Introduction to Linear Modelling

Anne Segonds-Pichon

v2019-03
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**(\( x \)) = \( y \)

\[
y = \beta_0 + \beta_1 \cdot x
\]

- **In R:**
  - Correlation: `cor()`
  - Linear regression: `lm()`
Linear regression

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

```r
conifer <- read.csv("coniferlight.csv")
```

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

  \[
  \text{light} = \beta_0 + \beta_1 \times \text{depth}
  \]

  ```r
  plot(conifer$Light ~ conifer$Depth)
  ```
Linear regression

- Linear modelling in R: `lm(y~x)`
- Regression: `lm(conifer$Light~conifer$Depth)`
- or: `lm(Light~Depth, data=conifer)`

\[
\text{light} = \beta_0 + \beta_1 \times \text{depth}
\]

\[
\text{light} = 5014 - 292 \times \text{depth}
\]
Linear regression

```
linear.conifer<-lm(conifer$Light~conifer$Depth)
```

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

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Class</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear.conifer</td>
<td>list [12] (S3: lm)</td>
<td>12</td>
</tr>
<tr>
<td>coefficients</td>
<td>double [2]</td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>double [1]</td>
<td></td>
</tr>
<tr>
<td>conifer$Depth</td>
<td>double [1]</td>
<td></td>
</tr>
<tr>
<td>residuals</td>
<td>double [13]</td>
<td></td>
</tr>
<tr>
<td>effects</td>
<td>double [13]</td>
<td></td>
</tr>
<tr>
<td>rank</td>
<td>integer [1]</td>
<td></td>
</tr>
<tr>
<td>fitted.values</td>
<td>double [13]</td>
<td></td>
</tr>
<tr>
<td>assign</td>
<td>integer [2]</td>
<td></td>
</tr>
<tr>
<td>qr</td>
<td>list [5] (S3: qr)</td>
<td></td>
</tr>
<tr>
<td>df.residual</td>
<td>integer [1]</td>
<td></td>
</tr>
<tr>
<td>xlevels</td>
<td>list [0]</td>
<td></td>
</tr>
<tr>
<td>call</td>
<td>language</td>
<td></td>
</tr>
<tr>
<td>terms</td>
<td>formula</td>
<td></td>
</tr>
<tr>
<td>model</td>
<td>list [13 x 2] (S3: data.frame)</td>
<td></td>
</tr>
</tbody>
</table>
Linear regression `abline()`

`abline(a, b …) a=intercept, b=slope`

`abline(linear.conifer)`

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

light = 5014 - 292*depth
Linear regression summary()

\[ \text{lm}(\text{conifer}$Light \sim \text{conifer}$Depth) \]

```r
linear.conifer <- lm(conifer$Light ~ conifer$Depth)
summary(linear.conifer)
```

**Call:**
\[ \text{lm}(\text{formula} = \text{conifer}$Light \sim \text{conifer}$Depth) \]

**Coefficients:**
- (Intercept) \[ 5014.0 \]
- conifer$Depth \[ -292.2 \]

**Residuals:**
- Min \[ -819.9 \]
- 1Q \[ -330.5 \]
- Median \[ -192.3 \]
- 3Q \[ 431.2 \]
- Max \[ 1014.1 \]

**Coefficients:**
- Estimate \[ 5014.0 \]
- Std. Error \[ 342.15 \]
- t value \[ 14.654 \]
- \( p \)-value \[ 0.0000000000000000 \]

**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

**Coefficient of determination**

**p-value**
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: 0.000263
```

• $r$: coefficient of correlation between $x$ (depth) and $y$ (light)
  – e.g. here: $r = -0.846$ so $r^2 = -0.846 \times -0.846 = 0.716 = R$-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
> cor(conifer)
          Light Depth
Light 1.000000 -0.8464529
Depth -0.8464529 1.000000
```
Linear regression \texttt{anova()} \texttt{summary(linear.conifer)}

\begin{verbatim}
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    14.65   4.3e-08 ***
conifer$Depth     -292.16      55.41    -5.27    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
\end{verbatim}

