DSCI561: Regression I

Lecture 2: November 20, 2017

Gabriela Cohen Freue
Department of Statistics, UBC

In Lect 1: two-sample t-test *vs* ANOVA

- In both cases we are interested in studying the expected number of runs per game
  - **Quantitative response variable**: tax assessment value for a property built in period i: Y_i
  - **Population mean**: $\mu_i = E[Y_i]$
- **two-sample t-test:** compares the means of two populations (groups)

$$H_0 : \mu_C = \mu_M$$

- **one-way ANOVA**: compares the means of *K* groups

  1 factor, K levels)
  - **1 factor**: age period; **K=3 levels**: C, M, O $\quad H_0 : \mu_C = \mu_M = \mu_O$
  - **K=2**: it is equivalent to a two-sample t-test

- ANOVA: Study the effect of one or more ***qualitative variables (factors)*** on a ***quantitative variable (response)***:
  - **Quantitative response**: tax property assessment value

Two-sample t-test as a special case of ANOVA

```r
#t-test vs ANOVA
#responses within each group
tax.M <-dat.small %>% subset(age_factor =="M", select=assessment_k)
tax.C <-dat.small %>% subset(age_factor =="C", select=assessment_k)

t.test(tax.M,tax.C,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  tax.M and tax.C
## t = -2.2034, df = 18, p-value = 0.04083
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -180.111457   -4.288543
## sample estimates:
## mean of x mean of y
##     428.8     521.0
```

$(-2.2034)^2 = 4.855$

```r
#subset of 2 age periods
summary(aov(assessment_k~age_factor,data=subset(dat.small,age_factor %in% c("M","C"))))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## age_factor   1  31878   31878   4.855 0.0408 *
## Residuals   18 118188    6566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#t-test vs ANOVA
#responses within each group
tax.M <-dat.small %>% subset(age_factor =="M", select=assessment_k)
tax.C <-dat.small %>% subset(age_factor =="C", select=assessment_k)

t.test(tax.M,tax.C,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  tax.M and tax.C
## t = -2.2034, df = 18, p-value = 0.04083
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -180.111457   -4.288543
## sample estimates:
## mean of x mean of y
##     428.8     521.0
```

$$H_0 : \mu_C = \mu_M$$

same test

```r
#subset of 2 age periods
summary(aov(assessment_k~age_factor,data=subset(dat.small,age_factor %in% c("M","C"))))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## age_factor   1  31878   31878   4.855 0.0408 *
## Residuals   18 118188    6566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA as a special case of Regression

Two groups

```
summary(lm(assessment_k~age_factor,data=subset(dat.small,age_factor %in% c("M","C"))))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor, data = subset(dat.small,
##     age_factor %in% c("M", "C")))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -178.80  -17.55  -10.90   39.45  197.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   521.00      36.24  14.377 2.62e-11 ***
## age_factorM   -92.20      41.84  -2.203   0.0408 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.03 on 18 degrees of freedom
## Multiple R-squared:  0.2124, Adjusted R-squared:  0.1687
## F-statistic: 4.855 on 1 and 18 DF,  p-value: 0.04083
```

same as t-test

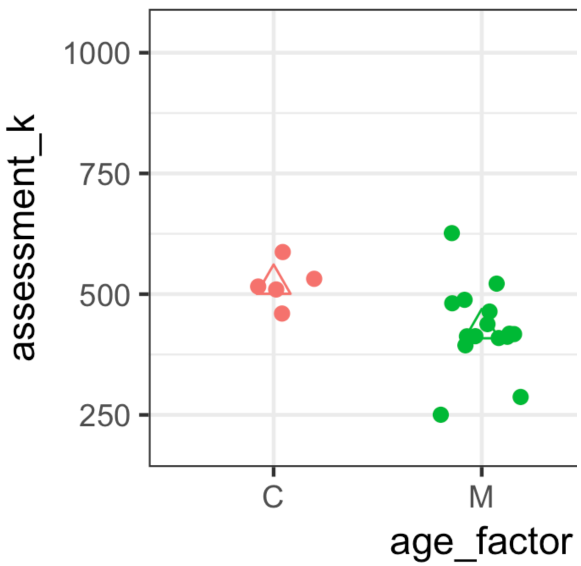
same as ANOVA

More than 2 groups

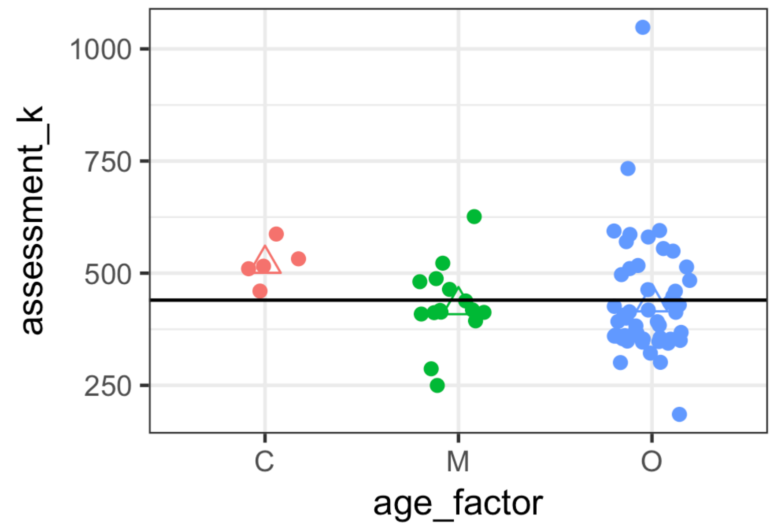
```
#More than 2 groups

#LM with 3 age periods
summary(lm(assessment_k~age_factor,data=dat.small))

##
## Call:
## lm(formula = assessment_k ~ age_factor, data = dat.small)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -250.14   -74.89   -16.97    51.36   612.86
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   521.00      54.10   9.631 2.87e-14 ***
## age_factorM   -92.20      62.46  -1.476    0.145
## age_factorO   -85.86      56.74  -1.513    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121 on 67 degrees of freedom
## Multiple R-squared:  0.0353, Adjusted R-squared:  0.006498
## F-statistic: 1.226 on 2 and 67 DF,  p-value: 0.3001
#ANOVA with 3 age periods
summary(aov(assessment_k~age_factor,data=dat.small))

