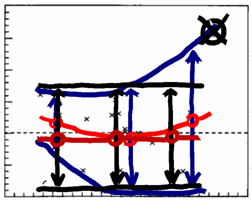
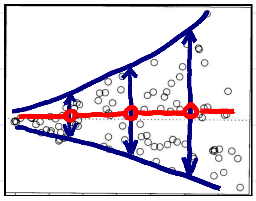
p. 7-9





➢ unfortunately, in real data set, it's rare the pattern is so clear (**Q**: what will you conclude from the residual plot on the right?)

➢ in models with many terms or models with complex non-linear mean structure, cannot necessarily associate shapes in a residual plot with a particular problem with the assumptions, e.g.,

(LNp.7-2)

① true model: $E(Y)=|x_1|/[2+(1.5+x_2)^2]$ with constant variance

② fitted model: $E(Y)=\beta_0+\beta_1x_1+\beta_2x_2$ approximate

nonlinear mean structure

• possible remedies for unsatisfactory residual plots

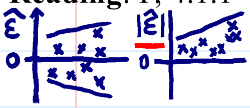
unsatisfactory residual plot	plot residuals against ...		
	$\hat{y}$	x_k	time order
non-constant variance	1. weighted least square 2. transform y	1. weighted least square 2. transform y	weighted least square
curvature in mean structure	1. add extra term 2. transform y	1. add extra term of x_k 2. transform y	add term of time in model

different philosophy

change current model

❖ **Reading**: F, 4.1.1 ❖ **Further reading**: D&S, 2.5

11/24

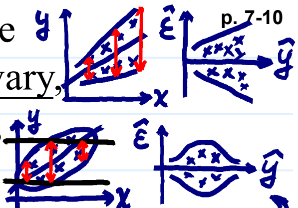


① (LNp.2) Non-constant variance — overall pattern

• if not sure, plot absolute values of residuals against $\hat{y}$, x_k's, time order

• when non-constant variance exists, $\hat{\beta}_{OLS}$ will be more variable than the best estimates ($\hat{\beta}_{WLS}$, $\hat{\beta}_{GLS}$) ($\hat{\beta}_{OLS}$ unbiased but not BLUE) and $\hat{\sigma}$ wrong ($\Rightarrow$ test and C.I. inaccurate)

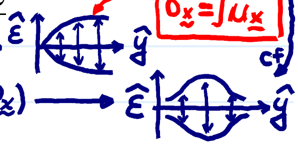
p. 7-10



• It's better try to understand the cause of non-constant variance before taking any remedies, e.g., (1) larger response have more "room" to vary, (2) response constrained to lie between a maximum and a minimum, (3) response from Poisson distribution or binomial distribution, …

count data ① ②

$\Rightarrow$ discovering reasons to support the remedies you are going to take



• remedies for non-constant variance

① $Y_{\underline{x}} \sim P(\mu_{\underline{x}})$, $E(Y_{\underline{x}})=Var(Y_{\underline{x}}) \Rightarrow \mu_{\underline{x}}=\sigma_{\underline{x}}^2$ $\sigma_{\underline{x}}=\sqrt{\mu_{\underline{x}}}$ cf.

② $Y_{\underline{x}} \sim Bin(n, p_{\underline{x}})$, $E(Y_{\underline{x}})=np_{\underline{x}}$, $Var(Y_{\underline{x}})=np_{\underline{x}}(1-p_{\underline{x}})$

➢ weighted least squares

▪ need weighting information (may from plotting residual vs. x_k) or

▪ model the form of Σ and using IRWLS ← check LNp 6-5 ($\sigma_{\underline{x}} \leftrightarrow x_k$) cf. $\sigma_{\underline{x}}^2 \propto g(\mu_{\underline{x}})$ g: variance function

➢ transform Y (may use information from plotting residual vs. $\hat{y}$) ($\sigma_{\underline{x}} \leftrightarrow \mu_{\underline{x}}$)

variance stablizing transformation

$Y=X\underline{\beta}+\underline{\varepsilon}$, $h(Y)=X\underline{\beta}'+\underline{\varepsilon}'$