\texttt{anova(linear.conifer)}

\begin{verbatim}
Analysis of Variance Table

Response: conifer$Light
           Df Sum Sq Mean Sq F value    Pr(>F)
conifer$Depth  1 8738553 8738553 27.798 0.0002633 ***
Residuals     11 3457910   31435
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
\end{verbatim}

Total amount of variability: \(8738553 + 3457910 = 12196463\)

Proportion explained by depth: \(\frac{8738553}{12196463} = 0.716\)
Linear regression the error $\varepsilon$

- 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.

$$3627 = 5014 - 292 \times 4.75$$

$y = \beta_0 + \beta_1 \times x + \varepsilon$
linear.conifer <- lm(conifer$Light ~ conifer$Depth)

light = 5014 - 292*depth

3627 = 5014 - 292*4.75
3627 + 1014 = 4641

light = 5014 - 292*depth + \epsilon
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

- Linearity
- Homogeneity of variance
- Normality
- Outliers
- Influential cases

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

- Linearity
- Homogeneity of variance
- Normality
- Outliers
- Influential cases

Bad model
The linear model perspective

Coyotes = Body length ~ Gender

Protein = Expression ~ Cell line

Goggles = Attractiveness ~ Alcohol and Gender
The linear model perspective

Coyotes body length

- 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

coyote<-read.csv("coyote.csv")  # remember to use tab!
head(coyote)

stripchart(coyote$length~coyote$gender,
        vertical=TRUE,
        method="jitter",
        las=1,
        ylab="Lengths",
        pch=16,
        col=c("darkorange","purple"),
        cex=1.5)

length.means <- tapply(coyote$length, coyote$gender, mean)

segments(1:2-0.15,
        length.means,
        1:2+0.15,
        length.means,
        lwd=3)
The linear model perspective
Comparing 2 groups

t.test(coyote$length~coyote$gender, var.equal=T)

The linear model perspective
Comparing 2 groups

lm(coyote$length~coyote$gender)

Females=89.71 cm, Males=89.71 + 2.34=92.05
The linear model perspective

Comparing 2 groups

```r
lm(coyote$length ~ coyote$gender)
```

Body length = \( \beta_0 + \beta_1 \cdot \text{Gender} \)

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

Body length = 89.712 + 2.344*Gender
The linear model perspective
Comparing 2 groups

\[ y = \beta_0 + \beta_1 \times x \]

**conifer.csv**

\[ \text{light} = 5014 - 292 \times \text{depth} \]

**coyote.csv**

Body length = 89.712 + 2.344*Gender

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

\[ y = \beta_0 + \beta_1 \times x \]
The linear model perspective
Comparing 2 groups
The linear model perspective
Comparing 2 groups

```r
linear.coyote <- lm(coyote$length ~ coyote$gender)
linear.coyote
```

Coefficients:
(Intercept) 89.712 coyote$gendermale 2.344

86 coyotes

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
The linear model perspective
Comparing 2 groups

`t.test(coyote$length~coyote$gender)`

```
Two Sample t-test

data: coyote$length by coyote$gender
t = -1.6411, df = 84, p-value = 0.1045
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-5.814747 0.496375
sample estimates: mean in group female mean in group male
89.71163 92.005881
```

`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 2.6931 0.1045
Residuals 84 3664.9 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 -4.0155  0.2884  3.9442 12.9442

Coefficients:                 Estimate Std. Error   t value Pr(>|t|)
(Intercept)                 89.7121     1.0105  88.8200  < 2e-16 ***
coyote$gendermale          2.3444      1.1428   2.0411    0.0477    
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1

Residual standard error: 5.363 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
```

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

**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 2.69320   0.1045
Residuals      84 3684.9  43.868

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**

Assumptions

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

**Linearity**

![Linearity Plot](image)

**Normality**

~ `shapiro.test()`

**Equality of Variance**

`leveneTest()`

**Outliers**

![Outliers Plot](image)
Example: *coyote.csv*

**One last thing**

- `t.test(coyote$length~coyote$gender, var.equal = TRUE)`

```
> head(coyote)
length gender
1  93.0 female
2  97.0 female
3  92.0 female
4 101.6 female
5  93.0 female
6  84.5 female
```

**Long format**

- `t.test(coyote.wide$female, coyote.wide$male, var.equal = TRUE)`

