e.

$$RR = \frac{P(Y = 1|X = x + 1)}{P(Y = 1|X = x)} = \frac{\mu(x + 1)}{\mu(x)}.$$

($Y = 1$ is the "bad outcome", so $P(Y = 1)$ can be interpreted as "risk".)
The β-coefficients in GLM with a log link can be interpreted as log relative risks:

$$\log \mu(x) = \beta_0 + \beta_1 x \implies \beta_1 = \log \frac{\mu(x + 1)}{\mu(x)}.$$

So it's natural to fit a GLM with log link and Bernoulli mean-variance relationship: $\text{var}(Y) = v(\mu) = \mu(1 - \mu)$.
But the estimation procedure for this GLM does not always converge.[5] What can we do with this?

[5] The reason is $\mu(X_i)$ can be greater than 1 for certain X_i for certain β in the iterative algorithm, in which case $\text{var}(Y)$ is undefined.

Estimating Relative Risk/Risk Ratio (RR) for Binary Responses (continued)

When setting $\text{var}(Y) = v(\mu) = \mu$ (i.e. Poisson regression), the procedure always converges. Therefore, we can intentionally misspecify $\text{var}(Y)$ and use Poisson regression with robust SEs.

Central role of mean model in GLM with large sample size when inferring association

Don't let the type of response variable (continuous vs discrete vs binary etc.) determine the GLM you fit!
Let your question/hypothesis decide. The mean model $g(\mu) = X^\top \underline{\beta}$ is the core, and let robust SEs take care of the rest (variance or the distribution of $Y|X$).[6]
Want to infer the difference in proportions? Linear regression is perfectly fine.
Want to infer the difference in rates for count response? Linear regression is perfectly fine.

[6] Prediction is another story.

Summary of assumptions needed in GLMs

- Independence
- Correct mean model: $g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}$ (recall "all models are wrong"...)
- Large sample size
- (Correct mean-variance relationship or correct conditional distribution $Y|X$ if we don't use robust SEs.)

Confidence intervals and Z-tests

Confidence intervals and Z-tests for a coefficient β are based on normal approximation of $\hat{\beta}$.
Given a valid SE, the 95% CI for β is $\hat{\beta} \pm 1.96 \times SE$. If $H_0 : \beta = \beta_{H_0}$, at significance level 5%, we would reject H_0 if $|(\hat{\beta} - \beta_{H_0})/SE| > 1.96$.

Teeth data with robust SE

Confidence intervals:

```
coef(fit)+robust.SE%o%c(-1.96,1.96)
```

```
##                   [,1]      [,2]
## (Intercept) -7.119818 -5.098074
## X            1.085818  1.588702
```

Confidence interval for [the mean number of teeth extracted when severity meansure is 0 (e^{β_0})] and [the ratio of means per unit difference in severity measure (e^{β_1})]:

```
exp(coef(fit)+robust.SE%o%c(-1.96,1.96))
```

```
##                     [,1]        [,2]
## (Intercept) 0.0008089139 0.006108501
## X           2.9618609433 4.897389014
```

P-values of Z-tests for $H_0 : \beta_0 = 0$ and for $H_0 : \beta_1 = 0$:

```
2*pnorm(abs(coef(fit)/robust.SE),lower.tail=FALSE)
```

```
##  (Intercept)            X
## 2.290824e-32 1.927087e-25
```

