Statistical Learning

ISLR

Book website

http://www-bcf.usc.edu/~gareth/ISL/

Answers to the ISLR questions

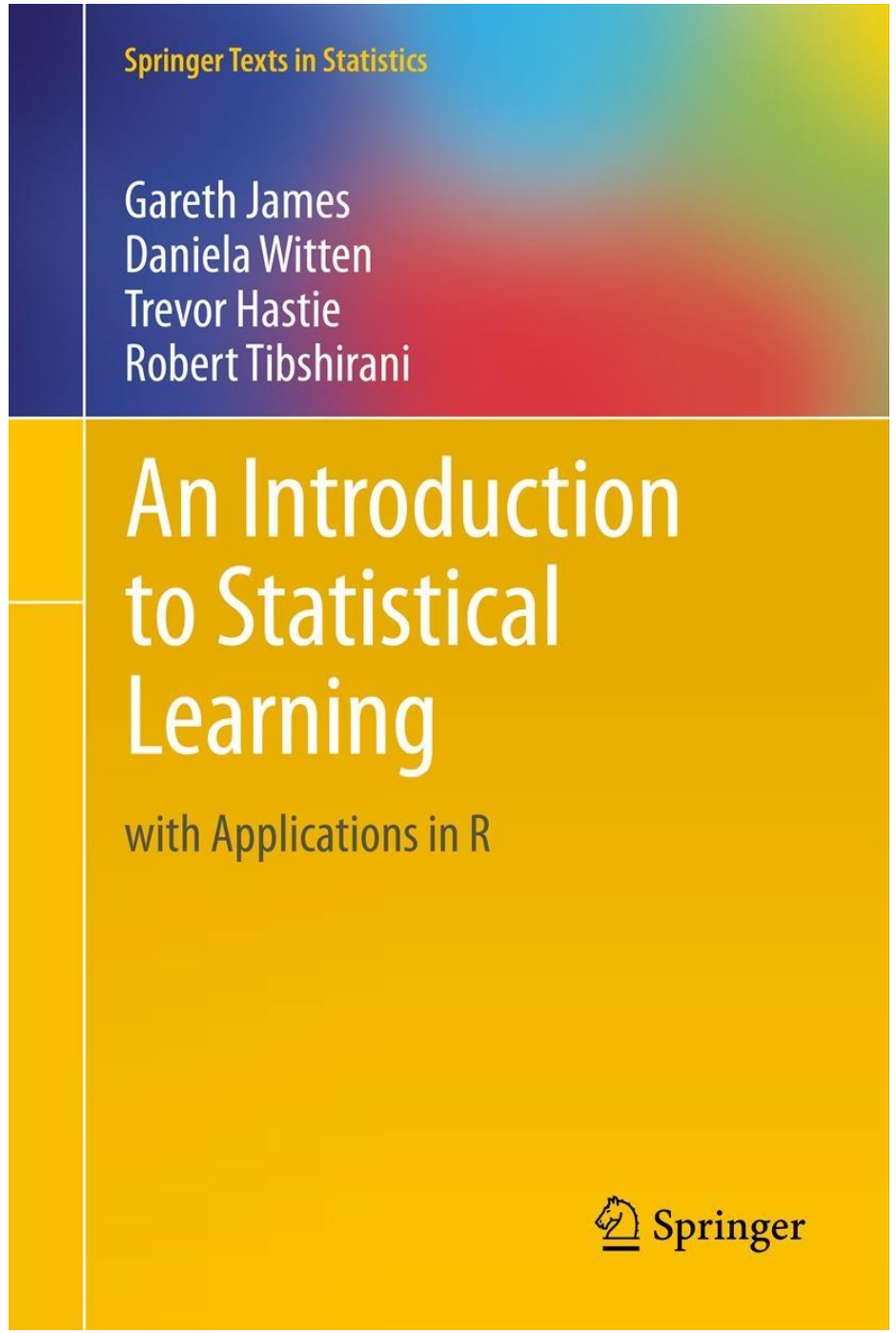
https://github.com/asadoughi/stat-learning

