Biostatistics 140.754
Advanced Methods in Biostatistics IV

Jeffrey Leek

Assistant Professor
Department of Biostatistics
jleek@jhsph.edu

Lecture 6

Tip + Paper

Tip When doing a real data analysis with collaborators - try to get the "rawest" form of the data you can. This will help you out in a couple of different ways: (1) you will learn about the science, (2) you will be able to identify problems that got smoothed over in creating the pretty data set, (3) you can make your analyses more reproducible. Sometimes, it saves a lot of time to start with the pretty data set - this is ok as long as you take the time to learn about what data processing happened to produce your pretty data set.

Paper of the Day: "Regulation of aging and age-related disease by daf-16 and heat shock factor"
http://www.molbio1.princeton.edu/labs/murphy/murphy_pubs/HsuMurphyKenyon.pdf

Outline For Today

- Log-linear models for tables
- Motivating EE with vector valued observations
- EE for vectors approach (leading toward GEE)

Survey Results

Response rate: 8/14 or 8/16
Paper of the day: 2/8 are reading it
Speed of class: Too fast (4), just right (4)
Reading assignments: Very helpful (1), Helpful (6), Not helpful -but interesting (1)

- I plan to keep giving papers - if nothing else just cause they fire me up.
- I think I may slow down just a notch - but not too much. If you need extra help, definitely email me.
- The questions on the reading assignments are meant to help you understand how to construct a paper

EE for univariate data

General form of the EE's [1], which we can interpret as fitting some line/surface to univariate data

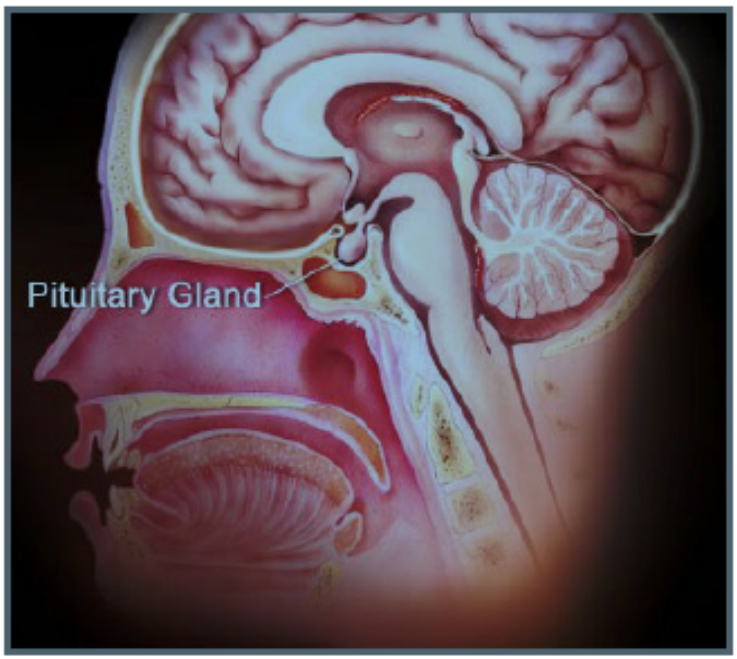
$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\frac{\partial g(\mathbf{X}_i^T \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right]^T w(\mathbf{X}_i^T \boldsymbol{\beta})(Y_i - g(\mathbf{X}_i^T \boldsymbol{\beta})) = \mathbf{0}$$

- Assumes Y_i, X_i is an independent sample
- Increasing assumptions can be applied - do we believe the mean model? do we have a model for the variability? do we have a model for the whole distribution?
- Interpretations change as we make different choices

[1] Formally, the class of linear unbiased estimating equations

Motivation: Dental Growth †

An example of renown - and distress, and misery. The distance from the pterygomaxillary fissure to the center of the pituitary gland is easily obtained, from xrays:



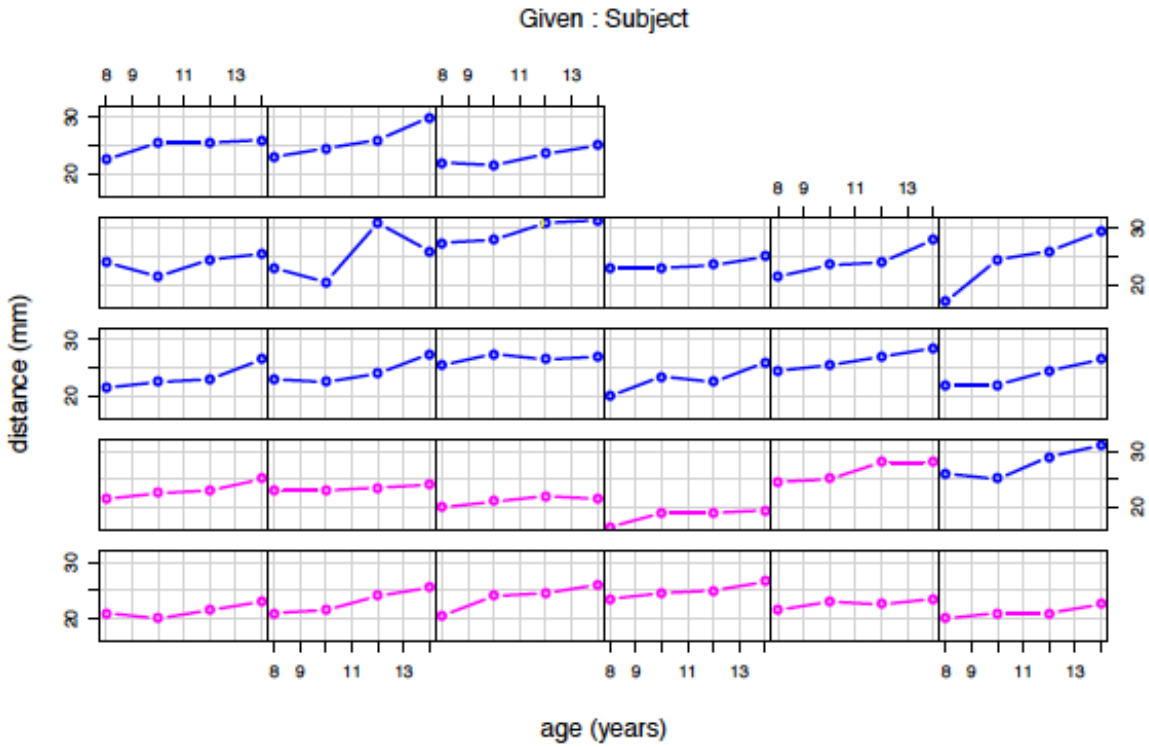
Motivation: Dental Growth †

It's a measure of growth, useful for e.g. orthodontists. Here is the data, in mm, for $n = 11$ girls and 16 boys.

