Analysis of Variance (ANOVA tests)

16 March, 2015
Analysing Differences

Correlation

such as: linear regression, Pearson’ coefficient
Measure

- Strength of association
- Look for relationships between variables
- World made of relationships
  - What is the relationship between gender and height in humans?

Difference statistic

such as: t.test, anova, correlation test
Measure

- Differences in central tendency
- Look for difference in variables
- World made of differences
  - What is the difference between heights of human males and females?
Definition:

- Generalization of the *t-test*

- Use for analyzing differences between **more than 2 groups**

- Analyses and interprets observations of several groups, treatments, conditions, etc.

- Decomposes the total variance present in the data into contributions of the single sources of variation:
  - systematic contributions = differences of means
  - random contribution = variability around group mean
Analysis of Variance

ANOVA

Definition:

- Generalization of the *t-test*
- Use for analyzing differences between **more than 2 groups**
- Analyses and interprets observations of several groups, treatments, conditions, etc.
- Decomposes the total variance present in the data into contributions of the single sources of variation:
  - systematic contributions = differences of means
  - random contribution = variability around group mean

Idea:

- Explore the effect of different groups on an response variable
Research Questions

**Medicine**  *Arness et al 1978* 22 bypass-patients are randomly divided into 3 treatments groups (different respiration). Differ the values of folic acid in red blood cells after 24h?

**Economy** Check whether the power consumption is the same over all supermakets of the same brand.

**Sociology** Does satisfaction significantly differ by location of institution (rural, urban, suburban?)

**More examples I**  Is there any significant difference in weight loss when dieting, exercising and dieting and exercising?

**More Examples II**  Is there any significant difference in children self-esteem by parenting style (authoritative, permissive, authoritarian)?
Example

Let’s analyze the Cushings dataset from the package MASS.

**Info**  Cushing’s syndrome is a hormone disorder associated with high level of cortisol secreted by the adrenal gland.

**Variables**  Type: Categorical, define the underlying type of syndrome
- Tetrahydrocortisone: Numerical, urinary excretion rate (mg/24h)
- Pregnanetriol: Numerical, urinary excretion rate (mg/24h)

**Aim**  Find whether the 4 groups are different with respect to the urinary excretion rate of Tetrahydrocortisone
Data Exploration I

Let’s look at the data:

- Look at the distribution of Tetrahydrocortisone according to disease type
- Use boxplots and dotplots

```r
data(Cushings, package="MASS")
plot(Tetrahydrocortisone~Type, data=Cushings)
```
Data Exploration II

points(3, Cushings[c4, "Tetrahydrocortisone"], col="red", pch=19, cex=2)
ANalyses Of VAriances (ANOVA)

Given $N$ groups define:

**Response variable** $y_{ij}$ (the urinary excretion for each individual/group)

**Explanatory variable** $x_{ij}$ (the type of disease)

**Group description** Mean and variance for each group.

NB These are extraction from the whole population

- $G_1 (m_1, s_1) \neq (\mu_1, \sigma_1)$
- $G_2 (m_2, s_2) \neq (\mu_2, \sigma_2)$
- \ldots
- $G_N (m_N, s_N) \neq (\mu_N, \sigma_N)$

**Global mean** The overall mean of the response variable. Do not take group categories into account

$$\bar{y}$$

**Group Mean** Response variable mean for each group

$$(\bar{y}_1, \ldots, \bar{y}_N)$$
Look at the distribution of the means

```r
dotplot(Tetrahydrocortisone Type, data=Cushings, panel=function(x,y,..., m=mean(Cushings[,"Tetrahydrocortisone"]){ panel.abline(h=m, lty=2) panel.dotplot(x,y,..., cex=1.8) for (i in levels(x)){ panel.dotplot(unique(x[x==i]), mean(y[x==i]),col="red", pch=17, cex=2) } })
```

**NB:** Look at the variation
- from the mean between groups
- from the mean within groups

Both the variations contribute to the total variation of the observations
Anova

Take into account variances

Anova partition the sums of squares (variance from the mean) into:

- **Explained variance** (*between groups* \( SS_b \))
  - Systematic variance, it reflects differences among groups due to the experimental treatment of characteristic for group membership
  - This is the sum of square between groups, hence the squared differences between each individual group mean from the grand mean