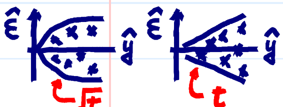
▪ idea: find a transformation h such that $\mathrm{var}(h(y_{\underline{x}}))$ is a constant, (**Q**: how? Hint: δ-method) over $\mu_{\underline{x}}$

$$h(y_x) \approx h(\underbrace{E(y)}_{\mu_x})+(y-E(y))h'(E(y)) \cancel{+\cdots} \Rightarrow \mathrm{var}(h(y)) \approx \mathrm{var}(y)[h'(E(y))]^2=c$$

Taylor expansion

hope var($h(y)$) to be a constant $c \Rightarrow h'(\underbrace{E(y_x)}_{\mu_x}) \propto 1/(\underbrace{\mathrm{var}(y_x)}_{\sigma_x^2})^{1/2}$

$$h(\underbrace{E(y)}_{\mu_x}) \equiv \int 1/(\underbrace{\mathrm{var}(y)}_{g(\mu_x)})^{1/2}\, d(\underbrace{E(y)}_{\mu_x}) = \int \frac{1}{\sqrt{g(\mu_x)}}\, d\mu_x \quad \propto \frac{1}{\sqrt{g(\mu_x)}}$$



$g(t)=t$ $g(t)=t^2$

v ($\sigma_{\underline{x}} \leftrightarrow \mu_{\underline{x}}$)

x ($\sigma_{\underline{x}}^2 \leftrightarrow \mu_{\underline{x}}$)

$g(t)=t^2$ — $\sigma_x^2 \propto \mu_x^2$

▫ Example 1: $\mathrm{var}(y_x) \propto [E(y_x)]^2 \Rightarrow$ suggest $h(y)=\log(y)$ $h(t)=\int \frac{1}{\sqrt{t^2}}dt = \ln(t)$

$g(t)=t$ — $\sigma_x^2 \propto \mu_x$

▫ Example 2: $\mathrm{var}(y_x) \propto E(y_x) \Rightarrow$ suggest $h(y)=y^{1/2}$ $h(t)=\int \frac{1}{\sqrt{t}}dt \propto t^{1/2}$

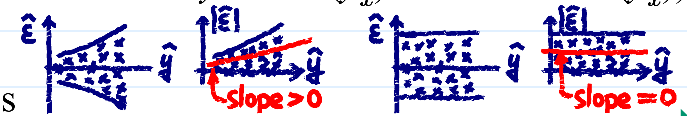
▪ Note: in residual plot, tend to see $[\mathrm{var}(y_x)]^{1/2}$ rather than $\mathrm{var}(y_x)$ (example?)

- practical problems:
  - if $y_i \le 0$, for some i, square root or log transformations fail ⇒ can do transformation on y_i+d, where d is some small amount s.t. $y_i+d>0$ for all i
  - transformation may make interpretation difficult — interpretation of $\hat{\beta}'$ (unit=?) cf. interpretation of $\hat{\beta}$
- example of transformations ← Weisberg (2005). Applied Linear Regression, Sec. 8.3.1

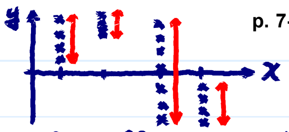
sqrt(y_i)	var(y_i) ∝ E(y_i)^{1}	useful for count data from Poisson distribution
log(y_i) — $Y\|x$ — ratio is large — y_x — log(y_x)	var(y_i) ∝ [E(y_i)]^2	very common, good candidate if the range of Y is very broad, say y_x's value ranges from 10 to $10^5, 10^6$ — σ_x/μ_x = a constant — coefficient of variation
$1/y_i$ — t^2 — $\hat{\varepsilon}$ — $\hat{y}$	var(y_i) ∝ [E(y_i)]^4 — $g(t)=t^4$ — t — $R(t)=\int \frac{1}{\sqrt{t^4}}dt = \int \frac{1}{t^2}dt \propto t^{-1}$ — most data	appropriate when responses are "bunched" near zero, but, in markedly decreasing numbers, large responses do occur — most — very few — y_x — heavy-tailed dist. — very few data
$\sin^{-1}$(sqrt(y_i))	var(y_i) ∝ E(y_i)(1 - E(y_i)) — $g(t)=t(1-t)$ — t	for binomial proportions — $R(t)=\int \frac{1}{\sqrt{t(1-t)}}dt$

