# Biostat 202: Opportunies and Challenges of "Big Data": Prediction Algorithms 1: Regression

**John Kornak**

**Division of Biostatistics**

**Department of Epidemiology and Biostatistics**

**08/15/2016**

# Overview of machine learning (prediction and clustering) lectures

- 3 lectures
- Machine Learning and the big picture
- Supervised vs. unsupervised Machine Learning
- Regression (prediction)
- Classification (prediction)
- Validation and testing
- Cross-validation
- Data reduction techniques
- Clustering
- IBM SPSS Data Modeler examples throughout

# Overview of this lecture: Prediction 1

- Machine Learning and the big picture
- Supervised vs. Unsupervised Machine Learning
- Linear regression
- Multiple regression
- Validation

# The big picture

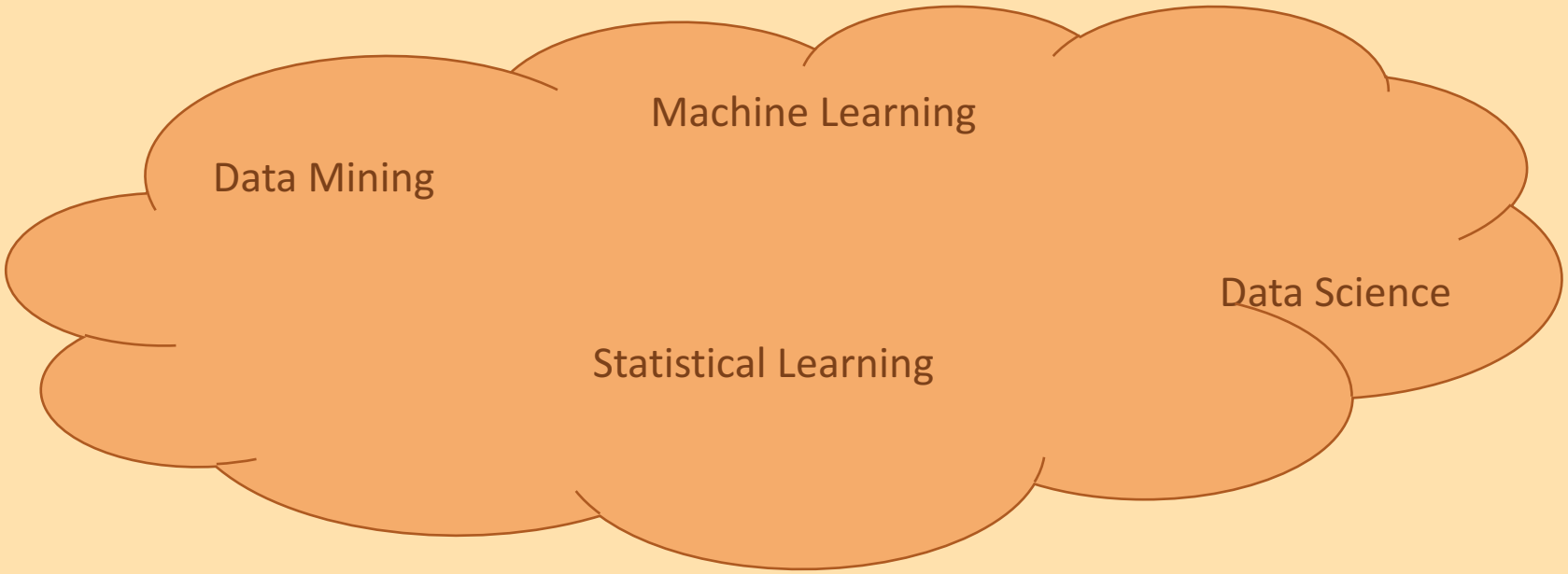
Machine Learning

Data Mining

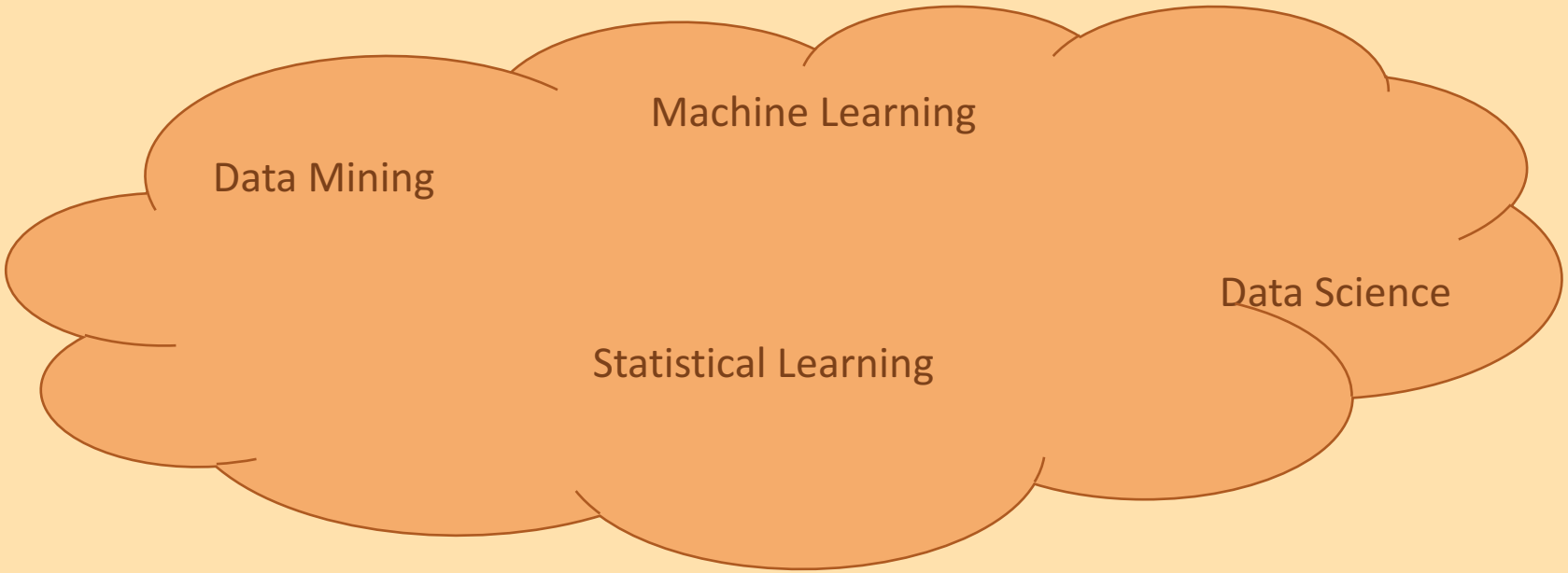
Data Science

Statistical Learning

# The big picture



# The big picture




Deep Learning

# Machine learning

- Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. In 1959,  Arthur Samuel defined machine learning as a “Field of study that gives computers the ability to learn without being explicitly programmed”

- Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

- We are not trying to do AI, but rather use some machine learning algorithms for the purpose of prediction or identifying patterns in data.

# Prediction

- Predict the CD4 count in a patient 6 months from now based on current CD4, viral load, demographic and lifestyle variables
- Predict whether a patient will develop breast cancer based on genetic, demographic, and lifestyle variables
- Predict whether or not, and what type of dementia a patient has based on MRI of the brain (see case study lecture)

# Supervised vs unsupervised

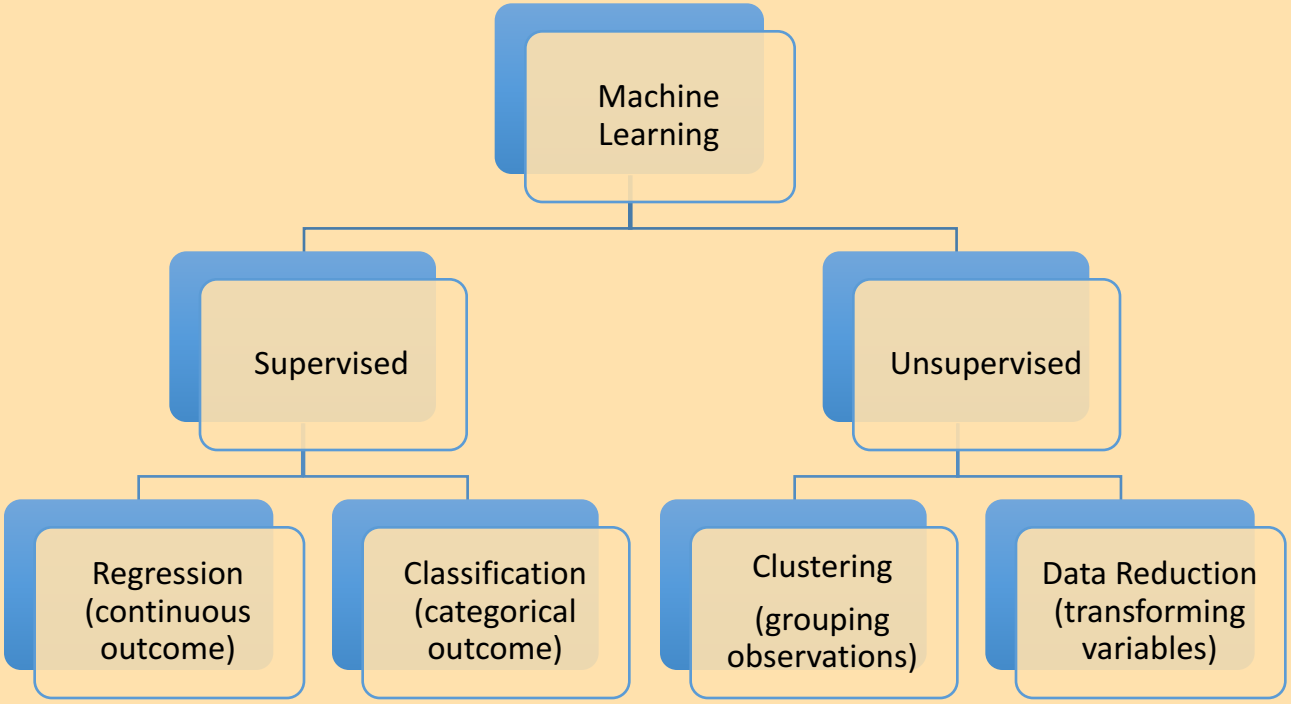
- **Supervised Learning**
- Have data where you know outcome in order to *train* the model
- Train the model to predict future outcomes from sets of predictors

**Prediction: Regression (continuous outcomes) and classification (categorical outcomes)**

- **Unsupervised Learning**
- No desired outcomes
- Separate data points into groups with "similar" properties (clustering)
- Or reduce data to fewer variables in some way that still captures important structure of the data

**Clustering and data reduction**

# Supervised vs. unsupervised



# Definitions

- *Instance* (also *Item*, *Record*) = observation:
  - an example, described by a number of attributes,
  - e.g. a day can be described by temperature, humidity and cloud status
  - e.g. a person can be described by age, gender, blood pressure and level of physical activity
- *Attribute* or *Field* = variable
  - measuring aspects of the Instance, e.g. temperature
- *Class* or *Label* = class or group
  - grouping of instances/observations, e.g. disease type

# Definitions

- *Target*. Also called *outcome*, *response*, *dependent variable* – the variable that you want to predict
- *Input*. Also called *predictor*, *covariate*, *independent variable* – the variables you will use to predict
- Modeler also has variable roles of *Record ID* (identification variable like medical record number), *split* (for separate analyses of subgroups), and *partition* (for assessing training/testing).

# Regression

- Start simple with one predictor
- We will move onto multiple predictors later…

# Linear Regression

Consider a response variable $y$ as linearly related to an attribute/predictor $x$, then:

$$y = ax + c$$

e.g. $y$ = Systolic BP (SysBP)

$x$ = Age

$c$= intercept (SysBP at age 0)

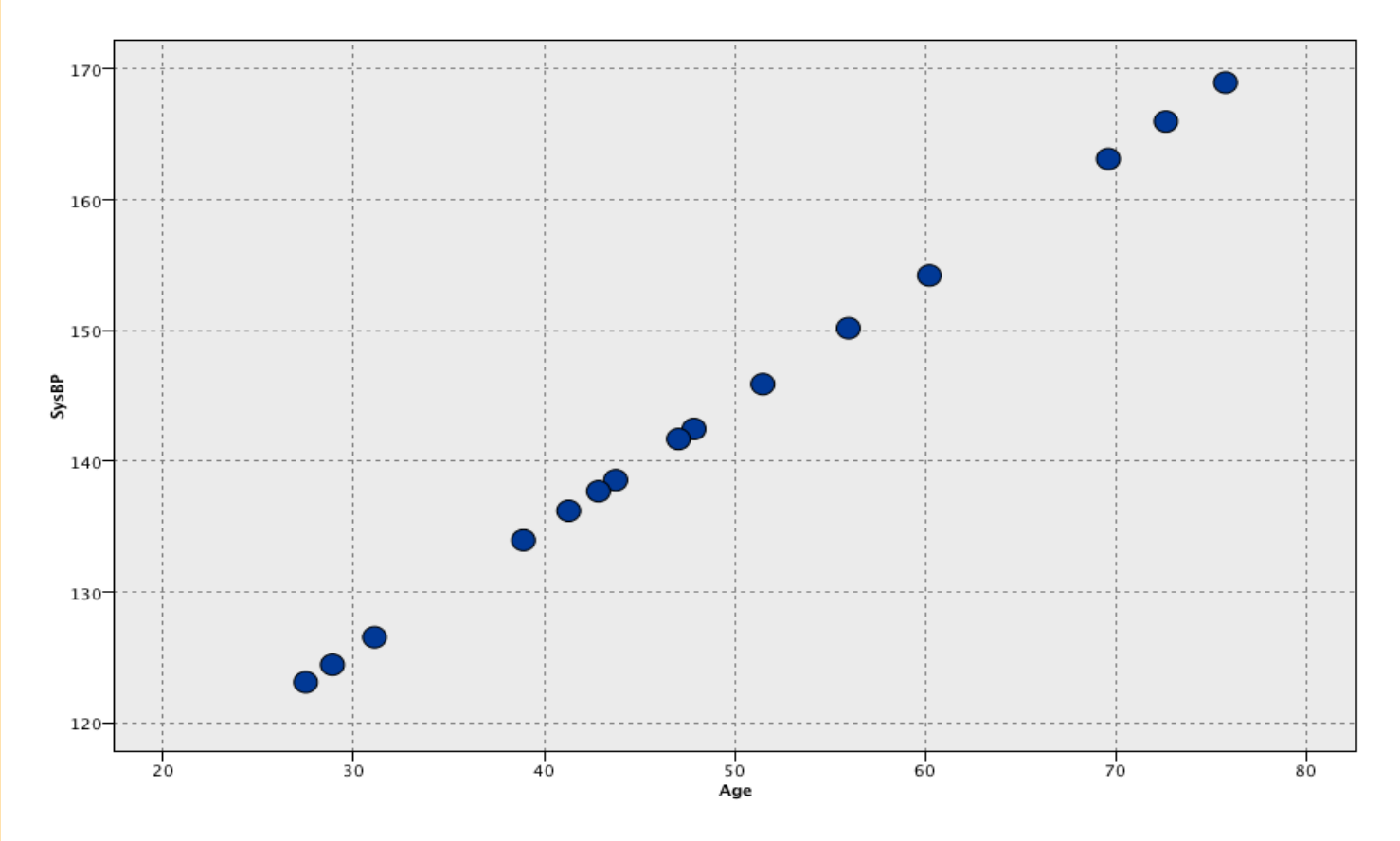
$a$ = change SysBP per year in Age

Question: Is Age predictive of SysBP?

