Knowledge Management and Business Intelligence

Supervised Learning:
Regression

Jan Vanthienen

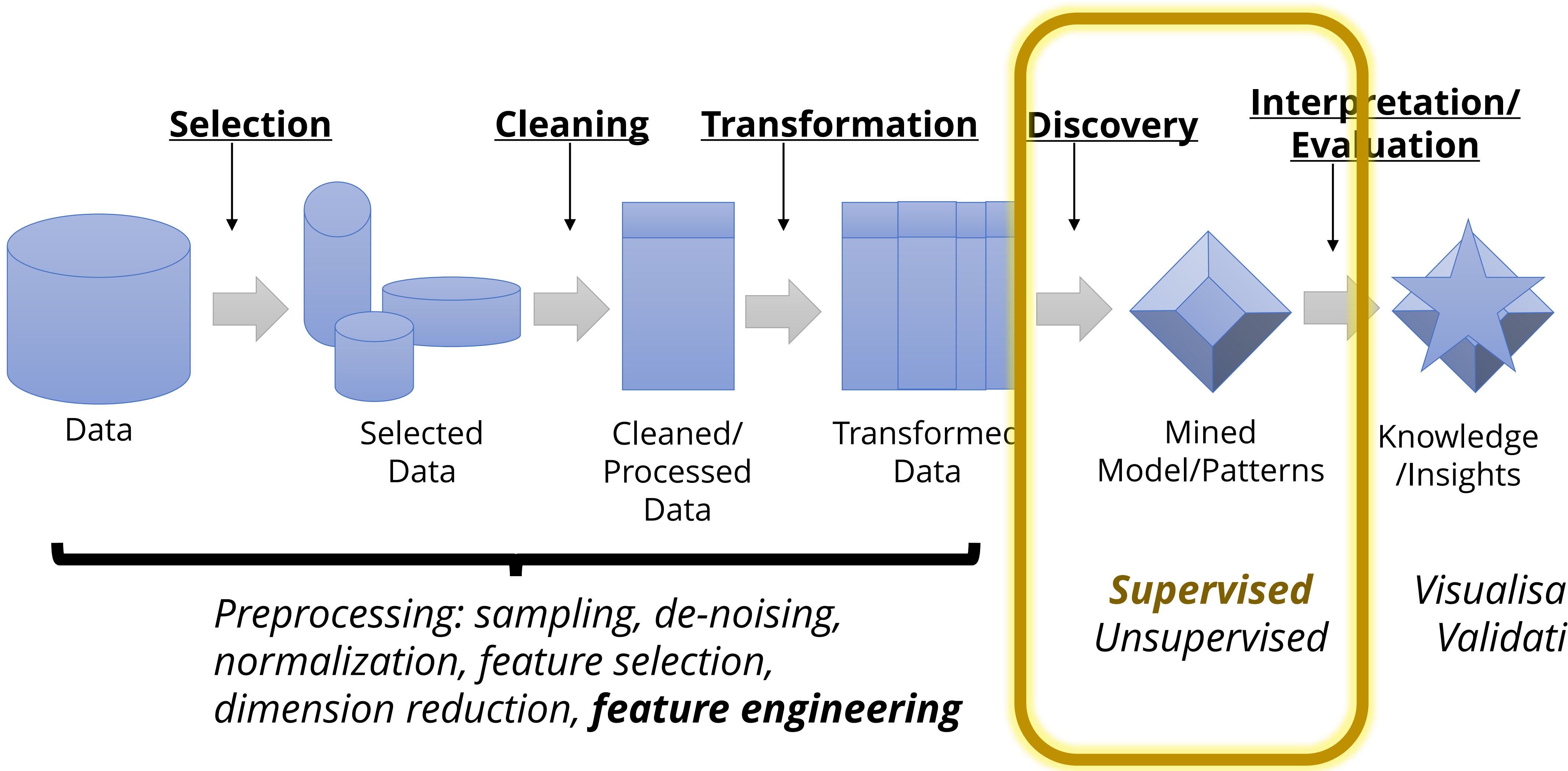
Business Intelligence

Knowledge Based Systems,
Business Rules, Intelligent Systems
& Knowledge Discovery

Acco

Ch 10

This course



Selection **Cleaning** **Transformation** **Discovery** **Interpretation/ Evaluation**

Data

Selected Data

Cleaned/ Processed Data

Transformed Data

Mined Model/Patterns

Knowledge /Insights

Preprocessing: sampling, de-noising, normalization, feature selection, dimension reduction, ***feature engineering***

Supervised
Unsupervised

Visualisation
Validation

Recall

Predictive data mining methods

- Use some variables to predict unknown or future values of other variables
- Example: classification, regression, recommender systems

Descriptive data mining methods

- Find human-interpretable patterns that describe the data
- Example: clustering, associations, sequences

Supervised data mining:

- You have a **labelled** data set at your disposal
- I.e. list of customers with outcome *yes* if bought product and *no* if they did not

Unsupervised data mining:

- You don't have the labels at your disposal

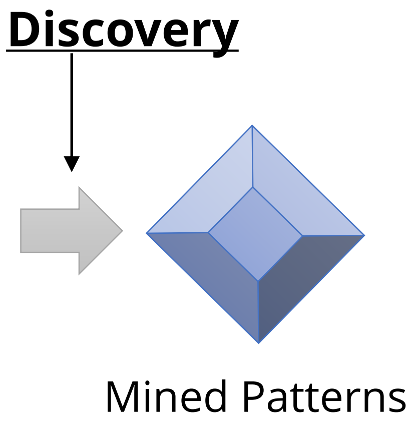
Predictive, supervised data mining

A data set with:

- Instances and variables
- And now: also a label we wish to learn and explain, and also predict and forecast

Two main types of supervised data mining can be distinguished:

- Regression: continuous, numeric label (well... except logistic regression)
- Classification: categorical label



Regression

Definition: “regression”:

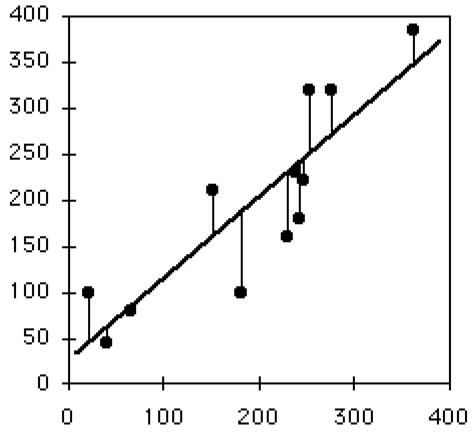
- In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables
- The relationship between a dependent variable and one or more independent variables
  - E.g.: predict price of house based on some variables: number of bedrooms, distance to capital city, etc.
- Regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed
- Regression analysis is widely used for prediction and forecasting
- Strong statistical, parametric underpinnings

Parametric statistics assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters

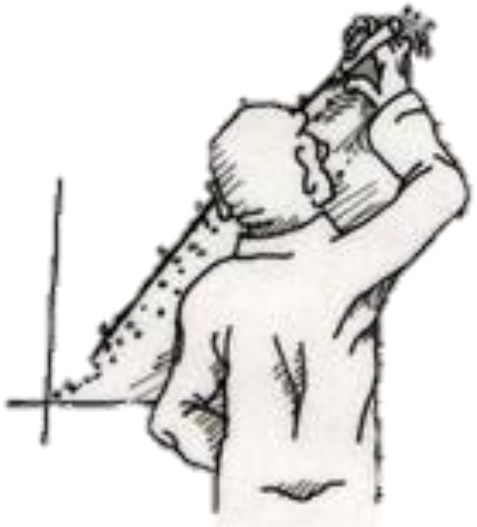
Regression

- Linear regression, multiple linear regression
- Lasso and ridge regression
- Logistic regression
- Nonparametric regression

Linear regression



- $y = \beta_0 + \beta_1 x + \epsilon$ with $\epsilon \sim N(0, \sigma)$
- E.g. university-gpa = 50 + 0.3 * highschool-gpa
- β_0: mean response when *x* = 0 (the *y*-intercept)
- β_1: change in mean response when *x* increased by one unit (the slope)
- $\beta_0 + \beta_1 x$: mean response when variable takes on the value *x*
- How to determine / estimate the values of β_0 and β_1?



Linear regression

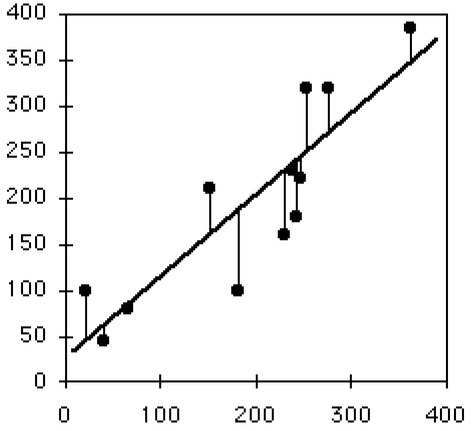
- $y = \beta_0 + \beta_1 x + \epsilon$ with $\epsilon \sim N(0, \sigma)$

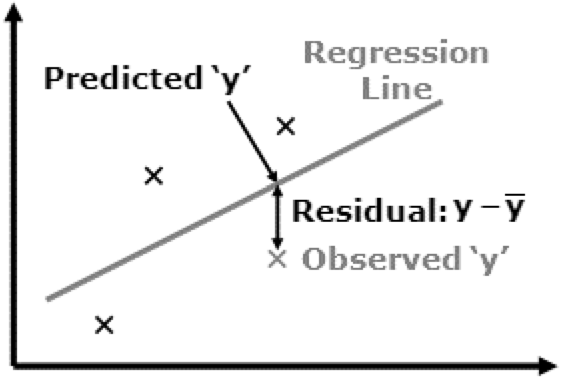
- How to determine / estimate the values of β_0 and β_1?

