LINEAR REGRESSION

Friday, April 5, 2019 10:13 AM

INTRODUCTION

REGRESSION MODEL IS A STATISTICAL METHOD FOR PREDICTING VALUES of ONE OR MORE RESPONE VARIABLES FROM A SET of PREDICTOR VARIABLES ON THE RESPONES.

MATHEMATICAL FORMULATION

WITH r PREDITOR VAR.S
n RESPONE VAR.S

CLASSICAL LINEAR REG. MODEL

$$\underline{y} = Z\underline{\beta} + \underline{\varepsilon}$$

IN WHICH, $\mathbb{E}(\underline{\varepsilon}) = \underline{0}$ AND $\mathrm{COV}(\underline{\varepsilon}) = \sigma^2 I$.
OR MORE EXPLICITLY, WE CAN WRITE THE MODEL AS:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1r} \\ 1 & z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

SIZE $n \times 1$ $\quad$ $n \times (r+1)$ $\quad$ $(r+1) \times 1$ $\quad$ $n \times 1$

REMARK AT THIS STAGE WE DONT ASSUME THAT $\underline{\varepsilon}$ IS DISTRIBUTED AS NORMAL; WE ONLY NEED IT TO BE ZERO-MEAN AND COMPONENT-WISLY UNCORRELATED.

METHODS FOR DERIVING ESTIMATION

WE WANT TO PREDICT THE RESPONE FOR GIVEN VALUES of PREDICTOR VARIABLES, THEN IT HAS TO "FIT" THE MODEL DESCRIBLED AS ABOVE. WE MAINLY HAVE TWO IDEAS TO REACH OUR GOAL:

1) LEAST SQUARES ESTIMATION

IDEA: MINIMIZE THE DISTANCE BETWEEN THE OBSERVED RESPONES AND PREDICTIONS OF RESPONES.

OPTIMIZATION MODEL:

OPTIMIZATION MODEL:

$$(L.E.S.)\quad \hat{\beta} = \arg\min_{\beta \in \mathbb{R}^{r+1}} (y - Z\beta)'(y - Z\beta)$$

RESULT 1 THE SOLUTION OF (L.E.S.) IS $\hat{\beta} = (Z'Z)^{-1}Z'y$

PROOF JUST BY BASIC COMPUTATION:

$$\begin{aligned} f(\beta) &:= (y - Z\beta)'(y - Z\beta) \\ &= (y' - \beta'Z')(y - Z\beta) \\ &= y'y - y'Z\beta - \beta'Z'y + \beta'Z'Z\beta \\ &= y'y - 2\beta'Z'y + \beta'Z'Z\beta \end{aligned}$$

$y'Z\beta \in \mathbb{R}$, $\beta'Z'y \in \mathbb{R}$, AND $(y'Z\beta)' = \beta'Z'y$

MOREOVER, WE STATE THAT:

$$① \quad \frac{\partial \beta'Z'y}{\partial \beta} = Z'y \quad \text{AND} \quad \frac{\partial \beta'Z'Z\beta}{\partial \beta} = 2Z'Z\beta \quad ②$$

TILL NOW THE DERIVATIVE WITH RESPECT TO A VECTOR IS NOT RIGOROUSLY DEFINED, THUS WE HAVE TO SHOW THE DETAIL

SUB-PROOF OF ①

$$\beta'Z'y = (\beta_1, \dots, \beta_{r+1}) \begin{pmatrix} (Z'y)_1 \\ \vdots \\ (Z'y)_{r+1} \end{pmatrix}$$

$$= \sum_{i=1}^{r+1} \beta_i (Z'y)_i$$

$$\Rightarrow \frac{\partial \beta'Z'y}{\partial \beta_i} = (Z'y)_i$$

$\Rightarrow$ WE HAVE THE COMPATIBLE NOTATION:

$$\frac{\partial \beta'Z'y}{\partial \beta} := Z'y$$

SUB-PROOF OF ② OBVIOUSLY $Z'Z$ IS A SYMMETRIC MATRIX, WE SHALL DENOTE $A := Z'Z$.

$$\Rightarrow \beta'A\beta = \sum^{r+1} \beta \cdot A_{ij} \beta$$

WE SHALL DENOTE $A := Z'Z$.

$$\Rightarrow \beta' A \beta = \sum_{i,j=1}^{r+1} \beta_i A_{ij} \beta_j$$

$$\Rightarrow \frac{\partial \beta' A \beta}{\partial \beta_i} = 2 \sum_{j=1}^{r+1} A_{ij} \beta_j = 2 (A\beta)_i$$

$\Rightarrow$ IN THIS CASE WE ALSO HAVE THE COMPATIBLE NOTATION:

$$\frac{\partial \beta' A \beta}{\partial \beta} := 2 Z'Z \beta$$

RETURN BACK TO $f(\beta) = y'y - 2\beta' Z'y + \beta' Z'Z \beta$, AND LET

$$\frac{\partial f}{\partial \beta} = -2Z'y + 2Z'Z\beta = 0$$

$$\Rightarrow \beta = (Z'Z)^{-1} Z'y =: \hat{\beta}$$

BESIDES PROVING $\hat{\beta}$ IS A STATIONARY POINT OF f, WE HAVE ALSO TO STUDY THE SECOND ORDER DERIVATIVE:

$$\frac{\partial^2 f}{\partial \beta^2} = 2Z'Z$$

RECALL A VERY IMPORTANT LEMMA WITHOUT PROOF:

WITH THE CLASSICAL DEF: POS. + DEFINITIVE = SYM. + $x'Ax > 0$, $\forall x$.

