Generalized Least Square (GLS)

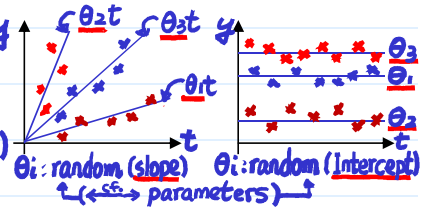
Recall OLS estimator is BLUE under Gauss-Markov condition

- model: $Y=X\beta+\varepsilon$, $E(\varepsilon)=0$ and $\text{var}(\varepsilon)=\sigma^2 I \Rightarrow \varepsilon$: uncorrelated and constant variance
  (no normality; $\text{Var}(Y)=\sigma^2 I$ [Y])

  Q: what if $\text{var}(\varepsilon)\neq\sigma^2 I$? ε may have non-constant variance and/or are correlated, e.g., → correlation between observations in Y [$\text{Var}(\varepsilon)=\text{Var}(Y)$]

  Q how to model correlation in sampling model? (X,Y) vs. Y|X (LN p.5-7)

  - ➢ time series correlation [e.g., $\varepsilon_t \sim$ ARMA(r,m)] — ① data observed over time ② previous observations correlated with future observation (Auto-regression)
  - ➢ growth curve model, repeated measurement model [e.g., several observations taken from same person, or same unit] — data y_i, y_j from same unit: $\text{cov}(y_i, y_j) = \text{cov}(\theta_{t_i}+\delta_i, \theta_{t_2}+\delta_j) = t_i t_2 \text{Var}(\theta) \neq 0$

    $\text{cov}(y_i, y_j) = \text{cov}(\theta+\delta_i, \theta+\delta_j) = \text{Var}(\theta)$
  - ➢ spatial correlation [e.g., data taken over contiguous geographical areas: census tracts, countries, or states in a country. Nearby areas are often much alike] (split plot design; cf. → nearby time points)
  - ➢ nested errors [e.g., M sets of observations, each set from common production run, from same/common equipment, or from same survey-taker] (random effect model in DOE; no parameters; parameter)

    $\text{Var}(\varepsilon) = \text{Var}(Y) =$ block-diagonal matrix

- 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 (off-diagonal part of Σ; diagonal part of Σ)

(Σ is a covariance matrix ($\because \text{Var}(\varepsilon/\sigma)=\Sigma$))

- 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) ($\Sigma^{1/2}(\Sigma^{1/2})^T$; not unique)

  Note: If $\text{Var}(Y)=\sigma^2\Sigma$, we can find a matrix A s.t. $\text{Var}(AY)=E^*(AYY^TA^T)=AE^*(YY^T)A^T=(A\Sigma A^T)\sigma^2=\sigma^2 I$ ($A=S^{-1}$)

  $Y=X\beta+\varepsilon \Rightarrow S^{-1}Y=S^{-1}X\beta+S^{-1}\varepsilon \Rightarrow Y'=X'\beta+\varepsilon'$ (identical β), where (S^{-1} known; still a linear model (known response & predictors))

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

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

  $\Rightarrow$ For Y' and X', the assumption in ordinary least square is satisfied → can do OLS on Y' & X'

- GLS: find β that minimize (Gauss-Markov conditions; $(\Sigma^{-1/2})^T(\Sigma^{-1/2})=\Sigma^{-1}$)

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

  $0=\hat{\varepsilon}'^T\hat{Y}'=\hat{\varepsilon}^T\Sigma^{-1}\hat{Y}$

  $\Rightarrow \hat\beta=(X'^TX')^{-1}X'^TY'=(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$ ($\equiv \hat\beta_{GLS}$ for original data X, 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 $X'=S^{-1}X$ ($\hat\varepsilon=Y-X\hat\beta$, $\hat Y$; $\varepsilon'=S^{-1}\varepsilon$; $\sigma^2 I$; $\sigma^2\Sigma$)

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

  Q: What depends on the choice of S?

