Part I:

Introductory topics

Modeling independent data

- Generalized linear regression models aim to learn about how variation in some univariate response, Y, depends on a set of p covariates, X
  - ⋆ linear regression
  - ⋆ generalized linear models
- Regression models generally have two components
  - ⋆ a *systematic* component
  - ⋆ a *random* component
- The systematic component provides structure for understanding mechanisms that generate the data as well as underlying associations
- The random component provides a means to ‘explain’ everything else

Linear regression

- Let i index units in the sample or the population

- By a *linear regression model* we mean a statistical model with the following elements/assumptions:

  (1) mean model: $\mathrm{E}[Y_i|X_i] = \mu_i = X_i^T\boldsymbol{\beta}$

  (2) error term: $\epsilon_i = Y_i - \mathrm{E}[Y_i|X_i]$

  (3) the ϵ_i's are independent

  (4) $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{V}[\epsilon_i] = \sigma_i^2$

  ⋆ (1) is the systematic component

  ⋆ (2), (3) and (4) jointly specify the random component

- Given a sample of size N, the *ordinary least squares* (OLS) estimator of $\boldsymbol{\beta}$ is:

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} \;=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

  - ⋆ $\boldsymbol{Y} \;=\; (Y_1,\ \ldots,\ Y_N)$
  - ⋆ $\boldsymbol{X}$ is an $N \times (p+1)$ matrix with row entries given by X_i

- It's straightforward to verify that $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is unbiased as an estimator of $\boldsymbol{\beta}$

- If the errors are homoskedastic we have that:

$$\text{Cov}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}] \;=\; \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

  - ⋆ exploit independence assumption to obtain a relatively simple expression
  - ⋆ require a plug-in estimate of σ^2

- If the errors are heteroskedastic we have that:

$$\text{Cov}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}] \; = \; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

  - ⋆ again require a plug-in estimator but now for $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$
  - ⋆ the independence assumption means that we 'only' require N elements to be estimated
  - ⋆ Huber-White estimator or the bootstrap

- The *weighted least squares* (WLS) estimator is:

$$\widehat{\boldsymbol{\beta}}_{\text{WLS}} \; = \; (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$$

- When we set $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$, we have the *generalized least squares* (GLS) estimator:

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} \; = \; (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

  - ⋆ BLUE by the Gauss-Markov Theorem

Generalized linear models (GLMs)

- The set-up for a GLM requires specification of three elements:

  (1) probability distribution, $Y_i \sim f_Y(y; \mu_i, \phi)$

  (2) linear predictor, $\eta_i = X_i^T \boldsymbol{\beta}$

  (3) link function, $g(\mu_i) = \eta_i$

  ⋆ element (1) is the random component

  ⋆ elements (2) and (3) jointly specify the systematic component

  ⋆ ϕ is a dispersion parameter, which may or may not be needed

- Given an i.i.d sample of size N, estimation/inference could proceed via:

  ⋆ maximum likelihood

  ⋆ quasi-likelihood

- Towards maximum likelihood-based estimation/inference, letting $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$ and assuming independence across study units, one can write down the *likelihood*:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) \;=\; \prod_{i=1}^{N} f_Y(y_i|\boldsymbol{\theta}),$$

  from which the *score function* can be derived:

$$\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) \;=\; \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}),$$

  and inference based on the inverse of the observed *information matrix* with $(j,k)^{th}$ element:

$$\boldsymbol{I}(\boldsymbol{\theta})_{j,k} \;=\; \sum_{i=1}^{N} -\frac{\partial^2 \ell(\boldsymbol{\theta}|y_i)}{\partial \theta_j \partial \theta_k}$$

- The mechanics for quasi-likelihood are essentially the same
  - ⋆ plug-in estimator for dispersion parameters

Dependence

- The independence assumption facilitated an initial simplification of the joint distribution of $\boldsymbol{Y}$
  - ⋆ structure of Cov$[\boldsymbol{Y}]$ in linear regression
  - ⋆ decomposition of the likelihood in GLMs

- In this course, we are going to consider settings where the independence assumption may not hold across all N study units
  - ⋆ some study units exhibit *dependence* with other study units

- Two questions:

Q: What does it mean for study units to be dependent?

Q: Why should we care?

- Suppose Y is continuous and consider the variance-covariance matrix for the responses from a sample of size N:

$$\text{Cov}[\boldsymbol{Y}] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \ldots & \sigma_{1N} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \ldots & \sigma_{2N} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \ldots & \sigma_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1N} & \sigma_{2N} & \sigma_{2N} & \ldots & \sigma_N^2 \end{bmatrix}$$

  - ⋆ $N + N(N-1)/2$ terms

- In practice, we'll seldom want to adopt the position that each of the N study units exhibits dependence with each of the other study units
  - ⋆ estimation of (asymptotic) variances is considerably simplified when we have some independent 'replication'

- Rather, we typically adopt some simplifying assumption(s) for the dependence structure across the N study units

- To guide this, it's worth thinking about how dependence might arise

- One way of thinking about dependence is that there is some phenomenon that 'connects' study units
  - ⋆ such that their responses co-vary or depend on each other
  - ⋆ either positively or negatively

- Conceptually, one might think of these 'connections' as arising due to one or more shared characteristics

- In this course, we will frame these connections by conceiving of the N study units as being naturally *clustered* in some way

- Examples include:
  - ⋆ repeated blood pressure measurements on an individual
  - ⋆ patients within a hospital
  - ⋆ gene expression measurements obtained in batches
  - ⋆ individual residences within geographic areas

- Note how the term ‘study unit’ refers to quite different entities in each of these examples, as does the notion of a ‘cluster’

- Reflected in the often-interchangeable terminology:

Clusters	Study units
individual/patient/subject	measurement/observation/time
physician/hospital	individual/patient/subject
county/state	people/hospital

- Knowing the scientific context will typically clarify whether or not we are talking about a cluster or a study unit

- Given a clustering, we then place structure on how study units are:
  1. dependent across clusters
  2. dependent within clusters

- In practice, we typically assume:
  - ⋆ study units between clusters to be independent
  - ⋆ some structure for dependence within clusters

- Intuitively, the ‘independence’ assumption can lead us to view the observed clusters as a random i.i.d sample from some population of such clusters

Q: Settings where the ‘independence’ assumption is unlikely to hold?

- For example, if there are N=8 study units across 3 clusters, one might specify the following dependence structure:

$$\text{Cov}[\boldsymbol{Y}] = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_2\sigma_3\rho & 0 & 0 & 0 & 0 & 0 \\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & 0 & 0 & 0 & 0 & 0 \\ \sigma_1\sigma_3\rho & \sigma_2\sigma_3\rho & \sigma_3^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4^2 & \sigma_4\sigma_5\rho & \sigma_5\sigma_6\rho & 0 & 0 \\ 0 & 0 & 0 & \sigma_4\sigma_5\rho & \sigma_5^2 & \sigma_5\sigma_6\rho & 0 & 0 \\ 0 & 0 & 0 & \sigma_4\sigma_6\rho & \sigma_5\sigma_6\rho & \sigma_6^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_7^2 & \sigma_7\sigma_8\rho \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_7\sigma_8\rho & \sigma_8^2 \end{bmatrix}$$

  - ⋆ study units across clusters are uncorrelated.
  - ⋆ study units within clusters are correlated, with the same degree of correlation across the 3 clusters: ρ.

- Notice how there is only one more parameter in this structure than what is required in the heteroskedastic setting for linear regression with independent study units
  - ⋆ i.e. the correlation ρ

- Clearly, there are many ways of structuring the dependence between and within clusters

- An important aspect of this course, therefore, will be to learn how to do so

- As we'll see, knowing how to structure dependence helps ensure valid estimation and efficient inference

- In some settings, knowing how to structure dependence will be important because it is of intrinsic scientific interest

Notation

- Suppose we observe a sample of K randomly selected clusters/subjects from some population of such clusters/subjects

- Let n_k denote the number of study units in the k^{th} cluster
  - ⋆ $N = \sum_{k=1}^{K} n_k$

- For the k^{th} cluster, we observe a vector response:

$$\boldsymbol{Y}_k \;=\; (Y_{k1}, \ldots, Y_{kn_k})^T$$

- Associated with the i^{th} measurement on the k^{th} cluster is a vector of covariates:

$$\boldsymbol{X}_{ki} \;=\; (X_{ki,1}, \ldots, X_{ki,p})$$

  - ⋆ first element of $\boldsymbol{X}_{ki}$ will typically be '1.0' for the intercept

- Let $\boldsymbol{X}_k$ denote the corresponding $n_k \times p$ matrix of covariate values for the k^{th} cluster:

$$\boldsymbol{X}_k \;=\; (\boldsymbol{X}_{k1}, \ldots, \boldsymbol{X}_{kn_k})^T$$

- Some covariates may take on the <u>same</u> value across all study units within a cluster:

$$X_{k1,j} \;=\; X_{k2,j} \;=\; \ldots \;=\; X_{kn_k,j}$$

  ⋆ *cluster-specific* or *between-subject* or *time-independent*

- Other covariates may take on <u>different</u> values:

$$X_{ki,j} \;\neq\; X_{kl,j} \qquad \text{for some } i \neq l$$

  ⋆ *subject-specific* or *within-cluster* or *time-dependent*

- In some instances, covariates could in principle vary within cluster but don't during the course of the study

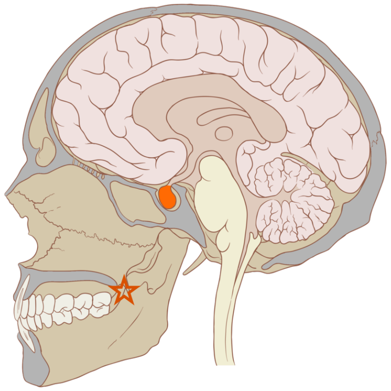
  ⋆ e.g. education

Data examples

Dental growth curve data

- K=26 children were followed every 2 years from 8 to 14 years of age
  - ⋆ total of N=104 visits
  - ⋆ data collected at the UNC Dental School in the 60's

