

Modèles Linéaires

Correction TD 2 : Régression linéaire simple Licence 3 MIASHS

Guillaume Metzler, Francesco Amato, Alejandro Rivera

Institut de Communication (ICOM)

Université de Lyon, Université Lumière Lyon 2

Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr ; francesco.amato@univ-lyon2.fr

alejandro.rivera@univ-lyon2.fr

Résumé

Dans la précédente fiche, nous avons étudié trois façons différentes d'obtenir les solutions à notre problème de régression linéaire simple.

Dans la présente fiche, nous allons maintenant nous intéresser à la propriété de ces estimateurs et tester l'influence de la variable explicative sur la variable à expliquer.

Plus précisément :

- on étudiera le biais et la variance de la pente du modèle, donné par $\hat{\beta}_1$
- on étudiera le biais et la variance de l'ordonnée à l'origine du modèle $\hat{\beta}_0$
- on construira des intervalles de confiance sur ces derniers
- on testera la significativité de la pente du modèle.

Modèle de régression simple

On rappelle que le modèle linéaire gaussien simple s'écrit sous la forme

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

où Y est la variable à expliquer, X est la variable explicative et ε est une variable aléatoire représentant les erreurs du modèle que l'on supposera normalement distribuée avec une variance inconnue égale à σ^2

L'estimation des paramètres se fait à l'aide d'un échantillon $S = \{(y_i, x_i)\}_{i=1}^n$.

Dans ce TD, on se concentre sur l'étude des paramètres du modèle et nous appliquerons ensuite cela sur les données présentées ci-dessous, pour obtenir la droite de

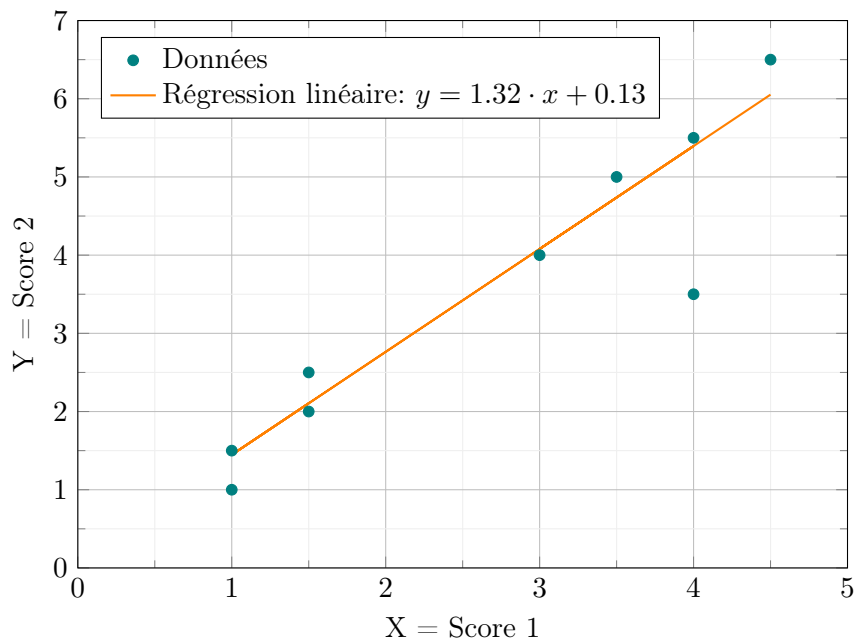


FIGURE 1 – Application de la régression linéaire simple gaussien sur les données présentées dans la table associée. On cherche alors à expliquer le score obtenu au deuxième examen en fonction du score obtenu au premier examen.

régression associée, présentée en Figure 1.

