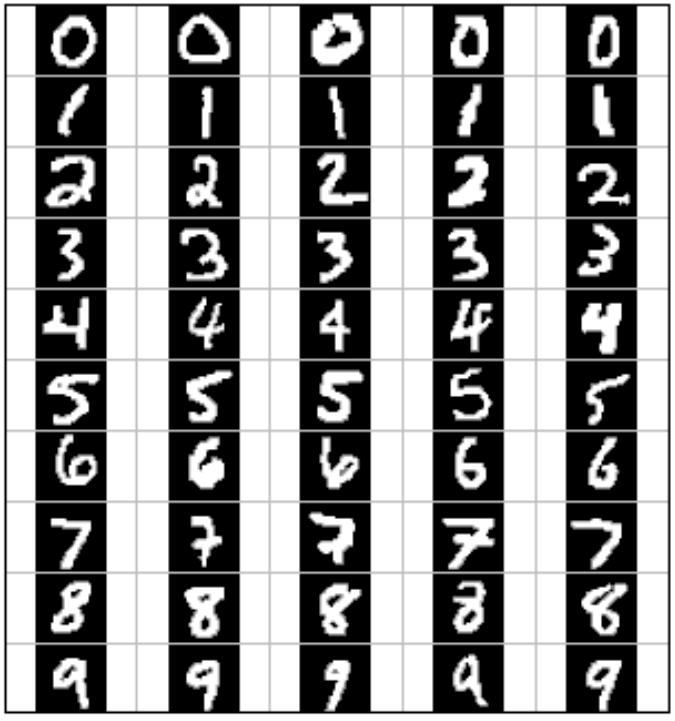
Regression avancée
Chapitre 1 : introduction

Agathe Guilloux
Professeure au LaMME - Université d'Évry - Paris Saclay

Exemples

Des chiffres

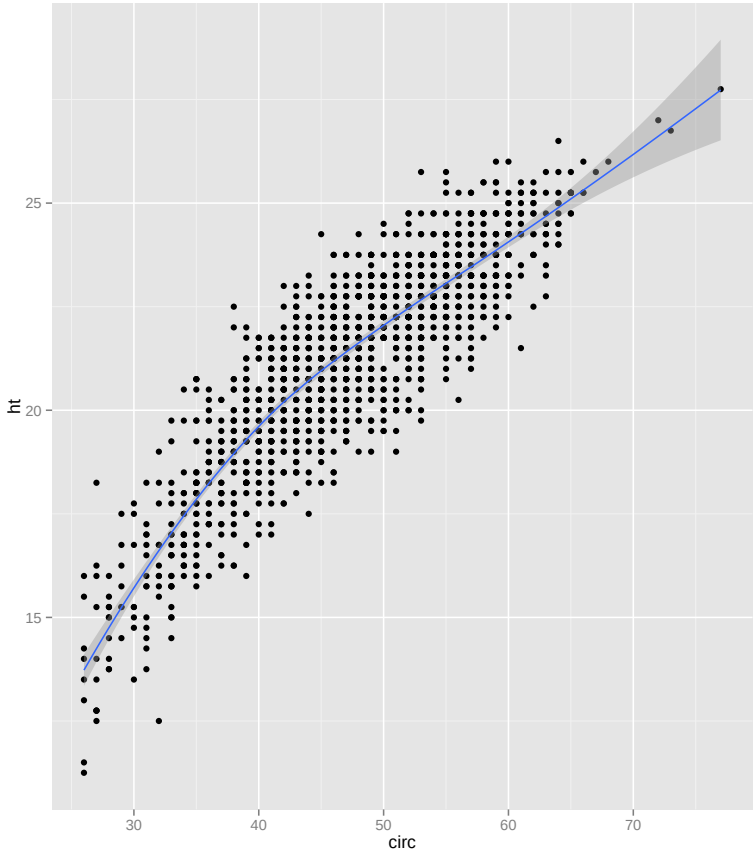


- Lire un code postal sur une enveloppe.
- But : assigner un chiffre à une image.
- Input : image.
- Output : chiffre correspondant.

Spam detection

- Données : emails.
- Input : email.
- Output : Spam or No Spam.

Eucalyptus



- But : prédire la hauteur en fonction de la circonférence.
- Input : circonférence.
- Output : hauteur.

Données “Vulnerability”, Patt et al., PNAS (2009)

Les pays les moins développés sont-ils plus vulnérables aux changements climatiques ?

Les auteurs ont voulu expliquer `ln_death_risk`, log du risque mortel dû aux évènement climatiques en fonction

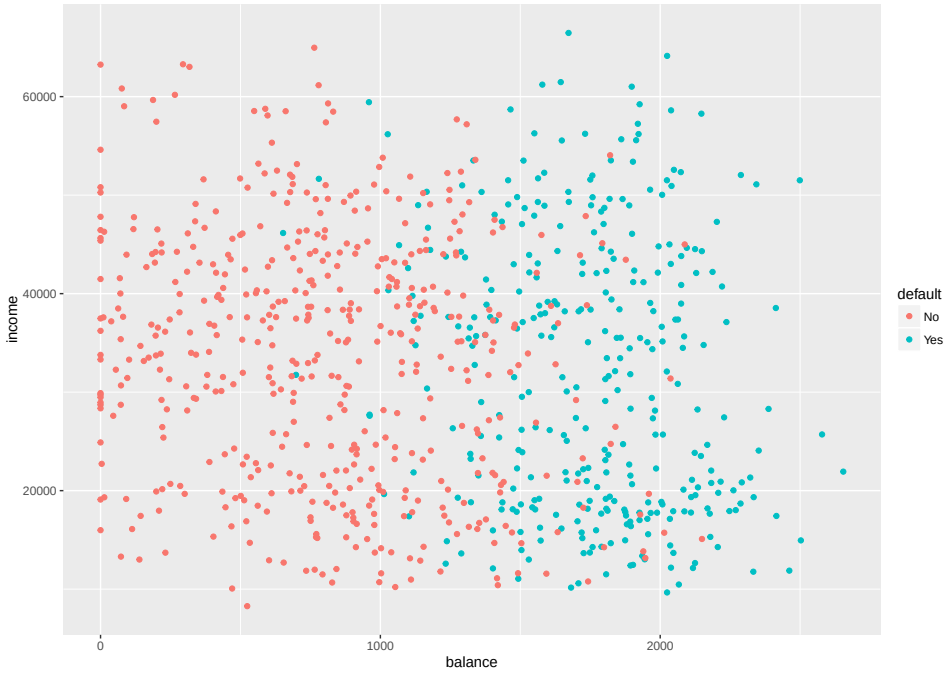
- du log du nombre d'évènements climatiques `ln_death_risk`
- du log de la fertilité `ln_fertility`
- de l'indice de développement humain `hdi` (United Nations)
- du log de la population `ln_pop`

Ils concluent que le développement socio-économique a un lien sur la fragilité aux événements climatiques, et ce lien pourrait se révéler dans le deuxième quart du 21ième siècle.