Girls				Boys			
8 yrs	10 yrs	12 yrs	14 yrs	8 yrs	10 yrs	12 yrs	14 yrs
21	20	21.5	23	26	25	29	31
21	21.5	24	25.5	21.5	22.5	23	26.5
20.5	24	24.5	26	23	22.5	24	27.5
23.5	24.5	25	26.5	25.5	27.5	26.5	27
21.5	23	22.5	23.5	20	23.5	22.5	26
20	21	21	22.5	24.5	25.5	27	28.5
21.5	22.5	23	25	22	22	24.5	26.5
23	23	23.5	24	24	21.5	24.5	25.5
20	21	22	21.5	23	20.5	31	26
16.5	19	19	19.5	27.5	28	31	31.5
24.5	25	28	28	23	23	23.5	25
				21.5	23.5	24	28
				17	24.5	26	29.5
				22.5	25.5	25.5	26
				23	24.5	26	30
				22	21.5	23.5	25

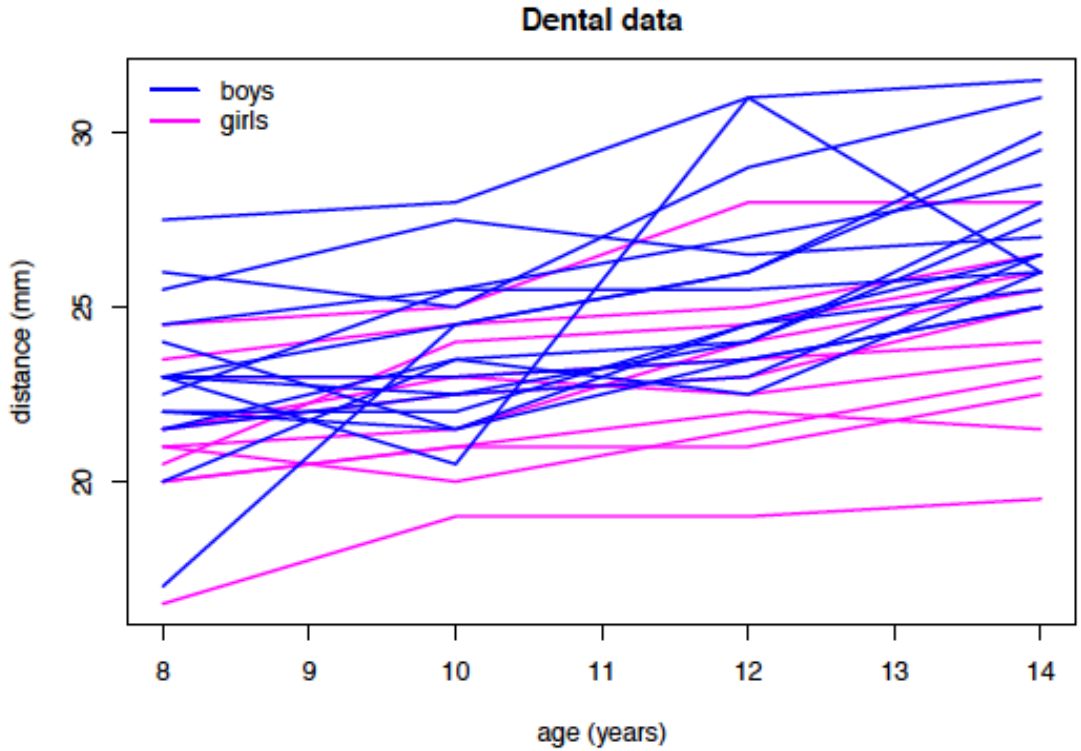
Motivation: Dental Growth †

Displaying the data with coplot - `coplot(distance ~ age | Subject)`



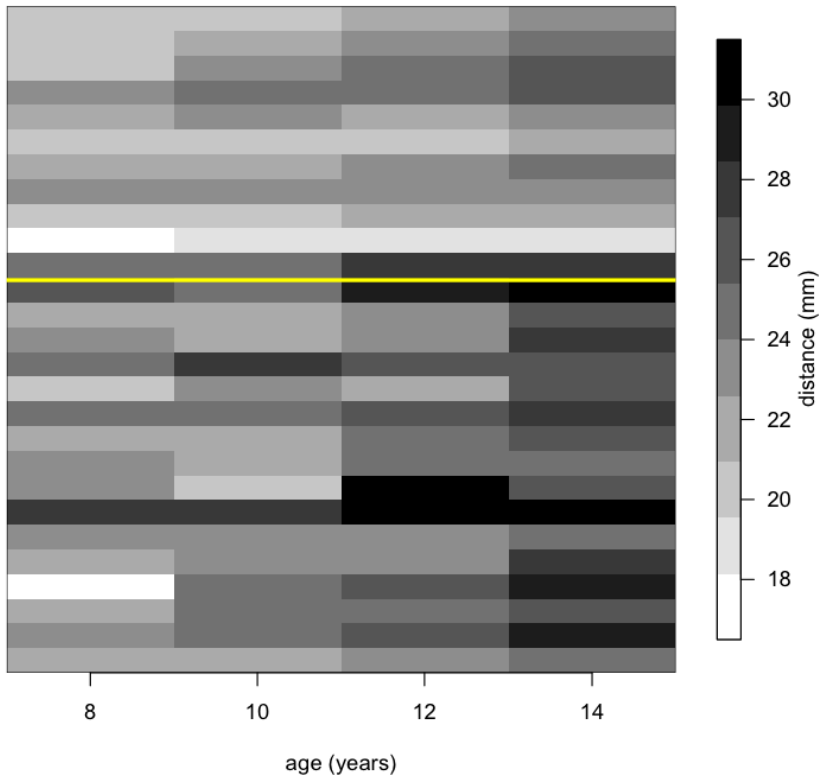
Motivation: Dental Growth †

`matplot()` makes "spaghetti plots" - for big n use transparent colors



Motivation: Dental Growth †

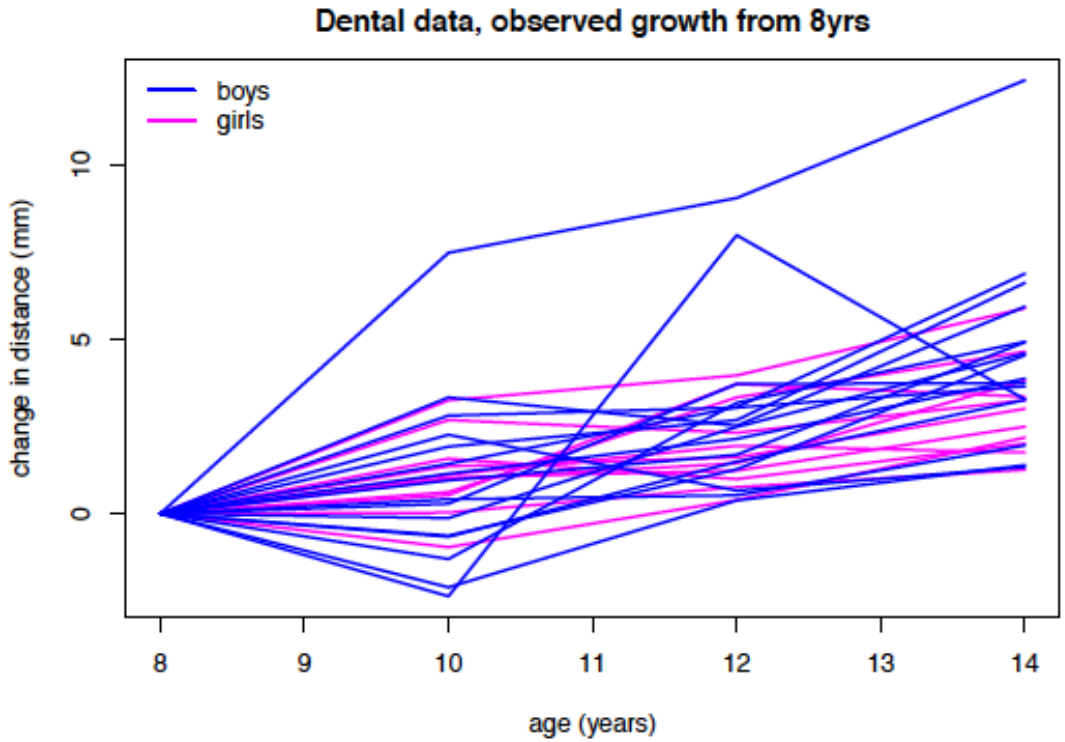
Lasagna plots!! Invented here by B. Swihart + B. Caffo ("Lasagna plots: a saucy alternative to spaghetti plots") [2]



[2] http://www.bepress.com/jhubiostat/paper141/

