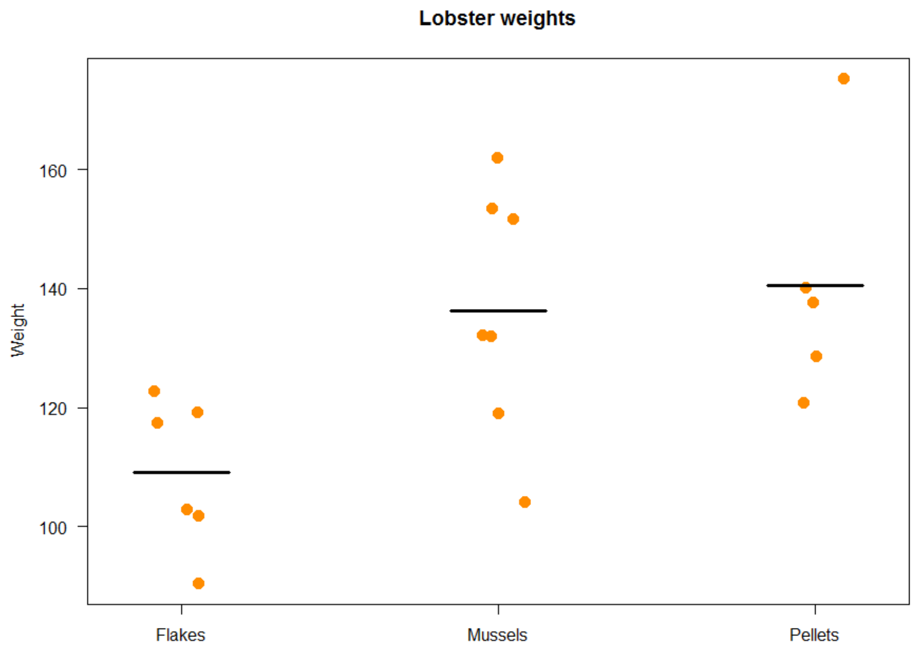
Introduction to Linear Modelling

Anne Segonds-Pichon

v2020-09

Linear modelling is about language

Is there a difference between the 3 diets?

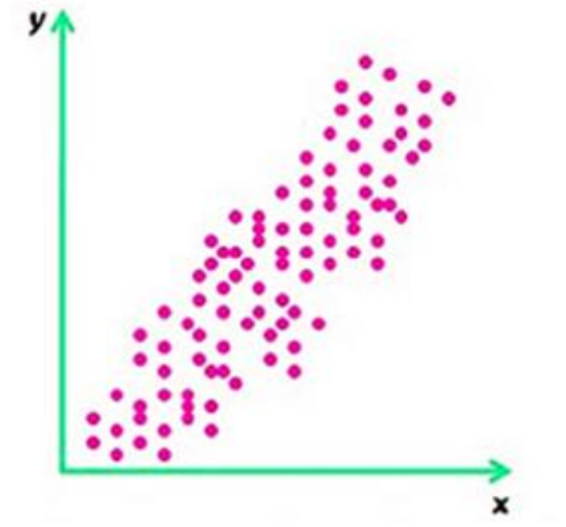


Can diet predict lobster weight?

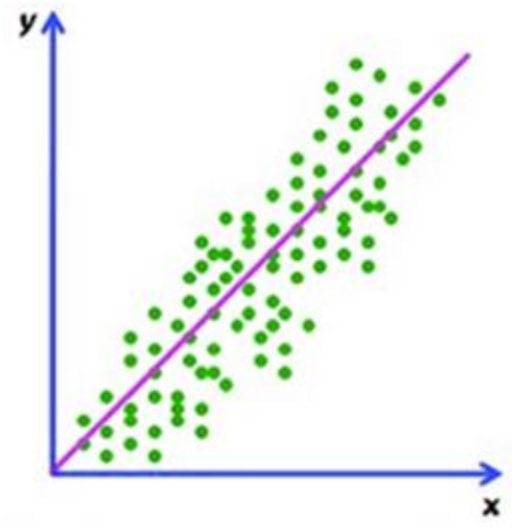
Model(Diet) = Weight

Simple linear model

- Linear regression
  - Correlation: is there an **association** between 2 variables?
  - Regression: is there an **association** and can one variable be used to **predict** the values of the other?



Correlation = Association



Regression = Prediction

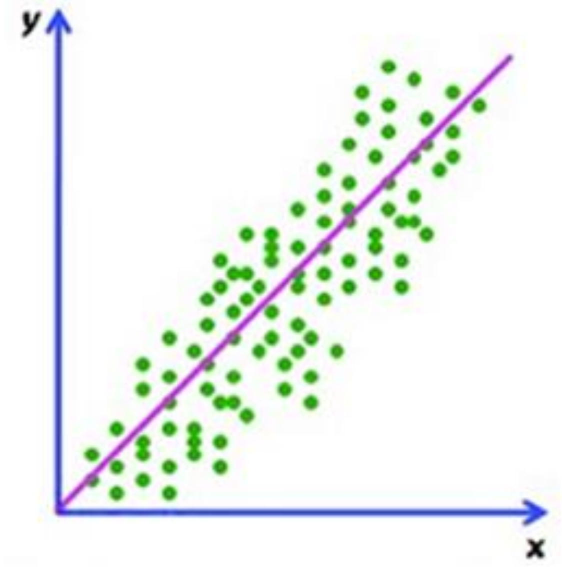
Simple linear model

- Linear regression models the dependence between 2 variables:
  a **dependent y** and a **independent x**.
  - **Causality**
    - **Model(x) = y**

response → y, predictor → x

$$y = \beta_0 + \beta_1 * x$$

Model ($\beta_0 + \beta_1$)



- In R:

Correlation: `cor()`

Linear regression: `lm()`

Linear regression

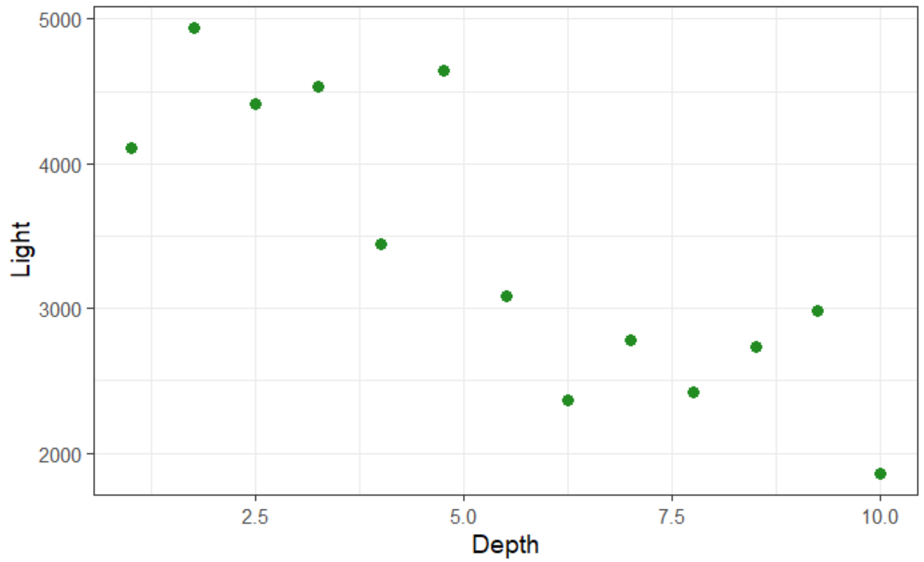
- **Example**: **coniferlight.csv**

```
conifer<-read_csv("coniferlight.csv")
```

Light <dbl>	Depth <dbl>
4105.646	1.00
4933.925	1.75
4416.527	2.50
4528.618	3.25
3442.610	4.00
4640.297	4.75
3081.990	5.50
2368.113	6.25
2776.557	7.00
2419.193	7.75

- Question: how is **light** (lux) affected by the **depth** (m) at which it is measured from the top of the canopy?

$$\textbf{light} = \beta_0 + \beta_1 * \textbf{depth}$$




```
conifer %>%
      ggplot(aes(Depth, Light))+
      geom_point(colour="forestgreen", size=3)
```

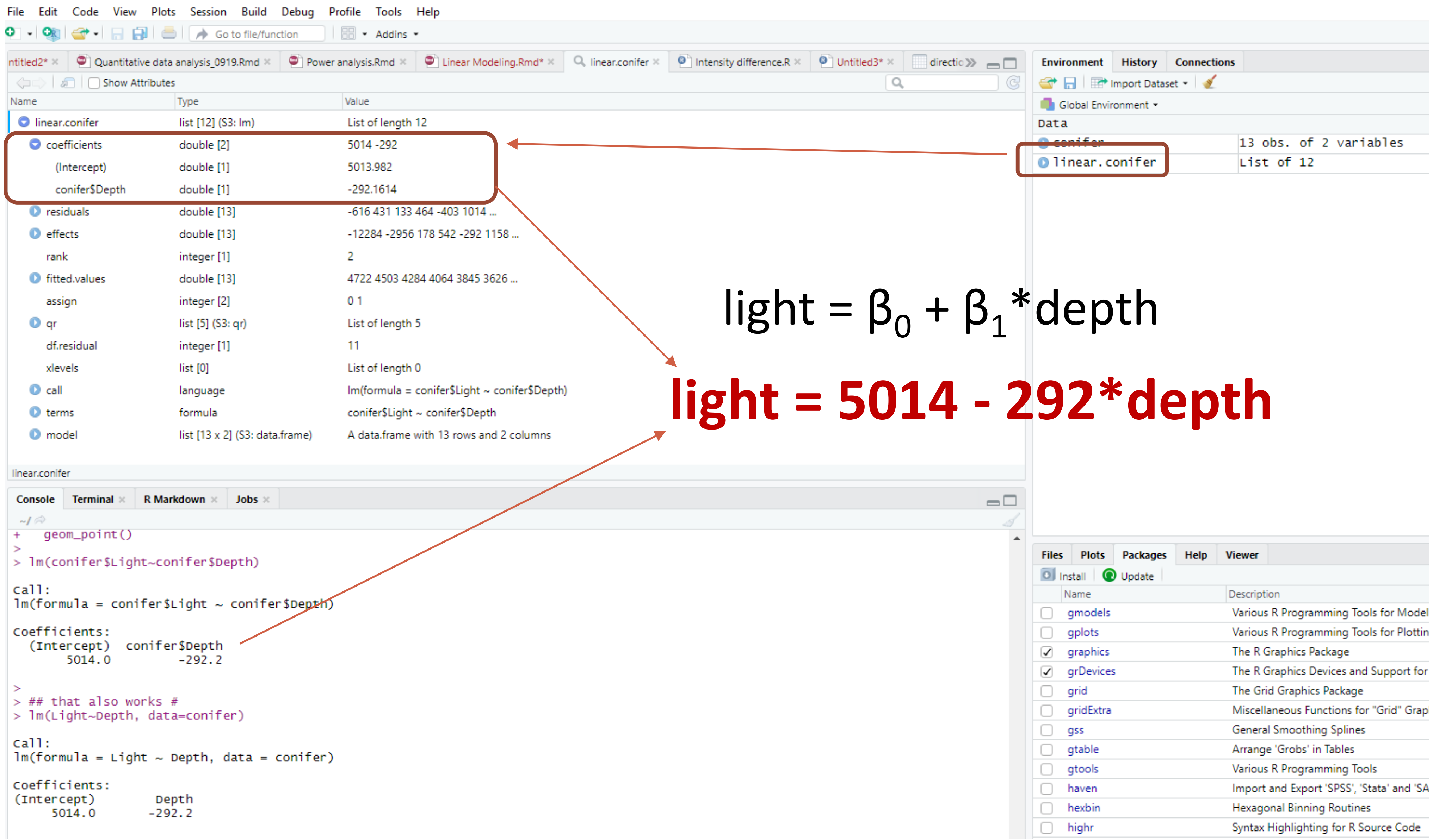
Linear regression

- Linear modelling in R: **`lm(y~x)`**

- Regression: `lm(conifer$Light~conifer$Depth)`
- or: `lm(Light~Depth, data=conifer)`

```
lm(Light~Depth, data=conifer) -> linear.conifer
```

