In [1]:
```
library(Stat2Data)
library(leaps)
```

In [2]:
```
options(repr.plot.width=8, repr.plot.height=8)
```

Problem 1

In [3]:
```
data(HighPeaks)
head(HighPeaks)
```

A data.frame: 6 × 6

	Peak	Elevation	Difficulty	Ascent	Length	Time
	<fct>	<int>	<int>	<int>	<dbl>	<dbl>
1	Mt. Marcy	5344	5	3166	14.8	10.0
2	Algonquin Peak	5114	5	2936	9.6	9.0
3	Mt. Haystack	4960	7	3570	17.8	12.0
4	Mt. Skylight	4926	7	4265	17.9	15.0
5	Whiteface Mtn.	4867	4	2535	10.4	8.5
6	Dix Mtn.	4857	5	2800	13.2	10.0

a.

$Peak$ contains the name of each mountain. This isn't a useful variable for developing a regression model.

In [4]:
```
models <- regsubsets(Time ~ Elevation + Difficulty + Ascent + Length,
                     data = HighPeaks, nbest = 2)
sum <- summary(models)
cbind(as.data.frame(sum$outmat), sum$rsq, sum$adjr2, sum$cp)
```

A data.frame: 7 × 7

	Elevation	Difficulty	Ascent	Length	sum$rsq	sum$adjr2	sum$cp
	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
1 (1)				*	0.7370358	0.7310593	25.412218
1 (2)		*			0.6566249	0.6488209	46.025951
2 (1)		*		*	0.7962182	0.7867400	12.240486
2 (2)	*			*	0.7702826	0.7595980	18.889226
3 (1)	*	*		*	0.8272018	0.8148590	6.297702
3 (2)		*	*	*	0.7995560	0.7852385	13.384844
4 (1)	*	*	*	*	0.8400656	0.8244622	5.000000

The highest R^2 (and adjusted R^2) comes from model 4(1), with $Elevation$, $Difficulty$, $Ascent$, and $Length$ as explanatory variables. We fit the model below:

In [5]:
```
fit <- lm(Time ~ Elevation + Difficulty + Ascent + Length, data = HighPeaks)
summary(fit)
```

```
Call:
lm(formula = Time ~ Elevation + Difficulty + Ascent + Length,
    data = HighPeaks)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.77942 -0.81216 -0.08647  0.68962  3.06736 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.9567864  2.2307630   2.670  0.01082 *  
Elevation   -0.0016703  0.0005183  -3.223  0.00249 ** 
Difficulty   0.8654527  0.2285275   3.787  0.00049 ***
Ascent       0.0006011  0.0003310   1.816  0.07669 .  
Length       0.4440084  0.0812523   5.465 2.49e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.171 on 41 degrees of freedom
Multiple R-squared:  0.8401,	Adjusted R-squared:  0.8245 
F-statistic: 53.84 on 4 and 41 DF,  p-value: 8.738e-16
```

The fitted model is

$$Time = 5.9567864 - 0.0016703Elevation + 0.8654527Difficulty + 0.0006011Ascent \\ + 0.4440084Length$$

The R^2 is 0.8401.

b.

In [6]:
```
set.seed(2022)
train = sample(46, 36)
```

Build and fit the model using the training sample:

In [7]:
```
fit.training <- lm(Time ~ Length, data = HighPeaks[train,])
summary(fit.training)
```

```
Call:
lm(formula = Time ~ Length, data = HighPeaks[train, ])

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5011 -0.7828  0.0827  0.6107  3.9475 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.02529    0.90221   2.245   0.0314 *  
Length       0.68920    0.07006   9.838 1.77e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.466 on 34 degrees of freedom
Multiple R-squared:   0.74,	Adjusted R-squared:  0.7324 
F-statistic: 96.78 on 1 and 34 DF,  p-value: 1.77e-11
```

Predict $Time$ based on the holdout sample:

In [8]:
```
time.hat <- predict(fit.training, newdata = HighPeaks[-train,])
```

Compute the cross-validation correlation:

