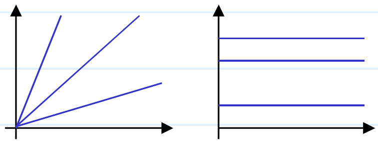
Generalized Least Square (GLS)

- model: $Y=X\beta+\varepsilon$, $E(\varepsilon)=0$ and $\text{var}(\varepsilon)=\sigma^2 I \Rightarrow \varepsilon$: uncorrelated and constant variance

Q: what if $\text{var}(\varepsilon)\neq\sigma^2 I$? ε may have non-constant variance and/or are correlated, e.g.,

- time series correlation [e.g., ε_t~ARMA(r,m)]
- growth curve model, repeated measurement model [e.g., several observations taken from same person, or same unit]



- spatial correlation [e.g., data taken over contiguous geographical areas: census tracts, countries, or states in a country. Nearby areas are often much alike]
- nested errors [e.g., M sets of observations, each set from common production run, from same/common equipment, or from same survey-taker]

- Consider the case $\text{var}(\varepsilon)=\sigma^2\Sigma$, where $\Sigma\,(\neq I)$ is *known* but σ^2 is unknown, i.e., we know the *correlation* and *relative variance* between the errors but we don't know the absolute scale

- Because $\Sigma_{n\times n}$ is symmetric and positive definite, we can write $\Sigma=SS^T$, where S is an $n\times n$ nonsigular matrix (by Cholesky or spectral decompositions)

$$Y=X\beta+\varepsilon \Rightarrow S^{-1}Y=S^{-1}X\beta+S^{-1}\varepsilon \Rightarrow Y'=X'\beta+\varepsilon', \text{ where}$$

$$Y'=S^{-1}Y,\ X'=S^{-1}X,\ \varepsilon'=S^{-1}\varepsilon, \text{ and}$$

$$E(\varepsilon')=0 \text{ and } \text{var}(\varepsilon')=\text{var}(S^{-1}\varepsilon)=S^{-1}\text{var}(\varepsilon)S^{-T}=S^{-1}\sigma^2 SS^T S^{-T}=\sigma^2 I$$

$\Rightarrow$ For Y' and X', the assumption in ordinary least square is satisfied

- GLS: find β that minimize

$$\varepsilon'^T\varepsilon'=(Y'-X'\beta)^T(Y'-X'\beta)=(Y-X\beta)^T S^{-T}S^{-1}(Y-X\beta)=(Y-X\beta)^T\Sigma^{-1}(Y-X\beta)$$

$$\Rightarrow \hat{\beta}=(X'^TX')^{-1}X'^TY'=(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

$$\Rightarrow \text{var}(\hat{\beta})=\sigma^2(X'^TX')^{-1}=\sigma^2(X^T\Sigma^{-1}X)^{-1}$$

GLS is like OLS regressing

$Y'=S^{-1}Y$ on $X'=S^{-1}X$

Note1: $\varepsilon'^T\varepsilon'$, $\hat{\beta}$, and $\text{var}(\hat{\beta})$ are invariant to the choice of S.

Note2: $\hat{\beta}=\hat{\beta}_{\text{OLS}}$ if $\Omega[\Sigma^{-1}X]=\Omega[X]$

- Q: why should not use ordinary least square when var(ε)$\neq\sigma^2 I$?

  model: $Y=X\beta+\varepsilon$, E(ε)=0 and var(ε)=$\sigma^2\Sigma$, OLS estimator $\hat{\beta}_{\text{OLS}}=(X^TX)^{-1}X^TY$

  $\boxed{\text{E}(\hat{\beta}_{\text{OLS}})=\beta,\ \text{var}(\hat{\beta}_{\text{OLS}})=\sigma^2(X^TX)^{-1}X^T\Sigma X(X^TX)^{-1} \Rightarrow \text{unbiased but variance not minimized (Note: } c^T\hat{\beta} \text{ is BLUE of } c^T\beta \Rightarrow \text{var}(c^T\hat{\beta})\leq\text{var}(c^T\hat{\beta}_{\text{OLS}}))}$

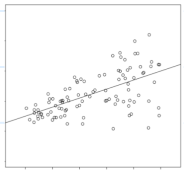
- diagnostics (residual analysis) should be applied on $Y'-X'\hat{\beta}=S^{-1}(Y-X\hat{\beta})=S^{-1}\hat{\varepsilon}$ because ε' are i.i.d. but not ε