  Ans: $\hat Y'=S^{-1}\hat Y$, $\hat\varepsilon'=S^{-1}\hat\varepsilon$

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

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

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

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

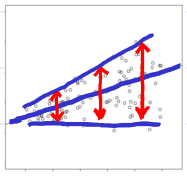
- diagnostics (residual analysis) should be applied on $Y'-X'\hat{\beta}=S^{-1}(Y-X\hat{\beta})=S^{-1}\hat{\varepsilon}$ because ε' are i.i.d. but not ε

- 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)

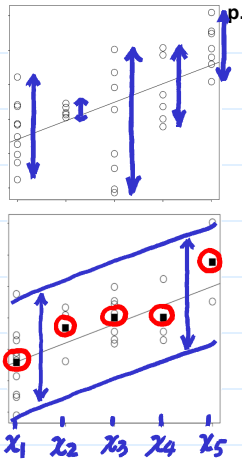
❖ **Reading**: Faraway (2005, 1st ed.), 6.1 ❖ **Futher reading**: D&S, 9.2

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., $var(\varepsilon_i)=x_i^2\sigma^2$ ($\Rightarrow w_i=1/x_i^2$)]



  - data with replicates, which show a pattern of unequal variance [e.g., $var(\varepsilon_i)\approx$ 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, $var(\varepsilon_i)=\sigma^2/n_i$ ($\Rightarrow w_i=n_i$)]



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

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

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

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

- $S=\text{diag}(1/\sqrt{w_1},\ldots,1/\sqrt{w_n})$, 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? Y_i's with small variance have more influence on $\hat{\beta}_{WLS}$ pull the fitted line toward these data points

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? Ans: residuals (Σ or Σ?)

- model the mean response for Y, $E(Y)=X\beta$
- model the variance in Y, $var(Y)=f(X, \rho)$, where ρ are parameters for the variance model (or $=f(E(Y), \rho)$)

procedure to estimate β in GLM

residuals → fit the mean model to estimate β ($\hat{\beta}_{WLS}$ or $\hat{\beta}_{GLS}$) → fit the variance model to estimate ρ → weights (or $\hat{\Sigma}$) → residuals

- Example: $var(\varepsilon_i)=\rho_0+\rho_1 x_{i1}$

  $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=E(\varepsilon_i^2)=\rho_0+\rho_1 x_{i1}$

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

quasi-likelihood in GLM — Garroll & Ruppert (1988)

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) — add distribution assumption

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

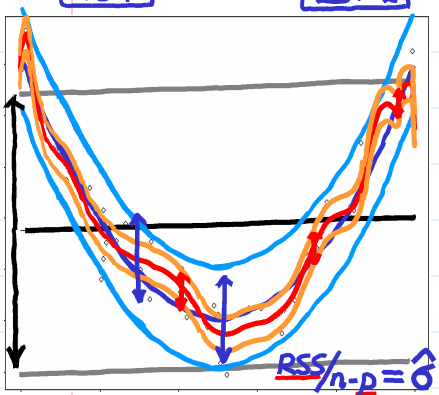
Testing for Lack of Fit

fitted model — Recall goodness of fit: R^2 & $\hat{\sigma}$ ($Y \approx \hat{Y}$)

Note: good fit → not necessary a right X

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

規律 (systematic) + 隨機 (random); $\hat{Y}=X\hat{\beta}$, $Y=X\beta+\varepsilon$, $X\beta$



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

- data generated from the "true model" --- a 2nd-order polynomial of x $E(y|x)=\beta_0+\beta_1 x+\beta_2 x^2$
- solid line (fitted model=1st-order): not capture the quadratic pattern in data (規律 → 隨機) ⇒ might produce biased $\hat{\beta}$
- dashed line (fitted model=2nd-order): OK
- dotted line (fitted model=8th-order): fluctuated, fitted values and data are too close (規律 ← 隨機)

Why not compare R^2 to true R^2? — R^2 increases — depends on the range of X (LNp.3-18)

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

If we have some information about how large σ^2 should be,

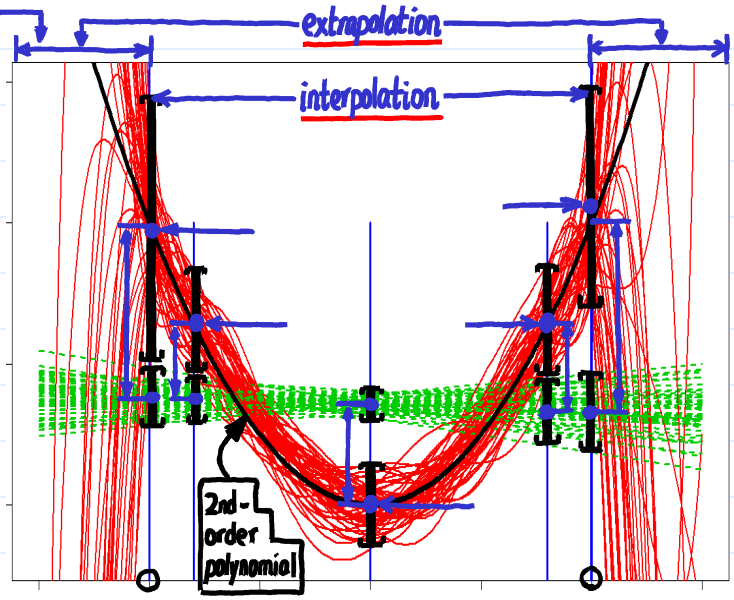
- X too complex (too simple & too complex could coexist in a fitted model)
  - ⇒ will explain the variation caused by errors, in addition to the mean structure
  - ⇒ overfit the data
  - ⇒ $\hat{\sigma}^2$ under-estimated

