# Statistics and learning

## Regression

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Friday 25th January 2013

# The regression model

- expresses a random variable $Y$ as a function of random variables $X \ in\mathbb{R}^p$ according to:

$$Y = f(X; \beta) + \epsilon,$$

where functional $f$ depends on **unknown parameters** $\beta_1, \ldots, \beta_k$ and the **residual** (or **error**) $\epsilon$ is an unobservable rv which accounts for random fluctuations between the model and $Y$.

# The regression model

- expresses a random variable $Y$ as a function of random variables $X \ in\mathbb{R}^p$ according to:

$$Y = f(X; \beta) + \epsilon,$$

where functional $f$ depends on **unknown parameters** $\beta_1, \dots, \beta_k$ and the **residual** (or **error**) $\epsilon$ is an unobservable rv which accounts for random fluctuations between the model and $Y$.
- Goal: from $n$ experimental observations $(x_i, y_i)$, we aim at

# The regression model

- expresses a random variable $Y$ as a function of random variables $X \ in \mathbb{R}^p$ according to:

$$Y = f(X; \beta) + \epsilon,$$

  where functional $f$ depends on **unknown parameters** $\beta_1, \dots, \beta_k$ and the **residual** (or **error**) $\epsilon$ is an unobservable rv which accounts for random fluctuations between the model and $Y$.
- Goal: from $n$ experimental observations $(x_i, y_i)$, we aim at
  - **estimate** unknown $(\beta_l)_{l=1 \dots k}$,

# The regression model

- expresses a random variable $Y$ as a function of random variables $X\ in\mathbb{R}^p$ according to:

$$Y = f(X;\beta) + \epsilon,$$

  where functional $f$ depends on **unknown parameters** $\beta_1, \dots, \beta_k$ and the **residual** (or **error**) $\epsilon$ is an unobservable rv which accounts for random fluctuations between the model and $Y$.
- Goal: from $n$ experimental observations $(x_i, y_i)$, we aim at
  - **estimate** unknown $(\beta_l)_{l=1\dots k}$,
  - evaluate the **fit** of the model

# The regression model

- expresses a random variable $Y$ as a function of random variables $X \ in \mathbb{R}^p$ according to:

$$Y = f(X; \beta) + \epsilon,$$

  where functional $f$ depends on **unknown parameters** $\beta_1, \ldots, \beta_k$ and the **residual** (or **error**) $\epsilon$ is an unobservable rv which accounts for random fluctuations between the model and $Y$.
- Goal: from $n$ experimental observations $(x_i, y_i)$, we aim at
  - **estimate** unknown $(\beta_l)_{l=1 \ldots k}$,
  - evaluate the **fit** of the model
  - if the fit is acceptable, tests on parameters can be performed and the model can be used for **predictions**

# Simple linear regression

- A single **explanatory variable** $X$ and an affine relationship to the **dependant variable** $Y$:

$$E[Y \mid X = x] = \beta_0 + \beta_1 x \text{ or } Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where $\beta_1$ is the slope of the adjusted regression line and $\beta_0$ is the intercept.

# Simple linear regression

- A single **explanatory variable** $X$ and an affine relationship to the **dependant variable** $Y$:

$$E[Y \mid X = x] = \beta_0 + \beta_1 x \text{ or } Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

  where $\beta_1$ is the slope of the adjusted regression line and $\beta_0$ is the intercept.

- Residuals $\epsilon_i$ are assumed to be centred (R1), have equal variances ($= \sigma^2$, R2) and be uncorrelated: $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \quad \forall i \neq j$ (R3).

# Simple linear regression

- A single **explanatory variable** $X$ and an affine relationship to the **dependant variable** $Y$:

$$E[Y \mid X = x] = \beta_0 + \beta_1 x \text{ or } Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

  where $\beta_1$ is the slope of the adjusted regression line and $\beta_0$ is the intercept.
- Residuals $\epsilon_i$ are assumed to be centred (R1), have equal variances ($= \sigma^2$, R2) and be uncorrelated: $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \quad \forall i \neq j$ (R3).
- Hence: $E[Y_i] = \beta_0 + \beta_1 x_i$, $\mathrm{Var}(Y_i) = \sigma^2$ and $\mathrm{Cov}(Y_i, Y_j) = 0, \quad \forall i \neq j$.