Teeth data with robust SE

Simpler method:

```
library(lmtest)
coefci(fit,vcov.=v)
```

```
##                 2.5 %    97.5 %
## (Intercept) -7.119799 -5.098092
## X            1.085822  1.588698
```

```
coeftest(fit,vcov.=v)
```

```
##
## z test of coefficients:
##
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -6.10895    0.51575 -11.845 < 2.2e-16 ***
## X            1.33726    0.12829  10.424 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wald tests

Wald tests are a generalized version of Z-tests that can test composite hypotheses with robust SE (similar to F-tests in linear regression). Suppose we want to test the null hypothesis that k β-coefficients (say $\underline{\beta} = (\beta_1, \ldots, \beta_k)^\top$) are all 0. Given a valid estimator of $\text{cov}(\underline{\hat{\beta}})$, $\hat{\Sigma}$, the Wald test statistic is

$$\underline{\hat{\beta}}^\top \hat{\Sigma}^{-1} \underline{\hat{\beta}}.$$

It has an approximate distribution of χ^2_k with large sample size. H_0 is rejected when the test statistic is too large (like F-tests).

Teeth data with robust SE

Suppose we include both severity and age in the model:

$$\log(\mu) = \beta_0 + \beta_1 X + \beta_2 \text{Age}$$

```
fit2<-glm(Y~X+AGE,data=data,family=poisson)
```

Wald test for $H_0 : \beta_1 = \beta_2 = 0$:

```
beta2<-coef(fit2)
v2<-vcovHC(fit2)
(wald.statistic<-t(beta2[c(2,3)])%*%
        solve(v2[c(2,3),c(2,3)],beta2[c(2,3)]))
```

```
##           [,1]
## [1,] 119.5995
```

```
pchisq(wald.statistic,df=2,lower.tail=FALSE)
```

```
##               [,1]
## [1,] 1.069791e-26
```

Teeth data with robust SE

Simpler method:

```
library(lmtest)
waldtest(fit2,glm(Y~1,data=data,family=poisson),vcov=v2,test="Chisq")
```

```
## Wald test
##
## Model 1: Y ~ X + AGE
## Model 2: Y ~ 1
##   Res.Df Df Chisq Pr(>Chisq)
## 1    793
## 2    795 -2 119.6  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#need to specify test="Chisq"
```

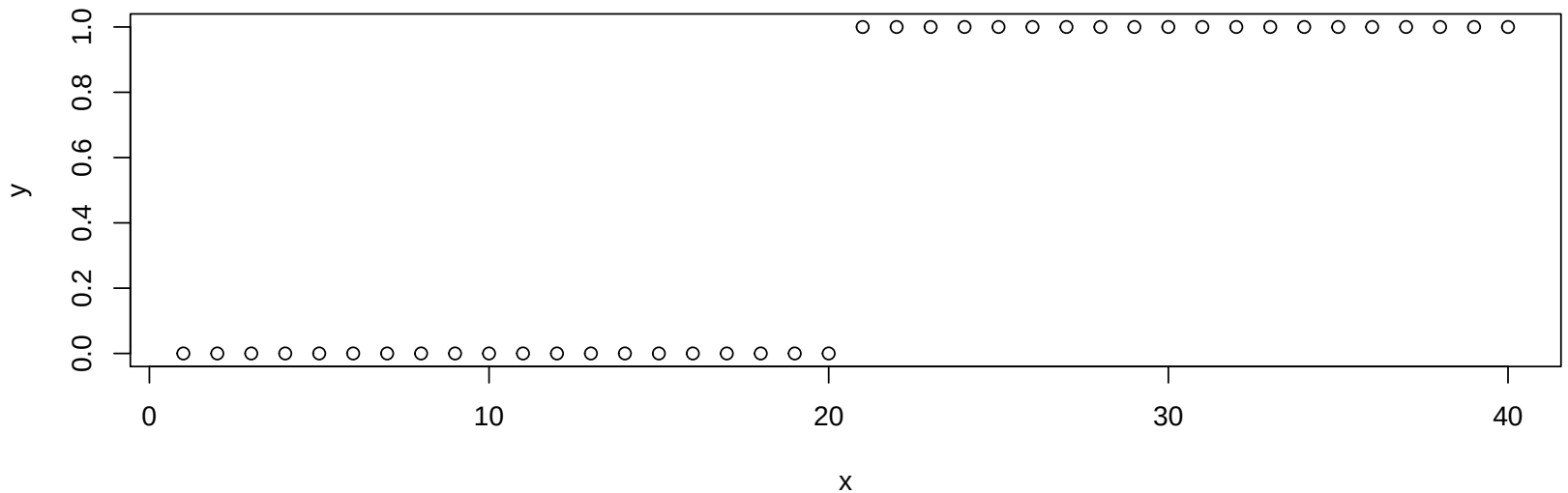
Comparison with Likelihood Ratio Test (LRT)/Analysis of Deviance

There are LRT/Analysis of Deviance (more similar to F-tests) for GLMs, but they assume the distributional assumption (or the variance function in most cases) is correct and do not allow the use of robust SE.
We can use `anova` function for LRT/Analysis of Deviance.

Comparison with Likelihood Ratio Test (LRT)/Analysis of Deviance (continued)

However, LRTs can be very useful when the true β coefficient is very far from 0, i.e. when X and Y have extremely strong association. This can happen when X alone can (almost) determine Y (e.g. diseases caused by genetic mutations). Consider this example:

