Regression Analysis

(1) Linear Model:

A regression model is linear model: $Y = X\beta + \varepsilon$.

where $X_{n \times p}$ is usually full rank p. $C(X)$ contains categorical variables and at least one contine variables. Thus. $X^T X$ is nonsingular.

① Simple linear Regression:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \quad 1 \le i \le n. \quad E(\varepsilon_i) = 0. \quad Var(\varepsilon_i) = \sigma^2.$$

$$\Rightarrow X = \begin{pmatrix} J_n & \begin{matrix} x_1 \\ \vdots \\ x_n \end{matrix} \end{pmatrix} \Rightarrow \hat{\beta} = \frac{1}{\sum_1^n (X_i - \bar{X})^2} \begin{pmatrix} n\bar{Y} \sum_1^n X_i^2 - n\bar{X} \sum_1^n X_i Y_i \\ -n^2 \bar{X} \bar{Y} + n \sum_1^n X_i Y_i \end{pmatrix}$$

More formly. $\hat{\beta}_1 = \dfrac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_1^n (X_i - \bar{X})^2} = \dfrac{\widehat{Cov}(X, Y)}{\widehat{Var}(X)}$

and $\hat{\beta}_0 = \bar{Y} - \bar{X} \hat{\beta}_1$

which follows from $M = (m_{ij})$. $m_{ij} = \dfrac{1}{n} + \dfrac{(X_i - \bar{X})(X_j - \bar{X})}{\sum_1^n (X_i - \bar{X})^2}$

② Multiple Linear Regression:

$$Y_i = \beta_0 + \beta_1 X_{i1} \cdots + \beta_p X_{ip} + \varepsilon_i. \quad 1 \le i \le n. \quad \begin{cases} E(\varepsilon_i) = 0 \\ \sigma^2(\varepsilon_i) = \sigma^2 \end{cases} \text{ i.i.d.}$$

More generally. consider $Y = X\beta + \varepsilon$. $X \in M^{n \times p}$.

Note that $C(M) = C(X)$ has rank p.

Decompose into sum of p orthogonal subspace of $\dim = 1$.

$$C(M) = \sum_{i=1}^p C(M_i). \quad r(M_i) = 1. \quad M_i M_j = 0. \; i \neq j$$

<u>Def</u>: $Z_i = (X_1, X_2 \cdots X_i)$ $1 \le i \le p$. $Z_0 = 0$.

Set $P_{Z_i} = Z_i (Z_i^T Z_i)^{-1} Z_i^T$. Then $M_i = P_{Z_i} - P_{Z_{i-1}}$. $1 \le i \le p$.

It easy to check $\{M_i\}_1^p$ is what we want.

<u>Denote</u>: $SSR(X_j | X_1 \cdots X_{j-1}) = Y^T M_j Y$.

<u>Rmk</u>: It can be intepreted as regression sum of squares due to add X_i when $X_1 \cdots X_{i-1}$ are already in model.

$$\Rightarrow SSR(X) = \sum_i^p SSR(X_i | X_1 \cdots X_{i-1})$$

<u>Rmk</u>: The orthogonal Decomposition of $C(X)$. depends on the order of variables X_i being fit into the model. So the decomposition may not be unique (c.f. Unbalance ANCOVA)

(2) <u>Best Linear Prediction</u>:

One of the main goal and use of regression model is for prediction but not only to make inference on parameters.

That is, predict Y on basis of information of $X_1, X_2 \cdots X_p$. Find $f(X)$ to minimize:

$E(Y - f(X))^2$ which need joint dist of (X, Y).

i) dist of (X, Y) is known:

Thm. Set $m(X) = m(Y|X)$. Then $m(X) = \arg\min E(Y - f(X))^2$

Pf: Check $E((Y - m(X))(m(X) - f(X))) = 0$

prop. $\mathrm{cov}(Y, f(X)) = \mathrm{cov}(m(X), f(X))$

Pf: Check: $E((Y - m(X)) f(X)) = 0$

Thm. (Criteria of Best Predictor)

$\tilde{\eta}(X)$ is any predictor. Then $\tilde{\eta}(X) = m(X)$. a.s.

