Logistic Regression

standard (often first) way to model binary outcomes

most common way to make inferences on coefficients for binary outcome to understand something about the world

Our first GLM will be the logit.

The pieces:

1. A density for $Y_i|X_i$. Here, $Y_i|X_i \sim Bern(\pi_i)$
2. Structural linear component for conditional mean, $X_i^\top \beta$
   N.B. $\mathbb{E}[Y_i|X_i] = Pr(Y_i = 1|X_i) = \pi_i$

3a. A link function connecting π_i (our $\mathbb{E}[Y_i]$) to $X_i^\top \beta$. Here:

$$g(\pi_i) = logit(\pi_i) = log\frac{\pi_i}{1-\pi_i} = X_i^\top \beta$$

3b. Equivalently, inverse link *from* the linear component,

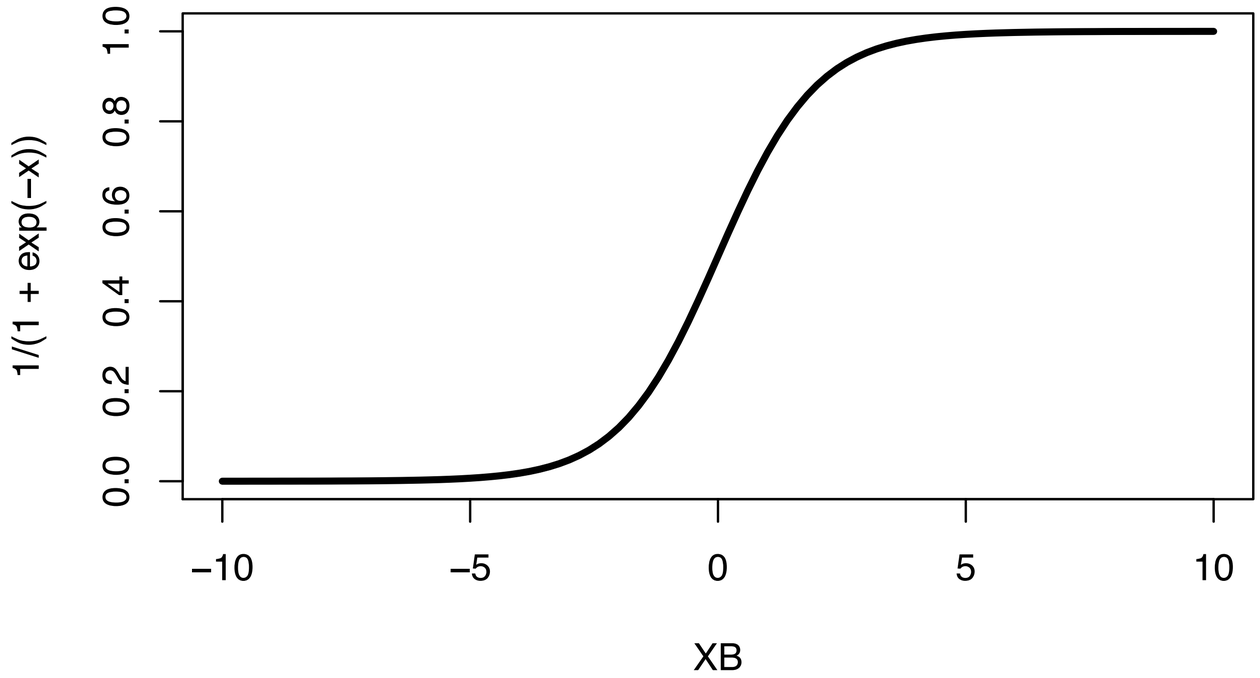
$$\pi_i = logit^{-1}(X_i^\top \beta) = \frac{1}{1 + e^{-X_i^\top \beta}}$$

All together,

$$Y_i|X_i \sim Bern(logit^{-1}(X_i^\top \beta))$$

$X_i^\top \beta$ is in $[-\infty, \infty]$, but you need a result in $[0, 1]$.

So you need a squashing function:



Why this link? No great reason, and we'll see others, but...

- odds: $\frac{\pi}{1-\pi} \in [0, \infty]$, so can't model that as $X_i^\top \beta$
- log-odds: $log(\frac{\pi}{1-\pi}) \in [-\infty, \infty]$, so you're good

Outcome: binary

Distribution. $Y_i|X_i \sim \textit{Bernoulli}(Y_i|\pi_i)$

Inverse Link. Some options:

- Logit: $\pi_i = \frac{1}{1+exp(-X_i^\top \beta)}$
- Probit: $\pi_i = \Phi(X_i^\top \beta)$
- Complementary-log-log: $\pi_i = 1 - exp(-exp(X_i^\top \beta))$

Outcome: $\mathbb{R}$

Distribution.

- $Y_i|X_i \sim \mathcal{N}(Y_i|\theta_i)$
- $Y_i|X_i \sim t(Y_i|\theta_i)$

Inverse Link. Identity: $\mathbb{E}[Y_i|X_i] = g(X_i^\top \beta) = X_i^\top \beta$

Outcome: count

Distribution.

- $Y_i|X_i \sim \textit{Poisson}(\lambda_i)$
- $Y_i|X_i \sim \textit{Negbin}(\mu_i, \kappa)$

Inverse Link

- For Poisson, $\mathbb{E}[Y_i|X_i] = \lambda_i = exp(X_i^\top \beta)$, (log link)
- For Negbin, $\mathbb{E}[Y_i|X_i] = \mu_i = exp(X_i^\top \beta)$, (log link)

Score

- For sample: $S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$
- For individual, $S_i(\theta|\boldsymbol{X}) = \frac{\partial \ell_i(\theta|\boldsymbol{X})}{\partial \theta}$
- Sample score is sum of individual scores

The Hessian

$$H(\theta|\boldsymbol{X}) = \frac{\partial^2 \ell(\theta|\boldsymbol{x})}{\partial \theta \partial \theta^\top}$$

- Is a $dim(\theta) \times dim(\theta)$ matrix
- Again, Hessian for sample is sum of individual Hessians

Preview: Information Matrix, $I(\theta|\boldsymbol{X})$

- $I(\theta|\boldsymbol{X}) = -\mathbb{E}[H(\theta|\boldsymbol{X})]$
- Under correct specification, also

$$I(\theta|\boldsymbol{x}) = -\mathbb{E}[H(\theta|\boldsymbol{x})] = \mathbb{E}[S(\theta|\boldsymbol{x})S(\theta|\boldsymbol{x})^\top]$$

- Especially useful because: $var(\theta_{MLE}|\boldsymbol{x}) = [I(\theta|\boldsymbol{x})]^{-1}$

For most non-linear link functions no closed-form solution to $S(\theta) = 0$

So we use an optimization to find

$$\underset{\theta}{\textit{argmax}}\ \ell_N(\theta|\boldsymbol{X})$$

Intuition:

- You're at θ_t. Climb "uphill": $\theta_{t+1} = \theta_t + \alpha S(\theta_t)$ for some $\alpha > 0$
- Climb slower when curvature ($-H(\theta)$) is larger:

$$\theta_{t+1} = \theta_t + (-H(\theta_t)^{-1})S(\theta_t)$$

Downsides:

- must assume distribution
- must assume $\mu_i = g^{-1}(X_i^\top \beta)$
- often inconsistent if assumptions fail

Upside:

- unified approach to model wide variety of outcome variables
- when correct, best unbiased estimator (coming soon)
- often a good approach when you have a limited dependent variable

Wells in Bangladesh

we will consider a data set involving modeling the decisions of households in Bangladesh about whether to change their source of drinking water

many of the wells used for drinking water in Bangladesh and other South Asian are contaminated with naturally occurring arsenic, affecting an estimated 100 million people

arsenic is a cumulative poison, with risks of cancer and other diseases thought to be

Wells in Bangladesh

a research team from the United States and Bangladesh measured arsenic levels for all wells in a certain area, labeled the well with its arsenic level, and encouraged households drinking from unsafe wells (> 0.5μg/L) to switch to a safer well

a few years later, the researchers returned to find out who had switched wells, Switch=1 and who had not, Switch=0

the file wells.txt contains information on well switching for 3,020 households

We consider the following explanatory variables:

Arsenic: The arsenic level of the household's well

Dist: The distance to the nearest safe well

Community: Whether any members of the household are active in community organizations

Education: Years of education of the head of the household

R-squared for logistic regression?