```
y<-c(rep(0,20),rep(1,20))
x<-1:40
plot(x,y)
```



Comparison with Likelihood Ratio Test (LRT)/Analysis of Deviance (continued)

```
model<-glm(y~x,family=binomial)
```

```
## Warning: glm.fit: algorithm did not converge
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
coef(summary(model))
```

```
##               Estimate Std. Error      z value  Pr(>|z|)
## (Intercept) -776.03478  226643.12 -0.003424038 0.9972680
## x             37.85535   11052.47  0.003425057 0.9972672
```

```
anova(model,glm(y~1,family=binomial),test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ x
## Model 2: y ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        38      0.000
## 2        39     55.452 -1  -55.452 9.578e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Checking Assumptions for GLMs

For checking assumptions about linear regression models the "ordinary" residuals (observed - fitted values) are used. Patterns in the residuals reflect departures from the model assumptions.
However, for a GLM with a non-linear link function (e.g., log or logit) and/or a non-constant variance function, the ordinary residuals will not correctly reflect departures from the model assumptions. For example, with Poisson regression an assumption is that the variance is equal to the mean (hence not constant), so we should expect the residuals to display non-constant variance.

Residuals for GLMs

The usual type of residual for a GLM is the Pearson residual defined as

$$r_P = \frac{Y - \hat{\mu}}{\sqrt{v(\hat{\mu})}} \text{ or } \frac{Y - \hat{\mu}}{\sqrt{\hat{\sigma}^2 v(\hat{\mu})}}$$

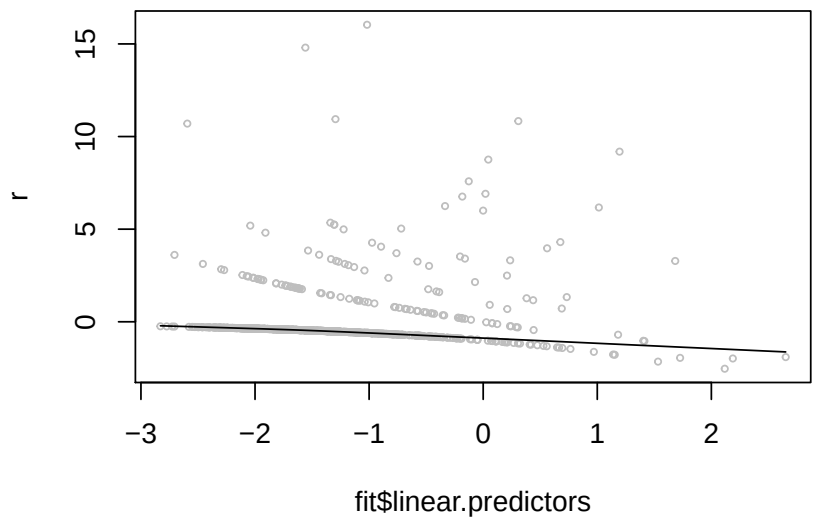
Note that the Pearson residual is equal to the ordinary residual divided by an estimate of its SD. We expect the Pearson residuals to have no pattern in its mean when plotting against covariates or fitted values. Because it is standardized in this way, we expect the Pearson residuals to have variance 1 if the assumed var(Y) is correct.

