Regression

Predictor: Variable $X_1 \cdots X_d$ ($n \times 1$ feature vectors)

Response: RV Y ($n \times 1$)

- Conditional expectation of Y given values $x_1 \cdots x_d$ of var $X_1 \cdots X_d$ is called the regression function of Y on $X_1 \cdots X_d$, or simply the regression of Y on $X_1 \cdots X_d$. In symbols, $E(Y | x_1 \cdots x_d)$

Assumption of LR. (General Linear Model)

1. Predictor is known / for multiple reg. requires $\text{rank}(X) = d$)
2. Normality: conditional distr of Y_i is normal.
3. Linear Mean: $\exists$ parameter $\beta_{d \times 1}$, $E(Y | x_1 \cdots x_d) = \sum_i \beta_i x_i$
4. Common Variance (homoscedasticity): $\exists \sigma^2$, $\text{Var}(Y_i | x_1 \cdots x_d) = \sigma^2$, $i = 1 \cdots n$.
5. Independence: Given $x_1 \cdots x_d$, RV $Y_1 \cdots Y_n$ are independent.

Given the above assumptions, the conditioned joint distr of $Y_1 \cdots Y_n$ given vector $x_1 \cdots x_d$, parameters β, σ^2 is:

$$f(Y | X, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum (Y_i - \sum \beta_j x_j)^2\right)$$

Formulation of LR

- Simple LR

$$y = \beta_0 + \beta_1 X + \varepsilon$$

- Multiple LR

$$y = X_0 \beta_0 + \cdots + X_{d+1} \beta_{d+1} + \varepsilon \qquad X_i : \ n \times 1$$

Note, constant term is generalized with $X_0 = 1$.

Matrix notation: $y = X\beta + \varepsilon$, $X: n \times d$, $\beta: d \times 1$

Solution of LR

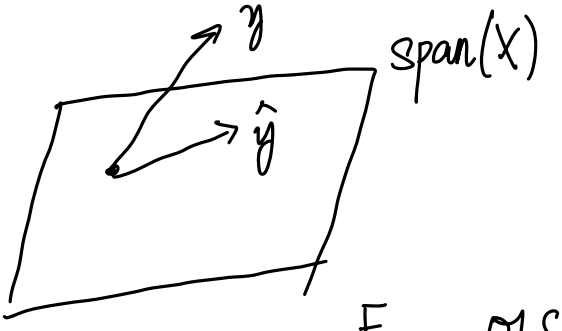
Under L2 loss: $L(\beta) = \|y - X\beta\|_2^2$

$\det(X^TX) \neq 0$

$$\hat{\beta} = \begin{cases} (X^TX)^{-1}X^Ty & \text{when } \det(X) \neq 0 \text{ (invertible)} \\ V\Sigma^{+}U^*y & \text{when } \det(X) = 0, \text{ SVD: } X = U\Sigma V^* \\ R^{+}Q^Ty & \text{when QR: } X = QR \end{cases}$$

Geometrical Perspective of OLS

$y = X\beta$, OLS $\hat{\beta} = X^+y$



Note: XX^+ is a linear orthorgonal operator that projects any vector onto span(X).

From OLS, $\hat{y} = X\hat{\beta} = XX^+y$, which is projection of y.

Gauss-Markov Theorem

Least square solution is of least variance among all linear unbiased estimators of parameter β.

Formal proof:

Let $\tilde{\beta} = Cy$ be an arbitrary unbiased linear estimator of β.

Choose D, s.t., $C = X^{+} + D$

$\mathbb{E}(\tilde{\beta}) = \beta$: unbiasedness

$$\mathbb{E}(\tilde{\beta}) = \mathbb{E}(Cy) = \mathbb{E}\left((X^{+}+D)(X\beta+\varepsilon)\right)$$
$$= \beta + DX\beta + (X^{+}+D)\underbrace{\mathbb{E}(\varepsilon)}_{=0}$$
$$= (I+DX)\beta$$

Thus, in order for $(I+DX)\beta = \beta$, $DX = 0$.

$$\mathrm{Var}(\tilde{\beta}) = \mathrm{Var}(Cy) = C\,\mathrm{Var}(y)\,C^{T} = \sigma^2 CC^{T}$$
$$= \sigma^2\left[(X^{+}+D)(X^{+}+D)^{T}\right] = \sigma^2\left(X^{+}X^{+T} + X^{+}D^{T} + DX^{+T} + DD^{T}\right)$$
$$= \sigma^2\left((X^{T}X)^{-1}X^{T}X(X^{T}X)^{-1T} + (X^{T}X)^{-1}X^{T}D^{T} + DX(X^{T}X)^{-1T} + DD^{T}\right)$$
$$(DX=0) \quad = \sigma^2\left((X^{T}X)^{-1} + DD^{T}\right) = \mathrm{Var}(\hat{\beta}) + \sigma^2 DD^{T}$$

Since DD^{T} p.s.d, $\mathrm{Var}(\tilde{\beta}) \geq \mathrm{Var}(\hat{\beta})$. ☒

Alt. explanation for naive case:

We will use bias-variance tradeoff for estimators.

