Multiple Regression Part 2

STAT 315, 26/03

Question: what is the purpose of fitting a statistical model to data?

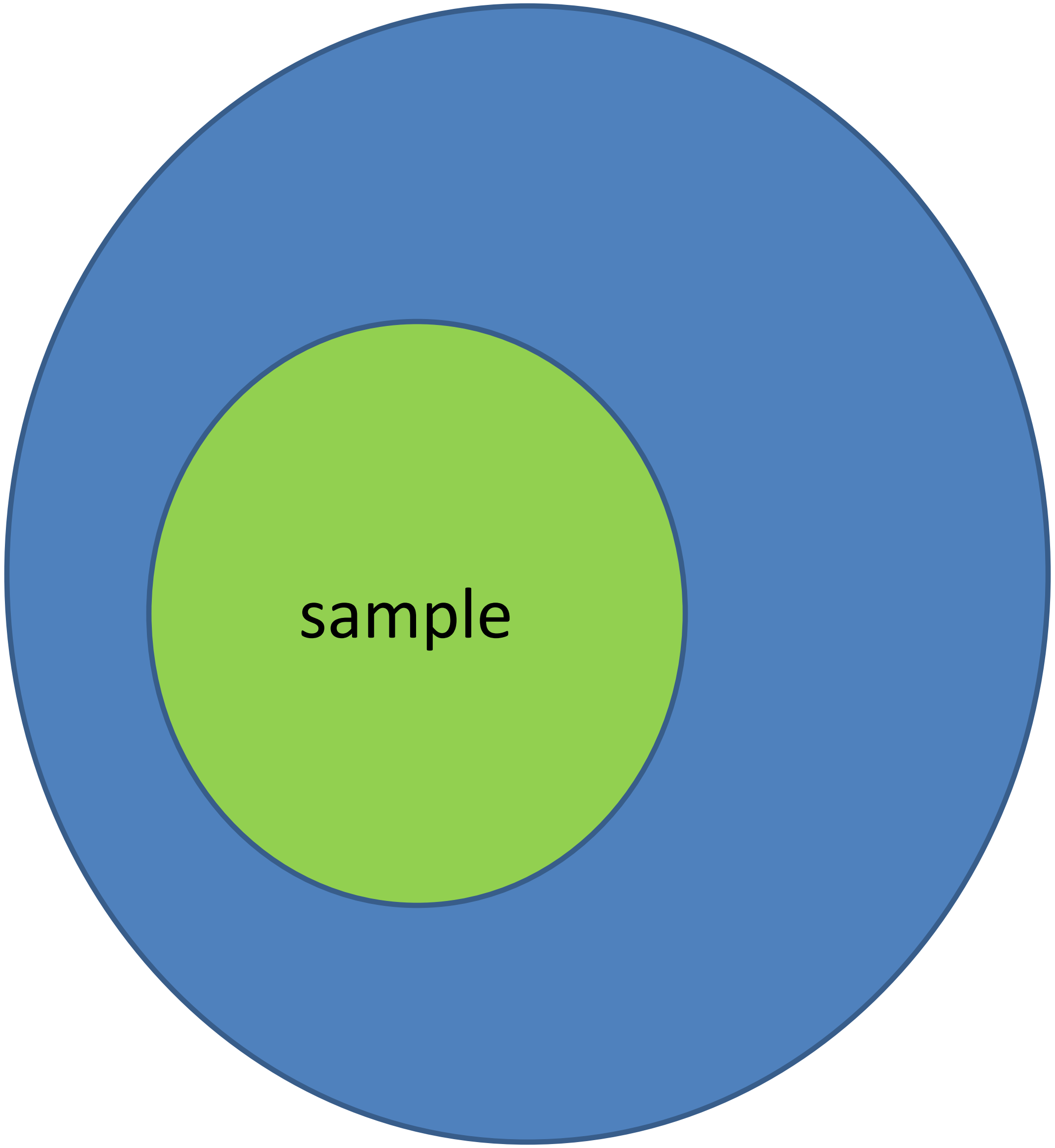
```
Call:
lm(formula = stack.loss ~ ., data = stackloss)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2377 -1.7117 -0.4551  2.3614  5.6978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.9197    11.8960  -3.356  0.00375 **
Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680   3.520  0.00263 **
Acid.Conc.   -0.1521     0.1563  -0.973  0.34405
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared:  0.9136,    Adjusted R-squared:  0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

How should we use this output? What is it telling us about stack loss?