```
> head(coyote.wide)
female  male
1   93.0  97.0
2   97.0  95.0
3   92.0  96.0
4 101.6  91.0
5   93.0  95.0
6   84.5  84.5
```

**Wide format**
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% (R²=0.031)
Exercises 10 and 11: 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")`
  - Prepare the file using `melt()`*, `colnames()` and `na.omit()`
  - Log-transformed the expression `log10()`
  - Run again the anova using `aov()` and `summary()`
  - Compare the 2 outputs
  - Use `lm()` and `summary()` for the linear model approach
  - 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

* # reshape2 package #
Exercise 11: 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 11: protein.expression.csv - Answers

```r
anova.log.protein <- aov(log10.expression ~ line, data = protein.stack.clean)
summary(anova.log.protein)
```

```
                DF  Sum Sq  Mean Sq F value  Pr(>F)
line            4 2.69100  0.67275  8.123  0.0002 **
Residuals        73 6.04625  0.08281
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```r
pairwise.t.test(protein.stack.clean$log10.expression, protein.stack.clean$line, p.adj = "bonf")
```

```
  data: protein.stack.clean$log10.expression and protein.stack.clean$line
  A     B     C     D     E
A -0.3665  1.0000 -3.67 -3.67 -3.67
B  1.0000 -0.3665  1.90  1.90  1.90
C  3.6700  1.9000 -3.67 -3.67 -3.67
D  3.6700  1.9000 -3.67 -3.67 -3.67
E  3.6700  1.9000 -3.67 -3.67 -3.67

TukeyHSD(anova.log.protein,"line")
```

```r
  diff  lwr  upr p adj
B-A -0.25024832 -0.57882494 0.07838558 0.2187264
B-C -1.07499726 -2.37497820 0.22003329 0.3420847
B-D -1.32599726 -2.62497820 0.07308461 0.2187264
B-E -1.32599726 -2.62497820 0.07308461 0.2187264
C-A 0.07499726 0.37497820 1.22500329 0.3420847
C-B -1.07499726 -2.37497820 0.22003329 0.3420847
C-D -0.32599726 -1.62497820 0.97308461 0.2187264
C-E -0.32599726 -1.62497820 0.97308461 0.2187264
D-A 1.32599726 0.60502820 2.04497820 0.0343876
D-B 0.07499726 0.37497820 1.22500329 0.3420847
D-C -1.07499726 -2.37497820 0.22003329 0.3420847
D-E -0.32599726 -1.62497820 0.97308461 0.2187264
E-A -1.32599726 -2.62497820 0.07308461 0.2187264
E-B -0.32599726 -1.62497820 0.97308461 0.2187264
E-C 0.07499726 0.37497820 1.22500329 0.3420847
E-D 1.32599726 0.60502820 2.04497820 0.0343876
```
Exercise 11: protein.expression.csv - Answers

protein<-read.csv("protein expression.csv",header=T)
## more elegant way ##  ## reshape2 package ##
protein.stack.clean<-melt(protein, variable.name = "line", value.name = "expression", na.rm=TRUE)
head(protein.stack.clean)

##

stripchart(protein.stack.clean$expression~protein.stack.clean$line,
  vertical=TRUE,
  method="jitter",
  las=1,
  ylab="Protein Expression",
  pch=16,
  col=rainbow(5),
  log="y"
)

expression.means<-tapply(protein.stack.clean$expression,protein.stack.clean$line,mean)
segments(1:5-0.15,expression.means,1:5+0.15, expression.means, lwd=3)

protein.stack.clean$log10.expression <- log10(protein.stack.clean$expression)
Exercise 11: protein.expression.csv - Answers

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

```
          Df Sum Sq Mean Sq  F value   Pr(>F)
line        3 1.0436  0.3479  8.123 0.0003 **
Residuals   73 9.6232  0.1319
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1

summary(linear.protein)
```

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

residuals:
     Min      1Q  Median      3Q     Max
-0.62471  0.02284  0.18263  0.69537

coefficients:                     Estimate  Std. Error t value Pr(>|t|)
(Intercept)               -0.03144    0.08038  -0.3905   0.70617
lineB                     -0.25025    0.11749  -2.1300   0.03655 *
lineC                     -0.07500    0.10725  -0.6994   0.48661
lineD                     0.30549    0.10725   2.8448   0.00571 **
lineE                     0.13328    0.10725   1.2435   0.21798
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 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.stack.clean)
```