Visualisation des données "Vulnerability"

country_name	ln_events	ln_fert	hdi	ln_pop	ln_death_risk
Albania	2.3025850	1.2383740	0.7530000	4.0061200	-0.7102835002
Algeria	3.4965080	1.5993880	0.7025000	6.2838850	0.8961844999
Angola	3.0445230	1.9459100	0.4460000	5.5560560	0.2246879996
Argentina	3.6375860	1.0116010	0.8525001	6.4835150	-1.1036180004
Armenia	1.3862940	0.7654679	0.7380000	3.9765620	-2.3671239981
Australia	4.3944490	0.7654679	0.9480000	5.8379250	-1.0504329996
Austria	3.0910430	0.5306283	0.9330000	4.9908860	-1.4073670018
Azerbaijan	1.7917590	1.0986120	0.7460000	4.9572340	-2.1846459984
Bahamas	2.3025850	1.0116010	0.8325000	1.6226830	1.3217560000
Bangladesh	4.8362820	1.5475620	0.5000000	7.8287280	4.1112999999
Belarus	1.6094380	0.5596158	0.7795000	5.1651670	-3.2192569975

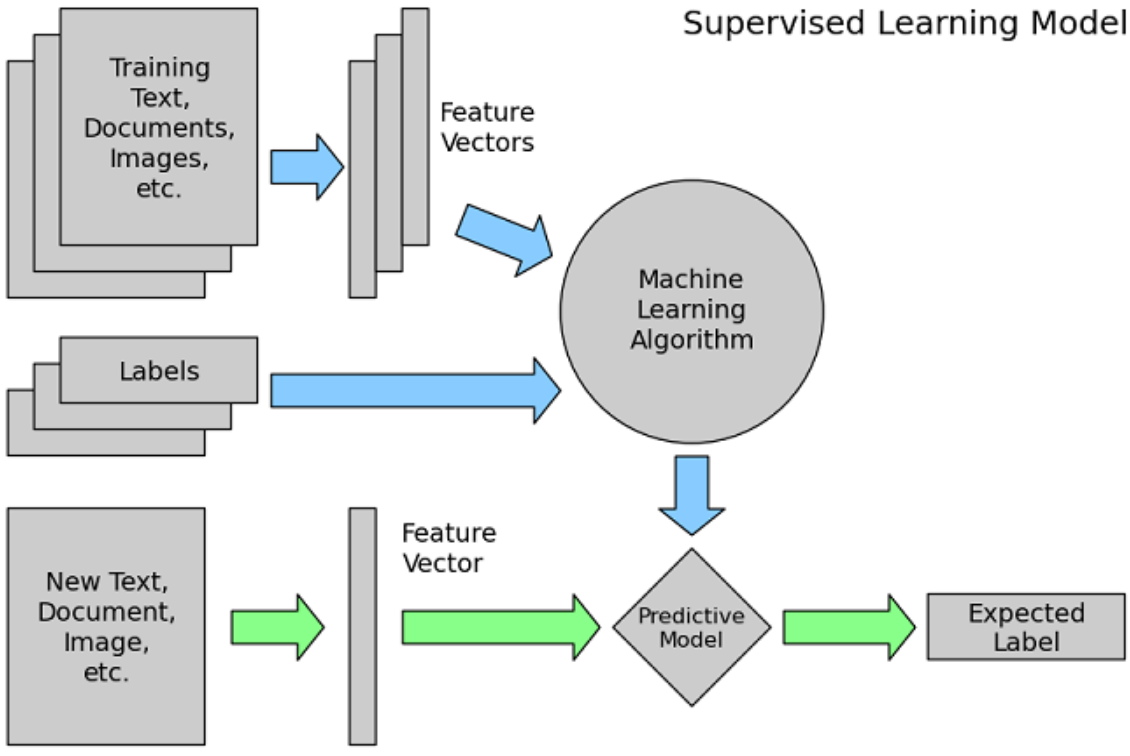
Défaut de crédit



- But : prédire le défaut (ou non de paiement)
- Input : `balance` : le solde du compte client, `income` : les revenus du client
- Output : défault `Yes` ou `No`

Apprentissage statistique supervisé

Apprentissage statistique supervisé



- Input : covariables, variables explicatives, features $\mathbf{X} = (X^1, \ldots, X^p)$
- Ouput : variable à expliquer, variable dépendante, réponse, label $\mathbf{Y}$

Données "Vulnerability"

country_name	ln_events	ln_fert	hdi	ln_pop	ln_death_risk
Albania	2.3025850	1.2383740	0.7530000	4.0061200	-0.7102835002
Algeria	3.4965080	1.5993880	0.7025000	6.2838850	0.8961844999
Angola	3.0445230	1.9459100	0.4460000	5.5560560	0.2246879996
Argentina	3.6375860	1.0116010	0.8525001	6.4835150	-1.1036180004
Armenia	1.3862940	0.7654679	0.7380000	3.9765620	-2.3671239981
Australia	4.3944490	0.7654679	0.9480000	5.8379250	-1.0504329996
Austria	3.0910430	0.5306283	0.9330000	4.9908860	-1.4073670018
Azerbaijan	1.7917590	1.0986120	0.7460000	4.9572340	-2.1846459984
Bahamas	2.3025850	1.0116010	0.8325000	1.6226830	1.3217560000
Bangladesh	4.8362820	1.5475620	0.5000000	7.8287280	4.1112999999
Belarus	1.6094380	0.5596158	0.7795000	5.1651670	-3.2192569975

Dans notre exemple :

- X^1 est `ln_events`
- X^2 est `ln_fert`
- X^3 est `hdi`
- X^4 est `ln_pop` et
- Y est `ln_death_risk`

Données

Données d'entrainement / learning set / training set

$\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ avec pour chaque individu $i = 1, \ldots, n$

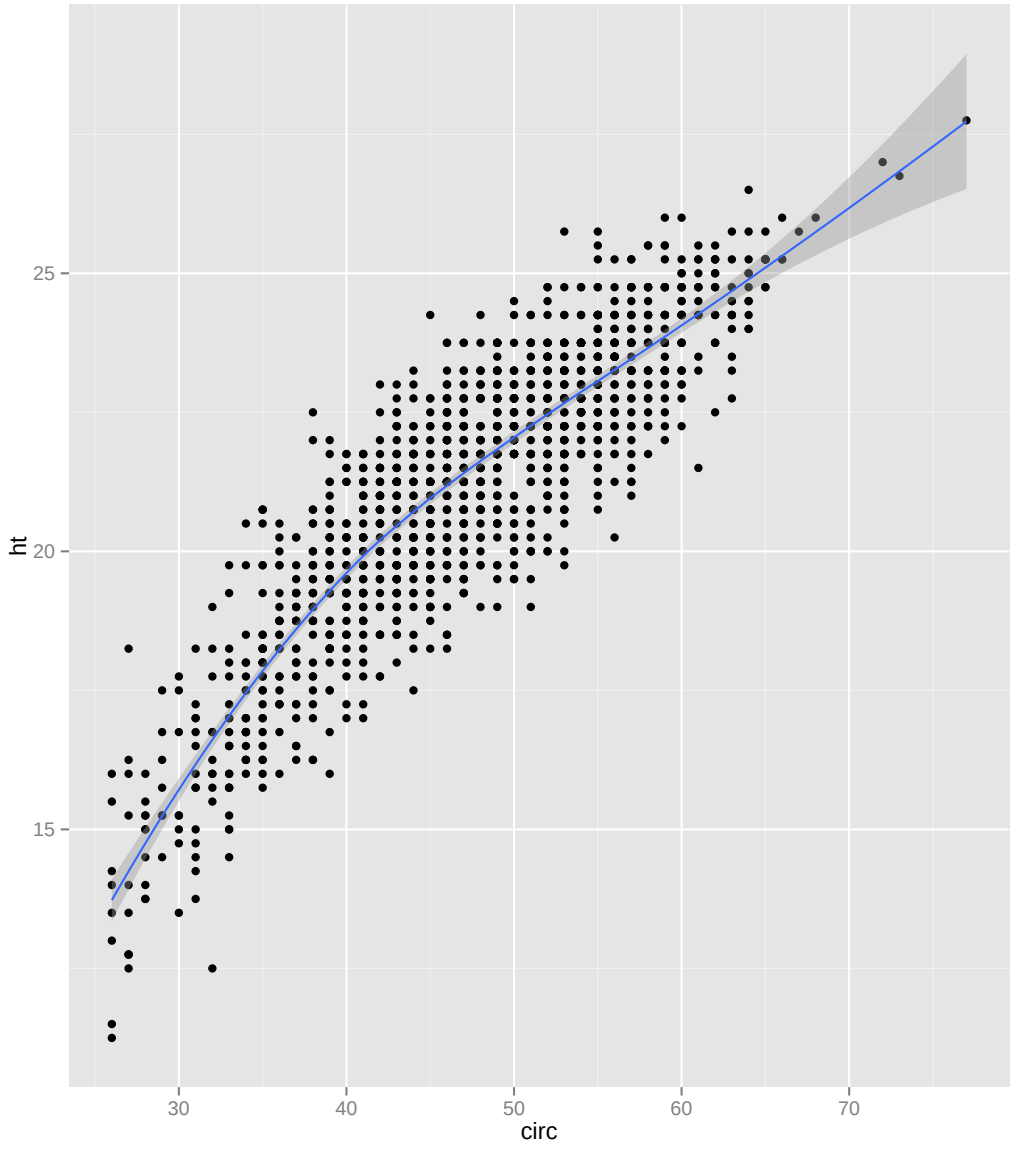
- des features
$$\mathbf{X}_i = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(d)}) \in \mathcal{X}$$
- un label $Y_i \in \mathcal{Y}$