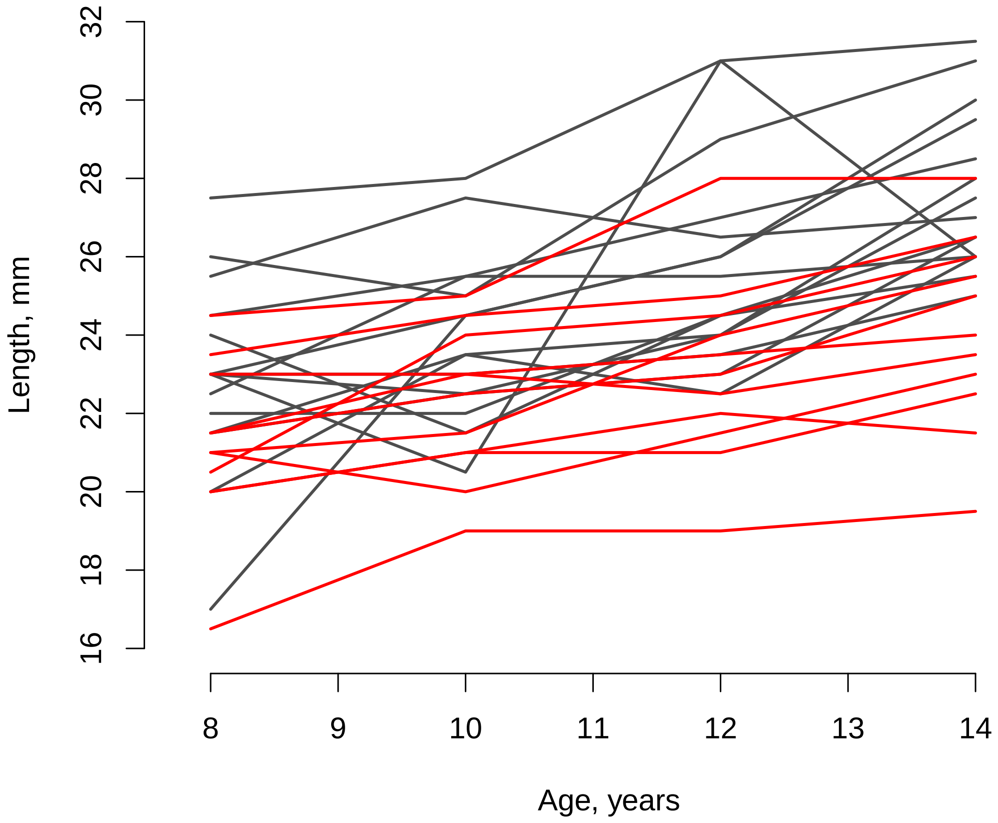
- Response of interest is the distance (mm) between the pituitary gland and the pterygomaxillary fissure:



```
> ##
> load("Growth.RData")
>
> ##
> growth
  id gender age length
1  1 female   8   21.0
2  1 female  10   20.0
3  1 female  12   21.5
4  1 female  14   23.0
5  2 female   8   21.0
...
>
> ##
> nrow(growth)
[1] 104
> length(unique(growth$id))
[1] 26
> table(growth$age)

 8 10 12 14
26 26 26 26
```

- Growth trajectories for the 26 children:
  - ⋆ grey lines are for the 15 boys
  - ⋆ red lines are for the 11 girls



- Note that gender is a cluster-specific covariate while age is a study unit-specific covariate

- Specific goals might include:

  1. estimation of the average growth curve
     - ∗ for all children
     - ∗ among all boys and all girls, separately
  2. identification of factors associated with growth
  3. testing of whether boys and girls differ in their average growth curves
  4. characterization of the degree of heterogeneity in the growth trajectories across children
  5. prediction of an individual childs' growth trajectory

CD4+ count data

- The Multicenter AIDS cohort study (MACS) was the first large study designed to investigate the natural history of AIDS

- Since 1984, MACS has enrolled ≈7,000 homosexual men
  - ⋆ UCLA, Northwestern, University of Pittsburgh, Johns Hopkins
  - ⋆ Kaslow et al (1987, AJE)

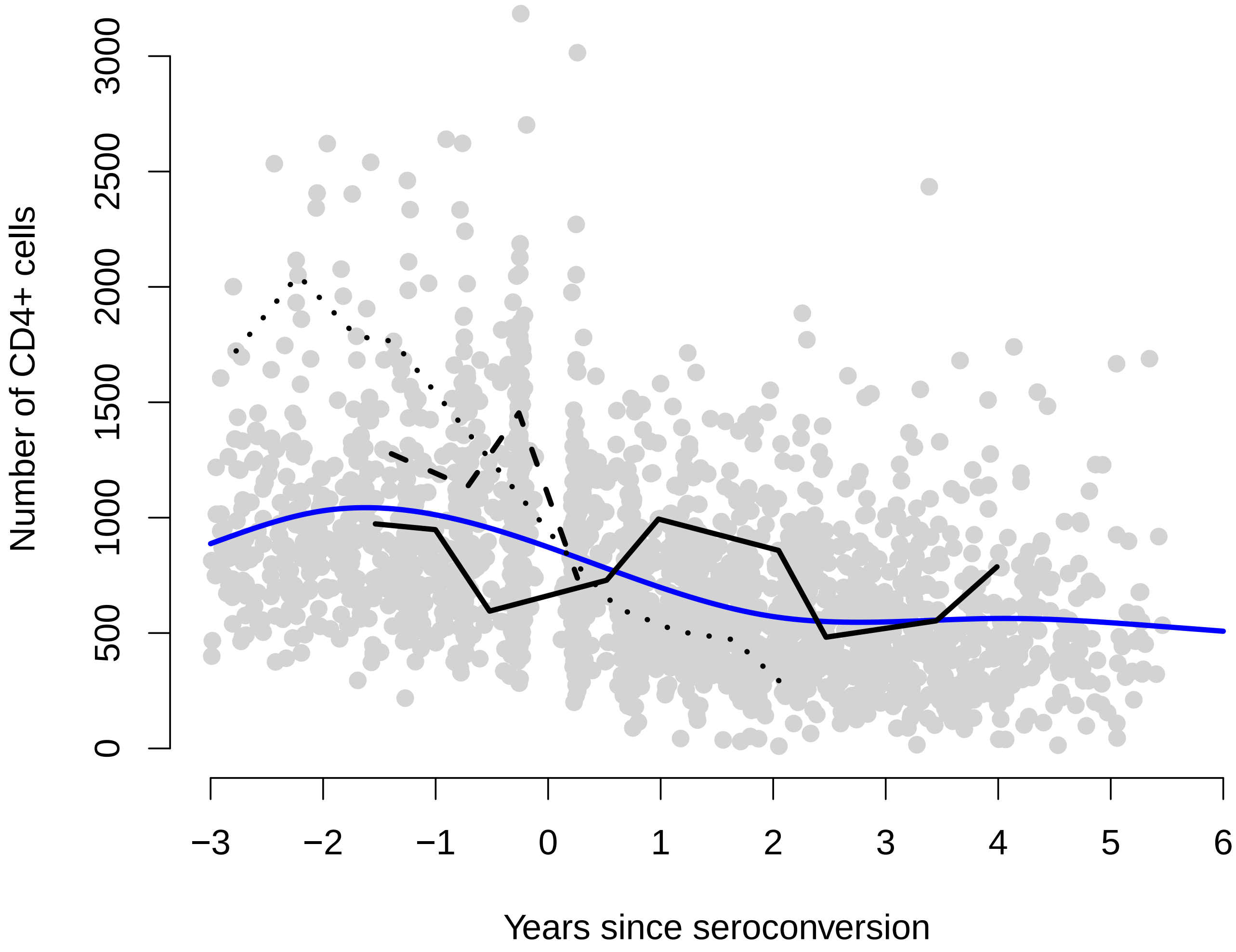
- We are going to focus on the CD4+ cell count trajectory over time
  - ⋆ CD4+ cells orchestrate the body's immune response

- N=2,376 measurements from K=369 men from the original cohort who seroconverted during follow-up
  - ⋆ 'seroconversion' corresponds to when HIV becomes detectable in the blood

```
> ##
> load("MACS.RData")
>
> ##
> macs
      id      time  age packs drug partners cesd cd4
1  10002 -0.741958 6.57     0    0       10   15 548
2  10002 -0.246407 6.57     0    1       10    9 893
3  10002  0.243669 6.57     0    1       10    6 657
4  10005 -2.729637 6.95     0    1       10   11 464
5  10005 -2.250513 6.95     0    1       10    3 845
...
>
> ##
> nrow(macs)
[1] 2376
> length(unique(macs$id))
[1] 369
> table(table(macs$id))

 1  2  3  4  5  6  7  8  9 10 11 12
 5 24 25 47 43 52 40 41 38 21 23 10
```

- CD4+ cell counts by time
  - ⋆ blue line is the overall (smoothed) average
  - ⋆ black lines are for three individual men



- In regard to CD4+ cell count trajectories, specific goals might include:

  1. estimation of the average CD4+ cell count trajectory among all men
  2. identification of factors associated with changes in CD4+ cell count
  3. testing of whether CD4+ count trajectories are associated with age
  4. characterization of the degree of heterogeneity in the CD4+ cell count trajectory across men
  5. prediction of the CD4+ cell count trajectory for an individual man

CMS data

- Investigate outcomes among patients diagnosed with pancreatic cancer
  - ⋆ Medicare Part A hospitalization data

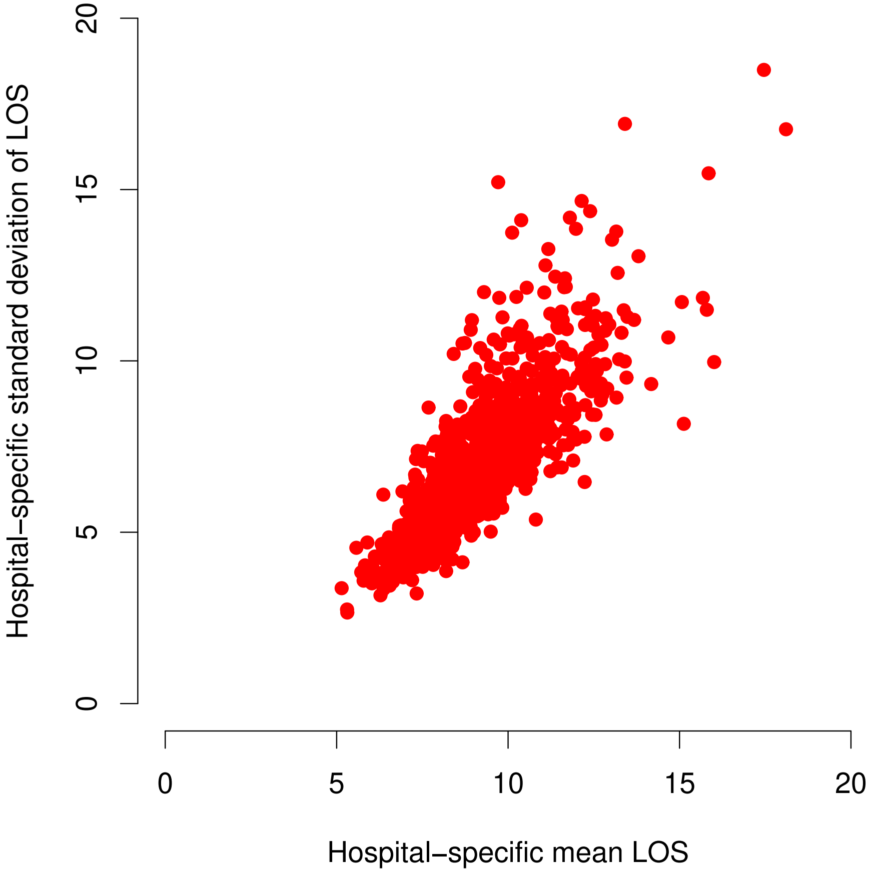
- Focus attention on patients:
  - ⋆ aged 65 years or older
  - ⋆ diagnosed between 2000-2009
  - ⋆ successfully discharged
    - ∗ did not die during the initial hospitalization
    - ∗ were not transferred to another hospital