Y : Score examen 2	3.5	4	5	1	2	1.5	2.5	5.5	6	6.5
X : Score examen 1	4	3	3.5	1	1.5	1	1.5	4	3.5	4.5

On rappelle que les paramètres du modèle de régression sont données par les relations

$$\hat{\beta}_1 = \frac{\text{Cov}[X, Y]}{\text{Var}[X]} \quad \text{et} \quad \hat{\beta}_0 = \mathbb{E}[Y] - \hat{\beta}_1 \mathbb{E}[X].$$

On utilisera également le fait que la variance des erreurs σ^2 peut être estimée par

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Régression linéaire avec

Pour effectuer la régression linéaire avec , vous pouvez utiliser le code suivant

```
# Pour charger un jeu de données
data = read.csv("../data/reglin.csv", sep=";")
# Régression linéaire
mymodel = lm(Y~X,data)
```

Un résumé statistiques de la régression linéaire peut être obtenu à l'aide de la commande

```
summary(mymodel)
```

Notre objectif sera d'expliquer comment ces valeurs sont obtenues au fur et à mesure des séances.

On pourra extraire les coefficient de la régression à l'aide de la commande

```
coeff = mymodel$coefficients
```

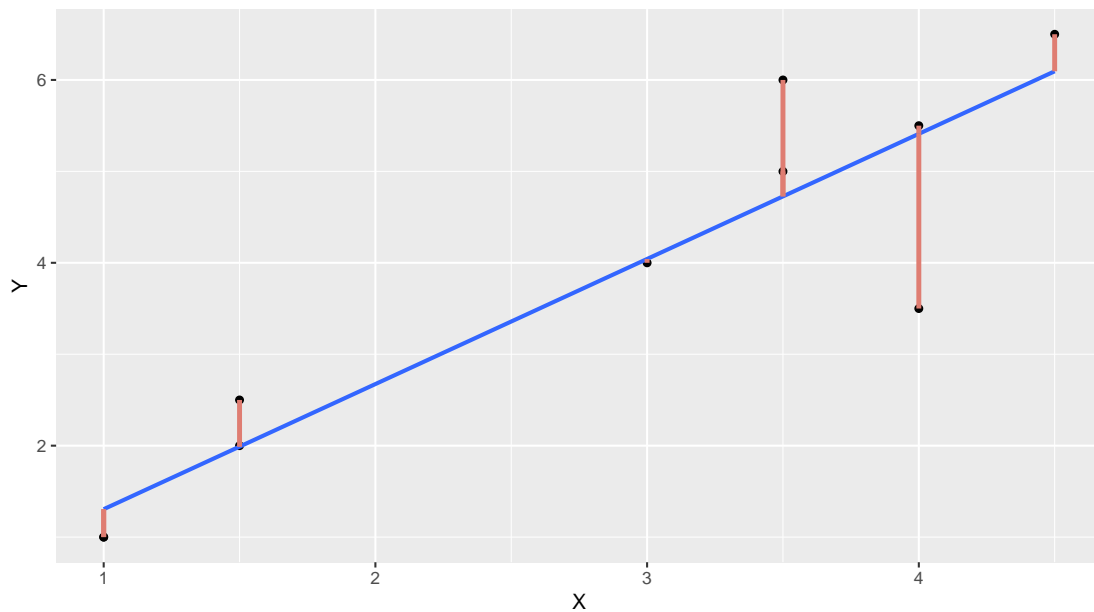
De la même façon, nous pourrions obtenir les résidus de la régression comme suit :

```
mymodel$residuals
```

On pourra enfin représenter nos données, la droite de régression, ainsi que les résidus graphiquement

```
# Graphical Representation of the Model and Residuals

library(ggplot2)
ggplot(data, aes(x=X, y=Y)) +
  geom_point() +
  geom_smooth(method=lm,se=FALSE) +
  geom_segment(aes(x = X, y = Y, xend = X,
                  yend = coeff[1] + coeff[2]*X,
                  col = "Residuals"),
              col = "#DF7D72", lwd= 1.2, data = data)
```



Etude des propriétés de la pente du modèle

On cherche à étudier le biais et la variance de cet estimateur.