- do nothing ⇒ because (i) $\hat{\boldsymbol{\beta}}_{OLS}$ is still unbiased, although not BLUE; (ii) tests and C.I. inaccurate, but bootstrap may be used to get more accurate results
- use generalized linear model (e.g., Poisson/binomial y ⇒ var(y_x): function of E(y_x))

• formal test for non-constant variance
- regressing absolute residuals (or $\hat{\varepsilon}^2$) on $\hat{y}$ or x_k's



- data with replication ⇒ can estimate variances of distinct $\boldsymbol{x}_i$'s and test their homogeneity (see D&S, 2.2)
  $H_0: \sigma^2_{x_1} = \sigma^2_{x_2} = \cdots = \sigma^2_{x_n}$ ← Bartlett's test



(Recall. $\hat{\sigma}^2_{PE}$ in LNp.6-9)

- data without replication ⇒ assign variance a model, test whether parameters in the model equal zero (see Weinberg (2005), 8.3.2) — e.g. $\sigma_z^2 = \sigma^2 \cdot \exp(\lambda z)$ — parameters
- formal test may be good at detecting a particular kind of non-constant variance $\lambda \overset{?}{=} 0$ (depending on the alternative hypothesis), but always do the residual plots — test H_0 vs. H_1, $H_0 \cup H_1$ = all possible models

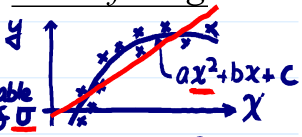
❖ **Reading**: Faraway (2005, 1st ed.), 4.1.1 ❖ **Further reading**: D&S, 2.2, 13.6

② (LNp.2) → Curvature in the mean of residuals ← overall pattern

• related to the concept of lack-of-fit (tests for lack-of-fit can be used if possible), i.e., the current model, E($\boldsymbol{Y}$)=$\boldsymbol{X\beta}$, may need to be modified for *achieving better fitting* — too simple

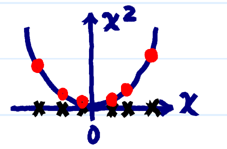
Weisberg (2005). Sec. 8.2

⊙ A simple test for curvature: test whether a plot of residuals versus a quantity $\boldsymbol{U}$ (e.g., $\hat{y}$ or x_k's) is a null plot or has curvature (can be any reasonable known function of U)

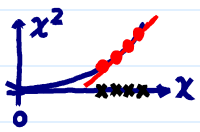


⇒ refit the original mean structure with an additional term $\boldsymbol{U}^2$ added

⇒ significant t-test for $\boldsymbol{U}^2$ suggests curvature (be aware of collinearity between $\boldsymbol{U}^2$ and other terms in original mean structure)



• **Q**: how to identify *why* the non-linearity happened? high cor(x, x^2) → ⇒ can use orthogonal polynomial to remove collinearity (future lecture)
- plot residuals against $\hat{y}$ ⇒ can tell you whether some problems exist, but cannot tell you why



possible variables/effects to be added into model matrix X

➢ plot residuals against x_k's or y against x_k's ⇒ may tell you why this problem happened, but in multivariate regression there may exist correlation between predictors, then it's difficult to find why

e.g. $cor(x_k, x_i) \approx 1$, x_i already appears in the model matrix X. Then, $\hat{\varepsilon} \perp x_i$ and $cor(\hat{\varepsilon}, x_k) \approx 0$

➢ *added variable* (*partial regression*) plots

$\hat{\varepsilon}_Y(x_k) = \beta_2 \hat{\varepsilon}_{x_k} + \varepsilon$

Fitted model: $Y = X_1\beta_1 + \varepsilon$

- recall:

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon = (X_1\beta_1 + H_1X_2\beta_2) + ((I-H_1)X_2\beta_2 + \varepsilon)$$

1. regress Y on all predictors except x_k ⇒ get residuals $\hat{\varepsilon}_Y(x_k)$
2. regress x_k on all predictors except x_k ⇒ get residuals $\hat{\varepsilon}_{x_k}$

remove collinearity btw x_k & the other predictors

□ $\hat{\varepsilon}_Y(x_k)$: part of Y not explained by all predictors except x_k

□ $\hat{\varepsilon}_{x_k}$: part of x_k not explained by all predictors except x_k

