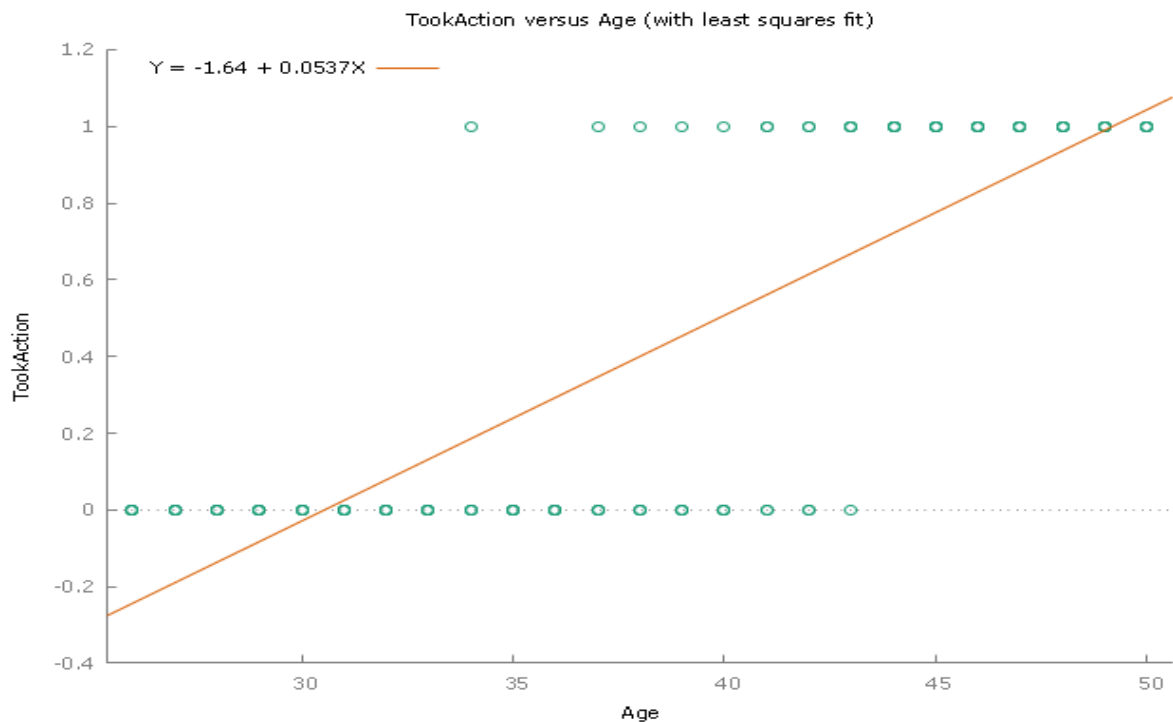
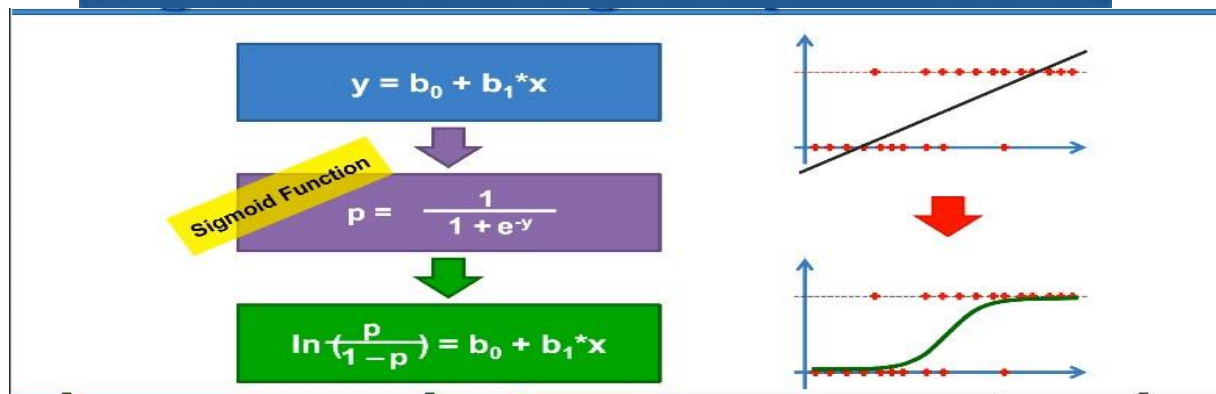
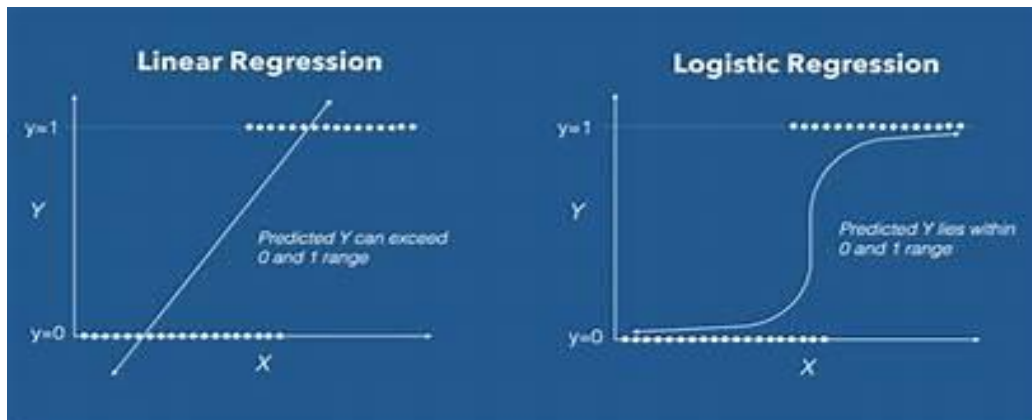


