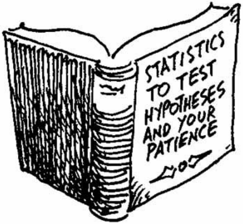
Quick Overview of Traditional Methods: Linear regression & hypothesis testing

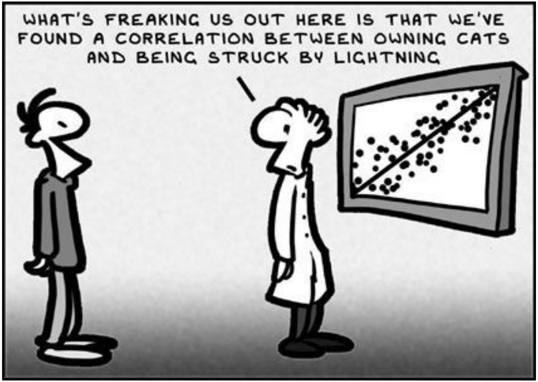
Dr. Marijke Welvaert



What I will cover today

1. Testing for associations between continuous variables
2. Testing for mean differences
3. Within-subject designs: old school
4. Within-subject designs extension

Testing for associations between continuous variables



Correlations

Pearson's ρ

```
> x <- c(165,109,132,121,115,143,119,128,132,158)
> y <- c(124,135,119,121,139,153,101,125,106,99)
> cor(x, y)

[1] -0.1790753
```

Spearman's r

```
> cor(x, y, method="spearman")

[1] -0.2553203
```

Kendall's τ

```
> cor(x, y, method="kendall")

[1] -0.1797866
```

Neither `cor()` or `cov` produce tests of significance, although you can use the `cor.test()` function to test a single correlation coefficient.

General linear model: Overview

The general linear model is defined as

$$Y = \beta X + \varepsilon$$

with

- Y: Dependent variable (continuous)
- β: Estimated coefficients
- X: Design matrix with 1 to p independent variables (continuous/categorical)
- ε: Residual error

Assumptions

1. Linear relationship
2. $\varepsilon \overset{i.i.d.}{\sim} N(0,1)$: Normality, non-dependency, homoscedasticity

Multiple regression, ANOVA and ANCOVA all belong under this umbrella!

Linear regression

Example data: Manatees in Florida

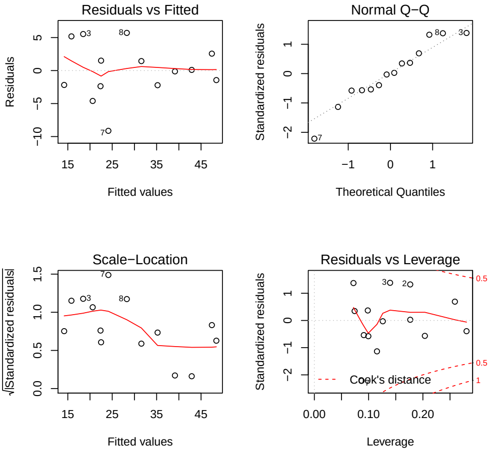
```
> speed <- c(447,460,481,498,513,512,526,559,585,614,
+       645,675,711,719)
> seacow <- c(12,21,24,16,24,20,15,34,33,33,39,43,50,47)
> fit <- lm(seacow ~ speed)
> fit

Call:
lm(formula = seacow ~ speed)

Coefficients:
(Intercept)        speed
    -42.125        0.126

> coef(fit)

(Intercept)       speed
 -42.124616    0.125959
```

Linear regression

```
> summary(fit)
Call:
lm(formula = seacow ~ speed)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.1298 -2.2054 -0.0084  2.3027  5.7135 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -42.12462    7.47213  -5.638 0.000109 ***
speed         0.12596    0.01301   9.682 5.07e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.311 on 12 degrees of freedom
Multiple R-squared:  0.8865,	Adjusted R-squared:  0.8771 
F-statistic: 93.75 on 1 and 12 DF,  p-value: 5.071e-07
> confint(fit)
                   2.5 %      97.5 %
(Intercept) -58.40498410 -25.8442488
speed         0.09761426   0.1543038
```

