# Week 4

## Clustering Example

Single Value Decomposition

```
svd1 = svd(scale(samsungData
        [SamsungData$subject == 1,
        -c(562, 563)] )

par(mfrow = c(1,2))

plot(svd1$u[,1], col = numericalActivity,
        pch=19)

plot(svd1$u[,2], col = ...., pch=19)..
```

left singular vector
& explains the most variation

find maximum contributor

```
plot(svd1$v[,2], pch=19)
```

weight with which these variables contribute to variance

So we may ~~calculate~~ the variables with maximal weight

max Contrib = which.max(svd1$V[, 2])

distanceMatrix ← dist(samsungData
    [sd$subject == 1, c(10:12, maxcontrib)])

↓

then draw it

Goals of Statistical Modelling
- describe the distribution
- describe the relationship
-

Regression

Basic Least Squares

how to describe the distribution

lm → lm1 (galton$child ~ galton$parent)

lm1$fitted

lm1$residual ← left overs

linear model

$$C_i = b_0 + b_1 P_i + \underbrace{e_i}_{\text{error term}}$$

$$\sum_{i=1}^{n} \underbrace{(C_i - \{b_0 + b_1 P_i\})^2}_{\text{smallest}}$$

how to summarize a distribution?

if you ~~would~~ had to choose only one number to describe some data, what would you use?

$\Downarrow$

mean (for symmetric distribution)

center of distribution

also, it minimizes

$$\sum_i (c_i - \mu)^2$$

$c_i$ - observation

value that minimizes this formula is avg of c

'jitter' - add some noise to a variable

lm(x ~ y)

quantative var

variables to include in your model

equation for a line

$$\underset{\substack{\uparrow \\ \text{intercept term} \\ \text{(where the line crosses 0)}}}{C_i = b_0} + \underset{\substack{\uparrow \\ \text{slope}}}{b_1 P_i} + \underset{\substack{\uparrow \text{ (error term)} \\ \text{everything we didn't} \\ \text{measure}}}{e_i}$$

best line is:

$$\sum_i \left( C_i - \{ \underbrace{b_0 + b_1 P_i}_{\text{smallest}} \} \right)^2$$

lm1$fitted - line, found with least squares

(how many variation around this line?)

↓

lm1$residuals - left overs

## Inference Basics (linlog, linlegence)

```
library(UsingR); data(galton);

plot(galton$parent, galton$child, pch=19,
     col="blue")

lm1 = lm(galton$child ~ galton$parent)

lines(galton$parent, lm1$fitted, col="red",
      lwd=3)
```

```
lm1

Coefficients

(Intercept)      galton$parent
   23.942            0.646
```

$C_i = b_0 + b_1 P_i$

Let's generate a population of families

```
newGalton = data.frame(parent = rep(NA, 1e6),
                       child = rep(NA, 1e6))
```

- `newGalton$parent = rnorm(1e6, mean = mean(galton$parent), sd = sd(galton$parent))` — **parent**

  with the same properties as our sample

- `newGalton$child =` — **child**

  Intercept, slope

  `lm1$coeff[1]` ($b_0$) `+ lm$coeff[2]` ($b_1$) `· newGalton$parent +`

  `+ rnorm(1e6, sd=sd(lm1$residuals))`

  error term (noise)

smoothScatter(newGalton$parent,
newGalton$child)

abline(lm1, col="red", lwd=3)

regression line from Galton's Data
fits into newly generated data!

Let's take a sample

(from newly generated data)

50 el. from generated data

sampleGalton1 = newGalton[sample(1:1e6,
size=50, replace=F), ]

sampleLm1 = lm(sampleGalton$child,
sampleGalton$parent)

linear model from ~~ge~~ sample

and it's different from original lm1!

plot(sampleGalton1$parent,
sampleGalton1$child, pch=19, col="blue")

different!

lines(sampleGalton1$parent, sampleLm1$fitted,
lwd=3, lty=2) ← new

abline(lm1, col="red", lwd=3) ← old

if we take another sample, we'll see again
different results

So line you get from sample isn't the same line for the population

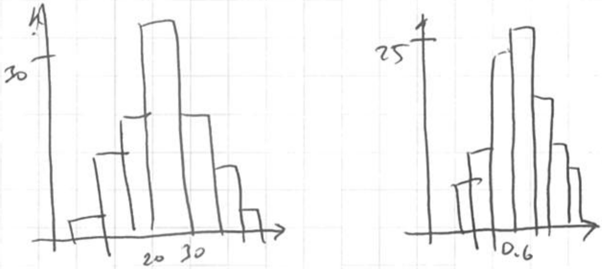
Histogram of estimates

```
par(mfrow = c(1, 2))
hist(interceptCoeffs, ...)
hist(slopeCoeffs, ...)
```

coefficients from lm run many times on different samples

slope coeffs

```
l = list of 100
for (i in 1:100) {
  sam = sample from population
  l[[i]] = lm(sam$ ~ sam$)
}
```




Intercept

Slope

How these coefficients are distributed?

