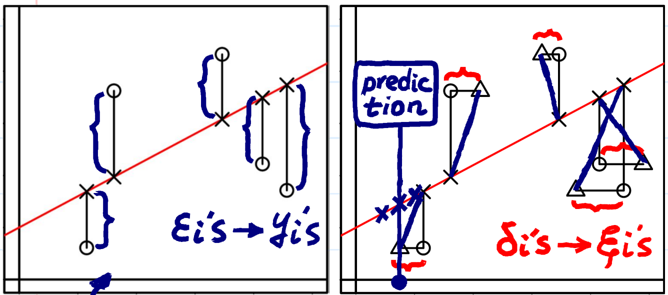
Errors in the predictor

- Recall: $Y = X\beta + \varepsilon$ (random: Y, ε), where ε is error that represent measurement error or unexplained variation in Y, and "it's assumed that X are fixed values measured without error."
- Q: what if X (← random) are measured or observed with error? (examples) (1) sampling Data (3) DOE data (2) observational data

  Note: compare the difference between the 2 statements: "X measured with errors" and "X are random variables, such as in sampling model" ⇒ Both X are random variables but, different causes of their randomness
- Q: what happens if you ignore errors in X and still use OLS estimator? (LNp. 5-6~7) Let us consider a simple example: $E(y_\xi) = \beta_0 + \beta_1 \xi$



×: "fine" value, no error in predictor and response, but not observed
○: observations with error only in response, but not error in predictor
Δ: observations with error in both response and predictor

- ➢ Q: Is minimizing *RSS* (i.e., OLS) still reasonable for Δ data? No. (not a linear model)
- ➢ a statistical model for Δ data: ideal "true" relationship is $E(y_\xi)$: $\eta_i = \beta_0 + \beta_1 \xi_i$, (not observed)

  but observe $y_i = \eta_i + \varepsilon_i$ and $x_i = \xi_i + \delta_i$,

  where ε and δ are errors of response and predictor respectively, i.e., (assume independent)

  (observed) $y_i = \beta_0 + \beta_1 \xi_i + \varepsilon_i = \beta_0 + \beta_1 x_i + (\varepsilon_i - \beta_1 \delta_i)$ (observed) ε_i^*

  ×: (ξ_i, η_i)
  ○: (ξ_i, y_i)
  Δ: (x_i, y_i)

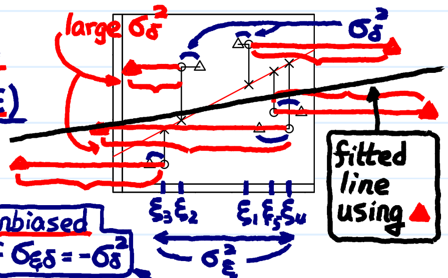
- ➢ Q: what problem if we use ordinary least square to estimate β_1 in the model?
  - Let's assume $E(\varepsilon_i) = E(\delta_i) = 0$, $var(\varepsilon_i) = \sigma^2$, $var(\delta_i) = \sigma_\delta^2$ and $cov(\varepsilon, \delta) = 0$ → cf. $Var(\varepsilon_i^*) = \sigma^2 + \beta_1^2 \sigma_\delta^2$
  - (ξ_i's are random in sampling model) Let $\sigma_\xi^2 = \Sigma(\xi_i - \bar{\xi})^2/n$ (Note: when ξ_i's are not random, we could regard it as a measure of the spread of the predictor), $\sigma_{\xi\delta} = cov(\xi, \delta)$ and assume $cov(\xi, \varepsilon) = 0$
  - the OLS estimator of β_1 is: (LNp.3-7) $\hat{\rho}\frac{\hat{\sigma}_y}{\hat{\sigma}_x} = \hat{\beta}_1 = [\Sigma(x_i - \bar{x})(y_i - \bar{y})] / \Sigma(x_i - \bar{x})^2$ ← $cov(x, y)/var(x)$
  - after some calculation, we can write $\dfrac{cov(\xi+\delta, \beta_1\xi+\varepsilon)}{Var(\xi+\delta)}$

