# Statistiques multivariées

Agathe Guilloux
agathe.guilloux@math.cnrs.fr

# Chapitre 2 : modèle linéaire généralisé

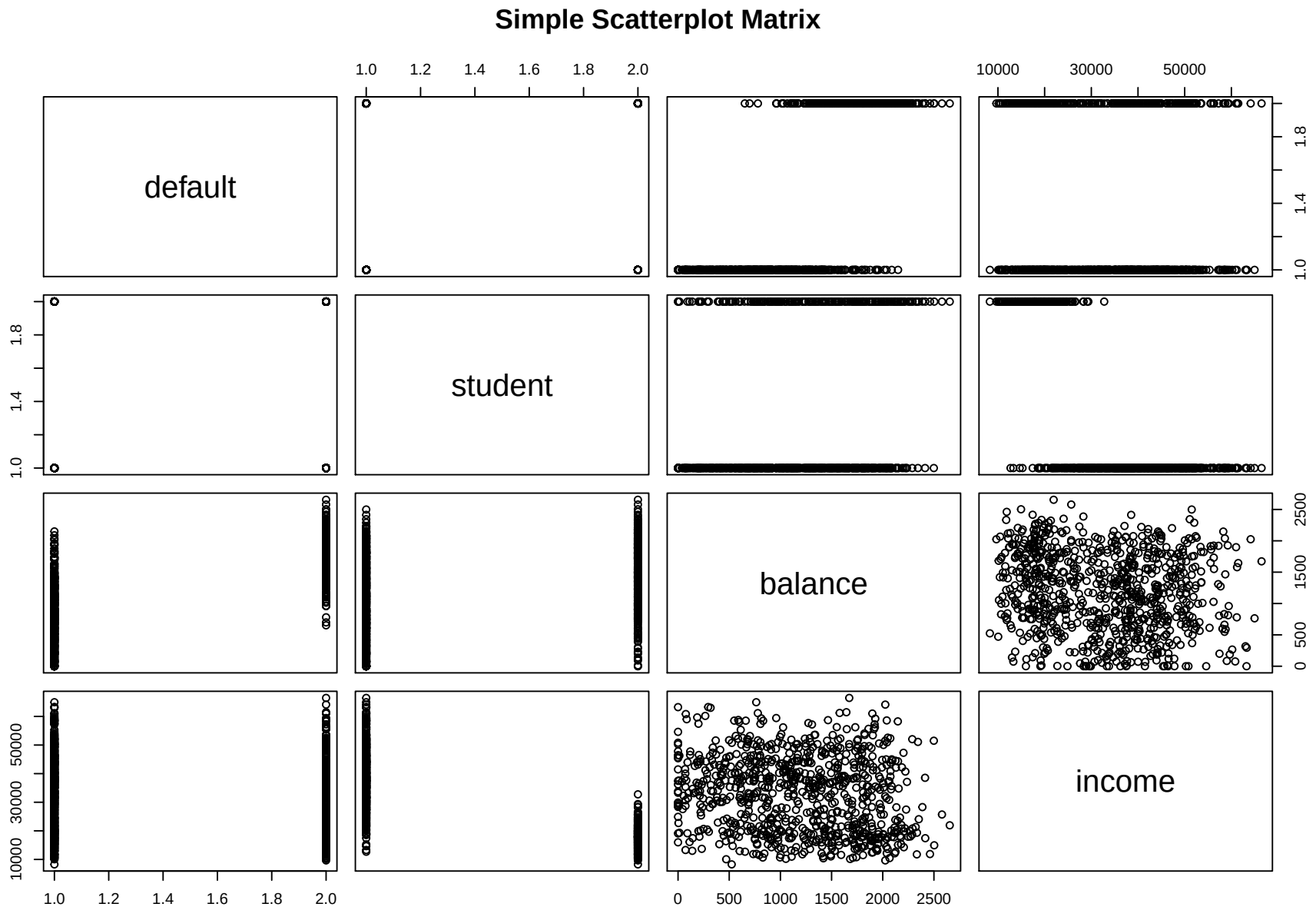
# Les jeux de données

"Default" : téléchargement, voir aussi [JWHT13]
On veut expliquer le défaut de paiement (`default=Yes`) ou non (`default=No`) de 833 clients avec les variables

- `student` : `Yes` ou `No`
- `balance` : le solde du compte client
- `income` : les revenus du client

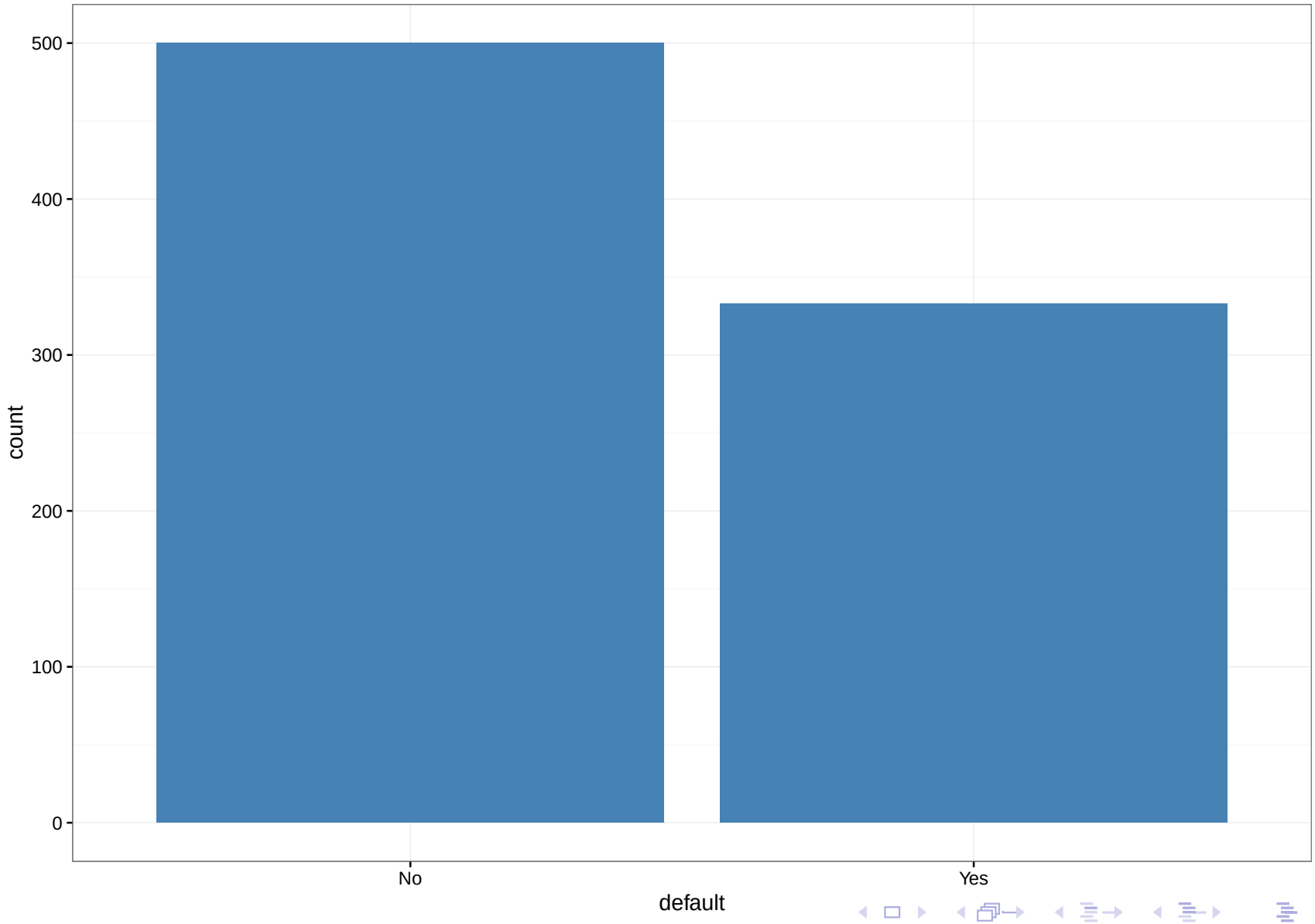
# scatterplot

```
pairs(Default,  main="Simple Scatterplot Matrix")
```



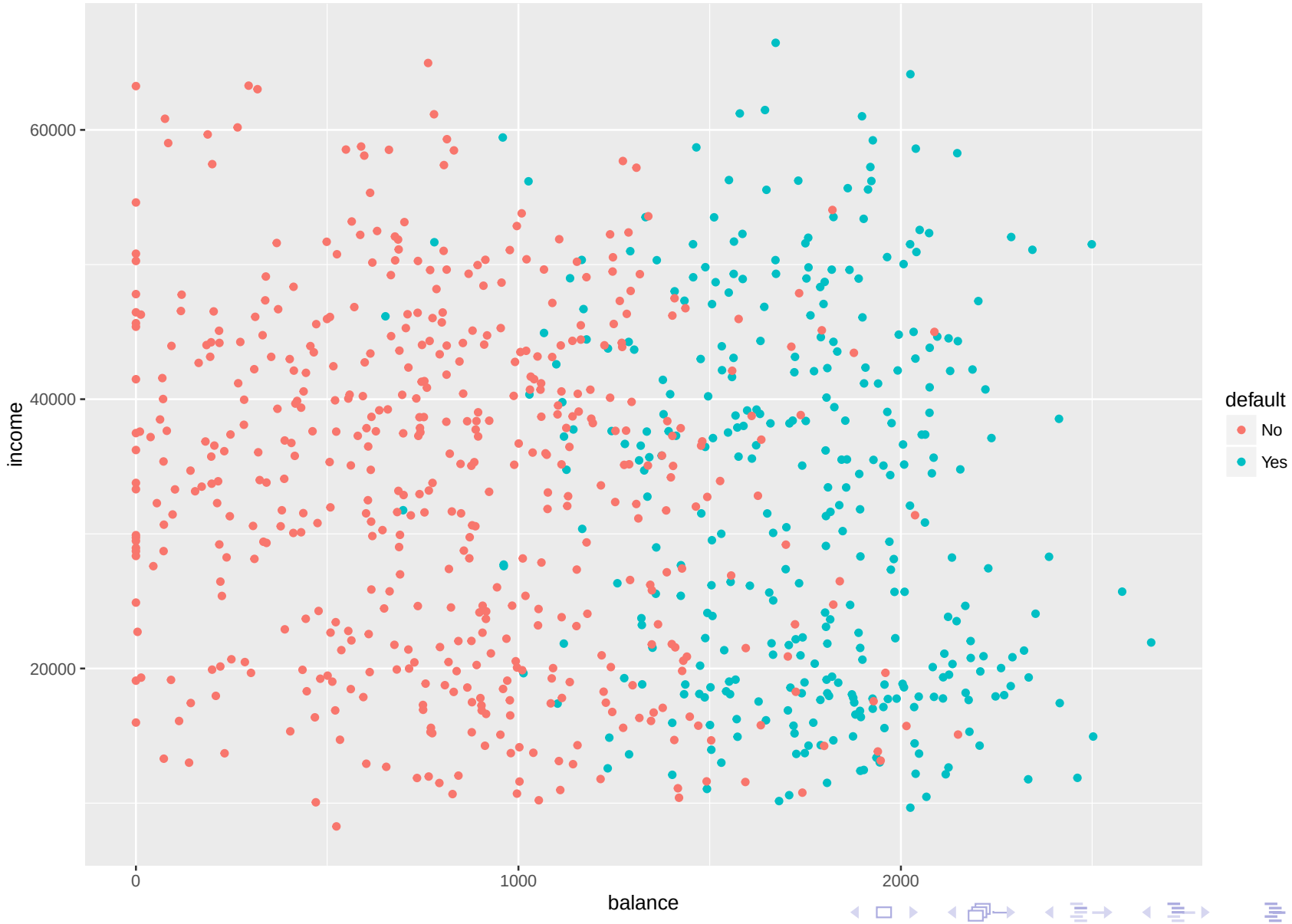
# histogram

```
ggplot(Default, aes(x=default))  +
  geom_bar( fill = "steelblue") + theme_bw()
```

