Biostatistics 140.754
Advanced Methods in Biostatistics IV

Jeffrey Leek

Assistant Professor
Department of Biostatistics
jleek@jhsph.edu

Tip + Paper

Tip Try to write your first paper as soon as you possibly can and try to do as much of it on your own as you can. You don't have to wait for faculty to give you ideas, read papers and think of what you think would have been better (you might check with a faculty member first so you don't repeat what's done, etc.). You will learn more writing your first paper than in almost any/all classes.

Paper of the Day: "An index to quantify an individual's scientific research output"
http://www.pnas.org/content/102/46/16569.full
(Jeff's $= 8$, Rafa's $= 36$, Bert Vogelstein > 160)
Open problem: how do you predict who will be a "productive" scientist based on their graduate/postdoc work.

From the Glynn Paper

Title: Alleviating linear ecological bias and optimal design with subsample data

Abstract:
We illustrate that combining ecological data with subsample data in situations in which a linear model is appropriate provides two main benefits. First, by including the individual level subsample data, the biases that are associated with linear ecological inference can be eliminated. Second, available ecological data can be used to design optimal subsampling schemes that maximize information about parameters.We present an application of this methodology to the classic problem of estimating the effect of a college degree on wages, showing that small, optimally chosen subsamples can be combined with ecological data to generate precise estimates relative to a simple random subsample.

Outline For Today

- Parametric examples
- GLMs - the details
- Model checking
- Interactions

GLMs: The three parts

Generalized linear models are a very common data analysis tool. As we have seen, they often estimate sensible parameters, even when the parametric assumptions don't hold [1] The three components of a GLM are

- The random component: outcome Y_i, covariates $x = (x_{i1}, \ldots, x_{ip})$
- Systematic component $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, $i = 1, \ldots, n$ or equivalently $\eta = x\boldsymbol{\beta}$
- The link function $\mu_i = \mathbb{E}(y_i)$ linked to the linear predictor by $\eta_i = g(\mu_i)$

[1] GLMs can also give you a place to start when trying to define the scale of differences for non-parametric regression. But you don't have to be limited to these choices!

GLMs: Some important properties

- $\ell(\theta_i, \phi | y_i) = \log f(y_i|\theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi)$
- $\mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right) = \mathbb{E}\left(\frac{f'(y_i|\theta_i,\phi)}{f(y_i|\theta_i,\phi)}\right) = \int f'(y_i|\theta_i, \phi) = \frac{\partial}{\partial \theta_i} \int f(y_i|\theta_i, \phi) = 0$
- So $\mathbb{E}\left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right) = 0$ or $\mu_i = \mathbb{E}[y_i] = b'(\theta_i)$
- $\frac{\partial^2 \ell}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a(\phi)}$ and recall $\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_i^2}\right) = -\mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta_i}\right)^2\right]$
- $\implies \frac{b''(\theta_i)}{a(\phi)} = \mathbb{E}\left(\frac{(y_i - b'(\theta_i))^2}{a(\phi)^2}\right) = \frac{1}{a(\phi)^2}\mathrm{Var}(y_i)$
- $\implies a(\phi)b''(\theta_i) = \mathrm{Var}(y_i)$

GLMs: example of properties

Example 1: $y_i \sim \text{Poisson}(\mu_i)$

$$\begin{aligned} f(y_i|\mu_i) &= \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i} \\ &= e^{y_i \log \mu_i - \mu_i - \log(y_i)} \\ &= e^{y_i\theta_i - e^{\theta_i} - \log(y_i!)} \end{aligned}$$

where $\theta_i = \log(\mu_i)$, $b(\theta_i) = e^{\theta_i}$, $c(y_i) = \log(y_i)$

$\implies \mathbb{E}[y_i] = b'(\theta_i) = e^{\theta_i} = \mu_i$ and $\text{Var}(y_i) = b''(\theta_i) = e^{\theta_i} = \mu_i$

For fun - work out the same things for the Binomial model.

GLMs: the link function

- $\mu_i = \mathbb{E}[y_i]$ linked to linear predictor by $\eta_i = g(\mu_i)$
- g is a monotone differentiable function, i.e. $g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j$
- The g that gives $g(\mu_i) = \theta_i$ is called the canonical link.

Example: $y_i \sim \text{Poisson}(\mu_i)$

$\mathbb{E}[y_i] = \mu_i = e^{\theta_i} \implies g(\mu_i) = \log(\mu_i)$

GLMs: ML parameter estimation

$$L(\theta_i, \phi) = \sum [y_i \theta_i - b(\theta_i)]/a(\phi) + \sum c(y_i|\phi)$$

For the canonical link, $\eta_i = \theta_i = \sum_j x_{ij}\beta_j$ and the kernel of the log likelihood simplifies to: $\sum_i y_i(\sum_j x_{ij}\beta_j) = \sum_j \beta_j \underbrace{\sum_i y_i x_{ij}}_{\text{sufficient stat.}}$

Likelihood equations Contribution of the ith observation is:
$\ell_i = [y_i\theta_i + b(\theta_i)]/a(\phi) + c(y_i, \phi)$

$$S_i(\beta_j) = \frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{(y_i - \mu_i)}{a(\phi)}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \left(\frac{V(y_i)}{a(\phi)}\right)^{-1} = \frac{a(\phi)}{V(y_i)}$$

$$\frac{\partial \mu_i}{\partial \beta_j} = x_{ij}$$

GLMs: ML parameter estimation (cont.)

$\ell(\beta) = \sum_{i=1}^{n} \ell_i(\beta)$ so the score equations are:

$$\sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{V(y_i)} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right) x_{ij} = 0$$

These are non=linear functions of β and must be solved iteratively. There are two ways of doing this that get used:

1. Newton-Raphson [2] $\hat{\beta}^{[k+1]} = \hat{\beta}^{[k]} - \left(H^{[k]}\right)^{-1} S^{[k]}$ where H is the Hessian matrix $\left(\frac{\partial^2 \ell(\beta)}{\partial \beta_j \partial \beta_k} \right)$ and S is the score $\nabla \ell(\beta)$ with elements $\frac{\partial \ell(\beta)}{\partial \beta_j}$
2. Fisher scoring $\hat{\beta}^{[k+1]} = \hat{\beta}^{[k]} + \left(I_{inf}^{[k]}\right)^{-1} S^{[k]}$ where $I_{inf}^{[k]} = -\mathbb{E}[H]^{[k]}$ is the the kth approximation for the estimated expected information matrix [3].

[2] Simple form is: $x^{[k+1]} = x_{[k]} - f(x^{[k]})/f'(x^{[k]})$

[3] like NR but uses the expected information rather than the observed

Fisher Scoring: calculating the expected Hessian

$$
\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 \ell_i}{\partial \beta_i \beta_k}\right] &= -\mathbb{E}\left[\frac{\partial \ell_i}{\partial \beta_j} \cdot \frac{\partial \ell_i}{\partial \beta_k}\right] \\
&= -\mathbb{E}\left\{\left(\frac{(y_i - \mu_i)x_{ij}}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right) \times \left(\frac{(y_i - \mu_i)x_{ik}}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right)\right\} \\
&= -\frac{x_{ij}x_{ik}}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2
\end{aligned}
$$

$$
\implies \mathbb{E}\left[\frac{\partial^2 \ell}{\partial \beta_i \beta_k}\right] = \sum_{i=1}^{n} -\frac{x_{ij}x_{ik}}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \sum_{i=1}^{n} w_i x_{ij} x_{ik}
$$

where $w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 / V(y_i)$

So the information matrix is $I = X'WX$ where W is a diagonal matrix with elements w_i [4] [5]

[4] The asymptotic covariance matrix of $\hat{\beta}$ is the inverse of the information matrix

[5] ML estimation using N-R or F-S corresponds to IRLS for a linearized version of the link function

GLMs: family's and link functions

	Family Name				
Link	`binomial`	`Gamma`	*`gaussian`*	`inverse-gaussian`	`poisson`
`logit`	D				
`probit`	•				
`cloglog`	•				
`identity`		•	D		
`inverse`		D			
`log`		•			D
$1/\mu^2$				D	
`sqrt`					•

GLMs: family's and link functions

	Family Name				
Link	`binomial`	`Gamma`	*`gaussian`*	`inverse-gaussian`	`poisson`
`logit`	D				
`probit`	•				
`cloglog`	•				
`identity`		•	D		
`inverse`		D			
`log`		•			D
$1/\mu^2$				D	
`sqrt`					•

Model-checking†

Up to now, we have distinguished key differences between modes of frequentist inference[6]. Under any of these modes, in many applied settings it will be appropriate to perform some form of 'check' on what you did:

- In nonparametric work, we want to ensure that the regression's summary of the underlying super-population appears to be useful (a.k.a. not missing the point)
- In semiparametric and parametric work, we want to check modeling assumptions (... so we don't get yelled at)

The mainly-graphical checks we use in either case are similar, although the interpretation changes.

The goal of these checks is to spot major aberrations - note there is typically very little power to spot anything else, without strong assumptions.

[6] non-,semi-, parametric

Model-checking†

In nonparametric work, there's always the option to maintain your single-minded interest in the pre-specified parameter $\theta(F)$ and to do no checks.

