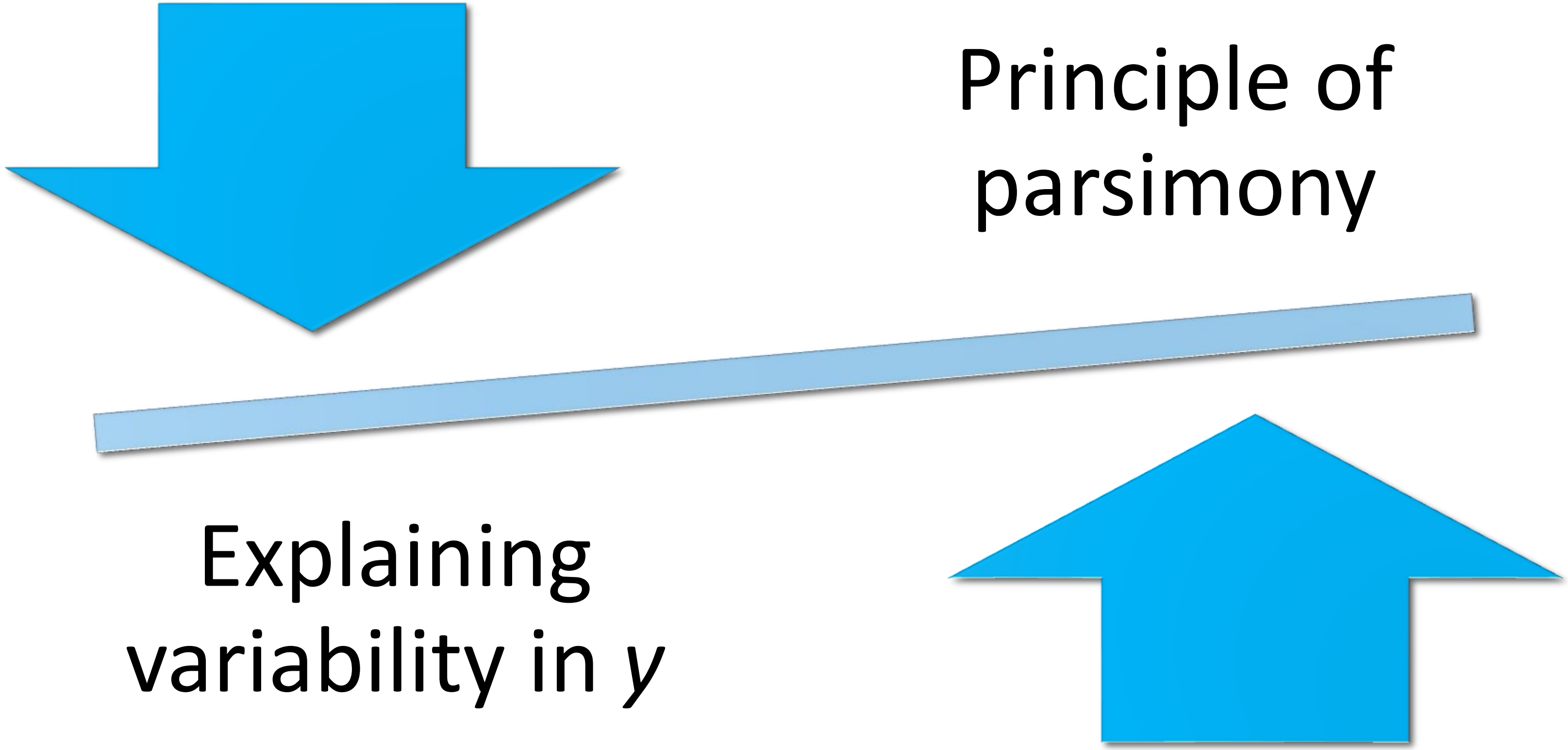
Model Building: All Subsets Regression

CHRIS SUTTON, MARCH 2024

Conflicting objectives



Approaches

There are a number of techniques to help decide which explanatory variables to keep in a multiple linear regression model:

1. Using F tests to delete variables
2. Considering All-Subsets Regression
3. Backward Elimination
4. Stepwise Regression or Modified Forward Regression
5. Akaike's Information Criterion (AIC)

Subset deletion by F test

- evaluate q parameter (reduced) alternative to p parameter (full) model
- produce ANOVA for full and reduced models to get SS
- calculate ExtraSS (increase in regression or reduction in residual SS)
- test H_0: $\beta_q = \cdots = \beta_{p-1} = 0$ through a modified F test
- the F statistic used ExtraSS, p – q and full model S^2
- if we cannot reject H_0 we can work with the reduced model

All Subsets Regression

Different multiple linear regression models we may wish to consider

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{3i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_2 x_{2i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_3 x_{3i} + \varepsilon_i$$

$$y_i = \beta_0 + \varepsilon_i$$

With 3 explanatory variables there are 8 potential linear models

In general there will be 2^{p-1} models

All subsets

With $p - 1$ explanatory variables there are 2^{p-1} linear models

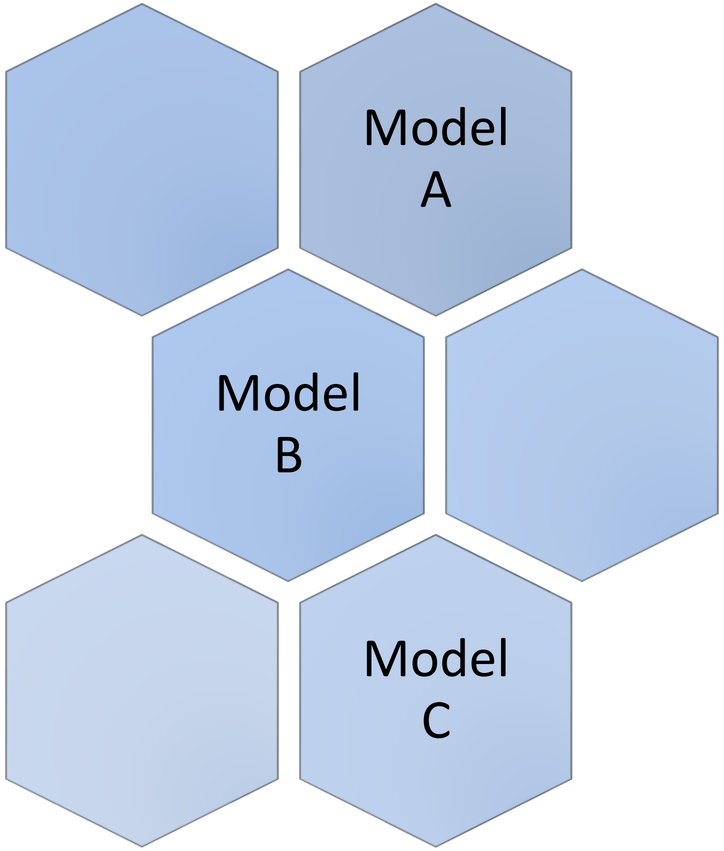
We would like some methods for evaluating them all

and then selecting the "best" one

The obvious method is to calculate some Statistic for each and compare these, selecting the model with the "best" properties as measured by the Statistic

Would allow the creation of a 'League Table' for all the models

What characterises a good / better / best model?



