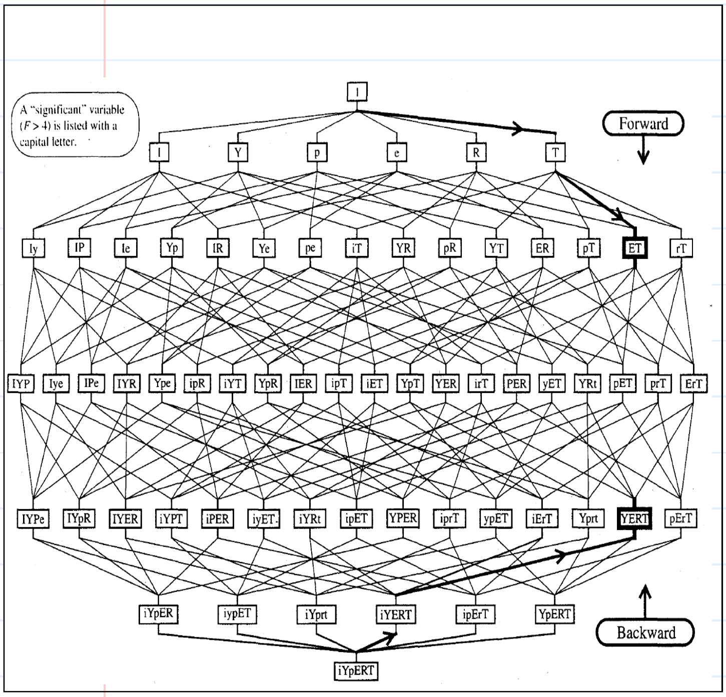
Model (variable) selection

- **Q**: what is a "model selection" problem?
  - consider the *full model*:

$$y = \beta_0 + \beta_1 g_1(x_1, \ldots, x_m) + \beta_2 g_2(x_1, \ldots, x_m) + \cdots + \beta_{k-1} g_{k-1}(x_1, \ldots, x_m) + \epsilon$$

For $1 \le i \le k-1$, should the term $\beta_i g_i$ be included in the *final fitted model*?

Example: 6 terms, I, Y, p, E, R, T



- (sub-)model: a subset set of all $k-1$ terms, e.g.,
  $\{\boldsymbol{1}, g_1, g_2\}$,
  $\{\boldsymbol{1}, g_2, g_4, g_5, g_{k-1}\}$, …
  - p = # of parameters in a sub-model
  - # of different sub-models = 2^{k-1}
- hierarchical structure of all sub-models (see graph)
- objective of model selection: select a "best" sub-model

Q: what is a good sub-model? We usually hope a good model to have
- high R^2
- $\hat{\sigma} \approx \sigma$
- not too many terms
- terms with significant t-tests

- **Q**: why bother to select a best subset of all terms?
  - simplicity: principle of Occam's Razor, removal of redundant terms results in a simpler model
  - unnecessary terms will cost d.f. and add noise to the estimation of other quantities ⇒ less precise test/C.I. and tend to increase the standard error
  - collinearity reduction: collinearity is caused by having too many terms trying to do same job
  - save cost : if model is used for prediction, can save time and/or money by not measuring redundant terms
- preliminary steps before performing variable selection
  - identify outliers and influential points --- may exclude them temporarily
  - add any terms, transformations, or (linear) combinations of the predictors or extra predictors that seem appropriate
- two types of variable selection procedures: *testing-based* and *criterion-based*
- testing-based procedure
  - **Recall**:
    - the p-value of t-test is an index of effect significance/importance
    - cannot simultaneously remove terms with insignificant p-values
  - backward elimination: (1) start with full model (all terms); (2) eliminate the term with the largest p-value greater than "α-to-remove" preset value; (3) refit the model and go to step (2); (4) stop when all p-values < α-to-remove

- forward selections: (1) start with no terms in the model (y~$\mathbf{1}$); (2) For terms not in the model, check their p-values if they are added to the model. Add the term with the smallest p-value less than "α-to-enter" preset value; (3) refit the model and go to step (2); (4) stop when all the p-values > α-to-enter
- stepwise regression: a combination of forward and backward and there are several variations on exactly how this is done. Roughly speaking, at any step, it can (1) select a new term, according to "α-to-enter", or (2) remove a term from model, according to "α-to-remove", or (3) stop
- drawbacks:
  - may miss "optimal" model because of its "one-at-a-time" adding/dropping
  - α-values (α-to-enter and α-to-remove) should not be treated too literally: because of multiple testing occurring
  - removal of less significant terms tends to increase the significance of the remaining terms $\Rightarrow$ may lead to overstate the importance of the remaining terms
  - The procedure is not directly linked to final objectives of regression, such as prediction or interpretation. It's only based on statistical significance of testing in its selection.
  - for prediction purpose, testing-based procedure tends to pick smaller models than desired