Diagnostic plots

```
> par(mfrow=c(2,2))
> plot(fit)
```



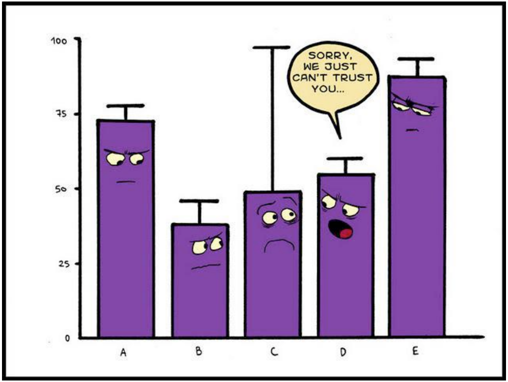
Linear regression

Predicting a new value

What is the expected number of killed Manatees if 800,000 speed boats are registered in a year?

```
> new <- data.frame(speed=c(800))
> predict(fit, new, interval="prediction")

       fit      lwr      upr
1 58.64262 46.89716 70.38808
```

Testing for differences between means



Comparing 2 means: t-test

Toy data:

```
> x <- c(165,109,132,121,115,143,119,128,132,158)
```

One Sample t-test

```
> t.test(x, mu=120, alternative="two.sided",
+                    conf.level=0.95)

        One Sample t-test

data:  x
t = 2.1097, df = 9, p-value = 0.0641
alternative hypothesis: true mean is not equal to 120
95 percent confidence interval:
 119.1185 145.2815
sample estimates:
mean of x
    132.2
```

Comparing 2 means: t-test

Toy data:

```
> x <- c(165,109,132,121,115,143,119,128,132,158)
> y <- c(124,135,119,121,139,153,101,125,106,99)
```

Two Sample t-test

```
> t.test(x, y, alternative="two.sided",
+                    conf.level=0.95)

        Welch Two Sample t-test


data:  x and y
t = 1.2591, df = 17.935, p-value = 0.2241
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -6.690478 26.690478
sample estimates:
mean of x mean of y
    132.2     122.2
```

Comparing 2 means: t-test

Equality of variances assumption:

```
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 1.1282, num df = 9, denom df = 9, p-value = 0.8603
alternative hypothesis: true ratio of variances is not equal t
95 percent confidence interval:
 0.2802302 4.5421473
sample estimates:
ratio of variances
          1.128205
```

Comparing 2 means: t-test

Two Sample t-test

```
> t.test(x, y, alternative="two.sided",
+          var.equal=TRUE, conf.level=0.95)

        Two Sample t-test

data:  x and y
t = 1.2591, df = 18, p-value = 0.2241
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -6.686136 26.686136
sample estimates:
mean of x mean of y
    132.2     122.2
```

Comparing 2 means: t-test

Toy data:

```
> x1 <- c(165,109,132,121,115,143,119,128,132,158)
> x2 <- c(124,135,119,121,139,153,101,125,106,99)
```

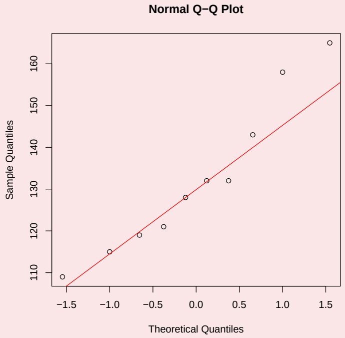
Paired Sample t-test

```
> t.test(x1, x2, alternative="greater",
+                  paired=TRUE, conf.level=0.95)

        Paired t-test


data:  x1 and x2
t = 1.1597, df = 9, p-value = 0.138
alternative hypothesis: true difference in means is greater th
95 percent confidence interval:
 -5.806872       Inf
sample estimates:
mean of the differences
                     10
```

Normality assumption

QQ-plot

```
> qqnorm(x)
> qqline(x, col="red")
```



Normality assumption

Testing for normality

```
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.93085, p-value = 0.4563
```

Note: This test behaves badly with small sample sizes!