Let $\breve{\theta}$ be an estimator of θ.

Note that, we should not treat this as a Bayesian estimation where θ is RV, but rather use frequentist view that is conditioned on θ, i.e., θ is regarded as constant.

$$MSE(\breve{\theta}) = E(\breve{\theta}-\theta)^2 = E(\breve{\theta}^2 - 2E(\breve{\theta})\theta + \theta^2)$$
$$= E(\breve{\theta}^2) - E^2(\breve{\theta}) + E^2(\breve{\theta}) - 2E(\breve{\theta})\theta + \theta^2$$
$$= Var(\breve{\theta}) + [E(\breve{\theta}) - \theta]^2$$

Variance of estimator || bias squared

We will showcase Gauss-Markov for a linear combination $\theta = a^T\beta$.
(Here, we can regard both a and β as constant)

Now, least square estimator is $\hat{\theta} = a^T\hat{\beta} = a^T X^+ y$

It's unbiased if we assume the linear model is correct:

$$E(a^T\hat{\beta}) = a^T X^+ E(y) = a^T X^+ X\beta = a^T\beta$$

For an arbitrary linear estimator wrt y, $c^T y = \breve{\theta}$.

By bias-var decomposition, $MSE(\breve{\theta}) = Var(\breve{\theta}) + bias^2(\breve{\theta})$

Thus for unbiased estimators, $Var(\breve{\theta}) = MSE(\breve{\theta})$

which means the least square estimator which minimizes MSE also minimizes variance.

Estimators of β and σ^2

treat RSS as constant. min f gives $\hat{\sigma}^2$.
min RSS, or least square gives $\hat{\beta}$.

MLE: comes from data likelihood function $f(Y|\beta,\vec{\sigma}) = c\cdot(\frac{1}{\sigma^2})^{n/2} e^{-\frac{1}{2\sigma^2}RSS}$

$\hat{\beta}$ = least square estimate

$\hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - (\sum \beta_j X_j))^2 = \frac{1}{n}RSS$ — $\hat{\beta}, \hat{\sigma}^2$ are still RVs

Unbiased:

$$\beta' = \hat{\beta}, \quad \sigma'^2 = \frac{1}{n-d}RSS = \frac{n}{n-d}\hat{\sigma}^2$$

Distribution of $\hat{\beta}_i$ (Simple LR)

Simple LR, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$, $\vec{x}$ fixed, $\vec{Y}$ RV.

$$\hat{\beta}_1 = \frac{SXY}{SXX}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

↳ Intuition: slope need to be of form Y/X

From the above definition of $\hat{\beta}_i$, taking $\mathbb{E}$ and Var of $\hat{\beta}_i$ and using assumption $\mathbb{E}(Y) = \beta_0 + \beta_1 x$, $Var(Y) = \sigma$,

We get distr of $\hat{\beta}_i$:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SXX}\right), \quad \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)\right),$$

$$cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{SXX}$$

- Estimator of $Var(\beta_i)$ can be obtained by substituting σ with estimators of σ, e.g., $\hat{\sigma}$, σ'.

Joint Distribution of $\hat{\beta}$, $\hat{\sigma}^2$ (Simple LR)

- Joint distr of $(\hat{\beta}_0, \hat{\beta}_1)$ is bivariate normal w/ μ, Σ above.

- For $n \geq 3$, $\hat{\sigma}^2$ is independent of $(\hat{\beta}_0, \hat{\beta}_1)$
  RSS is also independent of $\hat{\beta}$, because $\hat{\sigma}^2 = \frac{1}{n} RSS$

- $n\hat{\sigma}^2/\sigma^2 = \frac{RSS}{\sigma^2}$ has χ^2 distr w/ $n-2$ dof.

Intuition: $\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{\sum (Y_i - x_i^T \hat{\beta})^2}{\sigma^2}$

We use $\hat{\beta}$ to estimate β as it's unknown.
Suppose we have known β, and use that in the above eq,

$$\Rightarrow \frac{n\hat{\sigma}^2}{\sigma^2} = \frac{\sum (Y_i - x_i^T \beta)}{\sigma^2}$$

By assumption of LR, $Y_i \sim N(x_i^T \beta, \sigma^2)$.
Thus $(Y_i - x_i^T \beta)^2/\sigma^2 \sim \chi^2$ w/ 1 dof.
$n\hat{\sigma}^2/\sigma^2 = \sum (Y_i - x_i^T \beta)/\sigma^2 = \sum (\chi^2 \text{ w/ 1 dof}) = \chi^2$ w/ n dof.

