# VARIABLE SELECTION

DATA FOLKZ
CATAPULT DATA LEADERS

# VARIABLE SELECTION

- ***To describe some techniques for selecting the explanatory variables for a regression***
- ***To describe the consequences of making an incorrect choice***
- ***To apply these techniques to an example***

DATA FOLKZ
CATAPULT DATA LEADERS

# VARIABLE SELECTION

- Often there are several (perhaps a large number) of potential explanatory variables available to build a regression model. Which ones should we use?

- We could, of course, use them all. However, sometimes this turns out to be not such a good idea.

# OVERFITTING

- If we put too many variables in the model, including some unrelated to the response, we are ***overfitting***.
  Consequences are:
  - Fitted model is not good for prediction of new data – prediction error is underestimated
  - Model is too elaborate, models "noise" that will not be the same for new data
  - Variances of regression coefficients inflated

DATA FOLKZ
CATAPULT DATA LEADERS

# UNDERFITTING

- If we put too few variables in the model, leaving out variables that could help explain the response, we are ***underfitting***. Consequences:
  - Fitted model is not good for prediction of new data – prediction is biased
  - Regression coefficients are biased
  - Estimate of error variance is too large

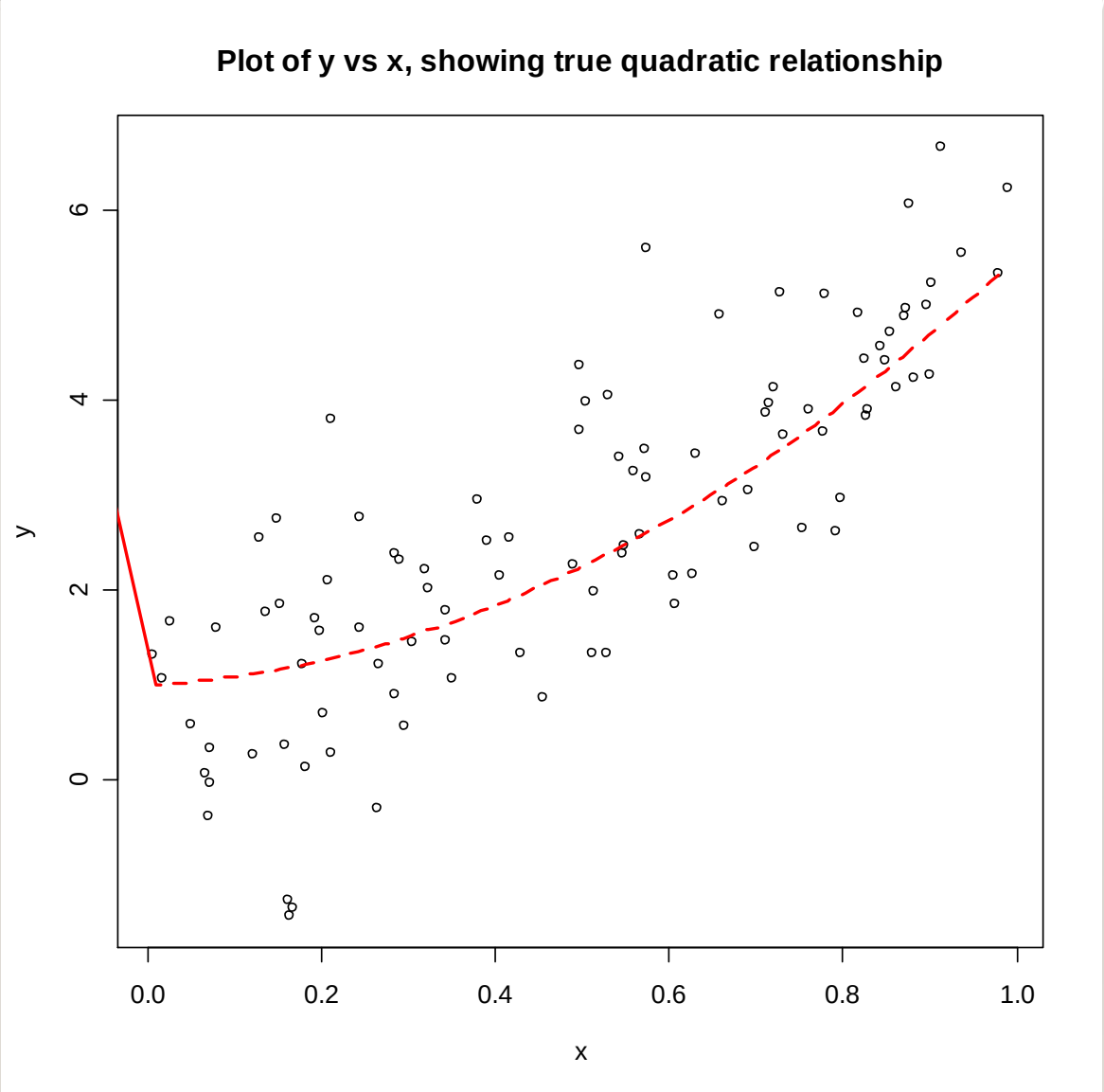
# EXAMPLE

- Suppose we have some data which follow a quadratic model

$$Y = 1 + 0.5\,x + 4\,x^2 + N(0,1)$$

where the x's are uniform on [0,1]

The next slide shows the data, with the true regression shown as a dotted line.


Plot of y vs x, showing true quadratic relationship
y
x
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0