    (exercise) → $E(\hat{\beta}_1) \approx \beta_1 \times [(\sigma_\xi^2 + \sigma_{\xi\delta})/(\sigma_\xi^2 + \sigma_\delta^2 + 2\sigma_{\xi\delta})]$ ← unbiased if $\sigma_{\xi\delta} = -\sigma_\delta^2$
  - if no relation between ξ and δ (i.e., $cor(\xi, \delta) = 0$),

    $E(\hat{\beta}_1) \approx \beta_1 \times [\sigma_\xi^2/(\sigma_\xi^2 + \sigma_\delta^2)] = \beta_1 \times [1/(1 + \sigma_\delta^2/\sigma_\xi^2)]$ ⇒ $\hat{\beta}_1$ is biased

    (≤ 1) (if x & δ are uncorrelated, $\sigma_{\xi\delta} = cov(\xi, \delta) = cov(x - \delta, \delta) = -\sigma_\delta^2$)
  - typically, bias in $\hat{\beta}_1$ is towards zero
  - size of the bias depends mainly on the ratio $\sigma_\delta^2/\sigma_\xi^2$ (i.e., variability in the errors of predictor relative to the spread of predictor) (Q: why reasonable?)
    - ratio is small ⇒ no worry
    - ratio is large, $|\hat{\beta}_1|$ is underestimated ⇒ use *measurement error model*

    (How? ① reduce σ_δ^2 ② increase σ_ξ^2)



- For multiple predictors, the usual effect of measurement errors on predictors is to bias the estimator of β in the direction of zero But, $Var(\hat{Y})$↑ ∵ error in X
- Prediction is not biased since future X will also be measured with errors. So, model for prediction should be built on X's measurement with error. (Note. $y = \beta_0 + \beta_1 x + \varepsilon^*$)

❖ **Reading**: Faraway (1st ed.), 5.1; W, 4.6.3 ❖ **Further reading**: D&S, 3.4, 9.7

Collinearity ← Recall. Identifiability (LNp.5-11 ~ 12)

model: $y = \sum_j \beta_j g_j(\underline{x}) + \varepsilon$

- collinearity: predictors are (linearly) related to each other
  - X^TX is singular ⇒ some predictors are linear combinations of others
    ⇒ (exact) collinearity ⇒ no unique estimate of β
    unidentifiable: $\exists\, a_1, \ldots, a_p$ s.t. $\sum_{j=1}^{p} a_j g_j(\underline{x}_i) = 0$, $\forall i$
  - X^TX close to singular ⇒ close to linear dependent among some predictors
    ⇒ (approximate) collinearity or multicollinearity $\exists\, a_1, \ldots, a_p$ s.t. $\sum_{j=1}^{p} a_j g_j(\underline{x}_i) \approx 0$
- effect of collinearity: (exercise by Schur complement) ← $(X^TX)^{-1}_{jj} = [\underline{g}_j^T\underline{g}_j - \underline{g}_j^T X_{(-j)}(X_{(-j)}^T X_{(-j)})^{-1} X_{(-j)}^T \underline{g}_j]^{-1}$ ← RSS_j^{-1}
  - estimated effects are unstable (can change magnitude or sign depending on the other predictors in the model) ⇒ interpretation of estimated coefficients difficult
  - cause numerical problem in estimating β and associated quantities ← calculate $(X^TX)^{-1}$
  - $S_j \cdot TSS_j$
  - var($\hat{\beta}_j$) = σ^2 $(1/(1-R_j^2))$ $(1/S_j)$, where $S_j = \Sigma_i (g_{ij} - \bar{g}_j)^2$ and R_j^2 is the coefficient of determination obtained from regressing g_j on all other predictors ⇒ when $R_j^2 \approx 1$, var($\hat{\beta}_j$) large ⇒ t-test may fail to reveal significance, i.e., miss important g_j
    $(X^TX)^{-1}_{jj} \times \sigma^2$
  - *variance inflation factor*: $VIF_j = 1/(1-R_j^2)$ ⇒ when S_j is fixed, VIF_j represents the increase in variance due to the collinearity (e.g., interpret VIF_j=16?)
    compared to the case of orthogonality (i.e., $R_j^2 = 0$)
    $s.e.(\hat{\beta}_j) \approx \sqrt{16} = 4$ times larger than being orthogonal.
- detection of collinearity:
  - examine correlations between predictors, i.e., cor(g_k, g_j) ← from X^TX
    ⇒ any values close to 1 or −1 reveal *pairwise* correlation
  - for each g_j, regress g_j on all other predictors and compute R_j^2 or VIF_j
    ⇒ R_j^2 close to one or VIF_j much larger than one indicate a problem of collinearity

