Association

Evan Misshula

2018-04-26

NYC notifies defendants who have a DAT

	no-warrant-issued	fta
no-call	484	146
called	1954	358

NYC notifies defendants who have a DAT

	no-warrant-issued	fta
no-call	484	146
called	1954	358

- $H_0 : \pi_1 = \pi_2$ vs $H_1 : \pi_1 \neq \pi_2$

Let's dod the computation

```
noCall <- c(484,146)
called <- c(1954,358)
shows <- c(noCall[1],called[1])
prop.test(shows,c(sum(noCall),sum(shows)))


2-sample test for equality of proportions with continuity corre

data:  shows out of c(sum(noCall), sum(shows))
X-squared = 3.1858, df = 1, p-value = 0.07428
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.070777010  0.004331706
sample estimates:
   prop 1    prop 2
0.7682540 0.8014766
```

The chi-squared distribution

- The χ^2-distribution is the distribution of a the sum of n squared Gaussian variables

$$\chi^2 = \sum_{i=1}^{n} Z_i^2 \text{ where } Z_i \text{ is } N(0,1) \text{ i.i.d.}$$

- We call n the degrees of freedom

Am I being cheated?

Am I being cheated?

- What if we toss a dice 150 times and get this . . .

Am I being cheated?

- What if we toss a dice 150 times and get this . . .

face	1	2	3	4	5	6
Count	22	21	22	27	22	36

- Actual

Am I being cheated?

- What if we toss a dice 150 times and get this . . .

face	1	2	3	4	5	6
Count	22	21	22	27	22	36

- Actual

face	1	2	3	4	5	6
Count	25	25	25	25	25	25

- Expected

Am I being cheated?

- What if we toss a dice 150 times and get this . . .

face	1	2	3	4	5	6
Count	22	21	22	27	22	36

- Actual

face	1	2	3	4	5	6
Count	25	25	25	25	25	25

- Expected
- If there is a big discrepancy?

We use the χ^2 test to check goodness of fit

- Goodness of fit checks if something comes from a known population (Verzani, 2004)

We use the χ^2 test to check goodness of fit

- Goodness of fit checks if something comes from a known population (Verzani, 2004)
- Let O_i be the observed count of category i
  - How many 1's,2's,. . . ,6's did we get?

We use the χ^2 test to check goodness of fit

- Goodness of fit checks if something comes from a known population (Verzani, 2004)
- Let O_i be the observed count of category i
  - How many 1's,2's,. . . ,6's did we get?
- Let E_i be the expected count of category i
  - How many 1's,2's,. . . ,6's did we expect?

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Let's do the test in R

```
diceCounts <- c(22,21,22,27,22,36)
myExpProb <- rep(1,6)/6
chisq.test(diceCounts,p=myExpProb)
```

```
Chi-squared test for given probabilities

data:  diceCounts
X-squared = 6.72, df = 5, p-value = 0.2423
```

Association

is a property of two or more variables where a change a deviation from the mean in one makes a deviation of the other more likely.

Positive and negative

- Two variables are *positively associated* if cases with higher/lower than average values of one variable also tend to have higher/lower than average values of the other variable.
- Two variables are *negatively associated* if individuals with higher than average values of one variable tend to have lower than average values of the other variable, and vice versa

Warning

Warning

- People often confuse association with a causal relationship
- This is known formally as *Post hoc propeter hoc*
  - For example in the US:
    - There is a negative association between the amount a person spends in a the number of subsequent years a person survives (Stark, 2018).

Discussion question

1. (True or False) Across countries, there is a strong positive association between per capita dietary sugar intake and various cancers. Therefore, sugar probably causes cancer (Stark, 2018).

Discussion question

1. (True or False) Across countries, there is a strong positive association between per capita dietary sugar intake and various cancers. Therefore, sugar probably causes cancer (Stark, 2018).

2. (True or False) Across countries, there is a strong positive association between per capita tobacco intake and various cancers. Therefore, tobacco probably causes cancer.

Discussion question

1. (True or False) Across countries, there is a strong positive association between per capita dietary sugar intake and various cancers. Therefore, sugar probably causes cancer (Stark, 2018).

2. (True or False) Across countries, there is a strong positive association between per capita tobacco intake and various cancers. Therefore, tobacco probably causes cancer.

- Answer: Even when there is strong evidence that something is true, we need more than association to assert causation.

Correlation definition

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- n is the sample size
- x_i and y_i are the two variables observed
- $\hat{x}$ is the sample standard deviation

correlation properties

- ranges over -1 to 1
- positive correlation $\Rightarrow$ positive association
- negative correlation $\Rightarrow$ negative association

Simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- β_0 and β_1 are estimated from the data

Simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- β_0 and β_1 are estimated from the data
- ϵ_i is the difference between the model and reality

The solution to finding β_0 and β_{1b}

$$b1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \bar{y} = b_0 + b_1 \bar{x}$$

Linear Regression with R

Linear Regression with R

- Dependant variable, y, is a person's max heart rate
- Independant variable x, is a person's age

Linear Regression with R

- Dependant variable, y, is a person's max heart rate
- Independant variable x, is a person's age
- Can we predict Max heart rate from age?

