

Modèles Linéaires Appliqués / Régression

Régression Logistique: x catégorielle

Arthur Charpentier

UQAM

Hiver 2020 - COVID-19 # 4



Test d'Indépendance de Pearson

On considère deux variables qualitatives

$$x_1 \in \{a_1, \dots, a_I\} \text{ et } x_2 \in \{b_1, \dots, b_J\}$$

et on convertit en un tableau de contingence,

$$n_{i,j} = \sum_k \mathbf{1}(x_{1,k} = a_i) \mathbf{1}(x_{2,k} = b_j)$$

```
1 > loc = "http://freakonometrics.free.fr/titanic.RData"
2 > download.file(loc, "titanic.RData")
3 > load("titanic.RData")
4 > base = base[,1:7]
5 > table(base$Survived, base$Pclass)
```

	1	2	3
0	64	90	270
1	120	83	85

Test d'Indépendance de Pearson

Test : $H_0 : X_1 \perp\!\!\!\perp X_2$, i.e. (cf [définition](#)), $\forall i, j$

$$\mathbb{P}[X_1 = a_i, X_2 = b_j] = \mathbb{P}[X_1 = a_i] \cdot \mathbb{P}[X_2 = b_j]$$

i.e. sous H_0 , on espère avoir

$$\frac{n_{i,j}}{n} \approx \frac{n_{i,\cdot}}{n} \cdot \frac{n_{\cdot,j}}{n} = \frac{n_{i,j}^{\perp}}{n} \text{ où } n_{i,\cdot} = \sum_{j=1}^J n_{i,j}, \quad n_{\cdot,j} = \sum_{i=1}^I n_{i,j}$$

La statistique de test (de Pearson) est

$$Q = \sum_{i,j} \frac{(n_{i,j} - n_{i,j}^{\perp})^2}{n_{i,j}^{\perp}} \sim \chi^2((I-1)(J-1))$$

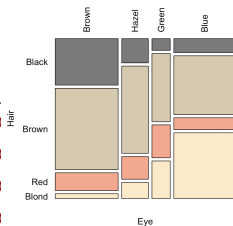
où $u_{i,j} = \frac{n_{i,j} - n_{i,j}^{\perp}}{\sqrt{n_{i,j}^{\perp}}}$ est la contribution du couple (i, j) .

Test d'Indépendance de Pearson

	brown	hazel	green	blue	
black	63.0%	13.9%	4.6%	18.5%	100.0%
brown	41.6%	18.9%	10.1%	29.4%	100.0%
red	36.6%	19.7%	19.7%	23.9%	100.0%
blond	5.5%	7.9%	12.6%	74.0%	100.0%
	37.2%	15.7%	10.8%	36.3%	

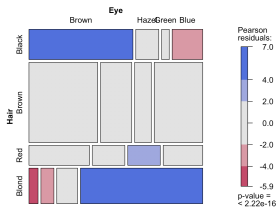


	brown	hazel	green	blue	
black	30.9%	16.1%	7.8%	9.3%	18.2%
brown	54.1%	58.1%	45.3%	39.1%	48.3%
red	11.8%	15.1%	21.9%	7.9%	12.0%
blond	3.2%	10.8%	25.0%	43.7%	21.5%
	100.0%	100.0%	100.0%	100.0%	



Test d'Indépendance de Pearson

	brown	hazel	green	blue	
black	68	15	5	20	108
brown	119	54	29	84	286
red	26	14	14	17	71
blond	7	10	16	94	127
	220	93	64	215	



	brown	hazel	green	blue	
black	40	17	12	39	108
brown	106	45	31	104	286
red	26	11	8	26	71
blond	47	20	14	46	127
	220	93	64	215	

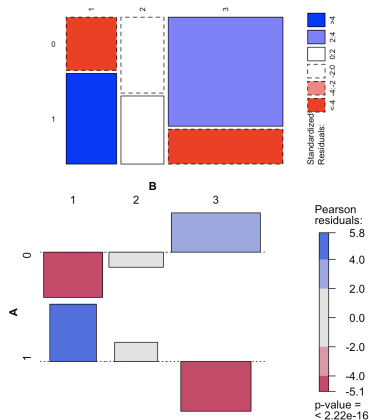
Compare $n_{i,j}$ and $n_{i,j}^\perp$

$$n_{i,j}^\perp = \frac{n_{i,\cdot} \cdot n_{\cdot,j}}{n}$$

Régression sur x_1

$$x_1 \in \{A, B, C\}, \mathbf{x} = (\mathbf{1}_A, \mathbf{1}_B, \mathbf{1}_C)$$

```
1 > (T=table(base$Survived,  
2     base$Pclass))  
3     Pclass  
4 Survived    1    2    3  
5           0   80   97  372  
6           1  136   87  119  
7  
8 > chisq.test(T)  
9  
10 Pearson's Chi-squared test  
11 X-squared = 91.081, df = 2,  
12   p-value < 2.2e-16  
13 > library("graphics")  
14 > mosaicplot(T)  
15 > library("vcd")  
16 > assoc(T)
```