- criterion-based procedure (k: # of all parameters, including intercept; p: # of parameters in a sub-model; m_p: a sub-model with p parameters):
  - pick a criterion for judging the worth of a sub-model, consider all possible sub-models and pick those with best values of the criterion
  - # of all possible sub-models=2^{k-1} $\Rightarrow$ if k is large, computation may be too expensive, clever algorithm like "branch-and-bound" method can avoid it
  - adjusted R^2 (denoted by R_a^2) criterion: for a sub-model m_p,

    $R^2 = 1 - (RSS/TSS)$: not good, adding terms always increase R^2

    $R_a^2 = 1 - \{[RSS/(n-p)]/[TSS/(n-1)]\} = 1 - [(n-1)/(n-p)](1-R^2) = 1 - \left(\hat{\sigma}^2_{m_p}/\hat{\sigma}^2_{\text{null}}\right)$
    - will only increase when a term has some value
    - larger R_a^2 is better [notice the connection between R_a^2 and $\hat{\sigma}^2_{m_p}$]
  - PRESS (Predicted REsidual Sum of Square) criterion

    PRESS $= \sum_i \hat{\varepsilon}_{(i)}^2$, where $\hat{\varepsilon}_{(i)}$ are non-standardized jacknife residuals
    - smaller value of PRESS is better
    - more expensive computation than R_a^2
    - tends to pick bigger models ($\Rightarrow$ may be desirable for prediction purpose)
  - Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC)

    $$\text{AIC} = -2(\text{maximized log-likelihood}) + 2\,p$$
    $$\text{BIC} = -2(\text{maximized log-likelihood}) + \log(n)\,p$$

    for linear model, -2(maximized log-likelihood) $= n\log(RSS_{m_p}/n) +$ constant

- smaller value of AIC or BIC is better
- get a balance between model fit and model size: BIC penalizes larger models more heavily than AIC $\Rightarrow$ BIC tends to prefer smaller models

➢ Mallow's C_p statistics: MSE of prediction,

$$(1/\sigma^2)\ \Sigma_i\ \mathrm{E}[\ \hat{y}_i - \mathrm{E}(y_i)\ ]^2$$

would be a good criterion, which can be estimated by:

$$C_p = RSS_{m_p}/\hat{\sigma}^2_{\text{full}} + 2p - n,$$

where $\hat{\sigma}^2_{\text{full}}$ estimated from the model with all terms (full model) and RSS_{m_p} is obtained from a sub-model m_p

- cheap to compute
- closely related to R_a^2 and AIC, BIC
- under full model: $RSS_{\{full\ model\}} = (n-k)\hat{\sigma}^2_{\text{full}}$, so $C_k = k$ for full model
- for sub-models that fit: $\mathrm{E}(RSS_{m_p}) = (n-p)\sigma^2$, so $C_p \approx p$, i.e., C_p close to p implies the sub-model fits
- for sub-models that do not fit: $\mathrm{E}(RSS_{m_p}) >> (n-p)\sigma^2$ and $C_p >> p$
- it's usual to plot C_p against p. Models with small p and C_p around the $C_p=p$ line or less than p are desirable

➢ Note: C_p, R_a^2, AIC, BIC all trade-off fit in terms of RSS_{m_p} against complexity (p) $\Rightarrow$ we prefer models with smaller RSS_{m_p} and smaller p; however, $RSS_{m_p}\downarrow$ as $p\uparrow$

