General Linear Model I: Regression

Philippe Rast

PSC 204B:
General Linear Model
January, 09

UC Davis, Winter 2018

PSC 204 B: Overview

- Class on Tuesday & Thursday
- Labs on Friday
  - Kristine O'Laughlin & Jared Stokes
  - R statistical software (get R and a decent editor)
- Course notes on Canvas
- Homework: Assigned Thursday, due following Thursday
  - Revisited in Friday's lab
- Final Exam
  - Take home, work alone
  - Given out on the last day of class
  - Due one week later
- Grade: 50% homework, 50% final exam
- Reading
  - McElrath (2014) Statistical Rethinking
  - Gelman & Hill (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models

! Discuss Friday Lab composition

Topics

Single-level Regression:

Week 1 Linear Regression (G&H: 3,4)
Week 2 Multiple Regression
Week 3 Violation of Assumptions
Week 4 Logistic Regression and GLM (G&H: 5, 6)
Week 5 Model comparison, Over-fitting, Information Criteria (McE: 6)
Week 6 Regression inference via simulations (G&H: 7–10)

Multilevel Regression:

Week 7 Multilevel Linear Models (G&H: 11–13)
Week 8 Multilevel Generalized Models (G&H: 14, 15)
Week 9 Bayesian Inference (G&H: 18 / McE: 1, 2, 3)
Week 10 Fitting Models in Stan and brms (G&H: 16, 17 / McE: 11)

Aims of The Course

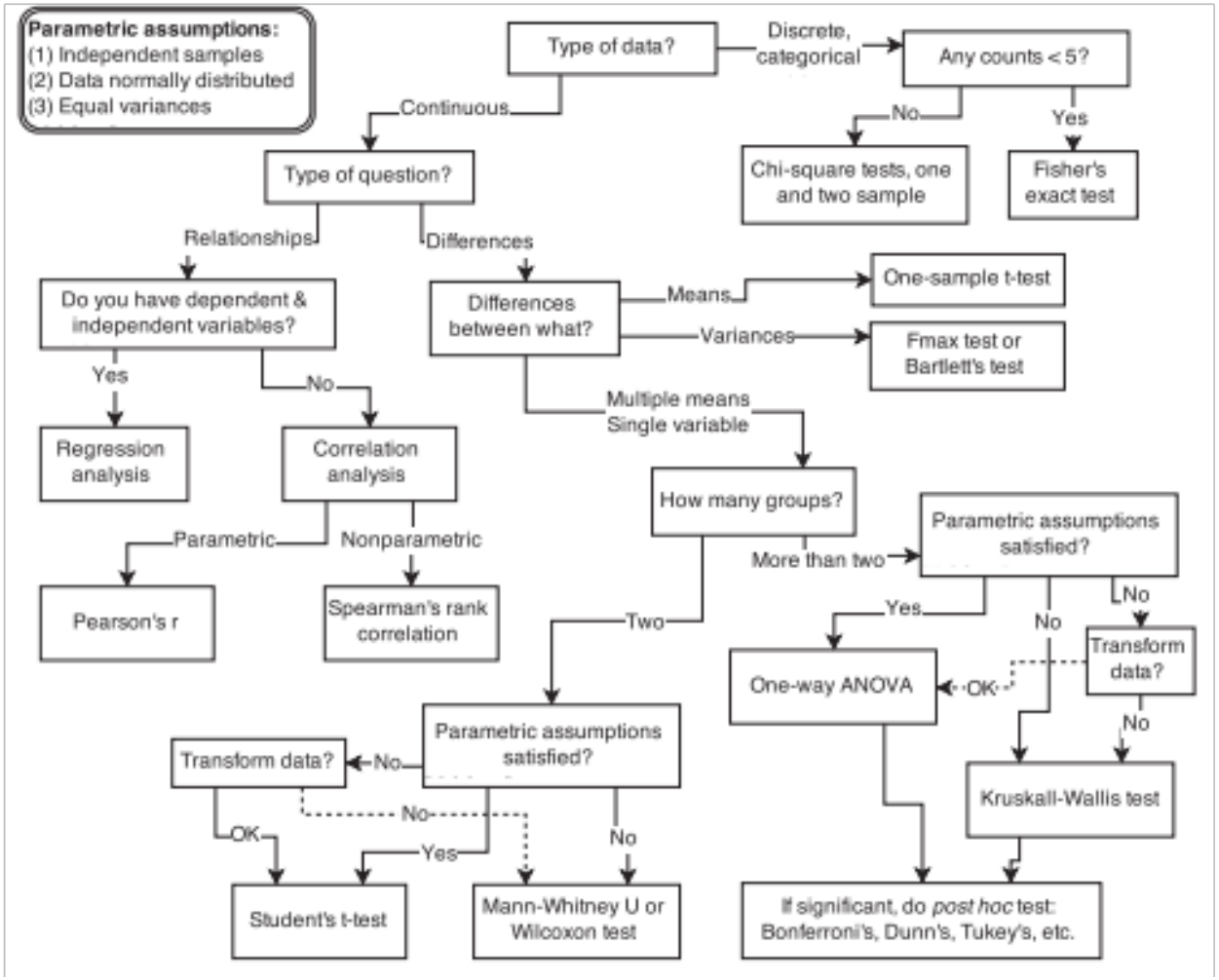
- Know the basic analysis methods, conditions, and recognize violations of assumptions of the general linear model and be able to partly reproduce and comment these methods
- Run own empirical analyses and be able to adequately interpret results in the realm of the general linear model.
- Be able to understand and critically evaluate the use of general linear model methods in the literature and empirical work.
- The GLM is the basis for more complex models – provide a ground to access other modeling techniques

Some Thoughts about Statistical Models

- Models are meant to simplify big world problems
- Recreate a "small world"
- A t-test or correlation is a model
- A model is neither "true" nor "false"
- A model is a tool, a robot (Jayne, 2013), a Golem (Collins & Pinch 1998; McElrath, 2014)

- There are a number of models that are commonly used for many different research questions
- Whole set of pre-fabricated models

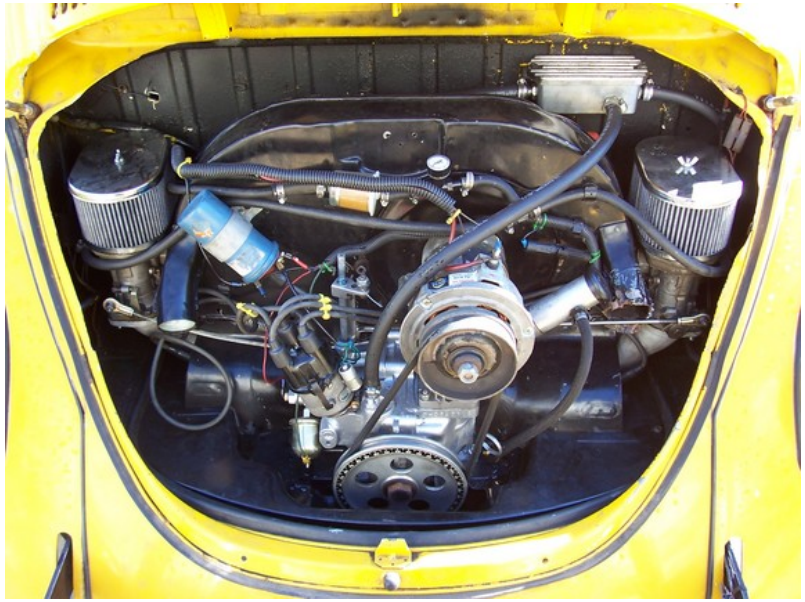
Decision Tree



What is the underlying structure, what are the assumptions?

Decision Tree

- Decision trees and flowcharts hide the machine within
- Statistical tools are specialized tools
- Classical tools are not diverse enough to handle many common research questions.



- We need a set of principles for *designing*, *building*, and *refining* special-purpose statistical procedures.

Models And Hypotheses

- Our starting point: Hypothesis
- Model represents hypothesis
- Inference from model is used to test hypothesis

Problem:

- Hypotheses are not models - and models are not hypotheses
- All models are false, so what does it mean to falsify a model?
- Consequence of the requirement to work with models
- ▷ Not possible to deduce that a hypothesis is false, just because we reject a model derived from it.

Models And Hypotheses

Attempting to mimic falsification is not a generally useful approach to statistical methods

- What are we to do?
- We are to model!
- Models can be made into testing procedures
- ▷ All statistical tests are also models–

  but they can also be used to measure, forecast, and argue
- Models will not answer what is true or false

- Doing research benefits from the ability to produce and manipulate statistical models
  - Scientific problems are more general than "testing"
  - The pre-fabricated models are probably ill-fit for many research contexts
  - Approach research question with different tools/models

Approach

- Learn about models
- General(-ized) linear model
- Different ways of making inferences
- Bayesian and frequentist data analysis

Modeling: Information Compression

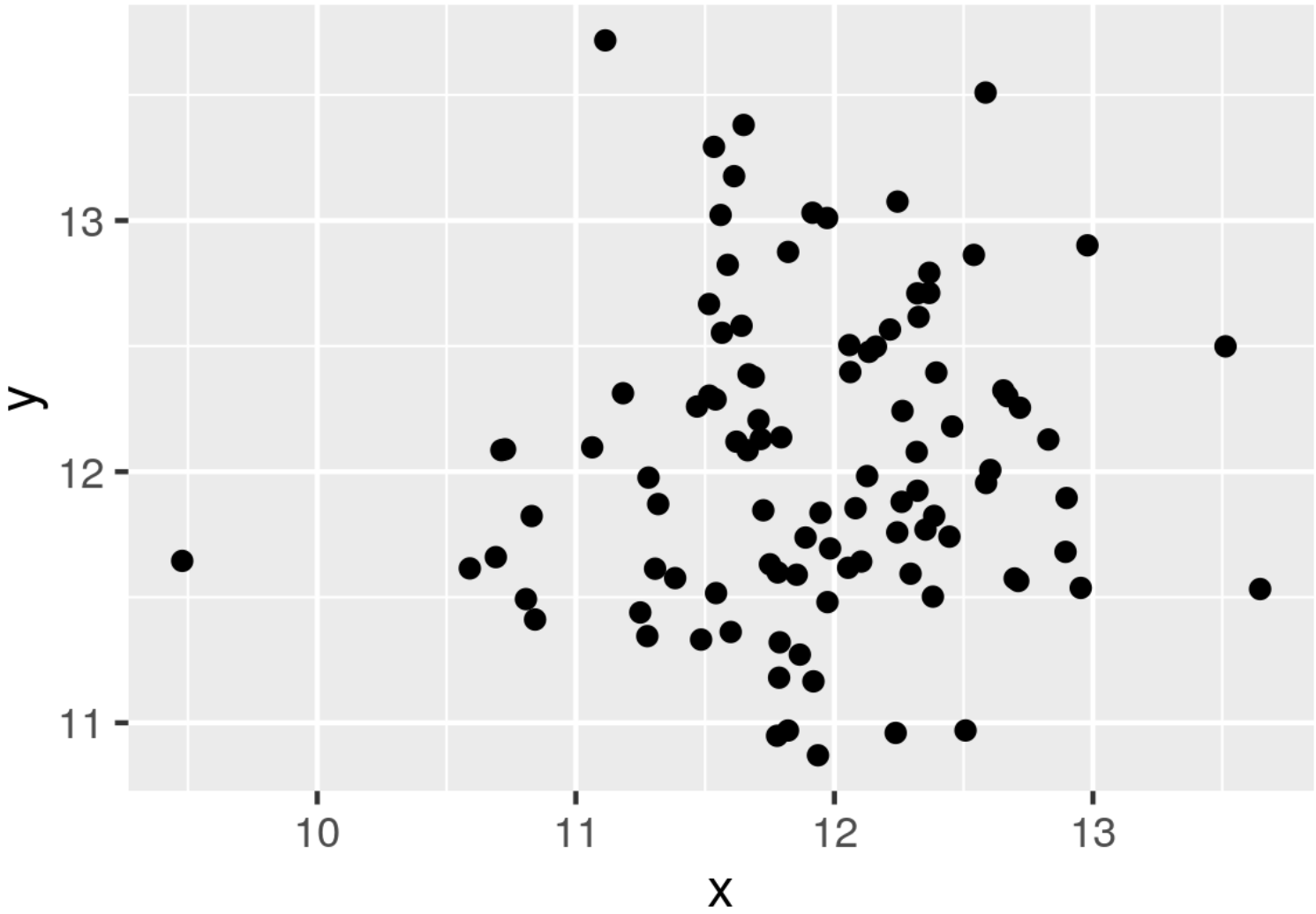
- Building a model also entails reducing information
- Simple things are easier to understand
- How much reduction?
  - ▷ Overfitting vs. underfitting

- One of the simplest and most frequent models: Correlation
  - ▷ Reduce relationship among potentially huge amount of information to one value.

Let's think a about this:

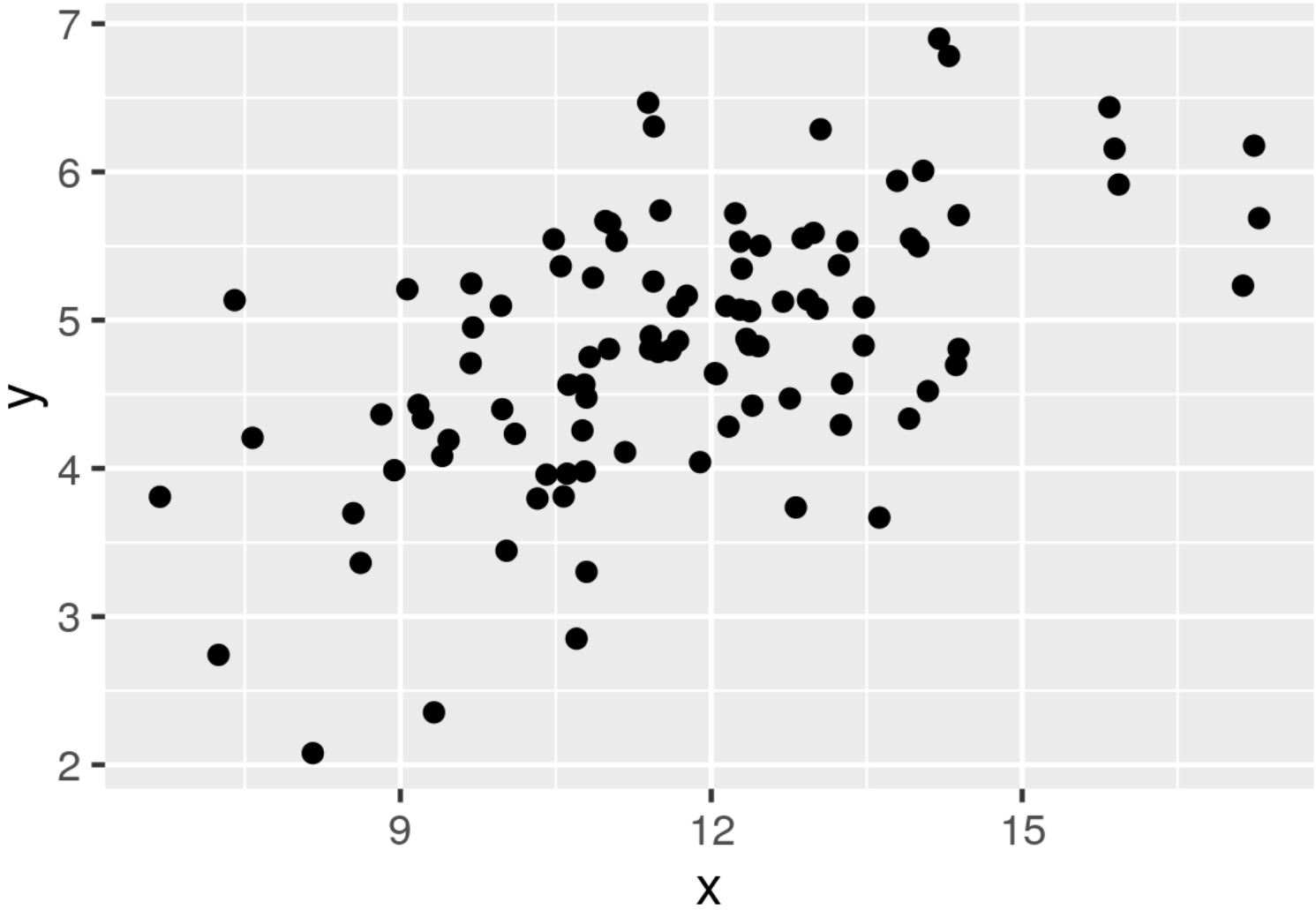
Graphic Representations of Relationships

No correlation



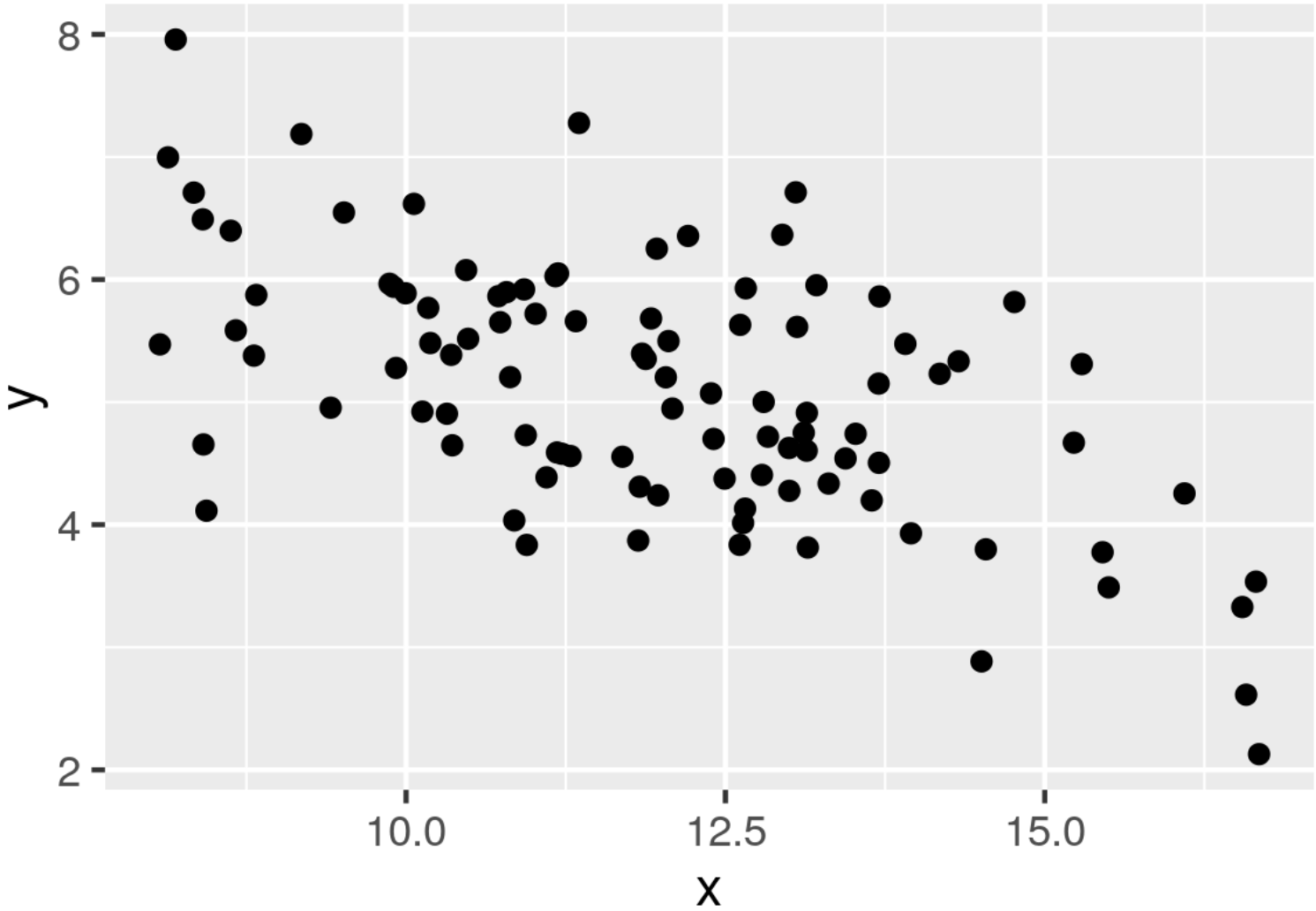
Graphic Representations of Relationships