Download heart.R from misshula.org

```
x <- c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37)
y <- c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,
summary(lm(y~x))


Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9258 -2.5383  0.3879  3.1867  6.6242

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
x            -0.79773    0.06996  -11.40 3.85e-08 ***
---
```

plot the regression line

```
myTitle <- paste0("y=",lm(y~x)$coefficients[2],
  "x + ",lm(y~x)$coefficients[1])
plot(x,y, main=myTitle)
abline(lm(y~x))
```

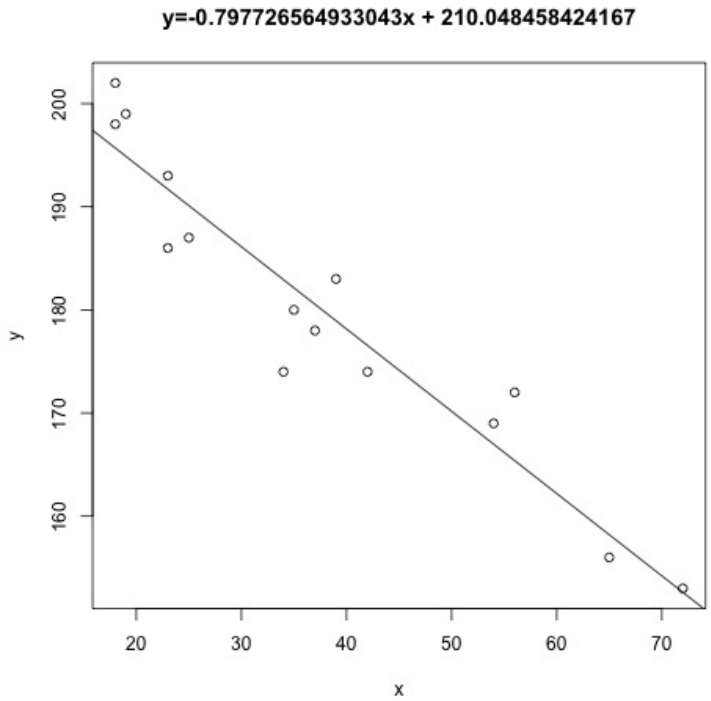
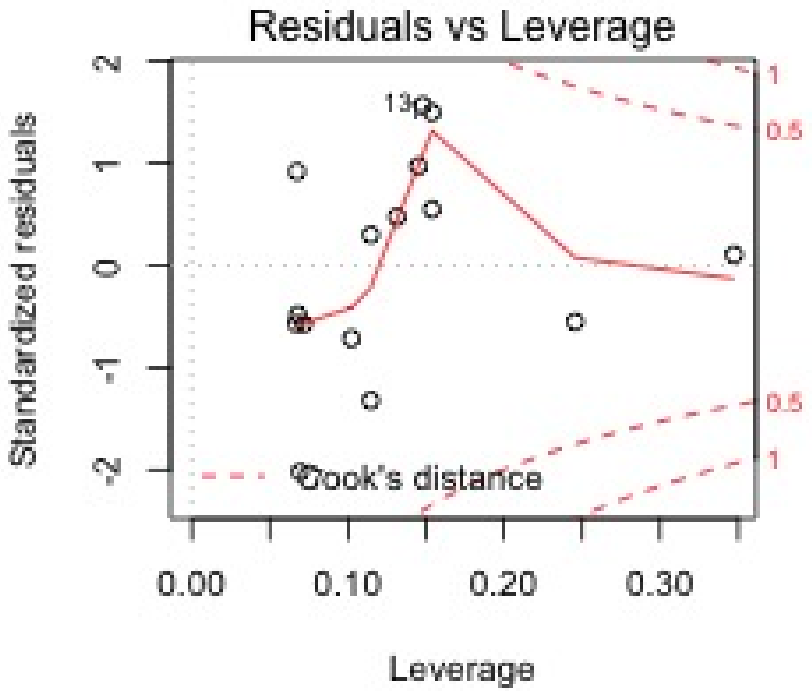


Figure: Heartrate vs age with regression line

Testing the model

```
library(UsingR)
par(mfrow=c(2,2))
lm.result=simple.lm(x,y)
plot(lm.result)
```



What to look for

What to look for

- Residuals v. fitted $\hat{y} vs \hat{\epsilon}$ no trend

What to look for

- Residuals v. fitted $\hat{y} vs \hat{\epsilon}$ no trend
- Normal qq-plot should fall near a straight line

What to look for

- Residuals v. fitted $\hat{y} vs \hat{\epsilon}$ no trend
- Normal qq-plot should fall near a straight line
- Scale-Location where are the biggest errors

What to look for

- Residuals v. fitted $\hat{y} vs \hat{\epsilon}$ no trend
- Normal qq-plot should fall near a straight line
- Scale-Location where are the biggest errors
- Cook's distance Does any point have a big influence

Multiple Linear regression