Now by estimating β with $\hat{\beta}$, we removed
$\begin{cases} 2 \text{ dof } (\hat{\beta}_0, \hat{\beta}_1) \text{ for simple LR} \\ d \text{ dof } (\hat{\beta}_{d\times 1}) \text{ for multiple LR} \end{cases}$ from χ^2_n
or, $n\hat{\sigma}^2/\sigma^2 \sim \chi^2$ w/ $n-2$ dof for simple LR.

Draft Proof:

Let's construct orthogonal matrix $A_{n\times n}$:

1st column of A: $a_{1j} = \frac{1}{\sqrt{n}}$

2nd column of A: $a_{2j} = \frac{1}{\sqrt{SXX}}(x_j - \bar{x})$

Pad other columns with Gram-Schmidt.

Now, let $Z = AY$ ($Z_{n\times 1}$, $Y_{n\times 1}$)

It's possible to show:

- $Z_1 \cdots Z_n$ are independent and has normal distr w/ σ^2.

Now, since $Z_1 = \sum a_{1j} Y_j = \sqrt{n}\,\bar{Y}$, $\hat{\beta}_0 = \bar{Y} - \bar{x}\hat{\beta}_1$,

$$Z_1 = \sqrt{n}(\hat{\beta}_0 + \bar{x}\hat{\beta}_1).$$

Similarly, $Z_2 = \sum a_{2j} Y_j = \frac{1}{\sqrt{SXX}} \sum (x_j - \bar{x}) Y_j = \sqrt{SXX}\,\hat{\beta}_1$

which implies

$$\hat{\beta}_0 = \frac{1}{\sqrt{n}} Z_1 - \frac{\bar{x}}{\sqrt{SXX}} Z_2$$

$$\hat{\beta}_1 = \frac{1}{\sqrt{SXX}} Z_2$$

Since Z_1, Z_2 are independent normal RV, $(\hat{\beta}_0, \hat{\beta}_1)$ ~ bivariate normal.

Besides, we can show that $RSS = \sum_n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \cdots$

$$= \sum_{i=3} Z_i^2 \quad (n-2 \text{ dof})$$

$\hat{\sigma}^2 = RSS/n$ is independent of (Z_1, Z_2), and thus $(\hat{\beta}_0, \hat{\beta}_1)$.

$n\hat{\sigma}^2/\sigma^2 = RSS/\sigma^2 = \sum Z_i^2/\sigma^2$, thus it has χ^2 distr w/ $n-2$ dof.

Distribution of $\hat{\beta}$ (Multiple LR)

Multiple LR, $\hat{\beta}_{d\times 1}$, $X_{n\times d}$ fixed, $y_{n\times 1}$ RV

$$\hat{\beta} = X^{+}y, \quad \hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (X^T X)^{-1}\right)$$

Calculation of $\mathrm{Var}(\beta)$:

Theorem: If $z = Py$, $\mathrm{Var}(z) = P\,\mathrm{Var}(y)\,P^T$

Proof:
$$\mathrm{Var}(z) = \mathrm{Var}(Py) = \mathbb{E}\left((Py)(Py)^T\right) - \mathbb{E}(Py)\,\mathbb{E}(Py)^T$$
$$= P\,\mathbb{E}(yy^T)P^T - P\,\mathbb{E}(y)\mathbb{E}(y)^T P^T$$
$$= P\,\mathrm{Var}(y)\,P^T$$

Using theorem above,
$$\mathrm{Var}(\hat{\beta}) = (X^T X)^{-1} X^T \cdot \sigma^2 I \cdot X (X^T X)^{-1}$$
$$= \sigma^2 I \cdot (X^T X)^{-1}$$

Joint Distribution of $\hat{\beta}$, $\hat{\sigma}^2$ (Multiple LR)

- $\hat{\beta} \sim$ multivariate normal $\mathcal{N}\left(\beta, \sigma^2 (X^T X)^{-1}\right)$
- $\hat{\sigma}^2$ is independent of $\hat{\beta}$
- $n\hat{\sigma}^2/\sigma^2 \sim \chi^2$ w/ $n-d$ dof.

(Can be proved with similar procedure as in simple LR, but requires some more advanced technique.)

Testing Hypothesis (Simple LR)

- H_0: $\beta_i = v$ with t-test

WLOG, consider H_0: $\beta_1 = v$:

We know: $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{SXX})$, and we observe $\hat{\beta}_1$.

Then statistic

$$Z = \frac{\hat{\beta}_1 - v}{\sigma / \sqrt{SXX}} \sim \mathcal{N}(0,1).$$

That is enough if true var σ^2 is known.