# Fitting linear regression

$y$

SysBP



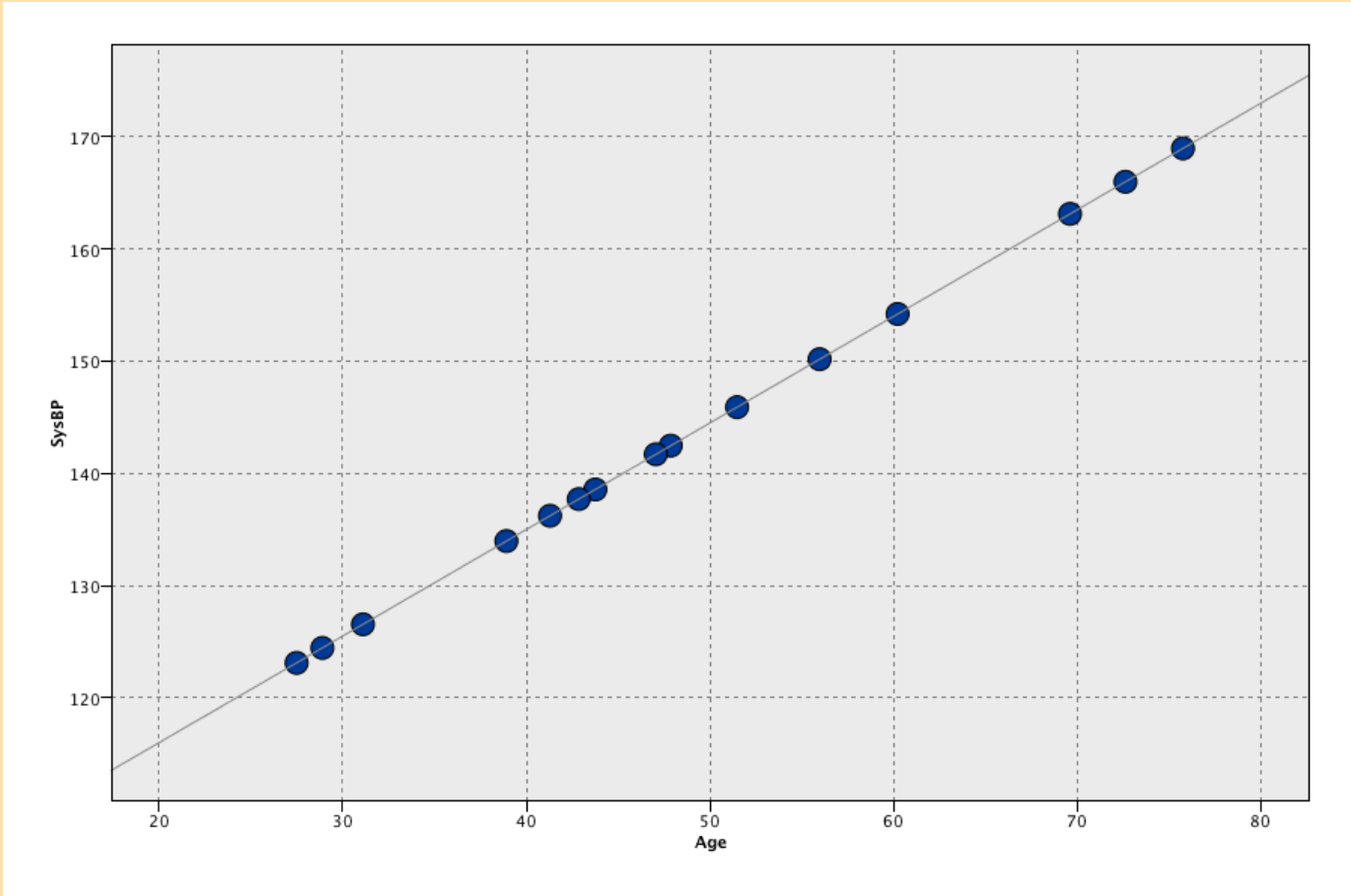

Age

$x$

# Fitting linear regression

$y$

SysBP




Age

$x$

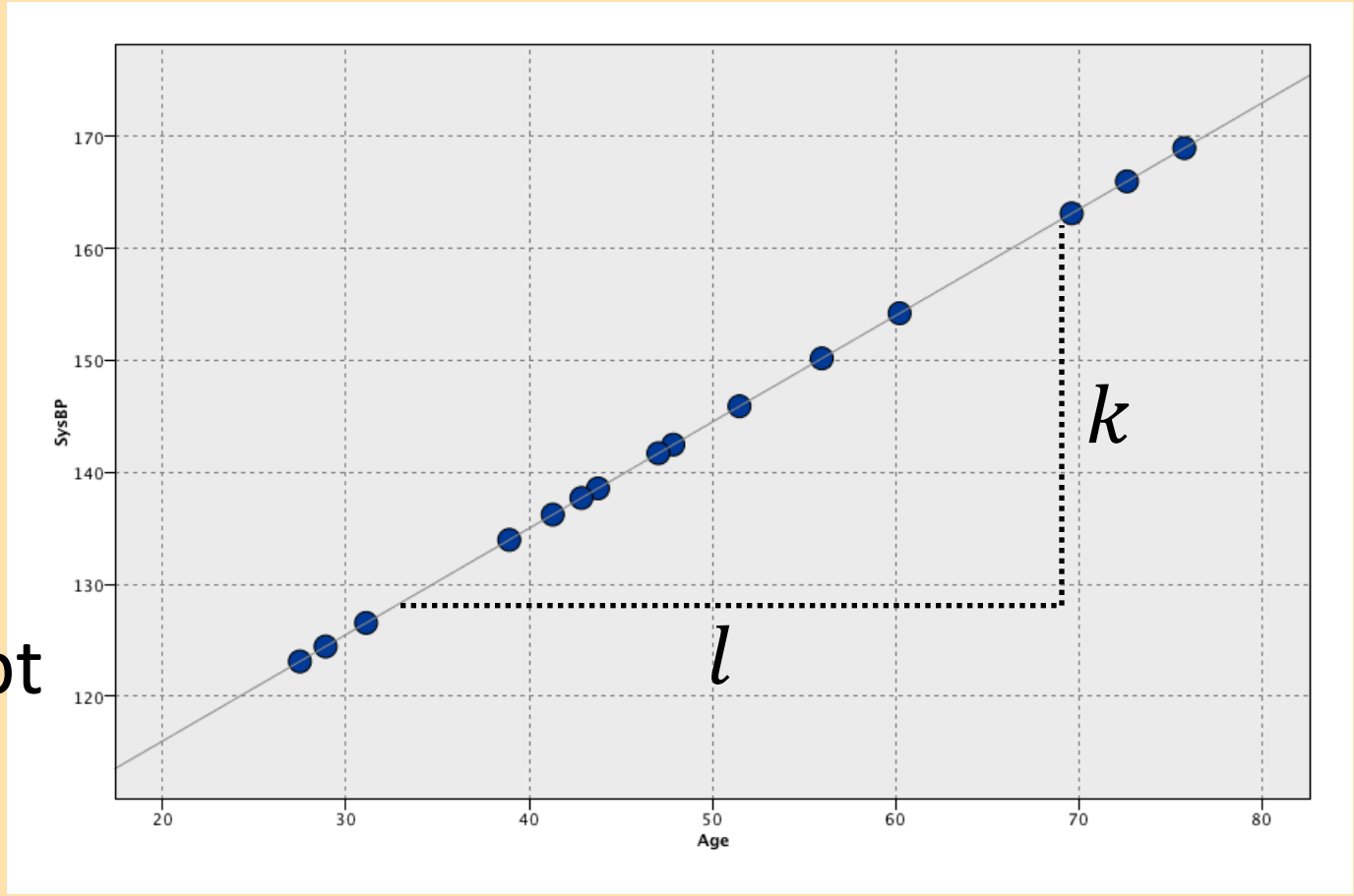
# Fitting linear regression

$y$

SysBP

intercept

$c$




slope

$$a = \frac{k}{l}$$

Age

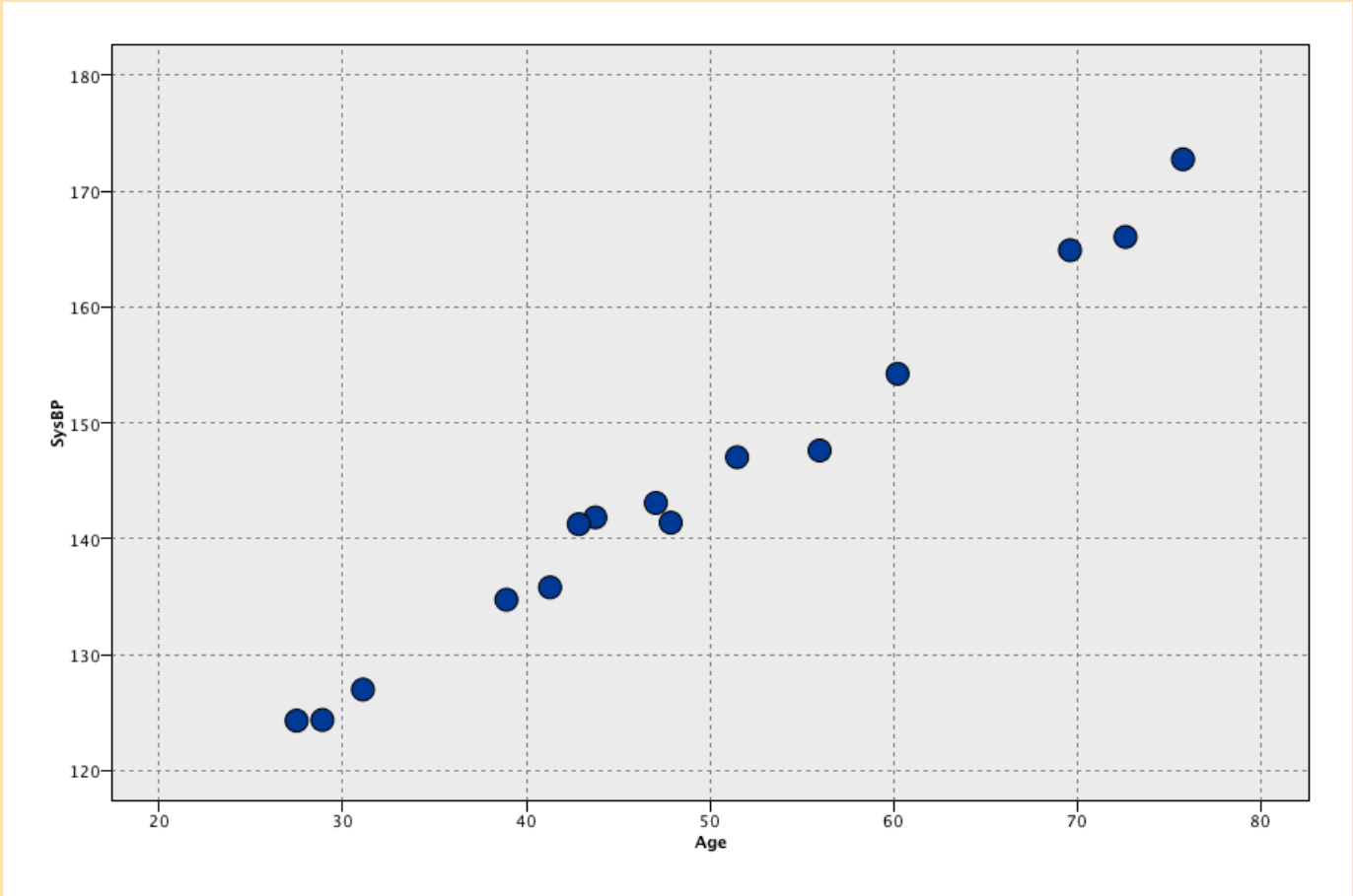
$x$

# Fitting linear regression

- But data (almost) never lie on a perfect straight line
  - Measurement error
  - Between subject variability (e.g. think about height vs. age – not all people of the same age have the same height)

# Data with measurement error

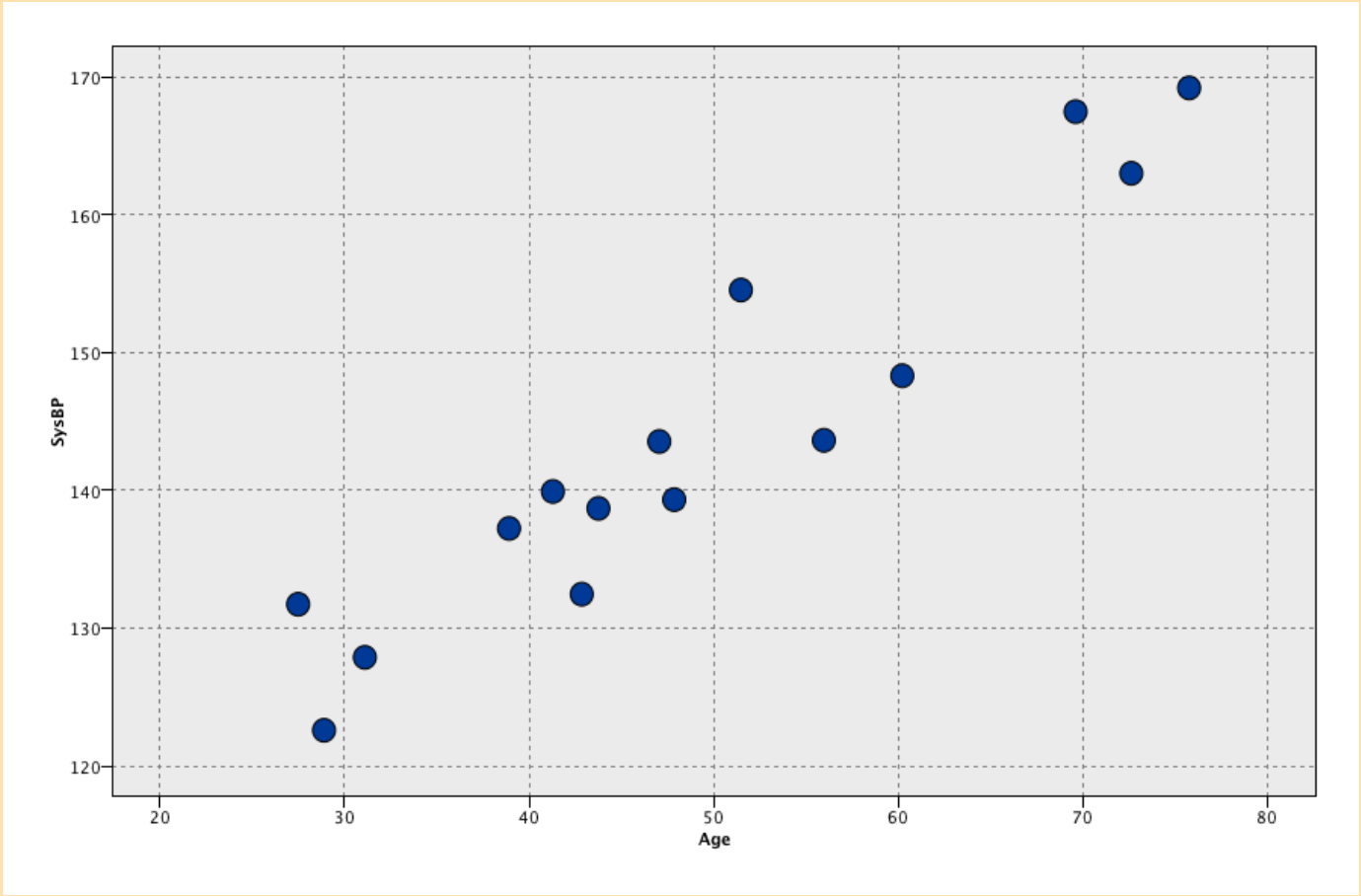
SysBP




Age

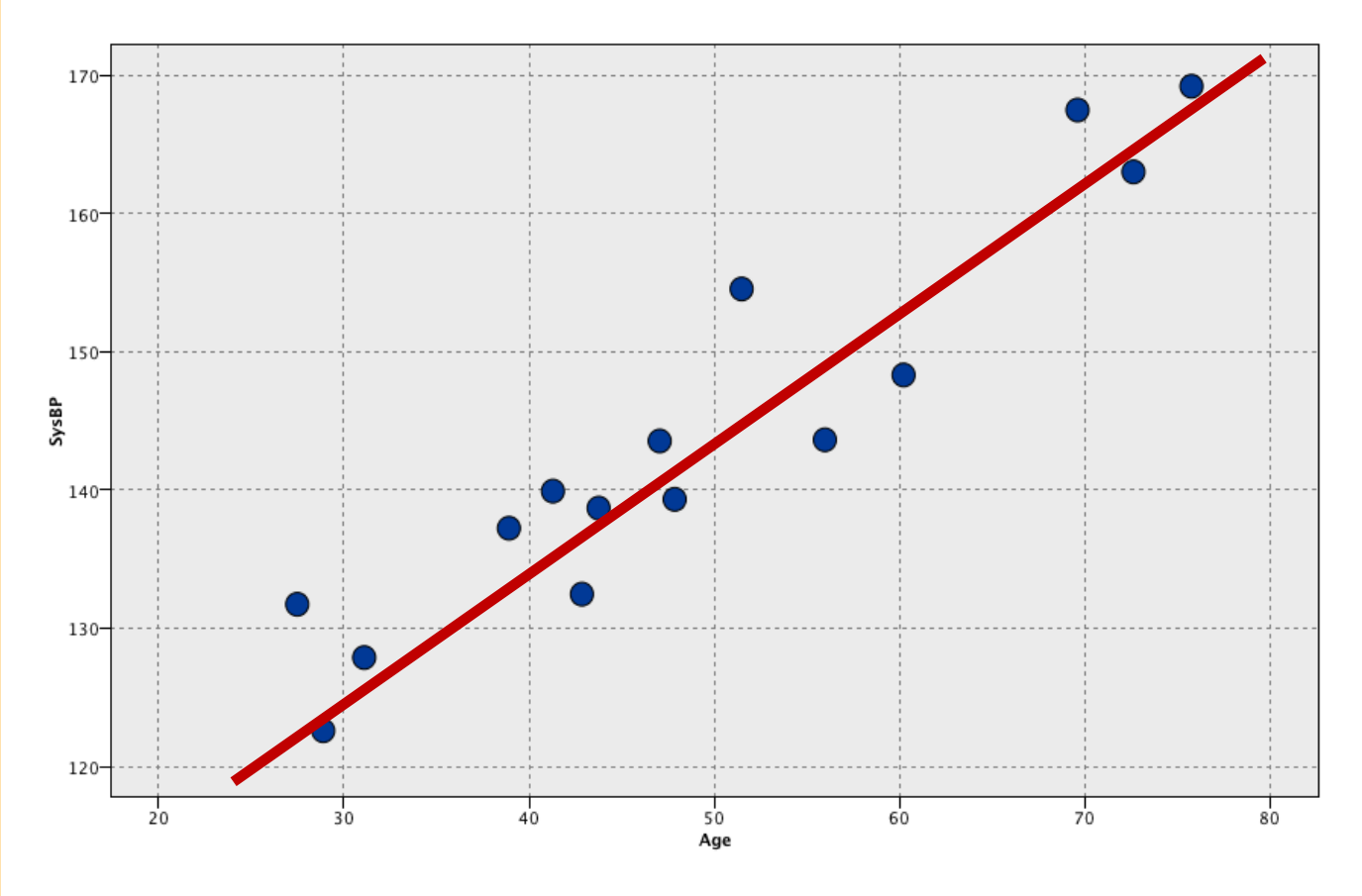
# Data with measurement error plus biological variability