Hypothèse : les $(\mathbf{X}_i, Y_i)$ sont i.i.d. de loi $\mathbf{P}$ inconnue.

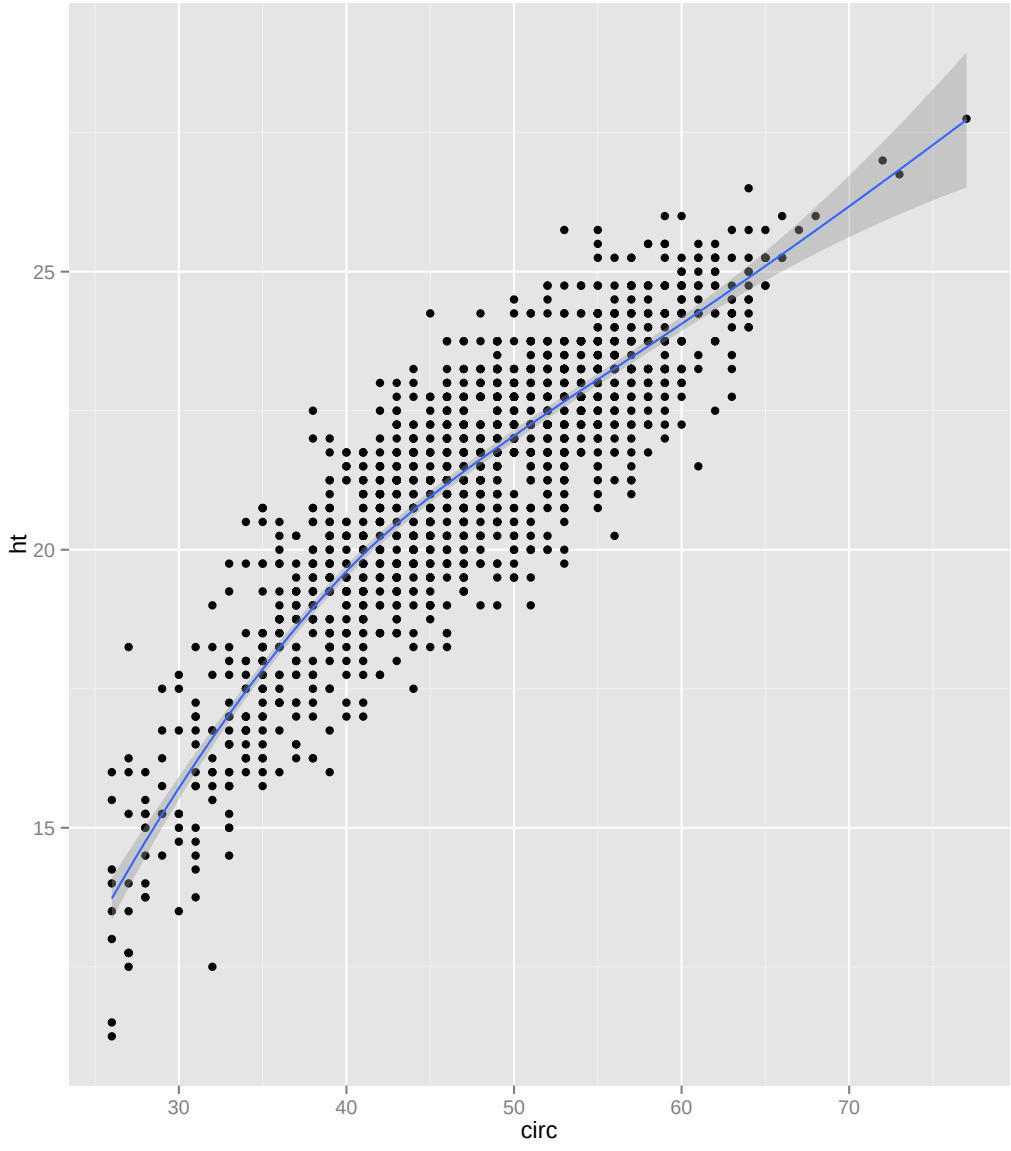
Deux exemples courants :

- $\mathbf{X} \in \mathbb{R}^d$ and $Y \in \{-1, 1\}$ (classification)
- et $\mathbf{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ (régression).