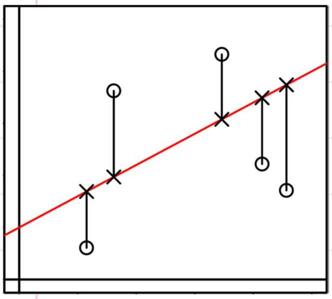
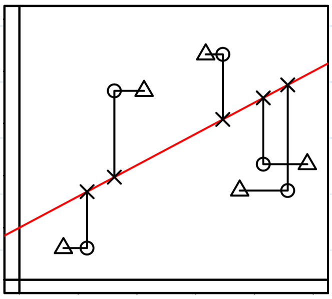
- ambiguity about the best model is possible. When several candidate models exist:
  - ➢ check if models make similar predictions? if yes, can make decision on the basis of cost; if no, do not pick one model arbitrary. Report a range of models.
  - ➢ interpretations qualitatively similar? if not, avoid strong conclusion and report a range of models
  - ➢ examine which has the best diagnostics
- Notes:
  - ➢ terms not in final model can still be correlated with the response
    $\Rightarrow$ not to say they are unrelated to the response;
    $\Rightarrow$ better to say they provide no additional explanatory effect beyond those terms included in final model
  - ➢ It's important to keep in mind that model selection should not be divorced from the underlying purpose of investigation
  - ➢ automatic variable selection are not guaranteed to be consistent with your goals. Use these methods as a guide only.
  - ➢ these methods do not consider the natural hierarchy in some models: For example, in polynomial model, higher-order terms (such as x_1^2, x_2^2, x_1x_2) should be considered only when corresponding lower-order terms (such as x_1 and x_2) have been included in the model $\Rightarrow$ not all sub-models are candidate models

❖ **Reading**: Faraway (1st ed.), chapter 8; W, 10.2, 10.3, 10.4 ❖ **Further reading**: D&S, chapter 15

Errors in the predictor

- Recall: $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is error that represent measurement error or unexplained variation in $\boldsymbol{Y}$, and "it's assumed that $\boldsymbol{X}$ are fixed values measured without error."
- **Q**: what if $\boldsymbol{X}$ are measured or observed with error? (examples)

  Note: compare the difference between the 2 statements: "$\boldsymbol{X}$ measured with errors" and "$\boldsymbol{X}$ are random variables, such as in sampling model"
- **Q**: what happens if you ignore errors in $\boldsymbol{X}$ and still use OLS estimator? Let us consider a simple example:

×: "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?
- a statistical model for Δ data:

  ideal "true" relationship is $\eta_i=\beta_0+\beta_1\xi_i$,

  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.,

  $$y_i=\beta_0+\beta_1\xi_i+\varepsilon_i=\beta_0+\beta_1 x_i+(\varepsilon_i-\beta_1\delta_i)$$

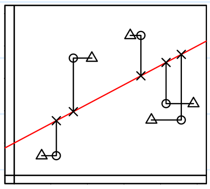
×: (ξ_i, η_i)
○: (ξ_i, y_i)
Δ: (x_i, y_i)

- **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 $cov(\varepsilon,\delta)=0$
  - 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:

    $$\hat{\beta}_1=[\Sigma(x_i-\bar{x})(y_i-\bar{y})]/\Sigma(x_i-\bar{x})^2$$
  - after some calculation, we can write

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

    $$E(\hat{\beta}_1)\approx\beta_1\times[\sigma_\xi^2/(\sigma_\xi^2+\sigma_\delta^2)]=\beta_1\times[1/(1+\sigma_\delta^2/\sigma_\xi^2)]\quad\Rightarrow\hat{\beta}_1 \text{ is biased}$$
  - 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 $\Rightarrow$ no worry
    - ratio is large, $|\hat{\beta}_1|$ is underestimated $\Rightarrow$ use *measurement error model*



- For multiple predictors, the usual effect of measurement errors on predictors is to bias the estimator of $\boldsymbol{\beta}$ in the direction of zero
- Prediction is not biased since future $\boldsymbol{X}$ will also be measured with errors. So, model for prediction should be built on $\boldsymbol{X}$'s measurement with error.