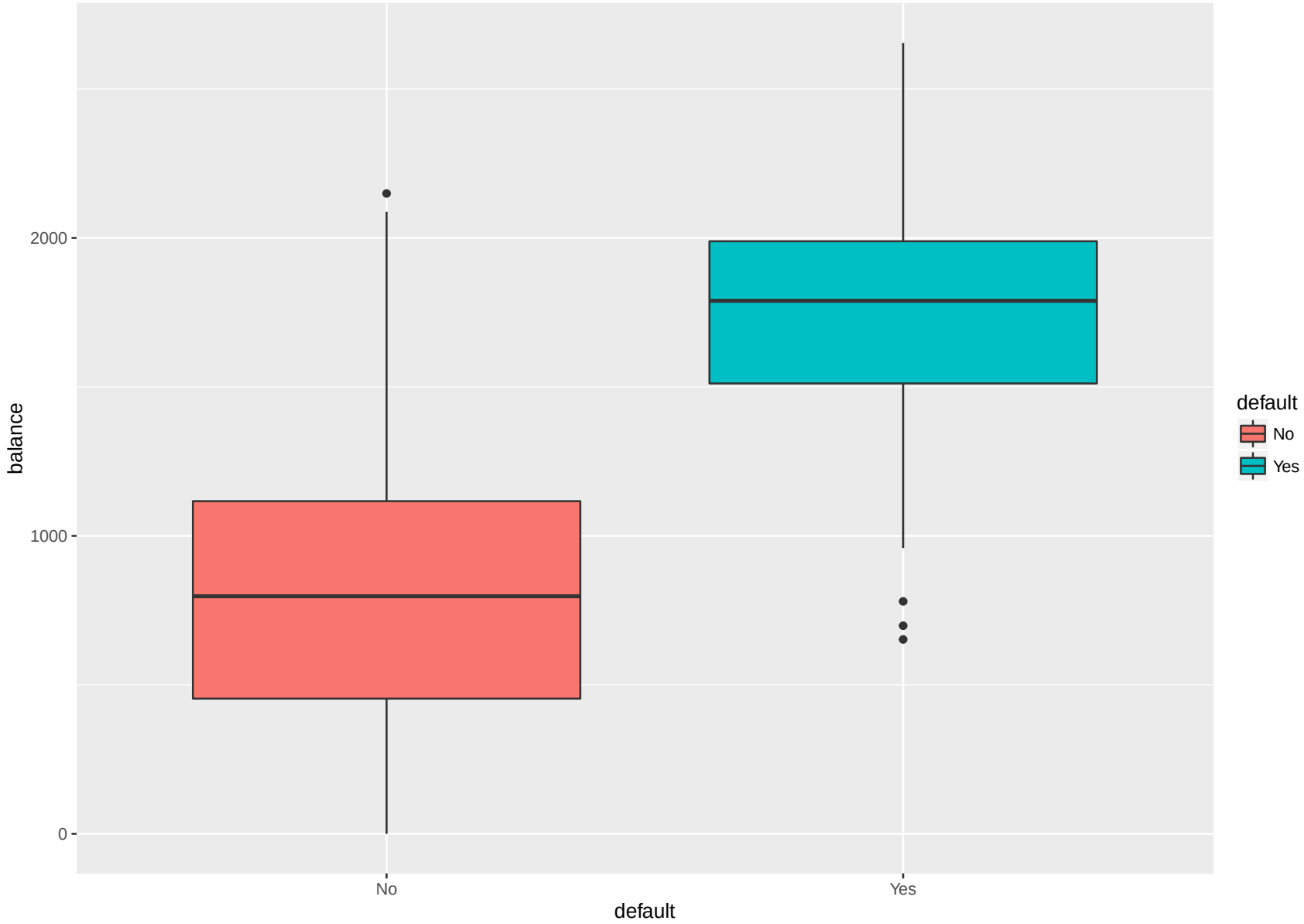


# pour les variables balance et income

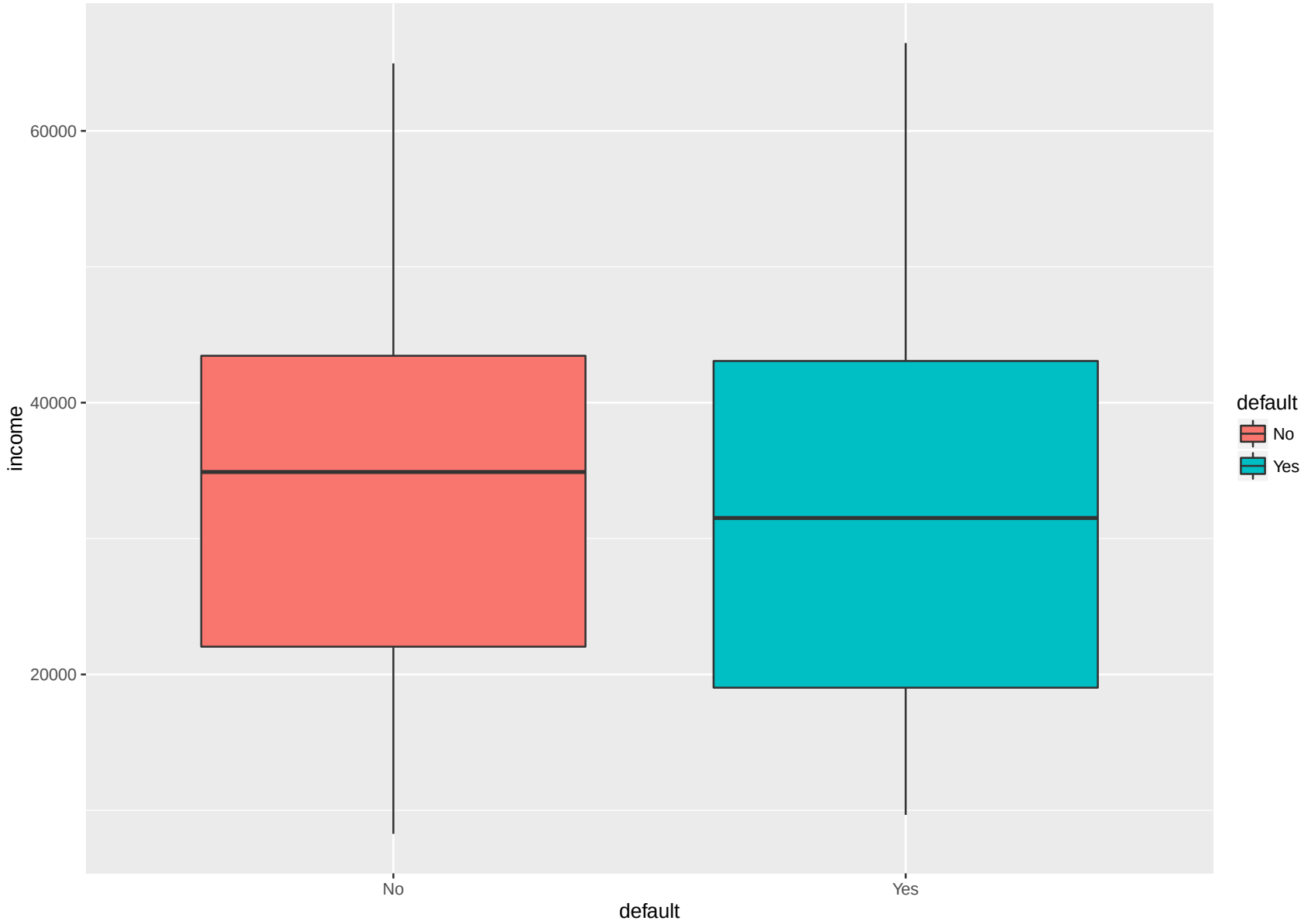
```
ggplot(Default, aes(balance, income))   +
  geom_point(aes(colour = default))
```



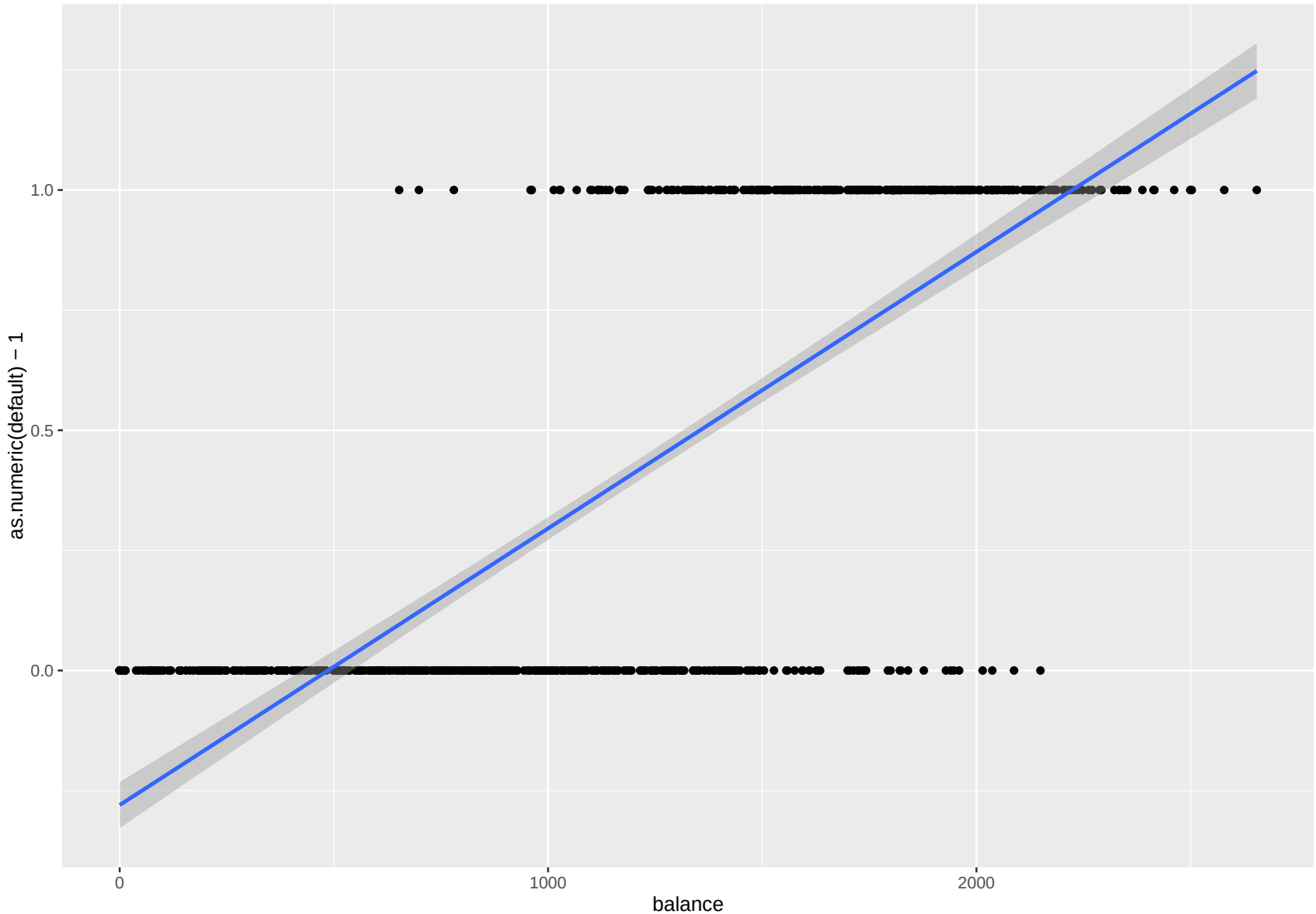
```
ggplot(Default, aes(default, balance)) +
  geom_boxplot(aes(fill = default))
```



