The Problem of Model Selection

The Problem of Model Selection

A given data set can conceivably have been generated from uncountably many models. Identifying the true model is like finding a piece of hay in a haystack. Said another way, the model space is massive and the criterion for what constitutes the "best" model is ill-defined.

The Problem of Model Selection

Best strategy: Use domain knowledge to constrain the model space and/or build models that help you answer specific scientific questions.

Another common strategy:

1. Pick a criterion for "best".
2. Decide how to explore the model space.
3. Select "best" model in search area.

Tread Carefully!!! The second strategy can lead to myopic analysis, overconfidence, and wrong-headed conclusions.

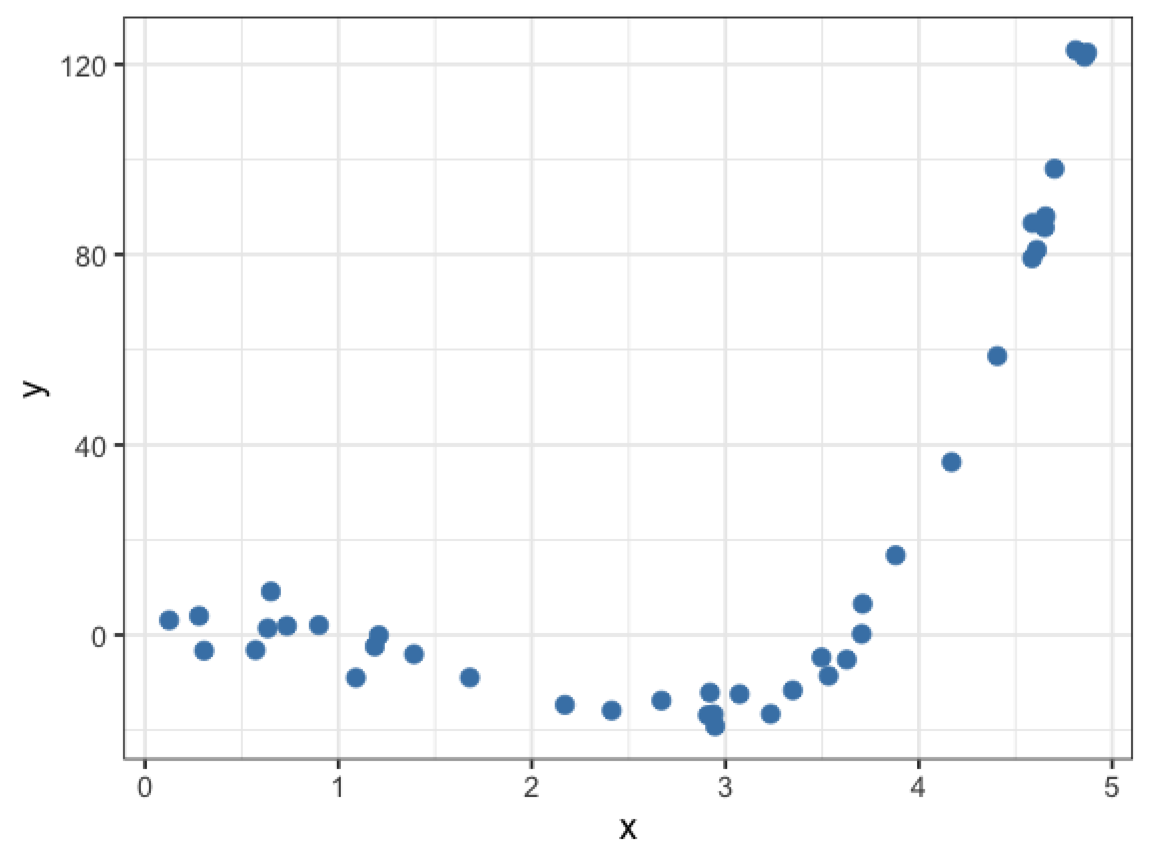
What do we mean by "best"?

While we'd like to find the "true" model, in practice we just hope we're doing a good job at:

1. Prediction
2. Description

Synthetic example

How smooth should our model be?

