· (Deviations from Means) Let $S = X_1^\top X_1 - n\bar{X}\bar{X}^\top = Z^\top Z$ and $Z = X_1 - 1\bar{X}^\top$ then $\hat\beta_0 = \bar y - \bar{X}^\top \hat b = \frac{1}{n}1^\top Y - \bar{X}^\top \hat b$,
$\hat b = S^{-1}(X_1^\top Y - n\bar y \bar X) = (Z^\top Z)^{-1}Z^\top Y$ and $\mathrm{var}(\hat b) = (Z^\top Z)^{-1}\sigma^2$,
$\mathrm{var}(\hat\beta_0) = \frac{\sigma^2}{n} + \bar X^\top \mathrm{var}(\hat b)\bar X$, $\mathrm{cov}(\hat\beta_0, \hat b^\top) = -\bar X^\top \mathrm{var}(\hat b)$.
· $\frac{[n-r(X)]\hat\sigma^2}{\sigma^2} = \frac{SSE}{\sigma^2} = \frac{1}{\sigma^2}Y^\top(I-H)Y \sim \chi^2_{n-r(X)}$,
$\frac{Y^\top Y}{\sigma^2} \sim \chi^2(n, \frac{\beta^\top X^\top X\beta}{2\sigma^2})$, $\frac{SSR}{\sigma^2} \sim \chi^2(r(X), \frac{\beta^\top X^\top X\beta}{2\sigma^2})$,
$\frac{MSR}{MSE} \sim F[r(X), n-r(X), \frac{\beta^\top X^\top X\beta}{2\sigma^2}]$.
· $\mathrm{SST}_m = SST - SSM = Y(I - \frac{1}{n}11^\top)Y$. $\frac{SSM}{\sigma^2} \sim \chi^2(1, \frac{(1^\top X\beta)^2}{2n\sigma^2})$,
$\mathrm{SSR}_m = SSR - SSM = \hat b^\top Z^\top Y = \hat b^\top (Z^\top Z)\hat b$,
$\frac{SST_m}{\sigma^2} \sim \chi^2(n-1, \frac{\beta^\top X^\top X\beta - \frac{1}{n}(1^\top X\beta)^2}{2\sigma^2})$, $\frac{\mathrm{SSR}_m}{\sigma^2} \sim \chi^2(r-1, \frac{b^\top(Z^\top Z)b}{2\sigma^2})$, $\frac{MSM}{MSE} \sim F[1, n-r(X), \frac{(1^\top X\beta)^2}{2n\sigma^2}]$,
$\frac{MSR_m}{MSE} \sim F[r(X)-1, n-r(X), \frac{b^\top(Z^\top Z)b}{2\sigma^2}]$.
· (General linear hypothesis) The general hypothesis we consider is $H_0: K^\top\beta = m$ where β is the $(k+1)$-dimensional vector of parameters of the model, $K^\top$ is any full row rank matrix of size $s\times(k+1)$ and m is a $s\times 1$ vector of specified constants. Define
$Q = (K^\top\hat\beta - m)^\top[K^\top(X^\top X)^{-1}K]^{-1}(K^\top\hat\beta - m)$,
$\frac{Q}{\sigma^2} \sim \chi^2(s, \frac{1}{2\sigma^2}(K^\top\beta - m)^\top[K^\top(X^\top X)^{-1}K]^{-1}(K^\top\beta - m))$,
$F(H) = \frac{Q/s}{SSE/[n-r(X)]} \sim F(s, n-r(X), (K^\top\beta-m)^\top[K^\top(X^\top X)^{-1}K]^{-1}(K^\top\beta - m))$.
· (Estimation under the null hypothesis) When H_0 is true, $\tilde\beta$ is derived so as to minimize the least squares objective function subject to the constraint $K^\top\beta = m$, $\tilde\beta = \hat\beta - (X^\top X)^{-1}K(K^\top(X^\top X)^{-1}K)^{-1}(K^\top\hat\beta - m)$ and $SSE_{H_0} = SSE + Q$.

