Statistical Learning and Prediction

Marjolein Fokkema

Methodology and Statistics Unit
Leiden University
The Netherlands

m.fokkema@fsw.leidenuniv.nl

This Course

- New methodology for data analysis
- Same models, but different focus
- Machine Learning/Computer Science
- Statistics and Machine Learning: Statistical Learning
- New name: Data Science

Five Sessions

Day 1:

S. 1 Intro, Bias-Variance Tradeoff, Prediction (Chapters 1, 2, 3)

S. 2 Resampling (Chapter 4, 5)

S. 3 Regularization (Chapter 6)

Five Sessions

Day 2:

S. 4 Unsupervised Learning (Chapter 10)

S. 5 Beyond Linearity and Main Effects:

- Generalized Additive Models / Splines (Chapter 7)
- Decision Trees and Ensembles (Chapter 8)
- Support Vector Machines (Chapter 9)

Statistical Learning

- Statistical learning refers to a vast set of tools for understanding data.

  - Supervised: $Y \leftarrow f(X_1, \ldots, X_p)$; predict Y on the basis of X

  - Unsupervised: $X_1, \ldots, X_p$; finding structure (underlying dimensions/groups)

Statistical Learning

- The linear regression model

$$Y = f(X_1, \ldots, X_p) = a + b_1 X_1 + \cdots + b_p X_p$$

  can be used for explanation and/or prediction:

- Explanation: understanding how the X's are related to the Y; possibly causally.

- Prediction: if we have new observations with known values of the X's, what is the expected (predicted) value of Y and how accurate are these predictions?

Explanatory Regression

- Suppose we have data and obtained estimates:

$$\hat{y}_i = 2 + 0.5x_{i1} + 1.5x_{i2}$$

  with both coefficients statistically significantly different from 0.

  - Estimated coefficients indicate the magnitude of the effects.
  - Standard errors idicate the variability of the estimated coefficients.
  - Statistical tests indicate whether the explanatory variables really affect the response, given the other variables.
  - Adequate estimation of the coefficients (b_1 and b_2) is crucial: Accurate estimates = unbiased estimates!

Predictive Regression

- Suppose we have data and obtain estimates:

$$\hat{y}_i = 2 + 0.5x_{i1} + 1.5x_{i2}$$

  - Suppose we have a new observation with $X_1 = 2$ and $X_2 = 3$
  - With these values we can predict Y: $2 + 0.5 \times 2 + 1.5 \times 3 = 7.5$
  - Prediction focusses on accuracy of $\hat{y}$. We do not need to recover parameters that generated the data, but merely obtain a model that yields as accurate as possible $\hat{y}$.

Regression

More general, consider we have a population and within this population the conditional means of the response variable ($Y \in \mathbb{R}$) are given by a function of the predictor variables ($X \in \mathbb{R}^p$), that is

$$Y = f(X) + \epsilon.$$

Often we collect data for a sample of n persons. These data are given by $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ used to <u>train</u> a model