- Minimize the sum of squared errors (SSE) = $\sum_{i=1}^{n}(y_i - \hat{y_i})^2$

- Standard error of estimate $\sigma_{est} = \sqrt{SSE/n}$

- Or, for a sample rather than whole population: $s_{est} = \sqrt{\frac{SSE}{n-2}}$ (because two parameters were estimated in order to estimate SSE)





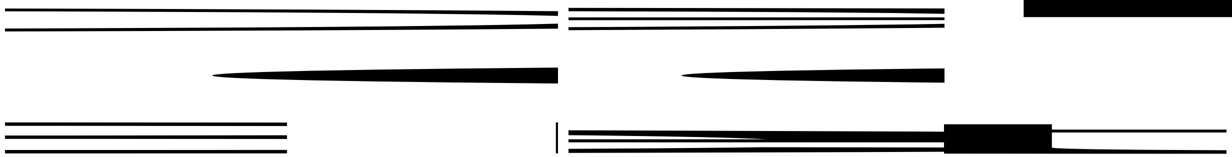
Regression model validation

- Hypothesis tests
  - 2-sided: $H_o: \beta_1 = 0$, $H_A: \beta_1 \neq 0$
  - 1-sided: $H_o: \beta_1 = 0$, $H_A^+: \beta_1 > 0$, $H_A^-: \beta_1 < 0$

- Confidence interval for β_1:
  - Conclude positive association if entire interval above zero
  - Negative if entire interval below zero
  - If interval contains zero: cannot conclude significant association
  - Same as using 2-sided hypothesis test

$$\hat{\beta}_1 \pm t_{\alpha/2}\,\hat{\sigma}_{\hat{\beta}_1} \equiv \hat{\beta}_1 \pm t_{\alpha/2}\,\frac{s}{\sqrt{S_{xx}}}$$

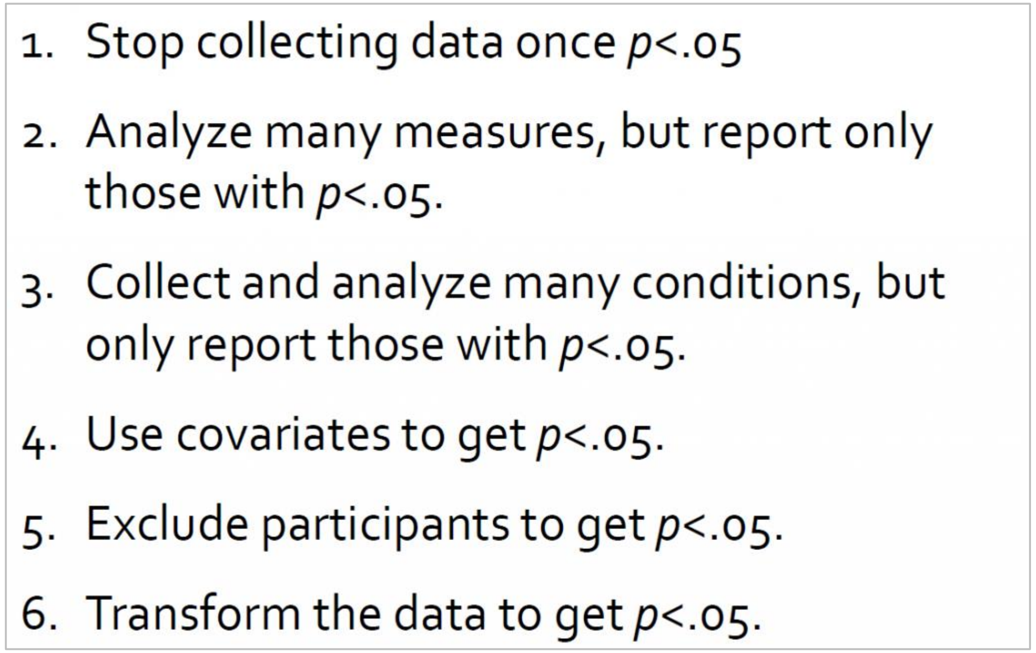
Also look at “effect size”



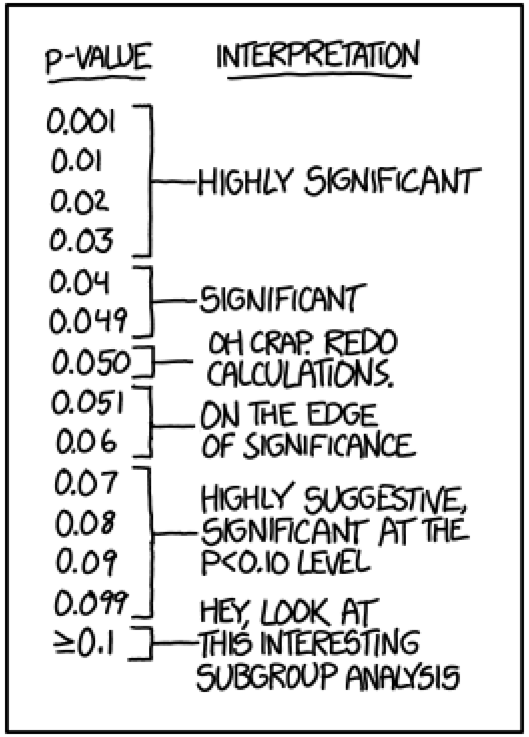
intercept
x

P-value hacking

Data torture, fishing, dredging, massage… :(



1. Stop collecting data once *p*<.05
2. Analyze many measures, but report only those with *p*<.05.
3. Collect and analyze many conditions, but only report those with *p*<.05.
4. Use covariates to get *p*<.05.
5. Exclude participants to get *p*<.05.
6. Transform the data to get *p*<.05.


