MAT 2777
Probabilités et statistique pour ingénieur.e.s

Chapitre 7
La régression linéaire et la corrélation

P. Boily (uOttawa)

Hiver 2023

Aperçu

Scénario – Motivation (p.3)

7.1 – Le coefficient de corrélation (p.5)

- Les propriétés de ρ_{XY} (p.6)
- Le calcul de ρ_{XY} avec R (p.8)

7.2 – La régression linéaire simple (p.9)

- L'estimation de la variance σ^2 (p.18)
- Les propriétés des estimateurs des moindres carrés (p.23)

7.3 – Tests d'hypothèses pour la régression linéaire (p.26)

- L'ordonnée à l'origine (p.27)
- La pente (p.29)
- La signification de la régression (p.33)

7.4 – Les intervalles de confiance et de prédiction pour la régression linéaire (p.36)

- L'ordonnée à l'origine et la pente (p.37)
- La réponse moyenne (p.39)
- La prédiction de nouvelles observations (p.43)

7.5 – L'analysis de la variance (p.47)

7.6 – Le coefficient de détermination (p.50)

Annexe – Résumés et exemples (p.52)

- Les arrestations aux États-Unis (p.53)
- Les données de compagnies aériennes (p.59)

Scénario – Motivation

Considérons les données suivantes, constituées de $n = 20$ observations appariées (x_i, y_i) des teneurs en hydrocarbures (x) et en oxygène pur (y) dans des carburants :

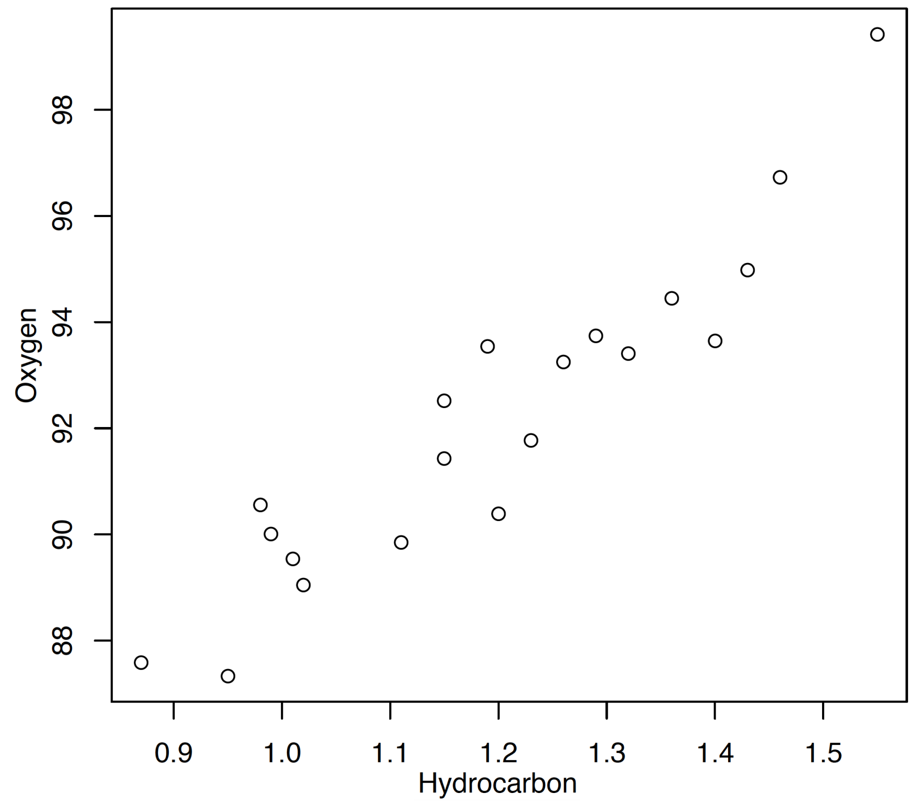
```
x:  0.99  1.02  1.15  1.29  1.46  1.36  0.87  1.23  1.55  1.40
y: 90.01 89.05 91.43 93.74 96.73 94.45 87.59 91.77 99.42 93.65

x:  1.19  1.15  0.98  1.01  1.11  1.20  1.26  1.32  1.43  0.95
y: 93.54 92.52 90.56 89.54 89.85 90.39 93.25 93.41 94.98 87.33
```

Objectifs :

- mesurer la **force de l'association** entre x et y.
- décrire la **relation** entre x et y.

On fournit une première description de la relation à l'aide d'un graphique.



Il semblerait que les points se situent autour d'une droite cachée !

7.1 – Le coefficient de corrélation

Pour les observations appariées (x_i, y_i), $i = 1, \ldots, n$, on défini le **coefficient de corrélation** de x et y par

$$\rho_{XY} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}}.$$

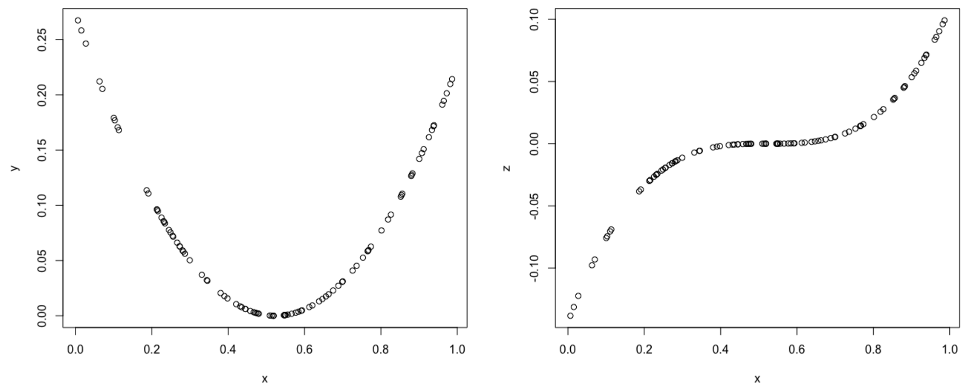
Le coefficient ρ_{XY} n'est défini que si $S_{xx} \neq 0$ et $S_{yy} \neq 0$, c'est-à-dire que ni x_i ni y_i ne sont constants. Les variables x et y sont **sans corrélation** si $\rho_{XY} = 0$ (ou très petit, en pratique), et sont **corrélées** si $\rho_{XY} \neq 0$ (ou $|\rho_{XY}|$ est "grand", en pratique).

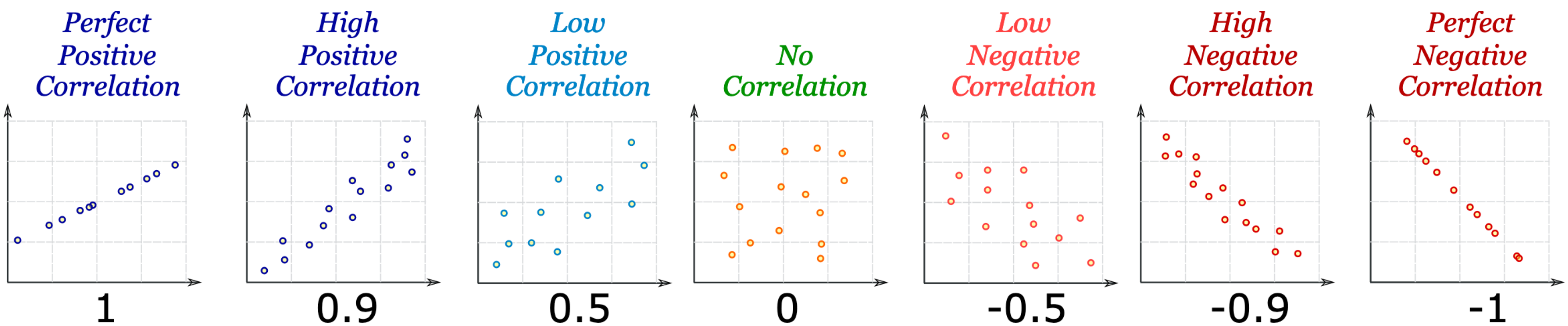
Exemple : pour les données de la diapositive précédente, nous avons $S_{xy} \approx 10.18$, $S_{xx} \approx 0.68$, $S_{yy} \approx 173.38$, et $\rho_{XY} \approx \frac{10.18}{\sqrt{0.68 \cdot 173.38}} \approx 0.94$.

