Analyse d'un dataset

Isadora Bartkowiak, Perla Benhamou, Maxime Bonan, Elyora Trabelsi Raphael Uzan

Introduction

L'objectif de ce projet est d'appréhender des méthodes statistiques simples et complexes. L'intérêt est également de mettre en application les éléments de cours.

La base de données retenue représente les pourboires reçus par un serveur dans un restaurant américain dans les années 1990. Les données ont été récoltées par le serveur pendant une période de 2 mois et demi durant l'année 1994. Ce jeu de données est utilisé à titre d'exemple dans dans l'ouvrage Practical Data Analysis: Case Studies in Business Statistics de Peter G. Bryant, Marlène A. Smith publié en 1995.

La base de données nommée Tips a été récupérée sur le site http://www.info.univ-angers.fr/~gh/Datasets/tips.zip, elle comporte 244 occurrences représentant une addition dans le restaurant. Voici un extrait de notre base:

```
"Bill","Tip","Credit","Guests","Day","Server","PctTip"
10.17,1.83,"n",1,"W","A",18
18.4,2.75,"n",2,"M","B",14.9
11.72,2.28,"y",1,"W","A",19.5
9.2,1.8,"n",1,"W","A",19.6
18.14,4,"n",3,"W","C",22.1
20.87,3.13,"y",2,"W","B",15
25.09,5,"y",2,"R","C",19.9
18.62,3.35,"y",2,"T","A",18
39.75,7.25, »y",2,"W","A",18.2
```

Une addition est caractérisée par 8 variables, 4 quantitatives et 4 qualitatives.

Nom de la variable	**Description**	**Type**	**Valeurs**
IDEN	Identifiant de l'addition.	Quantitative	
TOTBILL	Prix total de l'addition en dollar.	Quantitative discrète	
TIP	Montant du pourboire en dollar.	Quantitative discrète	
SEX	Sexe du client ayant payé l'addition.	Qualitative	{0:Homme, 1:Femme}
SMOKER	Indique si le client est fumeur ou non fumeur.	Qualitative	{0:Non-fumeur, 1:Fumeur}

DAY	Jour de la semaine.	Qualitative	{3 Jeudi, 4 Vendredi, 5 Samedi, 6 Dimanche}
TIME	Période dans la journée	Qualitative	{0:Journée, 1:Soirée}
SIZE	Nombre de personnes.	Quantitative discrète	

Notre choix s'est porté sur cette base en raison de sa simplicité, l'objectif de ce projet étant d'expérimenter une analyse statistique, nous avons préféré partir sur une base et un sujet peu complexe. La base comporte sufisament de variables pour y effectuer des analyses.

Pour un serveur, les pourboires représentent une composante importante de leur rémunération. Il est intéressant pour un directeur de restaurant et pour le moral de son personnel de déterminer quels facteurs peuvent influer sur le montant du pourboire. L'objet de notre analyse part du postulat que le montant du pourboire n'est pas exclusivement déterminé par la qualité du service. La problèmatique générale consiste donc à se demander : Est ce que le montant du pourboire est uniquement motivé ou influencé par la qualité du service ? Quels facteurs peuvent influencer ce montant ?

Est-il vrai que le sexe du client influe sur le montant du pourboire ? L'un des deux sexes aurait-il tendance à donner plus ?

Les stratistiques descriptives nous inquent que les femmes donnent en moyenne 2,83 dollars de pourboire et les hommes 3,09 dollars. On pourrait penser que les hommes ont tendance à donner plus de pourboire que les femmes. Nous allons effectuer un test de student afin de vérifier si cette différence de moyenne est statistiquement significative. On émet l'hypothèse nulle qu'il n'existe pas de différence significative entre les moyennes des deux sexes.

```
> t.test(tips$TIP ~tips$SEX)
data:  tips$TIP by tips$SEX
t = 1.4895, df = 215.71, p-value = 0.1378
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0828057  0.5951448
sample estimates:
mean in group 0 mean in group 1
       3.089618        2.833448
```

Le résultat du test nous indique une p-value de 0.1378, soit environ 13% de chance de se tromper si l'on rejete l'hypothèse nulle. La différence entre les deux groupes n'est donc pas significative. Autrement dit, les résultats observés pour notre échantillon ne nous permettent pas d'affirmer que l'ensemble des hommes donnent plus de pourboire que l'ensemble des femmes.