Non-parametric alternative

```
> wilcox.test(x1, x2, paired=TRUE)

        Wilcoxon signed rank test with continuity correction

data:  x1 and x2
V = 31.5, p-value = 0.3135
alternative hypothesis: true location shift is not equal to 0
```

ANOVA

Example data: Medicine trial in depressed patients

```
> condition <- c(rep(1,8), rep(2,8), rep(3,8), rep(4,8))
> fcond <- factor(condition, labels=c("placebo", "10mg", "20mg", "30mg"))
> vertigo <- c(25,27,28,23,21,25,29,28,23,20,29,27,26,28,29,25,19,
+       16,21,20,18,19,21,22,16,17,21,24,23,22,18,25)
> fit <- lm(vertigo ~ fcond)
```

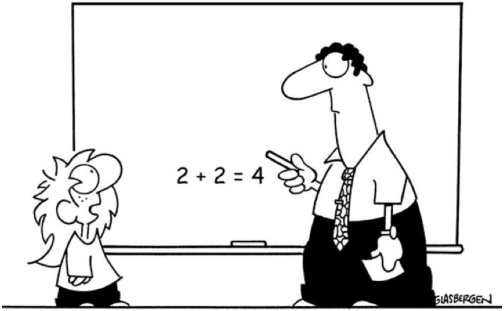
ANOVA

```
> summary(fit)

Call:
lm(formula = vertigo ~ fcond)

Residuals:
   Min     1Q Median     3Q    Max 
-5.875 -1.812  0.375  2.250  4.250 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   25.750      1.009  25.530  < 2e-16 ***
fcond10mg      0.125      1.426   0.088  0.93079    
fcond20mg     -6.250      1.426  -4.382  0.00015 ***
fcond30mg     -5.000      1.426  -3.505  0.00155 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.853 on 28 degrees of freedom
Multiple R-squared:  0.5377,	Adjusted R-squared:  0.4882 
F-statistic: 10.86 on 3 and 28 DF,  p-value: 6.644e-05
```

ANOVA

```
> anova(fit)

Analysis of Variance Table

Response: vertigo
          Df Sum Sq Mean Sq F value    Pr(>F)
fcond      3 265.09  88.365  10.858 6.644e-05 ***
Residuals 28 227.88   8.138
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> par(mfrow=c(2,2))
> plot(fit)
```

Within-subject designs: old school



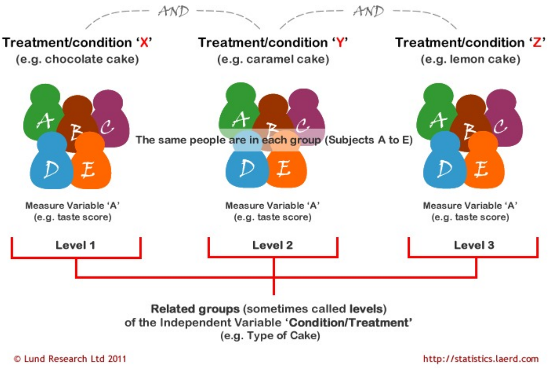
"How can I trust your information when you're using such outdated technology?"