Les propriétés de ρ_{XY}

- ρ_{XY} n'est pas affecté par les changements d'**échelle** ou d'**origine** ;
- ρ_{XY} est **symétrique** en x et y ($\rho_{XY} = \rho_{YX}$) et $-1 \leq \rho_{XY} \leq 1$;
- si $\rho_{XY} = \pm 1$, alors les observations (x_i, y_i) se retrouvent toutes sur une droite avec une pente **positive (1)** et **négative (−1)** ;
- le signe de ρ_{XY} reflète la **tendance** des points ;
- si la valeur de $|\rho_{XY}|$ est élevée, cela n'implique pas nécessairement qu'il y une **relation de causalité** entre x and y ;

- x et y peuvent avoir une relation **non linéaire** très prononcée sans que ρ_{XY} ne le reflète (-0.12 à gauche, 0.93 à droite)





Le calcul de ρ_{XY} avec R

```
> x=c(0.99, 1.02, 1.15, 1.29, 1.46, 1.36, 0.87, 1.23, 1.55, 1.40,
      1.19, 1.15, 0.98, 1.01, 1.11, 1.20, 1.26, 1.32, 1.43, 0.95)
> y=c(90.01, 89.05, 91.43, 93.74, 96.73, 94.45, 87.59, 91.77, 99.42, 93.65,
      93.54, 92.52, 90.56, 89.54, 89.85, 90.39, 93.25, 93.41, 94.98, 87.33)

> plot(x,y) # will produce the scatterplot on slide 3
> cor(x,y)
    0.9367154

> Sxy=sum((x-mean(x))*(y-mean(y)))
> Sxx=sum((x-mean(x))^2)
> Syy=sum((y-mean(y))^2)
> rho=Sxy/(sqrt(Sxx*Syy))
> rho
    0.9367154
```

7.2 – La régression linéaire simple

L'**analyse de régression** peut être utilisée pour décrire la relation entre une **variable prédictive** X et une **variable réponse** Y. Supposons qu'elles sont liées par le modèle d'**ajustement linéaire**

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

où ε est l'**erreur aléatoire** et β_0, β_1 sont les **coefficients de régression**.

On suppose que $\mathrm{E}[\varepsilon] = 0$, et que la variance de l'erreur $\sigma_\varepsilon^2 = \sigma^2$ est constante. Le modèle peut alors se ré-écrire sous la forme

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X.$$

Supposons que les observations $(x_i, y_i), i = 1, \ldots, n$ sont telles que

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

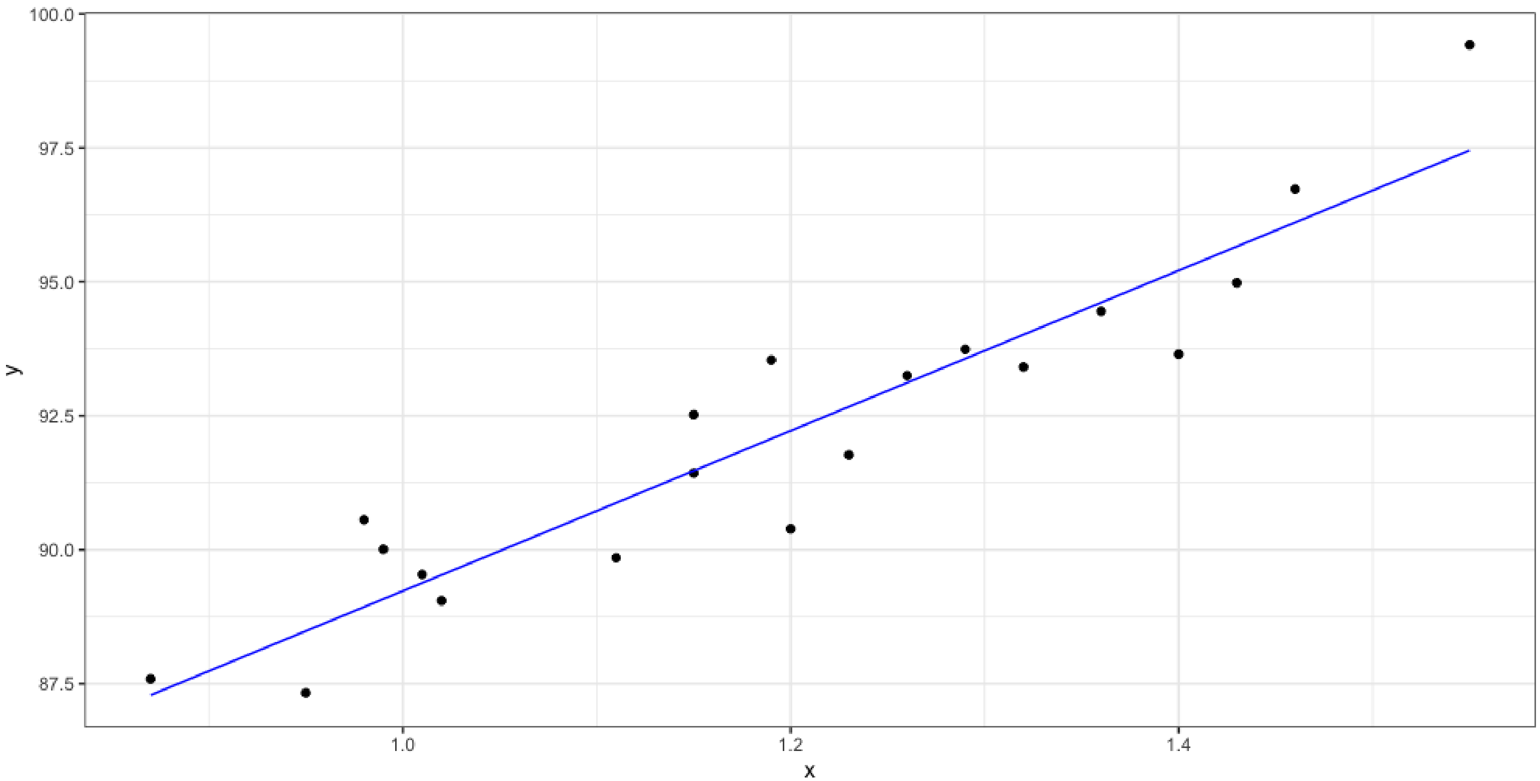
On cherche des **estimateurs** b_0, b_1 des paramètres **inconnus** β_0, β_1, afin d'obtenir la **droite d'ajustement estimée**

$$\hat{y}_i = b_0 + b_1 x_i.$$

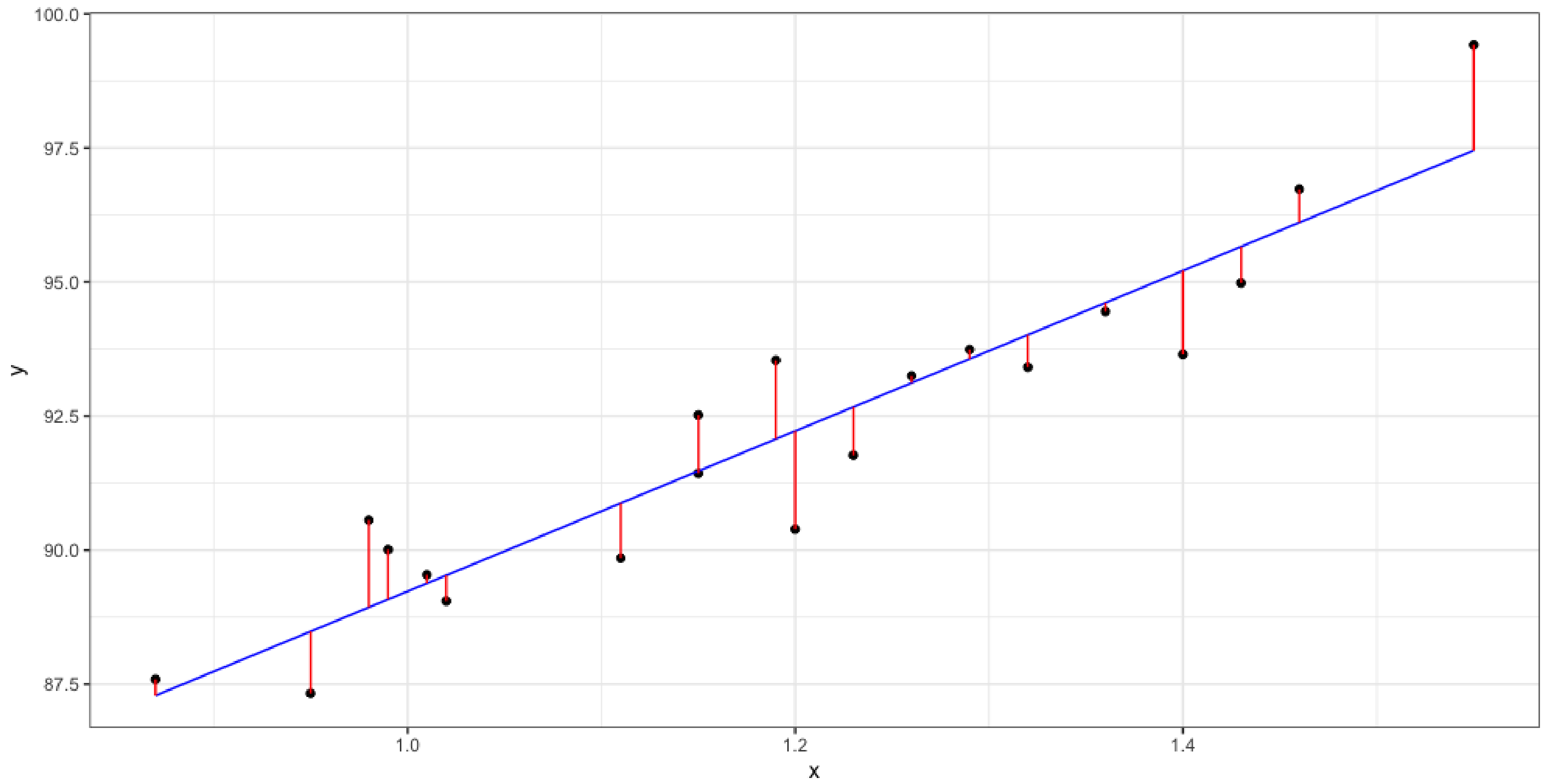
L'**erreur résiduelle** obtenue en prédisant y_i à l'aide de $\hat{y}_i$ est alors