Lectures by Hastie and Friedman

http://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/

Slides and R videos by Abbass Al Sharif

http://www.alsharif.info/#!iom530/c21o7



Supervised statistical learning

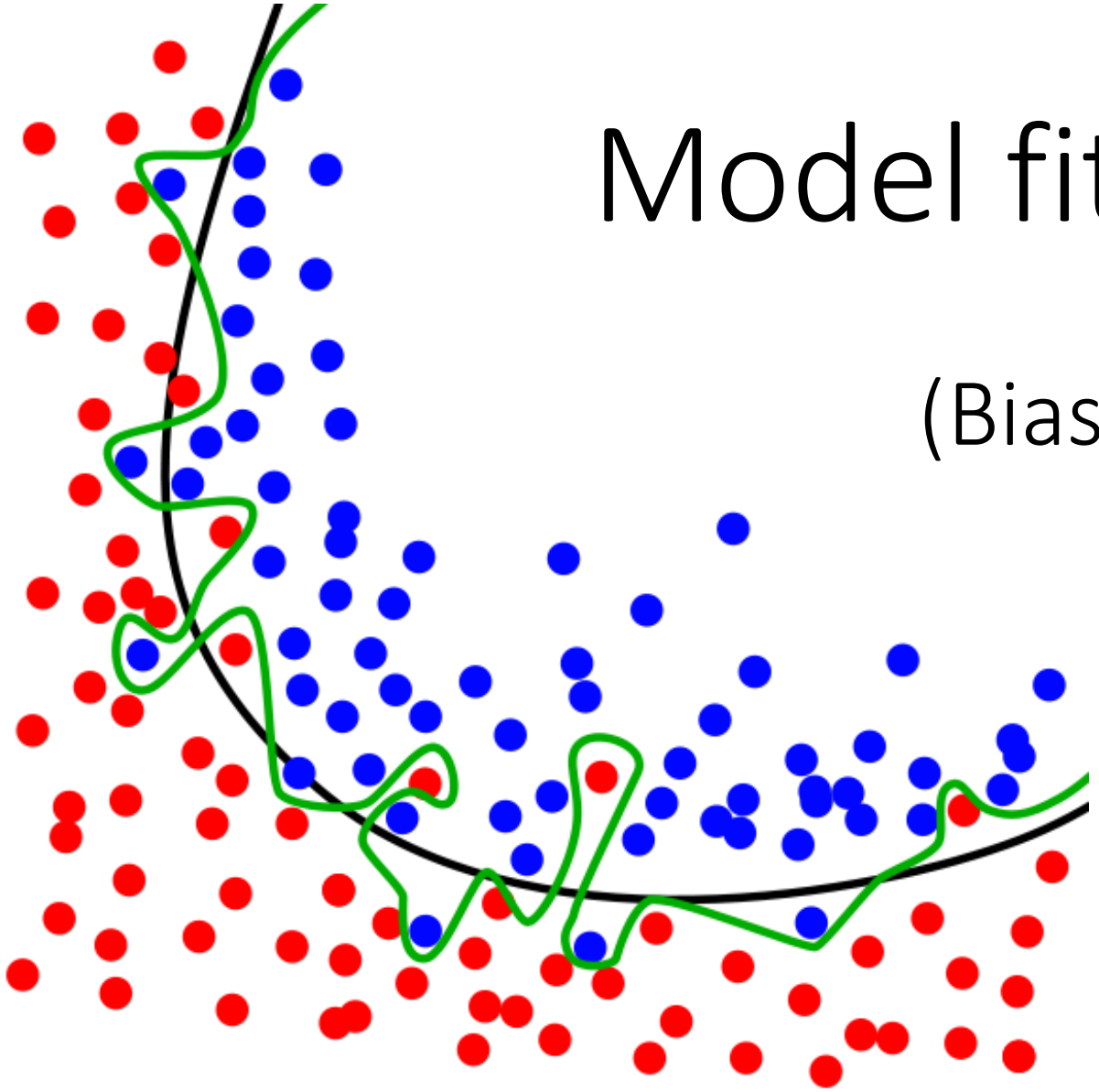
quantity of interest *function* *predictors* *error*

Assume: $Y = f(X) + \epsilon$

Estimate: f to get: $\hat{f}$

Such that: $\hat{Y} = \hat{f}(X)$

For prediction and/or inference

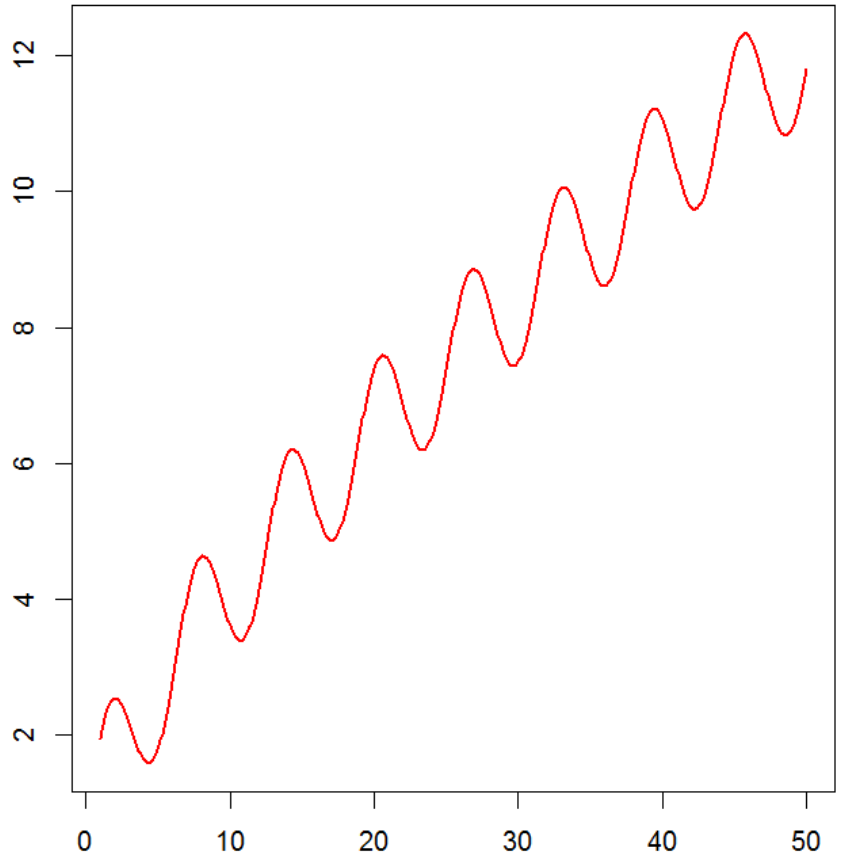


Model fit vs. Model stability

(Bias – variance trade-off)

