Count Data

count data is another common type of data in observational and epidemiological studies

this type of data naturally arises from studies investigating the incidence or mortality of diseases in a population

the Poisson distribution is a natural choice to model the distribution of such data

as with the binomial distribution leading to logistic regression, a simple Poisson model is quite limited

We want to allow each sampling unit (person, county, etc.) to have a unique rate parameter, depending on the explanatory variables

The random and systematic components are as follows:

$$\text{Random component: } y_i \sim \text{Pois}(\lambda_i)$$
$$\text{Systematic component: } \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

The Poisson distribution has a log link function

$$\log(\lambda_i) = \eta_i$$
$$\lambda_i = \exp(\eta_i)$$

the meaning of λ often requires additional thought

When we employ a Poisson model, what we are modeling is the rate of events

We need to be careful about specifying what we are estimating: a rate per what?

For example, if we are modeling motor vehicle crashes, we may be estimating a rate per 1,000 population, a rate per 1,000 licensed drivers, a rate per 1,000 registered motor vehicles, or a rate per 100,000 miles traveled

a kind of rate that is particularly common in epidemiological studies is a rate per person-years of follow-up

for example, consider the classic study by Doll et al. in which all British male doctors were sent a questionnaire about their age and whether they smoked tobacco

the doctors were then followed up for a number of years to see whether or not they had died from coronary heart disease

suppose, then, that we wish to model $\lambda(\mathbf{x})$ the rate per 1,000 person-years of follow-up, given the explanatory variables Age and Smoking

now,

$$\mathrm{E}(Y_i) = t_i \lambda_i,$$

where t_i denotes the person-years of follow-up for observation i

this implies that

$$\begin{aligned} \log(\mu_i) &= \log(t_i) + \log(\lambda_i) \\ &= \log(t_i) + \eta_i; \end{aligned}$$

thus the usual relationship between μ_i and the linear predictor is offset by the amount $\log(t_i)$

the estimated rates from our Poisson regression model

	Smokers	Non-smokers
35–44	0.52	0.36
45–54	2.29	1.60
55–64	7.17	5.03
65–74	14.78	10.37
75–84	20.97	14.71

note that, by fitting a model with no interaction between age and smoking, we enforce that the rate ratio (RR) between smokers and non-smokers are the same in each age group ($0.52/0.36 = \cdots = 20.97/14.71 = 1.43$)

if we allow for interaction we obtain

	Smokers	Non-smokers	RR
35–44	0.61	0.11	5.5
45–54	2.40	1.12	2.1
55–64	7.20	4.90	1.5
65–74	14.69	10.83	1.4
75–84	19.18	21.20	0.9

Poisson regression is an adequate tool for analyzing cohort studies; however, if one has detailed individual-level data, one can apply the more sophisticated approaches that have been developed in the field of survival analysis

Overdispersion

one of the defining characteristics of Poisson regression is its lack of a scale parameter: E(Y) = Var(Y), and no parameter is available to adjust that relationship

In practice, when working with Poisson regression, it is often the case that the variability of y_i about $\hat{\lambda}_i$ is larger than what $\hat{\lambda}_i$ predicts

this implies that there is more variability around the model's fitted values than is consistent with the Poisson distribution

the term for this phenomenon is overdispersion

data for which this phenomenon manifests itself are often called “overdispersed”

there are two common approaches to correcting for overdispersion:

Quasi-likelihood
Negative binomial regression

the score arising from a Poisson regression model is

$$\frac{\partial \ell}{\partial \theta} = \sum_i \{y_i - \hat{\lambda}_i\}$$

where $\theta = \log(\lambda)$ is the parameter

there is no scale parameter, which would show up in the denominator on the right hand side

now suppose we add one

$$\frac{\partial \ell}{\partial \theta} = \sum_i \frac{y_i - \hat{\lambda}_i}{\phi}$$

then $\mathrm{Var}(Y) = \phi V(\mu)$ so we now have a parameter that allows the variance to be larger or smaller than the mean by a multiplicative factor ϕ

This will not change $\widehat{\boldsymbol{\beta}}$,

however, it will affect inference, since

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}\left(\boldsymbol{\beta}, \phi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\right)$$

so what distribution is this, that gives rise to this score?

there isn’t one (at least, not one for which you can write down the distribution in closed form)

this approach, where you modify the score directly and never actually specify a distribution, is known as quasi-likelihood

typically, the scale parameter ϕ is estimated using the method of moments estimator