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

coefficients:                     Estimate  Std. Error t value Pr(>|t|)
(Intercept)               -0.03144    0.08038  -0.3905   0.70617
lineB                     -0.25025    0.11749  -2.1300   0.03655 *
lineC                     -0.07500    0.10725  -0.6994   0.48661
lineD                     0.30549    0.10725   2.8448   0.00571 **
lineE                     0.13328    0.10725   1.2435   0.21798
```
Exercise 11: protein.expression.csv - Answers

\[ \text{Expression} = \beta_0 + \beta_1 \cdot \text{Line} \]

Example:
Line B = -0.03 - 0.25 = -0.28
Exercise 11: protein.expression.csv - Answers

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

Linearity

Normal Q-Q

Equality of Variance

Outliers

shapiro.test()

leveneTest()
Exercise 11: protein.expression.csv - Answers

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

anova.log.protein<-aov(log10.expression~line, data=protein.stack.clean)
summary(anova.log.protein)

Proportion of variance explained by cell lines: 31%

2.691 + 6.046 = 8.737: total amount of variance in the data
Proportion explained by gender: 2.691/8.737 = 0.308
Exercise 11: protein.expression.csv

- **Questions**: *is there a difference in protein expression between the 5 cell lines?*
  - does cell line predict protein expression?
    - **Answer**: Yes $p=1.78e^{-5}$

- 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

- 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

```r
goggles <- read.csv("goggles.csv")
head(goggles)
```
The linear model perspective

Two factors

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

```
Df  Sum Sq Mean Sq  F value  Pr(>F)
alcohol       2    3332.3  1666.15   20.065 7.64e-07 ***
gender         1     168.7   168.75    2.032   0.161
alcohol:gender 2    1978.1   989.06  11.911 7.99e-05 ***
Residuals      42    3488.5    83.04
---
Signif. codes:  < 0.001 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1
```

```
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   20.065 7.64e-07 ***
gender         1     168.7   168.75    2.032   0.161
alcohol:gender 2    1978.1   989.06  11.911 7.99e-05 ***
Residuals      42    3487.5    83.04
---
Signif. codes:  < 0.001 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1
```

\[
y = \beta_0 + \beta_1 * x + \beta_2 * x^2 + \beta_3 * x_1 x_2
\]

\[
\frac{(3332.3 + 168.7 + 1978.1)}{(3332.3 + 168.7 + 1978.1 + 3487.5)} = 0.611
\]

\(R^2 = 61\%\)

Model

> tapply(goggles$attractiveness, goggles$subgroup, mean) 

<table>
<thead>
<tr>
<th>Female None</th>
<th>Male None</th>
<th>Female 2 Pints</th>
<th>Male 2 Pints</th>
<th>Female 4 Pints</th>
<th>Male 4 Pints</th>
</tr>
</thead>
<tbody>
<tr>
<td>60.625</td>
<td>66.875</td>
<td>62.500</td>
<td>66.875</td>
<td>57.500</td>
<td>35.625</td>
</tr>
</tbody>
</table>
The linear model perspective

Two factors

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

\[
\text{Attractiveness} = 60.625 \begin{pmatrix} 0 \\ 1.875 \\ -3.125 \end{pmatrix} \begin{pmatrix} \text{if None} \\ \text{if 2 Pints} \\ \text{if 4 Pints} \end{pmatrix} + 6.250 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} \text{if Female} \\ \text{if Male} \end{pmatrix} + \begin{pmatrix} 0 \\ -1.875 \\ -28.125 \end{pmatrix} \begin{pmatrix} \text{if Male and 2 Pints} \\ \text{if male and 4 Pints} \end{pmatrix}
\]
Exercise 12: goggles.csv

- goggles.csv goggles<-read.csv("goggles.csv")

- Run again the 2-way ANOVA aov()
- Run the same analysis using a linear model approach lm()
- Work out R² from the anova() output
- Work out the equation of the model from the summary() output
  - Hint: Attractiveness= β₀ + β₁Gender + β₂Alcohol + β₃Gender*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 12: goggles.csv - Answers

\[
\text{Attractiveness} = 60.625 + \begin{pmatrix} 
0 & 1.875 & 0 \\
-3.125 & 0 & 6.250 \\
-1.875 & 6.250 & -28.125 \\
\end{pmatrix} \begin{pmatrix} 
\text{if None} \\
\text{if 2 Pints} \\
\text{if Male and 2 Pints} \\
\end{pmatrix} + \begin{pmatrix} 
0 \\
6.250 \\
-28.125 \\
\end{pmatrix} \begin{pmatrix} 
\text{if Female} \\
\text{if Male} \\
\end{pmatrix}
\]

- goggles.csv

