Variable Selection

Consider : $Y = X\beta + \varepsilon$. $\varepsilon \sim N_n(0, \sigma^2 I_n)$ where

$Y \in \mathbb{R}^n$. $\beta \in \mathbb{R}^p$ (unknown). $X \in M^{n\times p}$. $r(X) = p$.

H_0 : $(0\ I_{p-1})\beta = 0$ is test for whether all the variables influence Y (data). $\beta = (\beta_0\ \beta_1 \cdots \beta_{p-1})$

$X = (J_n\ X_1 \cdots X_{p-1})$. β_0 is fixed.

H_0: $\beta_i = 0$ is test for single one.

(1) Consequence of Selection:

- We want to reduce the number of variables but retain the accuracy of estimation at the same time.

suppose $X_q \in M^{n\times q}$. $\beta_q \in \mathbb{R}^q$. $X = (X_q\ X_t)$.

$\beta = \binom{\beta_q}{\beta_t}$. rewrite the model in:

$$Y = X_q\beta_q + X_t\beta_t + \varepsilon$$

Choose $q-1$ variables X_q from $p-1$ variables X then generate a alternative model:

$Y = X_q\beta_q + \varepsilon$. correspond : H_0 : $(0\ I_t)\beta = 0$ $(\Lambda = (0, I_t))$

$$R = \{ \frac{(\Lambda^T\hat\beta)^T(\Lambda^T(X^TX)^{-1}\Lambda)^{-1}\Lambda^T\hat\beta / r(\Lambda)}{MSE} = \frac{\hat\beta_t^T D^{-1}\hat\beta_t}{t\,MSE} \geq F_\alpha(t, n-p)\}$$

where $D = (X_t^T(I - P_{X_q})X_t)^{-1}$. $P_{X_q} = X_q(X_q^TX_q)^{-1}X_q^T$

By: $\begin{pmatrix} I_q & 0 \\ -X_t^TX_q(X_q^TX_q)^{-1} & I_t \end{pmatrix} X^TX \begin{pmatrix} I_q & -(X_q^TX_q)^{-1}X_q^TX_t \\ 0 & I_t \end{pmatrix}$

$= \begin{pmatrix} X_q^TX_q & 0 \\ 0 & D^{-1} \end{pmatrix}$. Denote $A = (X_q^TX_q)^{-1}X_q^TX_t$

$\Rightarrow (X^TX)^{-1} = \begin{pmatrix} (X_q^TX_q)^{-1} + ADA^T & -AD \\ -DA^T & D \end{pmatrix}$

<u>Denote</u>: i) $\hat{\beta} = (X^TX)^{-1}X^TY$. $\hat{\sigma}^2 = \dfrac{Y^T(I-M)Y}{n - r(X)}$. $\hat{y} = \vec{x}^T\hat{\beta} \in \mathbb{R}^1$

ii) $\tilde{\beta}_q = (X_q^TX_q)^{-1}X_q^TY$. $\tilde{\sigma}_q^2 = \dfrac{Y^T(I-M_q)Y}{n - r(X_q)}$ $\tilde{y} = \vec{x}_q^T\tilde{\beta}_q \in \mathbb{R}^1$

where $\vec{x} = (\vec{x}_q \;\; \vec{x}_t)$. y is estimate

iii) $MSE(\hat{\theta}) = E((\hat{\theta} - \theta)(\hat{\theta} - \theta)^T)$

<u>Thm.</u> (Influence on estimation)

i) $E(\hat{\beta}) = \beta$. if the full model is correct.

$\beta_t = 0$ or $X_q^TX_t = 0 \Longleftrightarrow E(\tilde{\beta}_q) = \beta_q$

ii) $Var(\hat{\beta})_q - Var(\tilde{\beta}_q) \geq 0$

iii) $Var(\hat{\beta})_t - \beta_t\beta_t^T \geq 0 \Rightarrow$

$Var(\hat{\beta})_q - E((\tilde{\beta}_q - \beta_q)(\tilde{\beta}_q - \beta_q)^T) \geq 0$

iv) $E(\tilde{\sigma}_q^2) \geq E(\hat{\sigma}^2) = \sigma^2$. "=" holds if and only if $\beta_t = 0$.