Candidate Statistics

Variance

MS_E

R-sq

Adjusted R-sq

Mallow's

Variance

We would like a model with the lowest possible variance of the residuals

But σ^2 is unknown

This leads us to MS_E our unbiased estimator for σ^2

Mean Square of Residuals

If we simply select the model with lowest MS_E that will often be the full model

So this is a very conservative method of model selection

Better might be to find a model that

- keeps MS_E close to full model MS_E
- with the smallest number of explanatory variables

A plot of all the model MS_E against number of variables is good way to judge this

R-squared

The Coefficient of Determination or R^2 is

$$R^2 = 100\% \frac{SS_R}{SS_T} = 100\%(1 - \frac{SS_E}{SS_T})$$

Adding more explanatory variables will always increase R^2

So we cannot simply find the model that maximises R^2 as that will always be the full model

Again we could plot R^2 against number of variables for all the models and see where increases in R^2 start to level off

Adjusted R-squared

R^2 does not take account of the number of explanatory variables

Therefore is not a “fair” way of comparing a 5 variable model with a 8 variable one

We have seen Adjusted R-sq alongside [Multiple] R-sq in `summary()` output

Adjusted $R^2 = 100\%(1 - (n-1)\frac{MS_E}{SS_T})$

R-sq versus Adjusted R-sq

- R^2 always increases when we add a new explanatory variable
- Adjusted R^2 only increases if the new variable's parameter is significant
- specifically Adjusted R^2 only increases if the F statistic associated with the parameter for the new variable is > 1
- selecting the model with highest Adjusted R^2 does not automatically lead to the full model and is better way of comparing models of different sizes

Mallow's Statistic

Mallow's Statistic (or sometimes Mallow's Cp) or C_k

For a model with k parameters using n observations and $\varepsilon_i \sim N(0, \sigma^2)$

$$C_k = \frac{SS_E^{(k)}}{\sigma^2} + 2k - n$$

where $SS_E^{(k)}$ is the residual sum of squares for the linear regression model with those k parameters

Using Mallow's Statistic

If the k parameter model has all the statistically significant parameters in it

$E[SS_E^{(k)}] = (n - k)\,\sigma^2$

and then

$C_k = (n - k) + 2k - n = k$

If the model excludes one or more statistically significant parameters

$E[SS_E^{(k)}] > (n - k)\,\sigma^2$ and then $C_k > k$

This suggests choosing the model with C_k closest to k

2nd use of Mallow

It can also be shown that Mallow's Statistic is also an estimator of the mean square error of prediction in a linear regression model with k parameters

This would suggest choosing the model with smallest C_k

So we have two possible selection rules:

- closest to k
- minimise

[we did say there was no one correct answer in model selection]

Practical issues with Mallow's

σ^2 used in the calculation of C_k is unknown

We usually replace it with $S^2 = MS_E^{full}$

- Note we take S^2 from the full model not the *k* parameter model
- This is how R estimates C_k

In R, if we have `full_model` and say `model_k` both constructed with `lm()`

Then Mallow's Statistic is found by

```
ols_mallows_cp(model_k, full_model)
```

Model building example
UK CPI inflation

Modelling objective

Can we build a multiple linear regression model for CPI inflation using other economic indicators as explanatory variables

Data on QM Plus `UK_Economic_CPI_Model_Data.csv`

Quarterly data on CPI and 6 potential explanatory variables 1989 – 2021

Potential explanatory variables

- GDP Growth
- M4 Money Supply
- Unemployment
- Household Income
- Savings Ratio
- FTSE100 value

Importing data and constructing the full model

```
> UK_Economic_CPI_Model_Data <-
read.csv("~/UK_Economic_CPI_Model_Data.csv")
>   View(UK_Economic_CPI_Model_Data)
> y = UK_Economic_CPI_Model_Data$CPI
> x1 = UK_Economic_CPI_Model_Data$GDP_Growth
> x2 = UK_Economic_CPI_Model_Data$M4_Growth
> x3 = UK_Economic_CPI_Model_Data$Unemployment
> x4 = UK_Economic_CPI_Model_Data$Household_Income
> x5 = UK_Economic_CPI_Model_Data$Savings
> x6 = UK_Economic_CPI_Model_Data$FTSE100
> full_model = lm(y~x1+x2+x3+x4+x5+x6)
> summary(full_model)
```

Full model output

```
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6)
 Coefficients:
              Estimate Std. Error t value
 (Intercept)  4.3825900  1.4831094   2.955
 x1          -0.0969511  0.0544612  -1.780
 x2           0.0397468  0.0294816   1.348
 x3           0.3728978  0.1205571   3.093
 x4          -0.1453724  0.0584997  -2.485
 x5          -0.1743909  0.0522764  -3.336
 x6          -0.0004899  0.0001330  -3.682
```

```
              Pr(>|t|)
 (Intercept) 0.003753 **
 x1          0.077535 .
 x2          0.180094
 x3          0.002456 **
 x4          0.014310 *
 x5          0.001127 **
 x6          0.000346 ***
```

Full model output continued

```
Multiple R-squared:  0.4364, Adjusted R-squared:  0.4086
F-statistic: 15.74 on 6 and 122 DF,  p-value: 2.51e-13
```

```
> qf(0.05, 6, 122, lower.tail = FALSE)
[1] 2.173733
```

Full model ANOVA

```
> anova(full_model)
Analysis of Variance Table
Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x1          1   2.139   2.139  1.2194 0.2716512
x2          1  13.856  13.856  7.8980 0.0057669
x3          1 101.654 101.654 57.9428 6.333e-12
x4          1   5.457   5.457  3.1105 0.0802916
x5          1  18.806  18.806 10.7195 0.0013797
x6          1  23.789  23.789 13.5598 0.0003456
Residuals 122 214.034   1.754
```

Overall (full) model significance

H_0: $\beta_1 = \beta_2 = \dots = \beta_6 = 0$

H_1: at least one of the parameters is not zero

F = Variance Ratio = 15.74

Under H_0: $F \sim F^{6}_{122}$ and $F^{6}_{122}(0.05)$ = 2.17 < 15.74

Therefore we reject H_0 at 95% significance

There is evidence that at least some of the parameters are non zero and therefore the model has overall significance

Consider 2 variables for subset deletion

GDP Growth	M4 Money Supply	Unemployment
Household Income	Savings Ratio	FTSE100 value

Reduced Model

Delete `x1 x4` keep `x2 x3 x5 x6`

```
> reduced_model = lm(y~x2+x3+x5+x6)
> summary(reduced_model)
```