# Simple linear regression

- A single **explanatory variable** $X$ and an affine relationship to the **dependant variable** $Y$:

$$E[Y \mid X = x] = \beta_0 + \beta_1 x \text{ or } Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

  where $\beta_1$ is the slope of the adjusted regression line and $\beta_0$ is the intercept.
- Residuals $\epsilon_i$ are assumed to be centred (R1), have equal variances ($= \sigma^2$, R2) and be uncorrelated: $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \quad \forall i \neq j$ (R3).
- Hence: $E[Y_i] = \beta_0 + \beta_1 x_i$, $\mathrm{Var}(Y_i) = \sigma^2$ and $\mathrm{Cov}(Y_i, Y_j) = 0, \quad \forall i \neq j$.
- Fitting (or adjusting) the model = estimate $\beta_0$, $\beta_1$ and $\sigma$ from the $n$-sample $(x_i, y_i)$.

# Least square estimate

- Seeking values for $\beta_0$ and $\beta_1$ minimising the sum of quadratic errors:

$$(\hat{\beta_0}, \hat{\beta_1}) = \text{argmin}_{(\beta_0}, \beta_1) \hat{\in} \mathbb{R}^2 \sum [y_i - (\beta_0 + \beta_1 x_i)]^2$$

# Least square estimate

- Seeking values for $\beta_0$ and $\beta_1$ minimising the sum of quadratic errors:

$$(\hat{\beta_0}, \hat{\beta_1}) = \text{argmin}_{(\beta_0}, \beta_1) \hat{\in} \mathbb{R}^2 \sum [y_i - (\beta_0 + \beta_1 x_i)]^2$$

# Least square estimate

- Seeking values for $\beta_0$ and $\beta_1$ minimising the sum of quadratic errors:

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin}_{(\beta_0}, \beta_1) \hat{\in} \mathbb{R}^2 \sum [y_i - (\beta_0 + \beta_1 x_i)]^2$$

# Least square estimate

- Seeking values for $\beta_0$ and $\beta_1$ minimising the sum of quadratic errors:

$$(\hat{\beta_0}, \hat{\beta_1}) = \text{argmin}_{(\beta_0}, \beta_1) \hat{\in} \mathbb{R}^2 \sum [y_i - (\beta_0 + \beta_1 x_i)]^2$$

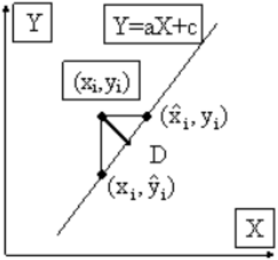
Note that $Y$ and $X$ do not play a symetric role !

## Least square estimate

- Seeking values for $\beta_0$ and $\beta_1$ minimising the sum of quadratic errors:

$$(\hat{\beta_0}, \hat{\beta_1}) = \mathrm{argmin}_{(\beta_0}, \beta_1) \hat{\in} \mathbb{R}^2 \sum [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Note that $Y$ and $X$ do not play a symetric role !

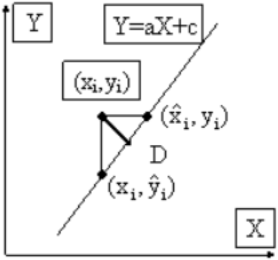


# Least square estimate

- Seeking values for $\beta_0$ and $\beta_1$ minimising the sum of quadratic errors:

$$(\hat{\beta_0}, \hat{\beta_1}) = \text{argmin}_{(\beta_0}, \beta_1) \hat{\in} \mathbb{R}^2 \sum [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Note that $Y$ and $X$ do not play a symetric role !




- In matrix notation (useful later): $Y = X.B + \epsilon$, with $Y = {}^\top(Y_1 \ldots Y_n)$, $\beta = {}^\top(\beta_1, \beta_2)$, $\epsilon = {}^\top(\epsilon_1 \ldots \epsilon_n)$ and $X = {}^\top \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{pmatrix}$.

# Estimator properties

- useful notations: $\bar{x} = 1/n \sum_i x_i$, $\bar{y}$, $s_x^2$, $s_y^2$ and $s_{xy} = 1/(n-1) \sum_i (x_i - \bar{x})(y_i - \bar{y})$.