Correlation $r = .6$



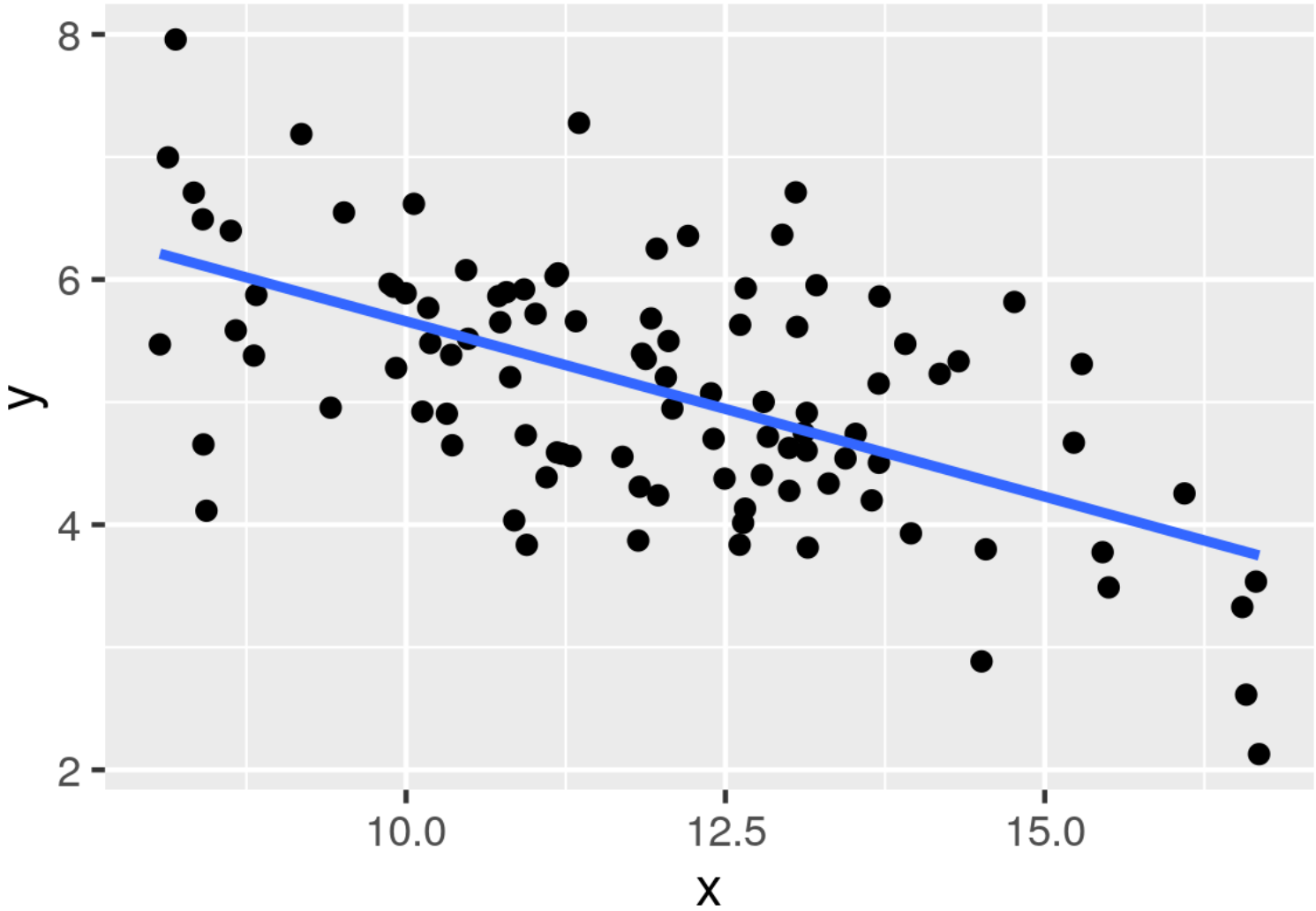
Graphic Representations of Relationships

Correlation $r = -.6$



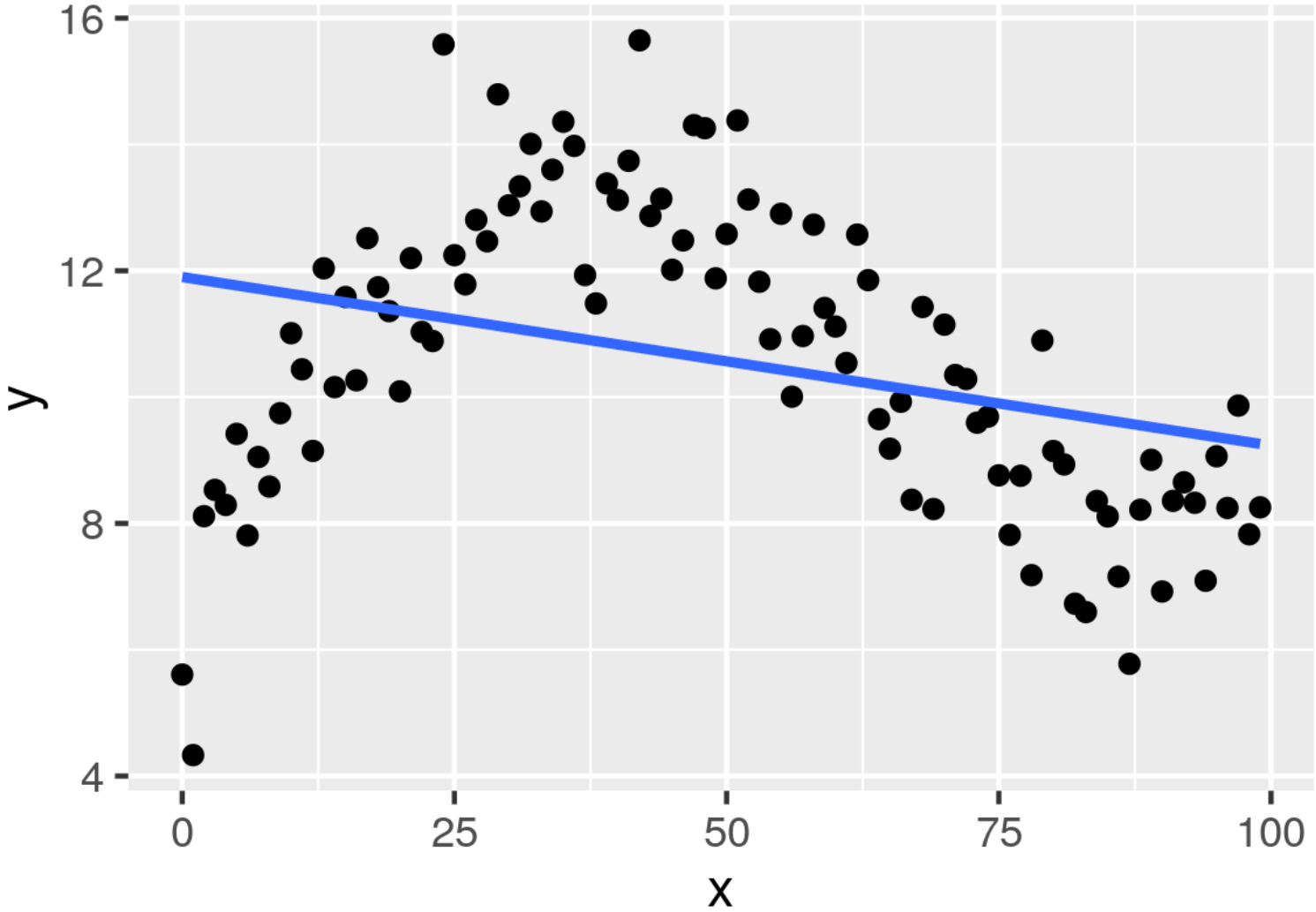
Graphic Representations of Relationships

Correlation $r = -.6$



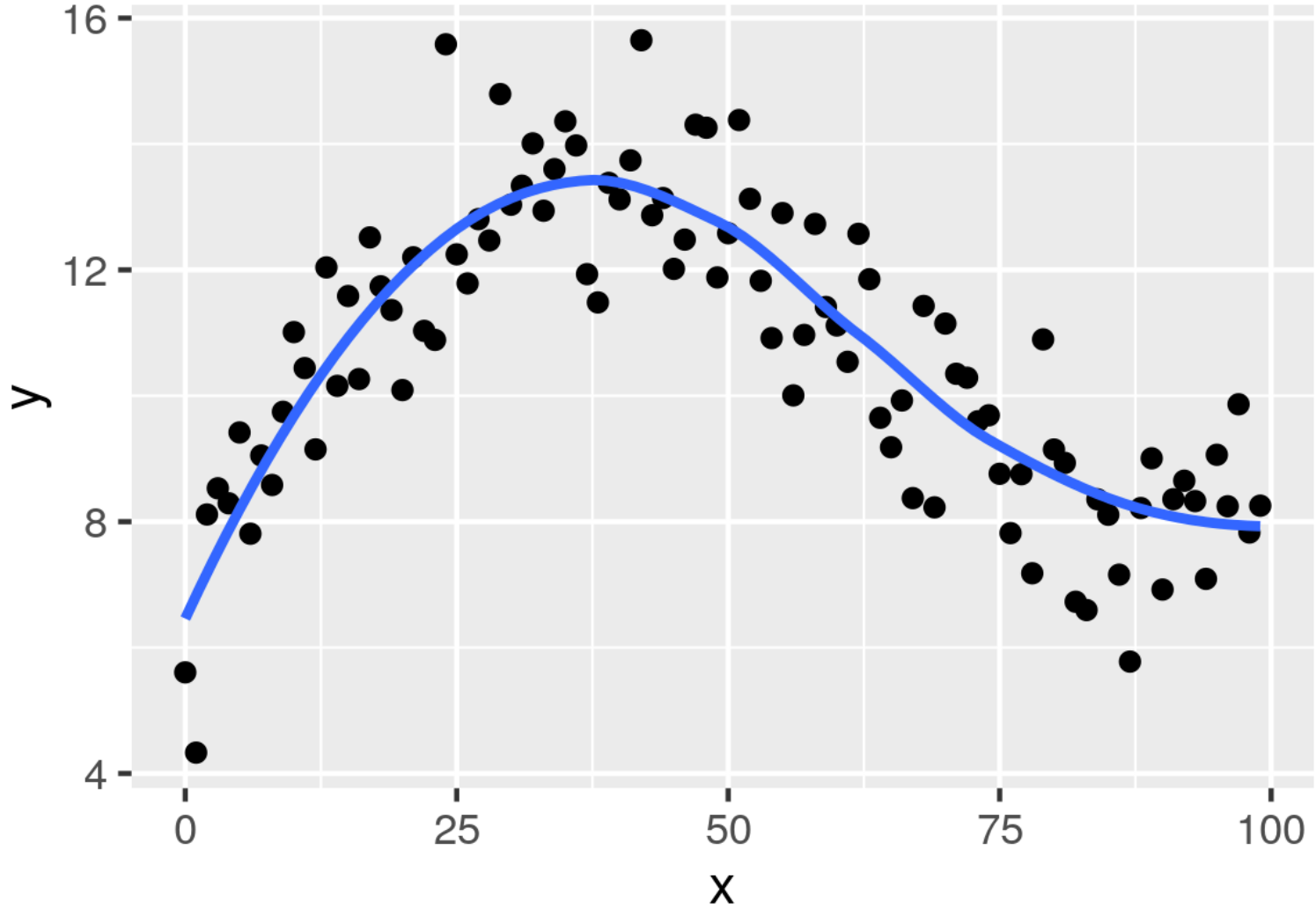
Graphic Representations of Relationships

Correlation $r = -.4$



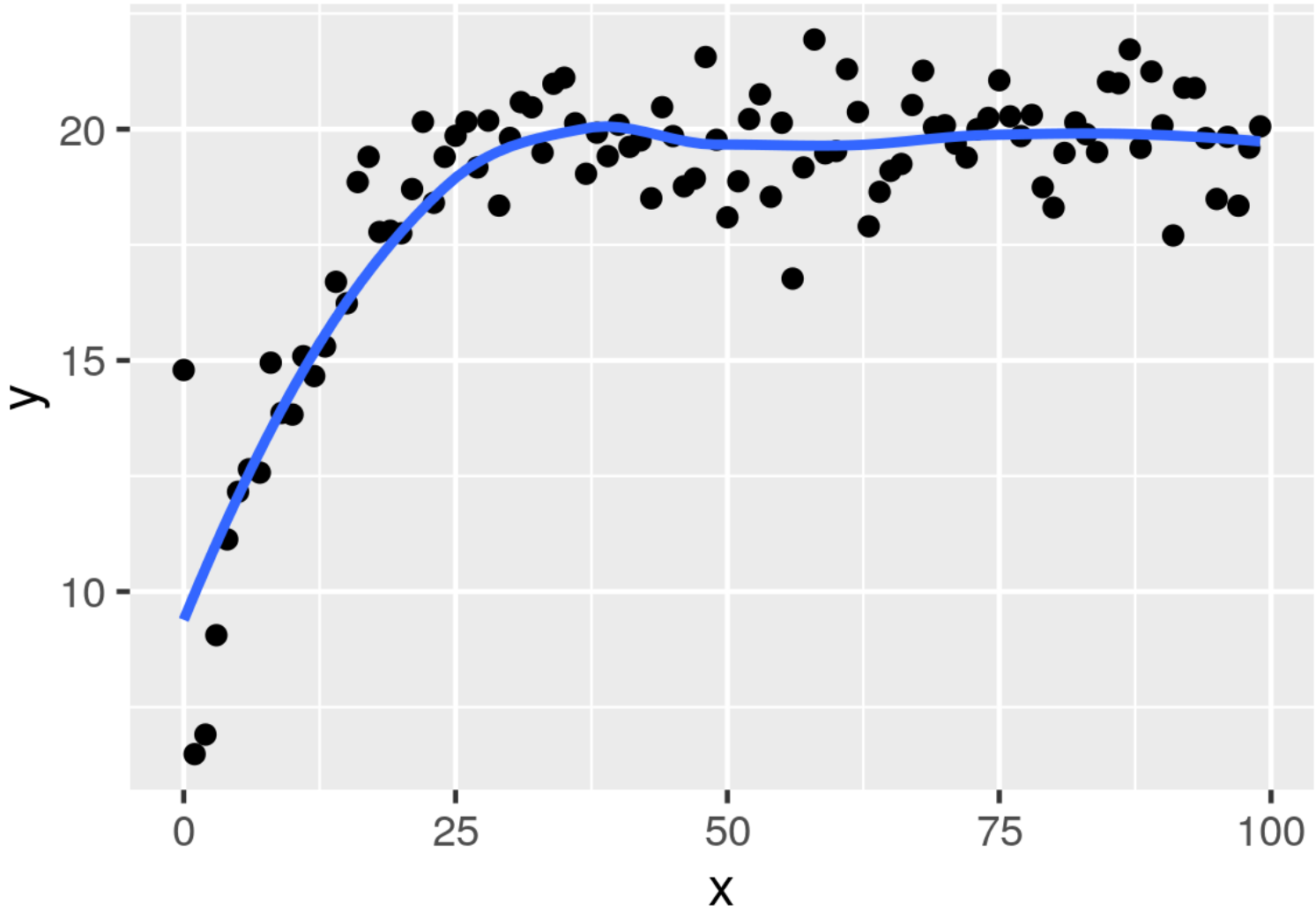
Graphic Representations of Relationships

Cubic model ($y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^2 + \epsilon$)

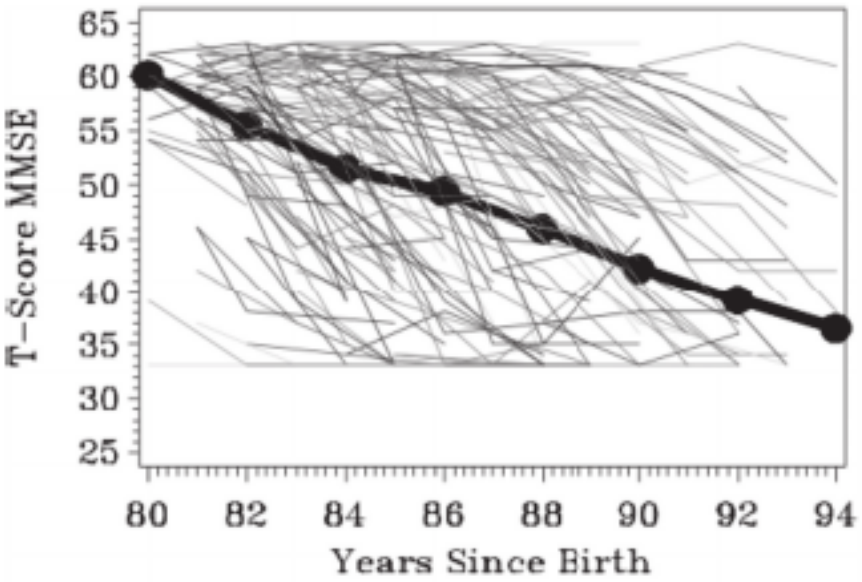


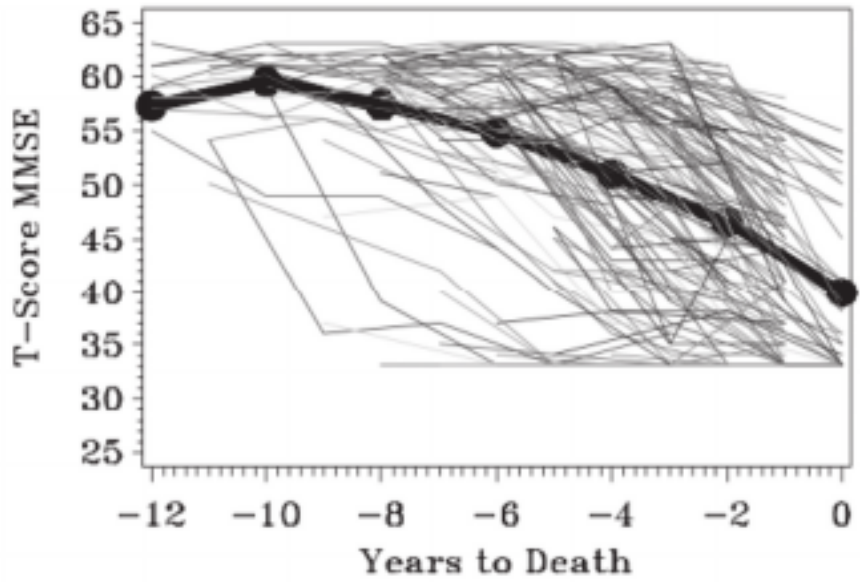
Graphic Representations of Relationships