##             Df Sum Sq Mean Sq F value Pr(>F)
## age_factor   2  35867   17934   1.226    0.3
## Residuals   67 980328   14632
```

different from t-test!!

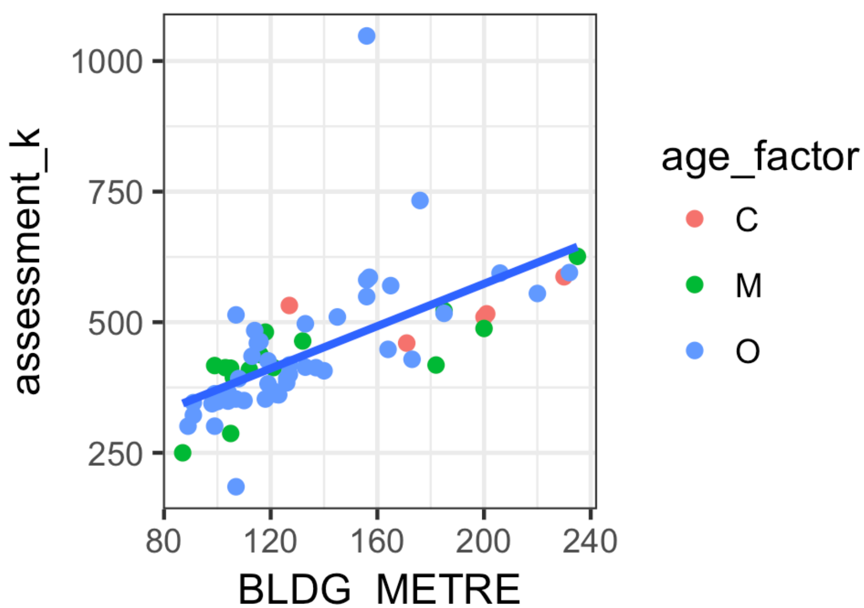
same test



two-samples t-test: 2 groups



one-way ANOVA: more than 2 groups



Linear regression: quantitative and qualitative explanatory variables

In today's lecture

- Comparison with the output of the `lm()` function in R
- Review of linear algebra operations
- Review the mathematical notation of a linear model and its connection with the R-code
- Build the matrix notation of a linear model

Some linear algebra…

Sum of matrices

- Let **A** and **B** be n x m matrices (n rows, m columns)
- **A+B** is an n x m matrix with ijth element equal to $a_{ij} + b_{ij}$

$$\underbrace{\begin{bmatrix} \boxed{a_{11}} & a_{12} \\ a_{21} & a_{22} \\ \vdots & \vdots \\ a_{m1} & a_{m2} \end{bmatrix}}_{\mathbf{A}} + \underbrace{\begin{bmatrix} \boxed{b_{11}} & b_{12} \\ b_{21} & b_{22} \\ \vdots & \vdots \\ b_{m1} & b_{m2} \end{bmatrix}}_{\mathbf{B}} = \underbrace{\begin{bmatrix} \boxed{a_{11} + b_{11}} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \\ \vdots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} \end{bmatrix}}_{\mathbf{A+B}}$$

$$\begin{bmatrix} 1 & 0 & 3 & -1 \\ 0 & 1 & 1 & -1 \\ 0 & \boxed{4} & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & -1 & 1 \\ 2 & 0 & 0 & 0 \\ 0 & \boxed{-2} & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 2 & 1 & 1 & -1 \\ 0 & \boxed{2} & 0 & 0 \end{bmatrix}$$

http://en.wikipedia.org/wiki/Matrix_addition

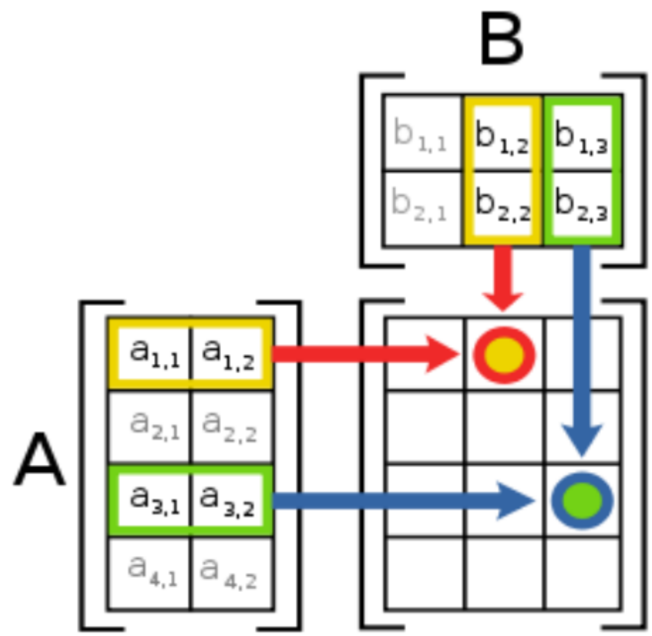
Multiplication of matrices

- If $\mathbf{A}$ is an n x m matrix, $\mathbf{AB}$ is defined only if $\mathbf{B}$ has m rows (number of columns in $\mathbf{B}$ doesn't matter)

- $\mathbf{AB}$ is an n x m matrix

Dot product

$$\boldsymbol{a} \cdot \boldsymbol{b} = \sum a_i b_i$$



http://en.wikipedia.org/wiki/Matrix_multiplication

Example

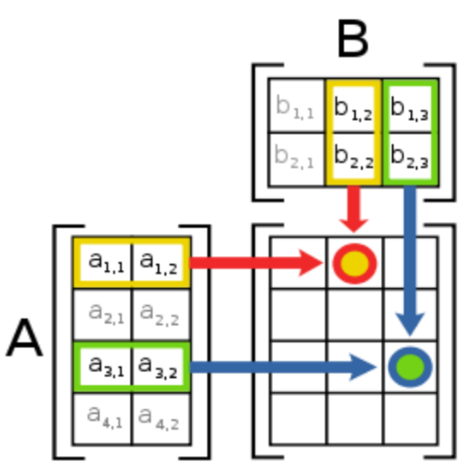
$$\begin{bmatrix} \color{red}{1} & \color{red}{2} \\ 1 & 4 \\ 0 & 5 \end{bmatrix} * \begin{bmatrix} \color{red}{-1} & -2 \\ \color{red}{-2} & 1 \end{bmatrix} = \begin{bmatrix} \color{red}{-5} & 0 \\ -9 & 2 \\ -10 & 5 \end{bmatrix}$$

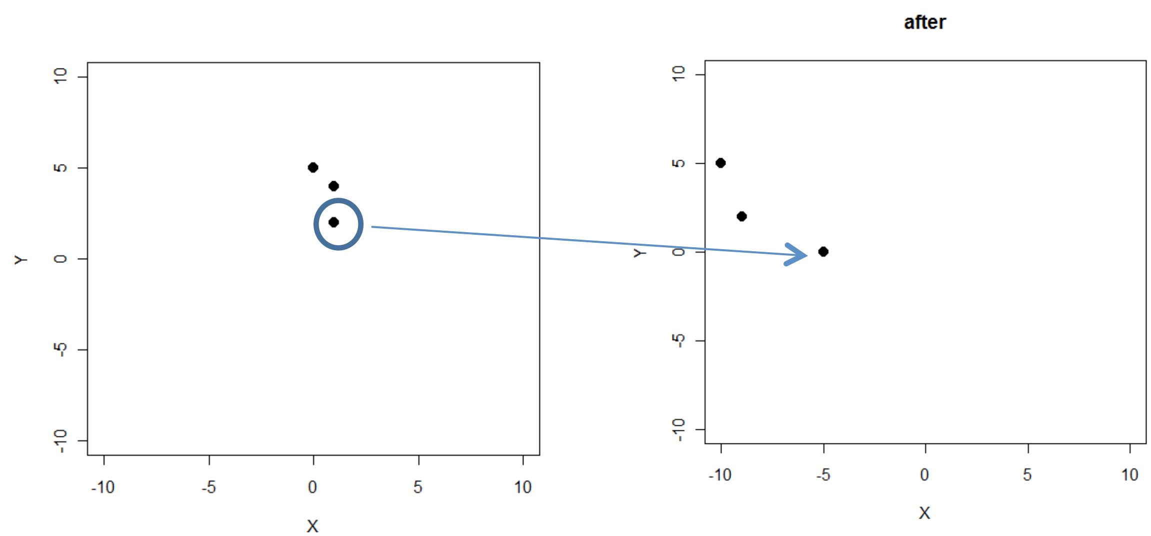
(3x**2**) (**2**x2) (3x2)