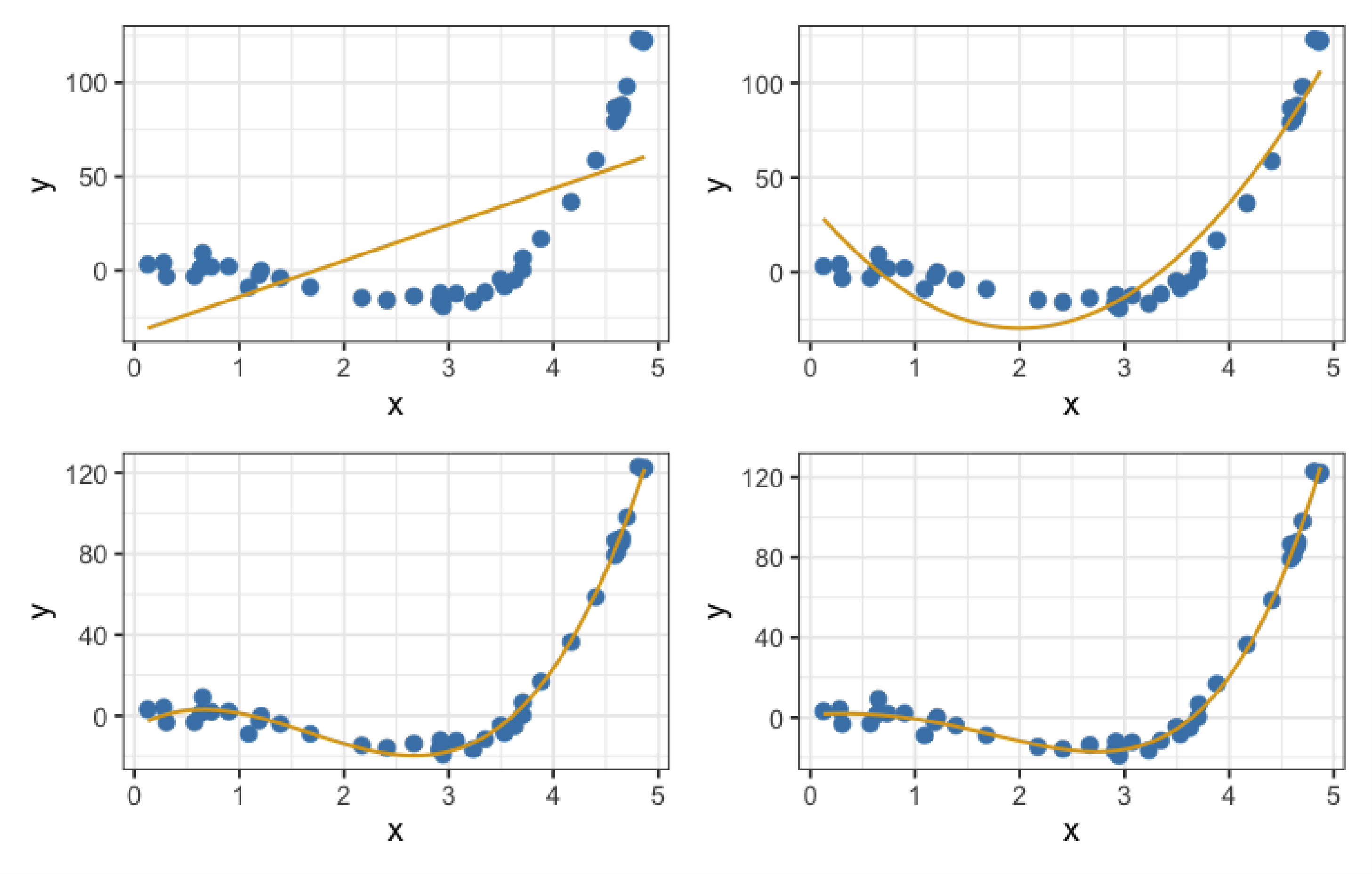


Four candidates

```
m1 <- lm(y ~ x)
m2 <- lm(y ~ x + I(x^2))
m3 <- lm(y ~ x + I(x^2) + I(x^3))
m4 <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4))
```

We can add *polynomial* terms to account for non-linear trends.

Four candidates



R^2

One way to quantify the explanatory power of a model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

This captures the proportion of variability in the y explained by our regression model.