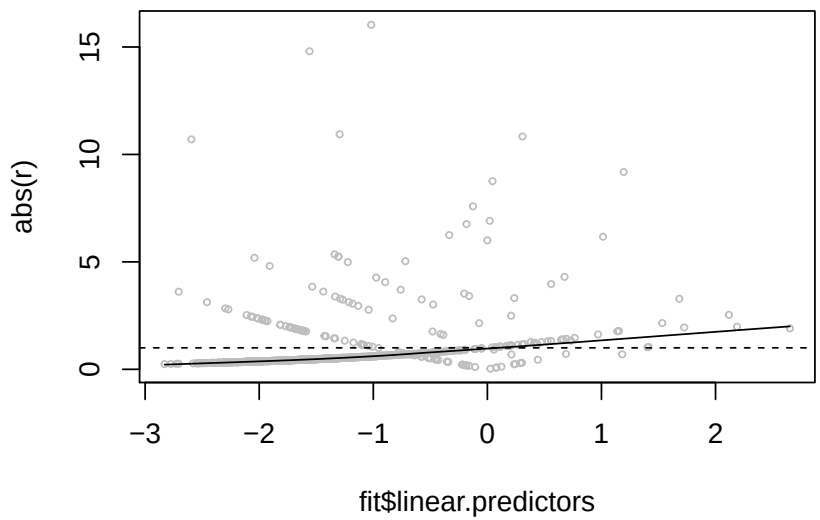
There is another type of residual sometimes used for GLMs, called the *deviance* residual. It is based on the form of the underlying probability distribution. The default for the 'glm' function is the deviance residual. The two types of residuals will often give similar conclusions.

Checking the assumptions of the Poisson regression model for Teeth data

Residual plots can be difficult to interpret, especially when the counts are mostly small.

```
r<-residuals(fit,type="pearson")
par(mfrow=c(1,2),mar=c(5,4,4,1))
scatter.smooth(fit$linear.predictors,r,cex=0.5,col="gray")
scatter.smooth(fit$linear.predictors,abs(r),cex=0.5,col="gray")
abline(h=1,lty=2)
```





Assessment of the assumptions

Residuals versus linear predictors

There is a bit non-linearity in the smooth curve.

Absolute residuals versus linear predictors

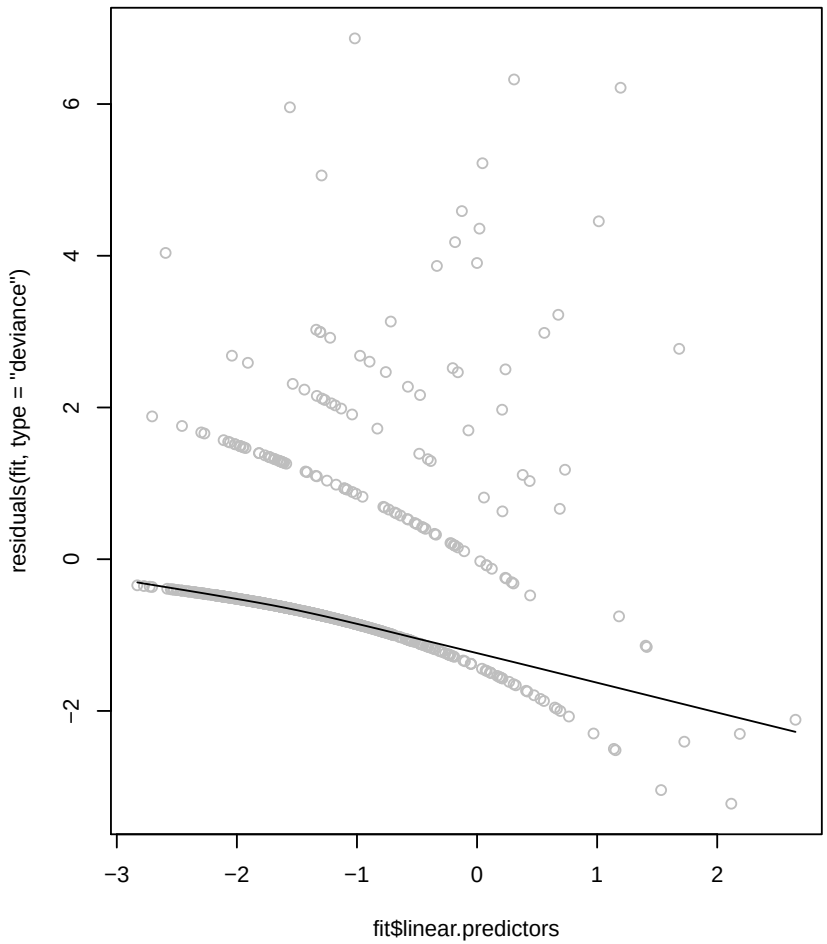
The smooth curve in the plot of absolute residuals suggests that the Poisson mean-variance relationship is not quite right.
Besides, there seems some outliers.