LEMMA 2. A REAL SYMMETRIC MATRIX A IS POSITIVE DEFINITIVE IF AND ONLY IF THERE EXISTS A NON-SINGULAR MATRIX Z SUCH THAT

$$A = Z'Z.$$

THANKS TO THIS LEMMA; AND WITH THE THEORY IN BASIC OPTIMIZATION, WE KNOW: FOR OBTAINING THE MINIMUM AT $\hat{\beta}$, IT REQUIRES $Z'Z$ TO BE POSITIVE DEFINITIVE, WHICH IS EQUIVALENT TO REQUIRING Z TO BE FULL-RANK. ■

2) MAXIMUM LIKELIHOOD ESTIMATION

IDEA: CHOOSE THE PARAMETER WHICH MAXIMIZE THE LIKE-

IDEA: CHOOSE THE PARAMETER WHICH MAXIMIZE THE LIKELIHOOD AS THE ESTIMATION OF TRUE PARAMETER.

REMARK FOR STUDYING THE LIKELIHOOD, WE MUST INTRODUCE THE PROBABILISTIC DISTRIBUTIONS INTO THE MODEL. THEREFORE, FROM NOW ON WE ASSUME

$$\underline{\varepsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I)$$ ← $\mathcal{N}$ - multi-variate Gaussian distribution

THE MODEL IS AS $\underline{y} = Z\underline{\beta} + \underline{\varepsilon}$, WHICH IMPLIES $\underline{y} \sim \mathcal{N}(Z\underline{\beta}, \sigma^2 I)$, FROM WHICH WE CAN CONSTRUCT THE MAXIMUM LIKELIHOOD FUNCTION

$$\mathcal{L}(\underline{y} \mid \underline{\beta}) = \frac{1}{(2\pi)^{n/2} |\sigma^2 I|^{1/2}} e^{\frac{-(\underline{y}-Z\underline{\beta})'(\sigma^2 I)^{-1}(\underline{y}-Z\underline{\beta})}{2}}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot e^{\frac{-(\underline{y}-Z\underline{\beta})'(\underline{y}-Z\underline{\beta})}{2\sigma^2}}$$

THEN LET $\hat{\underline{\beta}} = \underset{\underline{\beta} \in \mathbb{R}^{r+1}}{\arg\max}\ \mathcal{L}(\underline{y} \mid \underline{\beta})$, WHICH IS EXACTLY EQUIVALENT TO

$$\hat{\underline{\beta}} = \underset{\underline{\beta} \in \mathbb{R}^{r+1}}{\arg\min}\ (\underline{y}-Z\underline{\beta})'(\underline{y}-Z\underline{\beta})$$

THEREFORE WE WILL GET THE SAME RESULT AS IN THE LEAST SQUARES ESTIMATION.

PROPERTIES OF ESTIMATORS

LAST SECTION WE HAVE USED TWO METHODS, OR WE SHALL SAY, TWO DIFFERENT PHILOSOPHIES TO DERIVE THE ESTIMATOR

$$\hat{\underline{\beta}} = (Z'Z)^{-1} Z'\underline{y}$$

HERE WE STILL JUST ASSUME $\mathbb{E}[\underline{\varepsilon}] = \underline{0}$ AND $\mathrm{cov}(\underline{\varepsilon}) = \sigma^2 I$, RATHER THAN GIVE THE ERROR ANY SPECIFIC DISTRIBUTION.

MOREOVER, HERE WE INTRODUCE THE RESIDUAL $\hat{\underline{\varepsilon}}$

MOREOVER, HERE WE INTRODUCE THE RESIDUAL $\hat{\underline{\varepsilon}}$ AS THE ESTIMATOR OF ERROR $\underline{\varepsilon}$. IT IS DEFINED AS:

$$\hat{\underline{\varepsilon}} := \underline{y} - \hat{\underline{y}} = \underline{y} - Z(Z'Z^{-1})Z'\underline{y} = \left[I - Z(Z'Z^{-1})Z'\right]\underline{y}$$

THEN WE HAVE THE FOLLOWING RESULT:

RESULT 3 UNDER THE BASIC ASSUMPTION ON $\underline{\varepsilon}$, THE ESTIMATOR $\hat{\underline{\beta}} = (Z'Z)^{-1}Z'\underline{y}$ HAS

$$E[\hat{\underline{\beta}}] = \underline{\beta}$$ ← UNBIASED ESTIMATOR.

AND

$$\mathrm{COV}(\hat{\underline{\beta}}) = \sigma^2(Z'Z)^{-1}$$

THE RESIDUAL $\hat{\underline{\varepsilon}}$ HAS

$$E[\hat{\underline{\varepsilon}}] = 0$$

AND

$$\mathrm{COV}(\hat{\underline{\varepsilon}}) = \sigma^2\left[I - Z(Z'Z)^{-1}Z'\right]$$

MOREOVER, $\hat{\underline{\beta}}$ AND $\hat{\underline{\varepsilon}}$ ARE UNCORRELATED.

Proof

$$\begin{aligned} E[\hat{\underline{\beta}}] &= E[(Z'Z)^{-1}Z'\underline{y}] \\ &= E[(Z'Z)^{-1}Z'(Z\underline{\beta} + \underline{\varepsilon})] \\ &= E[(Z'Z)^{-1}Z'Z\cdot\underline{\beta}] + E[(Z'Z)^{-1}Z'\underline{\varepsilon}] \\ &= \underline{\beta} \end{aligned}$$

↑ IN OUR SETTING, ALL THE THINGS ARE DETERMINISTIC, EXCEPT $\underline{\varepsilon}$.