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i, \quad i = 1, \ldots, n.$$

Comment trouve-t-on les estimateurs ? Comment déterminons-nous si la droite ajustée est un bon modèle pour les données ?



droite ajustée: $\hat{y} = 74.28 + 14.95x$



erreurs résiduelles: $e_i = y_i - \hat{y}_i$

Considérons la **somme des erreurs quadratiques** (SSE) :

$$\mathrm{SSE}=\sum_{i=1}^{n} e_i^2=\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2$$

(on peut montrer que $\mathrm{SSE}/\sigma^2 \sim \chi^2(n-2)$, mais cela n'est pas dans le cadre du cours). Les valeurs optimales de b_0 et b_1 sont celles qui minimisent l'SSE. À ce titre, on résoud

$$0=\frac{\mathrm{dSSE}}{\mathrm{d}b_0}=-2\sum(y_i-b_0-b_1x_i)=-2n(\overline{y}-b_0-b_1\overline{x})$$

$$0=\frac{\mathrm{dSSE}}{\mathrm{d}b_1}=-2\sum(y_i-b_0-b_1x_i)x_i=-2\left(\sum x_iy_i-nb_0\overline{x}-b_1\sum x_i^2\right)$$

afin d'obtenir les **estimateurs des moindres carrés** b_0, b_1 de β_0, β_1, resp.

Puisque $\frac{\mathrm{dSSE}}{\mathrm{d}b_0}=0$, nous obtenons

$$\overline{y}-b_0-b_1\overline{x}=0 \Longrightarrow b_0=\overline{y}-b_1\overline{x}.$$

Pour le second coefficient, on note que

$$\begin{aligned} S_{xy} &= \sum(x_i-\overline{x})(y_i-\overline{y})=\sum x_iy_i-n\overline{xy} \\ S_{xx} &= \sum(x_i-\overline{x})^2=\sum x_i^2-n\overline{x}^2, \end{aligned}$$

d'où

$$\begin{aligned} \sum x_iy_i &= S_{xy}+n\overline{xy} \\ \sum x_i^2 &= S_{xx}+n\overline{x}^2. \end{aligned}$$

Puisque $\frac{\mathrm{dSSE}}{\mathrm{d}b_1} = 0$, nous obtenons

$$\sum x_i y_i - nb_0\overline{x} - b_1 \sum x_i^2 = 0$$

$$(S_{xy} + n\overline{xy}) - nb_0\overline{x} - b_1(S_{xx} + n\overline{x}^2) = 0$$

$$S_{xy} + n\overline{xy} - n(\overline{y} - b_1\overline{x})\overline{x} - b_1 S_{xx} - nb_1\overline{x}^2 = 0$$

$$S_{xy} + n\overline{xy} - n\overline{xy} + nb_1\overline{x}^2 - b_1 S_{xx} - nb_1\overline{x}^2 = 0$$

$$S_{xy} - b_1 S_{xx} = 0$$

$$b_1 = \frac{S_{xy}}{S_{xx}}.$$

Les estimateurs sont des **combinaisons linéaires** des réponses observées y_i :

$$b_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} u_i y_i, \quad b_0 = \overline{y} - b_1\overline{x} = \sum_{i=1}^{n} v_i y_i.$$

Exemple : pour les données sur les carburants, nous avons déjà trouvé que

$$S_{xy} \approx 10.18,\ \ S_{xx} \approx 0.68,\ \ \text{et } S_{yy} = 173.38.$$

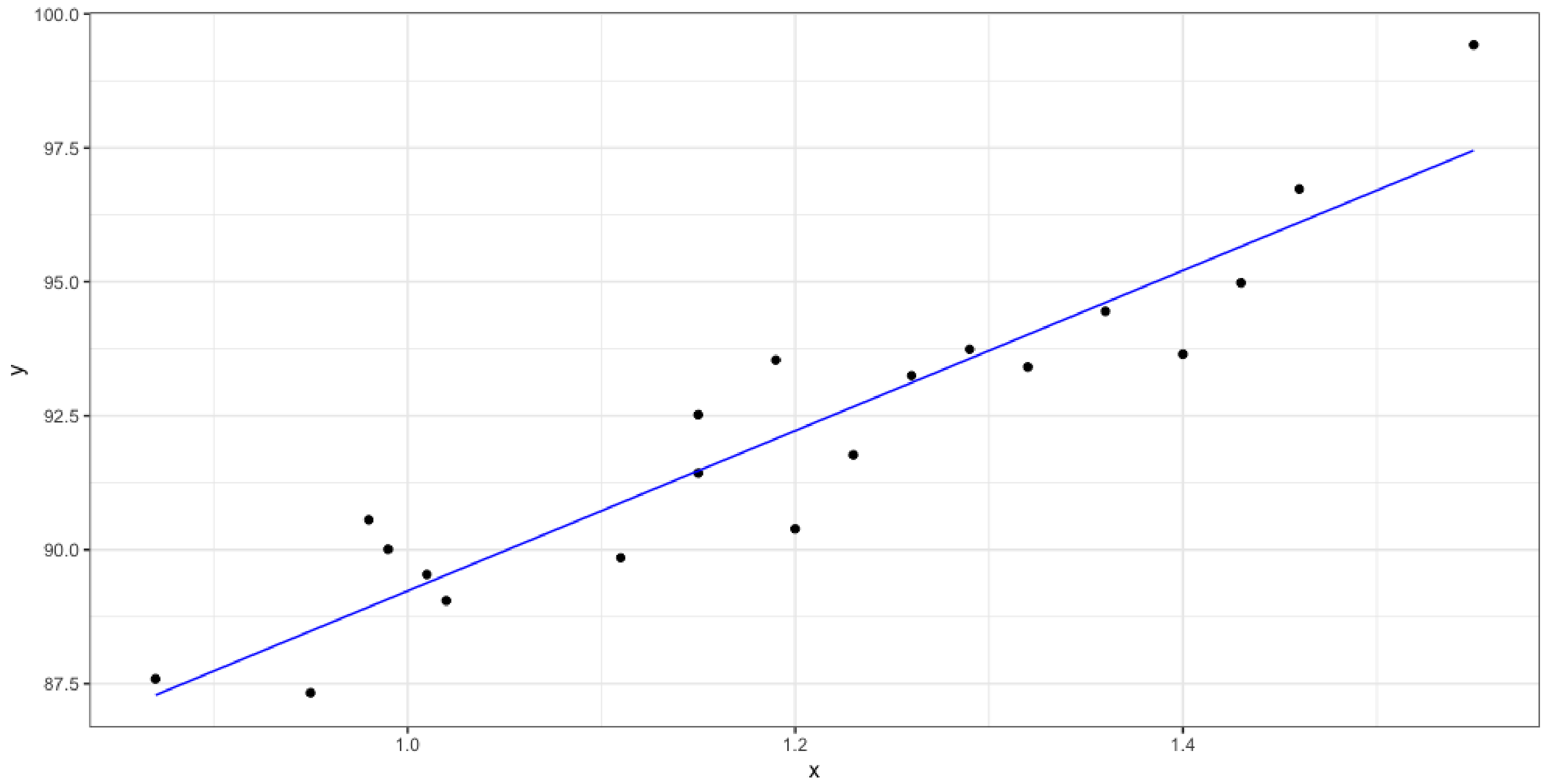
Ainsi, $b_1 = \frac{10.18}{0.68} = 14.95$. De plus, puisque

$$n = 20,\ \ \overline{x} = 1.20,\ \ \text{et } \overline{y} = 92.16,$$

nous avons aussi $b_0 = 92.16 - 14.95(1.20) = 74.28$.

Par conséquent, la **droite de régression ajustée** est

$$\hat{y} = 74.28 + 14.95x.$$



droite de régression : $\hat{y} = 74.28 + 14.95x$

L'estimation de la variance σ^2

Rappelons que la variance du terme d'erreur est $\sigma_\varepsilon^2 = \sigma^2$. Pour estimer σ^2, nous utilisons

$$\mathrm{SSE} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

La question est la suivante : quel dénominateur devons-nous utiliser ?

Pour une population, nous utilisons n. Pour un échantillon, $n-1$. Pour l'erreur de régression, l'**estimateur non biaisé** de σ^2 est en fait

$$\hat{\sigma}^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2},$$

puisque SSE a $n-2$ **degrés de liberté** : on doit trouver une estimation de 2 paramètres afin d'obtenir $\hat{y}_i$, b_0 **et** b_1.

Exemple : quelle est la variance de l'erreur dans le modèle linéaire pour les données sur les carburants ?

