MODEL SELECTION CRITERIA IN R:

1. R^2 **statistics:** We may use

$$R^2 = \frac{\text{SS}_\text{R}}{\text{SS}_\text{T}} = 1 - \frac{\text{SS}_\text{Res}}{\text{SS}_\text{T}} \qquad \text{or} \qquad R^2_\text{Adj} = 1 - \frac{\text{SS}_\text{Res}/(n-p)}{\text{SS}_\text{T}/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2).$$

where p is the total number of parameters. R^2 does not take into account model complexity (that is, the number of parameters fitted), whereas R^2_Adj does.

2. **Mean Square Residual:** We consider

$$\text{MS}_\text{Res} = \frac{\text{SS}_\text{Res}}{(n-p)}$$

and note that

$$R^2_\text{Adj} = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - \frac{\text{SS}_\text{Res}}{\text{SS}_\text{T}}\right) = 1 - \frac{\text{MS}_\text{Res}}{\text{SS}_\text{T}/(n-1)}$$

so that maximizing R^2_Adj corresponds exactly to minimizing MS_Res.

3. **Mallows's C_p statistic:** Let $\mu_i = \mathbb{E}_{Y_i|X_i}[Y_i|\mathbf{x}_i]$ and $\widehat{\mu}_i = \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\widehat{Y}_i|\mathbf{x}_i]$ be the *modelled* and *fitted* expected values of response Y_i at predictor values $\mathbf{x}_i$ respectively. The expected (or mean) squared error (MSE) of the fit for datum i is

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[(\widehat{Y}_i - \mu_i)^2|\mathbf{x}_i]$$

which can be decomposed

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[(\widehat{Y}_i - \mu_i)^2|\mathbf{x}_i] = \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[(\widehat{Y}_i - \widehat{\mu}_i)^2|\mathbf{x}_i] + (\widehat{\mu}_i - \mu_i)^2 &= \text{Var}_{\mathbf{Y}|\mathbf{X}}[\widehat{Y}_i|\mathbf{x}_i] + (\widehat{\mu}_i - \mu_i)^2 \\ &= \text{variance for datum } i + (\text{bias for datum } i)^2\end{aligned}$$

Let

$$\text{SS}_\text{B} = \sum_{i=1}^{n}(\widehat{\mu}_i - \mu_i)^2 = (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}})^\top(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}) = \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{\mu}$$

say, denote the total squared bias, aggregated across all data points, and

$$\text{FMSE} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\text{Var}_{\mathbf{Y}|\mathbf{X}}[\widehat{Y}_i|\mathbf{x}_i] + (\widehat{\mu}_i - \mu_i)^2\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{Var}_{\mathbf{Y}|\mathbf{X}}[\widehat{Y}_i|\mathbf{x}_i] + \frac{\text{SS}_\text{B}}{\sigma^2}.$$

Recall that if $\mathbf{H}$ is the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ then

$$\text{Var}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}|\mathbf{x}] = \text{Var}_{\mathbf{Y}|\mathbf{X}}[\mathbf{H}\mathbf{Y}|\mathbf{x}] = \sigma^2\mathbf{H}^\top\mathbf{H} = \sigma^2\mathbf{H}$$

and so

$$\sum_{i=1}^{n}\text{Var}_{\mathbf{Y}|\mathbf{X}}[\widehat{Y}_i|\mathbf{x}_i] = \text{Trace}(\sigma^2\mathbf{H}) = \sigma^2\text{Trace}(\mathbf{H}) = p\sigma^2$$

Also by previous results for quadratic forms

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\text{SS}_\text{Res}\mathbf{X}] &= \mathbb{E}_{\mathbf{Y}|\mathbf{X}}\left[\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{Y}\middle|\mathbf{X}\right] \\ &= \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{\mu} + \text{Trace}(\sigma^2(\mathbf{I}_n - \mathbf{H})) \\ &= (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}})^\top(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}) + (n-p)\sigma^2 \\ &= \text{SS}_\text{B} + (n-p)\sigma^2.\end{aligned}$$

Therefore we may rewrite

$$\text{FMSE} = \frac{1}{\sigma^2}\left[p\sigma^2 + \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\text{SS}_{\text{Res}}|\mathbf{X}] - (n-p)\sigma^2\right] = \frac{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\text{SS}_{\text{Res}}|\mathbf{X}]}{\sigma^2} - n + 2p$$

An estimator of this quantity is

$$C_p = \frac{\text{SS}_{\text{Res}}}{\widehat{\sigma}^2} - n + 2p$$

where $\widehat{\sigma}^2$ is some estimator of σ^2 derived, say, from the the 'largest' model that is being considered.