$$\begin{aligned} \mathrm{Cov}(\hat{\underline{\beta}}) &= E[(\hat{\underline{\beta}} - E[\hat{\underline{\beta}}])(\hat{\underline{\beta}} - E[\hat{\underline{\beta}}])'] \\ &= E[(\hat{\underline{\beta}} - \underline{\beta})(\hat{\underline{\beta}} - \underline{\beta})'] \\ &= E[\hat{\underline{\beta}}\hat{\underline{\beta}}' - \hat{\underline{\beta}}\underline{\beta}' - \underline{\beta}\hat{\underline{\beta}}' + \underline{\beta}\underline{\beta}'] \\ &= E[(Z'Z)^{-1}Z'\underline{y}\cdot\underline{y}'Z(Z'Z)^{-1}] + \underline{\beta}\underline{\beta}' \\ &\quad - E[(Z'Z)^{-1}Z'\underline{y}\cdot\underline{\beta}' + \underline{\beta}\cdot\underline{y}'Z(Z'Z)^{-1}] \end{aligned}$$

RECALL THAT $\underline{y} = Z\underline{\beta} + \underline{\varepsilon}$

$$\begin{aligned} &= E[(Z'Z)^{-1}Z'(Z\underline{\beta}+\underline{\varepsilon})(Z\underline{\beta}+\underline{\varepsilon})'Z(Z'Z)^{-1}] + \underline{\beta}\underline{\beta}' \\ &\quad - E[(Z'Z)^{-1}Z'(Z\underline{\beta}+\underline{\varepsilon})\cdot\underline{\beta}' + \underline{\beta}\cdot(Z\underline{\beta}+\underline{\varepsilon})'Z(Z'Z)^{-1}] \\ &= E[(Z'Z)^{-1}Z'(Z\underline{\beta}\cdot\underline{\beta}'Z' + Z\underline{\beta}\cdot\underline{\varepsilon}' + \underline{\varepsilon}\cdot\underline{\beta}'Z' + \underline{\varepsilon}\underline{\varepsilon}')Z(Z'Z)^{-1}] + \underline{\beta}\underline{\beta}' \end{aligned}$$

$$= \mathbb{E}\left[(Z'Z)^{-1}Z'(Z\beta\cdot\beta'Z' + Z\beta\cdot\varepsilon' + \varepsilon\cdot\beta'Z' + \varepsilon\varepsilon')Z(Z'Z)^{-1} + \beta\beta'\right.$$
$$\left. - \mathbb{E}\left[(Z'Z)^{-1}Z'Z\beta\beta' + \beta\cdot\beta'Z'Z(Z'Z)^{-1}\right]\right.$$
$$= \mathbb{E}\left[(Z'Z)^{-1}Z'Z\beta\cdot\beta'Z'Z(Z'Z)^{-1}\right] + \mathbb{E}\left[(Z'Z)^{-1}Z'\varepsilon\varepsilon'Z(Z'Z)^{-1}\right]$$
$$+ \beta\beta' - \beta\beta' - \beta\beta'$$
$$= (Z'Z)^{-1}Z'\cdot\mathbb{E}[\varepsilon\varepsilon']\cdot Z(Z'Z)^{-1}$$
$$= (Z'Z)^{-1}Z'\cdot\sigma^2 I\cdot Z(Z'Z)^{-1}$$
$$= \sigma^2(Z'Z)^{-1}$$

LUCKILY, ε HAS ZERO MEAN, THIS EXACTLY IS ε'S COVARIANCE MATRIX.

NOW WE CAN EASILY OBTAIN THAT

$$\mathbb{E}[\hat{\varepsilon}] = \mathbb{E}[y - Z\hat{\beta}]$$
$$= \mathbb{E}[y - Z\cdot(Z'Z)^{-1}Z'y]$$
$$= Z\beta - Z\cdot(Z'Z)^{-1}Z'\cdot Z\beta = 0$$

ONE CAN ALSO SHOW THAT $\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2[I - Z(Z'Z)^{-1}Z']$ AND $\mathrm{Cov}(\hat{\beta}, \hat{\varepsilon}) = 0$ BY THE SIMILAR COMPUTATION, BUT WE OMIT THE DETAIL HERE. ∎

GEOMETRY OF LEAST SQUARES

A GEOMETRICAL INTERPRETATION OF LEAST SQUARES TECHNIQUE HIGHLIGHTS THE NATURE OF THE CONCEPT. FIRST WE SHALL REFRESH US WITH THE NOTATION OF PROJECTION IN THE HILBERT SPACE. IN FINITELY DIMENSIONAL CASE, THE PROJECTION OPERATOR CAN BE REPRESENTED AS CANONICAL MATRIX. WE GIVE THE FOLLOWING DEFINITION:

DEF 4 A REAL-VALUED MATRIX P IS CALLED ORTHOGONAL PROJECTION MATRIX IF IT IS:

1) SYMMETRIC,
2) $P^2 = P$.

IN OUR CASE, BOTH RESPONSE VARIABLE y AND PREDICTED RESPONSE VARIABLE $\hat{y}$ BELONG TO FINITE DIMENSIONAL HILBERT SPACE $\mathbb{R}^n$. THE LINK BETWEEN THEM IS THE LINEAR

RESPONSE VARIABLE y BELONG TO FINITE DIMENSIONAL HILBERT SPACE $\mathbb{R}^n$, THE LINK BETWEEN THEM IS THE LINEAR OPERATOR (OR WE CALL IT MATRIX):

$$\hat{y} = Z\hat{\beta} = Z\cdot(Z'Z)^{-1}Z'y$$

← ALWAYS ASSUME THAT Z IS OF FULL RANK

WE DENOTE $P := Z(Z'Z)^{-1}Z'$, AND OBVIOUSLY:

$$P: \mathbb{R}^n \to \mathbb{R}^n, \quad P: y \longmapsto Z(Z'Z)^{-1}Z'y$$

THEN, WE CAN CHECK THAT

$$P' = [Z(Z'Z)^{-1}Z']' = Z(Z'Z)^{-1}Z' = P$$

$$P^2 = P\cdot P = Z(Z'Z)^{-1}Z'\cdot Z(Z'Z)^{-1}Z' = Z(Z'Z)^{-1}Z' = P$$

ACCORDING TO DEFINITION 4, WE KNOW THAT $P = Z(Z'Z)^{-1}Z$ IS A PROJECTION OPERATOR (MATRIX) ON $\mathbb{R}^n$. SINCE WE HAS ASSUMED THAT Z IS OF FULL RANK, RECALL LEMMA 2, WE KNOW THAT $Z'Z$ MUST BE POSITIVE DEFINITE. THEN WE CAN USE THE POWERFUL TOOL: SPECTRAL THEOREM, FOR STUDY THE PROPERTY OF OUR PROJECTION MATRIX P.

