Statistics for Life Sciences

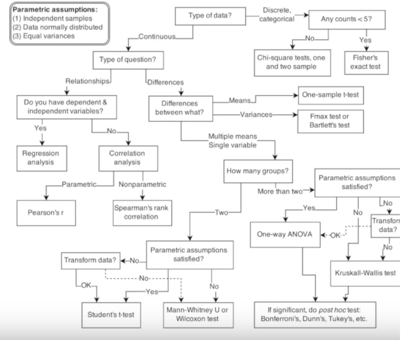
Linear models and interactions

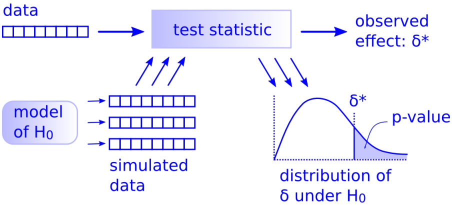
Nathan Harmston

2022-03-18

Hypothesis testing

there is only one test





data

test statistic

observed effect: δ*

model of H_0

simulated data

δ*

p-value

distribution of δ under H_0

http://allendowney.blogspot.com/2016/06/there-is-still-only-one-test.html

Alternatives to the t-test

assume the underlying data is normally distributed AND you aren't sure your samples are large enough to invoke CLT?

Alternatives to the t-test

assume the underlying data is normally distributed AND you aren't sure your samples are large enough to invoke CLT?

- use a nonparametric test - wilcoxon rank sum test - Mann Whitney - ?wilcox.test
- permutation tests
- bootstrap

Bootstrapping

1. make a bootstrapped dataset - take the data and sample with replacement
2. calculate a statistic
3. keep track of it
4. repeat lots of times

histogram tells us what might happen if we repeated the experiment lots of times

assumption: your sample is random

Bootstrapping

what is sampling with replacement?

Bootstrapping

what is sampling with replacement?

how is this different to permutation testing?

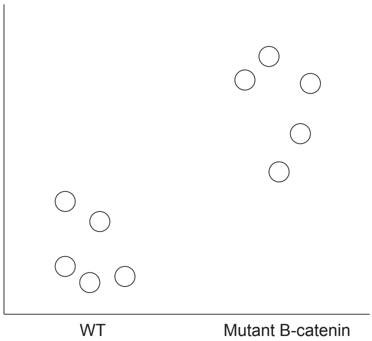
lets try coding bootstrapping

- Bootstrapping is great for estimating the confidence intervals of test statistics.
- Permutation testing is best used for testing hypotheses.

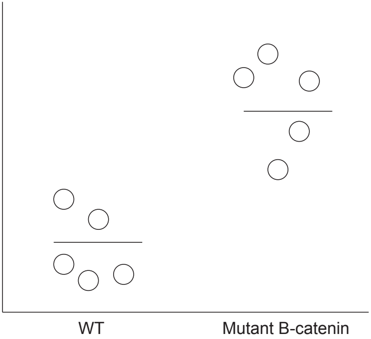
Relationships between linear models and ANOVAs

an ANOVA is a special case of a linear model a t-test is a special case of a linear model

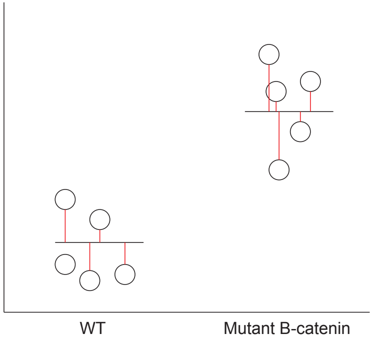
a t-test is a linear model?



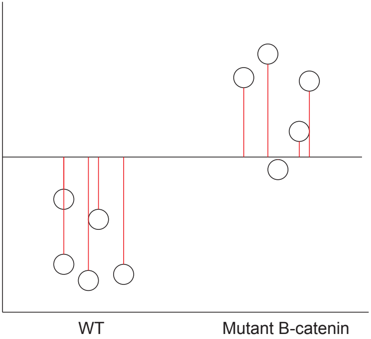
a t-test is a linear model?



a t-test is a linear model?



a t-test is a linear model?



a t-test is a linear model?

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} \tag{1}$$

$$F = \frac{\frac{U}{r_1}}{\frac{V}{r_2}} \tag{2}$$

r_1 = 2 - 1
r_2 = n - 2

lets try this out

- Lecture13.Rmd

ANOVA is a linear model?