Linear regression



$$light = \beta_0 + \beta_1 * depth$$

light = 5014 - 292*depth

Linear regression

- Line of best fit (= regression line)   light = 5014 - 292*depth

```
geom_abline(intercept= , slope= )
```

```
coefficients()
```

linear.conifer	list [12] (S3: lm)	List of length 12
coefficients	double [2]	5014 -292
(Intercept)	double [1]	5013.982
conifer$Depth	double [1]	-292.1614
residuals	double [13]	-616 431 133 464 -403 1014 ...
effects	double [13]	-12284 -2956 178 542 -292 1158 ...
rank	integer [1]	2
fitted.values	double [13]	4722 4503 4284 4064 3845 3626 ...
assign	integer [2]	0 1
qr	list [5] (S3: qr)	List of length 5
df.residual	integer [1]	11
xlevels	list [0]	List of length 0
call	language	lm(formula = conifer$Light ~ conifer$Depth)
terms	formula	conifer$Light ~ conifer$Depth
model	list [13 x 2] (S3: data.frame)	A data.frame with 13 rows and 2 columns

```
Coefficients:
  (Intercept)  conifer$Depth
       5014.0         -292.2
```

```
cf.abline[1]
```

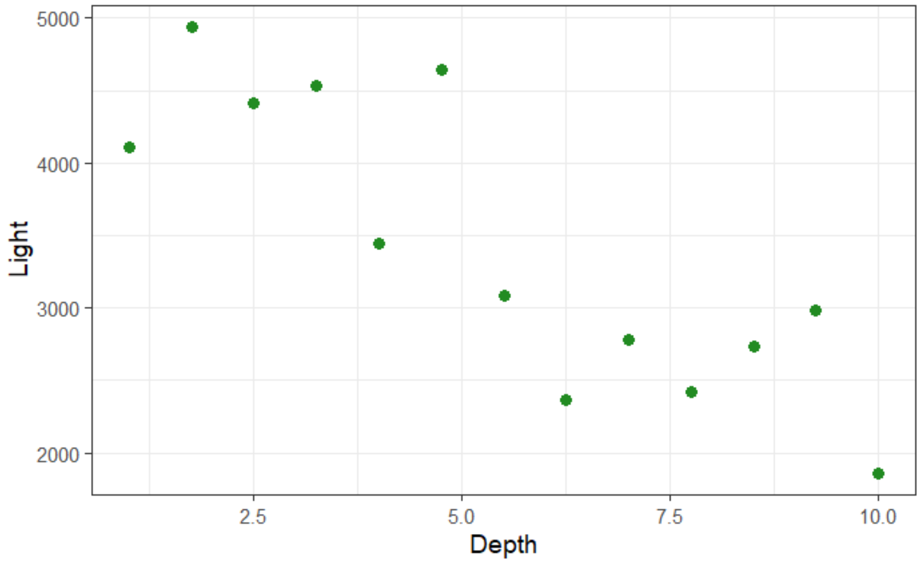
```
(Intercept)
     5014.0
```

It's a vector!

Linear regression

```
coefficients(linear.conifer) -> cf.abline

conifer %>%
      ggplot(aes(Depth, Light))+
            geom_point( colour="forestgreen", size=3)+
            geom_abline(aes(intercept=cf.abline[1], slope=cf.abline[2]))
```

```
Coefficients:
  (Intercept)  conifer$Depth
       5014.0         -292.2
```



light = 5014 - 292*depth

Linear regression

```
lm(Light~Depth, data=conifer) -> linear.conifer
summary(linear.conifer)
```

```
Coefficients:
 (Intercept)  conifer$Depth
      5014.0         -292.2
```

```
Call:
lm(formula = conifer$Light ~ conifer$Depth)

Residuals:
   Min     1Q Median     3Q    Max
-819.9 -330.5 -192.3  431.2 1014.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5013.98     342.15  14.654 1.46e-08 ***
conifer$Depth  -292.16      55.41  -5.272 0.000263 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 560.7 on 11 degrees of freedom
Multiple R-squared:  0.7165,    Adjusted R-squared:  0.6907
F-statistic:  27.8 on 1 and 11 DF,  p-value: 0.0002633
```

p-value

Coefficient of determination

Linear regression

- Coefficient of determination:
  - R-squared **(r^2)**:
    - It quantifies the proportion of variance in Y that can be explained by X, it can be expressed as a percentage.
    - e.g. here **71.65%** of the variability observed in light is explained by the depth at which it is measured in a conifer tree.

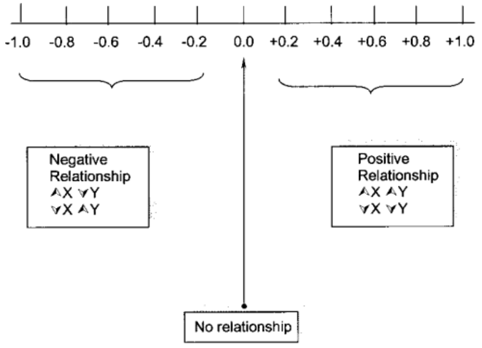
```
Residual standard error: 560.7 on 11 degrees of freedom
Multiple R-squared:  0.7165,	Adjusted R-squared:  0.6907
F-statistic:  27.8 on 1 and 11 DF,  p-value: 0.0002633
```

- r: coefficient of correlation between x (depth) and y (light)
  - e.g. here: **r = -0.846** so **r^2** = -0.846 * -0.846 **= 0.716 = R-squared**

```
> head(conifer)
     Light Depth
1 4105.646  1.00
2 4933.925  1.75
3 4416.527  2.50
4 4528.618  3.25
5 3442.610  4.00
6 4640.297  4.75
```

```
conifer %>%
  cor_test(Light, Depth)
```

var1	var2	cor	statistic	p
<chr>	<chr>	<dbl>	<dbl>	<dbl>
Light	Depth	-0.85	-5.272411	0.000263



Linear regression

```
summary(linear.conifer)
```

```
Call:
lm(formula = conifer$Light ~ conifer$Depth)

Residuals:
   Min     1Q Median     3Q    Max 
-819.9 -330.5 -192.3  431.2 1014.1 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    5013.98     342.15  14.654 1.46e-08 ***
conifer$Depth  -292.16      55.41  -5.272 0.000263 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 560.7 on 11 degrees of freedom
Multiple R-squared:  0.7165,	Adjusted R-squared:  0.6907 
F-statistic:  27.8 on 1 and 11 DF,  p-value: 0.0002633
```

```
anova(linear.conifer)
```

```
      Analysis of Variance Table

      Response: conifer$Light
                    Df  Sum Sq Mean Sq F value    Pr(>F)    
Model conifer$Depth  1 8738553 8738553  27.798 0.0002633 ***
Error Residuals     11 3457910  314355                      
      ---
      Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Total amount of variability:
8738553 + 3457910 = 12196463

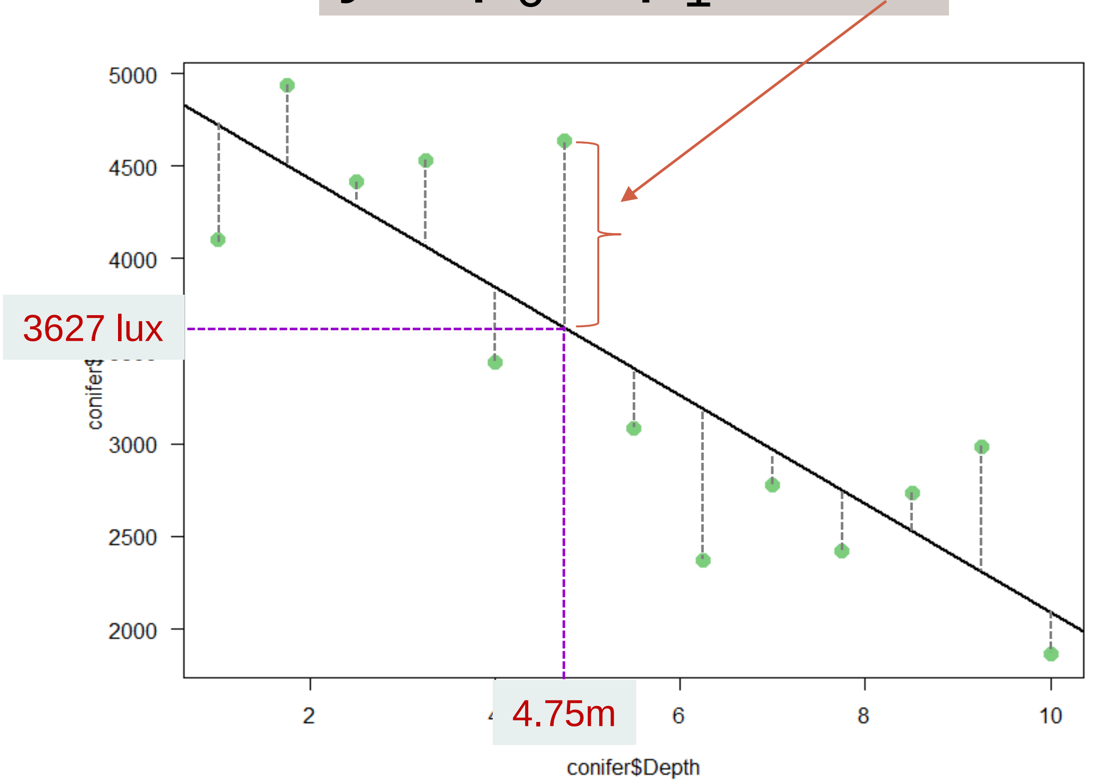
Proportion explained by depth:
8738553/12196463 = **0.716**

Linear regression the error Ɛ

- Depth predicts about 72% (R-Squared) of the variability of light
  - so 28% is explained by other factors (e.g. Individual variability…)
- Example: the model predicts 3627 lux at a depth of 4.75 m in a conifer.

3627 = 5014 - 292*4.75

$$y = \beta_0 + \beta_1 * x + \varepsilon$$