O/W we need to leverage the fact $n\hat{\sigma}^2/\sigma^2 \sim \chi^2$.

t-test statistic:

$$Z = \frac{\hat{\beta}_1 - v}{\sigma' / \sqrt{SXX}} = \frac{\hat{\beta}_1 - v}{\sqrt{RSS}/\sqrt{(n-2)SXX}}$$

- H_0: $\beta_0 = v_0$, $\beta_1 = v_1$ with F-test

First use centered version of LR to get independent $\hat{\beta}_0^*$, $\hat{\beta}_1^*$.

$$y_i = \beta_0^* + \beta_1^*(x_i - \bar{x}) + \varepsilon_i \quad (\text{In fact, } \beta_1^* = \beta_1)$$

$$\hat{\beta}_0^* = \bar{y}, \quad \hat{\beta}_1^* = SXY/SXX$$

$$\hat{\beta}_0^* \sim \mathcal{N}(\beta_0^*, \frac{\sigma^2}{N}), \quad \hat{\beta}_1^* \sim \mathcal{N}(\beta_1^*, \frac{\sigma^2}{SXX})$$

For known σ^2, T statistics for $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ are independent.

Denote them as t_0, t_1, $t_i \sim \chi^2$ w/ 1 dof.

Combine them by $t = t_0 + t_1 \sim \chi^2$ w/ 2 dof.

t contains σ^2 which is unknown. To cancel σ^2, leverage $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$ to get a new statistic

$$Z = t / (n\hat{\sigma}^2/\sigma^2)$$ which contains only observable RVs.

$Z \sim F^2_{n-2}$. (because of the form χ^2/χ^2)

Testing Hypothesis (Multiple LR)

- $H_0: \beta_i = v$

  We know that $\hat{\beta}_i \sim \mathcal{N}\left(\beta_i, \underbrace{\text{diag}_i\left(\sigma^2 (X^T X)^{-1}\right)}_{\sigma_i^2}\right)$

  Thus we can construct statistics

  $$Z = \frac{\hat{\beta}_i - v}{\sigma_i} \sim \chi^2$$

- $H_0: R_{(d-s)\times d}\,\beta = r$

  This is a general linear hypothesis that involves multiple dimensions of β. It's possible to construct an F-statistic for this case.

Decomposition of Variation (for simple LR)

Let $SXX = \Sigma(x_i - \bar{x})(x_i - \bar{x})$, $S\{X,Y\}\{X,Y\}$ can be defined similarly.

Now, $SYY = \Sigma(y_i - \bar{y})^2$

$= \Sigma(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$

$= \underbrace{\Sigma(y_i - \hat{y}_i)^2}_{RSS} + \underbrace{\Sigma(\hat{y}_i - \bar{y})^2}_{SSReg} + 2\underbrace{\Sigma(y_i - \hat{y}_i)}_{e_i}\underbrace{(\hat{y}_i - \bar{y})}_{= \hat{y}_i - \bar{y}_i}$

RSS: (residual sum-of-sq)

SSReg: (regression sum-of-sq. also explained portion of variability by regression.)

e_i: (error term)

$= \hat{y}_i - \bar{y}_i$: ($\hat{y}$ deviation from mean)

$= \langle \vec{e}, \hat{y} - \bar{y} \rangle$

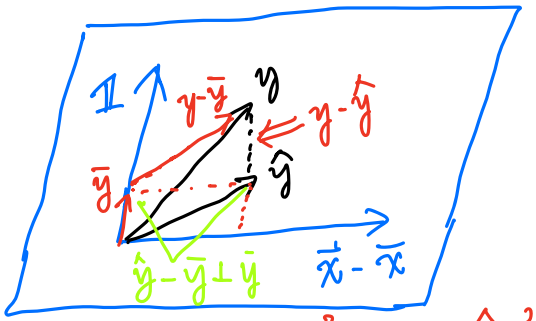
$= b_1 \langle \vec{e}, \vec{x} - \bar{x} \rangle$

$= 0$

By det, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

By choosing $\hat{\beta}_0$, $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$.

Thus $\hat{y} - \bar{y} = \hat{\beta}_1 (x - \bar{x})$.



$SYY = \|y - \bar{y}\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - \bar{y}\|^2$

This is an orthogonal dcmp of SYY.

(*) (will be referred to)

Thus, $SYY = RSS + SSReg$.

Note, the dcmp above is just the orthogonal decomposition with constant dimension removed from every term.

$SYY = \|y - \bar{y}\|_2^2$ fit with constant term

$RSS = \|y - \hat{y}\|_2^2 = \|y - \bar{y}\cdot 1 - (\hat{y} - \bar{y})\|_2^2$ fit with $x - \bar{x}$ term

$SSReg = \|\hat{y} - \bar{y}\|_2^2 \Rightarrow$ norm of fitting w/ $x - \bar{x}$ term

Analysis of Variance (ANOVA)

(used to compare the means of multiple normal distr)

(also useful for testing adequacy of LR)

Setting (One-way layout)

Random samples of d normal distr are available.

