Models for Dependent Responses

All of the models that we have seen so far have this form:

$$\underline{Y} \sim N(X\underline{b}, \sigma^2 I)$$

We have worked out some theory for estimating $\underline{b}$ and σ^2

We have considered various variations on what can go in X:

- numeric covariates, factors, both, interactions

Key assumptions are $\mathrm{Var}(Y_i) = \sigma^2$ and $\mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j$

There are important and useful models that relax these assumptions:

Examples:

① $Y_i = b_0 + \sum_{j=1}^{p} x_{ij} b_j + z_i \varepsilon_i$

$z_i > 0$ (observed), and $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$

② $Y_i = b_0 + B_{j(i)} + \varepsilon_i$

$j(i)$ = factor level, $B_1, \dots, B_J \overset{ind}{\sim} N(0, \sigma_1^2)$

$\varepsilon_1, \dots, \varepsilon_n \overset{ind}{\sim} N(0, \sigma_2^2)$

Both models can be written in the more general form

$$\underline{Y} \sim N(X\underline{b}, \Sigma_\theta)$$

① $\Sigma_\theta = \sigma^2 \mathrm{Diag}(z_1^2, \dots, z_n^2)$ $\qquad \theta = \sigma_2$

② $\Sigma_\theta = \sigma_1^2 M + \sigma_2^2 I$ (we will derive M later) $\qquad \theta = (\sigma_1^2, \sigma_2^2)$

Maximum Likelihood Estimation

Recall that the joint density for $\underline{Y} \sim N(X\underline{b}, \Sigma_\theta)$ is

$$p(\underline{z}) = \frac{1}{(2\pi)^{n/2}} \det(\Sigma_\theta)^{1/2} \exp\left(-\frac{1}{2}(\underline{z} - X\underline{b})^T \Sigma_\theta^{-1} (\underline{z} - X\underline{b})\right)$$

The joint density has the property that it integrates to 1

$$\int_{\mathbb{R}^n} p(\underline{z})\, d\underline{z} = 1$$

The likelihood function is the joint density, evaluated at the data y, viewed as a function of the parameters

$$L(\underline{b}, \underline{\theta}) = p(\underline{y}) = \frac{1}{(2\pi)^{n/2}} \det(\Sigma_\theta)^{1/2} \exp\left(-\frac{1}{2}(\underline{y} - X\underline{b})^T \Sigma_\theta^{-1} (\underline{y} - X\underline{b})\right)$$

The maximum likelihood estimates (MLE) are the parameters that maximize the likelihood

We often work with the loglikelihood:

$$\ell(\underline{b}, \theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det(\Sigma_\theta) - \frac{1}{2}(\underline{y} - X\underline{b})^T \Sigma_\theta^{-1} (\underline{y} - X\underline{b})$$

<u>Example:</u> $Y_i = \sum_{j=0}^{p} b_j x_{ij} + z_i \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$

$$\underline{Y} \sim N(X\underline{b}, \Sigma_\theta)$$

$$\Sigma_\theta = \operatorname{diag}(\sigma^2 z_1^2, \sigma^2 z_2^2, \ldots, \sigma^2 z_n^2) = \sigma^2 \begin{bmatrix} z_1^2 & & 0 \\ & z_2^2 & \\ 0 & & \ddots & z_n^2 \end{bmatrix}$$

– $Var(Y_i) = z_i^2 \sigma^2$ and $Cov(Y_i, Y_k) = 0$

Multivariate normal log likelihood

$$\ell(\underline{b}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma_\theta - \frac{1}{2}(\underline{y} - X\underline{b})^T \Sigma_\theta^{-1} (\underline{y} - X\underline{b})$$

$$\det(\Sigma) = \sigma^{2n} \prod_{i=1}^{n} z_i^2, \qquad \log\det(\Sigma) = n\log\sigma^2 + \sum_{i=1}^{n} \log z_i^2$$

$$\Sigma^{-1} = \frac{1}{\sigma^2} \begin{bmatrix} z_1^{-2} & & 0 \\ & z_2^{-2} & \\ 0 & \ddots & z_n^{-2} \end{bmatrix}$$

