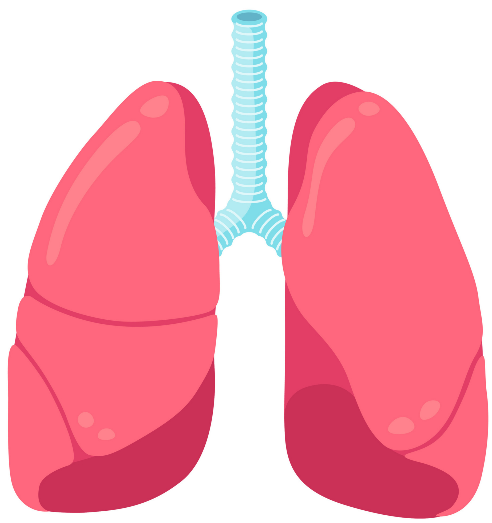
STAT 331 – Lecture 5 (Data analysis)

Here we study a data set that studies the lung function in children and teens. The data is taken from Kahn, Michael (2005). "An Exhalent Problem for Teaching Statistics", The Journal of Statistical Education, 13(2).



Consider a data set from $n = 655$ children between 3 and 19 years old. The variables in the data set include Forced Exhalation Volume (`FEV`) (the response variable), which is a measure of the amount of air an individual can forcibly exhale from their lungs, and `age` (the explanatory variable) in years. Other explanatory variables collected also include `ht` (height in inches), `sex` (1 = male, 0 = female) and `smoke` (1 = yes, 0 = no).

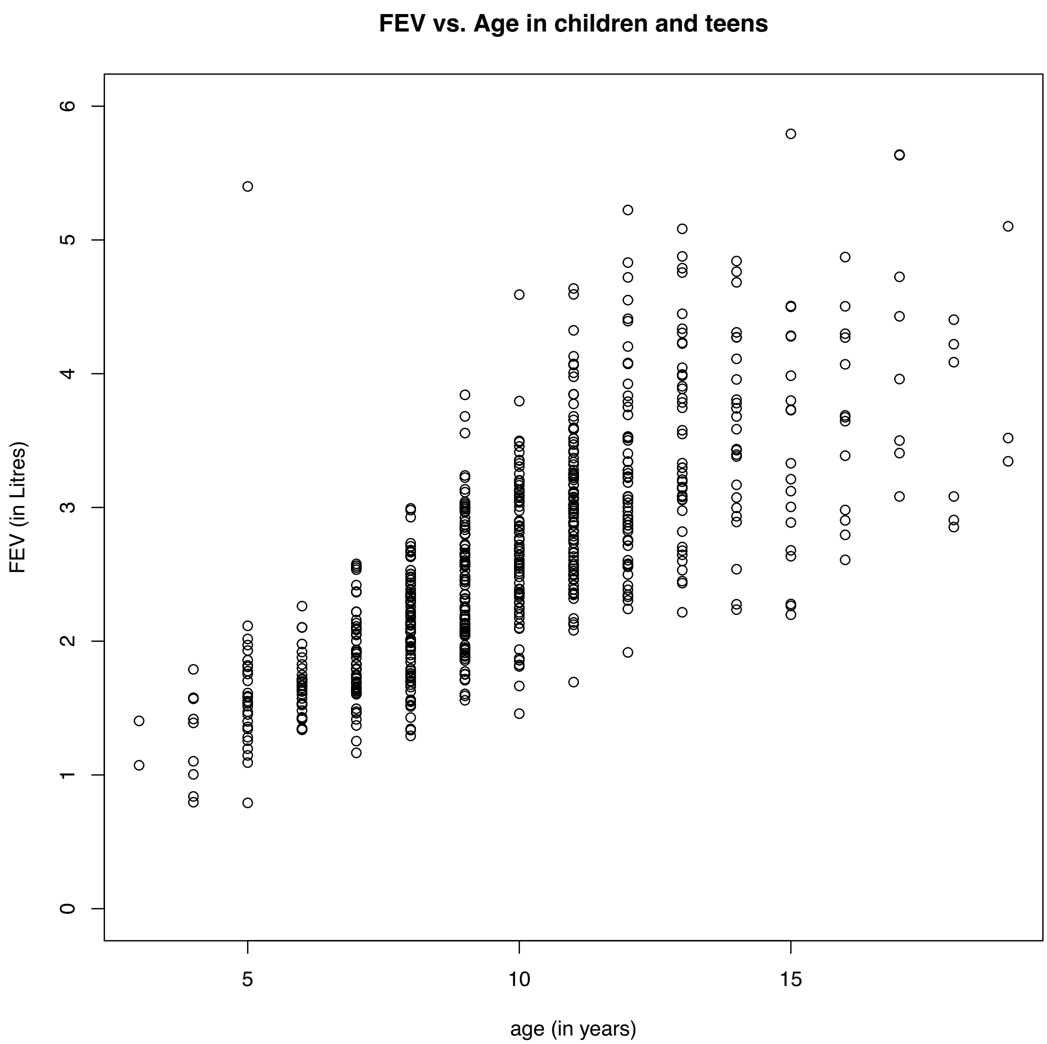
1 Read and view data from csv file

```
> lungdat = read.csv("lung_dat.csv", header=T)
> head(lungdat) ## View only the first 6 rows of data

  age   FEV   ht sex smoke
1   9 1.708 57.0   0     0
2   8 1.724 67.5   0     0
3   7 1.720 54.5   0     0
4   9 1.558 53.0   1     0
5   9 1.895 57.0   1     0
6   8 2.336 61.0   0     0
```

