Regression

Vinish Shrestha

08/27/2021

Summation

$\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$, where n is a positive integer.

- X is a random variable, i.e. flip of a coin
- x_i is a realized outcome

- Say, flip a coin; *head* $= \$1$ and *tail* $= -1$.
  - $n = 5$
  - $x_1 = h,\ x_2 = t,\ x_3 = h,\ x_4 = t,\ x_5 = h$

Then, $\sum_{i=1}^{5} x_i = 1 - 1 + 1 - 1 + 1 = 1$

Some laws about summation

- $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$
- given that a and b are constants
- $\sum_{i=1}^{n}(ax_i + bx_i) = \sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} bx_i$
- $\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} \neq \sum_{i=1}^{n} \frac{x_i}{y_i}$
- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$; (sample) average
- $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}$

Average height of 10 people

```r
set.seed = 1
height = rnorm(10, 5.7, 0.5)
print(round(height, 2))
```

```
##  [1] 5.68 5.73 6.44 5.69 6.01 6.04 5.54 6.32 6.28 5.99
```

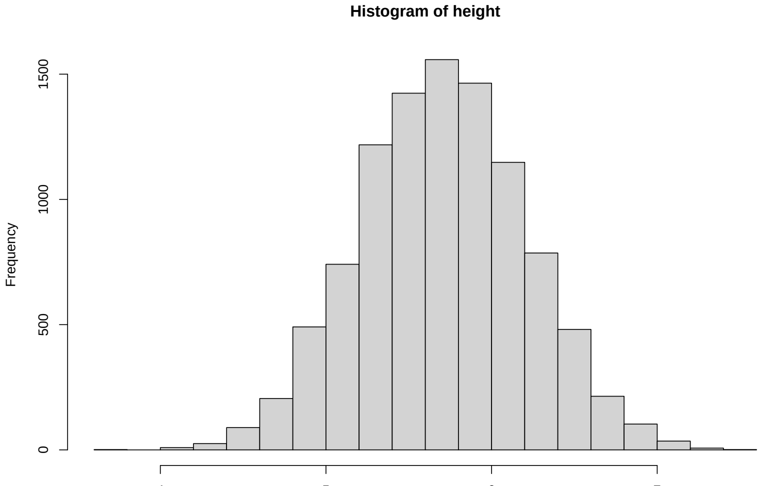
- check the average height of the sample
- is it close to 5.7?
- Why is it not exactly 5.7?

```r
mean(height)
```

```
## [1] 5.97194
```

Large n

```
set.seed = 1
height = rnorm(10000, 5.7, 0.5)
hist(height)
```



Large n

```
mean(height) #for a large n
```

```
## [1] 5.703207
```

Expected value

$$E(X) = p_1x_1 + p_2x_2 + p_3x_3... + p_nx_n$$

where, p_i is probability associated with outcome x_i.

- say, get \$1 if head and -\$1 if tail. What is the expected payoff?

Population vs Sample

- Population: entire group
  - population of this university student body: **all students**
  - you'd fall in it
- Sample: a subset of the population
  - you could either be or not be in the sample
- expectation, $E()$ is a population concept

Additional properties of the expectation operator

Consider two random variables W and H

- $E(aW + b) = aE(W) + b$
- $E(W + H) = E(W) + E(H)$; linear operator
- $E(W - E(W)) = E(W) - E(E(W)) = 0$

- $Variance(W) = \sigma^2 = E(W - E(W))^2$: population concept
- $E[(W^2) - (E(W))^2]$: population concept

- $\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$; sample variance
- $\hat{S} = \frac{1}{n-1} [\sum_{i=1}^{n} (x_i - \bar{x})^2]^{1/2}$; sample standard deviation

How two variables move ..

Very often we are concerned with how variables are related to one another

- Temperature and crime.
- GPA and earnings.
- Mobility and COVID19 cases.

Covariance and correlation describes how variables are *linearly* related to one another.

Covariance

Consider random variables X and Y

- $Cov(X, Y)$
  - $= E[(X - E(X))(Y - E(Y))]$
  - $= E(X)E(Y) - E(XY)$
- 0 covariance does not necessarily mean that X and Y are independent
- But if X and Y are independent, $Cov(X, Y) = 0$.
  - if X and Y are independent, $E(X)E(Y) = E(XY)$
- $\sum_{i=1}^{n} \frac{(x_i - \bar{x}_i)(y_i - \bar{y}_i)}{(n-1)}$; sample covariance
  - $(n - 1)$ is used in the denominator for unbiasedness of the estimator when $E()$ is unknown.

