ÉCOLE
POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

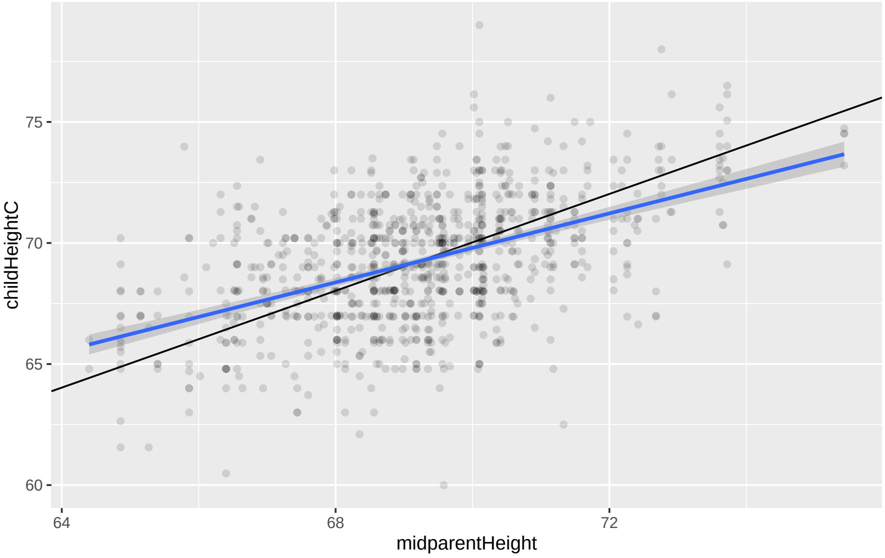
E. Le Pennec

Fall 2017

THE ART OF LINEAR REGRESSION

Contents

Regression to Mediocrity



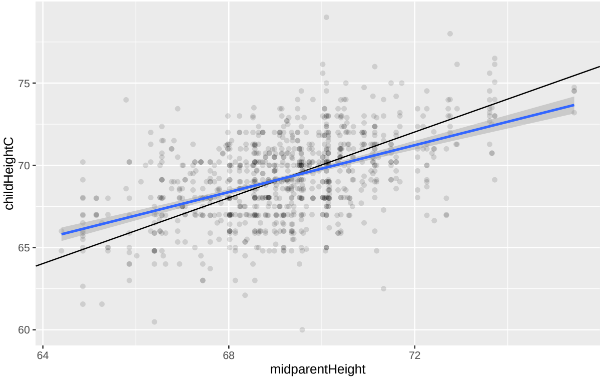
F. Galton (1877)

- Size of 934 children with respect to the one the parents.
- Scatterplot of 934 couples (X, Y):
  - X axis: size of the parents (corrected for the gender...)
  - Y axis: size of the children (corrected for the gender...)

- Asymmetric behavior of $(X, Y - X)$!
- Comp. between (X, Y) and a slope 1 line.and the *best* line fit.
- How to find the *best* fit? Are those two lines (statistically) different?

- Apparent *regression* to the *mediocrity* phenomena!

Regression to Mediocrity



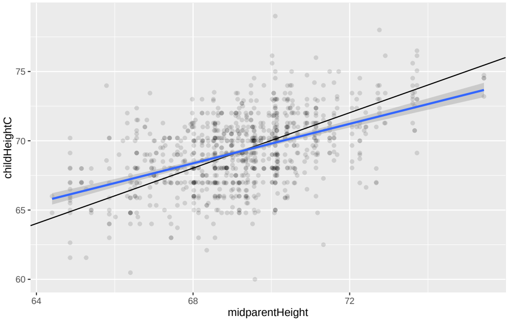
- Tall fathers tend to have tall sons but the sons are not, on average, as tall as the fathers. Also, short fathers have short sons who are not, on average, as short as their fathers.

- Galton called this effect *regression to the mediocrity.*

- In other words, the son's height tends to be *closer to the overall mean* height than the father's height was.

- Nowadays, the term *regression* is used more generally in statistics to refer to the process of fitting a line to data.

Regression



Setting

- Some samples (X_i, Y_i) of a numerical variable Y related to another numerical variable X.

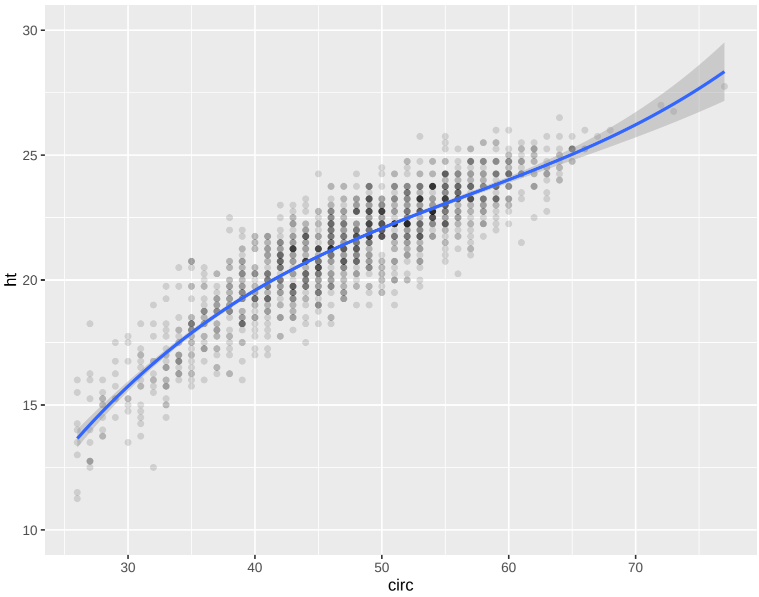
Goal

- Exhibit a link between Y and X of type $Y \simeq f(X)$
  - to predict Y with respect to X,
  - to understand the influence of X

Goal of the Statistician

- Propose a statistical model,
- Propose numerical methods to estimates the unknown parameters,
- Assess parameter uncertainty, assess model assumptions, choose a proper model !

Eucalyptus



Dataset - P.A. Cornillon

- Real dataset of 1429 eucalyptus obtained by P.A. Cornillon:
  - X: circumference / Y: height

- Can we predict the height from the circumference?
  - by a *line*? by a more complex formula?
  - by also taking account of the block and the clone type?

Boston Housing

By Jameslwoodward - Own work, CC BY-SA 3.0 https://commons.wikimedia.org/w/index.php?curid=8371157

Dataset - Harrison and Rubinfeld (1978)

- Contains 20 socio-economical variables for all the 506 tracts of Boston.
- Classical task: predict the median value of houses (Y) from the other information (town, position, crime rate, presence of Charles river, proportion of old houses, nox pollution...) ($\underline{X}$)
- Two simultaneous goal:
  - Obtain a good prediction,
  - Understand the influence of each variable

Birth Weight

By Drpoulette from Mexico City, Mexico - Flickr, CC BY 2.0 https://commons.wikimedia.org/w/index.php?curid=1372104

Dataset - Baystate Medical Center, Springfield, Mass (1986)

- Study of 189 child:
  - Y: birth weight (or indicator of weight less than 2.5 kg)
  - $\underline{X}$: age, mother's weight, ethnic category, smoking status, previous premature labors, hypertension, uterine irritability, visits during the first trimester
- Lots of categorical variables.
- Two simultaneous goal:
  - Obtain a good prediction,
  - Understand the influence of each variable

Linear Regression

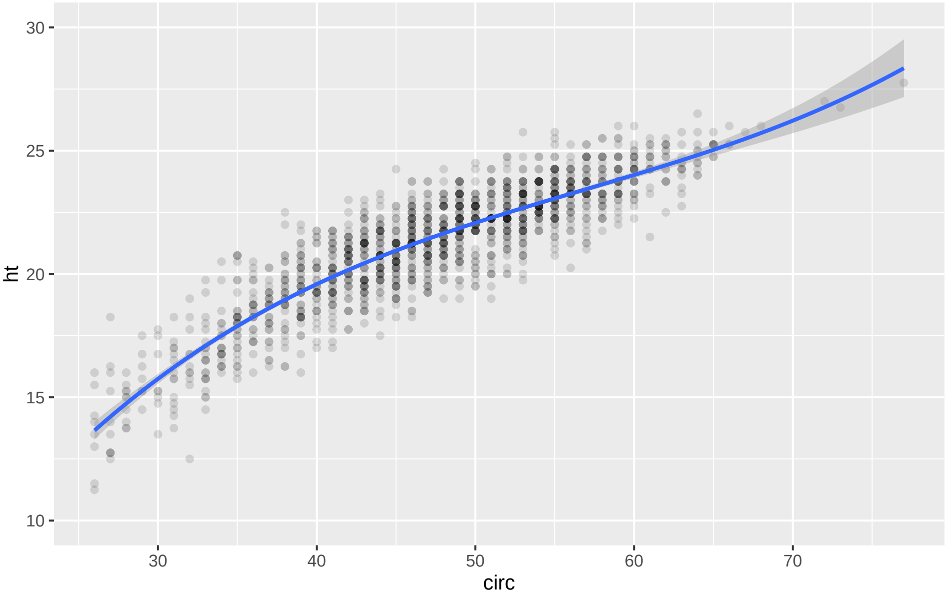
Linear Regression as a Model

- Strong assumption that the *relationship between $\underline{X}$ and Y is linear* (possible after a suitable transformation).
- *Formalization* by a model specifying how the systematic and random components come together to produce our data.

All models are wrong but some are useful!

- A model is simply a *set of assumptions* about how the world works.
- Of course, *all models are wrong*!
- But they can help us in *understanding* our data, and (if judiciously selected) may serve as *useful approximations* of the truth.

Theory and Practice



By the end of this course, you should be able to

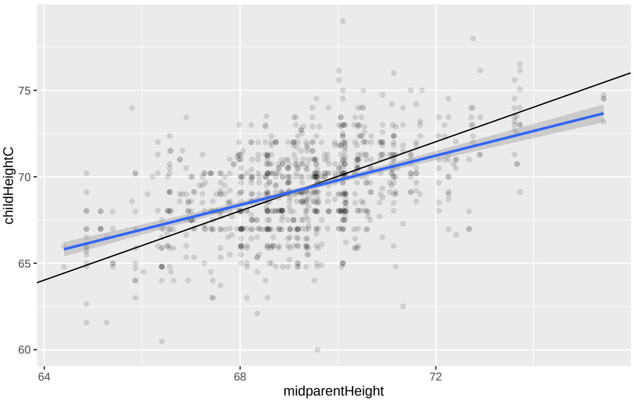
- Understand the statistical model used in (linear) regression.
- Know the properties of the estimated model.
- Interpret and comment the results.
- Use a computer program to perform the analysis.

Proposed R Framework

Working environment

- RStudio:
  - Well designed IDE
  - Platform independent
  - Rmarkdown!
- Packages:
  - `dplyr` (and friends) for data frame management
  - `ggplot2` (and friends) for graphics
  - `lm`, `glm` (and friends) for statistical model
- Not necessarily the best choice for everything or everyone....
- Choice of a systematic and coherent syntax...
- *R* seen as a *glue tool* more than a programming language...
- as *S* was designed by J. Chambers at Bell Labs in 1976!
- Importance of *literate programming and reproducible science*!
- All the datasets used are available in *R*! *Play with them!*

Galton



Lab with *R*

- You will learn:
  - How to access and view the dataset,
  - How to correct the height with regard to the gender,
  - How to visualize the dataset,
  - How to perform the linear regression,
  - How to read the summary of a linear regession in R,
  - How to use the estimation of R to compute a confidence interval.
- You will be able to say whether or not the data support the claims that there is a *regression to the mediocrity* phenomena.

Chapter 1

Linear Regression

Gaussian Variables

Gaussian Properties 1

Definition and Basic Properties

- *Def:* A r.v. $Z \in \mathbb{R}$ follows a standard Gaussian law $\mathrm{N}(\mu, \sigma^2)$ if and only if it has a density w. r. to the Lebesgue measure dt
$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$
- *Prop:* If $Z \sim \mathrm{N}(\mu, \sigma^2)$, $\mathbb{E}\left[Z\right] = \mu$ and $\mathbb{V}\mathrm{ar}\left[Z\right] = \sigma^2$.
- *Prop:* If $Z \sim \mathrm{N}(0, 1)$, $Z' = \mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$.
- *Prop:* If $Z \sim \mathrm{N}(\mu, \sigma^2)$, $Z' = (Z - \mu)/\sigma \sim \mathrm{N}(0, 1)$.

Characteristic function and Gaussian Law Characterization

- *Prop:* $Z \sim \mathrm{N}(\mu, \sigma^2) \Rightarrow \forall c \in \mathbb{C}, \quad \mathbb{E}\left[e^{cZ}\right] = e^{c\mu + c^2\frac{\sigma^2}{2}}$
- *Prop:* $Z \sim \mathrm{N}(\mu, \sigma^2) \Leftrightarrow \forall \omega \in \mathbb{R}, \quad \mathbb{E}\left[e^{i\omega Z}\right] = e^{i\omega\mu - \frac{\sigma^2\omega^2}{2}}$

Sum of Independent Gaussians

- *Prop:* If $Z_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ are independent
$$\alpha_0 + \sum_{i=1}^{n} \alpha_i Z_i \sim \mathrm{N}\left(\alpha_0 + \sum_{i=1}^{,} \alpha_i \mu_i, \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2\right)$$

Gaussian Properties 2

- If $Z \sim \mathrm{N}(\mu, \sigma^2)$, for any $a \in \mathbb{R}_+$
$$\mathbb{P}\left(\sigma^{-1}(Z - \mu) \in [-a, a]\right) = \Phi(a) - \Phi(-a)$$
where Φ is the cumulative distribution function of a standard normal distribution [note that $\sigma^{-1}(Z - \mu) \sim \mathrm{N}(0, 1)$]

- Because the Gaussian distribution is symmetric, for any $a > 0$,

$$1 - \Phi(a) = \Phi(-a) \Longrightarrow \Phi(a) - \Phi(-a) = 1 - 2(1 - \Phi(a))$$

- Therefore:

$$\begin{aligned}\mathbb{P}\left(Z \in [\mu - a\sigma, \mu + a\sigma]\right) &= \mathbb{P}\left(\sigma^{-1}(Z - \mu) \in [-a, a]\right) \\ &= 1 - 2(1 - \Phi(a))\end{aligned}$$

Gaussian tails

- $\forall t > 0$,

$$\frac{t^2}{1+t^2}\frac{e^{-t^2/2}}{t\sqrt{2\pi}} \leq \mathbb{P}\left(N > t\right) \leq \min\left(\frac{e^{-t^2/2}}{t\sqrt{2\pi}}, \frac{e^{-t^2/2}}{2}\right) \leq e^{-t^2/2}$$

  where $N \sim \mathrm{N}(0,1)$

Proof

- Chernoff bound:

$$\begin{aligned}\mathbb{P}\left(N > t\right) &= \min_{\lambda} \mathbb{P}\left(e^{\lambda N} > e^{\lambda t}\right) \\ &\leq \min_{\lambda} \frac{\mathbb{E}\left[e^{\lambda N}\right]}{e^{\lambda}t} \\ &\leq \min_{\lambda} e^{\lambda^2/2 - \lambda t} \\ &\leq e^{\min_{\lambda} \lambda^2/2 - \lambda t} = e^{-t^2/2}\end{aligned}$$

- Less accurate but generic technique.
- Upper bound:

$$\begin{aligned}\mathbb{P}\left(N > t\right) &= \int_t^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\ &= \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(t+u)^2}{2}} du \\ &= \int_0^{+\infty} e^{-t^2/2} e^{-2ut} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\ &= e^{-t^2/2} \int_0^{+\infty} e^{-ut} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\ &\leq e^{-t^2/2} \min\left(\int_0^{+\infty} e^{-ut} \frac{1}{\sqrt{2\pi}} du, \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du\right) \\ &\leq e^{-t^2/2} \min(\frac{1}{t\sqrt{2\pi}}, \frac{1}{2})\end{aligned}$$

- Lower bound:

$$\begin{aligned}\mathbb{P}\left(N > t\right) &= \int_t^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\ &\geq \int_t^{+\infty} \frac{t^2}{u^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du\end{aligned}$$

$$\begin{aligned}
&\geq \left[\frac{-t^2}{u}\frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}\right]_t^{+\infty} - \int_t^{+\infty}\left(-\frac{t^2}{u}\right)\left(-u\frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}\right)du\\
&\geq t^2\frac{e^{-\frac{t^2}{2}}}{t\sqrt{2\pi}} - t^2\mathbb{P}\left(N > t\right)
\end{aligned}$$

hence

$$\mathbb{P}\left(N > t\right) \geq \frac{t^2}{1+t^2}\frac{e^{-\frac{t^2}{2}}}{t\sqrt{2\pi}}$$

Gaussian Properties 3

Student Law

- *Def:* If Z and S^2 are two independent r.v. such that $Z \sim \mathrm{N}(0,1)$ and $S^2 \sim \chi^2(d)/d$, the law of Z/S is called the Student law of degree d denoted $\mathrm{T}(d)$.
- *Prop:* If Z and S^2 are two independent r.v. such that that $Z \sim \mathrm{N}(\mu,\sigma^2)$ and $S^2/\sigma^2 \sim \chi^2(d)/d$ then

$$\frac{Z-\mu}{S} \sim \mathrm{T}(d)$$

Student Law and Confidence Interval

- *Prop:* If Z and S^2 are two independent r.v. such that that $Z \sim \mathrm{N}(\mu,\sigma^2)$ and $S^2/\sigma^2 \sim \chi^2(d)/d$ then

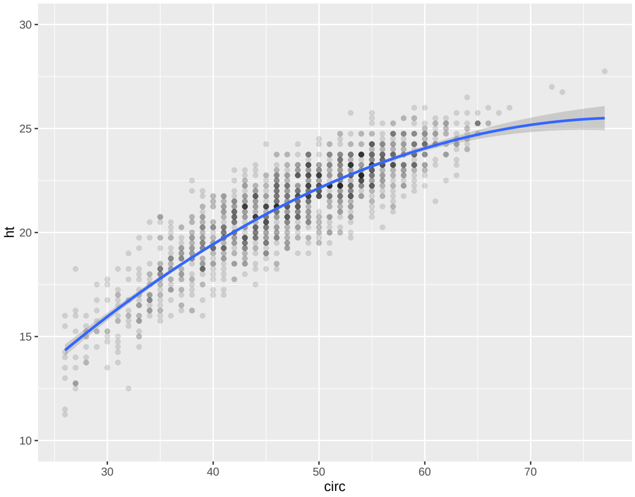
$$\begin{aligned}
&\mathbb{P}\left(\mu \in [Z - t_- S, Z + t_+ S]\right)\\
&\quad = \mathbb{P}\left(T \in [-t_+, t_-]\right) = 1 - \left(\mathbb{P}\left(T \geq t_+\right) + \mathbb{P}\left(T \leq t_-\right)\right)
\end{aligned}$$

with $T \sim \mathrm{T}(d)$.

1.1 Least Square Multivariate Regression

1.1.1 Least Square and Projection

Eucalyptus



Dataset - P.A. Cornillon

- Real dataset of 1429 eucalyptus obtained by P.A. Cornillon:
  - X: circumference / Y: height

- How can we predict the height from the circumference?
  - Univariate Linear Model: $Y \simeq \beta^{(2)}X + \beta^{(1)}$
  - Polynomial Model: $Y \simeq \beta^{(3)}X^2 + \beta^{(2)}X + \beta^{(1)}$
  - Transformed Polynomial Model: $Y \simeq \beta^{(3)}X + \beta^{(2)}\sqrt{X} + \beta^{(1)}$

Multivariate Regression

- *Setting:* link between $Y \in \mathbb{R}$ and $\underline{X} \in \mathbb{R}^p$.

Linear Model

$$Y \simeq \sum_{k=1}^{p} \beta^{(k)}\underline{X}^{(k)} = \langle \underline{X}, \beta \rangle = \underline{X}^t\beta$$

- Examples:
  - Univariate regression: $\underline{X} = (1, X)^t$:

$$Y \simeq \underline{X}^t\beta = \beta^{(1)}\underline{X}^{(1)} + \beta^{(2)}\underline{X}^{(2)} = \beta^{(1)} + \beta^{(2)}X$$

  - Polynomial regression: $\underline{X} = (1, X, X^2)^t$

$$Y \simeq \underline{X}^t\beta = \beta^{(1)}\underline{X}^{(1)} + \beta^{(2)}\underline{X}^{(2)} + \beta^{(3)}\underline{X}^{(3)} = \beta^{(1)} + \beta^{(2)}X + \beta^{(3)}X^2$$

  - Transformed Polynomial regression $\underline{X} = (1, \sqrt{X}, X)^t$

$$Y \simeq \underline{X}^t\beta = \beta^{(1)}\underline{X}^{(1)} + \beta^{(2)}\underline{X}^{(2)} + \beta^{(3)}\underline{X}^{(3)} = \beta^{(1)} + \beta^{(2)}\sqrt{X} + \beta^{(3)}X$$

- Models are linear in β (and not in the original variable X)!

Least Square

- *Linear model:* $Y \sim \underline{X}^t\beta = \sum_{j=1}^{p} \underline{X}^{(j)}\beta^{(j)}$.
- How to estimate the regression coefficients β from a sample $(\underline{X}_i, Y_i)_{1\leq i\leq n}$?

Least Square

- Choose $\hat{\beta}$ as a minimizer of the sum of squared error

$$L(\beta) = \sum_{i=1}^{n} |Y_i - \underline{X}_i^t \beta|^2$$

- Approach advocated by Legendre (1805) and Gauss (1809 or 1795)!
- *Prop:* L is a quadratic function .
- *Cor:* $\hat{\beta}$ exists (but is not necessarily unique...)

Proof

- *Coercivity*: if $\beta \notin \operatorname{span}(\underline{X}_1, \dots, \underline{X}_n)^\perp$, it exists i_0 such that $\underline{X}_{i_0}^t \beta \neq 0$ and

$$\begin{aligned} L(\lambda\beta) &= \sum_{i=1}^{n} |Y_i - \lambda \underline{X}_i^t \beta|^2 \\ &\geq |Y_{i_0} - \lambda \underline{X}_{i_0}^t \beta|^2 \\ &\geq \lambda^2 |\underline{X}_{i_0}^t \beta|^2 - 2\lambda Y_{i_0} \underline{X}_{i_0}^t \beta + |Y_{i_0}|^2 \xrightarrow[\lambda \to \pm\infty]{} +\infty \end{aligned}$$

- *Gradient*:

$$\begin{aligned} \nabla L(\beta) = \nabla \left(\sum_{i=1}^{n} |Y_i - \underline{X}_i^t \beta|^2 \right) &= \sum_{i=1}^{n} \nabla \left(|Y_i - \underline{X}_i^t \beta|^2 \right) \\ &= \sum_{i=1}^{n} 2 \underline{X}_i (\underline{X}_i^t \beta - Y_i) \\ &= 2 \left(\sum_{i=1}^{n} \underline{X}_i \underline{X}_i^t \right) \beta - 2 \sum_{i=1}^{n} Y_i \underline{X}_i \end{aligned}$$

- If $\sum_{i=1}^{n} \underline{X}_i \underline{X}_i^t$ is invertible, then the equation

$$\nabla L(\beta) = 0$$

  admits a unique solution, which is given by

$$\hat{\beta} = \left(\sum_{i=1}^{n} \underline{X}_i \underline{X}_i^t \right)^{-1} \sum_{i=1}^{n} \underline{X}_i Y_i \ .$$

Matrix Formulation

- *Idea:* Stack the observations:

$$\mathbb{Y}_{(n)} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \mathbb{X}_{(n)} = \begin{pmatrix} \underline{X}_1^t \\ \vdots \\ \underline{X}_n^t \end{pmatrix}$$

- Linear model:

$$\mathbb{Y}_{(n)} \simeq \mathbb{X}_{(n)} \beta \Leftrightarrow \begin{cases} Y_1 \simeq \underline{X}_1^t \beta \\ \vdots \\ Y_n \simeq \underline{X}_n^t \beta \end{cases}$$

Least Square and Minimizers

- Least Square:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\beta\|^2 \Leftrightarrow \hat{\beta} = \operatorname{argmin} \sum_{i=1}^{n} |Y_i - \underline{X}_i^t \beta|^2$$

- *Prop.:* If $\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}$ is invertible then

$$\hat{\beta} = \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \mathbb{X}_{(n)}^t \mathbb{Y}_{(n)}$$

Proof

- Least Square:

$$L(\beta) = \|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\beta\|^2$$

- Gradient:

$$\nabla L(\beta) = 2\mathbb{X}_{(n)}^t (\mathbb{X}_{(n)}\beta - \mathbb{Y}_{(n)})$$

- First order condition:

$$\begin{aligned} \nabla L(\beta) = \mathbb{0}_{(p)} &\Leftrightarrow 2\mathbb{X}_{(n)}^t (\mathbb{X}_{(n)}\beta - \mathbb{Y}_{(n)}) = \mathbb{0}_{(p)} \\ &\Leftrightarrow \mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\beta = \mathbb{X}_{(n)}^t \mathbb{Y}_{(n)} \end{aligned}$$

- First order condition if $\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}$ is invertible:

$$\nabla L(\beta) = \mathbb{0}_{(p)} \Leftrightarrow \beta = \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \mathbb{X}_{(n)}^t \mathbb{Y}_{(n)}$$

Invertibility and Identifiability

Proposition 1 (Invertibility, Injectivity and Column Freedom)**.** *$\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}$ invertible $\Leftrightarrow$ $\mathbb{X}_{(n)}$ full rank $\Leftrightarrow$ the columns of $\mathbb{X}_{(n)}$ are linearly independent*

Proposition 2 (Uniqueness of the Least Square)**.** *$\mathbb{X}_{(n)}$ full rank $\Leftrightarrow$ the minimizer of $\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\beta\|^2$ is unique.*

- If $\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}$ is not invertible, there are an infinity of solution to the Least Square problems!
- *Remark:* Do not use a (simple) linear regression when $\mathbb{X}_{(n)}$ is not full rank.

Proof

$$\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \text{ invertible} \Longleftrightarrow (\mathbb{X}_{(n)}\beta \neq \mathbb{0}_{(p)} \text{ for all } \beta \neq \mathbb{0}_{(p)})$$

Assume that there exists $\beta \neq \mathbb{0}_{(p)}$ such that $\mathbb{X}_{(n)}\beta = \mathbb{0}_{(n)}$. Then

$$\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\beta = \mathbb{0}_{(p)} \Longrightarrow \operatorname{Ker}(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}) \neq \{\mathbb{0}_{(p)}\}$$

Therefore

$$\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \quad \text{is not invertible}$$

Conversely, if $\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}$ is not invertible, then $\operatorname{Ker}(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}) \neq \{\mathbb{0}_{(p)}\}$ and there exists $\beta \neq \mathbb{0}_{(p)}$ such that $\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\beta = \mathbb{0}_{(n)}$. Then

$$\beta^t \mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\beta = 0 \Longrightarrow \|\mathbb{X}_{(n)}\beta\| = 0 \Longrightarrow \mathbb{X}_{(n)}\beta = \mathbb{0}_{(n)} .$$

$\mathbb{X}_{(n)}$ full rank $\Rightarrow$ $\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}$ invertible $\Rightarrow$ $\hat{\beta} = \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \mathbb{X}_{(n)}^t \mathbb{Y}_{(n)}$

- $\mathbb{X}_{(n)}$ is not full rank $\Rightarrow$ there exists $\delta \neq \mathbb{0}_{(p)}$, such that $\mathbb{X}_{(n)}\delta = \mathbb{0}_{(n)}$.
- thus If β_1 is a minimizer of $\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\beta\|^2$, then $\beta_1 + \delta$ is another minimizer as $\mathbb{X}_{(n)}(\beta_1 + \delta) = \mathbb{X}_{(n)}\beta_1$.
- The minimizer is thus not unique.

Projection

- Least square: $\hat{\beta} = \operatorname{argmin} \|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\beta\|^2$
- Least square and prediction: $\mathbb{X}_{(n)}\hat{\beta} = \operatorname{argmin} \|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\beta\|^2$

Proposition 3 (Projection)**.**

$$\mathbb{X}_{(n)}\hat{\beta} = \mathrm{Proj}_{\mathbb{X}_{(n)}} \mathbb{Y}_{(n)}$$

where $\mathrm{Proj}_{\mathbb{X}_{(n)}}$ *is the orthogonal projection on* $\mathrm{span}\{\mathbb{X}_{(n)}\}$*.*

$$\mathrm{Proj}_{\mathbb{X}_{(n)}} = \mathbb{X}_{(n)} \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \mathbb{X}_{(n)}^t$$

- Even if $\mathbb{X}_{(n)}$ is not full rank, the projection $\mathbb{X}_{(n)}\hat{\beta} = \mathrm{Proj}_{\mathbb{X}_{(n)}} \mathbb{Y}_{(n)}$ remains unique.
- But the estimator $\hat{\beta}$ is not unique!

Projection and Geometry

Proposition 4 (Geometrical Interpretation)**.**
- *predictor on the design* $\mathbb{X}_{(n)}\hat{\beta} = \mathrm{Proj}_{\mathbb{X}_{(n)}} \mathbb{Y}_{(n)}$
- *residual:* $\widehat{\epsilon}_{(n)} = \mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta} = \mathrm{Proj}_{\mathbb{X}_{(n)}^{\perp}} \mathbb{Y}_{(n)}$ *where* $\mathrm{Proj}_{\mathbb{X}_{(n)}^{\perp}}$ *is the orthogonal projection on the orthogonal space of* $\mathrm{span}\{\mathbb{X}_{(n)}\}$*.*
- *Pythagoras:*

$$\|\mathbb{Y}_{(n)}\|^2 = \|\mathbb{X}_{(n)}\hat{\beta}\|^2 + \|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2$$

SST, SSR and ESS

SST, SSR and ESS

- If $\mathrm{span}\{\mathbb{X}_{(n)}\}$ includes the constant vector $\mathbb{1}_{(n)} = (1, \ldots, 1)^t$,

$$\underbrace{\left\|\mathbb{Y}_{(n)} - \mathbb{1}_{(n)}\mathbb{E}_n\left[\mathbb{Y}_{(n)}\right]\right\|^2}_{\text{SST}}$$

$$= \underbrace{\left\|\mathbb{X}_{(n)}\hat{\beta} - \mathbb{1}_{(n)}\mathbb{E}_n\left[\mathbb{X}_{(n)}\hat{\beta}\right]\right\|^2}_{\text{ESS}} + \underbrace{\left\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\right\|^2}_{\text{SSR}}$$

- Acronyms:
  - SST: Sum of Squared Total,
  - ESS: Explained Sum of Squared,
  - SSR: Sum of Squared Residuals
- Pythagoras...

Proof

$$\begin{aligned}
&\|\mathbb{Y}_{(n)} - \mathbb{1}_{(n)}\mathbb{E}_n\left[\mathbb{Y}_{(n)}\right]\|^2 \\
&= \|\mathbb{Y}_{(n)} - \mathbf{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} + \mathbf{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} - \mathbb{1}_{(n)}\mathbb{E}_n\left[\mathbb{Y}_{(n)}\right]\|^2 \\
&= \|\mathbb{Y}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)}\|^2 + \|\mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} - \mathbb{1}_{(n)}\mathbb{E}_n\left[\mathbb{Y}_{(n)}\right]\|^2
\end{aligned}$$

since

$$\begin{aligned}
&\mathbb{Y}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} \perp \mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} \\
&\mathbb{Y}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} \perp \mathbb{1}_{(n)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_n\left[\mathbb{Y}_{(n)}\right] &= n^{-1}\mathbb{1}_{(n)}^t\mathbb{Y}_{(n)} \\
&= n^{-1}\mathbb{1}_{(n)}^t\{\mathbb{Y}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)}\} + \mathbb{1}_{(n)}^t\mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} \\
&= n^{-1}\mathbb{1}_{(n)}^t\mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} = \mathbb{E}_n\left[\mathbb{X}_{(n)}\hat{\beta}\right] .
\end{aligned}$$

since

$$\mathbb{Y}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} \perp \mathbb{1}_{(n)}$$

$\mathbf{R^2}$

Quantify the fraction of the *variance* explained by the linear prediction.

$\mathbf{R^2}$

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{SST}} = \frac{\left\|\mathbb{X}_{(n)}\hat{\beta} - \mathbb{E}_n\left[\mathbb{X}_{(n)}\hat{\beta}\right]\mathbb{1}_{(n)}\right\|^2}{\left\|\mathbb{Y}_{(n)} - \mathbb{E}_n\left[\mathbb{Y}_{(n)}\right]\mathbb{1}_{(n)}\right\|^2}$$

$$R^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\left\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\right\|^2}{\left\|\mathbb{Y}_{(n)} - \mathbb{E}_n\left[\mathbb{Y}_{(n)}\right]\mathbb{1}_{(n)}\right\|^2} \in [0,1]$$

Measure of the *quality* of the prediction: fraction of the variance is explained by the linear prediction.

1.1.2 First Order and Second Order Statistical Model

Statistical Model

- So far no statistical model!

First order probabilistic (statistical) model

Observations:

$$Y_i = \underline{X}_i^t\beta^\star + \epsilon_i$$

with $\{\epsilon_i\}_{i=1}^n$ a *noise* satisfying $\mathbb{E}\left[\epsilon_i|\underline{X}_i\right] = 0$

- *Equivalent formulation:* $\mathbb{E}\left[Y_i|\underline{X}_i\right] = \underline{X}_i^t\beta^\star$.

Matrix formulation

- $\mathbb{Y}_{(n)} = \mathbb{X}_{(n)}\beta^\star + \epsilon_{(n)}$ with $\epsilon_{(n)} = (\epsilon_1, \ldots, \epsilon_n)^t$.
- *Prop:* $\mathbb{E}\left[\epsilon_{(n)}\middle|\mathbb{X}_{(n)}\right] = \mathbb{0}_{(n)}$

Bias

Unbiased Estimate

- *Parameter:* If $\mathbb{E}\left[\epsilon_i | \underline{X}_i\right] = 0$ and $\mathbb{X}_{(n)}$ full-rank,
$$\mathbb{E}\left[\hat{\beta}\middle|\mathbb{X}_{(n)}\right] = \beta^\star$$
- *Prediction:* If $\mathbb{E}\left[\epsilon_i | \underline{X}_i\right] = 0$,
$$\mathbb{E}\left[\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\middle|\mathbb{X}_{(n)}\right] = \mathbb{0}_{(n)}$$
- *Prediction at a new point:* If $\mathbb{E}\left[\epsilon_i | \underline{X}_i\right] = 0$, $\mathbb{X}_{(n)}$ full-rank, $\widetilde{Y} = \underline{\widetilde{X}}^t\beta^\star + \widetilde{\epsilon}$ with $\mathbb{E}\left[\tilde{\epsilon}\middle|\underline{\widetilde{X}}\right] = 0$,
$$\mathbb{E}\left[\widetilde{Y} - f_{\hat{\beta}}(\underline{\widetilde{X}})\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right] = 0$$
with $f_{\hat{\beta}}(\underline{\widetilde{X}}) = \underline{\widetilde{X}}^t\hat{\beta}$.

Proof

- If A a *deterministic* matrix, $\mathbb{E}\left[AZ\right] = A\mathbb{E}\left[Z\right]$
- Projection:
$$\begin{aligned}\mathbb{X}_{(n)}\hat{\beta} &= \operatorname{Proj}_{\mathbb{X}_{(n)}} \mathbb{Y}_{(n)}\\ &= \operatorname{Proj}_{\mathbb{X}_{(n)}} \left(\mathbb{X}_{(n)}\beta^\star + \epsilon_{(n)}\right)\\ &= \mathbb{X}_{(n)}\beta^\star + \operatorname{Proj}_{\mathbb{X}_{(n)}} \epsilon_{(n)}\end{aligned}$$
- Expectation:
$$\begin{aligned}\mathbb{E}\left[\mathbb{X}_{(n)}\hat{\beta}\middle|\mathbb{X}_{(n)}\right] &= \mathbb{E}\left[\mathbb{X}_{(n)}\beta^\star + \operatorname{Proj}_{\mathbb{X}_{(n)}} \epsilon_{(n)}\middle|\mathbb{X}_{(n)}\right]\\ &= \mathbb{X}_{(n)}\beta^\star + \operatorname{Proj}_{\mathbb{X}_{(n)}} \mathbb{E}\left[\epsilon_{(n)}\middle|\mathbb{X}_{(n)}\right]\\ &= \mathbb{X}_{(n)}\beta^\star\end{aligned}$$
- Prediction:
$$\begin{aligned}\mathbb{E}\left[\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\middle|\mathbb{X}_{(n)}\right] &= \mathbb{E}\left[\mathbb{Y}_{(n)}\middle|\mathbb{X}_{(n)}\right] - \mathbb{E}\left[\mathbb{X}_{(n)}\hat{\beta}\middle|\mathbb{X}_{(n)}\right]\\ &= \mathbb{X}_{(n)}\beta^\star - \mathbb{X}_{(n)}\beta^\star\\ &= \mathbb{0}_{(n)}\end{aligned}$$
- If $\mathbb{X}_{(n)}$ is full-rank, $\beta = \left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t\right)\mathbb{X}_{(n)}\beta$
$$\begin{aligned}\mathbb{E}\left[\hat{\beta}\middle|\mathbb{X}_{(n)}\right] &= \mathbb{E}\left[\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t\right)\mathbb{X}_{(n)}\hat{\beta}\middle|\mathbb{X}_{(n)}\right]\\ &= \left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t\right)\mathbb{E}\left[\mathbb{X}_{(n)}\hat{\beta}\middle|\mathbb{X}_{(n)}\right]\\ &= \left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t\right)\mathbb{X}_{(n)}\beta^\star\\ &= \beta^\star\end{aligned}$$
- Now,
$$\begin{aligned}\mathbb{E}\left[\widetilde{Y} - \underline{\widetilde{X}}^t\hat{\beta}\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right] &= \mathbb{E}\left[\widetilde{Y}\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right] - \mathbb{E}\left[\underline{\widetilde{X}}^t\hat{\beta}\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right]\\ &= \underline{\widetilde{X}}^t\beta^\star - \underline{\widetilde{X}}^t\beta^\star\\ &= 0\end{aligned}$$

Homoscedastic Statistical Model

So far minimalist statistical model!

Second order statistical model

Observations:

$$Y_i = \underline{X}_i{}^t \beta^\star + \epsilon_i$$

with

- $\mathbb{E}\left[\epsilon_i \middle| \underline{X}_i\right] = 0$,
- $\mathbb{V}\mathrm{ar}\left[\epsilon_i \middle| \underline{X}_i\right] = \sigma_\star^2$ (Homoscedasticity),
- $\mathbb{C}\mathrm{ov}\left[\epsilon_i, \epsilon_j \middle| \underline{X}_i, \underline{X}_j\right] = 0$ if $i \neq j$ (Decorrelation).

- *Equiv. reformulation:* $\mathbb{E}\left[Y_i \middle| \underline{X}_i\right] = \underline{X}_i{}^t \beta^\star$ and $\mathbb{C}\mathrm{ov}\left[Y_i, Y_j \middle| \underline{X}_i, \underline{X}_j\right] = \sigma_\star^2 \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker symbol.