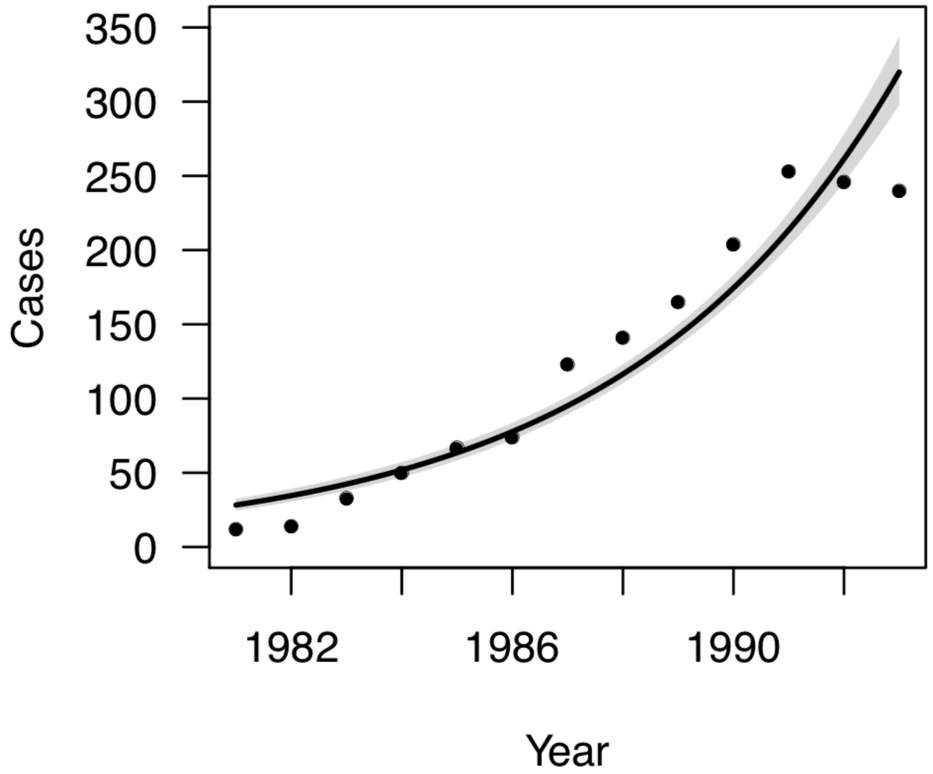
$$\hat{\phi} = \frac{X^2}{n-p}$$

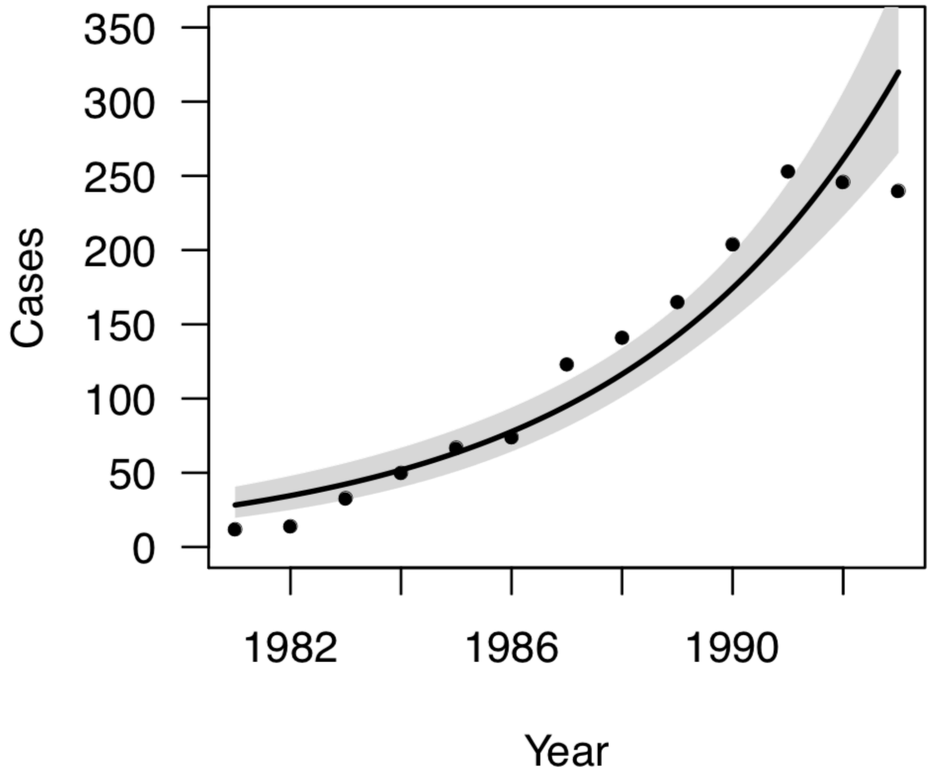
to use this approach in R, one can specify family=quasipoisson

For our Belgian AIDS data, $\hat{\phi}$ = 6.7, implying that the variance was nearly 7 times larger than that implied by the Poisson distribution

Again, the fit is the same

However, our standard errors are $\sqrt{6.7} \approx 2.6$ times larger


Cases
350
300
250
200
150
100
50
0
1982
1986
1990
Year


Cases
350
300
250
200
150
100
50
0
1982
1986
1990
Year

the quasi-Poisson approach is attractive for several reasons, but its big drawback is that lacks a log-likelihood

this prevents you from using any of the likelihood-based tools we have discussed for GLMs: likelihood ratio tests, AIC/BIC, deviance explained, deviance residuals

an alternative approach that allows all those maximum likelihood tools is based on the negative binomial distribution

the negative binomial distribution has other uses in probability and statistics, but for our purposes we can think about it as arising from a two-stage hierarchical process

$$\mathrm{E}(Y) = \lambda$$
$$\mathrm{Var}(Y) = \lambda + \lambda^2/\theta$$

the marginal distribution of Y is then negative binomial, with

$$Z \sim \mathrm{Gamma}(\theta, \theta)$$
$$Y|Z \sim \mathrm{Poisson}(\lambda Z)$$

thus, like the Poisson distribution, the negative binomial has support only on the positive integers, but unlike the Poisson, its variance is larger than its mean

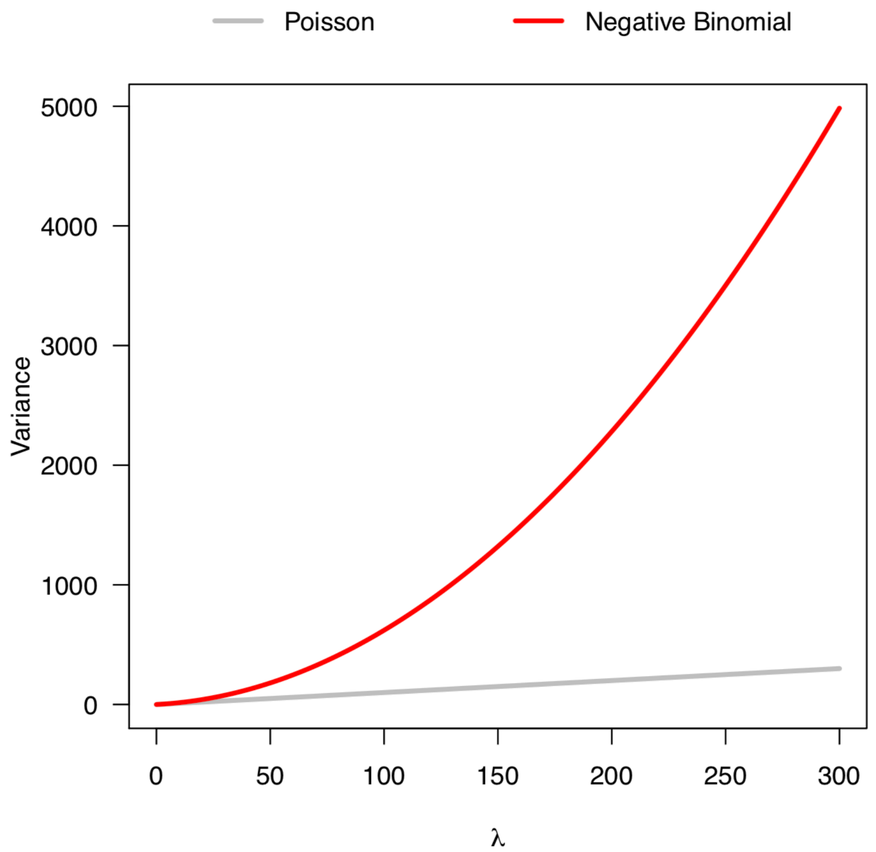
note, however, that the negative binomial distribution is not a member of the exponential family

thus, the theory and fitting procedures we have developed for GLMs do not directly apply here