Logistic Regression



Régression linéaire SIMPLE de TookAction sur Age

On remarque que les points d'observation sont très loin des valeurs prédites par la droite. Alors La droite de régression linéaire n'est pas le modèle adéquate pour prédire l'action TookAction avec Age.



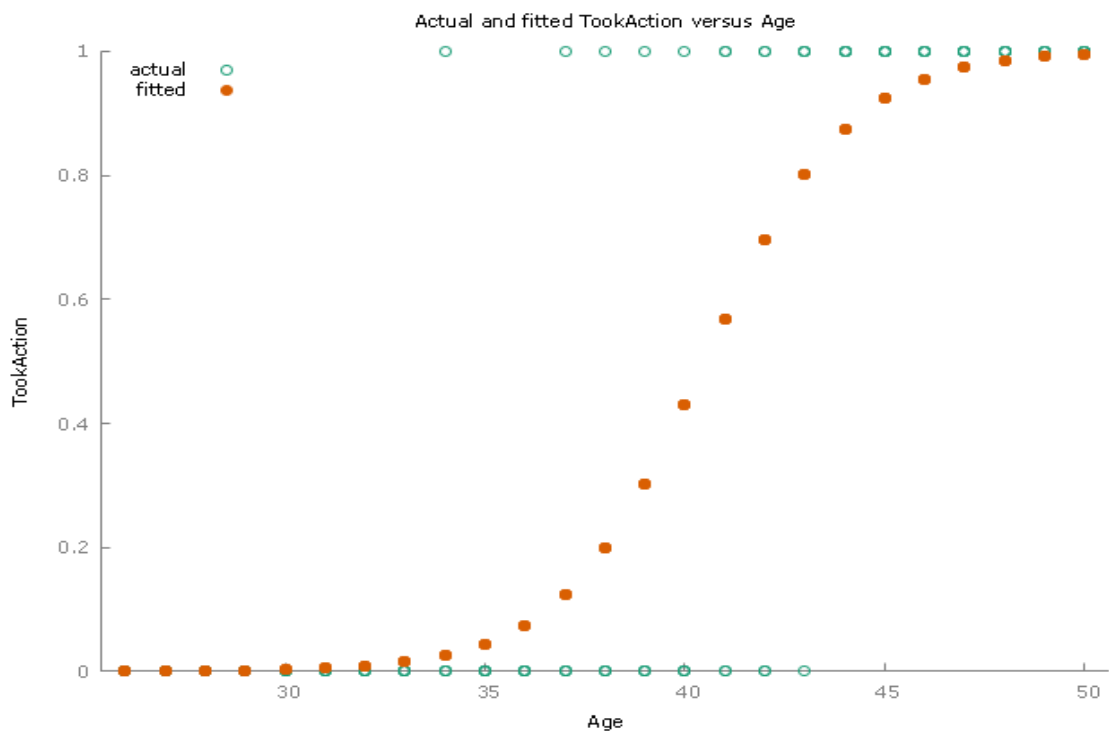
$$\ln \left(\frac{p}{1-p} \right) = b_0 + b_1 \cdot x$$

$$\ln \left(\frac{p}{1-p} \right) = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_n \cdot x_n$$

12

Avec p : la probabilité que l'évènement se réalise.