```
linear.conifer <-lm(Light~Depth, data=conifer)
```

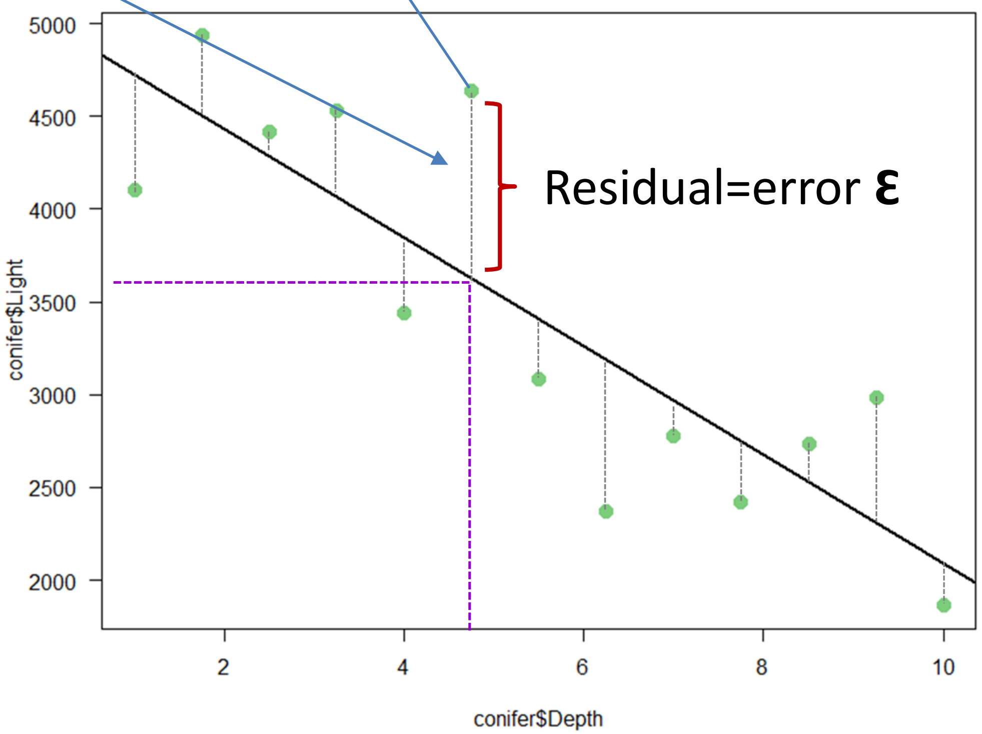
linear.conifer	list [12] (S3: lm)	List of length 12
coefficients	double [2]	5014 -292
residuals	double [13]	-616 431 133 464 -403 1014 ...
1	double [1]	-616.1747
2	double [1]	431.2254
3	double [1]	132.9487
4	double [1]	464.1605
5	double [1]	-402.7264
6	double [1]	1014.081
7	double [1]	-325.1044
8	double [1]	-819.861
9	double [1]	-192.2953
10	double [1]	-330.5381
11	double [1]	203.3228
12	double [1]	671.0097
13	double [1]	-230.0484
effects	double [13]	-12284 -2956 178 542 -292 1158 ...
rank	integer [1]	2
fitted.values	double [13]	4722 4503 4284 4064 3845 3626 ...

	Light	Depth
1	4105.646	1.00
2	4933.925	1.75
3	4416.527	2.50
4	4528.618	3.25
5	3442.610	4.00
6	4640.297	4.75
7	3081.990	5.50
8	2368.113	6.25
9	2776.557	7.00
10	2419.193	7.75
11	2733.933	8.50
12	2982.499	9.25
13	1862.320	10.00

light = 5014 - 292*depth

3627 = 5014 - 292*4.75

3627+1014=4641



light = 5014 - 292*depth+ ε

Assumptions

- The usual ones: normality, homogeneity of variance, linearity and independence.
- **Outliers**: the observed value for the point is very different from that predicted by the regression model
- **Leverage points**: A leverage point is defined as an observation that has a value of x that is far away from the mean of x
- **Influential observations**: change the slope of the line. Thus, have a large influence on the fit of the model. One method to find influential points is to compare the fit of the model with and without each observation.
  - The Cook's distance statistic is a measure of the influence of each observation on the regression coefficients.
- Bottom line: **influential outliers** are problematic.

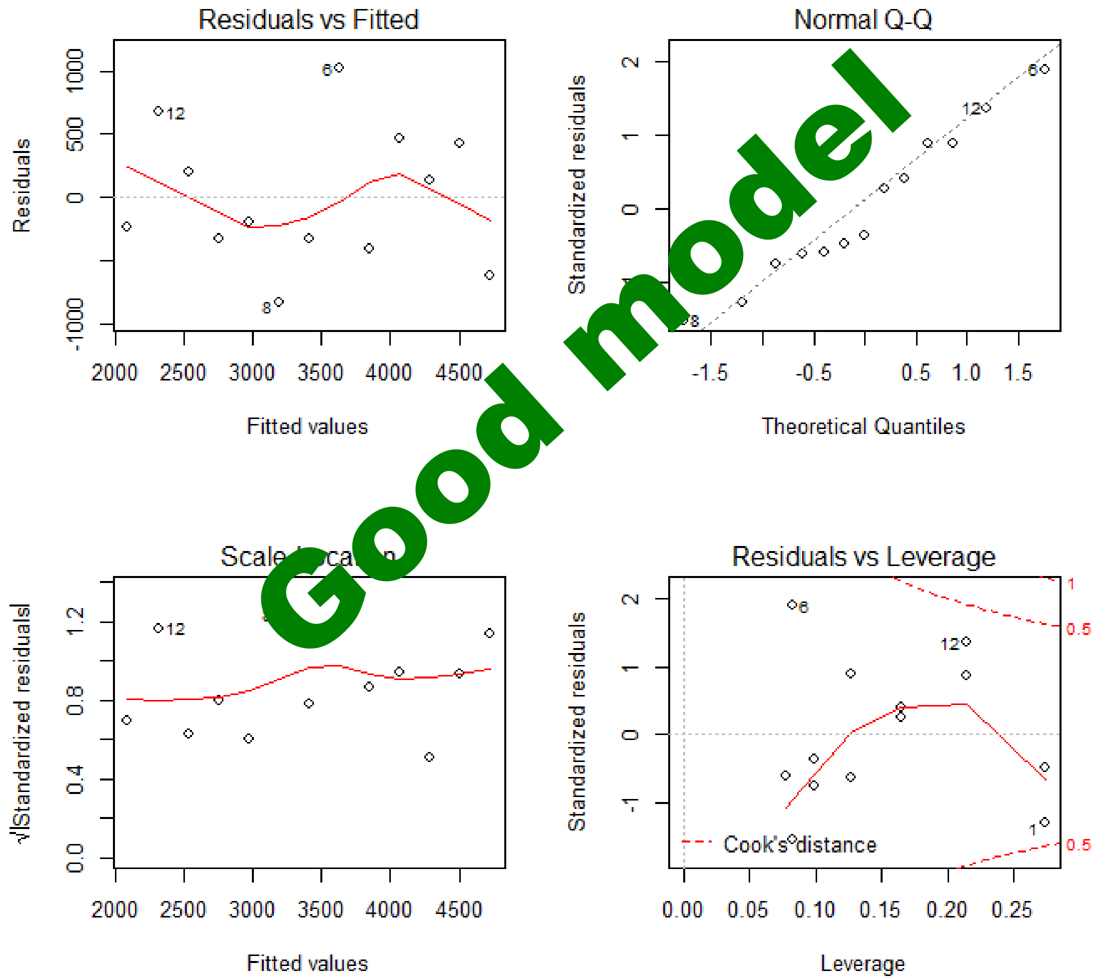
Assumptions

linear.conifer	list [12] (S3: lm)	List of length 12
coefficients	double [2]	5014 -292
residuals	double [13]	-616 431 133 464 -403 1014 ...

```
par(mfrow=c(2,2))
plot(linear.conifer)
```

Linearity

Normality



Homogeneity of variance

Outliers
Influential cases

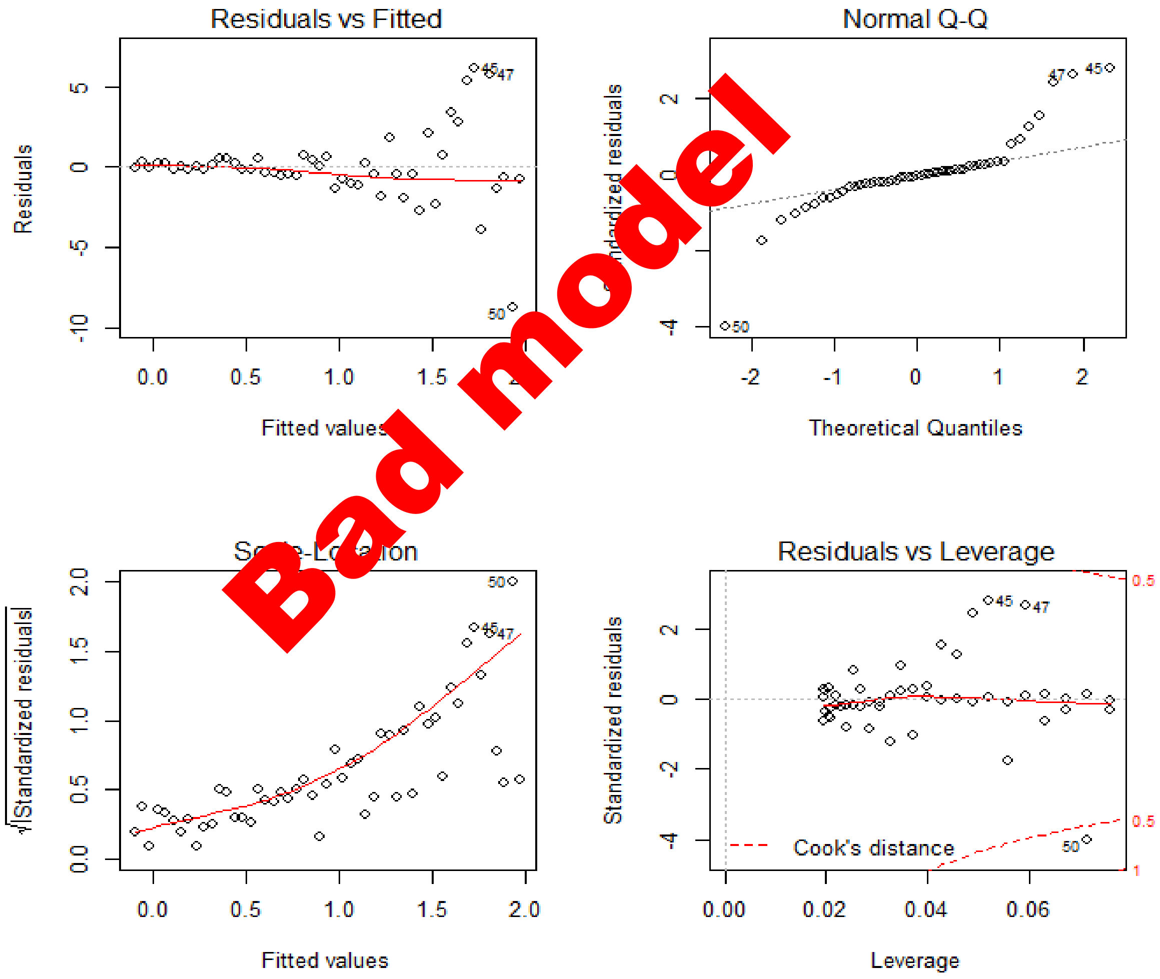
Assumptions

Linearity

Normality

Homogeneity of variance

Outliers
Influential cases



Linear regression

Your turn!

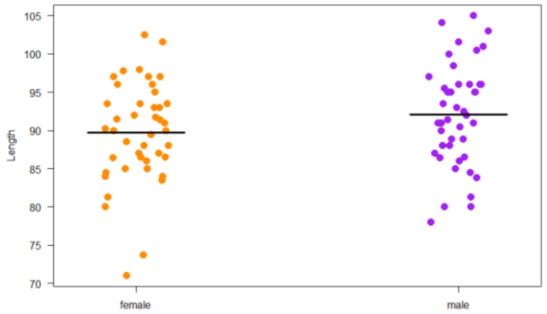
- Load `coniferlight.csv -> conifer`
- Plot the data `geom_point()`
- Build the model: `lm(Light~Depth, data=conifer)-> linear.conifer`
- Identify the coefficients of the model
- Add a line of best-fit

```
coefficients(linear.conifer) -> cf.abline
geom_abline(intercept=cf.abline[1], slope=cf.abline[2])
```

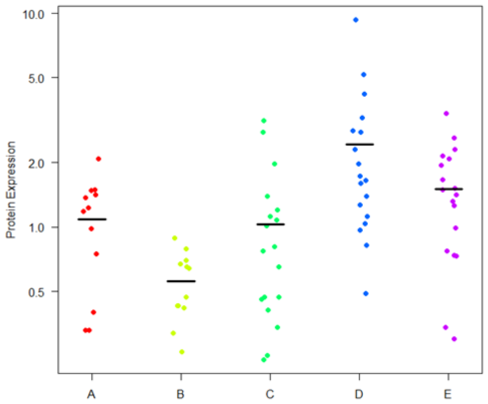
- Is the relationship between Depth and Light significant? `summary(linear.conifer)`
- How much of the variance is explained? $\mathbf{R^2}$
- What is the coefficient of correlation? `cor_test(conifer)`
- Compare the outputs of `summary(linear.conifer)` and `anova(linear.conifer)`
- Check out the assumptions

