# Data Analysis and Visualization

## Linear regression

Matthias Heinig, Jan Krumsiek

# Overview: Linear regression

- A simple linear model
- Parameter estimation
- Hypothesis testing
- Multiple linear regression
- Model selection
- Diagnostic plots

# A simple linear model

# A simple linear model

A *simple linear model* allows to study the relationship between two continuous variables

- one variable `x` is the *predictor*, *explanatory* or *independent* variable
- the other variable `y` is the *response*, *outcome* or *depdentent* variable
- the model is called *simple* because we study only one predictor variable

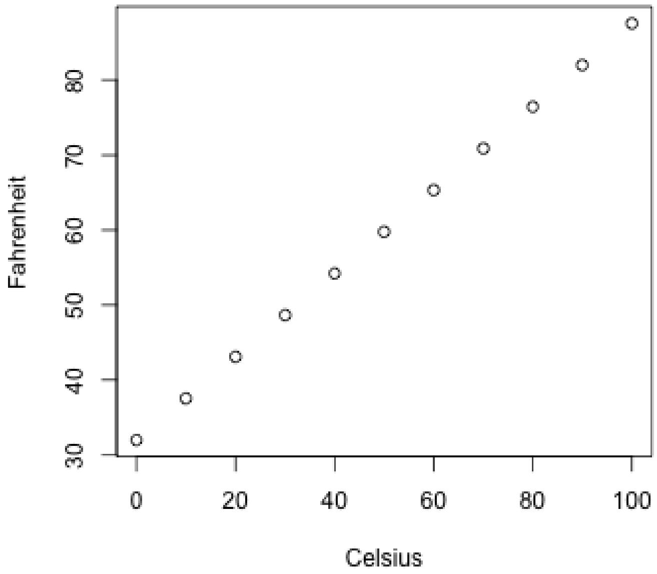
*Goals* of the analysis are

1. Prediction of future observations.
2. Assessment of the effect of, or relationship between, explanatory variables on the response.
3. A general description of data structure.

Further reading:
https://rafalab.github.io/dsbook/case-study-is-height-hereditary.html
(https://rafalab.github.io/dsbook/case-study-is-height-hereditary.html)

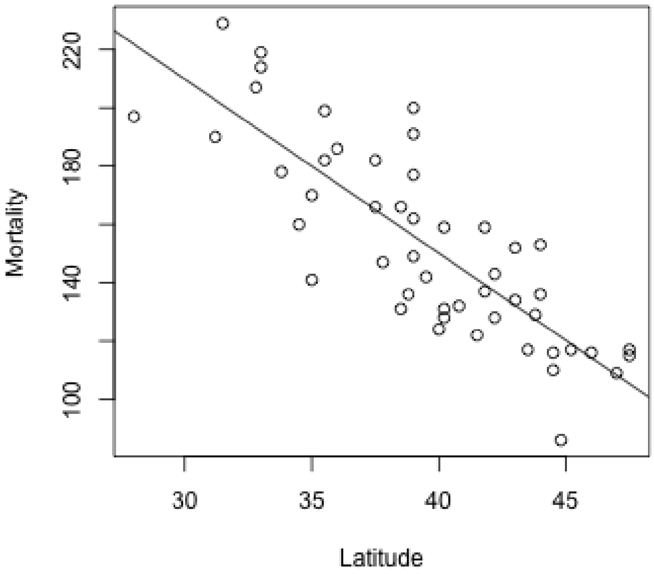
# Deterministic vs statistical model

Deterministic



$$F = \frac{5}{9}C + 32$$

Statistical

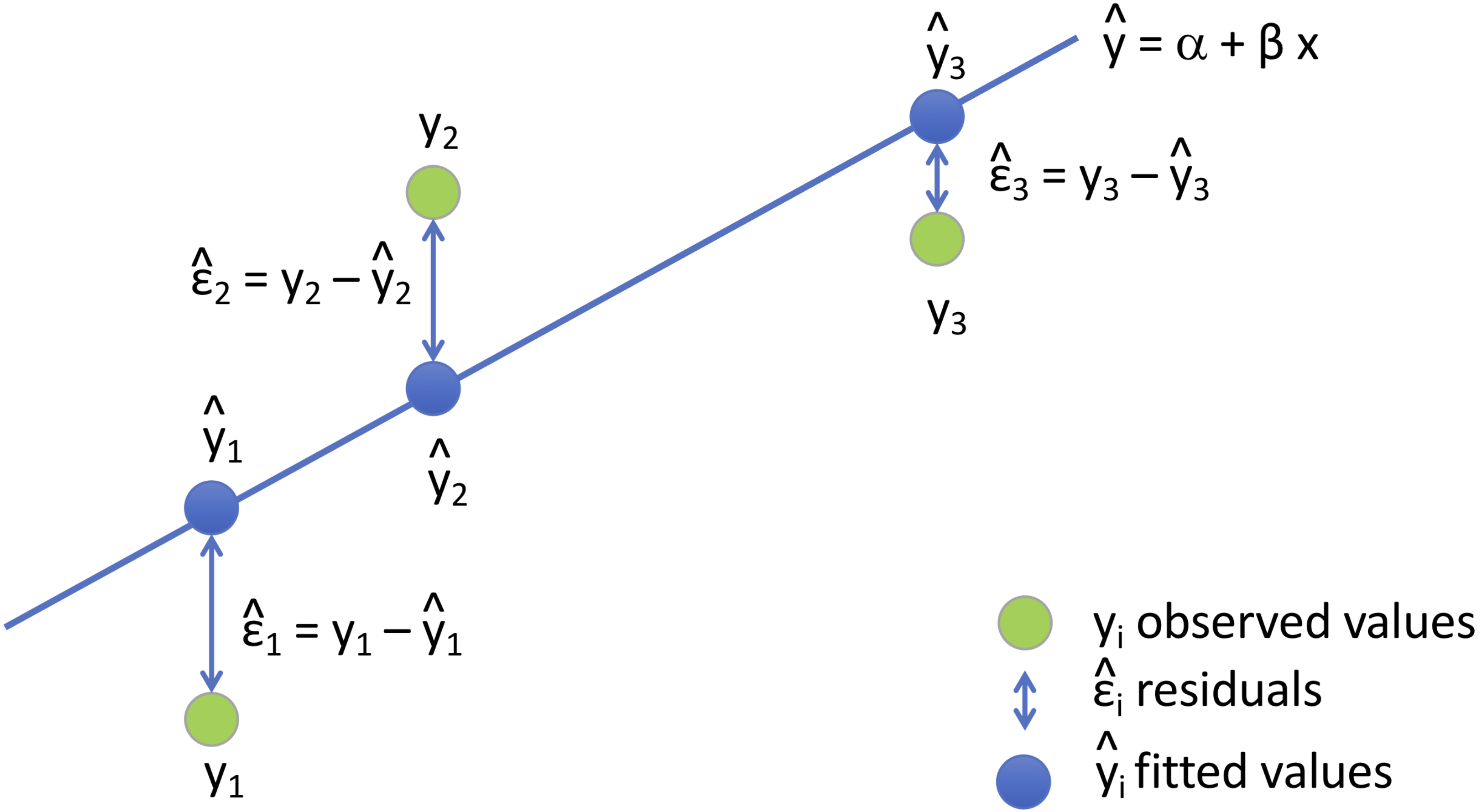


$$M = 389 - 6L$$

# Model specification

> For a data set $(x, y)_i$ with $i \in \{1 \ldots N\}$ the simple linear model is defined as
>
> $$y_i = \alpha + \beta x_i + \epsilon_i$$
>
> with free parameters $\alpha$ and $\beta$ and a random error $\epsilon_i \sim N(0, \sigma^2)$ that is i.i.d. (independently and indentically distributed)