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Regression model validation

- $\boldsymbol{r^2}$**: coefficient of determination**: the proportion of variation in *y* explained / “captured” by the regression model

- $r^2 = \frac{S_{yy} - SSE}{S_{yy}} \quad \in [0,1]$

- 1 minus the fraction of unexplained variance

- An r-squared of 1 means that the regression line perfectly fits the data

- With $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y_i})^2$ and $SSE = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$

Mean

Predicted

Regression model validation

- $\boldsymbol{r^2}$**: coefficient of determination**: the proportion of variation in *y* explained / “captured” by the regression model

- **AIC (Akaike Information Criterion):** $\sigma_{est}^2(\frac{n+k}{n-k})$ = $\frac{SSE}{n}$ $(\frac{n+k}{n-k})$

- **Bayesian Information Criterion (BIC)**, also known as Schwarz criterion

- **Correlation** between predicted and true *y* value

- $\boldsymbol{r^2}$**-adjusted**: $r_a^2 = 1 - (1 - r^2)\left(\frac{n-1}{n-k}\right)$ a modified version of r-squared adjusted for the number of predictors in the model: the adjusted r-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

- With $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y_i})^2$ and $SSE = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$

Multiple linear regression

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$ with $\epsilon \sim N(0, \sigma)$

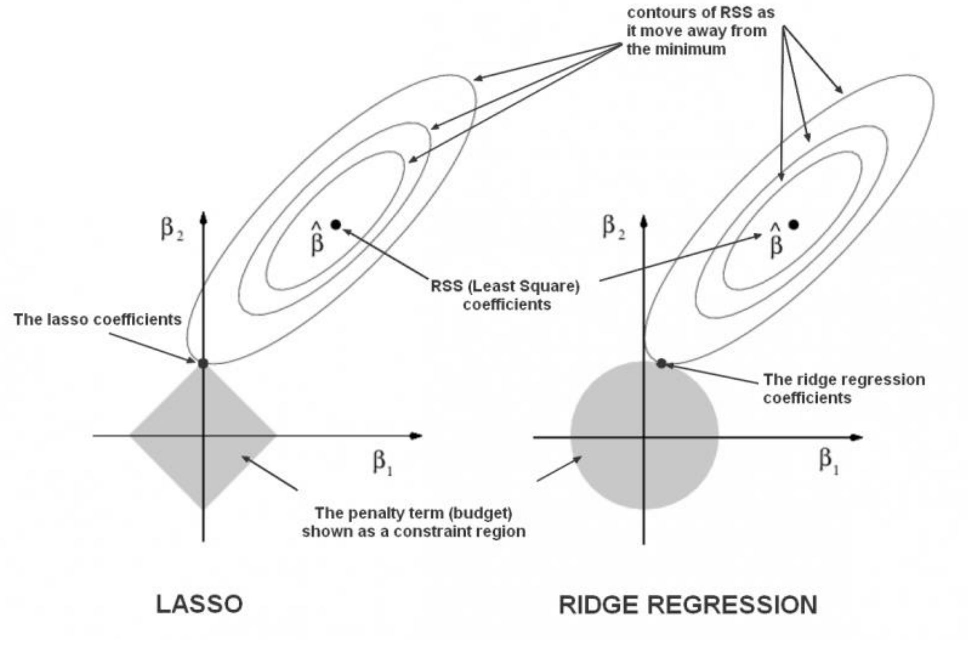
- Similar concepts apply, try to minimize $SSE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$

- In general, **smaller** models are preferred:
  - Use smallest subset of variables as possible whilst still keeping r-squared high
  - Stepwise introduction or removal of variables
  - Keep significant variables only

- Other **validation** checks:
  - Check residuals/fit of model: “fitting the majority vs. spotting the minority”
  - Variables with too extreme coefficients
  - Sign of the coefficients, e.g. house price decreases if number of bedrooms increases?

Lasso and ridge regression

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$ with $\epsilon \sim N(0, \sigma)$ try to minimize $SSE = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$
- **Lasso estimation** (least absolute shrinkage and selection operator) regression: minimize $SSE = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \sum_{j=1}^{p}|\beta_j|$
- **Ridge regression estimation**: minimize $SSE = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \sum_{j=1}^{p}|\beta_j^2|$
- As penalties (loss function on ϵ):
  - L1 penalty: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \boldsymbol{\lambda}\sum|\boldsymbol{\beta_i}|$ lasso
  - L2 penalty: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \boldsymbol{\lambda}\sum\boldsymbol{\beta_i^2}$ ridge regression
  - Increasing lambda pushes the coefficients to zero during SSE estimation
  - For **regularization (L1 and L2)** and **variable selection (L1): lasso will force coefficients to become zero (selection), ridge only to keep them within bounds**
  - Why ridge, then? Easier to implement (slightly) and faster to compute (slightly), or when you have a limited number of variables to begin with


contours of RSS as it move away from the minimum
β_2
$\hat{\beta}$
RSS (Least Square) coefficients
The lasso coefficients
β_1
β_2
$\hat{\beta}$
The ridge regression coefficients
β_1
The penalty term (budget) shown as a constraint region
LASSO
RIDGE REGRESSION