THEOREM 5 [REAL SPECTRAL THEOREM IN MATRIX PRESENTATION] SUPPOSE P IS AN $n\times n$ MATRIX, THEN $\mathbb{R}^n$ HAS AN ORTHONORMAL BASIS CONSISTING OF EIGENVECTOR OF P IF AND ONLY IF P IS SYMMETRIC.

↑ FOR THE PROOF PLEASE CHECK THE BOOK OF AXLER - LINEAR ALGEBRA DONE RIGHT.

THEN, WE CAN WRITE

$$Z'Z = \sum_{i=1}^{r+1} \lambda_i e_i e_i'$$

WHERE $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{r+1} > 0$ ARE THE EIGENVALUES OF $Z'Z$ AND $e_1, e_2, \ldots, e_{r+1}$ ARE THE CORRESPONDING EIGENVECTORS. FURTHER,

$$(Z'Z)^{-1} = \sum_{i=1}^{r+1} \frac{1}{\lambda_i} e_i e_i'$$

CONSIDER $q_i = \lambda_i^{-1/2} Z e_i$, WHICH IS A LINEAR COMBINATION OF THE COLUMNS OF Z. THEN WE HAVE

CONSIDER $q_i = \lambda_i^{-1/2} Z e_i$, WHICH IS A LINEAR COMBINATION OF THE COLUMNS OF Z. THEN WE HAVE

$$\begin{aligned} q_i' q_k &= \lambda_i^{-1/2} e_i' Z' \cdot \lambda_k^{-1/2} Z e_k \\ &= \lambda_i^{-1/2} \cdot \lambda_k^{-1/2} \cdot e_i' \left(\sum_{j=1}^{r+1} \lambda_j e_j e_j' \right) e_k \\ &= \begin{cases} 0 & \text{if } i \neq k \\ 1 & \text{if } i = k \end{cases} \end{aligned}$$

THAT IS, THE $r+1$ VECTORS q_i ARE MUTUALLY PERPENDICULAR AND HAVE THE UNIT LENGTH. THEIR LINEAR COMBINATION SPAN THE SPACE OF ALL LINEAR COMBINATIONS OF THE COLUMNS OF Z. MOREOVER,

$$Z(Z'Z)^{-1}Z' = \sum_{i=1}^{r+1} \lambda_i^{-1} Z e_i e_i' Z' = \sum_{i=1}^{r+1} q_i q_i'$$

THE PROJECTION OF y ON A LINEAR COMBINATION OF $[q_1, \ldots, q_{r+1}]$ IS

$$\sum_{i=1}^{r+1} (q_i' y) q_i = \left(\sum_{i=1}^{r+1} q_i q_i' \right) y = Z(Z'Z)^{-1} Z' y = Z\hat{\beta}$$

THUS, THE MULTIPLICATION BY $Z(Z'Z)^{-1}Z$ PROJECTS A VECTOR ONTO THE SPACE SPANNED BY THE COLUMNS OF Z.

EVALUATION: SUM-OF-SQUARES DECOMPOSITION

FIRST WE CAN OBSERVE THAT

$$\begin{aligned} Z'\hat{\varepsilon} &= Z'(y - \hat{y}) \\ &= Z'[I - Z(Z'Z)^{-1}Z']y \qquad (1) \\ &= [Z' - Z'Z(Z'Z)^{-1}Z']y = 0 \end{aligned}$$

AND $\hat{y}'\hat{\varepsilon} = \hat{\beta}' Z'\hat{\varepsilon} = 0$.

THEN WE CAN DERICTLY GET THAT

$$\begin{aligned} y'y &= (\hat{y} + y - \hat{y})'(\hat{y} + y - \hat{y}) \\ &= (\hat{y} + \hat{\varepsilon})'(\hat{y} + \hat{\varepsilon}) \\ &= \hat{y}'\hat{y} + \hat{y}'\hat{\varepsilon} + \hat{\varepsilon}'\hat{y} + \hat{\varepsilon}'\hat{\varepsilon} \qquad (2) \\ &= \hat{y}'\hat{y} + \hat{\varepsilon}'\hat{\varepsilon} \end{aligned}$$

$$= \underline{y}'\underline{y} + \underline{y}'\underline{z} + \underline{\varepsilon}'\underline{y} + \underline{z}'\underline{z}$$
$$= \hat{\underline{y}}'\hat{\underline{y}} + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$$

FURTHERMORE, SINCE THE FIRST COLUMN OF Z IS $\underline{1}$, THEN ① TELLS US ALSO $\underline{1}'\hat{\underline{\varepsilon}} = 0$, THEN

$$0 = \underline{1}'\hat{\underline{\varepsilon}} = \sum_{j=1}^{n} \hat{\varepsilon}_n = \sum_{j=1}^{n} (y_j - \hat{y}_j) = \sum_{j=1}^{n} y_j - \sum_{j=1}^{n} \hat{y}_j \quad ③$$

$$\Leftrightarrow \bar{y} = \bar{\hat{y}}, \text{ IF WE DEFINE } \bar{y} := \frac{1}{n}\sum_{j=1}^{n} y_j \text{ AND } \bar{\hat{y}} := \frac{1}{n}\sum_{j=1}^{n} \hat{y}_j$$

