Regression

$Y \in \mathbb{R}$ → target variable, $x \in \mathbb{R}^p$ → features

Def (Regression function)

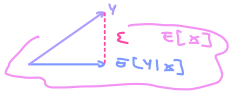
$\mathbb{E}[Y|x] : \mathbb{R}^p \to \mathbb{R}$

Obs: if $(Y, x) \sim N_{p+1} \Rightarrow \mathbb{E}[Y|x] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ (linear regression)

Data: $\mathcal{X} = \begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$ $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ $i = 1, \dots, n$ $\Rightarrow$ $\hat{f}$ ← estimate of regress. function $\hat{f} : \mathbb{R}^p \to \mathbb{R}$

General model: $Y = \mathbb{E}[Y|x] + \varepsilon$ $\quad \varepsilon \perp\!\!\!\perp x$

$Y = f(x) + \varepsilon$

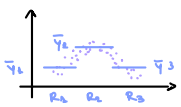


Two different approaches to deal with it:

1. Totally data driven, no model for f: "Let the data speak!" → CART (good in prediction, bad for interpretation)
2. Structured models → linear regression models (good for interpretation)

CART (Classification and regression tree)

Idea: Split the feature space $\mathbb{R}^p$ in $R_1, \dots, R_J$ partition of $\mathbb{R}^p$ and predict y in R_i by means of $\bar{y}_i = \frac{1}{|R_i|} \sum_{x_i \in R_i} y_i$



$\hat{f}(x) = \sum_{j=1}^{J} \bar{y}_j \mathbb{1}[x \in R_j]$ step-wise function

Problems and Goals: how to find $R_1, \dots, R_J$ s.t.

1. $\bigcup_{j=1}^{J} R_j = \mathbb{R}^p$
2. $R_i \cap R_j = \emptyset$ for $i \neq j$ $\quad i, j = 1, \dots$
3. $R_i \cap \{x_i,\ i = 1, \dots, n\} \neq \emptyset$ $\quad i = 1, \dots, J$

Optimal criterion: find $R_1, \dots, R_J$ s.t. $\sum_{j=1}^{J} \sum_{x_i \in R_j} (y_i - \bar{y}_j)^2$ is minimum (too many subsets → computationally unsolvable)

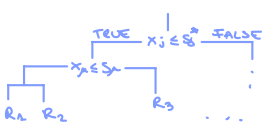
CART looks for rectangles $R_1, \dots, R_J$

have to optimize through previous criteria

↳ optimal can be an R_i for each x_i ← overfitting

Algorithm:

1. look at x_2 and find s_2^* (split) which maximizes:
   $\sum_{i=1}^{n} (y_i - \bar{y})^2 - \left[\sum_{x_{2i} \le s_2} (y_i - \bar{y}_-)^2 + \sum_{x_{2i} > s_2} (y_i - \bar{y}_+)^2 \right]$ ⊕ with $\bar{y}_+ = \frac{1}{|\{\dots\}|} \sum_{i : x_i > s_2} y_i$ and $y_- = \dots$
2. repeat (1) for $x_2, x_3, \dots, x_p$ → $s_2^*, s_3^*, \dots, s_p^*$
3. choose s_j^* which maximizes (*). We now have more dataset:
   $R_1 = \{x_i \in \mathcal{X} : x_{ij} \le s_j^*\}$, $R_2 = \{\dots : x_{ij} > s_j^*\}$
4. Repeat (1), (2), (3) for all R_i.
5. Stop splitting an R_i if number of data points in R_i is below treshold



Whenever I stop earlier I get anyways a separation of the features space. A goal could be to maximize purity in my rectangles. Useful for categorical and continuous variables

To control overfitting: introduce penalization for J and minimize: $\boxed{\sum_{j=1}^{J} \sum_{x_i \in R_j} (y_i - \bar{y}_j)^2 + \alpha J = W(\alpha)}$

⇒ Introduce an extra param α ⇒ chosen by cross-validation

Linear model for REGRESSION

Data $X = \begin{bmatrix} (x_{11} \dots x_{1p})' = x_1' & y_1 \\ (x_{21} \dots x_{2p})' = x_2' & y_2 \\ \vdots & \vdots \\ (x_{n1} \dots x_{np})' = x_n' & y_n \end{bmatrix}$

Design matrix: $Z = \begin{bmatrix} 1 & z_{11} & \dots & z_{1r} \\ 1 & z_{21} & \dots & z_{2r} \\ \vdots & & & \vdots \\ 1 & z_{n1} & \dots & z_{nr} \end{bmatrix}$ $n \times (r+1)$

with: $z_1 = h_1(x_1 \dots x_p) \dots z_r = h_r(x_1 \dots x_p)$ (special case $r=p$ $z_1 = x_1 \dots z_p = x_p$)

and $h_1 \dots h_r$ are KNOWN

We will have: $\mathbb{E}[Y \mid x_1, \dots, x_p] = \beta_0 + \beta_1 h_1(x_1 \dots x_p) + \beta_2 h_2(x_1 \dots x_p) + \dots + \beta_r h_r(x_1 \dots x_p) = \beta_0 + \beta_1 z_1 + \dots + \beta_r z_r$

Model for the phenomenon: $Y = \beta_0 + \beta_1 z_1 + \dots + \beta_r z_r + \varepsilon$ ε r.v. $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$

Model for the data: $y = (y_1 \dots y_n)'$ $y = Z\beta + \varepsilon$ with: Z $n \times (r+1)$ known

ε r.v. s.t. $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Cov}(\varepsilon) = \sigma^2 I$

i.e. for $i = 1, \dots, n$ $y_i = \beta_0 + \beta_1 z_{i1} + \dots + \beta_p z_{ip} + \varepsilon_i$ ε_i s.t. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$ $\mathbb{E}[\varepsilon_i] = 0$

EXAMPLE (ANOVA + linear regression)

$$\mathbb{I} = \begin{cases} x_{11}, \dots, x_{1n_1} & \text{iid} \sim N(\mu, \sigma^2) \\ \vdots \\ x_{g1}, \dots, x_{gn_g} & \text{iid} \sim N(\mu, \sigma^2) \end{cases}$$

$y = (x_{11}, \dots, x_{1n_1}, x_{21}, \dots, x_{2n_2}, \dots, x_{g1}, \dots, x_{gn_g})' \in \mathbb{R}^n$ $n = n_1 + n_2 + \dots + n_g$

$Z = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & \vdots & 0 & & \vdots \\ \vdots & 1 & \vdots & & \vdots \\ \vdots & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix}$ (n_1, n_2, n_g blocks)

$\beta = (\mu, \tau_1, \dots, \tau_g)$

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

$\Rightarrow y = Z\beta + \varepsilon \iff \begin{cases} x_{ij} = \mu + \tau_i + \varepsilon_{ij} \\ \varepsilon_{ij} \quad \text{iid} \sim N(0, \sigma^2) \end{cases}$ $i = 1, \dots, g$, $j = 1, \dots n_i$

$\Rightarrow$ it's an ANOVA as CRM

notice: Z is not full rank $\longrightarrow$ problem of identification of parameter.