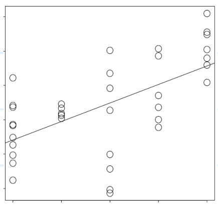
- The practical problem is that Σ may not be known. It's usually necessary to make some assumptions and examine the residuals to estimate Σ (check lab for an example, IRWLS)

❖ **Reading**: Faraway (2005, 1st ed.), 6.1 ❖ **Futher reading**: D&S, 9.2

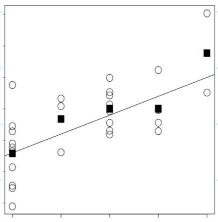
Weighted Least Square (WLS)

- Sometimes, the errors are uncorrelated, but have unequal variance where the form of the inequality is known ($\Rightarrow\Sigma$ is diagonal, it's a special case of GLS), example:
  - error variance proportional to a function of predictors [e.g., var(ε_i)=$x_i^2\sigma^2$ ($\Rightarrow w_i=1/x_i^2$)]





  - data with replicates, which show a pattern of unequal variance [e.g., var(ε_i) ≈ sample variance of observations with same x_i ($\Rightarrow w_i$ = 1/(sample variance))]
  - the observed y_i's are actually averages of several observations. [e.g., suppose y_i is the average of n_i observations, var(ε_i)=σ^2/n_i ($\Rightarrow w_i=n_i$)]



- ε: uncorrelated, but not constant variance $\Rightarrow$ Σ is diagonal. Write

$$\Sigma=\begin{pmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & 1/w_n \end{pmatrix} \Rightarrow \Sigma^{-1}=\begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$

  where w_i's ($\propto$ 1/var(ε_i)) are called *weights*.

  low weight $\Leftrightarrow$ high variance; high weight $\Leftrightarrow$ low variance

- S=diag($1/\sqrt{w_1}$, ..., $1/\sqrt{w_n}$), then $\Sigma=SS^T$

  $\Rightarrow$ OLS regress $S^{-1}Y$ (i.e., $\sqrt{w_i}\,y_i$) on $S^{-1}X$ ($\sqrt{w_i}\,x_i$) (Note. the column of ones, i.e., intercept needs to be replaced with $\sqrt{w_i}$)

  $\Rightarrow$ convenient for regression package without a weighted options

- **Q**: Why observations with smaller variance should be multiplied by heavier weight? intuitive interpretation?
- iteratively re-weighted least squares (IRWLS): In all the previous examples, weights (or Σ in GLS) are assumed known. **Q**: what if var(ε_i) is not completely known, what weights should we use? **Q**: where can you find the information of weights?
  - ➢ model the mean response for $\boldsymbol{Y}$, $E(\boldsymbol{Y})=\boldsymbol{X\beta}$
  - ➢ model the variance in $\boldsymbol{Y}$, $\text{var}(\boldsymbol{Y})=f(\boldsymbol{X}, \rho)$, where ρ are parameters for the variance model

residuals — fit the mean model to estimate β → fit the variance model to estimate ρ — weights → fit the mean model to estimate β

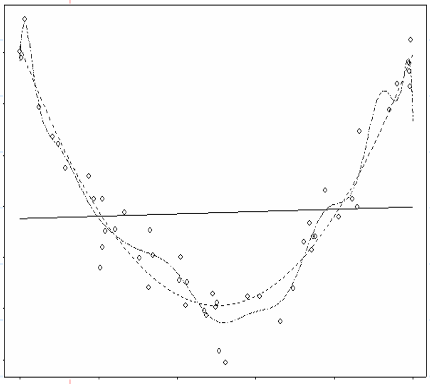
  - ➢ Example: $\text{var}(\varepsilon_i)=\rho_0+\rho_1 x_{i1}$
    1. start with $w_i=1$
    2. use weighted least square to estimate β
    3. use the residuals to estimate ρ_0 and ρ_1, perhaps by regressing residuals^2 on x_1
    4. re-compute the weights and go to 2. Continue until convergence

    Problems: converge? how is the inference about β affected? d.f.=? ...etc
  - ➢ alternative approach: jointly estimate the mean and variance parameters using likelihood based method (in R, use `gls()` function in the `nlme` library)

❖ **Reading**: Faraway (2005, 1st ed.), 6.2 ❖ **Futher reading**: D&S, 9.2

Testing for Lack of Fit

- model: $\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, **Q**: many choices of $\boldsymbol{X}$, how can we tell the chosen $\boldsymbol{X}$ *fits* the data?