C_p is *Mallows's statistic*. We choose the model that minimizes C_p. We have that

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[C_p|\mathbf{X}] = p.$$

4. **Akaike's Information Criterion (AIC):** We define for a probability model with parameters θ

$$\text{AIC} = -2\ell(\widehat{\theta}) + 2\text{dim}(\theta)$$

where $\ell(\theta)$ is the log-likelihood function, $\widehat{\theta}$ is the maximum likelihood estimate of the parameter θ, and $\text{dim}(\theta)$ is the dimension of θ.

For linear regression models under a normality assumption, we have that $\theta = (\beta, \sigma^2)$ with

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\beta)^2$$

Plugging in $\widehat{\beta}$ and $\widehat{\sigma}^2_{\text{ML}}$, we obtain

$$\ell(\widehat{\beta}, \widehat{\sigma}^2_{\text{ML}}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{\text{SS}_{\text{Res}}}{n}\right) - \frac{n\text{SS}_{\text{Res}}}{2\text{SS}_{\text{Res}}}$$

so therefore, writing

$$c(n) = n\log(2\pi) + n$$

for the constant function of n, we have

$$\text{AIC} = c(n) + n\log\left(\frac{\text{SS}_{\text{Res}}}{n}\right) + 2(p+1).$$

This is Akaike's Information Criterion: we choose the model with the lowest value of AIC. The constant $c(n)$ need not be included in the calculation as it is constant across all models considered.

5. **Bayesian Information Criterion (BIC):** The Bayesian Information Criterion (BIC) is a modification of AIC. We define

$$\text{BIC} = n\log\left(\frac{\text{SS}_{\text{Res}}}{n}\right) + (p+1)\log(n).$$

and again choose the model with the smallest BIC.