Reduced Model output

```
Call:
lm(formula = y ~ x2 + x3 + x5 + x6)
Coefficients:
              Estimate Std. Error t value
(Intercept)  3.3416968  1.4877039   2.246
x2           0.0397521  0.0305786   1.300
x3           0.3539582  0.1215354   2.912
x5          -0.1276026  0.0503458  -2.535
x6          -0.0004266  0.0001352  -3.155
```

Reduced model output continued

```
Multiple R-squared:  0.3821, Adjusted R-squared:  0.3621
F-statistic: 19.17 on 4 and 124 DF,  p-value: 2.7e-12
```

We now need to complete a Subset deletion F test on the reduced versus the full model using the Extra Sum of Squares principle

$p - q = 7 - 5 = 2$

Reduced Model ANOVA

```
> anova(reduced_model)
Analysis of Variance Table
Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x2          1  14.435  14.435  7.6282  0.006621
x3          1 100.351 100.351 53.0291 3.325e-11
x5          1  11.457  11.457  6.0544  0.015248
x6          1  18.836  18.836  9.9537  0.002015
Residuals 124 234.655   1.892
```

Subset deletion test

H_0: $\beta_1 = \beta_4 = 0$ H_1: at least one of them is not zero

$ExtraSS = SS_E^{red} - SS_E^{full}$ = 234.655 – 214.034 = 20.621

$S^2 = MS_E^{full}$ = 1.754

$F^* = \frac{ExtraSS/(p-q)}{S^2}$ = (20.621/2)/1.754 = 5.878

Under H_0 $F^* \sim F_{n-p}^{p-q} = F_{122}^{2}$

```
> qf(0.05,2,122, lower.tail = FALSE)
[1] 3.070512
```

at 95% significance

F^*= 5.878 > 3.071 therefore we reject H_0 and cannot delete both variables

Consider just 1 variable for deletion

GDP Growth	M4 Money Supply	Unemployment
Household Income	Savings Ratio	FTSE100 value

Single variable deletion

The subset deletion of 2 variables did not pass the F test at 95%

Look at whether we can omit just `x1`

Can use a *t* test for this as p – q = 1

$H_0\colon \beta_1 = 0$ $H_0\colon \beta_1 \neq 0$

$t = {}^{\widehat{\beta}_1}/_{s.e.(\widehat{\beta}_1)} = -0.0969511/$ 0.0544612 = -1.780 from the full model

```
> qt(0.025, 122)
[1] -1.9796
```

t test results

Under H_0 $t \sim t_{n-p}$ in a two sided test at 95% significance $t_{122}(0.025)$ = 1.98

$|t|$ = 1.78 < 1.98

Therefore we cannot reject H_0

Hence we conclude β_1 is not significantly different from zero

And we can omit variable `x1` and move to a 5 variable model

All subsets regression

With 5 variables and $p = 6$ we have 32 potential multiple regression models

- null model
- 5 simple linear regression models
- 10 two variable models
- 10 three variable models
- 5 four variable models
- full model

Statistics we will consider for each of the 32 models

- Mean Square for Residuals
- R-squared
- Adjusted R-squared
- Mallow's Statistic C_k

```
UK_Economic_CPI_Model_Statistics <-
read.csv("~//UK_Economic_CPI_Model_Statistics.csv")

View(UK_Economic_CPI_Model_Statistics)
```

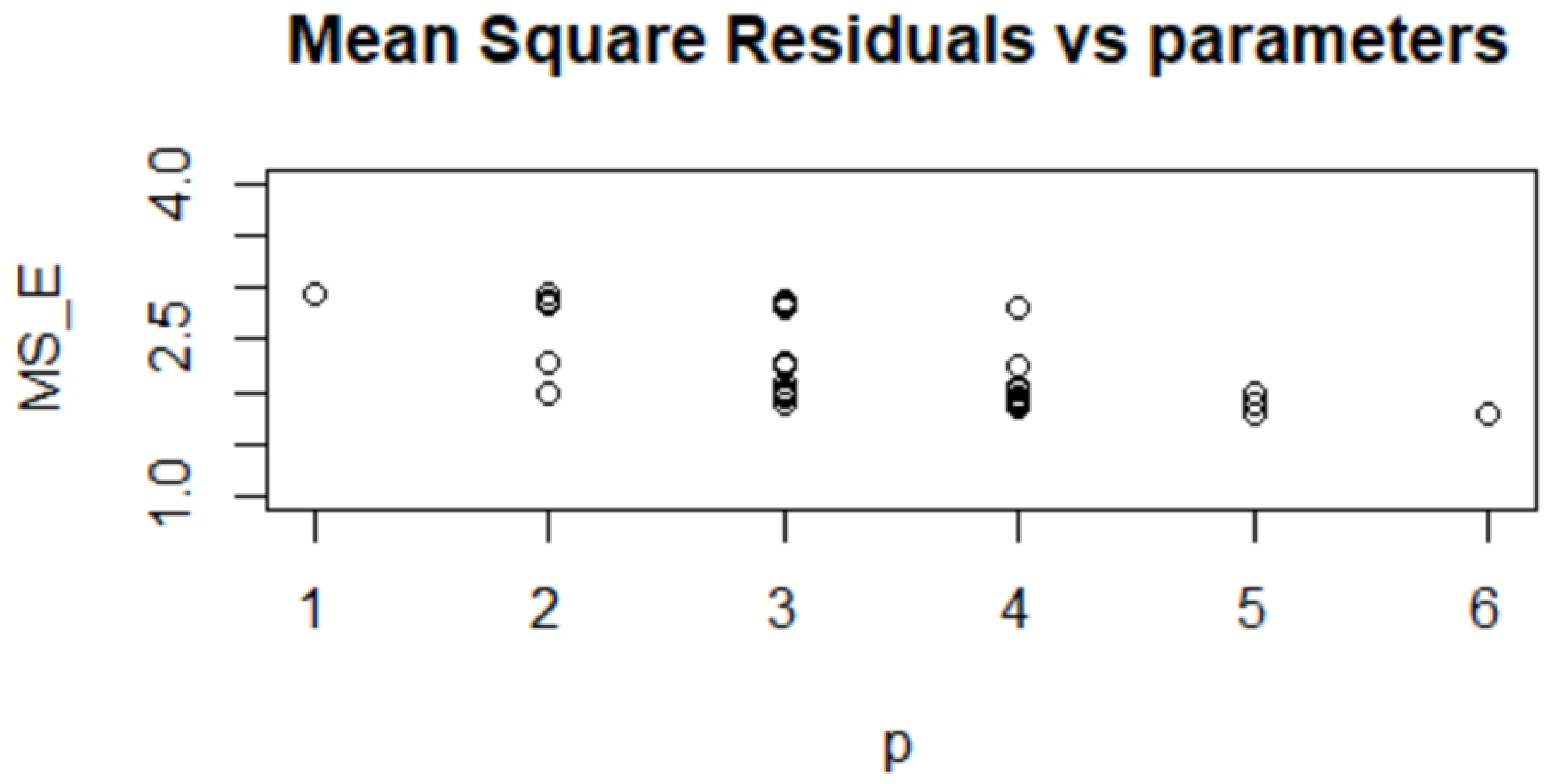
32 models constructed with `lm()`

```
> tail(UK_Economic_CPI_Model_Statistics,12)
   Model p   MSE     R2  AdjR2        Ck
21  m246 4 1.904 0.3684 0.3534 12.400000
22  m256 4 2.006 0.3398 0.3239 19.600000
23  m345 4 2.256 0.2573 0.2394 37.247059
24  m346 4 1.871 0.3794 0.3646 10.070588
25  m356 4 1.903 0.3736 0.3586 12.329412
26  m456 4 1.866 0.3857 0.3710  9.717647
27 m2345 5 1.974 0.3553 0.3345 18.235294
28 m2346 5 1.876 0.3828 0.3630 11.372549
29 m2356 5 1.892 0.3821 0.3621 12.492997
30 m2456 5 1.879 0.3865 0.3667 11.582633
31 m3456 5 1.794 0.4143 0.3954  5.630252
32  full 6 1.785 0.4217 0.3982  6.000000
```