Assume known function *f*

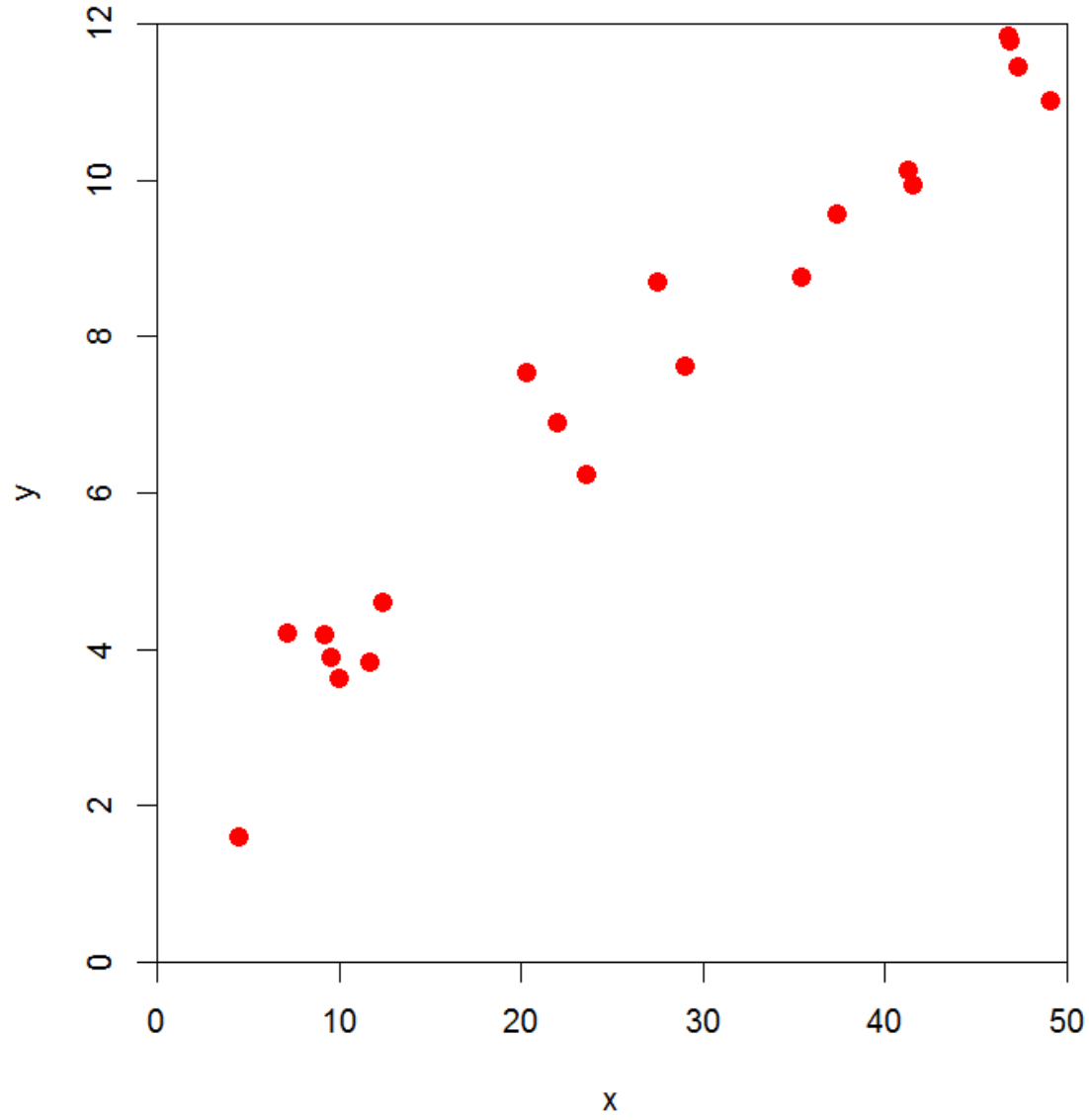
$$Y = f(X) = \frac{X}{10} + \sin(X) + \sqrt{X}$$




```
f <- function(x) x/10 + sin(x) + sqrt(x)
X <- seq(1,50,0.1)
Y <- f(X)
plot(X, Y)
```