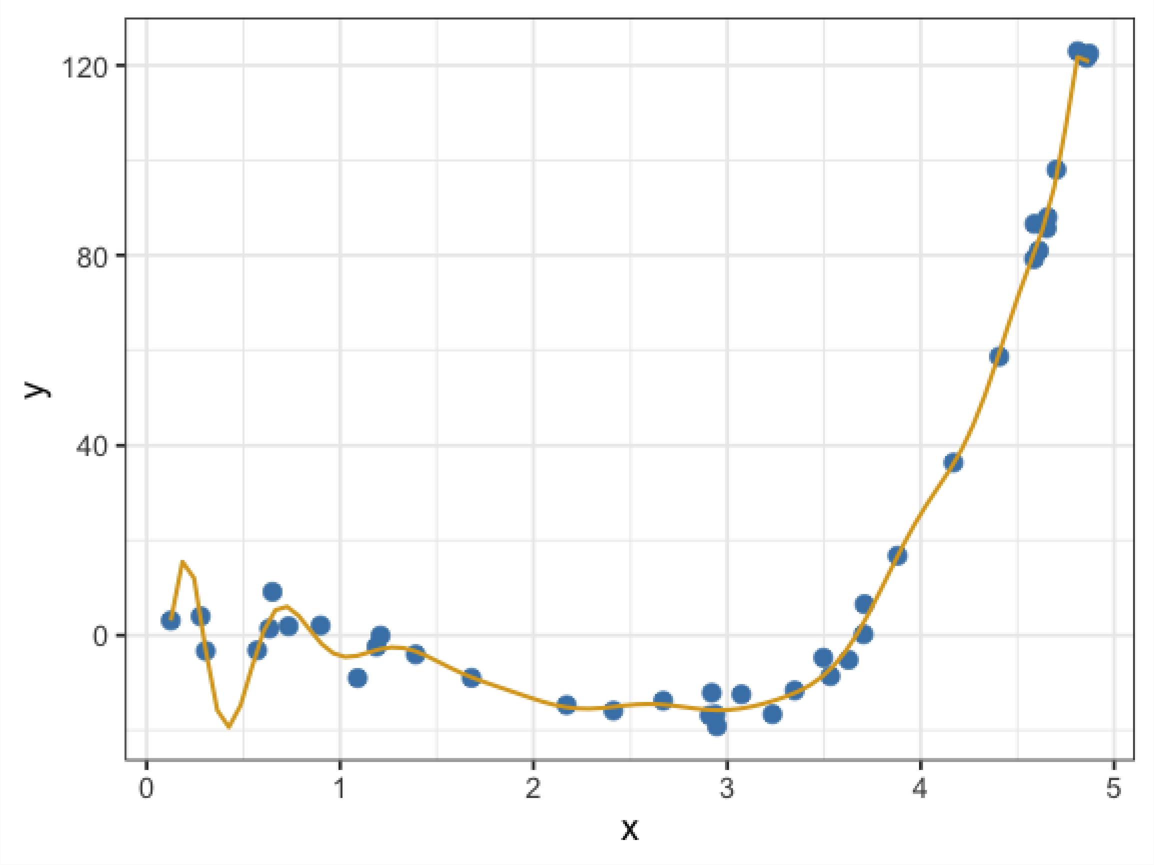
Comparing R^2

```
c(summary(m1)$r.squared,
  summary(m2)$r.squared,
  summary(m3)$r.squared,
  summary(m4)$r.squared)
```

```
## [1] 0.439 0.919 0.992 0.994
```

The observed data is best explained by the quartic model. So that's the best model, right?

The BEST model!



The BEST model!

```
mBEST <- lm(y ~ poly(x, 20))
c(summary(m1)$r.squared,
  summary(m2)$r.squared,
  summary(m3)$r.squared,
  summary(m4)$r.squared,
  summary(mBEST)$r.squared)
```

```
## [1] 0.439 0.919 0.992 0.994 0.997
```

But surely that's not the best model...

Three Criteria

1. R^2
2. R^2_{adj}
3. p-values

There are many others (AIC, BIC, AIC_C, ...).

R^2_{adj}

A measure of explanatory power of model:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

But it only goes up with added predictors, therefore we add a penalty.

$$R^2_{adj} = 1 - \frac{SS_{res}/(n-(p+1))}{SS_{tot}/(n-1)}$$

R^2 vs. R^2_{adj}

```
summary(mBEST)$r.squared
```

```
## [1] 0.997
```

```
summary(mBEST)$adj.r.squared
```

```
## [1] 0.994
```

The Signal and the Noise

live coding