# Model visualized


$\hat{y}_3$
$\hat{y} = \alpha + \beta x$
$y_2$
$\hat{\varepsilon}_3 = y_3 - \hat{y}_3$
$\hat{\varepsilon}_2 = y_2 - \hat{y}_2$
$y_3$
$\hat{y}_1$
$\hat{y}_2$
$\hat{\varepsilon}_1 = y_1 - \hat{y}_1$
$y_1$
$y_i$ observed values
$\hat{\varepsilon}_i$ residuals
$\hat{y}_i$ fitted values

# Model likelihood

$$y_i = \alpha + \beta x_i + \epsilon_i$$

The normal distribution is defined as

$$N(\epsilon, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\epsilon^2}{2\sigma^2})$$

with this model we can now compute the likelihood of the data $(x, y)_i$ with $i \in \{1 \ldots N\}$ as a function of the model parameters $\alpha, \beta, \sigma^2$

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^{N} N(\epsilon_i, \sigma^2) \\ &= \prod_{i=1}^{N} N(y_i - \hat{y}_i, \sigma^2) \\ &= \prod_{i=1}^{N} N(y_i - (\alpha + \beta x_i), \sigma^2) \end{aligned}$$

# Quiz

Which assumption allows to factorize the Likelihood of the data under the linear model

$$y_i = \alpha + \beta x_i + \epsilon_i$$

as

$$L(\alpha, \beta, \sigma^2) = \prod_{i=1}^{N} N(\epsilon_i, \sigma^2)?$$

○ A Independence and identical distribution of the predictors $x_i$

○ B Independence and identical distribution of the responses $y_i$

○ C Independence and identical distribution of the errors $\epsilon_i$

Submit Show Hint Show Answer Clear

# Parameter estimation

**Problem**: How do we find the best parameters of our model?
**Solution**: maximize the (log) likelihood of our data

$$\log(L(\alpha, \beta, \sigma^2)) = -0.5N \log(2\pi\sigma^2) + \sum_{i=1}^{N} -\frac{(y_i - (\alpha + \beta x_i))^2}{2\sigma^2}$$

How to maximize a quadratic function?
We compute gradient and set it to zero, this yields:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

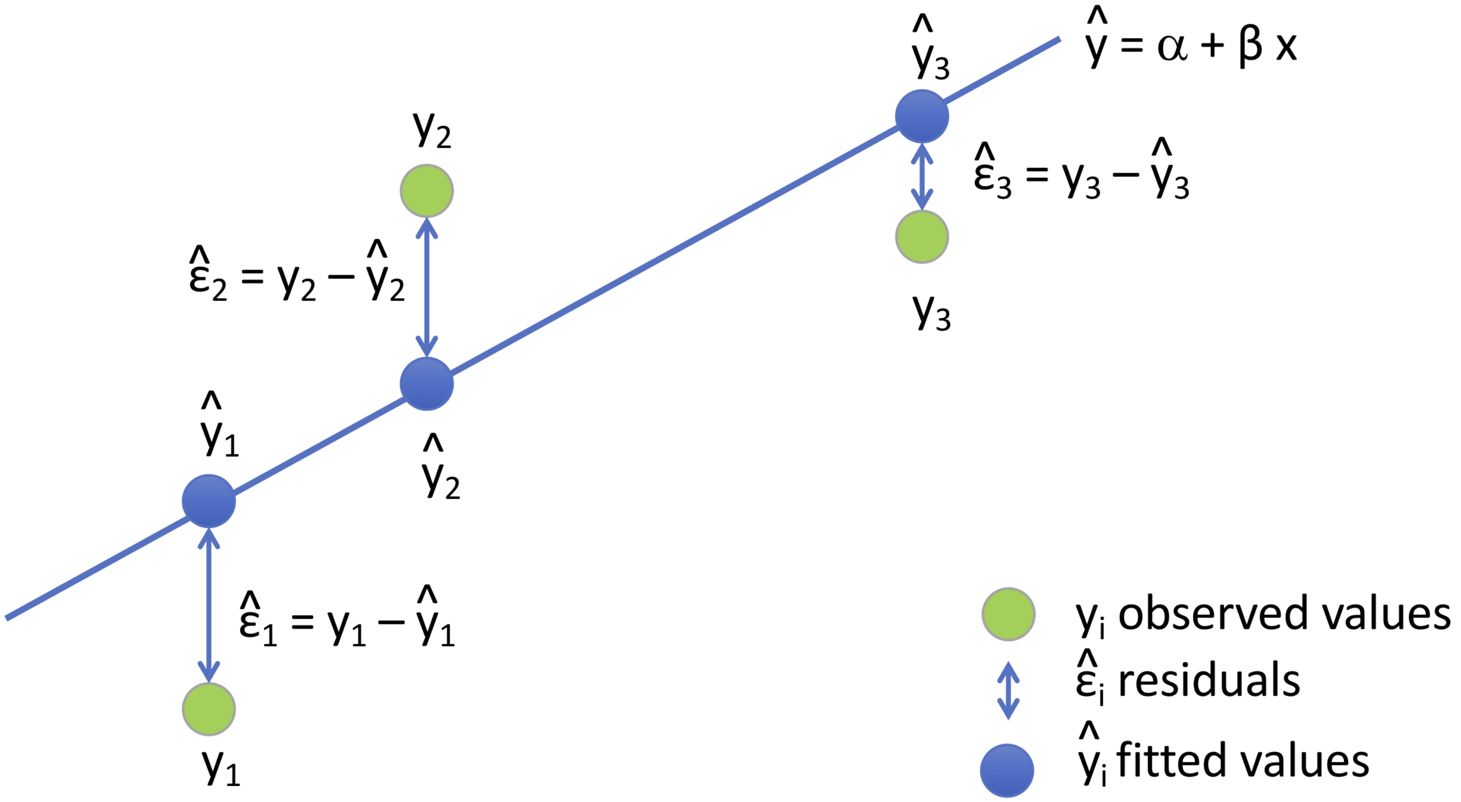
$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - (\hat{\alpha} + \hat{\beta}x_i)^2)$$

with means denoted by $\bar{x}$ and $\bar{y}$.