- Additionally restrict to hospitals with at least 50 admissions

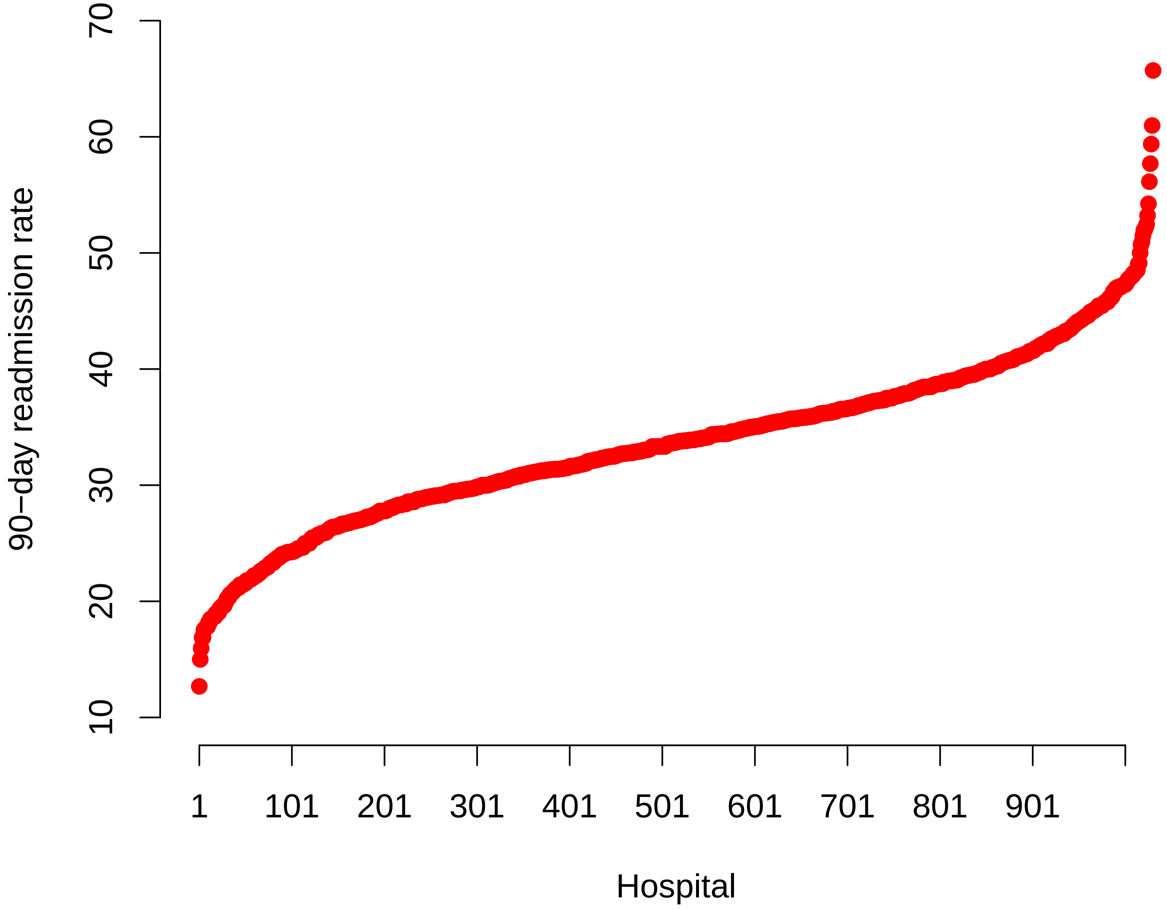
- Results in N=121,577 patients diagnosed at one of K=1,031 hospitals

```
> ##
> load("CMS.RData")
>
> ##
> CMS[1:5,]
  hospID hospVol year state female age  race admission deyo LOS  discharge T1 T2
1      1     228 2008    IL      1  72 Other        ER    1  14     1.Home  8 NA
2      1     228 2000    IL      1  65 White        ER    1   3    9.Other NA 45
3      1     228 2005    IL      0  77 White        ER    1  10 2.HomeCare  6 NA
4      1     228 2009    IL      1  67 White        ER    1  10  3.SNF/ICF NA 16
5      1     228 2000    IL      1  78 White     Other    1   4  3.SNF/ICF NA 19
>
> ##
> nrow(CMS)
[1] 121577
> length(unique(CMS$hospID))
[1] 1031
> summary(as.numeric(table(CMS$hospID)))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   51.0    66.0    88.0   117.9   134.0  1035.0
```

- Consider the response ‘length of hospital stay’
  - ⋆ continuous(ish)
  - ⋆ summarize hospitals by the mean and standard deviation



- Consider the response ‘readmission within 90 days’
  - ⋆ binary
  - ⋆ summarize hospitals by the mean or rate



- If we take length of hospital stay as the response, specific goals might include:

  1. estimation of the average length of stay
     - among all adults aged 65 years or older
     - among males and females aged 65 years or older, separately
  2. identification of factors associated with the average length of hospital stay
  3. testing whether the average length of hospital stay is associated with
     - hospital volume, a cluster-specific covariate
     - age, a patient-specific covariate
  4. characterization of the degree of heterogeneity in the average length of stay across hospitals
     - beyond that explained by covariates
  5. prediction of a given patients length of stay

Benefits of analyzing dependent data

Change over time

- A key benefit of longitudinal data is that one can investigate changes in the response over time
  - ⋆ within an individual or patient
- This, in turn, means that we can distinguish *cohort effects* from *age effects*
- For example, in the MACS study we might hypothesize and investigate
  - ⋆ a cohort effect: at the time of seroconversion, younger men have higher CD4+ count
  - ⋆ an age effect: post-seroconversion, the trajectory of CD4+ count is steeper for older men

Cross-sectional vs. longitudinal effects

- Beyond the ‘effect’ of time, we can also examine the effect of changes in some exposure or risk factor over time
  - ⋆ again within an individual or patient

- To be concrete, suppose Y_{ki} is the response for the k^{th} subject at the i^{th} time point in a longitudinal study and consider the model:

$$\mathsf{E}[Y_{ki}] \;=\; \beta_0 \;+\; \beta_C X_{k1} \;+\; \beta_L(X_{ki} - X_{k1}).$$

- Notice that

$$\mathsf{E}[Y_{k1}] \;=\; \beta_0 \;+\; \beta_C X_{k1}$$

  - ⋆ β_C is the difference in the expected response at the first time point (baseline) between two populations that differ in X by one unit
  - ⋆ referred to as a *cross-sectional contrast*

- Also notice that

$$\mathsf{E}[Y_{ki} - Y_{k1}] \; = \; \beta_L(X_{ki} - X_{k1}).$$

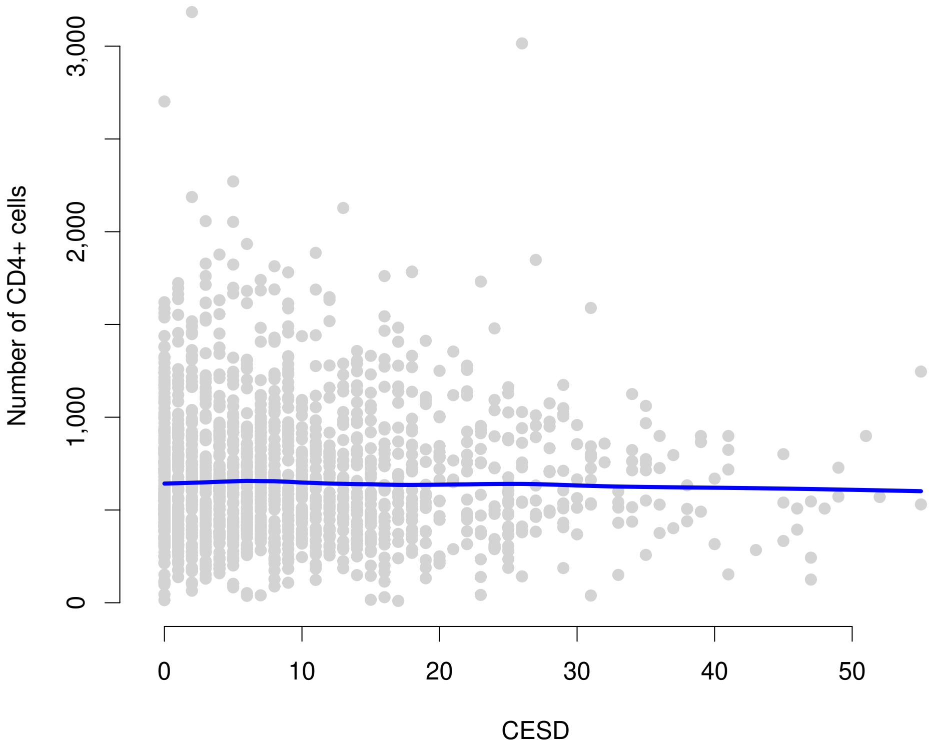
  - ⋆ β_L is the change in the expected response per unit change in X for populations with the same baseline value of X
  - ⋆ referred to as a *longitudinal contrast*

- Note, in the absence of longitudinal data we would not be able to estimate β_L and, therefore, could not distinguish β_C from β_L
  - ⋆ only have (Y,X) at a single time point

- To see this idea a little more concretely, consider the association between CD4$+$ cell count and CESD, a measure of depression, in the MACS data

- Restrict attention to K^*=266 patients with at least one pre- and one post-seroconversion measurement
  - ⋆ pre measurement had to have been within 6 months of seroconversion

- A simple (unadjusted) scatterplot indicates that there is little-to-no evidence of an association:



Q: Can we disentangle a cross-sectional effect of CESD from a longitudinal effect?