Geometrically speaking is not a problem, but if I want a unique solution it is!

To get it as full rank is "easy": including the constraint: $\sum_{i=1}^{g} n_i \tau_i = 0$, $\tau_g = \sum_{i=1}^{n-1} \frac{n_i}{n_g} \tau_i$, $n_g = 0$

We have:

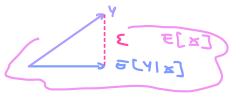
$Z = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & \vdots & 0 & & \vdots \\ \vdots & 1 & \vdots & & \vdots \\ \vdots & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & -n_1/n_g & -n_2/n_g & \dots & -n_{g-1}/n_g \end{bmatrix}$ $n \times g$

$\beta = (\mu, \tau_1, \dots, \tau_{g-1}) \in \mathbb{R}^g$

$y = Z\beta + \varepsilon$

full rank makes computation easier, but not required

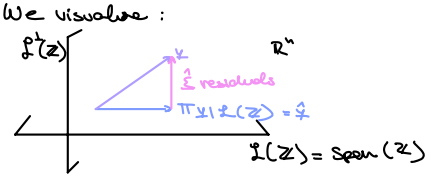
How to estimate β and σ^2?



$y = Z\beta + \varepsilon$ $Z\beta = \mathbb{E}[Y \mid x]$

$Z = [c_1, \dots, c_{r+1}]$ $c_i \in \mathbb{R}^n$, $Z\beta = \beta_0 c_1 + \dots + \beta_r c_{r+1}$

We visualize:



$\pi_{Y \mid L(Z)} = Z\hat{\beta}$, $Z\hat{\beta}$ is the solution of OLS (ordinary least square):

$\underset{\beta \in \mathbb{R}^{r+1}}{\operatorname{argmin}} \| y - Z\beta \|^2 = \underset{\beta \in \mathbb{R}^{r+1}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 z_{i1} - \dots - \beta_r z_{ir})^2 = \hat{\beta}$

Proposition

If Z is full rank:

1. $\hat{y} = \pi_{Y \mid L(Z)} = Z(Z'Z)^{-1} Z' y = H y$
2. $\hat{\beta} = (Z'Z)^{-1} Z' y$

Proof: $Z'Z$ $(r+1) \times (r+1)$

$Z'Z = \sum_{i=1}^{r+1} \lambda_i e_i e_i'$ $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{r+1} > 0$ (Z full rank)

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

For $i = 1, \dots, r+1$ $\quad q_i = \frac{1}{\sqrt{\lambda_i}} Z e_i$ s.t.

1. $q_i \in \mathcal{L}(Z)$, $i = 1, \dots, r+1$
2. $q_i' q_j = \frac{1}{\sqrt{\lambda_i \lambda_j}} e_i' Z' Z e_j = \sqrt{\frac{\lambda_j}{\lambda_i}} e_i' e_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$

$\Rightarrow \{q_1, q_2, \dots, q_{r+1}\}$ orthonormal basis for $\mathcal{L}(Z)$

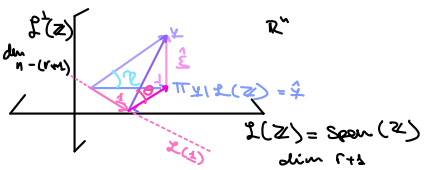
$$\pi_{y|Z} = \sum_{i=1}^{r+1} \frac{q_i q_i'}{q_i' q_i} y = \sum_{i=1}^{r+1} \frac{1}{\lambda_i} Z e_i e_i' Z' y = Z \Big(\sum_{i=1}^{r+1} \frac{1}{\lambda_i} e_i e_i' \Big) Z' y = \underbrace{Z (Z'Z)^{-1} Z'}_{H} y$$

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

Obs:

1. If $\operatorname{rank}(Z) = k < r+1 \leq n$ in the proof: $\lambda_1 \geq \lambda_2 \geq \dots \lambda_k > 0$, $\lambda_{k+1} = \lambda_{k+2} = \dots = \lambda_{r+1} = 0$
   $\Rightarrow$ cannot divide $\Rightarrow$ 1st step summation at step k
   $\Rightarrow$ I get the generalized inverse $\Rightarrow$ $H = Z (Z'Z)^{-} Z'$ and $\hat{\beta} = (Z'Z)^{-} Z y$
2. If $\operatorname{rank}(Z) = r+1 = n \Rightarrow y = \hat{y}$
   NOT GOOD, enormous generalization error (unless $y = Z\beta + \varepsilon$ $\operatorname{Cov}(\varepsilon) = 0$ stupid!!)

EXAMPLE



$\hat{y} = H y$
$\hat{\varepsilon} = (I - H) y$
$H = Z (Z'Z)^{-1} Z'$ orth proj on $\mathcal{L}(Z)$ $(\dim = r+1)$
$I - H$ orth project. on $\mathcal{L}^\perp(Z)$ $\dim n-(r+1)$
$\|y\|^2 = \|\hat{y}\|^2 + \|\hat{\varepsilon}\|^2$
$SS_{tot} = SS_{reg} + SS_{res}$ 1st decomposition of variance formula

$\pi_{y|1} = \bar{y} \cdot 1$

$\pi_{\hat{y}|1} = \frac{1 1'}{1'1} \hat{y} = \frac{1 1'}{1'1} H y = \frac{1 1'}{1'1} y = \bar{y} \cdot 1$
$\quad \hookrightarrow (1'H) = (H'1)' = (H1)' = 1'$

$\|y - \bar{y} 1\|^2 = \|\hat{y} - \bar{y} \cdot 1\|^2 + \|\hat{\varepsilon}\|^2$
$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{\varepsilon}_i^2$
$SS = SS_{reg} + SS_{res}$ 2nd decomposition of variance formula
$1 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} + \frac{\sum \hat{\varepsilon}_i^2}{\sum (y_i - \bar{y})^2}$
$\quad \hookrightarrow R^2$ coeff. of determination $\longrightarrow R^2 = 1 - \frac{\sum \hat{\varepsilon}_i^2}{\sum (y_i - \bar{y})^2} = 1 - \sin^2\theta = \cos^2\theta$

Obs:

1. $R^2 = 1 \Rightarrow \theta = 0 \Rightarrow y = \hat{y}$
2. $R^2 = 0 \Rightarrow \theta = \frac{\pi}{2} \Rightarrow \bar{y} \cdot 1 = \hat{y}$

Rmk:

$$Z = \begin{bmatrix} 1 & z_{11} & \dots & z_{1r} \\ 1 & z_{21} & \dots & z_{2r} \\ \vdots & & & \vdots \\ 1 & z_{n1} & \dots & z_{nr} \end{bmatrix}$$