# Estimator properties

- useful notations: $\bar{x} = 1/n \sum_i x_i$, $\bar{y}$, $s_x^2$, $s_y^2$ and $s_{xy} = 1/(n-1) \sum_i (x_i - \bar{x})(y_i - \bar{y})$.
- Linear correlation coefficient: $r_{xy} = \frac{s_{xy}}{s_x s_y}$.

# Estimator properties

- useful notations: $\bar{x} = 1/n \sum_i x_i$, $\bar{y}$, $s_x^2$, $s_y^2$ and $s_{xy} = 1/(n-1) \sum_i (x_i - \bar{x})(y_i - \bar{y})$.
- Linear correlation coefficient: $r_{xy} = \frac{s_{xy}}{s_x s_y}$.

## Theorem

1. *Least Square estimators:*$= \hat{\beta}_1 = s_{xy}/s_x^2$ *and* $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$*.*
2. *These estimators are unbiased and efficient.*
3. $s^2 = \frac{1}{n-2} \sum_i \left[ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$ *is an unbiased estimator of* $\sigma^2$*. It is however not efficient.*
4. $\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{(n-1)s_x^2}$ *and* $\mathrm{Var}(\hat{\beta}_1) = \bar{x}^2 \mathrm{Var}(\hat{\beta}_1) + \sigma^2/n$

# Simple Gaussian linear model

- In addition to R1 (centred noise), R2 (equal variance noise) and R3 (uncorrelated noise), we assume (R3') $\forall i \neq j$, $\epsilon_i$ and $\epsilon_j$ independent and (R4) $\forall i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ or equivalently $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$.

# Simple Gaussian linear model

- In addition to R1 (centred noise), R2 (equal variance noise) and R3 (uncorrelated noise), we assume (R3') $\forall i \neq j$, $\epsilon_i$ and $\epsilon_j$ independent and (R4) $\forall i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ or equivalently $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$.
- **Theorem**: under (R1, R2, R3' and R4), Least Square estimators $=$ MLE.

# Simple Gaussian linear model

- In addition to R1 (centred noise), R2 (equal variance noise) and R3 (uncorrelated noise), we assume (R3') $\forall i \neq j$, $\epsilon_i$ and $\epsilon_j$ independent and (R4) $\forall i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ or equivalently $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$.
- **Theorem**: under (R1, R2, R3' and R4), Least Square estimators = MLE.

## Theorem (Distribution of estimators)

1. $\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \sigma^2_{\hat{\beta}_0})$ *and* $\hat{\beta}_1 \sim \mathcal{N}(\beta_0, \sigma^2_{\hat{\beta}_1})$*, with* $\sigma^2_{\hat{\beta}_0} = \sigma^2 \left(\bar{x}^2 / \sum_i (x_i - \bar{x})^2 + 1/n\right)$ *and* $\sigma^2_{\hat{\beta}_1} = \sigma^2 / \sum_i (x_i - \bar{x})^2$
2. $(n-2)s^2/\sigma^2 \sim \chi^2_{n-2}$
3. $\hat{\beta}_0$ *and* $\hat{\beta}_1$ *are independent of* $\hat{\epsilon}_i$*.*
4. *Estimators of* $\sigma^2_{\hat{\beta}_0}$ *and* $\sigma^2_{\hat{\beta}_1}$ *are given in 1. by replacing* $\sigma^2$ *by* $s^2$*.*

# Tests, ANOVA and determination coefficient

- Previous theorem allows us to build CI for $\beta_0$ and $\beta_1$.

# Tests, ANOVA and determination coefficient

- Previous theorem allows us to build CI for $\beta_0$ and $\beta_1$.
- $SST/n = SSR/n + SSE/n$, with $SST = \sum_i (y_i - \bar{y})^2$ (total sum of squares), $SSR = \sum_i (\hat{y}_i - \bar{y})^2$ (regression sum of squares) and $SSE = \sum_i (y_i - \bar{y_i})^2$ (sum of squared errors).