- This is fine, in e.g. a clinical trial (Why?)
- "Switching horses" to some $\theta'(F)$ and starting again messes up frequentist calibration, at least to some extent (Why? How?)
- Nevertheless, not-quite-honest inference on $\theta'(F)$ may be more scientifically useful than accurate inference on $\theta(F)$. (When?)

So, sometimes they're justified, sometimes they're not...

Model-checking†

Non-trivial issues to consider before doing checks

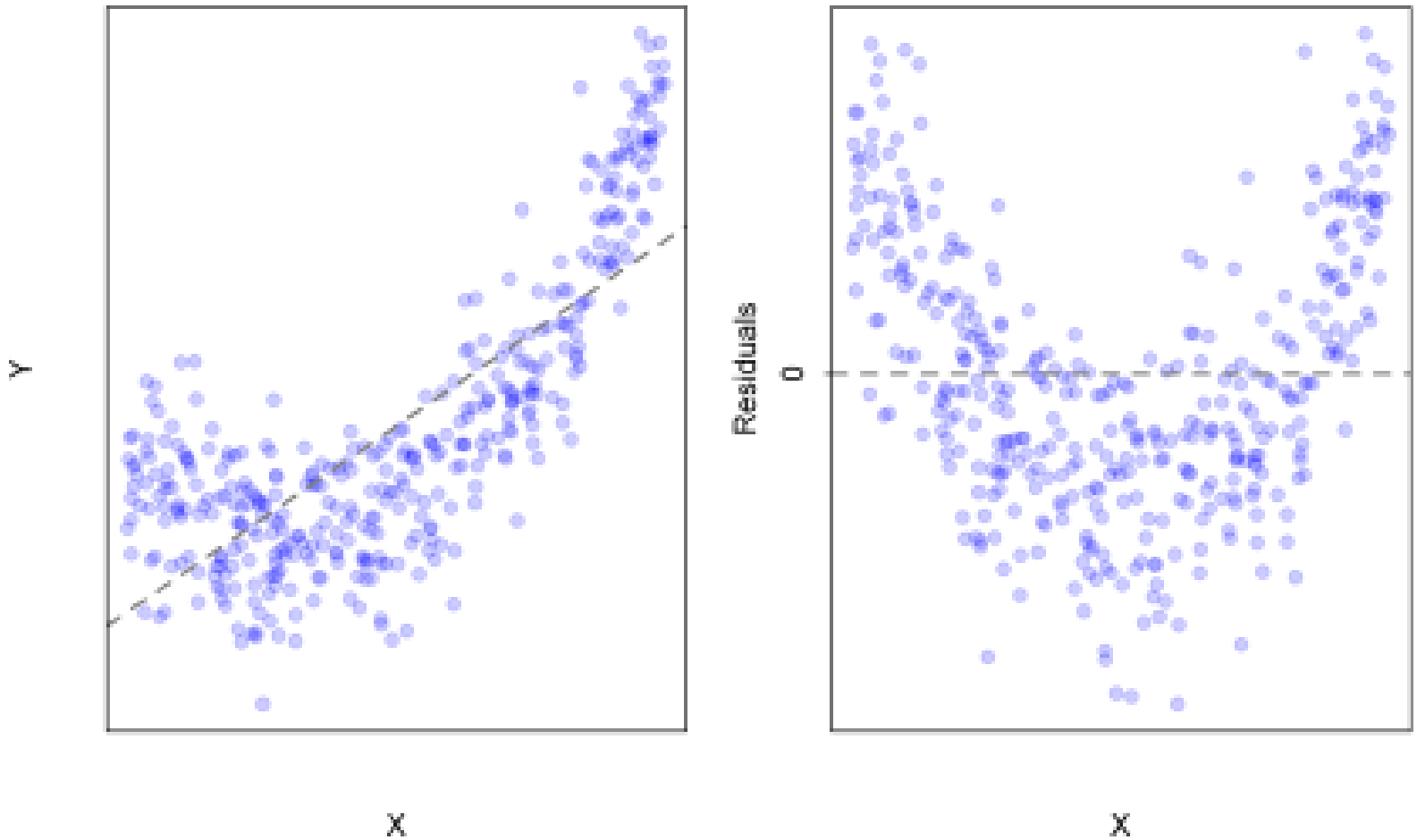
- In some applied areas, you are simply not allowed to do model checking. Model checking's ad hoccery introduces "wiggle room" into the inferential process. If sufficient money/prestige is available, it's naïve to think people will not "wiggle" to the most lucrative answer available
- Not every statistician or scientist is aware of this problem and understandably, those doing checks in good faith resent being called deceitful. Be tactful.
- "Missing" something important in your data is Very Bad

Rarer Referees may try forcing checking procedures on you, even when not needed. Your choice may be to comply, or argue back, or publish elsewhere

Trading these off is application-specific, and also a matter of your personal taste

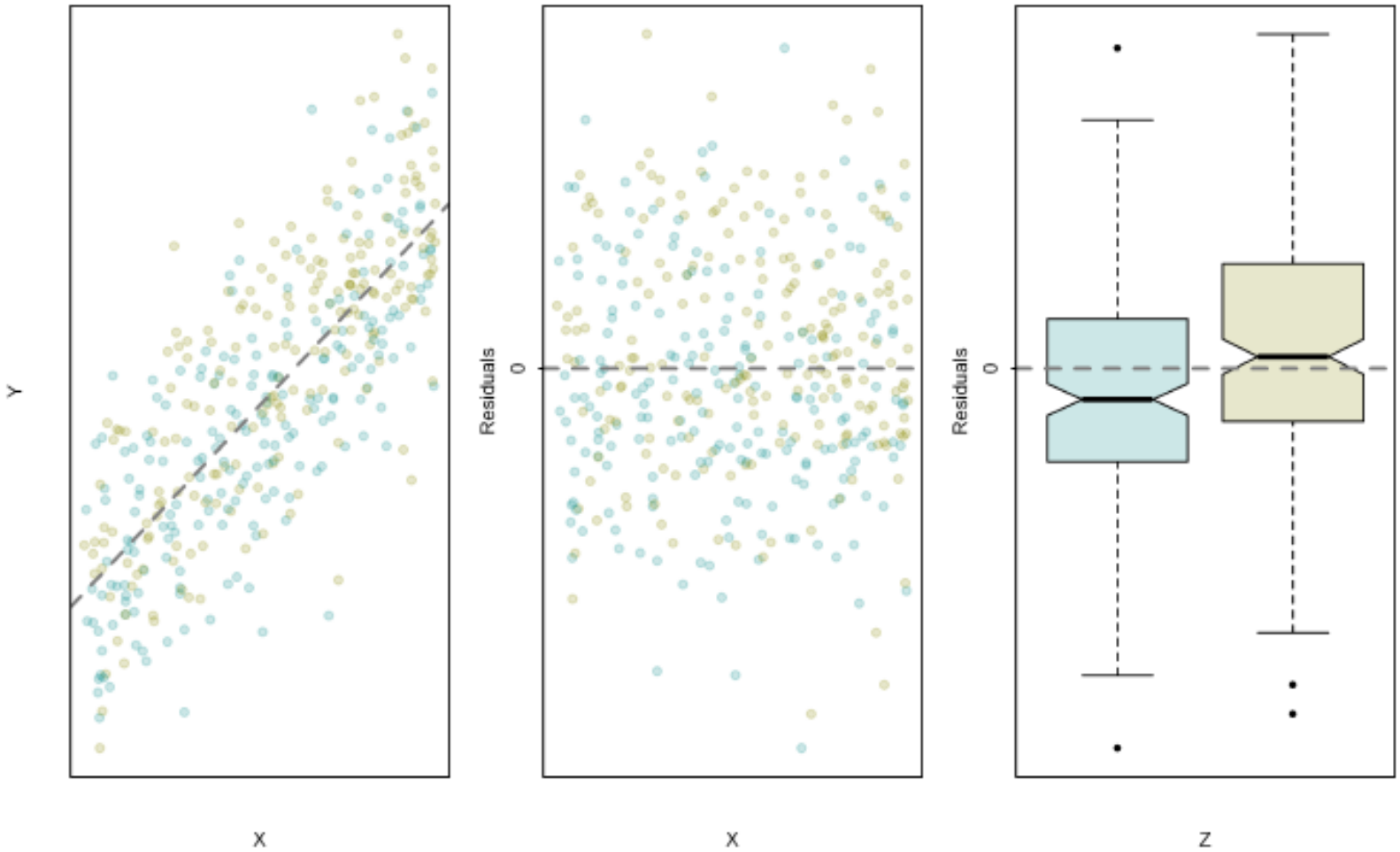
Model-checking: without the "model" bit†

It's likely you'd want to know about these:



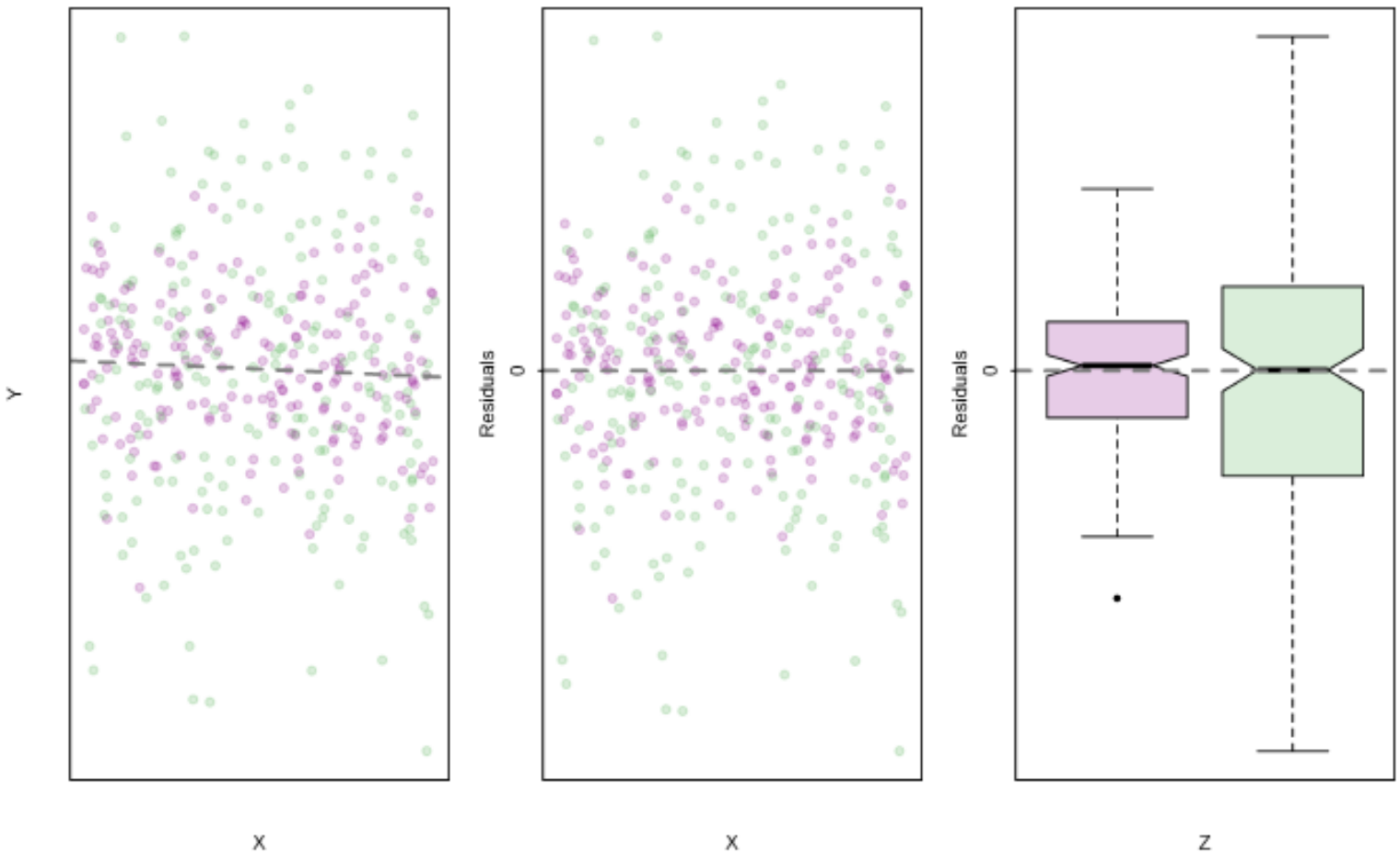
Model-checking: without the "model" bit†

It's likely you'd want to know about these:



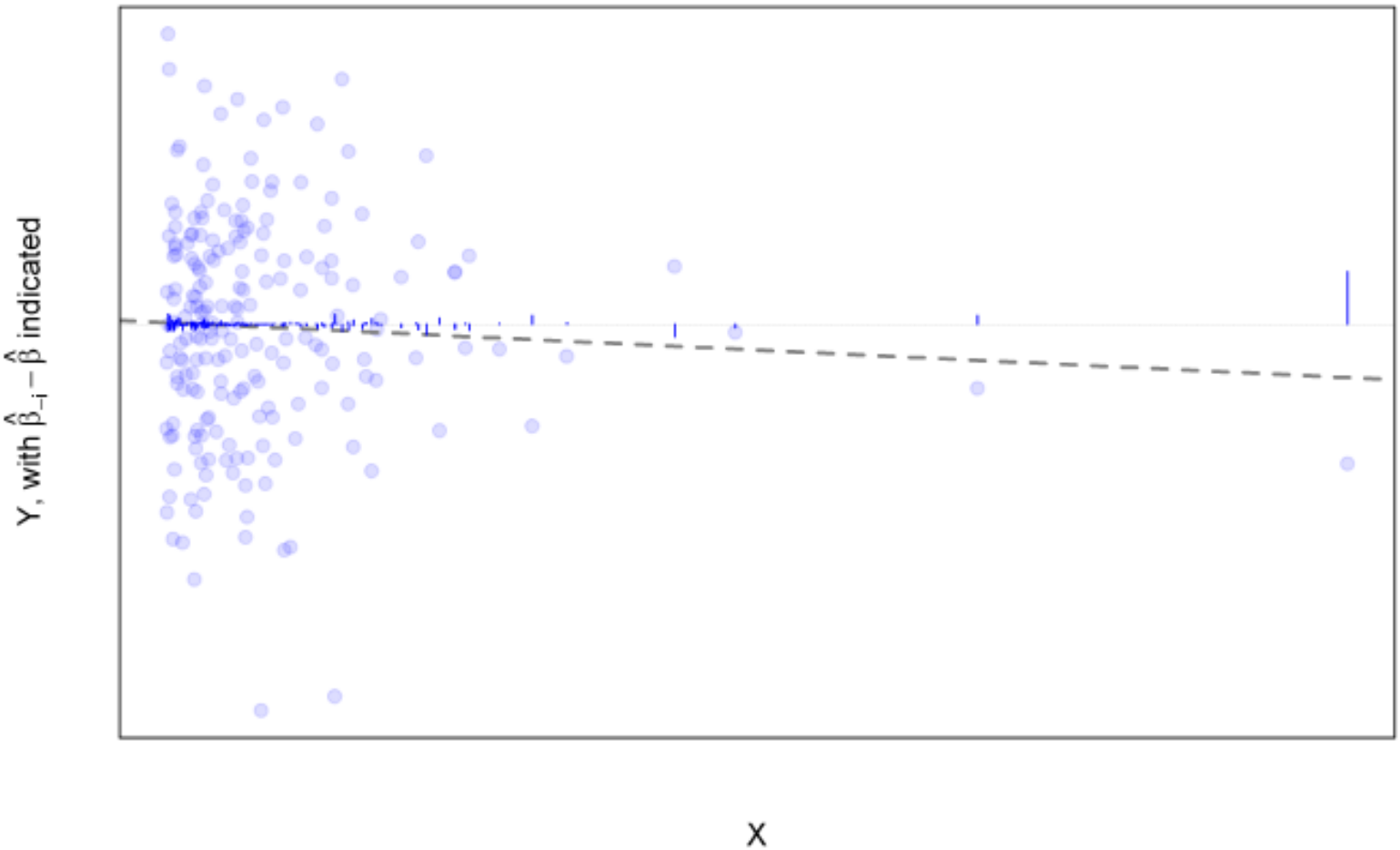
Model-checking: without the "model" bit†

It's likely you'd want to know about these:



Model-checking: without the "model" bit†

It's likely you'd want to know about these:



Model-checking: without the "model" bit†

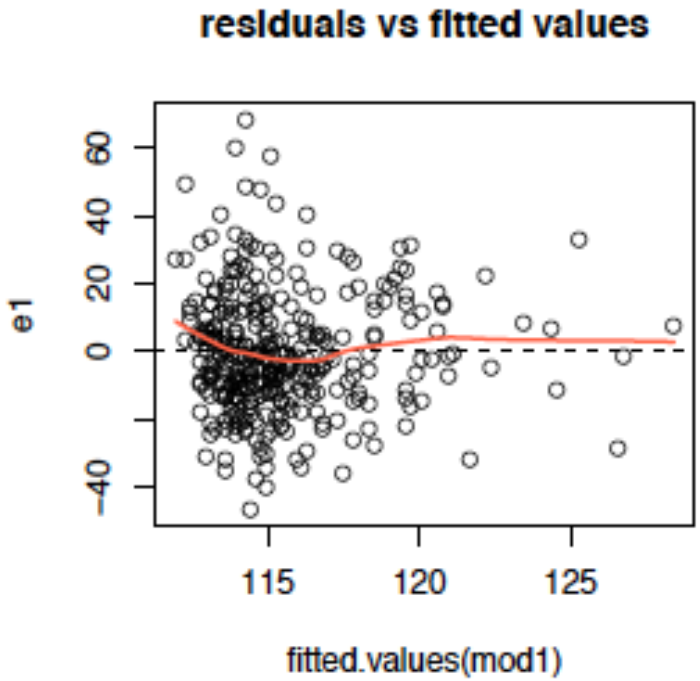
In nonparametric work, the lack of a specified variance structure for $\mathbf{Y}$ means it is essentially impossible to say how much variability residuals $\mathbf{Y} - g(\mathbf{X}\hat{\beta})$, "should" have by chance alone.