If $1 \notin \mathcal{L}(Z)$, e.g. regression through the origin

$\tilde{R}^2 = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|y\|^2} = \cos^2\eta$

$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2 / (n-(r+1))}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)}$ $\quad R^2$ adjusted

Properties (of $\hat{\beta}$ and $\hat{\varepsilon}$)

Assume Z to be full rank $(r+1)$

$\hat{\beta} = (Z'Z)^{-1} Z' y$
$\hat{\varepsilon} = (I - H) y$
$H = Z (Z'Z)^{-1} Z'$

1. $E[\hat{\beta}] = \beta$ (unbiased)
2. $\operatorname{Cov}(\hat{\beta}) = \sigma^2 (Z'Z)^{-1}$
3. $E[\hat{\varepsilon}] = 0$ (recall $E[\varepsilon] = 0$)
4. $\operatorname{Cov}[\hat{\varepsilon}] = \sigma^2 (I - H)$ (recall $\operatorname{cov}(\varepsilon) = \sigma^2 I$)
5. $E[\hat{\varepsilon}'\hat{\varepsilon}] = E[\|\hat{\varepsilon}\|^2] = (n-(r+1)) \sigma^2 \Longrightarrow E\left[\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-(r+1)}\right] = \sigma^2$ (unbiased)

Proof:

1. $\mathbb{E}[\hat{\beta}] = \mathbb{E}[(Z'Z)^{-1}Z'Y] = (Z'Z)^{-1}Z\,\mathbb{E}[Y]$
   $Y = Z\beta + \varepsilon$
   $\mathbb{E}[Y] = Z\beta + \mathbb{E}[\varepsilon] = Z\beta$
   $\mathbb{E}[\hat{\beta}] = (Z'Z)^{-1}Z'Z\beta = \beta$

2. $\mathrm{Cov}(\hat{\beta}) = (Z'Z)^{-1}Z'\,\mathrm{Cov}(Y)\,Z(Z'Z)^{-1} = \sigma^2(Z'Z)^{-1}Z'Z\,(Z'Z)^{-1} = \sigma^2(Z'Z)^{-1}$
   $\mathrm{Cov}(Y) = \mathrm{Cov}(\varepsilon) = \sigma^2 I$

3. $\mathbb{E}[\hat{\varepsilon}] = (I-H)\mathbb{E}[Y] = (I-H)Z\beta = 0$

4. $\mathrm{Cov}(\hat{\varepsilon}) = (I-H)\mathrm{Cov}(Y)(I-H)' = \sigma^2(I-H)(I-H)' = \sigma^2(I-H)$

5. $\mathbb{E}[\hat{\varepsilon}'\hat{\varepsilon}] = \mathrm{tr}\,\mathbb{E}[\hat{\varepsilon}'\hat{\varepsilon}] = \mathbb{E}[\mathrm{tr}(\hat{\varepsilon}'\hat{\varepsilon})] = \mathbb{E}[\mathrm{tr}(\hat{\varepsilon}\hat{\varepsilon}')]$
   $\hat{\varepsilon} = (I-H)Y = (I-H)(Z\beta+\varepsilon) = (I-H)\varepsilon$
   $\mathbb{E}[\hat{\varepsilon}'\hat{\varepsilon}] = \mathbb{E}[\mathrm{tr}((I-H)\varepsilon\varepsilon'(I-H))] = \mathrm{tr}[(I-H)\mathbb{E}[\varepsilon\varepsilon'](I-H)] = \mathrm{tr}(\sigma^2(I-H)) =$
   $\overset{!}{=} \sigma^2\,\mathrm{tr}(I-H) = \sigma^2(n - \mathrm{tr}(H))$
   $\mathrm{tr}(H) = \mathrm{tr}(Z(Z'Z)^{-1}Z') = \mathrm{tr}(Z'Z(Z'Z)^{-1}) = \mathrm{tr}(I_{r+1}) = r+1$
   $\mathbb{E}[\hat{\varepsilon}'\hat{\varepsilon}] = \sigma^2(n-(r+1))$

this tr trick is going to be in the exam to be used

Obs: $\mathrm{Cov}(\hat{\beta}) = \sigma^2(Z'Z)^{-1}$

→ $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_r)'$ → are not uncorrelated UNLESS $Z'Z = \alpha I$

$Y = \hat{\beta}_0 + \hat{\beta}_1 z_1 + \dots + \hat{\beta}_r z_r$

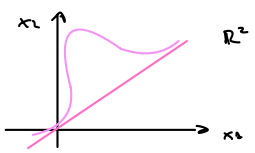
↳ $\beta_1/2$ like that if they are correllated, then also the other ones are changed

→ You can control the variab. of $\hat{\beta}$ by controlling the design matrix

Assumption UNTIL NEW NOTICE: $\varepsilon \sim N_n(0, \sigma^2 I)$
$Y = Z\beta + \varepsilon$

Proposition

1. $\hat{\beta}$ and $\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$ are MLE est of β and σ^2
2. $\hat{\beta} \sim N_{r+1}(\beta, \sigma^2(Z'Z)^{-1})$
3. $\hat{\varepsilon} \sim N_n(0, \sigma^2(I-H))$
4. $\hat{\varepsilon} \perp\!\!\!\perp \hat{\beta}$
5. $\hat{\varepsilon}'\hat{\varepsilon} \sim \sigma^2 \chi(n-(r+1))$



Proof: Note: $Y = Z\beta + \varepsilon \sim N_n(Z\beta, \sigma^2 I)$

1. write the Lik and short differentiating

2/3/4 $\begin{pmatrix}\hat{\beta}\\ \hat{\varepsilon}\end{pmatrix} = \underbrace{\begin{pmatrix}(Z'Z)Z'\\ I-H\end{pmatrix}}_{A} Y \sim N\left(\begin{pmatrix}\beta\\ 0\end{pmatrix}, \begin{bmatrix}(Z'Z)^{-1} & 0\\ 0 & I-H\end{bmatrix}\right)$

$N(AZ\beta, \sigma^2 AIA')$

$\hat{\varepsilon} \sim N_n(0, \sigma^2(I-H))$

$\frac{1}{\sigma^2}(\hat{\varepsilon}-0)'(I-H)^{-1}(\hat{\varepsilon}-0) \sim \chi^2_{(n-(r+1))}$

Exercise: $x \sim N_p(\mu, \Sigma)$ $\quad \mathrm{rank}(\Sigma) = k < p \implies (x-\mu)'\Sigma(x-\mu) \sim \chi^2(k)$

Linear models

$\mathbb{R}^n \ni \boxed{Y = Z\beta + \varepsilon}$

Z design matrix $(r+1) \times n$ (full rank)
$\beta \in \mathbb{R}^{r+1}$
$\varepsilon \sim N_n(0, \sigma^2 I)$

$\hat{\beta} = (Z'Z)^{-1} Z' Y$
$S^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-(r+1)}$
$\hat{\varepsilon} = (I - H) Y$
$H = Z(Z'Z)^{-1}Z'$
$\hat{Y} = H Y$