Eucalyptus

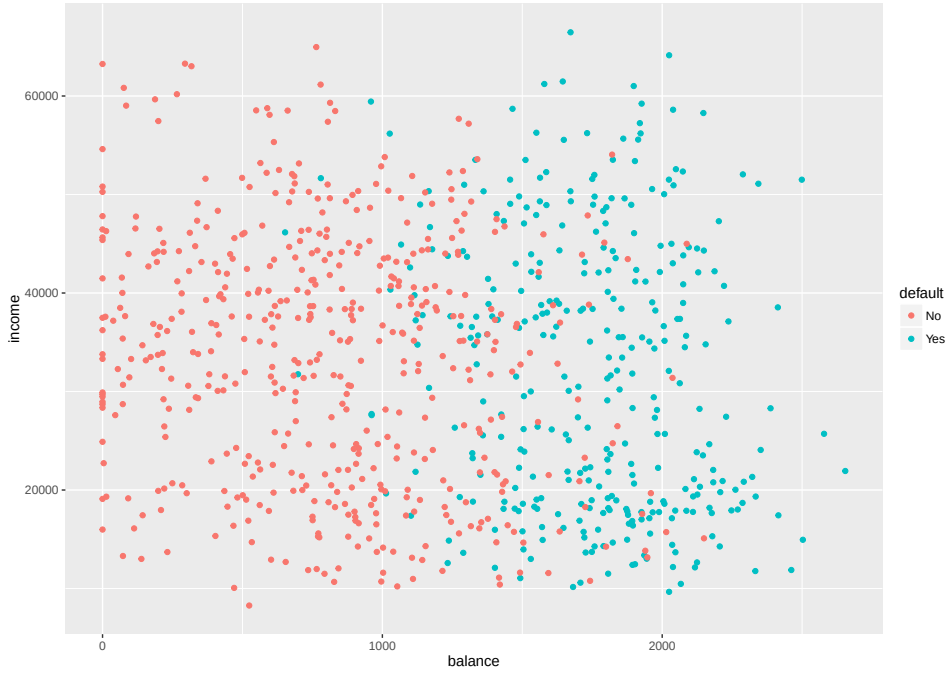


Eucalyptus

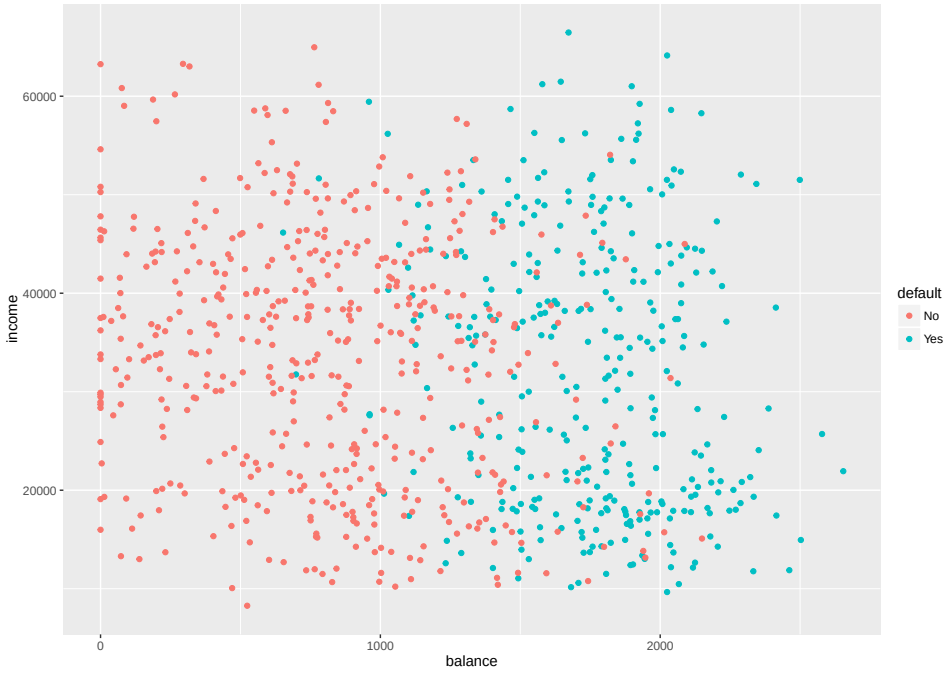


- $\mathcal{X} = \mathbb{R}$
- $\mathcal{Y} = \mathbb{R}$

Défaut de crédit



Défaut de crédit



- $\mathcal{X} = \mathbb{R}^2$
- $\mathcal{Y} = \{\text{Yes}, \text{No}\}$ ou $\{0, 1\}$

Données "Vulnerability"

country_name	ln_events	ln_fert	hdi	ln_pop	ln_death_risk
Albania	2.3025850	1.2383740	0.7530000	4.0061200	-0.7102835002
Algeria	3.4965080	1.5993880	0.7025000	6.2838850	0.8961844999
Angola	3.0445230	1.9459100	0.4460000	5.5560560	0.2246879996
Argentina	3.6375860	1.0116010	0.8525001	6.4835150	-1.1036180004
Armenia	1.3862940	0.7654679	0.7380000	3.9765620	-2.3671239981
Australia	4.3944490	0.7654679	0.9480000	5.8379250	-1.0504329996
Austria	3.0910430	0.5306283	0.9330000	4.9908860	-1.4073670018
Azerbaijan	1.7917590	1.0986120	0.7460000	4.9572340	-2.1846459984
Bahamas	2.3025850	1.0116010	0.8325000	1.6226830	1.3217560000
Bangladesh	4.8362820	1.5475620	0.5000000	7.8287280	4.1112999999
Belarus	1.6094380	0.5596158	0.7795000	5.1651670	-3.2192569975

Données "Vulnerability"

country_name	ln_events	ln_fert	hdi	ln_pop	ln_death_risk
Albania	2.3025850	1.2383740	0.7530000	4.0061200	-0.7102835002
Algeria	3.4965080	1.5993880	0.7025000	6.2838850	0.8961844999
Angola	3.0445230	1.9459100	0.4460000	5.5560560	0.2246879996
Argentina	3.6375860	1.0116010	0.8525001	6.4835150	-1.1036180004
Armenia	1.3862940	0.7654679	0.7380000	3.9765620	-2.3671239981
Australia	4.3944490	0.7654679	0.9480000	5.8379250	-1.0504329996
Austria	3.0910430	0.5306283	0.9330000	4.9908860	-1.4073670018
Azerbaijan	1.7917590	1.0986120	0.7460000	4.9572340	-2.1846459984
Bahamas	2.3025850	1.0116010	0.8325000	1.6226830	1.3217560000
Bangladesh	4.8362820	1.5475620	0.5000000	7.8287280	4.1112999999
Belarus	1.6094380	0.5596158	0.7795000	5.1651670	-3.2192569975

- $\mathcal{X} = \mathbb{R}^4$
- $\mathcal{Y} = \mathbb{R}$

Spam detection

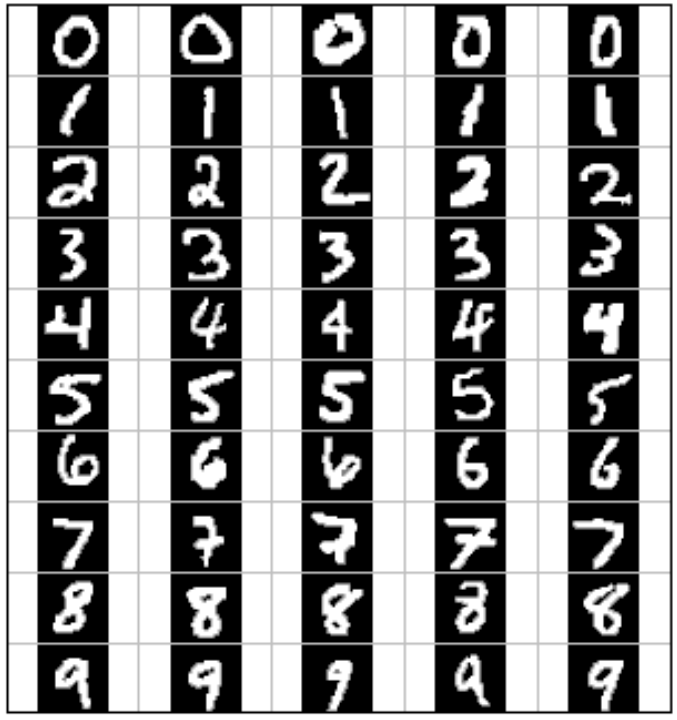
- Input : email.
- Output : Spam or No Spam.

Spam detection

- Input : email.
- Output : Spam or No Spam.