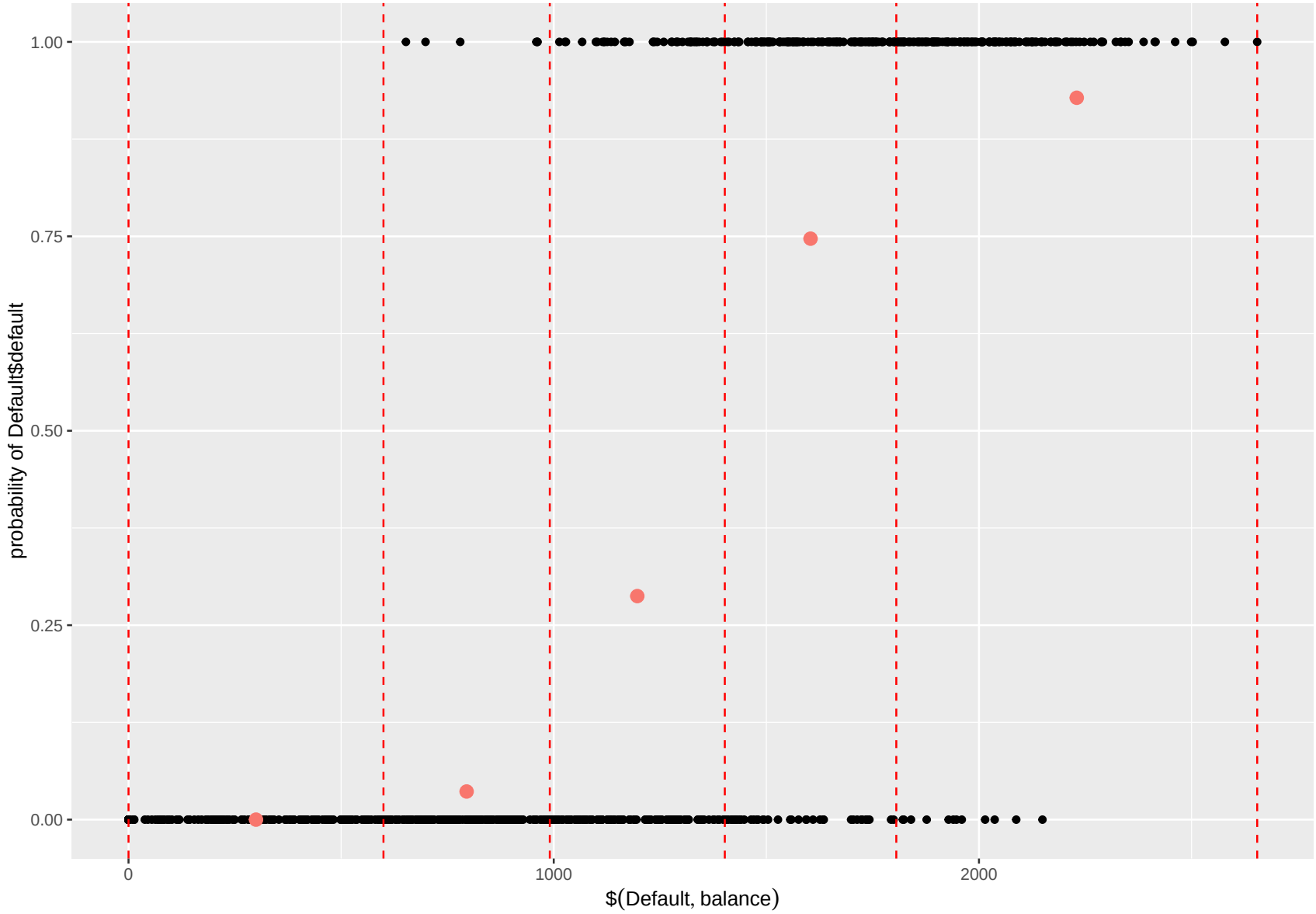
```
ggplot(Default, aes(default, income)) +
  geom_boxplot(aes(fill = default))
```



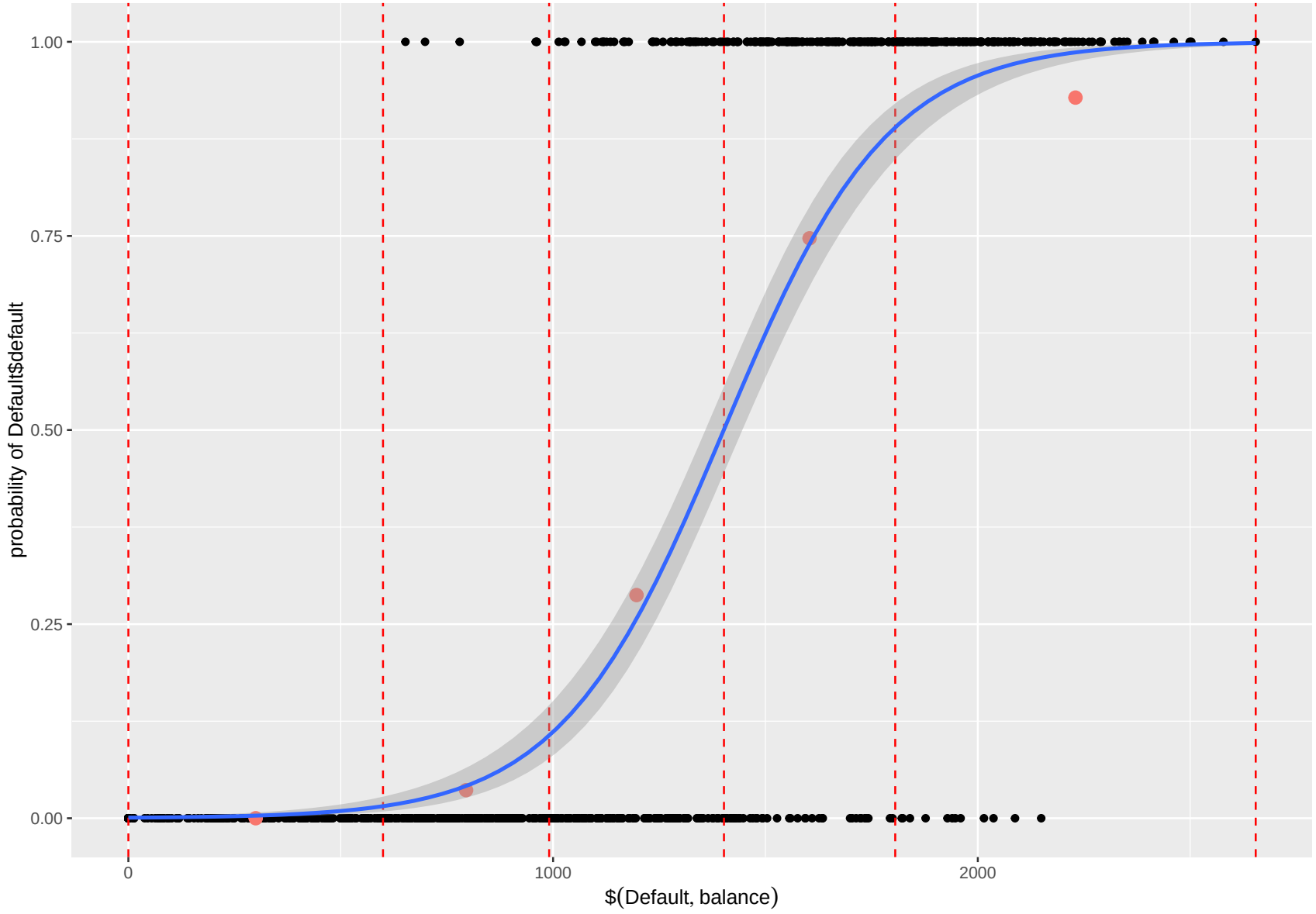
```
ggplot(Default, aes(balance, as.numeric(default)-1))   +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "gaussian"))
```



```
plt = illustration_logistique(Default,Default$balance,Default$default)
plt
```



```
plt = illustration_logistique(Default,Default$balance,Default$default, log_reg =
plt
```



"crab" : téléchargement
On veut expliquer le nombre de mâles "satellites" (`Sa`) autour de $n = 173$ crabes femelles par

- la couleur de la femelle (`C`)
- l'état de sa colonne ("spine condition") (`S`)
- le poids (`Wt`)
- la largeur de sa carapace (`W`)

# Les données “crab”

```
crab = read.table("crab.txt",header = TRUE)
glimpse(crab)
```

```
## Observations: 173
## Variables: 6
## $ Obs <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,...
## $ C   <int> 2, 3, 3, 4, 2, 1, 4, 2, 2, 2, 1, 3, 2, 2, 3, 3, 2, 2, 2, 2...
## $ S   <int> 3, 3, 3, 2, 3, 2, 3, 3, 1, 3, 1, 3, 1, 3, 3, 3, 3, 3, 3, 3...
## $ W   <dbl> 28.3, 26.0, 25.6, 21.0, 29.0, 25.0, 26.2, 24.9, 25.7, 27.5...
## $ Wt  <dbl> 3.05, 2.60, 2.15, 1.85, 3.00, 2.30, 1.30, 2.10, 2.00, 3.15...
## $ Sa  <int> 8, 4, 0, 0, 1, 3, 0, 0, 8, 6, 5, 4, 3, 4, 3, 5, 8, 3, 6, 4...
```