- **Unexplained variance** (*within groups* \( SS_w \))
  - Random variance, or noise. It reflects the variance from the mean between individuals in the same group.
  - Sum of squares of the differences of each individual from the individual's group mean

- The **linear model** conceptually is:

\[
SS_t = SS_b + SS_w
\]
Recall the formula of variance:

\[
\sigma^2 = \frac{\sum (x - \bar{x})^2}{N - 1}
\]

We can rewrite it using the Sum of Squares and the degree of freedom as:

\[
\sigma^2 = \frac{SS}{df}
\]

Hence:

**Between** \( SS_b = \sum_{i=1}^{N} n_i (\bar{y}_i - \bar{y})^2 \)

In order to take into account the different sample size in each group, note the weight \( n_i \). Groups with more observation are weighted more heavily.

**Within** \( SS_w = \sum_{i=1}^{N} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \)

**Total** \( SS = \sum_{i=1}^{N} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \) Deviation of each individual from the overall average
ANOVA Tests

Assumptions:
- Observation over groups are independent
- Observations for each group come from a gaussian distribution
- Variances should be homogeneous

H0
- $\mu_1 = \mu_2 = \ldots = \mu_N$

H1 = not(H0)
- At least 2 means over all groups should be different

Compute the F-statistic for N groups and n observation:

$$F = \frac{SS_b/(N - 1)}{SS_w/n - N}$$

Comparing variation between groups and variation within groups. If the null hypothesis is true, then the test statistic F has an F-distribution $F(df_b, df_w)$

Degree of freedom:
- Between $df_b = N - 1$
- Within $df_w = n - N$

There is any significant difference between at least two of the groups in variable x
F-distribution

```r
xyplot(Y ~ X, groups = df1 + df2, data = df, type = "l", lwd = 2,
   auto.key = list(TRUE, columns = 3,
   text = paste("df1,df2", unique(paste(df$df1, df$df2, sep = ",")), sep = " = "), lines = TRUE, points = FALSE,
   ylab = "Density",
   xlab = ",",
   main = "F-Distribution",
)
```

![F-Distribution plots](image)

**Analysis of Variance (ANOVA tests)**
Test variable $F$:

$$F_{obs} = \frac{SS_{between}/(N-1)}{SS_{within}/(n-N)}$$

If $F_{obs} > \theta$ or $p_{obs} < \alpha$ reject H0
If $F_{obs} < \theta$ or $p_{obs} > \alpha$ accept H0

Accept H0 $\Rightarrow \mu_1 = \mu_2 = \ldots = \mu_N$
Reject H0 $\Rightarrow \mu_i \neq \mu_j$

**NB:** $i$ and $j$ could be any couple of groups!
A working example

Example

Let's analyze the Cushings dataset from the package MASS.

Variables

* Type: Categorical, define the underlying type of syndrome
  This is the grouping variable $N = 4$
  Tetrahydrocortisone: Numerical, urinary excretion rate (mg/24h); the response variable $y$

Aim

Find whether the 4 groups are different with respect to the urinary excretion rate of Tetrahydrocortisone

Compute the overall mean $\bar{y}$:

```r
mean(Cushings$Tetrahydrocortisone)
```

## [1] 10.456

Compute the mean for each group $\bar{y}_i$

```r
tapply(Cushings$Tetrahydrocortisone, Cushings$Type, mean)
```

## a  b  c  u
## 2.9667 8.1800 19.7200 14.0167

Compute Degree of freedom $df_b$, $df_w$:

```r
length(levels(Cushings$Type)) - 1  ## df_1
nrow(Cushings) - length(levels(Cushings$Type))  ## df_2
```

## [1] 3
## [1] 23
Manual Computation

Compute $SS_b$ and $SS_w$ and the $f$ statistic:

\[
SS_b <- \text{sum}\left(\text{table}(\text{Cushings$Type}) \times \text{tapply}(\text{Cushings$Tetrahydrocortisone, Cushings$Type, mean}) - \text{mean}(\text{Cushings$Tetrahydrocortisone})^2\right)
\]

\[
SS_w <- \text{sum}\left(\text{unlist}\left(\text{by}(\text{Cushings, Cushings$Type, function(x){sum}}((x$Tetrahydrocortisone - \text{mean}(x$Tetrahydrocortisone)^2))\right)\right)
\]

\[
f <- (SS_b / 3) / (SS_w / 23)
\]