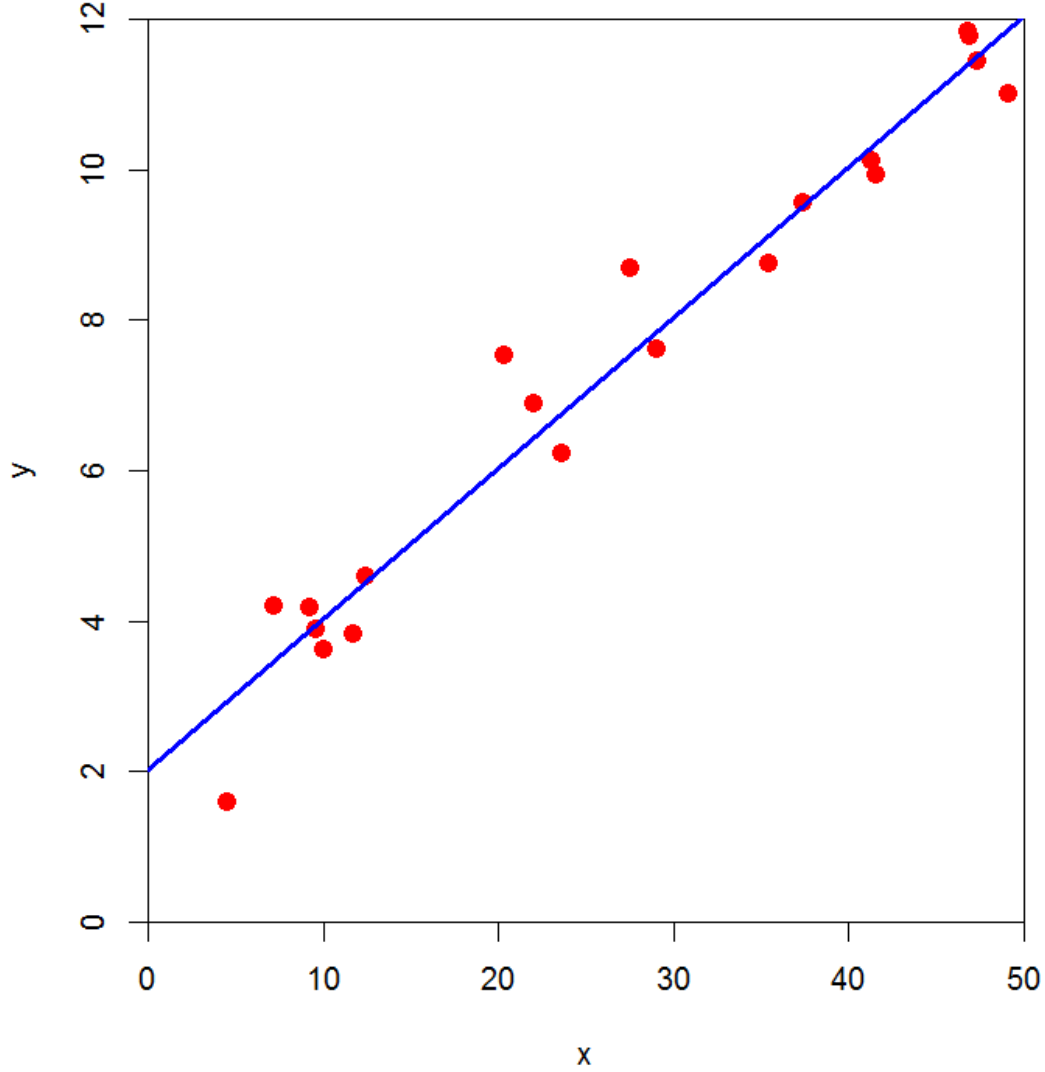
Collect a sample (here without error)

```
set.seed(2)
x <- sample(X, 20)
y <- f(x)
plot(x, y)
```



Estimate *f*

```
> m <- lm(y ~ x)
> plot(x, y)
> abline(m)
> coefficients(m)

(Intercept)                x
   2.032109    0.200007

> summary(m)$r.squared
[1] 0.9573415
```



Linear (parametric) model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\hat{f}: \quad y = 2.03 + 0.2\,x$$

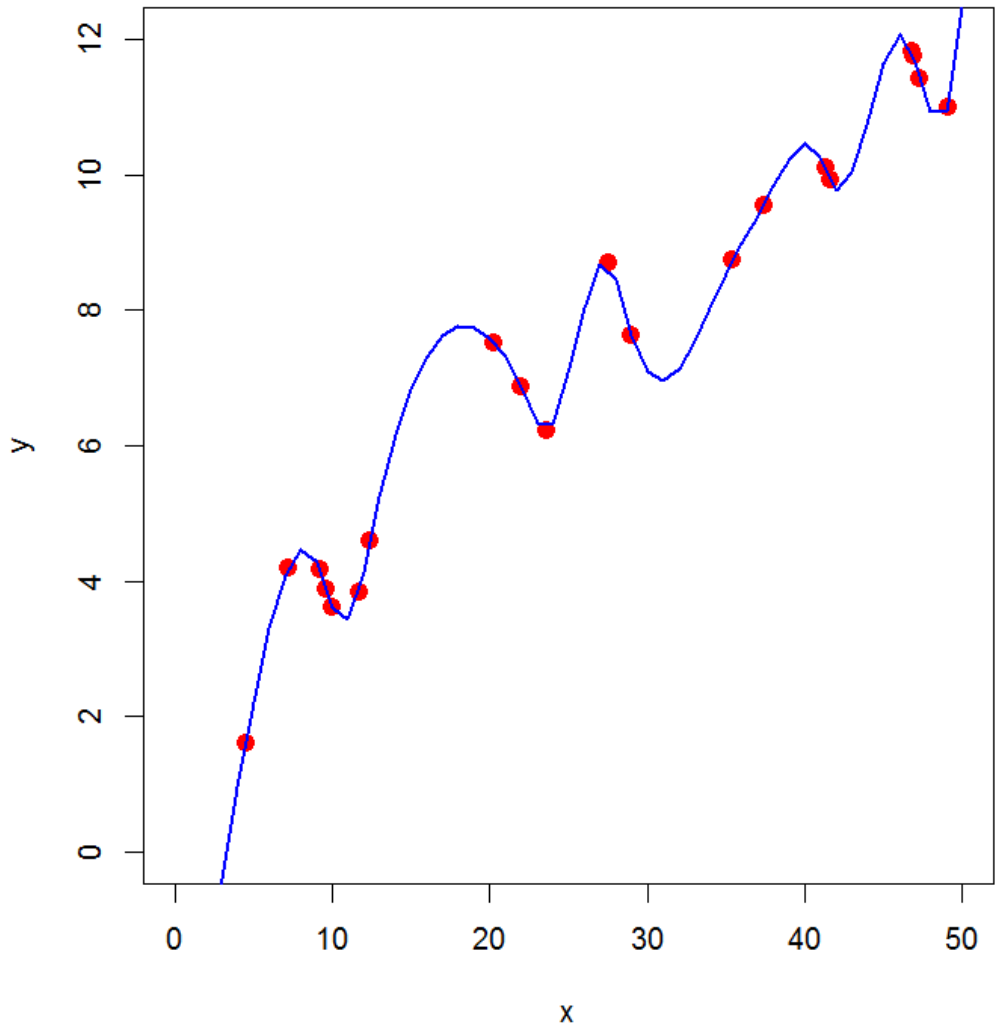
```
> coefficients(m)

(Intercept)                x
   2.032109         0.200007
```