SysBP




Age

# How to choose a fitted line?



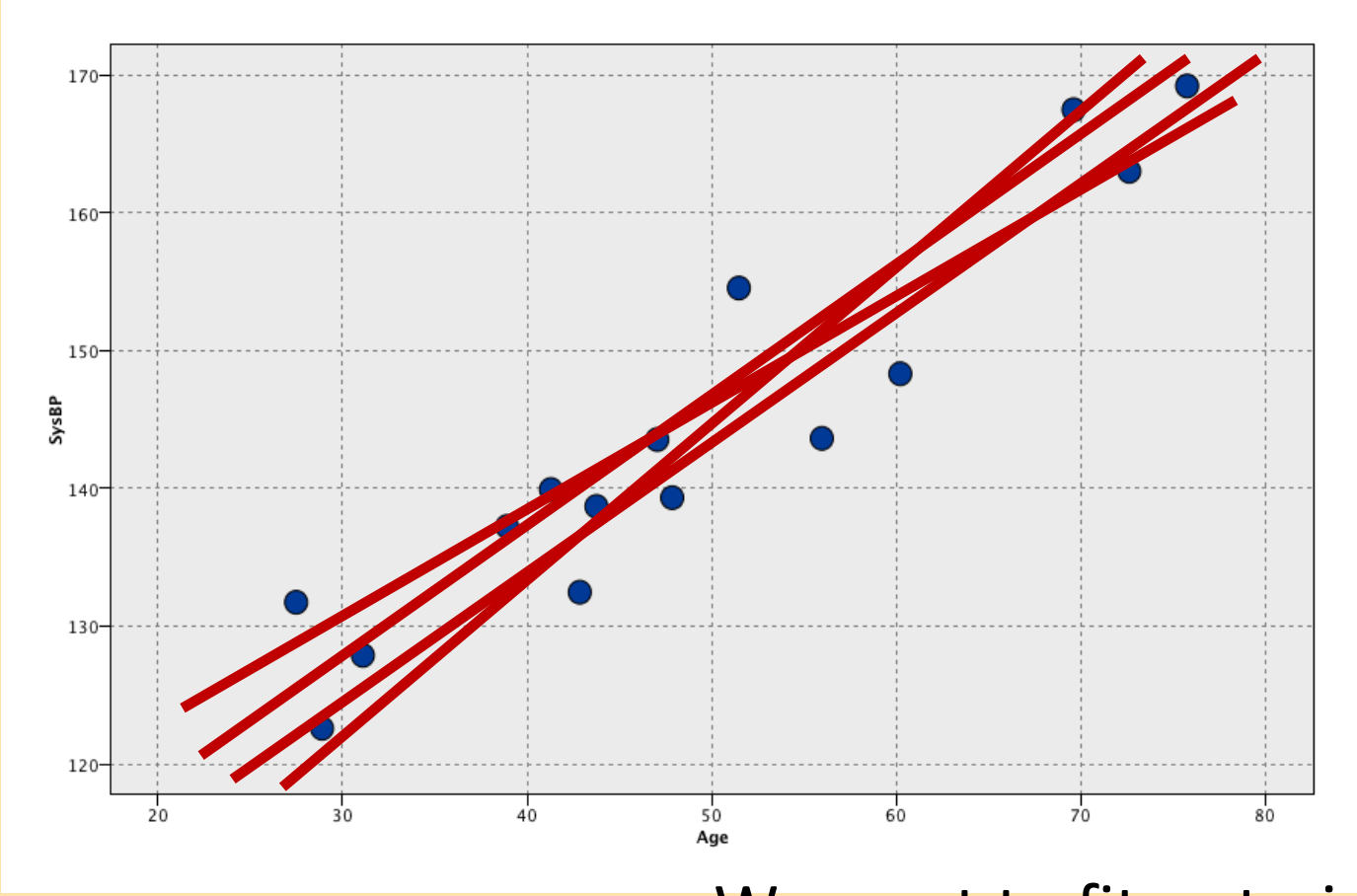

SysBP

Age

One possible line

# How to choose a fitted line?

SysBP

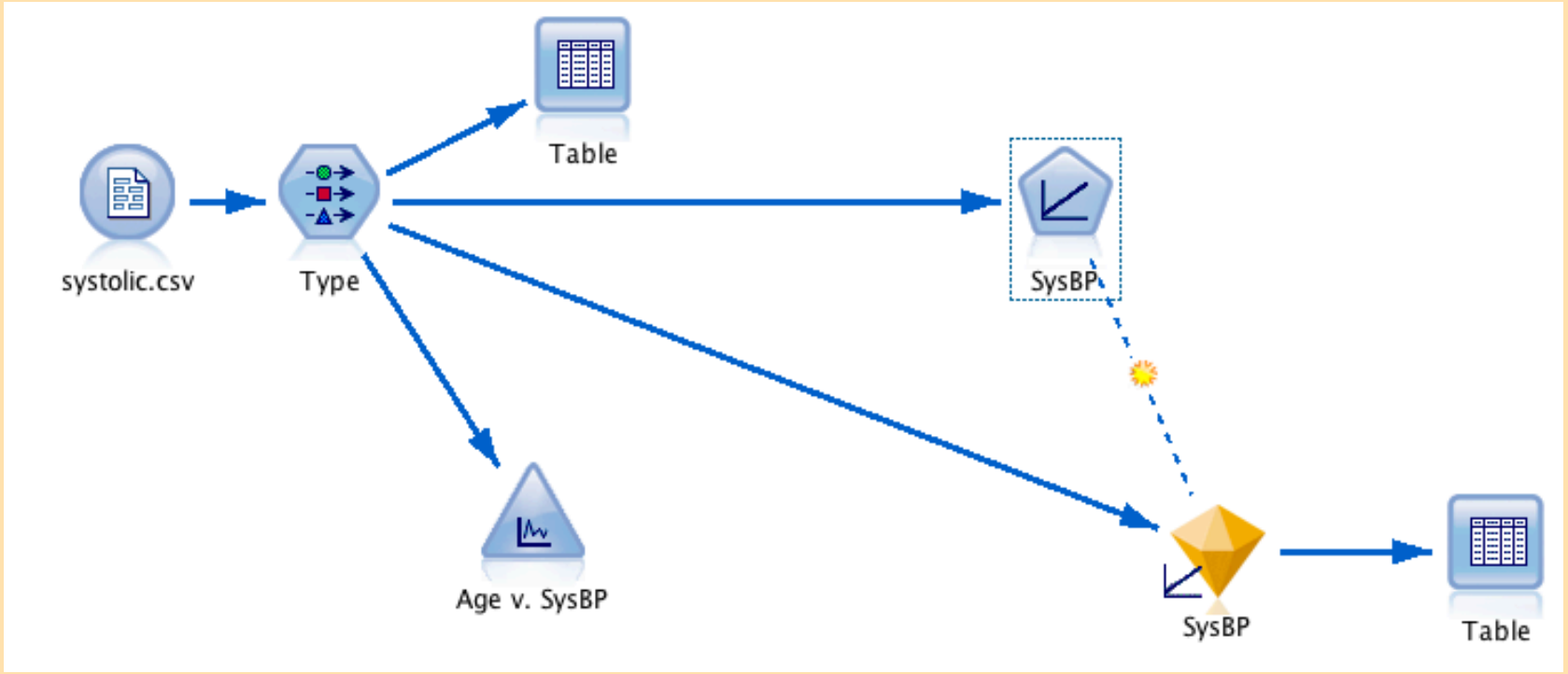



Age

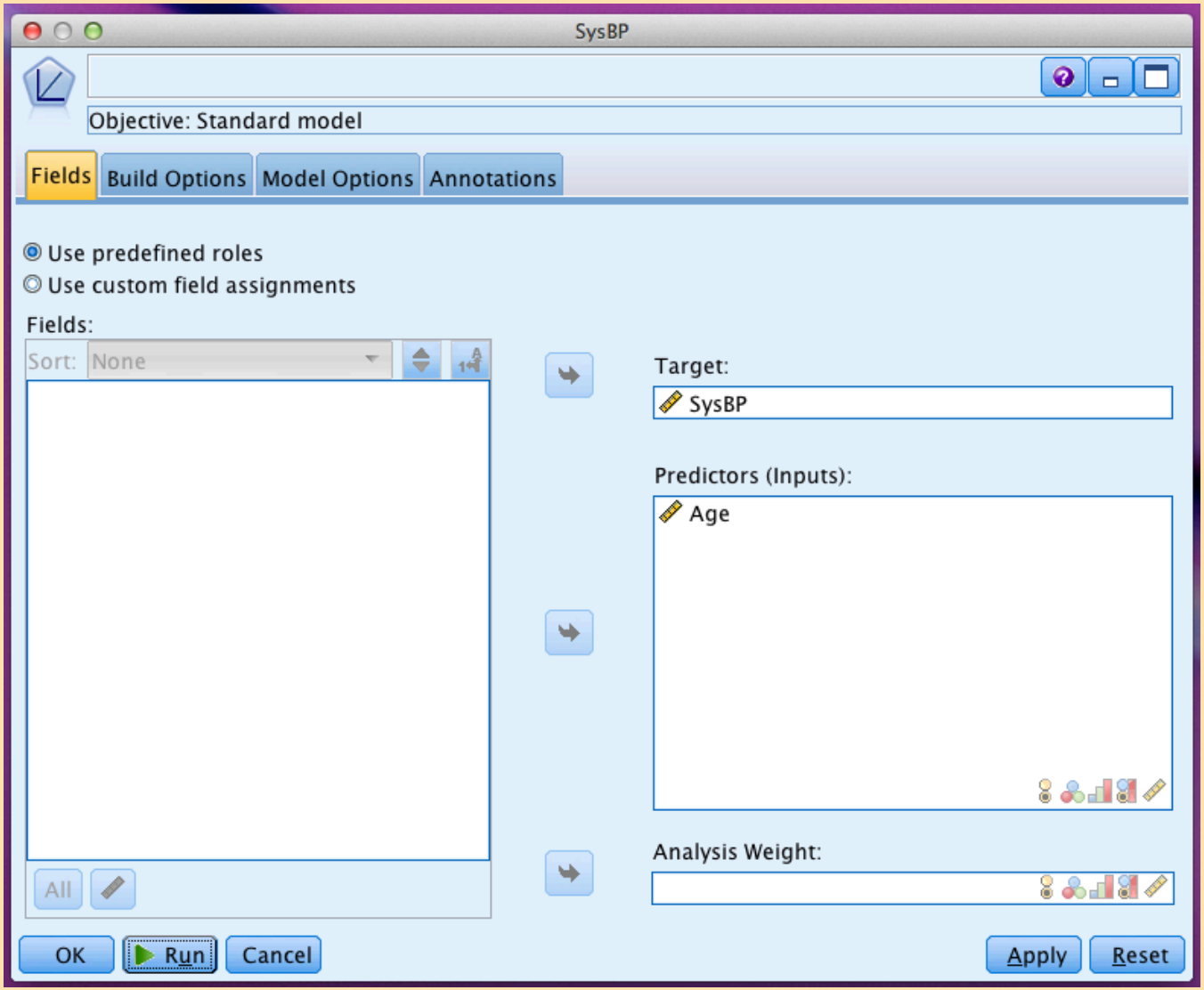
We want to fit a straight line to give us a relationship, but which one is ‘best’?

# IBM Modeler linear regression stream

Modeling tab

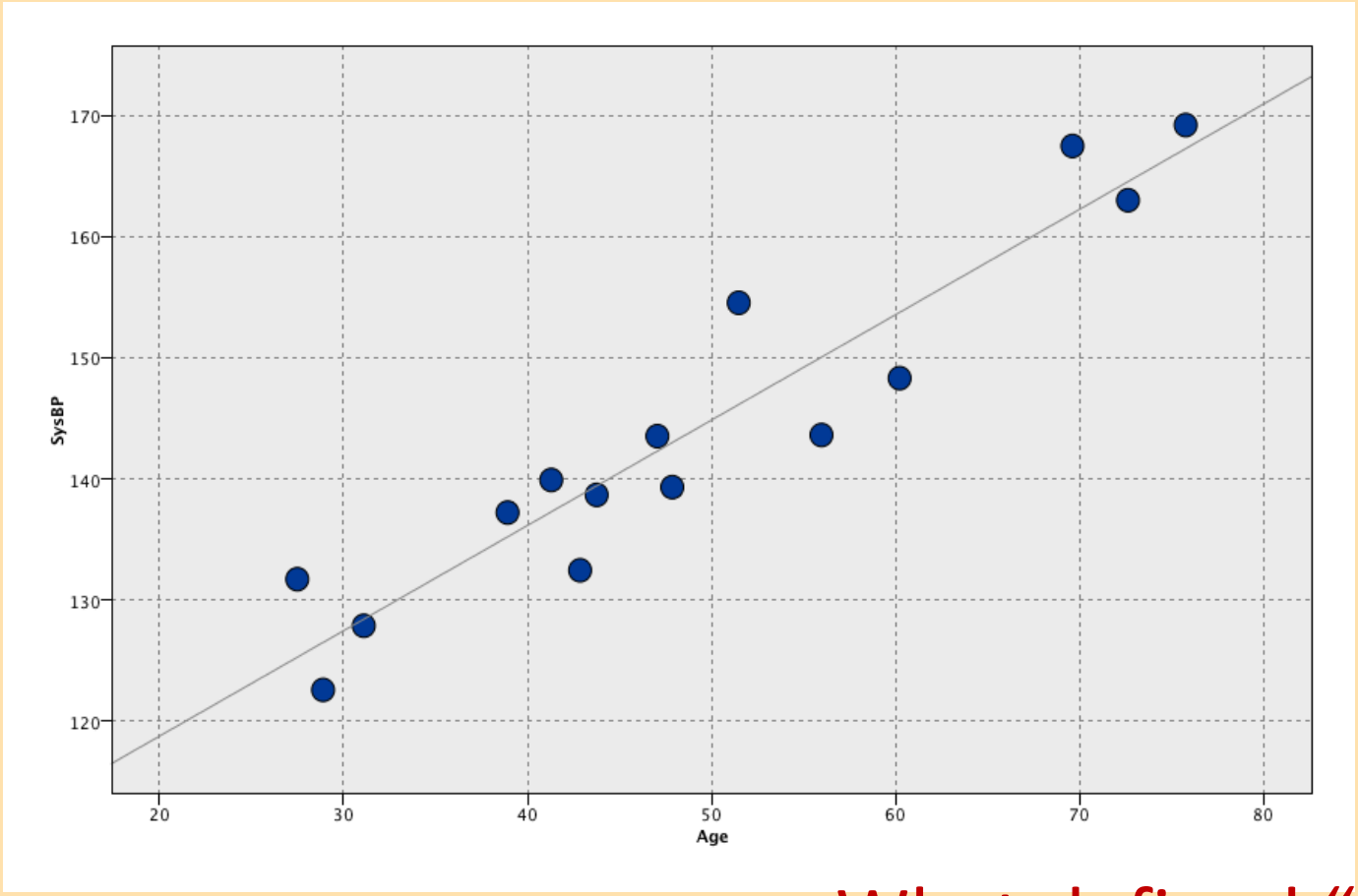



# IBM Modeler Linear Node



# Fitted line from Linear node

SysBP




What defined “best”?

Age

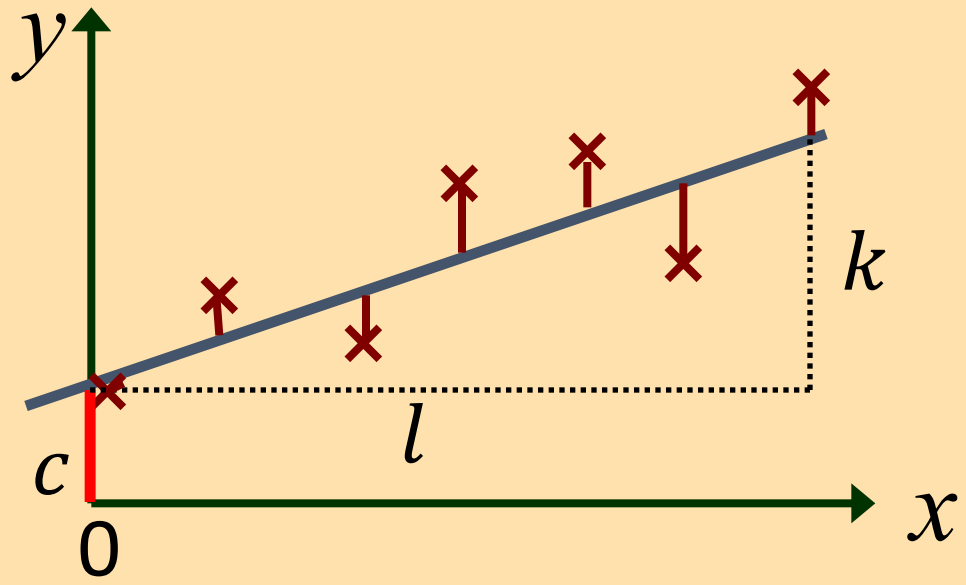
# Model Fitting

- Classical statistical models are conventionally fit by *maximum likelihood*
- Equivalent to *least squares fitting* for linear regression/model

  Find the line such that the sum of all the square vertical distances from data points to the line is minimized
- Classical statistical inference focuses on the regression relationship between $x$ and $y$ -- usually concerned with whether there is evidence that the slope $a$ is non-zero and what is its estimated value (and uncertainty)?
- In statistical/machine learning *emphasis* is more on using estimates of $a$ and $c$ to predict future values of $y$ from future values of $x$ -- interested in predictors that add predictive value, not evidence for non-zero relationship

# Least Squares

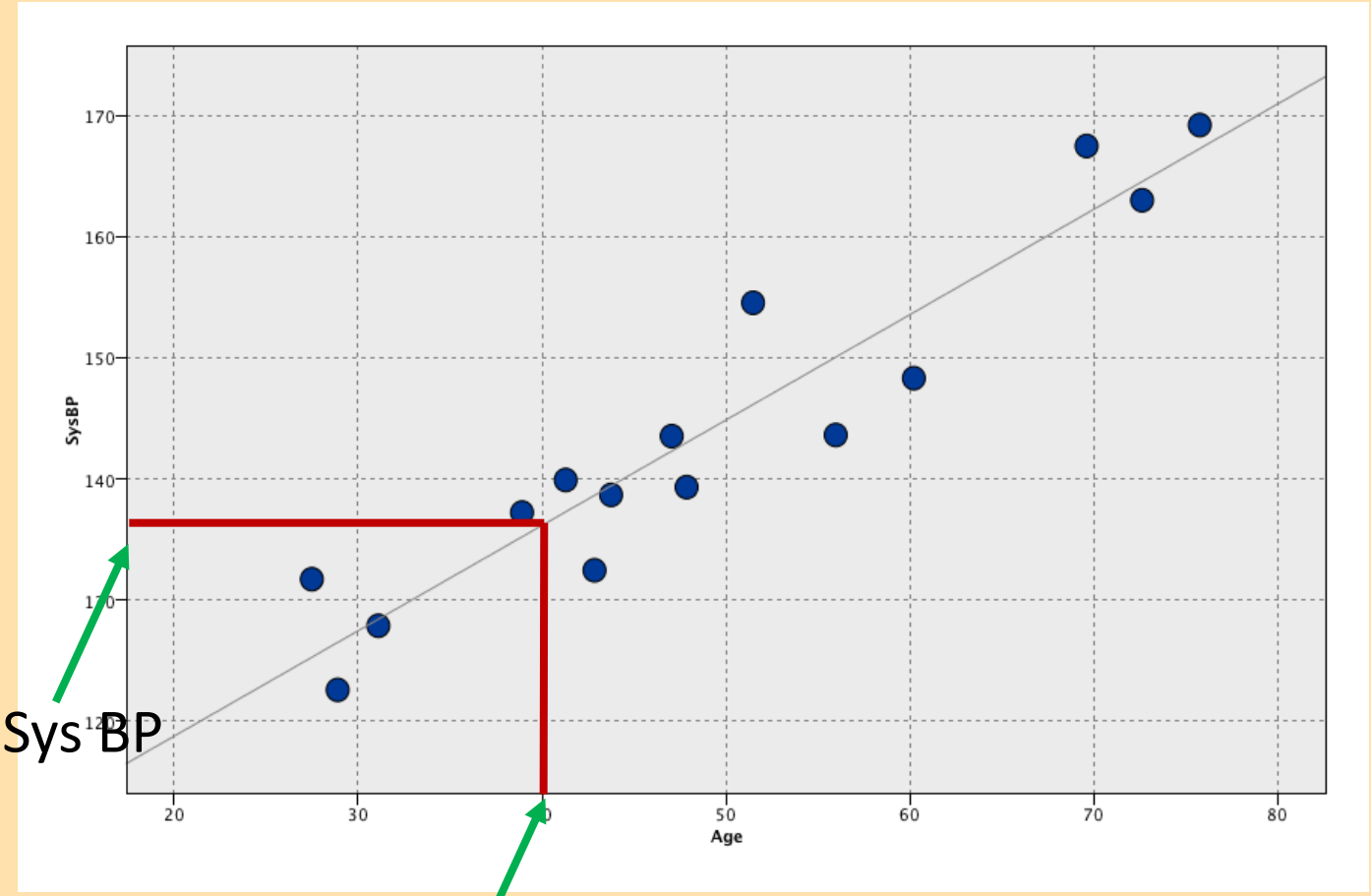
- Model relationship between $x$ and $y$ (e.g. linear)
- Assume error is in $y$, not in $x$
- Find the line such that it gives you the smallest *sum of squared* errors




$$a = \frac{k}{l}$$

Add the squared values of the vertical red lines together – choose line that minimizes this total