3. plot $\hat{\varepsilon}_Y(x_k)$ versus $\hat{\varepsilon}_{x_k}$

Fit simple linear regression: $\hat{\varepsilon}_Y(x_k) = \beta_0 + \beta_k \hat{\varepsilon}_{x_k} + \varepsilon$

the estimated coefficient of x_k when x_k is included in model matrix

- the slope of a fitted line to the added variable plot is $\hat{\beta}_k$ and intercept=0 (the line passes (0, 0))

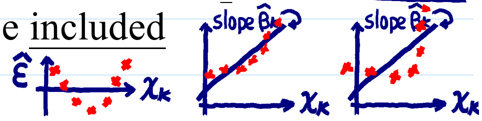
check LNp.7-2

$\because \sum \hat{\varepsilon}_Y(x_k) = 0$, $\sum \hat{\varepsilon}_{x_k} = 0$ if X_1 contains the intercept

- a strong relationship between the plotted quantities corresponds to a strong adjusted relationship between y and x_k

Fitted model: $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$

- can be used to check if new predictors should be included



➢ partial residual plots

fitted linear line (Note: $\hat{\varepsilon} \perp x_k$): slope $\hat{\beta}_k$, pass $(\bar{x}_k, \hat{\beta}_k \bar{x}_k)$

- plot $\hat{\varepsilon} + \hat{\beta}_k x_k$ versus x_k ⇒ same interpretation as added variable plots
- $y - \Sigma_{j \neq k} \hat{\beta}_j x_j = \hat{y} + \hat{\varepsilon} - \Sigma_{j \neq k} \hat{\beta}_j x_j = \hat{\varepsilon} + \hat{\beta}_k x_k$

can use Lowess (future lecture) to find how x_k affect y

- remedies for curvature ⇒ adjust the mean structure, E(Y)=$X\beta$, for better fitting

➢ many many modeling techniques in addition to linear regression can be adopted (GLM, additive model, nonparametric regression, ACE, AVAS, regression trees, regression spline, MARS)

link function

additivity & variance stabilization

multivariate adaptive regression spline

alternating coditional expectation

Hastie & Tibshirani (1990). Generalized additive models

➢ add more (polynomial or cross product) terms

interaction

- may identify required terms from residual plot, added variable plot, or partial residual plot (polynomial model will be introduced in further lecture)

➢ transformation of response or predictors. idea behind the approach:

Linear model use linear structure $X\beta$ to approximate the functional relationship between $y_x \leftrightarrow x$ 規律

(i) a statistical model is a local approximation of the underlying system

(ii) when the mean structure of the underlying system is non-linear and complex, a linear approximation over a relatively wide range of X may be inadequate (e.g.,

a good fitted model would require a lot of effects (e.g., check LNp 7-9)

$$E(Y) = \beta_0 x_1^{\beta_1} x_2^{\beta_2}) \rightarrow \log(E(Y)) = \log(\beta_0) + \beta_1 \log(x_1) + \beta_2 \log(x_2)$$

$E[\log(Y)]$ ≠

cf. variance stablizing transformation (LNp.10)

(iii)we sometimes can find suitable transformations of data that will permit a non-linear model to be better approximated (after transformation) by a linear one (e.g.,

use fewer terms

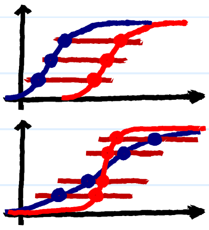
$$E(\log(Y)) \approx \log(\beta_0) + \beta_1 \log(x_1) + \beta_2 \log(x_2))$$

a model with only a few terms

$\frac{z_1-\mu_1}{\sigma_1} \sim \frac{z_2-\mu_2}{\sigma_2} \sim N(0,1)$

If (z_1, z_2) such that $\Phi\left(\frac{z_1-\mu_1}{\sigma_1}\right) = \Phi\left(\frac{z_2-\mu_2}{\sigma_2}\right) \Rightarrow (z_1-\mu_1)/\sigma_1 = (z_2-\mu_2)/\sigma_2$

