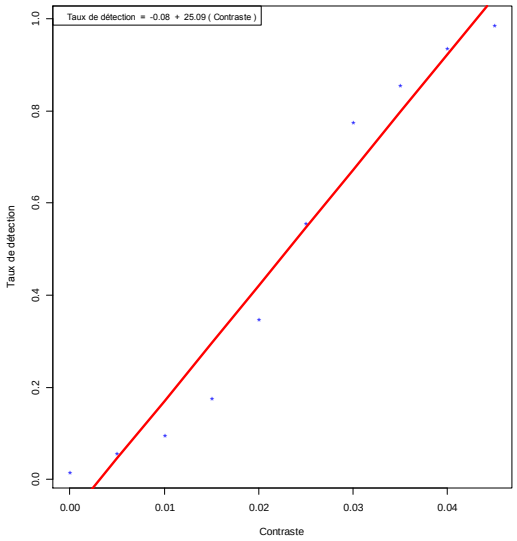
4.22 [Graybill, Franklin A. et Iyer, Hariharan K., *Regression Analysis, concepts and applications*, (1994), Duxbury Press] Dans le cadre d'une étude sur la perception visuelle, on soumet 10 sujets à l'expérience suivante. On présente à chaque sujet une image sur diapositive 100 fois. Le sujet doit déclarer à chaque fois s'il a perçu l'objet ou non. Les diapositives présentées aux dix sujets portaient la même image, mais le contraste optique entre l'objet et le fond variait. Le tableau ci-dessous montre les valeurs de *x*, une mesure du contraste optique, et *y*, la proportion des fois où l'objet a été perçu par le sujet.

y *Taux de détection de l'objet*	*x* *Contraste optique entre l'objet et le fond*	*y* *Taux de détection de l'objet*	*x* *Contraste optique entre l'objet et le fond*
0,02	0,000	0,56	0,025
0,06	0,005	0,78	0,030
0,10	0,010	0,86	0,035
0,18	0,015	0,94	0,040
0,35	0,020	0,99	0,045

a) Faites un graphique pour voir si une droite pourrait décrire la relation entre x et y.



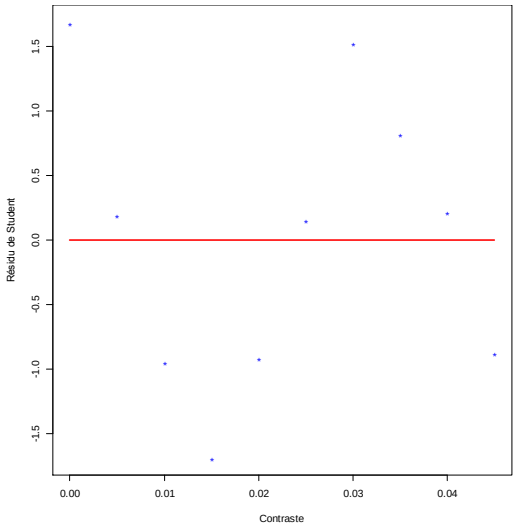
Bien qu'une droite puisse bien, éventuellement, fournir un modèle acceptable (le coefficient de corrélation est un impressionnant 0,98), il est évident que la relation n'est pas linéaire.

b) Déterminer une droite de régression de *y* sur *x*, puis examiner le graphique des résidus normés.

Données d'une régression simple:

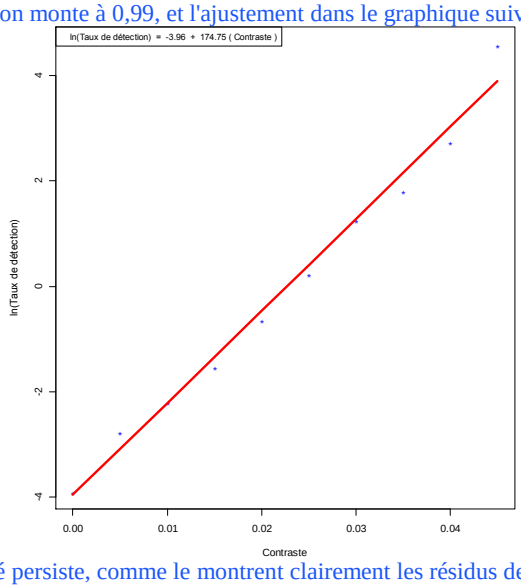
```
> summary(lm(y~x))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08055    0.04792  -1.681    0.131
x           25.09091    1.79537  13.975 6.66e-07 ***

Residual standard error: 0.08154 on 8 degrees of freedom
Multiple R-squared: 0.9607,     Adjusted R-squared: 0.9557
F-statistic: 195.3 on 1 and 8 DF,  p-value: 6.661e-07
```