Prop:
1. $\hat{\beta} \sim N_{r+1}(\beta, \sigma^2 (Z'Z)^{-1})$
2. $\hat{\varepsilon} \sim N_n(0, \sigma^2 (I - H))$ } $\perp\!\!\!\perp$
3. $\hat{\varepsilon}'\hat{\varepsilon} \sim \sigma^2 \chi^2(n-(r+1))$

Note:
$\frac{1}{\sigma^2}(\hat{\beta}-\beta)'(Z'Z)(\hat{\beta}-\beta) \sim \chi^2(r+1)$ } $\perp\!\!\!\perp$
$\frac{1}{\sigma^2}\hat{\varepsilon}'\hat{\varepsilon} \sim \chi^2(n-(r+1))$

$$\frac{(\hat{\beta}-\beta)'(Z'Z)(\hat{\beta}-\beta)/r+1}{\underbrace{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-(r+1)}}_{=S^2}} \sim F(r+1, n-(r+1)) \implies \frac{1}{S^2}(\hat{\beta}-\beta)'(Z'Z)(\hat{\beta}-\beta) \sim (r+1) F(r+1, n-(r+1))$$

$\alpha \in (0,1)$
$CR_{1-\alpha}(\beta) = \{ \eta \in \mathbb{R}^{r+1} : (\hat{\beta}-\eta)' Z'Z (\hat{\beta}-\beta) \le (r+1) S^2 F_{1-\alpha}(r+1, n-(r+1)) \}$

$a \in \mathbb{R}^{r+1}$
$a'\hat{\beta} \sim N_1(a'\beta, \sigma^2 a'(Z'Z)^{-1}a)$ $\implies$ $\frac{a'(\hat{\beta}-\beta)}{\sigma\sqrt{a'(Z'Z)^{-1}a}} \sim N(0,1)$

$$\implies \frac{\frac{a'(\hat{\beta}-\beta)}{\sigma\sqrt{a'(Z'Z)^{-1}a}}}{\sqrt{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-(r+1)}\frac{1}{\sigma^2}}} \sim t(n-(r+1)) \implies \frac{a'(\hat{\beta}-\beta)}{S\sqrt{a'(Z'Z)^{-1}a}} \sim t(n-(r+1))$$

$\alpha \in (0,1)$ $\quad CI_{1-\alpha}(a'\beta) = \left[a'\hat{\beta} \pm S\sqrt{a'(Z'Z)^{-1}a}\; t_{1-\frac{\alpha}{2}}(n-(r+1)) \right]$ are at the two intervals for $a'\beta$

In particular
$a = (0 \cdots 0 \underset{i}{1} 0 \cdots 0)$
$CI_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i \pm S\sqrt{\text{diag}_i (Z'Z)^{-1}}\; t_{1-\frac{\alpha}{2}}(n-(r+1)) \right]$
$\alpha \longrightarrow \frac{\alpha}{r+1}$ (Bonferroni)

Test:
$H_0: \beta_i = 0$ vs $H_1: \beta_i \neq 0$
Reject H_0 at level $\alpha \in (0,1)$ if $T_i = \frac{|\hat{\beta}_i|}{S\sqrt{\text{diag}_i (Z'Z)^{-1}}} > t_{1-\frac{\alpha}{2}}(n-(r+1)) \implies$ p-value

$$\max_{a \in \mathbb{R}^{r+1}} \frac{[a'(\hat{\beta}-\beta)]^2}{[S\sqrt{a'(Z'Z)^{-1}a}]} = \max_{a \in \mathbb{R}^{r+1}} \frac{[a'(\hat{\beta}-\beta)]^2}{S^2[a'(Z'Z)^{-1}a]} = \frac{1}{S^2}(\hat{\beta}-\beta)(Z'Z)(\hat{\beta}-\beta) \sim F(r+1, n-(r+1))$$

$SimCI_{1-\alpha}(a'\beta) = \left[a'\hat{\beta} \pm S(a'(Z'Z)^{-1}a)\sqrt{F_{1-\alpha}(r+1, n-(n+1))} \right]$

$\frac{1}{\sigma^2}\hat{\varepsilon}'\hat{\varepsilon} \sim \chi^2[n-(n+1)]$
$\frac{(n-(r+1))}{\sigma^2} S^2 \sim \chi^2(n-(r+1))$

$P\left[\chi^2_{\frac{\alpha}{2}}(n-(r+1)) < \frac{(n-(r+1))S^2}{\sigma^2} < \chi_{1-\frac{\alpha}{2}}(n-(r+1)) \right] = 1-\alpha$
$\alpha \in (0,1)$
$CI_{1-\alpha}(\sigma^2) = \left[\frac{(n-(r+1))S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-(r+1))}, \frac{(n-(r+1))S^2}{\chi^2_{\frac{\alpha}{2}}(n-(r+1))} \right]$

C $p\times(r+1)$ matrix

$H_0: C\underline{\beta} = \underline{0}$ vs $C\underline{\beta} \neq \underline{0}$ (Ex $H_0: C\beta = \underline{K}_0$ vs $C\beta \neq \underline{K}_0$)

$$C\underline{\beta} = \begin{bmatrix} C_{11}\beta_0 + \dots + C_{1,r+1}\beta_r \\ \vdots \\ C_{p1}\beta_0 + \dots + C_{p,r+1}\beta_r \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$C\hat{\underline{\beta}} \sim N_p\left(C\underline{\beta}, \sigma^2 C(Z'Z)^{-1}C'\right)$$

Under H_0

$$\frac{\frac{1}{\sigma^2}\frac{(C\hat{\beta})'[C(Z'Z)^{-1}C']^{-1}(C\hat{\beta})}{p} \sim \chi^2(p)}{\frac{1}{\sigma^2}\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-(r+1)} \sim \chi^2(n-(r+1))} \sim F(p, n-(r+1))$$

$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-(r+1)} = S^2$

$$F = \frac{1}{S^2}(C\hat{\beta})'[C(Z'Z)^{-1}C']^{-1}(C\hat{\beta}) \sim pF(p, n-(r+1))$$

Reject $H_0: C\beta = \underline{0}$ at level $\alpha \in (0,1)$ if $F > pF_{1-\alpha}(p, n-(r+1))$

$H_0: \beta_0 = \beta_1 = \dots = \beta_{r-(p-1)} = 0$ vs $H_1: \exists\, \beta_i \neq 0$, $i = r-(p-1), \dots, r$

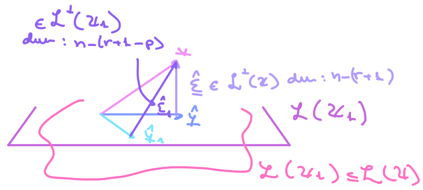
$$C = \left[\begin{array}{ccc|ccc} 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & 0 & 1 \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{array}\right]\Big\}p = [\underline{0}\ \ I_p] \quad p\times(r+1)$$