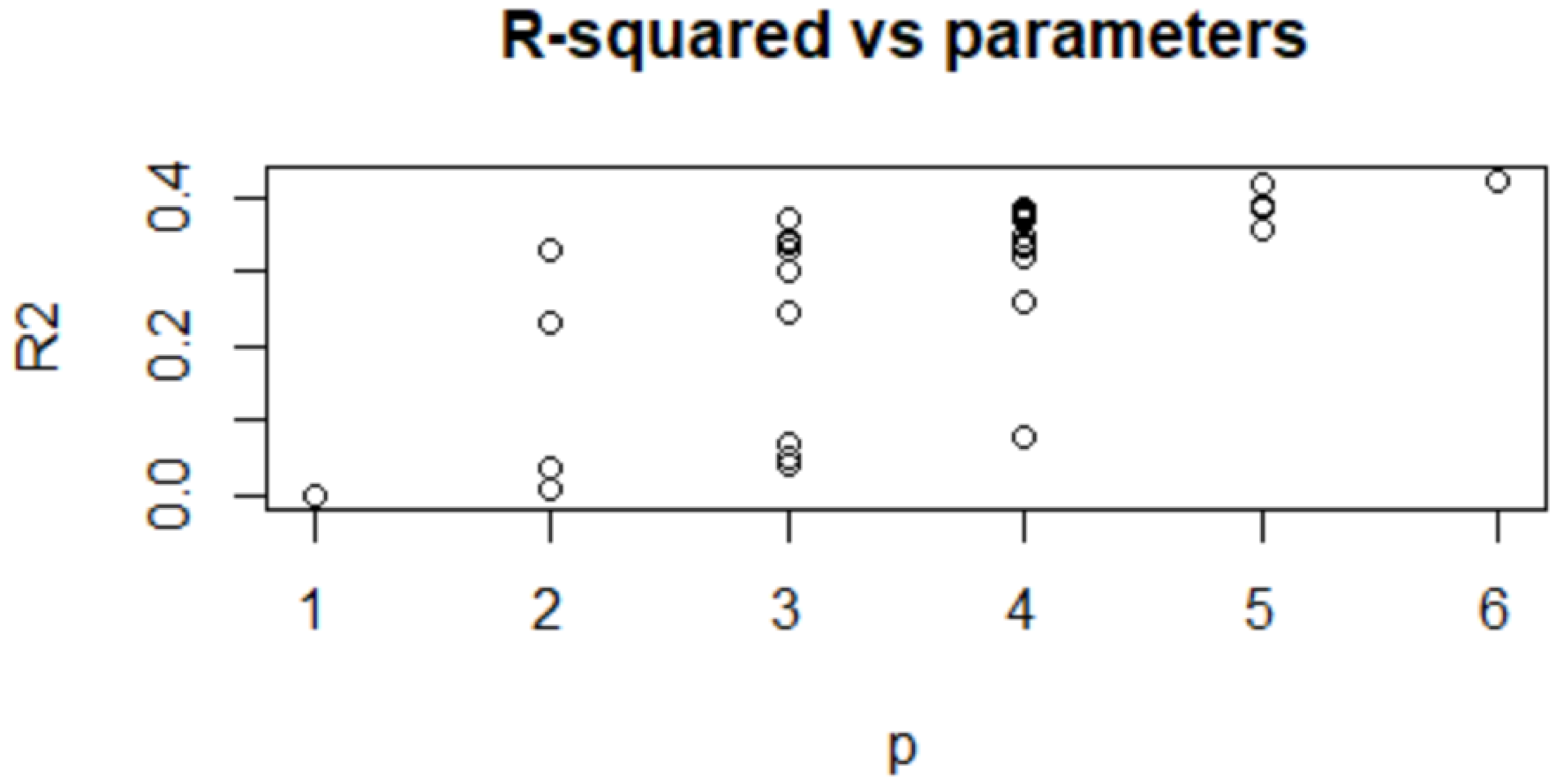
```
m245 = lm(y~x2+x4+x5)
anova(m245)
summary(m245)
```