$$(\underline{y} - X\underline{b})^T \Sigma_\theta^{-1} (\underline{y} - X\underline{b}) = \frac{1}{\sigma^2}(\underline{y} - X\underline{b})^T \begin{bmatrix} z_1^{-2} & 0 \\ 0 & \ddots \; z_n^{-2} \end{bmatrix} (\underline{y} - X\underline{b})$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{1}{z_i^2} \left(y_i - \sum_{j=0}^{p} x_{ij} b_j \right)^2$$

$$\ell(\underline{b}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log z_i^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{1}{z_i^2} \left(y_i - \sum_{j=0}^{p} x_{ij} b_j \right)^2$$

maximizing $\ell(\underline{b}, \sigma^2)$ w.r.t $\underline{b}$ means minimizing

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{1}{z_i^2}\left(y_i - \sum_{j=0}^{p} x_{ij} b_j\right)^2$$

weighted least squares

Taking derivative of $\ell(\underline{b}, \sigma^2)$ w.r.t σ^2 and setting $= 0$:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{z_i^2}\left(y_i - \sum_{j=0}^{p} x_{ij} \hat{b}_j\right)^2$$

why this makes sense:

$$E\left(\left(Y_i - \sum_{j=0}^{p} b_j x_{ij}\right)^2\right) = Var(Y_i) = z_i^2 \sigma^2$$

$$E\left(\frac{1}{z_i^2}\left(Y_i - \sum_{j=0}^{p} b_j x_{ij}\right)^2\right) = \frac{Var(Y_i)}{z_i^2} = \sigma^2$$

$\hat{\sigma}^2$ is the average of $\frac{1}{z_i^2}\left(y_i - \sum_{j=0}^{p} x_{ij} b_j\right)^2$

$\underline{Y} \sim N(X\underline{b}, \Sigma_\theta)$ observe y, Σ_θ depends on parameter θ.

$$\ell(\underline{b}, \theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\det\Sigma_\theta - \frac{1}{2}(y - X\underline{b})^T \Sigma_\theta^{-1}(y - X\underline{b})$$

$$\frac{\partial \ell(\underline{b}, \theta)}{\partial b_k} = -\frac{1}{2}\frac{\partial}{\partial b_j}\left(\sum_{i=1}^{n}\sum_{\ell=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}b_j\right)\left(y_\ell - \sum_{j=0}^{p} x_{\ell j}b_j\right)(\Sigma_\theta^{-1})_{i\ell}\right)$$

$$= +\frac{1}{2}\sum_{i=1}^{n}\sum_{\ell=1}^{n}(\Sigma_\theta^{-1})_{i\ell}\left(y_i - \sum_{j=0}^{p} x_{ij}b_j\right)x_{\ell k}$$

$$+\frac{1}{2}\sum_{i=1}^{n}\sum_{\ell=1}^{n}(\Sigma_\theta^{-1})_{i\ell}\left(y_\ell - \sum_{j=0}^{p} x_{\ell j}b_j\right)x_{ik}$$

$$= \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}b_j\right)\sum_{\ell=1}^{n}(\Sigma_\theta^{-1})_{i\ell}\, x_{\ell k}$$

$$= \sum_{i=1}^{n}(y - X\underline{b})_i(\Sigma_\theta^{-1}X)_{ik} = \left[(y - X\underline{b})^T\Sigma_\theta^{-1}X\right]_k$$

$$\begin{bmatrix} \frac{\partial \ell(\underline{b}, \theta)}{\partial b_0} \\ \vdots \\ \frac{\partial \ell(\underline{b}, \theta)}{\partial b_p} \end{bmatrix} = X^T\Sigma_\theta^{-1}(y - X\underline{b})$$

$$X^T\Sigma_\theta^{-1}(y - X\hat{\underline{b}}) = 0 \qquad X^T\Sigma_\theta^{-1}y = X^T\Sigma_\theta^{-1}X\,\underline{b}$$

$$\boxed{\hat{\underline{b}} = (X^T\Sigma_\theta^{-1}X)^{-1}X^T\Sigma_\theta^{-1}y}$$

generalized least squares estimate

sometimes need to use a numerical optimization procedure to find $\hat{\theta}$. (e.g. Newton-Raphson, BFGS, etc.)