From the central limit theorem

estimate of int → $\hat{b}_0 \sim N(b_0 \sim Var(\hat{b}_0))$

$\hat{b}_1 \sim N(b_1 \sim \underbrace{Var(\hat{b}_1)}))$ } follow normal distribution

estimate of slope

we don't know this var. exactly, but we can estimate them as well

$\hat{b}_0$ – $b_0$ estimated

estimation:

$$\hat{b}_0 \approx N(b_0, \hat{Var}(\hat{b}_0))$$
$$\hat{b}_1 \approx N(\hat{b}_1, \hat{Var}(\hat{b}_1))$$

$\sqrt{\hat{Var}(\hat{b}_0)}$ – "standart error" of the estimate $\hat{b}_0$

$(S.E.(\hat{b}_0))$

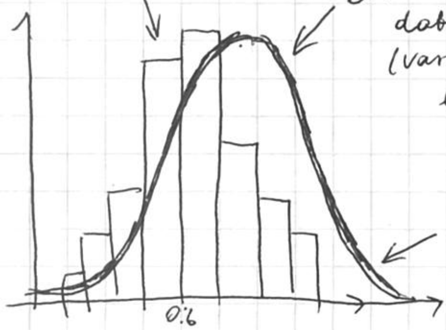
summary(sampleLm1)

⇓

gives you residuals
and estimates for the coefficent
as well as S.E.s for $b_0$ and $b_1$

distribution of
Slopes

estimated slope using one data sample
(variance is the same,
but little bit shifted)




```
lines(seq(0,5, length=100),
  dnorm(seq(0,5, length=100),
    mean=coef(sampleLm1)[2
    sd=summary(sLm1)$coeff[2,2
    lwd=3,
  col="red")
```

Standardized coefficients

$$\hat{b}_0 \approx N(b_0, \hat{Var}(\hat{b}_0))$$

dt in R

$\dfrac{\hat{b}_0 - b_0}{S.E.(\hat{b}_0)} \sim t_{n-2}$ ← center ← t-distribution (as n grows, it gets closer to $N(0,1)$) ← 2 degrees of freedom

↑ number of samples you had

degrees of freedom tell you how much variation you have left over after estimating your parameters

we loose 2 degrees of freedom when we calculate mean and the slope

$$\hat{b}_1 \approx N(b_1, \hat{Var}(\hat{b}_1))$$

$$\frac{\hat{b}_1 - b_1}{S.E.(\hat{b}_1)}$$

Confidence Intervals

We have an estimated $\hat{b}_1$ and want to know how good our estimate is

⇓

Create a "level $\alpha$ $\underbrace{\text{confidence interval}}_{\text{set of plausible values for } b_1}$"

So a confidence interval will include the real parameter $b_1$ in

$\alpha$% of the time in repeated studies (i.e. sampling)

(so if we sample 100 times, in $\alpha$ samples $b_1$ will be in the calculated interval)

- $\alpha$-confidence interval

$$\left(\hat{b}_1 - T_{\alpha/2} \cdot SE(\hat{b}_1) \; , \; \hat{b}_1 + T_{\alpha/2} \cdot SE(\hat{b}_1)\right)$$

can be calculated using lm command

summary(sampleLm1)$coeff

particular quantile of t-distribution

- confint
~~confint~~ (sampleLm1, level = 0.95)

will calculate confidence interval

# How to report the inference

A one inch increase
in parental height
is associated with
a 0.77 inch increase
in child's height
(95%, CI: 0.42 - 1.12 inches)

confint(sampleLm,
level = 0.95)

| | | |
|---|---|---|
| (Intercept) | -7.8 | 39.5 |
| parent | 0.42 | 1.13 |

sampleLm$coeff

| (Intercept) | parent |
|---|---|
| 15.86 | 0.7698 |

## P-values

most commonly reported measure of
„statistical significance"

Idea: suppose we see nothing

how likely / unlikely that there is
a relationship between variables?