Rmk: i) means $\tilde{\beta}_q$ isn't a unbiased estimator generally

ii) means if we used the reduced model but the origin model is true. then variance of estimator will reduce.

iii) means the variables β_t discarded exactly effects the estimation. Since β_t can't be estimated and have large variance ($Var(\hat{\beta})_t \geq \beta_t \beta_t^T$). Delete β_t can reduce variance.

iv) means $\tilde{\sigma}_q^2$ isn't unbiased estimator for σ^2 if the origin model is true.

Pf: i) $E(\hat{\beta}) = (X^TX)^{-1} X^T X \beta = \beta$.

$$E(\tilde{\beta}_q) = (X_q^T X_q)^{-1} X_q^T X\beta = (X_q^T X_q)^{-1} X_q^T (X_q \ X_t) \begin{pmatrix} \beta_q \\ \beta_t \end{pmatrix}$$

$$= \beta_q + (X_q^T X_q)^{-1} X_q^T X_t \beta_t = \beta_q + A\beta_t$$

ii) $$Var(\hat{\beta}) = (X^TX)^{-1} X^T Var(Y) X (X^TX)^{-1}$$

$$= \sigma^2 (X^TX)^{-1} = \sigma^2 \begin{pmatrix} (X_q^TX_q)^{-1} + ADA^T & -AD \\ -DA^T & D \end{pmatrix}$$

$$\therefore Var(\hat{\beta})_q = \sigma^2 ((X_q^T X_q)^{-1} + ADA^T)$$

with $Var(\tilde{\beta}_q) = \sigma^2 (X_q^T X_q)^{-1}$

$$\Rightarrow Var(\hat{\beta})_q - Var(\tilde{\beta}_q) = \sigma^2 ADA^T \geq 0$$

$$MSE(\tilde{\beta}_q) = Var(\tilde{\beta}_q) + (E(\tilde{\beta}_q) - \beta_q)(E(\tilde{\beta}_q) - \beta_q)^T$$

$$= Var(\tilde{\beta}_q) + A\beta_t\beta_t^T A^T$$

$$\therefore Var(\hat{\beta})_q - MSE(\tilde{\beta}_q) = A(\sigma^2 D - \beta_t\beta_t^T)A^T$$

iv) $$E(\tilde{\sigma}_q^2) = \frac{1}{n-r(X_q)} E(Y^T(I-M_q)Y)$$

$$= \frac{1}{n-r(X_q)} tr((I-M_q)E(YY^T))$$

$$= \frac{1}{n-r(X_q)} tr((I-M_q)(\sigma^2 I + X\beta\beta^T X^T))$$

$$= \sigma^2 + \frac{1}{n-r(X_q)} \beta^T X^T (I-M_q) X\beta$$

$$= \sigma^2 + \frac{1}{n-r(X_q)} \beta_t^T X_t^T (I-M_q) X_t\beta_t$$

Thm. (Influence on estimate)

i) $E(\tilde{\eta}) = x^T\beta$ if the origin model is true

$\beta_t = 0 \iff E(\tilde{\eta}) = x^T\beta$

ii) $Var(\eta - x^T\hat{\beta}) \geq Var(\eta - x_q^T\tilde{\beta}_q)$

iii) $Var(\hat{\beta})_t - \beta_t\beta_t^T \geq 0 \Rightarrow Var(\eta - x^T\hat{\beta}) \geq E(\eta - x_q^T\tilde{\beta}_q)^2$

Rmk: It means if origin model is true. then $\tilde{\beta}_q$ isn't unbiased. But its MSE will reduce if $Var(\hat{\beta})_t \geq \beta_t\beta_t^T$. i.e. its variance is large.

Pf: i) $E(\hat{\eta}) = x^T E(\hat{\beta}) = x^T \beta$.