Compute the p-value = $P(F \geq 3.2)$

\[
pvalue <- \text{df}(3.2, df1=3, df2=23)
\]

- Define a threshold $\alpha$ that if $p_{obs} \geq \alpha$ we reject $H_0$. Normally $\alpha = 0.1, 0.05, 0.01$

```
pvalue
## [1] 0.041074
```

(UNITN-FBK)
Visualizeing the testing variable

```r
plot(df$X, df$Y, yla="Density", xlab="", type="l", main="F(3,23), f=3.2, p=0.041")
polygon(x=c(x, rev(x)), y=c(rep(0, length(y)), rev(y)), col="grey80")
arrows(x0=2, x1=3.2, y0=0.05, y1=0, lwd=1.5)
arrows(x0=3.5, x1=3.8, y0=0.2, y1=0.01, lwd=1.5)
text(x=2, y=0.05, labels="f", pos=3, cex=1.8)
```

```
F(3,23), f=3.2, p=0.041
```

![Graph showing the visualization of the testing variable with a density plot, a shaded area indicating observed value (Pobs), and annotations for f and Pobs.](image)
Let R work for us

How to compute everything in R?

```r
myanova <- aov(Tetrahydrocortisone~Type, data=Cushings)
summary(myanova)
```

```
## Df Sum Sq Mean Sq F value Pr(>F)
## Type    3 894   297.8 3.23  0.041 *
## Residuals 23 2124   92.3
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
Until now we have seen an explanatory variable (categorical) vs a response variable (numerical). This is called **One-way Anova**.

What if we have **more than one** categorical variable?

**Two-way Anova:**

- This is an anova with 2 factors
- Investigate the effect of two (or more) combining factors
- Normally one of them is the main factor. The second one is believed to be important in the study of the relationship between response variable and explanatory factor.
Example

Let's work on the rat genotype dataset from package MASS. Rat litters were separated by their natural mothers and were nurtured by foster mothers.

**Info** Investigate the inheritance of maternal influences on the growth of the rat

**Variables**
- Wt: Weight of the litter
- Litter: Litter identifier
- Mother: Mother genotype

**Aim** In this example, we are mainly interested in the variation of weight gain across different genotypes of mothers. However, we need to account for possible weight gain variation due to the genotype of the litters.
Effect of interaction:

Given the previous example:

- Look at the model with group separated: **no interaction** between factors.
  - The linear model for the ANOVA will be:
    \[ SS = SS_M + SS_L + SS_E \]
    where \( SS_m \) is the variation according to mother genotype, \( SS_L \) the variation according to litter group and \( SS_E \) is the random variation within the group.

- If we believe the interaction between factors will be important than:
  - The linear model for the ANOVA will be:
    \[ SS = SS_M + SS_L + SS_M \times L + SS_E \]

Variation the response variable due to specific combinations of the two factors is usually referred to as the **interaction effect**
In contrast, the variation in the response variable due to one of the factors alone is called the **main effect**.
Two-way anova in R

Get the data from package MASS and look at the dataset

```r
data(genotype, package="MASS")
head(genotype)
##    Litter Mother  Wt
## 1: A     A   61.5
## 2: A     A   68.2
## 3: A     A   64.0
## 4: A     A   65.0
## 5: A     A   59.7
## 6: A     B   55.0

str(genotype)
## 'data.frame': 61 obs. of 3 variables:
##  $ Litter: Factor w/ 4 levels "A","B","I","J": 1 1 1 1 1 1 1 1 1 1 ...  
##  $ Mother: Factor w/ 4 levels "A","B","I","J": 1 1 1 1 1 2 2 2 3 3 ...  
##  $ Wt   : num 61.5 68.2 64 65 59.7 55 42 60.2 52.5 61.8 ...  