Est-il vrai que le montant du tip varie en fonction du jour de la semaine ? Les pourboires sont-ils supérieurs un jour en particulier ?

H0 = 3 = 4 =5 = 6
H0 = Mardi = Vendredi = Samedi = Dimanche

Testons la normalité des distributions avec le test de Shapiro Wilk :

```
> shapiro.test(tips$TIP)
Shapiro-Wilk normality test

data:  tips$TIP
W = 0.89781, p-value = 8.2e-12
```

Le résultat n'est pas concluant.

Quel facteur influe le plus sur le montant du pourboire ?

Le montant du pourboire peut être influencé par une multitude de facteur qu'ils soient internes ou externe à notre base, nous allons maintenant nous intéresser à déterminer quel facteur (mesuré) a le plus d'impact sur ce montant. A quelle proportion participe-il au montant du pourboire ? Et enfin quelle est la précision de ce modèle ? Pour tenter de répondre à cette hypothèse nous allons effectuer une regression multiple afin d'expliquer la variable TIP. Cette étude comportera une partie décrivant le déroulement de l'analyse et une autre qui expliquera les résultats obtenus.

La régression linéaire multiple consiste à expliquer une variable Y (ici TIP) par d'autres variables explicatives. Le but de cette régression linéaire est de déterminer la significativité des variables, tout en minimisant le risque d'erreur pour prédire la précision du modèle.

Nous allons dans un premier temps, décrire un modèle linéaire contenant 4 variables explicatives afin de tenter d'expliquer le montant d'un pourboire. Ce modèle est défini par une équation de la forme :

$$TIP = \alpha + b_1 \times SMOKER + b_2 \times SEX + b_3 \times TIME + b_4 \times SIZE$$

Afin d'estimer les paramètres, nous allons décrire un modèle en utilisant la fonction lm() qui va permettre d'ajuster le modèle linéaire.

```
> summary(lm(formula = tips$TIP ~ tips$SMOKER + tips$SEX + tips$TIME + tips
$SIZE))

Residuals:
    Min      1Q  Median      3Q     Max
-2.9335 -0.8038 -0.1065  0.5090  6.4896

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.99752    0.27744   3.595 0.000393 ***
tips$SMOKER  0.19237    0.16108   1.194 0.233569
tips$SEX    -0.09839    0.16536  -0.595 0.552402
tips$TIME    0.18579    0.17736   1.048 0.295909
tips$SIZE    0.71157    0.08292   8.581 1.21e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.208 on 239 degrees of freedom
Multiple R-squared:  0.2501,     Adjusted R-squared:  0.2376
F-statistic: 19.93 on 4 and 239 DF,  p-value: 3.569e-14
```

Ce modèle nous donne l'estimation suivante :

$$TIP = 0.99 + 0.19 \times SMOKER - 0.09 \times SEX - 0.18 \times TIME + 0.71 \times SIZE$$

Nous remarquons que sur ces 4 variables seul le nombre de convives influent sur le pourcentage de pourboire donné avec une p-value faible. Le R carré ajusté mesure le pourcentage de la variance de la variable expliquée par la variance de toutes les variables explicatives. On remarque alors que notre modèle n'est précis qu'à 23%.

Essayons ensuite de décrire un modèle contenant uniquement le nombre de convives lors du repas.