Motivation: Dental Growth †

Growth from 8 years old (slightly jittered, at 10+ years old)



Motivation: Dental Growth †

Here are some scientific questions - match them to the graphs

- How much "longer" are the distances, comparing 12 year olds and 10 year olds?
- What shape is the "average" curve?
- Do boys or girls grow faster from age 8?
- Do boys or girls grow faster?
- What distance do you expect your patient (a boy, age 10, currently with $d = 23$mm) to attain by age 14?
- Is variability in d better explained by within-child or between-child differences?

All of these are reasonable questions - but they require different analyses, even given the same data.

Motivation: Dental Growth †

We distinguish two important types of questions:

- What ist he average, e.g. growth curve? **[marginal]**
- For a specific child, what is the growth curve? **[conditional]**

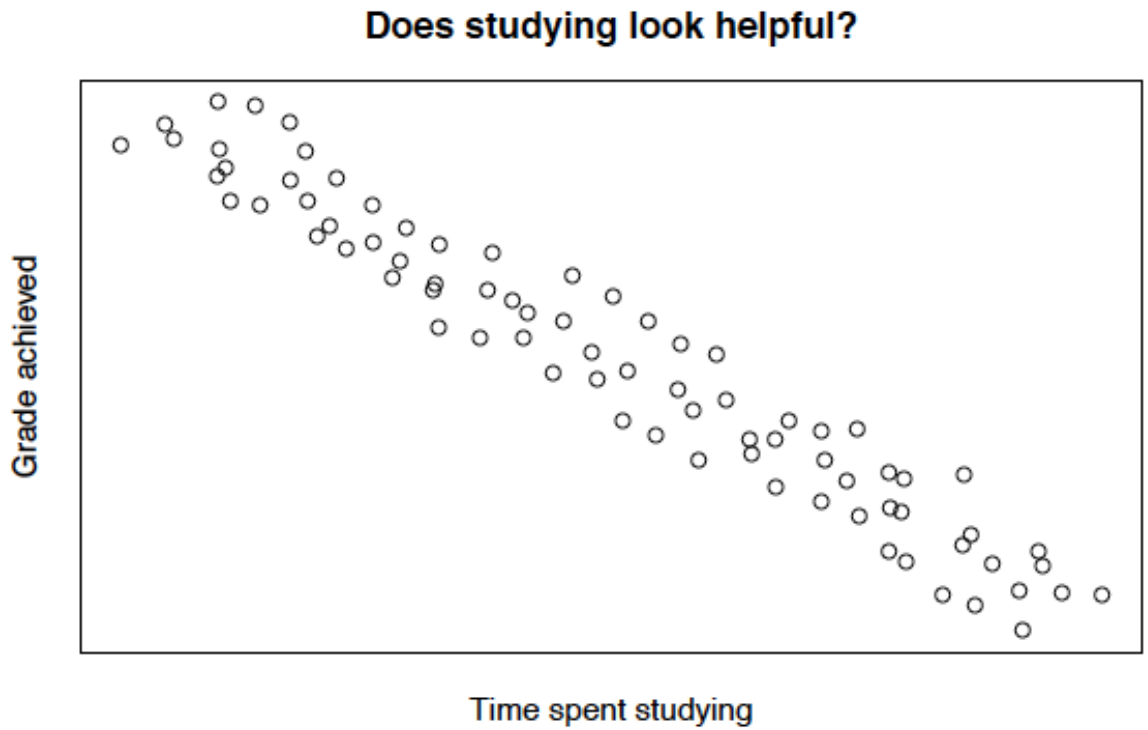
The former is a question about whole populations; the latter is about particular subjects (or types of subjects).