COMBINE ② AND ③ WE CAN GET

$$\underline{y}'\underline{y} - n\bar{y}^2 = \hat{\underline{y}}'\hat{\underline{y}} - n(\bar{\hat{y}})^2 + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$$

$$\Leftrightarrow \sum_{j=1}^{n} (y_j - \bar{y})^2 = \sum_{j=1}^{n} (\hat{y}_j - \bar{\hat{y}})^2 + \sum_{j=1}^{n} \hat{\varepsilon}_j^2$$

$\Rightarrow$ TOTAL SUM OF SQUARES ABOUT MEAN = REGRESSION SUM OF SQUARES + RESIDUAL SUM OF SQUARES

THE PRECEDING SUM OF SQUARES DECOMPOSITION SUGGESTS THAT THE QUALITY OF THE MODELS FIT CAN BE MEASURED BY COEFFICIENT OF DETERMINATION:

$$R^2 = 1 - \frac{\sum_{j=1}^{n} \hat{\varepsilon}_j^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2} = \frac{\sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2}$$

ONE CAN OBSERVE THAT $R^2 = 1$ IF THE FITTED EQUATION PASSES THROUGH ALL THE DATA POINTS; $R^2 = 0$ IF $\hat{\beta}_0 = \bar{y}$ AND $\hat{\beta}_1 = \cdots = \hat{\beta}_r = 0$, THAT IS, THE PREDICTOR VARIBLES IN FACT HAS NO INFLUENCE ON THE RESPONE VARIABLES.

INFERENCES ABOUT REGRESSION MODEL

FROM NOW ON WE FORMLY INTRODUCE A SPECIFIC GAUSSIAN DISTRIBUTION TO THE ERROR $\underline{\varepsilon}$, AS

$$\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I)$$

← IN THE ROOT, THE "n" REPRESENTS THE DIMENSION

THEN WE HAVE THE FOLLOWING IMPORTANT RESULT:

THEN WE HAVE THE FOLLOWING IMPORTANT RESULT:

RESULT 6 LET $\underline{y} = Z\underline{\beta} + \underline{\varepsilon}$, WHERE Z HAS FULL RANK $r+1$ AND $\underline{\varepsilon}$ IS DISTRIBUTED AS $\underline{\varepsilon} \sim \mathcal{N}_n(\underline{0}, \sigma^2 I)$, THEN THE MAXIMAL LIKELIHOOD ESTIMATOR OF $\underline{\beta}$ IS AS THE LEAST SQUARES ESTIMATOR $\hat{\underline{\beta}}$. MOREOVER,

1) $\hat{\underline{\beta}} = (Z'Z)^{-1}Z'\underline{y} \sim \mathcal{N}_{r+1}(\underline{\beta}, \sigma^2(Z'Z)^{-1})$

2) $\hat{\underline{\varepsilon}} = \underline{y} - Z\hat{\underline{\beta}}$, $\hat{\underline{\varepsilon}} \perp \hat{\underline{\beta}}$. FURTHER $n\hat{\sigma}^2 = \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} \sim \sigma^2 \chi^2_{n-r-1}$ ①

WHERE $\hat{\sigma}^2$ IS THE MAXIMAL LIKELIHOOD ESTIMATOR OF σ^2.

PROOF WE HAVE SHOWN THAT THE ESTIMATOR DERIVED BY LEAST SQUARES ESTIMATION AND MAXIMAL LIKELIHOOD METHOD COINCIDE UNDER THE SPECIFIC ASSUMPTION ON $\underline{\varepsilon}$.

FOR THE POINT 1, IT IS A DIRECT CONCLUSION OF RESULT 3. FOR THE POINT 2, ALSO FROM RESULT 3 WE KNOW THAT $\mathrm{COV}(\hat{\underline{\varepsilon}}, \hat{\underline{\beta}}) = 0$, AND UNDER THE ASSUMPTION OF GAUSSIAN DISTRIBUTION, NOW THEY ARE ALSO INDEPENDENT. THEREFORE, WE ONLY HAVE TO DEMONSTRATE ①, THIS NEEDS SOME CALCULATIONS:

$P := Z(Z'Z)^{-1}Z'$

$$\hat{\underline{\varepsilon}} = \underline{y} - \hat{\underline{y}} = (I - P)\underline{y}, \text{ WHERE } \underline{y} \sim \mathcal{N}_n(Z\underline{\beta}, \sigma^2 I)$$

$$\Rightarrow \hat{\underline{\varepsilon}} \sim \mathcal{N}_n\big([I-P]Z\underline{\beta}, (I-P)'\sigma^2 I(I-P)\big).$$

NOTICE $[I-P]Z\underline{\beta} = (I - Z(Z'Z)^{-1}Z')Z\underline{\beta} = Z\underline{\beta} - Z\underline{\beta} = \underline{0}$,

$(I-P)'\cdot\sigma^2 I\cdot(I-P) = \sigma^2(I-P)$ — P IS PROJ. MATRIX, THEN I−P ALSO IS, ONE CAN USE DEFINITION TO CHECK. GEOMETRICALLY SPEAKING, IT'S A PROJ. OP. TO THE PERPENDICULAR SPACE.

$$\Rightarrow \hat{\underline{\varepsilon}} \sim \mathcal{N}_n(\underline{0}, \sigma^2(I-P)).$$

WE DO ALREADY KNOW THAT $I-P$ IS ALSO A PROJECTION MATRIX, NOW WE HAVE TO USE AN IMPORTANT PROPERTY OF EIGENVALUES OF A PROJECTION MATRIX:

SUPPOSE H IS AN (ORTHOGONAL) PROJECT MATRIX, WE SHALL STUDY ITS EIGENVALUES AND EIGENVECTORS AS:

SUPPOSE H IS AN (ORTHOGONAL) PROJECT MATRIX, WE SHALL STUDY ITS EIGENVALUES AND EIGENVECTORS AS:

$$H\underline{x} = \lambda \underline{x} \quad \text{FOR SOME } \underline{x} \neq 0 \text{ AND } \lambda \in \mathbb{R}$$

SINCE H IS IDEMPOTENT, THAT IS $H = H^2$, THEREFORE:

$$\lambda \underline{x} = H\underline{x} = H^2\underline{x} = H \cdot H\underline{x} = H \cdot \lambda \underline{x} = \lambda^2 \underline{x} \quad \text{FOR SOME } \underline{x} \neq 0.$$

$\Rightarrow$ λ MUST BE EITHER 0 OR 1.