Logistic Regression of TookAction sur Age



Model 1: Logit, using observations 1-100

Dependent variable: TookAction

Standard errors based on Hessian

	Coefficient	Std. Error	z	p-value	
const	-22.5768	4.79476	-4.709	<0.0001	***
Age	0.557366	0.118128	4.718	<0.0001	***
Mean dependent var	0.400000	S.D. dependent var		0.492366	
McFadden R-squared	0.652483	Adjusted R-squared		0.622766	
Log-likelihood	-23.38827	Akaike criterion		50.77655	
Schwarz criterion	55.98689	Hannan-Quinn		52.88526	

Number of cases 'correctly predicted' = 90 (90.0%)

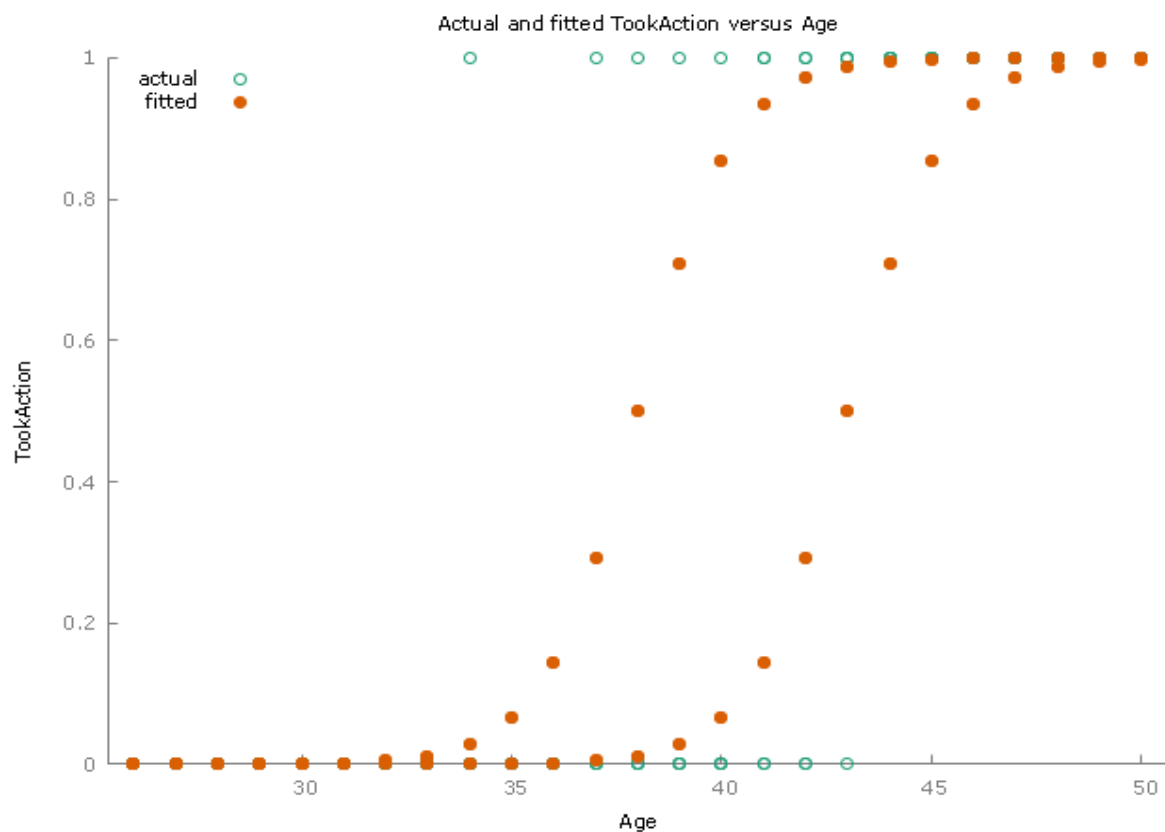
f(beta'x) at mean of independent vars = 0.492

Likelihood ratio test: Chi-square(1) = 87.8258 [0.0000]

	Predicted	
	0	1
Actual	0	55
	1	5
	55	35

Logistic Regression of TookAction sur Age et Gender(Male)

On a dummify la variable Gender en gardant Male.