- examine eigenvalues, $\lambda_1 \geq \ldots \geq \lambda_p$, of X^TX ⇒ small eigenvalues indicate a problem
  ↳ eigenvector $(a_1, \cdots, a_p)$
  LNp.5-12
  - condition number: $k = (\lambda_1/\lambda_p)^{1/2}$
  - rough rule: $k > 30$ is considered large ⇒ $\lambda_1/\lambda_p > 900$
    Then, $a_1\underline{g}_1 + \cdots + a_p\underline{g}_p \approx \underline{0}$
  - for each i, $(\lambda_1/\lambda_i)^{1/2}$ are worth considering ⇒ there may exist more than one linear combination relationship between predictors
  - eigenvectors of small eigenvalues indicate possible source of collinearity
- how to deal with collinearity:
  - identify the cause of collinearity in data
    Check LNp.5-11 ~ 12
    explain why collinearity occurs, not only to detect whether it occurs.
  - amputate some predictors if you can --- remember that collinearity happens because too many variables try to do the *same job* of explaining the response ← cf.
    check 1st Note in LNp.9-6
  - do not conclude the predictors we drop have nothing to do with the response ← cf.
  - techniques such as principal component regression, ridge regression, partial least squares, …, may help → to reduce the impact of collinearity, e.g., PC use linear combinations of g_j's

❖ **Reading**: Faraway (1st ed.), 5.3; W, 10.1
❖ **Further reading**: D&S, 16.1, 16.4, 16.5

主成分 → Principal components

every column of Z is a linear combination of the columns of X

β, β' are different parameters

- Recall: $Y = X\beta + \varepsilon$. If X is orthogonal (i.e., X^TX is a diagonal matrix), then estimation, testing, and parameter interpretation are greatly simplified.
- idea: For non-orthogonal X, replace $Y = X\beta + \varepsilon$ by $Y = Z\beta' + \varepsilon$, where Z is a linear combinations of X (i.e., $Z_{n\times q} = X_{n\times p}U_{p\times q}$, $p \geq q$) and Z is orthogonal (Z^TZ is diagonal)

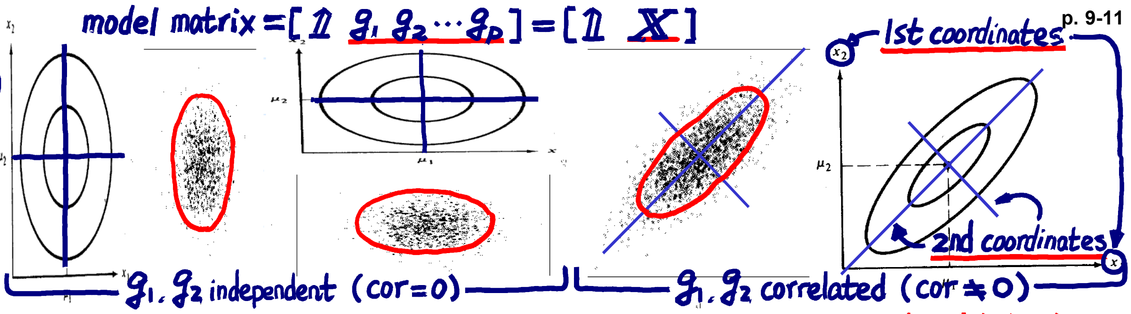
- orthogonality

Recall: (LNp.4-5) bivariate normal $\bar{g}=0$

orthogonal g_j, g_k

$\Rightarrow cor(g_j, g_k)=0$

model matrix $= [\mathbb{1}\ g_1\ g_2 \cdots g_p] = [\mathbb{1}\ X]$



- $\underline{Z}_{n\times q} = \underline{X}_{n\times p}\underline{U}_{p\times q}$, $p \geq q$, e.g., take a look of the first column of Z