Models not of Full Rank

· (Properties of $\beta^0 = GX^\top y$) (1) $E(\beta^0) = GX^\top X\beta$. (2) $\mathrm{Var}(\beta^0) = GX^\top XG^\top\sigma^2$. (3) $\hat y = X\beta^0 = XGX^\top y$. (4) $E(\hat y) = XGX^\top X\beta = X\beta$. (5) $\mathrm{SSE} = y^\top(I - XGX^\top)y$. (6) $SSR = y^\top XGX^\top y = (\beta^0)^\top X^\top y$. (7) $\hat\sigma^2 = \frac{SSE}{n-r(X)}$ is an unbiased estimator of σ^2. (8) $SSR_m = y^\top(XGX^\top - \frac{11^\top}{n})y$.
· (Distributional Properties) (1) $\beta^0 \sim N(GX^\top X\beta, GX^\top XG^\top\sigma^2)$. The covariance matrix is singular. (2) $\frac{SSE}{\sigma^2} \sim \chi^2_{n-r(X)}$, $\frac{SSR}{\sigma^2} \sim \chi^2(r(X), \frac{1}{\sigma^2}\beta^\top X^\top X\beta)$, $\frac{MSR}{MSE} \sim F(r(X), n-r(X), \frac{1}{2\sigma^2}\beta^\top X^\top X\beta)$.
· (Identifiability) Formally, the parameter β is identifiable if $f(\beta_1) = f(\beta_2)$ implies that $\beta_1=\beta_2$ for any β_1 and β_2. More generally, the vector-valued function $g(\beta)$ is identifiable if $f(\beta_1)=f(\beta_2)$ implies that $g(\beta_1)=g(\beta_2)$.
[**Proposition**] In a linear model for which X is of full rank, β is identifiable.
[**Proposition**] A function $g(\beta)$ is identifiable if and only if $g(\beta) = (h\circ f)(\beta)$ for some function h.
· (Estimable functions) If a vector t exists such that $t^\top E(y) = q^\top\beta$, then $q^\top\beta$ is said to be estimable. Linear combinations of estimable functions are estimable.
· $E(t^\top y) = t^\top X\beta = q^\top\beta \Rightarrow t^\top X = q^\top$ for some t. This is equivalent to saying that q is in the row space of X.
[Theorem] (Gauss-Markov Theorem) The best linear unbiased estimator of the **estimable** function $q^\top\beta$ is $q^\top\beta^0$.
(Sketch of proof) $\mathrm{var}(q^\top\beta^0) = t^\top XGX^\top t\sigma^2 = q^\top Gq\sigma^2$, Suppose $k^\top y$ is another linear unbiased estimator of $q^\top\beta$, $\mathrm{cov}(q^\top\beta^0, k^\top y) = \mathrm{cov}(q^\top GX^\top y, k^\top y) = q^\top Gq\sigma^2$. $\mathrm{var}(q^\top\beta^0 - k^\top y) = \mathrm{var}(q^\top\beta^0) + \mathrm{var}(k^\top y) - 2\mathrm{cov}(q^\top\beta^0, k^\top y) = \mathrm{var}(k^\top y) + q^\top Gq\sigma^2 - 2q^\top Gq\sigma^2 = \mathrm{var}(k^\top y) - \mathrm{var}(q^\top\beta^0) \ge 0$.
[Theorem] (Test of Estimability) The linear function $q^\top\beta$ is estimable if and only if $q^\top H = q^\top$ where $H = GX^\top X$.
· (Testable hypothesis) A testable hypothesis is $H_0: K^\top\beta = m$ such that $K^\top = T^\top X$ for some matrix $T^\top$ of order $r\times n$. The matrix $K^\top$ of size $r\times p$ is always of full-row rank.
· (Hypothesis testing) $K^\top\beta^0 - m \sim N(K^\top\beta - m, K^\top GK\sigma^2)$, $Q = (K^\top\beta^0 - m)^\top(K^\top GK)^{-1}(K^\top\beta^0 - m)$. Then $\frac{Q}{\sigma^2} \sim \chi^2(r, (K^\top\beta - m)^\top(K^\top GK)^{-1}(K^\top\beta - m)/2\sigma^2)$,
$F(H) = \frac{Q/r}{SSE/(n-r(X))} \sim F(r, n-r(X), (K^\top\beta-m)^\top(K^\top G K)^{-1}(K^\top\beta - m)/2\sigma^2)$.