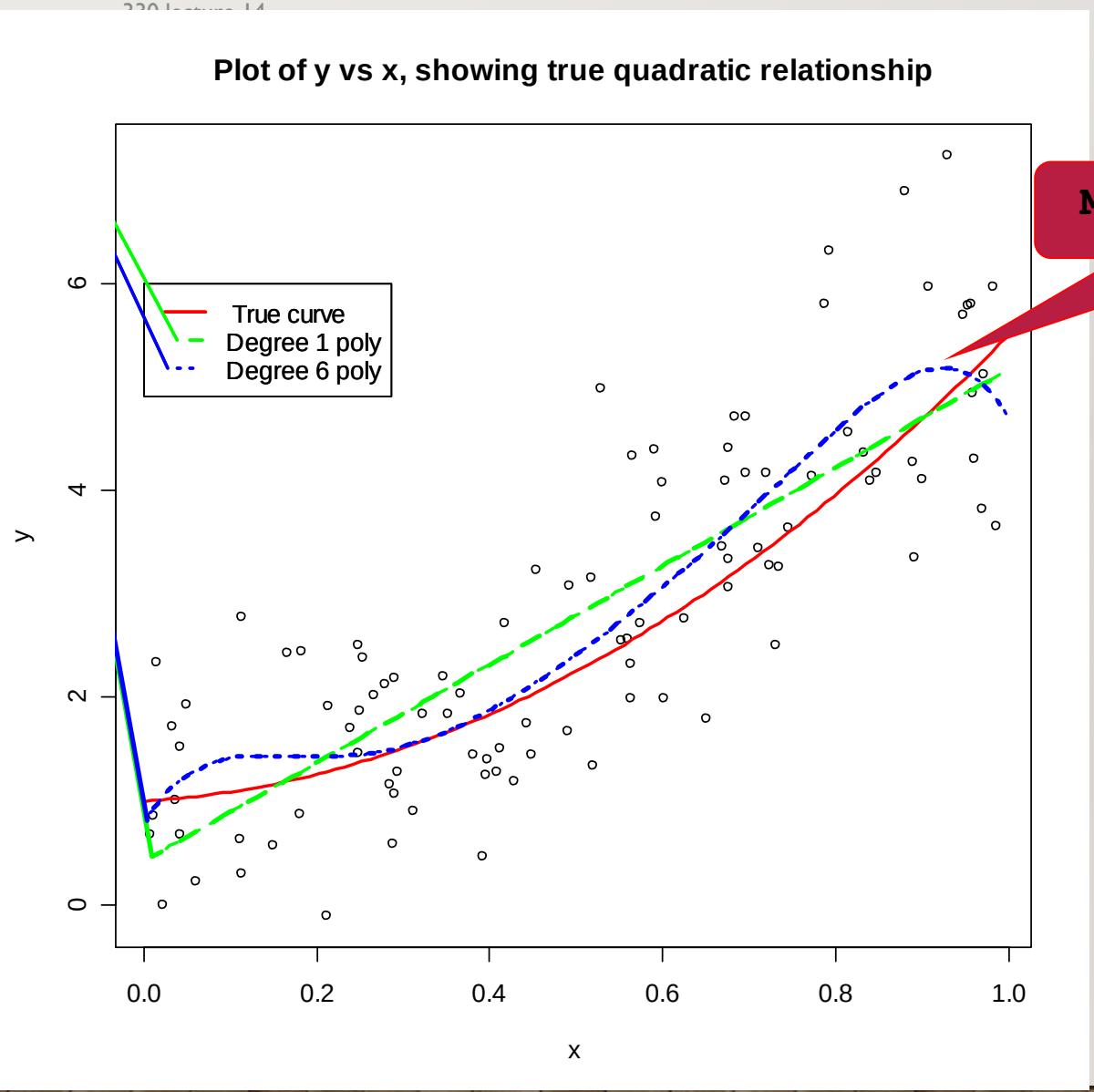
# 8 UNDER/OVER FITTING

- Suppose we fit a straight line. This is underfitting, since we are not fitting the squared term. The fitted line (in green) is shown on the next slide.

- Alternatively, we could fit a 6-degree polynomial. This is overfitting, since there are unnecessary terms in $x^3$, $x^4$, $x^5$ and $x^6$. The fitted polynomial is shown in blue on the next slide. Fit using

```
lm(y~poly(x,6))
```

DATA FOLKZ
CATAPULT DATA LEADERS




Modelling noise!

# 10 POINTS TO NOTE

- Straight line is biased: can't capture the curvature in the true regression
- 6-degree line :too variable, attracted to the errors which would be different for a new set of data
- Moral: For good models we need to choose variables wisely to avoid overfitting and underfitting. This is called ***variable selection***

# 11 USES OF REGRESSION

Two main uses

1. To explain the role(s) of the explanatory variables in influencing the response
2. To construct a prediction equation for predicting the response

Consequences of over/under fitting are different in each case

# USING REGRESSION FOR EXPLANATION

- Consider the example of heart disease (D), and two risk factors, alcohol (A) and smoking (S).
- Studies have found an association between A and D ( a significant regression coef if we regress D on A)
- There is also an association between A and S.

# 13 EXPLANATION (2)

Possible explanations for the significant alcohol coefficient:

1. Alcohol consumption causes heart disease