summary(genotype)
##    Litter  Mother  Wt
##  A:17    A:16  Min. :36.3
##  B:15    B:14  1st Qu.:48.2
##  I:14    I:16  Median :54.0
##  J:15    J:15  Mean  :54.0
##        3rd Qu.:60.3
##        Max.  :69.8
```

(UNITN-FBK)
Compute the ANOVA model **without interaction**, use the sign `+` in the formula to add factors

```r
myanova.noint <- aov(Wt~Mother + Litter, data=genotype)
summary(myanova.noint)
```

```
## Df Sum Sq Mean Sq F value Pr(>F)
## Mother 3 772 257.2  4.25 0.0091 **
## Litter 3  64  21.2  0.35 0.7887  
## Residuals 54 3265  60.5
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
Compute the ANOVA model with interaction, use the sign * to add factors and include the interaction between them

```r
myanova.int <- aov(Wt~Mother * Litter, data=genotype)
summary(myanova.int)
```

```
## Df Sum Sq Mean Sq F value Pr(>F)
## Mother 3 772 257.2 4.74 0.0059 **
## Litter 3 64 21.2 0.39 0.7600
## Mother:Litter 9 824 91.6 1.69 0.1201
## Residuals 45 2441 54.2
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
Check the assumptions

Recall:

Assumptions:
- Observation over groups are independent
- Observations for each groups come from a gaussian distribution
- Variances should be **homogeneous**
Let's use the *poisons* dataset from package MASS.

**Info** This is a study on survival time of rats after poison which undergone different treatments. A total amount of 48 rats, 3 poisons and 4 treatments. Each combination poison-treatment was used for 4 animals.

**Variables**
- **time**: survival time of the rat
- **poison**: type of poison (1,2,3)
- **treat**: type of treatment (A,B,C,D)

**Aim** Explore the effect on survival of poison, treatment and the combined effect of these two groups.
Explore the data first

Look at the boxplot of survival time at all combinations of poison and treatment

```
boxplot(time ~ treat + poison, data=rats, ylab="survival time", cex.lab=1.5)
```

I is quite apparent that the assumption of equal variances is not true...
we go ahead any way
Explore the data first

One may assume additive effects of the two factors, or an interaction among the two.

Graphical analysis of interactions (if no interactions, lines should be parallel)

```r
interaction.plot(rats$treat, rats$poison, rats$time)  # graphical analysis of interaction (if no interaction, lines should be parallel)
```

![Graph showing interaction plot](attachment:graph.png)
The lines are mean survival at different treatments for the poisons. Actually, they are not so far from parallel.
Build the model with interactions I

Model with interaction terms
Note the use of * for the interaction of the factors

g <- lm(time ~ poison * treat, data = rats)
anova(g)

## Analysis of Variance Table
## Response: time
## Df Sum Sq Mean Sq F value Pr(>F)
## poison 2 1.033 0.517 23.22 3.3e-07 ***
## treat 3 0.921 0.307 13.81 3.8e-06 ***
## poison:treat 6 0.250 0.042 1.87 0.11
## Residuals 36 0.801 0.022
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Build the model with interactions II

summary(g)

## Call:
## lm(formula = time ~ poison * treat, data = rats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3250 -0.0488  0.0050  0.0431  0.4250
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.4125     0.0746   5.53 2.9e-06  ***
## poison2                   -0.0925     0.1055  -0.88  0.386     .
## poison3                   -0.2025     0.1055  -1.92  0.063     .
## treatB                    0.4675     0.1055   4.43 8.4e-05  ***
## treatC                    0.1550     0.1055   1.47  0.150     .
## treatD                    0.1975     0.1055   1.87  0.069     .
## poison2:treatB            0.0275     0.1491   0.18  0.855     .
## poison3:treatB            -0.3425     0.1491  -2.30  0.028     *
## poison2:treatC            -0.1000     0.1491  -0.67  0.507     .
## poison3:treatC            -0.1300     0.1491  -0.87  0.389     .
## poison2:treatD            0.1500     0.1491   1.01  0.321     .
## poison3:treatD            -0.0825     0.1491  -0.55  0.584     .

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.149 on 36 degrees of freedom
## Multiple R-squared:  0.734, Adjusted R-squared:  0.652
## F-statistic: 9.01 on 11 and 36 DF,  p-value: 1.99e-07
Build the additive model I

Model without interaction term. Only additive effect.

\[
g_{\text{add}} \leftarrow \text{lm}(\text{time} \sim \text{poison} + \text{treat}, \text{data} = \text{rats})
\]

\[
\text{anova}(g_{\text{add})
\]