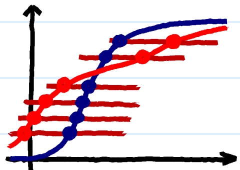
- some examples:



transformation		non-linear model
log(y)	log(x)	$E(y) = \alpha \prod x_j^{\beta_j}$
log(y)	x	$E(y) = \alpha \exp(\Sigma\beta_j x_j)$
y	log(x)	$E(y) = \alpha + \Sigma \beta_j \log(x_j)$
1/y	1/x	$E(y) = 1/[\alpha + \Sigma (\beta_j/x_j)]$
1/y	x	$E(y) = 1/(\alpha + \Sigma \beta_j x_j)$
y	1/x	$E(y) = \alpha + \Sigma \beta_j (1/x_j)$

- There exists numerical method for finding a suitable transformation to improve the fit and/or remedy non-constant variance (e.g., Box-Cox transformation, future lectures)

empirical cdf $\frac{1}{n}$ $\frac{1}{n}$ $z_{(1)}$ $z_{(2)}$ ...



❖ **Reading**: F, 4.3, 7.2.4 ❖ **Further reading**: D&S, 8.2

quantile vs. quantile

Various plots and tests for diagnostics

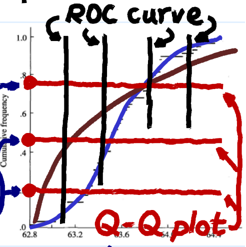
- Q-Q plot

  ➢ **Q**: we often see the statement "$z_1, z_2, \ldots, z_m$ are i.i.d. from a cdf F", how to examine if F is an appropriate distribution assumption for z_i's? (Hint: examine the similarity btw cdf and empirical cdf)

  $z_{(1)}, \cdots, z_{(n)} \leftarrow$ each probability $\frac{1}{n}$

  ROC curve

  $\frac{i}{n+1}$

  Q-Q plot

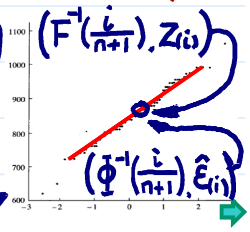
  ➢ normal (probability) plot: assessing normality assumption of ε (Note: tests and C.I. depend on normality assumption)

  slope=1 intercept=0

  $\left(F^{-1}\left(\frac{i}{n+1}\right), z_{(i)}\right)$

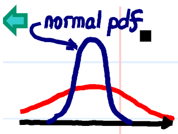
  1. sort the data $\hat{\varepsilon}_{(1)} \le \hat{\varepsilon}_{(2)} \le \cdots \le \hat{\varepsilon}_{(n)}$ $\hat{\varepsilon} \sim N(0, (I-H)\sigma^2)$ if $\varepsilon \sim N(0, \sigma^2 I)$
  2. plot $\hat{\varepsilon}_{(i)}$ against $\Phi^{-1}(i/(n+1))$, where Φ is the cdf of N(0, 1)

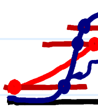
  - If the residuals are normally distributed, an approximately straight-line relationship will be observed (null plot) cf

  $\left(\Phi^{-1}\left(\frac{i}{n+1}\right), \hat{\varepsilon}_{(i)}\right)$

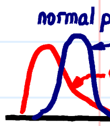
normal pdf



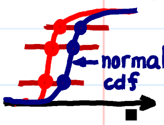


normal cdf

normal pdf

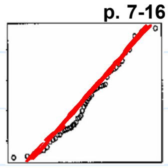


skewed

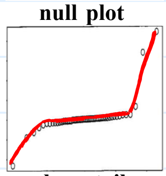


normal cdf

- non-normality: long-tail, short-tail, asymmetric — tend to generate data with large residuals ⇒ can remove outliers then plot again
  - worst case is long-tail; mild non-normality can safely be ignored; the larger the sample size, the less troublesome the non-normality ← Why? CLT for $\hat{\beta}$
  - for long-tail, (i) use test based on other distributions, or bootstrap, or permutation tests (ii) for estimation, use robust methods (e.g., least absolute deviation instead of least square)
  - asymmetric, transform Y (e.g., Box-Cox method)
  - short-tail can be reasonably ignored
- formal tests exists (such as Kolmogorov-Smirnov test), but not as flexible as the Q-Q plot