```
par(mfrow=c(2,2))
plot(linear.conifer)
```

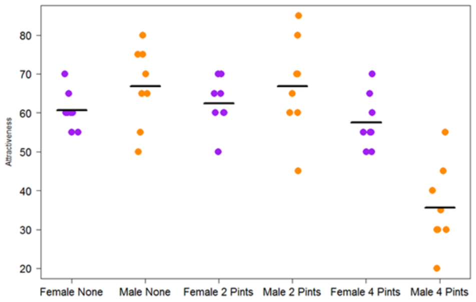
The linear model perspective



Coyotes= Body length~Gender

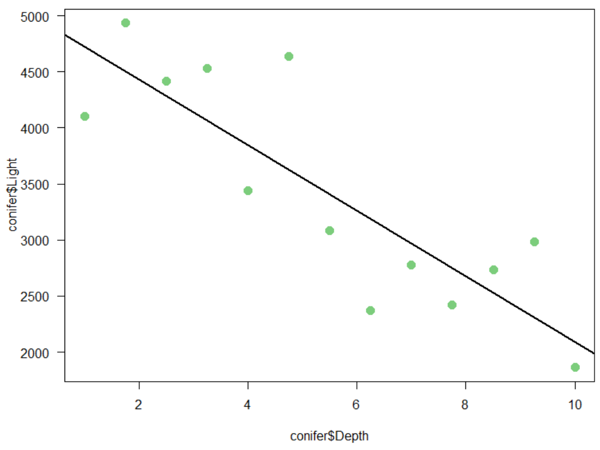


Protein = Expression~Cell line

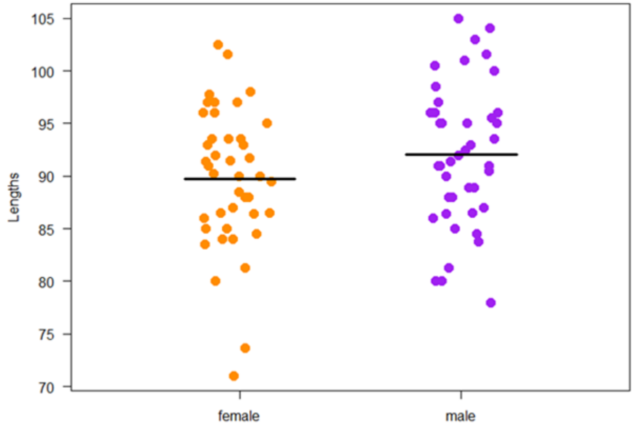


Goggles = Attractiveness~Alcohol and Gender

The linear model perspective



Continuous predictor



Categorical predictor

<u>Coyotes body length</u>

- Is there a difference between the 2 genders?

becomes

- **Does gender predict coyote body length?**

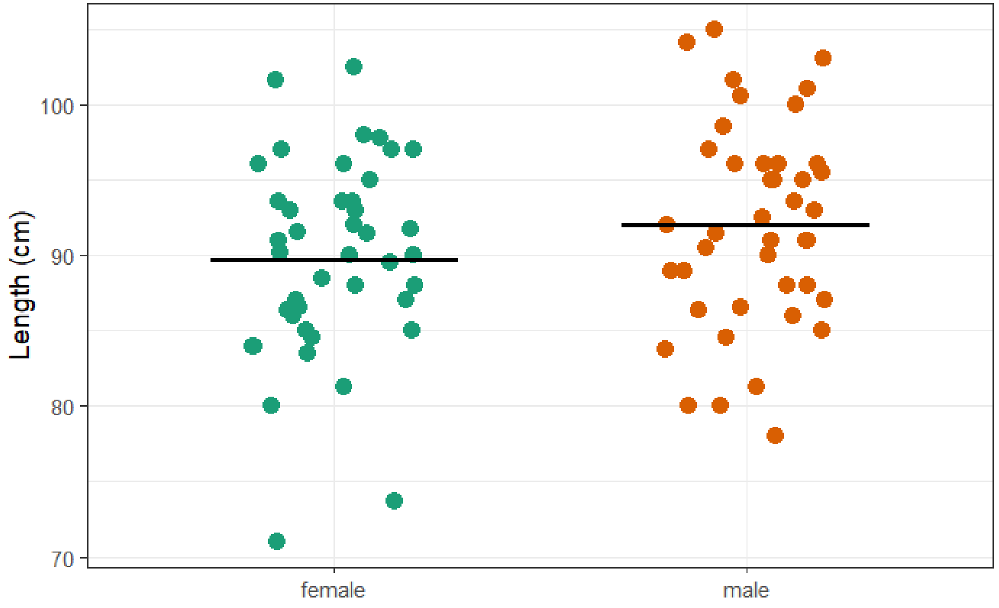
Example: coyotes

- Questions: *do male and female coyotes differ in size?*
  - does gender predict coyote body length?
  - how much of body length is predicted by gender?

The linear model perspective

Comparing 2 groups

```
read_csv("coyote.csv") -> coyote

coyote %>%
  ggplot(aes(gender,length, colour=gender)) +
    geom_jitter(height=0, size=4, width=0.2) +
    theme(legend.position = "none")+
    ylab("Length (cm)")+
    scale_colour_brewer(palette="Dark2")+
    xlab(NULL)+
    stat_summary(fun=mean, fun.min=mean, fun.max=mean, geom="errorbar",colour="black", size=1.2, width=0.6)
```



The linear model perspective

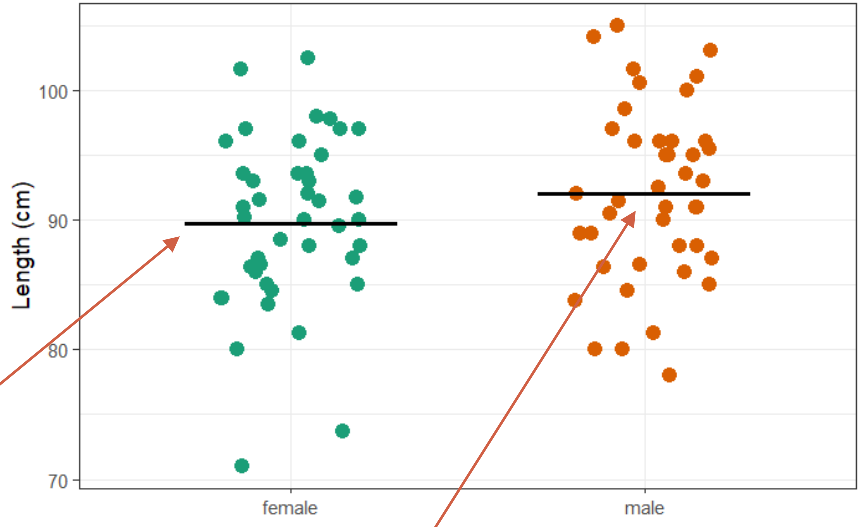
Comparing 2 groups

```
coyote %>%
  t_test(length~gender, var.equal=T)
```

.y. <chr>	group1 <chr>	group2 <chr>	n1 <int>	n2 <int>	statistic <dbl>	df <dbl>	p <dbl>
length	female	male	43	43	-1.641109	84	0.105

```
lm(length~gender, data=coyote)
```

```
Call:
lm(formula = coyote$length ~ coyote$gender)

Coefficients:
      (Intercept)  coyote$gendermale
           89.712              2.344
```



Females=89.71 cm, Males=89.71 + 2.34=92.05

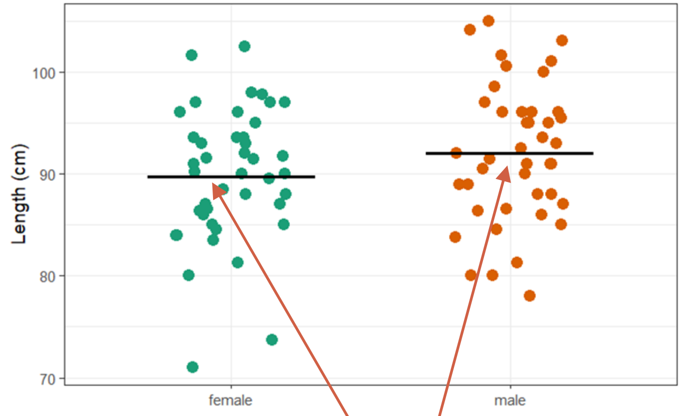
The linear model perspective

Comparing 2 groups

```
lm(length~gender, data=coyote)
```

```
Call:
lm(formula = coyote$length ~ coyote$gender)

Coefficients:
      (Intercept)  coyote$gendermale
           89.712              2.344
```

Body length = $\beta_0 + \beta_1$*Gender



Model

$$\text{Body Length} = \begin{pmatrix} 89.71 \\ 92.06 \end{pmatrix} \begin{pmatrix} \text{If Female} \\ \text{If Male} \end{pmatrix}$$

$$\text{Body Length} = 89.71 + \begin{pmatrix} 0 \\ 2.344 \end{pmatrix} \begin{pmatrix} \text{If Female} \\ \text{If Male} \end{pmatrix}$$

Body length = 89.712 + 2.344*Gender

The linear model perspective

Comparing 2 groups

$$y = \beta_0 + \beta_1 * x$$

continuous

conifer.csv

light = 5014 - 292*depth

categorical

coyote.csv

Body length = 89.712 + 2.344*Gender

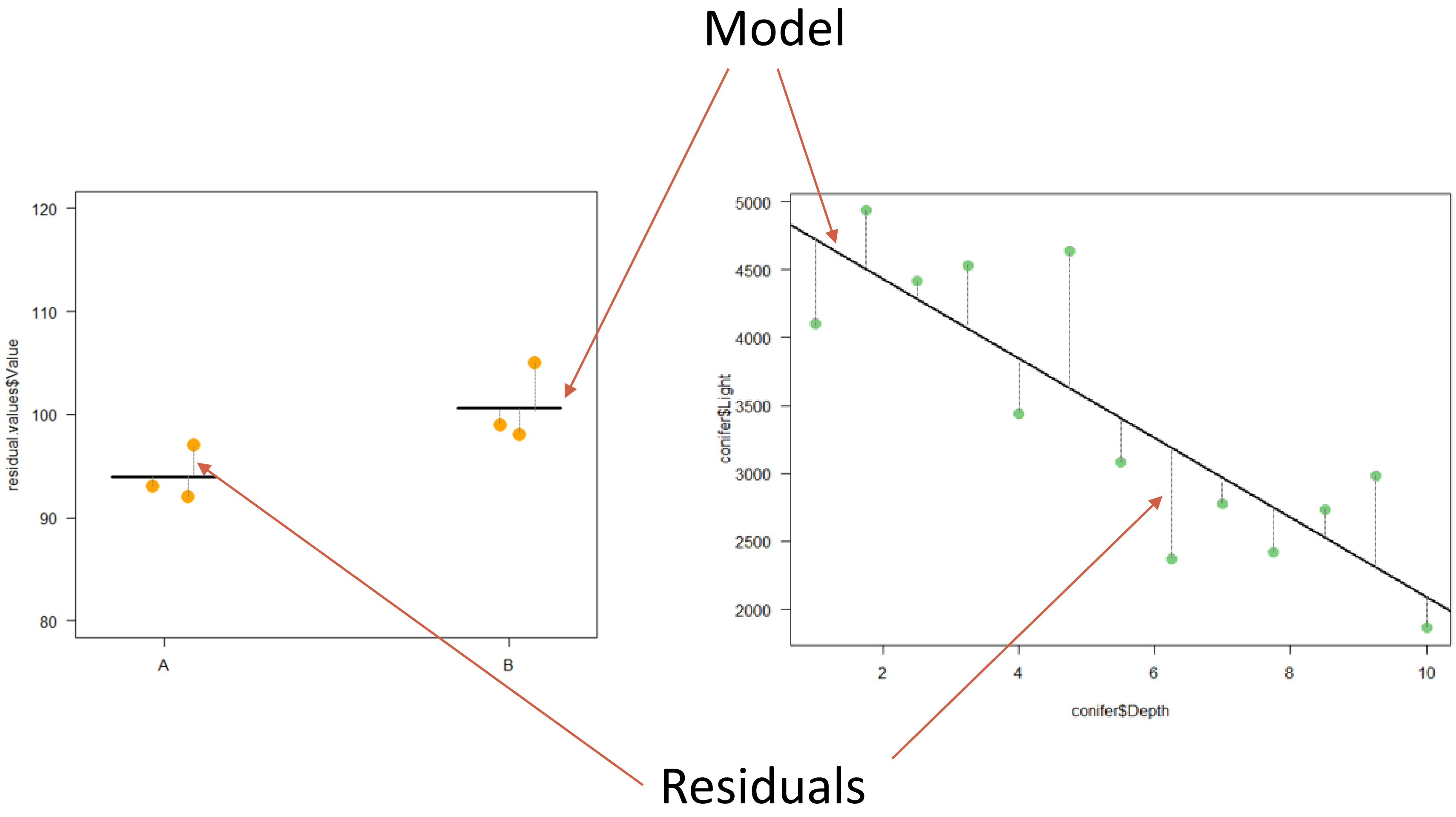
$$\text{Body Length} = 89.71 + \begin{pmatrix} 0 \\ 2.344 \end{pmatrix} \begin{pmatrix} \text{If Female} \\ \text{If Male} \end{pmatrix}$$

vector

$$y = \beta_0 + \beta_1 * x$$

The linear model perspective

Comparing 2 groups



The linear model perspective

Comparing 2 groups

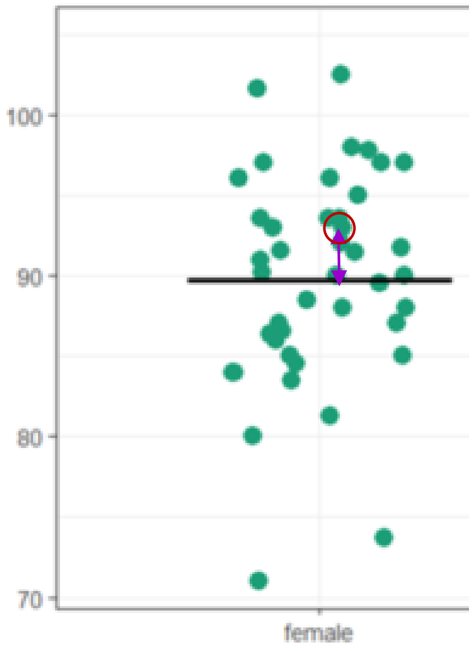
```
linear.coyote<-lm(length~gender, data=coyote)
```