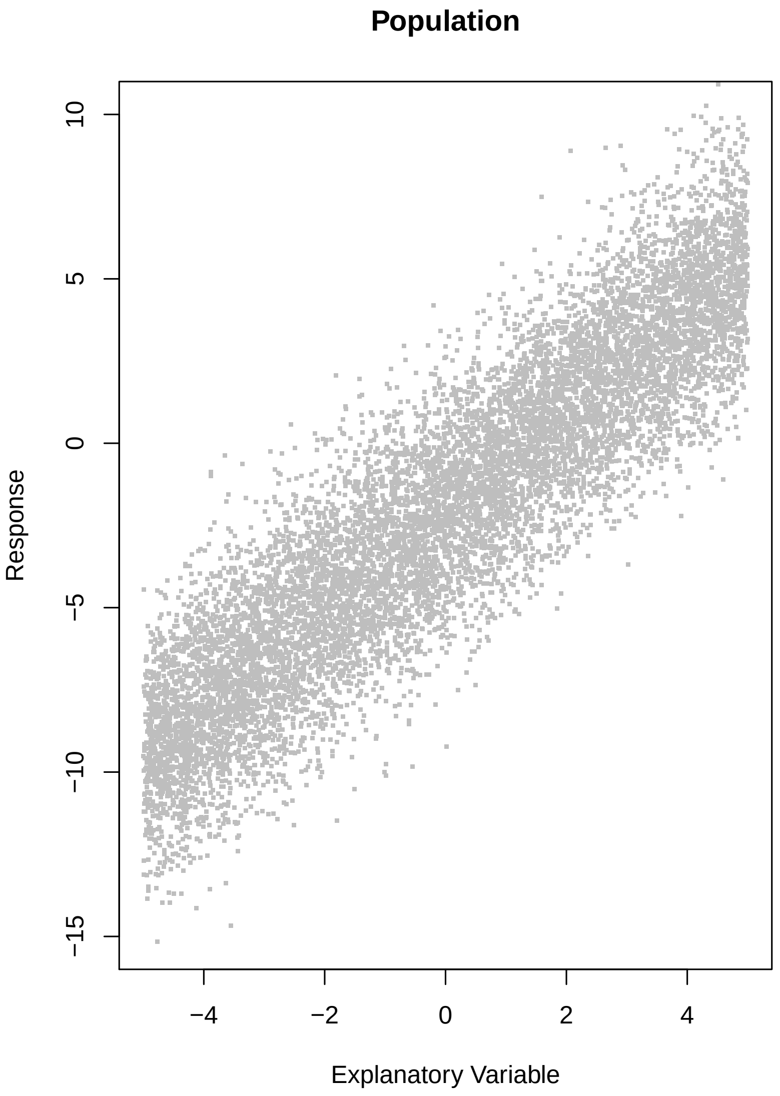
$$y_i = \hat{f}(x_i) + \epsilon_i$$

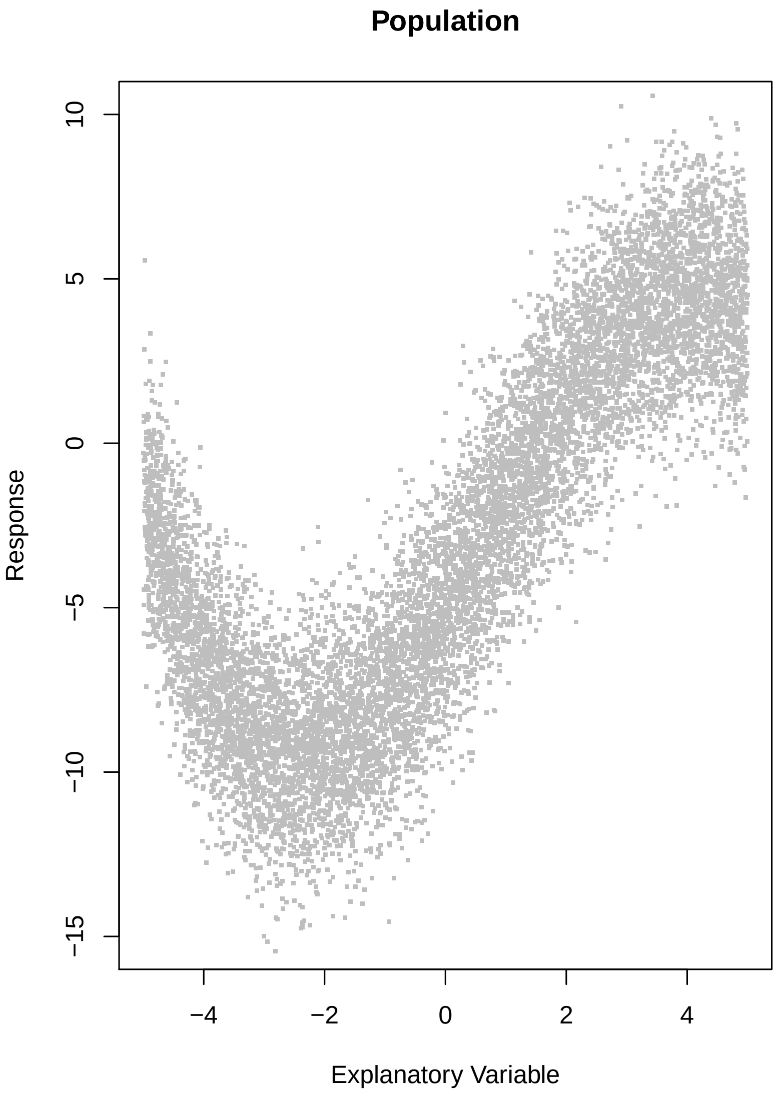
Regression

The function $f(\cdot)$ returns a single value for any given $\mathbf{x}$ but can be any function, for example