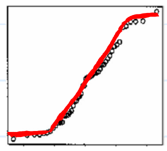


null plot



long tail

asymmetric

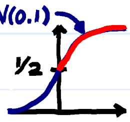


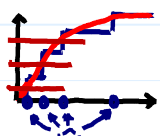
short tail

➢ normal plot can be applied to identify extreme values (e.g., in residuals, → outliers; influential observations ← leverages, Cook's statistics, …): in the case, not interested in a straight line ← overall pattern relationship, but rather looking for points *that depart from the straight line* ← individual pattern

- half-normal plot ← absolute values

  cdf of N(0,1)

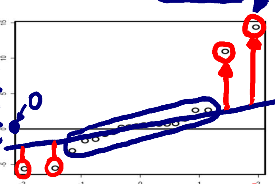
  $|\hat{\varepsilon}|_{(i)}$

  1. sort the absolute data $|\hat{\varepsilon}|_{(1)} \le |\hat{\varepsilon}|_{(2)} \le \cdots \le |\hat{\varepsilon}|_{(n)}$
  2. plot $|\hat{\varepsilon}|_{(i)}$ against $\Phi^{-1}((n+i)/(2n+1))$ ← cdf of $|X|$, where $X \sim N(0,1)$

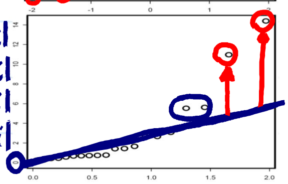
  ➢ usually used to identify "extreme" values

  ➢ can be used to examine residuals, leverages, Cook's statistics, treatment effects (especially for experimental data without replicates) ← $|\hat{\beta}_i|$'s

$\hat{\varepsilon}_i$ R_i C_i



$|\hat{\varepsilon}_i|$ $|R_i|$ $|C_i|$ $|\hat{\beta}_i|$



- diagnostic of correlated errors when a time order is available
  - plot $\hat{\varepsilon}$ against time   lag 2, lag 3, ... ← lag 1
  - plot $\hat{\varepsilon}_{i+1}$ against $\hat{\varepsilon}_i$, when i related to time
  - use formal tests like the Durbin-Watson or runs test

AR(1): $y \sim x$

$x \rightarrow \hat{\varepsilon}_1\ \hat{\varepsilon}_2\ \hat{\varepsilon}_3\ \hat{\varepsilon}_4 \cdots \hat{\varepsilon}_{n-1}\ \hat{\varepsilon}_n \leftarrow y$

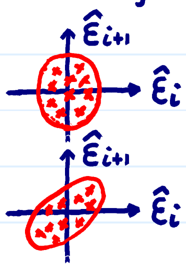
time: $t_1\ t_2\ t_3\ t_4 \cdots t_{n-1}\ t_n$ (equally spaced)

$x_1\ y_1 x_2\ y_2 x_3\ y_3 \quad x_{n-1}\ y_{n-1}$

$\sum(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2 = \sum \hat{\varepsilon}_i^2 + \sum \hat{\varepsilon}_{i-1}^2 - 2\sum \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}$

$$\boxed{DW = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^{n}\hat{\epsilon}_i^2}}$$

Information of Δ



a correlation matrix of ε_i's (y_i's)

block-diagonal correlation matrix (LNp.6-1) cf.

$$\begin{array}{c|cccc} & \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 \cdots \\ \hline \varepsilon_1 & 1 & \Delta & \times & \square \\ \varepsilon_2 & \Delta & 1 & \Delta & \times \\ \varepsilon_3 & \times & \Delta & 1 & \Delta \\ \varepsilon_4 & \square & \times & \Delta & 1 \\ \vdots & & & & \end{array}$$

- $0 \le DW \le 4$
- positively correlated $\Rightarrow DW \to 0$
- negatively correlated $\Rightarrow DW \to 4$
- under null (i.e., correlation=0) $\Rightarrow DW \approx 2$
- null distribution depends on $\boldsymbol{X}$ ← ∵ $\hat{\varepsilon} = (I-H)\varepsilon$ and $\varepsilon \sim N(0, \sigma^2 I)$ under null.

$1 > |\Delta| > |\times| > |\square| > \cdots$

cor ↓ as time duration ↑

- use GLS when you have correlated errors

❖ **Reading**: F, 4.1.2, 4.1.3   ❖ **Further reading**: D&S, 2.4, 2.7, chapter 7