```
linear.coyote
```

```
Coefficients:
(Intercept)  coyote$gendermale
     89.712              2.344
```

linear.coyote	list [13] (S3: lm)	List of length 13
coefficients	double [2]	89.71 2.34
(Intercept)	double [1]	89.71163
coyote$gendermale	double [1]	2.344186
residuals	double [86]	3.29 7.29 2.29 11.89 3.29 -5.21 ...

86 coyotes



$$\text{Body Length} = 89.71 + \begin{pmatrix} 0 \\ 2.344 \end{pmatrix} \begin{pmatrix} \text{If Female} \\ \text{If Male} \end{pmatrix}$$

Female 1: 89.71 + 3.29 = 93 cm

length	gender
93.0	female
97.0	female
92.0	female
101.6	female
93.0	female
84.5	female
102.5	female
97.8	female
91.0	female
98.0	female

The linear model perspective

Comparing 2 groups

```
coyote %>%
  t_test(length~gender, var.equal=T)
```

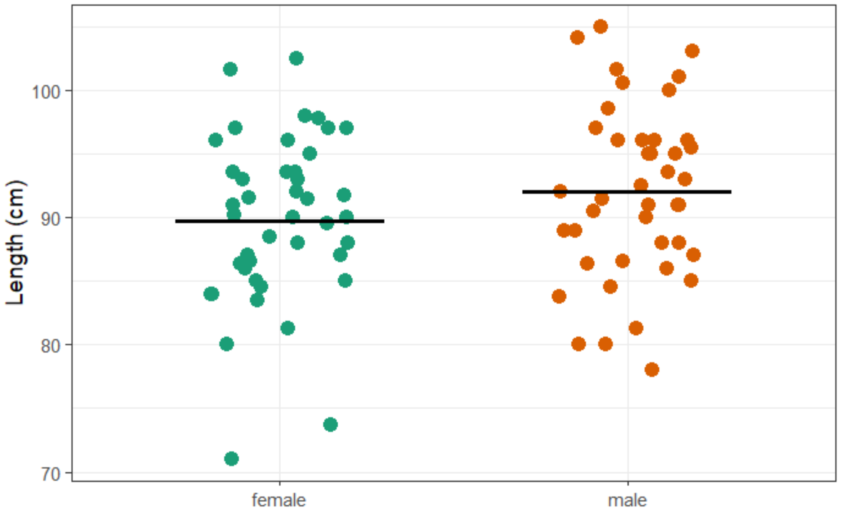
.y. <chr>	group1 <chr>	group2 <chr>	n1 <int>	n2 <int>	statistic <dbl>	df <dbl>	p <dbl>
length	female	male	43	43	-1.641109	84	0.105

```
summary(linear.coyote)
```

```
Call:
lm(formula = coyote$length ~ coyote$gender)

Residuals:
     Min       1Q   Median       3Q      Max
-18.7116  -4.0558   0.2884   3.9442  12.9442

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         89.712      1.010  88.820   <2e-16 ***
coyote$gendermale    2.344      1.428   1.641    0.105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.623 on 84 degrees of freedom
Multiple R-squared:  0.03107,	Adjusted R-squared:  0.01953
F-statistic: 2.693 on 1 and 84 DF,  p-value: 0.1045
```

```
anova(linear.coyote)
```

```
Analysis of Variance Table

Response: coyote$length
              Df Sum Sq Mean Sq F value Pr(>F)
coyote$gender  1  118.1 118.147  2.6932 0.1045
Residuals     84 3684.9  43.868
```



The linear model perspective

Comparing 2 groups

```
summary(linear.coyote)
```

```
Call:
lm(formula = coyote$length ~ coyote$gender)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.7116  -4.0558   0.2884   3.9442  12.9442 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)          89.712      1.010  88.820   <2e-16 ***
coyote$gendermale     2.344      1.428   1.641    0.105    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.623 on 84 degrees of freedom
Multiple R-squared:  0.03107,	Adjusted R-squared:  0.01953 
F-statistic: 2.693 on 1 and 84 DF,  p-value: 0.1045
```

```
anova(linear.coyote)
```

```
Analysis of Variance Table

Response: coyote$length
              Df Sum Sq Mean Sq F value Pr(>F)
coyote$gender  1  118.1 118.147  2.6932 0.1045
Residuals     84 3684.9  43.868
```

About 3% of the variability in body length is explained by gender.

118.1 + 3684.9 = 3803: total amount of variance in the data

Proportion explained by gender: 118.1/3803 = **0.031**

The linear model perspective

Comparing 2 groups

```
linear.coyote
```

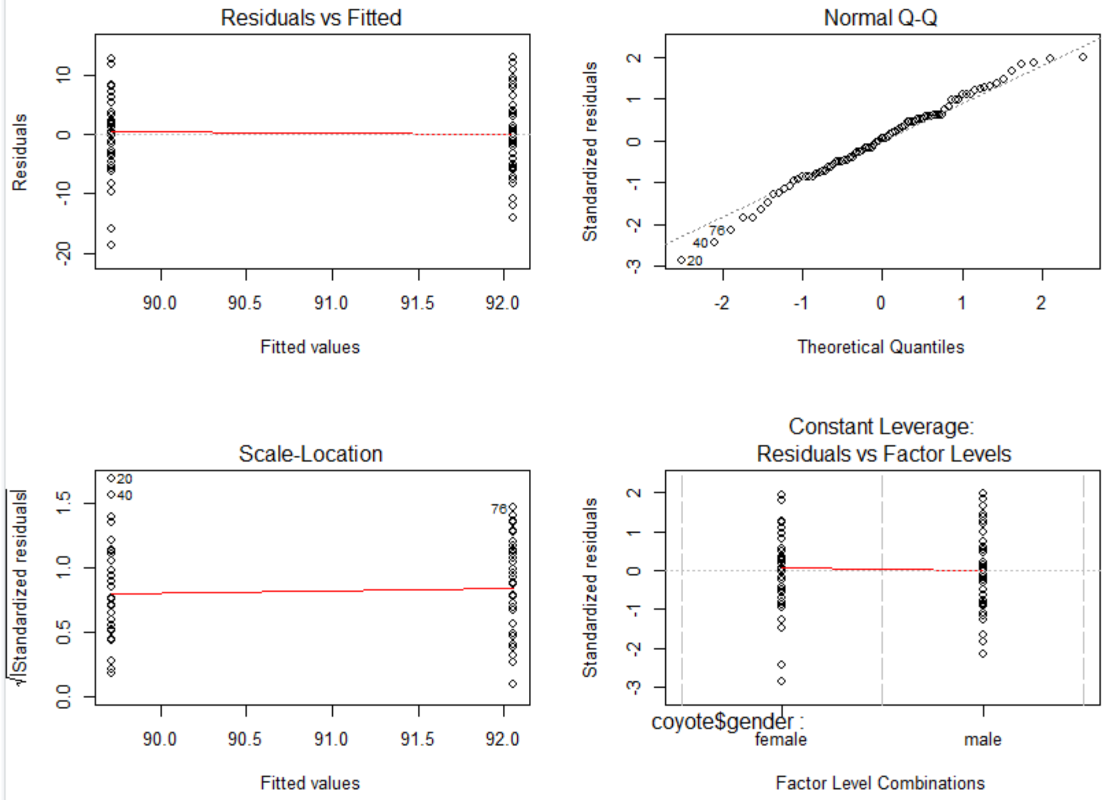
```
linear.coyote          List of 13
  coefficients : Named num [1:2] 89.71 2.34
  ..- attr(*, "names")= chr [1:2] "(Intercept)" "coyote$gendermale"
  residuals : Named num [1:86] 3.29 7.29 2.29 11.89 3.29 ...
  ..- attr(*, "names")= chr [1:86] "1" "2" "3" "4"
```

Assumptions

```
par(mfrow=c(2,2))
plot(linear.coyote)
```

Linearity

Normality
~ shapiro_test()



Equality of Variance
levene_test()

Outliers !

Example: coyote.csv

- Questions: *do male and female coyotes differ in size?*
  - does gender predict body length?
    - Answer: Quite unlikely: $p = 0.105$

  - how much of body length is predicted by gender?
    - Answer: About 3% (R^2=0.031)

Exercises 9 and 10: coyotes and protein expressions

- **coyote.csv** `coyote<-read_csv("coyote.csv")`
    - Run the t-test again `t_test()`
    - Run the same analysis using a linear model approach `lm()`
    - Compare the outputs and understand the coefficients from `lm()`
    - Use `summary()` and `anova()` to explore further the analysis
    - Work out R^2 from the `anova()` output
    - Don't forget to check the assumptions

- **protein.expression.csv** `protein<-read_csv("protein.expression.csv")`
    - Log-transformed the expression `log10()`
    - Run again the anova using `anova_test()`
    - Use `lm()` and `summary()` for the linear model approach
    - Compare the 2 outputs
    - Work out the means log10.expression for the 5 cell lines
    - Compare the outputs and understand the coefficients from `lm()`
    - Work out R^2 from the `anova()` output
    - Don't forget to check out the assumptions

Exercise 10 : protein.expression.csv

- **Questions**: *is there a difference in protein expression between the 5 cell lines?*
  - does cell line predict protein expression?
  - how much of the protein expression is predicted by the cell line?

Exercise 10 : protein.expression.csv - *Answers*