Tous les calculs semblent dire que la relation est parfaitement linéaire. Sauf que le graphique suivant (résidus de Student) expose l'absence de linéarité plus clairement encore que le graphique ci-dessus.

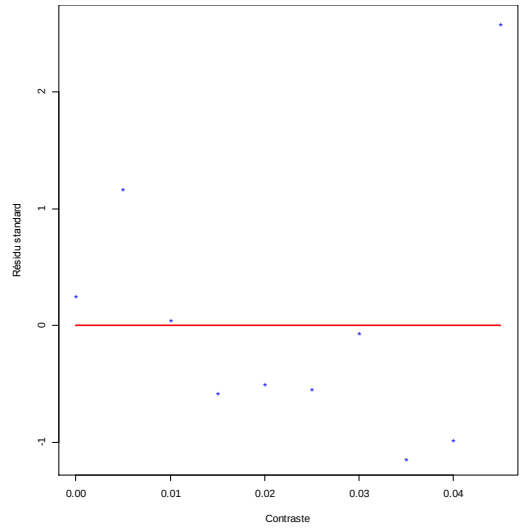


c) Considérez le modèle logistique : déterminez une régression de $z = \ln\left(\frac{y}{1-y}\right)$ sur x. Ce modèle semble-t-il meilleur ? [*La somme des carrés* $\sum_{i=1}^{n}(z_i - \hat{z}_i)^2$ *ne peut pas être comparée à la somme* $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ *issue du premier modèle. Le taux de détection estimé par ce deuxième modèle est* $\hat{\hat{y}}_i = \frac{e^{\hat{z}_i}}{1+e^{\hat{z}_i}}$. *On doit donc comparer* $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ *à* $\sum_{i=1}^{n}(y_i - \hat{\hat{y}}_i)^2$.]

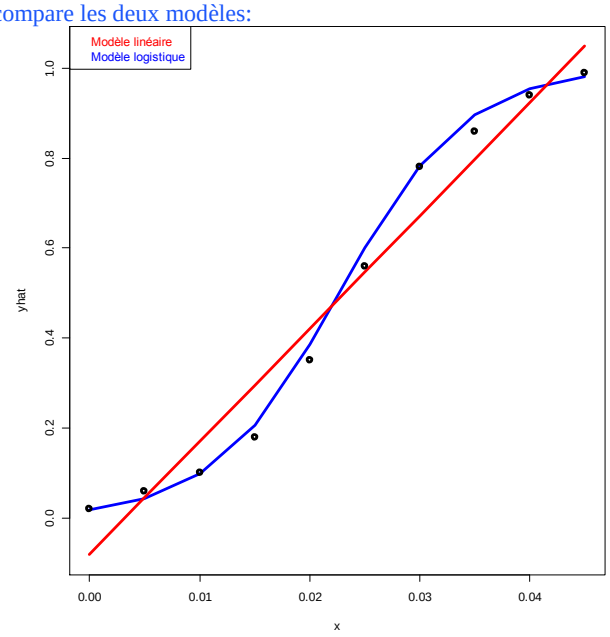
Le coefficient de corrélation monte à 0,99, et l'ajustement dans le graphique suivant semble meilleur.



Mais l'absence de linéarité persiste, comme le montrent clairement les résidus de Student, que voici :



Le graphique suivant compare les deux modèles:



$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0{,}0532$; $\sum_{i=1}^{n}(y_i - \hat{\hat{y}}_i)^2 = 0{,}0055$

Il est évident donc que la fonction logistique constitue un meilleur ajustement. Une absence lancinante de linéarité dérange un peu, mais on ne peut pas faire mieux pour le moment. Il faudrait qu'on dispose de plus de données.