ADVANCED STATISTICAL MODELING

Version 1, 2018

Jacolien van Rij [j.c.van.rij@rug.nl]
Hermine Berberyan, and Stefan Huijser

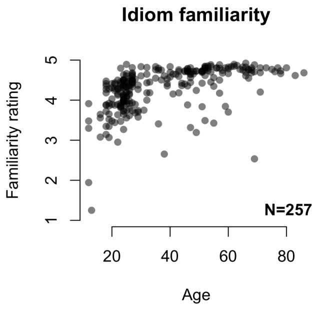
Today's topic

INTRODUCTION LINEAR MODEL

EXAMPLE DATA

- ❑ Idiom familiarity rating

Does idiom vocabulary change when getting older?



(Sprenger, la Roi, & van Rij, submitted)

HYPOTHESIS TESTING

- ❑ Statistical hypothesis ≠ research hypothesis
- ❑ Example:
  - RH: Idiom vocabulary increases with age.

    Older participants will know more idioms than younger participants.
  - SH: Difference idiom vocabulary$_{old}$ – Idiom vocabulary$_{young}$ is larger than 0.
    - $H_0 = 0$
    - $H_1 > 0$ ← older participants know more idioms
    - $H_2 < 0$ ← younger participants know more idioms

STATISTICAL TEST

- Calculates the probability that H_0 is true
  - α : significance level of test
- Errors in statistical conclusions:

	decision: retain H_0	reject H_0
H_0 is true	correct decision	**type I error**
H_0 is false	**type II error**	correct decision

STATISTICAL TEST

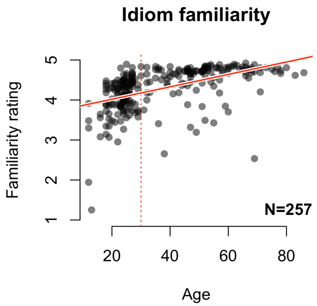
- Calculates the probability that H_0 is true
  - α : significance level of test
  - β : power of test
- Errors in statistical conclusions:

	decision: retain H_0	reject H_0
H_0 is true	$1-\alpha$	α (type I error)
H_0 is false	β (type II error)	$1-\beta$

EXAMPLE DATA

- Idiom familiarity rating

Does idiom vocabulary change when getting older?



(Sprenger, la Roi, & van Rij, submitted)

LINEAR REGRESSION

- **Formula:**

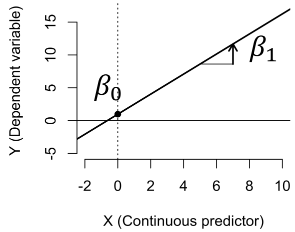
model

$$y_i = \boxed{\beta_0 + \beta_1 x_i} + \varepsilon_i$$

β_0= intercept

β_1= slope

$\varepsilon_i \sim N(0, \sigma)$



LINEAR REGRESSION

```
m1 <- lm(Rating ~ Age, data=subdat)
summary(m1)
```


LINEAR REGRESSION

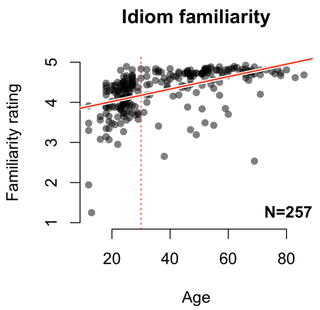
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

```
Call:
lm(formula = meanRating ~ Age, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6610 -0.1779  0.1022  0.2911  0.7960 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.709178   0.068551  54.108   <2e-16 ***
Age         0.015522   0.001651   9.403   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4607 on 255 degrees of freedom
Multiple R-squared:  0.2575,	Adjusted R-squared:  0.2546 
F-statistic: 88.42 on 1 and 255 DF,  p-value: < 2.2e-16
```

estimated coefficients

LINEAR REGRESSION

y = 3.709 + 0.0155*Age



(Sprenger, la Roi, & van Rij, submitted)

CENTERING

```
subdat$cAge <-
      subdat$Age – median(subdat$Age)

m1 <- lm(Rating ~ cAge, data=subdat)
summary(m1)
```

LINEAR REGRESSION

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

```
Call:
lm(formula = meanRating ~ cAge, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6610 -0.1779  0.1022  0.2911  0.7960

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.174834   0.031423 132.858   <2e-16 ***
cAge        0.015522   0.001651   9.403   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4607 on 255 degrees of freedom
Multiple R-squared:  0.2575,	Adjusted R-squared:  0.2546
F-statistic: 88.42 on 1 and 255 DF,  p-value: < 2.2e-16
```

$t = \frac{b}{SE(b)}$

estimated coefficients

P-VALUES

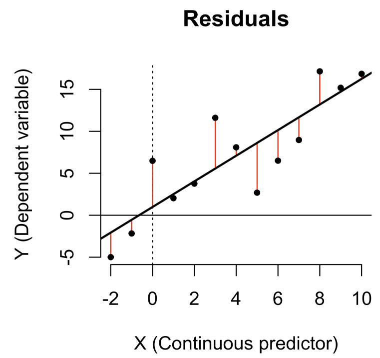
- p = Type I error rate if rejecting H_0
- Navarro, 2017; Table 11.1:

Usual notation	Signif. stars	English translation	The null is...
$p > .05$		The test wasn't significant	Retained
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.	Rejected
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.	Rejected
$p < .001$	***	The test was significant at all levels	Rejected

LINEAR REGRESSION

- Selection coefficients
  - minimizing sum of squared residuals: $\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$
  - residuals = unexplained part of data

LINEAR REGRESSION



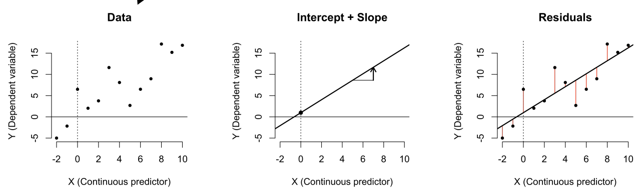
REGRESSION ANALYSIS

DATA = MODEL + RESIDUALS

- **Model:**
  - Explanation / description of data
  - Fitted effects, model estimates
- **Error:**
  - Unexplained part of data
  - Residuals

REGRESSION ANALYSIS

DATA = MODEL + RESIDUALS



LINEAR REGRESSION

model

$$y_i = \boxed{\beta_0 + \beta_1 x_i} + \varepsilon_i$$

```
Call:
lm(formula = meanRating ~ cAge, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6610 -0.1779  0.1022  0.2911  0.7960

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.174834   0.031423 132.858   <2e-16 ***
cAge        0.015522   0.001651   9.403   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4607 on 255 degrees of freedom
Multiple R-squared:  0.2575,	Adjusted R-squared:  0.2546
F-statistic: 88.42 on 1 and 255 DF,  p-value: < 2.2e-16
```

distribution residuals

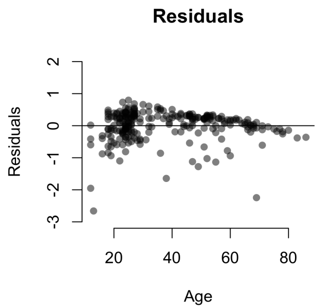
REGRESSION ANALYSIS

- **Assumptions regression analysis:**
  - Residuals are normally distributed
  - No structure in residuals
  - No heteroscedasticity → assumption for Gaussian data
    - Variance should not change with mean

reflects model fit

LINEAR REGRESSION

- ❑ Checking assumptions



LINEAR REGRESSION

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

```
Call:
lm(formula = meanRating ~ cAge, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6610 -0.1779  0.1022  0.2911  0.7960

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.174834   0.031423 132.858   <2e-16 ***
cAge        0.015522   0.001651   9.403   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4607 on 255 degrees of freedom
Multiple R-squared:  0.2575,    Adjusted R-squared:  0.2546
F-statistic: 88.42 on 1 and 255 DF,  p-value: < 2.2e-16
```

distribution residuals

R² VALUE

- ❑ sum of squared residuals: $SS_{\text{res}} = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$
- ❑ total variability: $SS_{\text{tot}} = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$
- ❑ $R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$
- ❑ adjusted $R^2 = 1 - \left(\frac{SS_{\text{res}}}{SS_{\text{tot}}} \times \frac{N-1}{N-K-1}\right)$

degrees of freedom of model (k=number of predictors)

CATEGORICAL PREDICTOR

```
m2 <- lm(Rating ~ cAge+gender,      data=subdat)
summary(m2)
```

CATEGORICAL PREDICTOR

```
Call:
lm(formula = meanRating ~ cAge + gender, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6667 -0.1834  0.1007  0.2924  0.7975 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.180243   0.060327  69.293   <2e-16 ***
cAge         0.015500   0.001667   9.299   <2e-16 ***
gendervrouw -0.007016   0.066749  -0.105    0.916    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4615 on 254 degrees of freedom
Multiple R-squared:  0.2575,	Adjusted R-squared:  0.2517 
F-statistic: 44.04 on 2 and 254 DF,  p-value: < 2.2e-16
```


MODEL COMPARISONS

anova(m1, m2)

```
Analysis of Variance Table

Model 1: meanRating ~ cAge
Model 2: meanRating ~ cAge + gender
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    255 54.111                          
2    254 54.109  1 0.0023534 0.011 0.9164
```


MODEL COMPARISONS

m0 <- lm(Rating ~ 1, data=subdat)

anova(m0, m1)

```
Analysis of Variance Table

Model 1: meanRating ~ 1
Model 2: meanRating ~ cAge
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)    
1    256 72.874                                 
2    255 54.111  1    18.763 88.42 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


MODEL COMPARISONS

anova(m2) ← Do not use the function like this, order matters!

```
Analysis of Variance Table

Response: meanRating
           Df Sum Sq Mean Sq F value Pr(>F)    
cAge        1 18.763 18.7629  88.078 <2e-16 ***
gender      1  0.002  0.0024   0.011 0.9164    
Residuals 254 54.109  0.2130                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```