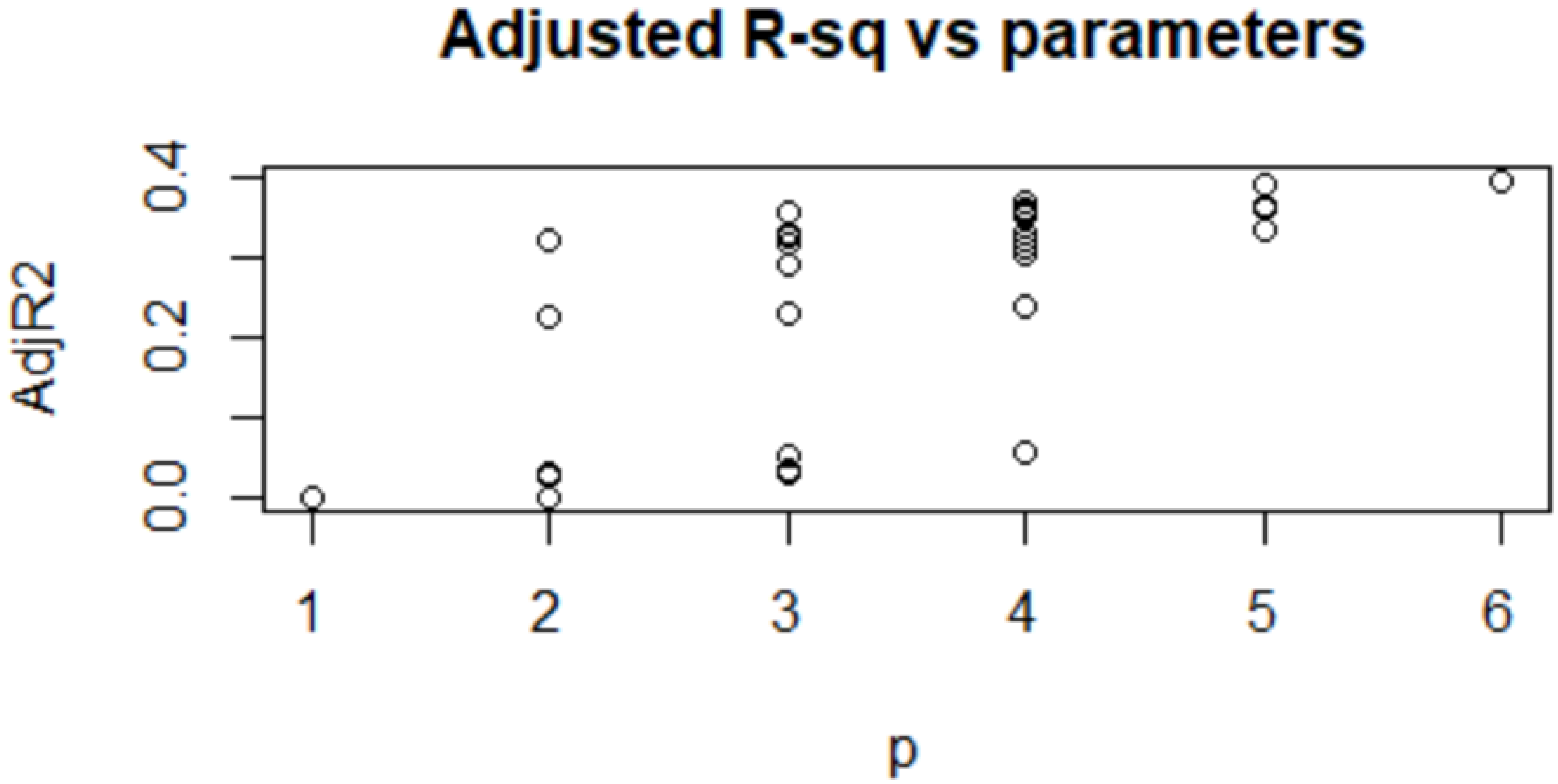
estimate σ^2 with MS_E from `anova(full)`

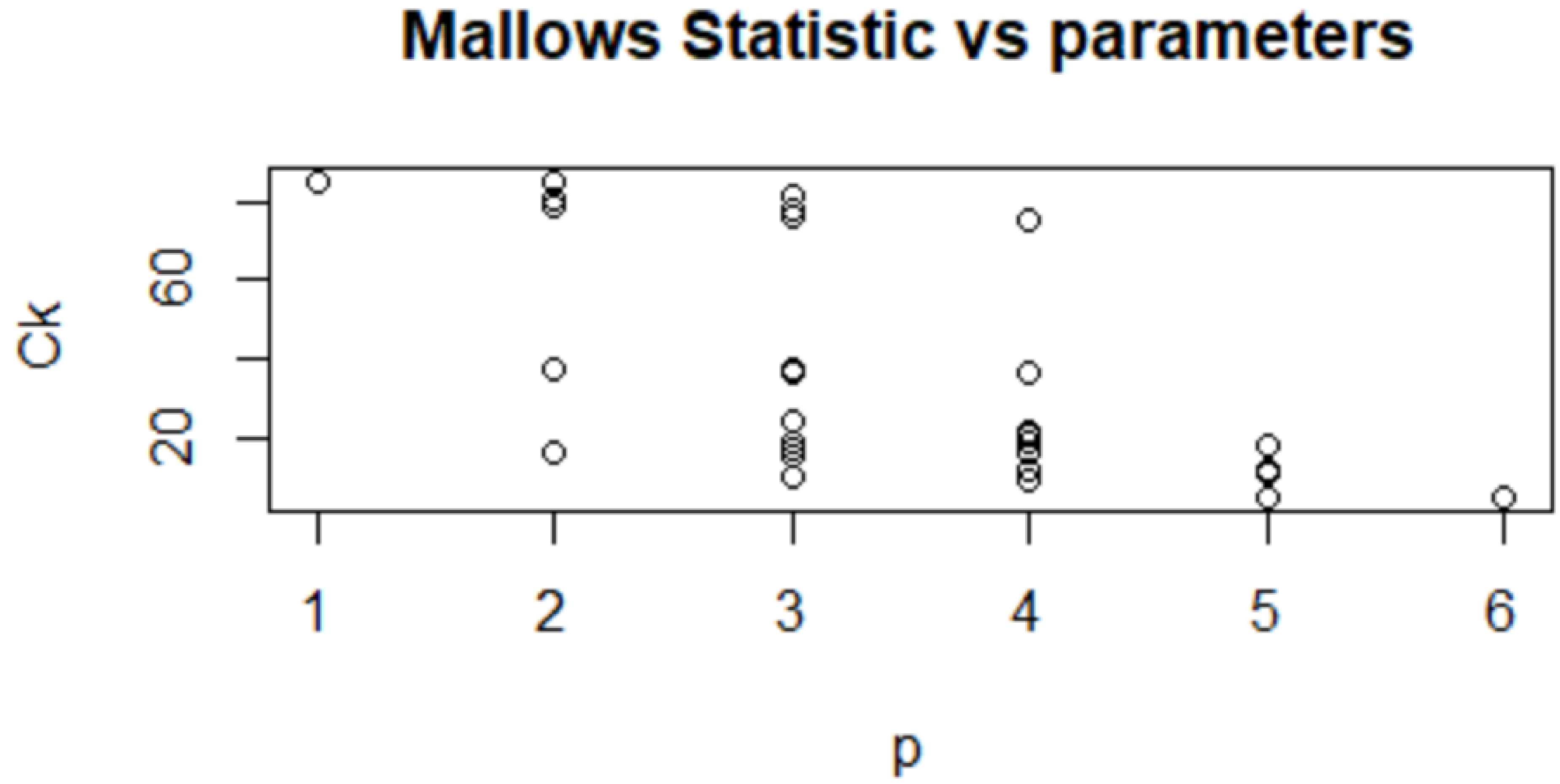
use $C_k = \frac{SS_E^{m(k)}}{\sigma^2} + 2k - 130$