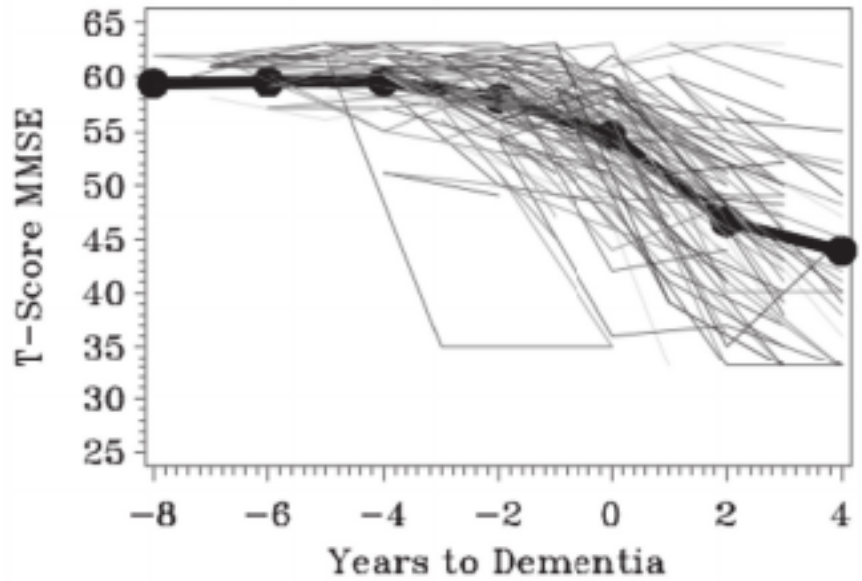
Correlation $r = .6$; Exponential model $(y = \theta_3 - (\theta_3 - \theta_0)e^{(-x\theta_2)})$

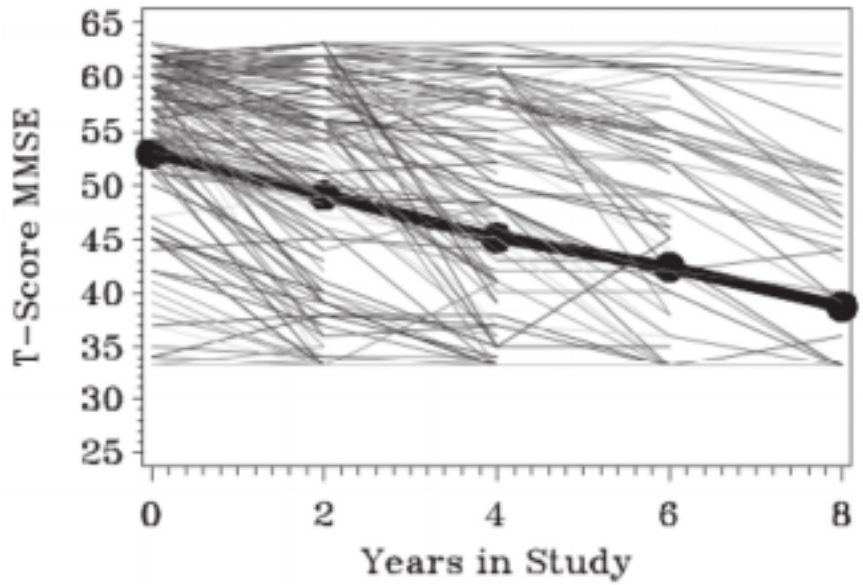


Same Data – Different Figure (*Fig 10, Hoffman 2012*)









Covariance and Correlation

- Linear relation between two variables
  - Covariance: Unstandardized measure of the relationship between two variables
  - Sample covariance:

$$\mathrm{C}(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

  - Correlation: "Standardized" covariance with variances fixed at 1.
  - Pearson's correlation for a sample:

$$r_{xy} = \frac{\mathrm{C}(x, y)}{\sqrt{s_x^2 s_y^2}}$$

  - with sample variance $s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$ and $s_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2$

Assumptions

The correlations is just a coefficient – can always be obtained
The test for significance of Pearson's r among X and Y:

- Related pairs
- Each variable should be continuous (Spearman r for ordinal)
- No outliers
- Linear relationship
- Homogeneous variance
- Normal distribution for all variables in the population sampled.

Simpson's Paradox

Sometimes it all looks good but...

R code: Simpson.R

Concepts and Methods from Basic Probability and Statistics

Probability distributions

Simple probability distributions are the building blocks for elaborate models.

Probability distributions

- Distributions of data (e.g., heights of adults).
  Notation $y_i, i = 1, ..., n$
- Distributions of parameter values.
  Notation $\theta_j, j = 1, ..., J$, or other Greek letters such as α, β, γ.
- Distributions of error terms, which we write as $\epsilon_i, i = 1, ..., n$.

Concepts and Methods from Basic Probability and Statistics

Probability distributions

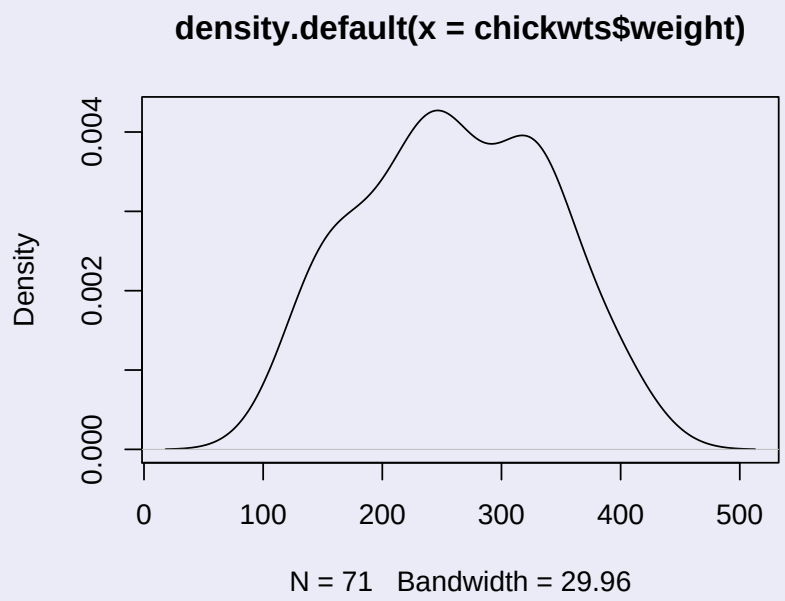
- The basic way that distributions are used in statistical modeling
  - Start by fitting a distribution to data y
  - Then, get predictors X and model y given X with errors ϵ.
- Further information in X can change the distribution of the ϵ's

Distribution of Chick Weights

Data included in R

R code: Chicks.R

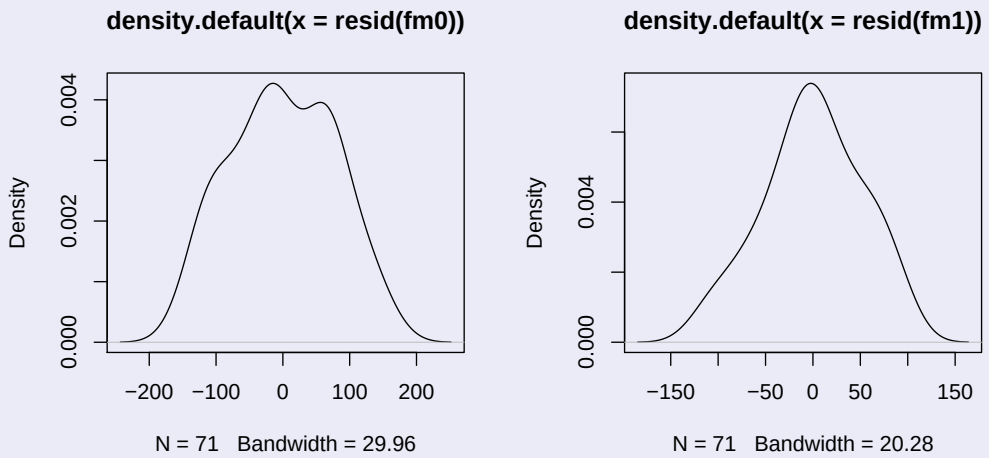
```
## Look at distributions of chick weights on different diets
plot(density(chickwts$weight))
```



Distribution of Chick Weights

R code: Chicks.R

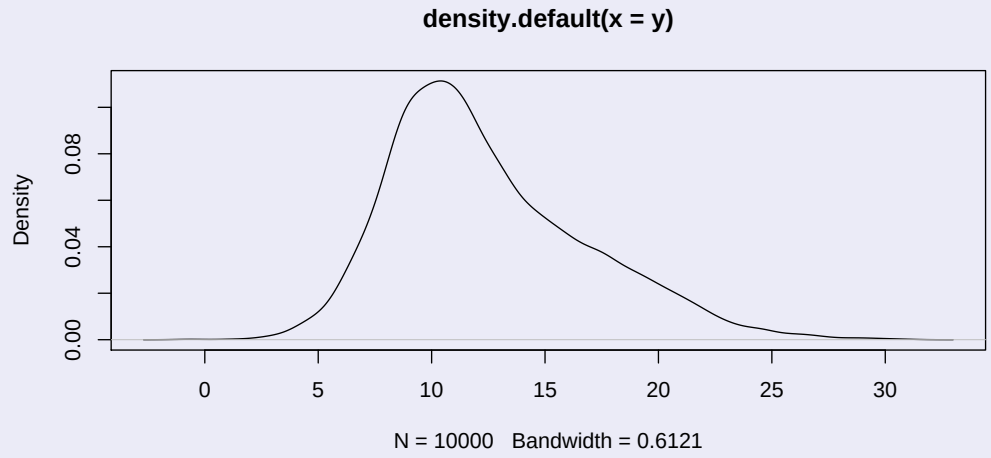
```
op <- par(mfcol = c(1,2))
fm0 <- lm(weight ~ 1, data = chickwts)
plot(density(resid(fm0)))
fm1 <- lm(weight ~ feed, data = chickwts)
plot(density(resid(fm1)))
```



Simulate Data

R code: Chicks.R

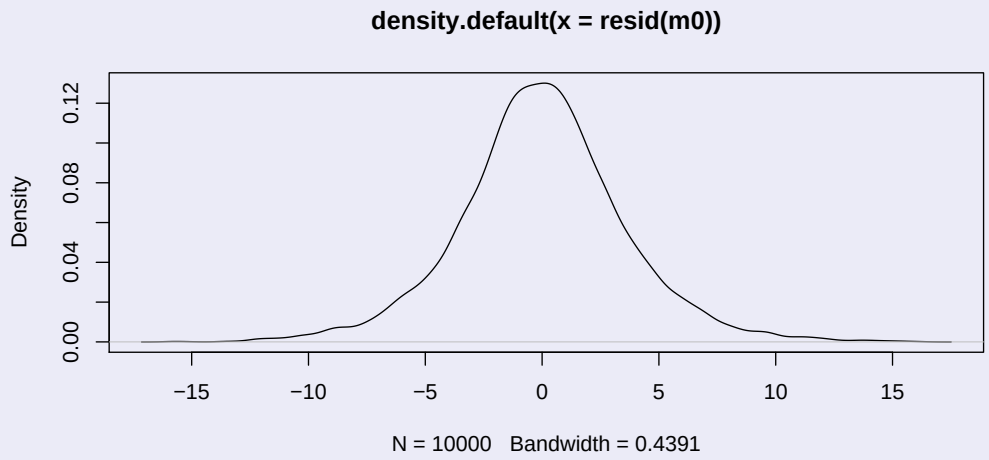
```
## Generate own distributions:
n = 1e4
group = sample(c(0, 1), n, replace = TRUE)
y = rnorm(n, 10, 2) + group*rnorm(n, 5, 4) + rnorm(n, 0, 1)
plot(density(y))
```



Simulate Data

R code: Chicks.R

```
m0 <- lm(y ~ group)
plot(density(resid(m0)))
```



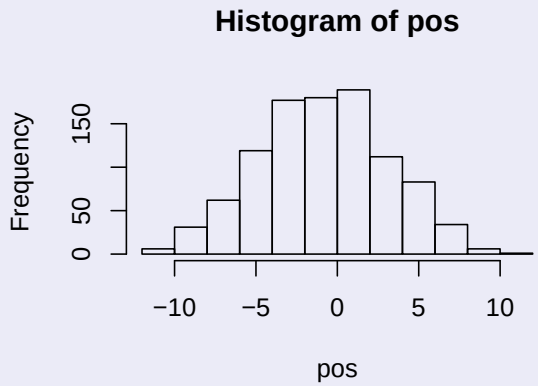
Normal distribution; means and variances

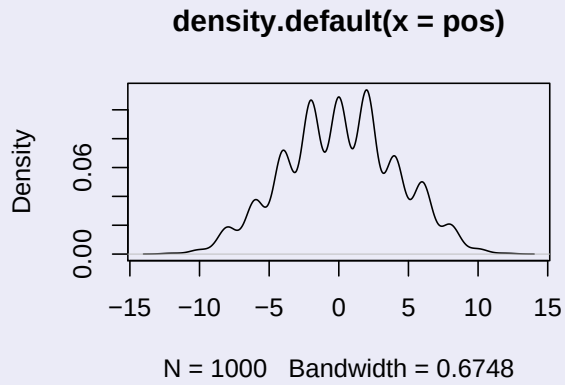
$y \sim \mathcal{N}(\mu, \sigma^2)$

- Central Limit Theorem of probability: Sum of many small independent random variables will be a random variable with an approximate normal distribution.

R code: Normal.R

```
pos <- replicate(1000, sum(sample(c(-1,1), size = 16,
replace = TRUE))) op <- par(mfcol = c(1,2)) hist(pos)
plot(density(pos))
```



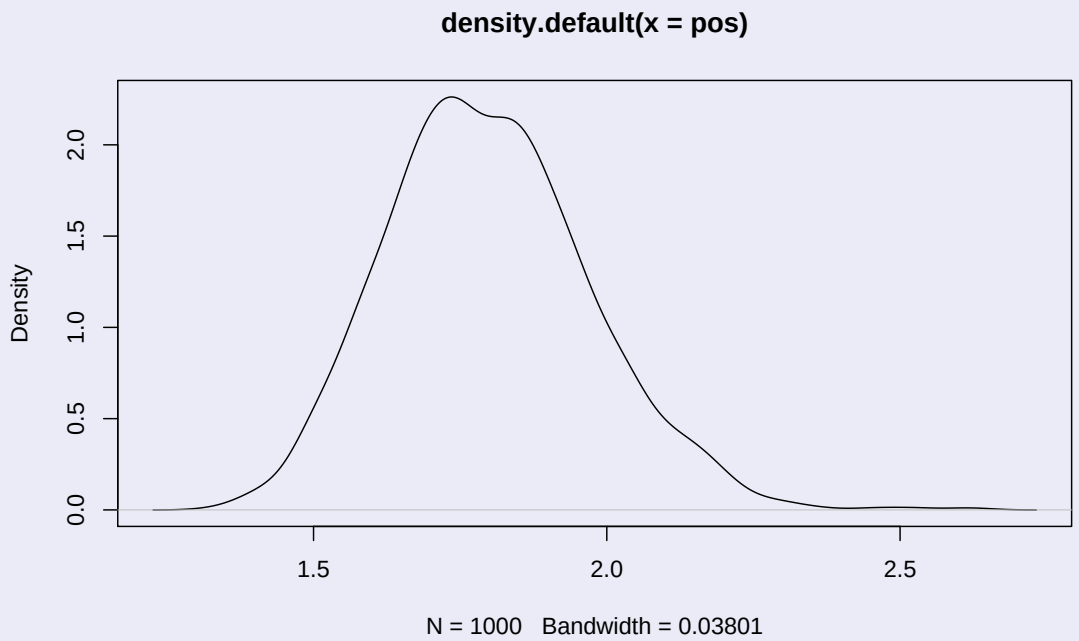


Normal distribution; means and variances

- Normal by multiplication
- Multiplying small numbers is approx. the same as addition

R code: Small increments

```
pos <- replicate(1000, prod( 1 + runif(12, 0, 0.1)))
plot(density(pos))
```

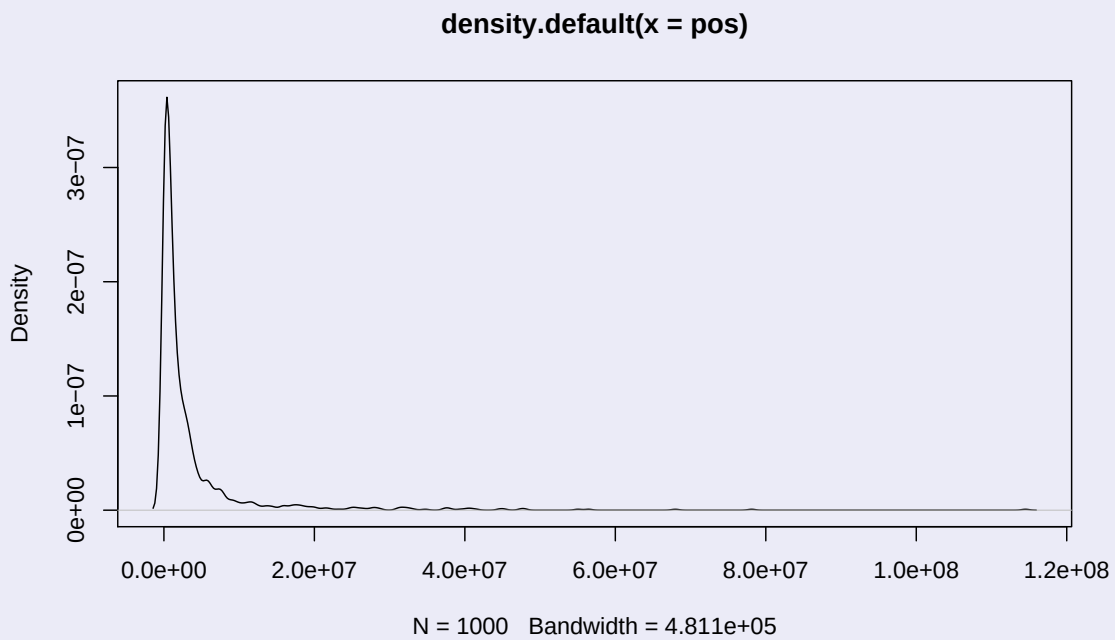


Normal distribution; means and variances

- Does not work with bigger numbers:

R code: Big increments

```
pos <- replicate(1000, prod( 1 + runif(12, 0, 5)))
plot(density(pos))
```

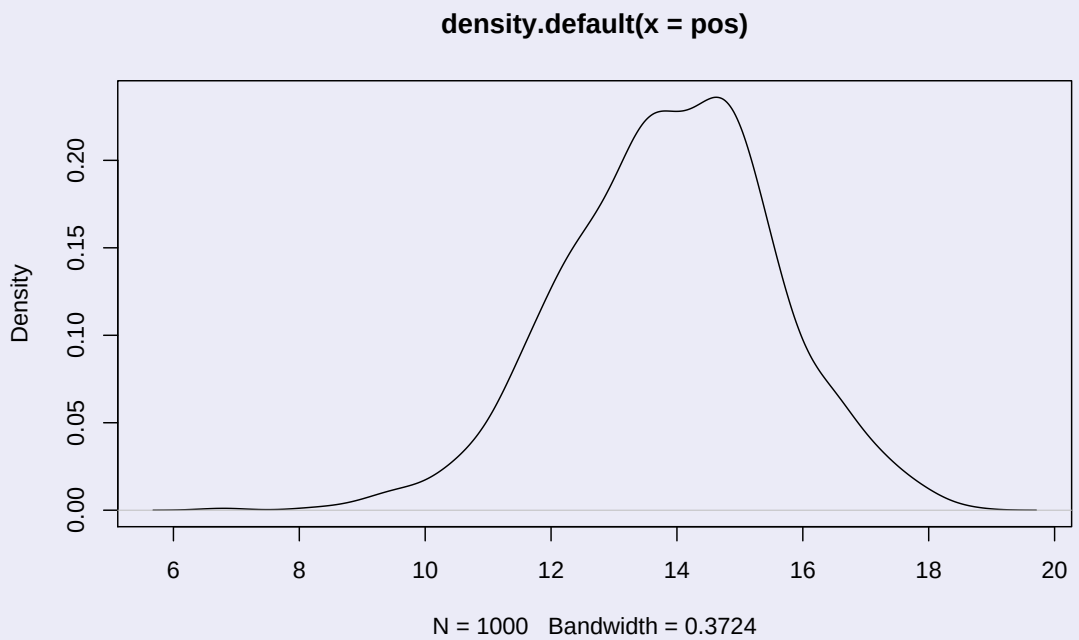


Normal distribution; means and variances

- Normal by log-multiplication

R code: Log increments

```
pos <- replicate(1000, log( prod( 1 + runif(12, 0, 5)) ))
plot(density(pos))
```



Gaussian Distributions

Justification for using Gaussian distribution

Natural processes

- World is full of Gaussian distributions, approximately.
  - ▷ Addition of small fluctuations
  - ▷ Measurement errors, variations in growth etc. tend towards Gaussian distributions

but There are other patterns arising from natural processes: Exponential, gamma, Poisson

- Little assumptions: Only need to assume that a measure has a mean and a finite variance
- Maximum entropy

Gaussian Distributions

Probability density of y given mean μ and standard deviation σ is

$$p(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Dissecting PDF:

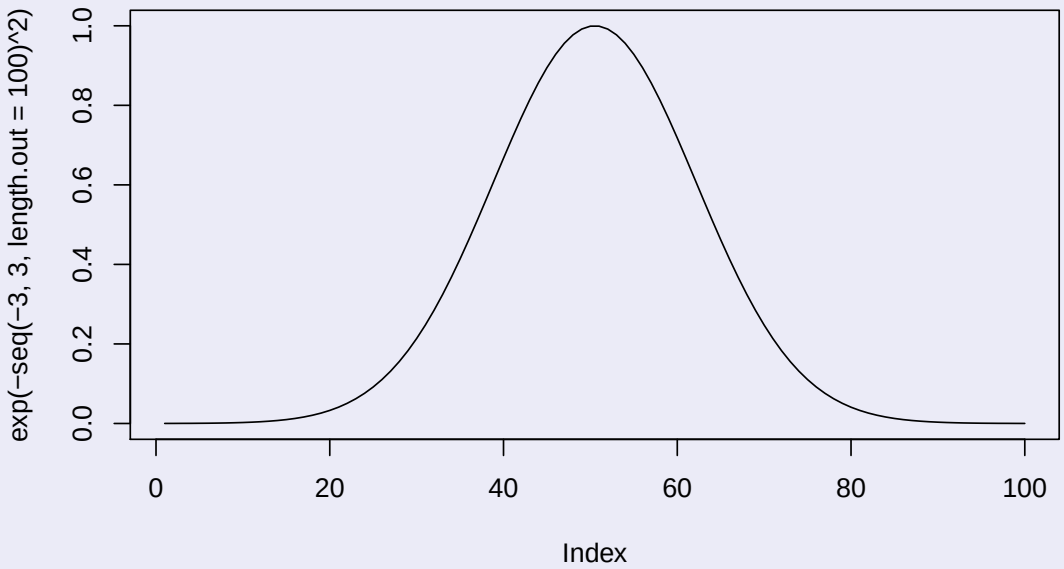
- Important bit: $(y-\mu)^2$
  This gives it the fundamental quadratic shape
- Taking the exponent, turns it into the bell shape
- All other terms are meant scale and standardize the function so that it sums to 1

Gaussian Distributions

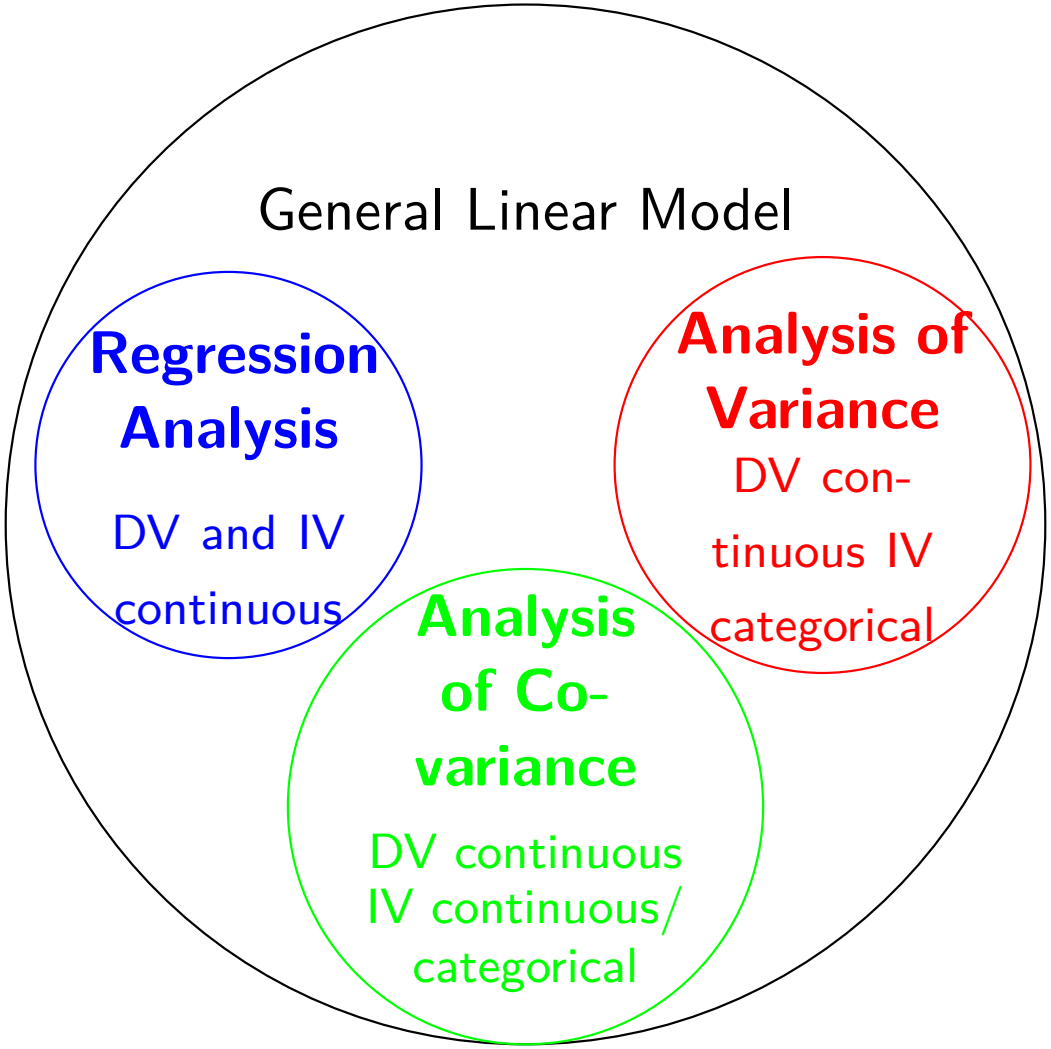
$$p(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Gaussian Bell

```
plot( exp( -seq(-3, 3, length.out = 100)^2), type = 'l')
```



The GLM Overview



GLM: Regression, Analysis of Variance and -Covariance

The GLM Overview

The common theme among all these analysis methods is the assumption that

- relations among variables are *linear*
- The GLM can always be expressed in the form of a linear equation system.
- The GLM is a system of linear equations that contains a number of unknown parameters (β-parameters to be estimated).

The GLM Overview

Typical research questions addressed with the GLM:

- Regression: Can we predict income from GRE scores?
- Analysis of variance: Does memory performance depend on the number of stimulus presentations and visual components of test words.
- Analysis of covariance: Does income depend on job satisfaction and gender?

We always assume linearity among the outcome and the predictors.

Curvi*linear* relations can be addressed by raising the predictor to a higher polynomial.

Equations of the GLM

The GLM can be expressed by a small number of equations, those that we will encounter most often:

(a) Basic equation:

$$\underset{N\times 1}{\mathbf{y}} = \underset{N\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{N\times 1}{\boldsymbol{\epsilon}}$$

where N is the sample size and p is the number of independent predictors.

Equations of the GLM

- This basic equation represents the GLM in matrix notation
  - In the vector $\boldsymbol{y}$ the N observed values are written into one column.
  - In $\mathbf{X}$, the design matrix, the independent variables are written into the $1., 2., ..., p$ columns.
  - $\boldsymbol{\beta}$ contains the p parameters to be estimated in the regression and it defines the strength with which the independent and dependent variable are associated.
  - $\boldsymbol{\epsilon}$ contains the N disturbances or residuals in one column.

Equations of the GLM

(b) Equation for estimating the β parameters

$$\underset{p\times 1}{\hat{\boldsymbol{\beta}}} = \underset{p\times N}{(\mathbf{X}'\mathbf{X})^{-1}}\ \underset{N\times p}{\mathbf{X}'}\ \underset{p\times N}{\mathbf{y}}\ \underset{N\times 1}{}$$

This equation indicates how the basic equation must be solved with regard to the parameters to be estimated so that the variance of the disturbances is minimized.

- $\mathbf{X}'$ is a transposed matrix, i.e., a matrix in which rows and columns have been swapped.
- $\mathbf{X}'\mathbf{X}$ Is the inner product of a matrix. The resulting matrix is smaller than the original matrices.
- $\mathbf{X}'\mathbf{X}^{-1}$ Is an inverse, the matrix analog to a division.

Equations of the GLM

(c) Equation of the variances for the estimated β-parameters

$$\mathrm{V}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}_{\epsilon}^{2}(\mathbf{X}'\mathbf{X}^{-1}).$$

This equation is needed to statistically test the estimates of the β-parameters (significance test)

- $\hat{\sigma}_{\epsilon}^{2}$ Is the estimation of the disturbance or error variance.
- For significance testing, the estimate of a parameter is divided by the root of its variance (standard error).
- This quotient is (under certain assumptions) t-distributed.

On the Significance of the GLM

- Its submodels (regression, [co-] variance analysis) capture a large part of univariate and multivariate statistical evaluation procedures.
- The GLM is a framework model and established analysis method – which is a curse and a blessing...
- The GLM offers approaches that are compatible with conventional statistical methods (e.g., different ways of partitioning sums of squares)
- A variety of hypotheses can be tested using the GLM (which typically exceeds the number of substantive hypotheses).
- The GLM is itself a submodel of the Generalized Linear Model.

Example and Data

- Is perceptual processing speed related to the age of the respondent?
- Such a question will typically be aimed at a population (and not only on a specific sample)
  - We are looking for a method (inferential statistics) which allows us to
  - infer certain relations in a population
  - from relations that we observed in a random sample (as a subset from that population)
- The inferences drawn from the sample to the population will *never* be certain, but will always bear a certain amount of probability of error, which depends on the precision of our estimation.

Example and Data

Table: Data from the Longitudinal Aging Study Amsterdam

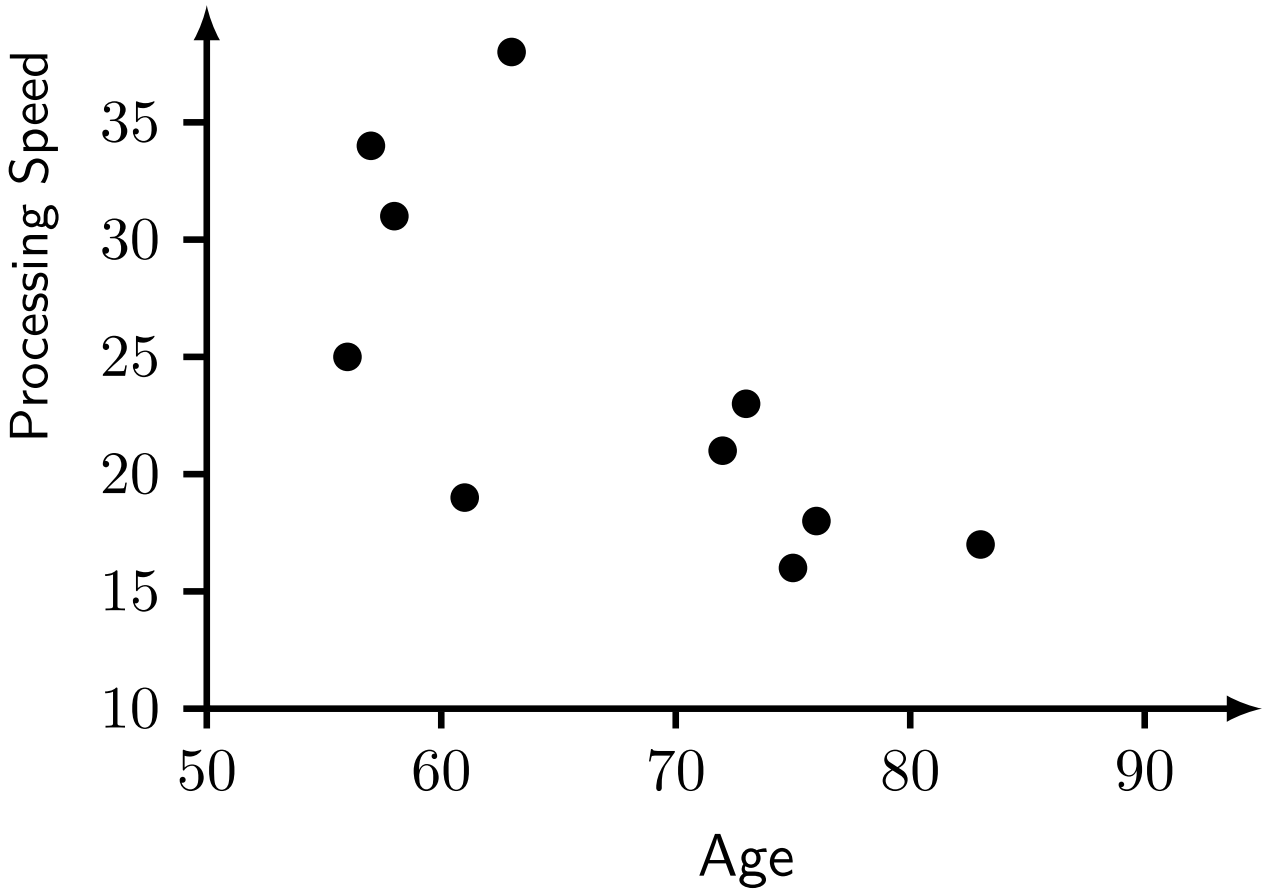
Person	1	2	3	4	5	6	7	8	9	10
x: Age	56	57	58	61	63	72	73	75	76	83
y: PS[†]	25	34	31	19	38	21	23	16	18	17

[†]Processing speed.

- Descriptive statistics:
  $\hat{\mu}_x = 67.4$, $s_x^2 = 81.44$, $\hat{\sigma}_x^2 = 90.48$,
  $\hat{\mu}_y = 24.2$, $s_y^2 = 52.96$, $\hat{\sigma}_y^2 = 58.84$,
  $s_{xy} = -45.18$, $\hat{\sigma}_{xy} = -50.20$, $r = -0.69$.

Example and Data

- Graphical representation



Defining a Model: Functional specification

- Is processing speed related to age, ie. does it depend on it?
- Formally we can express this as

$$y = f(x)$$

- y corresponds to processing speed or
  - the dependent variable, the endogenous variable, the criterion variable.
  - This variable should be at least on an interval scale.
- x is the age of the respondent
  - the independent variable, the exogenous variable, the predictor variable.
  - Most often it will be on an interval scale but it can be on a nominal scale.
  - If not stated otherwise we will assume x to be on an interval scale.

Defining a Model: Functional specification

- **Assumption 1**
  The relation among y and x is linear, i.e.

$$y = \beta_0 + \beta_1 x.$$

- This assumption holds (obviously only) for linear models.
- Meaning:
  - Used to simplify real-world problems.
  - β_0 is the intercept.
  - β_1 is the slope.
  - As such we assume the relation among x and y to be linear.

Defining a Model: Functional specification

- **Assumption 2**
  The parameters β_0 and β_1 are for all N observations the same, i.e.,

$$y_i = \beta_0 + \beta_1 x_i \quad (i = 1 \dots N)$$

- Meaning:
  - In principle, the relation among x and y is equally representative for all observations.
  - Each observation provides the same amount of information regarding the relation among x and y.
  - Observations must be independent from each other (e.g. *not* from couples).

Defining a Model: Functional specification

- **Assumption 3**
  No relevant exogenous and endogenous variables are missing in the regression equation and the exogenous variables x is not irrelevant.
- Meaning:
  - If a relevant exogenous variable is ignored then the parameter estimates for β_1 is biased (unless the exogenous variables are completely independent, i.e., uncorrelated)
  - This specification error is probably the most common "error" (not residual) in regression type analyses.
  - In case x is irrelevant, we may bias the estimation of the standard error of β_1 (estimate is inefficient).

Defining a Model: Error (Disturbance)

- The prediction of y on the basis of x is not perfect due to several reasons.
- For example:
  - The data used in the model may contain unsystematic error (different sources such as measurement, test setting etc.)
  - Human behavior is not completely deterministic but only predictable to some degree.
  - A number of "irrelevant" distractors may differentially affect participants (e.g. fatigue, motivation, focus, etc.).
- Typically there will be some deviation from the actual y-value and from the one predicted on the basis of x.
- This deviation is referred to as statistical error or disturbance.