Matrix formulation

- $\mathbb{Y}_{(n)} = \mathbb{X}_{(n)} \beta^\star + \epsilon_{(n)}$ with $\epsilon_{(n)} = (\epsilon_1, \dots, \epsilon_n)^t$.
- *Prop:* $\mathbb{E}\left[\epsilon_{(n)} \middle| \mathbb{X}_{(n)}\right] = \mathbb{0}_{(n)}$ and $\mathbb{C}\mathrm{ov}\left[\epsilon_{(n)} \middle| \mathbb{X}_{(n)}\right] = \sigma_\star^2 \mathrm{Id}_{(n)}$

- *Rk:* $\mathbb{E}\left[\epsilon_{(n)} \middle| \mathbb{X}_{(n)}\right] = \mathbb{0}_{(n)}$ and $\mathbb{C}\mathrm{ov}\left[\epsilon_{(n)} \middle| \mathbb{X}_{(n)}\right] = \sigma_\star^2 \mathrm{Id}_{(n)}$ weaker than $\mathbb{E}\left[Y_i \middle| \underline{X}_i\right] = \underline{X}_i{}^t \beta^\star$ and $\mathbb{C}\mathrm{ov}\left[Y_i, Y_j \middle| \underline{X}_i, \underline{X}_j\right] = \sigma_\star^2 \delta_{i=j}$

Estimate Covariance

- Residual: $\widehat{\epsilon}_{(n)} = \mathbb{Y}_{(n)} - \mathbb{X}_{(n)} \hat{\beta}$

Proposition 5 (Covariance)**.**

$$\begin{aligned}
\mathbb{E}\left[\begin{pmatrix} \hat{\beta} \\ \widehat{\epsilon}_{(n)} \end{pmatrix} \middle| \mathbb{X}_{(n)}\right] &= \begin{pmatrix} \beta^\star \\ \mathbb{0}_{(n)} \end{pmatrix} \\
\mathbb{C}\mathrm{ov}\left[\begin{pmatrix} \hat{\beta} \\ \widehat{\epsilon}_{(n)} \end{pmatrix} \middle| \mathbb{X}_{(n)}\right] &= \sigma_\star^2 \begin{pmatrix} (\mathbb{X}_{(n)}^t \mathbb{X}_{(n)})^{-1} & \mathbb{0}_{(p \times n)} \\ \mathbb{0}_{(n \times p)} & \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\begin{pmatrix} \mathbb{X}_{(n)} \hat{\beta} \\ \widehat{\epsilon}_{(n)} \end{pmatrix} \middle| \mathbb{X}_{(n)}\right] &= \begin{pmatrix} \mathbb{X}_{(n)} \beta^\star \\ \mathbb{0}_{(n)} \end{pmatrix} \\
\mathbb{C}\mathrm{ov}\left[\begin{pmatrix} \mathbb{X}_{(n)} \hat{\beta} \\ \widehat{\epsilon}_{(n)} \end{pmatrix} \middle| \mathbb{X}_{(n)}\right] &= \sigma_\star^2 \begin{pmatrix} \mathrm{Proj}_{\mathbb{X}_{(n)}} & \mathbb{0}_{(n \times n)} \\ \mathbb{0}_{(n \times n)} & \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \end{pmatrix}
\end{aligned}$$

Proof

- If A is a *deterministic* matrix, then $\mathbb{C}\mathrm{ov}\left[AZ\right] = A\,\mathbb{C}\mathrm{ov}\left[Z\right]A^t$
- Note that

$$\begin{pmatrix} \mathbb{X}_{(n)}\hat{\beta} \\ \hat{\beta} \\ \widehat{\epsilon}_{(n)} \end{pmatrix} = \begin{pmatrix} \mathrm{Proj}_{\mathbb{X}_{(n)}} \\ \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t \\ \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \end{pmatrix} \mathbb{Y}_{(n)}$$

- Expectation:

$$\mathbb{E}\left[\begin{pmatrix} \mathbb{X}_{(n)}\hat{\beta} \\ \hat{\beta} \\ \widehat{\epsilon}_{(n)} \end{pmatrix}\middle|\mathbb{X}_{(n)}\right] = \begin{pmatrix} \mathrm{Proj}_{\mathbb{X}_{(n)}} \\ \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t \\ \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \end{pmatrix} \mathbb{E}\left[\mathbb{Y}_{(n)}\middle|\mathbb{X}_{(n)}\right]$$
$$= \begin{pmatrix} \mathrm{Proj}_{\mathbb{X}_{(n)}} \\ \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t \\ \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \end{pmatrix} \mathbb{X}_{(n)}\beta^\star = \begin{pmatrix} \mathbb{X}_{(n)}\beta^\star \\ \beta^\star \\ \mathbb{0}_{(n)} \end{pmatrix}$$

- Covariance:

$$\mathbb{C}\mathrm{ov}\left[\begin{pmatrix} \mathbb{X}_{(n)}\hat{\beta} \\ \hat{\beta} \\ \widehat{\epsilon}_{(n)} \end{pmatrix}\middle|\mathbb{X}_{(n)}\right]$$
$$= \begin{pmatrix} \mathrm{Proj}_{\mathbb{X}_{(n)}} \\ \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t \\ \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \end{pmatrix} \mathbb{C}\mathrm{ov}\left[\mathbb{Y}_{(n)}\middle|\mathbb{X}_{(n)}\right] \begin{pmatrix} \mathrm{Proj}_{\mathbb{X}_{(n)}} \\ \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t \\ \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \end{pmatrix}^t$$
$$= \sigma_\star^2 \begin{pmatrix} \mathrm{Proj}_{\mathbb{X}_{(n)}} & \mathbb{X}_{(n)}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1} & \mathbb{0}_{(n\times n)} \\ \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t & \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1} & \mathbb{0}_{(p\times n)} \\ \mathbb{0}_{(n\times n)} & \mathbb{0}_{(n\times p)} & \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \end{pmatrix}$$

where we have used

$$\begin{aligned} \mathrm{Proj}_{\mathbb{X}_{(n)}}^2 &= \mathrm{Proj}_{\mathbb{X}_{(n)}} \\ \left(\mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\right)^2 &= \mathrm{Proj}_{\mathbb{X}_{(n)}^\perp}^2 = \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \\ \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1} &= \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1} \\ \mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{X}_{(n)}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1} &= \mathbb{X}_{(n)}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1} \\ &= \mathbb{X}_{(n)}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1} \end{aligned}$$

Prediction

- *New observation:* $\widetilde{Y} = \underline{\widetilde{X}}^t\beta^\star + \tilde{\epsilon}$ with $\mathbb{E}\left[\tilde{\epsilon}\middle|\underline{\widetilde{X}}\right] = 0$, $\mathbb{V}\mathrm{ar}\left[\tilde{\epsilon}\middle|\underline{\widetilde{X}}\right] = \sigma_\star^2$ and $\mathbb{C}\mathrm{ov}\left[\tilde{\epsilon}, Y\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right] = \mathbb{0}_{(1\times n)}$
- *Prediction:* $f_{\hat{\beta}}(\underline{\widetilde{X}}) = \underline{\widetilde{X}}^t\hat{\beta}$

Bias and Variance

- Expectation:

$$\mathbb{E}\left[\widetilde{Y} - f_{\hat{\beta}}(\underline{\widetilde{X}})\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right] = 0$$

- Variance:

$$\mathbb{V}\mathrm{ar}\left[\widetilde{Y} - f_{\hat{\beta}}(\underline{\widetilde{X}})\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right] = \sigma_\star^2\left(1 + \underline{\widetilde{X}}^t\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\underline{\widetilde{X}}\right)$$

Proof

Define the *pseudo-inverse* of $\mathbb{X}_{(n)}$:

$$\mathbb{X}_{(n)}^{\#} = \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t$$

- By construction,

$$\begin{aligned}\widetilde{Y} - f_{\hat{\beta}}(\underline{\widetilde{X}}) &= \widetilde{Y} - \underline{\widetilde{X}}^t\mathbb{X}_{(n)}^{\#}\mathbb{Y}_{(n)}\\ &= \underline{\widetilde{X}}^t\beta^\star + \tilde{\epsilon} - \underline{\widetilde{X}}^t\left(\beta^\star + \mathbb{X}_{(n)}^{\#}\epsilon_{(n)}\right)\\ &= \begin{pmatrix}1 & \underline{\widetilde{X}}^t\mathbb{X}_{(n)}^{\#}\end{pmatrix}\begin{pmatrix}\tilde{\epsilon}\\ \epsilon_{(n)}\end{pmatrix}\end{aligned}$$

- Expectation:

$$\begin{aligned}\mathbb{E}\left[\widetilde{Y} - f_{\hat{\beta}}(\underline{\widetilde{X}})\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right] &= \begin{pmatrix}1 & \underline{\widetilde{X}}^t\mathbb{X}_{(n)}^{\#}\end{pmatrix}\mathbb{E}\left[\begin{pmatrix}\tilde{\epsilon}\\ \epsilon_{(n)}\end{pmatrix}\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right]\\ &= \begin{pmatrix}1 & \underline{\widetilde{X}}^t\mathbb{X}_{(n)}^{\#}\end{pmatrix}\begin{pmatrix}0\\ \mathbb{0}_{(n)}\end{pmatrix}\\ &= 0\end{aligned}$$

- Variance:

$$\begin{aligned}&\mathbb{V}\mathrm{ar}\left[\widetilde{Y} - f_{\hat{\beta}}(\underline{\widetilde{X}})\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right]\\ &= \mathbb{V}\mathrm{ar}\left[\begin{pmatrix}1 & \underline{\widetilde{X}}^t\mathbb{X}_{(n)}^{\#}\end{pmatrix}\begin{pmatrix}\tilde{\epsilon}\\ \epsilon_{(n)}\end{pmatrix}\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right]\\ &= \begin{pmatrix}1 & \underline{\widetilde{X}}^t\mathbb{X}_{(n)}^{\#}\end{pmatrix}\mathbb{C}\mathrm{ov}\left[\begin{pmatrix}\tilde{\epsilon}\\ \epsilon_{(n)}\end{pmatrix}\middle|\mathbb{X}_{(n)}, \underline{\widetilde{X}}\right]\begin{pmatrix}1\\ \mathbb{X}_{(n)}^{\#\,t}\underline{\widetilde{X}}\end{pmatrix}\\ &= \begin{pmatrix}1 & \underline{\widetilde{X}}^t\mathbb{X}_{(n)}^{\#}\end{pmatrix}\sigma_\star^2\mathrm{Id}_{(n+1)}\begin{pmatrix}1\\ \mathbb{X}_{(n)}^{\#\,t}\underline{\widetilde{X}}\end{pmatrix}\\ &= \sigma_\star^2\left(1 + \underline{\widetilde{X}}^t\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\underline{\widetilde{X}}\right)\end{aligned}$$

Residual

$$\widehat{\epsilon}_{(n)} = \mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta} = (\mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}})\mathbb{Y}_{(n)}$$

Residual Properties

- Expectation:

$$\mathbb{E}\left[\widehat{\epsilon}_{(n)}\middle|\mathbb{X}_{(n)}\right] = \mathbb{0}_{(n)}$$

- Covariance:

$$\mathbb{C}\mathrm{ov}\left[\widehat{\epsilon}_{(n)}\middle|\mathbb{X}_{(n)}\right] = \sigma_\star^2(\mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}) = \sigma_\star^2\,\mathrm{Proj}_{\mathbb{X}_{(n)}^\perp}$$

- Expectation of the norm:

$$\mathbb{E}\left[\|\widehat{\epsilon}_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right] = \sigma_\star^2(n-p)$$

Proof

- Norm Expectation:

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{\epsilon}_{(n)}\|^2 \middle| \mathbb{X}_{(n)}\right] &= \mathbb{E}\left[\widehat{\epsilon}^{\,t}_{(n)}\widehat{\epsilon}_{(n)} \middle| \mathbb{X}_{(n)}\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left(\widehat{\epsilon}^{\,t}_{(n)}\widehat{\epsilon}_{(n)}\right) \middle| \mathbb{X}_{(n)}\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left(\widehat{\epsilon}_{(n)}\widehat{\epsilon}^{\,t}_{(n)}\right) \middle| \mathbb{X}_{(n)}\right] \\
&= \operatorname{tr}\left(\mathbb{C}\mathrm{ov}\left[\widehat{\epsilon}_{(n)} \middle| \mathbb{X}_{(n)}\right]\right) \\
&= \operatorname{tr}\left(\sigma_\star^2(\mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}})\right) \\
&= \sigma_\star^2(n-p)
\end{aligned}$$

Variance Estimation

The expectation of the norm of the residual is given by: $\mathbb{E}\left[\|\widehat{\epsilon}_{(n)}\|^2 \middle| \mathbb{X}_{(n)}\right] = \sigma_\star^2(n-p)$.

Proposition 6. *The estimator*

$$\widehat{\sigma^2} = \frac{\|\widehat{\epsilon}_{(n)}\|^2}{n-p} = \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{n-p}$$

is an unbiased estimator of the variance:

$$\mathbb{E}\left[\widehat{\sigma^2} \middle| \mathbb{X}_{(n)}\right] = \sigma_\star^2$$

- Unbiasedness is not preserved by nonlinear transformation; for example, the "natural" estimator of the standard deviation $\sqrt{\widehat{\sigma^2}}$ is biased!

$$\mathbb{E}\left[\sqrt{\widehat{\sigma^2}} \middle| \mathbb{X}_{(n)}\right] < \sqrt{\mathbb{E}\left[\widehat{\sigma^2} \middle| \mathbb{X}_{(n)}\right]} = \sigma_\star \quad \text{(Jensen!)}$$

1.1.3 Asymptotic Analysis

Asymptotic in n

What's going on when n goes to $+\infty$?

- Does $\hat{\beta}$ converge to $\beta^\star$?
- Is there a limiting distribution for $(\hat{\beta} - \beta^\star)$ and what is the proper normalization?

- So far all the results are obtained conditionally to $\mathbb{X}_{(n)}$!
- *Missing information:* behavior of $\mathbb{X}_{(n)}$ when n goes to $+\infty$.

Two models for $\mathbb{X}_{(n)}$

- Fixed design:
  - No statistical model for $\underline{X}_i$ / Results conditionally to $\mathbb{X}_{(n)}$
  - Need to impose some restrictions on the asymptotic behavior of $\mathbb{X}_{(n)}$.
- Random design:
  - $\{\underline{X}_i\}_{i=1}^{\infty}$ are i.i.d. / Probabilistic results.
  - Need to impose some (mild) restriction on the law of $\underline{X}_i$.

Consistency

- *Prop:*

$$\mathbb{E}\left[\|\hat{\beta}-\beta^\star\|^2\middle|\mathbb{X}_{(n)}\right]=\sigma_\star^2\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\right)$$

Consistency

- *Cor:* (Fixed design)

$$\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\right)\xrightarrow[n\to+\infty]{}0\Rightarrow\hat{\beta}|\mathbb{X}_{(n)}\xrightarrow[n\to+\infty]{L^2}\beta^\star$$

- *Cor:* (Random Design) If $\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\right)\xrightarrow{P}0$ then

$$\hat{\beta}\xrightarrow[n\to+\infty]{P}\beta^\star$$

- *Rk:* The L^2 implies also the convergence in probability.
- *Random design:* A sufficient condition for the convergence in probability: $\mathbb{E}\left[\underline{X}\underline{X}^t\right]$ definite positive.

Proof

- Reusing the previous bias and variance computation:

$$\begin{aligned}\mathbb{E}\left[\|\hat{\beta}-\beta^\star\|^2\middle|\mathbb{X}_{(n)}\right]&=\mathbb{E}\left[\operatorname{tr}\left(\left(\hat{\beta}-\beta^\star\right)^t\left(\hat{\beta}-\beta^\star\right)\right)\middle|\mathbb{X}_{(n)}\right]\\&=\mathbb{E}\left[\operatorname{tr}\left(\left(\hat{\beta}-\beta^\star\right)\left(\hat{\beta}-\beta^\star\right)^t\right)\middle|\mathbb{X}_{(n)}\right]\\&=\operatorname{tr}\left(\mathbb{E}\left[\left(\hat{\beta}-\beta^\star\right)\left(\hat{\beta}-\beta^\star\right)^t\middle|\mathbb{X}_{(n)}\right]\right)\\&=\operatorname{tr}\left(\mathbb{C}\mathrm{ov}\left[\hat{\beta}\middle|\mathbb{X}_{(n)}\right]\right)\\&=\sigma_\star^2\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\right)\end{aligned}$$

- The fixed design case is a direct application of this property.
- For the random design, using the previous result

$$\begin{aligned}\mathbb{P}\left(\|\hat{\beta}-\beta^\star\|^2>\epsilon\middle|\mathbb{X}_{(n)}\right)&\leq\frac{\mathbb{E}\left[\|\hat{\beta}-\beta^\star\|^2\middle|\mathbb{X}_{(n)}\right]}{\epsilon}\\&\leq\frac{\sigma_\star^2}{\epsilon}\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\right)\end{aligned}$$

  which implies by conditioning on $\frac{\sigma_\star^2}{\epsilon}\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\right)>\eta$

$$\begin{aligned}\mathbb{P}\left(\|\hat{\beta}-\beta^\star\|^2>\epsilon\right)\leq\mathbb{P}&\left(\frac{\sigma_\star^2}{\epsilon}\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\right)>\eta\right)\\&+\eta\mathbb{P}\left(\frac{\sigma_\star^2}{\epsilon}\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\right)\leq\eta\right)\\\leq\mathbb{P}&\left(\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\right)>\frac{\eta\epsilon}{\sigma_\star^2}\right)+\eta\end{aligned}$$

  that can be make arbitrarily small if

$$\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)\right)\xrightarrow{P}0$$

- As

$$\begin{aligned}\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} &= \sum_{i=1}^n \underline{X}_i \underline{X}_i{}^t \\ &= n\mathbb{E}_n \left[\underline{X}\underline{X}^t\right]\end{aligned}$$

  then if we assume that $\mathbb{E}\left[\underline{X}\underline{X}^t\right] = Q$ definite positive, by the weak Law of Large Number,

$$\frac{1}{n}\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \xrightarrow{P} Q$$

  and then by Slutsky Lemma

$$n\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \xrightarrow{P} Q^{-1}$$

  and thus

$$\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\right) \xrightarrow{P} 0$$

1.2 Gaussian Linear Regression

1.2.1 Gaussian Case

Gaussian Statistical Model

- Natural extension of the *second order statistical model*: *Gaussian* structure on the noise, which is thus *entirely characterized by its second order properties.*

Gaussian probabilistic (statistical) model

- Observation:

$$Y = \underline{X}^t \beta^\star + \epsilon$$

  with ϵ a *noise* satisfying $\epsilon \sim \mathrm{N}(0, \sigma_\star^2)$.

- Observations:

$$Y_i = \underline{X}_i{}^t \beta^\star + \epsilon_i$$

  with ϵ_i i.id. satisfying $\epsilon_i \sim \mathrm{N}(0, \sigma_\star^2)$ (equivalent to jointly Gaussian and uncorrelated)

- Equivalent reformulation: independent

$$Y_i | \underline{X}_i \sim \mathrm{N}(\underline{X}_i{}^t \beta^\star, \sigma_\star^2)$$

Matrix formulation

- $\mathbb{Y}_{(n)} = \mathbb{X}_{(n)}\beta^\star + \epsilon_{(n)}$ with

$$\epsilon_{(n)} = (\epsilon_1, \ldots, \epsilon_n)^t \sim \mathrm{N}(\mathbb{0}_{(n)}, \sigma_\star^2 \mathrm{Id}_{(n)})$$

- Equivalent reformulation:

$$\mathbb{Y}_{(n)} \Big| \mathbb{X}_{(n)} \sim \mathrm{N}\left(\mathbb{X}_{(n)}\beta^\star, \sigma_\star^2 \mathrm{Id}_{(n)}\right)$$

Gaussian distribution of Least Square Estimators

Proposition 7 (Gaussian distribution of the estimators)**.**

$$\left.\begin{pmatrix} \mathbb{X}_{(n)}\hat{\beta} \\ \hat{\beta} \\ \widehat{\epsilon}_{(n)} \end{pmatrix}\right| \mathbb{X}_{(n)} \sim \mathrm{N}\left(\begin{pmatrix} \mathbb{X}_{(n)}\beta^{\star} \\ \beta^{\star} \\ \mathbb{0}_{(n)} \end{pmatrix}, \sigma_{\star}^2 V\right)$$

with

$$V = \begin{pmatrix} \mathrm{Proj}_{\mathbb{X}_{(n)}} & \mathbb{X}_{(n)}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1} & \mathbb{0}_{(n\times n)} \\ \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t & \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1} & \mathbb{0}_{(p\times n)} \\ \mathbb{0}_{(n\times n)} & \mathbb{0}_{(n\times p)} & \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \end{pmatrix}$$

The distribution of the estimator is entirely characterized by its expectation and its covariance.

Proof

- We have

$$\begin{pmatrix} \mathbb{X}_{(n)}\hat{\beta} \\ \hat{\beta} \\ \widehat{\epsilon}_{(n)} \end{pmatrix} = \begin{pmatrix} \mathrm{Proj}_{\mathbb{X}_{(n)}} \\ \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t \\ \mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \end{pmatrix} \mathbb{Y}_{(n)}$$

- Thus

$$\left.\begin{pmatrix} \mathbb{X}_{(n)}\hat{\beta} \\ \hat{\beta} \\ \widehat{\epsilon}_{(n)} \end{pmatrix}\right| \mathbb{X}_{(n)}$$

  is a Gaussian vector (the distribution of which is characterized by its expectation and its covariance).

Cochran Theorem

Proposition 8. *If* $Z \sim \mathrm{N}(\mathbb{0}_{(n)}, \sigma^2\mathrm{Id}_{(n)})$ *and* Proj *is an orthogonal projector on a space of dimension* p *then*

- *The random variables*

$$\mathrm{Proj}\, Z \sim \mathrm{N}(\mathbb{0}_{(n)}, \sigma^2\,\mathrm{Proj})$$
$$(\mathrm{Id}_{(n)} - \mathrm{Proj})Z \sim \mathrm{N}\left(\mathbb{0}_{(n)}, \sigma^2(\mathrm{Id}_{(n)} - \mathrm{Proj})\right)$$

  are independent.

- $\|(\mathrm{Id}_{(n)} - \mathrm{Proj})Z\|^2 \sim \sigma^2\,\chi^2(n-p)$ *a* χ^2 *variable with* $n-p$ *degrees of freedom..*

Proof

- If $Z \sim \mathrm{N}(\mathbb{0}_{(n)}, \sigma^2\mathrm{Id}_{(n)})$ then

$$\begin{pmatrix} \mathrm{Proj}\, Z \\ (\mathrm{Id}_{(n)} - \mathrm{Proj})Z \end{pmatrix} = \begin{pmatrix} \mathrm{Proj} \\ \mathrm{Id}_{(n)} - \mathrm{Proj} \end{pmatrix} Z$$

- Hence

$$\begin{aligned}\begin{pmatrix}\operatorname{Proj} Z\\(\mathrm{Id}_{(n)}-\operatorname{Proj})Z\end{pmatrix} &\sim \mathrm{N}\left(\begin{pmatrix}\operatorname{Proj}\\\mathrm{Id}_{(n)}-\operatorname{Proj}\end{pmatrix}\mathbb{E}\left[Z\right],\right.\\ &\qquad\left.\begin{pmatrix}\operatorname{Proj}\\\mathrm{Id}_{(n)}-\operatorname{Proj}\end{pmatrix}\mathbb{C}\mathrm{ov}\left[Z\right]\begin{pmatrix}\operatorname{Proj}\\\mathrm{Id}_{(n)}-\operatorname{Proj}\end{pmatrix}^t\right)\\ &\sim \mathrm{N}\left(\mathbb{0}_{(n)},\sigma^2\begin{pmatrix}\operatorname{Proj} & \mathbb{0}_{(n\times n)}\\\mathbb{0}_{(n\times n)} & \mathrm{Id}_{(n)}-\operatorname{Proj}\end{pmatrix}\right)\end{aligned}$$

  where we have used $\operatorname{Proj}\operatorname{Proj}^t=\operatorname{Proj}^2=\operatorname{Proj}$, $(\mathrm{Id}_{(n)}-\operatorname{Proj})(\mathrm{Id}_{(n)}-\operatorname{Proj})^t=(\mathrm{Id}_{(n)}-\operatorname{Proj})^2=\mathrm{Id}_{(n)}-\operatorname{Proj}$ and $\operatorname{Proj}\,({}^t\mathrm{Id}_{(n)}-\operatorname{Proj})=\operatorname{Proj}(\mathrm{Id}_{(n)}-\operatorname{Proj})=\mathbb{0}_{(n\times n)}$.

- Now

$$\begin{pmatrix}\operatorname{Proj} Z\\(\mathrm{Id}_{(n)}-\operatorname{Proj})Z\end{pmatrix}\sim \mathrm{N}\left(\mathbb{0}_{(n)},\sigma^2\begin{pmatrix}\operatorname{Proj} & \mathbb{0}_{(n\times n)}\\\mathbb{0}_{(n\times n)} & \mathrm{Id}_{(n)}-\operatorname{Proj}\end{pmatrix}\right)$$

  which implies that

  - $\operatorname{Proj} Z$ and $(\mathrm{Id}_{(n)}-\operatorname{Proj})Z$ are jointly Gaussian with $\operatorname{Proj} Z\sim\mathrm{N}(\mathbb{0}_{(n)},\sigma^2\operatorname{Proj})$ and $(\mathrm{Id}_{(n)}-\operatorname{Proj})Z\sim\mathrm{N}(\mathbb{0},\sigma^2(\mathrm{Id}_{(n)}-\operatorname{Proj}))$
  - $\operatorname{Proj} Z$ and $(\mathrm{Id}_{(n)}-\operatorname{Proj})Z$ are jointly Gaussian and uncorrelated $\mathbb{C}\mathrm{ov}\left[\operatorname{Proj} Z,(\mathrm{Id}_{(n)}-\operatorname{Proj})Z\right]=\mathbb{0}_{(n\times n)}$. Therefore, $\operatorname{Proj} Z$ and $(\mathrm{Id}_{(n)}-\operatorname{Proj})Z$ are independent.

- Proj being an orthogonal projection, it can diagonalized in an orthonormal basis, i.e. there exits an orthogonal matrix B ($B^tB=BB^t=\mathrm{Id}$) satisfying

$$\operatorname{Proj}=BDB^t$$

  where D is a diagonal matrix where all the elements are equal to 0 except the p first one that are equal to 1.

- Since B is orthogonal matrix

$$\begin{aligned}\|(\mathrm{Id}_{(n)}-\operatorname{Proj})Z\|^2&=\|B(\mathrm{Id}_{(n)}-D)B^tZ\|^2\\&=\|(\mathrm{Id}_{(n)}-D)B^tZ\|^2\\&=\sum_{i=p+1}^{n}\left|(B^tZ)_i\right|^2\end{aligned}$$

- Now

$$B^tZ\sim\mathrm{N}(\mathbb{0}_{(n)},\sigma^2B^tB)=\mathrm{N}(\mathbb{0}_{(n)},\sigma^2\mathrm{Id}_{(n)})\ ,$$

  which implies that the $(B^tZ)_i$ are i.i.d. $\mathrm{N}(0,\sigma^2)$.

- The distribution $\|(\mathrm{Id}_{(n)}-\operatorname{Proj})Z\|^2$ is thus the law of the sum of $n-p$ independent Gaussian $\mathrm{N}(0,\sigma^2)$, i.e. $\sigma^2\,\chi^2(n-p)$.

Student distribution

Definition 1. Let Z and S^2 be two independent random variable such that $Z\sim\mathrm{N}(0,1)$ and $S^2\sim\chi^2(d)/d$. The law of Z/S is called the Student distribution with d degrees of freedom. d denoted $\mathrm{T}(d)$.

Proposition 9. *If Z and S^2 are two independent r.v. such that that $Z\sim\mathrm{N}(\mu,\sigma^2)$ and $S^2/\sigma^2\sim\chi^2(d)/d$ then*

$$\frac{Z-\mu}{S}\sim\mathrm{T}(d)$$

Distribution of the Variance Estimator

- Unbiased variance estimator:

$$\widehat{\sigma^2} = \frac{\|\widehat{\epsilon}_{(n)}\|^2}{n-p} = \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{n-p}$$

Properties

- $\widehat{\sigma^2}\Big|\mathbb{X}_{(n)} \sim \sigma_\star^2\, \chi^2(n-p)/(n-p)$
- $\widehat{\sigma^2}\Big|\mathbb{X}_{(n)}$ is independent of $\begin{pmatrix}\mathbb{X}_{(n)}\hat{\beta}\\ \hat{\beta}\end{pmatrix}$
- For any $a \in \mathbb{R}^p$,

$$\left.\frac{a^t\hat{\beta} - a^t\beta^\star}{\sqrt{\widehat{\sigma^2}a^t\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}a}}\right|\mathbb{X}_{(n)} \sim \mathrm{T}(n-p)$$

where $\mathrm{T}(n-p)$ is the Student law with $n-p$ *degrees of freedom*.

Proof

- The properties of $\widehat{\sigma^2}$ are a direct application of Cochran theorem to

$$\epsilon_{(n)} = \mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\beta^\star \sim \mathrm{N}(\mathbb{0}_{(n)}, \sigma_\star^2\mathrm{Id}_{(n)})$$

and $\mathrm{Proj} = \mathrm{Proj}_{\mathbb{X}_{(n)}}$.

- Now

$$\begin{aligned}\mathrm{Proj}_{\mathbb{X}_{(n)}}\epsilon_{(n)} &= \mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{X}_{(n)}\beta^\star\\ &= \mathbb{X}_{(n)}(\hat{\beta} - \beta^\star) \sim \mathrm{N}(\mathbb{0}_{(n)}, \sigma^2\,\mathrm{Proj}_{\mathbb{X}_{(n)}})\end{aligned}$$

and

$$\begin{aligned}&(\mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}})\epsilon_{(n)}\\ &= \left(\mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\right)\mathbb{Y}_{(n)} - \left(\mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\right)\mathbb{X}_{(n)}\beta^\star\\ &= \mathbb{Y}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\mathbb{Y}_{(n)} = \mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\\ &= \widehat{\epsilon}_{(n)} \sim \mathrm{N}(\mathbb{0}_{(n)}, \sigma^2(\mathrm{Id} - \mathrm{Proj}_{\mathbb{X}_{(n)}}))\end{aligned}$$

are independent

Prediction

- New observation:

$$\widetilde{Y} = \underline{\widetilde{X}}^t\beta^\star + \widetilde{\epsilon}$$

with $\widetilde{\epsilon}$ $\mathrm{N}(0, \sigma^2)$ indep. from $\epsilon_{(n)}$ conditionally to $\mathbb{X}_{(n)}$ and $\widetilde{X}$.

- Prediction:

$$\begin{aligned} f_{\hat\beta}(\underline{\widetilde{X}}) &= \underline{\widetilde{X}}^t \hat\beta \\ &= \underline{\widetilde{X}}^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \mathbb{X}_{(n)}^t \mathbb{Y}_{(n)} \end{aligned}$$

Laws

- Prediction:

$$f_{\hat\beta}(\underline{\widetilde{X}})\Big|\mathbb{X}_{(n)}, \underline{\widetilde{X}} \sim \mathrm{N}\left(\underline{\widetilde{X}}^t \beta^\star, \sigma_\star^2 \underline{\widetilde{X}}^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \underline{\widetilde{X}}\right)$$

- Error:

$$\widetilde{Y} - f_{\hat\beta}(\underline{\widetilde{X}})\Big|\mathbb{X}_{(n)}, \underline{\widetilde{X}} \sim \mathrm{N}\left(0, \sigma_\star^2\left(1 + \underline{\widetilde{X}}^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \underline{\widetilde{X}}\right)\right)$$

- Error norm:

$$\|\widetilde{Y} - f_{\hat\beta}(\underline{\widetilde{X}})\|^2\Big|\mathbb{X}_{(n)}, \underline{\widetilde{X}} \sim \sigma_\star^2\left(1 + \underline{\widetilde{X}}^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \underline{\widetilde{X}}\right)\chi^2(1)$$

1.2.2 Link with Maximum Likelihood

Gaussian Likelihood

- Gaussian model:

$$\mathbb{Y}_{(n)}\Big|\mathbb{X}_{(n)} \sim \mathrm{N}\left(\mathbb{X}_{(n)}\beta^\star, \sigma_\star^2 \mathrm{Id}_{(n)}\right)$$

- Both $\beta^\star$ and $\sigma_\star$ are unknown to the statistician!

Pdf and Likelihood of the observation

- Probability Density Function:

$$\mathbb{P}_{\beta,s^2}\left(\mathbb{Y}_{(n)}\Big|\mathbb{X}_{(n)}\right) = \frac{1}{\left(2\pi s^2\right)^{n/2}} e^{-\frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\beta\|^2}{2s^2}}$$

- Likelihood:

$$\mathcal{L}_{\mathbb{Y}_{(n)}|\mathbb{X}_{(n)}}\left(\beta, s^2\right) = \frac{1}{\left(2\pi s^2\right)^{n/2}} e^{-\frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\beta\|^2}{2s^2}}$$

Maximum Likelihood

- *Maximum Likelihood Principle* (Fisher): estimate the parameters by the ones that maximimize the likelihood.

Maximum Likelihood Estimation

- ML:

$$\begin{aligned}(\hat{\beta},\widehat{s^2}) &= \operatorname*{argmax}_{\beta,s^2} \mathcal{L}_{\mathbb{Y}_{(n)}|\mathbb{X}_{(n)}}\left(\beta,s^2\right)\\ &= \operatorname*{argmax}_{\beta,s^2} \frac{1}{\left(2\pi s^2\right)^{n/2}} e^{-\frac{\|\mathbb{Y}_{(n)}-\mathbb{X}_{(n)}\beta\|^2}{2s^2}}\end{aligned}$$

- ML and opposite of the Log-Likelihood:

$$\begin{aligned}(\hat{\beta},\widehat{s^2}) &= \operatorname*{argmin}_{\beta,s^2} -\log \mathcal{L}_{\mathbb{Y}_{(n)}|\mathbb{X}_{(n)}}\left(\beta,s^2\right)\\ &= \operatorname*{argmin}_{\beta,s^2} \frac{n}{2}\log\left(2\pi s^2\right) + \frac{\|\mathbb{Y}_{(n)}-\mathbb{X}_{(n)}\beta\|^2}{2s^2}\end{aligned}$$

- Least Square minimization for β!

Maximum Likelihood Estimate

ML Estimate

- Formulas:

$$\begin{aligned}\hat{\beta} &= \left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t\mathbb{Y}_{(n)}\\ \widehat{s^2} &= \frac{\|\mathbb{Y}_{(n)}-\mathbb{X}_{(n)}\hat{\beta}\|^2}{n}\end{aligned}$$

- Differs only in the estimation of the variance!
- *Rk:* In the family of estimate $\frac{\|\mathbb{Y}_{(n)}-\mathbb{X}_{(n)}\hat{\beta}\|^2}{m}$, the best for the quadratic risk is neither $m=n-p$ or $m=n$ but $m=n-p+2$!
- *Rk:* $\widehat{s^2}$ is better than $\widehat{\sigma^2}$ for the quadratic risk when $p\leq 4$ or $p\geq n-2$.

Proof

- For any s^2,

$$-\log \mathcal{L}_{\mathbb{Y}_{(n)}|\mathbb{X}_{(n)}}\left(\beta,s^2\right) = \frac{n}{2}\log\left(2\pi s^2\right) + \frac{\|\mathbb{Y}_{(n)}-\mathbb{X}_{(n)}\beta\|^2}{2s^2}$$

is minimized when $\|\mathbb{Y}_{(n)}-\mathbb{X}_{(n)}\beta\|^2$ is minimum thus $\hat{\beta}$ is the least square estimate.

- Now

$$-\log \mathcal{L}_{\mathbb{Y}_{(n)}|\mathbb{X}_{(n)}}\left(\hat{\beta}, s^2\right) = \frac{n}{2}\log\left(2\pi s^2\right) + \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{2s^2}$$

and thus

$$\frac{\partial - \log \mathcal{L}_{\mathbb{Y}_{(n)}|\mathbb{X}_{(n)}}\left(\hat{\beta}, s^2\right)}{ds^2} = \frac{n}{2}\frac{1}{s^2} - \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{2\left(s^2\right)^2}$$

The first order optimality condition yields

$$\frac{n}{2}\frac{1}{\widehat{s^2}} - \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{2\left(\widehat{s^2}\right)^2} = 0$$

which implies

$$\widehat{s^2} = \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{n}$$

- Let

$$\widehat{\sigma^2}_\kappa = \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{\kappa},$$

as $\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2|\mathbb{X}_{(n)} \sim \sigma_\star^2\,\chi^2(n-p)$

$$\begin{aligned}\mathbb{E}\left[\widehat{\sigma^2}_\kappa\middle|\mathbb{X}_{(n)}\right] &= \sigma_\star^2\frac{n-p}{\kappa}\\ \mathbb{V}\mathrm{ar}\left[\widehat{\sigma^2}_\kappa\middle|\mathbb{X}_{(n)}\right] &= \sigma_\star^4\frac{2(n-p)}{\kappa^2}\end{aligned}$$

This implies that

$$\begin{aligned}&\mathbb{E}\left[|\widehat{\sigma^2}_\kappa - \sigma_\star^2|^2\middle|\mathbb{X}_{(n)}\right]\\ &\quad= \left|C\mathbb{E}\left[\widehat{\sigma^2}_\kappa\right]\mathbb{X}_{(n)} - \sigma_\star^2\right|^2 + \mathbb{V}\mathrm{ar}\left[\widehat{\sigma^2}_\kappa\middle|\mathbb{X}_{(n)}\right]\\ &\quad= \sigma_\star^4\left(\left(\frac{n-p}{\kappa} - 1\right)^2 + \frac{2(n-p)}{\kappa^2}\right)\\ &\quad= \sigma_\star^4\left(1 - 2\frac{n-p}{\kappa} + \frac{(n-p)(n-p+2)}{\kappa^2}\right)\\ &\quad= \sigma_\star^4\left(\left(\frac{\sqrt{(n-p)(n-p+2)}}{\kappa} - \frac{\sqrt{n-p}}{\sqrt{n-p+2}}\right)^2 + 1 - \frac{n-p}{n-p+2}\right)\\ &\quad= \sigma_\star^4\left((n-p)(n-p+2)\left(\frac{1}{\kappa} - \frac{1}{n-p+2}\right)^2 + 1 - \frac{n-p}{n-p+2}\right)\end{aligned}$$

which implies that the best choice is $\kappa = n - p + 2$.

- Now $\widehat{s^2}$ is better than $\widehat{\sigma^2}$ when

$$\begin{aligned}&\left(\frac{1}{n-p} - \frac{1}{n-p+2}\right)^2 > \left(\frac{1}{n} - \frac{1}{n-p+2}\right)^2\\ \Leftrightarrow &\left(\frac{2}{(n-p)(n-p+2)}\right)^2 > \left(\frac{-p+2}{n(n-p+2)}\right)^2\end{aligned}$$

$$\begin{aligned} &\Leftrightarrow \frac{4}{(n-p)^2} > \frac{(p-2)^2}{n^2} \\ &\Leftrightarrow 4n^2 > (n-p)^2(p-2)^2 \end{aligned}$$

This strict inequality is always satisfied when $0 \leq p \leq 4$ or $n-2 \leq p \leq n$.

A tedious analysis shows that this is the only case when $n \geq 16$. Otherwise,

- when $n = 15$, we have an equality for $p = 5$ and $n = 12$ and a strict inequaliy when $0 \leq p \leq 4$ or $13 \leq p \leq n$.
- when $13 \leq n \leq 14$, we have a strict inequality when $0 \leq p \leq 5$ or $n-3 \leq p \leq n$.
- when $n = 12$, we have an equality for $p = 6$ and $n = 8$ and a strict inequality when $0 \leq p \leq 5$ or $n-3 \leq p \leq n$.
- when $n \leq 11$, we always have a strict inequality.

1.2.3 Confidence Region

Confidence Region and Test

Confidence Region

- Construct a data dependent region Z (resp. an interval I) to which $\beta^\star$ (resp. $a^t\beta^\star$) belongs with a probability close to $1-\alpha$

- Example: Confidence interval of the univariate case.

Test

- Given an hypothesis H_0 (and possibly an alternative H_1).
- Construct a test which may disprove the assumption H_0 with a control on the probability of error.

- Examples: Student based tests of nullity of a coefficient in the univariate case.

Known variance $\sigma_\star^2$

$$\left(\hat{\beta} - \beta^\star\right) \sim \mathrm{N}\left(\mathbb{0}_{(p)}, \sigma_\star^2 \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\right)$$

Consequences

- Scalar product:

$$\frac{\left(a^t\hat{\beta} - a^t\beta^\star\right)}{\sqrt{\sigma_\star^2 a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} a}} \sim \mathrm{N}\left(0,1\right)$$

- Let $\mathbb{M}$ be a $q \times p$ matrix of rank q:

$$\frac{1}{q\sigma_\star^2} \left\| \left(\mathbb{M}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2} \left(\mathbb{M}\hat{\beta} - \mathbb{M}\beta^\star\right) \right\|^2 \sim \chi^2(q)/q$$

Fisher distribution

Definition 2. Let V^2 and V'^2 be two independent variables such that $V^2 \sim \chi^2(q)$ and $V'^2 \sim \chi^2(q')$. The distribution of the ratio

$$\frac{V^2}{q} \Big/ \frac{V'^2}{q'}$$

is called the Fisher distribution with (q,q') degrees of freedom. This distribution is denoted $\mathrm{F}(q,q')$.

Proposition 10. *Let V^2 and V'^2 be two independent variables such that $V^2/\sigma^2 \sim \chi^2(q)$ and $V'^2/\sigma^2 \sim \chi^2(q')$. Then*

$$\frac{V^2}{q} \Big/ \frac{V'^2}{q'} \sim \mathrm{F}(q,q')$$

Proof

- The first property is straightforward as

$$\frac{V^2}{q} \Big/ \frac{V'^2}{q'} = \frac{V^2}{\sigma^2 q} \Big/ \frac{V'^2}{\sigma^2 q'} .$$

- To prove $\mathrm{t}_{1-\alpha/2}(n-p) = \sqrt{\mathrm{f}_{1-\alpha}(1,n-p)}$, we let $Z \sim \mathrm{N}(0,1)$ and $S^2 \sim \chi^2(n-p)/n-p$ independent of Z. By definition, $Z/S \sim \mathrm{T}(n-p)$ while $\|Z\|^2/S^2 \sim \mathrm{F}(1,n-p)$. This implies thus that $\mathbb{P}(|Z/S| \leq t) = \mathbb{P}\left(\|Z\|^2/S^2 \leq t^2\right)$, which in turn implies $\mathrm{t}_{1-\alpha/2}(n-p) = \sqrt{\mathrm{f}_{1-\alpha}(1,n-p)}$.

Unknown σ^2

$\widehat{\sigma^2} \sim \chi^2(n-p)/(n-p)$ and is independent of $\hat\beta$.

Consequences

- Scalar product:

$$\frac{\left(a^t\hat\beta - a^t\beta^\star\right)}{\sqrt{\widehat{\sigma^2} a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} a}} \sim \mathrm{T}(n-p)$$

- Norms: Let $\mathbb{M}$ be a $q \times p$ matrix of rank q:

$$\frac{1}{q\widehat{\sigma^2}} \left\| \left(\mathbb{M} \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \mathbb{M}^t \right)^{-1/2} \left(\mathbb{M}\hat\beta - \mathbb{M}\beta^\star\right) \right\|^2 \sim \mathrm{F}(q,n-p)$$

  where $\mathrm{F}(q,n-p)$ is the Fisher law of degrees q and $n-p$.