We are really interested in generalising from a sample to a larger population. This is known as **inference** because we want to infer things about the population from the sample.

- The reason why we *need* p-values, confidence intervals and things is because we can't usually take another sample from the population. We just have to make do with the data we have.

- The reason why we need regression diagnostics is because our p-values and things will not be valid if the assumptions of regression are violated.

- "Valid" means that they won't accurately reflect what would happen if we got a different sample from the population that our sample came from.

- (If *only* there was some way to get more data from the same population!)

1. CRIM per capita crime rate by town
2. ZN proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS proportion of non-retail business acres per town
4. CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX nitric oxides concentration (parts per 10 million)
6. RM average number of rooms per dwelling
7. AGE proportion of owner-occupied units built prior to 1940
8. DIS weighted distances to five Boston employment centres
9. RAD index of accessibility to radial highways
10. TAX full-value property-tax rate per $10,000
11. PTRATIO pupil-teacher ratio by town
12. B 1000(Bk - 0.63)^2 where Bk is the proportion of African-Americans by town
13. LSTAT % lower status of the population
14. MEDV Median value of owner-occupied homes in $1000's

Boston housing example:

We want to predict MEDV given the other variables. Throwing all the variables into a regression looked a bit suspicious. How to improve?

Ockham’s Razor:

When two theories fit the facts, choose the simpler one.

Usually the simpler theory will have better predictive power on new data.

We need a way to measure the fit of a model with includes a penalty for models which are too complicated.

For regression, one way of doing this is using Mallows' C_p

$$C_p = \frac{RSS(model)}{MSE(full\ model)} - n + 2(p + 1)$$

penalty

By making C_p small, we can try to find a model which fits well but which has a small value of p.

For linear regression, C_p is the same as the AIC (Akaike Information Criterion).

```
proc reg data = boston;
  model MEDV = CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX
PTRATIO B LSTAT / vif selection=cp;
  run;
```

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	36.34115	5.06749	7.17	<.0001	0
CRIM	1	-0.10841	0.03278	-3.31	0.0010	1.78970
ZN	1	0.04584	0.01352	3.39	0.0008	2.23923
CHAS	1	2.71872	0.85424	3.18	0.0016	1.05982
NOX	1	-17.37602	3.53524	-4.92	<.0001	3.77801
RM	1	3.80158	0.40632	9.36	<.0001	1.83481
DIS	1	-1.49271	0.18573	-8.04	<.0001	3.44342
RAD	1	0.29961	0.06340	4.73	<.0001	6.86113
TAX	1	-0.01178	0.00337	-3.49	0.0005	7.27239
PTRATIO	1	-0.94652	0.12907	-7.33	<.0001	1.75768
B	1	0.00929	0.00267	3.47	0.0006	1.34156
LSTAT	1	-0.52255	0.04742	-11.02	<.0001	2.58198

INDUS and AGE were dropped. Eleven variables are left.