Solution : puisque $S_{xy} \approx 10.18$, $S_{yy} = 173.38$, $b_1 = 14.95$, et $n = 20$, nous avons

$$\hat{\sigma}^2 = \frac{173.38 - 14.95(10.18)}{20-2} \approx 1.18.$$

Le code suivant montre comment tracer la droite de meilleur ajustement, obtenir les estimateurs de β_1, β_2, et extraire l'**erreur quadratique moyenne** (MSE) dans `R`, en supposant que `x`, `y`, `Sxx`, et `Sxy` ont été déjà été calculés.

```
> library(ggplot2)   ### preambule
> fuels=data.frame(x,y)
> ### fonction R pour le meilleur ajustement lineaire
> model <- lm(y ~ x, data=fuels)
```

```
> summary(model)   ### a expliquer plus tard
    Call: lm(formula = y ~ x, data = fuels)

    Residuals:
         Min        1Q   Median        3Q       Max
    -1.83029 -0.73334  0.04497  0.69969  1.96809

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept)   74.283      1.593   46.62  < 2e-16 ***
    x             14.947      1.317   11.35 1.23e-09 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
    Residual standard error: 1.087 on 18 degrees of freedom
    Multiple R-squared: 0.8774, Adjusted R-squared: 0.8706
    F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09

> ### graphique de la droite d'ajustement
> ggplot(model) + geom_point(aes(x=x, y=y)) +
    geom_line(aes(x=x, y=.fitted), color="blue" ) +
    theme_bw()

> ### graphiques des erreurs residuelles
> ggplot(model) + geom_point(aes(x=x, y=y)) +
    geom_line(aes(x=x, y=.fitted), color="blue" ) +
    geom_linerange(aes(x=x, ymin=.fitted, ymax=y), color="red") +
    theme_bw()
```

```
> ### calcul direct de la variance
> n=length(x)
> sigma2 = (Syy-as.numeric(model$coefficients[2])*Sxy)/(n-2)
> sigma2
    1.180545

> ### calcul de MSE a partir du resume
> summary(model)$sigma^2
    1.180545
```

Il serait a votre avantage d'apprendre à utiliser des logiciels afin de trouver la droite de meilleur ajustement, et surtout, d'apprendre à lire les résultats de ces logiciels.

Les propriétés des estimateurs des moindres carrés

Le modèle d'ajustement linéaire simple est

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \text{avec} \quad \mathrm{E}[\varepsilon] = 0, \; \sigma_\varepsilon^2 = \sigma^2.$$

Étant donné X, Y est une variable aléatoire :

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X, \quad \mathrm{Var}[Y|X] = \sigma^2.$$

Notez que b_0 et b_1 dépendent des x et y observés, qui sont des réalisations des v.a. X et Y. Par conséquent, les estimateurs sont des **variables aléatoires**, c-à-d que différentes réalisations (données observées) conduisent à différentes estimations b_0, b_1 de β_0, β_1.

On peut montrer que

$$\mathrm{E}[b_0] = \beta_0, \qquad \sigma^2_{b_0} = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right] = \sigma^2\frac{\sum_{i=1}^n x_i^2}{nS_{xx}},$$
$$\mathrm{E}[b_1] = \beta_1, \qquad \sigma^2_{b_1} = \sigma^2/S_{xx}.$$

Nous disons des estimateurs b_0, b_1 de β_0, β_1 qu'ils sont **sans biais**. les erreurs types estimées (obtenues en remplaçant σ^2 par $\mathrm{MSE} = \hat{\sigma}^2$ dans les expressions pour $\sigma^2_{b_1}$ et $\sigma^2_{b_0}$) sont :

$$\mathrm{se}(b_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]} \quad \text{et} \quad \mathrm{se}(b_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

Exemple : trouvez les erreurs types estimées de b_0 et b_1 dans les données sur les carburants.

Solution : puisque $n = 20$, $\overline{x} = 1.20$, $S_{xx} = 0.68$, et $\hat{\sigma}^2 = 1.18$, nous avons t

$$\text{se}(b_0) = \sqrt{1.18\left[\frac{1}{20} + \frac{1.20^2}{0.68}\right]} \approx 1.593 \quad \text{et} \quad \text{se}(b_1) = \sqrt{\frac{1.18}{0.68}} \approx 1.317.$$

Ces informations sont également disponibles dans les résultats de `R` :

```
> summary(model)$coefficients

            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 74.28331   1.593473 46.61723 3.171476e-20
x           14.94748   1.316758 11.35173 1.227314e-09
```

7.3 – Tests d'hypothèses pour la régression linéaire

Avec les erreurs types, nous pouvons **tester des hypothèses** sur les paramètres de régression.

Les étapes sont les mêmes que celles du chapitre 6 :

1. définissez une **hypothèse nulle** H_0 et une **hypothèse alternative** H_1 ;
2. calculez une **statistique de test** (souvent à l'aide d'une normalisation) ;
3. trouvez la **région critique/valeur**$-p$ correspondante, sous H_0 ;
4. **rejetez** H_0 **ou non** en se basant sur la **région critique/valeur**$-p$.

Test d'hypothèse pour l'ordonnée β_0

Nous pourrions nous intéresser à vérifier si la véritable ordonnée à l'origine β_0 est égale à une certaine **valeur candidate** $\beta_{0,0}$, c'est-à-dire

$$H_0 : \beta_0 = \beta_{0,0} \quad \text{par rapport à} \quad H_1 : \beta_0 \neq \beta_{0,0}.$$

Le modèle de régression linéaire requiert des erreurs normales $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, ce qui implique que $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$, $i = 1, \ldots, n$.

Puisque le coefficient b_0 est une fonction linéaire des y_i, il suit une loi normale de moyenne β_0 et de variance $\sigma^2 \frac{\sum x_i^2}{nS_{xx}}$. Par conséquent, si H_0 est valide,

$$Z_0 = \frac{b_0 - \beta_{0,0}}{\sqrt{\sigma^2 \frac{\sum x_i^2}{nS_{xx}}}} \sim \mathcal{N}(0, 1).$$

Mais la variance σ^2 est inconnue ; la statistique de test (avec $\hat{\sigma}^2 = \mathrm{MSE}$)

$$T_0 = \frac{b_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \frac{\sum x_i^2}{nS_{xx}}}} \sim t(n-2)$$

suit une loi t de Student avec $n-2$ degrés de liberté.

Hypothèse alternative	**Région critique**
$H_1 : \beta_0 > \beta_{0,0}$	$t_0 > t_\alpha(n-2)$
$H_1 : \beta_0 < \beta_{0,0}$	$t_0 < -t_\alpha(n-2)$
$H_1 : \beta_0 \neq \beta_{0,0}$	$\lvert t_0 \rvert > t_{\alpha/2}(n-2)$

où t_0 est la valeur observée de T_0 et $t_\alpha(n-2)$ est la valeur critique satisfaisant $P(T > t_\alpha(n-2)) = \alpha$, lorsque $T \sim t(n-2)$.

Rappel : **rejetez H_0 lorsque t_0 se retrouve dans la région critique**.

Test d'hypothèse pour la pente β_1

Nous pourrions nous intéresser à vérifier si la véritable pente β_1 est égale à une certaine **valeur candidate** $\beta_{1,0}$, c'est-à-dire

$$H_0 : \beta_1 = \beta_{1,0} \quad \text{par rapport à} \quad H_1 : \beta_1 \neq \beta_{1,0}.$$

Le modèle de régression linéaire requiert des erreurs normales $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, ce qui implique que $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$, $i = 1, \ldots, n$.

Puisque le coefficient b_1 est une fonction linéaire des y_i, il suit une loi normale de moyenne β_1 et de variance $\frac{\sigma^2}{S_{xx}}$. Par conséquent, si H_0 est valide,

$$Z_0 = \frac{b_1 - \beta_{1,0}}{\sqrt{\sigma^2/S_{xx}}} \sim \mathcal{N}(0, 1).$$

Mais la variance σ^2 est inconnue ; la statistique de test (avec $\hat{\sigma}^2 = \mathrm{MSE}$)

$$T_0 = \frac{b_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t(n-2)$$

suit une loi t de Student avec $n-2$ degrés de liberté.

Hypothèse alternative	**Région critique**
$H_1 : \beta_1 > \beta_{1,0}$	$t_0 > t_\alpha(n-2)$
$H_1 : \beta_1 < \beta_{1,0}$	$t_0 < -t_\alpha(n-2)$
$H_1 : \beta_1 \neq \beta_{1,0}$	$\lvert t_0\rvert > t_{\alpha/2}(n-2)$

où t_0 est la valeur observée de T_0 et $t_\alpha(n-2)$ est la valeur critique satisfaisant $P(T > t_\alpha(n-2)) = \alpha$, lorsque $T \sim t(n-2)$.

Rappel : **rejetez H_0 lorsque t_0 se retrouve dans la région critique**.

Exemples : utilisez l'ensemble de données sur les carburants et supposez que les quantités ont été attribuées/calculées lors d'une étape précédente.

a) Testez $H_0 : \beta_0 = 75$ par rapport à $H_1 : \beta_0 < 75$ lorsque $\alpha = 0.05$.

b) Testez $H_0 : \beta_1 = 10$ par rapport à $H_1 : \beta_1 > 10$ lorsque $\alpha = 0.05$.

c) Testez $H_0 : \beta_1 = 0$ par rapport à $H_1 : \beta_1 \neq 0$ lorsque $\alpha = 0.05$.