- is this the same for ANOVA and linear models ?

Birth Weight and Smoking

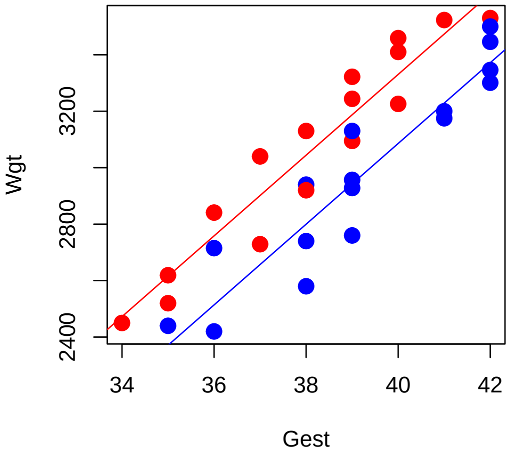
- birth weight (Weight) in grams of baby
- Smoking status (Smoke) of mother (yes or no)
- length of gestation (Gest) in weeks

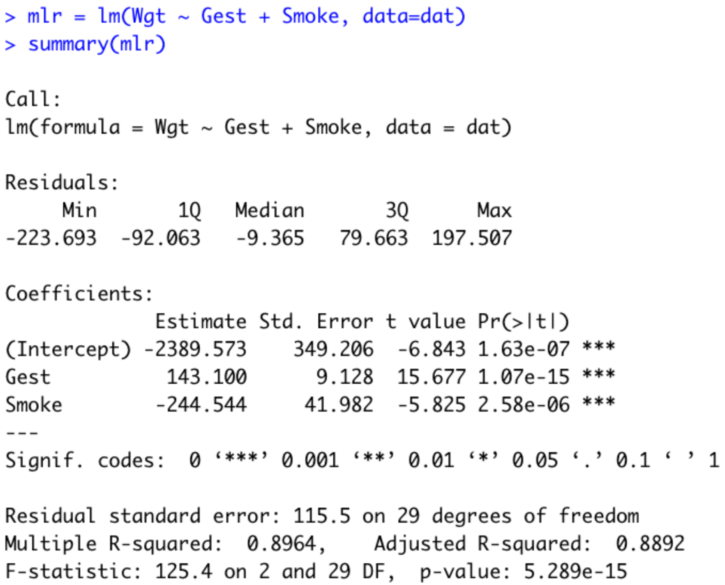
Daniel, (1999)

$$Wgt = -2389.573 + 143Gest - 244.544Smoke$$

(3)

$$Wgt = -2634.117 + 143Gest \tag{4}$$


Wgt
2400
2800
3200
34
36
38
40
42
Gest



```
> mlr = lm(Wgt ~ Gest + Smoke, data=dat)
> summary(mlr)

Call:
lm(formula = Wgt ~ Gest + Smoke, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-223.693  -92.063   -9.365   79.663  197.507 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2389.573    349.206  -6.843 1.63e-07 ***
Gest          143.100      9.128  15.677 1.07e-15 ***
Smoke        -244.544     41.982  -5.825 2.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 115.5 on 29 degrees of freedom
Multiple R-squared:  0.8964,	Adjusted R-squared:  0.8892 
F-statistic: 125.4 on 2 and 29 DF,  p-value: 5.289e-15
```

so smoking has an effect on weight

and gestation length has an effect on weight

both of these lines are parallel

so ...

- the effect of gest on birth weight is the same regardless of smoker or not
- the effect of smoking on birth weight is the same regardless of length of gestation

additive model

If only the world was that easy ....

the world is not ...

- linear
- additive
- normal

If only the world was that easy ....

the world is not ...

- linear
- additive
- normal

need to consider how things may modify a response

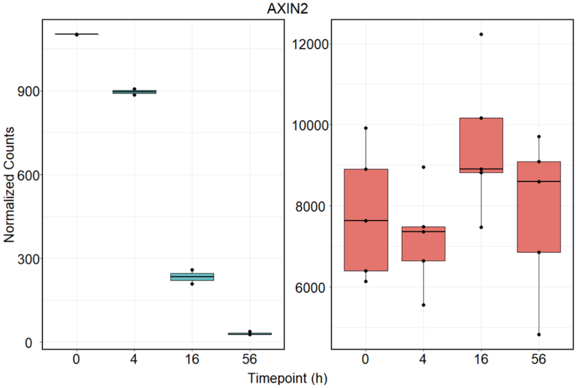
how is an dependent variable changed by each independent variable and their combination?