Logistic regression

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$ with $\epsilon \sim N(0, \sigma)$
- Same basic formula, but now not with a continuous *y*
- Two possible outcomes: either 0 or 1, no or yes – a **categorical, binary label, not continuous**
- Logistic regression is thus **more like a classification technique rather than regression**
- Instead of simply predicting 0 or 1 (like a linear model), we want to predict the probability that *y* = 1

Logistic regression

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$ with $\epsilon \sim N(0, \sigma)$

- Why couldn't we just use linear regression? Just consider *y* as either numerical 0 or 1, right?

- But linear regression penalizes large errors quadratically!
  - E.g: If *y* = 1 and model predicts 1.2: ok
  - If *y* = 1 and model predicts 10.0: huge error, even although here this would still be considered ok

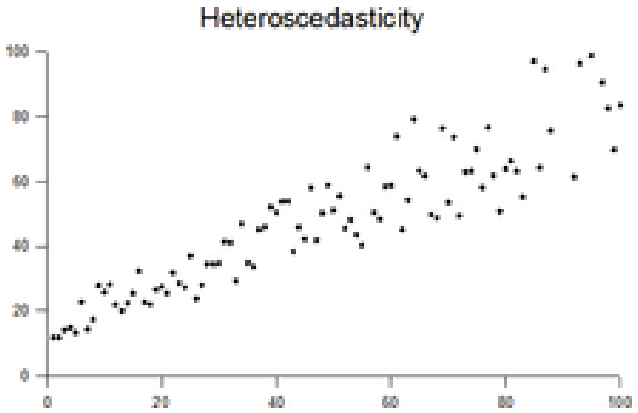
- Logistic model thus better suited to deal with the nature of the response variable

Logistic regression

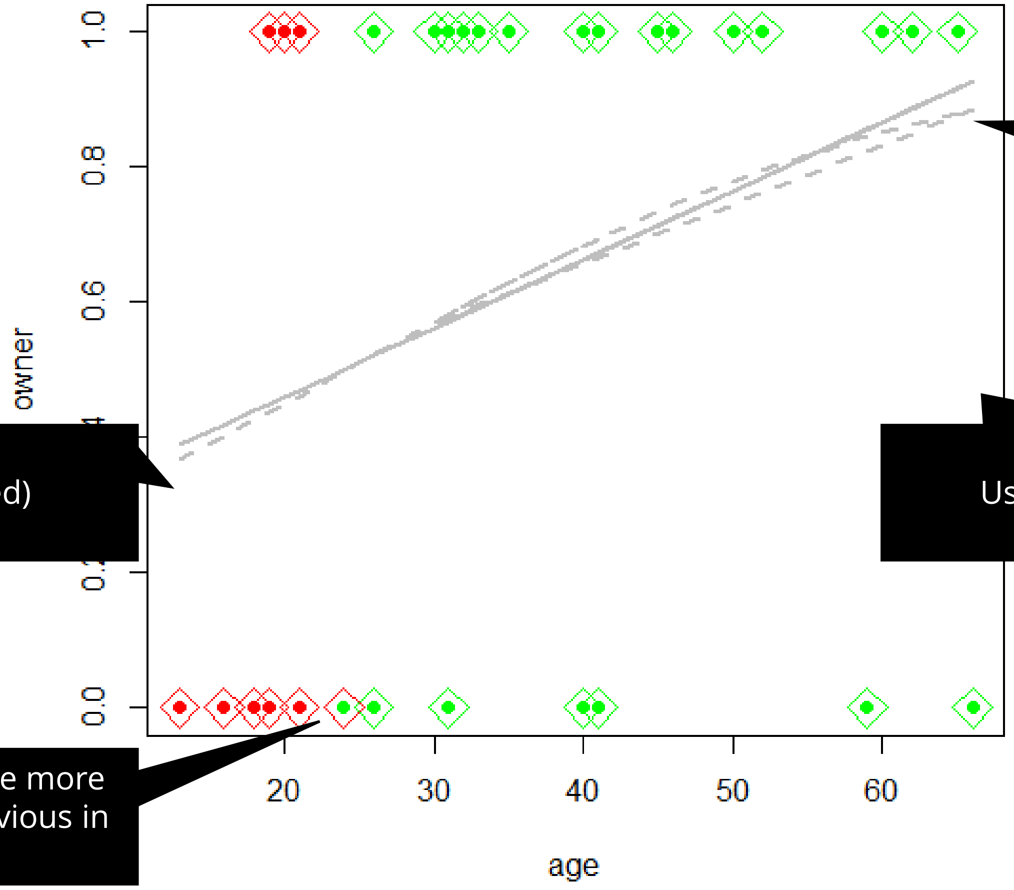
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$ with $\epsilon \sim N(0, \sigma)$

- Logistic model thus better suited to deal with the nature of the response variable