A Random effects Model

factor : J levels

model : $Y_i = b_0 + B_{j(i)} + \varepsilon_i$

$B_1, \dots, B_J \overset{iid}{\sim} N(0, \theta_1^2)$

$\varepsilon_i \overset{iid}{\sim} N(0, \theta_2^2)$, $B_j \perp\!\!\!\perp \varepsilon_i \;\; \forall j, i$

$\underline{Y} = (Y_1, \dots, Y_n)^T$ is MVN. need to get exp. and cov.

$$E(Y_i) = E\left(b_0 + B_{j(i)} + \varepsilon_i\right) = b_0 + E(B_{j(i)}) + E(\varepsilon_i) = b_0$$

$$\Rightarrow E(\underline{Y}) = b_0 \mathbb{1}$$

$$Cov(Y_i, Y_k) = E\left[\left(b_0 + B_{j(i)} + \varepsilon_i - b_0\right)\left(b_0 + B_{j(k)} + \varepsilon_k - b_0\right)\right]$$

$$= E\left[\left(B_{j(i)} + \varepsilon_i\right)\left(B_{j(k)} + \varepsilon_k\right)\right]$$

$$= E\left(B_{j(i)} B_{j(k)}\right) + E\left(B_{j(i)} \varepsilon_k\right) + E\left(\varepsilon_i B_{j(k)}\right) + E\left(\varepsilon_i \varepsilon_k\right)$$

$$= \begin{Bmatrix} \theta_1^2 & \text{if } j(i) = j(k) \\ 0 & \text{if } j(i) \neq j(k) \end{Bmatrix} + 0 + 0 + \begin{Bmatrix} \theta_2^2 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{Bmatrix}$$

There are 3 possibilities:

1. $i=k \Rightarrow j(i)=j(k) \Rightarrow Cov(Y_i, Y_k) = \sigma_1^2+\sigma_2^2$
2. $i\neq k$ but $j(i)=j(k) \Rightarrow Cov(Y_i, Y_k) = \sigma_1^2$
3. $j(i)\neq j(k) \Rightarrow i\neq k \Rightarrow Cov(Y_i, Y_k) = 0$

$$Cov(Y_i, Y_k) = \begin{cases} \sigma_1^2+\sigma_2^2 & \text{if } i=k \\ \sigma_1^2 & \text{if } i\neq k,\ j(i)=j(k) \\ 0 & \text{if } i\neq k,\ j(i)\neq j(k) \end{cases}$$

The random effects model looks very similar to the one-factor model, and it will often give similar results, but it is different.

one factor model:

$$Y_i = b_0 + b_{j(i)} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$$

$b_0, b_1, \ldots, b_J$ are numbers (no assumptions)

Random effects model

$$Y_i = b_0 + B_{j(i)} + \varepsilon_i$$

$\varepsilon_i \overset{iid}{\sim} N(0,\sigma_2^2)$

$B_1, \ldots, B_J \overset{iid}{\sim} N(0,\sigma_1^2)$ (follow a normal dist).

The random effects model is the same as the one factor model aside from the fact that it places a <u>stricter</u> assumption on the effects, i.e. a normal distribution.

When should you use the random effects model vs the one-factor non-random effects (fixed effects) model?

You will hear a lot of <u>folklore</u> about this:

✗ when the effects are random.

✗ when the effects come from a population.

✗ when you don't care about the values of the effects

Ignore the <u>folklore</u>. Treat the normal assumption like any other assumption. You should impose an assumption when:

✓ you think the assumption is close to correct

✓ you think the assumption will help the analysis.

For example: factor = patient. Each patient takes 3 systolic blood pressure measurements.

$$Y_i = b_0 + B_{j(i)} + \varepsilon_i$$

we want to estimate the distribution of BP for a new patient, not in our sample.