Defining a Model: Error or Disturbance

- Errors or disturbances are unsystematic and different for each person.
- Formally, we need to expand our model to

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1 \ldots N). \tag{1}$$

- Meaning:
  - ϵ_i denotes the error for person i.
  - ϵ_i corresponds to the deviation from the actual value y and the predicted value $\hat{y}_i$, i.e., $\epsilon_i = y_i - \hat{y}_i$.
- Hence, Equation (1) can be expressed as

$$\hat{y}_i = \beta_0 + \beta_1 x_i. \tag{2}$$

Defining a Model: Error or Disturbance

- **Assumption 1**
  The expected value of the error is 0 for all N participants, i.e.

$$\mathrm{E}(\epsilon_i) = 0. \tag{3}$$

- Meaning:
  - The probability, of ϵ_i being zero, is highest.
  - This does not mean that y_i and $\hat{y}_i$ coincide for all participants, not even for one single person!
  - It holds however: Were we to compute an infinite number of times y_i and $\hat{y}_i$, then we would obtain an average value of 0 for $y_i - \hat{y}_i$ (according to the law of large numbers)

Defining a Model: Error or Disturbance

- **Assumption 2**
  The variance of the error is equal and constant for all N persons, i.e.

$$\mathsf{V}(\epsilon_i) = \sigma_\epsilon^2 \quad (< \infty). \tag{4}$$

- Meaning:
  - The probability of a deviation among y_i and $\hat{y}_i$ is equally large for all persons.
  - If we obtained an infinite number of measurements for all persons then (according to the law of large numbers) the variance among y_i and $\hat{y}_i$ would be the same for all persons.
  - The precision of the prediction of y given x is (in principle) constant for each person i.

Defining a Model: Error or Disturbance

- **Assumption 3**
  The errors of two arbitrarily chosen data points i and j are independent of each other, i.e.

$$\mathrm{C}(\epsilon_i, \epsilon_j) = 0 \quad (i \neq j; i, j = 1 \ldots N).$$

- Meaning:
  - The covariance among ϵ_i and ϵ_j is 0.
  - Knowing the error of person i will not inform us by any means about the error of person j
  - Hence, each person provides the same amount of information for predicting y on the basis of x.

Defining a Model: Error or Disturbance

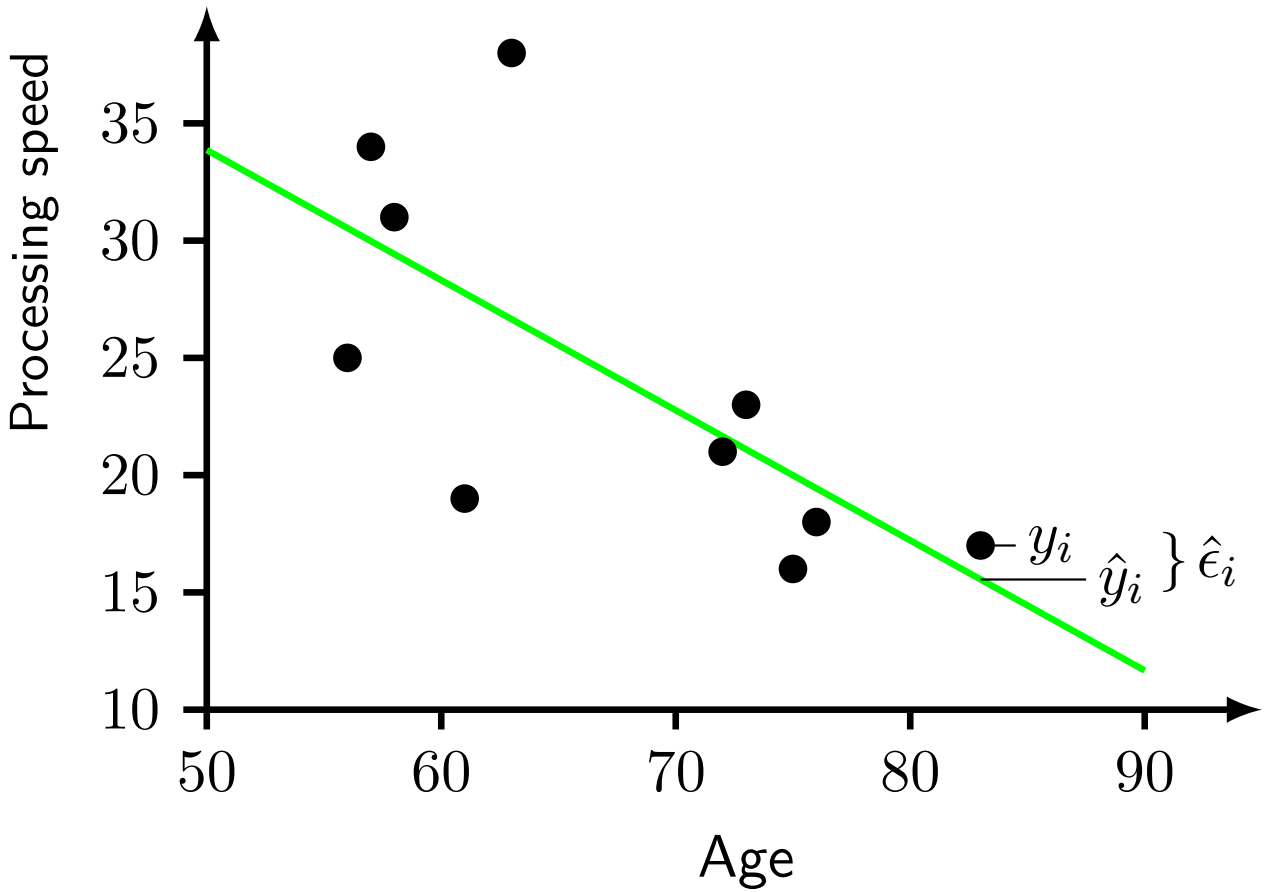
- **Assumption 4**
  The error of any person i follows a normal (or Gaussian) distribution, i.e.

$$\epsilon_i \sim N(\mathsf{E}[\epsilon_i], \mathsf{V}[\epsilon_i]) \quad \equiv \quad \epsilon_i \sim N(0, \sigma^2_\epsilon).$$

- Meaning:
  - $\sim$ reads "is distributed as", N for normal distribution.
  - Given the symmetry of the normal distribution this means that positive and negative deviations among y_i and $\hat{y}_i$ are equally probable.
  - Given the shape of the normal distribution, small deviations among y_i and $\hat{y}_i$ are much more probable than large deviations.