Variance of each distr is σ^2, we want to compare their means given observed values of the samples.

Notation: For $i=1 \cdots d$, we observe $Y_{i1}, Y_{i2} \cdots Y_{in_i} \sim N(\mu_i, \sigma^2)$

To find μ_i, we can solve the following:

$$Y = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{dn_d} \end{bmatrix} = \begin{bmatrix} 1\ 0 \cdots 0 \\ \cdots \\ 1\ 0 \cdots 0 \\ 0\ 1\ 0 \cdots 0 \\ 0\ 1\ 0 \cdots 0 \\ \vdots \\ 0 \cdots 0\ 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \vdots \\ 1 \\ \\ \\ \mu_d \end{bmatrix}$$

which is a multiple regression.

Let $\bar{Y}_{i+}$ be sample mean of n_i i-th population,

i.e., $\bar{Y}_{i+} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$.

MLE of σ^2, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{d} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i+})^2$

(sample var using mean of each corresponding population)

Target Hypothesis

H_0 : $\mu_1 = \mu_2 = \cdots = \mu_d$

Partitioning of Sum of Squares

Let $\bar{Y}_{++} = \frac{1}{n}\sum\sum Y_{ij}$, $S^2_{tot} = \sum_{i=1}^{d}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{++})^2$

We can partition S_{tot} as:

$$S^2_{tot} = S^2_{Res} + S^2_{Betw}$$

where $S^2_{Res} = \sum\sum(Y_{ij} - \bar{Y}_{i+})^2$ (in-population variation)

$S^2_{Betw} = \sum_{i=1}^{d} n_i(\bar{Y}_{i+} - \bar{Y}_{++})^2$ (between-population variation)

variation that can't be explained by it's population identity

variation explained by population id

Besides, $S^2_{Res}/\sigma^2 \sim \chi^2$ w/ $n-d$ dof,

S^2_{Res} and S^2_{Betw} are independent.

Also, if $\mu_1 = \cdots = \mu_d$, which is a useful scenario to test,

$S^2_{Betw}/\sigma^2 \sim \chi^2$ w/ $d-1$ dof.

Intuition of $d-1$ dof: S^2_{Betw} is of form $\sum(X_i - \bar{X})^2$

can use similar argument when proving sample var $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$.

(ANOVA table)

This partitioning can be summarized as a table:

Source of Variation	DOF	Sum of Square	Mean Square
Between population	$d-1$	S^2_{Betw}	$S^2_{Betw}/(d-1)$
Residual	$n-d$	S^2_{Res}	$S^2_{Res}/(n-d)$
Total	$n-1$	S^2_{tot}	

Mean Square: Sum of square / DOF, used in hypo testing.

Note: Residual mean square is σ'^2 of regression.

ANOVA Test

Let test statistic

$$U^2 = \frac{S^2_{Betw}/(d-1)}{S^2_{Res}/(n-d)}, \quad U^2 \sim F(d-1, n-d).$$

Intuition: If $\mu_i \neq \mu_j$, S^2_{Betw} will become large, and U^2 fall into reject region.

ANOVA and LR

As developed earlier, ANOVA is just a special type of multiple LR, with design matrix of specific form.

The partition of variation in ANOVA is backed by that of LR.

Thus the F-test statistics also has its counterpart in LR.

(checking adequacy of LR)

ANOVA as Significance Test in LR

When we apply LR, we made the linear assumption which is not warranted.

We can use ANOVA as a significance test for LR.

$S^2_{tot} \leftrightarrow SYY \quad S^2_{Betw} \leftrightarrow SS_{Reg} \quad S^2_{Res} \leftrightarrow RSS$

Thus we can develop ANOVA style test for LR:

H_0: β entries except for the constant term equals 0. (# group = 1 in ANOVA)

$$U^2 = \frac{SS_{Reg}/dof(SS_{Reg})}{RSS/dof(RSS)} = \frac{SS_{Reg}/(d-1)}{RSS/(n-d)}$$

Intuition: SS_{Reg}/RSS measures how much of variation is explained by LR and thus reflects significance of regression.

R^2 (coefficient of determination) and Goodness of Fit (Simple LR)

$$R^2 = 1 - \frac{RSS}{SYY} = \frac{SSReg}{SYY}$$

ratio of unexplained var

This step, as derived above (*), requires being able to choose β_0 such that $\langle \vec{e}, \hat{y} - \bar{y} \rangle = 0$.

Thus $R^2 = \frac{SSReg}{SYY}$ is only valid for LR with nonzero intercept.

