Winter Course Statistical Learning
Session 1: Introduction

Session 1 Topics

- Supervised versus unsupervised
- Explanation versus prediction
- Bias, variance and irreducible error
- k Nearest Neighbors (kNN)
- Curse of dimensionality

Statistical Learning

Vast set of tools for understanding data.

- Supervised: $Y \leftarrow f(X_1, \dots, X_p)$; predict Y on the basis of the Xs.
    - Sessions 1, 2, 3 and 5 of this course.
- Unsupervised: $X_1, \dots, X_p$; finding structure in the Xs (underlying dimensions/groups).
    - Session 4 of this course.

Explanation versus Prediction

The linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

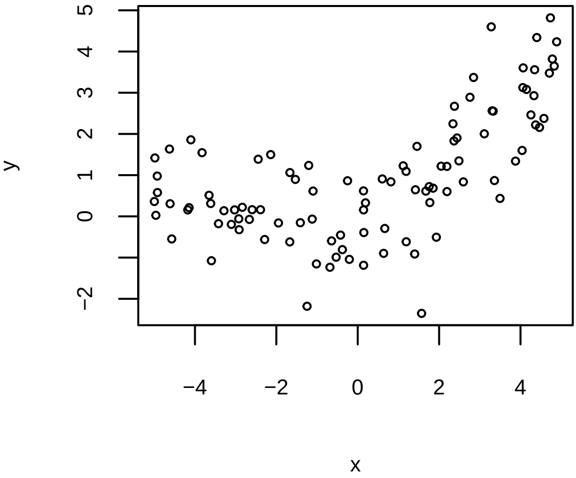
can be used for explanation and/or prediction.

- Explanation: Understanding how the Xs are related to Y (possibly causally).
- Prediction: If we have new observations with known values of the Xs, what is the expected (predicted) value of Y and how accurate are these predictions?

A sample of data

```r
set.seed(42)
n <- 100
x <- runif(n, min = -5, max = 5)
y <- 0.25*x + 0.1*x^2 + rnorm(n)
plot(x, y, cex.lab = .7, cex.axis = .7, main = "")
```

A sample of data



First course in regression: Explanation focus

```r
lmod <- lm(y ~ x)
summary(lmod)
```

- What is the direction and strength of the effect of predictor(s)?
- How uncertain is the estimated effect?
- Is the effect significant?