2. Alcohol consumption does not cause heart disease but is associated with smoking that does.

# 14 EXPLANATION (3)

- To decide among these, we can fix S and see if A is related to D for fixed S. This is measured by the coefficient of A in the model including S. Leaving S out gives a biased estimate of the appropriate beta.
- Variables like S are called *confounders*, omitting them leads to misleading conclusions.
- Thus, underfitting is potentially more serious than overfitting when interpreting coefficients – see example in the next lecture

# 15 PREDICTION

- The situation is simpler when we are predicting. We choose the model that will give us the smallest prediction error. This is often not the full model.
- We will discuss methods for estimating the prediction error later in the lecture, and in the next lecture

# 16 VARIABLE SELECTION

- If we have k variables, and assuming a constant term in each model, there are $2^k$-1 possible subsets of variables, not counting the null model with no variables. How do we select a subset for our model?

- Two main approaches: All possible regressions (APR, this lecture) and stepwise methods (SWR, next lecture)

# ALL POSSIBLE REGRESSIONS

- For each subset of variables, define a criterion of "model goodness" which tries to balance over-fitting (model too complex) with under-fitting (model doesn't fit very well).
- Calculate the criterion for each of the $2^k$-1 models (subsets)
- Pick the best one according to the criterion.
- One difficulty: there are several possible criteria, and they don't always agree.