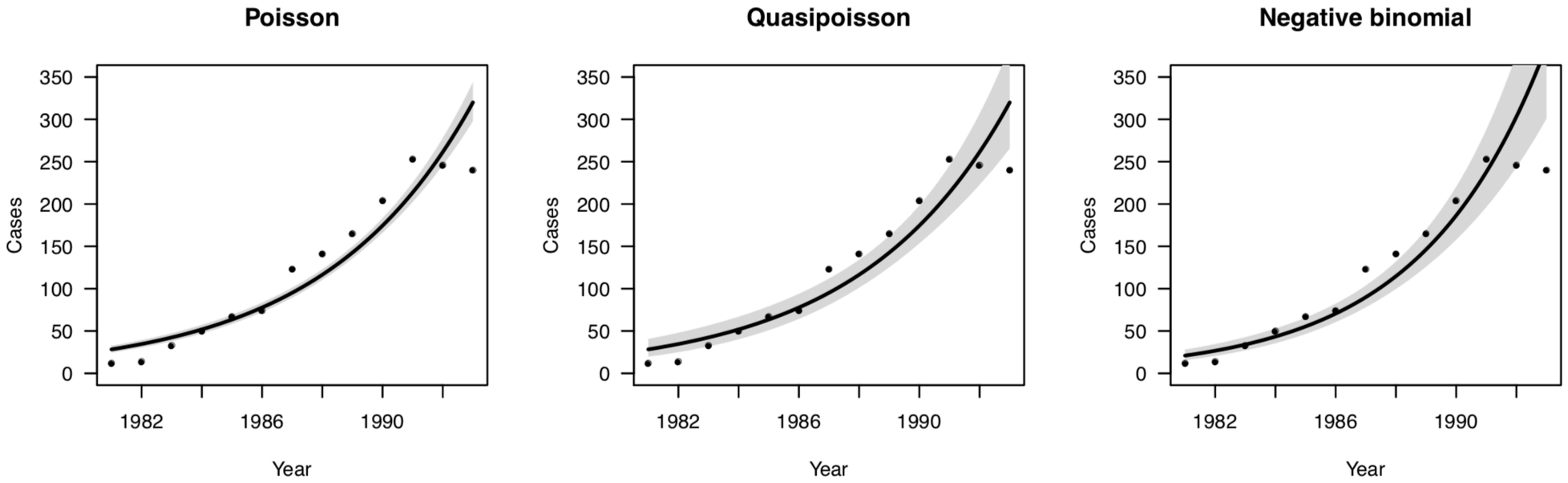
however, it is customary to employ a log link to make negative binomial regression look like Poisson regression

in R, one must use the glm.nb function in the MASS package

for the Belgian AIDS data, $\hat{\theta} = 19.2$, implying the following mean-variance relationship:



this leads to the following



arguably, the negative binomial estimates are even worse than the Poisson estimates, and certainly drastically worse than the quadratic Poisson model

however, its “goodness of fit” measures are much better

this is why I remarked earlier that it’s wrong to think of the data as overdispersed – if the data show more variability than the model can explain, the most likely explanation is a bad model

the quadratic Poisson fit shows no overdispersion (the residuals are actually slightly “underdispersed”)

accounting for overdispersion is a good idea – if the model doesn't fit the data, this should be reflected with larger standard errors and wider confidence intervals

many analysts have the view that quasi-Poisson or negative binomial regression automatically "fixes" the overdispersion problem

this is a potentially dangerous misconception – surely, accurately modeling the mean is of greater priority than modeling the variance

while quasi-Poisson and negative binomial approaches are useful, they are certainly no substitute for careful consideration of the systematic component of the model

lets take a look at another example where the negative binomial is a good fit for the data

hurdle model

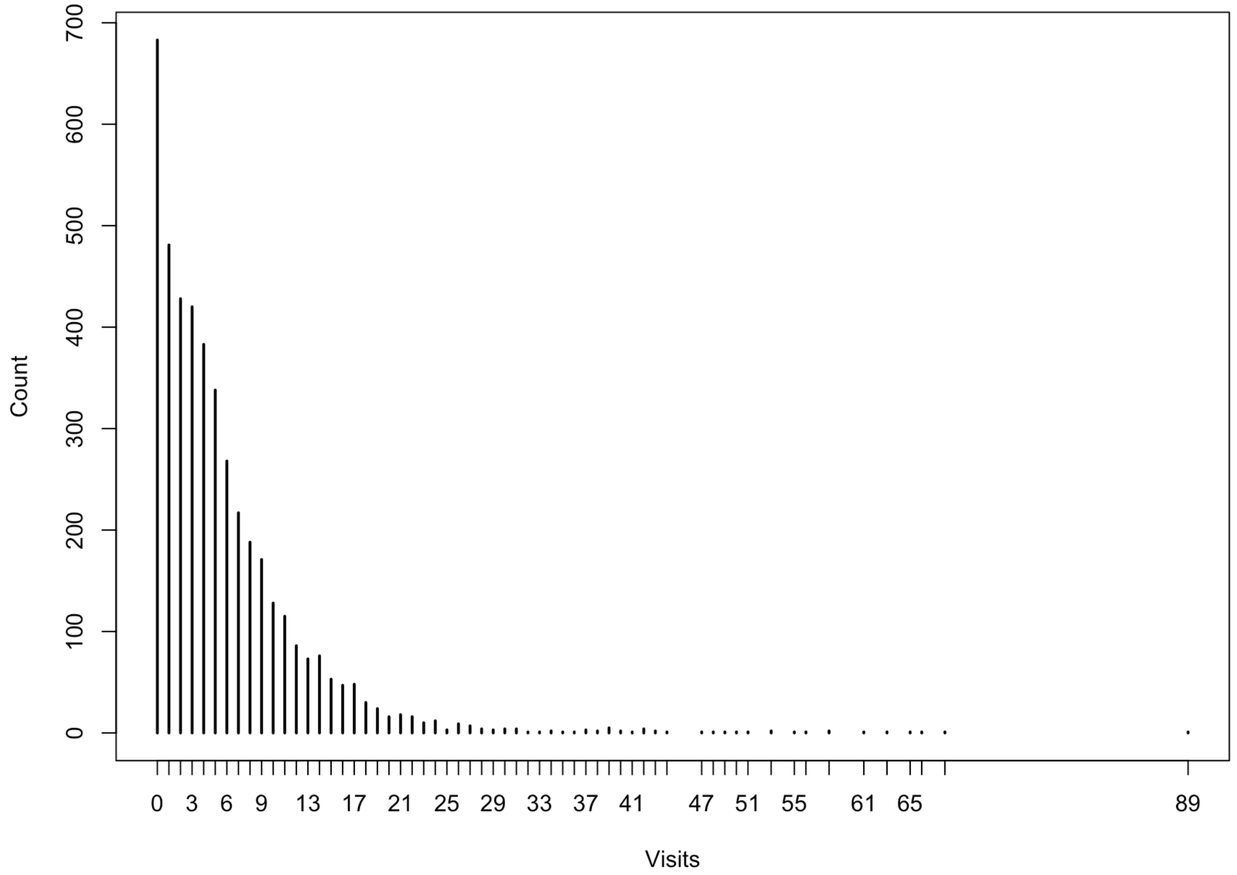
a hurdle model is a model to help handle excess zeros and also can consider overdispersion

let’s take a look at a number of predictors that affect the total number of physician office visits

Individuals, aged 66 and over, covered by Medicare in 1988 with the number of physician office visits

We have a sample of 4,406 individuals from which we will consider the following variables

hospital, health, chronic, gender, school, insurance



there are a significant number of individuals who had 0 physician visits

as this is count data we can first consider a poisson regression, and then we will follow by also fitting a negative binomial regression

as initially noted we had a large number of zeroes so we can compare the number zeroes predicted by each model to the observed count

do that by first predicting the expected mean count for each observation, and then using those expected mean counts to predict the probability of a zero count.

then we can sum those expected probabilities to get an estimate of how many zero counts we might expect to see.