# Making a prediction




SysBP

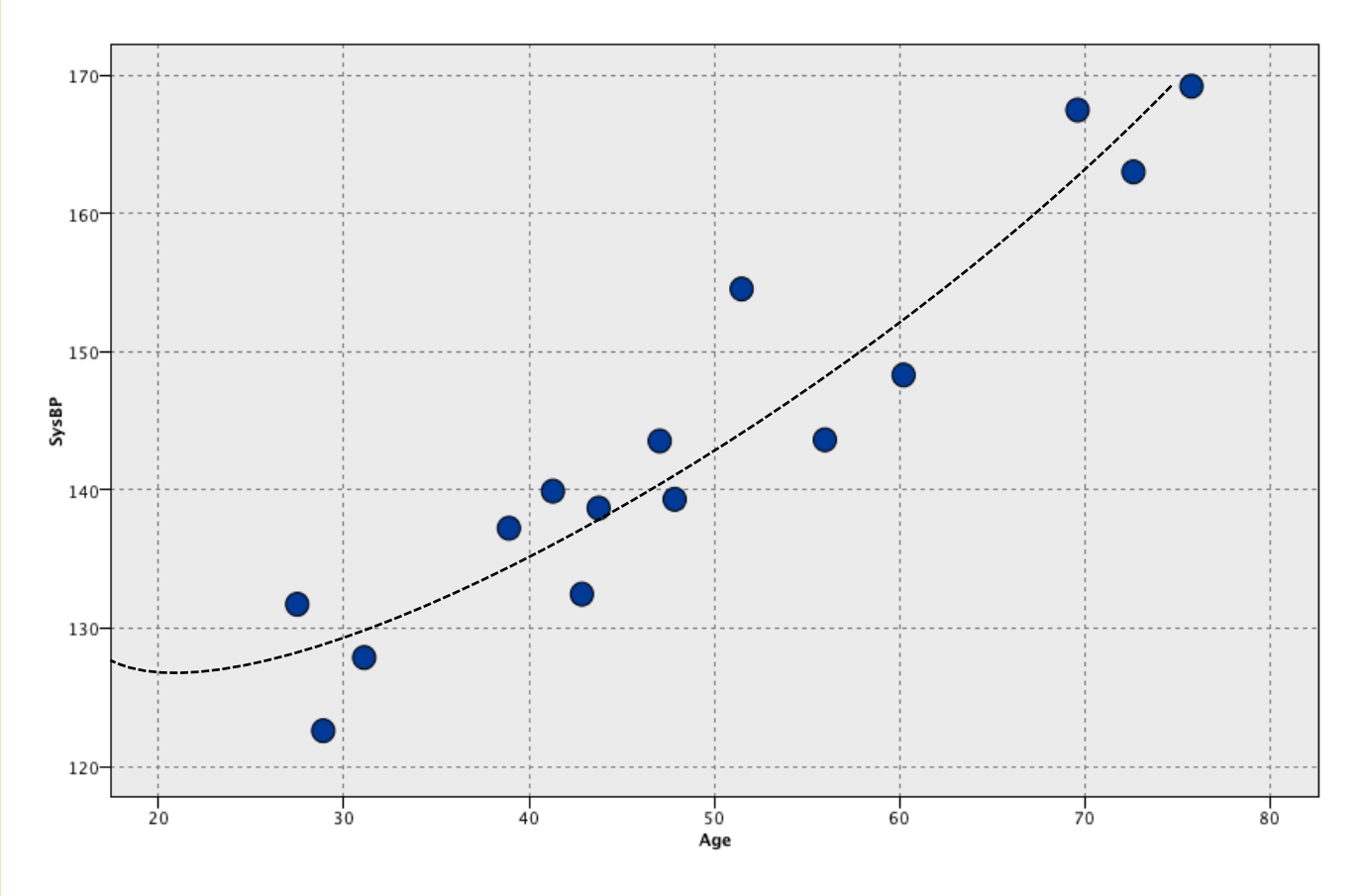
Predicted Sys BP

New subject age 40

Age

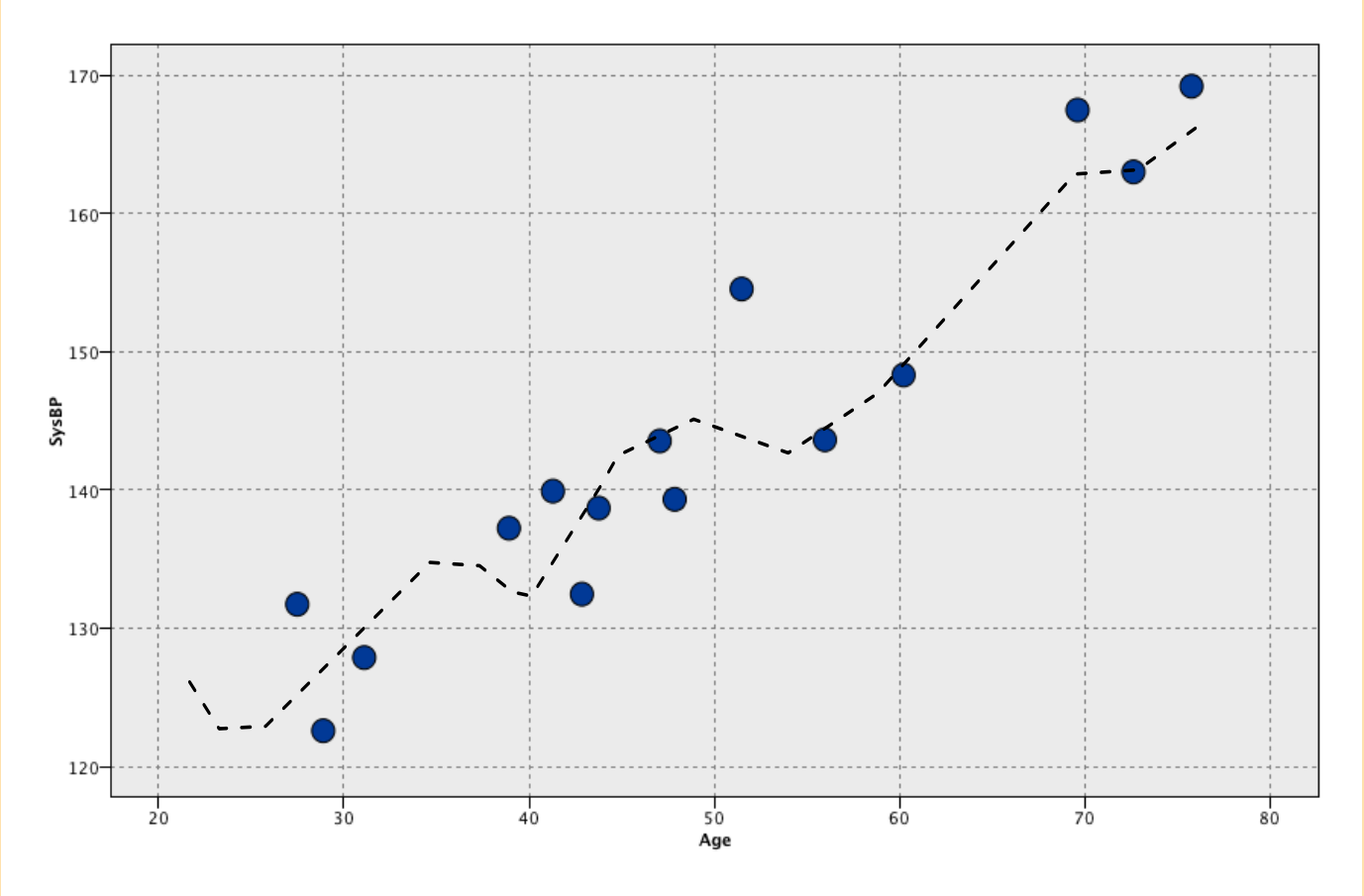
# Why straight line? What about?

SysBP




Age

# Or even?

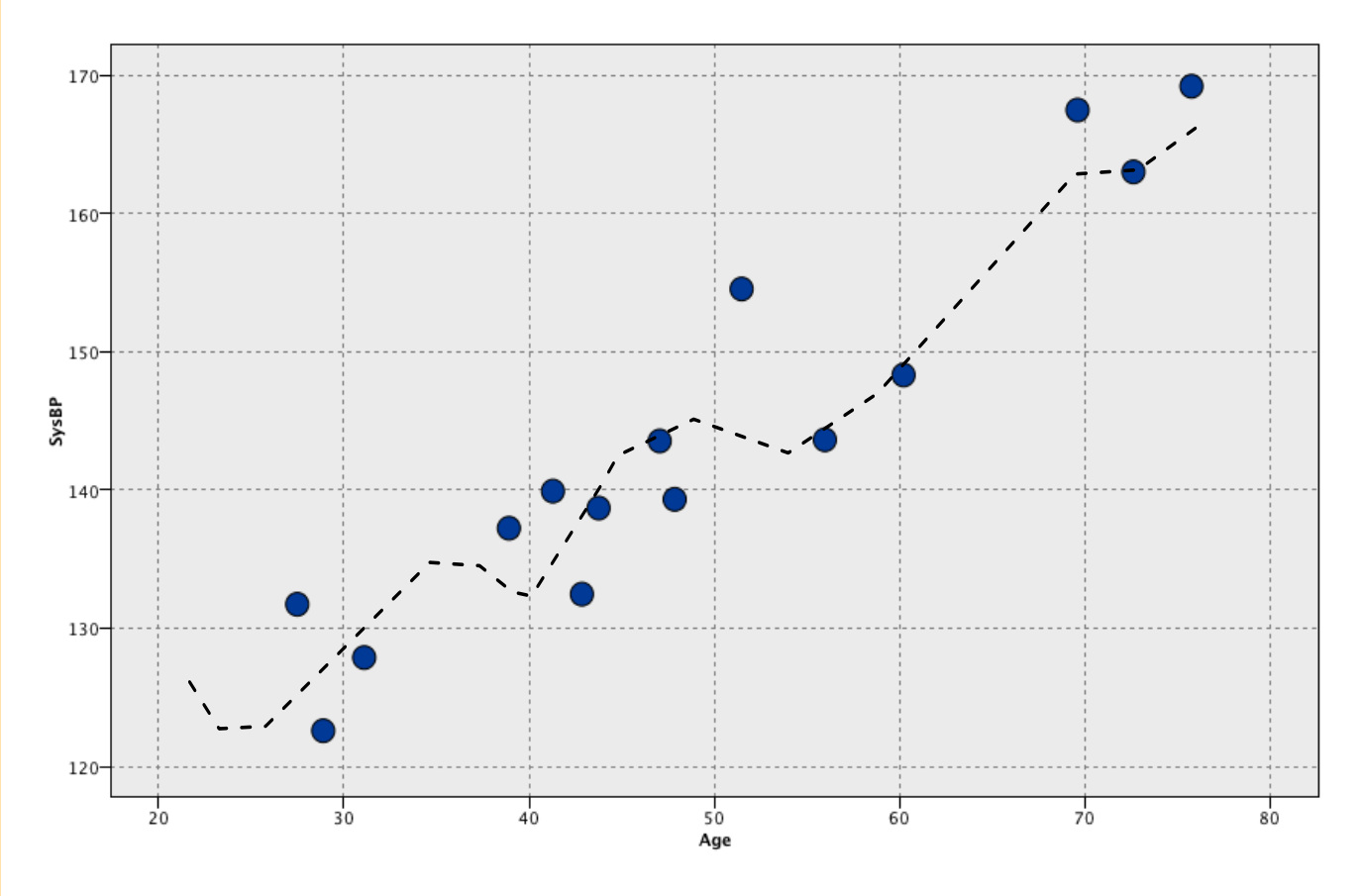



SysBP

Age

# Or even?

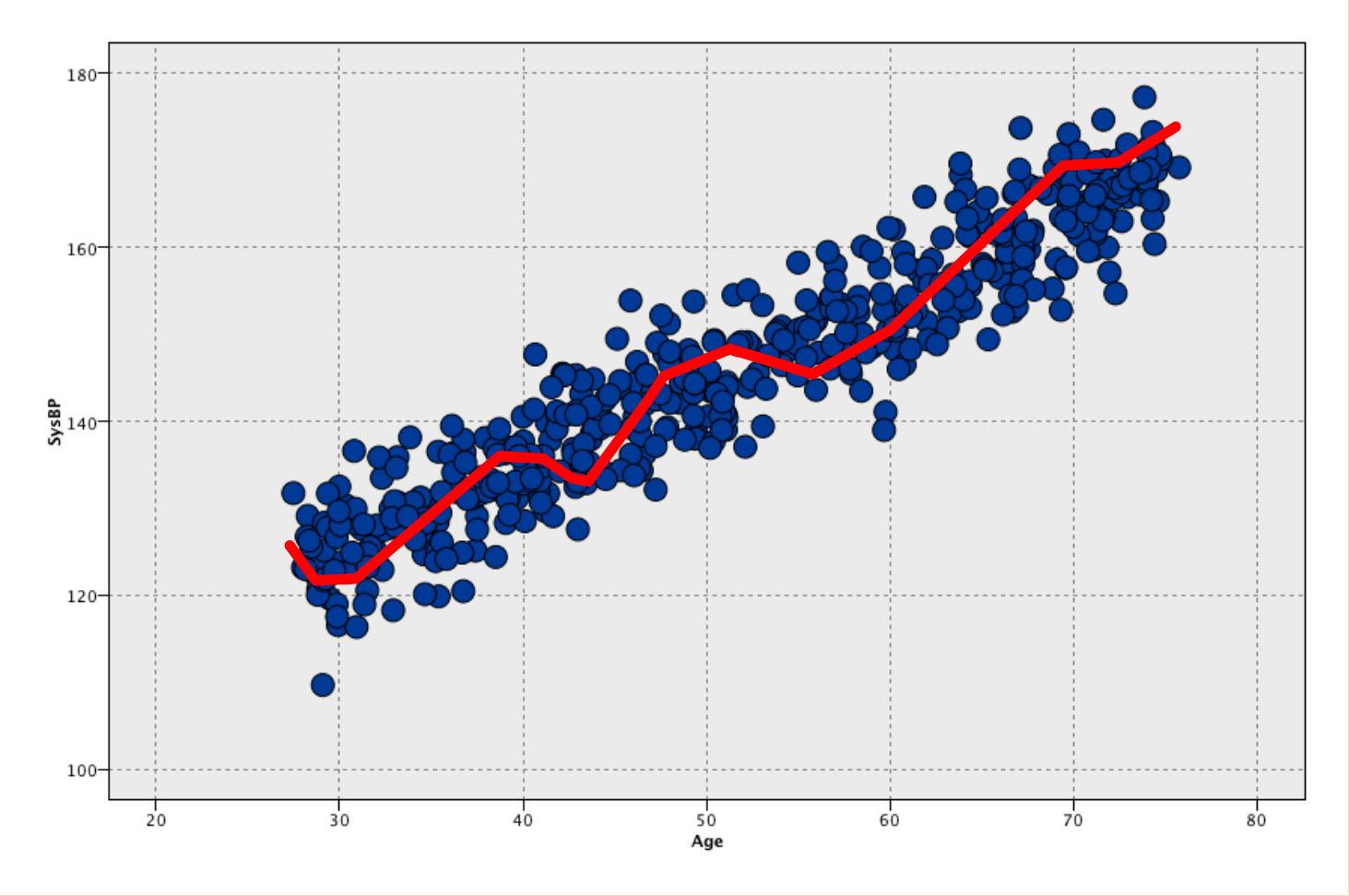
SysBP




Age

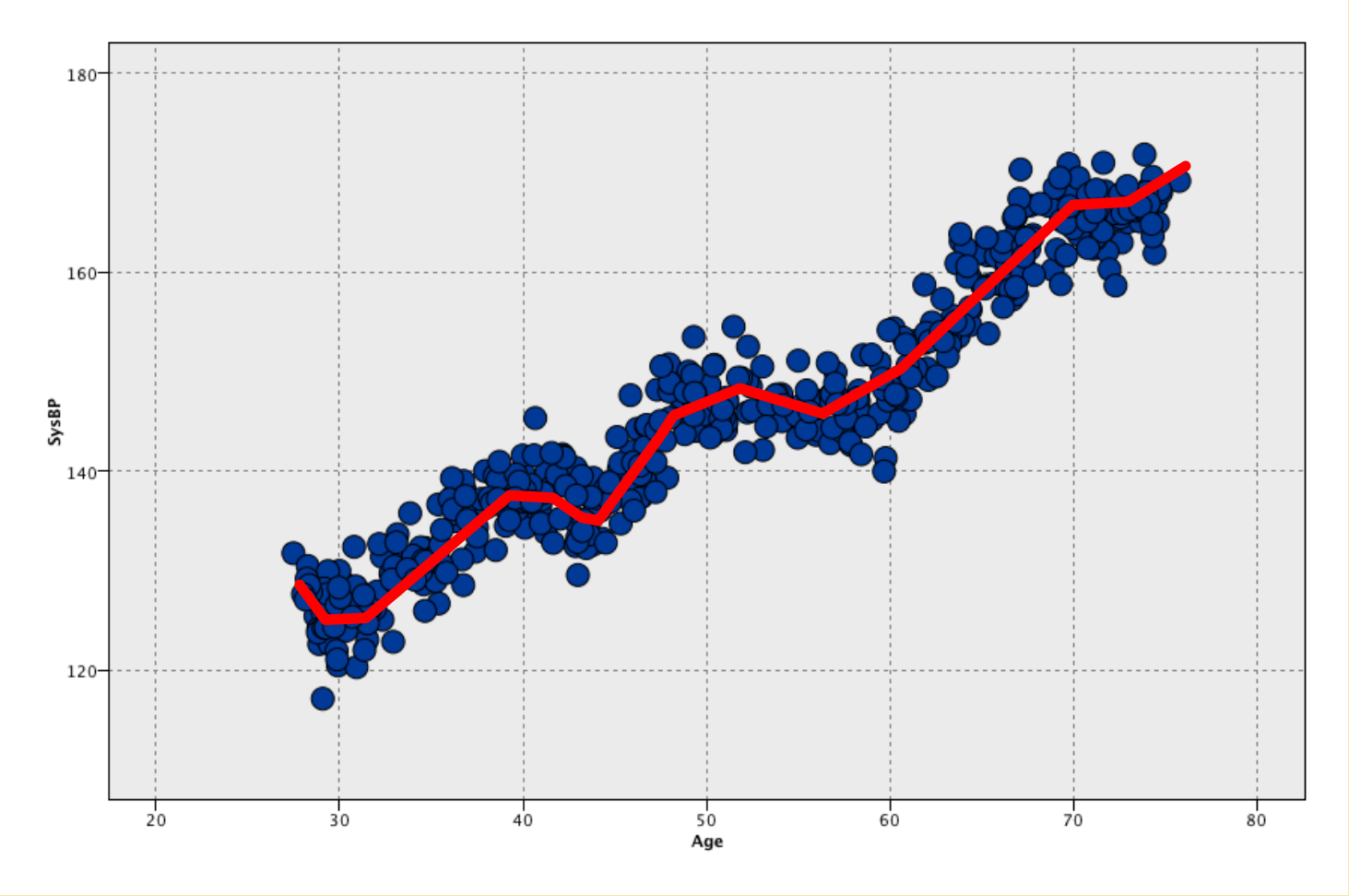
**Overfitted?**

# What about with more data?




Age

# What about this data?




Age

# Big Data – Large N and Large P

- Big data can be thought of in two ways
- Large N = large number of subjects – implies more information for fitting complex models or ones with many predictors
- Large P = large number of predictors, or complex relationships between predictor and outcome, or between predictors (interaction terms)

