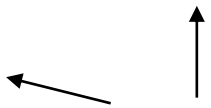
Statistical modeling in logistics

MMS075

Lecture 7a – Training error vs test error,
Validation set, K-fold cross-validation, LOOCV,
Ethical analysis of big data



These are new compared to Lecture 6b

Acknowledgement: Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Outline

- Training error vs test error
  - Mean squared error
  - Error rate for classification
  - Overfitting
- Methods for estimating test error:
  - Validation set approach
  - K-fold cross-validation
  - Leave-one-out cross-validation (LOOCV)
- Ethical analysis of big data

Recommended resources

Reading in ISL: Sections 2.2.1, beginning of 2.2.3, and 5.1 for theory, 5.3.1-5.3.3 for R codes

The videos from the Statistical Learning course are available at this link. Relevant videos for the new material today:

- **Assessing Model Accuracy and Bias-Variance Trade-off** (10:04)
- **Estimating Prediction Error and Validation Set Approach** (14:01)
- **K-fold Cross-Validation** (13:33)
- **Cross-Validation: The Right and Wrong Ways** (10:07)

For the discussion of ethical analysis of big data:

- Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. (2017). Ten simple rules for responsible big data research. PLoS Comput Biol 13(3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Training error vs test error

Mean squared error

Error rate for classification

Overfitting

What is a good model?

- We have n observations in form of predictor-response pairs:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- Based (**trained**) on these observations, we define a model $\hat{f}$ and hope that the model approximates the true connection between predictor and response, i.e. $\hat{f}(X) \approx Y$

- **Training mean squared error (training MSE)** measures how well this holds for training points, i.e. measures quality of fit:

$$MSE = \frac{(y_1 - \hat{f}(x_1))^2 + (y_2 - \hat{f}(x_2))^2 + \ldots (y_n - \hat{f}(x_n))^2}{n}$$

What do we really want?

- Why do we want a model with good quality of fit, i.e. small MSE?

- Why do we need a model at all? We KNOW the response values for the training set → why should we estimate them?

- Because we hope that the model gives useful information for **new data (test data)**: if $(x_1^{\text{new}}, y_1^{\text{new}}), (x_2^{\text{new}}, y_2^{\text{new}}), \ldots$ are new, previously unseen observations that were not used to train (i.e. define) the model, we want a model with small average prediction error

→ We want to minimize $\text{Average}\left[(y^{\text{new}} - \hat{f}(x^{\text{new}}))^2\right]$

Training error vs test error

Mean squared error

Error rate for classification

Overfitting

How to define error for classification?

- We have n observations in form of predictor-response pairs: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$; the responses are categories

  → cannot use definition of MSE; what to take instead?

- **Training error rate** for a model $\hat{f}$ trained on the above observations is the proportion of mistakes our classifier makes when applying it on the training observations:

  $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$: the predicted response classes by $\hat{f}$ for $x_1, x_2, \ldots, x_n$;

  $I(y_i \neq \hat{y}_i)$: predicted response class for observation i differs from observed;

$$\text{Training error rate} = \frac{I(y_1 \neq \hat{y}_1) + I(y_2 \neq \hat{y}_2) + \ldots I(y_n \neq \hat{y}_n)}{n}$$

Good classifier has low **test** error rate

- Analogously to regression, the training error rate is less important
- We want a small average error rate for **new data (test data)**: if $(x_1^{\text{new}}, y_1^{\text{new}}), (x_2^{\text{new}}, y_2^{\text{new}}), \ldots$ are new, previously unseen observations that were not used to train the model

→ want to minimize **test error rate**: $\text{Average}\left[I\left(y^{\text{new}} \neq \hat{y}^{\text{new}}\right)\right]$

- We discuss regression henceforth, but classification is analogous

Training error vs test error

Mean squared error

Error rate for classification

Overfitting

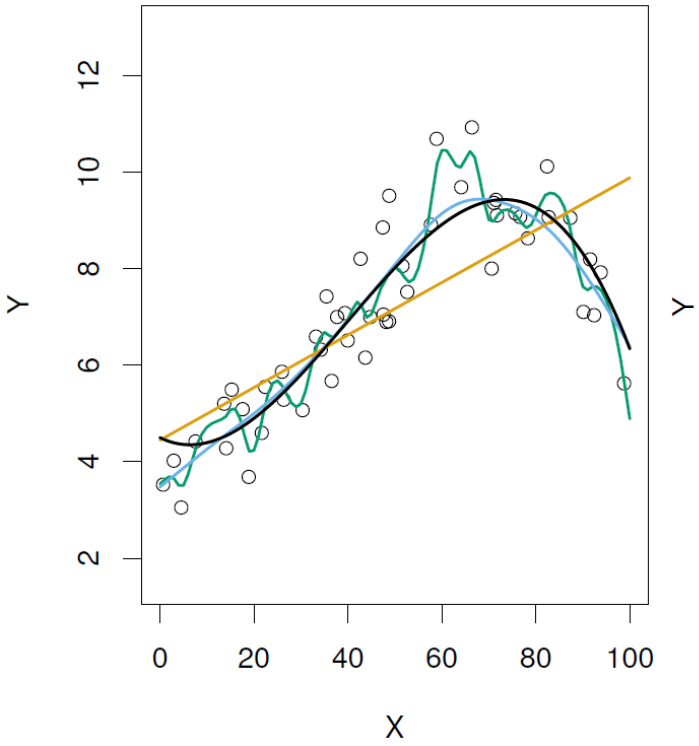
Is it best to minimize training MSE?

- Model with a good fit on training data is a natural aim
- **Overfitting**: if model follows training points too closely & reproduces noise effects (i.e. pick up patterns caused by chance rather than a meaningful relationship) that may not be present in new data
- This is caused by overly flexible models. Examples on next slides
- However, among models of given complexity (e.g. linear models), the one that fits training points best should be chosen

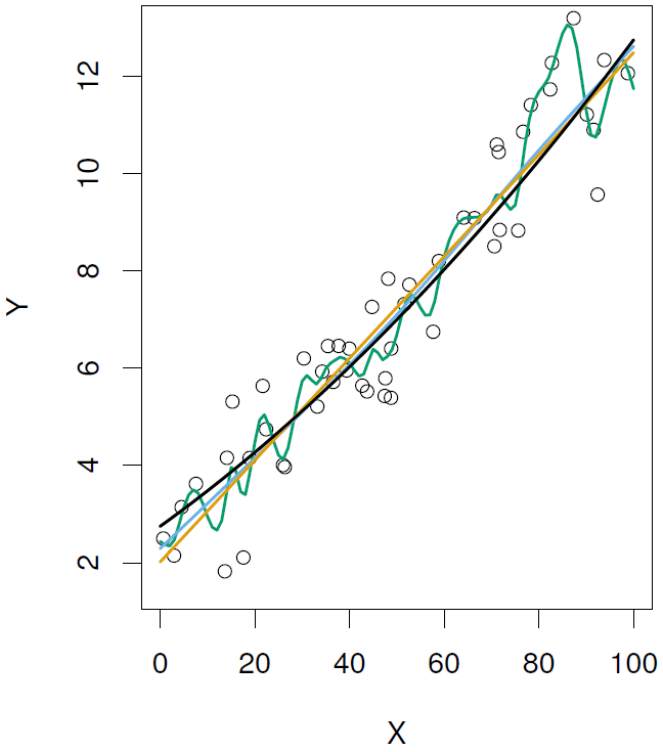
How flexible model do we need in general?

Structure (shape) of points looks close to...