Q: what is "fit"? what "fit" is appropriate? example:

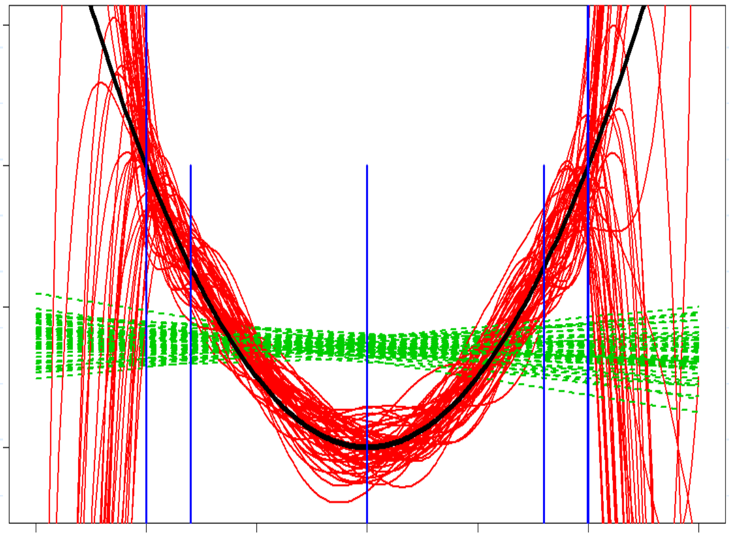
- data generated from the "true model" --- a 2nd-order polynomial of x
- solid line (fitted model=1st-order): not capture the quadratic pattern in data
- dashed line (fitted model=2nd-order): OK
- dotted line (fitted model=8th-order): fluctuated, fitted values and data are too close

➢ $\boldsymbol{X}$ too simple
⇒ not enough to explain the mean structure in data
⇒ lack of fit (under-fitting)
⇒ $\hat{\sigma}^2$ over-estimated

➢ $\boldsymbol{X}$ too complex
⇒ will explain the variation caused by errors, in addition to the mean structure
⇒ overfit the data
⇒ $\hat{\sigma}^2$ under-estimated

➢ **Q**: what statistic carry the information about lack of fit or overfit? Ans: $\hat{\sigma}^2$

- Repeat the simulation 50 times
  - Black line: true model, 2nd-order polynomial
  - Green lines: fitted lines under under-fitting model, 1st-order polynomials
  - Red lines: fitted lines under over-fitting model, 8th-order polynomials



  - Comparison
    - Under-fitting lines: small variance, large bias
    - Over-fitting lines: large variance, small bias
    - Interpolation vs. extrapolation

- Note: high R^2 (a measure of "goodness of fit") doesn't imply the model is a good fit (check lab for examples of high R^2 but lack of fit)
- graphical checking (based on residuals ←informal) (check a lab example)

• a possible testing procedure for lack of fit (⇒ compare $\hat{\sigma}^2$ to σ^2)

- when σ^2 is known
  - examples:
    - σ^2 is known from past experience
    - source of variation in ε is only measurement error, measuring device well understood, knowledge of measurement error inherent in an instrument or by definition
    - each y_i is an average of a very large number of observations, σ_i^2 effectively estimated (see an example in lab)
  - test for lack of fit:

    H_0: $E(\boldsymbol{Y})=\boldsymbol{X\beta}$ is correct against H_1: $E(\boldsymbol{Y})=\boldsymbol{X\beta}$ is too simple

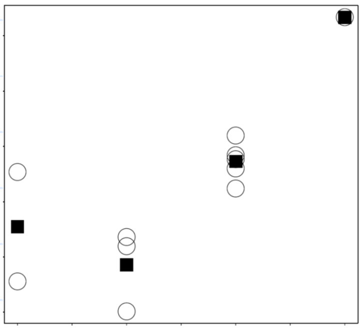
    - $(n-p)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}$ if model is correct (under H_0), where n=# of observations, p=# of parameters in $\boldsymbol{\beta}$
    - test statistic: $(n-p)\hat{\sigma}^2/\sigma^2 = RSS/\sigma^2$, compared with χ^2_{n-p}
    - reject if $(n-p)\hat{\sigma}^2/\sigma^2 > \chi^{2\,(1-\alpha)}_{n-p}$
    - if a lack of fit is found, a new (more complex) model is needed

➢ when σ^2 is unknown ($\Rightarrow$ can we use data to estimate it?)