$E(\tilde{\eta}_q) = x_q^T (\beta_q + A\beta_t) = x_q^T \beta_q + x_q^T A \beta_t$

ii) $Var(\eta - x^T \hat{\beta}) = Var(\eta) + x^T Var(\hat{\beta}) x$

$= \sigma^2 (1 + x^T (X^T X)^{-1} x)$

$Var(\eta - x_q^T \tilde{\beta}_q) = Var(\eta) + x_q^T Var(\tilde{\beta}_q) x_q$

$= \sigma^2 (1 + x_q^T (X_q^T X_q)^{-1} x_q)$

$\Rightarrow Var(\eta - x^T \hat{\beta}) - Var(\eta - x_q^T \tilde{\beta}_q) =$

$\sigma^2 (A^T x_q - x_t)^T D (A^T x_q - x_t) \geq 0$

iii) $E(\eta - x_q^T \tilde{\beta}_q)^2 = Var(\eta - x_q^T \tilde{\beta}_q) + E^2(\eta - x_q^T \tilde{\beta}_q)$

$= Var(\eta - x_q^T \tilde{\beta}_q) + (x^T \beta - x_q^T (\beta_q + A\beta_t))^2$

$= Var(\eta - x_q^T \tilde{\beta}_q) + (A^T x_q - x_t)^T \beta_t \beta_t^T (A^T x_q - x_t)$

$\Rightarrow Var(\eta - x^T \hat{\beta}) - E(\eta - x_q^T \tilde{\beta}_q)^2 =$

$\sigma^2 (A^T x_q - x_t)^T (\sigma^2 D - \beta_t \beta_t^T)(A^T x_q - x_t) \geq 0$

with $Var(\hat{\beta})_t - \beta_t \beta_t^T = \sigma^2 D - \beta_t \beta_t^T$.

Rmk: Back to the hypothesis testing: Replace para. in $\sigma^2 D - \beta_t \beta_t^T \geq 0$ with estimator, then,

$\hat{\sigma}^2 D - \hat{\beta}_t \hat{\beta}_t^T \geq 0 \iff \hat{\beta}_t^T D^{-1} \hat{\beta}_t / t\hat{\sigma}^2 \leq \frac{1}{t}$

i.e. if $\frac{1}{t} < F_\alpha(t, n-p)$, then we accept

$H_0: \beta_t = 0$. (set $\alpha = D^{-\frac{1}{2}} \hat{\beta}_t$. prove: $\hat{\sigma}^2 I \geq \alpha\alpha^T$

$\iff \hat{\sigma}^2 \geq \alpha^T \alpha$. it's easy to see)

(2) Principles:

If we have p variables to be selected.

Then there're $2^p - 1$ possible regression equation

Denote: i) $SST = Y^T Y$, $SSE = Y^T (I - M) Y$

$SSE_q = Y^T (I - M_q) Y$.

ii) $R^2 = 1 - \frac{SSE}{SST}$. $R_q^2 = 1 - \frac{SSE_q}{SST}$

Rmk: $\forall q \leq p$. $SSE \leq SSE_q$. $R^2 \geq R_q^2$

① For fitting the model:

i) Minimize the mean of SSE_q:

i.e. find q. $\hat{\sigma}_q^2 = \min_r \frac{SSE_r}{n-r}$

Rmk: $SSE_q \uparrow$ as $q \downarrow$. $n-q \uparrow$ as $q \downarrow$

if $\{x_i\}_i^q$ effects y significantly

then $SSE_q \downarrow$ fast as $q \uparrow$.

if not. then $SSE_q \downarrow$ slowly as $q \uparrow$

$\frac{1}{n-q} \uparrow$ as $q \uparrow$. is penalty of increase

of number of variables.

ii) Maximize $\bar{R}_q^2 =: 1 - \frac{SSE_q / n-q}{SST / n-1} = 1 - (1 - R_q^2) \frac{n-1}{n-q}$

Rmk: $\beta_t = 0 \Rightarrow \bar{R}_q^2 \geq \bar{R}_{q+t}^2$

We call $\bar{R}_q^2$ adjustment complex decision coefficient.

② **For Prediction:**

Denote: $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{1q}^T & x_{1t}^T \\ \vdots & \vdots \\ x_{nq}^T & x_{nt}^T \end{pmatrix}$, $x_i = \begin{pmatrix} x_{iq} \\ x_{it} \end{pmatrix}$

$$\Rightarrow X^T X = \begin{pmatrix} X_q^T X_q & X_q^T X_t \\ X_q^T X_t & X_t^T X_t \end{pmatrix}$$

$$X_k^T X_j = \sum_{i=1}^n x_{ik} x_{ij}^T, \quad k, j \in \{q, t\}.$$

i) Minimize the estimator of JJ_t:

$$JJ_t = \sum_{i=1}^n Var(Y_i - x_{iq}^T \tilde{\beta}_q) = \sum_{i=1}^n (1 + x_{iq}^T (X_q^T X_q)^{-1} x_{iq}) \sigma^2$$