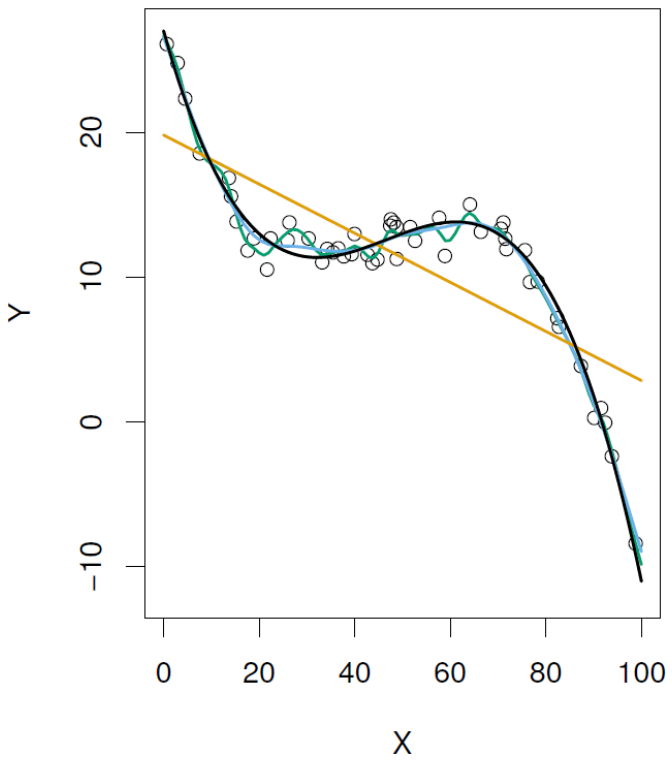
... a quadratic (or low-degree) curve



... a line



... a cubic (or high-degree) curve

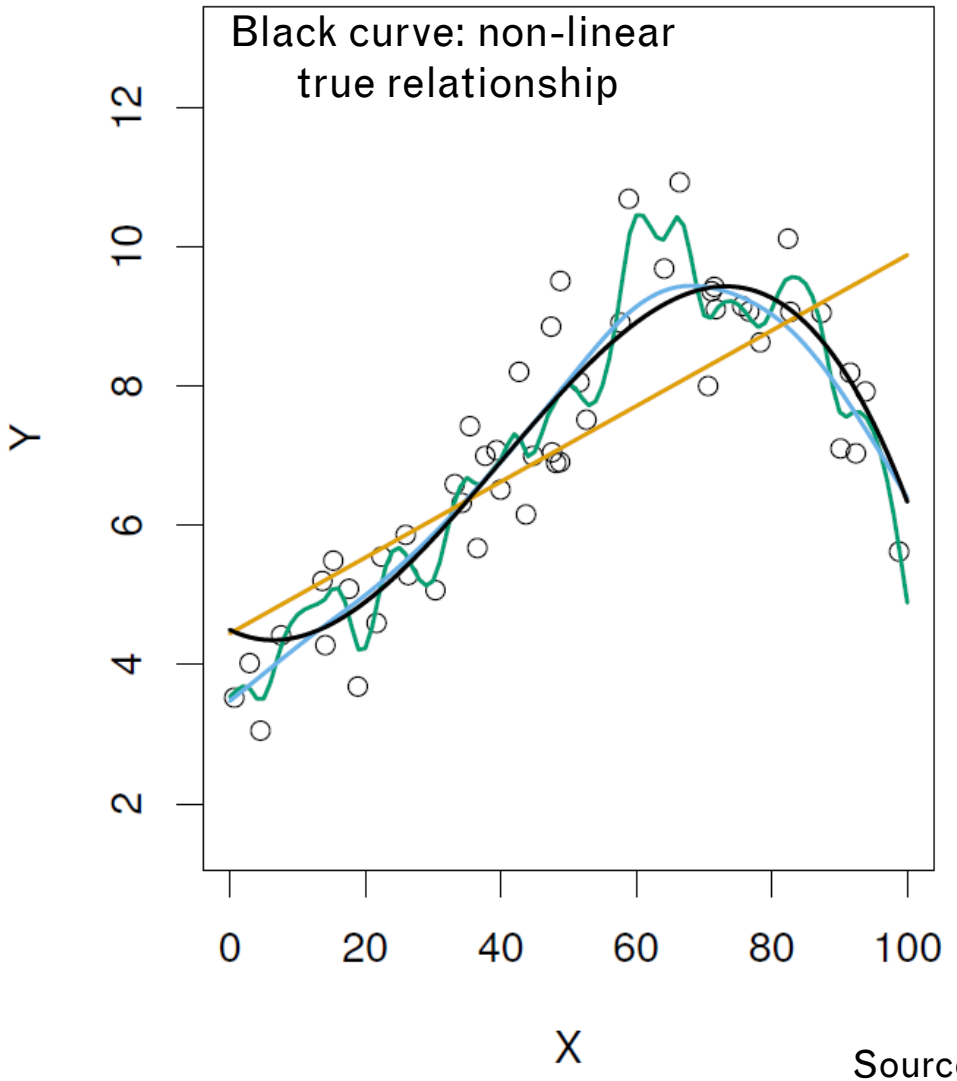


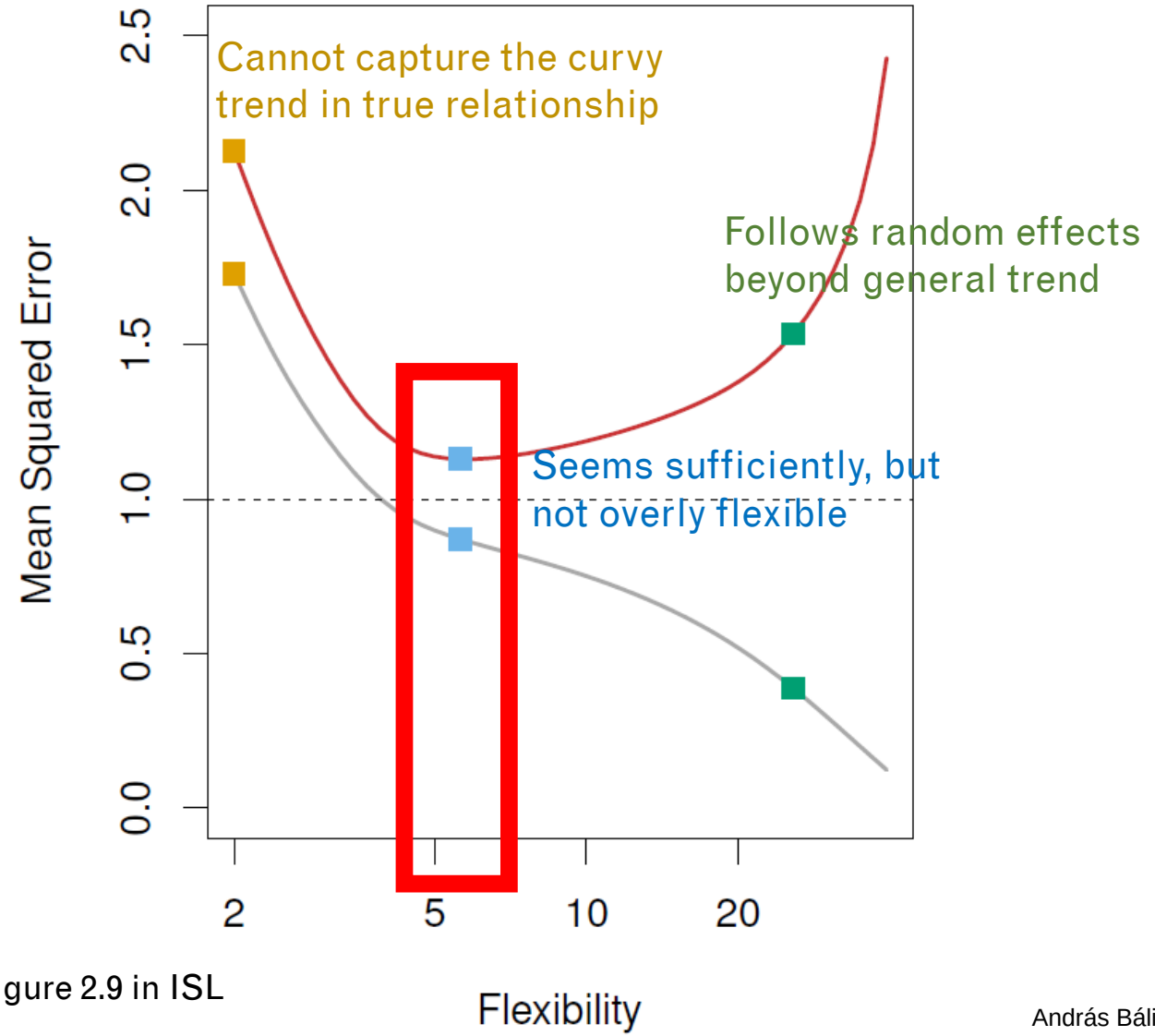
Source: Figures 2.9, 2.10 and 2.11 in ISL

What do we expect?

- Flexibilility corresponding to general shape should work best:
  - Less complex models cannot reproduce observed shape
  - More complex models are too wiggly and can/will follow noise in data
- This was simulated data → we can create a large test set with the true distribution & check test error values for various models
- Next slides show test MSE (red curves) and training MSE (grey curves) for models of different flexibility.
- Training MSE always decreases with larger flexibility; test MSE is U-shaped, as anticipated, indicating a trade-off

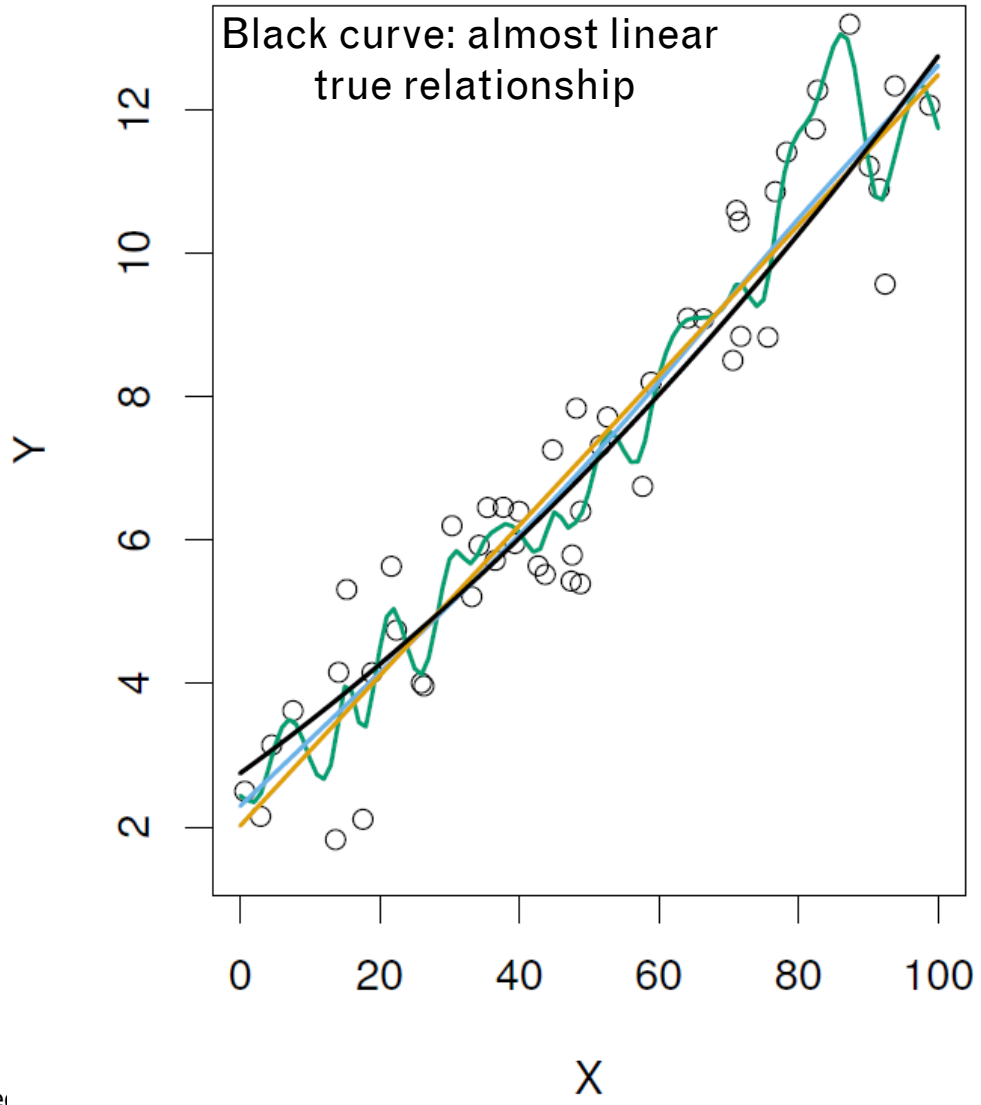
Test error curves by model complexity (1)