```
#Multiplication
A <- matrix(c(1, 2 , 1,4 , 0,5),3,2,byrow=T)
B <- matrix(c(-1, -2 , -2,1),2,2,byrow=T)

A%*%B
```

```
##       [,1] [,2]
## [1,]   -5    0
## [2,]   -9    2
## [3,]  -10    5
```

```
#Note: A%*%B is not the same as A*B
```





Matrix operations as transformations in space

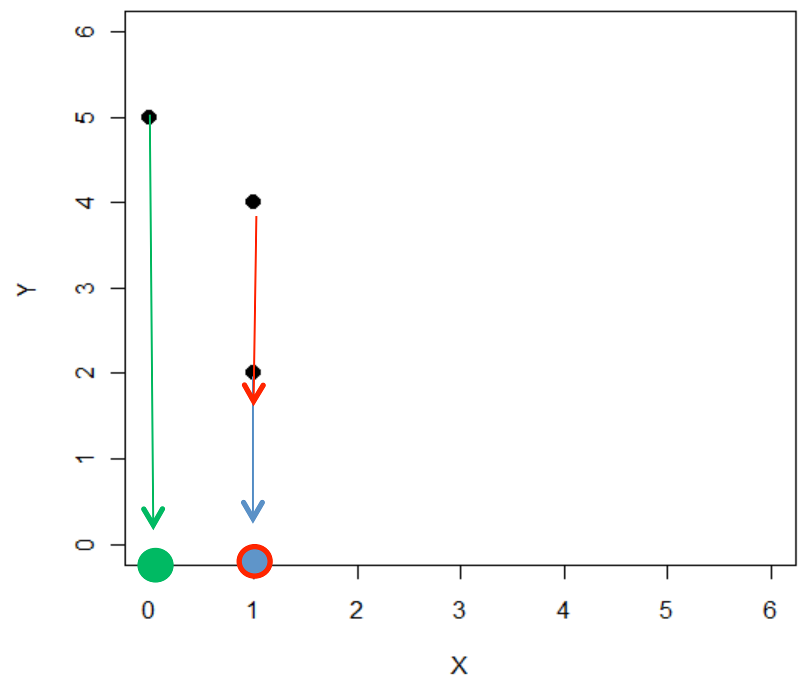
- Multiply by a scalar: moving the points further apart in space (or closer together)
- Multiply by another matrix: e.g., rotation, or projection
- Projection in particular is a fundamental operation: often want to project from original space to a reduced space that is “explanatory”

Examples: Projections

$$\begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 0 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

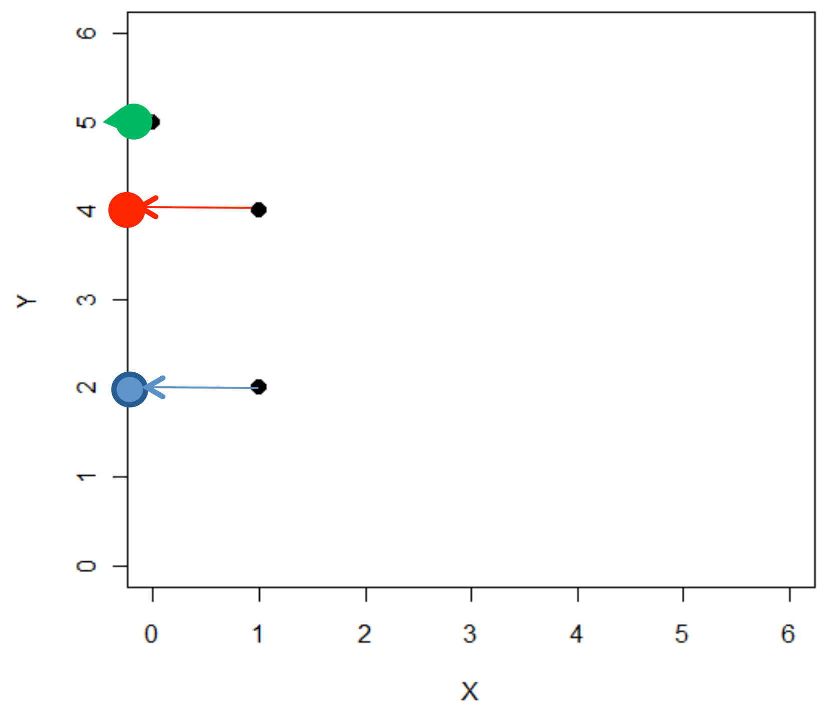
```r
#Projection
A <- matrix(c(1,2,1,4,0,5),3,2,byrow=T)
Px <- matrix(c(rep(0,3),1),2,2)
A%*%Px
```

```
##      [,1] [,2]
## [1,]    0    2
## [2,]    0    4
## [3,]    0    5
```



Examples: Projections

$$\begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 0 & 5 \end{bmatrix} \times \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 0 & 4 \\ 0 & 5 \end{bmatrix}$$




```
Py <- matrix(c(1,rep(0,3)),2,2)
A%*%Py
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    1    0
## [3,]    0    0
```

We can project onto any line we like

Inverse of a matrix

- An n x n square matrix $\mathbf{A}$ is invertible, if there exist an n x n square matrix $\mathbf{B}$ such that $\mathbf{AB=BA=I_n}$
- This matrix $\mathbf{B}$ is called the inverse of $\mathbf{A}$: $\mathbf{A^{-1}}$

$$\begin{bmatrix} 7 & 2 & 1 \\ 0 & 3 & -1 \\ -3 & 4 & -2 \end{bmatrix}^{-1} = \begin{bmatrix} -2 & 8 & -5 \\ 3 & -11 & 7 \\ 9 & -34 & 21 \end{bmatrix}$$

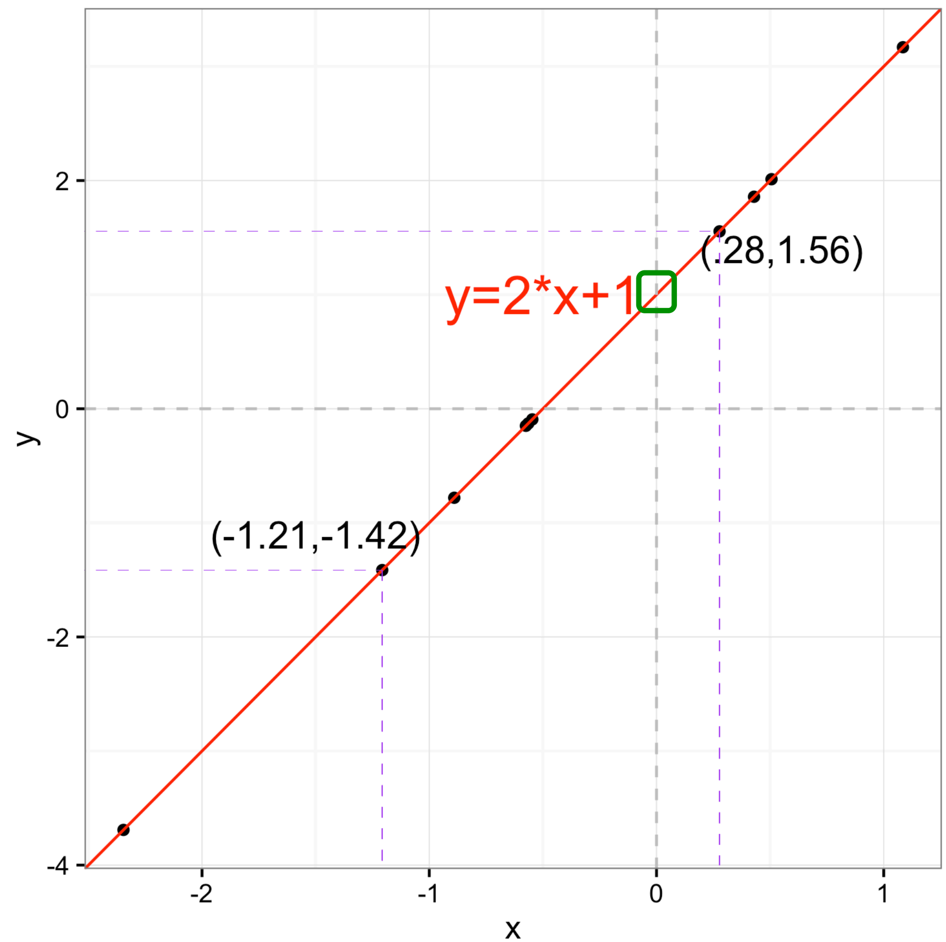
```
#Inverse
A <- matrix(c(7, 2 , 1,0 , 3 ,-1, -3, 4 , -2),3,3,byrow=T)
solve(A)
```