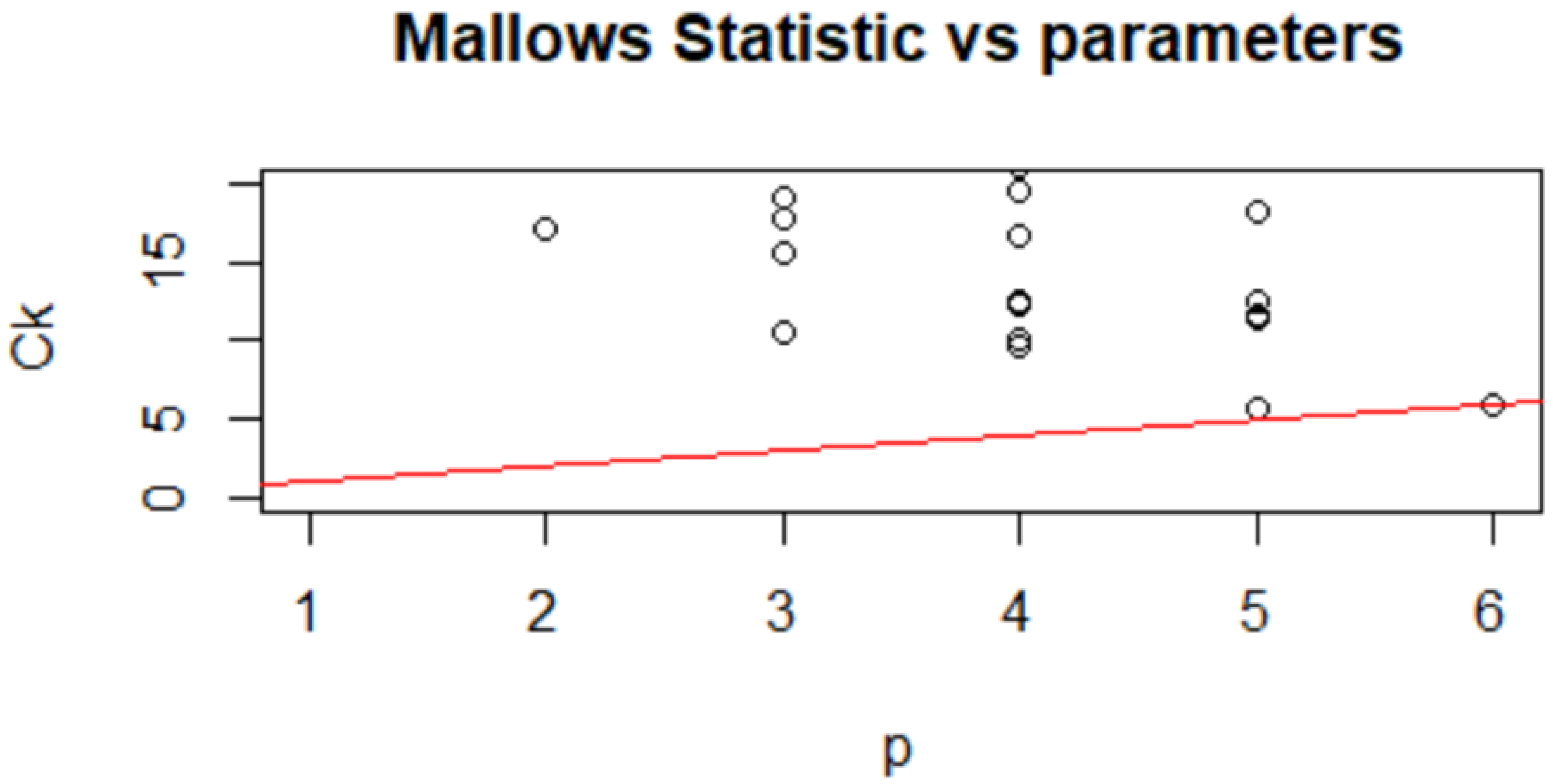



```
plot(p, MS_E, main = "Mean Square Residuals vs parameters", ylim = c(1,4))
```


R-squared vs parameters
R2
0.0
0.2
0.4
1
2
3
4
5
6
p


Adjusted R-sq vs parameters
AdjR2
0.0
0.2
0.4
1
2
3
4
5
6
p


Mallows Statistic vs parameters
Ck
20
60
1
2
3
4
5
6
p




```
> plot(p, Ck, main = "Mallows Statistic vs parameters", ylim = c(0,20))
> abline(0,1, col = "red")
```

Some conclusions

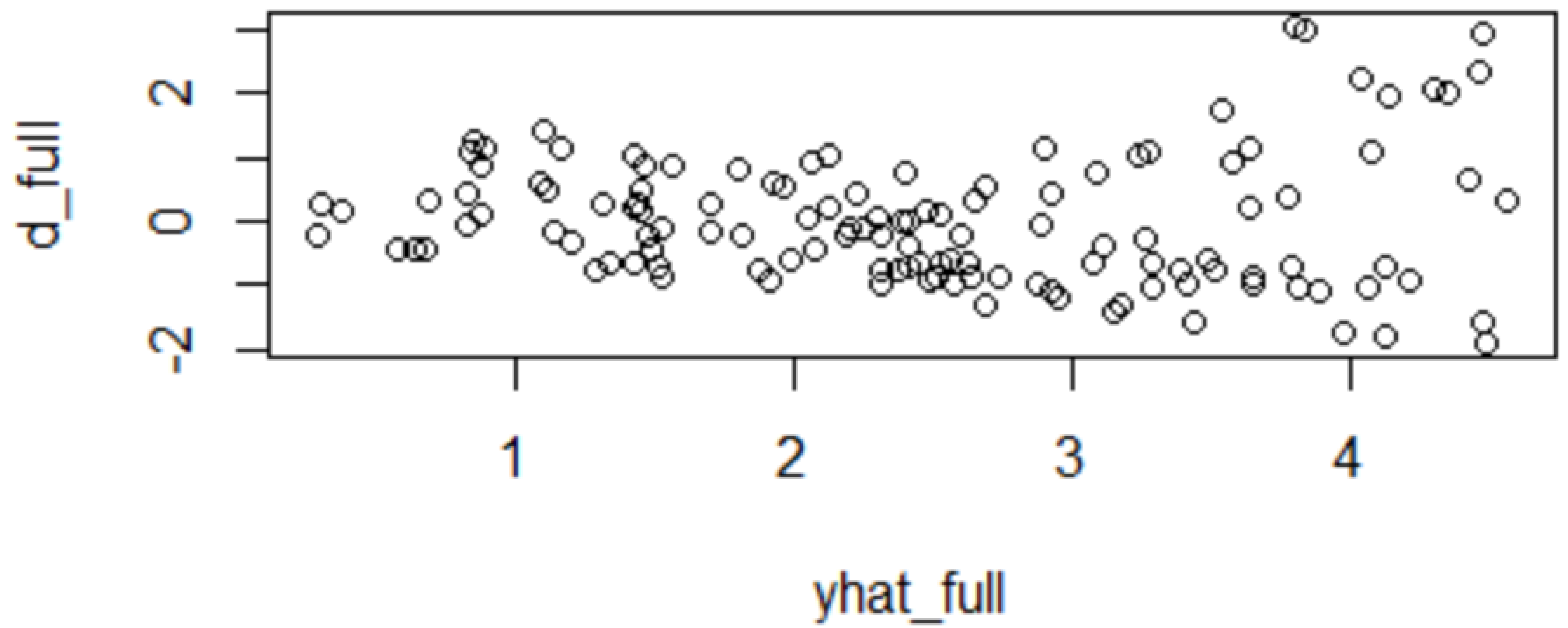
- the full model has lowest MS_E and highest Adjusted R^2
- model m3456 has lowest C_k and close to $C_k = k = 5$
- this model has 2nd lowest MS_E and 2nd highest Adjusted R^2
- a lot of the progress in reducing MS_E can be achieved through the best simple linear regression model m6 (the FTSE100 variable)
- none of these models has a very good R^2 (maximum 42%)