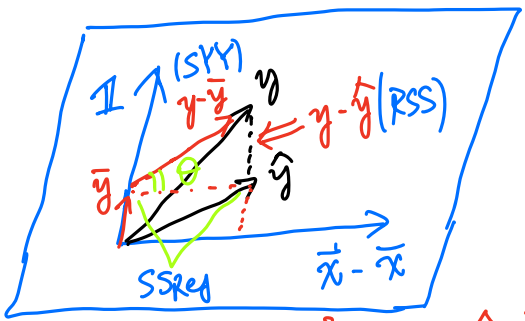
For LR with no intercept,

$$R^2 = 1 - \frac{RSS}{\sum y_i^2}$$

$R^2 \in [0, 1]$

measures goodness of fit.

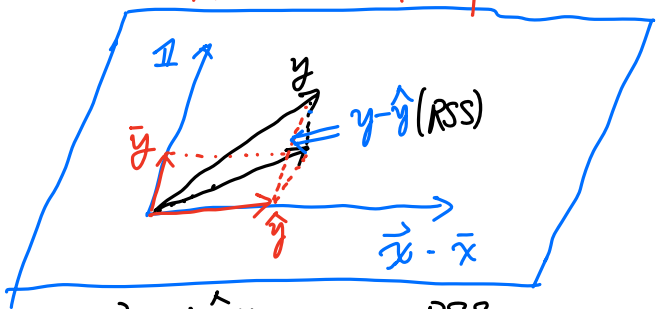
LR w/ intercept



$$SYY = \|y - \bar{y}\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - \bar{y}\|^2$$

$$R^2 = 1 - \frac{RSS}{SYY} = \frac{SSReg}{SYY}$$

LR w/o intercept



$$R^2 = \frac{\|\hat{y}\|}{\|y\|} = 1 - \frac{RSS}{\sum y_i^2}$$

Goodness of Fit for Multiple Regression

$$R^2 = 1 - RSS/SYY$$

where $RSS = \|y - \hat{y}\|_2^2$

$= y^T y - y^T (X\hat{\beta})$

$= y^T y - y^T \boxed{X X^+ y}$

$= SYY - SSReg$

$SYY = \|y - \bar{y}\|_2^2$

orthogonal projection of y on span(X)

$y^T \cdot proj(y) \Rightarrow$ component of 2-norm that's covered by span(X)

$\Rightarrow$ amount explained by reg

$\Rightarrow$ SSReg

(Decomposition of err arise naturally from geometric perspective)

Adjusted R^2: $\bar{R}^2$

$$\bar{R}^2 = 1 - \frac{RSS/(n-d)}{SYY/(n-1)} = 1 - \frac{n-1}{n-d}(1-R^2)$$

obtained from dividing RSS and SYY with their degree of freedom.

R^2 and Pearson Correlation (For simple LR)

Note, $R^2 = Corr(\vec{x}, \vec{y})^2$.

Proof:

$$\begin{aligned} R^2 &= SSReg / SYY \\ &= \sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2 \\ &= \sum(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 / \sum(y_i - \bar{y})^2 \\ &= \sum((\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 x_i - \bar{y})^2 / \sum(y_i - \bar{y})^2 \\ &= \hat{\beta}_1^2 \sum(x_i - \bar{x})^2 / \sum(y_i - \bar{y})^2 \\ &= \frac{SXY^2}{SXX^2} \cdot SXX / SYY \\ &= SXY^2 / (SXX \cdot SYY) \\ &= [SXY / (\sqrt{SXX} \cdot \sqrt{SYY})]^2 \\ &= corr^2(X, Y) \end{aligned}$$

Alt. Definition of R^2 (For simple LR)

R^2 is correlation of $\vec{x}$ and $\vec{y}$ squared. $R^2 = corr^2(\vec{x}, \vec{y})$

For general LR, $R^2 = Corr^2(y, \hat{y})$.

Multicollinearity

Multiple Regression from Simple Reg

Let $\vec{x}_{i\ n\times 1}$ $(i=1\cdots d)$ be i-th dim vector, if $\langle x_i, x_j\rangle = 0$.

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \begin{pmatrix} x_1^T \\ \vdots \\ x_d^T \end{pmatrix}(x_1 \cdots x_d)^{-1} \begin{pmatrix} x_1^Ty \\ \vdots \\ x_d^Ty \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{x_1^Tx_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{x_d^Tx_d} \end{pmatrix} \cdot \begin{pmatrix} x_1^Ty \\ \vdots \\ x_d^Ty \end{pmatrix}$$

or, $\hat{\beta}_i = \dfrac{\langle x_i, y\rangle}{\langle x_i, x_i\rangle}$ same as simple reg of y on x_i.

i.e., orthorgonal x_i $\Rightarrow$ can reduce multiple reg to simple reg on each dimension.