```r
## Analysis of Variance Table
##
## Response: time
## Df Sum Sq Mean Sq F value Pr(>F)
## poison 2 1.033  0.517  20.6 5.7e-07 ***
## treat  3 0.921  0.307  12.3 6.7e-06 ***
## Residuals 42 1.051  0.025
##---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
Build the additive model II

```r
summary(g_add)
```

```r
##
## Call:
## lm(formula = time ~ poison + treat, data = rats)
##
## Residuals:
## Min 1Q Median 3Q Max
## -0.2517 -0.0962 -0.0149 0.0618 0.4983
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.4523     0.0559  8.09  4.2e-10 ***
## poison2              -0.0731     0.0559  -1.31   0.1981
## poison3              -0.3412     0.0559  -6.10  2.8e-07 ***
## treatB               0.3625     0.0646   5.61  1.4e-06 ***
## treatC               0.0783     0.0646   1.21   0.2319
## treatD               0.2200     0.0646   3.41   0.0015 **
##
## ---
## Signif. codes:  < 0.001 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.158 on 42 degrees of freedom
## Multiple R-squared:  0.65, Adjusted R-squared:  0.609
## F-statistic: 15.6 on 5 and 42 DF,  p-value: 1.12e-08
```
Tests one model against the other
The result is exactly what we had seen in `anova(g)` as test of the interaction

```r
anova(g, g_add)
```

```r
## Analysis of Variance Table
## Model 1: time ~ poison * treat
## Model 2: time ~ poison + treat
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 36 0.801
## 2 42 1.051 -6 -0.25 1.87 0.11
```
Check the assumptions I

What can we do about unequal variances? Check the normality of errors

```r
qqnorm(g$res, main="Quantiles of residuals vs. normal ones", xlab="", ylab="")
qqline(g$res)
```

Boxplot of residuals
Check the assumptions II

`boxplot(g$res ~ round(g$fitted, 2), main = "Residuals at each poison+treat combination", cex.main=1.5)`
Transform the data

Try to apply a transformation on the data.
1/time can be interpreted as mortality

```r
boxplot(1/time ~ poison + treat, data = rats) # look at data under this transformation
```

![Boxplot](image-url)
Build the model on transformed data

With interaction

```r
gl <- lm(1/time ~ poison * treat, data = rats) #analysis of 1/survival (with interaction)
anova(gl) #interaction not significant
```

```
## Analysis of Variance Table
##
## Response: 1/time
## Df Sum Sq Mean Sq F value Pr(>F)
## poison 2 34.9  17.44  72.63 2.3e-13 ***
## treat  3 20.4   6.80  28.34 1.4e-09 ***
## poison:treat 6  1.6   0.26  1.09  0.39
## Residuals 36  8.6  0.24
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Without interaction

```r
gl <- lm(1/time ~ poison + treat, data = rats) #analysis of 1/survival (without interaction)
anova(gl)
```

```
## Analysis of Variance Table
##
## Response: 1/time
## Df Sum Sq Mean Sq F value Pr(>F)
## poison  2 34.9  17.44  71.7  2.9e-14 ***
## treat   3 20.4   6.80  28.0  4.2e-10 ***
## Residuals 42 10.2  0.24
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
Check the assumption I

```r
qqnorm(g1$res, main="Q-Q plot of residuals of ANOVA of inverses", xlab="", ylab="", cex.main=1.5) # visual test for the normality of residuals
qqline(g1$res)
```

```r
boxplot(g1$res ~ round(g1$fitted,2), main = "Box-plots of residuals of ANOVA of inverses", cex.main=1.5)
```

(Q–Q plot of residuals of ANOVA of inverses)
Check the assumption II

Boxplots of residuals of ANOVA of inverses

# This seems the best one
How to get the groups which differs?

1. Sort the means ($m_1 > m_2 > ... > m_N$)

2. Compare groups using t.test
   
   NB correct for multiple testing
   
   - $G_N$ vs $G_1$; $G_N$ vs $G_2$ ...
   - $G_{N-1}$ vs $G_1$ ...
   - ...