SIMULATION STUDY

We have the model for three continuous predictors X_1, X_2, X_3

$$Y_i = 2 + 2x_{i1} + 2x_{i2} - 2x_{i1}x_{i2} + \epsilon_i$$

with $\sigma^2 = 1$. We have $n = 200$. Here is the simulation code:

```
set.seed(798)
n<-200; p<-3
Sig<-rWishart(1,p+2,diag(1,p)/(p+2))[,,1]
library(MASS)
x<-mvrnorm(n,mu=rep(0,p),Sigma=Sig)
be<-c(2,2,2,0,-2)
xm<-cbind(rep(1,n),x,x[,1]*x[,2])
Y<-xm %*% be + rnorm(n)

x1<-x[,1]
x2<-x[,2]
x3<-x[,3]

fit0<-lm(Y~1)
fit1<-lm(Y~x1)
fit2<-lm(Y~x2)
fit3<-lm(Y~x3)
fit12<-lm(Y~x1+x2)
fit13<-lm(Y~x1+x3)
fit23<-lm(Y~x2+x3)
fit123<-lm(Y~x1+x2+x3)
fit12i<-lm(Y~x1*x2)
fit13i<-lm(Y~x1*x3)
fit23i<-lm(Y~x2*x3)
fit123i<-lm(Y~x1*x2*x3)

criteria.eval<-function(fit.obj,nv,bigsig.hat){
        cvec<-rep(0,5)
        SSRes<-sum(residuals(fit.obj)^2)
        p<-length(coef(fit.obj))
        cvec[1]<-summary(fit.obj)$r.squared
        cvec[2]<-summary(fit.obj)$adj.r.squared
        cvec[3]<-SSRes/bigsig.hat^2-n+2*p

        #AIC in R computes
      # n*log(sum(residuals(fit.obj)^2)/n)+2*(length(coef(fit.obj))+1)+n*log(2*pi)+n
        cvec[4]<-AIC(fit.obj)

        #BIC in R computes
      # n*log(sum(residuals(fit.obj)^2)/n)+log(n)*(length(coef(fit.obj))+1)+n*log(2*pi)+n
        cvec[5]<-BIC(fit.obj)

        return(cvec)
}

bigs.hat<-summary(fit123i)$sigma
cvals<-matrix(0,nrow=12,ncol=5)
cvals[1,]<-criteria.eval(fit0,n,bigs.hat)
cvals[2,]<-criteria.eval(fit1,n,bigs.hat)
cvals[3,]<-criteria.eval(fit2,n,bigs.hat)
cvals[4,]<-criteria.eval(fit3,n,bigs.hat)
cvals[5,]<-criteria.eval(fit12,n,bigs.hat)
cvals[6,]<-criteria.eval(fit13,n,bigs.hat)
cvals[7,]<-criteria.eval(fit23,n,bigs.hat)
```

```
cvals[8,]<-criteria.eval(fit123,n,bigs.hat)
cvals[9,]<-criteria.eval(fit12i,n,bigs.hat)
cvals[10,]<-criteria.eval(fit13i,n,bigs.hat)
cvals[11,]<-criteria.eval(fit23i,n,bigs.hat)
cvals[12,]<-criteria.eval(fit123i,n,bigs.hat)

Criteria<-data.frame(cvals)
names(Criteria)<-c('Rsq','Adj.Rsq','Cp','AIC','BIC')

rownames(Criteria)<-c('1','x1','x2','x3','x1+x2','x1+x3','x2+x3','x1+x2+x3',
                      'x1*x2','x1*x3','x2*x3','x1*x2*x3')

round(Criteria,4)
```

```
:              Rsq Adj.Rsq       Cp      AIC      BIC
: 1         0.0000  0.0000 799.1174 875.3679 881.9646
: x1        0.2505  0.2467 551.3719 819.7068 829.6018
: x2        0.5189  0.5164 283.7367 731.0417 740.9366
: x3        0.1196  0.1151 681.8659 851.8930 861.7880
: x1+x2     0.7055  0.7026  99.6020 634.8392 648.0325
: x1+x3     0.3890  0.3828 415.2121 780.8275 794.0208
: x2+x3     0.5239  0.5190 280.7558 730.9543 744.1476
: x1+x2+x3 0.7058  0.7013 101.3825 636.6897 653.1813
: x1*x2     0.8032  0.8001   4.2736 556.2961 572.7877
: x1*x3     0.4074  0.3983 398.9377 776.7363 793.2279
: x2*x3     0.5240  0.5167 282.6702 732.9183 749.4098
: x1*x2*x3 0.8074  0.8004   8.0000 559.8933 589.5782
```

This reveals the model $X_1 * X_2 = X_1 + X_2 + X_1 : X_2$ as most appropriate model.

```
summary(fit12i)
```

```
:
: Call:
: lm(formula = Y ~ x1 * x2)
:
: Residuals:
:      Min       1Q   Median       3Q      Max
: -2.43675 -0.68819 -0.01849  0.68452  2.18404
:
: Coefficients:
:             Estimate Std. Error t value Pr(>|t|)
: (Intercept)  2.02079    0.06895  29.310   <2e-16 ***
: x1           1.91766    0.12823  14.954   <2e-16 ***
: x2           2.05010    0.10398  19.717   <2e-16 ***
: x1:x2       -1.91633    0.19438  -9.859   <2e-16 ***
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
:
: Residual standard error: 0.9578 on 196 degrees of freedom
: Multiple R-squared:  0.8032,Adjusted R-squared:  0.8001
: F-statistic: 266.6 on 3 and 196 DF,  p-value: < 2.2e-16
```

The parameter estimates are therefore

$$\widehat{\beta}_0 = 2.0208 \quad \widehat{\beta}_1 = 1.9177 \quad \widehat{\beta}_2 = 2.0501 \quad \widehat{\beta}_{12} = -1.9163$$

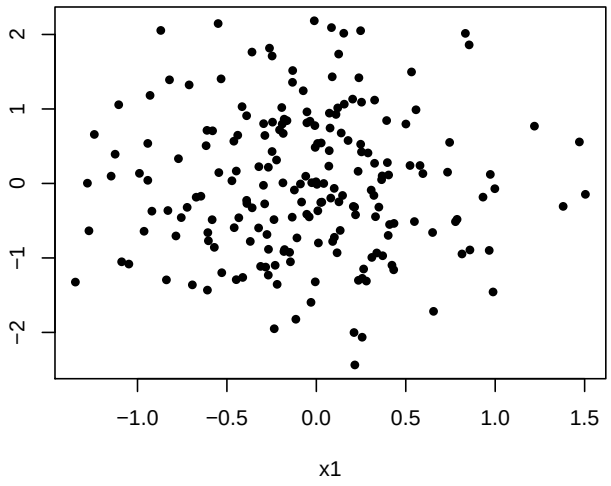
which are close to the data generating values.