- **Not to be confused with logarithmic transformation:**
  - $\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$ with $\epsilon \sim N(0, \sigma)$
  - When residuals have a skewed distribution (transformation will give symmetrically distributed residuals around zero)
  - The spread of the residuals changes systematically with the values of the dependent variable (going from heteroscedasticity to homoscedasticity, i.e. if there are sub-populations that have different variabilities from others)



Logistic regression



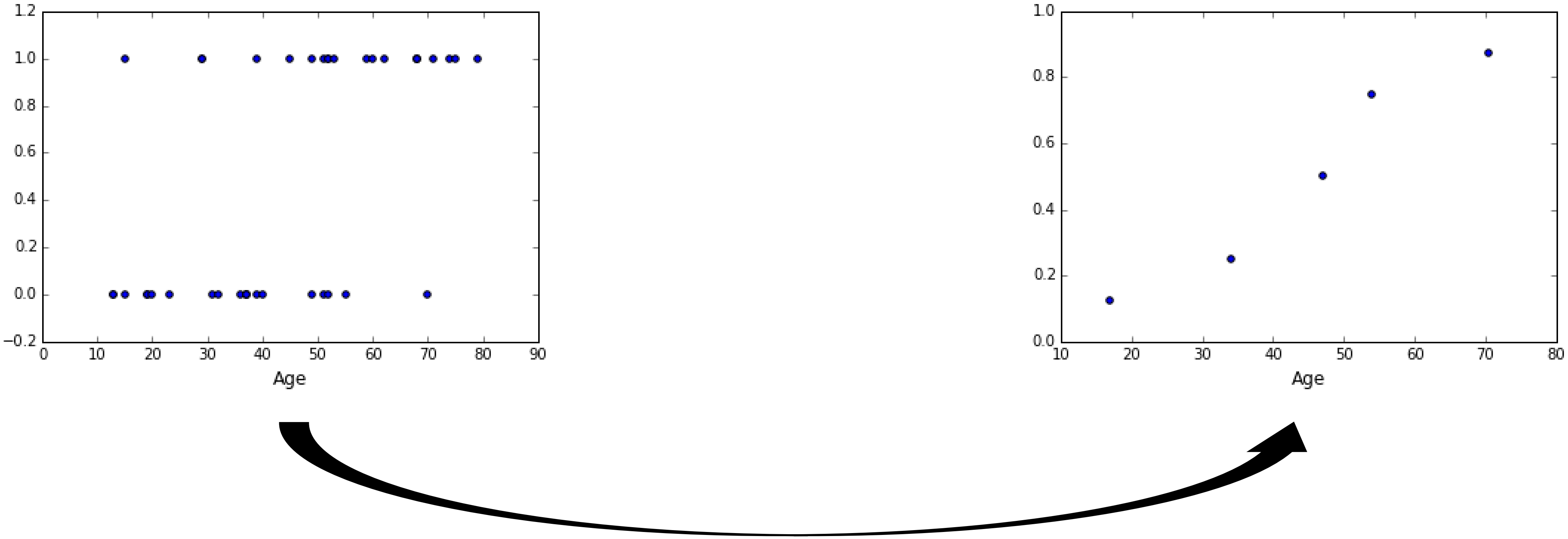
Normal linear regression (solid)

Logistic regression (dotted)

Using 0.5 as the cutoff point

Logistic regression catches one more example (difference not so obvious in this toy-example)