- need "model-free method" to estimate σ^2 (i.e., free of the $E(\boldsymbol{Y})=\boldsymbol{X\beta}$ assumption)
- because we want to use estimated σ^2 to justify whether $E(\boldsymbol{Y})=\boldsymbol{X\beta}$ is suitable, the estimated σ^2 should have no relationship with the choice of $\boldsymbol{X}$
- denote the estimated σ^2 under model-free method by $\hat{\sigma}^2_{\text{p.e.}}$, where p.e. stands for "pure error")
- usually, only possible for data with *replication* (**Q**: why?)
  - how to estimate σ^2 (model-free)?
  - $SS_{\text{p.e.}} = \Sigma_{\text{distinct x}} \Sigma_{\text{within an x}} (y_{x,i} - \bar{y}_x)^2$
  - d.f. of p.e. $= \Sigma_{\text{distinct x}}$ (# of replications – 1)

    e.g. $(2-1)+(3-1)+(5-1)+(1-1)=7$
  - $\hat{\sigma}^2_{\text{p.e.}} = SS_{\text{p.e.}}/df_{\text{p.e.}}$
  - test statistic:

    (n=# of observations, p=# of parameters in $\boldsymbol{\beta}$, RSS calculated from the model $E(\boldsymbol{Y})=\boldsymbol{X\beta}$)

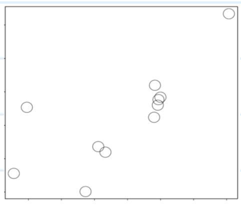


	$d.f.$	SS	MS	F
Residual	$n-p$	RSS		
Lack of fit	$n-p-df_{\text{p.e.}}$	$RSS-SS_{\text{p.e.}}$	$(RSS-SS_{\text{p.e.}})/(n-p-df_{\text{p.e.}})$	ratio of MS (compared to $F_{n-p-df_{\text{p.e.}},\, df_{\text{p.e.}}}$)
Pure error	$df_{\text{p.e.}}$	$SS_{\text{p.e.}}$	$SS_{\text{p.e.}}/df_{\text{p.e.}}$	

- Note: $\hat{\sigma}^2_{\text{p.e.}}$ is the estimate of σ^2 when we fit a saturated model to the data
- alternative view:

  this is a comparison between the model of interest (i.e., ω: $\boldsymbol{X\beta}$) and a saturated model (Ω, whose R^2 reaches the maximum) that assigns a parameter to each *unique* combination of the predictors $\Rightarrow$ standard F-testing for H_0: ω v.s. H_1: Ω\ω

- need replication to make the test, but it's rare in obs'nal data
  - possible solution: grouping (could be questionable $\Rightarrow$ different grouping schemes may cause different conclusions)



- **Q**: what's a conservative conclusion when H_0 is accepted?
  $\Rightarrow$ may not conclude $\boldsymbol{X\beta}$ is the true model. We may say the true E($\boldsymbol{Y}$) $\approx \boldsymbol{X\beta}$ on the observed data points
- **Q**: can the procedure be modified to test overfitting?

- Note that fitting is not everything
  - it often possible to fit data perfectly by introducing more effects/predictors
  - for data without replication, you can fit a model with R^2=1 and zero $\hat{\sigma}^2$
  - a very complex model can fit data perfectly (even exactly), but ...
    - may have no explanation (may learn nothing beyond the data itself)
    - prediction unstable
      (e.g, on region without data points, MSE=Var+Bias^2)
- **Q**: what is the source of variation in your data? ($\boldsymbol{X\beta}$ and $\boldsymbol{\varepsilon}$)
  what σ^2 is estimated (i.e., what is the source of variation in $\boldsymbol{\varepsilon}$)? example:
  - replication generated from different units v.s. repeated measures of same unit
  - repeatability v.s. reproducibility in measurement system analysis

❖ **Reading**: Faraway (2005, 1st ed.), 6.3 ❖ **Futher reading**: D&S, 2.1