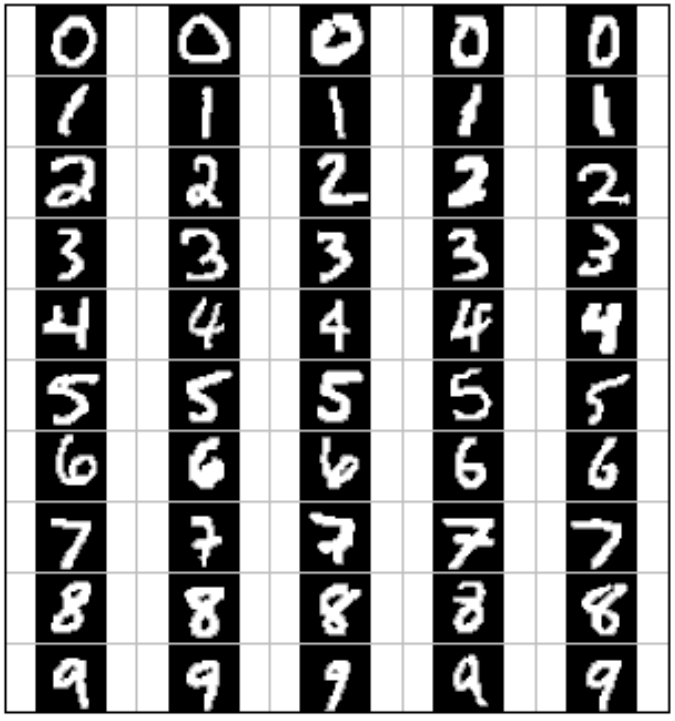
- $\mathcal{X}$ =??? –> features design
- $\mathcal{Y} = \{0, 1\}$

Des chiffres



- Input : image.
- Output : chiffre correspondant.

Des chiffres



- Input : image.
- Output : chiffre correspondant.

- $\mathcal{X}$ =??? –> features design
- $\mathcal{Y} = \{0, 1, 2, \ldots, 9\}$

Prédicteur / predictor

Prédicteur

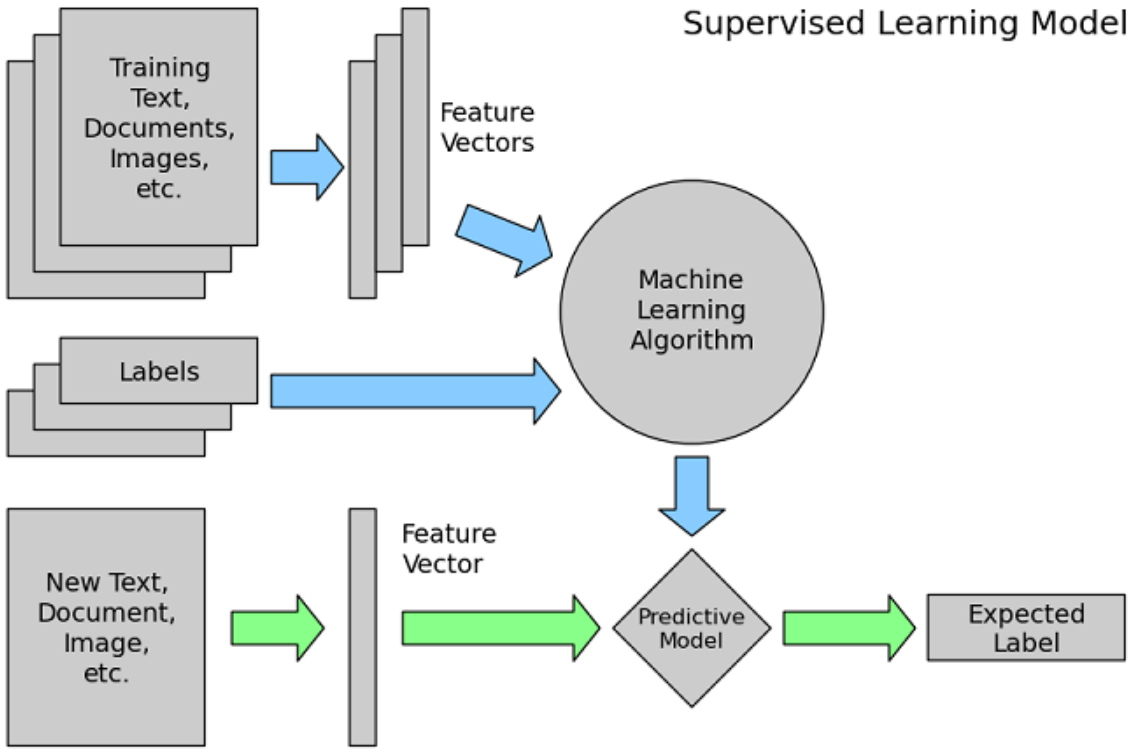
Un prédicteur est une fonction dans $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y} \text{ measurable}\}$

But de l'apprentissage supervisé : construire un **bon prédicteur** $\widehat{f}$ à partir des données d'entraînement.

Attention : il nous faudra préciser le terme **bon** !! En un certain sens (à définir), on veut que

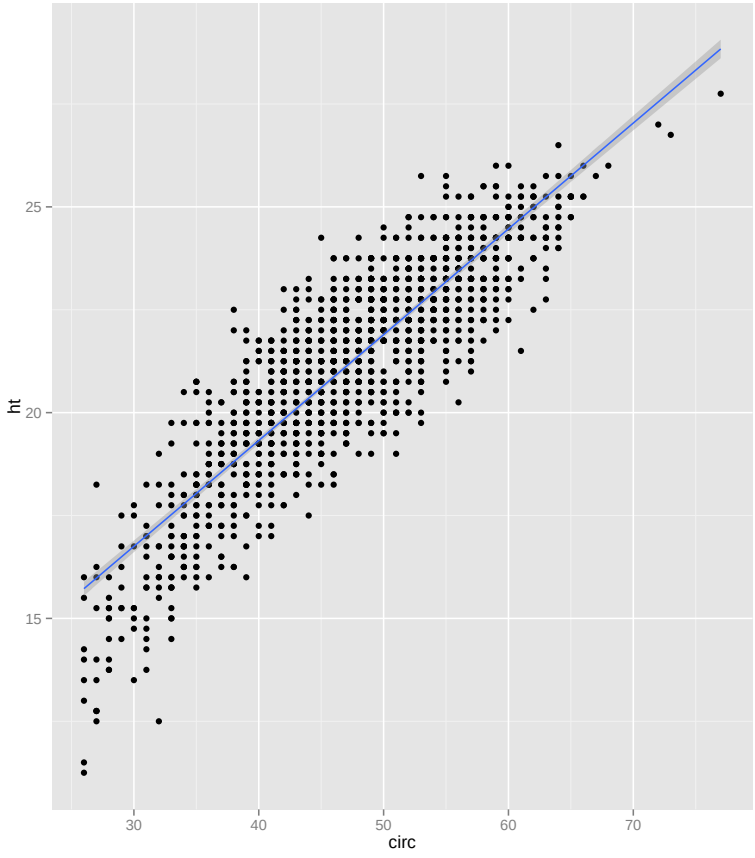
$$Y_i \simeq \widehat{f}(\mathbf{X}_i).$$

Apprentissage statistique supervisé



Régression linéaire univariée

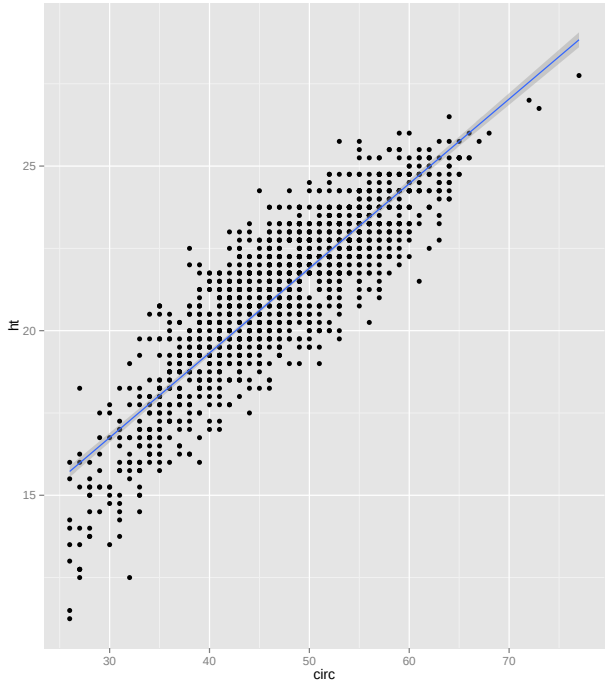
Eucalyptus



- But : prédire la hauteur en fonction de la circonférence.
- Input : circonférence **circ**.
- Output : hauteur ht.