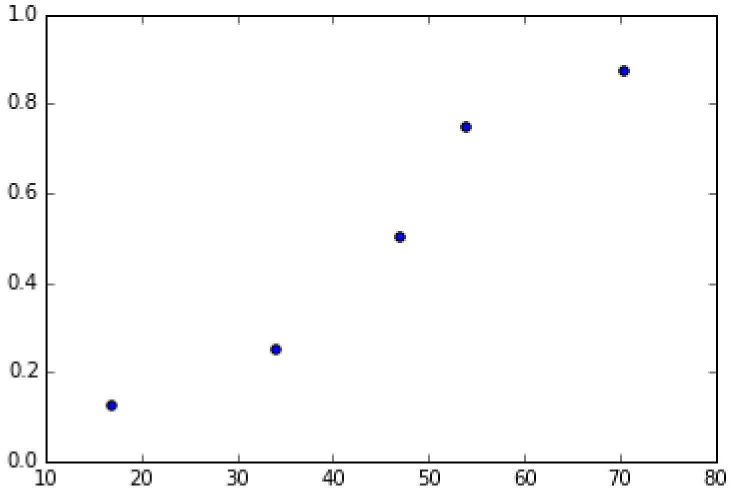
Logistic regression

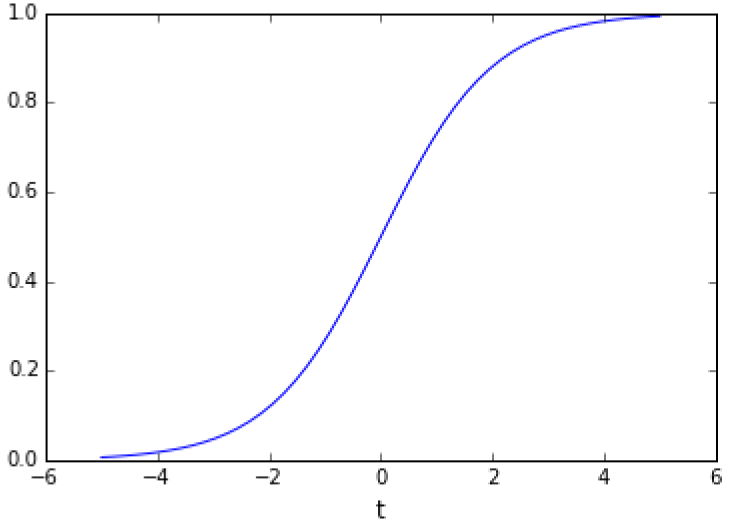


Equally sized bins

http://gormanalysis.com/logistic-regression-fundamentals/

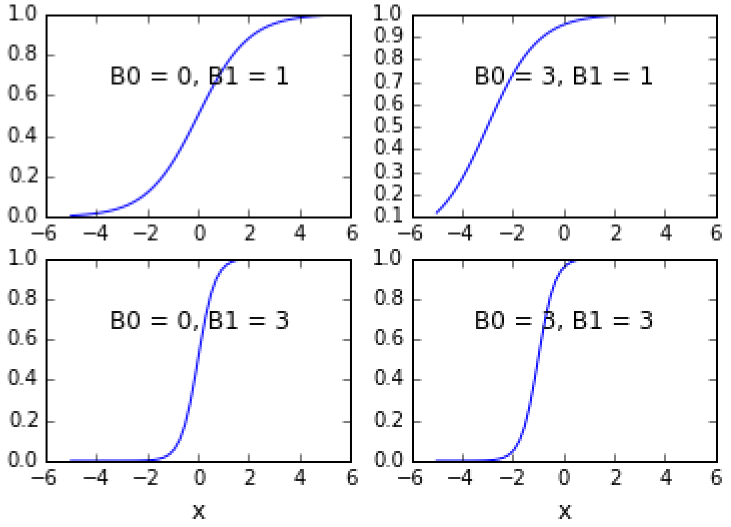
Logistic regression





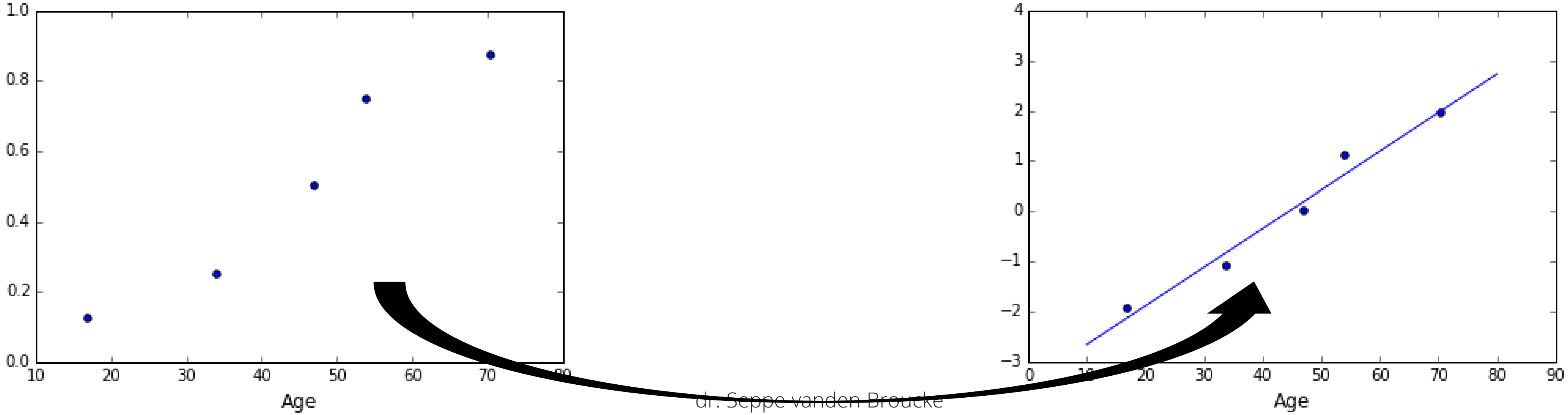
- Like this, our data takes an S-curve shape
- How can we describe such a shape?
- Logistic function: $\sigma = \frac{1}{1+e^{-t}}$
- We can fit this function to our data points!
- $F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

Logistic regression



- $F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$
  - The probability that *y* = 1 given *x*, P(y = 1 | x)

- We can rewrite this as $\ln\left(\frac{F(x)}{1-F(x)}\right) = \beta_0 + \beta_1 x$
  - Nice, this is just a linear function as we saw before
  - We can take our existing data, and convert it to an *y'*
  - And fit this using a normal linear regression



Logistic regression

- $F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots)}}$

- Normally, the coefficients are estimated using maximum likelihood estimation (and not using the binning trick we applied before)

- A Bernoulli variable is a binary random variable (0 or 1) with probability of success *p*. The Bernoulli distribution is defined by a single parameter, *p*, and its expected value equals *p*

- At each point, the samples close to that point follow a Bernoulli distribution whose expected value *p* is *p(x)*