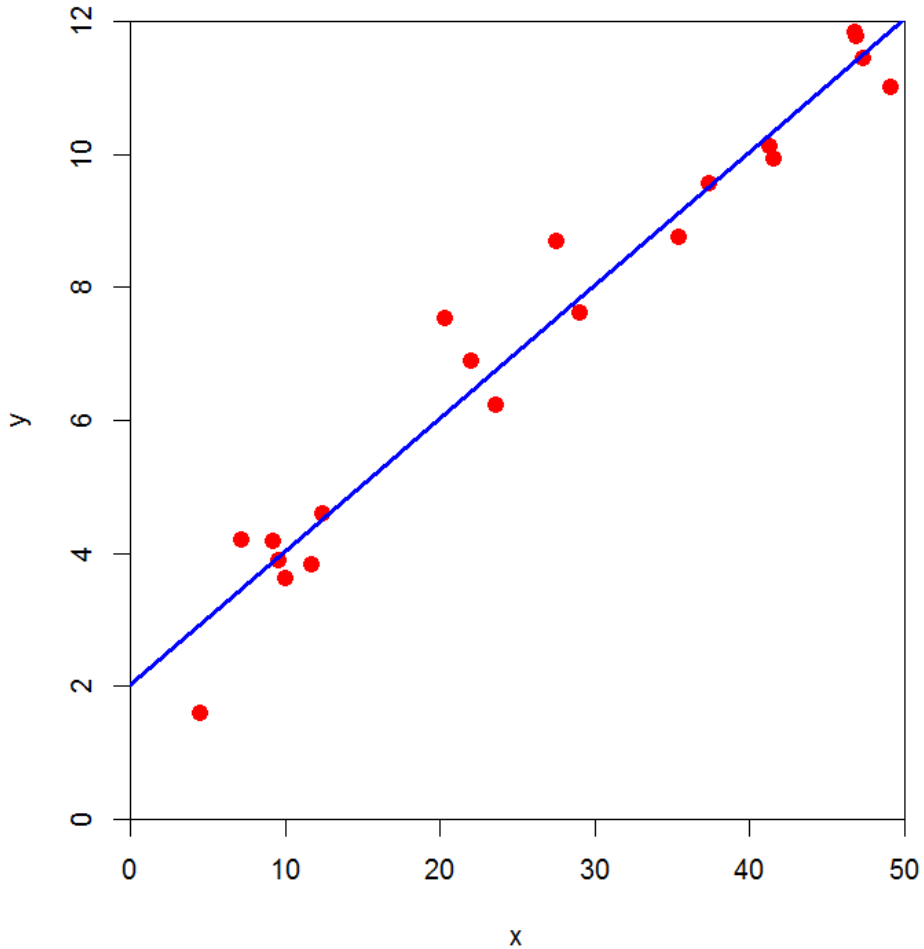
$$(f: \; y = \frac{X}{10} + \sin(X) + \sqrt{X}\,)$$