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

Collinearity

- collinearity: predictors are (linearly) related to each other
  - $X^T X$ is singular $\Rightarrow$ some predictors are linear combinations of others $\Rightarrow$ (exact) collinearity $\Rightarrow$ no unique estimate of β
  - $X^T X$ close to singular $\Rightarrow$ close to linear dependent among some predictors $\Rightarrow$ (approximate) collinearity or multicollinearity
- effect of collinearity:
  - estimated effects are unstable (can change magnitude or sign depending on the other predictors in the model) $\Rightarrow$ interpretation of estimated coefficients difficult
  - cause numerical problem in estimating β and associated quantities
  - $\text{var}(\hat{\beta}_j) = \sigma^2 (1/(1-R_j^2))(1/S_j)$, 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 $\Rightarrow$ when $R_j^2 \approx 1$, $\text{var}(\hat{\beta}_j)$ large $\Rightarrow$ t-test may fail to reveal significance, i.e., miss important g_j
  - *variance inflation factor*: $VIF_j = 1/(1-R_j^2)$ $\Rightarrow$ when S_j is fixed, VIF_j represents the increase in variance due to the collinearity (e.g., interpret VIF_j=16?)
- detection of collinearity:
  - examine correlations between predictors, i.e., $\text{cor}(g_k, g_j)$ $\Rightarrow$ 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 $\Rightarrow$ R_j^2 close to one or VIF_j much larger than one indicate a problem of collinearity

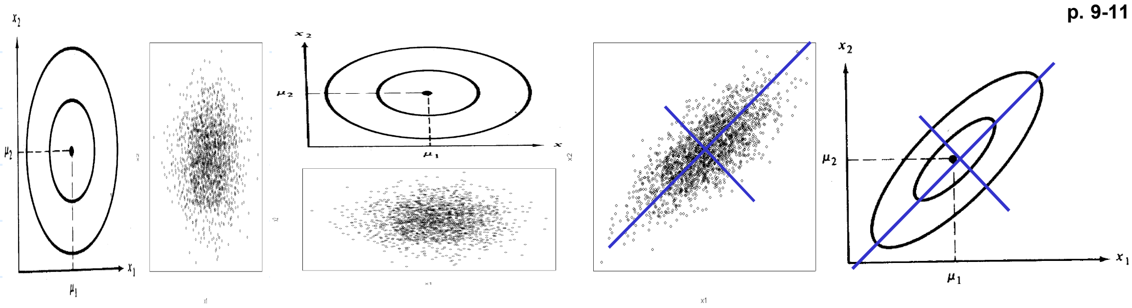
  - examine eigenvalues, $\lambda_1 \geq \ldots \geq \lambda_p$, of $X^T X$ $\Rightarrow$ small eigenvalues indicate a problem
    - condition number: $k=(\lambda_1/\lambda_p)^{1/2}$
    - rough rule: $k > 30$ is considered large
    - for each i, $(\lambda_1/\lambda_i)^{1/2}$ are worth considering $\Rightarrow$ 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
  - amputate some predictors if you can --- remember that collinearity happens because too many variables try to do the *same job* of explaining the response
  - do not conclude the predictors we drop have nothing to do with the response
  - techniques such as principle component regression, ridge regression, partial least squares, …, may help

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

Principal components

- Recall: $Y=X\beta+\varepsilon$. If X is orthogonal (i.e., $X^T X$ 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 ($Z^T Z$ is diagonal)

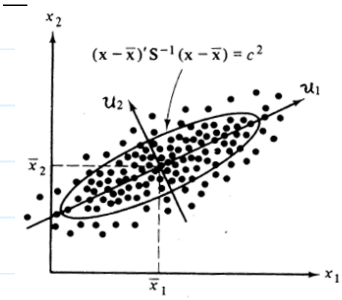
- orthogonality
  Recall: bivariate normal



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

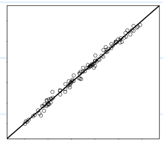
$$\begin{bmatrix} z_{11} \\ z_{21} \\ \cdots \\ z_{n1} \end{bmatrix} = u_{11}\begin{bmatrix} g_{11} \\ g_{21} \\ \cdots \\ g_{n1} \end{bmatrix} + u_{21}\begin{bmatrix} g_{12} \\ g_{22} \\ \cdots \\ g_{n2} \end{bmatrix} + \cdots + u_{p1}\begin{bmatrix} g_{1p} \\ g_{2p} \\ \cdots \\ g_{np} \end{bmatrix}$$

$$(g_{11}, g_{12}, \ldots, g_{1p}) \longrightarrow (z_{11}, z_{12}, \ldots, z_{1q})$$



➢ 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

- concept of *dimension reduction*:



➢ **Q**: take a look of the graph, the points are of 1-dim or of 2-dim?
  ⇒ very similar to a line ⇒ high correlation ⇒ data is 2-dim, but close to 1-dim

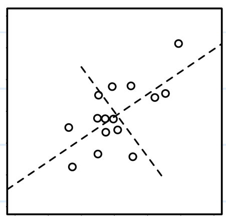
➢ replace large number of columns in $\boldsymbol{X}$ with small number of columns in $\boldsymbol{Z}$
  ⇒ simpler model, especially useful (1) when few linear combinations of $\boldsymbol{X}$ are enough to represent the variation in $\boldsymbol{X}$; (2) when $p > n$

- principal component (PC):

➢ transform $\boldsymbol{X}$ to $\boldsymbol{Z}$ which is orthogonal, but how?