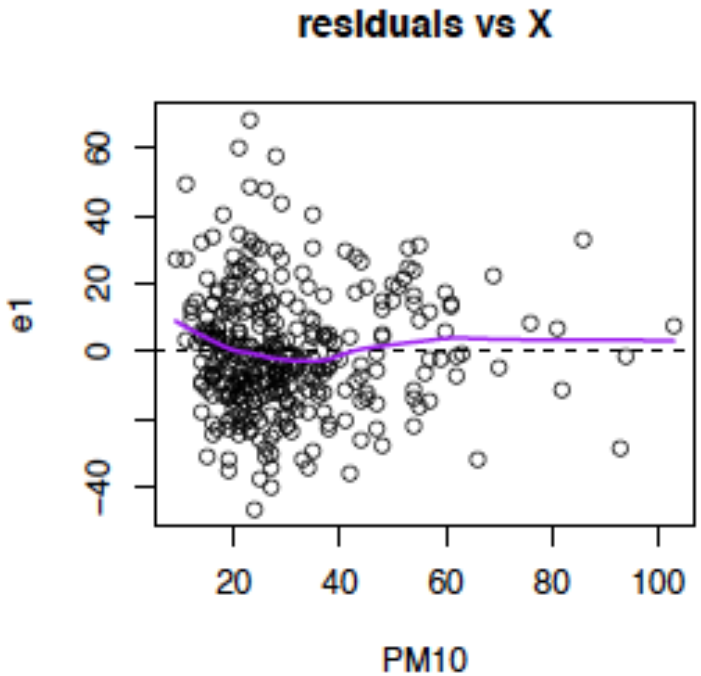
- When plotting residuals against fitted values and/or covariate values, use `lowess( ... iter=0)`
- Compare results from deletion diagnostics (leave-one-out-checks) to the original $\hat{\beta}$ and its estimated standard error - are the changes big, relative to how precisely we can specify β?
- For binary data, just plotting $\mathbf{Y} - g(\mathbf{X}\hat{\beta})$ versus fitted values is brain-dead [7]. However, a smoothed line (or something similar) through those points may have value.

[7] ...even though your software may do it

Model-checking: without the "model" bit†

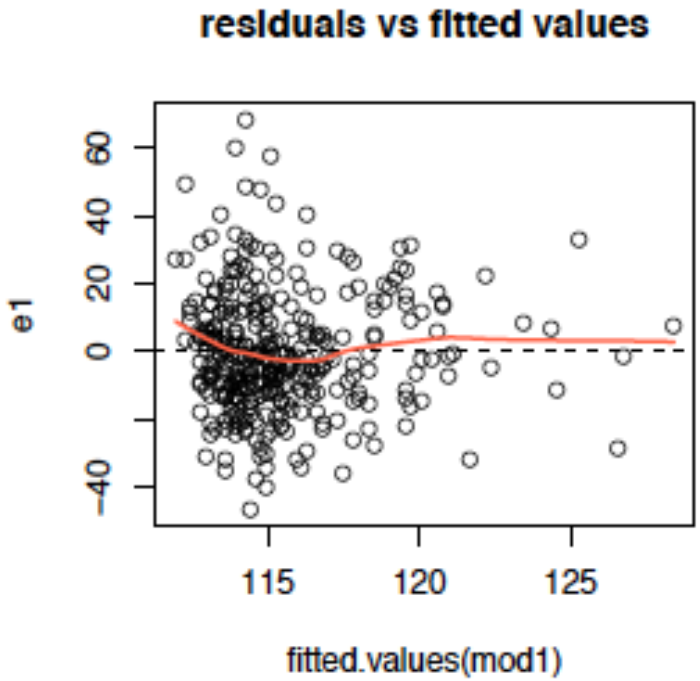
```
> mod1 <- glm(Deaths ~ PM10, data=dat, family=poisson)
> e1 <- residuals(mod1, type="response") # for types, see ?residuals.glm
> lowess1 <- lowess(fitted.values(mod1), e1, iter=0)
```

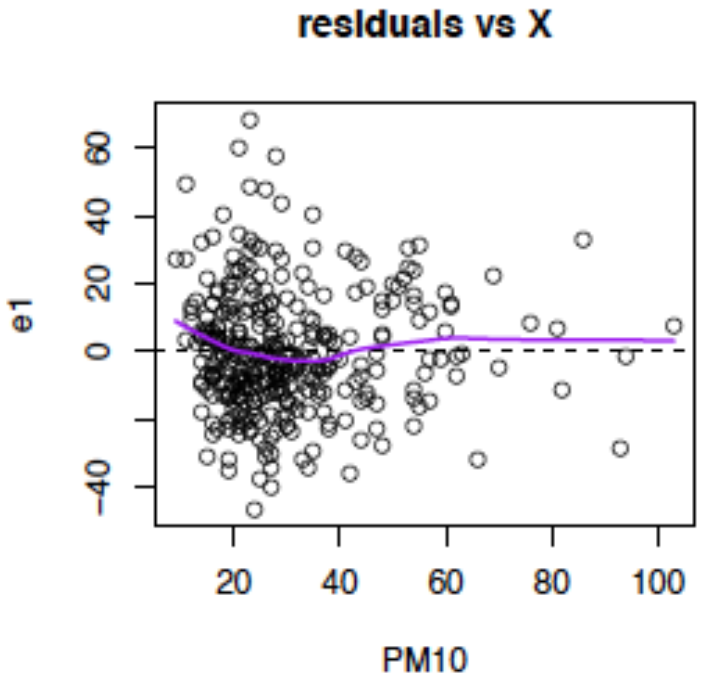




Model-checking: without the "model" bit†

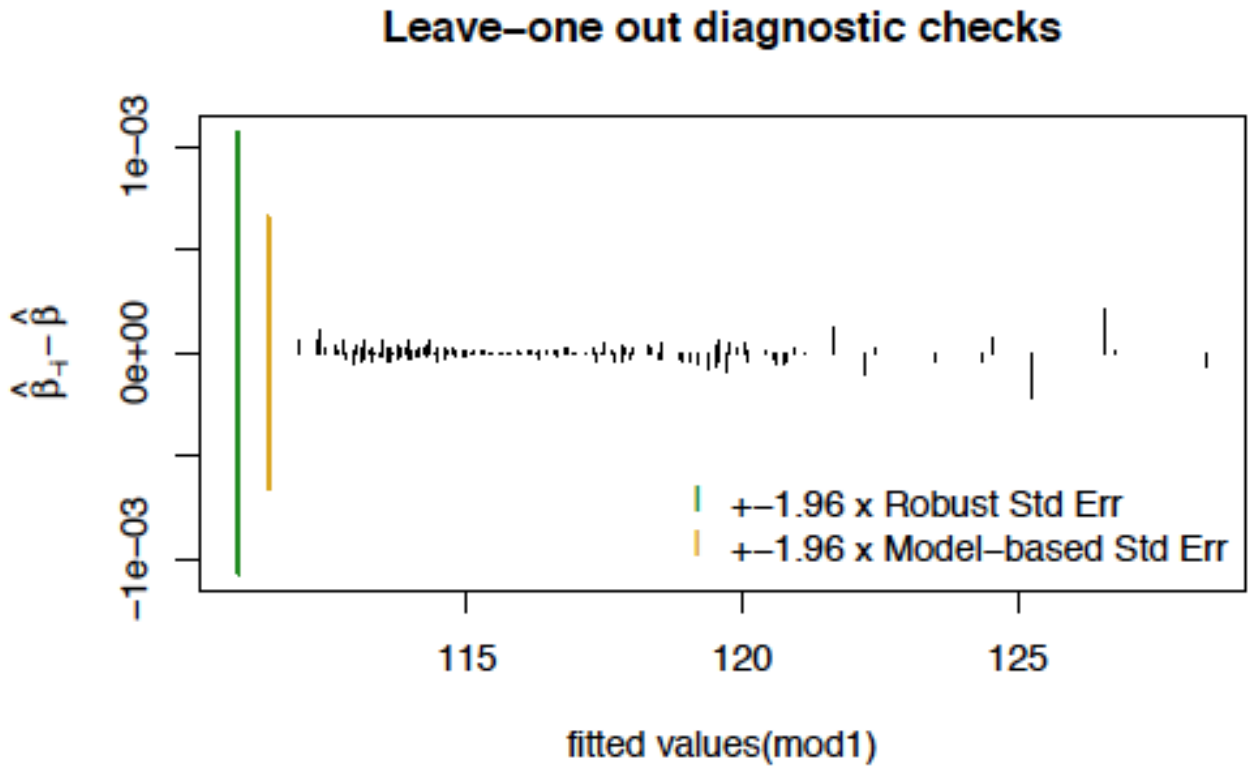
```
> mod1 <- glm(Deaths ~ PM10, data=dat, family=poisson)
> e1 <- residuals(mod1, type="response") # for types, see ?residuals.glm
> lowess1 <- lowess(fitted.values(mod1), e1, iter=0)
```





Model-checking: without the "model" bit†

```
beta.subi <- sapply(1:dim(dat)[1], function(i){
coef(glm(Deaths ~ PM10, data=dat[-i,], family=poisson))
}) # see ?dfbeta for a quick approximation
plot( x=fitted(mod1), y=(beta.subi[2,]-coef(mod1)[2])[is.na(dat$PM10)], type="h" )
```