$\Leftrightarrow$ $\mathrm{cov}(f(X), Y - \tilde{\eta}(X)) = 0$ for any function f.

and $E(\tilde{\eta}(X)) = \mu_Y$.

Pf: ($\Leftarrow$) It's from prop. above

($\Rightarrow$) Prove: $\sigma^2(\tilde{\eta}(X) - m(X)) = 0$. Since $E(\tilde{\eta}(X) - m(X)) = 0$.

$LHS = \mathrm{cov}(\eta - m(X), \tilde{\eta}(X) - m(X)) - \mathrm{cov}(\eta - \tilde{\eta}(X), \tilde{\eta}(X) - m(X))$

$= ① + ② = 0$. $① = 0$ by prop. $② = 0$ by condition

ii) dist of (X, Y) is unknown:

We only can find best linear predictor of Y if we only know means. vars of X, Y and $\mathrm{cov}(X, Y)$.

i.e. minimize $E(Y - f(X))^2$. $f(X) = \alpha + X^T\beta$.

Rmk: If (X, Y) is multinomal. Then the best prediction is linear predictor:

$$E(Y|X) = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$$

Thm. β_* is solution of $\Sigma_{XX}\beta = \Sigma_{XY}$. Then: $\hat{E}(Y|X) =:$

$\mu_Y + (X-\mu_X)^T\beta_*$ is best linear predictor

Pf: Show: $E((Y-\hat{E}(Y|X))(\hat{E}(Y|X) - f(X))) = 0$

where $f(X) = \eta + (X-\mu_X)\beta^T$.

$\Rightarrow E(Y-\alpha-X^T\beta)^2 = E(Y-\hat{E}(Y|X))^2 + E(\hat{E}(Y|X)-\square)^2$

Rmk: BLP is unique. i.e. to minimize:

$E(\hat{E}(Y|X)-\eta-(X-\mu_X)^T\beta)^2$. $\eta = \mu_X$ and β must

be solution of $\Sigma_{XX}\beta = \Sigma_{XY}$.

Note it's: $(\mu_Y-\eta)^2 + E((X-\mu_X)^T(\beta_*-\beta))^2$

So minimization require: $\eta = \mu_Y$. $E(\square)^2 = 0$

Since $E(\square)^2 = (\beta_*-\beta)^T\Sigma_{XX}(\beta_*-\beta)$. $\Leftrightarrow$

$\Sigma_{XX}^{\frac{1}{2}}(\beta_*-\beta) = 0 \Leftrightarrow \Sigma_{XX}(\beta_*-\beta) = 0$.

iii) Apply to multiple linear regression:

Denote: $S_{XX} = \dfrac{X^T(I-P_n)X}{n-1}$, $S_{XY} = \dfrac{X^T(I-P_n)Y}{n-1}$

$\hat{\mu}_X = \frac{1}{n}J_n^T X$. $\hat{\mu}_Y = \frac{1}{n}J_n^T Y$.

$\Rightarrow$ Best linear prediction of Y_i: $\hat{Y}_i = \hat{\mu}_Y + (X_i - \hat{\mu}_X)^T\hat{\beta}_*$

where $\hat{\beta}_*$ is solution of $S_{XX}\hat{\beta} = S_{XY}$

Rmk: $\hat{\beta}_*$ is LSE of $Y_i = \alpha + (X_i - \hat{\mu}_X)^T\beta + \varepsilon_i$. $1 \le i \le p$.

$E(\varepsilon_i) = 0$. $Var(\varepsilon_i) = \sigma^2$. i.e. $Y = (J_n\ (I-P_n)X)\binom{\alpha}{\beta} + \varepsilon$.

We can obtain LSE of α: $\hat{\alpha} = P_nY = \bar{Y}$.