For general case (non orthorgonal), x_i s can be orthogonalized, e.g., w/ Gram Schmidt.

Let $X_{n\times d} = Q_{n\times d} R_{d\times d}$ ← upper triangular

$\text{diag}(R) = \mathbb{1}$ for Gram-Schmidt

$Q = (z_1 \cdots z_d)$, where z_i is residual of regressing x_i on $x_{<i}$

Note, if consider constant term, set $x_0 = z_0 = \mathbb{1}$.

Now, $\hat{\beta} = (X^T X)^{-1} X^T y = R^{-1} Q^T y$

$\hat{y} = Q Q^T y$

Consider last dim $\hat{\beta}_d$, $\hat{\beta}_d = \dfrac{\langle z_d, y\rangle}{\langle z_d, z_d\rangle}$

(We consider $\hat{\beta}_d$ here, but no loss of generality because we can rearrange to let any x_i be x_d.)

$$\text{Var}(\hat{\beta}_d) = \mathbb{E}(\hat{\beta}_d^2) - \mathbb{E}^2(\hat{\beta}_d)$$

$$= \frac{\mathbb{E}(z_d^T y y^T z_d) - [z_d^T \mathbb{E}(y)]^2}{(z_d^T z_d)^2} = \frac{z_d^T \left[\mathbb{E}(yy^T) - \mathbb{E}(y)\mathbb{E}(y)^T\right] z_d}{(z_d^T z_d)^2}$$

$$= \frac{z_d^T \text{cov}(y,y)\, z_d}{(z_d^T z_d)^2} = \frac{\sigma^2}{\|z_d\|_2^2}$$

Now, if x_d is correlated with some x_i, $\|z_d\|$ will be small and $\text{Var}(\hat{\beta}_d)$ will be large.

Or, more generally, the precision with which we can estimate $\hat{\beta}_i$ depends on how much x_i is unexplained by other x_i s.

Multicollinearity

Definition: X being rank deficient, or feature vectors are not linearly independent. This will cause LR estimation to have large variance.

The Canonical Form

(useful when X is rank deficient, causing X^TX to be singular)

Consider spectral decomposition of symmetric matrix

$$X^TX = P\Lambda P^T$$ (symmetric matrix can be orthogonally diagonalized)

$P = (p_1, \cdots, p_d)$, $PP^T = I$

The model becomes: $y = XPP^T\beta + \varepsilon$

$$= \tilde{X}\tilde{\beta} + \varepsilon$$

where $\tilde{X} = XP$, $\tilde{\beta} = P^T\beta$, and $\tilde{X}^T\tilde{X} = P^TX^TXP = \Lambda = \text{diag}(\lambda_1 \cdots \lambda_d)$,

which means columns of $\tilde{X}$ are orthorgonal.

Note: not $\hat{\tilde{\beta}}$

$$\text{Var}(\hat{\beta}) = \sigma^2(X^TX)^{-1} = \sigma^2 \sum \lambda_i^{-1} p_i p_i^T$$

$$\hat{\tilde{\beta}} = \tilde{X}^+ y = \Lambda^{-1}\tilde{X}^T y$$

$$\text{Var}(\hat{\tilde{\beta}}) = \sigma^2\Lambda^{-1}$$

Handling Exact Multicollinearity

Consider canonical form $y = \tilde{X}\tilde{\beta} + \varepsilon$

$\tilde{x}_i = X p_i$ is the i-th principle component

$\tilde{x}_i^T \tilde{x}_i = p_i^T X^T X p_i = \lambda_i$.

Assume $\text{rank}(X) = d - k$, then $\lambda_{d-k+1} \cdots \lambda_d = 0$.

Divide the matrices by $d-k, k$:

$P = (P_1, P_2)$ $\quad \Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}$ $\quad \tilde{X} = (\tilde{X}_1, \tilde{X}_2) = (XP_1, XP_2)$

$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} P_1^T \beta \\ P_2^T \beta \end{pmatrix}$, $\tilde{X}_2 = 0$ because $\tilde{x}_i^T \tilde{x}_i = 0$ for $\lambda_{d-k+1} \cdots \lambda_d$.

$$y = \tilde{X}_1 \tilde{\beta}_1 + \tilde{X}_2 \tilde{\beta}_2 + \varepsilon$$
$$= \tilde{X}_1 \tilde{\beta}_1$$

$$\hat{\tilde{\beta}}_1 = \tilde{X}_1^+ y, \quad \hat{\tilde{\beta}} = \begin{pmatrix} \hat{\tilde{\beta}}_1 \\ 0 \end{pmatrix} = X^+ y = (P \Lambda^{-1} P^T) X^T y$$

where $\Lambda^{-1} = \begin{pmatrix} \Lambda_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}$

Soft Solution to Multicollinearity

Includes Ridge, Shrinkage, Lasso ...