```
##      [,1] [,2] [,3]
## [1,]   -2    8   -5
## [2,]    3  -11    7
## [3,]    9  -34   21
```

A line

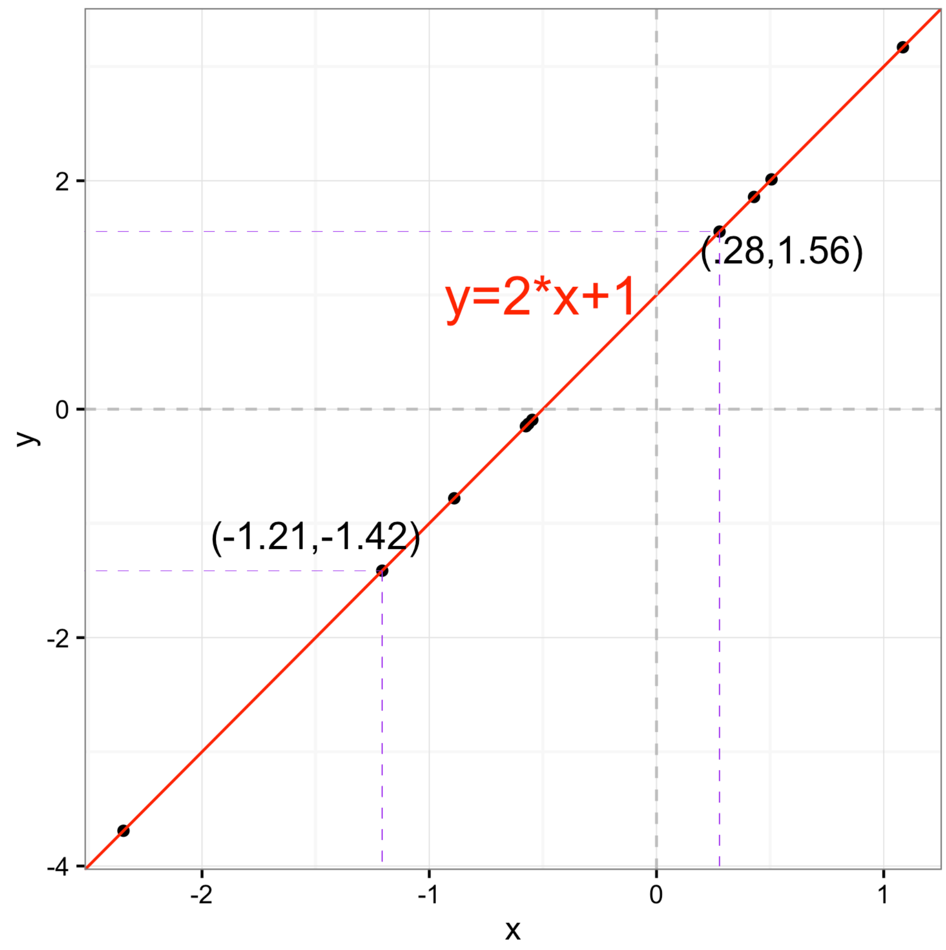
$$y = a + m * x$$

intercept



A line

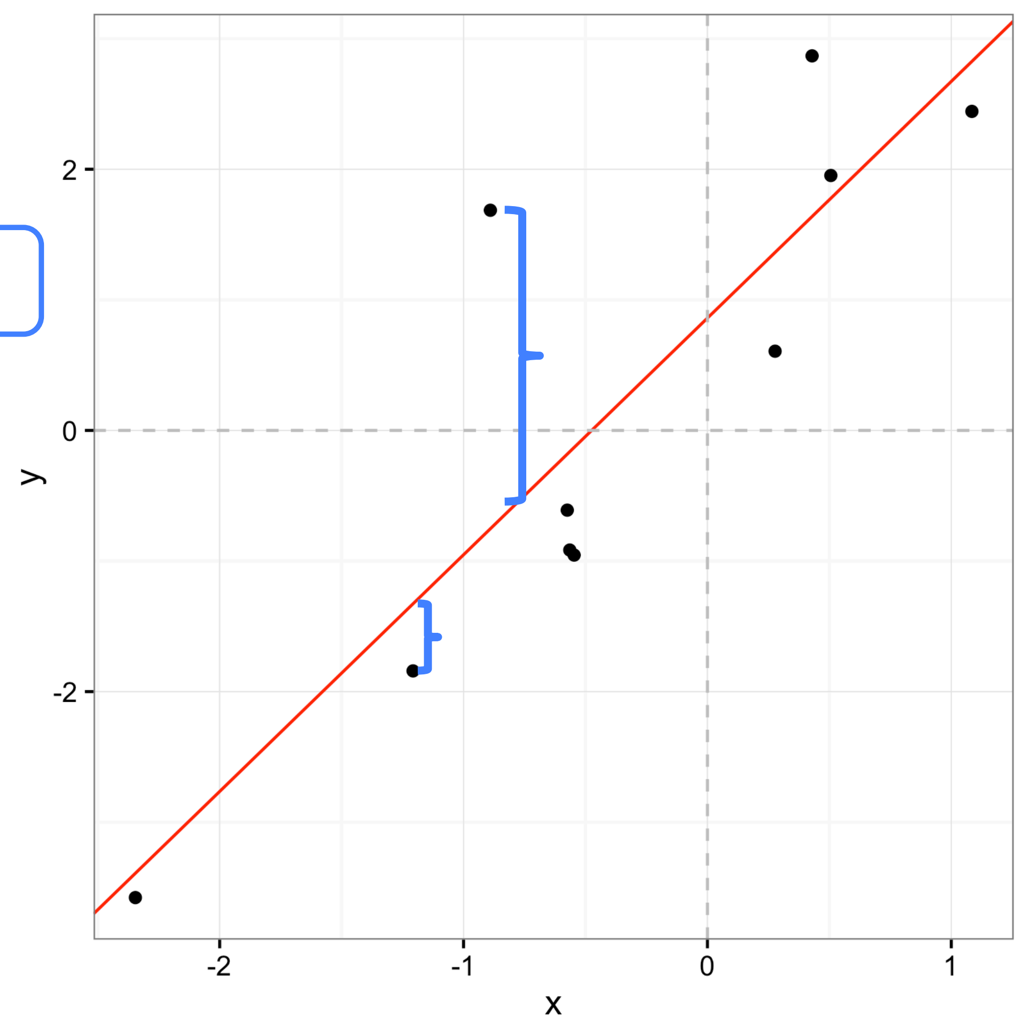
$$y = a + \underbrace{m}_{\text{slope}} * x$$

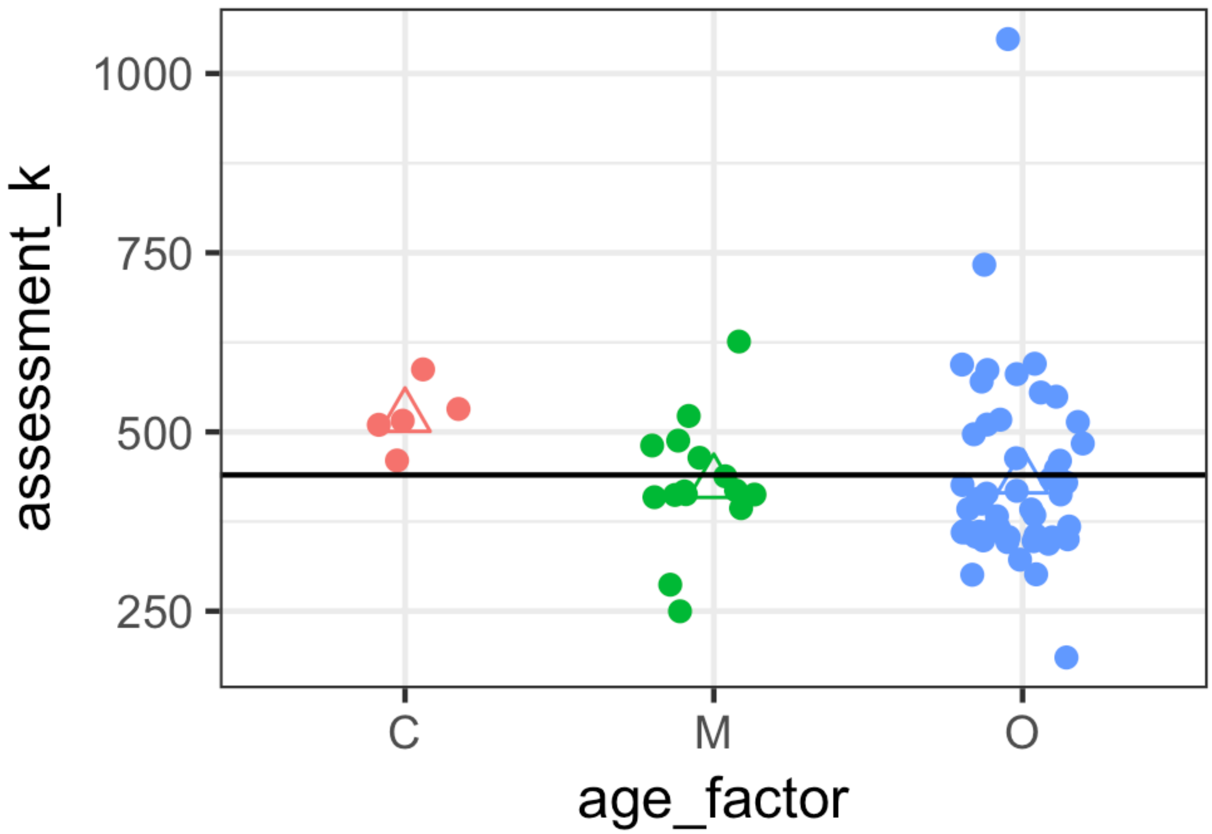


A regression line

$$y_i = a + m * x_i + \varepsilon_i$$

error





Where is the linear regression?

Glance at the data

YEAR_BUILT	age_factor	assessment_k
2013	C	510
2003	C	516
2002	C	460
2002	C	532
2005	C	587
1995	M	481
1989	M	409
1991	M	522
1998	M	413
1989	M	413
1988	M	394
1990	M	418
1998	M	464
1990	M	412
1989	M	488
1990	M	626
1984	M	417
1989	M	250
1997	M	438
1980	M	287

$$Y_C;\ Y_{C1}, \ldots, Y_{C5},\ n_C = 5$$

$$Y_M;\ Y_{M1}, \ldots, Y_{M15},\ n_M = 15$$

$$\boxed{H_0 : \mu_C = \mu_M}$$