<u>prop.</u> (Criteria of BLP)

$\tilde{\eta}(X)$ is any linear predictor. Then $\tilde{\eta}(X) = \hat{E}(Y|X)$ a.s

$\Leftrightarrow$ $Cov(f(X), Y - \tilde{\eta}(X)) = 0$. for any linear function f

and $E(\tilde{\eta}(X)) = \mu_Y$.

<u>Pf:</u> Set $f(X) = \eta + (X - \mu_X)^T \beta$.

($\Rightarrow$) $Cov(f(X), Y - \tilde{\eta}(X)) = \beta^T \Sigma_{XY} - \beta^T \Sigma_{XX} \beta_X = 0$

($\Leftarrow$) Write $\tilde{\eta}(X) = \mu_Y + (X - \mu_X)^T \delta$ for some δ.

$\therefore Cov(f(X), Y - \tilde{\eta}(X)) = \beta^T \Sigma_{XY} - \beta^T \Sigma_{XX} \delta = 0$

iv) <u>For vector $Y = (\eta_1 \cdots \eta_p)$:</u>

If we have data $X = (X_1 \cdots X_q)$

<u>Thm.</u> $f(X) = (E(Y_1|\vec{X}), \cdots E(Y_p|\vec{X}))$ minimizes:

$E((Y - f(X))^T (Y - f(X)))$

<u>Thm.</u> $f_i(X) = \mu_{Y_i} + (X - \mu_X)^T \beta_i^*$. $\Sigma_{XY_i} \beta_i^* = \Sigma_{XX}$. Then

$\vec{f}(X)$ is best linear predictor of $\vec{Y}$.

(3) <u>Coefficient of Determination:</u>

① <u>Multiple Correlation Coefficient:</u>

Consider: $Y = J_n \beta_0 + X_{n \times p} \beta + \varepsilon$. $r(X) = p$.

Written in $Y = J_n \delta_0 + (X - \frac{J_n^n}{n}) X \beta + \varepsilon$

where $\delta_0 = \beta_0 + \frac{J_n^T}{n} X \beta$.

<u>Def.</u> $R^2 = \frac{SSR}{SSTOT - C}$ is Coefficient of Determination

where $SSTOT = Y^T Y$. $C = n\bar{Y}^2$. $SSR = Y^T(M^* - \frac{J_n^n}{n})Y$

M^* is $P_{C(J_n\ X)}$ orthonormal proj.

<u>Rmk:</u> i) $SYY = SSTOT - C = Y^T(I - P_n)Y = \sum_i^n (Y_i - \bar{Y})^2$

ii) $SSR = SYY - SSE$.

iii) $R^2 = 1 - \frac{SSE}{SYY}$. $0 \le R^2 \le 1$. R^2 is interpreted as the proportion of total variability in Y explained by indept variable. $(X_1 \cdots X_p)$

It can measure the predictive ability of a model. $R^2 \uparrow$. The better fit of model.

To motivate use of R^2 to acess fit:

R^2 is actually estimate of the square of multiple correlation coefficient.

<u>Def.</u> The multiple correlation coefficient between $Y_{1\times 1}$ and $\vec{X} = (X_1 \cdots X_p)$ is $\max_{(\alpha,\beta)} \{ Corr^2(Y, \alpha + X^T\beta) \}$.

<u>Rmk:</u> It's $\max_\beta \frac{(\Sigma_{YX}\beta)^2}{\beta^T \Sigma_{XX}\beta \cdot \sigma_{YY}} = \max_\beta \frac{(\Sigma_{YX}\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XX}^{\frac{1}{2}}\beta)^2}{\sigma_{YY}\cdot\beta^T\Sigma_{XX}\beta}$

$\le \frac{\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}}{\sigma_{YY}}$. So $MCC(Y, \vec{X}) = \frac{\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}}{\sigma_{YY}}$

Its estimation is $\frac{S_{Yx} S_{xx}^{-1} S_{xY}}{\hat{\sigma}_{YY}}$. $\hat{\sigma}_{YY}^2 = Y^T(I-P_n)Y = SYY$.

$S_{Yx}^2 = Y^T(I-P_n)X$. $S_{xx}^2 = X^T(I-P_n)X$. $S_{xY}^2 = S_{Yx}^T$.