(UNITN-FBK) Analysis of Variance (ANOVA tests) 16 March, 2015 42 / 56
Can we say anything about the difference among treatments?
Summary makes comparisons (via tests of single factor)
It is not correct testing directly all possible differences thus there is an approximate procedure that works well (TukeyHSD)

```r
TukeyHSD(aov(l/time ~ poison + treat, data = rats))
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level

## Fit: aov(formula = l/time ~ poison + treat, data = rats)

## $poison
## diff lwr upr  p adj
## 2-1 0.46864 0.045056 0.89223 0.02716
## 3-1 1.99642 1.572839 2.42001 0.00000
## 3-2 1.52778 1.104198 1.95137 0.00000

## $treat
## diff lwr upr  p adj
## B-A -1.65740 -2.19593 -1.118870 0.00000
## C-A -0.57214 -1.11067 -0.033604 0.03352
## D-A -1.35834 -1.89687 -0.819806 0.00000
## C-B 1.08527 0.54674 1.623799 0.00002
## D-B 0.29906 -0.23947 0.837596 0.45509
## D-C -0.78620 -1.32473 -0.247671 0.00184
```
Testing the variance

How do we check if the variances are homogeneous?

```r
library(lawstat, quietly=TRUE)
levene.test(1/rats$time, rats$treat, location="mean")
```

```
##
## classical Levene's test based on the absolute deviations from the
## mean ( none not applied because the location is not set to median
## )
##
## data: 1/rats$time
## Test Statistic = 0.9185, p-value = 0.4398
```

```r
levene.test(rats$time, rats$treat, location="mean")
```

```
##
## classical Levene's test based on the absolute deviations from the
## mean ( none not applied because the location is not set to median
## )
##
## data: rats$time
## Test Statistic = 6.0747, p-value = 0.001494
```
Select the “best” subset of predictors.

1. Simplest way to explain the data. Among several explanations for a phenomenon, the simplest is the best.
2. Unnecessary predictors will add noise to the estimation. Waste of degree of freedom.
3. Collinearity (too many variables doing the same job).
Stepwise Procedure
Backward Elimination and Forward Selection

Backward Elimination

1. Start with all the predictors in the model.
2. Remove the predictor with highest p-value greater than $\alpha_{crit}$.
3. Refit the model and goto 2
4. Stop when all p-values are less than $\alpha_{crit}$.

The $\alpha_{crit}$ is sometimes called the “p-to-remove” and does not have to be 5%. If prediction performance is the goal, then a 15-20% cut-off may work best, although methods designed more directly for optimal prediction should be preferred.

* Forward Selection is the reverse of the backward elimination.
Example

Try the variable selection procedure on the states dataset collected by US Bureau of the Census. Variables are:

- Population estimate as of July 1, 1975
- Per capita income (1974)
- illiteracy (1970, percent of population)
- life expectancy in years (1969-71)
- murder and non-negligent manslaughter rate per 100,000 population (1976)
- percent high-school graduates (1970)
- mean number of days with min temperature 32 degrees (1931-1960) in capital or large city
- land area in square miles

Consider life expectancy as the response variables and the remaining variables as predictors and fit a linear model.
**Stepwise Procedure**

**Examples**

```r
data(state)
states <- data.frame(state.x77, row.names=state.abb)
regtot <- lm(Life.Exp ~ ., data = states)  # we do not need to write them all.
## It's ok writing ~ . and this will use all variables but the response...
summary(regtot)
```

```r
## Call:
## lm(formula = Life.Exp ~ ., data = states)
#### Residuals:
##    Min     1Q  Median     3Q    Max
## -1.4890 -0.5123 -0.0275  0.5700  1.4945
#### Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.09e+01  1.75e+00  40.59   < 2e-16 ***
## Population 5.18e-05  2.92e-05   1.77   0.083 .
## Income     -2.18e-05  2.44e-04  -0.09   0.929
## Illiteracy  3.38e-02  3.66e-01   0.09   0.927
## Murder     -3.01e-01  4.66e-02  -6.46   8.7e-08 ***
## HS.Grad     4.89e-02  2.33e-02   2.10   0.042 *
## Frost      -5.74e-03  3.14e-03  -1.82   0.075 .
## Area      -7.38e-08  1.67e-06  -0.04   0.965
##---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#### Residual standard error: 0.745 on 42 degrees of freedom
## Multiple R-squared:  0.736, Adjusted R-squared:  0.692
## F-statistic: 16.7 on 7 and 42 DF,  p-value: 2.53e-10
```
Stepwise Procedure