On peut commencer par chercher $\widehat{f}$ parmi des fonctions très simples : les fonctions linéaires.

Eucalyptus



- Modèle linéaire :

$$\mathcal{F} = \{f_\beta(\mathbf{circ}) = \beta_1 + \beta_2 \mathbf{circ}, \beta_1, \beta_2 \in \mathbb{R}\}$$

- Comment choisir $\beta = (\beta_1, \beta_2)$?

Moindres carrés

On veut que

$$Y_i \simeq \widehat{f}(X_i).$$

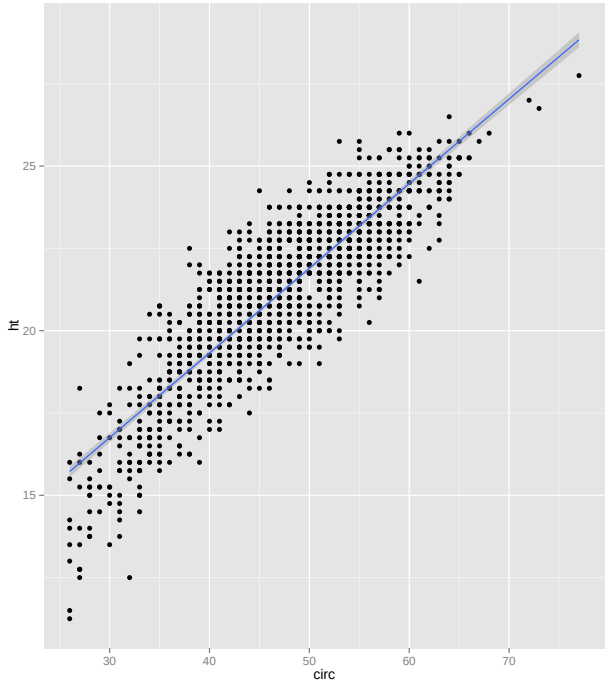
- Critère naturel

$$\sum_{i=1}^{n} |Y_i - f_\beta(\mathbf{X}_i)|^2 = \sum_{i=1}^{n} |\mathrm{ht}_i - f_\beta(\mathbf{circ}_i)|^2$$
$$= \sum_{i=1}^{n} |\mathrm{ht}_i - (\beta_1 + \beta_2 \mathbf{circ}_i)|^2$$

- On choisit β qui minimise ce critère :

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^2}{\mathrm{argmin}} \sum_{i=1}^{n} |h_i - (\beta_1 + \beta_2 \mathbf{circ}_i)|^2$$

Prédiction

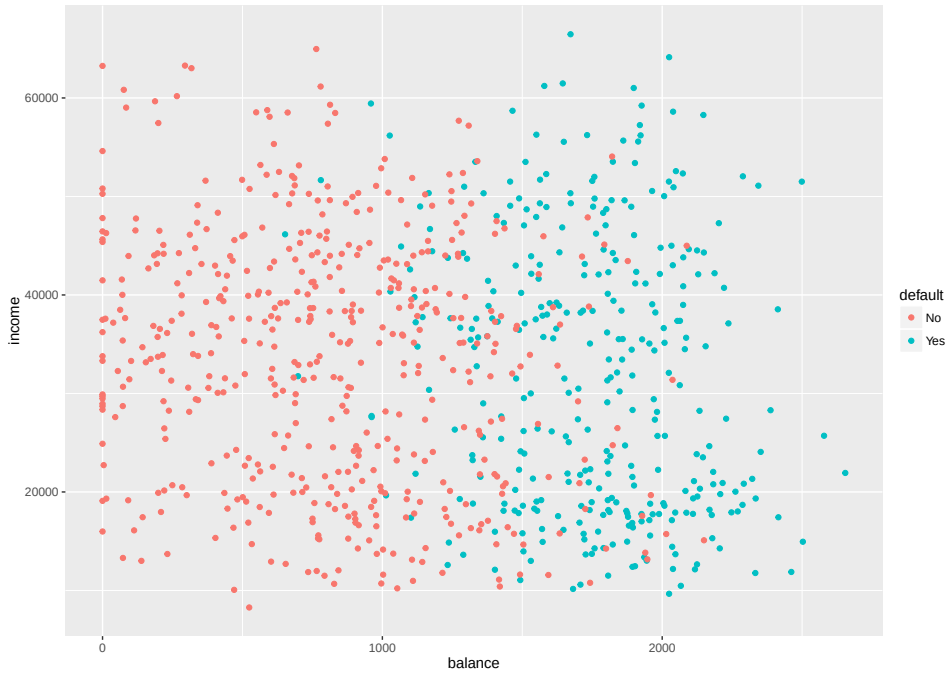


- Prédicteur linéaire pour la hauteur :

$$\widehat{\mathsf{ht}} = f_{\widehat{\beta}}(\mathbf{circ}) = \widehat{\beta}_1 + \widehat{\beta}_2 \mathbf{circ}$$

Pas si simple...