(the right block is p columns wide)

$H_0: \beta_r = \beta_{r+1} = \dots = \beta_{r-(p+1)} = 0$

is testing: full model $\underline{Y} = Z\underline{\beta} + \underline{\varepsilon}$

vs reduced model $\underline{Y} = Z_1\underline{\beta}_1 + \underline{\varepsilon}$



Reject H_0 if $SS_{res}(Z_1) - SS_{res}(Z)$ is large

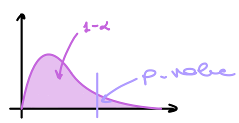
but $\dfrac{SS_{res}(Z_1) - SS_{res}(Z)}{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-(r+1)}} \sim F(p, n-(r+1))$ ← $S^2 \cdot p$

Very particular case:

$H_0: \beta_1 = \beta_2 = \dots = \beta_r = 0$ vs $H_1: \exists\, \beta_i \neq 0$

$\Longrightarrow Z_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ $\quad SS_{res}(Z_1) = \sum(y_i - \bar{y})^2$

$$\frac{SS_{res}(Z_1) - SS_{res}(Z)}{S^2 \cdot r} = \frac{\sum(y_i - \bar{y})^2 - \sum\hat{\varepsilon}_i^2}{S^2 \cdot r} = \frac{\sum(\hat{y}_i - \bar{y})^2 / r}{\sum(\hat{\varepsilon}_i)^2 / n-(r+1)} \sim F(r, n-(r+1))$$



Reject H_0 (i.e. there is something in the model) if $F > F_{1-\alpha}(r, n-(r+1))$

PREDICTION

Model for the phenomenon $Y_0 = \underline{z}_0'\underline{\beta} + \varepsilon_0$

↳ $\underline{z}_0' = (1\ z_{01} \dots z_{0r})$ new case, not in the training set

predict: $\mathbb{E}[Y_0 | \underline{z}_0] = \underline{z}_0'\underline{\beta}$

Unbiased predictor of $\underline{z}_0'\underline{\beta}$: $\underline{z}_0'\hat{\underline{\beta}} = \underline{z}_0'(Z'Z)^{-1}Z'\underline{Y}$

Indeed: $\mathbb{E}[\underline{z}_0'\hat{\underline{\beta}}] = \underline{z}_0'\underline{\beta}$

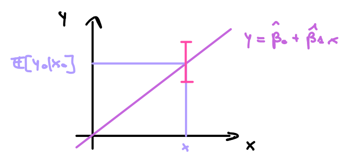
Gauss-Markov thm

$\underline{z}_0'\hat{\beta}$ is BLUE i.e. Best Linear Unbiased Estimator of $\underline{z}_0'\underline{\beta}$

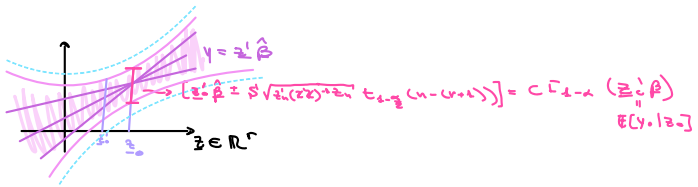
Point prediction

$\underline{z}_0'\underline{\beta}$

$$CI_{1-\alpha}(\underline{z}_0'\underline{\beta}) = \left[\underline{z}_0'\hat{\underline{\beta}} \pm S\sqrt{\underline{z}_0'(Z'Z)^{-1}\underline{z}_0}\; t_{1-\frac{\alpha}{2}}(n-(r+1))\right]$$



Note



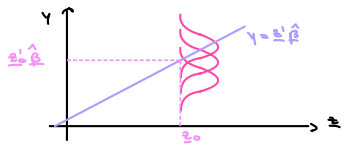
"The bands contain the real model with conf $1-\alpha$"
↳ wrong! each interval 95% not together → Sim CI

$[z_0'\hat{\beta} \pm s(z_0'(Z'Z)^{-1}z_0)\sqrt{F_{1-\alpha}(r+1, n-(r+1))}]$ all possible simultaneously

In this case not Bonferroni because $\to +\infty$, so F-test

Sometimes we want to estimate for $\mathbb{E}[y_0|z_0] \to \widehat{\mathbb{E}[y_0|z_0]}$

$y_0 = z_0'\beta + \varepsilon_0$
$\mathbb{E}[y_0|z_0] = z_0'\hat{\beta}$
$\varepsilon_0 \perp\!\!\!\perp \varepsilon_1, \varepsilon_2, \dots$



not true → estimate has distribution

Question: I interval s.t. $\mathbb{P}[y_0 \in I \mid z_0] = 1-\alpha$

We know $y_0 \sim N(z_0'\beta, \sigma^2)$, $\perp\!\!\!\perp$, $z_0'\hat{\beta} \sim N(z_0'\beta, \sigma^2(Z'Z)^{-1})$
so $y_0 - z_0\hat{\beta} \sim N(0, \sigma^2(1 + z_0'(Z'Z)^{-1}z_0)$
so $\dfrac{y_0 - z_0\hat{\beta}}{\sigma\sqrt{1+(Z'Z)^{-1}}} \Big/ \sqrt{1/\sigma^2\left(\hat{\varepsilon}'\hat{\varepsilon}/(n-(r+1))\right)} \sim t(n-(r+1)) \implies I = \left[z_0'\hat{\beta} \pm s\sqrt{1 + z_0'(Z'Z)^{-1}z_0}\, t_{1-\frac{\alpha}{2}}(n-(r+1))\right]$

Lin model

$y = Z\beta + \varepsilon$
$\mathbb{E}[\varepsilon] = 0$, $\mathrm{Cov}(\varepsilon) = \sigma^2 I$ ie: $\begin{cases} \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 & \text{if } i \neq j \\ \mathrm{Var}(\varepsilon_i) = \sigma^2 \end{cases}$

? $\mathrm{Cov}(\varepsilon) = \sigma^2 \Sigma$ (Σ $n \times n$ cov.)
(σ^2: unknown, Σ: known)

$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} (y - Z\beta)' \Sigma^{-1} (y - Z\beta)$ GLS: Generalized Least Squares

$(y - Z\beta)' \Sigma^{-1/2}\Sigma^{-1/2}(y - Z\beta) = \left[\Sigma^{-1/2}y - \Sigma^{-1/2}Z\beta\right]'\left[\Sigma^{-1/2}y - \Sigma^{-1/2}Z\beta\right] = \|\tilde{y} - \tilde{Z}\beta\|$

$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|\tilde{y} - \tilde{Z}\beta\| \implies \hat{\beta} = (\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\tilde{y} = (Z'\Sigma^{-1/2}\Sigma^{-1/2}Z)^{-1}Z'\Sigma^{-1/2}\Sigma^{-1/2}y = (Z'\Sigma^{-1}Z)^{-1}Z'\Sigma^{-1}y$