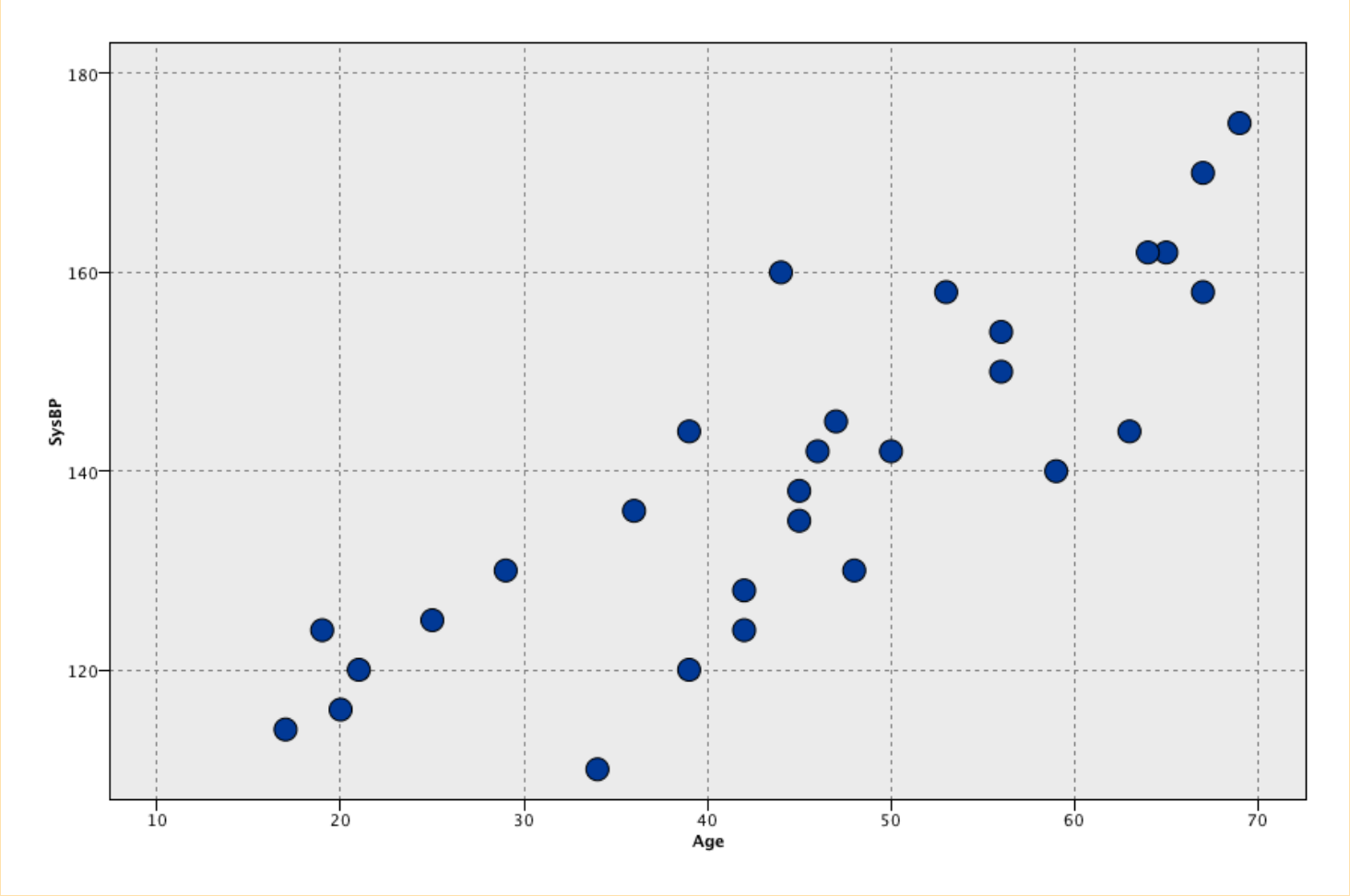
# The trade-off

- Want to find true predictive patterns in the data
- Idea is to use some measure of “prediction quality” in order to assess what level of complexity is best given a particular dataset
- Want to avoid overfitting spurious detail: i.e. don’t want to make the model capture apparent patterns due to noise
- Solution – independent validation – reserve some data that has not been used in model fitting for the purpose of model evaluation and comparison – we will look at this briefly later, and in more detail on Thursday

# The real SysBP dataset

- Helmut Spaeth, Mathematical Algorithms for Linear Regression, Academic Press, 1991, page 304, ISBN 0-12-656460-4.
- Data available at: http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html as the x03.txt dataset
- I will put Modeler friendly version on CLE
- The systolic blood pressure was measured for 30 people of different ages.