Within-subject designs

- Repeated measures designs
- Longitudinal designs
- Cross-over designs

Any dataset in which you measured the same subject at least twice on a certain variable

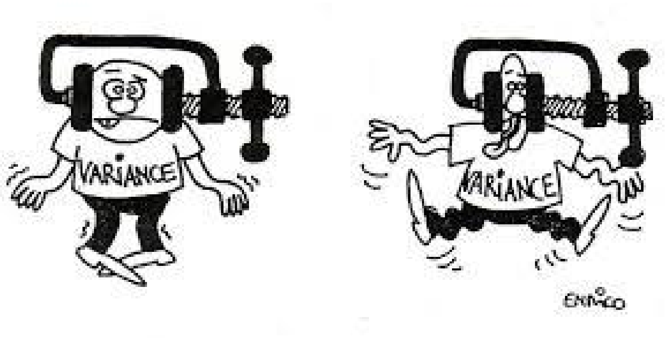
Repeated Measures ANOVA



Assumptions

1. Balanced design
2. Multivariate normal distribution
3. Sphericity (≈ variance homogeneity)

Within-subject designs extension



The linear mixed model as an extension of linear regression

Standard linear regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

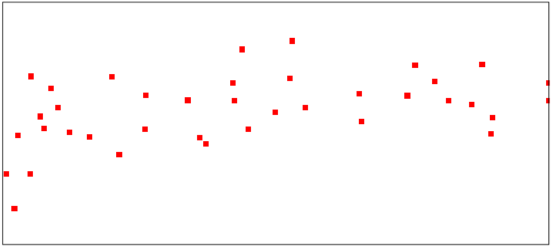
with $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma_\varepsilon)$

Linear mixed model

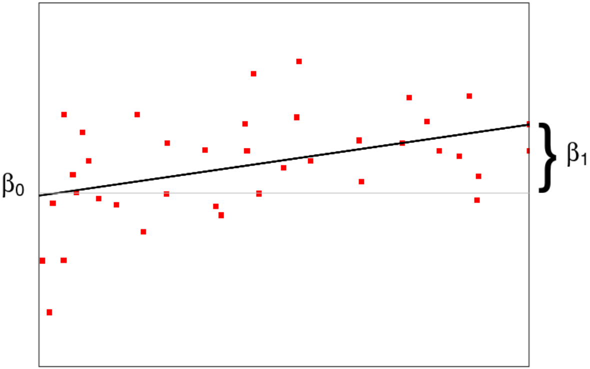
$$Y_i = \beta_0 + \beta_1 X_i + b_i + \varepsilon_i$$

with $b_i \sim N(0, \sigma_b)$ and $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma_\varepsilon)$

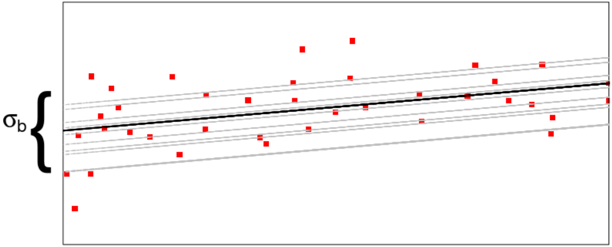
The linear mixed model: Graphical interpretation



The linear mixed model: Graphical interpretation



The linear mixed model: Graphical interpretation



When to consider a Linear Mixed Model?

- missing data
- unbalanced designs
- more measurements than subject ($n < p$)
- control of "unwanted" variation
- ...

Example: Supernova 10K walking trials

Burke et al. (2017). Low Carbohydrate, High Fat diet impairs exercise economy and negates the performance benefit from intensified training in elite race walkers. *The Journal of Physiology*, 595(9), 2785–2807.

- Effect of diet on walking economy
- 3 diet interventions implemented over 2 camps
- Pre- and Post-camp 10K races
- Partial cross-over of subjects between camps
- Substantially different racing conditions within and between camps
- Missing data