Note: $y = Z\beta + \varepsilon$ $\mathrm{cov}(\varepsilon) = \sigma^2\Sigma$
$\Sigma^{-1/2}y = \Sigma^{-1/2}Z\beta + \Sigma^{-1/2}\varepsilon$ → $\tilde{y} = \tilde{Z}\beta + \tilde{\varepsilon}$
($\tilde{y}$, $\tilde{Z}$, $\tilde{\varepsilon}$)
$\mathrm{Cov}(\tilde{\varepsilon}) = \sigma^2\Sigma^{-1/2}\Sigma\Sigma^{-1/2} = \sigma^2 I$

Note: If Σ is unknown:
- model Σ parametrically;
- iteratively estimate Σ, from the residuals, initializing $\Sigma_0 = I$
- Transform y or Z

EXAMPLES

1. Y_i is the mean of n_i obs indep. with same var σ^2
$\mathrm{Var}(Y_i) = \frac{\sigma^2}{n_i}$

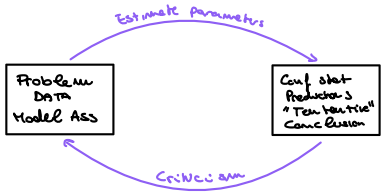
Σ known $\Sigma = \begin{bmatrix} 1/n_1 & & \\ & 1/n_2 & \\ & & \ddots \\ & & & 1/n_n \end{bmatrix}$ → $\Sigma^{-1} = \begin{bmatrix} n_1 & & \\ & n_2 & \\ & & \ddots \\ & & & n_n \end{bmatrix}$ WLS

2. y_i sum of n_i obs indep. with same var σ^2

$\mathrm{Var}(y_i) = n_i \sigma^2$

Σ known $\quad \Sigma = \begin{bmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_n \end{bmatrix} \longrightarrow \Sigma^{-1} = \begin{bmatrix} 1/n_1 & & & \\ & 1/n_2 & & \\ & & \ddots & \\ & & & 1/n_n \end{bmatrix}$ WLS

Box



Diagnostic for LM

- Residuals analysis, outliers, heteroscedasticity, normality, autocorrelation, ...
- influential cases
- collinearity
- others ..

Model

$Y = Z\beta + \varepsilon$

ε : $\quad \mathbb{E}[\varepsilon] = 0$

$\mathrm{Cov}(\varepsilon) = \sigma^2 I$

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

dist on $\mathbb{R}^n$

Fitted model

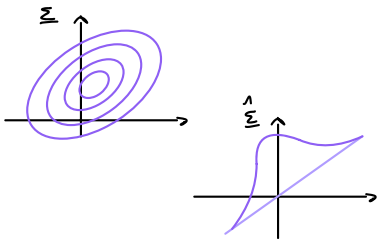
$\hat{Y} = Z\hat{\beta} \Longrightarrow Y = Z\hat{\beta} + \hat{\varepsilon}$

$\hat{\varepsilon}$: $\quad \mathbb{E}[\hat{\varepsilon}] = 0$

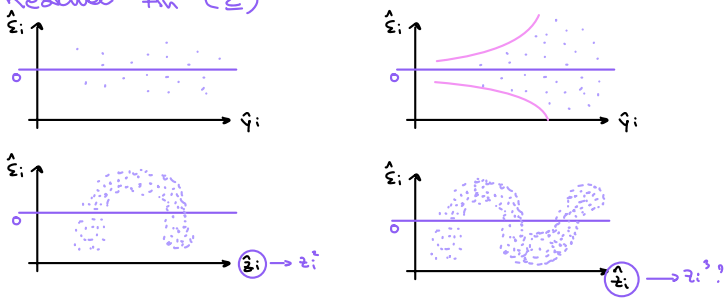
$\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2 (I - H)$

$\hat{\varepsilon} \sim N(0, \sigma^2 (I - H))$

dist on $\mathcal{L}^\perp(Z)$



Residual An ($\hat{\varepsilon}$)



But: $\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2 (I - H)$

$\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2 (1 - h_{ii})$

h_{ii} diag of H

$H = Z (Z'Z)^{-1} Z'$

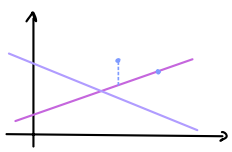
$\hat{\varepsilon}_i \longrightarrow \dfrac{\hat{\varepsilon}_i}{s\sqrt{1 - h_{ii}}}$ Standardized residuals

If assume $\varepsilon \sim N \longrightarrow$ Check Gauss of $\hat{\varepsilon}$: QQ plots, testing, ..

h_{ii} leverage

1. $0 \le h_{ii} \le 1$ because $H = H'$ and $H \cdot H = H$

$h_{ii} \uparrow 1 \Longrightarrow \left.\begin{array}{l} \mathrm{Var}(\varepsilon_i) \downarrow 0 \\ \mathbb{E}[\varepsilon_i] = 0 \end{array}\right\} \Longrightarrow \hat{\varepsilon}_i = 0$



But $\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2 (I - H)$

$\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2 (1 - h_{ii})$

h_{ii} diag H: $H = Z(Z'Z)^{-1}Z'$

$\hat{\varepsilon}_i \longrightarrow \dfrac{\hat{\varepsilon}_i}{s(1 - h_{ii})}$ standardized residuals

Influential case

$X = \begin{bmatrix} x_1' & y_1 \\ x_2' & y_2 \\ \vdots & \vdots \\ x_n' & y_n \end{bmatrix} \Longrightarrow Z = \begin{bmatrix} 1 & z_{11} \cdots z_{1r} \\ \vdots & \\ 1 & z_{n1} \cdots z_{nr} \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

$Z_{-i} = \begin{bmatrix} 1 & z_{21} & \cdots & z_{2r} \\ \vdots & & & \\ 1 & z_{i-1,1} & \cdots & z_{i-1,r} \\ 1 & z_{i+1,1} & \cdots & z_{i+1,r} \\ \vdots & & & \\ 1 & z_{n1} & \cdots & z_{nr} \end{bmatrix}$ Holdout unit i

$y_{-i} = Z_{-i}\beta_{-i} + \varepsilon_{-i} \longrightarrow \hat{\beta}_{-i}$

if $\hat{\beta}$ and $\hat{\beta}_{-i}$ are very different $\Longrightarrow$ unit i is influential

$D_i = \dfrac{(\hat{\beta} - \hat{\beta}_{-i})(Z'Z)(\hat{\beta} - \hat{\beta}_{-i})}{S^2(r+1)}$ Cook's distance for case i

D_i large $\Longrightarrow$ influential cases

- Compare with a quantile of $F(r+1, n-(r+1))$
- $D_i > 1$

In fact $D_i = \left(\dfrac{\hat{\varepsilon}_i}{S\sqrt{1-h_{ii}}}\right)^2 \cdot \dfrac{h_{ii}}{1-h_{ii}} \cdot \dfrac{1}{r+1}$

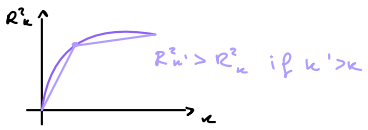
↑ std res. ↑ with h_{ii}

Model selection

$y = \beta_0 + \beta_1 z_1 + \dots + \beta_r z_r + \varepsilon$ "Can I remove some of them?"