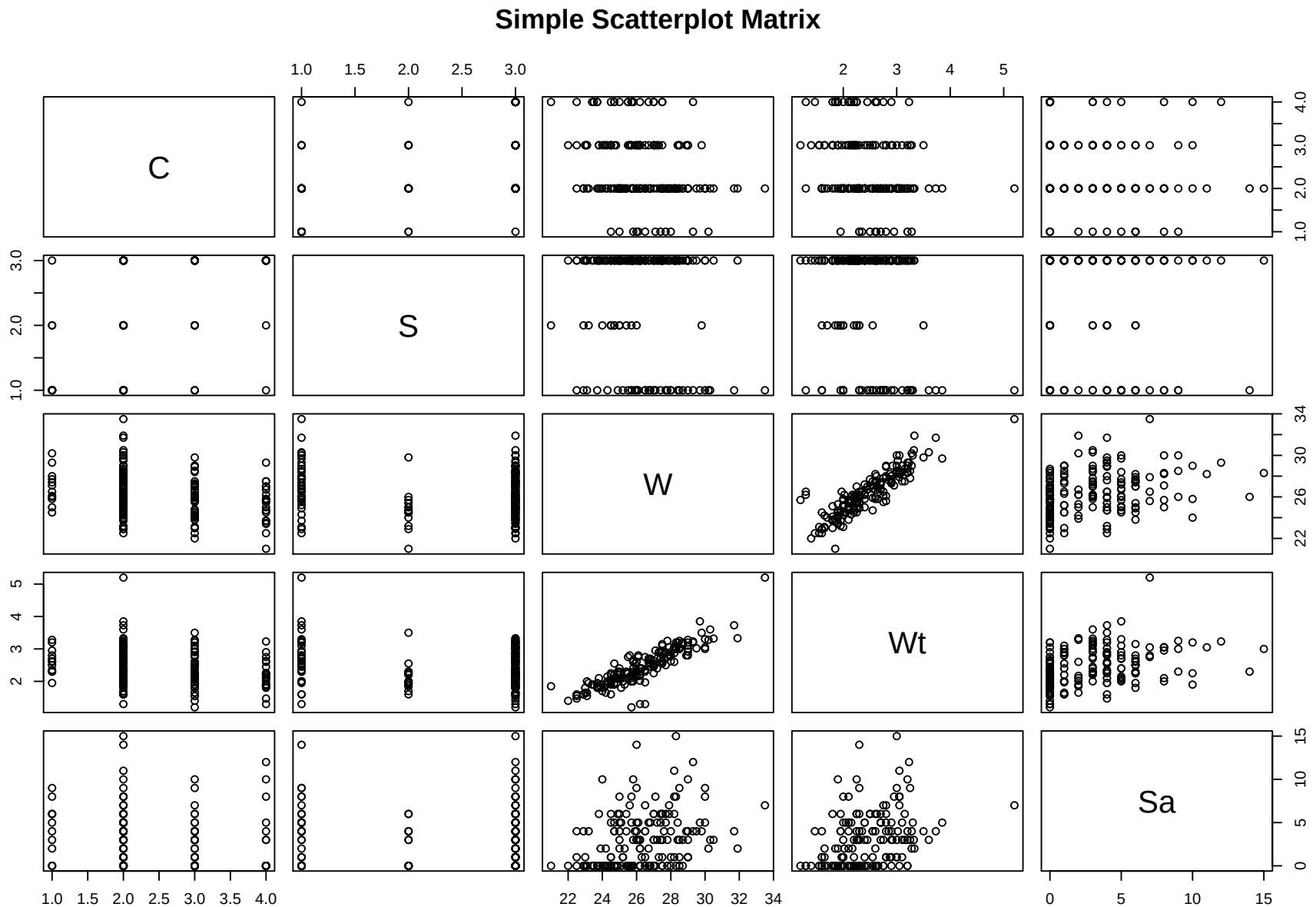
```
crab = mutate(crab , C = factor(C))
crab = mutate(crab , S = factor(S))
glimpse(crab)
```

```
## Observations: 173
## Variables: 6
## $ Obs <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,...
## $ C   <fctr> 2, 3, 3, 4, 2, 1, 4, 2, 2, 2, 1, 3, 2, 2, 3, 3, 2, 2, 2, ...
## $ S   <fctr> 3, 3, 3, 2, 3, 2, 3, 3, 1, 3, 1, 3, 1, 3, 3, 3, 3, 3, 3, ...
## $ W   <dbl> 28.3, 26.0, 25.6, 21.0, 29.0, 25.0, 26.2, 24.9, 25.7, 27.5...
## $ Wt  <dbl> 3.05, 2.60, 2.15, 1.85, 3.00, 2.30, 1.30, 2.10, 2.00, 3.15...
## $ Sa  <int> 8, 4, 0, 0, 1, 3, 0, 0, 8, 6, 5, 4, 3, 4, 3, 5, 8, 3, 6, 4...
```

```
attach(crab)
```

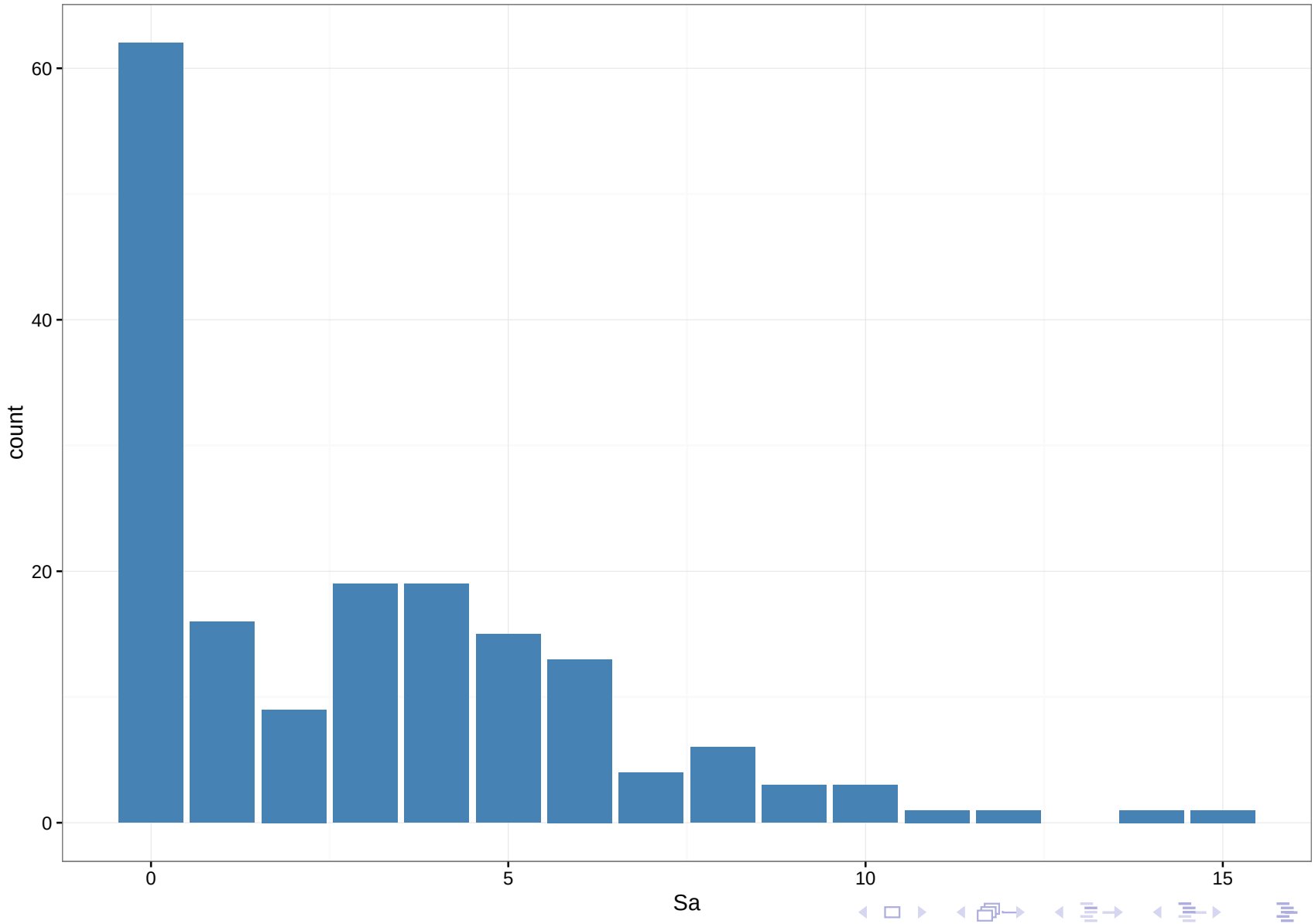
# scatterplot

```
pairs(crab[-1],  main="Simple Scatterplot Matrix")
```



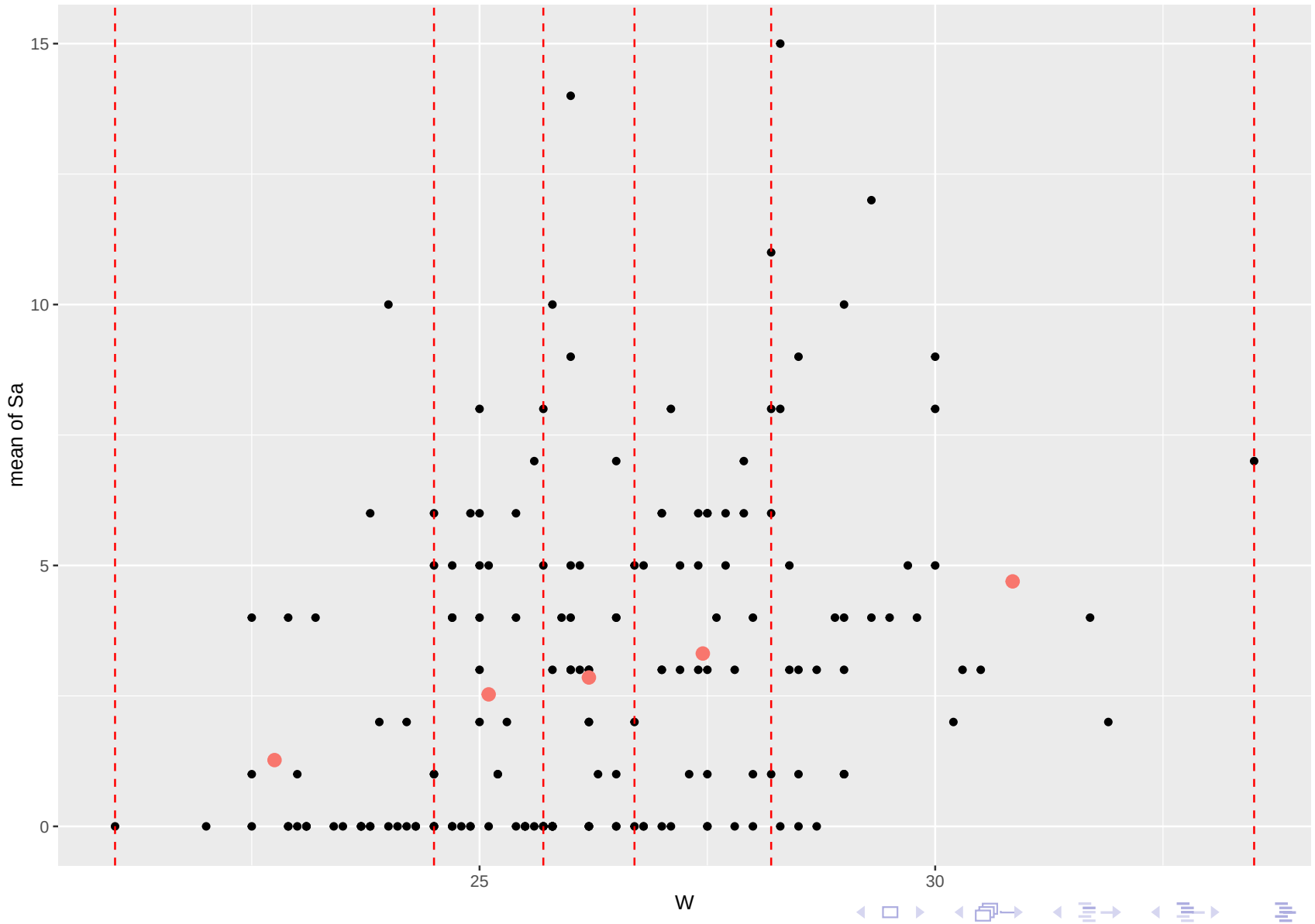
# histogram

```
ggplot(crab, aes(x=Sa))  +
  geom_bar( fill = "steelblue") + theme_bw()
```