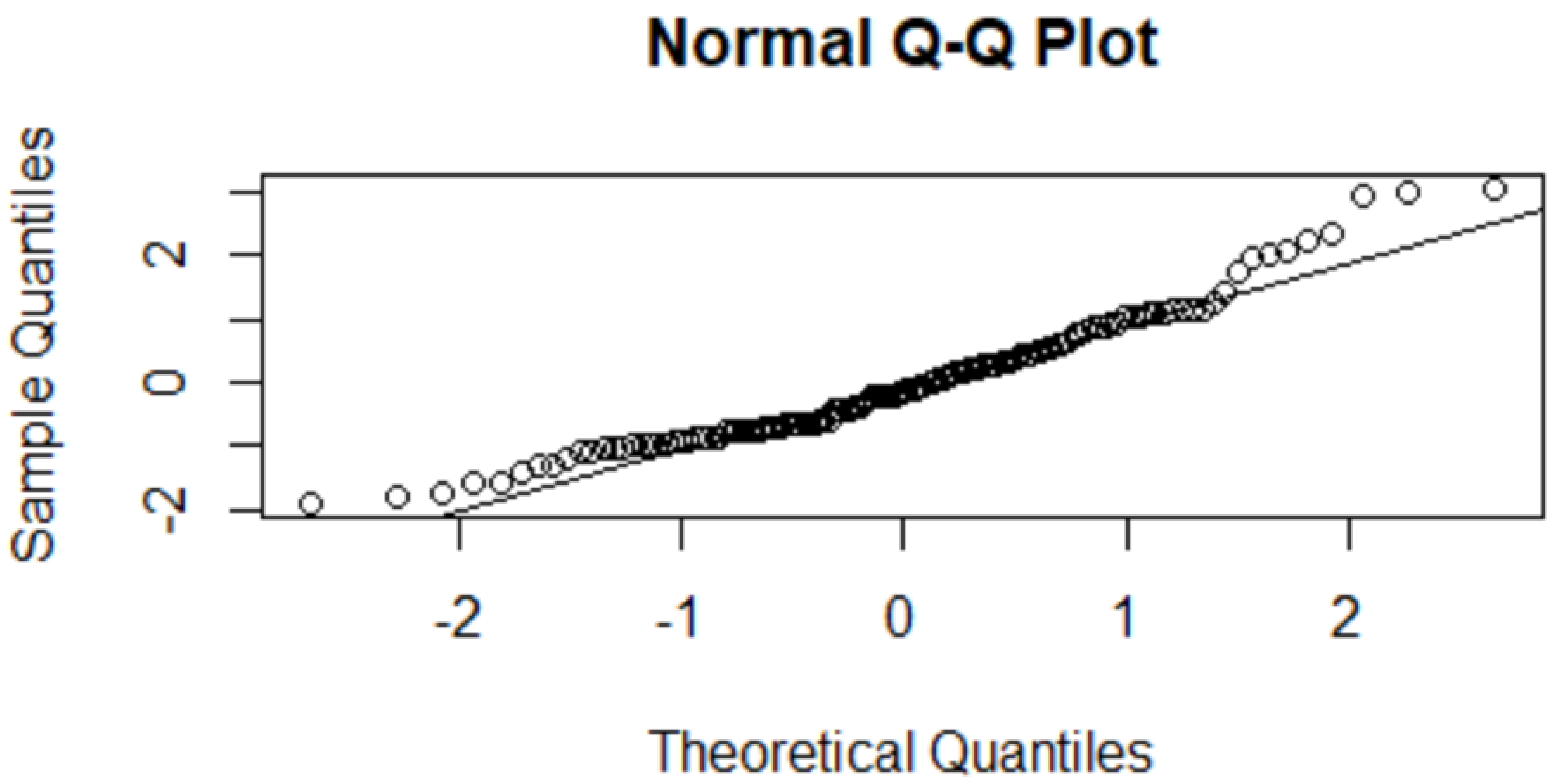
Further investigations

It seems that we should investigate the full model further

```
> full = lm(y~x2+x3+x4+x5+x6)
> d_full = rstandard(full)
> yhat_full = fitted(full)
> plot(yhat_full, d_full, main = "Std Residuals vs
Fitted, full model p=6")
> qqnorm(d_full)
> qqline(d_full)
```

Std Residuals vs Fitted, full model p=6




Normal Q-Q Plot
Sample Quantiles
-2
0
2
-2
-1
0
1
2
Theoretical Quantiles

Residual plot conclusions

- We do not have a constant variance
- There are reasons to question the Normal distribution assumption

```
> shapiro.test(d_full)

     Shapiro-Wilk normality test

data:  d_full
W = 0.95416, p-value = 0.0002547
```

We should investigate transforming the response variable

Linear model of $\sqrt{CPI}$

A large number of transformations of *y* are possible

Of the straightforward ones, $\sqrt{y}$ is the most promising

```
> y2 = sqrt(y)
> transform_model = lm(y2~x2+x3+x4+x5+x6)
> summary(transform_model)
```

Model output after transforming *y*

```
lm(formula = y2 ~ x2 + x3 + x4 + x5 + x6)
```

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	1.700e+00	4.381e-01	3.881
x2	2.154e-02	8.704e-03	2.474
x3	1.533e-01	3.471e-02	4.418
x4	-7.093e-02	1.689e-02	-4.199
x5	-6.248e-02	1.438e-02	-4.344
x6	-1.138e-04	3.925e-05	-2.901

	Pr(>|t|)	
(Intercept)	0.000169	***
x2	0.014709	*
x3	2.16e-05	***
x4	5.11e-05	***
x5	2.90e-05	***
x6	0.004414	**