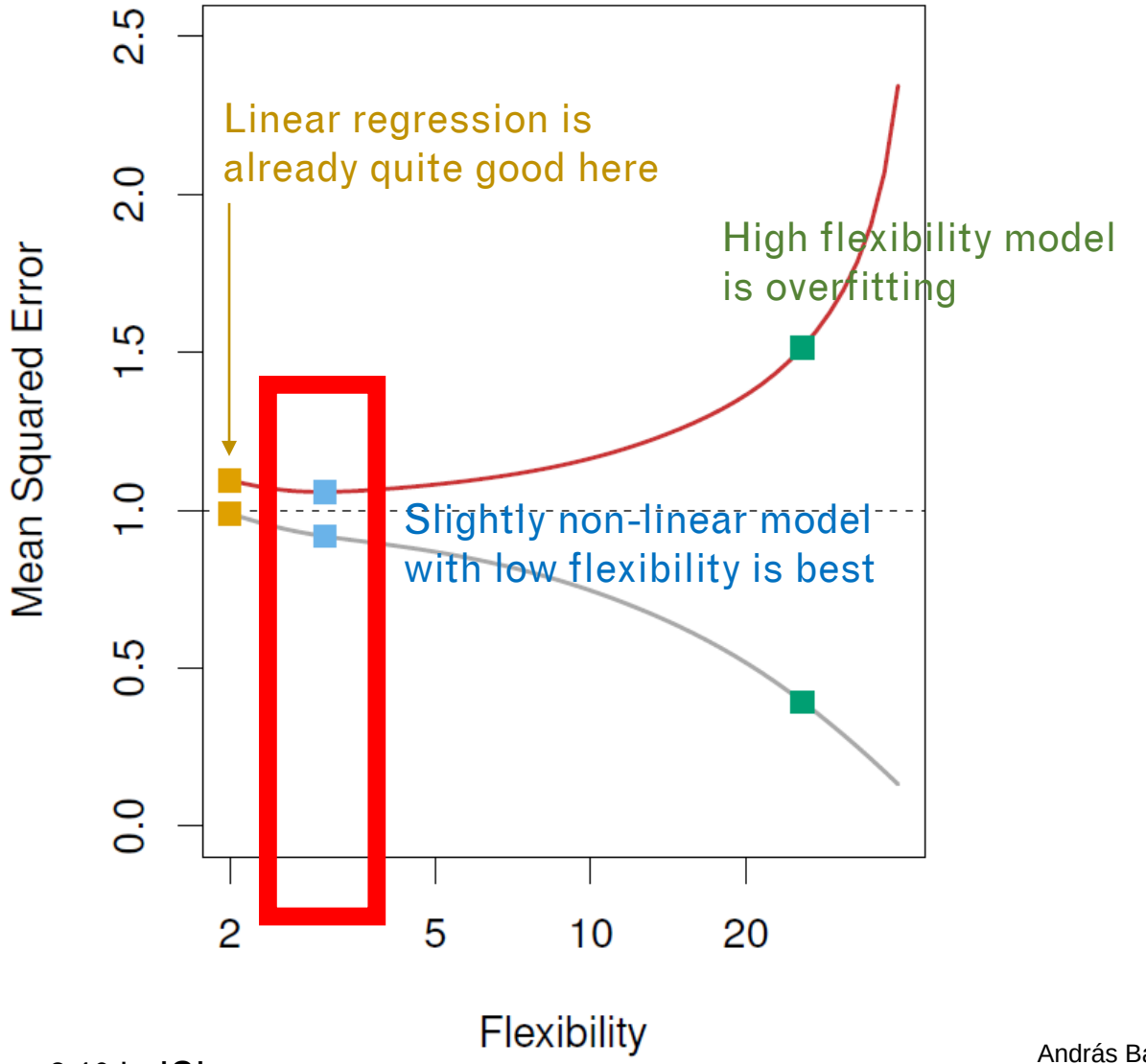




Source: Figure 2.9 in ISL

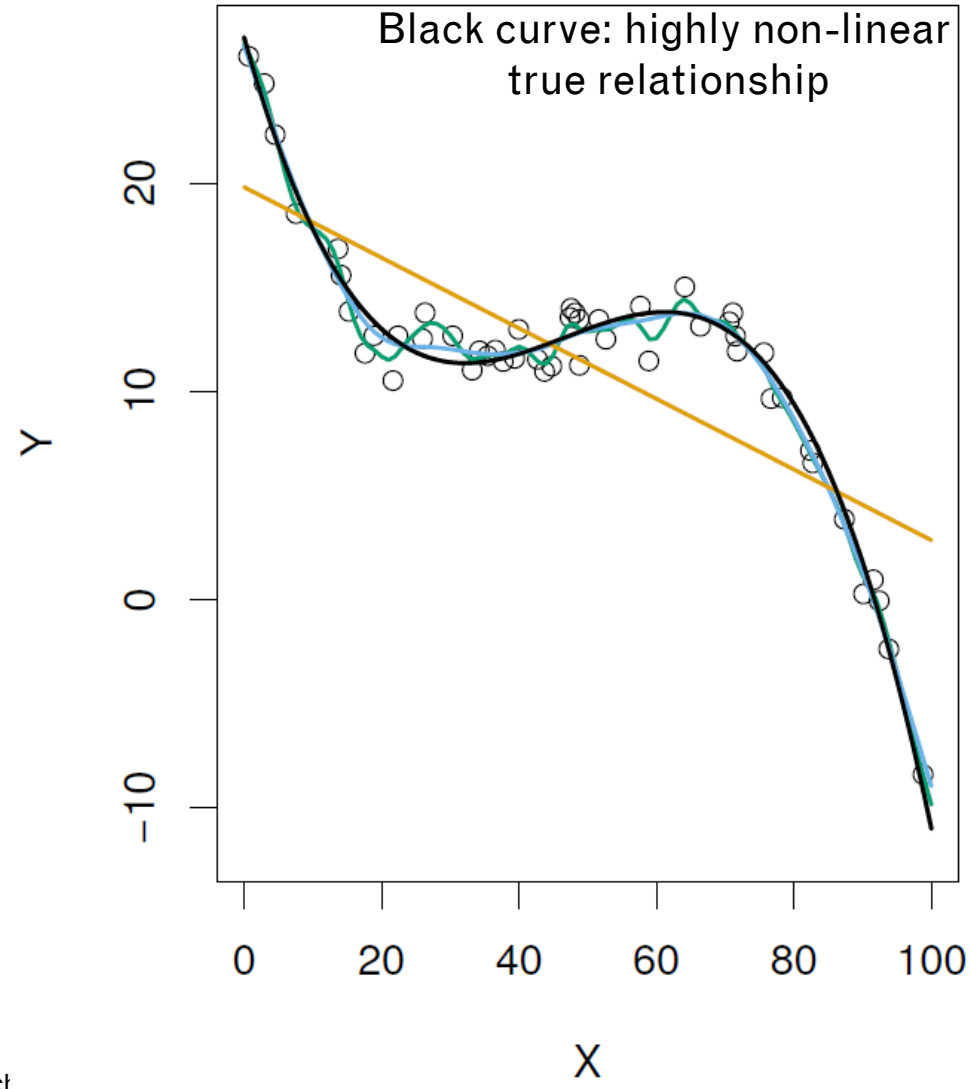
Test error curves by model complexity (2)

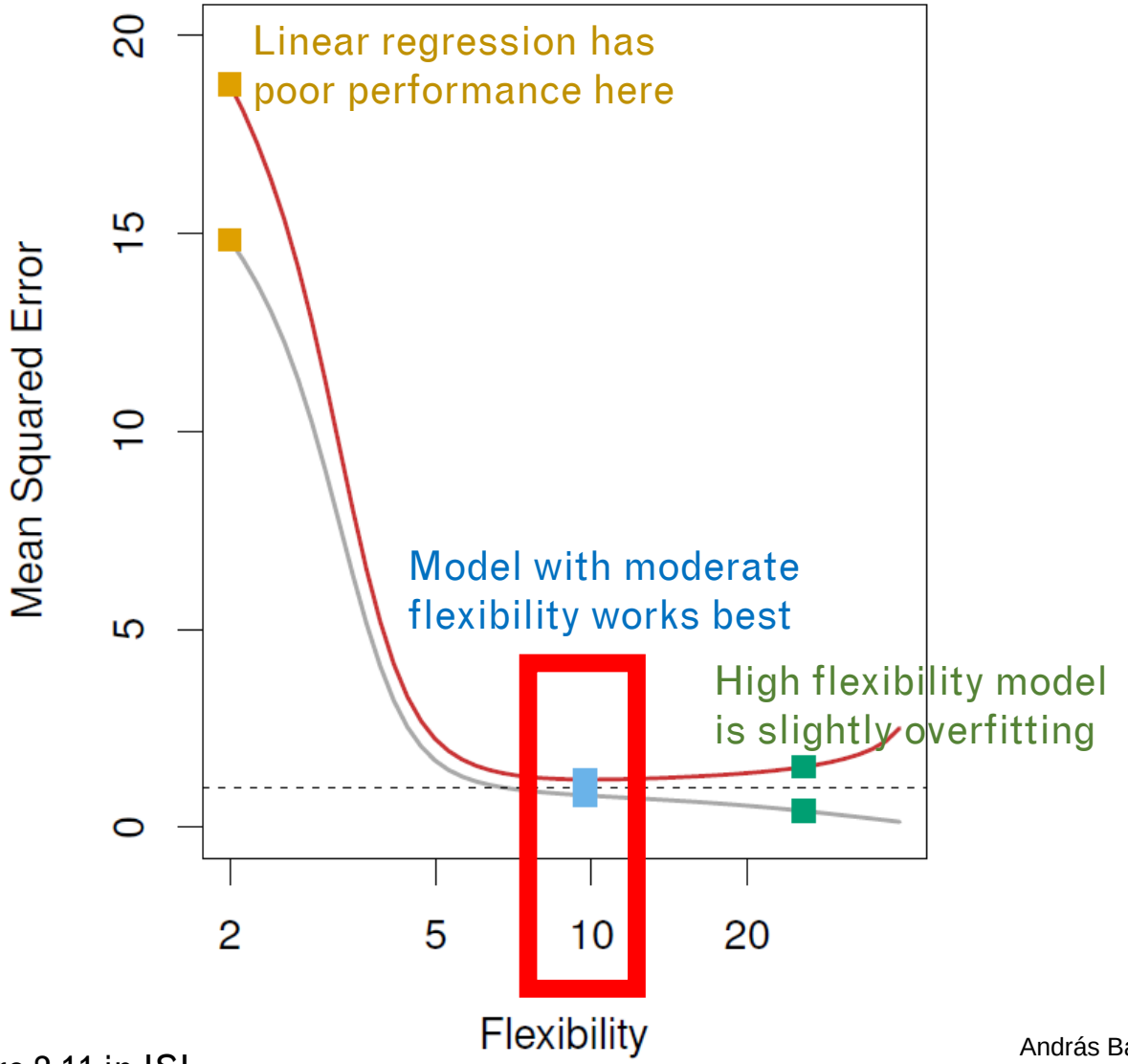




Source: Figure 2.10 in ISL

Test error curves by model complexity (3)





Source: Figure 2.11 in ISL

Simulated data is uncommon

- If data are simulated → as much test data can be produced as needed → we can have a full understanding of test error
- This is not a usual case! We typically have only one set of observations – how can we then estimate test error?
- Three methods are discussed in the next slides

Methods estimating test error

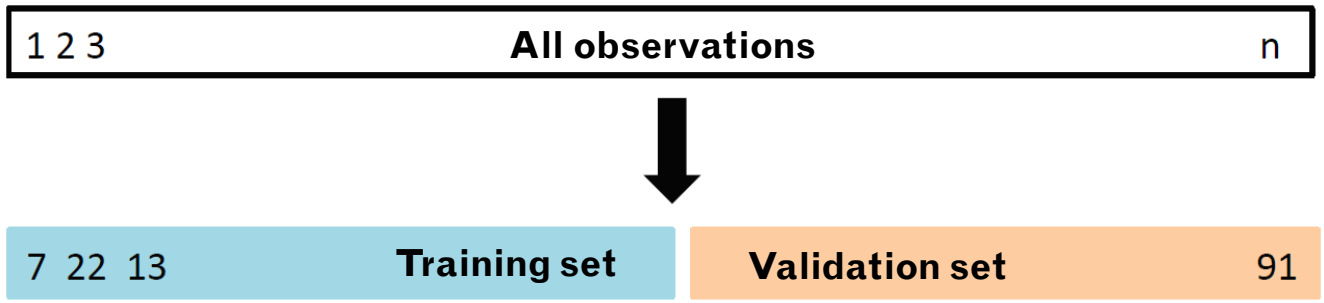
Validation set approach

K-fold cross-validation

LOOCV

Validation set approch: basic idea

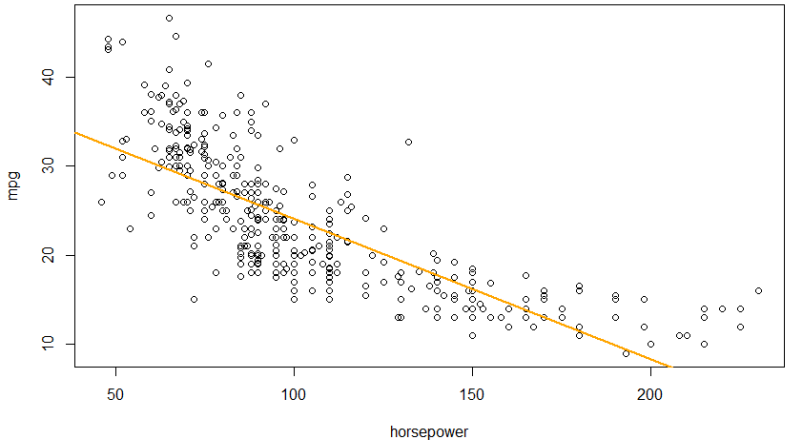
- Reserve some part of observations that will not be used in model building
- This set of points (called **hold out set** or **validation set**) is unseen by the model while the model is defined → it can play the role of test data to see how well the model can predict unseen points
- The set that was used for model definition (e.g. get coefficient estimates) is the **training set**