# Intuition

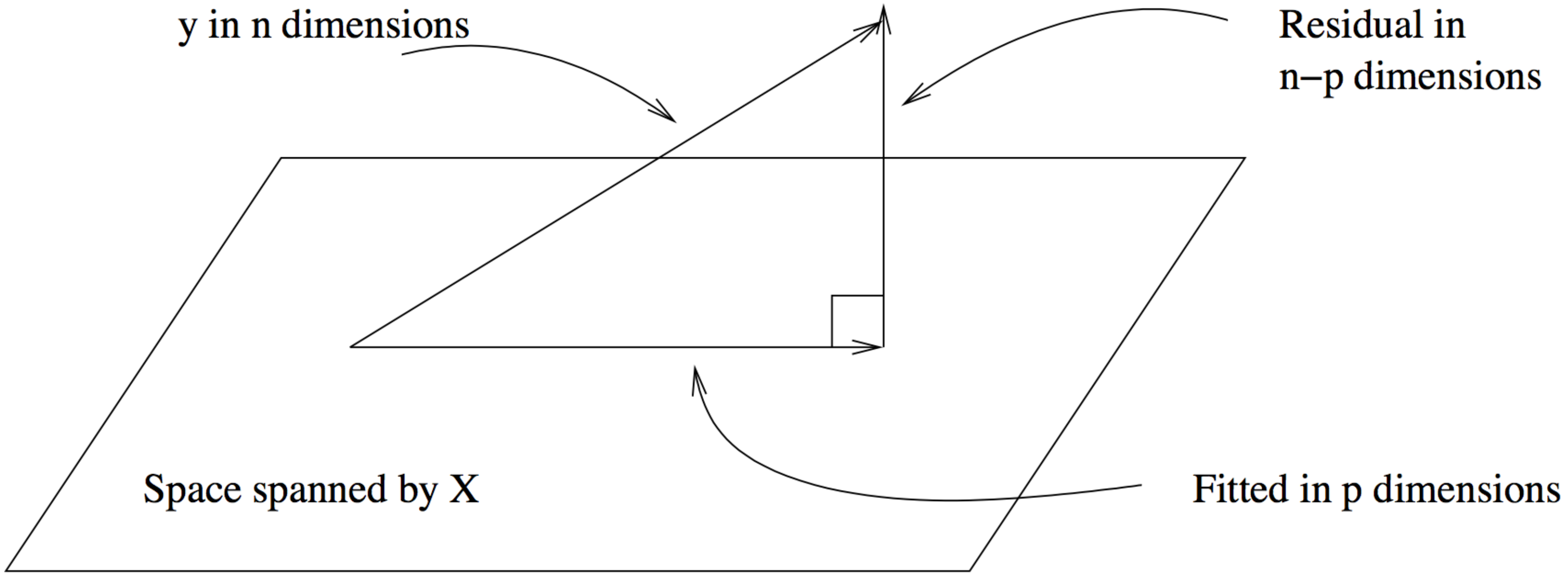
Maximizing the likelihood is actually equivalent to minimizing the residual sum of squares (*RSS*).


$\hat{y} = \alpha + \beta x$
$\hat{y}_3$
$y_2$
$\hat{\varepsilon}_3 = y_3 - \hat{y}_3$
$\hat{\varepsilon}_2 = y_2 - \hat{y}_2$
$y_3$
$\hat{y}_1$
$\hat{y}_2$
$\hat{\varepsilon}_1 = y_1 - \hat{y}_1$
$y_1$
$y_i$ observed values
$\hat{\varepsilon}_i$ residuals
$\hat{y}_i$ fitted values

$$RSS = \sum_{i=1}^{N} (y_i - (\alpha + \beta x_i))^2$$

# Geometric representation

Maximizing the likelihood is actually equivalent to minimizing the squared distance between observation and prediction



# Fit a simple linear model in R

```r
skincancer = read.table("extdata/skincancer.txt", header=T)
head(skincancer)
```

```
##         State  Lat Mort Ocean  Long
## 1     Alabama 33.0  219     1  87.0
## 2     Arizona 34.5  160     0 112.0
## 3    Arkansas 35.0  170     0  92.5
## 4  California 37.5  182     1 119.5
## 5    Colorado 39.0  149     0 105.5
## 6 Connecticut 41.8  159     1  72.8
```

```r
m = lm(Mort ~ Lat, data=skincancer)
coef(m)
```

```
## (Intercept)         Lat
##  389.189351   -5.977636
```

Other useful functions for `lm` objects

- `predict` compute the fitted values or predict response for new data
- `resid` compute the residuals

# How well does our model fit the data?

Any data set will give us estimates of the model, but how good is the model overall?

- Compute model predictions

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

- Compute the *residuals* (compare predictions with the actual values)

$$\hat{\epsilon}_i = \hat{y}_i - y_i$$

- Compute the *residual sum of squares*

$$RSS = \sum_{i=1}^{N} \hat{\epsilon}_i^2$$

- Compare the residual sum of squares to the total sum squares (*SS*) of $y$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = 1 - \frac{RSS}{SS}$$

$R^2$ is called the *coefficient of determination* and represents the percentage of variance explained by the model.

# Quiz

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = 1 - \frac{RSS}{SS}$$

What is the range of values that $R^2$ can take?

○ A -infinity < R2 < infinity

○ B 0 < R2 < infinity

○ C -infinity < R2 < 1

○ D 0 < R2 < 1

Submit Show Hint Show Answer Clear

# Hypothesis testing

Now that we can estimate the parameters and assuming the Gaussian noise model we can ask:

> Is there a linear relationship between $y$ and $x$?

# Quiz

$$y = \alpha + \beta x + \epsilon$$

Which expression would indicate a linear relationship?