Proof

- Recall that

$$\left(\hat\beta - \beta^\star\right) \sim \mathrm{N}\left(\mathbb{0}_{(p)}, \sigma_\star^2 \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\right)$$

  so that

$$\mathbb{M}\left(\hat\beta - \beta^\star\right) \sim \mathrm{N}\left(\mathbb{0}_{(q)}, \sigma_\star^2 \mathbb{M} \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \mathbb{M}^t\right)$$

- Using $\mathbb{M} = a^t$, we obtain

$$\begin{aligned} \left(a^t\hat{\beta} - a^t\beta^\star\right) &\sim \mathrm{N}\left(0, \sigma_\star^2 a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} a\right) \\ \Leftrightarrow \frac{\left(a^t\hat{\beta} - a^t\beta^\star\right)}{\sqrt{\sigma_\star^2 a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} a}} &\sim \mathrm{N}\left(0,1\right) \end{aligned}$$

- As $\widehat{\sigma^2} \sim \sigma_\star^2 \, \chi^2(n-p)\big/(n-p)$ and is independent of $\hat{\beta}$, we obtain immediately

$$\begin{aligned} \frac{\left(a^t\hat{\beta} - a^t\beta^\star\right)}{\sqrt{\widehat{\sigma^2} a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} a}} &= \frac{\left(a^t\hat{\beta} - a^t\beta^\star\right)}{\sqrt{\sigma_\star^2 a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} a}} \Bigg/ \sqrt{\frac{\widehat{\sigma^2}}{\sigma_\star^2}} \\ &\sim \mathrm{T}(n-p) \end{aligned}$$

- Using, instead of $\mathbb{M}$, $\left(\mathbb{M}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\mathbb{M}$

$$\begin{aligned} &\left(\mathbb{M}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\mathbb{M}\left(\hat{\beta} - \beta^\star\right) \\ &\sim \mathrm{N}\Bigg(\mathbb{0}_{(q)}, \sigma_\star^2 \left(\mathbb{M}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\mathbb{M}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \\ &\qquad\qquad \left(\left(\mathbb{M}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\mathbb{M}\right)^t\Bigg) \\ &\sim \mathrm{N}\left(\mathbb{0}_{(q)}, \sigma_\star^2 \mathrm{Id}_{(q)}\right) \end{aligned}$$

Hence

$$\left\|\left(\mathbb{M}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\mathbb{M}\left(\hat{\beta} - \beta^\star\right)\right|^2 \sim \sigma_\star^2 \, \chi^2(q)$$

- For the last law, it suffices to notice that $\frac{1}{q}\left\|\left(\mathbb{M}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\mathbb{M}\left(\hat{\beta} - \beta^\star\right)\right|^2$ and $\widehat{\sigma^2}$ are independent because $\hat{\beta}$ and $\widehat{\sigma^2}$ are independent and of laws respectively $\sigma_\star^2 \, \chi^2(q)\big/q$ and $\sigma_\star^2 \, \chi^2(n-p)\big/(n-p)$ which implies by the definition of the Fisher law

$$\frac{1}{q\widehat{\sigma^2}}\left\|\left(\mathbb{M}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\left(\mathbb{M}\hat{\beta} - \mathbb{M}\beta^\star\right)\right\|^2 \sim \mathrm{F}(q, n-p)$$

Confidence Interval

Confidence Interval for Scalar Product

- Let $\mathrm{t}_{1-\alpha/2}\,(n-p)$ be the quantile $1-\alpha/2$ of the Student distribution $\mathrm{T}(n-p)$. Then, for all $a \in \mathbb{R}^p$,

$$\mathcal{I}(a, \hat{\beta}, \widehat{\sigma^2}, \alpha) = \left[a^t\hat{\beta} \pm \mathrm{t}_{1-\alpha/2}\,(n-p)\sqrt{\widehat{\sigma^2} a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} a}\right]$$

is a confidence interval for $a^t\beta^\star$ of coverage property $1-\alpha$ cond. to $\mathbb{X}_{(n)}$.

- This is equivalent to say that, *if the model is true*,

$$\mathbb{P}\left(a^t\beta^\star \in \mathcal{I}(a, \hat{\beta}, \widehat{\sigma^2}, \alpha)\middle|\mathbb{X}_{(n)}\right) = 1-\alpha$$

- Interval $\mathcal{I}(a, \hat{\beta}, \widehat{\sigma^2}, \alpha)$ can be computed!

Confidence Region

Confidence Region

- Let $\mathrm{f}_{1-\alpha}(q,n-p)$ be the quantile $1-\alpha$ of the Fisher law $\mathrm{F}(q,n-p)$, for any $q\times p$ matrix of rank q $\mathbb{M}$,
$$\mathcal{E}(\mathbb{M},\hat{\beta},\widehat{\sigma^2}\alpha)=$$
$$\left\{\beta\in\mathbb{R}^p,\left\|\left(\mathbb{M}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\left(\mathbb{M}\hat{\beta}-\mathbb{M}\beta\right)\right\|^2\leq z_{\widehat{\sigma^2},q,n-p,\alpha}\right\}$$
with $z_{\widehat{\sigma^2},q,n-p,\alpha}g=q\widehat{\sigma^2}\,\mathrm{f}_{1-\alpha}(q,n-p)$ is an ellipsoid of confidence $1-\alpha$ for $\mathbb{M}\beta^\star$ conditionally to $\mathbb{X}_{(n)}$.

- This is equivalent to say that, *if the model is true*,
$$\mathbb{P}\left(\mathbb{M}\beta^\star\in\mathcal{E}(\mathbb{M},\hat{\beta},\widehat{\sigma^2},\alpha)\middle|\mathbb{X}_{(n)}\right)=1-\alpha$$

- Ellipse $\mathcal{E}(\mathbb{M},\hat{\beta},\widehat{\sigma^2},\alpha)$ can be computed!

Proof

- The confidence interval is a direct consequence of
$$\frac{\left(a^t\hat{\beta}-a^t\beta^\star\right)}{\sqrt{\widehat{\sigma^2}a^t\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}a}}\sim\mathrm{T}(n-p)$$
Indeed this implies that with probability $1-(\alpha_-+\alpha_+)$
$$\begin{aligned}\mathrm{t}_{\alpha_-}(n-p)\leq&\frac{\left(a^t\hat{\beta}-a^t\beta^\star\right)}{\sqrt{\widehat{\sigma^2}a^t\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}a}}\leq\mathrm{t}_{1-\alpha_+}(n-p)\\&\Leftrightarrow a^t\hat{\beta}+\mathrm{t}_{\alpha_-}(n-p)\sqrt{\widehat{\sigma^2}a^t\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}a}\leq a^t\beta^\star\\&\quad\text{and } a^t\beta^\star\leq a^t\hat{\beta}+\mathrm{t}_{1-\alpha_+}(n-p)\sqrt{\widehat{\sigma^2}a^t\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}a}\end{aligned}$$
which gives the results by using $\alpha_-=\alpha_+=\alpha/2$ and that, thanks to the symmetry of $\mathrm{T}(n-p)$, $\mathrm{t}_{\alpha/2}(n-p)=-\mathrm{t}_{1-\alpha/2}(n-p)$.

- Asymmetric confidence interval of level $1-\alpha$ can be obtained by any choice such that $\alpha_-+\alpha_+=\alpha$. For instance, $\alpha_-=0$ and $\alpha_+=\alpha$ or $\alpha_-=\alpha$ and $\alpha_+=0$

- Along the same line,
$$\frac{1}{q\widehat{\sigma^2}}\left\|\left(\mathbb{M}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\left(\mathbb{M}\hat{\beta}-\mathbb{M}\beta^\star\right)\right\|^2\sim\mathrm{F}(q,n-p)$$
which implies that with probability $1-\alpha$
$$\begin{aligned}&\frac{1}{q\widehat{\sigma^2}}\left\|\left(\mathbb{M}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\left(\mathbb{M}\hat{\beta}-\mathbb{M}\beta^\star\right)\right\|^2\leq\mathrm{f}_{1-\alpha}(q,n-p)\\&\Leftrightarrow\left\|\left(\mathbb{M}\left(\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\left(\mathbb{M}\hat{\beta}-\mathbb{M}\beta^\star\right)\right\|^2\leq q\widehat{\sigma^2}\,\mathrm{f}_{1-\alpha}(q,n-p)\\&\Leftrightarrow\mathbb{M}\beta^\star\in\mathcal{E}(\mathbb{M},\hat{\beta},\widehat{\sigma^2},\alpha)\end{aligned}$$

- For the last properties, with $\mathbb{M} = a^t$ of rank 1

$$\begin{aligned}
& a^t\beta \in \mathcal{E}(a^t, \hat{\beta}, \widehat{\sigma^2}, \alpha) \\
& \Leftrightarrow \left\| \left(a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \right)^{-1} a \right)^{-1/2} \left(a^t\hat{\beta} - a^t\beta \right) \right\|^2 \leq \widehat{\sigma^2}\, \mathrm{f}_{1-\alpha}(1, n-p) \\
& \Leftrightarrow \left(a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \right)^{-1} a \right)^{-1} \left\| a^t\hat{\beta} - a^t\beta \right) \|^2 \leq \widehat{\sigma^2}\, \mathrm{f}_{1-\alpha}(1, n-p) \\
& \Leftrightarrow \left\| a^t\hat{\beta} - a^t\beta \right\|^2 \leq \sqrt{\widehat{\sigma^2} a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \right)^{-1} a} \sqrt{\mathrm{f}_{1-\alpha}(1, n-p)}
\end{aligned}$$

Now using $\mathrm{t}_{1-\alpha/2}(n-p) = \sqrt{\mathrm{f}_{1-\alpha}(1, n-p)}$ so that

$$\begin{aligned}
& \Leftrightarrow \left\| a^t\hat{\beta} - a^t\beta \right\|^2 \leq \sqrt{\widehat{\sigma^2} a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \right)^{-1} a}\, \mathrm{t}_{1-\alpha/2}(n-p) \\
& \Leftrightarrow a^t\beta \in \mathcal{I}(a, \hat{\beta}, \widehat{\sigma^2}, \alpha)
\end{aligned}$$

1.3 Tests in the Gaussian Model

1.3.1 Tests

Hypothesis and Tests

Fisher Approach

- Define a null hypothesis H_0,
- Construct a test $\mathcal{T}$ that may reject H_0
- Control the probability of (wrongly) rejecting H_0 when it holds.

Neyman-Pearson Approach

- Define a null hypothesis H_0 and an alternative hypothesis H_1.
- Construct a test $\mathcal{T}$ that may reject H_0 compared to H_1.
- Control the probability of (wrongly) rejecting H_0 when it holds and of (wrongly) not rejecting H_0 when H_1 holds.

- Rejecting H_0 does not mean accepting H_1!
- Fisher approach requires only a model for H_0 but does not allow a comparison with another hypothesis.

Test Constructions

- Constructions for a test $\mathcal{T}$ of level α, i.e. such that the probability of wrongly rejecting H_0 when it holds is smaller or equal to α (at least asymptotically).

Critical Region Approach

- Select a test statistic T whose behavior is known (at least asymptotically) under H_0.
- Compute a *critical region* $\mathcal{R}(\alpha)$ such that, under H_0,

$$\mathbb{P}\left(T \in \mathcal{R}(\alpha)\right) \geq 1 - \alpha.$$

- Reject H_0 if $T \notin \mathcal{R}(\alpha)$.

- Alternative hypothesis H_1 can be used to select a *good* test, i.e. such that, under H_1, $\mathbb{P}\left(T \in \mathcal{R}(\alpha)\right) = \beta$ is small.

- If T is a scalar that is assumed to be small under H_0, often

$$\mathcal{R}(\alpha) = \{t, t \leq t_{1-\alpha}\}$$

where t_α denote the quantile of order α of T under H_0.

p-value Approach

- Select a scalar T whose behavior is known (at least asympt.) under H_0 and which is assumed to be small under H_0.

- Compute

$$p = \mathbb{P}\left(T' > T | T\right)$$

with T' ind. of T such that $T' \sim T$

- Reject H_0 if $p < \alpha$.

- The two approaches coincides for $\mathcal{R}(\alpha) = \{t, t \leq t_{1-\alpha}\}$:

$$T \notin \mathcal{R}(\alpha) \Leftrightarrow p = \mathbb{P}\left(T' > T | T\right) < \alpha$$

up to some small modification at the boundary.

- Alternative hypothesis H_1 can be used to select a *good* test, i.e. such that, under H_1, $\mathbb{P}\left(p \geq \alpha\right) = \beta$ is small.

- Soft decision: allow to choose α afterwards.

Proof

- Let T' and T be two independent real r.v. of the same law,

$$p = \mathbb{P}\left(T' > T \middle| T\right) < \alpha \Leftrightarrow T > t_{1-\alpha}$$

where t_α is the quantile of order α of T. Hence

$$\begin{aligned}\mathbb{P}\left(p < \alpha\right) &= \mathbb{P}\left(\mathbb{P}\left(T' > T \middle| T\right) < \alpha\right)\\ &= \mathbb{P}\left(T > t_{1-\alpha}\right)\\ &= 1 - \mathbb{P}\left(T \leq t_{1-\alpha}\right) = \alpha\end{aligned}$$

- Critical Region and o-values:

$$\begin{aligned}&T \notin \mathcal{R}(\alpha)\\ &\Leftrightarrow T > t_{1-\alpha}\\ &\Leftrightarrow \mathbb{P}\left(T' \leq T \middle| T\right) \geq 1-\alpha\\ &\Leftrightarrow \mathbb{P}\left(T' > T \middle| T\right) < \alpha\end{aligned}$$

Student Test

- Gaussian statistical model
- Consider the test

$$H_0 : \beta^\star_{(k)} = 0, \qquad \text{against} \qquad H_1 : \beta^\star_{(k)} \neq 0.$$

Proposition 11 (Student test)**.** *Under H_0*

$$\frac{\hat{\beta}_{(k)}}{\sqrt{\widehat{\sigma^2} {e^{(k)}}^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} e^{(k)}}} \sim \mathrm{T}(n-p)$$

where $e^{(k)}$ is a vector in $\mathbb{R}^p$ with all coordinates equal to 0 except the kth one.

Critical Region

- Denote

$$\mathcal{R}(\alpha) = \left\{ \beta \in \mathbb{R} \; : \; |\beta| \leq \mathrm{t}_{1-\alpha/2}(n-p) \sqrt{\widehat{\sigma^2} \left[\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \right]_{k,k}} \right\}$$

Critical Region

- Under the null hypothesis H_0,

$$\mathbb{P}\left(\hat{\beta}_{(k)} \in \mathcal{R}(\alpha)\right) = 1 - \alpha$$

- The test

$$\mathbb{1}\left\{\hat{\beta}_{(k)} \notin \mathcal{R}(\alpha)\right\}$$

  is of level α (the test rejects H_0 when $\hat{\beta}_{(k)} \notin \mathcal{R}(\alpha)$)

Critical Region and Confidence Interval

Duality between Critical Region and Confidence Interval

$$\hat{\beta}_{(k)} \in \mathcal{R}(\alpha) \Leftrightarrow 0 \in \mathcal{I}(e^{(k)}, \hat{\beta}, \widehat{\sigma^2}, \alpha)$$

- Confidence interval: for all $\beta^\star, \sigma_\star$,

$$\mathbb{P}_{\beta^\star,\sigma_\star}\left(\beta^\star, \sigma_\star \in \mathcal{I}(e^{(k)}, \hat{\beta}, \widehat{\sigma^2}, \alpha)\right) = 1 - \alpha$$

- Consider the test $H_0 : \beta^\star_{(k)} = 0, \qquad$ against $\qquad H_1 : \beta^\star_{(k)} \neq 0..$
- By construction

$$\mathbb{P}_{0,\sigma_\star}\left(\beta^\star, \sigma_\star \in \mathcal{I}(e^{(k)}, \hat{\beta}, \widehat{\sigma^2}, \alpha)\right) = 1 - \alpha$$

  which is equivalent to

$$\mathbb{P}_{0,\sigma_\star}\left(\hat{\beta}_{(k)} \in \mathcal{R}(\alpha)\right) = 1 - \alpha$$

p-value of the Student's test

Denote

$$T = \frac{\left|\hat{\beta}_{(k)}\right|}{\sqrt{\widehat{\sigma^2}\left[\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\right]_{k,k}}} \sim |\mathrm{T}(n-p)|$$

Proposition 12 (p-value t-test)**.** *The p-value of the test is*

$$p = Q(T) \quad Q(t) = \mathbb{P}\left(|\mathrm{T}(n-p)| > t\right) .$$

under H_0, $\mathbb{P}_{0,\sigma_\star}(p \geq \alpha) = 1-\alpha$.

Proof

- Critical Region and Confidence Interval:

$$\begin{aligned}
&\hat{\beta}_{(k)} \in \mathcal{R}(\alpha) \\
&\Leftrightarrow \left|\hat{\beta}_{(k)}\right| \leq \mathrm{t}_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2} e^{(k)t}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} e^{(k)}} \\
&\Leftrightarrow 0 \in \left[\hat{\beta}_{(k)} - \mathrm{t}_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2} e^{(k)t}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} e^{(k)}},\right. \\
&\qquad\qquad \left.\hat{\beta}_{(k)} + \mathrm{t}_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2} e^{(k)t}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} e^{(k)}}\right] \\
&\Leftrightarrow 0 \in \mathcal{I}(e^{(k)}, \hat{\beta}, \widehat{\sigma^2}, \alpha)
\end{aligned}$$

Generalization of Student Test

Consider the test

$$H_0 : \ a^t\beta = 0, \qquad \text{against} \qquad H_1 : \ a^t\beta \neq 0.$$

Proposition 13. *The interval*

$$\mathcal{R}(\alpha) = \left\{\beta \in \mathbb{R}^p \ : \ |a^t\beta| \leq C_{n-p}(a, \alpha, \mathbb{X}_{(n)}, \widehat{\sigma^2})\right\}$$

where

$$C_{n-p}(a, \alpha, \mathbb{X}_{(n)}) = \mathrm{t}_{1-\alpha/2}(n-p)\sqrt{\widehat{\sigma^2} a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} a}$$

is the critical region of a test of level α, i.e. for all $\beta^\star \in \mathbb{R}^p$ such that $a^t\beta^\star = 0$ and $\sigma_\star > 0$,

$$\mathbb{P}_{\beta^\star,\sigma_\star}\left(a^t\hat{\beta} \in \mathcal{R}(\alpha)\right) = 1-\alpha$$

p-value of the generalized Student test

Let

$$T = \frac{\left|a^t\hat{\beta}\right|}{\sqrt{\widehat{\sigma^2} a^t \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} a}} \sim |\mathrm{T}(n-p)|$$

Proposition 14 (p-value t-test)**.** *The p-value of the test is*

$$p = Q(T) \quad Q(t) = \mathbb{P}\left(|\mathrm{T}(n-p)| > t\right) .$$

under H_0, i.e. for all $\beta^\star \in \mathbb{R}^p$ and $\sigma_\star > 0$ such that $a^t\beta^\star = 0$, $\mathbb{P}_{\beta^\star,\sigma_\star}(p \geq \alpha) = 1-\alpha$.

Wald Test

- Test for $H_0 = \{\mathbb{M}\beta^\star = a\}$ with $\mathbb{M}$ a matrix of rank $q \leq p$.
- *Prop:* Under H_0,

$$\frac{1}{q\widehat{\sigma^2}} \left\| \left(\mathbb{M} \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \right)^{-1} \mathbb{M}^t \right)^{-1/2} \left(\mathbb{M}\hat{\beta} - a \right) \right\|^2 \sim \mathrm{F}(q, n-p)$$

where $\mathrm{F}(q, n-p)$ is the Fisher law of degrees q and $n-p$.

Critical Region

- Define $\mathcal{R}(\alpha)$ by

$$\begin{aligned} \mathcal{R}(\alpha) = \Bigg\{ t, \left\| \left(\mathbb{M} \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \right)^{-1} \mathbb{M}^t \right)^{-1/2} (t - a) \right\|^2 \\ \leq \mathrm{f}_{1-\alpha}(q, n-p) \Bigg\} \end{aligned}$$

- *Prop:* Under H_0,

$$\mathbb{P}\left(\mathbb{M}\hat{\beta} \in \mathcal{R}(\alpha) \right) = 1 - \alpha$$

p-value

- Let

$$T = \frac{1}{q\widehat{\sigma^2}} \left\| \left(\mathbb{M} \left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \right)^{-1} \mathbb{M}^t \right)^{-1/2} \left(\mathbb{M}\hat{\beta} - a \right) \right\|^2$$

and

$$p = \mathbb{P}\left(\mathrm{F}(q, n-p) > T | T \right).$$

- *Prop:* Under H_0,

$$\mathbb{P}\left(p \geq \alpha \right) = 1 - \alpha$$

- Those two tests are equivalent but the second one allows to choose α at the end...
- *Prop:* For $\mathbb{M} = a^t$, the symmetric Fisher test and Wald test are equivalent.

Fisher Test

Test on embedded models

- Null hypothesis $H_0 = \{\mathbb{Y}_{(n)} = \mathbb{X}_{(n)}\beta^\star + \epsilon_{(n)}\}$
- Alternative hypothesis $H_1 = \{\mathbb{Y}_{(n)} = \mathbb{Z}_{(n)}\gamma^\star + \epsilon_{(n)}\}$ with $\mathrm{span}\{\mathbb{X}_{(n)}\} \subset \mathrm{span}\{\mathbb{Z}_{(n)}\}$ with rank $\mathbb{Z}_{(n)} = q > p$.

- The model of H_0 is a submodel of the model of H_1.
- *Natural idea:*
  - Estimate $\hat{\beta}$ and $\widehat{\gamma}$ by LS.
  - Compare the prediction $\mathbb{X}_{(n)}\hat{\beta}$ and $\mathbb{Z}_{(n)}\widehat{\gamma}$.

Prediction Difference and Fisher Law

- *Prop:* Under H_0,

$$T = \frac{(n-q)\|\mathbb{X}_{(n)}\hat{\beta} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}{(q-p)\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2} \sim \mathrm{F}(q-p, n-q)$$

Tests

- *Critical Region:*

$$\mathcal{R}(\alpha) = \{t, t \leq \mathrm{f}_{1-\alpha}(q-p, n-q)\}$$

- *Prop:* Under H_0,

$$\mathbb{P}\left(T \in \mathcal{R}(\alpha)\right) = 1 - \alpha$$

- *p-value:*

$$p = \mathbb{P}\left(\mathrm{F}(q-p, n-q) > T | T\right)$$

- *Prop:* Under H_0,

$$\mathbb{P}\left(p \geq \alpha\right) = 1 - \alpha$$

- Reject H_0 if the difference of prediction is sufficiently large!
- *Prop:* Equivalent to Wald test of $\mathbb{W}\mathbb{Z}\gamma = 0$ with $\mathbb{W}_{(n)}$ the matrix of a basis of the orthogonal of span$\{\mathbb{X}_{(n)}\}$ in span$\{\mathbb{Z}_{(n)}\}$.

Proof

- By construction, $\mathbb{X}_{(n)}\hat{\beta} = \mathrm{Proj}_{\mathbb{X}_{(n)}} \mathbb{Y}_{(n)}$, $\mathbb{Z}_{(n)}\widehat{\gamma} = \mathrm{Proj}_{\mathbb{Z}_{(n)}} \mathbb{Y}_{(n)}$. Let $\mathbb{W}_{(n)}$ the matrix of a basis of the orthogonal of span$\{\mathbb{X}_{(n)}\}$ in span$\{\mathbb{Z}_{(n)}\}$. One verify that

$$\begin{aligned} \mathbb{Z}_{(n)}\widehat{\gamma} - \mathbb{X}_{(n)}\hat{\beta} &= \mathrm{Proj}_{\mathbb{Z}_{(n)}} \mathbb{Y}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}} \mathbb{Y}_{(n)} \\ &= \mathrm{Proj}_{\mathbb{W}_{(n)}} \mathbb{Y}_{(n)}. \end{aligned}$$

  Similarily,

$$\begin{aligned} \mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma} &= \mathbb{Y}_{(n)} - \mathrm{Proj}_{\mathbb{Z}_{(n)}} \mathbb{Y}_{(n)} \\ &= \mathrm{Proj}_{\mathbb{Z}^{\perp}_{(n)}} \mathbb{Y}_{(n)}. \end{aligned}$$

  Hence under H_0, as $\mathbb{Y}_{(n)} = \mathbb{Z}_{(n)}\gamma^{\star} + \epsilon_{(n)}$ while $\mathrm{Proj}_{\mathbb{W}_{(n)}} \mathbb{Z}_{(n)}\gamma^{\star} = 0$,

$$\mathbb{Z}_{(n)}\widehat{\gamma} - \mathbb{X}_{(n)}\hat{\beta} = \mathrm{Proj}_{\mathbb{W}_{(n)}} \epsilon_{(n)}$$

$$\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma} = \mathrm{Proj}_{\mathbb{Z}_{(n)}^{\perp}} \epsilon_{(n)}$$

As those two space are orthogonal, one obtains that

$$\begin{aligned} \|\mathbb{Z}_{(n)}\widehat{\gamma} - \mathbb{X}_{(n)}\hat{\beta}\|^2 &\sim \sigma_\star^2 \chi^2(q-p) \\ \|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2 &\sim \sigma_\star^2 \chi^2(n-q) \end{aligned}$$

and that they are independent. It follows thus that

$$T = \frac{(n-q)\|\mathbb{X}_{(n)}\hat{\beta} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}{(p-q)\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2} \sim \mathrm{F}(q-p, n-q)$$

Test on $\sigma_\star^2$

- Null hypothesis $H_0 = \{\sigma_\star^2 = \sigma_0^2\}$.
- *Prop:* Under H_0,

$$(n-p)\frac{\widehat{\sigma^2}}{\sigma_0^2} \sim \chi^2(n-p)$$

Tests

- Critical Region:

$$\mathcal{R}(\alpha) = \{t, \chi^2{}_{\alpha/2}(n-p) \leq (n-p)\frac{t}{\sigma_0^2} \leq \chi^2{}_{1-\alpha/2}(n-p)\}$$

- p-value: $T = (n-p)\dfrac{\widehat{\sigma^2}}{\sigma_0^2}$ and

$$p = 2\min\left(\mathbb{P}\left(\chi^2(n-p) > T\middle|T\right), \mathbb{P}\left(\chi^2(n-p) < T\middle|T\right)\right)$$

- More complicated p-value due to the asymmetry of the distribution!

Proof

- Let T' be an independent copy of T:

$$\begin{aligned} &p = 2\min\left(\mathbb{P}\left(T' > T\middle|T\right), \mathbb{P}\left(T' < T\middle|T\right)\right) < \alpha \\ &\Leftrightarrow \min\left(\mathbb{P}\left(T' > T\middle|T\right), \mathbb{P}\left(T' < T\middle|T\right)\right) < \alpha/2 \\ &\Leftrightarrow T > 1 - t_{1-\alpha/2} \text{ or } T \leq t_{\alpha/2} \end{aligned}$$

where we have used that T has a continuous density. Hence

$$\begin{aligned} \mathbb{P}\left(p < \alpha\right) &= \mathbb{P}\left(T > 1 - t_{1-\alpha/2} \text{ or } T \leq t_{\alpha/2}\right) \\ &= \mathbb{P}\left(T > 1 - t_{1-\alpha/2}\right) + \mathbb{P}\left(T \leq t_{\alpha/2}\right) \\ &= \alpha/2 + \alpha/2 = \alpha. \end{aligned}$$

Wald vs Likelihood Ratio Test

Two different systematic construction

- *Wald:*
  - Null hypothesis of type $H_0 = \{\mathbb{M}\beta^\star = c\}$
  - Compute an estimate $\hat{\beta}$.
  - Test if the difference $\mathbb{M}\hat{\beta} - c$ is large enough to reject H_0
- *Likelihood Ratio Test:*
  - Null hypothesis of type $H_0 = \{\beta^\star \in \Theta_0\}$ vs an alternative of type $H_1 = \{\beta^\star \in \Theta_1\}$.
  - Compute the maximum likelihood estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ in the two models.
  - Test whether the deviance $2\left(\log \mathcal{L}\left(\hat{\beta}_1\right) - \log \mathcal{L}\left(\hat{\beta}_0\right)\right)$ is large enough to reject H_0
- *Key:* control of the difference or the deviance under H_0!
- *Rk:* Under mild assumptions,
  - $\hat{\beta}$ is asymptotically Gaussian and thus the difference.
  - The deviance follows asymptotically a $\chi^2()$ of degree $\dim \Theta_1 - \dim \Theta_0$ if $\Theta_0 \subset \Theta_1$

Wald and Student

- Here, $\hat{\beta}$ is Gaussian... but the variance is unknown.
- Use of the Student and the Fisher laws to quantify the difference.
- *Rk* Require to compute only one estimate.

Fisher

- Let $\Theta_0 = \{(\gamma, s^2) \in \mathbb{R}^q \times \mathbb{R}^+, \mathbb{Z}_{(n)}\gamma \in \operatorname{span}\{\mathbb{X}_{(n)}\}\}$, $\Theta_1 = \{(\gamma, s^2) \in \mathbb{R}^q \times \mathbb{R}^+\}$,

$$(\hat{\beta}, \widehat{s_0^2}) = \operatorname*{argmin}_{\Theta_0} \log \mathcal{L}_{\mathbb{Y}_{(n)},\mathbb{Z}_{(n)}}\left(\gamma, s^2\right)$$
$$(\widehat{\gamma}, \widehat{s_1^2}) = \operatorname*{argmin}_{\Theta_1} \log \mathcal{L}_{\mathbb{Y}_{(n)},\mathbb{Z}_{(n)}}\left(\gamma, s^2\right)$$

- *Prop:*

$$\begin{aligned}
&\log \mathcal{L}_{\mathbb{Y}_{(n)},\mathbb{Z}_{(n)}}\left(\widehat{\gamma}, \widehat{s_1^2}\right) - \log \mathcal{L}_{\mathbb{Y}_{(n)},\mathbb{Z}_{(n)}}\left(\hat{\beta}, \widehat{s_0^2}\right) \\
&= n \log \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2} \\
&= n \log \left(1 + \frac{\|\mathbb{Z}_{(n)}\widehat{\gamma} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}\right)
\end{aligned}$$

- Test based either on the asymptotic $\chi^2(n-q)$ behavior...
- Deviance is large when

$$\frac{\|\mathbb{Z}_{(n)}\widehat{\gamma} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}$$

  is large...

Proof

- Thanks to the Gaussian model,

$$-\log \mathcal{L}_{\mathbb{Y}_{(n)},\mathbb{Z}_{(n)}}\left(\gamma, s^2\right) = \frac{n}{2}\log(2\pi\sigma_\star^2) + \frac{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\gamma\|^2}{2\sigma^2}$$

- ML in H_1: this the unconstrained minimizer and thus

$$\begin{aligned}
\widehat{\gamma}_1 &= \widehat{\gamma} \\
\widehat{s}_1^2 &= \frac{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}{n} \\
-\log \mathcal{L}_{\mathbb{Y}_{(n)},\mathbb{Z}_{(n)}}\left(\widehat{\gamma}_1, \widehat{s}_1^2\right) &= \frac{n}{2}\left(\log(2\pi) + \log\frac{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}{n} + 1\right)
\end{aligned}$$

- $H_0 = \{(\gamma, s^2) \in \mathbb{R}^q \times \mathbb{R}^+, \mathbb{Z}_{(n)}\gamma \in \operatorname{span}\{\mathbb{X}_{(n)}\}\}$ so that $\gamma = \left(\mathbb{Z}_{(n)}{}^t\mathbb{Z}_{(n)}\right)^{-1}\mathbb{Z}_{(n)}\mathbb{X}_{(n)}\beta$. We can thus also minimize in β and derives the minimum in γ. It suffices to notice that $\mathbb{Z}_{(n)}\gamma = \mathbb{X}_{(n)}\beta$ to deduce that

$$\begin{aligned}
\widehat{\gamma}_0 &= \left(\mathbb{Z}_{(n)}{}^t\mathbb{Z}_{(n)}\right)^{-1}\mathbb{Z}_{(n)}\mathbb{X}_{(n)}\hat{\beta} \\
\widehat{s}_0^2 &= \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{n} \\
-\log \mathcal{L}_{\mathbb{Y}_{(n)},\mathbb{Z}_{(n)}}\left(\widehat{\gamma}_0, \widehat{s}_0^2\right) &= \frac{n}{2}\left(\log(2\pi) + \log\frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{n} + 1\right)
\end{aligned}$$

- Thus

$$\begin{aligned}
&\log \mathcal{L}_{\mathbb{Y}_{(n)},\mathbb{Z}_{(n)}}\left(\widehat{\gamma}_0, \widehat{s}_0^2\right) - \log \mathcal{L}_{\mathbb{Y}_{(n)},\mathbb{Z}_{(n)}}\left(\widehat{\gamma}_1, \widehat{s}_1^2\right) \\
&= n \log \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2} \\
&= n \log\left(1 + \frac{\|\mathbb{Z}_{(n)}\widehat{\gamma} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}\right)
\end{aligned}$$

1.3.2 Significant Coefficients and ANOVA

R summary

R summary for Galton regression

```
Call:
lm(formula = childHeightC ~ midparentHeight, data = GaltonFamilies)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4992 -1.4956  0.0939  1.5365  9.1303

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     19.91751    2.82130    7.06 3.26e-12 ***
midparentHeight  0.71258    0.04075   17.49  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.244 on 932 degrees of freedom
Multiple R-squared:  0.247,Adjusted R-squared:  0.2462
F-statistic: 305.8 on 1 and 932 DF,  p-value: < 2.2e-16
```

- Lots of information! Lots of tests!

Call

```
Call:
lm(formula = childHeightC ~ midparentHeight, data = GaltonFamilies)
```

- Gives the command used to generate the regression

Residuals

```
Residuals:
    Min      1Q  Median      3Q     Max 
-9.4992 -1.4956  0.0939  1.5365  9.1303 
```

- Gives some quantiles of the empirical residuals $\widehat{\epsilon}_i$

Coefficients

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     19.91751    2.82130    7.06 3.26e-12 ***
midparentHeight  0.71258    0.04075   17.49  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- For each coefficients in the model, R gives
  - `Estimate` : the estimated value,
  - `Std. Error` : the standard error, i.e. the standard deviation of the estimate,
  - `t value` : the value of the Student test statistic in a test of nullity of the coefficients,
  - `Pr(>|t|)`: the corresponding p-values.
  - The result of *tests of levels* .001, .01, .05 and .1 through a character code.

Model

```
Residual standard error: 2.244 on 932 degrees of freedom
Multiple R-squared:  0.247,Adjusted R-squared:  0.2462 
F-statistic: 305.8 on 1 and 932 DF,  p-value: < 2.2e-16
```

- A set of statistic for the whole model.
- `Residual standard error` : the estimate of the standard deviation obtained as the square root of the ususal estimate. The degree of freedom corresponds to $n - p$
- `R-squared` : two R^2, the first one corresponds to

  $$1 - \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{\|\mathbb{Y}_{(n)} - \bar{\mathbb{Y}_{(n)}}\|^2}$$

  and the second an adjusted one (more on this later)
- `F-statistic` : the value of the Fisher test comparing the model with only the intercept with the model used as well as the corresponding p-value.

ANOVA

(One Factor) ANalysis Of VAriance

- *Test of equality of means* of different subgroups.
  - K groups of cardinality n_k such that $\sum_{k=1}^{K} n_k = n$.
  - $Y_{i,k}$ measure of the individual i in the group k
- *Model:*
$$Y_{i,k} = \mu_k + \epsilon_{i,k}$$
where μ_k is the mean of the kth group and $\epsilon_{i,k}$ is a i.i.d. (Gaussian) noise.
- *Hypothesis to test:* $H_0 = \{\mu_1 = \cdots = \mu_K\}$
- *Rk:* Gaussian case amounts to say that we observe some independent
$$Y_{i,k} \sim \mathrm{N}\left(\mu_k, \sigma_\star^2\right).$$
- Two embedded models in competition:
$$\underbrace{Y_{i,k} \sim \mathrm{N}\left(\mu, \sigma_\star^2\right)}_{H_0} \text{ vs } \underbrace{Y_{i,k} \sim \mathrm{N}\left(\mu_k, \sigma_\star^2\right)}_{H_1}$$

Parameter Estimation and Likelihood

- In H_0,
$$\widehat{\mu}_0 = \frac{1}{n}\sum_{i,k} Y_{i,k} \text{ and } \widehat{\sigma^2}_0 = \frac{1}{n}\sum_{i,k} |Y_{i,k} - \widehat{\mu}|^2$$
$$-\log \mathcal{L}\left(\widehat{\mu}_0, \widehat{\sigma^2}_0\right) = \frac{n}{2}\left(\log(2\pi) + \log \widehat{\sigma^2}_0 + 1\right)$$
- In H_1,
$$\widehat{\mu_k}_1 = \frac{1}{n_k}\sum_{i} Y_{i,k} \text{ and } \widehat{\sigma^2}_1 = \frac{1}{n}\sum_{i,k} |Y_{i,k} - \widehat{\mu_k}_1|^2$$
$$-\log \mathcal{L}\left(\widehat{\mu}_1, \widehat{\sigma^2}_1\right) = \frac{n}{2}\left(\log(2\pi) + \log \widehat{\sigma^2}_1 + 1\right)$$

Likelihood Ratio Test

- Difference of - log *Likelihood*:
$$\begin{aligned} T &= 2\big(-\log \mathcal{L}\left(\widehat{\mu}_0, \widehat{\sigma^2}_0\right) + \log \mathcal{L}\left(\widehat{\mu}_1, \widehat{\sigma^2}_1\right)\big) \\ &= n\left(\log \frac{\widehat{\sigma^2}_0}{\widehat{\sigma^2}_1}\right) \\ &= n\left(\log\left(1 + \frac{\sum_k n_k |\widehat{\mu_k}_1 - \widehat{\mu}_0|^2}{\sum_{i,k} |Y_{i,k} - \widehat{\mu_k}_1|^2}\right)\right) \end{aligned}$$
- Asymptotically, $T \xrightarrow{D} \chi^2(|I| - 1)$
- Comparison of a (weighted) variance of the mean with a (weighted) sum of the variance within the group.
- This looks very similar to the formula of Fisher test!

ANOVA and Linear Model

One Factor ANOVA model is a Linear Model

- Model :

$$\begin{aligned} Y_{i,k} &= \mu_k + \epsilon_{i,k} \\ &= \sum_{k'=1}^{K} \mu_{k'} \mathbf{1}_{k'=k} + \epsilon_{i,k} \\ &= \underline{X}_{i,k}{}^t \mu + \epsilon_{i,k} \end{aligned}$$

  with $\underline{X}_{i,k}{}^t = (\mathbf{1}_{k=1}, \ldots, \mathbf{1}_{k=K})$ and $\mu^t = (\mu_1, \ldots, \mu_K)$

Model Matrix

- Matrix rewriting:

$$\mathbb{Y}_{(n)} = \mathbb{X}_{(n)}\mu + \epsilon_{(n)}$$

  with

$$\mathbb{Y}_{(n)} = \begin{pmatrix} Y_{1,2} \\ Y_{2,2} \\ \vdots \\ Y_{n_1,2} \\ Y_{1,3} \\ \vdots \\ Y_{n_{I-1},I-1} \\ Y_{1,I} \\ \vdots Y_{n_I,I} \end{pmatrix} \quad \mathbb{X}_{(n)} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ & & \vdots & & \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ & & \vdots & & \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ & & \vdots & & \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \quad \text{and} \quad \epsilon_{(n)} = \begin{pmatrix} \epsilon_{1,1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{n_1,1} \\ \epsilon_{1,2} \\ \vdots \\ \epsilon_{n_{I-1},I-1} \\ \epsilon_{1,I} \\ \vdots \epsilon_{n_I,I} \end{pmatrix}$$

- Previous test is exactly the classical Fisher!
- *Cor*:

$$\frac{(n-K)\sum_{k=1}^{K} n_k |\widehat{\mu_k}_1 - \widehat{\mu}_0|^2}{(K-1)\sum_{i,k} |Y_{i,k} - \widehat{\mu_k}_1|^2} \sim \mathrm{F}(K-1)n-K$$

ANOVA and Factor

Factor and Dummy Coding

- *Factor:* Categorical variable having K modalities.
- *Dummy Coding:* map $V \in \{1, \ldots, K\}$ to $\underline{X} = \mathcal{C}_d(V) = (\mathbf{1}_{V=1}, \ldots, \mathbf{1}_{V=K})^t \in \mathbb{R}^K$
- *One factor ANOVA:*
  - Simple groupwise model on the mean:

$$Y_{i,k} = \mu_k + \epsilon_{i,k}$$

  - Linear regression on a dummy coded factor having K modalities

$$Y_{i,k} = \underline{X}_{i,k}{}^t \mu + \epsilon_{i,k}$$

    with $\epsilon_{i,k}$ i.i.d. $\mathrm{N}(0, \sigma_\star^2)$.

- Fisher test of $\mu_1 = \cdots = \mu_K$.

Overparametrized Model

- A more structured model is given by:

$$\begin{aligned} Y_{i,k} &= \mu. + \mu_{k_1} + \epsilon_{i,k} \\ &= \begin{pmatrix} 1 \\ \mathcal{C}_d(V) \end{pmatrix}^t \mu + \epsilon_{i,k} \end{aligned}$$

- Overparametrized model as $\sum_{k=1}^K \mathcal{C}_d(V)^{(k)} = 1$
- Need to add some constrains to the coefficients or in the coding scheme.

Coefficients Constrains

- Add a linear constrain not satisfied by $\mu_1 = \cdots = \mu_K$ except for $\mu_1 = \cdots = \mu_k = 0$:
  - First modality ref: $\mu_1 = 0$
  - Zero average: $\sum_{k=1}^K \mu_k = 0$
  - General linear constraint : $a^t\mu = 0$ with $a^t\mathbb{1}_{(K)} \neq 0$.
- Formulation:

$$\widehat{\mu} = \operatorname*{argmin}_{a^t\mu=0} \|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\mu\|^2$$

with

$$\mathbb{X}_{(n)} = \begin{pmatrix} 1 & \mathcal{C}_d(V_{1,1})^t \\ \vdots & \vdots \\ 1 & \mathcal{C}_d(V_{n_K,K})^t \end{pmatrix}$$

- Require a minimization under a linear constraint...