2^r possible models with r regressors

- For $k = 0, \dots, r$ fit all $\binom{r}{k}$ with k regressors
- Choose the best one (e.g. the one with max R^2)



Collinearity and variable selection

$Y = Z\beta + \varepsilon$ Z design matrix, $n \times (r+1)$

ols: $\hat{\beta} = (Z'Z)^{-1}Z'Y$ if Z is full rank

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

If 2 or more regressors are close to be linearly dependent $\Longrightarrow$ $\mathrm{Cor}(\hat{\beta})$ will explode

$\mathrm{Var}(\hat{\beta}_j) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)^2} \cdot \dfrac{1}{1-R^2_j} = \mathrm{diag}_j\left(\sigma^2(Z'Z)^{-1}\right)$

R^2_j is coeff. of det. when Z_j is regressed on $z_1, z_2, \dots, \cancel{z_j}, \dots, z_r$

If $\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)^2 \uparrow$ $\Longrightarrow$ $\mathrm{Var}(\hat{\beta}_j) \downarrow$

If $R^2_j \uparrow 1$ $\Longrightarrow$ $\mathrm{Var}(\hat{\beta}_j) \uparrow$ Collinearity

Red alarm if: $\dfrac{1}{1-R^2_j} = \mathrm{VIF}$ = Variance inflation Factor > 5

Fitted model: $y_0 = z_0'\hat{\beta}$ $z_0 = (1\ z_{01}\ \dots\ z_{0r})'$

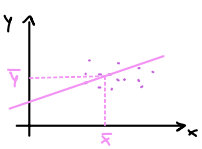
$z_0 = (1\ \bar{z}_1\ \bar{z}_2\ \dots\ \bar{z}_r)' = \dfrac{Z'1}{1'\cdot 1}$

$y_0 = \dfrac{1'Z}{1'1}(Z'Z)^{-1}Z'y = \dfrac{1'H1}{1'1} = \dfrac{1'\hat{y}}{1'1} = \bar{\hat{y}} = \bar{y}$ $\quad (1'H) = (H1)' = 1'$

That is: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{z}_1 + \hat{\beta}_2\bar{z}_2 + \dots + \hat{\beta}_r\bar{z}_r$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{z}_1 - \hat{\beta}_2\bar{z}_2 - \dots - \hat{\beta}_r\bar{z}_r$

$$\boxed{y_0 - \bar{y} = \hat{\beta}_1(z_{01} - \bar{z}_1) + \dots + \hat{\beta}_r(z_{0r} - \bar{z}_r)}$$



ols fitted models pass through the barycenter $(\bar{z}_1 \dots \bar{z}_r\ \bar{y})$

Centering: $Y \longrightarrow Y^* = Y - \bar{y}\cdot 1 = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$

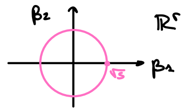
$Z \longrightarrow Z^* = \begin{bmatrix} z_{11} - \bar{z}_1 & z_{12} - \bar{z}_2 & \dots & z_{1r} - \bar{z}_r \\ & & & \\ z_{n1} - \bar{z}_1 & \dots & & z_{nr} - \bar{z}_r \end{bmatrix}$ $n \times r$

ols: $\begin{cases} \hat{\beta}^* = \underset{\beta \in \mathbb{R}^r}{\mathrm{argmin}} \| Y^* - Z^*\beta \|^2 \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}^*_1\bar{z}_1 - \dots - \hat{\beta}^*_r\bar{z}_r \\ \hat{\beta}_i = \hat{\beta}^*_i \quad i = 1, \dots, r \end{cases}$

Caution: From now on Y and Z $(n \times r)$ are centered

Ridge regression

$$(\star)\quad\begin{cases}\underset{\beta\in\mathbb{R}^r}{\operatorname{argmin}} \|y - Z\beta\|^2 \\ \|\beta\|^2 \le s\end{cases}$$

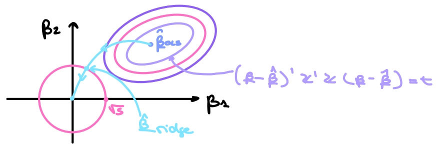


$$\|y - Z\beta\|^2 = \|y - \hat{y} - Z(\beta - \hat{\beta}_{OLS})\|^2 = Z\hat{\beta}_{OLS} = \|\hat{\varepsilon} - Z(\beta - \hat{\beta}_{OLS})\|^2 = \|\hat{\varepsilon}\|^2 + \|Z(\beta - \hat{\beta}_{OLS})\|^2$$

($\hat{\varepsilon} \in \mathcal{L}^{\perp}(Z)$, $\in \mathcal{L}(Z)$)

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

$$(\star)\quad\begin{cases}\underset{\beta\in\mathbb{R}^r}{\operatorname{argmin}} \|Z(\beta - \hat{\beta}_{OLS})\|^2 \\ \|\beta\|^2 \le s\end{cases} \qquad \mathbb{R}^r$$



$$\begin{cases}\underset{\beta}{\operatorname{argmin}} (\beta - \hat{\beta}_{OLS})'Z'Z(\beta - \hat{\beta}_{OLS}) \\ \|\beta\|^2 \le s\end{cases}$$

Consider the Lagrangian: $\underset{\beta\in\mathbb{R}^r}{\operatorname{argmin}} \|y - Z\beta\|^2 + \lambda\|\beta\|^2$ ($\lambda = \lambda(s)$) $\Longrightarrow$ $\hat{\beta}_{ridge} = (Z'Z + \lambda I)^{-1} Z'y$

Obs:
1. $\hat{\beta}_{ridge}$ is biased
2. For any regression problem $\Longrightarrow$ $\exists\, \lambda^*$ such $\mathbb{E}[\|\hat{\beta}_{ridge}\|^2] \le \mathbb{E}[\|\hat{\beta}_{OLS} - \beta\|^2]$
3. Find the λ^* by cross validation

If the problem is collinearity a different sol'n could have been $\Longrightarrow$ orthog. regressor

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{1r} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{nr}\end{bmatrix} \xrightarrow{PCA} Z^* = \begin{bmatrix} s_{11} & \cdots & s_{1r} \\ \vdots & & \vdots \\ s_{n1} & \cdots & s_{nr}\end{bmatrix}$$

(columns $z_1 \dots z_r$; $PC_1 \dots PC_r$, keep k)

$$Z^* = \begin{bmatrix} s_{11} & \cdots & s_{1k} \\ \vdots & & \vdots \\ s_{n1} & \cdots & s_{nk}\end{bmatrix} \quad (PC_1 \dots PC_k)$$