○ A alpha = 0

○ B beta = 0

○ C alpha != 0

○ D beta != 0

Submit Show Hint Show Answer Clear

# Hypothesis testing

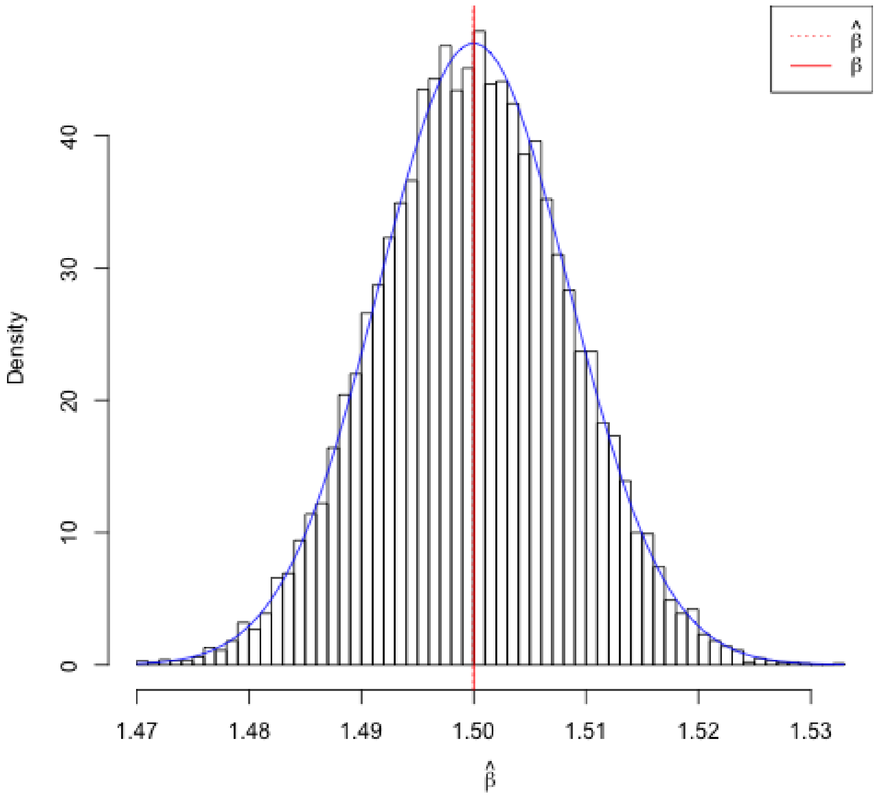
Using the assumption of indenpendent Gaussian noise we can derive the theretical distributions of our estimates.

$$\hat{\beta} \sim N(\beta, \sigma^2/Ns_X^2)$$

where $s_X^2$ is the variance of $X$.
Note that the true value of $\beta$ and $\sigma^2$ are usually not known and need to be estimated. Using the estimate $\hat{\sigma}^2$ to compute the so called standard error $\hat{se}(\hat{\beta}) = \hat{\sigma}^2/Ns_X^2$ we obtain the following distribution

$$\frac{\hat{\beta} - \beta}{\hat{se}(\hat{\beta})} \sim t_{N-2}$$

where $t_{N-2}$ denotes the student's $t$ distribution with $N-2$ degrees of freedom.



# Hypothesis testing: P-value

The P-value of a statistical test is the **probability** of the value of a **test statistic** at least as extreme as the one observed in our data **under the null hypothesis**.

In our case:

- Null hypothesis $H_0 : \beta_0 = 0$
- test statistic is $\hat{t} = \frac{\hat{\beta}-\beta_0}{\hat{se}(\hat{\beta})} = \frac{\hat{\beta}}{\hat{se}(\hat{\beta})}$
- probability under the null model: $P(t \geq \hat{t}) \sim t_{N-2}$

To confirm a liner relation between $y$ and $x$ we need to reject the null hypothesis at significance level $\alpha (= 0.05)$:

- Accept $H_0$ if $P(|t| \geq |\hat{t}|) > \alpha$
- Reject $H_0$ if $P(|t| \geq |\hat{t}|) \leq \alpha$

# Quiz

When would we speak of a linear relationship with $H_0 : \beta = 0$ at significance level $\alpha$?

○ A Reject H0 if P(|t| >|\hat{t}|) > alpha

○ B Reject H0 if P(|t| > |\hat{t}|) < alpha

Submit Show Hint Show Answer Clear

# Hypothesis testing in R

```
m = lm(Mort ~ Lat, data=skincancer)
summary(m)
```

```
##
## Call:
## lm(formula = Mort ~ Lat, data = skincancer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.972 -13.185   0.972  12.006  43.938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 389.1894    23.8123   16.34  < 2e-16 ***
## Lat          -5.9776     0.5984   -9.99 3.31e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.12 on 47 degrees of freedom
## Multiple R-squared:  0.6798, Adjusted R-squared:  0.673
## F-statistic:  99.8 on 1 and 47 DF,  p-value: 3.309e-13
```

# Relation to the *classical* t-test

## Two group t-test with equal variance

- mean group 0: $\mu_0$
- mean group 1: $\mu_1$
- **H0**: $\mu_0 = \mu_1$
- $t = \frac{\bar{X}_0 - \bar{X}_1}{s_p}$
- $s_p$: pooled standard deviation

## Linear model with one indicator variable

- $y = \alpha + \beta x$
- group 0: $x = 0$
- group 1: $x = 1$
- mean group 0: $\mu_0 = \alpha$
- mean group 1: $\mu_1 = \alpha + \beta$
- **H0**: $\beta = 0 \Leftrightarrow \alpha = \alpha + \beta \Leftrightarrow \mu_0 = \mu_1$
- $t = \frac{\beta}{se}$
- $se$: standard error of $\beta$

# Multiple linear regression

# Multiple linear regression

For a data set $(\mathbf{x}, y)_i$ with $i \in \{1 \dots N\}$ and $\mathbf{x}$ a vector of length $p$ the multiple linear regression model is defined as

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

with free parameters $\alpha$ and $\beta$ and a random error $\epsilon_i \sim N(0, \sigma^2)$ that is i.i.d. (independently and indentically distributed)

The model can be written in matrix notation

$$\mathbf{y} = X\beta + \epsilon$$

here the matrix $X$ is of dimension $(N \times p + 1)$ where each row corresponds to the vector $\mathbf{x}$ with a 1 prepended to accomodate the intercept. The error is distributed as $\epsilon \sim N(\mathbf{0}, \Sigma)$ as a multivariate Gaussian with covariance $\Sigma = \sigma^2 I$ (i.i.d).

# Parameter estimation

By the method of maximum likelihood (also for least squares) we obtain

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$$

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}^T \hat{\epsilon}}{N - p}$$