Solution : nous ne rejetons pas H_0 en a), mais nous le faisons en b) et c).

```
> b0 = as.numeric(model$coefficients[1])  ### parametre de regression
> b1 = as.numeric(model$coefficients[2])  ### parametre de regression
> beta00 = 75  ### pour a)
> beta10 = 10  ### pour b)
```

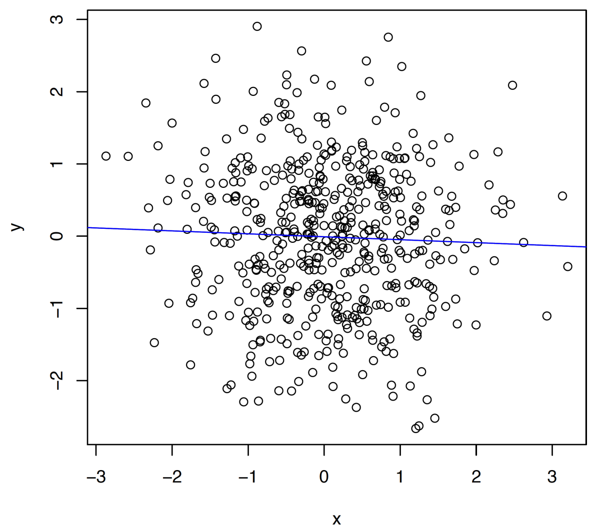
```
# a)
> t0a = (b0-beta00)/sqrt(sigma2*sum(x^2)/n/Sxx)  ### statistique de test
> crit_t005_18a = qt(0.05,n-2)                   ### valeur critique
> t0a < crit_t005_18a                            ### test de region critique
    FALSE                                        ### on ne rejete pas H0

# b)
> t0b = (b1-beta10)/sqrt(sigma2/Sxx)  ### statistique de test
> crit_t005_18b = - qt(0.05,n-2)      ### valeur critique
> t0b > crit_t005_18b                 ### test de region critique
    TRUE                              ### on rejete H0 pour H1

# c)
> t0c = b1/sqrt(sigma2/Sxx)           ### statistique de test
> crit_t0025_18c = - qt(0.025,18)     ### valeur critique
> abs(t0c) > crit_t0025_18c           ### test de region critique
    TRUE                              ### on rejete H0 pour H1
```

La signification de la régression

Tant que $S_{xx} \neq 0$, nous pouvons ajuster une ligne de régression aux observations en utilisant l'**approche des moindres carrés**. Rappelons que l'un des objectifs de la régression linéaire est de **décrire une relation linéaire** entre X et Y... tant qu'une telle relation existe.



La droite de régression pour l'ensemble de données de la diapositive précédente est

$$\hat{y} = -0.01 - 0.04x,$$

mais cette droite **ne décrit pas du tout** l'ensemble de données, qui ressemble plutôt à une tache diffuse. La relation entre X et Y dans cet ensemble de données **n'est tout simplement pas linéaire**.

Étant donné une ligne de régression, nous pouvons vouloir tester si elle est **significative**. Le test de **signification de la régression** est

$$H_0 : \beta_1 = 0 \text{ par rapport à } H_1 : \beta_1 \neq 0.$$

Si nous rejetons H_0 en faveur de H_1, alors l'évidence suggère qu'il existe une **relation linéaire** entre X et Y.

Exemple : dans l'ensemble de données sur les carburants, nous avons $b_1 = 14.95$, $n = 20$, $S_{xx} = 0.68$, $\hat{\sigma}^2 = 1.18$. Nous testons pour la signification de la régression lorsque $\alpha = 0.01$:

$$H_0 : \beta_1 = 0 \quad \text{par rapport à} \quad H_1 : \beta_1 \neq 0.$$

Puisque la valeur observée de la statistique de test est

$$t_0 = \frac{b_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = 11.35 > 2.88 = t_{0.01/2}(18)\,,$$

où $t_{0.01/2}(18)$ est la valeur critique de la loi de Student avec 18 degrés de liberté lorsque $\alpha = 0.01$ dans le cadre d'un test bilatéral, nous **rejetons** H_0 et concluons qu'il **existe une relation linéaire** entre X et Y (lorsque $\alpha = 0.01$).

7.4 – Les intervalles de confiance et de prédiction pour la régression linéaire

Nous pouvons également construire des **intervalles de confiance** (I.C.) pour les paramètres de régression et des **intervalles de prédiction** (I.P.) pour les valeurs prédites. Les étapes sont les mêmes que celles du chapitre 5 :

1. calculez une **estimation** W pour un paramètre β ou une prédiction Y ;
2. déterminez l'**erreur type appropriée** $\text{se}(W)$;
3. sélectionnez un **niveau de confiance** α et trouvez la valeur critique correspondante $k_{\alpha/2}$;
4. construisez l'**intervalle à environ** $100(1-\alpha)\%$: $W \pm k_{\alpha/2} \cdot \text{se}(W)$.

L'ordonnée à l'origine et la pente

Puisque nous estimons la variance de l'erreur à l'aide de $\hat{\sigma}^2 = \text{MSE}$, nous devons utiliser la loi de Student avec $n-2$ degrés de liberté (i.e., nous utilisons les données pour estimer 2 paramètres).

Les I.C. de β_0, β_1 à environ $100(1-\alpha)\%$ sont :

$$\beta_0 : \qquad b_0 \pm t_{\alpha/2}(n-2)\text{se}(b_0) = b_0 \pm t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2 \frac{\sum x_i^2}{nS_{xx}}}$$

$$\beta_1 : \qquad b_1 \pm t_{\alpha/2}(n-2)\text{se}(b_1) = b_1 \pm t_{\alpha/2}(n-2)\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

La mise en garde concernant l'interprétation des I.C. reste valable.

Exemple : trouvez les I.C. de β_0, β_1 à environ $95\%, 99\%$ dans l'exemple des carburants.

Solution : nous avons vu que $b_0 = 74.283$, $b_1 = 14.947$, $\text{se}(b_0) = 1.593$, $\text{se}(b_1) = 1.317$, $t_{0.025}(18) = 2.10$, et $t_{0.005}(18) = 2.88$.

Ainsi, lorsque $\alpha = 0.05$, nous obtenons

$$\begin{aligned}\beta_0 &: \ 74.283 \pm 2.10(1.593) = (70.93, 77.63)\\ \beta_1 &: \ 14.497 \pm 2.10(1.317) = (12.18, 17.71),\end{aligned}$$

et lorsque $\alpha = 0.01$, nous obtenons

$$\begin{aligned}\beta_0 &: \ 74.283 \pm 2.88(1.593) = (69.70, 78.87)\\ \beta_1 &: \ 14.497 \pm 2.88(1.317) = (11.15, 18.74).\end{aligned}$$

La réponse moyenne

Nous pourrions également nous intéresser à $\mu_{Y|x_0} = \mathrm{E}[Y|x_0]$, la **réponse moyenne** à un x_0 donné (en pratique, il pourrait y avoir réplication dans une expérience pour x_0, disons).

La valeur prédite peut être lue directement à partir de la droite d'ajustement :

$$\hat{\mu}_{Y|x_0} = b_0 + b_1 x_0.$$

Lorsque $x = x_0$, la distance entre la valeur estimée et la droite d'ajustement est

$$\hat{\mu}_{Y|x_0} - \mu_{Y|x_0} = (b_0 - \beta_0) + (b_1 - \beta_1)\, x_0.$$

Mais $\mathrm{E}[\hat{\mu}_{Y\,|\,x_0}] = \mu_{Y\,|\,x_0}$ et

$$\mathrm{Var}[\hat{\mu}_{Y\,|\,x_0}] = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right].$$

On remarque que

$$\mathrm{Var}[\hat{\mu}_{Y\,|\,x_0}] = \mathrm{Var}[b_0 + b_1 x_0] \neq \mathrm{Var}[b_0] + \mathrm{Var}[b_1 x_0]$$

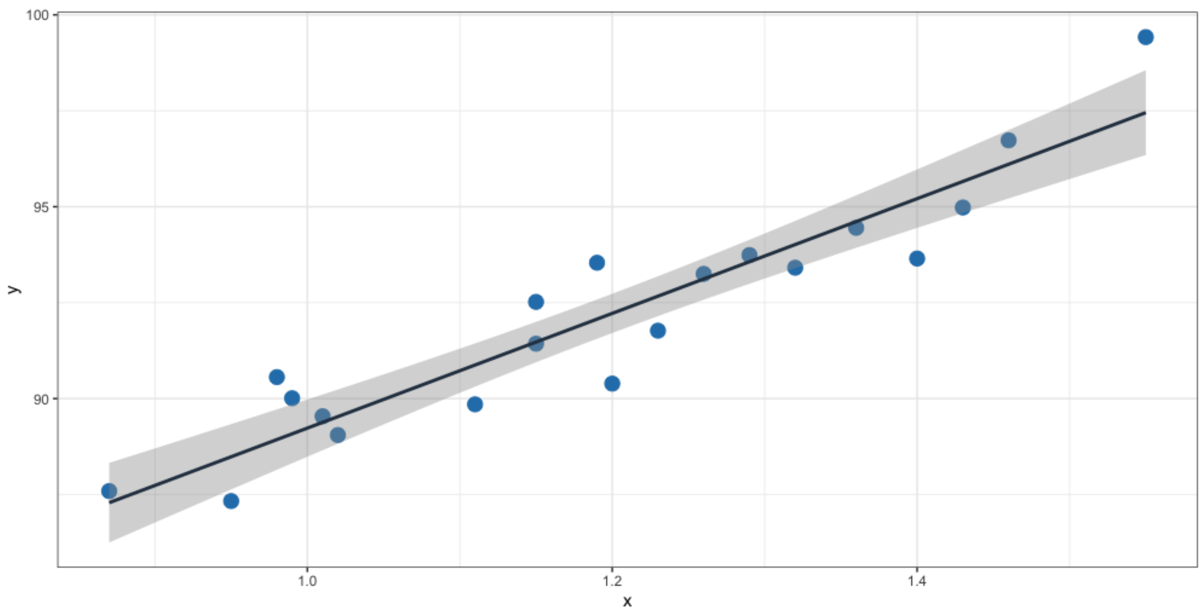
car b_0 et b_1 ne sont pas indépendants.