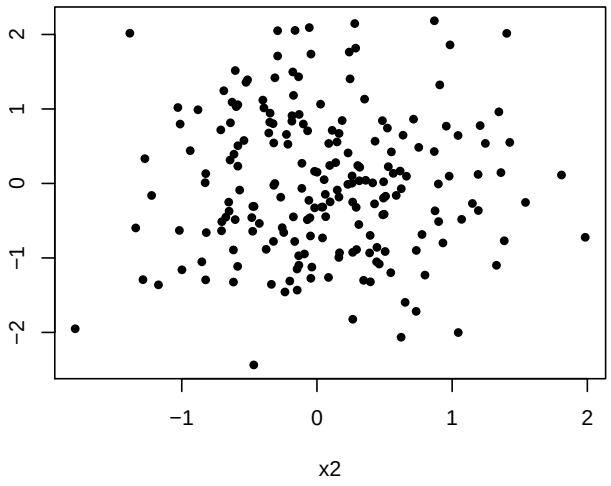
For an equivalent ANOVA test to the one in the summary output:

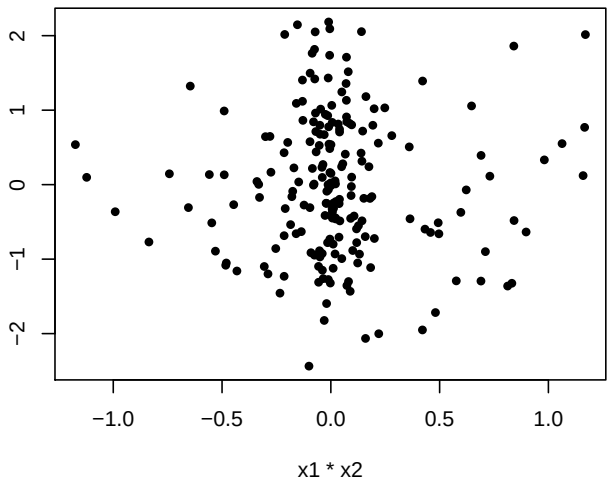
```
anova(fit12,fit12i)

: Analysis of Variance Table
:
: Model 1: Y ~ x1 + x2
: Model 2: Y ~ x1 * x2
:   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
: 1    197 268.98
: 2    196 179.81  1    89.166 97.193 < 2.2e-16 ***
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2),mar=c(4,2,1,2))
plot(x1,residuals(fit12i),pch=19,cex=0.75)
plot(x2,residuals(fit12i),pch=19,cex=0.75)
plot(x1*x2,residuals(fit12i),pch=19,cex=0.75)
```







Finally, for an **incorrect** model we obtain misleading results:

```
summary(fit13i)

:
: Call:
: lm(formula = Y ~ x1 * x3)
:
: Residuals:
:     Min      1Q  Median      3Q     Max
: -5.3750 -1.0790  0.0121  0.9794  4.5081
:
: Coefficients:
:             Estimate Std. Error t value Pr(>|t|)
: (Intercept)   2.0229     0.1186  17.057  < 2e-16 ***
: x1            2.0842     0.2193   9.503  < 2e-16 ***
: x3            0.9138     0.1337   6.834 1.02e-10 ***
: x1:x3        -0.5377     0.2184  -2.462   0.0147 *
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
:
: Residual standard error: 1.662 on 196 degrees of freedom
: Multiple R-squared:  0.4074,Adjusted R-squared:  0.3983
: F-statistic: 44.91 on 3 and 196 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2),mar=c(4,2,1,2))
plot(x1,residuals(fit13i),pch=19,cex=0.75)
plot(x3,residuals(fit13i),pch=19,cex=0.75)
plot(x1*x3,residuals(fit13i),pch=19,cex=0.75)
```