Correlation

- magnitude of covariance difficult to interpret
- instead use correlation
- consider: $W = \frac{X - E(X)}{V(X)}$ and $Z = \frac{Z - E(Z)}{V(Z)}$, normalized to $mean = 0$ and $sd = 1$

- $Corr(W, Z) = Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{(Var(X)Var(y))}}$
  - note that $Cov(a)$, if a is a constant, is zero
  - $E(a) = a$, so $E(a - E(a)) = 0$
- correlation coefficient bounded between -1 and 1
- **Note: Just because two variables lead to a covariance of zero it does not mean that the two variables are independent. These variables can still be related non-linearly. So in this regard, correlation is really a linear concept. This may not be suitable for non-linear analysis.**

Covariance

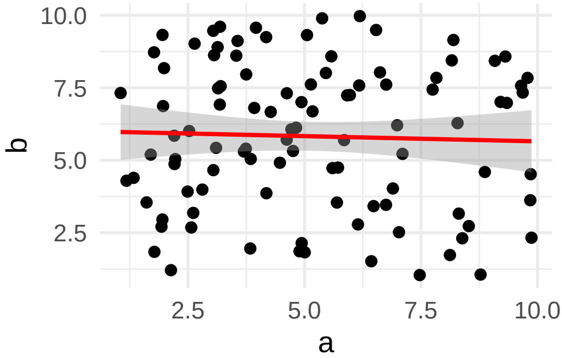
Lets look at some simulations

```
#uniform distribution
set.seed(14825) # allows replicability
a <- runif(100, min = 1, max = 10)
b <- runif(100, min = 1, max = 10)
unrelated <- data.frame(a,b)
```

Covariance

```
library(ggplot2)
ggplot(unrelated, aes(x= a, y=b)) + geom_point() + theme_mi
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Covariance

```r
cov(a, b)
```

```
## [1] -0.226395
```

```r
cor(a,b)
```

```
## [1] -0.03612904
```

- *correlation pretty close to zero!*

Regression

Let's start with a population model

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

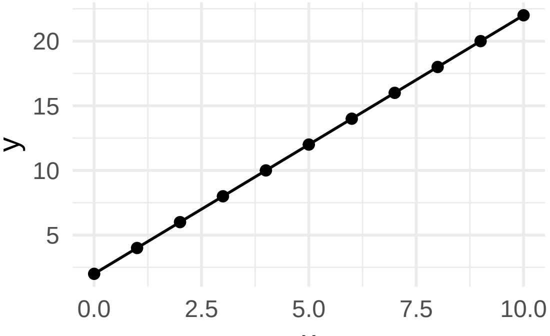
where,

- subscript i is the unit (person, state)
- y is the dependent variable, x is independent variable
- β_0 is the y-intercept
- β_1 is the coefficient (of interest)
- u is the error term; random element

Let's compare with

$y = mx + c$

```
x <- seq(0,10)
c <- 2
m <- 2
y <- m*x + c
eqline <- data.frame(x, y)
ggplot(eqline, aes(x = x, y = y)) + geom_point() + geom_li
```



Let's compare with

- No random element – given a value of x, if you know m and c you can perfectly figure out the value of y

Regression

- In the regression specification u introduces randomness
- This means that there are other factors than x which influences y

Note that: $y_i = \beta_0 + \beta_1 x_i + u_i$

defines a population concept. We now want to empirically estimate β_0 and β_1

- By construction, all other factors that determine y except x are thrown into u
  - think of u as a trash can

Assumptions for estimation

- We need some assumptions:

1) $E(u) = 0$; having β_0 (intercept) in the specification allows this
2) $E(u|x) = E(u)$ (Mean Independence)
    - implies that $E(ux) = 0$
    - think of $P(A|B) = \frac{P(A \cap B)}{P(B)}$

What does 2) imply?

Mean independence

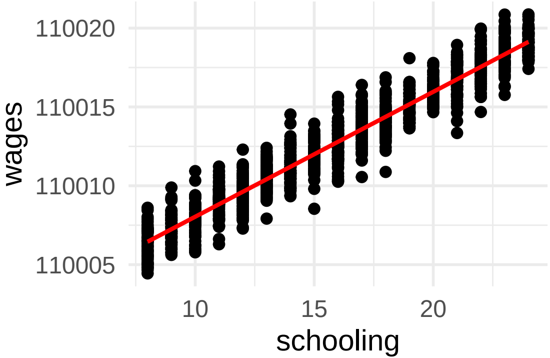
- think of relationship between schooling and wages