- Even for marginal inferences, it will be invalid to treat the data as $27 \times 4 = 108$ independent outcomes; we actually have 27 sets of 4 correlated outcomes - your values of d_8 and d_{14} will be closer to each other than to my d_8, d_{14}[3]
- This holds even if you and I are the same sex - though in e.g. males only, our data might be "closer" than otherwise
- For either question, naïvely assuming independence can easily lead to underestimates of standard errors (anti-conservative) though over-estimation is also possible (loss of power)

[3] Blame your parents - or mine

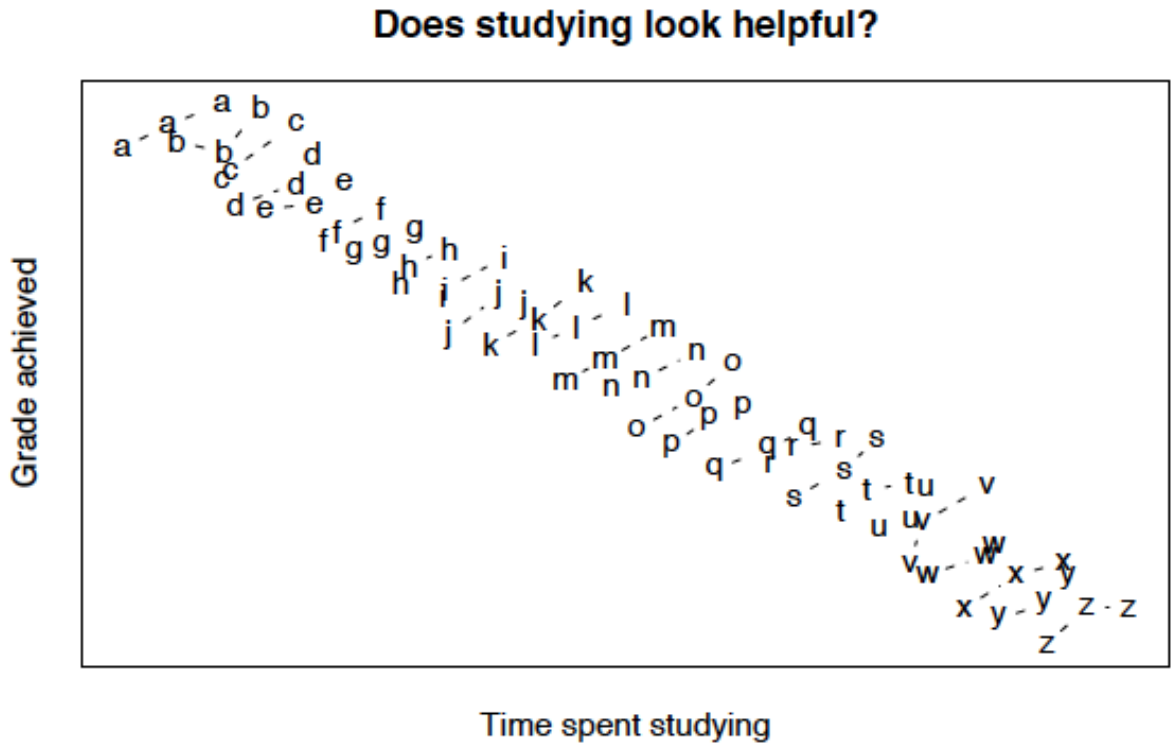
Motivation: Conditional and marginal †

Some made-up data, on a topic that concerns you



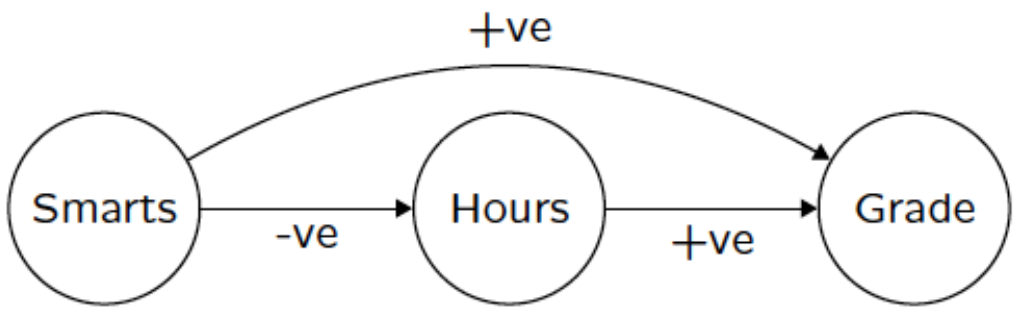
Motivation: Conditional and marginal †

Some made-up data, on a topic that concerns you



Motivation: Conditional and marginal †

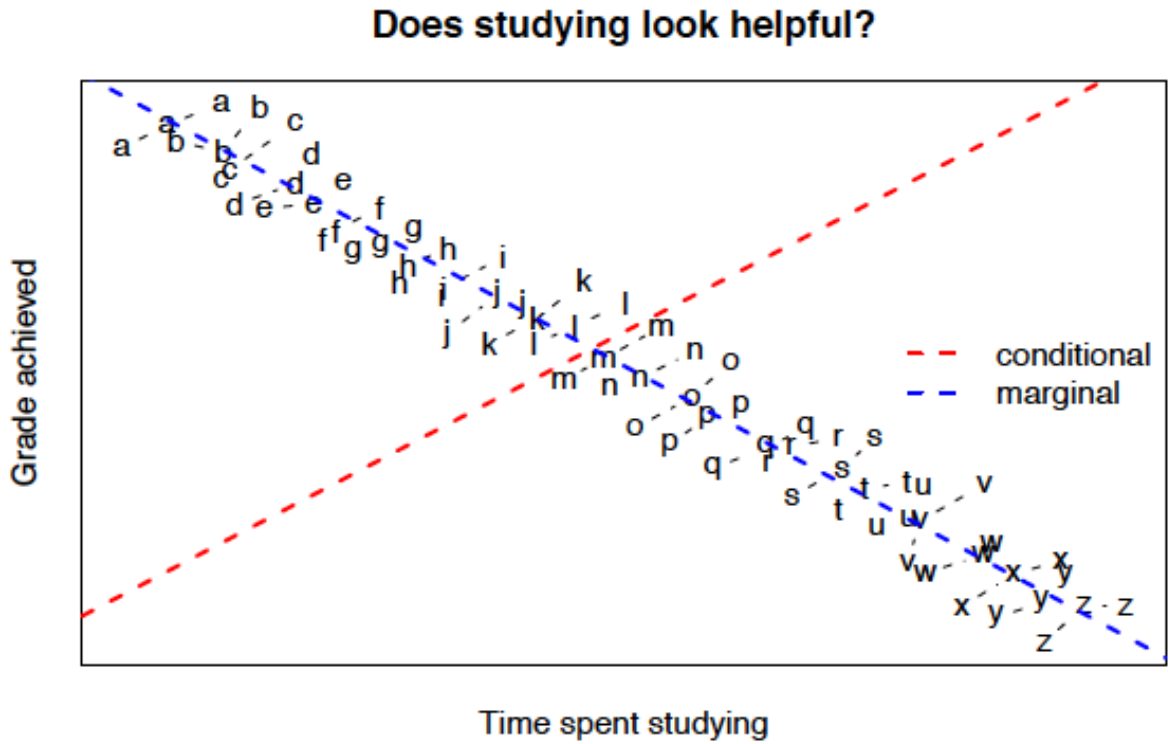
Plausibly here's what's going on:



- On its own, Hours **is** negatively associated with Grade - note I only said "associated"
- The causal version of this statement (Hours decreases Grade) is wrong - a.k.a confounding
- Keeping person fixed, Hours **is** positively associated with Grade
- Averaging over the whole population, Hours **is** negatively associated with grade.