Avec la valeur critique habituelle $t_{\alpha/2}(n-2)$, l'I.C. de la **réponse moyenne à** x_0 à environ $100(1-\alpha)\%$ est

$$\hat{\mu}_{Y\,|\,x_0} \pm t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}.$$

Exemple : dans les données de carburant, l'I.C. de $\mu_{Y|x_0}$ à environ 95% est

$$74.28 + 14.95x_0 \pm 2.10\sqrt{1.18\left[\frac{1}{20} + \frac{(x_0 - 1.12)^2}{0.68}\right]}.$$



Un bon nombre d'observations se trouvent en dehors de l'I.C. de la réponse moyenne à environ 95% pour la réponse moyenne, ce qui peut s'expliquer par la taille relativement faible de l'échantillon (et nous avons ignoré la procédure de correction de Bonferroni pour les intervalles simultanés, qui est hors du cadre de ce cours).

Le code `R` permettant de produire ce graphique est présenté ci-dessous :

```
> ggplot(fuels, aes(x=x, y=y)) +
    geom_point(color='#2980B9', size = 4) +
    geom_smooth(method=lm, color='#2C3E50') +
    theme_bw()
```

La prédiction de nouvelles observations

Si x_0 est la valeur d'intérêt pour le régresseur (prédicteur), alors la **valeur estimée de la variable réponse Y** est

$$\hat{y} = \hat{Y}_0 = b_0 + b_1 x_0.$$

Si Y_0 est la véritable observation éventuelle lorsque $X = x_0$ (c-à-d que $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon$) et si $\hat{Y}_0$ est la valeur **prédite** telle que donnée par l'équation ci-dessus, alors l'**erreur de prédiction**

$$e_{\hat{p}} = Y_0 - \hat{Y}_0 = \beta_0 + \beta_1 x_0 + \varepsilon - (b_0 + b_1 x_0) = (\beta_0 - b_0) + (\beta_1 - b_1) x_0 + \varepsilon$$

suit une loi normale de moyenne **nulle** et de variance $\sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]$.

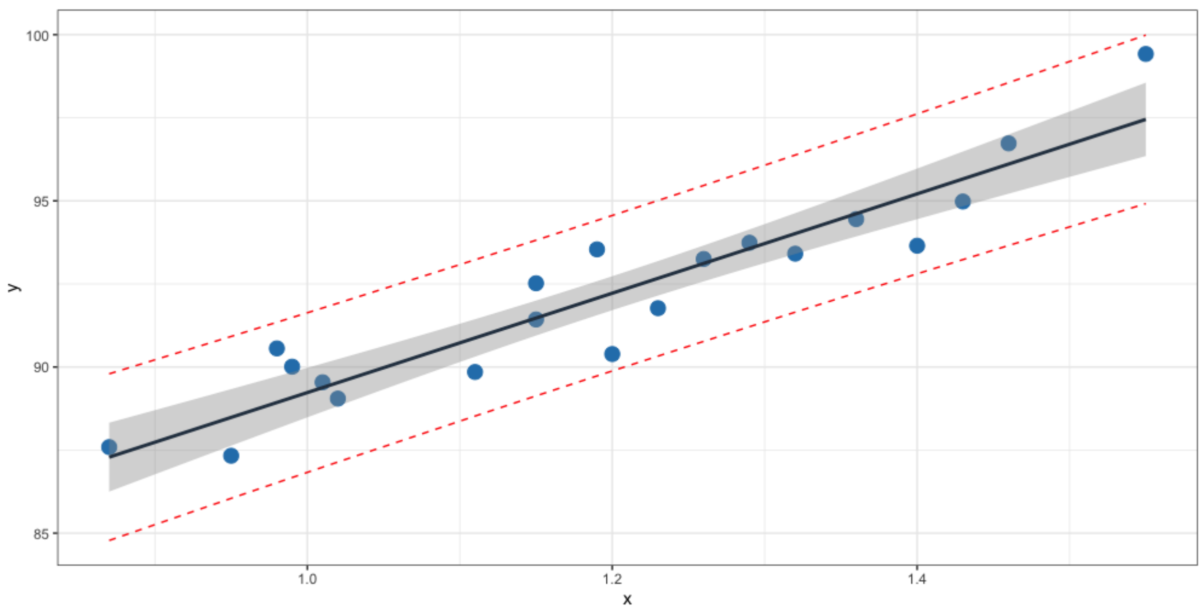
En substituant σ^2 par son estimateur $\hat{\sigma}^2 = \mathrm{MSE}$, nous obtenons un **intervalle de prédiction (I.P.)** de Y_0 à environ $100(1-\alpha)\%$:

$$b_0 + b_1 x_0 \pm t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]},$$

où $t_{\alpha/2}$ est la valeur critique de la loi de Student avec $n-2$ degrés de liberté pour un α donné.

Exemple : pour l'ensemble de données sur les carburants, l'I.P. de $\mu_{Y|x_0}$ à environ 95% est

$$74.28 + 14.95x_0 \pm 2.10\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(x_0 - 1.12)^2}{0.68}\right]}.$$



Aucune des observations ne se trouve en dehors de l'I.P. de nouvelles observations. En général, pour un α donné, l'I.P. est plus large que l'I.C., ce qui n'est pas surprenant : le TLC implique que la réponse moyenne a une plus faible variance celle des les réponses prédites.

Le code R qui produit le graphique sur la diapo précédente est

```
## construction de l'I.P.
> preds <- predict(model, interval="prediction")

## placer les donnees dans un dataframe
> new.fuels <- cbind(fuels, preds)

## tracer le graphique
> ggplot(new.fuels, aes(x=x, y=y)) +
    geom_point(color='#2980B9', size = 4) +
    geom_smooth(method=lm, color='#2C3E50') +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed") +
    geom_line(aes(y=upr), color = "red", linetype = "dashed") +
    theme_bw()
```

7.5 – L'analyse de la variance

Le test de **signification de la régression**,

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

peut être reformulé en terme d'**analyse de la variance** (ANOVA), donnée par le tableau suivant :

Source de variation	**Somme de carrés**	**deg lib**	**Moyenne quad**	F^*	**Valeur**$-p$
Régression	SSR	1	MSR	$\frac{\text{MSR}}{\text{MSE}}$	$P(F > F^*)$
Erreur	SSE	$n-2$	MSE		
Total	SST	$n-1$			

Dans ce tableau, la statistique F^* suit une loi $F(1, n-2)$, et

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad \text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2, \quad \text{SST} = \sum_{i=1}^{n} (y_i - \overline{y})^2,$$

$$\text{MSR} = \frac{\text{SSR}}{1}, \quad \text{MSE} = \frac{\text{SSE}}{n-2}, \quad \text{et} \quad F^* = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/1}{\text{SSE}/n-2}$$

La **zone de rejet** pour l'hypothèse nulle $H_0 : \beta_1 = 0$ est encore

$$|T^*| = \left| \frac{b_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \right| > t_{\alpha/2}(n-2),$$

mais elle peut aussi s'écrire comme $F^* > f_\alpha(1, n-2)$, où $f_\alpha(1, n-2)$ est la valeur critique de la loi F avec $\nu_1 = 1$ et $\nu_2 = n - 2$ degrés de liberté.

Exemple : la statistique F peut se lire à partir du résumé de la régression linéaire dans `R`. Pour l'ensemble de données sur les carburants, il s'agit de :

```
Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared: 0.8774, Adjusted R-squared: 0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

La valeur critique lorsque $\alpha = 0.05$ est

$$f_{0.05}(1,18) = \texttt{qf(0.95,1,18)} = 4.41.$$

Puisque

$$F^* = 128.9 > f_{0.05}(1,18) = 4.4,$$

nous **rejetons** l'hypothèse nulle H_0 **en faveur** d'une régression significative lorsque $\alpha = 0.05$.

7.6 – Le coefficient de détermination

Le **coefficient de détermination** du modèle d'ajustement linéaire est

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}},$$

où SSE et SST sont définies comme dans l'ANOVA.

Cette valeur représente la **proportion de la variabilité** dans la réponse qui s'explique par le modèle ajusté. Il se situe toujours entre 0 et 1 ; lorsque $R^2 \approx 1$, on consider que l'ajustement est "très bon" (mais c'est relatif).

IL FAUT RESTER VIGILANT : en pratique, R^2 n'est pas toujours le meilleur moyen de déterminer la **adéquation de l'ajustement**. Certains facteurs (tels que le nombre d'observations) peuvent aussi venir affecter le coefficient de détermination.

Exemple : le coefficient de détermination R^2 peut se lire à partir du résumé de la régression linéaire dans R. Pour l'ensemble de données sur les carburants, il s'agit de :

```
Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared: 0.8774, Adjusted R-squared: 0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

Puisque $R^2 = 0.8774$, le modèle d'ajustement linéaire explique environ 88% de la variabilité dans la réponse ... ce n'est pas bien surprenant puisque les données semblent réellement provenir d'un mécanisme linéaire.