- ⋆ consider, specifically, the cross-sectional effect at the time of seroconversion

- Let $X_{k'}$ be the CESD score that is closest to but just before seroconversion for the k^{th} patient
  - ⋆ the ‘baseline’ measurement

- $Y_{k'}$ is the corresponding CD4 cell count

- Consider the model:

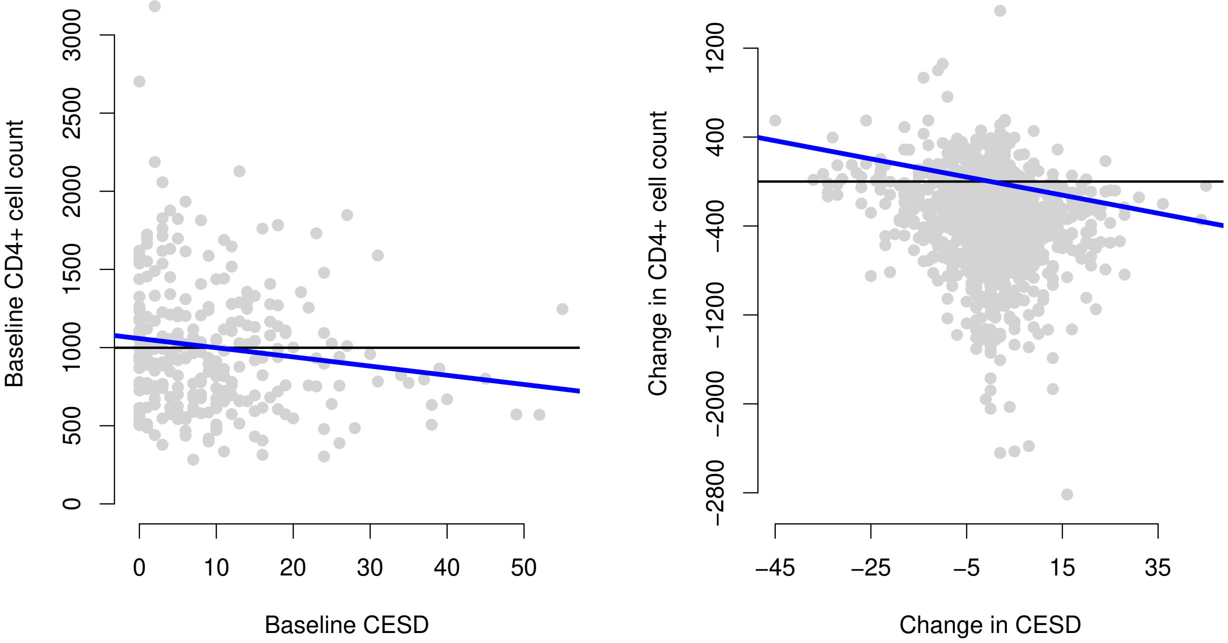
$$\mathsf{E}[Y_{ki}] \;=\; \beta_0 \;+\; \beta_C X_{k'} \;+\; \beta_L(X_{ki} - X_{k'}).$$

from which we have:

$$\begin{aligned} \mathsf{E}[Y_{k'}] \;&=\; \beta_0 \;+\; \beta_C X_{k'} \\ \mathsf{E}[Y_{ki} - Y_{k'}] \;&=\; \beta_L(X_{ki} - X_{k'}) \end{aligned}$$

- Note, we could also get at these with two separate regressions

- We can visualize the effects graphically as:



- Suggests that:
  - ⋆ men with more depressive symptoms at seroconversion tend to have lower CD4+ cell counts
  - ⋆ men who experience increases in depressive symptoms over time tend to experience greater decreases in CD4+ cell count

Efficiency

- Beyond expanding the range of questions that one can address, the analysis of repeated measurements on the same cluster can also provide efficiency gains

- Consider a randomized trial of some active treatment versus a control
  - ⋆ $K/2$ study participants in each arm
  - ⋆ X_k=0/1 is a binary indicator of treatment assignment (control/active) for the k^{th} participant

- Suppose the response of interest, Y, is measured at baseline and at some follow-up visit
  - ⋆ n_k=2 measurements on each patient
  - ⋆ let T_{ki}=0/1 be an indicator of whether the i^{th} observation on the k^{th} patient is a baseline or follow-up measurement

- Consider the following three models:

$$\begin{aligned} \mathsf{E}[Y_{ki}|\ X_k, T_{ki}] &= \beta_0 + \beta_1 X_k + \beta_2 T_{ki} + \gamma X_k T_{ki} \\ \mathsf{E}[Y_{ki}|\ X_k, T_{ki}] &= \beta_0 + \beta_2 T_{ki} + \gamma X_k T_{ki} \\ \mathsf{E}[Y_{ki}|\ X_k, T_{ki}=1] &= \alpha_0 + \gamma X_k \end{aligned}$$

  - ⋆ γ is the same in each of these models

- Interpretation of γ?
  - ⋆ consider the change in expected response from baseline to the follow-up time
  - ⋆ γ is the difference in the change in expected response, comparing two populations defined by treatment allocation

- Finally, suppose $\boldsymbol{Y}_k = (Y_{k1}, Y_{k2})$ is distributed according to a bivariate Normal with
  - ⋆ $\mathsf{V}[Y_{ki}] = \sigma^2$, for i=1,2
  - ⋆ $\mathsf{Cor}[Y_{k1}, Y_{k2}] = \rho$

- The MLEs of γ based on the three models are:

$$\begin{aligned}\widehat{\gamma}^1 &= [\hat{\mu}_{12} - \hat{\mu}_{11}] - [\hat{\mu}_{02} - \hat{\mu}_{01}] \\ \widehat{\gamma}^2 &= [\hat{\mu}_{12} - \rho\hat{\mu}_{11}] - [\hat{\mu}_{02} - \rho\hat{\mu}_{01}] \\ \widehat{\gamma}^3 &= [\hat{\mu}_{12}] - [\hat{\mu}_{02}]\end{aligned}$$

  ⋆ μ_{xt} is the mean for arm $X = x$ at time $T = t$

- The variances of the MLEs of γ based on the three models are, respectively:

$$\begin{aligned}\mathsf{V}[\widehat{\gamma}^1] &= \frac{4\sigma^2}{K} \times 2(1-\rho) \\ \mathsf{V}[\widehat{\gamma}^2] &= \frac{4\sigma^2}{K} \times (1-\rho^2) \\ \mathsf{V}[\widehat{\gamma}^3] &= \frac{4\sigma^2}{K}\end{aligned}$$

- If $\rho > 0$ one can take advantage of using both measurements to increase efficiency

The analysis of dependent data

- While there are important benefits associated with the use of dependent data, the main drawback is that the analysis is typicaly more complex

- There is a massive literature on the analysis of dependent data

- Key point:

  There is no single approach that is appropriate for all settings

- The fact that this is the case can be frustrating in practice

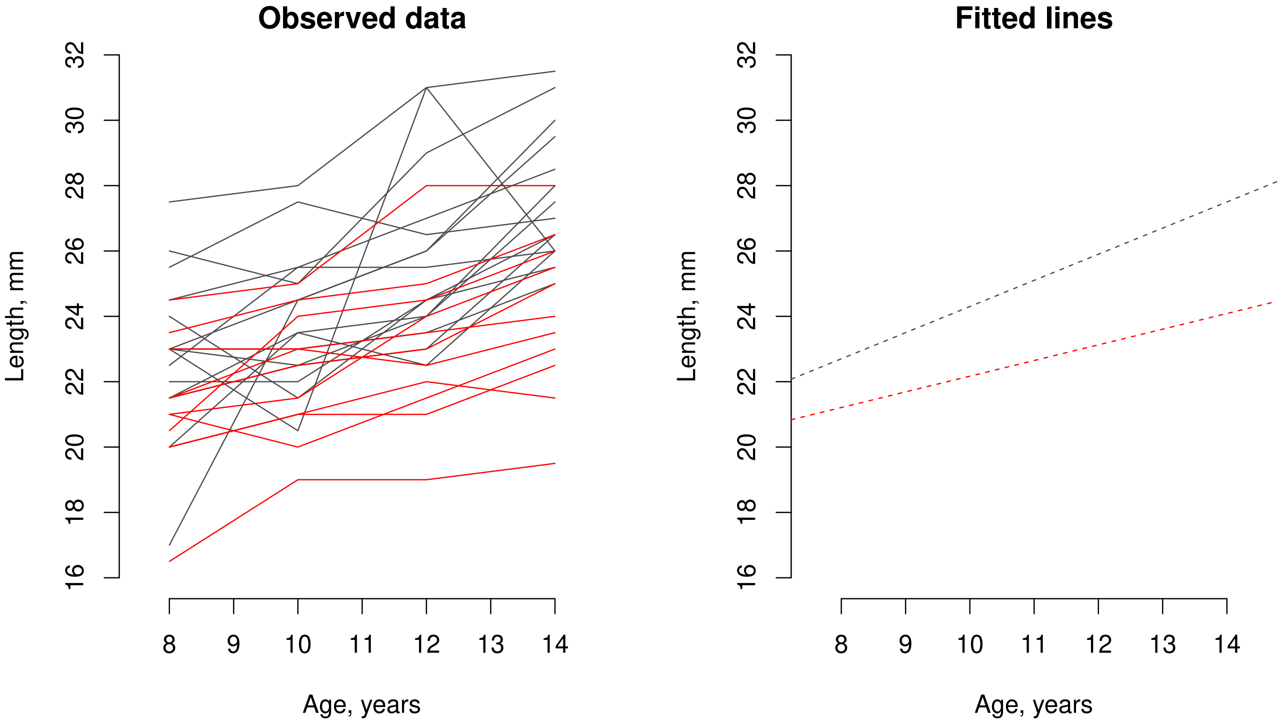
- How one moves forward may depend on:
  - ⋆ the framing of the scientific goals
  - ⋆ the assumptions one is willing to make
  - ⋆ your own personal philosophy and those of your collaborators

- One option would be to fit a single curve to all of the data

$$\mathsf{E}[Y_{ki}] \;=\; X_{ki}^T\boldsymbol{\beta}$$

- Returning to the dental growth data, one could fit the model:

$$\mathsf{E}[Y_{ki}] \;=\; \beta_0 \;+\; \beta_1 A_{ki} \;+\; \beta_1 G_k \;+\; \beta_3 A_{ki} G_k$$



- Appealing in that the interpretation of $\boldsymbol{\beta}$ follows the interpretation of models that were covered in Methods I
  - ⋆ differences in average responses between different populations of study units

- Note such an interpretation does not condition on cluster membership
  - ⋆ 'averages' are across clusters