The Bias-Variance Trade-off

· (The bias-variance trade-off) Consider the expected prediction error of an estimate $\hat f(z)$ at a particular point z_0,
$\mathrm{PE}(z_0) = E_{Y|Z=z_0}[(Y - \hat f(Z))^2 \mid Z = z_0] = \sigma_\epsilon^2 + \mathrm{Bias}^2(\hat f(z_0)) + \mathrm{Var}(\hat f(z_0)) = \sigma_\epsilon^2 + \mathrm{MSE}(\hat f(z_0))$.

Ridge Regression

· (Ridge regression) Minimize $(Y - Z\beta)^\top(Y-Z\beta)$ s.t. $\sum_{j=1}^p\beta_j^2 \leqslant t$. $\hat\beta^{\text{ridge}} = \arg\min_\beta \|Y - Z\beta\|_2^2 + \lambda\|\beta\|_2^2$. Its solution may have smaller average PE than $\hat\beta^{ls}$. Note that $\mathrm{PRSS}(\beta)_{\ell_2}$ is convex, and hence has a unique solution. $\hat\beta_\lambda^{\text{ridge}} = (Z^\top Z + \lambda I_p)^{-1}Z^\top Y$.
· By convention, (1) Z is assumed to be standardized (mean = 0, variance = 1) and (2) Y is assumed to be centered.
· ($\hat\beta_\lambda^{\text{ridge}}$ is biased). Let $R = Z^\top Z$ and Z has full column rank. $\hat\beta_\lambda^{\text{ridge}} = (Z^\top Z + \lambda I_p)^{-1}Z^\top Y = (R + \lambda I_p)^{-1}R(R^{-1}Z^\top Y) = [R(I_p + \lambda R^{-1})]^{-1}R\hat\beta^{ls} = (I_p + \lambda R^{-1})^{-1}\hat\beta^{ls}$. Thus, $E(\hat\beta_\lambda^{\text{ridge}}) = E[(I_p+\lambda R^{-1})^{-1}\hat\beta^{ls}] = (I_p + \lambda R^{-1})^{-1}\beta \ne \beta \quad (\forall\beta\ne0, \lambda\ne0)$ Note that 1 is not an eigenvalue of the matrix $(I_p + \lambda R^{-1})^{-1}$ as $\lambda \ne 0$ and R is positive definite.
· (SVD) By the singular value decomposition, $Z = UDV^\top$ where U is an $n\times p$ orthogonal matrix, D is $p\times p$ diagonal matrix and $V^\top$ is a $p\times p$ orthogonal matrix. $\hat\beta_\lambda^{\text{ridge}} = V\,\mathrm{diag}(\frac{d_j}{d_j^2+\lambda})U^\top Y$, by using $Z^\top Z = (UDV^\top)^\top(UDV^\top) = VD^\top U^\top UDV^\top = VD^2V^\top$.
· (Orthonormal Z) Z is orthonormal, then $Z^\top Z = I_p$. Let $\hat\beta^{ls}$ denote the LS solution, then $\hat\beta_\lambda^{\text{ridge}} = \frac{1}{1+\lambda}\hat\beta^{ls}$ and $\hat\beta_j^{\text{ridge}} = \frac{\hat\beta_j^{ols}}{1+\lambda}$. The optimal choice of λ minimizing the expected prediction error is $\lambda^* = \frac{p\sigma^2}{\sum_{j=1}^p\beta_j^2}$ where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the true coefficient vector.
· Variance of the ridge regression estimate is $\mathrm{Var}(\hat\beta^{\text{ridge}}) = \sigma^2 W_\lambda(Z^\top Z)^{-1}W_\lambda = \sigma^2(Z^\top Z + \lambda I_p)^{-1}Z^\top Z(Z^\top Z+\lambda I_p)^{-1}$ where $W_\lambda = (Z^\top Z + \lambda I_p)^{-1}Z^\top Z$. Recall that $\hat\beta^{\text{ridge}} = W_\lambda\hat\beta^{ls}$
· The bias of the ridge regression estimate is $\mathrm{bias}(\hat\beta^{\text{ridge}}) = -\lambda W_\lambda(Z^\top Z)^{-1}\beta = -\lambda(Z^\top Z + \lambda I_p)^{-1}\beta$.
· It can be shown that (1) the total variance $\sum_{j=1}^p \mathrm{Var}(\hat\beta_j^{\text{ridge}})$ is a monotone decreasing sequence with respect to λ. (2) the total squared bias $\sum_{j=1}^p\mathrm{Bias}^2(\hat\beta_j^{\text{ridge}})$ is a monotone increasing sequence with respect to λ.
· (Existence Theorem) There always exists a λ such that the MSE of $\hat\beta^{\text{ridge}}$ is less than the MSE of $\hat\beta^{\text{OLS}}$.