2 Question: What is the association between age and FEV?

```
> plot(lungdat$age, lungdat$FEV, xlab = "age (in years)", ylab = "FEV (in Litres)",
  ylim=c(0,6), main = "FEV vs. Age in children and teens")
```



Trend seems linear

2.1 Fit a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, 655, \quad \epsilon_i \sim N(0, \sigma^2) \text{ iid}$$

```
> myfit = lm(FEV ~ age, data = lungdat)
> summary(myfit) ## Shows a summary of our fitted model

Call:
lm(formula = FEV ~ age, data = lungdat)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5545 -0.3578 -0.0601  0.3182  3.8359

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.470482   0.080314   5.858 7.43e-09 ***
age         0.218721   0.007756  28.199  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5867 on 653 degrees of freedom
Multiple R-squared:  0.5491,    Adjusted R-squared:  0.5484
F-statistic: 795.2 on 1 and 653 DF,  p-value: < 2.2e-16

> n = nrow(lungdat)
> qt (0.975 , n-2)
[1] 1.963603
```

may be blocked out

From the table,

$\hat{\beta}_0 = 0.470$, $SE(\hat{\beta}_0) = 0.080$

$\hat{\beta}_1 = 0.219$, $SE(\hat{\beta}_1) = 0.008$

$\hat{\sigma}^2 = 0.5867^2 = 0.3442$

$\therefore$ Line of best fit:

$\hat{\mu} = 0.470 + 0.219x$,

where x is age (in yrs).

To answer the question if there is a linear relationship between FEV and age, we test the hypothesis $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$.

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.219}{0.008} = 28.200$$

$\therefore |t| >> t_{0.975, 653} = 1.96$

$\left(p\text{-value} = 2P(T \geq 28.200) \approx 0\right)$

$\therefore$ Reject H_0. Conclude a significant relationship.

Intepretation $\hat{\beta}_1$: For one year increase in a child's age, on average their FEV increases by 0.219 Litres.

3 Q: What is the average FEV (in L) for a 5 year old child? Give a 95% confidence interval

```
> myfit = lm(FEV ~ age, data = lungdat)
> summary(myfit) ## Shows a summary of our fitted model

Call:
lm(formula = FEV ~ age, data = lungdat)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5545 -0.3578 -0.0601  0.3182  3.8359

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.470482   0.080314   5.858 7.43e-09 ***
age         0.218721   0.007756  28.199  < 2e-16 ***
---
...
> sum(myfit$residuals^2)
[1] 224.8002

> xbar = mean(lungdat$age); Sxx = sum( (lungdat$age - xbar)^2 )
> xbar; Sxx
[1] 9.923664
[1] 5722.183

> n = nrow(lungdat)
> qt (0.975 , n-2)
[1] 1.963603
```

$\hat{\sigma} \left(= \sqrt{\sum_{i=1}^{n} e_i^2 / n-2} \right)$

or from table on page 3: "residual standard error" $= 0.5867$

For a child who is 5 years old, on average we have

$$\hat{\mu} = 0.470 + 0.219(5) = 1.56$$

$$SE(\hat{\mu}) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(5-\bar{x})^2}{Sxx} \right)} = 0.0445$$

$\therefore$ 95% CI for mean FEV for $x = 5$ is

$$\hat{\mu} \pm t_{0.975,\,653}\, SE(\hat{\mu}) = (1.48,\ 1.65)$$

We can also predict using R:

```
> predict(object = myfit, newdata = data.frame(age = 5),
  interval = "confidence", level = 0.95)
       fit      lwr      upr
1 1.564089 1.476624 1.651553
```

$\underbrace{1.564089}_{\hat{\mu}}$ $\underbrace{(1.476624\ \ 1.651553)}_{95\%\ CI.}$

4 Q: What is the predicted FEV (in L) for newly observed child who is 10 year old? Give a 95% prediction interval

```
> myfit = lm(FEV ~ age, data = lungdat)
> summary(myfit) ## Shows a summary of our fitted model

Call:
lm(formula = FEV ~ age, data = lungdat)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5545 -0.3578 -0.0601  0.3182  3.8359

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.470482   0.080314   5.858 7.43e-09 ***
age         0.218721   0.007756  28.199  < 2e-16 ***
---
...
> sum(myfit$residuals^2)
[1] 224.8002
```

$\hat{\sigma} \left(= \sqrt{\sum_{i=1}^{n} e_i^2 / n-2} \right)$

or from table on page 3: residual standard error $= 0.5867$

```
> xbar = mean(lungdat$age); Sxx = sum( (lungdat$age - xbar)^2 )
> xbar; Sxx
[1] 9.923664
[1] 5722.183

> n = nrow(lungdat)
> qt (0.975 , n-2)
[1] 1.963603
```

For a new observation with $x_0 = 10$:

$$\hat{y}_0 = 0.470 + 0.219(10) = 2.66.$$

$$SE(Y_0 - \hat{Y}_0) = \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(10-\bar{x})^2}{Sxx}\right)} = 0.587$$

$\therefore$ 95% PI for new observation w/ $x_0 = 10$ is

$$\hat{y}_0 \pm t_{0.975,653} \, SE(\hat{y}_0) = (1.50, 3.81)$$

We can also predict using R:

```
> predict(object = myfit, newdata = data.frame(age = 10),
  interval = "prediction", level = 0.95)
       fit      lwr      upr
1 2.657695 1.504701 3.810689
```

What if we wanted a 90% PI?

Instead of $t_{0.975,653}$, we need $\underset{\displaystyle 1.647}{\underset{\parallel}{t_{0.95,653}}}$ as critical value.