```
## Call:
## lm(formula = assessment_k ~ age_factor, data = subset(dat.small,
##     age_factor %in% c("M", "C")))
```

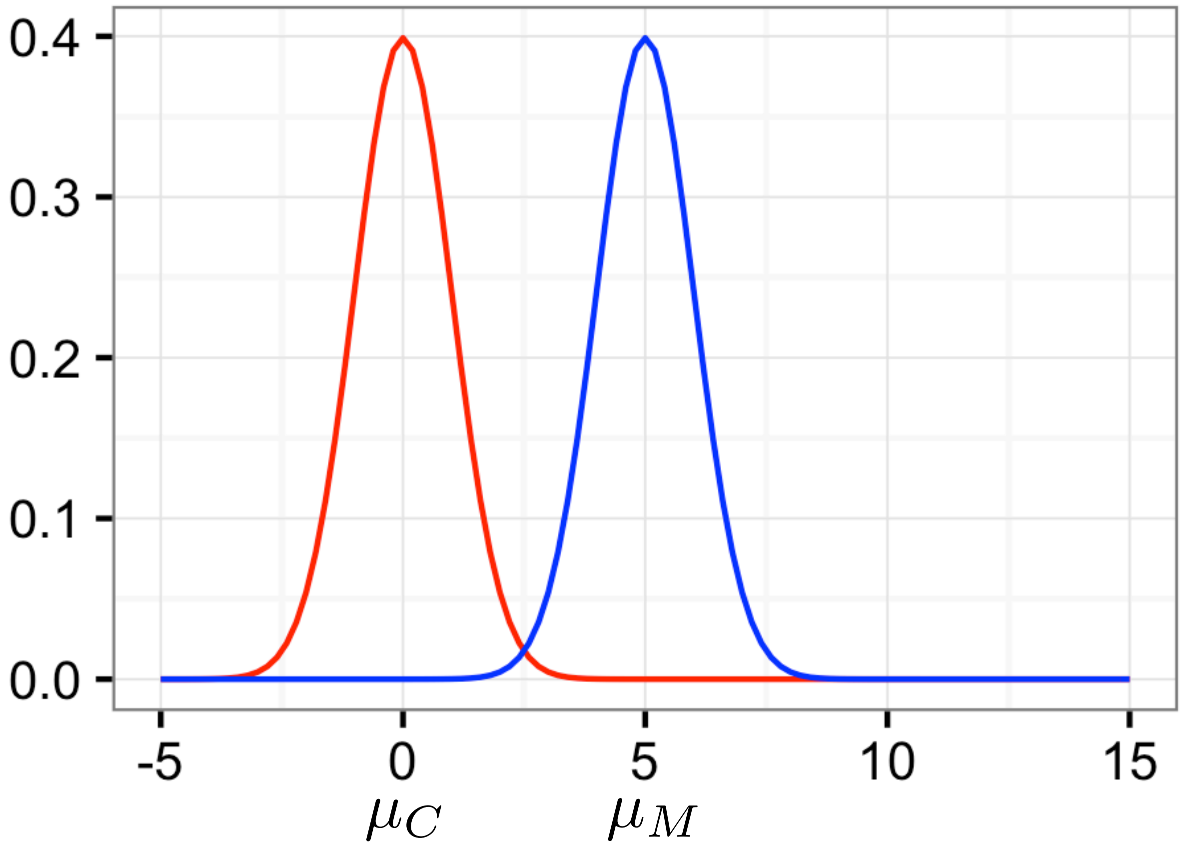
$$Y_{ij} = \mu_j + \varepsilon_{ij}, \text{ where } \varepsilon_{ij} \sim F_j, E(\varepsilon_{ij}) = 0$$

YEAR_BUILT	age_factor	assessment_k
2013	C	510
2003	C	516
2002	C	460
2002	C	532
2005	C	587
1995	M	481
1989	M	409
1991	M	522
1998	M	413
1989	M	413
1988	M	394
1990	M	418
1998	M	464
1990	M	412
1989	M	488
1990	M	626
1984	M	417
1989	M	250
1997	M	438
1980	M	287

$$\begin{bmatrix} \boxed{\begin{matrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \end{matrix}} \\ \boxed{\begin{matrix} Y_{12} \\ \vdots \\ Y_{n_2 2} \end{matrix}} \end{bmatrix} = \begin{bmatrix} \boxed{\begin{matrix} \mu_1 \\ \vdots \\ \mu_1 \end{matrix}} \\ \boxed{\begin{matrix} \mu_2 \\ \vdots \\ \mu_2 \end{matrix}} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2 2} \end{bmatrix}$$

$$\boxed{H_0 : \mu_C = \mu_M}$$

Change in notation…



$$Y_{ij} = \mu_j + \varepsilon_{ij},\ \varepsilon_{ij} \sim F_j,\ E[\varepsilon_{ij}] = 0$$
$$Y_M \sim F_M;\ E[Y_M] = \mu_M$$
$$Y_C \sim F_C;\ E[Y_B] = \mu_C$$

$$H_0 : \mu_C = \mu_M$$

We don't know or observe these curves and parameters

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \text{ where } \varepsilon_{ij} \sim F_j, E(\varepsilon_{ij}) = 0$$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ \vdots \\ Y_{n_2 2} \end{bmatrix} = \qquad\qquad = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2 2} \end{bmatrix}$$

$$Y_{ij} = \mu_j + \varepsilon_{ij},\ \text{where } \varepsilon_{ij} \sim F_j, E(\varepsilon_{ij}) = 0$$