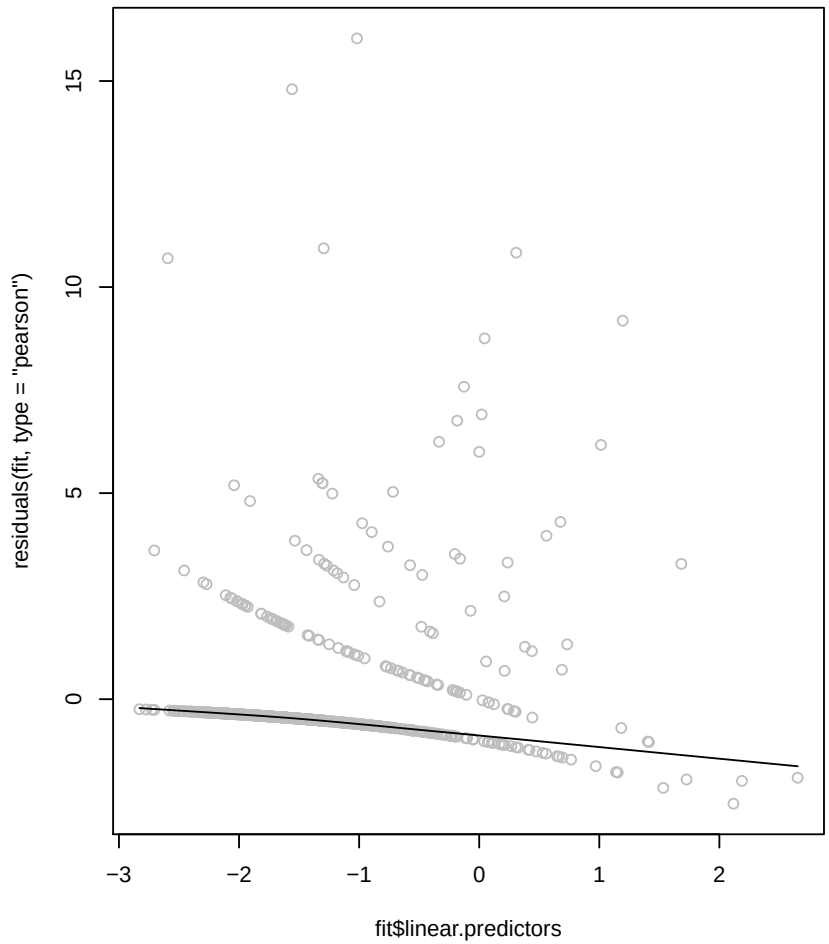
Overall summary

The model assumptions are not satisfied very well.

Pearson and deviance residuals for Teeth data

Model: Y~X





Pearson and deviance residuals for Teeth data (continued)

Model: Y~X+AGE