Coding Constrains

- Avoid the over parameterization: $\operatorname{span}\{\mathcal{C}(V_1)_{(n)}\} \cap \operatorname{span}\{\mathbb{1}_{(n)}\} = \{\mathbb{0}_{(n)}\}$
  - Use the first modality as a ref.: $\mathcal{C}(V) = (\mathbf{1}_{X=2}, \ldots, \mathbf{1}_{X=K})^t$
  - Use a diff. effect: $\mathcal{C}(V) = (\mathbf{1}_{X=2} - \mathbf{1}_{X=1}, \ldots, \mathbf{1}_{X=K} - \mathbf{1}_{X=1})^t$
  - More generally, use

$$\mathcal{C}(V) = \mathbb{M}_c{}^t\mathcal{C}_d(V)$$

with $\operatorname{rank}\mathbb{M}_c = K - 1$ and $\mathbb{1}_{(K)} \notin \operatorname{span}\mathbb{M}_c$

- Formulation:

$$\widehat{\beta} = \operatorname{argmin} \|\mathbb{Y}_{(n)} - \mathbb{X}'_{(n)}\beta\|^2$$

with

$$\mathbb{X}'_{(n)} = \begin{pmatrix} 1 & \mathcal{C}(V_{1,1})^t \\ \vdots & \vdots \\ 1 & \mathcal{C}(V_{n_K,K})^t \end{pmatrix}$$

- Non constraint optimization.
- Equivalent results if $a \in \operatorname{span}\{\mathbb{M}_c\}^\perp$!

ANOVA and Factors

- How to handle two (or more) categorical variables?
- Let V_1 be a categorical variable with K_1 modalities and V_2 a categorical variable with K_2 modalities.

The Dummy Coding Model

- A natural model is

$$\begin{aligned} Y_{i,k_1,k_2} &= \mu_{k_1,k_2} + \epsilon_{i,k_1,k_2} \\ &= \mathcal{C}_d(V_1,V_2)^t \mu + \epsilon_{i,k_1,k_2} \end{aligned}$$

which amounts to consider the categorical variable $V=(V_1,V_2)$ with $K_1 \times K_2$ modalities...

- *Issue:* No easy way to test if V_1 (or V_2) has no influence...

Overparametrized Model

- A more structured model is given by:

$$\begin{aligned} Y_{i,k_1,k_2} &= \mu_{\cdot,\cdot} + \mu_{k_1,\cdot} + \mu_{k_2,\cdot} + \mu_{k_1,k_2} + \epsilon i, k_1, k_2 \\ &= \left(1, \mathcal{C}_d(V_1)^t, \mathcal{C}_d(V_2)^t, \mathcal{C}_d(V_1,V_2)^t\right) \mu + \epsilon_{i,k_1,k2} \end{aligned}$$

- Overparametrized model as $\sum_{k=1}^K \mathcal{C}_d(V)^{(k)} = 1$ for $V=V_1$, $V=V_2$ and $V=(V_1,V_2)$
- Requires *constrains* on the coeff. or in the coding scheme.
- *Coding:* Same idea than in the single factor case but with a *hierarchical structure*:
  - Code the individual variable as if there was a single variable.
  - Code the interaction by a code $\mathbb{M}_c\mathcal{C}(V_1,V_2)$ of rank $K_1 \times K_2 - (K_1-1) - (K_2-1) - 1$ and which is orthogonal to the space spanned by $\mathcal{C}(V_1)$ and $\mathcal{C}(V_2)$.
- Tedious *linear algebra*... handled automatically and efficiently by R or any statistical software.

1.3.3 Asymptotic Analysis

Non Gaussian Case

- Previous analysis only valid under Gaussianity.
- *Very strong assumption!*
- Can we do something with the second order model?

Asymptotic Analysis

- If there is a limiting distribution for $(\hat{\beta} - \beta^\star)$ a proper normalization one can construct asymptotic test based on the limit law!
- *Main result:* Under some (mild) assumptions, if $Q_n = \frac{1}{n}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right) \xrightarrow{P} Q$ s.d.p. then

$$\sqrt{n}\frac{Q_n^{-1/2}(\hat{\beta} - \beta^\star)}{\sqrt{\widehat{\sigma^2}}} \xrightarrow{D} \mathrm{N}\left(\mathbb{0}_{(p)}, \mathrm{Id}_{(p)}\right)$$

- Asymptotic Wald test and Asymptotic Likelihood Ratio test...

Asymptotic in n

What's going on when n goes to $+\infty$?

- Does $\hat{\beta}$ converge to $\beta^\star$?
- Is there a limiting distribution for $(\hat{\beta} - \beta^\star)$ and what is the proper normalization?

- *Missing information:* behavior of $\mathbb{X}_{(n)}$ when n goes to $+\infty$.

Two models for $\mathbb{X}_{(n)}$

- Fixed design:
  - No statistical model for $\underline{X}_i$ / Results conditionally to $\mathbb{X}_{(n)}$
  - Need to impose some restrictions on the asymptotic behavior of $\mathbb{X}_{(n)}$.
- Random design:
  - $\underline{X}_i$ are i.i.d. / Probabilistic results.
  - Need to impose some (mild) restriction on the law of $\underline{X}_i$.

Consistency

- *Prop:*

$$\mathbb{E}\left[\|\hat{\beta} - \beta^\star\|^2 \middle| \mathbb{X}_{(n)}\right] = \sigma_\star^2 \operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\right)$$

Consistency

- *Cor:* (Fixed design)

$$\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\right) \xrightarrow[n\to+\infty]{} 0 \Rightarrow \hat{\beta}|\mathbb{X}_{(n)} \xrightarrow[n\to+\infty]{L^2} \beta^\star$$

- *Cor:* (Random Design) If $\operatorname{tr}\left(\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\right) \xrightarrow{P} 0$ then

$$\hat{\beta} \xrightarrow[n\to+\infty]{P} \beta^\star$$

- *Rk:* The L^2 implies also the convergence in probability.
- *Random design:* A sufficient condition for the convergence in probability is given by $\mathbb{E}\left[\underline{X}\underline{X}^t\right]$ definite positive.

Gaussian Limit

- To obtain a (Gaussian) limit law, one should add more assumptions:
  - ϵ_i independent
  - $Q_n = \frac{1}{n}\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right) \xrightarrow{P} Q$ s.d.p.
  - $\dfrac{1}{n^{1+\delta/2}} \displaystyle\sum_{i=1}^{n} \mathbb{E}\left[\|\underline{X}_i\|^{2+\delta}\|\epsilon_i\|^{2+\delta}\right] \to 0$ (Technical condition!)

Gaussian Limit

- *Thm:* Under the previous assumption,

$$\sqrt{n}\left(\hat{\beta}-\beta^\star\right)\Big|\mathbb{X}_{(n)} \xrightarrow{D} \mathrm{N}\left(0, \sigma_\star^2 Q^{-1}\right)$$

- *Cor:*

$$\sqrt{n} Q_n^{1/2}\left(\hat{\beta}-\beta^\star\right)\Big|\mathbb{X}_{(n)} \xrightarrow{D} \mathrm{N}\left(0, \sigma_\star^2 \mathrm{Id}_{(p)}\right)$$

- Quite technical proof!

Proof

- First, notice that

$$\begin{aligned} \mathbb{E}\left[\hat{\beta}\middle|\mathbb{X}_{(n)}\right] &= \beta^\star \\ \mathbb{V}\mathrm{ar}\left[\hat{\beta}\middle|\mathbb{X}_{(n)}\right] &= \sigma_\star^2\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \end{aligned}$$

  so that

$$\begin{aligned} \mathbb{E}\left[\sqrt{n}\left(\hat{\beta}-\beta^\star\right)\middle|\mathbb{X}_{(n)}\right] &= 0 \\ \mathbb{E}\left[\sqrt{n}\left(\hat{\beta}-\beta^\star\right)\middle|\mathbb{X}_{(n)}\right] &= \sigma_\star^2\left(\frac{1}{n}\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1} \end{aligned}$$

  which gives an hint on the limiting mean and variance.

- Now,

$$\begin{aligned} \sqrt{n}\left(\hat{\beta}-\beta^\star\right) &= \sqrt{n}\left(\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t \mathbb{Y}_{(n)} - \beta^\star\right) \\ &= \sqrt{n}\left(\left(\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\mathbb{X}_{(n)}^t \epsilon_{(n)}\right) \\ &= \left(\frac{1}{n}\mathbb{X}_{(n)}^t \mathbb{X}_{(n)}\right)^{-1}\sqrt{n}\left(\frac{1}{n}\mathbb{X}_{(n)}^t \epsilon_{(n)}\right) \end{aligned}$$

- By Slutsky theorem, as $\frac{1}{n}\mathbb{X}_{(n)}^t \mathbb{X}_{(n)} \xrightarrow{P} Q$, it suffices thus to prove that $\frac{1}{n}\mathbb{X}_{(n)}^t \epsilon_{(n)} \xrightarrow{D} \mathrm{N}\left(\mathbb{0}_{(p)}, \sigma_\star^2 Q\right)$ to obtain the result.

- This in turn is equivalent to prove that $\forall a \in \mathbb{R}^p$,

$$\sqrt{n}\left(\frac{1}{n} a^t \mathbb{X}_{(n)}^t \epsilon_{(n)}\right) \xrightarrow{D} \mathrm{N}\left(0, \sigma_\star^2 a^t Q a\right)$$

- Now

$$\begin{aligned}\frac{1}{n}a^t\mathbb{X}_{(n)}^t\epsilon_{(n)} &= \frac{1}{n}\mathbb{X}_{(n)}a^t\epsilon_{(n)}\\ &= \frac{1}{n}\sum_{i=1}^n \underline{X}_i{}^t a\epsilon_i\end{aligned}$$

- If the $\underline{X}_i$ were i.i.d. the results would thus be direct application of the TCL. Here we should use a stronger TCL type result that ensures a Gaussian limiting behavior for a sum of only independent variables.

- We rely on the Lindeberg theorem and Lyapunov condition:
  - *Thm:* Assume Z_i is a sequence of independent random variable such that $\mathbb{E}[Z_i] = 0$, $\mathbb{V}\mathrm{ar}[Z_i] = [\sigma_\star^2]_i$ and, letting

$$\begin{aligned}s_n^2 &= \frac{1}{n}\sum_{i=1}^n \sigma_i^2\\ \kappa_n(\epsilon) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[|Z_i|^2\mathbf{1}_{|Z_i|>\sqrt{n}\epsilon}\right],\end{aligned}$$

    if $s_n^2 \xrightarrow{P} \sigma_\star^2$ and for all $\epsilon, \kappa_n(\epsilon) \to 0$ then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Z_i\right) \xrightarrow{D} \mathrm{N}(0,\sigma_\star^2).$$

  - *Lyapunov condition:* If, letting

$$\gamma_n = \frac{1}{n^{1+\delta/2}}\sum_{i=1}^n \mathbb{E}\left[|Z_i|^{2+\delta}\right],$$

    $\gamma_n \to 0$ then $\forall\epsilon, \kappa_n(\epsilon) \to 0$.

- Proof of the Lyapunov condition:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[|Z_i|^2\mathbf{1}_{|Z_i|>\sqrt{n}\epsilon}\right] &\leq \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}\left[|Z_i|^{2+\delta}\right]}{(\sqrt{n}\epsilon)^\delta}\\ &\leq \frac{1}{\epsilon^\delta}\frac{1}{n^{1+\delta/2}}\sum_{i=1}^n \mathbb{E}\left[|Z_i|^{2+\delta}\right] = \frac{\gamma_n}{\epsilon^\delta}\end{aligned}$$

- Let $Z_i = \underline{X}_i{}^t a\epsilon_i$. They are independent and they satisfy

$$\mathbb{E}\left[Z_i\middle|\mathbb{X}_{(n)}\right] = 0 \qquad [\sigma_\star^2]_i = \mathbb{V}\mathrm{ar}\left[Z_i\middle|\mathbb{X}_{(n)}\right] = \sigma_\star^2 a^t\underline{X}_i\underline{X}_i{}^t a.$$

  Now

$$\begin{aligned}s_n^2 &= \frac{1}{n}\sum_{i=1}^n [\sigma_\star^2]_i\\ &= \frac{1}{n}\sum_{i=1}^n \sigma_\star^2\left(a^t\underline{X}_i\underline{X}_i{}^t a\right)\\ &= \sigma_\star^2 a^t\left(\frac{1}{n}\sum_{i=1}^n \underline{X}_i\underline{X}_i{}^t\right)a\\ &= \sigma_\star^2 a^t\left(\frac{1}{n}\mathbb{X}_{(n)}^t\mathbb{X}_{(n)}\right)a \xrightarrow{P} a^tQa\end{aligned}$$

while

$$\begin{aligned}\gamma_n &= \frac{1}{n^{1+\delta/2}} \sum_{i=1}^{n} \mathbb{E}\left[|Z_i|^{2+\delta}\right] \\ &= \frac{1}{n^{1+\delta/2}} \sum_{i=1}^{n} \mathbb{E}\left[|\underline{X}_i{}^t a \epsilon_i|^{2+\delta}\right] \\ &\leq \frac{a^{2+\delta}}{n^{1+\delta/2}} \sum_{i=1}^{n} \mathbb{E}\left[\|\underline{X}_i\|^{2+\delta}|\epsilon_i|^{2+\delta}\right]\end{aligned}$$

which goes to 0 by hypothesis.

Variance Estimate

- Estimate:

$$\widehat{\sigma^2_m} = \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{m}$$

Properties

- *Thm:* Under the assumption of the Gaussian limit thm, if $m/n \to 1$ then

$$\widehat{\sigma^2_m}\Big|\mathbb{X}_{(n)} \xrightarrow{P} \sigma^2_\star.$$

- If furthermore $\mathbb{E}\left[|\epsilon|^4\right] < +\infty$

$$\sqrt{n}\left(\widehat{\sigma^2}_m - \sigma^2_\star\right) \xrightarrow{D} \mathrm{N}\left(0, \mathbb{V}\mathrm{ar}\left[|\epsilon|^2\right]\right)$$

- *Cor:* Under the assumption of the Gaussian limit thm, if $\sqrt{n}\left(\frac{n}{m} - 1\right) \to 0$ then

$$\sqrt{n}\frac{Q_n^{1/2}\left(\hat{\beta} - \beta^\star\right)}{\sqrt{\widehat{\sigma^2}_m}} \xrightarrow{D} \mathrm{N}\left(\mathbb{0}_{(p)}, \mathrm{Id}_{(p)}\right)$$

Proof

- By definition,

$$\begin{aligned}\widehat{\sigma^2_m} &= \frac{\|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}\hat{\beta}\|^2}{m} \\ &= \frac{\|\left(\mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\right)\mathbb{Y}_{(n)}\|^2}{m} \\ &= \frac{\|\left(\mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}_{(n)}}\right)\epsilon_{(n)}\|^2}{m} \\ &= \frac{\|\epsilon_{(n)}\|^2}{m} - \frac{\|\mathrm{Proj}_{\mathbb{X}_{(n)}} c\epsilon_{(n)}\|}{m} \\ &= \frac{\|\epsilon_{(n)}\|^2}{m} - \frac{\|\mathbb{X}_{(n)}\left(\hat{\beta} - \beta^\star\right)\|^2}{m}\end{aligned}$$

Now

$$\frac{\|\epsilon_{(n)}\|^2}{m} = \frac{n}{m}\frac{1}{n}\sum_{i=1}^{n}|\epsilon_i|^2$$

where $n/m \to 1$ by hypothesis while $\frac{1}{n}\sum_{i=1}^{n}|\epsilon_i|^2 \xrightarrow{P} \mathbb{E}\left[\epsilon\right] = \sigma_\star^2$ thanks to the Law of Large Numbers so that

$$\frac{\|\epsilon_{(n)}\|^2}{m} \xrightarrow{P} \sigma_\star^2$$

It remains to notice that thank to the Gaussian limit

$$\sqrt{n}\left(\frac{1}{\sqrt{n}}\mathbb{X}_{(n)}\right)\left(\hat{\beta}-\beta^\star\right) \xrightarrow{D} \mathrm{N}\left(0, \sigma_\star^2\,\mathrm{Proj}_{\mathbb{X}_{(n)}}\right)$$

i.e.

$$\mathbb{X}_{(n)}\left(\hat{\beta}-\beta^\star\right) \xrightarrow{D} \mathrm{N}\left(0, \sigma_\star^2\,\mathrm{Proj}_{\mathbb{X}_{(n)}}\right)$$

which implies

$$\|\mathbb{X}_{(n)}\left(\hat{\beta}-\beta^\star\right)\|^2 \xrightarrow{D} \|\,\mathrm{N}\left(0, \sigma_\star^2\,\mathrm{Proj}_{\mathbb{X}_{(n)}}\right)\|^2 = \sigma_\star^2\,\chi^2(p)$$

which leads to

$$\frac{\|\mathbb{X}_{(n)}\left(\hat{\beta}-\beta^\star\right)\|^2}{m} \xrightarrow{P} 0$$

We can now conclude

$$\begin{aligned}\widehat{\sigma^2_m} &= \frac{\|\epsilon_{(n)}\|^2}{m} - \frac{\|\mathbb{X}_{(n)}\left(\hat{\beta}-\beta^\star\right)\|^2}{m}\\ &\xrightarrow{P} \sigma_\star^2 - 0 = \sigma_\star^2\end{aligned}$$

- Proof:

$$\begin{aligned}\sqrt{n}\left(\widehat{\sigma^2_m} - \sigma_\star^2\right) = \sqrt{n}\left(\frac{\|\epsilon_{(n)}\|^2}{n} - \sigma_\star^2\right) + \sqrt{n}\left(\frac{n-m}{m}\right)\frac{\|\epsilon_{(n)}\|^2}{n}\\ + \sqrt{n}\frac{\|\mathbb{X}_{(n)}(\hat{\beta}-\beta^\star)\|^2}{m}\end{aligned}$$

For the first term,

$$\frac{\|\epsilon_{(n)}|^2}{n} = \frac{1}{n}\sum_{i=1}^{n}|\epsilon_i|^2$$

hence

$$\sqrt{n}\left(\frac{\|\epsilon_{(n)}\|^2}{n} - \sigma_\star^2\right) \xrightarrow{D} \mathrm{N}\left(0, \mathbb{V}\mathrm{ar}\left[|\epsilon|^2\right]\right)$$

For the second term

$$\sqrt{n}\left(\frac{n-m}{m}\right)\frac{\|\epsilon_{(n)}\|^2}{n},$$

By the LLN $\frac{\|\epsilon_{(n)}\|^2}{n} \xrightarrow{P} \sigma_\star^2$ while by hypothesis $\sqrt{n}\left(\frac{n-m}{m}\right) \to 0$ hence

$$\sqrt{n}\left(\frac{n-m}{m}\right)\frac{\|\epsilon_{(n)}\|^2}{n} \xrightarrow{P} 0$$

For the last term,

$$\sqrt{n}\frac{\|\mathbb{X}_{(n)}(\hat{\beta}-\beta^\star)\|^2}{m} = \frac{\sqrt{n}}{m}\|\mathbb{X}_{(n)}(\hat{\beta}-\beta^\star)\|^2$$

we have seen that $\|\mathbb{X}_{(n)}(\hat{\beta}-\beta^\star)\|^2 \xrightarrow{D} \sigma_\star^2\chi^2(p)$ while $\sqrt{n}\left(\frac{n-m}{m}\right) \to 0$ implies $\frac{\sqrt{n}}{m} \to 0$ so that

$$\sqrt{n}\frac{\|\mathbb{X}_{(n)}(\hat{\beta}-\beta^\star)\|^2}{m} \xrightarrow{P} 0$$

which concludes the proof.

- For the corrolary, it suffices to notice that

$$\sqrt{n}\frac{Q_n^{1/2}\left(\hat{\beta}-\beta^\star\right)}{\sigma_\star} \xrightarrow{D} \mathrm{N}\left(\mathbb{0}_{(p)}, \mathrm{Id}_{(p)}\right)$$

while $\sqrt{\widehat{\sigma^2}_m} \xrightarrow{P} \sigma_\star$ so that thanks to Slutsky Lemma

$$\sqrt{n}\frac{Q_n^{1/2}\left(\hat{\beta}-\beta^\star\right)}{\sqrt{\widehat{\sigma^2}_m}} \xrightarrow{D} \mathrm{N}\left(\mathbb{0}_{(p)}, \mathrm{Id}_{(p)}\right)$$

Asymptotic Tests

Student Test

- For any $a \in \mathbb{R}^p$,

$$\sqrt{n}\frac{\left(a^tQ_n^{-1}a\right)^{-1/2}\left(a^t\hat{\beta}-a^t\beta^\star\right)}{\sqrt{\widehat{\sigma^2}}} \xrightarrow{D} \mathrm{N}\left(\mathbb{0}_{(p)}, \mathrm{Id}_{(p)}\right)$$

- Test of asympotic level α:

$$\left|\sqrt{n}\frac{\left(a^tQ_n^{-1}a\right)^{-1/2}\left(a^t\hat{\beta}-a^t\beta^\star\right)}{\sqrt{\widehat{\sigma^2}}}\right| \leq \mathrm{N}_{1-\alpha/2}$$

with $\mathrm{N}_{\alpha'}$ the quantile α' of a $\mathrm{N}(0,1)$

- Asymptotic p-value:

$$p = \mathbb{P}\left(|\,\mathrm{N}(0,1)| > \left|\sqrt{n}\frac{\left(a^tQ_n^{-1}a\right)^{-1/2}\left(a^t\hat{\beta}-a^t\beta^\star\right)}{\sqrt{\widehat{\sigma^2}}}\right|\right)$$

- No need to use the Student correction...

Asymptotic Wald Test

- Let $\mathbb{M}$ be a $q \times p$ matrix of rank q:

$$\frac{n}{\widehat{\sigma^2}}\left\|\left(\mathbb{M}Q_n^{-1}\mathbb{M}^t\right)^{-1/2}\left(\mathbb{M}\hat{\beta}-\mathbb{M}\beta^\star\right)\right\|^2 \xrightarrow{D} \chi^2(q)$$

- Test of asymptotic level α:

$$\frac{n}{\widehat{\sigma^2}} \left\| \left(\mathbb{M} Q_n^{-1} \mathbb{M}^t\right)^{-1/2} \left(\mathbb{M}\hat{\beta} - \mathbb{M}\beta^\star\right) \right\|^2 \leq {\chi^2}_{1-\alpha}(q)$$

  where ${\chi^2}_{\alpha'}(d)$ is the quantile α' of a $\chi^2(d)$

- Asymptotic p-value:

$$p = \mathbb{P}\left(\chi^2(q) > \frac{n}{\widehat{\sigma^2}} \left\| \left(\mathbb{M} Q_n^{-1} \mathbb{M}^t\right)^{-1/2} \left(\mathbb{M}\hat{\beta} - \mathbb{M}\beta^\star\right) \right\|^2\right)$$

- No need to use Fisher correction...

Asymptotic Fisher Test

- If $\mathrm{span}\{\mathbb{X}_{(n)}\} \subset \mathrm{span}\{\mathbb{Z}_{(n)}\}$

$$\frac{(n-q)\|\mathbb{X}_{(n)}\hat{\beta} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2} \xrightarrow{D} \chi^2(q-p)$$

- Test of asymptotic level α:

$$\frac{(n-q)\|\mathbb{X}_{(n)}\hat{\beta} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2} \leq {\chi^2}_{1-\alpha}(q-p)$$

- Asymptotic p-value:

$$p = \mathbb{P}\left(\chi^2(q-p) > \frac{(n-q)\|\mathbb{X}_{(n)}\hat{\beta} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}{\|\mathbb{Y}_{(n)} - \mathbb{Z}_{(n)}\widehat{\gamma}\|^2}\right)$$

- No need to use Fisher correction...

Chapter 2

Non Parametric Setting and Model Selection

2.1 Projection Based Linear Regression

2.1.1 Linear Regression and Models

Non Parameteric POV and Models

Parametric POV

- *Observation:* $Y_i = f_{\beta^\star}(\underline{X}_i) + \epsilon_i = \underline{X}_i{}^t \beta^\star + \epsilon_i$
- *Parametric model assumptions:*
  - $\{\epsilon_i\}_{i=1}^n$ are i.i.d.
  - $\mathbb{E}\left[Y_i | \underline{X}_i\right] = \underline{X}_i{}^t \beta^\star$ for an unknown $\beta^\star$
- *Estimation:* estimation of $\beta^\star$ and hence of $f_{\beta^\star}$.

- This is a model for the observation...

All Models are Wrong but Some are Useful

G. Box (1919-2013)

- Now it would be very *remarkable* if any system existing in the real world could be *exactly represented* by any simple model.
- However, cunningly chosen parsimonious models often do provide remarkably *useful approximations.*
- For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an *ideal* gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules.
- For such a model there is no need to ask the question *Is the model true?.* If *truth* is to be the *whole truth* the answer must be *No.* The only question of interest is *Is the model illuminating and useful?*

Non Parametric POV

Non parametric POV

- *Observation:*
$$Y_i = f^\star(\underline{X}_i) + \epsilon_i$$

- *Non Parametric model assumption:*
  - $\{\epsilon_i\}_{i=1}^n$ are i.i.d.

- *Estimation:* estimation of $f^\star$ through a model of type f_β

- Rk: $f^\star(\underline{X}_i) = \mathbb{E}\left[Y_i|\underline{X}_i\right]$ as soon as $\mathbb{E}\left[\epsilon_i|\underline{X}_i\right] = 0$.

- We will in the sequel no longer assume that the model(s) is (are) correct, i.e. we do not assume that there exists $\beta^\star$ such that $\mathbb{E}\left[Y_i|\underline{X}_i\right] = \underline{X}_i{}^t\beta^\star$!

- *Models are only approximations of the truth!*

Linear Models

Linear models

- Linear models:
$$f(\underline{X}) = \Phi\left(\underline{X}\right)^t\beta$$
with Φ a suitable transformation.

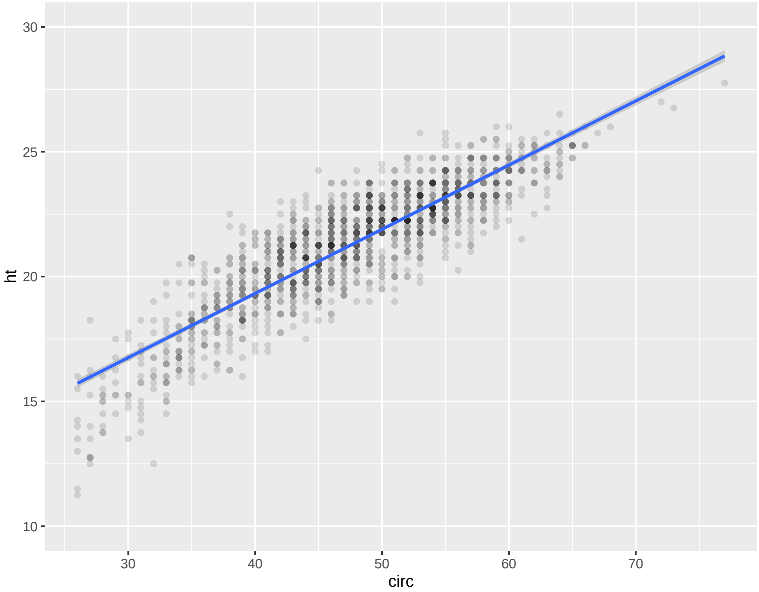
- *Pros:*
  - simple to fit,
  - easy to understand,
  - wide variety of useful techniques for testing the assumptions involved

- *Rk:* There are obviously cases when the linear model fails, for instance because of an intrinsic nonlinearity in the data!

- *Nonparametric regression* (a topic not covered in this course) provides a means for modeling such data.

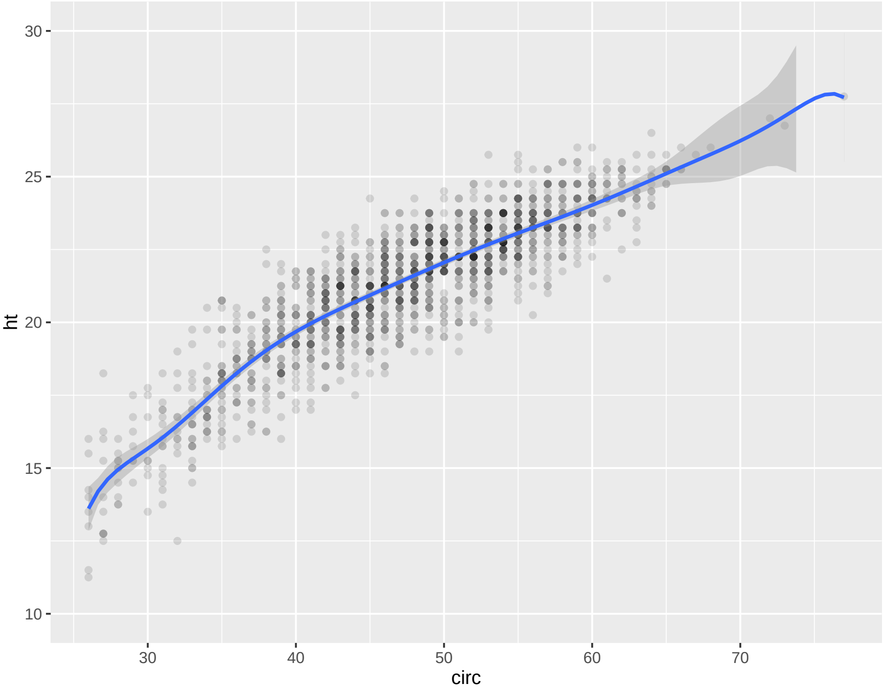
- *Rk:* Nonparametric regression can be used as a benchmark for linear models against which to test the linearity assumption.

Example 1: Polynomial Regression



- Polynomial model: $f_\beta(\underline{X}_i) = \sum_{l=1}^{p} \beta_l \underline{X}_i^{l-1}$
- Linear in β!
- Amounts to use $\underline{X}_i^\Phi = \Phi\left(\underline{X}_i\right) = \left(1, \underline{X}_i, \ldots, \underline{X}_i^{p-1}\right)^t$
- Easy least squares estimation!

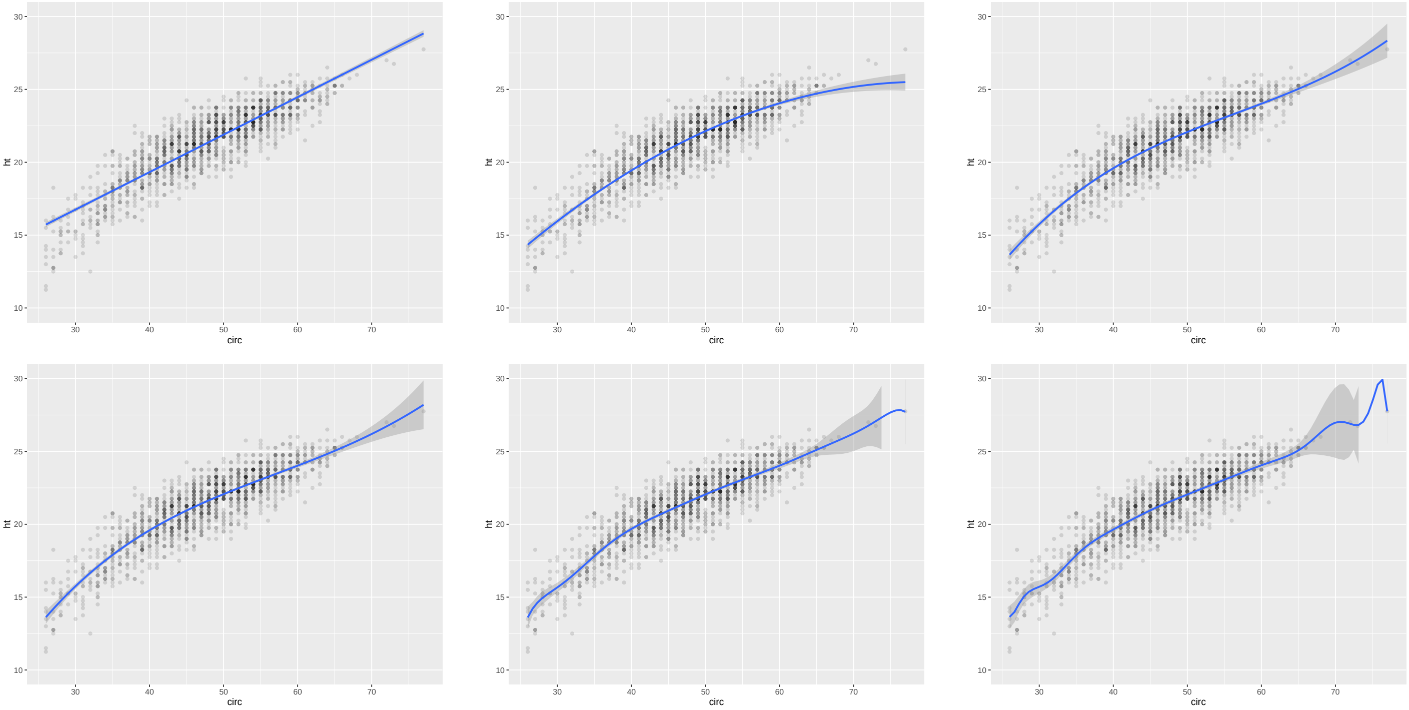
Example 1: Which Degree?



Models

- Increasing degree = increasing complexity and better fit on the data

Example 1: Which Degree?



Best Degree?

- How to choose among those solution?

Example 2: 1D Regression in a Dictionary

- $Y = f^\star(X) + \epsilon$ with $X \in [0,1]$.
- Estimation in a dictionary of functions ϕ_j by

$$f_\beta(X) = \sum_{j=1}^{p} \beta_j \phi_j(X)$$

- For instance the d first elements of the Fourier basis or a polynomial basis...
- Linear model with $\underline{X}^{\Phi} = \Phi\left(X\right) = \left(\phi_1(X), \ldots, \phi_p(X)\right)^t$:

$$f_\beta(X) = f_{\Phi,\beta}(X) = \underline{X}^{\Phi\,t}\beta = \Phi\left(X\right)^t\beta$$

- We may also assume that $\underline{X}_i = \frac{i}{n}$...

Models

- Choice of dictionary,
- Choice of p.

Example 3: Birth Weight Regression

Dataset

- Study on risk factors associated with low infant birth weight conducted at Baystate Medical Center, Springfield, Mass during 1986
- 189 observations of 10 variables

Goal

- Predict the birth weight from those variables.
- Infer the one having an impact by looking at the coefficients!

Variables

- Raw data:

Variable	Content
low	indicator of birth weight less than 2.5 kg
age	mother's age in years
lwt	mother's weight in pounds at last menstrual period
cat.	mother's ethnic category (1 = european descent, 2 = african decent, 3 = other)
smoke	smoking status during pregnancy
ptl	number of previous premature labors
ht	history of hypertension
ui	presence of uterine irritability
ftv	number of physician visits during the first trimester
bwt	birth weight in grams

- Preprocessed data:
  - define $\underline{X}$ by removing low (and bwt), replacing the ethnic category (cat.) by two variables and adding a constant variable equal to 1.
  - set $Y = \text{bwt}$

Linear Model

- Model $f_\beta(\underline{X}) = \underline{X}^t\beta =$

$$\begin{aligned} f_\beta(\underline{X}) = {} & \beta_1 \times 1 + \beta_2 \times \text{age} + \beta_3 \times \text{lwt} \\ & + \beta_4 \times \text{category==1} + \beta_5 \times \text{category==2} \\ & + \beta_6 \times \text{smoke} + \beta_7 \times \text{ptl} + \beta_8 \times \text{ht} + \beta_9 \times \text{ui} + \beta_{10} \times \text{ftv} \end{aligned}$$

- Each coordinate of β corresponds to a variable.

Variable Selection

- Why not using only some variables?
  - The smaller the number of variables the easier the estimation...
  - If a variable is (almost) useless it can be safely removed...
  - Important variables are the ones useful in prediction...
- Transformations: for a subset I of $\{1, \dots, 10\}$

$$\underline{X}^\Phi = \Phi\left(\underline{X}\right) = (\underline{X}^{(k)})^t_{k \in I}$$

- Models: $f_{\Phi,\beta}(\underline{X}) = \Phi\left(\underline{X}^\Phi\right)\beta$.

Example 4: Variable Selection

Linear Model

- Model $f_\beta(\underline{X}) = \underline{X}^t\beta$
- Each coordinate in β corresponds to a variable.

Variable Selection

- Why not using only some variables?
  - The smaller the number of variables the easier the estimation...
  - If a variable is (almost) useless it can be safely removed...
  - Important variables are the ones that are useful to obtain a good prediction...
- Transformation(s): for a subset I of $\{1,\ldots,p\}$

$$\underline{X}^\Phi = \Phi\left(\underline{X}\right) = (\underline{X}^{(k)})^t_{k\in I}$$

- Model(s):

$$f_{\Phi,\beta}(\underline{X}) = \underline{X}^{\Phi t}\beta = \Phi\left(\underline{X}\right)^t\beta$$

  with $\beta \in \mathbb{R}^{|I|}$.

Example 5: Transformed Representation and Feature Design

- Linear model $f_\beta(\underline{X}) = \underline{X}^t\beta$.

Transformed Representation

- From $\underline{X}$ to $\underline{X}^\Phi = \Phi(\underline{X})$!
- New description of $\underline{X}$ leads to a different *linear* model:

$$f_{\Phi,\beta}(\underline{X}) = \underline{X}^{\Phi t}\beta = \Phi\left(\underline{X}\right)^t\beta$$

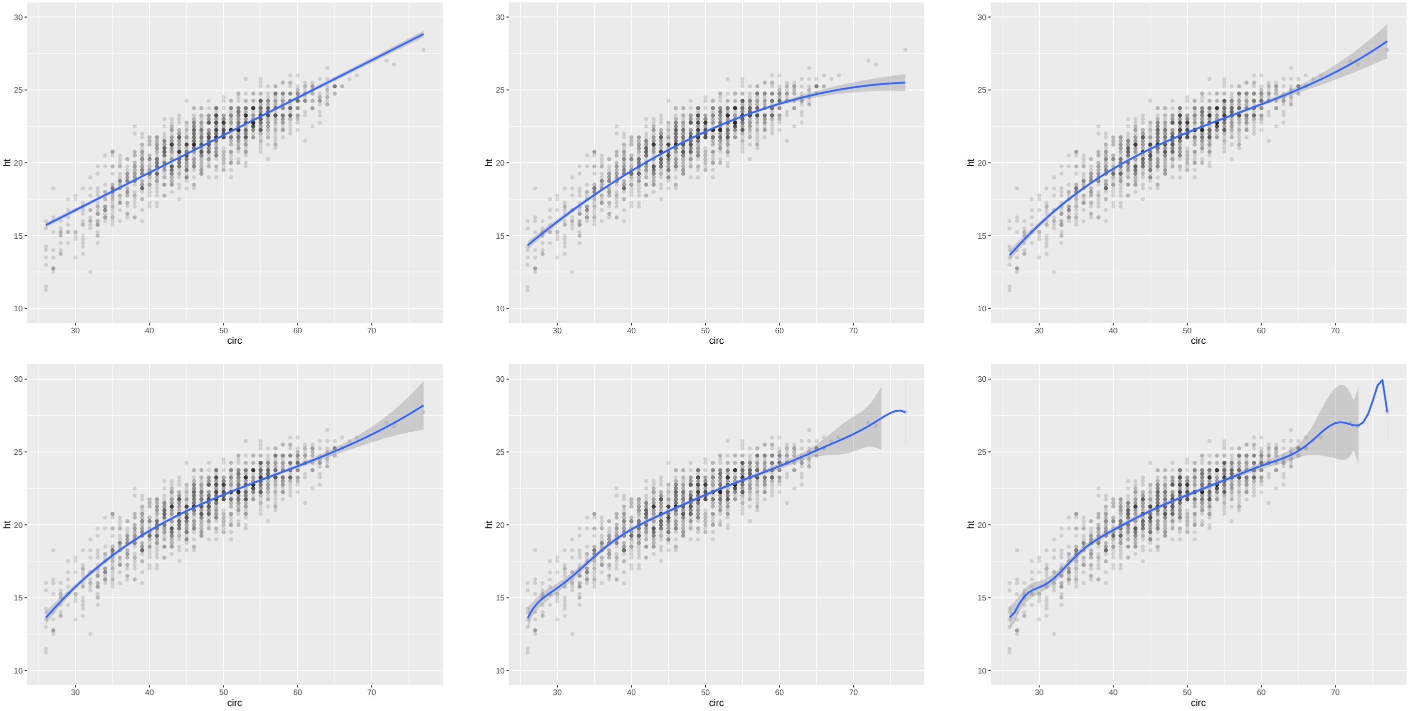
  with $\beta \in \mathbb{R}^{p'}$.

Feature Design

- Art of choosing Φ.
- Examples:
  - Renormalization, (domain specific) transform
  - Basis decomposition
  - Interaction between different variables...
- Need to select a good transformation.

2.1.2 Choice Criterion: Prediction Error or Prob. Distance

Choice Criterion



Choice Criterion

- *Prediction error* between $f^\star(\underline{X}) = \mathbb{E}\left[Y|\underline{X}\right]$ (or $Y|\underline{X}$) and $\widehat{f}(\underline{X})$,
- *Probability distance* between $Y|\underline{X}$ and $\mathrm{N}(\widehat{f}(\underline{X}), \widehat{\sigma^2})$

Fixed or Random Design

- Observation: $Y_i = f^\star(\underline{X}_i) + \epsilon_i$ at $\underline{X}_i$.
- Prediction: $\widehat{f}(\underline{X})$ at a generic $\underline{X}$.

Two settings

- *Fixed Design:*
  - The explanatory variables $\{\underline{X}_i\}_{i=1}^n$s are fixed (deterministic).
  - Error measured on the design variables $\{\underline{X}_i\}_{i=1}^n$
- *Random Design:*
  - The explanatory variables $\{\underline{X}_i\}_{i=1}^n$ are i.i.d. sampled from a unknown distribution.
  - Error averaged under the law of $\underline{X}$.

Fixed Design ($\underline{X}_i$ non random)

- Observation: $Y_i = f^\star(\underline{X}_i) + \epsilon_i$ at $\underline{X}_i$
- Prediction: $\widehat{f}(\underline{X})$ at a generic $\underline{X}$.
- New observation $Y_i' = f^\star(\underline{X}_i) + \epsilon_i'$ at the same $\underline{X}_i$.