Model-checking: first two moments†

In semiparametric work, we have more motivation to use Pearson residuals

$$r_i = \frac{Y_i - g(x_i^T \hat{\beta})}{\sqrt{\hat{\mathbf{V}}_i}}$$

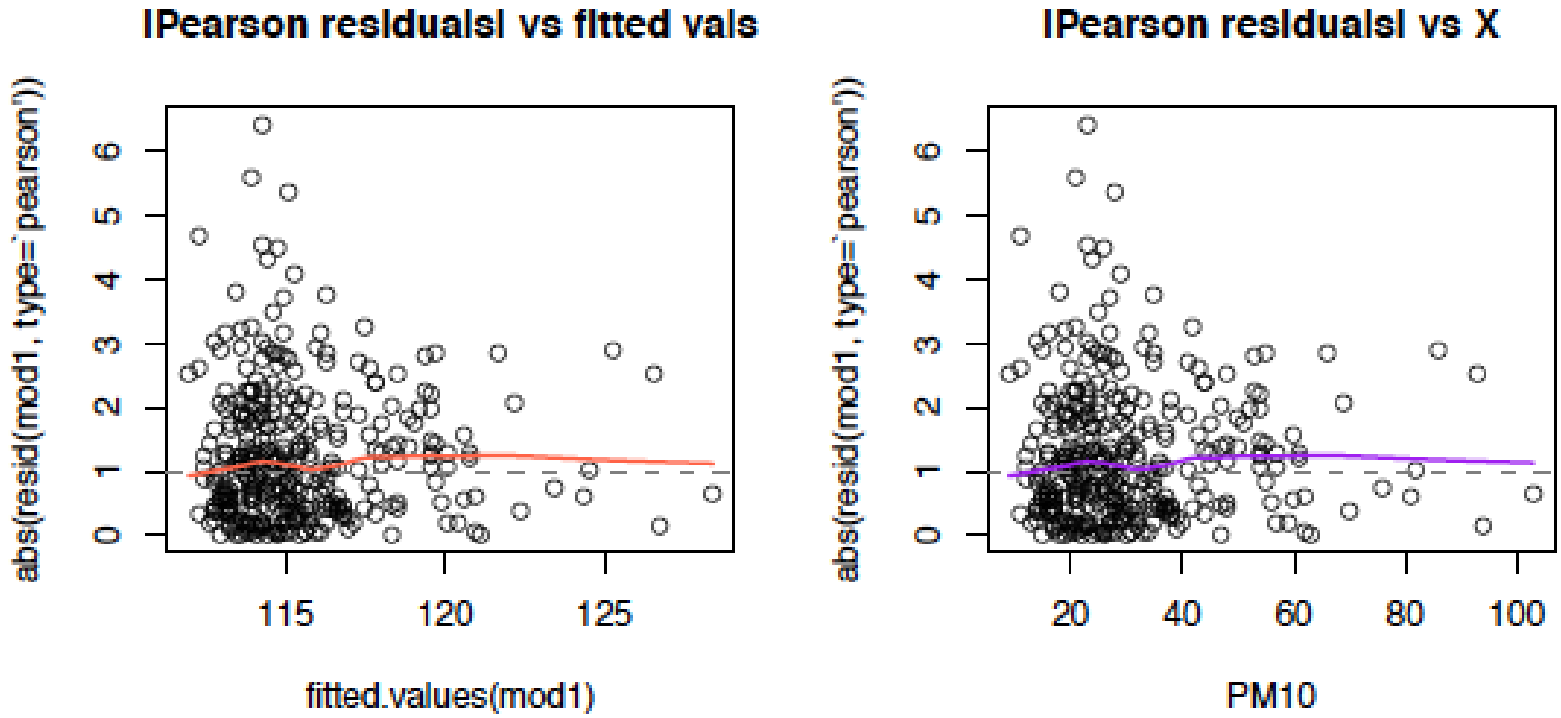
where $\hat{V}_i$ is the fitted value of the variance of Y_i.

- The distribution of these should end up roughly like $N(0, 1)$ if the mean and variance are correctly specified.
- Exact Normality of $\{r_i\}$ is not plausible (or even possible sometimes) - but recall that we didn't need it for inference anyway
- To assess the variance model, plot squared residuals - against fitted values and covariates

See also standardized residuals - which account for leverage and studentized residuals where the denominator is re-computed leaving out the data point in question.

Model-checking: first two moments†

Plots and smoothers of $|r_i|$ against x_i for the London PM10 data - using model-based standard errors (use of r_i^2 is also sane)



What's going on? What would happen using QL methods?

Model-checking: (gross Deviance)†

In parametric work, a formal frequentist assessment of model validity uses the "deviance"

$$D = -2\left[\ell(\hat{\theta}) - \ell(\tilde{\theta})\right]$$

where $\tilde{\theta}$ denotes fitting each data point Y_i with its own parameter, a so-called "saturated-model". In GLMs with "scale" parameter ϕ, asymptotically we find that:

$$D \sim \phi\chi^2_{n-p}$$

where $n - p$ is known as the "residual degrees of freedom" . (D is sometimes known as the "residual deviance")

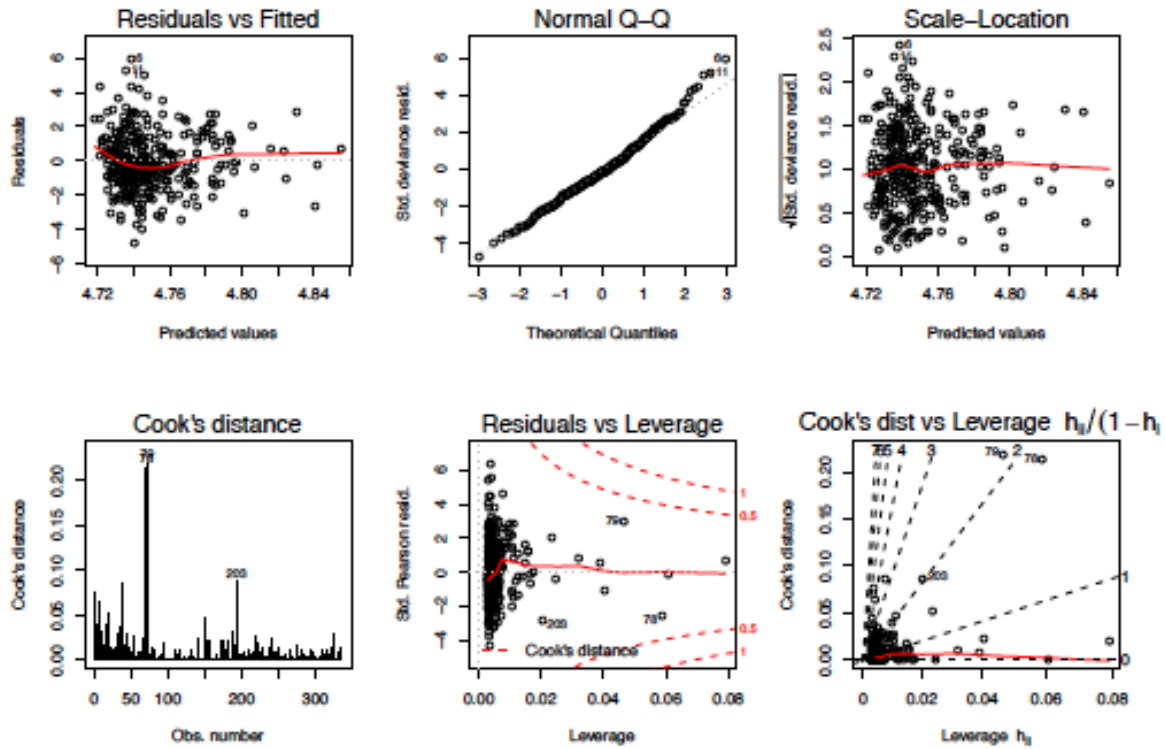
```
>mod1 <- glm(Deaths ~ PM10, data=dat, family=poisson)
>summary(mod1) # produces a lot of output, including;
(Dispersion parameter for poisson family taken to be 1)
Residual deviance:  908.65 on 333 degrees of freedom
>with(dat, # same thing 'by hand';
2*sum(dpois(Deaths[!is.na(PM10)], Deaths[!is.na(PM10)], log=T)-dpois(Deaths[!is.na(PM10)],
fitted(mod1), log=T)))
[1] 908.653
> qchisq(0.95, df=333)
[1] 376.5550
```

Model-checking: (gross) Deviance†

- On its own, deviance tells you nothing about what's going wrong. Is the mean wrong? Is it the variance? Or something else? (Use of deviance alone is not recommended)
- With large samples, large deviances can result from models which are "wrong but useful", e.g. a model with the right mean, but which gives slightly conservative inference
- Deviance-based residuals are also available
- If your different fitted models have different n (perhaps due to missing values being dropped) then comparing their deviances produces garbage
- "Analysis of Deviance" (which `anova` produces) compares deviances for a sequence of models. Be aware it estimates the dispersion parameter ϕ from the "biggest" model

Model-checking: (gross) Deviance†

Out put from `plot.glm(which=1:6)` uses all these ideas



Model-checking: a deviant example†

- In practice it is alarmingly easy to get deviances and likelihood ratio statistics (and Bayes factors) wrong by a factor of -1. Doing so is disastrous - for you and the science
- Almost always, your analysis involves parameter estimates and intervals, that will tell almost the same story as the tests. If your testing and estimation results do not "match" closely, be suspicious of coding errors
- This idea is not 100% guaranteed - particularly when contrasting 95% intervals on several paramters with one multi-df test -but it's sane
- Taking more than one attempt to get the code right for one statistical test does not incur a multiple-testing penalty - because you only do one statistical test.