We see that the poisson severely underfit zero counts.

the negative binomial get much closer to the correct number of zeros, but as we will see overfits at other values near zero (1,2,3) to account for this

This is where the Hurdle model comes in. This is a two-part model that specifies one process for zero counts and another process for positive counts

The idea being once a ‘hurdle’ is cleared (a threshold crossed) then the positive counts occur. If the hurdle is not cleared then the have a count of 0

The combination of process is models that we have seen previously

- First part, binary logit
- Second part, *truncated* poisson or negative binomial model

in the case of the physician visits data the first part governs whether a patient visits a doctor or not and then the second part governs the counts (how many) visits are made

the pscl package provides us with a hurdle function for fitting these models

first we use the predict function with `type = "prob"`. this returns a predicted probability for all possible observed counts for each observation.

in this case, that returns a 4406 x 90 matrix. That's 4406 rows for each observation, and 90 possible counts.

the first column contains the predicted probabilities for getting a 0 count. As before we can sum those probabilities to get an expected number of 0 counts.

we get output for two different models.

- The first section of output is for the positive-count process.
- The second section is for the zero-count process. we can interpret these just as we would for any other model.

Having fit a hurdle model, how many 0 counts does it predict? This is a little trickier to extract.

we can also predict the expected mean count using both components of the hurdle model. The mathematical expression for this is

$$\mathsf{E}[y|x] = \frac{1 - f_1(0|x)}{1 - f_2(0|x)}\mu_2(x)$$

This says the expected count given our predictors is a product of two things: a ratio and a mean.

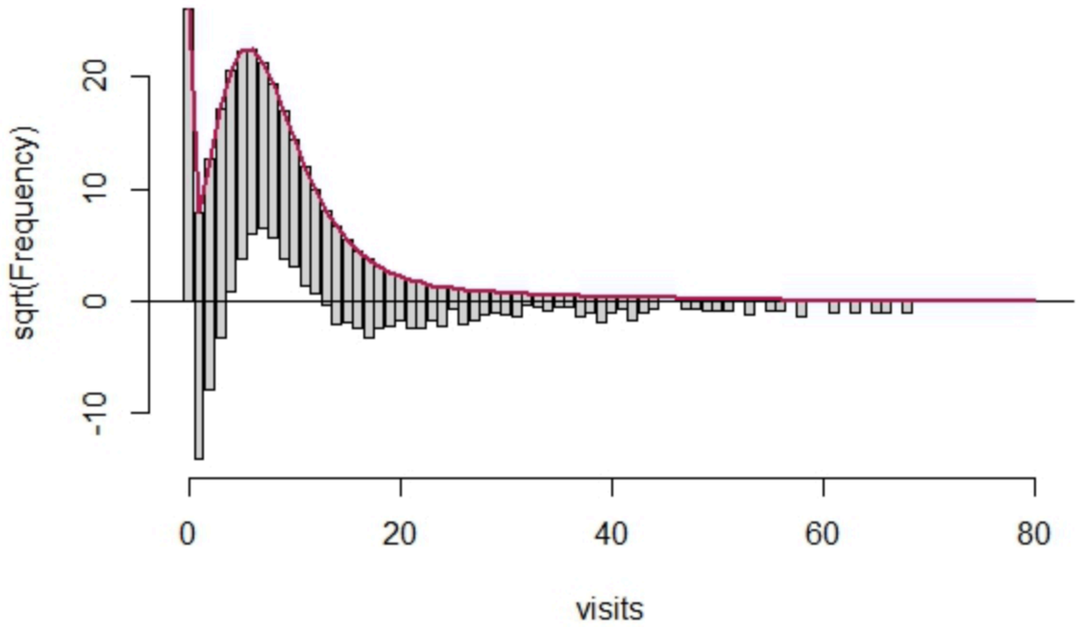
The ratio is the probability of a non-zero in the first process divided the probability of a non-zero in the second *untruncated* process. Recall these are logistic and Poisson, respectively, by default.

The mean is of the *untruncated* version of the positive-count process.

We can use the `predict` function to get these expected mean counts by setting `type = "response"`, which is the default.

It appears we have addressed the excess 0's, but what about the overdispersion?