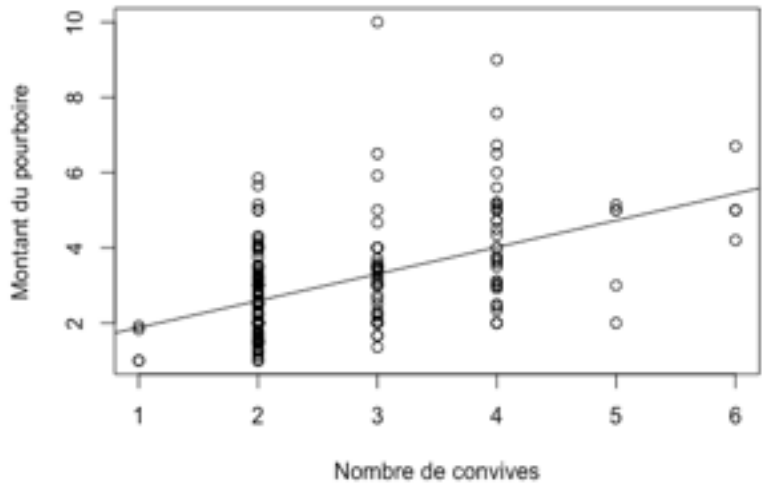
```
> summary(lm(formula = tips$TIP ~ tips$SIZE))

Residuals:
    Min      1Q  Median      3Q     Max
-2.7282 -0.8478 -0.0928  0.5872  6.6954

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.16913    0.22342   5.233 3.61e-07 ***
tips$SIZE    0.71182    0.08156   8.728 4.30e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.209 on 242 degrees of freedom
Multiple R-squared:  0.2394,     Adjusted R-squared:  0.2363
F-statistic: 76.18 on 1 and 242 DF,  p-value: 4.301e-16
```

Ce qui nous donne l'estimation suivante :

$$TIP = 1.16 + 0.71 \times SIZE$$

Notre modèle n'a toujours qu'une précision de 23%, nous pouvons néanmoins contaster que le montant du pourboire augmente en fonction du nombre de convives. Voici une illustration graphique de résultat:

```
> plot(tips$SIZE, tips$TIP, xlab="Nombre de convives", ylab="Montant du
pourboire")
> abline(reg2)
```

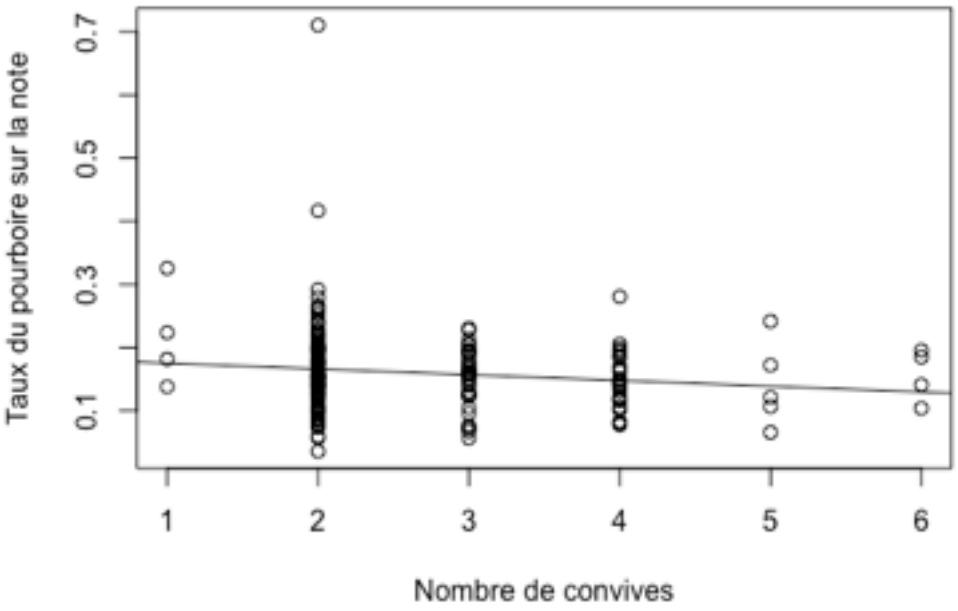


Que se passe t-il si nous effectuons la même démarche mais avec le taux de pourboire ?

```
> summary(lm(formula = tips$PTIP ~ tips$SIZE))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.184375   0.011191  16.475   <2e-16 ***
tips$SIZE   -0.009173   0.004085  -2.245   0.0256 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06057 on 242 degrees of freedom
Multiple R-squared:  0.02041,     Adjusted R-squared:  0.01636
F-statistic: 5.042 on 1 and 242 DF,  p-value: 0.02565
```

Le modèle nous donne l'équation suivante: PTIP = 0.18 - 0.009 x SIZE avec une précision de 0.01 %.

Selon les 2 modèles précédemment décrits, il est intéressant de constater que si le montant du pourboire semble augmenter en ajoutant des convives, la proportion par rapport à la note totale semble quant à elle diminuer.


Taux du pourboire sur la note
0.1
0.3
0.5
0.7
1
2
3
4
5
6
Nombre de convives