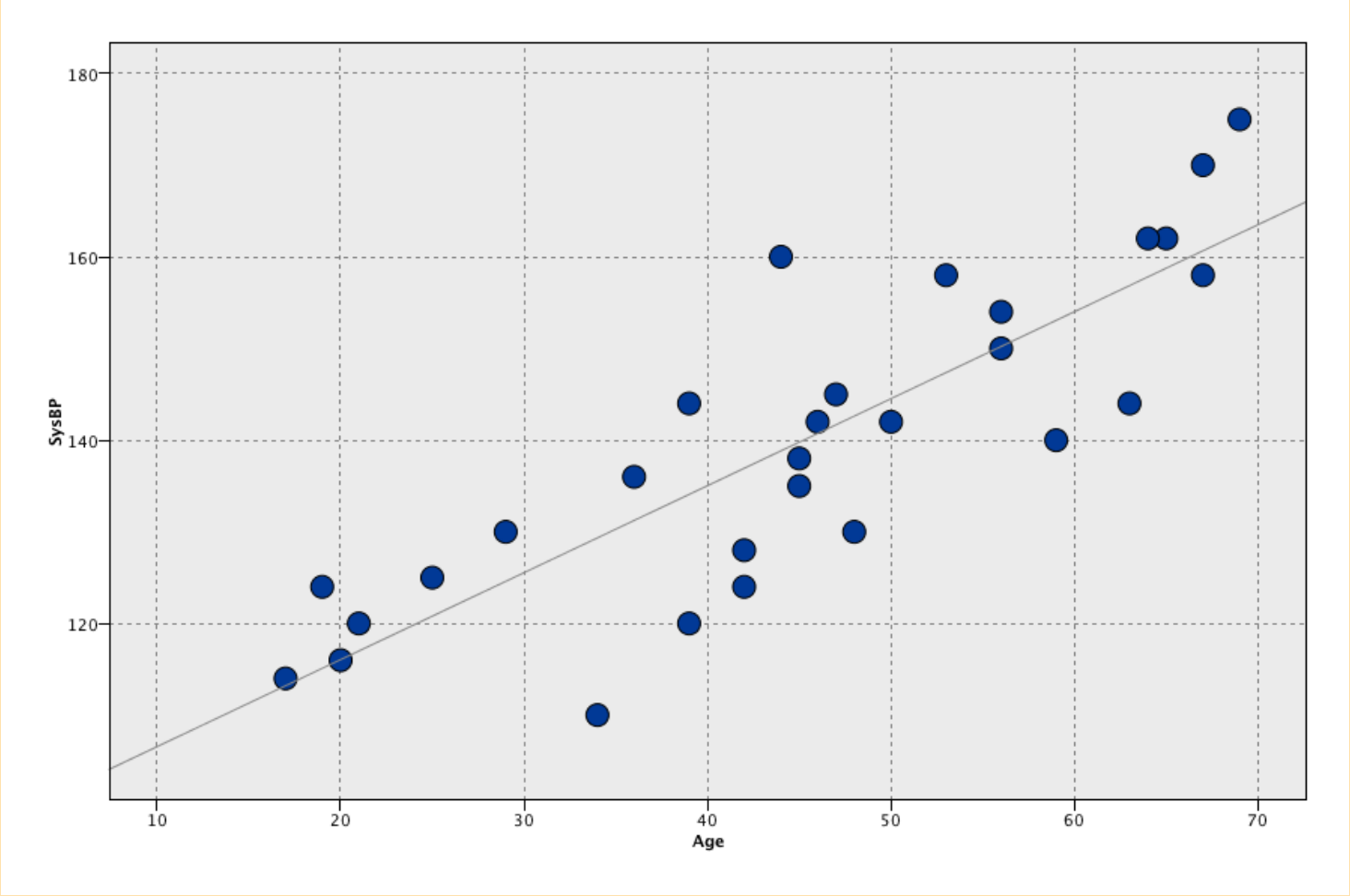
# The SysBP dataset




SysBP

Age

# The real dataset with fitted line

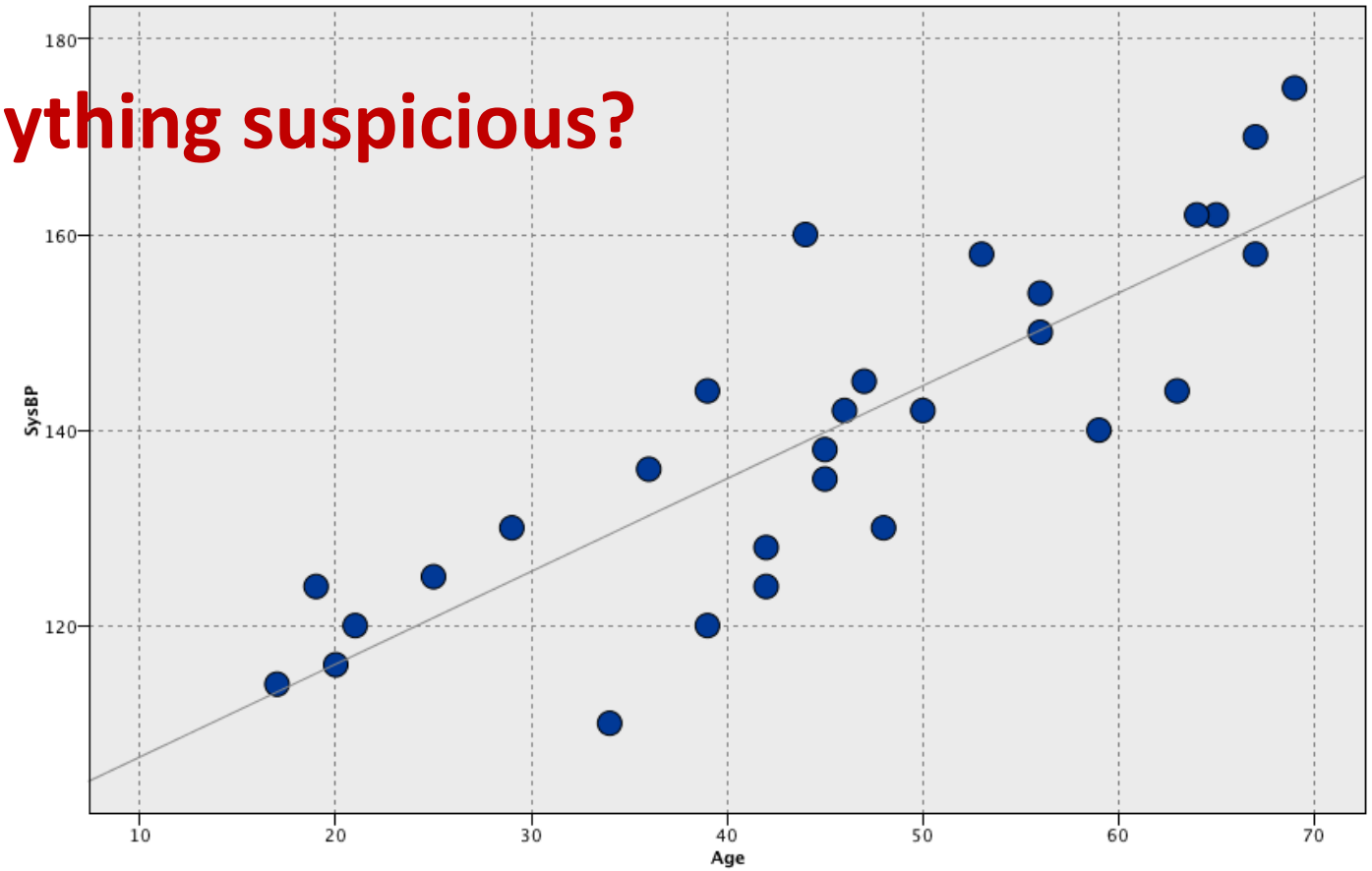



SysBP

Age

# The real dataset

**Anything suspicious?**

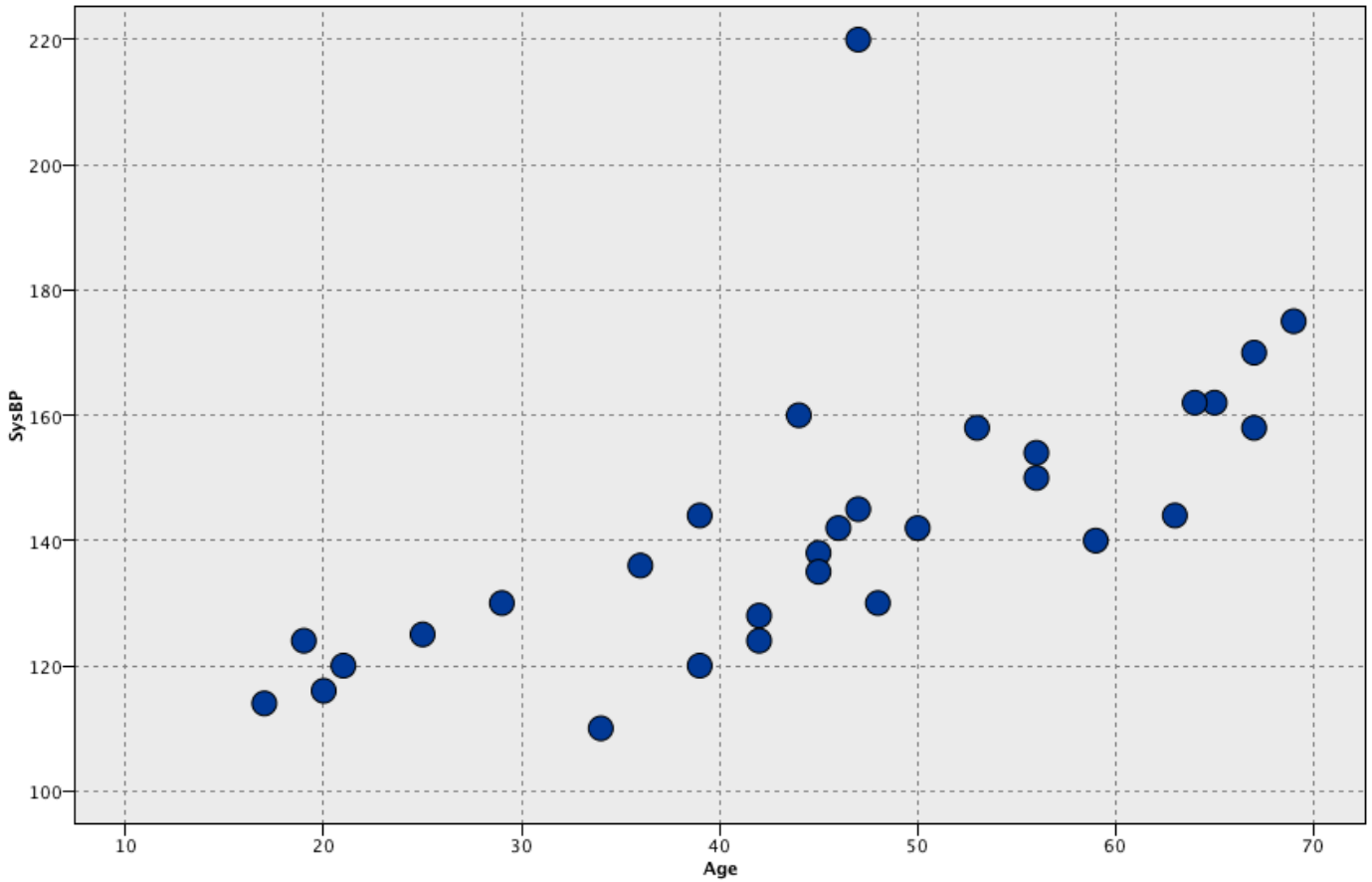



SysBP

Age

# High blood pressure subject

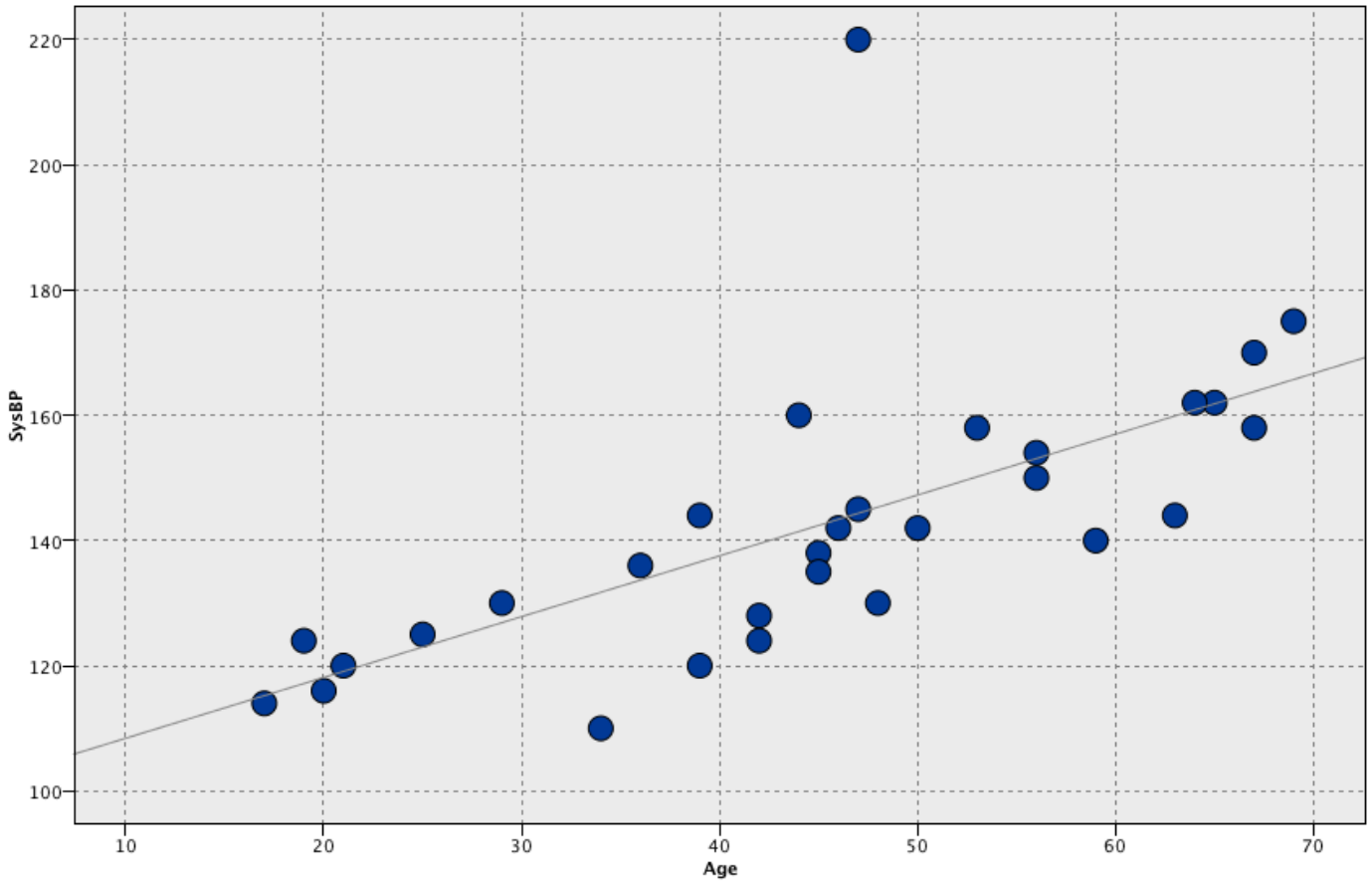
SysBP




Age

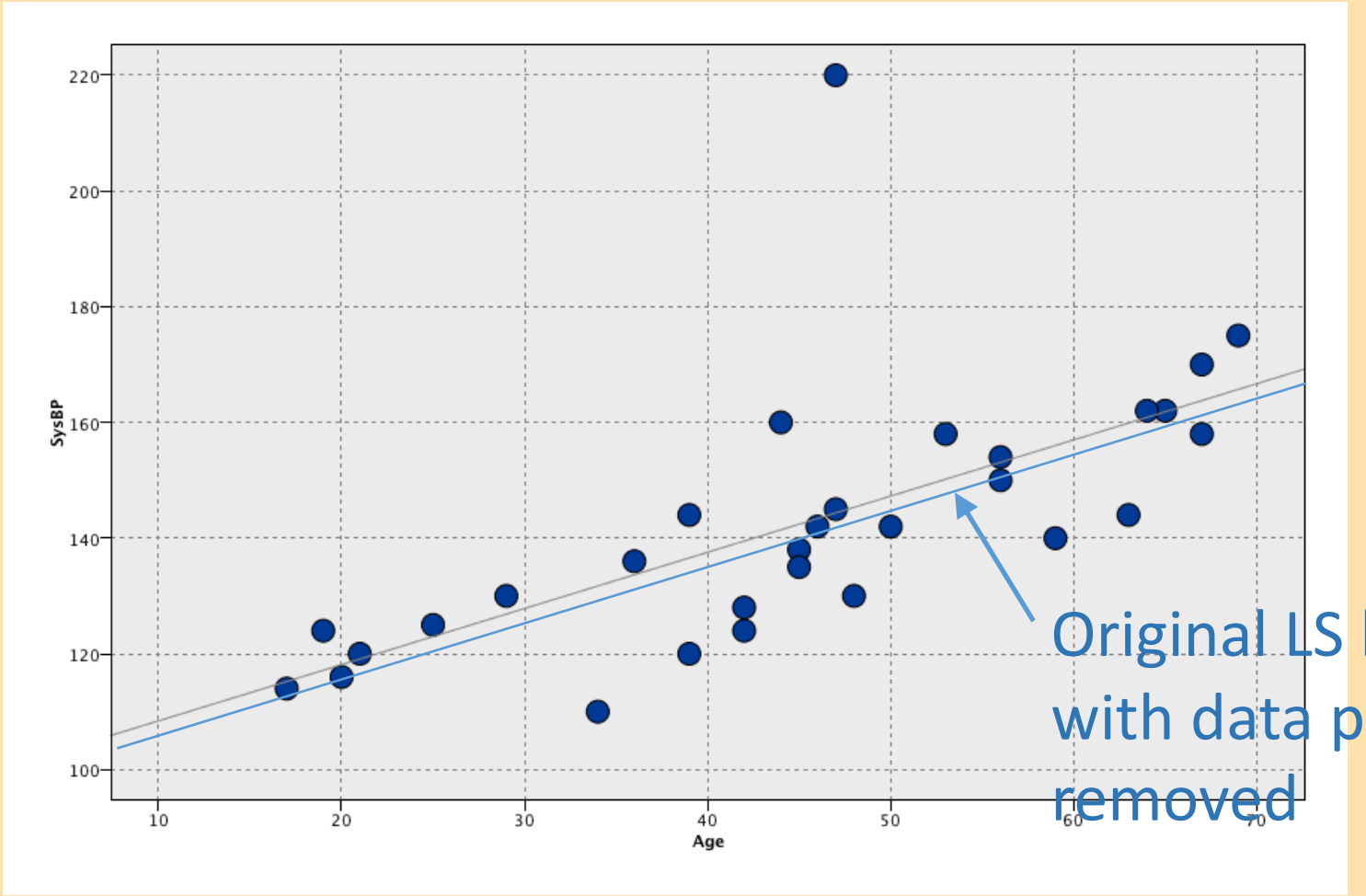
# Fitted line

SysBP




Age

# The real dataset



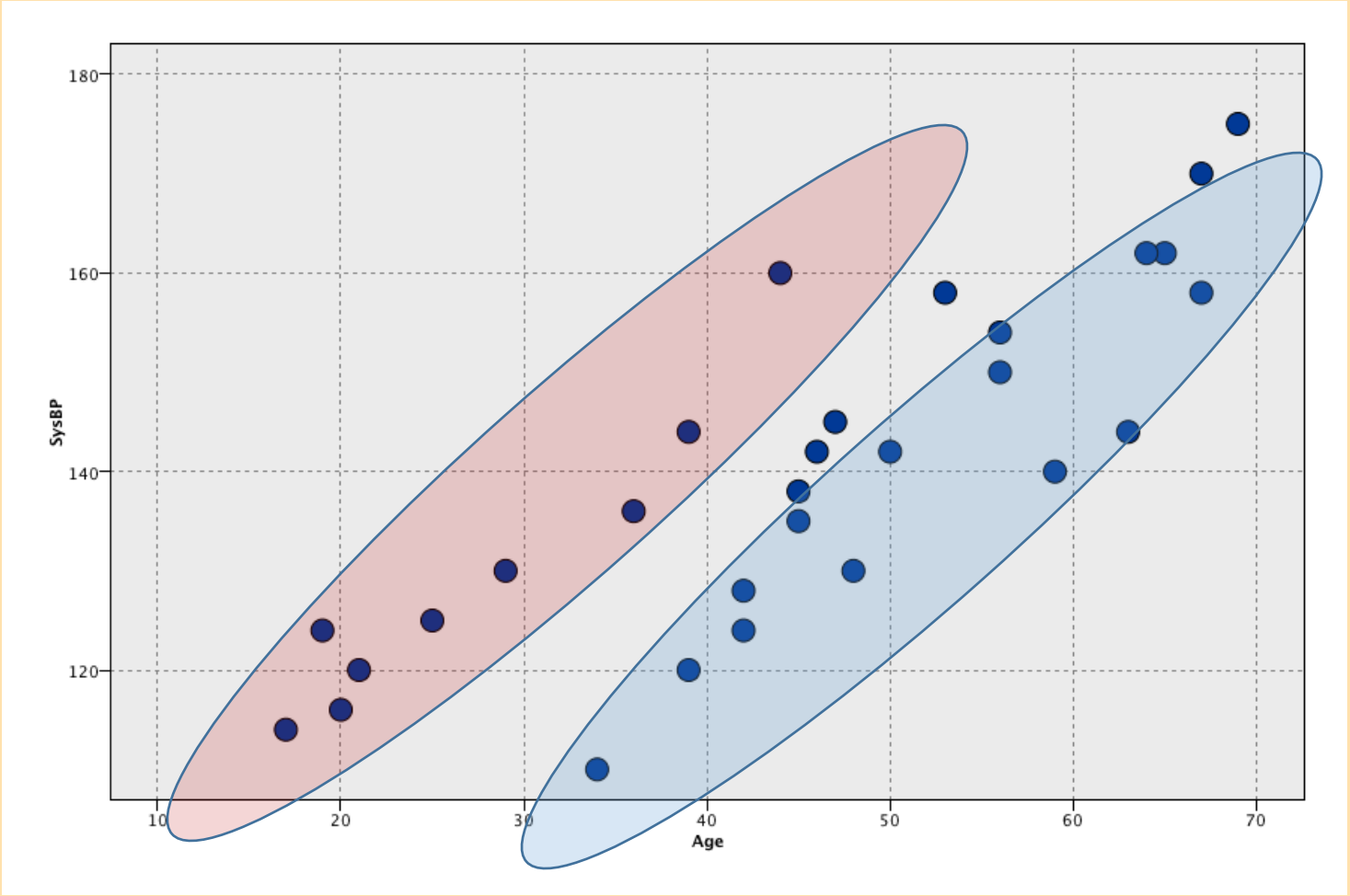

SysBP

Age

# Is anything going on here?

Continuing with outlier dropped…

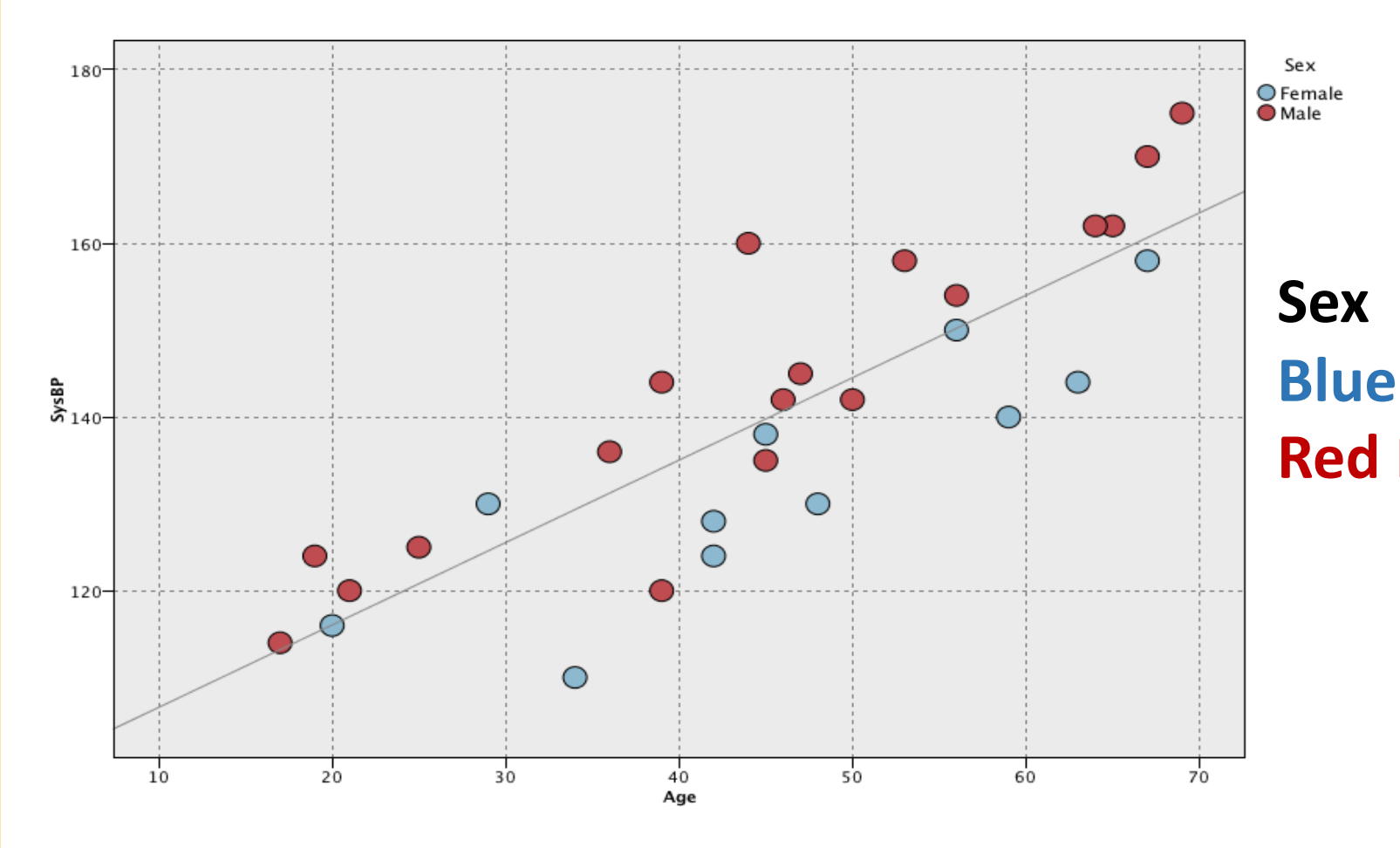
SysBP




Age

# Separate by Sex

SysBP




Age

# Multi-predictor models

Now we might entertain a more complicated model:

$$y = ax + bz + c$$

e.g. $y$ = Systolic BP (SysBP)

$x$ = Age

$z$ = Sex (1 = M, 0 = F)

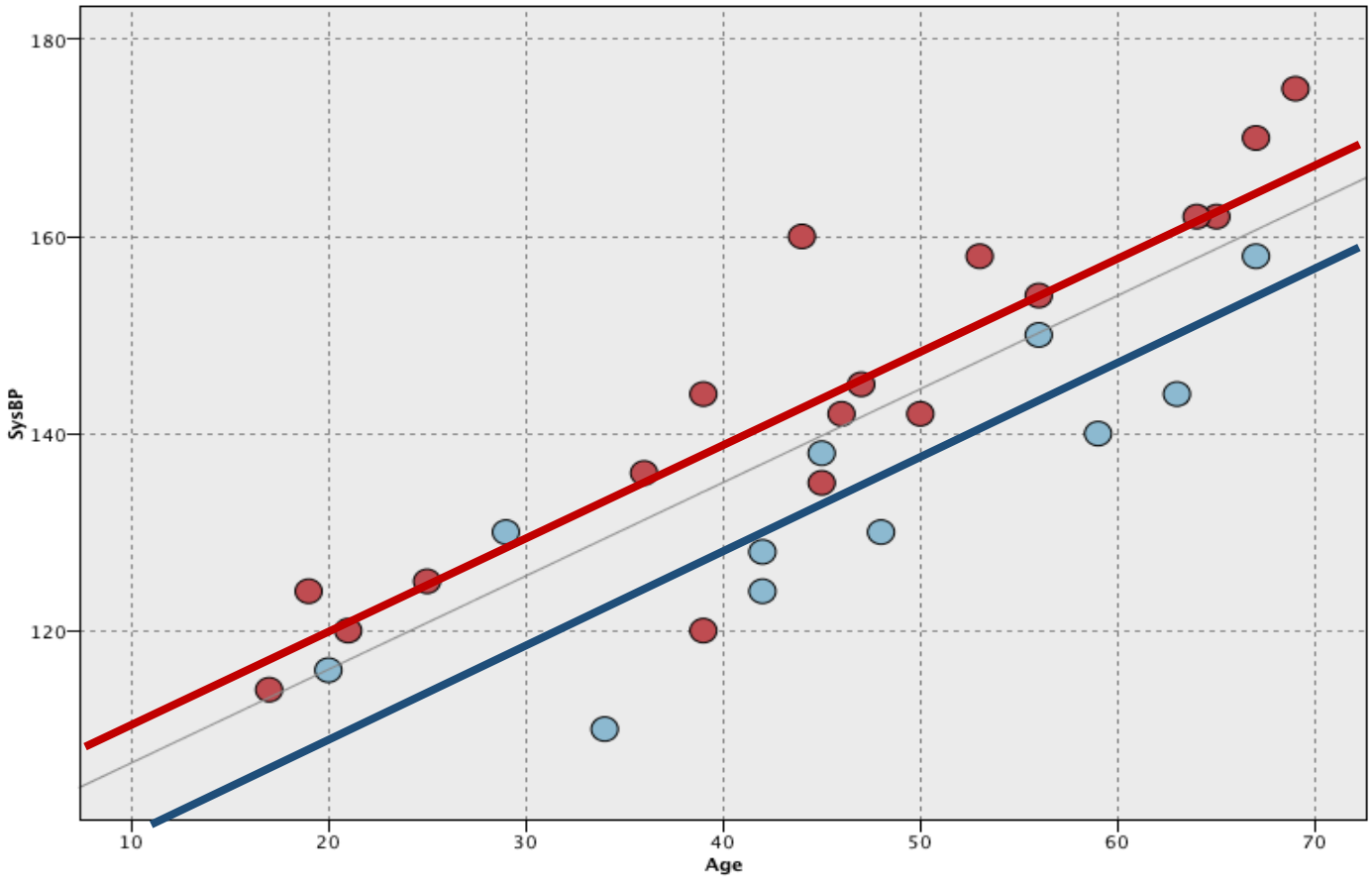
$c$ = intercept (SysBP at age 0 and Sex = F)

$a$ = change SysBP per year in Age

$b$ = difference in SysBP for M vs. F

Question: Are Age and Sex predictive of SysBP?

# Age *and* Sex as predictors




**Sex**

**Blue F**

**Red M**

Gray line = original fit, no sex predictor

SysBP

Age

# And interactions

But why not even more complicated such as:

$$y = ax + bz + abw + c$$

e.g. $y$ = Systolic BP (SysBP)

$x$ = Age

$z$ = Sex (1 = M, 0 = F)

$c$ = intercept (SysBP at age 0 and Sex = F)

$a$ = change in SysBP per year in Age for Sex = F

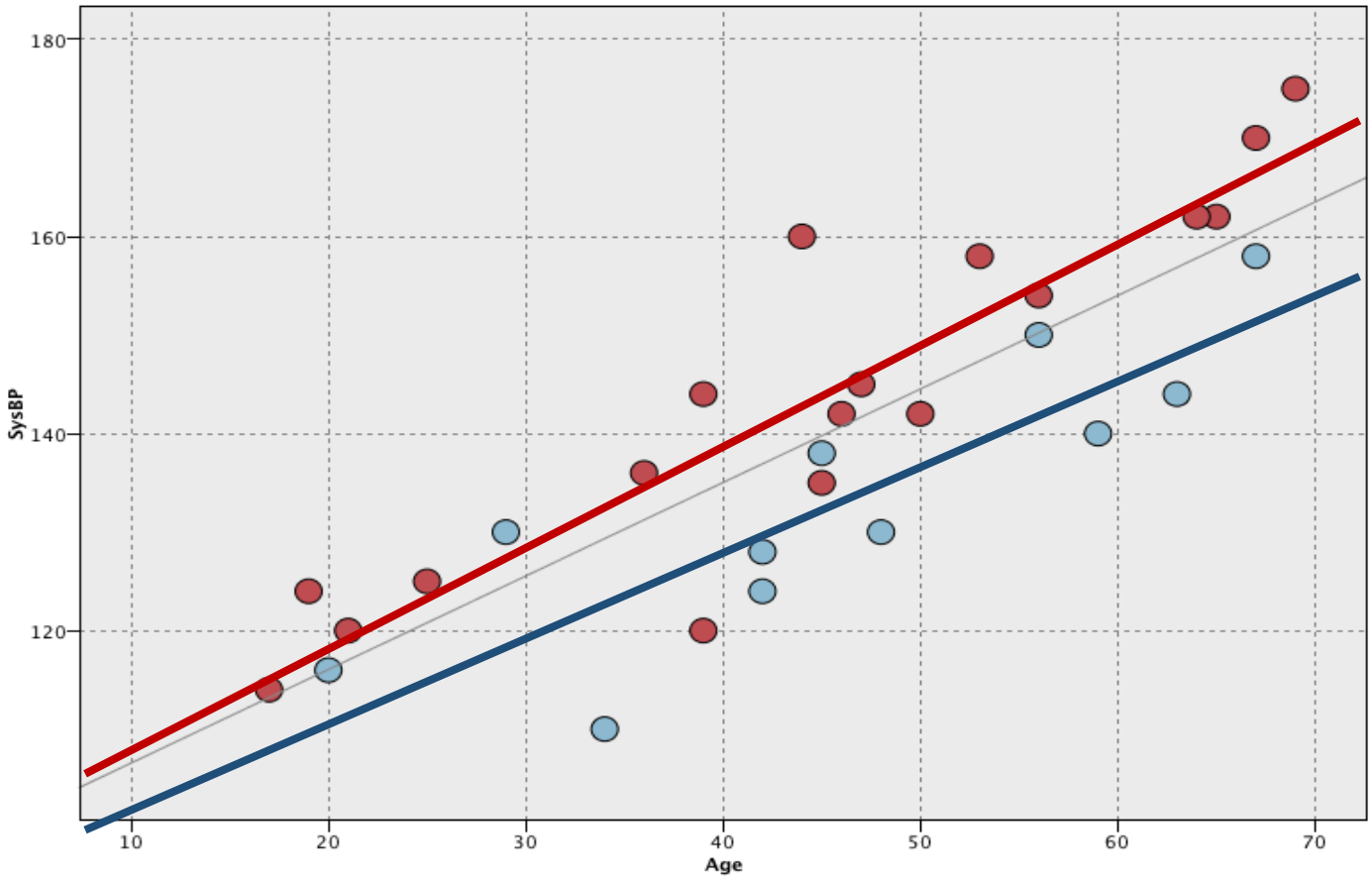
$b$ = difference in SysBP for M vs. F at Age = 0

$w$ = interaction effect - difference in SysBP per year in age for Sex = M vs. Sex = F

Question: Are Age and Sex and interaction of Age and Sex predictive of SysBP?