Motivation: Conditional and marginal †

Two lines illustrating different parameters:



Motivation: Conditional and marginal †

- Marginal = "averaging over everything else"
- Conditional = ""keeping everything else fixed"

Marginal and conditional statements are **not the same**

- This is true regardless of whether you infer causality or not (so mistaking one for the other is not confounding)
- Both associations can be useful. The data will not tell you which is "right", because "right" depends on what you want to know. You must pick
- Recall non-collapsibility, for parameters defined via linear operations, start-specific and population-averaged versions *may* be numerically identical
- Expect debate: *Our experience over more than 20 years of modelling longitudinal data in social science and medical applications is that [...the...] conditional formulation has a scientific supremacy* - Crouchley and Davies, JRSSA 2001

Vector Outcomes†

Initially, we will describe estimating equations for a fitted line/plane alone; this approach will view within-cluster correlation as a **nuissance**, i.e. of (at most) secondary interest.

Nonparametric estimation of standard errors relies on simple random sampling of clusters from the superpopulation of clusters, i.e. of independence between clusters, and on sandwich-style asymptotic approximations.

We will then extend the (semi-parametric) quasi-likelihood methods to **Generalized Estimating Equations** (GEE)[4], where assumptions of mean models mean we can draw stronger inference - and where knowledge of within-cluster correlation may be exploited to achieve efficiency.

[4] and closely-related approaches

Vector Outcomes: notation†

Notation generalizes what we used for univariate outcomes:

$(Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T = Y_i$	Outcomes for cluster i
$(X_{i1}, X_{i2}, \ldots, X_{in_i})^T = X_i$	Length p covariate vectors for cluster i
X_{ijk}	Value of covariate k, at observation j, in cluster i
$\boldsymbol{\beta}$	$p \times 1$ vector of parameters

As before, associations between Y_{ij} and their respective X_{ij} can be identified by line-fitting operations [5], where the "lines" we fit are of the form

$$y_{\cdot j} = g(x_{\cdot j}^T \boldsymbol{\beta})$$

where g^{-1} is the familiar "link" function.

[5] formally, we're fitting multiple lines, or (equivalently) a surface)

Vector Outcomes: nonparametric†

How do we fit these lines? For now, consider estimating equations of the (familiar) form:

$$\hat{\boldsymbol{\beta}} : \sum_{i=1}^{n} \frac{\partial g(X_i^T \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} (Y_i - g(X_i^T \boldsymbol{\beta})) = \mathbf{0}_p$$

where, by convention, $g(\cdot)$ is applied component-wise. The matrix of derivatives is $p \times n_i$, and it multiplies a $n_i \times 1$ vector of residuals

- $\hat{\boldsymbol{\beta}}$ estimates the population quantity $\boldsymbol{\beta}$, a (for now) unweighted least-squares line summarizing the super-population **of all clusters**
- Assume (for now) that n_i, the number of observations per cluster is uninformative about $\boldsymbol{\beta}$, e.g. n_i is constant for all i.
- For the identity link, interpretations as weighted pairwise slopes exists but are not pursued here.

Vector Outcomes: nonparametric†

Nonparametric interpretations of these fitted lines follows similarly to EE

- The difference (or fold change/rate ratio/relative risk, or odds ratio) in Y **associated** with a 1-unit difference in X_{ijk}...
- ... linearly adjusting for all the other $X_{ijk'}$
- We are giving a simple summary, for the whole population, of how the expectation of Y differs with X; describing β as a linear/log-linear/logistic-linear "trend" captures this idea well.
- We are not claiming anything about $\Delta(Y)$ for specific differences in X, e.g. $X_{ijk} = 5.2$ versus 6.3.
- Defined this way, β is agnostic about whether $\Delta(Y)$ is a within-cluster or between-cluster comparison - we are treating them in the same way.

Vector Outcomes: nonparametric†

To illustrate these ideas we first consider some simulated data. We consider the situation where the true data-generating mechanism is:

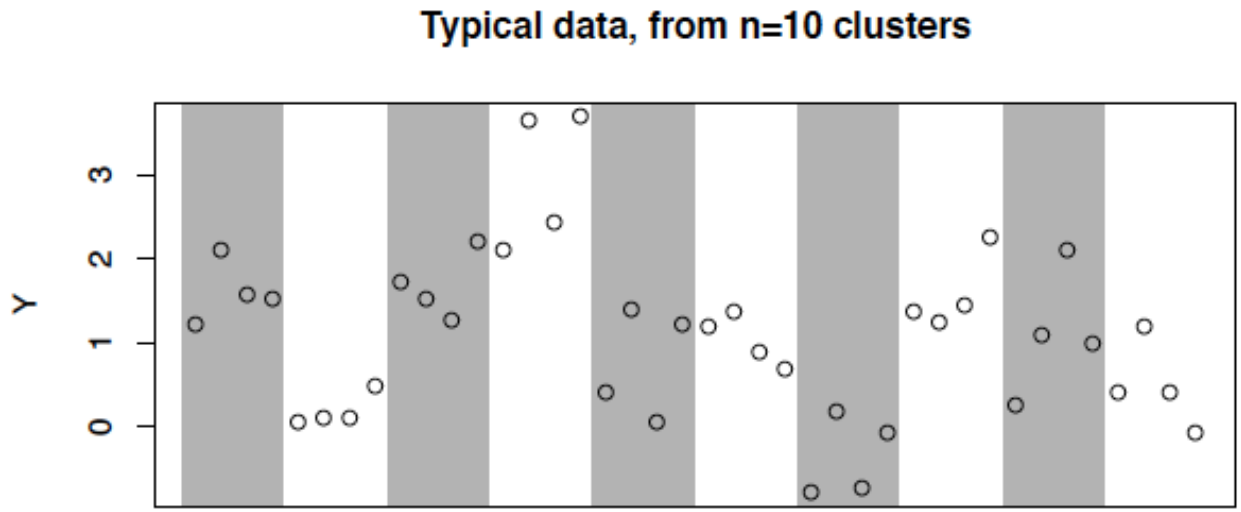
$$\begin{aligned} b_i &\sim N(0,1) \\ X_{ij} &\sim N(0,1) \\ Y_{ij}|\{X_{ij}=x, b_i=b\} &\sim N(b+\gamma_0+\gamma_1 x+\gamma_2 x^2, 0.5^2) \end{aligned}$$

for $1 \leq j \leq n_i$ and $1 \leq i \leq n$. Here we use $n_i = 4$ for all clusters, and $\gamma_0 = 1, \gamma_1 = 0.4, \gamma_2 = 0.1$.