In linear regression, R-squared is a very useful quantity, describing the fraction of the variability in the response that the explanatory variables can explain

there are a number of ways one can define an analog to R-squared in the logistic regression case, but none of them are as widely useful as R-squared in linear regression

one approach is to compute the correlation r between the observed outcomes and the fitted values

in linear regression, the square of this correlation is R-squared

thus, one reasonable way to define an R-squared for logistic regression is to square r, the Pearson correlation between the observed and fitted values

another approach is to measure the reduction in squared error:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

this approach has the advantage that it looks exactly like R-squared for linear regression, and we can therefore easily adjust for the number of parameters:

$$R^2_{\text{adj}} = 1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2 / (n - p)}{\sum_i (y_i - \bar{y})^2 / (n - 1)}$$

Does this make sense?

these two preceding measures have the advantage of working on the scale of the original variable and being easy to interpret However, one might question the logic of treating all differences (between observed and predicted) equally

compare predicted probabilities .9 with .99 for an observation with $y_i = 0$

the squared differences are similar, 0.99^2 = 0.9801 and 0.9^2 = 0.81, despite the fact that $Pr(y_i = 0)$ differs by a factor of 10 for the two models

instead of looking at the reduction in squared error, lets look at the reduction in deviance (differences on the likelihood scale)

deviance residual is the contribution of each point to the likelihood

$$d_i = s_i \sqrt{-2 \left\{ y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i) \right\}},$$

combining each of the residual deviances we get a RSS like statistic

deviance

$$D = \sum d_i^2$$

Also in other words this is -2*Loglikelihood

In principle, this could be compared to the chi-squared distribution as a rough goodness of fit test

Letting D_0 denote the null deviance (i.e., the deviance of the intercept-only, or simple binomial, model), another attempt at an Rsquared-like measure is

$$\frac{D_0 - D}{D_0} = 1 - \frac{D}{D_0},$$

the explained deviance (often reported as a percentage) Because deviance roughly follows χ^2_{n-p} distribution, it can also be adjusted for number of parameters:

$$1 - \frac{D/(n-p)}{D_0/(n-1)}$$

Letting D_0 denote the null deviance (i.e., the deviance of the intercept-only, or simple binomial, model), another attempt at an Rsquared-like measure is

$$\frac{D_0 - D}{D_0} = 1 - \frac{D}{D_0},$$

the explained deviance (often reported as a percentage) Because deviance roughly follows χ^2_{n-p} distribution, it can also be adjusted for number of parameters:

$$1 - \frac{D/(n-p)}{D_0/(n-1)}$$

Well-switching example

to get a sense of how these measures look, let's compare three models

$$\eta = \beta_0 + \beta_1 \texttt{Distance}$$
$$\eta = \beta_0 + \beta_1 \texttt{Distance} + \beta_2 \texttt{Arsenic}$$
$$\eta = \beta_0 + \text{all explanatory variables}$$

Classification

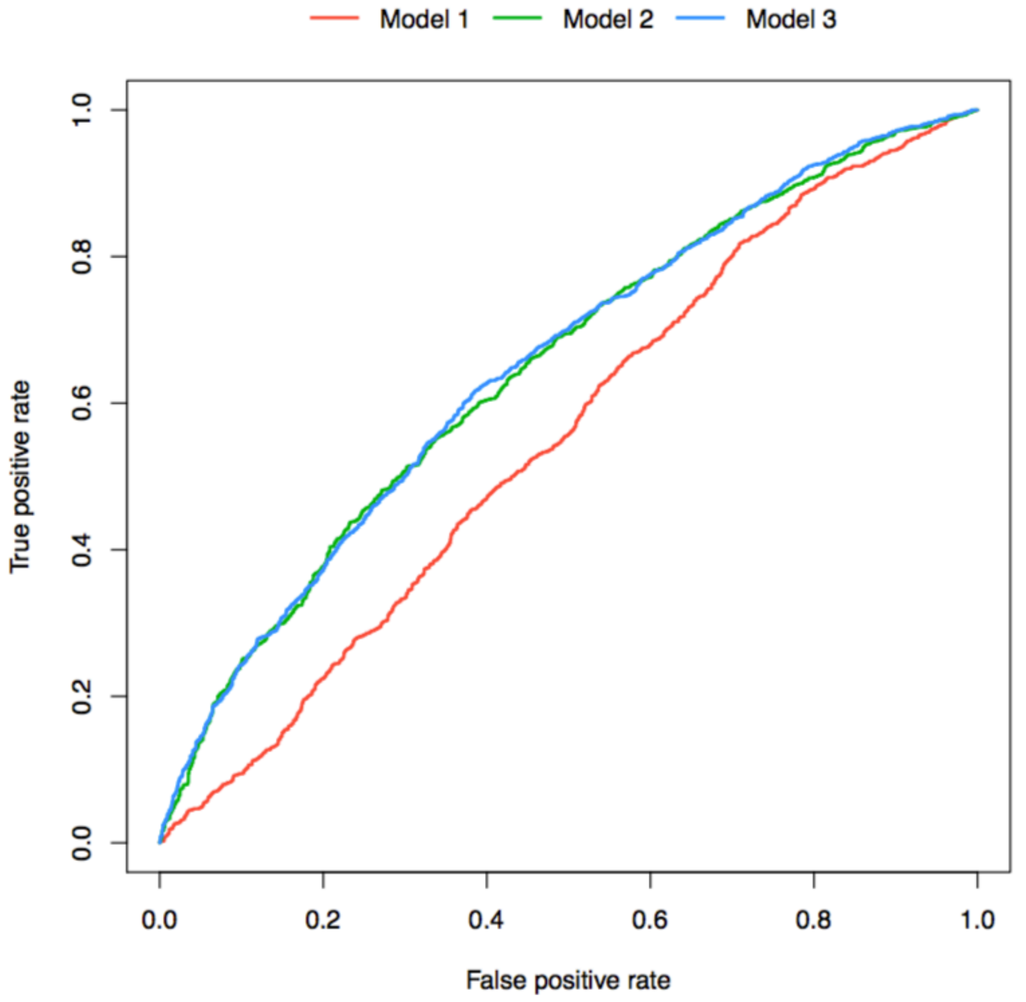
an alternative way of thinking about how well a model fits the data is with respect to classification

The obvious approach is to predict $y_i = 1$ if predicted probability is greater than 0.5, although other cutoffs could be used if, for example, the cost of false positive is larger than the cost of a false negative (or vice versa)