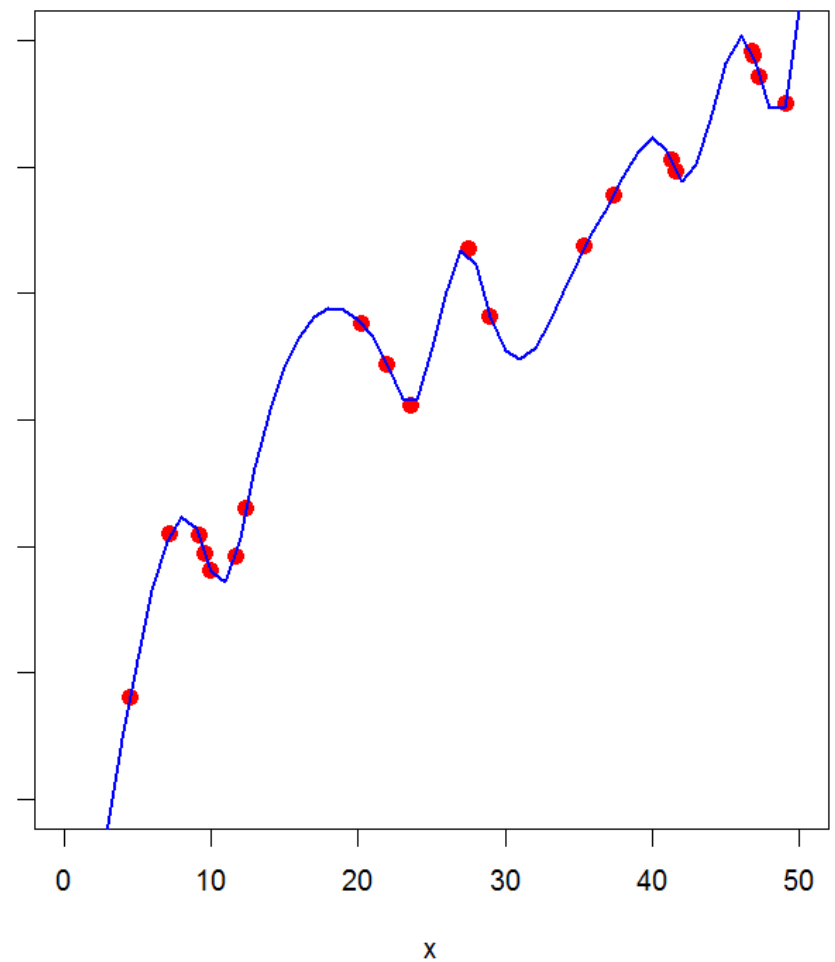
Flexible model

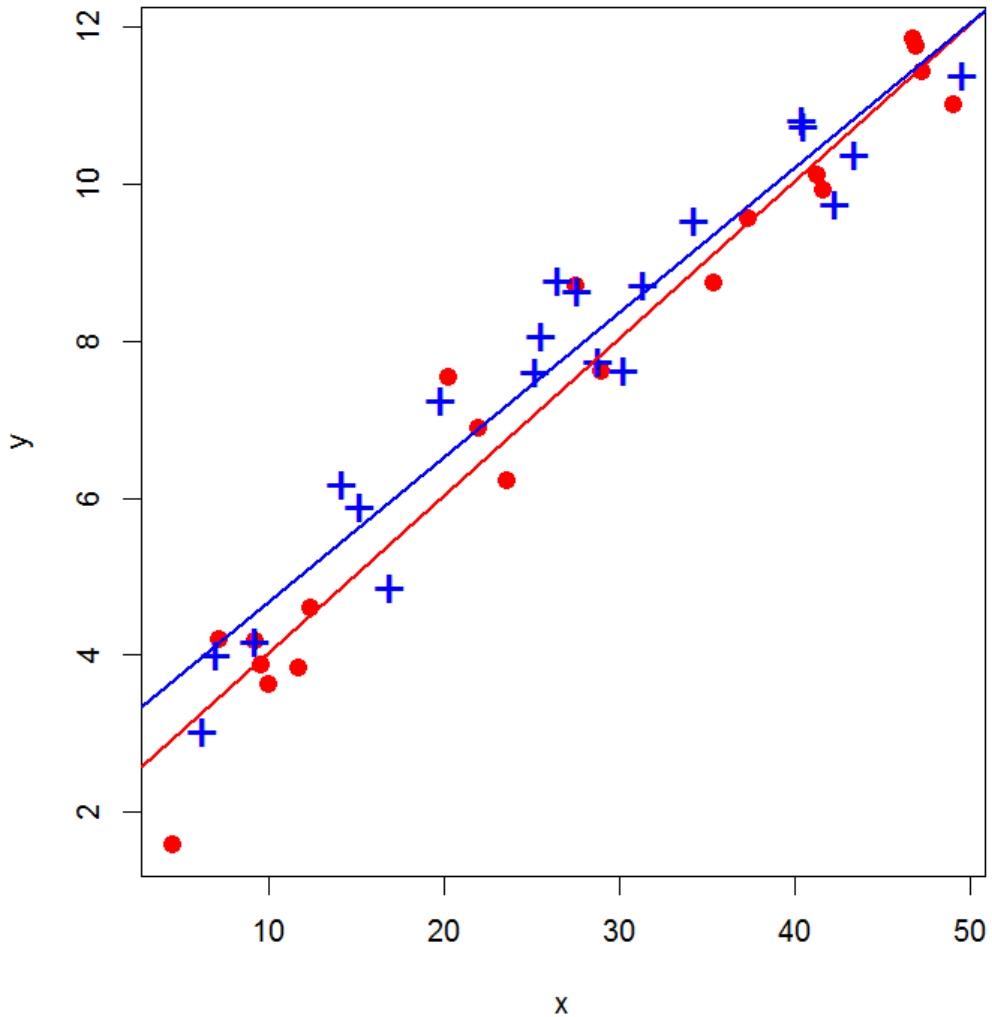
```
> sf <- splinefun(x, y)
> mse <- mean((y - sf(x))^2)
> mse
[1] 0
```

```
> plot(x, y)
> lines(1:50, sf(1:50))
```