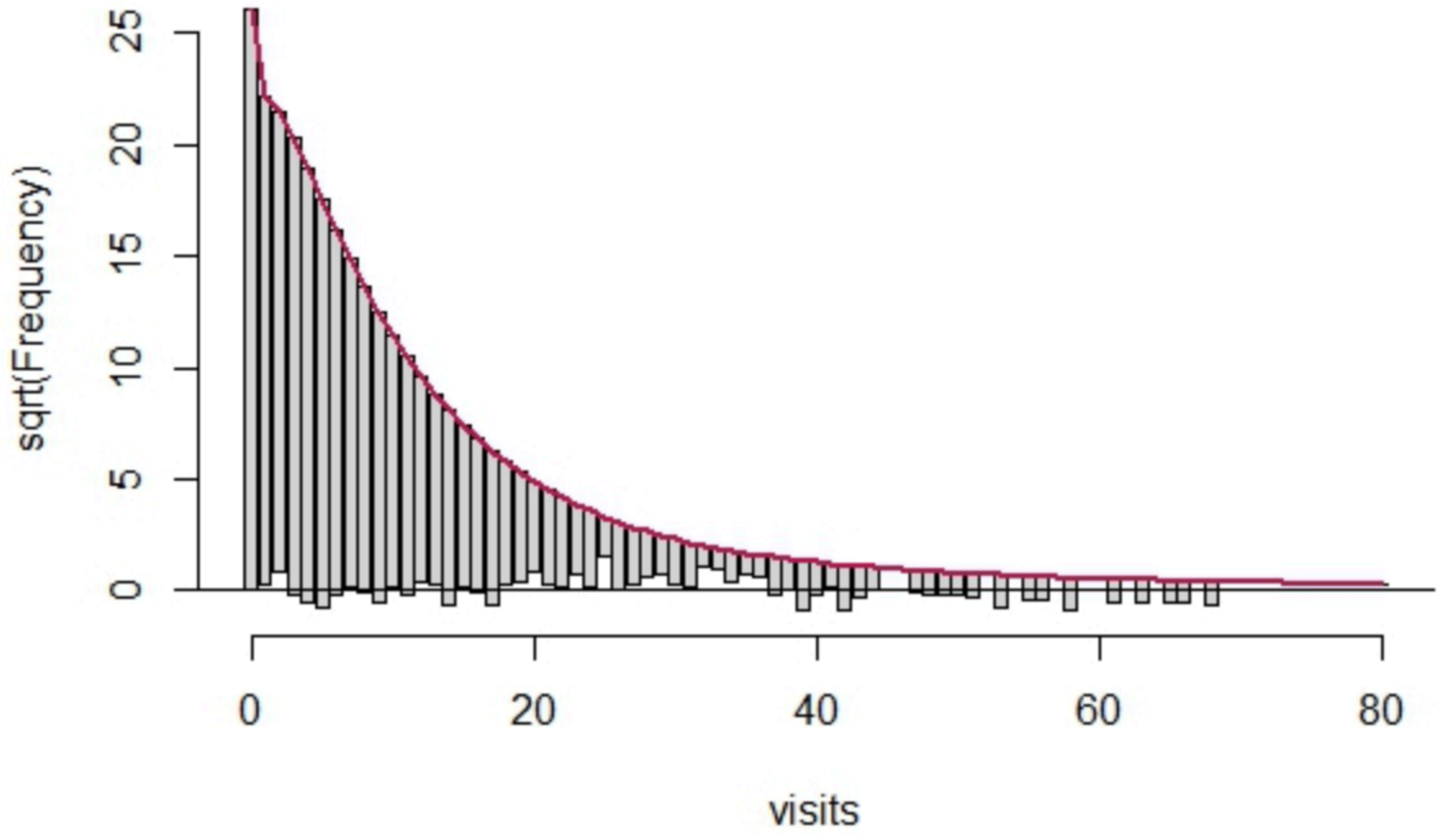
We can visualize the fit of this model using a rootogram



The line at 0 allows us to easily visualize where the model is over- or under-fitting.

At 0 it fits perfectly by design. But at counts 1, 2 and 3 we see dramatic under-fitting (under the line) and then pronounced over-fitting at counts 5 – 9 (over the line). We also see a great deal of under-fitting at the higher counts as well.

This points to overdispersion



traditional model-comparison criteria such as AIC show the negative binomial version is better fitting as well.

recall that each component of a hurdle model can have different sets of predictors.

we can do this in the `hurdle` function by using “|” in the model formula. For example, let’s say we want to fit the zero hurdle component using only the insurance and gender predictors

```
fit.hurdle.nb2 <- hurdle(visits ~ . | gender + insurance, data
= nmes, dist = “negbin")
```

this says fit the count data model (visits regressed on all other variables) conditional on the zero hurdle model (visits regressed on gender and insurance).

Multinomial Regression

we have used logistic regression to model binary (yes/no) data

what if we have multiple categories? For example, different forms of a disease, different types of species, or choices from among several alternatives?

now we will discuss the generalization of logistic regression (which involved a binomial outcome) to multinomial regression, in which the outcome is multinomial

to illustrate multinomial regression, we'll analyze a study of factors influencing the primary food choice of alligators

the study involved 219 alligators captured in four Florida lake

the outcome variable, Food, is the primary food type, and consists of five categories:

bird, fish, invertebrates, reptiles, other

in addition to the lake in which the alligator was captured, we also have information pertaining to the alligator's

- Size: Either small ($\leq$ 2.3 meters long) or large ($>$ 2.3 meters long)
- Sex

the question of interest is the effect that these factors have on the primary food type that an alligator chooses to eat

we will use the following notation in this lecture and the next to describe multi-class models:

Let Y be a random variable that can take on one of K discrete values (i.e., fall into one of K classes)

Number the classes 1,…,K

Let $\pi_{i2} = \Pr(Y_i = 2)$ denotes the probability that the ith individual's outcome belongs to the second class

More generally, $\pi_{ik} = \Pr(Y_i = k)$ denotes the probability that the ith individual's outcome belongs to the kth class

in case you have not seen it before, the multinomial distribution is defined as follows

$$p(Y = \mathbf{y}) = \frac{n!}{y_1! \cdots y_K!} \pi_1^{y_1} \cdots \pi_K^{y_K},$$

where $\sum_k y_k = n$ and $\sum_k \pi_k = 1$

note that for K = 2, this reduces to the binomial distribution

if the data were iid, we could simply fit the multinomial distribution to our data

however, the purpose of our analysis is to examine the ways in which factors (which vary from alligator to alligator) change π; hence the name multinomial regression

multinomial logistic regression is essentially equivalent to the following:

- Let k = 1 denote the reference category
- Fit separate logistic regression models for k = 2, . . . , K, comparing each outcome to the baseline:

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \mathbf{x}_i^T \boldsymbol{\beta}_k$$

note that this will result in K – 1 vectors of regression coefficients (we don't need to estimate the Kth vector because $\sum_k \pi_k = 1$

this is the multinomial regression model, although the estimation procedure is complicated by the constraint that $\sum_k \pi_k = 1$

the fitted class probabilities for an observation with explanatory variable vector x are therefore

$$\hat{\pi}_1 = \frac{1}{1 + \sum_k \exp(\mathbf{x}^T \widehat{\boldsymbol{\beta}}_k)}$$

$$\hat{\pi}_k = \frac{\exp(\mathbf{x}^T \widehat{\boldsymbol{\beta}}_k)}{1 + \sum_l \exp(\mathbf{x}^T \widehat{\boldsymbol{\beta}}_l)}$$

like logistic regression, odds ratios in the multinomial model are easily estimated as exponential functions of the regression coefficients

$$\begin{aligned}\mathrm{OR}_{kl} = \frac{\pi_k}{\pi_l} &= \frac{\pi_k/\pi_1}{\pi_l/\pi_1} \\ &= \frac{\exp\left((\mathbf{x}_2 - \mathbf{x}_1)^T \boldsymbol{\beta}_k\right)}{\exp\left((\mathbf{x}_2 - \mathbf{x}_1)^T \boldsymbol{\beta}_l\right)} \\ &= \exp\left((\mathbf{x}_2 - \mathbf{x}_1)^T (\boldsymbol{\beta}_k - \boldsymbol{\beta}_l)\right)\end{aligned}$$

In the simple case of changing x_j by δ_j and comparing k to the reference category,