Classification

an alternative way of thinking about how well a model fits the data is with respect to classification

The obvious approach is to predict $y_i = 1$ if predicted probability is greater than 0.5, although other cutoffs could be used if, for example, the cost of false positive is larger than the cost of a false negative (or vice versa)


Model 1
Model 2
Model 3
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

for the three models on the previous slide, no matter what the false positive rate, models 2 and 3 had higher true positive rates than model 1

however, comparing models 2 and 3, either model could be “on top” depending on where we are at on the curve

A useful summary of the overall quality of the curve is the area under the curve, or AUC

	Model 1	Model 2	Model 3
AUC	0.55	0.65	0.65

note that random guessing would yield an AUC of 0.5; perfect classification would yield an AUC of 1

Basic principles of model selection

let’s remind ourselves of the basic principles of model selection

simple models have low variance, but risk bias

more complicated models reduce bias and fit the sample data better, but can be highly variable and do not necessarily generalize to the population better

model selection criteria can be informative, but should not be applied blindly – there is no substitute for thinking carefully about the scientific meaning and plausibility of the models under consideration

AIC

consider the expected prediction accuracy of a model using log-likelihood as a measure of accuracy:

$$\mathrm{E} \sum_i \log \Pr_{\hat{\theta}}(Y_i),$$

where $\hat{\theta}$ is the MLE of the parameters of the distribution function for y and the $\{Y_i\}$ are out-of-sample random variables (i.e., not the $\{y_i\}$ used to fit the model)

Akaike showed that

$$-2\mathrm{E} \sum_i \log \Pr_{\hat{\theta}}(Y_i) \approx -2\mathrm{E}(\mathrm{loglik}) + 2p,$$

where loglik is the log-likelihood of the little model

this suggests the following criterion, named the Akaike information criterion:

$$\text{AIC} = -2\text{loglik} + 2p = D + 2p$$

certainly, a lower AIC is better than a higher AIC (we wouldn't want our expected deviance to be large), but suppose the AIC values for two models differ by, say, 1; is that a meaningful difference?

a useful rough guide is that AIC differences under 2 are not particularly meaningful, AIC differences of around 5 are fairly convincing, and AIC differences over 10 provide clear support for the model with the lower AIC

BIC

another common information model selection criterion for GLMs is called the Bayesian information criterion, or BIC

as you might guess, its derivation is Bayesian and beyond the scope of this course

however, its form happens to be very similar to AIC:

$$\text{BIC} = -2\text{loglik} + p\log(n) = D + p\log(n)$$

Note that because log(n) is bigger than 2 (unless n < 8), BIC penalizes model complexity more heavily than AIC, and thus tends to favor more parsimonious models

BIC has a direct Bayesian interpretation in that it allows you to calculate (approximately, given equal prior probability on each model) the posterior probability of each model under consideration:

$$P(M_j|\mathbf{y}) \approx \frac{\exp(-0.5\mathrm{BIC}_j)}{\sum_k \exp(-0.5\mathrm{BIC}_k)},$$

where the sum is over the models under consideration

Applying AIC and BIC to our three models from earlier:

	Model		
	1	2	3
AIC	4080	3937	3918
BIC	4092	3955	3948
$P(M_j\|\mathbf{y})$	0.00	0.03	0.97

both approaches agree that the most complex model is the best despite its extra parameters, although BIC is less enthusiastic about the difference between models 2 and 3

it is important to keep in mind the famous words of George Box:

‘All models are wrong, but some are useful.’

Certainly, a useful model should fit the data well, and information criteria are helpful guides here, but other considerations, such as interpretability and scientific justification are also important

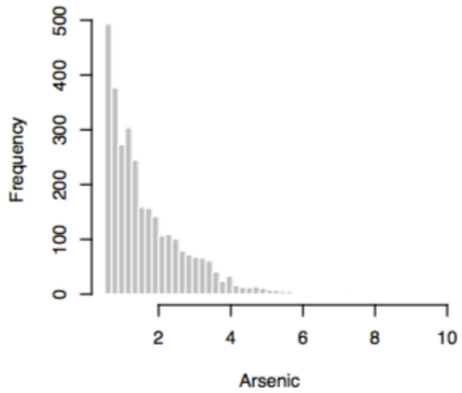
there are no explicit rules to follow when building a model – other than, perhaps, don’t build a model by blindly following rules – and different people could look at this same data and build different models

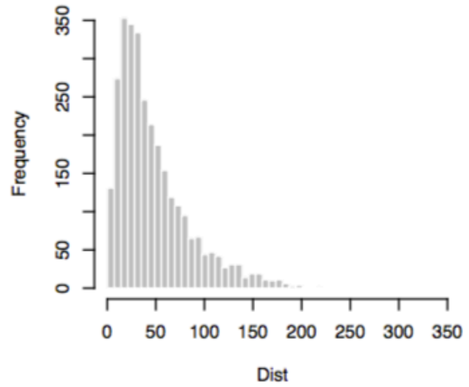
generally, I use AIC/BIC to judge the quality of a GLM's fit to the data, weighted roughly according to the scale discussed in the previous lecture

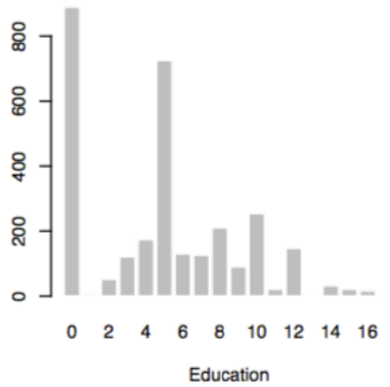
the other primary consideration I use is whether a variable and the sign of its coefficient makes sense (i.e., whether, before seeing the data, I thought it would have been important and act in the appropriate direction)

for example, if a variable leaves the AIC essentially unchanged, whether to include it or not depends on external considerations: if it makes sense and has a believable direction, leave it in; if not, we might as well remove it

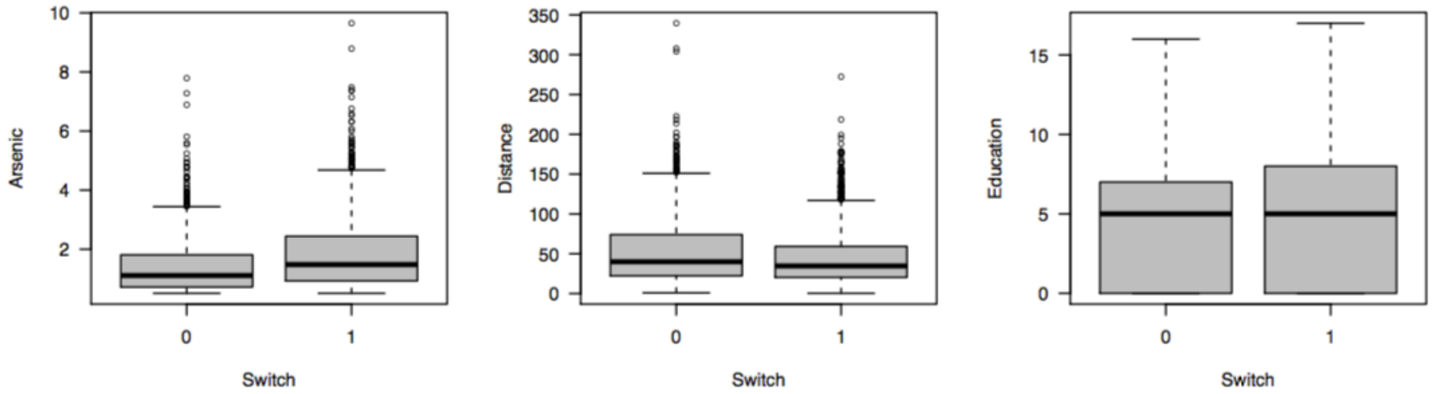
One should always start by looking at descriptive statistics and get a sense of each variable, Switch 58%; Community 42%