- interaction effects represent the combined effects of factors on the dependent variable.
- the effect of one factor depends on the level of the other factor

- response to drug is different over time between mutant and WT cells
- weight loss is different between male and female mice for different diets
- number of offspring is different for different species of flies at different temperatures

Example

AXIN2 is a β-catenin dependent gene

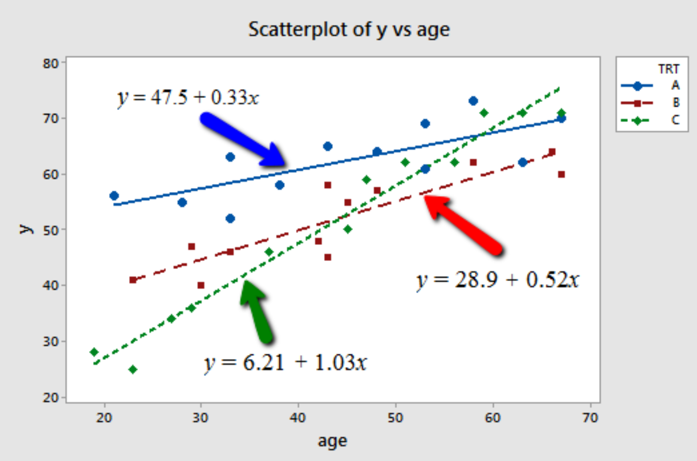


Interaction term

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3} + \epsilon \tag{5}$$

```
mod = aov(y ~ x + z + x*z)
summary(mod)
```

we’re actually building three different models for each treatment?



how would you talk about this?

esoph dataset - esophageal cancer cases.

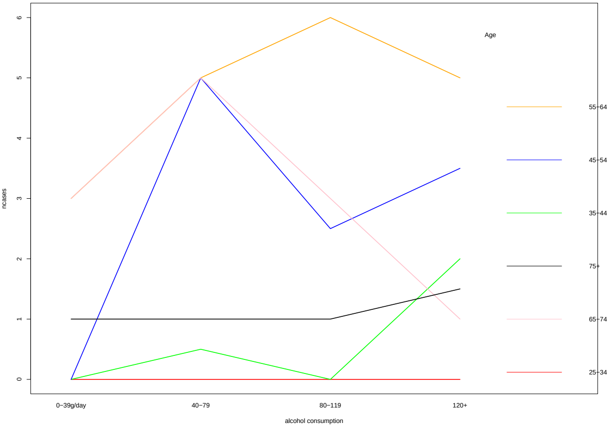
what is the effect of age group (agegp) and alcohol consumption (alcgp) on the number of cases of the cancer (ncases)?

Does the interaction between these two factors affect the number of cases?

what are our hypotheses?

- There is no interaction between the two categorical variables
- the response is the same across all groups for the first factor
- the response is the same across all the groups for the second factor

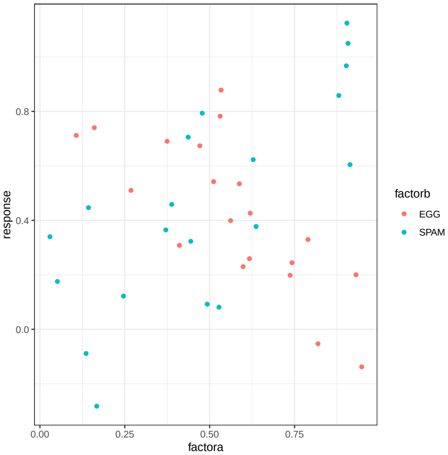
Interaction plots



the effect of A on the response depends on the treatment given

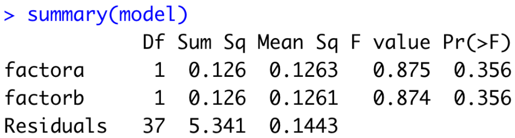
if the lines on the interaction plot are parallel then there is no interaction effect. If the lines intersect then there is likely to be an interaction effect.

