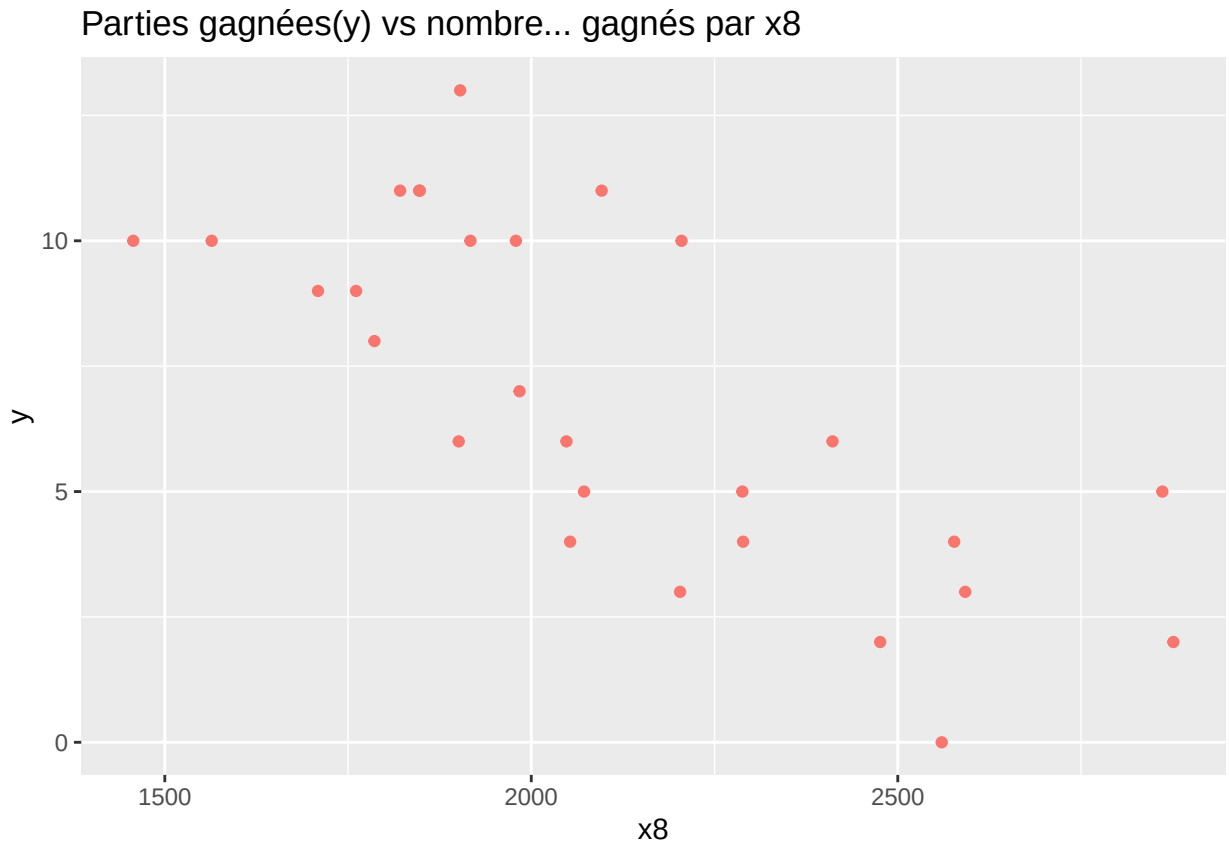
MAT3775–Devoir–1

Nom: Jonathan Domingue Numero etudiant: 300246863

2023-09-29 00:26:12

1 Question 1: Ajuster un modèle de régression linéaire simple reliant les matchs gagnés aux yards gagnés au sol par les adversaires

```
library(ggplot2)
ggplot(table.b1, aes(x = x8, y = y)) +
  geom_point(aes(color="dodgerblue"))+
  theme(legend.position="none")+
  labs( title= "Parties gagnées(y) vs nombre... gagnés par x8",
                  x = "x8", y = "y")
```



1.1 Régression linéaire simple de question 1

```
lm(formule, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset)
```

où :

- formule a la forme **réponse** ~ **termes** où **réponse** est le vecteur de réponse (numérique) **termes** est une série de termes qui spécifie un prédicteur linéaire pour la réponse.
- data est le jeu de données, etc. Pour en savoir plus, utilisez l'aide de R (`?lm()`).

Nous pouvons maintenant faire l'Analyse de régression à l'aide de la commande `lm(...)` dans R

```
#renvoie un objet modèle linéaire, qui est sauvegardé dans `model_fit`
model_fit = lm(formula = y ~ x8, data = table.b1)
model_fit # affiche les coefficients
```

```
##
## Call:
## lm(formula = y ~ x8, data = table.b1)
##
## Coefficients:
## (Intercept)           x8
##   21.788251    -0.007025
```

L'impression de l'objet `modèle linéaire` montre simplement les coefficients de régression estimés. Un rapport plus complet est obtenu par la fonction summary.

On peut donc clairement conclure des résultats que $b_0 = 21.788251$ et que $b_1 = -0.007025$.

Ainsi: $\hat{Y} = 21.788251 - 0.007025X$

1.2 Display the summary of the regression

As shown below we can also display the summary of the regression

```
# fournir des résumés alternatifs d'un ajustement de régression
summary(model_fit)
```

```
##
## Call:
## lm(formula = y ~ x8, data = table.b1)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -3.804 -1.591 -0.647  2.032  4.580
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.788251   2.696233   8.081 1.46e-08 ***
## x8          -0.007025   0.001260  -5.577 7.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 26 degrees of freedom
## Multiple R-squared:  0.5447, Adjusted R-squared:  0.5272
## F-statistic:  31.1 on 1 and 26 DF,  p-value: 7.381e-06
```

On peut dire que l'équation linéaire correspondante est $y = \cdots + \cdots x8$.