In [9]:
```
cor(time.hat, HighPeaks[-train,]$Time)
```

0.851089332424344

Problem 2

In [10]:
```
data(Leafhoppers)
head(Leafhoppers)
```

A data.frame: 6 × 3

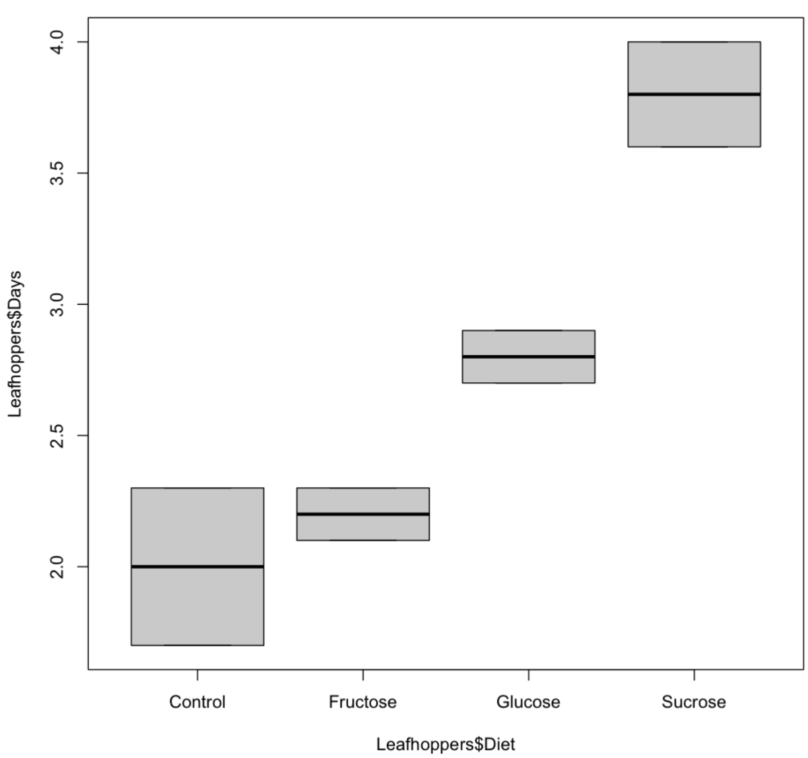
	Dish	Diet	Days
	<int>	<fct>	<dbl>
1	1	Control	2.3
2	2	Control	1.7
3	3	Sucrose	3.6
4	4	Sucrose	4.0
5	5	Glucose	2.9
6	6	Glucose	2.7

a.

This is an experiment: the researchers control the values of the explanatory variable $Diet$.

b.

In [11]:
```
boxplot(Leafhoppers$Days ~ Leafhoppers$Diet)
```



c.

In [12]:
```
y.bar <- mean(Leafhoppers$Days)
```

```
y.bar
```

2.7

d.

In [13]:
```
y.bar.k <- tapply(Leafhoppers$Days, Leafhoppers$Diet, mean)
alpha.k <- y.bar.k - y.bar

alpha.k
```

Control: -0.7 **Fructose:** -0.5 **Glucose:** 0.0999999999999996 **Sucrose:** 1.1

e.

Each population (group) has the same standard deviations

As we see below,

$$\frac{\text{max SD}}{\text{min SD}} \approx \frac{0.4242}{0.1414} \approx 3,$$

which is larger than the 2 that our rule of thumb from class allows. This condition is violated.

In [14]:
```
tapply(Leafhoppers$Days, Leafhoppers$Diet, sd)
```

Control: 0.424264068711928 **Fructose:** 0.141421356237309 **Glucose:** 0.141421356237309 **Sucrose:** 0.282842712474619

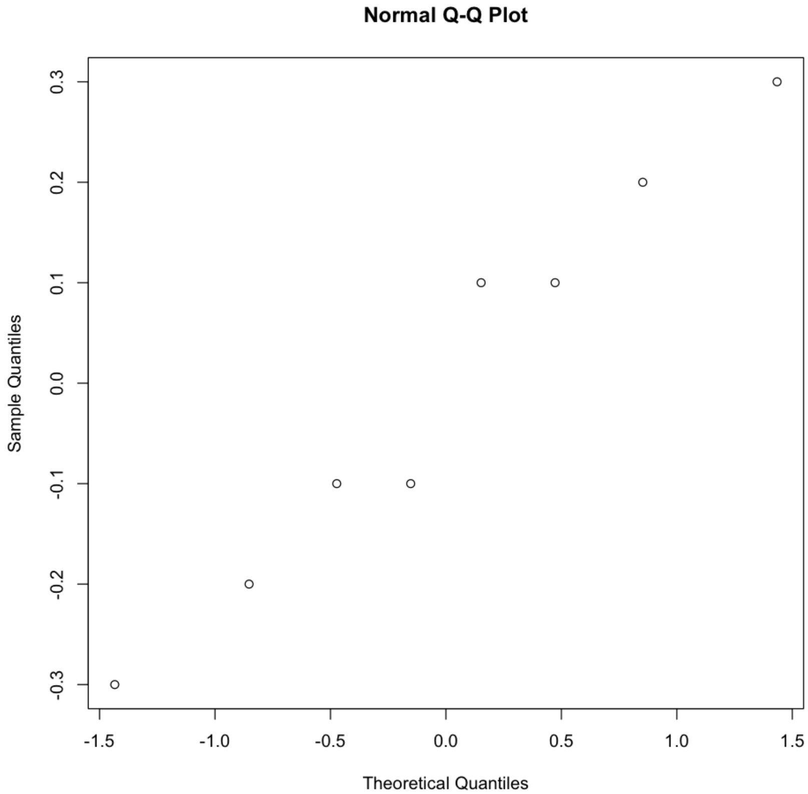
In [15]:
```
0.4242/0.1414
```

3

Each population (group) is Normal

As we see below, the Normal Q-Q plot of the residuals of the one-way ANOVA is an approximately straight line. This condition is satisfied.