what happens when we don't consider interaction terms?



```
data = read.delim("badinteraction.tsv", sep="\t")
```

not considering interaction term

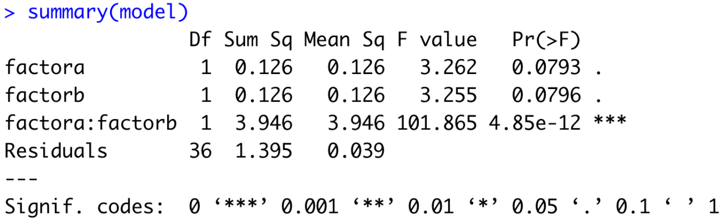


```
> summary(model)
            Df Sum Sq Mean Sq F value Pr(>F)
factora      1  0.126  0.1263   0.875  0.356
factorb      1  0.126  0.1261   0.874  0.356
Residuals   37  5.341  0.1443
```

not considering interaction term

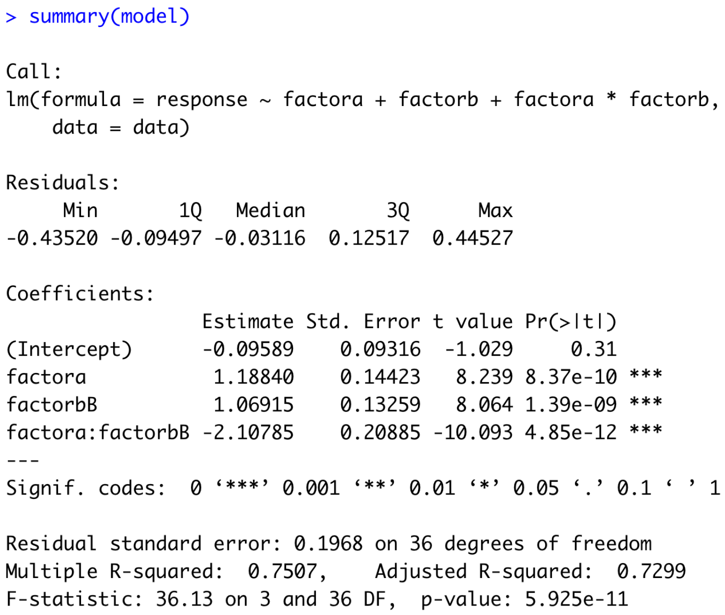
```
> summary(model)

Call:
lm(formula = response ~ factora + factorb, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.6752 -0.2260 -0.0220  0.2432  0.8192 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   0.4764     0.1427   3.339  0.00193 **
factora       0.1830     0.2014   0.909  0.36922   
factorbB     -0.1123     0.1202  -0.935  0.35604   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3799 on 37 degrees of freedom
Multiple R-squared:  0.04513,   Adjusted R-squared:  -0.006482 
F-statistic: 0.8744 on 2 and 37 DF,  p-value: 0.4255
```

considering an interaction term



```
> summary(model)
                Df Sum Sq Mean Sq F value   Pr(>F)    
factora          1  0.126   0.126   3.262   0.0793 .  
factorb          1  0.126   0.126   3.255   0.0796 .  
factora:factorb  1  3.946   3.946 101.865 4.85e-12 ***
Residuals       36  1.395   0.039                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

considering an interaction term



```
> summary(model)

Call:
lm(formula = response ~ factora + factorb + factora * factorb,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.43520 -0.09497 -0.03116  0.12517  0.44527 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -0.09589    0.09316  -1.029     0.31    
factora           1.18840    0.14423   8.239 8.37e-10 ***
factorbB          1.06915    0.13259   8.064 1.39e-09 ***
factora:factorbB -2.10785    0.20885 -10.093 4.85e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1968 on 36 degrees of freedom
Multiple R-squared:  0.7507,    Adjusted R-squared:  0.7299 
F-statistic: 36.13 on 3 and 36 DF,  p-value: 5.925e-11
```

So we can't make statements about the two factors independently

- when factor B is EGG - increasing factor A from low to high decreases the mean response.
- for factor B is SPAM increasing factor A from low to high increases the mean response
- how would you talk about the importance of either factor alone?

interaction effects

interaction effects are very hard to identify - we normally have low power to detect them - require large sample sizes

so typically you don't set out to study these normally

but its normally worth having a quick peak and set up the model with an interaction term and see if that is significant

if interaction term is not significant - examine the main effects, or just re-run without the interaction term

examine the interactions - **how?**

Friday

- logistic regression
- chapters 38 and 42
- problem set 3

next week ...

survival analysis