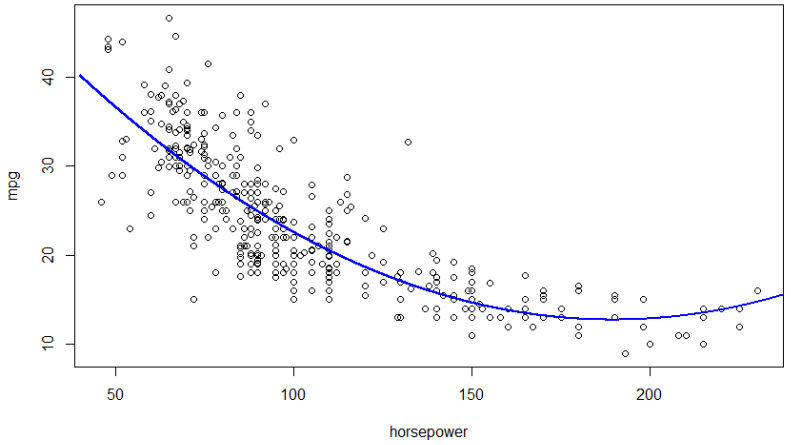
Validation set approach, source: Figure 5.1 in ISL, labels added

Recall mpg vs horsepower models

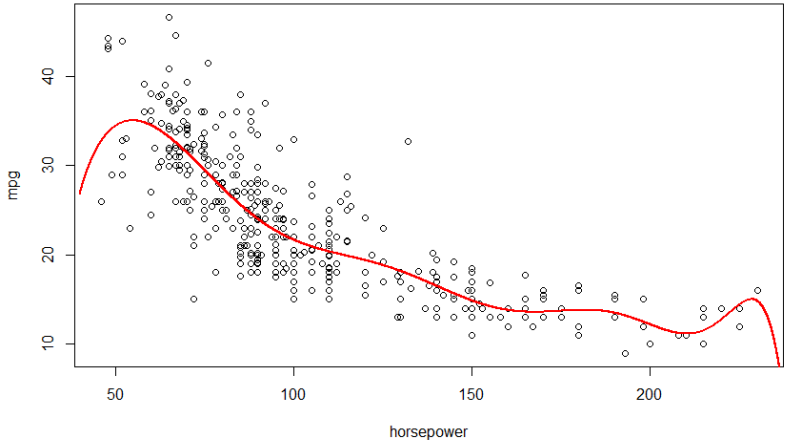
Linear model



Quadratic model (degree 2 polynomial)



High-degree model (degree 10 polynomial)



Plots and p-values suggest:

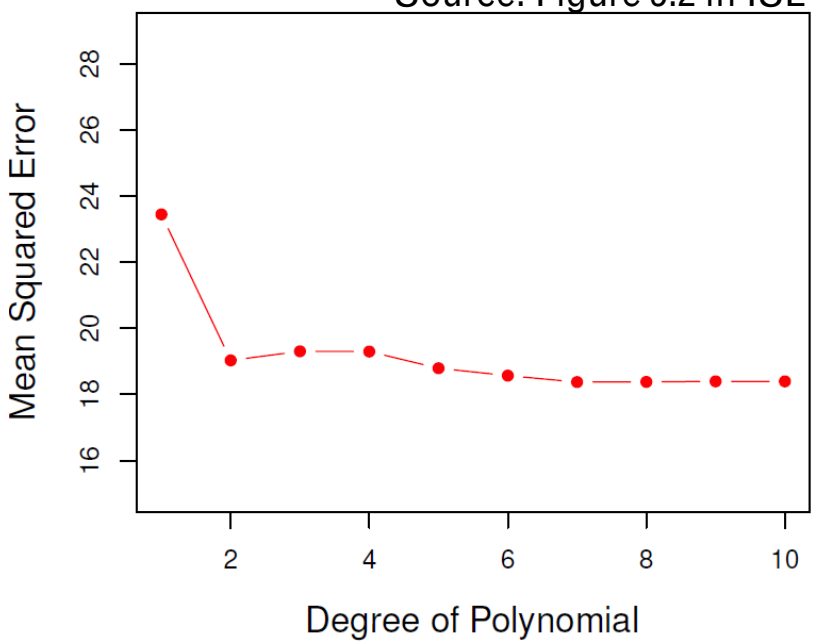
- Quadratic model is better than linear
- High-degree model may be overfitting

What can we conclude with the validation set approach?

Validation set error estimates

1. Divide the observations into a training set and a validation set
2. Using the points in the training set, fit a linear, quadratic and higher degree models
3. Using the points in the validation set, compute MSE for all these models and plot the error estimates:

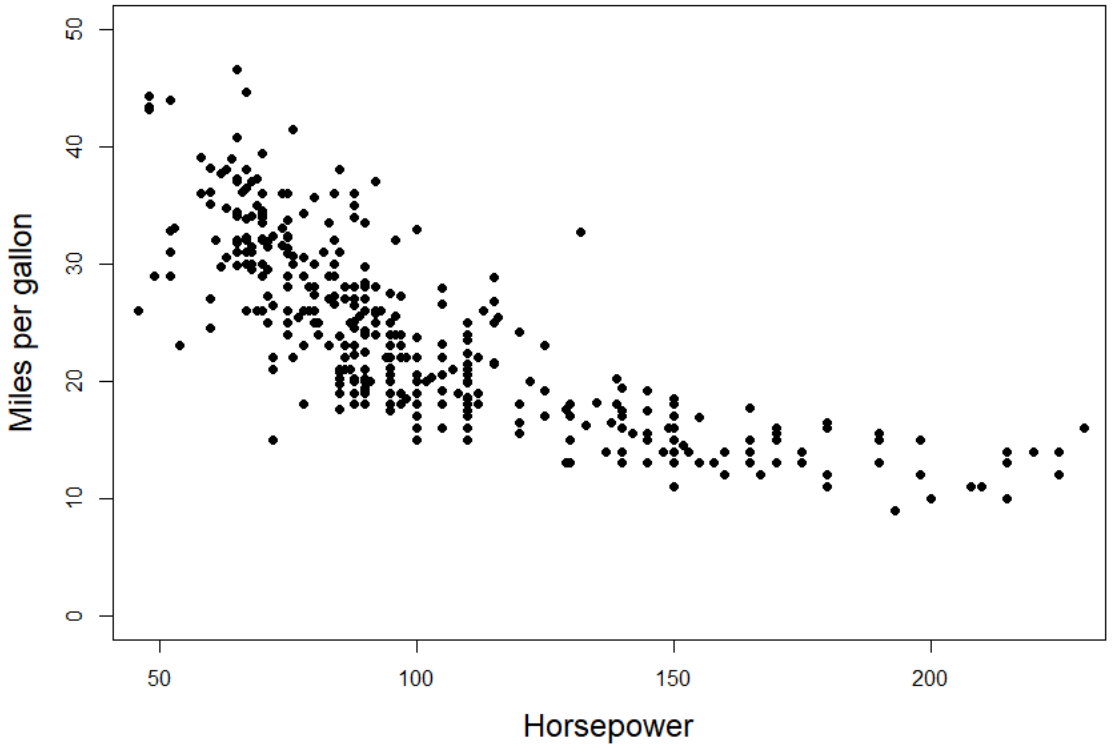
Source: Figure 5.2 in ISL



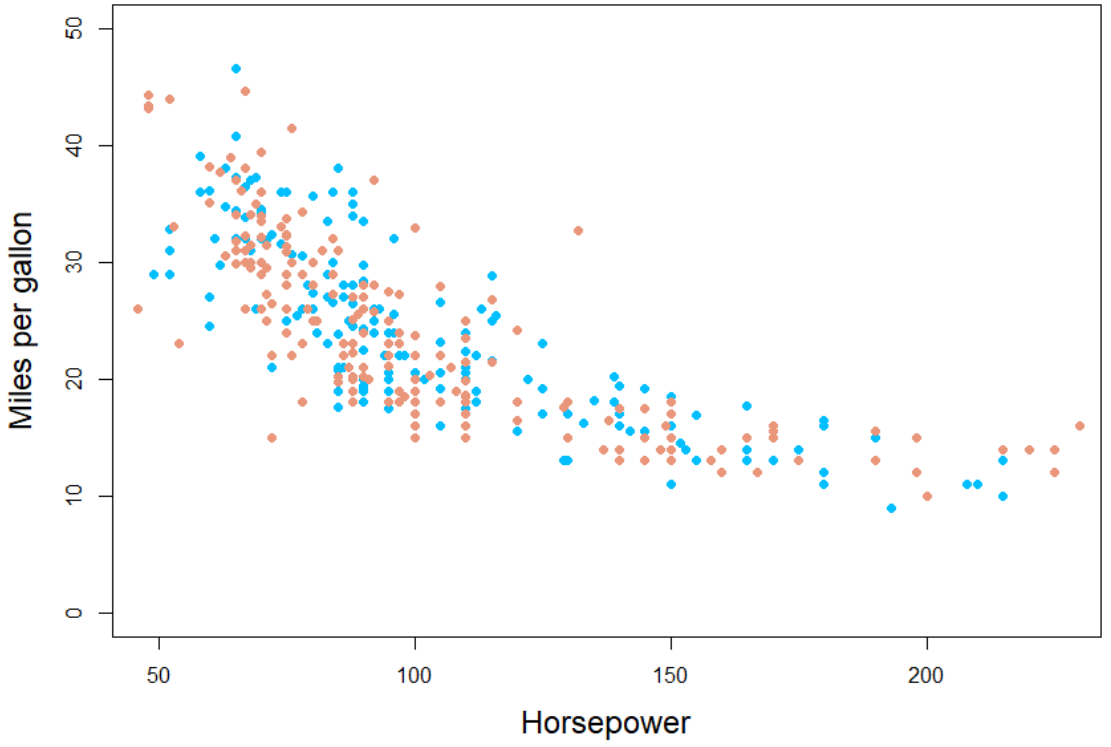
Detailed example for 1,2 and 10 degrees is shown on the next slides