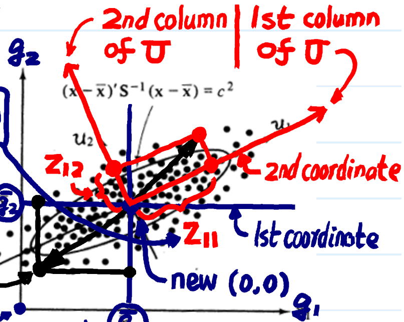
rotation matrix

$$\underline{z}_1 = \begin{bmatrix} z_{11} \\ z_{21} \\ \cdots \\ z_{n1} \end{bmatrix} \xleftarrow{\text{inner product}} \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1p} \\ g_{21} & g_{22} & \cdots & g_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ g_{n1} & g_{n2} & \cdots & g_{np} \end{bmatrix}$$

1st row of X "•"; projection; u_{11}, u_{21}, ..., u_{p1} ← 1st column of U (unit vector)

$(g_{11}, g_{12}, \ldots, g_{1p}) \xrightarrow{\text{rotation}} (z_{11}, z_{12}, \ldots, z_{1q})$ ← 1st row of Z "•"

row of X

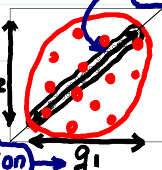


- see graph, in which z_1 is the projection of points on the direction u_1; z_2 is the projection of points on the direction u_2

cf. graph in LNp.3-3

cf. model selection.

- concept of *dimension reduction*: 1. smaller dimension, better 2. important information should be kept



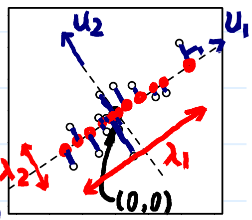
  - **Q**: take a look of the graph, the points are of 1-dim or of 2-dim? model selection

    ⇒ very similar to a line ⇒ high correlation ⇒ data is 2-dim, but close to 1-dim
  - replace large number (p) of columns in X with small number (q (usually ≤ p)) of columns in Z

    ⇒ simpler model, especially useful (1) when few linear combinations of X are enough to represent the variation in X; (2) when $p > n$ ← unidentifiable

- principal component (PC): $Z_j = XU_j \Leftarrow Z = XU$; $Z_j^T Z_j = U_j^T(X^TX)U_j = \lambda_j \|U_j\|^2 = \lambda_j$
  - transform X to Z which is orthogonal, but how? Z_j, Z_k uncorrelated, if $E(Z_j)=E(Z_k)=0$
  - find U such that * Z^TZ is diagonal, i.e., $Z^TZ = \text{diag}(\lambda_1, \ldots, \lambda_p)$, where $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$
  - since $Z^TZ = U^T(X^TX)U$, to make Z^TZ diagonal, we can choose columns of U are (symmetric) orthogonal eigenvectors of X^TX, then the $\lambda_1, \lambda_2, \ldots, \lambda_p$ are eigenvalues of X^TX
    - let U_j and λ_j be the j-th eigenvector and eigenvalue of X^TX, then $(X^TX)U_j = \lambda_j U_j$
    - $U_k^TU_j = 0$ for $k \neq j$ and $\|U_j\|=1$ for all j
    - $(Z^TZ)_{kj}$: $U_k^T(X^TX)U_j = \lambda_j U_k^TU_j$, which equals 0 if $k \neq j$ and equals λ_j if $k=j$
  - Z_1 (=1st column of Z) is called 1st principal component (PC),

    projection: Z_2 (=2nd column of Z) is called 2nd principal component (PC), ...



  - another way to look at it: $Z_j^TZ_j = (Z^TZ)_{jj} = \lambda_j$

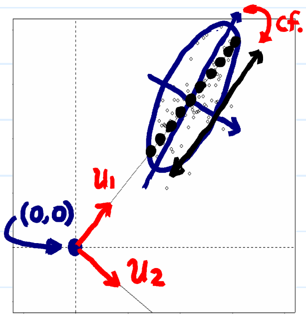
    U_j: unit vector
    - Z_1 = linear combination of columns of X that has maximum length², i.e, maximizing Σz_{i1}^2 (variation of Z_1)