Next its useful to compare each variable to the outcome:



Community = Yes: 55% switched; Community = No: 59% switched

A reasonable place to start is the additive model

As always, however, the regression coefficients have little meaning unless we consider the scale of each variable

For example, distance is measured in meters, so all the distance coefficient tells us is how much the log-odds of well-switching change if we compare a well 90 meters away vs. a well 91 meters away; obviously this will be an inconsequential change

On the other hand, if distance were measured in kilometers it would be enormous

Standardization

a reasonable rule is to set changes to be roughly two standard deviations (the reason for the factor of 2 is so that they may be compared to 0-1 variables, which have standard deviations of

this amounts to a change in distance of approximately 100 meters, a change in arsenic levels of approximately 2μg/L, and a change in education of 8 years

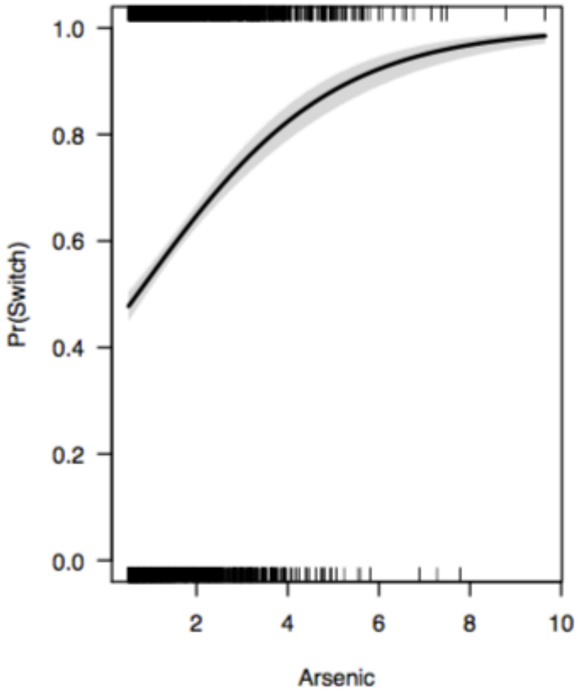
while we're at it, it would also be helpful to subtract off the mean of each of these variables; otherwise, the intercept,0, is essentially meaningless, as it corresponds to an arsenic level of 0

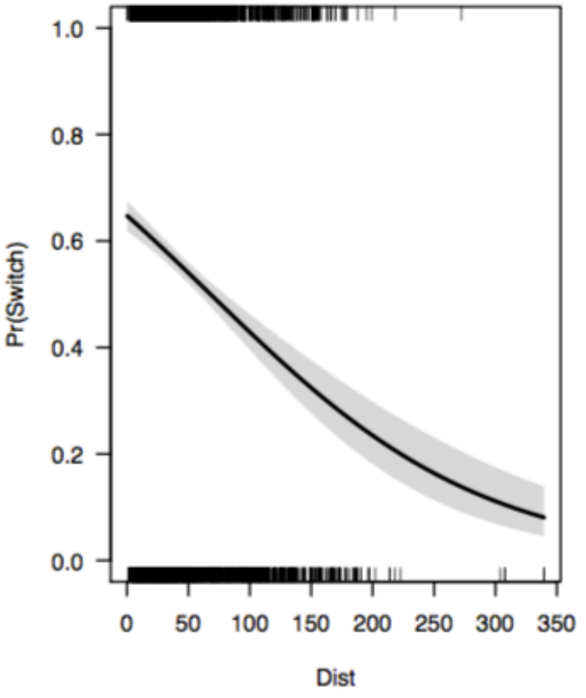
the following table, with variables sorted in terms of importance

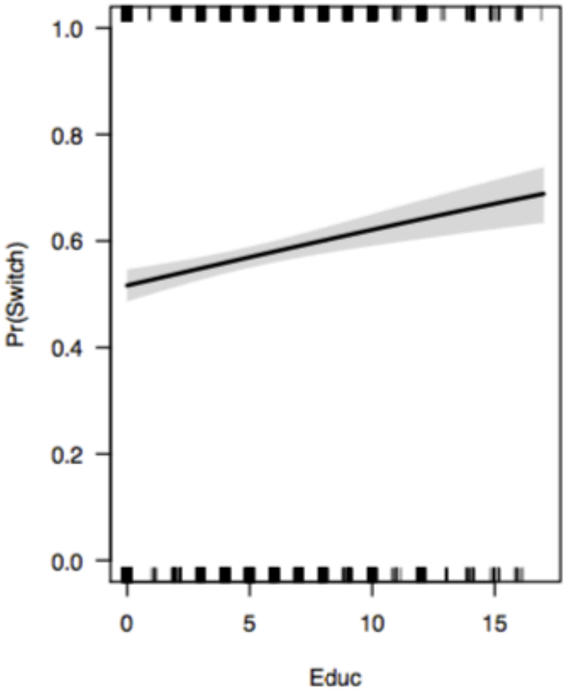
	β	SE
Arsenic	0.93	0.08
Distance	-0.90	0.10
Education	0.34	0.08
Community	-0.12	0.08

from here on, I'll discard Community from the model: there is no statistical justification for it (not significant, doesn't improve AIC), and I find its negative sign difficult to explain/interpret

In short, it seems to add nothing but clutter to our model


Pr(Switch)
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
Arsenic


Pr(Switch)
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
350
Dist


Pr(Switch)
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
Educ

the main assumptions that our model at this point is making are:

- there are no other (unmeasured) covariates that affect the probability of switching
- the effect of each variable is linear
- no interactions

The first assumption is almost certainly not true, but there's nothing we can do about it after the data has been collected

The second and third assumptions, however, we can investigate

transformations

given the highly skewed distributions for distance and arsenic, a log transformation would seem reasonable

gor education, the bar plot we looked at earlier would suggest the following education categories:

None: Educ=0
Low: Educ $\in$ [1-5]
Medium: Educ $\in$ [6-10]
High: Educ $\geq$ 11

unlike linear regression, it is diffcult (at least in my opinion) to sense an appropriate transformation in logistic regression by looking at residuals, but we can try fitting these models and seeing what happens with the AIC

the results

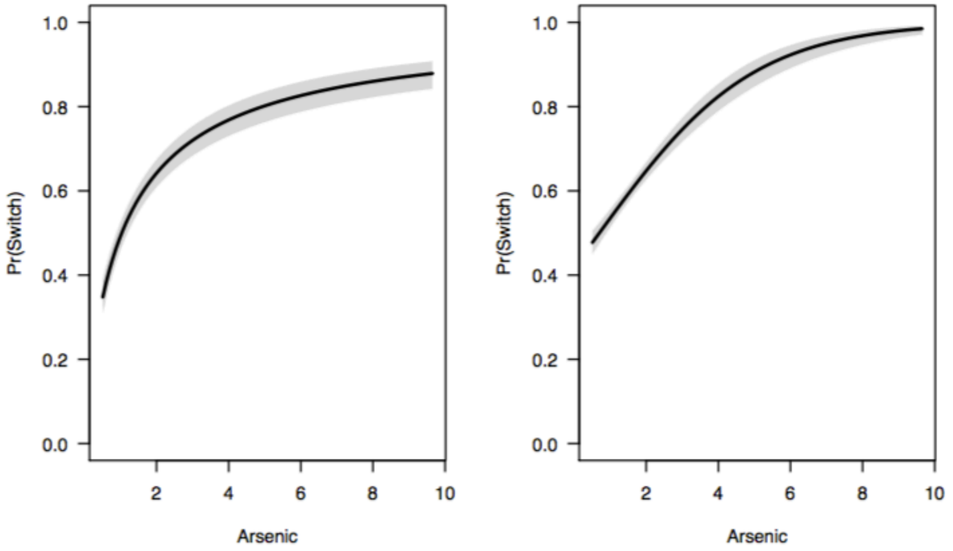
Transformation	AIC
None	3918
log(Arsenic)	3886
log(Distance)	3951
Ed. Categories	3910

this is fairly convincing evidence that we should adopt the log(Arsenic) and education category transformations, but stick with a linear effect of distance (this model has an AIC of 3878

note the effect of the log transformation: compared with the earlier model, the probability of switching changes more rapidly for low arsenic levels, but is flatter for high arsenic levels

this implies that people react more strongly to the difference between arsenic levels of 1 and 2 than they do between arsenic levels of 5 and 6 (which seems perfectly reasonable)

for education, our new model indicates that there is little difference between individuals with none or little education, but that higher levels of education are associated with increased probability of switching (again, this seems reasonable)



finally, let’s consider interactions

in this example, where we only have three variables, might as well just consider all three two-way interactions

if we had more terms, we would have to narrow our focus; typically this would involve:

- concentrating on interactions involving the most important main effects
- concentrating on interactions involving the primary treatment or exposure (if there is one)

the results

Interaction	AIC
None	3878
Arsenic × Distance	3879
Education × Distance	3860
Arsenic × Education	3878

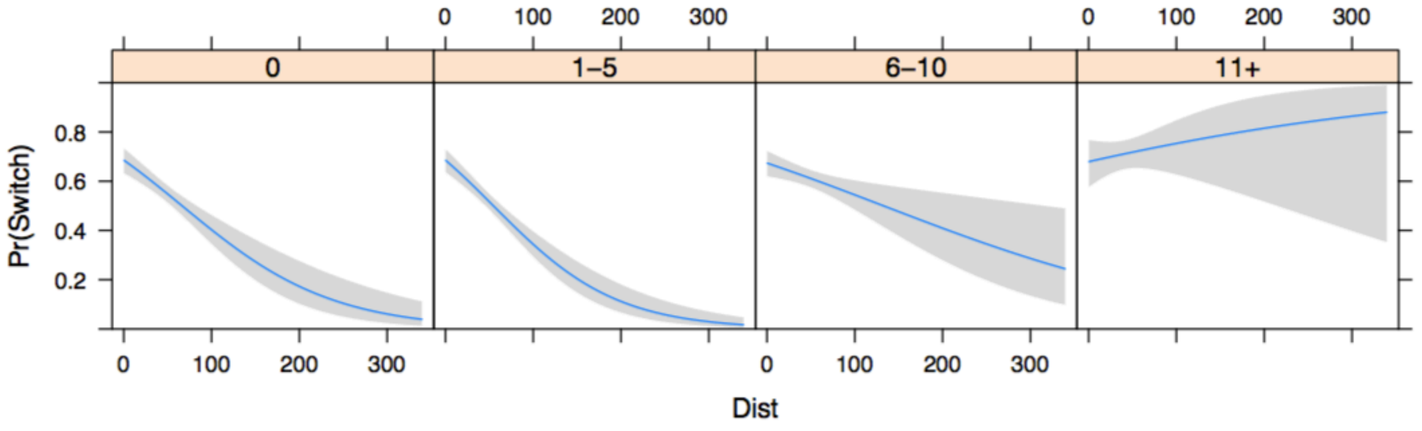
so it seems that we have pretty strong evidence that the effect of distance is not the same for all levels of education; the other interactions don't seem to add much

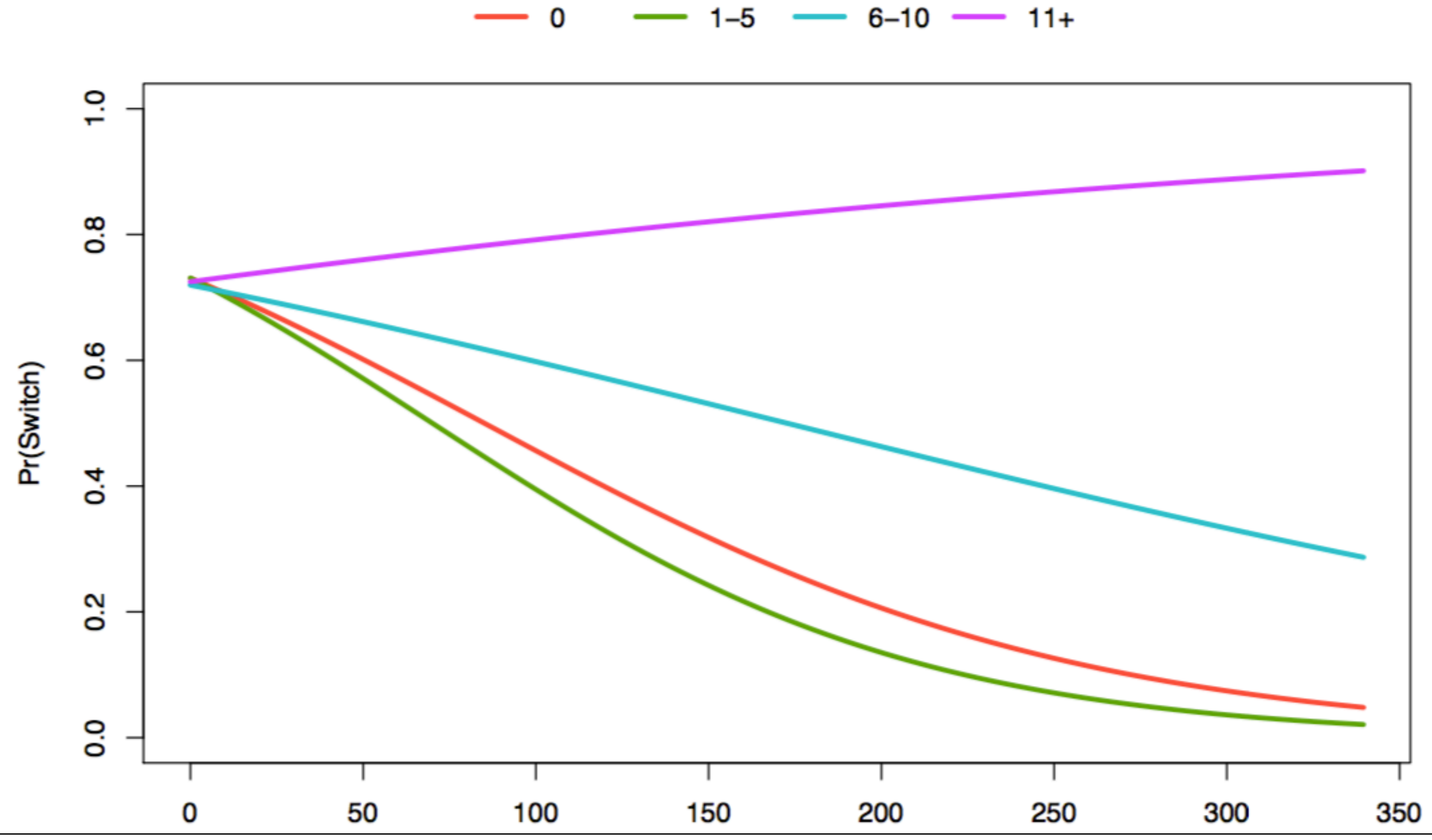
thus, distance has a large effect for low-education households, a moderate effect for medium-education households, and seemingly no e↵ect for high-education households