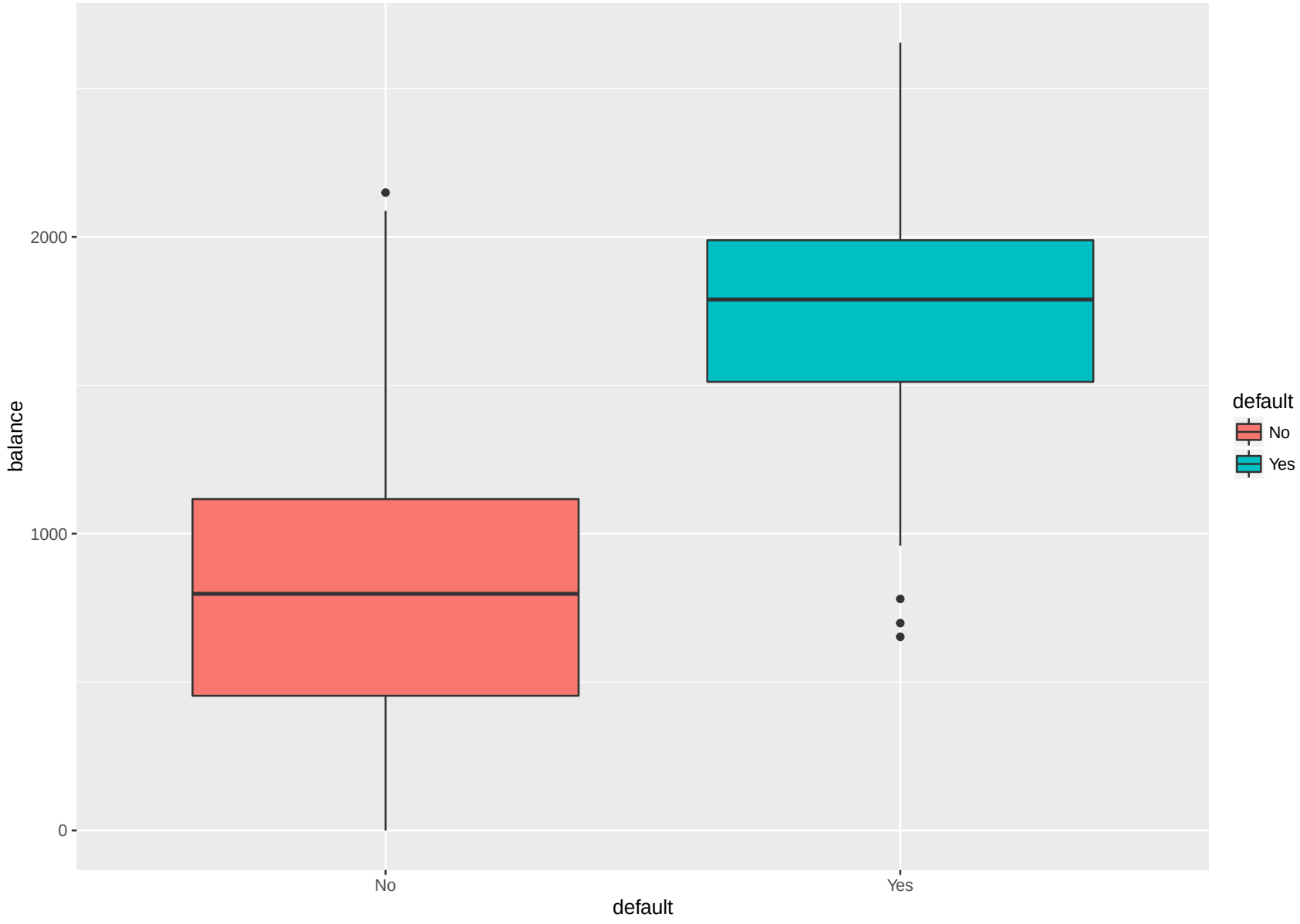
Défaut de crédit



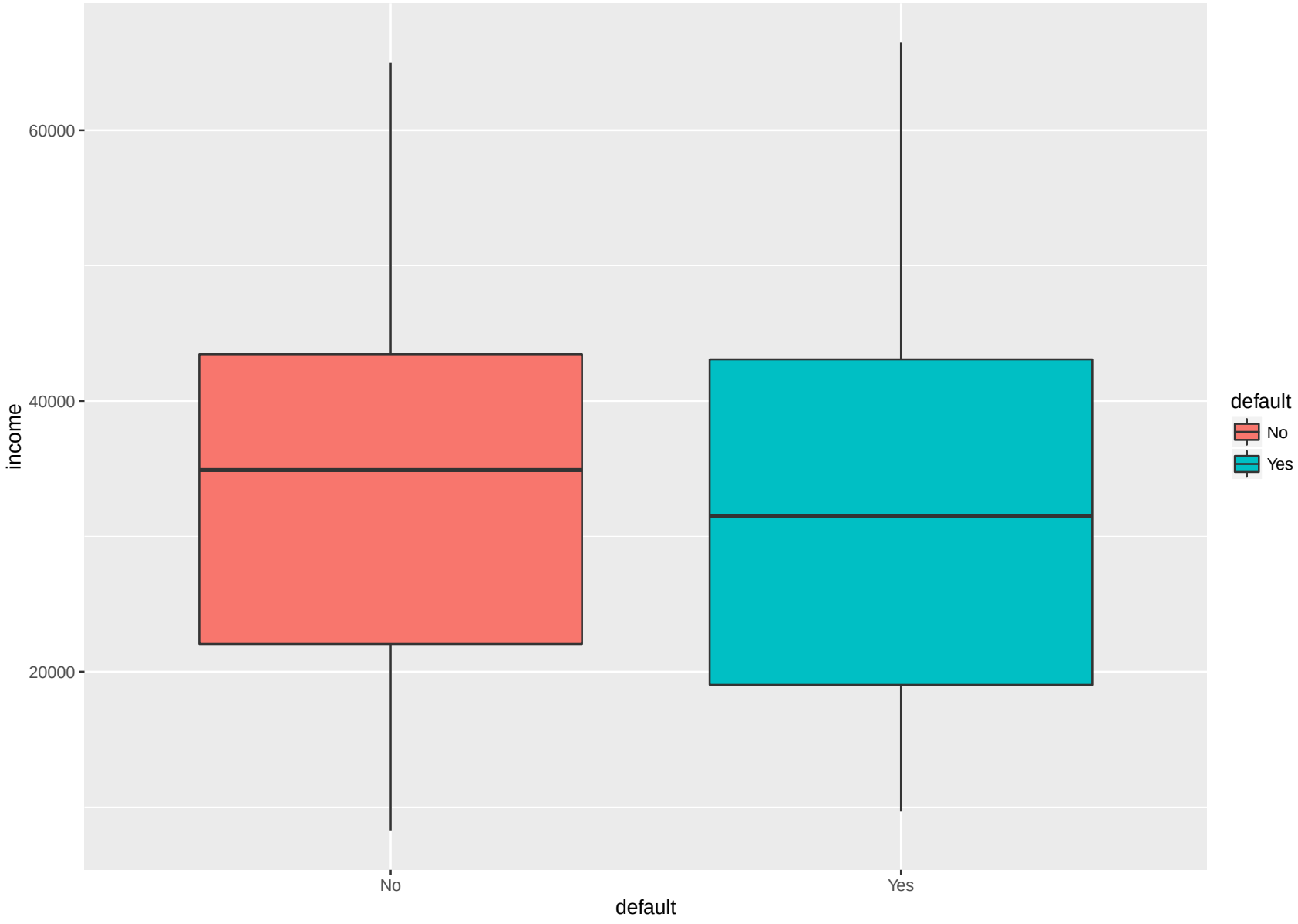
- $\mathcal{X} = \mathbb{R}^2$
- $\mathcal{Y} = \{0, 1\}$

On veut prédire le défaut (ou $\{0, 1\}$) à partir du solde du compte et du revenu.