In R:

```
model <- lm(MEDV ~. ,data=boston)
model2 <- step(model)
```

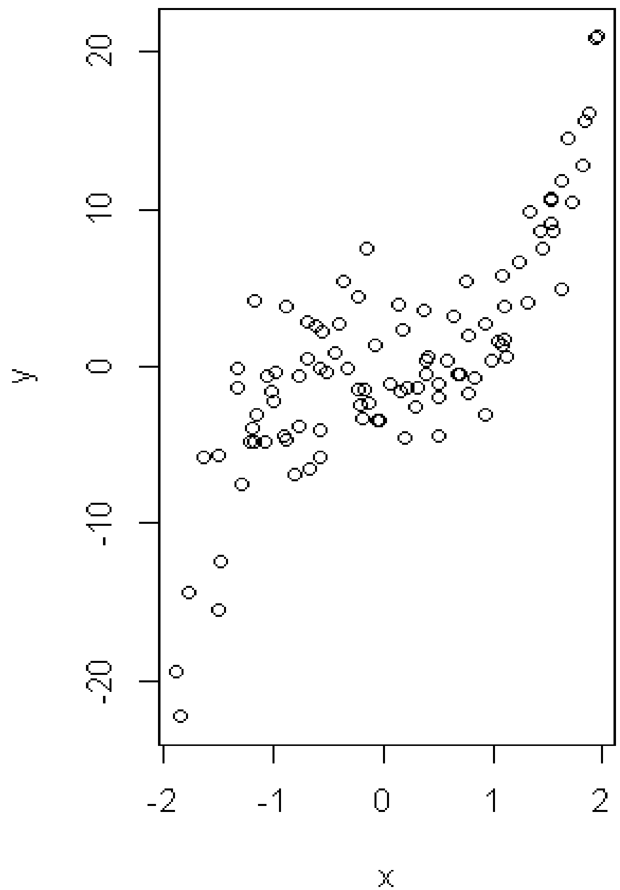
Note: the algorithm being followed is a **greedy algorithm**. At each stage the computer tries all possible ways of adding or removing a single variable, and selects the model with the smallest C_p. There is no guarantee that this will be the best possible value of C_p among all models!

Called **stepwise selection**

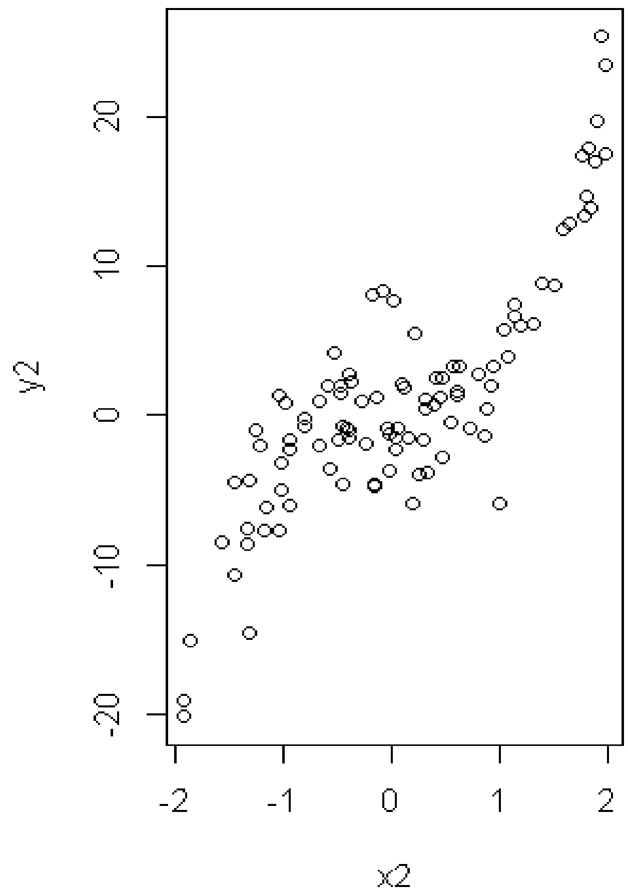
Other criteria for selecting a model include the AIC (popular) and the BIC (also popular, more conservative.) They tend to have the form

(Measure of goodness of fit) - (complexity penalty)

- Simpler models have high **bias** (they fit badly.)
- Complex models have high **variance** (they **over-fit**.)