1. Justifier que pour tout entier $i \in \llbracket 1, n \rrbracket$, nous avons

$$\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i.$$

On rappelle que pour tout entier $i \in \llbracket 1, n \rrbracket$, nous avons

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Or, on se rappelle que seule y_i et ε_i sont aléatoires dans cette expression. Ainsi

$$\begin{aligned} \mathbb{E}[y_i] &= \mathbb{E}[\beta_0 + \beta_1 x_i + \varepsilon_i], \\ &\quad \downarrow \text{linéarité de l'espérance} \\ &= \mathbb{E}[\beta_0] + \mathbb{E}[\beta_1] \mathbb{E}[x_i] + \mathbb{E}[\varepsilon_i], \\ &\quad \downarrow \text{toutes les quantités sont déterministes sauf } \varepsilon_i \text{ qui est de moyenne nulle} \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

2. En déduire la relation

$$\mathbb{E}[\bar{y}] = \beta_0 + \beta_1 \bar{x},$$

$$\text{où } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \text{ et } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

On procède comme dans la question précédente

$$\begin{aligned} \mathbb{E}[\bar{y}] &= \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n y_i \right], \\ &\downarrow \text{linéarité de l'espérance et question précédente} \\ &= \frac{1}{n} \sum_{i=1}^n \beta_0 + \beta_1 x_i, \\ &= \beta_0 + \beta_1 \bar{x}. \end{aligned}$$

3. A l'aide de l'expression de $\hat{\beta}_1$, montrer que ce dernier est un estimateur sans biais de β_1 .

On rappelle que l'on a

$$\begin{aligned} \hat{\beta}_1 &= \frac{\text{Cov}[X, Y]}{\text{Var}[X]} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \\ \mathbb{E}[\hat{\beta}_1] &= \frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}[y_i - \bar{y}]}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) (\mathbb{E}[y_i] - \mathbb{E}[\bar{y}])}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ &= \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ \mathbb{E}[\hat{\beta}_1] &= \beta_1. \end{aligned}$$

4. A l'aide des relations suivantes

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{et} \quad \bar{y} = \beta_0 + \beta_1 \bar{x}.$$

Montrer que l'on a

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

On repart de la définition de l'estimateur de β_1 .

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ &\quad \downarrow y_i - \bar{y} = \beta_1(x_i - \bar{x}) + \varepsilon_i \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_1(x_i - \bar{x}) + \varepsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ \hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.\end{aligned}$$

5. Dans la suite, on posera $\omega_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$. En déduire que la variance de l'estimateur $\hat{\beta}_1$ est égale à

$$\text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

$$\begin{aligned}\text{Var}[\hat{\beta}_1] &= \mathbb{E} \left[(\hat{\beta}_1 - \mathbb{E}[\hat{\beta}_1])^2 \right], \\ &= \mathbb{E} \left[(\beta_1 + \sum_{i=1}^n \omega_i \varepsilon_i - \beta_1)^2 \right], \\ &= \mathbb{E} \left[\left(\sum_{i=1}^n \omega_i \varepsilon_i \right)^2 \right], \\ &= \mathbb{E} \left[\sum_{i=1}^n (\omega_i \varepsilon_i)^2 + 2 \sum_{i < i'}^n \varepsilon_i \varepsilon_{i'} \omega_i \omega_{i'} \right], \\ &= \sum_{i=1}^n \underbrace{\mathbb{E}[\varepsilon_i^2]}_{=\text{Var}[\varepsilon_i]=\sigma^2} \omega_i^2 + 2 \sum_{i < i'}^n \underbrace{\mathbb{E}[\varepsilon_i \varepsilon_{i'}]}_{=0} \omega_i \omega_{i'}, \\ \text{Var}[\hat{\beta}_1] &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.\end{aligned}$$

Etude des propriétés de l'ordonnée à l'origine du modèle

Nous reprenons le même travail avec l'ordonnée à l'origine cette fois-ci.