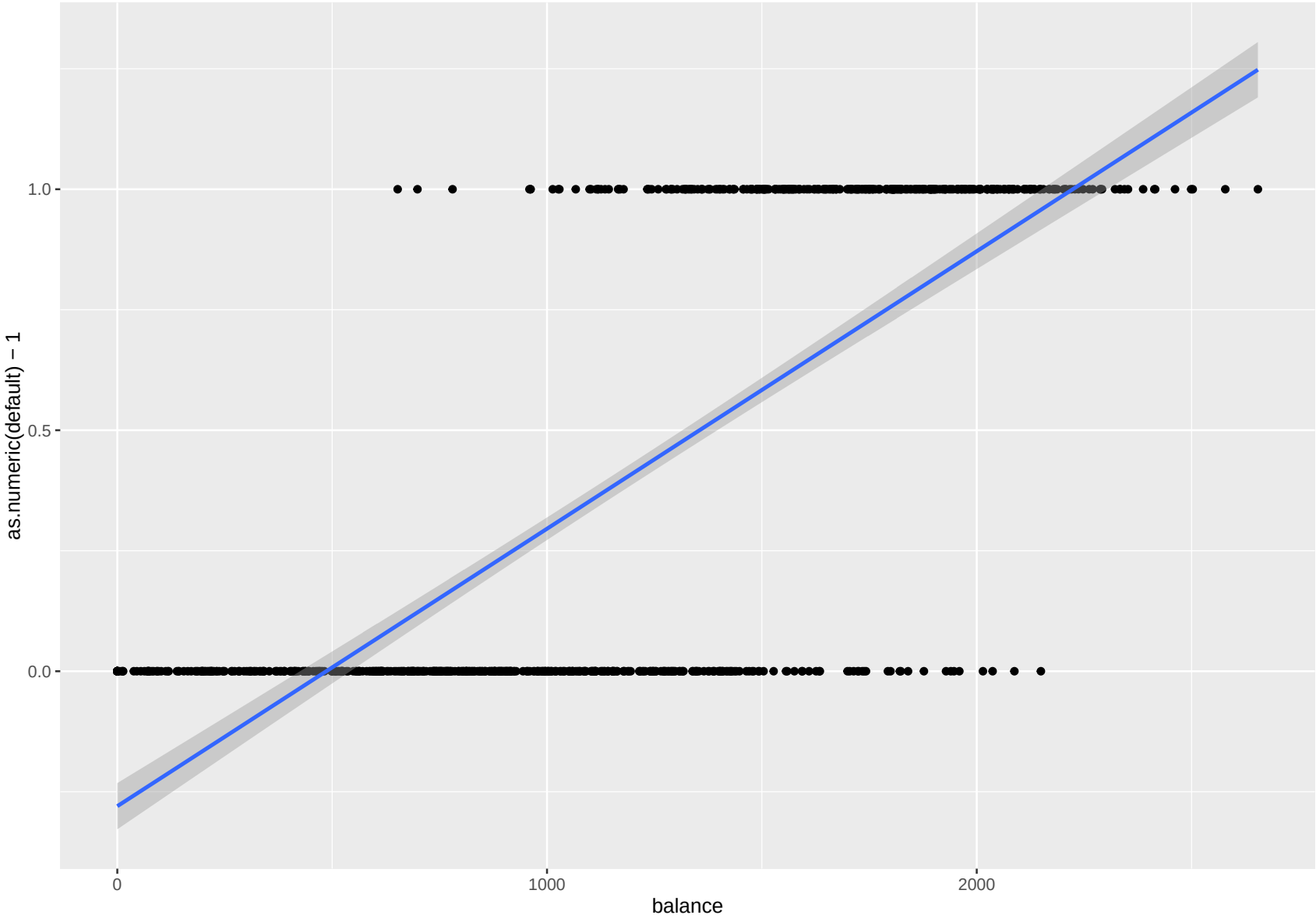
Pour le solde



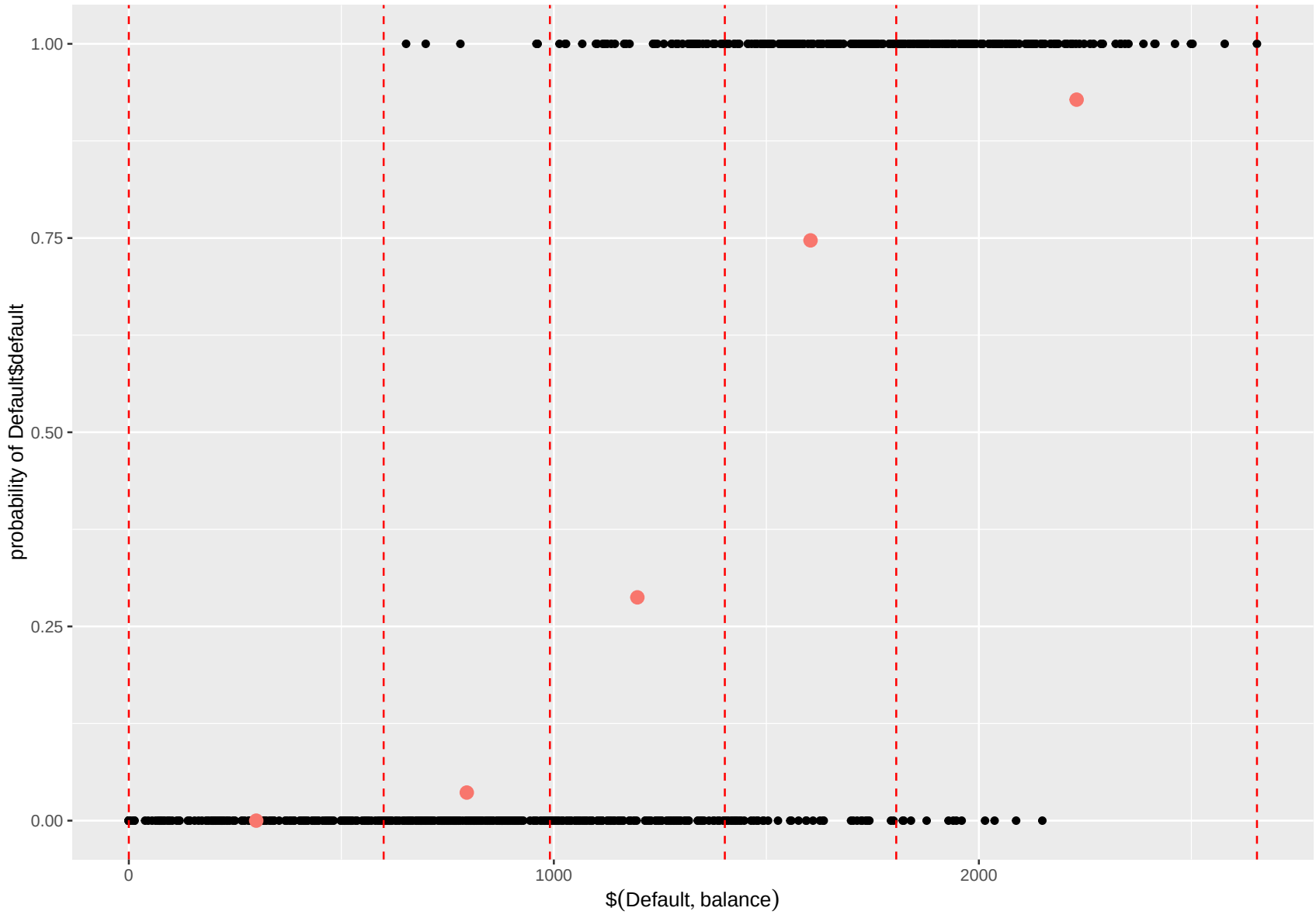
Pour le revenu



Par régression linéaire ?



L'espérance des Y_i sachant $\mathbf{X_i}$



Une approche possible : la modélisation statistique

La loi de Y_i sachant $\mathbf{X}_i$

Dans ces exemples, c'est l'espérance des Y_i qui dépend des X_i et

- pour les données "Défaut de crédit" : $Y_i \in \{0, 1\}$ suit une loi de Bernoulli
- pour les données "Eucalyptus" : $Y_i \in \mathbb{R}$, on pense à une loi normale.

On va modéliser la loi de Y_i sachant $\mathbf{X}_i$:

- on va supposer que $\mathbb{E}(Y_i|\mathbf{X}_i) = f^\star(\mathbf{X}_i)$
- bien on va chercher à construire un prédicteur $\hat{f}$ "proche" de $f^\star$ à partir des données.

Résumé

Dans ce cours

Nous allons

- adopter le point de vue de la régression, c'est-à-dire modéliser la loi de Y_i sachant $\mathbf{X}_i$,
- les familles de lois que nous allons considérer appartiennent à la **"famille exponentielle"** (cf. Chapitre 2)
- notre façon de définir le "bon" sera le **maximum de vraisemblance** (cf. Chapitre 2)

Nous parlerons dans ce cours de **"modèle linéaire généralisé"** (generalized linear model / GLM).

Quelques remarques :

- il y a d'autres approches possibles –> cf. **Cours de Machine learning du Semestre 2**
- le cas gaussien se traite plus simplement avec une approche géométrique –> cf. **Cours de Modélisation statistique**.

Outline