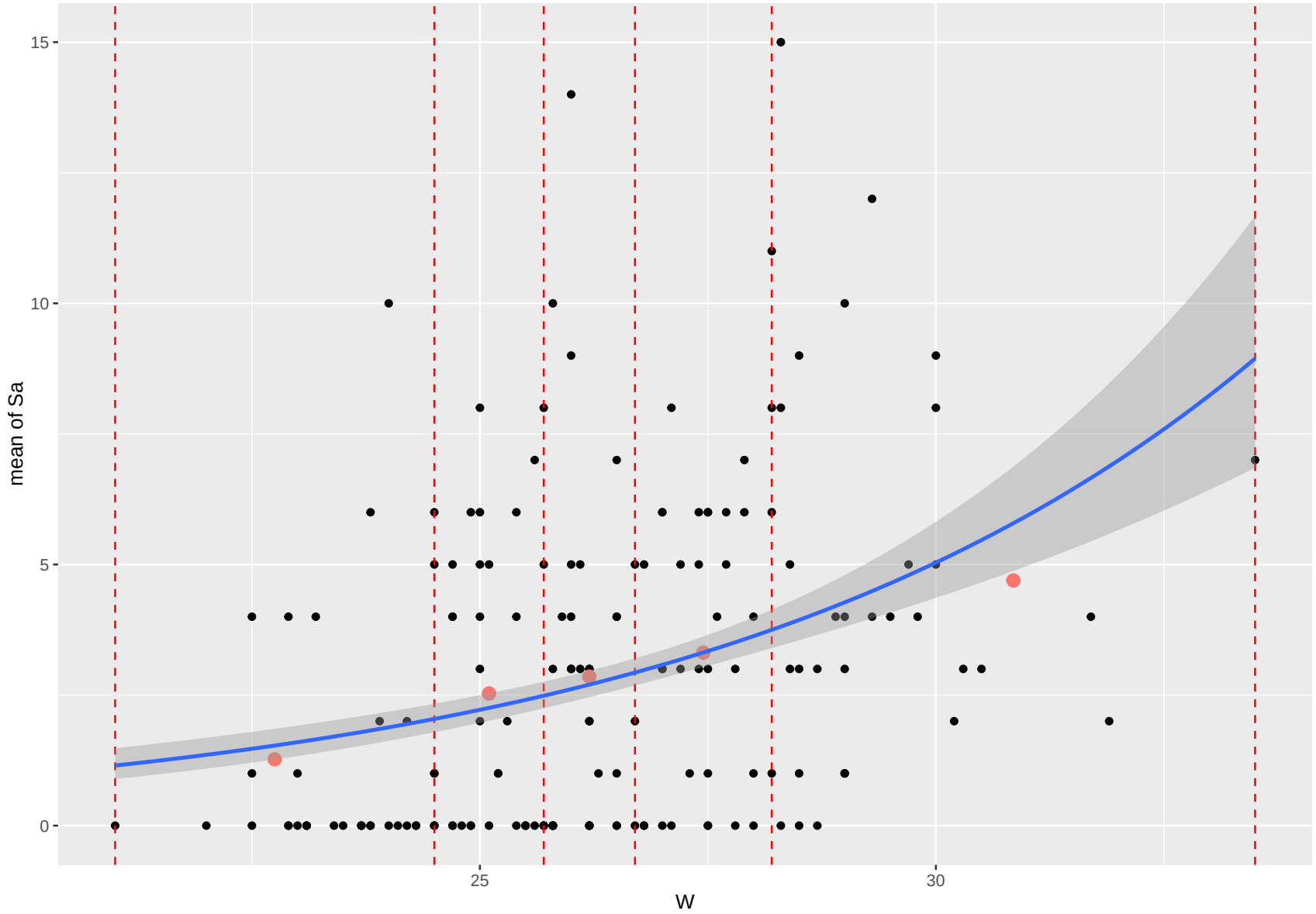


# pour la variable W

```
plt = illustration_poisson(crab,W,Sa)
plt
```



```
plt = illustration_poisson(crab,W,Sa, log_pois = TRUE)
plt
```



# Modèles de régression

- Dans le cas des données `Default`, $Y_i =$`Yes` ou `No` suit une loi de Bernoulli
- dans le cas des données `crab`, $Y_i \in \mathbb{N}$, on pense à la loi de Poisson

dans les deux cas, on veut lier $\mathbb{E}(Y_i)$ aux covariables $X_i$.

# Modèle linéaire gaussien

Dans le modèle $Y = X\beta + \varepsilon$ avec $\varepsilon$ de loi $\mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, on a

1. pour tout $i = 1, \ldots, n$, $\mathbb{E}(Y_i) = X_i\beta = \mathrm{Id}(X_i\beta)$
2. et la densité de $(Y_1, \ldots, Y_n)$ est

$$\begin{aligned} &\frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta)\right) \\ =&\frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right). \end{aligned}$$

Donc $-\log$ de la vraisemblance (aux points $\gamma$ et $\nu$) est

$$\log\left((\sqrt{2\pi\nu^2})^n\right) + \left(\frac{1}{2\nu^2}\|Y - X\gamma\|^2\right).$$

On définit l'estimateur de $\beta$ au maximum de vraisemblance (au minimum de $-\log$ de la vraisemblance) par

$$\hat{\beta} = \underset{\gamma \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y - X\gamma\|^2.$$

L'estimateur des moindes carrés $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ est aussi l'estimateur au maximum de vraisemblance.

3 Par ailleurs $-\log$ de la vraisemblance (aux points $\gamma$ et $\nu$) est

$$\|Y - X\gamma\|^2 = \sum_{i=1}^{n} (Y_i - X_i\gamma)^2 = \sum_{i=1}^{n} Y_i^2 - 2Y_iX_i\gamma + (X_i\gamma)^2.$$

# 3 points pour les modèles linéaires généralisés

1. pour tout $i = 1, \ldots, n$, $\mathbb{E}(Y_i) = g^{-1}(X_i\beta)$ ou $g(\mathbb{E}(Y_i)) = X_i\beta$
2. $\hat{\beta}$ est l'estimateur au maximum de vraisemblance
3. la log-vraisemblance est proportionnelle à

$$\sum_{i=1}^{n} Y_i X_i \gamma + b(X_i\gamma) + c(Y_i).$$

# Modèle logistique

$Y_i =$`Yes` ou `No` suit une loi de Bernoulli $\mathcal{B}(1, \pi_i)$ et on veut

1. pour tout $i = 1, \ldots, n$, $\mathbb{E}(Y_i) = g^{-1}(X_i\beta)$ ou $g(\mathbb{E}(Y_i)) = X_i\beta$ donc

$$\pi_i = g^{-1}(X_i\beta)$$

2. $\hat{\beta}$ est l'estimateur au maximum de vraisemblance. On écrit la log-vraisemblance au point $\gamma$

$$\begin{aligned}
&\log\left(\prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}\right) \\
&= \sum_{i=1}^{n}\left\{Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right\} \\
&= \sum_{i=1}^{n}\left\{Y_i \log\left(\frac{g^{-1}(X_i\gamma)}{1-g^{-1}(X_i\gamma)}\right) + \log\left(1-g^{-1}(X_i\gamma)\right)\right\}
\end{aligned}$$

La log-vraisemblance doit

3. être proportionnelle à

$$\sum_{i=1}^{n} Y_i X_i \gamma + b(X_i \gamma) + c(Y_i)$$

donc on doit choisir $g$ pour que

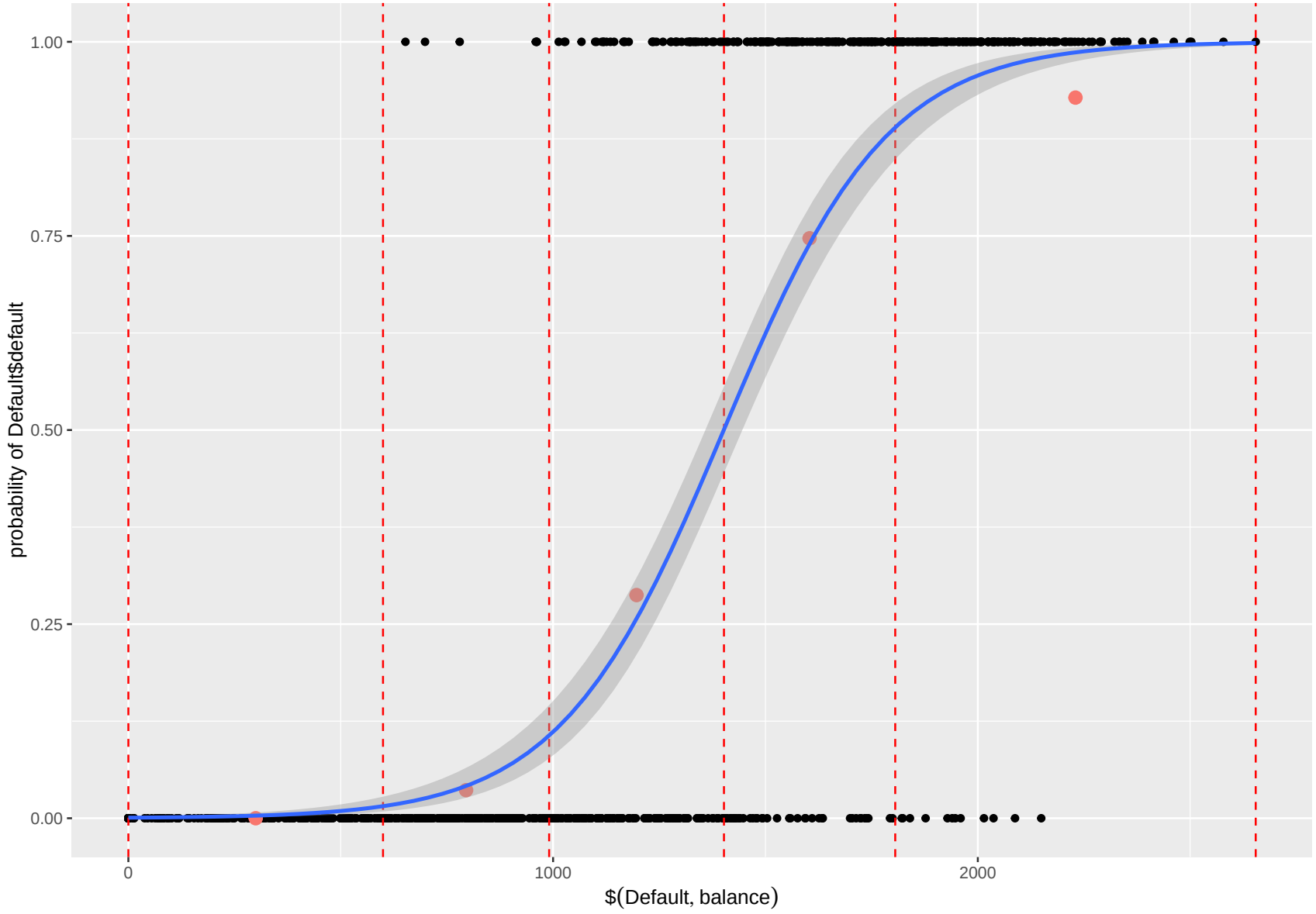
$$\log\left(\frac{g^{-1}(u)}{1-g^{-1}(u)}\right) = u.$$