Fixed effects model offers no help

$\hat{b}_0 + \hat{b}_1, \ldots, \hat{b}_0 + \hat{b}_J$ says nothing about $b_0 + b_{J+1}$.

Random effects model says that $b_0 + B_{J+1}$ has the same distribution as $b_0 + B_1, \ldots, b_0 + B_J$.

$$\Rightarrow b_0 + B_{J+1} \mathrel{\hat{\sim}} N(\hat{b}_0, \hat{\sigma}_1^2)$$

Remember, though, that "random" $\neq$ "normal"

If the effects are not normal, assuming normal can hurt you.

Two-factor random effects models

Factor 1 : $j(i) \in \{1, \ldots, J\}$

Factor 2 : $k(i) \in \{1, \ldots, K\}$

$Y_i = b_0 + B_{j(i)} + C_{k(i)} + \varepsilon_i$

$B_1, \ldots, B_J \overset{ind}{\sim} N(0, \sigma_1^2)$

$C_1, \ldots, C_K \overset{ind}{\sim} N(0, \sigma_2^2)$

$\varepsilon_1, \ldots, \varepsilon_n \overset{ind}{\sim} N(0, \sigma_3^2)$

"Nested" random effects models

Dataset on student achievement

i	school $j(i)$	classroom $k(i)$	school-class $\ell(i)$	score y_i
1	1	1	1	72
2	1	1	1	84
3	1	1	1	63
4	1	2	2	95
5	1	2	2	78
6	2	1	3	79
7	2	1	3	90
8	2	2	4	74
9	2	2	4	85
10	2	2	4	89
⋮	⋮	⋮	⋮	⋮

This model does not make sense. Why?

$$Y_i = b_0 + B_{j(i)} + C_{k(i)} + \varepsilon_i$$
$$B_j \overset{ind}{\sim} N(0, \sigma_1^2), \quad C_k \overset{ind}{\sim} N(0, \sigma_2^2), \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma_3^2)$$

This one is better:

$$Y_i = b_0 + B_{j(i)} + C_{j(i),k(i)} + \varepsilon_i$$
$$B_j \overset{ind}{\sim} N(0, \sigma_1^2), \quad C_{j,k} \overset{ind}{\sim} N(0, \sigma_2^2), \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma_3^2)$$

Classroom 1 in school 1 is different from classroom 1 in school 2.

This is simply a labeling problem. If we had given each classroom a unique id, $(1, 2, 3, 4, \ldots)$ then the first model would be OK (actually both models are the same)

Random intercepts and slopes model

response y_i, factor $j(i)$, Numeric covariate x_i.

Model: $Y_i = b_0 + b_1 x_i + C_{j(i)} + D_{j(i)} x_i + \varepsilon_i$

$C_1, \dots, C_J \overset{ind}{\sim} N(0, \sigma_1^2) \quad D_1, \dots, D_J \overset{ind}{\sim} N(0, \sigma_2^2) \quad \varepsilon_1, \dots, \varepsilon_n \overset{ind}{\sim} N(0, \sigma_3^2)$

Useful when you have repeated measurements for subject j at different values of the numeric covariate.

Interpretations: b_0 = expected intercept

b_1 = expected slope

$b_0 + C_j$ = intercept for level j.

$b_1 + D_j$ = slope for level j.

Other random effects models

Any model we have studied can be turned into a random effects model by placing a normal assumption on coefs

Your imagination is the limit

Remember to always think carefully about the assumptions you impose

Estimation for random effects models

use lmer function in lme4 package (also nlme, regress + others)

data vectors: y, fac1, fac2, num1, num2

lmer(y ~ 1 + (1 | fac1)) random effect in fac1

lmer(y ~ fac1 + (1 | fac2)) nonrandom in fac1

random in fac2

lmer(y ~ num1 + (1 | fac1)) random intercepts

lmer(y ~ num1 + (num1 | fac1)) random int + random slope.

what does lmer do? Max lik or Res Max Lik (REML)

Exercises this week: work out covariance matrices for some random effects models.