So $R^2 = \widehat{mcc}(Y, \vec{X})$.

Apply in Test:

Note that $\frac{SSR}{SSE} = \frac{R^2}{1-R^2}$. Test: $H_0: \vec{\beta} = 0$

Then $F = \frac{Y^T(M^* - \frac{1}{n}J_n^n)Y/p}{Y^T(I-M^*)Y/(n-p-1)} = \frac{n-p-1}{p}\frac{R^2}{1-R^2} \overset{H_0}{\sim} F(p, n-p-1)$

③ Partial Correlation Coefficients:

i) We're also interested in the correlation between 2 variables condition on a set of variables that are already in model.

Denote: $\rho_{y \cdot x}$ is partial correlation coefficient of Y_1, Y_2 given $X_1 \cdots X_{p-1}$.

Rmk: $\rho_{y \cdot x}$ is measure of linear relation between Y_1, Y_2 after taking the effect of $\vec{X}$ out of Y_1, Y_2.

Note that BLP of $Y = (Y_1, Y_2)^T$ given X is:

$\hat{E}(Y|X) = \mu_Y + \beta_*^T(X-\mu_X)$. $\Sigma_{xx}\beta_* = \Sigma_{xY} \in M^{(p-1)\times 2}$.

$\Rightarrow$ Take effects of X out of Y by looking at:

the prediction error: $Y - (\mu_Y + \beta_*^T (X - \mu_X))$

Def: $\rho_{y\cdot x}$ is correlation of two component of

$Y - (\mu_Y + \beta_*^T (X - \mu_X))$.

Rmk: $Cov(Y - \mu_Y - \beta_*^T (X - \mu_X)) =$

$$Cov(Y - \mu_Y) + \beta_*^T Cov(X - \mu_X)\beta_* - \square - \square$$

$$= \Sigma_{YY} + \beta_*^T \Sigma_{XX}\beta_* - \Sigma_{YX}\beta_* - \beta_*^T \Sigma_{XY}$$

$$= \Sigma_{YY} - \beta_*^T \Sigma_{XX}\beta_* \qquad (\text{Note } \Sigma_{XX} = \Sigma_{XX}\Sigma_{XX}^- \Sigma_{XX})$$

$$= \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^- \Sigma_{XY} \Rightarrow \text{Calculate } \rho_{y\cdot x}.$$

ii) In sample Case:

If we have sample: $Y = (Y_1\ Y_2) = \begin{pmatrix} Y_{11} & Y_{12} \\ \vdots & \vdots \\ Y_{n1} & Y_{n2} \end{pmatrix}$

$X = (X_{ij})_{n\times p}$. The estimate of $\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^- \Sigma_{XY}$ is

$$S_{YY} - S_{YX} S_{XX}^{-1} S_{XY} = \frac{1}{n-1} Y^T (I - M^*) Y$$

So $$\hat{\rho}_{y\cdot x} = r_{y|x} = \frac{Y_2^T (I - M^*) Y_1}{(Y_1^T (I-M^*) Y_1)^{\frac{1}{2}} (Y_2^T (I-M^*) Y_2)^{\frac{1}{2}}}$$

Def: $r_{y|x}$ is sample partial correlation coefficient

iii) Hypothesis Test:

Consider fitting: $Y_1 = J_n \gamma_0 + X\gamma_1 + Y_2\gamma_2 + \varepsilon$.