In [16]:
```
test <- aov(Days ~ Diet, data = Leafhoppers)
qqnorm(residuals(test))
```



After accounting for group membership, responses are independent

The groups – in this case, the diets – were randomly assigned, so responses should be independent after accounting for group membership. This condition is satisfied.

f.

We fit the one-way ANOVA model in part e, in order to check the Normality of the residuals.

In [17]:
```
summary(test)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
Diet         3   3.92   1.307   17.42 0.00925 **
Residuals    4   0.30   0.075
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We perform a one-way ANOVA F-test:

1. The hypotheses:

$$H_0 : \mu_{Control} = \mu_{Fructose} = \mu_{Glucose} = \mu_{Sucrose} \quad \text{vs.} \quad H_A : \text{at least one of the } \mu_k \text{ is diferent}$$

2. Test statistic: $F = 17.42$

3. p-value = 0.00925

4. Assume a significance level of 0.05. Since the p-value is less than 0.05, we reject H_0. We see significant evidence that the mean time until half the leafhoppers in a dish died differs by diet.

g.

In [18]:
```
alpha <- 0.05
```

```
n <- nrow(Leafhoppers)
K <- 4  # control, fructose, glucose, sucrose

t <- qt(1 - alpha/2, df = n - K)
sd <- sqrt(0.075)
n.k <- tapply(Leafhoppers$Days, Leafhoppers$Diet, length)

ci.lower <- y.bar.k - t * sd * sqrt(1 / n.k)
ci.upper <- y.bar.k + t * sd * sqrt(1 / n.k)

ci.lower
ci.upper
```

Control: 1.46234371729549 **Fructose:** 1.66234371729549 **Glucose:** 2.26234371729549 **Sucrose:** 3.26234371729549

Control: 2.53765628270451 **Fructose:** 2.73765628270451 **Glucose:** 3.33765628270451 **Sucrose:** 4.33765628270451

The 95% CI for the mean length of life for leafhoppers on the control diet is (1.46234371729549, 2.53765628270451).

Problem 3

a.

Option (b)

b.

Option (d)

c.

(a) The test statistic for the coefficient of Lot is $5.657/3.075 \approx 1.839$.

(b) Yes, at the $\alpha = 0.05$ level, we have significant evidence that the overall model is effective, because the p-value of the F-test (0.000985) is less than α.

(c) No, at the $\alpha = 0.05$ level, we do not have significant evidence that $Size$ is associated with $Price$, after accounting for Lot, because the p-value 0.2068 is larger than α.

d.

Option (c)

Problem 4

a.

$$\begin{aligned} Height &= \beta_0 + \beta_1 Water + \beta_2 FertA + \beta_3 (Water \times FertA) + \varepsilon \\ &= \beta_0 + \beta_1 Water + \beta_2 (1) + \beta_3 (Water)(1) + \varepsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Water + \varepsilon \end{aligned}$$

The intercept of for fertilizer A is $\beta_0 + \beta_2$.

b.

The slope of $Water$ for fertilizer A is $\beta_1 + \beta_3$.

c.

$$\begin{aligned} Height &= \beta_0 + \beta_1 Water + \beta_2 FertA + \beta_3 (Water \times FertA) + \varepsilon \\ &= \beta_0 + \beta_1 Water + \beta_2 (0) + \beta_3 (Water)(0) + \varepsilon \\ &= \beta_0 + \beta_1 Water + \varepsilon \end{aligned}$$

The slope of $Water$ for fertilizer B is β_1.

d.

The interaction term $Water \times FertA$.

Problem 5

a.

Null hypothesis $H_0 : \beta_2 = \beta_3 = 0$

Alternative hypothesis $H_A : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$

b.

The reduced model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

c.

Nested F-test.

d.

(i) The SSE for reduced model is 36.234. (Row 1 of the table corresponds to the first argument given to `anova`, and row 2 corresponds to the second argument.)

(ii) The degrees of freedom is $n - (k + 1)$, where k is the number of predictors.

So, using information about the reduced model, we have $n - (3 + 1) = 31$, or $n = 35$.

Equivalently, using information about the full model, we have $n - (5 + 1) = 29$, or $n = 35$.

(iii) Assuming a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis, because the p-value 0.6071 is greater than α.

We conclude that we do not see evidence that including X_2 and X_3 provides a significant improvement. We should use the reduced model.