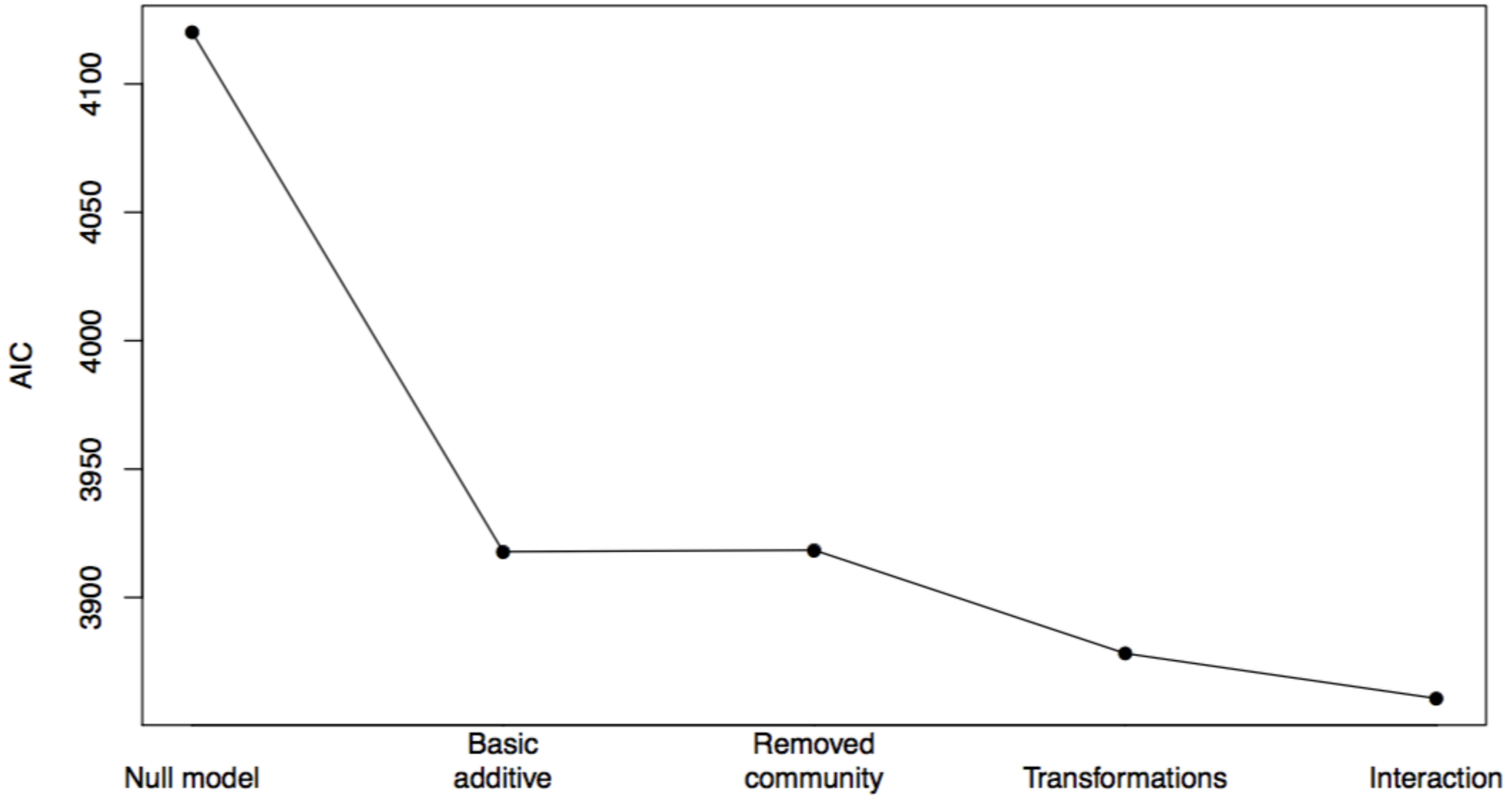
presumably this is due not so much to education per se, but because education serves as a marker for socioeconomic status

thus, while individuals with low education are presumably poorer and have to walk to the next well, better-educated individuals are more likely to be wealthier and able to afford other, less onerous means of transportation

when communicating the results of models with interactions, reporting regression coefficients directly is rarely (really depends on audience) a good idea – only a small fraction of your audience will usually be able to discern their meaning


0
1–5
6–10
11+
Pr(Switch)
0.8
0.6
0.4
0.2
0
100
200
300
Dist


0
1–5
6–10
11+
Pr(Switch)
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
350


AIC
4100
4050
4000
3950
3900
Null model
Basic additive
Removed community
Transformations
Interaction

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 $_i$, depending on the explanatory variables

The random and systematic components are as follows:

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

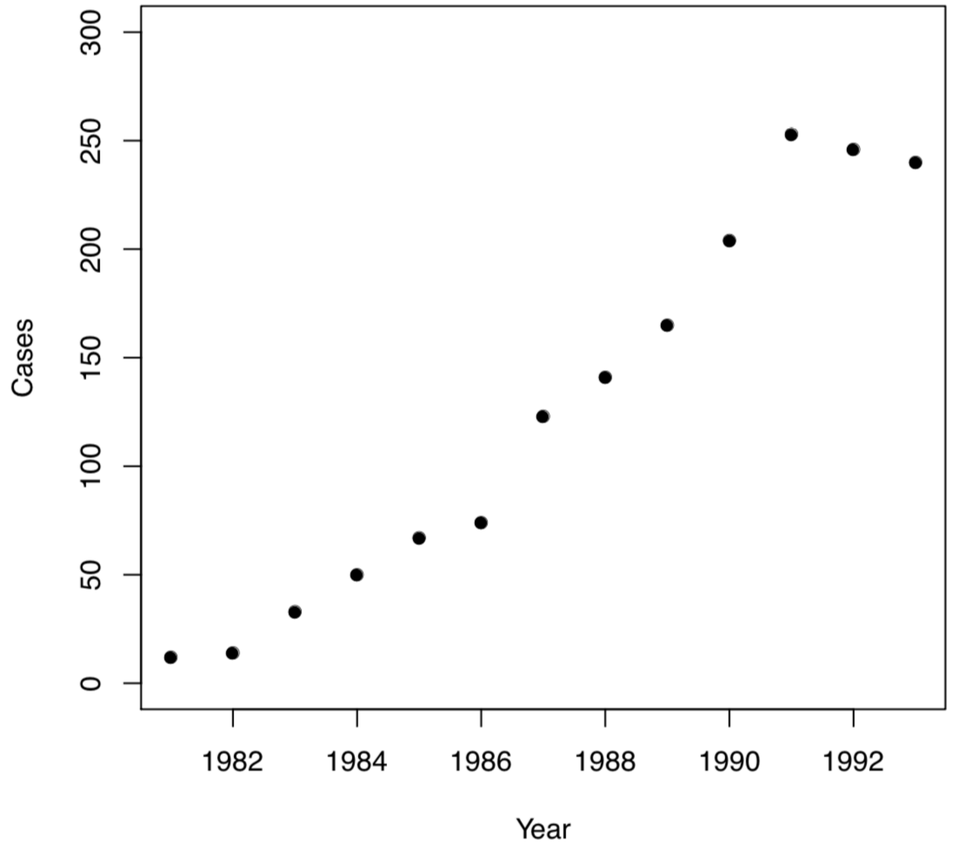
The Poisson distribution has a log link function