# POSSIBLE CRITERIA: $R^2$



- Since $R^2$ increases as we add more variables, picking the model with the biggest $R^2$ will always select the model with all the variables. This will often result in overfitting.
- However, $R^2$ is OK for choosing between models with the same number of variables.
- We need to modify $R^2$ to penalize overly complicated models. One way is to use the adjusted $R^2$ (p = number of coefficients in model)

$$\overline{R}_p^2 = 1 - \frac{(n-1)}{(n-p)}(1 - R_p^2)$$

# INTERPRETATION





- Suppose we have 2 models: model A with p-1 variables and model B with an additional q variables (we say A is a submodel of B)
- Then the adjusted $R^2$ is defined so that

$$\overline{R}_p^2 < \overline{R}_{p+q}^2 \text{ if and only if } F > 1$$

where F is the F statistic for testing that model A is adequate.

# RESIDUAL MEAN SQUARE (RMS)

- Recall the estimate of the error variance $\sigma^2$: estimated by $s^2$=RSS/(n-p), sometimes called the residual mean square (RMS)
- Choose model with the minimum RMS
- We can show that this is equivalent to choosing the model with the biggest adjusted $R^2$

# AIC AND BIC

- These are criteria that balance goodness of fit (as measured by RSS) against model complexity (as measured by the number of regression coefficients)

- AIC (Akaike Information Criterion) is, up to a constant depending on n , $AIC = n \log(RMS_p) + 2p$
- Alternative version is $AIC = RSS_p/RMS_{Full} + 2p$, equivalent to Cp

RMS = residual mean square

- BIC (Bayesian Information Criterion) is

$$RSS_p/RMS_{Full} + p \log(n)$$

- Small values = good model
- AIC tends to favour more complex models than BIC

# CRITERIA BASED ON PREDICTION ERROR

- Our final set of criteria use an estimate of prediction error to evaluate models

- They measure how well a model predicts *new* data

# ESTIMATING PREDICTION ERROR: CROSS-VALIDATION

- If we have plenty of data, we split the data into 2 parts
  - The "training set", used to fit the model and construct the predictor
  - The "test set", used to estimate the prediction error
- Test set error (=prediction error) estimated by

$$n^{-1} \sum_{\text{test set}} (y_i - \hat{y}_i)^2$$

**Predicted value using training set predictor with new data**

- Choose model with smallest prediction error
- NB: Using training set to estimate prediction error underestimates the error (old data)

# ESTIMATING PREDICTION ERROR: CROSS-VALIDATION (2)

- If we don't have plenty of data, we randomly split the data into 10 parts. One part acts as a test set, the rest as the training set. We compute the prediction error from the test set as before.
- Repeat another 9 times, using a different $10^{th}$ as the test set each time. Average the estimates to get a good estimate of prediction error
- Repeat for different "random splits"
- This is "10-fold cross-validation". Can do 5-fold, or n-fold, but 10-fold seems to be best.

# MALLOW'S CP: ESTIMATING PREDICTION ERROR

Suppose we have a model with p regression coefficients. "Mallows $C_p$" provides an estimate of how well the model predicts new data, and is given by