# Nested models and hypothesis testing

Example model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

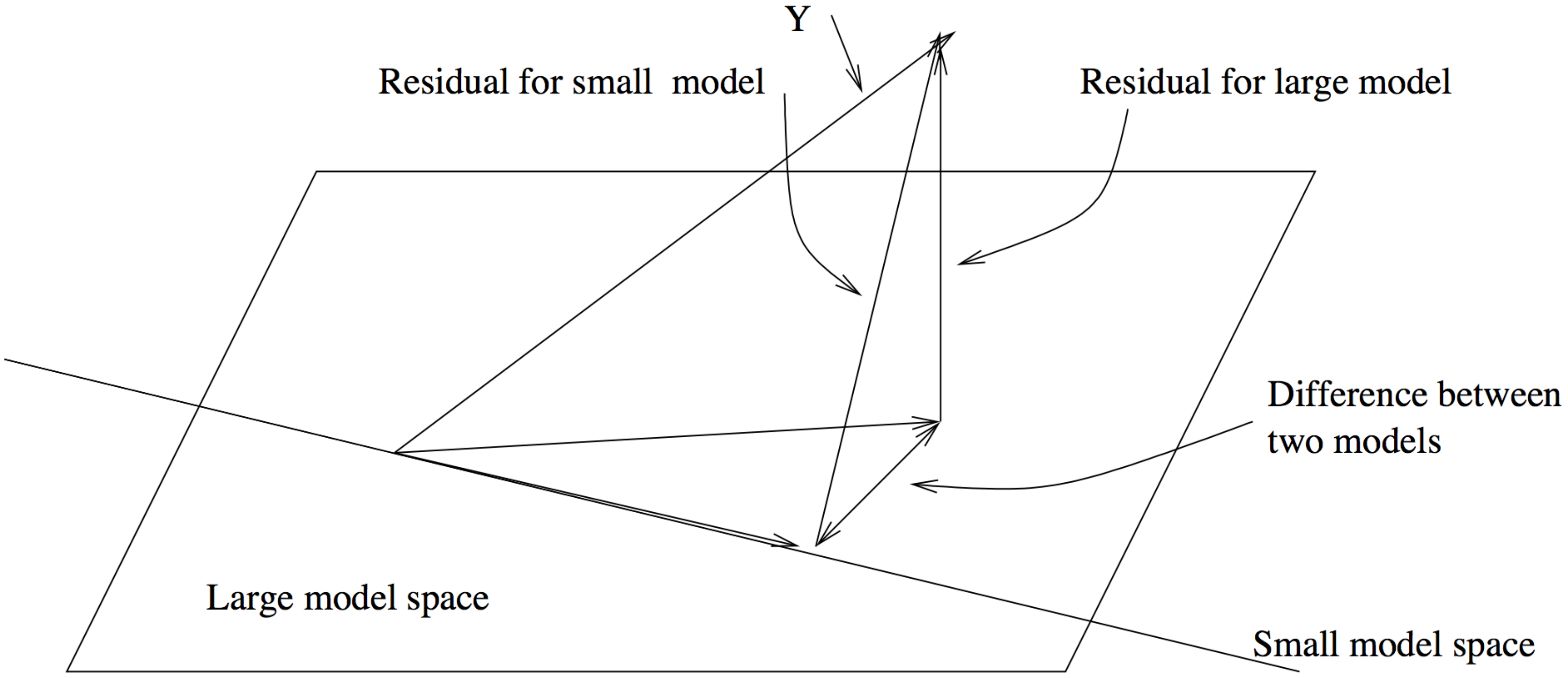
- General concept: **nested models**
- Comparison of a full model $\Omega$ to a reduced model $\omega$
- Model $\omega$ is a special case of the more general model $\Omega$

Examples

- test of individual predictor $x_1$
    - Full model: all $\beta$ can take any value
    - Reduced model: $\beta_1 = 0$ other $\beta$ can take any value
- test of two predictors $x_1$ and $x_2$
    - Full model: all $\beta$ can take any value
    - Reduced model: $\beta_1 = \beta_2 = \beta_3 = 0$ (only the mean $\beta_0$ can take any value)

# Geometric representation

Maximizing the likelihood is actually equivalent to minimizing the squared distance between observation and prediction



# Likelihood ratio test (LRT)

Define the ratio of the two maximized likehoods as test statistic and reject if the ratio is too large

$$\frac{\max_{\beta,\sigma\in\Omega} L(\beta,\sigma)}{\max_{\beta,\sigma\in\omega} L(\beta,\sigma)}$$

Looking at the details we find that $L(\hat{\beta},\hat{\sigma}) \propto (\hat{\sigma^2})^{-n/2}$ , which gives us a test that rejects if

$$\frac{\hat{\sigma}^2_\omega}{\hat{\sigma}^2_\Omega} > \text{a constant}$$

is too large. This is equivalent to

$$\frac{RSS_\omega}{RSS_\Omega} > \text{a constant}$$

$$\frac{RSS_\omega}{RSS_\Omega} - 1 > \text{a constant} - 1$$

$$\frac{RSS_\omega - RSS_\Omega}{RSS_\Omega} > \text{another constant}$$

# Distribution of the LRT

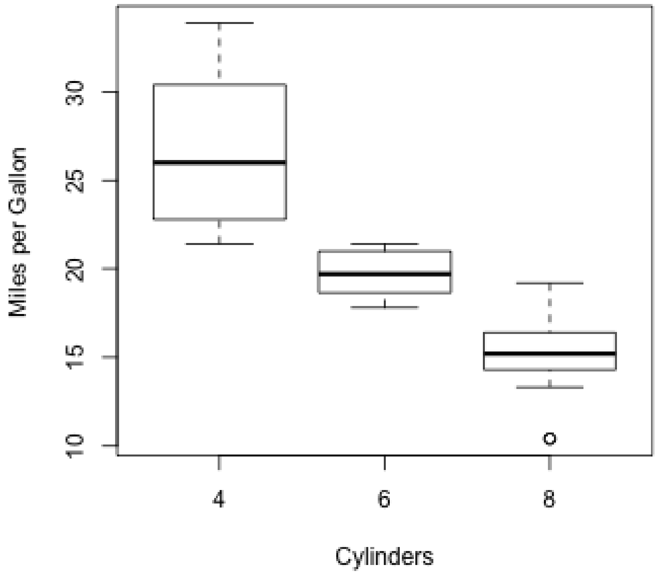
- $q$ number of parameters in (dimension of) model $\Omega$
- $p$ number of parameters in (dimension of) model $\omega$
- test statistic

$$F = \frac{(RSS_{\omega} - RSS_{\Omega})/(q-p)}{RSS_{\Omega}/(n-q)}$$

- $F$ is distributed according to the F distribution with (q - p) and (n - q) degrees of freedom
- Reject the LRT if $F$ is larger that the critital value corresponding to the significance level
- This analysis is also frequently referred to as "Analysis of Variance": **ANOVA**

# Example: testing the difference of means in 3 groups

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$



Indicator variables

- group "6 cylinders": $x_1 = 1$
- group "8 cylinders": $x_2 = 1$

Test the effect of both indicators at the same time

- **H0:** $\beta_1 = \beta_2 = 0$
- Full model: $\Omega$ is the space where all three $\beta$ can take any value
- Reduced model: $\omega$ is the space where only $\beta_0$ can take any value

# Example in R

```
data("mtcars")
## for the example we need a factor
## else it will be interpreted as number
mtcars$cyl <- as.factor(mtcars$cyl)
## fit the full model
full <- lm(mpg ~ cyl, data=mtcars)
## have a look at the model matrix
## which is automatically created
head(model.matrix(full))
```

```
##                   (Intercept) cyl6 cyl8
## Mazda RX4                   1    1    0
## Mazda RX4 Wag               1    1    0
## Datsun 710                  1    0    0
## Hornet 4 Drive              1    1    0
## Hornet Sportabout           1    0    1
## Valiant                     1    1    0
```

```
## fit the reduced model (only the intercept "1")
reduced <- lm(mpg ~ 1, data=mtcars)
```