ALSO THANKS TO REAL SPETRAL THEOREM, SINCE $I-P$ IS REAL AND SYMMETRIC, WE CAN FIND ORTHONORMAL BASIS OF $\mathbb{R}^n$ CONSISTING OF EIGENVECTORS of $I-P$. WE ALREADY KNOW THAT THE EIGENVALUES of $I-P$ ARE EITHER 0 OR 1, ACCORDING TO THIS NICE PROPERTY, WE CAN STUDY THE DIMENSION of CORRESPONDING EIGENSPACE V_0 AND V_1 ($V_0 \oplus V_1 = \mathbb{R}^n$) BY JUST COMPUTING THE TRACE of $I-P$:

$$\begin{aligned} Tr(I-P) &= Tr(I) - Tr(P) \\ &= Tr(I) - Tr\left(Z(Z'Z)^{-1}Z'\right) \\ &= Tr(I) - Tr\left((Z'Z)^{-1}Z'Z\right) \\ &= Tr(I) - Tr(\tilde{I}) \end{aligned}$$

CYCLIC PROPERTY of TRACE

NOTICE THAT I AND $\tilde{I}$ ARE BOTH IDENTITY MATRIX (OPERATOR), BUT $I: \mathbb{R}^n \to \mathbb{R}^n$; $\tilde{I}: \mathbb{R}^{r+1} \to \mathbb{R}^{r+1}$, THEREFORE

$$Tr(I-P) = n - (r+1) = n - r - 1$$

THIS TELLS US $\dim V_0 = r+1$, $\dim V_1 = n-r-1$, WHERE 0 AND 1 ARE CORRESPONDING EIGENVALUES OF MATRIX $I-P$. THEN AGAIN IN TERMS of REAL SPECTRAL THEOREM, WE HAVE THE FOLLOWING EQUALITY:

$$I - P = Q \Lambda Q'$$

WHERE:

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_n \end{bmatrix}, \quad Q = [e_1, e_2, \ldots, e_n]$$

$\lambda_1, \lambda_2, \ldots, \lambda_n$ ARE EIGENVALUES of MATRIX $I-P$, IN PARTI-

$$\Lambda = \begin{bmatrix} & & \\ & & \\ & & \lambda_n \end{bmatrix}$$

$\lambda_1, \lambda_2, \dots, \lambda_n$ ARE EIGENVALUES OF MATRIX $I-P$, IN PARTICULAR, $r+1$ OF THEM ARE ZERO, AND $n-r-1$ OF THEM ARE ONE; $\underline{e}_1, \underline{e}_2, \dots, \underline{e}_n$ ARE CORRESPONDING NORMED, MUTUALLY ORTHOGONAL EIGENVECTORS, SO WE ALSO KNOW $QQ' = Q'Q = I$.

NOW, WE CAN GO BACK TO STUDY $\hat{\underline{\varepsilon}}$, WE KNOW

$$\hat{\underline{\varepsilon}} \sim \mathcal{N}_n(\underline{0}, \sigma^2(I-P))$$

$$\Leftrightarrow \hat{\underline{\varepsilon}} \sim \mathcal{N}_n(\underline{0}, \sigma^2 Q \Lambda Q')$$

$$\Rightarrow Q'\hat{\underline{\varepsilon}} \sim \mathcal{N}_n(\underline{0}, \sigma^2 Q'Q \Lambda Q'Q)$$

$$\Leftrightarrow Q'\hat{\underline{\varepsilon}} \sim \mathcal{N}_n(\underline{0}, \sigma^2 \Lambda) \quad (*)$$

THEN WE OBSERVE THAT

$$\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} = \hat{\underline{\varepsilon}}' Q Q' \hat{\underline{\varepsilon}} = (Q'\hat{\underline{\varepsilon}})' \cdot Q'\hat{\underline{\varepsilon}}$$

WHERE $Q'\hat{\underline{\varepsilon}}$ IS A COLUMN VECTOR OF RANDOM VARIABLES WITH LENGTH n, AND OBSERVE $(*)$, WE KNOW: $r+1$ MANY COMPONENTS OF $Q'\hat{\underline{\varepsilon}}$ ARE OF ZERO MEAN, ZERO VARIANCE AND INDEPENDENT TO OTHER COMPONENTS (ZERO COVARIANCE IMPLIES THE UNCORRELATION, GAUSSIAN DISTRIBUTION GIVES THE INDEPENDENCE), THUS WE CAN REGARD THEM AS BEING "DEGENERATED" TO CONSTANT NUMBER 0, THIS GIVES US THE FINAL RESULT:

$$\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} = (Q'\hat{\underline{\varepsilon}})' \cdot Q'\hat{\underline{\varepsilon}} = \sum_{j=1}^{n-r-1} \left[Q'\hat{\underline{\varepsilon}}\right]_j^2 = \sigma^2 \sum_{j=1}^{n-r-1} \left(\frac{[Q'\hat{\underline{\varepsilon}}]_j}{\sigma}\right)^2$$

WHERE: $[Q'\hat{\underline{\varepsilon}}]_j \sim \mathcal{N}_1(0, \sigma^2) \Leftrightarrow \frac{1}{\sigma}[Q'\hat{\underline{\varepsilon}}]_j \sim (\mathcal{N}_1(0,1))$

$$\Rightarrow \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} \sim \sigma^2 \chi^2(n-r-1)$$ ∎

FROM RESULT 6 WE CAN IMMEDIATELY SEE $\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$ CAN BE WRAPPED TO BE A GOOD ESTIMATOR OF σ^2.