- An example of a *marginal* model
  - ⋆ marginal with respect to the cluster membership

- Possibly accompany this specification of the mean model with a (separate) model for the dependence structure

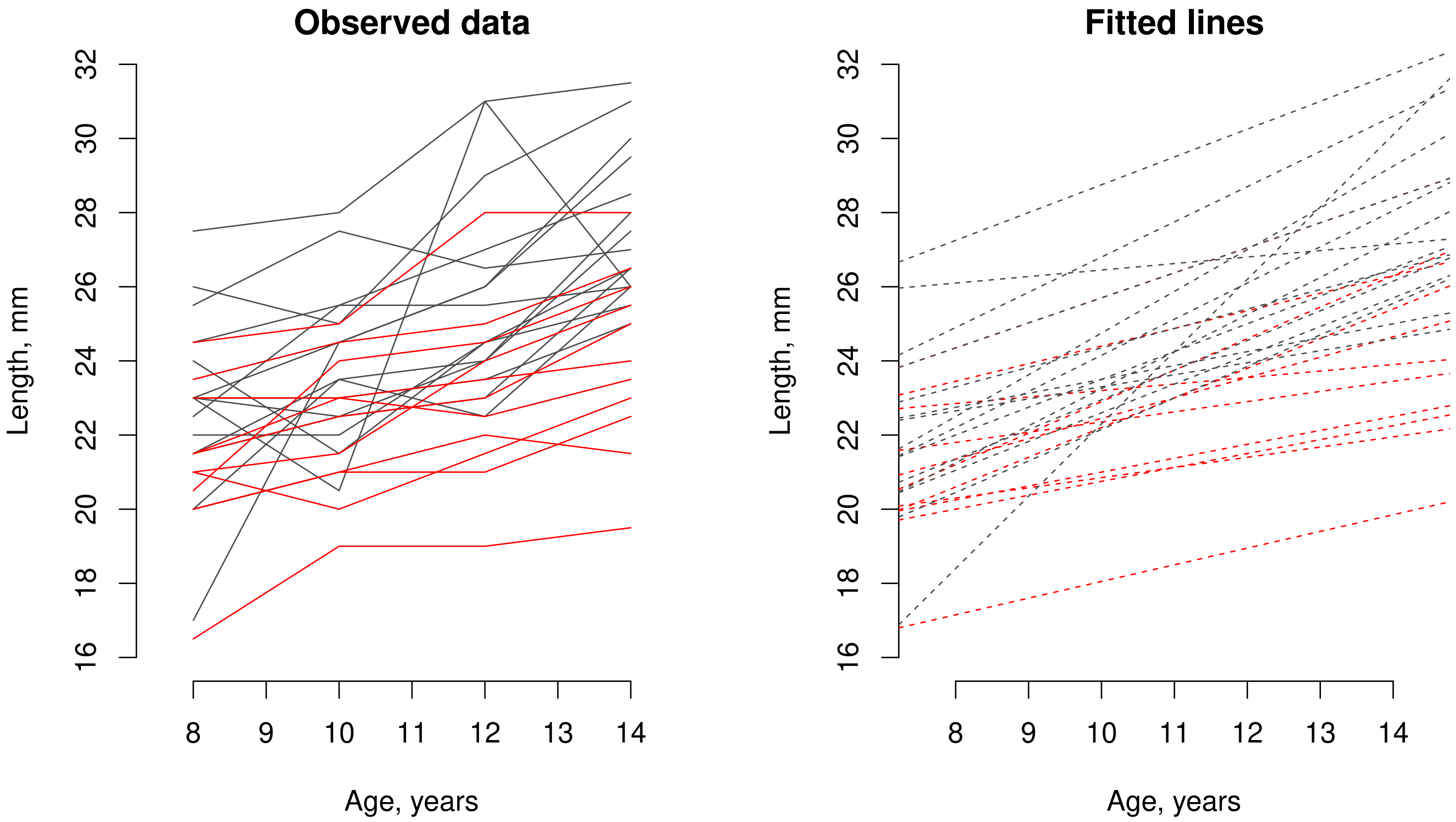
$$\text{Cov}[\boldsymbol{Y}_k] \; = \; \boldsymbol{\Sigma}_k(\boldsymbol{\alpha})$$

  - ⋆ $\boldsymbol{\alpha}$ would need to be estimated

- A second option would be to fit separate curves to each cluster:

$$\mathsf{E}[Y_{ki}] = X_{ki}^T \boldsymbol{\beta}_k$$

- For the dental data, this would amount to fitting K=26 separate regressions:



- Interpretation of $\boldsymbol{\beta}_k$ for this specification is in terms of averages across study units within a specific cluster

- An example of a *conditional* model
  - ⋆ conditional with respect to cluster membership

- Instead of adopting one of the two extremes:

$$
\begin{aligned}
(i)\ \mathsf{E}[Y_{ki}] &= X_{ki}^T\boldsymbol{\beta} \\
(ii)\ \mathsf{E}[Y_{ki}] &= X_{ki}^T\boldsymbol{\beta}_k
\end{aligned}
$$

  we might want to adopt some intermediary structure across the clusters
  - ⋆ enable borrowing of strength across clusters
  - ⋆ enable the characterization of variation across clusters
  - ⋆ enable distinguishing between- vs within-cluster effects

- One approach to doing this is to constrain certain parameters to be the same across clusters

- For example, one could assume that the covariate effects are the same across clusters but that they each have their own intercept:

$$\mathsf{E}[Y_{ki}] \;=\; \beta^{\dagger}_{0k} \;+\; \beta^{\dagger}_{1} X_{ki,1} \;+\; \ldots \;+\; \beta^{\dagger}_{p-1} X_{ki,p-1}$$

- Note, the interpretation of $\beta^{\dagger}_{j}$ requires holding 'cluster' fixed
  - ⋆ because of the cluster-specific intercepts

- Hence this model is also a *conditional* model

- Adopting this model ultimately requires estimation of $K + (p-1)$ parameters
  - ⋆ asymptotics get tricky because the number of parameters increase with 'sample size' K

- To mitigate this problem, one could incorporate additional structure across the β_{0k}
  - ⋆ reduce the dimension of unknown parameters

- Mixed effects models provide one approach to doing this

- For example, note that we can re-write the previous model as:

$$\begin{aligned} \mathsf{E}[Y_{ki}] &= \beta_{0k}^{\dagger} \ + \ \beta_1^{\dagger} X_{ki,1} \ + \ \ldots \ + \ \beta_{p-1}^{\dagger} X_{ki,p-1} \\ &= (\beta_0^{\dagger} + \gamma_k) \ + \ \beta_1^{\dagger} X_{ki,1} \ + \ \ldots \ + \ \beta_{p-1}^{\dagger} X_{ki,p-1} \\ &= X_{ki}^T \boldsymbol{\beta}^{\dagger} \ + \ \gamma_k \end{aligned}$$

- One could then assume that the γ_k arise from some distribution that characterizes variation in a (hypothetical) population of clusters from which we have a sample of size K
  - ⋆ by far the most common choice is that $\gamma_k \ \sim_{\text{i.i.d}} \ \text{Normal}(0, \sigma_\gamma^2)$
  - ⋆ other choices are certainly possible

- Yet another approach that is particularly useful in longitudinal studies is to use a transition model

$$\mathsf{E}[Y_{ki}] \;=\; X_{ki}^T\boldsymbol{\beta}^* \;+\; \mathcal{Y}_{ki}\boldsymbol{\alpha}$$

  - ⋆ $\mathcal{Y}_{ki}$ is the *history* prior to the i^{th} observation

- The interpretation of $\boldsymbol{\beta}^*$ requires conditioning on the history of the cluster

- Therefore, this is another example of a *conditional* model

Exploratory analysis

- While much of this course will focus on regression as a tool for answering scientific questions of interest, exploratory analyses are often a useful initial step
- Understand the nature of the available data
  - ⋆ make sure it is consistent with your understanding of how the data was supposed to have be collected
- Reveal unusual observations and/or missingness patterns
- Initial exploration of structure
  - ⋆ mean model
  - ⋆ dependence model

Summaries

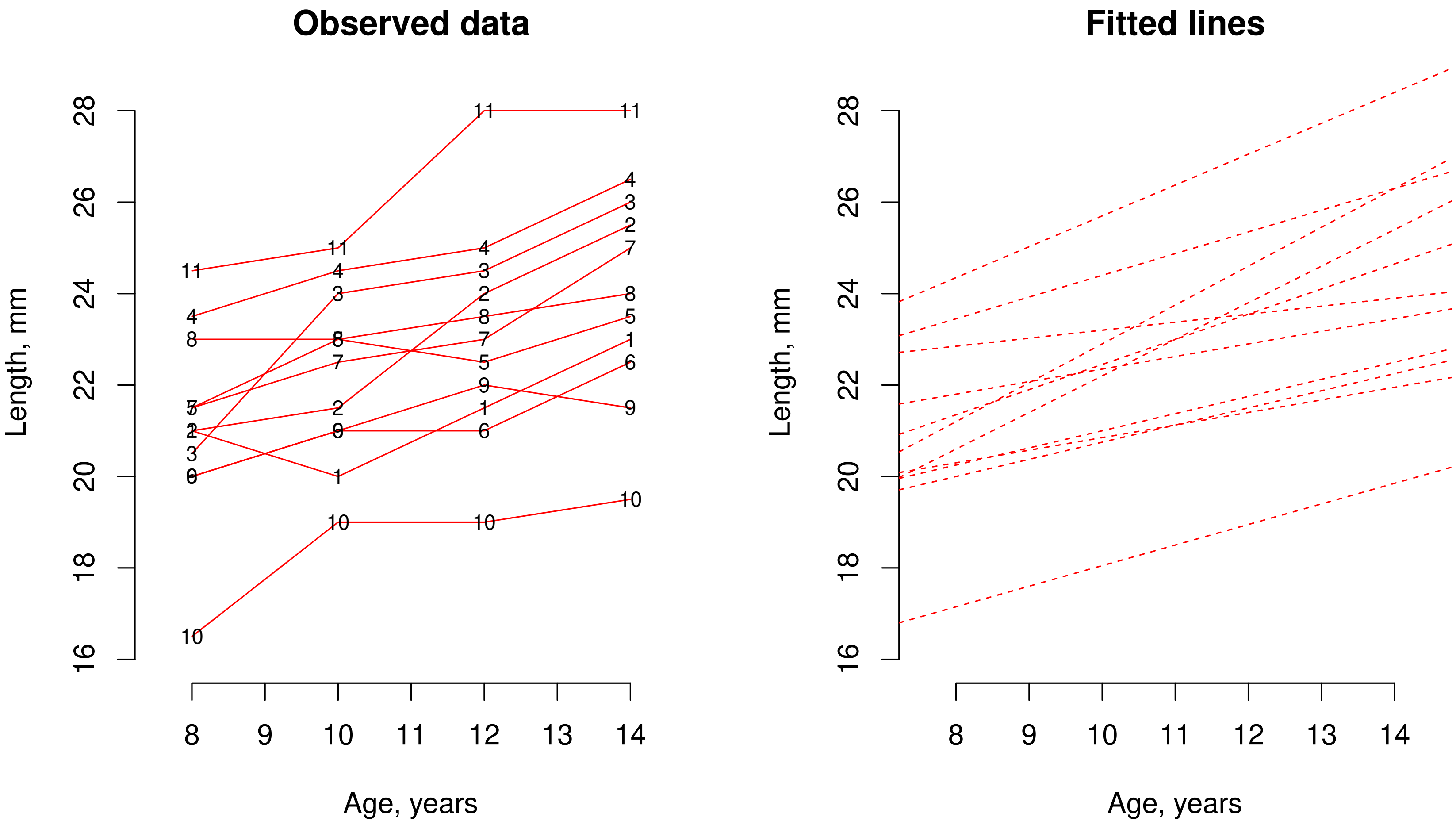
- Consider the dental length data from UNC