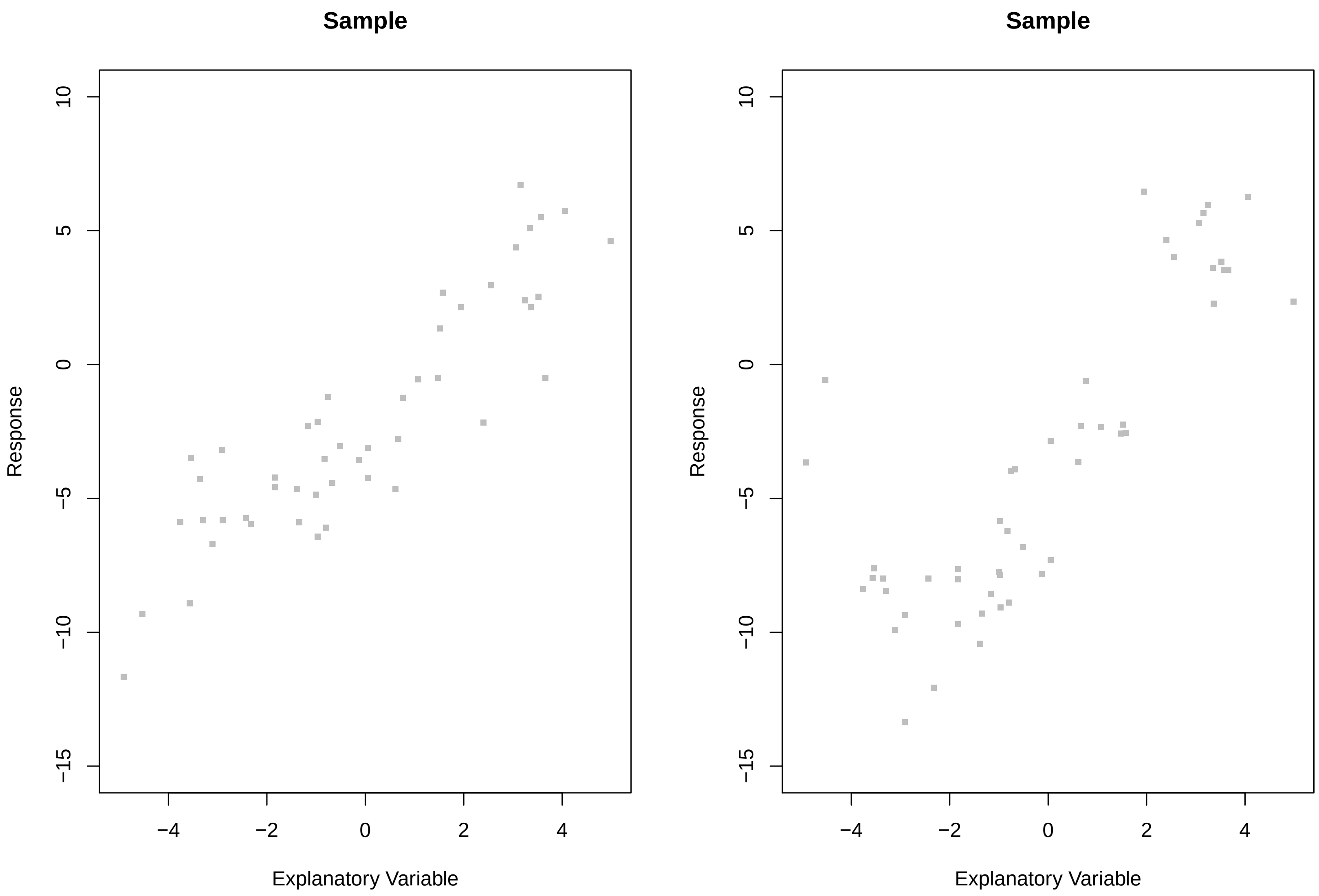
- Straight lines
- Polynomials
- Smooth lines
- Step functions
- Smooth functions with steps
- ......

Populations





Samples; $n = 50$


Sample
Response
10
5
0
−5
−10
−15
−4
−2
0
2
4
Explanatory Variable
Sample
Response
10
5
0
−5
−10
−15
−4
−2
0
2
4
Explanatory Variable

Predictive Regression

Suppose we have a training dataset of size n, to which we apply an algorithm to obtain an $\hat{f}(x)$ (e.g., OLS regression or kNN with $k = 5$).

The expected prediction error we would make on a new observation $(\mathbf{x}_0, y_0)$, is given by:

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] \;=\; \sigma^2 + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \mathrm{Var}\left(\hat{f}(x_0)\right)$$

where the expectation is taken over all possible values $\mathbf{x}_0$ and all possible training sets of size n.

(Derivation is beyond scope of this course, but provided here)

Predictive Regression

The prediction error of our model-fitting procedure thus decomposes into

- an irreducible term,
- the squared bias,
- and the variance.

Traditional statistical textbooks focus on obtaining unbiased coefficient estimates. This unbiasedness comes at the price of increased variance.

Modern statistical learning sometimes trades in some bias against variance. That is, we accept biased estimates as long as the variance decreases more than the squared bias increases.

Predictive Regression

- Conditional means in the population:

$$Y = a + bX + e$$

- In this case we have a univariate regression.

- Linear regression estimated using OLS gives unbiased estimates, $\hat{a}$ and $\hat{b}$.

- QUESTION: *Can we get better predictions with biased estimates?*

Experiment: Predictive Regression

I repeatedly generated $n = 50$ training observations using

$$y = bx + e$$

where x comes from a uniform distribution with min = -5 and max = 5; e is drawn from a standard normal distribution. I varied the ’true’ effect $b \in (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$.

I generated test observations from the same population.

On the training I fitted a univariate OLS regression, providing an estimated regression coefficient, $\hat{b}$.