1. Montrer que l'on

$$\hat{\beta}_0 = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x}.$$

En effet, repartons de l'expression de l'estimateur

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ &\downarrow \text{ on utilise le fait que } \bar{y} = \beta_1 \bar{x} + \beta_0 \\ &= \beta_1 \bar{x} + \beta_0 - \hat{\beta}_1 \bar{x}, \\ &= \beta_0 + (\beta_1 - \hat{\beta}_1) \bar{x}.\end{aligned}$$

2. En déduire que $\hat{\beta}_0$ est un estimateur sans biais de β_0 .

On peut alors déterminer l'espérance de l'estimateur

$$\begin{aligned}\mathbb{E}[\hat{\beta}_0] &= \mathbb{E}[\beta_0 + (\beta_1 - \hat{\beta}_1) \bar{x}], \\ &\downarrow \text{ linéarité de l'espérance } \\ &= \mathbb{E}[\beta_0] + \mathbb{E}[(\beta_1 - \hat{\beta}_1) \bar{x}], \\ &\downarrow \text{ seul } \hat{\beta}_1 \text{ est aléatoire ici } \\ &= \beta_0 + (\beta_1 - \mathbb{E}[\hat{\beta}_1]) \bar{x}, \\ &\downarrow \text{ on a vu que } \mathbb{E}[\hat{\beta}_1] = \beta_1 \\ &= \beta_0.\end{aligned}$$

3. Déterminer la variance de $\hat{\beta}_0$.

Pour cela on va repartir de la définition de départ

$$\begin{aligned}\text{Var}[\hat{\beta}_0] &= \text{Var}[\bar{y} - \hat{\beta}_1 \bar{x}], \\ &\downarrow \text{ définition de la variance } \\ &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^n y_i\right] + \text{Var}[\hat{\beta}_1 \bar{x}] - 2\text{Cov}\left(\frac{\sum_{i=1}^n y_i}{n}, \hat{\beta}_1 \bar{x}\right), \\ &\downarrow \text{ seul les } y_i \text{ et } \hat{\beta}_1 \text{ sont aléatoires } \\ &= \frac{1}{n^2} \text{Var}\left[\sum_{i=1}^n y_i\right] + \bar{x}^2 \text{Var}[\hat{\beta}_1] - \frac{2\bar{x}}{n} \text{Cov}\left[\sum_{i=1}^n y_i, \hat{\beta}_1\right], \\ &\downarrow \text{ or } \text{Var}[y_i] = \sigma^2 \text{ et les } y_i \text{ sont indépendants } \\ &\downarrow \text{ on a déjà calculé la variance de } \hat{\beta}_1 \\ &= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{2\bar{x}}{n} \text{Cov}\left[\sum_{i=1}^n y_i, \sum_{i=1}^n \omega_i \varepsilon_i\right],\end{aligned}$$

$$\begin{aligned}
& \downarrow \text{ on utilise le fait } y_i = \beta_1 x_i + \beta_0 + \varepsilon_i \\
&= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{2\bar{x}}{n} \text{Cov} \left[\sum_{i=1}^n \varepsilon_i, \sum_{i=1}^n \omega_i \varepsilon_i \right], \\
& \downarrow \text{ on utilise le fait que les erreurs sont indépendantes, i.e. } [\varepsilon_i, \varepsilon_j] = 0. \\
&= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{2\bar{x}}{n} \sum_{i=1}^n \text{Var}[\varepsilon_i] \omega_i, \\
& \downarrow \text{ or } \text{Var}[\varepsilon_i] = \sigma^2 \text{ et } \sum_{i=1}^n \omega_i = 0 \\
&\text{Var}[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

Intervalle de confiance

On peut montrer que les variables aléatoires suivantes

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\text{Var}[\hat{\beta}_0]}} \quad \text{et} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}[\hat{\beta}_1]}}$$

suivent une loi de Student à $n - 2$ degrés de liberté, la valeur 2 étant liée aux nombres de paramètres du modèle.

1. Donner l'expression des intervalles de confiance de niveau $1 - \alpha$ des paramètres β_0 et β_1 .

L'intervalle de confiance de niveau $1 - \alpha$ sont données par


$$I_{1-\alpha}(\beta_1) = \left[\hat{\beta}_1 - t_{1-\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}; \hat{\beta}_1 + t_{1-\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

et

$$I_{1-\alpha}(\beta_0) = \left[\hat{\beta}_0 - t_{1-\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}; \hat{\beta}_0 + t_{1-\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right],$$

où $t_{1-\alpha/2, n-2}$ désigne le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - 2$ degrés de liberté.