Step 1: Divide the observations

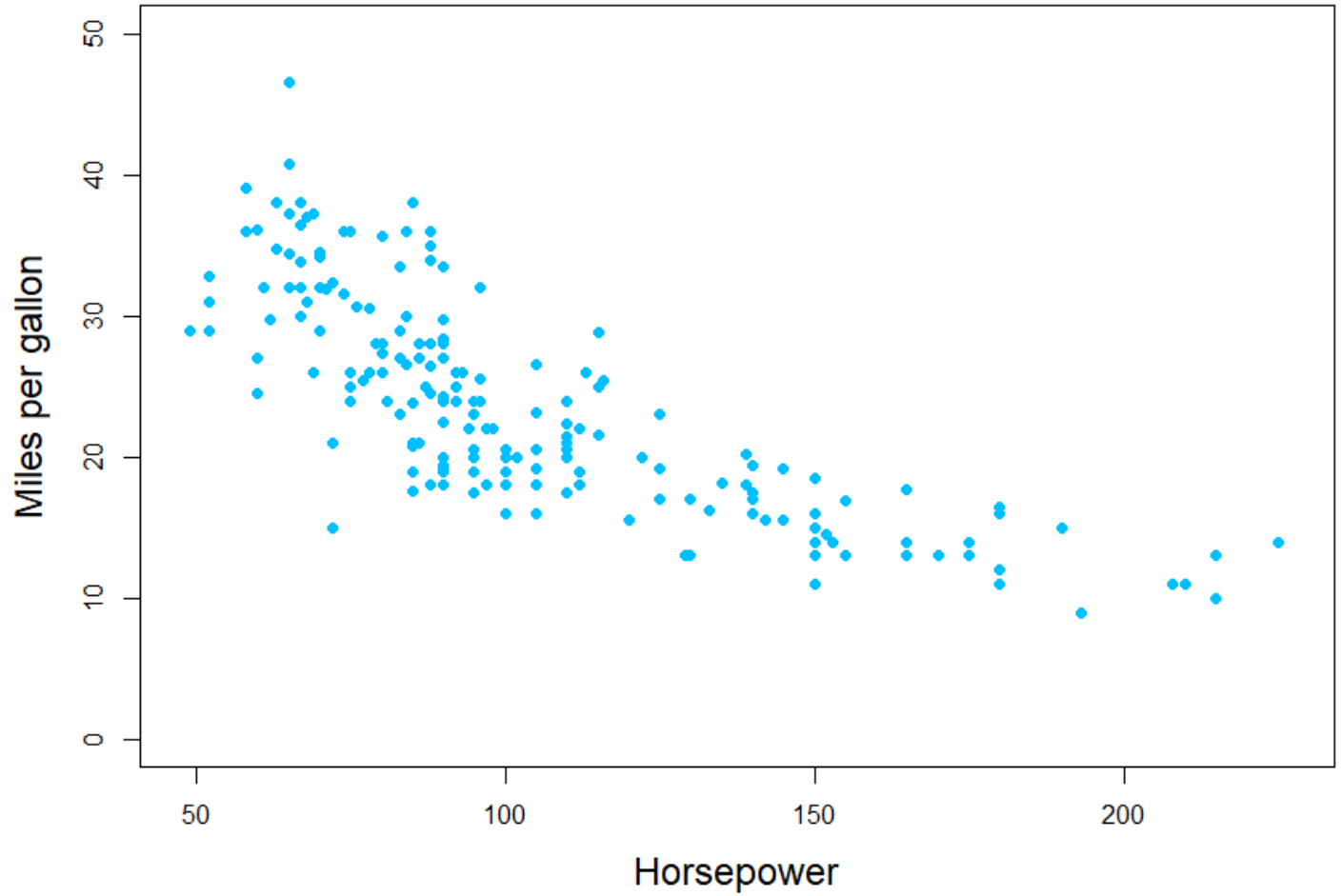
All observations



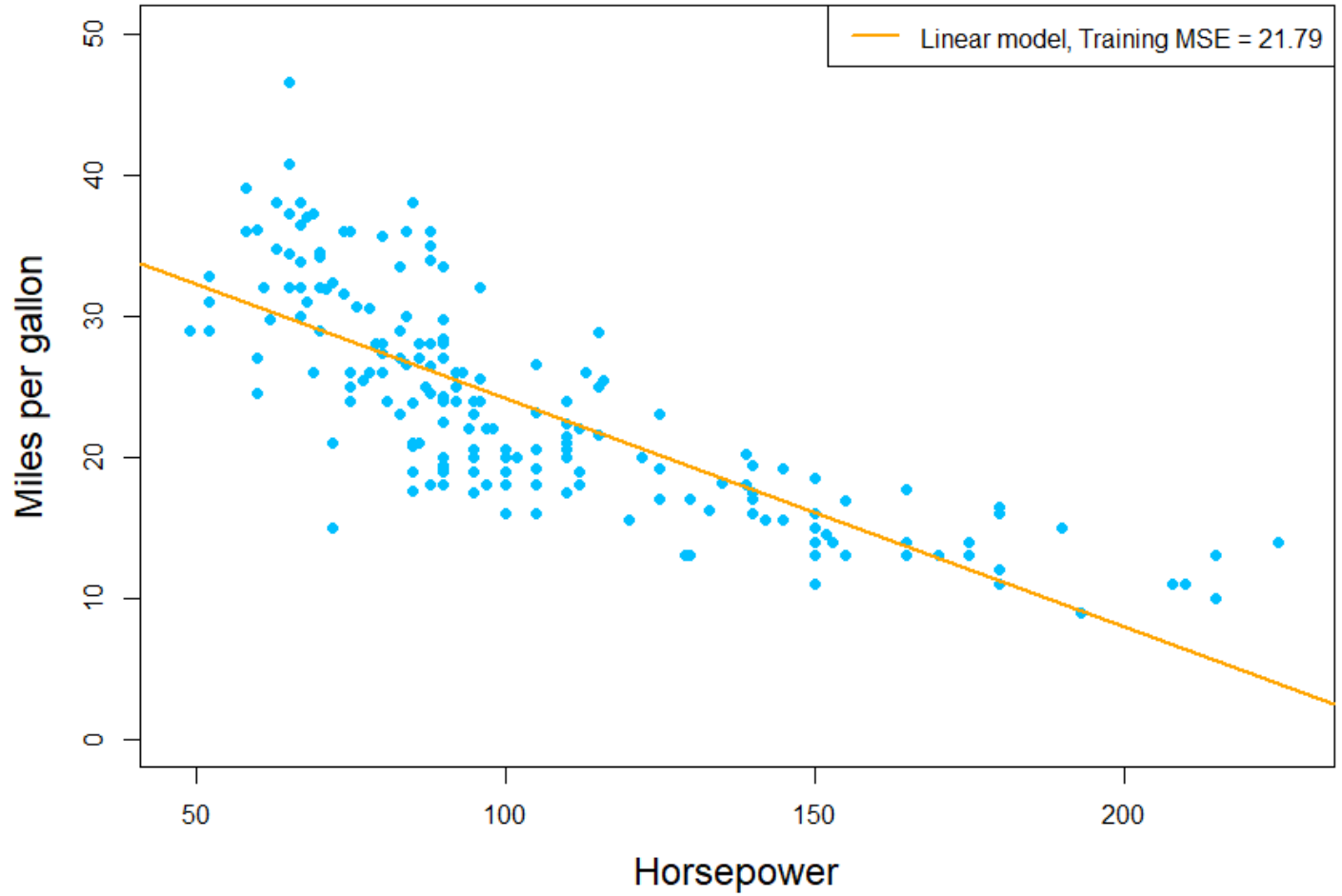
One possible split into a training set (blue points) and a test set (brown points)



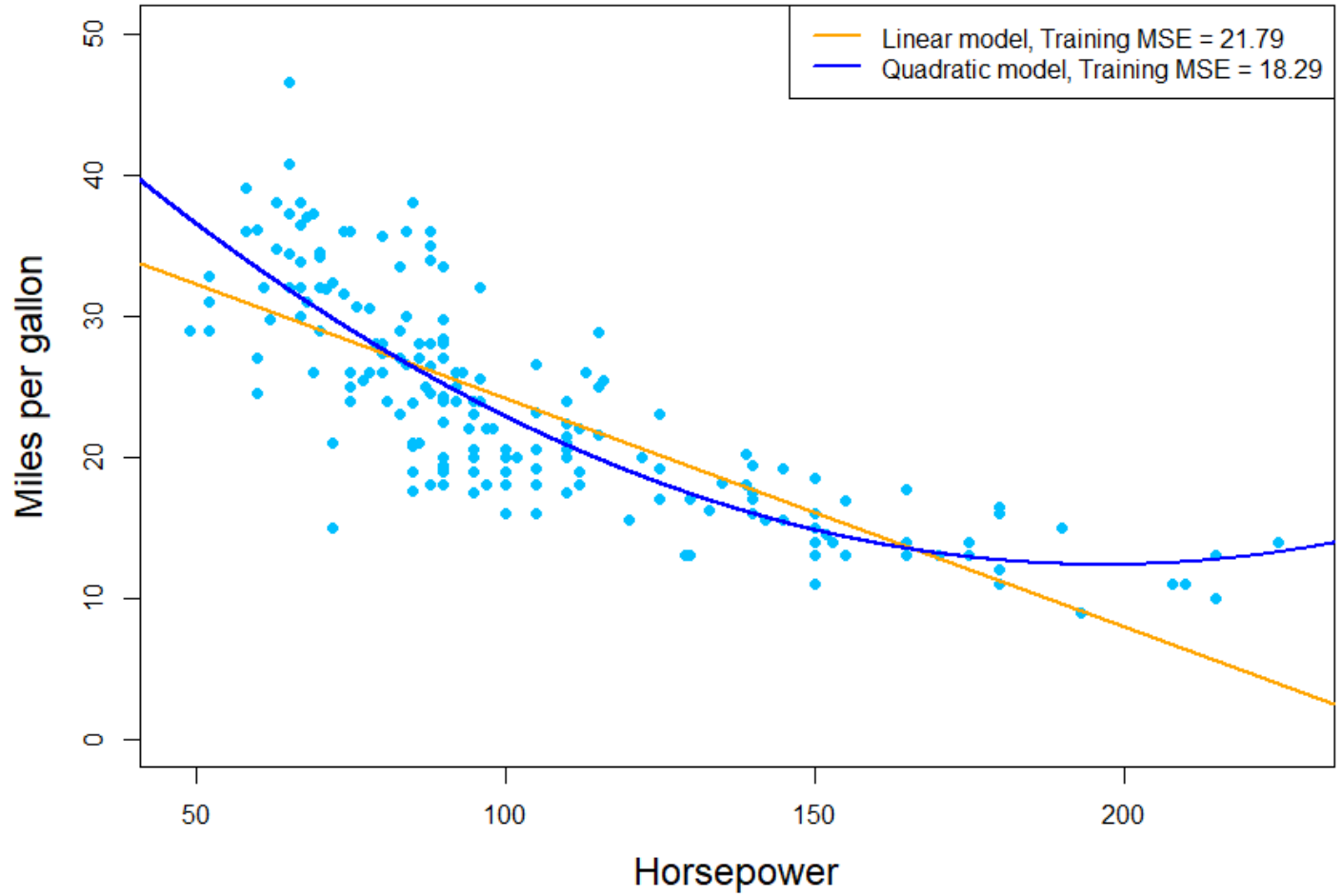
Step 2: Consider points in the training set