Time Series Models

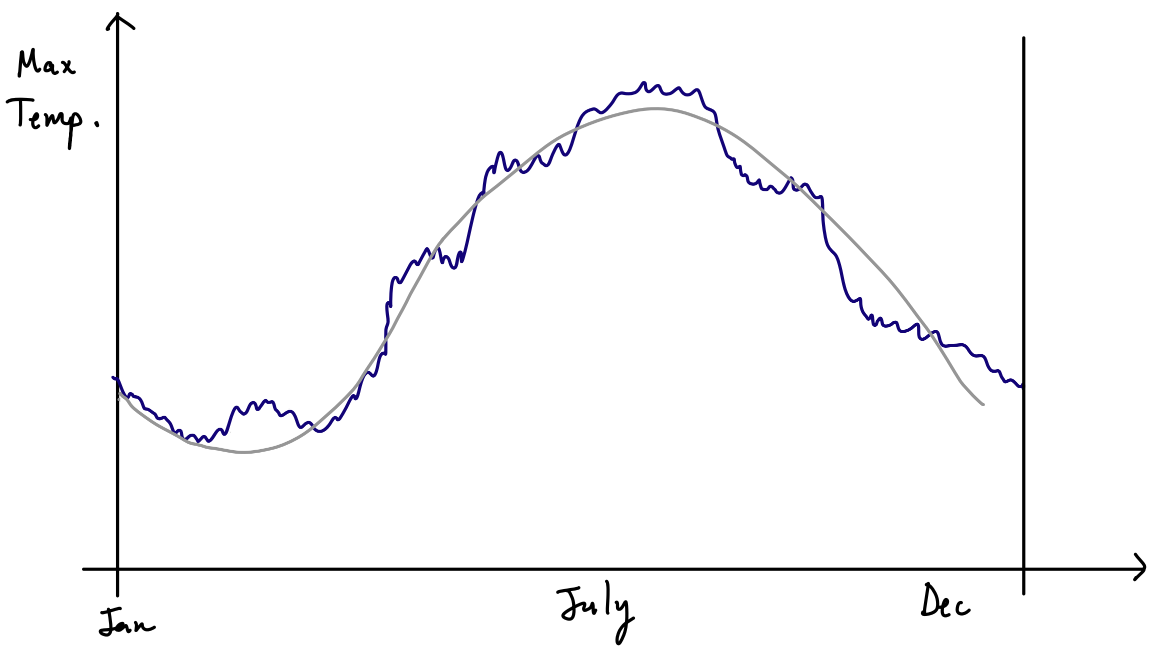
Remember the ithaca weather data.

y_i = maximum temp on day i.

Model 1: $Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$$x_{i1} = \cos\left(\frac{2\pi i}{365}\right) \qquad x_{i2} = \sin\left(\frac{2\pi i}{365}\right)$$

critique: model implies $Y_i \perp\!\!\!\perp Y_{i+1}$



Moving average model:

$\varepsilon_0, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$

$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \rho \varepsilon_{i-1} + \varepsilon_i$

$\underline{Y} = (Y_1, \ldots, Y_n)^T$ is MVN

$$E(Y_i) = b_0 + b_1 x_{i1} + b_2 x_{i2}$$

$$Var(Y_i) = \rho^2 \sigma^2 + \sigma^2 = (\rho^2 + 1)\sigma^2$$

$$\begin{aligned}
Cov(Y_i, Y_{i-1}) &= E\left([\rho \varepsilon_{i-1} + \varepsilon_i][\rho \varepsilon_{i-2} + \varepsilon_{i-1}] \right) \\
&= \rho^2 E(\varepsilon_{i-1}\varepsilon_{i-2}) + \rho E(\varepsilon_{i-1}^2) + \rho E(\varepsilon_i \varepsilon_{i-2}) + E(\varepsilon_i \varepsilon_{i-1}) \\
&= 0 + \rho\sigma^2 + 0 + 0 = \rho\sigma^2
\end{aligned}$$

