➢ ridge estimator is biased $\overset{cf.}{\longleftrightarrow}$ OLS est'or $\hat{\gamma}_0$ is BLUE, but $tr[cov(\hat{\gamma}_0)] = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$

$(F^TF+\lambda I)^{-1}F^TE(Z)=$ E$(\hat{\gamma}_\lambda) = \gamma - \lambda(F^TF+\lambda I)^{-1}\gamma \Rightarrow$ Bias$(\hat{\gamma}_\lambda) = -\lambda(F^TF+\lambda I)^{-1}\gamma$

$\boxed{(F^TF+\lambda I) - \lambda I = F^TF}$ $\quad F\gamma$

Total of variances: tr[cov$(\hat{\gamma}_\lambda)$] = $\sigma^2 \Sigma_j\, \lambda_j (\lambda_j+\lambda)^{-2}$ $\overset{cf.}{\longleftarrow}$ use (Δ) in LNp.15

trace = sum of eigenvalues $\quad cov(\hat{\gamma}_\lambda) = E^*(\hat{\gamma}_\lambda\hat{\gamma}_\lambda^T) = \sigma^2(F^TF+\lambda I)^{-1}(F^TF)(F^TF+\lambda I)^{-1}$

$E(\hat{\theta}-\theta)^2 = Var(\hat{\theta}) + Bias(\hat{\theta})^2$ ← Mean Square Error:

MSE$(\hat{\gamma}_\lambda)$ = E[$(\hat{\gamma}_\lambda - \gamma)(\hat{\gamma}_\lambda - \gamma)^T$] = cov$(\hat{\gamma}_\lambda)$ + Bias$(\hat{\gamma}_\lambda)$ Bias$(\hat{\gamma}_\lambda)^T$ $\overset{tr}{\longrightarrow} \|Bias(\hat{\gamma}_\lambda)\|^2$

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

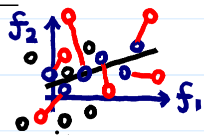
The total MSE of ridge estimator can be lower than OLS estimator when strong collinearity exists; the price we pay is, of course, the bias.

- why ridge regression can work? ⇒ add *additional information* to remove collinearity. The following conditions, that all lead to ridge estimator, can offer some insights:

➢ Suppose ∃ an $n\times p$ matrix V s.t. $V^TF=0$ and $V^TZ=0$. Let $W_{n\times p}=\lambda^{1/2}V(V^TV)^{-1/2}$. Then, (1) $W^TW=\lambda I$, (2) $W^TF=0$, and (3) $W^TZ=0$. The OLS estimator of the model: $Z = \underbrace{(F+W)}_{G}\gamma + \varepsilon$ is: ⟶ OLS est'or based on G: $(G^TG)^{-1}G^TZ$

$Z = F\gamma + \varepsilon \overset{cf.}{\longleftarrow}$

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



⇒ suitably *disturbing* F by a small amount to remove strong collinearity

➢ Consider the model $\mathcal{Z} = \mathcal{F}\gamma + \varepsilon$, where $\mathcal{Z}_{2n\times 1} = [Z^T\ 0]^T$, $\mathcal{F}_{2n\times p} = [F^T\ W^T]^T$. $\begin{bmatrix} F \\ W \end{bmatrix}$

Its OLS estimator is: $\begin{bmatrix} Z \\ 0 \end{bmatrix}$ $\quad \mathcal{F}^T\mathcal{Z} = F^TZ + W^T0$ $\quad \mathcal{F}^T\mathcal{F} = F^TF + W^TW = F^TF + \lambda I$

pseudo-data to present additional information

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

⇒ *adding additional "cases"* to the data set to remove strong collinearity

➢ Minimize $RSS=(Z-F\gamma)^T(Z-F\gamma)$ subject to this constraint $\gamma^T\gamma \le c^2$.

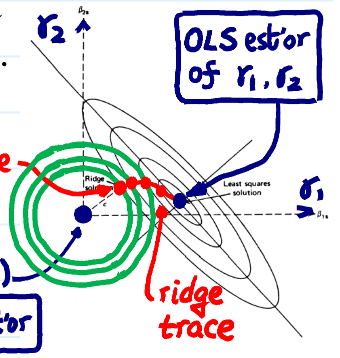
⟵ a quadratic form: $(\gamma-\hat{\gamma}_0)^TF^TF(\gamma-\hat{\gamma}_0)$

The solution is the ridge estimator that satisfies $\hat{\gamma}_\lambda^T\hat{\gamma}_\lambda = c^2$. ← L^2-norm $\overset{cf.}{\longleftrightarrow}$ L^1-norm $\sum_{i=1}^{p}|\hat{\gamma}_{i,\lambda}| = c$ → LASSO

shrinkage methods $\quad$ This explains why need to standardize X ($X \to F$) $\quad$ meaning? regularization $\quad \overset{cf.}{\longrightarrow}$ double exponential distribution

➢ Bayesian viewpoint: put a multivariate normal prior $N(0, \lambda^{-1}I)$ on γ. Then, the Bayes estimator is the ridge estimator. ⇒ choice of a larger λ implies γ 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 $\quad$ data → OLS est'or, prior → $\underline{0}$ } → Bayes est'or

➢ Parameter estimation and testing: LS estimator, generalized (weighted) LS, ridge estimator, robust 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 models, broken line, spline, principal component, dummy variables, ...

➢ 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 → final fitted model

  sensitive to unusual observations

  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.
  - 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

  CV: training data, validation data, test data

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

- 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 averaging for prediction purpose ← cf. Bayesian approach

- 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