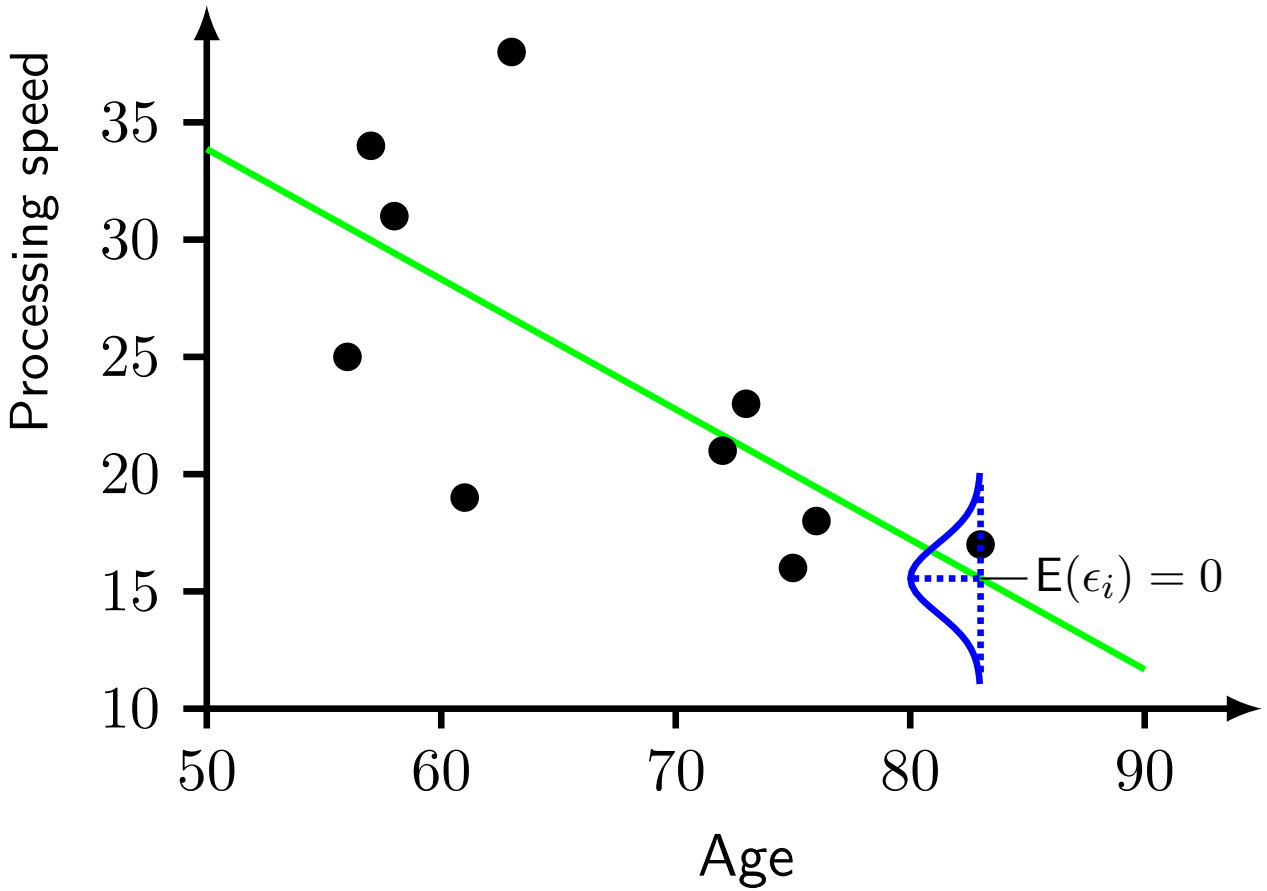
Defining a Model: Error or Disturbance

- Graphical representation



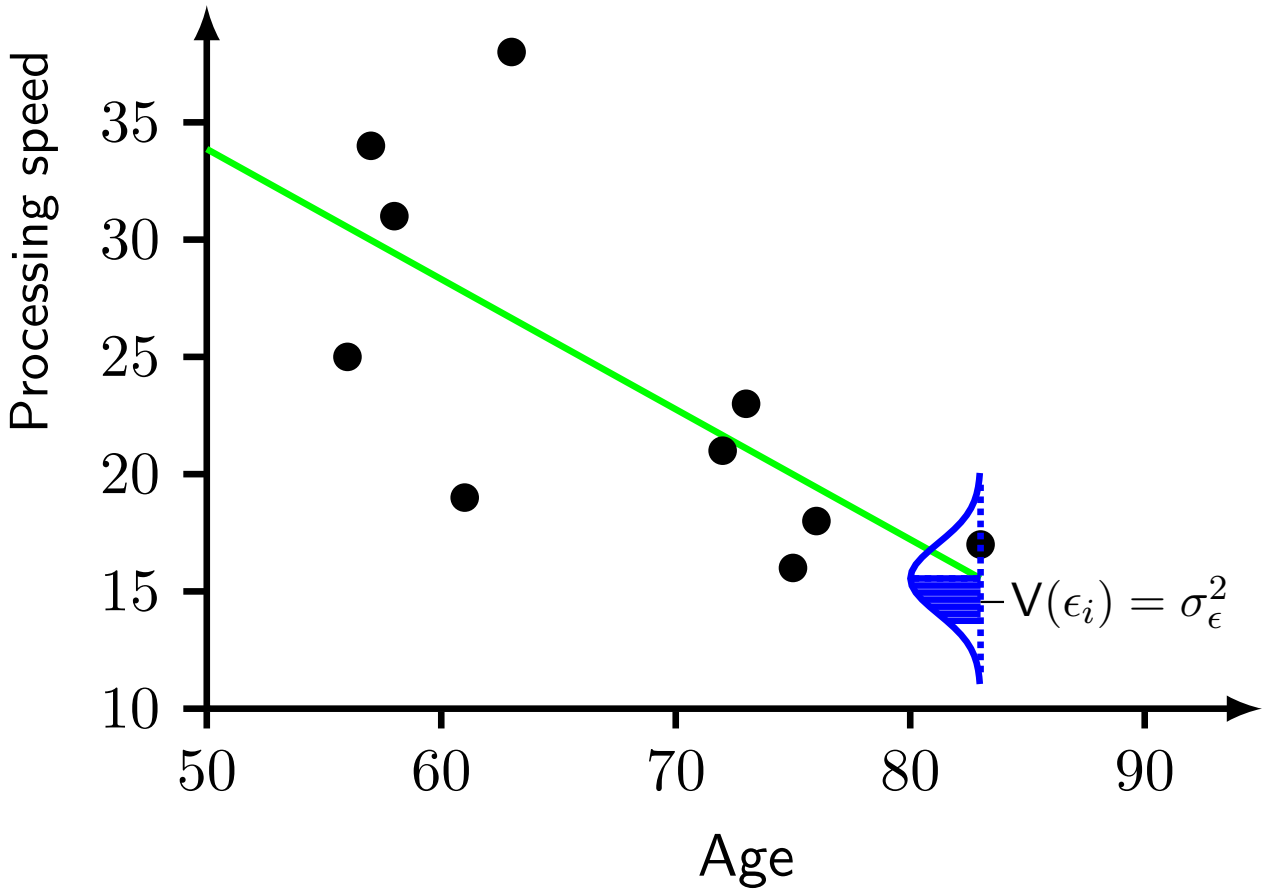
Defining a Model: Error or Disturbance

- Graphical representation



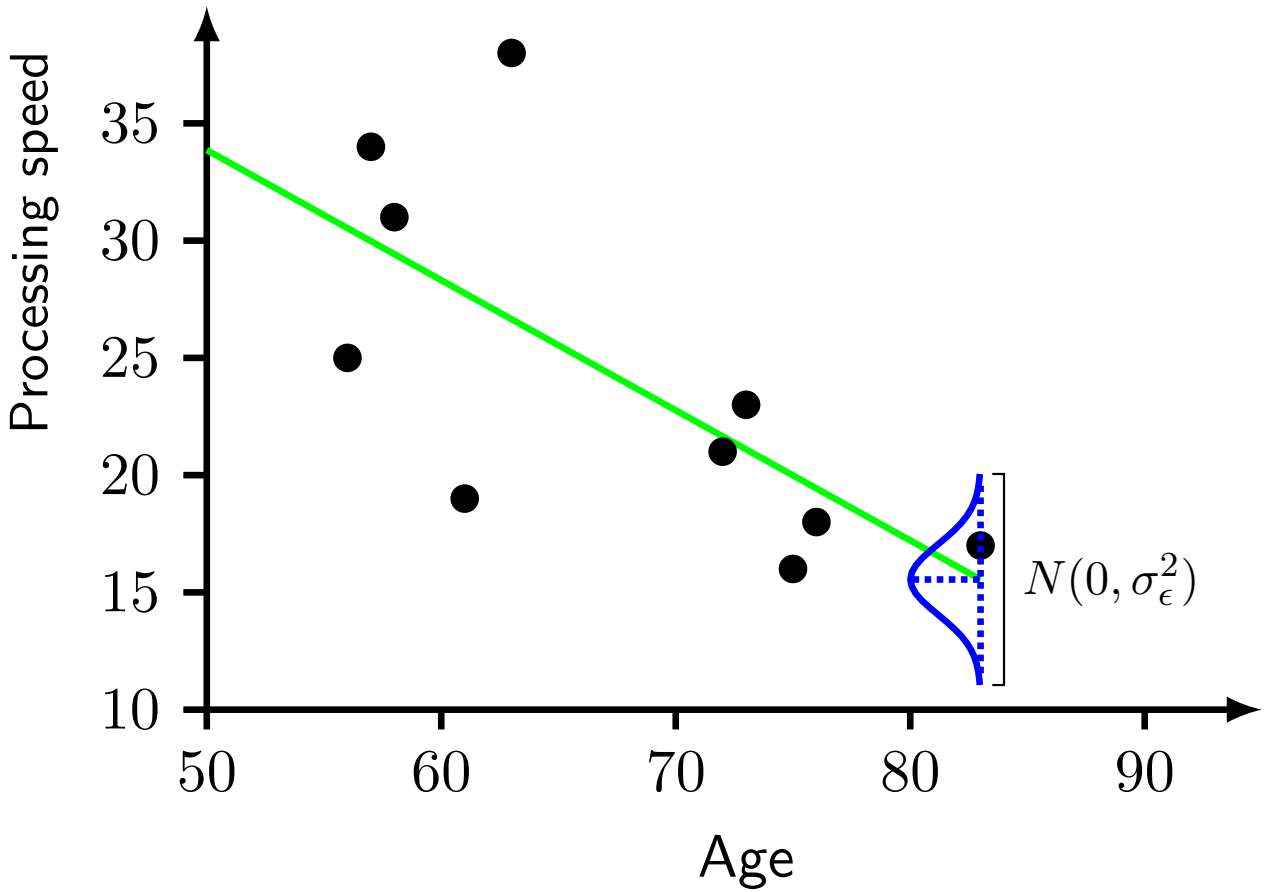
Defining a Model: Error or Disturbance

- Graphical representation



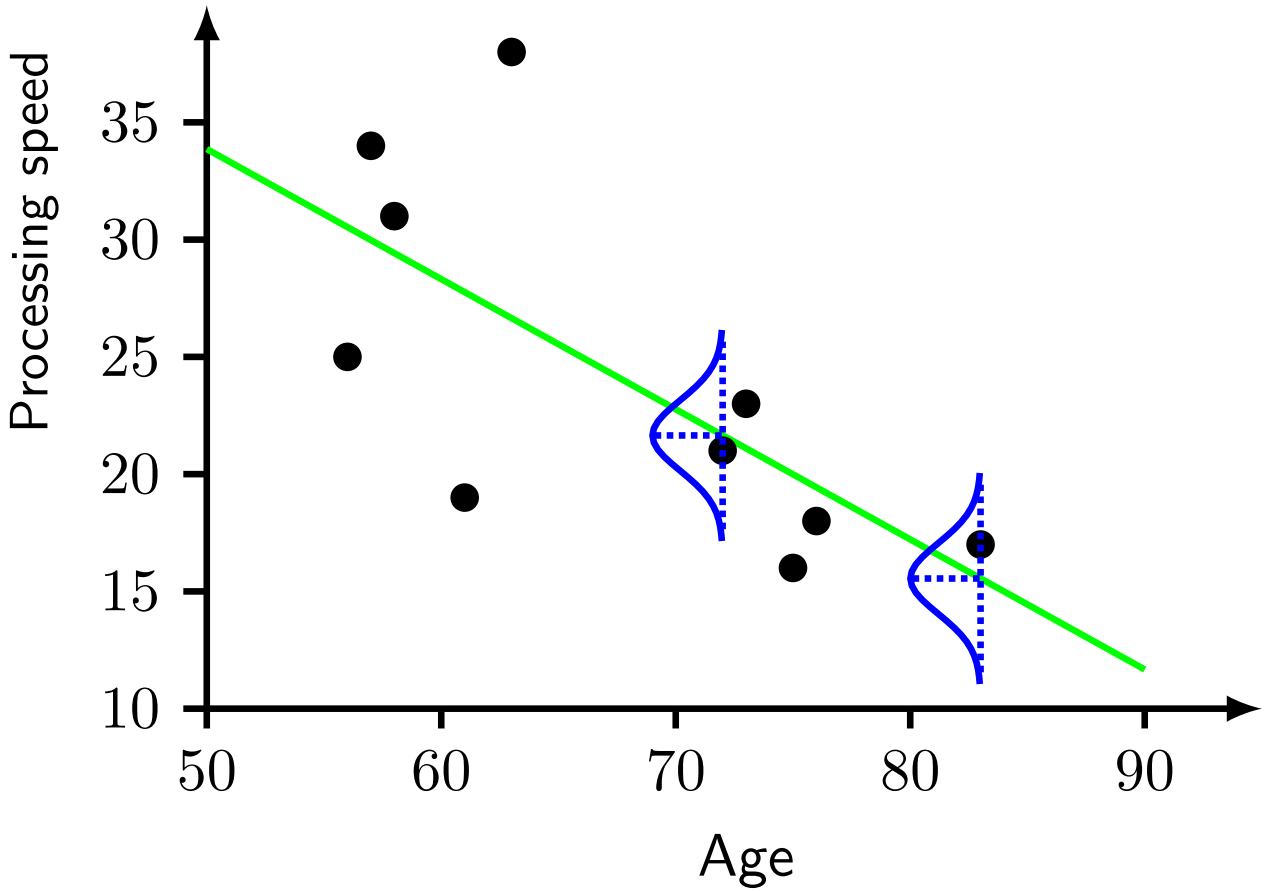
Defining a Model: Error or Disturbance

- Graphical representation



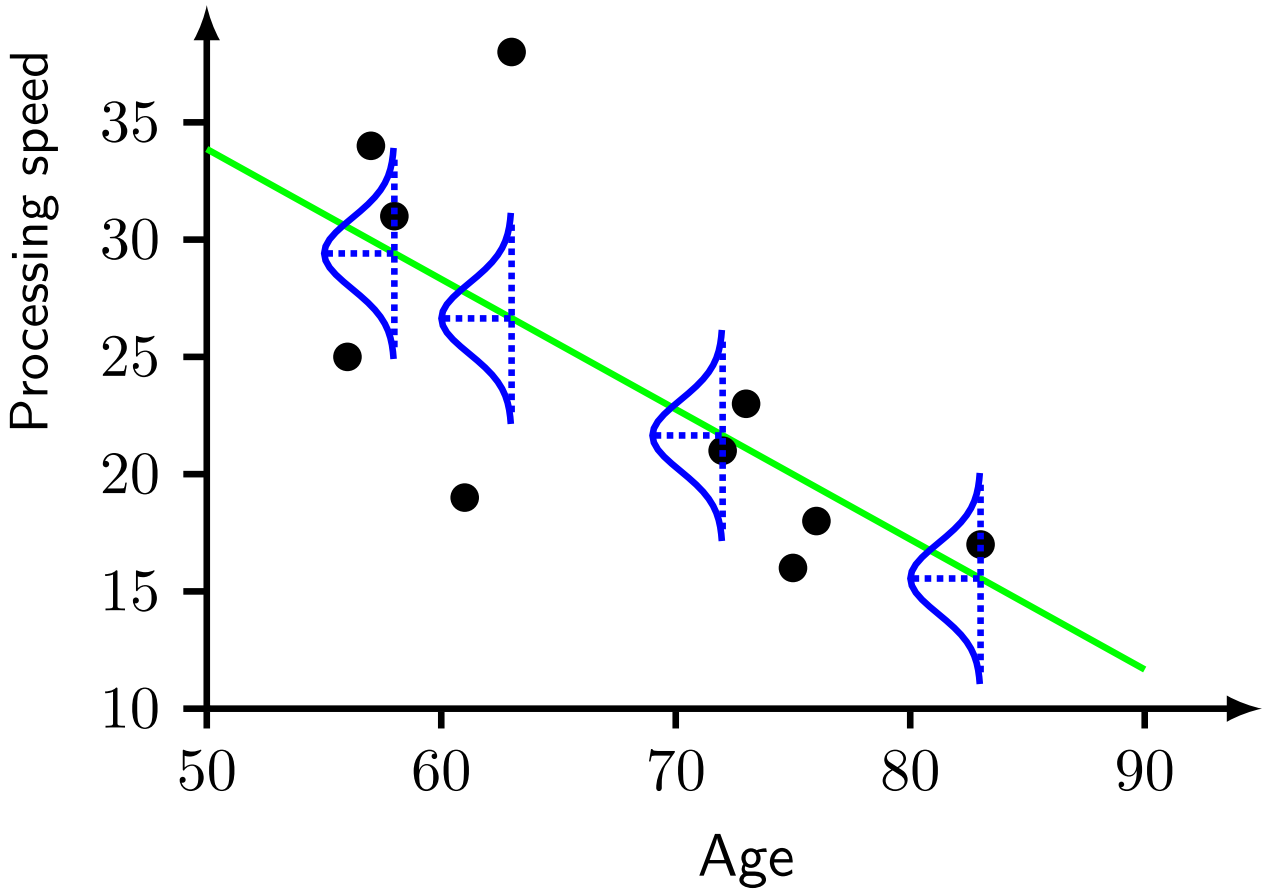
Defining a Model: Error or Disturbance

- Graphical representation



Defining a Model: Error or Disturbance

- Graphical representation



Defining a Model: Error or Disturbance

- Assumptions 1 to 3 regarding the error are referred to as **Gauss-Markov theorem**.
- They are not relevant for the estimation of the regression parameters β_0 and β_1 (but they are relevant for the BLUE properties).
- In order to statistically validate the estimated parameters β_0 and β_1 (significance testing and computation of confidence intervals) the normality assumption is indispensable.
- Assumption 4 results from the central limit theorem. Simplified: When many draws of independent random variables are summed up, their sum tends toward a normal distribution.

Defining a Model: Predictor Variable x

- **Assumption 1**
  The predictor variable x_i is not a random variable but deterministic for all i $(i = 1 \ldots N)$.
- Meaning:
  - We assume that x_i is free of any random influence.
  - Repeated measures of x_i should yield equal measurement values.
- **Assumption 2**
  The predictor variable x_i yields different measurements for different people i, i.e, the variance of x is larger than zero $(s_x^2 > 0)$.
  - If all values of x were the same, the regression line would be parallel to the y axis.
  - In this case the slope of the regression would not be defined.

Summary

- GLM is our framework for most of the statistics we encounter
- Covariances and correlations provide basic information on dependence or association among variables
- Generally a good idea to plot our data
- No functional form other than linear relation
- The GLM comes with a number of "implicit" assumption
  - Regarding its functional form
  - Regarding its error and its predictors

General Linear Model I: Regression

Philippe Rast

PSC 204B:
General Linear Model
January 11

UC Davis, Winter 2018

Overview

7 I Simple Linear Regression

- Overview
- Today
- Model Estimation
- Estimation
  - LS-Method
- Goodness of Fit
- Matrix Notation
  - LS-Method
- Conlusion
- Example and Data
- Model Specification
- Assumptions
- Predictors
- Goodness of fit

Fundamental Problem

- The *only* information we have are observed data
- The are a gazillion ways on how these data may have come about
- We want to know what ways are the most likely.

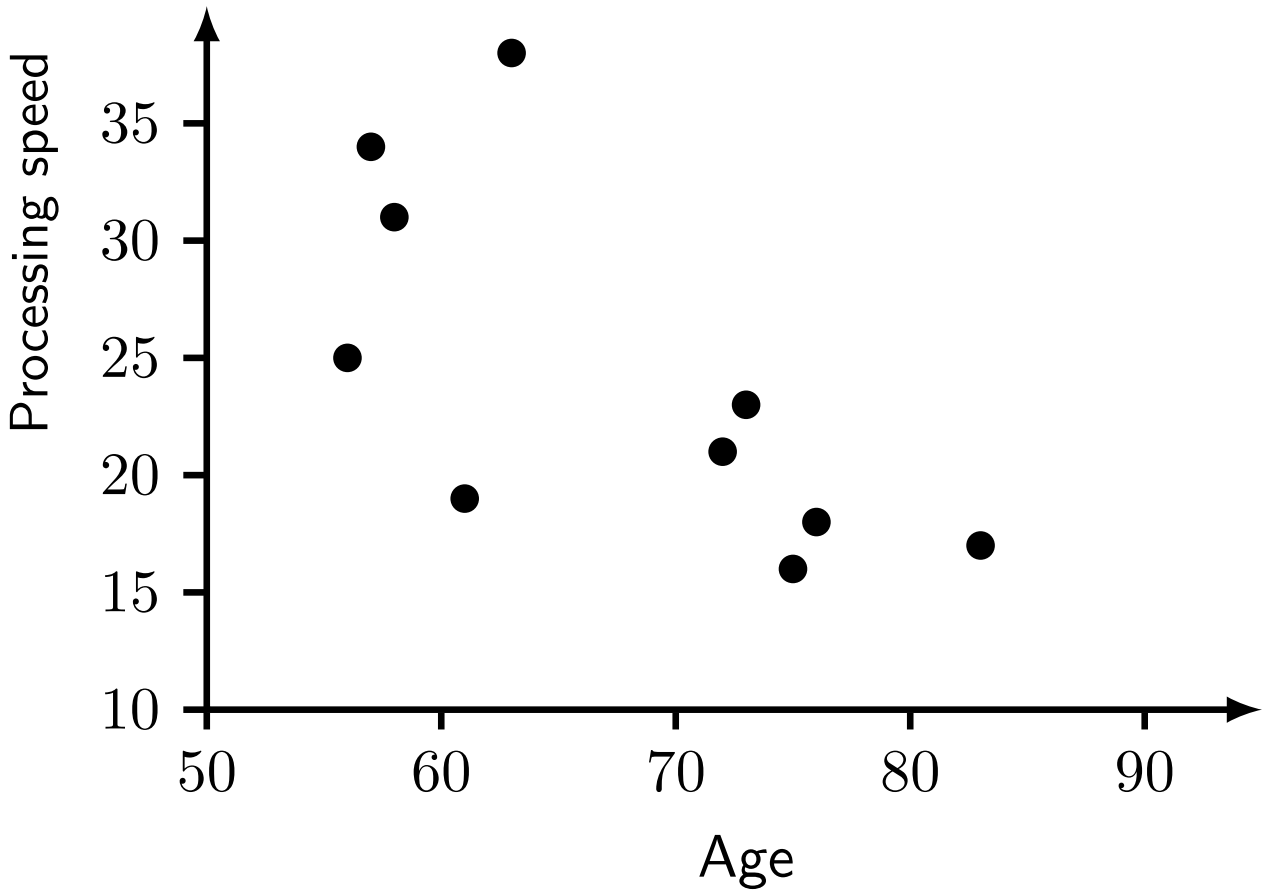
- What does statistics for us?
- Gives us a set of rules and techniques that are commonly used to approximate some common data types
- "rules and techniques" is nothing else then assumptions.
- All we have, besides our observed data are: Assumptions!

Today's Focus and Goals

- Find solution to the best regression line via the least squares method.
- Elaborate on effect sizes in the simple linear regression.
- Develop, carry out and interpret hypothesis tests with regard to β-parameters.

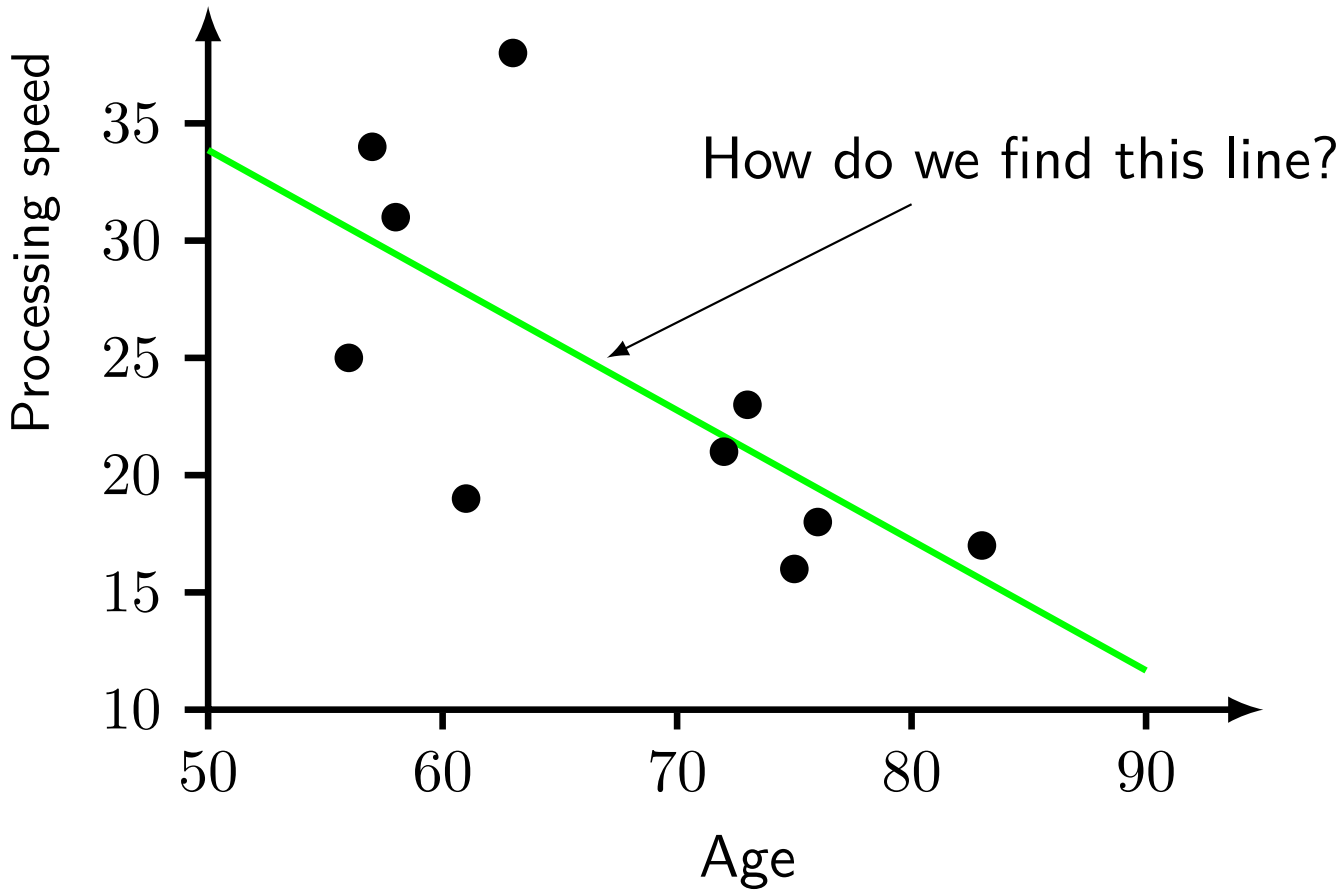
Defining a Model: Error or Disturbance

- Graphical representation



Defining a Model: Error or Disturbance

- Graphical representation



Model Estimation

- What is the best way to estimate the regression line?
- What does "best" mean in this context?
  - Alternative 1: The errors should be minimal over-all, i.e.,

$$\sum_{i=1}^{N} \hat{\epsilon}_i \rightarrow \text{Minimum}.$$

  - Alternative 2: The sum of squared errors (least squares) is to be minimized, i.e.,

$$\begin{aligned} \text{LS}_{\hat{\epsilon}} = \sum_{i=1}^{N} (\hat{\epsilon}_i - \hat{\mu}_\epsilon)^2 &\rightarrow \text{Minimum} \\ = \sum_{i=1}^{N} \hat{\epsilon}_i^2 &\rightarrow \text{Minimum}. \end{aligned}$$

Model Estimation: Least Squares (LS)

- Alternative 2 corresponds to the method of "least squares" as the sum of the *squared* errors is minimized.
  - For linear regression: **O**rdinary **L**east **S**quares (OLS).
  - By squaring, larger errors obtain more weight in the regression compared to smaller errors.
  - This approach avoids larger errors among y_i and $\hat{y}_i$
  - LS-estimators are BLUE (*B*est *L*inear *U*nbiased *E*stimator), i.e., for linear models there are no other estimators with better properties regarding bias (expected value – true value) and efficiency (smallest mean squared error).
  - Given the Gauss-Markov theorem there is no other estimator with smaller variance compared to the LS-estimator.

Model Estimation: Least Squares (LS)

- Derivation of estimator:

$$\begin{aligned} \mathrm{LS}_{\hat{\epsilon}} &= \sum_{i=1}^{N} \hat{\epsilon}_i^2 \\ &= \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned} \tag{5}$$

- Expression (5) needs to be minimized.
- Minimum: First derivative is zero and second derivative is larger than zero.