Model 2: Logit, using observations 1-100					
Dependent variable: TookAction					
Standard errors based on Hessian					
	<i>Coefficient</i>	<i>Std. Error</i>	<i>z</i>	<i>p-value</i>	
const	-38.1520	9.98690	-3.820	0.0001	***
Age	0.887226	0.231787	3.828	0.0001	***
Male	4.43738	1.49193	2.974	0.0029	***
Mean dependent var					
		0.400000	S.D. dependent var		0.492366
McFadden R-squared		0.779993	Adjusted R-squared		0.735417
Log-likelihood		-14.80675	Akaike criterion		35.61349
Schwarz criterion		43.42901	Hannan-Quinn		38.77657

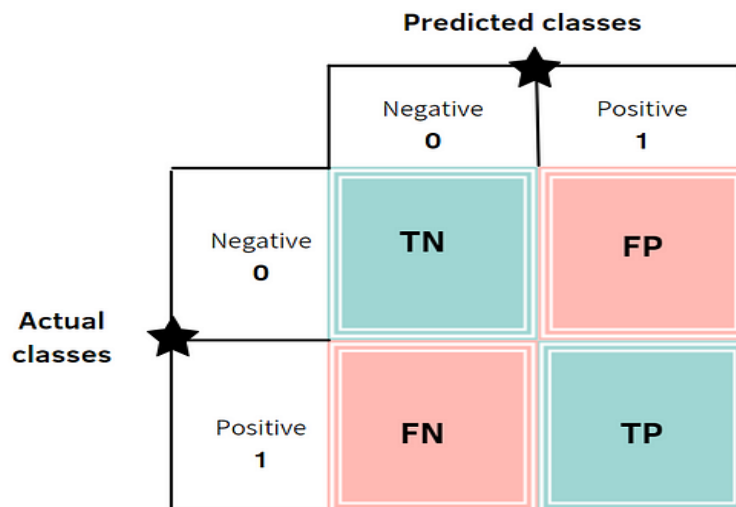
Number of cases 'correctly predicted' = 94 (94.0%)
f(beta'x) at mean of independent vars = 0.492

Likelihood ratio test: Chi-square(2) = 104.989 [0.0000]

Predicted	
0	1
Actual 0	58 2
1	4 36

Voir la table de prédiction dans tab_predict.pdf

- False Positive (FP : Erreur de 1ère espèce), c'est l'erreur commise par le modèle lorsqu'il prédit que qq va se passer alors qu'en réalité NON.
- False Negative (FN : Erreur de 2ème espèce), c'est l'erreur commise par le modèle lorsqu'il prédit que qq ne va pas se passer alors qu'en réalité OUI.
- Matrice de confusion



Interprétations des coefficients de la régression logistique

Prenons la Logistic Regression of TookAction sur Age et Gender(Female), on a:

Model 3: Logit, using observations 1-100					
Dependent variable: TookAction					
Standard errors based on Hessian					
	Coefficient	Std. Error	z	p-value	
const	-33.7146	8.83304	-3.817	0.0001	***
Age	0.887226	0.231787	3.828	0.0001	***
Female	-4.43738	1.49193	-2.974	0.0029	***

- Age > 0, i.e. si vous prenez 2 femmes, alors la femme la plus âgée a plus de chance de cliquer sur le lien de l'offre que la femme la moins âgée.
- Female < 0, i.e. si on maintient toutes les autres variables indépendantes constantes, i.e. si on prend 2 personnes du même âge, un homme et une femme du même âge alors la femme a moins de cliquer sur l'offre que l'homme.

Logistic Regression Coefficients

What you can and cannot do:

1. **You CAN** interpret coefficient signs: '+' contributes, '-' detracts
2. **You CANNOT** interpret magnitudes of the coefficients to quantify associations between the DV and IVs directly
3. **You CAN** still compare magnitudes of the coefficients to contrast the level of **per-unit** association of different IVs to the DV
4. **You CAN** use the Z-statistic to contrast the level of association of different IVs with the DV. Since the Z-statistic is standardized you don't need to worry about scale, however it is harder to interpret