$$\begin{bmatrix} \boxed{\begin{matrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \end{matrix}} \\ \boxed{\begin{matrix} Y_{12} \\ \vdots \\ Y_{n_2 2} \end{matrix}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \boxed{\mu_1} \\ \boxed{\mu_2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2 2} \end{bmatrix} = \begin{bmatrix} \boxed{\begin{matrix} \mu_1 \\ \vdots \\ \mu_1 \end{matrix}} \\ \boxed{\begin{matrix} \mu_2 \\ \vdots \\ \mu_2 \end{matrix}} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2 2} \end{bmatrix}$$

For example:

$$Y_{11} = 1 * \mu_1 + 0 * \mu_2 + \varepsilon_{11} = \mu_1 + \varepsilon_{11}$$

$$Y_{n_2 2} = 0 * \mu_1 + 1 * \mu_2 + \varepsilon_{n_2 2} = \mu_2 + \varepsilon_{n_2 2}$$

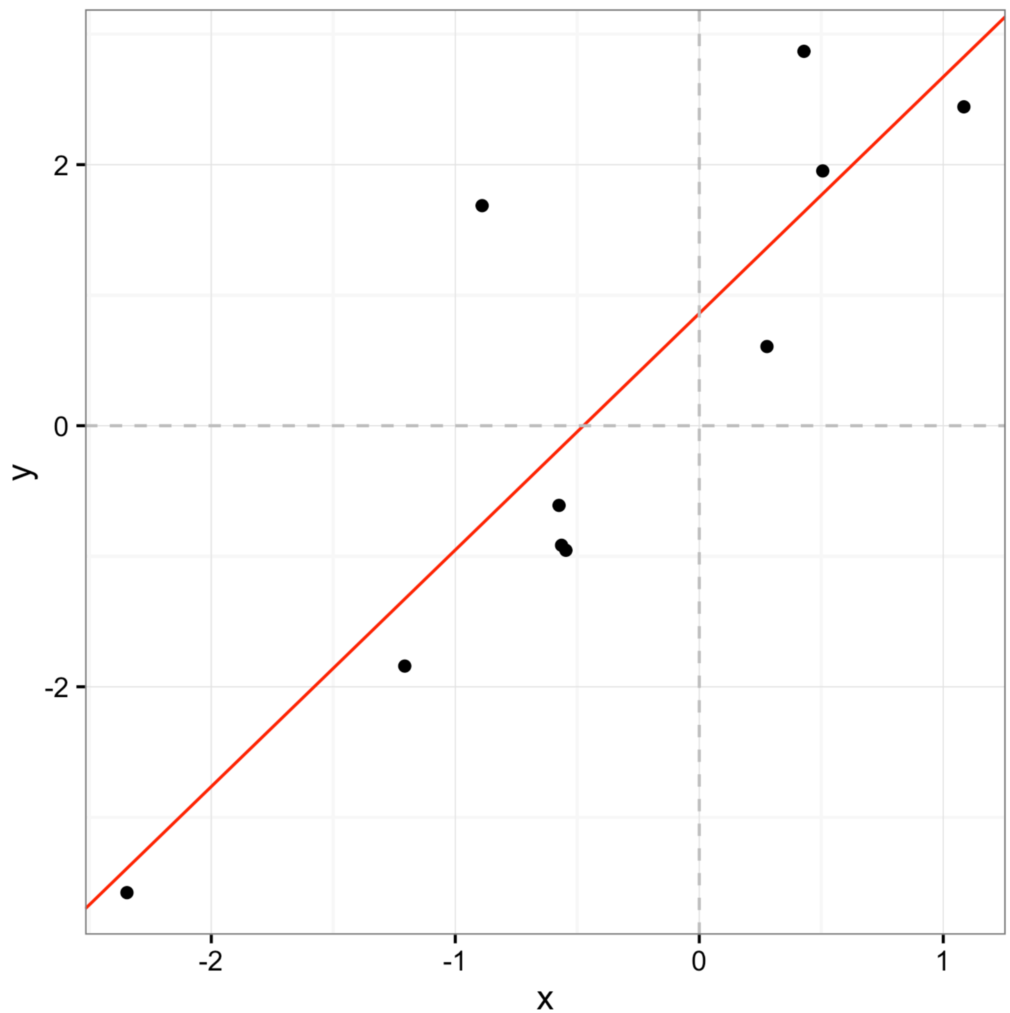
$$Y_{ij} = \mu_j + \varepsilon_{ij}, \text{ where } \varepsilon_{ij} \sim F_j, E(\varepsilon_{ij}) = 0$$

$$\underbrace{\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ \vdots \\ Y_{n_2 2} \end{bmatrix}}_{\text{response } \boldsymbol{Y}} = \underbrace{\begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}}_{\text{design matrix } \boldsymbol{X}} \underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}}_{\text{regression parameters}} + \underbrace{\begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2 2} \end{bmatrix}}_{\text{error term}}$$

$$Y = X\alpha + \varepsilon$$

A regression line

$$y_i = a + m * x_i + \varepsilon_i$$



the column vector of the responses
one element per experimental unit

a column vector
of the errors

$$Y = X\alpha + \varepsilon$$

a (design) matrix that represents covariate
info, one row per experimental unit

a column vector of the parameters in the
linear model

Generic linear model, using conventional matrix formulation

Slide by Prof Jenny Bryan

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ \vdots \\ Y_{n_2 2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2 2} \end{bmatrix}$$

Two groups

sample mean of C