First course in regression

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6938 -0.7524  0.0120  0.7382  2.7491 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.87152    0.12292   7.090 2.09e-10 ***
x            0.29819    0.04078   7.313 7.20e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 1.225 on 98 degrees of freedom
Multiple R-squared:  0.353, Adjusted R-squared:  0.3464 
F-statistic: 53.47 on 1 and 98 DF,  p-value: 7.199e-11
```

Explanatory focus in regression

- Adequate estimation is assumed to be *unbiased* estimation:

$$\mathbb{E}[\hat{\beta}] = \beta$$

.

- Estimated coefficients are accurate *on average* (over many replications of collecting a sample and fitting the model).
- Ordinary Least Squares (or maximum likelihood) yields unbiased estimation:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{N} (y_i - x_i^\top \beta)^2 \right)$$

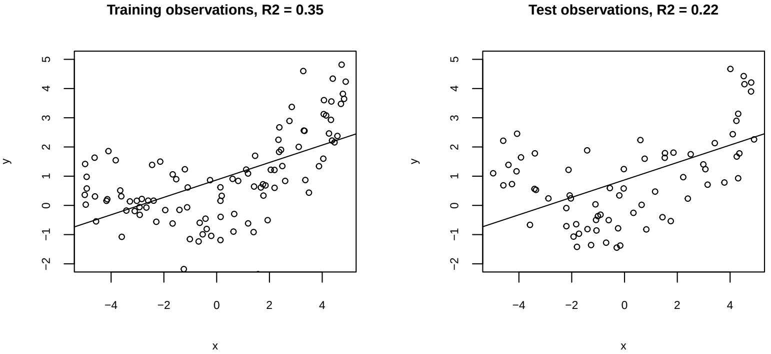
- Note that squared errors are minimized on *training* observations.

This course: From explanatory to predictive focus

- Unbiased estimation is optimal if:
    - we have huge samples (very rare in science),
    - or very many samples (only in meta analysis),
    - or perfect predictions are possible (extremely rare in science).
- In behavioral sciences, the number of observations tends to be limited, we only have a single dataset, predictions will always be (far from) perfect (random noise and measurement error).

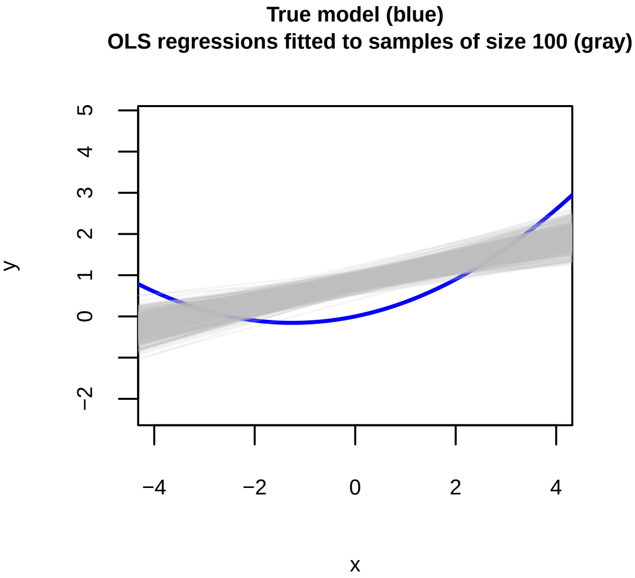
This course: Prediction focus in regression

- How well does the fitted model predict on *new* observations from the same population?



- Fit on training observations is overly optimistic.
  - Session 3 (penalized regression): Do not perfectly minimize errors on training observations, but add a penalty.

Repeat sampling and model fitting 500 times



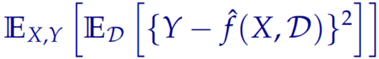
Minimizing the expected prediction error

We consider the performance of a statistical method, repeatedly applied to data samples from the same population P^*.

We'd like to find the method that is expected to provide the best prediction model $\hat{f}(X, \mathcal{D}) = \hat{Y}$,

where $\mathcal{D}$ is a training dataset of n observations drawn from the population $(X, Y) \sim P^*$.

The Expected Prediction Error (EPE) is:



$$\mathbb{E}_{X,Y}\left[\mathbb{E}_{\mathcal{D}}\left[\{Y - \hat{f}(X, \mathcal{D})\}^2\right]\right]$$

This EPE is a somewhat theoretical quantity, we cannot compute it in practice, only estimate it.

Decomposing EPE into Bias, variance, irreducible error

EPE $=$ Bias^2 $+$ Variance $+$ Irreducible Error. Let:

$f(X)$ be the 'true' model (unknown but given by P^*),

$\hat{f}(X;\mathcal{D}) = \hat{Y}$ be the model-fitting procedure applied to a single dataset $\mathcal{D}$ (i.e., a predictive model fitted to $\mathcal{D}$),

$\bar{f}(X) = \mathbb{E}_{\mathcal{D}}[\hat{f}(X)]$ be the average fitted model over many repetitions of sampling a dataset $\mathcal{D}$ and applying the method to it.

Decomposing EPE into Bias, variance, irreducible error

Then it follows:

- Bias$^2 = \mathbb{E}_X\left[\{\bar{f}(X) - f(X)\}^2\right]$, the average squared difference between the true model and the average fitted model,
- Variance $=$ $\mathbb{E}_X\left[\mathbb{E}_\mathcal{D}\left[\{\hat{f}(X;\mathcal{D}) - \bar{f}(X)\}^2\right]\right]$, the average squared difference between individual iterations of drawing a sample and fitting a model, and the average fitted model,
- Irreducible Error $= \mathbb{E}_{X,Y}\left[\{Y - f(X)\}^2\right] = \sigma_\epsilon^2$, the average squared difference between the true model and data points.

Informally, more bias is less flexibility, yielding higher variance.

Irreducible error is a property of the chosen population (data problem), beyond our control.

We must choose (squared) bias and variance so that their sum is minimized (thus EPE is minimized).

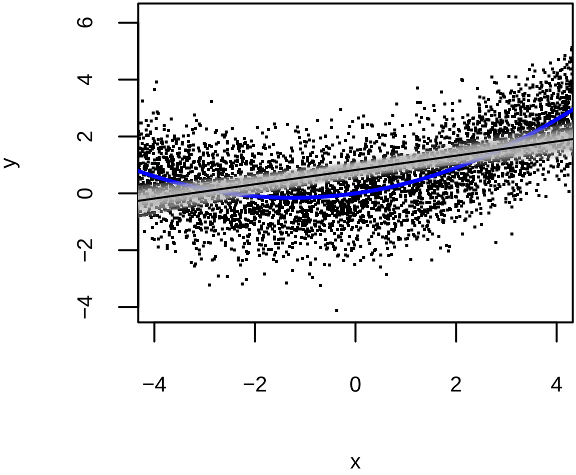
Estimating expected prediction error

In practice, we *estimate* EPE using test observations:

$$\mathsf{MSE}_{test} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2$$

- Model with lowest test MSE generalizes best and should (in most cases) be preferred.
- Other loss / error functions can also be used to select the best model, such as mean absolute error (but not easily decomposable into bias, variance and irreducible error).

Bias, variance, irreducible error: Example



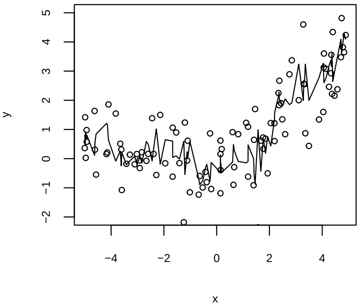
Bias, variance, irreducible error: Example

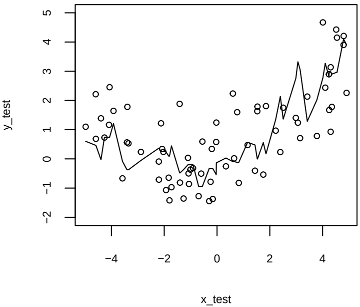
- Bias^2: Average squared difference between blue and black line.
- Variance: Average squared differences between gray lines.
- Irreducible error: Average squared differences between blue line and data points.

kNN: A completely non-parametric approach

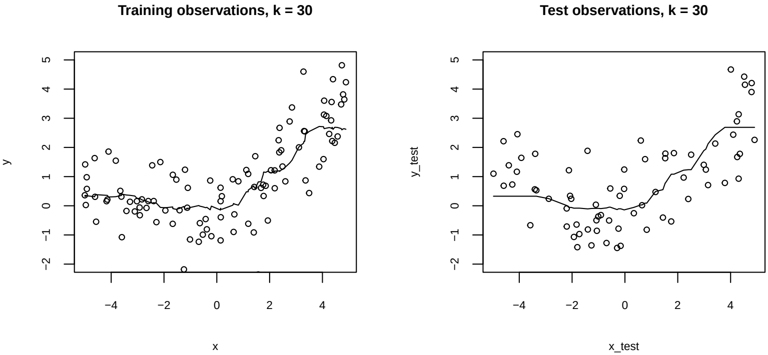
- k Nearest Neighbours (kNN) does not make assumptions about the distribution of data.
- As with nearly any non-parametric method, could also argue it is overparameterized.
    - with $k = 1$, as many means as training datapoint are estimated.

kNN: A completely non-parametric approach


Training observations, k = 3


Test observations, k = 3

kNN: A completely non-parametric approach



What happens to bias if k increases? What happens to variance if k increases?

If irreducible error increases, should a lower or higher k be preferred?

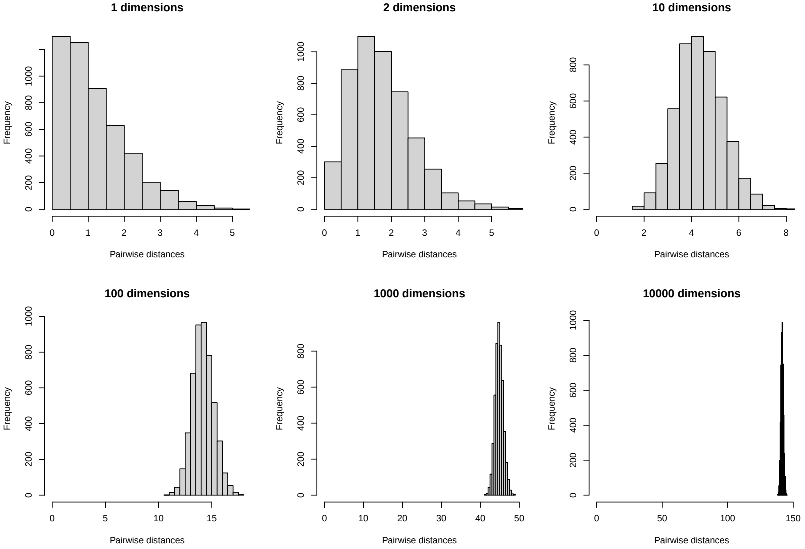
The curse of dimensionality

kNN (and other distance-based methods) assumes that nearness is meaningful.

Are distances in high dimensional problems just as meaningful as in low-dimensional problems?