Régression sur x_1

```
1 > with(base, aggregate(Survived, by=list(Pclass), mean))
  $x
2 [1] 0.6296296 0.4728261 0.2423625
```

$$\hat{p}_A = \frac{1}{n_A} \sum_{i:x_i=A} y_i = 62.96\%, \quad \hat{p}_B = \frac{1}{n_B} \sum_{i:x_i=B} y_i = 47.28\%,$$

```
1 > reg = glm(Survived ~ Pclass, family = "binomial",
  data = base)
2 > summary(reg)
3
4 Coefficients:
5             Estimate Std. Error z value Pr(>|z|)
6 (Intercept)   0.5306     0.1409   3.766 0.000166 ***
7 Pclass2      -0.6394     0.2041  -3.133 0.001731 **
8 Pclass3      -1.6704     0.1759  -9.496 < 2e-16 ***
```

Régression sur x_1

$$\hat{p}_A = \frac{1}{n_A} \sum_{i:x_i=A} y_i = 62.96\%, \quad \hat{p}_B = \frac{1}{n_B} \sum_{i:x_i=B} y_i = 47.28\%,$$

```
1 > beta = coefficients(reg)
2 > exp(beta[1]) / (1 + exp(beta[1]))
3 (Intercept)
4 0.6296296
5 > exp(beta[1] + beta[2]) / (1 + exp(beta[1] + beta[2]))
6 (Intercept)
7 0.4728261
```

$$\hat{p}_A = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}}, \quad \hat{p}_B = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}}, \quad \hat{p}_C = \frac{e^{\hat{\beta}_0 + \hat{\beta}_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_2}}$$

Régression sur x_1

$$\sum_{i=1}^n \hat{p}_i = \sum_{i=1}^n y_i \text{ et } \sum_{i:x_i=j} \hat{p}_i = \sum_{i:x_i=j} y_i, \forall j$$

```
1 > sum(predict(reg,type="response"))
2 [1] 342
3 > sum(base$Survived)
4 [1] 342
5 > sum(predict(reg,type="response")[base$Pclass=="1"])
6 [1] 136
7 > sum(base$Survived[base$Pclass=="1"])
8 [1] 136
```

Survie des Passagers du Titanic

y : indicatrice de survie d'un passager du Titanic

```
1 > reg = glm(Survived ~ Sex, family = "binomial", data
  = base)
2 > summary(reg)
3
4 Coefficients:
5             Estimate Std. Error z value Pr(>|z|)
6 (Intercept)   1.0566     0.1290   8.191 2.58e-16 ***
7 Sexmale      -2.5137     0.1672 -15.036 < 2e-16 ***
8 > confint(reg)
9
10              2.5 %      97.5 %
11 (Intercept)  0.8085881  1.314934
12 Sexmale     -2.8465007 -2.190728
```

Survie des Passagers du Titanic

y : indicatrice de survie d'un passager du Titanic

```
1 > with(base, table(Survived, Sex))
2           Sex
3 Survived female male
4           0      81  468
5           1     233  109
6
7 > with(base, chisq.test(table(Survived, Sex)))
8
9      Pearson's Chi-squared test with Yates' continuity
      correction
10
11 data:  table(Survived, Sex)
12 X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Régression & Données Groupées

La régression binomiale sur données catégorielles peut se faire

- sur les données individuelles, $(y_i, x_{1,i}, x_{2,i}, x_{3,i})$, où $y \in \{0, 1\}$
- sur les données groupes, sur tableau de contingence $N_{i,j,k}$

```
1 > reg = glm(Survived ~ Sex+Embarked+Pclass, family = "
  binomial", data = base)
2 > summary(reg)
3
4 Coefficients:
5             Estimate Std. Error z value Pr(>|z|)
6 (Intercept)   2.6394      0.2672   9.877 < 2e-16 ***
7 Sexmale      -2.6081      0.1855 -14.056 < 2e-16 ***
8 EmbarkedQ     -0.1454      0.3626  -0.401  0.68837
9 EmbarkedS     -0.5954      0.2278  -2.613  0.00897 **
10 Pclass2      -0.6691      0.2525  -2.649  0.00806 **
11 Pclass3      -1.8385      0.2247  -8.182 2.78e-16 ***
```

Régression & Données Groupées

```
1 > Y = xtabs(Survived ~ Embarked+Pclass+Sex,data=base)
2 > N = xtabs( (Survived>=0) ~ Embarked+Pclass+Sex,data=
    base)
3 > dfy = as.data.frame(Y)
4 > dfn = as.data.frame(N)
5 > df = data.frame(dfy[,1:3],Y=dfy$Freq,N=dfn$Freq)
6 > tail(df)
7      Embarked Pclass  Sex  Y    N
8 13          C      2 male  2   10
9 14          Q      2 male  0    1
10 15          S      2 male 15   97
11 16          C      3 male 10   43
12 17          Q      3 male  3   39
13 18          S      3 male 34  265
```

où $\tilde{Y}_i \sim \mathcal{B}(n_i, p_i)$, $p_i = \text{logit}^{-1}(\mathbf{x}^\top \boldsymbol{\gamma})$

Régression & Données Groupées

Les deux modèles sont équivalents, $\hat{\gamma} = \hat{\beta}$

```
1 > regg = glm(cbind(Y,N-Y) ~ Sex+Embarked+Pclass, data
2   = df, family=binomial)
3
4 Coefficients:
5 (Intercept)      Sexmale      EmbarkedQ      EmbarkedS
6      2.6394      -2.6081      -0.1454      -0.5954
7      Pclass2      Pclass3
8      -0.6691      -1.8385
9 > reg
10
11 Coefficients:
12 (Intercept)      Sexmale      EmbarkedQ      EmbarkedS
13      2.6394      -2.6081      -0.1454      -0.5954
14      Pclass2      Pclass3
15      -0.6691      -1.8385
```