    $\|Z_j\|^2 = Z_j^TZ_j = (XU_j)^T(XU_j) = U_j^TX^TXU_j \Rightarrow$ quadratic form
    - Z_2 = linear combination of columns of X that is orthogonal to Z_1 and has maximum length² ⊕ $U_2 \perp U_1$
    - Z_3 = linear combination of columns of X that is orthogonal to Z_1, Z_2 and has maximum length² ⊕ $U_3 \perp U_1$, $U_3 \perp U_2$
    - …
  - some properties:
    - * $U^TU = I_{q\times q}$ ← U: an orthogonal matrix
    - zero eigenvalue ⇒ unidentifiable
    - λ_j = length² of $Z_j = \Sigma_i z_{ij}^2$ [note: when $E(X_j)=0 \Rightarrow E(Z_j)=0 \Rightarrow \lambda_j \propto var(Z_j)$] sample mean; sample variance; $\bar{g}_j = 0\ \forall j$
    - $\lambda_1 + \ldots + \lambda_p = tr(X^TX) = \Sigma_j$(length² of X_j) $tr(Z^TZ)$

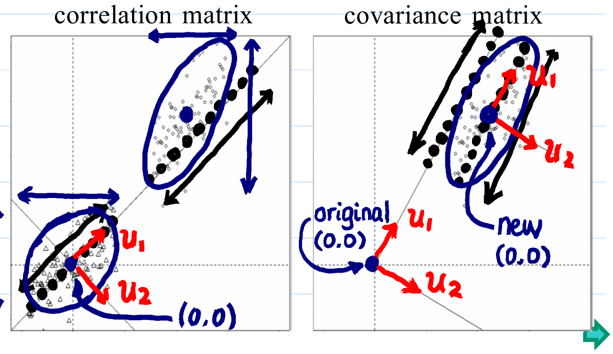
    [note: when $E(X_j)=0$, $\forall j$ $\lambda_1 + \ldots + \lambda_p \propto \Sigma_j var(X_j)$: total variation] of X; sum of variations of different units
    - $\lambda_j/(\lambda_1 + \ldots + \lambda_p)$ = proportion of total variation explained by the jth PC

- **Q**: how to interpret $Z_1, Z_2, \ldots, Z_p$? Ans: compare the coefficients in eigenvectors
  - Ex 1: Z_1=0.46GNP+0.32UnEm+0.46POP+0.46Year+... ⇒ hard to give meaning
    average
  - Ex 2: Z_1=0.63(hw1)+0.57(hw2)+0.52(hw3) ∝ average homework scores;
    Z_2 = 0.67(hw1)+0.08(hw2)-0.75(hw3) ∝ difference between hw 1 and 3 scores
    ≈ 0

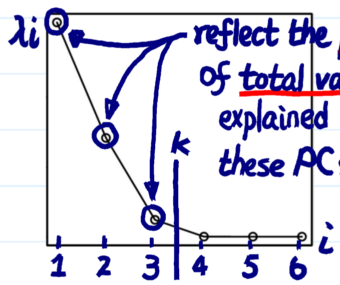
- variation on principal component regression
  - use X^TX with/without constant term (without constant term ⇒ PC's may not be orthogonal to constant term)
    $X \to X_v$ location change of g_j's
    $(Z^TZ)_{jk}=0$ ~~⇒~~ Z_j, Z_k uncorrelated
    X^TX without constant term
  - use covariance matrix of X (without constant term), i.e., $X_v^TX_v/(n-1)$ where X_v is formed by centering each g_j, to find eigenvectors U and eigenvalues. Then, λ_j = var(z_j). The transformation U can be applied on X or X_v [PC's are orthogonal to constant term if transformation is applied on X_v]
    $g_j-\bar{g}_j$
    $Z=XU$
    $Z_v=X_vU$



  - use correlation matrix of X (without constant term), i.e, $X_r^TX_r/(n-1)$, where X_r is formed by standardizing each g_j. To make sense, the transformation should be applied on X_r. Then, λ_j = var(z_j) and PC's are orthogonal to constant term
    $X \to X_r$ location & scale change of g_j's
    $\frac{g_j-\bar{g}_j}{s_{g_j}}$
    free of unit
    $Z_r = X_rU$