Example: Linear Mixed Model solution

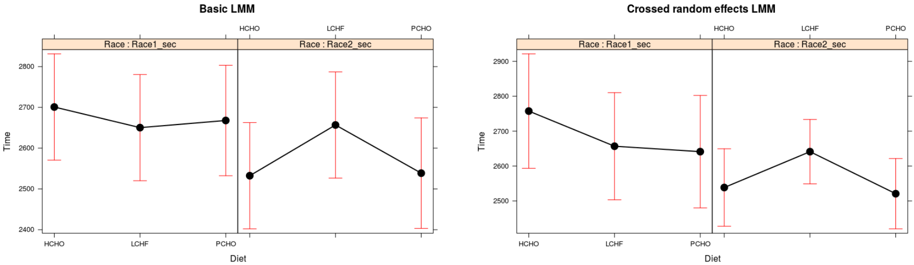
Accounting for 3 sources of variation: Subject, Race and Camp

$$
\begin{aligned}
Y_i = \beta_0 + \beta_1 \text{ Diet}_i + \beta_2 \text{ Race}_i + \beta_3 \text{ Camp}_i \\
+ \beta_4 \text{ Diet}_i \times \text{Race}_i + \beta_5 \text{ Race}_i \times \text{Camp}_i \\
+ b_{1i} + b_{2i} \text{ Race}_i + b_{3i^\star} + b_{4i^\star} \text{ Camp}_{i^\star} + \varepsilon_i
\end{aligned}
$$

lme4

Bates et al. 2015, JSS

Example: Results



Example: R code

```
> ## Load required packages
> library(lme4)
> library(reshape2)
> library(car)
> ## Read data
> data <- read.table("Diet_data_10Ktrial.csv",
+         header=TRUE, sep=",", dec=".")
> data.trim <- data[, -c(4:5,7:8)]
> ## Transform data from wide format to long format
> data.long <- melt(data.trim,
+           id.vars=c("subject", "camp", "Diet"),
+           measure.vars=c("Race1_sec","Race2_sec"),
+           variable.name="Race", value.name="Time")
> # convert camp into factor/make LCHF reference
> data.long$camp <- factor(data.long$camp)
> data.long$Diet <- relevel(data.long$Diet, ref="LCHF")
```

Example: R code

```
> fit <- lmer(Time ~ (Diet + Race + camp)^2 +
+        (Race|subject) + (camp|subject), data=data.long)
> print(fit, corr=FALSE)
Linear mixed model fit by REML ['lmerMod']
Formula:
Time ~ (Diet + Race + camp)^2 + (Race | subject) + (camp | sub
   Data: data.long
REML criterion at convergence: 558.5919
Random effects:
 Groups    Name          Std.Dev. Corr
 subject   (Intercept)   249.64
           RaceRace2_sec 180.95   -1.00
 subject.1 (Intercept)   192.87
           camp2         170.97   -1.00
 Residual                 77.77
Number of obs: 52, groups:  subject, 18
Fixed Effects:
```

Example: R code

```
> summary(fit)$coefficients
                           Estimate Std. Error    t value
(Intercept)              2749.87891  115.18479 23.8736291
DietHCHO                 -162.25439  166.00956 -0.9773798
DietPCHO                 -159.40113  152.26069 -1.0468962
RaceRace2_sec              70.28972   64.44298  1.0907272
camp2                    -142.40756  111.90451 -1.2725810
DietHCHO:RaceRace2_sec   -203.37860   63.10971 -3.2226196
DietPCHO:RaceRace2_sec   -104.86568   63.21595 -1.6588484
DietHCHO:camp2            402.11982  185.88650  2.1632545
DietPCHO:camp2            219.99762  170.50420  1.2902768
RaceRace2_sec:camp2      -131.34933   49.79539 -2.6377809
```

Example: R code

```
> Anova(fit, test="F")

Analysis of Deviance Table (Type II Wald F tests with Kenward-

Response: Time
                F Df Df.res   Pr(>F)
Diet       1.3728  2 16.683 0.28059
Race       6.8797  1 24.011 0.01491 *
camp       0.4602  1 12.199 0.51018
Diet:Race  4.3131  2 13.335 0.03592 *
Diet:camp  2.2025  2 15.468 0.14398
Race:camp  6.0796  1 10.109 0.03311 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```