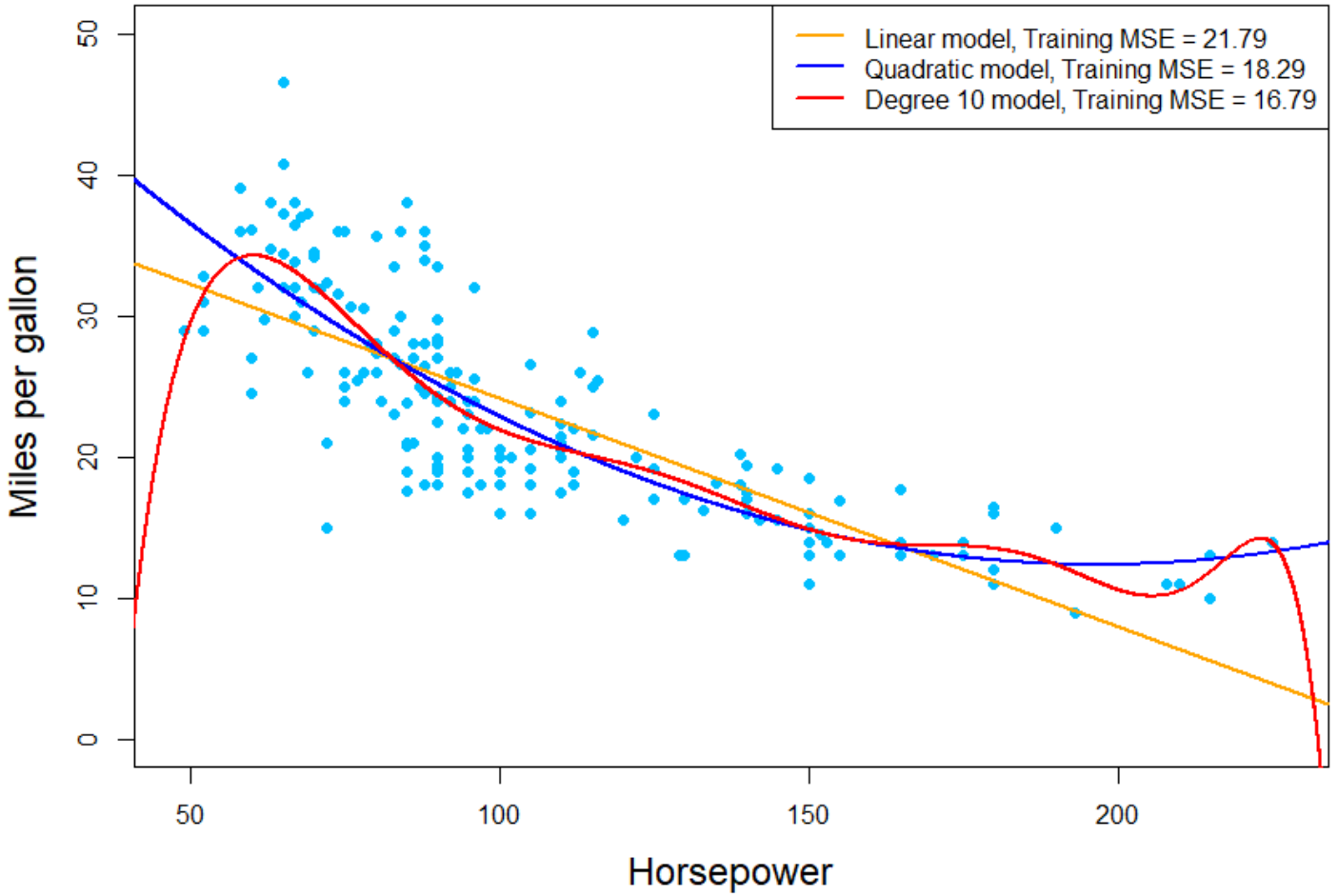
Step 2: Fit models on the training set (1)



Step 2: Fit models on the training set (2)



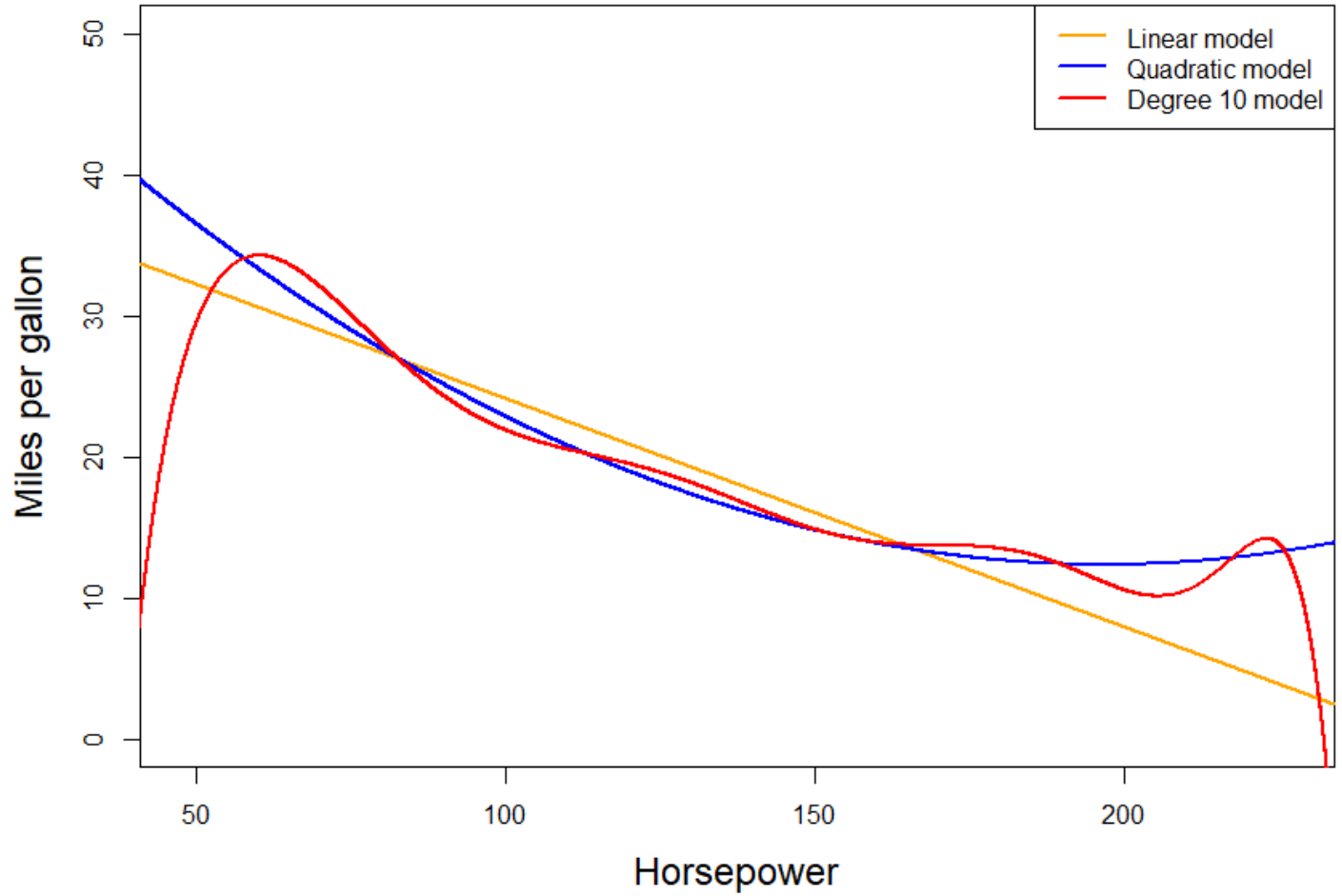
Step 2: Fit models on the training set (3)



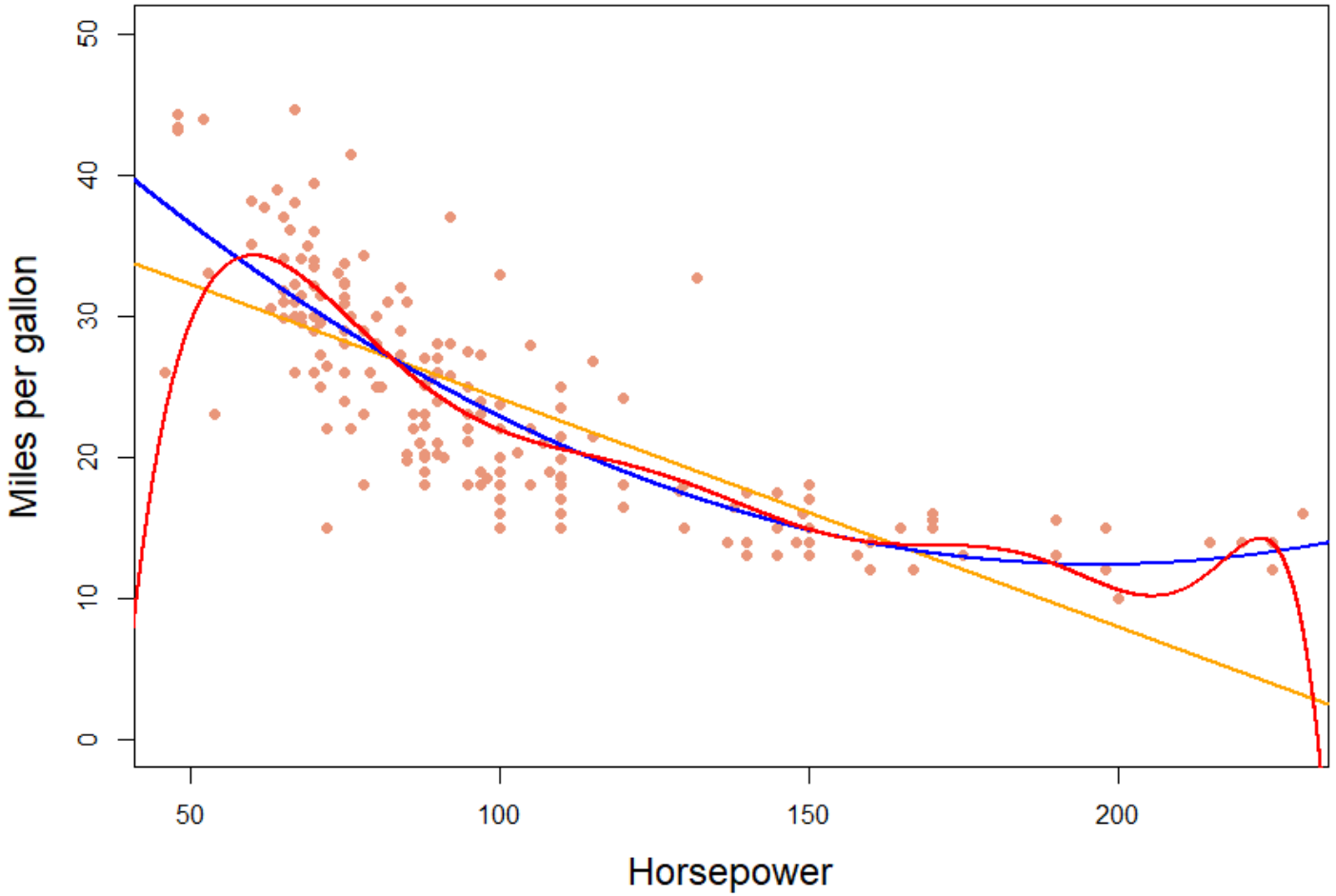
→ Training MSE is lowest for most flexible model

What really matters is not this, but test MSE

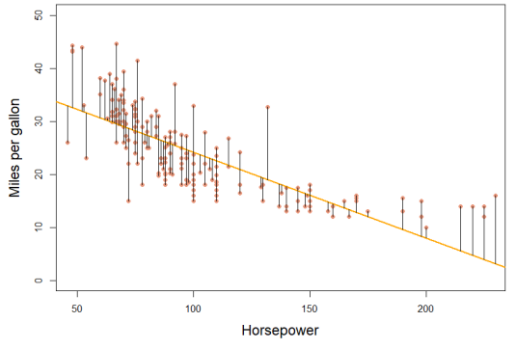
The models are defined – how good are they?