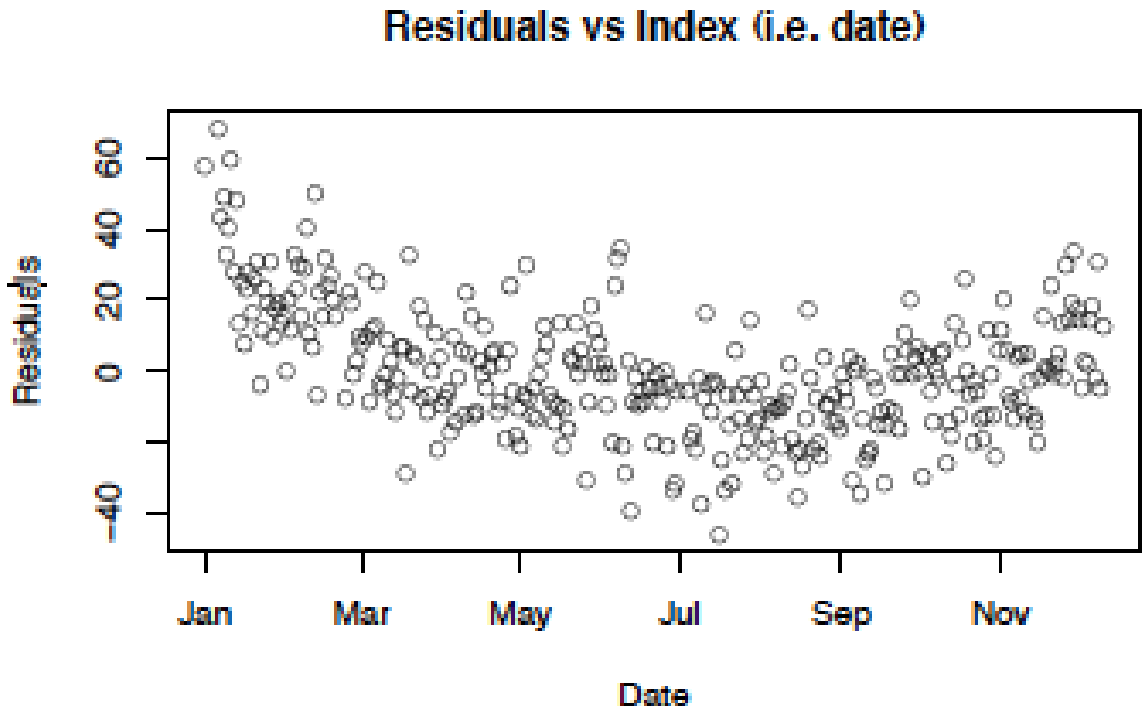
Model-checking: got independence?†

In many applications, the assumed independence of $Y_i|X_i$ is reasonable. But checking independence of the Y_i (or of the $Y_i|X_i = x_i$) has a role - strong dependence can severely invalidate confidence interval coverage - see later in the course.

- Plot e_i against time, or index number i
- (Better) plot e_i versus e_{i-1} for $i = 2, \ldots, n$. Independent residuals should produce a plot with a random scatter of points
- Plot a "correlogram" of the correlation of pairs of e_i versus their "lag" (use `acf()` in `R`)

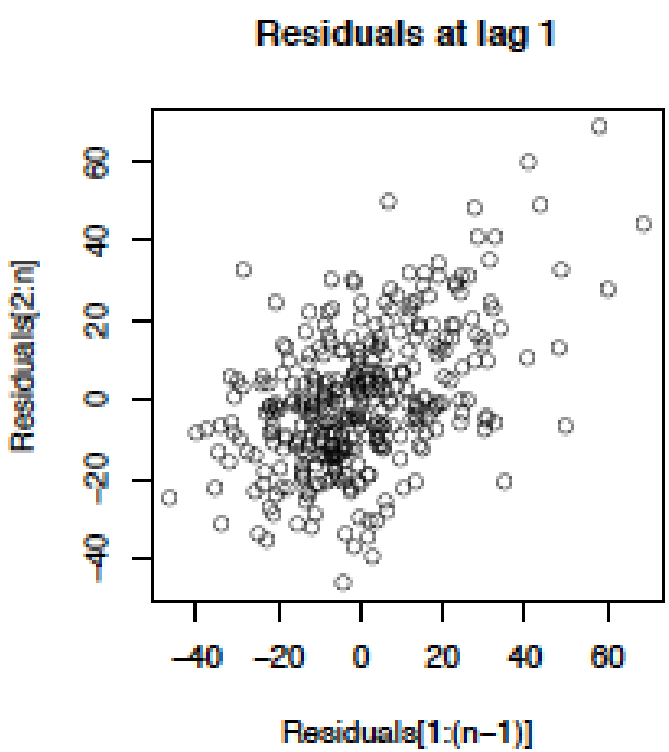
Examples follow with the London PM10 data, from a log-linear regression of Deaths on PM10.

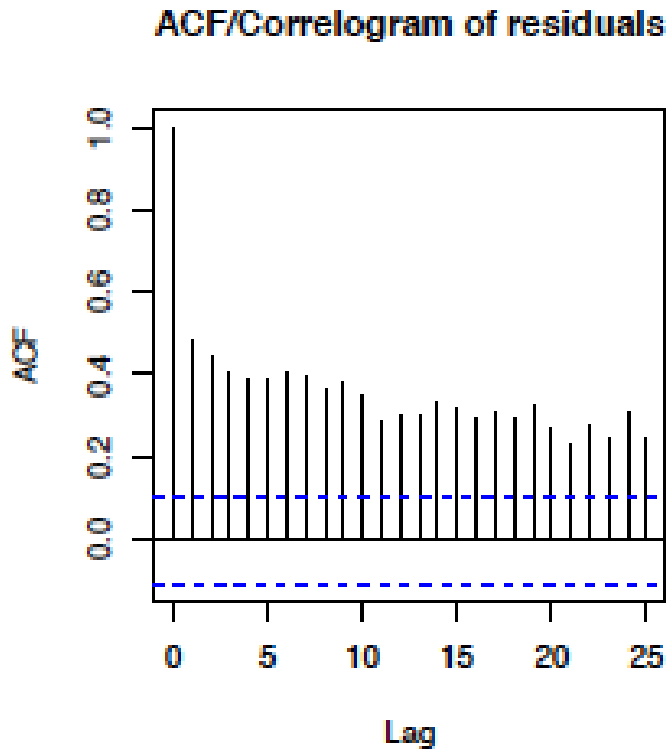
Model-checking: got independence?†



- What's going on?
- Note that this only "works" because the data is stored chronologically

Model-checking: got independence?†





- while these plots are less-intuitive, they can be more convincing; there are tests associated with the ACF plot

Model-checking: summary†

Most important points:

- Checking validity of various sorts can be a sane and useful part of analysis
- Say what you did, and why
- Similar checks can be used under different levels of assumptions
- Plots on a scale related to β will be easier to think about scientifically - though may not have an obvious statistical calibration
- Be aware of over-interpreting plots, some data point has to be most extreme, on any scale.

Model-checking: other notes†

- Checking that you have a simple random sample is essentially impossible, without strong assumptions - because the available data always look like a simple random sample from the population of those who end up in available data sets
- Checking finite-moment conditions is equally squirrely (...expect your data to have finite moments)
- "Leave one out" diagnostics can suggest whether asymptotic approximations are decent $\hat{\beta}_{n+1} \gg, \ll \hat{\beta}_n$ doesn't happen if $n \approx \infty$.
- Not all statisticians know about/believe in nonparametric interpretations. Referees tend to think you're assuming something stronger - so write carefully, and precisely.