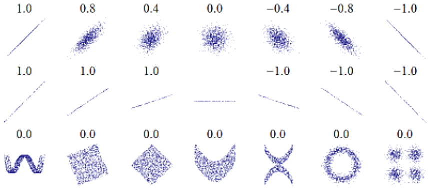
# Tests, ANOVA and determination coefficient

- Previous theorem allows us to build CI for $\beta_0$ and $\beta_1$.
- $SST/n = SSR/n + SSE/n$, with $SST = \sum_i (y_i - \bar{y})^2$ (total sum of squares), $SSR = \sum_i (\hat{y_i} - \bar{y})^2$ (regression sum of squares) and $SSE = \sum_i (y_i - \bar{y_i})^2$ (sum of squared errors).
- **Definition**: Determination coefficient $R^2 = \frac{\sum_i (\hat{y_i} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\text{Residual Variance}}{\text{Total variance}}$.

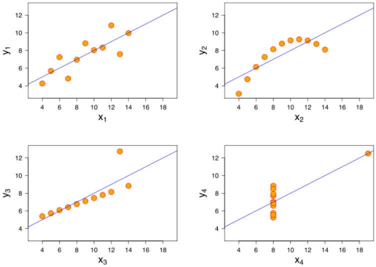
# Tests, ANOVA and determination coefficient

- Previous theorem allows us to build CI for $\beta_0$ and $\beta_1$.
- $SST/n = SSR/n + SSE/n$, with $SST = \sum_i (y_i - \bar{y})^2$ (total sum of squares), $SSR = \sum_i (\hat{y}_i - \bar{y})^2$ (regression sum of squares) and $SSE = \sum_i (y_i - \bar{y_i})^2$ (sum of squared errors).
- **Definition**: Determination coefficient
$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\text{Residual Variance}}{\text{Total variance}}$.

$\rightarrow$ Always use scatterplots to interpret linear model adequacy




same $R^2 = 0.667$



# Prediction

- Given a new $x^*$, what is the prediction $\tilde{y}$ ?

# Prediction

- Given a new $x^*$, what is the prediction $\tilde{y}$ ?
- It's simply $\widehat{y(x^*)} = \hat{\beta}_0 + \hat{\beta}_1 x^*$. But what is its precision ?

# Prediction

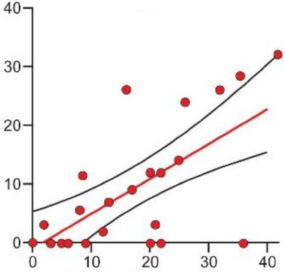
- Given a new $x^*$, what is the prediction $\tilde{y}$ ?
- It's simply $\widehat{y(x^*)} = \hat{\beta}_0 + \hat{\beta}_1 x^*$. But what is its precision ?
- Its CI is $\left[\hat{\beta}_0 + \hat{\beta}_1 x^* + / - t_{n-2;1-\alpha/2} s^*\right]$, where $s^* = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$.

# Prediction

- Given a new $x^*$, what is the prediction $\tilde{y}$ ?
- It's simply $\widehat{y(x^*)} = \hat{\beta}_0 + \hat{\beta}_1 x^*$. But what is its precision ?
- Its CI is $\left[\hat{\beta}_0 + \hat{\beta}_1 x^* + / - t_{n-2;1-\alpha/2} s^*\right]$, where $s^* = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$.
- Predictions are valid in the range of $(x_i)$'s.

# Prediction

- Given a new $x^*$, what is the prediction $\tilde{y}$ ?
- It's simply $\widehat{y(x^*)} = \hat{\beta}_0 + \hat{\beta}_1 x^*$. But what is its precision ?
- Its CI is $\left[\hat{\beta}_0 + \hat{\beta}_1 x^* + / - t_{n-2;1-\alpha/2} s^*\right]$, where $s^* = s\sqrt{1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$.
- Predictions are valid in the range of $(x_i)$'s.
- The precision varies according to the $x^*$ value you want to predict:



# Multiple linear regression

- Natural extension when several $(X_j)_{j=1\dots p}$ are used to explain $Y$.

# Multiple linear regression

- Natural extension when several $(X_j)_{j=1\dots p}$ are used to explain $Y$.
- Model simply writes: $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$. In matrix notations with obvious generalisation: $Y = X\beta + \epsilon$.

# Multiple linear regression

- Natural extension when several $(X_j)_{j=1\dots p}$ are used to explain $Y$.
- Model simply writes: $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$. In matrix notations with obvious generalisation: $Y = X\beta + \epsilon$.
- $x = (x_i^j)_{i,j}$ is the observed **design matrix**.