Note that $RHS = n\sigma^2 + tr((X_q^T X_q)^{-1} \sum_{i=1}^n (x_{iq} x_{iq}^T) \sigma^2)$

$$= (n+q)\sigma^2$$

Replace σ^2 by its estimator $\tilde{\sigma}_q^2$.

$$\Rightarrow \text{minimize} \quad \widehat{JJ}_t = (n+q)\tilde{\sigma}_q^2 = \frac{n+q}{n-q} SSE_q$$

ii) Minimize $S_q = \tilde{\sigma}_q^2 / (n-q-1)$

It's from: Consider $\eta_j = \beta_0 + \sum_i^{q-1} \beta_i x_{ji} + \varepsilon_j = \tilde{\beta}_0 + \sum \beta_i (x_{ji} - \bar{x}_j) + \varepsilon_j$

i.e. $Y = P_n X\beta + (I - P_n) X\beta + \varepsilon$. (centralization), $P_n = J_n J_n^T / n$

$\hat{Y} = P_n Y + (I - P_n) X\hat{\beta}$, $\hat{\beta} = (X^T (I - P_n) X)^{-} X^T (I - P_n) Y$

$\Rightarrow$ estimate $\hat{\eta} = \bar{\eta} + (x - \bar{x})^T \hat{\beta}$. $Var(\eta - \hat{\eta}) = (\frac{n+1}{n} + (x - \bar{x})^T S^{-1} (x - \bar{x})) \sigma^2$

where $S = X^T (I - P_n) X = \sum (X_k - \bar{x})(X_k - \bar{x})^T$. x from X. $\begin{pmatrix} Y \\ X \end{pmatrix} \sim N(m, \Sigma)$

Note that $\frac{n-q-1}{q-1} \frac{n}{n+1} (x - \bar{x})^T S^{-1} (x - \bar{x}) \sim F(q-1, n-q-1)$

$\Rightarrow E(Var(\eta - \hat{\eta} \mid X, x)) = \frac{n+1}{n} \frac{n-2}{n-q-1} (\Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY})$. Set $S_q = \frac{\tilde{\sigma}_q^2}{n-q-1}$.

iii) C_p statistics:

Consider to minimize $J_p = \frac{1}{\sigma^2}\sum_i^n (\hat{\eta}_i - E(Y_i))^2$

$= \frac{1}{\sigma^2}\sum (x_{iq}^T \hat{\beta}_q - x_i^T \beta)^2$. Take it expection:

$$E(J_p) = \frac{1}{\sigma^2}\sum_{i=1}^n Var(x_{iq}^T\hat{\beta}_q) + (E(x_{iq}^T\hat{\beta}_q) - x_i^T\beta)^2$$

$$= \sum x_{iq}^T (X_q^T X_q)^{-1} x_{iq} + \frac{1}{\sigma^2}\sum (x_{iq}^T(\beta_q + A\beta_t) - x_i^T\beta)^2$$

$$= q + \frac{1}{\sigma^2}\sum (x_{iq}^T A\beta_t - x_{it}^T\beta_t)^2$$

$$= q + \frac{1}{\sigma^2}\beta_t^T\left(\sum (A^T x_{iq} - x_{it})(A^T x_{iq} - x_{it})^T\right)\beta_t$$

$$= q + \frac{1}{\sigma^2}\beta_t^T D^{-1}\beta_t$$

$$= q + \frac{1}{\sigma^2}(n-q)(E(\hat{\sigma}_q^2) - \sigma^2)$$

(since $E(\hat{\sigma}_q^2) = \sigma^2 + \frac{1}{n-q}\beta_t^T X_t^T (I - M_q) X_t \beta_t$)

$$\Rightarrow E(J_p) = \frac{E(SSE_q)}{\sigma^2} + 2q - n$$

Choose $C_p = 2q - n + \frac{SSE_q}{\hat{\sigma}^2}$.

Rmk: $E(C_p) = q - t + \frac{n-p}{n-p-2}(t + \beta_t^T D^{-1}\beta_t/\sigma^2)$

if $n-p$ is large. When $\beta_t = 0$

then $E(C_p) \simeq q$.