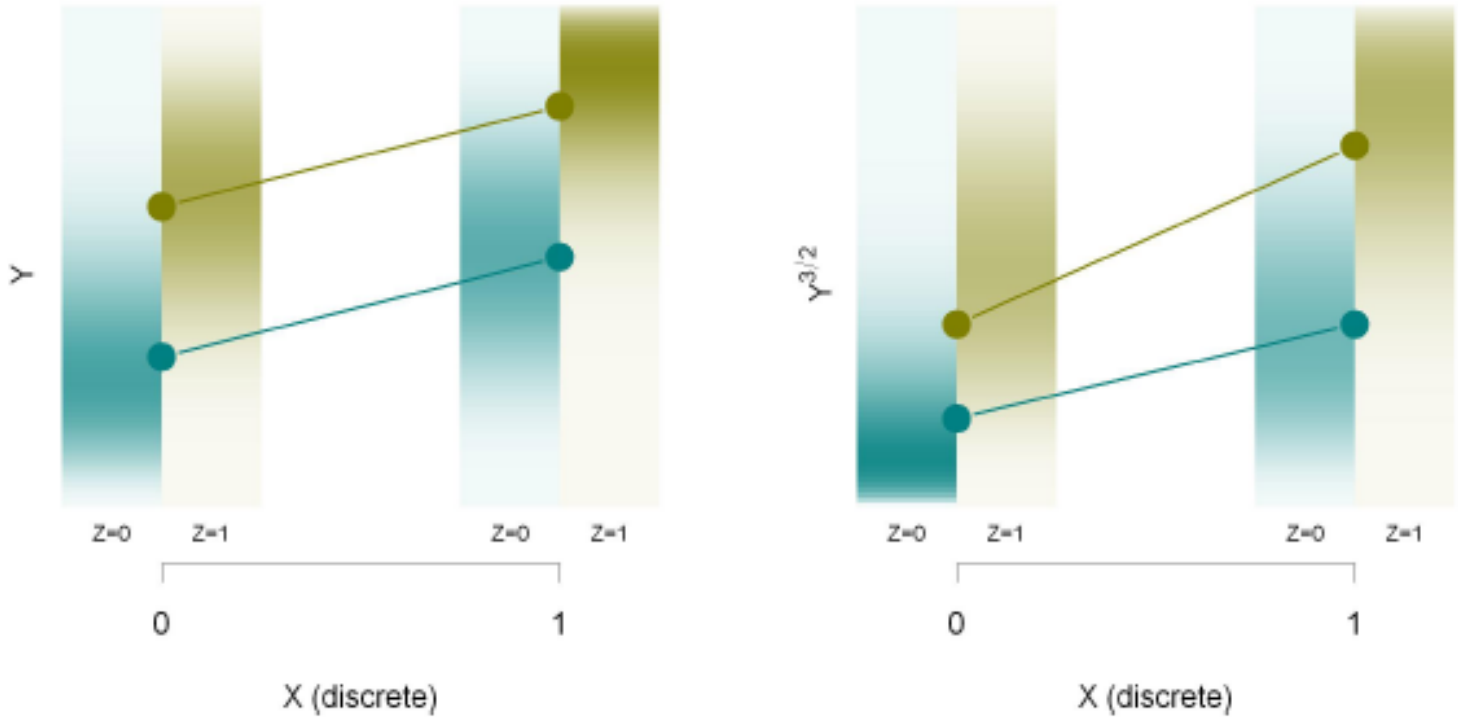
Interactions†

Getting the language right is hard for interactions:
$\mathbb{E}[Y|X = x, Z = z] = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$

- For every 1-unit difference in Z, β_3 tells you the difference in difference in expected Y per 1 unit difference in X' (!) An engineer might call this $\Delta\left(\frac{\Delta Y}{\Delta X}/\Delta Z\right)$
- β_3 tells you how the $Y - X$ slope differs, per unit difference in Z
- This is a subtle parameter! Expect to have less power than for main effects
- Note β_3 also tells you about the $Y - Z$ slope, over different X - describe whichever is simpler/more relevant.
- Also, e.g. β_2 tells you about the $Y - Z$ slope at $X = 0$. With centered X, this need not be an insane parameter

Interactions†

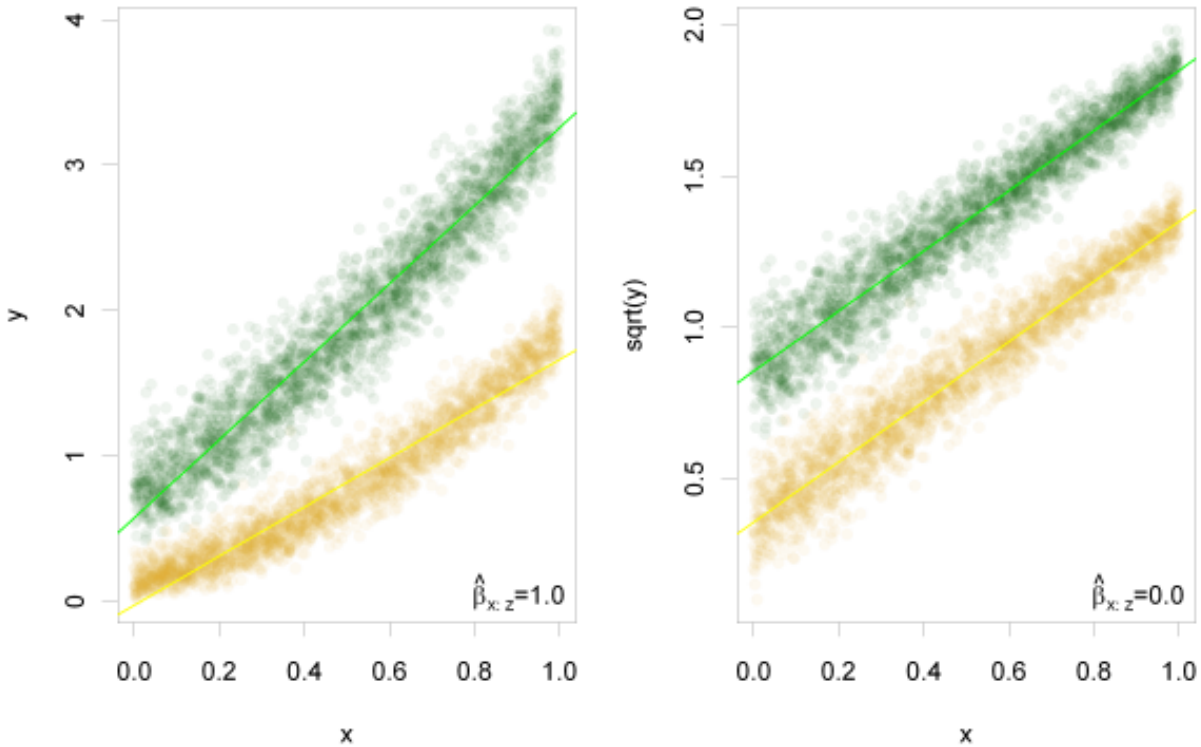
Interactions are squirrelly! They depend on the Y scale



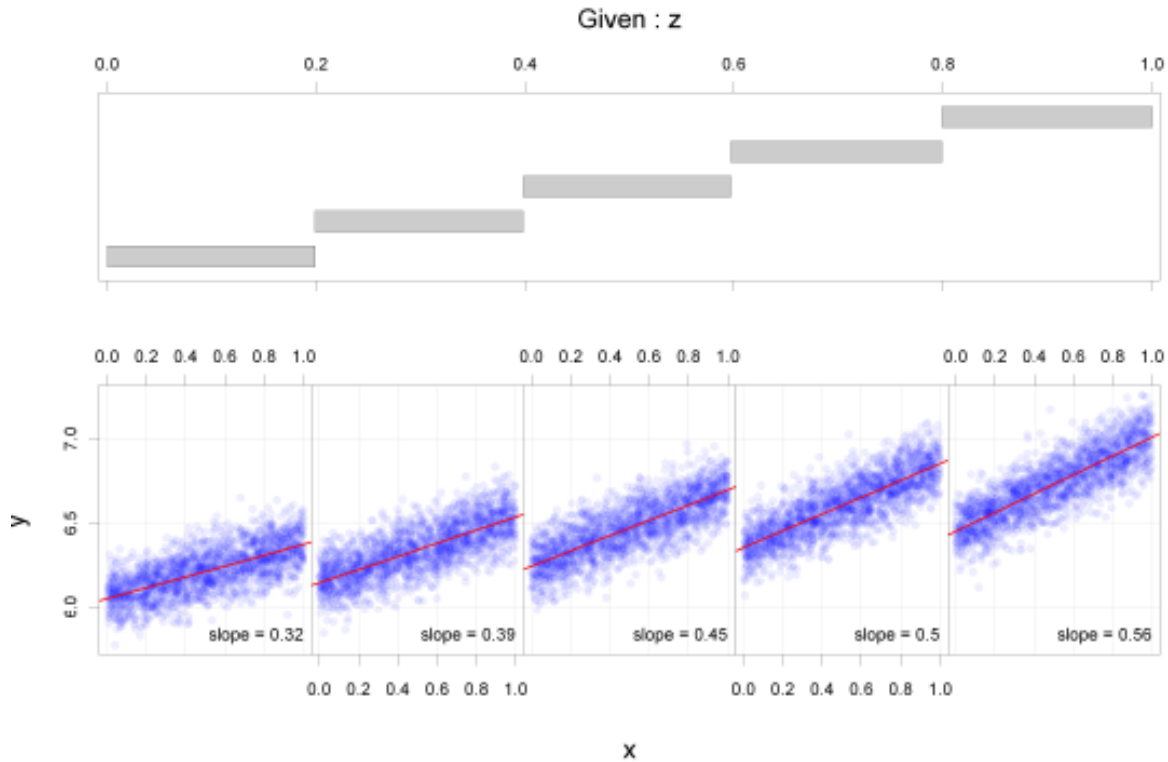
(Note that mean model-validity is not an issue here)