$$Cp = \frac{RSS_p}{RMS_{FULL}} + 2p - n$$

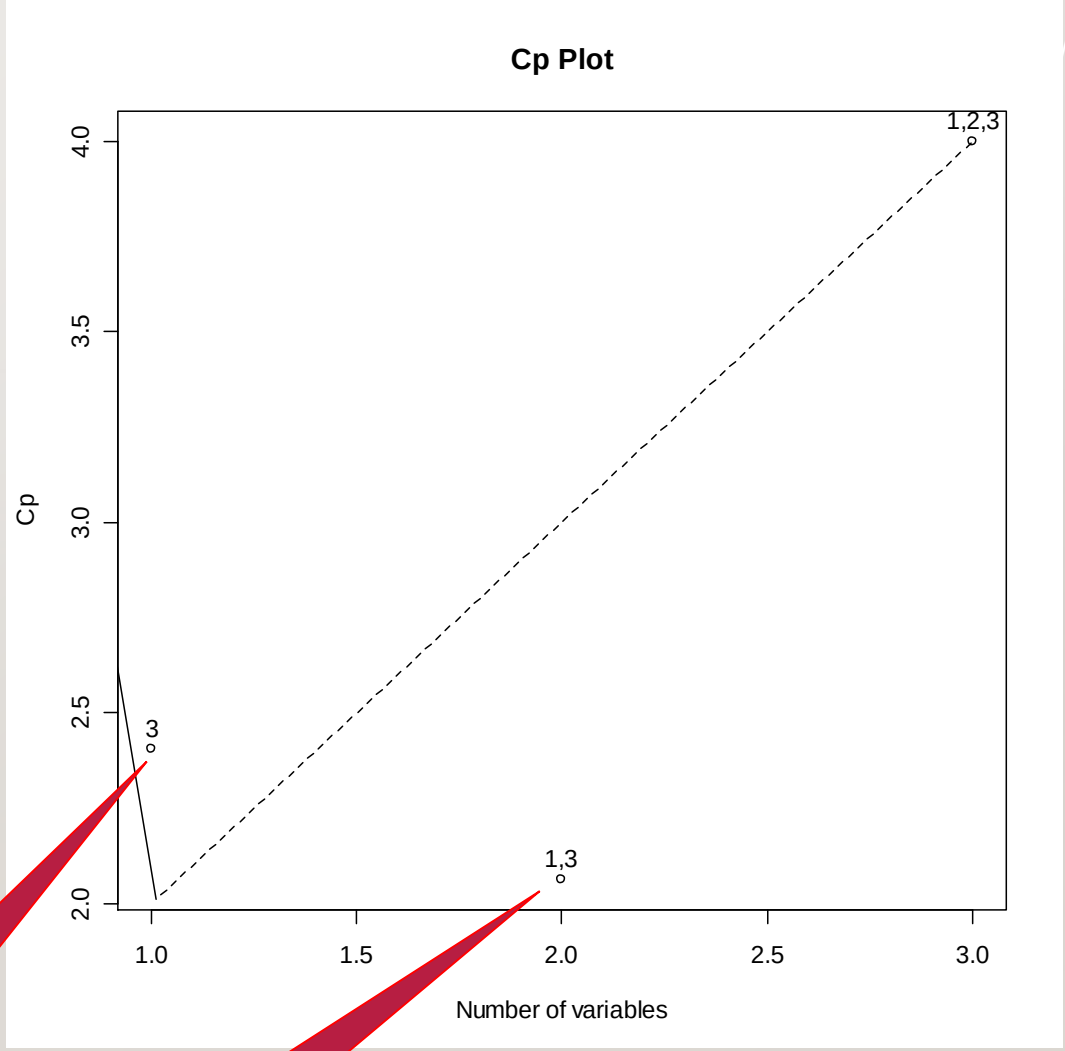
The subscript FULL refers to the "full model" with k variables. Small values of Cp with Cp about p are good. *Warning:* $C_{k+1}$=k+1 always, so don't take this as evidence that the full model is good unless all the other Cp's are bigger. Note similarity to AIC.

# EXAMPLE: THE FATTY ACID DATA

The R function allpossregs does the business: eg for the fatty acid data *NB This function requires the package "R330"*

```
> fatty.lm <- lm(ffa ~ age + skinfold + weight, data = fatty.df)
> library(R330)
> allpossregs(ffa ~ age + skinfold + weight, data = fatty.df)
```

| | rssp | sigma2 | adjRsq | Cp | AIC | BIC | CV | age | weight | skinfold |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.910 | 0.051 | 0.380 | 2.406 | 22.406 | 24.397 | 0.114 | 0 | 1 | 0 |
| 2 | 0.794 | 0.047 | 0.427 | 2.062 | 22.062 | 25.049 | 0.107 | 1 | 1 | 0 |
| 3 | 0.791 | 0.049 | 0.394 | 4.000 | 24.000 | 27.983 | 0.117 | 1 | 1 | 1 |