- We can treat *F(x)* and *p(x)* as the same, the probability that *y* = 1
  - So we can think of estimating our logistic function as defining an infinite set of Bernoulli distributions (one for every value of *x*)

Logistic regression

- $F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots)}}$

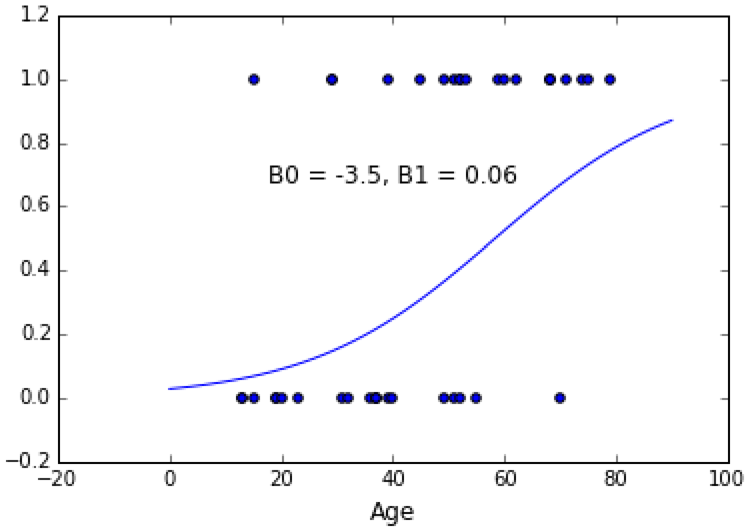
- Let's start with a guess for our coefficients

- We can calculate the **likelihood** of our model, meaning assuming our model is the true one, what is the probability of our sample data points occurring from our model?

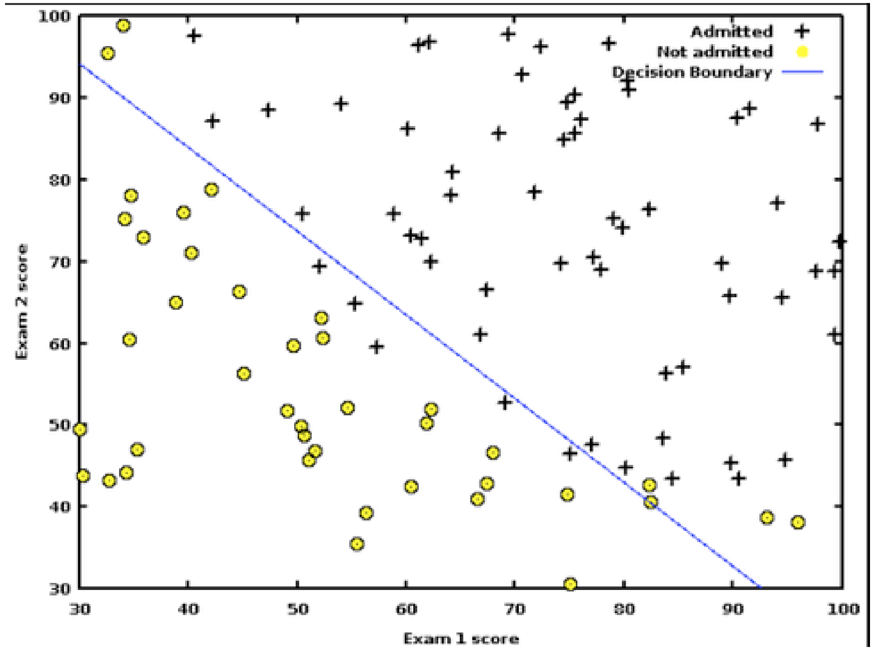
- P(y=0 | x=13) * P(y=1 | x=28) * … = (1-F(13)) * F(28) * …

- We want to **maximize the likelihood**: find the **estimates for our coefficients that maximize the probability that our same data points came from our estimated model**

- Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximize the likelihood function, so that an iterative approach must be used instead

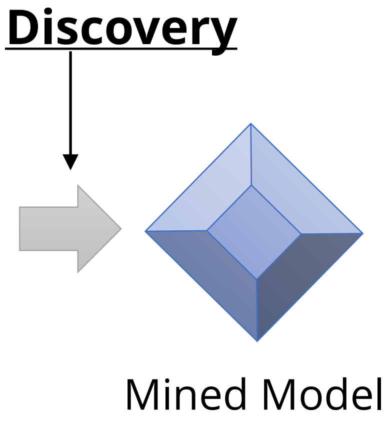


Logistic regression



- Iterative approach, but still very fast
- Easy to interpret, understand
- Statistical rigor
- General rule of thumb: for classification, start with logistic regression
- Linear decision boundary, though interaction effects can be taken into the model: $\beta_{12}x_1x_2$
- L1 and L2 penalization
- Sensitive to outliers, categorical variables need to be converted to dummies

Wrap-up


Discovery

Mined Model

Supervised learning: regression

- For continuous predictors
- Logistic regression: actually binary classification
- Logistic regression: most used classification technique!
- Very statistically sound underpinnings, huge amount of literature, knowledge available
- Challenges: sensitivity to outliers, categorical variables
- Many more aspects not discussed here