```
summary(lm(assessment_k~age_factor,data=subset(dat.small,age_factor %in% c("M","C"))))
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    521.00      36.24  14.377 2.62e-11 ***
## age_factorM    -92.20      41.84  -2.203   0.0408 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.03 on 18 degrees of freedom
## Multiple R-squared:  0.2124, Adjusted R-squared:  0.1687
## F-statistic: 4.855 on 1 and 18 DF,  p-value: 0.04083
```

NOT the sample mean of M

$$
\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ Y_{12} \\ \vdots \\ Y_{n_2 2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2 2} \end{bmatrix}
$$

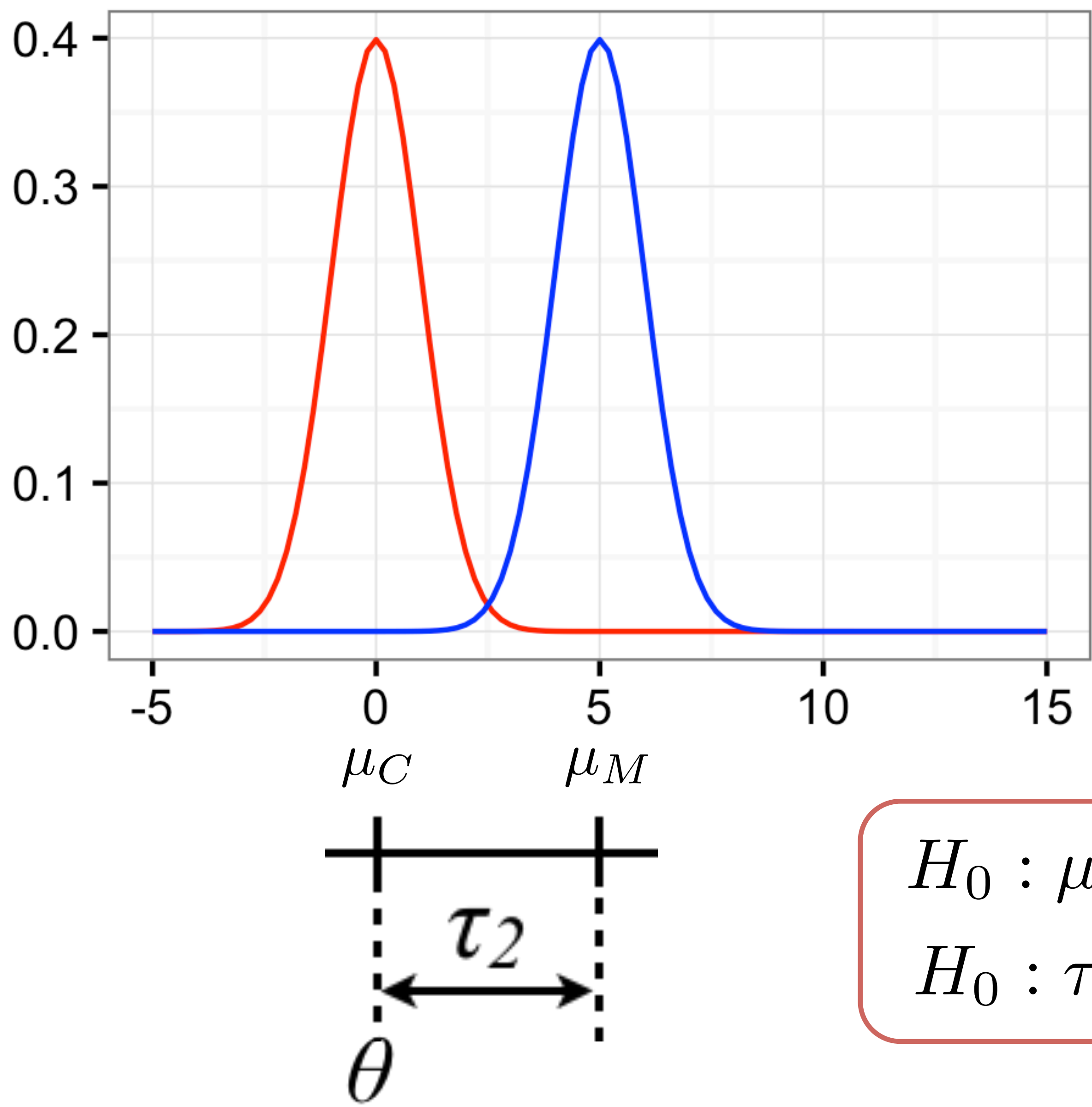
This is a one way of writing our problem as a linear regression

"cell means" parametrization

$$
Y = X\alpha + \varepsilon
$$

… but there are other ways!!

By default, R does not estimate these parameters



$$H_0 : \mu_M = \mu_C$$
$$H_0 : \tau_2 = 0$$

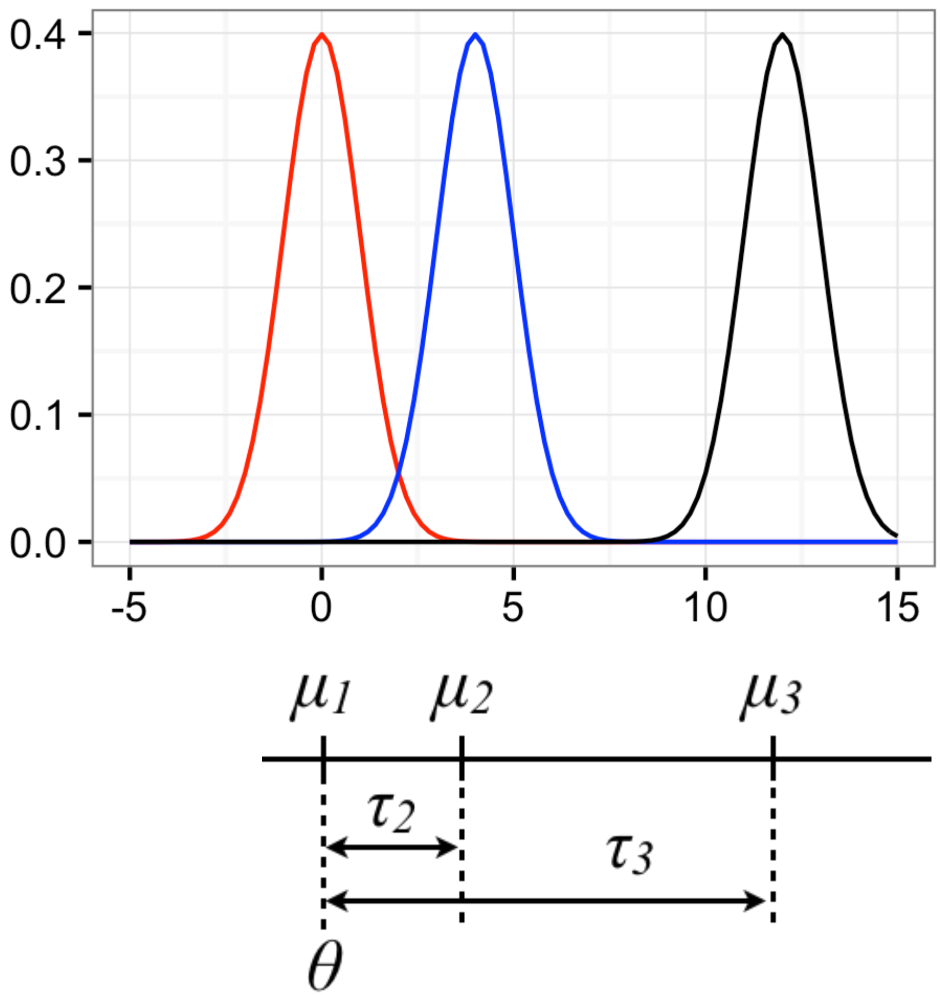
A different parametrization: "reference-treatment effect"

ANOVA-style, "cell means"

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

ANOVA-style, "ref + tx effects"

$$Y_{ij} = \theta + \tau_j + \varepsilon_{ij}, (\tau_1 = 0)$$



ANOVA-style, "cell means"

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

ANOVA-style, "ref + tx effects"

$$Y_{ij} = \theta + \tau_j + \varepsilon_{ij}, (\tau_1 = 0)$$

$$Y = X\alpha + \varepsilon$$

$$\begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ \\ y_{n_3 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \\ \varepsilon_{n_3 3} \end{bmatrix}$$

$$\begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ \\ y_{n_3 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \\ \varepsilon_{n_3 3} \end{bmatrix}$$

The design matrix specifies how the observed data relates to the regression parameters.