$\Rightarrow$ Choose q. s.t. C_p is small and

$|C_p - q|$ is small.

iv) Minimize PRESS:

Denote: $Y^{(-i)} = \begin{pmatrix} y_1 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{pmatrix}$, $X^{(-i)} = \begin{pmatrix} x_1^T \\ \vdots \\ x_{i-1}^T \\ x_{i+1}^T \\ \vdots \\ x_n^T \end{pmatrix}$ (remove i^{th} component)

Consider $Y^{(-i)} = X^{(-i)}\beta + \varepsilon$

$\hat{\beta}^{(-i)} = (X^{(-i)T}X^{(-i)})^{-1}X^{(-i)T}Y^{(-i)}$, $\hat{e}^{(-i)} = y_i - x_i^T\hat{\beta}^{(-i)}$

<u>Def</u>: $PRESS = \sum_{i=1}^n (\hat{e}^{(-i)})^2$, prediction of sum of square of error.

<u>Denote</u>: $h_{ii} = x_i^T(X^TX)^{-1}x_i$ (i^{th} diagonal element of $X(X^TX)^{-1}X^T$)

To calculate PRESS:

$$\hat{e}^{(-i)} = y_i - x_i^T(X^{(-i)T}X^{(-i)})^{-1}(X^{(-i)T}Y^{(-i)})$$

$$= y_i - x_i^T(X^TX - x_ix_i^T)^{-1}(X^TY - x_iy_i)$$

$$(X^TX - x_ix_i^T)^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1-h_{ii}} \quad \text{(easy to check)}$$

$$\therefore \hat{e}^{(-i)} = \hat{e}_i + h_{ii}y_i - \frac{h_{ii}x_i^T\hat{\beta}}{1-h_{ii}} + \frac{h_{ii}y_i}{1-h_{ii}} \quad (\hat{e}_i = y_i - x_i^T\hat{\beta})$$

$$= \frac{\hat{e}_i}{1-h_{ii}}.$$

<u>Rmk</u>: For $PRESS_q = \sum_{i=1}^n (\hat{e}_q^{(-i)})^2$. We have:

$\hat{e}_q^{(-i)} = \dfrac{\hat{e}_{iq}}{1-h_{iiq}}$. h_{iiq} is i^{th} diagonal element of $H_q = X_q(X_q^TX_q)^{-1}X_q$

③ <u>By MLE</u>:

i) AIC: (Akaike information criteria)

If $\varepsilon \sim N(0, \sigma^2 I_n)$. Then:

$$\ln L(\beta, \sigma^2 | Y) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)$$

$$\ln L_{max} = \ln L((X^TX)^{-1}X^TY, \frac{SSE}{n})$$

$$= -\frac{n}{2}\ln(\frac{2\pi}{n}) - \frac{n}{2}\ln(SSE) - \frac{n}{2}$$

Def: $AIC = -2\log L(\hat{\theta}|X) + 2p$. $p = \dim\theta$.

$\Rightarrow AIC_q = n\ln SSE_q + 2q$ in this case

Find q to minimize AIC_q

ii) BIC (Bayesian information criterion)

$$BIC_q = n\ln SSE_q + 2q\ln n$$

(3) Selection:

By above, we have three common criteria:

i) adjrsp criteria: $\max_q \bar{R}_q^2$

ii) C_p criteria: $\min_q C_q = \min_q(\frac{SSE_q}{SSE_p/(n-p)} + 2q - n)$

iii) AIC: $\min_q n\ln SSE_q + 2q$

① Global Selection:

Check $2^p - 1$ possible regression models for optimal q. But it's low efficient when p is large.

e.g. choose $A(k) = \{x_{i_1} \cdots x_{i_k}\} \subset \{x_j\}_1^n$.

Calculate criterion on $A(k)$.

② Stepwise method:

i) Test for Significant:

· Consider $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$. i.e. test whether $\{x_i\}_1^p$ influences y a lot.

Denote: $SST = Y^T(I - P_n^n)Y$. $df_T = n-1$.

$SSR = Y^T(M - P_n^n)Y$. $df_R = p$

$SSE = Y^T(I - M)Y$. $df_E = n-p-1$

$\Rightarrow SST = SSR + SSE$. SSE indept with SSR

$$F = \frac{SSR/p}{SSE/(n-p-1)} \sim F(p, n-p-1). \text{ under } H_0.$$

We obtain p-value: $p = P(F(p, n-p-1) > F)$

· Consider $H_0: \beta_{i_0} = 0$. i.e. test the influence of individual variable.