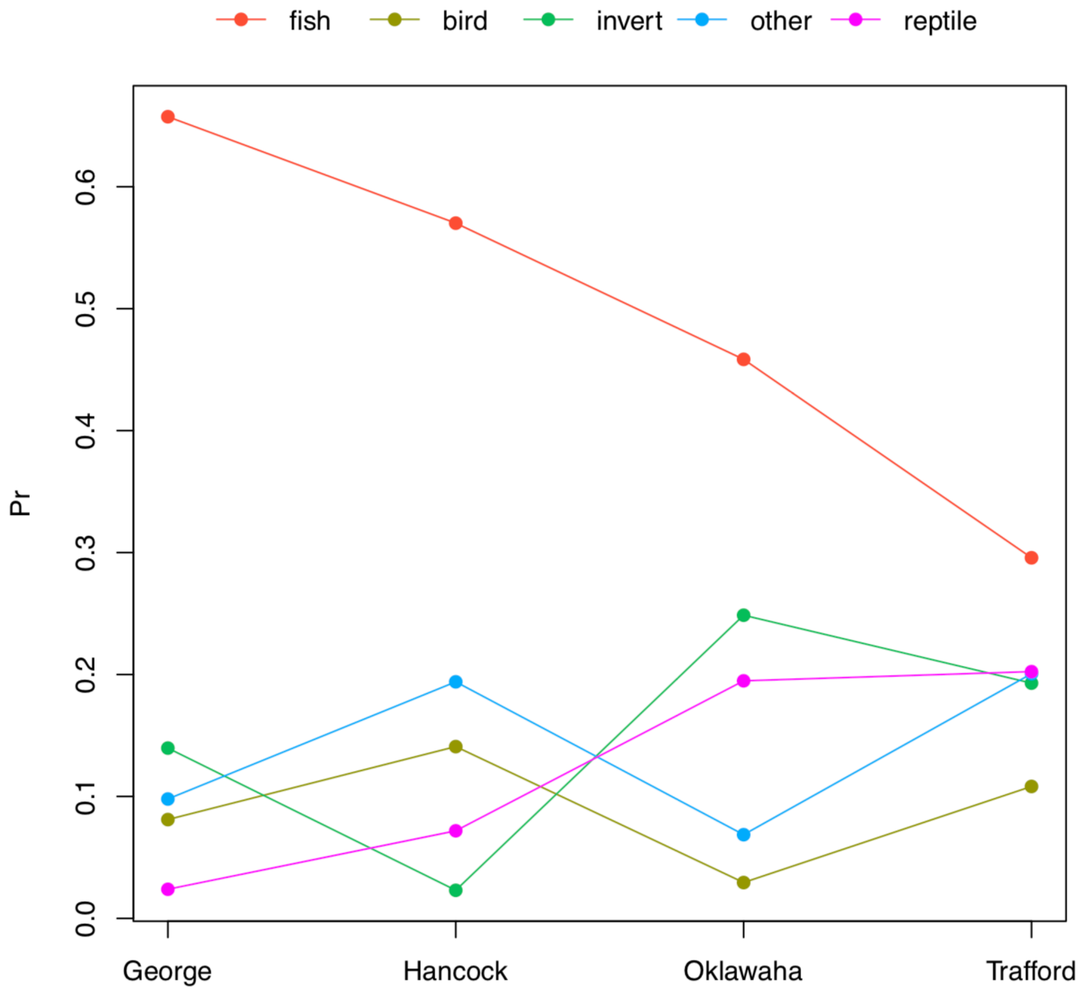
$$\mathrm{OR}_{kl} = \exp(\delta_j \beta_{kj})$$

some model selection

Model	AIC
Null	612
Size	605
Size + Lake	580
Size + Lake + Sex	586
Size × Lake	587

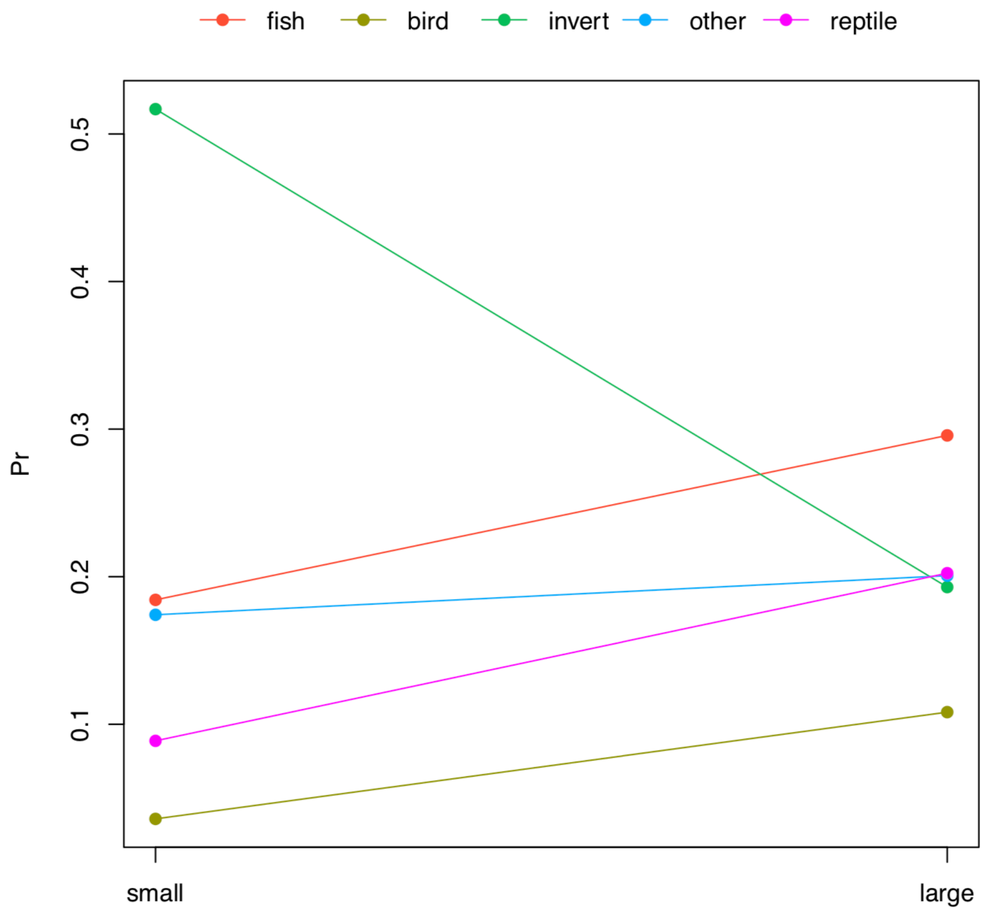
it would seem, therefore, preferences, although there is no evidence of an interaction between the two, or any meaningful differences in the eating preferences of male and female gators

lake


fish
bird
invert
other
reptile
Pr
0.0
0.1
0.2
0.3
0.4
0.5
0.6
George
Hancock
Oklawaha
Trafford