Examples

```r
reg <- update(regtot, ~ . - Area)  # take Area away from predictor variables
summary(reg)
```

```
##
## Call:  
## lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder +  
##     HS.Grad + Frost, data = states)
##
## Residuals:   
##    Min      1Q  Median      3Q     Max  
## -1.4905 -0.5253 -0.0255  0.5716  1.5037
##
## Coefficients:  
##                Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 7.10e+01 1.39e+00  51.17   < 2e-16 ***
## Population  5.19e-05 2.88e-05   1.80    0.079 .
## Income     -2.44e-05 2.34e-04  -0.10    0.934
## Illiteracy  2.85e-02 3.42e-01   0.08    0.934
## Murder     -3.02e-01 4.33e-02  -6.96   1.5e-08 ***
## HS.Grad    4.85e-02 2.07e-02    2.35    0.024 *
## Frost      -5.78e-03 2.97e-03  -1.94    0.058 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.736 on 43 degrees of freedom  
## Multiple R-squared: 0.736, Adjusted R-squared: 0.699      
## F-statistic: 20 on 6 and 43 DF, p-value: 5.36e-11
```

(UNITN-FBK)
Stepwise Procedure

Examples

Remove all the variables with **high** p-values.

```r
reg <- update(regtot, ~ . - Illiteracy - Income - Area - Population)
## remove all variables with high p-value
summary(reg)$r.squared; summary(regtot)$r.squared

## [1] 0.71266
## [1] 0.73616

summary(reg)$adj.r.squared; summary(regtot)$adj.r.squared

## [1] 0.69392
## [1] 0.69218
```

**NB:** The $R^2$ for the full model of 0.736 is reduced only slightly to 0.713 in the final model. Thus the removal of four predictors causes only a **minor reduction** in fit. Beside, removing the `Population` variable could be a “close call” since the p-value is close to the $\alpha_{crit}$ of 5%.

We may want a more robust method to select the **best** variables.
Criterion-based procedures

Idea

With \( p \) predictors there are \( 2^p \) possible models. Fit all the possible models and chose the best one based on some criteria.

The criteria should balance the overfitting and the complexity of the model. Larger models will fit better and so have smaller RSS but use more parameters.

**AIC** Akaike Information Criterion

\[
AIC = -2\log\text{likelihood} + 2p
\]

**BIC** Bayes Information Criterion

\[
BIC = -2\log\text{likelihood} + p\log(n)
\]

NB: For linear regression models, \(-2\log\text{likelihood} = n\log\left(\frac{RSS}{n}\right)\)

These criteria can be used for other models too.

**BIC** penalizes larger models more heavily -> prefers smaller models
Criterion-based procedures

Example

Recall the `regtot` model defined as `regtot <- lm(Life.Exp ~., data=state)`.
Apply AIC (or BIC) criterion to compare models sequentially.

```r
step(regtot)
```

```
## Start: AIC=-22.18
## Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + Frost + Area
##
## Df Sum of Sq RSS  AIC
## - Area 1 0.00 23.3 -24.2
## - Income 1 0.00 23.3 -24.2
## - Illiteracy 1 0.00 23.3 -24.2
## <none> 1 23.3 -22.2
## - Population 1 1.75 25.0 -20.6
## - Frost 1 1.85 25.1 -20.4
## - HS.Grad 1 2.44 25.7 -19.2
## - Murder 1 23.14 46.4 10.3
##
## Step: AIC=-24.18
## Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + Frost
##
## Df Sum of Sq RSS  AIC
## - Illiteracy 1 0.00 23.3 -26.2
## - Income 1 0.01 23.3 -26.2
## <none> 1 23.3 -24.2
## - Population 1 1.76 25.1 -22.5
## - Frost 1 2.05 25.3 -22.0
## - HS.Grad 1 2.98 26.3 -20.2
## - Murder 1 23.14 46.4 10.3
##
## Step: AIC=-26.17
## Life.Exp ~ Population + Murder + HS.Grad + Frost
##
## Df Sum of Sq RSS  AIC
## - Income 1 0.0 23.3 -28.2
## <none> 1 23.3 -26.2
## - Population 1 2.1 25.4 -25.9
## - Frost 1 3.1 26.4 -23.9
## - HS.Grad 1 5.1 28.4 -20.2
## - Murder 1 34.8 58.1 15.5
##
## Call:
## lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost, data = states)
##
## Coefficients:
## (Intercept) Population Murder HS.Grad Frost
## 7.10e+01 5.01e-05 -3.00e-01 4.66e-02 -5.94e-03
```
Adjusted $R^2$