# Multiple linear regression

- Natural extension when several $(X_j)_{j=1\dots p}$ are used to explain $Y$.
- Model simply writes: $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$. In matrix notations with obvious generalisation: $Y = X\beta + \epsilon$.
- $x = (x_i^j)_{i,j}$ is the observed **design matrix**.
- Identifiability of $\beta$ is equivalent to the linear independence of the columns of $x$ *i.e.* $\mathrm{Rank}(X) = p + 1$. This is equivalent to ${}^\top XX$ being invertible.

# Multiple linear regression

- Natural extension when several $(X_j)_{j=1\dots p}$ are used to explain $Y$.
- Model simply writes: $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$. In matrix notations with obvious generalisation: $Y = X\beta + \epsilon$.
- $x = (x_i^j)_{i,j}$ is the observed **design matrix**.
- Identifiability of $\beta$ is equivalent to the linear independence of the columns of $x$ *i.e.* $\mathrm{Rank}(X) = p + 1$. This is equivalent to ${}^\top XX$ being invertible.
- Parameter estimation: $\mathrm{argmin}_\beta \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_i^j - \beta_0 \right)^2 \Leftrightarrow \mathrm{argmin}_\beta \sum_i \hat{\epsilon_i}^2 \Leftrightarrow \mathrm{argmin}_\beta \|Y - X\beta\|_2^2$.

# Multiple linear regression

- Natural extension when several $(X_j)_{j=1\dots p}$ are used to explain $Y$.
- Model simply writes: $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$. In matrix notations with obvious generalisation: $Y = X\beta + \epsilon$.
- $x = (x_i^j)_{i,j}$ is the observed **design matrix**.
- Identifiability of $\beta$ is equivalent to the linear independence of the columns of $x$ *i.e.* $\mathrm{Rank}(X) = p + 1$. This is equivalent to ${}^\top X X$ being invertible.
- Parameter estimation: $\mathrm{argmin}_\beta \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_i^j - \beta_0 \right)^2 \Leftrightarrow \mathrm{argmin}_\beta \sum_i \hat{\epsilon_i}^2 \Leftrightarrow \mathrm{argmin}_\beta \|Y - X\beta\|_2^2$.
- **Theorem** The Least Square Estimator of $\beta$ is $\hat{\beta} = ({}^\top X X)^{-1}\, {}^\top X\, Y$.

# Properties of the least square estimate

**Theorem**

*The estimator $\hat{\beta}$ previously defined is s.t.*

1. $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2({}^\top X X)^{-1})$ *and*
2. $\hat{\beta}$ *efficient: among all unbiased estimator, it has the smallest variance.*

# Properties of the least square estimate

**Theorem**

*The estimator $\hat{\beta}$ previously defined is s.t.*

1. $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2({}^\top X X)^{-1})$ *and*
2. $\hat{\beta}$ *efficient: among all unbiased estimator, it has the smallest variance.*

- few control on $\sigma^2$. So the structure of ${}^\top X\, X$ dictates the quality of estimator $\hat{\beta}$: optimal experimental design subject.

# Properties of the least square estimate

**Theorem**

*The estimator $\hat{\beta}$ previously defined is s.t.*

1. $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2({}^\top X X)^{-1})$ *and*
2. $\hat{\beta}$ *efficient: among all unbiased estimator, it has the smallest variance.*

- few control on $\sigma^2$. So the structure of ${}^\top X\, X$ dictates the quality of estimator $\hat{\beta}$: optimal experimental design subject.

**Theorem**

$\hat{Y} = X\hat{\beta}$*: predicted values. Then* $\hat{Y} = H\,Y$*, with* $H = X\,({}^\top X\,X)^{-1}\,{}^\top X$*;* $\epsilon = Y - \hat{Y} = (Id - H)\,Y$*. Note that $H$ is the orthogonal projection on* $\mathrm{Vect}(X) \subset \mathbb{R}^n$*. We have:*

1. $\mathrm{Cov}(\hat{Y}) = \sigma^2 H$,
2. $\mathrm{Cov}(\epsilon) = \sigma^2(Id - H)$ *and*
3. $\hat{\sigma^2} = \frac{\|\epsilon^2\|}{n-p-1}$.