# Age and Sex interaction

SysBP




**Sex**

**Blue F**

**Red M**

Gray line = original fit, no sex predictor

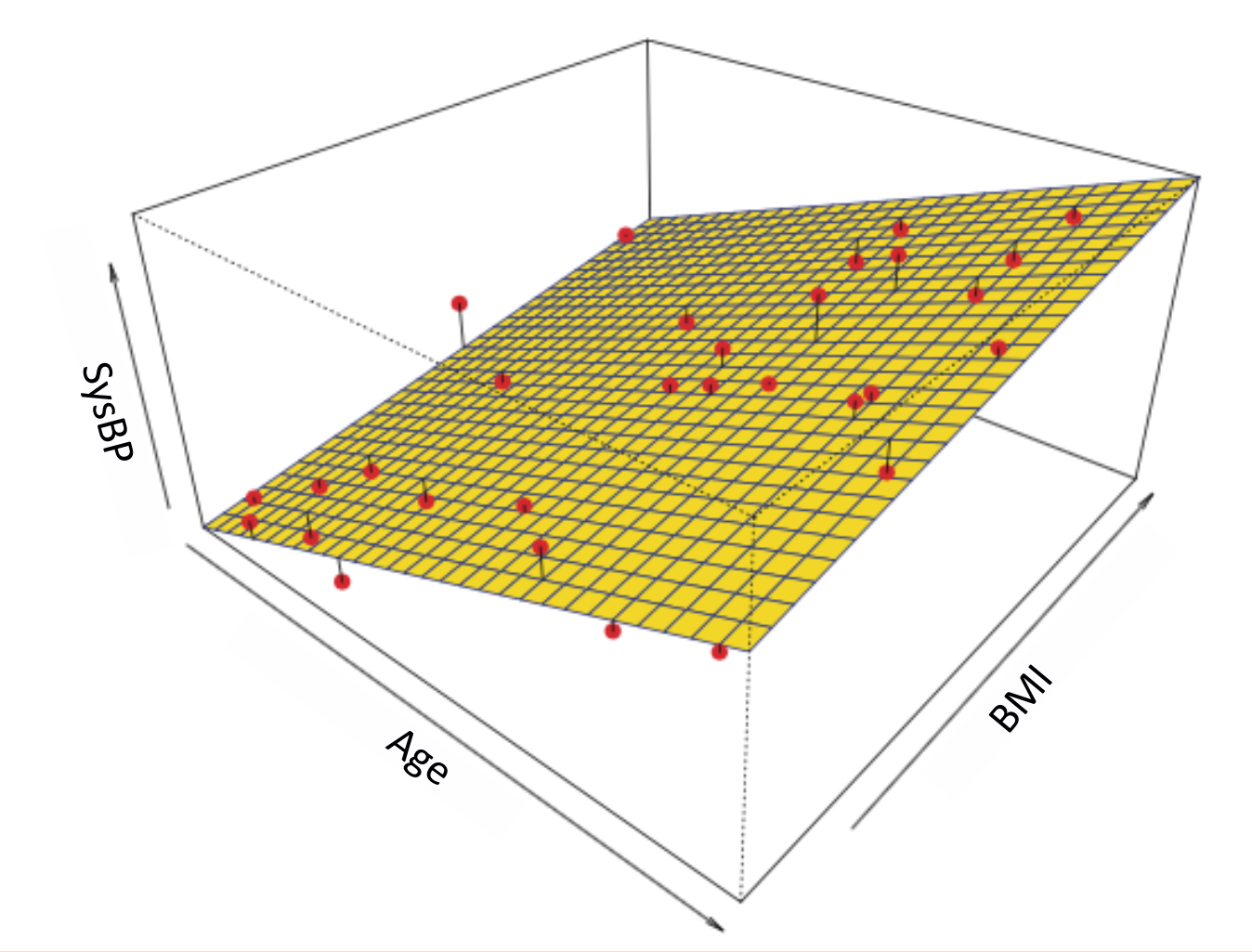
Age

# Sorry, I lied – there is no sex variable – I made it up to illustrate

I will make the dataset available on CLE in case you want to play with it…

# Linear least squares with 2 (continuous) predictors

Adapted From ISLR



# Many predictors

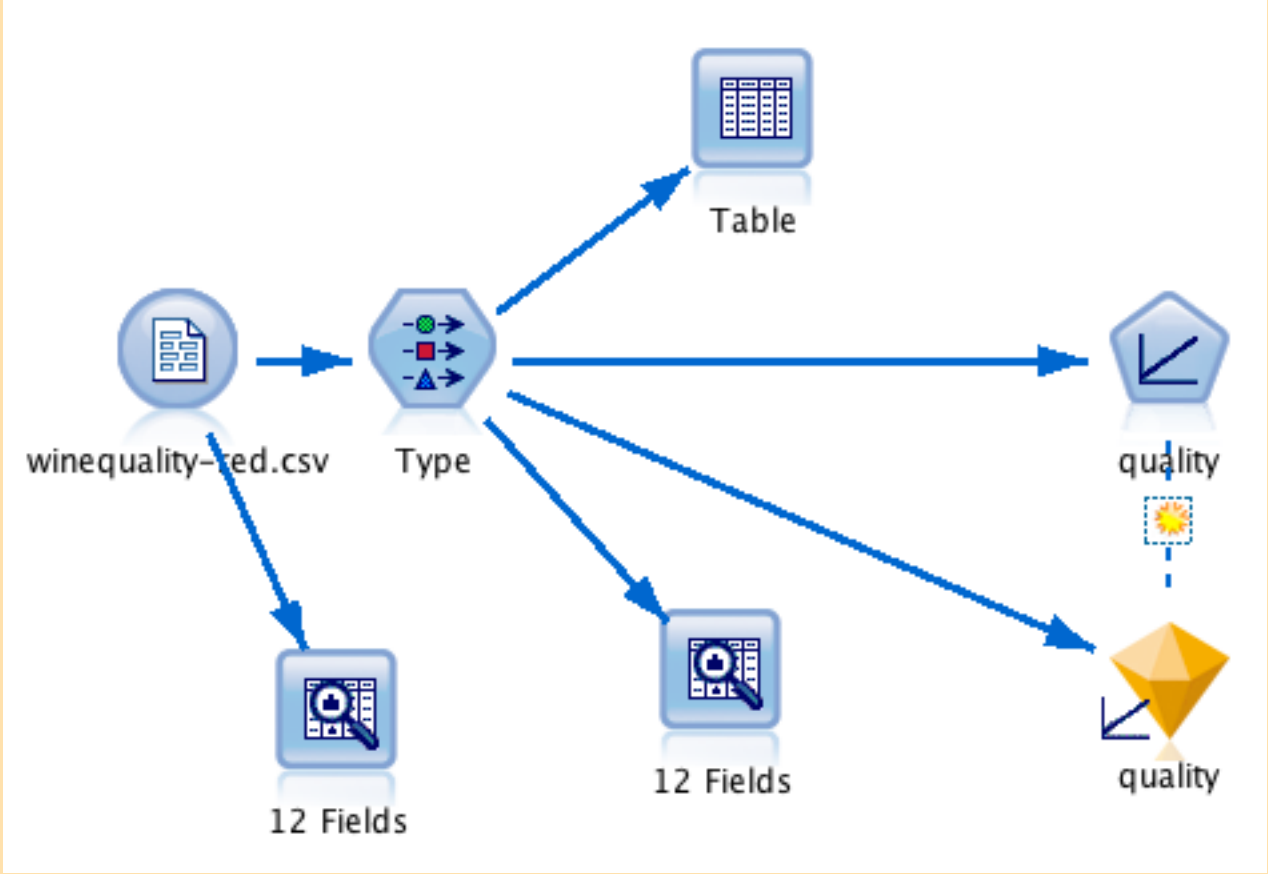
$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

- Fit $n$-dimensional hyper-plane in $(n + 1)$-dimensional space
- The goal of some methods is to find manageable subset of parameters that captures most of predictive information – variable selection method -- Question: Which predictors $x_1, x_2, \ldots, x_n$ are predictive of $y$?
- Other methods don't worry about number of predictors explicitly, but rather just build a predictive "black box"

# Wine dataset

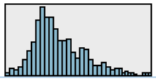
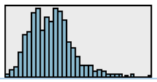
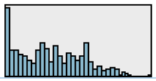
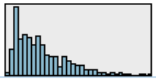
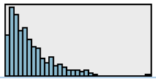
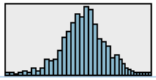
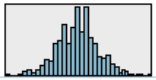
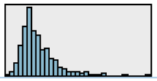
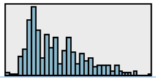
- *Wine Quality* dataset from UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets.html (I will also place on CLE)
- Freely available dataset relating to Portuguese “Vinho Verde”
- P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.
- Use objective test predictors (alcohol level, PH, acidity, etc.) to predict wine expert subjective assessments: median of 3 experts scoring 0 (very bad) to 10 (very excellent)
- No missing attributes
- We will focus on the “red variant” dataset

# Multiple Regression Stream



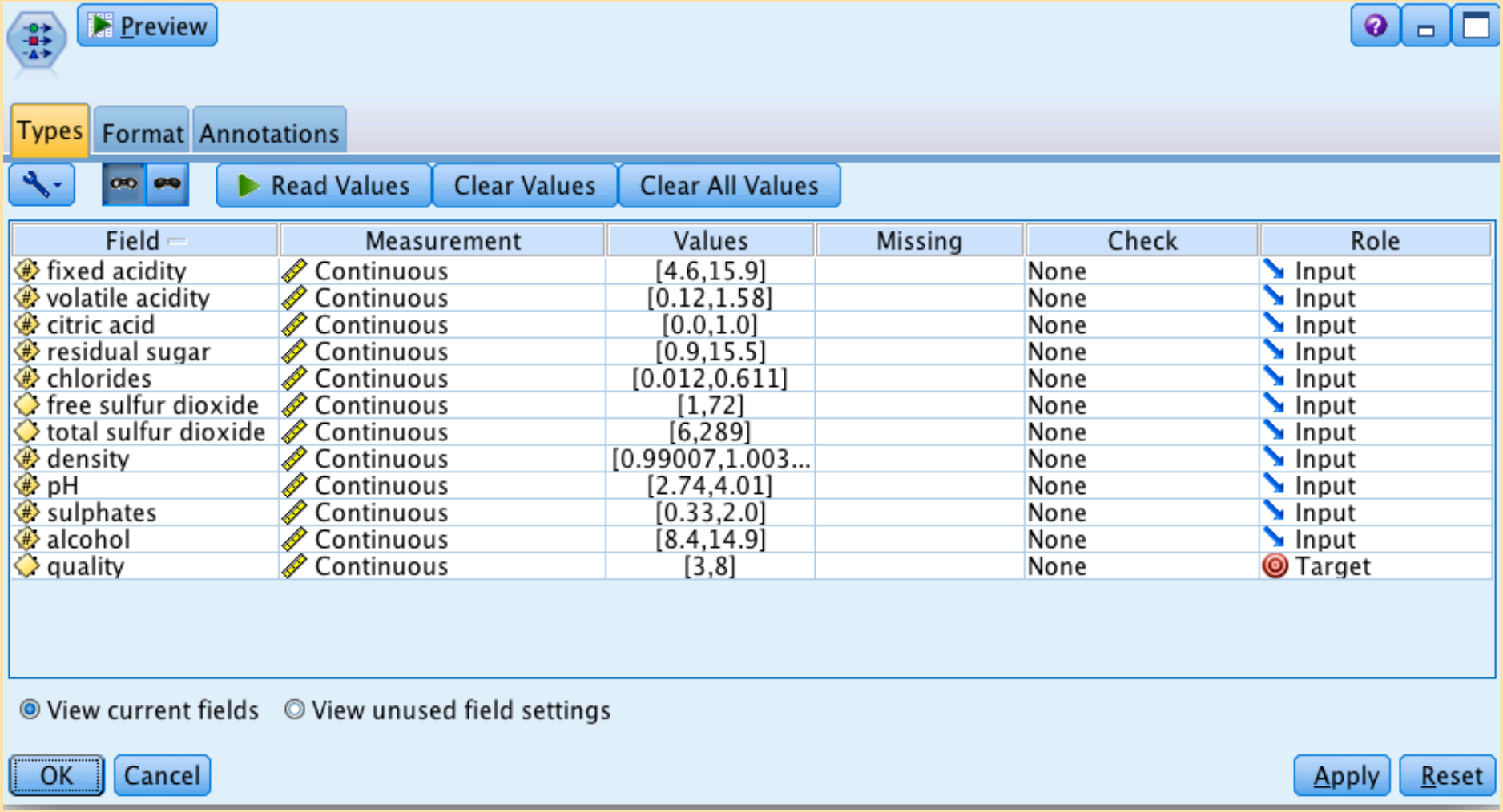
# Data Audit prior to Type

Audit | Quality | Annotations

| Field | Graph | Measurement | Min | Max | Mean | Std. Dev | Skewness | Unique | Valid |
|---|---|---|---|---|---|---|---|---|---|
| fixed acidity |  | Continuous | 4.600 | 15.900 | 8.320 | 1.741 | 0.983 | -- | 1599 |
| volatile acidity |  | Continuous | 0.120 | 1.580 | 0.528 | 0.179 | 0.672 | -- | 1599 |
| citric acid |  | Continuous | 0.000 | 1.000 | 0.271 | 0.195 | 0.318 | -- | 1599 |
| residual sugar |  | Continuous | 0.900 | 15.500 | 2.539 | 1.410 | 4.541 | -- | 1599 |
| chlorides |  | Continuous | 0.012 | 0.611 | 0.087 | 0.047 | 5.680 | -- | 1599 |
| free sulfur dioxi... |  | Continuous | 1 | 72 | 15.874 | 10.458 | 1.250 | -- | 1599 |
| total sulfur diox... |  | Continuous | 6 | 289 | 46.467 | 32.895 | 1.516 | -- | 1599 |
| density |  | Continuous | 0.990 | 1.004 | 0.997 | 0.002 | 0.071 | -- | 1599 |
| pH |  | Continuous | 2.740 | 4.010 | 3.311 | 0.154 | 0.194 | -- | 1599 |
| sulphates |  | Continuous | 0.330 | 2.000 | 0.658 | 0.170 | 2.429 | -- | 1599 |
| alcohol |  | Continuous | 8.400 | 14.900 | 10.423 | 1.066 | 0.861 | -- | 1599 |
| quality |  | Continuous | 3 | 8 | 5.636 | 0.808 | 0.218 | -- | 1599 |

[1] Indicates a multimode result [2] Indicates a sampled result

# Type Node




| Field | Measurement | Values | Missing | Check | Role |
|---|---|---|---|---|---|
| fixed acidity | Continuous | [4.6,15.9] | | None | Input |
| volatile acidity | Continuous | [0.12,1.58] | | None | Input |
| citric acid | Continuous | [0.0,1.0] | | None | Input |
| residual sugar | Continuous | [0.9,15.5] | | None | Input |
| chlorides | Continuous | [0.012,0.611] | | None | Input |
| free sulfur dioxide | Continuous | [1,72] | | None | Input |
| total sulfur dioxide | Continuous | [6,289] | | None | Input |
| density | Continuous | [0.99007,1.003... | | None | Input |
| pH | Continuous | [2.74,4.01] | | None | Input |
| sulphates | Continuous | [0.33,2.0] | | None | Input |
| alcohol | Continuous | [8.4,14.9] | | None | Input |
| quality | Continuous | [3,8] | | None | Target |