2. Effectuer l'application numérique en prenant $\alpha = 0.05$ et en utilisant les données de l'énoncé.

On va effectuer notre application numérique avec  directement.
Pour cela, on va d'abord calculer la variance de notre estimateur

```
summary(mymodel)

##
## Call:
## lm(formula = Y ~ X, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9105 -0.2395  0.0500  0.3724  1.2737
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06316    0.64341  -0.098  0.924219
## X            1.36842    0.21184   6.460  0.000196 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8637 on 8 degrees of freedom
## Multiple R-squared:  0.8391, Adjusted R-squared:  0.819
## F-statistic: 41.73 on 1 and 8 DF, p-value: 0.0001963

# Risque d'erreur
alpha = 0.05

# On commence par calculer la variance de nos estimateurs

# Nombre de paramètres du modèle
k = 2
# Nombre d'individus
n = length(data$Y)

# Estimation de la variance des erreurs
sigma_hat = (1/(n-k))*sum(mymodel$residuals^2)

# Calcul des variances
var_b0 = sigma_hat*sum(data$X^2)/(n*(n-1)*var(data$X))
var_b1 = sigma_hat/((n-1)*var(data$X))
```

On peut maintenant déterminer les bornes inférieures et supérieures de nos in-

tervalles de confiance.

```
# Pour la pente du modèle
borne_inf_b1 = coeff[2] - qt(1-alpha/2,n-k)*sqrt(var_b1)
borne_sup_b1 = coeff[2] + qt(1-alpha/2,n-k)*sqrt(var_b1)

# Pour l'ordonnée à l'origine du modèle
borne_inf_b0 = coeff[1] - qt(1-alpha,n-k)*sqrt(var_b0)
borne_sup_b0 = coeff[1] + qt(1-alpha,n-k)*sqrt(var_b0)
```

Test de la significativité du modèle

On cherche enfin à savoir si le modèle appris est significatif. Pour cela on va tester la significativité de la pente du modèle et regarder si cette dernière est significativement différente de 0.

Cette vérification nous permettra d'affirmer que la variable X permet, au moins en partie, d'expliquer les valeurs de la variable aléatoire Y .

1. A l'aide de l'intervalle de confiance précédemment construit, peut-on dire que la pente du modèle est significative ?

```
# On teste si 0 appartient ou non à notre intervalle de confiance
ifelse( (0<borne_inf_b1)|(0>borne_sup_b1),
  "0 n'appartient à l'IC, on rejette H_0",
  "0 appartient à l'IC, on ne rejette pas H_0")

## X
## "0 n'appartient à l'IC, on rejette H_0"
```

2. On peut également procéder au test statistique suivant :

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0.$$

Après avoir déterminé la valeur de la statistique de test, déterminer la p-value associée au test et conclure au risque d'erreur $\alpha = 0.05$.

On effectue ici un test bilatéral donc notre p-value sera calculée comme étant deux fois la probabilité qu'une loi de student à $n - 2$ degrés de libertés prennent des valeurs plus grandes que la valeur absolue de notre statistique de test, soit

$$2\mathbb{P}[T \geq |t_{test}|],$$

où $t_{test} = \frac{\hat{\beta}_1}{\sqrt{\text{Var}[\hat{\beta}_1]}}$ et T désigne une variable aléatoire suivant une loi de student à $n - 2$ degrés de libertés.

```
# Calcul de la statistique de test
t_test = coeff[2]/sqrt(var_b1)

# Calcul de la p-value
p_value = 2*(1-pt(abs(t_test),n-k))
p_value

##           X
## 0.000196273

ifelse(p_value < alpha,
"On rejette H0",
"On ne rejette pas H0")

##           X
## "On rejette H0"

summary(mymodel)

##
## Call:
## lm(formula = Y ~ X, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9105 -0.2395  0.0500  0.3724  1.2737
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06316    0.64341  -0.098  0.924219
## X           1.36842    0.21184   6.460  0.000196 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8637 on 8 degrees of freedom
## Multiple R-squared:  0.8391, Adjusted R-squared:  0.819
## F-statistic: 41.73 on 1 and 8 DF,  p-value: 0.0001963
```