# Practical uses

- CI for $\beta_j$: $[\hat{\beta}_j + / - t_{n-p-1;1-\alpha/2}\sigma_{\hat{\beta}_j}]$, with $t_{n-p-1;1-\alpha/2}$ a Student-quantile and $\sigma_{\hat{\beta}_j}$ the squareroot of the $j^{\text{th}}$ element of $\mathrm{Cov}(\hat{\beta})$.

# Practical uses

- CI for $\beta_j$: $[\hat{\beta}_j + / - t_{n-p-1;1-\alpha/2}\sigma_{\hat{\beta}_j}]$, with $t_{n-p-1;1-\alpha/2}$ a Student-quantile and $\sigma_{\hat{\beta}_j}$ the squareroot of the $j^{\text{th}}$ element of $\mathrm{Cov}(\hat{\beta})$.
- Tests on $\beta_j$: the rv $\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}}$ has a Student distribution.

# Practical uses

- CI for $\beta_j$: $[\hat{\beta}_j + / - t_{n-p-1;1-\alpha/2}\sigma_{\hat{\beta}_j}]$, with $t_{n-p-1;1-\alpha/2}$ a Student-quantile and $\sigma_{\hat{\beta}_j}$ the squareroot of the $j^{\text{th}}$ element of $\mathrm{Cov}(\hat{\beta})$.
- Tests on $\beta_j$: the rv $\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}}$ has a Student distribution.
- Confidence region for $\beta = (\beta_0 \dots \beta_p)$:

$$R_{1-\alpha}(\beta) = \left\{ z \in \mathbb{R}^{p+1} | \,^\top(z - \hat{\beta})\,^\top X\, X\,(z - \hat{\beta}) \leq (p+1)s^2 f_{k;n-p-1;1-\alpha} \right\}.$$

It is an ellipsoid centred on $\hat{\beta}$ with volume, shape and orientation depending upon $^\top X\, X$.

# Practical uses

- CI for $\beta_j$: $[\hat{\beta}_j + / - t_{n-p-1;1-\alpha/2}\sigma_{\hat{\beta}_j}]$, with $t_{n-p-1;1-\alpha/2}$ a Student-quantile and $\sigma_{\hat{\beta}_j}$ the squareroot of the $j^{\text{th}}$ element of $\mathrm{Cov}(\hat{\beta})$.
- Tests on $\beta_j$: the rv $\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}}$ has a Student distribution.
- Confidence region for $\beta = (\beta_0 \dots \beta_p)$:

  $$R_{1-\alpha}(\beta) = \left\{ z \in \mathbb{R}^{p+1} | \,^{\top}(z - \hat{\beta})\,^{\top}X\,X\,(z - \hat{\beta}) \leq (p+1)s^2 f_{k;n-p-1;1-\alpha} \right\}.$$

  It is an ellipsoid centred on $\hat{\beta}$ with volume, shape and orientation depending upon $^{\top}X\,X$.
- CI for previsions on $y^*$:

  $$[y^* + / - t_{n-p-1;1-\alpha/2} s \left(1 +^{\top} x^* (^{\top}X\,X)^{-1}\right)^{1/2}].$$

# Usual diagnosis

- residual plot: variance homogeneity (weights can be used if not), model validation...

# Usual diagnosis

- residual plot: variance homogeneity (weights can be used if not), model validation. . .
- QQ-plots: to detect outliers . . .

# Usual diagnosis

- residual plot: variance homogeneity (weights can be used if not), model validation. . .
- QQ-plots: to detect outliers . . .
- model selection. $R^2$ for model with same number of regressors. $R^2_{adj} = \frac{(n-1)R^2-(p-1)}{n-p}$. Maximising $R^2_{adj}$ is equivalent to maximising the mean quadratic error.

# Usual diagnosis

- residual plot: variance homogeneity (weights can be used if not), model validation...
- QQ-plots: to detect outliers ...
- model selection. $R^2$ for model with same number of regressors. $R^2_{adj} = \frac{(n-1)R^2-(p-1)}{n-p}$. Maximising $R^2_{adj}$ is equivalent to maximising the mean quadratic error.
- test by ANOVA: $F = \frac{SSR/p}{SSE/(n-p-1)}$ has a Fisher distribution with $p, (n-p-1)$ df. Since testing (H0) $\beta_1 = \ldots = \beta_p = 0$ has little interest (rejected asa one of the variable is linked to $Y$), one can test (H0') $\beta_{i_1} = \ldots = \beta_{i_q} = 0$, with $q < p$ and $\frac{(SSR-SSR_q)/q}{SSE/(n-p-1)}$ has a Fisher distribution with $q, (n-p-1)$ df.