QUESTION: *what do we know about* $\hat{b}$*?*

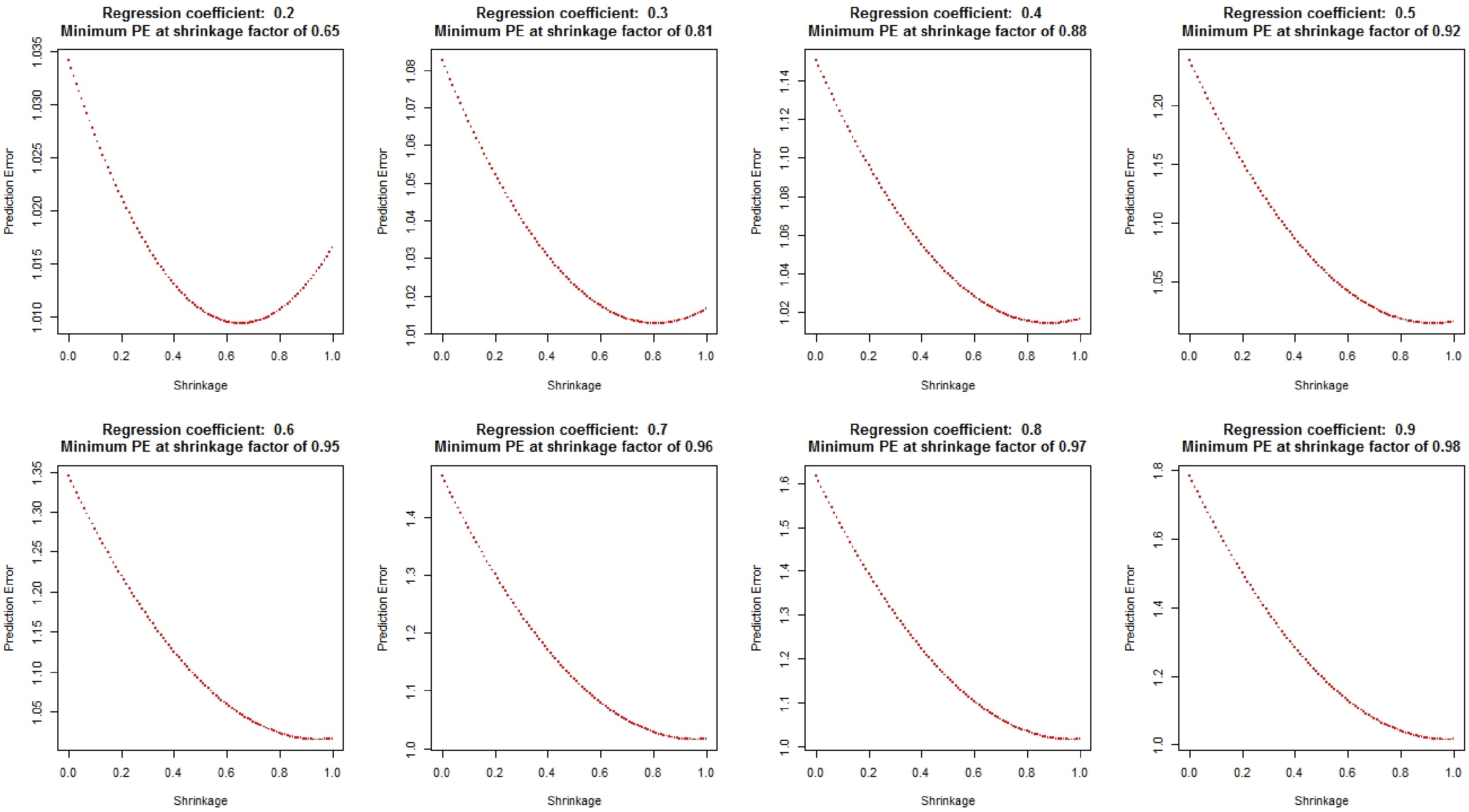
Experiment: Predictive Regression

I took the OLS estimated coefficient and applied a shrinkage factor $c \in (0, 0.1, \ldots, 0.9, 1)$, to compute predictions:

$$\hat{y} = x * c * \hat{b}$$

Then I compared predicted $(\hat{y})$ with observed response values (y) . I averaged over 1,000 replications:

Experiment: Results for training sample size $n = 50$

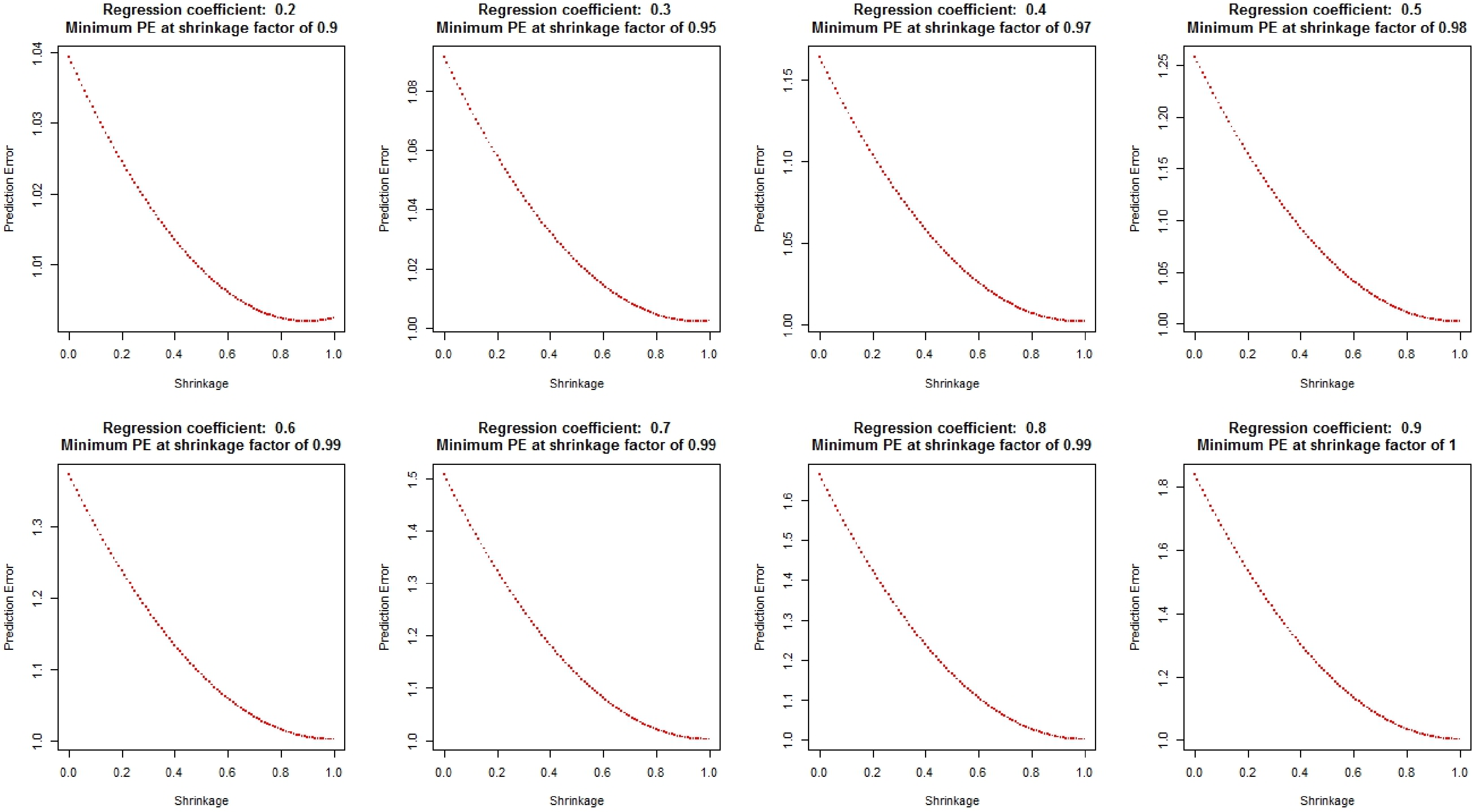


Experiment: Predictive Regression

We see that shrinkage (a biased coefficient) is beneficial for prediction. For lower effect sizes, stronger shrinkage is optimal.

Now we repeat the same experiment with a larger training sample:

Experiment: Results for training sample size $n = 200$



Regression coefficient: 0.2
Minimum PE at shrinkage factor of 0.9

Regression coefficient: 0.3
Minimum PE at shrinkage factor of 0.95

Regression coefficient: 0.4
Minimum PE at shrinkage factor of 0.97

Regression coefficient: 0.5
Minimum PE at shrinkage factor of 0.98

Regression coefficient: 0.6
Minimum PE at shrinkage factor of 0.99

Regression coefficient: 0.7
Minimum PE at shrinkage factor of 0.99

Regression coefficient: 0.8
Minimum PE at shrinkage factor of 0.99

Regression coefficient: 0.9
Minimum PE at shrinkage factor of 1

Experiment: Predictive Regression

With larger sample size, shrinkage can still be beneficial for prediction, but lower amount of shrinkage is optimal than with smaller sample size.

Amount of shrinkage required for optimal prediction accuracy depends on amount of information present in sample:

- Effect size (stronger effect provides more information; i.e., larger value of b, or lower variance of e).

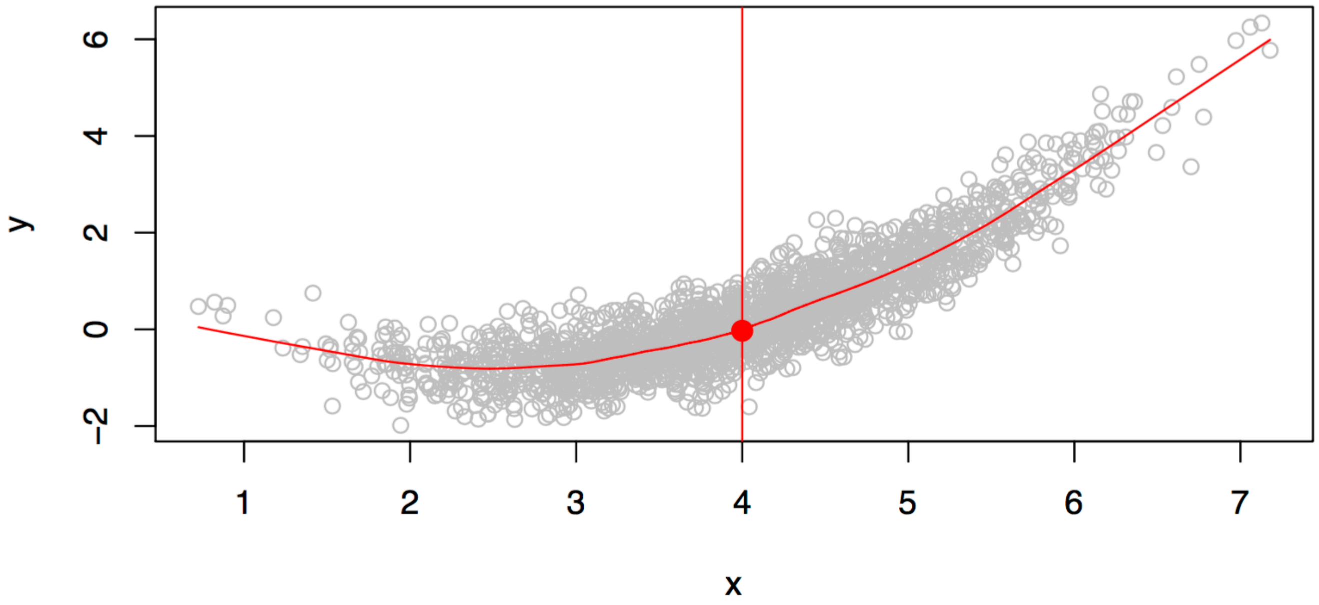
- Sample size (larger sample size provides more information).