➢ find $\boldsymbol{U}$ such that $\boldsymbol{Z}^T\boldsymbol{Z}$ is diagonal, i.e., $\boldsymbol{Z}^T\boldsymbol{Z} = \text{diag}(\lambda_1, \ldots, \lambda_p)$, where $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$

➢ since $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{U}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{U}$, to make $\boldsymbol{Z}^T\boldsymbol{Z}$ diagonal, we can choose columns of $\boldsymbol{U}$ are orthogonal eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$, then the $\lambda_1, \lambda_2, \ldots, \lambda_p$ are eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$
  - let $\boldsymbol{U}_j$ and λ_j be the j-th eigenvector and eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}$, then $(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{U}_j = \lambda_j \boldsymbol{U}_j$
  - $\boldsymbol{U}_k^T\boldsymbol{U}_j = 0$ for $k \neq j$ and $\|\boldsymbol{U}_j\| = 1$ for all j
  - $\boldsymbol{U}_k^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{U}_j = \lambda_j \boldsymbol{U}_k^T\boldsymbol{U}_j$, which equals 0 if $k \neq j$ and equals λ_j if $k = j$



➢ $\boldsymbol{Z}_1$ (=1st column of $\boldsymbol{Z}$) is called 1st principal component (PC),
  $\boldsymbol{Z}_2$ (=2nd column of $\boldsymbol{Z}$) is called 2nd principal component (PC), …

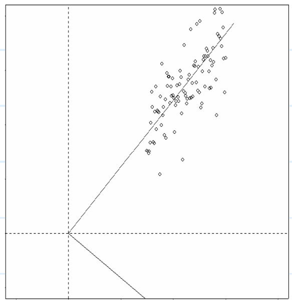
➢ another way to look at it:
  - $\boldsymbol{Z}_1$ = linear combination of columns of $\boldsymbol{X}$ that has maximum length², i.e, maximizing Σz_{i1}^2 (variation of $\boldsymbol{Z}_1$)
  - $\boldsymbol{Z}_2$ = linear combination of columns of $\boldsymbol{X}$ that is orthogonal to $\boldsymbol{Z}_1$ and has maximum length²
  - $\boldsymbol{Z}_3$ = linear combination of columns $\boldsymbol{X}$ that is orthogonal to $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$ and has maximum length²
  - …

➢ some properties:
  - $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{q\times q}$
  - zero eigenvalue ⇒ unidentifiable
  - λ_j = length² of $\boldsymbol{Z}_j = \Sigma_i z_{ij}^2$ [note: when $E(\boldsymbol{X}_j)=0 \Rightarrow E(\boldsymbol{Z}_j)=0 \Rightarrow \lambda_j \propto \text{var}(\boldsymbol{Z}_j)$]
  - $\lambda_1 + \ldots + \lambda_p = \text{tr}(\boldsymbol{X}^T\boldsymbol{X}) = \Sigma_j(\text{length}^2 \text{ of } \boldsymbol{X}_j)$ [note: when $E(\boldsymbol{X}_j)=0$, $\lambda_1 + \ldots + \lambda_p \propto \Sigma_j \text{var}(\boldsymbol{X}_j)$: total variation]
  - $\lambda_j/(\lambda_1 + \ldots + \lambda_p)$ = proportion of total variation explained by the jth PC

- **Q**: how to interpret $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_p$? Ans: compare the coefficients in eigenvectors
  - Ex 1: $\boldsymbol{Z}_1$=0.46GNP+0.32UnEm+0.46POP+0.46Year+... $\Rightarrow$ hard to give meaning
  - Ex 2: $\boldsymbol{Z}_1$=0.63(hw1)+0.57(hw2)+0.52(hw3) $\propto$ average homework scores;
    $\boldsymbol{Z}_2$ = 0.67(hw1)+0.08(hw2) - 0.75(hw3) $\propto$ difference between hw 1 and 3 scores