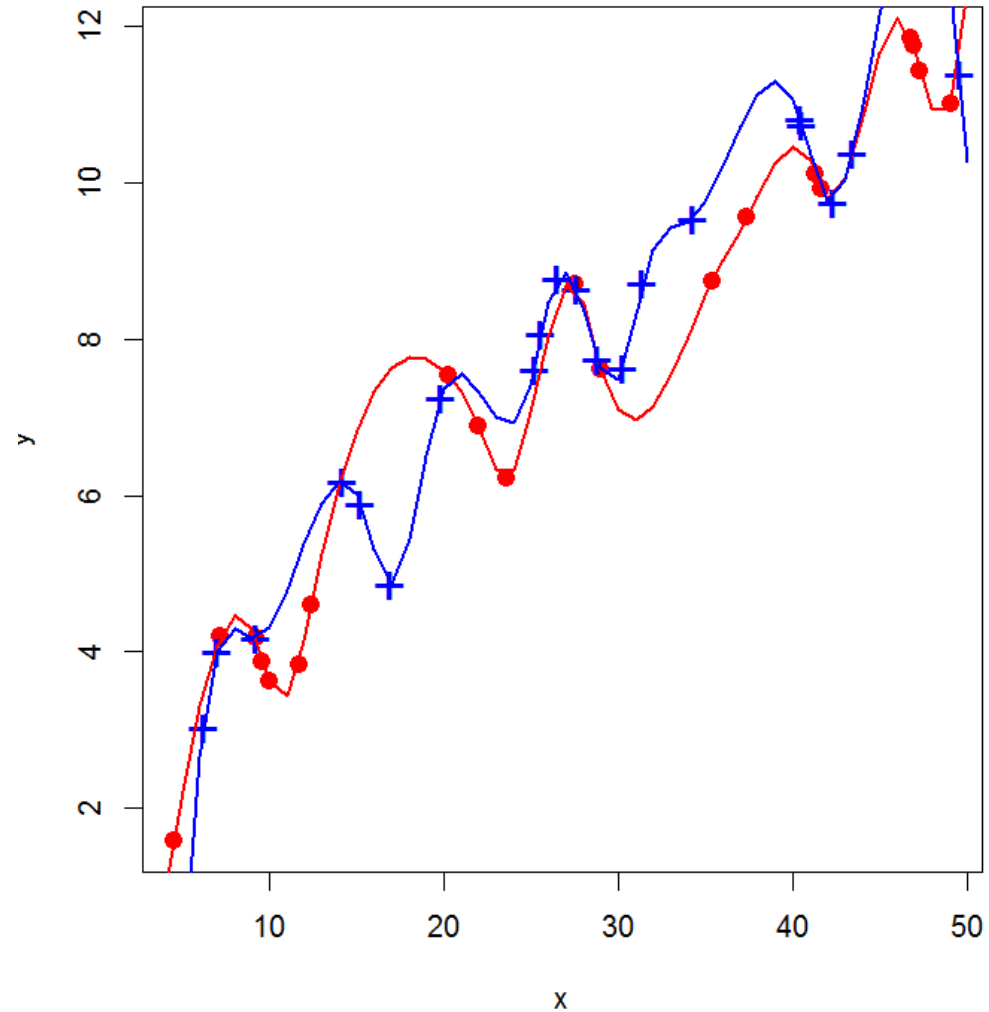



y
x
0
2
4
6
8
10
12
0
10
20
30
40
50


x
0
10
20
30
40
50



MSE = 0.59



MSE = 0.74

Use simple models if you can

Easy interpretation & understanding
Not prone to overfitting

Use model training and testing data for model evaluation

Beware the variance-bias tradeoff

Simple linear regression if great.
But, you can get in trouble with:

1. **Non-linearity of the data (!!)**
2. **Non-independence of the error terms (perhaps spatially)**
3. Non-constant variance of error terms
4. Outliers
5. High leverage points
6. **Collinearity**

Alternatives to (simple) least squares regression

For increased accuracy and interpretability

Subset selection --- Stepwise, Lasso

Shrinkage --- Ridge, Lasso

Dimension reduction --- Principal components regression, PLS

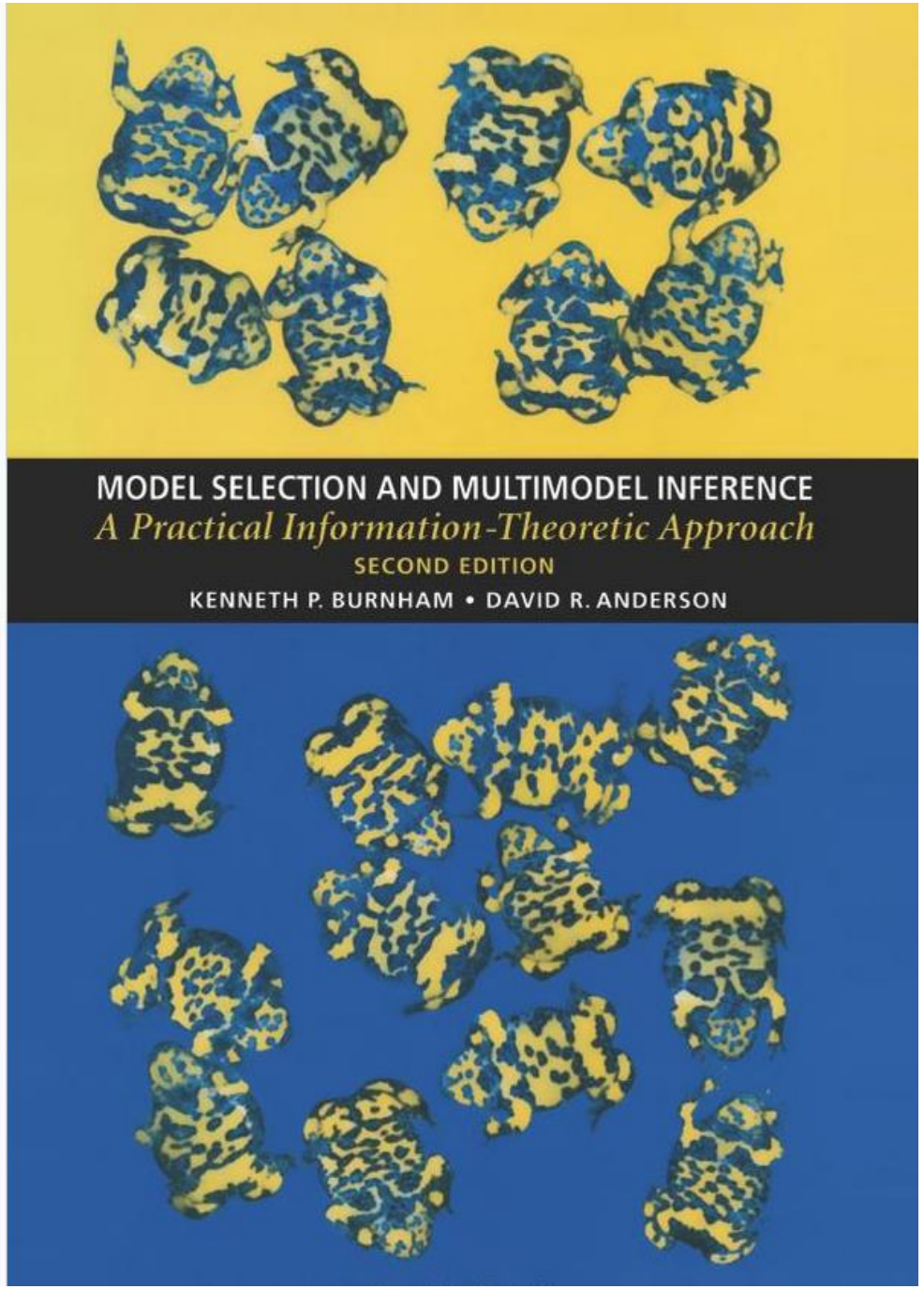
AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The statement's six principles, many of which address misconceptions and misuse of the *p*-value, are the following:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*