Population: Nonlinear Regression

Often we fit a linear model, assuming that the conditional means in the population lie on a straight line.

This assumption is most likely false!

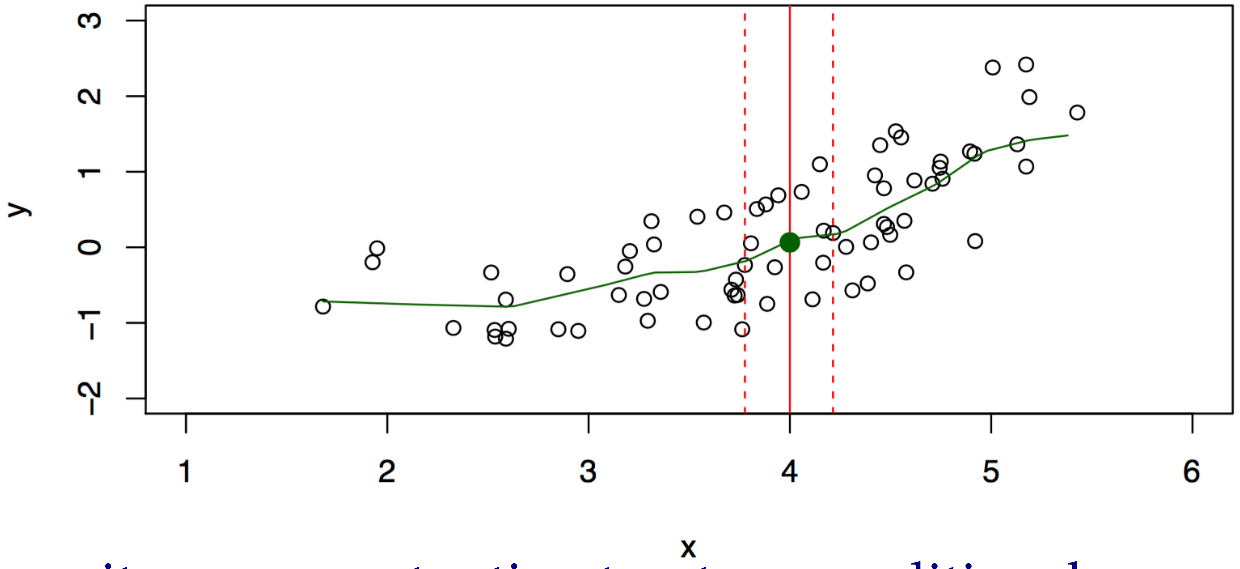
Population: Nonlinear Regression



The regression line in the population combines the conditional means at each point x

Regression: Some data

We want to get an estimate $\hat{f}(X)$ of $f(X)$.



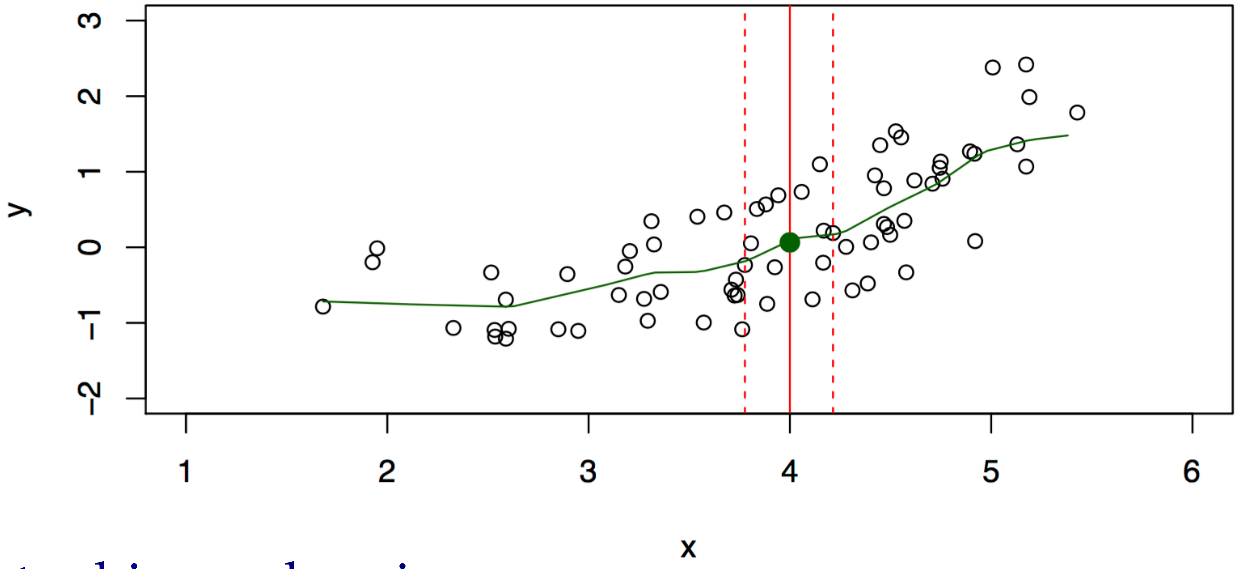
Due to sparsity we cannot estimate a true conditional mean at all values of X.