true model: $Y=X^*\beta^*+\varepsilon^*$; fitted model: $=X\beta+\varepsilon=X\hat{\beta}+\hat{\varepsilon}$, $\hat{Y}$

problem of overfitting: current data / future data → cross validation

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

- Repeat the simulation 50 times $\hat{Y}=P_\Omega Y=P_\Omega(X\beta+\varepsilon)$
  - 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



$\text{span}\{x_1\}=\Omega_1$
$\text{span}\{x_1,x_2\}=\Omega_2^*=\Omega_1\oplus\Omega_2$
$\text{span}\{x_1,x_2,x_3\}=\Omega_3^*=\Omega_1\oplus\Omega_2\oplus\Omega_3$

very large variance in the red lines when extrapolating

under-fitting: $Y=X_1\beta_1+\varepsilon_1$
correct: $Y=X_1\beta_1+X_2\beta_2+\varepsilon_2$
over-fitting: $Y=X_1\beta_1+X_2\beta_2+X_3\beta_3+\varepsilon_3$

Define $\Omega_1=\text{span}\{x_1\}$ $\Omega_1\perp\Omega_2\perp\Omega_3$
$\Omega_2^*=\text{span}\{x_1,x_2\}$, $\Omega_2=\Omega_2^*\cap\Omega_1^\perp$
$\Omega_3^*=\text{span}\{x_1,x_2,x_3\}$, $\Omega_3=\Omega_3^*\cap\Omega_2^{*\perp}$

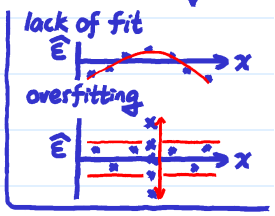
Define $\hat{Y}_1=P_{\Omega_1}Y$, $\hat{Y}_2=P_{\Omega_2}Y$, $\hat{Y}_3=P_{\Omega_3}Y$.

under-fitting: $\hat{Y}=\hat{Y}_1$
correct: $\hat{Y}=\hat{Y}_1+\hat{Y}_2$
over-fitting: $\hat{Y}=\hat{Y}_1+\hat{Y}_2+\hat{Y}_3$

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

Recall. MSE = variance + bias²

- 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) cf. overfitting
- 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 (check LNp.6-11)
      - 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: fitted model: $Y=X\beta+\varepsilon$, $\varepsilon\sim N(0,\sigma^2 I)$, σ^2: known
      H_0: $E(Y)=X\beta$ is correct against H_1: $E(Y)=X\beta$ 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 β (RSS; null distribution)
      - test statistic: $(n-p)\hat{\sigma}^2/\sigma^2 = RSS/\sigma^2$, compared with χ^2_{n-p}
      - reject if $(n-p)\hat{\sigma}^2/\sigma^2 > \chi^2_{n-p}(1-\alpha)$ (known; exam over-fitting?) Why? Why not 2-sided test?
      - if a lack of fit is found, a new (more complex) model is needed

variance estimate under a fitted model: $Y=X\beta+\varepsilon$ — $\text{Var}(\varepsilon^*)$ true error — check $\hat{\sigma}^2\overset{?}{=}\sigma^2$

$x_i\to(y_{i1},\cdots,y_{in_i})$, obtain $(\bar{x}_i,\bar{y}_i)$, $\hat{\sigma}_i^2\Rightarrow\text{Var}\left(\bar{y}_i/\hat{\sigma}_i/\sqrt{n_i}\right)\approx 1$
$z=X\beta+\varepsilon\Rightarrow\text{Var}(\varepsilon)=I$
z_i