# Example in R

```
## compare the models
anova(reduced, full)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ 1
## Model 2: mpg ~ cyl
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     31 1126.05
## 2     29  301.26  2    824.78 39.697 4.979e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Adjusting for continuous confounding variables

- in the car example a large fuel consumption might be due to
  - strong engine
  - heavy cars
- many cylinders might be used in heavy cars with strong engines?
  - include the confounders in the reduced model

```
full <- lm(mpg ~ cyl + hp + wt, data=mtcars)
reduced <- lm(mpg ~ hp + wt, data=mtcars)
anova(reduced, full)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp + wt
## Model 2: mpg ~ cyl + hp + wt
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     29 195.05
## 2     28 176.62  1    18.427 2.9213 0.09848 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Model selection

# How do we know which variables need to be in the model?

- We want to explain the data in the simplest way
- Unnecessary predictors will add noise to the estimation of other quantities that we are interested in
- Collinearity is caused by having too many variables trying to do the same job
- Procedures
  - Backward Elimination
  - Forward Selection
- Decision to keep / drop variables based on
  - Hypothesis tests
  - Information criteria (AIC, BIC)
- All procedures are heuristics, so try out!

# Backward elimination

1. Start with all the predictors in the model
2. Remove the predictor with highest p-value greater than $\alpha_{crit}$
3. Refit the model and goto 2
4. Stop when all p-values are less than $\alpha_{crit}$

# Forward selection

This just reverses the backward method.

1. Start with no variables in the model.
2. For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than $\alpha_{crit}$.
3. Continue until no new predictors can be added.

# Diagnostic plots

# Quiz

Which are the most important assumptions of the linear model for hypothesis testing?

○ A Relation between y and x is linear

○ B Errors (residuals) are identically and independently distributed

○ C Errors (residuals) follow a normal distribution

○ D A, B and C

Submit    Show Hint    Show Answer    Clear

# Diagnostic plots

**How to check our assumptions graphically?**

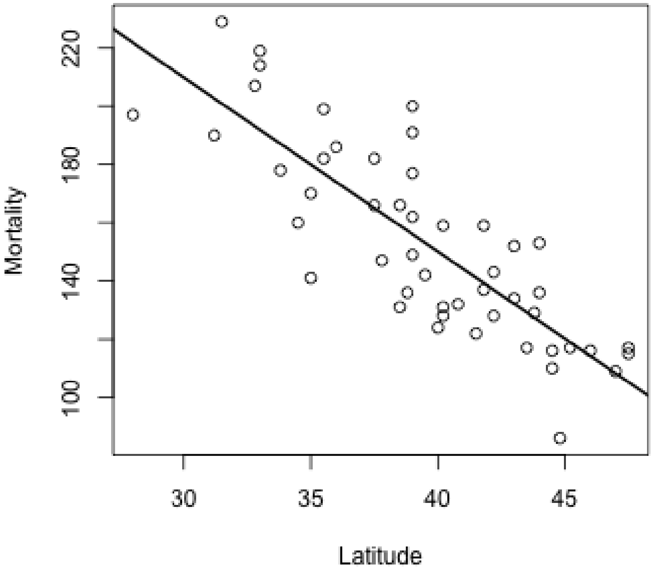
- Relation between $y$ and $x$ is linear
- Errors (residuals) are identically and independently distributed
- Errors (residuals) follow a normal distribution

# Relation between $y$ and $x$ is linear

## Simple linear regression

Scatter plot of $y$ versus $x$

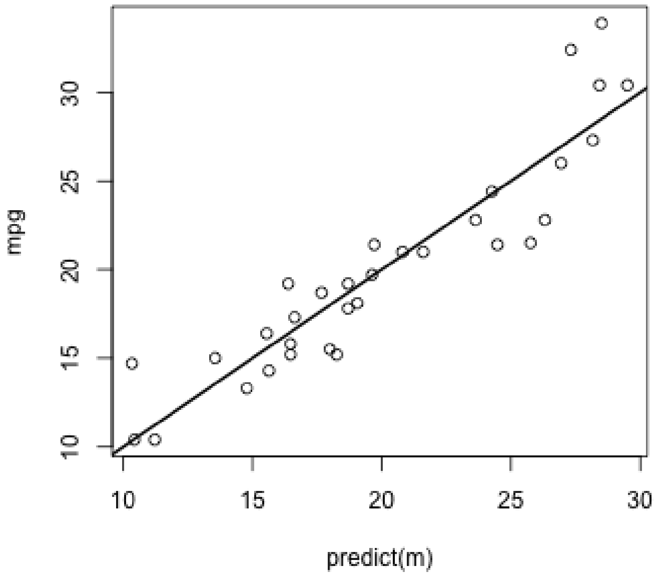
```
plot(Mort ~ Lat, data=skincancer,
xlab="Latitude", ylab="Mortality")
m = lm(Mort ~ Lat, data=skincancer)
abline(m, lwd=2)
```



## Multiple linear regression

Scatter plot of $y$ versus $\hat{y}$

```
m <- lm(mpg ~ cyl + hp + wt, data=mtcars)
plot(mpg ~ predict(m), data=mtcars)
abline(a=0, b=1, lwd=2)
```



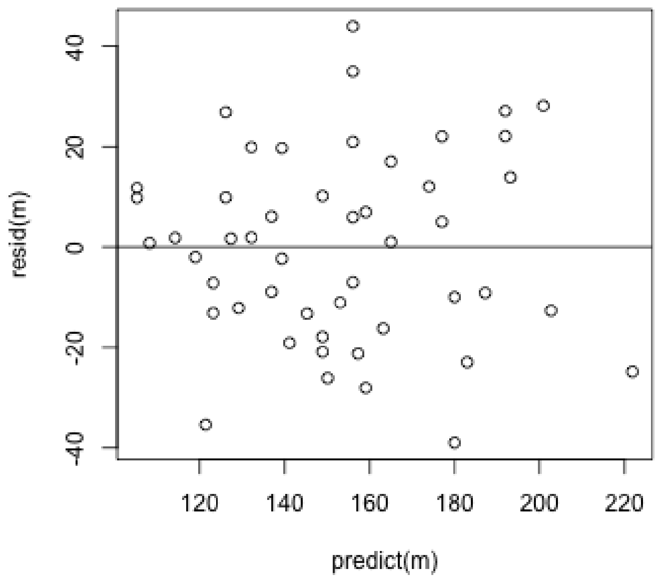
# Residuals are identically and independently distributed

The residuals across all data points come from the same distribution with the same parameters.