Errors

- *Prediction error:*

$$\mathcal{R}^{\otimes n}(\widehat{f}) = n^{-1} \sum_{i=1}^{n} \mathbb{E}\left[|\widehat{f}(\underline{X}_i) - Y_i'|^2 \middle| \mathbb{X}_{(n)}, \epsilon_{(n)} \right]$$

- *Estimation error:*

$$\mathcal{E}^{\otimes n}(\widehat{f}) = n^{-1} \sum_{i=1}^{n} \mathbb{E}\left[|\widehat{f}(\underline{X}_i) - f^\star(\underline{X}_i)|^2 \middle| \mathbb{X}_{(n)}, \epsilon_{(n)} \right]$$

Errors

- Link between the Prediction and Estimation errors

$$\mathcal{R}^{\otimes n}(\widehat{f}) = \mathcal{E}^{\otimes n}(\widehat{f}) + \underbrace{n^{-1} \sum_{i=1}^{n} \mathbb{E}\left[|\epsilon_i'|^2 \right]}_{\text{indep. of } \widehat{f}}$$

- Since the predictor $\widehat{f}$ is random, these two quantities are also *random*.

Random Design ($\underline{X}_i$ random)

- Observation: $Y_i = f^\star(\underline{X}_i) + \epsilon_i$ at $\underline{X}_i$
- Prediction: $\widehat{f}(\underline{X})$ at a generic $\underline{X}$.
- New observation $Y' = f^\star(\underline{X}) + \epsilon'$ at a generic $\underline{X}$.

Errors

- *Prediction error:*

$$\mathcal{R}(\widehat{f}) = \mathbb{E}\left[|\widehat{f}(\underline{X}) - Y'|^2 \middle| \mathbb{X}_{(n)}, \epsilon_{(n)} \right]$$

- *Estimation error:*

$$\mathcal{E}(\widehat{f}) = \mathbb{E}\left[|\widehat{f}(\underline{X}) - f^\star(\underline{X})|^2 \middle| \mathbb{X}_{(n)}, \epsilon_{(n)} \right]$$

- Link:

$$\mathcal{R}(\widehat{f}) = \mathcal{E}(\widehat{f}) + \underbrace{\mathbb{E}\left[|\epsilon'|^2 \right]}_{\text{indep. of } \widehat{f}}$$

- *Rk:* As $\widehat{f}$ is random, those quantities are *random*.

Probabilistic Distance (Gaussian Model)

- Observation: $Y_i = f^\star(\underline{X}_i) + \epsilon_i$ at $\underline{X}_i$
- Estimation: $\widehat{Y|\underline{X}} \sim \mathrm{N}(\widehat{f}(\underline{X}), \widehat{\sigma^2})$ at a generic $\underline{X}$.

Kullback-Leibler Divergences

- *Fixed Design:*

$$\mathrm{KL}^{\otimes n}(Y|\underline{X}, \widehat{Y|\underline{X}}) = n^{-1} \sum_{i=1}^{n} \mathrm{KL}(Y|\underline{X}_i, \mathrm{N}(\widehat{f}(\underline{X}_i), \widehat{\sigma^2}))$$

- *Random Design:*

$$\mathrm{KL}^{\otimes}(Y|\underline{X}, \widehat{Y|\underline{X}}) = \mathbb{E}\left[\mathrm{KL}(Y|\underline{X}, \mathrm{N}(\widehat{f}(\underline{X}), \widehat{\sigma^2}))\middle|\mathbb{X}_{(n)}, \epsilon_{(n)}\right]$$

- *Rk:* As $\widehat{f}$ and $\widehat{\sigma^2}$ are random, those quantities are *random.*
- *Gaussian error:* If $\epsilon_i \sim \mathrm{N}(0, \sigma_\star^2)$
  - Kullback-Leiber divergence

$$\mathrm{KL}(Y|\underline{X}, \mathrm{N}(\widehat{f}(\underline{X}), \widehat{\sigma^2})) = \frac{1}{2}\log\frac{\widehat{\sigma^2}}{\sigma_\star^2} + \frac{\sigma_\star^2 + |\widehat{f}(\underline{X}) - f^\star(\underline{X})|^2}{2\widehat{\sigma^2}} - \frac{1}{2}$$

  - Similar criterion than prediction error (cf Least Square vs ML)

Small Error?

- A *good* model means an estimate $\widehat{f}$ with a *small error* $\mathcal{R}(\widehat{f})$!
- What does *small* means when $\mathcal{R}(\widehat{f})$ is random?

Average Behavior

- *Criterion:* Average risk (for i.i.d. copies of the dataset)

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right]$$

- *Bound:* $\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \leq \dots$

PAC Behavior

- *Criterion:* Quantile of order $1-\alpha$ of the risk (for i.i.d. copies of the dataset)

$$q = \min\{q', \mathbb{P}\left(\mathcal{R}(\widehat{f}) \leq q'\right) \geq 1-\alpha\}$$

- *Bound:* With probability larger than $1-\alpha$, $\mathcal{R}(\widehat{f}) \leq \dots$

2.1.3 Prediction Error Analysis

Fixed Design Analysis

- *Fixed Design:* estimation and prediction at the same fixed covariates $\{\underline{X}_i\}_{i=1}^n$.
- *Simpler* framework than the random design as we predict at the covariates values.

Matrix Notation

- Original Design Matrix:

$$\mathbb{X}_{(n)} = (\underline{X}_1, \dots, \underline{X}_n)^t$$

- Transformed Design Matrix:

$$\mathbb{X}^{\Phi}_{(n)} = (\underline{X}^{\Phi}_1, \dots, \underline{X}^{\Phi}_n)^t$$

- Observation: $\mathbb{Y}_{(n)} = (Y_1, \dots, Y_n)$ and

$$\mathbb{Y}_{(n)} = f^{\star}(\mathbb{X}_{(n)}) + \epsilon_{(n)}$$

- Linear prediction: $f_{\Phi,\beta}(\underline{X}^{\Phi}) = \underline{X}^{\Phi t}\beta$
- Least Square: $\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|\mathbb{Y}_{(n)} - \mathbb{X}^{\Phi}_{(n)}\beta\|^2$
- New observation:

$$\mathbb{Y}'_{(n)} = f^{\star}(\mathbb{X}_{(n)}) + \epsilon'_{(n)}$$

Errors

- *Prediction Error:*

$$\mathcal{R}^{\otimes n}(f_{\Phi,\beta}) = \mathbb{E}\left[n^{-1}\|\mathbb{Y}'_{(n)} - \mathbb{X}^{\Phi}_{(n)}\beta\|^2 \middle| \mathbb{X}_{(n)}, \epsilon_{(n)}\right]$$

- *Estimation Error:*

$$\mathcal{E}^{\otimes n}(f_{\Phi,\beta}) = n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) - \mathbb{X}^{\Phi}_{(n)}\beta\|^2$$

Link between Prediction and Estimation Errors

- Link:

$$\mathcal{R}^{\otimes n}(f_{\Phi,\beta}) = \mathcal{E}^{\otimes n}(f_{\Phi,\beta}) + \underbrace{\mathbb{E}\left[n^{-1}\|\epsilon'_{(n)}\|^2\right]}_{\text{ind. of } \beta}$$

- *Rk:* Those quantities are *random* as soon as β depends on the observations.

Least Square Error

- Least Square and Projection:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|\mathbb{Y}_{(n)} - \mathbb{X}^{\Phi}_{(n)}\beta\|^2 \Leftrightarrow \mathbb{X}^{\Phi}_{(n)}\hat{\beta} = \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \mathbb{Y}_{(n)}$$

where $\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}}$ is the projection matrix on $\operatorname{span}(\mathbb{X}^{\Phi}_{(n)})$.

Prediction and errors

- *Prediction*:

$$\begin{aligned}\mathbb{X}^{\Phi}_{(n)}\hat{\beta} &= \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} Y_{(n)} \\ &= \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)}) + \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\end{aligned}$$

- *Prediction difference*:

$$\mathbb{Y}'_{(n)} - \mathbb{X}_{(n)}\hat{\beta} = f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)} + \epsilon'_{(n)}$$

- *Estimation difference*:

$$f^{\star}(\mathbb{X}_{(n)}) - \mathbb{X}_{(n)}\hat{\beta} = f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}$$

Least square Error Theorem

Theorem 3. • *Prediction Error:*

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}^{\otimes n}(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] &= n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)})\|^2 \\ &\quad + \mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right] + \mathbb{E}\left[n^{-1}\|\epsilon'_{(n)}\|^2\right]\end{aligned}$$

- *Estimation Error:*

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}^{\otimes n}(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] &= n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)})\|^2 \\ &\quad + \mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]\end{aligned}$$

Proof

$$\begin{aligned}\mathcal{R}^{\otimes n}(f_{\Phi(,)\hat{\beta}}) &= \mathbb{E}\left[n^{-1}\|\mathbb{Y}'_{(n)} - \mathbb{X}^{\Phi}_{(n)}\hat{\beta}\|^2\middle|\mathbb{X}_{(n)}, \epsilon_{(n)}\right] \\ &= \mathbb{E}\left[n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) + \epsilon'_{(n)} - \mathbb{X}^{\Phi}_{(n)}\hat{\beta}\|^2\middle|\mathbb{X}_{(n)}, \epsilon_{(n)}\right] \\ &= n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) - \mathbb{X}^{\Phi}_{(n)}\hat{\beta}\|^2 + \mathbb{E}\left[n^{-1}\|\epsilon'_{(n)}\|^2\middle|\mathbb{X}_{(n)}, \epsilon_{(n)}\right] \\ &\quad + 2\mathbb{E}\left[n^{-1}\langle f^{\star}(\mathbb{X}_{(n)}) - \mathbb{X}^{\Phi}_{(n)}\hat{\beta}, \epsilon'_{(n)}\rangle\middle|\mathbb{X}_{(n)}, \epsilon_{(n)}\right]\end{aligned}$$

By independence between $\mathbb{X}_{(n)}, \epsilon_{(n)}$ and $\epsilon'_{(n)}$:

$$\mathcal{R}^{\otimes n}(f_{\Phi,\hat{\beta}}) = \mathcal{E}^{\otimes n}(f_{\Phi,\hat{\beta}}) + \mathbb{E}\left[n^{-1}\|\epsilon'_{(n)}\|^2\right]$$

$$\begin{aligned}\mathcal{E}^{\otimes n}(f_{\Phi,\hat\beta}) &= \|f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)}) + \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} \epsilon_{(n)}\|^2\\ &= n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)})\|^2 + n^{-1}\|\operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} \epsilon_{(n)}\|^2\\ &\quad + 2n^{-1}\langle f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)}), \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} \epsilon_{(n)}\rangle\\ &= n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)})\|^2 + n^{-1}\|\operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} \epsilon_{(n)}\|^2\end{aligned}$$

Taking the expectation leads then to

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}^{\otimes n}(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right] &= n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)})\|^2\\ &\quad + \mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} \epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]\end{aligned}$$

Bias / Variance tradeoff

Assumption $\mathbb{C}\mathrm{ov}\left[\epsilon_{(n)}\right] = \sigma_\star^2 \mathrm{Id}_{(n)}$.

- *Prediction Error:*

$$\mathbb{E}\left[\mathcal{R}^{\otimes n}(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right] = n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)})\|^2 + \sigma_\star^2 \frac{\dim \mathbb{X}^\Phi_{(n)}}{n} + \sigma_\star^2$$

- *Estimation Error:*

$$\mathbb{E}\left[\mathcal{E}^{\otimes n}(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right] = n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)})\|^2 + \sigma_\star^2 \frac{\dim \mathbb{X}^\Phi_{(n)}}{n}$$

Bias / Variance Interpretation

- *Estimation Error:*

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}^{\otimes n}(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right] &= n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)})\|^2\\ &\quad + \mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} \epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]\end{aligned}$$

- *Bias:* $n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)})\|^2$
  - Error of the best approximation of $f^\star(\mathbb{X}_{(n)})$ in $\operatorname{span}(\mathbb{X}^\Phi_{(n)})$.
  - Diminishes when the *dimension* increases.
- *Variance (Estimation error):* $\mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} \epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]$
  - Average norm of the projection of $\epsilon_{(n)}$ onto $\operatorname{span}(\mathbb{X}^\Phi_{(n)})$.
  - Increases when the *dimension* increases.
- If $\mathbb{C}\mathrm{ov}\left[\epsilon_{(n)}\right] = \sigma_\star^2 \mathrm{Id}_{(n)}$, variance: $\sigma_\star^2 \frac{\dim(\mathbb{X}^\Phi_{(n)})}{n}$.
- Similar decomposition for $\mathbb{E}\left[\mathcal{R}^{\otimes n}(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right]$.

Example 4: Variable Selection

Variable Selection

- Transformation(s): for a subset I of $\{1,\dots,p\}$

$$\underline{X}^{\Phi} = \Phi\left(\underline{X}\right) = \left(\underline{X}^{(k)}\right)^{t}_{k\in I}$$

- Estimate $f_{\Phi,\hat{\beta}}(\underline{X})$: least square estimate on the space spanned by a subset of variables.

- *Estimation Error:*

$$\mathbb{E}\left[\mathcal{E}^{\otimes n}(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] = n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) - \mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)})\|^2 + \sigma_{\star}^2\frac{|I|}{n}.$$

Bias/Variance Tradeoff

- Bias: ability to approximate the function $f^{\star}$ with a subset of variables.
- Variance: proportional to the number of variables used.
- Tradeoff depends on the unknown function $f^{\star}$!

Example 2: 1D Regression in a Dictionary

Setting

- $Y = f^{\star}(X) + \epsilon$ with $X \in [0,1]$.
- Estimation in a dictionary of functions ϕ_j by

$$f_{\Phi_p,\beta}(X) = \sum_{j=1}^{p}\beta_j\phi_j(X)$$

- Two choices:
  - function family (Fourier basis, polynomial basis, splines...)
  - ordered functions (first elements) / family subset ?
- For instance the p first elements of the Fourier basis or a polynomial basis...
- Linear model with $\underline{X}^{\Phi_p} = \left(\phi_1(X),\dots,\phi_p(X)\right)^{t}$!
- Strong assumption: ϵ i.i.d with variance $\sigma_{\star}^2$
- *Fixed design*: X_k not random.
- *Uniform design*: $X_k = k/n$, $k\in\{1,\dots,n\}$.

Projection Estimator

- $\hat{f} = f_{\Phi_p,\hat{\beta}}$ with

$$\hat{\beta} = \operatorname*{argmin}_{\beta} n^{-1} \sum_{i=1}^{n} |Y_i - \sum_{j=1}^{p} \beta_j \phi_j(X_i)|^2$$

Least-square error

$$\begin{aligned}&\mathbb{E}\left[\mathcal{R}^{\otimes n}(f_{\Phi_p,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right]\\ &= \underbrace{n^{-1}\|(I - \operatorname{Proj}_{\mathbb{X}_{(n)}^{\Phi_p}}) f^\star(\mathbb{X}_{(n)})\|^2}_{\text{Approx. error}} + \underbrace{\frac{p}{n}\sigma_\star^2}_{\text{Estim. error}} + \underbrace{\sigma_\star^2}_{\text{Variability}}\end{aligned}$$

Error for the Projection Estimator

Least-square error

$$\begin{aligned}&\mathbb{E}\left[\mathcal{R}^{\otimes n}(f_{\Phi_p,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right]\\ &= \underbrace{n^{-1}\|(I - \operatorname{Proj}_{\mathbb{X}_{(n)}^{\Phi_p}}) f^\star(\mathbb{X}_{(n)})\|^2}_{\text{Approx. error}} + \underbrace{\frac{p}{n}\sigma_\star^2}_{\text{Estim. error}} + \underbrace{\sigma_\star^2}_{\text{Variability}}\end{aligned}$$

- Approximation error in the basis *decays* with the number of elements p...
- but estimation error *grows* with p (more coefficients mean more variance) !
- What is the *optimal trade-off* for the projection order p?

Approximation Error and Regularity

Regularity Condition

- Regularity can be translated in term of approximation error!
- Typical approximation decay:

$$\frac{1}{n}\|(I - \operatorname{Proj}_{\mathbb{X}_{(n)}^{\Phi_p}}) f^\star(\mathbf{X}_{(n)})\|^2 \leq C p^{-2\alpha}$$

- Example:
  - Fourier decomposition with p first coefficients of a $\mathcal{C}^\alpha$ function.
  - Wavelet decomposition with p largest coefficients of a piecewise $\mathcal{C}^\alpha$ function.
- Consequence:

$$\begin{aligned}&\frac{1}{n}\mathbb{E}\left[\|\mathbf{Y}'_{(n)} - \mathbf{X}_{(n)}\widehat{\beta}\|^2\right]\\ &= \frac{1}{n}\|(I - \operatorname{Proj}_{\mathbb{X}_{(n)}^{\Phi_p}}) f^\star(\mathbf{X}_{(n)})\|^2 + \frac{p}{n}\sigma^2 + \sigma^2\\ &\leq C p^{-2\alpha} + \frac{p}{n}\sigma^2 + \sigma^2\end{aligned}$$

Optimized Approximation Errror

Optimization in p

- If p were free:

$$\min_{p\geq 1} Cp^{-2\alpha} + \frac{p}{n}\sigma^2 + \sigma^2 \leq C_\alpha \left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$$

- Similar bound for $p \in \mathbb{N}^*$:

$$\min_{p\in\mathbb{N}} Cp^{-2\alpha} + \frac{p}{n}\sigma^2 + \sigma^2 \leq C'_\alpha \left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$$

- Obtained for $p \sim \left(\frac{n}{\sigma^2}\right)^{2\alpha+1}$
- Issue: best choice depends of the unknown regularity α!
- Similar setting for variable selection but less theoretical control...

Example 2: 1D Regression in a Dictionary

Dictionary Choice

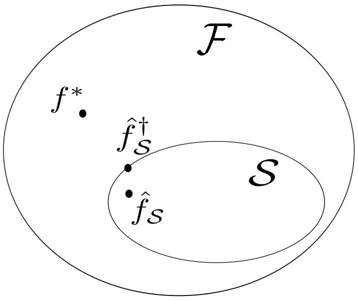
- Find a family function ϕ_j such that

$$\frac{1}{n}\|(I - \mathrm{Proj}_{\mathbb{X}_{(n)}^{\Phi_p}})f^\star(\mathbf{X}_{(n)})\|^2 \leq Cp^{-2\alpha}$$

  with a certain ordering of the function ϕ_i, a small constant C and a large exponent α.

- Set $p \sim \left(\frac{n}{\sigma^2}\right)^{2\alpha+1}$
- Questions:
  - Which family (polynomial, Fourier...) ?
  - Which dimension? Which ordering? Which subset?
- There is a *best* choice that depends on $f^\star$, n and σ!

Generic Bias-Variance Dilemma



- Precise analysis valid only for linear regression with fixed design but the *conclusions remain valid in full generality*!
- General setting:
  - $\mathcal{F} = \{\text{measurable functions } \mathcal{X} \to \mathcal{Y}\}$
  - Best solution: $f^\star = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$
  - Class $\mathcal{S} \subset \mathcal{F}$ of functions
  - Ideal target in $\mathcal{S}$: $f_{\mathcal{S}}^\dagger = \operatorname{argmin}_{f \in \mathcal{S}} \mathcal{R}(f)$
  - Estimate in $\mathcal{S}$: $\widehat{f}_{\mathcal{S}}$ obtained with some procedure

Approximation error and estimation error (Bias/Variance)

$$\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^\star) = \underbrace{\mathcal{R}(f_{\mathcal{S}}^\star) - \mathcal{R}(f^\star)}_{\text{Approximation error}} + \underbrace{\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f_{\mathcal{S}}^\star)}_{\text{Estimation error}}$$

- Approx. error can be large if the model $\mathcal{S}$ is not suitable and is independent of n.
- Estimation error can be large if the model is complex and decreases with n.

- *Best* choice *depends* on the *unknown best function* and ***n***!

2.2 Kernel Based Linear Regression

2.2.1 Kernel Estimation

Local Averaging

A naive idea

- $\mathbb{E}[Y|\mathbf{X}]$ can be approximated by a local average:

$$\hat{f}(\mathbf{X}) = \frac{1}{|\{\mathbf{X}_i \in \mathcal{N}(\mathbf{X})\}|} \sum_{\mathbf{X}_i \in \mathcal{N}(\mathbf{X})} Y_i$$

  where $\mathcal{B}(\mathbf{X})$ is a neighborhood of $\mathbf{X}$.

- Heuristic:
  - On the one hand, if $\mathbf{X} \to \mathbb{E}\left[Y|\mathbf{X}\right]$ is regular then
  $$\mathbb{E}\left[Y|\mathbf{X}\right] \simeq \mathbb{E}\left[\mathbb{E}\left[Y|\mathbf{X}'\right]|\mathbf{X}' \in \mathcal{N}(\mathbf{X})\right] = \mathbb{E}\left[Y|\mathbf{X}' \in \mathcal{N}(\mathbf{X})\right]$$
  - On the other hand,
  $$\mathbb{E}\left[Y|\mathbf{X}' \in \mathcal{N}(\mathbf{X})\right] \simeq \frac{1}{|\{\mathbf{X}_i \in \mathcal{N}(\mathbf{X})\}|} \sum_{\mathbf{X}_i \in \mathcal{N}(\mathbf{X})} Y_i$$

Neighborhood and Size

- Most classical choice: $\mathcal{N}(\mathbf{X}) = \{\mathbf{X}', \|\mathbf{X} - \mathbf{X}'\| \leq h\ \}$ where $\|.\|$ is a (pseudo) norm and h a size (bandwidth) parameter.
- In principle, the norm and h could vary with $\mathbf{X}$, and the norm can be replaced by a (pseudo) distance.
- Focus here on a fixed distance with a fixed bandwidth h cased.

Bandwidth

- Heuristic:
  - A large bandwidth ensures that the average is taken on many samples and thus the variance is small...
  - A small bandwidth is such that the approximation $\mathbb{E}\left[Y|\mathbf{X}\right] \simeq \mathbb{E}\left[Y|\mathbf{X}' \in \mathcal{N}(\mathbf{X})\right]$ is more accurate.
- Parameter choice issue!

Weighted Local Averaging

Weighted Local Average

- Replace the neighborhood $\mathcal{N}(\mathbf{X})$ by a decaying window function $w(\mathbf{X}, \mathbf{X}')$.
- $\mathbb{E}\left[Y|\mathbf{X}\right]$ can be approximated by a weighted local average:
$$\widehat{f}(\mathbf{X}) = \frac{\sum_i w(\mathbf{X}, \mathbf{X}'_i) Y_i}{\sum_i w(\mathbf{X}, \mathbf{X}'_i)}.$$

Kernel

- Most classical choice: $w(\mathbf{X}, \mathbf{X}') = K\left(\frac{\mathbf{X}-\mathbf{X}'}{h}\right)$ where h the bandwidth is a scale parameter.
- Examples:
  - Box kernel: $K(t) = \mathbf{1}_{\|t\| \leq 1}$ (Neighborhood)
  - Triangular kernel: $K(t) = \max(1 - \|t\|, 0)$.
  - Gaussian kernel: $K(t) = e^{-t^2/2}$
- **Rk:** K and λK yields the same estimate.

Link with Density Estimation

Density Estimation

- How to estimate the density p of $\mathbf{X}$ with respect to the Lebesgue measure from an i.i.d. sample $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$.
- Parametric approach: assume that the density has a known parametrized shape and estimate those parameters.
- Nonparametric approach: do not assume that the density has a known parametrized shape and
  - Approximate it by a parametric one, whose parameters can be estimated
  - Estimate directly the density
- Important nonparametric statistic topic!

Kernel Density Estimation (Parzen)

- Choose a positive kernel K such that $\int K(x)dx = 1$
- Use as an estimate

$$\widehat{p}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} K(\mathbf{X} - \mathbf{X}_i)$$

- If $K = \frac{1}{Z_h} \mathbf{1}_{\|t\| \leq h}$ this can be easily interpreted in term of local empirical density of samples!
- General K corresponds to the same smoothing idea we have used before.
- We will often use $K_h(t) = \frac{1}{h^d} K(t/h)$ and let

$$\widehat{p}_h(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{X} - \mathbf{X}_i)$$

Properties

- Error decomposition:

$$\mathbb{E}\left[|p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})|^2\right] = \mathbb{E}\left[p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})\right]^2 + \mathbb{V}\mathrm{ar}\left[p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})\right]$$

- Bias:

$$\mathbb{E}\left[p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})\right] = p(\mathbf{X}) - (K_h * p)(\mathbf{X})$$

- Variance: if p is upper bounded by $p_{\max}$ then

$$\mathbb{V}\mathrm{ar}\left[p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})\right] \leq \frac{p_{\max} \int K^2(x)dx}{nh^d}$$

Bandwidth choice

- A small h leads to a small bias but a large variance...
- A large h leads to a small variance but a large bias...
- Theoretical analysis possible!

Regularity, Bias and Bandwidth

Regularity

- Hölder type regularity: for $\alpha \in \mathbb{R}^+$, it exist C such that $\forall \mathbf{X}$ there is a polynomial $P_{\mathbf{X}}$ of degree smaller than α satisfying

$$\sup_{\mathbf{X}'} |p(\mathbf{X}) + P_{\mathbf{X}}(\mathbf{X}') - p(\mathbf{X}')| \leq C \|\mathbf{X} - \mathbf{X}'\|^\alpha$$

- Consequence: If K is such that $\int K(x)P(x)dx = P(0)$ for all polynomials of degree less than α then

$$|p(\mathbf{X}) - (K_h * p)(\mathbf{X})| \leq C' h^\alpha$$

Bias / Variance tradeoff

- Upper bound on the error:

$$\begin{aligned}\mathbb{E}\left[|p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})|^2\right] &= \mathbb{E}\left[p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})\right]^2 + \mathbb{V}\text{ar}\left[p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})\right] \\ &\leq (C')^2 h^{2\alpha} + \frac{p_{\max} \int K^2(x)dx}{nh^d}\end{aligned}$$

Upper bound optimization

- Upper bound on the error:

$$\begin{aligned}\mathbb{E}\left[|p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})|^2\right] &= \mathbb{E}\left[p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})\right]^2 + \mathbb{V}\text{ar}\left[p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})\right] \\ &\leq (C')^2 h^{2\alpha} + \frac{p_{\max} \int K^2(x)dx}{nh^d}\end{aligned}$$

- Optimized by $h^{2\alpha+d} = \frac{d p_{\max} \int K^2(x)dx}{2\alpha n}$.
- Resulting upper bound

$$\mathbb{E}\left[|p(\mathbf{X}) - \widehat{p}_h(\mathbf{X})|^2\right] \leq C''_{d,\alpha} \left(\frac{1}{n}\right)^{2\alpha/(2\alpha+d)}$$

- Adaptation issue!

From Density Estimation to Regression

Nadaraya-Watson Heuristic

- Provided all the densities exist

$$\mathbb{E}\left[Y|\mathbf{X}\right] = \frac{\int Y p(\mathbf{X}, Y) dY}{\int p(Y, \mathbf{X}) dY} = \frac{\int Y p(\mathbf{X}, Y) dY}{p(\mathbf{X})}$$

- Replace the unknown densities by their estimates:

$$\begin{aligned}\widehat{p}(\mathbf{X}) &= \frac{1}{n} \sum_{i=1}^n K(\mathbf{X} - \mathbf{X}_i) \\ \widehat{p}(\mathbf{X}, Y) &= \frac{1}{n} \sum_{i=1}^n K(\mathbf{X} - \mathbf{X}_i) K'(Y - Y_i)\end{aligned}$$

- Now if K' is a kernel such that $\int YK'(Y)dY = 0$ then

$$\int Y\widehat{p}(\mathbf{X}, Y)dY = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{X} - \mathbf{X}_i)Y_i$$

Nadaraya-Watson

- Resulting estimator of $\mathbb{E}\left[Y|\mathbf{X}\right]$

$$\widehat{f}(\mathbf{X}) = \frac{\sum_{i=1}^{n} Y_i K_h(\mathbf{X} - \mathbf{X}_i)}{\sum_{i=1}^{n} K_h(\mathbf{X} - \mathbf{X}_i)}$$

- Same local weighted average estimator!

Bandwidth Choice

- Bandwidth h of K allows to balances between bias and variance.
- Similar but more complex theoretical analysis of the error is possible.
- Same conclusion: the smoother the densities the easier the estimation but the optimal bandwidth depends on the unknown regularity!

Local Linear Estimation

Another Point of View on Kernel

- Nadaraya-Watson estimator:

$$\widehat{f}(\mathbf{X}) = \frac{\sum_{i=1}^{n} Y_i K_h(\mathbf{X} - \mathbf{X}_i)}{\sum_{i=1}^{n} K_h(\mathbf{X} - \mathbf{X}_i)}$$

- Can be view as a minimizer of

$$\sum_{i=1}^{n} |Y_i - \beta|^2 K_h(\mathbf{X} - \mathbf{X}_i)$$

- Local regression of order 0!

Local Linear Model

- Estimate $\mathbb{E}\left[Y|\mathbf{X}\right]$ by $\widehat{f}(\mathbf{X}) = \langle \Phi(\mathbf{X}), \widehat{\beta}(\mathbf{X})\rangle$ where Φ is any function of $\mathbf{X}$ and $\widehat{\beta}(\mathbf{X})$ is the minimizer of

$$\sum_{i=1}^{n} |Y_i - \langle \Phi(\mathbf{X}_i), \beta\rangle|^2 K_h(\mathbf{X} - \mathbf{X}_i).$$

LOESS: LOcal polynomial regrESSion

1D Nonparametric Regression

- Assume that $\mathbf{X} \in \mathbb{R}$ and let $\Phi(\mathbf{X}) = (1, \mathbf{X}, \ldots, \mathbf{X}^d)$.
- LOESS estimate: $\hat{f}(\mathbf{X}) = \sum_{j=0}^{d} \widehat{\beta}(\mathbf{X}_j \mathbf{X}^j$ with $\widehat{\beta}(\mathbf{X})$ minimizing
$$\sum_{i=1}^{n} |Y_i - \sum_{j=1}^{d} \beta \mathbf{X}_i^j|^2 K_h(\mathbf{X} - \mathbf{X}_i).$$
- Most classical kernel used: Tricubic kernel
$$K(t) = \max(1 - |t|^3, 0)^3$$
- Most classical degree: 2...
- Local bandwidth choice such that a proportion of points belongs to the window.

Linear Smoother

Linearity?

- On easily verify that for all the previous estimators
$$\widehat{f}(\mathbf{X}) = \sum_{i=1}^{n} \omega_i(\mathbf{X}) Y_i = \langle \omega(\mathbf{X}), Y_{(n)} \rangle$$
where $\omega(\mathbf{X})$ is independent of $Y_{(n)}$ but not of $\mathbf{X}_{(n)}$.
- Examples:
  - Nadaraya-Watson: $\omega_i(\mathbf{X}) = \frac{K(\mathbf{X},\mathbf{X}_i)}{\sum_{i'=1}^{n} K(\mathbf{X},\mathbf{X}_i)}$
  - Linear regression: $\omega_i(\mathbf{X}) = \left(\mathbf{X}^t (\mathbf{X}_{(n)}^t \mathbf{X}_{(n)})^{-1} \mathbf{X}^t\right)_i$

Linear Smoother

- We will call Linear Smoother any estimator that can be written as
$$\widehat{f}(\mathbf{X}) = \sum_{i=1}^{n} \omega_i(\mathbf{X}) Y_i = \langle \omega(\mathbf{X}), Y_{(n)} \rangle$$
where $\omega(\mathbf{X})$ is independent of $Y_{(n)}$ but not of $\mathbf{X}_{(n)}$.
- We may further impose that $\sum_{i=1}^{n} \omega_i(\mathbf{X}) = 1$!
- Large class of estimators!

Model Selection

- Model selection within a family (bandwidth choice, variable selection) or across different techniques...
- What is the best choice of $\omega_m(\mathbf{X})$ with $m \in \mathcal{M}$?
- Model selection technique: CV and penalization!

2.2.2 Model Selection and Empirical Error

Model Selection

Approximation error and estimation error (Bias/Variance)

$$\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^\star) = \underbrace{\mathcal{R}(f^\star_{\mathcal{S}}) - \mathcal{R}(f^\star)}_{\text{Approximation error}} + \underbrace{\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^\star_{\mathcal{S}})}_{\text{Estimation error}}$$

- Approx. error can be large if the model $\mathcal{S}$ is not suitable and is independent of n.
- Estimation error can be large if the model is complex and decreases with n.

- Good model collection and efficient automatic tradeoff = Key to obtain good performance!

Statistical Learning

- How to design models? (Model/feature design)
- How to chose among several models? (Model/feature selection)

Empirical Risk Analysis

Empirical Risk

- A natural proxy for $\mathcal{R}(f) = \mathbb{E}\left[|Y - f(\underline{X})|^2\right]$ is the empirical risk:

$$\mathcal{R}_n(f) = n^{-1}\sum_{i=1}^{n} |Y_i - f(\underline{X}_i)|^2$$

- *Question:* Can we assess the quality of a model by looking at $\mathcal{R}_n(\hat{f})$?

f is a deterministic function (does not depend of $(\underline{X}_i, Y_i)$)

- Empirical loss satisfies

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_n(f)\right] &= \mathcal{R}(f)\\ &= \mathcal{E}(f) + \mathbb{E}\left[|\epsilon|^2\right]\\ \mathbb{V}\text{ar}\left[\mathcal{R}_n(f)\right] &= \frac{C_f}{n}\end{aligned}$$

- Justify the approximation of $\mathcal{R}(f)$ by $\mathcal{R}_n(f)$.

But $\widehat{f}$ depends of $(\underline{X}_i, Y_i)$!

- Empirical loss satisfies

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_n(\widehat{f})\right] &\neq \mathcal{R}(\widehat{f})\\ \mathbb{V}\text{ar}\left[\mathcal{R}_n(\widehat{f})\right] &= ?\end{aligned}$$

- No justification of the approximation of $\mathcal{R}(f)$ by $\mathcal{R}_n(f)$.

Least Square and Error Optimism

- Linear model set: $\mathcal{S}_\Phi = \{f_{\Phi,\beta}, \beta \in \mathbb{R}^{p'}\}$
- Least Square: $f_{\Phi,\hat{\beta}} = \underset{f_{\Phi,\beta} \in \mathcal{S}_\Phi}{\operatorname{argmin}} \mathcal{R}_n(f_{\Phi,\beta})$.
- Best ideal Solution: $f_{\Phi,\beta^\dagger} = \underset{f_{\Phi,\beta} \in \mathcal{S}_\Phi}{\operatorname{argmin}} \mathcal{R}(f_{\Phi,\beta})$.

Empirical Risk Optimism

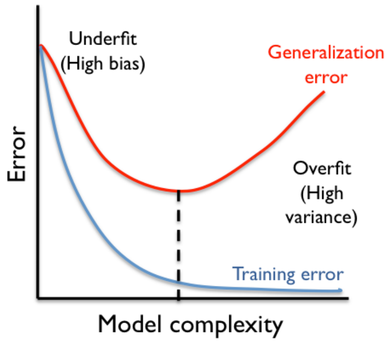
$$\mathbb{E}\left[\mathcal{R}_n(f_{\Phi,\hat{\beta}})\right] \leq \mathcal{R}(f_{\Phi,\beta^\dagger}) \leq \mathcal{R}(f_{\Phi,\hat{\beta}})$$

- *Proof:*
  - $\mathcal{R}_n(f_{\Phi,\hat{\beta}}) \leq \mathcal{R}_n(f_{\Phi,\beta^\dagger})$
  - $\mathcal{R}(f_{\Phi,\hat{\beta}}) \geq \mathcal{R}(f_{\Phi,\beta^\dagger})$
  - $\mathbb{E}\left[\mathcal{R}_n(f_{\Phi,\beta^\dagger})\right] = \mathcal{R}(f_{\Phi,\beta^\dagger})$
- *Fixed design:*

$$\mathbb{E}\left[\mathcal{R}_n^{\otimes n}(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] \leq \mathbb{E}\left[\mathcal{R}^{\otimes n}(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] - \mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} \epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]$$

- *Empirical error (almost) useless to assess the quality!*

Over-fitting Issue



Error behavior

- Empirical risk (error made on the training set) decays when the complexity of the model increases.
- Quite different behavior when the error is computed on new observations (true risk / generalization error).

- Overfit for complex models: parameters learned are too specific to the learning set!
- General situation! (Think of polynomial fit...)
- Need to use another criterion than the training error!

Cross Validation and Penalization

Two directions

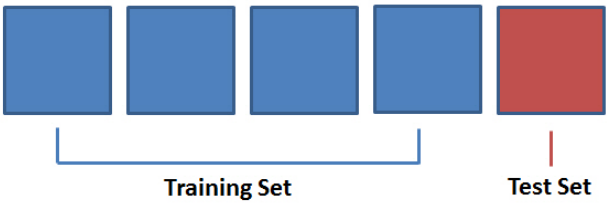
- *Estimate* $\mathcal{R}(\hat{f})$ in a different way?
- Find a way to *correct* $\mathcal{R}_n(\hat{f})$?

Two Approaches

- *Cross validation:* Estimate the error on a different dataset:
  - Very efficient (and almost always used in practice!)
  - Need more data for the error computation.
- *Penalization approach:* Correct the optimism of the empirical error:
  - Requires to find the correction (penalty).
- Outline:
  - Cross Validation,
  - Prediction error correction,
  - Probabilistic distance correction
  - Practical Penalization.

2.2.3 Cross Validation

Cross Validation



- *Very simple idea:* use a second learning/verification set to compute a verification error.
- Sufficient to remove the dependency issue!
- Implicit random design setting...

Cross Validation

- Use $(1-\epsilon)n$ observations to train and ϵn to verify!
- Validation for a learning set of size $(1-\epsilon)\times n$ instead of n!
- Unstable error estimate if ϵn is too small ?
- Most classical variations:
  - Hold Out,
  - Leave One Out,
  - V-fold cross validation.

Hold Out

Principle

- Split the dataset $\mathcal{D}$ in 2 sets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ of size $n(1-\epsilon)$ and $n\epsilon$.
- Learn $\hat{f}^{HO}$ from the subset $\mathcal{D}_{\text{train}}$.
- Compute the empirical error on the subset $\mathcal{D}_{\text{test}}$:[-.45cm]

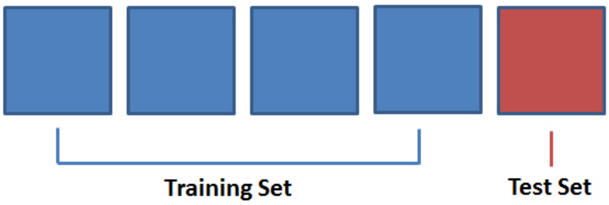
$$\mathcal{R}_n^{HO}(\hat{f}^{HO}) = \frac{1}{n\epsilon} \sum_{(\underline{X}_i, Y_i) \in \mathcal{D}_{\text{test}}} |Y_i - \hat{f}^{HO}(\underline{X}_i)|^2$$

Model Selection by Cross Validation

- Compute $\mathcal{R}_n^{HO}(\hat{f}_{\mathcal{S}}^{HO})$ for all possible models $\mathcal{S}$,
- Select the model with the smallest CV error,
- Reestimate the $\hat{f}_{\mathcal{S}}$ with all the data.

- Does not take into account the variability of $\mathcal{R}_n^{HO}(\hat{f}^{HO})$ ($\hat{f}$ and error estimate)
- Biased toward simpler model as the estimation does not use all the data initially.

V-fold Cross Validation



Principle

- Split the dataset $\mathcal{D}$ in V sets $\mathcal{D}_v$ of almost equal size.
- For $v \in \{1, \dots, V\}$:
  - Learn $\hat{f}^{-v}$ from the dataset $\mathcal{D}$ from which the data in $\mathcal{D}_v$ have been removed.
  - Compute the empirical error on $\mathcal{D}_v$:

$$\mathcal{R}_n^{-v}(\hat{f}^{-v}) = \frac{1}{n_v} \sum_{(\underline{X}_i, Y_i) \in \mathcal{D}_v} |Y_i - \hat{f}^{-v}(\underline{X}_i)|^2$$

- Compute the average empirical error over all the blocks:

$$\mathcal{R}_n^{CV}(\hat{f}) = \frac{1}{V} \sum_{v=1}^{V} \mathcal{R}_n^{-v}(\hat{f}^{-v})$$

- Leave One Out : V = n.

Analysis (when n is a multiple of V)

- The $\mathcal{R}_n^{-v}(\hat{f}^{-v})$ are identically distributed variable but are not independent!
- Consequence:

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_n^{CV}(\hat{f})\right] &= \mathbb{E}\left[\mathcal{R}_n^{-v}(\hat{f}^{-v})\right]\\ \mathbb{V}\mathrm{ar}\left[\mathcal{R}_n^{CV}(\hat{f})\right] &= \frac{1}{V}\,\mathbb{V}\mathrm{ar}\left[\mathcal{R}_n^{-v}(\hat{f}^{-v})\right]\\ &\quad + (1-\frac{1}{V})\,\mathbb{C}\mathrm{ov}\left[\mathcal{R}_n^{-v}(\hat{f}^{-v}), \mathcal{R}_n^{-v'}(\hat{f}^{-v'})\right]\end{aligned}$$

- Average risk for a sample of size $(1-\frac{1}{V})n$.
- Variance term much more complex to analyze!
- Accuracy/Speed tradeoff: $V = 5$ or $V = 10$!

Cross Validation and Confidence Interval

- How to replace pointwise estimation by a confidence interval?
- Can we use the variability of the CV estimates?
- *Negative result:* No unbiased estimate of the variance!

Gaussian Interval (Quasi independence assumption)

- Compute the empirical variance and divide it by the number of folds to construct an asymptotic Gaussian confidence interval,
- Select the simplest model whose values falls into the confidence interval of the model having the smallest CV error.