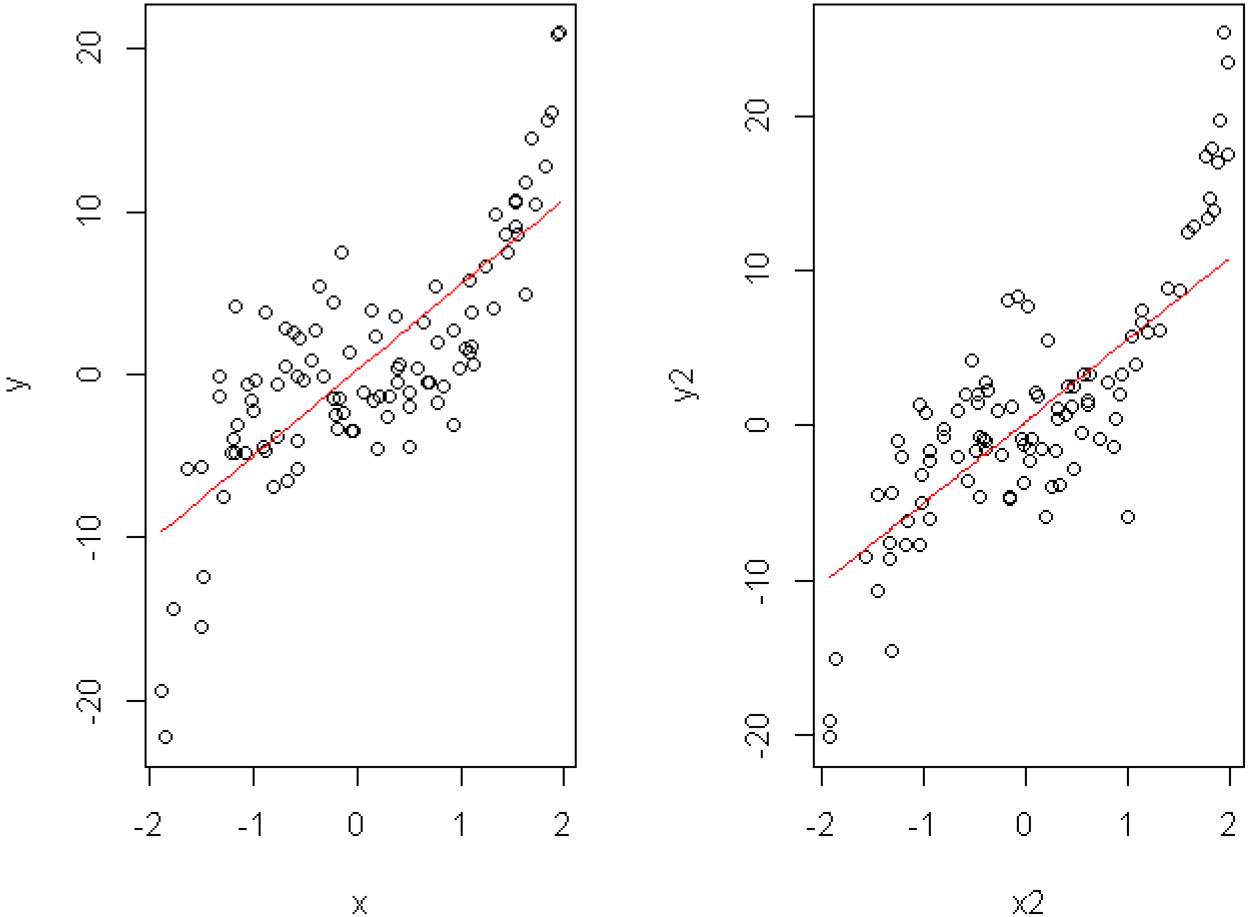


Training data

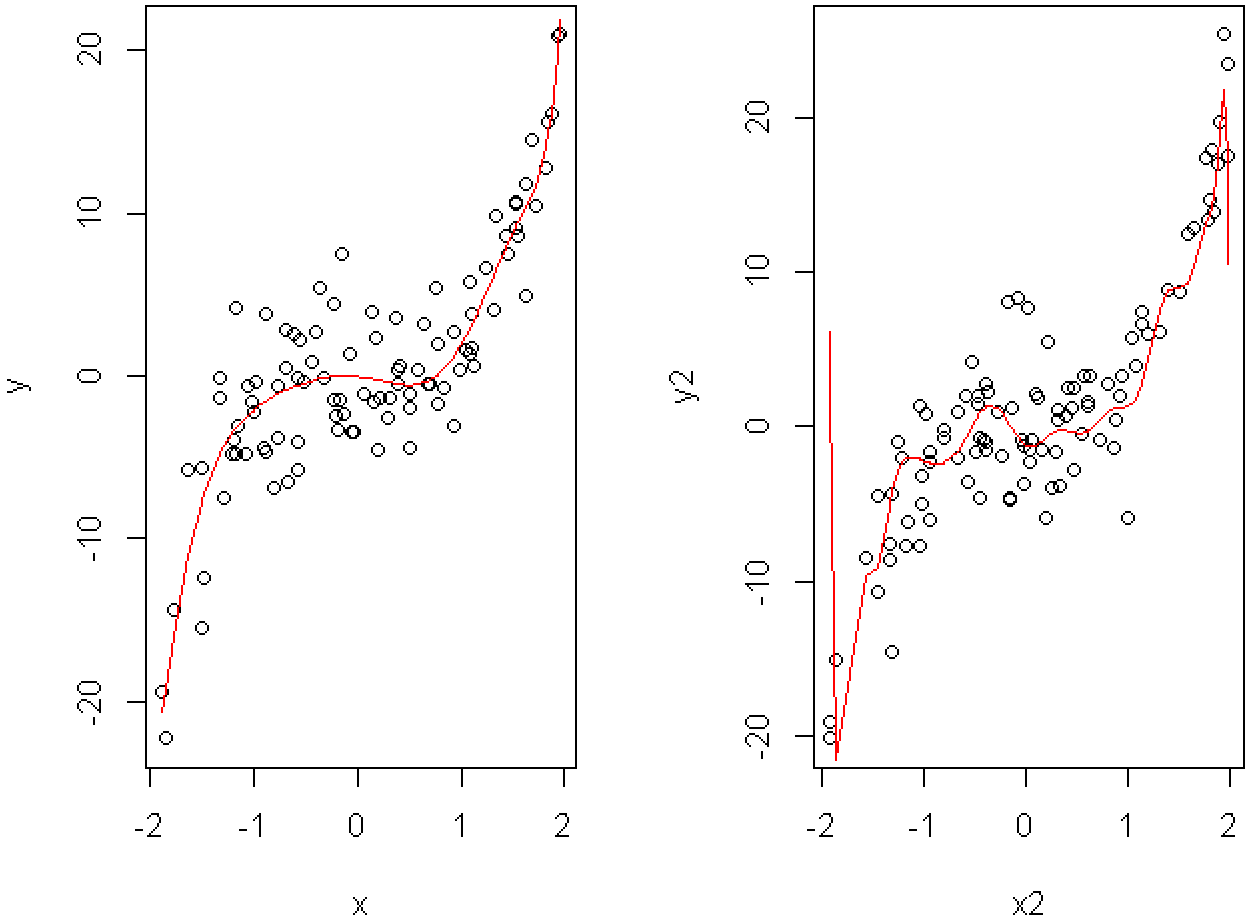


Test data