$Cov(Y_i, Y_j) = 0$ if $|i-j| > 1$ ⟵ exercise

$$\underline{Y} \sim N(X\underline{b}, \Sigma_\theta)$$

$$\Sigma_\theta = \sigma^2 \begin{bmatrix} \rho^2+1 & \rho & 0 & \cdots & 0 \\ \rho & \rho^2+1 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \rho \\ 0 & \cdots & 0 & \rho & \rho^2+1 \end{bmatrix}$$

$\theta = (\sigma^2, \rho)$

use REML to estimate θ

Autoregressive model:

$$E(Y_i) = b_0 + b_1 x_{i1} + b_2 x_{i2}$$

$$Y_i - E(Y_i) = \rho \left(Y_{i-1} - E(Y_{i-1}) \right) + \varepsilon_i$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad 0 < \rho < 1$$

The deviation between Y_i and its mean equals a fraction of the deviation between Y_{i-1} and its mean, plus independent error.

$\underline{Y} = (Y_1, \ldots, Y_n)$ is MVN

already said $E(Y_i) = b_0 + b_1 x_{i1} + b_2 x_{i2}$

$$E(\underline{Y}) = X\underline{b}, \quad X = [\mathbb{1} \ \underline{x}_1 \ \underline{x}_2], \quad \underline{b} = (b_0, b_1, b_2)^T$$

$$Var(Y_i) = E\left[(Y_i - E(Y_i))^2 \right]$$

$$= E\left[\rho^2 (Y_{i-1} - E(Y_{i-1}))^2 + \rho(Y_{i-1} - E(Y_{i-1}))\varepsilon_i + \varepsilon_i^2 \right]$$

$$= \rho^2 Var(Y_{i-1}) + \sigma^2$$

if $Var(Y_{i-1}) = Var(Y_i)$ (stationary process)

$$\Rightarrow Var(Y_i) = \frac{\sigma^2}{1-\rho^2}$$

$$\mathrm{Cov}(Y_i, Y_{i-1}) = E\left[(Y_i - E(Y_i))(Y_{i-1} - E(Y_{i-1}))\right]$$

$$= E\left[\left(\rho(Y_{i-1} - E(Y_{i-1})) + \varepsilon_i\right)\left(Y_{i-1} - E(Y_{i-1})\right)\right]$$

$$= \rho E\left[(Y_{i-1} - E(Y_{i-1}))^2\right] + 0$$

$$= \rho \mathrm{Var}(Y_i) = \frac{\rho \sigma^2}{1-\rho^2}$$

Turns out that $\mathrm{Cov}(Y_i, Y_k) = \dfrac{\rho^{|i-k|}\sigma^2}{1-\rho^2}$

$$\Sigma_{\theta, ik} = \frac{\rho^{|i-k|}\sigma^2}{1-\rho^2}$$

<u>Trick</u> for fitting intercept-only autoregressive models

Model: $Y_i - b_0 = \rho(Y_{i-1} - b_0) + \varepsilon_i$

$$Y_i = \underbrace{(1-\rho)b_0}_{a_0} + \underbrace{\rho}_{a_1} Y_{i-1} + \varepsilon_i$$

$$Y_i = a_0 + a_1 Y_{i-1} + \varepsilon_i$$

* Create dataset where first column is y_i, second column is $x_i = y_{i-1}$
* Do regression of y_i on x_i
* solve for b_0 and ρ

Higher-order autoregressive models

$E(Y_i) = b_0 + b_1 x_{i1}$ (or add more covariates)

$Y_i - E(Y_i) = \rho_1 (Y_{i-1} - E(Y_{i-1})) + \rho_2 (Y_{i-2} - E(Y_{i-2})) + \varepsilon_i$

$\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$

Autoregressive model of order 2, AR(2)

We can have AR(m):

$$Y_i - E(Y_i) = \sum_{j=1}^{m} \rho_j \left[Y_{i-j} - E(Y_{i-j}) \right] + \varepsilon_i$$

$\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$

Exercises this week: Fit some autoregressive models to daily temperature data.