1.3 Partie b) Construire le tableau d'analyse de la variance et tester la signification de la régression.

```
## Construction du tableau ANOVA
anova(model_fit)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x8         1 178.09 178.092  31.103 7.381e-06 ***
## Residuals 26 148.87   5.726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.4 c)Intervalles de confiance

```
confint(model_fit,level=0.95)
```

```
##                    2.5 %       97.5 %
## (Intercept) 16.246064040 27.330437725
## x8          -0.009614347 -0.004435854
```

1.5 d)Quel pourcentage de la variabilité totale est expliqué par ce modéle

Pour cette question afin de determiner le pourcentage de variabilite va falloir calculer R^2

$$R^2 = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2}$$

```
summary(model_fit)
```

```
##
## Call:
## lm(formula = y ~ x8, data = table.b1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.804 -1.591 -0.647  2.032  4.580
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.788251   2.696233   8.081 1.46e-08 ***
## x8          -0.007025   0.001260  -5.577 7.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 26 degrees of freedom
## Multiple R-squared:  0.5447, Adjusted R-squared:  0.5272
## F-statistic:  31.1 on 1 and 26 DF,  p-value: 7.381e-06
```

Nous pouvons voir par le output du summary que la valeur de R^2 est de 0.5447.

1.6 e)Trouver une intervalle de confiance de 95% sur le nombre moyen de matches gagnés si les yards au sol de l'adversaire sont limités a 2000 yards

```
new.dat = data.frame(x8 = 2000)
predict(model_fit, newdata = new.dat, interval = 'confidence')
```

```
##       fit      lwr      upr
## 1 7.73805 6.765753 8.710348
```

##Question 2

2 Problème 2.10:

Le poids et la tension artérielle systolique de 26 hommes sélectionnés au hasard dans la tranche d'âge de 25 à 30 ans sont indiqués dans le tableau correspondant à ce problème. Supposons que le poids et la tension artérielle (TA) soient conjointement distribués normalement.

2.1 Partie a) Trouver une droite de régression reliant la pression artérielle systolique au poids normalement

```
y= p2.10$sysbp
x= p2.10$weight
model_fit = lm(y~x)
model_fit
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     69.1044       0.4194
```

Les valeurs obtenus peuvent etre représentés par $b_0 = 69.1044$ et $b_1 = 0.4194$.

Donc: $\hat{Y} = 69.1044 + 0.4194X$.

2.2 Partie b:

$$r = \frac{\sum(X_i - \overline{X})Y_i}{\sqrt{\sum(X_i - \overline{X})^2 \cdot \sum(Y_i - \overline{Y})^2}}$$

```
# On calcule le coéfficient de corrélation
cor(x,y)
```

```
## [1] 0.7734903
```

Nous pouvons voir que le facteur de correlation est: $r = 0.7734903$ Ce qui montre qu'il y a une forte relation lineaire, positive entre les variables X et Y.

3 Partie c) Tester l'hypothèse selon laquelle rho = 0.

```
#On utilise une Lois T comme vu en cours
cor.test(x, y, NULL, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
```

```
## t = 5.9786, df = 24, p-value = 3.591e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5513214 0.8932215
## sample estimates:
##       cor
## 0.7734903
```

nous pouvons voir que la loi t donne une valeur de 5.9786 et la valeur de p est tres petite donc nous pouvons rejetter H_0

3.1 Partie d:

Testons les hypotheses suivantes :

Vu que l'hypothese nulle $H_0 : \rho = 0.6$ nous devrons utiliser le test suivant:

$$z = (arctanh(r) - arctanh(0.6))(26-3)^{1/2} = 1.61$$

Prenons un $\alpha = 0.5$, nous pouvons voir que le test z est inferieur a $z_{\alpha/2} = 1.96$.

On ne peut pas donc rejeter H0 en faveur de H1

3.2 Partie e)

Un 95% interval de confiance ρ est

$$[tanh(arctanh(r) - \frac{z_{0.025}}{\sqrt{n-3}}), tanh(arctanh(r) + \frac{z_{0.025}}{\sqrt{n-3}})]$$

```
cor.test(x, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 5.9786, df = 24, p-value = 3.591e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5513214 0.8932215
## sample estimates:
##       cor
## 0.7734903
```

IC: $[0.5513214; 0.8932215]$

4 Question 3

Considérer les données d'oxydation du méthanol dans le tableau B.20. Le chimiste estime que le rapport entre l'oxygène d'entrée et le méthanol d'entrée contrôle le processus de conversion. Effectuez une analyse approfondie de ces données. Les données confirment-elles la conviction du chimiste

```
x = table.b20$x5
y = table.b20$y

my_linear_model = lm(y~x)
summary(my_linear_model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30.29 -24.15 -16.76  29.42  63.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    21.25      20.99   1.013    0.326
## x               7.80      16.78   0.465    0.648
##
## Residual standard error: 35.76 on 16 degrees of freedom
## Multiple R-squared:  0.01333,    Adjusted R-squared:  -0.04834
## F-statistic: 0.2161 on 1 and 16 DF,  p-value: 0.6483
```

```
cor(x,y)
```

```
## [1] 0.1154477
```

Dans le calcul du coéfficient de corrélation, on peut voire que la valeur est proche de 0 b_1 approximativement 0 Puisque la valeur absolue du coéfficient de corrélation entre les deux variabless (notamment l'oxygene d'entrée et le methanol d'entrée) on peut dire que l'association linéaire entre nos deux variables est faible

Donc on en vient que le chemist etait pas correcte