thus

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

note that the link ensures that lambda > 0, as it must be for the Poisson distribution

as a first example of Poisson regression, consider the following data on the number of new cases of AIDS in Belgium, 1981–1993:



exponential growth models are reasonable in the early stages of an epidemic

the simple linear model

$$\eta_i = \beta_0 + \beta_1 \texttt{Year},$$

when combined with a log link, is equivalent to fitting the exponential growth model

$$\lambda_i = \gamma \exp(\delta t_i),$$

where $\beta_0 = \log(\gamma)$ and $\beta_1 = \delta$

the standard R model output results in

	β	SE
Intercept	-397.06	15.46
Year	0.20	0.01

what does the intercept mean here?

re-centering year so that it begins at the start of the study (1981), we obtain a meaningful intercept:

	Estimate	Std. Error
Intercept	3.34	0.07
Year	0.20	0.01

recall that we are modeling with a log link; thus in 1981 the model estimates

$$e^{3.34} = 28.2$$

consider two hypothetical observations with different explanatory variables x_1 and x_2; the Poisson GLM with log link implies that

$$\begin{aligned}\frac{\lambda_2}{\lambda_1} &= \frac{\exp(\eta_2)}{\exp(\eta_1)} \\ &= \exp((\mathbf{x}_2 - \mathbf{x}_1)^T \boldsymbol{\beta})\end{aligned}$$

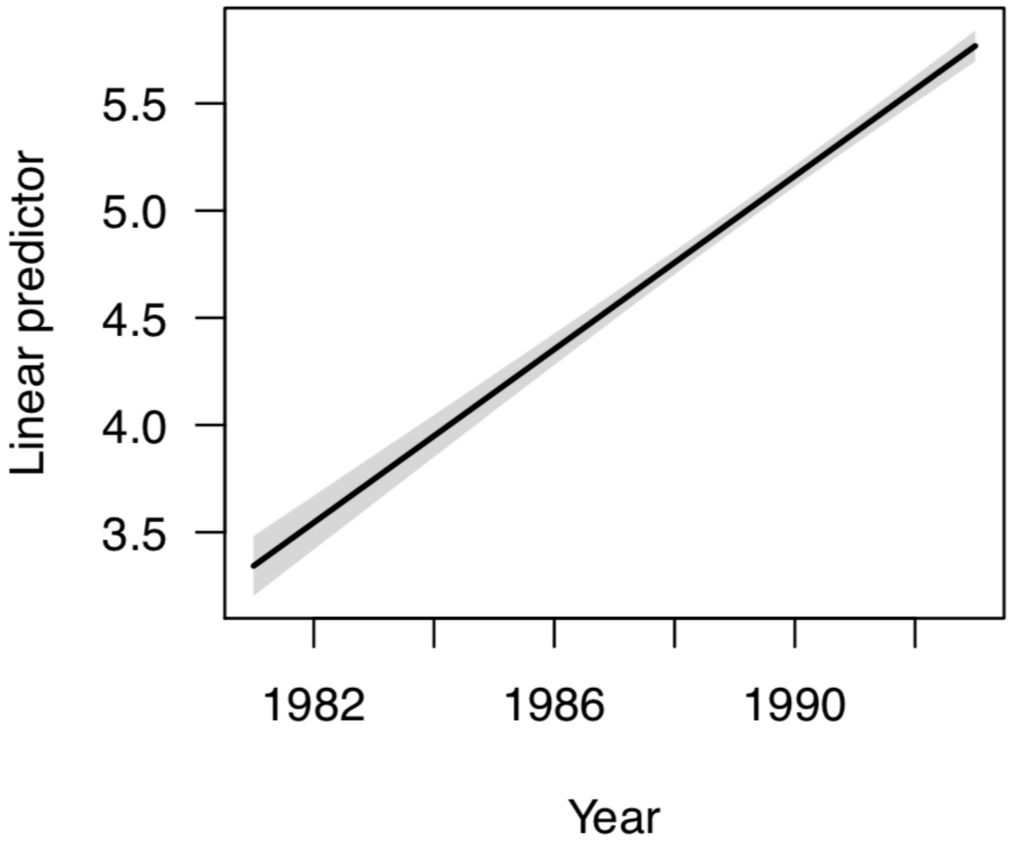
In particular, if variable j changes by an amount δ_j, then the rate *ratio* λ_2/λ_1 is $\exp(\delta_j \beta_j)$

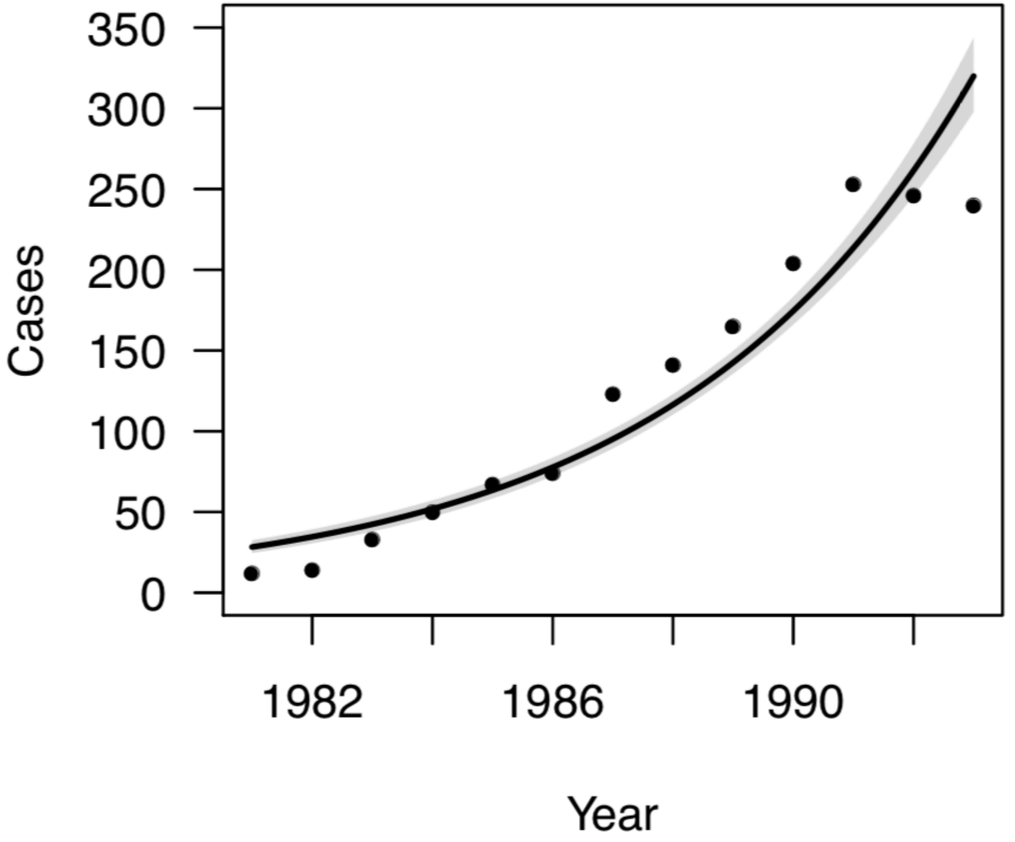
rate ratios (RR) are a common way of describing the coefficients of a Poisson regression model, putting them on a scale that is more interpretable, analogous to the use of odds ratios in logistic regression models

so, our regression coe cient of 0.20 implies that the rate ratio is $e^{0.\bar{2}0} = 1.2$ the number of AIDS cases in Belgium increased by 20% each year over the time span 1981-1993