$p < 0.0001$

size



$p = 0.0003$

odds ratios comparing large vs. small alligators, with invertebrates as reference group

		95% CI		
	OR	Lower	Upper	p
Bird	8.1	2.1	31.5	0.003
Reptile	6.1	1.9	19.9	0.003
Fish	4.3	2.0	9.3	0.0002
Other	3.1	1.1	8.3	0.03

odds ratios comparing Lake Trafford vs. Lake George, with fish as reference group

		95% CI		
	OR	Lower	Upper	p
Reptile	18.8	2.1	167.9	0.009
Other	4.6	1.3	15.4	0.01
Invert	3.1	1.2	8.0	0.02
Bird	3.0	0.6	15.4	0.2

Proportional Odds Models

multinomial regression requires the estimation of $(K - 1)p$ parameters, and assumes nothing about the relationship between the categories
this is very flexible of course, but has two downsides:

- the large number of parameters can be cumbersome to interpret
- estimating a large number of parameters can result in high variability in the estimates

when the categories are ordered, making assumptions about the relationships between them allows us to introduce some structure and estimate fewer parameters, decreasing variability and increasing interpretability

in October 2000, a coal slurry impoundment ruptured, emptying more than 300 million gallons of toxic coal waste into the streams of Martin County, Kentucky

researchers from the sociology department at the University of Kentucky carried out a study of the disaster’s effect on trust in the community

in the months following the disaster, and then again 10 years later, a survey was administered to residents of Martin County

among the survey items was the statement, “I have trust in the local government”; respondents were asked to choose their reaction to that statement from among the following options:

Strongly disagree | Disagree | Neutral | Agree | Strongly agree

such items are often referred to as being measured on a Likert scale after their inventor, Rensis Likert

in addition to trust in local government and year (2011 vs. 2001), we will also consider the demographic variable Education, recorded as Less than high school/High school/Some college/College degree

unlike the case with alligator food choices, the response categories here are ordered, which suggests a certain relationship between them

it would be odd, for example, if comparing 2011 vs. 2001 we came to the conclusion that the number of “strongly agree” and “disagree” responses went up significantly, but that the number of “agree” and “strongly disagree” responses went down significantly

to address this ordering, we can focus on the cumulative logits:

$$\log\left(\frac{\Pr(Y \leq k)}{\Pr(Y > k)}\right) = \log\left(\frac{\pi_1 + \cdots + \pi_k}{\pi_{k+1} + \cdots + \pi_K}\right)$$

the proportional odds model assumes that each explanatory variable exerts the same effect on each cumulative logit regardless of the cutoff k

$$\log\left(\frac{\Pr(Y \leq k)}{\Pr(Y > k)}\right) = \alpha_k + \mathbf{x}^T\boldsymbol{\beta}$$

each cumulative logit has its own intercept, but each explanatory variable only has a single coefficient β; thus, the model has fewer terms than a multinomial regression model

writing down the proportional odds model requires us to modify previous notation. in the above, x and β do not include a term for the intercept

the description of the model given on the previous slide is perhaps a bit counterintuitive, in that high values of $\eta = \alpha_k + \mathbf{x}^T\boldsymbol{\beta}$ are associated with low values of Y

as a results many people like to formulate the model as

$$\log\left(\frac{\Pr(Y \leq k)}{\Pr(Y > k)}\right) = \alpha_k - \mathbf{x}^T\boldsymbol{\beta},$$

so that the sign of β has the usual meaning (i.e., if positive, an increase in x is associated with an increase in Y)

this is how R already formulates this model

suppose we wish to calculate Pr(Y = k|x) based on our model For the simple case where k = 1, the calculations are identical to logistic regression:

$$\Pr(Y = 1|\mathbf{x}) = \Pr(Y \leq 1|\mathbf{x}) = \frac{\exp(\alpha_1 + \mathbf{x}^T\boldsymbol{\beta})}{1 + \exp(\alpha_1 + \mathbf{x}^T\boldsymbol{\beta})}$$

Calculating the probabilities for other categories requires a bit more work:

$$\Pr(Y = k|\mathbf{x}) = \Pr(Y \leq k|\mathbf{x}) - \Pr(Y \leq k - 1|\mathbf{x})$$

as with regular logistic regression, a key advantage of the logit link is that additive models yield constant odds ratios

consider comparing two arbitrary individuals, with covariates x_2 and x_1

$$\frac{\Pr(Y > k|\mathbf{x}_2)/\Pr(Y \leq k|\mathbf{x}_2)}{\Pr(Y > k|\mathbf{x}_1)/\Pr(Y \leq k|\mathbf{x}_1)} = (\mathbf{x}_2 - \mathbf{x}_1)^T\boldsymbol{\beta}$$

thus, for the usual case where we consider changing only a single parameter at a time by one unit, e^{β} represents the odds ratio, as usual in logistic regression

now the difference, however, is that e^{β} now represents a cumulative odds ratio: the odds of “at least k” under two different conditions

we get exactly the same odds ratio for comparing {Neutral, Agree, Strongly agree} vs {Disagree, Strongly disagree} as when comparing {Agree, Strongly agree} vs {Neutral, Disagree, Strongly disagree}

the odds ratio is constant across each split; hence the name “proportional odds model”

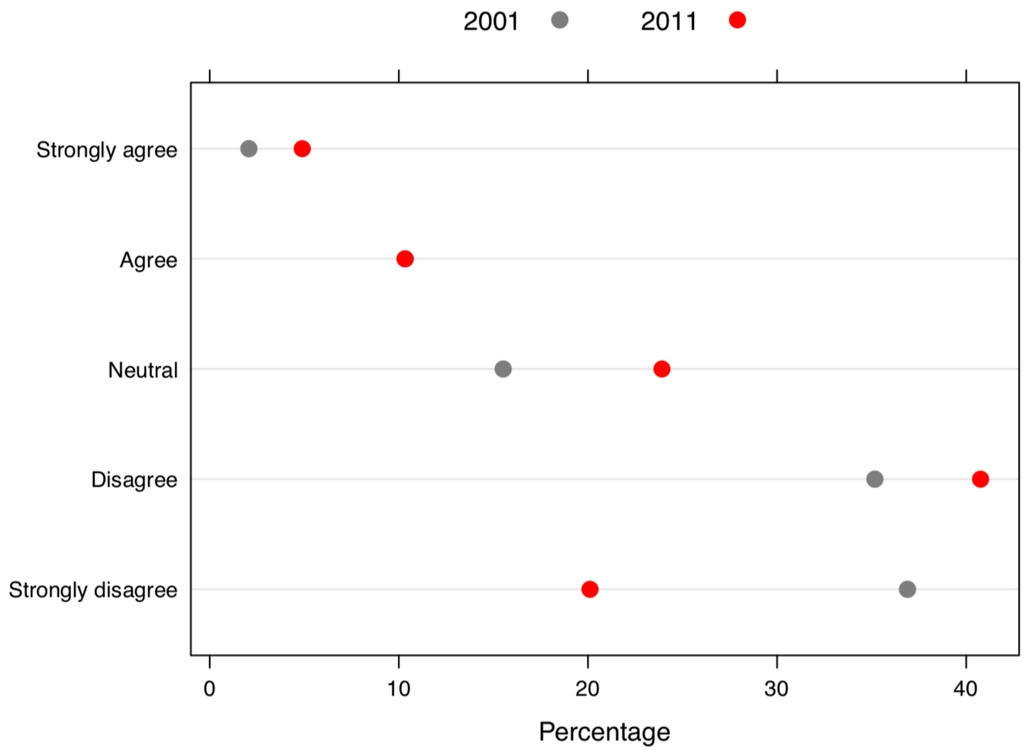
for the model with Year as the only explanatory variable, we have $\hat{\beta} = 0.618$ and