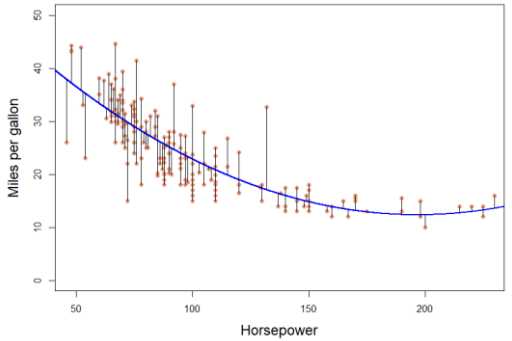


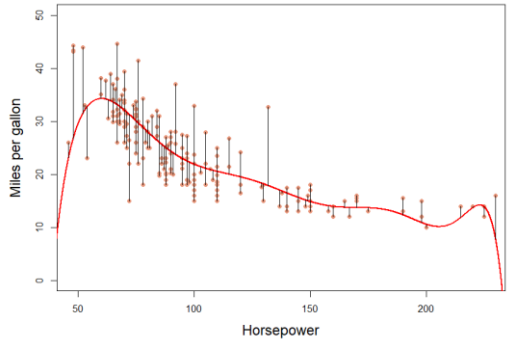
Step 3: Test models on the validation set

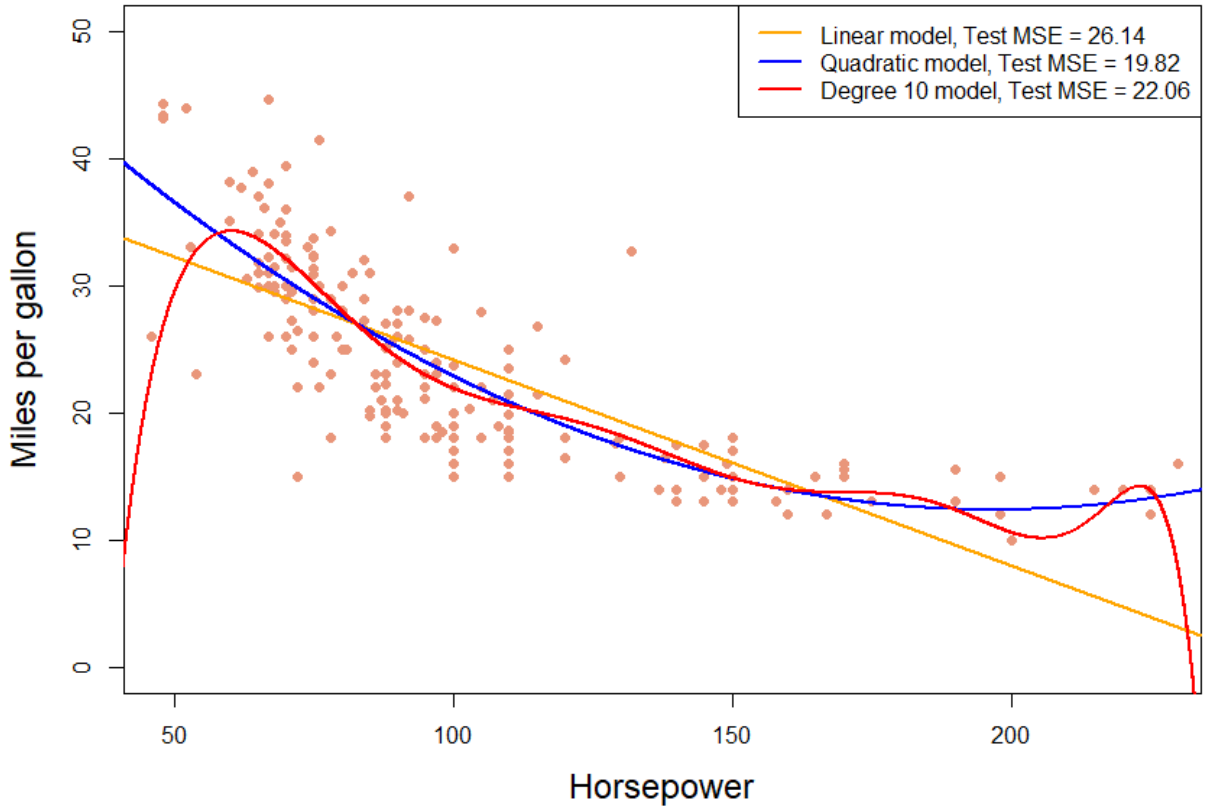


Step 3: Validation set MSE is error estimate

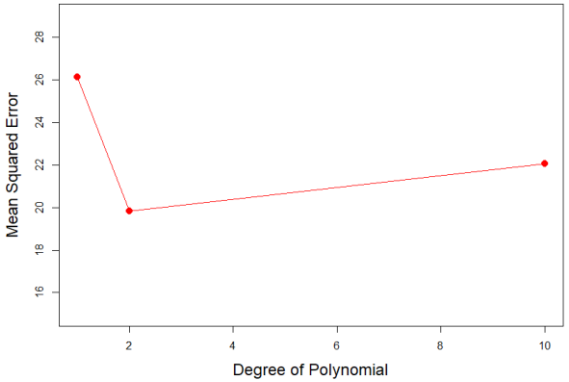








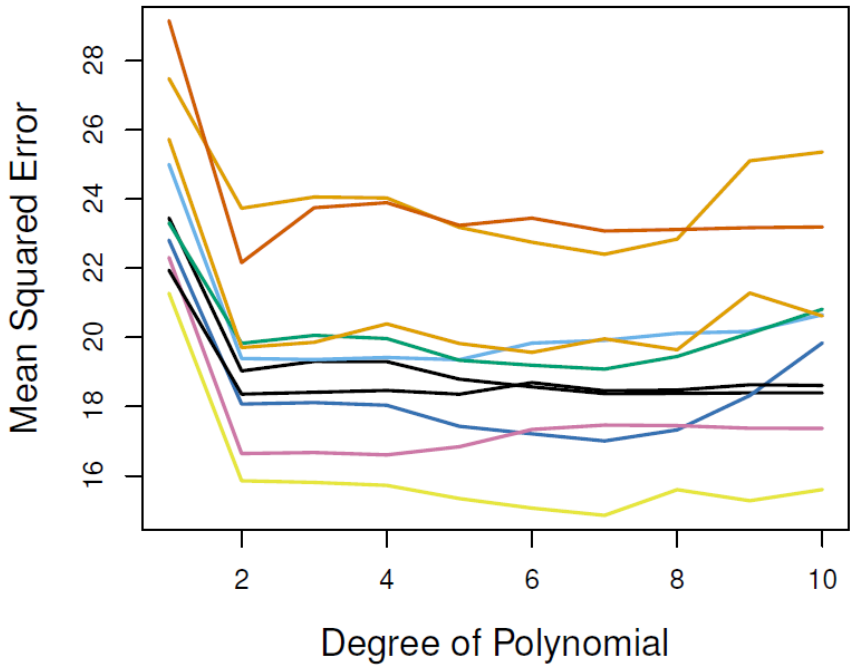
With this split of points, the test MSE for quadratic model is lowest of the three



Issue with validation set approach

- Results depend on how the division of observations into training set and validation set was made

→ Validation set estimate of test error is highly variable



Validation set test error estimates with 10 different splits, source: Figure 5.2 in ISL

Note: while the value of the MSE varies wildly, all divisions show some similar patterns:

- Degree 2 polynomial is better than degree 1 (i.e. quadratic model is better than linear)
- No large difference between error estimates for different degrees when using ≥2 degrees

→ These results support using quadratic model (if more complex model gives little gain, then choose simpler one for interpretability)

Methods estimating test error

Validation set approach

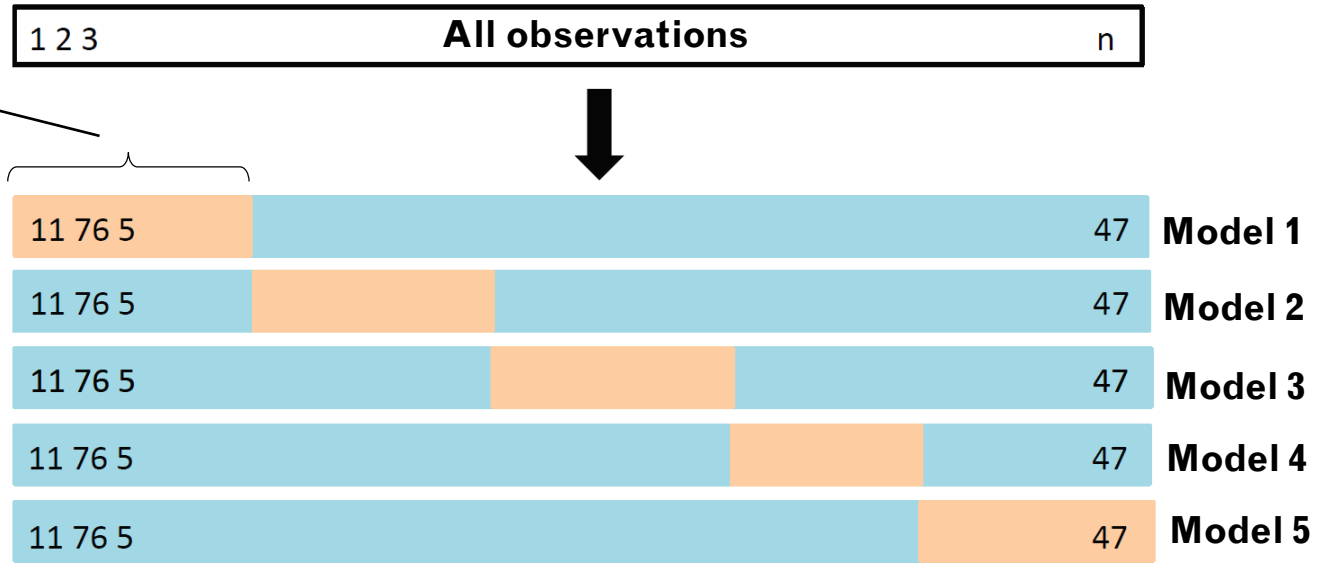
K-fold cross-validation

LOOCV

K-fold cross-validation: basic idea

- Divide the n observations into K equal parts (as equal as possible)
- Consider K different models, considering each part once as validation set and the other K-1 parts combined as training set