Model Estimation: Least Squares (LS)

- Expand $\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$:

$$\mathrm{LS}_{\hat{\epsilon}} = \sum_{i=1}^{N}(y_i^2 + \hat{\beta}_0^2 + \hat{\beta}_1^2 x_i^2 - 2y_i\hat{\beta}_0 - 2y_i\hat{\beta}_1 x_i + 2\hat{\beta}_0\hat{\beta}_1 x_i)$$

- First partial derivatives of both parameters in Equation (5):

$$\begin{aligned} \frac{\partial \mathrm{LS}_{\hat{\epsilon}}}{\partial \hat{\beta}_0} &= -2\sum_{i=1}^{N} y_i + 2N\hat{\beta}_0 + 2\hat{\beta}_1 \sum_{i=1}^{N} x_i \qquad (6) \\ \frac{\partial \mathrm{LS}_{\hat{\epsilon}}}{\partial \hat{\beta}_1} &= -2\sum_{i=1}^{N} y_i x_i + 2\hat{\beta}_0 \sum_{i=1}^{N} x_i + 2\hat{\beta}_1 \sum_{i=1}^{N} x_i^2 \end{aligned}$$

Model Estimation: Least Squares (LS)

- Set to zero and solve for $\hat{\beta}$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2} \tag{8}$$

- Insert (7) in (8):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} y_i x_i - \bar{y} \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2 - \bar{x} \sum_{i=1}^{N} x_i} \quad \text{[Extend with 1/N]}$$

$$= \frac{s_{xy}}{s_x^2} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \frac{\text{Covariance of } x \text{ and } y}{\text{Variance of } x}.$$

Model Estimation: Least Squares (LS)

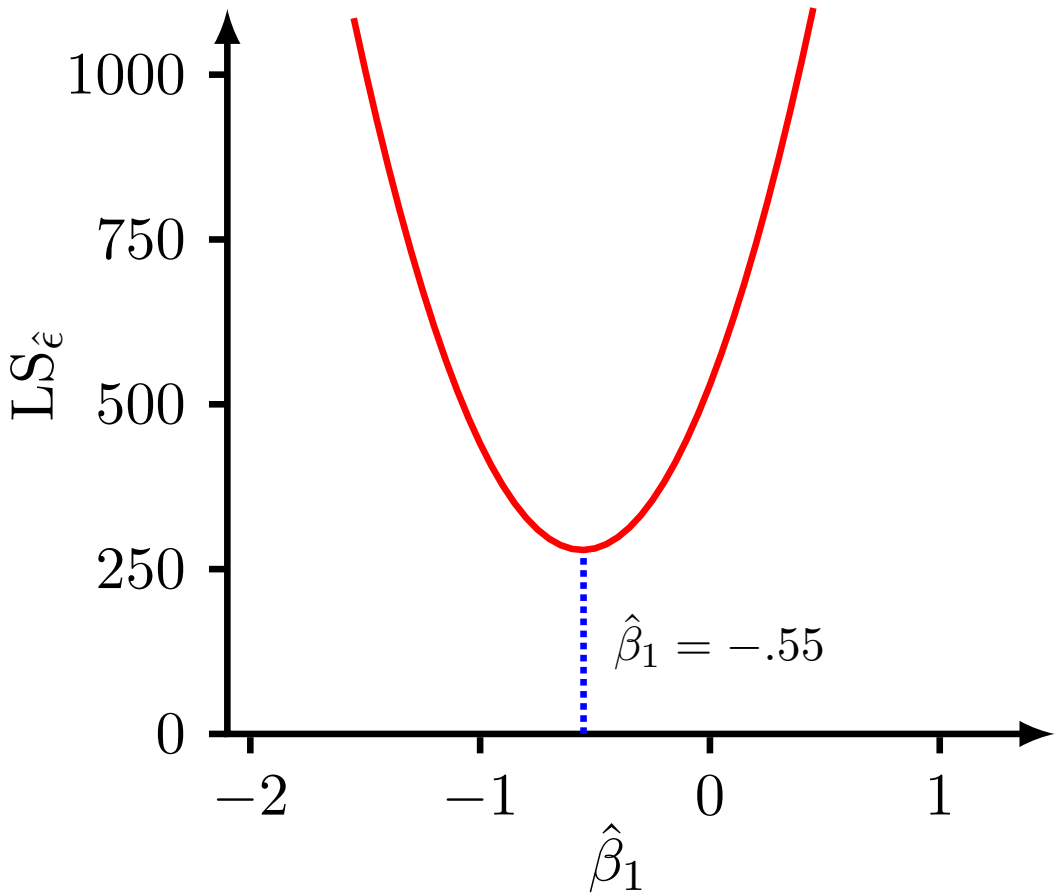
- Given that the second partial derivatives ($2N$ and $2\sum x_i^2$) are greater than zero, the sum of squared errors is minimal.
- Inserting values from our example from Table 1
  - $\bar{x} = 67.4$, $\bar{y} = 24.2$, $\hat{\sigma}_x^2 = 90.48$, $\hat{\sigma}_{xy} = -50.2$.
  - $\hat{\beta}_1 = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \frac{-50.2}{90.48} \approx -0.55$.
  - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 24.2 + .55 \times 67.4 = 61.27$.
  - The full regression equation is

$$\hat{y}_i = 61.27 - .55x_i.$$

- Interpretation:
  - For a person with age 0, we predict a processing speed value of 61.27.
  - For each additional year of age we predict a loss in processing speed of $-.55$.

Model Estimation: Least Squares (LS)

- Graphical representation: $\text{LS}_{\hat{\epsilon}}$ in dependence of $\hat{\beta}_1$



Goodness of Fit: Explained Variance (R^2)

- How good is our prediction of y given x?
  - If the prediction was perfect the variance of the predicted values would correspond exactly to the variance of the original values i.e., $\sigma^2_{\hat{y}} = \sigma^2_y$.
  - Typically, the variance of the predicted values will be smaller than the variance of the original values because our prediction of y using x also includes errors and unsystematic influences.
  - It holds:
$$\hat{\sigma}^2_y = \hat{\sigma}^2_{\hat{y}} + \hat{\sigma}^2_\epsilon.$$
  - An index for the ratio of "explained" variance or the effect size thus is
$$R^2 = \frac{\hat{\sigma}^2_{\hat{y}}}{\hat{\sigma}^2_y} = \frac{\hat{\sigma}^2_{\hat{y}}}{\hat{\sigma}^2_{\hat{y}} + \hat{\sigma}^2_\epsilon} = 1 - \frac{\hat{\sigma}^2_\epsilon}{\hat{\sigma}^2_y}.$$

Goodness of Fit: Explained Variance (R^2)

- How well can we predict processing speed due to age of respondent?
  - We obtain
$$\hat{\sigma}^2_\epsilon = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-1} = 30.99$$
  - Hence, the explained variance is:
$$R^2 = 1 - \frac{\hat{\sigma}^2_\epsilon}{\hat{\sigma}^2_y} = 1 - \frac{30.99}{58.84} \approx .47$$
- Interpretation:
  - Age explains approximately 47% of the variance in processing speed.
  - Individual differences in age explain approximately 47% of the individual differences in processing speed.

Model Estimation: Adjusted R^2

- R^2 tends to increase spuriously when additional parameters are included in the regression
- Adjusted R^2 ($\bar{R}^2$) takes into account the number of predictors:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-p-1}$$

where N is sample size and p is the number of predictors (excluding intercept)

- In our example

$$\bar{R}^2 = 1 - (1 - 0.47)\frac{10-1}{10-1-1} = 0.47\frac{9}{8} = .40$$

Goodness of Fit: Explained Variance (R^2)

- 40% explained Variance. Is this a lot? Is this little?
- Following a very broad classification by Cohen (1988)
  - R^2 of 1% is a small effect.
  - R^2 of 10% is a medium effect.
  - R^2 of 25% is a large effect.
- Following this classification we are looking at a (very) large effect.
- This classification is only a reference point.
  - If previous studies were only able to show effect sizes of .05, then an R^2 of .10 may be interpreted as a large (in the sense of: relevant) effect.

Conclusions

- What did we gain so far?
  - Under the given assumptions we have found a linear relation among age and processing speed (PS) in sample data.
  - The relation is based on the assumption that with increasing age PS declines.
  - Our calculations yielded $\hat{\beta}_0 = 61.27$ and $\hat{\beta}_1 = -.55$.
  - In the sample we have found a linear relation among age and PS of ($R^2 = 47\%$).

Simple Linear Regression in Matrix Notation: Example

Table: Data from the Longitudinal Aging Study Amsterdam

Person	1	2	3	4	5	6	7	8	9	10
x: Age	56	57	58	61	63	72	73	75	76	83
y: PS†	25	34	31	19	38	21	23	16	18	17

†Processing speed.

- We define

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 25 \\ 34 \\ \vdots \\ 18 \\ 17 \end{bmatrix} ; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} ; \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} = \begin{bmatrix} 1 & 55 \\ 1 & 57 \\ \vdots & \vdots \\ 1 & 76 \\ 1 & 83 \end{bmatrix} ; \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_N \end{bmatrix} .$$

Simple Linear Regression in Matrix Notation

- In matrix notation we write

$$\underset{N\times 1}{\mathbf{y}} = \underset{N\times 2}{\mathbf{X}}\ \underset{2\times 1}{\boldsymbol{\beta}} + \underset{N\times 1}{\boldsymbol{\epsilon}}\ . \tag{9}$$

- This is equivalent to the linear system of equations

$$\begin{aligned}
y_1 &= 1\times\beta_0 + x_1\times\beta_1 + \epsilon_1\\
y_2 &= 1\times\beta_0 + x_2\times\beta_1 + \epsilon_2\\
\vdots &= \vdots\\
y_i &= 1\times\beta_0 + x_i\times\beta_1 + \epsilon_i\\
\vdots &= \vdots\\
y_N &= 1\times\beta_0 + x_N\times\beta_1 + \epsilon_N.
\end{aligned}$$

Simple Linear Regression in Matrix Notation

- For the predicted values we have

$$\underset{N\times 1}{\hat{\mathbf{y}}} = \underset{N\times 2}{\mathbf{X}}\ \underset{2\times 1}{\hat{\boldsymbol{\beta}}}\ . \tag{10}$$

- This is equivalent to the linear system of equations

$$\begin{aligned}
\hat{y}_1 &= 1 \times \hat{\beta}_0 + x_1 \times \hat{\beta}_1 \\
\hat{y}_2 &= 1 \times \hat{\beta}_0 + x_2 \times \hat{\beta}_1 \\
\vdots &= \vdots \\
\hat{y}_i &= 1 \times \hat{\beta}_0 + x_i \times \hat{\beta}_1 \\
\vdots &= \vdots \\
\hat{y}_N &= 1 \times \hat{\beta}_0 + x_N \times \hat{\beta}_1.
\end{aligned}$$

Simple Linear Regression in Matrix Notation: LS-Method

- The least squares criterion (LS) in matrix notation

$$\underset{1\times N}{\hat{\boldsymbol{\epsilon}}'}\ \underset{N\times 1}{\hat{\boldsymbol{\epsilon}}} \rightarrow \text{Minimum} \tag{11}$$

- In matrix notation we obtain for the error

$$\begin{aligned} \hat{\boldsymbol{\epsilon}} &= \mathbf{y} - \hat{\mathbf{y}} \qquad \text{[Eq. 10]} \\ &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \end{aligned} \tag{12}$$

- equivalent to the linear system of equations

$$\begin{aligned} \hat{\epsilon}_1 &= y_1 - (1 \times \hat{\beta}_0 + x_1 \times \hat{\beta}_1) \\ \vdots &= \vdots \\ \hat{\epsilon}_i &= y_i - (1 \times \hat{\beta}_0 + x_i \times \hat{\beta}_1) \\ \vdots &= \vdots \\ \hat{\epsilon}_N &= y_N - (1 \times \hat{\beta}_0 + x_N \times \hat{\beta}_1). \end{aligned}$$

Parameter Estimation: Least Squares (LS)

- Parameter estimate

$$\begin{aligned} \mathrm{LS}_{\hat{\epsilon}} &= \sum_{i=1}^{N} \hat{\epsilon}_i^2 \\ &= \hat{\boldsymbol{\epsilon}}' \hat{\boldsymbol{\epsilon}} && \text{[Eq. 12]} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}. \end{aligned}$$

- Derive, set to zero and solve for $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{13}$$

  - Requirement for the estimation is that the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ exists. We will see that, especially in the case of multicollinearity, this requirement is not always fulfilled.

Summary Simple Linear Regression

- The simple linear regression serves to analyze the degree of a linear relation among two variables (dependent in independent variable; criterion and predictor)
- The two parameters (intercept and slope) are estimated by means of least squares (the sum of squared errors is being minimized)
- The crucial benefit of matrix algebra is that
  - the computation is more direct, compared to the computation via means and variances.
  - Equation (13) also holds for cases with more than one predictor variable (*multiple* regression) in our model.
- We can represent models of any level of complexity in a compact manner and always compute them the same way.

Multiple Regression: Example and Data

Table: Data from the Longitudinal Aging Study Amsterdam

Person	1	2	3	4	5	6	7	8	9	10
x_1: Age	56	57	58	61	63	72	73	75	76	83
x_2: Education†	9	13	11	9	12	8	10	8	11	12
y: PS‡	25	34	31	19	38	21	23	16	18	17

†Years in school ‡Processing speed.

Descriptive statistics:

Age: $\hat{\mu}_{x_1} = 67.4$, $s^2_{x_1} = 81.44$, $\hat{\sigma}^2_{x_1} = 90.48$,

Education: $\hat{\mu}_{x_2} = 10.3$, $s^2_{x_2} = 2.81$, $\hat{\sigma}^2_{x_2} = 3.12$,

PS: $\hat{\mu}_y = 24.2$, $s^2_y = 52.96$, $\hat{\sigma}^2_y = 58.84$,

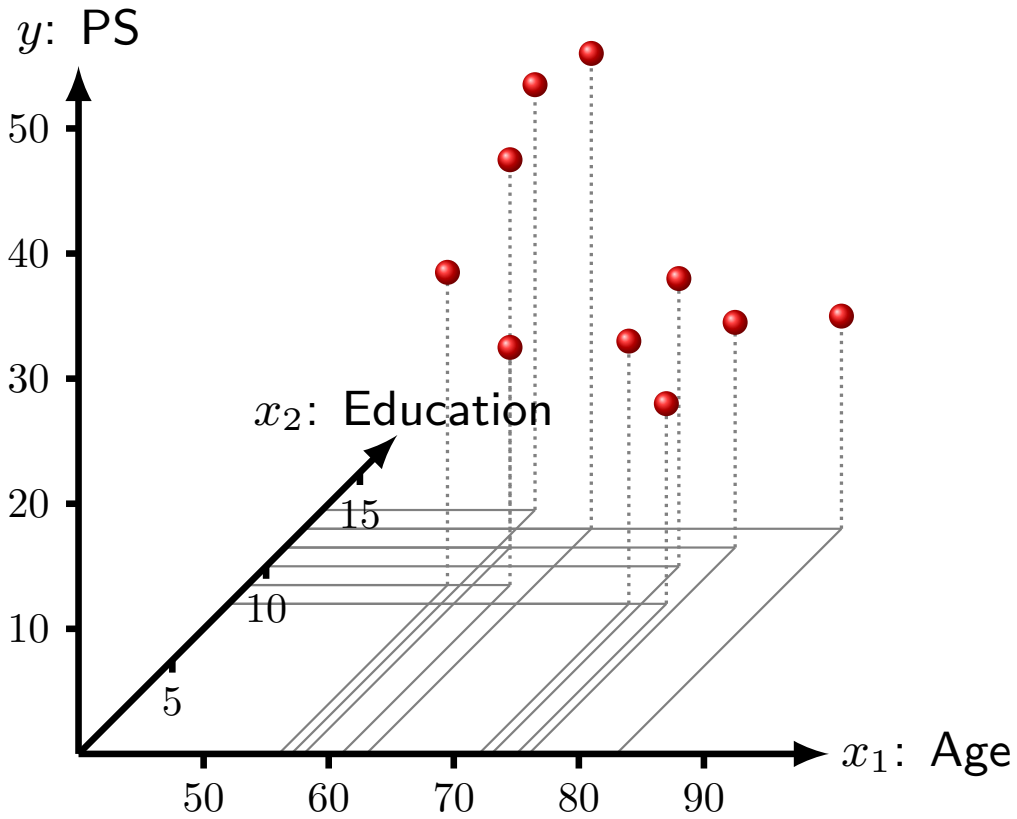
Age·PS: $s_{x_1y} = -45.18$, $\hat{\sigma}_{x_1y} = -50.20$, $r_{x_1y} = -.69$,

Education·PS: $s_{x_2y} = 7.04$, $\hat{\sigma}_{x_2y} = 7.82$, $r_{x_2y} = .58$,

Age·Education: $s_{x_1x_2} = -1.62$, $\hat{\sigma}_{x_1x_2} = -1.8$, $r_{x_1x_2} = -.11$.

Multiple Regression: Example and Data

- Graphical representation



Specifying a Model: Functional Specification

As we did for the simple linear regression, we will define a set of assumptions regarding the functional specification of our model:

- **Assumption 1**
  The relation among $\mathbf{y}$ and $\mathbf{X}$ is linear, i.e. in the population it holds

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}. \tag{14}$$

- **Assumption 2**
  The parameters in vector $\boldsymbol{\beta}$ are the same for for all N observations, i.e.,

$$y_i = \mathbf{x}_i{}'\boldsymbol{\beta} \quad (i = 1 \dots N). \tag{15}$$

- **Assumption 3**
  No relevant exogenous/independent variables are missing in the regression equation and the exogenous variables in $\mathbf{X}$ are not irrelevant.

Specifying a Model: Error or Disturbance

The assumptions regarding the error are:

- **Assumption 1**
  The expectation of the error is 0 for all N individuals, i.e. for the $N \times 1$-vector $\boldsymbol{\epsilon}$ it holds
  $$\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}. \tag{16}$$
- **Assumption 2**
  The variance of the error is the same and constant for all N individuals *and*
- **Assumption 3**
  The errors of two arbitrary individuals i and j are independent. This results in a covariance matrix of errors
  $$\mathrm{V}(\boldsymbol{\epsilon}) = \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma_\epsilon^2 \mathbf{I}, \tag{17}$$
  with $\mathbf{I}$ as a $N \times N$-identity matrix (diagonal!).

Specifying a Model: Error or Disturbance

- **Assumption 4**

  The errors of the N individuals are multivariate normal distributed, i.e. for the $N \times 1$-vector $\boldsymbol{\epsilon}$ it holds

  $$\boldsymbol{\epsilon} \sim MVN(\mathrm{E}[\boldsymbol{\epsilon}], \mathrm{V}[\boldsymbol{\epsilon}]).$$

  MVN is short for *multivariate* normal distribution.
  Assumptions 1 to 3 can be expressed in

  $$\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}). \tag{18}$$

- Multivariate normally distributed means that the normals distribution holds for multiple dimensions In the example (with two predictors) in that would be two dimensions. I.e. that would be a bivariate normal distribution.

Specifying a Model: Predictor Variables in $\mathbf{X}$

- **Assumption 1**
  The predictor variables in $\mathbf{X}$ are not random variables but deterministic for all i ($i = 1 \ldots N$).
- **Assumption 2**
  The predictor variables in $\mathbf{X}$ contain (1) for different individuals i different values and (2) are not linearly dependent (not collinear) from each other.
  Another equivalent form for assumption 2 would be

$$\det(\mathbf{X}'\mathbf{X}) \neq 0,$$

  where $\det(\cdot)$ represents the determinant.
  - Basically, assumption 2 makes sure that the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ actually exists – without it we would not be able to estimate the parameters.