Good model

Good model

# EXAMPLE: THE EVAPORATION DATA

- This was discussed in Tutorial 2: the variables are
  - **evap:** the amount of moisture evaporating from the soil in the 24 hour period (response)
  - **maxst:** maximum soil temperature over the 24 hour period
  - **minst:** minimum soil temperature over the 24 hour period
  - **avst:** average soil temperature over the 24 hour period
  - **maxat:** maximum air temperature over the 24 hour period
  - **minat**: minimum air temperature over the 24 hour period
  - **avat:** average air temperature over the 24 hour period
  - **maxh:** maximum humidity over the 24 hour period
  - **minh:** minimum humidity over the 24 hour period
  - **avh:** average humidity over the 24 hour period
  - **wind:** average wind speed over the 24 hour period.

# 29 VARIABLE SELECTION

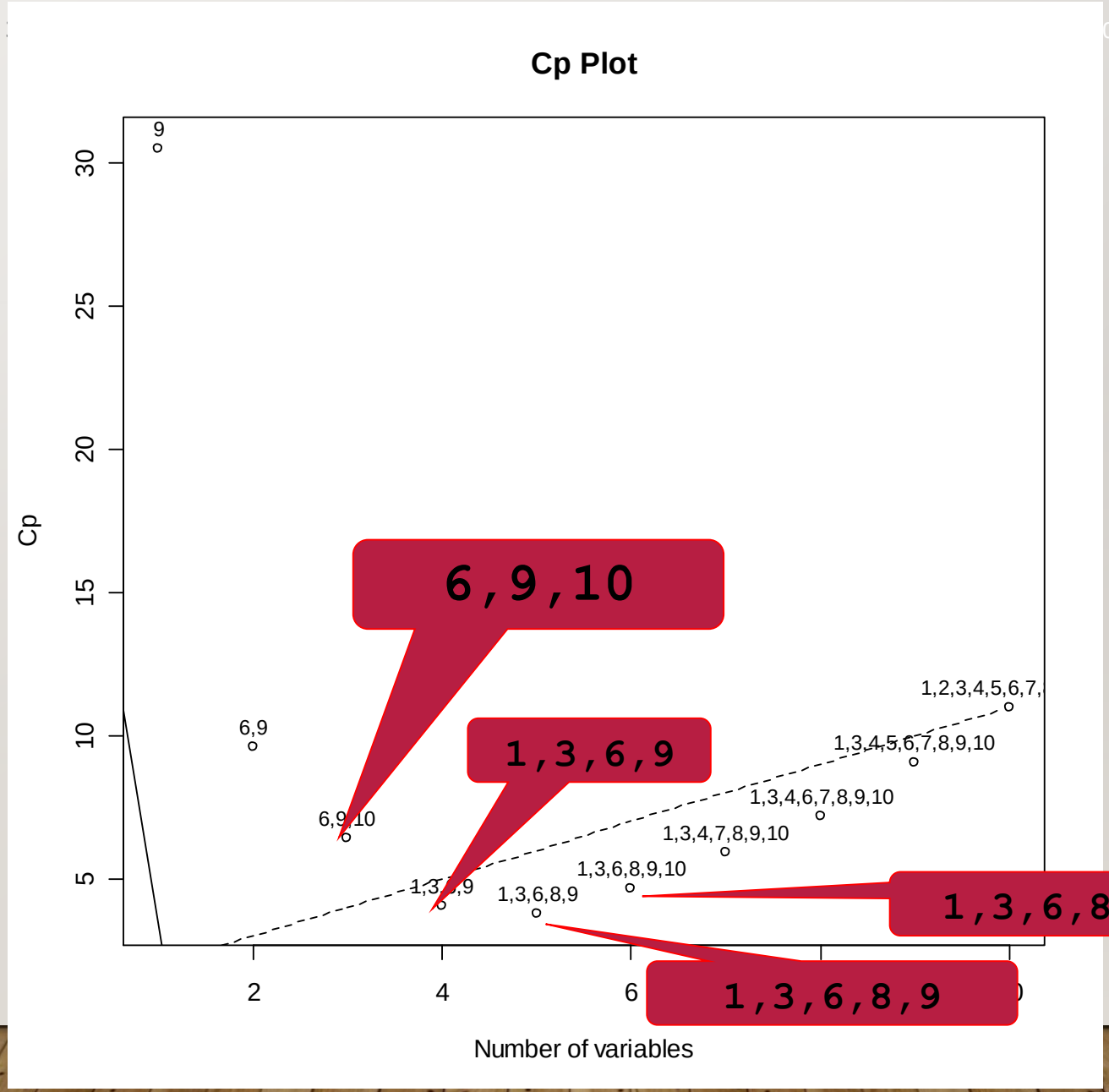
- There are strong relationships between the variables, so we probably don't need them all. We can perform an all possible regressions analysis using the code

```
evap.df = read.table(
"http://www.stat.auckland.ac.nz/
   ~lee/330/datasets.dir/evap.txt",
   header=TRUE)
evap.lm = lm(evap~.,data=evap.df)
library(R330)
allpossregs(evap~.,data=evap.df)
```

```
Call:
lm(formula = evap ~ ., data = evap.df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -54.074877 130.720826  -0.414  0.68164
avst          2.231782   1.003882   2.223  0.03276 *
minst         0.204854   1.104523   0.185  0.85393
maxst        -0.742580   0.349609  -2.124  0.04081 *
avat          0.501055   0.568964   0.881  0.38452
minat         0.304126   0.788877   0.386  0.70219
maxat         0.092187   0.218054   0.423  0.67505
avh           1.109858   1.133126   0.979  0.33407
minh          0.751405   0.487749   1.541  0.13242
maxh         -0.556292   0.161602  -3.442  0.00151 **
wind          0.008918   0.009167   0.973  0.33733
Residual standard error: 6.508 on 35 degrees of freedom