Cela revient à choisir

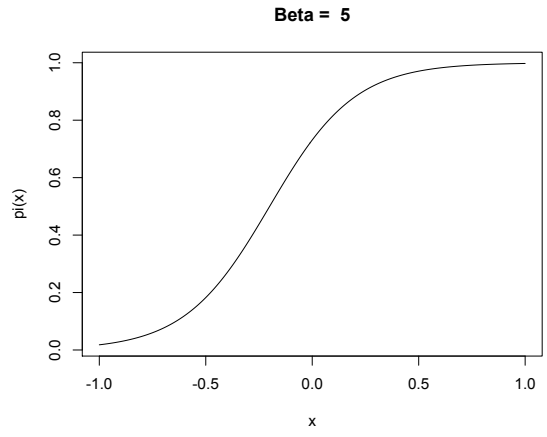
$$g^{-1}(u) = \frac{\exp(u)}{1+\exp(u)} \Leftrightarrow g(v) = \log(\frac{v}{1-v})$$

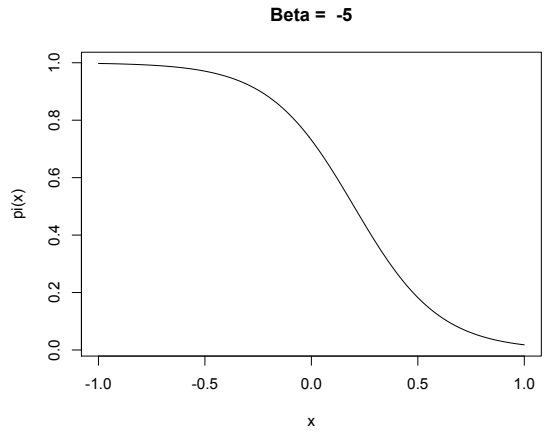
$g$ s'appelle la fonction logistique et $g^{-1}$ la fonction sigmoïde.

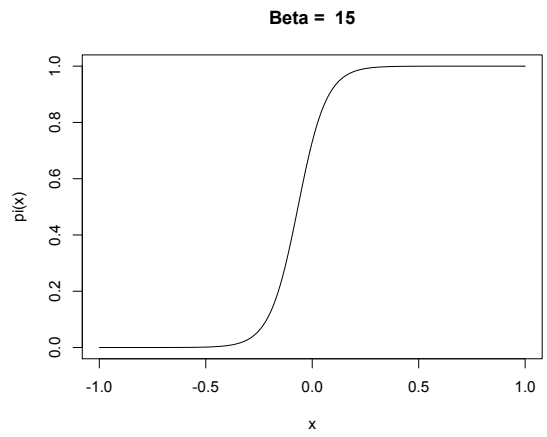
```
plt = illustration_logistique(Default,Default$balance,Default$default, log_reg =
plt
```

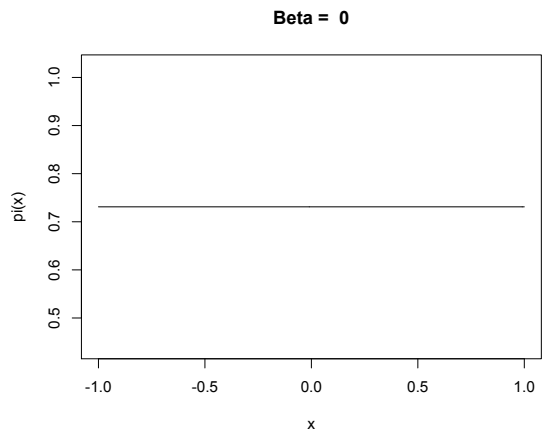


# Illustration



## Modèle logistique

On observe pour $i = 1, \ldots, n$

- des variables explicatives $X_i$ en dimension $p + 1$
- une variable $Y_i$ de loi de Bernoulli $\mathcal{B}\left(1, \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}\right)$.

On définit $\hat{\beta}$ comme l'estimateur au maximum de vraisemblance.

## Loi des $\hat{\beta}$ et tests de Wald

Asymptotiquement, pour toute covariable $j$

$$\frac{\hat{\beta}_j - \beta_j}{\hat{s}(\hat{\beta}_j)} \sim \mathcal{N}(0,1),$$

$\hat{s}(\hat{\beta}_j)$ est l'estimateur de l'écart-type de $\hat{\beta}_j$ . Donc pour tester $\mathcal{H}_0 : \beta_j = 0$, on utilise la statistique

$$\hat{s}(\hat{\beta}_j)^{-1}\hat{\beta}_j$$

que l'on compare aux fractiles de la loi $\mathcal{N}(0\,,1)$ .

# Estimation des coefficients

```
fit_logistic = glm(default ~ student + balance + income, family = "binomial",
                   data=Default)
summary(fit_logistic)
```

```
##
## Call:
## glm(formula = default ~ student + balance + income, family = "binomial",
##     data = Default)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.80908  -0.35761  -0.07953   0.40687   2.73280
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.725e+00  7.843e-01  -9.849   <2e-16 ***
## studentYes  -3.678e-01  3.958e-01  -0.929    0.353
## balance      5.330e-03  3.672e-04  14.516   <2e-16 ***
## income       1.165e-05  1.430e-05   0.815    0.415
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1121.08  on 832  degrees of freedom
## Residual deviance:  492.14  on 829  degrees of freedom
```

# Rapport de côtes ou odd-ratios

## Définition : odds ou côte

La quantité $\pi(X_i)/1-\pi(X_i)$ est appelé odds ou côte.

Dans le modèle logistique, on a définit l'odds (ou la côte) par

$$\frac{\pi(X_i)}{1-\pi(X_i)} = \exp(X_i\beta) = \exp(\beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p).$$

On considère deux individus $i_1$ et $i_2$ dont la valeur des covariables ne diffère que pour la $j$-ième covariable avec $X_{i_1}^j - X_{i_2}^j = 1$, on calcule l'odds-ratio (ou le rapport des côtes)

$$\frac{\pi(X_{i_1})}{1-\pi(X_{i_1})} \Big/ \frac{\pi(X_{i_2})}{1-\pi(X_{i_2})} = \exp(\beta_j)$$

On dira alors qu'une augmentation de $1$ de la variable $j$ entraîne une multiplication de l'odds ratio de $\exp(\beta_j)$ .

# Prédiction

On prédit en régression logistique en calculant pour un nouvel individu avec les covariables $X_+$

$$\hat{\pi}(X_+) = \frac{\exp(X_+\hat{\beta})}{\exp(X_+\hat{\beta}) + 1},$$

cela nous donne une valeur entre $0$ et $1$, si on a besoin de prédire $0$ ou $1$, on compare $\hat{\pi}(X_+)$ à $1/2$, si

$$\hat{\pi}(X_+) > 1/2,$$

on prédit $Y_+^P = 1$ et $0$ sinon.

## Intervalle de prédiction

On peut également définir un intervalle de confiance pour $\pi(X_i)$ au niveau $0\ .95$ par

$$\left[\frac{\exp(X_i\hat{\beta} - 1.96\hat{s})}{1 + \exp(X_i\hat{\beta} + 1.96\hat{s})}; \frac{\exp(X_i\hat{\beta} + 1.96\hat{s})}{1 + \exp(X_i\hat{\beta} - 1.96\hat{s})}\right]$$

où $\hat{s}$ est un estimateur de l'écart-type de $X_i\hat{\beta}$.

# Matrice de confusion

## Définitions : matrice de confusion

Pour chaque individu $i = 1, \ldots, n$ de notre échantillon, on note $Y_i^P$ la prédiction de $Y_i$, on peut construire une matrice de confusion

| | | Valeurs observées | |
|---|---|---|---|
| | | $Y_i = 0$ | $Y_i = 1$ |
| Valeurs prédites | $Y_i^P = 0$ | TN | FN |
| | $Y_i^P = 1$ | FP | TP |
| | total | N | P |

où P=POSITIVE, N=NEGATIVE, F=FALSE, T=TRUE.

On définit alors

- le **true positive rate ou sensibilité** comme TP/P
- le **false discovery rate** comme FP/(FP+TP)
- le **true negative rate ou spécificité** comme TN/N
- le **false positive rate** comme FP/(FP+TN)=FP/N = 1 - spécificité

# Prédictions

```
# Prédictions
predictions = predict(fit_logistic,type = "response")
predictions_01 = predict(fit_logistic,type = "response") > 1/2
# Matrice de confusion
table(predictions_01,default)
```

```
##                default
## predictions_01   No Yes
##          FALSE 451  52
##          TRUE   49 281
```

# Dans notre exemple

| | | Valeurs observées | |
|---|---|---|---|
| | | $Y_i = 0$ | $Y_i = 1$ |
| Valeurs prédites | $Y_i^P = 0$ | 451 | 52 |
| | $Y_i^P = 1$ | 49 | 281 |
| | total | 500 | 333 |

donc

- le **true positive rate ou sensibilité** vaut environ 0.84
- le **false positive rate** vaut environ 0.1
- le **false discovery rate** vaut environ 0.15
- le **true negative rate ou spécificité** vaut environ 0.9

# Courbe ROC

Pour construire les prédictions $(Y_i^P)$, nous avons pris un seuil $1/2$. Si, maintenant, nous faisons varier ce seuil, nous obtenons de nouvelles prédictions définies par

$$\text{si } \hat{\pi}(X_+) > s, \text{ on prédit } Y_+^{P,s} = 1 \text{ et } 0 \text{ sinon.}$$

Si $s = 0$

| | | 0bs. | |
|---|---|---|---|
| | | $Y_i = 0$ | $Y_i = 1$ |
| Pred. | $Y_i^P = 0$ | 0 | 0 |
| | $Y_i^P = 1$ | 500 | 333 |

donc la sensibilité vaut $1$ et la spécificité vaut $0$.