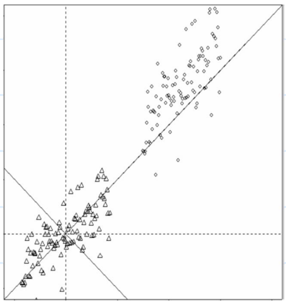
- variation on principal components
  - use $\boldsymbol{X}^T\boldsymbol{X}$ with/without constant term (without constant term $\Rightarrow$ PC's may not be orthogonal to constant term)

$\boldsymbol{X}^T\boldsymbol{X}$ without constant term

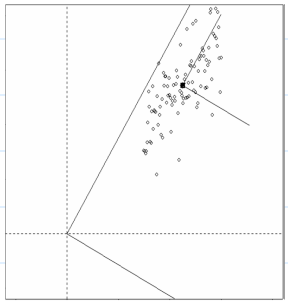


  - use covariance matrix of $\boldsymbol{X}$ (without constant term), i.e., $\boldsymbol{X}_v^T\boldsymbol{X}_v/(n-1)$ where $\boldsymbol{X}_v$ is formed by centering each g_j, to find eigenvectors $\boldsymbol{U}$ and eigenvalues. Then, $\lambda_j = \text{var}(z_j)$. The transformation $\boldsymbol{U}$ can be applied on $\boldsymbol{X}$ or $\boldsymbol{X}_v$ [PC's are orthogonal to constant term if transformation is applied on $\boldsymbol{X}_v$]
  - use correlation matrix of $\boldsymbol{X}$ (without constant term), i.e, $\boldsymbol{X}_r^T\boldsymbol{X}_r/(n-1)$, where $\boldsymbol{X}_r$ is formed by standardizing each g_j. To make sense, the transformation should be applied on $\boldsymbol{X}_r$ .Then, $\lambda_j = \text{var}(z_j)$ and PC's are orthogonal to constant term

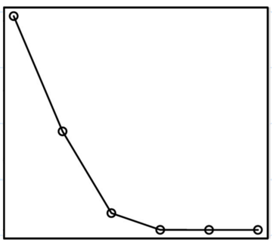
correlation matrix



covariance matrix



- Notes:
  - interpretation is a problem --- little is gained if principal components are not interpretable
  - 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 $\Rightarrow$ $(\lambda_1+\ldots+\lambda_k)/(\lambda_1+\ldots+\lambda_p)$ = 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
    - partial least square
    - sliced inverse regression (SIR)
    - principal Hessian directions (pHd)
    - projection pursuit regression
    - canonical correlation analysis
    - LASSO



❖ **Reading**: Faraway (1st ed.), 9.1

Ridge regression

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

- ridge regression: a method of combating strong collinearity [Note: It would be better to find out how collinearity occurs before doing ridge regression.]
  - ➢ centering and scaling predictors: $\boldsymbol{X} \rightarrow \boldsymbol{F}$, i.e., $\boldsymbol{F}^T\boldsymbol{F}$ = correlation matrix of $\boldsymbol{X}$ (**Q**: why?), and centering response: $\boldsymbol{Y} \rightarrow \boldsymbol{Z}$, i.e., $\boldsymbol{Z} = \boldsymbol{Y} - \bar{\boldsymbol{Y}}$,

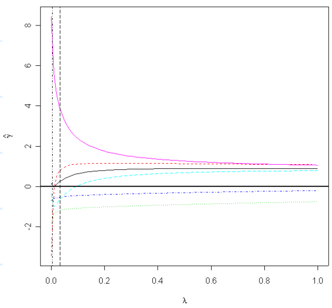
$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \quad \Rightarrow \quad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$z = \gamma_1 f_1 + \ldots + \gamma_p f_p + \varepsilon \quad \Rightarrow \quad \boldsymbol{Z} = \boldsymbol{F}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