We can take a small neighbourhood around $X = x$ and take mean in that neighbourhood as the predicted value at $X = x$: *nearest neighbour averaging*.

This creates bias and variance.

Regression: Some data

We want to get an estimate $\hat{f}(X)$ of $f(X)$.

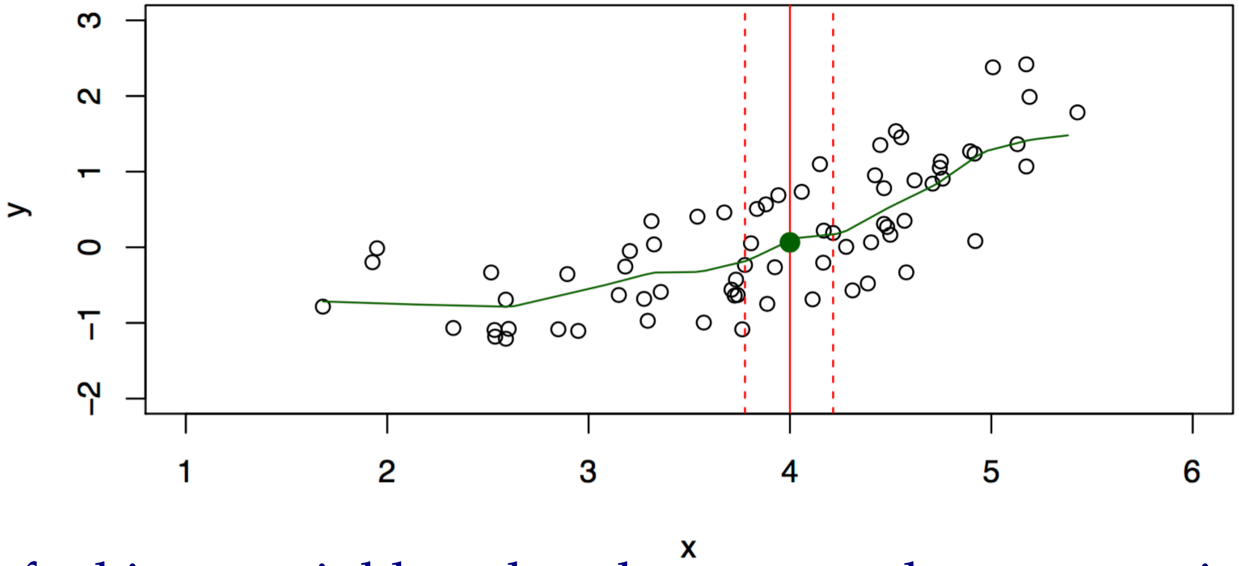


This creates bias and variance:

- The larger the neighbourhood, the larger the bias
- The larger the neighbourhood, the smaller the variance