Annexe – Résumés et exemples

1. Tracez le nuage de points
2. Trouvez la droite d'ajustement
3. Vérifiez la pertinence d'un ajustement linéaire (coefficient de corrélation, test de la régression significative, etc.)
4. Vérifiez la qualité de l'ajustement ou l'I.C. de la droite
5. Vérifiez les hypothèses du modèle (à l'aide des erreurs résiduelles)
6. Proposez des prédictions, si c'est approprié de le faire

Exemple: les arrestations aux États-Unis

US Arrests contient des statistiques, # d'arrestations par 100,000 résidents, pour divers types de crime en 1973 dans chacun des $n = 50$ états américains.

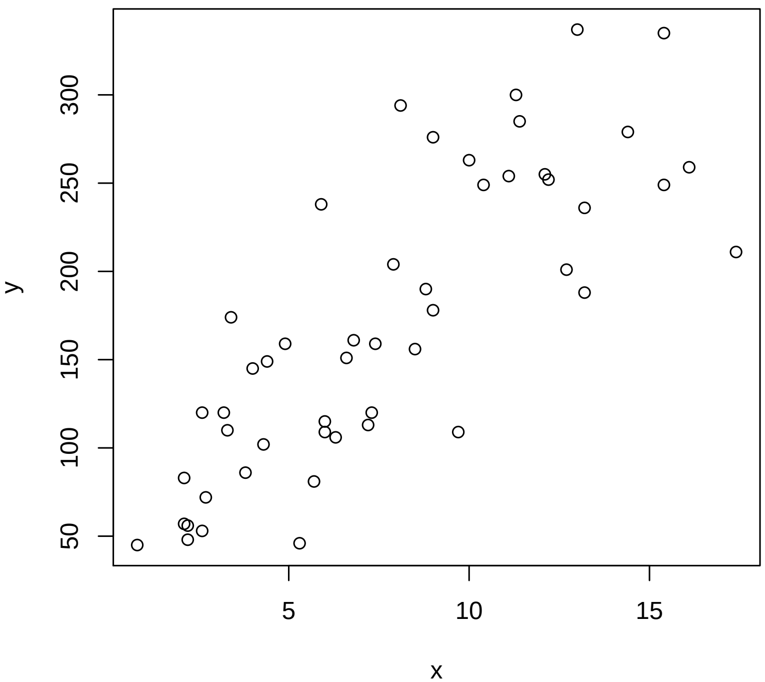
1. La réponse est y (# d'agressions) et le régresseur est x (# de meurtres) pour chacun des 50 états.

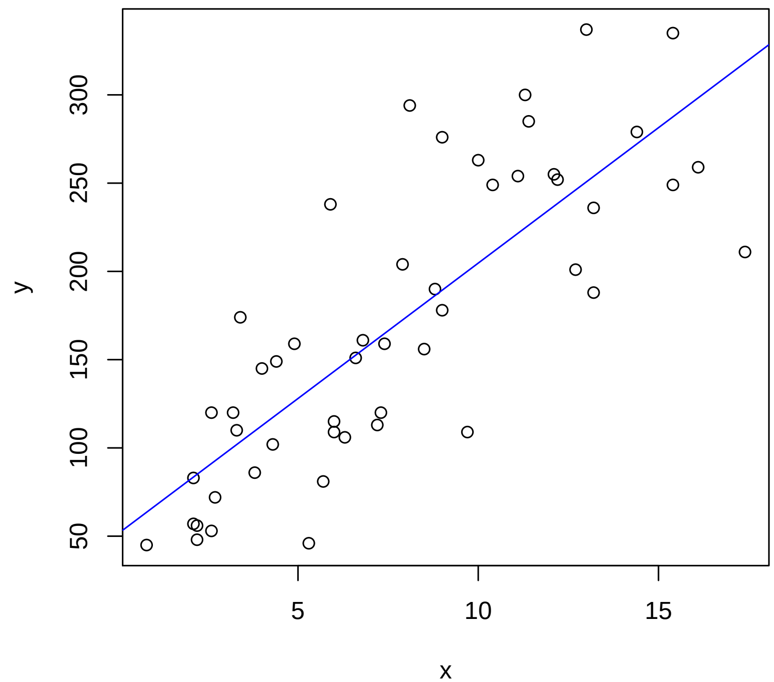
2. Nous avons

$$\sum_{i=1}^{n} x_i = 389.4, \ \sum_{i=1}^{n} y_i = 8538$$

$$\sum_{i=1}^{n} x_i^2 = 3962.2, \ \sum_{i=1}^{n} y_i^2 = 1,798,262, \ \sum_{i=1}^{n} x_i y_i = 80,756.$$

La droite de meilleur ajustement est ainsi $\hat{y} = 51.27 + 15.34x$.





3. La valeur relativement élevée du coefficient de corrélation $\rho = 0.802$ suggère une **relation linéaire entre x et y**. Nous testons

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

la statistique de test est

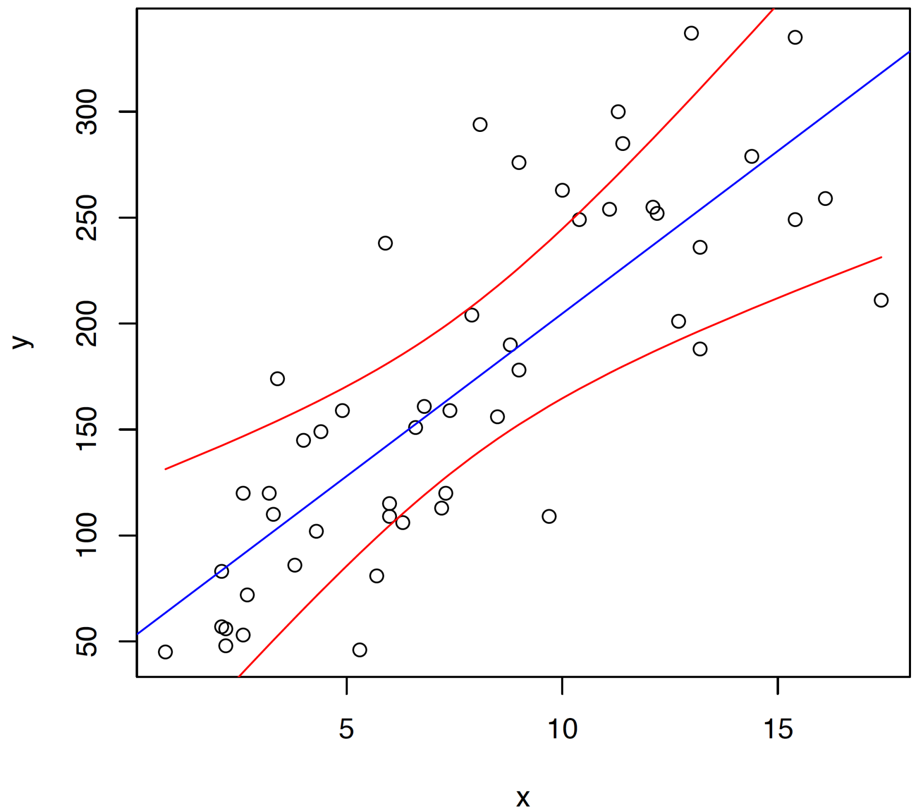
$$T_0 = \frac{b_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}},$$

où $\hat{\sigma}^2 = 2531.73$ and $S_{xx} = 929.55$. Sa valeur observée est $t_0 = 9.30$; puisque

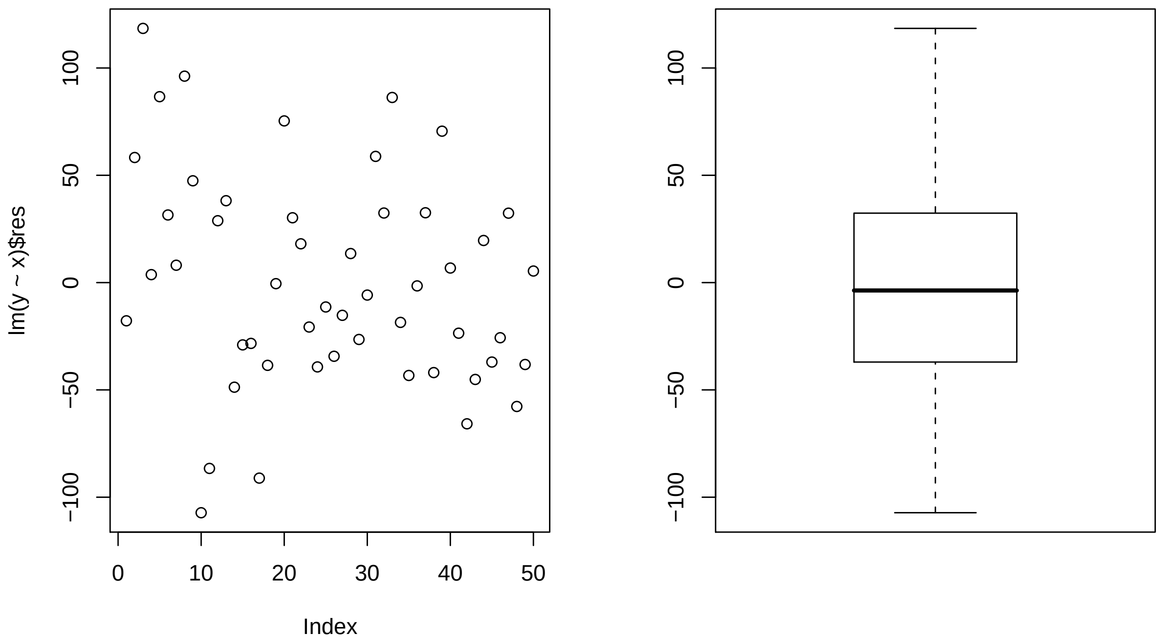
$$t_{0.05/2}(50-2) \approx 2.01 < t_0 = 9.30,$$

nous **rejetons** H_0 en faveur **d'une relation linéaire entre x et y**.