PAC approach (Small risk estimation error assumption)

- Compute the raw medians or (or a larger raw quantiles)
- Select the model having the smallest quantiles to ensure a small risk with high probability.
- Always reestimate the chosen model with all the data.

Linear Regression and Leave One Out

- Leave One Out = V fold for $V = n$: very expensive in general.

A fast LOO formula for the linear regression

- *Prop:* for the least squares linear regression,

$$\hat{f}^{-i}(\underline{X}_i) = \frac{\hat{f}(\underline{X}_i) - h_{ii}Y_i}{1-h_{ii}}$$

  with h_{ii} the ith diagonal coefficient of the *hat* (projection) matrix $\mathbb{X}^{\Phi}_{(n)}(\mathbb{X}_{(n)}{}^{\Phi t}\mathbb{X}^{\Phi}_{(n)})^{-1}\mathbb{X}_{(n)}{}^{\Phi t}$.
- Proof based on linear algebra!
- Leads to a fast formula for LOO:

$$\mathcal{R}_n^{LOO}(\hat{f}) = n^{-1}\sum_{i=1}^{n}\frac{|Y_i - \hat{f}(\underline{X}_i)|^2}{(1-h_{ii})^2}$$

Proof of LOO formula

- By construction,

$$\hat{f}^{-i}(\underline{X}_i) = \underline{X}_i^{\Phi\,t}\hat{\beta}^{-i} = \underline{X}_i{}^t(\underline{X}^{\Phi}_{(n)-i}{}^t\,\underline{X}^{\Phi}_{(n)-i})^{-1}\underline{X}^{\Phi}_{(n)-i}{}^t\,\mathbb{Y}_{(n)-i}$$

- Now $\underline{X}^{\Phi}_{(n)-i}{}^t\,\underline{X}^{\Phi}_{(n)-i} = \mathbb{X}_{(n)}{}^{\Phi\,t}\mathbb{X}^{\Phi}_{(n)} - \underline{X}_i^{\Phi}\,\underline{X}_i{}^t$ and $\underline{X}^{\Phi}_{(n)-i}{}^t\,\underline{\mathbf{Y}}_{(n)-i} = \mathbb{X}_{(n)}{}^{\Phi\,t}\mathbb{Y}_{(n)} - \underline{X}_i^{\Phi}Y_i$

- Using $(M+uv^t)^{-1} = M^{-1} - \frac{M^{-1}uv^tM^{-1}}{1+u^tM^{-1}v}$ with $M = \mathbb{X}^t_{(n)}\mathbb{X}_{(n)}$, $u = -v = \underline{X}_i$ yields:

$$\hat{f}^{-i}(\underline{X}_i) = \underline{X}_i^{\Phi\,t}\left(M^{-1} + \frac{M^{-1}\underline{X}_i^{\Phi}\,\underline{X}_i^{\Phi\,t}M^{-1}}{1-\underline{X}_i^{\Phi\,t}M^{-1}\underline{X}_i^{\Phi}}\right)\left(\mathbb{X}_{(n)}{}^{\Phi\,t}\mathbb{Y}_{(n)} - \underline{X}_i^{\Phi}Y_i\right)$$

using $h_{ii} = \underline{X}_i^{\Phi\,t}M^{-1}\underline{X}_i^{\Phi}$

$$= \hat{f}(\underline{X}_i) + \frac{h_{ii}}{1-h_{ii}}\hat{f}(\underline{X}_i) - h_{ii}Y_i - \frac{h_{ii}^2}{Y}{}_i$$

$$\hat{f}^{-i}(\underline{X}_i) = \frac{\hat{f}(\underline{X}_i) - h_{ii}Y_i}{1-h_{ii}}$$

Linear Smoother and LOESS

- Generalization to linear smoother:

$$\widehat{f}(\underline{X}) = \sum_{i=1}^{n}\omega_i(\underline{X})Y_i = \omega(\underline{X})^t\mathbb{Y}_{(n)}$$

where $\omega(\underline{X})$ is independent of $\mathbb{Y}_{(n)}$ but not of $\mathbb{X}_{(n)}$.

LOESS

- Most commonly used linear smoothther: *LOWESS* (or *LOESS*) procedure first developed by Cleveland (1979)
- *LOWESS*: locally weighted scatterplot smoother / *LOESS*: local regression.
- Essentially the same thing: local polynomial regression fits blended together.

A fast LOO formula for most linear smoother

- *Assumption:* The linear smoother is such that

$$\hat{f}^{-i}(\underline{X}) = \frac{\sum_{j\neq i}\omega_j(\underline{X})Y_j}{\sum_{j\neq i}\omega_j(\underline{X})} = \frac{\hat{f}(\underline{X}) - \omega_i(\underline{X})Y_j}{1-\omega_i(\underline{X})}$$

- True for linear least square, LOESS and LOWESS.

LOO for Linear Smoothers

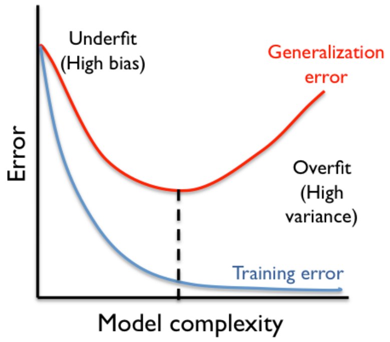
- Leads to a fast formula for LOO:

$$\mathcal{R}_n^{LOO}(\hat{f}) = n^{-1}\sum_{i=1}^{n}\frac{|Y_i - \hat{f}(\underline{X}_i)|^2}{(1-\omega_i(\underline{X}_i))^2}$$

- Generalized CV heuristic: Replace $\omega_i(\underline{X}_i)$ by $n^{-1}\sum_{i=1}^{n}\omega_i(\underline{X}_i)$!

2.3 Prediction Error Correction

Over-fitting Issue



Error behavior

- Empirical risk (error made on the training set) decays when the complexity of the model increases.
- Quite different behavior when the error is computed on new observations (true risk / generalization error).

- Overfit for complex models: parameters learned are too specific to the learning set!
- General situation! (Think of polynomial fit...)
- Need to use another criterion than the training error!

Cross Validation and Penalization

Two directions

- *Estimate* $\mathcal{R}(\hat{f})$ in a different way?
- Find a way to *correct* $\mathcal{R}_n(\hat{f})$?

Two Approaches

- *Cross validation:* Estimate the error on a different dataset:
  - Very efficient (and almost always used in practice!)
  - Need more data for the error computation.
- *Penalization approach:* Correct the optimism of the empirical error:
  - Requires to find the correction (penalty).

- Outline:
  - Cross Validation,
  - Prediction error correction,
  - Probabilistic distance correction
  - Practical Penalization.

2.3.1 Prediction Error and Its Unbiased Correction

Least Square and Error

- Regression model:

$$\mathbb{Y}_{(n)} = f^\star + \epsilon_{(n)}$$

- Linear model: $f_{\Phi,\beta} = \mathbb{X}^\Phi_{(n)}\beta$.
- Least square:

$$\hat{\beta} = \operatorname*{argmin}_\beta \|\mathbb{Y}_{(n)} - \mathbb{X}^\Phi_{(n)}\beta\|^2$$

Errors

- *Prediction Error:*

$$\mathcal{R}^{\otimes n}(f_{\Phi,\hat{\beta}}) = \mathbb{E}\left[n^{-1}\|\mathbb{Y}'_{(n)} - \mathbb{X}^\Phi_{(n)}\hat{\beta}\|^2\middle|\mathbb{X}_{(n)},\epsilon_{(n)}\right]$$

- *Estimation Error:*

$$\mathcal{E}^{\otimes n}(f_{\Phi,\hat{\beta}}) = n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \mathbb{X}^\Phi_{(n)}\hat{\beta}\|^2$$

- Rk:$\mathcal{R}^{\otimes n}(f_{\Phi,\hat{\beta}}) = \mathcal{E}^{\otimes n}(f_{\Phi,\hat{\beta}}) + \mathbb{E}\left[n^{-1}\|\epsilon'_{(n)}\|^2\right]$
- *Empirical Error:*

$$\mathcal{R}_n(f_{\Phi,\hat{\beta}}) = \|\mathbb{Y}_{(n)} - \mathbb{X}^\Phi_{(n)}\hat{\beta}\|^2$$

Least Square Error Theorem

- *Prediction Error:*

$$\mathbb{E}\left[\mathcal{R}^{\otimes n}(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] = n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)})\|^2 \\ + \mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}}\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right] + \mathbb{E}\left[n^{-1}\|\epsilon'_{(n)}\|^2\right]$$

- *Estimation Error:*

$$\mathbb{E}\left[\mathcal{E}^{\otimes n}(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] = n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)})\|^2 \\ + \mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}}\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]$$

- *Empirical Error:* ?

Empirical Error Analysis

Proposition 15 (Empirical Error for linear least-squares estimator). - *If $f_{\Phi,\hat{\beta}}$ is the least square estimate:*

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_n(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] &= n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)})\|^2 \\ &+ \mathbb{E}\left[n^{-1}\|(I - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]\end{aligned}$$

- *If* $\mathbb{C}\mathrm{ov}\left[\epsilon_{(n)}\right] = \sigma_{\star}^2 \mathrm{Id}_{(n)}$,

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_n(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] &= n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)})\|^2 \\ &+ \left(1 - \frac{\dim \mathbb{X}^{\Phi}_{(n)}}{n}\right)\sigma_{\star}^2\end{aligned}$$

Proof

- By construction,

$$\begin{aligned}\mathcal{R}_n(f_{\Phi,\hat{\beta}}) = n^{-1}\|\mathbb{Y}_{(n)} - f_{\Phi,\hat{\beta}}(\mathbb{X}_{(n)})\|^2 &= n^{-1}\|\mathbb{Y}_{(n)} - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \mathbb{Y}_{(n)}\|^2 \\ &= n^{-1}\|(I - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})f^{\star}(\mathbb{X}_{(n)}) + \epsilon_{(n)}\|^2 \\ &= n^{-1}\|(I - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})f^{\star}(\mathbb{X}_{(n)})\|^2 + n^{-1}\|(I - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})\epsilon_{(n)}\|^2 \\ &+ \frac{2}{n}\langle (I - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})f^{\star}(\mathbb{X}_{(n)})(I - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})\epsilon_{(n)}\rangle\end{aligned}$$

- Since $\mathbb{E}\left[\epsilon_{(n)}\middle|\mathbb{X}_{(n)}\right] = 0$, we get

$$\begin{aligned}&\mathbb{E}\left[\mathcal{R}_n(f_{\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] \\ &= n^{-1}\|(I - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})f^{\star}(\mathbb{X}_{(n)})\|^2 + \mathbb{E}\left[n^{-1}\|(I - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right] \\ &= n^{-1}\|(I - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})f^{\star}(\mathbb{X}_{(n)})\|^2 + \left(1 - \frac{\dim \mathbb{X}^{\Phi}_{(n)}}{n}\right)\sigma_{\star}^2\end{aligned}$$

Least Square Error Theorem

- *Prediction Error:*

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}^{\otimes n}(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] &= n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)})\|^2 \\ &+ \mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right] + \mathbb{E}\left[n^{-1}\|\epsilon'_{(n)}\|^2\right]\end{aligned}$$

- *Estimation Error:*

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}^{\otimes n}(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] &= n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)})\|^2 \\ &+ \mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]\end{aligned}$$

- *Empirical Error:*

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_n(f_{\Phi,\hat{\beta}})\middle|\mathbb{X}_{(n)}\right] &= n^{-1}\|f^{\star}(\mathbb{X}_{(n)}) - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} f^{\star}(\mathbb{X}_{(n)})\|^2 \\ &+ \mathbb{E}\left[n^{-1}\|(I - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]\end{aligned}$$

Prediction Error and Empirical Error

Proposition 16.

$$\mathbb{E}\left[\mathcal{R}^{\otimes n}(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right] = \mathbb{E}\left[\mathcal{R}_n(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right] + 2\mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}}\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]$$

If $\mathbb{C}\mathrm{ov}\left[\epsilon_{(n)}\right] = \sigma_\star^2\mathrm{Id}_{(n)}$,

$$\mathbb{E}\left[\mathcal{R}^{\otimes n}(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right] = \mathbb{E}\left[\mathcal{R}_n(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right] + 2\sigma_\star^2\frac{\dim\mathbb{X}^{\Phi}_{(n)}}{n}$$

Proof

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}^{\otimes n}(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right]\\
&\quad= \mathbb{E}\left[\mathcal{R}_n(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right] - \mathbb{E}\left[n^{-1}\|(\mathrm{Id}_{(n)} - \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}})\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]\\
&\quad+ \mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}}\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]\\
&= \mathbb{E}\left[\mathcal{R}_n(f_{\Phi,\hat\beta})\middle|\mathbb{X}_{(n)}\right] + 2\mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}}\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]\\
&\quad- \mathbb{E}\left[n^{-1}\|\epsilon_{(n)}\|^2\right]
\end{aligned}$$

Unbiased Risk Estimation

- *Idea:* Correct the empirical error to obtain an unbiased error estimate.

Unbiased Risk Estimate

- *Prediction Error:*

$$\mathcal{R}_n(f_{\Phi,\hat\beta}) + 2\mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}}\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right]$$

- *Estimation Error:*

$$\mathcal{R}_n(f_{\Phi,\hat\beta}) + 2\mathbb{E}\left[n^{-1}\|\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}}\epsilon_{(n)}\|^2\middle|\mathbb{X}_{(n)}\right] - \mathbb{E}\left[n^{-1}\|\epsilon'_{(n)}\|^2\right]$$

Unbiased Risk Estimate (if $\mathbb{C}\mathrm{ov}\left[\epsilon_{(n)}\right] = \sigma_\star^2\mathrm{Id}_{(n)}$)

- *Prediction Error:*

$$\mathcal{R}_n(f_{\Phi,\hat\beta}) + 2\sigma_\star^2\frac{\dim\mathbb{X}^{\Phi}_{(n)}}{n}$$

- *Estimation Error:*

$$\mathcal{R}_n(f_{\Phi,\hat\beta}) + 2\sigma_\star^2\frac{\dim\mathbb{X}^{\Phi}_{(n)}}{n} - \sigma_\star^2$$

Unbiased Risk Model Selection

Principle

- Compute the least square estimate $\widehat{f}_\Phi$ for all transformations Φ in a collection as well as the corresponding empirical risk

$$\mathcal{R}_n(\widehat{f}_\Phi) = n^{-1} \sum_{i=1}^{n} |Y_i - \widehat{f}_\Phi(\underline{X}_i)|^2.$$

- Estimate the variance $\sigma_\star^2$ of the model by $\widehat{\sigma^2}$.

- Correct those risk thanks to the previous formula:

$$\mathcal{R}_n(\widehat{f}_\Phi) + 2\widehat{\sigma^2} \frac{\dim \mathbb{X}_{(n)}^\Phi}{n}$$

  where $\dim \mathbb{X}_{(n)}^\Phi$ is the dimension of the least square model.

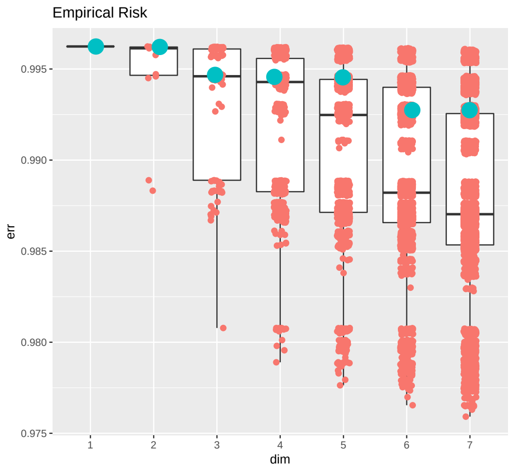
- Select the one with the smallest corrected risk.

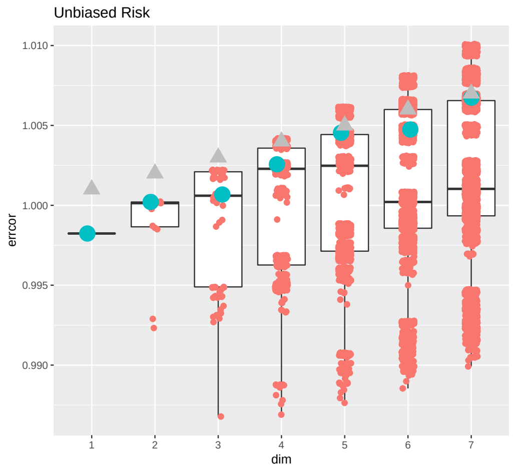
- *Rk:* Estimation of $\widehat{\sigma^2}$ (often done in the most complex model to limit the bias effect) $\frac{1}{n - \dim \mathbb{X}_{(n)}^\Phi} \|\mathbb{Y}_{(n)} - \operatorname{Proj}_{\mathbb{X}_{(n)}^\Phi} \mathbb{Y}_{(n)}\|^2$

- No theoretical guarantee so far...

2.3.2 Penalized Model Selection

Bias Correction Optimism





- Unbiased heuristic is *optimistic*:
  - Does not take into account the *variance* of the estimate!
  - Does not take into account the fact that we use *several* estimates!

Confidence Bound

- Unbiased heuristic is *optimistic*:
  - Does not take into account the *variance* of the estimate!
- Idea: replace the value by a confidence bound!
- Possible using an asymptotic analysis...

Asymptotic Analysis ($\mathbb{C}\mathrm{ov}\left[\epsilon_{(n)}\right] = \sigma_\star^2 \mathrm{Id}_{(n)}$)

- For any $\epsilon > 0$ and $\eta > 0$, the probability of

$$\begin{aligned}(1-\eta)\left(\mathcal{R}^{\otimes n}(\widehat{f}_\Phi) - \mathcal{R}^{\otimes n}(f^\star)\right) \leq & \left(\mathcal{R}_n(\widehat{f}_\Phi) - \mathcal{R}_n(f^\star)\right)\\ & + (2+2\eta)\sigma_\star^2 \frac{\dim \mathbb{X}^\Phi_{(n)}}{n}\\ & + \frac{2\sigma_\star^2(-\log \epsilon)}{\eta n}\end{aligned}$$

  tends to be larger than $1-\epsilon$ when n goes to infinity.

- Asymptotic upper bound of the estimation risk for a single model!

Concentration

- *Concentration*: control of the deviation of a random variable around its expectation.
- Thm: Bienaymé-Tchebyshev, Hoeffding, Bernstein, Gaussian tails...

Gaussian tails

- $\forall t > 0$,

$$\begin{aligned}\frac{t^2}{1+t^2}\frac{e^{-t^2/2}}{t\sqrt{2\pi}} \leq & \mathbb{P}\left(N > \mu + t\sigma\right)\\ & \leq \min\left(\frac{e^{-t^2/2}}{t\sqrt{2\pi}}, \frac{e^{-t^2/2}}{2}\right) \leq e^{-t^2/2}\end{aligned}$$

  where $N \sim \mathrm{N}(\mu, \sigma)$

- Asymptotic analysis based on an asymptotic Gaussian behavior with σ small with respect to μ...

Proof

- Chernoff bound:

$$\begin{aligned}\mathbb{P}\left(N > t\right) &= \min_\lambda \mathbb{P}\left(e^{\lambda N} > e^{\lambda t}\right)\\ &\leq \min_\lambda \frac{\mathbb{E}\left[e^{\lambda N}\right]}{e^{\lambda t}}\\ &\leq \min_\lambda e^{\lambda^2/2 - \lambda t}\\ &\leq e^{\min_\lambda \lambda^2/2 - \lambda t} = e^{-t^2/2}\end{aligned}$$

- Less accurate but generic technique.
- Upper bound:

$$\begin{aligned}
\mathbb{P}\left(N>t\right) &= \int_t^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\
&= \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(t+u)^2}{2}} du \\
&= \int_0^{+\infty} e^{-t^2/2} e^{-2ut} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\
&= e^{-t^2/2} \int_0^{+\infty} e^{-ut} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\
&\leq e^{-t^2/2} \min\left(\int_0^{+\infty} e^{-ut} \frac{1}{\sqrt{2\pi}} du, \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \right) \\
&\leq e^{-t^2/2} \min(\frac{1}{t\sqrt{2\pi}}, \frac{1}{2})
\end{aligned}$$

- Lower bound:

$$\begin{aligned}
\mathbb{P}\left(N>t\right) &= \int_t^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\
&\geq \int_t^{+\infty} \frac{t^2}{u^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\
&\geq \left[\frac{-t^2}{u} \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}} \right]_t^{+\infty} - \int_t^{+\infty} \left(-\frac{t^2}{u}\right) \left(-u \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}\right) du \\
&\geq t^2 \frac{e^{-\frac{t^2}{2}}}{t\sqrt{2\pi}} - t^2 \mathbb{P}\left(N>t\right)
\end{aligned}$$

  hence

$$\mathbb{P}\left(N>t\right) \geq \frac{t^2}{1+t^2} \frac{e^{-\frac{t^2}{2}}}{t\sqrt{2\pi}}$$

Sketch of Proof

- Let

$$\begin{aligned}
\widehat{f}_\Phi &= \operatorname*{argmin}_{f_{\Phi,\beta}} \mathcal{R}_n(f_{\Phi,\beta}) = \operatorname*{argmin}_{f_{\Phi,\beta}} n^{-1} \|\mathbb{Y}_{(n)} - f_{\Phi,\beta}(\mathbb{X}_{(n)})\|^2 \\
\widehat{f}_\Phi(\mathbb{X}_{(n)}) &= \mathrm{Proj}_{\mathbb{X}^\Phi_{(n)}} \mathbb{Y}_{(n)} \\
\widetilde{f}_\Phi &= \operatorname*{argmin}_{f_{\Phi,\beta}} \mathcal{R}(f_{\Phi,\beta}) \\
&= \operatorname*{argmin}_{f_{\Phi,\beta}} n^{-1} \|f^\star(\mathbb{X}_{(n)}) - f_{\Phi,\beta}(\mathbb{X}_{(n)})\|^2 + \mathbb{E}\left[n^{-1}\|\epsilon_{(n)}\|^2\right] \\
\widetilde{f}_\Phi(\mathbb{X}_{(n)}) &= \mathrm{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star(\mathbb{X}_{(n)})
\end{aligned}$$

- Rk: $\widehat{f}_\Phi(\mathbb{X}_{(n)}) = \mathrm{Proj}_{\mathbb{X}^\Phi_{(n)}} \mathbb{Y}_{(n)} = \mathrm{Proj}_{\mathbb{X}^\Phi_{(n)}} (f^\star(\mathbb{X}_{(n)}) + \epsilon_{(n)}) = \widetilde{f}_\Phi + \mathrm{Proj}_{\mathbb{X}^\Phi_{(n)}} \epsilon_{(n)}$
- By construction,

$$\Delta_n(\widehat{f}_\Phi) = \mathcal{R}_n(\widehat{f}_\Phi) - \mathcal{R}_n(f^\star)$$

$$\begin{aligned}
&= n^{-1}\|\mathbb{Y}_{(n)} - \mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \mathbb{Y}_{(n)}\|^2 - n^{-1}\|\mathbb{Y}_{(n)} - f^\star(\mathbb{X}_{(n)})\|^2 \\
&= n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)})\|^2 + n^{-1}\|(I - \mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}})\epsilon_{(n)}\|^2 \\
&\qquad + \frac{2}{n}\langle f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)}), (I - \mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}})\epsilon_{(n)}\rangle \\
&\quad - n^{-1}\|\epsilon_{(n)}\|^2 \\
&= n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)})\|^2 - n^{-1}\|\mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 \\
&\qquad + \frac{2}{n}\langle f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)}), \epsilon_{(n)}\rangle
\end{aligned}$$

- Along the same line,

$$\begin{aligned}
\Delta(\widehat{f}_\Phi) &= \mathcal{R}^{\otimes n}(\widehat{f}_\Phi) - \mathcal{R}^{\otimes n}(f^\star) \\
&= \mathbb{E}\left[n^{-1}\|\mathbb{Y}'_{(n)} - \mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \mathbb{Y}_{(n)}\|^2 - \|\mathbb{Y}'_{(n)} - f^\star(\mathbb{X}_{(n)})\|^2 \middle| \mathbb{X}_{(n)}, \epsilon_{(n)}\right] \\
&= n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)})\|^2 + \mathbb{E}\left[n^{-1}\|\epsilon'_{(n)} - \mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 \middle| \mathbb{X}_{(n)}, \epsilon_{(n)}\right] \\
&\quad + \mathbb{E}\left[n^{-1}\langle f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)}), \epsilon'_{(n)} - \mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\rangle\| \middle| \mathbb{X}_{(n)}, \epsilon_{(n)}\right] \\
&\quad - \mathbb{E}\left[n^{-1}\|\epsilon'_{(n)}\|^2\right]
\end{aligned}$$

  and using the independence between $\epsilon_{(n)}$ and $\epsilon'_{(n)}$, the orthogonality between $f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)}) = (\mathrm{Id}_{(n)} - \mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}})f^\star(\mathbb{X}_{(n)})$ and $\mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}$, as well as $\mathbb{E}\left[\epsilon'_{(n)}\right] = \mathbb{0}_{(n)}$

$$= n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)})\|^2 + n^{-1}\|\mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2$$

- Thus

$$\begin{aligned}
\Delta(\widehat{f}_\Phi) - \Delta_n(\widehat{f}_\Phi) = \mathcal{R}(\widehat{f}_\Phi) - \mathcal{R}(f^\star) - \left(\mathcal{R}_n(\widehat{f}_\Phi) - \mathcal{R}_n(f^\star)\right) \\
= \frac{2}{n}\|\mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 \\
- \frac{2}{n}\langle f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)}), \epsilon_{(n)}\rangle
\end{aligned}$$

- We verify that

$$\mathbb{E}\left[\Delta(\widehat{f}_\Phi) - \Delta_n(\widehat{f}_\Phi) - 2\sigma_\star^2\frac{\dim \mathbb{X}^{\Phi}_{(n)}}{n} \middle| \mathbb{X}_{(n)}\right] = 0$$

- Now, as $\mathbb{C}\mathrm{ov}\left[\epsilon_{(n)}\right] = \sigma_\star^2\mathrm{Id}_{(n)}$, asymptotically
  - $\frac{2}{n}\|\mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 \sim \frac{2\sigma_\star^2}{n}\chi^2(\dim \mathbb{X}^{\Phi}_{(n)})$
  - $\frac{2}{n}\langle f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)}), \epsilon_{(n)}\rangle \sim \mathrm{N}(0, \frac{4\sigma_\star^2}{n}n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)})\|^2)$
  - Those variables are asymptotically independent.

- Asymptotically,

$$\frac{2}{n}\|\mathrm{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 \sim \mathrm{N}\left(2\sigma_\star^2\frac{\dim \mathbb{X}^{\Phi}_{(n)}}{n}, \frac{8\sigma^4}{n}\frac{\dim \mathbb{X}^{\Phi}_{(n)}}{n}\right)$$

and thus asymptotically $\Delta(\widehat{f}_\Phi) - \Delta_n(\widehat{f}_\Phi) - 2\sigma_\star^2 \frac{\dim \mathbb{X}^\Phi_{(n)}}{n}$ follows a Gaussian law

$$\begin{aligned}
&\mathrm{N}\left(0, \frac{4\sigma_\star^2}{n} n^{-1}\|f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)})\|^2 + \frac{8\sigma_\star^4}{n}\frac{\dim \mathbb{X}^\Phi_{(n)}}{n}\right)\\
&= \mathrm{N}\left(0, \frac{4\sigma_\star^2}{n}\left(n^{-1}\|(f^\star(\mathbb{X}_{(n)}) - \widetilde{f}_\Phi(\mathbb{X}_{(n)})\|^2 + 2\sigma_\star^2\frac{\dim\mathbb{X}^\Phi_{(n)}}{n}\right)\right)\\
&= \mathrm{N}\left(0, \frac{4\sigma_\star^2}{n}\widetilde{\Delta}\right)
\end{aligned}$$

- Using now $\mathbb{P}\left(\mathrm{N}(0,\sigma_\star) \geq \sqrt{2t\sigma_\star}\right) \leq e^{-t}$, we derive

$$\mathbb{P}\left\{\Delta(\widehat{f}_\Phi) - \Delta_n(\widehat{f}_\Phi) - 2\sigma_\star^2\frac{\dim\mathbb{X}_{(n)}}{n} \geq 2\sqrt{2t}\sqrt{\frac{\sigma_\star^2}{n}}\sqrt{\widetilde{\Delta}}\right\} \leq e^{-t}$$

- Now for any $\eta > 0$

$$\begin{aligned}
2\sqrt{2t}\sqrt{\frac{\sigma_\star^2}{n}}\sqrt{\widetilde{\Delta}} &= 2\sqrt{\frac{2t\sigma^2}{n}}\sqrt{\widetilde{\Delta}}\\
&\leq \frac{2t\sigma_\star^2}{\eta n} + \eta\widetilde{\Delta}\\
&\leq \frac{2t\sigma_\star^2}{\eta n} + \eta\left(n^{-1}\|(I - \mathrm{Proj}_{\mathbb{X}^\Phi_{(n)}})f^\star(\mathbb{X}_{(n)})\|^2 + 2\sigma_\star^2\frac{\dim\mathbb{X}^\Phi_{(n)}}{n}\right)\\
&\leq \frac{2t\sigma_\star^2}{\eta n} + \eta\left(\Delta(\widehat{f}_\Phi) + 2\sigma_\star^2\frac{\dim\mathbb{X}^\Phi_{(n)}}{n}\right)
\end{aligned}$$

- Hence, asymptotically,

$$(1-\eta)\Delta(\widehat{f}_\Phi) \leq \Delta_n(\widehat{f}_\Phi) + (2+2\eta)\sigma_\star^2\frac{\dim\mathbb{X}_{(n)}}{n} + \frac{2\sigma_\star^2 t}{\eta n}$$

  holds with a probability smaller than e^{-t},

- Equivalently, asymptotically,

$$(1-\eta)\Delta(\widehat{f}_\Phi) \leq \Delta_n(\widehat{f}_\Phi) + (2+2\eta)\sigma_\star^2\frac{\dim\mathbb{X}_{(n)}}{n} + \frac{2\sigma_\star^2(-\log\epsilon)}{\eta n}$$

  holds with a probability larger than $1-\epsilon$.

Confidence Bounds

- Unbiased heuristic is *optimistic*:
  - Does not take into account the fact that we use *several* estimates!

From One Confidence Bound to Several

- *Key tool*: union bound.
- For any (finite) collection of transformation Φ, any sequence ϵ_Φ and any $\eta > 0$, asymptotically, with probability larger than $1 - \sum_\Phi \epsilon_\Phi$ *simultaneously for all* Φ

$$\begin{aligned}
(1-\eta)\left(\mathcal{R}^{\otimes n}(\widehat{f}_\Phi) - \mathcal{R}^{\otimes n}(f^\star)\right) \leq \left(\mathcal{R}_n(\widehat{f}_\Phi) - \mathcal{R}_n(f^\star)\right)&\\
+ (2+2\eta)\sigma_\star^2\frac{\dim\mathbb{X}^\Phi_{(n)}}{n}&\\
+ \frac{2\sigma_\star^2(-\log\epsilon_\Phi)}{\eta n}&
\end{aligned}$$

- *Simultaneous Confidence Bounds Heuristic:* Assume $\sum_\Phi \epsilon_\Phi \leq \epsilon \ll 1$ then minimizing

$$\mathcal{R}_n(\widehat{f}_\Phi) + (2+2\eta)\sigma_\star^2 \frac{\dim \mathbb{X}^\Phi_{(n)}}{n} + \frac{2\sigma_\star^2(-\log \epsilon_\Phi)}{\eta n}$$

  is almost equivalent to minimizing $\mathcal{R}(\widehat{f}_\Phi)$

Union bound

- *Key property:*

$$\mathbb{P}\left(\cup_i A_i\right) \leq \sum_{i=1} \mathbb{P}\left(A_i\right)$$
$$\mathbb{P}\left(\cap_i A_i\right) \geq 1 - \sum_i (1 - \mathbb{P}\left(A_i\right))$$

- Rk: equality if the A_i (respectively $\bar{A}_i$ are disjoints.

Union bound

- If the A_is are such that $\mathbb{P}\left(A_i\right) \geq 1 - \epsilon_i$
- then $\mathbb{P}\left(\cap_i A_i\right) \geq 1 - \sum_i \epsilon_i$

- Canonical example with $Z_i \sim \mathrm{N}(\mu, \sigma^2)$:
  - $\forall i,\ \mathbb{P}\left(Z_i \leq \mu_i + \sqrt{2|I| \log \epsilon^{-1}}\sigma_i\right) \geq 1 - \frac{\epsilon}{|I|}$
  - $\mathbb{P}\left(\forall i,\ Z_i \leq \mu_i + \sqrt{2|I| \log \epsilon^{-1}}\sigma_i\right) \geq 1 - \epsilon$

Penalized Model Selection

- Theoretical result can be obtained for the Gaussian i.i.d. noise setting!

Model Selection

- Assume $\sum_\Phi \epsilon_\Phi \leq \epsilon$, let $0 < \eta < 1$ and let $\mathrm{pen}(\Phi)$ be an arbitrary penalty satisfying

$$\mathrm{pen}(\Phi) \geq \frac{\sigma_\star^2}{n}\left((2+2\eta)\dim \mathbb{X}^\Phi_{(n)} + (4 + \frac{4}{\eta})(-\log \epsilon_\Phi)\right)$$

- Let

$$\widehat{\Phi} = \operatorname*{argmin}_\Phi \mathcal{R}_n(\widehat{f}_\Phi) + \mathrm{pen}(\Phi)$$

  then

$$\begin{aligned}&\mathbb{E}\left[\mathcal{E}^{\otimes n}\left(\widehat{f}_{\widehat{\Phi}}\right)\middle|\mathbb{X}_{(n)}\right]\\ &\quad\leq \frac{1}{1-\eta}\left(\min_\Phi \min_{f_{\Phi,\beta}} n^{-1}\|f_{\Phi,\beta} - f^\star\|^2 + \mathrm{pen}(\Phi) + \frac{\sigma_\star^2}{n}(4 + \frac{4}{\eta})\epsilon\right)\end{aligned}$$

- Almost as well as the best bias and variance tradeoff...

Penalty

Theoretical Penalty

- Pen: $\operatorname{pen}(\Phi) > \frac{\sigma_\star^2}{n}\Big((2+2\eta)\dim \mathbb{X}^\Phi_{(n)} + (4+\frac{4}{\eta})(-\log \epsilon_\Phi)\Big)$.
- Slightly larger than AIC $(2\sigma_\star^2\frac{\dim(\mathbb{X}^\Phi_{(n)})}{n})$:
  - $2+2\eta$ instead of 2 in front of $\sigma_\star^2\frac{\dim(\mathbb{X}^\Phi_{(n)})}{n}$ (Variance)
  - Union bound term $\sigma_\star^2(4+\frac{4}{\eta})\frac{-\log\epsilon_\Phi}{n}$ (Simultaneous control)

Bounds for ϵ_Φ

- Finite collection of size M:

$$\epsilon_\Phi = \epsilon/M \Leftrightarrow -\log\epsilon_\Phi = \log M - \log\epsilon$$

- Embedded collection of increasing dimension up to p :

$$\epsilon_\Phi = \frac{\epsilon}{\log(p)\dim\mathbb{X}^\Phi_{(n)}} \Leftrightarrow -\log\epsilon_\Phi = \log\dim\mathbb{X}^\Phi_{(n)} + \log\log p - \log\epsilon$$

- All spaces spanned by a subset of coordinates in dimension p

$$\epsilon_\Phi = \frac{\epsilon}{Cp^{\dim\mathbb{X}^\Phi_{(n)}}} \Leftrightarrow -\log\epsilon_\Phi = \log p(\dim\mathbb{X}^\Phi_{(n)}) + \log C - \log\epsilon$$

- Negligible with respect to $\dim\mathbb{X}^\Phi_{(n)}$ except for the last case...