Regression: Some data

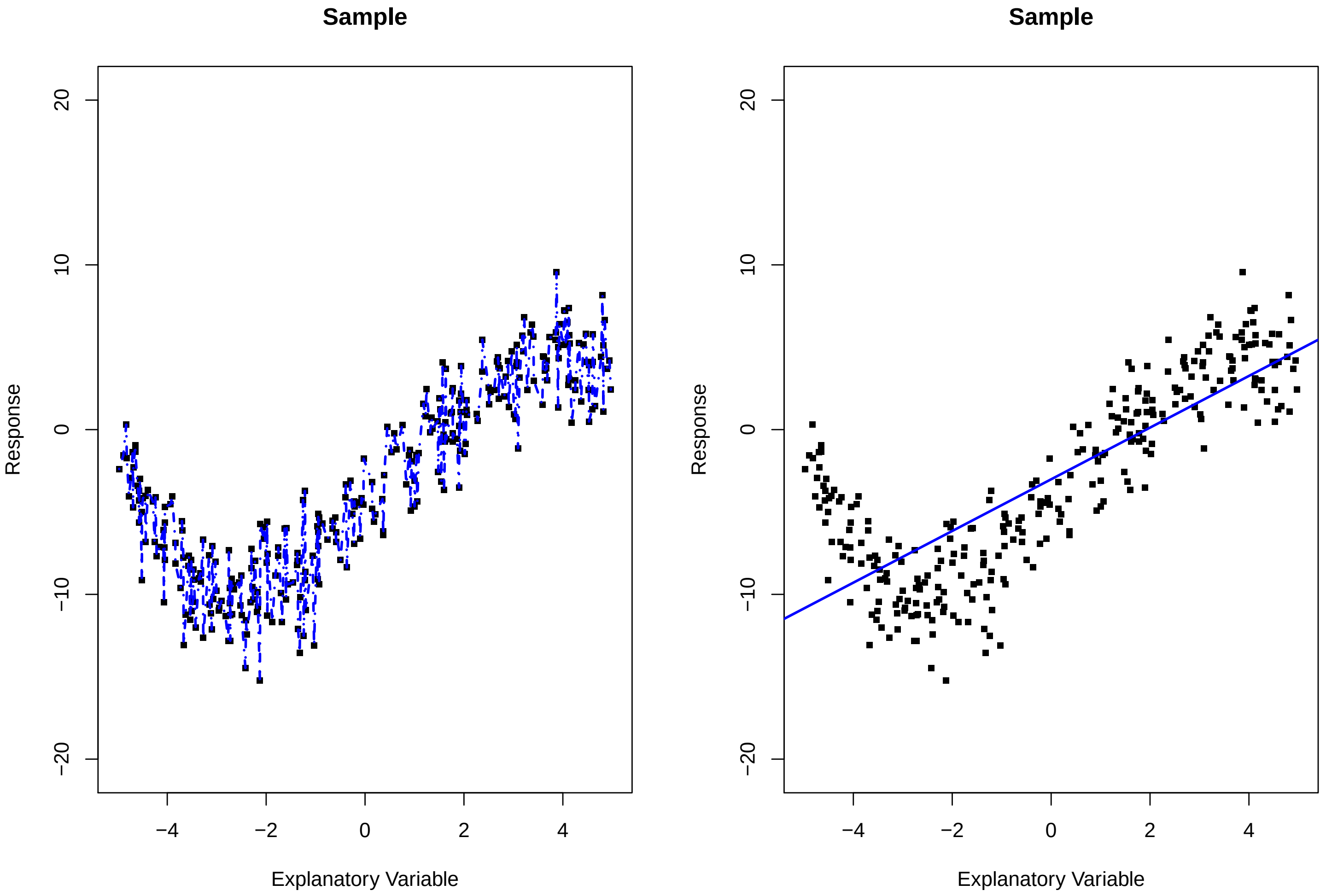
We want to get an estimate $\hat{f}(X)$ of $f(X)$.



Instead of taking a neighbourhood we can make parametric assumptions (for example: linear association), this also creates bias.

QUESTION: *What about the variance?*

Nearest Neighbour *vs* Linear Regression



Exercises

- Exercises from book (Chapter 2, page 52-53): Exercise 1

- Exercise Nonlinear Regression:

Generate a training and test set (size 100) of data following with a single predictor X (uniformly distributed from -5 till 5) following

$$f(x) = x + 8\sin(x/2) + \epsilon$$

with $\sigma_\epsilon \sim N(0,1)$.

Fit polynomial regression models to the training data of degree 1 to 15, make predictions on the test set and compute the prediction error for each degree.

Make a plot with the degree of the polynomial against the prediction error in the test set.

Exercises

Hints:

- Make sure to use the set.seed() function, first.
- Use functions runif() and rnorm() to generate the data.
- Use functions lm() and poly() to fit a linear regression model with polynomial terms to the training data.
- To obtain predictions, apply function predict() to the fitted model, and supply the test data to the newdata argument.
- To compute prediction error on training data:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Use a for loop to fit polynomials of order 1 to 15.

Multiple Predictor Variables

- With multiple predictors the observations are further spread out through the space

- The three nearest neighbours might not be near at every point

- Then flexible regression models become very wild

- This is known as the *curse of dimensionality*

- More structure in f is needed

- QUESTION: *How can we impose structure?*

Classification

- Y is a categorical variable with categories $\mathcal{C} = \{1, \ldots, K\}$
- Again we have predictors X
- Build a classifier $C(X)$ that assigns a class label of $\mathcal{C}$ to future unlabeled observations X.
- Conditional probabilities

$$p_k(X) = Pr(Y = k|X = x), k = 1, \ldots, K$$

- *Bayes Classifier* assigns

$$C(X) = j \text{ if } p_j(x) = \max\{p_1(x), \ldots, p_K(x)\}$$

Classification

- Again, with data we cannot estimate the probability at all possible values of X.

- Take a neighbourhood.

- K nearest neighbour (KNN) analysis: take the K nearest points and classify using a majority vote.

- For classification the error is usually measured using the misclassification error rate:
$$\mathrm{Err}_{\mathrm{Te}} = \mathrm{Ave}_{i \in \mathrm{Te}}(y_i \neq \hat{C}(x_i))$$

- This quantifies the accuracy of the class predictions. If a method returns predicted probabilities, average squared error loss (c.f., MSE for continuous responses) on predicted probabilities may be a more appropriate measure of accuracy.

Exercises

- What are the similarities and differences between nearest neighbour regression and nearest neighbour classification?

- Exercises from book (Chapter 2, page 52-53): Exercise 7