  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 $\lambda>0$, $\hat{\boldsymbol{\gamma}}_\lambda = (\boldsymbol{F}^T\boldsymbol{F}+\lambda\boldsymbol{I})^{-1}\boldsymbol{F}^T\boldsymbol{Z} = (\boldsymbol{F}^T\boldsymbol{F}+\lambda\boldsymbol{I})^{-1}\boldsymbol{F}^T\boldsymbol{Y}$ [note: $\boldsymbol{F}^T\boldsymbol{1}=\boldsymbol{0}$]
    - $\lambda=0 \Rightarrow \hat{\boldsymbol{\gamma}}_0$ is the OLS estimator and $\lambda\rightarrow\infty \Rightarrow \hat{\boldsymbol{\gamma}}_\infty = \boldsymbol{0}$
    - for an eigenvector $\boldsymbol{u}_i$ of $\boldsymbol{F}^T\boldsymbol{F}$ and its corresponding eigenvalue λ_i, $(\boldsymbol{F}^T\boldsymbol{F}+\lambda\boldsymbol{I})\boldsymbol{u}_i = (\lambda_i+\lambda)\boldsymbol{u}_i \Rightarrow \boldsymbol{u}_i$ is an eigenvector of $(\boldsymbol{F}^T\boldsymbol{F}+\lambda\boldsymbol{I})$ with corresponding eigenvalue $\lambda_i+\lambda$ $(>\lambda_i)$
    - ridge estimator can remedy the problems caused by strong collinearity
  - ➢ how to choose an appropriate λ?

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

  plot $\hat{\boldsymbol{\gamma}}_\lambda$ against λ

  Find a minimum value of λ (usually chosen in [0, 1]) after which $\hat{\boldsymbol{\gamma}}_\lambda$ are moderately stable.



  - ➢ ridge estimator is biased

$$\mathrm{E}(\hat{\boldsymbol{\gamma}}_\lambda) = \boldsymbol{\gamma} - \lambda(\boldsymbol{F}^T\boldsymbol{F}+\lambda\boldsymbol{I})^{-1}\boldsymbol{\gamma} \;\Rightarrow\; \mathrm{Bias}(\hat{\boldsymbol{\gamma}}_\lambda) = -\lambda(\boldsymbol{F}^T\boldsymbol{F}+\lambda\boldsymbol{I})^{-1}\boldsymbol{\gamma}$$

$$\text{Total of variances: } \mathrm{tr}[\mathrm{cov}(\hat{\boldsymbol{\gamma}}_\lambda)] = \sigma^2 \Sigma_j \lambda_j (\lambda_j+\lambda)^{-2}$$

  Mean Square Error:

$$\mathrm{MSE}(\hat{\boldsymbol{\gamma}}_\lambda) = \mathrm{E}[(\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma})(\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma})^T] = \mathrm{cov}(\hat{\boldsymbol{\gamma}}_\lambda) + \mathrm{Bias}(\hat{\boldsymbol{\gamma}}_\lambda)\mathrm{Bias}(\hat{\boldsymbol{\gamma}}_\lambda)^T$$

  Total Mean Square Error: $\mathrm{tr}[\mathrm{MSE}(\hat{\boldsymbol{\gamma}}_\lambda)] = \mathrm{tr}[\mathrm{cov}(\hat{\boldsymbol{\gamma}}_\lambda)] + \lambda^2\boldsymbol{\gamma}^T(\boldsymbol{F}^T\boldsymbol{F}+\lambda\boldsymbol{I})^{-2}\boldsymbol{\gamma}$

  The total MSE of ridge estimator can be lower than OLS estimator when strong collinearity exists; the price we pay is, of course, the bias.

- why ridge regression can work? $\Rightarrow$ add *additional information* to remove collinearity. The following conditions, that all lead to ridge estimator, can offer some insights:
  - ➢ Suppose $\exists$ an $n\times p$ matrix $\boldsymbol{V}$ s.t. $\boldsymbol{V}^T\boldsymbol{F}=\boldsymbol{0}$ and $\boldsymbol{V}^T\boldsymbol{Z}=\boldsymbol{0}$. Let $\boldsymbol{W}_{n\times p}=\lambda^{1/2}\boldsymbol{V}(\boldsymbol{V}^T\boldsymbol{V})^{-1/2}$. Then, (1) $\boldsymbol{W}^T\boldsymbol{W}=\lambda\boldsymbol{I}$, (2) $\boldsymbol{W}^T\boldsymbol{F}=\boldsymbol{0}$, and (3) $\boldsymbol{W}^T\boldsymbol{Z}=\boldsymbol{0}$. The OLS estimator of the model: $\boldsymbol{Z} = (\boldsymbol{F}+\boldsymbol{W})\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ is:

$$[(\boldsymbol{F}+\boldsymbol{W})^T(\boldsymbol{F}+\boldsymbol{W})]^{-1}(\boldsymbol{F}+\boldsymbol{W})^T\boldsymbol{Z} = (\boldsymbol{F}^T\boldsymbol{F}+\lambda\boldsymbol{I})^{-1}\boldsymbol{F}^T\boldsymbol{Z}$$