This data generating mechanism (F) is an example of a linear mixed model - although our analysis will not assume we know the parametric form of F, i.e. no blinding light from a UFO told us any of the above, except independence of clusters.

Vector Outcomes: nonparametric†

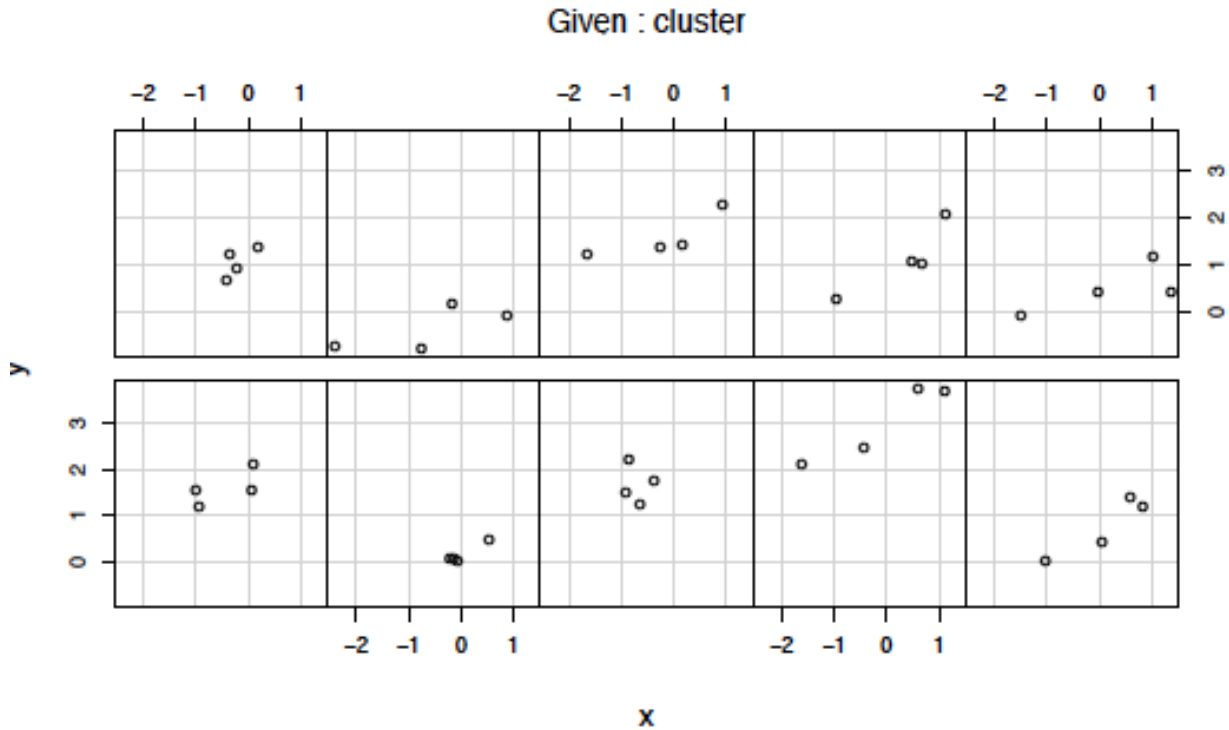
In our setup, *within*-cluster dependence is induced by "random effects" $\{b_i\}$



... outcomes within a cluster correlate more closely with each other than with outcomes in other clusters.

Vector Outcomes: nonparametric†

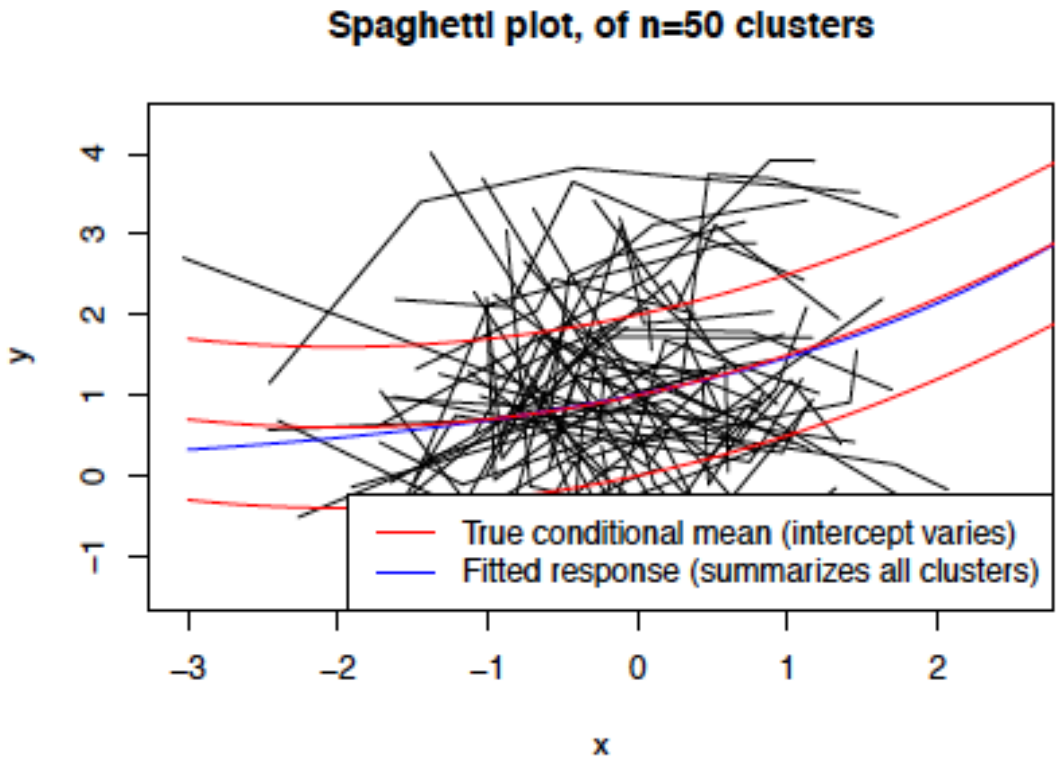
Within each cluster, the true mean of $Y|X = x$ is quadratic in x - although with four data points per cluster, this is not obvious by inspection:



Vector Outcomes: nonparametric†

We will fit a line of the form

$$y_{\cdot j} = \exp(\beta_0 + \beta_1 x_{\cdot j})$$



- i.e. we estimate a log-linear summary of the clusters

Vector Outcomes: nonparametric†

The estimating equations we will use are:

$$\sum_{i=1}^{n} \left[\frac{\partial e^{X_i^T \boldsymbol{\beta}}}{\partial \boldsymbol{\beta}} \right]^T \left(Y_i - e^{X_i^T \boldsymbol{\beta}} \right) = \mathbf{0}_p$$