4. L'I.C. de la réponse moyenne à environ 95% est :



5. L'ajustement est **adéquat** car les résidus ne présentent aucune **tendance systématique** : ils sont **uniformément distribués** autour de 0.



6. Comme la régression semble être un bon modèle de la situation, elle pourrait avoir un bon pouvoir prédictif (sur son domaine). Nous pouvons prédire le nombre d'agressions dans un état américain, si son nombre de meurtres est de $x_0 = 20$, par exemple :

$$\hat{y}_0 = 51.27 + 15.34(20) = 358.07.$$

Une façon équivalente de poser la question : chercher une estimation ponctuelle du nombre d'agressions dans un état américain si son nombre de meurtres est de 20.

De même, l'I.P. du # d'assauts dans un état américain si $x_0 = 20$ est :

$$358.07 \pm 2.01\sqrt{2531.73\left[1 + \frac{1}{48} + \frac{(20-7.78)^2}{929.55}\right]} = 358.07 \pm 40.64.$$

Exemple: les données de compagnie aérienne

Ces données mesurent le total mensuel des passagers des compagnies aériennes internationales de 1949 à 1960 (`AirPassengers` dans `R`).

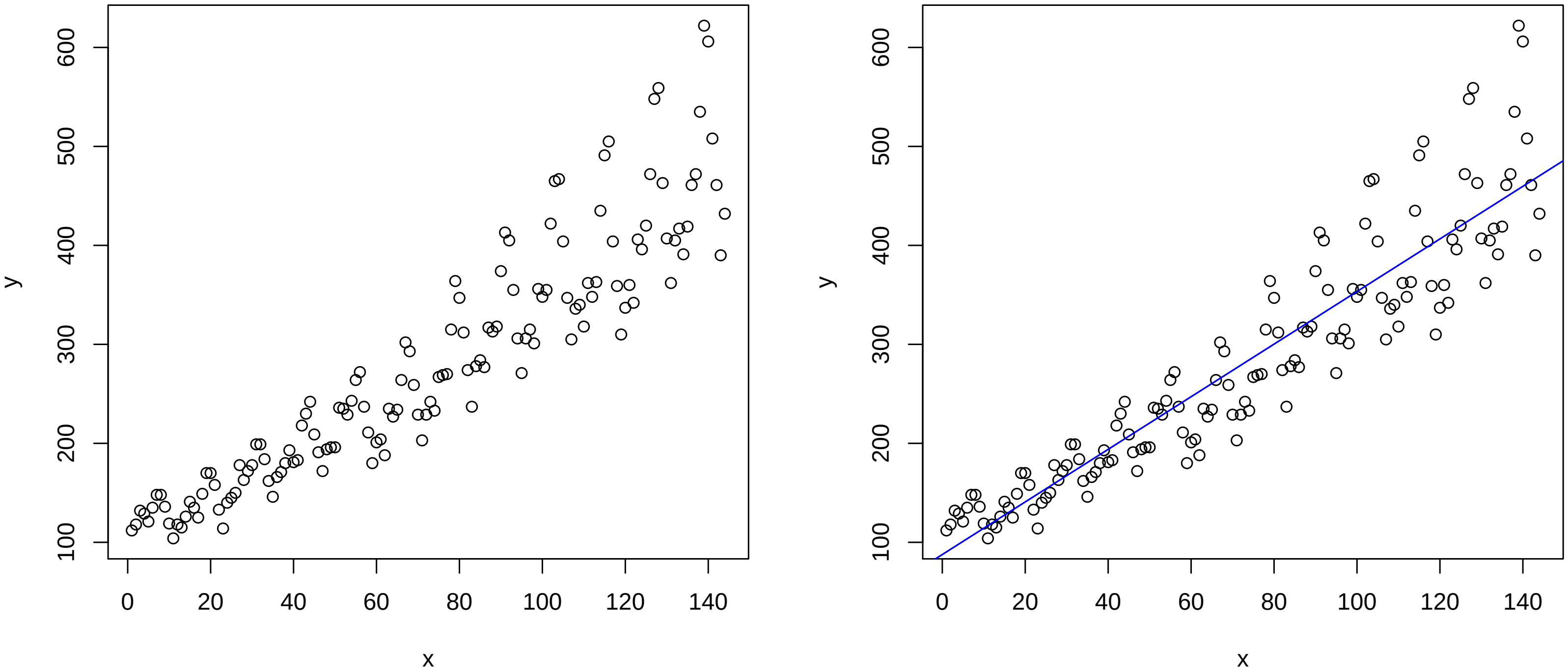
1. La réponse est y (# mensuel de passagers), et le prédicteur est x (# mois depuis le 1 janvier 1949) : $x = (1, 2, \ldots, 144)$.

2. Nous avons

$$\sum_{i=1}^{n} x_i = 10,440, \ \sum_{i=1}^{n} y_i = 40,363$$

$$\sum_{i=1}^{n} x_i^2 = 1,005,720, \ \sum_{i=1}^{n} y_i^2 = 13,371,737, \ \sum_{i=1}^{n} x_i y_i = 3,587,478.$$

La droite de meilleur ajustement est $\hat{y} = 87.653 + 2.657x$.



3. Le coefficient de corrélation est $\rho = 0.924$, ce qui suggère **une forte relation linéaire entre x et y**. Nous testons

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

la statistique de test est

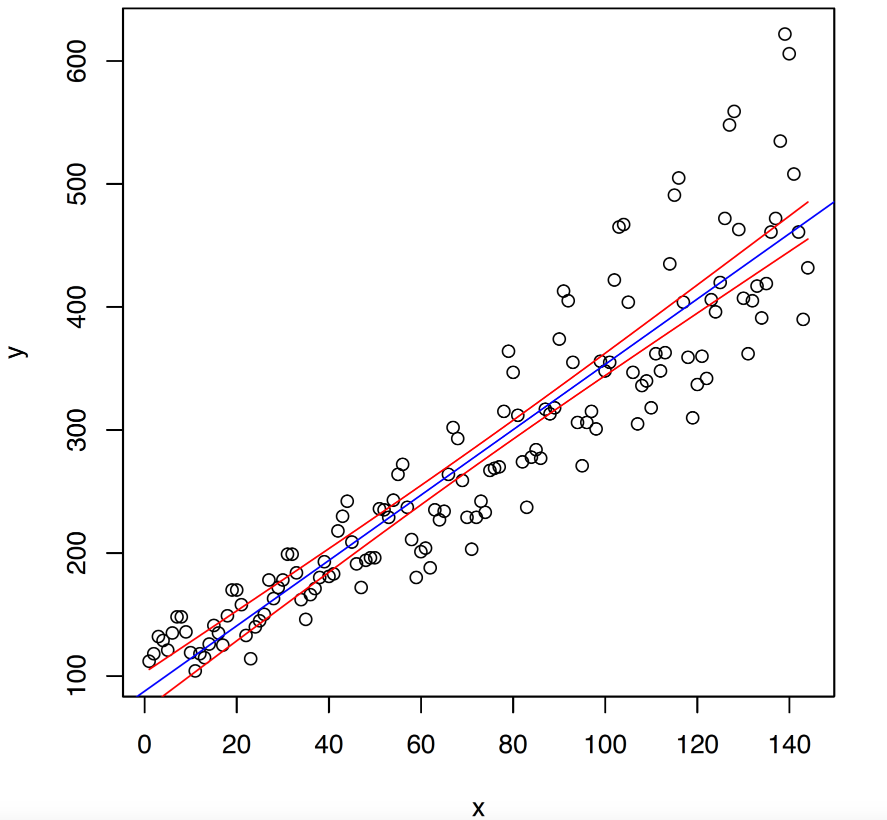
$$T_0 = \frac{b_1 - 0}{\sqrt{\hat{\sigma}^2/S_{xx}}},$$

où $\hat{\sigma}^2 = 2121.261$ et $S_{xx} = 248820$. Nous observons $t_0 = 28.77644$; puisque

$$t_{0.05/2}(144-2) \approx 1.97 < t_0 = 28.78,$$

nous **rejetons** H_0 en faveur d'**une relation linéaire entre x et y**.

4. L'I.C. de la réponse moyenne à environ 95% est :



5. Les erreurs résiduels présentent une **tendance** : la variance de l'erreur **n'est pas constante** et elle **augmente avec** x. Nous devons procéder à des **transformations de données** avant d'utiliser la régression linéaire.