where $\varepsilon \sim N(0, \sigma^2 I)$

Since $C((I-M^*)Y_2)$ is $C^{\perp}(J_n, X) \cap C(J_n, X, Y_1)$

Then: $SSR(Y_2 \mid J_n, X) = Y_1^T P_{C((I-M_*)Y_2)} Y_1$

$$= Y_1^T (I-M^*) Y_2 (Y_2^T (I-M^*) Y_2)^{-1} Y_2^T (I-M^*) Y_1$$

With $SSE(J_n, X) = Y_1^T (I-M^*) Y_1$

$\Rightarrow r_{y|x}^2 = SSR(Y_2 \mid J_n, X) / SSE(J_n, X)$

To test: $H_0: \rho_{y \cdot x} = 0 \iff H_0: \gamma_2 = 0 \quad (\iff Cov(Y_1, Y_2 \mid X) = 0)$

$$F = \frac{Y_1^T (P_{C(J_n, X, Y_2)} - P_{C(J_n, X)}) Y_1 / 1}{Y_1^T (I - P_{C(J_n, X, Y_2)}) Y_1 / (n-p-2)}$$

$$= \frac{Y_1^T P_{C((I-P_{C(J_n,X)})Y_2)} Y_1}{Y_1^T (I - P_{C(J_n,X)} - P_{C((I-P_{C(J_n,X)})Y_2)}) Y_1 / (n-p-2)}$$

$$= \frac{r_{y|x}^2}{(1-r_{y|x}^2)/(n-p-2)} \overset{H_0}{\sim} F(1, n-p-2)$$

④ Squared Predictive Correlation:

i) S.P.C of predictor $\tilde{\eta}(X)$ is $Corr(Y, \tilde{\eta}(X))$

The higher SPC. The more predictive $\tilde{\eta}(x)$ is.

Thm. $Corr^2(Y, \tilde{\eta}(X)) \leq Corr^2(Y, m(X)) =: R^2$

Pf: $\sigma_{Y\tilde{\eta}}^2 = \sigma_{m\tilde{\eta}}^2 \leq \sigma_{mm} \sigma_{\tilde{\eta}\tilde{\eta}}$. By Schwarz.

And $R^2 = \frac{\sigma_{Ym}^2}{\sigma_{YY}\sigma_{mm}} = \frac{\sigma_{mm}^2}{\sigma_{YY}\sigma_{mm}} = \sigma_{mm}/\sigma_{YY}$.

Rmk: In fact, high SPC can also be attained by bad predictors. Since $\tilde{\eta}(X)$ is high correlated with Y doesn't mean $\tilde{\eta}(X)$ is close enough to Y.

ii) Linearized Predictor:

If we have a predictor $\tilde{\eta}(X)$. We can construct a linear predictor which is at least as good as $\tilde{\eta}(X)$: $\hat{\eta}(X) = \mu_Y + \frac{\sigma_{Y\tilde{\eta}}}{\sigma_{\tilde{\eta}\tilde{\eta}}}(\tilde{\eta}(X) - \mu_{\tilde{\eta}})$. i.e.

the BLP based on $\tilde{\eta}(X)$ rather than X.

Rmk: Note $\sigma_{\hat{\eta}\hat{\eta}} = \sigma_{Y\tilde{\eta}}^2 / \sigma_{\tilde{\eta}\tilde{\eta}} = \sigma_{Y\hat{\eta}}$. We obtain:

$$Corr^2(Y, \hat{\eta}(X)) = \sigma_{\hat{\eta}\hat{\eta}} / \sigma_{YY}$$

Def: We measure the goodness of such pred. $\hat{\eta}(X)$

by: $E(Y - \hat{\eta}(X))^2 = \sigma_{YY} - \sigma_{\hat{\eta}\hat{\eta}}$

prop. For two linearized predictor $\hat{\eta}_1(X)$, $\hat{\eta}_2(X)$.

$\hat{\eta}_2(X)$ is better $\Leftrightarrow$ SPC of $\hat{\eta}_2(X)$ is higher

Pf: $Corr^2(Y, \hat{\eta}_1(X)) \leq Corr^2(Y, \hat{\eta}_2(X)) \Leftrightarrow$

$\sigma_{\hat{\eta}_1\hat{\eta}_1} / \sigma_{YY} \leq \sigma_{\hat{\eta}_2\hat{\eta}_2} / \sigma_{YY} \Leftrightarrow$

$\sigma_{YY} - \sigma_{\hat{\eta}_2\hat{\eta}_2} \leq \sigma_{YY} - \sigma_{\hat{\eta}_1\hat{\eta}_1}$. by Def.