Because (for now) we are not weighting, these can be re-written as:

$$\sum_{ij} \frac{\partial e^{X_{ij}^T \boldsymbol{\beta}}}{\partial \boldsymbol{\beta}} \left(Y_{ij} - e^{X_i^T \boldsymbol{\beta}} \right) = \mathbf{0}_p$$

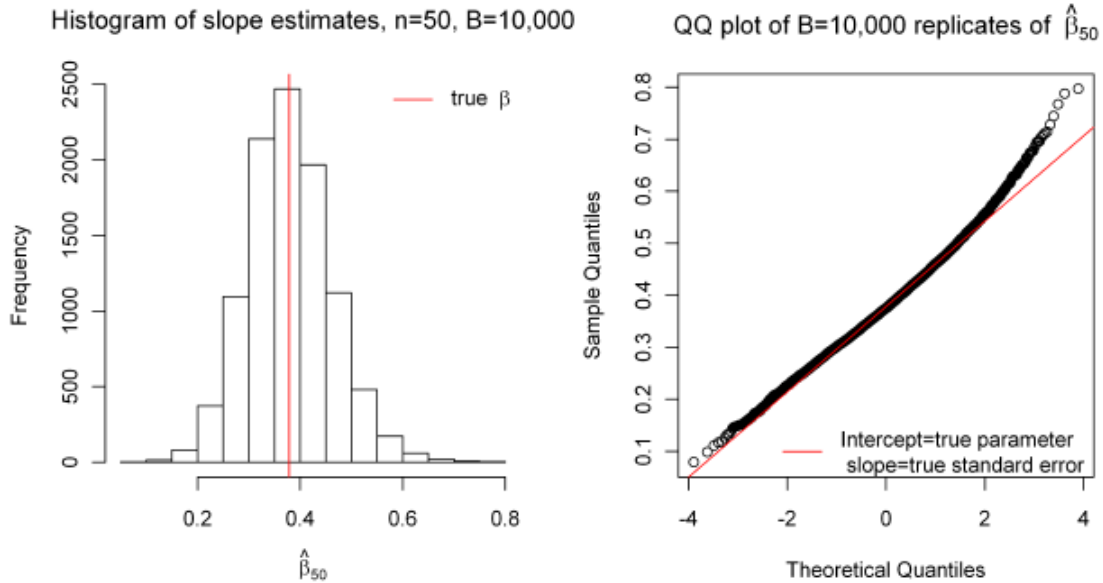
where now the terms in the summands are $p \times 1$ and 1×1.

Q. To what use of `glm()` does this correspond? (i.e. what `family` and `link` functions?) Hint: remember we're only using `glm()` to solve equations...

A. Your answer here.

Vector Outcomes: nonparametric†

The EE can be solved numerically (e.g. by Newton Raphson)[6] For some sample data sets:



[6] The EFs are still p functions of $\hat{\beta}$ with derivatives

Vector Outcomes: nonparametric†

We see that, despite not having a closed form, or being based on i.i.d. $\{Y \in \mathbb{R}, X \in \mathbb{R}^p\}$, the estimate $\hat{\boldsymbol{\beta}}$ really is approximately Normal - it is slightly heavy tailed in the RH tail, light tailed in the LH

- The Central Limit Theorem still applies, we sum n terms involve $Y_i \in \mathbb{R}^{n_i}$ and $n_i \times p$ matrix X_i, but we still have p estimating functions
- In this example, with $n = 50$ bias is not a major problem (although formally, for some datasets a finite root may not exist)
- The true log-linear slope parameter is $\beta = 0.378$; note this is not equal to any of the γ parameters, nor is it a simple function of them.
- The standard error of $\hat{\beta}_{50}$ is 0.082.

Vector Outcomes: sandwiches†

The same theory applies as for independent outcomes, asymptotically, the distribution of $\hat{\boldsymbol{\beta}}$ is given by:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_D N(\mathbf{0}_p, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})$$

where

$$\begin{aligned}
\mathbf{A} &= \mathbb{E}_F\left[\frac{\partial}{\partial \boldsymbol{\beta}}\left(\frac{\partial g(X_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(Y_i - g(X_i^T\boldsymbol{\beta}))\right)\right] \\
\mathbf{B} &= \mathbb{E}_F\left[\left(\frac{\partial g(X_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(Y_i - g(X_i^T\boldsymbol{\beta}))\right)\left(\frac{\partial g(X_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(Y_i - g(X_i^T\boldsymbol{\beta}))\right)^T\right]
\end{aligned}$$

While not stressed here, with weighted estimating equations the same form applies, but also involves inserting a $n_i \times n_i$ matrix W_i between the $\partial g/\partial \beta$ and residuals terms. "Pretty" choices of weight make nice cancellation happen - as in canonical link GLMs.

Vector Outcomes: sandwiches†

The estimators:

$$\hat{\mathbf{A}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} \left(\frac{\partial g(X_i^T \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} (Y_i - g(X_i^T \hat{\boldsymbol{\beta}})) \right)$$

$$\hat{\mathbf{B}} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial g(X_i^T \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} (Y_i - g(X_i^T \hat{\boldsymbol{\beta}})) \right) \left(\frac{\partial g(X_i^T \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} (Y_i - g(X_i^T \hat{\boldsymbol{\beta}})) \right)^T$$

which are combined in the usual way:

$$\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}(\hat{\mathbf{A}})^{-1}\hat{\mathbf{B}}(\hat{\mathbf{A}})^{-1}$$

Particularly when n_i is not constant, coding up these expression can require considerable care

- If possible, check your code on some data where $n_i = 1$
- Break the job into small chunks you understand
- Become adept with `by()`, `aggregate()`, and related functions