```
schooling = seq(8, 24, 1)
#get 1000 observations from the pot with replacement
schooling = sample(schooling, 1000, replace = T)
error <- rnorm(1000, mean = 0, sd = 1)
wages = 110000 + 0.8*schooling + error
schooling <- data.frame(schooling, wages)
```

Mean independence

```
ggplot(schooling, aes(x=schooling, y=wages)) + geom_point(
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Mean independence

- $E(u|x) = E(u)$ means that at every slice of schooling expectation of the error term is the same
- *ability* falls under u
  - So, $E(ability|schooling = 10) = E(ability|schooling = 16) = E(ability|schooling = 20)$
- But if people choose schooling based on their ability, the assumption that $E(u|x) = E(u)$ may be violated
  - Related to concept: **Correlation is not causality**
  - A point which will be addressed in later lectures

Use two assumptions

- To find the estimates of β_0 and β_1, use two assumptions:

1) $E(y - \beta_0 - \beta_1 x) = 0$
2) $E[x(y - \beta_0 - \beta_1 x)] = 0$

There are two equations and two unknowns ($\beta_0\ and \beta_1$). First setup sample counterparts:

a) $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

b) $\frac{1}{n}\sum_{i=1}^{n} x(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

Use two assumptions for Ordinary Least Square

Solve for β_0 from equation a)

- $\frac{1}{n}\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_0 - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1 x_i) = 0$
- $\bar{y}_i - \hat{\beta}_0 - \beta_1 \hat{x}_i = 0$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Replace $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ in equation b)

- $\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) = 0$
- $\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} x_i(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})$
- $\sum_{i=1}^{n}(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})$
- $\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2$
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $\hat{\beta}_1 = \frac{Sample\ Cov(X,Y)}{Sample\ Var(X,Y)}$
- OLS estimator

Next

- replace $\hat{\beta}_1$ in $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ to find $\hat{\beta}_0$
- The fitted value is given as $\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$
- The residual is written as $\hat{u}_i = y_i - \hat{\beta}_1 x_i - \hat{\beta}_0$
- Sum of the squares of residuals (SSR) $\sum_{i=1}^{n} \hat{u}_i{}^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)^2$

Our goal is to obtain estimates of β_0 and β_1 such that it minimizes SSR. Will yield same result as before.

Consider a short simulation

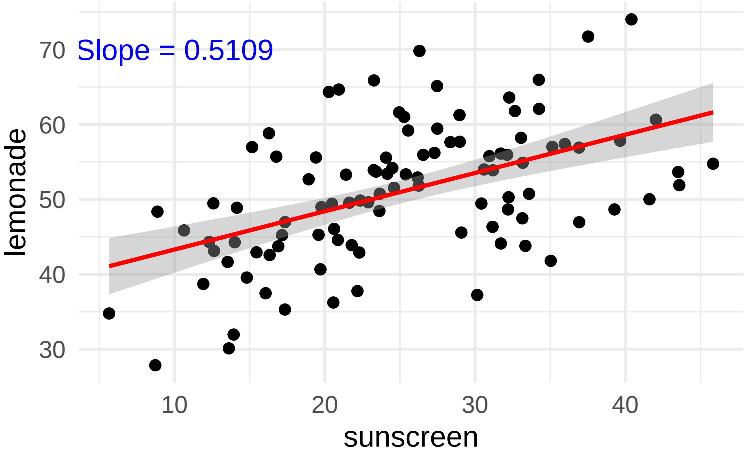
```
set.seed(1)
#simulate quantity demanded of lemonades
lemonade = rnorm(100, 50, 10)
error = rnorm(100, 0, 1)
sunscreen = 1/2*lemonade + 8*error #quantity demanded for s
data <- data.frame(cbind(lemonade, sunscreen))
reg1 <- lm(sunscreen ~ lemonade, data)
reg2 <- lm(lemonade ~ sunscreen, data)
```

- lm is linear regression model; sunscreen $\sim$ lemonade is formula, and data is the dataframe

Sunscreen and lemonade

```
library(ggplot2)
ggplot(data, aes(x=sunscreen, y = lemonade)) + geom_point()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Sunscreen and lemonade

```r
reg2
```

```
##
## Call:
## lm(formula = lemonade ~ sunscreen, data = data)
##
## Coefficients:
## (Intercept)    sunscreen
##     38.1930       0.5109
```

```r
coef(summary(reg2))
```

```
##               Estimate Std. Error   t value     Pr(>|t|
## (Intercept) 38.1930197  2.3589577 16.190633 1.886954e-2
## sunscreen    0.5108893  0.0882082  5.791858 8.421458e-0
```

Next

- Correlation does not mean causality.