$y = Z^*\beta + \varepsilon \qquad \beta \in \mathbb{R}^k$

$\hat{\beta}_{PCA} = \underset{\beta\in\mathbb{R}^k}{\operatorname{argmin}} \|y - Z^*\beta\|^2$

PCA-regression fitted model: $\hat{\beta}_{PCA}$ (forget about the PCA etichetta)

$$y = \hat{\beta}_1 PC1 + \hat{\beta}_2 PC2 + \dots + \hat{\beta}_k PCk$$

$$\begin{cases} PC1 = e_{11}z_1 + \dots + e_{r1}z_r \\ \vdots \\ PCk = e_{1k}z_1 + \dots + e_{rk}z_r\end{cases}$$

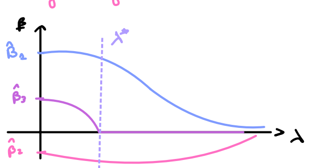
$$\hat{y} = z_1(\hat{\beta}_1 e_{11} + \hat{\beta}_2 e_{12} + \dots + \hat{\beta}_k e_{1k}) + z_2(\hat{\beta}_1 e_{21} + \hat{\beta}_2 e_{22} + \dots + \hat{\beta}_k e_{2k}) + \dots + z_r(\hat{\beta}_1 e_{r1} + \hat{\beta}_2 e_{r2} + \dots + \hat{\beta}_k e_{rk})$$

(γ_1, γ_2, γ_r)

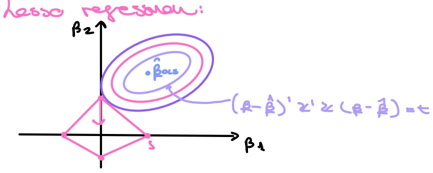
$$= z_1\gamma_1 + z_2\gamma_2 + \dots + z_r\gamma_r$$

Obs: PCA-reg., ridge reg., don't have spars sol'n in terms of $z_1 \dots z_r$

Ridge regression:



Lasso regression:

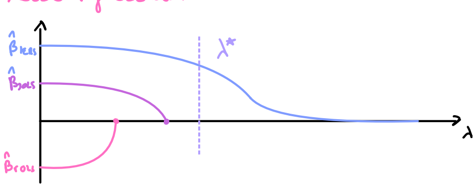


$$\begin{cases}\underset{\beta\in\mathbb{R}^r}{\operatorname{argmin}} (\beta - \hat{\beta}_{OLS})'Z'Z(\beta - \hat{\beta}_{OLS}) \\ \|\beta\|_1 \le s \qquad (\Sigma|\beta_i| < s)\end{cases}$$

Lagrangian: $\underset{\beta}{\operatorname{argmin}} \|y - Z\beta\|^2 + \lambda\Sigma|\beta_i|$

Problem: Lagrangian $\underset{\beta}{\operatorname{argmin}} \|y - Z\beta\|^2 + \lambda\sum_i|\beta_i|$ not identically solvable

Lasso regression:

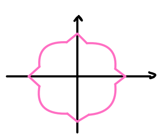


only one to select

Now plenty possibilities $\longrightarrow$ boundary even more spikey

A compromise: $\underset{\beta}{\operatorname{argmin}} \|y - Z\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2$

Two variables:



ELASTIC NET

ENSAMBLE METHODS

$\mathcal{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ $\quad x_i \in \mathbb{R}^p$, $y_i \in \begin{cases} \mathbb{R} \\ \{\text{labels}\} \end{cases}$ (regressor)

$M_{\mathcal{X}} : \mathbb{R}^p \longrightarrow \begin{cases} \mathbb{R} \\ \{\text{labels}\} \end{cases}$, given a new vector of features $x \in \mathbb{R}^p$ we predict $M_{\mathcal{X}}(x)$

Learning $M_{\mathcal{X}}$ means dealing with bias-variance trade-off

How to reduce the variance of the prediction without losing too much in terms of bias?

We must separate concepts of measurement y and true measure μ.

$y = \mu + \varepsilon$, $\quad \varepsilon : \mathbb{E}[\varepsilon] = 0$, $\quad \text{Var}(\varepsilon) = \sigma^2$

Take one measurement: $y_1 \leftarrow y_1 = \mu + \varepsilon_1$ with $\mathbb{E}[y_1] = \mu$, $\text{var}(y_1) = \varepsilon_1$

To reduce variability we can take more meas., we take B indep. meas.

$y_1, y_2, \dots, y_B$ (realization of)

$Y_1, Y_2, \dots, Y_B$

$Y_i = \mu + \varepsilon_i \quad i = 1, \dots, B$, $\quad \varepsilon_i$ iid : $\mathbb{E}[\varepsilon_i] = 0$, $\quad \text{Var}(\varepsilon_i) = \sigma^2$

Take the average: $\bar{y} = \frac{1}{B} \sum_{i=1}^{B} y_i$ realization of $\bar{Y} = \frac{1}{B} \sum_{i=1}^{B} Y_i$

with $\mathbb{E}[\bar{Y}] = \mu$ but now $\text{Var}(\bar{Y}) = \frac{1}{B} \sigma^2$ (reduced variability)

This was the basic idea to apply to our argument

$\mathcal{X}$ n units $\rightsquigarrow M_{\mathcal{X}}$ and I generate $\mathcal{X}_1 \rightsquigarrow M_{\mathcal{X}_1}$, $\mathcal{X}_2 \rightsquigarrow M_{\mathcal{X}_2}$, $\dots$, $\mathcal{X}_B \rightsquigarrow M_{\mathcal{X}_B}$

and we aggregate them $M = \frac{1}{B} \sum_{i=1}^{B} M_{\mathcal{X}_i}$

Why don't I use a big data $\mathcal{X}$ $B \times n$ units to learning? Not practical

Bootstrap:

$\mathcal{X}$ n units training set, want to generate B data sets, pseudocode:

For $i = 1, \dots, n$

sample the unit (row) out of $\mathcal{X}$

$\rightarrow (x_i^*, y_i^*)$

sample with replacement

$\rightarrow \mathcal{X}^*$ obtained by resampling with replacement from $\mathcal{X}$.

Repeat it for all B sets, from each of them we can extract a model and aggregate them

$\longrightarrow$ BAGGING = BOOTSTRAP + AGGREGATION

check some prob:

Unit $u \in \mathcal{X}$: $\mathbb{P}[u \notin \mathcal{X}^*] = (1 - 1/n)^n \xrightarrow{n \to \infty} e^{-1} \simeq 1/3$

So 2/3 of sample in $\mathcal{X}$ can appear in $\mathcal{X}^* \Longrightarrow$ not possible to say independent

$\text{Var}(M) > \frac{\sigma^2}{B}$ if $\sigma^2 = \text{Var}(M_{\mathcal{X}_i})$ but in any case $\sigma^2 > \text{Var}(M)$