Another way of putting it is that $e^{5(0.20)} = 2.7$ the number of AIDS cases increased by 170% every five years

Or yet another way of putting it, $e^{3.5(0.20)} = 2$ the number of AIDS cases doubled every 3.5 years


Linear predictor
5.5
5.0
4.5
4.0
3.5
1982
1986
1990
Year


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

so how effective is our model at predicting the outcome?

as with logistic regression, two measures are commonly used: reduction in squared error and deviance explained

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{\lambda}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

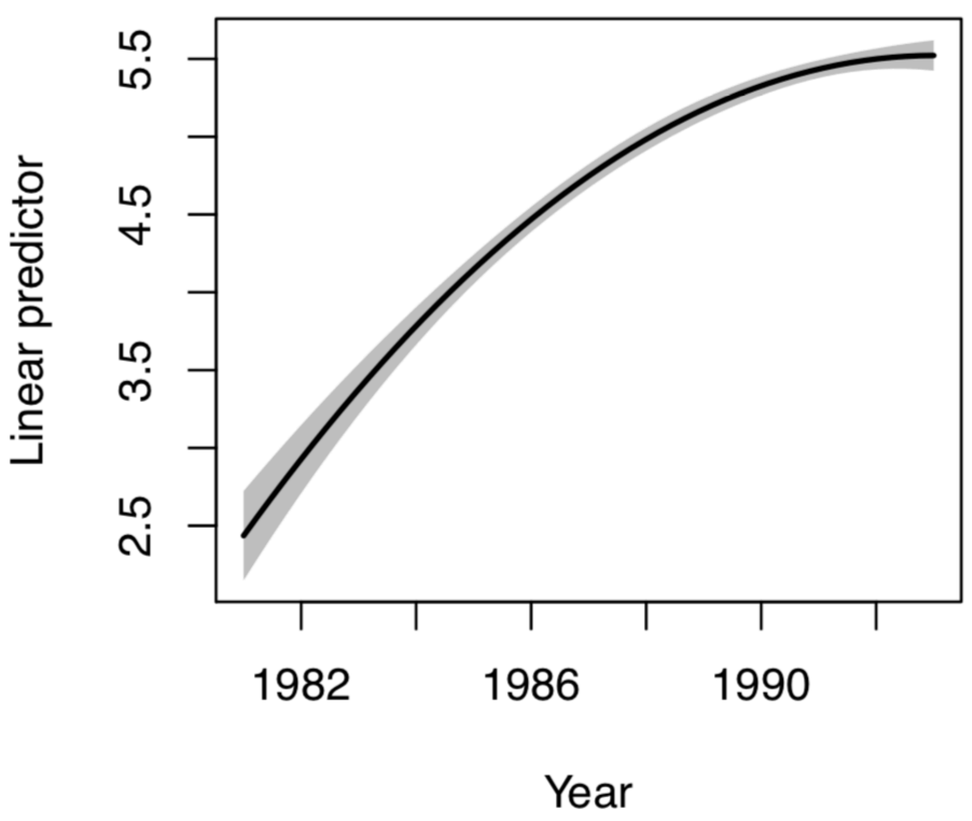
the reduction in squared error is

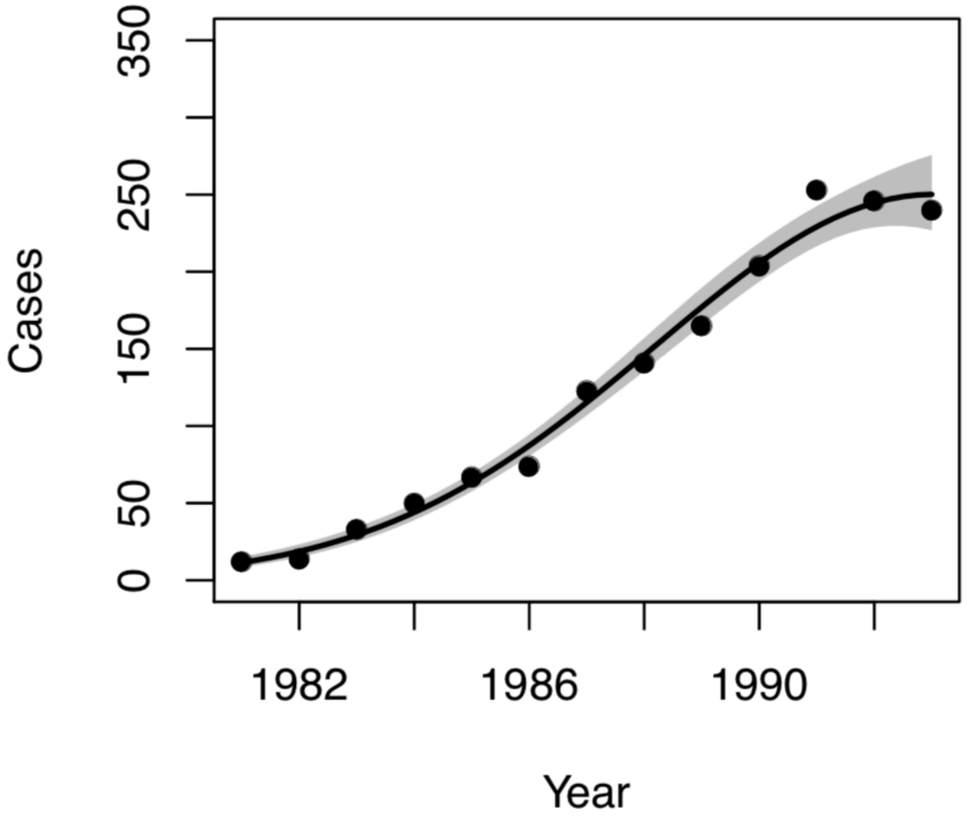
$$1 - \frac{D}{D_0}$$

Once again, both measures can be adjusted for number of parameters by dividing the numerator by n p and the denominator by n 1

		R^2	R^2_{adj}	DE	DE_{adj}
1981–1993	Linear	0.880	0.869	0.907	0.899
1981–1991	Linear	0.973	0.970	0.964	0.960
1981–1993	Quadratic	0.988	0.986	0.989	0.987

AIC also strongly favors a quadratic model (166 vs. 97)


Linear predictor
2.5
3.5
4.5
5.5
1982
1986
1990
Year


Cases
0
50
150
250
350
1982
1986
1990
Year