The Gaussian Process Model

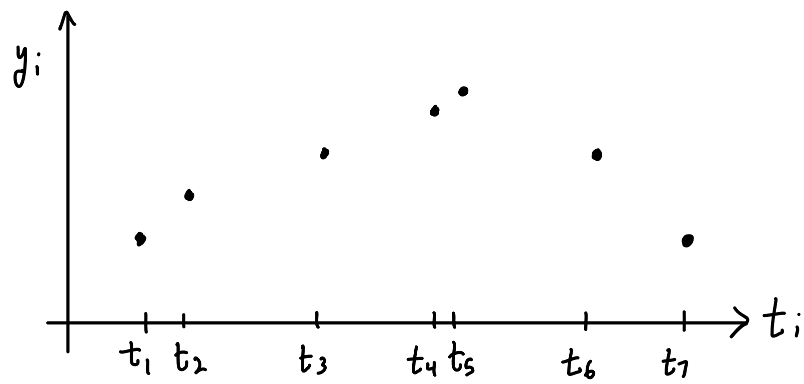
Sounds really fancy and complicated, but it's not

Setup:

Responses : $y_1, \ldots, y_n$

Inputs : $t_1, \ldots, t_n$ (times, spatial locations, etc.)

Covariate : $x_1, \ldots, x_n$



Goals : infer relationship between y and x

interpolate or extrapolate y at new time t_{n+1}

Model for y_i : $Y_i = b_0 + b_1 x_i + Z_i$

b_0, b_1 non-random

$(Z_1, \ldots, Z_n)^T \sim N(\underline{0}, \Sigma)$

$\Rightarrow (Y_1, \ldots, Y_n)^T \sim N(X\underline{b}, \Sigma)$

So far, nothing new here. You've seen this before

The new part is how we get Σ.

Since we need a model for observations at an arbitrary set of locations, our method for constructing Σ should allow us to get $Cov(z_1, z_2)$ for <u>any</u> pair of inputs t_1 and t_2

Need a **covariance function**

$$Cov(z_1, z_2) = K(t_1, t_2)$$

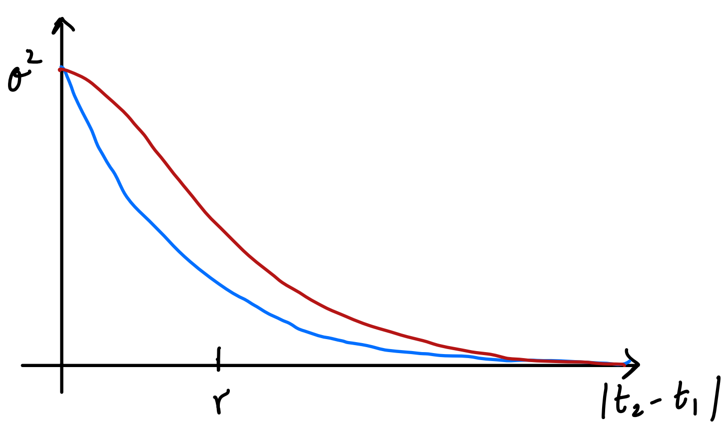
This will allow us to construct Σ from $t_1, \ldots, t_n$ <u>and</u> get $Cov(z_i, z_{n+1})$ for prediction

Can't just pick any old function K. We need a K that is guaranteed to give us a positive definite Σ.

<u>Some examples</u>:

$$K(t_1, t_2) = \sigma^2 \exp(-|t_2 - t_1|/r)$$

$$K(t_1, t_2) = \sigma^2 (1 + |t_2 - t_1|/r) \exp(-|t_2 - t_1|/r)$$



We typically use maximum likelihood to estimate any parameters in the model (e.g. b_0, b_1, σ^2, r)

We use the multivariate normal distribution to make predictions (conditional distributions from multivariate normal)

Active area of research in statistics and machine learning

* How do we construct realistic models?
* How do we do computations for large datasets?