Shrinkage Methods

Ridge (L2) Regression

β do not have 0 index

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum \left(y_i - \beta_0 - \beta^T x_i \right)^2 + \lambda \| \beta \|_2^2 \right\}$$

Dual form:

$$\underset{\text{s.t.}}{\operatorname{argmin}} \sum (y_i - \beta_0 - \beta^T x_i)^2 \quad \text{s.t.} \quad \| \beta \|_2^2 \leq t$$

where there is 1 to 1 correspondence between λ and t.

Why regularization:

When there are correlated variables, there may exist large coeff for one var cancelled by large negative coeff of another variable, causing large variance.

Note, ridge is not equivalent under scaling of inputs.
Thus need to standardize before fitting.

Solution of Ridge

From primal form, $L(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$

$$\Rightarrow \quad \beta^* = (X^T X + \lambda I)^{-1} X^T y$$

Interpretation from SVD:

Let $X = UDV^T$,

$$X\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y = UU^T y = \sum u_j u_j^T y$$

$$X\hat{\beta}_{RIDGE} = X\beta^* = \sum_{j=1}^{d} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y$$

shrink SVD coordinate by $\frac{d_j^2}{d_j^2 + \lambda}$

dimension w/ smaller d_j^2 is shrinked more.

Lasso (L1)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum (y_i - \beta_0 - \beta^T x_i)^2 + \lambda \|\beta\|_1 \right.$$

Dual:

$$\underset{\beta}{\operatorname{argmin}} \sum (y_i - \beta_0 - \beta^T x_i)^2 \quad \text{s.t.} \quad \|\beta\|_1 \le t$$

Making t sufficiently small will cause certain dimension of β to be exactly zero.

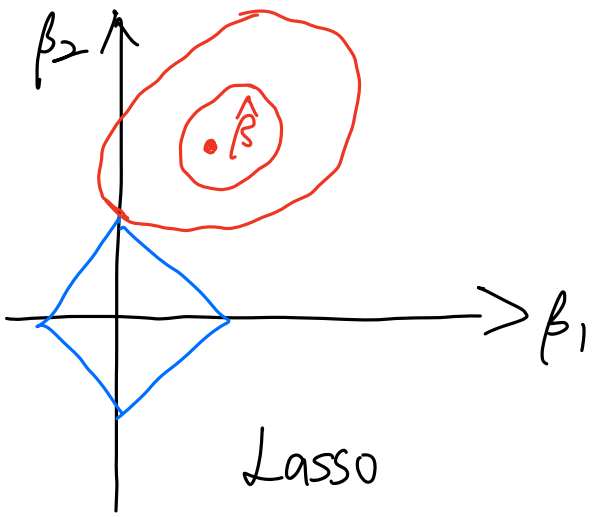
Varying $t \Rightarrow$ continuous subset selection

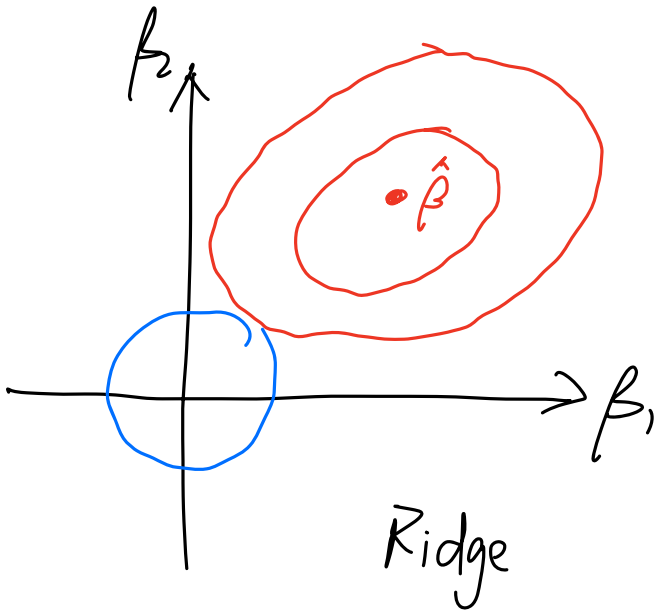
Subset Selection vs. Shrinkage Method

For orthorgonal X, $\hat{\beta}_j =$:

Best subset (size M)	$\hat{\beta}_j \cdot \mathbb{1}(\lVert\hat{\beta}_j\rVert \geq \lVert\hat{\beta}_M\rVert)$
Ridge	$\hat{\beta}_j / (1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j) \cdot (\lvert\hat{\beta}_j\rvert - \lambda)_+$

Illustration:





—— : constraint

—— : loss function contour

Note, the illustration shows intuitively why L1 (Lasso) tend to completely eliminate smaller dimensions (due to shape of contour)