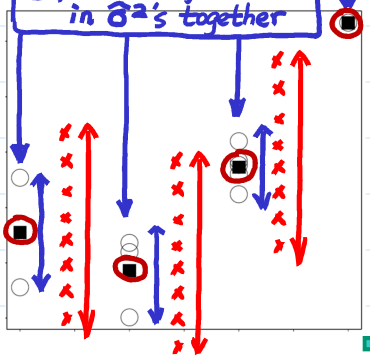
- when σ^2 is unknown ($\Rightarrow$ can we use data to estimate it?)
  - ↳ information is in residuals
  - need "model-free method" to estimate σ^2 (i.e., free of the $E(\boldsymbol{Y})=\boldsymbol{X\beta}$ assumption) — residuals of what model?
    - ↑ Why? $\hat{\varepsilon} = Y - X\hat{\beta}$ ← depends on the assignment of X
  - 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") — can treat the y_i's on each distinct x_i as a one-sample case
  - usually, only possible for data with *replication* (Q: why?)
    - how to estimate σ^2 (model-free)? Recall $\hat{\sigma}^2 = RSS/df$
    - $SS_{\text{p.e.}} = \Sigma_{\text{distinct x}} \Sigma_{\text{within an x}} (y_{x,i} - \bar{y}_x)^2$ — n - (# of distint x_i)
    - d.f. of p.e. $= \Sigma_{\text{distinct x}}$ (# of replications $-$ 1)

      e.g. $(2-1)+(3-1)+(5-1)+(1-1)=7$ — Why?
    - $\hat{\sigma}^2_{\text{p.e.}} = SS_{\text{p.e.}}/df_{\text{p.e.}}$ — $\hat{\sigma}^2_{\text{pool}} = \dfrac{(m_1-1)\hat{\sigma}_1^2 + \cdots + (m_K-1)\hat{\sigma}_K^2}{(m_1-1)+\cdots+(m_K-1)}$ — # of distinct x_i's
    - test statistic:

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

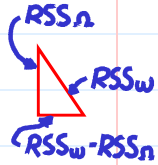
Recall: one-sample case
$Y_1, \cdots, Y_m \overset{iid}{\sim} N(\mu, \sigma^2)$
$\bar{Y} \to \mu$
$S^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y})^2 \to \sigma^2$

① treat each group of data as one sample
② obtain $\hat{\sigma}^2$ of each sample
③ pool the information in $\hat{\sigma}^2$'s together



$E(Y)=X\beta$	*d.f.*	*SS*	*MS*	*F*
Residual (ω)	$n-p$	RSS_ω	$RSS_\omega/n-p = \hat{\sigma}^2_\omega$	
Lack of fit	$n-p-df_{\text{p.e.}}$ ($k-p$)	$RSS-SS_{\text{p.e.}}$	$(RSS-SS_{\text{p.e.}})/(n-p-df_{\text{p.e.}})$ $(RSS_\omega - RSS_\Omega)/k-p$	ratio of MS (compared to $F_{n-p-df_{\text{p.e.}},\, df_{\text{p.e.}}}$)
Pure error (Ω)	$df_{\text{p.e.}}$ ($n-k$)	$SS_{\text{p.e.}}$ (RSS_Ω)	$SS_{\text{p.e.}}/df_{\text{p.e.}} = \hat{\sigma}^2_{p.e.}$	null dist.

Recall. General form of F-test in LN p.4-10



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

k-sample model: $\Omega: E(Y_{ij}|x_i) = \mu_i$, $i = 1, \cdots, k$ (# of distinct x_i's); $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_k) \in \mathbb{R}^k$; $j = 1, \cdots, n_i$ $(n_1 + \cdots + n_k = n)$

$Y = X'\beta' + \varepsilon$, $X' = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & & 1 \end{bmatrix}$, $\beta' = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix} \Rightarrow \hat{\sigma}^2 = \hat{\sigma}^2_{p.e.}$

cf. LN p. 5-11 ($n=p$)

$\omega: E(Y_{ij}|x_i) = X\beta$ — $\hat{Y} = ?$

reduce the dimension of $\boldsymbol{\mu}$ from k to p — # of parameters $= p < k$

$\omega = \text{span}(X) \subseteq \text{span}(X') = \Omega$ — not one if there are replicates (why? $\hat{\varepsilon} \approx 0$)

- 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: $\Omega \backslash \omega$

Why cannot this test apply to no replicate data?

- 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)