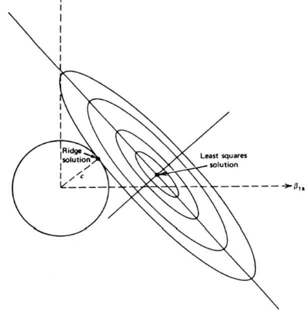
  $\Rightarrow$ suitably *disturbing* $\boldsymbol{F}$ by a small amount to remove strong collinearity
  - ➢ Consider the model $\mathcal{Z} = \mathcal{F}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, where $\mathcal{Z} = [\boldsymbol{Z}^T\ \boldsymbol{0}]^T$, $\mathcal{F} = [\boldsymbol{F}^T\ \boldsymbol{W}^T]^T$. Its OLS estimator is:

$$(\mathcal{F}^T\mathcal{F})^{-1}\mathcal{F}^T\mathcal{Z} = (\boldsymbol{F}^T\boldsymbol{F}+\lambda\boldsymbol{I})^{-1}\boldsymbol{F}^T\boldsymbol{Z}$$

  $\Rightarrow$ *adding additional "cases"* to the data set to remove strong collinearity

- Minimize $RSS=(\boldsymbol{Z}-\boldsymbol{F}\boldsymbol{\gamma})^T(\boldsymbol{Z}-\boldsymbol{F}\boldsymbol{\gamma})$ subject to this constraint $\boldsymbol{\gamma}^T\boldsymbol{\gamma}\leq c^2$.
  The solution is the ridge estimator that satisfies $\hat{\boldsymbol{\gamma}}_\lambda^T\hat{\boldsymbol{\gamma}}_\lambda=c^2$

- Bayesian viewpoint: put a multivariate normal prior $\boldsymbol{N}(\boldsymbol{0}, \lambda^{-1}\boldsymbol{I})$ on $\boldsymbol{\gamma}$. Then, the Bayes estimator is the ridge estimator. $\Rightarrow$ choice of a larger λ implies $\boldsymbol{\gamma}$ were more likely to be small, and vice versa.



- an implicit pre-assumption in ridge regression: coefficients (after normalizing) are not likely to be very large

❖ **Reading**: Faraway (1st ed.), 9.3 ❖ **Further reading**: D&S, chapter 17

Analysis strategy and model uncertainty

- you have learned
  - Parameter estimation and testing: LS estimator, generalized (weighted) LS, ridge estimator, t-test, F-tests, lack-of-fit, C.I., R^2, prediction, ...
  - Diagnostics (checking assumptions): such as constant variance, linearity, normality, outliers, influential points, serial correlation, collinearity, ...
  - Transformation: transforming the response and/or the predictors, Box-Cox, polynomial regression, broken line, spline, principal component, ...
  - Variable selection: testing-based and criterion-based procedures
- **Q**: what order should these be done? should procedures be repeated at later stage? when should we stop?

- a recommended analysis strategy:

a reasonable regression model → Diagnostics → Transformation → Variable Selection → Diagnostics → Stop

  Note: there is no hard-and-fast rules about how it should be done. Regression analysis is a search for structure in data. Better to try a variety of orders.

- Danger of doing too much analysis. More transformations, permutations of leaving out influential points and outliers you have done, better fitting model you will find --- however, may lead to over-fitting or no guarantee that the model is a good representation of the underlying system.

  "If you torture the data long enough, it will confess to anything."

  - avoid complex models for small dataset
  - try to obtain new data to validate your proposed model
  - use past experience with similar data to guide the choice of model
- model multiplicity: Same data can support different models, that sometimes lead to different conclusions. Personal preference, different analysis strategy, or changes in order of analysis components may result in different models. Always try to take a second independent look at the data.
- model uncertainty: Usually, inferences are based on the assumption that the selected final model was fixed in advance and so only reflect uncertainty concerning the parameters of that fixed model. **Q**: should we consider the variation caused by model multiplicity? From this viewpoint, the reported standard errors are usually too small.

❖ **Reading**: Faraway (1st ed.), chapter 10