$$\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

Plurality must never be posited without necessity
(**simpler is better**)

OPEN ACCESS

ESSAY

969,082 VIEWS | 1,187 CITATIONS

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

- Abstract
- Modeling the Framework for False Positive Findings
- Bias
- Testing by Several Independent Teams
- Corollaries
- Most Research Findings Are False for Most Research Designs and for Most Fields
- Claimed Research Findings May Often Be Simply Accurate Measures of the Prevailing Bias
- How Can We Improve the Situation?
- References

Reader Comments (32)

Media Coverage (38)

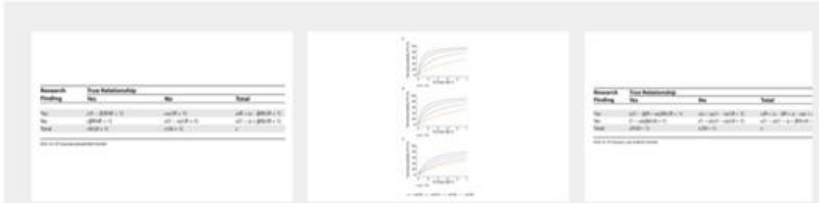
Figures

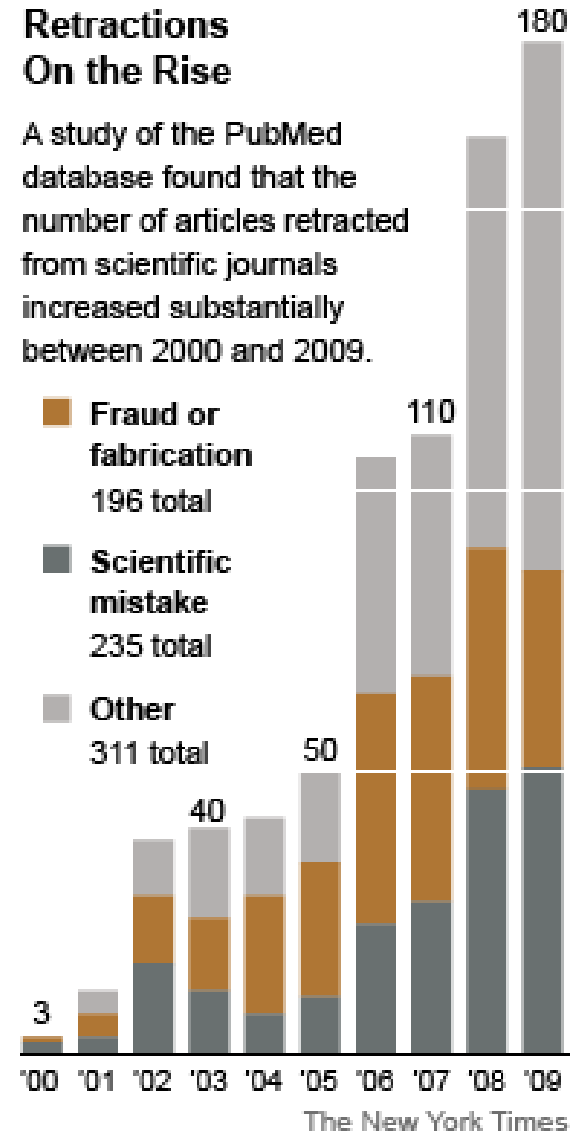
Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Figures





The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.

The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true. ((Empirical evidence on expert opinion shows that it is extremely unreliable)).

The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

Inference

Science: Is it true that?

experimental, theory, generalization

Prediction

Practice: What if?

observational, computational, data intensive