Si $s = 1$

| | | Obs. | |
|---|---|---|---|
| | | $Y_i = 0$ | $Y_i = 1$ |
| Pred. | $Y_i^P = 0$ | 500 | 333 |
| | $Y_i^P = 1$ | 0 | 0 |

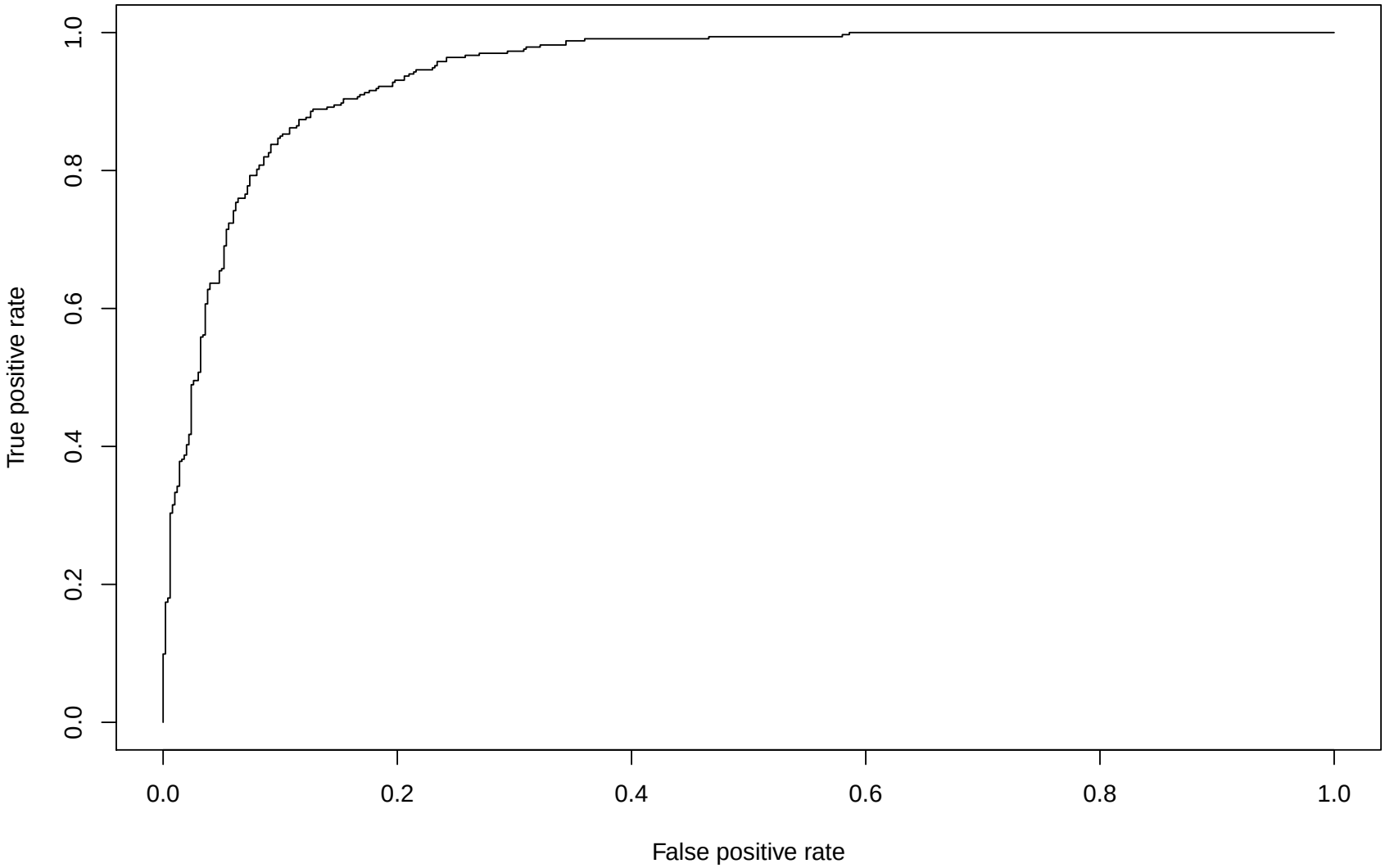
donc la sensibilité vaut $0$ et la spécificité vaut $1$.

## Définition : la courbe ROC et l'AUC

La courbe ROC (receiver operating characteristic) représente la sensibilité (qui vaut TP/P) contre $1-$ la spécificité (qui vaut FP/N) pour toutes les valeurs du seuil entre $0$ et $1$.
L'AUC (area under the ROC curve) est l'aire sous la courbe ROC.

Une courbe ROC idéale sera collée au coin supérieur gauche, donc plus l'AUC est grande meilleur est le classifieur. Une règle de classification au hasard aura un AUC d'environ 0.5.

```
plot( perf )
```


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

# AUC

```
ROC_auc <- performance( pred,"auc")
AUC <- ROC_auc@y.values[[1]]
print(AUC)
```

```
## [1] 0.943976
```

# La déviance

## Définition : les déviances

On définit la déviance résiduelle par

$$D(\hat{\beta}) = 2 \log \left( \mathcal{V}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) \right)$$

On définit la déviance nulle (null deviance) par

$$D^{\text{null}} = 2 \log \left( \mathcal{V}(\hat{\beta}_0^{\text{null}}) \right),$$

où $\hat{\beta}_0^{\text{null}}$ a été calculé dans le modèle sans covariable, donc avec l'intercept seul.

## Test de nullité simultanée des coefficients

Pour tester $\mathcal{H}_0 : \beta_1 = \ldots = \beta_p = 0$

- On utilise la statistique du rapport de vraisemblance $D(\hat{\beta}) - D^{\text{null}}$
- qui suit, sous $\mathcal{H}_0$, si $n$ est grand, une loi du $\chi^2(p)$
- on rejette $\mathcal{H}_0$ au niveau $\alpha$ si $D(\hat{\beta}) - D^{\text{null}} > z'$ où $z'$ est un fractile de la loi du $\chi^2(p)$.

C'est l'équivalent du test de Fisher.

# Test de nullité simultanée des coefficients

```
fit_null = glm(default ~ 1, family = "binomial", data=Default)
anova(fit_null,fit_logistic,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: default ~ 1
## Model 2: default ~ student + balance + income
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       832    1121.08
## 2       829     492.14  3   628.94 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Résidus (de déviance)

A nouveau par analogie avec le modèle linéaire gaussien, on construit les résidus de déviance, en identifiant

$$D(\hat{\beta}) = 2\log\left(\mathcal{V}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)\right) = \sum_{i=1}^{n} r_i^2.$$

On définit alors

$$r_i = sign(Y_i - \hat{Y}_i)\sqrt{r_i}.$$

## Cas de la régression logistique

En régression logistique, on obtient

$$\begin{aligned} r_i &= \sqrt{-2\log(\hat{\pi}(X_i)} \text{ si } Y_i = 1 \\ &= -\sqrt{-2\log(1-\hat{\pi}(X_i))} \text{ si } Y_i = 0. \end{aligned}$$

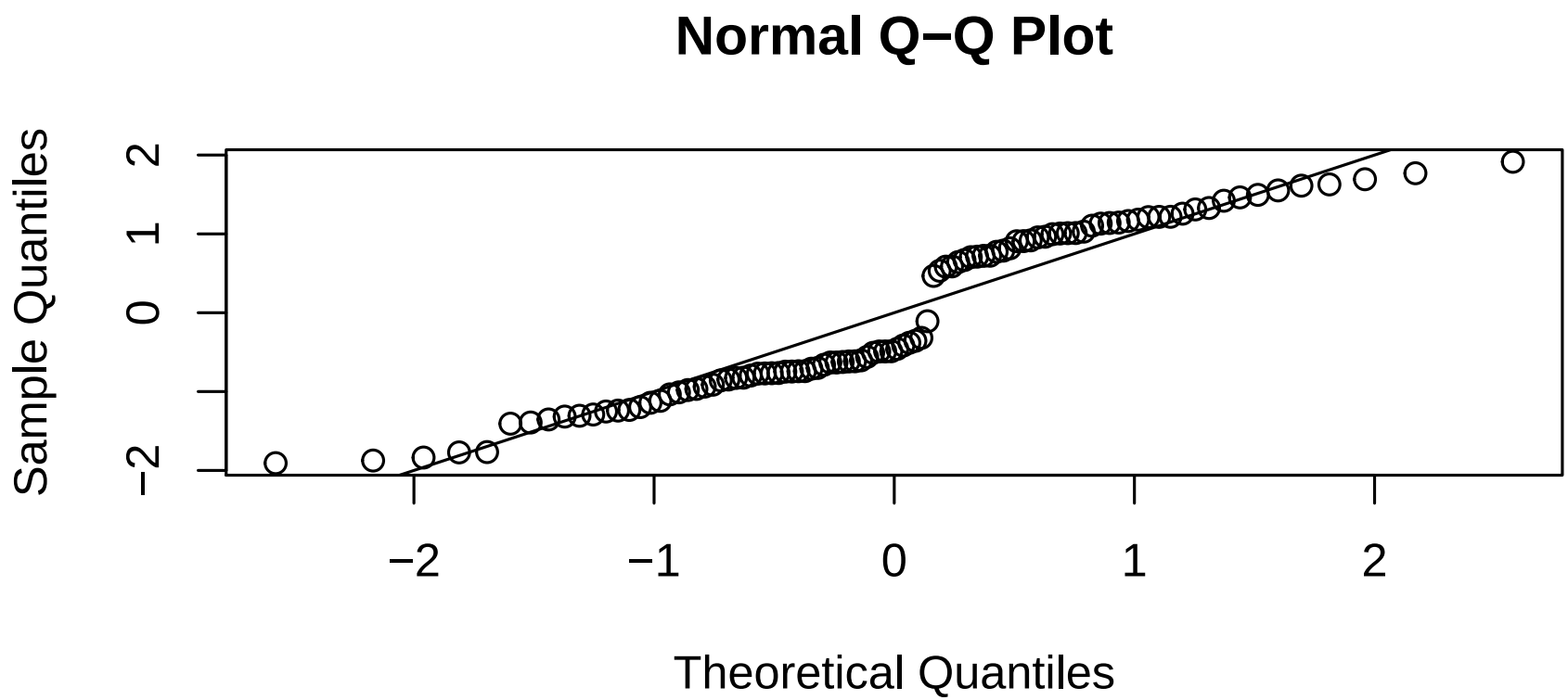
# Résidus de déviance / observations aberrantes

```
which(abs(resid(fit_logistic,type="deviance"))>2)
```

```
##  14  22 146 190 195 276 310 318 319 362 394 448 451 458 468 470 581 645
##  14  22 146 190 195 276 310 318 319 362 394 448 451 458 468 470 581 645
## 649 706 744 759 760 765 781
## 649 706 744 759 760 765 781
```

# Attention pas de diagnostic gaussien !