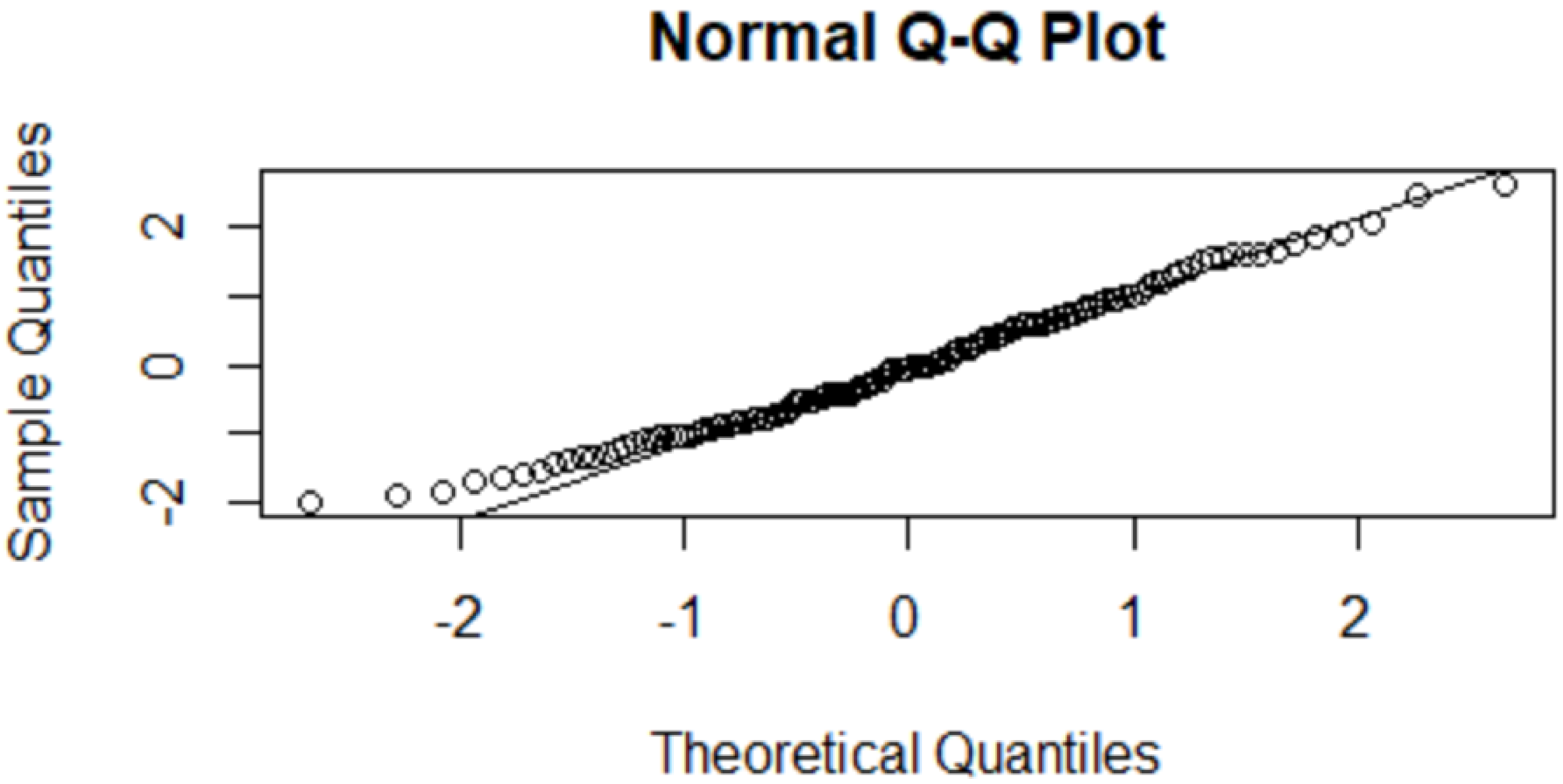
```
Multiple R-squared:  0.4825,  Adjusted R-squared:  0.4615
F-statistic: 22.94 on 5 and 123 DF,  p-value: 3.268e-16
```

Effect of the transformation of *y*

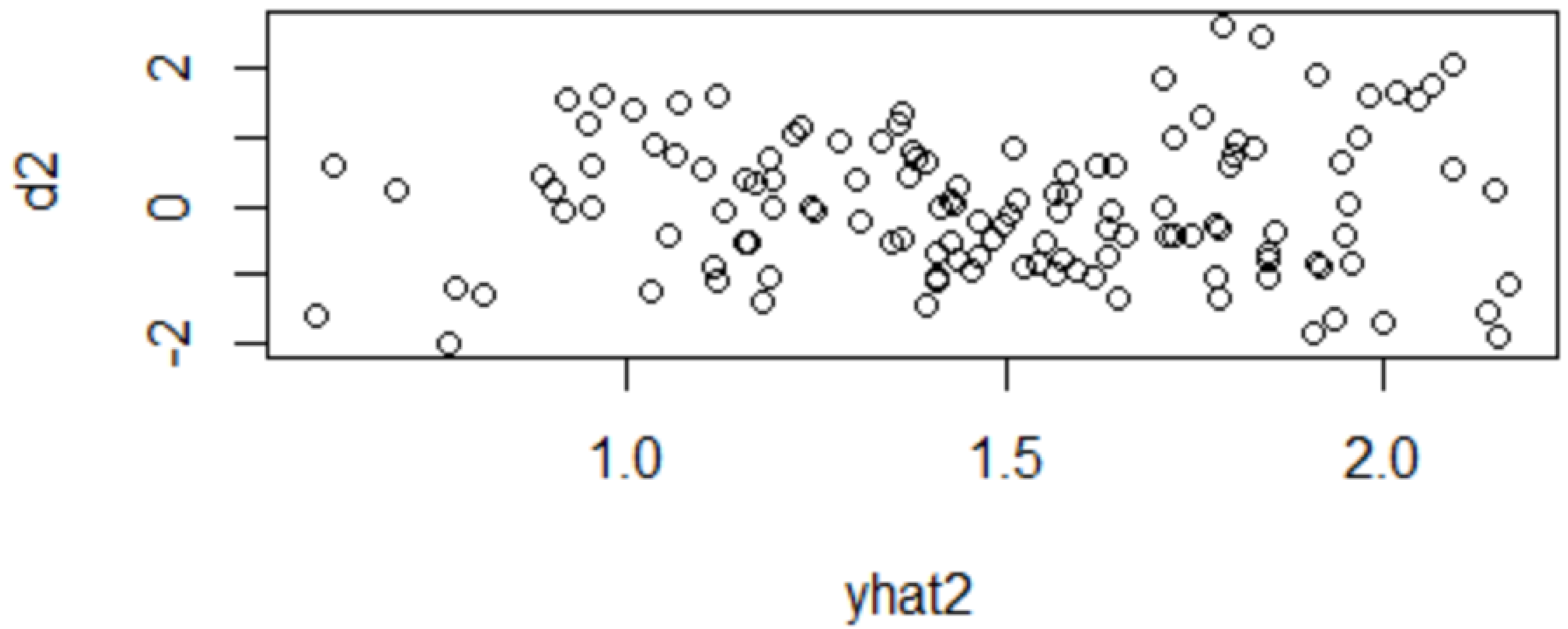
Square root transformation

- improves R-sq a little (42% to 48%)
- improves the nature of the residuals considerably

```
> d2 = rstandard(transform_model)
> yhat2 = fitted(transform_model)
> shapiro.test(d2)
      Shapiro-Wilk normality test
data:  d2
W = 0.98582, p-value = 0.2012
```


Normal Q-Q Plot
Sample Quantiles
-2
0
2
-2
-1
0
1
2
Theoretical Quantiles

Std Res vs Fitted, transformed model



More work still needed

- Missing explanatory variables
  - Exchange Rate
  - Industrial output
  - Consumer confidence
  - Commodities
  - Housing
- Relationships might not be linear
- Was always unlikely that CPI inflation would be straightforward to model