	95% CI			
	OR	Lower	Upper	p
Year	1.9	1.3	2.6	0.0003

one valid interpretation of this finding would be that the odds of agreeing with the statement “I trust the local government” nearly doubled between 2001 and 2011

or, equally valid, the odds of disagreeing with the statement fell by nearly half

this is something of an oversimplification, in that the actual increases were not perfectly proportional across all categories



estimated probability of each response, subject to model restrictions

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
2001	0.35	0.38	0.16	0.08	0.02
2011	0.23	0.37	0.22	0.14	0.04

observed (“raw”) proportions

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
2001	0.37	0.35	0.16	0.10	0.02
2011	0.20	0.41	0.24	0.10	0.05

multinomial regression avoids the proportional odds assumption, allowing the possibility of capturing non-proportional trends in the explanatory variables

the downsides, however, are that we don’t simply get an OR for Year, we have separate ORs for Agree vs. Disagree, Strongly agree vs. neutral, neutral vs. disagree, etc.

this is a bit cumbersome to interpret; for example, the estimated agree vs. disagree odds ratio for the multinomial model is 0.86, suggesting the opposite trend (a decline in trust over time) from most other comparisons

furthermore, the multinomial model estimates have higher variance

for example, $SE_{\hat{\beta}} = 0.17$ for the proportional odds model, $SE_{\hat{\beta}} = 0.33$ for the multinomial model comparing agree to disagree

to compare the models, we could use AIC; here, the proportional odds model has AIC 1312, while the multinomial model has AIC 1313, indicating that the violation of proportional odds is not substantial enough to warrant all the extra parameters that the multinomial model introduces

comparing the fit of various models:

	AIC
(Intercept)	1324
`Year`	1312
`Year` + `Educ`	1314
`Year` × `Educ`	1309

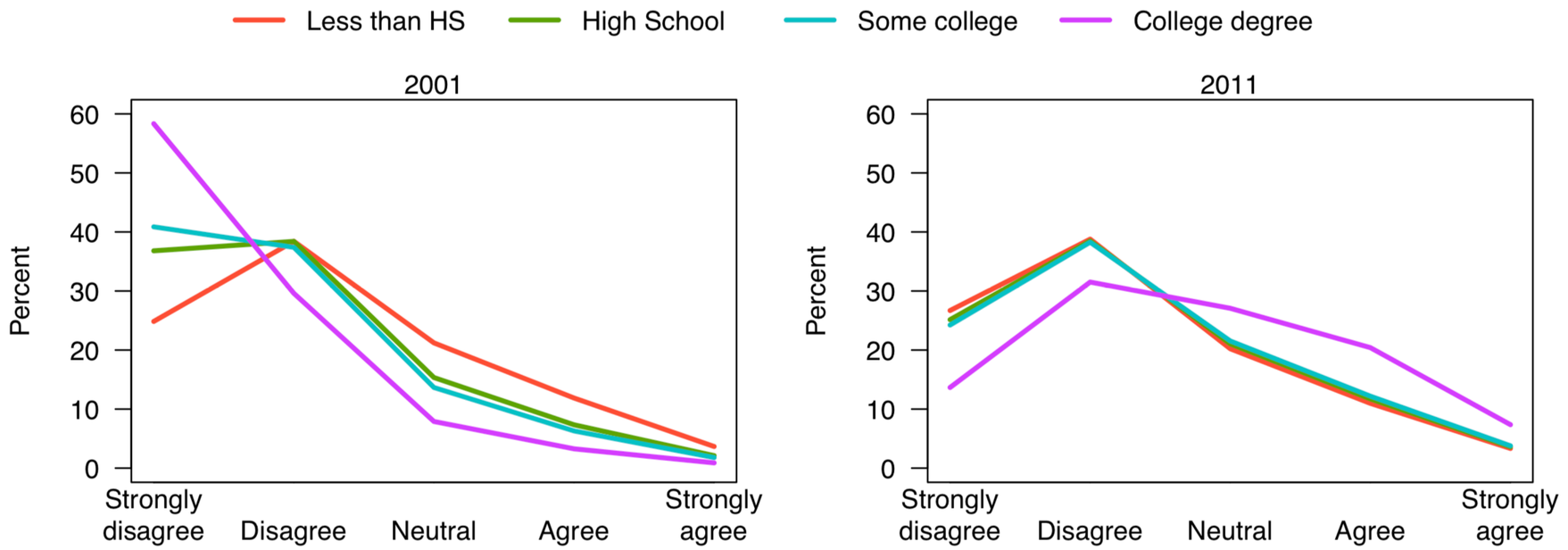
interestingly, education seems to add very little as an additive effect, but has a somewhat important interaction with Year

It is worth noting that assuming a linear trend for education has an even better fit (AIC 1303), but for the sake of illustration, I'll continue treating education as a categorical variable

		95% CI	
Education	OR	2.5 %	97.5 %
(No interaction)	1.9	1.3	2.6
Less than HS	0.9	0.4	1.9
High school	1.7	1.0	3.0
Some college	2.2	1.1	4.3
College degree	8.9	2.8	30.1

including an interaction reveals that the change in trust over time was not the same in each demographic group – trust changed very little among those with less than a high school education and dramatically for those with a college degree

probability estimates



ordinal responses are very common in the medical, epidemiological, and social sciences; I have been asked to analyze ordinal data on many occasions

the proportional odds model is a rather elegant (and popular) way to handle ordinal data, respecting both its ordering as well as its categorical nature without any substantial increase in the difficulty of interpretation, as individual coefficients have odds ratio interpretations very similar to logistic regression

finally, there is also a rather interesting connection between the proportional odds model and nonparametric testing: with a single binary covariate, the (score) test of a proportional odds model is equivalent to the Wilcoxon rank sum test