	Age, years			
	8	10	12	14
Mean length (mm)				
Males	22.9	24.0	25.9	27.6
Females	21.2	22.2	23.1	24.1
Difference (mm)	1.7	1.8	2.8	3.5

- There seems to be preliminary evidence for:
  - ⋆ **trends** in that average dental length increases with age for males and females
  - ⋆ **cross-sectional differences** in that males have larger average dental length at each age
  - ⋆ **longitudinal differences** in that the increase in average dental length over time is greater for males

Individual trajectory plots

- Now consider the individual trajectories among the 11 females:



Observed data

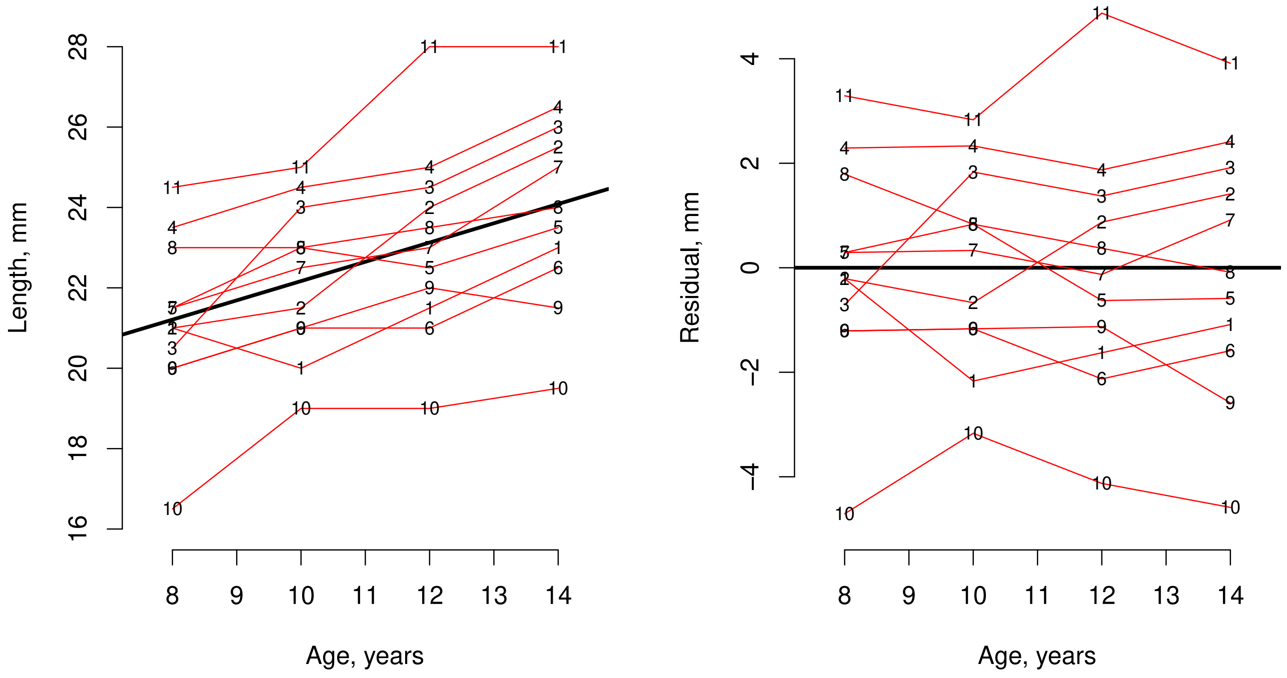
Fitted lines

- From the plots we find that there is preliminary evidence for:
  - ⋆ **a trend** in that dental length increases with age for females
  - ⋆ **tracking** in that females with large dental length at younger ages tend to have large dental length at older ages
  - ⋆ **comparable variation over time** in that variation across females is (roughly) similar across the ages

- Despite only have 11 females, we also find a number of potential outliers/strange observations:
  - ⋆ subject 10 appears to have quite small dental lengths, relative to the other females
  - ⋆ subject 11 appears to have quite large dental lengths, relative to the other females
  - ⋆ subjects 1, 5 and 9 appear to experience decreases in dental length during some intervals

Individual trajectory plots of the residuals

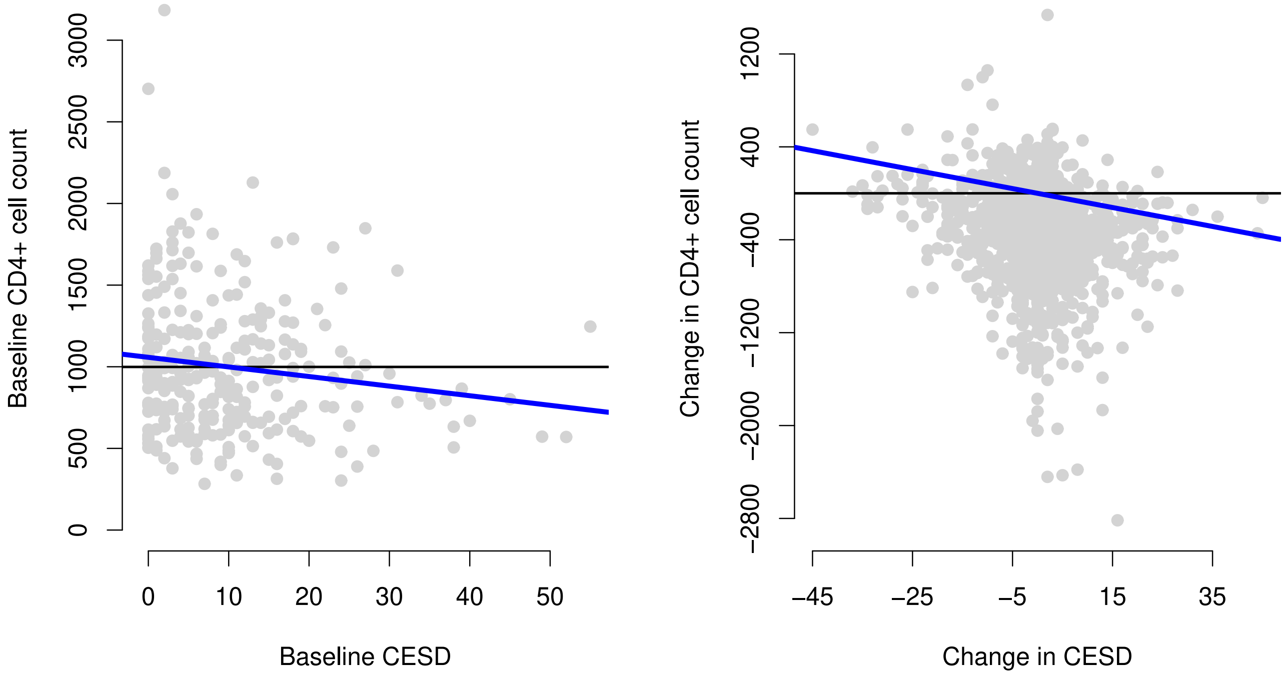
- Sometimes it is easier to identify patterns and/or unusual observations if one removes the average trend
  - ⋆ generally easier to see variation around a flat line than around a slope



- Typically, we consider the random component of any given model after having ‘accounted for’ the systematic component

Exposures other than time

- Returning to the MACS data, consider again the association between CD4+ cell count and CESD



- We may want to revisit these preliminary findings by removing time trends
  - ⋆ consider the associations ‘adjusting’ for time

- Towards this, consider again the model

$$\mathsf{E}[Y_{ki}] \; = \; \beta_0 \; + \; \beta_C X_{k'} \; + \; \beta_L (X_{ki} - X_{k'}).$$

  from which we have:

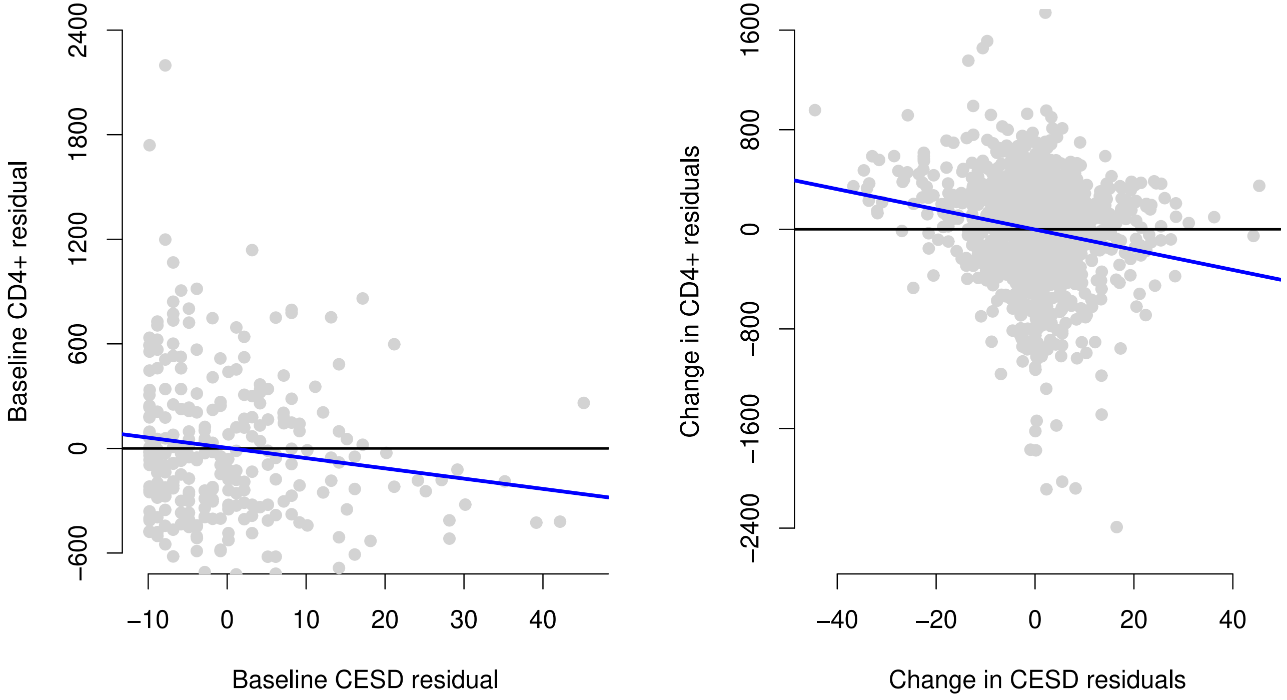
$$\begin{aligned} \mathsf{E}[Y_{k'}] \; &= \; \beta_0 \; + \; \beta_C X_{k'} \\ \mathsf{E}[Y_{ki} - Y_{k'}] \; &= \; \beta_L (X_{ki} - X_{k'}) \end{aligned}$$