```
p <- 10000
N <- 100
set.seed(42)
X <- matrix(rnorm(p*N), ncol = p, nrow = N)
par(mfrow = c(2, 3))
for (p in c(1, 2, 10, 100, 1000, 10000)) {
  distances <- dist(X[ , 1:p])
  hist(distances, main = paste(p, "dimensions"),
       xlim = c(0, max(distances)), xlab = "Pairwise distan
}
```

The curse of dimensionality



The curse of dimensionality

Distance is more meaningful in lower dimensions.

With very high dimensions, all observations are far apart, there are no real neighbours. Being nearer by 1 or 2 is likely to reflect only chance fluctuations.

High-dimensional space is lonely!

Would you prefer a high bias (low flexibility) or low bias (high flexibility) method for high-dimensional data problems?

How can one reduce dimensionality? (also: session 4)

Exercise: Flexibility and predictive performance

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible (low bias) statistical learning method to be better or worse than an inflexible (high bias) method:

a) Sample size N is extremely large, and the number of predictors p is small.

b) The number of predictors p is extremely large, and the number of observations N is small.

c) The relationship between the predictors and response is highly non-linear.

d) The variance of the error terms, $\sigma^2 = Var(\epsilon)$, is extremely high.

We see eachother again in Session 5: Non-linear models

- Generalized Additive Models and Smoothing Splines
- Support Vector Machines
- Decision Trees
- Tree ensembles