R code

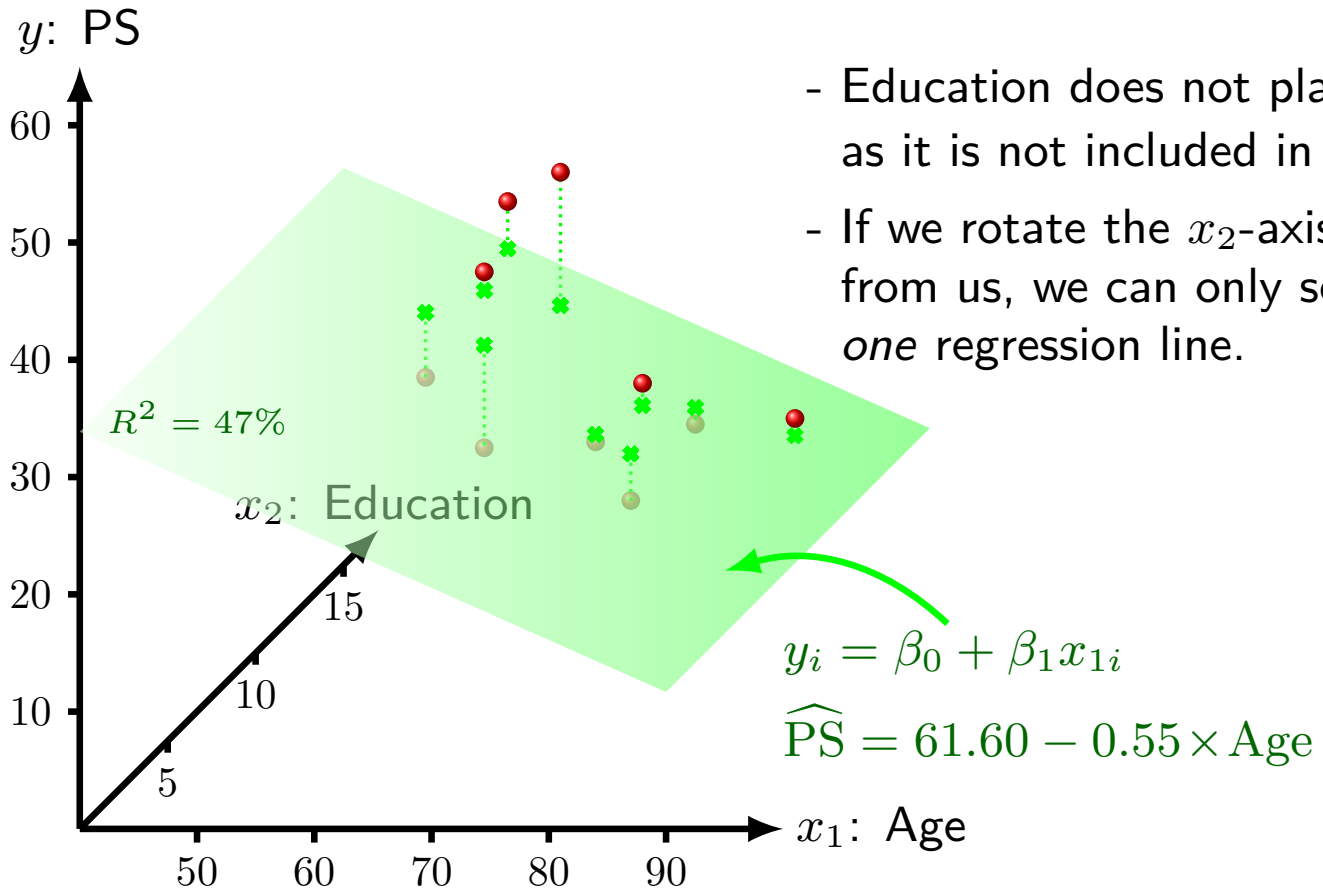
```
x1 = c(56, 57, 58, 61, 63, 72, 73, 75, 76, 83)
x2 = c( 9, 13, 11,  9, 12,  8, 10,  8, 11, 12)
y = c(25, 34, 31, 19, 38, 21, 23, 16, 18, 17)
lm(y ~ x1 + x2)

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##     35.8551      -0.5108       2.2109

X = cbind(rep(1,10), x1, x2)
solve( t(X) %*% X ) %*% t(X) %*% y

##           [,1]
##    35.8550716
## x1 -0.5107859
## x2  2.2108637
```

Parameter Estimation: Least Squares (LS)

- Interpretation of multiple regression:
  - $\hat{\beta}_0$: If all predictor variables are 0, then 35.86 is the value that we predict for y
  - $\hat{\beta}_1$: If all other predictors are held constant (here: Education) then we predict a negative change of -.51 in y for each additional year in chronological age.
  - $\hat{\beta}_2$: If all other predictors are held constant (here: Age) then we predict a positive change in y of 2.21 units for each additional year spent in school.
- What does "holding constant" mean?
  - Visually speaking, we reduce the k-dimensional space to two dimensions.
  - In our example we are looking at the plot from either a complete x_1 (age) perspective and thus ignoring x_2 or, from a x_2 (Education) and ignoring x_1.

Parameter Estimation

- Graphical representation: **Simple regression** of y on x_1



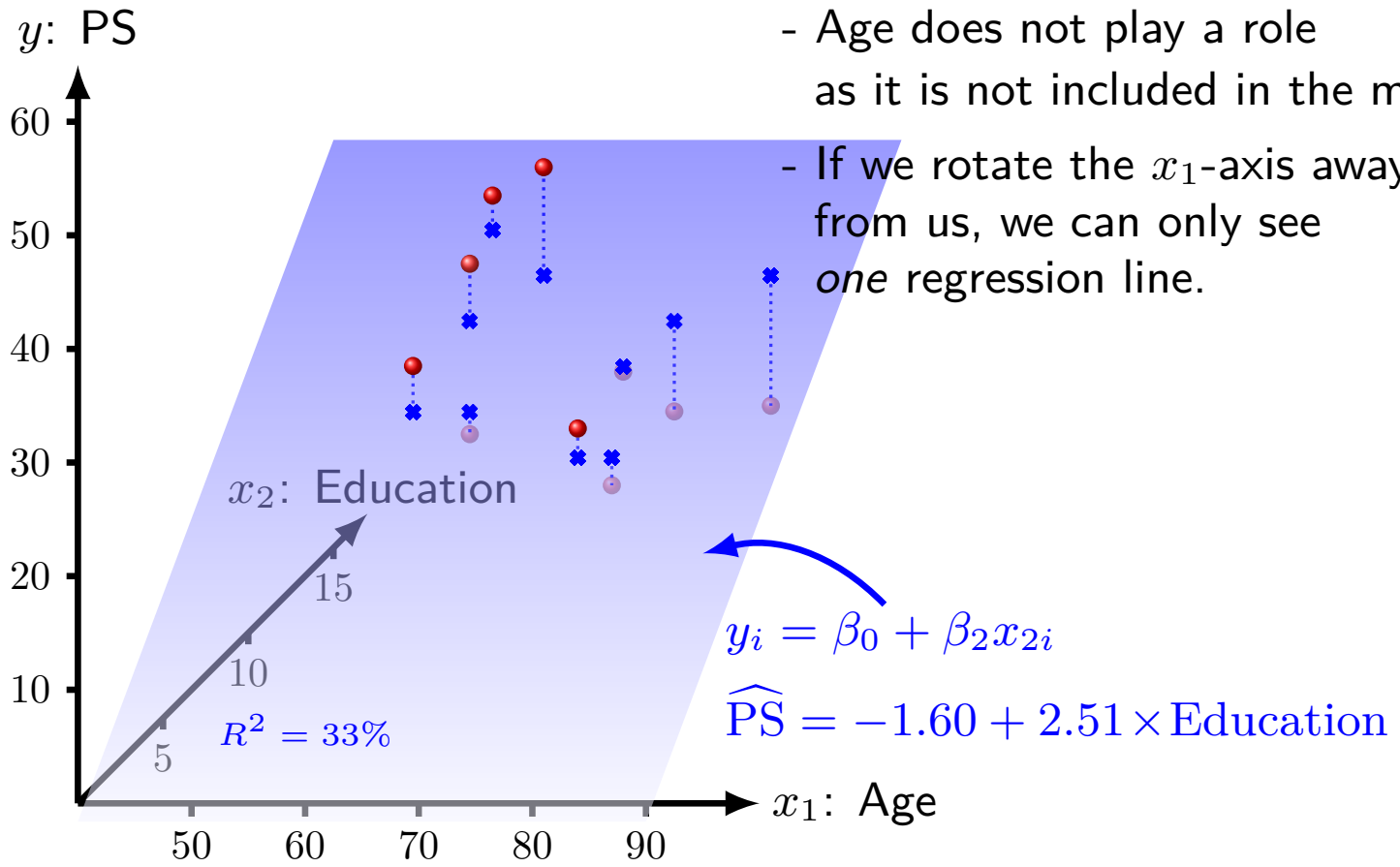
- Education does not play a role as it is not included in the model.
- If we rotate the x_2-axis away from us, we can only see *one* regression line.

$$y_i = \beta_0 + \beta_1 x_{1i}$$

$$\widehat{\text{PS}} = 61.60 - 0.55 \times \text{Age}$$

Parameter Estimation

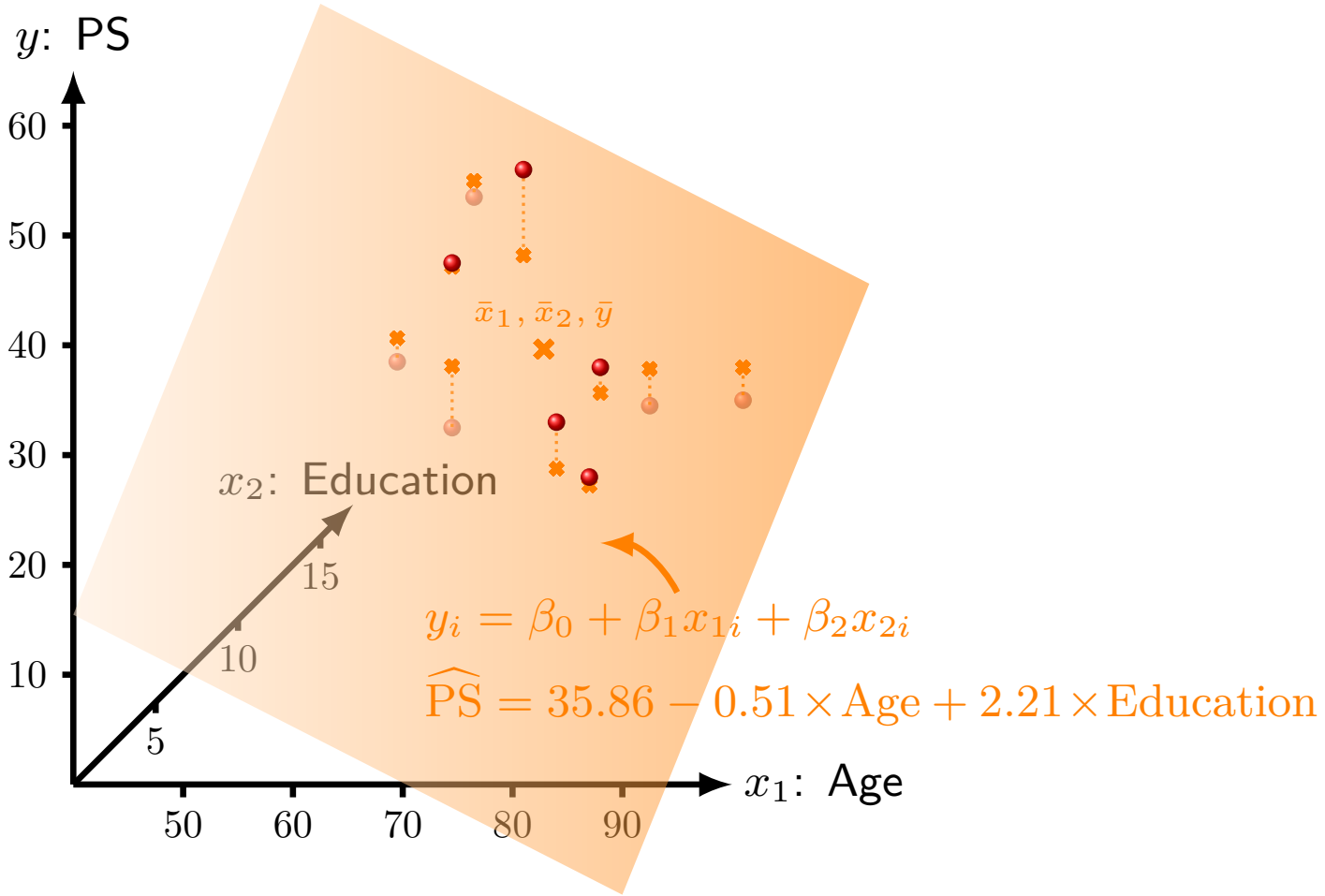
- Graphical representation: **Simple regression** of y on x_2



- Age does not play a role as it is not included in the model.
- If we rotate the x_1-axis away from us, we can only see *one* regression line.

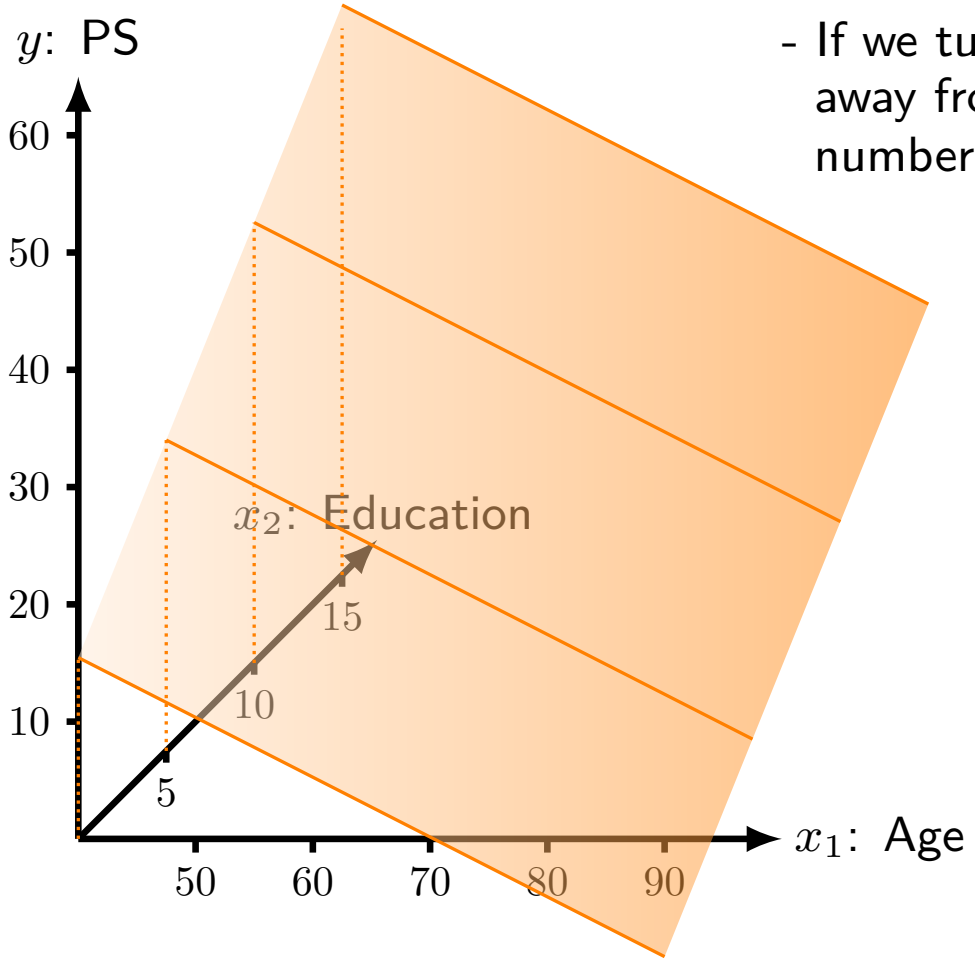
Parameter Estimation

- Graphical representation: **Multiple regression** of y on x_1, x_2



Parameter Estimation

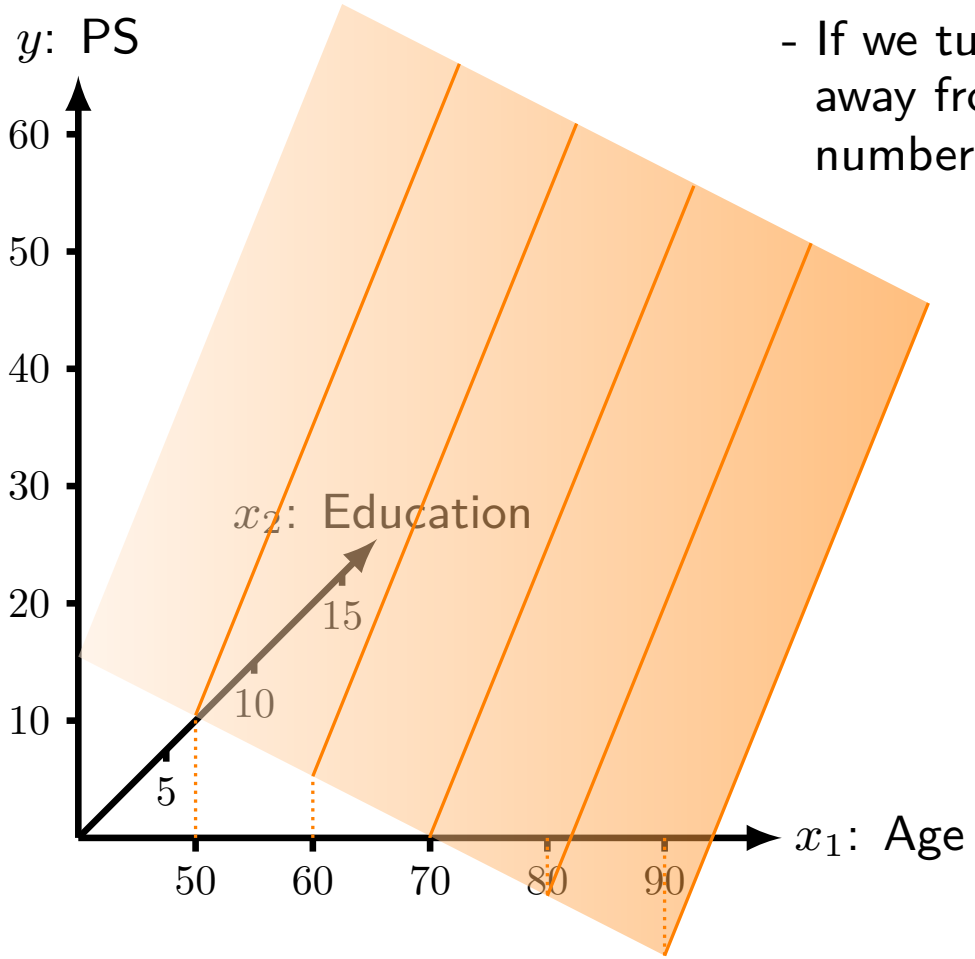
- Multiple regression of y on x_1, x_2; *Holding constant* x_2



- If we turn the x_2-axis away from us, we se a number of *parallel* lines.

Parameter Estimation

- Multiple regression of y on x_1, x_2; *Holding constant* x_1



- If we turn the x_1-axis away from us, we see a number of *parallel* lines.

Goodness of Fit: Variance Explained (R^2)

- How good is our prediction of y given x_1 and x_2?
- Analogous to the simple linear regression, variance explained is defined as

$$R^2 = \frac{\hat{\mathbf{y}}^{*\prime}\hat{\mathbf{y}}^*}{\mathbf{y}^{*\prime}\mathbf{y}^*} = 1 - \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{\mathbf{y}^{*\prime}\mathbf{y}^*}.$$

- If we use our values from the example we obtain

$$R^2 = \frac{386.42}{529.60} = 1 - \frac{143.18}{529.60} = 0.7296 \approx 73\%.$$

- Age and education explain 73% of the variance in processing speed.

"Guessing" as alternative to LS

LS is not the only method to obtain parameter estimates

- Some combinations of parameters will be more likely to produce y
- PDF: $p(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$

  $\mu = \mathbf{X}\boldsymbol{\beta}$

  $\sigma^2 = \mathsf{V}(\boldsymbol{\epsilon})$

  $\mathbf{y}$
- We have $\mathbf{X}$ and $\mathbf{y}$

- Let's guess, $\boldsymbol{\beta}$ and σ^2!

R code: LL.R