COROLLARY 7 DEFINE $\delta^2 := \frac{\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}}{n-r-1}$, THEN δ^2 IS AN UNBIASED

COROLLARY 7 DEFINE $s^2 := \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-r-1}$, THEN s^2 IS AN UNBIASED ESTIMATOR OF σ^2.

PROOF SINCE $\hat{\varepsilon}'\hat{\varepsilon} \sim \sigma^2 \chi^2(n-r-1)$, THEN DIRECTLY

$$\mathbb{E}[s^2] = \mathbb{E}\left[\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-r-1}\right] = \frac{\sigma^2}{n-r-1}\cdot(n-r-1) = \sigma^2 \quad \blacksquare$$

FROM RESULT 6, WE GET $\hat{\beta} \sim \mathcal{N}_{r+1}(\beta, \sigma^2(Z'Z)^{-1})$. BASE ON THIS WE COULD DO MANY STATISTIC INFERENCES, THE MOST IMPORTANT ONE IS GIVEN AS FOLLOWING:

RESULT 8 LET $Y = Z\beta + \varepsilon$, WHERE Z HAS FULL RANK $r+1$ AND $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. THEN A $100(1-\alpha)$ PERCENT CONFIDENCE REGION FOR β IS GIVEN BY

IGNORE THE TERMINOLOGY LIKE "SIMULTANEOUS CONFIDENCE INTERVALS" IF YOU DON'T KNOW IT

$$(\beta-\hat{\beta})'Z'Z(\beta-\hat{\beta}) \le (r+1)s^2 F_{r+1,n-r-1}(\alpha)$$

WHERE $F_{r+1,n-r-1}(\alpha)$ IS THE UPPER (100α)-TH PERCENTILE OF AN F-DISTRIBUTION WITH $r+1$ AND $n-r-1$ DEGREE OF FREEDOM.

ALSO, SIMULTANEOUS $100(1-\alpha)$ PERCENT CONFIDENCE INTERVALS FOR THE β_i ARE GIVEN BY

$$\hat{\beta}_i \pm \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}\cdot\sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}, \quad i = 0, 1, \ldots, r$$

WHERE $\widehat{\mathrm{Var}}(\hat{\beta}_i)$ IS THE DIAGONAL ELEMENT OF ESTIMATED COVARIANCE MATRIX $s^2(Z'Z)^{-1}$ CORRESPONDING TO $\hat{\beta}_i$.

PROOF PLEASE REFER TO BOOK "APPLIED MULTIVARIATE STATISTICAL ANALYSIS", WRITTEN BY R. JOHNSON ET. AL. $\blacksquare$

IN FACT, RESULT 8 PLAYS A CRUCIAL ROLE WHEN WE WANT TO KNOW IF THE COEFFICIENT OF PREDICTOR VARIABLE IS STATISTICALLY SIGNIFICANT. OR IN OTHER WORDS, IF THE COEFFICIENT OF A PREDICTOR VARIABLE IS "VERY NEAR" TO ZERO, WE MAY DELETE THIS PREDICTOR VARIABLE FROM THE MODEL, RECONSTRUCT THE DESIGN MATRIX Z. REDO THE REGRESSION

MAY DELETE THIS PREDICTOR VARIABLE FROM THE MODEL, RECONSTRUCT THE DESIGN MATRIX $\underset{\sim}{Z}$, REDO THE REGRESSION.

PRACTITIONERS OFTEN IGNORE THE "SIMULTANEOUS" CONFIDENCE PROPERTY OF INTEVAL ESTIMATES IN RESULT 8. INSTEAD, THEY REPLACE $(r+1)F_{m,n-r-1}(\alpha)$ WITH ONE-AT-A-TIME t VALUE $t_{n-r-1}(\alpha/2)$ AND USE THE INTERVALS

$$\hat{\beta} \pm t_{n-r-1}(\alpha/2)\sqrt{\widehat{Var}(\hat{\beta}_i)}$$

WHEN SEARCHING FOR IMPORTANT PREDICTOR VARIABLES.

MOREOVER, HYPOTHESIS TESTS METHOD COULD BE IMPLEMENTED FOR TESTING HYPOTHESIS LIKE $H_0: \beta_{n_1} = \beta_{n_2} = \cdots = \beta_{n_k} = 0$, WHERE $n_1, n_2, \ldots n_k \in \{0, 1, \ldots, r\}$. BUT WE DONT INTRODUCE THE DETAIL HERE.

INFERENCES FROM THE ESTIMATED REGRESSION FUNCTION

FROM PRECEDING SECTIONS WE GET OUR ESTIMATED REGRESSION FUNCTION. NOW, SUPPOSE THAT WE GET A NEW DATA

$$\underset{\sim}{z}_0' = [1, z_{01}, \ldots, z_{0r}]$$

AND WE WANT TO LEARN THE BEHAVIOR of RESPONCE of THE ESTIMATED REGRESSION FUNCTION, WE SHALL DENOTE IT AS Y_0. THEN, HERE ARE TWO SLIGHTLY DIFFERENT STRATEGIES:

1. ESTIMATE THE EXPECTED VALUE of Y_0.
2. PREDICT Y_0. (MORE UNCERTAIN)

ESTIMATION THE EXPECTED VALUE of Y_0

ACCORDING TO THE REGRESSION MODEL $\underset{\sim}{Y} = \underset{\sim}{Z}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$, IN WHICH $\underset{\sim}{\varepsilon} \sim N_n(\underset{\sim}{0}, \sigma^2 I)$, THE EXPECTED VALUE of Y_0 IS

$$E[Y_0 | \underset{\sim}{z}_0] = \beta_0 + \beta_1 z_{01} + \cdots + \beta_r z_{0r} = \underset{\sim}{z}_0' \underset{\sim}{\beta}$$

ITS LEAST SQUARE ESTIMATION IS $\underset{\sim}{z}_0' \hat{\underset{\sim}{\beta}}$.