Proof

- Let Φ be a transformation, we define, with a slight abuse of notation,

$$\begin{aligned}\widehat{f}_\Phi &= \operatorname*{argmin}_{f_{\Phi,\beta}} \mathcal{R}_n(f_{\Phi,\beta}) = \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} \mathbb{Y}_{(n)}\\ \widetilde{f}_\Phi &= \operatorname*{argmin}_{f_{\Phi,\beta}} \mathcal{R}(f_{\Phi,\beta}) = \operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} f^\star\end{aligned}$$

  and

$$\begin{aligned}\widehat{\Delta}_n(\Phi) &= \mathcal{R}_n(\widehat{f}_\Phi) - \mathcal{R}_n(f^\star) = n^{-1}\|\mathbb{Y}_{(n)} - \widehat{f}_\Phi\|^2 - n^{-1}\|\mathbb{Y}_{(n)} - f^\star\|^2\\ &= \|f^\star - \widehat{f}_\Phi\|^2 + \frac{2}{n}\langle\epsilon_{(n)}, f^\star - \widehat{f}_\Phi\rangle\\ \widetilde{\Delta}_n(\Phi) &= \mathcal{R}_n(\widetilde{f}_\Phi) - \mathcal{R}_n(f^\star)\\ &= \|f^\star - \widetilde{f}_\Phi\|^2 + \frac{2}{n}\langle\epsilon_{(n)}, f^\star - \widetilde{f}_\Phi\rangle\\ \widehat{\Delta}(\Phi) &= \mathcal{R}(\widehat{f}_\Phi) - \mathcal{R}(f^\star)\\ &= \|f^\star - \widehat{f}_\Phi\|^2 = \|f^\star - \widetilde{f}_\Phi\|^2 + \|\operatorname{Proj}_{\mathbb{X}^\Phi_{(n)}} \epsilon_{(n)}\|^2\\ \widetilde{\Delta}(\Phi) &= \mathcal{R}(\widetilde{f}_\Phi) - \mathcal{R}(f^\star)\\ &= \|f^\star - \widetilde{f}_\Phi\|^2\end{aligned}$$

- Now $\widehat{\Phi}$ is defined as

$$\begin{aligned}\widehat{\Phi} &= \operatorname*{argmin}_{\Phi} \mathcal{R}_n(\widehat{f}_\Phi) + \mathrm{pen}(\Phi)\\ &= \operatorname*{argmin}_{\Phi} \widehat{\Delta}_n(\Phi) + \mathrm{pen}(\Phi)\end{aligned}$$

- By construction,

$$\begin{aligned}\widehat{\Delta}(\widehat{\Phi}) &\leq \widehat{\Delta}_n(\widehat{\Phi}) + \widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi})\\ &\leq \widehat{\Delta}_n(\widehat{\Phi}) + \mathrm{pen}(\widehat{\Phi}) + \widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi}) - \mathrm{pen}(\widehat{\Phi})\\ &\leq \widetilde{\Delta}_n(\Phi) + \mathrm{pen}(\Phi) + \widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi}) - \mathrm{pen}(\widehat{\Phi})\\ &\leq \widetilde{\Delta}(\Phi) + \mathrm{pen}(\Phi) + \widetilde{\Delta}_n(\Phi) - \widetilde{\Delta}(\Phi) + \widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi}) - \mathrm{pen}(\widehat{\Phi})\end{aligned}$$

  so that

$$\begin{aligned}\mathbb{E}\left[\widehat{\Delta}(\widehat{\Phi})\right] &\leq \mathbb{E}\left[\widetilde{\Delta}(\Phi) + \mathrm{pen}(\Phi) + \widetilde{\Delta}_n(\Phi) - \widetilde{\Delta}(\Phi) + \widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi}) - \mathrm{pen}(\widehat{\Phi})\right]\\ &\leq \widetilde{\Delta}(\Phi) + \mathrm{pen}(\Phi) + \mathbb{E}\left[\widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi}) - \mathrm{pen}(\widehat{\Phi})\right]\\ &\leq \min_{\Phi}\left(\widetilde{\Delta}(\Phi) + \mathrm{pen}(\Phi)\right) + \mathbb{E}\left[\widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi}) - \mathrm{pen}(\widehat{\Phi})\right]\end{aligned}$$

  which yields the result if $\mathbb{E}\left[\widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi}) - \mathrm{pen}(\widehat{\Phi})\right] \leq 0$

- We will not be able to prove such a result!
- We will only be able to bound $(1-\eta)\widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi})$.
- We start from

$$\begin{aligned}(1-\eta)\widehat{\Delta}(\widehat{\Phi}) \leq \widetilde{\Delta}(\Phi) + \mathrm{pen}(\Phi) &+ \widetilde{\Delta}_n(\Phi) - \widetilde{\Delta}(\Phi)\\ &+ (1-\eta)\widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi}) - \mathrm{pen}(\widehat{\Phi})\end{aligned}$$

- The term $\widetilde{\Delta}_n(\Phi) - \widetilde{\Delta}(\Phi) = \frac{2}{n}\langle \epsilon_{(n)}, f^\star - \widetilde{f}_\Phi\rangle \sim \mathrm{N}(0, \frac{4\sigma_\star^2}{n} n^{-1}\|f^\star - \widetilde{f}_\Phi\|^2)$ and thus satisfy

$$\mathbb{E}\left[\widetilde{\Delta}_n(\Phi) - \widetilde{\Delta}(\Phi)\right] = 0$$

$$\mathbb{P}\left(\widetilde{\Delta}_n(\Phi) - \widetilde{\Delta}(\Phi) > \frac{2\sigma}{\sqrt{n}}\sqrt{\widetilde{\Delta}(\Phi)}\sqrt{2t}\right) \leq e^{-t}$$

  and also

$$\mathbb{P}\left(\widetilde{\Delta}_n(\Phi) - \widetilde{\Delta}(\Phi) > \eta\widetilde{\Delta}(\Phi) + \frac{2\sigma^2 t}{n\eta}\right) \leq e^{-t}$$

- For the term, $(1-\eta)\widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi}) - \mathrm{pen}(\widehat{\Phi})$, we notice that

$$(1-\eta)\widehat{\Delta}(\widehat{\Phi}) - \widehat{\Delta}_n(\widehat{\Phi}) - \mathrm{pen}(\widehat{\Phi}) \leq \sup_{\Phi}\left((1-\eta)\widehat{\Delta}(\Phi) - \widehat{\Delta}_n(\Phi) - \mathrm{pen}(\Phi)\right)$$

- Now

$$\begin{aligned}&(1-\eta)\widehat{\Delta}(\Phi) - \widehat{\Delta}_n(\Phi) - \mathrm{pen}(\Phi)\\ &\quad= \widehat{\Delta}(\Phi) - \widehat{\Delta}_n(\Phi) - \left(\mathrm{pen}(\Phi) + \eta\widehat{\Delta}(\Phi)\right)\\ &\quad= \frac{2}{n}\|\operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 + \frac{2}{n}\langle \widetilde{f}_\Phi - f^\star, \epsilon_{(n)}\rangle - \left(\mathrm{pen}(\Phi) + \eta\widehat{\Delta}(\Phi)\right)\end{aligned}$$

$$\begin{aligned}&\le \frac{2}{n}\| \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 + \frac{2}{n}\langle \widetilde{f}_\Phi - f^\star, \epsilon_{(n)}\rangle \\ &\quad - \left(\operatorname{pen}(\Phi) + \eta n^{-1}\|\widetilde{f}_\Phi - f^\star\|^2\right)\end{aligned}$$

with the two random term independent and of known laws

$$\begin{aligned}\frac{2}{n}\| \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 &\sim \frac{2\sigma^2}{n}\,\chi^2(\dim \mathbb{X}^{\Phi}_{(n)}) \\ \frac{2}{n}\langle \widetilde{f}_\Phi - f^\star, \epsilon_{(n)}\rangle &\sim \mathrm{N}\left(0, \frac{4\sigma^2}{n} n^{-1}\|\widetilde{f}_\Phi - f^\star\|^2\right)\end{aligned}$$

- Note that

$$\mathbb{P}\left(\mathrm{N}(0,\sigma^2) > \sigma\sqrt{2t}\right) \le e^{-t}$$

while (we will not prove it...)

$$\mathbb{P}\left(\chi^2(p) > p + 2\sqrt{pt} + 2t\right) \le e^{-t}$$

- This implies that

$$\begin{aligned}\mathbb{P}\left(\frac{2}{n}\| \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 > \frac{2\sigma^2}{n}\dim\Phi + \frac{4\sigma^2}{n}\sqrt{\dim \mathbb{X}^{\Phi}_{(n)} t} + \frac{4\sigma^2}{n} t\right) &> e^{-t} \\ \mathbb{P}\left(\frac{2}{n}\| \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 > \frac{\sigma^2}{n}(2+2\eta)\dim\Phi + \frac{\sigma^2}{n}(4+\frac{2}{\eta})t\right) &> e^{-t}\end{aligned}$$

along the same line

$$\begin{aligned}\mathbb{P}\left(\frac{2}{n}\langle \widetilde{f}_\Phi - f^\star, \epsilon_{(n)}\rangle > 2\sqrt{\frac{\sigma^2}{n}}\sqrt{n^{-1}\|\widetilde{f}_\Phi - f^\star\|^2}\sqrt{2t}\right) &> e^{-t} \\ \mathbb{P}\left(\frac{2}{n}\langle \widetilde{f}_\Phi - f^\star, \epsilon_{(n)}\rangle > \eta n^{-1}\|\widetilde{f}_\Phi - f^\star\|^2 + \frac{2\sigma^2}{\eta n} t\right) &> e^{-t}\end{aligned}$$

- Combining the two bounds yields that the event

$$\begin{aligned}\frac{2}{n}\| \operatorname{Proj}_{\mathbb{X}^{\Phi}_{(n)}} \epsilon_{(n)}\|^2 + \frac{2}{n}\langle \widetilde{f}_\Phi - f^\star, \epsilon_{(n)}\rangle \ge \frac{\sigma^2}{n}&(2+2\eta)\dim \mathbb{X}^{\Phi}_{(n)} + \frac{\sigma^2}{n}(4+\frac{2}{\eta})t \\ &+ \eta n^{-1}\|\widetilde{f}_\Phi - f^\star\|^2 + \frac{2\sigma^2}{\eta n}t'\end{aligned}$$

occurs with a probability smaller than $e^{-t} + e^{-t'}$

- In turn, this implies that

$$\begin{aligned}&(1-\eta)\widehat{\Delta}(\Phi) - \widehat{\Delta}_n(\Phi) - \operatorname{pen}(\Phi) \\ &\quad\ge \frac{\sigma^2}{n}(2+2\eta)\dim \mathbb{X}^{\Phi}_{(n)} + \frac{\sigma^2}{n}(4+\frac{2}{\eta})t \\ &\qquad + \eta n^{-1}\|\widetilde{f}_\Phi - f^\star\|^2 + \frac{2\sigma^2}{n}t' - \left(\operatorname{pen}(\Phi) + \eta n^{-1}\|\widetilde{f}_\Phi - f^\star\|^2\right) \\ &\quad\ge \frac{\sigma^2}{n}(2+2\eta)\dim \mathbb{X}^{\Phi}_{(n)} + \frac{\sigma^2}{n}(4+\frac{2}{\eta})t + \frac{2\sigma^2}{\eta n}t' - \operatorname{pen}(\Phi)\end{aligned}$$

occurs with a probability smaller than $e^{-t} + e^{-t'}$

- Let $t = t' = u - \log \epsilon_\Phi$, we deduce that
$$\begin{aligned}(1-\eta)\widehat{\Delta}(\Phi) - \widehat{\Delta}_n(\Phi) - \text{pen}(\Phi) \\ \geq \frac{\sigma^2}{n}(2+2\eta)\dim \mathbb{X}^{\Phi}_{(n)} - \frac{\sigma^2}{n}(4+\frac{4}{\eta})\log\epsilon_\Phi - \text{pen}(\Phi) \\ + \frac{\sigma^2}{n}(4+\frac{4}{\eta})u\end{aligned}$$
occurs with a probability smaller than $\epsilon_\Phi e^{-u}$
- Recall now that
$$\frac{\sigma^2}{n}\left((2+2\eta)\dim\mathbb{X}^{\Phi}_{(n)} + (4+\frac{4}{\eta})(-\log\epsilon_\Phi)\right) \leq \text{pen}(\Phi)$$
and $\sum_\Phi \epsilon_\Phi = \epsilon$ so that the previous inequality implies that
$$\mathbb{P}\left(\sup_\Phi\left((1-\eta)\widehat{\Delta}(\Phi) - \widehat{\Delta}_n(\Phi) - \text{pen}(\Phi)\right) \geq \frac{\sigma^2}{n}(4+\frac{4}{\eta})u\right) \leq \epsilon e^{-u}$$
which implies
$$\mathbb{E}\left[\sup_\Phi\left((1-\eta)\widehat{\Delta}(\Phi) - \widehat{\Delta}_n(\Phi) - \text{pen}(\Phi)\right)\right] \leq \frac{\sigma^2}{n}(4+\frac{4}{\eta})\epsilon$$

2.4 AIC/BIC, General Penalization Scheme and Practical Model Selection

2.4.1 Probabilistic Distance and AIC / BIC Heuristics

Probabilistic Distance (Gaussian Model)

- Observation: $Y_i = f^\star(\underline{X}_i) + \epsilon_i$ at $\underline{X}_i$
- Estimation: $\widehat{Y|\underline{X}} \sim \text{N}(\widehat{f}(\underline{X}), \widehat{\sigma^2})$ at a generic $\underline{X}$.

Kullback-Leibler Divergences

- *Fixed Design:*
$$\text{KL}^{\otimes n}(Y|\underline{X}, \widehat{Y|\underline{X}}) = n^{-1}\sum_{i=1}^{n}\text{KL}(Y|\underline{X}_i, \text{N}(\widehat{f}(\underline{X}_i), \widehat{\sigma^2}))$$
- *Random Design:*
$$\text{KL}^{\otimes}(Y|\underline{X}, \widehat{Y|\underline{X}}) = \mathbb{E}\left[\text{KL}(Y|\underline{X}, \text{N}(\widehat{f}(\underline{X}), \widehat{\sigma^2}))\middle|\mathbb{X}_{(n)}, \epsilon_{(n)}\right]$$
- *Rk:* As $\widehat{f}$ and $\widehat{\sigma^2}$ are random, those quantities are *random*.
- *Gaussian error:* If $\epsilon_i \sim \text{N}(0, \sigma^2_\star)$
  - Kullback-Leiber divergence
$$\text{KL}(Y|\underline{X}, \text{N}(\widehat{f}(\underline{X}), \widehat{\sigma^2})) = \frac{1}{2}\log\frac{\widehat{\sigma^2}}{\sigma^2_\star} + \frac{\sigma^2_\star + |\widehat{f}(\underline{X}) - f^\star(\underline{X})|^2}{2\widehat{\sigma^2}} - \frac{1}{2}$$
  - Similar criterion than prediction error (cf Least Square vs ML)

Kullback-Leibler Divergence and Likelihood

Kullback-Leibler Divergence between $P^\star$ and P

$$\mathrm{KL}(P^\star, P) = \begin{cases} -\int \log \frac{dP}{dP^\star} dP^\star & \text{if } P \ll P^\star \\ +\infty & \text{otherwise} \end{cases}$$

- Not a distance but $\mathrm{KL}(P^\star, P) \geq 0$ and $\mathrm{KL}(P^\star, P) = 0 \Leftrightarrow P^\star = P$.
- Rescaled Log Likelihood: $L_n\left(\frac{dP}{d\mu}\right) = n^{-1} \log \mathcal{L}_{\mathbb{X}_{(n)}}\left(\frac{dP}{d\mu}\right) = n^{-1} \sum_{i=1}^{n} \log \frac{dP}{d\mu}(X_i)$

Kullback-Leibler Divergence and Likelihood

- If $X \sim P^\star$, $P \ll \mu$ and $P^\star \ll \mu$,

$$\mathbb{E}\left[-L_n\left(\frac{dP}{d\mu}\right) + L_n\left(\frac{dP}{d\mu}\right)\right] = \mathrm{KL}(P^\star, P)$$

- KL is the *natural distance* when using Maximum Likelihood!

Conditional Density Setting

Kullback-Leibler Divergences

- *Fixed Design:*

$$\mathrm{KL}^{\otimes n}(P^\star(Y|\underline{X}), P(Y|\underline{X})) = n^{-1} \sum_{i=1}^{n} \mathrm{KL}(P^\star(Y|\underline{X}_i), P(Y|\underline{X}_i))$$

- *Random Design:*

$$\mathrm{KL}^{\otimes}(P^\star(Y|\underline{X}), P(Y|\underline{X})) = \mathbb{E}\left[\mathrm{KL}(P^\star(Y|\underline{X}), P^\star(Y|\underline{X}))\middle|\mathbb{X}_{(n)}, \epsilon_{(n)}\right]$$

- Rescaled log-Likelihood: $L_n\left(\frac{dP}{d\mu}\right) = n^{-1} \sum_{i=1}^{n} \log \frac{dP}{d\mu}(Y_i|\underline{X}_i)$

Kullback-Leibler Divergence and Likelihood

- If $X|Y \sim P^\star$, $P \ll \mu$ and $P^\star \ll \mu$,

$$\mathbb{E}\left[-L_n\left(\frac{dP}{d\mu}\right) + L_n\left(\frac{dP}{d\mu}\right)\right] = \mathrm{KL}^{\otimes}(P^\star, P)$$

- True for random and fix design!

Maximum Likelihood

- Family $\{P_\theta\}_{\theta\in\Theta}$ of law such that $\frac{dP_\theta}{d\mu}(Y|\underline{X}) = p_\theta(Y|\underline{X})$.
- Gaussian Reg.: $P_\theta(Y|\underline{X}) = \mathrm{N}(\underline{X}^t\beta, s^2)$ with $\theta = (\beta, s^2)$ and $\Theta = \mathbb{R}^p \times \mathbb{R}^+$.

Maximum Likelihood Principle

- Choose

$$\widehat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} -L_n(p_\theta) = \operatorname*{argmin}_{\theta\in\Theta} -n^{-1}\sum_{i=1}^n \log p_\theta(Y_i|\underline{X}_i)$$

- Ideal target in the family

$$\theta^\dagger = \operatorname*{argmin}_{\theta\in\Theta} \mathrm{KL}^{\otimes}(P^\star, P_\theta) = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}\left[-L_n(p_\theta)\right] = \operatorname*{argmin}_{\theta\in\Theta} -L(p_\theta)$$

Empirical Loss Optimism

- By construction

$$\mathbb{E}\left[-L_n(p_{\widehat{\theta}})\right] \leq -L(p_{\theta^\dagger}) \leq -L(p_{\widehat{\theta}})$$

AIC Heuristic

AIC Correction

- Under some regularity assumptions, if $P^\star = P_{\theta^\dagger}$ then asymptotically

$$-L(p_{\widehat{\theta}}) \sim -L_n(p_{\widehat{\theta}}) + \frac{\dim\Theta}{n}$$

- Formula exists even if $P^\star \neq P_{\theta^\dagger}$ but correction depends on the unknown $\theta^\dagger$...

Principle

- Compute the ML estimate $P_{\widehat{\theta}_{\mathcal{S}}}$ for all model $\mathcal{S}$
- Compare the corrected risk

$$-L_n(p_{\widehat{\theta}_{\mathcal{S}}}) + \frac{\dim\Theta_{\mathcal{S}}}{n}$$

  where $\dim\Theta_{\mathcal{S}}$ is the dimension of the model $\mathcal{S}$.
- Select the one with the smallest corrected risk.

Gaussian case

- If $P^\star = \mathrm{N}(\Phi\left(\underline{X}\right)^t \beta^\dagger, \sigma_\star^2)$

$$\begin{aligned}
-L(p_{\widehat{\theta}}) &\sim \frac{1}{2}\log(2\pi\widehat{\sigma^2}) + n^{-1}\sum_{i=1}^{n}\frac{|Y_i - \Phi\left(vecX_i\right)^t\hat{\beta}|^2}{2\widehat{\sigma^2}} + \frac{dim\mathbb{X}^{\Phi}_{(n)}+1}{n} \\
&\sim \frac{1}{2}\log(2\pi\widehat{\sigma^2}) \\
&\quad + \frac{1}{2\widehat{\sigma^2}}\left(\underbrace{n^{-1}\sum_{i=1}^{n}|Y_i - \Phi\left(\underline{X}_i\right)^t\hat{\beta}|^2 + 2\widehat{\sigma^2}\frac{dim\mathbb{X}^{\Phi}_{(n)}}{n}}_{\text{Unbiased Risk Estimate}} + \frac{2\widehat{\sigma^2}}{n}\right)
\end{aligned}$$

- Same criterion than Unbiased Risk Estimate if $\sigma_\star^2$ is known!
- AIC heuristic is often used even if the $P^\star = P_{\beta^\dagger}$ assumption does not hold... but without any guarantee...

Sketch of Proof

- With a slight abuse of notation, we denote

$$\begin{aligned}
L_n(\theta) &= L_n(p_\theta) \\
L(\theta) &= L(p_\theta) = \mathbb{E}\left[L_n(p_\theta)\right] = \mathbb{E}\left[L_n(\theta)\right]
\end{aligned}$$

- Taylor expansion around $\theta^\dagger$

$$\begin{aligned}
L_n(\theta) &\sim L_n(\theta^\dagger) + \nabla L_n(\theta^\dagger)^t(\theta-\theta^\dagger) + \frac{1}{2}(\theta-\theta^\dagger)^t HL_n(\theta^\dagger)(\theta-\theta^\dagger) \\
L(\theta) &\sim L(\theta^\dagger) + \frac{1}{2}(\theta-\theta^\dagger)HL(\theta^\dagger)(\theta-\theta^\dagger)
\end{aligned}$$

- We deduce

$$\widehat{\theta} \sim \theta^\dagger - (HL_n(\theta^\dagger))^{-1}\nabla L_n(\theta^\dagger)$$

  and thus

$$\begin{aligned}
L_n(\widehat{\theta}) &\sim L_n(\theta^\dagger) - \frac{1}{2}\nabla L_n(\theta^\dagger)(HL_n(\theta^\dagger))^{-1}\nabla L_n(\theta^\dagger) \\
L(\widehat{\theta}) &\sim L(\theta^\dagger) + \frac{1}{2}\nabla L_n(\theta^\dagger)(HL_n(\theta^\dagger))^{-1}HL(\theta^\dagger)(HL_n(\theta^\dagger))^{-1}\nabla L_n(\theta^\dagger)
\end{aligned}$$

- Thus

$$\begin{aligned}
L(\widehat{\theta}) - L_n(\widehat{\theta}) \sim L(\theta^\dagger) - L_n(\theta^\dagger)& \\
+ \frac{1}{2}\nabla L_n(\theta^\dagger)^t(HL_n(\theta^\dagger))^{-1}HL(\theta^\dagger)(HL_n(\theta^\dagger))^{-1}\nabla L_n(\theta^\dagger)& \\
+ \frac{1}{2}\nabla L_n(\theta^\dagger)^t(HL_n(\theta^\dagger))^{-1}\nabla L_n(\theta^\dagger)&
\end{aligned}$$

- As $HL_n(\theta^\dagger)$ tends to $HL(\theta^\dagger)$, we have

$$L(\widehat{\theta}) - L_n(\theta) \sim L(\theta^\dagger) - L_n(\theta^\dagger) + \nabla L_n(\theta^\dagger)^t(HL(\theta^\dagger))^{-1}\nabla L_n(\theta^\dagger))$$

- Now, by the CLT,

$$\sqrt{n}\nabla L_n(\theta^\dagger) \to Z \sim \mathrm{N}(0, J(\theta^\dagger))$$

  with $J(\theta) = \mathbb{V}\mathrm{ar}\left[\nabla \log p_\theta(Y|\underline{X})\right]$ and thus

$$L(\widehat{\theta}) - L_n(\theta) \sim L(\theta^\dagger) - L_n(\theta^\dagger) + n^{-1}Z^tHL(\theta^\dagger)^{-1}Z.$$

- Taking the expectation leads to

$$\mathbb{E}\left[L(\widehat{\theta})-L_n(\widehat{\theta})\right]\sim n^{-1}\mathrm{Tr}(HL(\theta^\dagger)^{-1}J(\theta^\dagger))$$

- Now,

$$\begin{aligned}HL(\theta^\dagger)_{i,j}&=\mathbb{E}\left[\frac{\partial^2}{d\theta_i d\theta_j}(\log p_{\theta^\dagger}(Y|\underline{X}))\right]\\&=-\mathbb{E}\left[\frac{\frac{\partial}{d\theta_j}p_{\theta^\dagger}(Y|\underline{X})\times\frac{\partial}{d\theta_j}p_{\theta^\dagger}(Y|\underline{X})}{p^2_{\theta^\dagger}(Y|\underline{X})}\right]+\mathbb{E}\left[\frac{\frac{\partial^2}{d\theta_j d\theta_i}p_{\theta^\dagger}(Y|\underline{X})}{p_{\theta^\dagger}(Y|\underline{X})}\right]\\&=-\mathbb{E}\left[\left(\frac{\partial}{d\theta_i}\log p_{\theta^\dagger}(Y|\underline{X})\right)\left(\frac{\partial}{d\theta_j}\log p_{\theta^\dagger}(Y|\underline{X})\right)\right]+\Delta(\theta^\dagger)_{i,j}\end{aligned}$$

hence

$$HL(\theta^\dagger)=-J(\theta^\dagger)+\Delta(\theta^\dagger)$$

with $\Delta(\theta^\dagger)=\mathbb{E}\left[\frac{\frac{\partial^2}{d\theta_j d\theta_i}p_{\theta^\dagger}(Y|\underline{X})}{p_{\theta^\dagger}(Y|\underline{X})}\right]$

Asymptotic Control

- We have thus

$$\mathbb{E}\left[L(\widehat{\theta})-L_n(\widehat{\theta})\right]\sim-\frac{\dim\Theta}{n}+n^{-1}\mathrm{Tr}(HL(\theta^\dagger)^{-1}\Delta(\theta^\dagger))$$

- Now if $p_{\theta^\dagger}$ is the true law then

$$\begin{aligned}\Delta(\theta^\dagger)&=\mathbb{E}\left[\frac{\frac{\partial^2}{d\theta_j d\theta_i}p_{\theta^\dagger}(Y|\underline{X})}{p_{\theta^\dagger}(Y|\underline{X})}\right]\\&=\mathbb{E}\left[\int\frac{\partial^2}{d\theta_j d\theta_i}p_{\theta^\dagger}(y|\underline{X})dy\right]=0\end{aligned}$$

and we obtain a *bias of* p/n! (Usual AIC)

Bayesian Approach and BIC

- Model rewriting: $\mathbb{P}_{\theta_\mathcal{S}}\left(\mathbb{Y}_{(n)}\middle|\mathbb{X}_{(n)}\right)=\mathbb{P}\left(\mathbb{Y}_{(n)}\middle|\mathbb{X}_{(n)},\theta_\mathcal{S},\mathcal{S}\right)$

Bayesian Approach

- Define a prior law on $\mathcal{S}$ and $\theta_\mathcal{S}|\mathcal{S}$
- Use the Bayes formula to obtain the law of $(\theta_\mathcal{S},\mathcal{S})|\mathbb{Y}_{(n)},\mathbb{X}_{(n)}$ or $\mathcal{S}|\mathbb{Y}_{(n)},\mathbb{X}_{(n)}$
- Use this law to characterize the uncertainty on the estimates or to do prediction.
- Very powerful framework!

Bayesian Information Criterion

$$\log\mathbb{P}\left(\mathcal{S}\middle|\mathbb{Y}_{(n)},\mathbb{X}_{(n)}\right)\sim n\left(L_n(\widehat{\theta}_\mathcal{S})-\frac{\log n}{2}\frac{\dim\Theta_\mathcal{S}}{n}\right)$$

$$-\log\mathbb{P}\left(\mathcal{S}\middle|\mathbb{Y}_{(n)},\mathbb{X}_{(n)}\right)\sim n\left(-L_n(\widehat{\theta}_\mathcal{S})+\frac{\log n}{2}\frac{\dim\Theta_\mathcal{S}}{n}\right)$$

- Select the best model through a corrected empirical risk...

Sketch of Proof

- Bayesian Approach:

$$\begin{aligned}\log \mathbb{P}\left(\mathcal{S}\middle|\mathbb{Y}_{(n)},\mathbb{X}_{(n)}\right) &= \log \int \mathbb{P}\left(\mathcal{S},\theta\middle|\mathbb{Y}_{(n)},\mathbb{X}_{(n)}\right)d\theta\\ &= \log \int \frac{\mathbb{P}\left(Y|\theta,\mathcal{S}\right)\mathbb{P}\left(\theta|\mathcal{S}\right)\mathbb{P}\left(\mathcal{S}\right)}{\mathbb{P}\left(Y\right)}d\theta\\ &= \log \int e^{n(L_n(\theta)+n^{-1}\log\mathbb{P}(\theta|\mathcal{S}))}d\theta\\ &\quad + \log\mathbb{P}\left(\mathcal{S}\right) - \log\mathbb{P}\left(Y\right)\end{aligned}$$

- Using a Taylor expansion around the ML parameter estimate $\widehat{\theta}$ yields $L_n(\theta) \sim L_n(\widehat{\theta})+\frac{1}{2}(\theta-\widehat{\theta})^t(HL_n(\widehat{\theta}))(\theta-\widehat{\theta})$

- We deduce that

$$\begin{aligned}\log\int e^{n(L_n(\theta)+n^{-1}\log\mathbb{P}(\theta|\mathcal{S}))}d\theta &\sim nL_n(\widehat{\theta})\\ &+ \log\int e^{-\frac{1}{2}(\theta-\widehat{\theta})^t(-nHL_n(\widehat{\theta}))(\theta-\widehat{\theta})+\log\mathbb{P}(\theta|\mathcal{S}))}d\theta\end{aligned}$$

- If we assume the prior is flat around $\widehat{\theta}$, we obtain

$$\begin{aligned}\log\int e^{n(L_n(\theta)+n^{-1}\log\mathbb{P}(\theta|\mathcal{S}))}d\theta \sim nL_n(\theta) &+ \log\mathbb{P}\left(\widehat{\theta}\middle|\mathcal{S}\right)\\ &+\frac{\dim\Theta_{\mathcal{S}}}{2}\log 2\pi - \frac{\dim\mathcal{S}}{2}\log n\\ &-\frac{1}{2}\log\det(-HL_n(\widehat{\theta}))\end{aligned}$$

- Hence

$$\begin{aligned}\log\mathbb{P}\left(\mathcal{S}\middle|\mathbb{Y}_{(n)},\mathbb{X}_{(n)}\right) &\sim n\left(L_n(\widehat{\theta}) - \frac{\log n - \log 2\pi}{2}\frac{\dim\Theta_{\mathcal{S}}}{n}\right)\\ &\quad -\frac{1}{2}\log\det(HL_n(\widehat{\theta}))\\ &\quad +\log\mathbb{P}\left(\widehat{\theta}\middle|\mathcal{S}\right)+\log\mathbb{P}\left(\mathcal{S}\right)-\log\mathbb{P}\left(Y\right)\\ &\sim n\left(L_n(\widehat{\theta})-\frac{\log n}{2}\frac{\dim\Theta_{\mathcal{S}}}{n}\right)\end{aligned}$$

Bayesian Information Criterion

$$\log\mathbb{P}\left(\mathcal{S}\middle|\mathbb{Y}_{(n)},\mathbb{X}_{(n)}\right)\sim n\left(L_n(\widehat{\theta})-\frac{\log n}{2}\frac{\dim\Theta_{\mathcal{S}}}{n}\right)$$

$$-\log\mathbb{P}\left(\mathcal{S}\middle|\mathbb{Y}_{(n)},\mathbb{X}_{(n)}\right)\sim n\left(-L_n(\widehat{\theta})+\frac{\log n}{2}\frac{\dim\Theta_{\mathcal{S}}}{n}\right)$$

Penalized Model Selection

- Two settings but similar methodology.

Penalized Model Selection

- Compute a ML estimate $P_{\widehat{\theta}_{\mathcal{S}}}$ for all model $\mathcal{S}$
- Compute the corresponding empirical risks $-L(P_{\widehat{\theta}_{\mathcal{S}}})$.
- Correct those risks by adding a penalty proportional to the dimension

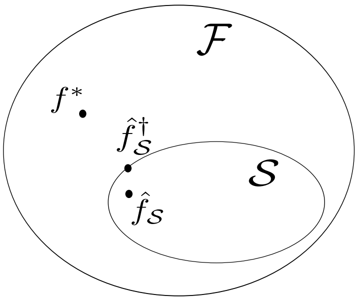
$$-L(P_{\widehat{\theta}_{\mathcal{S}}}) + \lambda \frac{\dim \Theta_{\mathcal{S}}}{n}$$

  where $\dim \Theta_{\mathcal{S}}$ is the dimension of the model $\mathcal{S}$.
- Select the one with the smallest corrected risk.

- Variation on the choice of λ!
- Can we use a data driven choice?

2.4.2 Linear Model(s) and General Penalization

Generic Bias-Variance Dilemma



- General setting:
  - $\mathcal{F} = \{$measurable fonctions$\mathcal{X} \to \mathcal{Y}\}$
  - Best solution: $f^\star = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$
  - Class $\mathcal{S} \subset \mathcal{F}$ of functions
  - Ideal target in $\mathcal{S}$: $f^\dagger_{\mathcal{S}} = \operatorname{argmin}_{f \in \mathcal{S}} \mathcal{R}(f)$
  - Estimate in $\mathcal{S}$: $\widehat{f}_{\mathcal{S}}$ obtained with some procedure

Approximation error and estimation error (Bias/Variance)

$$\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^\star) = \underbrace{\mathcal{R}(f^\star_{\mathcal{S}}) - \mathcal{R}(f^\star)}_{\text{Approximation error}} + \underbrace{\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^\star_{\mathcal{S}})}_{\text{Estimation error}}$$

- Approx. error can be large if the model $\mathcal{S}$ is not suitable and is independent of n.
- Estimation error can be large if the model is complex and decreases with n.

- *Best* choice *depends* on the *unknown best function* and $\boldsymbol{n}$!
- How to define different models $\mathcal{S}$?

Variable Selection

- *Setting*: Linear model = prediction of Y by $\underline{X}^{\Phi t}\beta$.

Model coefficients

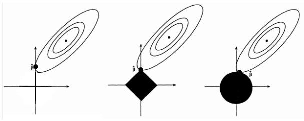
- Model entirely specified by β.
- Coefficientwise:
  - $\beta_i = 0$ means that the ith covariate is not used.
  - $\beta_i \sim 0$ means that the ith covariate as a *low* influence...

- If some covariates are useless, better use a simpler model...

Submodels

- *Simplify* the model through a constraint on β!
- Examples:
  - Support: Impose that $\beta_i = 0$ for $i \notin I$.
  - Support size: Impose that $\|\beta\|_0 = \sum_{i=1}^{p} \mathbf{1}_{\beta_i \neq 0} < C$
  - Norm: Impose that $\|\beta\|_q < C$ with $1 \leq q$ (Often $q = 2$ or $q = 1$)

Norms and Sparsity



Sparsity

- β is sparse if its number of non-zero coefficients (ℓ_0) is small...
- Easy interpretation in term of dimension/complexity.

Norm Constraint and Sparsity

- Sparsest solution obtained by definition with the ℓ_0 norm.
- No induced sparsity with the ℓ_2 norm...
- Sparsity with the ℓ_1 norm (can even be proved to be the same than with the ℓ_0 norm under some assumptions).
- Geometric explanation.

Constraint and Penalization

Constrained Optimization

- Choose a constant C.
- Compute β as

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^p, \|\beta\|_q \leq C} n^{-1} \sum_{i=1}^{n} |Y_i - \underline{X}_i^{\Phi t} \beta|^2$$

- Corresponds to a model $\mathcal{S} = \{f_{\Phi,\beta}, \beta \in \mathbb{R}^p, \|\beta\|_q \leq C\}$

Lagrangian Reformulation

- Choose λ and compute β as

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^p} n^{-1} \sum_{i=1}^{n} |Y_i - \underline{X}_i^{\Phi t} \beta|^2 + \lambda \|\beta\|_q^{q'}$$

  with $q' = q$ except if $q = 0$ where $q' = 1$.
- Easier calibration through λ... but no explicit model $\mathcal{S}$...

- *Rk:* $\|\beta\|_q$ is not scaling invariant if $q \neq 0$...
- Initial rescaling issue.

Penalization

Penalized Linear Model

- Minimization of

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^d} n^{-1} \sum_{i=1}^{n} |Y_i - \underline{X}_i^{\Phi t} \beta|^2 + \operatorname{pen}(\beta)$$

  where $\operatorname{pen}(\beta)$ is a (sparsity promoting) penalty
- Variable selection if β is sparse.

Classical Penalties

- AIC: $\operatorname{pen}(\beta) = \lambda \|\beta\|_0$ (non convex / sparsity)
- Ridge: $\operatorname{pen}(\beta) = \lambda \|\beta\|_2^2$ (convex / no sparsity)
- Lasso: $\operatorname{pen}(\beta) = \lambda \|\beta\|_1$ (convex / sparsity)
- Elastic net: $\operatorname{pen}(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$ (convex / sparsity)

- Easy optimization if pen (and the loss) is convex...
- *Need to specify* λ*!*

Penalization and Cross-Validation

Practical Selection Methodology

- Choose a penalty shape $\widetilde{\text{pen}}$.
- Compute a CV error for a penalty $\lambda\widetilde{\text{pen}}$ for all $\lambda \in \Lambda$.
- Determine $\widehat{\lambda}$ the λ minimizing the CV error.
- Compute the parameters with a penalty $\widehat{\lambda}\widetilde{\text{pen}}$.

Why not using only CV?

- *If* the penalized likelihood minimization is easy, much cheaper to compute the CV error for all $\lambda \in \Lambda$ than for all possible estimators...
- CV performs best when the set of candidates is not too big (or is structured...)

2.4.3 Practical Model Selection

Smooth Optimization

Explicit Least Square Solution

Empirical Loss Minimization

- Minimization of

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^p} n^{-1} \sum_{i=1} |Y_i - \underline{X}_i^{\Phi t}\beta|^2 \Leftrightarrow \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|\mathbb{Y}_{(n)} - \mathbb{X}_{(n)}^{\Phi}\beta\|^2$$

- Explicit solution:

$$\hat{\beta} = \left(\mathbb{X}_{(n)}{}^{\Phi t}\mathbb{X}_{(n)}^{\Phi}\right)^{-1} \mathbb{X}_{(n)}{}^{\Phi t}\mathbb{Y}_{(n)}$$

- Invertibility assumption...
- Computational cost $O(n \times p^2 + p^3)$:
  - Computation of $\mathbb{X}_{(n)}{}^{\Phi t}\mathbb{Y}_{(n)}$: $O(n \times p)$
  - Computation of $\mathbb{X}_{(n)}{}^{\Phi t}\mathbb{X}_{(n)}^{\Phi}$: $O(n \times p^2)$
  - Resolution of the p dimensional linear system: p^3 (or rather $O(p^{2.4})$ with a ridiculously big constant)
- Can be prohibitive!

Iterative Solution

Empirical Loss Minimization

- Minimization of

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^p} n^{-1}\sum_{i=1} |Y_i - \underline{X}_i^{\Phi\,t}\beta|^2 \Leftrightarrow \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|\mathbb{Y}_{(n)} - \mathbb{X}^{\Phi}_{(n)}\beta\|^2$$

- Iterative gradient descent solution:

$$\hat{\beta}_{k+1} = \hat{\beta}_k - h_k \underbrace{2{\mathbb{X}_{(n)}}^{\Phi\,t}\left(\mathbb{Y}_{(n)} - \mathbb{X}^{\Phi}_{(n)}\hat{\beta}_k\right)}_{\nabla\|\mathbb{Y}_{(n)} - \mathbb{X}^{\Phi}_{(n)}\beta\|^2}$$

- Convexity guarantees the convergence up to the very important choice of h_k (cf optimization theory)!
- Computation cost per iteration $O(n \times p)$:
  - Update of the residual $\mathbb{Y}_{(n)} - \mathbb{X}^{\Phi}_{(n)}\hat{\beta}_k$: $O(n \times p)$
  - Computation of the matrix product: $O(p \times n)$
  - Computation of the difference $O(p)$
- Advantageous if
  - good solution obtained in less than $\max(p, p^2/n)$ steps,
  - p too large for the numerical resolution of the equation.

Other Iterative Solutions

- Several variation of gradient descent!
- *Specific Quadratic Form Algorithm:*
  - Conjugate gradient descent
  - Less than p steps required...
- *(Block) Coordinate Descent:*
  - Only modify one (or a small number of) coordinate(s) at a time
  - Cost per iteration reduced to $O(n)$
- *Stochastic Gradient Descent:*
  - Replace the *true* gradient by a stochastic approximation

$$\frac{1}{|I|}\sum_{i\in I} 2\underline{X}_i^{\Phi}(Y_i - \underline{X}_i^{\Phi\,t}\hat{\beta}_k)$$

    with I a small random subset of $\{1, \ldots, n\}$
  - Cost per iteration reduced to $O(d|I|)$
- *Combination* of those methods and *dual* approaches...

ℓ^2 Penalization (Ridge Regression)

ℓ^2 Penalized Empirical Loss Minimization

- Minimization of

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^p} n^{-1}\sum_{i=1}|Y_i - \underline{X}_i^{\Phi t}\beta|^2 + \lambda\|\beta\|_2^2$$

- Introduced originally in inverse problem

- Explicit solution:

$$\hat{\beta} = \left(\mathbb{X}_{(n)}{}^{\Phi t}\mathbb{X}^{\Phi}_{(n)} + \lambda \mathrm{Id}_{(p)}\right)^{-1}\mathbb{X}_{(n)}{}^{\Phi t}\mathbb{Y}_{(n)}$$

- Gradient Descent:

$$\nabla\left(n^{-1}\sum_{i=1}|Y_i-\underline{X}_i^{\Phi t}\beta|^2+\lambda\|\beta\|_2^2\right) = n^{-1}\sum_{i=1} 2\underline{X}_i^{\Phi}(Y_i - \underline{X}_i^{\Phi t}\beta) + 2\lambda\beta$$

- Same algorithms than for the non regularized solution!
- Need to choose λ (to guaranty good performance)!

Exploration

ℓ^0 Penalization

ℓ^0 Penalized Empirical Loss Minimization

- Minimization of

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^p} n^{-1}\sum_{i=1}^{n}|Y_i - \underline{X}_i^{\Phi t}\beta|^2 + \lambda\|\beta\|_0$$

- Equivalent model selection reformulation:
  - For every $I \subset \{1,\ldots,p\}$, compute

$$\hat{\beta}_I = \operatorname*{argmin}_{\beta_J, \beta_{J,i}=0 \forall i \notin I} n^{-1}\sum_{i=1}^{n}|Y_i - \underline{X}_i^{\Phi t}\beta|^2$$

  - Determine

$$\hat{I} = \operatorname*{argmin}_{i} n^{-1}\sum_{i=1}^{n}|Y_i - \underline{X}_i^{\Phi t}\hat{\beta}_I|^2 + \lambda|I|$$

- Amounts to use a *projection* Φ for each support.
- Need to perform the optimization (non convex/non smooth)!
- Need to choose λ (to guaranty good performance)!

Practical ℓ^0 Penalization

Exact Minimization

- Easy optimization for a given support!
- Very different situation for the support...
- Bruteforce exploration of the support = combinatorial problem.
- 2^p models (supports) to be explored!
- Only possible if p is (very) small or the design matrix is orthogonal!
- *Complete* exploration possible in a special case!

Orthogonal Matrix Design

- If $\mathbb{X}^{\Phi}_{(n)}$ is a orthogonal matrix:

$$\begin{aligned} & n^{-1}\|\mathbb{Y}_{(n)} - \mathbb{X}^{\Phi}_{(n)}\beta\|^2 + \lambda\|\beta\|_0 \\ & \quad = \sum_{k=1}^{p} \left[n^{-1} \left(\left({\mathbb{X}_{(n)}}^{\Phi t} \mathbb{Y}_{(n)} \right)_k - \beta_k \right)^2 + \lambda \mathbf{1}_{\beta_k \neq 0} \right] \end{aligned}$$

- Penalized Least Square solution:

$$\hat{\beta}_k = \begin{cases} \left(\mathbb{X}^{t}_{(n)} \mathbb{Y}_{(n)} \right)_k & \text{if } \left| \left({\mathbb{X}_{(n)}}^{\Phi t} \mathbb{Y}_{(n)} \right)_k \right|^2 > n\lambda \\ 0 & \text{otherwise} \end{cases}$$

- *Thresholding:* set to 0 of the coefficients of the classical Least Square solution smaller than $\sqrt{n\lambda}$ in absolute values.
- May occurs when the design matrix is related to an orthonormal basis (polynomials, Fourier... or PCA)

Clever Exploration

- Minimization of the criterion but without an exhaustive exploration of the subsets.
- Generic strategy:
  - Start with a pool of subsets of size P
  - Create a larger pool of size PC by adding and/or removing variables from the previous subset
  - Keep only the best P subset according to the criterion and iterate
- Variations on the size of the subsets, the initial subsets, the rule to add and remove variables, the criterion...
- Forward, Backward, Forward/Backward, Stochastic (Genetic) Algorithm...

Forward strategy

- Start with an empty model
- At each step, create a larger collection by creating models equal to the current one plus any variable not used in the current model (one at a time)
- Modify the current model if the best model within the new collection leads to a reduction of the criterion.

Backward strategy

- Start with the full model.
- At each step, create a larger collection by creating models equal to the current one minus any variable used in the current model (one at a time)
- Modify the current model if the best model within the new collection leads to a reduction of the criterion.

Forward/Backward strategy

- Start with the full model.
- At each step, create a larger collection by creating models equal to the current one plus any variable not used in the current model (one at a time) and to the current one minus any variable used in the current model (one at a time)
- Modify the current model if the best model within the new collection leads to a reduction of the criterion.

- Various Stochastic (Genetic) Algorithm...
- Stability issue...

Convex Optimization

ℓ^1 Penalization

ℓ^1 Penalized Empirical Loss Minimization

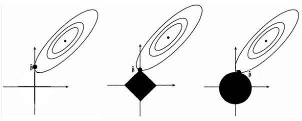
- Minimization of

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^p} n^{-1} \sum_{i=1} |Y_i - \underline{X}_i^{\Phi\, t} \beta|^2 + \lambda \|\beta\|_1$$

- Introduced originally as a convexification of the ℓ^0 loss...

- Non smooth but convex function and thus existing fast optimization algorithm.
- Need to choose λ (to guaranty good performance)!

ℓ^1 Penalization and Sparsity



Sparsification Properties

- Let

$$C(\beta) = n^{-1} \sum_{i=1} |Y_i - {\underline{X}_i^{\Phi}}^{t}\beta|^2 + \lambda\|\beta\|_1$$

- Convex subgradient property: $\hat{\beta} = \operatorname{argmin} C(\beta) \Leftrightarrow 0 \in \delta C(\hat{\beta})$:

$$n^{-1} \sum_{i=1}^{n} 2\underline{X}_{i,k}(Y_i - {\underline{X}_i^{\Phi}}^{t}\beta) \begin{cases} = \lambda & \text{if } \hat{\beta}_k < 0 \\ \in [-\lambda, \lambda] & \text{if } \hat{\beta}_k = 0 \\ = -\lambda & \text{if } \hat{\beta}_k > 0 \end{cases}$$

- More *flexibility* at $\beta_k = 0$...

Convex Penalties

Penalized Likelihood

- Minimization of

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^p} n^{-1} \sum_{i=1} |Y_i - {\underline{X}_i^{\Phi}}^{t}\beta|^2 + \lambda\|\beta\|_1$$

- Convex function in $\beta \in \mathbb{R}^d$!