```
n = 100
x <-rnorm(n)
p<-exp(x)/(1+exp(x))
#simulate response
y <- rbinom(n,1,p)
#fit model
model = glm(y~x,family="binomial")
qqnorm(resid(model, type="deviance"))
abline(0,1)
```



# Leviers d'observations

```
influences = influence(fit_logistic)
hat = influences$hat
which(hat > 3*ncol(Default) /nrow(Default))
```

```
##  11  18  79 111 122 125 130 144 145 153 154 159 167 183 290 299 537 604
##  11  18  79 111 122 125 130 144 145 153 154 159 167 183 290 299 537 604
## 653 702
## 653 702
```

# Diagnostic sur X

```
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
vif(fit_logistic)
```

```
##  student  balance   income
## 2.664428 1.063226 2.643437
```

# Modèle poissonien

$Y_i \in \mathbb{N}$, on pense à la loi de Poisson $\mathcal{P}(\lambda_i)$ et on veut

1. pour tout $i = 1, \ldots, n$, $\mathbb{E}(Y_i) = g^{-1}(X_i\beta)$ ou $g(\mathbb{E}(Y_i)) = X_i\beta$ donc

$$\lambda_i = g^{-1}(X_i\beta)$$

2. $\hat{\beta}$ est l'estimateur au maximum de vraisemblance. On écrit la log-vraisemblance au point $\gamma$

$$\log\left(\prod_{i=1}^{n} e^{-\lambda_i}\frac{\lambda_i^{Y_i}}{Y_i!}\right) = \sum_{i=1}^{n}\{-\lambda_i + Y_i\log(\lambda_i) - log(Y_i!)\}$$

$$= \sum_{i=1}^{n}\left\{-g^{-1}(X_i\gamma) + Y_i\log\left(g^{-1}(X_i\gamma)\right) - log(Y_i!)\right\}$$

La log-vraisemblance doit

3. être proportionnelle à

$$\sum_{i=1}^{n} Y_i X_i \gamma + b(X_i \gamma) + c(Y_i)$$

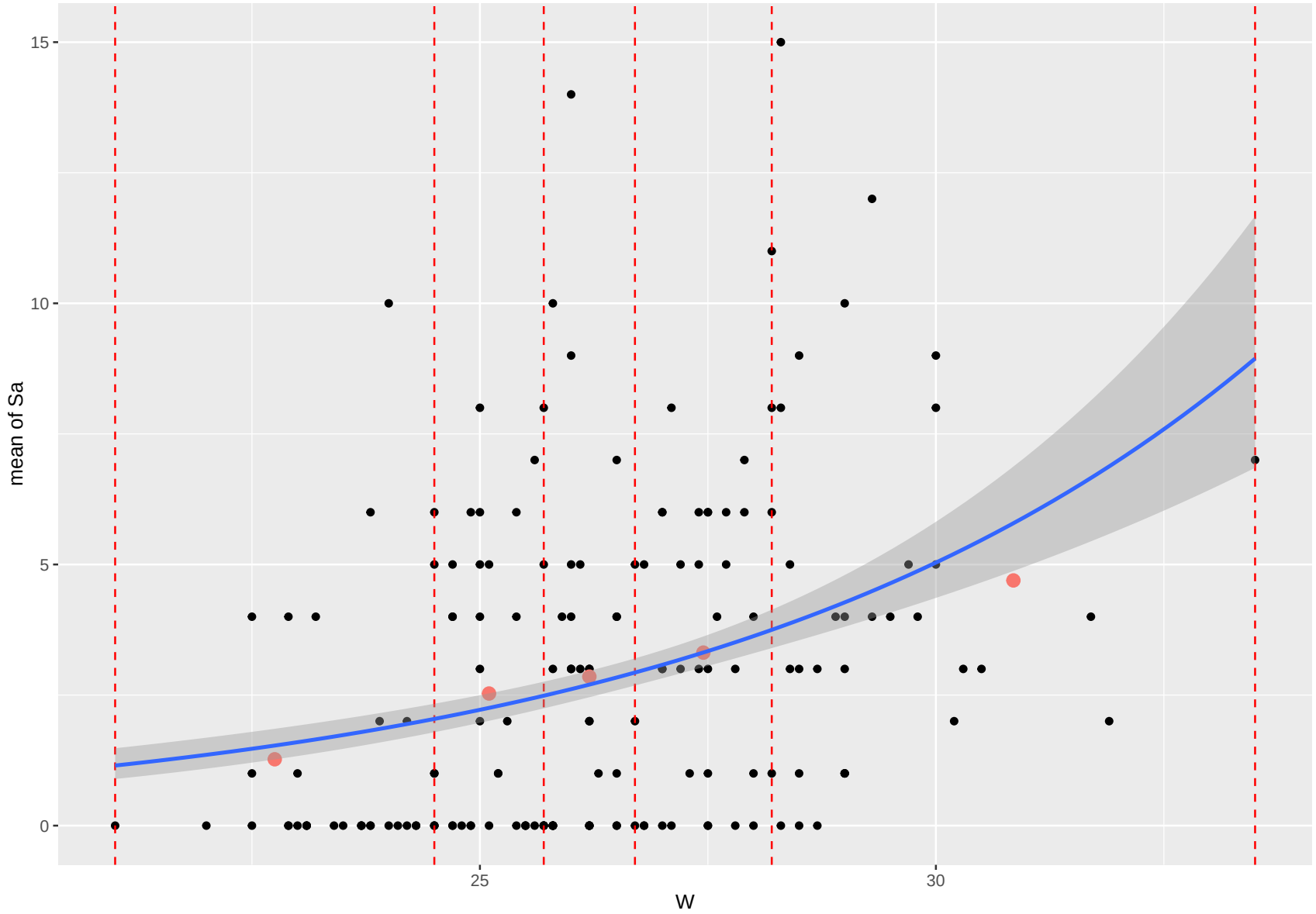
donc on doit choisir $g$ pour que

$$\log\left(g^{-1}(u)\right) = u.$$

Cela revient à choisir

$$g^{-1}(u) = \exp(u) \Leftrightarrow g(v) = \log(v).$$

```
plt = illustration_poisson(crab,W,Sa, log_pois = TRUE)
plt
```



## Modèle poissonnien

On observe pour $i = 1, \ldots, n$

- des variables explicatives $X_i$ en dimension $p + 1$
- une variable $Y_i$ de loi de Poisson $\mathcal{P}(\exp(X_i\beta))$.

On définit $\hat{\beta}$ comme l'estimateur au maximum de vraisemblance.

## Loi des $\hat{\beta}$ et tests de Wald

Asymptotiquement, pour toute covariable $j$

$$\frac{\hat{\beta}_j - \beta_j}{\hat{s}(\hat{\beta}_j)} \sim \mathcal{N}(0,1),$$

$\hat{s}(\hat{\beta}_j)$ est l'estimateur de l'écart-type de $\hat{\beta}_j$ . Donc pour tester $\mathcal{H}_0 : \beta_j = 0$, on utilise la statistique

$$\hat{s}(\hat{\beta}_j)^{-1}\hat{\beta}_j$$

que l'on compare aux fractiles de la loi $\mathcal{N}(0\,,1)$ .

# Estimation des coefficients

```
glimpse(crab)
```

```
## Observations: 173
## Variables: 6
## $ Obs <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,...
## $ C   <fctr> 2, 3, 3, 4, 2, 1, 4, 2, 2, 2, 1, 3, 2, 2, 3, 3, 2, 2, 2, ...
## $ S   <fctr> 3, 3, 3, 2, 3, 2, 3, 3, 1, 3, 1, 3, 1, 3, 3, 3, 3, 3, 3, ...
## $ W   <dbl> 28.3, 26.0, 25.6, 21.0, 29.0, 25.0, 26.2, 24.9, 25.7, 27.5...
## $ Wt  <dbl> 3.05, 2.60, 2.15, 1.85, 3.00, 2.30, 1.30, 2.10, 2.00, 3.15...
## $ Sa  <int> 8, 4, 0, 0, 1, 3, 0, 0, 8, 6, 5, 4, 3, 4, 3, 5, 8, 3, 6, 4...
```

```
fit_poisson = glm(Sa ~ C + S + W + Wt, family = "poisson",
                  data=crab)
summary(fit_poisson)
```

```
##
## Call:
## glm(formula = Sa ~ C + S + W + Wt, family = "poisson", data = crab)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0291  -1.8632  -0.5991   0.9331   4.9449
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.35722    0.96700  -0.369  0.71182
## C2          -0.26491    0.16811  -1.576  0.11507
```

# Taux relatifs

On considère deux individus $i_1$ et $i_2$ dont la valeur des covariables ne diffère que pour la $j$-ième covariable avec $X^j_{i_1} - X^j_{i_2} = 1$, on calcule alors les espérances

$$\mathbb{E}(Y_{i_1}) = exp(X_{i_1}\beta)$$
$$\mathbb{E}(Y_{i_2}) = exp(X_{i_2}\beta)$$

et leur rapport

$$\frac{\mathbb{E}(Y_{i_1})}{\mathbb{E}(Y_{i_2})} = exp(\beta_j)$$

ainsi $\exp(\beta_j)$ est la valeur par laquelle est multipliée l'espérance de la variable à expliquée quand $X^j$ augmente d'une unité, on l'appelle taux relatif.

# Prédiction

On prédit en régression poissonnienne en calculant pour un nouvel individu avec les covariables $X_+$

$$\hat{Y}_+^P = \hat{\lambda}(X_+) = \exp(X_+\hat{\beta}).$$

## Intervalle de prédiction

On peut également définir un intervalle de confiance pour $\pi(X_i)$ au niveau $0\ .95$ par

$$\left[\exp(X_i\hat{\beta} - 1.96\hat{s}); \exp(X_i\hat{\beta} + 1.96\hat{s})\right]$$

où $\hat{s}$ est un estimateur de l'écart-type de $X_i\hat{\beta}$.

# Prédictions

```
# Prédictions
predictions = predict(fit_poisson,type = "response")
```

## Test de nullité simultanée des coefficients

Pour tester $\mathcal{H}_0 : \beta_1 = \ldots = \beta_p = 0$

- On utilise la statistique du rapport de vraisemblance $D^0(\hat{\beta}) = -2\log(\mathcal{L}^{\text{null}}) + 2\log(\mathcal{L}(\hat{\beta}))$ où $\mathcal{L}$ est la log-vraisemblance du modèle
- qui suit, sous $\mathcal{H}_0$, si $n$ est grand, une loi du $\chi^2(p)$
- on rejette $\mathcal{H}_0$ au niveau $\alpha$ si $D^0(\hat{\beta}) > z'$ où $z'$ est un fractile de la loi du $\chi^2(p)$.

C'est l'équivalent du test de Fisher.

# Test de nullité simultanée des coefficients

```
fit_null = glm(Sa ~ 1, family = "poisson", data=crab)
anova(fit_null,fit_poisson,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Sa ~ 1
## Model 2: Sa ~ C + S + W + Wt
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       172     632.79
## 2       165     549.56  7   83.228 3.021e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Résidus (de déviance)

A nouveau par analogie avec le modèle linéaire gaussien, on construit les résidus de déviance, en identifiant

$$D^{\text{sat}} - D(\hat{\beta}) = 2\Big(\sum_{i=1}^{n} -Y_i + Y_i \log(Y_i)\Big) - 2\log\left(\mathcal{V}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)\right) = \sum_{i=1}^{n} r_i^2.$$

On définit alors

$$r_i = sign(Y_i - \hat{Y}_i)\sqrt{r_i}.$$

## Cas de la régression poissonnienne

On obtient

$$r_i = sign(Y_i - \exp(X_i\hat{\beta}))\sqrt{2}\left(\exp(X_i\hat{\beta}) - Y_i - Y_i\big(X_i\hat{\beta} - \log(Y_i)\big)\right)$$

# Résidus de déviance / observations aberrantes

```
which(abs(resid(fit_poisson,type="deviance"))>2)
```

```
##   3   8   9  21  24  32  34  37  42  44  46  48  49  59  61  62  66  75
##   3   8   9  21  24  32  34  37  42  44  46  48  49  59  61  62  66  75
##  79  87  88  90  95 101 102 104 109 110 120 121 123 125 130 132 133 137
##  79  87  88  90  95 101 102 104 109 110 120 121 123 125 130 132 133 137
## 139 141 145 151 153 154 159 160 166 167 168 169
## 139 141 145 151 153 154 159 160 166 167 168 169
```

# References I


Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An introduction to statistical learning*, vol. 6, Springer, 2013.