ITS LEAST SQUARE ESTIMATION IS $\underline{z}_0'\hat{\beta}$.

RESULT 9 FOR THE CLASSICAL LINEAR REGRESSION MODEL WITH $\underline{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I)$, $\underline{z}_0'\hat{\beta}$ IS THE UNBIASED LINEAR ESTIMATOR OF $E[Y_0 | \underline{z}_0]$ WITH MINIMUM VARIANCE,

$$\text{Var}(\underline{z}_0'\hat{\beta}) = \underline{z}_0'(\underline{z}'\underline{z})^{-1}\underline{z}_0\sigma^2$$

MOREOVER, A $100(1-\alpha)\%$ CONFIDENCE INTERVAL FOR $E[Y_0 | \underline{z}_0]$ IS PROVIDED BY

$$\underline{z}_0'\hat{\beta} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{(\underline{z}_0'(\underline{z}'\underline{z})^{-1}\underline{z}_0)s^2}$$

WHERE $t_{n-r-1}(\alpha/2)$ IS THE UPPER $100(\alpha/2)$-TH PERCENTILE OF A t-DISTRIBUTION WITH $n-r-1$ DEGREE OF FREEDOM.

FORCASTING A NEW OBSERVATION AT $\underline{z}_0$

PREDICTION OF A NEW OBSERVATION, WE CALL IT Y_0, IS MORE UNCERTAIN THAN ESTIMATING THE EXPECTED VALUE OF Y_0. ACCORDING TO THE CLASSICAL REGRESSION MODEL,

$$Y_0 = \underline{z}_0'\beta + \varepsilon_0$$

WHERE $\varepsilon_0 \sim \mathcal{N}_1(0, \sigma^2)$ AND IS INDEPENDENT OF $\underline{\varepsilon}$, HENCE OF $\hat{\beta}$ AND σ^2. THE ERRORS $\underline{\varepsilon}$ INFLUENCE THE ESTIMATORS $\hat{\beta}$ AND s^2 THROUGH THE RESPONSES $\underline{Y}$, BUT ε_0 DOES NOT.

RESULT 10 GIVEN A CLASSICAL LINEAR REGRESSION MODEL, A NEW OBSERVATION Y_0 HAS THE UNBIASED PREDICTOR

$$\underline{z}_0'\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 z_{01} + \cdots + \hat{\beta}_r z_{0r}$$

THE VARIANCE OF THE FORECAST ERROR $Y_0 - \underline{z}_0'\hat{\beta}$ IS

$$\text{Var}(Y_0 - \underline{z}_0'\hat{\beta}) = \sigma^2(1 + \underline{z}_0'(\underline{z}'\underline{z})^{-1}\underline{z}_0)$$

WHEN THE ERRORS $\underline{\varepsilon}$ HAVE A NORMAL DISTRIBUTION, A $100(1-\alpha)\%$ PREDICTION INTEVAL FOR Y_0 IS GIVEN BY

$$\underline{z}_0'\hat{\beta} \pm t_{n-r-1}(\alpha/2)\sqrt{s^2(1 + \underline{z}_0'(\underline{z}'\underline{z})^{-1}\underline{z}_0)}$$

$$\underline{z}_0^T \hat{\beta} \pm t_{n-r-1}(\alpha/2) \sqrt{\delta^2 (1 + \underline{z}_0^T (Z^T Z)^{-1} \underline{z}_0)}$$

WHERE $t_{n-r-1}(\alpha/2)$ IS THE UPPER $100(\alpha/2)$-TH PERCENTILE OF A t-DISTRIBUTION WITH $n-r-1$ DEGREE OF FREEDOM.

WE CAN OBSERVE THAT: THE PREDICTION INTERVAL FOR y_0 IS WIDER THAN THE CONFIDENCE INTERVAL FOR ESTIMATING THE VALUE OF $E[Y_0 | \underline{z}_0]$. THE ADDITIONAL UNCERTAINTY IN FORECASTING y_0, WHICH IS REPRESENTED BY EXTRA TERM δ^2 IN THE EXPRESSION $\delta^2(1 + \underline{z}_0^T (Z^T Z)^{-1} \underline{z}_0)$, COMES FROM THE PRESENCE OF UNKNOWN ERROR ε_0.

COLLINEARITY

THE COLLINEARITY IS A PHENOMENON IN WHICH ONE PREDICTOR VARIABLE IN A LINEAR REGRESSION MODEL CAN BE LINEARLY PREDICTED FROM THE OTHERS WITH A LARGE DEGREE OF ACCURACY. OR MATHEMATICALLY SPEAKING, THE DESIGN MATRIX IS "NEARLY" NON-FULL-RANK, THEN THE CALCULATION OF $(Z^T Z)^{-1}$ IS NUMERICALLY UNSTABLE. HERE WE GIVE SOME REMEDIES:

1. DELETING. DELETE ONE OF A PAIR OF PREDICTOR VARIABLES THAT ARE STRONGLY CORRELATED.

2. PRINCIPAL COMPONENT ANALYSIS. RECALL THAT THIS METHOD PRODUCES NEW UNCORRELATED DATA.

3. RIDGE REGRESSION.