\[
R^2_a = 1 - \frac{RSS}{TSS} = 1 - \left( \frac{n-1}{n-p} \right) (1 - R^2)
\]

Adding a variable to a model can only decrease the RSS and so only increase the $R^2$ so $R^2$ by itself is not a good criterion because it would always choose the **largest possible model**.

Adding a predictor will only increase $R^2_a$ if it has some value.

Recall $R^2$:

\[
R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (\hat{y}_i - \bar{y})^2}
\]
Predicted residual sum of squares (PRESS)

\[ RSS = \| Y - \hat{Y} \|^2 \]

may not represent adequately the predictive power of the model on other data. Predicted residual sum of squares (PRESS) is

\[ \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 \]

where \( \hat{y}_{(i)} \) is the value predicted for \( x = x_i \) using the model on a dataset without the point \( x_i \), i.e.

\[ \hat{\beta}_{(i)} = (X_{(i)}^t X_{(i)})^{-1} X_{(i)}^t Y_{(i)} \]

where the subscript \( (i) \) means that the \( i \)-th row and column have been deleted. It turns out

\[ y_i - \hat{y}_{(i)} = \frac{y_i - \hat{y}_i}{1 - H_{ii}}. \]
Data science theoretical projects

- Deep Learning mathematical analysis applied on -omics or other big data.
- Multiplex networks mathematical analysis and distance analysis and application on real datasets.
- Embedding network models in machine learning for data integration.

Data science projects

- Integration and analysis of genetic data and geographical information.
- Project on http://opendata.unesco.org data. Prediction of expenses and project evaluation.
- Energy consumption models modulated by environmental factors.

Contact: filosi@fbk.eu
1. We would like to investigate the effectiveness of various feed supplements (feed) on the growth rate (weight) of chickens. Use boxplots and a plot of means to visualize the difference between feed types. Use ANOVA to examine the effectiveness of feed supplements. Comment on your findings and appropriateness of your assumptions.

2. We believe that mean urinary excretion rate of Pregnanetriol changes based on the underlying type of Cushing's syndrome. Investigate whether there is statistically significant mean difference for this steroid metabolites.

3. For the rat genotype data discussed during the lesson, use one-way ANOVA to investigate whether weight gain (Wt) of the litter (in grams) at age 28 days is related to mother's genotype (Mother). Repeat the analysis for the relationship between weight gain (Wt) and genotype of the litter (Litter). Compare the plot of means for the first analysis to that of the second one.

4. Load the anorexia data set from the MASS package. This data set was collected to investigate the effectiveness of different treatments (Treat) on increasing weight for young female anorexia patients. Create a new variable called Difference by subtracting the weight of patient before study period (Prewt) from her weight after the study period (Postwt): Difference = Postwt - Prewt. Use a plot of means to visualize how this variable changes depending on the type of treatment. Use ANOVA to investigate whether the type of treatment makes a difference in the amount of weight gain.

5. The data set cabbages available from the MASS package include a study on comparing ascorbic acid content between two different cultivars of cabbage. In this data set, the two different cultivars were planted on three different dates, denoted as d16, d20, or d21. The variable Data is a factor that specifies the planting date for each cabbage. Use two-way ANOVA to evaluate the relationship between the vitamin C content and cultivars while controlling for the effect of planting dates.