Vector Outcomes: sandwiches†

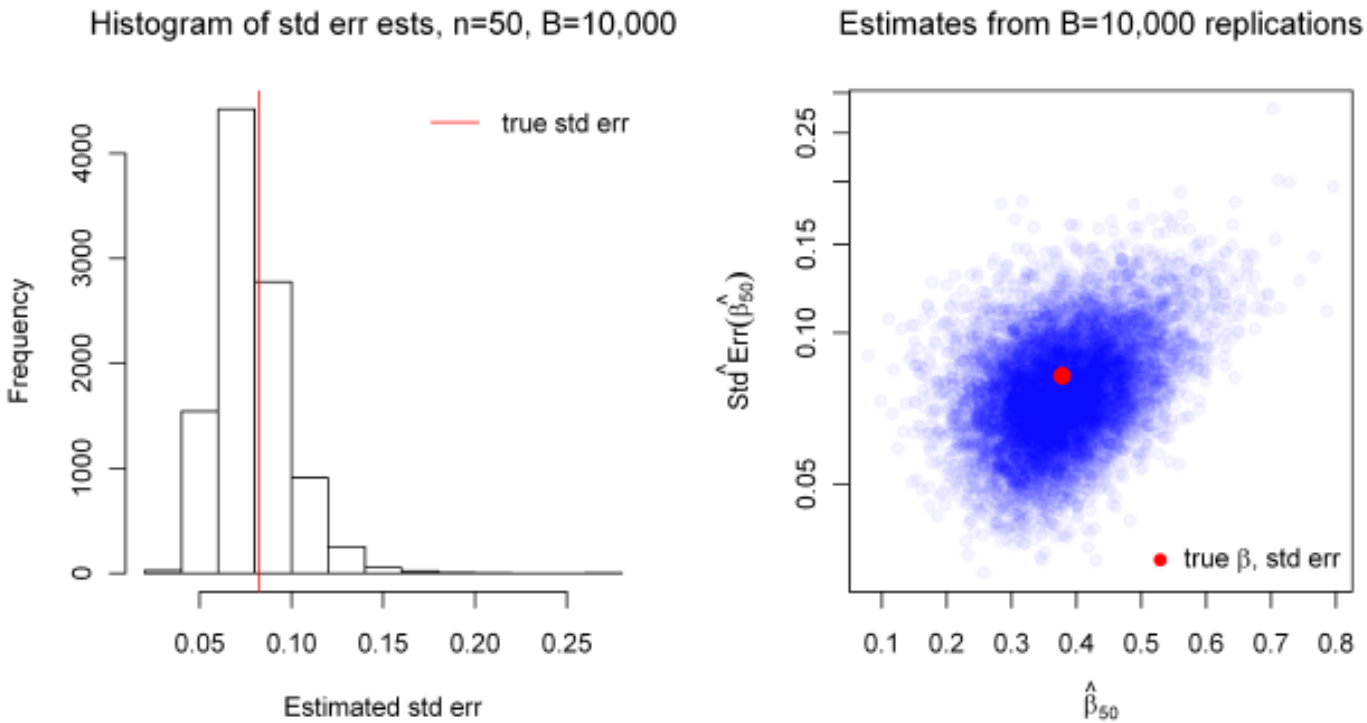
If the summands are $p \times p$ matrices, one way to add them up is with R code like this (for the log-link, unweighted example)

```
m1 <- make.one(100,4,c(0.4,0.1)) # data frame with int=1, x, y
b1 <- fit.one(m1)$coef # beta.hat - obtained using glm()
# a helper function
# - it adds up a list - a list of matrices, in our case
sum.list <- function(myL){ n <- length(myL); out <- myL[[1]]
for(i in 2:n){ out <- out + myL[[i]] }
out}
# evaluate the A summands;
allAi <- by(m1, m1$id, function(cluster){ # look up ?by
Xi <- as.matrix(cluster[,c("int","x")])
Yi <- cluster$y
mui <- as.vector(exp(Xi %*% b1))
resid <- Yi - mui
dgdbeta <- Xi * mui
t(Xi) %*% diag(mui * resid) %*% Xi - t(dgdbeta) %*% dgdbeta
})
Ahat <- sum.list(allAi)
```

... see class site for $\hat{\mathbf{B}}$. Turning this list of n matrices into a $n \times (p^2)$ data frame can speed things up - a little -but it is work

Vector Outcomes: sandwiches†

For $n = 50$, plots of the standard error estimates, and the point estimates, for the "slope" parameter β_1



Vector Outcomes: sandwiches†

- The approach is asymptotic in n, the number of clusters. For small n, it's not perfect, but it does work eventually. In many circumstances, practitioners used to small n dislike this approach
- The variability in the estimated standard error is *much* smaller than that of $\hat{\beta}$ (compare the histograms) - which is important, for accurate asymptotic approximations[7]
- The estimated standard error tends to be "too big" when $\hat{\beta}$ is also "too big". This makes the asymptotics work faster than we'd see for independent $\hat{\beta}$ and est. std. err., keeping the same marginal distributions. If the inaccuracies worked against each other, the asymptotics would work more slowly.
- The est'd std. err is not Normal - nor do we need it to be

[7] It may be helpful to recall that $t_{df} \approx N(0,1)$ and classical linear models

Vector Outcomes: describing output†

How to describe the output? Without further assumptions

- The point estimate $\hat{\beta}$ estimates the log-linear trend summarizing how Y varies with X
- ... weighting all observations equally
- With other covariates this gets trickier; it's common to refer to the association of Y and X "adjusting for Z"; we are describing one element of multidimensional "trend"

These "trend" parameters describe the population, but do not directly describe repeated sampling of a specific cluster, or clusters with $X = x$. In this sense, weak assumptions $\leftrightarrow$ weak inference (Our non-parametric inference for univariate Y could also be called "weak")

Vector Outcomes: describing output†

How to describe the estimates standard errors?

- The estimated standard error tells us about the variability in $\hat{\beta}$, if we repeated the whole experiment (sampling clusters) many times over
- ... yes, this estimate allows for within-cluster correlation of outcomes

As long as we are sampling in this way, the sandwich approach *does* describe the standard errors in terms strong enough for practical use.

(For frequentist replications where clusters and X are fixed, in large samples the sandwich will be a little conservative, at worst)

Vector Outcomes: describing output†

We can (justifiably) claim to be "allowing for" within-cluster correlations. But it's more accurate to say we are being *agnostic* to within-cluster correlation; we don't need any particular correlation structure to hold, nor does this approach derive any benefit from knowing that structure.

- To see the robustness, note that in the "middle" of $\hat{\mathbf{B}}$ we see terms:

$$(Y_i - g(X_i^T\hat{\beta}))(Y_i - g(X_i^T\hat{\beta}))^T$$

  This is an empirical estimate of $\mathrm{Var}[Y_i]$. For each individual i, it's likely a **terrible** estimate - but construction of the sandwich for $\mathrm{Var}[\hat{\beta}]$ uses an average of n of these empirical estimates - which will be very accurate, if n is big.
- Also, we'd get more efficiency if we knew and used information about the form of the correlations - used in GEE

Vector Outcomes: nonparametric approach summary†

Summary of EEs/sandwich, for vector outcomes:

- Relies on the same math as EEs for univariate outcomes - though now we consider a cluster's-worth of data as an independent unit, not each observation
- Methods work well with large n, though not so great with small n - which can be a serious problem, in some applications
- Can be hard to describe the trend correctly, i.e. sufficiently weakly
- No standard software exists, except where the math coincides with e.g. GEE - so expect to roll up your sleeves.