```
protein %>%
  anova_test(log10.expression~line)
```

```
ANOVA Table (type II tests)

  Effect DFn DFd     F        p p<.05   ges
1   line   4  73 8.123 1.78e-05     * 0.308
```

generalised **e**ffect **s**ize (Eta squared η^2) = R^2 ish

```
protein %>%
  tukey_hsd(log10.expression~line)
```

Tukey correction

	term <chr>	group1 <chr>	group2 <chr>	estimate <dbl>	conf.low <dbl>	conf.high <dbl>	p.adj <dbl>	p.adj.signif <chr>
1	line	A	B	-0.25024832	-0.578882494	0.07838585	2.19e-01	ns
2	line	A	C	-0.07499724	-0.374997820	0.22500335	9.56e-01	ns
3	line	A	D	0.30549397	0.005493391	0.60549456	4.39e-02	*
4	line	A	E	0.13327517	-0.166725416	0.43327575	7.27e-01	ns
5	line	B	C	0.17525108	-0.124749499	0.47525167	4.81e-01	ns
6	line	B	D	0.55574230	0.255741712	0.85574288	1.83e-05	****
7	line	B	E	0.38352349	0.083522904	0.68352407	5.48e-03	**
8	line	C	D	0.38049121	0.112162532	0.64881989	1.54e-03	**
9	line	C	E	0.20827240	-0.060056276	0.47660108	2.02e-01	ns
10	line	D	E	-0.17221881	-0.440547487	0.09610987	3.84e-01	ns

Exercise 10 : protein.expression.csv - *Answers*

```
linear.protein<-lm(log10.expression~line, data=protein)
```

```
anova(linear.protein)
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)    
line       4  2.691  0.6728   8.123 1.78e-05 ***
Residuals 73  6.046  0.0828                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
summary(linear.protein)
```

```
Call:
lm(formula = log10.expression ~ line, data = protein.stack.clean)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.62471 -0.21993  0.02264  0.18263  0.69537 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.03144    0.08308  -0.378  0.70617   
lineB       -0.25025    0.11749  -2.130  0.03655 * 
lineC       -0.07500    0.10725  -0.699  0.48661   
lineD        0.30549    0.10725   2.848  0.00571 **
lineE        0.13328    0.10725   1.243  0.21798   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2878 on 73 degrees of freedom
Multiple R-squared:  0.308,	Adjusted R-squared:  0.2701 
F-statistic: 8.123 on 4 and 73 DF,  p-value: 1.784e-05
```

```
lm(log10.expression~line,data=protein)
```

```
Call:
lm(formula = log10.expression ~ line, data = protein.stack.clean)

Coefficients:
(Intercept)        lineB        lineC        lineD        lineE  
   -0.03144     -0.25025     -0.07500      0.30549      0.13328  
```

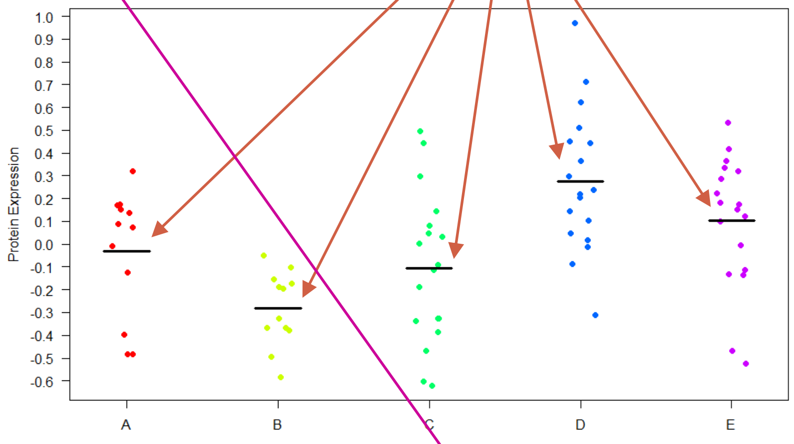
Exercise 10 : protein.expression.csv - *Answers*

```
protein %>%
  group_by(line) %>%
    summarise(mean=mean(log10.expression))
```

line <chr>	mean <dbl>
A	-0.03144412
B	-0.28169245
C	-0.10644136
D	0.27404985
E	0.10183104

$$\text{Expression} = \begin{pmatrix} -0.03144 \\ -0.28169 \\ -0.10644 \\ 0.27405 \\ 0.10183 \end{pmatrix} \begin{pmatrix} \text{Line A} \\ \text{Line B} \\ \text{Line C} \\ \text{Line D} \\ \text{Line E} \end{pmatrix}$$

Model




```
lm(log10.expression~line,data=protein)
```

```
Call:
lm(formula = log10.expression ~ line, data = protein.stack.clean)

Coefficients:
(Intercept)        lineB        lineC        lineD        lineE
   -0.03144     -0.25025     -0.07500      0.30549      0.13328
```

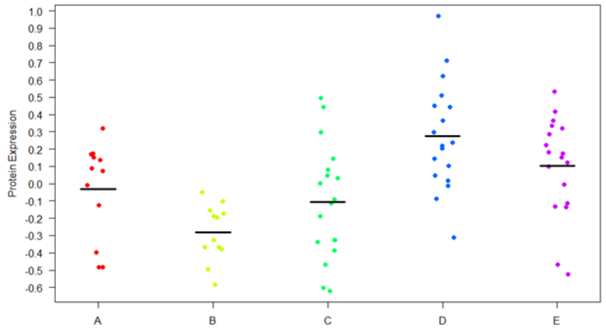
Expression= β_0 + β_1*Line

$$\text{Expression} = -0.03144 + \begin{pmatrix} 0 \\ -0.25025 \\ -0.07500 \\ 0.30549 \\ 0.13328 \end{pmatrix} \begin{pmatrix} \text{Line A} \\ \text{Line B} \\ \text{Line C} \\ \text{Line D} \\ \text{Line E} \end{pmatrix}$$

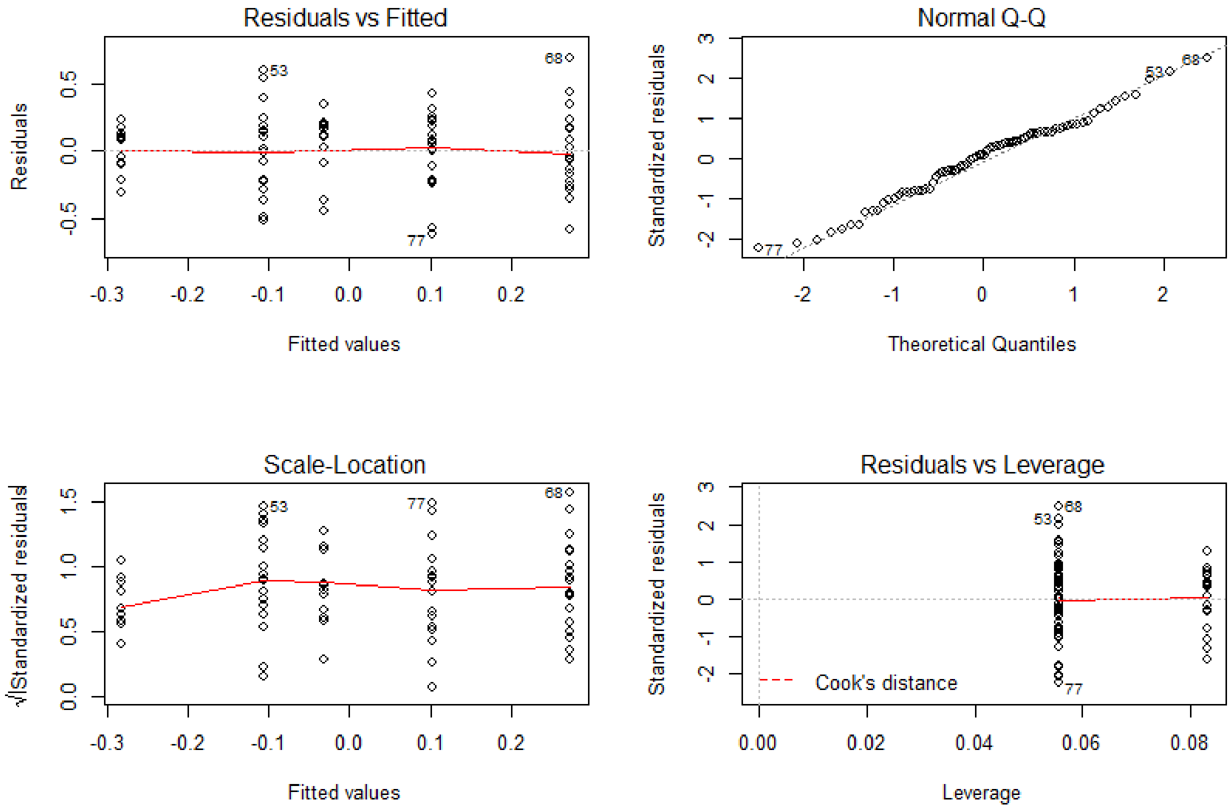
Example:
Line B = -0.03-0.25 = -0.28

Exercise 10: protein.expression.csv - *Answers*

```
par(mfrow=c(2,2))
plot(linear.protein)
```



Linearity



Normality

`shapiro_test()`

Equality of Variance

`levene_test()`

Outliers !

Exercise 10 : protein.expression.csv - *Answers*

```
linear.protein<-lm(log10.expression~line,data=protein)
summary(linear.protein)
```

```
Call:
lm(formula = log10.expression ~ line, data = protein.stack.clean)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.62471 -0.21993  0.02264  0.18263  0.69537 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.03144    0.08308  -0.378  0.70617   
lineB       -0.25025    0.11749  -2.130  0.03655 * 
lineC       -0.07500    0.10725  -0.699  0.48661   
lineD        0.30549    0.10725   2.848  0.00571 **
lineE        0.13328    0.10725   1.243  0.21798   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2878 on 73 degrees of freedom
Multiple R-squared:  0.308,	Adjusted R-squared:  0.2701 
F-statistic: 8.123 on 4 and 73 DF,  p-value: 1.784e-05
```

Proportion of variance explained by cell lines: 31%

```
protein %>%
  anova_test(log10.expression~line, detailed = TRUE)
```

```
  Effect   SSn   SSd DFn DFd     F        p p<.05   ges
1   line 2.691 6.046   4  73 8.123 1.78e-05     * 0.308
```

SSn

SSd

Source of variation	Sum of Squares	df	Mean Square	F	p-value
Between Groups	2.691	4	0.673	8.12	<0.0001
Within Groups	6.046	73	0.083		
Total	8.637				

2.691 + 6.046 = 8.737: total amount of variance in the data

Proportion explained by gender: 2.691/8.737 = **0.308**

Exercise 10 : protein.expression.csv

- **Questions**: *is there a difference in protein expression between the 5 cell lines?*
  - does cell line predict protein expression?
    - Answer: Yes p=1.78e-05
  - how much of the protein expression is predicted by the cell line?
    - Answer: About 31% (R^2=0.308)

Two-way Analysis of Variance

Example: goggles.csv

- The ‘beer-goggle’ effect

Alcohol	None		2 Pints		4 Pints	
Gender	Female	Male	Female	Male	Female	Male
	65	50	70	55	45	30
	70	55	65	65	60	30
	60	80	60	70	85	30
	60	65	70	55	65	55
	60	70	65	55	70	35
	55	75	60	60	70	20
	60	75	60	50	80	45
	55	65	50	50	60	40

- Study: effects of alcohol on mate selection in night-clubs.
- Pool of independent judges scored the levels of attractiveness of the person that the participant was chatting up at the end of the evening.
- **Question**: is subjective perception of physical attractiveness affected by alcohol consumption?
  - Attractiveness on a scale from 0 to 100