Approach

1. Define the hypothetical distribution of a
data summary (statistic)

(null hypothesis)

"when nothing is going on"

2. Calculate the summary / test statistics with the data we have

3. Compare what we calculated to our hypothetical distribution and see if the value is "extreme" (p-value)

null hypothesis

$$\frac{\hat{b}_1 - b_1}{S.E.(\hat{b}_1)} \sim t_{n-2}$$

↓

$H_0$ : there is no relationship between parent and child height

(i.e. $b_1 = 0$)

Under the null hypothesis the distribution is:

$$\frac{\hat{b}_1}{S.E.(\hat{b}_1)} \sim t_{n-2}$$

↑

now we compare it with $t_{n-2}$ and we if there are any surprizes x = seq(-20, 20, length = 100)

plot(x, dt(x, df = (928 - 2)), col = "blue",
lwd = 3, type = "l")

t-distribution

degrees of freedom

null-distribution

summary(~~sample.lm1~~ lm1)

...

Coefficients

| | Estimate | S.E. | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 23.94 | 2.81 | 8.52 | <2e-16 |
| galton$parent | 0.64 | 0.04 | 15.71 | <2e-16 |

estimate divided by S.E.

Some typical values

used to see if there are any relationships between 2 variables

$p < 0.05$ Statistically significant ← usually used

$p < 0.01$ strongly significant

$p < 0.001$ very significant

Usually in report both confidence interval and p-value are mentioned

How to interpret?

summary(lm(g$child ~ g$parent))$coeff

| | Estimate | S.E. | t | Pr(>\|t\|) |
|---|---|---|---|---|
| (intercept) | 23.94 | 2.81 | 8.5 | 6.53e-17 |
| g$parent | 0.64 | 0.04 | 15.7 | 1.73e-49 |

A one inch increase in parental height is associated with a 0.77 inch increase in child's height

(95% CI: 0.42–1.12 inches)

This difference was statistically significant ($p < 0.001$)

# Regression with factor variables

rotten tomatoes score vs rating

```
plot(movies$score ~ jitter(as.numeric(
        movies$rating)), col="blue",
        xaxt="n", pch=19)
```

factor

```
meanRatings = tapply(movies$score,
        movies$rating, mean)
```

```
points(1:4, meanRatings, col="red",
        pch="-", cex=5)
```

Another way to write it down:

$$\mathbb{1}(Ra_i = \text{"PG"}) = \begin{cases} 1 & \text{if } Ra_i == \text{"PG"} \\ 0 & \text{if } Ra_i \neq \text{"PG"} \end{cases}$$

regression line

$$S_i = b_0 + b_1 \mathbb{1}(Ra_i = \text{"PG"}) + b_2 \mathbb{1}(Ra_i = \text{"PG-13"}) + b_3 \mathbb{1}(Ra_i = \text{"R"}) + e_i$$

$b_0$ - avg of G

$b_0 + b_1$ = avg of PG  $b_0 + b_2$ = avg of PG-13

$b_0 + b_3$ = avg of R

```
lm1 = lm(movies$score ~ as.factor(movies$rating))
summary(lm1)
```

$\Downarrow$

and we get estimate for each movie type

```
lm1 = ..
anova(lm1)
```
← analysis of variance table

(check what anova is)

# Multiple Regressions

- regression with multiple covariates
- still using least square / central limit theorem

lmBoth = lm ( hunger$Numeric ~
hunger$Year + $\underbrace{\text{hunger\$Sex}}$ )

categorize by sex

same slope

lmBoth = lm ( hunger$Numeric ~
hunger$Year + hunger$Sex +
+ hunger$Sex * hunger$Year )

2 slopes

# Regression in the Real World

ideal data: galton data set

(~~dot~~ cloud-shape)

- confounders

  a variable that is correlated with both outcome and the covariates

  - confounders can change the regression line
  - change the sign of the line!
  - can be detected by careful exploration

    ~~data~~ visualization

- right-skew

  (income) - log it

  better for regression

hist

- outliers

outliers – data points that do not appear to follow the pattern of the other data points

can have drammatic impact on how you visualize data

and on your analysis

what can you do?

- you have to know if it's real or not

  If not – remove it

- or

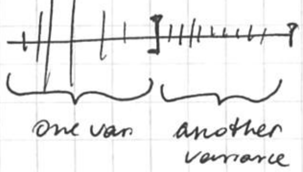
  - logarithm it
  - sensitivity analysis (leave out for awhile)
  - robust methods –

- A line isn't always the best summary

(wiki – Linear_Regression)

- changing variance:

when there are 2 variances




- Box-Cox transform
- Variance stabilizing transform
- Weighted least squares
- etc

- Units

Absolute vs Relative

- standardize, but keep track

but it affects

- model fits
- interpretation
- inference

- overloading regression

too many params in regression

- correlation and causation

be critical / consider alternative

Was X indeed the cause for Y?
(chocolate consumption vs nobel prize winners)