LASSO

· (LASSO) Minimize $(Y - Z\beta)^\top(Y - Z\beta)$, s.t. $\sum_{j=1}^p|\beta_j| \leqslant t$. It is often convenient to rewrite the LASSO problem in the so-called Lagrangian form : $\text{minimize}_\beta (Y-Z\beta)^\top(Y-Z\beta) + \lambda\|\beta\|_1$, where $\lambda\ge0$ is the so-called tuning parameter. The resulting lasso estimate is denoted by $\hat\beta_\lambda^{\text{lasso}}$.
· There is usually a factor $1/(2n)$ or $1/n$ appearing in front of $(Y-Z\beta)^\top(Y-Z\beta)$. This kind of standardization makes λ values comparable for different sample sizes (useful for cross-validation).
· We typically center the response and standardize the predictors so that each column is centered ($\frac1n\sum_{i=1}^n Z_{ij} = 0$) and has unit variance ($\frac1n\sum_{i=1}^n Z_{ij}^2 = 1$).
· (KKT condition) The necessary and sufficient conditions for a solution to LASSO take the form

$$-\langle z_j, Y - Z\beta\rangle + \lambda s_j = 0, \quad j = 1,\ldots,p$$

Here each s_j is an unknown quantity equal to $\mathrm{sign}(\beta_j)$ if $\beta_j\ne0$ and some value lying in $[-1,1]$ otherwise.
· (Orthonormal Z) Consider an orthogonal design case with $Z^\top Z = \mathrm{I}$. The LASSO method is equivalent to : solve β_j's componentwisely by solving : $\min_{\beta_j}(\beta_j - \hat\beta_j^{ls})^2 + \lambda|\beta_j|$. The solution to the above problem is

$$\hat\beta_j^{\text{lasso}} = \mathrm{sgn}(\hat\beta_j^{ls})(|\hat\beta_j^{ls}| - \frac{\lambda}{2})_+ = \begin{cases} \hat\beta_j^{ls} - \frac{\lambda}{2} & \text{if} \quad \hat\beta_j^{ls} > \frac{\lambda}{2} \\ 0 & \text{if} \quad |\hat\beta_j^{ls}| \leqslant \frac{\lambda}{2} \\ \hat\beta_j^{ls} + \frac{\lambda}{2} & \text{if} \quad \hat\beta_j^{ls} < -\frac{\lambda}{2} \end{cases}$$