```
goggles<-read_csv("goggles.csv")
head(goggles)
```

```
gender alcohol attractiveness
Female    None             65
Female    None             70
Female    None             60
Female    None             60
Female    None             60
Female    None             55
```

The linear model perspective

Two factors

```
goggles %>%
  anova_test(attractiveness~alcohol+gender+alcohol*gender)
```

```
ANOVA Table (type II tests)

          Effect DFn DFd      F        p p<.05   ges
1        alcohol   2  42 20.065 7.65e-07     * 0.489
2         gender   1  42  2.032 1.61e-01       0.046
3 alcohol:gender   2  42 11.911 7.99e-05     * 0.362
```

```
goggles %>%
  group_by(alcohol, gender) %>%
    summarise(means=mean(attractiveness))
```

alcohol <chr>	gender <chr>	means <dbl>
0 Pints	Female	60.625
0 Pints	Male	66.875
2 Pints	Female	62.500
2 Pints	Male	66.875
4 Pints	Female	57.500
4 Pints	Male	35.625

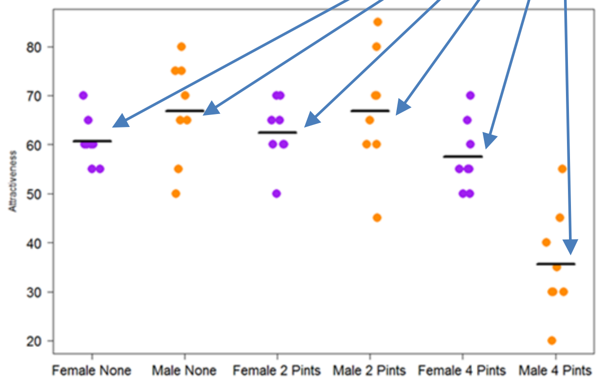
```
linear.goggles<-lm(attractiveness~alcohol+gender+alcohol*gender,data=goggles)
anova(linear.goggles)
```

```
Analysis of Variance Table

Response: attractiveness
               Df Sum Sq Mean Sq F value    Pr(>F)    
alcohol         2 3332.3 1666.15 20.0654 7.649e-07 ***
gender          1  168.7  168.75  2.0323    0.1614    
alcohol:gender  2 1978.1  989.06 11.9113 7.987e-05 ***
Residuals      42 3487.5   83.04                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(3332.3+168.7+1978.1)/(3332.3+168.7+1978.1+3487.5) = 0.611

R^2 = 61%

Model



$$y = \beta_0 + \beta_1 * x + \beta_2 * x_2 + \beta_3 * x_1 x_2$$

The linear model perspective

Two factors

```
linear.goggles<-lm(attractiveness~alcohol+gender+alcohol*gender,data=goggles)
summary(linear.goggles)
```

```
Call:
lm(formula = attractiveness ~ alcohol * gender, data = goggles)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.875  -5.625  -0.625   5.156  19.375 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 60.625      3.222  18.818  < 2e-16 ***
alcohol2 Pints               1.875      4.556   0.412    0.683    
alcohol4 Pints              -3.125      4.556  -0.686    0.497    
genderMale                   6.250      4.556   1.372    0.177    
alcohol2 Pints:genderMale   -1.875      6.443  -0.291    0.772    
alcohol4 Pints:genderMale  -28.125      6.443  -4.365 8.12e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.112 on 42 degrees of freedom
Multiple R-squared:  0.6111,	Adjusted R-squared:  0.5648 
F-statistic:  13.2 on 5 and 42 DF,  p-value: 9.609e-08
```

$$\text{Attractiveness} = \beta_0 + \beta_1\text{Alcohol} + \beta_2\text{Gender} + \beta_3\text{Gender*Alcohol}$$

$$\text{Attractiveness} = 60.625 + \begin{pmatrix} 0 \\ 1.875 \\ -3.125 \end{pmatrix} \begin{pmatrix} \text{if None} \\ \text{if 2 Pints} \\ \text{if 4 Pints} \end{pmatrix} + \begin{pmatrix} 0 \\ 6.250 \end{pmatrix} \begin{pmatrix} \text{if Female} \\ \text{if Male} \end{pmatrix} + \begin{pmatrix} 0 \\ -1.875 \\ -28.125 \end{pmatrix} \begin{pmatrix} \text{Otherwise} \\ \text{if Male and 2 Pints} \\ \text{if male and 4 Pints} \end{pmatrix}$$

Exercise 11: goggles.csv

- **goggles.csv** `goggles<-read_csv("goggles.csv")`

  - Run again the 2-way ANOVA `anova_test()`
  - Run the same analysis using a linear model approach `lm()`
  - Work out R^2 from the `anova()` output
  - Work out the equation of the model from the `summary()` output
    - *Hint: Attractiveness=* $\beta_0 + \beta_1 Gender + \beta_2 Alcohol + \beta_3 Gender*Alcohol$

  - Predict the attractiveness of a date:
    - for a female with no drinks
    - for a male with no drinks
    - for a male with 4 pints

Exercise 11: goggles.csv - *Answers*

$$\text{Attractiveness} = 60.625 + \begin{pmatrix} 0 \\ 1.875 \\ -3.125 \end{pmatrix} \begin{pmatrix} \text{if None} \\ \text{if 2 Pints} \\ \text{if 4 Pints} \end{pmatrix} + \begin{pmatrix} 0 \\ 6.250 \end{pmatrix} \begin{pmatrix} \text{if Female} \\ \text{if Male} \end{pmatrix} + \begin{pmatrix} 0 \\ -1.875 \\ -28.125 \end{pmatrix} \begin{pmatrix} \text{Otherwise} \\ \text{if Male and 2 Pints} \\ \text{if male and 4 Pints} \end{pmatrix}$$

- **goggles.csv**
  - Predict the attractiveness of a date:
    - for a female with no drinks

    60.625+0+0 = **60.625**
    - for a male with no drinks

    60.625+0+6.250= **66.875**
    - for a male with 4 pints

    60.625-3.125+6.250-28.125= **35.625**

```
goggles %>%
  group_by(alcohol, gender) %>%
    summarise(means=mean(attractiveness))
```

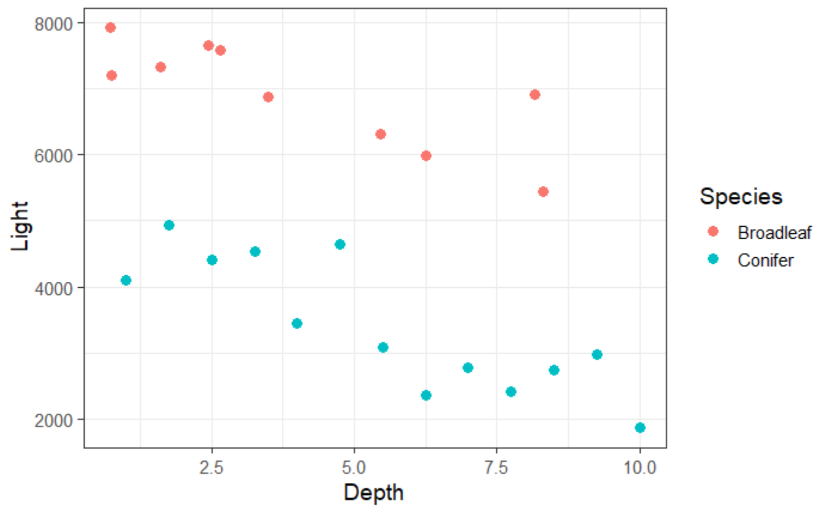
alcohol <chr>	gender <chr>	means <dbl>
0 Pints	Female	60.625
0 Pints	Male	66.875
2 Pints	Female	62.500
2 Pints	Male	66.875
4 Pints	Female	57.500
4 Pints	Male	35.625

The linear model perspective

Categorical and continuous factors

- Nothing special stats-wise with a mix of categorical and continuous factors
  - Same logic
  - But R makes it a little tricky to plot the model

```
treelight<-read_csv("treelight.csv")

treelight %>%
      ggplot(aes(x=Depth, y=Light, colour=Species))+
      geom_point(size=3)
```



The linear model perspective

Categorical and continuous factors

```
lm(Light~Depth+Species+Depth*Species, data=treelight)
```

```
Call:
lm(formula = Light ~ Depth * Species, data = treelight)

Coefficients:
          (Intercept)                  Depth         SpeciesConifer  Depth:SpeciesConifer
              7798.57                -221.13               -2784.58                -71.04
```

```
linear.treelight<-lm(Light~Depth*Species, data=treelight)
summary(linear.treelight)
```

```
Call:
lm(formula = Light ~ Depth * Species, data = treelight)

Residuals:
   Min     1Q Median     3Q    Max 
-819.9 -366.6 -161.3  377.1 1014.1 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           7798.57     298.62  26.115 2.38e-16 ***
Depth                 -221.13      61.80  -3.578  0.00201 ** 
SpeciesConifer       -2784.58     442.27  -6.296 4.82e-06 ***
Depth:SpeciesConifer   -71.04      81.31  -0.874  0.39321    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 534.6 on 19 degrees of freedom
Multiple R-squared:  0.9379,	Adjusted R-squared:  0.9281 
F-statistic: 95.71 on 3 and 19 DF,  p-value: 1.195e-11
```

Complete model

The linear model perspective

Categorical and continuous factors

- Additive model:

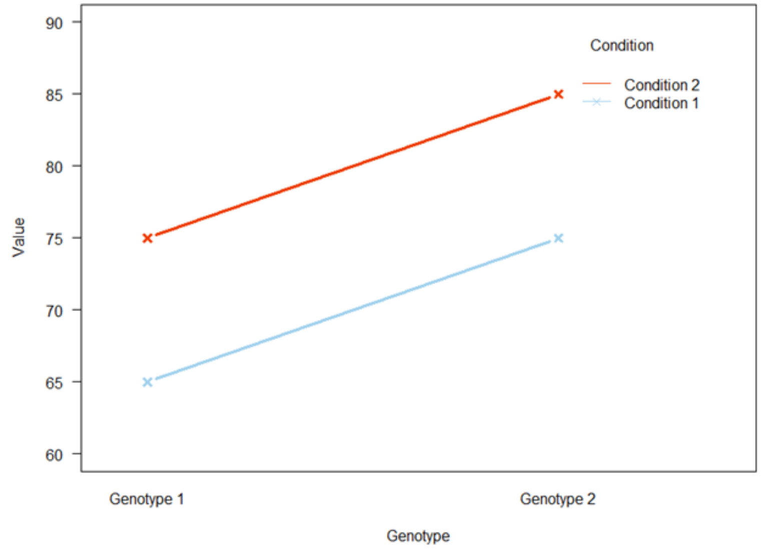
```
linear.treelight.add<-lm(Light~Depth+Species, data=treelight)
summary(linear.treelight.add)
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       7962.03     231.36  34.415  < 2e-16 ***
Depth             -262.17      39.92  -6.567 2.13e-06 ***
SpeciesConifer  -3113.03     231.59 -13.442 1.78e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 531.4 on 20 degrees of freedom
Multiple R-squared:  0.9354,   Adjusted R-squared:  0.929
F-statistic: 144.9 on 2 and 20 DF,  p-value: 1.257e-12
```

$$y = \beta_0 + \beta_1 * x$$

```
> lm(Light~Depth+Species, data=treelight)

Call:
lm(formula = Light ~ Depth + Species, data = treelight)

Coefficients:
   (Intercept)           Depth  SpeciesConifer
        7962.0          -262.2         -3113.0
```