- Normal distribution
- Mean $\mu = 0$
- Standard deviation $\sigma$

# Scatter plot of residuals $\hat{\epsilon}$ and predicted values $\hat{y}$

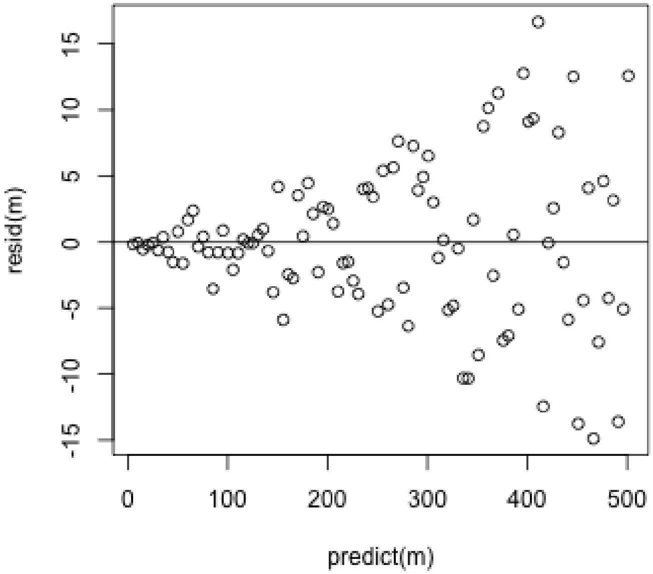
```
m = lm(Mort ~ Lat, data=skincancer)
plot(resid(m) ~ predict(m))
abline(h=0)
```



# What could go wrong?

Variance not constant: **heteroscedascity**

```
x <- 1:100
y <- rnorm(100, mean=5 * x, sd=0.1*x)
m <- lm(y ~ x)
plot(resid(m) ~ predict(m))
abline(h=0)
```



# What to do when the variance is not constant?

- transformation of the response $y$
  - log transformation
  - square root transformation
  - variance stabilizing transformation

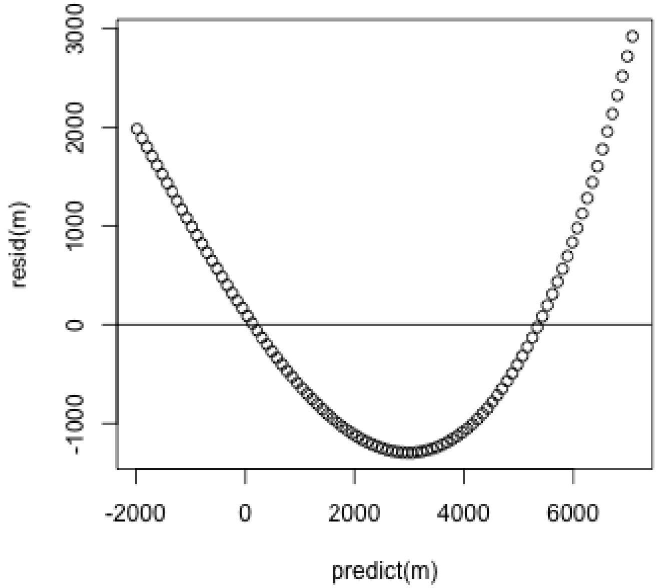
# Consequence of heteroscedascity

- Hypothesis tests are invalid because standard errors of estimates are inconsistent

# What could go wrong?

Relation between $x$ and $y$ is not linear

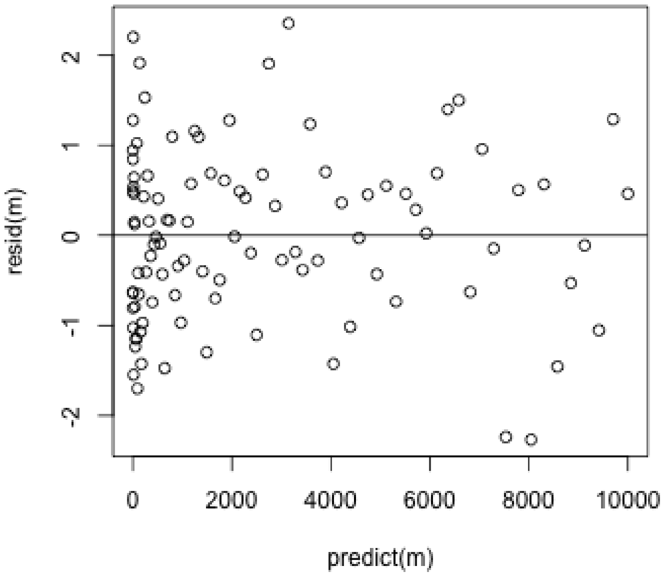
```
x <- 1:100
y <- rnorm(100, mean=0.01 * x^3)
m <- lm(y ~ x)
plot(resid(m) ~ predict(m))
abline(h=0)
```



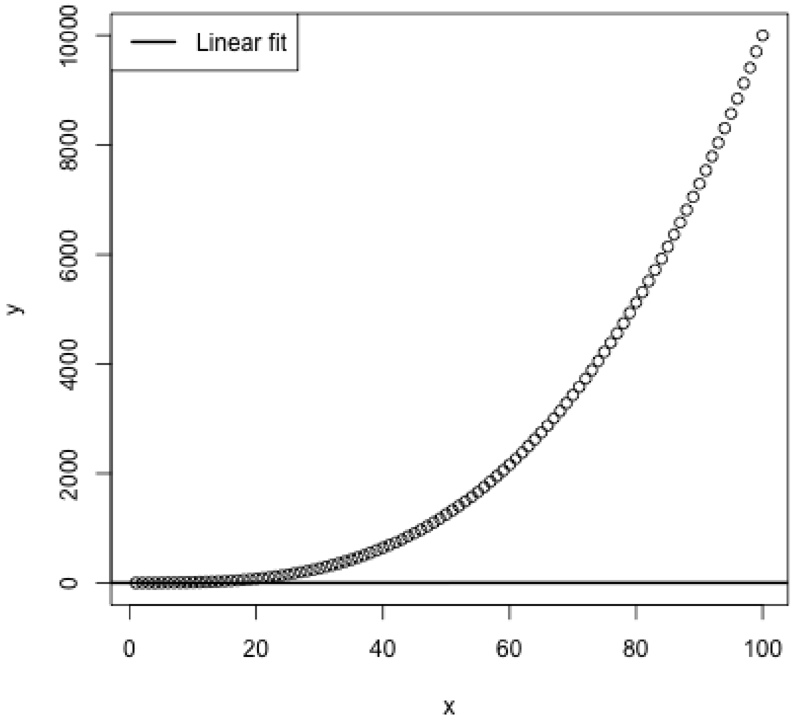
# What to do when relation is not linear?