Multiple R-Squared: 0.8463,     Adjusted R-squared: 0.8023
F-statistic: 19.27 on 10 and 35 DF,  p-value: 2.073e-11
```

```
> library(R330)  # NB Load R330 library
> allpossregs(evap~., data=evap.df)
       rssp sigma2 adjRsq     Cp    AIC    BIC       CV
1  3071.255 69.801  0.674 30.519 76.519 80.177 308.052
2  2101.113 48.863  0.772  9.612 55.612 61.098 208.962
3  1879.949 44.761  0.791  6.390 52.390 59.705 191.622
4  1696.789 41.385  0.807  4.065 50.065 59.208 206.449
5  1599.138 39.978  0.813  3.759 49.759 60.731 223.113
6  1552.033 39.796  0.814  4.647 50.647 63.448 233.692
7  1521.227 40.032  0.813  5.920 51.920 66.549 260.577
8  1490.602 40.287  0.812  7.197 53.197 69.654 271.771
9  1483.733 41.215  0.808  9.034 55.034 73.321 302.781
10 1482.277 42.351  0.802 11.000 57.000 77.115 325.410
    avst minst maxst avat minat maxat avh minh maxh wind
1      0     0     0    0     0     0   0    0    1    0
2      0     0     0    0     0     1   0    0    1    0
3      0     0     0    0     0     1   0    0    1    1
4      1     0     1    0     0     1   0    0    1    0
5      1     0     1    0     0     1   0    1    1    0
6      1     0     1    0     0     1   0    1    1    1
7      1     0     1    1     0     0   1    1    1    1
8      1     0     1    1     0     1   1    1    1    1
9      1     0     1    1     1     1   1    1    1    1
10     1     1     1    1     1     1   1    1    1    1
```



```
> sub.lm = lm(evap~maxat + maxh + wind,data=evap.df)
> summary(sub.lm)
 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 123.901800  24.624411   5.032 9.60e-06 ***
maxat         0.222768   0.059113   3.769 0.000506 ***
maxh         -0.342915   0.042776  -8.016 5.31e-10 ***
wind          0.015998   0.007197   2.223 0.031664 *
---

Residual standard error: 6.69 on 42 degrees of freedom
Multiple R-squared: 0.805,      Adjusted R-squared:
0.7911
F-statistic:  57.8 on 3 and 42 DF,  p-value: 5.834e-15
```

Full model was 0.8463