- We can 'remove' the time trends in Y and X by taking the residuals from models of each with time as a predictor

- For example:

```
>##
> fitX <- lm(cesd ~ ns(time, knots=c(-2, 0, 2, 4)), data=macsSub)
> fitY <- lm(cd4 ~ ns(time, knots=c(-2, 0, 2, 4)), data=macsSub)
>
>##
> residX <- residuals(fitX)
> residY <- residuals(fitY)
```

- Plot:
  - ⋆ residuals that correspond to the baseline measurement
  - ⋆ change in residuals, relative to the baseline measurement



- In this instance, we draw the same general conclusions

Exploring the covariance structure

- Recall the marginal model: $\mathsf{E}[Y_{ki}] \;=\; X_{ki}^T\boldsymbol{\beta}$
- Suppose $\widehat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$ and consider the (marginalized) residuals:

$$R_{ki} \;=\; Y_{ki} \;-\; X_{ki}^T\widehat{\boldsymbol{\beta}}, \qquad i = 1, \ldots, n_k, \;\; k = 1, \ldots, K$$

- The standard deviation and correlation matrix of the residuals for the k^{th} cluster is:

$$\boldsymbol{S} \;=\; \begin{bmatrix} \sigma_1 & & & \\ \rho_{12} & \sigma_2 & & \\ \vdots & \vdots & \ddots & \\ \rho_{1n_k} & \rho_{2n_k} & \ldots & \sigma_{n_k} \end{bmatrix}$$

where

$$\sigma_i \;=\; \sqrt{\mathsf{V}[R_{ki}]} \quad \text{and} \quad \rho_{ij} \;=\; \frac{\mathsf{Cov}[R_{ki}, R_{kj}]}{\sqrt{\mathsf{V}[R_{ki}]\mathsf{V}[R_{kj}]}}$$

- Returning to the dental growth data, let's fit a marginal model to the females as a linear function of age and examine the residuals:

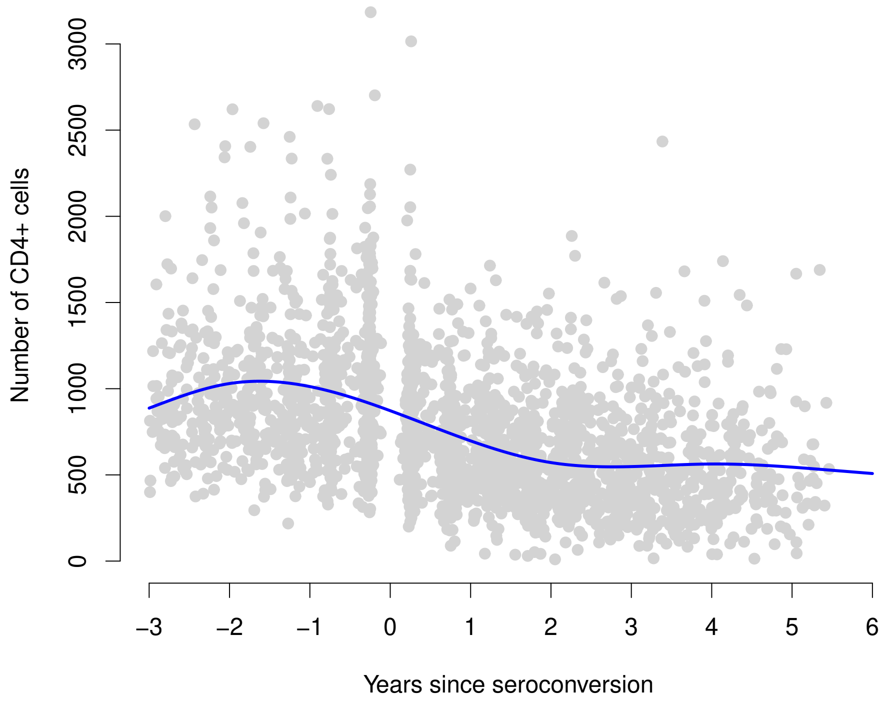
```
> ##
> fitF   <- lm(length ~ age, data=growth, subset=(growth$gender == "female"))
> resMat <- matrix(residuals(fitF), ncol=4, byrow=TRUE)
> round(sqrt(diag(cov(resMat))), 2)
> round(cor(resMat), 2)
```

- Yields:

$$\widehat{\boldsymbol{S}} = \begin{bmatrix} \mathbf{2.12} & & & \\ 0.83 & \mathbf{1.90} & & \\ 0.86 & 0.90 & \mathbf{2.36} & \\ 0.84 & 0.88 & 0.95 & \mathbf{2.44} \end{bmatrix}$$

- Suggests:
  - ⋆ heteroskedasticity, in that variation increases with age
  - ⋆ strong correlation among observations within a child

- Repeating this in the MACS data requires creating categorical bins for the timing of the observed CD4+ counts



- Categorize time:

```
> ##
> tcat <- round(macs$time)
> table(tcat)
 -3  -2  -1   0   1   2   3   4   5
 71 198 315 529 431 346 254 163  69
```

- Collapse years -3 and 5 into adjacent categories:

```
> ##
> tcat <- ifelse(tcat == -3, -2, ifelse(tcat == 5, 4, tcat)) + 3
> table(tcat)
  1   2   3   4   5   6   7
269 315 529 431 346 254 232
```

- Fit a flexible marginal model:

```
> ##
> fitMM  <- lm(cd4 ~ ns(time, knots=c(-2, 0, 2, 4)), data=macs)
```

- As we consider the residuals, we should note that, in contrast to the dental growth data, the observed data are not *balanced*
  - ⋆ n_k is not the same across all clusters
  - ⋆ patients contribute to different (categorized) time points
  - ⋆ patients contribute multiple observations to the same (categorized) time point

- For simplicity, consider the mean residual for any given individual during any given (categorized) time point

```
> ##
> resMat <- tapply(residuals(fitMM), list(macs$id, tcat), FUN=mean)
> round(resMat)
         1    2    3    4    5    6   7
10002   NA -436  -96   NA   NA   NA  NA
10005 -316   NA -266 -389   NA   NA  NA
10029   NA  -34  -59   99  172   NA  NA
10039   NA  194  225   NA   NA   NA  NA
10048   NA   NA -131   37 -224 -324  NA
10052   NA  -95   NA -364   NA   NA  50
10079   NA -555 -338 -563   NA   NA  NA
10088   NA   NA  -50  -85  394   86  11
...
```

- Estimate variance/covariance matrix, and correlation matrix, specifying the `use="pairwise.complete.obs"` option in R

```
> ##
> sqrt(diag(cov(resMat, use="pairwise.complete.obs")))
> cor(resMat, use="pairwise.complete.obs")
```

- Yields:

$$\widehat{S} = \begin{bmatrix} \mathbf{379} & & & & & & \\ 0.70 & \mathbf{397} & & & & & \\ 0.61 & 0.58 & \mathbf{349} & & & & \\ 0.47 & 0.55 & 0.59 & \mathbf{264} & & & \\ 0.30 & 0.46 & 0.51 & 0.75 & \mathbf{301} & & \\ 0.50 & 0.56 & 0.46 & 0.67 & 0.81 & \mathbf{296} & \\ 0.89 & 0.47 & 0.49 & 0.59 & 0.73 & 0.83 & \mathbf{323} \end{bmatrix}$$

- Find:
  - ⋆ no clear indication of a mean-variance relationship
  - ⋆ some indication that correlation decays as the distance between two observations increases

- When there is a lack of balance, it is also worth characterizing how many data points one has to estimate any given variance/correlation component:

```
> ##
> nS <- matrix(NA, nrow=7, ncol=7)
> for(i in 1:7){
+   for(j in 1:7) nS[i,j] <- nrow(na.omit(resMat[,c(i,j)]))
+ }
> nS
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  145  114  101   90   76   36    9
[2,]  114  211  171  157  121   80   47
[3,]  101  171  307  236  192  144  106
[4,]   90  157  236  279  195  149  104
[5,]   76  121  192  195  226  142   95
[6,]   36   80  144  149  142  167  101
[7,]    9   47  106  104   95  101  116
```

- Suggests that we shouldn't put too much stock into the estimate $\hat{\rho}_{17} = 0.89$

The variogram

- The categorization of time was (fairly) arbitrary and it would be good if one could investigate correlation as a function of continuous time

- Towards this, consider the *autocorrelation function* for the residuals:

$$\rho(u) \;=\; \text{Cor}[R(t),\; R(t-u)]$$

  ⋆ correlation between time points that are u units apart

- The empirical autocorrelation function for the CD4+ cell count residuals is:

u	1	2	3	4	5	6
$\hat{\rho}(u)$	0.57	0.52	0.44	0.41	0.41	0.89
# pairs	2,639	1,878	1,271	791	176	9

  ⋆ assuming stationarity one can pool across observation pairs that differ by u units in time

- The autocorrelation function is most effective for studying equally spaced data that are (roughly) stationary
  - ⋆ estimation relies on a categorization of time to form pairs

- An alternative function that describes association among repeated observations is the *variogram*
  - ⋆ as we'll see it is also easily estimated from irregularly spaced data

- For a stochastic process $R(t)$, the variogram is defined as

$$\gamma(u) \;=\; \frac{1}{2}\mathsf{E}\left[\{R(t) - R(t-u)\}^2\right], \qquad u > 0$$