Slide by Prof Jenny Bryan

ANOVA-style, "cell means"

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

ANOVA-style, "ref + tx effects"

$$Y_{ij} = \theta + \tau_j + \varepsilon_{ij}, (\tau_1 = 0)$$

$$Y = X\alpha + \varepsilon$$

$$\begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ \\ y_{n_3 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \\ \varepsilon_{n_3 3} \end{bmatrix}$$

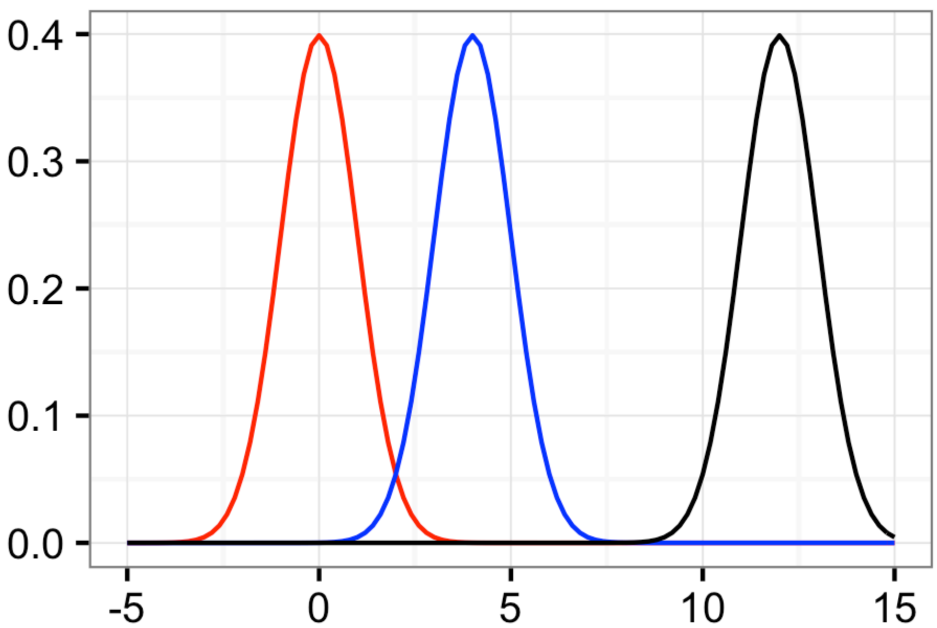
reference

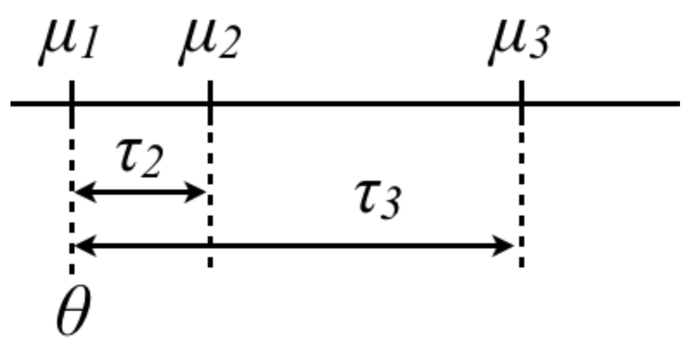
$$\begin{bmatrix} \boxed{y_{11}} \\ \boxed{y_{21}} \\ \vdots \\ \\ y_{n_3 3} \end{bmatrix} = \begin{bmatrix} \boxed{1} & 0 & 0 \\ \boxed{\vdots} & \vdots & \vdots \\ \boxed{1} & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \boxed{\theta} \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \\ \varepsilon_{n_3 3} \end{bmatrix}$$

For example:

$$Y_{11} = 1 * \theta + 0 * \tau_2 + 0 * \tau_3 + \varepsilon_{11} = \theta + \varepsilon_{11} \implies E[Y_{11}] = \theta$$

$$Y_{13} = 1 * \theta + 0 * \tau_2 + 1 * \tau_3 + \varepsilon_{13} = \theta + \tau_3 + \varepsilon_{13} \implies E[Y_{13}] = \theta + \tau_3$$





$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_0 : \tau_2 = \tau_3 = 0$$

treatment effect

$$Y_{i1} \sim F_1;\ E[Y_{i1}] = \mu_1 \qquad E[Y_{i1}] = \theta$$

reference

$$Y_{i2} \sim F_2;\ E[Y_{i2}] = \mu_2 \qquad E[Y_{i2}] = \theta + \tau_2 \implies E[Y_{i2}] - E[Y_{i1}] = \tau_2$$
$$Y_{i3} \sim F_3;\ E[Y_{i3}] = \mu_3 \qquad E[Y_{i3}] = \theta + \tau_3 \implies E[Y_{i3}] - E[Y_{i1}] = \tau_3$$

$$Y = X\alpha + \varepsilon$$

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{n_3 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n_3 3} \end{bmatrix}$$

Reference period: C

M vs C

O vs C

difference in ***population*** means