```
goggles <- read.csv("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
    \[60.625 - 3.125 + 6.250 - 28.125 = 35.625\]

```r
> tapply(goggles$attractiveness, goggles$subgroup, mean)  
Female None  Male None Female 2 Pints  Male 2 Pints Female 4 Pints  Male 4 Pints  
60.625  66.875  62.500  66.875   57.500   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

```r
# Load data
treelight <- read.csv("treelight.csv")

# Plot
plot(treelight$Light ~ treelight$Depth, col=treelight$Species, pch=16, cex=1.5)
legend("topright", c("Broadleaf", "Conifer"), pch=16, col=1:2)
```
The linear model perspective

Categorical and continuous factors

```r
lm(Light~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

summary(linear.treelight)
```

Complete model
The linear model perspective
Categorical and continuous factors

• Additive model:

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

```
Coefficients: 
Estimate Std. Error  t value     Pr(>|t|) 
(Intercept)  7962.0     231.3 34.153 < 2e-16 ***
Depth        -262.2      39.9  -6.567 2.13e-06 ***
SpeciesConifer -3113.0    231.6 -13.442 1.78e-11 ***
---
signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 . ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 5314 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

`abline()` does not work directly here 😞

\[
y = \beta_0 + \beta_1 x
\]

> `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

No interaction

Both Effect
The linear model perspective
Categorical and continuous factors

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

\[
\begin{array}{ccc}
(\text{Intercept}) & \text{Depth} & \text{Species} \\
7962.0316 & -262.1656 & -3113.0265 \\
\end{array}
\]

cf.add[1]

It’s a vector!

abline(cf.add[1], cf.add[2], col="black")
abline(cf.add[1]+cf.add[3], cf.add[2], col="red")

Broadleaf:
Light = 7962.03 -262.17*Depth

Conifer:
Light = (7962.03-3113.03) -262.17*Depth
Exercise 13: 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 13: treelight.csv

cf<-coefficients(linear.treelight)

plot(treelight$Light~treelight$Depth, col=treelight$Species, pch=16, las=1)
legend("topright", c("Broadleaf", "Conifer"), pch=16, col=1:2)
abline(cf[1], cf[2], col="black", lty="dashed")

cf.add<-coefficients(linear.treelight.add)
abline(cf.add [1], cf.add [2], col="black")
abline(cf.add[1]+cf.add [3], cf.add [2], col="red")
Exercise 13: treelight.csv

• Extract the parameters from the complete model

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

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>Depth</th>
<th>SpeciesConifer</th>
<th>Depth:SpeciesConifer</th>
</tr>
</thead>
<tbody>
<tr>
<td>7798.56552</td>
<td>-221.12564</td>
<td>-2784.58333</td>
<td>-71.03575</td>
</tr>
</tbody>
</table>

```r
plot(treelight$Light~treelight$Depth, col=treelight$Species, pch=16, las=1)
legend("topright", c("Broadleaf", "Conifer"), pch=16, col=1:2)
abline(cf[1], cf[2], col="black", lty="dashed")
```

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
    \[(7798.57-2784.58)-(221.13+71.04)\times4 = 4413.63\]
  • In a broadleaf species, 6 metres from the top of the canopy
    \[7798.57-221.13\times6 = 6471.79\]
Linear model

Simplest
\[ y = \beta_0 + \beta_1 x \]

With 2 factors
\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \]

Let's not forget the error
\[ y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i \]

General formula
\[ y_i = (\text{model}) + \text{error}_i \]
\[ y_i = (\text{model}) + \text{error}_i \]