- Notes:
  - interpretation is a problem --- little is gained if principal components are not interpretable
    Note. In the development of PCs, the criteria (LNp.12) do not consider interpretation
  - how many principal components are worth considering? plot λ_i, often the plot has a noticeable "elbow" --- the point, say k, at which further eigenvalues are negligible in size compared to the earlier ones ⇒ $(\lambda_1+\ldots+\lambda_k)/(\lambda_1+\ldots+\lambda_p)$ = proportion of total variation explained by the first k principal components
  - principal components do not use information from the response in dimension reduction. It is possible that a lesser principal component is actually very important in explaining/predicting the response. Dimension-reduction methods that utilize information about the response exist, such as
    Recall criteria in LNp 12
    explain small variation of X
    - partial least square ← relax one ★
    - sliced inverse regression (SIR)
    - principal Hessian directions (pHd)
    - projection pursuit regression
    - canonical correlation analysis
    - LASSO ← cf.



2 constraints ★ in PCs: (LNp.12)
① orthogonal coordinates (U) ∵ $U^TU=I$
② orthogonal projected data (Z) $Z^TZ=\text{diag}(\lambda)$

❖ **Reading**: Faraway (1st ed.), 9.1

Ridge regression

other better estimator than OLS ?

- **Q**: what is the problem? strong collinearity (i.e., X^TX close to singular) causes (1) numerical problem in calculating $(X^TX)^{-1}$; (2) $\hat{\beta}$ unstable; (3) large variance in $\hat{\beta}$
  OLS estimator

- ridge estimator: a method of combating strong collinearity [Note: It would be better to find out how collinearity occurs before doing ridge regression.] can compare $\hat{\gamma}$

different range of an x in different data

➢ centering and scaling predictors: $X \to F$, i.e., F^TF = correlation matrix of X (**Q**: why?), and centering response: $Y \to Z$, i.e., $Z = Y - \bar{Y}$,

Standardization

$\because + \lambda I$

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \;\Rightarrow\; Y = X\beta + \varepsilon$$

$$z = \boxed{\text{gone}}\; \gamma_1 f_1 + \ldots + \gamma_p f_p + \varepsilon \;\Rightarrow\; Z = F\gamma + \varepsilon$$

OLS est'or of γ $= (F^TF)^{-1}F^TZ$ unstable if there exists strong collinearity

Recall Location & Scale change (LN p.8-2)

Note: $\beta_i = \gamma_i / sd_i$, where sd_i is the sample standard deviation of x_i, $i=1, \ldots, p$

➢ ridge estimator: for $\lambda > 0$, $\hat{\gamma}_\lambda = (F^TF + \lambda I)^{-1}F^TZ = (F^TF + \lambda I)^{-1}F^TY$ [note: $F^T\mathbf{1} = \mathbf{0}$] no intercept

- $\lambda = 0 \Rightarrow \hat{\gamma}_0$ is the OLS estimator and $\lambda \to \infty \Rightarrow \hat{\gamma}_\infty = \mathbf{0}$

OLS est'or, $\lambda\uparrow$, shrinkage

- for an eigenvector u_i of F^TF and its corresponding eigenvalue λ_i, $(F^TF + \lambda I)u_i = (\lambda_i + \lambda)u_i \Rightarrow u_i$ is an eigenvector of $(F^TF + \lambda I)$ with corresponding eigenvalue $\lambda_i + \lambda\ (> \lambda_i)$ ⟵ strong collinearity $\Leftrightarrow$ some λ_i's ≈ 0

F^TF & $F^TF + \lambda I$ have same eigenvectors

- ridge estimator can remedy the problems caused by strong collinearity

(△) ➢ how to choose an appropriate λ? ⟶ criteria, e.g., crossvalidation, …

$F^TF = U\Lambda U^T$, $F^TF + \lambda I = U(\Lambda + \lambda I)U^T$, $UU^T = I$

There exists various methods *automatically* choosing a λ. However, the most popular method is through *ridge trace*: ⟶

why? ⟶ plot $\hat{\gamma}_\lambda$ against λ

Find a minimum value of λ (usually chosen in [0, 1]) (cf. $\sum_{i=1}^{p} \lambda_i = \text{trace}(F^TF) = p$, average of λ_i's $= 1$) after which $\hat{\gamma}_\lambda$ are moderately stable.

can be used in model selection