Non-Normal Case

· Assume that $Y = X\beta + \epsilon$, $f_\epsilon(\cdot)$ is unknown. The score function is defined as $\nabla_\beta \log f_\epsilon(Y - X\beta) = (-X)^\top(\dot f_\epsilon(Y-X\beta)/f_\epsilon(Y-X\beta))$. Let $g_n(\beta) = \sum_i(\dot f_\epsilon(y_i - x_i^\top\beta)/f_\epsilon(y_i - x_i^\top\beta))x_i$. Then $g_n(\beta)=0$ gives $\hat\beta_n$.
· $\sqrt n(\hat\beta_n - \beta_0) = -(\frac1n\sum_i\dot g_i(\beta_0))^{-1}(\frac{1}{\sqrt n}\sum_i g_i(\beta_0))$. Then $\sqrt n(\hat\beta_n - \beta_0) \xrightarrow{d} \mathcal N(0, A^{-1}B(A^{-1})^\top)$, where $A = E\dot g(\beta_0)$ (Hessian) and $B = \mathrm{Var}(g(\beta))$. To ensure consistency, there should be $Eg(\beta_0) = 0$.

Matrix Derivative

Formula : $df = \mathrm{tr}\left(\frac{\partial f}{\partial X}^\top dX\right)$
1. $d(XY) = (dX)Y + XdY; d\left(X^\top\right) = (dX)^\top; d\,\mathrm{tr}(X) = \mathrm{tr}(dX)$
2. $dX^{-1} = -X^{-1}dXX^{-1}$.
3. $d|X| = \mathrm{tr}\left(X^{\#}dX\right)$, $X^{\#}$ represents the adjugate matrix of X. When X is invertible $d|X| = |X|\,\mathrm{tr}\left(X^{-1}dX\right)$.
4. $d(X\odot Y) = dX\odot Y + X\odot dY$, $\odot$ represents the element-wise multiplication of matrices X and Y of the same size.
5. $d\sigma(X) = \sigma'(X)\odot dX, \sigma(X) = [\sigma(X_{ij})]$ is an element-wise scalar function operation, $\sigma'(X) = [\sigma'(X_{ij})]$ is the element-wise derivative.

$u_t = \Lambda_{ut} w_{vt} + \varepsilon_{ut} \quad t = 1, \cdots, T$

$w_k = (w_{u1k}, \cdots w_{uTk})^T, \ \varepsilon_{wk} = (\varepsilon_{u1k} \cdots \varepsilon_{uTk})^T$

$w_k = \Lambda_{\xi}\xi_{v+k} + u_k \quad k = 1, \cdots q$

$\xi = (\xi_1^T \cdots \xi_v^T, \xi_{v+1}^T \cdots \xi_{v+q}^T)$

$$\begin{pmatrix} u_{t1} \\ u_{t2} \\ \vdots \\ u_{tp} \end{pmatrix} = \begin{pmatrix} \lambda_{ut,11} & \lambda_{ut,12} & \cdots & \lambda_{ut,1q} \\ \lambda_{ut,21} & \lambda_{ut,22} & \cdots & \lambda_{ut,2q} \\ \vdots & \vdots & & \vdots \\ \cdots & \cdots & \cdots & \lambda_{ut,pq} \end{pmatrix} \begin{pmatrix} w_{ut1} \\ w_{ut2} \\ \vdots \\ w_{utq} \end{pmatrix} + \begin{pmatrix} \varepsilon_{ut1} \\ \varepsilon_{ut2} \\ \vdots \\ \varepsilon_{utp} \end{pmatrix}$$

$\eta = B + \Pi\eta + \Gamma F(\xi) + \delta$

$\eta = B + \Lambda_\delta G(\xi) + \delta$