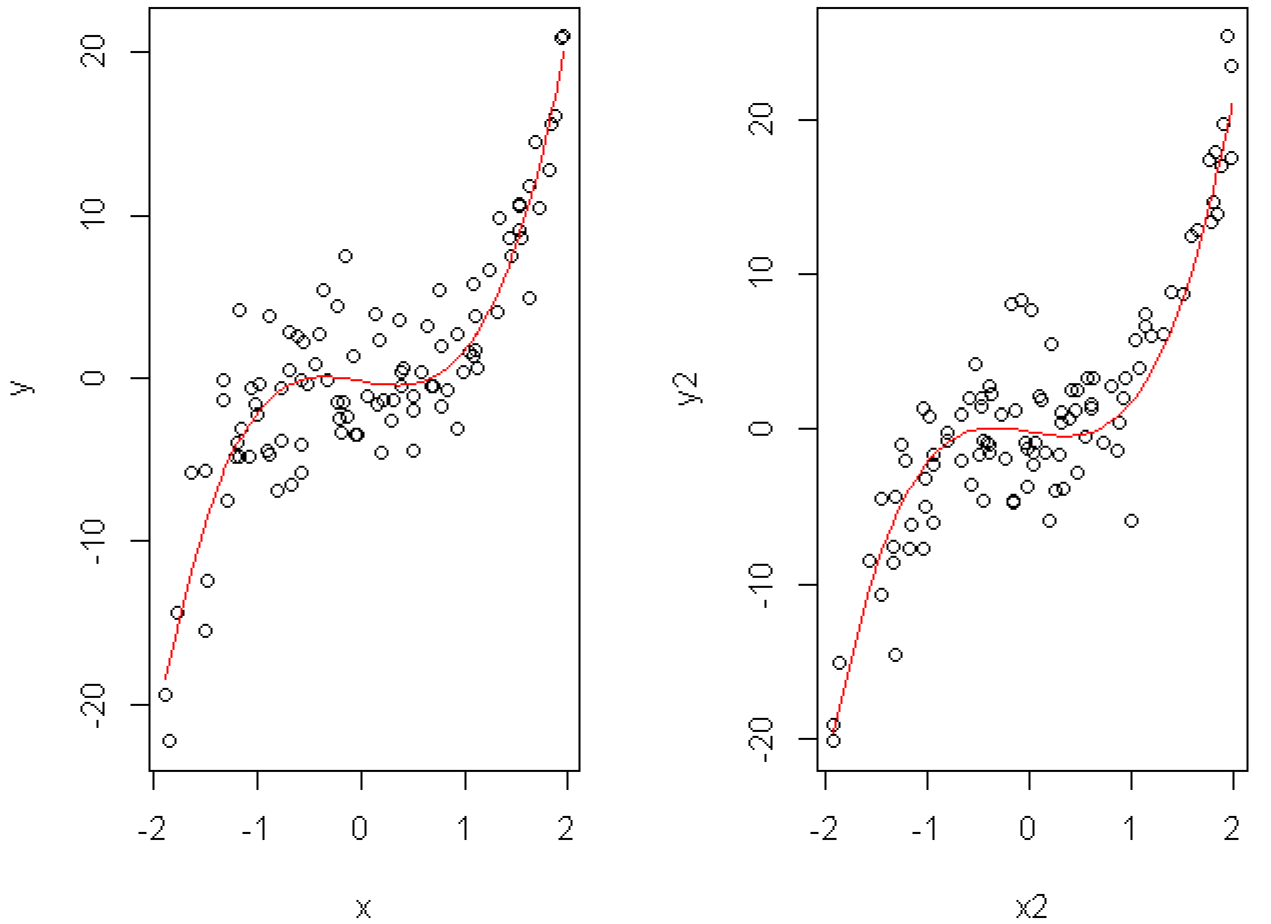
$$Y = 3x^3 - x + N(0, 3)$$



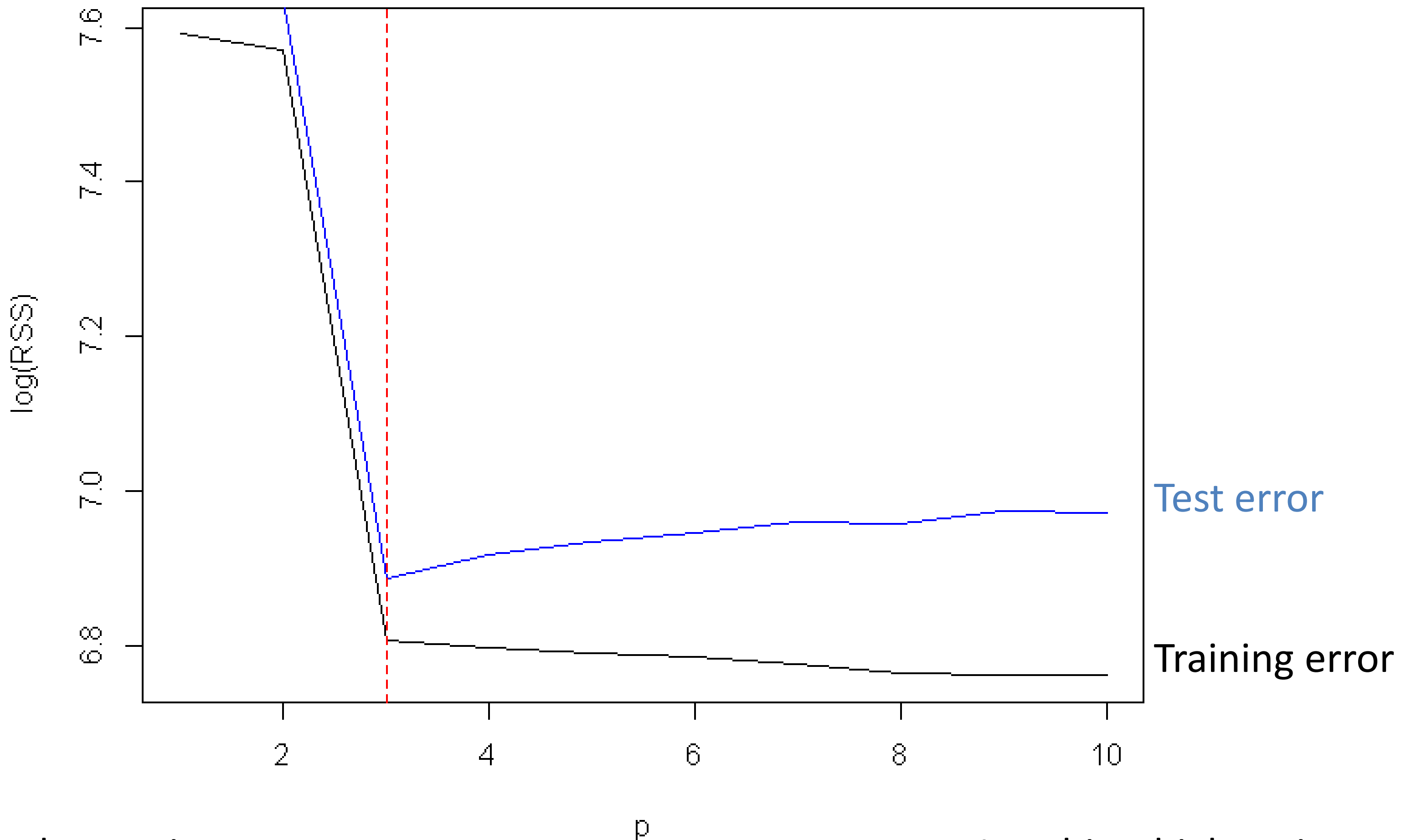
y
20
10
0
-10
-20
-2
-1
0
1
2
x
y2
20
10
0
-10
-20
-2
-1
0
1
2
x2



y
20
10
0
-10
-20
-2
-1
0
1
2
x
y2
20
10
0
-10
-20
-2
-1
0
1
2
x2


y
20
10
0
-10
-20
-2
-1
0
1
2
x
y2
20
10
0
-10
-20
-2
-1
0
1
2
x2

Bias-variance tradeoff



High bias, low variance

Low bias, high variance