No interaction



Both Effect

The linear model perspective

Categorical and continuous factors

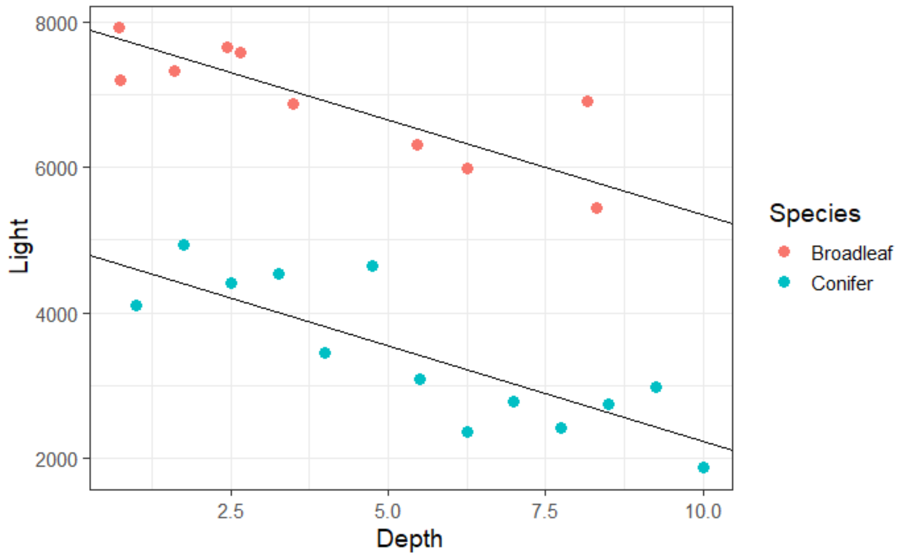
```
cf.add<-coefficients(linear.treelight.add)

 (Intercept)        Depth SpeciesConifer
   7962.0316    -262.1656     -3113.0265

cf.add[1]
```

It's a vector!

```
 (Intercept)
   7962.032

geom_abline(intercept=cf.add[1], slope=cf.add[2])+
geom_abline(intercept=(cf.add[1]+cf.add[3]), slope=cf.add[2])
```

Broadleaf:
Light = 7962.03 -262.17*Depth

Conifer:
Light = (7962.03-3113.03) -262.17*Depth

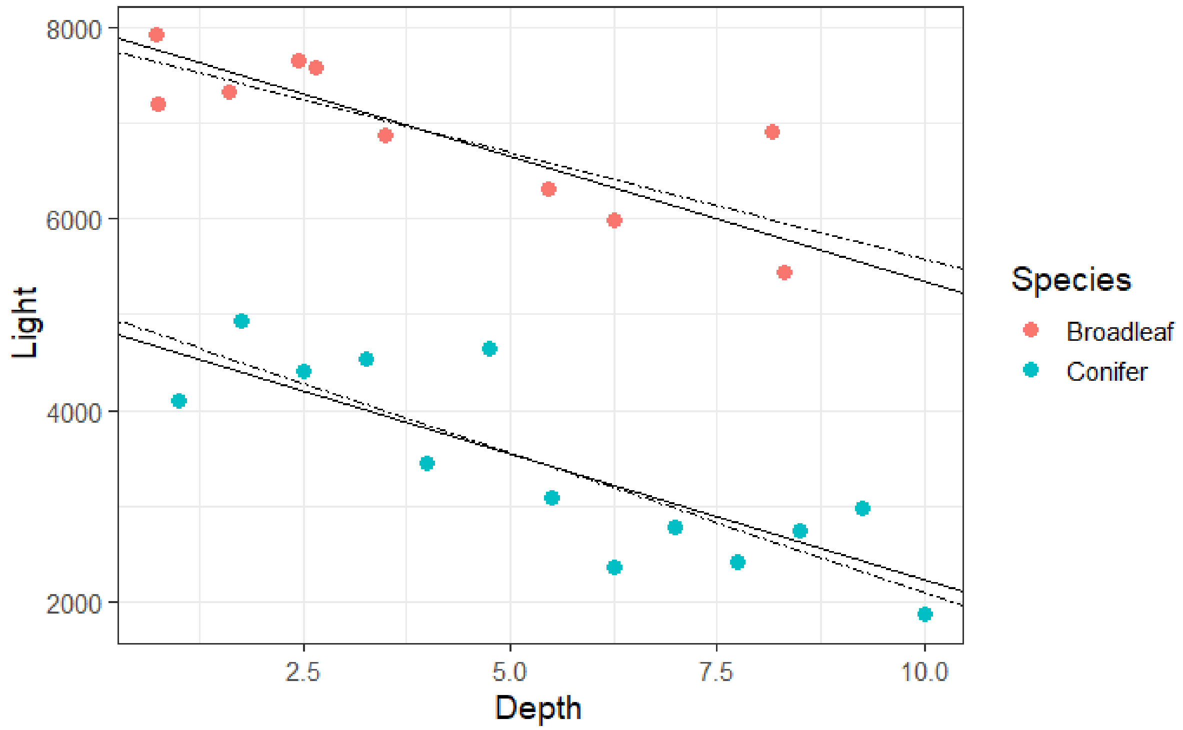


Exercise 12: treelight.csv

- **treelight.csv** `treelight<-read_csv("treelight.csv")`
  - Plot the data
  - Run a linear model `lm()`
  - Extract the parameters from the additive model
  - Plot a line of best fit for each species
  - Extract the parameters from the complete model
  - Write the new equations for broadleaf and conifer species.
  - Plot a line of best fit for each species (use dashed lines to distinguish between the 2 models).
  - Calculate the amount of light predicted:
    - In a conifer, 4 metres from the top of the canopy
    - In a broadleaf tree, 6 metres from the top of the canopy
  - How much of the variability of light is predicted by the depth and the species?

Exercise 12 : treelight.csv

```
cf<-coefficients(linear.treelight)

ggplot(treelight, aes(x=Depth, y=Light, group=Species, colour=Species))+
  geom_point(size=3)+
  geom_abline(intercept=cf.add[1], slope=cf.add[2])+
  geom_abline(intercept=(cf.add[1]+cf.add[3]), slope=cf.add[2])+
  geom_abline(intercept=(cf[1]), slope=cf[2], linetype="twodash")+
  geom_abline(intercept=(cf[1]+cf[3]), slope=(cf[2]+cf[4]), linetype="twodash")
```



Exercise 12 : treelight.csv

- Extract the parameters from the complete model

```
cf<-coefficients(linear.treelight)
```

```
    (Intercept)                  Depth        SpeciesConifer  Depth:SpeciesConifer
     7798.56552             -221.12564           -2784.58333             -71.03575
```

```
ggplot(treelight, aes(x=Depth, y=Light, group=Species, colour=Species))+
  geom_point(size=3)+
  geom_abline(intercept=cf.add[1], slope=cf.add[2])+
  geom_abline(intercept=(cf.add[1]+cf.add[3]), slope=cf.add[2])+
  geom_abline(intercept=(cf[1]), slope=cf[2], linetype="twodash")+
  geom_abline(intercept=(cf[1]+cf[3]), slope=(cf[2]+cf[4]), linetype="twodash")
```

Broadleaf: Light = 7798.57 – 221.13*Depth

Conifer: Light = (7798.57-2784.58) – (221.13-71.04)*Depth

- Calculate the amount of light predicted:
    - In a conifer, 4 metres from the top of the canopy

    *(7798.57-2784.58)-(221.13+71.04)*4 = 4413.63*

    - In a broadleaf species, 6 metres from the top of the canopy

    *7798.57-221.13*6 = 6471.79*

Linear model

Simplest

$$y = \beta_0 + \beta_1 * x$$

With 2 factors

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_1 x_2$$

With n factors

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_n * x_n$$

Simplest

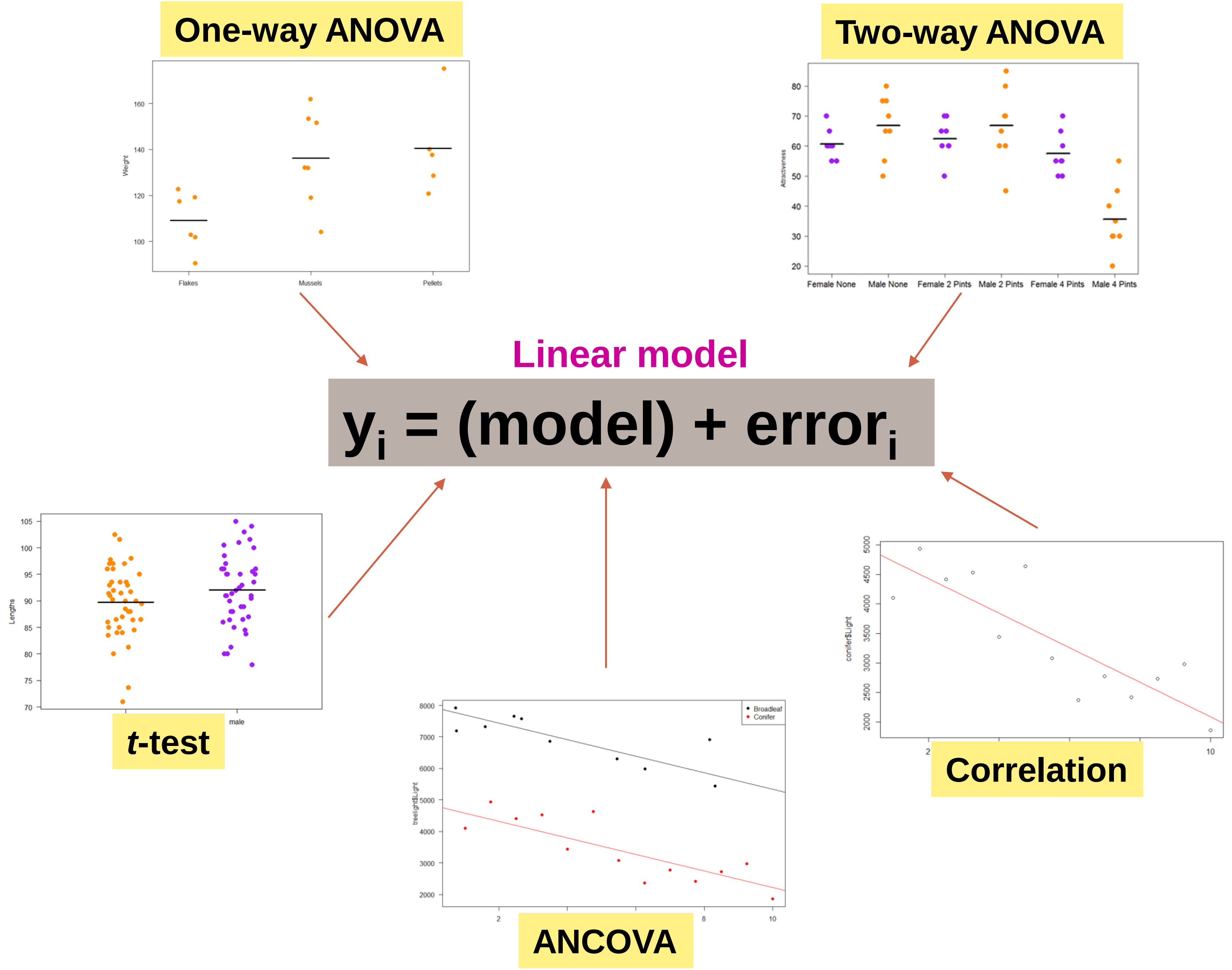
$$y = \beta_0 + \beta_1 * x$$

Let’s not forget the error

$$y_i = (\beta_0 + \beta_1 * x_i) + \varepsilon_i$$

General formula

$$\mathbf{y_i = (model) + error_i}$$


One-way ANOVA
Flakes
Mussels
Pellets
Weight
Two-way ANOVA
Female None
Male None
Female 2 Pints
Male 2 Pints
Female 4 Pints
Male 4 Pints
Linear model
y_i = (model) + $error_i$
Lengths
male
t-test
Broadleaf
Conifer
ANCOVA
Correlation