Denote $RSS^{(-i_0)}$ is RSS on $\{x_k\}_{k \neq i_0}$.

$P_{i_0} = RSS - RSS^{(-i_0)} = \hat{\beta}_{i_0}^2 / l_{ii}$. (†) (prove it later)

where l_{ii} is the i^{th} diagonal element of $L^{-1} = (X^T(I - P_n^n)X)^{-1}$.

$$F_{i_0} = \frac{P_{i_0}}{SSE/(n-p-1)} \sim F(1, n-p-1) \text{ under } H_0$$

Rmk: $SST = SSE + SSR = SSE^{(-i)} + SSR^{(-i)} \Rightarrow P_{i_0} = Q(-i_0) - Q$

ii) Forward Selection:

Establish p regressional equations with one variable. Then calculate each p-value ($F \sim F_{(1, n-2)}$)

$\Rightarrow$ Choose the variable with smallest p-value and introduce it into equation. (Denote x_1)

$\Rightarrow$ Establish p-1 regression equations w.r.t $\{(x_1, x_i)\}_{i=2}^{p}$. Choose the pair with smallest p-value.

$\Rightarrow$ Repeat the process until we obtain the target number of variables.

Rmk: The later introduction of variables may reduce the significant of former variables.

iii) Backward Selection:

Put all the variables into the equations

$\Rightarrow$ Calculate F_i. discard the max one.

$\Rightarrow$ Put p-1 variables into equations. then repeat the process before.

Rmk: It needs a lot computations.

iv) Stepwise Selection:

Step one: discard variable. suppose we have introduced $\{x_{ik}\}_{k=1}^{r}$.

Calculate $P_{ik} = SSR\{ij\}_{j=1}^{r} - SSR\{ij\}_{j\neq k}$. if

x_0 is the variable corresponds $\min\limits_{1\le k\le r} P_{ik}$.

then test: $F_0 = \dfrac{P_0}{SSE/(n-r-1)}$. $P = P\{F(1, n-r-1) > F_0 \mid \beta_0 = 0\}$

if $P \geq \alpha_{out}$. $\Rightarrow$ discard x_0

$P < \alpha_{out}$ $\Rightarrow$ consider to introduce other variables

Step two: introduce new variables. suppose we have

$\{x_{jk}\}_{k=1}^{p-r}$ out of the equation.

$P_{jk} = SSR\{ij\}_i^r \cup \{jk\} - SSR\{ij\}_i^r$. if

x_{j0} is the variable correspond $\max\limits_{1\le k\le p-r} P_{jk}$

$$F_{j0} = \frac{P_{j0}}{SSE\{ij\}_i^r \cup \{jk\}/(n-r-2)} \overset{H_0}{\sim} F(1, n-r-2)$$

$P = P(F(1, n-r-2) > F_{j0} \mid H_0)$. $H_0: \beta_{j0} = 0$.

if $P < \alpha_{out}$ $\Rightarrow$ introduce x_{j0} into equation

$P \geq \alpha_{out}$ $\Rightarrow$ Selection is over.

Pf of (*):
$$\begin{cases} Q = Y^T(I - X(X^TX)^{-1}X^T)Y \\ Q(i) = Y^T(I - X_i(X_i^TX_i)^{-1}X_i^T)Y \end{cases} \quad X = (X_i \;\; x_i)$$

$$(X^TX)^{-1} = \begin{pmatrix} (X_i^TX_i)^{-1} + AD^{-1}A^T & -AD^{-1} \\ -D^{-1}A^T & D^{-1} \end{pmatrix}$$

$A = (X_i^TX_i)^{-1}X_i^Tx_i$. $D = x_i^T(I - P_{(X_i)})x_i = 1/l_{ii}$

$\Rightarrow P_i = Y^T(A^TX_i^T - x_i^T)^T D^{-1}(A^TX_i^T - x_i^T)Y$

$\hat{\beta}_i = (0 \;\; I)(X^TX)^{-1}X^TY = (-A^TX_i^T + x_i^T)Y/D$

$\Rightarrow P_i$ indept with Q. as well.