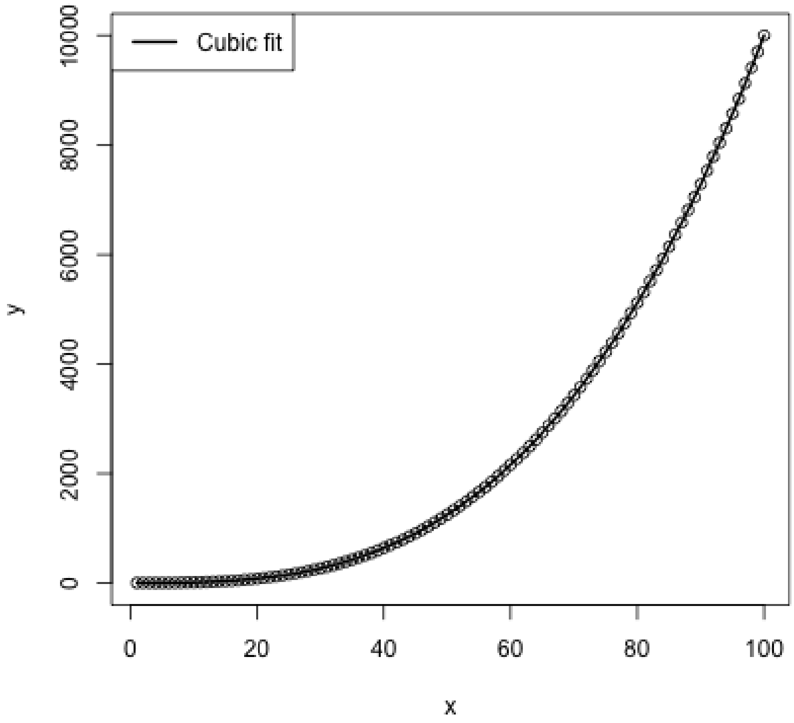
- transformation of the data $x$
  - in our example $x^3$, in practice difficult to know! Try out!

```
x <- 1:100
y <- rnorm(100, mean=0.01 * x^3)
xt <- x^3
m <- lm(y ~ xt)
plot(resid(m) ~ predict(m))
abline(h=0)
```



# Consequence of non-linear relation: bad fit

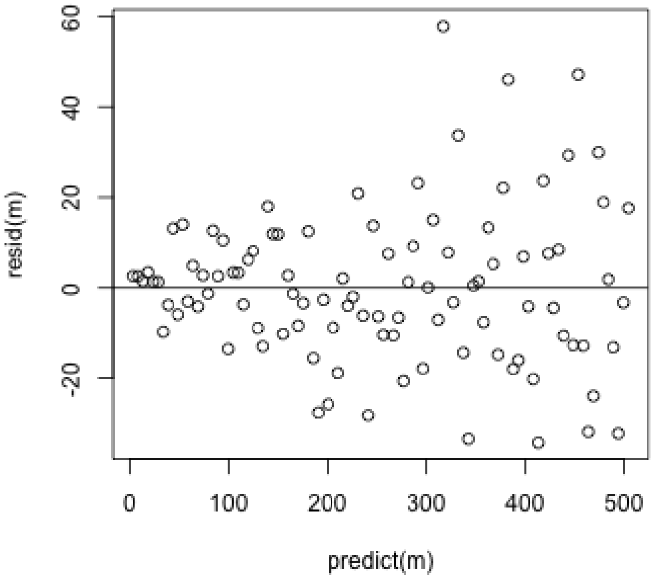




# What could go wrong?

Residuals are not normal

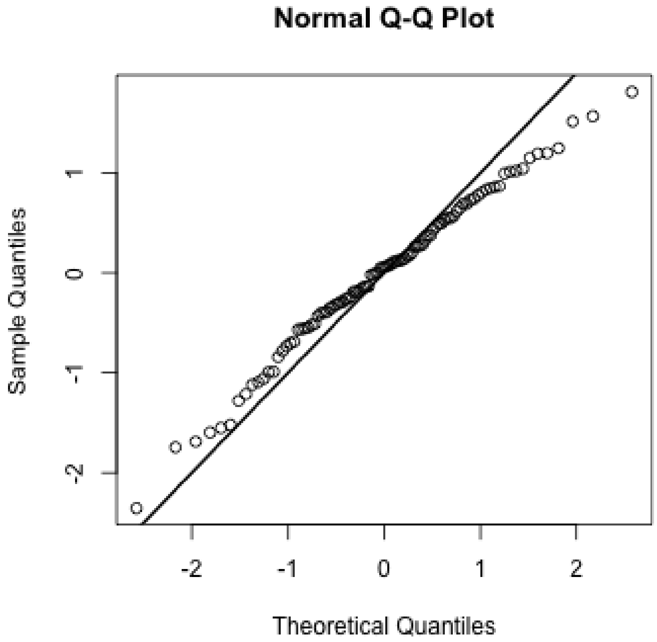
```
x <- 1:100
y <- rpois(100, lambda=5 * x)
m <- lm(y ~ x)
plot(resid(m) ~ predict(m))
abline(h=0)
```



# QQ-Plot to check for normal residuals

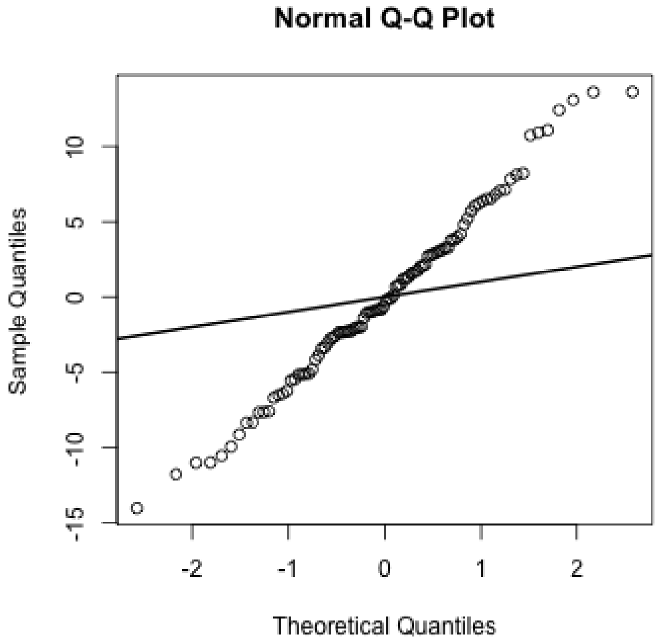
Normally distributed residuals

```
x <- 1:100
y <- rnorm(100, mean=5 * x)
m <- lm(y ~ x)
qqnorm(resid(m))
abline(a=0, b=1, lwd=2)
```



Poisson distributed residuals

```
x <- 1:100
y <- rpois(100, lambda=x)
m <- lm(y ~ x)
qqnorm(resid(m))
abline(a=0, b=1, lwd=2)
```



# Consequence of non-normal residuals

- Hypothesis test may not be valid
- Unreliable error rates
  - True type-I error (false positive) is not $\alpha$
  - True type-II error (false negative) is not $\beta$
- In particular problematic for small data sets

# Summary and conclusion

Linear models

- are a powerful and versatile tool
- can be used to
  - predict future data
  - assess linear relations between variables (hypothesis testing)
  - describe the structure of the data (multiple linear regression)
  - control for confounding (multiple linear regression)
- assumptions need to be checked (diagnostic plots)
- can be generalized for different distributions (GLMs for classification: next lecture)