- When $R(t)$ is stationary, we have:

$$\gamma(u) \;=\; \sigma^2\{1 - \rho(u)\}$$

  - ⋆ σ^2 is the variance of $R(t)$
  - ⋆ $\rho(u)$ is the autocorrelation function

- In the longitudinal context, the *sample variogram* can be calculated by smoothing the observed half-squared differences between pairs of residuals:

$$\hat{\gamma}_{k,ij} \;=\; \frac{1}{2}(R_{ki} - R_{kj})^2$$

against the corresponding time-differences:

$$u_{k,ij} \;=\; |t_{ki} - t_{kj}|$$

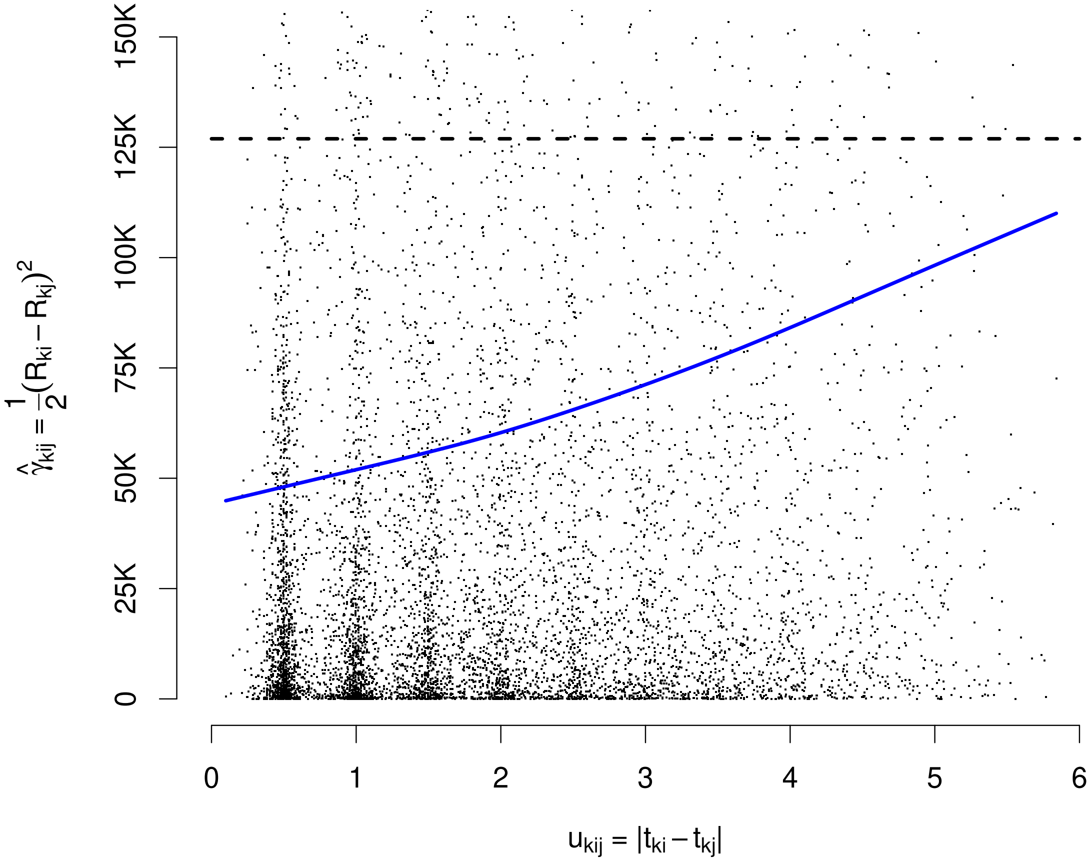
  - ⋆ t_{ki} is the time at which the i^{th} study unit on the k^{th} cluster is observed

- As we'll see, the sample variogram can be used to investigate different sources of variation in the residuals

- Before doing so it useful to characterize the *total variation*, σ^2:
  - ⋆ can be estimated using pairs of residuals <u>across</u> clusters:

$$\hat{\sigma}^2 \;=\; \text{mean}\left\{\frac{1}{2}(R_{ki} - R_{lj})^2\right\}$$

- Sample variogram for the MACS CD4+ cell count data:
  - ⋆ solid line indicates a smoothed trend
  - ⋆ dashed line indicates $\hat{\sigma}^2 = 126{,}927.4$
  - ⋆ $\max(\hat{\gamma}_{k,ij}) = 3{,}001{,}000$, so the y-axis has been truncated



- Certain features of this sample variogram have intuitive interpretations in the context of the following model:

$$Y_{ki} \;=\; X_{ki}^T\boldsymbol{\beta} \;+\; \gamma_k \;+\; W_k(t_{ki}) \;+\; \epsilon_{ki}$$

- This model contains three sources of random variation:

1. γ_k, a *subject-specific random effect*
   - ∗ captures variation between subjects
   - ∗ indicates some ‘trait’ that is specific to the subject
2. $W_k(t_{ki})$, a term that captures *serial correlation*
   - ∗ variation due to some underlying time-varying stochastic process
   - ∗ describes the current ‘state’
3. ϵ_{ki}, a standard *measurement error* term
   - ∗ usual sources of residual ‘noise’

- If we assume that $\gamma_k \sim N(0,\, \sigma^2_\gamma)$, that the serial dependence follows an autoregressive structure such that

  $$\text{Cov}[W_k(t_{ki}), W_k(t_{kj})] = \sigma^2_W \times \rho^{|t_{ki}-t_{kj}|},$$

  and, finally, that $\epsilon_{ki} \sim N(0,\, \sigma^2_\epsilon)$, then one can show that

  $$\text{V}[Y_{ki}|\boldsymbol{\beta}] \;=\; \sigma^2 \;=\; \sigma^2_\gamma + \sigma^2_W + \sigma^2_\epsilon.$$

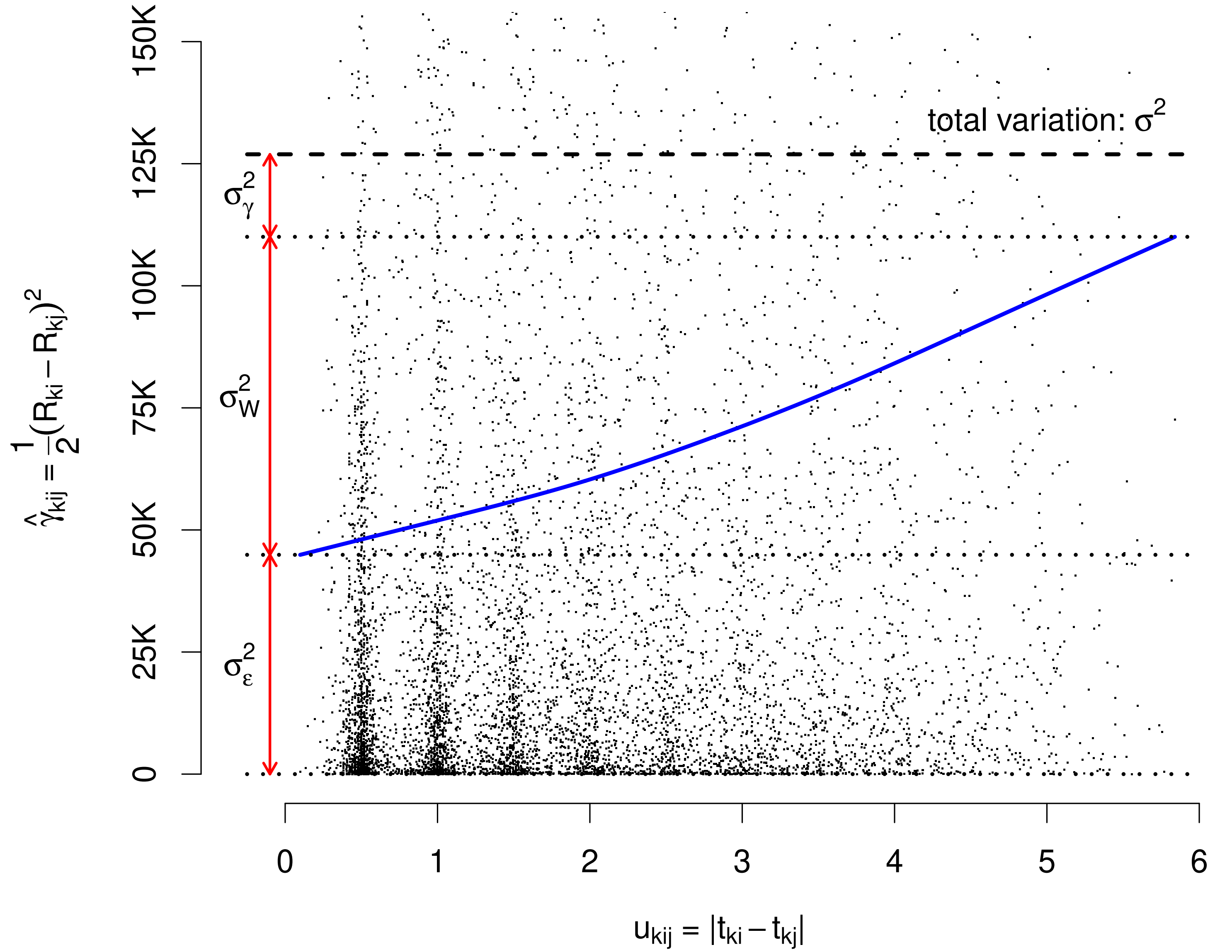
  ⋆ hence, under these assumptions, the total variation can be naturally broken down into contributions from the three components

- One can also show that:

  $$\begin{aligned} \gamma(u_{k,ij}) \;&=\; \frac{1}{2}\text{E}\left[\{R_{ki} - R_{kj}\}^2\right] \\ &=\; \sigma^2_W(1-\rho^{|u_{k,ij}|}) \;+\; \sigma^2_\epsilon \end{aligned}$$

  so that $\gamma(u_{k,ij}) \longrightarrow \sigma^2_\epsilon$ as $u_{k,ij} \longrightarrow 0$

- Given these results, we can visualize each of these components in the sample variogram:



Summary

- Data examples
  - ⋆ dental growth data from UNC
  - ⋆ CD4+ cell count data from MACS
  - ⋆ outcomes among patients diagnosed with pancreatic cancer

- Benefits of dependent data
  - ⋆ expands the range of questions that one can address
  - ⋆ exploit correlation to get efficiency gains

- Exploratory analyses
  - ⋆ missing data and unusual observations/outliers
  - ⋆ preliminary exploration of the mean structure
  - ⋆ preliminary exploration of the dependence structure