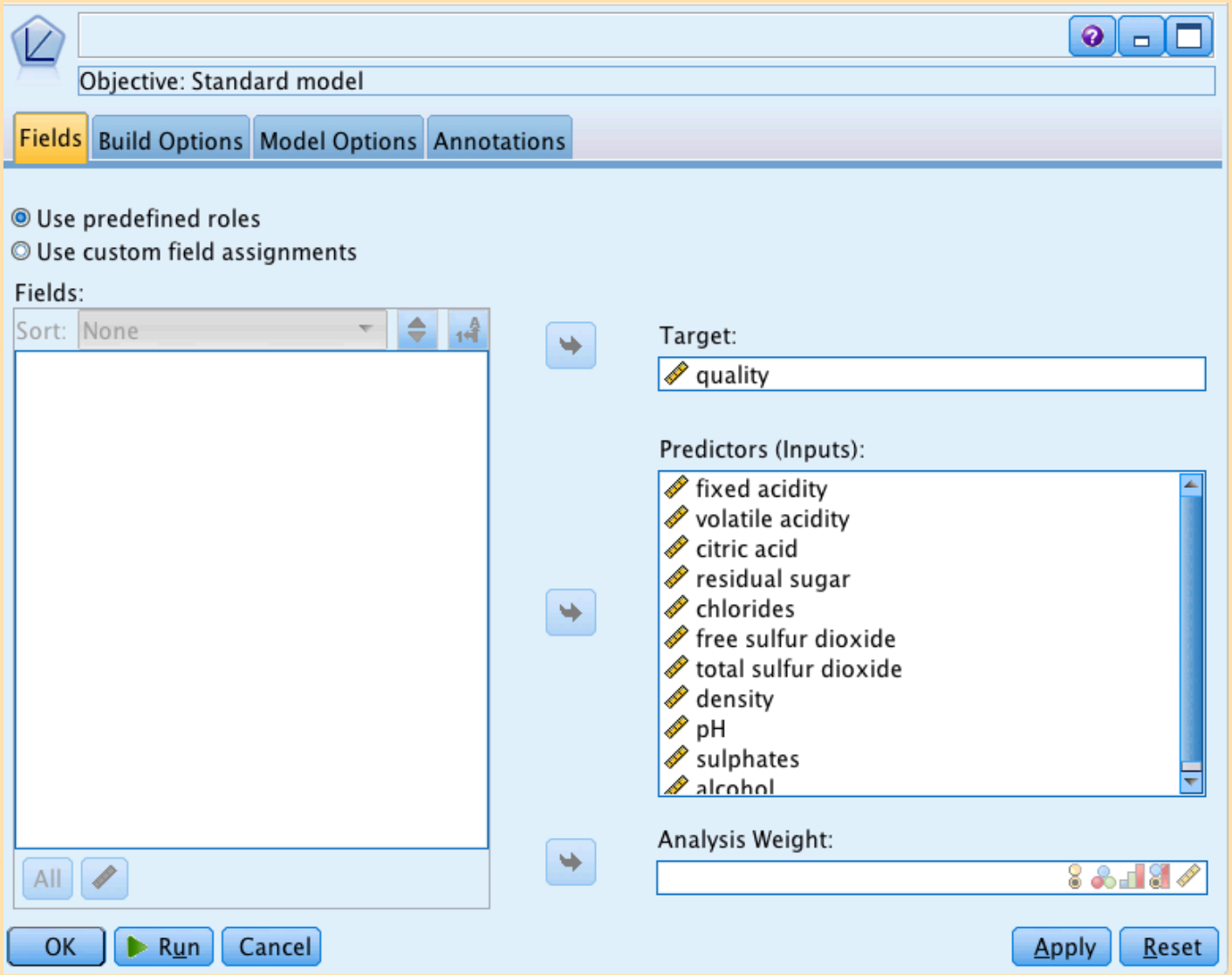
# Output of Table node

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.400 | 0.700 | 0.000 | 1.900 | 0.076 | 11 | 34 | 0.998 | 3.5... | 0.560 | 9.400 | 5 |
| 2 | 7.800 | 0.880 | 0.000 | 2.600 | 0.098 | 25 | 67 | 0.997 | 3.2... | 0.680 | 9.800 | 5 |
| 3 | 7.800 | 0.760 | 0.040 | 2.300 | 0.092 | 15 | 54 | 0.997 | 3.2... | 0.650 | 9.800 | 5 |
| 4 | 11.200 | 0.280 | 0.560 | 1.900 | 0.075 | 17 | 60 | 0.998 | 3.1... | 0.580 | 9.800 | 6 |
| 5 | 7.400 | 0.700 | 0.000 | 1.900 | 0.076 | 11 | 34 | 0.998 | 3.5... | 0.560 | 9.400 | 5 |
| 6 | 7.400 | 0.660 | 0.000 | 1.800 | 0.075 | 13 | 40 | 0.998 | 3.5... | 0.560 | 9.400 | 5 |
| 7 | 7.900 | 0.600 | 0.060 | 1.600 | 0.069 | 15 | 59 | 0.996 | 3.3... | 0.460 | 9.400 | 5 |
| 8 | 7.300 | 0.650 | 0.000 | 1.200 | 0.065 | 15 | 21 | 0.995 | 3.3... | 0.470 | 10.000 | 7 |
| 9 | 7.800 | 0.580 | 0.020 | 2.000 | 0.073 | 9 | 18 | 0.997 | 3.3... | 0.570 | 9.500 | 7 |
| 10 | 7.500 | 0.500 | 0.360 | 6.100 | 0.071 | 17 | 102 | 0.998 | 3.3... | 0.800 | 10.500 | 5 |
| 11 | 6.700 | 0.580 | 0.080 | 1.800 | 0.097 | 15 | 65 | 0.996 | 3.2... | 0.540 | 9.200 | 5 |
| 12 | 7.500 | 0.500 | 0.360 | 6.100 | 0.071 | 17 | 102 | 0.998 | 3.3... | 0.800 | 10.500 | 5 |
| 13 | 5.600 | 0.615 | 0.000 | 1.600 | 0.089 | 16 | 59 | 0.994 | 3.5... | 0.520 | 9.900 | 5 |
| 14 | 7.800 | 0.610 | 0.290 | 1.600 | 0.114 | 9 | 29 | 0.997 | 3.2... | 1.560 | 9.100 | 5 |
| 15 | 8.900 | 0.620 | 0.180 | 3.800 | 0.176 | 52 | 145 | 0.999 | 3.1... | 0.880 | 9.200 | 5 |
| 16 | 8.900 | 0.620 | 0.190 | 3.900 | 0.170 | 51 | 148 | 0.999 | 3.1... | 0.930 | 9.200 | 5 |
| 17 | 8.500 | 0.280 | 0.560 | 1.800 | 0.092 | 35 | 103 | 0.997 | 3.3... | 0.750 | 10.500 | 7 |
| 18 | 8.100 | 0.560 | 0.280 | 1.700 | 0.368 | 16 | 56 | 0.997 | 3.1... | 1.280 | 9.300 | 5 |
| 19 | 7.400 | 0.590 | 0.080 | 4.400 | 0.086 | 6 | 29 | 0.997 | 3.3... | 0.500 | 9.000 | 4 |
| 20 | 7.900 | 0.320 | 0.510 | 1.800 | 0.341 | 17 | 56 | 0.997 | 3.0... | 1.080 | 9.200 | 6 |
| 21 | 8.900 | 0.220 | 0.480 | 1.800 | 0.077 | 29 | 60 | 0.997 | 3.3 | 0.530 | 9.400 | 6 |

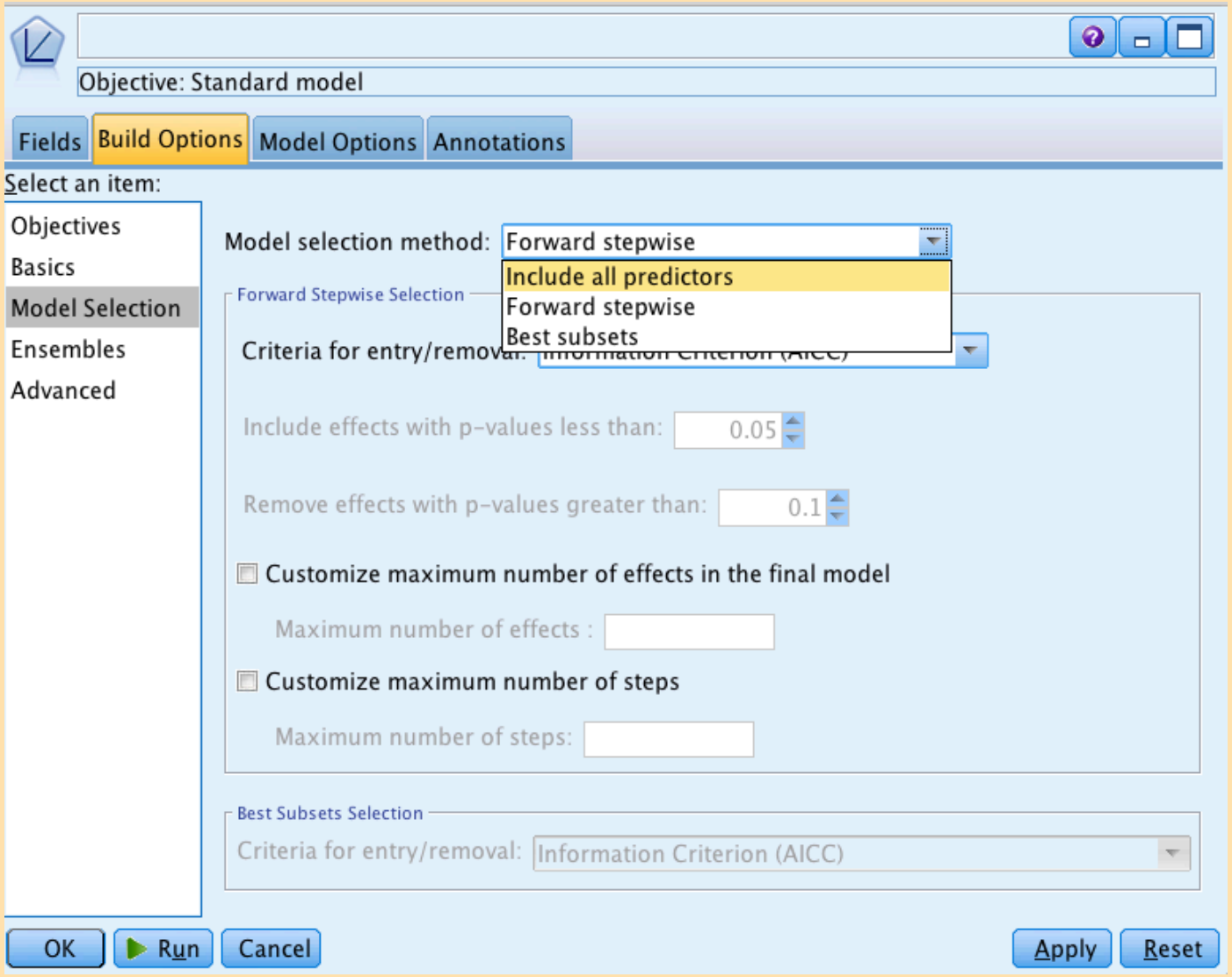
. . .

. . .

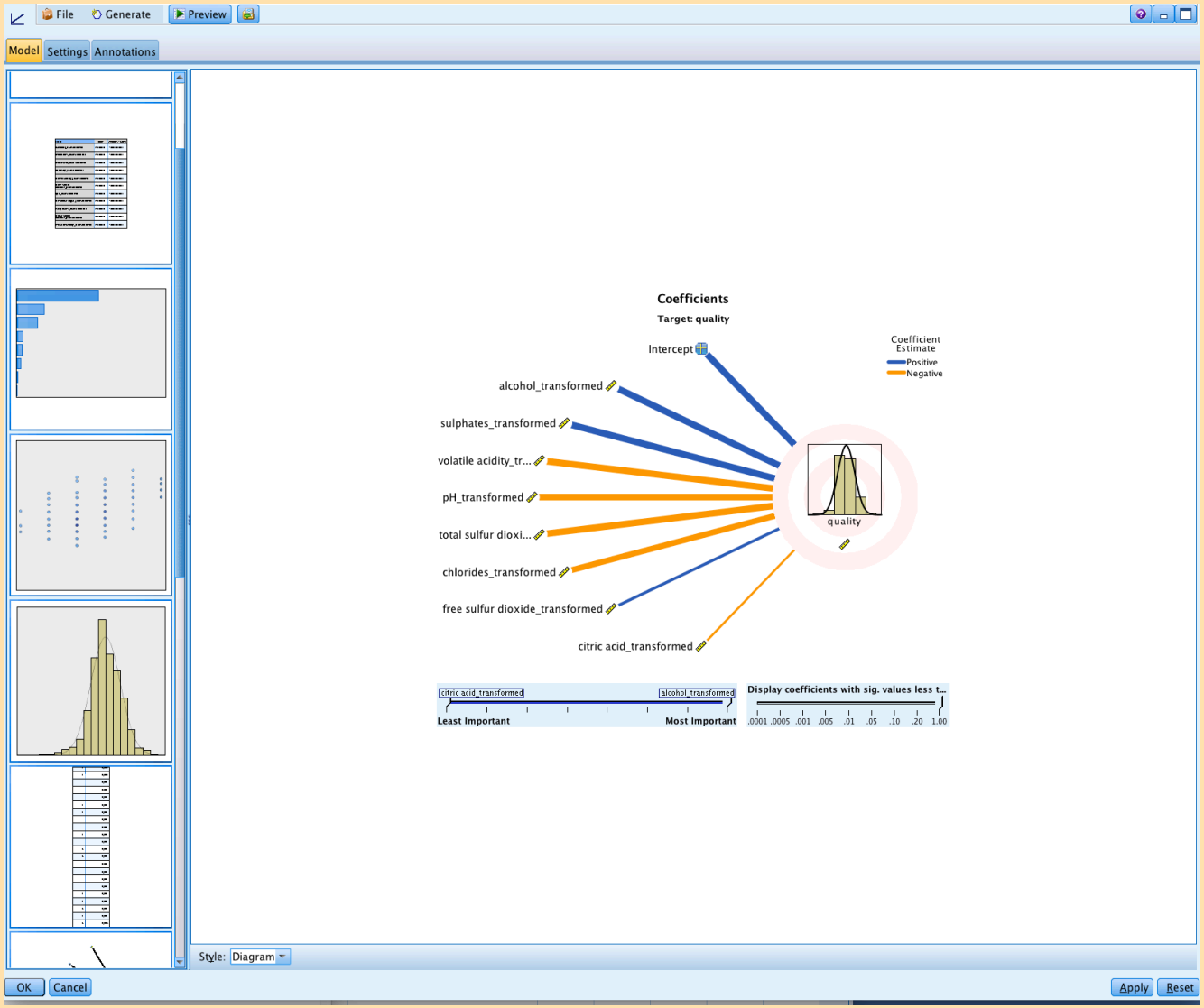
# Linear node


Objective: Standard model
Fields
Build Options
Model Options
Annotations
Use predefined roles
Use custom field assignments
Fields:
Sort:
None
All
Target:
quality
Predictors (Inputs):
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
Analysis Weight:
OK
Run
Cancel
Apply
Reset

# Linear node


Objective: Standard model
Fields
Build Options
Model Options
Annotations
Select an item:
Objectives
Basics
Model Selection
Ensembles
Advanced
Model selection method:
Forward stepwise
Include all predictors
Forward stepwise
Best subsets
Forward Stepwise Selection
Criteria for entry/removal:
Include effects with p-values less than:
0.05
Remove effects with p-values greater than:
0.1
Customize maximum number of effects in the final model
Maximum number of effects :
Customize maximum number of steps
Maximum number of steps:
Best Subsets Selection
Criteria for entry/removal:
Information Criterion (AICC)
OK
Run
Cancel
Apply
Reset

# Golden Nugget

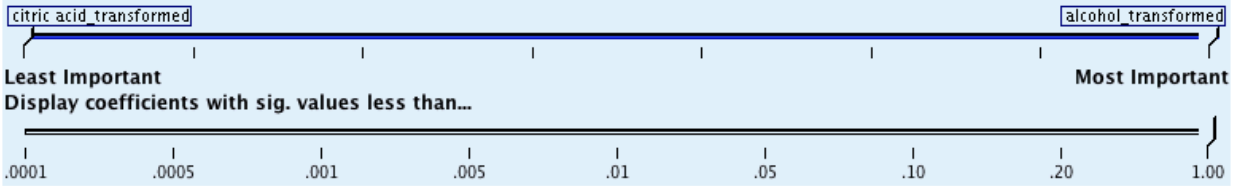


# Coefficients

**Coefficients**

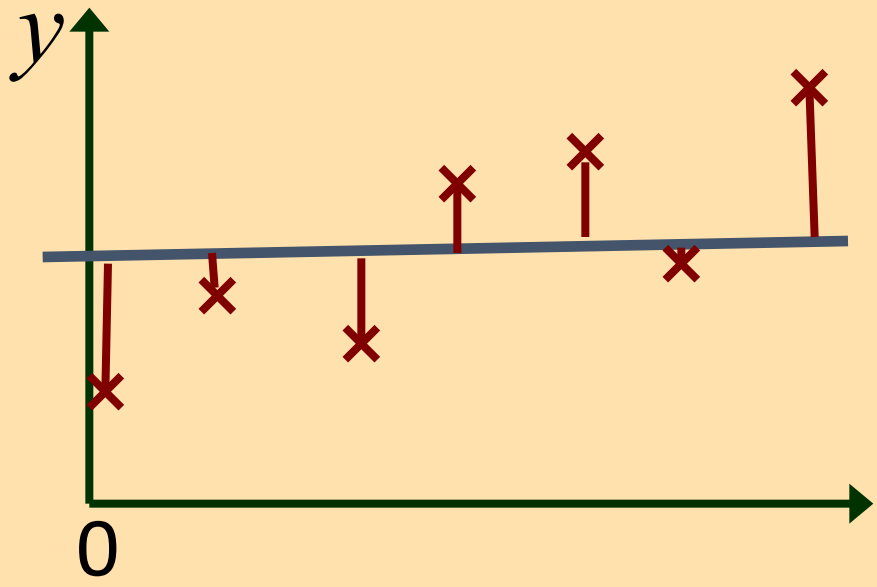
**Target: quality**

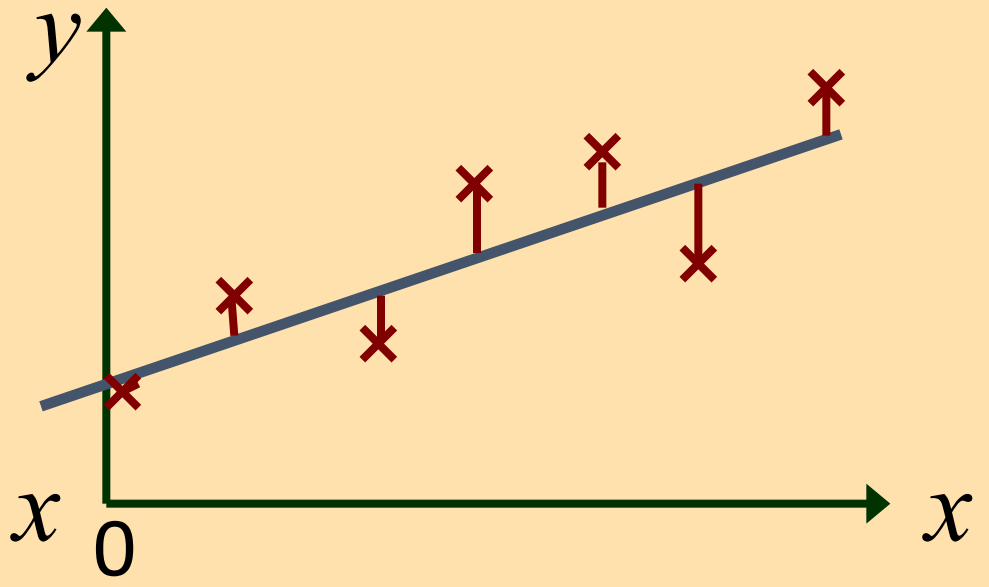
| Model Term | Coefficient ▼ | Std.Error | t | Sig. | 95% Confidence Interval | | Importance |
|---|---|---|---|---|---|---|---|
| | | | | | Lower | Upper | |
| Intercept | 4.818 | 0.483 | 9.974 | .000 | 3.871 | 5.766 | |
| alcohol_transformed | 0.294 | 0.018 | 16.695 | .000 | 0.260 | 0.329 | 0.549 |
| sulphates_transformed | 1.255 | 0.130 | 9.680 | .000 | 1.000 | 1.509 | 0.184 |
| volatile acidity_transformed | -1.029 | 0.121 | -8.474 | .000 | -1.267 | -0.791 | 0.141 |
| pH_transformed | -0.634 | 0.136 | -4.679 | .000 | -0.900 | -0.368 | 0.043 |
| total sulfur dioxide_transformed | -0.003 | 0.001 | -4.461 | .000 | -0.005 | -0.002 | 0.039 |
| chlorides_transformed | -3.383 | 0.890 | -3.803 | .000 | -5.128 | -1.638 | 0.028 |
| free sulfur dioxide_transformed | 0.005 | 0.002 | 1.990 | .047 | 0.000 | 0.009 | 0.008 |
| citric acid_transformed | -0.228 | 0.123 | -1.852 | .064 | -0.469 | 0.013 | 0.007 |




Style: Table

# $R^2$




Null Model $R^2 = 0$

RegressionModel $R^2 = 0.4$
40% of variance is explained by regression line

# Estimating best model without overfitting

- How to choose between different models/algorithms?
- Idea is to use “prediction quality” in order to assess what level of complexity is best given a particular dataset
- But we can’t use the same data we fit the model to in order to assess how well it did – could be perfect fit but not generalize to other data – each model is subject to overfitting – e.g., in genetics project we have > 6500 candidate genes but only 72 observations!
- So we deliberately leave data aside for “testing”…

# Training/Test

- Randomly split data into training and test sets (often ~2/3 for train, 1/3 for test)
- Build a regression model using the *training* set and evaluate it using the *test* set
- Best if you have a lot of instances/observations (at least 100, preferably > 1000)
- Alternative when data set is not so big is to use *cross-validation* – discuss Thursday
- Need a metric by which to evaluate…

# Evaluation

- Possible evaluation measures in Modeler:
  - Correlation -- correlation coefficient between the observed target values and model estimated target values
  - Relative Error  $(1- R^2)$
  - Number of fields – number of variables included in the model – “smaller will be simpler model with less chance of overfitting”

# IBM Demo – Auto Numeric Node

# Friendly competition – prize $5 Starbuck's cards

- Figure out how to
  - Increase label size on plots
  - Change line thickness on plots
  - Overlay more than one line on plots

# Thursday's Lecture

- The classification problem
- Logistic regression
- Classification trees
- Support Vector Machines