These observations are in:
- Validation set in Model 1
- Training set in Model 2
- Training set in Model 3
- Training set in Model 4
- Training set in Model 5



5-fold CV, source: Figure 5.5 in ISL, labels added

K-fold CV error estimates for mpg example

1. Divide the observations into K equal sets

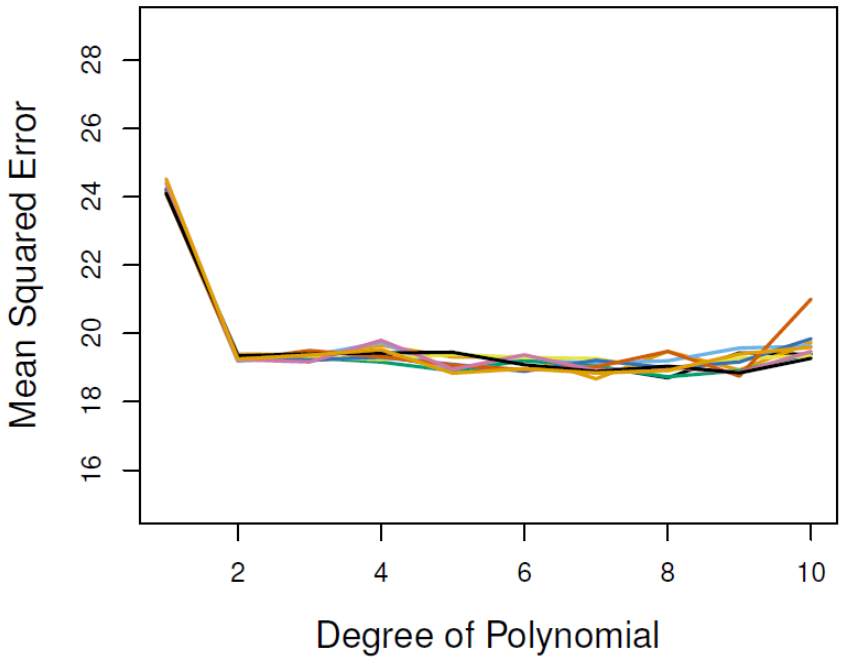
2. Changing the role of training set as shown on previous slide, fit a linear, quadratic and higher degree models K times

3. Compute MSE for each of the K linear models, K quadratic models, K higher degree models on the corresponding validation set (which is different for each of the K models)

4. Plot the average of the K error estimates, for each type of model (e.g. average of the K mean squared error values for linear models gives the MSE estimate at degree = 1 on next slide)

Less variablility in test error estimate

- Results depend somewhat on how the K folds were defined
- However, the variability in the estimate of test error is much smaller than it was with the validation set approach



10-fold CV test error estimates with
10 different splits, source: Figure 5.4 in ISL

In this case, the same patterns are even clearer:

- Degree 2 polynomial is better than degree 1 (i.e. quadratic model is better than linear)
- No large difference between error estimates for different degrees when using ≥2 degrees

→ 10-fold CV supports quadratic model

Methods estimating test error

Validation set approach

K-fold cross-validation

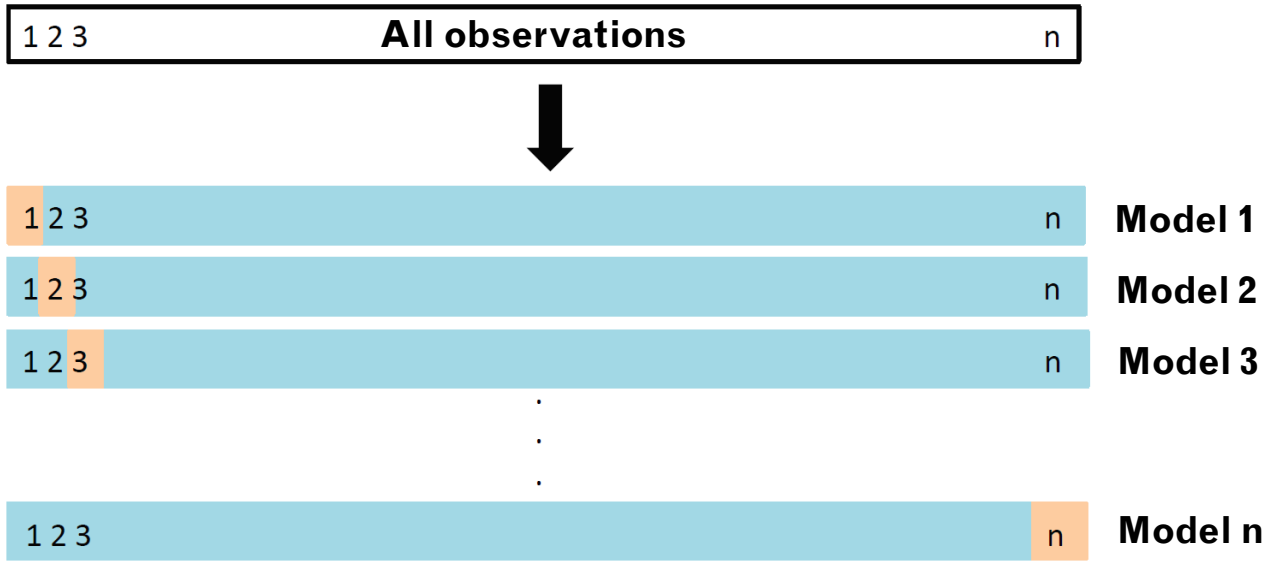
LOOCV

Special case: n-fold CV, called **LOOCV**

- What happens if we use n folds? Each fold consists of 1 observation

- In each of the n models in n-fold cross-validation:
  - The training set contains n-1 observations
  - The validation set is a single observation

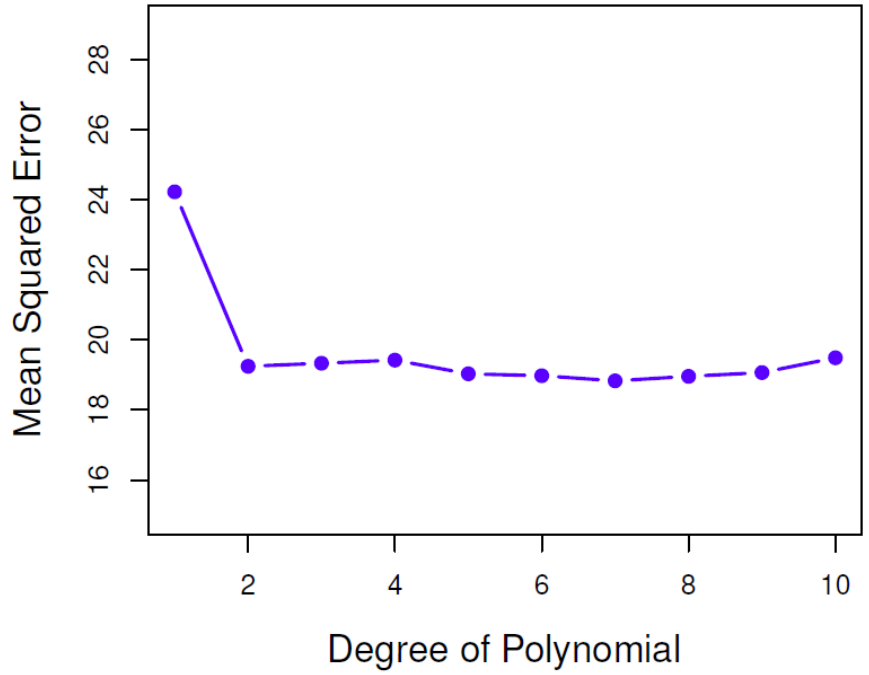
Leave-One-Out CV



LOOCV, source: Figure 5.5 in ISL, labels added

No variablility in test error estimate

- Only one way to do LOOCV: n steps, leave one observation out in each step, take average MSE → no variability in test error estimate



Leave-one-out cross-validation test error estimates, source: Figure 5.4 in ISL

Same patterns again:

- Degree 2 polynomial is better than degree 1 (i.e. quadratic model is better than linear)
- No large difference between error estimates for different degrees when using ≥2 degrees

→ LOOCV also supports quadratic model

Ethical analysis of big data

How to do analysis in an ethical way?

- Recommendations are made in the following article:

Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399

- For a full understading, read the article. Here: some highlights are given, with some own edits and additions

- Ethical aspects addressed in Assignment 3, but not in the exam

Data are people and can do harm

- Start with the assumption that all data are people until proven otherwise

- Apparently neutral data can lead to discrimination: categorization based on zip codes resulted in less access to Amazon Prime same-day delivery service for African-Americans in United States cities*

* Ingold D, Spencer S. Amazon Doesn't Consider the Race of Its Customers. Should It? Bloomberg.com, 21 April 2016. http://www.bloomberg.com/graphics/2016-amazon-same-day/. Accessed 12 June 2016

Privacy is more than a binary value

- Breaches of privacy are key means by which big data research can do harm
- Privacy is contextual* and situational** (e.g. just because something has been shared publicly does not mean any subsequent use would be unproblematic)
- Marketing based on search patterns have been perceived by some to be “creepy” or even outright breaches of privacy

* Nissenbaum H. Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press; 2009.
** Marwick AE. boyd d. Networked privacy: How teenagers negotiate context in social media. New Media & Society. 2014:1461444814543995.

Guard against data re-identification

- It is problematic to assume that data cannot be re-identified.
- When datasets thought to be anonymized are combined with other variables, it may result in unexpected re-identification
- Birthdate, gender, and zip code can identify people today*, but even seemingly harmless factors like battery usage may aid personal identification tomorrow**

*Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002;10(05):557–70.
** Michalevsky Y, Schulman A, Veerapandian GA, Boneh D, Nakibly G. Powerspy: Location tracking using mobile device power analysis.
In 24th USENIX Security Symposium (USENIX Security 15) 2015 (pp. 785–800).

Context & multiple meanings

Context affects every stage:

- data acquisition
- data cleaning
- interpretation of findings
- dissemination of the results

Context and evolution of data needs to be documented

Multiple meanings and uses:

- interpretation of those (re)using your data may differ from your own
- consider & describe potential multiple meanings (e.g. for models with confounding!)
- use clear & high quality figures*

→ Do not overstate clarity, acknowledge multiple meanings and uses
→ Document strengths & weaknesses of data and analysis

* See e.g. the following article for recommendations: Rougier, N.P., Droetboom, M., Bourne, P.E. (2014). Ten Simple Rules for Better Figures. PLoS Comput Biol. 10(9): e1003833, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4161295/

Long-term strategy

Discuss tough, ethical choices

- Make grappling with ethical questions part of standard workflow
- Ethics is often about finding a good or better, but not perfect, answer

Make code of conduct

- Address issues that might be ignored until they blow up
- Make researchers and representatives of affected communities active contributors

Auditability

- Plan for & welcome audits of big data practices
- Explicit about decisions → understandability and replicability