Convex Optimization

- A local minimum is a global minimum!
- No possibility to be trapped in a local minimum!
- Several very efficient minimization algorithm exists.
- Huge progress recently (motivated by big data...).
- Canonical algorithm: *(sub)gradient descent.*

Subgradient Descent Algorithm

- Start with a point $\hat{\beta}_0$
- for $k = 1, \ldots$ until *convergence* repeat:
  - $\hat{\beta}_{k+1} \leftarrow \hat{\beta}_k - h_k \delta f(\hat{\beta}_k)$ where $\delta f(\beta^k)$ is any subgradient of f at β^k

Step/Learning Rate Choice

- Choice of h_k crucial!
- Provable convergence toward a minimum for suitable choice!

- Better schemes exist (proximal algorithms, accelerated algorithms, primal/dual algorithms...)
- Convex Optimization is a subject on his own.

Chapter 3

Extensions of the Linear Model

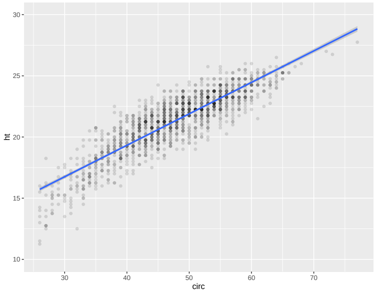
Outline

Contents

3.1 Binary Outcome and Logistic Regression

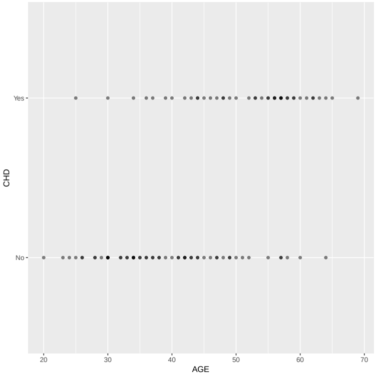
3.1.1 Binary outcome, Bernoulli Models and Parameterization

Eucalyptus



- Goal: predict the height from circumference
- $\underline{X} = \boldsymbol{\mathit{circ}} = \text{circumference} \in \mathbb{R}$.
- $Y = ht = \text{height} \in \mathbb{R}$.
- *Linear* model: $\mathbb{E}\left[Y|\underline{X}\right] \simeq \underline{X}^t\beta$
- *Gaussian* model: $Y|\underline{X} \simeq \mathrm{N}(\mathbb{E}\left[Y|\underline{X}\right], \sigma_\star^2)$

Heart Disease



- Goal: predict the presence or the absence of a disease given the age
- $\underline{X} = \boldsymbol{\mathit{age}} = \text{age} \in \mathbb{R}$.
- $Y = CHD = \text{Heart Disease} \in \{\text{Yes}, \text{No}\}$

- *Linear* model ??? / *Gaussian* model ???

Bernoulli Law

- *Goal:* predict $Y \in \{-1, 1\}$ given $\underline{X} \in \mathbb{R}^d$
- Bernoulli $\mathcal{B}(p)$: law on $\{-1, 1\}$ such that

$$Y \sim \mathcal{B}(p) \Leftrightarrow \begin{cases} \mathbb{P}\left(Y = 1\right) = p \\ \mathbb{P}\left(Y = -1\right) = 1 - p \end{cases}$$

Conditional Bernoulli Law

- If $Y \in \{-1, 1\}$ then

$$Y | \underline{X} \sim \mathcal{B}\left(\mathbb{P}\left(Y = 1 | \underline{X}\right)\right)$$

- No approximation at all!
- Better situation than in the regression case where the Gaussian model was an approximation!

Conditional Bernoulli Models

- To any function $p : \mathbb{R}^d \mapsto [0, 1]$, one can associate the conditional Bernoulli law $Y | \underline{X} \sim \mathcal{B}(p(\underline{X}))$

Parametric Conditional Bernoulli Model

- To any function $\Theta \mapsto (\mathbb{R}^d \mapsto [0, 1])$, one can associate the parametric model

$$\mathcal{S} = \{\mathcal{B}(p_\theta(\underline{X})), \theta \in \Theta\}$$

- Two questions arise:
  - How to choose the parameterization?
  - How to choose the parameter given the data?

Parameterization

Linear Parameterization

- $p_{\Phi,\beta}(\underline{X}) = h(\Phi(\underline{X})^t \beta)$ with h a non decreasing link function from $\mathbb{R}$ to $[0, 1]$
- Most classical choices: $\Phi = \mathrm{Id}_{(p)}$ and

$$\begin{aligned} h(t) &= \frac{e^t}{1 + e^t} && \text{logit or logistic} \\ h(t) &= F_{\mathrm{N}}(t) && \text{probit} \\ h(t) &= 1 - e^{-e^t} && \text{log-log} \end{aligned}$$

- More complex parameterizations are possible... but we will only consider with the linear ones...

Maximum Likelihood Estimate

Probabilistic Model

- By construction, $Y|\underline{X}$ follows $\mathcal{B}(\mathbb{P}\left(Y=+1|\underline{X}\right))$
- *Modelization:* Approximation of $Y|\underline{X}$ by $\mathcal{B}(h({\underline{X}^{\Phi}}^t\beta))$
- *Natural* probabilistic choice for β: β minimizing the $\mathrm{KL}^{\otimes}$ divergence between $\mathcal{B}(\mathbb{P}\left(Y=1|X\right))$ and $\mathcal{B}(h({\underline{X}^{\Phi}}^t\beta))$.

Kullback-Leibler Divergence

$$\begin{aligned}
&\mathrm{KL}^{\otimes}(\mathcal{B}(\mathbb{P}\left(Y=1|\underline{X}\right)),\mathcal{B}(h({\underline{X}^{\Phi}}^t\beta))\\
&\quad=\mathbb{E}\Big[-\mathbb{P}\left(Y=1|\underline{X}\right)\log(h({\underline{X}^{\Phi}}^t\beta))\\
&\qquad\qquad-(1-\mathbb{P}\left(Y=1|\underline{X}\right))\log(1-h({\underline{X}^{\Phi}}^t\beta))\Big]+C_{\underline{X},Y}
\end{aligned}$$

- Proof:

$$\begin{aligned}
&\mathrm{KL}^{\otimes}(\mathcal{B}(\mathbb{P}\left(Y=1|\underline{X}\right)),\mathcal{B}(h({\underline{X}^{\Phi}}^t\beta))\\
&\quad=\mathbb{E}\Bigg[\mathbb{P}\left(Y=1|\underline{X}\right)\log\frac{\mathbb{P}\left(Y=1|\underline{X}\right)}{h({\underline{X}^{\Phi}}^t\beta)}\\
&\qquad\qquad+(1-\mathbb{P}\left(Y=1|\underline{X}\right))\log\frac{1-\mathbb{P}\left(Y=1|\underline{X}\right)}{1-h({\underline{X}^{\Phi}}^t\beta)}\Bigg]\\
&\quad=\mathbb{E}\Big[-\mathbb{P}\left(Y=1|\underline{X}\right)\log(h({\underline{X}^{\Phi}}^t\beta))\\
&\qquad\qquad-(1-\mathbb{P}\left(Y=1|\underline{X}\right))\log(1-h({\underline{X}^{\Phi}}^t\beta))\Big]+C_{\underline{X},Y}
\end{aligned}$$

log-likelihood

- *Target:*

$$\begin{aligned}
&\mathrm{KL}^{\otimes}(\mathcal{B}(\mathbb{P}\left(Y=1|\underline{X}\right)),\mathcal{B}(h({\underline{X}^{\Phi}}^t\beta))\\
&\quad=\mathbb{E}\Big[-\mathbb{P}\left(Y=1|\underline{X}\right)\log(h({\underline{X}^{\Phi}}^t\beta))\\
&\qquad\qquad-(1-\mathbb{P}\left(Y=1|\underline{X}\right))\log(1-h({\underline{X}^{\Phi}}^t\beta))\Big]+C_{\underline{X},Y}
\end{aligned}$$

- *Empirical counterpart* of $\mathrm{KL}^{\otimes}$ = opposite of the log-likelihood:

$$-n^{-1}\sum_{i=1}^{n}\left(\mathbf{1}_{y_i=1}\log(h({\underline{X}_i^{\Phi}}^t\beta))+\mathbf{1}_{y_i=-1}\log(1-h({\underline{X}_i^{\Phi}}^t\beta))\right)$$

- Minimization possible if h is regular leading to a value $\hat{\beta}$.
- Provides an *estimate of* $\boldsymbol{Y|\underline{X}}\sim\mathcal{B}(\mathbb{P}\left(Y=1|\underline{X}\right))$:

$$\boldsymbol{\mathcal{B}(h({\underline{X}^{\Phi}}^t\hat{\beta}))}.$$

Link with Classification

- *Classification:* predict Y for a given $\underline{X}$ by $f(\underline{X}) \in \{-1, 1\}$.
- *Quality*: *average prediction loss* $\mathbb{E}\left[\ell(Y, f(\underline{X}))\right]$ with ℓ a loss measuring the *misclassification cost*:

True negative:	$\ell(-1,-1) = 0$	False positive:	$\ell(-1,1)) = C_{FP}$
False negative:	$\ell(1,-1) = C_{FN}$	True positive:	$\ell(1,1) = 0$

Optimal Choice and Conditional Law

- The best possible f is the Bayes classifier:

$$f(\underline{X}) = \begin{cases} -1 & \text{if } C_{FN}\mathbb{P}\left(Y = 1|\underline{X}\right) \leq C_{FP}\mathbb{P}\left(Y = -1|\underline{X}\right) \\ & \quad \Leftrightarrow \mathbb{P}\left(Y = 1|\underline{X}\right) \leq \frac{C_{FP}}{C_{FN}+C_{FP}} \Leftrightarrow \frac{\mathbb{P}(Y=1|\underline{X})}{\mathbb{P}(Y=-1|\underline{X})} \leq \frac{C_{FP}}{C_{FN}} \\ 1 & \text{otherwise} \end{cases}$$

- If $C_{FP} = C_{FN}$: choice of the most probable class!
- Natural *plug-in rule* where $\mathbb{P}\left(Y = 1|\underline{X}\right)$ is replaced by its estimate $\widehat{\mathbb{P}\left(Y = 1|\underline{X}\right)} = h(\underline{X}^{\Phi t}\hat{\beta})$

Proof

- For any decision function f,

$$\begin{aligned} \mathbb{E}\left[\ell(Y, f(\underline{X}))\right] &= \mathbb{E}\left[\mathbb{E}\left[\ell(Y, f(\underline{X}))|\underline{X}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[C_{FN}\mathbf{1}_{f(\underline{X})=-1}\mathbf{1}_{Y=1} + C_{FP}\mathbf{1}_{f(\underline{X})=1}\mathbf{1}_{Y=-1}\middle|\underline{X}\right]\right] \\ &= \mathbb{E}\Big[C_{FN}\mathbf{1}_{f(\underline{X})=-1}\mathbb{P}\left(Y = 1|\underline{X}\right) \\ &\qquad + C_{FP}\mathbf{1}_{f(\underline{X})=1}\mathbb{P}\left(Y = 1|\underline{X}\right)\Big] \end{aligned}$$

- For a given $\underline{X}$, one should thus minimize

$$C_{FN}\mathbb{P}\left(Y = 1|\underline{X}\right)\mathbf{1}_{f(\underline{X})=-1} + C_{FP}\mathbb{P}\left(Y = -1|\underline{X}\right)\mathbf{1}_{f(\underline{X})=1}$$

and thus choose

$$f(\underline{X}) = \begin{cases} -1 & \text{if } C_{FN}\mathbb{P}\left(Y = 1|\underline{X}\right) \leq C_{FP}\mathbb{P}\left(Y = -1|\underline{X}\right) \\ 1 & \text{otherwise} \end{cases}$$

3.1.2 Logistic Regression

Logistic Model and Odds

Logistic model

- *Natural* choice for the link function h:

$$h(t) = \frac{e^t}{1 + e^t} \quad \text{logit or logistic}$$

- Logistic model: corresponding *linear* model

$$\mathbb{P}\left(Y = 1|\underline{X}\right) = \frac{e^{\underline{X}^{\Phi t}\beta}}{1 + e^{\underline{X}^{\Phi t}\beta}} = h(\underline{X}^{\Phi t}\beta)$$

- Special case of the Bernoulli models...
- Interpretation in terms of *odd* $\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=-1)}$ (cf bookmaker)

Logistic and Odd

- With the logistic link function, Bernoulli law $\mathcal{B}(h(t))$ satisfies

$$\underbrace{\frac{\mathbb{P}\left(Y=1\right)}{\mathbb{P}\left(Y=-1\right)}}_{\text{Odd}} = e^t \Leftrightarrow \log \frac{\mathbb{P}\left(Y=1\right)}{\mathbb{P}\left(Y=-1\right)} = t$$

- *Logistic model*: linear model on the logarithm of the conditional odd

$$\log \frac{\mathbb{P}\left(Y=1|\underline{X}\right)}{\mathbb{P}\left(Y=-1|\underline{X}\right)} = \underline{X}^{\Phi t}\beta$$

- Increase of $\underline{X}^{\Phi t}\beta \sim$ an increase of the log of the odd.
- Interpretation of the sign of the coefficients.

Logistic Model and Classification

Optimal Choice and Conditional Law

- The best possible f is the Bayes classifier:

$$f(\underline{X}) = \begin{cases} -1 & \text{if } \frac{\mathbb{P}(Y=1|\underline{X})}{\mathbb{P}(Y=-1|\underline{X})} \leq \frac{C_{FP}}{C_{FN}} \\ 1 & \text{otherwise} \end{cases}$$

Logistic Model Classifier

- Plugin strategy:

$$f_\beta(\underline{X}) = \begin{cases} -1 & \text{if } \log \frac{\mathbb{P}_\beta(Y=1|\underline{X})}{\mathbb{P}_\beta(Y=-1|\underline{X})} \leq \log \frac{C_{FP}}{C_{FN}} \\ & \quad \Leftrightarrow \Phi(\underline{X})^t\beta \leq \log \frac{C_{FP}}{C_{FN}} \\ 1 & \text{otherwise} \end{cases}$$

- *Linear boundary decision.*

Logistic Regression and Minimization

Likelikood Rewriting

- Opposite of the log-likelihood:

$$\begin{aligned} & - n^{-1} \sum_{i=1}^{n} \left(\mathbf{1}_{Y_i=1} \log(h(\underline{X}_i^{\Phi t}\beta)) + \mathbf{1}_{Y_i=-1} \log(1 - h(\underline{X}_i^{\Phi t}\beta)) \right) \\ & = -n^{-1} \sum_{i=1}^{n} \left(\mathbf{1}_{Y_i=1} \log \frac{e^{\underline{X}_i^{\Phi t}\beta}}{1 + e^{\underline{X}_i^{\Phi t}\beta}} + \mathbf{1}_{Y_i=-1} \log \frac{1}{1 + e^{\underline{X}_i^{\Phi t}\beta}} \right) \\ & = n^{-1} \sum_{i=1}^{n} \log \left(1 + e^{-Y_i(\underline{X}_i^{\Phi t}\beta)} \right) \end{aligned}$$

- *Convex and smooth function of β*
- Easy optimization: gradient descent,...

- Logistic regression: $\mathcal{C}^2$ log-Likelihood
- Use of a second order algorithm (Newton Algorithm)
- Notation:

$$\Pi_{\Phi,\beta}(\underline{X}_i) = \mathbb{P}_{\Phi,\beta}\left(Y_i = 1|\underline{X}_i\right) = \frac{e^{\underline{X}_i^{\Phi\,t}\beta}}{1+e^{\underline{X}_i^{\Phi\,t}\beta}}$$

$$W_{\Phi,\beta}(\mathbb{X}_{(n)}) = \mathrm{Diag}(\Pi_{\Phi,\beta}(\mathbb{X}_{(n)})(1-\Pi_{\Phi,\beta}(\mathbb{X}_{(n)}))$$

Iterative Reweighted Least Squares

- Set an initial β_0
- At each step,
  - Define the adjusted residual

$$\mathbb{Z}_{\Phi,\beta_k}(\mathbb{X}_{(n)}) = \left(\mathbb{X}^{\Phi}_{(n)}\beta_k + W^{-1}_{\Phi,\beta_k}(\mathbb{X}_{(n)})\left(\frac{\mathbb{Y}_{(n)}+1}{2} - \Pi_{\Phi,\beta_k}(\mathbb{X}_{(n)})\right)\right)$$

  - Solve the (re)weighted least squares problem:

$$\beta_{k+1} = \mathrm{argmin}\left\|\mathbb{Z}_{\Phi,\beta_k}(\mathbb{X}_{(n)}) - \mathbb{X}^{\Phi}_{(n)}\beta\right\|^2_{W_{\Phi,\beta_k}(\mathbb{X}_{(n)})}$$

- Not the best algorithm in high dimension...

Proof

- Probability:

$$\Pi_{\Phi,\beta}(\underline{X}_i) = \mathbb{P}_{\Phi,\beta}\left(Y_i = 1|\underline{X}_i\right) = g\left(\underline{X}_i^{\Phi\,t}\beta\right) = \frac{e^{\underline{X}_i^{\Phi\,t}\beta}}{1+e^{\underline{X}_i^{\Phi\,t}\beta}}$$

- Matrix notation:

$$\Pi_{\Phi,\beta}(\mathbb{X}_{(n)}) = \begin{pmatrix}\Pi_{\Phi,\beta}(\underline{X}_1)\\ \vdots \\ \Pi_{\Phi,\beta}(\underline{X}_n)\end{pmatrix}$$

- Log-Likelihood:

$$L_n(\beta) = -n^{-1}\sum_{i=1}^{n}\log\left(1+e^{-Y_i(\underline{X}_i^{\Phi\,t}\beta)}\right)$$

- Gradient:

$$\begin{aligned}
\nabla L_n(\beta) &= n^{-1} \sum_{i=1}^{n} Y_i {\underline{X}_i^{\Phi}}^{t} \frac{e^{-Y_i({\underline{X}_i^{\Phi}}^{t}\beta)}}{1+e^{-Y_i({\underline{X}_i^{\Phi}}^{t}\beta)}} \\
&= n^{-1} \sum_{i=1}^{n} {\underline{X}_i^{\Phi}}^{t} \left(\frac{Y_i+1}{2} - \frac{e^{{\underline{X}_i^{\Phi}}^{t}\beta}}{1+e^{{\underline{X}_i^{\Phi}}^{t}\beta}} \right) \\
&= n^{-1} \sum_{i=1}^{n} {\underline{X}_i^{\Phi}}^{t} \left(\frac{Y_i+1}{2} - \Pi_{\Phi,\beta}(\underline{X}_i) \right) \\
&= n^{-1} {\mathbb{X}_{(n)}}^{\Phi\,t} \left(\frac{\mathbb{Y}_{(n)}+1}{2} - \Pi_{\Phi,\beta}(\mathbb{X}_{(n)}) \right)
\end{aligned}$$

- Hessian:

$$\begin{aligned}
HL_n(\beta) &= -n^{-1} \sum_{i=1}^{n} \underline{X}_i^{\Phi} \underline{X}^{\Phi\,t} \left(\frac{e^{{\underline{X}_i^{\Phi}}^{t}\beta}}{1+e^{{\underline{X}_i^{\Phi}}^{t}\beta}} - \frac{(e^{{\underline{X}_i^{\Phi}}^{t}\beta})^2}{(1+e^{{\underline{X}_i^{\Phi}}^{t}\beta})^2} \right) \\
&= -n^{-1} \sum_{i=1}^{n} \underline{X}_i^{\Phi} \underline{X}^{\Phi\,t} \frac{e^{{\underline{X}_i^{\Phi}}^{t}\beta}}{(1+e^{{\underline{X}_i^{\Phi}}^{t}\beta})^2} \\
&= -n^{-1} \sum_{i=1}^{n} \underline{X}_i^{\Phi} \underline{X}^{\Phi\,t} \Pi_{\Phi,\beta}(\underline{X}_i^{\Phi})(1-\Pi_{\Phi,\beta}(\underline{X}_i^{\Phi})) \\
&= -n^{-1} \mathbb{X}_{(n)}^{\Phi} W_{\Phi,\beta}(\mathbb{X}_{(n)}) {\mathbb{X}_{(n)}}^{\Phi\,t}
\end{aligned}$$

with

$$W_{\Phi,\beta}(\mathbb{X}_{(n)}) = \mathrm{Diag}(\Pi_{\Phi,\beta}(\mathbb{X}_{(n)})(1-\Pi_{\Phi,\beta}(\mathbb{X}_{(n)}))$$

- Local quadratic form approximation around β_0:

$$\begin{aligned}
L(\beta) &\sim L(\beta_0) + \nabla L(\beta_0)(\beta-\beta_0) + \frac{1}{2}(\beta-\beta_0)^t HL(\beta_0)(\beta-\beta_0) \\
&\sim L(\beta_0) + n^{-1} {\mathbb{X}_{(n)}}^{\Phi\,t} \left(\frac{\mathbb{Y}_{(n)}+1}{2} - \Pi_{\Phi,\beta_0}(\mathbb{X}_{(n)}) \right) (\beta-\beta_0) \\
&\quad - \frac{1}{2n}(\beta-\beta_0)^t \mathbb{X}_{(n)}^{\Phi} W_{\Phi,\beta_0}(\mathbb{X}_{(n)}) {\mathbb{X}_{(n)}}^{\Phi\,t} (\beta-\beta_0)
\end{aligned}$$

- Newton step: optimization of the quadratic approximation

$$\begin{aligned}
\beta &= \beta_0 + \left(\mathbb{X}_{(n)}^{\Phi} W_{\Phi,\beta_0}(\mathbb{X}_{(n)}) {\mathbb{X}_{(n)}}^{\Phi\,t} \right)^{-1} {\mathbb{X}_{(n)}}^{\Phi\,t} \left(\frac{\mathbb{Y}_{(n)}+1}{2} - \Pi_{\Phi,\beta_0}(\mathbb{X}_{(n)}) \right) \\
&= \left(\mathbb{X}_{(n)}^{\Phi} W_{\Phi,\beta}(\mathbb{X}_{(n)}) {\mathbb{X}_{(n)}}^{\Phi\,t} \right)^{-1} {\mathbb{X}_{(n)}}^{\Phi\,t} W_{\Phi,\beta_0}(\mathbb{X}_{(n)}) \\
&\quad \underbrace{\left(\mathbb{X}_{(n)}^{\Phi} \beta_0 + W_{\Phi,\beta_0}^{-1}(\mathbb{X}_{(n)}) \left(\frac{\mathbb{Y}_{(n)}+1}{2} - \Pi_{\Phi,\beta_0}(\mathbb{X}_{(n)}) \right) \right)}_{\mathbb{Z}_{\Phi,\beta_0}(\mathbb{X}_{(n)}) = \text{Adjusted Residual}}
\end{aligned}$$

- Newton solution: solution of the weighted least square:

$$\begin{aligned}
\beta &= \operatorname{argmin} \left\| \mathbb{Z}_{\Phi,\beta_0}(\mathbb{X}_{(n)}) - \mathbb{X}_{(n)}^{\Phi}\beta \right\|^2_{W_{\Phi,\beta}(\mathbb{X}_{(n)})} \\
&= \operatorname{argmin} n^{-1} \sum_{i=1}^{n} \Pi_{\Phi,\beta}(\underline{X}_i^{\Phi})(1-\Pi_{\Phi,\beta}(\underline{X}_i^{\Phi})) \\
&\qquad \left((\mathbb{Z}_{\Phi,\beta_0}(\mathbb{X}_{(n)}))_i - {\underline{X}_i^{\Phi}}^{t}\beta \right)^2
\end{aligned}$$

3.1.3 Logistic Regression, Confidence Zones and Tests

Asymptotic

- What happens when n goes to $+\infty$?
- Assumptions:
  - True model assumption: $\exists \beta^\star$ such that $\mathbb{P}\left(Y=1|\underline{X}\right) = g(\underline{X}^{\Phi t}\beta^\star)$
  - Design assumption: random design such that $\mathbb{E}\left[\underline{X}^{\Phi}\underline{X}^{\Phi t}\right]$ s.d.p.

Asymptotic Results

- $\hat{\beta}$ converges toward $\beta^\star$ in probability.
- CLT: if $\mathbb{M}$ is a $q \times p$ matrix of rank q

$$\sqrt{n}\left(\mathbb{M}\hat{\beta} - \mathbb{M}\beta^\star\right) \xrightarrow{P} \mathrm{N}\left(0, \mathbb{M}\left({\mathbb{X}_{(n)}}^{\Phi t} W_{\Phi,\beta^\star}(\mathbb{X}_{(n)})\mathbb{X}^{\Phi}_{(n)}\right)^{-1}\mathbb{M}^t\right)$$

- Norm version: if $\mathbb{M}$ is a $q \times p$ matrix of rank q

$$n\left\|\left(\mathbb{M}\left({\mathbb{X}_{(n)}}^{\Phi t} W_{\Phi,\hat{\beta}}(\mathbb{X}_{(n)})\mathbb{X}^{\Phi}_{(n)}\right)^{-1}\mathbb{M}^t\right)^{-1/2}\left(\mathbb{M}\hat{\beta} - \mathbb{M}\beta^\star\right)\right\|^2 \xrightarrow{P} \chi^2(q)$$

- Likelihood:

$$2\left(\log \mathcal{L}_n(\hat{\beta}) - \log \mathcal{L}_n(\beta^\star)\right) \sim \chi^2(p)$$

- Proof left to the reader...

Asymptotic Confidence Zones

- Based on the asymptotic results...
- Let $\Sigma_{\Phi,\hat{\beta}} = \left({\mathbb{X}_{(n)}}^{\Phi t} W_{\Phi,\hat{\beta}}(\mathbb{X}_{(n)})\mathbb{X}^{\Phi}_{(n)}\right)^{-1}$

Gaussian Asymptotic Confidence Zone (Wald)

- Asymptotically if $\mathbb{M}$ is a $q \times p$ matrix of rank q

$$Z_\alpha = \left\{\mathbb{M}\beta, n\left\|\left(\mathbb{M}\Sigma_{\Phi,\hat{\beta}}\mathbb{M}^t\right)^{-1/2}\left(\mathbb{M}\hat{\beta} - \mathbb{M}\beta\right)\right\|^2 \leq {\chi^2}_{1-\alpha}(q)\right\}$$

  is a confidence zone of level α for $\mathbb{M}\beta^\star$.

Likelihood Asymptotic Confidence Zone

- Likelihood based confidence zones: Asymptotically

$$Z_\alpha = \left\{\beta, 2\left(\log \mathcal{L}_n(\hat{\beta}) - \log \mathcal{L}_n(\beta)\right) \leq {\chi^2}_{1-\alpha}(p)\right\}$$

  is a confidence zone of level α for $\beta^\star$.

- Different geometry between the two type of zones when $\mathbb{M} = \mathrm{Id}_{(p)}$.

Wald vs Likelihood Ratio Test

Two different systematic construction

- *Wald:*
  - Null hypothesis of type $H_0 = \{\mathbb{M}\beta^\star = c\}$
  - Compute an estimate $\hat{\beta}$.
  - Test if the difference $\mathbb{M}\hat{\beta} - c$ is large enough to reject H_0
- *Likelihood Ratio Test:*
  - Null hypothesis of type $H_0 = \{\beta^\star \in \Theta_0\}$ vs an alternative of type $H_1 = \{\beta^\star \in \Theta_1\}$.
  - Compute the maximum likelihood estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ in the two models.
  - Test whether the deviance $2\left(\log \mathcal{L}\left(\hat{\beta}_0\right) - \log \mathcal{L}\left(\hat{\beta}_1\right)\right)$ is large enough to reject H_0
- *Key:* control of the difference or the deviance under H_0!
- *Rk:* Under mild assumptions,
  - $\hat{\beta}$ is asymptotically Gaussian and thus the difference.
  - The deviance follows asymptotically a χ^2 of degree $\dim \Theta_1 - \dim \Theta_0$ if $\Theta_0 \subset \Theta_1$

Asymptotic Tests

Gaussian Asymptotic Test (Wald)

- Test of $\mathbb{M}\beta^\star = c$ asymptotically of level α

$$T_\alpha = \begin{cases} 0 & \text{if} n \left\| \left(\mathbb{M}\Sigma_{\Phi,\hat{\beta}}\mathbb{M}^t\right)^{-1/2} \left(\mathbb{M}\hat{\beta} - c\right) \right\|^2 \leq {\chi^2}_{1-\alpha}(q) \\ 1 & \text{otherwise} \end{cases}$$

- Dual p-value approach:

$$p = \mathbb{P}\left(\chi^2(q) > n \left\| \left(\mathbb{M}\Sigma_{\Phi,\hat{\beta}}\mathbb{M}^t\right)^{-1/2} \left(\mathbb{M}\hat{\beta} - c\right) \right|^2\right)$$

Likelihood Asymptotic Test

- Asymptotic test of $\beta^\star \in \Theta_0$ vs $\beta^\star \in \Theta_1$ of level α

$$T_\alpha = \begin{cases} 0 & \text{if } 2\left(\log \mathcal{L}_n(\hat{\beta}_1) - \log \mathcal{L}_n(\hat{\beta}_0)\right) \leq {\chi^2}_{1-\alpha}(\dim \Theta_1 - \dim \Theta_0) \\ 1 & \text{otherwise} \end{cases}$$

- Dual p-value approach:

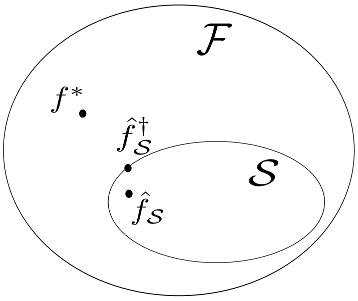
$$p = \mathbb{P}\left(\chi^2(\dim \Theta_1 - \dim \Theta_0) > 2\left(\log \mathcal{L}_n(\hat{\beta}_1) - \log \mathcal{L}_n(\hat{\beta}_0)\right)\right)$$

3.1.4 Logistic Regression and Models

Bias-Variance Dilemna

Non parametric setting

- $\mathcal{F} = \{\text{measurable fonctions } \mathcal{X} \to \mathcal{Y}\}$
- Best solution: $f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$
- Class $\mathcal{S} = \{f_\beta, \beta \in \mathbb{R}^p\} \subset \mathcal{F}$ of functions
- Ideal target in $\mathcal{S}$: $f^\dagger_\mathcal{S} = \operatorname{argmin}_{f_\beta, \beta \in \mathbb{R}^p} \mathcal{R}(f_\beta)$
- Estimate in $\mathcal{S}$: $\widehat{f}_\beta$ obtained by the logistic regression...



Approximation error and estimation error (Bias/Variance)

$$\mathcal{R}(\widehat{f}_\mathcal{S}) - \mathcal{R}(f^*) = \underbrace{\mathcal{R}(f^*_\mathcal{S}) - \mathcal{R}(f^*)}_{\text{Approximation error}} + \underbrace{\mathcal{R}(\widehat{f}_\mathcal{S}) - \mathcal{R}(f^*_\mathcal{S})}_{\text{Estimation error}}$$

- Approx. error can be large if the model is not suitable or p is small.
- Estimation error can be large if p is large.
- *Rule of thumb*: at least $p/n < 1/20\ldots$

Simplified Logistic Models

Logistic Coefficients

- Logistic regression entirely specified by β.
- Coefficientwise:
  - $\beta_i = 0$ means that the ith covariate is not used.
  - $\beta_i \sim 0$ means that the ith covariate as a low influence...

Simplified Logistic Models

- Enforce simplicity through a constraint on β!
- Support constraint: $\|\beta\|_0 = \sum_{i=1}^{d} \mathbf{1}_{\beta_i \neq 0} < C$
- Size constraint: $\|\beta\|_p < C$ with $1 \leq p$ (Often $p = 2$ or $p = 1$)

- *Rk:* $\|\beta\|_p$ is not scaling invariant if $p \neq 0$...
- Initial rescaling issue.

Penalization

Penalized Likelihood

- Minimization of

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^d} n^{-1} \sum_{i=1}^{n} \log(1 + e^{-Y_i(\underline{X}_i{}^t \beta)}) + \operatorname{pen}(\beta)$$

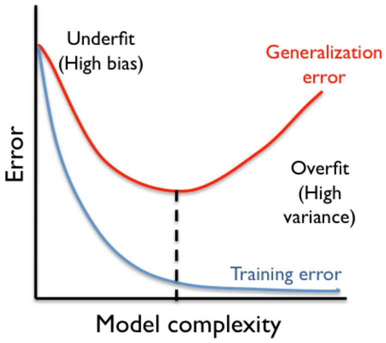
  where $\operatorname{pen}(\beta)$ is a (sparsity promoting) penalty
- Variable selection if β is sparse.

Classical Penalties

- AIC: $\operatorname{pen}(\beta) = \lambda \|\beta\|_0$ (non convex / sparsity)
- Ridge: $\operatorname{pen}(\beta) = \lambda \|\beta\|_2^2$ (convex / no sparsity)
- Lasso: $\operatorname{pen}(\beta) = \lambda \|\beta\|_1$ (convex / sparsity)
- Elastic net: $\operatorname{pen}(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$ (convex / sparsity)

- Easy optimization if pen (and the loss) is convex...
- *Need to specify λ!*

Regularization Parameter Issue



- Need to choose λ from the data!

Error behaviour

- Learning/training error (error made on the learning/training set) decays when the regularization parameter decreases.

- Quite different behavior when the error is computed on new observations (generalization error).

- *Overfit* for complex models: parameters learned are too specific to the learning set!
- General situation! (Think of linear regression...)
- Need another criterion than the training error!

Cross Validation and Penalization

Two Approaches

- **Cross validation:** Very efficient (and almost always used in practice!) but slightly biased as it target uses only a fraction of the data.
- **Penalization approach:** use empirical loss criterion but penalize it by a term increasing with the complexity of $\mathcal{S}$

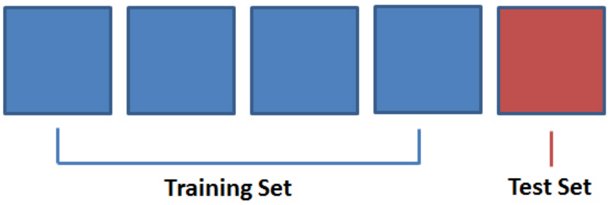
$$R_n(\widehat{f_{\mathcal{S}}}) \to R_n(\widehat{f_{\mathcal{S}}}) + \mathrm{pen}(\mathcal{S})$$

and choose the model with the smallest penalized risk.

Which loss to use?

- The loss used in the risk: most natural!
- The loss used to estimate $\widehat{\beta}$: penalized estimation!

Cross Validation



- *Very simple idea:* use a second learning/verification set to compute a verification error.
- Sufficient to avoid over-fitting!

Cross Validation

- Use $\frac{V-1}{V}n$ observations to train and $\frac{1}{V}n$ to verify!
- Validation for a learning set of size $(1-\frac{1}{V}) \times n$ instead of n!

- Most classical variations:
  - Leave One Out,
  - V-fold cross validation.
- Accuracy/Speed tradeoff: $V = 5$ or $V = 10$!

Penalization

Penalization as a risk correction

- The empirical loss computed on an estimator selected in a family according to the data is biased!
- Optimistic estimation of the risk...
- Estimate an upper bound of this optimism for a given family, called the penalty.
- Add it to the empirical loss

- One can also think of the penalty as a way to force the use of *simple* models...
- **Rk:** Interpretability with both the statistical and the optimization point of view.

Penalization

Penalized Loss

- Minimization of

$$\operatorname*{argmin}_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\beta(\underline{X}_i)) + \text{pen}(\beta)$$

  where $\text{pen}(\theta)$ is a penalty.

Penalties

- Upper bound of the optimism of the empirical loss
- Depends on the loss and the framework!

Instantiation

- AIC / BIC:
- Structural Risk Minimization...

Penalization and Cross-Validation

Practical Selection Methodology

- Choose a penalty shape $\widetilde{\text{pen}}(\beta)$.
- Compute a CV error for a penalty $\lambda\widetilde{\text{pen}}(\beta)$ for all $\lambda \in \Lambda$.
- Determine $\widehat{\lambda}$ the λ minimizing the CV error.
- Compute the final logistic regression with a penalty $\widehat{\lambda}\widetilde{\text{pen}}(\beta)$.

Why not using only CV?

- *If* the penalized likelihood minimization is easy, much cheaper to compute the CV error for all $\lambda \in \Lambda$ than for all $\beta \in \mathbb{R}^d$!
- CV performs best when the set of candidates is not too big (or is structured...)

3.1.5 Generalized Linear Model

Generalized Linear Model

How to generalize the classical linear model?

- Linear model:

$$\mathbb{E}\left[Y|\underline{X}\right] = \underline{X}^{\Phi^t}\beta$$

- Generalized linear model:

$$g(\mathbb{E}\left[Y|\underline{X}\right]) = \underline{X}^{\Phi^t}\beta$$

  with g a link function.

Family of law P_θ such that θ is characterized by the mean

- $P_\theta = \mathrm{N}(\theta, \sigma_\star^2)$ and $g(x) = x$
- $P_\theta = \mathcal{B}(\theta)$ and $g(x) = \ln \frac{x}{1-x}$ (logit)
- $P_\theta = \mathcal{P}(\theta)$ and $g(x) = \ln(x)$

Exponential family

Exponential family

- Probability law family P_θ such that the density can be written

$$f(y,\theta,\varphi) = e^{\frac{y\theta - v(\theta)}{\varphi} + w(y,\varphi)}$$

  where φ is a nuisance parameter and w a function independent of θ.

- Examples:
  - Gaussian: $f(y,\theta,\varphi) = e^{\frac{y\theta - \theta^2/2}{\varphi} + \frac{y^2/2}{\varphi}}$
  - Bernoulli: $f(y,\theta) = e^{z\theta - \ln(1+e^\theta)}$ $(\theta = \ln p/(1-p))$
  - Poisson: $f(y,\theta) = e^{(y\theta - e^\theta) + \ln(y!)}$ $(\theta = \ln\lambda)$

Properties

- $\mathbb{E}_\theta\left[Y\right] = v'(\theta)$ and $\mathbb{V}\mathrm{ar}_\theta\left[Y\right] = \varphi v''(\theta)$
- The maximum likelihood estimate satisfies $v'(\widehat{\theta}) = n^{-1}\sum_{i=1}^n Y_i$

Canonical parameterization

- $v''(\theta) = \frac{1}{\phi}\mathbb{V}\mathrm{ar}_\theta\left[Y\right] > 0$ thus $v'(\theta) = \mathbb{E}_\theta\left[Y\right]$ yields a bijection between θ and $\mathbb{E}_\theta\left[Y\right]$.
- Examples:
  - Gaussian:

$$\mathbb{E}_\theta\left[Y\right] = \theta \Leftrightarrow \theta = \mathbb{E}_\theta\left[Y\right]$$

  - Bernoulli:

$$\mathbb{E}_\theta\left[Y\right] = \frac{e^\theta}{1-e^\theta} \Leftrightarrow \theta = \ln \frac{\mathbb{E}_\theta\left[Y\right]}{1-\mathbb{E}_\theta\left[Y\right]}$$

  - Poisson:

$$\mathbb{E}_\theta\left[Y\right] = e^\theta \Leftrightarrow \theta = \ln \mathbb{E}_\theta\left[Y\right]$$

- φ: nuisance parameter more complex to estimate.

Generalized Linear Model

Ingredients

- Exponential family P_θ
- Link function g such that

$$g(\mathbb{E}_{\theta(\underline{X})}\left[Y|\underline{X}\right]) = {\underline{X}^{\Phi}}^t\beta$$

- Beware: g is not necessary equal to v' the canonical link function!
- Property: $g^{-1}({\underline{X}^{\Phi}}^t\beta) = \mathbb{E}_{\theta(\underline{X})}\left[Y|\underline{X}\right] = v'(\theta(\underline{X}^{\Phi}))$

Examples

- Gaussian and $g(x) = x$ ($g^{-1}(x) = x$) [Gaussian linear model]
- Bernoulli and
  - $g(x) = \ln x/(1-x)$ ($g^{-1}(x) = e^x/(1-e^x)$) [logit model]
  - $g(x) = \Phi^{-1}(x)$ ($g^{-1}(x) = \Phi(x)$) [probit model]
  - $g(x) = \ln(-\ln(1-x))$ ($g^{-1}(x) = 1 - e^{-e^x}$) [log-log model]
- Poisson and $g(x) = \ln x$ ($g^{-1} = e^x$) [Poisson model with logarithmic link]

Likelihood and Optimization

- Def: $\theta({\underline{X}^{\Phi}}^t\beta) = v'^{-1}\left(g^{-1}({\underline{X}^{\Phi}}^t\beta)\right)$

Opposite of the log-likelihood

- Expression:

$$\log \mathcal{L}_n(\beta) = \frac{1}{\varphi}\sum_{i=1}^n \left(Y_i\theta({\underline{X}_i^{\Phi}}^t\beta) - v(\theta({\underline{X}_i^{\Phi}}^t\beta))\right) + \sum_{i=1}^n \frac{w(Y_i,\varphi)}{\varphi}$$

- Derivative:

$$\begin{aligned}\frac{\partial \log \mathcal{L}_n}{d\beta_j}(\beta) &= \frac{1}{\varphi}\sum_{i=1}^n (Y_i - v'(\theta({\underline{X}_i^{\Phi}}^t\beta))\frac{\partial\theta}{d\beta_j}\\ &= \frac{1}{\varphi}\sum_{i=1}^n \frac{Y_i - g^{-1}({\underline{X}_i^{\Phi}}^t\beta)}{V(g^{-1}({\underline{X}_i^{\Phi}}^t\beta))}(g^{-1})'({\underline{X}_i^{\Phi}}^t\beta)\underline{X}_{i,j}^{\Phi}\end{aligned}$$

  with $V(g^{-1}({\underline{X}_i^{\Phi}}^t\beta)) = v''(\theta({\underline{X}_i^{\Phi}}^t\beta))$.

- Numerical scheme possible...