# Usual diagnosis

- residual plot: variance homogeneity (weights can be used if not), model validation...
- QQ-plots: to detect outliers ...
- model selection. $R^2$ for model with same number of regressors. $R^2_{adj} = \frac{(n-1)R^2-(p-1)}{n-p}$. Maximising $R^2_{adj}$ is equivalent to maximising the mean quadratic error.
- test by ANOVA: $F = \frac{SSR/p}{SSE/(n-p-1)}$ has a Fisher distribution with $p, (n-p-1)$ df. Since testing (H0) $\beta_1 = \ldots = \beta_p = 0$ has little interest (rejected asa one of the variable is linked to $Y$), one can test (H0') $\beta_{i_1} = \ldots = \beta_{i_q} = 0$, with $q < p$ and $\frac{(SSR-SSR_q)/q}{SSE/(n-p-1)}$ has a Fisher distribution with $q, (n-p-1)$ df.
- Application: variable selection for model interpretation: backward (remove 1 by 1 least significative with t-test), forward (include 1 by 1 most significative with F-test), stepwise (variant of forward).

# Collinearity and model selection

- detecting colinearity between the $x_i$'s. Inverting ${}^\top X\, X$ if $\det({}^\top X\, X) \approx 0$ is difficult. Moreover, the inverse will have a huge variance !

# Collinearity and model selection

- detecting colinearity between the $x_i$'s. Inverting ${}^{\top}X\,X$ if $\det({}^{\top}X\,X) \approx 0$ is difficult. Moreover, the inverse will have a huge variance !
- to detect collinearity, compute $VIF(x_j) = \frac{1}{1-R_j^2}$, with $R_j^2$ the determination coefficient of $x_j$ regressed againt $x \setminus \{x_j\}$. Perfect othrogonality is $VIF(x_j) = 1$ and the stronger the collinearity, the larger the value for $VIF(x_j)$.

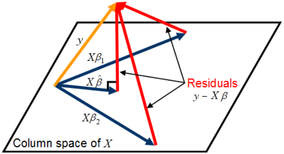
# Collinearity and model selection

- detecting colinearity between the $x_i$'s. Inverting ${}^\top X\,X$ if $\det({}^\top X\,X) \approx 0$ is difficult. Moreover, the inverse will have a huge variance !
- to detect collinearity, compute $VIF(x_j) = \frac{1}{1-R_j^2}$, with $R_j^2$ the determination coefficient of $x_j$ regressed againt $x \setminus \{x_j\}$. Perfect othrogonality is $VIF(x_j) = 1$ and the stronger the collinearity, the larger the value for $VIF(x_j)$.
- Ridge regression introduces a bias but reduces the variance (keeps all variables). Lasso regression does the same but also does a selection on variables. Issue here: penalty term to tune...

# Last generalisations

## Multiple outputs, GLM and non-linear regressions

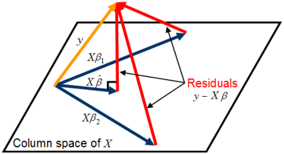
- Multiple output regression



# Last generalisations

Multiple outputs, GLM and non-linear regressions

- Multiple output regression



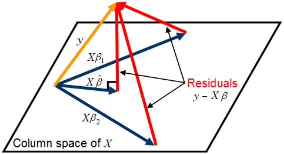

- GLM are of the form

$$Y = \beta_0 + \sum_j \beta_j x^j + \sum_{k,l} \beta_{k,l} x^k x^l + \epsilon.$$

# Last generalisations

## Multiple outputs, GLM and non-linear regressions

- Multiple output regression




- GLM are of the form

$$Y = \beta_0 + \sum_j \beta_j x^j + \sum_{k,l} \beta_{k,l} x^k x^l + \epsilon.$$

- Non-linear (parametric) regression has the form $Y = f(x;\theta) + \epsilon$. Examples include exponential or logistic models.

# Today's session is over

Next time: A practical R session to be studied by you !