Interactions†

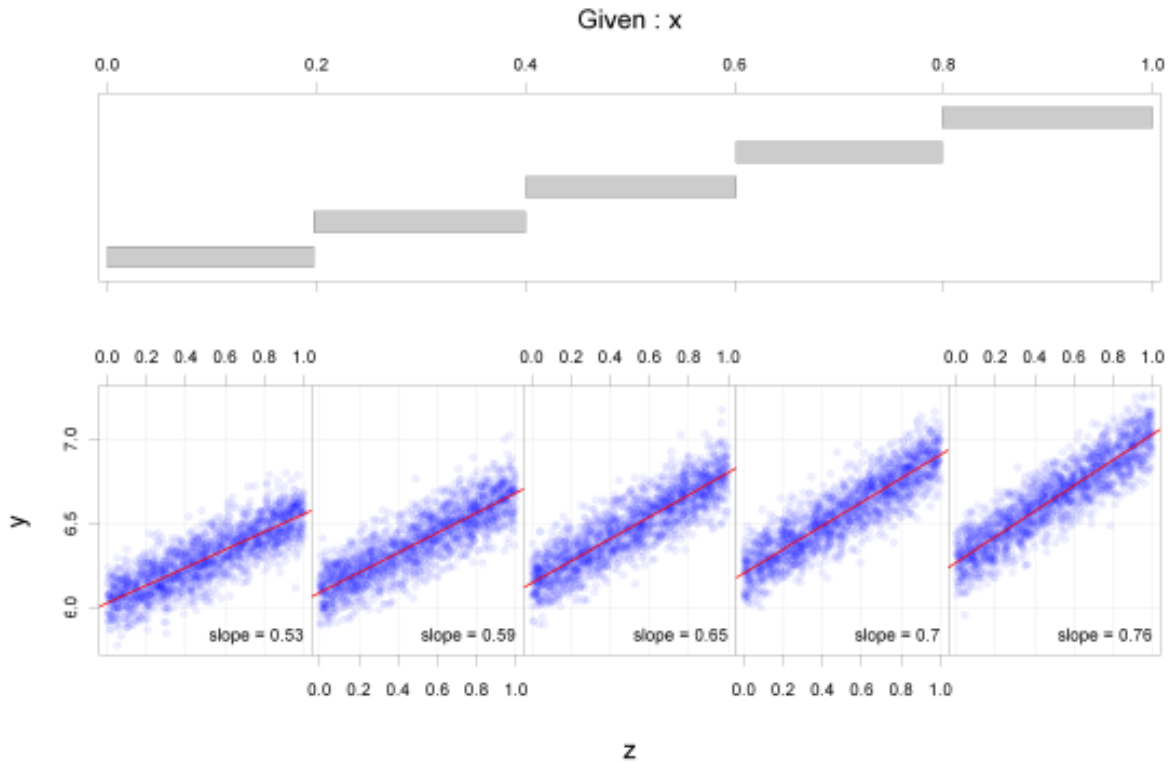
Some data, loosely based on the crabs example



Another set of data, with X and Z continuous

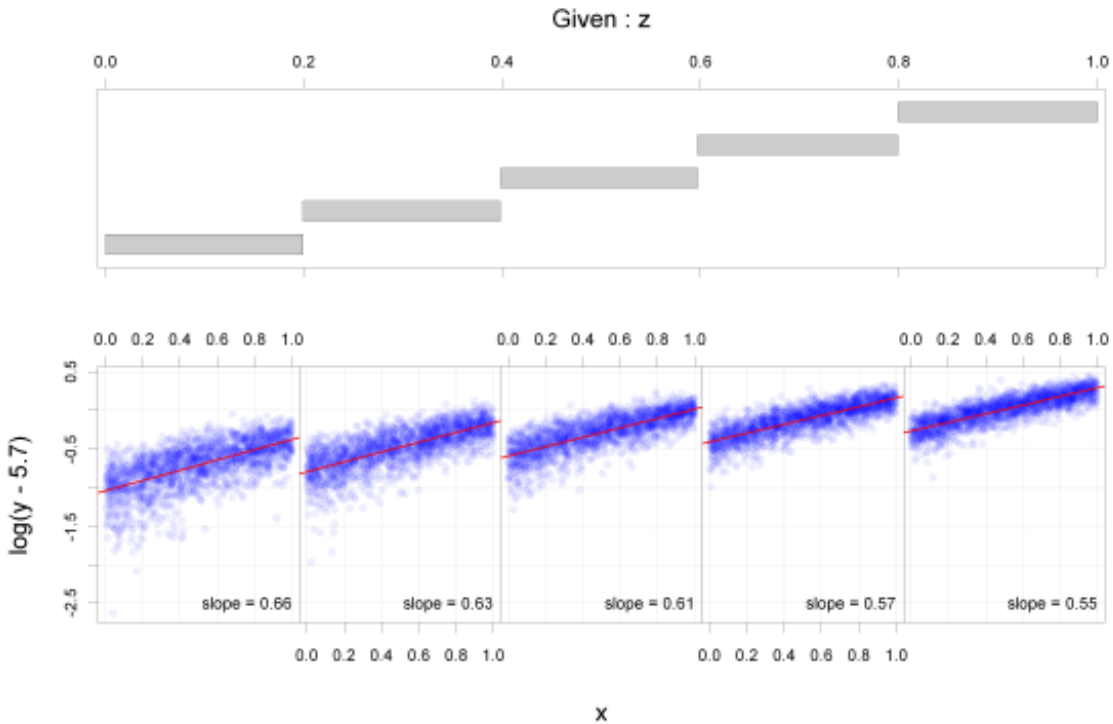


Another set of data, with X and Z continuous



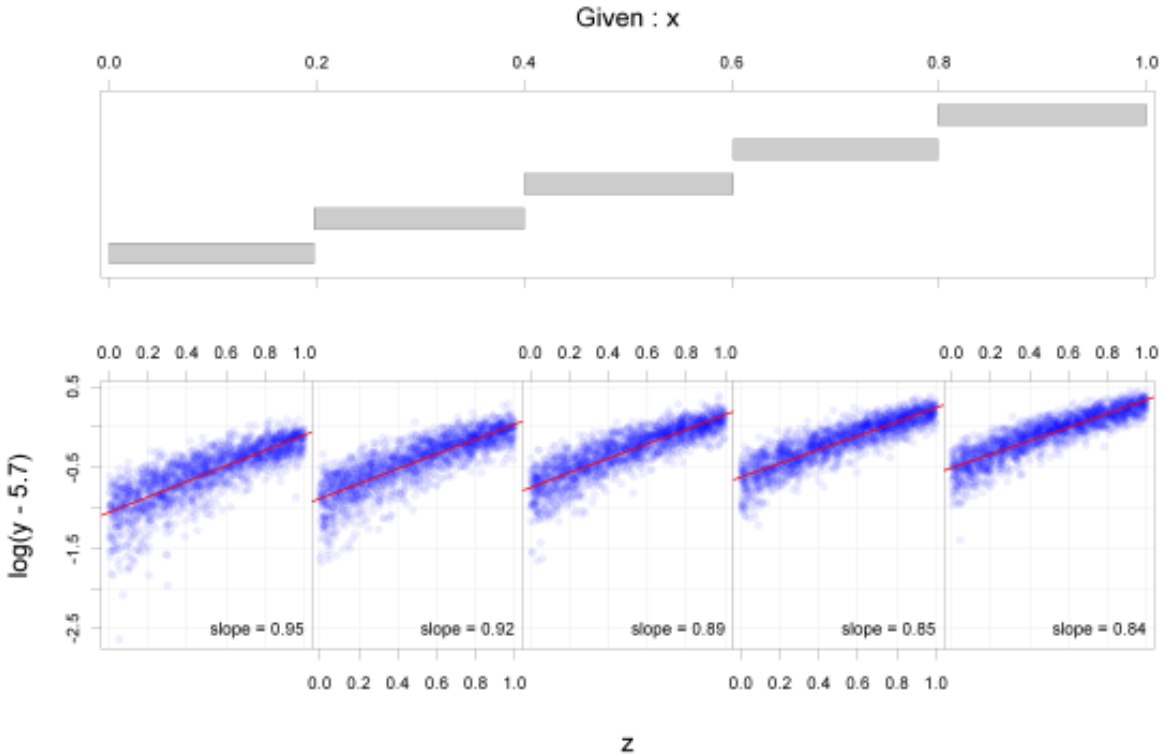
Interactions†

... where transforming Y "flips" the interaction



Interactions†

... where transforming Y "flips" the interaction



Interactions†

- Although you may see an interaction fitting `y ~ x*z`, it may be absent for e.g. `y ~ x*z, link = ‘‘log’’` - even for infinite n. (This is a “quantitative” or “removable” interaction)
- “Effect modification” can be scientifically “real”, but statistically removeable - or bogus and statistically “findable” (!)
- If (say) Z is binary and “the lines cross” the interaction is “qualitative”, and not removable by transformations
- Referring to “deviation from linearity on the chosen scale” would be more appropriate but sounds feeble.
- Some scientific interpretations use terms like “interaction” to defined e.g. behavior you could never capture looking at $Y - Z$ and $Y - X$ relationships on their own. The statistical meaning is different, and weaker.